Bioinformatics, 2023, **39(6)**, btad370
https://doi.org/10.1093/bioinformatics/btad370
Advance access publication 7 June 2023
Original Paper

OXFORD

Genome analysis

networkGWAS: a network-based approach to discover genetic associations

Giulia Muzio [1,2,*], Leslie O'Bray [1,2], Laetitia Meng-Papaxanthos[1,2,3], Juliane Klatt [1,2], Krista Fischer [4,5], Karsten Borgwardt [1,2,6,*]

[1]Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland
[2]Swiss Institute for Bioinformatics (SIB), 1015 Lausanne, Switzerland
[3]Google Research, Brain Team, 8002 Zürich, Switzerland
[4]Institute of Mathematics and Statistics, University of Tartu, 51009 Tartu, Estonia
[5]Institute of Genomics, University of Tartu, 51010 Tartu, Estonia
[6]Department of Machine Learning and Systems Biology, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

*Corresponding authors. Department of Biosystems Science and Engineering, ETH Zurich, Machine Learning and Computational Biology Lab, Mattenstrasse 26, 4058 Basel, Switzerland. E-mail: giulia.muzio@bsse.ethz.ch (G.M.); Department of Machine Learning and Systems Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Planegg/Martinsried, Germany. E-mail: borgwardt@biochem.mpg.de (K.B.)

Associate Editor: Alfonso Valencia

Abstract

Motivation: While the search for associations between genetic markers and complex traits has led to the discovery of tens of thousands of trait-related genetic variants, the vast majority of these only explain a small fraction of the observed phenotypic variation. One possible strategy to overcome this while leveraging biological prior is to aggregate the effects of several genetic markers and to test entire genes, pathways or (sub)networks of genes for association to a phenotype. The latter, network-based genome-wide association studies, in particular suffer from a vast search space and an inherent multiple testing problem. As a consequence, current approaches are either based on greedy feature selection, thereby risking that they miss relevant associations, or neglect doing a multiple testing correction, which can lead to an abundance of false positive findings.

Results: To address the shortcomings of current approaches of network-based genome-wide association studies, we propose `networkGWAS`, a computationally efficient and statistically sound approach to network-based genome-wide association studies using mixed models and neighborhood aggregation. It allows for population structure correction and for well-calibrated *P*-values, which are obtained through circular and degree-preserving network permutations. `networkGWAS` successfully detects known associations on diverse synthetic phenotypes, as well as known and novel genes in phenotypes from *Saccharomyces cerevisiae* and *Homo sapiens*. It thereby enables the systematic combination of gene-based genome-wide association studies with biological network information.

Availability and implementation: https://github.com/BorgwardtLab/networkGWAS.git.

1 Introduction

Genome-wide association studies (GWAS) aim to identify statistical associations between genetic variants, most commonly in the form of single nucleotide polymorphisms (SNPs), and disease risk or other phenotypes. However, most of the phenotypes of interest are complex traits in the sense that they do not follow a Mendelian pattern of inheritance since they are controlled by multiple SNPs and genes, and are influenced by environmental factors. With respect to such traits, traditional GWAS face the fundamental obstacle of "missing heritability," i.e. single SNPs which were found to be significantly associated often account for a small portion of the variation of heritable phenotypes. In fact, when the development of a certain phenotype involves the interplay of multiple pathways, large parts of missing heritability could be due to genetic interactions rather than directly corresponding to undetected association with genetic variants (Zuk *et al.* 2012). Therefore, a great effort has been undertaken to develop more comprehensive and powerful GWAS methodologies, aiming at understanding and incorporating biological mechanisms underlying the genetics of complex traits. To date, the rich knowledge about biological networks, which is already available, such as protein–protein interaction (PPI) and gene regulatory networks, is rarely leveraged in a statistical and rigorous way in GWAS. Including such contextual and functional information, representing processes relevant to the phenotype under study, constitutes a promising approach to overcome the problem of missing heritability. This strategy, in fact, can enable an increase in statistical power as well as an improvement in interpretability in GWAS aimed at complex traits.

The problem of limited power in GWAS is generally rooted in both a large marker-to-sample ratio and low heritability of complex traits. In order to mitigate that, two strategies have

Received: July 29, 2022. **Revised:** April 11, 2023. **Editorial Decision:** May 21, 2023. **Accepted:** June 6, 2023

been pursued: (i) to group genetic markers and test set of markers at once, thereby reducing the multiplicity of markers tested (Holden *et al.* 2008, Li and Leal 2008, Schwender *et al.* 2011, Listgarten *et al.* 2013), or (ii) to employ biological networks in order to conduct a *post hoc* aggregation of association (Ideker *et al.* 2002, Akula *et al.* 2011, Azencott *et al.* 2013, Greene *et al.* 2015, Shim and Lee 2015, Wang *et al.* 2015, Shim *et al.* 2017, Carlin *et al.* 2019). Both approaches amplify the signal of SNPs or genes, which are collectively phenotype-related but would not pass the significance threshold on their own. However, within the set-based test strategy, so far, SNP sets are typically chosen based on membership to a functional unit on the genome. Hence, this strategy lacks a principled procedure to select SNP sets that goes beyond single genes or mere regions on the genome. The *post hoc* aggregation strategies, on the other hand, present other limitations, depending on the algorithm they are based upon. For example, the methods based on greedy search risk to miss potential associations. They, in fact, start the search for associated subnetworks from "seed genes" on the network, and then expand toward the most associated genes in their neighborhoods, until some conditions are met. Therefore, signal located on genes far away from the seeds can potentially be ignored. Furthermore, multiple testing correction is generally not applicable to such methods given the dynamics in the search and the presence of 2^n hypothesis, with n being the number of nodes. Lastly, other *post hoc* tools provide scores rather than *P*-values, which do not allow to assess the results in a statistically rigorous way.

We propose to combine both strategies and thereby overcome their respective weaknesses. More precisely, our approach entails testing sets of SNPs, as done e.g. by the FaST-LMM-Set method (Listgarten *et al.* 2013), but we guide the SNP selection by means of biological networks. Specifically, we define the set of subnetworks to test for association by means of a neighborhood aggregation-based strategy. Despite limiting the search space *a priori*, this approach is efficient to compute and results in having a clear number of hypothesis, allowing to perform proper multiple testing correction, in contrast to *post hoc* methods. Thus, we arrive at a strategy, `networkGWAS`, that incorporates both a biologically meaningful way to select SNP sets that goes beyond functional units, and that yields statistically rigorous *P*-values for the SNP sets tested, obtained through circular and degree-preserving network permutations.

2 `networkGWAS`

In this section, we detail how we exploit the biological network information, we discuss the mathematical model we use and the details of how we obtain *P*-values.

2.1 Neighborhood aggregation

We test pre-defined sets of SNPs, rather than single SNPs, in order to both reduce the number of markers tested and to account for gene interaction in addition to mere genetic variance. Multiple methods performing SNP set-based tests already exist, including gene enrichment analysis (Holden *et al.* 2008), collapsing methods (Li and Leal 2008), multivariate regression (Schwender *et al.* 2011) and linear mixed models (LMMs) (Listgarten *et al.* 2013). However, none of them incorporates biological network structure in order to guide the SNP-set selection, thereby choosing SNP sets that are not representative of biological mechanisms. In our approach, instead, we select SNP sets to be tested based on PPI networks, i.e. the graph representation of the interactions between proteins. PPIs are essential for almost all biological mechanisms, and are defined as the specific, non-generic, physical contact between proteins in a particular biological context (Rivas and Fontanillo 2010). These interactions can be both stable (e.g. as in multi-enzyme complexes) or transient [e.g. as in interaction with kinases (Junker and Schreiber 2011)]. Since we focus on complex phenotypes, we employ the entire PPI network, including both stable and transient interactions, thus capturing effects of molecular mechanisms taking place in diverse cells and tissues, and of various kinetics.

More precisely, each sample i in the GWAS dataset is represented as a graph $G_i = (V, E)$, where V is the set of nodes and E is the set of edges in the PPI network. In these graphs, the nodes V represent genes, and edges E indicate any kind of PPI between gene products of the two nodes they connect. Since each sample uses the same PPI network, the topology is shared, i.e. (V, E) is the same for each sample. The node features, however, vary depending on the sample. Each node $v \in V$ is attributed with a feature vector $\vec{a}_i(v)$ comprising the values of all SNPs overlapping with the corresponding gene. Based on this representation, one SNP set per gene is constructed by means of concatenating the feature vector of the gene itself as well as its k-hop neighbor genes according to the PPI network. As a result, the node feature vector $\vec{l}_i(v)$ of a node v and a sample i is now represented by the union of its own SNPs and those from its k-hop neighborhood $\mathcal{N}_k$,

$$\vec{l}_i(v) = \bigcup_{v' \in \mathcal{N}_k(v)} \vec{a}_i(v'). \tag{1}$$

This neighborhood aggregation operation is visualized in Fig. 1a. We thereby directly test the significance of biological subnetworks to identify pathways underlying complex phenotypes. In summary, our neighborhood aggregation approach is akin to the idea underlying graph kernels (Borgwardt *et al.* 2020) or graph convolutional networks (GCNs) (Kipf and Welling 2017). All of these methods leverage localized first-order approximations of subgraph structure in order to avoid an exhaustive search of all subgraphs, which would scale exponentially in the size of the network at hand, whereas our approach is linear in the number of nodes, even when the k-hop neighborhood is defined to be >1. Hence, this approach represents a good compromise between leveraging the biological network information and enumerating all subgraphs. Note that in all the experiments we perform, we employ 1-hop neighborhoods to define our SNP sets.

2.2 Model

Once SNP sets have been selected in the aforementioned manner, we employ a FaST-LMM-Set like model (Lippert *et al.* 2014) to estimate the statistical associations with the phenotype of choice. The LMM we use,

$$\vec{y} = X_\mathrm{f} \times \vec{\beta}_\mathrm{f} + V_\mathrm{s} \times \vec{w}_\mathrm{s} + \vec{\epsilon}, \tag{2}$$

features one random effect $V_\mathrm{s} \times \vec{w}_\mathrm{s}$, which accounts for the similarity among the SNPs of the set to be tested. The precise form of V_s depends on the similarity measure chosen and will be specified by means of Equations (4)–(6) below. Above, $\vec{y}$ contains the continuous phenotype values of the n individuals

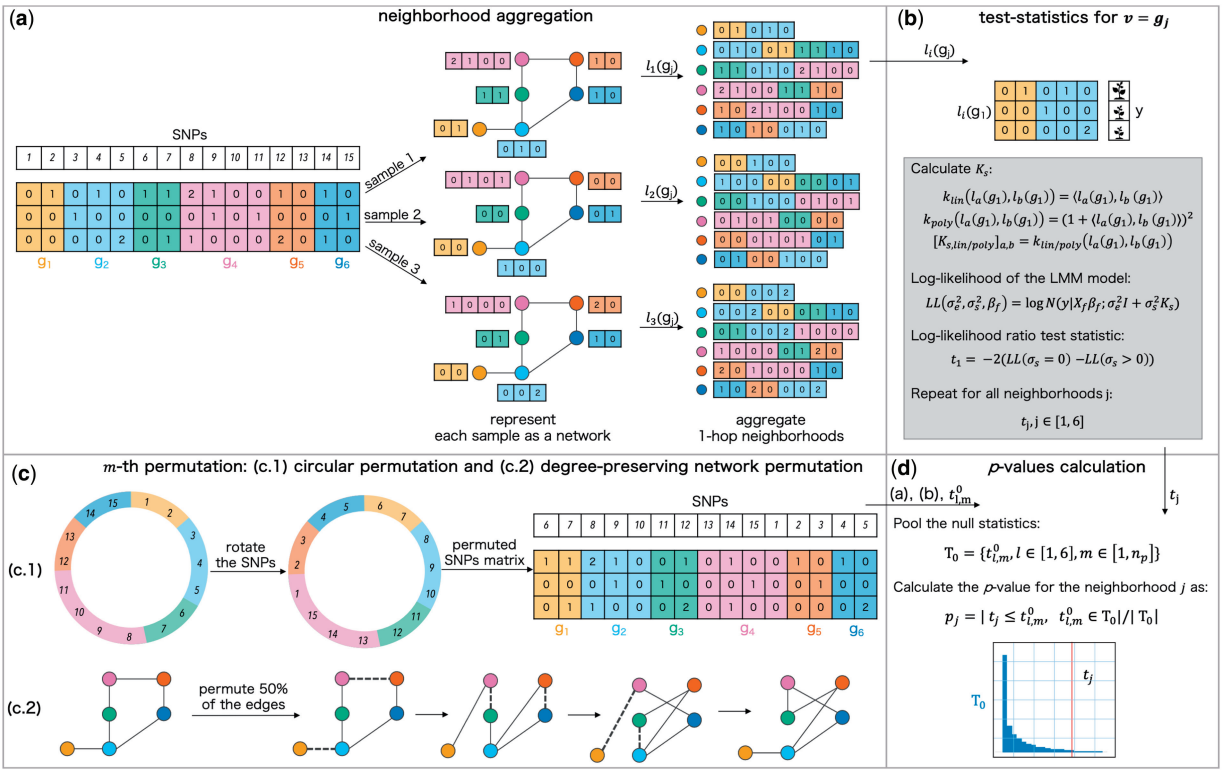

Figure 1. Overview of `networkGWAS`: (a) given a SNP matrix, a PPI network, and a mapping of SNPs onto the genes g_j (color-coded), the first step, i.e. the neighborhood aggregation, begins by representing each of the samples as a network which shares a fixed topology from the PPI network but differs in the node feature values. Subsequently, the 1-hop neighborhood aggregation of the features is performed for each sample i, resulting in each node j being labeled with $l_i(g_j)$, i.e. the concatenation of its own features and the features of its 1-hop neighbors in the PPI network. (b) Per each gene j, either K_{lin} or K_{poly} is calculated from $l_i(g_j)$, $\forall i \in [1,3]$. Then, the values of Equation (2) for $\sigma_s = 0$ and $\sigma_s > 0$ are estimated via restricted maximum likelihood, and, subsequently, the likelihood-ratio test statistics t_j, $j \in [1,6]$ are obtained. (c) The distribution of the test statistics under the null hypothesis of no association signal between the neighborhoods and the phenotype is derived via a permutation procedure, which combines a circular permutation (c.1) with a degree-preserving network permutation (c.2). Having obtained the permuted settings, steps (a) and (b) are performed to obtain a test statistic $t^0_{l,m}$ per neighborhood l and permutation m. (d) $t^0_{l,m}, \forall l \in [1,6]$ and $\forall m \in [1, n_p]$ are pooled to obtain the null distribution $\mathcal{T}_0$. Afterwards, a P-value per each neighborhood is estimated by calculating the ratio between the number of null statistics $t^0_{l,m}$ that are greater than or equal to the t_j, as obtained on the non-permuted setting, divided by the total number of $t^0_{l,m} \in \mathcal{T}_0$.

studied, X_f is the $n \times n_f$ matrix of n_f fixed effects (e.g. a column of 1 s corresponding to the intercept and other covariates, see Supplementary Section S1; here and throughout the manuscript, all section, figure and table numbers starting with an "S" refer to the Supplementary), $\vec{\beta}_f$ denotes the vector of the n_f fixed effect weights, $\vec{w}_s$ comprises the signal, i.e. the random effects of the n_s SNPs of interest and included in the pre-defined SNP set to be tested, and $\vec{\epsilon}$ models residual noise. $\vec{w}_s$, and $\vec{\epsilon}$ are assumed to be drawn from multivariate Gaussian distributions, respectively $\mathcal{N}(\vec{0}; \sigma_s^2 \mathbb{I})$ and $\mathcal{N}(\vec{0}; \sigma_e^2 \mathbb{I})$, where σ_s^2 represents the genetic variance, σ_e^2 the residual variance and $\mathbb{I}$ the identity matrix. Marginalizing over fixed effects, the log-likelihood of the model (2) reads

$$LL(\sigma_e^2, \sigma_s^2, \vec{\beta}_f) = \log \mathcal{N}(\vec{y} | X_f \vec{\beta}_f; \sigma_e^2 \mathbb{I} + \sigma_s^2 K_s), \tag{3}$$

where K_s is the covariance matrix capturing the similarities among the SNPs in the test set. The parameter σ_s serves to distinguish the null model (i.e. $\sigma_s = 0$) from alternative models (i.e. $\sigma_s > 0$), and is estimated from the GWAS dataset by means of restricted maximum likelihood. While in the original FaST-LMM-Set, K_s measures similarity through a linear kernel $k^{(\text{lin})}$, we additionally explore the use of a quadratic kernel $k^{(\text{poly})}$ for K_s in our method:

$$k^{(\text{lin})}(\vec{l}_i(v), \vec{l}_j(v)) = \langle \vec{l}_i(v), \vec{l}_j(v) \rangle, \tag{4}$$

$$k^{(\text{poly})}(\vec{l}_i(v), \vec{l}_j(v)) = (1 + \langle \vec{l}_i(v), \vec{l}_j(v) \rangle)^2, \tag{5}$$

$$[K_s^{(\text{lin/poly})}]_{i,j} = k^{(\text{lin/poly})}(\vec{l}_i(v), \vec{l}_j(v)), \tag{6}$$

where $\vec{l}_i(v)$ and $\vec{l}_j(v)$ are defined according to Equation (1) for the i-th and the j-th sample. We chose an inhomogeneous polynomial kernel in order for it to be able to capture both linear and non-linear similarity. In an additional deviation from FaST-LMM-Set, we normalize the diagonal entries of our final kernel matrix $\tilde{K}_s$ to be 1, by means of the following equation:

$$[\tilde{K}_s]_{ij} = [K_s]_{ij} / \sqrt{[K_s]_{ii}[K_s]_{jj}}. \tag{7}$$

Note that the definition of V_s in Equation (2) varies depending on the type of the kernel we use. If the kernel is defined following Equation (4), e.g. the linear kernel, V_s corresponds to the $n \times n_s$ design matrix whose rows correspond to $l_i(v), i = 1, \ldots, n$, normalized to be consistent with Equation (7). When the kernel is instead calculated according to Equation (5), V_s constitutes a higher-dimensional vector in the reproducing kernel Hilbert space, which additionally

accounts for the $n \times n_s(n_s - 1)$ second-order interactions among the SNPs in the test set.

2.3 *P*-value computation

`networkGWAS` aims to identify neighborhoods that are statistically associated with a trait of interest. In the following, we refer to the k-hop neighborhood of gene v_j as $\mathcal{N}_k(v_j)$, with $j = 1, \ldots, n_g$ and n_g equal to the total number of genes. We then define a null hypothesis H_j for each neighborhood, which signifies that $\mathcal{N}_k(v_j)$ does not affect the phenotype. Explicitly, this states that the set of SNPs representing $\mathcal{N}_k(v_j)$, i.e. $l_i(v), i = 1, \ldots, n$, does not exhibit an association signal with the trait under study. Note that under the null hypothesis H_j, the patterns of linkage disequilibrium (LD) among the SNPs and the population structure, if present, must be preserved. LD patterns represent the non-random correlations between the SNPs due to, e.g. spatial proximity on the genome, while population structure refers to the complex relatedness among the individuals, which impacts the values and structures of both the SNPs and the phenotypes. Having defined H_j, we further need to

1) define a measure, i.e. a test statistic, to quantify the association signal between each neighborhood and the phenotype,
2) obtain the null distribution of the test statistics underlying the null hypothesis,
3) estimate the *P*-values and
4) define a strategy to identify the statistically significantly associated neighborhoods.

Since we rely on FaST-LMM-Set for our SNP set-based test, the association between $\mathcal{N}_k(v_j)$ and the phenotype is quantified by calculating the log-likelihood ratio between the maximum restricted likelihood estimate of the alternative and null models from Equation (3). The test statistic obtained in this way for $\mathcal{N}_k(v_j)$ is referred to as t_j. Unlike the original FaST-LMM-Set method, which uses a parametric null distribution, we adhere to a non-parametric distribution for `networkGWAS` instead. The reasoning behind this choice is detailed in Supplementary Section S3. We determine the distribution of test statistics under the null hypothesis by means of a permutation strategy that allows to destroy the association signal between the sets of SNPs (i.e. the neighborhoods) and the phenotype, while simultaneously preserving LD patterns and population structure. We achieve this by performing permutations on both the SNPs and the network level, namely a circular permutation of the SNPs and a degree-preserving permutation of the network, both of which are detailed in the following.

Circular permutation of the SNPs. The implementation of the circular permutation procedure is inspired by Cabrera *et al.* (2012). We consider the SNPs to be ordered according to their genomic position in a circular way, namely that after the last SNP on the last chromosome, one restarts from the first SNP on the first chromosome. A single permutation is then performed by randomly selecting a number between 1 and the total number of SNPs (which in our case would be the total number of SNPs across all the n_g neighborhoods), and then rotating the SNPs by that value, while keeping the genomic coordinates fixed. A visualization of this is reported in Fig. 1c.1. This operation causes the SNPs to be assigned to different genomic positions compared to their original location on the genome (therefore positionally mapping them to a different gene compared to the non-rotated scenario), while conserving the same position with respect to each other, hence preserving the LD patterns among the SNPs. This operation sufficiently preserves any confounding population structure since the relative position between SNPs is again maintained. Furthermore, the relatedness signal is also preserved when constructing the null distribution since the arrangement of the individuals in the SNPs to test V_s (or interactions of SNPs to test), phenotype (y) and fixed effects (X_f) is not varied. Note that when performing multiple circular permutations, only non-repeating rotations are considered.

Degree-preserving network permutation. We permute the network on top of permuting the SNPs to add a level of randomization when performing the neighborhood aggregation operation. The aim is to prevent the network structure from creating spurious association signal. This network permutation consists of a methodology that allows us to shuffle the edges while maintaining the degree of each node preserved, thereby enforcing a comparable graph structure despite the permutations. Consider a network $G = (V, E)$, where V is the set of nodes $V = \{v_j, j = 1, \ldots, n_g\}$, and E is the set of edges. To generate one permutation of the network, the following steps are performed until 50% of the edges are rearranged:

1) Randomly select two pairs of connected genes, (v_a, v_b) and (v_c, v_d).
2) Check whether v_a or v_b are connected with v_c or v_d; if so, return to Point 1, otherwise proceed to Point 3.
3) Remove the edge between v_a and v_b, and v_c and v_d.
4) Connect v_a with v_c, and v_b with v_d.

Figure 1c.2 shows an example of this permutation technique. Note that the only parameter to set for this two-level permutation strategy is the percentage of edges to shuffle, which we set to 50%.

After having permuted the SNPs and the network following the aforementioned circular and degree-preserving permutations, respectively, we can again define neighborhoods on these permuted scenarios (as detailed in Section 2.1), and, subsequently, we can calculate a test statistic for each of them. Specifically, we can obtain $t^0_{l,m}$ for the l-th neighborhood and the m-th permutation as described above. By pooling these test statistics obtained from all neighborhoods under all permutations, we obtain an empirical test-statistic distribution under the null hypothesis (Zhang *et al.* 2010). Note that with this procedure, we obtain n_g statistics per each permutation, decreasing the total number of permutations n_p required (discussed in Supplementary Section S4).

Having determined t_j and the null distribution of these statistics under the null hypothesis, i.e. $\mathcal{T}_0 = \{t^0_{l,m}, l = 1, \ldots, n_g, m = 1, \ldots, n_p\}$, we calculate the *P*-value for $\mathcal{N}_k(v_j)$ as $p_j = |t_j \leq t^0_l, t^0_l \in \mathcal{T}_0|/|\mathcal{T}_0|$. Note that for avoiding significance values of zero in case $t^0_{l,m} < t_j, \forall t^0_{l,m} \in \mathcal{T}_0$, a pseudocount can be added. This procedure allows us to obtain calibrated *P*-values, which measure the significance of the statistical association of a particular neighborhood of interacting genes and the phenotype. For further details on the obtained *P*-values in the different experiments and applications, the reader is referred to Supplementary Section S3.

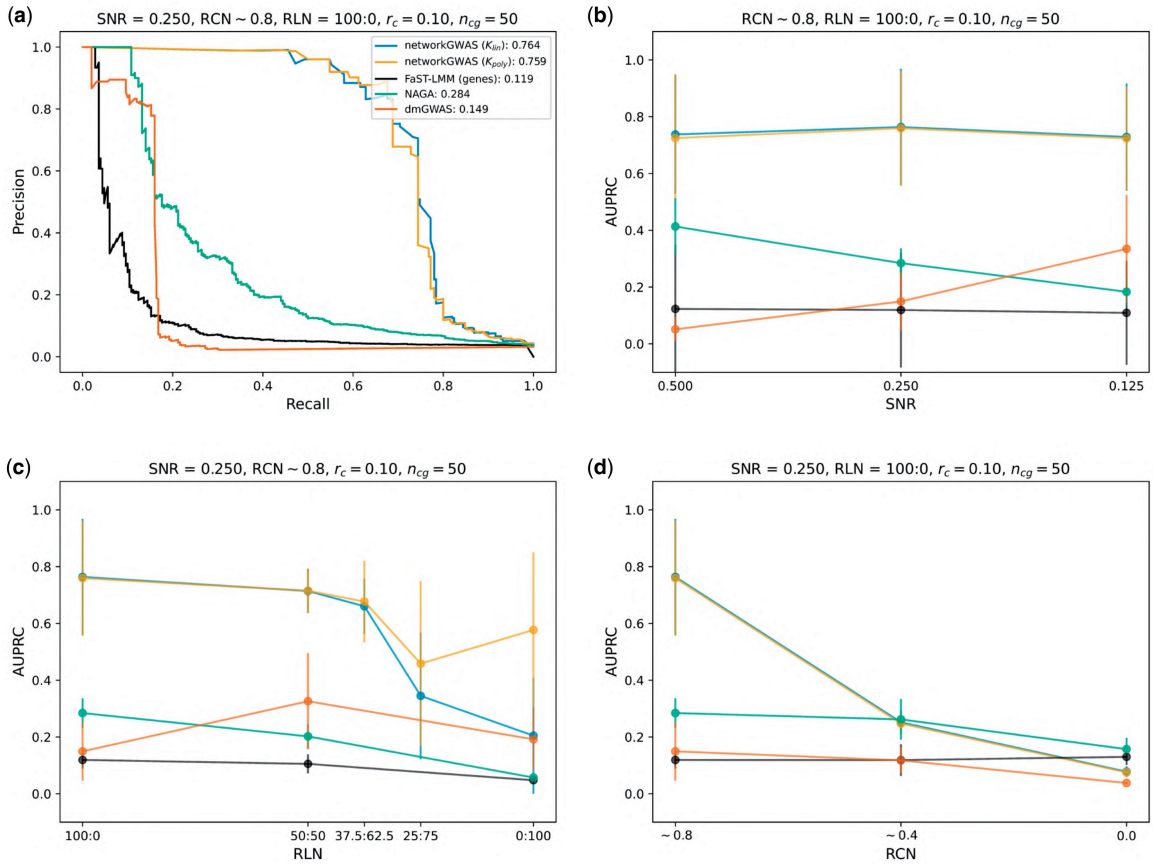

Figure 2. Results from simulating the phenotypes of *A.thaliana*. We present results from our method (`networkGWAS`), using either a linear (K_{lin}) or a polynomial (K_{poly}) kernel, as well as the performance of other comparison methods. In (a), we show the AUPRC of the baseline scenario (S0). In (b–d), we vary one variable while keeping the other four fixed: the SNR (b), the RLN (c) and the RCN (d). The AUPRC in function of the ratio of causal SNPs on a causal gene (r_c) and the number of causal genes (n_{cg}) is shown in the Supplementary Material.

In a last step, we identify the neighborhoods that are statistically associated with the phenotype of interest. When studying a particular organism, often multiple related phenotypes are available. Hence, we are testing multiple hypotheses on two levels, namely n_g neighborhoods for T traits. Therefore, to account for this aspect when correcting our significance level for multiple testing, we employ the hierarchical testing procedure proposed by Peterson *et al.* (2016) (detailed in Supplementary Section S5), which is based on the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995) and allows to effectively control the false-discovery rate (FDR) in such scenarios.

Lastly, the computational cost is in the order of $\mathcal{O}(n_p n_g n n_s^2)$. Supplementary Section S2 provides the details and Supplementary Table S1 reports the running times.

3 Simulations

In this section, we detail the simulations studies we perform, designed such to best demonstrate the robustness and limitations of `networkGWAS` under varying conditions, and we introduce the state-of-the-art methods we contrast our results to. In particular, we apply our method and the comparison partners on semi-simulated common variants settings and fully synthetic rare variants settings. While the first is presented in the following, the rare variants use case is discussed in Supplementary Section S8.

3.1 Experimental setup

In order to apply our method, one needs a GWAS dataset consisting of genotypes and a phenotype of interest, as well as a PPI network relevant to the phenotype chosen. In our experiments, we employ natural genotypes and a PPI network in combination with simulated phenotypes. This allows us to test our method on genotypes with realistic LD and MAF patterns and on a biological network with sensible structure, while also enabling us to have a ground truth to compare our results to. We use the genotype dataset for *Arabidopsis thaliana* from the AraGWAS Catalog (Togninalli *et al.* 2018) and the PPI network from The Arabidopsis Information Resource (TAIR) database (Lamesch *et al.* 2012). Supplementary Section S6.1 provides details on the data and processing. Note that, we chose the TAIR network for our semi-simulated setting since its smaller size allows us to run a large number of fast experiments.

To simulate the phenotypes, we firstly define the following parameters: (i) the number of genes n_{cg} that carry causal SNPs, henceforth called causal genes, (ii) the ratio of causal SNPs on a causal gene, r_c, (iii) the mean ratio of causal neighbors (RCN) of a causal gene, (iv) the signal-to-noise ratio (SNR) and (v) the mixing ratio of linear-to-non-linear (RLN) signal. Note that the causal genes (SNPs) are selected among the 1327 genes (37 458 SNPs) that remain after the processing. Therefore, the simulated genotype–phenotype associations are modeled within this subset of the genome. All the above defined parameters are systematically varied in our

experiments, as exemplified in Supplementary Section S7. Thus, we investigate the robustness of our method's and comparison partners' performance as we depart from the most amenable scenario (S0) of a purely linear signal, spread across a single or very few causal subgraph(s) composed of a high number of causal genes, with a high RCN, a realistic SNR and a high ratio of causal SNPs on the causal genes. Supplementary Table S2 summarizes the simulated scenarios, starting from our anchor scenario S0 and moving toward more challenging/different RCN, RLN, SNR, r_c or n_{cg}. For each scenario, we simulate five phenotypes as:

$$\vec{y} = cX \times \vec{\beta} + (1-c)X^{(2)} \times \vec{\beta}^{(2)} + \vec{\epsilon}, \tag{8}$$

where X is the n by p matrix of all SNPs across the genes considered in the analysis (i.e. 37 458 in this case), $\vec{\beta}$ represents the fixed effect of these SNPs, $\vec{\epsilon}$ models the noise, $X^{(2)}$ is the n by $p(p-1)/2$ second-order design matrix of all SNP interactions, $\vec{\beta}^{(2)}$ comprises the fixed effects of all SNP interactions, and the coefficient c serves to tune between the RLN signal. The choice of c and $\vec{\beta}$ leads to the realization of the different scenarios, as detailed in Supplementary Section S7.

3.2 Comparison partners and performance evaluation

We compare networkGWAS's performance to: (i) FaST-LMM-Set approach (Listgarten *et al.* 2013), however, with sets based on single genes rather than neighborhoods of interacting genes as defined by the PPI network, (ii) NAGA (Carlin *et al.* 2019) and (iii) dmGWAS (Wang *et al.* 2015). Both NAGA and dmGWAS incorporate PPI information following a *post hoc* strategy. In fact, they both commence with a classical GWAS analysis to obtain single-SNP *P*-values. Subsequently, dmGWAS employs a greedy-selection based, dense module searching, aiming to find PPI subnetworks enriched in low *P*-value SNPs. NAGA, on the other hand, first represents and scores entire genes based on their most significant SNP and then relies on a PPI-network propagation approach in order to spread and revise scores across gene-neighborhoods. While networkGWAS and FaST-LMM-Set approach return a *P*-value per each neighborhood or gene, dmGWAS and NAGA provide a score for the subnetworks and genes, respectively. Having obtained these *P*-value-based and score-based rankings, we evaluate the performance of linear and non-linear networkGWAS as well as the comparison partners by means of their respective mean area under the precision-recall curve (AUPRC) in terms of causal genes. Here, linear and non-linear refers to the SNP-set kernel employed [see Equation (5)], and the mean refers to the average with respect to the different random realizations of the various simulation settings.

3.3 Results

An overview of the results is shown in Fig. 2. The top left panel depicts the full mean precision-recall curves for all methods studied under the conditions of our anchor scenario S0, which is most amenable to network-guided search for genotypic–phenotypic associations (see Supplementary Table S2). In this scenario both linear and non-linear networkGWAS substantially outperform all comparison partners by achieving an AUPRC of 76.4% ± 20.5% and 75.9% ± 20.3%, more than doubling the average AUPRC of the strongest competitor, NAGA, which averagely reports an AUPRC of 28.4% ± 5.2%. When dividing networkGWAS's AUPRC by the prevalence, 3.8% in this scenario, we obtain a factor of about 20 for both linear and non-linear networkGWAS, further highlighting networkGWAS's strong performance in our anchor scenario. Furthermore, it is worth noting that networkGWAS presents higher recall in comparison to dmGWAS per each precision value, highlighting the improvement in the identification of relevant associations achievable by employing neighborhood aggregation instead of greedy search-based algorithm.

As demonstrated in the top right panel of Fig. 2, this dominance in performance is invariant as one departs from the conditions of S0 by varying the SNR while keeping the other simulation parameters fixed. Similarly, and as shown in the bottom left panel, the performance of non-linear networkGWAS is robust with respect to tuning the signal from purely linear to purely non-linear, while keeping the SNR, RCN, r_c and n_{cg} identical to those of scenario S0, again outperforming all comparison partners across the entire range of RLNs. Remarkably, the performance of linear networkGWAS starts to differ from that of its non-linear counter-part, only for signals which are dominantly non-linear and pars the non-linear networkGWAS up until an RLN of 37.5 : 62.5. This is in line with observations made elsewhere: approaches designed to detect statistical significance of single loci will miss those with modest marginal effects and large interactions (Marchini *et al.* 2005). Lastly, the strong performance of networkGWAS and its dominance over the comparison partners breaks down as we tune the RCN from ∼0.8 to 0.0, while keeping the SNR, RLN, r_c and n_{cg} the same as in scenario S0. This is shown in the bottom right panel of Fig. 2, and corresponds to a gradual transition from a few, large causal subgraphs in the PPI network, via multiple medium-sized causal subgraphs, to many isolated causal genes. This shows that networkGWAS' performance decreases with the decrease in the average percentage of causal neighbors in the neighborhoods. In the extreme scenario of this percentage being 0%, the signal, by construction, is independent of the PPI-network structure, resulting in the fact that network-guided methods cannot exploit the network structure to enhance performance. The reduction of the ratio of causal SNPs on a causal gene does not decrease the performance of any method since the SNR remains constant, as visible in Supplementary Fig. S1, left panel. Instead, when the signal is concentrated on a fewer number of genes, i.e. when we reduce the number of causal genes while maintaining the SNR unvaried, the performance of networkGWAS remains untouched, while FaST-LMM-Set and NAGA drastically improve, suggesting that incorporating the biological network structure can substantially improve the results when the signal is spread across a high number of causal genes (e.g. 50), as it is for complex traits. These results are reported in Supplementary Fig. S1.

4 Applications

To illustrate networkGWAS's ability to allow for the discovery of new statistically significant genotype–phenotype associations, we apply networkGWAS to natural phenotypes from *A.thaliana*, *Saccharomyces cerevisiae* and *Homo sapiens*. Specifically, we study these phenotypes with networkGWAS

and its comparison partners, i.e. NAGA, dmGWAS, FaST-LMM-Set, and we additionally employ traditional univariate GWAS (Lippert *et al.* 2011). To identify statistical associations in presence of *P*-values, we use the hierarchical procedure (Peterson *et al.* 2016) detailed in Supplementary Section S5, which allows to correct for multiple testing while controlling the FDR, whose level we set at 0.05. In the absence of a ground truth, we evaluate the findings of our method by (i) comparing our results to already published studies, (ii) comparing the gene-neighborhoods identified as significantly associated by our method with the genes identified by the comparison partners and (iii) investigating the potential biological relevance of identified genes in processes related to the phenotype.

4.1 Application to *A.thaliana*

When searching for associations with respect to natural phenotypes, we continue to use the *A.thaliana* genotype data, but rely on the larger STRING database for the PPI network (Szklarczyk *et al.* 2021). The phenotypes are selected from the AraPheno (Seren *et al.* 2017) database (see Supplementary Section S6.3 for details).

For most phenotypes, there are no statistically significant associations found by any method. However, for phenotype 704, the univariate GWAS alone returns statistically significant SNPs. These SNPs are located on seven genes not connected through the PPI network (see Supplementary Section S10.1). This scenario of association signal located on genes not connected through the PPI network is reminiscent of the simulation setting S6, where `networkGWAS` suffers since it cannot exploit the biological network information. This shows that, depending on the nature of the phenotype under study, it might not always be beneficial to incorporate the biological network information.

4.2 Application to *S.cerevisiae*

In addition to *A.thaliana*, we test our method on *S.cerevisiae*. Both the GWAS dataset and the phenotypes have been obtained from the study by Peter *et al.* (2018). The PPI network has been downloaded from the STRING database (Szklarczyk *et al.* 2021). More details in Supplementary Section S6.4.

Linear `networkGWAS` returns two statistically significant neighborhoods on the phenotype YPGALACTOSE, which represents the growth ratio between a stress condition and the standard growing condition for *S.cerevisiae* (see Supplementary Table S4 for details). These two identified neighborhoods represent a connected subgraph composed of 265 genes. We analyze these findings from different perspectives. As a preliminary investigation, we contrast them with what has been surfaced by the comparison partners and with the results presented in the original study (Peter *et al.* 2018). As for *P*-value providing methods, FaST-LMM-Set (i.e. the analysis of genes) reports no statistically significant findings for any of the phenotypes, while traditional univariate GWAS identifies statistical associations for nine of them [Supplementary Table S5 compares these findings with what reported by Peter *et al.* (2018)]. For YPGALACTOSE, however, the latter returns no associations. Since NAGA and dmGWAS do not provide *P*-values, but rather scores, we identify the associated genes and subnetworks for YPGALACTOSE by following the strategies proposed by the respective authors, namely selecting the first 100 top-ranked genes for NAGA and the top 1% subnetworks for dmGWAS. This leads us to have a collective interception of 10 genes with `networkGWAS`' findings. The comparison with the traditional GWAS reported in Peter *et al.* (2018) shows no overlap between the genes surfaced by `networkGWAS` and the genes reported by this study. The latter are isolated genes according to the PPI network, hence violating the fundamental assumption of network-based methods designed for finding interacting genes that collectively carry association signal.

To biologically interpret `networkGWAS`' results on YPGALACTOSE, we rely on the PANTHER Classification System (Thomas *et al.* 2003). Specifically, we use the PANTHER Over-representation Test (PANTHER version 17.0), with PANTHER GO-slim Biological Process as the annotation dataset and Fisher's exact test with FDR correction as the test type. We find that the genes identified by `networkGWAS` are significantly enriched (maximum *P*-value 4.70e-02) in processes related to (i) DNA replication, (ii) chromatin organization, (iii) the cell cycle and (iv) DNA transcription. All of these categories of processes are known to be affected when the *S.cerevisiae* organism undergoes a stress condition (Pardo *et al.* 2016, Crawford and Pavitt 2019). This demonstrates that `networkGWAS` is capable of identifying neighborhoods of genes that are involved in biological processes related to the analyzed phenotype.

We furthermore explore the identified neighborhoods by means of a *post hoc* linear association. We begin with a univariate association analysis on the 3549 SNPs included in the two neighborhoods, which led to no significance. Motivated by this result, we hypothesized the signal being of multivariate nature, and performed a Lasso analysis. We obtained the signal coming from 32 SNPs located on 25 genes (Supplementary Table S6). Interestingly, the locations of these genes are distributed across 10 different chromosomes, which highlights the benefits of including the PPI-network information as a means to lead the creation of the sets of SNPs to test.

4.3 Application to *H.sapiens*

To further validate our approach, we test `networkGWAS` and its comparison partners on human genetics data from the Estonian BioBank GWAS data (Leitsalu *et al.* 2015). We focus on three phenotypes, i.e. type II diabetes (T2D) diagnosis, height and body mass index (BMI). The numbers of the samples are 199 466, 148 144 and 136 772, respectively, proving the scalability of our method. As for the PPI network, we select high-confidence PPIs for the *H.sapiens* protein coding genes from the STRING database (Szklarczyk *et al.* 2021). Supplementary Section S6.5 reports additional details on the data and the preprocessing steps.

On the T2D phenotype, linear `networkGWAS` finds nine significantly associated neighborhoods (FDR$<$0.01) composed of 144 genes (see Supplementary Section S10.3), which represent a connected subgraph on the PPI network. To assess the biological relevance of these findings, we utilize the WEB-based GEne SeT AnaLysis Toolkit (WEBGESTALT) (Wang *et al.* 2013), since it directly allows to query diseases databases. The results show that 42 genes are already known to be involved in T2D-related processes. The remaining 102 candidate genes identified by `networkGWAS` are only partially surfaced by the comparison partners. In fact, traditional GWAS, NAGA and dmGWAS combined can only identify eight of them. Instead, the analysis of genes with FaST-LMM-Set reports no significant results, highlighting the benefit of

adding the network information in this context. A preliminary analysis of these candidate genes reports that many of them are involved in the Wnt signaling, which is known to be related to T2D (Chen *et al.* 2021). Additional results in Supplementary Section S9.

On height and BMI, `networkGWAS` reports no statistically significant neighborhoods. The traditional GWAS analysis, however, finds statistical associations for both phenotypes, suggesting that the beneficial effect of incorporating the PPI-network information might depend on the biology underlying the phenotypes under analysis.

5 Discussion

We have defined a principled way to perform gene-based GWAS utilizing network information as a prior in the process of testing statistical associations, allowing us to directly and in a statistically rigorous manner obtain *P*-values for entire PPI-based gene-neighborhoods, which represent biological pathways. This conceptually differs from the state-of-the-art network-based GWAS methods, which use the PPI information as a way of post-selection. We have demonstrated the superior performance of our PPI-network-based SNP set-based test, `networkGWAS`, compared to state-of-the-art SNP set-based methods (Listgarten *et al.* 2013) and approaches that incorporate PPIs (Wang *et al.* 2015, Carlin *et al.* 2019). Moreover, we have done so in a wide range of simulation settings for rare and common variants including very low SNRs, i.e. very low heritability, different numbers of SNPs/genes carrying the association signal and various mixtures of linear and non-linear signal. Concerning the latter, it is worth noting that none of the comparison partners can incorporate an explicit search for SNP interactions, which our method is capable of thanks to the use of a non-linear SNP-set kernel. `networkGWAS` is only outperformed if our underlying assumption, that the SNPs in neighborhoods of interacting genes are collectively related to the phenotype of interest, is strongly violated.

We furthermore have employed `networkGWAS` to study various phenotypes of *A.thaliana*, *S.cerevisiae* and *H.sapiens*. On the *S.cerevisiae* and *H.sapiens* phenotypes, `networkGWAS` finds collectively significantly associated genes that were almost entirely undiscovered by its strongest competitor, NAGA, demonstrating the complementarity of the methods. In addition, it is particularly noteworthy that, in both cases, `networkGWAS` is capable of finding biologically plausible associations when the single-gene-based SNP-set method does not, highlighting the value of incorporating the PPI-network information. When analyzing the *A.thaliana* phenotypes, instead, `networkGWAS` finds no statistically associated neighborhoods. Although this might be perceived as a discouraging result, the studied phenotypes do not necessarily present a genetic basis or association signal in the selected searching space. Hence, the fact that `networkGWAS` provides a statistically sound approach to identify the presence (if any) of significant associations is an actual advantage in comparison to the network-based comparison partners, which would output the top genes or subnetworks even on a phenotype presenting noise rather than actual signal. Furthermore, the presence of *P*-values allows to rigorously account for multiple testing correction.

As already pointed out, the application of `networkGWAS` on *S.cerevisiae* and *H.sapiens* exemplifies the benefit of exploiting the biological network information. However, PPIs supported by multiple evidence, i.e. high-confidence interactions, are not available (yet) for all the genes of a particular organism. For example, for *A.thaliana* only ∼56% of the genes participate in known high-confidence PPIs, while for *S.cerevisiae*, high-confidence interactions are known for ∼92% of the genes. This may explain the results on *A.thaliana* in the sense that the current PPI network is likely incomplete and potentially associated neighborhoods of genes have not yet been identified as forming a neighborhood due to yet undiscovered interactions. Fortunately, as the knowledge of biological pathways and gene–gene interaction increases, more and more evidence will be gained and PPIs discovered, thus further approaching the state of complete PPI networks. Another aspect to consider is that `networkGWAS` can only discover the collective signal of SNPs, which can be mapped onto genes. This drawback can be addressed by additionally applying traditional methods, such as univariate GWAS, to such SNPs, benefiting from increased test power due to a reduced search space.

`networkGWAS` provides *P*-values for the association of entire gene-neighborhoods, and cannot single-out precise genes or SNPs within such neighborhoods as more or less strongly contributing to that association signal. If one is interested in exploring this aspect, *post hoc* analysis needs to be performed and should be guided by the kernel K_s chosen in `networkGWAS`. That is, when a linear kernel was applied, one linear *post hoc* analysis is appropriate, as is done for the findings obtained on YPGALACTOSE. When, instead, a polynomial kernel was applied, the associated neighborhoods present non-linear signal and epistasis search represents a viable route for *post hoc* analysis.

A potential limitation of the approach presented here lies within the choice of testing one-hop neighborhoods. While testing only direct interactions of a given gene is meaningful from a biological perspective, we acknowledge that the neighborhood depth technically constitutes a choice of hyperparameter. While the use of k-hop neighborhoods for $k \geq 2$ needs to be investigated in the future, based on our simulations, we are confident that at least in medium-to-high RCN scenarios, `networkGWAS` is already capable of detecting signal that is spread further than across the 1-hop neighbors of a causal center-gene. This is supported by `networkGWAS`' findings on the *S.cerevisiae* phenotype YPGALACTOSE, as well as on T2D phenotype from *H.sapiens*. Indeed, the statistically associated neighborhoods are connected through the PPI network in both cases. We note that even if $k \geq 2$, our method still scales linearly with the number of nodes in the network.

Another area for further research is to exploit the kernelized nature of `networkGWAS` and use more complex kernels matrices K_s in our LMM given in Equation (2). When designing such kernels, one may either (i) continue to focus on the SNP content of genes, and experiment with the type of non-linearity, or (ii) depart from solely using SNPs as features and instead leverage the information in gene properties, such as the number of minor alleles on the SNPs belonging to a gene, in combination with graph kernels or GCNs. Lastly, `networkGWAS` as presented here can be considered an instance of a more fundamental framework, which can be naturally extended and be used, e.g. to study the same phenotype under different perspectives: (i) by utilizing various biological networks (e.g. gene co-expression or multi-omics networks), and (ii) by employing diverse ways to map the SNPs to the genes (e.g. chromatin mapping). In this sense, `networkGWAS` represents a very versatile tool.

Acknowledgements

The analysis on the Estonian BioBank data was performed under the supervision of K.F. during G.M.'s visit at the University of Tartu. Data collection, genotyping, quality control and imputation on the EstBB data was performed by the Estonian Biobank research team (Andres Metspalu, Lili Milani, Tõnu Esko, Reedik Mägi, Mari Nelis, and Georgi Hudjashov).

Ethics statement

The activities of the Estonian Biobank (EstBB) are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of the EstBB. Individual level data analysis in the EstBB was carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs).

Supplementary data

Supplementary data is available at *Bioinformatics* online.

Conflict of interest

K.B. is a co-founder and scientific advisory board member of Computomics GmbH, Tübingen.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813533. This study was supported in part by the Alfried Krupp Prize for Young University Teachers of the Alfried Krupp von Bohlen und Halbach-Stiftung (Borgwardt); and in part by the Estonian Research Council (Fischer) [grant number PRG-1197].

Data availability

The S. cerevisiae and A. thaliana data (genotype, phenotypes, gene annotations and PPI networks) are publicly available, as detailed in the respective sections and Supplementary sections. The gene annotations and PPI network from H. sapiens as well. Access to the EstBB genotype and phenotype data can be requested following the instructions at https://genomics.ut.ee/en/content/estonian-biobank. The code is public and available on GitHub at https://github.com/BorgwardtLab/networkGWAS.git.

References

Akula N, Baranova A, Seto D *et al.*; Bipolar Disorder Genome Study (BiGS) Consortium. A network-based approach to prioritize results from genome-wide association studies. *PLoS One* 2011;**6**:e24220.

Azencott C-A, Grimm D, Sugiyama M *et al.* Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics* 2013;**29**:i171–9.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Stat Methodol* 1995;**57**:289–300.

Borgwardt K, Ghisu E, Llinares-López F *et al.* Graph kernels: state-of-the-art and future challenges. *FNT in Mach Learn* 2020;**13**:531–712.

Cabrera CP, Navarro P, Huffman JE *et al.* Uncovering networks from genome-wide association studies via circular genomic permutation. *G3 (Bethesda)* 2012;**2**:1067–75.

Carlin DE, Fong SH, Qin Y *et al.* A fast and flexible framework for network-assisted genomic association. *iScience* 2019;**16**:155–61.

Chen J, Ning C, Mu J *et al.* Role of Wnt signaling pathways in type 2 diabetes mellitus. *Mol Cell Biochem* 2021;**476**:2219–32.

Crawford RA, Pavitt GD. Translational regulation in response to stress in *Saccharomyces cerevisiae*. *Yeast* 2019;**36**:5–21.

Greene CS, Krishnan A, Wong AK *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* 2015;**47**:569–76.

Holden M, Deng S, Wojnowski L *et al.* GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 2008;**24**:2784–5.

Ideker T, Ozier O, Schwikowski B *et al.* Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002;**18**:S233–40.

Junker BH, Schreiber F. *Analysis of Biological Networks*, Vol. **2**. Hoboken, New Jersey, U.S: John Wiley & Sons, 2011.

Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations, April 24-26, 2017, Toulon, France*. 2017. https://arxiv.org/abs/1609.02907.

Lamesch P, Berardini TZ, Li D *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 2012;**40**:D1202–10.

Leitsalu L, Haller T, Esko T *et al.* Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* 2015;**44**:1137–47.

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;**83**:311–21.

Lippert C, Listgarten J, Liu Y *et al.* FaST linear mixed models for genome-wide association studies. *Nat Methods* 2011;**8**:833–5.

Lippert C, Xiang J, Horta D *et al.* Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* 2014;**30**:3206–14.

Listgarten J, Lippert C, Kang EY *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* 2013;**29**:1526–33.

Marchini J, Donnelly P, Cardon LR *et al.* Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005;**37**:413–7.

Pardo B, Crabbé L, Pasero P *et al.* Signaling pathways of replication stress in yeast. *FEMS Yeast Res* 2016;**17**:fow101.

Peter J, De Chiara M, Friedrich A *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* 2018;**556**:339–44.

Peterson CB, Bogomolov M, Benjamini Y *et al.* Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies. *Genet Epidemiol* 2016;**40**:45–56.

Rivas JDL, Fontanillo C. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* 2010;**6**:e1000807.

Schwender H, Ruczinski I, Ickstadt K *et al.* Testing SNPs and sets of SNPs for importance in association studies. *Biostatistics* 2011;**12**:18–32.

Seren Ü, Grimm D, Fitz J *et al.* AraPheno: a public database for *Arabidopsis thaliana* phenotypes. *Nucleic Acids Res* 2017;**45**:D1054–9.

Shim JE, Bang C, Yang S *et al.* GWAB: a web server for the network-based boosting of human genome-wide association data. *Nucleic Acids Res* 2017;**45**:W154–61.

Shim JE, Lee I. Network-assisted approaches for human disease research. *Anim Cells Syst* 2015;**19**:231–5.

Szklarczyk D, Gable AL, Nastou KC *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**:D605–12.

Thomas PD, Campbell MJ, Kejariwal A *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 2003;**13**:2129–41.

Togninalli M, Seren Ü, Meng D *et al.* The AraGWAS catalog: a curated and standardized *Arabidopsis thaliana* GWAS catalog. *Nucleic Acids Res* 2018;**46**:D1150–6.

Wang J, Duncan D, Shi Z *et al.* WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 2013;**41**:W77–83.

Wang Q, Yu H, Zhao Z *et al.* EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics* 2015;**31**:2591–4.

Zhang X, Huang S, Zou F *et al.* TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* 2010;**26**:i217–27.

Zuk O, Hechter E, Sunyaev SR *et al.* The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2012;**109**:1193–8.