networkGWAS: A network-based approach for genome-wide association studies in structured populations

Working Paper

Author(s):
Muzio, Giulia; O'Bray, Leslie; Meng-Papaxanthos, Laetitia; Klatt, Juliane; Borgwardt, Karsten

Publication date:
2021-11-13

Permanent link:
https://doi.org/10.3929/ethz-b-000525237

Originally published in:
bioRxiv, https://doi.org/10.1101/2021.11.11.468206

Funding acknowledgement:
813533 - Machine Learning Frontiers in Precision Medicine (EC)

networkGWAS: A network-based approach for genome-wide association studies in structured populations

Giulia Muzio[1,2][0000−0001−5999−2030], Leslie O'Bray[1,2][0000−0001−8999−9962], Laetitia Meng-Papaxanthos[1,2,3][0000−0002−0521−621X], Juliane Klatt[1,2][0000−0003−4096−566X], and Karsten Borgwardt[1,2][0000−0001−7221−2393]

[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland
[3] Google Research, Brain Team

Abstract. While the search for associations between genetic markers and complex traits has discovered tens of thousands of trait-related genetic variants, the vast majority of these only explain a tiny fraction of observed phenotypic variation. One possible strategy to detect stronger associations 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 suffers from a huge 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, and/or neglect doing a multiple testing correction, which can lead to an abundance of false positive findings. 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 gene-based genome-wide association studies based on mixed models and neighborhood aggregation. It allows for population structure correction and for well-calibrated p-values, which we obtain through a block permutation scheme. **networkGWAS** successfully detects known or plausible associations on simulated rare variants from *H. sapiens* data as well as semi-simulated and real data with common variants from *A. thaliana* and enables the systematic combination of gene-based genome-wide association studies with biological network information.
Availability: `https://github.com/BorgwardtLab/networkGWAS.git`

Keywords: GWAS · biological networks · computational biology · graph kernels · neighborhood aggregation.

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 are controlled by multiple SNPs and genes and do not follow Mendelian inheritance, or are influenced by environmental factors. Traditional GWAS face the fundamental obstacle of missing heritability with respect to such traits, that is, the fact that the variation of heritable phenotypes may only be poorly explained by the single SNPs found to be significantly associated. As previously argued, large parts of missing heritability could be due to genetic interactions–if the development of a certain phenotype involves interaction among multiple pathways–rather than directly correspond to undetected association with genetic variants [43]. 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, GWAS rarely make use of the already available and rich knowledge about biological networks–such as protein-protein interaction (PPI) and gene regulation networks–representing processes relevant to the respective phenotype under study. Including such contextual and functional information can enable an increase in statistical power as well as interpretability in GWAS aimed at complex traits, thus representing a promising approach to overcome the missing heritability problem.

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 been pursued: (i) to group genetic markers and test them at once, thereby reducing the multiplicity of markers tested [13, 18, 21, 28, 40], or (ii) to employ biological networks in order to conduct a *post hoc* aggregation of association [5, 14, 2]. 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 strategy, on the other hand, suffers from the absence of statistically sound p-values for the set of aggregated SNPs. We propose to combine both strategies and thereby overcome their respective weaknesses. More precisely, our approach entails testing sets of SNPs, as done for example by the FaST-LMM-Set method [21], but we guide the SNP selection by means of biological networks. Thus, we arrive at a strategy 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.

The remainder of the manuscript is structured as follows: In Section 2, we detail the model we employ as well as explain how we do the p-value computation. We then provide an overview of the characteristics of the *A. thaliana* GWAS data set and PPI networks we apply our method to. Section 3 then lays out the specifics of the GWAS we conduct with respect to (semi-)simulated and natural data, as well as the baseline methods we compare against. Lastly, results are summarized in Section 4, and the limitations as well as key ingredients to overcoming them are discussed in the concluding Section 5.

2 `networkGWAS`

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 [13], collapsing methods [18], multivariate regression [28], and linear mixed models (LMMs) [21, 40]). 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 protein-protein interaction (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 [26]. These interactions can be both stable (e.g., as in multi-enzyme complexes) or transient (e.g., as in interaction with kinases [15]). Since we focus on complex phenotypes, such as the growth of the *A. thaliana* model organism, 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 data set 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. Since each sample uses the same PPI network, the topology is shared, i.e. (V, E) is the same for each sample. 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. Each node $v \in V$ is attributed with a feature vector $a(v)$ comprising the values of all SNPs overlapping with the corresponding gene. The samples are thus differentiated by these node attributes. 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 label vector $l(v)$ of a node v is now represented by the union of its own SNPs and those from its k-hop neighborhood $\mathcal{N}_k$,

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

We thereby directly test biological subnetworks' significance to identify pathways underlying complex phenotypes. In summary, our neighborhood aggregation approach is akin to the idea underlying graph kernels [32, 31, 4] or graph convolutional networks (GCNs) [16]. 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 greater than 1.

2.2 Model

In the following we discuss our mathematical model and details of how we obtain p-values as well as the computational cost involved.

Linear mixed model Once SNP sets have been selected in the aforementioned manner, we employ a FaST-LMM-Set like model [21] in order to estimate statistical association with the phenotype of choice. The LMM underlying the original FaST-LMM-Set method,

$$\vec{y} = X \cdot \vec{\beta} + \tfrac{1}{\sqrt{n_c}} V_c \cdot \vec{w}_c + \tfrac{1}{\sqrt{n_s}} V_s \cdot \vec{w}_s + \vec{\epsilon}, \tag{2}$$

features two random effects: one to capture confounders (V_c) and another to account for similarity among the SNPs of the set to be tested (V_s). Above, the vector $\vec{y}$ contains continuous phenotype values of the n individuals studied, X is the $n \times p$ design matrix, $\vec{\beta}$ comprises the fixed effects of all p SNPs included in the PPI network, w_c contains the random effects of the n_c SNPs from which relatedness is estimated, 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. $\vec{w}_c$, $\vec{w}_s$, and $\vec{\epsilon}$ are assumed to be drawn from multivariate Gaussian distributions $\mathcal{N}(\vec{0}; \sigma_c^2 I)$, $\mathcal{N}(\vec{0}; \sigma_s^2 I)$, and $\mathcal{N}(\vec{0}; \sigma_e^2 I)$, respectively. Marginalizing over random effects, and re-parametrizing random effects as a convex combination of two variance components, the log-likelihood of the model (2) reads

$$LL = \log \mathcal{N}(\vec{y} | X\vec{\beta}; \sigma_e^2 I + \sigma_g^2 [(1-\tau)K_c + \tau K_s]) \,, \tag{3}$$

with covariance matrices $K_c = \frac{1}{n_c} V_c V_c^T$ and $K_s = \frac{1}{n_s} V_s V_s^T$. As introduced in [21], the parameter $\tau \in [0, 1]$ serves to distinguish the null model (i.e., $\tau = 0$) from alternative models (i.e., $\tau \neq 0$), and is estimated from the GWAS data set by means of restricted maximum likelihood.

SNP-set kernel Both K_c and K_s are kernel matrices, the latter of which effectively regresses the set of SNPs of interest in a multivariate manner to estimate the statistical dependence between these genetic markers and the target phenotype. While in the original FaST-LMM-Set both K_c and K_s measure similarity through a linear kernel k_{lin}, we additionally use a quadratic kernel k_{poly} for K_s in our method,

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

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

$$[K_{s,\text{lin/poly}}]_{i,j} = k_{\text{lin/poly}}(\vec{v}_i, \vec{v}_j) \,. \tag{6}$$

Note that above we chose an inhomogeneous polynomical 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:

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

Population structure correction The kernel matrix K_c serves as a genetic similarity matrix (GSM) measuring and correcting for population structure in the form of a realized relationship matrix (RRM) [10, 12]. In order to yield the computational savings that reside at the core of the FaST-LMM procedure [19], the sum of the dimensions of the kernel matrices must be kept well below the number n of samples studied. To this end, the GSM K_c is constructed from a limited number of SNPs chosen on the basis of their p-values in an uncorrected linear regression with respect to the phenotype of interest. To identify the precise number of SNPs to be included in the GSM, the latter is constructed with an increasing number of SNPs–starting with those associated with the highest p-values–until the first minimum in the genomic inflation factor λ–defined as the ratio of the median observed to median theoretical test statistic–is met [19]. Note that in contrast to the original FaST-LMM-Set procedure, we do not include any SNPs to correct for population structure if $\lambda \leq 1.2$ in the uncorrected case, and deflation–if present–is only observed in the high p-value range. Lastly, in order to avoid reduced power caused by proximal contamination [20, 27], and identical to the procedure described in [21], SNPs included in the set of interest, plus those within a two-centimorgan buffer zone, are removed from the GSM if population-structure correction is applied.

2.3 p-value computation

Since we rely on FaST-LMM-Set for our SNP set-based test, p-values on the SNP sets are obtained based on a maximum-likelihood test statistic, comparing the maximum restricted likelihood of the alternative and null models as defined above. The original FaST-LMM-Set method follows the spirit of Wilks' theorem [38] and results by Greven *et al.* [11] and employs a mixture of χ^2 distributions,

$$p_0(x) = a\chi_0(x) + b\chi_d(x) \tag{8}$$

to serve as parametric distribution of the test statistic under the null hypothesis. The parameters are determined by fitting $p_0(x)$ to the 10% most significant tail of the null distribution of test statistics obtained by permuting individuals for only the SNPs in the set of interest.

While the above strategy saves computation time by decreasing the number of permutations needed, we found the resulting estimate of the distribution $p_0(x)$ of test statistics under the null hypothesis to not reflect the true distribution in our case. Similar limitations of the above parametric approach in LMMs have been discussed in [24], and hence we adhere to a non-parametric distribution for `networkGWAS` instead. More specifically, we determine the distribution of test statistics under the null hypothesis by means of a bootstrap that is realized through a block permutation strategy. To generate a single null test statistic for a neighborhood, we permute blocks of SNPs which are located next to each other on a chromosome and do so only for genes included in the SNP set of interest, i.e., the PPI neighborhood to be tested. Because we do not permute single SNPs but blocks of them, the pattern of linkage disequilibrium (LD) is preserved if the chosen block size is large enough. By permuting only SNP blocks within the test set while keeping those outside the network as they were, we also sufficiently preserve any confounding population structure. Note that with this procedure, we obtain as many null test statistics as we have genes in the PPI network, decreasing the number of permutations required. Pooling these test-statistics obtained by all SNP sets under any permutation, we obtain an empirical test-statistic distribution under the null hypothesis [42]. The latter then provides us with calibrated p-values measuring the significance of the statistical association of a particular neighborhood of interacting genes and the phenotype. Both the SNP-block size n_b and the number of permutations n_p needed are determined by visual inspection of quantile-quantile plots with respect to the observed calibrated p-values and the expected p-values under the null hypothesis. Both n_b and n_p are increased until the p-values look calibrated and $\lambda \leq 1.2$. Lastly, we correct our significance level for testing multiple SNP sets by means of the Benjamini-Hochberg procedure [3], which is known to control the false-discovery rate (FDR).

2.4 Computational cost

We base our method on the FaST-LMM-Set procedure [21], which in turn exploits the factored spectrally transformed linear mixed models (FaST-LMM) algorithm [19]. The latter achieves GWAS that scale linearly–instead of cubically–with the number of samples n in both run time and memory use given that (i) the sum of the number n_s of SNPs in the test set and number n_c of SNPs used to construct the the GSM is less than the cohort size, (ii) a factorizing genetic similarity matrix, such as the RRM, is used, and (iii) the SNP-set kernel is linear. In this best-case scenario, performing one test per neighborhood has a runtime of $\mathcal{O}(n_g n)$, where n_g is the number of genes in the PPI network and n is the cohort size. Before applying FaST-LMM-Set, however, networkGWAS requires the aggregation of gene neighborhoods which scales linearly in the number of edges of the PPI employed. As biological networks can be assumed to be only sparsely connected, this leads to a runtime of order $\mathcal{O}(n_g)$. Secondly, in order to obtain calibrated p-values from the test statistics, we have to repeat both the neighborhood aggregation and the n_g set tests for each of the n_p permutations. Lastly, for the polynomial kernel, run times change again as either the trivial factorization in the reproducing kernel Hilbert space of dimension $n_s(n_s-1)/2$ has to be employed or a factorization at cost $\mathcal{O}(n_3)$ has to be performed. This results in a computational cost of $\mathcal{O}\left(n_p n_g n\right)$ if $n > n_c + n_s$ for the linear kernel or $n > n_c + n_s(n_s-1)/2$ for the quadratic kernel. Otherwise, the cost amounts to $\mathcal{O}\left(n_p n_g n^3\right)$. Hence, while the cubic scaling is not favorable for large cohort sizes, asymptotically, run time scales linearly in n. For the *A. thaliana* data, the median number n_s of SNPs per neighborhood is 295 and of the same magnitude as n_c which ranges in the low hundreds of SNPs. Hence, for the majority of *A. thaliana* PPI neighborhoods tested, the cubic-to-linear speed up sets in for cohort sizes as small as $n \gtrsim 500$ for linear networkGWAS, while for non-linear networkGWAS that speed-up would be achieved for cohorts of $n \gtrsim 50,000$ samples.

3 Experiments

In order to apply our method, one needs a GWAS data set consisting of genotypes and a phenotype of interest, as well as a PPI network relevant to the phenotype chosen. In the following, we present results obtained with our method on three different data sets: semi-simulated data for the *A. thaliana* model organism, natural data for the *A. thaliana* model organism, and fully simulated rare variant scenarios for the *H. sapiens* data. While the (semi-)simulations are designed such as to best demonstrate the robustness and limitations of our method under varying conditions, the application to natural phenotypes serves to illustrate networkGWAS's ability to allow for the discovery of new statistically significant genotype-phenotype associations which could not have been found by means of traditional GWAS nor existing network or gene-based methods. In the fully and semi-simulated scenarios, we compare our results against (i) the original FaST-LMM-Set approach [21], however, with sets based on single genes rather than neighborhoods of interacting genes as defined by the PPI network, (ii) NAGA [5], and (iii) dmGWAS [14], whilst we leave out dmGWAS and, instead, employ univariate GWAS when studying the *A. thaliana* phenotypes. Both NAGA and dmGWAS commence with a classical GWAS analysis to obtain single-SNP p-values. Subsequently, dmGWAS employs 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. Hence, while both dmGWAS and NAGA incorporate PPI information, the conceptual difference between these two methods and ours is that they use such information as a way of post-selection, rather than exploiting it as a prior in the process of testing statistical associations. Therefore, unlike NAGA and dmGWAS, we directly and in a statistically rigorous manner obtain p-values for entire PPI-based gene neighborhoods, which represent biological pathways. Furthermore, none of the three comparison partners can incorporate an explicit search for SNP interactions significantly associated with the phenotype, which our method–by means of employing a non-linear SNP-set kernel–is capable of. In both the (semi-)simulated and natural phenotype experiments, we employ one-hop neighborhoods to define our SNP sets.

3.1 Simulated phenotypes for *A. thaliana*

As a GWAS data set, we make use of the AraGWAS Catalog [37]. AraGWAS constitutes a manually curated and standardized GWAS catalog for all publicly available phenotypes from AraPheno [29], which is a central repository of population-scale phenotypes for *A. thaliana* inbred lines. For the PPI network in our semi-simulated scenarios, where we use natural genotypes in combination with simulated phenotypes, we used the one provided by The Arabidopsis Information Resource (TAIR) [17], which due to its smaller size, allowed us to run a large number of fast experiments.

We start off with the fully-imputed data set provided by AraGWAS, which comprises 2,029 samples of 10,709,466 SNPs each and then filter for at least 5% minor-allele frequency, which leaves us with 1,763,004 remaining SNPs. Out of those, 37,458 can be mapped – using the strict positional mapping of the genes downloaded from TAIR database [17] – to the 1,327 genes included in the TAIR PPI network [17]. Our experiments are set up in a semi-simulated manner. With the objective to test our method on data presenting realistic LD patterns, the genotypes used are the ones extracted from AraGWAS, while the phenotypes are artificial. In order to simulate phenotypes, we firstly define the following parameters: (a) the number of genes n_{cg} that carry causal SNPs, henceforth called causal genes, (b) the ratio of causal SNPs on a causal gene, r_c, (c) the mean ratio of causal neighbors (RCN) of a causal gene, (d) the signal-to-noise-ratio (SNR), and (e) the mixing ratio of linear-to-nonlinear (RLN) signal.

While we keep the number of causal genes ($n_{cg} = 50$) and ratio of causal SNPs ($r_c = .1$) within such genes constant, the RCN, SNR and RLN are systematically varied in our experiments. 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) with a high ratio of causal neighbors, and high signal-to-noise ratio. Table 1 summarizes the scenarios simulated, starting from our anchor scenario S0 and deviating from its conditions by (i) varying the SNR while keeping the RCN and RLN constant, (ii) varying the RLN while keeping the SNR and RCN constant, and (iii) varying the RCN while keeping the SNR and RLN constant. We thereby study the isolated effects of moving towards more challenging RCN, RLN and SNR, respectively. In order to realize the predefined scenarios, we follow the procedure below for the simulation of the artificial phenotype:

1. Choice of causal genes:
   (a) For RCN equal to 0./0.4/0.8, randomly select n_{cg}/5/1 node(s) in the network and define them as causal.
   (b) Randomly include 0%/40%/100% of the k-hop neighbors of the starting node(s), then 0%/40%/100% of the k-hop neighbors of the newly included nodes and so on, until the pre-defined number n_{cg} of causal genes is reached.
   (c) Should the starting node(s) belong to a disconnected subgraph(s) smaller than as to allow for the inclusion of n_{cg} causal genes, repeat (a)-(c) until n_{cq} causal genes have been defined.
   (d) Compute RCN, and if RCN is not in $[0.]$ / $[0.4, 0.5]$ / $[0.8, 1.0]$, disregard and repeat (a)-(d).

2. Choice of causal SNPs and causal interactions:
   (a) Randomly select a ratio of r_c of the SNPs of each causal gene as carrying the signal.
   (b) For each causal subgraph, define one of the nodes with highest degree as center of the subgraph. For each causal SNP belonging to a non-center node of the same causal subgraph, randomly select a causal SNP on the center node and define the interaction between the two SNPs as causal. Ensure that each center-node causal SNP has at least one interaction partner.

3. Computation of the artificial phenotype:
   (a) Simulate the phenotype as
   $$\vec{y} = cX \cdot \vec{\beta} + (1-c)X^{(2)} \cdot \vec{\beta}^{(2)} + \vec{\epsilon} \tag{9}$$
   with X, $\vec{\beta}$, and $\vec{\epsilon}$ defined as in Equation (2), $X^{(2)}$ representing the n by $p(p-1)/2$ second-order design matrix of all SNP interactions, $\vec{\beta}^{(2)}$ comprising the fixed effects of all SNP interactions, and the coefficient c serving to tune between the ratio of linear to non-linear signal. $\vec{\epsilon}$ is drawn from a multivariate normal distribution $\mathcal{N}(\vec{0}, I)$.
   (b) Set β_i to a fixed and positive value b if the i^{th} SNP is a causal one, or to zero otherwise. Choose b such that for $c = 1$, the SNR equals the desired value.
   (c) Set $\beta_i^{(2)}$ to a fixed and positive value $b^{(2)}$ if the i^{th} SNP-interaction is a causal one, or to zero otherwise. Choose $b^{(2)}$ such that for $c = 0$ the SNR equals the desired value.
   (d) Choose c such that the desired RLN is realized.

Table 1: Overview of simulation settings for artificial phenotypes. While we keep the number of causal genes ($n_{cg} = 50$) and ratio of causal SNPs ($r_c = .1$) constant, the RCN, SNR and RLN are systematically varied.

	S0	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Ratio of causal neighbors (RCN)	0.8	0.8	0.8	0.8	0.8	0.4	0.0
Mixing ration of linear to non-linear signal (RLN)	100 : 0	100 : 0	100 : 0	50 : 50	0 : 100	100 : 0	100 : 0
Signal-to-noise ratio (SNR)	0.250	0.125	0.500	0.250	0.250	0.250	0.250

Table 2: Overview of *A. thaliana* phenotypes. Abbreviations: RLF = relative lifetime fitness, RSP = relative seed production, PT = precipitation treatment.

	id	**phenotype**	**cohort size n**	**original study**
N01	518	RLF in Spain for high PT and low density	512	[9]
N02	522	RLF in Germany for high PT and low density	512	[9]
N03	523	RLF in Germany for high PT and high density	512	[9]
N04	524	RLF in Germany for low PT and low density	512	[9]
N05	534	RSP in Spain for high PT and low density	511	[9]
N06	535	RSP in Spain for high PT and high density	512	[9]
N07	538	RSP in Germany for high PT and low density	512	[9]
N08	539	RSP in Germany for high PT and high density	511	[9]
N09	700	length of main flowering stem	680	[6]
N10	704	rosette leaf number	850	[6]
N11	705	cauline leaf number	904	[6]
N12	707	diameter of rosette	656	[6]

Tuning the signal-to-noise ratio (SNR) in our simulation allows us to mimic a range of heritability values h by means of the relation SNR $= h/(1-h)$. With low but realistic heritability values ranging from $h = 0.1$ to $h = 0.4$, SNRs to be investigated stretch from SNR $= 0.1$ to SNR $= 0.67$, which includes all the values investigated by us. We evaluate the performance of all methods by means of their respective median precision and recall in terms of causal genes, where the median is obtained with respect to the different random realizations corresponding to one simulation setting. Lastly, note that we create 5 random realizations of each scenario S0-S8, in order to estimate the variance in performance.

3.2 *A. thaliana* phenotypes

When searching for associations with respect to natural phenotypes, we continue to use the *A. thaliana* model organism, but rely on the larger STRING database for the PPI network. [34]. We select phenotypes from the AraPheno database based on the criteria that they are quasi-continuous (such as stem length, and flower diameter) and be known for at least 500 out of the 2,029 *A. thaliana* samples that we use the genotypes of. This leads to the selection of twelve phenotypes which are summarized in Table 2. For each phenotype, we collect the fully-imputed genotypes provided by AraGWAS for which the phenotype of interest has been determined. We then filter for at least 5% minor-allele frequency among those samples which–depending on the phenotype–leaves us with 1,837,440±27,477 remaining SNPs. Out of those, 535,177±9,875 could be positionally mapped [17] to the 25,490 genes included in the PPI network provided by the STRING database [34]. Note that interactions in the STRING PPI network are annotated with one or more 'scores' which are indicators of confidence, i.e. how likely STRING judges an interaction to be true, given the available evidence. These scores rank from 0 to 1, with 1 being the highest possible confidence. For our analysis we only consider high-confidence interactions with a score larger than or equal to 0.7.

In the absence of the ground truth, we evaluate the performance of our method in contrast to the comparison partners by (i) counting how many more genes we identify as significantly associated with the phenotype of interest, and (ii) investigating the potential biological relevance of these additional genes in processes related to the phenotype. Note that, strictly speaking, for (i) we are comparing the number of genes belonging to a gene neighborhood that is identified as significantly associated by our method with the number of genes directly identified as significantly associated by the comparison partners.

3.3 Simulated rare variants in *H. sapiens*

With the aim of simulating rare variants with realistic LD and MAF distributions, we use the sim1000G package [8], which requires as unique input the variant call format (VCF) file of the genomic region of interest. We choose the VCF file for chromosome 10 from Phase III 1000 genomes sequencing data [1], and we use PLINK [25, 30] to extract the variant calls for the European population. We employ bcftools [7] to further manipulate the VCF file, namely to filter out the variants that are not mapped – according to the MyGene database [39, 41] – onto the 398 interacting genes through the human STRING PPI network [34], restricted to chromosome 10. Having obtained the desired VCF file, we therefore use sim1000G to simulate 10,000 rare variants, e.g. MAF ≤ 0.1, from 500 unrelated subjects. The generation of the synthetic phenotypes follows the same steps and parameter choices as detailed in Section 3.1, apart from n_{cg}, which is set to 15 to have the same causal/non-causal genes ratio as for the *A. thaliana* use case. Lastly, the simulation settings S0 to S6 are chosen identical to the common-variant semi-simulations and are listed in Table 1.

4 Results

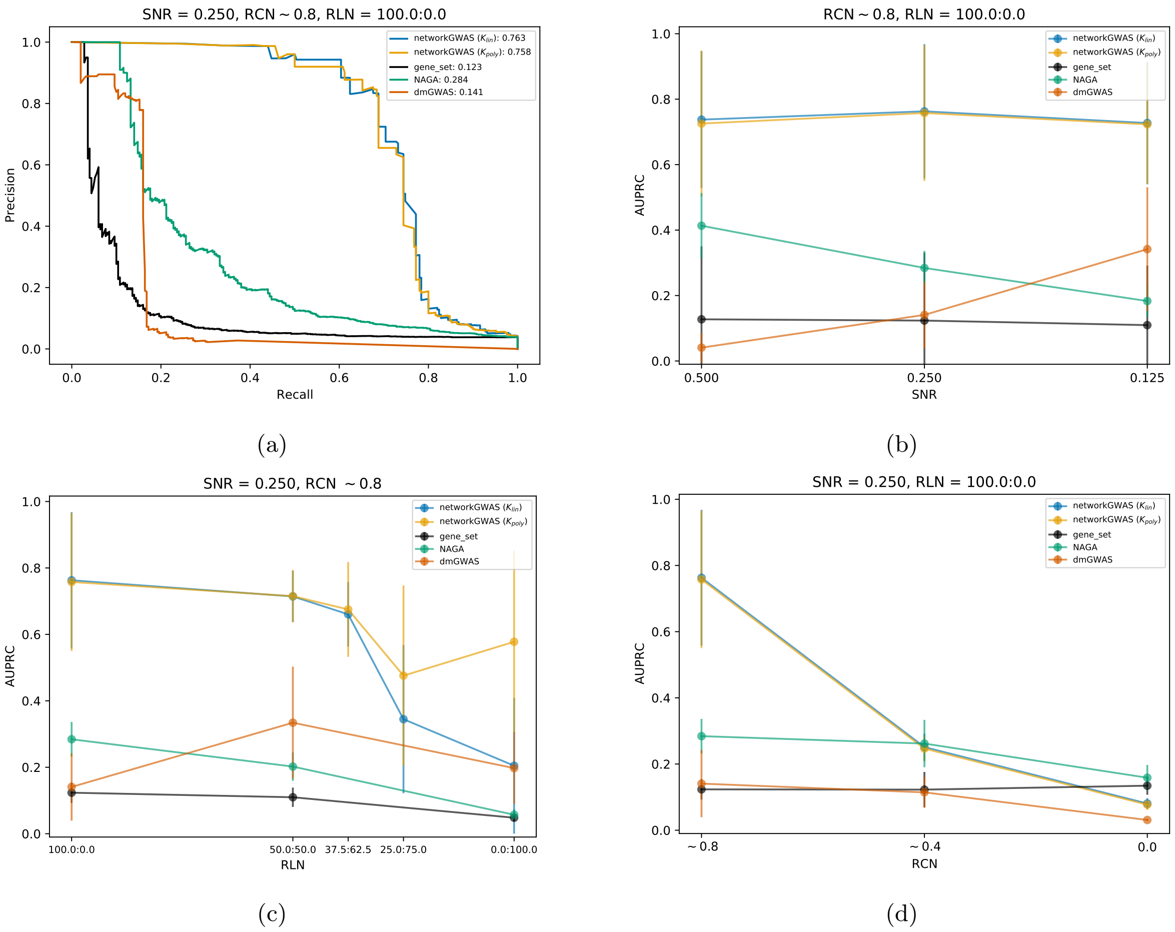
In the following section, we present the results we obtained on our simulated and natural phenotypes, namely on the simulated phenotypes for *A. thaliana*, the *A. thaliana* natural (i.e. true) phenotypes, and on the simulated rare variants in *H. sapiens*.

4.1 Simulated phenotypes for *A. thaliana*

Within our simulation study, we evaluate the performance of linear and non-linear `networkGWAS` as well as the comparison partners: gene-based FaST-LMM-Set [21], NAGA [5], and dmGWAS [14], 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 (4)), and the mean refers to the average with respect to the different random realizations of the various simulation settings. An overview of the results is compiled in Figure 1. 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 Table 1). In this scenario of a purely linear signal, spread across a single or very few causal subgraph(s) with a high ratio of causal neighbors, and high signal-to-noise ratio, both linear and non-linear `networkGWAS` outperform all the comparison partners by achieving an AUPRC of 76.3%±20.5% and 75.8%±20.8%, respectively. As demonstrated in the top right panel of Figure (1), this dominance in performance is invariant as one departs from the conditions of S0 by varying the SNR while keeping the ratio of causal neighbors, and the purely linear nature of the signal. 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 and RCN 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, that approaches designed to detect statistical significance of single loci will miss those with modest marginal effects and large interactions [22]. Lastly, the strong performance of `networkGWAS` and its dominance over the comparison partners breaks down as we tune the ratio of causal neighbors from ~ 0.8 to 0.0 while keeping the SNR and RLN the same as in scenario S0. This is shown in the bottom right panel of Figure (1), 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. In the latter case, the signal–by construction–is independent of the PPI network structure, and hence that structure cannot be exploited by any network-guided method to enhance performance.

4.2 *A. thaliana* phenotypes

We now turn to the results of our method on the natural, i.e. true, phenotypes associated with the *A. thaliana* model organism. We investigated whether `networkGWAS` was able to find any neighborhoods of genes that were statistically significantly associated with the phenotypes, and compared the results to the findings of



(a)

(b)

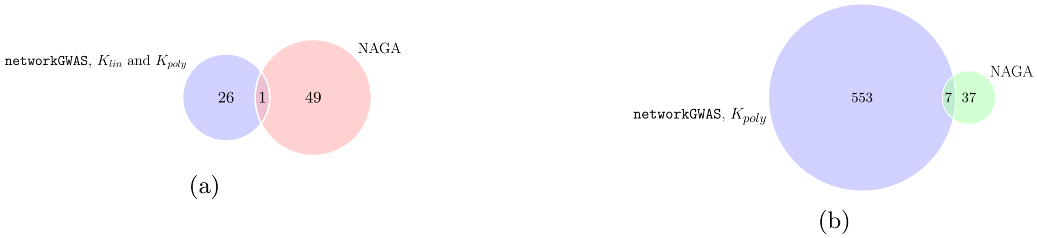
(c)

(d)

Fig. 1: 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 Subfigure 1a, we show the AUPRC of the baseline scenario (S0). In Subfigures 1b-1d, we vary one variable while keeping the other two fixed: SNR (1b), RLN (1c), and RCN (1d).

NAGA, FaST-LMM-Set, a simple GWAS of the individual SNPs that were mapped onto the PPI network, and a GWAS of the SNPs that were not mapped on the network (and thus not used in networkGWAS). For most of the phenotypes, there were no statistically significant genes found by any method. However, for two phenotypes, 523, and 705, networkGWAS alone returned statistically significant results. Since NAGA does not provide p-values, but rather scores, we defined a strategy to decide which genes to consider as associated with the studied phenotypes. By analysing the trend of the scores, we observe a strong change in the slope in the range of high scores: the analysis of the first derivative allows us to find the cutoff scores, 2.498 for phenotype 523 and 2.747 for phenotype 705, respectively. We used this threshold to identify genes that were associated with the phenotypes of interest, and then compared the findings to the genes that were surfaced by networkGWAS. A venn diagram showing the overlap can be found in Figure 2. Interestingly, networkGWAS returns genes that were almost entirely additional to the genes surfaced by NAGA, showcasing the added value that a neighborhood-based approach can bring.

Our method also allows one to investigate the biological interpretation of the genes found by networkGWAS. We use phenotype 523 for illustration, since 27 genes were found and are easy to investigate. We use the gene ontology (GO) Term Enrichment for Plants [35] provided by TAIR, which relies on the PANTHER Classification System [23, 36]. We use the PANTHER Overrepresentation Test (PANTHER version 16.0), with PANTHER GO-slim Biological Process as the annotation dataset and Fisher's exact test with FDR correction as the test type. We then perform the overrepresentation test for the significantly associated PPI



(a)

(b)

Fig. 2: Venn diagrams showing the overlap of the genes that were found to be associated with a phenotype of interest on phenotype 523 (Figure 2a) and 705 (Figure 2b) using `networkGWAS` and NAGA.

neighborhood of gene *AT4G18130*, yielding significant enrichment in processes related to (i) the cicardian rhythm, which prepares the plants for upcoming challenges by anticipating environmental factors [33], (ii) the response to light stimuli, and (iii) the transition from the vegetative to reproductive phase. As phenotype 523 represents the lifetime fitness under high precipitation and high density treatment, i.e., under great environmental strain, these findings of processes related to the plant's resilience and reproduction are biologically plausible.

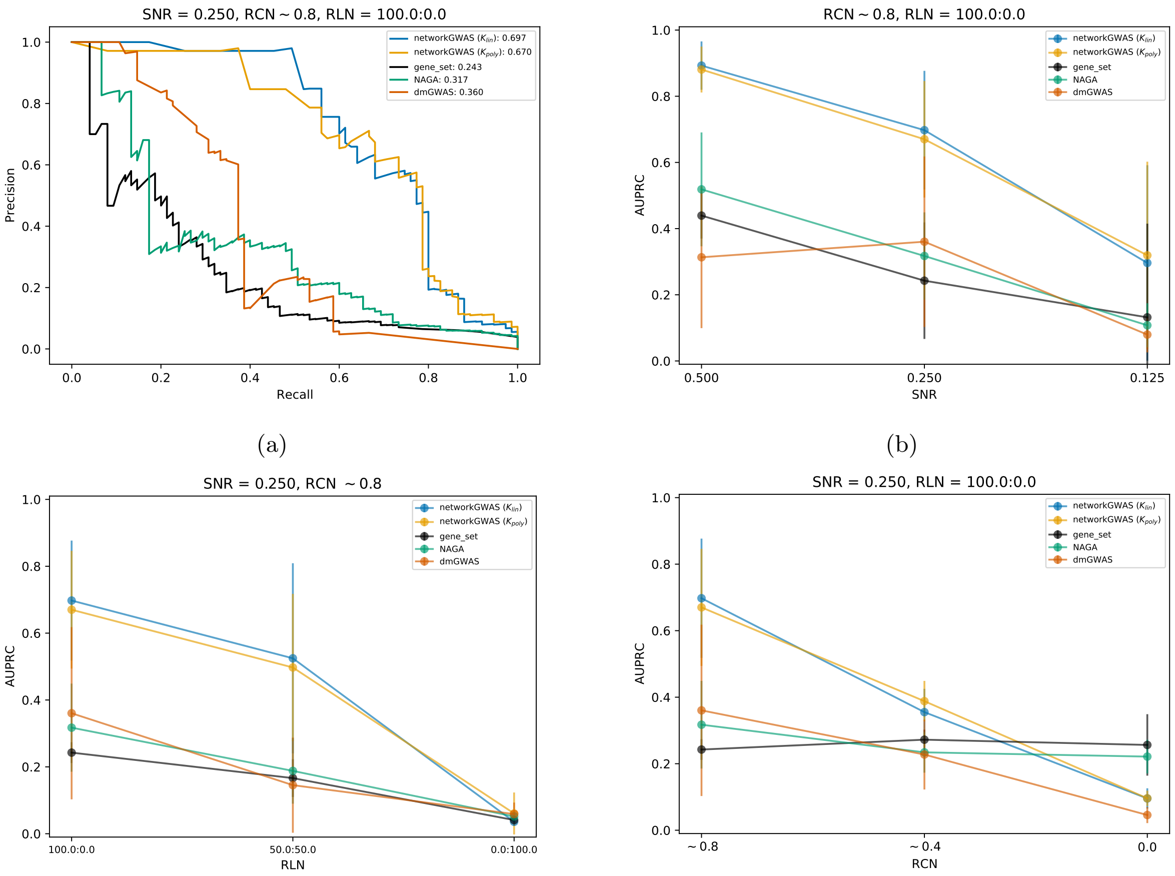
4.3 Simulated rare variants in *H. sapiens*

The results obtained for the rare-variant (RV) simulations are summarized in Figure 3 and are very similar in nature to the results obtained on common variants. As can be seen in the top left panel, in our scenario S0 of a purely linear signal, spread across a single or very few causal subgraph(s) with a high ratio of causal neighbors, and high signal-to-noise ratio, both linear and non-linear `networkGWAS` outperform all the comparison partners by achieving an AUPRC of 69.7%±17.9% and 67.0%±17.6%, respectively. As demonstrated in the top right and bottom left panel of Figure 3, this dominance is maintained over varying SNR and linear-to-nonlinear signal mixtures, and only vanishes for a fully nonlinear phenotype. Note, however, that in the RV case, `networkGWAS` is less robust to both increasing SNR and increasing nonlinearity of the signal than for the common variants. Furthermore, for RVs, the non-linear SNP-set kernel does not perform better than the linear one in any of the scenarios studied. As in the common-variant simulations, in the rare-variant case, too, `networkGWAS` cannot outperform its comparison partners in the case of isolated causal-genes, while `networkGWAS` does outperform comparison partners as soon as the PPI network holds phenotype-relevant information. This is shown in the bottom right panel of Figure 3.

5 Discussion

We defined a principled way to perform gene based genome wide association studies utilizing network information. 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 [21] and approaches that incorporate PPIs [5, 14]. 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, and various mixtures of linear and non-linear signal. Only if our underlying conjecture, that the SNPs in neighborhoods of interacting genes are collectively related to the phenotype of interest, is strongly violated, is `networkGWAS` outperformed. When studying natural phenotypes, `networkGWAS` finds collectively significantly associated and biologically plausible genes that were almost entirely undiscovered by its strongest competitor, NAGA, demonstrating the complementarity of the methods. In the following, we discuss the limitations of our method and potential means to mitigate them.

Limitations By construction, `networkGWAS` can only discover the collective signal of SNPs which are included in a biological network. Fortunately, as the knowledge of biological pathways and gene-gene interaction increases, fewer and fewer genes and SNPs will be excluded from such networks, which will negate this current limitation. Furthermore, this drawback can be addressed by applying traditional methods such as univariate GWAS or gene-based SNP-set tests to such SNPs, benefiting from increased test power due to a reduced search space. Another consideration is that by construction, `networkGWAS` provides p-values



(a)

(b)

(c)

(d)

Fig. 3: Results from our simulated *H. sapiens* use case. 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 Subfigure 3a, we show the AUPRC of the baseline scenario (S0). In Subfigures 3b-3d, we vary one variable while keeping the other two fixed: SNR (3b), RLN (3c), and RCN (3d).

for the association of entire gene neighborhoods, and cannot single-out precise genes or even SNPs within such neighborhoods as more or less strongly contributing to that association signal. If one were interested in SNP-level p-values, this could be addressed by means of applying traditional methods to the SNPs and genes comprised in the associated neighborhood. Another potential limitation lies within the choice of testing one-hop neighborhoods, which 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. We note that even if $k \geq 2$, our method still scales linearly with the number of nodes in the network.

Future Work An 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 nonlinearity, 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.

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).

References

1. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E., Kang, H.M., Korbel, J.O., Marchini, J., McCarthy, S., McVean, G., Abecasis, G.: A global reference for human genetic variation. Nature **526**(7571), 68–74 (2015)
2. Azencott, C.A., Grimm, D., Sugiyama, M., Kawahara, Y., Borgwardt, K.M.: Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics **29**(13), i171–i179 (06 2013). https://doi.org/10.1093/bioinformatics/btt238, https://doi.org/10.1093/bioinformatics/btt238
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
4. Borgwardt, K., Ghisu, E., Llinares-Lopez, F., O'Bray, L., Rieck, B.: Graph kernels: State-of-the-art and future challenges. Foundations and Trends® in Machine Learning **13**, 531–712 (2020). https://doi.org/10.1561/2200000076, http://dx.doi.org/10.1561/2200000076
5. Carlin, D.E., Fong, S.H., Qin, Y., Jia, T., Huang, J.K., Bao, B., Zhang, C., Ideker, T.: A fast and flexible framework for network-assisted genomic association. Iscience **16**, 155–161 (2019)
6. Consortium, T..G.: 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell **166(2)** (2016). https://doi.org/10.1016/j.cell.2016.05.063
7. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S., Davies, R., Li, H.: Twelve years of samtools and bcftools. Gigascience **10**(2), giab008 (2021)
8. Dimitromanolakis, A., Xu, J., Krol, A., Briollais, L.: sim1000g: a user-friendly genetic variant simulator in r for unrelated individuals and family-based designs. BMC Bioinformatics **20**(26) (2019)
9. Exposito-Alonso, M., 500 Genomes Field Experiment Team, Burbano, H.A., Bossdorf, O., Nielsen, R., Weigel, D.: Natural selection on the arabidopsis thaliana genome in present and future climates. Nature **None** (2019). https://doi.org/10.1038/s41586-019-1520-9
10. Goddard, M.E., Wray, N.R., Verbyla, K., Visscher, P.M., et al.: Estimating effects and making predictions from genome-wide marker data. Statistical science **24**(4), 517–529 (2009)
11. Greven, S., Crainiceanu, C.M., Küchenhoff, H., Peters, A.: Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. Journal of Computational and Graphical Statistics **17**(4), 870–891 (Dec 2008). https://doi.org/10.1198/106186008X386599, https://doi.org/10.1198/106186008X386599, publisher: Taylor & Francis _eprint: https://doi.org/10.1198/106186008X386599
12. Hayes, B.J., Visscher, P.M., Goddard, M.E.: Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research **91**(1), 47–60 (2009)
13. Holden, M., Deng, S., Wojnowski, L., Kulle, B.: GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics **24**(23), 2784–2785 (Dec 2008). https://doi.org/10.1093/bioinformatics/btn516, https://doi.org/10.1093/bioinformatics/btn516
14. Jia, P., Zheng, S., Long, J., Zheng, W., Zhao, Z.: dmgwas: dense module searching for genome-wide association studies in protein–protein interaction networks. Bioinformatics **27**(1), 95–102 (2011)
15. Junker, B.H., Schreiber, F.: Analysis of biological networks, vol. 2. John Wiley & Sons (2011)
16. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
17. Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., et al.: The arabidopsis information resource (tair): improved gene annotation and new tools. Nucleic acids research **40**(D1), D1202–D1210 (2012)
18. Li, B., Leal, S.M.: Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. The American Journal of Human Genetics **83**(3), 311–321 (Sep 2008). https://doi.org/10.1016/j.ajhg.2008.06.024, http://www.sciencedirect.com/science/article/pii/S0002929708004084
19. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D.: FaST linear mixed models for genome-wide association studies. Nat Methods **8**(10), 833–835 (Oct 2011). https://doi.org/10.1038/nmeth.1681, https://www.nature.com/articles/nmeth.1681, number: 10 Publisher: Nature Publishing Group
20. Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., Heckerman, D.: Improved linear mixed models for genome-wide association studies. Nature methods **9**(6), 525–526 (2012)
21. Listgarten, J., Lippert, C., Kang, E.Y., Xiang, J., Kadie, C.M., Heckerman, D.: A powerful and efficient set test for genetic markers that handles confounders. Bioinformatics **29**(12), 1526–1533 (Jun 2013). https://doi.org/10.1093/bioinformatics/btt177, https://academic.oup.com/bioinformatics/article/29/12/1526/291743, publisher: Oxford Academic

22. Marchini, J., Donnelly, P., Cardon, L.R.: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics **37**(4), 413–417 (2005)
23. Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., Thomas, P.D.: Panther version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium. Nucleic Acids Research **38**, D204–D210 (2009)
24. Pinheiro, J.C., Bates, D.M.: Linear Mixed-Effects Models: Basic Concepts and Examples. In: Mixed-Effects Models in S and S-PLUS. Statistics and Computing, Springer, New York, NY (2000). https://doi.org/10.1007/0-387-22747-4_1
25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M., Sham, P.C.: Plink: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics **81** (2007)
26. Rivas, J.D.L., Fontanillo, C.: Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. PLOS Computational Biology **6**(6), e1000807 (Jun 2010). https://doi.org/10.1371/journal.pcbi.1000807, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000807, publisher: Public Library of Science
27. Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E., et al.: Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature **476**(7359), 214 (2011)
28. Schwender, H., Ruczinski, I., Ickstadt, K.: Testing SNPs and sets of SNPs for importance in association studies. Biostatistics **12**(1), 18–32 (Jan 2011). https://doi.org/10.1093/biostatistics/kxq042, https://doi.org/10.1093/biostatistics/kxq042
29. Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., Korte, A.: Arapheno: a public database for arabidopsis thaliana phenotypes. Nucleic Acids Research p. gkw986 (2016)
30. Shaun Purcell: PLINK 2.0. http://pngu.mgh.harvard.edu/purcell/plink/ (2007)
31. Shervashidze, N., Schweitzer, P., van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler–lehman graph kernels. Journal of Machine Learning Research (12), 2539–2561 (2011)
32. Shervashidze, N., Borgwardt, K.: Fast subtree kernels on graphs. In: NeurIPS. pp. 1660–1668 (2009)
33. Srivastava, D., Shamim, M., Kumar, M., Mishra, A., Maurya, R., Sharma, D., Pandey, P., Singh, K.: Role of circadian rhythm in plant system: An update from development to stress response. Environmental and Experimental Botany **162**, 256–271 (2019)
34. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al.: The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research **49**(D1), D605–D612 (2021)
35. TAIR: GO Term Enrichment for Plants. https://www.arabidopsis.org/tools/go_term_enrichment.jsp (2021)
36. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A.: Panther: A library of protein families and subfamilies indexed by function. Genome Research **13**, 2129–2141 (2003)
37. Togninalli, M., Seren, Ü., Meng, D., Fitz, J., Nordborg, M., Weigel, D., Borgwardt, K., Korte, A., Grimm, D.G.: The aragwas catalog: a curated and standardized arabidopsis thaliana gwas catalog. Nucleic acids research **46**(D1), D1150–D1156 (2018)
38. Wilks, S.S.: The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals of Mathematical Statistics **9**(1), 60–62 (Mar 1938). https://doi.org/10.1214/aoms/1177732360, https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-9/issue-1/The-Large-Sample-Distribution-of-the-Likelihood-Ratio-for-Testing/10.1214/aoms/1177732360.full, publisher: Institute of Mathematical Statistics
39. Wu, C., MacLeod, I., Su, A.I.: BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic Acids Research **41**(D1), D561–D565 (2013)
40. Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X.: Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. The American Journal of Human Genetics **89**(1), 82–93 (Jul 2011). https://doi.org/10.1016/j.ajhg.2011.05.029, http://www.sciencedirect.com/science/article/pii/S0002929711002229
41. Xin, J., Mark, A., Afrasiabi, C., Tsueng, G., Juchler, M., Gopal, N., Stupp, G.S., Putman, T.E., Ainscough, B.J., Griffith, O.L., Torkamani, A., Whetzel, P.L., Mungall, C.J., Mooney, S.D., Su, A.I., Wu, C.: High-performance web services for querying gene and variant annotation. Genome Biology **17**(1), 1–7 (2016)
42. Zhang, X., Huang, S., Zou, F., Wang, W.: Team: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics **26**(12), i217–i227 (2010)
43. Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S.: The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences **109**(4), 1193–1198 (2012)