BMC Bioinformatics

SOFTWARE

Open Access

PRED-LD: efficient imputation of GWAS summary statistics

Georgios A. Manios[1], Aikaterini Michailidi[1], Panagiota I. Kontou[2] and Pantelis G. Bagos[1*]

*Correspondence:
pbagos@compgen.org

[1] Department of Computer Science and Biomedical Informatics, University of Thessaly, 35131 Lamia, Greece
[2] Department of Mathematics, University of Thessaly, 35131 Lamia, Greece

Abstract

Background: Genome-wide association studies have identified connections between genetic variations and diseases, but they only examine a small portion of single nucleotide polymorphisms. To enhance genetic findings, researchers suggest imputing genotypes for unmeasured SNPs to improve coverage and statistical power. When this is not possible, summary statistics imputation can be used as an alternative. The available summary statistics imputation tools rely on reference panels, such as the 1000 Genomes Project, to estimate linkage disequilibrium (LD) between variants for accurate imputation. Tools like FAPI and SSIMP use these reference panels in variant call format (VCF) for this purpose, though this process can be time-consuming. A more effective approach for processing reference panels in summary statistics imputation was proposed in RAISS. In this approach, the LD among the variants is precomputed from the reference panel, prior to imputation, thereby reducing computational time.

Results: We present PRED-LD, an imputation method for GWAS summary statistics that aims to enhance the resolution of genetic association analyses. The proposed method uses precomputed linkage disequilibrium statistics from HapMap, Pheno Scanner and TOP-LD to impute summary statistics, given beta coefficients and standard errors. The single-point approach that we describe provides a fast and accurate way to estimate associations for untyped single nucleotide polymorphisms that exhibit high linkage disequilibrium (LD). The proposed method is faster, provides accurate imputation compared to existing tools, and has been implemented in both a web service (https://compgen.dib.uth.gr/PRED-LD/) and a command-line tool (https://github.com/pbagos/PRED-LD), making it a useful resource for the research community.

Conclusions: PRED-LD offers an efficient and accurate method for GWAS summary statistics imputation, providing faster performance, direct result interpretation, and the ability to use multiple reference panels. Also, the online version of PRED-LD simplifies obtaining LD information and performing imputation tasks without downloading reference panels and will be continuously updated to support tools for meta-analysis and fine-mapping in GWAS.

Keywords: Genome-wide association studies, Summary statistics, Imputation, Linkage disequilibrium

Background

Genome-wide association studies (GWAS) have been successful in identifying links between genetic variations and diseases [1]. However, it is important to note that GWAS only explores a fraction of single nucleotide polymorphisms and, depending on the platform used, most studies include information for different typed markers. To further enhance genetic association discoveries, researchers have suggested imputing genotypes for many unmeasured SNPs to increase the coverage, thereby enhancing statistical power, increasing the accuracy of fine-mapping, and enabling effective meta-analyses [2]. When genotype imputation is not feasible for practical or ethical reasons, summary statistics imputation provides a practical alternative. Tools such as DIST [3], ImpG [4], FAPI [5], SSIMP [6] and RAISS [7] are designed to perform summary statistics imputation, with the use of reference panels, such as the 1000 Genomes Project [8]. Moreover, DISTMIX [9] is employed to execute summary statistics imputation in admixed populations. It is based on the same algorithm as DIST and weights to each under study GWAS population can be applied. GAUSS [10], a recent R package, offers a range of functions for the estimation of ancestry proportions of study cohorts, calculation of linkage disequilibrium, imputation of summary statistics, and the conducting of transcriptome-wide association studies. Its imputation functions are based on DIST and DISTMIX algorithms and utilize a reference panel [11] comprising 32,953 genomes from 29 ethnic groups, thereby enhancing the accuracy of the results, particularly in the case of rare variants. These panels offer haplotype information from individuals with the same ancestry as the population under study and can achieve an accurate imputation, although the imputation process for an entire GWAS can be a time-consuming task. We present here a simpler yet efficient and very fast method for imputing summary statistics using precalculated linkage disequilibrium (LD). The proposed method is available as an open-source tool, PRED-LD, which also features a web service version, for easy use in summary statistics imputation tasks.

Methods/Implementation

Within the framework of PRED-LD, LD information along with the respective variant allele frequencies and LD patterns can be derived from three different sources used as reference panels (Fig. 1.), from HapMap [12], Pheno Scanner [13], and TOP-LD [14]. Pheno Scanner provides LD statistics with $r^2 > 0.8$ and Minor Allele Frequency (MAF) > 0.01 that have been computed using the super-ancestries in 1000 Genomes project phase 3 reference panels corresponding to Europeans, East Asians, South Asians, Africans, and Admixed Americans. HapMap LD data involves a collection of linkage disequilibrium data compiled from merged genotype data from phases I + II + III submitted by HapMap genotyping centers to the DCC. These LD data were generated from the HaploView [15] software. HapMap LD data include samples from various populations. TOP-LD is an online platform for investigating LD patterns, which leverages high-coverage whole genome sequencing (WGS) data from European, African, East Asian and South Asian individuals participating in the NHLBI TOPMed [16] program, with $r^2 \geq 0.2$. TOP-LD is an advanced tool for exploring LD that provides a comprehensive view of genetic variations through the TOPMed WGS data, particularly rare variants,

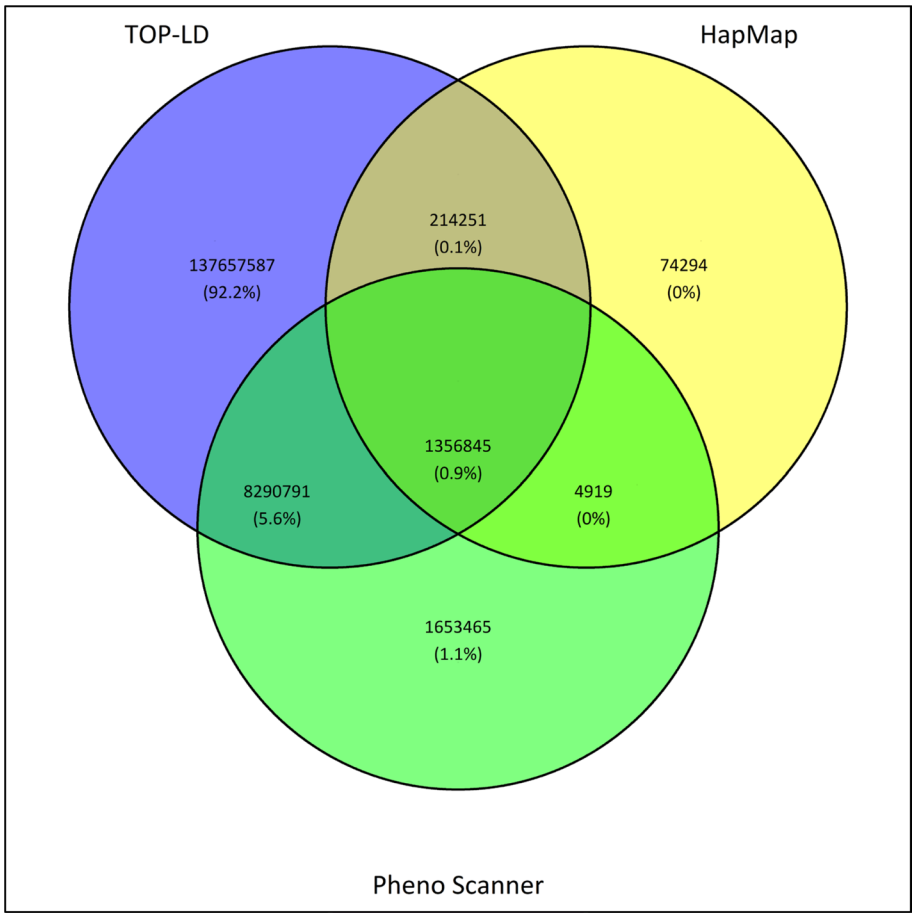

Fig. 1 Venn diagram of all the variants included in the LD reference panels that PRED-LD employs, in all ancestries

within specific populations. Compared to other LD resources such as HaploReg [17] or LDlink [18], TOP-LD represents a 2.6- to 9.1-fold increase in variant coverage. Regarding the selection of the reference panel, an intuitive solution would be to use TOP-LD as the primary reference for the imputation tasks, given that it encompassed the most extensive collection of variants. The imputation results with TOP-LD as reference panel, were both accurate and rapid. However, we investigated also the use of additional panels (Pheno Scanner, HapMap) in order to explore potential improvements.

PRED-LD, contrary to other method uses a single-point imputation method that relies on beta coefficients ($\beta = \log(OR)$) and standard errors from GWAS summary statistics. To estimate imputed beta coefficients for variants that were not typed (u), we first identify in the panel the typed SNP (t) with the maximum r^2 with the untyped one and then we use a well-known result by Zondervan and Cardon [19]. Zondervan and Cardon expanded an earlier finding, presented by Ackerman and coworkers [20], which demonstrated that for a trait locus with alleles T and t, having allele frequencies 1-p_t and p_t respectively, and a marker locus with alleles U and u and allele frequencies 1-p_u and p_u respectively, the odds ratio for an association involving the indirect allele u can be derived using the haplotype frequencies, as displayed in Eq. (1):

$$OR_u = \frac{(1-p_u)(OR_t p_{tu} + p_{tU})}{p_u(OR_t p_{Tu} + p_{TU})} \tag{1}$$

where OR_t is the trait or disease allelic *OR* of the typed variant, and p_{tu}, p_{tU}, p_{Tu} and p_{TU} correspond to the relevant haplotype frequencies. Given that $D = p_{tu} - p_t p_u$, Zondervan and Cardon showed that Eq. (1) can be reformulated as follows:

$$OR_u = 1 + \frac{D(OR_t - 1)}{p_u[(1-p_u) + (p_t(1-p_u) - D)(OR_t - 1)]} \tag{2}$$

where $D = \pm r\sqrt{p_t(1-p_t)p_u(1-p_u)}$, is the LD coefficient between the typed (t) and the untyped SNP (u), and r is the pairwise Pearson's correlation coefficient between the typed and untyped SNPs. Since the primary data is logOR and their standard errors, it is useful to rewrite Eq. (1) as:

$$\beta_u = \log\left(1 + \frac{D\left(e^{\beta_t} - 1\right)}{p_u\left[(1-p_u) + (p_t(1-p_u) - D)\left(e^{\beta_t} - 1\right)\right]}\right) \tag{3}$$

Some reference panels, like HapMap, provide information only on r^2 and D', so in such cases we also need to determine the sign of D. In doing so, we utilize the D' information from the LD panels. Given that $D' = D/D_{\max}$ [21] and

$$D_{\max} = \begin{cases} \min\{p_t p_u, (1-p_t)(1-p_u)\} \text{ when } D < 0 \\ \min\{p_t(1-p_u), p_u(1-p_t)\} \text{ when } D > 0 \end{cases} \tag{4}$$

it is now possible to ascertain which case, whether D is positive or negative, yields the corresponding D' value. In other words, using the known allele frequencies, we enumerate the two expressions in the right-hand side of Eq. (4) and decide which one holds. We need to mention that Eqs. from (1) to (4) all refer to population parameters. When we try to estimate the respective quantities from the sample, we need to denote them as estimates (for instance $\hat{\beta}_t$ and so on). Afterwards, by noticing that Eq. (3) is a function of β_t, an estimate of the variance and the standard errors of the imputed beta coefficients can be calculated using the Delta Method [22]:

$$\widehat{\text{var}}\left(f\left(\hat{\beta}_t\right)\right) \approx \left[f'\left(\hat{\beta}_t\right)\right]^2 \widehat{\text{var}}\left(\hat{\beta}_t\right) \tag{5}$$

with the derivative of f being given by:

$$f'(\beta_t) = \frac{\partial f(\beta_t)}{\partial \beta_t} = \frac{\frac{De^{\beta_t}}{(1+(e^{\beta_t}-1)(-D+p_t(1-p_u)-p_u))p_u} - \frac{De^{\beta_t}(e^{\beta_t}-1)(-D+p_t(1-p_u))}{(1+(e^{\beta_t}-1)(-D+p_t(1-p_u)-p_u))^2 p_u}}{1 + \frac{D(e^{\beta_t}-1)}{(1+(e^{\beta_t}-1)(-D+p_t(1-p_u)-p_u))p_u}} \tag{6}$$

Obviously, we use in Eq. (6) the sample estimates of the population parameters (D, β_t, p_t, p_u) and we plug the estimate of $f'(\beta_t)$ in Eq. (5) to obtain the estimated variance. This approach leverages the linkage disequilibrium and the allelic frequency information from the panels to assign the effect (logOR and its standard error) of the untyped marker. It is of importance to note that for each SNP to be imputed we utilize information of a single typed SNP, the one with the highest r^2. This approach allows the simultaneous use

of multiple panels and the inclusion of the SNP with the highest r^2. This contrasts with other methods that use all SNPs within a given window utilizing a multivariate approach and offers a number of significant advantages as we will see below.

Implementation

The Python source code of PRED-LD is accessible via a public GitHub repository at https://github.com/pbagos/PRED-LD. Users of PRED-LD can explore linkage disequilibrium information from various populations of the HapMap, Pheno Scanner, and TOP-LD precalculated LD panels, along with the results of the imputation process. Moreover, the users can conduct a more targeted imputation on specific rsIDs, giving a list of variants as an additional input argument and conduct whole GWAS imputation tasks. In addition, the web tool version includes Manhattan plots and QQ plots to depict the imputation results. The web version of PRED-LD (Figs. 2 and 3.) is publicly available at: https://compgen.dib.uth.gr/PRED-LD/. It is important to note that the web version of PRED-LD has a limitation of 20,000 rows for the input file. This restriction must be considered, to ensure that the imputation process will not be computationally intensive.

Datasets

To measure the accuracy of our method, we used eight distinct GWAS datasets. We collected a diverse set of Genome-Wide Association Studies (GWAS) focused on various traits, derived from open databases. The case–control ADHD dataset [23] was obtained from dbGaP [24], while datasets for traits such as urinary albumin to creatinine ratio (UACR) and glomerular filtration rate (GFR) [25] were obtained from GWAS Atlas [26]. The GFR data includes studies from both European and African populations. Additional datasets include studies on epilepsy [27], colorectal cancer [28], double eyelid [29] and coronary artery disease (CAD) [30], all derived from GWAS Atlas. This collection reflects a wide range of traits, populations, and genotyping platforms. The details of each.

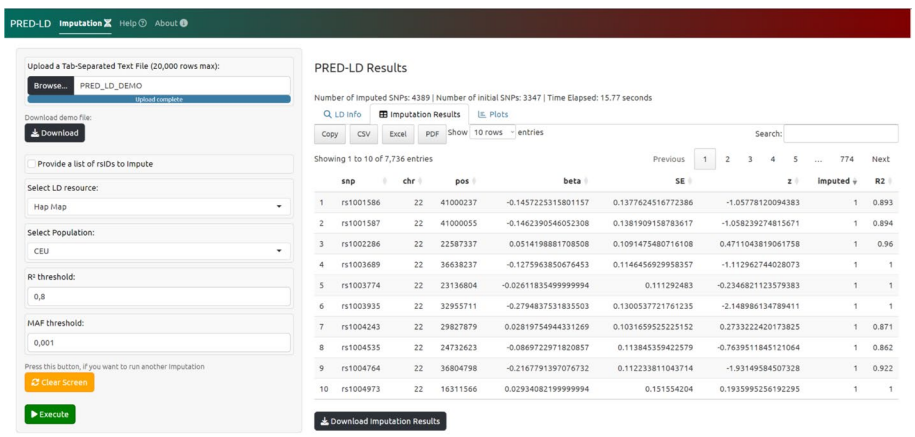

Fig. 2 Screenshot from the web interface of PRED-LD. In the sidebar panel, the user can select the desired options to perform an imputation task and in the main panel, the imputation results, LD information and plots are displayed

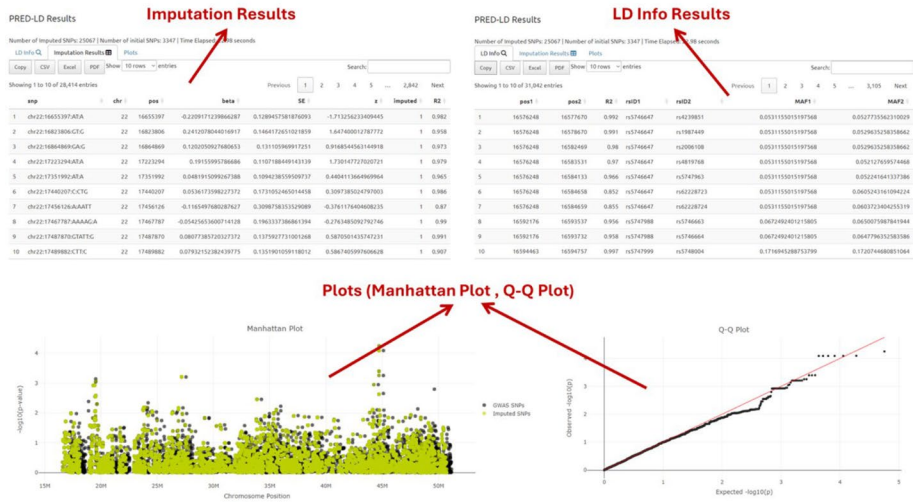

Fig. 3 Screenshot of results and plots of the web version of PRED-LD

Table 1 Overview of collected GWAS datasets

GWAS	Database	Database ID	Sample size	Number of SNPs	Population	Platform
ADHD	dbGaP	phs001869.v1.p1	1,001	247,475	EUR	Illumina Infinium PsychArray-24
UACR	GWAS Atlas	4061	31,164	90,283	EUR	Illumina HumanExome chip
GFR	GWAS Atlas	4053	14,308	64,719	EUR	Illumina HumanExome chip
GFR	GWAS Atlas	4054	2,162	59,504	AFR	Illumina HumanExome chip
Epilepsy	GWAS Atlas	4160	24,927	4,948,714	EUR	Illumina OmniExpress-24 v1.1
Colorectal Cancer	GWAS Atlas	4097	33,870	7,492,477	EAS	Illumina OmniExpress
Double Eyelid	GWAS Atlas	4021	5,614	549,759	EAS	Custom Affymetrix Axiom array
CAD	GWAS Atlas	3925	148,815	9,024,593	EUR	UK Biobank SNP array

study, including the specific traits, populations, and genotyping platforms, are provided in Table 1.

Results

The initial hypothesis was that the entire GWAS imputation tasks could be performed using all the available summary statistic imputation tools, thereby obtaining imputed values for the input variants provided in the input files. The first measure we use is the "number of imputed SNPs". That is, given the entire GWAS, the total number of additional SNPs whose effect could be predicted. This approach, however, does not provide

any clues as to whether these predictions are good or not. Thus, we need to perform predictions also on the SNPs that are already in the dataset and evaluate the performance. This approach would normally allow for a straightforward leave-one-out cross-validation. This was only possible with DIST and SSIMP (and PRED-LD), since the other tools do not offer such an option and performing the analysis repeatedly would require an enormous amount of time. In order to provide a fair comparison of all available methods, the performance of each tool was assessed according to the following procedure. For each GWAS dataset, we randomly removed (masked) 20% of the SNPs in chromosome 1 from the original dataset, performed summary statistic imputation on the removed variants, and computed the R^2 correlation coefficient between the observed z-scores and the imputed z-scores, as well as the observed and the imputed -— $\log_{10}(p)$ values. These measures account for the two measures of "imputation accuracy".

For DIST, FAPI and SSIMP, the imputation tasks were performed with the default settings. For RAISS, the subcommand "performance-grid-search" was performed to select its optimal performance parameters (eigen threshold and min-ld) prior to the imputation process, setting the same window length as the other methods (1000kbp). It is important to note that DIST had only two reference panels available for European populations (1000 Genomes Phase 1 Release 3 European and UK10K). Consequently, summary statistics imputation tasks with DIST for non-European populations were conducted using the 1000 Genomes Phase 1 Release 3 European reference panel, which included 386 samples and 9,544,788 total variants, despite the inherent bias in the results. Furthermore, FAPI performs p-value imputation, so only the $-\log_{10}(p)$ values were compared. GAUSS uses the same algorithms as DIST and DISTMIX, but it is designed for a different purpose and performs imputation in a narrow region. Finally, DISTMIX was excluded from the comparisons as it is designed for summary statistics imputation in admixed populations, whereas all the GWAS datasets consisted of discrete populations. To provide a clear understanding of the population representation and variant coverage within each reference panel, all the reference panels used in this study are described in Table 2. For all methods an important post-processing step was necessary, since in many cases the GWAS uses the alternative allele for reference and vice versa, which results in some of the beta coefficients to be given with the opposite sign. In such cases the reference and alternative allele for each marker were harmonized in order to have the GWAS under investigation to match those of the reference panel.

In the case of PRED-LD, we initially performed an evaluation in order to choose the best option regarding the reference panel. We thus investigated the use of the different panels separately, as well as in combination. In all cases we use an r^2 threshold of 0.5 which is regarded as an appropriate and impartial threshold for high LD (but we also investigated this, see below). Moreover, no minor allele frequency threshold was employed. The imputation tasks conducted on individual panels yielded promising results within a short time frame, particularly when using TOP-LD and Pheno Scanner, as illustrated in Table 3. The HapMap reference panel, being the smaller one, yields lower accuracy. However, combining all available panels resulted in slight improvements in the overall performance so we decided to include it as the default option for the method. The user, however, may choose differently (see below), especially when computation

Table 2 Description of the reference panels used by the different methods. We list a summary of the sample sizes and the total variants across the available populations

Reference panel	Label	population sample	Number of samples	Total Variants
TOP-LD ($r^2 \geq 0.2$)	EUR	European	13,160	69,524,944
	AFR	African	1,335	60,392,677
	SAS	South Asian	239	22,309,649
	EAS	East Asian	844	35,538,656
Pheno scanner (MAF > 1% and $r^2 > 0.8$,1000 Genomes project—Phase 3, hg19 and hg38)	EUR	European	503	11,159,862
	AFR	African	661	11,159,862
	SAS	South Asian	489	11,159,862
	EAS	East Asian	504	11,159,862
	AMR	American	347	11,159,862
HapMap	ASW	African ancestry in Southwest USA	90	1,561,113
	CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	180	1,412,161
	CHB	Han Chinese in Beijing, China	90	1,328,283
	CHD	Chinese in Metropolitan Denver, Colorado	100	1,305,880
	GIH	Gujarati Indians in Houston, Texas	100	1,407,540
	JPT	Japanese in Tokyo, Japan	91	1,296,969
	LWK	Luhya in Webuye, Kenya	100	1,529,438
	MEX	Mexican ancestry in Los Angeles, California	90	1,409,947
	MKK	Maasai in Kinyawa, Kenya	180	1,419,626
	TSI	Toscans in Italy	100	1,419,920
	YRI	Yoruba in Ibadan, Nigeria	180	1,501,085
RAISS (1000 Genomes Project Phase 3, hg38)	EUR	European	632	8,193,280
	AFR	African	893	13,876,891
	SAS	South Asian	601	8,579,150
	EAS	East Asian	585	7,245,426
	AMR	American	490	9,347,814
SSIMP & FAPI (1000 Genomes Project Phase 3, hg19)	EUR	European	503	81,271,745
	AFR	African	661	81,271,745
	SAS	South Asian	489	81,271,745
	EAS	East Asian	504	81,271,745
	AMR	American	347	81,271,745

time is of essence, since as it is apparent from the results, using only one of the panels results in a significant decrease in the execution time.

The comparisons of the summary statistics imputation tools across the aforementioned GWAS datasets revealed that PRED-LD demonstrated superior efficiency in terms of speed. On average PRED-LD, with the default option for combining all panels, completed the imputation task 3–20 times faster than other tools, including DIST, FAPI, and SSIMP, while maintaining superior imputation accuracy. To illustrate this, in certain datasets, SSIMP requires more than 18 h to complete the imputation process, whereas PRED-LD achieves the same result in less than 20 min. RAISS is also fast, but nevertheless PRED-LD is approximately 27.56% faster in overall execution time and 76.44% faster in time per 1,000 SNPs imputed. While tools such as DIST and

Table 3 Comparison of summary statistics imputation performance across each linkage disequilibrium panels, that PRED-LD utilizes. These results were obtained using an r^2 threshold of 0.5 for TOP-LD and HapMap LD panels and the r^2 threshold of 0.8 for Pheno Scanner, as its data inherently provide information using this r^2 threshold. When using the HapMap LD panel, we performed imputation for each subpopulation separately based on the respective GWAS population

LD Panel	GWAS	Imputed SNPs	$R^2(z)$	$R^2\left(-\log_{10}(p)\right)$	Imputation percentage in masked SNPs (%)	Time	Time per 1000 SNPs imputed (s)
TOP-LD	ADHD (EUR)	243,062	0.719	0.545	39.14	6 m 1 s	1.5
	UACR (EUR)	64,371	0.866	0.748	6.15	5 m 13 s	4.9
	GFR (EUR)	57,200	0.674	0.625	8.97	5 m 11 s	5.4
	GFR (AFR)	41,834	0.851	0.592	7.15	3 m 9 s	4.5
	Epilepsy (EUR)	440,870	0.924	0.860	91.82	9 m 50 s	1.3
	Colorectal Cancer (EAS)	378,395	0.823	0.772	62.68	7 m 7 s	1.1
	Double Eyelid (EAS)	274,356	0.628	0.427	47.27	6 m 8 s	1.3
	CAD (EUR)	682,449	0.925	0.894	77.87	12 m 21 s	1.1
Pheno Scanner	ADHD (EUR)	277,745	0.913	0.859	23.13	2 m 44 s	0.6
	UACR (EUR)	53,201	0.941	0.945	6.15	2 m 3 s	2.3
	GFR (EUR)	52,243	0.857	0.789	9.05	1 m 57 s	2.2
	GFR (AFR)	19,155	0.876	0.756	3.44	1 m 24 s	4.4
	Epilepsy (EUR)	460,371	0.942	0.894	95.06	8 m 7 s	1.1
	Colorectal Cancer (EAS)	339,280	0.917	0.876	61.25	3 m 40 s	0.6
	Double Eyelid (EAS)	281,966	0.803	0.661	41.81	2 m 54 s	0.6
	CAD (EUR)	589,504	0.960	0.938	77.65	10 m 13 s	1.0
HapMap (CEU)	ADHD (EUR)	44,510	0.568	0.472	30.22	1 m 20 s	1.8
HapMap (TSI)		43,810	0.662	0.509	29.62	1 m 51 s	2.5
HapMap (CEU)	UACR (EUR)	7,161	0.859	0.818	2.58	1 m 20 s	11.2
HapMap (TSI)		7,061	0.545	0.816	2.74	1 m 51 s	15.7
HapMap (CEU)	GFR (EUR)	7,200	0.530	0.490	3.93	1 m 9 s	9.6
HapMap (TSI)		7,076	0.725	0.605	4.01	1 m 32 s	13.0
HapMap (ASW)	GFR (AFR)	3,799	0.586	0.527	2.15	2 m 44 s	43.2
HapMap (LWK)		3,289	0.702	0.488	1.81	2 m 18 s	42.0
HapMap (MKK)		3,694	0.642	0.463	1.98	1 m 53 s	30.6
HapMap (YRI)		3,400	0.840	0.530	1.72	1 m 27 s	25.6
HapMap (CEU)	Epilepsy (EUR)	73,686	0.745	0.673	18.48	1 m 9 s	0.9
HapMap (TSI)		73,101	0.795	0.700	18.31	2 m 17 s	1.9
HapMap (CHB)	Colorectal Cancer (EAS)	40,336	0.760	0.732	9.54	1 m 14 s	1.8
HapMap (JPT)		40,609	0.772	0.761	9.43	2 m 26 s	3.6
HapMap (CHD)		39,193	0.820	0.790	9.00	2 m 56 s	4.5
HapMap (CHB)	Double Eyelid (EAS)	42,239	0.598	0.546	19.29	1 m 13 s	1.7
HapMap (JPT)		42,829	0.624	0.543	20.20	2 m 26 s	3.4
HapMap (CHD)		41,244	0.703	0.551	18.68	2 m 54 s	4.2
HapMap (CEU)	CAD (EUR)	82,190	0.836	0.806	11.73	1 m 42 s	1.2
HapMap (TSI)		82,278	0.795	0.799	11.75	2 m 30 s	1.8
Mean performance	**TOP-LD**	**272,817**	**0.801**	**0.683**	**42.63**	**6 m 52 s**	**2.6**
	Pheno Scanner	**259,183**	**0.901**	**0.840**	**39.69**	**4 m 7 s**	**1.6**
	HapMap	**34,435**	**0.705**	**0.631**	**11.36**	**1 m 54 s**	**11.0**
	All panels combined	**338,803**	**0.817**	**0.728**	**49.90**	**20 m 12 s**	**8.2**

Bold values denote the average performance across the datasets

SSIMP may achieve higher imputation coverage, they are associated with substantially longer run times and lower imputation accuracy. To provide fair runtime comparisons among the compared tools, we also took into consideration calculating the execution time per 1,000 SNPs imputed. Once again, PRED-LD is the faster among the tools considered here. The detailed comparison results are presented in Table 4 and Fig. 4.

The use of PRED-LD is transparent, and the user can choose different options regarding the reference panels or the r^2 threshold, in order to accomplish different tasks. To showcase the inverse relationship between accuracy and coverage we performed prediction in the test datasets under different LD thresholds for PRED-LD and R^2 thresholds for the other tools (Figs. 5 and 6). Thus, using a threshold of 0.8 we obtain smaller coverage but increased accuracy, whereas using a threshold of 0.5 we have lower accuracy but increased coverage for each tool. On the other hand, selecting only one panel may result to even faster imputations (3 to 10 times faster compared to the default option), with a moderate decrease in accuracy (Table 3). The only metric in which PRED-LD does not clearly outperform the other tools is coverage (and the number of imputed SNPs). In Table 4 we showed that DIST and SSIMP surpass PRED-LD in this regard, but PRED-LD can increase its coverage to almost match that of DIST, simply by lowering the LD threshold. Also, FAPI shows slightly better overall performance than PRED-LD in terms of $R^2\left(-\log_{10}(p)\right)$, with a small difference, whereas PRED-LD achieves 11.04% higher average coverage. In order to perform a head-to-head comparison against DIST and SSIMP, which show the higher coverage among the methods, in an unbiased manner, we performed the following additional evaluations. We filtered the results of DIST and SSIMP using the reported coefficient of determination (R^2) and we kept only the imputed SNPs with reported $R^2 > 0.5$. We also performed two additional comparisons of PRED-LD against DIST and SSIMP. The first considers the same number of imputed SNPs of PRED-LD (ranked in descending order of R^2 for DIST and SSIMP), while the second focuses on all the common imputed variants in every intersection combination. This way, we have results as comparable as possible to the ones obtained by PRED-LD with the $r^2 > 0.5$ default option for selecting the SNPs. The results are given in Table 5, where we can see that PRED-LD gives comparable results with DIST and SSIMP, except for speed, since it is still up to 3 and 20 times faster, respectively.

All the comparisons were conducted on a server equipped with an Intel Xeon E5-2660 v4 processor operating at a base frequency of 2.00 GHz, supported by 64 GB of RAM. The analysis code, for the execution commands and the presented results, is available in the GitHub repository of PRED-LD at the following link: (https://github.com/pbagos/PRED-LD/tree/main/paper).

Table 4 Comparison of the performance of the summary statistics imputation methods across the GWAS datasets

GWAS	Tool	Imputed SNPs	$R^2_{(z)}$	$R^2(-\log_{10}(p))$	Imputation percentage in masked SNPs	Time	Time per 1000 SNPs imputed (s)
ADHD (EUR)	PRED-LD	345,047	0.701	0.583	1,894/3,791 = 49.96%	21 m 37 s	3.8
	RAISS	213,003	0.576	0.366	1,504/3,791 = 39.67%	7 m 57 s	2.2
	DIST	737,618	0.558	0.402	3,784/3,791 = 99.81%	33 m 6 s	2.7
	SSIMP	1,875,319	0.614	0.452	3,763/3,791 = 99.26%	126 m 6 s	4.0
	FAPI	262,431	–	0.761	949/3,791 = 25.03%	146 m 45 s	33.5
UACR (EUR)	PRED-LD	91,989	0.885	0.874	166/1,819 = 9.12%	14 m 56 s	9.7
	RAISS	30,052	0.836	0.853	80/1,819 = 4.39%	7 m 24 s	14.7
	DIST	742,568	0.245	0.348	620/1,819 = 34.08%	25 m 1 s	2.0
	SSIMP	1,879,310	0.198	0.195	1,256/1,819 = 69.04%	117 m 54 s	3.8
	FAPI	47,068	–	0.947	90/1,819 = 4.94%	139 m 10 s	177.4
GFR (EUR)	PRED-LD	84,689	0.723	0.683	165/1,270 = 12.99%	14 m 20 s	10.2
	RAISS	24,284	0.853	0.826	69/1,270 = 5.43%	7 m 15 s	17.9
	DIST	742,593	0.185	0.171	639/1,270 = 50.31%	25 m 5 s	2.0
	SSIMP	1,879,896	0.145	0.070	1,111/1,270 = 87.48%	116 m 38 s	3.7
	FAPI	44,498	–	0.899	84/1,270 = 6.61%	139 m 10 s	187.6
GFR (AFR)	PRED-LD	51,633	0.838	0.585	98/1,160 = 8.44%	29 m 16 s	34.0
	RAISS	6,465	0.405	0.121	18/1,160 = 1.55%	9 m 27 s	87.7
	DIST	743,044	0.068	0.022	527/1,160 = 45.43%	9 m 17 s	0.8
	SSIMP	3,245,785	0.210	0.070	1,139/1,160 = 98.18%	153 m 50 s	2.8
	FAPI	47,232	–	0.514	83/1,160 = 7.15%	164 m 44 s	209.3
Epilepsy (EUR)	PRED-LD	516,837	0.927	0.873	72,589/73,759 = 98.41%	15 m 24 s	1.8
	RAISS	148,826	0.872	0.737	54,835/73,759 = 74.34%	60 m 22 s	24.3
	DIST	451,355	0.943	0.889	71,503/73,759 = 96.94%	232 m 28 s	30.9
	SSIMP	1,627,415	0.945	0.892	73,109/73,759 = 99.11%	1,096 m 24 s	40.4
	FAPI	182,768	–	0.922	66,251/73,759 = 89.82%	131 m 21 s	43.1
Colorectal Cancer (EAS)	PRED-LD	473,097	0.862	0.815	12,658/17,240 = 73.42%	9 m 15 s	1.2
	RAISS	240,299	0.877	0.770	8,839/17,240 = 51.27%	10 m 2 s	2.5
	DIST	689,032	0.759	0.506	14,028/17,240 = 81.36%	93 m 50 s	8.2
	SSIMP	1,783,481	0.850	0.787	16,512/17,240 = 95.77%	185 m 1 s	6.2
	FAPI	312,408	–	0.912	10,851/17,240 = 62.94%	92 m 6 s	17.7
Double Eyelid (EAS)	PRED-LD	367,707	0.664	0.499	4,709/7,820 = 60.21%	7 m 48 s	1.3
	RAISS	238,279	0.705	0.470	4,418/7,820 = 56.49%	7 m 53 s	2.0
	DIST	716,061	0.552	0.282	7,261/7,820 = 92.85%	64 m 20 s	5.4
	SSIMP	1,810,525	0.815	0.684	7,775/ 7,820 = 99.42%	183 m 27 s	6.1
	FAPI	279,728	–	0.760	3,437/7,820 = 43.95%	135 m 29 s	29.1
CAD (EUR)	PRED-LD	779,429	0.935	0.908	120,683/139,282 = 86.64%	49 m 2 s	3.8
	RAISS	139,507	0.892	0.812	74,897/139,282 = 53.77%	112 m 44 s	48.5
	DIST	251,662	0.905	0.872	112,862/139,282 = 81.03%	135 m 10 s	32.2
	SSIMP	1,450,754	0.867	0.835	136,174/139,282 = 97.76%	1,877 m 55 s	77.7
	FAPI	170,791	–	0.239	98,148/139,282 = 70.46%	142 m 7 s	49.9
Mean performance	**PRED-LD**	**338,803**	**0.817**	**0.728**	**49.90%**	**20 m 12 s**	**8.2**
	RAISS	**130,089**	**0.752**	**0.619**	**35.86%**	**27 m 53 s**	**25.0**
	DIST	**634,241**	**0.527**	**0.436**	**72.72%**	**76 m 15 s**	**10.5**
	SSIMP	**1,944,061**	**0.581**	**0.498**	**93.25%**	**428 m 9 s**	**18.1**
	FAPI	**168,365**	**–**	**0.744**	**38.86%**	**136 m 21 s**	**93.5**

Bold values denote the average performance across the datasets

Conclusions

PRED-LD offers an efficient method that performs GWAS summary statistics imputation. We showed that it is significantly faster compared to other methods, being at the

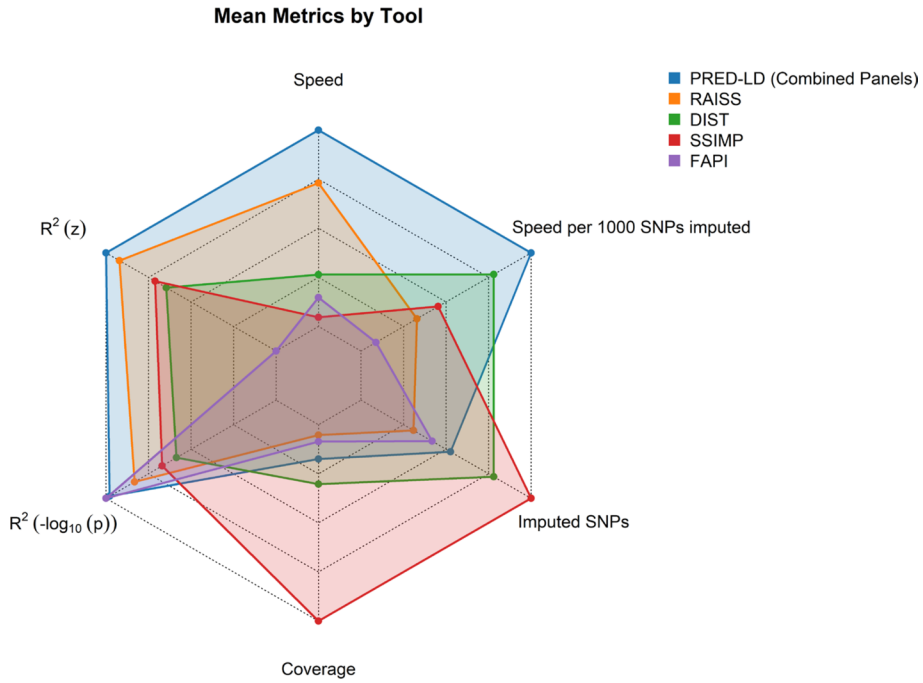

Fig. 4 Radar plot comparing the overall summary statistics imputation performance across the GWAS datasets. PRED-LD demonstrates the highest accuracy and coverage ratio and it is faster compared to the other tools. Only DIST and SSIMP outperform it in terms of imputation coverage and number of imputed SNPs

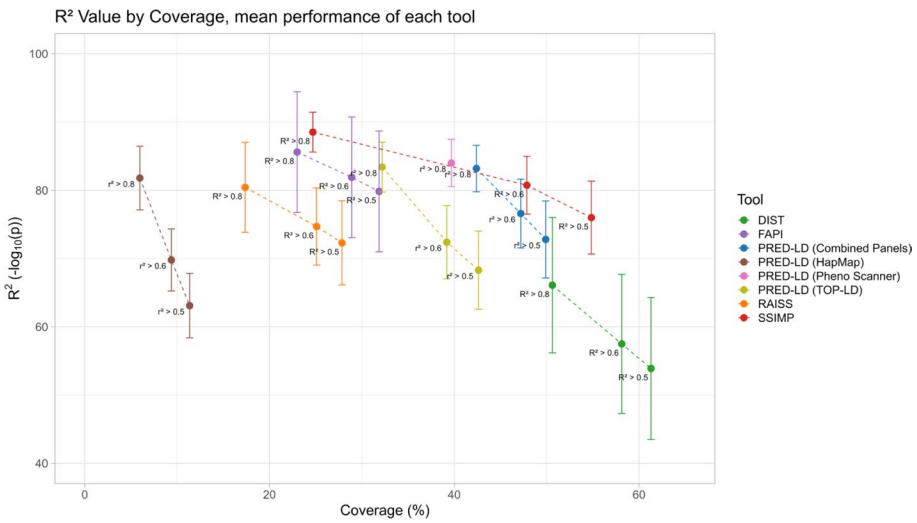

Fig. 5 Plot showing the inverse relationship of accuracy and coverage across different r^2 and R^2 thresholds for PRED-LD (combined and separate panels) and the compared tools, respectively. The results are obtained from the GWAS datasets as described in the text. We show the mean and the standard error of the mean for the predictions across the different datasets. PRED-LD allows both for a definition of a strict LD threshold or a lower one, resulting either in a more accurate imputation or a broader coverage, respectively. RAISS achieves high accuracy, but with the least coverage across all choices of thresholds. FAPI exhibits high R^2, but with more limited coverage compared to SSIMP, DIST and PRED-LD. Notably, SSIMP offers the highest accuracy and coverage ratio for $R^2 > 0.5$ and $R^2 > 0.6$ at the expense of execution time (see also Table 4). DIST delivers the best coverage overall, at the expense of the lowest R^2 among the tools evaluated

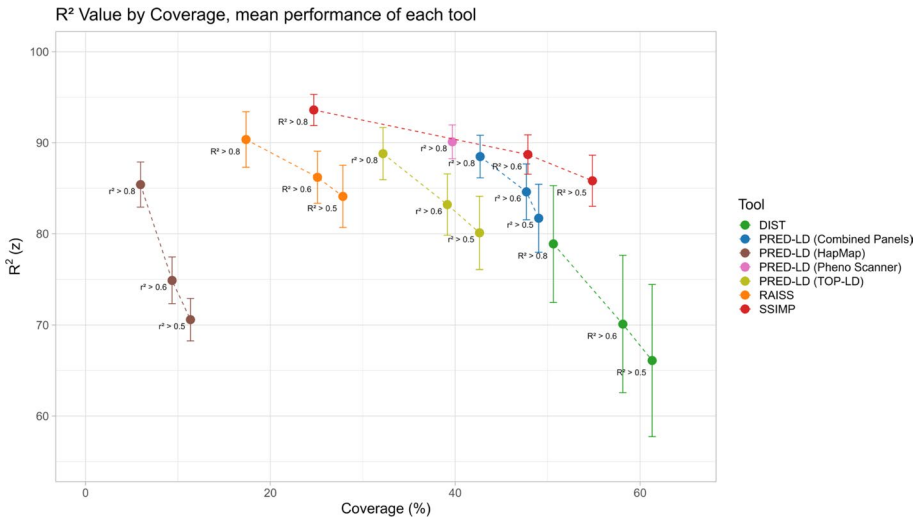

Fig. 6 Plot showing the inverse relationship of accuracy (*z*-values) and coverage across different r^2 and R^2 thresholds for PRED-LD (combined and separate panels) and the compared tools, respectively. The results are comparable to those of Fig. 5, with the absence of FAPI which reports only *p*-values. Regarding PRED-LD, the combined panels option demonstrates a high accuracy-coverage ratio, whilst when selecting the Pheno Scanner panel, PRED-LD exhibits the best results among all tools and PRED-LD panels. RAISS achieves high accuracy but moderate coverage in all thresholds. SSIMP offers the highest accuracy and coverage for $R^2 > 0.5$ and $R^2 > 0.6$. DIST delivers the best coverage overall, but lower R^2 among RAISS, SSIMP, and PRED-LD (in all its panels)

Table 5 Comparison of the mean performance of PRED-LD, DIST and SSIMP when the results of the latter two are filtered using (i) the reported coefficient of determination and keeping only the imputed SNPs with reported $R^2 > 0.5$, (ii) retaining the same number of imputed SNPs with PRED-LD (ranked in descending order of R^2 for DIST and SSIMP) and (iii) considering only the common imputed SNPs of PRED-LD across all intersection combinations. The three methods show comparable performance across all metrics except for speed. Execution time in other methods cannot be reduced since the R^2 can only be calculated after the imputation is performed

Filter	Tool	Imputed SNPs	$R^2(z)$	$(R^2 - \log_{10}(p))$	Imputation percentage in masked SNPs (%)	Time
$R^2 > 0.5$ (DIST, SSIMP) and $r^2 > 0.5$ (PRED-LD)	DIST	314,074	0.661	0.539	61.30	76 m 15 s
	SSIMP	310,107	0.858	0.760	54.84	428 m 9 s
	PRED-LD	338,803	0.817	0.728	49.90	20 m 12 s
Same number of imputed SNPs of PRED-LD	DIST	338,803	0.765	0.628	52.69	76 m 15 s
	SSIMP		0.879	0.808	53.86	428 m 9 s
	PRED-LD		0.817	0.728	49.90	20 m 12 s
Common SNPs (DIST-SSIMP-PRED-LD)	DIST	158,650	0.785	0.666	46.13	76 m 15 s
	SSIMP		0.837	0.737		428 m 9 s
	PRED-LD		0.816	0.725		20 m 12 s
Common SNPs (DIST-PRED-LD)	DIST	162,354	0.787	0.666	46.47	76 m 15 s
	PRED-LD		0.816	0.725		20 m 12 s
Common SNPs (SSIMP-PRED-LD)	SSIMP	219,732	0.837	0.739	49.39	428 m 9 s
	PRED-LD		0.817	0.727		20 m 12 s

same time equally accurate. The simplicity of the method offers a number of additional significant advantages. First, this approach allows both for the definition of a strict r^2 LD threshold or a lower one, resulting either in a more accurate imputation or a broader coverage, respectively. The user may perform an imputation and then filter the results according to the respective needs. A high threshold for the r^2 will produce smaller coverage but higher accuracy, whereas using a lower threshold will yield low accuracy but increased coverage. Second, the method can use and combine different reference panels with ease in a wide range of populations, since for each imputation only information from one SNP is used. When the computation time is an essential parameter, the user may choose one of the panels and significantly speed up the calculations (3 to 10 times faster compared to the default option), with a slight decrease in accuracy. Of course, this means that the method can easily take advantage of additional reference panels that will appear in the future.

The only downside of the method, compared to methods that use multiple markers, seems to be a somewhat reduced coverage (at least compared to SSIMP). This is easily understood if we imagine a situation in which none of the typed markers pass the r^2 threshold, but there are several markers that may contribute information through the multivariate normal distribution. However, to perform a fair comparison, we have shown that altering the r^2 threshold for PRED-LD, or the R^2 threshold for DIST and SSIMP, results in imputation accuracies and coverages that are comparable, with PRED-LD still clearly outperforming these methods in terms of speed. The methods using the multivariate normal distribution need additional computations in order to regularize the variance–covariance matrix, or to avoid multicollinearity. Thus, it seems that the single marker approach with the direct imputation from Eq. (2) or Eq. (3) is preferable, especially when the SNPs in the GWAS and the panel are dense.

We also need to comment on the use of different panels. A direct comparison of the methods that use different panels is not so easy to perform, given that each tool uses different file formats and specifications, but some observations can be made. For instance, we have shown that PRED-LD performance increases with larger and denser panels. On the other hand, the multiple marker methods use panels of different size (DIST uses the smaller one, whereas FAPI and SSIMP use a larger one, even compared to TOP-LD) but this does not directly translate to increased performance; they all seem to be less efficient compared to PRED-LD.

Finally, we need to emphasize that PRED-LD imputes beta coefficients and standard errors, from which the other statistics can be produced (z-values or p-values). Furthermore, the imputation accuracy of PRED-LD is high, either regarding z-values or p-values. In contrast, other methods (RAISS and DIST) can impute only z-scores or p-values, whereas FAPI imputes only p-values. This may be restrictive in some cases where the downstream analysis requires beta coefficients. Thus, PRED-LD is suitable both for applications that can utilize p-values, such as gene-based tests, as well as for applications in which the effect size is needed, such as random effects meta-analysis.

Taken together, PRED-LD is an optimal choice for large-scale GWAS imputation tasks, in which both computation efficiency and imputation accuracy are critical. The online version of PRED-LD can assist users in obtaining LD information from various sources and performing various imputation tasks with ease, without the need to download reference panels

for multiple populations and chromosomes. PRED-LD will be continuously updated, for instance by adding new reference panels, or performing optimizations in speed (parallelization and so on), and we believe that it will be widely used. In particular, we are planning to incorporate PRED-LD in various tools that will facilitate, for instance, meta-analysis allowing for non-overlapping sets of variants, in tools that perform analysis of multiple traits, or for statistical fine-mapping of causal variants in GWAS.

Acknowledgements
The authors would like to thank the editor and the two reviewers whose comments and constructive criticism helped in improving the quality of the manuscript.

Author contributions
The authors confirm contribution to the paper as follows: study conception and design: P.B.; data collection: G.M., A.M.; software implementation: G.M., analysis and interpretation of results: G.M., A.M., P.K. PB; draft manuscript preparation: G.M., A.M., P.K., P.B. All authors reviewed the results and approved the final version of the manuscript.

Funding
This project is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union –NextGenerationEU. The research conducted by Georgios A. Manios is carried out within the operating framework of the Center of Research Innovation and Excellence of the University of Thessaly and was funded by the Special Account of Research Grants of University of Thessaly.

Availability of data and materials
All data supporting this study are available at the following links, in the order provided in Table 1: ADHD (EUR): https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs001869/analyses/Fulldata/, UACR (EUR): http://ckdgen.imbi.uni-freiburg.de/files/Li2017/Published_UACR_EA.csv.gz, GFR (EUR): http://ckdgen.imbi.uni-freiburg.de/files/Li2017/Published_eGFRcrea_DM_EA.csv.gz, GFR (AFR): http://ckdgen.imbi.uni-freiburg.de/files/Li2017/Published_UACR_AA.csv.gz, Epilepsy (EUR): http://www.epigad.org/gwas_ilae2018_16loci/CAE_BOLT-LMM_final.gz, Colorectal Cancer (EAS): http://jenger.riken.jp/en/result, ID: 11, Study: Colorectal Cancer, Double Eyelid (EAS): https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-27145-2/MediaObjects/41598_2018_27145_MOESM6_ESM.txt, Coronary Artery Disease (EUR): http://www.cardiogramplusc4d.org/media/cardiogramplusc4d-consortium/data-downloads/UKBB.GWAS1KG.EXOME.CAD.SOFT.META.PublicRelease.300517.txt.gz, The web server of PRED-LD is freely available at https://compgen.dib.uth.gr/PRED_LD/. The source code of PRED-LD is available through a public GitHub repository at https://github.com/pbagos/PRED-LD.

Availability and requirements
Project name: PRED-LD. Project home page: https://github.com/pbagos/PRED-LD. Operating system(s): Platform independent. Programming Language: Python. Other requirements: Python 3.8.2 or higher, pandas 1.5.3, NumPy 1.24.1, Dask 2023.9.1. License: GNU GPL-3.0. Any restrictions to use by non-academics: None.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 October 2024 Accepted: 21 March 2025
Published online: 16 April 2025

References

1. Seng KC, Seng CK. The success of the genome-wide association approach: a brief story of a long struggle. Eur J Hum Genet. 2008;16:554–64.
2. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
3. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu S-A. DIST: direct imputation of summary statistics for unmeasured SNPs. Bioinformatics. 2013;29:2925–7.
4. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics. 2014;30:2906–14.
5. Kwan JSH, Li M-X, Deng J-E, Sham PC. FAPI: fast and accurate *P*-value imputation for genome-wide association study. Eur J Hum Genet. 2016;24:761–6.
6. Rüeger S, McDaid A, Kutalik Z. Improved imputation of summary statistics for admixed populations. BioRxiv. 2018;4:1158.

7. Julienne H, Shi H, Pasaniuc B, Aschard H. RAISS: robust and accurate imputation from summary statistics. Bioinformatics. 2019;35:4837–9.
8. Siva N. 1000 Genomes project. Nat Biotechnol. 2008;26:256–7.
9. Lee D, Bigdeli TB, Williamson VS, Vladimirov VI, Riley BP, Fanous AH, et al. DISTMIX: direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. Bioinformatics. 2015;31:3099–104.
10. Lee D, Bacanu S-A. GAUSS: a summary-statistics-based R package for accurate estimation of linkage disequilibrium for variants, Gaussian imputation, and TWAS analysis of cosmopolitan cohorts. Bioinformatics. 2024;40(4):btae203. https://doi.org/10.1093/bioinformatics/btae203.
11. Chatzinakos C, Lee D, Cai N, Vladimirov VI, Webb BT, Riley BP, et al. Increasing the resolution and precision of psychiatric genome-wide association studies by re-imputing summary statistics using a large, diverse reference panel. Am J Med Genet B Neuropsychiatr Genet. 2021;186:16–27.
12. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang HM, et al. The international HapMap project. 2003.
13. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype–phenotype associations. Bioinformatics. 2016;32:3207–9.
14. Huang L, Rosen JD, Sun Q, Chen J, Wheeler MM, Zhou Y, et al. TOP-LD: a tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. Am J Human Genetics. 2022;109:1175–81.
15. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
16. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590:290–9.
17. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930–4.
18. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31:3555–7.
19. Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. Nat Rev Genet. 2004;5:89–100.
20. Ackerman H, Usen S, Mott R, Richardson A, Sisay-Joof F, Katundu P, et al. Haplotypic analysis of the TNF locus by association efficiency and entropy. Genome Biol. 2003;4:1–13.
21. Lewontin RC. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics. 1964;49:49.
22. Oehlert GW. A note on the delta method. Am Stat. 1992;46:27–9.
23. Duan K, Chen J, Calhoun VD, Lin D, Jiang W, Franke B, et al. Neural correlates of cognitive function and symptoms in attention-deficit/hyperactivity disorder in adults. Neuroimage Clin. 2018;19:374–83.
24. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014;42:D975–9.
25. Li M, Li Y, Weeks O, Mijatovic V, Teumer A, Huffman JE, et al. SOS2 and ACP1 loci identified through large-scale exome chip analysis regulate kidney development and function. J Am Soc Nephrol. 2017;28:981–94.
26. 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. Nat Genet. 2019;51:1339–48.
27. Epilepsies ILAEC on C. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun. 2018;9:5269.
28. Tanikawa C, Kamatani Y, Takahashi A, Momozawa Y, Leveque K, Nagayama S, et al. GWAS identifies two novel colorectal cancer loci at 16q24.1 and 20q13.12. Carcinogenesis. 2018;39(5):652–60. https://doi.org/10.1093/carcin/bgy026.
29. Endo C, Johnson TA, Morino R, Nakazono K, Kamitsuji S, Akita M, et al. Genome-wide association study in Japanese females identifies fifteen novel skin-related trait associations. Sci Rep. 2018;8:8974.
30. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat Genet. 2017;49:1385–91.

Publisher's Note