Combining network-guided GWAS to discover susceptibility mechanisms to breast cancer

Héctor Climente-González, Christine Lonjou, Chloé-Agathe Azencott

September 3, 2019

Abstract

TODO This is the abstract.

Introduction

In human health, genome-wide association studies (GWAS) aim at quantifying how genetic variants predispose to complex diseases like diabetes or some forms of cancer [1]. The most commonly studied variants are single-nucleotide polymorphisms (SNPs). Once the relevant SNPs have been discovered, they can be used for diagnosis, computing the risk, and deepen our understanding of the disease. To that end, in a typical retrospective study thousands of unrelated samples are genotyped: the cases, suffering the disease of interest, and the controls from the general population. Then, a per-SNP statistical association test is conducted (e.g. χ^2). Those SNPs with a P-value lower than a conservative FWER threshold are candidates to further studies in an independent cohort.

GWAS has been successful at identifying the main variants underlying many common diseases. However, the experimental setting also presents intrinsic challenges. Some of them stem from the high-dimensionality of the problem, as every GWAS to date study more variants than samples are genotyped. This limits the statistical power of the experiment, as only variants with large and moderate effects can be detected. However, most of the heritability can be explained by variants with small effects. Additionally, to avoid false positives, a conservative multiple test correction is applied, typically Bonferroni. However, Bonferroni is known to be overly conservative when the statistical tests are correlated. Another open issue is the interpretability of the results, as the functional consequences of most common

variants are not well understood. The fact that SNPs in non-coding regions often top the results of GWAS portrays well this challenge. On top of that, recent large-sampled studies suggest that most of the genome contributes to a degree to any complex trait. This opposes to the traditional polygenic model of complex diseases, which postulated that only key genes in key pathways need to be affected. In the recently postulated omnigenic model [2], genes are highly interconnected, and influence each others behavior. Hence, a more comprehensive statistical framework is required.

In this context, many authors turn to network biology to handle the complex interplay of biomolecules that lead to disease [3]. As its name suggests, network biology models biology as a network, where the biomolecules under study, often genes, are nodes, and selected functional relationships between them are the edges that link them. These functional relationships come from evidence that the genes jointly contribute to a biological function; for instance, their expression is correlated, or a protein-protein interaction between their gene products have been discovered. Under this view, complex diseases are not the consequence of a single altered gene, but of the interaction of multiple interdependent biomolecules [4]. In fact, an examination of biological networks shows that disease genes have differential properties [4] [5]. This was particularly true for cancer driver genes, which tended to be key players in connecting different densely-connected communities of genes.

Crucially, the genes that contribute to a disease tend to participate in similar biological functions. Thus, guilt-by-association strategies have proved effective at identifying disease genes [6]. Network-based, biomarker discovery methods exploit this differential characteristic to identify disease genes on GWAS data. In essence, each SNP has a measure of association with the disease, given by the experiment, and a biological context, given by the network. Then, the problem becomes finding a functionally-related set of genes that is highly associated with the disease. Different solutions have been proposed to this problem, often stemming from divergent considerations of what the optimal solution set look like. Some methods strongly constrain the problem to certain kinds of subnetworks. Such is the extreme case of LEAN [7], which focuses on 'star subnetworks' i.e. instances were a gene and its direct interactors are associated with the disease. Other algorithms, like dmGWAS [8] and heinz [9], focus on interconnected genes with high association with the disease. However, they differ in their tolerance to include lowly associated nodes. Lastly, other methods explicitly consider the topology of the solution, favoring densely interconnected subnetworks; such is the case of HotNet2 [10], SConES [11], and SigMod [12].

TODO Check classification in Network-guided biomarker discovery.

In this work, we analyze the effectiveness of these six methods to discover new biomarkers on GWAS data. We focus on the GENESIS dataset [13], a homogeneous dataset of familial breast cancer in French population of European ancestry. After following a classical GWAS approach, we use these network-based methods to recover additional familial breast cancer biomarkers. Some of them are known, while others are specific to this dataset. Lastly, we carry out a comparison of the solutions obtained by the different methods, and aggregate them to obtain a consensus network of predisposition to familial breast cancer.

Methods

GENESIS

GENESIS is a GWAS study of familial breast cancer on the French population [13]. The index cases are patients with infiltrating mammary or ductal adenocarcinoma, who had sister with breast cancer, and negative for BRCA1 and BRCA2 mutations. The controls are unaffected colleagues and/or friends of the cases, born around the year of birth of the corresponding case ($\pm$ 3 years). In total, the GENESIS dataset consists of 2577 samples, of which 1279 are controls and 1298 are cases. The genotyping platform was iCOGS, a custom Illumina array designed to study genetic susceptibility of hormone-related cancers [14]. It contains 210918 SNPs previously associated with cancer susceptibility, survival, or other cancer-related traits; and in selected genes and pathways.

Preprocessing and quality control

We discarded SNPs with a minor allele frequency lower than 0.1%, not in Hardy - Weinberg equilibrium (P-value $<$0.001), and/or those with missing values on more than 10% of the samples. In addition, we removed the samples with more than 10% missing genotypes. 28 samples with TODO were removed. A subset of 20 duplicated SNPs in FGFR2 were also removed. The final dataset included 1271 controls and 1280 cases, genotyped over 197083 SNPs.

We looked for population structure that could create confounding associations. A PCA revealed no differential population structure between cases and controls (Supplementary Figure 1). Independently, we did not find evidence of genomic inflation ($\lambda = 1.05029$) either, thus further dismissing the presence of confounding population structure.

High-weight subnetwork discovery algorithms

SNP and gene association

To measure association between a genotype and the phenotype, we performed a per-SNP 1df χ^2 allelic test using PLINK v1.90 [15]. Then, we used VEGAS2v2 to compute the gene-level association score [16] from the SNP P-values. In order to map SNPs to genes we relied used their overlap on the sequence: all SNPs located within the boundaries of a gene, $\pm$ 50 kb, were mapped to that gene. To compute the gene association we used the 10% of SNPs with lowest P-values. We computed the association 62193 genes described in GENCODE 31 [17]; only 54612 had a SNP mapped to them. Then, we focused exclusively on the 32767 that could be mapped to an HGNC symbol. Out of the SNPs 197083 in iCOGS after quality control, 164037 mapped to at least one of these genes.

Gene-gene network

Out of the six methods tested, five use a gene-gene interaction network (Section Methods used), and their respective statistical frameworks are compatible with any type of network (protein interactions, gene coexpression, regulatory, etc.). However, in order to make the results comparable, we needed to apply all the methods to the same network. Hence, for practical reasons, we focused on a protein-protein interaction network (PPIN), as most of the methods were designed to scale appropriately to them. We built our PPIN from both binary and co-complex interactions stored in the HINT database (release April 2019) [18]. Unless specified otherwise, we used only interactions coming from high-throughput experiments to avoid biasing the topology of the network by well-studied genes with more known interactions on average. Out of the 146722 interactions from high-throughput experiments that HINT stores, we were able to map 142541 to a pair of HGNC symbols, which we used as node identifier.

Additionally, we compared the results of the aforementioned network with those obtained on a network built using interactions coming from both high-throughput and targeted studies. In that case, out of the 179332 interactions in HINT, we mapped 173797 to a pair of HGNC symbols.

The scoring function for the nodes changed from method to method (Section Methods used).

SNP networks

SConES [11] is the only of the studied methods designed to handle SNP networks. As in gene networks, two SNPs are linked in a SNP network when there is evidence of shared functionality between two SNPs. The authors suggested three ways of building these networks: connecting the genotyped SNPs consecutive in the genomic sequence ("GS network"); interconnecting all the SNPs mapped to the same gene, on top of GS ("GM network"); and interconnecting all SNPs mapped to two genes for which a protein-protein interaction exists ("GI network"). We used all three. For the GM network, we used the mapping described in Section SNP and gene association. For the GI network, we used the PPI as described in Section Gene-gene network.

For all three networks the node score used is the association of the individual SNPs with the phenotype; specifically, we used the 1 d.f. χ^2.

Mathematical notation

In this article, we refer to undirected, vertex-weighted networks, or graphs, G = (V,E,w). V = $\{v_1, \ldots, v_n\}$ refers to the vertices, with weights w: V $\rightarrow$ R. Equivalently, E $\subseteq \{\{x, y\} | x, y \in V \wedge x \neq y\}$ refers to the edges. When referring to a subnetwork S, V_S is the set of nodes in S and E_S is the set of edges in S. A special case of subgraphs are *connected* subgraphs, which occur when every node in the subgraph can be reached from any other node.

In addition, we use several matrices that describe different properties of a graph. The described matrices are square, and have as many rows and columns as nodes are in the network. In fact, the element i,j represent a selected relationship between v_i and v_j. The adjacency matrix W_G contains a 1 when the corresponding nodes are connected through an edge, and 0 otherwise; the diagonal is zero. The degree matrix D_G is a diagonal matrix which contains the degree of the different nodes. Lastly, the Laplacian matrix L_G is defined as $L_G = D_G - W_G$.

TODO Methods used

TODO explain why it's an open problem i.e. which score should be used (SNP, gene?), what the solution looks like, the problem is NP-hard. TODO specify how nodes are scored.

Finding the highest-scoring, densely interconnected subnetwork on a graph is an open problem in the field. Hence, several solutions have been proposed to the problem. In this paper, we apply six methods designed to explore the protein-protein interaction network, and one method, SConES,

Table 1: Summary of the differences between the studied algorithms. The colums are: Field, the field in which the algorithm was developed; Node type, the type of network, either gene (protein-protein interaction network usually) or a SNP network; Exhaustive, if all the possible solutions given the selected hyperparameters are explored, or not; Solution, referring properties that are enforced on the solution, other than being dense in high scores and connected; and Input, referring to whether the methods require genotype data or GWAS summary statistics.

Algorithm	Field	Node type	Exhaustive	Solution	Input
heinz	Omics	Gene	Yes?	-	Summary
HotNet2	Omics	Gene	Yes?	Modular	Summary
dmGWAS	GWAS	Gene	No	-	Summary
LEAN	Omics	Gene	Yes	Star-shaped	Summary
SConES	GWAS	SNP	Yes	Modular	Genotypes
SigMod	GWAS	Gene	Yes	Modular	Summary

which explores SNP-networks. We selected methods that had a readily available, programmatically accessible implementation. Their main differences are summarized in Table 1.

TODO Re-read heinz paper. It's the solution heuristic? If so, how good is it? Efficient enough to be used in SNP network? TODO Reformulate heinz to show similarities to SConES.

heinz The goal of heinz is identifying the highest-scored connected subgraph on the network [9]. The problem has a trivial solution when all scores are positive: the whole network; however, it becomes NP-complete when scores are both positive and negative. The authors propose a transformation of the nodes' P-value into a score which takes a negative value when no association with the phenotype is detected, and a positive value when it is. The distinction between both is determined though an FDR approach. Then, the problem is re-casted as the Prize-Collecting Steiner Tree Problem (PCST). This is the problem of selecting the connected subnetwork S that maximizes the *profit* p(S):

$$p(S) = \sum_{v \in V_S} p(v) - \sum_{e \in E_S} c(e).$$

were p(v) is called profit of adding a node, and c(e) is the cost of the

edge, both positive values. These quantities are defined from w' = $\min_{v \in V_G}$ w(v):

p(v) = w(v) - w',
c(e) = w'.

PCST has a heuristic, efficient solution [19]. We used the implementation of heinz from BioNet [20], available on Bioconductor [21].

HotNet2 HotNet2 was developed in the context of tumor driver identification, as a tool to find connected subgraphs of genes mutated more often than expected by chance [10]. To that end, it considers both the local topology of the network and the scores of the nodes. The former is captured by an insulated heat diffusion process, modeled by a random walk with restart. At the beginning, the score of the node determines its initial heat. In an iterative procedure, each node gives heat to its "colder" neighbors, and receives heat from its "hotter" neighbors, while retaining part of its heat (hence, *insulated*). This process continues until equilibrium is reached, and results in a similarity matrix F. This matrix is used to compute the similarity matrix E that accounts also for similarities in node scores as

$$E = F\,\mathrm{diag}(w(V)),$$

where diag($w(V)$) is a diagonal matrix with the node scores in its diagonal. HotNet2 explores the similarity network built from E to find densely connected subnetworks. Specifically, it only connects a pair of nodes i and j when E(i,j) $> \delta$. Lastly, HotNet2 evaluates the statistical significance of the subnetworks by comparing their size to the size of networks obtained by permuting the node scores.

HotNet2 has two parameters: the restart probability β, and the threshold heat δ. Both parameters are set automatically by the algorithm, and are robust [10].

HotNet2 is implemented in Python [22].

TODO Read Gene and Network Analysis of Common Variants Reveals Novel Associations in Multiple Complex Diseases

dmGWAS dmGWAS aims at identifying the connected subgraph with the largest amount of low P-values [8]. To that end, it first searches several candidate subnetwork solutions using a greedy procedure involving the following steps:

1. Select a seed node.
2. Compute Stouffer's Z-score Z_m for the current subgraph as

$$Z_m = \frac{\sum z_i}{\sqrt{k}}$$

   where k is the number of genes in the subgraph, $z_i = \phi^{-1}(1 - P_i)$, and ϕ^{-1} is the inverse normal distribution function.
3. Identify neighboring nodes i.e. nodes at shortest path $\leq d$. We set d = 2.
4. Add the neighboring nodes whose inclusion increases the Z_{m+1} more than $Z_m \times (1 + r)$. In our experiments, we set r = 0.1.
5. Repeat 2-4 until no increment $Z_m \times (1 + r)$ is possible.

Lastly, the module's Z-score is normalized as

$$Z_N = \frac{Z_m - \text{mean}\left(Z_m(\pi)\right)}{\text{SD}\left(Z_m(\pi)\right)}$$

where $Z_m(\pi)$ represent a vector with 100000 random subsets of the same number of genes.

We used the implementation of dmGWAS in the dmGWAS 3.0 R package [23]. We used the function *simpleChoose* to select the solution subnetwork, which aggregates the top 1% modules into the solution subnetwork.

LEAN Local enrichment analysis (LEAN) searches disregulated "star" gene subnetworks i.e. subnetworks composed by one central node and all its interactors [7]. By imposing this restriction, LEAN is able to exhaustively test all possible solution subnetworks (one per node in the network). For a particular subnetwork of size m, the P-values corresponding to the involved nodes are ranked as $p_1 \leq \ldots \leq p_m$. Then, k binomial tests are conducted, to compute the probability of having k out of m P-values lower or equal to p_k under the null hypothesis. The minimum of these k P-values is the score of the subnetwork. This score is transformed into a P-value through an empirical distribution obtained via a subsampling scheme, where sets of m genes are selected randomly, and their score computed. Lastly, P-values are corrected for multiple testing through a Benjamini-Hochberg correction. We used the implementation of LEAN from the LEANR R package [24].

SConES SConES searches the minimal, maximally interconnected, maximally associated subnetwork in a SNP graph [11]. Specifically, it solves the problem

$$\arg\max_{S \subseteq G} \underbrace{\sum_{v \in V_S} w(v)}_{\text{association}} - \underbrace{\lambda \sum_{v \in V_S} \sum_{u \notin V_S} W_{vu}}_{\text{connectivity}} - \underbrace{\eta |V_S|}_{\text{sparsity}} \tag{1}$$

where λ and η are hyperparameters that control the sparsity and the connectivity of the model. For two hyperparameters, the aforementioned problem has a unique solution, that SConES finds using a graph min-cut procedure. We used the version on SConES implemented in R package martini [25]. We selected λ and η by cross-validation, choosing the values that produce the most stable solution across folds. Note that the solution to the above problem can consist of several connected subnetworks which are disconnected from each other. In this case, the selected hyperparameters were $\eta = 3.51$, $\lambda = 210.29$ for SConES GS; $\eta = 3.51$, $\lambda = 97.61$ for SConES GM; and $\eta = 3.51$, $\lambda = 45.31$ for SConES GI.

TODO Comment similarity with heinz

SigMod SigMod aims at identifying the most densely connected gene subnetwork that is most strongly associated to the phenotype [12]. It addresses an optimization problem similar to that of SConES (Equation 1), but replacing the Laplacian matrix my the adjacency matrix (Section Mathematical notation).

$$\arg\max_{S \in G} \underbrace{\sum_{v \in V_S} w(v)}_{\text{association}} + \underbrace{\lambda \sum_{v \in V_S} \sum_{u \in V_S} W_{vu}}_{\text{connectivity}} - \underbrace{\eta |V_S|}_{\text{sparsity}} .$$

As SConES, this optimization problem can also be solved by a graph min-cut approach.

SigMod presents two important additional differences with SConES. First it is designed for gene-gene networks. Second, it returns a single connected subnetwork, which it achieves by exploring a grid of hyperparameters and processing their respective solutions. Specifically, for the range of $\lambda = \lambda_{\text{min}}, \ldots, \lambda_{\text{max}}$ for the same η, it prioritizes

the solution with the largest change in size from λ_n to λ_{n+1}. Such a large change implies that the network is strongly interconnected. This results in one candidate solution for each η, which are processed by removing any node not connected to any other. A score is assigned to each candidate solution by summing their node scores and normalizing by size. The candidate solution with the highest standardized score is the chosen solution. SigMod is implemented in an R package [26].

Mapping back and forth between gene methods and SConES

In this work dealt with multiple methods, which use GWAS data at different levels. VEGAS2 compute gene statistics from SNP statistics, which are then used by five gene-based network methods to find a subnetwork associated with familial breast cancer. In order to obtain a list of SNP biomarkers from these gene subnetworks, we consider all the genes that can be mapped to that gene as selected by the method. SConES is in the opposite case: it performs selection on a network of SNPs. In this case, when analyzing the genes selected by SConES, we consider any gene that can be mapped to any of the selected SNPs as selected as well.

Consensus network

The different high-weight subnetwork discovery algorithms make different assumptions on the nature of the solutions, and employ different strategies to find them. Hence, combining the outcome of the different approaches might provide a more complete outlook on the specific alterations on the GENESIS dataset. We built such consensus network by retaining the nodes that were selected by at least two of the methods. We combined the results of 6 methods: heinz, Hierarchical HotNet, dmGWAS, LEAN, SConES on the GM network, and SigMod.

Validation of selected biomarkers

Classification accuracy of selected biomarkers

To evaluate the quality of the solutions offered by the different algorithms, we used their predictor power. We reasoned that a desirable solution is one that is sparse, while offering a good predictor power. To evaluate the predicting power of the SNPs selected by the different methods, we used the performance of an L1-penalized logistic regression trained exclusively on those SNPs to predict the outcome (case/control). The L1 penalty helps to

account for LD to reduce the size of the active set, while improving the generalization of the classifier. The value of the λ, which controls the size of the coefficients, was set by cross-validation. To that end, we used the different network-methods on a random 80% of the samples and trained our predictor exclusively on the SNPs selected by a particular method, on these samples. When the method retrieved a list of genes (all of them except SConES), all the SNPs mapped to any of those genes were used. Then we evaluated performance of the classifier on the remaining 20% of the dataset. We repeated this procedure 5 times to estimate the average and the deviation of the different performance measures. The different performance measures we used where: size of the solution, size of the active set, specificity, sensitivity and average Jaccard similarity between different runs. In addition, we repeated the procedure without applying a network-based feature selection method.

Another desirable property is that the method retrieves a good candidate causal subnetwork. In consequence, we compared the outcome of each of the methods to the consensus subnetwork of all the solutions (Section Consensus network).

TODO Machine learning & SNP paper.

Biological relevance of the genes

An alternative way to validate the results is comparing our results to an external dataset. For that purpose, we recovered a list of 153 genes known to be associated to familial breast cancer from DisGeNET [27]. Across this article, we refer to these genes as *familial breast cancer genes*.

Additionally, we used the summary statistics from the Breast Cancer Association Consortium (BCAC) [28]. BCAC is one of the largest efforts in GWAS, with over 120000 women from European ancestry, albeit from different countries. As opposed to GENESIS, samples were not selected based on family history, and hence is enriched in sporadic breast cancers. Another difference is that BCAC is a relatively heterogeneous study on a pan-European sample, while GENESIS is a homogeneous dataset focused on the French population. Despite these differences, there should be shared genetic architecture. On top of that, that overlap should become more notorious when the results are aggregated at the gene level. For that purpose, we computed the gene association as in Section SNP and gene association. iCOGS array was used for genotyping in BCAC [14], the same array as for GENESIS [13]. Although imputed data is available, we used exclusively the SNPs available on GENESIS after quality control to make the results comparable.

Code availability

This work here presented required developing computational pipelines for several GWAS analyses, such as physical mapping of SNPs, computing gene scores, and perform six different network-based analyses. For each of those processes, a streamlined, project-agnostic pipeline with a clear interface was created. They are compiled in the following GitHub repository: `https://github.com/hclimente/gwas-tools`. The code that applies these pipelines to the GENESIS project, as well as the code that reproduces all the analyses in this article are available at `https://github.com/hclimente/genewa`.

Results

FGFR2 is strongly associated with familial breast cancer

We conducted association analyses both at the SNP level and at the gene level in the GENESIS dataset (Section SNP and gene association). Two genomic regions have a P-value lower than the Bonferroni threshold in chromosomes 10 and 16 (Figure 1A). The former overlaps with gene FGFR2; the latter with CASC16, and its located near the protein-coding gene TOX3. Variants in both FGFR2 and TOX3 were related to breast cancer susceptibility in other cohorts negative for BRCA1/2 [29]. Only the peak in chromosome 10 replicated in the gene-level analysis, with FGFR2 just below threshold of significance (Figure 1B).

These results show the overlap between the genetic architecture of the disease between the French population and other cohorts, especially at the gene level. In addition, there are other regions highly associated with familial breast cancer, albeit well above the conventional threshold of significance. The most prominent regions, which have been associated to breast cancer susceptibility in the past, are 3p24 [30], and 8q24 [31]. This motivates exploring network methods, which trade statistical association for biological significance.

TODO Network methods successfully identify genes linked to breast cancer

We applied six network methods to the GENESIS dataset (Section Methods used). We obtained eight solutions (Supplementary files 1 and 2): one for each of the gene-based methods (Section Gene-gene network), and one for each of the SNP networks of SConES (Section SNP networks). The solutions

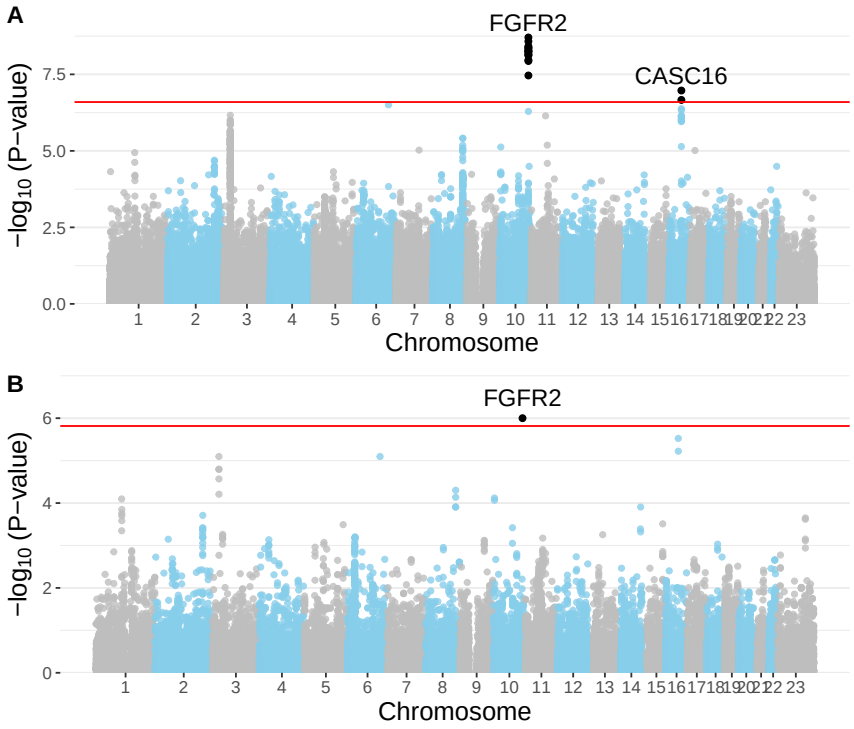

Figure 1: Association in GENESIS. The red line represents the Bonferroni threshold. **(A)** SNP association, measured from the outcome of a 1df χ^2 allelic test. SNPs that are within a coding gene, or within 50 kilobases of its boundaries are annotated. The Bonferroni threshold is 2.54×10^{-7}. **(B)** Gene association, measured by P-value of VEGAS2v2 [16] using the 10% of SNPs with the lowest P-values. The Bonferroni threshold is 1.53×10^{-6}.

Table 2: Summary statistics on the results of multiple network methods on the gene-gene interaction network. The first row contains the summary statistics on the whole network.

Network	Genes	Edges	Mean betweenness	Median P_{gene}	$Jaccard_{consensus}$
HT HINT	13619	142541	16706	0.46	0.004
Consensus	55	117	74062	0.0051	1
dmGWAS	194	450	49115	0.19	0.26
heinz	4	3	113633	0.0012	0.073
HotNet2					
LEAN	0	0	-	-	0
SConES GS	5	0	9805	2.7×10^{-5}	0.071
SConES GM	28	2	4267	0.067	0.078
SConES GI	0	0	-	-	0
SigMod	142	249	92603	0.0083	0.33

Table 3: Summary statistics on the results of SConES on the three SNP-SNP interaction networks. The first row within each block contains the summary statistics on the whole network.

Network	SNPs	Edges	Subnetworks	Mean betweenness	Median P_{SNP}
GS	197083	197060	-	2.03×10^7	0.49
SConES GS	1590	1585	5	2.52×10^7	0.023
GM	197083	6442446	-	3.99×10^6	0.49
SConES GM	1692	177611	5	4.40×10^6	0.055
GI	197083	28733720	-		0.49
SConES GI	408	539	5		0.076

were very heterogeneous (Tables 2 and 3): none of the subnetworks examined by LEAN was significant (adjusted P-value < 0.05), while dmGWAS produced the largest solution subnetwork with 194 genes. However, they succeed at recovering genes involved in the disease: five solution subnetworks are significantly enriched in familial breast cancer genes (dmGWAS, heinz, SConES GS, SConES GM, and SigMod, Fisher's exact test one-sided P-value < 0.03). We also compared the outcome of the network methods to the association tests conducted on the European cohort of the Breast Cancer Association Consortium (BCAC) [28] (Supplementary Figure 4). Encouragingly, every solution subnetwork was enriched in genes or SNPs that were Bonferroni-significant in BCAC. This shows that in practice network methods can find the same signal than more conservative analyses, by leveraging on the association of the biological context as a whole.

In fact, the solution subnetworks present other desirable properties. First, all solution subnetworks except LEAN's are, on average, more strongly associated to familial breast cancer than the whole HINT protein-protein interaction network. In our experiments, we observed that SConES GS strongly favor highly associated genes (median gene P-value $= 2.7 \times 10^{-5}$), while dmGWAS is less conservative (median gene P-value $= 0.19$). This exemplifies the differences between the methods: dmGWAS performs a greedy search that examines all neighbors at distance 2, and hence considering adding a weakly associated gene if it has a strongly associated neighbor. Also, the genes in five solution subnetworks display on average a higher betweenness centrality than the rest of the genes, a difference that is significant in two solutions (dmGWAS and SigMod, Wilcoxon rank-sum test one-sided P-value $< 6.9 \times 10^{-22}$). This agrees with the notion that disease genes are more central than other, non-essential genes [5].

Due to the differences between solutions, it is hard to draw joint conclusions. The 4-gene solution selected by heinz includes TOX3, in region 16q12, a gene that was linked to breast cancer susceptibility [**?**]. This region is also picked by SConES GS - which captures the structure of the genome -, and GM - which, on top of it, captures gene membership. These two also share other breast cancer related regions and genes: 3p24 (NEK10 [32]), 5p12 (FGF10, MRPS30 [33]), and 10q26 (FGFR2, Section FGFR2 is strongly associated with familial breast cancer). On the other hand, they select different regions: only SConES GS selects region 8q24 (PCAT1 [**?**]), while only SConES GM selects 10q24 ([**?**]). By dealing with SNP networks, SConES studies the association of non-coding regions, as well as SNPs in any gene, coding or else. In fact, SConES GI, which adds to GM the interactions between genes, retrieves 4 subnetworks in intergenic regions, and 1

overlapping an RNA gene (RNU6-420P). SigMod, despite being related to SConES, produces a vastly different, large solution. On top of recovering familial breast cancer genes, a part of its subnetwork composed of keratins is focused on cytoskeleton (*structural constituent of cytoskeleton*, GO enrichment's adjusted P-value = 9.10×10^{-4}), a potentially novel susceptibility mechanisms to cancer.

heinz retrieves a small, highly informative set of biomarkers in a fast and stable fashion

As the methods produced such different results, we compared their solutions in a 5-fold subsampling setting (Section Classification accuracy of selected biomarkers). Specifically, we measured the following properties (Figure 2): (i) size of the solution subnetwork; (ii) stability; (iii) sensitivity and specificity of an L1-penalized logistic regression on the selected SNPs; and (iv) computational runtime.

Both solution size and active set of SNPs selected by Lasso varies greatly between the different methods (Figure 2A). heinz has the smallest solutions, with an average of 182 selected our of which 5.6% (10.2) are selected by Lasso. The largest solutions come from SConES GM (4548.6 SNPs), and dmGWAS (4307.4 SNPs). Interestingly, SigMod and SConES GI have the highest proportion of the selected SNPs that go into the active set (11.47 and 10.3% respectively). This suggests those methods are selecting more informative SNPs on average.

The sensitivity and specificity of the classifier on the testing data informs us about the usefulness of the selected SNPs as patient classification (Figure 2B). All classifiers' sensitivities were in the 0.38 - 0.69 range; the specificities, between 0.40 and 0.70. On average, SConES GS had the highest sensitivity (0.57); heinz, the highest specificity (0.56). Both SConES GS and SConES GM had on average better sensitivity than the classifier trained on all the SNPs, and dmGWAS and heinz superior specificities. However, the differences them were negligible, well within the 95% confidence interval.

The stability of a method measures its ability to select the same SNPs in face of perturbations on the data. We measured it by computing the pairwise Jaccard similarities between all pairs of solutions (Figure 2C). Heinz's displayed a high stability in our benchmark, consistently selecting the same SNPs over the 5 subsamples. LEAN also showed a high stability consistently selecting no SNP.

In terms of computational runtime, the fastest method was heinz (Figure 2D), which leverages on its ability to find efficiently the solution in a

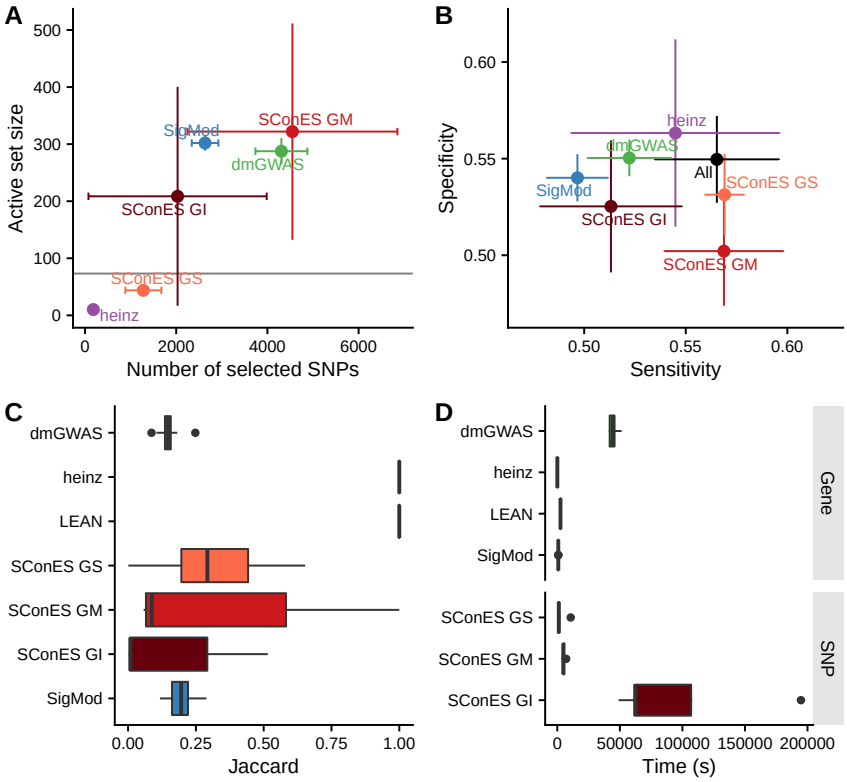

Figure 2: Comparison of network-based GWAS methods on GENESIS. Each method was run 5 times of a random subset of the samples, and tested on the remaining samples (Section Classification accuracy of selected biomarkers). **(A)** Number of SNPs selected by each method and number of SNPs on the active set used by the Lasso classifier. Points are the average over the 5 runs; lines represent the standard error of the mean. The horizontal grey line represents the average active set of Lasso using all the SNPs. **(B)** Sensitivity and specificity on testing set of the L1-penalized logistic regression trained on the features selected by each of the methods. In addition, the performance of the classifier trained on all SNPs is displayed. Points are the average over the 5 runs; lines represent the standard error of the mean. **(C)** Pairwise Jaccard similarities of the solutions used by different methods. A Jaccard similarity of 1 means the two solutions are the same. A Jaccard similarity of 0 means that there is no SNP in common between the two solutions. **(D)** Runtime of the evaluated methods, by type of network used (gene or SNP). The gene network-based methods required an additional 119980 seconds (1 day and 9.33 hours) on average to compute the gene scores from SNP summary statistics (not included in the displayed Time).

few seconds. The slowest method was SConES using the GI network, with approximately 1 day and 2.38 hours on average. Including the time required to compute the gene scores, however, slows down considerably gene-based methods; on this benchmark, that step took on average 1 day and 9.33 hours. Considering that time, dmGWAS is the slowest method, taking 1 day and 21.81 hours on average.

No such thing as perfect

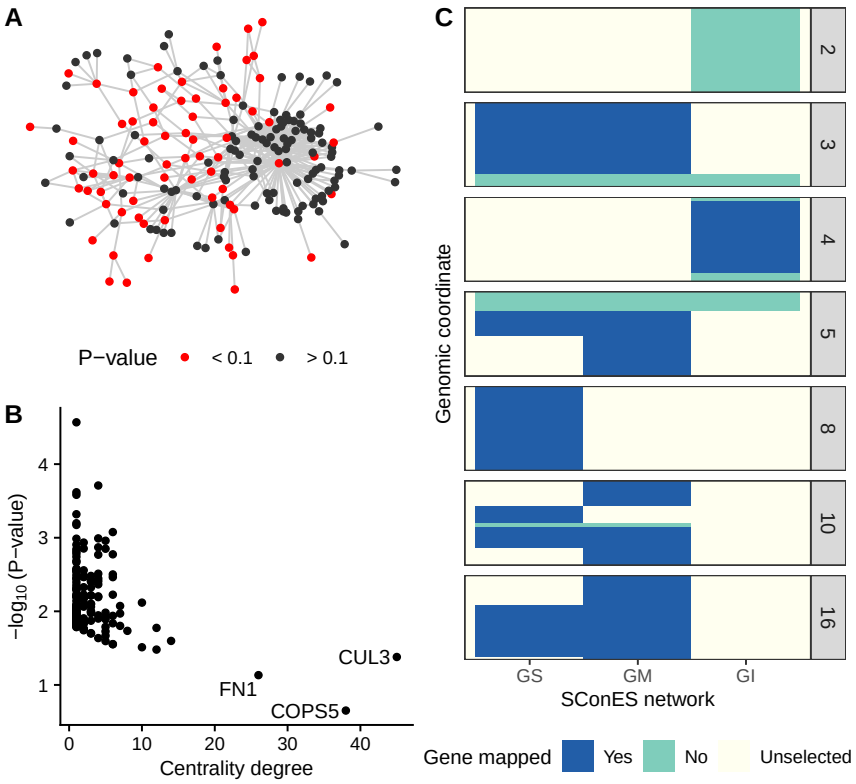

Figure 3: Drawbacks confronted when using network guided methods. **(A)** dmGWAS solution subnetwork. Genes with a P-value < 0.1 are highlighted in red. **(B)** Centrality degree and -$\log_{10}$ of the VEGAS P-value for the nodes in SigMod solution subnetwork. **(C)** Genomic regions where either SConES GS, GM or GI select SNPs.

In practice, and despite their similarities and their involvement in cancer mechanisms, the solutions are remarkably different (Supplementary figure

2A). That is due to the particulars of the methods, and directly or indirectly, they provide information about the dataset. For instance, the fact that LEAN did not provide any biomarkers implies that there is no gene such that both itself and its environment are on average strongly associated with the disease.

In this dataset, heinz's solution is very conservative, providing a small solution with the lowest median P-value for the subnetwork (Table 2). Due to this parsimonious and highly associated solution, it was the best method to select a set of good biomarkers for classification. (Figure 2B). Its conservativeness stems from its preprocessing step, which models the gene p-values as a mixture model of a beta and a uniform distribution, controlled by an FDR parameter. Due to the limited signal at the gene level in this dataset (Figure 1B), only 36 of them are considered to be associated to the disease. Hence, heinz's solution subnetwork consists only of 4 genes, which does not provide much insight of the biology of cancer. Importantly, it ignores genes that are strongly associated to cancer in this dataset like FGFR2.

On the other end of the spectrum, we have large, less conservative solutions provided by dmGWAS and SigMod. In fact both solutions present a relatively large overlap (Jaccard similarity = 0.16). However, they are also among the least associated subnetworks on average. In the case of dmGWAS, this is due to the greedy framework used to solve the problem: considering neighbors at distance 2, it accepts genes weakly associated to cancer if they are linked to another, strongly associated gene. This compounds when aggregating the results of successive greedy searches, and in fact we observe a large cluster of unassociated genes (Figure 3A). We observe the same tendency in SigMod's network, where the most central genes are the least associated to the disease (Figure 3B). Additionally, the relatively low signal-to-noise ratio combined with the large solution requires additional analyses to draw conclusions, such as enrichment analyses. Lastly, SigMod misses some of the most strongly associated, familial breast cancer genes in the dataset, like FGFR2 and TOX3.

By virtue of using a SNP subnetwork, SConES analyzes each SNP in their context. Thanks to that, it selects SNPs in genes none of whose interactors are associated to the disease, as well as SNPs in non-coding regions or in non-interacting genes. In fact, due to linkage disequilibrium, such genes are favored by SConES, as selecting one favors selecting another one. This might explain why the GS and GM networks, heavily affected by linkage disequilibrium, produce similar results (Supplementary figure 2B). On the other hand, SConES penalizes selecting SNPs and not their neighbors. This makes it conservative regarding SNPs with many interactions, for instance

those mapped to hubs in the PPIN. Influenced by this, SConES GI did not select any protein coding gene, despite selecting similar regions as SConES GS (Figure 3C). Also, the iCOGS platform is not a real GWAS experiment: the genome is not unbiasedly surveyed, some regions are fine-mapped - which might distort gene structure in GM and GI networks- while others are under studied - hurting the accuracy with which the GS network captures the genome structure.

TODO Aggregating solutions provides insights into the biology of cancer

To leverage on the strengths of each of the methods and compensate their respective weaknesses, we built a consensus subnetwork that captures the mechanisms most shared among the solution subnetworks (Section Consensus network). The consensus subnetwork (Figure 4) contains 53 genes and is enriched in familial breast cancer genes (Fisher's exact test P-value = 1.63×10^{-10}). Due to the limited overlap between methods, only 7 genes were common to more than two of them (Supplementary figure 6A). Interestingly, the more methods selected a gene, the higher its association was (Supplementary figure 6B). Disease genes have a higher betweenness centrality than non-disease genes [5]. We observe that to be the case in the disease under study (one-tailed Wilcoxon rank-sum test P-value = 2.64×10^{-5}, Supplementary figure 6C). Accordingly, the genes on the consensus network also had a higher betweenness centrality than the rest of the genes (one-tailed Wilcoxon rank-sum test P-value = 4.77×10^{-14}). Interestingly, cancer genes in the consensus network are less central than non-cancer genes in the consensus network (Wilcoxon rank-sum test P-value = 0.05). We studied if highly central genes were selected not because they were associated themselves, but because they were involved in the shortest path between two highly associated genes. As we found a weak correlation supporting this (Pearson correlation coefficient = -0.30, Supplementary figure 6D), we hypothesize that genes highly associated with the disease in this dataset and highly central might contribute to the heritability in the French population.

The consensus subnetwork is not completely connected: out of the 53 genes, the largest connected subnetwork includes only 40 genes. A GO enrichment analysis showed that this component is related to three major cellular processes: DNA helicase activity (adjusted P-value = 0.005), unfolded protein binding (adjusted P-value = 0.01), and poly(U) RNA binding (adjusted P-value = 0.01). We found support in the literature of the involvement of each of these functions in the development of cancer. *DNA*

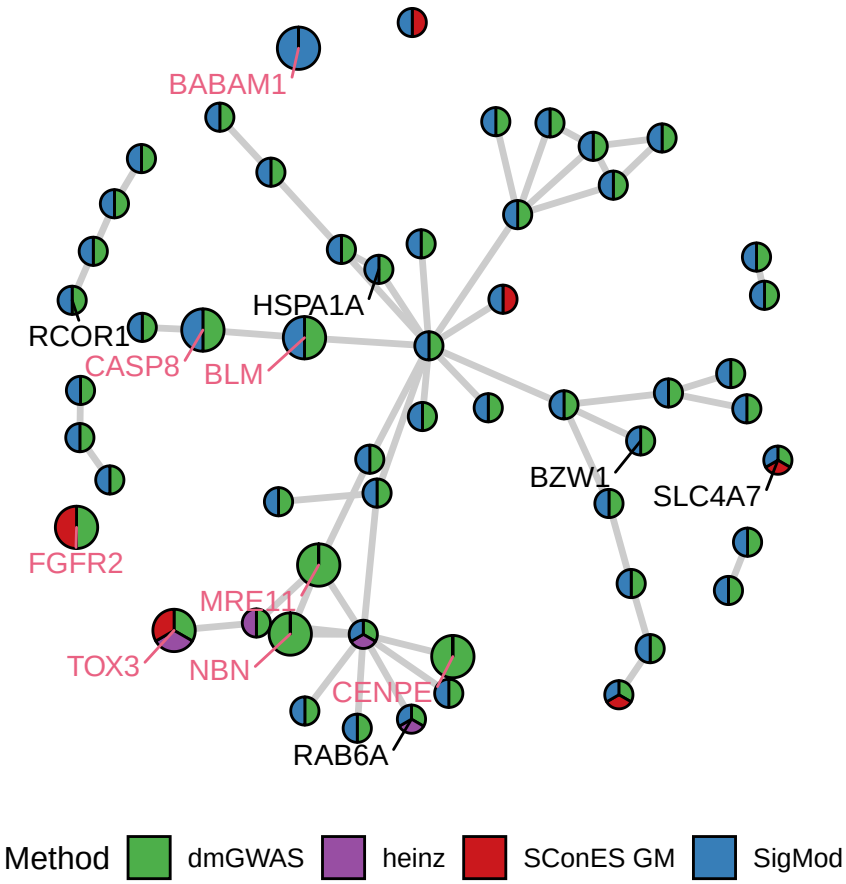

Figure 4: Consensus subnetwork on GENESIS (Section Consensus network). Each node is represented by a pie chart, which accounts the methods that selected it. The labeled genes have a VEGAS2 P-value < 0.001 and/or are known familial breast cancer genes (colored in pink).

helicase activity, for instance, is crucial for DNA repair [**?**]. Disruption of the DNA repair machinery is long-known to increase the likelihood of cancer, since mutations of BRCA1/2 were discovered [**?**]. Unsurprisingly, it involves three familial breast cancer genes (BLM, NBN, MRE11). Another enriched activity, *unfolded protein binding*, inhibits caspase-dependent apoptosis, improving the chances of developing cancer [**?**]. Three heat-shock proteins (HSPA1A, HSPA1B, HSPA1L) participate in this biological function.

Remarkably, 4 of the 53 genes are not linked to any other gene. Two of the latter are known familial breast cancer genes: FGFR2 (Section FGFR2 is strongly associated with familial breast cancer), and BABAM1 (VEGAS P-value = 3.22×10^{-3}). The other two are SLC4A7 and MRPS30. SLC4A7 (VEGAS P-value = 2.70×10^{5}) is a gene encoding a sodium bicarbonate cotransporter, which is selected by dmGWAS, SigMod, and SConES GM. It has linked to BRCA in the past [**?**]. MRPS30 (VEGAS P-value = 0.001) is a nuclear gene encoding a mitochondrial ribosomal protein.

TODO Hindrances of network analyses

The strength of network-based analyses comes from leveraging prior knowledge to boost discovery. However, they falter in front of genes with no prior knowledge, in other words, genes not in the network. Out of the 32767 genes that we can map the genotyped SNPs, 60.7% (19887) are not in the protein-protein interaction network. Out of those 5227 are protein coding (Supplementary figure 3). Among them, we find NEK10 (P-value 1.6×10^{-5}), linked to breast cancer susceptibility in the past [32], and POU5F1B, linked to prostate cancer [34]. However SNPs in NEK10 are selected by both SConES GS and GM, which do not use PPIs. Broadly speaking, protein coding genes absent from the PPIN are less associated with the phenotype on average (Wilcoxon rank-sum P-value = 2.79×10^{-8}). As we are dealing with high-throughput interactions, such difference cannot be due to the focus on well-known genes. Likely, it speaks to the fact that interactions involving genes with more interactions are more likely, and disease genes tend to be more central than average [5]. However, the difference is rather small: protein-coding genes in the network have a median P-value of 0.43, versus the 0.47 of those absent from it.

As not all databases compile the same interactions, the choice of the PPIN determines the final output. Specifically, in this work we used exclusively interactions from HINT from high-throughput experiments. This responds to concerns of some authors about biases introduced by adding interactions coming from targeted studies in the literature. It can be summa-

rized as "the rich get richer", where popular genes have a higher proportion of their interactions described. It has been reported that using such interactions might introduce biases in topological analyses [18]. On the other hand, one study found that the best predictor of the performance of a network for disease gene discovery is the size of the network [6]. This would support using the largest amount of interactions. To clarify their impact on this study, we compared the impact of using only physical interactions from high-throughput experiment versus interactions from both high-throughput and the literature (Section Gene-gene network). However, we found that for most of the methods using this expanded network did not have a great impact in size of the solution, classification accuracy, stability of the solution or runtime (Supplementary figure 5).

Lastly, we cannot forget the genes that are left out of the network due to our choice of focusing on PPIs. In fact, the largest group of unanalyzed genes are non-coding genes, mainly lncRNA, miRNA, and snRNA (Supplementary figure 3). The importance of these genes, like CASC16, is highlighted at the SNP-level, gene-level and again in SConES GS and GM analyses.

TODO Discussion

In this article we evaluate the viability of systems biology approach to GWAS, and examine a GWAS dataset on familial breast cancer focused on BRCA1/2 negative French women. Systems biology addresses two of the largest GWAS issues: interpretability and an overly conservative statistical framework that hinders discovery. This is achieved by considering the biological context of each of the genes and SNPs, and selecting a threshold of association based on it. However, the method of choice is unclear. Based on divergent considerations of what the desired set of biomarkers is, several methods for network-guided biomarker discovery have been proposed. In this article we reviewed the performance of six network-guided of them on GWAS. Despite their differences, most of them produced a relevant subset of biomarkers, recovering known familial breast cancer genes.

To overcome the problems posed by the individual methods, we propose combining them into a consensus subnetwork.

Each method had a different interface, required different preprocessing steps, and some exhibited unexpected behaviors. To facilitate their other authors applying them to new datasets and aggregating their solutions, we built nextflow pipelines [35] with a consistent interface and, whenever possible, parallelize the computation. They are available on GitHub:

`https://github.com/hclimente/gwas-tools`. Importantly, those methods that had a permissive license were compiled into a Docker image for easier use hclimente/gwas-tools.

References

[1] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Computational Biology*, vol. 8, p. e1002822, Dec. 2012. 00001.

[2] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An Expanded View of Complex Traits: From Polygenic to Omnigenic," *Cell*, vol. 169, pp. 1177–1186, June 2017. 00586.

[3] L. I. Furlong, "Human diseases through the lens of network biology," *Trends in Genetics*, vol. 29, pp. 150–159, Mar. 2013. 00128.

[4] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, pp. 56–68, Jan. 2011. 02826.

[5] J. Piñero, A. Berenstein, A. Gonzalez-Perez, A. Chernomoretz, and L. I. Furlong, "Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing," *Scientific Reports*, vol. 6, p. 24570, Apr. 2016. 00016.

[6] J. K. Huang, D. E. Carlin, M. K. Yu, W. Zhang, J. F. Kreisberg, P. Tamayo, and T. Ideker, "Systematic Evaluation of Molecular Networks for Discovery of Disease Genes," *Cell Systems*, vol. 6, pp. 484–495.e5, Apr. 2018. 00024.

[7] F. Gwinner, G. Boulday, C. Vandiedonck, M. Arnould, C. Cardoso, I. Nikolayeva, O. Guitart-Pla, C. V. Denis, O. D. Christophe, J. Beghain, E. Tournier-Lasserve, and B. Schwikowski, "Network-based analysis of omics data: The LEAN method," *Bioinformatics*, p. btw676, Oct. 2016. 00007.

[8] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao, "dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks," *Bioinformatics*, vol. 27, pp. 95–102, Jan. 2011. 00205.

[9] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Muller, "Identifying functional modules in protein-protein interaction networks: an integrated exact approach," *Bioinformatics*, vol. 24, pp. i223–i231, July 2008. 00429.

[10] M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nature Genetics*, vol. 47, pp. 106–114, Feb. 2015. 00411.

[11] C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt, "Efficient network-guided multi-locus association mapping with graph cuts," *Bioinformatics*, vol. 29, pp. i171–i179, July 2013. 00047.

[12] Y. Liu, M. Brossard, D. Roqueiro, P. Margaritte-Jeannin, C. Sarnowski, E. Bouzigon, and F. Demenais, "SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network," *Bioinformatics*, p. btx004, Jan. 2017. 00007.

[13] O. M. Sinilnikova, M.-G. Dondon, S. Eon-Marchais, F. Damiola, L. Barjhoux, M. Marcou, C. Verny-Pierre, V. Sornin, L. Toulemonde, J. Beauvallet, D. Le Gal, N. Mebirouk, M. Belotti, O. Caron, M. Gauthier-Villars, I. Coupier, B. Buecher, A. Lortholary, C. Dugast, P. Gesta, J.-P. Fricker, C. Noguès, L. Faivre, E. Luporsi, P. Berthet, C. Delnatte, V. Bonadona, C. M. Maugard, P. Pujol, C. Lasset, M. Longy, Y.-J. Bignon, C. Adenis, L. Venat-Bouvet, L. Demange, H. Dreyfus, M. Frenay, L. Gladieff, I. Mortemousque, S. Audebert-Bellanger, F. Soubrier, S. Giraud, S. Lejeune-Dumoulin, A. Chevrier, J.-M. Limacher, J. Chiesa, A. Fajac, A. Floquet, F. Eisinger, J. Tinat, C. Colas, S. Fert-Ferrer, C. Penet, T. Frebourg, M.-A. Collonge-Rame, E. Barouk-Simonet, V. Layet, D. Leroux, O. Cohen-Haguenauer, F. Prieur, E. Mouret-Fourme, F. Cornélis, P. Jonveaux, O. Bera, E. Cavaciuti, A. Tardivon, F. Lesueur, S. Mazoyer, D. Stoppa-Lyonnet, and N. Andrieu, "GENESIS: a French national resource to study the missing heritability of breast cancer," *BMC Cancer*, vol. 16, p. 13, Dec. 2016. 00005.

[14] L. C. Sakoda, E. Jorgenson, and J. S. Witte, "Turning of COGS moves forward findings for hormonally mediated cancers," *Nature Genetics*, vol. 45, pp. 345–348, Apr. 2013. 00060.

[15] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: rising to the challenge of larger and richer datasets," *GigaScience*, vol. 4, p. 7, Dec. 2015. 01610.

[16] A. Mishra and S. Macgregor, "VEGAS2: Software for More Flexible Gene-Based Testing," *Twin Research and Human Genetics*, vol. 18, pp. 86–91, Feb. 2015. 00125.

[17] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek, "GENCODE reference annotation for the human and mouse genomes," *Nucleic Acids Research*, vol. 47, pp. D766–D773, Jan. 2019. 00063.

[18] J. Das and H. Yu, "HINT: High-quality protein interactomes and their applications in understanding human disease," *BMC Systems Biology*, vol. 6, no. 1, p. 92, 2012. 00204.

[19] I. Ljubić, R. Weiskircher, U. Pferschy, G. W. Klau, P. Mutzel, and M. Fischetti, "An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem," *Mathematical Programming*, vol. 105, pp. 427–449, Feb. 2006. 00223.

[20] D. Beisser, G. W. Klau, T. Dandekar, T. Muller, and M. T. Dittrich, "BioNet: an R-Package for the functional analysis of biological networks," *Bioinformatics*, vol. 26, pp. 1129–1130, Apr. 2010. 00188.

[21] M. Dittrich and D. Beisser, "Bionet." `https://bioconductor.org/packages/BioNet/`, 2008. Accessed: 2019-07-16.

[22] M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, "Hotnet2." `https://github.com/raphael-group/hotnet2`, 2018. Accessed: 2019-07-16.

[23] Q. Wang and P. Jia, "dmgwas 3.0." `https://bioinfo.uth.edu/dmGWAS/`, 2014. Accessed: 2019-07-16.

[24] F. Gwinner, "Leanr." `https://cran.r-project.org/web/packages/LEANR/`, 2016. Accessed: 2019-07-16.

[25] H. Climente-González and C.-A. Azencott, "martini." `https://www.bioconductor.org/packages/martini/`, 2019. Accessed: 2019-07-16.

[26] Y. Liu, "Sigmod v2." `https://github.com/YuanlongLiu/SigMod`, 2018. Accessed: 2019-07-16.

[27] J. Piñero, Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, pp. D833–D839, Jan. 2017. 00369.

[28] K. Michailidou, J. Beesley, S. Lindstrom, S. Canisius, J. Dennis, M. J. Lush, M. J. Maranian, M. K. Bolla, Q. Wang, M. Shah, B. J. Perkins, K. Czene, M. Eriksson, H. Darabi, J. S. Brand, S. E. Bojesen, B. G. Nordestgaard, H. Flyger, S. F. Nielsen, N. Rahman, C. Turnbull, O. Fletcher, J. Peto, L. Gibson, I. dos Santos-Silva, J. Chang-Claude, D. Flesch-Janys, A. Rudolph, U. Eilber, S. Behrens, H. Nevanlinna, T. A. Muranen, K. Aittomäki, C. Blomqvist, S. Khan, K. Aaltonen, H. Ahsan, M. G. Kibriya, A. S. Whittemore, E. M. John, K. E. Malone, M. D. Gammon, R. M. Santella, G. Ursin, E. Makalic, D. F. Schmidt, G. Casey, D. J. Hunter, S. M. Gapstur, M. M. Gaudet, W. R. Diver, C. A. Haiman, F. Schumacher, B. E. Henderson, L. Le Marchand, C. D. Berg, S. J. Chanock, J. Figueroa, R. N. Hoover, D. Lambrechts, P. Neven, H. Wildiers, E. van Limbergen, M. K. Schmidt, A. Broeks, S. Verhoef, S. Cornelissen, F. J. Couch, J. E. Olson, E. Hallberg, C. Vachon, Q. Waisfisz, H. Meijers-Heijboer, M. A. Adank, R. B. van der Luijt, J. Li, J. Liu, K. Humphreys, D. Kang, J.-Y. Choi, S. K. Park, K.-Y. Yoo, K. Matsuo, H. Ito, H. Iwata, K. Tajima, P. Guénel, T. Truong, C. Mulot, M. Sanchez, B. Burwinkel, F. Marme, H. Surowy, C. Sohn,

A. H. Wu, C.-c. Tseng, D. Van Den Berg, D. O. Stram, A. González-Neira, J. Benitez, M. P. Zamora, J. I. A. Perez, X.-O. Shu, W. Lu, Y.-T. Gao, H. Cai, A. Cox, S. S. Cross, M. W. R. Reed, I. L. Andrulis, J. A. Knight, G. Glendon, A. M. Mulligan, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, A. Lindblom, S. Margolin, S. H. Teo, C. H. Yip, N. A. M. Taib, G.-H. Tan, M. J. Hooning, A. Hollestelle, J. W. M. Martens, J. M. Collée, W. Blot, L. B. Signorello, Q. Cai, J. L. Hopper, M. C. Southey, H. Tsimiklis, C. Apicella, C.-Y. Shen, C.-N. Hsiung, P.-E. Wu, M.-F. Hou, V. N. Kristensen, S. Nord, G. I. G. Alnaes, G. G. Giles, R. L. Milne, C. McLean, F. Canzian, D. Trichopoulos, P. Peeters, E. Lund, M. Sund, K.-T. Khaw, M. J. Gunter, D. Palli, L. M. Mortensen, L. Dossus, J.-M. Huerta, A. Meindl, R. K. Schmutzler, C. Sutter, R. Yang, K. Muir, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, M. Hartman, H. Miao, K. S. Chia, C. W. Chan, P. A. Fasching, A. Hein, M. W. Beckmann, L. Haeberle, H. Brenner, A. K. Dieffenbach, V. Arndt, C. Stegmaier, A. Ashworth, N. Orr, M. J. Schoemaker, A. J. Swerdlow, L. Brinton, M. Garcia-Closas, W. Zheng, S. L. Halverson, M. Shrubsole, J. Long, M. S. Goldberg, F. Labrèche, M. Dumont, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, M. Grip, H. Brauch, U. Hamann, T. Brüning, P. Radice, P. Peterlongo, S. Manoukian, L. Bernard, N. V. Bogdanova, T. Dörk, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, P. Devilee, R. A. E. M. Tollenaar, C. Seynaeve, C. J. Van Asperen, A. Jakubowska, J. Lubinski, K. Jaworska, T. Huzarski, S. Sangrajrang, V. Gaborieau, P. Brennan, J. McKay, S. Slager, A. E. Toland, C. B. Ambrosone, D. Yannoukakos, M. Kabisch, D. Torres, S. L. Neuhausen, H. Anton-Culver, C. Luccarini, C. Baynes, S. Ahmed, C. S. Healey, D. C. Tessier, D. Vincent, F. Bacot, G. Pita, M. R. Alonso, N. Álvarez, D. Herrero, J. Simard, P. P. D. P. Pharoah, P. Kraft, A. M. Dunning, G. Chenevix-Trench, P. Hall, and D. F. Easton, "Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer," *Nature Genetics*, vol. 47, pp. 373–380, Apr. 2015. 00000.

[29] E. S. Rinella, Y. Shao, L. Yackowski, S. Pramanik, R. Oratz, F. Schnabel, S. Guha, C. LeDuc, C. L. Campbell, S. D. Klugman, M. B. Terry, R. T. Senie, I. L. Andrulis, M. Daly, E. M. John, D. Roses, W. K. Chung, and H. Ostrer, "Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation," *Human Genetics*, vol. 132, pp. 523–536, May 2013. 00019.

[30] A. G. Brisbin, Y. W. Asmann, H. Song, Y.-Y. Tsai, J. A. Aakre, P. Yang, R. B. Jenkins, P. Pharoah, F. Schumacher, D. V. Conti, D. J. Duggan, M. Jenkins, J. Hopper, S. Gallinger, P. Newcomb, G. Casey, T. A. Sellers, and B. L. Fridley, "Meta-analysis of 8q24 for seven cancers reveals a locus between NOV and ENPP2 associated with cancer development," *BMC Medical Genetics*, vol. 12, p. 156, Dec. 2011. 00033.

[31] SEARCH, The GENICA Consortium, kConFab, Australian Ovarian Cancer Study Group, S. Ahmed, G. Thomas, M. Ghoussaini, C. S. Healey, M. K. Humphreys, R. Platte, J. Morrison, M. Maranian, K. A. Pooley, R. Luben, D. Eccles, D. G. Evans, O. Fletcher, N. Johnson, I. dos Santos Silva, J. Peto, M. R. Stratton, N. Rahman, K. Jacobs, R. Prentice, G. L. Anderson, A. Rajkovic, J. D. Curb, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, W. R. Diver, S. Bojesen, B. G. Nordestgaard, H. Flyger, T. Dörk, P. Schürmann, P. Hillemanns, J. H. Karstens, N. V. Bogdanova, N. N. Antonenkova, I. V. Zalutsky, M. Bermisheva, S. Fedorova, E. Khusnutdinova, D. Kang, K.-Y. Yoo, D. Y. Noh, S.-H. Ahn, P. Devilee, C. J. van Asperen, R. A. E. M. Tollenaar, C. Seynaeve, M. Garcia-Closas, J. Lissowska, L. Brinton, B. Peplonska, H. Nevanlinna, T. Heikkinen, K. Aittomäki, C. Blomqvist, J. L. Hopper, M. C. Southey, L. Smith, A. B. Spurdle, M. K. Schmidt, A. Broeks, R. R. van Hien, S. Cornelissen, R. L. Milne, G. Ribas, A. González-Neira, J. Benitez, R. K. Schmutzler, B. Burwinkel, C. R. Bartram, A. Meindl, H. Brauch, C. Justenhoven, U. Hamann, J. Chang-Claude, R. Hein, S. Wang-Gohrke, A. Lindblom, S. Margolin, A. Mannermaa, V.-M. Kosma, V. Kataja, J. E. Olson, X. Wang, Z. Fredericksen, G. G. Giles, G. Severi, L. Baglietto, D. R. English, S. E. Hankinson, D. G. Cox, P. Kraft, L. J. Vatten, K. Hveem, M. Kumle, A. Sigurdson, M. Doody, P. Bhatti, B. H. Alexander, M. J. Hooning, A. M. W. van den Ouweland, R. A. Oldenburg, M. Schutte, P. Hall, K. Czene, J. Liu, Y. Li, A. Cox, G. Elliott, I. Brock, M. W. R. Reed, C.-Y. Shen, J.-C. Yu, G.-C. Hsu, S.-T. Chen, H. Anton-Culver, A. Ziogas, I. L. Andrulis, J. A. Knight, J. Beesley, E. L. Goode, F. Couch, G. Chenevix-Trench, R. N. Hoover, B. A. J. Ponder, D. J. Hunter, P. D. P. Pharoah, A. M. Dunning, S. J. Chanock, and D. F. Easton, "Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2," *Nature Genetics*, vol. 41, pp. 585–590, May 2009. 00000.

[32] S. Ahmed, G. Thomas, M. Ghoussaini, C. S. Healey, M. K. Humphreys,

R. Platte, J. Morrison, M. Maranian, K. A. Pooley, R. Luben, D. Eccles, D. G. Evans, O. Fletcher, N. Johnson, I. dos Santos Silva, J. Peto, M. R. Stratton, N. Rahman, K. Jacobs, R. Prentice, G. L. Anderson, A. Rajkovic, J. D. Curb, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, W. R. Diver, S. Bojesen, B. G. Nordestgaard, H. Flyger, T. Dörk, P. Schürmann, P. Hillemanns, J. H. Karstens, N. V. Bogdanova, N. N. Antonenkova, I. V. Zalutsky, M. Bermisheva, S. Fedorova, E. Khusnutdinova, D. Kang, K.-Y. Yoo, D. Y. Noh, S.-H. Ahn, P. Devilee, C. J. van Asperen, R. A. E. M. Tollenaar, C. Seynaeve, M. Garcia-Closas, J. Lissowska, L. Brinton, B. Peplonska, H. Nevanlinna, T. Heikkinen, K. Aittomäki, C. Blomqvist, J. L. Hopper, M. C. Southey, L. Smith, A. B. Spurdle, M. K. Schmidt, A. Broeks, R. R. van Hien, S. Cornelissen, R. L. Milne, G. Ribas, A. González-Neira, J. Benitez, R. K. Schmutzler, B. Burwinkel, C. R. Bartram, A. Meindl, H. Brauch, C. Justenhoven, U. Hamann, J. Chang-Claude, R. Hein, S. Wang-Gohrke, A. Lindblom, S. Margolin, A. Mannermaa, V.-M. Kosma, V. Kataja, J. E. Olson, X. Wang, Z. Fredericksen, G. G. Giles, G. Severi, L. Baglietto, D. R. English, S. E. Hankinson, D. G. Cox, P. Kraft, L. J. Vatten, K. Hveem, M. Kumle, A. Sigurdson, M. Doody, P. Bhatti, B. H. Alexander, M. J. Hooning, A. M. W. van den Ouweland, R. A. Oldenburg, M. Schutte, P. Hall, K. Czene, J. Liu, Y. Li, A. Cox, G. Elliott, I. Brock, M. W. R. Reed, C.-Y. Shen, J.-C. Yu, G.-C. Hsu, S.-T. Chen, H. Anton-Culver, A. Ziogas, I. L. Andrulis, J. A. Knight, J. Beesley, E. L. Goode, F. Couch, G. Chenevix-Trench, R. N. Hoover, B. A. J. Ponder, D. J. Hunter, P. D. P. Pharoah, A. M. Dunning, S. J. Chanock, and D. F. Easton, "Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2," *Nature Genetics*, vol. 41, pp. 585–590, May 2009. 00000.

[33] D. A. Quigley, E. Fiorito, S. Nord, P. Van Loo, G. G. Alnaes, T. Fleischer, J. Tost, H. K. Moen Vollan, T. Tramm, J. Overgaard, I. R. Bukholm, A. Hurtado, A. Balmain, A.-L. Børresen-Dale, and V. Kristensen, "The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors," *Molecular Oncology*, vol. 8, pp. 273–284, Mar. 2014. 00000.

[34] J. Breyer, D. Dorset, T. Clark, K. Bradley, T. Wahlfors, K. McReynolds, W. Maynard, S. Chang, M. Cookson, J. Smith, J. Schleutker, W. Dupont, and J. Smith, "An Expressed Retrogene of the Master Embryonic Stem Cell Gene POU5f1 Is Associated with Prostate Can-

cer Susceptibility," *The American Journal of Human Genetics*, vol. 94, pp. 395–404, Mar. 2014. 00018.

[35] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature Biotechnology*, vol. 35, pp. 316–319, Apr. 2017. 00176.

Supplementary materials

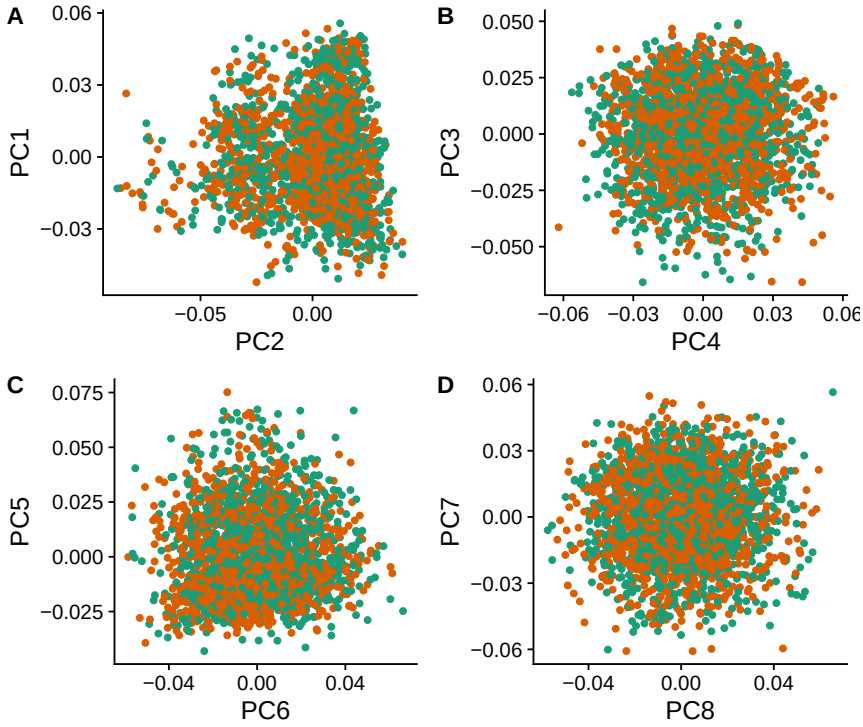

Figure 1: **(A,B,C,D)** Eight main principal components computed on the genotypes of GENESIS. Cases are colored in green, controls in orange.

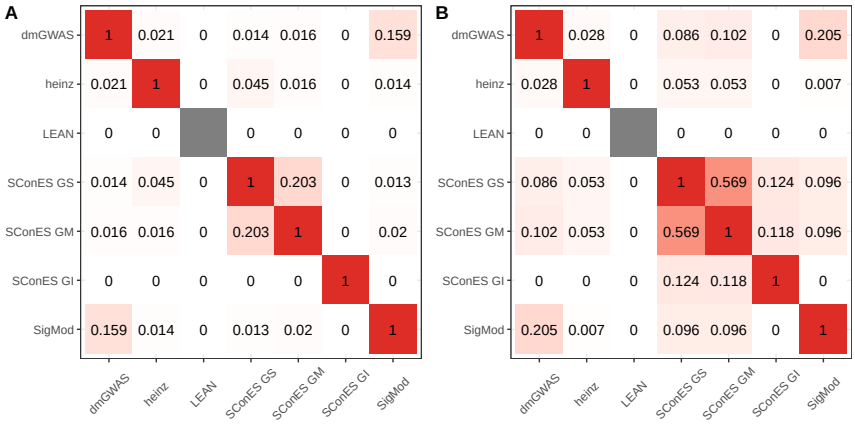

Figure 2: Jaccard similarity between the different solution gene subnetworks.

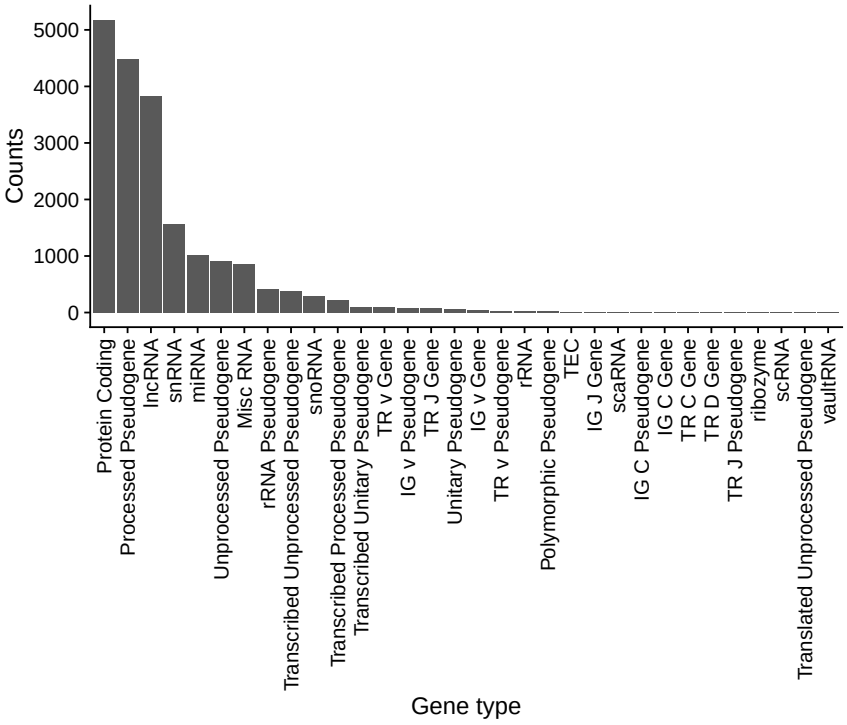

Figure 3: Biotypes of genes from the annotation that are not present in the HINT protein-protein interaction network.

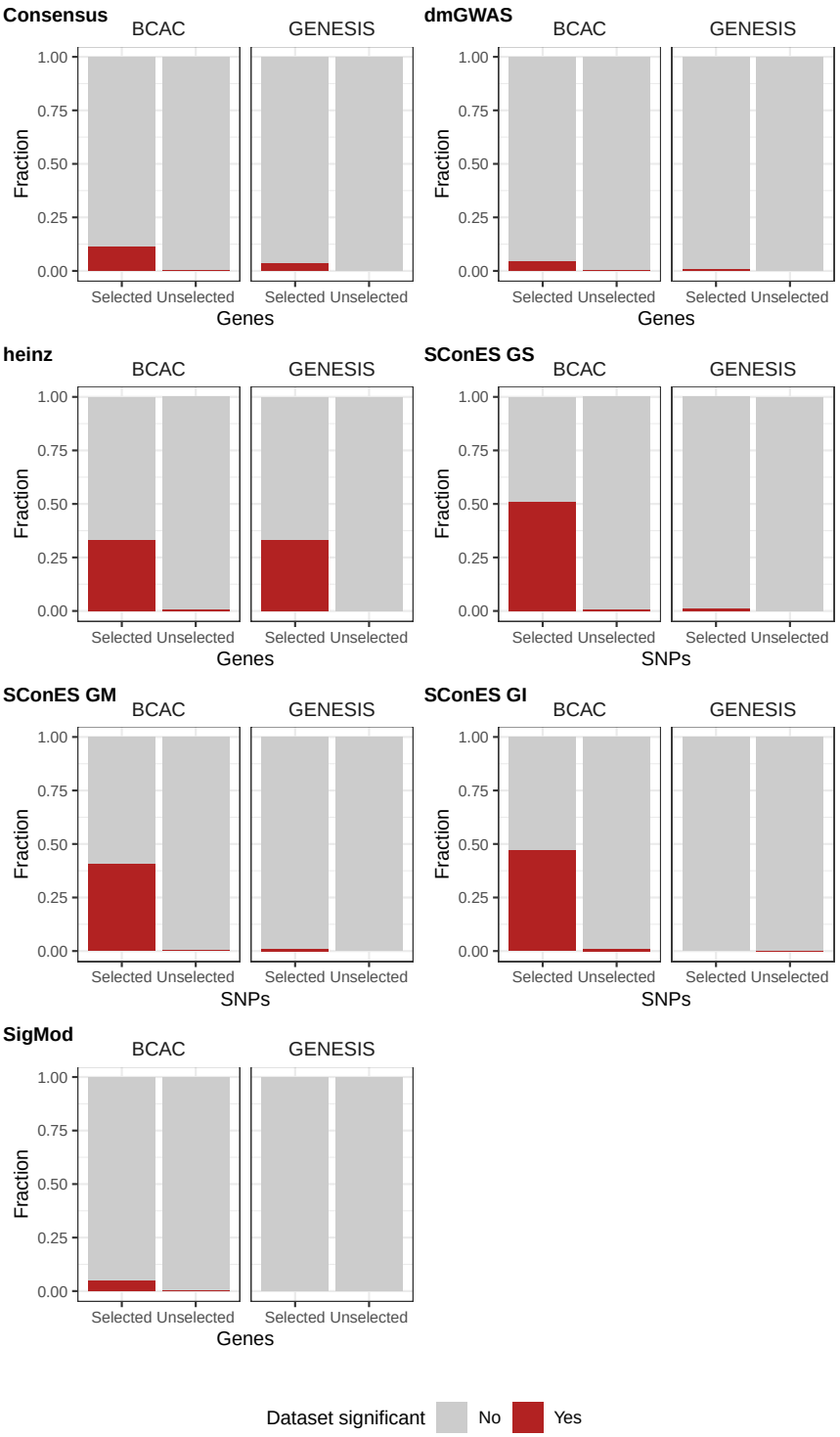

Figure 4: Bonferroni significance, in either the GENESIS or the BCAC datasets, of the genes (and SNPs in the case of SConES) detected by the network methods, and in the consensus subnetwork. LEAN was excluded, as it did not select any gene.

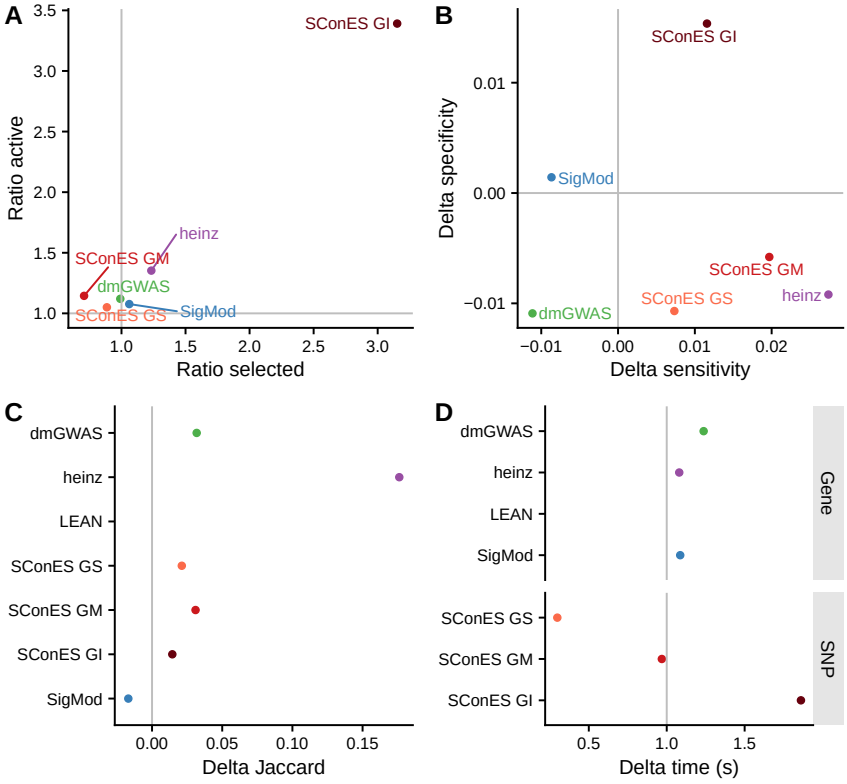

Figure 5: Comparison of benchmark on high-throughput interactions to benchmark on both high-throughput and literature curated interactions. Grey lines represent no change between the benchmarks (1 for ratios, 0 for differences). **(A)** Ratios of the selected features between both benchmarks and of the active set. **(B)** Shifts in sensitivity and specificity. **(C)** Shift in Jaccard similarity between benchmarks. **(D)** Ratio between the runtimes of the benchmarks.

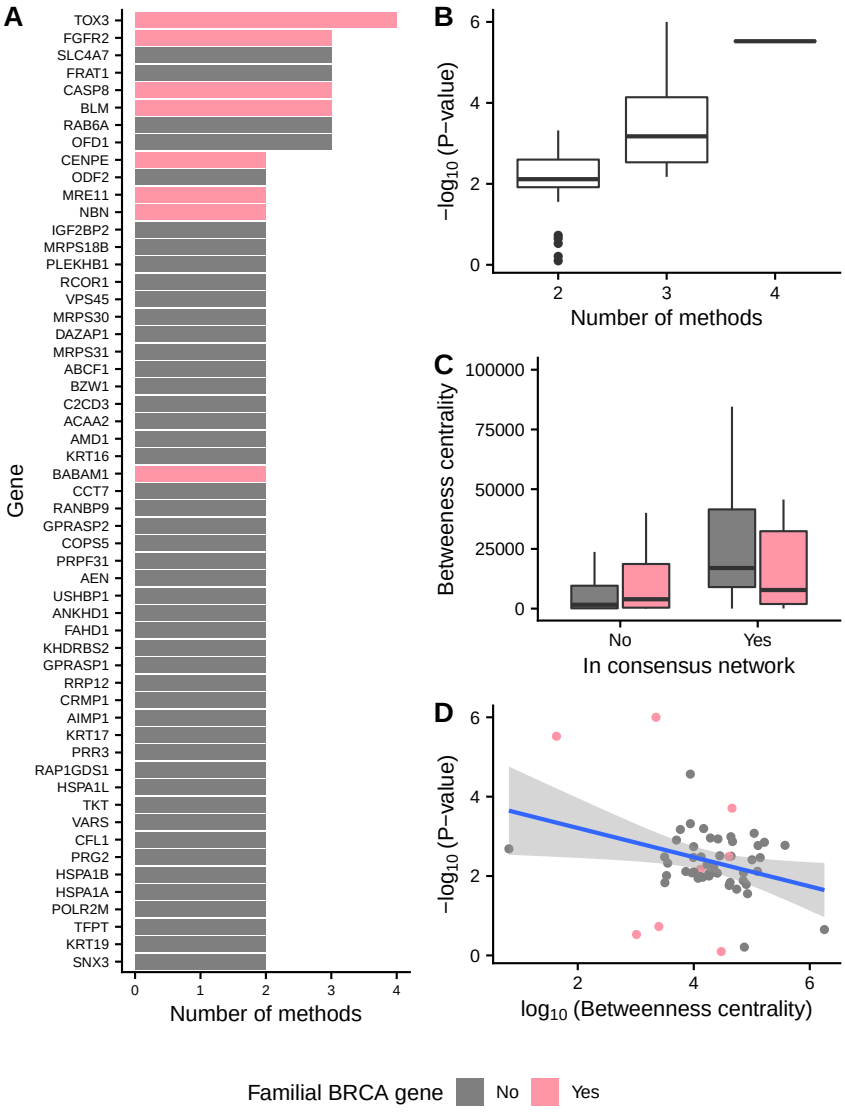

Figure 6: Genes on the consensus network. Familial breast cancer genes are colored in pink; the rest are colored in grey. **(A)** Number of methods selecting every gene in the subnetwork. **(B)** VEGAS P-values of association of the genes, with regards to the number of methods that selected them. **(C)** Comparison of betweenness centrality of the genes in the consensus network and the other genes in the PPIN and not in the consensus network. To improve visualization, we removed outliers. **(D)** Relationship between the $\log_{10}$ of the betweenness centrality and the -$\log_{10}$ of the VEGAS P-value of the genes in the consensus network. The blue line represents a fitted generalized linear model.