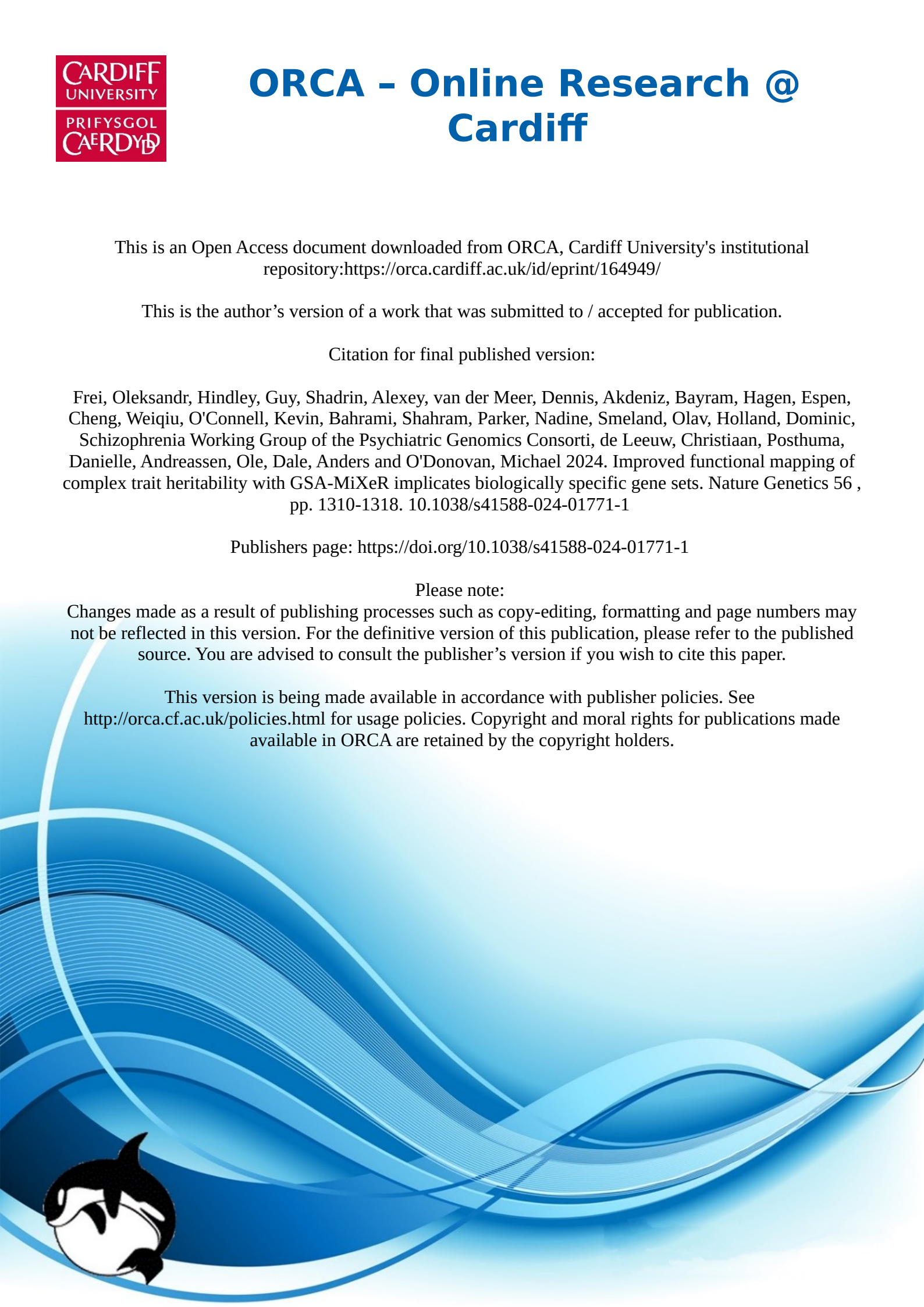


Improved functional mapping with GSA-MiXeR implicates biologically specific gene-sets and estimates enrichment magnitude

Oleksandr Frei[1,2*], Guy Hindley[1], Alexey A. Shadrin[1], Dennis van der Meer[1,3], Bayram C. Akdeniz[1,2], Espen Hagen[1], Weiqiu Cheng[1], Kevin S. O'Connell[1], Shahram Bahrami[1], Nadine Parker[1], Olav B. Smeland[1], Dominic Holland[4], Schizophrenia Working Group of the Psychiatric Genomics Consortium, Christiaan de Leeuw[5], Danielle Posthuma[5], Ole A. Andreassen[1], Anders M. Dale[4]

1. NORMENT Centre, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway
2. Centre for Bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway
3. School of Mental Health and Neuroscience, Faculty of Health, Medicine and Life Sciences, Maastricht University, The Netherlands
4. Center for Multimodal Imaging and Genetics, University of California San Diego, USA
5. Department of Complex Trait Genetics, Centre for Neurogenomics and Cognitive Research, VU University, Amsterdam, The Netherlands

*Corresponding author:
Oleksandr Frei
e-mail: oleksandr.frei@medisin.uio.no

Abstract

While genome-wide association studies (GWAS) are increasingly successful in discovering genomic loci associated with complex human traits and disorders, the biological interpretation of these findings remains challenging. We developed the GSA-MiXeR analytical tool for gene-set analysis (GSA), which fits a model for the heritability of individual genes, accounting for linkage disequilibrium across variants, and allowing the quantification of partitioned heritability and fold enrichment for small gene-sets. We validated the method using extensive simulations and sensitivity analyses. When applied to a diverse selection of complex traits and disorders, including schizophrenia, GSA-MiXeR prioritizes gene-sets with greater biological specificity compared to standard GSA approaches, implicating voltage-gated calcium channel function and dopaminergic signaling for schizophrenia. Such biologically relevant gene-sets, often with less than ten genes, are more likely to provide new insights into the pathobiology of complex diseases and highlight potential drug targets.

Introduction

Genome-wide association studies (GWAS) have discovered thousands of genomic loci associated with complex human traits and disorders, highlighting their polygenic nature and the predominance of small individual effects of common genetic variants[1]. Gene-set analysis (GSA) has become a powerful tool for understanding the biological implications of GWAS findings[2]. Building upon large databases of gene-sets such as the Molecular Signatures Database (MSigDB)[3], Gene Ontology (GO)[4], Synaptic Gene Ontologies (SynGO)[5] and others, GSA has provided many prominent findings[6-8], implicating the role of biological pathways and yielding relevant tissue- and cell type-specific insights into complex human traits and disorders.

Some of the popular analytical tools for GSA analysis of GWAS data are MAGMA[9], Fisher's exact (hypergeometric) test[10], and the stratified linkage disequilibrium (LD) score regression (sLDSC)[11]. MAGMA implements a two-stage approach, where GWAS p-values of single nucleotide polymorphisms (SNPs) are first aggregated into gene-level p-values, which are further tested for association with predefined gene-sets. The hypergeometric test evaluates whether the genes implicated in GWAS are over-represented within predefined gene-sets. Both tools implement competitive GSA[12], testing the null hypothesis that the genes in question are no more strongly associated with the phenotype than other genes. A limitation to both approaches is that they only provide a measure of statistical significance ("enrichment p-value"), which largely depends on the sample size of the GWAS and on the size of a gene-set. Such enrichment p-values do not constitute a biologically meaningful measure of enrichment magnitude, limiting our ability to compare the relative strength of enrichment across or within traits. Importantly, enrichment p-values correlate with gene-set size, tending to be more significant for larger gene-sets, thus limiting our ability to discover more informative biological pathways, with greater specificity to a given phenotype.

The sLDSC method quantifies enrichment by partitioning heritability across genomic regions. Specifically, such partitioning allows one to compute a trait's "fold enrichment" score for a genomic region. This is given by the ratio of the trait's heritability attributed to the region (as estimated by the model) to the heritability of the region estimated from a baseline model which assumes that heritability is uniformly distributed across the genome. The sLDSC method uses GWAS summary statistics to quantify fold enrichments in functional categories. For example, regions conserved in mammals have a disproportionately large contribution to heritability for many traits and disorders[11]. However, for smaller annotations, it has been shown that block jackknife-based significance testing used in sLDSC does not always control type 1 error for annotations with approximately 0.5% of SNPs or less[13]. Furthermore, sLDSC is based on an infinitesimal model with additional simplifying assumptions about the distribution of

genetic effects with respect to minor allele frequency (MAF) and LD[14], which in certain cases has been shown to bias fold enrichment estimates[15]. Finally, the sLDSC method has so far been applied only to tissue- and cell-type-specific annotations, but not to gene ontologies, both due to computational challenges (such as excessive memory usage and runtime) and algorithmic failures (such as non-invertible matrices arising in block jackknife-based significance testing) which has prevented the application of sLDSC to enrichment analysis of gene ontologies.

Here, building on a previously developed MiXeR framework[16-19], we introduce GSA-MiXeR, a novel method for competitive GSA which quantifies partitioned heritability attributed to each gene-set and its fold enrichment with respect to the baseline model (see Online Methods). Its efficient design and implementation, enabled by stochastic gradient-based log-likelihood optimization[20], allow GSA-MiXeR to jointly model more than 18,000 genes in a single model, while accounting simultaneously for the trait's polygenicity, MAF- and LD-dependency of genetic effects, and differential enrichment of functional categories, thus taking into account the unique genetic architecture of each trait. We evaluated the accuracy of GSA-MiXeR's fold enrichment estimates in extensive simulations using real UK Biobank genotypes. We also performed replication analysis using independent GWAS for discovery and replication, which confirmed that ranking gene-sets according to fold enrichment (GSA-MiXeR estimate) often results in a more stable order of gene-sets compared to ranking based on enrichment p-values from MAGMA, even though GSA-MiXeR promoted smaller gene-sets of potentially higher biological relevance. We also performed extensive sensitivity analyses, testing the robustness of the GSA-MiXeR method against a misspecified model of the genetic architecture and out-of-sample LD information.

To demonstrate GSA-MiXeR's performance with real phenotypic data, we applied it to the latest GWAS for schizophrenia[8], as well as to a selection of diverse human traits and disorders[21-35]. In our main analysis, we report GSA-MiXeR results for significant gene-sets first identified by MAGMA. This shows how the addition of fold enrichments enhances the biological interpretation of GSA results, allowing comparisons of gene-set effect size within and across phenotypes. In schizophrenia, GSA-MiXeR reveals that gene-sets related to calcium channel function have greater fold enrichment than larger gene-sets related to post-synaptic functioning. In a subsequent exploratory analysis of schizophrenia without filtering on significant MAGMA associations, we demonstrate that GSA-MiXeR implicates other highly enriched and biologically plausible gene-sets, including biological processes related to dopaminergic neurotransmission, pointing to potential key biological pathways perturbed in schizophrenia, and supporting prior hypotheses of schizophrenia pathogenesis[37]. When applied to 20 additional phenotypes, GSA-MiXeR tended to prioritize small, biologically relevant gene-sets.

Results

Simulation studies

To evaluate the accuracy of the GSA-MiXeR's fold enrichment estimates and standard errors (SEs), we conducted simulations by synthesizing a quantitative trait and its respective GWAS summary statistics using real UK Biobank genotypes (N=337,145 subjects after quality control). We simulated 9 distinct scenarios (see Online Methods and Supplementary Table 1 for an overview) of genetic architectures to validate the sensitivity of the fold enrichment estimates to the underlying assumptions of the GSA-MiXeR model. In simulations without enrichment, total heritability (h2=0.1, 0.4, or 0.7) of the trait was uniformly distributed across the genic regions. To simulate enrichment, a certain number of genes (ngenes=1, 3, 10, 30, or 100) received a 3-, 10- or 30-fold enrichment in heritability attributed to those genes. We used boundaries of real protein-coding genes in these simulations and selected an enriched subset of genes at random. Then we fitted two models: the baseline model with two variance parameters, allowing for different effect size variance in genic regions versus non-genic regions; and the full model, where, in

addition to the baseline model, each gene was modelled with its own effect size variance parameter. The estimated fold enrichment of a gene-set was calculated as a ratio of its heritability estimate from the full model versus the baseline model (see Online Methods).

The results (Figure 1, Supplementary Figure 1, and Supplementary Tables 1-2) show that GSA-MiXeR clearly differentiated between scenarios with 1-, 3-, 10-, and 30-fold enrichment, with estimated fold enrichment of 1.04, 2.67, 8.61, and 23.88, respectively, as measured by an average estimate across all simulated scenarios with a given true fold enrichment. This indicates that GSA-MiXeR estimates are somewhat conservative (biased downwards). However, for the least challenging simulated scenarios, the estimates were more accurate. For example, in six simulated scenarios with sufficiently large gene-sets (ngenes=10, 30, 100), sufficiently powered GWAS summary statistics (corresponding to h2=0.4 and 0.7 in simulations performed with N=337,145 subjects), and the "base" genetic architecture (causal variants uniformly distributed across protein-coding genes), the average fold enrichment estimate from GSA-MiXeR was 1.00, 2.90, 10.75, and 29.35, compared to the expected enrichment of 1, 3, 10, and 30, respectively. These results show that GSA-MiXeR gives accurate results if the underlying assumptions of its model are met, and if the region that is being tested for enrichment has a sufficiently powered signal in the GWAS summary statistics. The SEs generally matched the standard deviation across simulations (Supplementary Table 3); however, for genes with close to zero heritability estimate the hessian-based SE of the $\underline{\sigma_g^2}$ parameter was insufficiently robust and was truncated at 5 times the respective point estimate.

As a sensitivity analysis, we evaluated a "snpXfold" genetic architecture where heritability enrichment of the selected genes was modeled through a higher density of causal variants, while in the "base" architecture, following GSA-MiXeR assumptions, the enrichment was modeled through higher effect sizes. For the "snpXfold" architecture, the average fold-enrichment estimate across the six well-powered scenarios (ngenes=10, 30, 100 and h2=0.4, 0.7) was 0.94, 2.83, 9.90, and 23.76 (versus the expected values of 1, 3, 10 and 30, as above). Other simulated scenarios also represented challenging cases, for example, enrichment in a single gene (ngenes=1), small gene-sets (ngenes=3), testing traits with low GWAS signal (achieved by simulating low heritability of h2=0.1), or testing extreme scenarios of genetic architecture ("mafBelow10", "tldBelow50", "tldBelow25") in some cases lead to underestimating true enrichment by up to 50%. However, the scenarios without enrichment had no inflation in enrichment estimates, with an average fold enrichment close to 1.0.

We performed additional simulations to compare GSA-MiXeR and MAGMA's ability to sort gene-sets in a way that promotes enriched gene-sets (Figure 1, Supplementary Figure 2, Supplementary Table 4). For each of the previously conducted simulations, we generated 999 additional gene-sets, each with the same number of genes as the enriched gene-set. We then computed the position of the truly enriched gene-set on the list, sorted either according to GSA-MiXeR fold enrichment or according to MAGMA p-values. Both methods correctly rank the most enriched gene-sets at the 1st position if there was sufficient signal in the data, e.g., in a scenario with ngenes=100 and h2=0.7. For the "base" genetic architecture, GSA-MiXeR outperforms MAGMA in the less-powered scenarios. For instance, in simulations with 10x fold enrichment, ngenes=10, and h2=0.4, the average rank of the truly enriched gene-set was 1.1 for GSA-MiXeR, and 22.7 for MAGMA. However, for the "snpXfold" genetic architecture, GSA-MiXeR did not outperform MAGMA (Supplementary Figure 2b), showing that its advantage in ranking accuracy over MAGMA depends on how well the underlying assumptions about the trait's genetic architecture are met. At the same time, some genetic architectures (e.g., "CodingUCSC", "PromoterUCSC", "mafBelow10", "tldBelow50", "tldBelow25") where we previously observed up to 50% lower (conservative) fold enrichment estimates in GSA-MiXeR analysis, still show an advantage of GSA-MiXeR over MAGMA, ranking the truly enriched gene-set more accurately. The individual genes (ngenes=1) were generally ranked at approximately the 100-th position or higher by both MiXeR and MAGMA, indicating that the error of

estimating gene-level fold enrichment currently appears to be too large to make reliable inferences at the level of individual genes.

Application to real phenotypic data

We applied GSA-MiXeR to the latest GWAS of schizophrenia[39], as well as a selection of 20 traits and disorders representing a variety of different human phenotypes (Table 1), including somatic disorders, quantitative biophysical and biochemical measures, and mental traits. In our main analysis, we first identified all enriched gene-sets with MAGMA p-values below 0.05 after Bonferroni correction. Next, we re-ranked gene-sets based on GSA-MiXeR fold enrichment and compared them to the original MAGMA p-value-based ranking. We also compared fold enrichment and partitioned heritability estimated by GSA-MiXeR within and across phenotypes.

When applied to schizophrenia (Figure 2, Supplementary Table 5), the top ten GSA-MiXeR gene-sets had on average 236 genes, with an average fold enrichment of 2.49 (root mean square error (RMSE) 0.50) (Supplementary Table 5), versus an average of 620 genes and fold enrichment of 1.53 (RMSE 0.22) for the top ten MAGMA gene-sets. Interestingly, among the significant MAGMA gene-sets, the top four gene-sets after ordering based on GSA-MiXeR enrichment for schizophrenia were related to voltage-gated calcium channels and membrane depolarization (Figure 2), which are more specific biological processes compared to the larger synaptic and post-synaptic gene-sets prioritized by MAGMA. Furthermore, voltage-gated calcium channels gene-sets were approximately ten times more fold enriched than other MAGMA identified gene-sets (mean fold enrichment = 4.05 (RMSE 0.78) vs 1.33 (RMSE 0.09), respectively), with a marked drop in fold enrichment to the next most fold enriched gene-set (intrinsic component of post-synaptic membrane, fold enrichment 1.58 (SE 0.16). The larger enrichment was partly attributable to the *CACNA1C* gene which was the most enriched gene in all related gene-sets. To evaluate heterogeneity of gene fold enrichments within implicated gene-sets, we performed a leave-one-gene-out analysis, showing that the average fold enrichment across voltage-gated calcium channels gene sets reduced from 4.05 (RMSE 0.78) to 2.89 (RMSE 0.85) after exclusion of the *CACNA1C* gene (Supplementary Table 5), but the gene-sets remained enriched. The purpose of the leave-one-gene-out analysis is to evaluate heterogeneity of gene fold enrichments within implicated gene-sets, rather than to estimate uncertainty of the enrichment estimates.

In an exploratory analysis we considered all gene-sets with a positive value of GSA MiXeR's Akaike Information Criterion (AIC) (see Online Methods). For schizophrenia this implicated 36 gene-sets (Figure 3, Supplementary Table 7), with an average size of approx. 8 genes and fold enrichment of 7.28 (RMSE 3.41). These included several gene-sets related to dopaminergic neurotransmission, such as "dopamine neurotransmitter receptor activity" and "dopamine binding"[40], as well as "interneuron migration"[41] and "forebrain neuron fate commitment", none of which were identified by MAGMA. However, the enrichment in all of the top ten gene-sets besides "forebrain neuron fate commitment" was largely driven by the *DRD2* gene, rather than by the joint effect of the entire gene-set. After exclusion of DRD2, the average enrichment of these nine gene-sets reduced from 14.9 (RMSE 6.59) to 0.91 (RMSE 0.83) as opposed to 8.53 (SE 2.66) to 6.22 (2.26) for "forebrain neuron commitment" after exclusion of its most enriched gene "SATB" (Supplementary Table 7). This suggests large heterogeneity in enrichment across genes within highly enriched gene-sets, further demonstrated by QQ plots constrained to genetic variants within 10 kilobases (KB) up/down of each gene (Supplementary Figure 3). This revealed "null" QQ plots for a subset of genes, while other genes – particularly *DRD2* and *CACNA1C* – showed strong enrichment. Further, despite the high fold enrichments of the gene-sets discussed here, it is important to note that they individually contribute to less than 0.5% of schizophrenia's heritability, further emphasizing its broad polygenic architecture.

Among the 20 other phenotypes that were analyzed, thirteen were associated with significantly enriched gene-sets as estimated by MAGMA (Table 1, Supplementary Table 6), while all phenotypes had

at least one gene-set with a positive GSA-MiXeR AIC value (Table 1, Supplementary Table 8; see also Supplementary Note 2 for more extended description of the results). Overall, the gene-sets identified by MAGMA had a mean size of 400 genes, versus a mean size of 9 genes for gene-sets implicated by GSA-MiXeR. The average fold-enrichments were 3.5 (RMSE 1.42) and 10.3 (RSME 4.83), respectively. Additionally, GSA-MiXeR enabled cross-trait comparisons of gene-set enrichments. Across all 21 phenotypes there was large variation in the mean fold enrichment, ranging from 2.85 (RMSE 0.80) for educational attainment to 31.27 (RMSE 11.55) for heart failure. This was to some extent influenced by the number of gene-sets implicated by GSA-MiXeR for each phenotype, varying from a single gene-set for alcohol consumption, heart failure and hospitalized COVID, to 161 gene-sets for type 2 diabetes, 202 for systolic blood pressure, and 405 for height (Table 1, Supplementary Table 8). Differences in the number of gene-sets identified likely reflect the statistical power of the input GWAS as well as the extent to which the current set of Gene Ontology gene-sets corresponds to the biological mechanisms underlying each phenotype. Nevertheless, there was a trend for biochemical measures and somatic disorders, such as glycated haemoglobin and ulcerative colitis, to exhibit higher fold enrichments than complex mental phenotypes such as educational attainment and cognition.

We also compared GSA-MiXeR and LDAK-GBAT[43] estimates of heritability for individual genes (Supplementary Table 9-10 and Supplementary Figure 9). For a subset of 36 gene-sets implicated in the exploratory GSA-MiXeR analysis of schizophrenia, we also estimated fold enrichment with the stratified LDSC[44] method (Supplementary Table 11 and Supplementary Figure 10). The results from the different methods generally correlate well. There are certain discrepancies in the estimates for individual genes and gene-sets, which likely can be attributed to the differing methodology compounded by large uncertainty in the estimates. For example, LDAK-GBAT analyses all genes independently from each other, while GSA-MiXeR applies a joint model. Similarly, sLDSC uses an additive model across all categories (in this case 75 functional annotations plus additional categories for the gene-set and its genes), while GSA-MiXeR's heritability model separately accounts for functional annotations and genes, joining them as two factors in a multiplicative model.

Sensitivity analysis

We performed additional analyses to test the sensitivity of our results to GSA-MiXeR modeling assumptions, using real data GWAS on schizophrenia and height. For the enrichment model, we compared the three alternatives: (A) the default GSA-MiXeR model, allowing effect size variance to depend on genes; (B) a "gene-set" model, allowing effect size variance to depend on gene-sets, but not on individual genes; and (C) a "hybrid" model allowing effect size variance to depend on both genes and gene-sets. We expected that model (B) would yield less noisy results, at the cost of higher model misspecification as it doesn't model heterogeneity across individual genes. Another downside is that model (B) must deal with overlap across gene-sets, which we handle through an additive model (see Online Methods). Model (C) was expected to combine the flexibility of model (A) with the more robust inference of parameters in model (B). The results are presented in Supplementary Figure 4, also showing the difference in log-likelihood values of the models being compared. Models (A) and (C) yield similar results, with log-likelihoods suggesting that modeling enrichments at the gene level is superior to modeling enrichments at the gene-set level, in line with observations highlighted in Supplementary Figure 3. Our choice of model (A) over model (C) for the final analysis was to ensure that the enrichment estimate for a particular gene-set depends only on the set of genes included in the analysis and is independent of other gene-sets.

Further comparison of two baseline models (with and without accounting for enrichment in functional categories) showed that these two approaches are generally consistent, with slight inflation in estimates for schizophrenia and height if the baseline model does not account for functional categories (Supplementary Figure 5). This suggests that controlling for such enrichment is important, otherwise,

some gene-sets might appear enriched simply due to a higher content of enriched functional categories. We also evaluated an alternative strategy for capturing MAF- and LD-dependent architecture, using MAF- and LLD-bins, as modeled in sLDSC[45], instead the GSA-MiXeR's model using S- and L- parameters, showing that the fold enrichment estimates were very similar between these two models. However, the GSA-MiXeR's model had a better log-likelihood as compared to the model that explicitly includes MAF- and LLD- bins. Additional sensitivity analyses are presented in Supplementary Figure 6, showing only minor effects from applying an out-of-sample LD reference.

Replication analysis

In our analysis of real GWAS data, we first used MAGMA to filter on significantly associated gene-sets, followed by GSA-MiXeR to re-rank these gene-sets. Here, we present the results of an independent replication analysis using either MAGMA p-values or GSA-MiXeR fold enrichments to rank gene-sets. As a discovery sample, we used half of the UK Biobank panel for the height phenotype and half of the Psychiatric Genomics Consortium (PGC) schizophrenia sub-studies. The remaining subjects or studies were used as replication data sets. Using independently obtained results for MAGMA and GSA-MiXeR from the discovery and replication samples, we computed the fraction of top-N gene-sets in the discovery samples that also appeared within the top-N gene-sets in the replication samples, allowing N to be 10, 20, 50 or 100 gene-sets. Further, we investigated how the replication rate depended on gene-set size, by computing the replication rate for a stratum of gene-sets with up to 25 genes per set, and a second stratum of gene-sets with more than 25 genes per set.

The results are presented in Figure 4 (for schizophrenia) and in Supplementary Figure 7 (for height). GSA-MiXeR and MAGMA methods had a comparable replication rate when evaluated on all gene-sets, or in a stratum with more than 25 genes. The biggest differences were observed in the stratum of up to 25 genes, where we consistently observed better replication rates for GSA MiXeR ranking, compared to MAGMA ranking, both for schizophrenia and for height. Overall, the results indicate that replication of GSA-MiXeR-based ranking is equivalent to MAGMA-based ranking for all gene-sets and outperforms MAGMA-based ranking for smaller gene-sets. We also demonstrated that GSA-MiXeR may potentially improve the accuracy of polygenic risk scores in terms of Nagelkerke pseudo-R2 (Supplementary Figure 8, also Supplementary Note 2), however further evaluation of this approach is a subject of future work.

Discussion

Here, we present GSA-MiXeR, a novel method for competitive gene-set enrichment analysis that applies stochastic gradient-based optimization for maximum likelihood estimation from GWAS summary statistics. We show that GSA-MiXeR can accurately model the contribution of more than 18,000 genes to SNP-heritability of complex polygenic traits, while controlling for polygenicity, MAF- and LD-dependent genetic architecture, and functional categories. This enables the robust estimation of gene-set fold enrichment alongside partitioned heritability, thus estimating the magnitude of the enrichment, which was previously unfeasible due to computational and methodological limitations. While future work improving the validity of current gene-set definitions is required to fully exploit the power of GSA-MiXeR and GSA more generally, we show through real data from schizophrenia and 20 other traits how GSA-MiXeR, building on current GO gene-sets, enhances the biological interpretation of GWAS data.

Applying GSA-MiXeR to simulated and real data, we show its capability to reorder gene-sets in a way that promotes smaller gene-sets while yielding an equivalent or higher replication rate compared to current standards in the field. When applied to sufficiently powered GWAS, GSA-MiXeR provides an unbiased estimate of fold enrichments even for small gene-sets with 10 genes or less. GSA-MiXeR's ability to estimate fold enrichment of a gene-set, rather than its statistical significance (p-value), represents a key advance in GSA, allowing direct comparison of the relative biological effect of implicated gene-sets.

This facilitates the promotion of smaller gene-sets with more specific and well-defined functions that can be more readily interrogated in downstream experimental analysis.

When applied to schizophrenia, we demonstrate how the estimation of gene-set fold enrichments can enhance MAGMA-based GSA. GSA-MiXeR showed that gene-sets related to voltage-gated calcium channels were approximately ten times more fold-enriched than other larger gene-sets identified by MAGMA. This supports previous evidence indicating a role for voltage-gated calcium channels in schizophrenia, which has been hypothesized to affect schizophrenia risk through neurodevelopmental processes including hippocampal neurogenesis[47]. GSA-MiXeR's potential for novel mechanistic insights was further illustrated in our exploratory analysis without filtering on significant MAGMA associations. Further, despite recent advances in schizophrenia GWAS[8,48], results from GSA analyses have not yet mapped closely onto the primary clinical theories of schizophrenia pathoetiology[37]. It is therefore striking that the top two most fold-enriched gene-sets were related to dopaminergic neurotransmission, the leading theory of schizophrenia pathogenesis[49-51]. In addition, GSA-MiXeR implicated several other relevant gene-sets including "interneuron migration" which may affect parvalbumin-positive GABAergic interneurons that have been implicated in schizophrenia pathogenesis[52], and "forebrain neuron fate commitment"[53]. It is also worth noting that enrichment in the dopamine-related gene-sets appears to be driven largely by a single gene, *DRD2*. By contrast, the high enrichment of the gene-sets related to voltage-gated calcium channels and "forebrain neuron fate commitment" appears to be a property of multiple genes in each gene-set. When applied to 20 diverse human phenotypes, GSA-MiXeR also tended to prioritize smaller gene-sets of high biological relevance.

Previous methods, such as LDAK-GBAT[43] and h2gene[58], have estimated heritability attributed to individual genes, identifying individual genes that significantly contribute to trait's heritability, but without considering gene-set and without computing fold enrichment over a baseline. Existing methods also estimate disease heritability mediated by the cis genetic component of gene expression levels and its enrichment across gene-sets (MESC[59] and GCSC[60] methods), leveraging expression imputation from genetic data[61]. While these methods found both overall evidence of expression-mediated heritability and specific gene-sets showing disease-relevant enrichment, they have only accounted for a small fraction of estimated SNP-heritability. This can be either due to insufficient accuracy of gene expression imputation potentially related to low sample size of expression quantitative trait loci datasets for trait-relevant tissues and cell types, or due to other causal pathways beyond gene expression contributing to shaping genotype-phenotype associations. GSA-MiXeR complements these methods by estimating the fraction of the total SNP-based heritability attributed to a gene-sets without relying on gene expression data. Other methods, such as AI-MiXeR[18], RSS-E[62], and RSS-NET[63] have also computed enrichment in certain regions of the genome. However, they do not build a comprehensive model including all known human genes, a key advantage of GSA-MiXeR. To achieve this, the GSA-MiXeR model incorporates several parameters describing the distribution of additive genetic effects, fitting those parameters from the GWAS summary statistics. Furthermore, GSA-MiXeR evaluates fold enrichment of a gene-set versus a comprehensive baseline model, which accounts for differential enrichment of functional categories yet without gene-specific effects – an enrichment model which cannot be implemented using a fixed-effects approach[64].

The GSA-MiXeR tool has some limitations. First, it was not feasible to provide formal p-values due to technical reasons and difficulties in defining the null-hypothesis (see Supplementary Note 2), thus GSA-MiXeR relies on the MAGMA tool to pre-filter the set of gene-sets for the most conservative analysis. Our exploratory analysis (without pre-filtering by MAGMA) selects gene-sets based on AIC criteria, which does not allow for multiple testing correction; we however confirmed that ranking gene-sets according to GSA-MiXeR fold enrichment is at least as stable as ranking according to conservatively defined MAGMA p-values; additionally, all estimates have SEs to evaluate their uncertainty. Second, SEs are derived from the likelihood function, and may in some cases be not well calibrated, particularly for genes with close to zero

heritability estimate. Real-data analysis with GSA-MiXeR is unlikely to be affected by this due to filtering genes on a positive AIC value, which implies sufficient curvature of the log-likelihood around the MLE point and justifies hessian-based SEs estimation. Applying resampling approaches, such as bootstrap of jackknife, may further improve robustness of the SEs, however for GSA-MiXeR setting these techniques should resample individual subjects rather than SNPs, i.e. resampling should be implemented as an integral part of the GWAS analysis, leading to multiple instances of the GWAS summary statistics which can be later used to assess the uncertainty of model parameters. Third, GSA-MiXeR is based on certain assumptions, such as modeling enrichment through varying the effect sizes rather than varying the density of causal variants; we validated through sensitivity analysis that violation of these assumptions does not lead to significant biases. Finally, GSA-MiXeR does not handle continuous-valued genomic annotations as implemented in extended sLDSC[44]. In the future we are planning to extend GSA-MiXeR analysis to cover genes located on non-autosomes (including sex chromosomes and mitochondrial DNA), and provide a reference panel for GSA-MiXeR analysis for non-European populations.

To conclude, GSA-MiXeR estimates fold enrichment and implicates gene-sets with higher biological specificity than current standards. The gene-sets promoted by GSA-MiXeR can provide new insights into the pathobiology of complex diseases, with the potential for identifying new drug targets and the development of pathway-specific polygenic risk scores. This may help to advance the classification, diagnosis, and treatment of complex polygenic disorders.

Data availability

The datasets analyzed during the current study are available for download from the following URLs: GWAS from Psychiatric Genomics Consortium, https://www.med.unc.edu/pgc/results-and-downloads/downloads; GWAS on UK Biobank, https://github.com/Nealelab/UK_Biobank_GWAS; HRC reference data, https://ega-archive.org/datasets/EGAD00001002729 (upon application); UK Biobank accessed via application 27412, https://bbams.ndph.ox.ac.uk/ams/ (upon application); 1000 Genomes Phase3 data, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/; MsigDB v7.5, https://www.gsea-msigdb.org/gsea/msigdb/; functional categories, https://alkesgroup.broadinstitute.org/LDSCORE/.

Ethics statement

Each sample was collected with the participants' written informed consent and with approval by local institutional review boards. The use of summary statistics for genetic analysis was evaluated by The Norwegian Institutional Review Board: Regional Committees for Medical and Health Research Ethics (REC) South-East Norway and found that no additional ethical approval was required because no individual data were used. Individual level data from the TOP study were analyzed under ethical approvals from Norwegian REC (ref. 2009/2485), Data Inspectorate (ref. 03/02051), and The Norwegian Directorate of Health (ref. 05/5821).

Code availability

The MiXeR software and a tutorial example on how to use it will be available online upon publication (https://github.com/precimed/gsa-mixer, v2.0); PLINK software, https://www.cog-genomics.org/plink/2.0/; MAGMA software v1.09b, https://ctg.cncr.nl/software/magma; pipeline to simulate synthetic GWAS data from genotypes: https://github.com/precimed/simu (v0.9.3); https://github.com/ofrei/ldsc/tree/disable_jackknife - modified sLDSC software; https://dougspeed.com/ldak-gbat/ - LDAK v5.2 software.

Acknowledgements

The authors were funded by the Research Council of Norway (#324499, #324252, #273291, #223273, #276082), EU H2020 RealMent (#964874), KG Jebsen Stiftelsen, South East Norway Health Authority (#2022-073). This research has been conducted using the UK Biobank Resource under Application Number 27412. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), including Expanse and OASIS resources provided by San Diego Supercomputer Center at UC San Diego through allocation IBN200001. This work also used the TSD (Tjeneste for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo, IT-Department (USIT, tsd-drift@usit.uio.no), using resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway. C.d.L. was funded by Hoffman-La Roche.

Author's contributions

O.F. conceived the study; O.F., A.A.S., and K.O. pre-processed the data; O.F., D.v.d.M and E.H. performed all analyses, with conceptual input from G.H., C.d.L., O.A.A. and A.M.D; O.F. and G.H. drafted the manuscript; all authors contributed to and approved the final manuscript.

Competing financial interests

Dr. Dale is a Founder of and holds equity in CorTechs Labs, Inc, and serves on its Scientific Advisory Board. He is a member of the Scientific Advisory Board of Human Longevity, Inc. and receives funding through research agreements with General Electric Healthcare and Medtronic, Inc. The terms of these arrangements have been reviewed and approved by UCSD in accordance with its conflict of interest policies. Dr. Andreassen is a consultant for cortechs.ai, and received speaker's honorarium from Janssen, Lundbeck and Sunovion unrelated to the topic of this study. . The remaining authors have no competing interest.

Online Methods

GSA-MiXeR's full and baseline models

Under the assumptions of a simple additive genetic model, the contribution of the genotype to the phenotype is modeled as a sum of contributions across genetic variants: $y_k = \sum_{i=1}^{\bar{M}} g_{ik}\,\beta_i + e_k$, where y_k is a quantitative phenotype or disease liability of the k-th individual; g_{ik} is the number of reference alleles for the i-th variant (centered to zero mean, but not variance standardized), where the index i runs over the $\bar{M}$ variants in the reference panel; β_i is per-allele effect size (known also as an additive genetic effect of allele substitution). The parameters β_i and e_k are scaled so that phenotype y has zero mean and unit variance.

Following previous work[16-19], GSA-MiXeR builds on a causal mixture model[65] with a spike-and-slab prior distribution $\beta_i \sim (1-\pi_1)\delta_0 + \pi_1 N\left(0, \sigma_i^2\right)$,where δ_0 is the probability mass at zero (Dirac delta function), $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2, π_1 gives the prior probability of a variant having a non-zero effect size, and the corresponding effect size variance σ_i^2 is allowed to vary across genetic variants. To reduce the number of effective parameters being optimized, the full GSA-MiXeR model parametrizes σ_i^2 based on functional categories, genes, allele frequency, and LD score as follows:

$$\sigma_{i,full}^2 = \left(\sigma_{A,0}^2 + \sum_{p=1}^{N_A}[i \in A_p]\sigma_{A,p}^2\right)\left(\sigma_{G,0}^2 + \sum_{q=1}^{N_G}[i \in G_q]\sigma_{G,q}^2\right)H_i^S L_i^{\ell}. \quad (1)$$

Here, the H_i^S term allows the modeling of architectures dependent on allele frequency, the term $H_i = 2f_i(1-f_i)$ denotes heterozygosity of the i-th variant (f_i denotes minor allele frequency of the i-th variant), and the S parameter controls the effect size distribution along the allele frequency spectrum. Similarly, the L_i^ℓ term allows the modeling of LD-dependent genetic architectures, where L_i denotes the LD score of the i-th variant and the ℓ parameter controls the effect size distribution with respect to LD score. Each of the other two multiplicative factors in equation (1) controls the differential enrichment across functional categories and across genes, respectively. Index p runs across functional categories $\{A_1, A_2, \dots, A_{N_A}\}$, where parameter $\sigma^2_{A,p}$ represents the contribution of the p-th functional category to the variance of the i-th genetic variant (if that variant belongs to the category, as indicated by $[i \in A_p]$ term in equation (1), where square brackets maps true or false predicate to 1 or 0 – a notation known as Iverson Bracket). If genetic variant i belongs to multiple categories, the variances of those categories are added together. Similarly, index q runs across genes $\{G_1, G_2, \dots, G_{N_G}\}$, each contributing its own effect size variance $\sigma^2_{G,q}$ for genetic variants that belong to G_q. Parameters $\sigma^2_{A,0}$ and $\sigma^2_{G,0}$ allow for non-zero variance σ_i^2 for genetic variants that do not belong to any functional category or genes.

For the baseline GSA-MiXeR model, we used a single effect size variance parameter, σ^2_G, for all SNPs that belong to any gene, regardless of which gene they belong to:

$$\sigma^2_{i,base} = \left(\sigma^2_{A,0} + \sum_{p=1}^{N_A} [i \in A_p]\sigma^2_{A,p}\right)\left(\sigma^2_{G,0} + [i \in G]\sigma^2_G\right)H_i^S L_i^\ell \,,$$

where $G = \bigcup_{q=1}^{N_G} G_q$ is a union of all genes. The main results were reported with full and baseline models based on $N_A = 75$ functional annotations and $N_G = 18{,}201$ protein-coding genes. However, for the sensitivity analysis (Supplementary Figure 4) we allowed a full model where variance parameters $\sigma^2_{G,q}$ were modeled at the level of gene-sets, rather than at the level of individual genes. In this scenario heritability model $\sigma^2_{i,full}$ is still defined by the equation (1), however, each G_q corresponded to one of $N_G = 10{,}475$ gene-sets.

Heritability enrichment

The following expression defines SNP-based heritability for a phenotype standardized to a unit variance:

$$h^2_{SNP} = Var\left(\sum_i g_i\beta_i\right) = \sum_i Var(g_i\beta_i) + 2\sum_{i<j} Cov\left(g_i\beta_i, g_j\beta_j\right).$$

Following the h2gene approach[58], we employ the assumption that there is zero covariance between causal effects of different variants. This enables the partitioning of heritability across variants and, thereby, the definition of heritability of a gene-set G as

$$h^2_{full}(G) = \textstyle\sum_{i\in G} \pi_1 H_i \sigma^2_{i,full} \;\text{ and }\; h^2_{base}(G) = \sum_{i\in G} \pi_1 H_i \sigma^2_{i,base}\,,$$

where heterozygosity factor H_i converts variance σ_i^2 defined in the units of per-allele effects into variance expressed in the units of standardized genotypes. Fold enrichment of heritability can be computed as a ratio of $h^2_{full}(G)/h^2_{base}(G)$. However, we observed that fitting one parameter for each gene may lead to a slight inflation of heritability estimates, an issue that did not manifest itself when fitting the baseline model. To account for this, we define fold enrichment based on the fraction of total heritability attributed to the gene-set as:

$$enrich(G) = \frac{h^2(G)_{full}/h^2_{full}}{h^2(G)_{base}/h^2_{base}}. \qquad (2)$$

Log-likelihood computation and its gradients

Let z_j be the GWAS z-score of j-th variant out of in total M variants tested for association. We assume that association is driven by the underlying additive effects β_i, however, the set of GWAS variants does not need to be the same as the set of the underlying variants (hence we use different subscripts i and j to distinguish between them, and assume that all M variants tested for association are present among the $\bar{M}$ variants of the reference panel). Inference of GSA-MiXeR parameters is based on optimizing log-likelihood function $\log L(z_1, \dots, z_M|\theta) \rightarrow \max_{\theta}$ of observing a set of GWAS summary statistics given model parameters, while controlling for the LD structure among variants and their allele frequencies. The log-likelihood computation is based on previously derived methodology[16-18], which established the following additive relationship between genetic effects (β_i) and GWAS z-scores (z_j), highlighting that z_j depends on β_i of all i-th variants tagged by j-th variant due to LD:

$$z_j = \delta_j + \epsilon_j = \sum_i \sqrt{N_j H_i} r_{ij} \beta_i + \epsilon_j \,, \tag{3}$$

where N_j is sample size at the j-th variant, r_{ij} is the LD allele correlation between i-th underlying and j-th GWAS variants, and ϵ_j is the normally distributed residual $\epsilon_j \sim N(0, \sigma_0^2)$ with variance σ_0^2 that accounts for potential inflation in GWAS summary statistics. Our two approaches for computing the log likelihood using "fast" and "full" models are described in the Supplementary Note. The "fast" model is based on the method of moments, approximating $p(z_j)$ is a way that preserves the second and fourth moments of the distribution, thus preserving variance and kurtosis; the third moment (skewness) is irrelevant because of symmetry of the $p(z_j)$ distribution. The "full" model performs more accurate sampling from the prior distribution $p(\beta_i)$ at the expense of longer computation times. Based on these methods of evaluating log-likelihood, we found an efficient computation procedure for computing the gradients $\frac{\partial L(z_1,\dots,z_M|\theta)}{\partial \pi_1}$ and $\frac{\partial L(z_1,\dots,z_M|\theta)}{\partial \sigma_i^2}$ in the context of "fast" and "full" models. Finally, from equation (1) one can compute the Jacobian matrixes $\left[\frac{\partial \sigma_i^2}{\partial \sigma_{A,p}^2}\right]_{ip}$, $\left[\frac{\partial \sigma_i^2}{\partial \sigma_{G,q}^2}\right]_{iq}$, $\left[\frac{\partial \sigma_i^2}{\partial S}\right]_i$ and $\left[\frac{\partial \sigma_i^2}{\partial \ell}\right]_i$, thus allowing the computation of log-likelihood gradients with respect to the parameters of the final model: $\frac{\partial L(z_1,\dots,z_M|\theta)}{\partial \sigma_{A,p}^2} = \sum_i \frac{\partial L(z_1,\dots,z_M|\theta)}{\partial \sigma_i^2} \frac{\partial \sigma_i^2}{\partial \sigma_{A,p}^2}$, and similarly for all other parameters ($\sigma_{G,q}^2, S, \ell$ and π_1). This can be seen as an equivalent to error backpropagation in a neural network, with σ_i^2 representing the first layer of the network, and $\sigma_{A,p}^2$, $\sigma_{G,q}^2$ representing the second layer. The total amount of computations needed to calculate the entire gradient vector (for π_1 and σ_i^2 across all variants) is similar to that needed for computing the value of the likelihood function itself. This made it feasible to apply first-order optimization methods while avoiding computational burden despite adding tens of thousands of parameters to the model. For comparison, previous applications of the MiXeR model[16-19] had a fixed number of 3, 5, 9, and 12 parameters, respectively, and were fitted with zero-order methods including Differential evolution[66], Nelder-Mead[67], and Brent's method[68].

Log-likelihood optimization

The log-likelihood function $\log L(z_1, \dots, z_M|\theta) = \sum_j w_j \log p(z_j|\theta)$ is weighted across variants using inverse LD score[11] to avoid over-counting evidence from large LD blocks, in a way that is similar to `--w-ld` option in sLDSC. The fit is first performed for the baseline model, using the Adam algorithm[20] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$ in order to jointly fit parameters of the baseline GSA-MiXeR model. Optimization utilizes N=100 epochs, each of which passes over 22 batches (one batch per chromosome, re-shuffled in a random order for each epoch). The step-size parameter was gradually reduced from $\alpha = 0.064$ down to $\alpha = 0.0001$, reducing the value by a factor of 2 after every 10 epochs.

Optimization of the full model constrains $\sigma^2_{A,0}$, $\sigma^2_{A,p}$, $\sigma^2_{G,0}$, σ^2_0, S and ℓ parameters to their respective values from the baseline model. The optimization of the remaining parameters ($\sigma^2_{G,q}$, effect size variances for each gene) is initiated from the σ^2_G value from the baseline model and is done separately for each chromosome using $\beta_1 = 0.1$ and $\beta_2 = 0.8$ in Adam algorithm. While GSA-MiXeR implementation allows fitting all parameters including π_1 and S from the data, the main results were obtained constraining the model to an infinitesimal model ($\pi_1 = 1$), and $S = -0.25$ parameter chosen in line with the LDAK recommendation[14].

After fitting baseline and full GSA-MiXeR models, we apply Akaike Information Criterion (AIC) to identify genes and gene-sets that are improving the fit of the full model compared to the baseline model. More specifically, to assess whether in a certain genomic region (i.e. for a gene, or for a gene-set) our full model, θ_{full}, has a better likelihood than the baseline model, θ_{base}, we constructed a partial model, θ_{part}, which is equivalent to the full model θ_{full} except for the genomic region being tested, and in that region θ_{part} is equivalent to the baseline model θ_{base} ("equivalent" means that gene-level variance parameters σ^2_i in θ_{part} are set to their respective values from either from full or baseline models). AIC was computed as $2\log\left(\frac{L(\theta_{full})}{L(\theta_{part})}\right) - 2k$, where k represents the number of model parameters fitted for the genomic region being tested. In the case of non-overlapping genes, the k parameter equals to the number of genes in a gene-set being tested; in a more general case, this accounts for an overlap among adjacent genes occurring largely due to 10-kb up- and down-stream expansion.

SEs of fitted parameters were estimated from the observed Fisher's information matrix $-\Delta\Delta^T \log L|_{\theta^*}$ (the negative Hessian matrix of the log-likelihood function). For compound functions of multiple parameters, such as fold enrichment of heritability in a region, we sample N=100 realizations of the parameters vector θ from the posterior distribution $\theta \sim N(\theta^*, [-\Delta\Delta^T \log L|_{\theta^*}]^{-1})$, calculate the compound function (e.g., fold enrichment) for each realization of θ, and report the standard deviations across realization. If the Hessian matrix was not positive definite, we use marginal errors of fitted parameters. The Fisher's information matrix was computed efficiently using sparse matrices (see Supplementary Note 1 for further details). We validated SEs in simulations by comparing them to standard deviations of point estimates across simulation runs (Supplementary Table 3); for individual genes we also compared GSA-MiXeR SEs to SEs estimated by LDAK-GBAT, leading to satisfactory results (Supplementary Figure 9, Supplementary Table 9-10).

HRC and UKB references (LD structure)

We prepared two reference panels available for GSA-MiXeR analysis using Haplotype Reference Consortia (HRC) and UK Biobank (UKB) datasets.

Our HRC reference has 11,980,511 variants and 23,152 samples after basic QC procedure. During sample QC, we select individuals with EUR ancestry (including Finnish population) defined from the first two principal components of 1000 genomes Phase3 data. Further, we prune related individuals using KING software ("`king --unrelated -degree 2`"). For variant QC, we use "`plink --maf 0.001 --hwe 1e-10 --geno 0.1`" parameters, and further exclude markers without RS IDs and excluding duplicated RS IDs.

Our UKB reference has 12,926,669 variants and 337,145 subjects, derived from the UK Biobank imputed v3 dataset. During sample QC, we select unrelated individuals with white British ancestry, remove sex chromosome aneuploidy, and exclude participants who withdrew their consent. Variant QC was applied as follows: "`plink --maf 0.001 --hwe 1e-10 --geno 0.1`", in addition to filtering variants with imputation INFO score below 0.8 and excluding variants with duplicated RS IDs.

In both references, the calculation of allelic correlation (LD r2) was done from hard calls, separately for each chromosome, with a window size of 10 MB, and using a minimum LD r2 threshold of 0.01.

Functional categories and gene-sets

For functional categories, we used the set of 75 binary annotations designed for stratified LD score regression (baselineLD_v2.2).

To define gene boundaries, we used NCBI Entrez database, starting from a set of protein-coding (N=19,608 genes). After keeping only genes with known coordinates in the primary assembly of GRC37 human reference (annotation release 105.20201022), excluding genes on non-autosomes, and excluding genes without variants in MiXeR's UKB or HRC references, we obtained a final list of N=18,201 protein-coding genes. The boundaries of each gene were extended by 10 KB up- and downstream, both for the GSA-MiXeR and for the MAGMA analyses.

To define gene-sets, we used GO terms from MsigDB v7.5 (biological processes - c5.bp; cellular component - c5.cc; and molecular function - c5.mf; resulting in 10,402 gene-sets in total). This set was extended using the SynGO database (https://syngoportal.org/), which added 73 gene-sets after excluding those comprising 5 genes or less. Among 10,475 resulting gene-sets there were 19,366 unique genes present, of which 16,787 were present among the list of 18,201 genes available for analysis. To avoid reporting results for highly overlapping gene-sets, we pruned gene-sets that overlapped with a Dice coefficient of 0.8 or above, retaining smaller gene sets from each overlapping pair. This procedure excluded 1,771 gene-sets from the analysis, and the final list used in the analysis contained 8,704 gene-sets.

To validate whether gene-set enrichment was largely driven by a single gene, we performed additional leave-one-out analysis for all gene-sets with up to 25 genes, where enrichment was computed after removing one gene at a time and reporting the gene whose exclusion led to lowest enrichment.

Summary statistics and individual-level data

Table 1 gives an overview of publicly available GWAS summary statistics used for this analysis. In the replication analysis of SCZ, we used sub-study data to perform a variance-based meta-analysis on a randomly selected half of the cohorts. In our replication analysis of height, we performed GWAS on UK Biobank standing height (field 50), following the GWAS protocol from the Neale Lab (see URLs). To validate prediction (Supplementary Figure 8), we re-run GSA-MiXeR using summary statistics from PGC wave3 GWAS excluding the Thematically Organized Psychosis (TOP) cohort from Norway with N=743 schizophrenia cases and N=1074 controls, which we use to evaluate the accuracy of polygenic risk scores.

Simulations

Simulations used a panel of N=337,145 subjects and M=12,926,669 variants from UK Biobank. For each run, we used PLINK to obtain GWAS summary statistics of a quantitative phenotype $y_k = \sum_i g_{ik}\,\beta_i + e_k$, where β_i were drawn from a specific model (depending on simulation scenario, as described below), and e_k residual was drawn from a normal distribution with zero mean and variance chosen in a way that sets heritability $h^2 = var(G\,\beta)\,/\,var(y)$ to a predefined level (h2=0.1, 0.4 or 0.7).

A subset of causal variants in simulations was selected by randomly pruning all reference variants at LD r2=0.1 threshold so that causal variants were in weak LD with each other. This ensured that each genomic region had a well-defined heritability, as for a quantitative phenotype $y = \sum_i g_i\beta_i + \epsilon$ the variance of the genetic component under a fixed effects model is given by $Var(\sum_i g_i\beta_i) = \sum_i \beta_i^2\, Var(g_i) + \; 2\sum_i \sum_{j>i} \beta_i\beta_j Cov(g_i, g_j)$, where the second component is negligible if causal variants

are not in LD with each other. The remaining component allows the definition of genetic variance in a specific region g as $\sum_{i \in g} \beta_i^2 Var(g_i)$.

We considered 9 distinct scenarios for the genetic architecture of the simulated trait ("base", "snpXfold", "ncausals12926", "ncausals6463", "CodingUCSC", "PromoterUCSC", "mafBelow10", "tldBelow50" and "tldBelow25"), as outlined in Supplementary Table 1. In all scenarios, causal variants were selected from SNPs within protein-coding genes, expanded 10KB up- and downstream. In the "base" scenario, N=28,958 causal variants were uniformly distributed across protein-coding genes. In the "snpXfold" scenario the density of causal variants in the enriched gene-set was set to be 3, 10, or 30 times higher (depending on the geneXfold parameter) than the density of SNPs within non-enriched gene-sets, as such heritability enrichment in "snpXfold" scenario was modulated by the density of causal variants, rather than by enrichment of effect sizes. The "ncausals12926" and "ncausals6463" simulations were equivalent to the "base" scenario except that the polygenicity of the simulated traits was reduced, using N=12,926 and N=6,463 causal variants respectively, which is approximately 2- and 4-times lower polygenicity than in the "base" scenario. In "CodingUCSC" and "PromoterUCSC" scenarios all causal variants were drawn from the coding and promoter regions of the genes, respectively, as defined by baselineLD_v2.2 annotation from stratified LD score regression. In "mafBelow10" scenario all causal variants were drawn from SNPs with MAF<10%, as defined by the UK Biobank reference used in simulations. Finally, "tldBelow50" and "tldBelow25" scenarios represent an extreme case of LD-dependent genetic architecture where causal variants are drawn from SNPs with a total LD score below 50 and 25, respectively. In simulations, we used a basic model for MAF- and LD-dependency ($S = 0, \ell = 0$), also constraining respective GSA-MiXeR parameters to 0, and did not simulate enrichments across functional categories (except for "CodingUCSC" and "PromoterUCSC" scenarios where these functional categories were accounted for in the baseline and full GSA-MiXeR models; for "mafBelow10", "tldBelow50" and "tldBelow25" scenarios we included 10 MAF- and LD- bins in the model).

After selecting the set of causal variants, their effect sizes β_i were selected to ensure pre-defined fold enrichment for a selected set of genes. We consider two regions in a genome: G (representing the set of all genes) and $g \subset G$, representing the subset of N=1, 3, 10, 30, or 100 enriched genes. Let H_G and H_g be total heterozygosity for variants in G and g, respectively. To generate a trait with pre-defined fold enrichments, $f(g)$, we draw β_i from a normal distribution $N\left(0,\ h^2 \frac{f(g)}{H_G}\right)$ for variants in enriched gene-set g, and from $N\left(0,\ h^2 \frac{H_G - f(g)H_g}{H_G(H_G - H_g)}\right)$ for the remaining variants in $G \backslash g$. As shown in Supplementary Note 1, this translates to a true fold enrichment equal to the pre-selected $f(g)$ value. For example, simulations with h2=0.7 in the "base" scenario with geneXfold=30.0 yielded a simulated heritability of approx. 8% for the gene-set with 100 genes; for a scenario with 10 times smaller gene-sets the heritability for enriched gene-sets was 0.8%, i.e., lower proportionally to the size of the gene-set. Equivalently, in scenarios with lower fold enrichment (e.g. geneXfold=3.0) gene-sets of the same size (Ngene=100) also had a simulated heritability of 0.8%. For a single gene without enrichment its heritability was, on average, 5e-5 (in simulations with h2=0.7) or as low as 1e-5 (for h2=0.1).

We chose to keep the sample size constant throughout all simulations as the power of GWAS depends on z-scores, which in turn depend on the product of the sample size and heritability (as long as polygenicity is kept constant). Thus, our simulations using QCed UKB genotypes (N=337,145 subjects) and h2=0.1 yield results that are expected to be equivalent to simulations with N=1,000,000 for a trait with SNP heritability of h2=0.03, or to N=100,000 for a trait a with h2=0.3. Polygenicity in the "base" scenario is defined to match the polygenicity of schizophrenia, with its estimated 30,000 causal variants[17] (one of the highest estimates across complex traits). For traits with lower polygenicity GSA-MiXeR should generally yield more

precise results (among traits with equal SNP heritability and equal GWAS sample size), because lower polygenicity implies proportionally higher effects for causal variants, which in turn imply stronger GWAS signal in loci surrounding causal variants. However, in “ncausals12926” and “ncausals6463” simulations scenarios (2x and 4x lower polygenicity than the “base” scenario) we observed somewhat reduced accuracy of GSA-MiXeR compared to the “base” scenario. We attribute this to inaccuracies in our simulation framework, where sampling error in scenarios with a sufficiently low number of causal variants (much lower than the total number of genes) may cause actual enrichment to be lower than the pre-defined “geneXfold” value.

Display Items

Table 1. Summary of main and exploratory GSA-MiXeR results for schizophrenia and 20 other diverse human traits and disorders.

These main results are based on gene-sets selected through MAGMA analysis, annotated with fold enrichment estimates computed by GSA-MiXeR. Exploratory results are based on GSA-MiXeR analysis, implicating gene-sets with positive AIC (Akaike Information Criterion). Sample size indicates the total number of participants in GWAS for quantitative traits and the number of cases and controls for binary traits, respectively; RMSE, root mean square error.

Traits	Sample size	Main results			Exploratory results		
		Count of gene-sets	Fold enrichment average (RMSE)	Average size of gene-sets (#genes)	Count of gene-sets	Fold enrichment average (RMSE)	Average size of gene-sets (#genes)
Alcohol consumption[22]	112117	0			1	21.12 (8.96)	7.0
Alzheimer's disease[33]	86531/676386	13	10.30 (4.13)	65.5	51	9.36 (4.56)	11.1
Body mass index[35]	795640	16	1.28 (0.11)	900.8	41	4.86 (1.81)	9.0
Chronic kidney disease[34]	41395/439303	0			5	6.42 (3.03)	7.8
Chronotype[32]	345148	1	1.47 (0.16)	427.0	3	11.45 (5.49)	6.7
Cognition[29]	269867	1	3.49 (0.91)	5.0	6	3.54 (1.64)	6.8
Crohns disease[23]	12194/34915	77	2.59 (0.63)	513.6	52	15.93 (4.63)	9.1
Educational attainment[28]	765283	10	1.32 (0.07)	549.7	34	2.85 (0.80)	7.8
Fasting glucose[21]	281416	9	3.94 (1.00)	195.7	71	6.87 (2.46)	8.8
Fasting insulin[21]	281416	0			6	7.17 (4.63)	6.2
Glycated hemoglobin[21]	281416	3	5.35 (1.76)	460.7	57	8.41 (2.73)	10.9
Heart failure[30]	47309/930014	0			1	31.27 (11.55)	5.0
Height[32]	385748	40	2.26 (0.33)	636.1	405	4.69 (1.52)	17.5
Hospitalized covid[31]	32519/2062805	0			1	18.56 (6.92)	5.0
Migraine[25]	48975/540381	0			8	7.39 (3.08)	7.8
Schizophrenia[8]	53386/77258	17	1.97 (0.39)	441.4	36	7.28 (3.41)	7.8
Sleep[32]	384225	2	6.55 (1.96)	11.5	9	9.03 (3.31)	9.7
Stroke[27]	110182/1503898	0			3	14.82 (6.42)	7.0
Systolic Blood Pressure[24]	745820	25	1.82 (0.44)	706.8	202	4.70 (1.75)	14.5
Type 2 diabetes[26]	74124/824006	7	3.89 (0.73)	320.3	161	7.99 (2.75)	13.8
Ulcerative colitis[23]	12366/34915	33	2.06 (1.01)	353.8	20	11.98 (4.36)	7.2

Figure 1. Selected results from simulations

The top row of figures compares fold enrichment estimated by GSA-MiXeR versus ground truth (shown by the horizontal dashed line). The bottom row of figures visualizes the accuracy of sorting gene-sets according to the MAGMA enrichment p-values, or according to the GSA-MiXeR fold enrichment, showing the position of the truly enriched gene-set among 1000 randomly generated gene-sets in simulations (expected rank is 1, lower values are better). The simulations cover four levels of enrichment: 1 (null enrichment), and 3, 10, and 30-fold enrichment; and vary in terms of the size of the enriched gene-set: 1, 3, 10, 30, and 100 genes. In all simulations, the trait was simulated with the "base" scenario of genetic architecture (causal variants uniformly distributed across protein-coding genes), and total heritability h2=0.4. Additional simulations are presented in Supplementary Figures 1 and 2.

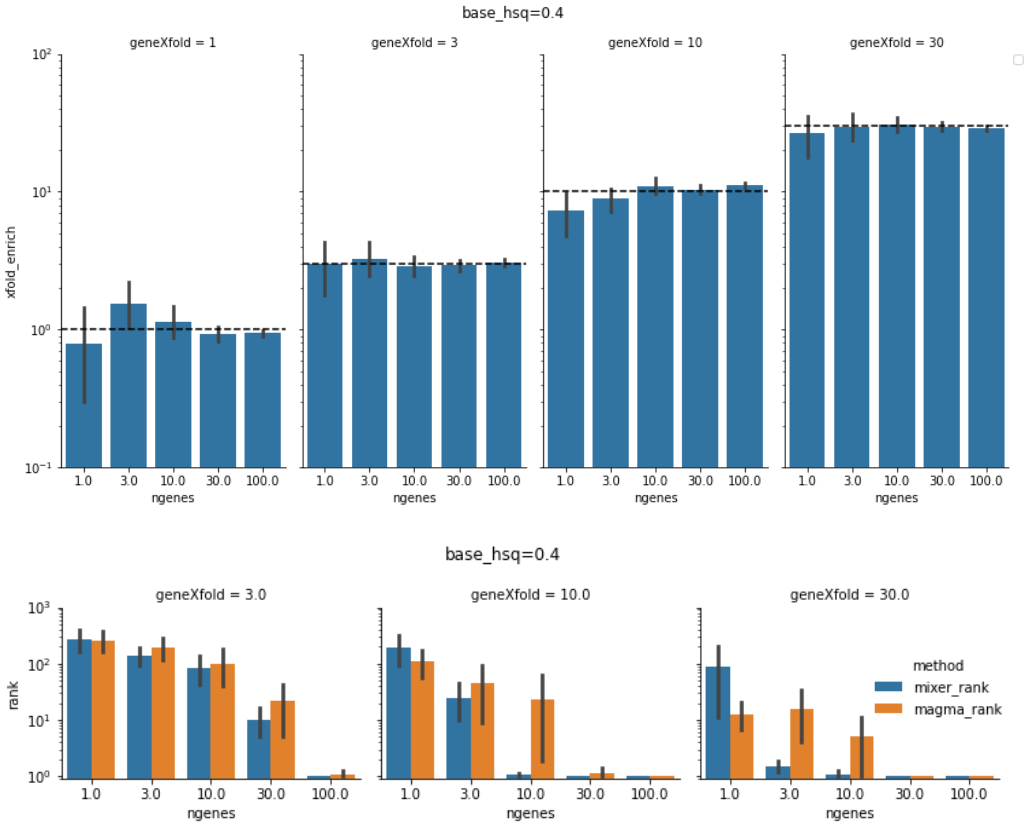


Figure 2. Main GSA-MiXeR results for schizophrenia

The set of gene-sets passing MAGMA significance threshold (p<0.05 after Bonferroni correction), ordered by GSA-MiXeR fold enrichment. The left-hand panel shows GSA-MiXeR fold enrichment estimates, with error bars showing standard errors of fold enrichment estimates. The right-hand panel shows log10 of MAGMA's enrichment p-value. The number of genes per gene-set is indicated in parenthesis after gene-set name. Vertical dashed lines indicate no enrichment (for GSA-MiXeR figures) and Bonferroni-corrected p-value threshold (for the MAGMA figure).

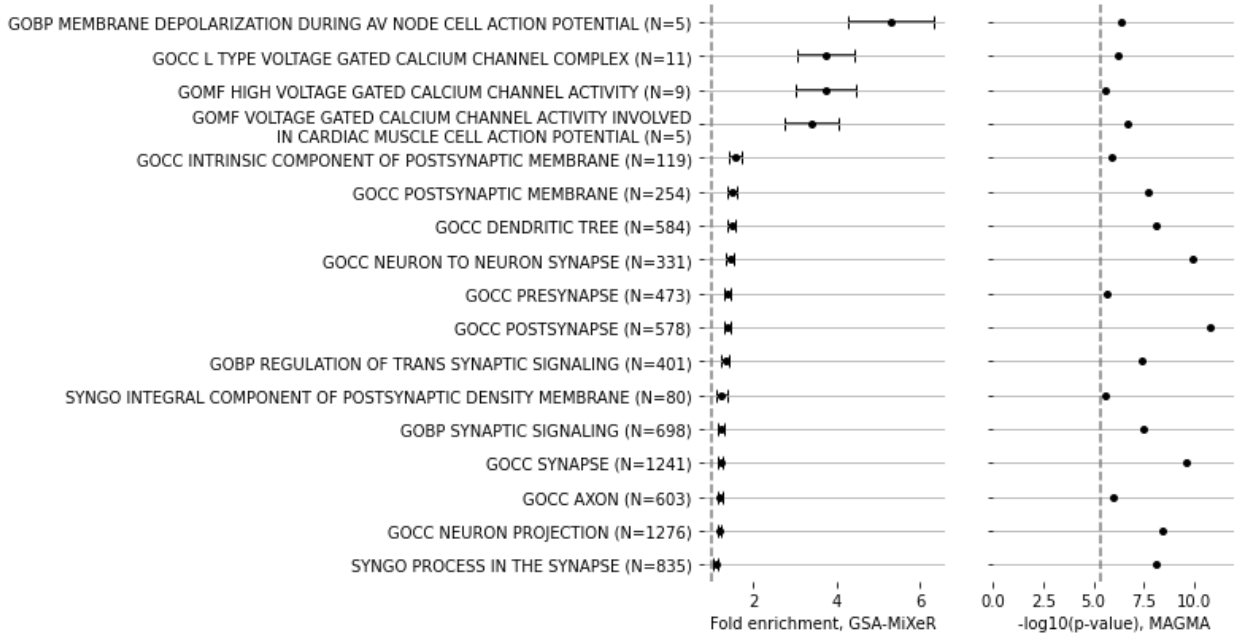


Figure 3. Exploratory GSA-MiXeR results for schizophrenia

Fold enrichment estimates for gene-sets with a positive AIC (Akaike information criterion) value in GSA-MiXeR analysis. Positive AIC indicates a better fit of the GSA-MiXeR full model compared to the baseline model for each of the implicated gene-sets. The number of genes per gene-set is indicated in parentheses after each gene-set name. GOMF – Gene Ontology Molecular Function; GOBP – Gene Ontology Biological Processes; GOCC – Gene Ontology Cellular Component).

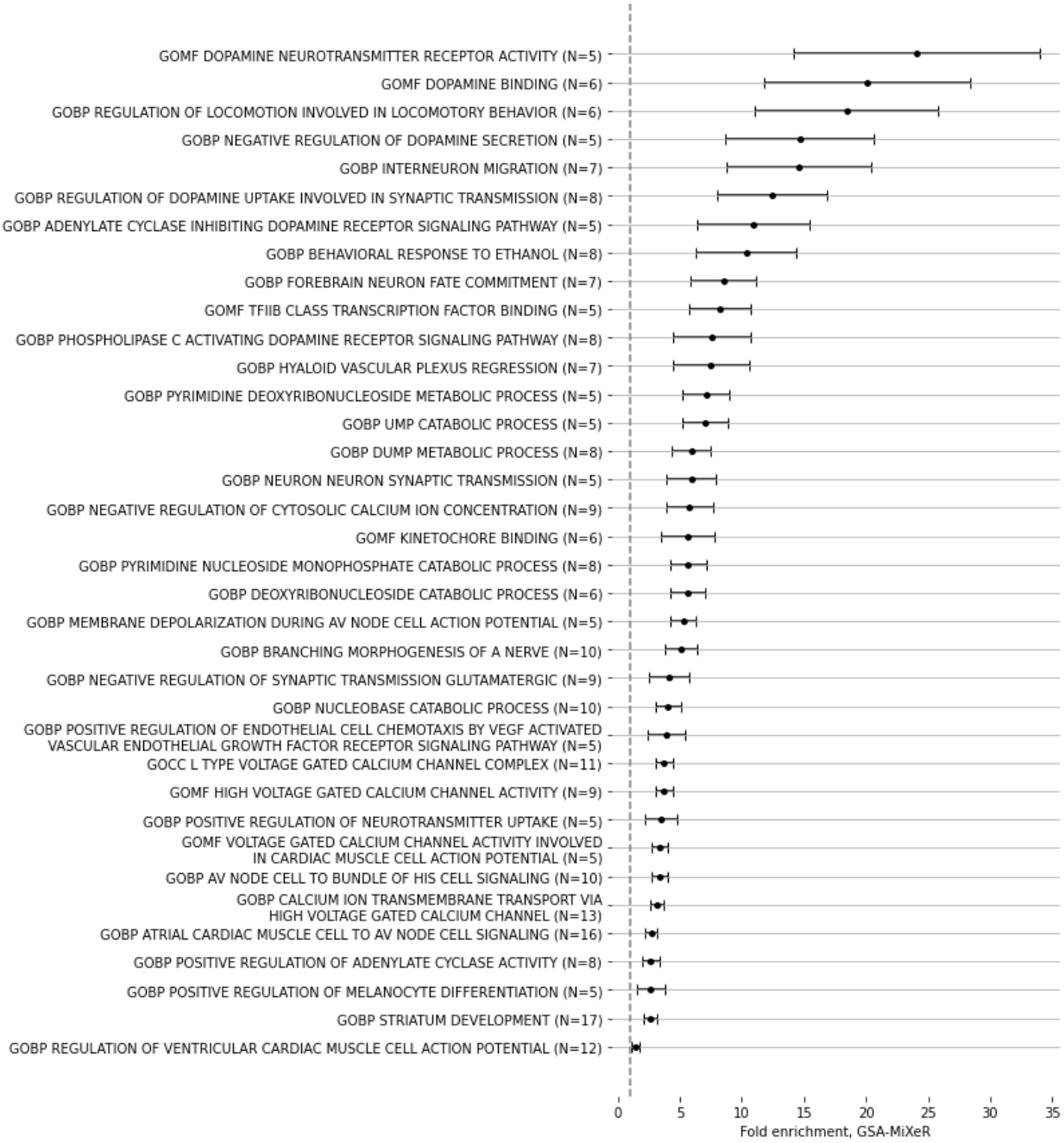
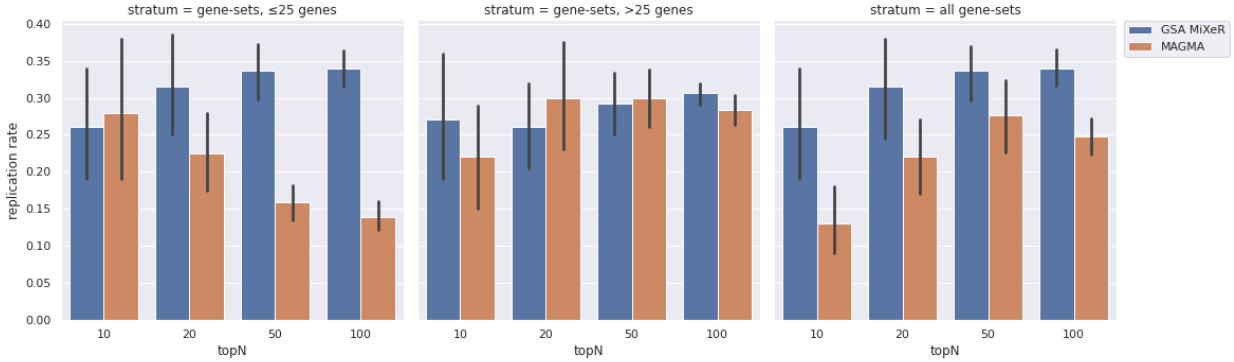


Figure 4. Replication analysis for schizophrenia

Replication analysis for schizophrenia (SCZ) using Psychiatric Genomic Consortia (PGC) substudies. Replication rate is assessed among gene-sets with up to 25 genes (left-hand figure), gene-sets with more than 25 genes (middle figure), and all gene-sets (right-hand figure). Replication rate is computed for GSA-MiXeR and MAGMA analyses using two-fold cross-validation, after applying both methods to two independent and equally sized GWAS sub-samples. After ranking gene-sets according to their fold enrichment of heritability (GSA-MiXeR) and enrichment p-value (MAGMA) obtained in the discovery sample, the replication rate (shown on the vertical axis) was then defined as the fraction of gene-sets that remain within top N gene-sets (N= 10, 20, 50, and 100 on the horizontal axis) in the replication half of the dataset.



References

1. Sullivan, P.F. & Geschwind, D.H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162-183 (2019).
2. de Leeuw, C.A., Neale, B.M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nature Reviews Genetics* **17**, 353-364 (2016).
3. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545 (2005).
4. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).
5. Koopmans, F. *et al.* SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103**, 217-234.e4 (2019).
6. Hill, W.D. *et al.* A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol Psychiatry* **24**, 169-181 (2019).
7. Howard, D.M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* **22**, 343-352 (2019).
8. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502-508 (2022).
9. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).
10. Simillion, C., Liechti, R., Lischer, H.E.L., Ioannidis, V. & Bruggmann, R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics* **18**, 151 (2017).
11. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
12. Goeman, J.J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980-987 (2007).
13. Tashman, K.C., Cui, R., O'Connor, L.J., Neale, B.M. & Finucane, H.K. Significance testing for small annotations in stratified LD-Score regression. *medRxiv*, 2021.03.13.21249938 (2021).
14. Speed, D., Cai, N., Johnson, M.R., Nejentsev, S. & Balding, D.J. Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**, 986-992 (2017).
15. Zabad, S., Ragsdale, A.P., Sun, R., Li, Y. & Gravel, S. Assumptions about frequency-dependent architectures of complex traits bias measures of functional enrichment. *Genet Epidemiol* **45**, 621-632 (2021).
16. Frei, O. *et al.* Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat Commun* **10**, 2417 (2019).
17. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet* **16**, e1008612 (2020).

18. Shadrin, A.A. *et al.* Phenotype-specific differences in polygenicity and effect size distribution across functional annotation categories revealed by AI-MiXeR. *Bioinformatics* **36**, 4749-4756 (2020).
19. Holland, D. *et al.* The genetic architecture of human complex phenotypes is modulated by linkage disequilibrium and heterozygosity. *Genetics* **217**(2021).
20. Kingma, D.P.a.B., Jimmy. Adam: A Method for Stochastic Optimization. *arxiv* (2014).
21. Chen, J. *et al.* The trans-ancestral genomic architecture of glycemic traits. *Nature Genetics* **53**, 840-860 (2021).
22. Clarke, T.K. *et al.* Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). *Molecular Psychiatry* **22**, 1376-1384 (2017).
23. de Lange, K.M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics* **49**, 256-261 (2017).
24. Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature Genetics* **50**, 1412-1425 (2018).
25. Hautakangas, H. *et al.* Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles. *Nature Genetics* **54**, 152-160 (2022).
26. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics* **50**, 1505-1513 (2018).
27. Mishra, A. *et al.* Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* **611**, 115-123 (2022).
28. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics* **54**, 437-449 (2022).
29. Savage, J.E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics* **50**, 912-919 (2018).
30. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nature Communications* **11**, 163 (2020).
31. The, C.-H.G.I. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics* **28**, 715-718 (2020).
32. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics* **51**, 1339-1348 (2019).
33. Wightman, D.P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics* **53**, 1276-1282 (2021).
34. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature Genetics* **51**, 957-972 (2019).

35. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Molecular Genetics* **27**, 3641-3649 (2018).
36. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704-712 (2022).
37. Smeland, O.B., Frei, O., Dale, A.M. & Andreassen, O.A. The polygenic architecture of schizophrenia - rethinking pathogenesis and nosology. *Nat Rev Neurol* **16**, 366-379 (2020).
38. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
39. The Schizophrenia Working Group of the Psychiatric Genomics, C., Ripke, S., Walters, J.T.R. & O'Donovan, M.C. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv*, 2020.09.12.20192922 (2020).
40. Nakazawa, K. *et al.* GABAergic interneuron origin of schizophrenia pathophysiology. *Neuropharmacology* **62**, 1574-83 (2012).
41. Stedehouder, J. & Kushner, S.A. Myelination of parvalbumin interneurons: a parsimonious locus of pathophysiological convergence in schizophrenia. *Mol Psychiatry* **22**, 4-12 (2017).
42. Marder, S.R. & Cannon, T.D. Schizophrenia. *New England Journal of Medicine* **381**, 1753-1761 (2019).
43. Berrandou, T.-E., Balding, D. & Speed, D. LDAK-GBAT: Fast and powerful gene-based association testing using summary statistics. *The American Journal of Human Genetics* **110**, 23-29 (2023).
44. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **49**, 1421-1427 (2017).
45. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* **49**, 1421-1427 (2017).
46. Hindley, G. *et al.* Charting the Landscape of Genetic Overlap Between Mental Disorders and Related Traits Beyond Genetic Correlation. *Am J Psychiatry*, appiajp21101051 (2022).
47. Moon, A.L., Haan, N., Wilkinson, L.S., Thomas, K.L. & Hall, J. CACNA1C: Association With Psychiatric Disorders, Behavior, and Neurogenesis. *Schizophrenia Bulletin* **44**, 958-965 (2018).
48. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509-516 (2022).
49. Howes, O.D. & Kapur, S. The Dopamine Hypothesis of Schizophrenia: Version III—The Final Common Pathway. *Schizophrenia Bulletin* **35**, 549-562 (2009).
50. Fusar-Poli, P. & Meyer-Lindenberg, A. Striatal Presynaptic Dopamine in Schizophrenia, Part II: Meta-Analysis of [18F/11C]-DOPA PET Studies. *Schizophrenia Bulletin* **39**, 33-42 (2013).

51. Huhn, M. *et al.* Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. *Lancet* **394**, 939-951 (2019).
52. Stedehouder, J. & Kushner, S.A. Myelination of parvalbumin interneurons: a parsimonious locus of pathophysiological convergence in schizophrenia. *Molecular Psychiatry* **22**, 4-12 (2017).
53. Harrison, P.J. Schizophrenia Susceptibility Genes and Neurodevelopment. *Biological Psychiatry* **61**, 1119-1120 (2007).
54. Roda, G. *et al.* Crohn's disease. *Nature Reviews Disease Primers* **6**, 22 (2020).
55. Ramasamy, S. & Subbian, S. Critical Determinants of Cytokine Storm and Type I Interferon Response in COVID-19 Pathogenesis. *Clinical Microbiology Reviews* **34**, 10.1128/cmr.00299-20 (2021).
56. Barrera-Chimal, J., Jaisser, F. & Anders, H.-J. The mineralocorticoid receptor in chronic kidney disease. *British Journal of Pharmacology* **179**, 3152-3164 (2022).
57. Kranzler, H.R. & Soyka, M. Diagnosis and Pharmacotherapy of Alcohol Use Disorder: A Review. *JAMA* **320**, 815-824 (2018).
58. Burch, K.S. *et al.* Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. *Am J Hum Genet* **109**, 692-709 (2022).
59. Yao, D.W., O'Connor, L.J., Price, A.L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics* **52**, 626-633 (2020).
60. Siewert-Rocks, K.M., Kim, S.S., Yao, D.W., Shi, H. & Price, A.L. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. *The American Journal of Human Genetics* **109**, 393-404 (2022).
61. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245-252 (2016).
62. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature Communications* **9**, 4361 (2018).
63. Zhu, X., Duren, Z. & Wong, W.H. Modeling regulatory network topology improves genome-wide analyses of complex human traits. *Nature Communications* **12**, 2851 (2021).
64. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics* **99**, 139-153 (2016).
65. Holland, D. *et al.* Estimating Degree Of Polygenicity, Causal Effect Size Variance, And Confounding Bias In GWAS Summary Statistics. *bioRxiv* (2017).
66. Storn, R. & Price, K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization* **11**, 341-359 (1997).
67. Nelder, J.A. & Mead, R. A Simplex Method for Function Minimization. *The Computer Journal* **7**, 308-313 (1965).

68. Brent, R.P. *Algorithms for minimization without derivatives*, (Prentice-Hall, 1973).