Accurate detection of shared genetic architecture from GWAS summary statistics in the small-sample context

Thomas W. Willis[1*], Chris Wallace [1,2,3],

1 MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom
2 Cambridge Institute of Therapeutic Immunology and Infectious Disease, University of Cambridge, Cambridge, United Kingdom **3** Department of Medicine, University of Cambridge, Cambridge, United Kingdom

* thomas.willis@mrc-bsu.cam.ac.uk

Abstract

Assessment of the genetic similarity between two phenotypes can provide insight into a common genetic aetiology and inform the use of pleiotropy-informed, cross-phenotype analytical methods to identify novel genetic associations. The genetic correlation is a well-known means of quantifying and testing for genetic similarity between traits, but its estimates are subject to comparatively large sampling error. This makes it unsuitable for use in a small-sample context. We discuss the use of a non-parametric test of genetic similarity first introduced by Li et al. for application to GWAS summary statistics. We establish that the null distribution of the test statistic is modelled better by an extreme value distribution than a transformation of the standard exponential distribution as originally recommended by Li and colleagues. We show with simulation studies and real data from GWAS of 18 phenotypes from the UK Biobank that the test is to be preferred for use with small sample sizes, particularly when genetic effects are few and large, outperforming the genetic correlation and another nonparametric statistical test of independence. We find the test suitable for the detection of genetic similarity in the rare disease context.

Author summary

The genome-wide association study (GWAS) is a method used to identify genetic variants which contribute to the risk of developing disease. These genetic variants are frequently shared between conditions, such that the study of the genetic basis of one disease can be informed by knowledge of another, similar disease. This approach can be productive where the disease in question is rare such that a GWAS has less power to associate variants with the disease, but there exist larger GWAS of similar diseases. Existing methods do not measure genetic similarity precisely when patients are few. Here we assess a previously published method of testing for genetic similarity between pairs of diseases using GWAS data, the 'GPS' test, against three other methods with the use of real and simulated data. We present a new computational procedure for carrying out the test and show that the GPS test is superior to its comparators in identifying genetic similarity when the sample size is small and when the genetic similarity signal is less strong. Use of the test will enable accurate detection of genetic similarity and the study of rarer conditions using data from better-characterised diseases.

Introduction

Genetic pleiotropy is the association of a genetic variant with more than one trait and is a pervasive property of the human genome[1]. A consequence of this phenomenon is the sharing of causal variants between phenotypes. The relationship of multiple phenotypic characters to a common heritable factor was first inferred from observations of the coinheritance of certain traits and diseases. Of these, particularly memorable are Darwin's remarks that hairless dogs have imperfect teeth[2] and that blue-eyed cats are 'invariably deaf'[3,4].

Genome-wide association studies (GWAS) have revealed the extent of human genetic pleiotropy. A 2011 systematic review estimated that 4.6% of SNPs and 16.9% of genes had cross-phenotype effects[1]. These were likely underestimates given the incomplete catalogue of significantly associated SNPs surveyed, the review's stringent criteria, and the growth in the number of published GWAS in the decade since[5]. Pleiotropy is frequently identified among groups of diseases exhibiting similar pathobiology. Work by Cotsapas and colleagues found that 44% of immune-mediated disease risk SNPs across seven conditions were associated with multiple immune-mediated diseases[6]. Sharing of genetic architecture has also been observed among psychiatric illnesses, most prominently between bipolar disorder and schizophrenia[7,8].

Cross-phenotype effects contribute to observed phenotypic correlations. Hazel introduced the notion of partitioning the phenotypic correlation of two characters into an additive genetic component, the coheritability, and a joint non-additive genetic and environmental component, the coenvironmentability[9,10]. The genetic correlation is a heritability-standardised form of coheritability and quantifies the sharing of additive genetic effects between traits. First estimated in groups of related individuals in twin and family studies, in the genomic era genetic correlation can be quantified in large cohorts of unrelated individuals using individual-level genotype data or genetic effect estimates from GWAS[11]. More precisely, these methods typically estimate the 'SNP-correlation', however, not the genetic correlation: their estimates include only the effects of common SNPs and exclude those of rarer SNPs and other forms of genetic variation, such as indels and structural variants.

In 2010 Yang and colleagues introduced a linear mixed model (LMM) approach to SNP-heritability estimation which made use of all genotyped SNPs, not just those attaining genome-wide significance[12]. This method is alternately called the genome-based restricted maximum likelihood (GREML) method or 'GCTA' after the accompanying software package[13]. It was later extended to the bivariate domain for the joint analysis of pairs of traits and estimation of their genetic correlation[14]. A drawback of the use of GCTA is its requirement for individual-level genotype data. Cross-trait linkage disequilibrium score regression (LDSC) allows the fast estimation of each trait's heritability and the genetic covariance of trait pairs using GWAS summary statistics instead[15,16]. LDSC correlation estimates for a large number of phenotype pairs are now publicly available online[17–19]. The more recently published SumHer program, part of the LDAK software package, offers similar functionality to LDSC[21]. Summary statistic-based methods like LDSC and SumHer have the advantage that they do not require access to individual-level genotype data, which can be difficult to obtain.

The relative merits of these methods in genetic correlation estimation has received less attention in the discourse than has their capacity to estimate heritability. In general, the sampling error of such estimates is relatively large as they combine estimates of coheritability and the heritability of both traits[22]. Wray et al. give the rule of thumb that at least 5,000 samples are needed per trait in order to identify as statistically significant a correlation of 0.2 with GCTA[23]. Guidance on the use of the LDSC software states that 5,000 samples are needed to achieve better than 'very noisy' results (`https://github.com/bulik/ldsc/wiki/FAQ`). The empirical standard errors

of LDSC estimates are approximately twice as large as those of GCTA[11] and LDSC thus requires much larger sample sizes to achieve estimates of the same precision[24].

The requirement for a large sample size for more precise correlation estimation is problematic when studying rare diseases, which yield only small numbers of cases. A recent GWAS of a broad phenotype of antibody-deficient primary immunodeficiency had only 722 cases[25] (PID has a minimum UK prevalence of around 1 in 17,000[26]), whilst the deeply-phenotyped UK Biobank (UKBB) cohort, consisting of around 500,000 middle-aged and elderly participants, featured only 415 and 391 cases for GWAS of self-reported systemic lupus erythematosus and schizophrenia[18,27], respectively, to take just two indicative diseases with prevalence approximately 1 in 1,000[28] and 4.6 in 1,000[29], respectively. This motivates the search for procedures with which to test for the presence of genetic sharing between traits in the small-sample size context. The need is particularly acute when making use of cross-phenotype, pleiotropy-informed methods to mitigate the problem of small sample sizes by conditioning on more highly-powered studies of related phenotypes[30–36]. Selection of these related traits must be driven by knowledge of their genetic similarity to the primary trait in the first place.

The nonparametric genome-wide pairwise-association signal sharing (GPS) test was devised by Li et al. to determine whether two traits are genetically related using GWAS p-values from marginal tests of the association of SNPs with each trait[37]. The test was originally applied to the detection of genetic similarity between a collection of paediatric autoimmune diseases. As a nonparametric test statistic, the GPS test statistic is not an estimator of a biologically meaningful parameter like the genetic correlation. Instead it is used only as evidence against a null hypothesis of the bivariate independence of the p-value vectors.

The GPS test is a candidate method for the detection of genetic similarity in the small-sample context. Here, we first establish the distribution of the GPS test statistic in the null case of bivariate independence, showing that it is better fit by a generalised extreme value distribution than a transformation of the standard exponential as originally proposed by Li and colleagues. We then compare the GPS test to tests of the genetic correlation as estimated by LDSC and SumHer using simulated GWAS and real GWAS from the UKBB. We show that the GPS test is more powerful than its comparators in the small-sample context and that it controls the type 1 error rate when applied to genetically uncorrelated phenotype pairs. We also show that the GPS test in this setting is superior to Hoeffding's test, a canonical statistical test of bivariate independence.

Results

The GPS test statistic is evaluated on a set of paired p-values $\{(u_i, v_i) | i = 1, \ldots, n\}$ in which u_i and v_i are p-values from tests of association of the same SNP i with the first and second traits, respectively, of a pair for which GWAS are available. The statistic is given by:

$$D = \sup_{(u,v)} \sqrt{\frac{n}{\ln n}} \frac{|F_{U,V}(u,v) - F_U(u)F_V(v)|}{\sqrt{F_U(u)F_V(v) - F_U(u)^2 F_V(v)^2}}, \tag{1}$$

where n is the total number of SNPs, $F_{U,V}$ is the empirical bivariate distribution function of (U, V), and F_U and F_V are the empirical distribution functions of U and V, respectively.

The null distribution of the GPS statistic is better fit by the generalised extreme value distribution

Li et al. proposed that the reciprocal of the squared GPS test statistic D is approximately distributed as a standard exponential random variable

$$\frac{1}{D^2} \sim \mathrm{Exp}(1) \tag{2}$$

under a null hypothesis of bivariate independence. We examined this proposal with the use of p-values from GWAS of 18 UKBB traits published by the Neale lab (Materials and methods)[18]. For each pair of traits, we randomly permuted the order of SNPs in one trait to generate realisations of the GPS test statistic under the null. For this we developed a computational procedure, making the generation of thousands of realisations feasible through the use of a recently published fast algorithm for the estimation and evaluation of multivariate ecdfs[38]. This procedure was sufficiently performant to allow the evaluation of the GPS test for a large number of trait pairs. We computed p-values for the GPS test according to (2) and found that the p-values were not uniformly distributed, in contravention of the expectation under the null (Fig 1A).

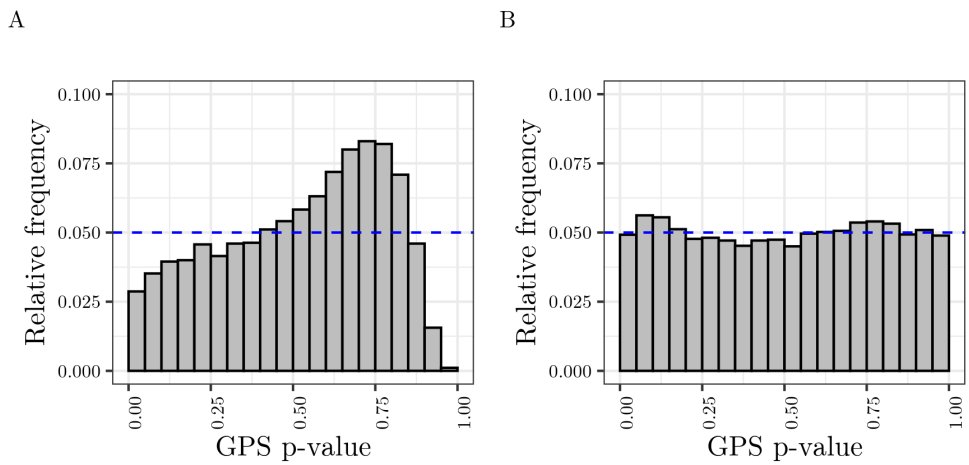

Fig 1. The empirical distribution of p-values under two null distributions. (A) The null from Li et al., the inverse square root of the standard exponential. (B) The null as a generalised extreme value distribution. The dashed line gives the expectation of p-values under the null. 10,000 null realisations of the GPS statistic were produced by permutation of the order of SNPs in GWAS of asthma and emphysema/chronic bronchitis from the UK Biobank collection discussed in Materials and methods.

The appearance of the supremum in the GPS test statistic suggested that its distribution might follow an extreme value distribution under the null and we found that the generalised extreme value distribution (GEVD) did indeed provide a much better fit as judged by uniformity of its p-values under the null hypothesis (Fig 1B). The GEVD fit null realisations of the GPS statistic well, diverging only in the right tail where the empirical distribution's tail was lighter than that of the fitted GEVD (Fig S1). The consequence of this behaviour is that the GEVD will generate conservative p-values for non-null values of the GPS statistic.

Estimates of each parameter of the GEVD varied as a function of the number of SNPs (Fig S2). We determined that, whilst the GEVD was the appropriate null distribution, it was necessary to obtain estimates of these parameters particular to each

pair of data sets; in other words, it was not possible to determine parameter values for a single reference GEVD null distribution to which to compare GPS test statistic values for any pair of data sets.

Despite the use of the fast ecdf algorithm the permutation procedure remained computationally burdensome, such that we sought to ascertain the smallest number of permutations needed to reliably estimate the GEVD parameters. We examined the value and precision of the estimates as a function of the number of permutations (Fig S3). We chose 3,000 permutations as a compromise between estimate accuracy and running time; parallelised across eight cores, 3,000 permutations took approximately one hour to obtain.

Simulation studies

To assess the GPS test's ability to detect genetic similarity, we compared it to a test of non-zero genetic correlation using either LDSC or SumHer, and to Hoeffding's test of bivariate independence. We simulated GWAS data for synthetic case-control 'traits' with five different genetic architectures (Table 1). For each sample size and genetic architecture, we specified the number of causal variants to be shared between traits. Each simulation condition was defined by this trio of values and 200 replicate pairs of GWAS were generated for each condition. As the induced heritability and coheritability depends not only on a fixed odds ratio but also the minor allele frequency, which varied between the randomly chosen causal variants, it was not possible to strictly specify the genetic correlation, only the number of shared causal variants. As a result, replicates have close but non-identical values of the genetic correlation.

Table 1. Details of the five simulation regimes used to generate simulated GWAS summary statistics with simGWAS.

Name	Odds ratio(s)	No. of causal variants	No. of shared causal variants
Large-effect I	1.2	25	5, 10, 15, 20, 25
Large-effect II	1.2	50	10, 20, 30, 40, 50
Mixed-effect I	1.05 and 1.2	200 (1.05) and 50 (1.2)	100 (1.05) and 0, 15, 25 (1.2)
Mixed-effect II	1.05 and 1.2	200 (1.05) and 50 (1.2)	200 (1.05) and 0, 15, 25 (1.2)
Small-effect	1.05	400	50, 100, 150, 200, 250, 300, 350, and 400

The GPS test, Hoeffding's test, SumHer, and LDSC controlled the type 1 error rate at $\alpha = 0.05$ (Fig 2). Both SumHer and LDSC failed to produce genetic correlation estimates for a proportion of simulated data sets with very small case numbers (Fig S4).

We found the GPS test exhibited appreciable power to detect genetic similarity in simulated data sets with as few as 1,000 cases across four of the five simulation regimes (Fig 3). In the small-effect regime, the GPS test's power remained low until the sample size was increased to 10,000 cases. Its power increased as the number of cases and the specified genetic correlation increased. For case numbers up to 50,000, the GPS was the most powerful method of those considered. At higher sample sizes, the GPS retained an advantage in its power to detect genetic similarity at the smallest non-zero values of the genetic correlation.

Application to immune and non-immune traits in the UK Biobank

The existence of pleiotropic loci manifesting cross-phenotype effects across many immune-mediated diseases (IMDs) is well-attested[6,37,39–42]. We surveyed UKBB disease traits to identify groups of immune- and non-immune-related diseases, including

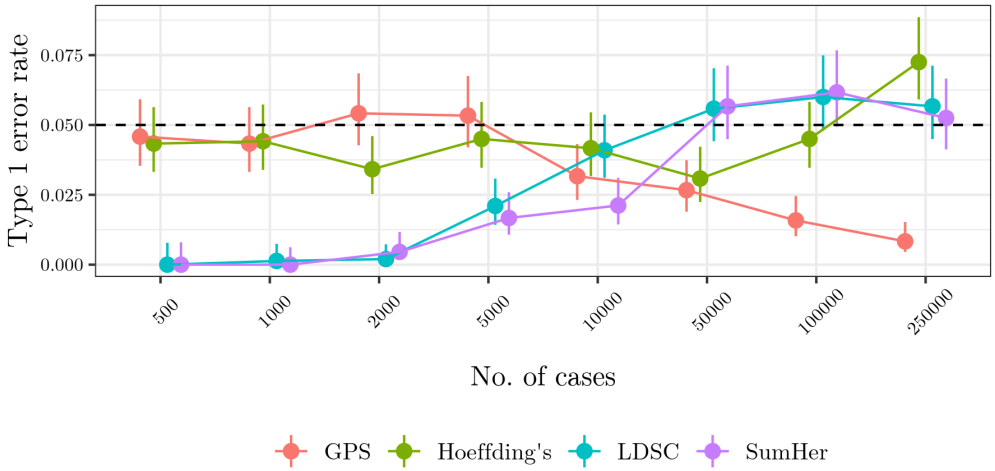

Fig 2. The type 1 error rates incurred by each method. The type 1 error rate for each method was estimated as the proportion of replicates for which $p \leq 0.05$ when the genetic correlation was set to zero. This proportion was measured only among those replicates producing a valid test statistic. 95% confidence intervals were calculated as Wilson score intervals. The dotted line depicts the size of the test, 0.05. A large proportion of replicates at case numbers smaller than 10,000 did not return valid test statistics for LDSC and SumHer (Fig S4) which caused the deflated type 1 error rates observed here at lower case numbers.

phenotypes with small, medium, and large sample sizes (Table 2). We hypothesised that a sensitive test would be more likely to give a positive signal where both diseases belonged to the IMD group than where diseases from the immune- and non-immune groups were paired together, or where both traits came from the group of largely unrelated non-immune diseases (collectively, the 'mixed' pairs). As before, we applied Hoeffding's test and tests of non-zero genetic correlation with LDSC and SumHer to serve as comparators for the GPS test.

The GPS test identified genetic similarity among IMD pairs across a range of sample sizes, generally attaining smaller p-values than its comparator methods for all but the largest sample sizes (Panel A of Fig S5). In particular, the GPS test produced nominally significant ($p \leq 0.05$) test statistics for 76% of immune/immune pairs in which the smaller of the pair of studies had fewer than 2,000 cases, whilst its comparators identified 10-29% (Table 3). For mixed pairs, both LDSC and SumHer outperformed the GPS test in identifying associated pairs for small and large samples alike (Table 3, Panel B of Fig S5 and Fig S6). We expected fewer significant results from the mixed pairs because of the pairing of ostensibly more aetiologically disparate phenotypes. This was borne out by the results of every method but LDSC, which identified a higher percentage of mixed pairs as significant than immune pairs.

The GPS statistic is computed as the supremum norm of a weighted distance between two distributions, rather than as an estimate of the integral of the squared distance as with a quadratic ecdf statistic like Hoeffding's test statistic, or as a heritability-standardised average of genetic effect covariance across the genome as with the genetic correlation. We investigated whether the GPS test might thus be sensitive to the isolated coincidence of strong association signals for two traits at a single locus in the absence of other evidence of genome-wide pleiotropy. The major histocompatibility complex (MHC) frequently exhibits highly significant association signals in GWAS of immune and non-immune diseases alike[43] and could plausibly account for significant

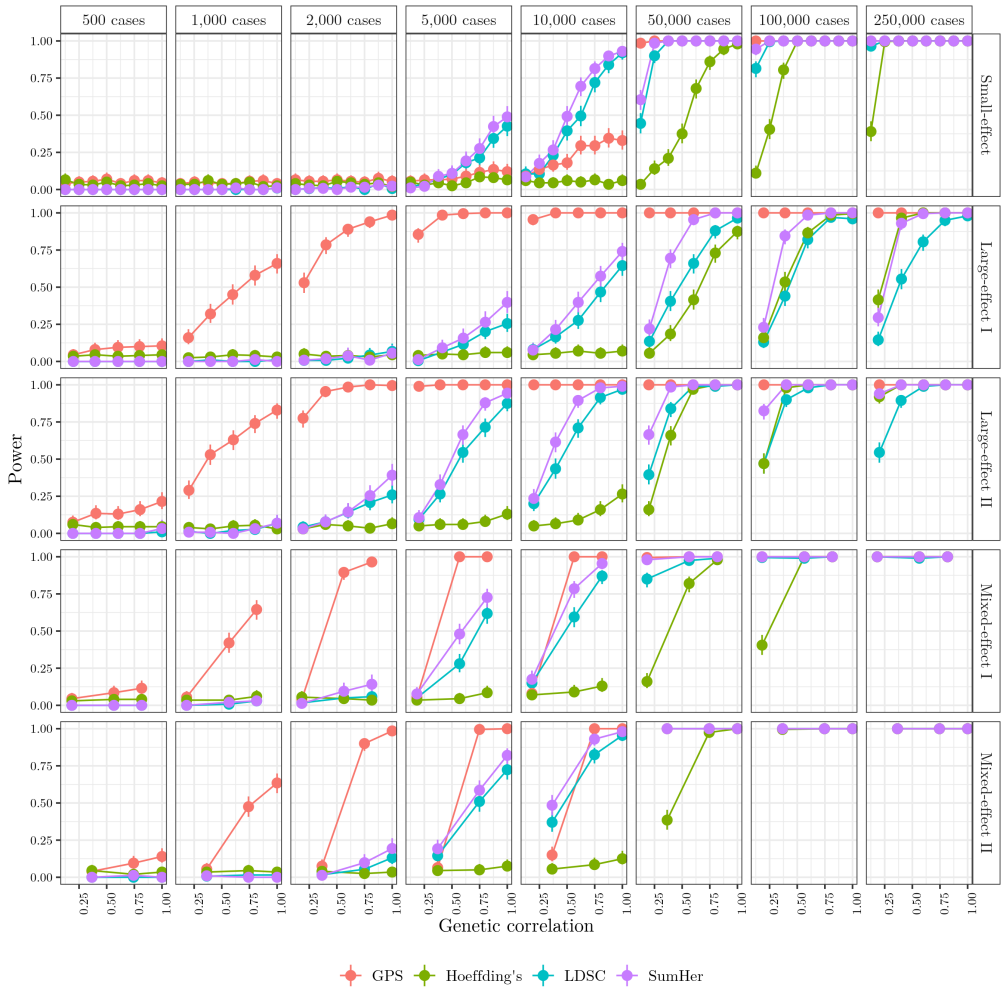

Fig 3. The power of each method to detect a non-zero genetic correlation in simulated data sets. Power was estimated as the proportion of replicate pairs for which the p-value reached nominal significance, $p \leq 0.05$, for a specified number of shared causal variants and sample size. This proportion was measured only among those replicates producing a valid test statistic. 95% confidence intervals were calculated as Wilson score intervals. As genetic correlation varies across replicate pairs according to the precise shared causal variants randomly chosen, we grouped observations according to the number of shared causal variants. The value of the genetic correlation for a point is the mean of the specified genetic correlation across replicate pairs grouped for that point's estimate.

GPS test statistics.

Removal of the MHC made GPS p-values less significant for 110 of 153 disease pairs tested. This effect was particularly noticeable for those immune/immune pairs which were significant when the MHC was included (Panel A of Fig S5 and Fig S7); only 15 of 31 of these remained significant after the removal of the MHC. We found that the pair of p-values accounting for the largest maximand, thus determining the value of the GPS test statistic, corresponded to a SNP located in the MHC in 80% of significant immune/immune pairs; this compared with 52% for significant mixed pairs (Table S1). Removal of the MHC from these data sets rendered the GPS test statistic

Trait	Abbreviation	Category	No. of cases	No. of controls
Lupus	-	immune	415	360,726
Type 1 diabetes	T1D	immune	583	360,611
Crohn's disease	CD	immune	1,096	360,045
Ulcerative colitis	UC	immune	2,143	359,051
Rheumatoid arthritis	-	immune	4,017	357,124
Eczema/dermatitis	-	immune	9,321	351,820
Hypothyroidism	-	immune	17,574	343,567
Hayfever	-	immune	20,904	70,883
Asthma	-	immune	41,934	319,207
Cardiomyopathy	-	non-immune	479	360,715
Endometriosis	-	non-immune	1,496	359,698
Macular degeneration	MD	non-immune	2,726	115,164
Glaucoma	-	non-immune	5,092	112,798
Leiomyoma	-	non-immune	5,507	355,687
Irritable bowel syndrome	IBS	non-immune	8,537	352,604
Cholelithiasis	-	non-immune	10,520	350,674
Osteoarthritis	-	non-immune	30,046	331,095
Hypercholesterolaemia	-	non-immune	43,957	317,184

Table 2. The selected UK Biobank traits used to evaluate the performance of the GPS test and its comparators.

Table 3. The percentage of disease pairs for which a significant test statistic was obtained, given by group, number of cases, method, and MHC inclusion status. The 'No. of cases' gives the smaller number of disease cases in each pair of case-control GWAS. 'Mixed' pairs were those with one immune and one non-immune disease or two non-immune diseases.

Group	MHC	No. of cases	No. of pairs	GPS (%)	Hoeffding's (%)	LDSC (%)	SumHer (%)
Immune/immune	with MHC	$\leq 2,000$	21	76	14	10	29
Immune/immune	with MHC	$> 2,000$	15	100	40	27	40
Mixed	with MHC	$\leq 2,000$	54	11	9	13	19
Mixed	with MHC	$> 2,000$	63	24	13	29	30
Immune/immune	sans MHC	$\leq 2,000$	21	43	14	10	10
Immune/immune	sans MHC	$> 2,000$	15	60	40	27	40
Mixed	sans MHC	$\leq 2,000$	54	4	4	11	6
Mixed	sans MHC	$> 2,000$	63	10	13	27	30

non-significant for 15 of 25 previously significant immune/immune pairs and all 11 previously significant mixed pairs. For three immune/immune and two mixed pairs, the GPS test statistic became significant only after the MHC was removed.

The impact of MHC removal on the performance of LDSC, SumHer, and Hoeffding's test on immune/immune pairs was far less pronounced. We found that 4 of 12 immune/immune pairs with a significant SumHer test statistic became non-significant with removal of the MHC. Nonetheless, the GPS test still identified more nominally significant pairs among the immune/immune pairs than did its comparators after removal of the MHC.

Discussion

We found the GPS test to be a superior means of identifying genetic similarity between disease traits in the small-sample context. For simulated data, the GPS test's power

was greater than its comparators when the sample size or genetic correlation was small. This greater power was also in evidence when applying the GPS test to real IMD GWAS from the UK Biobank.

60% of disease pairs for which genetic similarity was identified by the GPS test depended upon the inclusion of the MHC. Where the success of pleiotropy-informed methods depends upon the presence of broad genome-wide sharing of genetic effects across multiple loci, the GPS test may be less useful in identifying suitable traits to leverage. Additionally, the presence of extensive LD in the MHC means that LD pruning may not suffice to exclude LD between SNPs from the putatively independent set from which the GPS is computed. A significant test statistic may then arise from the presence of a SNP at which highly significant p-values coincide in each data set, but which is in fact in LD with causal variants distinct for each trait. This problem is compounded by the strength of MHC signals: p-values smaller than 1e-100 are frequently observed there. On the other hand, the discovery of shared MHC signals in early GWAS of immune-mediated diseases did precede later evidence of numerous shared causal variants in larger studies. Although this phenomenon may be particular to the genetic architecture of immune traits, the MHC acting as a sentinel of wider genetic similarity would go some way to vindicate this behaviour of the GPS test, at least in its application to immune-mediated diseases.

The GPS test statistic's construction as a supremum norm seemed to explain its better performance in the large- and mixed-effect regimes compared with the small-effect regime and its sensitivity to MHC signals. The statistic is best maximised by the presence of shared large-effect variants, such as those seen at the MHC, than by the cumulative signals of many smaller-effect variant.

We speculate that pervasive genetic similarity might be better detected with the use of a test of bivariate independence based on a quadratic ecdf statistic, where the distance criterion is the integral of the squared distance between two cdfs. Hoeffding's test uses one such statistic, but its adaptation to incorporate the Anderson-Darling distance's weight function[44] could improve its use in the GWAS context. This weight function emphasises divergence in the tails of the cdfs, that is, it would upweight the contribution of pairwise-dependent p-values where these are small and thus more significant. The GPS test statistic is a weighted supremum norm: the expression in its denominator is the square root of a bivariate function analogous to the univariate Anderson-Darling weight function and this component appears to accomplish just such an upweighting of distances observed at more significant p-values.

As a nonparametric test of bivariate independence, the GPS test does not provide a readily interpretable parameter estimate as does the genetic correlation. Whilst its null hypothesis can be precisely stated in statistical terms, the biological inference to be drawn from rejection of the null is unclear beyond a notion of some sharing of genetic effects between traits. As its test statistic is computed from p-values the GPS test is agnostic as to the direction of genetic effects: a significant test statistic can arise where a negative genetic correlation is observed between two traits, something we observed for cholelithiasis and hayfever, and hypercholesterolaemia and hayfever. Thus, whilst the GPS may be more sensitive in small sample cases, examination of the sign of effect estimates at SNPs with small p-values in both traits can also be informative.

We recommend the GPS test for the detection of shared genetic effects with small-sample data sets. It may be employed to detect genetic relationships between rare diseases and more common, better-understood traits or as a tool for the selection of genetically similar traits for analysis with pleiotropy-informed, cross-phenotype methods, which can be employed to overcome the problem of low power due to small sample size in association studies.

Materials and methods

LD reference panel

Both the GPS test and Hoeffding's test require independent observations from which to compute their test statistic. It was thus necessary to define a set of SNPs in approximate linkage equilibrium; summary statistics from this set can be regarded as independent observations. To serve as a reference data set with which to calculate linkage disequilibrium (LD) in terms of the SNP-SNP genotype squared correlation coefficient r^2, we obtained haplotype data for 503 European individuals from the Phase 3 release of the 1000 Genomes Project (1kGP)[48]. Using the parental identifiers supplied in the `ped` sample metadata file, we removed five individuals with one or both parents present in the cohort (sample identifiers `HG00155`, `NA07048`, `NA10847`, `NA10851`, and `NA12878`) to ensure all members of the cohort were unrelated. We subjected the data to quality control using `PLINK v2.00a2.3LM`[49,50], retaining only those SNPs with a minor allele frequency (MAF) ≥ 0.005, a Hardy-Weinberg equilibrium test p-value $>$ 1e-50, and a missingness rate ≤ 0.1.

We generated separate LD-pruned SNP sets for use with the UKBB data sets and the simulated GWAS data sets. The LD-pruned SNP set for use with the simulated GWAS data sets was generated by first filtering the 1kGP SNPs to retain those with MAF ≤ 0.01 in the European cohort (see 'Simulated GWAS data sets'), then pruning this set. The UKBB LD-pruned set was generated by first merging the 1kGP SNPs with the set of SNPs present in the UKBB data sets, then pruning the intersection of these two sets. For both UKBB and simulated data sets we carried out LD pruning with `PLINK`. The procedure randomly removes one SNP from each pair of highly correlated SNPs and produces an approximately independent set of markers. For the pruning procedure, we specified a 1000kbp window, a step size of 50bp, and a r^2 threshold of 0.2.

UK Biobank GWAS data sets

We downloaded UKBB GWAS summary statistics from the Neale lab's data repository, in particular the 'GWAS round 2' results released in August 2018 (`www.nealelab.is/uk-biobank`).

For analyses in which the MHC was removed from these data sets, we removed all SNPs in the base pair interval 24,000,000-45,000,000 on chromosome 6.

The GPS test and Hoeffding's test

We implemented the GPS test statistic in `C++` (`https://github.com/twillis209/gps_cpp`). We computed it using p-values only from SNPs in the LD-pruned set. To obtain a corresponding p-value for each trait pair, we repeatedly permuted the order of p-values for one trait and computed the GPS statistic each time. We used this permutation procedure to generate 3,000 realisations of the GPS statistic under the null hypothesis. We then fitted a GEV distribution to these null realisations through maximum likelihood estimation with the functions `fgev` and `fitdist` from the `R` packages `evd`[51] and `fitdistrplus`[52], respectively. The p-value was then obtained from the fitted GEV cdf. The permutation procedure requires repeated evaluation of the GPS statistic and we made this computationally feasible through use of a recently published divide-and-conquer algorithm for the estimation and evaluation of multivariate empirical cdfs[38]. We used the implementation of this algorithm from the STochastic OPTimization library [sic][53].

To study the dependence of the GEVD's location, shape, and scale parameters on the number of SNPs, we uniformly downsampled the LD-pruned UKBB SNP set to

obtain smaller subsets on which to evaluate the GPS statistic.

We evaluated Hoeffding's test statistic using the `hoeffding.D.test` function from the `R` package `independence`[54]. As with the GPS test statistic, Hoeffding's test statistic was evaluated using LD-pruned SNP sets.

Simulated GWAS data sets

We simulated GWAS summary statistics using the `R` package `simGWAS v0.2.0-4`[55]. Following the package vignette, we used haplotypes from European individuals from the Phase 3 1kGP data referenced above and included only those variants with MAF ≥ 0.01. We partitioned the genome into approximately independent LD blocks using `ldetect`[56] and simulated summary statistics in a blockwise manner in order to reduce the computational burden of simulation. We simulated at most one causal variant per LD block and selected, arbitrarily, the 2000th SNP in each LD block as the causal variant. We ran simulations under five regimes (Table 1). We also varied the sample size: we simulated GWAS with 500, 1,000, 2,000, 5,000, 10,000, 50,000, 100,000, and 250,000 cases. The number of controls was kept to a minimum of 10,000 and for simulations with more than 10,000 cases, the number of controls was increased to match the number of cases.

To induce a particular genetic correlation, we varied the number of causal variants shared between a pair of simulated GWAS. The number of causal variants per GWAS was kept constant in order to keep heritability per simulated trait constant. We simulated 400 replicate GWAS data sets for each sample size and genetic correlation, giving 200 replicate pairs of GWAS for each sample size-correlation pair. For a given replicate pair, sample size, and specified correlation, the identity of the LD blocks which were to contain causal variants was randomly chosen.

LD score regression

We estimated the genetic correlation using the LDSC software as directed in the LDSC GitHub repository's wiki (`https://github.com/bulik/ldsc/wiki`). When estimating the heritability from simulated summary statistics, we constrained the intercept to one as the simGWAS simulation model does not admit confounding. We used precomputed LD scores provided for download via the repository wiki (`https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2`).

SumHer

We downloaded the prebuilt LDAK binaries, version 5.2, from the package website (`https://dougspeed.com/ldak/`). We generated SNP taggings using Phase 3 1kGP data from European individuals and the 'LDAK-Thin' model as described on the LDAK website (`https://dougspeed.com/calculate-taggings/`). We estimated genetic correlation with the `SumHer` component of the LDAK package and the default 'SumHer-GC' model, which assumes that confounding inflation is multiplicative. The SumHer-GC model assumes the LDAK heritability model, not the LDAK-Thin model; the authors nonetheless recommend use of the latter owing to its simplicity and their observation that genetic correlation estimates are generally insensitive to the choice of heritability model (`http://dougspeed.com/genetic-correlations/`). We excluded SNPs which explained more than 1% of the phenotypic variance as per the authors' recommendation as analyses can be sensitive to large-effect loci. `SumHer` does not provide for hypothesis testing of the genetic correlation. We performed a chi-squared test against a null hypothesis of no genetic correlation (i.e. $H_0 : r_g = 0$) by dividing the genetic correlation estimate by its standard error, then squaring this quantity to obtain

a test statistic. We used the chi-squared survival function `chi2.sf` from the `scipy.stats` Python package to perform the test.

Code availability

Our implementation of the GPS test statistic in `C++` is available on GitHub (`https://github.com/twillis209/gps_cpp`). We have also made available a `snakemake`-based pipeline[57] for the reproduction of our results (`https://github.com/twillis209/gps_paper_pipeline`). This pipeline is intended to provide for an end-to-end reproduction, including the download of 1kGP reference data and UKBB GWAS data sets, the generation of our simulated data sets with `simGWAS`, and the analysis of these real and simulated data sets with the tests studied above.

Acknowledgments

We would like to thank Dr Xavier Warin for timely assistance with the use of the Stochastic Optimisation library `StOpt`[53]. We would also like to thank our colleague Dr Guillermo Reales for his curation of some of the GWAS data sets used in this work and creation of the `GWAS_tools` pipeline.

We wish to acknowledge all GWAS participants, in particular those of the UK Biobank, for their contribution to the data used herein. We also acknowledge the investigators who carried out these GWAS and made their summary statistics publicly available.

References

1. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant Pleiotropy in Human Complex Diseases and Traits. The American Journal of Human Genetics. 2011;89(5):607–618. doi:10.1016/j.ajhg.2011.10.004.

2. Drögemüller C, Karlsson EK, Hytönen MK, Perloski M, Dolf G, Sainio K, et al. A Mutation in Hairless Dogs Implicates FOXI3 in Ectodermal Development. Science. 2008;doi:10.1126/science.1162525.

3. Strain GM. Deafness in Blue-Eyed White Cats: The Uphill Road to Solving Polygenic Disorders. The Veterinary Journal. 2007;173(3):471–472. doi:10.1016/j.tvjl.2006.01.015.

4. Darwin C. On the Origin of Species, 1859; 2016.

5. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in Complex Traits: Challenges and Strategies. Nature Reviews Genetics. 2013;14(7):483–495. doi:10.1038/nrg3461.

6. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLOS Genetics. 2011;7(8):e1002254. doi:10.1371/journal.pgen.1002254.

7. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. Nature. 2009;460(7256):748–752. doi:10.1038/nature08185.

8. Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, et al. Genetic Relationship between Five Psychiatric Disorders Estimated from Genome-Wide SNPs. Nature Genetics. 2013;45(9):984–994. doi:10.1038/ng.2711.

9. Hazel LN. The Genetic Basis for Constructing Selection Indexes. Genetics. 1943;28(6):476–490.

10. NEI M. College of Agriculture. Plant Breeding. 1960;82(0).

11. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic Correlations of Polygenic Disease Traits: From Theory to Practice. Nature Reviews Genetics. 2019;20(10):567–581. doi:10.1038/s41576-019-0137-z.

12. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs Explain a Large Proportion of the Heritability for Human Height. Nature Genetics. 2010;42(7):565–569. doi:10.1038/ng.608.

13. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics. 2011;88(1):76–82. doi:10.1016/j.ajhg.2010.11.011.

14. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of Pleiotropy between Complex Diseases Using Single-Nucleotide Polymorphism-Derived Genomic Relationships and Restricted Maximum Likelihood. Bioinformatics. 2012;28(19):2540–2542. doi:10.1093/bioinformatics/bts474.

15. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. Nature Genetics. 2015;47(3):291–295. doi:10.1038/ng.3211.

16. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An Atlas of Genetic Correlations across Human Diseases and Traits. Nature Genetics. 2015;47(11):1236–1241. doi:10.1038/ng.3406.

17. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, et al. LD Hub: A Centralized Database and Web Interface to Perform LD Score Regression That Maximizes the Potential of Summary Level GWAS Data for SNP Heritability and Genetic Correlation Analysis. Bioinformatics. 2017;33(2):272–279. doi:10.1093/bioinformatics/btw613.

18. Neale lab. UK Biobank; 2018. http://www.nealelab.is/uk-biobank.

19. Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, et al. A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits. Nature Genetics. 2019;51(9):1339–1348. doi:10.1038/s41588-019-0481-0.

20. Speed D, Hemani G, Johnson MR, Balding DJ. Improved Heritability Estimation from Genome-wide SNPs. The American Journal of Human Genetics. 2012;91(6):1011–1021. doi:10.1016/j.ajhg.2012.10.010.

21. Speed D, Balding DJ. SumHer Better Estimates the SNP Heritability of Complex Traits from Summary Statistics. Nature Genetics. 2019;51(2):277–284. doi:10.1038/s41588-018-0279-5.

22. Falconer D, Mackay T. Introduction to Quantitative Genetics. Longmans Green, Harlow, Essex, UK. Introduction to quantitative genetics 4th ed Longmans Green, Harlow, Essex, UK. 1996;.

23. Wray NR, Lee SH, Mehta D, Vinkhuyzen AAE, Dudbridge F, Middeldorp CM. Research Review: Polygenic Methods and Their Application to Psychiatric Traits. Journal of Child Psychology and Psychiatry. 2014;55(10):1068–1087. doi:10.1111/jcpp.12295.

24. Ni G, Moser G, Wray NR, Lee SH. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. American Journal of Human Genetics. 2018;102(6):1185–1194. doi:10.1016/j.ajhg.2018.03.021.

25. Thaventhiran JED, Lango Allen H, Burren OS, Rae W, Greene D, Staples E, et al. Whole-Genome Sequencing of a Sporadic Primary Immunodeficiency Cohort. Nature. 2020;583(7814):90–95. doi:10.1038/s41586-020-2265-1.

26. Shillitoe B, Bangs C, Guzman D, Gennery AR, Longhurst HJ, Slatter M, et al. The United Kingdom Primary Immune Deficiency (UKPID) Registry 2012 to 2017. Clinical and Experimental Immunology. 2018;192(3):284–291. doi:10.1111/cei.13125.

27. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779.

28. Rees F, Doherty M, Grainge M, Davenport G, Lanyon P, Zhang W. The Incidence and Prevalence of Systemic Lupus Erythematosus in the UK, 1999–2012. Annals of the Rheumatic Diseases. 2016;75(1):136–141. doi:10.1136/annrheumdis-2014-206334.

29. Saha S, Chant D, Welham J, McGrath J. A Systematic Review of the Prevalence of Schizophrenia. PLOS Medicine. 2005;2(5):e141. doi:10.1371/journal.pmed.0020141.

30. Andreassen OA, Thompson WK, Schork AJ, Ripke S, Mattingsdal M, Kelsoe JR, et al. Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate. PLoS Genetics. 2013;9(4). doi:10.1371/journal.pgen.1003455.

31. Vuckovic D, Gasparini P, Soranzo N, Iotchkova V. MultiMeta: An R Package for Meta-Analyzing Multi-Phenotype Genome-Wide Association Studies. Bioinformatics. 2015;31(16):2754–2756. doi:10.1093/bioinformatics/btv222.

32. Liu J, Wan X, Ma S, Yang C. EPS: An Empirical Bayes Approach to Integrating Pleiotropy and Tissue-Specific Information for Prioritizing Risk Genes. Bioinformatics. 2016;32(12):1856–1864. doi:10.1093/bioinformatics/btw081.

33. Li X, Zhu X. Cross-Phenotype Association Analysis Using Summary Statistics from GWAS. Methods in molecular biology (Clifton, NJ). 2017;1666:455–467. doi:10.1007/978-1-4939-7274-6_22.

34. Qi G, Chatterjee N. Heritability Informed Power Optimization (HIPO) Leads to Enhanced Detection of Genetic Associations across Multiple Traits. PLOS Genetics. 2018;14(10):e1007549. doi:10.1371/journal.pgen.1007549.

35. Ray D, Boehnke M. Methods for Meta-Analysis of Multiple Traits Using GWAS Summary Statistics. Genetic Epidemiology. 2018;42(2):134–145. doi:10.1002/gepi.22105.

36. Liley J, Wallace C. Accurate Error Control in High-Dimensional Association Testing Using Conditional False Discovery Rates. Biometrical Journal. 2021;63(5):1096–1130. doi:10.1002/bimj.201900254.

37. Li YR, Li J, Zhao SD, Bradfield JP, Mentch FD, Maggadottir SM, et al. Meta-Analysis of Shared Genetic Architecture across Ten Pediatric Autoimmune Diseases. Nature medicine. 2015;21(9):1018–1027. doi:10.1038/nm.3933.

38. Langrené N, Warin X. Fast Multivariate Empirical Cumulative Distribution Function with Connection to Kernel Density Estimation. arXiv:200503246 [cs, stat]. 2020;.

39. Zhernakova A, van Diemen CC, Wijmenga C. Detecting Shared Pathogenesis from the Shared Genetics of Immune-Related Diseases. Nature Reviews Genetics. 2009;10(1):43–55. doi:10.1038/nrg2489.

40. Cotsapas C, Hafler DA. Immune-Mediated Disease Genetics: The Shared Basis of Pathogenesis. Trends in Immunology. 2013;34(1):22–26. doi:10.1016/j.it.2012.09.001.

41. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of Five Chronic Inflammatory Diseases Identifies 27 New Associations and Highlights Disease-Specific Patterns at Shared Loci. Nature Genetics. 2016;48(5):510–518. doi:10.1038/ng.3528.

42. Prince C, Mitchell RE, Richardson TG. Integrative Multiomics Analysis Highlights Immune-Cell Regulatory Mechanisms and Shared Genetic Architecture for 14 Immune-Associated Diseases and Cancer Outcomes. The American Journal of Human Genetics. 2021;doi:10.1016/j.ajhg.2021.10.003.

43. Kennedy AE, Ozbek U, Dorak MT. What Has GWAS Done for HLA and Disease Associations? International Journal of Immunogenetics. 2017;44(5):195–211. doi:10.1111/iji.12332.

44. Anderson TW, Darling DA. A Test of Goodness of Fit. Journal of the American Statistical Association. 1954;49(268):765–769. doi:10.1080/01621459.1954.10501232.

45. Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP Heritability in Complex Human Traits. Nature Genetics. 2017;49(7):986–992. doi:10.1038/ng.3865.

46. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics. Nature genetics. 2015;47(11):1228–1235. doi:10.1038/ng.3404.

47. Speed D, Holmes J, Balding DJ. Evaluating and Improving Heritability Models Using Summary Statistics. Nature Genetics. 2020;52(4):458–462. doi:10.1038/s41588-020-0600-y.

48. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A Global Reference for Human Genetic Variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393.

49. Purcell S, Chang C. PLINK 2.0;.

50. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. GigaScience. 2015;4(1):s13742–015–0047–8. doi:10.1186/s13742-015-0047-8.

51. Stephenson AG. Evd: Extreme Value Distributions. R News. 2002;2(2):0.

52. Delignette-Muller ML, Dutang C. fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software. 2015;64(4):1–34. doi:10.18637/jss.v064.i04.

53. Gevret H, Langren'e N, Lelong J, Warin X, Maheshwari A. STochastic OPTimization Library in C++; 2018.

54. Even-Zohar C. Independence: Fast Rank Tests. arXiv:201009712 [math, stat]. 2020;.

55. Fortune MD, Wallace C. simGWAS: A Fast Method for Simulation of Large Scale Case–Control GWAS Summary Statistics. Bioinformatics. 2019;35(11):1901–1906. doi:10.1093/bioinformatics/bty898.

56. Berisa T, Pickrell JK. Approximately Independent Linkage Disequilibrium Blocks in Human Populations. Bioinformatics. 2016;32(2):283–285. doi:10.1093/bioinformatics/btv546.

57. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable Data Analysis with Snakemake. 2021;doi:10.12688/f1000research.29032.2.

Supporting information

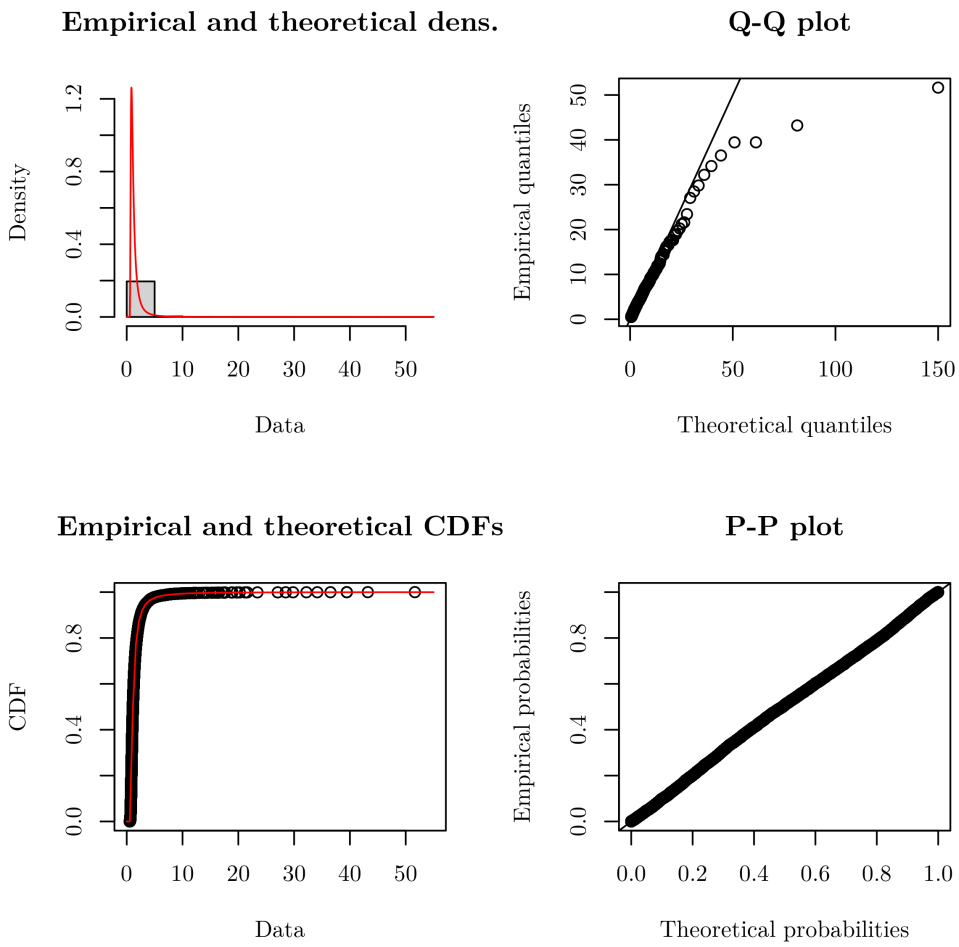

Fig S1. The goodness of fit of the generalised extreme value distribution (GEVD) to the GPS test statistic under the null hypothesis. Clockwise from the top left-hand plot: (1) a histogram of GPS statistics with fitted GEVD density superimposed as the red line, (2) a quantile-quantile plot of empirical and fitted GEVD ('theoretical') quantiles, (3) a percentile-percentile plot of the same data as (2), and (4) the empirical GPS stastistic cdf with fitted GEVD cdf superimposed as the red line. Null realisations of the GPS statistic were produced by permutation of the order of SNPs in GWAS of asthma and emphysema/chronic bronchitis from the UK Biobank collection discussed in Materials and methods.

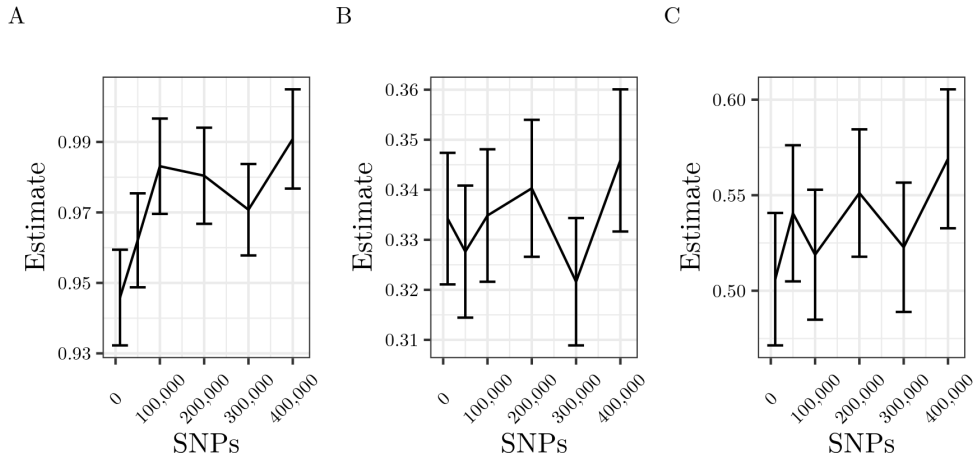

Fig S2. The dependence of estimates of the three generalised extreme value distribution parameters on the number of SNPs. (A) The location parameter. (B) The scale parameter. (C) The shape parameter. The error bars depict 95% confidence intervals for each estimate. The estimates were obtained by permuting the order of SNPs in a pair of GWAS summary statistics data sets and downsampling these same data sets to obtain the specified number of SNPs. The estimates shown were obtained using GWAS of asthma and emphysema/chronic bronchitis in the UKBB data set.

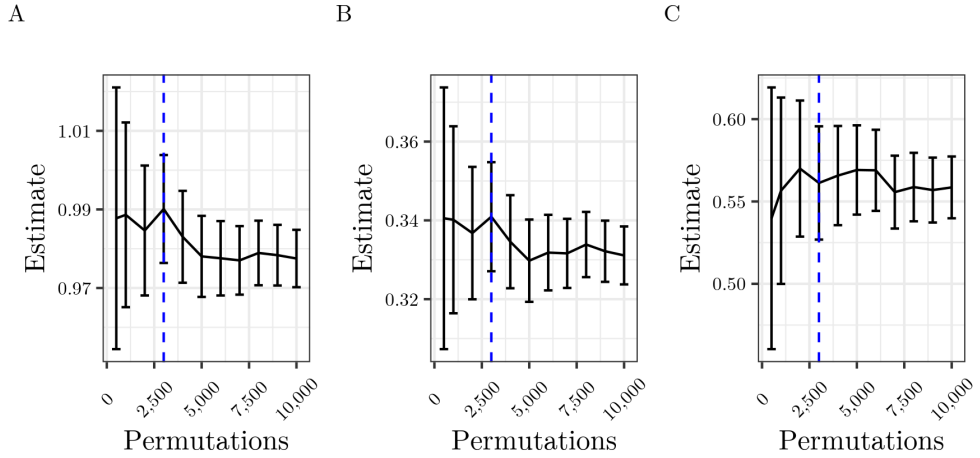

Fig S3. The dependence of estimates of the three generalised extreme value distribution parameters on the number of permutations n used to fit them. (A) The location parameter. (B) The scale parameter. (C) The shape parameter. Error bars depict 95% confidence intervals for each estimate. The dashed blue line indicates 3,000 permutations. The estimates were obtained by permuting the order of SNPs in a pair of GWAS summary statistic data sets. The estimates shown were obtained using GWAS of asthma and emphysema/chronic bronchitis.

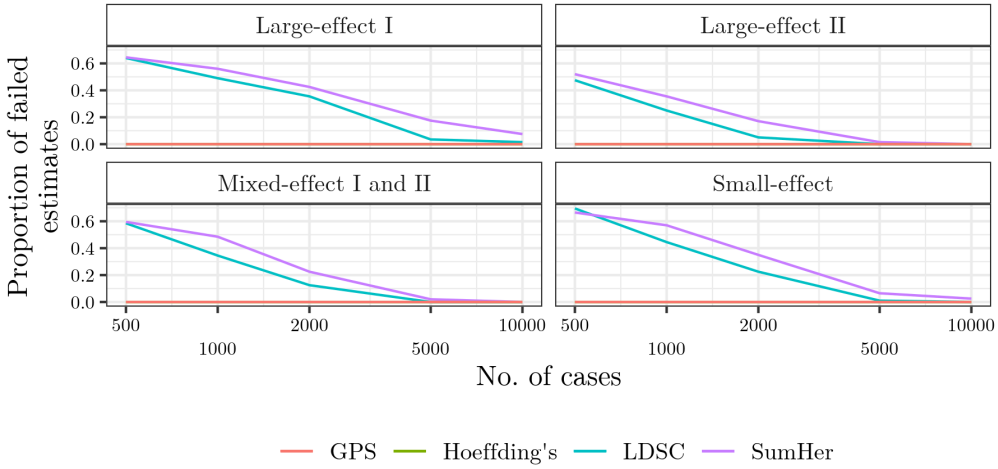

Fig S4. The proportion of simulated data sets for which a test statistic could not be obtained.

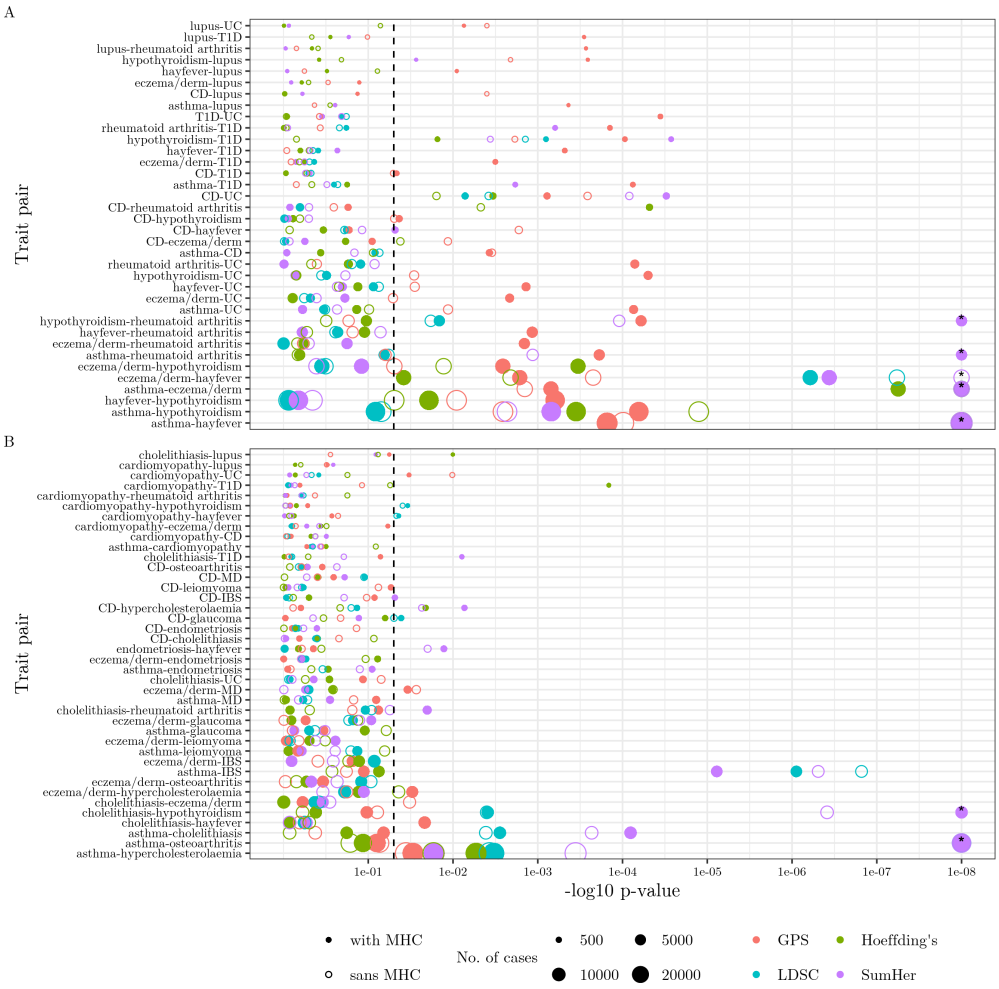

Fig S5. The results of significance testing of genetic similarity in UK Biobank GWAS data sets. (A) Testing pairs of immune-mediated diseases. (B) Testing pairs comprised of one immune- and one non-immune-mediated disease. The 'No. of cases' gives in terms of point size the smaller number of disease cases in each pair of case-control GWAS. The dashed line lies at 0.05. Filled points indicate the use of the entire data set, hollow points indicate the use of a data set with the MHC removed. Where p-values were smaller than 1e-8, they have been aliased to 1e-8 and indicated with an asterisk. For very small sample sizes, it was not possible to obtain genetic correlation estimates with LDSC and SumHer. (B) depicts 40 of the 81 immune-/non-immune disease pairs (Table 2)

.

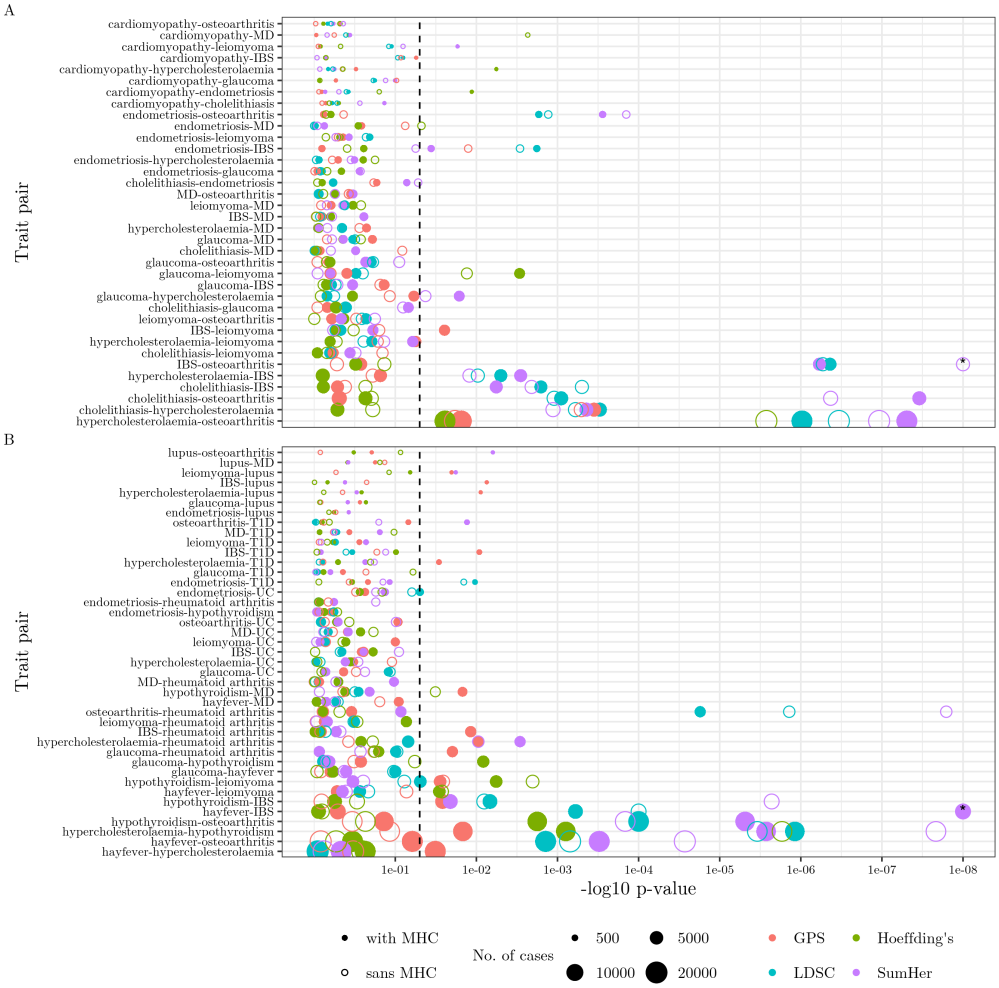

Fig S6. The results of significance testing of genetic similarity in the remaining pairs of selected UK Biobank GWAS data sets. (A) depicts results for all non-immune/non-immune pairs and (B) depicts the remaining 41 immune/non-immune pairs not depicted in Fig S5. The 'No. of cases' gives in terms of point size the smaller number of disease cases in each pair of case-control GWAS. The dashed line lies at 0.05. Filled points indicate the use of the entire data set, hollow points indicate the use of a data set with the MHC removed. Where p-values were smaller than 1e-8, they have been aliased to 1e-8 and indicated with an asterisk. For very small sample sizes, it was not possible to obtain genetic correlation estimates with LDSC and SumHer.

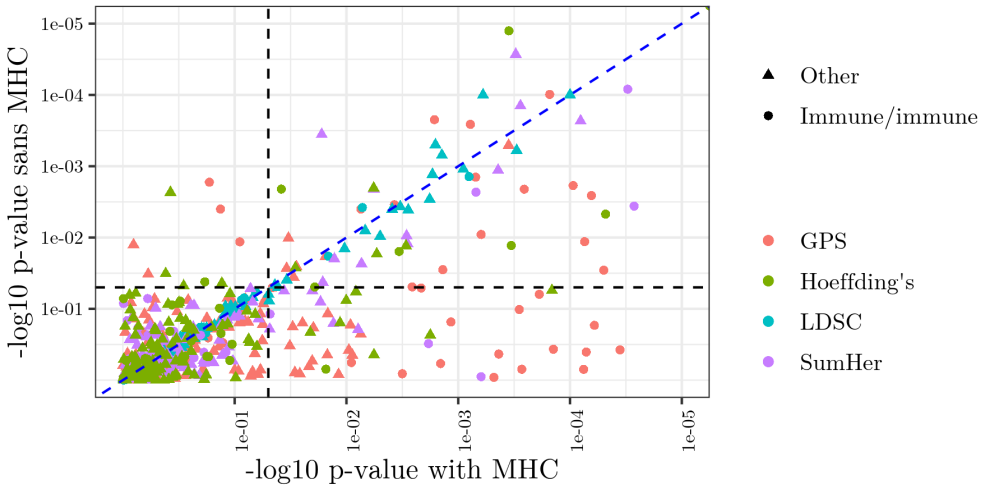

Fig S7. The effect of removing the MHC from UKBB data sets on test p-values. The black dashed lines lie at 0.05 on each axis. Circles indicate points corresponding to immune/immune pairs and triangles immune/non-immune or non-immune/non-immune pairs. The data are constrained to the range $[1e-6, 1]$ to more clearly depict the trend for significant GPS test p-values; a number of smaller-valued points for LDSC, SumHer, and Hoeffding's have been omitted.

Table S1. The effect of the removal of the major histocompatibility complex (MHC) from UKBB data sets on the number of significant GPS test statistics when the supremum was located in the MHC. 'Mixed' pairs were those with one immune and one non-immune disease or two non-immune diseases.

	Immune	Mixed
No. of pairs	36	117
No. of significant pairs	31	21
No. of significant pairs with MHC supremum	25	11
No. of significant pairs with MHC supremum still significant after MHC removal	10	0