The SumRank method efficiently detects valid pleiotropic variants using GWAS summary level data

Sander Lamballais[1]*; Gennady V. Roshchupkin[2,3]*; Raymond A. Poot[4]; Steven A. Kushner[5]; M. Arfan Ikram[3]; Hieab H. H. Adams[1,3,6†]; Henning Tiemeier[7,8†].

[1] Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, the Netherlands

[2] Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, the Netherlands

[3] Department of Radiology and Nuclear Medicine, Erasmus MC University Medical Center, Rotterdam, the Netherlands

[4] Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, the Netherlands

[5] Department of Psychiatry, Erasmus MC University Medical Center, Rotterdam, the Netherlands

[6] Latin American Brain Health (BrainLat), Universidad Adolfo Ibáñez, Santiago, Chile

[7] Department of Social and Behavioral Science, Harvard TH Chan School of Public Health, Boston, MA, USA

[8] Department of Child and Adolescent Psychiatry, Erasmus MC-Sophia Children's Hospital, Rotterdam, the Netherlands

* Equal contribution

† Equal contribution

CORRESPONDENCE

Hieab H. H. Adams, MD, PhD. Departments of Clinical Genetics and of Radiology and Nuclear Medicine. Erasmus MC University Medical Center, Wytemaweg 80, 3015 CE, Rotterdam, the Netherlands. Telephone number: +31 10 70 33559. Fax Number: +31 10 70 43489. E-mail address: h.adams@erasmusmc.nl.

ABSTRACT

Genome-wide association studies (GWASs) have illustrated that pleiotropy is the norm rather than the exception in the genetic architecture of complex traits. Previous methods have been proposed to investigate cross-trait associations; however, these typically suffer from substantially inflated false positive rates, are limited to associations between only two traits, or examine only a subset of genome-wide variants. Here, we introduce SumRank, a novel method for discovering cross-trait associations based on p-values from GWAS summary statistics. Using numerical and genotype simulations, we demonstrated that SumRank had a near-zero false positive rate, in contrast to contemporary meta-analytical approaches that yielded severely inflated rates. We applied SumRank to GWAS summary data of eight psychiatric disorders and identified 515 independent pleiotropic variants across 175 loci, of which 92 were novel and 71 had previously been associated with only one disorder. Overall, SumRank provides a valid and efficient method for genome-wide discovery of pleiotropic variants that underlie a scalable number of traits.

INTRODUCTION

Horizontal pleiotropy occurs when genetic variants are associated with multiple traits. Pleiotropy has received increasing attention because it explains how a limited number of genes map to numerous complex traits, allowing researchers to uncover novel genetic mechanisms and therapeutic targets of disease. Recent studies have identified pleiotropic variants in loci covering approximately 57% of the genome[1], thereby emphasizing the need for robust frameworks and tools that further understanding. However, the relevance of pleiotropy to human health and disease is still poorly understood.

The estimation of pleiotropic effects requires information on multiple traits and their genetic architectures. This problem can be addressed through genome-wide association studies (GWASs), which have been instrumental in uncovering the genetic architecture of numerous traits[2]. GWAS has been particularly insightful for traits that are highly polygenic. Multivariate extensions of GWAS, such as vbBIMBAM[3] and TATES[4], can examine traits jointly within the same population. However, these methods require individual-level data with information on all relevant traits. In addition, sharing individual-level data is subject to privacy and informed consent constraints, which limits accessibility to individual-level data by third parties. An alternative approach is to study horizontal pleiotropy using GWAS summary statistics. Over the past decade, summary-level results of thousands of GWASs have become widely available. Summary-level methods have been proposed to estimate the probability that a genetic variant is pleiotropic by performing calculations on the effect sizes or *p*-values across different GWASs. Examples include Fisher's combination test (FCT), Cross-Phenotype Meta-Analysis (CPMA)[5], ASsociation analysis based on subSETs (ASSET)[6], and conjunctional FDR (conjFDR)[7].

However, summary-level methods have several limitations. First, horizontal pleiotropy is defined as a genetic variant that has a non-zero effect on two or more traits under study. However, summary-level methods based on meta-analytic approaches can falsely detect variants as pleiotropic when they are associated only with a single trait, particularly when the association is strong. This leads to a high false positive rate and thus an overestimation of the number of pleiotropic variants. Second, GWASs provide results at the variant level for the whole genome; however, summary-level methods may not have the same level of detail. For example, the regional Bayes factor explores pleiotropy per independent genomic region rather than per variant[8], even though traits may be associated with the same region, but not with the

same causal variant. Third, genetic variants may be pleiotropic across any number of traits; however, methods are typically designed to examine pleiotropy across two traits at a time. Consequently, they can only detect pleiotropy at the pairwise level.

Here, we introduce SumRank, a method utilizing GWAS summary statistics that identifies variant-level genome-wide pleiotropy without limitation for the number of traits being examined simultaneously. Through numerical and genotype simulations, we showed that SumRank detects pleiotropic variants with higher specificity than other methods. We also showed that these findings hold under widely varying conditions. Finally, we demonstrated the discovery power of SumRank by exploring pleiotropy across eight different psychiatric disorders.

RESULTS

Overview of the method

SumRank is a combinatorial method for assessing genetic pleiotropy based on p-values of GWAS summary statistics. The proposed method has several key features. First, it explicitly tests the composite hypothesis that a variant is associated with all traits under study. Second, the method transforms the *p*-values to ranks, thus mitigating the effect of extremely small *p*-values in only one phenotype driving the detected effect. Third, SumRank can be applied to an unlimited number of traits. Finally, applying SumRank across the genome requires very modest computational power and memory and, more importantly, can easily be implemented in any programming or scripting language that allows basic arithmetic operations. Below, we address all these advantages.

The association between genetic variant g and trait Yj can be modelled using the following regression equation:

$$\mathbf{Y}_{\mathrm{j}} = \beta_0 + \beta_{1,\mathrm{j}}\mathrm{g} + \gamma\mathrm{w} + \varepsilon \qquad (1)$$

where γw represents a non-genetic component and ε is an error term. A pleiotropic variant is defined as one in which $\beta_{1,\mathrm{j}} \neq 0$ for all *j* in 1 … *m* traits under consideration.

Given *n* variants and *m* traits, ***P*** is an *n x m* matrix that contains the *p*-values from the corresponding GWASs. SumRank uses a combinatorial approach to assess the pleiotropy of variants based on ***P***. First, the *p*-values within each GWAS are sorted and ranked from smallest to largest. If a variant is associated with a trait, its rank should be low. Second, the ranks for

each variant across the traits under study are summed, resulting in $\boldsymbol{R}$, a vector of length n. Under the null hypothesis, low summed ranks are uncommon. Because pleiotropic variants are more likely to yield small p-values in a GWAS, their summed ranks should be low. Therefore, the third step is to determine the probability of $\boldsymbol{R_i}$ or smaller for a given variant i (**Methods**):

$$p_{sum,i} = \frac{m}{m!n^m} \sum_{k=1}^{R_i} \binom{k-1}{m-1} \quad (2)$$

Notably, psum can be approximated by summing the p-values directly within a given variant (**Methods section**):

$$p_{sum,i} = \frac{(\sum_{j=1}^{m} P_{i,j})^m}{m!} \quad (3)$$

The implications of SumRank are best understood when comparing the rejection regions with FCT and CPMA (Fig. 1a). When a variant has a small p-value in both GWASs (bottom left, Fig. 1a), the three methods yield correspondingly small p-values for pleiotropy. However, when the variant is associated with only one of the two traits, FCT and CPMA can still return a small p-value for pleiotropy regardless of the value of the other p-value. This will likely represent genetic loci that are not truly pleiotropic, that is, false positive findings. Given that SumRank relies on ranks rather than absolute p-values, it is robust to the influence of extremely small p-values. Both p-values must be small for SumRank to return a small p-value for pleiotropy.

We further tested the implications of the rejection regions by simulating pairs of p-values. First, we created 1 million pairs of p-values drawn from a uniform distribution, that is, under null conditions. In this scenario, none of the methods inflated the false positive rates for the two traits (Supplementary Fig. 1). Second, we created 1 million pairs where the first p-value was always set to $5 \cdot 10^{-8}$ (i.e., genome-wide significance), but the second p-value was again drawn from a uniform distribution. This represents a scenario where a variant is only associated with the first trait, and therefore should not be classified as pleiotropic. We tested the methods against a pleiotropy p-value threshold of $5 \cdot 10^{-8}$. Under these conditions, FCT and CPMA had false positive rates of 4.81% and 0.80%, respectively, compared to 0.03% for SumRank (Fig. 1b). Crucially, the false positive rates for FCT and CPMA reached 100% when the first p-value was set to $2.34 \cdot 10^{-9}$ and $4.07 \cdot 10^{-10}$, respectively (Fig. 1c). In contrast, the false positive rate of SumRank remained 0.03% even when the first p-value was set to $5 \cdot 10^{-100}$.

Performance in genotype simulations of two traits

We conducted a series of population-based genotype simulations with varying genetic architectures to evaluate the ability of SumRank and similar methods to detect pleiotropic variants. To this end, we simulated correlated traits with a known genetic architecture using data from the UK Biobank. From the UK Biobank genetic data, we selected independent variants with a minor allele frequency above 5% using PLINK 1.9. We then randomly assigned non-pleiotropic and pleiotropic causal effects to a subset of the independent variants, to effectively create hypothetical traits. For the primary simulations, the causal effects were drawn from a multivariate normal distribution. The heritability of the traits and the degree of genetic correlation between traits were set manually. We obtained summary statistics for the simulated traits through genome-wide analysis, which were used as input for SumRank and other methods. The summary statistics resulting from these methods were clumped against genome-wide significance ($5 \cdot 10^{-8}$) and $R > 0.2$, and the resulting lead and correlated variants were assessed for the presence of pleiotropic and non-pleiotropic variants. As the primary test we quantified the true pleiotropy rate, that is, the proportion of the variants marked as pleiotropic where the true effect was indeed pleiotropic. An extensive description of the procedure is provided in the Methods section.

For each simulation, we generated two traits for two independent cohorts of 100,000 individual samples from the UK Biobank. As shown in Fig. 2a, each trait had 1000 non-pleiotropic causal variants that were unique to the traits (red) and 1000 pleiotropic causal variants that were shared between them (green). Across 100 simulations, SumRank detected 21.1% (SD = 6.0%) of the pleiotropic causal variants, which was lower than that of FCT (34.2%, SD = 13.1%) and CPMA (32.4%, SD = 12.2%). However, SumRank labeled only 0.6% (SD = 0.2%) of the non-pleiotropic causal variants as pleiotropic, whereas FCT and CPMA labeled 18.9% (SD = 1.0%) and 17.8% (SD = 1.1%), respectively. In other words, SumRank had a true pleiotropy rate of 94.6% (SD = 1.6%) among the identified hits, compared with 47.6% (SD = 2.6%) for both FCT and CPMA.

True pleiotropy rates depend on the underlying genetic architectures, that is, the number of causal variants per trait and the overlap between them. Therefore, we reran the simulations where each trait had 2000 causal variants in total but varied the overlap between the traits. When the traits had a 25% overlap (i.e., 500 pleiotropic variants in total), the performance of FCT and CPMA worsened, with the true pleiotropy rate dropping to 22.1% (SD = 1.1%) and

22.3% (SD = 1.4%), respectively (Fig 2b). In contrast, SumRank yielded a true pleiotropy rate of 86.8% (SD = 4.6%). When the traits shared 75% of their variants (i.e., 1500 pleiotropic variants in total), the relative performance of FCT and CPMA improved, with true pleiotropy rates of 72.0% (SD = 1.8%) and 72.2% (SD = 1.8%), respectively (Fig 2c). Although SumRank still reached a true pleiotropy rate of 98.3% (SD = 1.0%), its advantage over the other methods was attenuated as SumRank identified fewer variants overall. Thus, the lower the ratio of pleiotropic to non-pleiotropic causal variants, the better SumRank will perform relative to the other methods.

As SumRank implicates fewer variants overall, it is possible that the methods prioritize the same variants, but that SumRank is simply more conservative. To examine this post-hoc hypothesis, we sorted the p-values for each method from lowest to highest and iteratively calculated for the most significant variants what the true pleiotropy rate was, up to 1000 variants (Fig. 3). SumRank had a 100% (SD = 0%) true pleiotropy rate when considering the top 1, 10, and 100 variants. In contrast, FCT had a true pleiotropy rate of 79.5% (SD = 44.7%) when considering the top variant, 57.5% (SD = 14.8%) in the top 10, and 54.3% (SD = 6.4%) in the top 100. CPMA followed a similar pattern as FCT (top 1: 80.0% (44.7%); top 10: 58.0% (17.9 %); top 100: 54.8% (6.7%)). Thus, our post-hoc hypothesis can be rejected: the methods prioritize vastly different variants, with SumRank prioritizing more true pleiotropic variants than FCT and CPMA.

Next, we tested another post-hoc hypothesis: SumRank's apparent advantages over existing methods might depend largely on the parameters used for the simulation. Therefore, we retested the simulations by varying the values of different parameters. We first set the sample size to 50,000, 100,000, and 150,000 individuals (Supplementary Fig. 2). Larger sample sizes led to a higher number of pleiotropic variants detected across all methods. The true pleiotropy rate for SumRank decreased to 91.2% compared to 46.1% for FCT and 46.5% for CPMA. We then varied the heritability of the traits to 0.25, 0.50, and 0.75 (Supplementary Fig. 3), where a higher heritability led to more identified variants, and relatively poorer performance by FCT and CPMA. We also changed the correlation between the effect sizes of the pleiotropic causal variants to 0, 0.5, or 1 (Supplementary Fig. 4). Counterintuitively, as the correlation of the effect sizes increased, the true pleiotropy rate for FCT decreased from 48.0% to 44.4%, and for CPMA from 48.5% to 45.0%. In contrast, the true pleiotropy rate for SumRank increased from 94.3% to 95.8%. Thus, we must also reject this post-hoc hypothesis.

Performance in genotype simulations with more than two traits

Our second hypothesis posited that SumRank can be simultaneously applied to more than two traits. Thus, SumRank can explicitly test the composite hypothesis across any number of m traits. To validate its use for multiple phenotypes, we tested the performance of SumRank using different simulations. First, we considered the distribution of SumRank *p*-values when performed on *p*-values drawn from a uniform distribution for 3, 5, 7, 10, 20, and 50 phenotypes, where a pleiotropic variant is defined as affecting all phenotypes simultaneously (Supplementary Fig 5). When the p-values for pleiotropy were tested against a threshold of 0.05, FCT and CPMA showed a false positive rate of approximately 5% across all conditions. In contrast, the false positive rate of SumRank declined with the number of phenotypes, and it was effectively 0% for 50 phenotypes. This is a direct consequence of the hypothesis being tested; it is extremely unlikely that 50 randomly sampled p-values all suggest an association between the variant and each trait.

As SumRank works with more than two traits, we simulated four traits to test whether the true pleiotropy rate for SumRank was also better than that for FCT and CPMA. Note that pleiotropy among the four traits can indicate any variant affecting two, three, or four traits. We sought to explicitly test how well the methods detected variants that only affect all four traits simultaneously. Therefore, we created three scenarios with varying numbers of pleiotropic variants for each combination of traits. The first scenario is shown in Fig 4a, where we simulated 1000 unique variants per trait (dark red), and 1000 pleiotropic variants that affected all four traits. In this scenario, SumRank reached a true pleiotropy rate of 99.2% (SD = 0.4%), with FCT and CPMA reaching 73.2% (SD = 1.0%) and 73.7% (SD = 1.0%), respectively. The second scenario is shown in Fig 4b, where we simulated 1000 variants unique to each trait (dark red), and 1000 pleiotropic variants for every combination of two (red), three (light red), or four traits (green). Under such conditions, SumRank still primarily identified causal variants that were pleiotropic for all four traits (79.6%, SD = 2.25%), and flagged few variants that were related to three traits (17.5%), two traits (2.4%), or even one trait (0.5%). In contrast, FCT and CPMA reached a true pleiotropy rates of 40.1% (SD = 1.3%) and 40.9% (SD = 1.3%), respectively. Crucially, 9.0% and 8.4% of the variants that FCT and CPMA identified as related to all four traits were simulated to be non-pleiotropic. This pattern of findings was most evident in the scenario where most variants were not pleiotropic and only 100 variants were pleiotropic for all four traits (Fig 4c). Of the loci that the methods found to be related to all traits, SumRank

was correct for 56% of the loci (SD = 7.8%), compared to 10.9% (SD = 1.3%) and 11.1% (SD = 1.3%) for FCT and CPMA, respectively. We argue that these results support the notion that SumRank is more robust to false positives when studying more than two traits.

Example study: pleiotropy across eight psychiatric disorders

To test the utility of SumRank with non-simulated data, we mirrored the design of the largest pleiotropy study to date across eight prominent psychiatric disorders (Supplementary Table 1)[9]: anorexia nervosa, attention-deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depression, obsessive-compulsive disorder, schizophrenia, and Tourette syndrome. Every possible combination of at least two traits (= 247) was tested against a significance threshold of ($5 \cdot 10^{-8}$ / 247 =) $2.02 \cdot 10^{-10}$. All calculations took approximately two minutes on a single core.

Overall, we identified 515 independent variants across 175 pleiotropic loci (Fig 5a). Compared to the underlying GWASs, 92 of these loci had previously not been associated with any trait, 71 loci with just one of the traits, and 12 loci were genome-wide significant in two or three of the GWASs. The independent variants were primarily located in the introns of protein-coding genes (47.7%), intergenic sequences (32.3%), or introns of non-coding genes (13.3%) (Fig 5b). Most independent variants were associated with two traits (n = 198); however, surprisingly we also identified large groups of independent variants related to three traits (n = 144), four traits (n = 101), and five or more traits (n = 28) (Fig. 5c). The largest group was related to bipolar disorder and schizophrenia (n = 117), followed by variants related to bipolar disorder, schizophrenia, and major depressive disorder (n = 81) (Fig. 6).

DISCUSSION

SumRank overcomes several shortcomings of the existing pleiotropy methods. Most importantly, as it assesses the composite hypothesis of whether individual variants associate with all traits, SumRank is not affected by a strong association with only one or a subset of the traits – which is reflected in its superior rate of true pleiotropic findings compared to other methods. In addition, it is robust, can be applied to any number of variants across the genome, and is readily computed as it relies on a simple algebraic formula.

The findings illustrate that the study of pleiotropy requires a delicate balancing act. Pleiotropy can be studied by only considering genome-wide significant hits[1], but this approach fails to

identify loci that did not reach genome-wide significance in all GWASs involved. Summary-level approaches can reveal such loci, but as is clear from our simulations this can come at the cost of an inflated false positive rate and thus having a higher rate of irrelevant loci, particularly when considering more than two traits. SumRank is a conservative approach that may yield fewer pleiotropic loci than other methods, but the loci that it does return have a much higher likelihood of being truly pleiotropic. Thus, SumRank strikes a balance that is more suitable for functional genomic and mechanistic biological follow-up studies of pleiotropic loci.

Composite hypotheses have been used before in the study of cross-trait associations, for example in conjFDR[7], the regional Bayes factor[8], and PLACO[10]. However, these approaches typically focus on pairs of traits. Through SumRank, we showed that among psychiatric disorders there is likely a significant amount of pleiotropy that extends beyond two traits, with most of the identified loci relating to three or more traits. This is in line with previous studies that focused only on genome-wide significant variants[1]. As more GWASs are conducted and their sample sizes increase, the relevance of multi-trait loci will become more apparent. Methods like SumRank are needed to properly model and disentangle such loci.

We tested SumRank extensively under various simulation conditions and demonstrated that it is a robust and efficient method. However, this method also has several shortcomings and areas for improvement. SumRank relies purely on GWAS p-values and cannot incorporate any information such as effect sizes, genetic correlations[11,12], omics data, or other external information. Nevertheless, the identification of loci can be sufficient to uncover relevant genes and accompanying gene pathways and physiological mechanisms. Furthermore, although untested in the current study, SumRank may be able to incorporate functional data such as eQTL data if such data also have accompanying *p*-values. Another limitation of SumRank is that like most summary-level methods, it does not distinguish between the different types of pleiotropy, that is, vertical and horizontal pleiotropy. Although variants may be marked as pleiotropic by any of the GWAS-based methods, we emphasize that further work is needed to elucidate the underlying mechanisms.

In conclusion, SumRank efficiently and robustly assesses the pleiotropy of genetic variants based on p-values from GWAS. Compared to other summary-level methods, it provides stringent control for false positives by using a composite hypothesis and allows for an unlimited number of traits to be studied with high computational efficiency. We believe that SumRank

can play a role in identifying multi-trait pleiotropic loci with high confidence, which in turn will prove invaluable for further elucidating the relevance of pleiotropy in health and disease.

METHODS

Theory

The SumRank method is based on the idea that pleiotropy can be assessed from GWAS summary statistics by ranking the p-values from the smallest (first rank) to the largest (maximum rank), and subsequently summing the ranks per variant. Considering m traits with N variants, $\boldsymbol{P}$ represents an N x m matrix, where each column contains the ranks of the p-values within each trait. If the null hypothesis of no association is based on a uniform distribution, then the probability that rank $P_{i,j}$ (variant i and trait j) is equal to any rank is 1/N. The ranks can subsequently be summed per variant to create $\boldsymbol{R}$, which is a vector of length N with $m \leq R_i \leq mN$.

SumRank tests the likelihood of $\boldsymbol{R_i}$ being equal to rank r or more extreme under a null hypothesis of randomly assigned ranks. Consider a set of vectors $x_1, \ldots, x_m$ such that $x_1 + \ldots + x_m = \boldsymbol{R_{sum,i}}$ with $1 \leq x_i \leq \boldsymbol{N}$. Let $\boldsymbol{S}$ be a set of solutions with $\boldsymbol{x_i} \geq 1$ and let $\boldsymbol{S}_i$ be the set of solutions with $\boldsymbol{x_i} < \boldsymbol{N}$ for exactly that i, then:

$$|T| = |\overline{S_1} \cap \overline{S_2} \ldots \overline{S_m}| \tag{4}$$

Based on the inclusion/exclusion criteria we can rewrite eq. (3) as:

$$|T| = |S| - \sum_{1 \leq i \leq m} |S_i| + \sum_{1 \leq i < j \leq m} |S_i \cap S_j| + \cdots + (-1)^m |S_i \cap S_j \ldots S_m| \tag{5}$$

where $|\boldsymbol{S}|$ is a combinatorial composition which equals $\binom{r-1}{m-1}$, or in other words, how many combinations of m-1 positive integer numbers can give us a sum equal to r-1. Therefore, $|S_i| = \binom{r-N-1}{m-1}$, and $|S_i \cap S_j| = \binom{r-2N-1}{m-1}$, etc., and:

$$\sum_{1 \leq i \leq m} |S_i| = \binom{r-N-1}{m-1}\binom{m}{1} \tag{6}$$

$$\sum_{1 \leq i < j \leq m} |S_i \cap S_j| = \binom{r-N-1}{m-1}\binom{m}{2} \tag{7}$$

etc. It is now possible to simplify eq. (4):

$$|T| = \sum_{k=0}^{d} (-1)^k \binom{r-kN-1}{m-1}\binom{m}{k} \tag{8}$$

where:

$$d = \left| \frac{(r-m)}{N} \right| \tag{9}$$

We can further divide $|\boldsymbol{T}|$ in eq. (7) by the total number of vectors of length m – as they arise with equal probability under the null hypothesis – to derive:

$$P(R_i = r) = \frac{1}{N^m} \sum_{k=0}^{d} (-1)^k \binom{r-kN-1}{m-1} \binom{m}{k} \tag{10}$$

Finally, we compute the p-value for SumRank:

$$P(R_i \leq r) = \frac{1}{N^m} \sum_{i=1}^{r} \sum_{k=0}^{d} (-1)^k \binom{r-kN-1}{m-1} \binom{m}{k} \tag{11}$$

In practice, we want to test the null hypothesis when $\boldsymbol{R_i}$ is much smaller than $\boldsymbol{N}$, and thus d will always be approximately zero. We can then simplify eq. (11):

$$P(R_i \leq r) = \frac{1}{N^m} \sum_{i=1}^{r} \binom{i-1}{m-1} \tag{12}$$

$$= \frac{\mathrm{m}}{m! N^m} \sum_{i=1}^{r} \frac{1}{i} \frac{i!}{(i-m)!} \tag{13}$$

$$= \frac{\mathrm{m}}{m! N^m} \sum_{i=1}^{r} (1-m+1) \ldots (i-1) \tag{14}$$

Further simplification yields eq. (3), shown in the Results section:

$$p_{sum,i} = \frac{\left(\sum_{j=1}^{m} P_{i,j}\right)^m}{m!}$$

Genotypic data

We used release version 3 for the imputed genotype datasets from the UK Biobank. The procedures for genotyping, imputation, and quality control have been described elsewhere[11]. In brief, blood samples from 488,377 participants were assayed with two different genotyping arrays: the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix and the Applied BioSystems UKB Axiom Array. After quality control of the markers and samples, the autosomes were phased with SHAPEIT3[13], using the 1000 Genomes phase 3 dataset as a reference[14], and subsequently imputed against the Haplotype Reference Consortium (HRC) dataset[15]. The data were also separately imputed against a combination of the 1000 Genomes phase 3 and the UK10K datasets[16], and all data were combined – using the HRC imputation when a variant was available in both – into a dataset consisting of 93,095,623 autosomal variants in 478,442 individuals.

Genotype simulations

The genotype simulations were conducted in a series of steps. First, we identified independent variants based on the full genetic dataset of the UK Biobank. We excluded individuals with a call rate of < 0.99; and we excluded variants with any missingness, a violation of the Hardy-Weinberg equilibrium ($p < 10^{-6}$), and less or more than two alleles. To obtain the independent variants we used PLINK2[17] with --indep-pairwise 50 5 0.05.

Second, we created 20 simulations where we randomly selected a subset of the independent variants that we assigned effect sizes for the traits that we aimed to simulate. In the case of two traits, we selected 1000 causal variants that only affected the first trait, 1000 causal variants that only affected the second trait, and 1000 causal variants that were pleiotropic and thus affected both traits. The effect sizes were drawn from a normal distribution $N(0,\ 1)$, and we set the Pearson's correlation between the effect sizes within a pleiotropic variant to 0.5.

Third, for each simulation we selected a subset of 100,000 individuals per trait, and we calculated the simulated traits based on the causal variants. The assigned trait value was set as the sum of the genetic component scaled by the heritability, and a complementary non-genetic component:

$$\boldsymbol{Y} = \sqrt{h^2}\boldsymbol{G} + N(0,\ 1 - h^2) \tag{15}$$

The Pearson's correlation between the traits was forced to 0.5, and the h^2 for both traits was set to 0.5.

Fourth, we conducted GWASs for the simulated traits and applied the pleiotropy methods. The GWASs were conducted with fastGWA[18]. Each GWAS was run on 100,000 individuals. The samples were non-overlapping, as overlap between the samples can lead to bias in the estimates[19], and correcting for this bias is beyond the scope of the current study. For the pleiotropy assessments we focused on SumRank, FCT, and CPMA.

Finally, to see whether the methods identified the non-pleiotropic or pleiotropic causal variants, we clumped the results using PLINK2[17] and assessed whether any causal variants were in linkage disequilibrium (LD) with the lead variants. We based the LD structure for clumping on the total sample that partook in the simulated GWASs. We set independent variants to be those that were genome-wide significant ($p < 5 \cdot 10^{-8}$) that were not in LD after clumping ($r^2 < 0.6$). A second round of clumping was performed against $r^2 < 0.1$ to define the lead variants. For each lead variant, we checked whether any causal variants in LD were pleiotropic or non-pleiotropic.

The simulations relied on several set parameters and modeling decisions. We performed multiple sensitivity analyses to understand how the behavior of the methods changed. For the sample size we reran the simulations with 50,000 and 150,000 individuals per GWAS. For heritability we also simulated h^2 being equal to 0.25 and 0.75. Finally, we changed the Pearson's correlation of the effect sizes within pleiotropic variants to 0.0 and 1.0.

Application to psychiatric disorders

For the psychiatric disorders we used data from the GWASs performed as part of the Psychiatric Genomics Consortium (PGC) (Supplementary Table 1). The GWASs were focused on anorexia nervosa[20], attention-deficit/hyperactivity disorder[21], autism spectrum disorder[22], bipolar disorder[23], major depressive disorder[24], obsessive-compulsive disorder[25], schizophrenia[26], and Tourette syndrome[27].

To assess pleiotropy, we applied SumRank to all possible combinations of the GWASs, i.e. 247. For each variant we identified the combination with the lowest p-value. The genome-wide significant threshold was set to $5 \cdot 10^{-8} / 247 = 2.02 \cdot 10^{-10}$. The results were further processed in FUMA to identify the loci and obtain information on the functional consequences[28].

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource under application number 23509. This work was supported by the Netherlands Organization for Scientific Research (NWO/ZonMW grant 016.VICI.170.200 [SL, HT], NWO VENI 916.19.151 [SL, HA]).We thank SURFsara (www.surfsara.nl) for their support in using the Dutch national supercomputer Cartesius.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

All summary statistics from the PGC can be obtained from https://www.med.unc.edu/pgc/download-results/. The raw UK Biobank data can be obtained through a data access application available at https://www.ukbiobank.ac.uk.

CODE AVAILABILITY

All code is freely available. The code for SumRank as well as FCT and CPMA are available through the "sumrank" R package (https://github.com/slamballais/sumrank). The code used to download the summary statistics, run the simulations, and perform the analyses can be found in a separate repository (https://github.com/slamballais/sumrank_paper).

REFERENCES

1. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**, 1339-1348 (2019).
2. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
3. Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PLoS One* **8**, e65245 (2013).
4. van der Sluis, S., Posthuma, D. & Dolan, C.V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* **9**, e1003235 (2013).
5. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* **7**, e1002254 (2011).
6. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet* **90**, 821-35 (2012).
7. Andreassen, O.A., Thompson, W.K. & Dale, A.M. Boosting the power of schizophrenia genetics by leveraging new statistical tools. *Schizophr Bull* **40**, 13-7 (2014).
8. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**, 709-17 (2016).
9. Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address, p.m.h.e. & Cross-Disorder Group of the Psychiatric Genomics, C. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469-1482 e11 (2019).
10. Ray, D. & Chatterjee, N. A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between Type 2 Diabetes and Prostate Cancer. *PLoS Genet* **16**, e1009218 (2020).
11. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
12. Lee, C.H., Shi, H., Pasaniuc, B., Eskin, E. & Han, B. PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. *Am J Hum Genet* **108**, 36-48 (2021).
13. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet* **48**, 817-20 (2016).
14. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
15. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).
16. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, 8111 (2015).
17. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
18. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet* **51**, 1749-1755 (2019).
19. LeBlanc, M. *et al.* A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC Genomics* **19**, 494 (2018).
20. Watson, H.J. *et al.* Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet* **51**, 1207-1214 (2019).

21. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet* **51**, 63-75 (2019).
22. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**, 431-444 (2019).
23. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet* **53**, 817-829 (2021).
24. Howard, D.M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* **22**, 343-352 (2019).
25. International Obsessive Compulsive Disorder Foundation Genetics, C. & Studies, O.C.D.C.G.A. Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol Psychiatry* **23**, 1181-1188 (2018).
26. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* (2022).
27. Yu, D. *et al.* Interrogating the Genetic Determinants of Tourette's Syndrome and Other Tic Disorders Through Genome-Wide Association Studies. *Am J Psychiatry* **176**, 217-227 (2019).
28. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

Figure 1. Performance of FCT, CPMA, and SumRank under theoretical conditions. (A) Rejection region plots for the different methods given two p-values. (B) Violin plot for the different methods under the condition where the methods are applied to a p-value of $5 \cdot 10^{-8}$ and uniformly sampled p-values. The y-axis indicates the p-value derived from the methods. (C) Similar to panel B, but the fixed p-value is varied across a range (x-axis) and plotted against the proportion of tests that yield statistically significant results (y-axis). CPMA = Cross-Phenotype Meta-Analysis; FCT = Fisher's combination test.

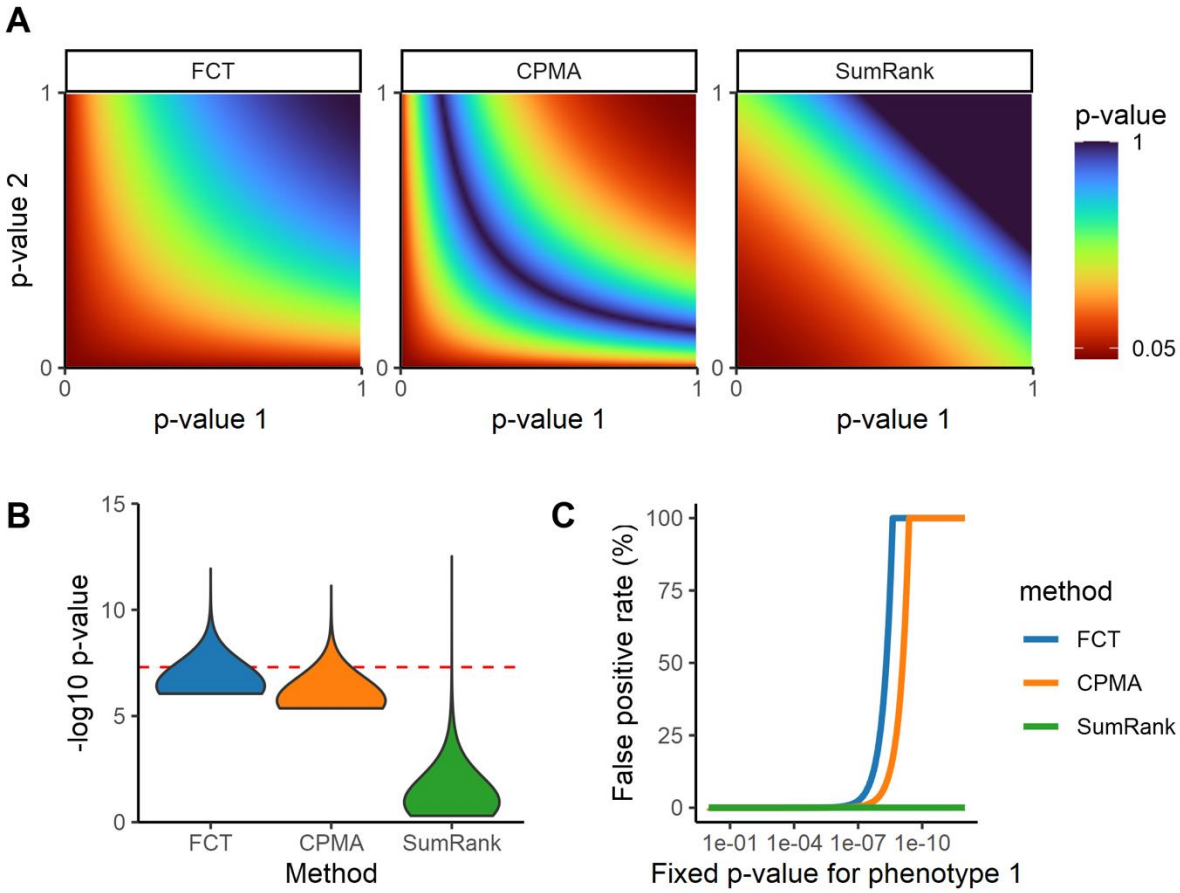

Figure 2. Results for FCT, CPMA, and SumRank for the genotype simulations of two traits, across 100 simulations, for various genetic architectures. Each trait was simulated to have 2000 causal variants, with each condition varying in the degree of overlap: (A) 50%, (B) 25%, (C) 75%. The left columns display the number of shared causal variants (green) and unique causal variants (red) for various simulation conditions. The right column contains bar plots of the variants that each method classified as pleiotropic, and therefore an overview of the number of correct and incorrect findings.

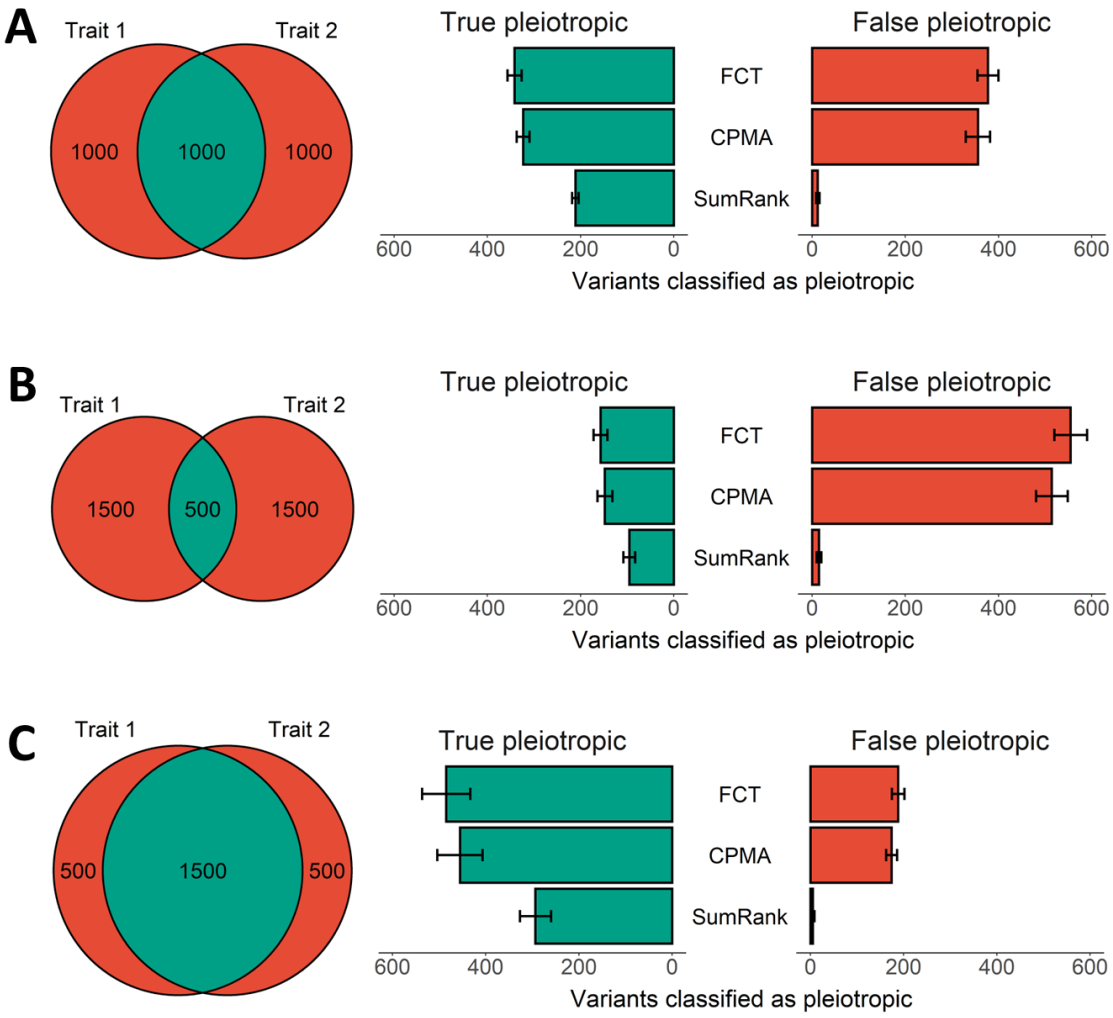

Figure 3. True pleiotropy rate for the most significant hits per method. The x-axis represents the p-values, sorted from lowest to highest. The y-axis represents the true pleiotropy rate for the top variants up to that point.

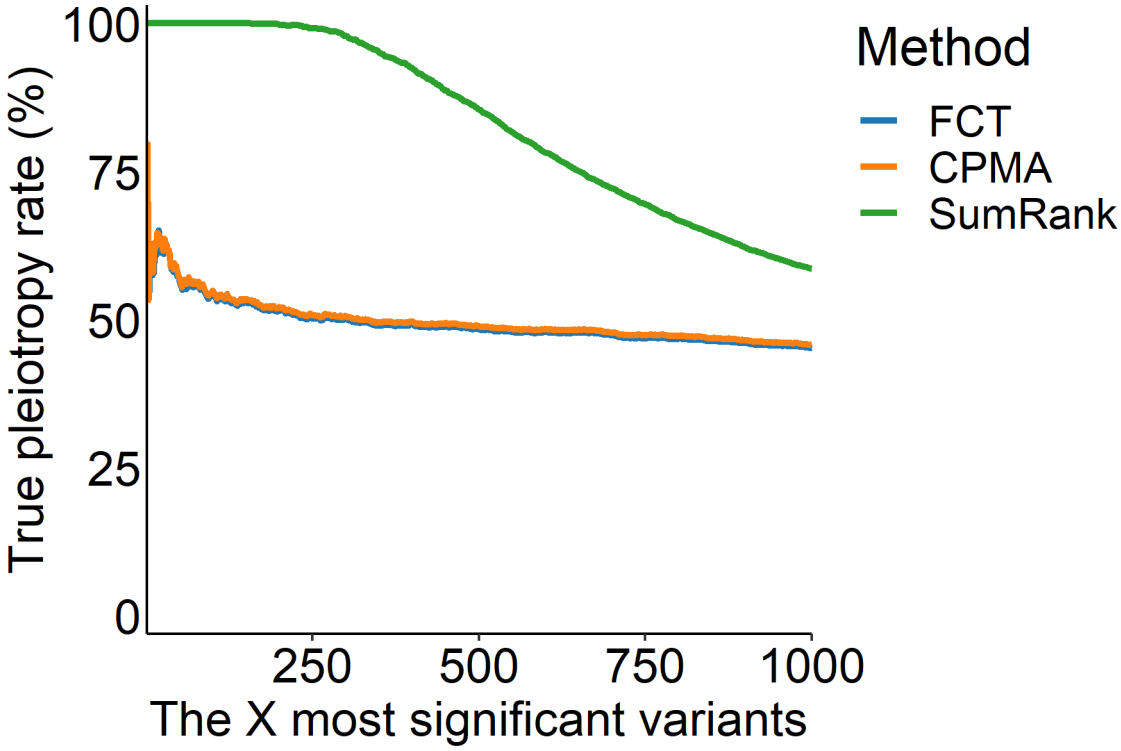

Figure 4. Results for FCT, CPMA, and SumRank for the genotype simulations of four traits, across 100 simulations, for various genetic architectures. The left column displays the genetic architectures. The right column contains bar plots of the variants that each method classified as pleiotropic for all four traits, and therefore an overview of the number of correct and incorrect findings. The conditions mimic settings where (A) each possible combination of traits has an equal number of causal variants, (B) variants either only influence one or four traits, (C) the number of causal variants that influence more traits decreases.

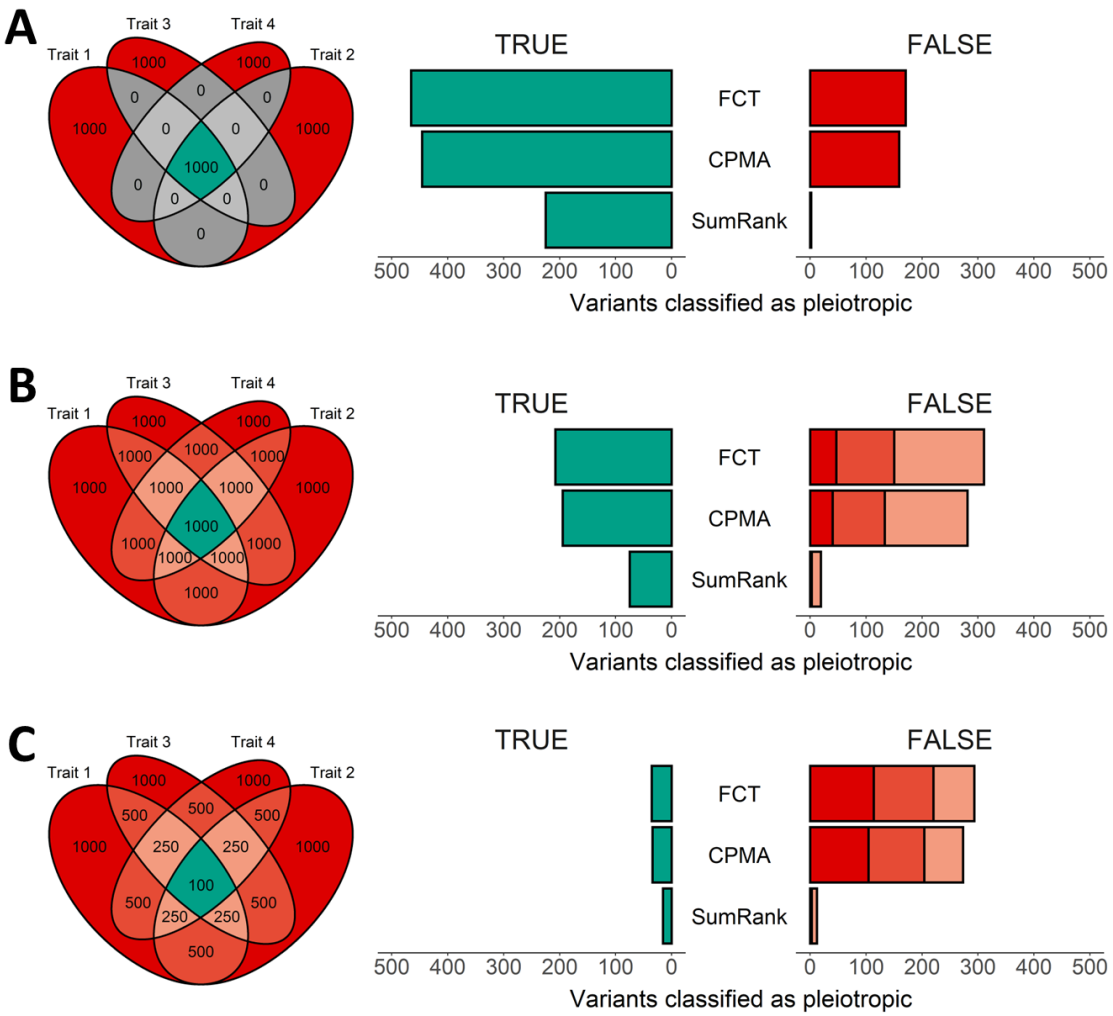

Figure 5. Results from SumRank on pleiotropy across eight psychiatric disorders. All combinations between the eight disorders were tested, totaling 247 combinations. (A) A Manhattan plot where each variant is represented by the lowest p-value across the 247 combinations. (B) Functional consequences of the pleiotropic variants. (C) The number of traits that independent variants associated with.

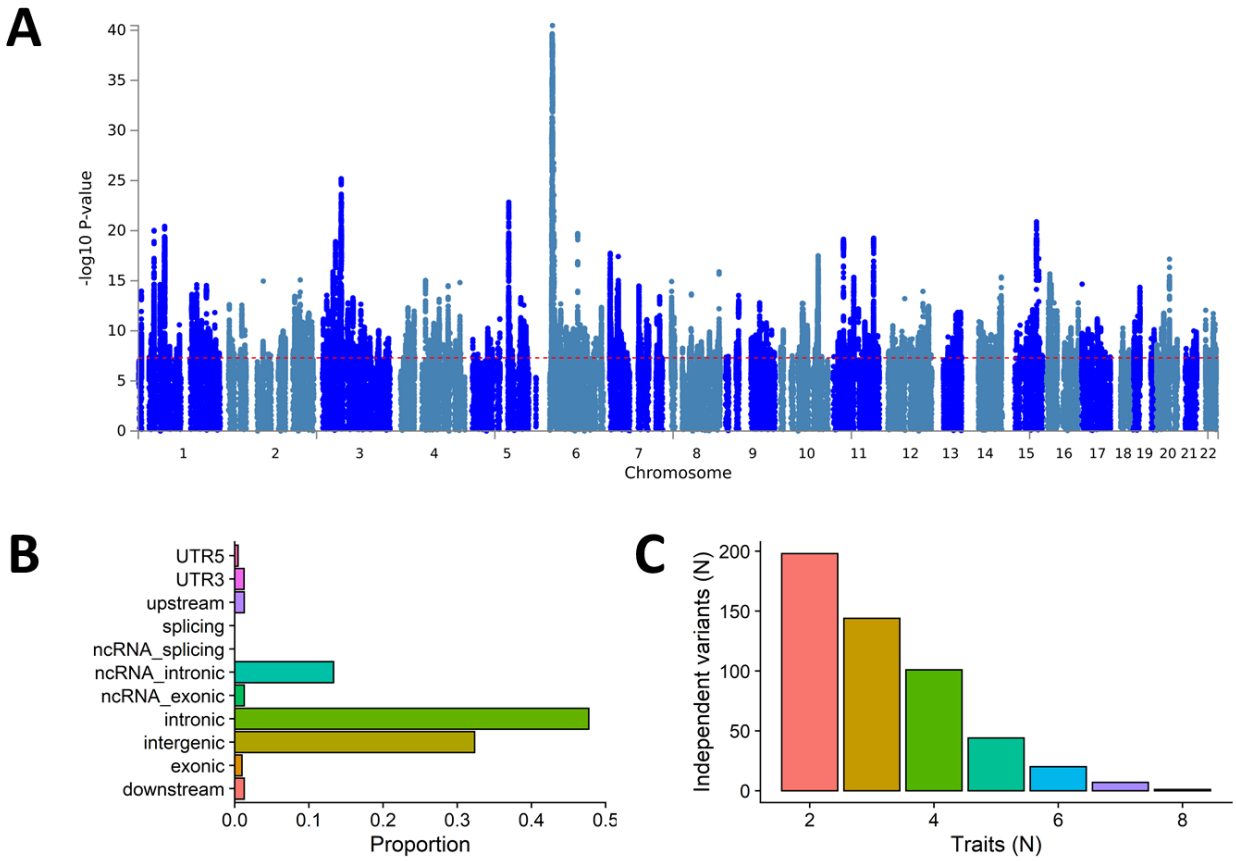
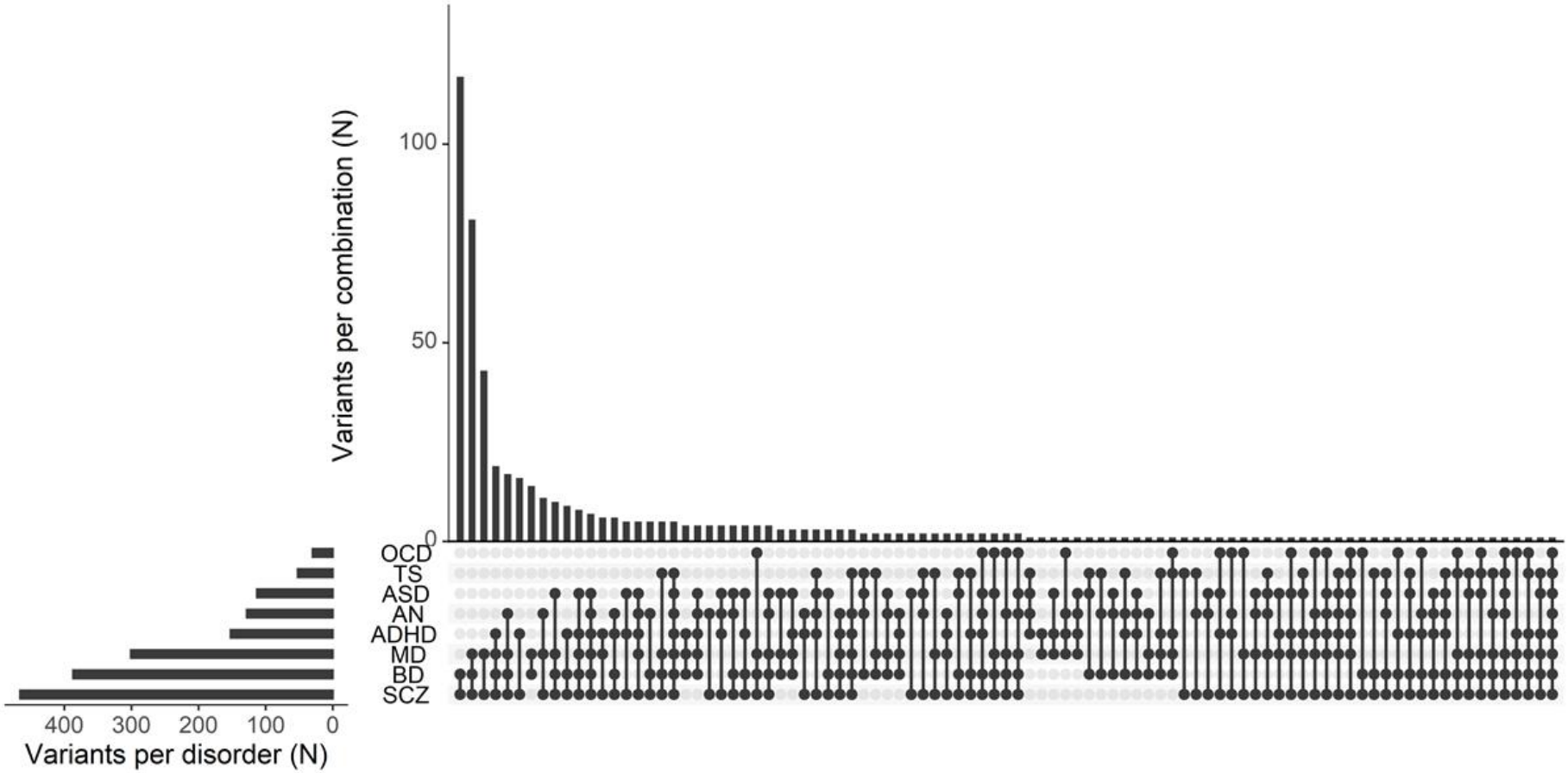
Figure 6. Upset plot of the SumRank pleiotropy results. ADHD = attention deficit hyperactivity disorder; AN = anorexia nervosa; ASD = autism spectrum disorder; BD = bipolar disorder; MD = Major depressive disorder; OCD = obsessive-compulsive disorder; SCZ = schizophrenia; TS = Tourette syndrome.