The power of pathway-based polygenic risk scores

Shing Wan Choi
Icahn School of Medicine at Mount Sinai
Judit Garcia-Gonzalez
Icahn School of Medicine at Mount Sinai
Yunfeng Ruan
Broad Institute
Hei Man Wu
https://orcid.org/0000-0003-1559-7586
Jessica Johnson
Clive Hoggart
Icahn School of Medicine at Mount Sinai
Paul O'Reilly (paul.oreilly@mssm.edu)
Icahn School of Medicine at Mount Sinai https://orcid.org/0000-0001-7515-0845

Biological Sciences - Article

Keywords: Genetic Liability, Functional Genome Sub-structure, GWAS Signal Enrichment, Heterogeneity

Posted Date: June 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-643696/v1

The power of pathway-based polygenic risk scores

Shing Wan Choi[1*], Judit García-González[1*], Yunfeng Ruan[2], Hei Man Wu[1], Jessica Johnson[1], Clive J Hoggart[1], Paul F. O'Reilly[1]

[1] Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA
[2] The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
* These authors contributed equally to this work

Polygenic risk scores (PRSs) have been among the leading advances in biomedicine in recent years. As a proxy of genetic liability[1], PRSs are utilised across multiple fields and applications[2]. While numerous statistical and machine learning methods have been developed to optimise their predictive accuracy[3–5], all of these distil genetic liability to a single number based on aggregation of an individual's genome-wide alleles. This results in a key loss of information about an individual's genetic profile, which could be critical given the functional sub-structure of the genome and the heterogeneity of complex disease. Here we evaluate the performance of pathway-based PRSs, in which polygenic scores are calculated across genomic pathways for each individual, and we introduce a software, PRSet, for computing and analysing pathway PRSs. We find that pathway PRSs have similar power for evaluating pathway enrichment of GWAS signal as the leading methods[6,7], with the distinct advantage of providing estimates of pathway genetic liability at the individual-level. Exemplifying their utility, we demonstrate that pathway PRSs can stratify diseases into subtypes in the UK Biobank with substantially greater power than genome-wide PRSs. Compared to genome-wide PRSs, we expect pathway-based PRSs to offer greater insights into the heterogeneity of complex disease and treatment response, generate more biologically tractable therapeutic targets, and provide a more powerful path to precision medicine.

As proxies for genetic liability to human traits or diseases [1], polygenic risk scores (PRSs) have been applied in numerous applications, including prediction of disease risk [8–13], patient stratification [14], investigation of treatment response [15–18] and genetically-informed experimental perturbation [19,20]. Most leading PRS methods, including those that incorporate functional annotation [21,22], are based on the classical polygenic model of disease, which assumes that individuals lie on a linear spectrum from low to high genetic risk and that summarises an individual's genetic profile to a single value estimate of liability. This incurs substantial loss of information about an individual's genetic profile, such as how the burden of genetic risk varies across biological processes and pathways. This information may be more informative for many applications of PRS, such as patient stratification and prediction of treatment response.

Here we evaluate a new polygenic risk score approach that accounts for genomic sub-structure, constitutes an extension to the classic polygenic model of disease, and may better reflect disease heterogeneity (**Fig. 1a**). Instead of aggregating the estimated effects of risk alleles across the entire genome, pathway-based PRSs aggregate risk alleles across *k* pathways (or gene sets) separately. Therefore, rather than a single genome-wide PRS, each individual has *k* PRSs corresponding to *k* pathways across the genome. Pathways can be defined in a variety of ways, including by existing canonical databases (e.g. KEGG, REACTOME [23,24]), by analytically derived modules of e.g. gene co-expression or protein-protein interactions, or from functional output of experimental perturbations [25–27]. Ideally, pathways reflect the encoding of different biological functions, separable in the same way that different environmental risk factors, such as smoking or dietary factors, are considered separately in epidemiological prediction models. From this perspective, GWAS results can be considered a composite of signal corresponding to function encoded by different genomic pathways (**Fig. 1b**).

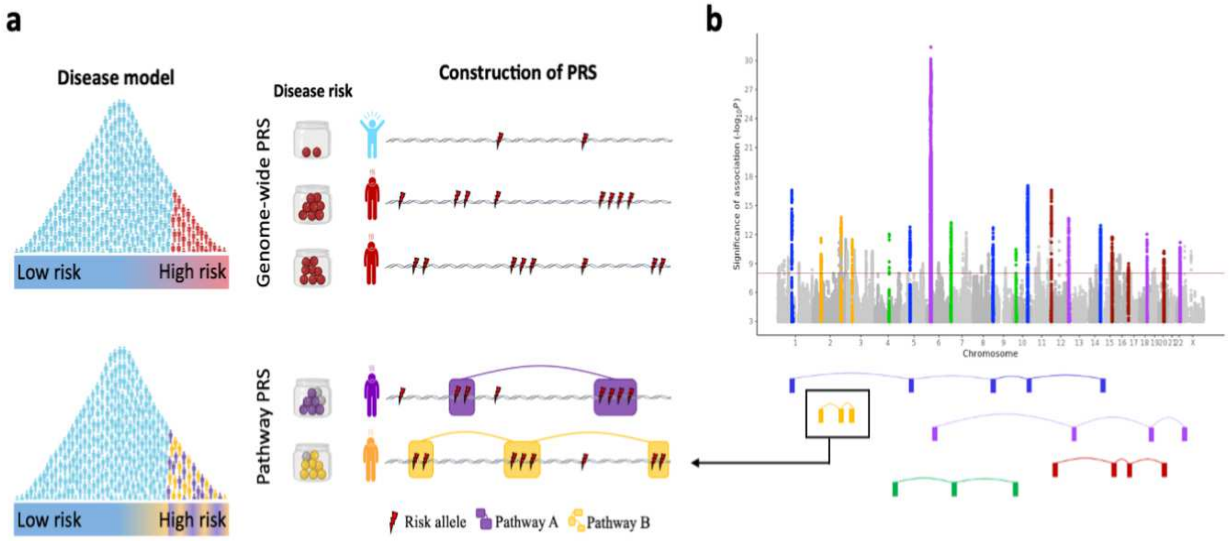


Fig. 1. The pathway polygenic risk score approach. Coloured boxes represent genes, lines link genes that are within the same genomic pathway. **a,** Upper model: Classical polygenic model of disease, in which individuals lie on a linear spectrum from low to high risk and genome-wide PRSs are constructed as the sum of risk alleles across the genome. Disease risk depicted by the Jar model [28]. Lower model: Pathway polygenic model of disease, in which multiple liabilities are considered and PRSs are constructed by aggregating risk alleles over different genomic pathways. **b,** GWAS results Manhattan plot illustrated as a hypothetical composite of signals that each correspond to alternative functional routes to disease. Pathways that only make a small contribution to disease risk across the population, or a contribution in a small fraction of individuals (e.g. nicotine receptor pathway in those individuals who smoke), are likely to harbour risk variants of relatively small effect.

A key concern in application of PRS computed over relatively short genomic regions is whether they are sufficiently powered to be useful. Here we show, for the first time, that polygenic risk scores capture genetic signal at the pathway-level with comparable power to that of leading pathway enrichment methods MAGMA [6] and LD score regression (LDSC) when applied to large target sample sizes [29,7]. Therefore, pathway PRS may be powered for a range of other

applications for which genome-wide PRS are presently used. To test this premise using real data, we perform a head-to-head performance comparison of pathway-specific versus genome-wide PRS for patient stratification, utilising the large number of individuals in the UK Biobank with common diseases such as type 2 diabetes (T2D), coronary artery disease (CAD), hypercholesterolemia (HC) and their comorbidities. We show that our pathway PRS method outperforms both genome-wide PRS alternatives, lassosum and PRSice, in 22 out of 23 supervised stratification scenarios tested, often by a wide margin. We expect the power of pathway PRS to improve substantially in the future with improved functional annotation and more context-specific definition of pathways. Our new method and accompanying software (called PRSet) builds on the popular PRSice genome-wide PRS tool [3,30] and is likewise user-friendly, fast, intuitive and openly available.

PRSet model overview

Our PRSet method for calculating pathway-based PRSs leverages the classical genome-wide PRS method[1] - clumping + thresholding (C+T) - to calculate *k* PRSs corresponding to *k* genomic pathways for an individual *i*, as follows:

$$PRS_{ik} = \sum_{j=1}^{m_k} \beta_j G_{ij}$$

where m_k is the number of clumped SNPs in pathway *k*, β_j is the SNP effect size estimated from a GWAS on the studied phenotype, G_{ij} is the genotype of individual *i*. In contrast to the C+T method, where SNPs are clumped across the whole genome, PRSet performs clumping on each pathway independently. This is done to retain pathway signal and to account for correlation between SNPs in nearby genes of the same pathway (**Supplementary Fig. 1**). The use of the *P*-value thresholding procedure is dependent on the use-case. For example, while *P*-value

thresholding is not performed in pathway enrichment analyses, it is performed in each pathway independently in the disease sub-typing application of this study (see **Methods**). Each pathway PRS can be tested for association with a phenotype of interest in a target sample by regressing the phenotype on the PRS, as in standard PRS analyses. Pathway enrichment is evaluated by computing an empirical "competitive" *P*-value, which accounts for pathway size using permutation (see **Methods**). Many applications of standard genome-wide PRS can be adapted to pathway PRS, the analyses of which can be evaluated and reported similarly.

In the next section, we benchmark the power of pathway PRS for assessing pathway enrichment, versus MAGMA and LDSC, using a range of comparisons that can be separated into (i) those that use canonical pathways, and (ii) those that define pathways by tissue and cell-type specific gene expression. These analyses are performed only to benchmark the methods' relative power as pathway enrichment tools, in particular to assess how well pathway PRS capture GWAS signal and, thus, their potential for wider use, and are not optimized to produce novel findings, e.g. for cell-type specificity.

Evaluating pathway enrichment

Canonical pathways: In this sub-section, pathways are defined using six publicly available databases (Biocarta [31], Pathway Interaction Database [32], Reactome [24], Mouse Genome Database [33], KEGG [23] and GO [34,35]) and pathway enrichment of genetic signal is tested by (i) a simulation study, (ii) real data using *MalaCards* gene scores (**Methods**). First, we performed simulations in which quantitative traits of varying heritability (h^2 = 0.1, 0.3, 0.5) were simulated from real genotype data of UK Biobank individuals, with causal variants randomly spread across 50 (real) pathways that were randomly selected from the pathway databases (**Methods**). GWAS were then performed on 250*k* individuals and their simulated traits, and an additional 1*k*, 10*k* and 100*k*

individuals were selected as target data. PRSet, MAGMA and LDSC were applied to the data (**Methods**) to test for pathway enrichment and the pathways were subsequently ranked by their inferred enrichment and compared to those of the known simulated enrichment via Kendall's correlation. This process was repeated 20 times. **Figure 2a** displays the results, showing best overall performance for MAGMA (Kendall $\tau^2 = 26.0\%$), then PRSet (Kendall $\tau^2 = 15.2\%$) and then LDSC (Kendall $\tau^2 = 8.62\%$). All methods perform better with larger h^2 and target sample sizes, with PRSet the best-performing method in large (100k) target data.

Next, we apply the three methods to the real data of the UK Biobank, and that of publicly available GWAS, across six traits: low-density lipoproteins (LDL), CAD, schizophrenia, BMI, Alzheimer's disease (proxy status) and alcohol consumption. Since the true GWAS pathway enrichment of each pathway is unknown, we produce a *disease relevance score* for each pathway by summing *MalaCards* gene scores (**Methods**), which assign values to genes based on systematic phenotype-specific text-mining of the literature (note that most genes are assigned a value of 0). In **Figure 2b**, we report the Kendall's correlations between the pathway enrichment ranked by the three methods versus the *MalaCards disease relevance scores*. While performance varies widely depending on pathway resource (**Fig. 2b**) and trait (**Supplementary Fig. 2**), the three methods show broadly similar results, with PRSet having the highest mean correlation ($\tau = 0.078$) between its pathway enrichment ranks and those of the *MalaCards* scores, followed by MAGMA ($\tau = 0.050$) and LDSC ($\tau = 0.043$). We also repeated the analysis removing all genes with *MalaCards* scores greater than 0 to examine evidence of pathway enrichment among genes not yet highlighted in the literature and found that the correlations were eliminated (**Supplementary Figs. 3, 4**). This may indicate that the methods have limited power to identify weak effects across pathways or that only a modest fraction of genes in pathways influence disease contribution to risk.

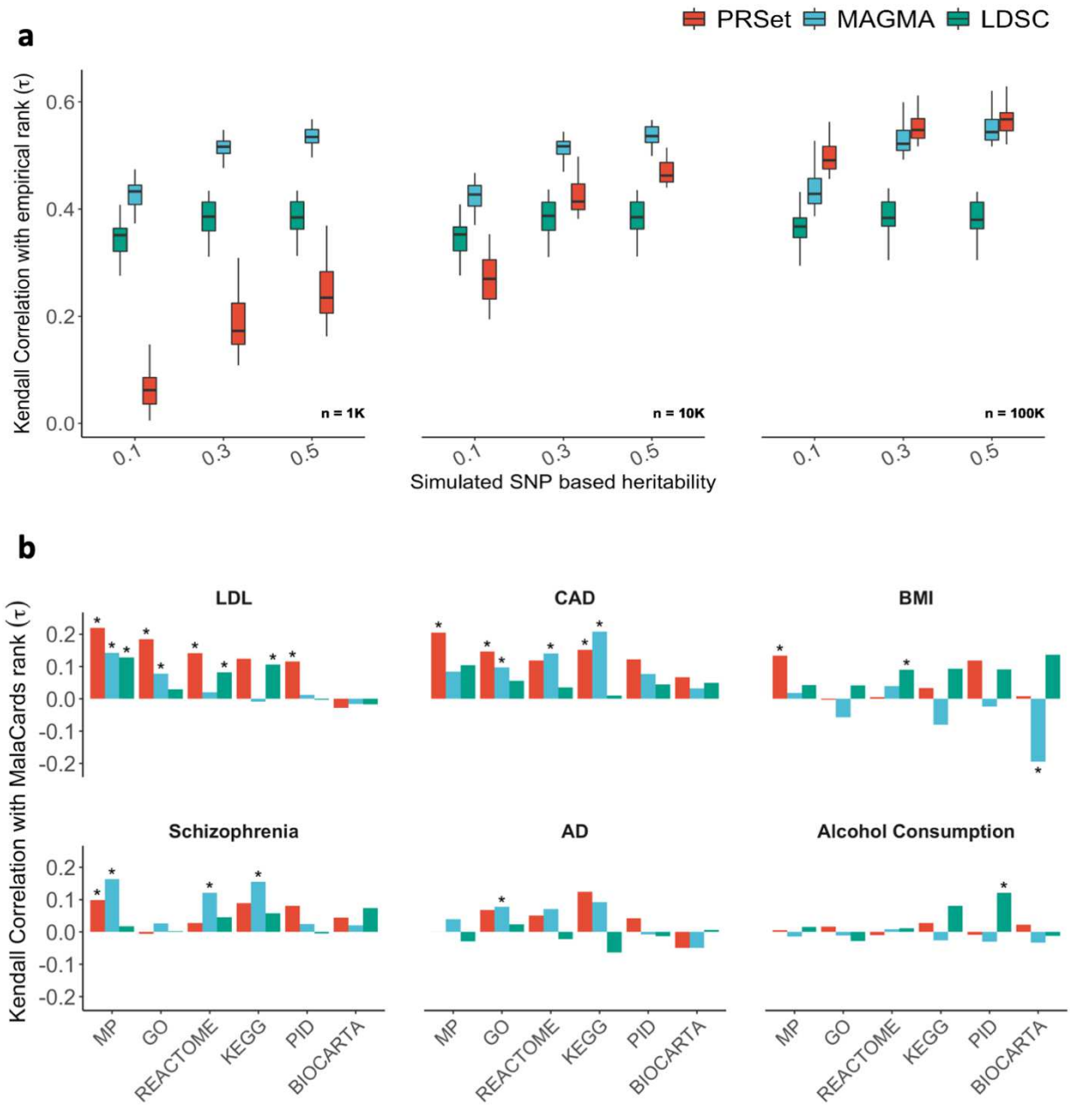


Fig. 2. **Evaluating pathway enrichment using canonical pathways. a,** Simulation study. Performance was defined as the Kendall correlation between the pathway ranks based on competitive *P*-values of enrichment computed by each software and the empirical pathway ranks based on the true (simulated) effects across the pathways. Boxplots illustrate the values of Kendall rank correlation coefficients (τ) for PRSet, MAGMA and LDSC for each combination of heritability (h^2 = 0.1, 0.3, 0.5) and target sample size n = (1K, 10K 100K). **b,** Real data study. Kendall correlation coefficients (τ) between pathway ranks based on

146 competitive *P*-values of enrichment computed by each software and pathway ranks based on *MalaCards*
147 disease relevance scores. *Empirical *P*-value < 0.05. MP, Mouse Genome Database; GO, Gene Ontology
148 database; KEGG, Kyoto Encyclopaedia of Genes and Genomes; PID, Pathway Interaction Database; AD,
149 Alzheimer's disease (proxy status).

150

151 **Pathways defined using tissue/cell-type specificity**: To further interrogate the power of PRS
152 to capture genetic signal at the pathway-level compared to MAGMA and LDSC, we compared
153 their performance in tissue/cell-type expression specificity analyses using the approach
154 introduced in *Skene et al* 2018 [36]. In this approach, genes are grouped into 11 quantiles of
155 increasing expression specificity based on expression reported across 47 bulk-tissues and 24
156 brain cell types (**Methods**). Tissue and cell types with greatest GWAS signal – as evaluated by
157 MAGMA and LDSC (and here PRSet) – among quantiles with the most specific expression, are
158 thus implicated in disease etiology. The pathway enrichment of the genes in the top quantile is
159 evaluated as standard for canonical pathways, which we refer to as the 'top quantile' test model,
160 while the linear trend of enrichment is also assessed and referred to as the 'linear' test model (see
161 **Extended Data Fig. 1**, **Methods**).

162 Here we perform these analyses in the same data and traits used in the previous section. In the
163 absence of well-established roles for individual tissue and cell types in these outcomes, we sought
164 *a priori* candidates from two domain experts for each outcome to provide an agnostic way to
165 evaluate the performance of the different methods in this setting (**Methods**). We observed
166 significant associations between expert opinion and the tissue-type specificity results, although
167 results varied substantially depending on the pathway method and test model used (**Fig. 3a**).
168 Associations were strongest for schizophrenia (**Fig. 3b-c, Extended Data Fig. 2**) and BMI
169 (**Extended Data Fig. 3**), in which MAGMA and LDSC had a higher correlation with expert's
170 opinion than PRSet, while PRSet had best performance in Alzheimer's disease (**Extended Data**

Fig. 4) and CAD (**Extended Data Fig. 5**). The associations relating to the cell-type specific analyses were relatively weak (**Fig. 3d**), with significant results only observed for MAGMA and PRSet in relation to schizophrenia. For AD, the strongest and only significant result was that of PRSet implicating microglia using the top quantile test model (**Fig. 3e**), which is notable since microglia has been extensively linked to AD etiology in the literature [37]. However, individual results reported here should be treated with caution, since they appear highly sensitive to the test model (top quantile / linear) and the number of quantiles used (**Extended Data Fig. 2-7**). Moreover, there have been several extensions of the *Skene et al* approach, including an extension of MAGMA designed specifically for tissue/cell-type analyses that likely has substantially higher power than the standard MAGMA enrichment tool used here [38]; the basic version of MAGMA as an enrichment tool was used here to enable like-for-like comparisons with PRSet and LDSC regarding power to capture pathway signal.

Our results benchmarking these pathway enrichment tools in multiple settings suggest that PRSet has broadly comparable power to capture genetic signal in pathways as MAGMA and LDSC, with the distinct advantage of providing individual-level estimates of pathway liability, which could be useful in a wide-range of applications. Below, we test the power of pathway PRS for one such application, that of disease stratification.

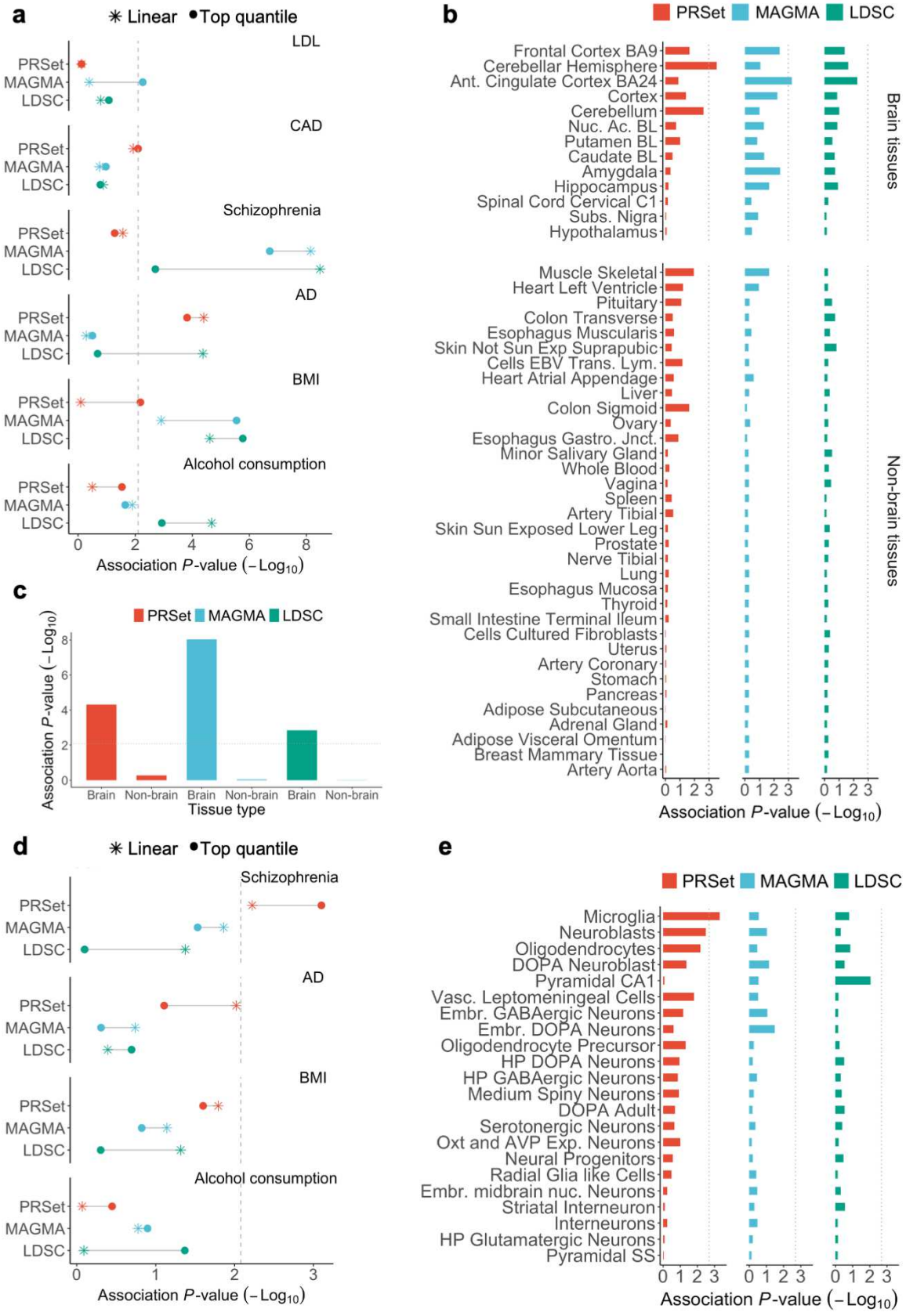

a
Linear
Top quantile
LDL
CAD
Schizophrenia
AD
BMI
Alcohol consumption
PRSet
MAGMA
LDSC
0
2
4
6
8
Association P-value (−Log10)
b
PRSet
MAGMA
LDSC
Frontal Cortex BA9
Cerebellar Hemisphere
Ant. Cingulate Cortex BA24
Cortex
Cerebellum
Nuc. Ac. BL
Putamen BL
Caudate BL
Amygdala
Hippocampus
Spinal Cord Cervical C1
Subs. Nigra
Hypothalamus
Brain tissues
Muscle Skeletal
Heart Left Ventricle
Pituitary
Colon Transverse
Esophagus Muscularis
Skin Not Sun Exp Suprapubic
Cells EBV Trans. Lym.
Heart Atrial Appendage
Liver
Colon Sigmoid
Ovary
Esophagus Gastro. Jnct.
Minor Salivary Gland
Whole Blood
Vagina
Spleen
Artery Tibial
Skin Sun Exposed Lower Leg
Prostate
Nerve Tibial
Lung
Esophagus Mucosa
Thyroid
Small Intestine Terminal Ileum
Cells Cultured Fibroblasts
Uterus
Artery Coronary
Stomach
Pancreas
Adipose Subcutaneous
Adrenal Gland
Adipose Visceral Omentum
Breast Mammary Tissue
Artery Aorta
Non-brain tissues
0 1 2 3
Association P-value (−Log10)
c
PRSet
MAGMA
LDSC
Association P-value (−Log10)
0
2
4
6
8
Brain
Non-brain
Brain
Non-brain
Brain
Non-brain
Tissue type
d
Linear
Top quantile
Schizophrenia
AD
BMI
Alcohol consumption
PRSet
MAGMA
LDSC
0
1
2
3
Association P-value (−Log10)
e
PRSet
MAGMA
LDSC
Microglia
Neuroblasts
Oligodendrocytes
DOPA Neuroblast
Pyramidal CA1
Vasc. Leptomeningeal Cells
Embr. GABAergic Neurons
Embr. DOPA Neurons
Oligodendrocyte Precursor
HP DOPA Neurons
HP GABAergic Neurons
Medium Spiny Neurons
DOPA Adult
Serotonergic Neurons
Oxt and AVP Exp. Neurons
Neural Progenitors
Radial Glia like Cells
Embr. midbrain nuc. Neurons
Striatal Interneuron
Interneurons
HP Glutamatergic Neurons
Pyramidal SS
0 1 2 3
Association P-value (−Log10)

Fig 3. **Performance of PRSet, MAGMA and LDSC for ranking of pathways defined by tissue and cell-type expression specificity. a,** Association between pathway enrichment *P*-value and expert opinion of tissue relevance for each software and six diseases. Colours indicate the software used to calculate enrichment: Red, *PRSet*; Blue, *MAGMA*; Green, *LDSC*. Results are shown for both the *top quantile* and *linear* specificity test methods (**Methods**). Dashed line corresponds to Bonferroni significance threshold of 0.05 for 6 tests (3 methods x 2 test models). **b,** Pathway enrichment results for schizophrenia under the top quantile test model. Bar plots show enrichment *P-value* for each tissue and pathway method. Dashed line corresponds to Bonferroni significance threshold for 47 tissues (-Log10(0.05/47) = 2.97). Ant, Anterior; Nuc. Ac., Nucleus Accumbens; BL, Basal Ganglia; Subs, Substantia; Exp, Exposed; EBV Trans. Lym., Epstein-Barr virus transformed lymphocytes; Gastro. Jnct, Gastroesophageal Junction. **c,** Enrichment of schizophrenia signal is higher in brain tissues vs non-brain tissues under the top quantile test model. Bar plots show the meta-analysis enrichment *P*-value using the Fisher's method for brain vs non-brain tissue and method. Dashed line corresponds to Bonferroni significance threshold for the 6 tests conducted. **d,** Associations between pathway enrichment *P*-value and expert opinion on cell type relevance for each software and four diseases. Colours indicate the software used to calculate tissue type enrichment: Red, *PRSet*; Blue, *MAGMA*; Green, *LDSC*. Dashed line indicates Bonferroni significance threshold of 0.05 for 6 tests (3 methods x 2 models). **e,** Pathway enrichment results for AD under the top quantile test model. Bar plots show enrichment *P*-values for each cell type and method. Dashed line corresponds to Bonferroni significance threshold for 22 cell types. DOPA, Dopaminergic, Vasc, Vascular; Emb, Embryonic; HP, hypothalamic; Oxt and AVP Exp Neurons, Oxytocin and Vasopressin Expressing Neurons; Nuc, Nucleus; SS, Somatostatin.

Pathway PRSs for disease stratification

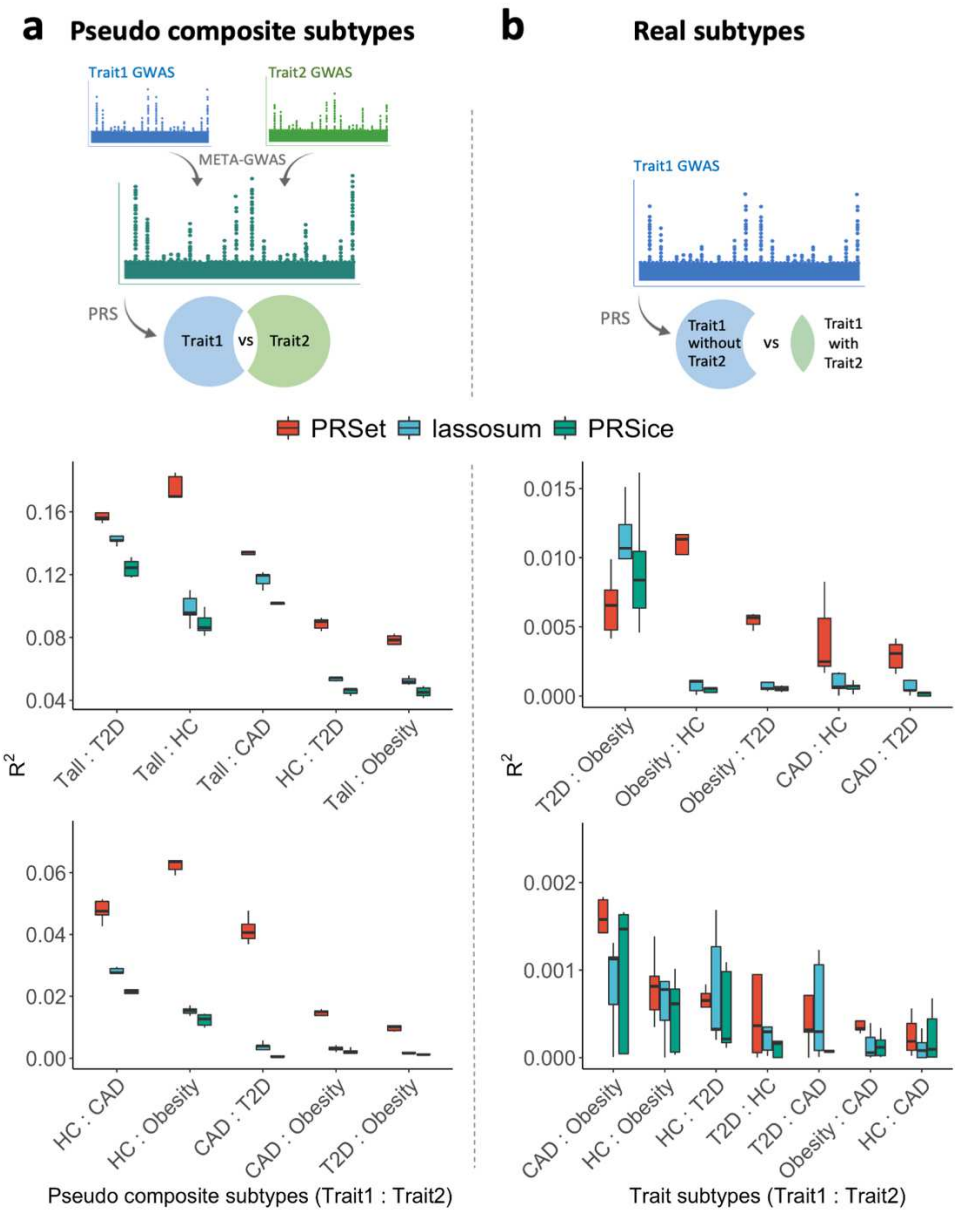
To evaluate the power of pathway PRSs, versus genome-wide PRSs, for performing disease stratification, or *sub-typing*, we leveraged the large number of individuals in the UK Biobank with major diseases: T2D (N = 19,668), CAD (N = 22,388), HC (N = 26,561), obesity (N = 92,818), and additionally "extreme height" (top 5^{th} percentile; N = 20,512) because height GWASs are presently the most powerful for PRS analyses (**Methods**). We compared the power of PRSet for sub-typing with two genome-wide PRS approaches, lassosum and PRSice (C+T), using two strategies. First, in the absence of well-established subtypes for these outcomes we produced "pseudo subtypes" by combining the 5 outcomes into pairs. We then meta-analysed the two GWAS of each pair for use as base data and applied each of the PRS methods to stratify the set of pooled UK Biobank patients (pooled across each pair, removing comorbid individuals), comparing inferred subtypes to the original patient groups (**Methods**). Supervised stratification was performed with 5-fold cross-validation, using lasso regression applied to the most enriched *k* pathway PRS predictors from PRSet (competitive *P*-value < 0.05), or standard regression with genome-wide PRS predictors generated by lassosum or PRSice (**Methods**).

PRSet had the highest subtype classification performance across all ten comparisons (**Fig. 4a**). Most of these showed strikingly higher discriminatory power for PRSet, notable since different PRS methods typically have similar predictive power [39–41]. We performed several sensitivity analyses to explore the causes of the relatively high power of PRSet here, repeating the analyses: i) with genes/pathways shifted by 5Mb to reduce their biological meaning (**Supplementary Note 2**), ii) using single diseases and testing case/control status instead (**Supplementary Note 3**), (iii) extending lassosum and PRSice to also utilize a lasso regression framework in their optimisation (**Supplementary Note 4**). Lower relative predictive power was observed for PRSet in each of these scenarios, indicating that the improvement in PRSet performance is partially due to (i) the

enrichment of signal in genes and/or pathways, (ii) the application to sub-typing, rather than case/control prediction, and (iii) the use of a regularisation framework (see **Discussion**).

Next, we defined subtypes of CAD, HC, T2D and obesity as the presence/absence of comorbidity within each pair of these four diseases (**Fig. 4b**). PRSs were calculated based on one of the disease GWASs before performing sub-typing. For example, PRSs based on CAD GWAS were used to discriminate between CAD patients with or without T2D, with or without HC, and with or without obesity. Classification performance estimates for these "real subtype" analyses were lower than for the pseudo subtypes, with R^2 estimates < 0.016 (**Fig. 4b**). In 11 of the 12 comparisons PRSet outperformed lassosum and PRSice, while for T2D with/without obesity, lassosum and PRSice outperformed PRSet.

Finally, we tested stratification performance in a disease with a relatively small number of patients in the UK Biobank but with well-established subtypes: inflammatory bowel disease (IBD) (N = 6,008) and its subtypes Crohn's disease (N = 2,101) and ulcerative colitis (N = 3,681). Supervised sub-typing was performed as above, with PRSet again showing higher discriminatory power (R^2 = 0.007, *P*-value = 1.2×10^{-6}) than PRSice (R^2 = 0.005, *P*-value = 0.01). Given the availability of a known number of subtypes for validation, we also performed unsupervised stratification using *K*-means clustering (K = 2; **Methods**) and found that neither PRSet (*P*-value = 0.69) or PRSice (*P*-value = 0.98) produced clusters that were associated with the two true subtypes.



257

258 **Figure 4. Performance of PRSs to classify trait subtypes.** Boxplots illustrate model R^2 from a 5-fold
259 cross validation approach. Pathway specific PRSs for 4,079 pathways were calculated using *PRSet*.
260 Genome-wide PRSs were calculated using lassosum and PRSice. **a,** Pseudo composite subtype analysis.
261 PRSs were calculated using effect sizes derived from meta-analysis of Trait1 and Trait2 GWASs. The target
262 sample was defined as pooled cases of Trait1 and Trait2, removing comorbid individuals. **b,** Real subtype
263 analysis. PRSs were calculated based on the GWAS of Trait1. The target sample was defined as cases of
264 Trait1 with/without comorbidity with Trait2.

Discussion

Here we have introduced a novel, pathway-based, polygenic risk score approach and software tool, PRSet, for performing pathway PRS analyses. We demonstrated that pathway PRSs can capture genetic signal across pathways with similar power to MAGMA and LDSC, with the distinct advantage of providing individual-level estimates of pathway liability. We do not presently recommend PRSet as an enrichment tool over these established methods, given its lower power under simulation in small target sample sizes (**Fig. 2a**). However, its capacity to capture genetic signal over pathways highlights the promise of pathway PRSs as higher-resolution, more biologically interpretable, alternatives to genome-wide PRSs.

Next, we directly compared the performance of the two PRS approaches in an application for which much hope is placed in genome-wide PRSs: disease stratification. In a series of supervised stratification analyses we found that PRSet outperformed genome-wide PRS methods, lassosum and PRSice, in almost all scenarios (22 out of 23) and often by striking margins. While we observed no significant unsupervised stratification of known IBD subtypes, we expect higher power when the corresponding pathway PRSs are trained in large case/control data. In contrast, genome-wide PRSs may be limited-by-design in their application to disease stratification because they are dominated by variants that affect multiple disease subtypes, while pathway PRSs allow for the independent influence of pathways that differentiate between subtypes.

Pathway PRSs, as outlined here, have two major limitations: (i) pathways are not well-defined and so are likely a poor proxy of biological function, (ii) variants are linked to pathways by location (i.e. inside genes) only (**Methods**). However, the rapid advances being made in functional genomics, via the integration of rich resources of multi-omics data now available, can help to address both issues. For example, future pathway PRSs could be enhanced so that pathways are defined according to robust differential gene co-expression or protein-protein interaction

networks and annotated by regulatory elements active in tissue and cell types relevant to the disease under study.

We believe that pathway-based PRSs may ultimately offer greater promise of delivering stratified medicine for complex diseases than genome-wide PRSs, which aggregate disparate forms of risk into a single number. However, despite promising early results for pathway PRSs reported here, they have several limitations that need addressing, some of which rely on field-level advances, before their potential can be realised. A better understanding of how genetics leads to biological function, and the role of pivotal genes in signalling and mechanistic cascades, will contribute to more accurate and powerful modelling of the multiple genetic liabilities underlying complex disease.

Our new method, PRSet, provides a novel approach to computing and analysing polygenic risk scores, motivated by the functional sub-structure of the genome and the heterogeneity of disease. Unlike genome-wide PRSs, pathway-based PRSs provide high-resolution information about an individual's genetic risk profile aligned to biological function, and thus have the potential to offer greater insights into disease and a more direct route to precision medicine.

Methods

Participants

UK Biobank

The UK Biobank (UKB) is a prospective multi-ethnic cohort of 502,493 participants, aged 40-69 years, initially recruited across the United Kingdom between 2006 and 2010, with follow up since. The UK Biobank genetic data used in this study included 488,377 samples and 805,426 SNPs.

Standard quality controls were performed, removing SNPs with genotype missingness > 0.02, minor allele frequency (MAF) < 0.01 and with Hardy Weinberg Equilibrium (HWE) *P*-value < $1x10^{-8}$. We removed all individuals who had withdrawn consent, who had a high degree of missingness or heterozygosity and who had mismatching genetically inferred and self-reported sex as reported by the UK biobank data processing team. We also removed individuals who were not of European ancestry based on a 4-mean clustering on the first two principal components, and related samples with kinship coefficient > 0.044 using a greedy algorithm, since present PRS methods have been shown to have relatively poor portability between global ancestries. A total of 387,392 individuals and 557,369 SNPs remained after quality control.

Sweden-Schizophrenia Population-Based cohort

Samples from the Sweden-Schizophrenia Population-Based cohort are a subset of the samples of the Psychiatric Genomics Consortium (PGC) Schizophrenia Working Group and were obtained from the database of Genotypes and Phenotypes (Study Accession: phs000473.v2.p2). Data processing and quality controls performed on these data are described elsewhere [42]. A total of 4,834 individuals diagnosed with schizophrenia and 6,128 controls were included.

Phenotypes

In order to optimise statistical power for benchmarking the performance of the methods tested in the study, we selected complex phenotypes with high SNP-heritability estimates, with publicly available summary statistics from large GWASs and that were measured in the UK Biobank or the Sweden-Schizophrenia Population-Based cohort (**Extended Data Table 1 and 2**). As such, we extracted data from the UK Biobank on the following phenotypes: BMI, LDL, CAD, alcohol consumption, standing height, T2D, and a proxy of AD based on parental history of the disease (**Supplementary Information**). Schizophrenia cases and controls were extracted from the Sweden-Schizophrenia Population-Based cohort.

GWAS Data Sets

GWAS data sets were downloaded on BMI [43], LDL [44], AD [45], CAD [46], T2D [47], alcohol consumption [48] and standing height [49] from public online databases and used without modification. Due to sample overlap between UKB and the Sweden-Schizophrenia Population-Based cohort, we used a version of the Sweden-Schizophrenia GWAS [42] with UKB data removed to prevent inflation of results.

Definition of pathways

KEGG [23], BioCarta [50], Pathway Interaction Database (PID) [51] and Reactome [52] canonical pathways were obtained from the Molecular Signatures Database (MSigDB v7.0) [53]. Pathways from the Gene Ontology database (GO, accessed on 2021-03-17) [34,35] and Mouse Genome Database (MGD, accessed on 2021-03-17) [54] were also included. For MGD pathways, we i) used the human-mouse homolog list provided by MGD to convert the mouse gene names to their human counterpart and ii) restricted our analyses to pathways with ontology level > 4 to avoid inclusion of pathways that are extremely specific. To avoid performing analyses on under-powered or overly genetic pathways, as standard in pathway analyses [6], we removed pathways

with fewer than 10 genes or more than 2000 genes to exclude over specific or too broad pathways. A total of 4,079 pathways across the six pathway database resources were included in the analyses.

Pathway enrichment analyses

PRSet

Pathway specific PRS analyses were performed using PRSice-2 (v2.3.5) on genotype data. The Major histocompatibility complex region (MHC, chr6:25Mb-34Mb) was removed for all the diseases assessed and the APOE region (chr19:44Mb-46Mb) was also removed for AD. SNPs were annotated to genes and pathways based on GTF files obtained from ENSEMBL (GRCh37.75). We extended the gene coordinates 35 kilobases (kb) upstream and 10 kb downstream of each gene to include potential regulatory elements. Ambiguous SNPs (A/T and G/C) and SNPs not present in both GWAS summary statistics and genotype data were excluded. 10,000 permutations were performed to obtain empirical "competitive" *P*-values, which account for pathway size, as follows: first, a "background" pathway containing all genic SNPs is constructed, and clumping is performed within this pathway. For pathways with *m* SNPs, *N* null pathways are generated by randomly selecting *m* "independent" SNPs from the "background" pathway. The competitive *P*-value can then be calculated as

$$competitive\ P-value = \frac{\sum_{n=1}^{N} I(P_n < P_o) + 1}{N + 1}$$

where *I*(.) is an indicator function, taking a value of 1 if the association *P*-value of the observed gene set (P_0) is larger than the one obtained from the n^{th} null set (P_n), and 0 otherwise. A pseudo-count of 1 is added to the numerator and denominator to avoid competitive *P*-values of 0 and conservatively counting the observed gene set as 1 potential null set [55]. One consideration of this

374 permutation procedure is that the smallest achievable competitive *P*-value is $1/(N+1)$, which
375 can lead to difficulties in ranking highly significant gene sets.

376 *MAGMA*

377 To directly compare the performance of PRSet vs MAGMA (v1.07b) given identical input data, we
378 removed all ambiguous SNPs and non-overlapping SNPs prior to MAGMA analyses. It is
379 important to note that this step is unnecessary for MAGMA and might negatively impact its
380 performance. After filtering, gene-based analyses were performed on summary statistics and
381 genotype data independently. Same as for PRSet analyses, a 35kb window upstream and a 10kb
382 window downstream were added to gene coordinates, the MHC region was excluded for all traits,
383 and the APOE region was excluded for AD. Gene-based results were meta-analysed using the
384 inbuilt `*--meta*` function and were subsequently used as input to the pathway analysis.

385 *LDSC*

386 Partitioned LD scores were calculated using the 1000 Genomes European genotype data [56].
387 Similar to PRSet, SNPs were annotated to genes and pathway with 35kb upstream and 10kb
388 downstream extension prior to calculation of LD scores. Ambiguous SNPs and non-overlapping
389 SNPs were removed prior to LDSC analyses to allow for direct comparison between PRSet and
390 LDSC. GWAS were performed on the target genotype data using PLINK v1.90b6.7 [57], and were
391 meta-analysed with the external GWAS summary statistics using METAL (2011-03-25)[58].
392 Partitioned LD score regression was then performed using LDSC v1.01 [7,29], with the MHC (all
393 traits) and APOE (AD only) regions excluded.

Evaluating pathway enrichment using canonical pathway definitions

Assessment of pathway enrichment by simulation

Generation of causal pathways

Out of 4,079 empirical pathways extracted from six publicly available collections (see "definition of pathways" section), we randomly select 50 pathways. These pathways were simulated with different levels of enrichment, ranging from 5-50% randomly assigned causal SNPs per pathway, with step size of 5%. The empirical enrichment level for all pathways included in our studies were calculated as the proportion of causal SNPs included. The simulation process was repeated 20 times and results over the 20 sets of simulations provided.

Phenotype simulation and sample selection

Simulation was performed using UKB genotype data to assess the performance of PRSet in enrichment analyses comparing to MAGMA and LDSC. Quantitative traits (*Y*) with SNP-based heritability (h^2) of 0.1, 0.3 or 0.5 were simulated as $Y=X\beta+\varepsilon$, where *X* is the standardized genotype matrix, ε is the random error defined as $\varepsilon \sim N\left(mean = 0, sd = \sqrt{var(X\beta)(1-h^2)}\right)$, and β is a vector of SNP effect sizes which follows a point-normal distribution $\beta \sim N\left(mean = 0, sd = \sqrt{h^2}\right)$, with non-causal SNPs assigned with $\beta = 0$.

For each trait, 250,000 individuals from European ancestry were randomly selected to generate the GWAS summary statistics using PLINK v1.90.b6.7. An independent set of either 1k, 10k or 100k individuals were then randomly selected as the target samples. Pathway analyses were performed as described in the previous sections.

Assessment of pathway enrichment using *MalaCards* relevance scores

To assess whether pathway enrichment results were in line with previous biological knowledge on the phenotypes of interest, disease-associated relevance scores for each pathway were constructed using information from the *MalaCards* database [59]. The *MalaCards* database provides a disease relevance score *for each gene* based on experimental evidence and co-citation in the literature. For the six diseases included in this analysis (schizophrenia, AD, alcohol consumption, LDL, CAD and BMI), we downloaded the *MalaCards* disease-associated relevance scores (Accessed on 2020-11-27, see **Extended Data Table 3** for disease terms used and number of genes). Next, we performed a rank normalization of the scores where, assuming that a disease has *n* genes with *MalaCards* scores, a score of (r+1)/(n+1) were assigned to each gene, with r being the inverse ranking of the gene with *MalaCards* score. Genes without a *MalaCards* score are assigned a score of 0. *MalaCards* provide gene information as gene symbols, which were transformed to ENSEMBL gene names.

To obtain disease-associated relevance scores *for each pathway*, we calculated the sum of the rank transformed *MalaCards* scores for the genes included in a pathway and divided by the number of genes in the pathway to account for pathway size.

Agreement between pathway enrichment results for PRSet, MAGMA and LDSC and the *Malacards disease relevance scores* was assessed by calculating the Kendall correlations between the $-\text{Log}_{10}$ competitive *P*-value generated by each pathway enrichment tool, and the *MalaCards* relevance score for each pathway.

Evaluating pathway enrichment using tissue/cell-type defined pathways

Calculation of tissue specificity from bulk-tissue RNA-sequencing data

To calculate tissue specificity across pathways, we obtained bulk-tissue RNA-sequencing gene expression data from 55 tissues from the GTEx consortium [60](v8, median across samples).

Tissues with less than 100 individuals, cancer related tissue types (e.g. EBV-transformed lymphocytes and Leukemia cell line), and testis (which were considered as an outlier [61]) were removed, retaining a total of 47 tissues. We filtered out all non-protein-coding genes and genes not expressed in any tissue.

Gene expression specificity was calculated by dividing the expression of each gene by its total expression across tissues [61]. The resulting gene expression specificity ranged from 0 (gene is not expressed) to 1 (gene is exclusively expressed in this tissue). Next, expression specificity of each tissue was divided into 11 quantiles following the approach introduced in *Skene et al 2018* [36], where the first quantile contained all non-expressed genes in a given tissue, and the 11th quantile contained the most specifically expressed genes. Genes within each quantile were grouped into a single pathway.

Defining the cell type specificity sets

Cell type specificity data were obtained from supplementary materials of Skene et al (2018) [36] which includes gene expression specificity information for 24 brain cell types obtained from single cell RNA-sequencing data. Again, expression specificity of each brain cell type was divided into 11 quantiles with the first quantile containing all non-expressed genes in a given cell type. Genes within each quantile were grouped into a single pathway.

Ranking the importance of cell type / tissue

To provide an objective estimate of tissue / cell-type importance for each phenotype, we invited two experts (per phenotype) who were blind to our experiment and algorithm design to provide their opinion on what cell-type(s) and tissue(s) are expected to be implicated for each disease context (**Extended Data Table 4**). The expert response was coded as “none” (both experts think tissue/cell type is not implicated), “single” (only one expert thinks a tissue/cell type is important) and “both” (both experts agree about the importance of a tissue/cell type).

Cell type and tissue specificity analyses

We used two testing strategies to assess the relationship between disease GWAS signals and tissue(s) / cell type(s) specificity. For the *Top quantile enrichment strategy* GWAS signals are enriched in the most specifically expressed genes [36,61,62]; whereas for the *Linear enrichment strategy* GWAS signals increase linearly with expression specificity [6,38]. The top quantile strategy reports the competitive *P*-value of the pathway defined by those genes in the top expression specificity quantile for each software and tissue/cell type. The linear enrichment strategy fits a linear regression with the -Log10 competitive *P*-value for each of the pathways defined by the expression specificity quantiles as dependent variable, and the quantile ranks as the predictor variable, and reports the one-sided *P-value* for a positive association.

The concurrence of the methods' ranking of the tissues / cell types with that of the experts within each disease for both the top quantile and linear enrichment strategies was measured by regressing the inverse normalized -Log10 *P*-value for the top quantile / linear enrichment strategies for each cell type / tissue against the expert opinion, coded as factor.

MAGMA has a specific model which accounts for expression specificity (`*--gene-covar*`). However, in favour of a more consistent analysis between the three software methods, this model was not used. It is thus possible that MAGMA can provide more powerful results using the dedicated model.

Results from the regressions against the expert confidence score assessed the association of the gene expression specificity and GWAS signal with the expert opinion for each pathway enrichment software under each of the two hypotheses.

Disease Stratification

Pseudo subtype analysis

Composite phenotypes were generated by combining two distinct phenotypes, including T2D, CAD, Obesity (defined as BMI > 30), extreme height (defined as top 5% each sex) and hypercholesterolemia (defined as LDL > 4.9 mmol/L) from the UK Biobank. To simplify the analysis, we removed controls and co-morbid cases so that the analysis was only performed in participants presenting either case outcome. Phenotypes were encoded mimicking sub-phenotypes of a given disease, for example, for the phenotype CAD-Obesity, samples with CAD (and not Obesity) were coded as 0 and those with Obesity (and not CAD) were coded as 1.

Comorbid subtype analysis

For the analysis of subtypes with presence/absence of comorbid diseases, we used T2D, CAD, Obesity and hypercholesterolemia, as these diseases present high comorbidity between them. For each disease, three traits were derived mimicking the subtypes of the disease as the presence/absence of the other disorders.

For computational expediency, sex, age, age of diagnosis (for CAD and T2D), genotyping batch, recruitment centre and first 15 principal components were adjusted for by using pseudo residuals from logistic regression analyses as the outcome variable in the PRS analyses.

Meta-analysis of GWASs

Previously published GWASs from each pair of traits were combined into a meta-analysis to obtain joint summary statistics for each composite phenotype. Meta-analyses were performed using METAL[58] with the sample-size weighted fixed-effects algorithm. To truly mimic a composite phenotype GWAS, only variants included in both GWAS summary statistics were retained.

Calculation of PRSs and evaluation of model performance

PRS analyses were performed using a 5-fold cross validation. For each iteration, samples were split into 80:20 training:validation subsets. Summary statistics of the "composite" phenotype were used as base sample, and UK Biobank samples with the residualised phenotypes were used as target sample (**Extended Data Tables 5 and 6**).

Pathway-specific PRSs for 4,079 pathways (see definition of pathways) were calculated using PRSet. Competitive *P*-values were calculated using 10,000 permutations and pathways with competitive *P*-value < 0.05 were defined as enriched. PRSs for the enriched pathways were recalculated using *P*-value thresholding, such that the predictive power of each PRS was maximized. The PRSs with the "best" predictive performance for each enriched pathway were then included in a generalized linear model with lasso regularization using the `*cv.glmnet*` function from the glmnet package (v4.0-2) in R with 5-fold cross-validation to select the lambda parameter that generates the smallest out-of-sample mean squared error (MSE). The resultant best fitting model was applied to the validation sample to calculate the model R^2.

We also performed genome-wide PRS analyses using lassosum and PRSice-2. Parameters (*P*-value thresholds for PRSice; penalty factor λ and soft-thresholding parameter *s* for lassosum) were optimised in the training samples and were applied to the validation samples to calculate the model R^2.

Stratification of inflammatory bowel disease subtypes

For the stratification of inflammatory bowel disease (IBD) subtypes, pathway and genome-wide PRS were calculated using IBD summary statistics as base sample, and cases of Crohn's disease vs cases of ulcerative colitis as target sample. Comorbid cases were excluded from the analysis. To adjust for sex, genotyping batch, recruitment centre and first 15 principal components, logistic

529 regressions were performed and the pseudo residuals were used as the outcome variable in the
530 PRS analyses.

531 PRS analyses were performed using a 5-fold cross validation approach. For each iteration,
532 samples were split into 80:20 training:validation subsets. For the training sample, we performed
533 a similar approach to that used for supervised classification of subtypes (See *Calculation of PRS*
534 *and evaluation of model performance*) but with the PRS standardised to have mean 0 and
535 standard deviation of 1. In this case, only PRSet and PRSice were used to calculate the PRSs,
536 and pathways with non-zero coefficients after the lasso regularisation were used in the
537 unsupervised algorithm step.

538 PRS for the pathways selected in the training step were calculated for the validation sample using
539 PRSet (for pathway PRS) and PRSice (for genome-wide PRS). The silhouette method was
540 applied to select the optimal number of centroids for K-mean clustering, and we used the optimal
541 number of centroids and also 2 centroids (when the optimal is not 2) for the down-stream k-mean
542 clustering. K-mean cluster groups were regressed against the subtypes to assess the
543 classification performance of the K-means algorithm.

544

545

546

547

548

Code Availability

The scripts used to perform quality control on UK Biobank data are available at https://gitlab.com/choishingwan/ukb_process. The scripts used in the current study are available at https://gitlab.com/choishingwan/prset_analyses. PRSet is a module within PRSice and is available on github repository [https://github.com/choishingwan/PRSice].

Acknowledgments

We thank the participants in the UK Biobank and the scientists involved in the construction of this resource. We thank Dr Kristen Brennand, Dr Jason Kovacic, Professor Alison Goate, Professor Ruth Loos, Dr Edoardo Marcora, Dr Alexander Charney, Dr Manav Kapoor and Dr Jacqueline Meyers for providing their expert knowledge for each specific disease. We thank Dr Conrad Iyegbe, Laura Sloofman, Collin Spencer, Chris Porras, Dr Zhe Wang and Dr Jiayi Xu for useful discussions and feedback. This research has been conducted using the UK Biobank Resource under application 18177 (P.F.O) and was supported by grants from the UK Medical Research Council (MR/N015746/1) and the National Institute of Health (R01MH122866) to PFO. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai, specifically the Minerva Supercomputer and the Mount Sinai Data Ark data commons. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, NHS, the NIHR or the Department of Health. Figure 1 was partially created using the resource BioRender.com.

Author Information

These authors contributed equally: Shing Wan Choi, Judit García-González

Corresponding authors

Correspondence to Shing Wan Choi and Paul O'Reilly.

Contributions

P.F.O conceived the design of the study, with S.W.C and Y.R providing critical contributions. S.W.C developed the computational methods and the PRSet software, with J.G.G, Y.R, H.M.W and P.F.O providing critical contributions. S.W.C, J.G.G and Y.R performed the analyses. J.J performed quality control of the Sweden-Schizophrenia Population-Based cohort data. P.F.O supervised the project, with C.J.H providing critical contributions. S.W.C, J.G.G and P.F.O wrote the manuscript, with critical feedback from all other authors.

References

1. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

2. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

3. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinforma. Oxf. Engl.* **31**, 1466–1468 (2015).

4. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

5. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).

6. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).

7. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

8. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

9. Musliner, K. L. *et al.* Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population. *JAMA Psychiatry* **76**, 516–525 (2019).

10. Zheutlin, A. B. *et al.* Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems. *Am. J. Psychiatry* **176**, 846–855 (2019).

11. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

12. Aung, N. *et al.* Genome-Wide Analysis of Left Ventricular Image-Derived Phenotypes Identifies Fourteen Loci Associated With Cardiac Morphogenesis and Heart Failure Development. *Circulation* **140**, 1318–1330 (2019).

13. Haas, M. E. *et al.* Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure. *Am. J. Hum. Genet.* **103**, 461–473 (2018).

14. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).

15. Zhang, J.-P. *et al.* Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First-Episode Psychosis. *Am. J. Psychiatry* **176**, 21–28 (2019).

16. Natarajan, P. *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* **135**, 2091–2101 (2017).

17. Mega, J. L. *et al.* Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet Lond. Engl.* **385**, 2264–2271 (2015).

18. Pain, O. *et al.* Antidepressant Response in Major Depressive Disorder: A Genome-wide Association Study. *medRxiv* 2020.12.11.20245035 (2020) doi:10.1101/2020.12.11.20245035.

19. Hoekstra, S. D., Stringer, S., Heine, V. M. & Posthuma, D. Genetically-Informed Patient Selection for iPSC Studies of Complex Diseases May Aid in Reducing Cellular Heterogeneity. *Front. Cell. Neurosci.* **11**, 164 (2017).

20. Dobrindt, K. *et al.* Publicly Available hiPSC Lines with Extreme Polygenic Risk Scores for Modeling Schizophrenia. *Complex Psychiatry* **6**, 68–82 (2020).

21. C, M.-L. *et al.* LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* (2018) doi:10.1101/375337.

22. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* **13**, e1005589 (2017).

23. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

24. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

25. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).

26. Szklarczyk, D. *et al.* STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).

27. Markowetz, F. How to Understand the Cell by Breaking It: Network Analysis of Gene Perturbation Screens. *PLOS Comput. Biol.* **6**, e1000655 (2010).

28. Austin, J. C. & Honer, W. G. Psychiatric genetic counselling for parents of individuals affected with psychotic disorders: a pilot study. *Early Interv. Psychiatry* **2**, 80–89 (2008).

29. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

30. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, (2019).

31. Nishimura, D. BioCarta. *Biotech Softw. Internet Rep.* **2**, 117–120 (2001).

32. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674-679 (2009).

33. Bult, C. J. *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.* **47**, D801–D806 (2019).

34. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

35. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

36. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

37. Hemonnot, A.-L., Hua, J., Ulmann, L. & Hirbec, H. Microglia in Alzheimer Disease: Well-Known Targets and New Opportunities. *Front. Aging Neurosci.* **11**, (2019).

38. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222 (2019).

673 39. Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-
674 standardized framework. *PLOS Genet.* **17**, e1009021 (2021).

675 40. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression
676 on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

677 41. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics*
678 **36**, 5424–5431 (2020).

679 42. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological
680 insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

681 43. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity
682 biology. *Nature* **518**, 197–206 (2015).

683 44. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat.*
684 *Genet.* **45**, 1274–1283 (2013).

685 45. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies
686 new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430
687 (2019).

688 46. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association
689 meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).

690 47. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in
691 Europeans. *Diabetes* **66**, 2888–2902 (2017).

692 48. Sanchez-Roige, S. *et al.* Genome-Wide Association Study Meta-Analysis of the Alcohol
693 Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am. J.*
694 *Psychiatry* **176**, 107–118 (2019).

695 49. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological
696 architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).

697 50. Nishimura, D. BioCarta. *Biotech Softw. Internet Rep.* **2**, 117–120 (2001).

698 51. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**,
699 D674–D679 (2009).

700 52. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**,
701 D649–D655 (2018).

702 53. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**,
703 1739–1740 (2011).

704 54. Bult, C. J. *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.* **47**, D801–
705 D806 (2019).

706 55. North, B. V., Curtis, D. & Sham, P. C. A Note on the Calculation of Empirical P Values
707 from Monte Carlo Procedures. *Am. J. Hum. Genet.* **71**, 439–441 (2002).

708 56. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

709 57. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
710 datasets. *GigaScience* **4**, (2015).

711 58. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of
712 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

713 59. Espe, S. Malacards: The Human Disease Database. *J. Med. Libr. Assoc. JMLA* **106**,
714 140–141 (2018).

60. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* 787903 (2019) doi:10.1101/787903.

61. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).

62. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

63. 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).

64. Sanchez-Roige, S. *et al.* Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am. J. Psychiatry* **176**, 107–118 (2018).

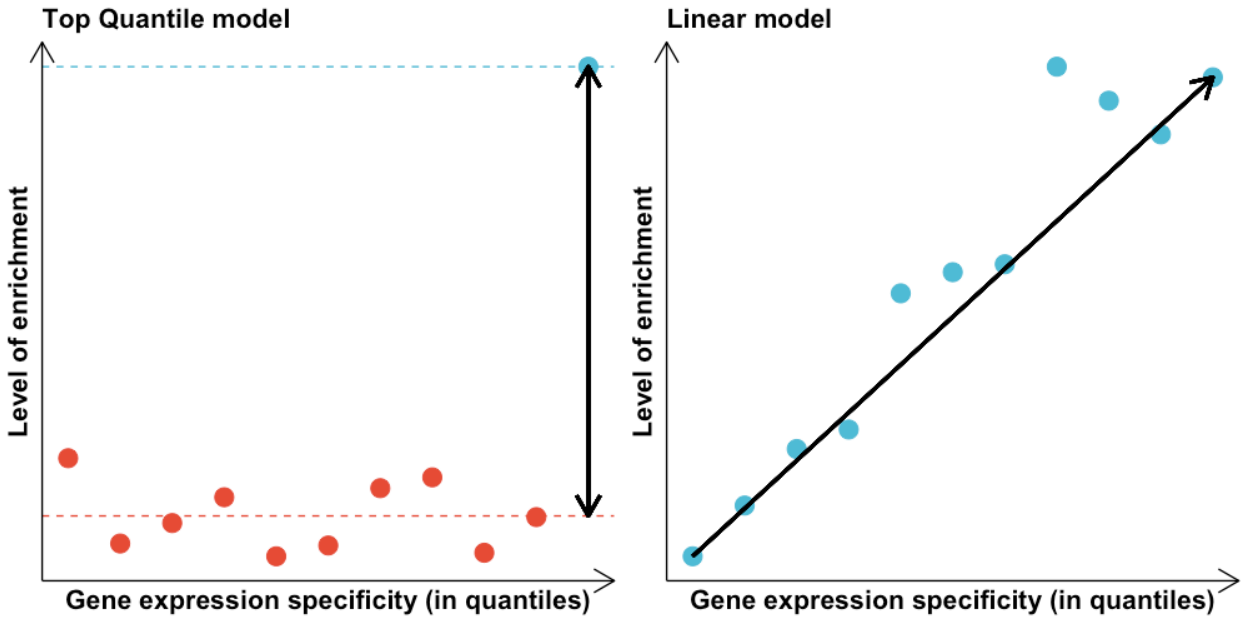
65. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).

66. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

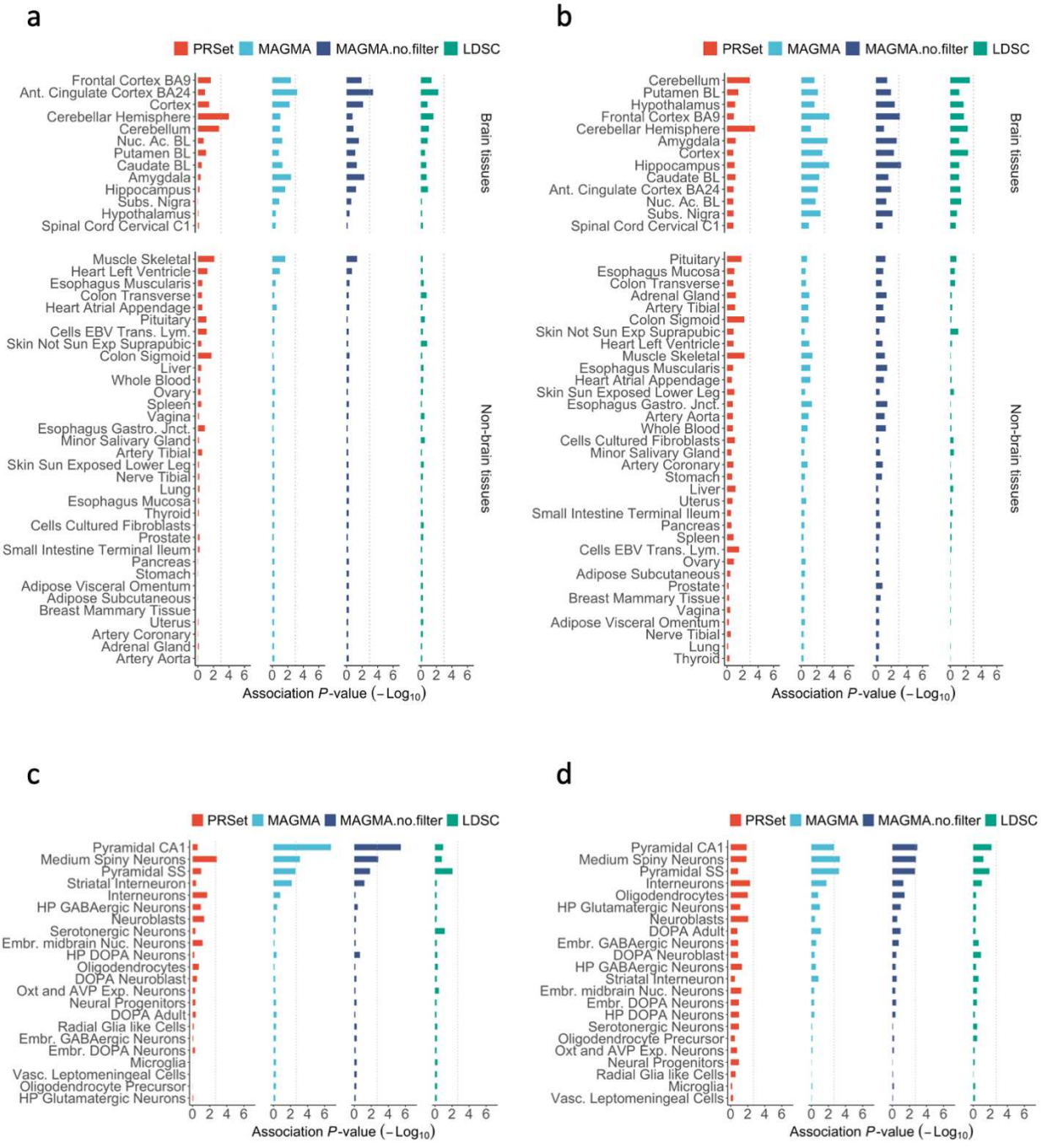
67. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).

68. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 1–7 (2018).

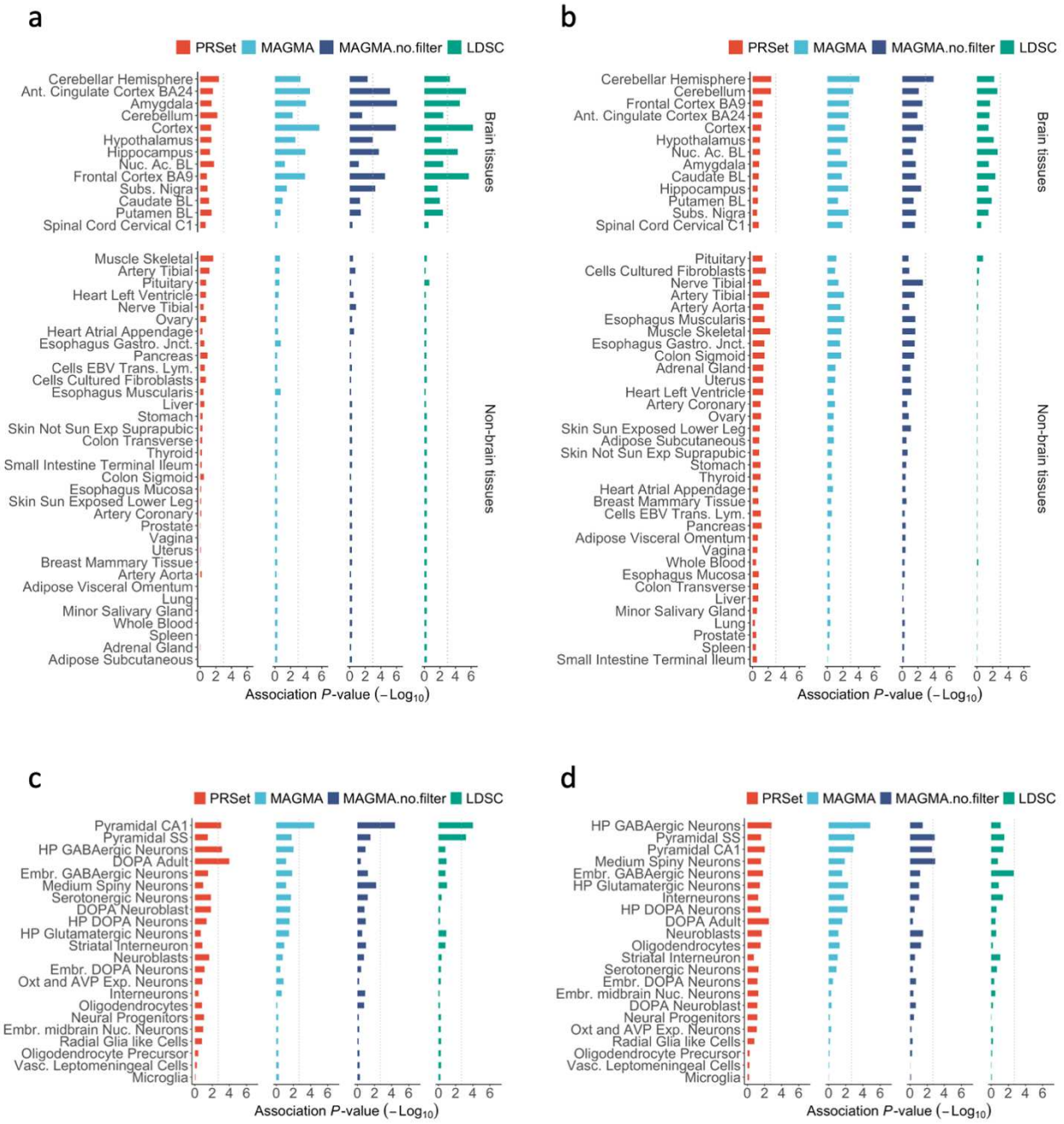
Extended Data Figures



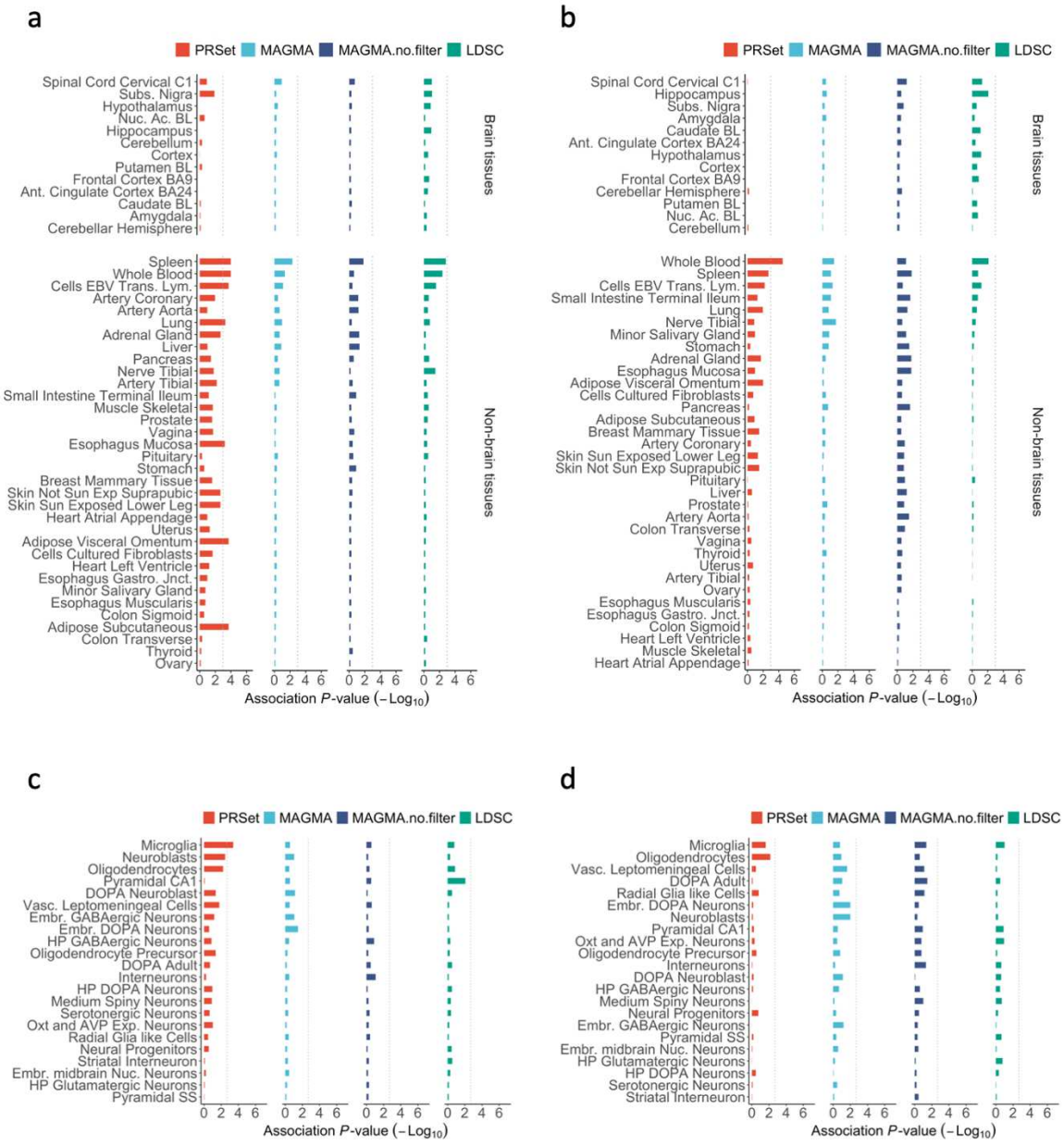
Extended Data Figure 1. Illustration of the test models used to assess cell type and tissue specificity. Left panel: illustrates the “top quantile” test model, which assumes that GWAS signal enrichment is concentrated in the most specifically expressed genes. Right panel: illustrates the “linear” test model, which assumes that enrichment of GWAS signal increases linearly with expression specificity.



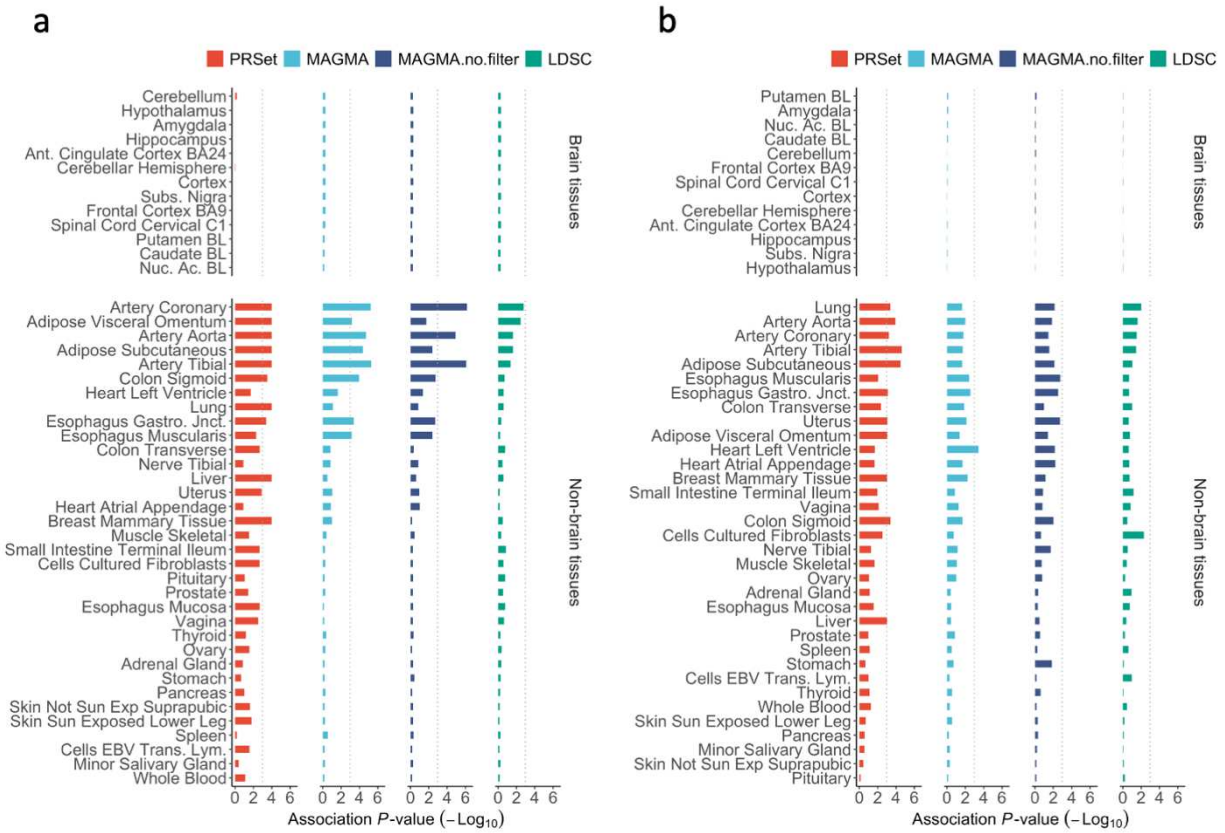
Extended Data Figure 2. Association between pathway enrichment *P*-value and expert opinion of tissue relevance (panels a and b) and cell type relevance (panels c and d) for each software and schizophrenia. Results for 'MAGMA' (light blue) were obtained after removing ambiguous and non-overlapping SNPs in the analysis to compare software performance given identical input data. Results for 'MAGMA.no.filter' (dark blue) show results including all available SNPs.



Extended Data Figure 3. Association between pathway enrichment *P*-value and expert opinion of tissue relevance (panels a and b) and cell type relevance (panels c and d) for each software and BMI. Results for 'MAGMA' (light blue) were obtained after removing ambiguous and non-overlapping SNPs in the analysis to compare software performance given identical input data. Results for 'MAGMA.no.filter' (dark blue) show results including all available SNPs.



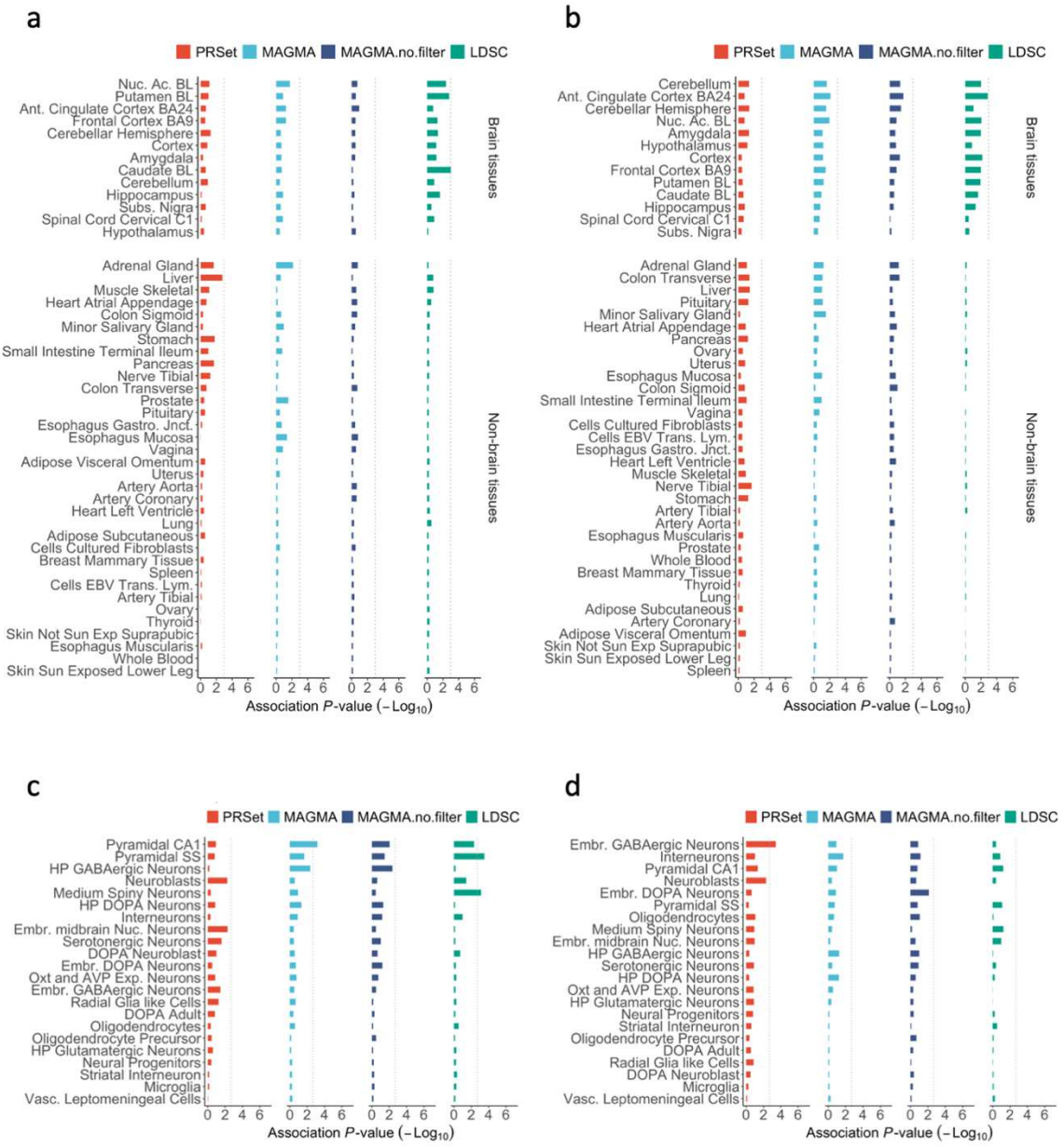
Extended Data Figure 4. Association between pathway enrichment *P*-value and expert opinion of tissue relevance (panels a and b) and cell type relevance (panels c and d) for each software and Alzheimer's disease. Results for 'MAGMA' (light blue) were obtained after removing ambiguous and non-overlapping SNPs in the analysis to compare software performance given identical input data. Results for 'MAGMA.no.filter' (dark blue) show results including all available SNPs.



766

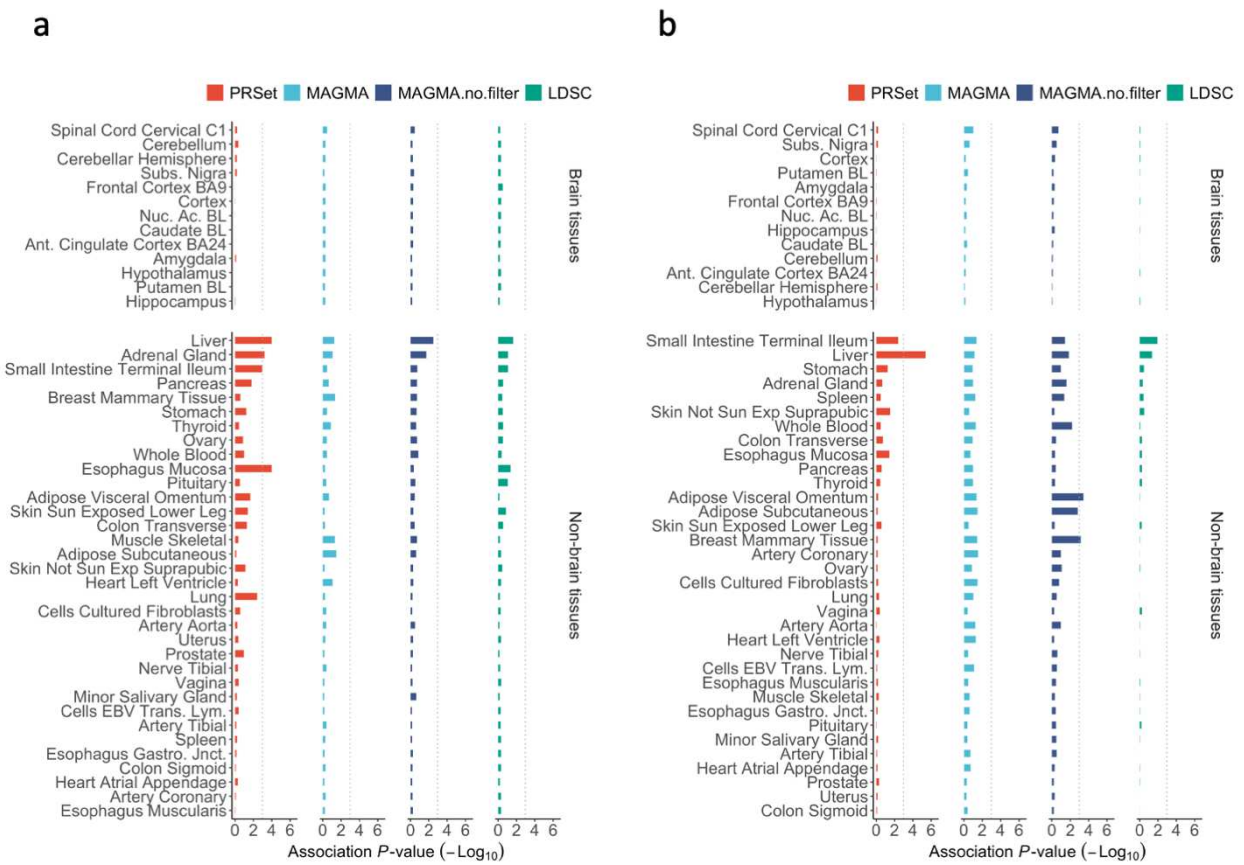
767 **Extended Data Figure 5.** Association between pathway enrichment *P*-value and expert opinion
768 of tissue relevance (panels a and b) and cell type relevance (panels c and d) for each software
769 and Coronary Artery Disease. Results for 'MAGMA' (light blue) were obtained after removing
770 ambiguous and non-overlapping SNPs in the analysis to compare software performance given
771 identical input data. Results for 'MAGMA.no.filter' (dark blue) show results including all available
772 SNPs.

773



774

775 **Extended Data Figure 6.** Association between pathway enrichment *P*-value and expert opinion
776 of tissue relevance (panels a and b) and cell type relevance (panels c and d) for each software
777 and alcohol consumption. Results for 'MAGMA' (light blue) were obtained after removing
778 ambiguous and non-overlapping SNPs in the analysis to compare software performance given
779 identical input data. Results for 'MAGMA.no.filter' (dark blue) show results including all available
780 SNPs.



781

782 **Extended Data Figure 7.** Association between pathway enrichment *P*-value and expert opinion
783 of tissue relevance (panels a and b) and cell type relevance (panels c and d) for each software
784 and low-density lipoproteins. Results for 'MAGMA' (light blue) were obtained after removing
785 ambiguous and non-overlapping SNPs in the analysis to compare software performance given
786 identical input data. Results for 'MAGMA.no.filter' (dark blue) show results including all available
787 SNPs.

788

Supplementary information for:

The power of pathway-based polygenic risk scores

Shing Wan Choi[1*], Judit García-González[1*], Yunfeng Ruan[2], Hei Man Wu[1], Jessica Johnson[1], Clive J Hoggart[1], Paul F. O'Reilly[1]

[1] Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA

[2] The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

* These authors contributed equally to this work

Supplementary Information Contents:

Supplementary Methods

Ascertainment of phenotypes in the UK Biobank cohort

Body Mass Index (BMI): BMI information was extracted from Field ID 21001. For the analysis of pathway enrichment and comparison with *MalaCards* disease relevance scores, residuals were calculated for men and women separately using linear regression, and were adjusted for age, recruitment centre, genotyping batch and 15 first principal components. Inverse normal transformation was carried out on the residuals[63].

Alcohol Consumption: Alcohol consumption scores were extracted using the consumption component of the Alcohol Use Disorder Identification Test (AUDIT-C), which was included in the mental health questionnaire[64]. AUDIT-C was based on three questions: frequency of drinking alcohol (Field ID 20414), amount of alcohol drunk on a typical drinking day (Field ID 20403), and frequency of consuming six or more units of alcohol (Field ID 20416). We used linear regression to obtain residuals of the AUDIT-C score adjusted for age, sex, recruitment centre, genotyping batch and 15 principal components [64].

Low-Density Lipoprotein (LDL): Levels of LDL across participants (Field ID 30780) were adjusted for stating use, following the procedure from Sinnott-Armstrong et al [65]. Medication use for 14 statins was extracted from Field ID 20003 (**Extended Data Table 2**). This information was then used to identify 1,382 individuals with LDL measurements that were not taking statins upon enrolment (years 2006-2010) but were taking statins at the time of first repeat assessment (years 2012-2013). For these individuals, we applied a statin correction factor on LDL measurements (*corrected LDL measurement = LDL upon enrolment / statin correction factor*). The statin correction factor was calculated as the mean of each individual's LDL measurement upon enrolment divided by LDL measurement at the time of first repeat assessment. Residuals of the LDL measurements were calculated by adjusting for age, sex, recruitment centre, genotyping

batch, fasting status (Field ID 74), dilution factor (Field ID 30897) and 15 PCs [66] using linear regression.

Coronary Artery Disease (CAD): CAD was defined as in Inouye et al (2018) [67]. CAD cases were ascertained as individuals who had suffered fatal or nonfatal myocardial infarction as indicated by their hospital records (Field IDs 41270, 41202, 41204, 41203, 41205 and 41271), death records (40001 and 40002) and medical history (Field IDs 6150 and 20002), and patients who have undergone percutaneous transluminal coronary angioplasty, or coronary artery bypass grafting (Field IDs 20004, 41200 and 41272). The age of event in cases was determined as the self-reported age and calculated age based on the earliest hospital record of the event or based on the death records; if more than one age were available, the smaller value was used. Age of the controls was determined as the latest self-reported age. For the analysis of pathway enrichment and comparison with MalaCards relevance scores, residuals of the CAD case control status were calculated by adjusting for age of event, sex, recruitment centre, genotyping batch and 15 principal components using logistic regression.

Alzheimer's Disease (AD): We generated a proxy phenotype for AD case control status based on family history of AD, similar to the procedure used in Marioni et al (2018) [68]. First, we removed participants diagnosed with AD as indicated by their hospital records (Field IDs 41202, 41204), death records (Field IDs 40001 and 40002) and primary care data, participants who were adopted (Field ID 1767), as well as participants whose parents were aged under 60 years (Field IDs 1845 and 2946), dead before age 60 years, or without age information (Field IDs 3526 and 1807). After merging with the genetic data, 41,164 participants had at least one parent who was affected by AD and 223,253 participants with parents non-affected by AD. Parental AD status was then defined as the number of parents affected by AD. Residuals of the parental AD status were calculated by adjusting for the sex of the participant, paternal age, maternal age, recruitment centre, genotyping batch and 15 principal components using linear regression.

Height: Standing height information was extracted from Field ID 50. Residuals were calculated by adjusting for sex, age, recruitment centre, genotyping batch and 15 principal components using linear regression.

Type 2 diabetes: Type 2 diabetes cases were ascertained as individuals diagnosed with non-insulin-dependent diabetes mellitus as indicated by their hospital records and primary care data (Field IDs 41202, 41204 and 41270) as well as death records (Field ID 40001 and 40002).

Inflammatory Bowel Disease: Cases of Inflammatory Bowel Disease were ascertained as individuals diagnosed with Crohn's disease or Ulcerative colitis as indicated by the self-reported questionnaire (Field ID 20002), their hospital records and primary care data (Field IDs 41202, 41204 and 41270) as well as death records (Field ID 40001 and 40002). Cases diagnosed with both Crohn's disease and Ulcerative Colitis were excluded. Residuals were calculated by adjusting for sex, recruitment centre, genotyping batch and 15 principal components using logistic regression.

PRSet implementation

Gene set based clumping

Genome-wide clumping can be suboptimal for gene set PRS calculation, as SNPs inside the gene set may be clumped out by SNPs outside the gene set. Therefore, clumping should be performed within each gene set to maximize signal retention. However, independently performing clumping on each gene set can be computationally intensive. PRSet optimizes the per-gene set clumping by utilizing a bit-flag system, where gene set membership is represented as a bit-flag (**Supplementary Fig. 1**). If the SNP is a member of the k^{th} gene set, then the k^{th} bit of the bit-flag will be set. During the clumping procedure, the index SNP will "remove" gene set membership from the clumped SNP if and only if they fall within the same gene set. At the end, SNPs without

any gene set memberships are removed from the analysis. This allows PRSet to perform the gene set based clumping without repeating the entire clumping procedure.

	Set A	Set B	Set C		Set A	Set B	Set C
Index SNP	1	0	1		1	0	1
Clumped SNP 1	1	1	0	⟶	0	1	0
Clumped SNP 2	1	0	1		0	0	0

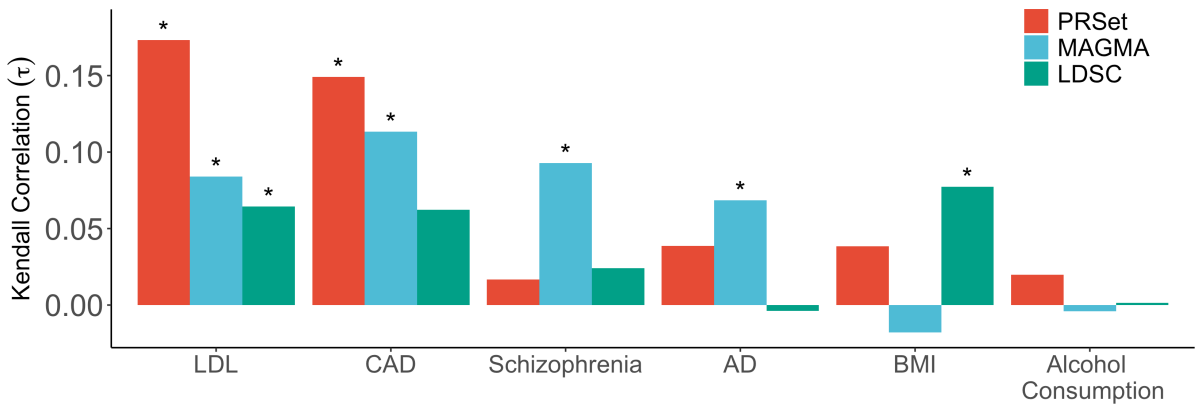
Supplementary Fig 1. Illustration of bit operation involved in PRSet clumping. The index SNP will "remove" gene set memberships from the clumped SNPs if and only if they fall within the same gene set. Clumped SNP without any gene set membership will be removed at the end of clumping. Here, clumped SNP 2 will be removed.

Permutation optimisation

PRSet calculates the competitive *P*-value via a permutation procedure, which is computationally expensive. To speed-up the permutation process, PRSet employes multiple techniques to optimize the computation procedure. First, the genotype of all "background" SNPs were loaded into the memory (using the --ultra parameter), such that the cost of repeated I/O during permutation is reduced. Then, gene sets with the same number of independent SNPs are grouped together and share the same null. In addition, during set-based permutation, an inverse regression is performed with *X* and *Y* switched. Inverted *X* and *Y* in the permutation results in the same z-score, but this allows pre-decomposition of the matrix and reuse for all permutations, drastically speeding up the performance.

Supplementary Notes

Supplementary Note 1. Kendall correlations between pathway enrichment and *MalaCards* scores



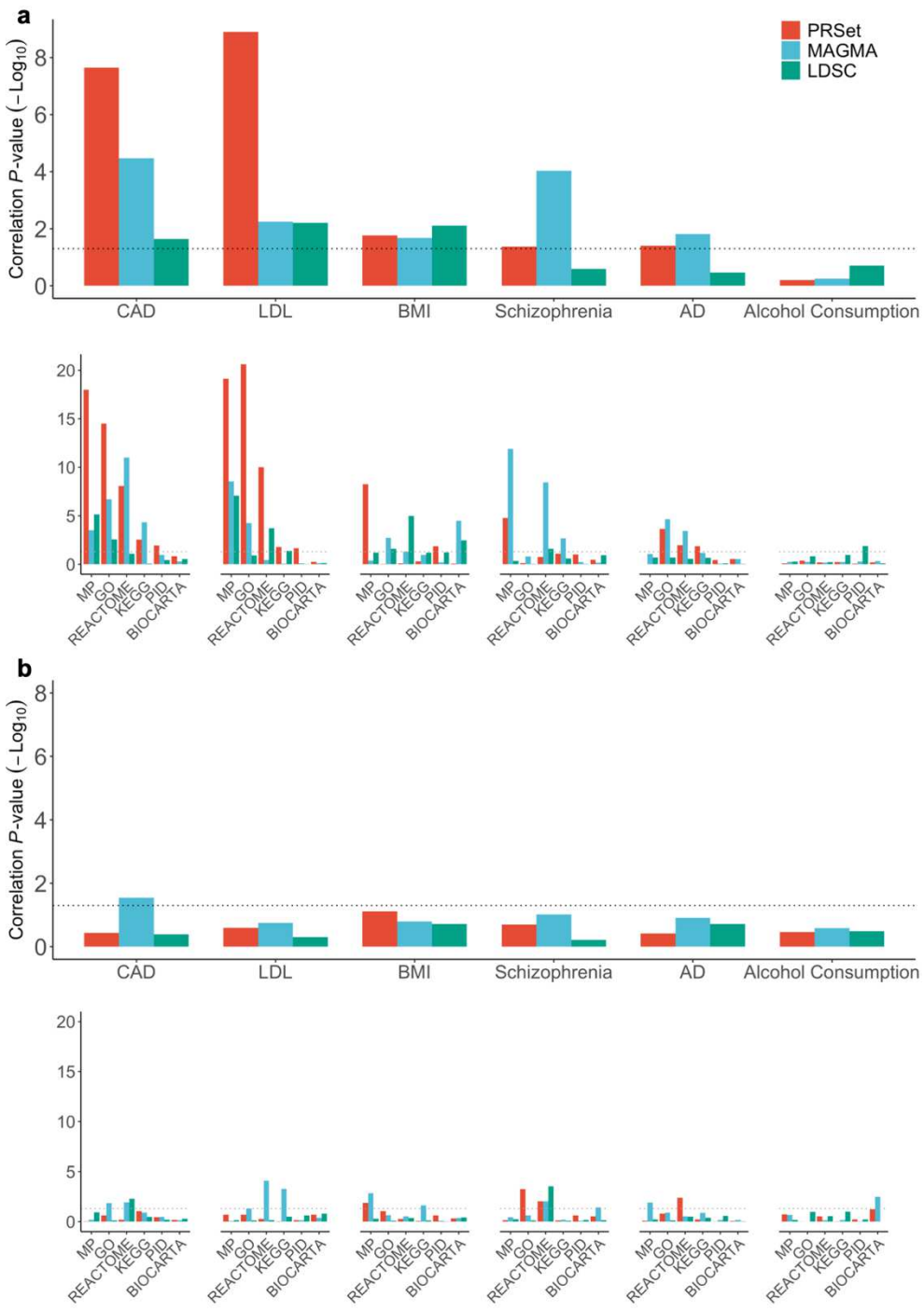
Supp Fig 2. Kendall correlation coefficients (τ) between pathway enrichment analyses and MalaCards relevance scores. Bar plots illustrate joint results of the six databases used to define pathways. *empirical *P*-value < 0.05.

Sensitivity analysis excluding genes in *MalaCards* database

One concern with using *MalaCards* as a proxy to empirical evidence is that the *MalaCards* scores may have inflated scores for genes with large effect sizes or genes that are well captured by GWAS because such genes may have particularly high experimental follow-up for this reason. To test whether potential biases in *MalaCards* scores could have a different impact across the enrichment algorithms, we ran a sensitivity analysis in which we removed all genes that have a *MalaCards* score above 0 from the pathway definitions and repeated the pathway enrichment analyses.

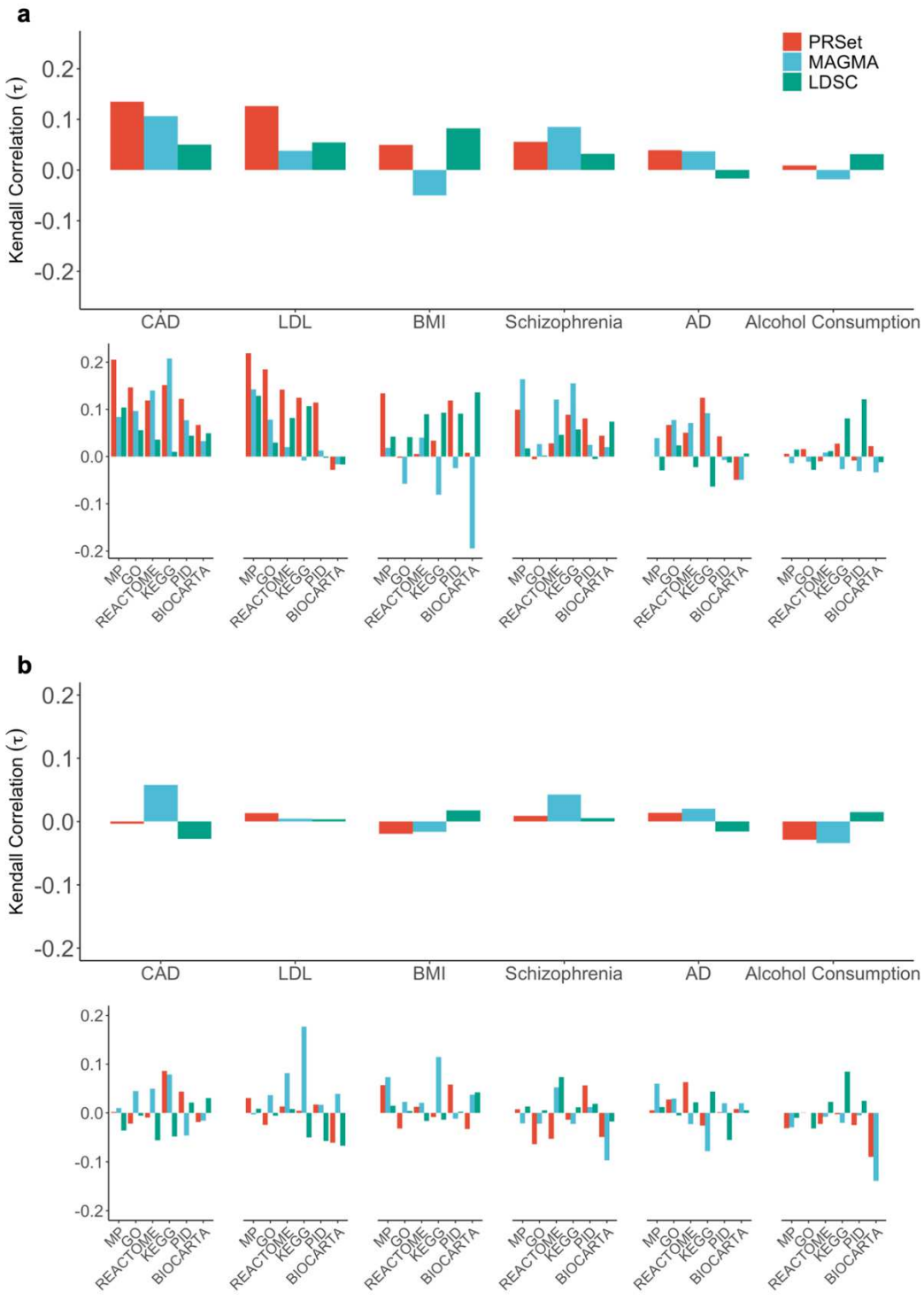
For the three methods, the correlation between pathway enrichment and the *MalaCards* relevance scores decreased substantially (**Supplementary Fig. 3 and 4**), confirming that *MalaCards*

925 pathway relevance scores include part of the disease signal captured by the three pathway
926 enrichment tools. The effect was stronger for PRSet and LDSC than for MAGMA. For PRSet and
927 LDSC correlations were no longer significant (**Supplementary Fig. 3b**), and correlation
928 coefficients across the six diseases decreased from $\tau = 0.077$ to $\tau = -0.0015$ for PRSet, and from
929 $\tau = 0.043$ to $\tau = 0.0076$ for LDSC (**Supplementary Fig. 4**); whereas for MAGMA, results were still
930 correlated with the original pathway relevance scores, albeit with reduced correlation and
931 significance ($\tau = 0.029$; *P*-value = 7.35×10^{-11}).



932

933 **Supplementary Fig 3**. Kendall correlation *P*-values between pathway enrichment analyses and
934 *MalaCards* relevance scores. **a,** Enrichment analyses use pathways that contain genes with
935 *MalaCards* scores **b,** Enrichment analyses use pathways where genes that have *MalaCards*
936 scores have been removed.



937

938 **Supplementary Fig 4**. Kendall correlation coefficients (τ) between pathway enrichment analyses

939 and *MalaCards* relevance scores. **a,** Enrichment analyses use pathways that contain genes with

MalaCards scores. **b,** Enrichment analyses use pathways where genes that have *MalaCards* scores are removed.

Supplementary Note 2. Performance of PRSet with a 5Mb shift in gene boundaries.

Rationale: Genome-wide PRS methods use the most associated SNPs (smallest *P*-value) to calculate PRSs. For heterogeneous diseases, the most associated SNPs are likely to contain risk alleles that are common across disease subtypes, as those will have more power to be detected. Whereas prioritizing SNPs that are common across subtypes can be useful for the prediction of the disease risk, those SNPs may not be useful for disease subtype classification.

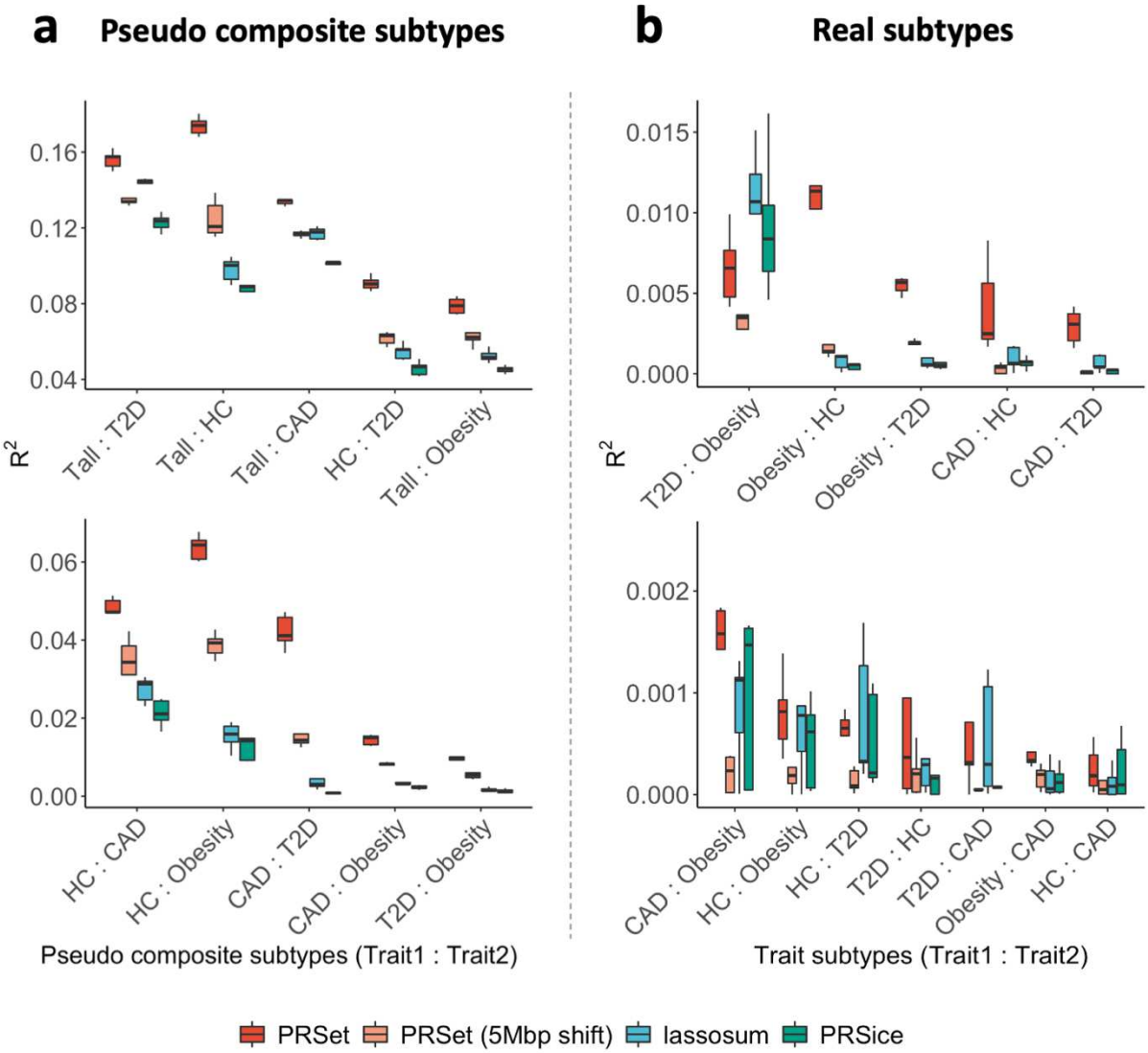
PRSet splits the genome into *k* pathways or 'chunks' and calculates PRSs for each pathway after clumping and thresholding. It is possible that splitting the genome into chunks and optimising parameters for each chunk improves classification: under this scenario, the most associated SNPs that are common across subtypes are less likely to mask other SNPs that -although less strongly associated- may be better subtype classifiers.

One important consideration is that the improvement in classification after splitting the genome into chunks may be true regardless of a chunk being part of a gene/biological pathway. To test whether the pathway structure is sufficient to explain the improvement in classification performance, or whether a *biologically informed* pathway structure is needed, we assessed the ability of PRSet to classify disease subtypes when we have reduced biological relevance of the pathway regions while retaining the pathway structures.

Methods: To keep the same pathway structure and overlap as the original analyses, we modified the gtf by shifting the gene boundaries 5Mbp, and we re-run the clustering analyses as previously described (Methods). Shifting was performed using the R packages 'genomation' (v.1.14.0) and

963 'GenomicRanges' (v.1.34.0). It is important to note that ~5% genes fall outside of chromosomal
964 boundary after the shift and they will be excluded from the analysis.

965 *Results*: After shifting gene boundaries 5Mb, 77.6% of the SNPs were out of genic regions. For
966 all the composite phenotypes, performance of PRSet with the 5Mb shift decreased substantially
967 as expected (**Supplementary Fig. 5**), suggesting that the biological information contained in the
968 pathway structure is useful for subtype classification.



969

Supplementary Fig 5. Classification performance of PRSet with a 5Mb shift in the gene coordinate boundaries versus standard version of PRSet and genome-wide PRS methods (lassosum and PRSice). **a,** pseudo composite subtype classification results. **b,** real subtype classification results.

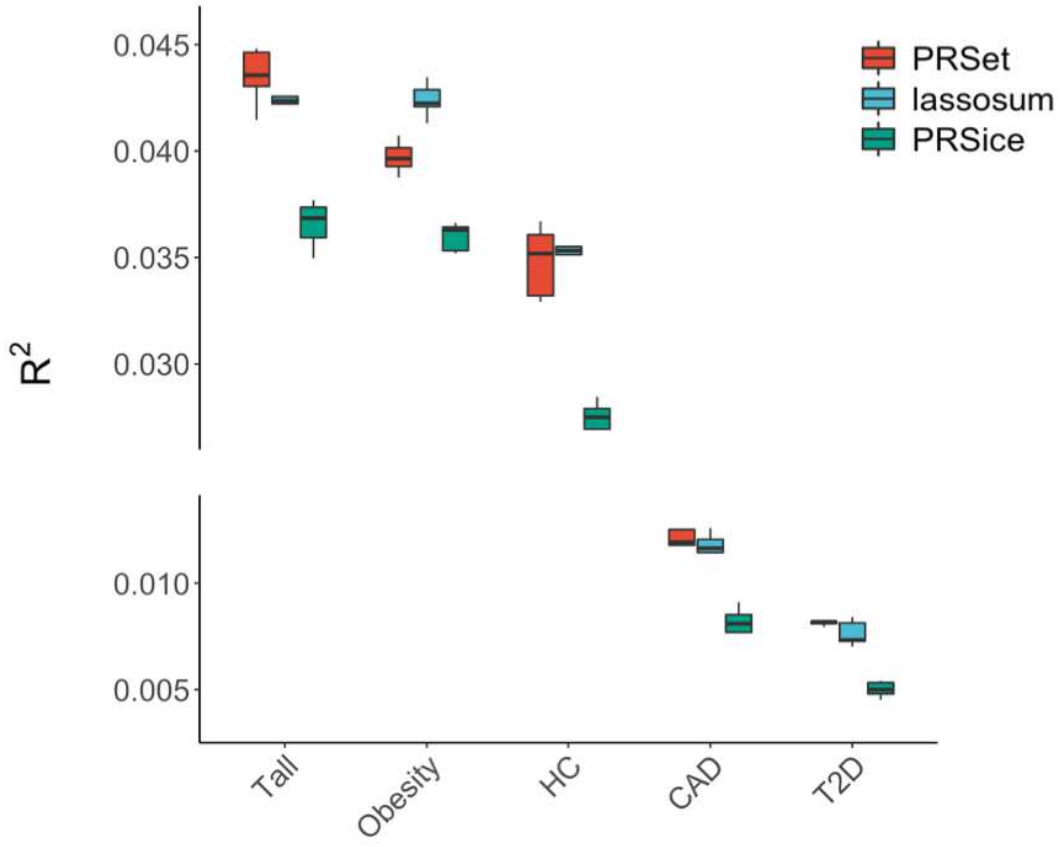
Supplementary Note 3. Performance of PRSet vs genome-wide PRS methods for single traits

Rationale: Following results from Supplementary Note 2, we hypothesise that PRSet outperforms genome-wide PRS methods at classifying disorders due to the use of pathways with biological information. PRSet calculates pathway specific PRSs, which allows SNPs with effect size in only one subtype to remain in the PRS, whereas PRSice and lassosum are more likely to optimize SNP signal that affects both subtypes -and therefore may not be so useful as classifiers.

If this hypothesis is true, PRSet would outperform genome-wide PRS tools for classification tasks, but its performance will be more limited when PRSs are used for the prediction of single, more homogeneous traits. In those cases, we expect genome-wide PRS to do a better job at trait prediction as they have more power to optimize genetic signal associated with disease. To test whether PRSet results are due to the use of pathways or due to methodological aspects that could have contributed to the improvement in performance, we assessed pathway and whole-genome PRSs for single traits (as opposed to than the composite phenotypes that were generated in the main analyses).

Methods: Whole-genome and pathway specific PRS were calculated for the same five traits/diseases that were used for the sub-phenotype classification in real data: type 2 diabetes, CAD, obesity (defined as BMI > 30), LDL and extreme height. We calculated PRS for these traits using publicly available GWAS data for individuals from the UK Biobank cohort (**Methods and Extended Data Tables 5 and 6**).

Results: For single trait analyses, the improvement in performance for PRSet vs the whole-genome methods were reduced, and in some cases (BMI and LDL) lassosum performed better than PRSet. For all the traits assessed, phenotypic variance explained by PRSice performance was the lowest (**Supplementary Fig. 6**).



Supplementary Fig 6. Performance of PRSet vs genome-wide PRS methods for prediction of single traits.

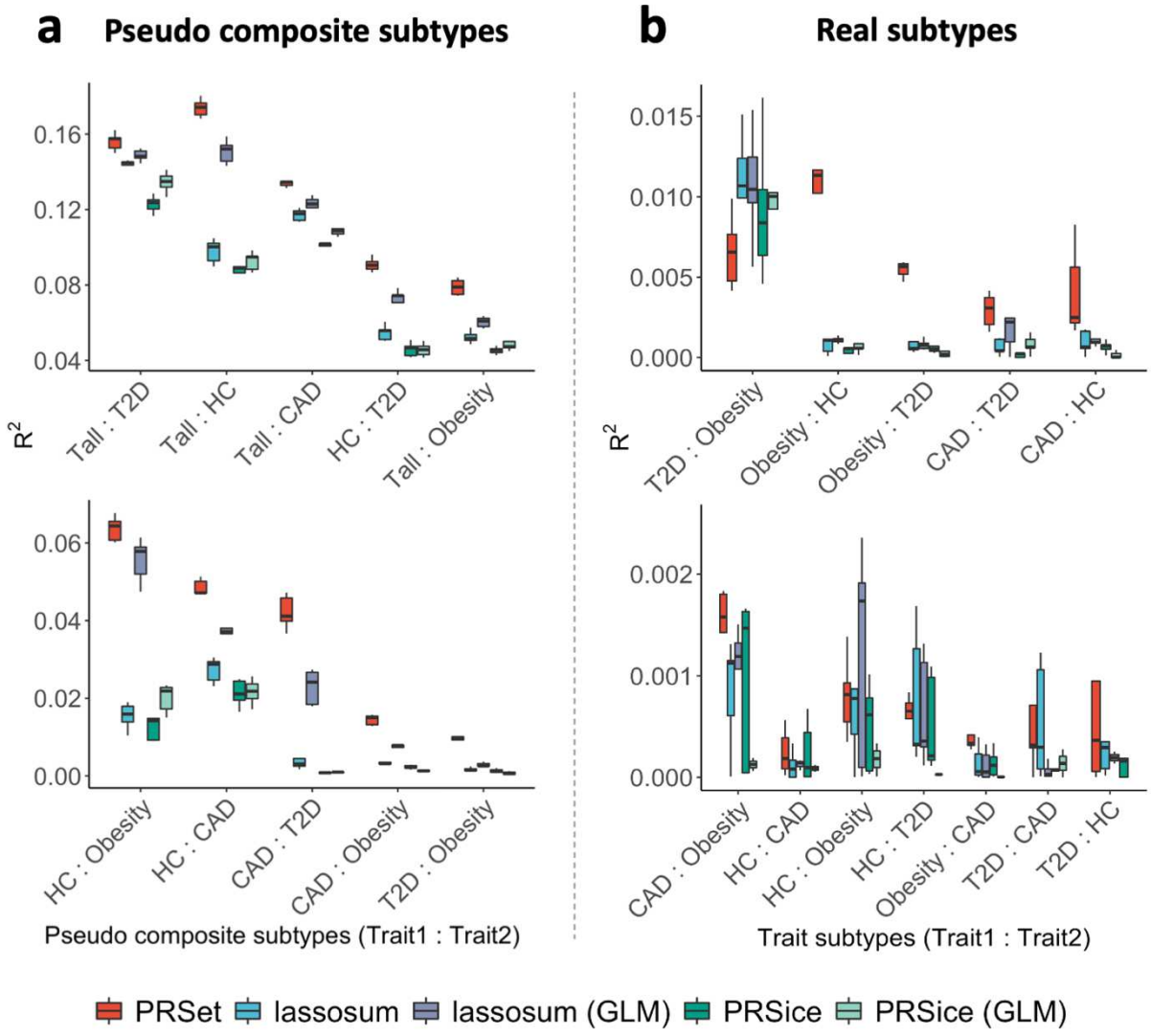
Supplementary Note 4. Performance of PRSet vs genome-wide PRS methods with lasso regression

Rationale: The inclusion of multiple pathway specific PRSs into a lasso regression may contribute to the improvement on PRSet classification performance. To test whether the regularization method also improves classification in genome wide PRSs, we calculated genome wide PRSs with lassosum and PRSice, and included the PRSs into a lasso regression.

Methods: Genome-wide PRS analyses were performed using lassosum and PRSice-2. For PRSice-2, ~500 PRSs were calculated at different *P*-value thresholds using high resolution scoring (step size of the threshold=0.001). Whereas in standard PRS calculations PRS at each *P*-value threshold includes all the SNPs with *P*-value below the threshold (i.e. each PRS is a subset of each other), in this analysis we calculated PRSs non-cumulatively, so that each PRS includes SNPs below the current *P*-value threshold, but it does not include SNPs at *P*-values higher than the previous threshold. Using non-cumulative PRSs followed by lasso regression is a better benchmarking strategy than using standard PRSs, because for PRSet the overlap of SNPs across pathways is only partial. For lassosum, PRSs were calculated at different values of penalty factor λ and soft-thresholding parameter *s*. PRSs for each software were then included in a generalized linear model with lasso regularization using the `*cv.glmnet*` function from the glmnet package (v4.0-2).

Results: For the pseudo composite subtype classification analyses, the application of lasso regularization improved the performance of both lassosum and PRSice (**Supplementary Fig. 7a**). Improvement in performance was stronger for lassosum (GLM) than for PRSice, but none of the two methods outperformed PRSet, giving supportive evidence that the use of pathways improves classification.

For the real subtype classification analyses, there was not a clear improvement in performance when GLM and regularization was applied (**Supplementary Fig. 7b**). In some cases, the best model for PRSice with GLM and regularization was the intercept, meaning that the R2 could not be obtained. For T2D with presence/absence of hypercholesterolemia (T2D : HC), intercept was the best model for the five-fold cross validations, therefore R2 estimates are omitted in the figure.



Supplementary Fig 7. Performance of PRSet vs genome-wide PRS methods with a generalized linear model and lasso regularization. **a,** pseudo composite subtype classification results. **b,** real subtype classification results. For the lassosum (GLM) and PRSice (GLM) models, whole-genome

1035 PRSs calculated with different parameters were subsequently used in a generalized linear model
1036 with lasso regularization. HC; hypercholesterolemia.

Supplementary References:

1. 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).

2. Sanchez-Roige, S. *et al.* Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *AJP* **176**, 107–118 (2018).

3. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics* **53**, 185–194 (2021).

4. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274–1283 (2013).

5. Inouye, M. *et al.* Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *Journal of the American College of Cardiology* **72**, 1883–1893 (2018).

6. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl Psychiatry* **8**, 1–7 (2018).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedDataTables.xlsx