Manuscript for review

Subset scanning for multi-trait analysis using GWAS summary statistics

Rui Cao,[1] Evan Olawsky,[1] Edward McFowland III,[2] Erin Marcotte,[3] Logan Spector[3] and Tianzhong Yang[1,3]

[1]Division of Biostatistics, University of Minnesota, 420 Delaware St. SE, 55455, MN, USA, [2]Technology and Operations Management, Harvard Business School, Soldiers Field, 02163, MA, USA and [3]Division of Epidemiology and Clinical Research, Department of Pediatrics, University of Minnesota, 420 Delaware St. SE, 55454, MN, USA

*To whom correspondence should be addressed. Email: yang3704@umn.edu

Abstract

Multi-trait analysis has been shown to have greater statistical power than single-trait analysis. Most of the existing multi-trait analysis methods only work with a limited number of traits and usually prioritize high statistical power over identifying relevant traits, which heavily rely on domain knowledge. To handle diseases and traits with obscure etiology, we developed TraitScan, a powerful and fast algorithm that agnostically searches and tests a subset of traits from a moderate or large number of traits (e.g., dozens to thousands) based on either individual-level or summary-level genetic data. We evaluated TraitScan using extensive simulations and found that it outperformed existing methods in terms of both testing power and trait selection when sparsity was low or modest. We then applied it to search for traits associated with Ewing Sarcoma, a rare bone tumor with peak onset in adolescence, among 706 traits in UK Biobank. Our analysis revealed a few promising traits worthy of further investigation, highlighting the use of TraitScan for more effective multi-trait analysis as biobanks emerge. Our algorithm is implemented in an R package 'TraitScan' available at `https://github.com/RuiCao34/TraitScan`.

Key words: Multi-trait analysis, Genome-wide association study (GWAS), Summary statistics, Biobank, Childhood cancer

1 Introduction

Genome-wide association studies (GWAS) have successfully improved the understanding of the genetic basis of many traits. The emergence of deeply phenotyped GWAS databases such as UK Biobank (Bycroft et al., 2018), eMERGE (Gottesman et al., 2013), and Vanderbilt BioVU (Roden et al., 2008), has facilitated studying associations between single nucleotide polymorphisms (SNPs) and a large number of traits. Phenome-wide association studies (PheWAS) have utilized this rich source of SNP-trait relationships to explore disease risks (Denny et al., 2010) and drug development (Diogo et al., 2018). By evaluating each trait individually, PheWAS is computationally fast to implement. However, it is well documented that joint multivariate analyses can be more powerful than univariate analyses such as PheWAS (Robinson et al., 2018). Many efforts have been devoted to multi-trait analyses that evaluate the relationships between a SNP and a set of traits simultaneously (O'Reilly et al., 2012; Kim et al., 2015; Zhu et al., 2015; Li and Zhu, 2017). Nonetheless, existing multi-trait analyses rely on domain knowledge to select a small number of related traits, and most of them only focus on obtaining high statistical power in hypothesis testing.

We are motivated to understand which risk factors contribute to childhood cancers, which are universally rare and have obscure disease etiology. For example, the etiology of Ewing sarcoma (EWS) remains unclear (Lahat et al., 2008), and conventional epidemiological methods to understand the disease are limited due to the extremely rare incidence rate of EWS (Spector et al., 2021). Recently, several genetic risk factors were identified in a GWAS of EWS (Machiela et al., 2018), potentiating the use of PheWAS to explore the risk factors agnostically. However, PheWAS can suffer from limited statistical power when scanning over a large number of traits, and simply applying the existing multi-trait methods to a large number of traits does not always yield meaningful results as the polygenic nature of some traits would eventually drive the statistical significance. To address the aforementioned limitations, we propose a novel multi-trait analysis method, TraitScan, that values both trait selection performance and high statistical power. Our method 'TraitScan' is based on a fast subset scan framework

(Neill, 2012) with a linear scan time of the number of traits and thus can handle high-dimensional trait selection. Our method contains three test statistics: higher criticism (HC), truncated chi-square (TC), and a combined test of HC and TC. We note a similar method ASSET (Bhattacharjee et al., 2012) with the same objective. However, it requires an exhaustive search of all possible subsets, which results in an exponential scan time and thus is not computationally efficient when the number of traits exceeds a few dozen.

Our proposed TraitScan algorithm is able to utilize summary-level GWAS data in situations where individual-level data are not available. Due to the logistical limitations and privacy concerns of sharing individual-level data, it has become a common practice to share GWAS summary statistics. Leveraging publicly available summary-level data, our method can filter relevant pleiotropic traits on any given SNP. Through simulations, we show that our method has high power and sensitivity in terms of trait selection under moderately sparse to sparse situations (i.e., the number of truly associated traits is smaller than or close to the square root of total traits). We evaluate traits associated with EWS through GWAS summary statistics on 706 traits filtered from the UK Biobank study (Sudlow et al., 2015). Besides single SNPs, we also show that our method can be extended to genetic scores, such as the predicted gene expression levels in transcriptome-wide association studies (TWAS) (Feng et al., 2021) and polygenic risk scores (PRS) (Torkamani et al., 2018). We implement our method in an R package 'TraitScan' that is publicly available at `https://github.com/RuiCao34/TraitScan`. The package provides an option to use the pre-calculated null distributions of the test statistics, which can handle screening 706 traits in 36 seconds.

2 Materials and methods

2.1 Models

Our method performs a scan for traits under the null hypothesis that none of the traits is associated with the genetic variant of interest (i.e., SNP). Assume that the data consist of p continuous traits and a minor allele dose for a SNP collected from n individuals, for $j = 1, \ldots, p$. Let $y_j = (y_{1j}, \ldots, y_{nj})^T$ be a vector of values of the j^{th} trait for each of the n individuals, $x = (x_1, \ldots, x_n)^T$ a vector of the minor allele doses of the SNP of interest, and $\epsilon_j = (\epsilon_{1j}, \ldots, \epsilon_{nj})^T$ a vector of error terms for the j^{th} trait. For continuous traits, we assume that each trait can be modeled as a linear function of the genetic variant, and without loss of generality, y_j is centered and standardized:

$$y_j = x\beta_j + \epsilon_j. \tag{1}$$

We define our null hypothesis as:

$$H_0 : \beta_1, \ldots, \beta_p = 0. \tag{2}$$

Let $\mathbf{Y} = (y_1, \ldots, y_p)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)$. We can stack the models together as:

$$\mathbf{Y} = x\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3}$$

To model the correlation structure between the traits, we assume that x is fixed and the rows of $\boldsymbol{\epsilon}$ represent $n \times p$ i.i.d. observations from an $MVN(\mathbf{0}, \boldsymbol{\Omega})$ distribution, where $\mathbf{0}$ is a vector of zeroes of length $n \times p$ and $\boldsymbol{\Omega} = I_{n\times n} \otimes \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is the potentially unknown $p \times p$ covariance matrix of the traits. When the individual-level data are available, the matrix can be estimated from the residuals of fitting separate linear regression models: $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-p}\sum_{j=1}^{n}(y_j - x\hat{\beta_j})(y_j - x\hat{\beta_j})^T$, where $\hat{\beta_j}$ is the ordinary least square estimate.

When only summary statistics are available, which typically consist of $\hat{\beta_j}$, $se(\hat{\beta_j})$ from Equation 1 and the z-score from a Wald test $z_j = \hat{\beta_j}/se(\hat{\beta_j})$, each entry of $\hat{\boldsymbol{\Sigma}}$, $\hat{\sigma}_{ij}$ can be approximated using the null SNPs (i.e. the SNPs with no association with any traits) by ignoring the estimation error of $\hat{\beta}$ (Kim et al., 2015; Liu and Lin, 2018):

$$cor(z_i, z_j) \approx cor(\hat{\beta}_i, \hat{\beta}_j) = cor(\frac{x^T}{x^T x}y_i, \frac{x^T}{x^T x}y_j) = cor(y_i, y_j) \equiv \sigma_{ij}. \tag{4}$$

Additionally, when the summary statistics come from overlapping but not identical samples, the correlation between z-scores is still proportional to the trait correlation σ_{ij} (Li et al., 2021):

$$cor(z_i, z_j) \approx cor(\hat{\beta}_i, \hat{\beta}_j) = \frac{n_{ij}}{\sqrt{n_i n_j}}\sigma_{ij}, \tag{5}$$

where n_i, n_j,and n_{ij} are the sample sizes of trait i, trait j, and their overlapping samples respectively. Since the $(i,j)^{th}$ z-score correlation is a constant across all null SNPs, it can be estimated empirically (Zhu et al., 2015):

$$\hat{cor}(z_i, z_j) = \frac{\sum_k (z_i^k - \bar{z}_i)(z_j^k - \bar{z}_j)}{\sqrt{\sum_k (z_i^k - \bar{z}_i)^2 \sum_k (z_j^k - \bar{z}_j)^2}}, \tag{6}$$

where z_i^k is the z-score for null SNP k and trait i, and $\bar{z}_i$ is the mean of vector $(z_i^1, ..z_i^k, ...)$ across all null SNPs.

For binary traits, a logistic regression model is usually fitted in GWAS:

$$log\frac{P(y_{ij} = 1|x_i)}{1 - P(y_{ij} = 1|x_i)} = \beta_{j0} + x_i\beta_{j1}. \tag{7}$$

When the effect size β_{j1} is small, which most often happens in GWAS, the logistic regression model can be approximated by a linear regression model based on the first-order Taylor expansion on β_{j1}:

$$P(y_{ij} = 1|x_i) = \frac{1}{1 + e^{-x_i\beta_{j1} - \beta_{j0}}} \approx \alpha_{j0} + \alpha_{j1}x_i + \xi, \tag{8}$$

where α_{j0}, α_{j1}, and ξ can be regarded as linear regression coefficients. The covariance for binary traits can be similarly derived using summary statistics of the null SNPs. In practice, null SNPs can be chosen based on GWAS p-values (i.e., > 0.05).

2.2 Scan Statistics

Our subset scanning algorithm relies on a score statistic F, which is a function of a non-empty subset $S \subseteq \{1, \ldots, p\}$. It quantifies the amount of anomalousness found in traits $\{y_j | j \in S\}$ under the null hypothesis that no trait is associated with the SNP. The most anomalous subset is found by maximizing $F(S)$ over all non-empty subsets of the traits. Calculating $F(S)$ over all possible subsets S is extremely burdensome when p is large. To ensure efficient maximization, we use subset scanning techniques and strive for a statistic accompanying priority function that satisfies the strong linear time subset scanning (LTSS) property:

Definition 1 (Neill (2012)) The score function $F(S)$ and priority function $G(j; \tilde{y})$ satisfy the strong LTSS property if and only if,

for all $j = 1, \ldots, p$, $\max_{S:|S|=j} F(S) = F(\{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(j)}\})$, where $\tilde{y}^{(j)}$ is the trait with the jth highest value of $G(\cdot; \tilde{y})$.

If $F(S)$ satisfies the strong LTSS property, the subset S^* that maximizes $F(S)$ must be the subset containing the c highest-priority traits $\{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(c)}\}$ for some c between 1 and p. Thus, to solve the global optimization problem, we can simply sort the traits by their priority value given by G and then compute $F(S)$ with S taken to be one of the p subsets $\{\tilde{y}^{(1)}\}, \{\tilde{y}^{(1)}, \tilde{y}^{(2)}\}, \ldots, \{\tilde{y}^{(1)}, \ldots, \tilde{y}^{(p)}\}$.

Neill (2012) gave a constructive theorem that produces a specific priority function $G(j; \tilde{Y})$ that follows directly from the score function $F(S)$ when certain properties hold. This pair of functions is then guaranteed to satisfy the strong LTSS property.

Theorem 1 (Neill (2012)) *Let $F(S) = F(T, |S|)$ be a function of one additive statistic of subset S, $T(S) = \sum_{j \in S} g(j; \tilde{y})$ (where $g(j; \tilde{y})$ depends only on trait $\tilde{y}_j$) and the cardinality of S, $|S|$. Assume that $F(S)$ is monotonically increasing with $T(S)$, then $F(S)$ satisfies the strong LTSS property with priority function $G(j; \tilde{y}) = g(j; \tilde{y})$.*

We thus construct two score statistics, HC and TC that satisfy the conditions of Theorem 1 while quantifying the amount of anomalousness found in a subset of the traits under the null hypothesis. We show by simulations that the HC or TC method had better performance than performing PheWAS under different scenarios. We also combine the two tests by taking the minimum p-value of the two tests, which enables us to achieve results comparable to the better-performed HC or TC method in terms of statistical power and trait selectivity.

2.2.1 Decorrelation

As in Theorem 1, a trait-level statistic is required to quantify the amount of association between a trait and a genetic variant. We used the p-value p_j of the Wald test z_j from the separate regression models (Equation 1), for $j = 1, \ldots, p$. As the traits are correlated and sampled from the same or overlapping individuals in our framework, the p_j's are correlated. Herein, we perform the ZCA-cor whitening method (Kessy et al., 2018) on z_j, ensuring that the whitened z_j^* remains maximally correlated with z_j. Then we obtain the p_j^* corresponding to z_j^*.

2.2.2 Higher Criticism Statistic

Following McFowland et al. (2013), we choose the HC score function: $F_{HC,\alpha}(S) = \frac{N_\alpha - \alpha|S|}{\sqrt{|S|\alpha(1-\alpha)}}$, where $N_\alpha = \sum_{j=1}^{|S|} n_\alpha(p_j^*) = \sum_{j=1}^{|S|} I(p_j^* < \alpha)$. The score function is the standardized difference between the observed count of p-values lower than a p-value threshold α and the expected count. According to Theorem 1, it can be easily seen that $F_{HC}(S)$ is a function of $|S|$ and one additive statistic N_α. $F_{HC}(S)$ is monotonically increasing with N_α and thus satisfies the strong LTSS property with priority function $n_\alpha(p_j^*) = I(p_j^* < \alpha)$.

As we do not know the optimal α, we define grid-based HC test statistic H_{HC}:

$$H_{HC} = \max_S F_{HC}(S) = \max_\alpha \max_S F_{HC,\alpha}(S) \tag{9}$$

over a grid of α and its corresponding subset

$$S_{HC} = \text{argmax}_S F_{HC}(S). \tag{10}$$

An ideal α grid should ensure all possible subsets in the search space, therefore, there should be no more than one p-value between two arbitrary adjacent α's. In practice, we recommend the α grid as a geometric sequence from the Bonferroni significant p-value threshold to overall Type I error with a sequence length of 200, i.e. $\alpha_1, ..., \alpha_{200} = \underbrace{0.05/p, ..., 0.05}_{n_\alpha = 200}$. The lower bound of α ensures that our test would always be more powerful than PheWAS in special scenarios where all traits are uncorrelated. In the meanwhile, we use an upper bound of 0.05 to decrease the search space.

2.2.3 Truncated Chi-squared Statistic

The HC test may not have ideal performance under non-sparse scenarios (Barnett et al., 2017) and does not take into account the strength of association. To make our method more robust, we propose an additional statistic, which is similar to the truncated z-score method (Bu et al., 2020) and also meets the strong LTSS property. First, we define γ as the $|z^*|$ threshold, which is closely related to α in HC statistics: $\gamma = \Phi^{-1}(1 - \alpha/2)$ and Φ as the cumulative distribution function of a standard normal distribution. The score function $F(S)$ is defined as $F_{TC}(S) = -log(P_{M_\gamma(S)}|H_0)$, i.e. the negative log p-value of the subset score function $M_\gamma(S) = \sum_{j=1}^{|S|} I(|z_j^*| > \gamma) z_j^{*2}$ with priority function $I(|z_j^*| > \gamma) z_j^{*2}$.

Note that M_γ is a non-decreasing function of S, and $F_{TC}(S)$ is monotonically increasing with M_γ, thus satisfying the strong LTSS property. Similarly, we test a grid of γ one-to-one mapping to the grid of α defined previously:

$$H_{TC} = \max_S F_{TC}(S) = \max_\gamma \max_S F_{TC,\gamma}(S) \tag{11}$$

$$S_{TC} = \text{argmax}_S F_{TC}(S) \tag{12}$$

γ grid is chosen in correspondence with the α values.

2.3 Assessing Significance

We have now obtained two subsets S_{HC} and S_{TC} that maximize HC or TC statistics. As described above, both subsets always include at least one trait. To determine whether the selected subset is sufficiently anomalous, we calculate the corresponding p-values p_{HC} and p_{TC} by comparing the two statistics H_{HC} and H_{TC} with their distributions under the null hypothesis.

p_{HC} can be calculated analytically (Supplementary Materials). As for p_{TC}, we use Monte Carlo (MC) simulations. We start by simulating p z-scores under the null, i.e. standard normal distribution for B iterations. For the b^{th} iteration, the $H_{TC,b}$ can be calculated, and empirical p-values p_{TC} can be estimated from the simulated H_{TC} distribution.

$$p_{TC} = \Sigma_b I(H_{TC} > H_{TC,b})/B. \tag{13}$$

To combine the HC and TC tests, we compare the p-values p_{HC} and p_{TC} and get the grid-based statistics:

$$H_{combined} = min(p_{HC}, p_{TC}), \tag{14}$$

and the traits selected by the combined test are also determined by the test with a smaller p-value:

$$S_{combined} = S_{\text{argmin}_M p_M}, \tag{15}$$

where M = HC or TC. The empirical null distribution of $H_{combined}$ can be similarly simulated by MC, and the p-value from the combined test $p_{combined}$ is calculated by comparing the test statistic $H_{combined}$ and its distribution under the null.

If the null hypothesis is rejected, we can conclude that the SNP is associated with at least one trait contained in S^*. Note that this MC simulation step only depends on the number of traits p and the choice of F-statistics. For SNPs sharing the same number of traits p, we do not need to recompute the test statistic distribution under the null.

2.4 Extension to Genetic Scores

Genetic scores integrate information from multiple SNPs. Linking traits with genetic scores could bring in more statistical power and provide a meaningful interpretation of the results. Genetic scores, which are usually the linear combinations of allele counts of multiple SNPs, have been extensively developed and distributed. Polygenic risk scores that predict the risk of clinical and epidemiological traits (Lewis and Vassos, 2020) or imputation models for gene expression levels in TWAS (Xu et al., 2023) are two types of commonly used genetic scores. We will show how TraitScan can be easily utilized on genetic scores using summary-level GWAS data and an external genetic reference panel.

2.4.1 Continuous Traits

Let X_{gs} denote the genetic score from q SNPs: $X_{gs} = \sum_{l=1}^{q} c_l X_l$, where $X_l = (x_{1l}, ..., x_{nl})$ is the genotype vector for n individuals at the l^{th} SNP, and c_l is the SNP weight vector. In GWAS models, the j^{th} trait y_j is marginally regressed on each SNP X_l, and regression coefficients $\hat{\beta}_{jl}$ and $se(\hat{\beta}_{jl})$ are estimated from the linear regression model

$$y_j = X_l \beta_{jl} + \epsilon_{jl}. \tag{16}$$

For the genetic score, we are interested in the regression model

$$y_j = X_{gs} \beta_{j,gs} + \epsilon_{j,gs} \tag{17}$$

and test the null hypothesis

$$H_{0,gs} : \beta_{1,gs}, \ldots, \beta_{p,gs} = 0. \tag{18}$$

When individual-level data are available, the regression coefficients $\hat{\beta}_{j,gs}$ and $se(\hat{\beta}_{j,gs})$ with z statistic from the Wald test $z_{j,gs} = \hat{\beta}_{j,gs}/se(\hat{\beta}_{j,gs})$ can be directly calculated. When only summary-level data are available, we have

$$\hat{\beta}_{j,gs} = (X_{gs}^T X_{gs})^{-1} X_{gs}^T y_j, \tag{19}$$

$$se(\hat{\beta}_{j,gs}) = \sqrt{\hat{\sigma}_{gs}^2 (X_{gs}^T X_{gs})^{-1}}, \tag{20}$$

$$\hat{\sigma}_{gs}^2 = \frac{y_j^T y_j - y_j^T X_{gs} (X_{gs}^T X_{gs})^{-1} X_{gs}^T y_j}{n_j - q}, \tag{21}$$

where $\hat{\sigma}_{gs}^2$ is the residual variance estimate and n_j is the sample size for j^{th} trait. The items $X_{gs}^T y_j$ and $y_j^T y_j$ can be derived from GWAS summary data (Pattee and Pan, 2020):

$$X_{gs}^T y_j = n_j (\hat{s}_1^2 \hat{\beta}_{j1}, \ldots, \hat{s}_p^2 \hat{\beta}_{jp}), \tag{22}$$

$$y_j^T y_j = n_j^2 \times \hat{s}_l^2 \times se(\hat{\beta}_{j,gs})^2 + n_j \times \hat{s}_l^2 \times \hat{\beta}_{j,gs}^2, \tag{23}$$

where $\hat{s}_l^2$ is the variance of SNP l. Both $\hat{s}_l^2$ and the genotype matrix $X_{gs}^T X_{gs}$ can be estimated from a reference panel comprising genotypic data of individuals from a general population (1000 Genomes Project Consortium, 2015). In practice, for Equation 23, we can calculate $y_j^T y_j$ across multiple SNPs and take the median as the estimate.

After z statistics $\{z_{j,gs}\}$ are computed, we could follow the same decorrelation and trait scanning steps as above since the genetic score can also be treated as a SNP, and the covariances between $\{z_{j,gs}\}$ are identical under the null hypothesis.

2.4.2 Binary Traits

We have the logistic regression for binary traits:

$$log \frac{P(y_{ij} = 1 | x_{il})}{1 - P(y_{ij} = 1 | x_{il})} = b_{0,jl} + x_{il} b_{jl} \tag{24}$$

for i^{th} individual, j^{th} trait, and l^{th} SNP, and $b_{0,jl}$ and b_{jl} are the regression coefficients. Following Pattee and Pan (2020), we could approximate $P(y_{ij} = 1 | x_{il})$ as a continuous outcome under a linear regression model and denote $\beta_{0,jl}$ and β_{jl} as the coefficients. The following equations hold:

$$\hat{\beta}_{jl} = \frac{e^{-\hat{b}_{0,jl}}}{(1 + e^{-\hat{b}_{0,jl}})^2} \hat{b}_{jl}, \tag{25}$$

$$se(\hat{\beta}_{jl}) = (\frac{e^{-\hat{b}_{0,jl}}}{(1 + e^{-\hat{b}_{0,jl}})^2})^2 se(\hat{b}_{jl}), \tag{26}$$

where $e^{-\hat{b}_{0,jl}} = \frac{P(y_{ij}=0)}{P(y_{ij}=1)}$ is the ratio of control and case sizes. The logistic regression coefficients can be thus converted to linear regression coefficients and handled by the steps mentioned above.

3 Real Data application

We used our method on UK Biobank GWAS data to find out potential traits linked to EWS. EWS is a type of rare childhood cancer in bone or soft tissue (Li and Chen, 2022). Previous studies (Postel-Vinay et al., 2012; Machiela et al., 2018) suggested that six SNPs rs113663169, rs7742053, rs10822056, rs2412476, rs6047482, and rs6106336 were significantly associated with EWS in individuals of European ancestry. We analyzed these six SNPs using the GWAS summary statistics of the UK Biobank data (team, 2020).

UK Biobank is a large-scale database encompassing a broad range of phenotypes, where individuals' genetic data are linked to electronic health records and survey measures (Sudlow et al., 2015). The phenotypes include population characteristics, biological markers, medical history, environments, dining habits, cognitive functions, etc. In the GWAS study, samples with sex discordance and SNPs with low minor allele counts or low imputation scores were filtered out. Summary statistics of GWAS were obtained from fitting generalized mixed models with a kinship matrix as a random effect and covariates as fixed effects within each genetic ancestry. The heritability of each trait was provided, which was estimated by the Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) (Zhou et al., 2018). More method and analysis details can be found on the Pan-UK Biobank website (`https://pan.ukbb.broadinstitute.org/`). In our analysis, we focused on the GWAS summary statistics for individuals of European ancestry and with a sufficient number of participants of both genders. We applied the following criteria which left us with 706 traits to perform the TraitScan algorithm:

Table 1. Use TraitScan to search among 706 traits in UK Biobank for EWS-linked SNPs.

SNP	Trait category	Most significant trait in the category	PheWAS p-value
rs113663169	Touchscreen questions	Natural hair color: blonde	$\leq 10^{-20}$
	Baseline characteristics	Seated height	3.19×10^{-7}
rs10822056	Biological samples	Monocyte count	1.19×10^{-13}
rs2412476	Biological samples	Aspartate aminotransferase	$\leq 10^{-20}$
		Erythrocyte distribution width	$\leq 10^{-20}$
rs6047482	Biological samples	Urea	1.02×10^{-7}
rs6106336	Biological samples	Insulin-like growth factor 1	6.16×10^{-7}

- Continuous traits with a sample size of at least 5,000 or binary traits with a sample size of at least 5,000 cases and 5,000 controls.
- Traits with genetic heritability estimated to be larger than 0.
- Traits with at least one genome-wide significant SNP (p-value$< 5 \times 10^{-8}$).
- Traits with the sample size of each sex larger than 20.
- Traits belonging to these categories were included: health-related outcomes, online follow-up, biological samples, X-ray absorptiometry (DXA), cognitive function, verbal interview, touchscreen questions (except traits related to eyes), and baseline characteristics.

As we intended to evaluate six SNPs, the significance level for TraitScan tests was set at $0.05/6 = 0.0083$ after the Bonferroni correction. SNPs rs113663169, rs10822056, rs2412476, rs6047482, and rs6106336 were shown to have significant associations with at least one trait out of 706 examined traits in UKBiobank by TraitScan (TraitScan combined tests p-value $\leq 1 \times 10^{-4}$). For SNP rs7742053, TraitScan combined test p-value was 0.863 and thus did not reach statistical significance. Table 1 summarizes the results of TraitScan combined test for the most significant trait in each category identified for the five SNPs. It showed traits that were highly significant in PheWAS were also captured by TraitScan. In fact, TraitScan identified a total of 21 UK Biobank traits related to the five EWS-linked SNPs, while 8 of the trait-SNP associations did not reach statistical significance in PheWAS (using a Bonferroni-significant threshold at $0.05/(706 \times 6) = 1.18 \times 10^{-5}$). A full list of selected traits is shown in Supplementary Materials (Table S1).

To demonstrate the TraitScan application on genetics scores, we further carried out our method on the transcriptomic scores of gene *KIZ* and gene *RREB1*, i.e., the SNP-imputed gene expression of the two genes, with the 706 UK Biobank traits. *KIZ* and *RREB1* had strong evidence to be linked to three top genome-wide significant SNPs in EWS GWAS (*KIZ* was linked to SNPs rs6047482 and rs6106336, and *RREB1* to rs7742053) (Machiela et al., 2018). The weights in the transcriptomic scores of human blood were obtained from Xu et al. (2023), and the internal r^2 of the scores were 0.206 and 0.031 for *KIZ* and *RREB1*, respectively. The significance levels for both genes were set at $0.05/2 = 0.025$ after the Bonferroni correction. In addition, we also tested the relationship between the number of risk alleles identified in Machiela et al. (2018) with EWS. It was shown that EWS cases had on average 1.08 more risk alleles than controls (p-value $= 2.44 \times 10^{-63}$).

After applying TraitScan on the genetic scores, neither of the two genes *KIZ* (p-value = 0.0826) and *RREB1* (p-value = 1) reached statistical significance in TraitScan, although the trait insulin-like growth factor 1 (IGF-1) picked up by *KIZ* was marginally significant. On the other hand, imputed gene expression of *RREB1*, of which rs7742053 was an expression quantitative trait loci (eQTL), had no evidence of linking to any of the 706 traits. The genetic score of six EWS SNPs, however, was significantly associated with three traits: blonde hair color, ease of skin tanning, and monocyte percentage (TraitScan p-value $< 1 \times 10^{-4}$), while the PheWAS identified two additional traits on the EWS score: patient care technician location and facial pain experienced in last month.

To investigate the causal relationship between EWS and the selected traits, we further performed bidirectional Mendelian randomization (MR) analysis on the traits selected by TraitScan using the TwoSampleMR package (Hemani et al., 2018, 2017). The instrumental variables were selected from either UK Biobank or EWS GWAS data and were clumped by $r^2 < 0.001$ and p-value $< 5 \times 10^{-5}$. For the selected traits with more than one instrumental variable, inverse-variance weighted (IVW), Egger regression, weighted median, simple mode, and weighted mode methods were applied, while the Wald ratio method was applied for the selected traits with one instrumental variable. Full results are reported in Table S2. EWS was shown to be causal for lower Alkaline phosphatase levels, the trait selected to be associated with SNP rs10822056 in TraitScan but not PheWAS (MR IVW coefficient $= -0.53$ with p-value $= 0.084$ and MR weighted median coefficient $= -1.10$ with p-value $= 1.31 \times 10^{-4}$). For the casual direction where EWS was the exposure, no MR test reached statistical significance after accounting for multiple testing, suggesting no causal effect from EWS to the selected UK Biobank traits.

4 Simulation

Throughout the simulations, we used five metrics to assess the performance of each method, i.e., power, size, recall, precision, and Jaccard similarity. Let p be the total number of traits, S^* be the subset of traits as chosen by a particular method and let S_0 be the true subset of pleiotropic traits. The size was defined as $E|S^*|$. Then, we defined precision to be $E\frac{|S^* \cap S_0|}{|S^*|}$, the proportion of traits identified by the method that was truly associated with the SNP. We defined recall to be $E\frac{|S^* \cap S_0|}{|S_0|}$, the proportion of the pleiotropic traits identified by the method. We defined Jaccard similarity, a combination of precision and recall, to be $E\frac{|S^* \cap S_0|}{|S^* \cup S_0|}$. Finally, power was assessed by comparing the observed statistic to the simulated distribution of null statistics.

We compared the performance of variable selection and testing for TraitScan using summary statistics with some existing methods: PheWAS (with Bonferroni adjustment), CPASSOC

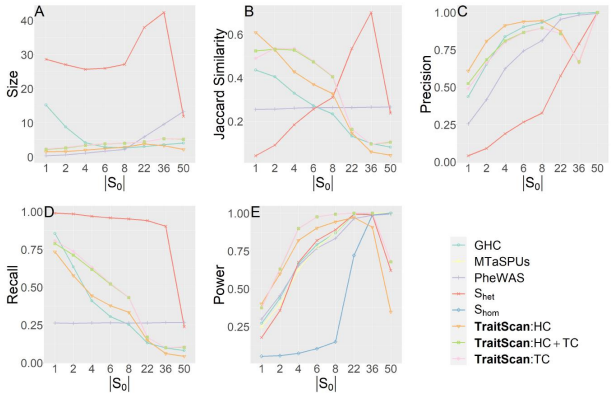

Fig. 1: Simulation scenario 1 with varying numbers of truly associated traits ($p = 50$)

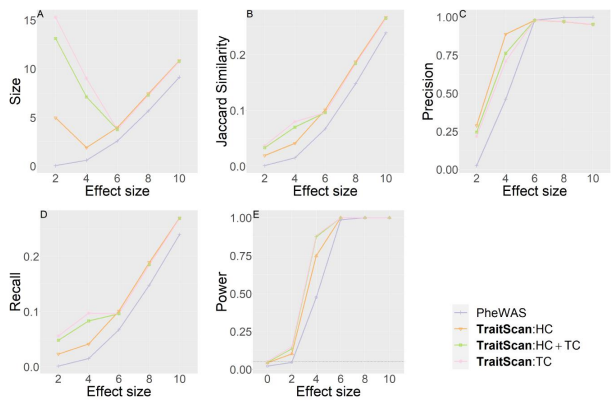

Fig. 2: Simulation scenario 2 with real data covariance matrix and effects from SNP rs113663169 ($p = 706, |S_0| = 38$). Effect size is in proportion to the estimated correlation observed in UK Biobank.

(S_{hom}, S_{het}) (Zhu et al., 2015), MTaSPUs (Kim et al., 2015), and generalized higher criticism (GHC) (Barnett et al., 2017). Among these methods, MTaSPUs and S_{hom} cannot select traits, and thus we only evaluated their performance in terms of statistical power. S_{het} includes a parameter grid as the thresholds for z-scores and can naturally select out the traits with absolute z-scores smaller than each threshold. As suggested by their package, the parameter grid was set as the observed trait p-values. The GHC was originally proposed for SNP set association with a single trait, and here we used it for a single SNP association with multiple traits. We did not compare our method with ASSET because it is not computationally efficient in our simulation settings (706 traits for the real data variance-covariance scenario and 50 traits for the rest of the scenarios).

We conducted simulations under multiple scenarios to show TraitScan has higher power and Jaccard similarity under moderately sparse ($|S_0|/p < 0.4$) and sparse situations ($|S_0|/p < 0.05$). We focus the discussion on the scenarios with varying numbers of truly associated traits (scenario 1) or with real data variance-covariance matrix (scenario 2) and briefly discuss the other four scenarios and their results.

We assessed the method performance by varying the number of truly associated traits $|S_0|$ in scenario 1 (Figure 1). In terms of statistical power, we observed that TraitScan test statistics (HC, TC, HC+TC) had the highest power under moderately sparse and sparse situations ($|S_0| < 22$, or $|S_0|/p < 0.44$). The HC test was more powerful than the TC test under extremely sparse situations ($|S_0| = 1$). When $|S_0| = 50$, the GHC, MTaSPUs, PheWAS, and S_{hom} had the highest power, and TraitScan and S_{het} were less powerful. In terms of variable selection performance, we found that TraitScan test statistics (HC, TC, HC+TC) also had the highest Jaccard similarity under moderately sparse and sparse situations ($|S_0| < 22$, or $|S_0|/p < 0.44$), while S_{het} had better Jaccard similarity as increasing proportion of truly associated traits. However, S_{het} tended to over-select traits and thus had the lowest precision under most situations.

We also tested method performance using the covariance matrix and effect sizes estimated from the 706 UK biobank traits (Figure 2), where 38 traits with marginal p-value ≤ 0.05 were set as a true subset of pleiotropic traits. In this scenario, we compared the performance of PheWAS and TraitScan under trait-SNP associations of different strengths. S_{het}, GHC, and MTaSPUs were not applied due to the computational burden, and S_{hom} was excluded due to the homogeneous effect assumption was not met in the simulation and unlikely to be met in real data analysis. By comparing the Type I errors ($\boldsymbol{\beta} = \mathbf{0}$), we showed that TraitScan tests were well calibrated, while PheWAS had a slightly deflated Type I error rate due to its independence assumption among traits. Similar to scenario 1 mentioned above, TraitScan had higher power and Jaccard similarity over PheWAS under all effect sizes. We also notice that the selected trait size did not grow with effect size. When the effect sizes were small, the p-values of truly associated traits and null traits were close, and TraitScan tended to select a large set of traits as the most anonymous trait subset.

We showed that TraitScan can handle mixed types of traits (i.e., continuous and binary) simultaneously (Figure S1) and the performance followed the same pattern as continuous traits. TraitScan still demonstrated the highest power and Jaccard Similarity under effects varying in both directions and magnitudes (Table S3), under block-diagonal correlated traits (Figure S2), or different correlation magnitudes or structures (Table S4). Moreover, we found that TraitScan had higher power and Jaccard similarity when handling more highly correlated traits. The detailed parameter settings of the simulation can be found in Supplementary Materials.

5 Discussion

We proposed a new method called TraitScan for post-GWAS trait subset scanning and testing. While most of the existing multi-trait methods rely on domain knowledge, our method allows agnostic search among a large number of traits and is able to identify a set of traits with the most anomalousness. TraitScan utilizes the fast subset scan framework (Neill, 2012), resulting in a linear scan time over the number of traits. Taking correlation among traits into consideration, TraitScan has demonstrated higher power and trait selectivity than PheWAS when sparsity was low or modest. The method is compatible with both individual-level and summary-level GWAS data, although we focus more on summary-level

Table 2. Computational time in seconds for 5,000 iterations (10 traits)

Method	Time
GHC	129.5
ASSET	5191.8
S_{het}	9.0
TraitScan-MC: HC + TC	46.1
TraitScan-analytic: HC	21.7
TraitScan-precalculated: HC+TC	28.6

GWAS data herein to allow an easy application to existing deeply phenotyped GWAS summary statistics databases.

In implementation, we recommend a grid of 200 $\alpha's$ with the minimum α as the Bonferroni significant p-value cutoff and maximum α of 0.05. A practical issue faced by TraitScan and other threshold-based multi-trait methods is the choice of the density of thresholds. For the statistic S_{het}, Zhu et al. (2015) recommended flexible thresholds which are the same as the input z-scores, while Bu et al. (2020) and TraitScan used fixed p-value thresholds. In simulations (not shown), we observed a dramatic power and precision loss of S_{het} using the same fixed threshold grid as in TraitScan. It is noteworthy that TraitScan always selects traits that are statistically significant in the decorrelated univariate analysis as we set the minimum α as the Bonferroni significant p-value cutoff. Based on our experience with simulations and real data analyses, we recommend a maximum α of 0.05 as traits with decorrelated p-values larger than 0.05 have never been selected by our algorithm. Since TraitScan includes a MC step simulating the null distribution of test statistics, an overly dense α grid may slow down the algorithm. We found a grid of 200 $\alpha's$ is sufficiently dense when handling hundreds of GWAS traits.

Given a fixed number of $\alpha's$, TraitScan has an $O(p)$ time complexity. If the number of $\alpha's$ is proportional to p, the time complexity is $O(p^2)$. While for the other multi-trait analysis methods, the S_{het} test in CPASSOC also has an $O(p)$ time complexity given a fixed p-value threshold, since it ranks the p-values and directly applies the threshold onto the raw p-values. When the thresholds are set as the input p-values, which is recommended in the CPASSOC pipeline, its time complexity is also $O(p^2)$. The ASSET has an $O(2^p)$ time complexity, meaning the computational time will be doubled once adding one more trait. Table 2 lists the computational time for analyzing 10 traits 5,000 times using different methods, including TraitScan with 10,000 iterations in the MC simulation (TraitScan-MC), TraitScan:HC test based on analytical null distribution (TraitScan-analytic), and TraitScan test using a precalculated null distribution estimated by MC with a given p (TraitScan-precalculated). The trait correlations and β parameters are the same as in Scenario 1.

In the examination of traits associated with EWS, TraitScan identified eight additional trait-SNP associations which did not reach the PheWAS significance level. One of these traits, alkaline phosphatase, measured by blood assays, also showed significance in MR analysis, suggesting it was causally related to EWS. Evidence has shown the presence of abundant alkaline phosphatase activity in EWS tumor cells (Sharada et al., 2006); however, the direction of association between alkaline phosphatase and EWS was previously unknown. Another trait, IGF-1, was selected by TraitScan for SNPs rs2412476 on chromosome 15, rs6047482 on chromosome 20, and rs10822056 on chromosome 10. IGF1-receptor is known to be upregulated in EWS, and anti-IGF1 is an experimental therapy (Gonzalez et al., 2020). Besides, the SNP rs7742053, the only SNP that failed to reach genome-wide significance in both TraitScan and PheWAS, has recently been reported to have a specific role in the increased binding of GGAA microsatellite alleles with the chromosomal translocation encoding chimeric transcription factors (Lee et al., 2023).

Examining the genetic score of *KIZ* and *RREB1* allowed us to investigate whether any trait was associated with EWS on the gene level. If a gene is associated with the same set of traits, then likely multiple SNPs in the gene will be associated with the traits, leading to higher power than the single SNP test. However, if the weight in the genetic score is not informative, such as imputing gene expression in a non-relevant tissue, or if multiple SNPs in the gene suggest a different association with the trait, we would have diminished power. We did not observe any traits significantly associated with the genetic scores (i.e., imputed gene expression) of *KIZ* and *RREB1*, potentially because blood may not be the most relevant tissue for EWS. We note that like other methods, our results relied on the quality of GWAS data. When handling real GWAS data, we applied a couple of filtering steps to exclude traits that are not heritable. However, after the filtering steps, there were still a few traits that lacked reasonable explanations of their genetic heritability such as the inpatient record format, or the potential mechanisms to be associated with EWS, such as fruit intake within the past 24 hours. We suspect it was due to the inadequate adjustment for confounding in the original GWAS analysis (Holmes et al., 2019). Without accessing the individual-level data, it is difficult to examine or correct the summary-level GWAS data, although there is some recent work performing quality control on GWAS errors using summary statistics and a reference panel (Chen et al., 2021; Darrous et al., 2021).

To use TraitScan in real data analysis, the following additional steps could help avoid potential power loss and increase the interpretability of the results. First of all, we suggest removing highly correlated traits from the pool of putative traits by examining the empirical trait correlation matrix. The LTSS property of TraitScan requires traits to be independent of each other. As shown in the simulation, our method had relatively low statistical power when the genetic variant had the effects and correlations of the same direction on most of the traits. We find that the decorrelation step on z-scores shifted the means of z-scores of the truly associated traits towards zero, resulting in a power loss. Therefore, removing such traits could potentially improve the statistical power. Future work may be focused on developing subset algorithms balancing the computational time and scan sensitivity. Secondly, we recommend checking the correlation between the decorrelated traits and raw traits. Due to the trait decorrelation in TraitScan, trait selection and testing are performed on the decorrelated z-scores, which are essentially linear combinations of raw z-scores. Although the ZCA-cor decorrelation method maximizes the average correlation between each dimension of the decorrelated and original data, the decorrelated traits might be considered to differ from the original traits. Therefore, this step could improve the interpretability of the findings. In our real data analysis, 99% of the 706 UK Biobank traits had an empirical correlation with the original trait greater than 0.7.

The understanding of rare diseases such as childhood cancer has long been limited. TraitScan is able to provide a list of possible traits associated with EWS through the disease-linked genetic

variants. As association does not imply causality, further biological experiments or additional data analysis approaches such as MR are required to study whether a trait and target disease are causally linked and whether the trait is a risk factor or a consequence of the disease.

Competing interests

No competing interest is declared.

Data Availability Statement

The summary level GWAS data for UK Biobank can be downloaded through instructions at `https://pan.ukbb.broadinstitute.org/downloads`. The algorithm for the proposed work is packaged in R, available at `https://github.com/RuiCao34/TraitScan`.

Acknowledgments

The authors thank Prof. Jim Hodges of the Division of Biostatistics, University of Minnesota, who helped with administrative matters early in this project. The authors also want to thank Dr. Aubrey K. Hubbard and Mitchell J. Machiela from the Division of Cancer Epidemiology and Genetics at National Cancer Institute for sharing the summary statistics of Ewing Sarcoma. This work is supported by the Children's Cancer Research Fund and by the Minnesota Supercomputing Institute at the University of Minnesota. Dr. Yang would like to further acknowledge St Baldrick Career Award for their support.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

I. Barnett, R. Mukherjee, and X. Lin. The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association*, 112 (517):64–76, 2017.

S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, P. Hartge, M. Yeager, C. C. Chung, S. J. Chanock, N. Chatterjee, et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics*, 90(5):821–835, 2012.

D. Bu, Q. Yang, Z. Meng, S. Zhang, and Q. Li. Truncated tests for combining evidence of summary statistics. *Genetic Epidemiology*, 44(7):687–701, 2020.

C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

W. Chen, Y. Wu, Z. Zheng, T. Qi, P. M. Visscher, Z. Zhu, and J. Yang. Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors. *Nature Communications*, 12(1):7117, 2021.

L. Darrous, N. Mounier, and Z. Kutalik. Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics. *Nature communications*, 12(1):7274, 2021.

J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

D. Diogo, C. Tian, C. S. Franklin, M. Alanne-Kinnunen, M. March, C. C. Spencer, C. Vangjeli, M. E. Weale, H. Mattsson, E. Kilpeläinen, et al. Phenome-wide association studies across large population cohorts support drug target validation. *Nature communications*, 9(1):1–13, 2018.

H. Feng, N. Mancuso, B. Pasaniuc, and P. Kraft. Multitrait transcriptome-wide association study (twas) tests. *Genetic Epidemiology*, 45(6):563–576, 2021. doi: 10.1002/gepi.22391.

E. Gonzalez, M. Bui, and A. A. Ahmed. IGF1R immunohistochemistry in ewing's sarcoma as predictor of response to targeted therapy. *International Journal of Health Sciences*, 14(4):17, 2020.

O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett, R. Li, T. A. Manolio, S. C. Sanderson, J. Kannry, R. Zinberg, M. A. Basford, et al. The electronic medical records and genomics (emerge) network: past, present, and future. *Genetics in Medicine*, 15(10):761–771, 2013.

G. Hemani, K. Tilling, and G. Davey Smith. Orienting the causal relationship between imprecisely measured traits using gwas summary data. *PLoS Genetics*, 13(11):e1007081, 2017. doi: 10.1371/journal.pgen.1007081. URL `https://doi.org/10.1371/journal.pgen.1007081`.

G. Hemani, J. Zheng, B. Elsworth, K. Wade, D. Baird, V. Haberland, C. Laurin, S. Burgess, J. Bowden, R. Langdon, V. Tan, J. Yarmolinsky, H. Shibab, N. Timpson, D. Evans, C. Relton, R. Martin, G. Davey Smith, T. Gaunt, P. Haycock, and The MR-Base Collaboration. The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, 7:e34408, 2018. doi: 10.7554/eLife.34408. URL `https://elifesciences.org/articles/34408`.

J. B. Holmes, D. Speed, and D. J. Balding. Summary statistic analyses can mistake confounding bias for heritability. *Genetic Epidemiology*, 43(8):930–940, 2019.

A. Kessy, A. Lewin, and K. Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

J. Kim, Y. Bai, and W. Pan. An adaptive association test for multiple phenotypes with gwas summary statistics. *Genetic epidemiology*, 39(8):651–663, 2015.

G. Lahat, A. Lazar, and D. Lev. Sarcoma epidemiology and etiology: potential environmental and genetic factors. *Surgical Clinics of North America*, 88(3):451–481, 2008.

O. W. Lee, C. Rodrigues, S.-H. Lin, W. Luo, K. Jones, D. W. Brown, W. Zhou, E. Karlins, S. M. Khan, S. Baulande, et al. Targeted long-read sequencing of the ewing sarcoma 6p25. 1 susceptibility locus identifies germline-somatic interactions with ewsr1-fli1 binding. *The American Journal of Human Genetics*, 110(3):427–441, 2023.

C. M. Lewis and E. Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Med*, 12(1):44, Dec. 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-00742-5. URL `https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00742-5`.

M. Li and C.-W. Chen. Epigenetic and transcriptional signaling in ewing sarcoma—disease etiology and therapeutic opportunities. *Biomedicines*, 10(6):1325, 2022.

T. Li, Z. Ning, and X. Shen. Improved estimation of phenotypic correlations using summary association statistics. *Frontiers in genetics*, 12:665252, 2021.

X. Li and X. Zhu. Cross-phenotype association analysis using summary statistics from gwas. In *Statistical Human Genetics*, pages 455–467. Springer, 2017.

Z. Liu and X. Lin. Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics*, 74(1):165–175, 2018.

M. J. Machiela, T. G. Grünewald, D. Surdez, S. Reynaud, O. Mirabeau, E. Karlins, R. A. Rubio, S. Zaidi, S. Grossetete-Lalami, S. Ballet, et al. Genome-wide association study identifies multiple new loci associated with Ewing sarcoma susceptibility. *Nature communications*, 9(1):1–8, 2018.

E. McFowland, S. Speakman, and D. B. Neill. Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research*, 14(1):1533–1561, 2013.

D. B. Neill. Fast subset scan for spatial pattern detection: Fast Subset Scan. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, Mar. 2012. ISSN 13697412. doi: 10.1111/j.1467-9868.2011.01014.x. URL `https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2011.01014.x`.

P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

J. Pattee and W. Pan. Penalized regression and model selection methods for polygenic scores on summary statistics. *PLOS Computational Biology*, 16(10):e1008271, Oct. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008271. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008271`. Publisher: Public Library of Science.

S. Postel-Vinay, A. S. Véron, F. Tirode, G. Pierron, S. Reynaud, H. Kovar, O. Oberlin, E. Lapouble, S. Ballet, C. Lucchesi, et al. Common variants near TARDBP and EGR2 are associated with susceptibility to ewing sarcoma. *Nature genetics*, 44(3):323–327, 2012.

J. R. Robinson, J. C. Denny, D. M. Roden, and S. L. Van Driest. Genome-wide and phenome-wide approaches to understand variable drug actions in electronic health records. *Clinical and translational science*, 11(2):112–122, 2018.

D. M. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser, and D. R. Masys. Development of a large-scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369, 2008.

P. Sharada, H. Girish, H. Umadevi, and N. Priya. Ewing's sarcoma of the mandible. *Journal of Oral and Maxillofacial Pathology*, 10(1):31, 2006.

L. G. Spector, A. K. Hubbard, B. J. Diessner, M. J. Machiela, B. R. Webber, and J. D. Schiffman. Comparative international incidence of Ewing sarcoma 1988 to 2012. *International journal of cancer*, 149(5):1054–1066, 2021.

C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12 (3):e1001779, 2015.

P.-U. team, 2020. URL `https://pan.ukbb.broadinstitute.org`.

A. Torkamani, N. E. Wineinger, and E. J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, 2018.

Y. Xu, S. C. Ritchie, Y. Liang, P. R. H. J. Timmers, M. Pietzner, L. Lannelongue, S. A. Lambert, U. A. Tahir, S. May-Wilson, C. Foguet, Johansson, P. Surendran, A. P. Nath, E. Persyn, J. E. Peters, C. Oliver-Williams, S. Deng, B. Prins, J. Luan, L. Bomba, N. Soranzo, E. Di Angelantonio, N. Pirastu, E. S. Tai, R. M. van Dam, H. Parkinson, E. E. Davenport, D. S. Paul, C. Yau, R. E. Gerszten, A. Mälarstig, J. Danesh, X. Sim, C. Langenberg, J. F. Wilson, A. S. Butterworth, and M. Inouye. An atlas of genetic scores to predict multi-omic traits. *Nature*, Mar. 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-05844-9. URL `https://www.nature.com/articles/s41586-023-05844-9`.

W. Zhou, J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. 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.

X. Zhu, T. Feng, B. O. Tayo, J. Liang, J. H. Young, N. Franceschini, J. A. Smith, L. R. Yanek, Y. V. Sun, T. L. Edwards, et al. Meta-analysis of correlated traits via summary statistics from gwass with an application in hypertension. *The American Journal of Human Genetics*, 96(1):21–36, 2015.