Phenotype projections accelerate biobank-scale GWAS

Michael Zietz[1], Undina Gisladottir[1], Kathleen LaRow Brown[1], and Nicholas Tatonetti*[1,2]

[1]Department of Biomedical Informatics, Columbia University
[2]Department of Computational Biomedicine, Cedars-Sinai

November 20, 2023

Abstract

Understanding the genetic basis of complex disease is a critical research goal due to the immense, worldwide burden of these diseases. Pan-biobank genome-wide association studies (GWAS) provide a powerful resource in complex disease genetics, generating shareable summary statistics on thousands of phenotypes. Biobank-scale GWAS have two notable limitations: they are resource-intensive to compute and do not inform about hand-crafted phenotype definitions, which are often more relevant to study. Here we present Indirect GWAS (indGWAS), a summary-statistic-based method that addresses these limitations. IndGWAS computes GWAS statistics for any phenotype defined as a linear combination of other phenotypes. Our method can reduce runtime by an order of magnitude for large pan-biobank GWAS, and it enables ultra-rapid (less than one minute) GWAS on hand-crafted phenotype definitions using only summary statistics. Overall, this method advances complex disease research by facilitating more accessible and cost-effective genetic studies using large observational data.

1 Introduction

Uncovering the genetic basis of complex diseases remains challenging despite decades of research [1]. Genome-wide association study (GWAS) is an important method for studying complex disease genetics. GWAS summary statistics provide a convenient summarization of the associations between genetic variation and complex disease phenotypes [2]. Because they do not contain protected genetic or health information, GWAS summary statistics are shared widely [3] and are used for purposes including heritability and genetic correlation estimation [4, 5], polygenic risk scoring [6, 7], and Mendelian randomization [8].

Performing GWAS using large-scale resources like the UK Biobank (UKBB) can be challenging and time-consuming for researchers. Hurdles to this research include restricted access to personally-identifying genetic and health information, resource-intensive computations, necessary expertise in defining phenotype data and filtering genetic data, and evolving GWAS best-practices. Pan-biobank GWAS (pbGWAS)—studies in which GWAS are run on hundreds to thousands of phenotypes—provide a solution to these challenges [9–12]. These

*Corresponding author: nicholas.tatonetti@cshs.org

studies pre-compute and publicly release GWAS summary statistics for all phenotypes in a dataset, eliminating the need for many GWAS to be duplicated. PbGWAS provide a useful starting point for genetic analysis of complex phenotypes, but they have two major limitations: they are highly resource intensive to compute, and they only release summary statistics on a limited number of predefined phenotypes. As a consequence, any new GWAS method, phenotype definition, or cohort selection requires a GWAS to be conducted from scratch, representing a major cost to research.

Here, we present Indirect GWAS (indGWAS), a method that aims to resolve these two limitations of pbGWAS. IndGWAS is a mathematical method for computing GWAS summary statistics for a phenotype defined as a linear combination of other phenotypes, using only summary information about the input phenotypes. Our approach has two aims: to increase the speed of pbGWAS and to enable researchers to compute GWAS for custom phenotype definitions using only pbGWAS summary statistics. IndGWAS enables large speed improvements for pbGWAS, potentially reducing computation time by an order of magnitude or more for the largest pan-biobank studies by reducing the dimension of phenotype data. Additionally, our method enables researchers to compute GWAS for arbitrarily-defined phenotypes using pbGWAS summary statistics extremely rapidly (less than one minute). Together, these improvements accelerate the computation of pbGWAS and increase the usefulness of pbGWAS results for studying specific traits and diseases.

2 Results

We created indGWAS, a method for performing GWAS on a linear combination of phenotypes using only summary statistics. We distinguish our indirect approach from a direct approach as follows: given y, a linear combination of phenotypes x_1, ..., x_m, a direct GWAS evaluates y for every individual in the research cohort and performs GWAS on it using a standard method. Conversely, our indirect method computes GWAS for y using only summary statistics about x_i, the input phenotypes. We used the closed-form solution to ordinary least squares (OLS) regression to derive the coefficient and standard error terms for an arbitrary linear combination of phenotypes (see Methods). IndGWAS is applicable to any GWAS method that can be framed as linear regression, including linear mixed models such as SAIGE and Regenie.

2.1 Validation of the method

We verified the mathematical correctness of our method by comparing direct to indirect approaches for several GWAS methods using data from the UKBB. First, we randomly sampled 100,000 individuals, 100,000 variants, and 10 quantitative phenotypes, and we generated 100 random projections of these phenotypes. Then, we computed GWAS summary statistics for the 100 random phenotypes using both direct and indirect approaches, applying four different GWAS methods: univariate OLS (only genotype and intercept), multivariate OLS (genotype, intercept, age, sex, 10 genetic PCs), SAIGE (with covariates), and Regenie (with covariates). For each method, we compared GWAS chi-squared statistics between direct and indirect approaches to verify that indGWAS produces equivalent results. All methods showed excellent correspondence between summary statistics ($R^2 > 0.99$, Figure 1), indicating that indGWAS is mathematically correct and applicable to both linear and linear mixed models.

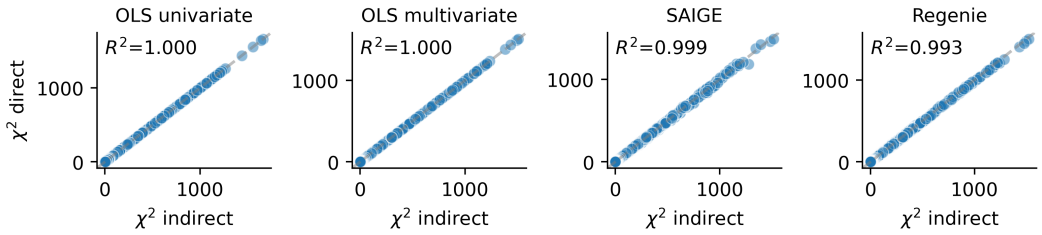

Figure 1: Validation of indGWAS using four GWAS methods. GWAS chi-squared statistics are compared between direct and indirect approaches for 100 randomly projected phenotypes, 100,000 genetic variants using data from the UK Biobank. Each point represents one genetic variant and one phenotype (10,000,000 total points in each facet). R^2 values indicate goodness of fit between the direct and indirect statistics.

2.2 IndGWAS enables faster pan-biobank GWAS

IndGWAS affords an approximation to traditional pbGWAS when quantitative phenotypes are correlated. In short, we reduce the dimensionality of a large set of phenotypes, perform direct GWAS on the latent space, then reconstruct full-dimensional summary statistics using indGWAS. By reducing the number of GWAS that must be computed using slower methods like linear mixed models, our approach saves a large fraction of computation time.

As an example, we applied our approach to 1378 quantitative phenotypes from UKBB (see Methods). We first determined how well these phenotypes could be compressed using principal component analysis (PCA), a linear dimensionality reduction method. We found that a large fraction (roughly 80%) of these quantitative phenotypes could be reconstructed with $R^2 > 0.8$ using 50% of the PCs, and with $R^2 > 0.9$ using 75% of the PCs (Figure 2A), confirming that these real phenotypes can be compressed using a linear method while retaining high reconstruction fidelity.

Next, we evaluated the correspondence between GWAS summary statistics as more principal components (PCs) are included in indGWAS (Figure 2B). We found that indGWAS is able to approximate statistically significant variant p-values using even very small fractions of the PCs, achieving very high recall (> 0.995) at all PC latent dimensions and good precision that improves as the latent dimension increases (Table 1).

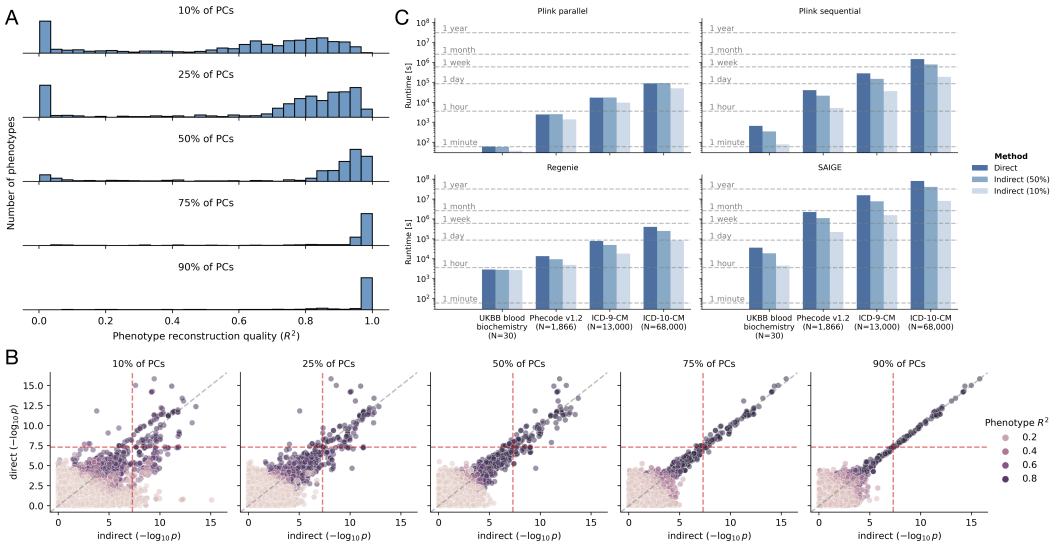

Figure 2: IndGWAS accelerates pan-biobank GWAS using a low-dimensional approximation. **A.** Real phenotypes can be compressed using a small fraction of PCs (using 1378 real quantitative phenotypes from the UKBB). **B.** GWAS summary statistics (here showing negative $\log_{10}$ p-values) from indGWAS become better approximations of the true (direct) summary statistics as more PCs are used. **C.** Using a low-dimensional indGWAS saves meaningful computation time for large multi-phenotype GWAS. Runtimes are quoted for a cohort of 100,000 individuals and 100,000 genetic variants.

	Time saved		# of significant variants			indGWAS performance		
% of PCs	**Hours**	**%**	**Direct**	**Indirect**	**Shared**	**Precision**	**Recall**	**F1**
10	7.3	87	2969	6496	2957	0.455	0.996	0.625
25	6.1	72	3354	4588	3342	0.728	0.997	0.842
50	3.9	47	3539	4212	3527	0.837	0.997	0.910
75	1.8	22	3646	3881	3634	0.936	0.997	0.966
90	0.5	6	3703	3756	3691	0.983	0.997	0.990

Table 1: Analysis of indGWAS performance across various dimensionality reductions. This analysis considered 100,000 variants, 100,000 samples, 1378 phenotypes, and used Plink 2 with sequential GWAS. Variants were considered genome-wide significant if their p-values were less than 5×10^{-8}. Time saved refers to the approximate computation time savings relative to a fully direct GWAS (approximately 8.4 hours under these conditions).

Finally, we evaluated the speed improvements afforded by indGWAS to various GWAS methods. We compared parallel OLS (Plink 2 runs multiple phenotypes simultaneously under certain ideal conditions), sequential OLS (Plink 2, one phenotype at a time), SAIGE, and REGENIE. For each method, we compared using the method directly for all phenotypes to using the method indirectly with indGWAS and various latent dimensions. As an example,

in the SAIGE 50% comparison, we compressed the 1378 phenotypes down to a latent space with half the dimensionality of the original phenotype space (N=689) using PCA, performed SAIGE GWAS on the latent space, then reconstructed full summary statistics using indGWAS. In each comparison involving indGWAS, the PCA and indGWAS computation times are included in the overall method runtime. Each method's quoted time accounts for 100,000 variants and 100,000 samples. We ran each method for 10, 50, and 100 phenotypes to evaluate their scaling behavior, then extrapolated runtimes from these points. We found that all methods were accelerated by using indGWAS with a sufficiently low latent space size (Figure 2C). As an example, the computation time for 68,000 ICD-10-CM GWAS using SAIGE on our data would be approximately 8×10^7 seconds, or roughly 926 days. Meanwhile, using indGWAS with 10 % of the PCs would reduce this computation time to approximately 8×10^6 seconds, or roughly 92 days.

2.3 IndGWAS performs GWAS on custom phenotype definitions using summary statistics

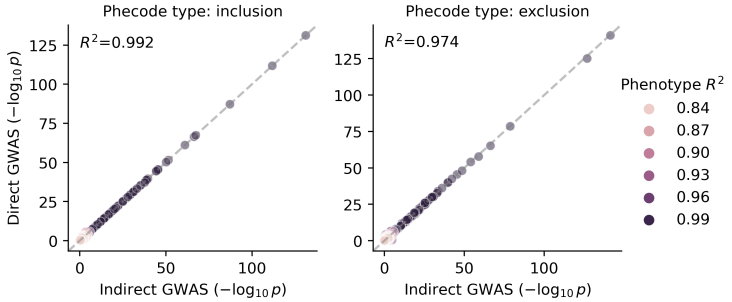

Figure 3: GWAS summary statistics for phecodes generated using indGWAS with summary statistics for 2616 binary and quantitative phenotypes. R^2 values indicate goodness of fit between the direct and indirect summary statistics.

The final goal of this work is to unlock pbGWAS data for the analysis of custom phenotype definitions that were not included in the pbGWAS. To address this challenge, we propose approximating a map between the pbGWAS phenotype space and the custom phenotype definition using linear regression, reconstructing the summary statistics with indGWAS, and using the goodness-of-fit of the regression to measure our confidence in the fidelity of the resulting GWAS summary statistics.

To illustrate this method, we considered phecodes, which group ICD codes into higher-level categories [13]. To explore the effects of more complex phenotype definitions, we compared phecodes using only inclusion definitions to using both inclusion and exclusion definitions. Phecode inclusion definitions are a simple maximum operation across ICD-10 codes (e.g. phecode 499: cystic fibrosis is present if a patient has any ICD-10 codes that start with E84). Exclusion definitions are designed to exclude individuals from the control group when they are not cases but have codes that are very similar to case inclusion codes and could represent mislabeling (e.g. exclude type 1 diabetes mellitus patients from type 2 diabetes mellitus control group). Inclusion/exclusion definitions are equivalent to inclusion definitions among individuals who are included in the analysis, but they result in different

cohorts for each phenotype, thereby violating the linearity assumption of indGWAS in a plausible scenario akin to real genetic analyses.

We used linear regression to approximate a linear map from the 2616 binary and quantitative phenotypes (see Methods) to the 567 resulting phecodes (for inclusion and inclusion/exclusion separately). For each phecode, we evaluated the phenotypic goodness-of-fit using R^2, then computed GWAS summary statistics using both direct and indirect approaches, each run with both inclusion and inclusion/exclusion definitions. Overall, both types of phecodes were well-fit using the features, with all phenotype fit R^2 values over 0.84. While both inclusion and inclusion/exclusion were well-fit, the missingness introduced by exclusion criteria reduced the correspondence of the GWAS summary statistics slightly (R^2 value of the negative log p-values reduced from 0.992 to 0.974 by adding exclusion codes, Figure 3).

2.4 Sensitivity to nonlinear phenotype definitions

Hand-crafted phenotype definitions used in observational research are often nonlinear, violating the indGWAS model. Nonetheless, many phenotypes can be approximated sufficiently using a linear model that indGWAS may be appropriate. For example, while none of the phecodes are strictly linear functions of ICD codes, their definitions can be approximated by linear models sufficiently well that indGWAS summary statistics match well (Figure 3). We evaluated how robust indGWAS is to nonlinear phenotype definitions by comparing phenotype fit quality (measured using mean squared error) to GWAS summary statistic correspondence (measured using the F1 score of genome-wide significant variants between direct and indirect GWAS methods), and we plot the results in Figure 4. We used F1 scores because variants with extreme p-values are the most relevant for this comparison. Phenotypes that are very well approximated using linear regression have low mean squared errors, and phenotypes whose indirect p-values match their direct p-values have high F1 scores.

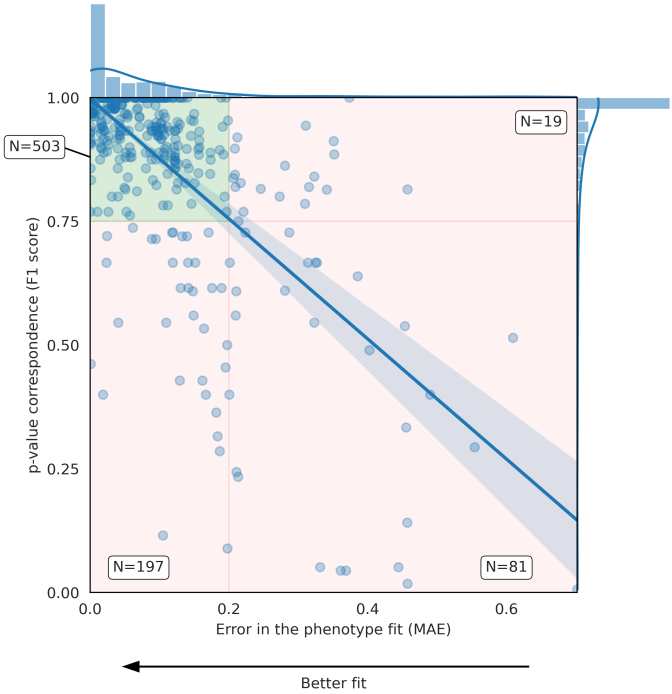

Figure 4: IndGWAS summary statistics correspond better for phenotypes that are well-fit using linear regression. Each point represents one phenotype. 1378 quantitative phenotypes are included, each with 100,000 tested variants on 100,000 samples using OLS GWAS in Plink. Phenotypes with zero genome-wide significant variants in either direct or indirect methods are not shown as F1 scores are undefined in these cases. We evaluated phenotype fit using mean absolute error (MAE) and evaluated the correspondence between direct and indirect summary statistics using an F1 score of genome-wide significant variants. We counted the number of variant effect sizes in four bins, according to cutoffs of 0.2 for MAE and 0.75 for F1. The line indicates a linear regression fit, and the shaded region indicates a bootstrap 95% confidence interval.

2.5 IndGWAS result quality across phenotype categories

We scrutinized the performance of indGWAS across various compression sizes using phenotype categories provided by the UKBB. The 1378 quantitative phenotypes we investigated fell into 48 different categories. Most phenotypes had few, if any, genome-wide significant variants, leading to undefined F1 scores. Of the 48 categories, only 12 had at least one compression fraction with more than two computable F1 scores. The indirect-direct p-value correspondences for these 12 categories are shown in Figure 5. Overall, performance increased broadly across categories as the latent dimension increased.

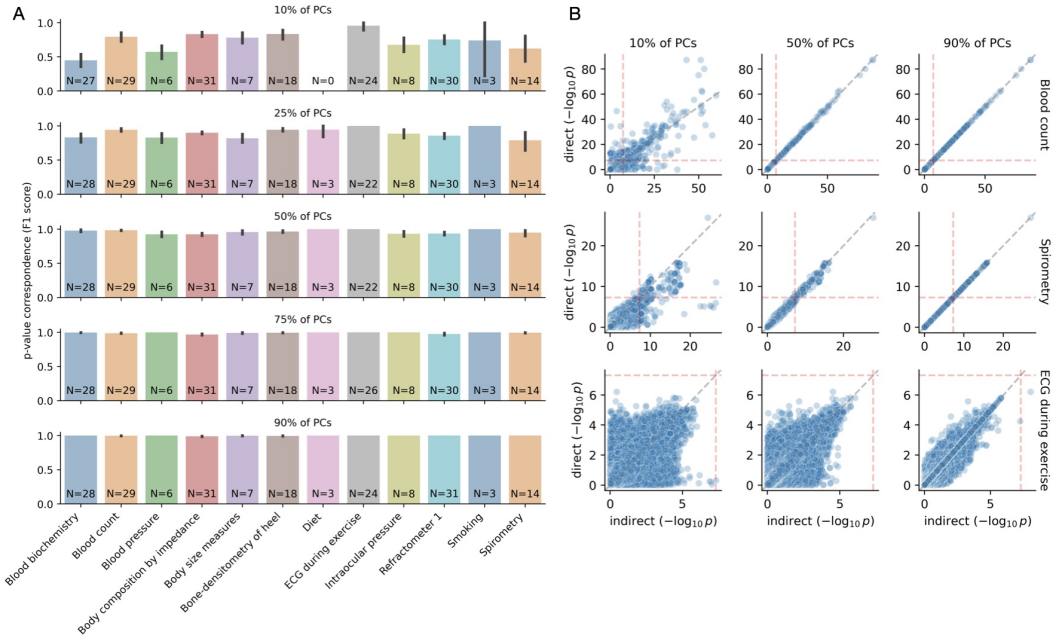

Figure 5: IndGWAS performance across phenotype categories defined by the UKBB. **A.** Each count (e.g. "N=27") gives the number of phenotypes within this category that had defined F1 scores for this compression fraction (e.g. "10% of PCs"). **B.** Select comparisons between direct and indirect GWAS summary statistics (here showing negative log p-values). IndGWAS becomes more accurate as a larger latent dimension is used and as p-values become more extreme.

3 Discussion

In this work, we introduced indGWAS, a mathematical method for reconstructing GWAS summary statistics for linear phenotype projections using only summary information. Our method aims to improve the speed and flexibility of pbGWAS by improving computation time and allowing the analysis of custom phenotype definitions, beyond the predefined pbGWAS phenotypes. To speed pbGWAS computation, we propose reducing the dimensionality of the phenotypes using a linear method such as PCA, performing GWAS on the lower-dimensional space, and then reconstructing high-dimensional GWAS summary statistics using the linear map from latent to original space. To allow the analysis of custom phenotypes, we propose using linear regression to approximate the chosen phenotype with the phenotypes included in a pbGWAS as features, applying indGWAS to reconstruct summary statistics, and using the goodness-of-fit for each phenotype as a measure of the fidelity in the summary statistics.

The biggest limitation of indGWAS is its reliance on linear projections. This restricts both the GWAS methods that can be used (e.g. no logistic regression or generalized linear mixed models) as well as the dimensionality reduction methods that can be used when attempting to accelerate pbGWAS (e.g. cannot use powerful nonlinear dimensionality reduction methods). Nonlinear functions such as the Boolean operators are typical for hand-crafted phenotype

definitions. While linear approximations may or may not perform well for all phenotypes, nonetheless the indGWAS approach can quantify the confidence we have in GWAS summary statistics by measuring the phenotypic goodness-of-fit (Figure 4). Moreover, as linear and logistic regressions produce summary statistics that match closely for variants with small or moderate effect [5], indGWAS's limitation here is not critical.

IndGWAS performs best when every phenotype and every feature phenotype use the same study cohort, as found in the comparison between inclusion and inclusion/exclusion phecode definitions (Figure 3). Future work that uses indGWAS should attempt to maintain the cohort as much as possible. An applicable method for this challenge is QTPhenProxy, in which the phenotype definition is replaced by an estimated phenotype probability using regression on other variables [14]. For the phecode case/control example, a regression could be fit on case/control definitions before being evaluated on the whole cohort. Individuals who would previously have been excluded would likely be assigned moderate disease probabilities, while cases would receive high probabilities and controls receive low probabilities.

We evaluated the runtime of various methods with and without indGWAS, and using a variety of latent space sizes. Given a sufficient reduction in phenotype dimensionality, indGWAS led to large speed improvements over fully direct methods (Figure 2). However, different applications could be more or less advantageous to indGWAS. For example, a study of all pairs of phenotypes would be much faster using indGWAS, as phenotype pairs are simple functions of the phenotypes themselves, and pairs could be well-approximated linearly. Conversely, a study of a small number of rare binary phenotypes would be unlikely to benefit much from the indGWAS approach, as these phenotypes are unlikely to be well-approximated linearly with few features. Additionally, our comparison did not use highly optimized parallel implementations such as SAIGE implemented in Hail [15] or REGENIE in Glow [16]. Our implementation of indGWAS could also be improved, and the majority of its runtime is spent reading and writing GWAS summary statistics from disk. Overall, we report timings to illustrate scaling behavior and to illustrate the advantages of using indGWAS, not to provide an exact reference for runtime expectations.

4 Methods

4.1 Indirect GWAS

Consider a regression involving a single genetic variant, a single phenotype, and N samples. Let $\boldsymbol{y}$ be the N-vector of phenotype values, $\boldsymbol{g}$ be the N-vector of genotype values (coded in any way), and $\boldsymbol{X}$ be the $N \times C$ matrix of covariates that includes an intercept and is assumed to be full-rank.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{g}\beta + \boldsymbol{\varepsilon}$$

Above, $\boldsymbol{\alpha}$ is the C-vector of covariate fixed effects, β is the scalar effect of the genotype on the phenotype, and $\boldsymbol{\varepsilon}$ is the N-vector of errors. Covariate effects can be removed using a residual projection matrix, $\boldsymbol{P} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Removing covariates results in residualized phenotype ($\tilde{\boldsymbol{y}} \equiv \boldsymbol{P}\boldsymbol{y}$), genotype ($\tilde{\boldsymbol{g}} \equiv \boldsymbol{P}\boldsymbol{g}$), and error ($\tilde{\boldsymbol{\varepsilon}} \equiv \boldsymbol{P}\boldsymbol{\varepsilon}$) vectors, and allows writing the regression more compactly.

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{g}}\beta + \tilde{\boldsymbol{\varepsilon}}$$

Estimates of the coefficient β and its standard error $\widehat{\mathrm{SE}(\hat{\beta})}$ can be written as follows:

$$\hat{\beta} = \frac{\tilde{\boldsymbol{g}}^T \tilde{\boldsymbol{y}}}{\tilde{\boldsymbol{g}}^T \tilde{\boldsymbol{g}}}$$

$$\widehat{\mathrm{SE}(\hat{\beta})}^2 = \frac{1}{N-C-1}\left(\frac{\mathrm{Var}(\tilde{y})}{\mathrm{Var}(\tilde{g})} - \hat{\beta}^2\right)$$

Now, suppose y can be written as a linear combination of some feature phenotypes. Let $\boldsymbol{Z}$ be an $N \times L$ matrix of feature phenotypes and let $\boldsymbol{\theta}$ be an L-vector of the projection coefficients.

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\theta}$$

Suppose that for every phenotype in $\boldsymbol{Z}$, the same genotype-phenotype regression has been performed. Let $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\omega}}$ be L-vectors holding coefficient estimates and standard errors, respectively. To estimate the standard error for $\hat{\beta}$, we can solve the equation for the standard error using $\hat{\boldsymbol{\omega}}$.

$$\mathrm{Var}(\tilde{g}_i) = \frac{\mathrm{Var}(\tilde{z}_i)}{\hat{\omega}_i^2(N-C-1) + \hat{\lambda}_i^{\,2}}$$

This equation for the genotypic partial variance is written in terms of variables that can be obtained from a GWAS study. It depends only on the phenotypic partial variance, the estimated coefficient, the estimated coefficient standard error, and the number of degrees-of-freedom ($N-C-1$). For indirect GWAS, we use the mean of these values across phenotypes as the genotype partial variance ($\mathrm{Var}(\tilde{g}) = \overline{\mathrm{Var}(\tilde{g}_i)}$). Unless cohorts are different for the different feature GWAS, this value will be the same for every phenotype. The final term that remains to be estimated is $\mathrm{Var}(\tilde{y})$, which can be computed using the partial covariance matrix of the feature traits, $\boldsymbol{K_{Z,Z}}$.

$$\mathrm{Var}(\tilde{y}) = \boldsymbol{\theta}^T \boldsymbol{K_{Z,Z}} \boldsymbol{\theta}$$

Finally, we can write the coefficient and standard error estimates for y as follows:

$$\hat{\beta} = \hat{\boldsymbol{\lambda}}^T \boldsymbol{\theta} \tag{1}$$

$$\widehat{\mathrm{SE}(\hat{\beta})} = \sqrt{\frac{1}{N-C-1}\left(\frac{\boldsymbol{\theta}^T \boldsymbol{K_{Z,Z}} \boldsymbol{\theta}}{\overline{\mathrm{Var}(\tilde{g}_i)}} - (\hat{\boldsymbol{\lambda}}^T \boldsymbol{\theta})^2\right)} \tag{2}$$

4.2 Empirical analyses

All analyses used genetic and phenotypic data from the UKBB. For genetic data, we used the UKBB's imputed variants version 3. All analyses used the same set of variants, which were selected using the following criteria: included in the 1000 Genomes Project phase 3 variants, <1% missingness (in 1000 Genomes and UKBB), minor allele frequency >5% (in 1000 Genomes and UKBB), no insertions or deletions, no ambiguous variants (i.e. A to T, T to A, C to G, or G to C). We randomly sampled 100,000 of the remaining variants to reduce overall computation time, and used the downsampled dataset for all but a few GWAS run to compare overall computation time between methods.

For sample selection and QC, we applied the following filters: individual must not have withdrawn their consent from the UK Biobank, individual must have at least one medical diagnosis (either self-reported, hospital inpatient, or outpatient), individual's reported sex must equal their genetic sex, individual must have been used in the computation of genetic principal components, the sample must neither exhibit sex chromosome aneuploidy nor be an outlier for heterozygosity or missingness, and the individual must have fewer than 10 third-degree relatives in the dataset. We used 12 covariates for each GWAS (excluding univariate OLS): age, sex, and the first 10 genetic principal components.

Quantitative phenotypes were selected using the data dictionary provided by the UK Biobank as of August 1, 2023. Specifically, we chose all phenotypes that met the following criteria: value type either continuous or integer, not a bulk data source, "unisex" trait, at minimum 75,000 participants, and we excluded all fields that contained genetic data (e.g. polygenic risk scores, genetic PCs, genotyping QC), information about data collection (e.g. number of samples attempted), and other QC information (e.g. accelerometer calibration). For fields that contained an array of data, each array item was treated as a separate field. We took the mean across instances for quantitative phenotypes. All quantitative phenotypes were mean imputed (to obtain a fully dense phenotype table), and transformed using the inverse rank normal transformation to ensure that all the traits are approximately distributed as a standard normal distribution for numerical stability.

For binary phenotypes, we used all three-letter ICD-10 codes which could be gleaned from the UK Biobank. These data are totally dense among individuals who appear in the data. Specifically, we gathered inpatient and outpatient records, self-reported medical conditions, and death register data, and we mapped all of these to ICD-10 using tables provided by the UK Biobank. We excluded all ICD-10 codes that had fewer than 100 occurrences in our data. All codes used as phenotypes, covariates, and filters are listed in the Supplementary Materials.

To evaluate indGWAS as a method to approximate summary statistics for phenotypes that were not included as part of a pbGWAS release, we used phecodes version 1.2, which maps ICD-10 to phecode groupings using both inclusion and exclusion criteria. Because we used summary diagnoses from the UK Biobank, we performed the phecode mapping using only single occurrences rather than the so-called rule of two that has been suggested for phecode mappings [17]. This approach was also used by the Pan-UK Biobank GWAS [10].

5 Code availability

IndGWAS is implemented in Python and C++ and is freely available at `https://github.com/tatonetti-lab/indirect-gwas`.

References

[1] Abdel Abdellaoui and Karin JH Verweij. Dissecting polygenic signals from genome-wide association studies on human behaviour. *Nature Human Behaviour*, 5(6):686–694, 2021.

[2] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[3] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1): D1005–D1012, 2019.

[4] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47 (3):291–295, 2015.

[5] Doug Speed and David J Balding. Sumher better estimates the snp heritability of complex traits from summary statistics. *Nature Genetics*, 51(2):277–284, 2019.

[6] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, 2018.

[7] Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7):1097–1103, 2021.

[8] Jean Morrison, Nicholas Knoblauch, Joseph H Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747, 2020.

[9] Wei Zhou, Jonas B Nielsen, Lars G Fritsche, Rounak Dey, Maiken E Gabrielsen, Brooke N Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A Gagliano, Aliya Gifford, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, 2018.

[10] Pan UKBB Team at Broad Institute. Pan-UK Biobank. `pan.ukbb.broadinstitute.org`, Accessed: 2023-08-01.

[11] Masahiro Kanai, Masato Akiyama, Atsushi Takahashi, Nana Matoba, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature Genetics*, 50(3):390–400, 2018.

[12] Kazuyoshi Ishigaki, Masato Akiyama, Masahiro Kanai, Atsushi Takahashi, Eiryo Kawakami, Hiroki Sugishita, Saori Sakaue, Nana Matoba, Siew-Kee Low, Yukinori Okada, et al. Large-scale genome-wide association study in a japanese population identifies novel susceptibility loci across different diseases. *Nature Genetics*, 52(7): 669–679, 2020.

[13] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[14] Phyllis M Thangaraj, Undina Gisladottir, and Nicholas P Tatonetti. Medical data and machine learning improve power of stroke genome-wide association studies. *bioRxiv*, pages 2020–01, 2020.

[15] Tim Poterba, et al. Hail-is/hail: 0.2.13. 0.2.13, Zenodo, 2019. doi:10.5281/zenodo.2646680.

[16] Karen Feng, et al. Projectglow/glow: V1.2.1. v1.2.1, Zenodo, 2022. doi:10.5281/zenodo.6474169.

[17] Lisa Bastarache. Using phecodes for research with the electronic health record: from phewas to phers. *Annual Review of Biomedical Data Science*, 4:1–19, 2021.

Supplementary information

Indirect GWAS runtimes

Direct GWAS approaches have computational costs related to the number of phenotypes. Indirect GWAS has computational costs related both to the number of input and output phenotypes. We measured the runtime for indGWAS alone using several input and output dimensions (Figure S1).

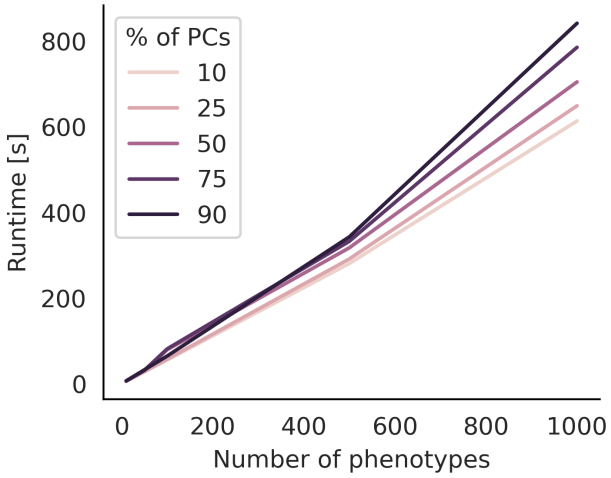

Figure S1: Runtimes for 10, 50, 100, 500, and 1000 phenotypes, for 10, 25, 50, 75, and 90 percent latent dimensions. The x-axis represents the number of output phenotypes (projections), while the "% of PCs" indicates the number of input phenotypes relative to the number of output phenotypes. Phenotypes and PCs here use the 1378 quantitative UKBB phenotypes.