A comprehensive benchmarking and validation study of cross-trait association methods

Sander Lamballais[1]*; Gennady V. Roshchupkin[2,3]*; Raymond A. Poot[4]; Steven A. Kushner[5,6]; M. Arfan Ikram[2]; Hieab H. H. Adams[7,8†]; Henning Tiemeier[9†].

[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] Department of Psychiatry, Columbia University, New York, New York, USA.

[7] Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands.

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

[9] Department of Social and Behavioral Science, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA.

* Equal contribution

† Equal contribution

CORRESPONDENCE

Hieab H. H. Adams, MD, PhD. Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands. E-mail address: hieab.adams@radboudumc.nl.

Abstract

Cross-trait analyses are a powerful approach for refining our understanding of the genetic architectures underlying complex traits. Although many cross-trait association methods exist, a systematic evaluation of their performance is lacking. We compare true and false positive rates of several other methods using numerical (up to 300 traits) and genotype (up to 4 traits) simulations and introduce a new cross-trait association method, SumRank. For two traits, ConjFDR, SumRank, and GPA showed high true positive rates while maintaining false positive rates below 5%. SumRank and ASSET outperformed other methods for more than two traits. Most other methods had false positive rates well above 5%, with the false positive rates rising with the number of traits. Application of SumRank to eight psychiatric disorders yielded 658 novel loci, tripling the number of known loci. We discuss the differences in performance in cross-trait analyses as well as the risk of inflated false positives.

Introduction

Pleiotropy occurs when a single gene or genetic locus influences more than one trait or disease. It explains how a limited number of genes can impact numerous traits simultaneously through shared biological pathways. By leveraging pleiotropy, insights into the genetic architecture of one disease can facilitate the elucidation of another. One powerful approach to detect pleiotropy is through cross-trait association analysis, which examines associations of genetic variants–such as single nucleotide polymorphism (SNPs)–with multiple traits. Such analyses capture both horizontal pleiotropy, where independent biological processes affect distinct traits, and vertical pleiotropy, in which one trait causally influences another. A recent study estimated that cross-trait associations span 57% of loci in the genome[1], but this is likely an underestimation due to stringent statistical thresholds and our still-evolving understanding of pleiotropy. A complete understanding of pleiotropy thus requires the development and validation of robust cross-trait association methods capable of reliably prioritizing truly pleiotropic SNPs.

Genome-wide association studies (GWASs) are commonly used to uncover genetic architectures by evaluating SNP-trait associations[2]. A common approach to studying cross-trait associations with GWAS is by using multivariate GWAS, which analyzes individual-level data on multiple traits in the same study population. However, such data are often either unavailable or impractical to obtain. An alternative approach is to use summary-level methods, which employ only summary-level GWAS statistics and can be conducted on independent, non-overlapping populations. Summary-level methods estimate the probability that a SNP is pleiotropic by integrating effect sizes or p-values across different GWASs. Notable examples include *Fisher's Combined Test* (FCT), *Cross-Phenotype Meta-Analysis* (CPMA)[3], *ASsociation analysis based on subSETs* (ASSET)[4], *Genetic analysis incorporating Pleiotropy and Annotation* (GPA)[5], *conjunctional FDR* (conjFDR)[6], *GWAS-pairwise* (GWAS-PW)[7], and *PLeiotropic Analysis under COmposite null hypothesis* (PLACO)[8] (**Table 1**). With the continuously growing number of GWASs available, these summary-level methods offer a straightforward and scalable approach to unravel the complex pleiotropic nature of the human genome.

Importantly, each of these methods relies on distinct statistical frameworks and assumptions, which lead to heterogenous performance across different scenarios. Moreover, most methods have only been tested under specific simulated conditions, and direct comparisons of their false positive and

true positive rates remain lacking. Therefore, a comprehensive benchmark is needed to evaluate the performance of current approaches, thereby guiding the development of novel methods. Additionally, certain properties of one method could potentially be adapted to enhance others, paving the way for more powerful and reliable detection of cross-trait associations.

Here, we introduce SumRank as a novel method for cross-trait analysis and systematically compared its performance with existing summary-level methods in detecting valid cross-trait associations under diverse conditions. To benchmark the performance of the various methods, we created a comprehensive simulation framework. First, we tested the behavior of eight methods on simple, numerical simulations comparing two traits, focusing on false positive and true positive rates. Second, to approximate the complexity of real-world scenarios, we performed genotype simulations based on UK Biobank data, modeling genetically related traits. Third, given the growing interest in cross-trait associations across numerous traits, we extended the numerical simulations to 300 traits for a subset of methods, exploring approaches to properly account for false positives. These simulations were further corroborated by genotype simulations based on the UK Biobank data. Finally, we applied the cross-trait methods to GWAS summary data of eight psychiatric traits as a real-world application.

Results

Overview of methods for cross-trait associations

The approaches to detect pleiotropic SNPs greatly differ between the summary-level methods examined (**Table 1**), including in their hypotheses, the number of traits that can be analyzed simultaneously, their reliance on information from other SNPs, and the type of input data required. Still, the methods are unified in how they model the association between a SNP *i* and a trait Y*j* using a regression formula:

$$\mathrm{Y_j} = \beta_0 + \beta_{1,\mathrm{j}} g_i + \gamma \mathrm{w} + \varepsilon \qquad (1)$$

where $\beta_{1,\mathrm{j}}$ is the effect of SNP *i* on trait Y*j*, *g* is the genotype, γw is a non-genetic component, and ε is an error term. In this context, a pleiotropic SNP is defined as one where $\beta_{1,\mathrm{j}} \neq 0$ for multiple traits *j* among the *m* traits under consideration.

70 Here we will consider the differences between the cross-trait methods, starting first with the
71 hypothesis. Methods such as FCT and CPMA test a non-composite null hypothesis where all $\beta_{1,1}$
72 to $\beta_{1,m}$ are zero. This means that any non-zero $\beta_{1,j}$ will lead to a rejection of the null hypothesis,
73 even if only one trait is associated. Alternatively, a composite null hypothesis posits that at least
74 one $\beta_{1,j}$ is zero, which leads to rejection of the null hypothesis if all traits are associated.

75 Second, although most cross-trait methods can only test associations across two traits, others can
76 handle multiple traits simultaneously (**Table 1**). In the latter case, multiple null hypotheses can be
77 formulated. The first is to test whether at least one of $\beta_{1,1} \ldots \beta_{1,m}$ are zero but only results in cross-
78 trait associations when all traits are associated. The second approach, as is done with the ASSET
79 method, is to test every possible subset of *m* traits and select the optimal subset, enabling the
80 detection of cross-trait associations in specific trait combinations.

81 Third, most methods calculate the cross-trait associations for each SNP separately. However,
82 methods like GWAS-PW instead incorporate information on linkage disequilibrium (LD) and
83 optimize the model per defined genomic region, thus considering a broader genomic context.

84 Lastly, the methods differ in their input requirements – p-values, effect sizes, standard errors and
85 genetic correlations (r_g) – and their output– (un)adjusted p-values, posterior probabilities, or simply
86 the selection of associated traits (**Table 1**).

87 In this study, we introduce SumRank, a novel combinatorial method for evaluating cross-trait
88 associations based on p-values from genome-wide association study (GWAS) summary statistics.
89 SumRank explicitly tests the composite hypothesis that a SNP is associated with all traits under
90 study and can be extended to test for subsets of traits. We will compare various methods across a
91 range of simulations and conditions, demonstrating how SumRank performs relative to existing
92 methods.

93 Numerical simulations for two traits

94 The differences between cross-trait methods are best illustrated by comparing their rejection plots
95 for two traits (**Fig. 1a**). We present the plots for methods without (FCT and CPMA) and with a
96 composite null hypothesis (SumRank and PLACO). When the SNP has low p-values in both
97 GWASs (**Fig. 1a**, top-right in each plot), all methods correctly identify the SNP as pleiotropic.
98 However, when the SNP is associated with only one trait (**Fig. 1a**, top-left or bottom-right in each

99 plot), FCT and CPMA will still yield low p-values (i.e. pleiotropic), irrespective of the other p-
100 value's magnitude. This behavior becomes more apparent near or below the genome-wide
101 significance value of $5 \cdot 10^{-8}$ (**Fig. 1a**, white line in each plot): even if the other p-value equals 1,
102 FCT and CPMA still assign the SNP to be pleiotropic. By design, composite methods like
103 SumRank and PLACO are resistant to this behavior and require evidence of association in both
104 traits before assigning pleiotropy. Notably, even composite methods show subtle differences in
105 how resistant they are. For example, PLACO's cross-trait p-value does decrease slowly as one of
106 the p-values becomes increasingly lower, while SumRank plateaus, reflecting a more conservative
107 approach.

108 The implications of the rejection regions were further explored by simulating pairs of p-values. We
109 created one million pairs of p-values drawn from a uniform distribution, representing null
110 conditions. In this scenario, none of the methods had an inflated false positive rate for the two traits
111 (**Supplementary Fig. 1**). Next, we created one million pairs where the first p-value was always set
112 to $5 \cdot 10^{-8}$ (i.e., genome-wide significance) and the second p-value was drawn from the uniform
113 distribution. This represents the case where a SNP is only associated with the first trait and therefore
114 should not be classified as pleiotropic. We classified any cross-trait association with a p-value
115 below $5 \cdot 10^{-8}$ identified by any method as a false positive finding. Under these conditions, FCT and
116 CPMA had false positive rates of 4.81% and 0.80%, respectively, compared to 0.03% for SumRank
117 and PLACO (**Fig. 1b**). Finally, we investigated the behavior of false positive rates as the fixed p-
118 value threshold was progressively reduced, reaching as low as $5 \cdot 10^{-300}$. (**Fig. 1c**). The false positive
119 rates for FCT and CPMA reached 100% if the first p-value was set to $2.34 \cdot 10^{-9}$ and $4.07 \cdot 10^{-10}$,
120 respectively. PLACO exhibited a more conservative approach yet still resulted in a false positive
121 rate of 25% when the first p-value was set to $1 \cdot 10^{-40}$, and 50% when set to $1 \cdot 10^{-111}$. In contrast, the
122 false positive rate of SumRank remained 0.03% even when the first p-value was set to $5 \cdot 10^{-300}$.

123 Genotype simulations for two traits

124 To emulate a real-world scenario, we conducted a series of population-based genotype simulations
125 with varying genetic architectures. We used UK Biobank genotype data to simulate two correlated
126 synthetic traits, for which we then defined the underlying genetic architectures. Independent SNPs
127 with a minor allele frequency above 5% were selected and a random subset was assigned non-
128 pleiotropic and pleiotropic causal effects, effectively creating hypothetical traits. For the primary

129 simulations, the causal effects were drawn from a multivariate normal distribution. We obtained
130 summary statistics by conducting GWASs, which served as the input for all included cross-trait
131 association methods (**Table 1**). For each method, we clumped the results to obtain lead and
132 correlated SNPs in LD. A detailed description of the procedure can be found in the **Methods**
133 section.

134 For each of twenty simulations, we generated two traits for two independent cohorts of 100,000
135 individuals from the UK Biobank. As shown in **Fig. 2a**, each trait had 1000 non-pleiotropic causal
136 SNPs unique to the traits (red) and 1000 pleiotropic causal SNPs shared between them (green). To
137 evaluate the methods, we focused on the number of true pleiotropic SNPs that were detected and
138 the true pleiotropy rate, defined as the proportion of detected SNPs for which the true effect was
139 pleiotropic. The methods varied greatly in performance. For example, the two methods identifying
140 most pleiotropic SNPs were FCT (n = 346) and CPMA (n = 325) (**Fig. 2b**), but these methods also
141 had the lowest true pleiotropy rates (FCT = 47.8%, CPMA = 48.2%) (**Fig. 2c**). In contrast, methods
142 that relied on composite hypotheses generally identified fewer pleiotropic SNPs (e.g., $n_{conjFDR}$ =
143 246, n_{GPA} = 229, $n_{SumRank}$ = 190), but their true pleiotropy rates were much higher (conjFDR =
144 98.0%, GPA = 98.5%, SumRank = 99.4%).

145 GPA and conjFDR inherently correct the false discovery rate in their models. When we applied the
146 Benjamini-Hochberg procedure to the SumRank results, it identified 277 pleiotropic SNPs with a
147 true pleiotropy rate of 96.4%. Crucially, all SNPs discovered by the unadjusted SumRank were
148 also found by conjFDR, which were in turn found by GPA, and all of these were contained within
149 the Benjamini-Hochberg-adjusted SumRank results, indicating a nested pattern of pleiotropic hits
150 across methods (**Supplementary Fig. 2**).

151 Two methods involving composite hypotheses performed poorly in the two-trait simulations:
152 GWAS-PW, which exhibited a true pleiotropy rate of roughly 50% regardless of the threshold
153 applied to the posterior probability (**Supplementary Fig. 3**); and ASSET, which barely detected
154 any pleiotropic SNPs under these conditions (n = 1).

155 Overall, most true pleiotropic variants were found by FCT and CPMA while conjFDR, SumRank,
156 and GPA had the highest true pleiotropy rates. This relative pattern of performance remained
157 similar when varying the degree of overlap between the traits (**Fig. 2d-i**), the heritability of the

158 traits (**Supplementary Fig. 4**), the correlation between the effect sizes within each pleiotropic SNP
159 (**Supplementary Fig. 5**), and the sample size of the GWASs (**Supplementary Fig. 6**).

160 Numerical simulations for more than two traits

161 Next, we assessed performance in numerical and genotype simulations with more than two traits.
162 For these simulations, we modified the FCT, CPMA, and SumRank methods to utilize the same
163 subset-based approach as the ASSET method. In brief, given a set of m traits, ASSET identifies the
164 subset of traits that is the most likely cross-trait association for a SNP g_i. Other methods can be
165 modified to similarly detect a subset of traits by testing every subset of traits $k \in m$ and selecting
166 the subset k with the lowest p-value that reaches statistical significance. Finally, ASSET also
167 applies a p-value filter: in the analysis of a given SNP g_i only traits with a p-value below 0.05 are
168 included. This drastically reduces the number of possible subsets to test, which reduces the false
169 positive rate.

170 We first performed numerical simulations for the subset-based versions of ASSET, FCT, CPMA,
171 and SumRank. For each SNP, we simulated 10 to 300 trait p-values in total, and the SNPs were
172 causal for 0 to 5 of those traits (see **Methods** section). In these models, we assessed how often the
173 methods classified SNPs to be causal for the non-associated traits (false positive), the truly
174 associated traits (true positive), or only a subset of the truly associated traits (partial positive). First,
175 to illustrate the effect of the p-value filter, we did not apply any filter to the methods, as is typically
176 done in all methods but ASSET. The results showed a striking pattern: in the extreme scenario
177 where 300 traits were analyzed, the false positive rates of all methods varied between 30% and
178 100% (**Fig. 3a**, **Supplementary Fig. 7**). We then applied various p-value thresholds to these
179 methods, which resulted in false positive rates that were near 0% for all methods except for CPMA
180 (**Fig. 3b**).

181 The methods differed in the rates of true positive and partial positive cross-trait associations. For
182 example, for SNPs that were causal for five traits, SumRank had a partial positive rate of 60%,
183 compared to 45% for ASSET. These findings show that ASSET works well, and that other cross-
184 trait association methods can be modified to explore numerous traits simultaneously by
185 implementing a subset-based approach and filtering input p-values.

186 ## Genotype simulations for more than two traits

187 Next, we repeated the genotype simulations with UK Biobank data. As we limited ourselves to
188 non-overlapping GWASs of 100,000 individuals, we simulated a scenario for four traits, for
189 ASSET as well as the subset-based versions of FCT, CPMA, and SumRank. Each possible
190 combination of traits had causal SNPs, with SNPs being associated with one (n = 1000 per trait),
191 two (n = 500 per combination), three (n = 250), or four traits (n = 100; **Fig. 3c**). The analyses
192 yielded three types of outcomes: pleiotropic SNPs for which all associated traits were detected (true
193 positive), pleiotropic SNPs for which only a subset of associated traits were detected (partial
194 positive), and SNPs for which at least one non-associated trait was detected (false positive). When
195 considering only true positive SNPs, FCT identified most (1169 SNPs), followed by CPMA (1127),
196 ASSET (954), and SumRank (915) (**Fig. 3d**). However, SumRank identified a much larger number
197 of partial positive SNPs (763) compared to FCT (393), CPMA (367), and ASSET (562).
198 Furthermore, the false positive rate for SumRank was only 0.5%, compared to 10.0% for FCT,
199 10.9% for CPMA, and 1.5% for ASSET.

200 ## Application to psychiatric traits

201 To further explore the differences between the methods in real-world data, we applied SumRank,
202 FCT, CPMA, and ASSET to the summary statistics for publicly available GWASs of eight
203 prominent psychiatric disorders (**Supplementary Table 1**): anorexia nervosa, attention-
204 deficit/hyperactivity disorder, autism spectrum disorder, bipolar disorder, major depressive
205 disorder, obsessive-compulsive disorder, schizophrenia, and Tourette syndrome. The methods
206 were applied to find the optimal subset of traits that were associated with each SNP.

207 The original eight GWASs identified 304 loci. Across the four cross-trait methods, 658 novel cross-
208 trait loci were identified. SumRank identified the most novel cross-trait loci (n = 565, **Fig. 4a**),
209 followed by FCT (n = 400), CPMA (n = 244), and finally ASSET (n = 184) (**Supplementary Fig.**
210 **8-10**). Of the 658 novel loci, 253 were unique to SumRank, 67 to FCT, and 43 to ASSET (**Fig. 4b**),
211 and 128 loci were found by all four methods. On average, the locus length of the novel loci were
212 similar across the methods (**Supplementary Fig. 11**).

213 Next, we considered the distribution of associated traits that each method found in the 304
214 previously identified loci. SNPs in most of these loci were only associated with one of the original
215 GWASs. SumRank found a similar pattern, where the majority of identified SNPs in these loci

216 were still associated with only one trait (**Fig. 4c**). In contrast, FCT and CPMA assigned the majority
217 of these SNPs to be pleiotropic across two or three traits. This suggests that FCT and CPMA are
218 more likely to assign SNPs as pleiotropic when the SNP is genome-wide significantly associated
219 with one trait. This confirms the rejection plots, and the higher rate of false positives observed in
220 our previous simulations (see **Fig. 1a** and **Fig. 3b**).

221 When considering the newly discovered loci, the distribution of associated traits was similar across
222 the different methods (**Fig. 4d**). Interestingly, the unique loci that SumRank identified were
223 primarily due to a relative enrichment of SNPs pleiotropic for two traits. This is in line with the
224 high partial positive rate that SumRank had in the simulations (see **Fig. 3d**), suggesting that
225 SumRank finds more true positive pleiotropic loci than other methods, although it will not identify
226 all the traits associated with these loci.

227 To further explore any downstream differences between the methods, we functionally annotated
228 the SNPs. As expected from previous literature, the SNPs that were implicated by the cross-trait
229 association methods were most likely to be intronic or intergenic (**Supplementary Fig. 12**). For
230 each method, as the number of associated traits increased, the SNPs were increasingly likely to be
231 located in introns. When considering tissue-specific gene enrichment for the mapped genes, the
232 results of all methods pointed towards enrichment in brain-related tissues from the GTEx v8 atlas
233 (**Supplementary Fig. 13**).

234 Discussion

235 The evaluation of summary-level cross-trait association methods revealed distinct differences in
236 their performance but also common limitations. First, in numeric and genotype simulations of two
237 traits we observed substantial inflation of false positive rates for various methods, warranting
238 careful review of previous work that has been published with these methods. The best performing
239 methods–conjFDR, SumRank, and GPA–all achieved high true positive rates; meanwhile, time the
240 methods strictly controlled the false positive rate by using a composite null hypothesis. Second, in
241 numeric and genotype simulations of up to 300 traits, we showed that multiple methods can achieve
242 reasonable power and control the false positive rate by adapting the subset approach introduced by
243 ASSET. Amongst these adapted methods, SumRank had the highest true positive rate while
244 maintaining a stringent false positive rate, closely followed by ASSET. Third, our application to

245 psychiatric traits reaffirmed that cross-trait association methods can drastically improve the yield
246 from GWASs by leveraging pleiotropy.

247 The widespread interest in pleiotropy and the accessibility of GWASs have led to studies focusing
248 on cross-trait associations across tens[7,10-13], hundreds[1,14,15], or even thousands of traits[16-18]. Overall,
249 three different approaches should be distinguished to study cross-trait associations across more
250 than two traits. The first and most stringent approach identifies SNPs as pleiotropic only if they
251 reach genome-wide significance in all underlying GWASs[1]. While this method minimizes false
252 positive findings, it risks many false negative cross-trait associations due to loci with weak signals.
253 The second approach is to analyze all traits pairwise. This approach successfully revealed robust
254 cross-trait loci in previous studies[7]. However, being an extension of the first approach, it
255 underestimates loci with more than two associated traits, especially if the signal is weak. Our
256 findings indicate that a third approach, examining every subset of possible traits[4], is most effective
257 if more than two traits are studied. We show that this approach, which was first implemented in
258 ASSET, can be effectively applied to other methods when combined with p-value filtering
259 approaches to mitigate severe inflation of false positives. Overall, our findings suggest that two
260 cross-trait association methods, SumRank and ASSET, can both be used to validly study hundreds
261 of traits simultaneously.

262 Regardless of the number of traits, our study showed significant variation in false positives rates
263 of different methods. Some methods provide strict control (e.g., SumRank), others lenient control
264 (e.g., PLACO), while some fail to control false positives adequately (e.g., non-composite
265 hypothesis methods). Even among methods with seemingly appropriate control, false positives
266 were inflated under specific conditions. For example, composite hypothesis methods like PLACO
267 can produce severe inflation of false positives if a SNP associates strongly with one trait but not
268 the other, a limitation noted by PLACO's authors[8]. Still, no singular method objectively
269 outperformed all other methods. SumRank, for example, performed well across all simulation
270 conditions, but the choice to use other methods can be justified depending on the characteristics of
271 the study and the relative importance ascribed to false positive and false negative rates. Stringent
272 methods like SumRank may be best suited for functional follow-up studies requiring precise SNP
273 prioritization, while more lenient methods may be suitable for explorative studies of less well-
274 powered GWASs. Given the centrality of pleiotropy in human health, there is an urgent need for
275 consensus on the minimal requirements to properly test and validate cross-trait methods,

standardization of benchmark simulations and ground truth datasets, and establishment of metrics by which these methods should be evaluated.

The results should be interpreted with several limitations in mind. First, while our simulations cover a broad spectrum of scenarios, they cannot capture all the complexities of real-world situations. For example, we selected causal SNPs that were reasonably independent, whereas genomic regions typically harbor multiple causal SNPs. Second, we focused on cross-trait associations rather than pleiotropy itself. Cross-trait association methods cannot distinguish horizontal from vertical pleiotropy. Thus, results from these methods should not be interpreted as causal without appropriate follow-up work using pleiotropy-specific methods and functional validation studies. Third, we only evaluated cross-trait association methods that are commonly used or cited.

This study elucidates the intricacies of cross-trait genetic association methods, highlighting significant shortcomings and strengths of various approaches. Our work emphasizes the need for more reliable cross-trait association methods and a standardized simulation framework to validate and develop them. By advancing such a framework and by introducing SumRank, we hope to stimulate methodological innovation in the study of cross-trait associations and pleiotropy.

Methods

SumRank

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 SNP. Considering m traits with N SNPs, $\boldsymbol{P}$ represents an N x m matrix, where each column contains the ranks of the p-values within each trait. The null hypothesis tested with the SumRank method represents the scenario where at most m - 1 traits are associated with the SNP. Under the null hypothesis, the distribution of ranks within a GWAS is uniform, and therefore the probability that rank $P_{i,j}$ (SNP i and trait j) is equal to any rank is 1/N. The ranks can subsequently be summed per SNP to create $\boldsymbol{R}$, which is a vector of length N with $m \leq R_i \leq mN$.

SumRank tests the probability 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):

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

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

327 $$= \frac{\mathrm{m}}{m!N^m} \sum_{i=1}^{r} (1 - m + 1) \dots (i - 1) \quad (14)$$

328 As the number of independent SNPs increases, the formula can be further simplified to eq. (3),
329 shown in the Results section:

330 $$p_{sum,i} = \frac{(\sum_{j=1}^{m} P_{i,j})^m}{m!}$$

331 ## Genotype data

332 We used release version 3 for the imputed genotype datasets from the UK Biobank. The design of
333 the UK Biobank resource has been approved by the North West Centre for Research Ethics
334 Committee, and written informed consent was provided by all participants. The procedures for
335 genotyping, imputation, and quality control have been described elsewhere[19]. In brief, blood
336 samples from 488,377 participants were assayed with two different genotyping arrays: the Applied
337 Biosystems UK BiLEVE Axiom Array by Affymetrix and the Applied BioSystems UKB Axiom
338 Array. After quality control of the markers and samples, the autosomes were phased with
339 SHAPEIT3[20], using the 1000 Genomes phase 3 dataset as a reference[21], and subsequently imputed
340 against the Haplotype Reference Consortium (HRC) dataset[22]. The data were also separately
341 imputed against a combination of the 1000 Genomes phase 3 and the UK10K datasets[23], and all
342 data were combined – using the HRC imputation when a SNP was available in both – into a dataset
343 consisting of 93,095,623 autosomal variants in 478,442 individuals.

344 ## Genotype simulations

345 The genotype simulations were conducted in a series of steps. First, we identified independent
346 SNPs based on the full genetic dataset of the UK Biobank. We excluded individuals with a call rate
347 of < 0.99; and we excluded SNPs with any missingness, a violation of the Hardy-Weinberg
348 equilibrium ($p < 10^{-6}$), and less or more than two alleles. To obtain the independent SNPs we used
349 PLINK2[24] with --indep-pairwise 50 5 0.05.

350 Second, we created 100 simulations where we randomly selected a subset of the independent SNPs.
351 For the traits that we aimed to simulate, we assigned each of these SNPs an effect size. In the case
352 of two traits, we selected 1000 causal SNPs that only affected the first trait, 1000 causal SNPs that
353 only affected the second trait, and 1000 causal SNPs that were pleiotropic and thus affected both

354 traits. The effect sizes were drawn from a normal distribution $N(0, 1)$, and we set the Pearson's
355 correlation between the effect sizes within a pleiotropic SNP to 0.5.

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

359 $$\boldsymbol{Y} = \sqrt{h^2}\boldsymbol{G} + N(0, \sqrt{1 - h^2}) \quad (15)$$

360 The phenotypic correlation between the traits was forced to 0.5, and the h^2 for both traits was set
361 to 0.5.

362 Fourth, we conducted GWASs for the simulated traits and applied the pleiotropy methods. The
363 GWASs were conducted with fastGWA[25]. Each GWAS was run on 100,000 individuals. The
364 samples were non-overlapping, as overlap between the samples can lead to bias in the estimates[26],
365 and correcting for this bias is beyond the scope of the current study.

366 Finally, to see whether the methods identified the non-pleiotropic or pleiotropic causal SNPs, we
367 clumped the results using PLINK2[24] and assessed whether any causal SNPs were in LD with the
368 lead SNPs. We based the LD structure for clumping on the total sample that partook in the
369 simulated GWASs. We set independent SNPs to be those that were genome-wide significant ($p <$
370 $5 \cdot 10^{-8}$) and not in LD after clumping ($r^2 < 0.6$). A second round of clumping was performed against
371 $r^2 < 0.1$ to define the lead SNPs. For each lead SNP, we checked whether any causal SNPs in LD
372 were pleiotropic or non-pleiotropic.

373 The simulations relied on several set parameters and modeling decisions. We performed multiple
374 sensitivity analyses to understand how the behavior of the methods changed. For the sample size
375 we reran the simulations with 50,000 and 150,000 individuals per GWAS. For heritability we also
376 simulated h^2 being equal to 0.1, 0.2, 0.3 and 0.4. Finally, we changed the Pearson's correlation of
377 the effect sizes within pleiotropic SNPs to 0.0 and 1.0.

378 Numerical simulations for more than two traits

379 To assess the false positive and true positive rates for more than two traits, we simulated
380 associations between causal SNPs and 10 to 100 traits. SNPs were defined as causal for zero, one,
381 or five traits. The effect sizes per minor allele were drawn from a normal distribution $\mathrm{N}(0, \tau)$ with:

$$\tau = \frac{h^2}{(1-h^2)f_i(1-f_i)m} \quad (16)$$

where h^2 is the SNP heritability and m is the total number of causal SNPs. For our simulations, the SNPs were simulated as having a minor allele frequency of 0.1, an h^2 of 0.25, and an m of 1000. The genotype matrix **G** was generated for 100,000 individuals given the minor allele frequency. Error terms and the traits were calculated as:

$$\mathbf{Y} = \sqrt{\frac{h^2 v_G}{v_e}}\mathbf{G} + \mathbf{E} \quad (17)$$

where **E** is a matrix for error terms per person (rows) per trait (columns) drawn from distibution $N(0, \tau)$, v_G is the genotypic variance per SNP, v_e is the column-wise variance of **E**, and $\sqrt{\frac{h^2 v_G}{v_e}}$ weighs the genetic and phenotypic components. The SNPs in **G** were then associated with 10 to 100 traits using linear regression models, where the non-causal traits were drawn from a normal distribution $N(0, 1)$. The resulting p-values were used as input for the cross-trait association methods.

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 grant 916.19.151 [SL, HA], NWO Veni grant 1936320 [GR]).We thank SURFsara (www.surfsara.nl) for their support in using the Dutch national supercomputer Snellius.

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

The code for SumRank is available through the *ctar* R package, which can be downloaded at https://github.com/slamballais/ctar.

410 # References

411 1. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex
412 traits. *Nat Genet* **51**, 1339-1348 (2019).
413 2. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation.
414 *Am J Hum Genet* **101**, 5-22 (2017).
415 3. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet*
416 **7**, e1002254 (2011).
417 4. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the
418 combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet*
419 **90**, 821-35 (2012).
420 5. Chung, D., Yang, C., Li, C., Gelernter, J. & Zhao, H. GPA: a statistical approach to
421 prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet* **10**,
422 e1004787 (2014).
423 6. Andreassen, O.A., Thompson, W.K. & Dale, A.M. Boosting the power of schizophrenia
424 genetics by leveraging new statistical tools. *Schizophr Bull* **40**, 13-7 (2014).
425 7. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human
426 traits. *Nat Genet* **48**, 709-17 (2016).
427 8. Ray, D. & Chatterjee, N. A powerful method for pleiotropic analysis under composite null
428 hypothesis identifies novel shared loci between Type 2 Diabetes and Prostate Cancer. *PLoS*
429 *Genet* **16**, e1009218 (2020).
430 9. Salinas, Y.D., Wang, Z. & DeWan, A.T. Statistical Analysis of Multiple Phenotypes in
431 Genetic Epidemiologic Studies: From Cross-Phenotype Associations to Pleiotropy. *Am J*
432 *Epidemiol* **187**, 855-863 (2018).
433 10. Mocci, E. *et al.* Genome wide association joint analysis reveals 99 risk loci for pain
434 susceptibility and pleiotropic relationships with psychiatric, metabolic, and immunological
435 traits. *PLoS Genet* **19**, e1010977 (2023).
436 11. Xue, H., Wu, C. & Pan, W. Leveraging existing GWAS summary data of genetically
437 correlated and uncorrelated traits to improve power for a new GWAS. *Genet Epidemiol* **44**,
438 717-732 (2020).
439 12. McGrath, I.M., International Endometriosis Genetics, C., Montgomery, G.W. & Mortlock,
440 S. Genomic characterisation of the overlap of endometriosis with 76 comorbidities
441 identifies pleiotropic and causal mechanisms underlying disease risk. *Hum Genet* **142**,
442 1345-1360 (2023).
443 13. Rashkin, S.R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis
444 in two large cohorts. *Nat Commun* **11**, 4423 (2020).
445 14. Li, Y. *et al.* Combined genome-wide association study of 136 quantitative ear morphology
446 traits in multiple populations reveal 8 novel loci. *PLoS Genet* **19**, e1010786 (2023).
447 15. Tanha, H.M., Sathyanarayanan, A., International Headache Genetics, C. & Nyholt, D.R.
448 Genetic overlap and causality between blood metabolites and migraine. *Am J Hum Genet*
449 **108**, 2086-2098 (2021).
450 16. Barrio-Hernandez, I. *et al.* Network expansion of genetic associations defines a pleiotropy
451 map of human cell biology. *Nat Genet* **55**, 389-398 (2023).
452 17. Oblong, L.M. *et al.* Principal and independent genomic components of brain structure and
453 function. *Genes Brain Behav* **23**, e12876 (2024).

454 18. Zhang, Z. *et al.* A scalable approach to characterize pleiotropy across thousands of human
455 diseases and complex traits using GWAS summary statistics. *Am J Hum Genet* **110**, 1863-
456 1874 (2023).
457 19. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.
458 *Nature* **562**, 203-209 (2018).
459 20. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet* **48**, 817-20
460 (2016).
461 21. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
462 *Nature* **526**, 68-74 (2015).
463 22. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat*
464 *Genet* **48**, 1279-83 (2016).
465 23. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K
466 haplotype reference panel. *Nat Commun* **6**, 8111 (2015).
467 24. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
468 datasets. *Gigascience* **4**, 7 (2015).
469 25. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale
470 data. *Nat Genet* **51**, 1749-1755 (2019).
471 26. LeBlanc, M. *et al.* A correction for sample overlap in genome-wide association studies in
472 a polygenic pleiotropy-informed framework. *BMC Genomics* **19**, 494 (2018).
473 27. Fisher, R.A. *Statistical Methods for Research Workers*, (Oliver and Boyd, Edinburgh,
474 1925).

475

Table 1. Overview of summary-level methods used to study cross-trait associations.

Method	Year	Ref.	Traits	Hypotheses	Subset testing	Input	Output
FCT	1925	[27]	≥2	H_0: The SNP is not associated with any trait	Possible	p-values	p-value given H_0
CPMA	2011	[3]	≥2	H_0: The SNP is not associated with any trait	Possible	p-values	p-value given H_0
ASSET	2012	[4]	≥2	H_0: The SNP is not associated with any subset of traits	Yes	β, SE_β, N	Subset of (positive/negative) associated traits and p-values
GPA	2014	[5]	2	4-group model: π_{00} (no association), π_{10} (trait 1 only), π_{01} (trait 2 only) and π_{11} (pleiotropy)	No	p-values	Most likely model (π_{00}, π_{10}, π_{01} or π_{11})
ConjFDR	2014	[6]	2	H_0: The SNP is not associated with both traits	No	p-values	q-value for cross-trait association of SNP
GWAS-PW	2016	[7]	2	4-group model: π_{00} (no association), π_{10} (trait 1 only), π_{01} (trait 2 only) and π_{11} (pleiotropy)	No / not relevant	Z, r_g	Posterior probability for π_{10}, π_{01} and π_{11}
PLACO	2020	[8]	2	H_0: The SNP is not associated with both traits	No	p-values	q-value for cross-trait association of SNP
SumRank	2023		≥2	H_0: The SNP is not associated with all traits under study	Possible	p-values	p-value given H_0

Figure 1. Numerical simulations for two traits. (A) Rejection plots for FCT, CPMA, SumRank, and PLACO. The gray dotted lines denote $5 \cdot 10^{-8}$ for the input GWAS p-values, and the white line denotes where the methods output a p-value of $5 \cdot 10^{-8}$ or lower. (B) Violin plot of the cross-trait p-values when testing pairs consisting of a p-value of $5 \cdot 10^{-8}$ and a uniformly sampled p-value. 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).

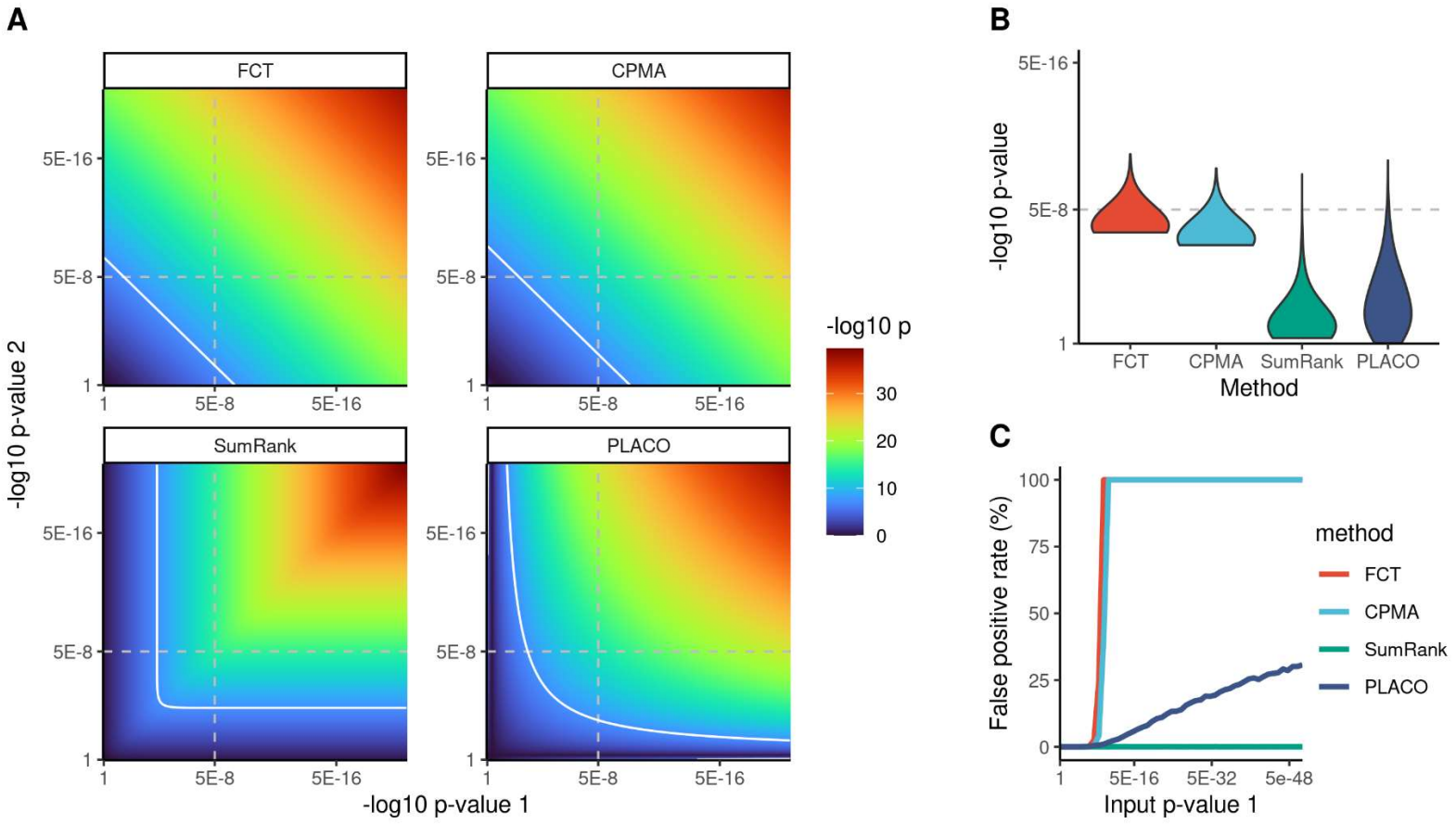

Figure 2. Genotype simulations for two traits. (A, D, G) Schematic of the causal SNPs for the two traits. Green and red indicate pleiotropic and non-pleiotropic SNPs, respectively. (B, E, H) The left columns display the number of shared causal SNPs (green) and unique causal SNPs (red) for various simulation conditions. The right column contains bar plots of the SNPs that each method classified as pleiotropic, and therefore an overview of the number of correct and incorrect findings. (C, F, I) The true pleiotropy rate per method, calculated as the proportion of true pleiotropic SNPs out of the SNPs detected by each method.

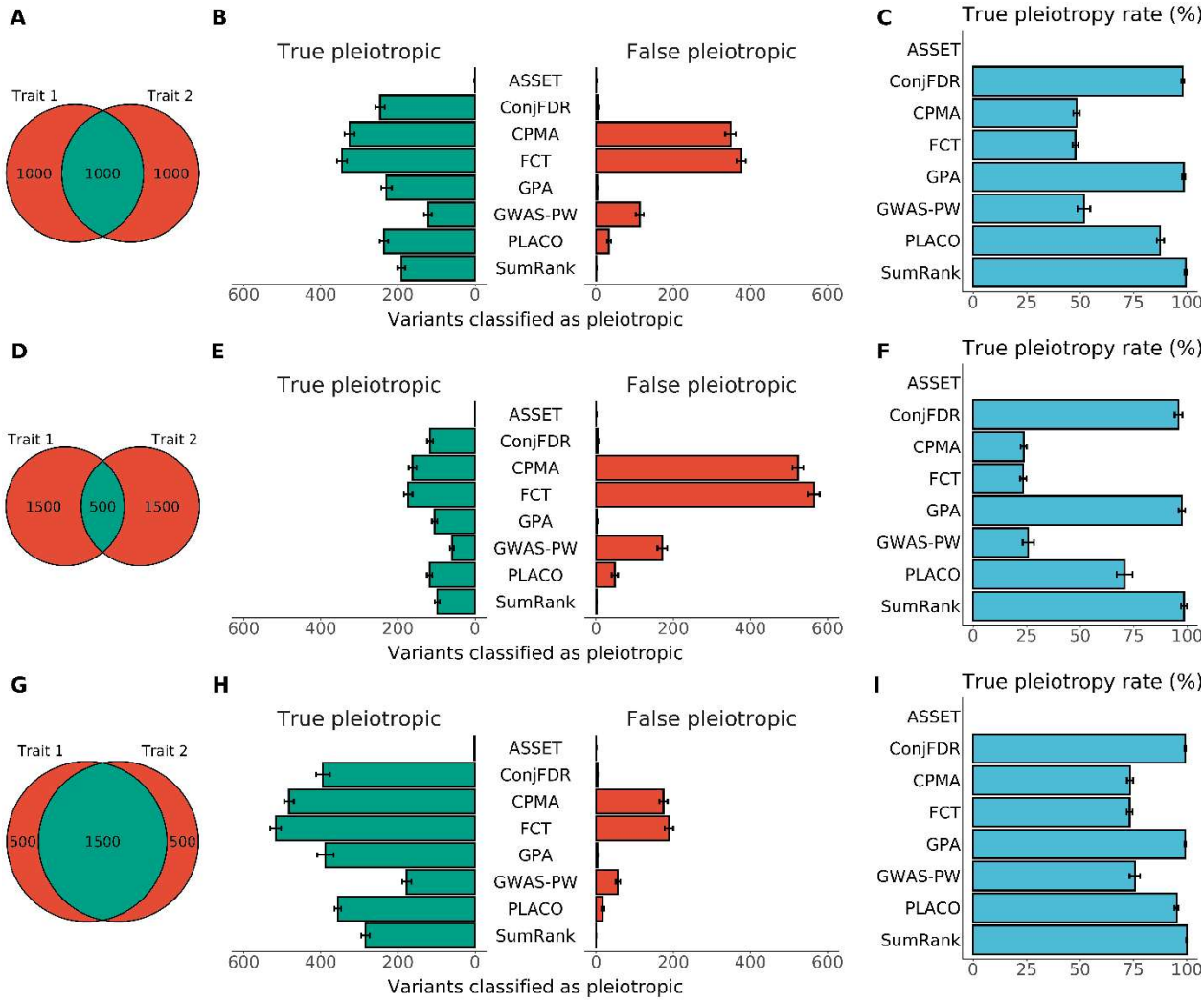

Figure 3. Numerical and genotype simulations for more than two traits. (A) Numerical simulations for SNPs that affect 0, 1, or 5 traits (columns) with metrics of false positives, true positives, and partial positive rates (rows), in analyses of 10 to 300 traits (x-axis). The partial positive rate denotes when a subset of the causal traits was identified, but not all. (B) When only those traits are included that have sufficiently low p-values (see Supplementary Figure 6). (C) Schematic of the causal SNPs for the genotype simulations of four traits. (D) Results from the genotype simulations.

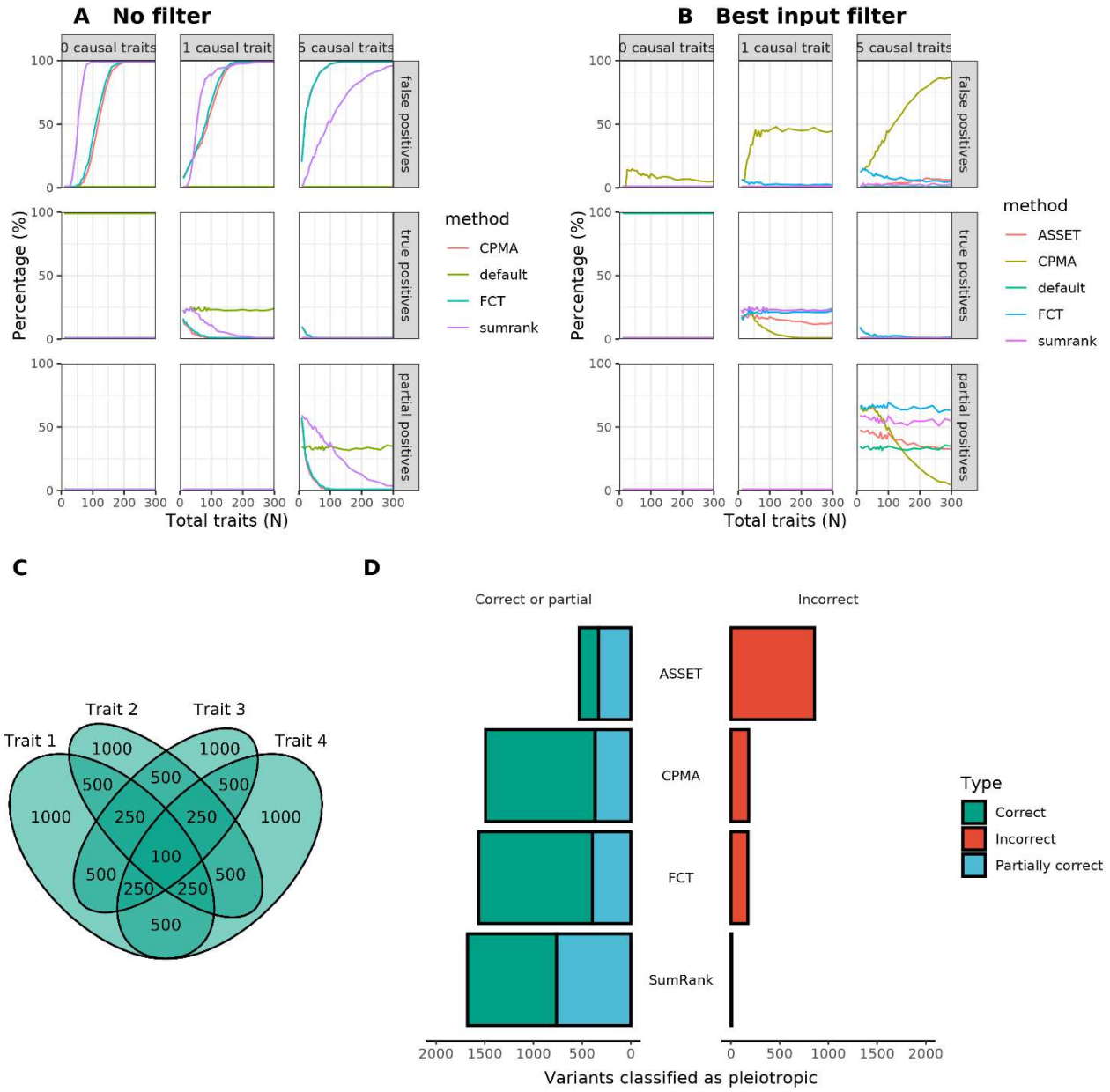
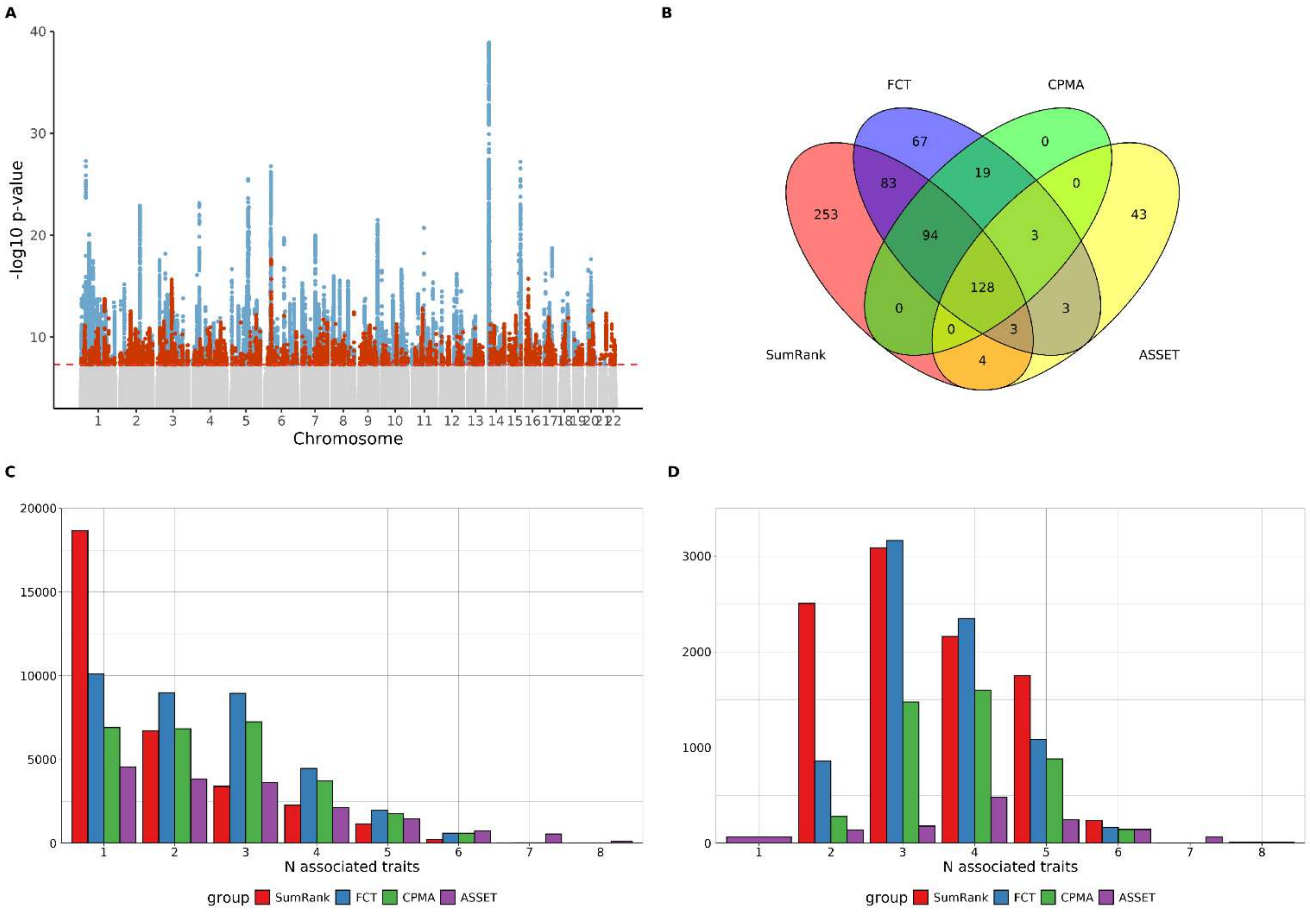
Figure 4. Cross-trait associations of eight psychiatric disorders. (A) Manhattan plot showing the results for the SumRank analysis. Each point represents the p-value for the optimal subset for a given SNP. Blue SNPs are genome-wide significant SNPs in loci that had previously been identified in one of the original GWASs. Dark orange SNPs represent genome-wide significant SNPs in novel loci. (B) Venn diagram of the overlap in detected regions by SumRank, FCT, CPMA, and ASSET. (C) Number of associated traits for the detected SNPs within the regions that were genome-wide significant in at least one of the original psychiatry GWASs. (D) Number of associated traits for the detected SNPs in novel loci.