Linkage Disequilibrium Score Statistic Regression for Identifying Novel Trait Associations

Rowland W. Pettit[1] · Christopher I. Amos[1,2,3]

Abstract

Purpose of Review In this paper, we review the linkage disequilibrium score regression (LDSR) statistical method for computing correlation and shared heritability between complex polygenetic traits.

Recent Findings Applications of the LDSR method have been prominent and wide ranging, building off the abundance of GWAS summary statistic data available both publicly and privately. LDSR has provided insight into the shared genetic architecture of many complex diseases, including anxiety disorders, psychiatric disorders, neurologic disorders, and many cancers including lung, breast, and pancreatic cancer. A recent HDL method development complements LDSR for more precise genetic correlation that estimates utilizing summary statistics with smaller sample sizes.

Summary Genetic correlation analysis provides a powerful tool for rapidly studying many potential genetically influenced factors to identify novel host factors that influence disease development. The LDSR method is a useful and effective approach for determining the genetic correlation between traits. The method is computationally efficient and robust to some sources of spurious association.

The Need for Genetic Correlation

Many human conditions and diseases are not single gene disorders but rather occur because of multifactorial biological processes [1]. For Mendelian-inherited diseases, understanding mutations that occur in a specific genetic sequence can fully explain the phenotypes observed in a human condition [2]. Examples of Mendelian-inherited diseases include sickle cell hemoglobinopathies [3], cystic fibrosis [4], and Duchenne muscular dystrophy [5]. Such direct mutation to phenotype relationships have made Mendelian disease processes easiest to diagnose, study, and develop treatments for their mitigation [6]. Such benefits are not afforded to polygenetic diseases, where multiple genetic regions are implicated in disease state progression [1]. The question arises—how do you study complex polygenetic diseases like lung cancer, bipolar disorder, or schizophrenia? Or, more broadly, what is the best way to understand complex human traits such intelligence, BMI, or genetically derived contributions to physical fitness? When a phenotype develops because of hundreds, or thousands, of small genetic effects scattered across the genome, it becomes challenging to investigate.

Approaches to studying polygenic traits have evolved [1]. Historically, multifactorial traits were studied with randomized controlled trials (RCT) and longitudinal studies [7]. In these efforts, cohorts were identified based on an exposure and followed for a duration of time to observe outcome differences between the cohorts that may arise. This forward in time analysis allowed for causal inferences to be made to help understand what exposures, or inherited traits, may cause a phenotype outcome of interest [8]. The longest active such studies is the

Framingham Heart study [9] which has been ongoing since 1948. The study has been a success and has led insights on the effect of diet, exercise, and medications on the development of hypertensive and arteriosclerotic heart disease [10]. In total, however, the study has cost $50 million and has required dedicated research personal oversight throughout its duration [11]. Cost and duration limitations prevent prospective RCT or longitudinal studies from being universally applicable, despite their utility. As a compromise, study methods have been developed to observe exposures and outcomes at a single time point, allowing for associative insights to be gained [7].
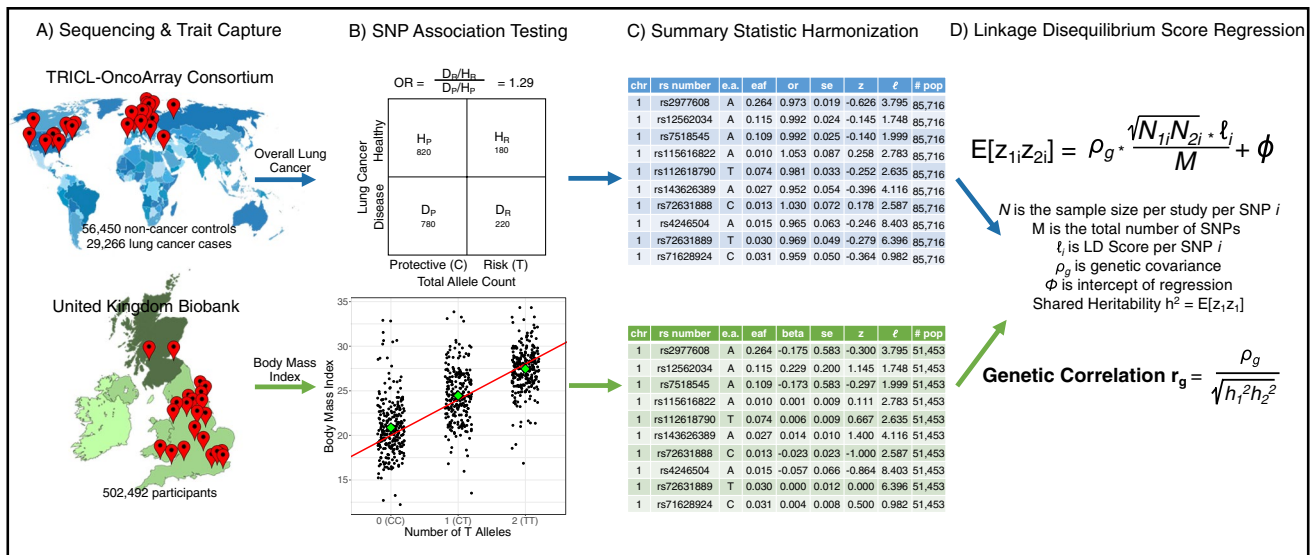
In 2000, the limitations of observational studies changed with the sequencing of the first human genome [12]. With increasing access to sequencing, a new concept could be applied to observational studies. As an individual's genotypes are set at conception, by also annotating genotype, a proxy of time could be reintroduced to observational studies. Studying the genetic differences between individuals and their trait outcomes allows for concurrent sampling, with an understanding that "genetics came first" [13]. To clarify, if genetic regions were associated with an outcome, then you could be confident that it was not the outcome that led to those genotypes being present, avoiding reverse causation biases [14]. The end-to-end total base pair count for the 23 human chromosomes measures nearly 3 billion [15]. Despite year over year decreasing costs, full genome sequence-based association analysis remains untenable computationally and economically [16]. Instead, single nucleotide polymorphism genetic annotation or "SNP" genotyping occurs, where individuals are strategically sampled at ½–10 million base pair positions. The SNP are selected to serve as tags or proxies that provide maximal information about nearby genetic variations. A single nucleotide polymorphism is a journaled substitution that occurs at a specific base pair in the genetic sequence [17]. With SNP genotyping, it became possible to measure effects occurring at millions of SNP sites [18] on the development of traits or human disease states with genome-wide association analysis (GWAS) [19]. Because of the low cost and high reliability of this clever analytical strategy, large numbers of samples could be queried, so that even small effects on risk could be identified.

While discovery of new variants influencing disease risk or trait variability has been a boon for understanding the genetic architecture of thousands of conditions, the availability of these scans has also supported novel strategies for evaluating similarities in the genetic architectures for multiple conditions. Conditions and traits could be genetically compared at genome-wide SNP sites for a comprehensive comparison of the shared genetic factors that may exist between two phenotypes. This concept of looking for shared trait genetics is known as "genetic correlation" and can be statistically measured to assess overall genetic similarity or differences between traits [20]. This method

has proven highly fruitful for identifying sets of genetically similar conditions that may have similar causes. New metrics including "shared heritability" [21] of traits and comparing genetic predisposition for trait development could be obtained through genetic correlation associative analysis [1]. Multiple methods for genetic correlations have been developed, including ones that look at individual level data, as well as ones that work on aggregated data or the "summary statistic" outputs from GWAS analysis [22]. In this paper, I will introduce the linkage disequilibrium score regression (LDSR) method [23••, 24••] to measure shared heritability and genetic correlation between traits. I will discuss its statistical framework, advantages and disadvantages, major applications, and emerging methods resulting as a consequence of LDSR in the field of genetic epidemiology.

Genome-Wide Association Studies The LDSR method uses GWAS summary statistics as input data [25]. In a GWAS [26], genetic variants across the whole genome are tested for association with a trait outcome of interest. The resulting per SNP effect estimates after GWAS analysis are reported and can be converted into SNP z-scores which we will be used in LDSR [27]. GWAS can be conducted on any reliably measured trait for which SNP sequencing has occurred. The trait studied can be binary, such as the "yes/no" outcome of lifetime development of lung adenocarcinoma. Traits studied with GWAS can also be quantitative, on a continuous scale, such as for an individuals measured body mass index (BMI). How association is tested for across the genome at each variant site changes given the type of the outcome. However, in general, an association test is performed at each SNP location across all participants in the GWAS, with and without the phenotype. If 11 million SNP are included in the GWAS, then an association testing is conducted 11 million times, once at each SNP variant, to determine if the allele ratios observed correlate with the expression of the studied phenotype. The output of these 11 million association tests is reported as the summary statistics of a GWAS.

In Fig. 1A–C, we present via visual abstract the method of obtaining z-scores for SNP measured for different traits. As a concrete example, we have included two large and publicly available GWAS consortiums and follow them throughout as an exemplary analysis. To demonstrate LDSR between two traits, we will introduce the TRICL-OncoArray Consortium [28, 29], which is the largest GWAS meta-analysis of lung cancer outcomes available to date. TRICL-OncoArray conducted a GWAS on 29,266 lung cancers cases and 56,450 non-cancer controls to identify SNP of interest related to lung oncogenesis. Additionally, we introduce and will follow traits from the United Kingdom Biobank [30] (UKB), which obtained 500,000 + volunteers at 22 sites in the UK, measuring 400 + physical, biomarker, and lab-related "traits" and conducting GWAS for each to determine genetic regions

A) Sequencing & Trait Capture

TRICL-OncoArray Consortium

56,450 non-cancer controls
29,266 lung cancer cases

Overall Lung Cancer

United Kingdom Biobank

502,492 participants

Body Mass Index

B) SNP Association Testing

$$OR = \frac{D_R/H_R}{D_P/H_P} = 1.29$$

	Protective (C)	Risk (T)
Healthy	H_P 820	H_R 180
Disease	D_P 780	D_R 220

Total Allele Count

Body Mass Index vs Number of T Alleles: 0 (CC), 1 (CT), 2 (TT)

C) Summary Statistic Harmonization

chr	rs number	e.a.	eaf	or	se	z	ℓ	# pop
1	rs2977608	A	0.264	0.973	0.019	-0.626	3.795	85,716
1	rs12562034	A	0.115	0.992	0.024	-0.145	1.748	85,716
1	rs7518545	A	0.109	0.992	0.025	-0.140	1.999	85,716
1	rs115616822	A	0.010	1.053	0.087	0.258	2.783	85,716
1	rs112618790	T	0.074	0.981	0.033	-0.252	2.635	85,716
1	rs143626389	A	0.027	0.952	0.054	-0.396	4.116	85,716
1	rs72631888	C	0.013	1.030	0.072	0.178	2.587	85,716
1	rs4246504	A	0.015	0.965	0.063	-0.246	8.403	85,716
1	rs72631889	T	0.030	0.969	0.049	-0.279	6.396	85,716
1	rs71628924	C	0.031	0.959	0.050	-0.364	0.982	85,716

chr	rs number	e.a.	eaf	beta	se	z	ℓ	# pop
1	rs2977608	A	0.264	-0.175	0.583	-0.300	3.795	51,453
1	rs12562034	A	0.115	0.229	0.200	1.145	1.748	51,453
1	rs7518545	A	0.109	-0.173	0.583	-0.297	1.999	51,453
1	rs115616822	A	0.010	0.001	0.009	0.111	2.783	51,453
1	rs112618790	T	0.074	0.006	0.009	0.667	2.635	51,453
1	rs143626389	A	0.027	0.014	0.010	1.400	4.116	51,453
1	rs72631888	C	0.013	-0.023	0.023	-1.000	2.587	51,453
1	rs4246504	A	0.015	-0.057	0.066	-0.864	8.403	51,453
1	rs72631889	T	0.030	0.000	0.012	0.000	6.396	51,453
1	rs71628924	C	0.031	0.004	0.008	0.500	0.982	51,453

D) Linkage Disequilibrium Score Regression

$$E[z_{1i}z_{2i}] = \rho_g * \frac{\sqrt{N_{1i}N_{2i}} * \ell_i}{M} + \phi$$

N is the sample size per study per SNP i
M is the total number of SNPs
ℓ_i is LD Score per SNP i
ρ_g is genetic covariance
Φ is intercept of regression
Shared Heritability $h^2 = E[z_1z_1]$

$$\text{Genetic Correlation } r_g = \frac{\rho_g}{\sqrt{h_1^2 h_2^2}}$$

Fig. 1 A graphical abstract annotating the workflow for conducting linkage disequilibrium score regression. In this figure, we include two exemplary consortia data, from the TRICL-OncoArray consortium and the UKB. The actual values in parts B and C provided are modified for teaching purposes

of interest. In our figure example, we will investigate the genetic correlation between the binary trait of "overall" lung cancer from TRICL-Oncoarray and the continuous trait of BMI from UKB.

SNP Association Testing In Fig. 1B, we define how association tests differ regarding the outcome of interest. At any given location, a diallelic SNP can take three different combinations: homozygous recessive, heterozygous, or homozygous dominant. When conducting GWAS on a binary outcome trait, an odds ratio is calculated at each SNP, measuring the odds of the trait outcome being present based on the number of alleles each study participant has given the trait outcome. Afterwards, one can view the summary statistic report from GWAS and quickly ascertain the odds of disease per SNP based on the number of a risk alleles that are present. A different approach is taken when performing SNP association tests for quantitative traits. Rather than conducting an odds ratio, with quantitative traits, a regression is performed. The number of the "effect" or "risk" alleles is placed on the x axis ("CC," "CT," "TT"), and the quantitative trait on the y axis, and a regression is performed. The beta value of the regression is reported as the per SNP effect estimate in GWAS summary statistics. Odds ratio and beta effect estimates can then be normalized to a directly comparable "z-score." For binary traits or diseases, the log of the odds ratio, divided by the standard error of the estimate, is the corresponding per SNP z-score [27]. For quantitative traits, a z-score is obtained simply by dividing the beta effect estimate by the standard error [27]. These z-scores

ultimately will be multiplied across traits and will serve as the "dependent variable" in LDSR.

Linkage Disequilibrium Scoring If the product of SNP z-scores across traits is the dependent variable in LDSR, what is the independent? The answer is the per SNP linkage disequilibrium score, or LD score. Genetic "linkage" is a term used to describe inheritance patterns of genetic sequences [31] on a chromosome during gamete formation or "gametogeneis" [32]. Linkage is the consequence of meiosis during cell division, where sequences of DNA that are physically close to each other on a chromosome are more likely to be inherited together in the offspring [18]. Specifically, during prophase 1 of gametogenesis, "crossing over" occurs to introduce genetic variation in the offspring [33]. This is where sister chromosomes align and switch parallel genetic segments with each other, known as recombination [34]. The functional consequence of recombination is that genetic regions that are close in proximity to each other are less likely to be separated during recombination, and are thus "linked." Genetic linkage can be quantified at each base pair position, describing this probability of separation of a SNP from its neighbor during crossing over as a "linkage score" [35]. A linkage score of "0" indicates "linkage equilibrium," implying statistical independence of an allele from a certain neighbor on the same chromosome [36]. When two loci are closely located, the probability that crossover will occur between alleles is low. Common causes for association of alleles at two different loci include bottlenecks, where limited numbers of individuals propagated communities,

migration among communities, and population expansion [37]. Population demographics have resulted in genome-wide associations within human populations. Because SNP loci that are closely located are unlikely to experience crossovers, these associations have persisted. Moreover, crossover does not occur randomly in the genome but is facilitated by PRDM9 [38], a zinc finger protein that recognizes specific chromosomal positions. Because crossover is non-random, the human genome comprises sets of correlated alleles at different SNP that are called haplotype blocks. Interestingly, the sequences that PRDM9 recognizes for crossover vary between African and non-African populations [39]. Because of the older origins in African-descent populations and differences in sites for recombination, haplotype blocks are smaller and may occur at different locations that are present in European-American descent populations [38]. Therefore, studying correlations in traits requires that African-descent and other populations are separately considered.

To assess overall linkage of a SNP with its neighbors, the LD score for a position can be calculated as a cumulative association metric indicating that alleles in the neighboring genomic sequence are inherited together more often in a study population than would be expected due to crossing over driven independence [36]. We will use LD score in the statistical measurement of genetic correlation between traits. Specifically, we will use LD score per SNP as the independent variable in a regression model. In Supplementary Fig. 1, we provide the full derivation for how to calculate the linkage disequilibrium metric (r^2) between a SNP and one of its neighbors. In brief, calculating the LD r^2 value between two SNP occurs by looking at the allele frequencies and the haplotype frequencies at each SNP and seeing if a difference exists in their observed values other than what would be expected based on equilibrium level crossing over events [36]. To calculate a total LD score for a given SNP, then, one simply sums up all LD r^2 values for the SNP with each of its neighbors in a predefined sampling "window" on the chromosome. Ultimately, we will use this LD score (sum of all neighbor LD r^2 values) as the dependent variable in a regression. The LD score, acting as the independent variable, serves to annotate and account for genetic homogeneity that exists only due to a lack of recombination (high LD score), and not due to significant association with a studied trait or outcome of interest [24••].

Summary Statistic Harmonization Ultimately, the goal of LDSR is to measure a genetic correlation between two traits, which necessitates bringing together two, often independent, GWAS summary statistic files for joint analysis. To do this, "harmonization" steps must be conducted [27]. We have already mentioned the calculation of z-scores from odds ratios or beta effect estimates, to have the same unit of comparison. An important next consideration includes

identifying the reference and effect alleles utilized per GWAS. It is common for a different reference allele to be used, leading to inversed effect estimates that will need to be corrected. Linkage between SNP can be exploited to maximize information available for LDSR method affording for the imputation of additional SNP for analysis. The long-term consequence of linkage disequilibrium evolutionarily is that genetic regions, also known as haplotype blocks, are conserved within populations [40]. It is possible to use haplotype and genetic reference information to safely impute base pair information around directly observed SNP instances [41]. SNP imputation is routinely conducted in genetic correlation analysis [42, 43]. During harmonization, it is a best practice [24••, 44] to identify an imputation quality threshold, such as conducted with an INFO score, and only use SNP from the two studies that meet a shared benchmark. Only including SNP with an INFO score greater than 0.9 is standard [45]. Additional harmonization procedures include filtering SNP included to only those which exist in a suitable large reference panel, such as HapMap3 [46]. SNP not included in HapMap3 can be removed, and the HapMap3 reference allele can additionally be used as the standard. It is also best practice to set a minor allele and maximal frequency threshold, such as $0.5 > MAF > 0.01$, and remove SNP that do not meet the criteria between the two datasets [47, 48]. Finally, LD scores either need to be calculated or appended to your GWAS summary statistics. Given the study population of origin, linkage disequilibrium scores can directly be annotated from a suitable reference. In our example of using TRICL-OncoArray consortium and the UKB, these studies are in individuals of European ancestry samples. Precomputed LD scores from a suitable dataset, such as the 1000 genomes European ancestry samples [49], may be obtained and joined to these summary statistic data. The related LDSR package has software to assist with these steps, and an annotated GitHub "wiki" page which can assist in incorporating suitable reference panels [50]. In Table 1, we have included example output after GWAS summary statistic harmonization for one of our example datasets that TRICL-OncoArray lung cancer. Once the summary statistics are harmonized, they have SNP-based LD scores as well z-scores as can be observed in Fig. 1C.

Linkage Disequilibrium Score Regression We have now discussed all the relevant pieces to conduct LDSR. To obtain the genetic correlation between two traits, such as overall lung cancer and BMI, you first calculate genetic covariance. Genetic covariance is obtained as the slope of the regression of the product of SNP z-scores against the per SNP linkage disequilibrium score for two traits using post-GWAS summary statistics. This is a slight simplification as in reality, the linkage disequilibrium score, which serves as the independent variable of the regression is normalized by the sample

Table 1 Example summary statistic output from GWAS summary statistics. This excerpt is taken from a GWAS of overall lung cancer, conducted and made available by the TRICL-OncoArray consortium

chr:pos	rs number	CHR	Position	Reference allele	Effect allele	EAF	OR	Std error	*P* value	*N* study	*N* cases	*N* controls
1:768,253	rs2977608	1	768,253	C	A	0.263621	0.972829	0.018705	0.140837	7	25,252	51,026
1:768,448	rs12562034	1	768,448	G	A	0.115154	0.991593	0.024271	0.72795	7	25,252	51,026
1:769,963	rs7518545	1	769,963	G	A	0.109147	0.992248	0.025075	0.756273	7	25,252	51,026
1:773,106	rs115616822	1	773,106	G	A	0.010418	1.0525	0.087128	0.55702	6	24,118	50,077
1:777,232	rs112618790	1	777,232	C	T	0.073764	0.980984	0.033359	0.564932	7	25,102	51,047
1:823,790	rs143626389	1	823,790	G	A	0.027093	0.952242	0.053516	0.360495	6	24,118	50,077
1:837,214	rs72631888	1	837,214	G	C	0.013284	1.030088	0.071702	0.679282	6	24,118	50,077
1:848,223	rs4246504	1	848,223	G	A	0.015336	0.964539	0.06255	0.563794	6	24,118	50,077
1:851,390	rs72631889	1	851,390	G	T	0.030402	0.969281	0.0494	0.527662	6	24,118	50,077

sizes of the GWAS studies being used, as well as the total number of SNP being included in the analysis [24••]. The LDSR regression equation slope, or genetic covariance, can be viewed in Fig. 1D. With a second normalization step, this genetic covariance estimate can be modified to a genetic correlation estimate by dividing the genetic covariance by the square root, of the product squared heritability of each individual trait [24••]. The heritability of a trait can be thought of as the amount of variance observed in a trait that can be explained through genetics [21]. In general, it is calculated in the same way as genetic covariance (Fig. 1D), however for the dependent variable, is the *z*-score of the trait squared, rather than the product of the *z*-scores per SNP between two traits [51]. To be precise, then, the heritability we estimate as part of the LDSR method is the proportion of the variance in a trait that can be explained by SNP which have been filtered down to be in common with the LD score standardized reference [52]. Dividing the genetic covariance between two traits by the square root of their product squared individual heritabilities yields the estimate of cross-trait genetic correlation [24••]. In conclusion then, we can calculate genetic correlation by normalizing the genetic covariance by each trait's heritability [53]. These equations can be viewed in Fig. 1D.

In reality, the actual package implementation of LDSR is a little more complicated than a simple regression. SNP utilized in GWAS that have high LD scores may be correlated with each other, which could inflate genetic correlation estimates inappropriately. Spurious association also could arise as a result of $E[z_1z_2]$ of SNP with high LD scores having higher variance than with low LD scores, known as heteroscedasticity. The full LDSR method accounts for these concerns by performing a weighted regression. Specifically, LDSR utilizes variable SNP weighting, adding an additional weight term of heteroscedasticity/LD score per SNP. During model fitting, SNP weights are correspondingly increased based on their heteroscedasticity and decreased by their LD

score to mitigate these concerns [24••]. A second concern for the LDSR exists regarding the upward or downward "pull" that large $E[z_1z_2]$ would have on the regression slope (ρ_g) and intercept terms. A few extremely large SNP $E[z_1z_2]$ outliers could greatly increase final genetic correlation estimates between two traits. The solution that LDSR takes to avoid this problem is to perform two sequential weighted regressions [24••]. First, a weighted regression is performed with large $E[z_1z_2]$ removed. This first regression is used to identify the regression intercept, without large $E[z_1z_2]$ bias. The intercept from this first regression is captured and fixed as a constant in a second, subsequent regression where all $E[z_1z_2]$ are used. The slope of the second regression then serves to estimate the genetic correlation between two traits, not skewed by disproportionately large $E[z_1z_2]$. A full discussion of conducting LDSR [54], including file formats, preprocessing steps, and troubleshooting, can be found online [55]. For your reference, we have included Table 2 an example of the output that is created when using the LDSR method. Following the example initiated earlier, Table 2 presents the output from comparing overall lung cancer from the TRICL-OncoArray Consortium to BMI and fitness related traits.

Advantages of LDSR When conducting GWAS population stratification [56] and cryptic heritability [57], both have the potential to inflate per SNP effect estimates. Population stratification occurs when similarities in allele frequencies exist within subpopulations of a study due to non-random mating within population subgroups [56]. Cryptic heritability occurs when individuals in cohort are related to each other [57]. However, LD score not *z*-scores is the independent variable in LDSR [24••]. While population stratification [58] and cryptic heritability may inflate *z*-scores, they will have no effect on LD term [24••]. Instead, the inflation from these confounding biases gets captured as part of the intercept term of the regression. The LDSR intercept (minus one)

Table 2 Example LDSR output from genetic correlation analysis between TRICL-OncoArray lung cancer outcomes with UKBB fitness-related traits. Full comparative analysis published in Pettit et al. 2021. UKB trait names are in bold. "rg" stands for pairwise genetic correlation. "se" stands for standard error of rg. "z" stands for z-score for rg. "p" stands for p-value for rg. "h2 obs" stands for observed scale heritability for trait. "h2 obs se" stands for standard error of observed scale heritability for trait. "h2 int" stands for single-trait ld score regression intercept for trait. "h2 int se" stands for standard error of single-trait ld score regression intercept for trait. "gcov int" stands for cross-trait ld score regression intercept. "gcov int se" stands for standard error of cross-trait ld score regression intercept

	rg	se	z	p	h^2	h^2 se	h^2 int	h^2 int se	gcov int	gcov int se
Body mass index										
Lung cancer	0.20	0.03	5.96	2.61E-09	0.25	0.01	1.07	0.02	0.01	0.01
Adenocarcinoma	0.10	0.04	2.75	5.91E-03	0.25	0.01	1.07	0.02	0.00	0.01
Small cell	0.24	0.05	5.09	3.54E-07	0.25	0.01	1.07	0.02	0.02	0.01
Squamous cell	0.27	0.04	6.11	9.91E-10	0.25	0.01	1.07	0.02	0.01	0.01
Body fat percentage										
Lung cancer	0.17	0.03	4.99	6.11E-07	0.23	0.01	1.07	0.02	0.01	0.01
Adenocarcinoma	0.07	0.04	1.82	6.95E-02	0.22	0.01	1.07	0.02	0.01	0.01
Small cell	0.21	0.04	4.74	2.13E-06	0.22	0.01	1.07	0.02	0.02	0.01
Squamous cell	0.23	0.04	5.21	1.85E-07	0.22	0.01	1.07	0.02	0.01	0.01
Waist circumference										
Lung cancer	0.19	0.03	5.59	2.27E-08	0.20	0.01	1.06	0.02	0.01	0.01
Adenocarcinoma	0.09	0.04	2.44	1.48E-02	0.20	0.01	1.06	0.02	0.00	0.01
Small cell	0.21	0.05	4.54	5.56E-06	0.20	0.01	1.06	0.02	0.02	0.01
Squamous cell	0.27	0.05	5.98	2.23E-09	0.20	0.01	1.06	0.02	0.00	0.01
Weight										
Lung cancer	0.14	0.03	4.30	1.68E-05	0.26	0.01	1.09	0.02	0.01	0.01
Adenocarcinoma	0.06	0.04	1.57	1.16E-01	0.26	0.01	1.08	0.02	0.00	0.01
Small cell	0.18	0.04	4.16	3.14E-05	0.26	0.01	1.09	0.02	0.02	0.01
Squamous cell	0.22	0.04	5.07	4.04E-07	0.26	0.01	1.09	0.02	0.01	0.01

can be used to estimate the contributions that population stratification and cryptic relatedness have on genetic covariance estimates [24••]. Thus, the LDSR method is robust and mostly mitigates inflation in associations otherwise expected by these two GWAS confounding study variables [24••]. Another advantage of LDSR is that it estimates genetic correlation from summary statistic data from GWAS, not the individual level data used to conduct the GWAS originally. This is helpful as many GWAS summary statistic data files are made publicly available online, while the individual data is not. This has positioned the LDSR method to be widely incorporated, drawing meaningful insight from the vast amount of available trait-specific GWAS results [59]. Finally, LDSR is computationally very efficient [24••]. The only inputs for LDSR are summary statistic files, as well as precomputed LD scores per chromosome. The method iteratively conducts weighted regressions on these data. Using a MacBook Pro with 2.6 GHz 6-core Intel Core i7 and 16 GB of 2400 MHz DDR4 onboard memory, it takes roughly 12 s to conduct LDSR between two traits after all files are properly formatted and this script is executed.

Disadvantages of LDSR LDSR is limited by the quality of underlying GWAS it incorporates [60]. Oftentimes, GWAS occurs in volunteers and a healthy volunteer bias may exist [61]. Thus, heritability estimates measured with LDSR of traits may not reflect a general population. Furthermore, heritability estimates are highly influenced by the robustness and consistency of the trait's measurement [52]. Increasing

the variability by which a trait is measured will decrease the heritability estimate as a result [62]. This fact is part of what explains why anthropomorphic traits (height, weight, etc.) demonstrate consistent and high heritability—they are both very heritable but also very accurately measured routinely [63]. The power to detect genetic correlation or measure trait heritability with LDSR will be directly influenced by how well that trait is routinely captured in your underlying GWAS [64]. Diving further, as LDSR is calculated using GWAS summary statistics, it does not have access to individual level data [65]. To achieve the same standard error estimates with summary statistic data, we need a much larger sample size than would be necessary when working with individual participant genotype level date. As a rough guide, it has been proposed that LDSR can only meaningfully be performed if the underlying GWAS sample sizes have enough power to detect a h^2 of ≥ 0.2 [65]. In practice, this corresponds to needing an underlying GWAS trait sample size of at least 4500 individuals to have 80% power to detect h^2 equal to 0.2 with nominal significance [65]. Additional limitations include the need for reference LD scores, limiting LDSR use from being applied to GWAS of recently admixed populations [61]. This often this leads to European only analysis.

Aside from intrinsic methodological limitations, we feel that some anecdotal observations are also worth noting. Due to the computational efficiency with which LDSR operates, the method is attractive for large multi-trait screening. It is not uncommon to compare hundreds of traits with a

potential outcome of interest to check for genetic correlation and shared heritability [66]. The literature has examples of both reporting of nominal and adjusted p significance values to account for multiple comparisons testing [67]. While either is acceptable if explicitly notated, it is best practice to provide a Bonferroni adjusted *p*-values when reporting significance of trait genetic correlations to account for all comparisons made during the study investigation [22, 67]. As mentioned, these shared heritability estimates are limited to the information captured in the underlying SNP-based genotyping methods [21]. A final, and important, limitation of LDSR is that it only compares the genetics of two traits, but does not capture potential confounding, mediating, or colliding effects that additional traits may have on two phenotypes being studied. Collider bias, which occurs from unforeseen study conditioning or selection by a third trait, can lead to apparent strong genetic correlation results when truly none exists [68, 69]. Mitigating such biases is a capability of multivariate Mendelian randomization methods [70] that is not yet developed at the summary statistic level for complex polygenic traits and LDSR genetic correlation analysis.

Applications of LDSR The LDSR method has been widely applied, capitalizing on the ever increasing publicly available GWAS summary statistic information that are published for complex traits. Examples of GWAS meta-analyses for polygenic traits for which LDSR can and have been applied include addressing anxiety disorders [71], PTSD [72], psychiatric disorders including attention-deficit/hyperactivity disorder [73], schizophrenia/bipolar [74], autism [75], neurologic disorders [76], educational attainment [77], acne [78], and many cancers including the lung [29], breast [79], and pancreas [80]. A full listing of available GWAS for analysis can be found in the NHGRI-EBI catalogue of published GWAS [81]. These disease-specific resources are ripe for pairing with biobanks and public consortia of gene-trait phenotypic modeling. A prominent example resource includes the UKB [30, 82]. Preprocessed UKB trait summary statistic files can be downloaded [83] directly from the Neale lab. A US counterpart, the All of Us study is building a 1,000,000-volunteer consortium combining genetic and phenotypic data [84]. As of the spring of 2022, the All of Us data has 340,000 samples and 450,000 participants [84] and is in the beta phase of testing with some research applications available. An additional resource is the Million Veteran Program [85, 86] which capitalizes on large Veterans Affairs healthcare system in the USA which already has a few prominent examples [87, 88] of utilization. Increasing access to privately held sequencing data for traits is emerging, and LDSR studies are being conducted utilizing 23andMe data [89, 90]. Stand-alone websites to guide implementation and use of LDSR exist, with LDhub [91] demonstrating notable

ease of use [92]. With burgeoning resources, it is not surprising that in the last year, we have gained novel insight into the shared genetic architectures of the placenta [93], obesity [94], polycystic ovary syndrome [94], cross-population associations [95], hypertrophic and dilated cardiomyopathies [96], carpal tunnel syndrome [97], osteoarthritis [98], and others with LDSR.

Emerging Methods and Future Directions A recent method improves upon LDSR through increasingly precise, and with less variance, genetic correlation estimations between two traits. The high-definition likelihood inference of genetic correlations, or HDL [99] method, was published in 2020 [99]. HDL produces congruent genetic correlation estimates as conducted by LDSR. However, by more fully accounting for LD across the genome, the HDL method is able to reduce the variance of genetic correlation estimates by 60%, which otherwise could only be accomplished by doubling underlying GWAS sample sizes with LDSR. Overall, the method is more computationally intensive, and with the same computational platform previously described, it takes roughly four minutes to compute genetic correlations between traits. This modest increase in processing time may make the method overall less attractive for large multi trait scans; however, its increased ability to measure significant correlation with smaller GWAS samples will ensure its future adoption. HDL will serve as an excellent method to confirm or rule out genetic correlation associations identified with LDSR that may be of borderline significance due to LDSR limitations. For a full introduction to the HDL method, we refer you to the original paper [99], and GitHub "wiki" page [100] for accessing the computational tool.

Finally, it would be interesting to understand the genetic correlation that exists between traits independent of potential third trait confounding. Finishing out our example, it would be interesting if an LDSR method could determine the genetic correlation between BMI and lung cancer independent of the effects of smoking. There is room for method development of a multi-trait linkage disequilibrium score regression, which could work with summary statistics and produce contingent genetic correlation estimates to remove or partially mitigate the effects that may exist with other related traits. In general, these and other future developments of LDSR will be most useful for traits that are polygenic. For traits or disease processes where a handful of significant SNP can describe the large amount proportion of heritability of the trait, then Mendelian randomization [101] and related methods [102] are more useful for modeling trait relationships. While this type of multi-trait modeling can be conducted with Mendelian randomization for strong SNP derived traits [101], no methods exist for genetic correlation analysis with complex traits to conduct such contingency analysis using summary statistic GWAS data. Despite these

avenues for future investigation, LDSR will continue to have substantial utility for estimating heritability and for gaining novel insight into understanding complex polygenic trait relationships.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Papers of particular interest, published recently, have been highlighted as:
● Of importance
●● Of major importance

1. 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 20:567–581
2. Radick G. Making sense of Mendelian genes. 2020. https://doi.org/10.1080/0308018820201794387. 45:299–314.
3. Pecker LH, Lanzkron S. Sickle cell disease. 2021. https://doi.org/10.7326/AITC202101190. 174:ITC1–ITC16
4. de Boeck K. Cystic fibrosis in the year 2020: a disease with a new face. Acta Paediatr. 2020;109:893–9.
5. Duan D, Goemans N, Takeda S, Mercuri E, Aartsma-Rus A. Duchenne muscular dystrophy. Nat Rev Dis Prim. 2021;7(1):1–19.
6. Bick D, Bick SL, Dimmock DP, Fowler TA, Caulfield MJ, Scott RH. An online compendium of treatable genetic disorders. Am J Med Genet C Semin Med Genet. 2021;187:48–54.
7. Pazoki R. Methods for polygenic traits. Methods Mol Biol. 2018;1793:145–56.
8. Bind MA. Causal modeling in environmental health. 2019. https://doi.org/10.1146/annurev-publhealth-040218-044048. 40:23–43
9. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular diseases: a historical perspective. Lancet. 2014;383:999.
10. O'Donnell C, Edition RE-RE de C (English, 2008 undefined Cardiovascular risk factors. Insights from framingham heart study. Elsevier
11. Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham Heart Study. Nat Rev Cardiol. 2019;16(11):687–98.
12. Hubbard T. The human genome project. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 2005.https://doi.org/10.1002/047001153X.G203112
13. Kim S, Misra A. SNP genotyping: technologies and biomedical applications. 2007. https://doi.org/10.1146/annurev.bioeng9060906152037. 9:289–320
14. Halldórsson BV, Istrail S, de La Vega FM. Optimal selection of SNP markers for disease association studies. Human Heredity. 2004;58:190–202.
15. Craig Venter J, Adams MD, Myers EW, et al. (2001) The sequence of the human genome. Science. 1979;291:1304–51.
16. Mardis ER. DNA sequencing technologies: 2006–2016. Nat Protoc. 2017;12(2):213–8.
17. Shastry BS, Shastry BS. SNP alleles in human disease and evolution. J Hum Genet. 2002;47(11):561–6.
18. Salisbury BA, Pungliya M, Choi JY, Jiang R, Sun XJ, Stephens JC. SNP and haplotype variation in the human genome. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2003;526:53–61.
19. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
20. Dudbridge F. Polygenic epidemiology. Genet Epidemiol. 2016;40:268–72.
21. Zaitlen N, Kraft P. Heritability in the genome-wide association era. Hum Genet. 2012;131(10):1655–64.
22. Kraft P, Chen H, Lindström S. The use of genetic correlation and Mendelian randomization studies to increase our understanding of relationships between complex traits. Curr Epidemiol Rep. 2020;7(2):104–12.
23.●● Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236–41. (**Introduction and validation of the LDSR method for calculating genetic correlation**)
24.●● Bulik-Sullivan B, Loh PR, Finucane HK, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291–5. (**Validation of LDSR capabilities for bias mitigation**)
25. MacArthur JAL, Buniello A, Harris LW, et al. Workshop proceedings: GWAS summary statistics standards and sharing. Cell Genomics. 2021;1: 100004.
26. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. Nature Nat Rev Methods Primers 2021. https://doi.org/10.1038/s43586-021-00056-9.
27. Tips for Formatting A Lot of GWAS Summary Association Statistics Data. https://huwenboshi.github.io/data%20management/2017/11/23/tips-for-formatting-gwas-summary-stats.html. Accessed 18 Feb 2022
28. Amos CI, Dennis J, Wang Z, et al. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. Cancer Epidemiol Biomark Prev. 2017;26:126–35.
29. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49:1126.
30. Sudlow C, Gallacher J, Allen N, 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 Med. 2015;12: e1001779.
31. Pulst SM. Genetic linkage analysis. Arch Neurol. 1999;56:667–72.
32. Weir BS. Inferences about linkage disequilibrium. Biometrics. 1979;35:235.
33. Muller HJ. The mechanism of crossing-over. 2015. https://doi.org/10.1086/279534. 50:193–221
34. Mather K. Crossing-over. Biol Rev. 1938;13:252–92.
35. Mueller JC. Linkage disequilibrium for different scales and applications. Brief Bioinform. 2004;5:355–64.
36. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477–85.

37. Harpending H, Rogers A. Genetic perspectives on human origins and differentiation. 2003. https://doi.org/10.1146/annurev.genom11361. 1:361–385

38. Hinch AG, Tandon A, Patterson N, et al. The landscape of recombination in African Americans. Nature. 2011;476:170–5.

39. Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet. 2010;42:859–63.

40. Gabriel SB, Schaffner SF, Nguyen H, et al. (2002) The structure of haplotype blocks in the human genome. Science. 1979;296:2225–9.

41. Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. Trends Genet. 2003;19:135–40.

42. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A. A comprehensive evaluation of SNP genotype imputation. Hum Genet. 2008;125(2):163–71.

43. Halperin E, Stephan DA. SNP imputation in association studies. Nat Biotechnol. 2009;27(4):349–51.

44. Altshuler DM, Durbin RM, Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.

45. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5: e1000529.

46. Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52–8.

47. Linck E, Battey CJ. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol Ecol Resour. 2019;19:639–47.

48. Tabangin ME, Woo JG, Martin LJ. The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proceedings. 2009;3(7):1–4.

49. Siva N. 1000 genomes project. Nat Biotechnol. 2008;26:256.

50. GitHub - bulik/ldsc: LD score regression (LDSC). https://github.com/bulik/ldsc. Accessed 27 Mar 2022

51. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. 2017. https://doi.org/10.1371/JOURNAL.PGEN.1006711.

52. Heritability 501: LDSR-based h2 in UKBB for the technically minded — Neale lab. http://www.nealelab.is/blog/2017/9/14/heritability-501-ldsr-based-h2-in-ukbb-for-the-technically-minded. Accessed 26 Mar 2022

53. Lynch M, Walsh B. Genetics and analysis of quantitative traits. 1998.

54. Home · bulik/ldsc Wiki · GitHub. https://github.com/bulik/ldsc/wiki. Accessed 26 Mar 2022

55. GitHub - bulik/ldsc: LD Score Regression (LDSC). https://github.com/bulik/ldsc. Accessed 26 Mar 2022

56. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. The Lancet. 2003;361:598–604.

57. Paaby AB, Rockman MV. Cryptic genetic variation: evolution's hidden substrate. Nat Rev Genet. 2014;15(4):247–58.

58. Browning SR, Browning BL. Population structure can inflate SNP-based heritability estimates. Am J Hum Genet. 2011;89:191–3.

59. Werme J, van der Sluis S, Posthuma D, de Leeuw CA. An integrated framework for local genetic correlation analysis. Nat Genet. 2022;54(3):274–82.

60. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27: e1608.

61. Alten S van, Domingue B, … TG-B, 2021 undefined The effects of demographic-based selection bias on GWAS results in the UK biobank. cupc.colorado.edu

62. Génin E. Missing heritability of complex diseases: case solved? Hum Genet. 2019;139(1):103–13.

63. Bourrat P, Lu Q, Jablonka E. Why the missing heritability might not be in the DNA. BioEssays. 2017. https://doi.org/10.1002/bies.201700067.

64. Aschard H. A perspective on interaction effects in genetic association studies. Genet Epidemiol. 2016;40:678–88.

65. Defining Confidence Levels for UKB Round 2 LDSR Analyses. https://nealelab.github.io/UKBB_ldsc/confidence.html. Accessed 21 Feb 2022

66. Byun J, Han Y, Ostrom QT, Edelson J, Walsh KM, Pettit RW, Bondy ML, Hung RJ, McKay JD, Amos CI. The shared genetic architectures between lung cancer and multiple polygenic phenotypes in genome-wide association studies. Cancer Epidemiol Biomark Prev. 2021;30:1156–64.

67. Pettit RW, Byun J, Han Y, Ostrom QT, Edelson J, Walsh KM, Bondy ML, Hung RJ, McKay JD, Amos CI. The shared genetic architecture between epidemiological and behavioral traits with lung cancer. Sci Rep. 2021;11(1):1–12.

68. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39:417–20.

69. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun. 2020;11(1):1–12.

70. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol. 2019;48:713–27.

71. Otowa T, Hek K, Lee M, et al. Meta-analysis of genome-wide association studies of anxiety disorders. Mol Psychiatry. 2016;21:1391.

72. Duncan LE, Ratanatharathorn A, Aiello AE, et al. Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Mol Psychiatry. 2018;23:666–73.

73. Faraone SV, Larsson Henrik, Org S. Genetics of attention deficit hyperactivity disorder. Mol Psychiatry. 2018;24(4):562–75.

74. Lo MT, Hinds DA, Tung JY, et al. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. Nat Genet. 2017;49:152.

75. Robinson EB, St Pourcain B, Anttila V, et al. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nat Genet. 2016;48:552.

76. Anttila V, Bulik-Sullivan B, Finucane HK, et al. (2018) Analysis of shared heritability in common disorders of the brain. Science. 1979. https://doi.org/10.1126/SCIENCE.AAP8757/SUPPL_FILE/AAP8757_TABLE_S7.XLSX.

77. Okbay A, Beauchamp JP, Fontana MA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016;533:539–42.

78. Mitchell B, Saklatvala J, … ND-N, 2022 undefined Genome-wide association meta-analysis identifies 29 new acne susceptibility loci. nature.com

79. Fachal L, Dunning AM. From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. Curr Opin Genet Dev. 2015;30:32–41.

80. Amundadottir LT. Pancreatic cancer genetics. Int J Biol Sci. 2016;12:314.

81. Buniello A, Macarthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12.

82. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. 2017;13: e1006711.

83. Results Files for LDSR of UK Biobank GWAS. https://nealelab.github.io/UKBB_ldsc/downloads.html. Accessed 21 Feb 2022

84. All of Us Research Hub. https://www.researchallofus.org/. Accessed 21 Feb 2022

85. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70:214–23.

86. Million Veteran Program. https://www.mvp.va.gov/pwa/. Accessed 21 Feb 2022

87. Klarin D, Lynch J, Aragam K, et al. Genome-wide association study of peripheral artery disease in the Million Veteran Program. Nature Medicine. 2019;25(8):1274–9.

88. Klarin D, Damrauer SM, Cho K, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nature Genetics. 2018;50(11):1514–23.

89. Warrier V, Team the 23andMe R, Bourgeron T, Baron-Cohen S. Genome-wide association study of social relationship satisfaction: significant loci and correlations with psychiatric conditions. bioRxiv. 2017;196071.

90. Warrier V, Grasby KL, Uzefovsky F, et al. Genome-wide meta-analysis of cognitive empathy: heritability, and correlates with sex, neuropsychiatric conditions and cognition. Mol Psychiatry. 2017;23(6):1402–9.

91. Zheng J, Erzurumluoglu AM, Elsworth BL, 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;33:272.

92. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431–44.

93. Bhattacharya A, Freedman A, … VA-N, 2022 undefined Placental genomics mediates genetic associations with complex health traits and disease. nature.com

94. Liu Q, Zhu Z, Kraft P, Deng Q, Stener-Victorin E, Jiang X. Genomic correlation, shared loci, and causal relationship between obesity and polycystic ovary syndrome: a large-scale genome-wide cross-trait analysis. BMC Med. 2022;20:66.

95. Sakaue S, Kanai M, Tanigawa Y, … JK-N, 2021 undefined A cross-population atlas of genetic associations for 220 human phenotypes. nature.com

96. Tadros R, Francis C, Xu X, Vermeer A, … AH-N, 2021 undefined Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. nature.com

97. Skuladottir ATh, Bjornsdottir G, Ferkingstad E, et al. A genome-wide meta-analysis identifies 50 genetic loci associated with carpal tunnel syndrome. Nature Communications. 2022;13(1):1–9.

98. Shen B, Xu J, Si H, Zeng Y, Wu Y, Li M. Evaluating the genetic correlation and causal relationship between osteoarthritis and human blood metabolites. 2022. https://doi.org/10.21203/rs.3.rs-1421690/v1

99. Ning Z, Pawitan Y, Shen X (2020) High-definition likelihood inference of genetic correlations across human complex traits. Nature Genetics 2020 52:8 52:859–864

100. Installation and update · zhenin/HDL Wiki · GitHub. https://github.com/zhenin/HDL/wiki/Installation-and-update. Accessed 28 Mar 2022

101. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res. 2007;16:309–30.

102. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Practice of Epidemiology. 2015. https://doi.org/10.1093/aje/kwu283.