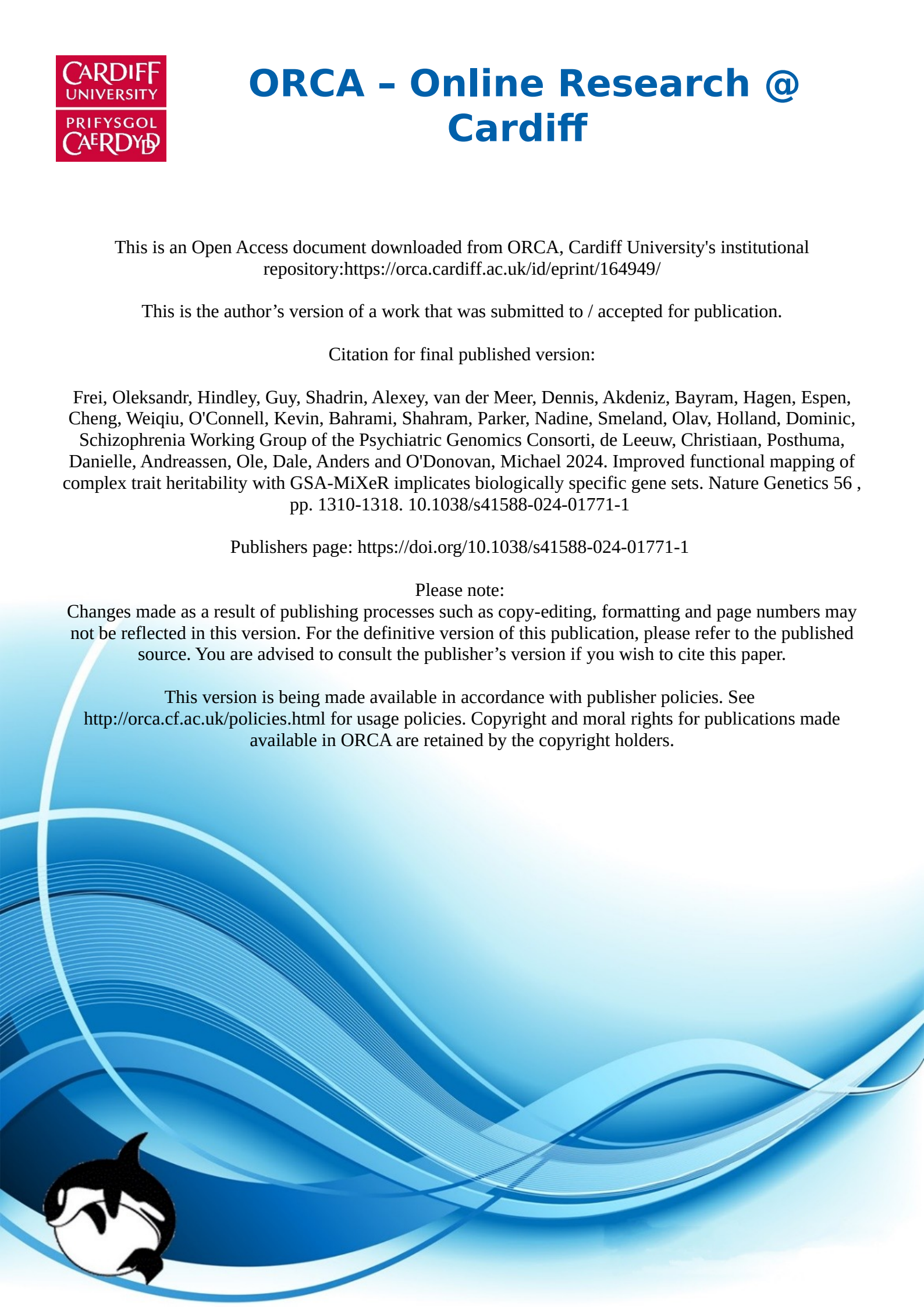


Supplementary Note 2

Table of Contents

Extended results

Main analysis

We further applied both GSA-MiXeR and MAGMA to 20 diverse human phenotypes (Supplementary Table 6), to demonstrate GSA-MiXeR utility for traits other than schizophrenia. After filtering on MAGMA-estimated statistical significance, 13 phenotypes were significantly associated with one or more gene-set. The number of gene-sets identified ranged from a single gene-set for chronotype ("regulation of cellular amide metabolic process") and cognition ("negative regulation of dendrite development") to 77 gene-sets for Crohn's disease. When comparing GSA-MiXeR prioritized gene-sets with MAGMA prioritized gene-sets, GSA-MiXeR frequently prioritized small gene-sets with high biological relevance. This included "negative regulation of amyloid fibril formation" (ngenes = 11) compared to "neurofibrillary tangle" (ngenes = 5) for Alzheimer's disease, "intrinsic component of postsynaptic membrane" (ngenes = 119) compared to "process in the synapse" (ngenes = 835) for educational attainment, "negative regulation of smoothened signaling pathway" (ngenes = 32) compared to "skeletal system development" (ngenes = 503) for height, "voltage gated calcium channel activity involved in cardiac muscle cell action potential" (ngenes = 5) compared to "negative regulation of transcription by RNA polymerase II" (ngenes = 1149) for systolic blood pressure, and "plasma cell differentiation" (ngenes = 6) compared to "lymphocyte activation" (ngenes = 695) for ulcerative colitis. There were also several phenotypes for which GSA-MiXeR prioritized the same gene-set as MAGMA, including "pancreas development" (ngenes = 79) for type 2 diabetes, "negative regulation of nucleobase containing compound metabolic process" (ngenes = 1367) for BMI, "negative regulation of type b pancreatic cell apoptotic process" (ngenes = 6) for both fasting glucose and glycated hemoglobin, and "hyaloid vascular plexus regression" (ngenes = 7) for sleep.

Main analysis for Crohn's disease

While MAGMA prioritized "cytokine production" (ngenes = 701) for Crohn's disease, the most fold enriched gene-set was "positive regulation of memory T cell differentiation" (ngenes = 9), followed by "positive regulation of interleukin 10 production" (ngenes = 38), "positive regulation of tumour necrosis factor…" (ngenes = 101), and "T-helper 17 cell lineage commitment" (ngenes = 13), the latter defined by the Gene Ontology database as "the process in which a CD4-positive, alpha-beta T cell becomes

committed to becoming a T-helper 17 cell, a CD4-positive, alpha-beta T cell … that produces IL-17"[1]. These four gene-sets are strikingly consistent with evidence that Crohn's disease may be driven by dysfunctional CD4-positive antigen-experienced T cells within the intestinal epithelium, which have been shown to switch from producing anti-inflammatory IL-10 to pro-inflammatory IL-17A, IFNγ, and tumor necrosis factor[2]. These findings illustrate how GSA-MiXeR can improve the granularity of GSA findings, helping to map GWAS findings to more informative neurobiological processes which can be tested experimentally. This may facilitate better characterization of the molecular mechanisms underlying diverse human phenotypes with the potential for identifying new druggable targets and clinical sub-groups.

Exploratory analysis

In our exploratory analysis, GSA-MiXeR frequently prioritized small, biologically relevant gene-sets which were not identified by MAGMA, including phenotypes that did not exhibit any significant MAGMA gene-sets (Supplementary Table 8). For example, "mineralocorticoid secretion" was the most highly enriched gene-set for chronic kidney disease, "interferon receptor activity" for hospitalized COVID, and "alcohol dehydrogenase activity zinc dependent" for alcohol consumption. All three gene-sets relate to core molecular mechanisms underlying each phenotype and have either been identified as potential treatment targets in the case of severe COVID[3] or are already targeted by widely available drugs such as spironolactone for chronic kidney disease[4] and disulfiram for alcohol dependence[5]. Nonetheless, the biological relevance of some high-ranking gene-sets within the exploratory analysis was less compatible with their respective phenotypes, including "peroxisome fission" for body mass index and "regulation of bile acid secretion" for fasting glucose. While filtering on GSA-MiXeR AIC differences may reveal novel pathophysiological insights, this emphasizes the importance of filtering on MAGMA estimated p-values to ensure robust findings. We additionally found that the magnitude of fold enrichments varied across phenotypes, with a tendency for mental traits such as cognition (mean fold enrichment = 3.54, Root Mean Square Error (RMSE) = 1.64) and educational attainment (mean fold enrichment = 2.86, RMSE = 0.80) to display lower fold enrichments than somatic disorders and biochemical measures such as Alzheimer's disease (mean fold enrichment = 9.36, RMSE = 4.56) and glycated hemoglobin (mean fold enrichment = 8.41, RMSE = 2.65). Cross-phenotype differences may reflect the heterogeneity and/or complexity of different phenotypes, as well as the extent to which the current gene-sets successfully map onto their underlying biological mechanisms.

Validate prediction accuracy in an independent schizophrenia sample

To further demonstrate the relevance of GSA-MiXeR findings, we tested whether GSA-MiXeR could improve the selection of SNPs for polygenic scoring (PGS) over standard selection based on p-values. We constructed an enhanced PGS model by re-ranking and clumping SNPs based on posterior effect size estimates from the baseline GSA-MiXeR model. The top-N SNPs were then included in the PGS model, using the original GWAS effect size estimates as weights. We compared our MiXeR-informed PGS to a conventional pruning and thresholding PGS in an independent clinical sample of 743 individuals with schizophrenia and 1074 healthy control subjects, with a mean age of 32.5 (SD 10.0) years, 52.6% male. This showed that the best-performing model achieved a 13% increase in accuracy, reaching 18.02% Nagelkerke pseudo-R2 with GSA-MiXeR versus 15.98% for conventional PGS (Supplementary Figure 8). Most notably, the based-performing GSA-MiXeR-enhanced PGS model included fewer SNPs than the conventional PGS. This indicates that GSA-MiXeR prioritizes SNPs which are more informative for case/control prediction in an independent sample.

Statistical significance testing

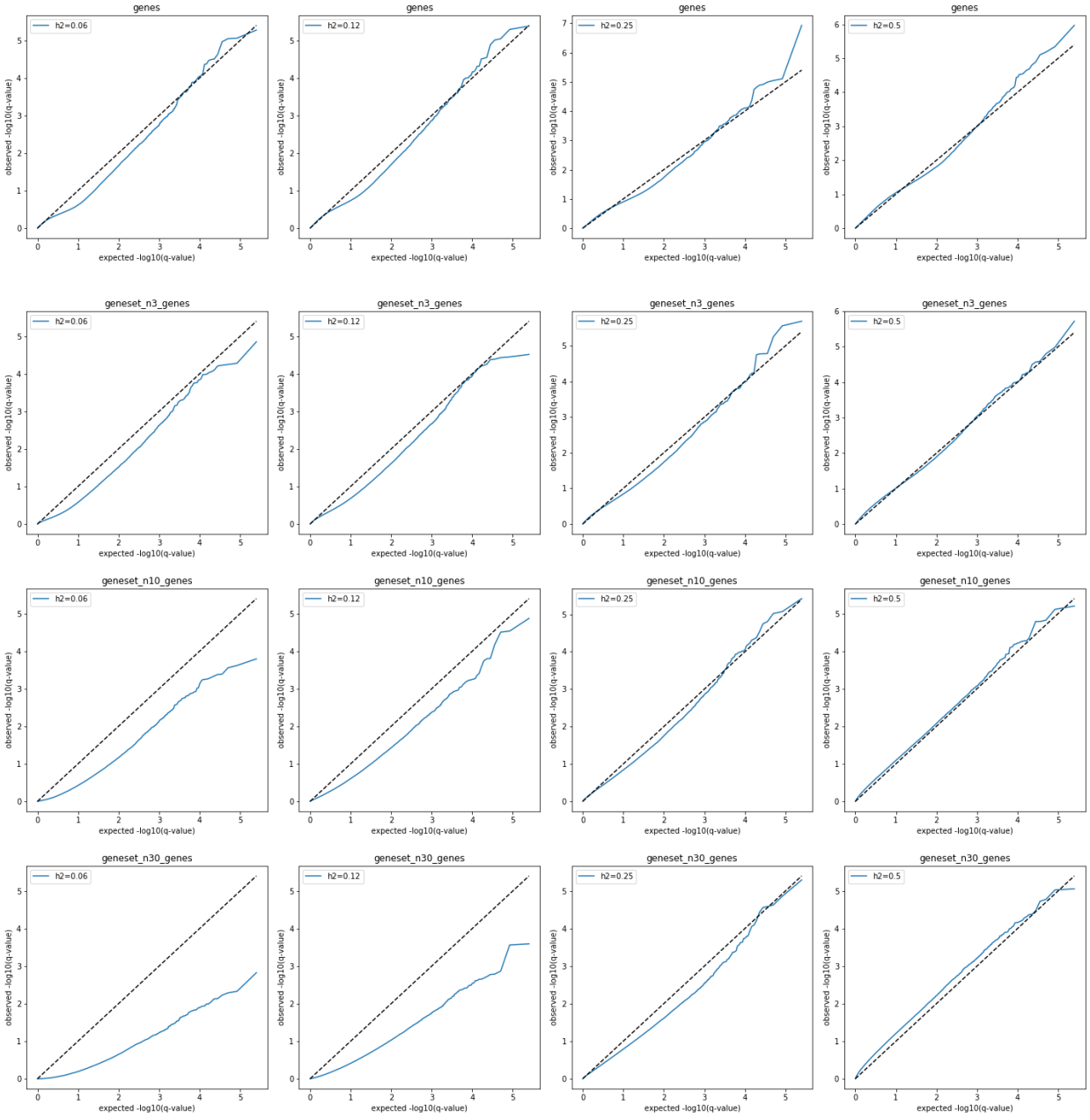
As introduced in the Online Methods, we apply Akaike Information Criterion (AIC) to identify genes and gene-sets that are improving the fit of the full model compared to the baseline model. We also evaluated

an alternative approach, deriving p-values using Wilks' theorem (chi-squared tail approximation), i.e. computing p-values as a tail probability of a χ_k^2 distribution for the likelihood ratio test (LRT) statistic: $P[\chi_k^2 > 2\log\left(\frac{L(\theta_{full})}{L(\theta_{part})}\right)]$, where θ_{full} and θ_{part} models are defined in the Online Methods, and k is the number of parameters optimized for the genomic region being tested. Wilks' theorem is known to have limitations, for example in a situation when the null hypothesis constrains parameters to a value that is close to the boundary of the parameter space[6]. Intuitively, this happens if some parameters of the full likelihood function are only partly effective, reducing effective degrees of freedom, and resulting in a null distribution of the LRT statistic that is different from the chi-square distribution.

To mitigate this limitation, one must fit an empirical null distribution, for example as implemented in LDAK-GBAT[7] for testing if individual genes have a non-zero heritable component. In LDAK-GBAT, the LRT statistic was parametrized as a mixture of the scaled gamma function and the Dirac delta function, with parameters of the mixture fitted from permutations. The key difference between statistical inference in LDAK-GBAT and GSA-MiXeR models lies in their null hypotheses. LDAK-GBAT aims to identify genes with a non-zero heritability, so under its null hypothesis GWAS z-scores have a well-defined distribution (a multivariate normal with covariance matrix equal to the LD matrix, which in practice can be approximated based on the reference dataset). GSA-MiXeR, in line with other competitive GSA tools, considers a null hypothesis whereby the trait's heritability is uniformly distributed across all genes, as modeled through a single effect-size variance parameter σ^2 shared across all genes. This null hypothesis precludes sampling GWAS z-scores under the null, and leaves no options for calibrating the LRT statistic distribution under the null hypothesis.

A naïve implementation of Wilk's theorem led to conservative p-values, over-corrected due to parameters that were ineffective. We observed this in simulations (see Supplementary Note Figure 1), where we synthesized phenotypes without gene enrichment, at varying levels of total heritability (h2=0.06, 0.12, 0.25, and 0.5). The simulations were conducted on a reduced sample of N=30,000 subjects from UK Biobank, and used chr1 only, performing 400 runs with N=300 randomly chosen genes, resulting in 1.2 million tests for each level of heritability, and using QQ plots to visualize the results of the LRT with p-values calculated using Wilk's theorem. The QQ plots drop below the null curve, indicating overly conservative p-values. While this does not preclude LRT application to identify individual genes, the issue became restrictive once such p-values are aggregated at the level of gene-sets. Our conclusion is that GSA MiXeR's LRT statistic is informative in differentiating noisy estimates vs truly enriched genes and gene-sets. However, it is not practical to find the exact shape of LRT distribution under the null hypothesis ("non-zero heritability uniformly distributed across genes") that GSA-MiXeR aims to test, precluding proper calibration of the p-values.

Supplementary Note Figure 1. *Quantile-quantile plots of gene and gene-set enrichment p-values in simulations under null, aggregated across N=400 runs using UKB chr1, with N=300 genes tested in each run, covering scenarios with simulated trait's heritability of 0.06, 0.12, 0.25, and 0.50. Gene-sets were generated by randomly selecting N=3, 10, or 30 genes. The p-values are computed from the LRT statistic with chi-squared tail probability (Wilk's theorem).*



Comparison of gene- vs gene-set heritability models

An important question raised for the GSA-MiXeR model is whether to allow individual variances of effect sizes at the level of individual genes, or only at the level of gene-sets. Stratified QQ plots (Supplementary Figure 3) show large heterogeneity in GWAS signal across genes. We also observed similar results comparing GSA-MiXeR with sLDSC (see Supplementary Table 11), where both tools again show a substantial variation in fold enrichment estimates across individual genes show within gene-sets. Given this heterogeneity, it's naturally more appropriate to fit an individual parameter for each gene, rather than an overall parameter for each gene-set.

Sensitivity analysis (Supplementary Figure 4) attempted to further evaluate the effect of modeling assumptions on the resulting fold enrichment estimates. We compared the three alternatives: (A) the default GSA-MiXeR model, allowing effect size variance to depend on genes; (B) "gene-set" model, allowing effect size variance to depend on gene-sets, but not on individual genes; (C) a "hybrid" model allowing effect size variance to depend on both genes and gene-sets.

Additionally, to compare models A, B and C, we have computed the difference in log-likelihood between these models, after making sure that all those models use exactly the same set of SNPs and the same SNP weights in their fit procedure. The log-likelihood values (shown in Supplementary Table 12) indicate that, for both schizophrenia and height, model (B) has substantially worse fit than models (A) and (C). For comparison, Supplementary Figure 5 includes a similar table comparing different strategies for modeling MAF- and LD- dependent architectures, as well as the impact of not modeling the functional annotations in the baseline GSA-MiXeR model. Both alternatives also negatively affected the log-likelihood of the model, but to a much smaller extent than the choice between models (A) and (B). This suggests that model (A) fits the data better than model (B). However, we also acknowledge that the direct comparison of the model's log-likelihood here did not account for the different number of parameters: 18,201 genes in model (A), and 10,475 gene-sets for model (B). Despite this, only a subset of parameters has a noticeable effect on the model's likelihood, e.g. only 363 genes (for schizophrenia) and 1131 genes (for height) had AIC>0 according to Supplementary Tables 9 and 10.

To account for the difference in effective number of parameters in the models (A) and (B), we re-run gene-level analysis restricting it to N=10,475 genes, yielding model (D) with same number of parameters as in the gene-set model (B). The genes were selected by sorting them on the length of the gene (in base-pairs) and retaining N=10,475 longest genes. This new model, now with the same number of parameters as the gene-set model, still had a substantially better log-likelihood, as shown in Supplementary Table 12, indicating that it is more appropriate to model effect size variance for individual genes, rather than for gene-sets as they are currently defined.

Details of MAGMA, LDAK-GBAT, and sLDSC analyses

MAGMA v1.09b analysis was conducted using the same reference, gene, and gene-set definitions as the GSA-MiXeR analysis, except for selecting a random subset of N=1,000 individuals from the HRC reference genotypes. MAGMA commands were as follows:

```
magma --annotate window=10 --snp-loc $BFILE.bim --gene-loc $GENE --out $BFILE
magma --bfile $BFILE --pval $SUMSTATS --gene-annot $BFILE.genes.annot --out $OUT.magma
magma --gene-results $OUT.magma.genes.raw --set-annot $GENESET ---out $OUT
```

For the LDAK-GBAT analysis, we used LDAK version 5.2, with HRC genotype panel as a reference (same set of 11,980,511 variants and 23,152 subjects as was used to great GSA-MiXeR LD reference), with `--extract` option to constrain the analysis to the overlapping SNPs that in GWAS summary statistics and HRC reference. Specific commands were as follows:

```
ldak5.2.linux --cut-genes $OUT --gene-buffer 10000 --bfile $BFILE --genefile $GENEFILE
--by-chr YES --extract $SUMSTATS.overlapHRC.justrs

ldak5.2.linux --calc-genes-reml $OUT --summary $SUMSTATS  --bfile $BFILE --ignore-weights
YES --allow-ambiguous YES --power -0.25 --gene-prune 0.5 --gene-permutations 10 --extract
$SUMSTATS.overlapHRC.justrs

ldak5.2.linux --join-genes-reml $OUT
```

For the sLDSC analysis, we noted the original software returns a "LinAlgError" error when it is applied to annotations based on individual genes, in which case all of the SNPs that belong to a single category fall

into a single block of the jackknife procedure. To avoid this error, we modified the sLDSC software (see URLs), allowing `--n-blocks 0` option to bypass its block-jackknife procedure, and reporting missing values (N/A) for all standard errors, test statistics, and p-values. Also, we included `--save-ldsc-reference` option in GSA-MiXeR package, allowing us to save binary and LD-weighted annotations based on the same gene and gene-set definitions as used in GSA-MiXeR analysis, as well as the same reference based on HRC genotypes.

This comparison between GSA-MiXeR and sLDSC was limited to one trait (schizophrenia) and a specific set of 36 gene-sets, implicated in exploratory GSA-MiXeR analysis. This is because expanding the comparison to all gene-sets and 21 traits faced practical difficulties with applying sLDSC tool to GSA. Based on our results for sensitivity analysis (see Supplementary Figure 4) we assumed that sLDSC model also needs an individual parameter for each gene within gene-set being tested. For this reason, each gene-set requires its own sLDSC run to estimate its fold enrichment of heritability – indeed, the annotation matrix that we use with sLDSC contains one column for the gene-set, and additional columns for each gene within gene-set (the columns are added to the baseline set of 75 functional annotation categories from sLDSC baseline-v2.2 model). sLDSC has not been evaluated with 100+ functional annotations, and additionally, it uses dense matrices to represent both binary & LD-weighted annotations, which makes it impractical to pull together multiple gene-sets in a single sLDSC analysis, and at the same time allow for each gene to have its own effect size variance parameter.

Computational requirements

GSA-MiXeR core functionality is implemented in C++, using OpenMP to parallelize execution across cores, with command-line interface and optimization procedures implemented in Python. The source code of the software is publicly available and is released with an open-source license (GPL-v3). We also provide a singularity container with the software and all dependencies. The tool is intended for high-performance computing (HPC) environments. An analysis requires 20 jobs, each requiring up to 36 GB of RAM and typically taking around 7 hours to complete using 8 CPU cores.

All analyses were performed on the Colossus HPC cluster provided by the Services for Sensitive Data (TSD) at the University of Oslo (UIO). Each compute node is equipped with two AMD EPYC 7551/7452 32-core CPUs, and a total of 501.5 GB RAM available per node. Based on the multithreaded scaling performance of the GSA-MiXeR implementation we currently recommend assigning 8 threads and 36 GB RAM to each analysis; the most up-to-date usage instruction will be provided in the software tutorial. We note that relatively high memory usage of the GSA-MiXeR is due to storing sparse LD matrices in memory, rather than storing a set of pre-computed LD scores (an approach that is more often implemented in other software packages). Despite increasing its memory requirements, sparse LD matrix representation offers a few advantages: for example, it enables “full” log-likelihood evaluation (see Supplementary Note 1) by randomly drawing causal SNPs; it also allows users to define their own functional annotations, genes, and gene-sets without a need to pre-computing LD scores towards these custom SNP categories.

References

1. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).
2. Roda, G. *et al.* Crohn’s disease. *Nature Reviews Disease Primers* **6**, 22 (2020).
3. 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).

4. Barrera-Chimal, J., Jaisser, F. & Anders, H.-J. The mineralocorticoid receptor in chronic kidney disease. *British Journal of Pharmacology* 179, 3152-3164 (2022).
5. Kranzler, H.R. & Soyka, M. Diagnosis and Pharmacotherapy of Alcohol Use Disorder: A Review. *JAMA* 320, 815-824 (2018).
6. Susko, E. Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika* 100, 1019-1023 (2013).
7. 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).