Comparing and combining network-guided biomarker discovery to discover genetic susceptibility mechanisms to breast cancer on the GENESIS study

Héctor Climente-González[1,2,3,4,♠], Christine Lonjou[1,2,3], Fabienne Lesueur[1,2,3], Dominique Stoppa-Lyonnet[5,6,7,♣], Nadine Andrieu[1,2,3], Chloé-Agathe Azencott[3,1,2]

[1]Institut Curie, PSL Research University, F-75005 Paris, France;
[2]INSERM, U900, F-75005 Paris, France;
[3]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France;
[4]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan;
[5]Service de Génétique, Institut Curie, F-75005 Paris, France;
[6]INSERM, U830, F-75005 Paris, France;
[7]Université Paris Descartes.
[♣]For the GENESIS study group
[♠]Corresponding author: hector.climente@riken.jp

Abstract

Systems biology provides a comprehensive approach to biomarker discovery and biological hypothesis building. Indeed, it allows to jointly consider the statistical association between gene variation and a phenotype, and the biological context of each gene, represented as a network. In this work, we study six network methods which identify subnetworks with high association scores to a phenotype. Specifically, we examine their utility to discover new biomarkers for breast cancer susceptibility by interrogating a genome-wide association study (GWAS) focused on French women with a family history of breast cancer and tested negative for pathogenic variants in *BRCA1* and *BRCA2*. We perform an in-depth benchmarking of the methods with regards to size of the solution subnetwork, their utility as biomarkers, and the stability and the runtime of the methods. By trading statistical stringency for biological meaningfulness, most network methods give more compelling results than standard SNP- and gene-level analyses, recovering causal subnetworks tightly related to cancer susceptibility. For instance, we show a general alteration of the neighborhood of *COPS5*, a gene related to multiple hallmarks of cancer. Importantly, we find a significantly large overlap between the genes in the solution networks and the genes significantly associated in the largest GWAS on susceptibility to breast cancer. Yet, network methods are notably unstable, producing different results when the input data changes slightly. To account for that, we produce a stable consensus subnetwork, formed by the most consistently selected genes. The stable consensus is composed of 68 genes, enriched in known breast cancer

susceptibility genes (*BLM*, *CASP8*, *CASP10*, *DNAJC1*, *FGFR2*, *MRPS30*, and *SLC4A7*, Fisher's exact test P-value = 3×10^{-4}) and occupying more central positions in the network than average. The network seems organized around *CUL3*, encoding an ubiquitin ligase related protein that regulates the protein levels of several genes involved in cancer progression. In conclusion, this article shows the pertinence of network-based analyses to tackle known issues with GWAS, namely lack of statistical power and of interpretable solutions. Project-agnostic implementations of each of the network methods are available at `https://github.com/hclimente/gwas-tools` to facilitate their application to other GWAS datasets.

Author summary

In genome-wide association studies (GWAS), thousands of genomes are scanned to identify variants associated with a complex trait. Over the last 15 years, GWAS have advanced our understanding of the genetics of complex diseases, and in particular of hereditary cancers. Yet, they have led to an apparent paradox: the more we perform such studies, the more it seems that the entire genome is involved in every disease. An elegant explanation has been proposed with the omnigenic model: only a limited number of core genes are directly involved in the disease; but gene functions are deeply interrelated, so that many other genes are able to alter the function of the core genes. These interrelations are often modeled as networks, and multiple algorithms have been proposed to use these networks to identify the subset of core genes involved in a specific trait. In this study, we characterize six such network methods on GENESIS, a GWAS dataset for familial breast cancer in the French population. Combining these approaches allows us to identify potentially novel breast cancer susceptibility genes, and provides a mechanistic explanation for their role in the development of the disease. Our pipeline can easily be applied to other diseases.

1 Introduction

In human health, genome-wide association studies (GWAS) aim at quantifying how single-nucleotide polymorphisms (SNPs) predispose to complex diseases, like diabetes or some forms of cancer [1]. To that end, in a typical GWAS 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. based on logistic regression). Those SNPs with a P-value lower than a conservative Bonferroni threshold are candidates to further studies in an independent cohort. Once the risk SNPs have been discovered, they can be used for risk assessment, and to deepen our understanding of the disease.

GWAS have successfully identified thousands of variants underlying many common diseases [2]. However, this experimental setting also presents intrinsic challenges. Some of them stem from the high dimensionality of the problem, as every GWAS to date studies more variants than samples are genotyped. This limits the statistical power of the experiment, as only variants with larger effects can be detected [3]. This is particularly problematic since the prevailing view is that most genetic architectures involve many variants with small effects [3]. Additionally, to avoid false positives, a conservative multiple test correction is applied, typically the previously mentioned Bonferroni correction. However, Bonferroni correction is overly conservative when the statistical tests are

correlated, as is the case in GWAS [4]. Another open issue is the interpretation of the results, as the functional consequences of most common variants are not well understood. On top of that, recent large-sampled studies suggest that numerous loci spread all along the genome contribute to a degree to any complex trait, in accordance with the infinitesimal model [5]. The recently proposed omnigenic model [6] offers an explanation: genes are very functionally inter-related and influence each other's behavior, which allows alterations in most genes to impact the subset of "core" genes directly involved in the disease's mechanism. Hence, a comprehensive statistical framework which includes the structure of biological data might address the aforementioned issues.

For this reason, many authors turn to network biology to handle the complex interplay of biomolecules that lead to disease [7]. As its name suggests, network biology models biology as a network, where the biomolecules under study, often genes, are nodes, and selected functional relationships are edges that link them. These relationships come from evidence that the genes jointly contribute to a biological function; for instance, their expression is correlated, or their products establish a protein-protein interaction. Under this view, complex diseases are not the consequence of a single altered gene, but of the interaction of multiple interdependent molecules [8]. In fact, an examination of biological networks shows that disease genes have differential properties [8, 9]. This is particularly true for cancer driver genes, which tend to be key players in connecting different, densely-connected communities of genes. Therefore, studying the neighborhood of disease-associated genes is effective at identifying new ones that are involved in the same biological functions [10].

Network-based biomarker discovery methods exploit this relatedness to identify disease genes on GWAS data [11]. In essence, each SNP has a score of association with the disease, given by the experiment, and functionally biological relationships, given by a network built on prior knowledge. Then, the problem becomes finding a functionally-related set of highly-scoring genes. Different solutions have been proposed to this problem, often stemming from divergent mathematical frameworks and considerations of what the optimal solution looks like. Some methods strongly constrain the problem to certain kinds of subnetworks. Such is the extreme case of LEAN [12], which focuses on "star" subnetworks, i.e. instances were both a gene and its direct interactors are associated with the disease. Other algorithms, like dmGWAS [13] and heinz [14], focus on larger subnetworks interconnecting genes with high association scores. However, they differ in their tolerance to the inclusion of low-scoring nodes, and the topology of the solution. Lastly, other methods also consider the topology of the network, favoring groups of nodes that are not only high-scoring, but also densely interconnected; such is the case of HotNet2 [15], SConES [16], and SigMod [17].

In this work, we analyze the effectiveness of these six network methods for biomarker discovery on GWAS data. While all of them capture susceptibility mechanisms resembling that postulated by the omnigenic model, they do so in different ways, and provide a representative view of the field. We worked on the GENESIS dataset [18], a study on familial breast cancer conducted in the French population. After a classical GWAS approach, we use these network methods to recover additional breast cancer biomarkers. 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.

2 Results

2.1 Conventional SNP- and gene-based analyses confirm that *FGFR2* locus is associated with familial breast cancer

We conducted association analyses in the GENESIS dataset at both SNP and gene levels (Section 4.3.1). Two genomic regions have a P-value lower than the Bonferroni threshold on chromosomes 10 and 16 (Supplementary Figure 1A). The former overlaps with gene *FGFR2*; the latter with *CASC16*, and it is located near the protein-coding gene *TOX3*. Variants in both *FGFR2* and *TOX3* have been repeatedly associated with breast cancer susceptibility in other case-control studies [19], *BRCA1* and *BRCA2* carrier studies [20], and in hereditary breast and ovarian cancer families negative for mutations in *BRCA1* and *BRCA2* [21]. In our studied population, only *FGFR2* was significantly associated with breast cancer at the gene-level (Supplementary Figure 1B).

Closer examination reveals two regions (3p24 and 8q24) having low, albeit not genome-wide significant, P-values. Both of them have been associated to breast cancer susceptibility in the past [22, 23]. We applied an L1-penalized logistic regression on the whole dataset (Section 4.5.2). It selected 100 SNPs, both from all aforementioned regions and new ones (Supplementary Figure 1C). Yet, it is unclear why those SNPs were selected, as emphasized by the high P-value of some of them, which further complicates the biological interpretation. Moreover, and in opposition to what would be expected under the omnigenic model, the genes to which these SNPs map to (Section 4.3.1) are not interconnected in the PPIN (Section 4.3.3). In addition, the classification performance of the method is very low, and L1-penalized methods select only one of several correlated variables and are prone to instability, which further complicates interpretation. This motivates exploring network methods, which trade statistical significance for biological relevance to find susceptibility subnetworks. In fact, such methods provided comparably (poor) classification performance to L1-penalized logistic regression (Figure 3B), while providing more interpretable solutions.

2.2 Network methods successfully identify genes associated with breast cancer

We applied six network methods to the GENESIS dataset (Section 4.3.4). As none of the networks examined by LEAN was significant (BH adjusted P-value < 0.05), we obtained six solutions (Figure 1): one for each of the remaining four gene-based methods, one for SConES GI (which works at the SNP level), and the consensus. These solutions differ in many aspects, making it hard to draw joint conclusions. For starters, the overlap between the genes featured in each solution is quite small (Figure 1B). Another prominent difference is their size: the largest solution, produced by HotNet2, contains 440 genes, while heinz's contained only 4 genes. While SConES GI failed to recover any protein coding gene, by dealing with SNP networks it retrieved four subnetworks in intergenic regions, and another one overlapping an RNA gene (*RNU6-420P*). Their topologies

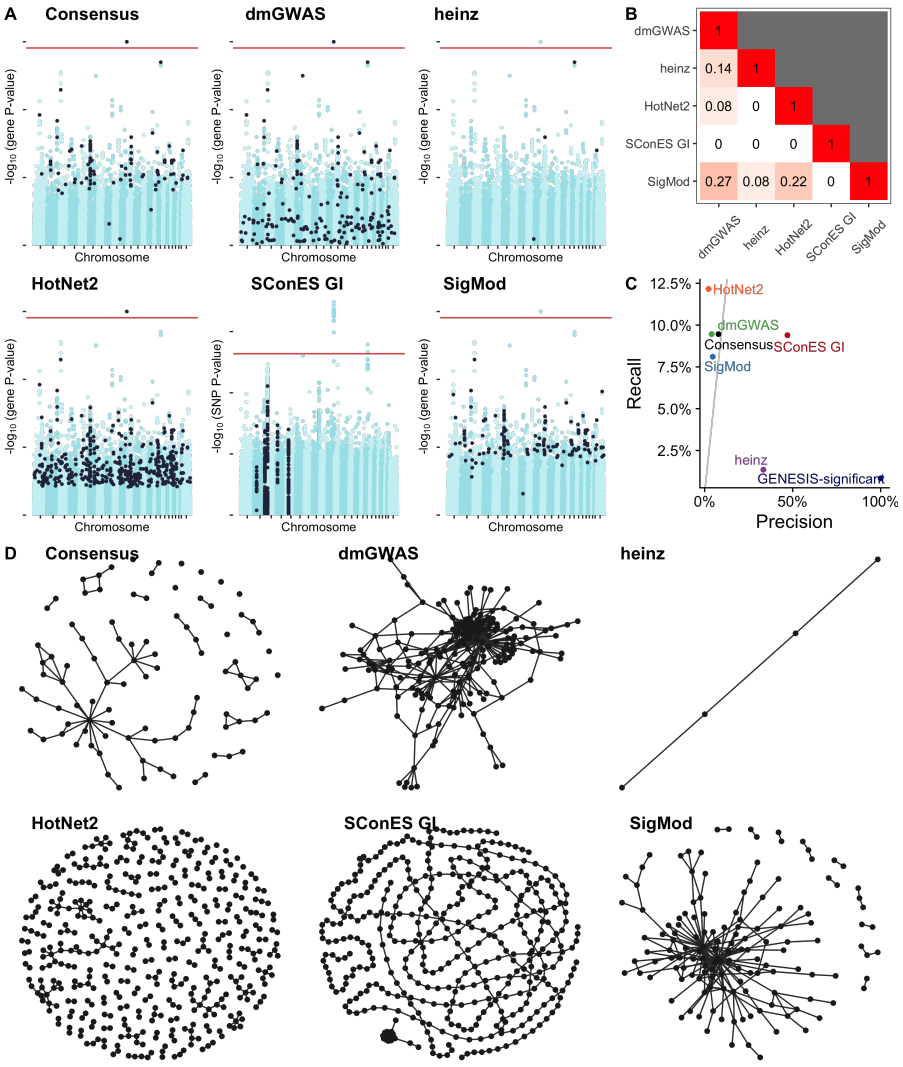

Fig. 1: Overview of the solutions produced by the different network methods (Section 4.3.4) on the GENESIS dataset. As LEAN did not produce any significant gene (BH adjusted P-value < 0.05), it was excluded. Unless indicated otherwise, results refer to genes, except for SConES GI which are at the SNP-level. **(A)** Manhattan plots of SNPs/genes; in black, the method's solution. The Bonferroni threshold is indicated by a red line (2.54×10^{-7} for SNPs, 1.53×10^{-6} for genes). **(B)** Overlap between the genes selected by each of the methods, measured by Pearson correlation between indicator vectors. **(C)** Precision and recall of the evaluated methods with respect to Bonferroni-significant SNPs/genes in BCAC. For reference, we added a gray line with a slope of 1. **(D)** Solution networks produced by the different methods.

Table 1: Summary statistics on the solutions of multiple network methods on the gene-gene interaction network. The first row contains the summary statistics on the whole network.

Network	# genes	# edges	$\overline{\text{Betweenness}}$	$\hat{\text{P}}_{\text{gene}}$	$\rho_{\text{consensus}}$
HINT HT	13 619	142 541	16 706	0.46	0.066
dmGWAS	194	450	49 115	0.19	0.41
heinz	4	3	113 633	0.001	0.21
HotNet2	440	374	7 739	0.048	0.31
LEAN	0	0	-	-	-
SConES GI	0 (1)	0	-	-	-
SigMod	142	249	92 603	0.008	0.73
Consensus	93	186	50 737	0.006	1
Stable consensus	68	49	94 854	0.005	0.54

 # genes: number of genes selected out of those that are part of the PPIN; for SConES GI the total number of genes, including RNA genes, was added in parentheses. $\overline{\textbf{Betweenness}}$: mean betweenness of the selected genes in the PPIN. $\hat{\textbf{P}}_{\textbf{gene}}$: median P-value of the selected genes. $\rho_{\textbf{consensus}}$: Pearson correlation between the subnetwork and the consensus network.

also differ, as measured by the median centrality (Table 1) and the number of connected components (Figure 1D). Only two methods have more than one connected component: SConES, as described above, and HotNet2. HotNet2 produced 135 subnetworks, 115 of which have less than five genes. The second largest subnetwork (13 nodes) contains the two breast cancer susceptibility genes *CASP8* and *BLM*. Lastly, a pathway enrichment analysis (Section 4.4) also showed similarities and differences in the underlying mechanisms. It linked different parts of SigMod's solution network to four processes: protein translation (including mitochondrial), mRNA splicing, protein misfolding, and keratinization (BH adjusted P-values < 0.03). Interestingly, dmGWAS solution is also related to protein misfolding (*attenuation phase*, BH adjusted P-value $= 0.01$). But, additionally, it includes submodules of proteins related to mitosis, DNA damage, and regulation of TP53 (BH adjusted P-values < 0.05), which match previously known mechanisms of breast cancer susceptibility [24]. As with SigMod, the genes in HotNet2's solution are involved in mitochondrial translation (BH adjusted P-value $= 1.87 \times 10^{-4}$), but also in glycogen metabolism and transcription of nuclear receptors (BH adjusted P-value < 0.04).

Despite their differences, there are additional common themes. All obtained solution subnetworks have lower association P-values than the whole PPIN (median P-value $\ll 0.46$, Table 1), despite containing genes with higher P-values as well (Figure 1A). This exemplifies the trade-off between statistical significance and biological relevance. However, there are nuances between solutions in this regard: heinz strongly favored genes with lower P-values, while dmGWAS was less conservative (median P-values 0.0012 and 0.19, respectively); SConES tended to select whole LD-blocks;

and HotNet2 and SigMod were less likely to select low scoring genes. Additionally, the solution subnetworks presented other desirable properties. First, five of them were enriched in known breast cancer susceptibility genes (consensus, dmGWAS, heinz, HotNet2, and SigMod, Fisher's exact test one-sided P-value < 0.03). Second, the genes in four solution subnetworks displayed on average a higher betweenness centrality than the rest of the genes, a difference that is significant in four solutions (consensus, dmGWAS, HotNet2, and SigMod, Wilcoxon rank-sum test P-value $< 1.4 \times 10^{-21}$). This agrees with the notion that disease genes are more central than other non-essential genes [9], an observation that holds in breast cancer (one-tailed Wilcoxon rank-sum test P-value $= 2.64 \times 10^{-5}$ when comparing the betweenness of known susceptibility genes versus the rest). Interestingly, SConES selected SNPs that are also more central than the average SNP (Supplementary table 1), suggesting that causal SNPs are also more central than non associated SNPs.

2.3 A case study: the consensus network

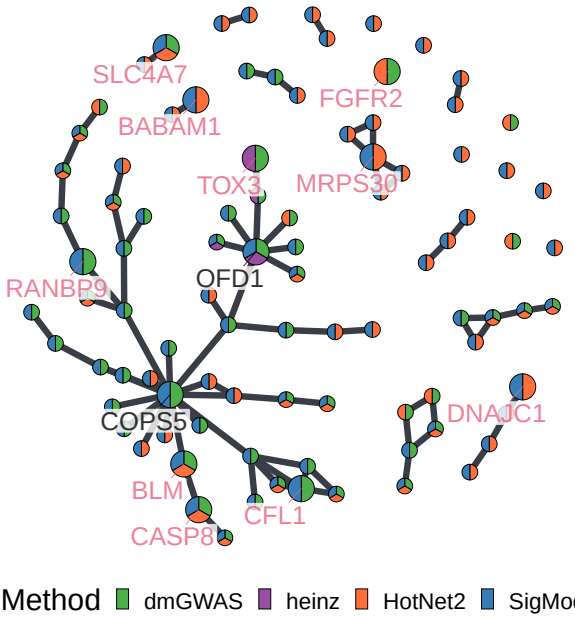

Fig. 2: Consensus subnetwork on GENESIS (Section 4.3.4). Each node is represented by a pie chart, which shows the methods that selected it. We labeled (and enlarged) the two most central genes (*COPS5* and *OFD1*) and those genes that are known breast cancer susceptibility genes and/or significantly associated with breast cancer susceptibility in the BCAC dataset. The latter ones are also colored in pink. All gene names are indicated in Supplementary Figure 3.

Despite the heterogeneity of the solutions, their shared properties suggest that each method captures different aspects of cancer susceptibility. Indeed, only 20 genes are common to more than two solutions (Supplementary Figure 4A), but encouragingly, the more methods selected a gene, the higher its association score to the phenotype (Supplementary Figure 4B). To leverage on their strengths and compensate their respective weaknesses, we built a consensus subnetwork that captures the mechanisms most shared among the solution subnetworks (Section 4.3.4). This subnetwork (Figure 2) contains 93 genes and exhibits the aforementioned properties of the individual solutions: enrichment in breast cancer susceptibility genes and higher betweenness centrality than the rest of the genes.

A pathway enrichment analysis of the genes in the consensus network also shows similar pathways as the individual solutions. We found two involved mechanisms: *mitochondrial translation* and *attenuation phase*. The former is supported by genes like *MRPS30* (VEGAS2 P-value = 0.001), which encode a mitochondrial ribosomal protein and was also linked to breast cancer susceptibility [25]. Interestingly, increased mitochondrial translation has been found in cancer cells [26], and its inhibition proposed as a therapeutic target. With regards to attenuation phase of heat shock response, it involves three Hsp70 chaperones: HSPA1A, HSPA1B, and HSPA1L. The genes encoding these proteins are all near each other at 6p21, in the region known as HLA. In fact, out of the 22 SNPs that map to any of these three genes, 9 map to all of them, and 4 to two, making hard to disentangle their effects. *HSPA1A* was the most strongly associated gene (VEGAS2 P-value = 8.37×10^{-4}).

Topologically the consensus consists of a connected component composed of 49 genes, and multiple smaller subnetworks. Among the latter, 19 genes are in subnetworks containing a single gene or two connected nodes, implying that they do not have a consistently altered neighborhood, but are strongly associated themselves and hence picked by at least two methods. The opposite would be the case of highly central genes in the PPIN, a property which is weakly anti-correlated with the P-value of association to the disease (Pearson correlation coefficient = -0.26, Supplementary Figure 4D). This suggests that they were selected because they were on the shortest path between two highly associated genes. In view of this, we hypothesize that highly central genes might contribute to the heritability through alterations of their neighborhood, consistently with the omnigenic model of disease [6]. For instance, the most central node in the consensus network is COPS5, a component of the COP9 signalosome which regulates multiple signalling pathways. *COPS5* is related to multiple hallmarks of cancer and is overexpressed in multiple tumors, including breast and ovarian cancer [27]. Despite its lack of association in GENESIS or BCAC (VEGAS2 P-value of 0.22 and 0.14 respectively), its neighbors in the consensus subnetwork have consistently low P-values (median VEGAS2 P-value = 0.006).

2.4 Network methods boost biomarker discovery

We compared the results of different network methods to the European sample of the Breast Cancer Association Consortium (BCAC) [19], the largest GWAS to date on breast cancer (Section 4.5.3).

Although BCAC case-control studies do not necessarily target cases with a family history of breast cancer, this comparison is pertinent since we expect a shared genetic architecture at the gene level, at which most network metods operate. This shared genetic architecture, together with BCAC's scale (90 times more samples than GENESIS) provides a reasonable counterfactual of what we would expect if GENESIS had a larger sample size. We computed a gene association score on BCAC, in an equivalent way to the one described in Section 4.3.1. The solutions provided by the different network approaches overlap significantly with BCAC findings (Fisher's exact test P-value < 0.019). The gene-based network methods achieve comparable precision (2%-25%) and recall (1.3-12.1%) at recovering BCAC-significant genes (Figure 1C). Interestingly, while SConES GI at the SNP-level achieves a similar recall (8.6%), it shows a much higher precision (47.3%).

2.5 Network methods share limitations

We compared the six network methods in a 5-fold subsampling setting (Section 4.5). Specifically, we measured five properties (Figure 3): size of the solution subnetwork; sensitivity and specificity of an L1-penalized logistic regression classifier on the selected SNPs; stability; and computational runtime. The solution size varies greatly between the different methods (Figure 3A). Heinz produced the smallest solutions, with an average of 182 selected SNPs. The largest solutions came from SConES GI (6 256.6 SNPs), and dmGWAS (4 255.0 SNPs). To determine whether the selected SNPs could be used for patient classification, we computed the performance of the classifier on the *test dataset* (Figure 3B). The different classifiers displayed similarly poor sensitivities and specificities, all in the $0.52 - 0.56$ range. Interestingly, the classifier trained on all the SNPs had a similar performance, despite being the only method aiming only at minimizing prediction error. It should be considered that, although these performances are low, we do not expect to separate cases from controls well using exclusively genetic data.

Another desirable quality of an algorithm is the stability of the solution with regards to small changes in the input (Section 4.5.1). Both heinz and LEAN displayed a high stability in our benchmark, consistently selecting the same genes and no genes over the 5 subsamples, respectively (Figure 3C). Conversely, the other methods displayed similarly low stabilities.

In terms of computational runtime, the fastest method was heinz (Figure 3D), which returned a solution in a few seconds. HotNet2 was the slowest (3 days and 14 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, it took 5 days on average for HotNet2 to produce a result.

2.6 Network topology matters, and might lead to ambiguous results

As shown above, and despite their similarities, the network methods produced remarkably different solutions. This is due to the particularities of each methods, and directly or indirectly provide information about the dataset. For instance, the fact that LEAN did not return any biomarker

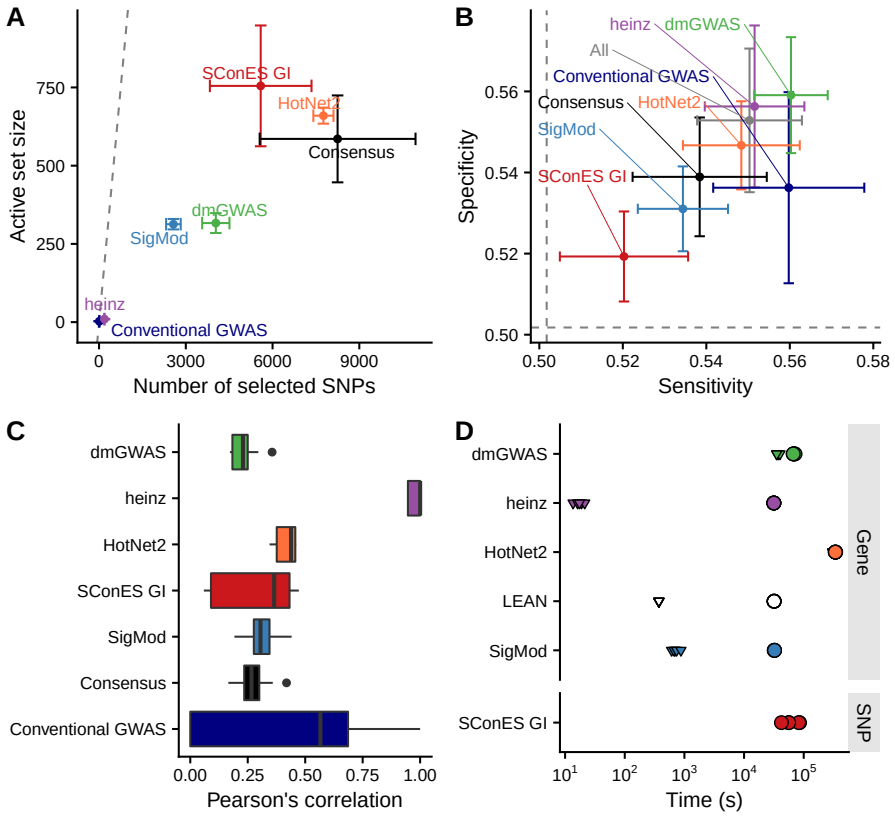

Fig. 3: Comparison of network-based GWAS methods on GENESIS. Each method was run 5 times on a random subset containing 80% of the samples, and tested on the remaining samples (Section 4.5). As LEAN did not select any gene, it was excluded from all panels except **D**. **(A)** Number of SNPs selected by each method and number of SNPs in the active set found by the classifier. Points are the average over the 5 runs; lines represent the standard error of the mean. A grey diagonal line with slope 1 is added for comparison. For reference, the active set of Lasso using all the SNPs included, on average, 154 117.4 SNPs. **(B)** Sensitivity and specificity on test 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 Pearson correlations of the solutions used by different methods. A Pearson correlation of 1 means the two solutions are the same. A Pearson correlation 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). For gene network-based methods, inverted triangles represent the runtime of the algorithm itself, and circles the total time, which includes the algorithm themselves and the additional 119 980 seconds (1 day and 9.33 hours) that VEGAS2 took on average to compute the gene scores from SNP summary statistics.

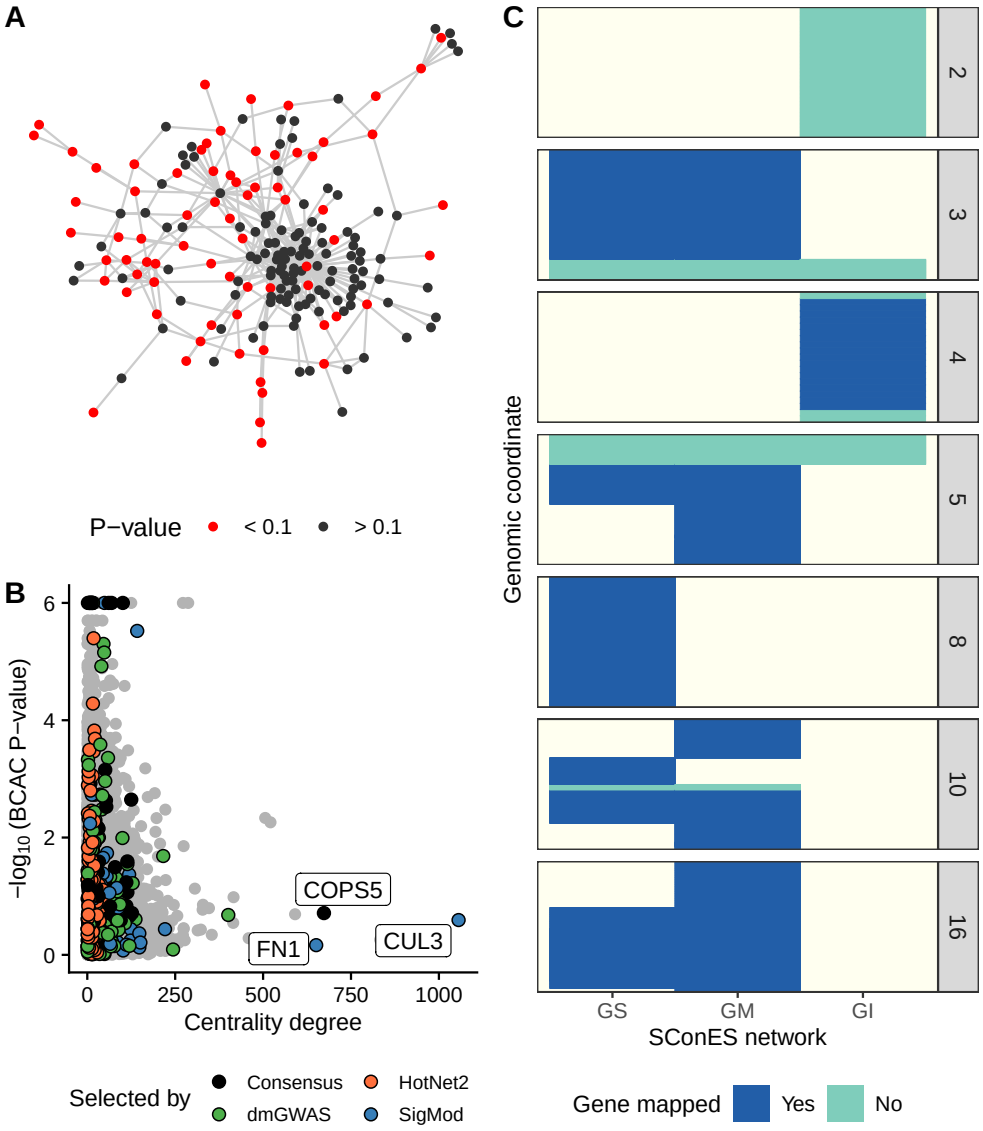

Fig. 4: Drawbacks encountered when using network guided methods. **(A)** DmGWAS solution subnetwork. Genes with a P-value < 0.1 are highlighted in red. **(B)** Centrality degree and -$\log_{10}$ of the VEGAS2 P-value in BCAC for each of the nodes in the PPIN. We highlighted the genes selected by each method, and the ones selected by more than one ("Consensus"). We labeled the three most central genes that were picked by any method. **(C)** Overlap between the solutions of SConES GS, GM or GI in the different genomic regions. SNPs that were not selected in the studied network, but were selected in another one, are displayed in background color.

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 1). Due to this parsimonious and highly associated solution, it was the best method to stably select a set of biomarkers (Figure 3C). Its conservativeness stems from its preprocessing step, which models the gene P-values as a mixture model of a beta distribution and a uniform distribution, controlled by an FDR parameter. Due to the limited signal at the gene level in this dataset (Supplementary Figure 1B), only 36 of all the genes retain a positive score after that transformation. Yet, this small solution does not provide much insight into the susceptibility mechanisms to cancer. Importantly, it ignores genes that are associated to cancer in this dataset like *FGFR2*.

On the other end of the spectrum, dmGWAS, HotNet2, and SigMod produced large solutions. dmGWAS' subnetwork is the least associated subnetwork on average. This is due to the greedy framework it uses, which has a bias for larger solutions [28]. It considered all nodes at distance 2 of the examined subnetwork, and accepted a weakly associated genes if it was linked to another, strongly associated one. This is exacerbated when the results of successive greedy searches are aggregated, leading to a large, tightly connected cluster of unassociated genes (Figure 4A). This relatively low signal-to-noise ratio combined with the large solution requires additional analyses to draw conclusions, such as enrichment analyses. In the same line, HotNet2's subnetwork is even harder to interpret, being composed of 440 genes divided into 135 subnetworks. Lastly, SigMod misses some of the most strongly associated, breast cancer susceptibility genes in the dataset, like *FGFR2* and *TOX3*.

Another peculiarity of network methods is their relationship to degree centrality. On the one hand we observed that highly central genes often had no association to disease (Figure 4B). On the other, network methods favor central genes, as they often connect high scoring nodes. This was specially the case of SigMod, which selected three highly central, unassociated genes: *COPS5*, *CUL3* and *FN1*. As we showed in Section 2.3, and will show in 2.7, there is evidence in the literature of the contribution of the first two to breast cancer susceptibility. With regards to *FN1*, it encodes a fibronectin, a protein of the extracellular matrix involved in cell adhesion and migration. Overexpression of *FN1* has been observed in breast cancer [29], and it is negatively correlated with poor prognosis in other cancer types [30, 31].

By virtue of using a SNP subnetwork, SConES analyzes each SNP in their functional context. It therefore can select SNPs in genes without any associated interactor, as well as SNPs in non-coding regions or in non-interacting genes. In fact, due to linkage disequilibrium, SConES favors such genes, as selecting SNPs in an LD-block which overlaps with a gene favors selecting the rest of the gene. This might explain why SConES produces similar results on the GS and GM networks, heavily affected by linkage disequilibrium (Supplementary Figure 2). On the other hand, SConES penalizes selecting SNPs and not their neighbors. This makes it conservative regarding SNPs with many interactions, like those mapped to hub genes in the PPIN. For this reason, SConES GI did not select any protein coding gene, despite selecting similar regions as SConES GS (Figure 4C). In fact

SConES GS and SConES GM select regions related to breast cancer, like 3p24 (*SLC4A7*/*NEK10* [32]), 5p12 (*FGF10*, *MRPS30* [25]), 10q26 (*FGFR2*), and 16q12 (*TOX3*). On top of those SConES GS selects region 8q24 (*POU5F1B* [33]). We hypothesize that the lack of results on the PPIN network of SConES GI and LEAN are due to the same cause: the absence of joint association of a gene and a majority of its neighbors. Although in the case of SConES other hyperparameters could lead to a more informative solution (e.g. a lower λ in Equation 1), it is unclear what the best strategy to find them is. In addition, due to the design of the iCOGS array, the genome of GENESIS participants has not been unbiasedly surveyed: some regions are fine-mapped — which might distort gene structure in GM and GI networks — while others are under studied — hindering the accuracy with which the GS network captures the genome structure.

2.7 Adjusting for instability preserves global network properties

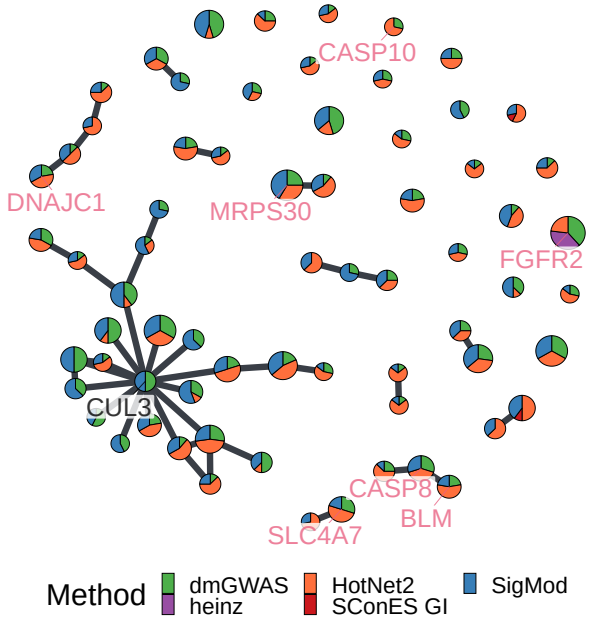

Fig. 5: Stable consensus subnetwork on GENESIS (Section 2.7). Each node is represented by a pie chart, which shows the methods that selected it. We labeled (and enlarged) the most central genes (*CUL3*) and those genes that are known breast cancer susceptibility genes and/or significantly associated with breast cancer susceptibility in the BCAC dataset. The latter ones are also colored in pink. All gene names are indicated in Supplementary Figure 5.

Most of the network methods, including the consensus, were highly unstable, raising questions

about the reliability of the results. We built a new, stable consensus network using the genes selected most often across the 30 solutions obtained by running the 6 methods on 5 different splits of the data (Section 4.5). Such a network is expected to capture the subnetworks more often found altered, and hence should be more resistant to noise. We used only genes selected in at least 7 of the solutions, which corresponded to 1% of all genes selected at least once. The resulting stablility-based consensus was composed of 68 genes (Figure 5). This network shares most of the proterties of the consensus: breast cancer susceptibility genes are overrepresented (P-value = 3×10^{-4}), as well as genes involved in mitochondrial translation and the attenuation phase (adjusted P-values 0.001 and 3×10^{-5} respectively); the selected genes are more central than average (P-value = 1.1×10^{-14}); and a considerable number of nodes (19) are isolated.

However, although this new network exhibits similar global properties as the previous one, the lack of stability results in different genes being selected. In this case, the most central gene is *CUL3*, which is absent from the previous consensus network and has a low association score in both GENESIS and BCAC (P-values of 0.04 and 0.26, respectively). This gene is a component of Cullin-RING ubiquitin ligases. Encouragingly, it impacts the protein levels of multiple genes relevant for cancer progression [34], and its overexpression was also linked to increased sensitivity to carcinogens [35].

3 Discussion

In recent years, the ability of GWAS to unravel the mechanisms leading to complex diseases has been called into question [6]. On the one hand, the omnigenic model proposes that gene functions are interwoven with each other in a dense co-function network. The practical consequence is that larger and larger GWAS will lead to the discovery of an uninformative wide-spread pleiotropy. On the other hand, discovery in GWAS is hindered by a conservative statistical framework. Network methods tackle these two issues by using both the association score and an interaction network to take into consideration the biological context of each of the genes and SNPs. Based on what could be considered diverse interpretations of the omnigenic model, several methods for network-guided biomarker discovery have been proposed in recent years. In this article we evaluated the relevance of six of them by examining a GWAS dataset on familial breast cancer.

Most of the network methods produced a relevant subset of biomarkers, recapitulating known breast cancer susceptibility genes. In general, the selected genes and SNPs were more central than average, in accordance with the observation that disease genes are more relatively central [9]. However, very central nodes are also more likely to be connecting any given random pair of nodes, making them more likely to be selected by these network methods. Across this article we show that highly central genes that were selected (*COPS5*, *CUL3* and *FN1*) could plausibly be involved in breast cancer susceptibility. Yet, further work is needed to characterize the impact of centrality on network methods' outputs. Despite these similarities, the solutions were notably different. At one end of the spectrum, SConES and heinz preferred small, highly associated solutions, providing a

conveniently short list of biomarkers, at the expense of not shedding much light on the etiology of the disease. On the other end, SigMod and dmGWAS gravitate towards larger, less associated solutions which provide a wide overview of the biological context. While this deepens our understanding of the disease and provide biological hypotheses, they require further analyses, which might deter unexperienced practitioners. HotNet2 balances both approaches at the expense of producing the largest solution: a constellation of many, highly associated, small subnetworks. Additionally, all solutions share two drawbacks. First, they are all equally bad at discriminating cases from controls. Yet, the classification accuracy of network methods is similar to that of a machine learning classifier, which suggests that cases and controls are difficult to separate in this dataset. This might be due to unaccounted for environmental factors, and limited statistical power, which reduces the ability to identify relevant SNPs. Second, all methods are remarkably unstable, yielding different solutions for slightly different inputs. This might partly be caused by the instability of the P-values themselves in low statistical power settings [36]. Hence, heinz's conservative transformation of P-values, which favors only the most extreme ones, leads to improved stability. Another source of instability might be the redundancy inherent to biological networks.

To overcome the limitations of the individual methods while exploiting their strengths, we proposed combining them into a consensus subnetwork. We use a straightforward strategy of including any node that was recovered by multiple methods. We proposed two networks: a consensus network that tackled the heterogeneity of the solutions in the full dataset, and a stable consensus network, that addressed the instability of the methods. They both synthesized the altered mechanisms: they both included the majority of the strongly associated smaller solutions and captured genes and broader mechanisms related to cancer. Thanks to their smaller size and their network structure, they provided compelling hypotheses on genes like *COPS5* and *CUL3*, which lack genome-wide association with the disease, but who are related to cancer at the expression level and whose neighborhood has consistent high association scores. Crucially, the consensus was as unstable as the tested methods, while the stable consensus shared these properties while accounting for instability. This supports that instability might be caused by redundant but equivalent biological mechanisms, and hence validates the conclusions obtained on the individual solutions and the consensus.

The strength of network-based analyses comes from leveraging prior knowledge to boost discovery. In consequence, they show their shortcomings with respect to understudied genes, especially those not in the network. Out of the 32 767 genes that we can map the genotyped SNPs to, 60.7% (19 887) are not in the protein-protein interaction network. The majority of those (14 660) are non-coding genes, mainly lncRNA, miRNA, and snRNA (Supplementary Figure 6). Yet, RNA genes like *CASC16* are associated to breast cancer (Section 2.1), reminding us of the importance of using networks beyond coding genes. In addition, even protein-coding genes linked to breast cancer susceptibility [32], like *NEK10* (P-value 1.6×10^{-5}, located near *SLC4A7*) or *POU5F1B*, were absent from the network. However, on average protein-coding genes absent from the PPIN are less associated with this phenotype (Wilcoxon rank-sum P-value = 2.79×10^{-8}, median P-values of 0.43 and 0.47). As we are using interactions from high-throughput experiments, such difference cannot be due to well-known genes having more known interactions. As disease genes tend to be

more central [9], we hypothesize that it is due to interactions between central genes being more likely. It is worth noting that network approaches that do not use PPIs, like SConES GS and GM, did recover SNPs in *NEK10* and *CASC16*. Moreover, both SConES GM and GI recovered intergenic regions, which might contain key regulatory elements [37] and, yet, are excluded from gene-centric approaches. This shows the potential of SNP networks, in which SNPs are linked when there is evidence of co-function, to perform network-guided GWAS even in the absence of gene-level interactions. Lastly, all the methods are heavily affected by how SNPs are mapped to genes. In Section 2.3 we highlight ambiguities that appear when genes overlap or are in linkage disequilibrium. In fact, the presented case is paradigmatic, since the genes are in the most gene-dense region of the genome [38]. Network methods are prone to selecting such genes when they are functionally related, and hence interconnected in the network, but might be more resilient to them when the overlapping genes are unrelated. Making use of more targeted mappings of SNPs to genes (e.g. eQTLs, SNPs associated to the expression of a gene), altogether with a stringent LD pruning, might address such problems.

As not all databases compile the same interactions, the choice of the PPIN determines the final output. In this work we used exclusively interactions from HINT from high-throughput experiments. This responds to concerns about adding interactions identified in targeted studies and prone to a "rich getting richer" phenomenon: popular genes have a higher proportion of their interactions described [39, 40], and they might bias discovery by reducing the average shortest path length between two random nodes. On the other hand, Huang et al. [10] found that the best predictor of the performance of a network for disease gene discovery is the size of the network, which supports using the largest amount of interactions. When we compared the impact of using a larger network containing interactions from both high-throughput experiments and the literature (Section 4.3.3), we found that for most of the methods it did not greatly change the size or the stability of the solution, the classification accuracy, or the runtime (Supplementary Figure 7). This supports using only interactions from high-throughput experiments, which produces apparently similar solutions and avoids falling into "circular reasonings", where the best known genes are artificially pushed into the solutions.

A crucial step for the gene based methods is the computation of the gene score. In this work we used VEGAS2 [41] due to the flexibility it offers to use user-specified gene annotations. However, it presents known problems (selection of an appropriate percentage of top SNPs, long runtimes and P-value precision limited to the number of permutations [42]), and other algorithms [42, 43, 44] might have more statistical power. Another important decision is how to handle LD in a GWAS. VEGAS2 accounts for LD patterns, and hence an LD pruning step would not impact gene-based network methods, although it would speed up VEGAS2's computation time. With regards to SConES, fewer SNPs would lead to simpler SNP networks and, possibly, shorter runtimes. However, as mentioned in Section 2.6, LD patterns seem paramount to SConES' solutions, and an LD pruning step could potentially alter them.

In order to produce the consensus networks, we faced the different interfaces, preprocessing steps, and unexpected behaviors of the various methods. To facilitate that other authors apply them to

new datasets and aggregate their solutions, we built six nextflow pipelines [45] with a consistent interface and, whenever possible, parallelized 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, which is available on Docker Hub hclimente/gwas-tools.

4 Materials and methods

4.1 GENESIS

The GENE Sisters (GENESIS) study was designed to investigate risk factors for familial breast cancer in the French population [18]. Index cases are patients with infiltrating mammary or ductal adenocarcinoma, who had a sister with breast cancer, and who have been tested negative for *BRCA1* and *BRCA2* pathogenic variants. Controls are unaffected colleagues and/or friends of the cases, born around the year of birth of their corresponding case ($\pm$ 3 years). We focused on the 2 577 samples of European ancestry, of which 1 279 are controls and 1 298 are cases. The genotyping was performed using the iCOGS array, a custom Illumina array designed to study genetic susceptibility of hormone-related cancers [46]. It contains 211 155 SNPs, including SNPs putatively associated with breast, ovarian, and prostate cancers, SNPs associated with survival after diagnosis, and SNPs associated to other cancer-related traits, as well as candidate functional variants in selected genes and pathways.

4.2 Preprocessing and quality control

We discarded SNPs with a minor allele frequency lower than 0.1%, those not in Hardy–Weinberg equilibrium in controls (P-value < 0.001), and those with genotyping data missing on more than 10% of the samples. A subset of 20 duplicated SNPs in *FGFR2* were also removed. In addition, we removed the samples with more than 10% missing genotypes. After controlling for relatedness, 17 additional samples were removed (6 for sample identity error, 6 controls related to other samples, 2 cases being related to an index case, and 3 additional controls having a high relatedness score). Lastly, based on study selection criteria, 11 other samples were removed (1 control having cancer, 4 index cases with no affected sister, 3 half-sisters, 1 sister with lobular carcinoma *in situ*, 1 with a *BRCA1* or *BRCA2* pathogenic variant detected in the family, 1 with unknown molecular diagnosis). The final dataset included 1 271 controls and 1 280 cases, genotyped over 197 083 SNPs.

We looked for population structure that could produce spurious associations. A principal component analysis revealed no visual differential population structure between cases and controls (Supplementary Figure 9). Independently, we did not find evidence of genomic inflation ($\lambda = 1.05$) either, further confirming the absence of confounding population structure.

4.3 High-score subnetwork search algorithms

4.3.1 SNP and gene association

To measure association between a genotype and the phenotype, we performed a per-SNP 1 d.f. χ^2 allelic test using PLINK v1.90 [47]. Then, we used VEGAS2 [41] to compute the gene-level association score from the P-values of the SNPs mapped to them. Specifically, for each gene we only used the 10% of SNPs mapped to it with lowest P-values. We mapped SNPs to genes through their genomic coordinates: all SNPs located within the boundaries of a gene, ±50 kb, were mapped to that gene. We used the 62 193 genes described in GENCODE 31 [48], although only 54 612 could be mapped to at least one SNP. Out of those, we focused exclusively on the 32 767 that had a gene symbol. Out of the 197 083 SNPs remaining after quality control, 164 037 were mapped to at least one of these genes.

We used such mapping to compare the outputs of methods that produce SNP-lists to those that produce gene-lists, and vice versa. For the former, we considered any gene that can be mapped to any of the selected SNPs as selected as well. For the latter, we considered all the SNPs that can be mapped to that gene as selected by the method.

4.3.2 Mathematical notations

In this article, we use undirected, vertex-weighted networks, or graphs, $G = (V, E, w)$. $V = \{v_1, \ldots, v_n\}$ refers to the vertices, with weights $w : V \to \mathbb{R}$. 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 to a weight, nodes have other properties, provided by the topology of the graph. In this article we focus on two of those: degree centrality, and betweenness centrality. The degree centrality, or degree, is the number of edges that a node has. The betweenness centrality, or betweenness, is the number of times a node participates in the shortest paths between two other nodes.

In addition, we use two 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. The element (i, j) represents a selected relationship between v_i and v_j. The *adjacency matrix* W_G contains a 1 when the corresponding nodes are connected, and 0 otherwise; the diagonal is zero. The *degree matrix* D_G is a diagonal matrix which contains the degree of the different nodes.

4.3.3 Networks

Gene network The statistical frameworks of the different network methods are compatible with any type of network (protein interactions, gene coexpression, regulatory, etc.). Yet, we used protein-protein interaction networks (PPIN) for all of them except SConES, as they are interpretable, well

characterized, and they were designed to run efficiently on networks of their size. We built our PPIN from both binary and co-complex interactions stored in the HINT database (release April 2019) [40]. Unless otherwise specified, we used only interactions coming from high-throughput experiments, leaving out targeted studies that might bias the topology of the network. Out of the 146 722 interactions from high-throughput experiments that HINT stores, we were able to map 142 541 to a pair of gene symbols, involving 13 619 genes. 12 880 of those mapped to a genotyped SNP after quality control, involving 127 604 interactions. The scoring function for the nodes changed from method to method (Section 4.3.4).

Additionally, we compared the results of the aforementioned PPIN with those obtained on another PPIN built using interactions coming from both high-throughput and targeted studies. In that case, out of the 179 332 interactions in HINT, 173 797 mapped to a pair of gene symbols. Out of those, 13 735 mapped to a genotyped SNP after quality control, involving 156 190 interactions.

SNP networks SConES [16] is the only network method designed to handle SNP networks. As in gene networks, two SNPs are connected in a SNP network when there is evidence of shared functionality between two SNPs. Azencott et al. [16] proposed three ways of building these networks: connecting the 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, on top of GM ("GI network"). We focused on the GI network, as it fits the scope of this work better, using the PPIN described above. However, at different stages of this work we also used GS and GM for comparison. For the GM network, we used the mapping described in Section 4.3.1. In all three the node scores are the association scores of the individual SNPs with the phenotype (1 d.f. χ^2). The properties of these three subnetworks are available in Supplementary Table 1.

4.3.4 Network methods

Genes that contribute to the same function are nearby in the PPIN, and can be topologically related to each other in diverse ways (densely interconnected modules, nodes around a hub, a path, etc.). But this is not the only aspect to model when developing a network method: how to score the nodes, whether the affected mechanisms form a single connected component or several, how to frame the problem in a computationally efficient fashion, which network to use, etc. Unsurprisingly, multiple solutions have been proposed. We examined six of them: five that explore the PPIN, and one which explores SNP networks. We selected methods that were open source, had an implementation available, and an accessible documentation. Their main differences are summarized in Table 2.

dmGWAS dmGWAS seeks the subgraph with the highest local density in low P-values [13]. To that end it searches candidate subnetwork solutions using a greedy, "seed and extend", heuristic:

1. Select a seed node i and form the subnetwork $S_i = \{i\}$.

Table 2: Summary of the differences between the network methods.

Method	Field	Nodes	Exhaustive	Solution	Comp.	Input	Scoring	Ref.
dmGWAS	GWAS	Genes	No	-	1	Summary	-$\log_{10}$(P)	[13]
heinz	Omics	Genes	Yes	-	1	Summary	BUM	[14]
HotNet2	Omics	Genes	Yes	Module	≥ 1	Summary	Local FDR	[15]
LEAN	Omics	Genes	Yes	Star	≥ 1	Summary	-$\log_{10}$(P)	[12]
SConES	GWAS	SNPs	Yes	Module	≥ 1	Genotypes	1 d.f. χ^2	[16]
SigMod	GWAS	Genes	Yes	Module	≥ 1	Summary	Φ^{-1}(1 - P)	[17]

Field: field in which the algorithm was developed. ***Nodes***: the type of nodes in the network, either genes (PPIN) or SNPs. ***Exhaustive***: whether all the possible solutions given the selected hyperparameters are explored. ***Solution***: additional properties are enforced on the solution subnetwork, other than containing high scoring, connected nodes. ***Comp.***: number of connected components in the solution. ***Input***: genotype data or GWAS summary statistics. ***Scoring***: how SNP/gene P-values are transformed into node scores. In the case of heinz, BUM stands for beta-uniform model, used to transform the P-values; for SigMod, Φ^{-1} represents the inverse of the cumulative distribution function of the standard Normal distribution. ***Ref.***: original publication featuring the algorithm.

2. Compute Stouffer's Z-score $\mathrm{Z_m}$ for S_i as

$$Z_m = \frac{1}{\sqrt{k}} \sum_{j \in S_i} z_j,$$

   where k is the number of genes in S_i; $\mathrm{z_j}$ is the Z score of gene j, computed as ϕ^{-1}(1 - $\mathrm{P\text{-}value_j}$); and ϕ^{-1} is the inverse normal distribution function.

3. Identify neighboring nodes of S_i, i.e. nodes at distance $\leq$ d.

4. Add the neighboring nodes whose inclusion increases the $\mathrm{Z_{m+1}}$ more than a threshold $\mathrm{Z_m} \times (1 + \mathrm{r})$.

5. Repeat 2-4 until no further enlargement is possible.

6. Add S_i to the list of subnetworks to return. Its Z-score is normalized as

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

   where $\mathrm{Z_m}(\pi)$ represents a vector containing 100 000 random subsets of the same number of genes.

DmGWAS carries out this process on every gene in the network. We used the implementation of dmGWAS in the dmGWAS 3.0 R package [49]. We used the suggested hyperparameters $d = 2$ and $r = 0.1$. We used the function *simpleChoose* to select the solution subnetwork, which aggregates the top 1% subnetworks.

heinz The goal of heinz is to identify the highest-scored connected subnetwork [14]. The authors propose a transformation of the genes' P-value into a score that is negative under no association with the phenotype, and positive when there is. This transformation is achieved by modelling the distribution of P-values by a beta-uniform model (BUM) parameterized by the desired false discovery rate (FDR). Thus formulated, the problem is NP-complete, and hence solving it would require a prohibitively long computational time. To solve it efficiently it is re-cast as the Prize-Collecting Steiner Tree Problem (PCST), which seeks to select the connected subnetwork S that maximizes the *profit* p(S), defined as:

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

were $p(v) = w(v) - w'$ is the *profit* of adding a node, $c(e) = w'$ is the *cost* of adding an edge, and $w' = min_{v \in V_G} w(v)$ is the smallest node weight of G. All three are positive quantities. Heinz implements the algorithm from Ljubić et al. [50] which, in practice is often fast and optimal, although neither is guaranteed. We used BioNet's implementation of heinz [51, 52].

HotNet2 HotNet2 was developed to find connected subgraphs of genes frequently mutated in cancer [15]. 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: at initialization, the score of the node determines its initial heat; iteratively each node yields heat to its "colder" neighbors, and receives heat from its "hotter" neighbors, while retaining part of its own (hence, *insulated*). This process continues until equilibrium is reached, and results in a diffusion matrix F. F is used to compute the similarity matrix E that models exchanged heat as

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

where $\mathrm{diag}(w(V))$ is a diagonal matrix with the node scores in its diagonal. For any two nodes i and j, E_{ij} models the amount of heat that diffuses from node j to node i, which can be interpreted as a (non-symmetric) similarity between those two nodes. To obtain densely connected subnetworks, HotNet2 prunes E, only preserving edges such that $w(E) > \delta$. Lastly, HotNet2 evaluates the statistical significance of the subnetworks by comparing their size to the size of networks obtained by permuting the node scores. We assigned initial node scores as in Nakka et al. [42], assigning a score of 0 for the genes with low probability of being associated to the disease, and -$\log_{10}$(P-value) to those likely to be. In this dataset, the threshold separating both was a P-value of 0.125, which was obtained using a local FDR approach [53]. HotNet2 has two parameters: the restart probability β, and the threshold heat δ. Both parameters are set automatically by the algorithm, which is robust to their values [15]. HotNet2 is implemented in Python [54].

LEAN LEAN searches altered "star" subnetworks, that is, subnetworks composed by one central node and all its interactors [12]. By imposing this restriction, LEAN is able to exhaustively

test all such subnetworks (one per node). For a particular subnetwork of size m, the P-values corresponding to the involved nodes are ranked as $\mathrm{p_1} \leq \ldots \leq \mathrm{p_m}$. Then, k binomial tests are conducted, to compute the probability of having k out of m P-values lower or equal to $\mathrm{p_k}$ 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 gene sets of the same size 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 [55].

SConES SConES searches the minimal, modular, and maximally associated subnetwork in a SNP graph [16]. 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. The connectivity term penalizes disconnected solutions, with many edges between nodes that are selected and nodes that are not. Given a λ and an η, the aforementioned problem has a unique solution, that SConES finds using a graph min-cut procedure. As in Azencott et al. [16], we selected λ and η by cross-validation, choosing the values that produce the most stable solution across folds. 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. We used the version on SConES implemented in the R package martini [56].

SigMod SigMod aims at identifying the highest-scoring, most densely connected gene subnetwork [17]. It addresses an optimization problem similar to that of SConES (Equation 1), but the connectivity term encourages connected solutions by favoring solutions where many edges connect two selected nodes, rather than penalizing disconnected ones.

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

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

SigMod presents three important differences with SConES. First it is designed for gene-gene networks. Second, it favors subnetworks containing many edges. SConES, instead, penalizes connections between the selected and unselected nodes. Third, it explores the grid of hyper-parameters differently, and processes their respective solutions. Specifically, for the range of $\lambda = \lambda_{\min}, \ldots, \lambda_{\max}$ 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 densely 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 [57].

Consensus We built a consensus network by retaining the nodes that were selected by at least two of the six methods (using SConES GI for SConES).

4.4 Pathway enrichment analysis

We searched for pathways enriched in the gene subnetworks produced by the above methods. We conducted an hypergeometric test on pathways from Reactome [58] using the R package ReactomePA [59]. The universe of genes included any gene that we could map to a SNP in the iCOGS array (Section 4.3.1). We adjusted the P-values for multiple testing as in Benjamini and Hochberg **(author?)** [60] (BH). Pathways with an BH adjusted P-value < 0.05 were deemed significant. As the significant pathways are often overlapping and redundant, the results were manually curated afterwards.

4.5 Evaluation of methods

We evaluated multiple properties (described below) of the different methods through a 5-fold subsampling setting. We applied each method to 5 random subsets of the original dataset containing 80% of the samples (*train set*). When pertinent, we evaluated the solution on the remaining 20% (*test set*). We used the 5 repetitions to estimate the average and the standard deviation of the different measures.

4.5.1 Properties ot the solution

We compared the runtime, the number of selected genes/SNPs, and the stability (sensitivity of the result to small changes in the input, here, using different *train* sets) of the different network methods. The stability was quantified using the Pearson correlation to measure the overlap between different runs as suggested by Nogueira and Brown [61].

4.5.2 Classification accuracy of selected SNPs

A desirable solution offers good predictive power on unseen *test* samples. We evaluated the predicting power of the SNPs selected by the different methods through the performance of an L1-penalized logistic regression classifier, a machine learning algorithm that searches a small subset of SNPs which provide good classification accuracy. We trained the classifier exclusively on those selected SNPs to predict the outcome (case/control). The L1 penalty helps to account for linkage

disequilibrium by reducing the number of SNPs included in the model (*active set*), while improving the generalization of the classifier. This penalty was set by cross-validation, choosing the value that minimized misclassification error. We applied each network method to each *train* set, and trained the classifier on it as well using only on the selected SNPs. When the method retrieved a list of genes (all of them except SConES), we considered as selected all the SNPs mapped to any of those genes. Then we evaluated the sensitivity and the specificity on the *test set*. The active set gave an estimate of a plausible, more sparse solution with a comparable predictive power to the original solution. To obtain a baseline, we also trained the classifier on all the SNPs. We do not expect a linear model on selected SNPs to be able to separate cases from controls well. Indeed, the lifetime incidence of breast cancer among women with a family history of breast or ovarian cancer, and no *BRCA1/2* mutations, is only 3.9 times more than in the general population [62]. However, classification accuracy may be one additional informative criterion on which to evaluate solutions.

4.5.3 Comparison to state-of-the-art

An alternative way to evaluate the results is comparing our results to an external dataset. For that purpose, we recovered a list of 153 genes associated to familial breast cancer from DisGeNET [63]. Across this article we refer to these genes as *breast cancer susceptibility genes*.

Additionally, we used the summary statistics from the Breast Cancer Association Consortium (BCAC), a meta-analysis of case-control studies conducted in multiple countries which included 13 250 642 SNPs genotyped or imputed on 228 951 women of European ancestry mostly from the general population [19]. Hence, a high proportion of breast cancer cases investigated in BCAC are sporadic (not selected according to family history), while GENESIS is a homogeneous dataset not included in BCAC and which focus on the French high-risk population attending the family cancer clinics. Despite these differences, we expect some degree of shared genetic architecture, especially at the gene level. For that purpose, we searched associated genes as in Section 4.3.1. We provided VEGAS2 with both the summary statistics of all available SNPs, and the genotypes from European samples from the 1000 Genomes Project [64] to compute the LD patterns.

4.6 Code availability

We developed computational pipelines for several steps of GWAS analyses, such as physically mapping SNPs to genes, computing gene scores, and performing six different network analyses. For each of those processes, we created a pipeline with a clear interface that should work on any GWAS dataset. They are compiled in `https://github.com/hclimente/gwas-tools`. Although the GENESIS data is not public, the code to apply the pipelines to this data, as well as the code that reproduces all the analyses in this article are available at `https://github.com/hclimente/genewa`. We deposited all the produced gene subnetworks on NDEx (`http://www.ndexbio.org`), under the UUID e9b0e22a-e9b0-11e9-bb65-0ac135e8bacf.

Author contributions

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

Data curation Christine Lonjou, GENESIS Study collaborators.

Formal Analysis Héctor Climente-González, Christine Lonjou.

Funding acquisition Dominique Stoppa-Lyonnet, Nadine Andrieu, Chloé-Agathe Azencott.

Investigation Héctor Climente-González, Christine Lonjou.

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

Project administration Chloé-Agathe Azencott.

Resources GENESIS Study collaborators, Dominique Stoppa-Lyonnet, Nadine Andrieu.

Software Héctor Climente-González, Christine Lonjou.

Supervision Christine Lonjou, Fabienne Lesueur, Nadine Andrieu, Chloé-Agathe Azencott.

Validation Christine Lonjou, Fabienne Lesueur.

Visualization Héctor Climente-González.

Writing – original draft Héctor Climente-González.

Writing – review & editing Héctor Climente-González, Christine Lonjou, Fabienne Lesueur, Nadine Andrieu, Chloé-Agathe Azencott.

Acknowledgments

We wish to thank Om Kulkarni for helpful discussion on gene-based GWAS and PPIN databases, and the genetic epidemiology platform (the PIGE, Plateforme d'Investigation en Génétique et Epidemiologie), the biological resource centre and all the GENESIS collaborating cancer clinics clinics.

References

[1] Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Computational Biology. 2012 Dec;8(12):e1002822. 00001. Available from: `https://dx.plos.org/10.1371/journal.pcbi.1002822`.

[2] Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research. 2019 Jan;47(D1):D1005–D1012. 00092. Available from: https://academic.oup.com/nar/article/47/D1/D1005/5184712.

[3] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics. 2017 Jul;101(1):5–22. 00634. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929717302409.

[4] Wang MH, Cordell HJ, Van Steen K. Statistical methods for genome-wide association studies. Seminars in Cancer Biology. 2018 May;00001. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1044579X1730278X.

[5] Barton NH, Etheridge AM, Véber A. The infinitesimal model: Definition, derivation, and implications. Theoretical Population Biology. 2017 Dec;118:50–73. 00054. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0040580917300886.

[6] Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun;169(7):1177–1186. 00586. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867417306293.

[7] Furlong LI. Human diseases through the lens of network biology. Trends in Genetics. 2013 Mar;29(3):150–159. 00128. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168952512001886.

[8] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011 Jan;12(1):56–68. 02826. Available from: http://www.nature.com/articles/nrg2918.

[9] Piñero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. Scientific Reports. 2016 Apr;6(1):24570. 00016. Available from: http://www.nature.com/articles/srep24570.

[10] Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Systems. 2018 Apr;6(4):484–495.e5. 00024. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405471218300954.

[11] Azencott CA. Network-Guided Biomarker Discovery. In: Machine Learning for Health Informatics. vol. 9605. Cham: Springer International Publishing; 2016. p. 319–336. 00000. Available from: http://link.springer.com/10.1007/978-3-319-50478-0_16.

[12] Gwinner F, Boulday G, Vandiedonck C, Arnould M, Cardoso C, Nikolayeva I, et al. Network-based analysis of omics data: The LEAN method. Bioinformatics. 2016 Oct;p. btw676. 00007.

[13] 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. 2011 Jan;27(1):95–102. 00205. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq615.

[14] Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. 2008 Jul;24(13):i223–i231. 00429. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn161.

[15] Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics. 2015 Feb;47(2):106–114. 00411. Available from: http://www.nature.com/articles/ng.3168.

[16] Azencott CA, Grimm D, Sugiyama M, Kawahara Y, Borgwardt KM. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics. 2013 Jul;29(13):i171–i179. 00047. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt238.

[17] Liu Y, Brossard M, Roqueiro D, Margaritte-Jeannin P, Sarnowski C, Bouzigon E, et al. SigMod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. Bioinformatics. 2017 Jan;p. btx004. 00007. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx004.

[18] Sinilnikova OM, Dondon MG, Eon-Marchais S, Damiola F, Barjhoux L, Marcou M, et al. GENESIS: a French national resource to study the missing heritability of breast cancer. BMC Cancer. 2016 Dec;16(1):13. 00005. Available from: http://bmccancer.biomedcentral.com/articles/10.1186/s12885-015-2028-9.

[19] Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017 Oct;551(7678):92–94. Available from: https://doi.org/10.1038/nature24284.

[20] Mulligan AM, , Couch FJ, Barrowdale D, Domchek SM, Eccles D, et al. Common breast cancer susceptibility alleles are associated with tumour subtypes in BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2. Breast Cancer Research. 2011 Nov;13(6). Available from: https://doi.org/10.1186/bcr3052.

[21] Rinella ES, Shao Y, Yackowski L, Pramanik S, Oratz R, Schnabel F, et al. Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation. Human Genetics. 2013 May;132(5):523–536. 00019. Available from: http://link.springer.com/10.1007/s00439-013-1269-4.

[22] Brisbin AG, Asmann YW, Song H, Tsai YY, Aakre JA, Yang P, et al. Meta-analysis of 8q24 for seven cancers reveals a locus between NOV and ENPP2 associated with cancer development. BMC Medical Genetics. 2011 Dec;12(1):156. 00033. Available from: http://bmcmedgenet.biomedcentral.com/articles/10.1186/1471-2350-12-156.

[23] SEARCH, The GENICA Consortium, kConFab, Australian Ovarian Cancer Study Group, Ahmed S, Thomas G, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nature Genetics. 2009 May;41(5):585–590. 00000. Available from: http://www.nature.com/articles/ng.354.

[24] Nielsen FC, van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. Nature Reviews Cancer. 2016 Sep;16(9):599–612. 00119. Available from: http://www.nature.com/articles/nrc.2016.72.

[25] Quigley DA, Fiorito E, Nord S, Van Loo P, Alnaes GG, Fleischer T, et al. The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. Molecular Oncology. 2014 Mar;8(2):273–284. 00000. Available from: http://doi.wiley.com/10.1016/j.molonc.2013.11.008.

[26] Yu M, Li R, Zhang J. Repositioning of antibiotic levofloxacin as a mitochondrial biogenesis inhibitor to target breast cancer. Biochemical and Biophysical Research Communications. 2016 Mar;471(4):639–645. Available from: https://doi.org/10.1016/j.bbrc.2016.02.072.

[27] Liu G, Claret FX, Zhou F, Pan Y. Jab1/COPS5 as a Novel Biomarker for Diagnosis, Prognosis, Therapy Prediction and Therapeutic Tools for Human Cancer. Frontiers in Pharmacology. 2018 Feb;9:135. 00005. Available from: http://journal.frontiersin.org/article/10.3389/fphar.2018.00135/full.

[28] Nikolayeva I, Guitart Pla O, Schwikowski B. Network module identification—A widespread theoretical bias and best practices. Methods. 2018 Jan;132:19–25. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1046202317300373.

[29] Ioachim E, Charchanti A, Briasoulis E, Karavasilis V, Tsanou H, Arvanitis DL, et al. Immunohistochemical expression of extracellular matrix components tenascin, fibronectin, collagen type IV and laminin in breast cancer: their prognostic value and role in tumour invasion and progression. European Journal of Cancer. 2002 Dec;38(18):2362–2370. Available from: https://doi.org/10.1016/s0959-8049(02)00210-1.

[30] Yi W, Xiao E, Ding R, Luo P, Yang Y. High expression of fibronectin is associated with poor prognosis, cell proliferation and malignancy via the NF-κB/p53-apoptosis signaling pathway in colorectal cancer. Oncology Reports. 2016 Oct;36(6):3145–3153. Available from: https://doi.org/10.3892/or.2016.5177.

[31] Sponziello M, Rosignolo F, Celano M, Maggisano V, Pecce V, Rose RFD, et al. Fibronectin-1 expression is increased in aggressive thyroid cancer and favors the migration and invasion of cancer cells. Molecular and Cellular Endocrinology. 2016 Aug;431:123–132. Available from: https://doi.org/10.1016/j.mce.2016.05.007.

[32] Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nature Genetics. 2009 May;41(5):585–590. 00000. Available from: http://www.nature.com/articles/ng.354.

[33] Breyer J, Dorset D, Clark T, Bradley K, Wahlfors T, McReynolds K, et al. An Expressed Retrogene of the Master Embryonic Stem Cell Gene POU5F1 Is Associated with Prostate Cancer Susceptibility. The American Journal of Human Genetics. 2014 Mar;94(3):395–404. 00018. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929714000573.

[34] Chen HY, Chen RH. Cullin 3 Ubiquitin Ligases in Cancer Biology: Functions and Therapeutic Implications. Frontiers in Oncology. 2016 May;6. Available from: https://doi.org/10.3389/fonc.2016.00113.

[35] Loignon M, Miao W, Hu L, Bier A, Bismar TA, Scrivens PJ, et al. Cul3 overexpression depletes Nrf2 in breast cancer and is associated with sensitivity to carcinogens, to oxidative stress, and to chemotherapy. Molecular Cancer Therapeutics. 2009 Jul;8(8):2432–2440. Available from: https://doi.org/10.1158/1535-7163.mct-08-1186.

[36] Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nature Methods. 2015 Mar;12(3):179–185. 00000. Available from: http://www.nature.com/articles/nmeth.3288.

[37] Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. The American Journal of Human Genetics. 2018 May;102(5):717–730. 00090. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929718301344.

[38] Xie T. Analysis of the Gene-Dense Major Histocompatibility Complex Class III Region and Its Comparison to Mouse. Genome Research. 2003 Dec;13(12):2621–2636. Available from: https://doi.org/10.1101/gr.1736803.

[39] Cai JJ, Borenstein E, Petrov DA. Broker Genes in Human Disease. Genome Biology and Evolution. 2010 Jan;2:815–825. 00060. Available from: https://academic.oup.com/gbe/article/doi/10.1093/gbe/evq064/581094.

[40] Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Systems Biology. 2012;6(1):92. 00204. Available from: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-92.

[41] Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. Twin Research and Human Genetics. 2015 Feb;18(1):86–91. 00125. Available from: https://www.cambridge.org/core/product/identifier/S1832427414000796/type/journal_article.

[42] Nakka P, Raphael BJ, Ramachandran S. Gene and Network Analysis of Common Variants Reveals Novel Associations in Multiple Complex Diseases. Genetics. 2016 Oct;204(2):783–798. 00015. Available from: http://www.genetics.org/cgi/doi/10.1534/genetics.116.188391.

[43] Ionita-Laza I, Lee S, Makarov V, Buxbaum J, Lin X. Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. The American Journal of Human Genetics. 2013 Jun;92(6):841–853. 00283. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929713001766.

[44] Wang M, Huang J, Liu Y, Ma L, Potash JB, Han S. COMBAT: A Combined Association Test for Genes Using Summary Statistics. Genetics. 2017 Nov;207(3):883–891. 00006. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.117.300257.

[45] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017 Apr;35(4):316–319. 00176. Available from: http://www.nature.com/articles/nbt.3820.

[46] Sakoda LC, Jorgenson E, Witte JS. Turning of COGS moves forward findings for hormonally mediated cancers. Nature Genetics. 2013 Apr;45(4):345–348. 00060. Available from: http://www.nature.com/articles/ng.2587.

[47] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015 Dec;4(1):7. 01610. Available from: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0047-8.

[48] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2019 Jan;47(D1):D766–D773. 00063. Available from: https://academic.oup.com/nar/article/47/D1/D766/5144133.

[49] Wang Q, Jia P. dmGWAS 3.0; 2014. Accessed: 2019-07-16. https://bioinfo.uth.edu/dmGWAS/.

[50] Ljubić I, Weiskircher R, Pferschy U, Klau GW, Mutzel P, Fischetti M. An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem. Mathematical Programming. 2006 Feb;105(2-3):427–449. 00223. Available from: http://link.springer.com/10.1007/s10107-005-0660-x.

[51] Beisser D, Klau GW, Dandekar T, Muller T, Dittrich MT. BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics. 2010 Apr;26(8):1129–1130. 00188. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq089.

[52] Dittrich M, Beisser D. BioNet; 2008. Accessed: 2019-07-16. https://bioconductor.org/packages/BioNet/.

[53] Scheid S, Spang R. twilight; a Bioconductor package for estimating the local false discovery rate. Bioinformatics. 2005 Jun;21(12):2921–2922. 00054. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti436.

[54] Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al.. HotNet2; 2018. Accessed: 2019-07-16. https://github.com/raphael-group/hotnet2.

[55] Gwinner F. LEANR; 2016. Accessed: 2019-07-16. https://cran.r-project.org/web/packages/LEANR/.

[56] Climente-González H, Azencott CA. martini; 2019. Accessed: 2019-07-16. https://www.bioconductor.org/packages/martini/.

[57] Liu Y. SigMod v2; 2018. Accessed: 2019-07-16. https://github.com/YuanlongLiu/SigMod.

[58] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2019 Nov;Available from: https://doi.org/10.1093/nar/gkz1031.

[59] Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Molecular BioSystems. 2016;12(2):477–479. Available from: https://doi.org/10.1039/c5mb00663e.

[60] 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). 1995 Jan;57(1):289–300. Available from: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[61] Nogueira S, Brown G. Measuring the Stability of Feature Selection. In: Machine Learning and Knowledge Discovery in Databases. vol. 9852. Cham: Springer International Publishing; 2016. p. 442–457. 00000. Available from: http://link.springer.com/10.1007/978-3-319-46227-1_28.

[62] Metcalfe KA, Finch A, Poll A, Horsman D, Kim-Sing C, Scott J, et al. Breast cancer risks in women with a family history of breast or ovarian cancer who have tested negative for a BRCA1 or BRCA2 mutation. British Journal of Cancer. 2008 Dec;100(2):421–425. Available from: https://doi.org/10.1038/sj.bjc.6604830.

[63] Piñero J, Bravo A, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2017 Jan;45(D1):D833–D839. 00369. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw943.

[64] The 1000 Genomes Project Consortium, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74. 00000. Available from: http://www.nature.com/articles/nature15393.

1 Supplementary materials

Table 1: 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	$\overline{\text{Betweenness}}$	$\hat{\text{P}}_{\text{SNP}}$
GS	197 083	197 060	-	2.03×10^7	0.49
SConES GS	1 590	1 585	5	2.52×10^7	0.023
GM	197 083	6 442 446	-	3.99×10^6	0.49
SConES GM	1 692	177 611	5	4.40×10^6	0.055
GI	197 083	28 733 720	-	1.46×10^6	0.49
SConES GI	408	539	5	9.33×10^6	0.076

$\overline{\text{Betweenness}}$: mean betweenness of the selected SNPs in the corresponding full network.$\hat{\text{P}}_{\text{SNP}}$: median P-value of the selected SNPs.

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	$\overline{\text{Betweenness}}$	$\hat{\text{P}}_{\text{gene}}$	$\rho_{consensus}$
SConES GS	5	0	9 805	2.7×10^{-5}	0.19
SConES GM	28	2	4 267	0.067	0.12

$\overline{\text{Betweenness}}$: mean betweenness of the selected genes in the full network. $\hat{\text{P}}_{\text{gene}}$: median P-value of the selected genes; $\rho_{consensus}$: Pearson correlation with the consensus network.

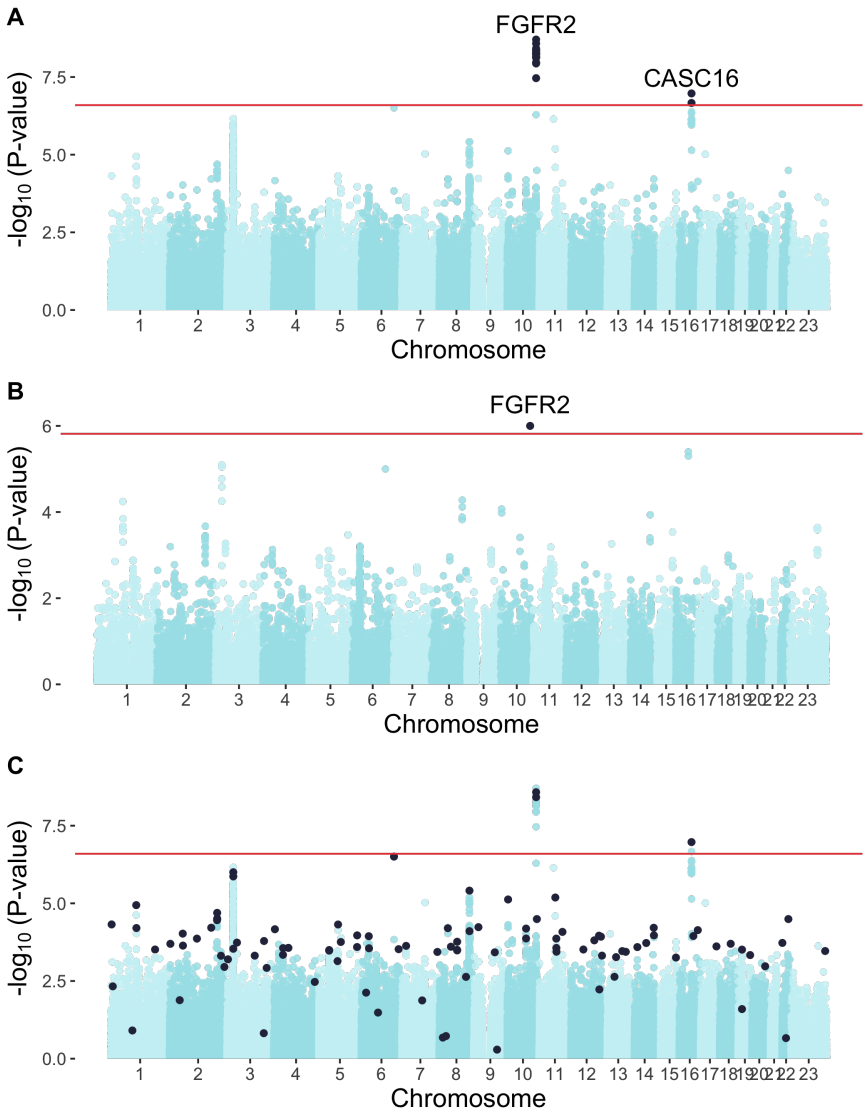

Fig. 1: Association in GENESIS. The red line represents the Bonferroni threshold. **(A)** SNP association, measured from the outcome of a 1 d.f. χ^2 allelic test. Significant SNPs that are within a coding gene, or within 50 kilobases of its boundaries, are annotated. The Bonferroni threshold is 2.54×10^{-7}. **(B)** Gene association, measured by P-value of VEGAS2 [41] using the 10% of SNPs with the lowest P-values. The Bonferroni threshold is 1.53×10^{-6}. **(C)** SNP association as in panel (A). The SNPs in black are selected by a L1-penalized logistic regression (Section 4.5.2, $\lambda = 0.03$).

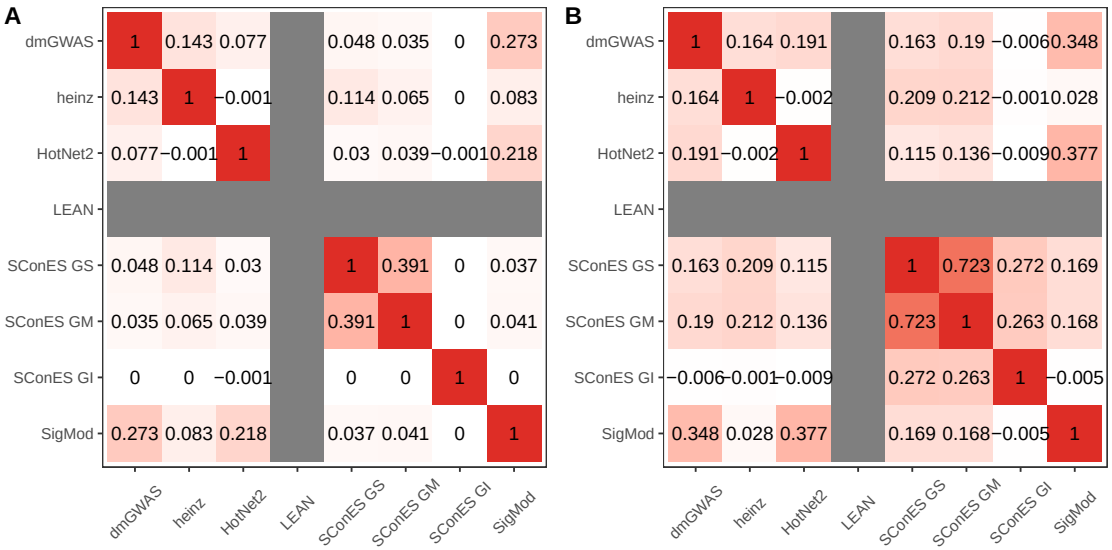

Fig. 2: Pearson correlation between the different solution subnetworks. **(A)** Correlation between selected SNPs. **(B)** Correlation between selected genes. In general, the solutions display a very low overlap.

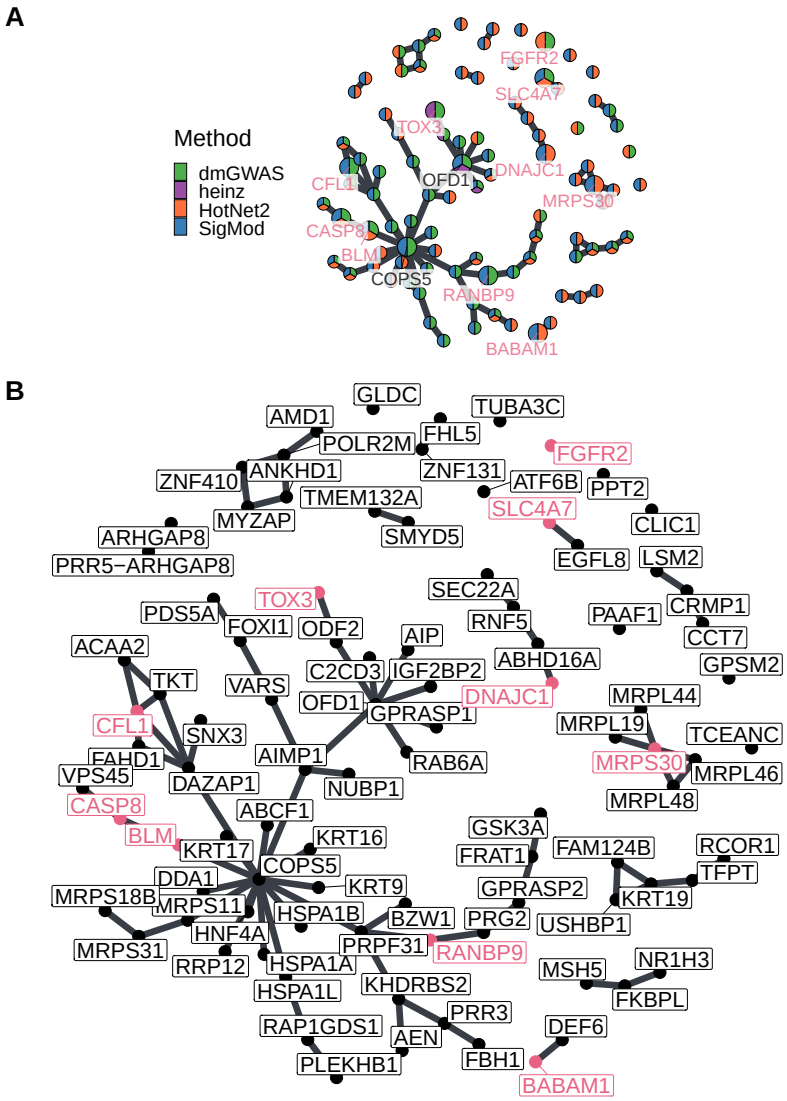

Fig. 3: Consensus subnetwork on GENESIS (Section 4.3.4). **(A)** Each node is represented by a pie chart, which shows the methods that selected it. We labeled (and enlarged) the two most central genes (*COPS5* and *OFD1*) and those genes that are known breast cancer susceptibility genes and/or significantly associated with breast cancer susceptibility in the BCAC dataset. The latter ones are also colored in pink. This panel is identical to Figure 2. **(B)** Same network, but every gene name is indicated.

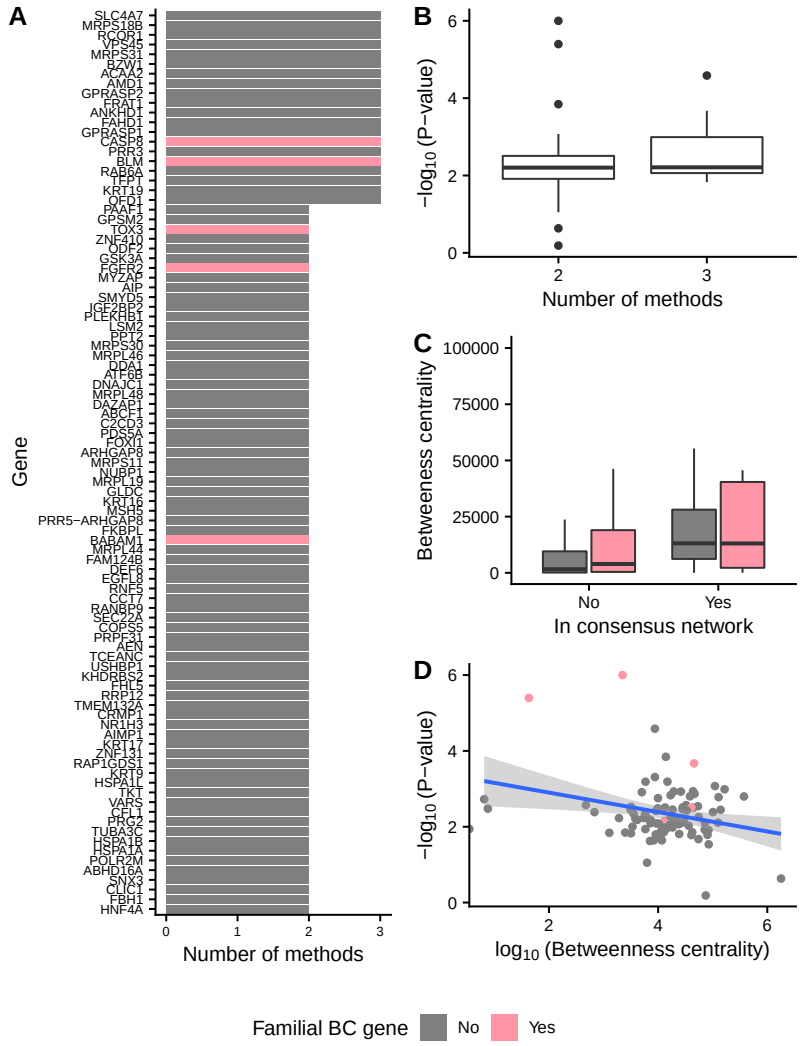

Fig. 4: Genes on the consensus network. Breast cancer susceptibility genes are colored in pink; the rest are colored in grey. **(A)** Number of methods selecting every gene in the subnetwork. **(B)** VEGAS2 P-values of association of the genes, with regards to the number of methods that selected them. **(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 VEGAS2 P-value of the genes in the consensus network. The blue line represents a fitted generalized linear model.

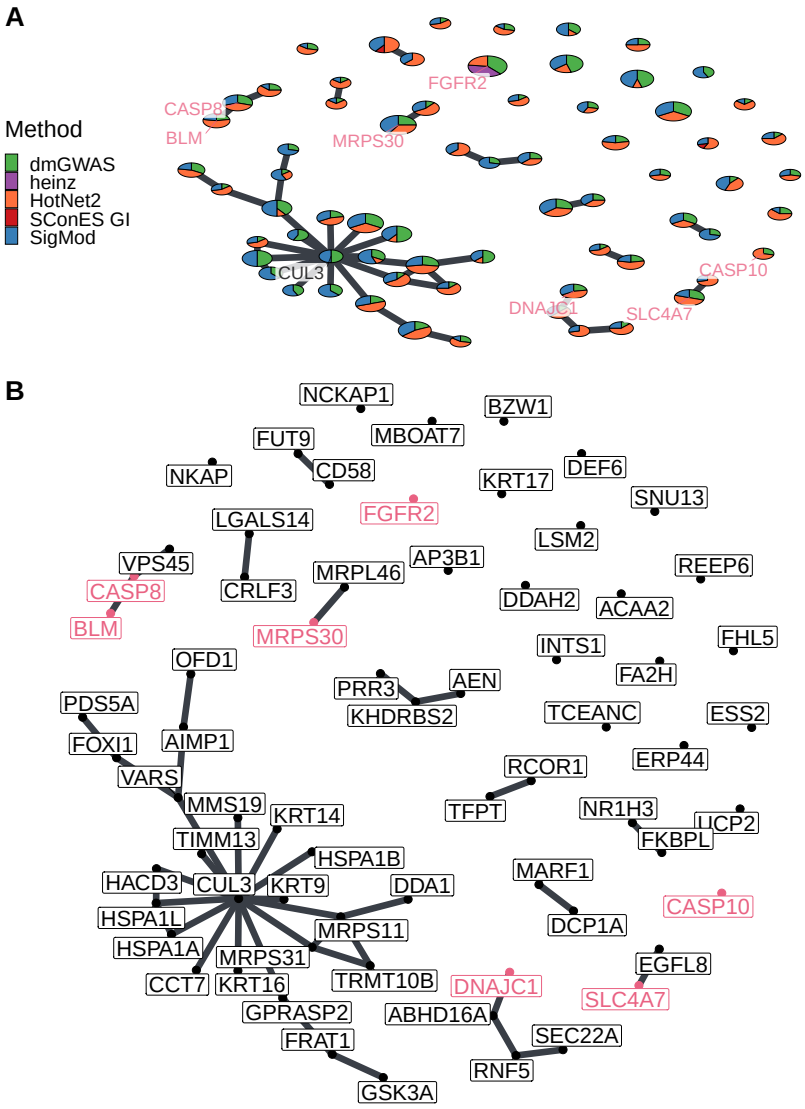

Fig. 5: Stable consensus subnetwork on GENESIS (Section 2.7). **(A)** Each node is represented by a pie chart, which shows the methods that selected it. We labeled (and enlarged) the two most central genes (*CUL3*) and those genes that are known breast cancer susceptibility genes and/or significantly associated with breast cancer susceptibility in the BCAC dataset. The latter ones are also colored in pink. This panel is identical to Figure 5. **(B)** Same network, but every gene name is indicated.

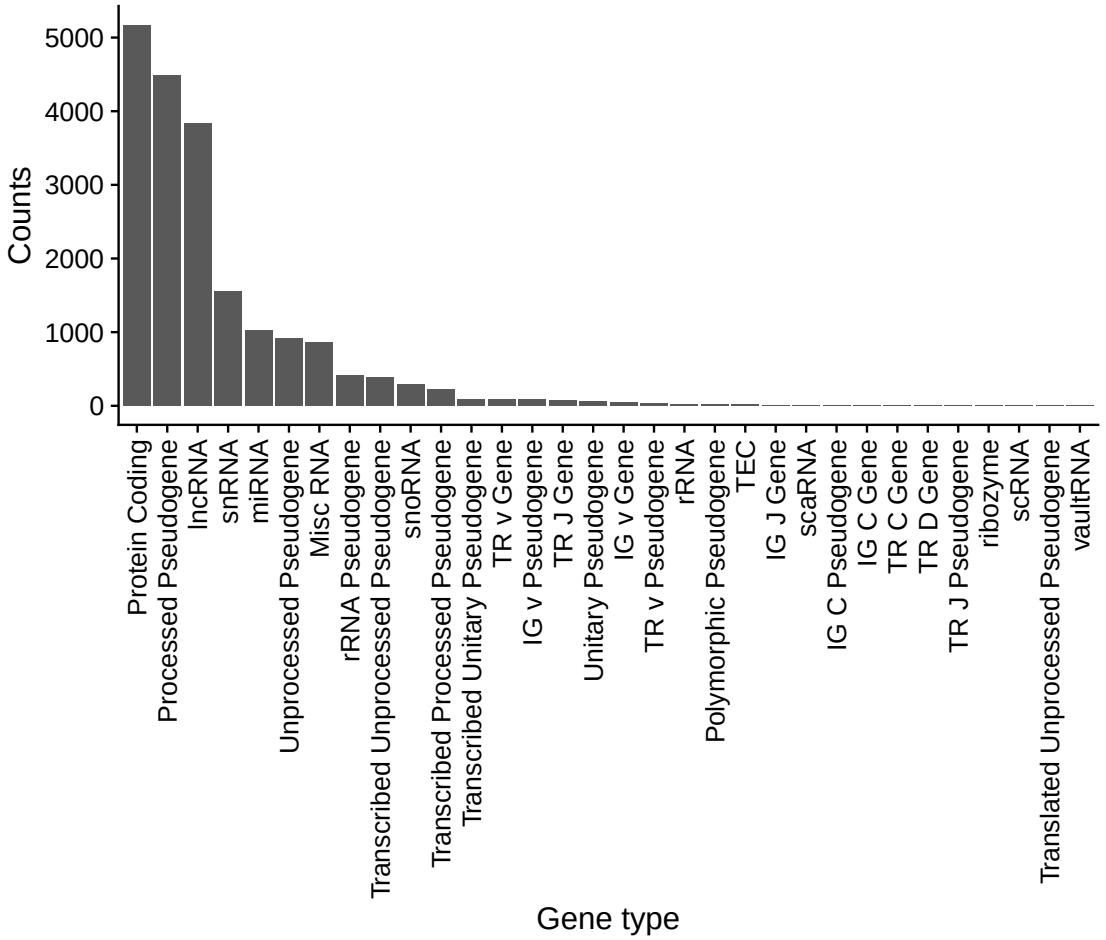

Fig. 6: Biotypes of genes from the annotation that are not present in the HINT protein-protein interaction network.

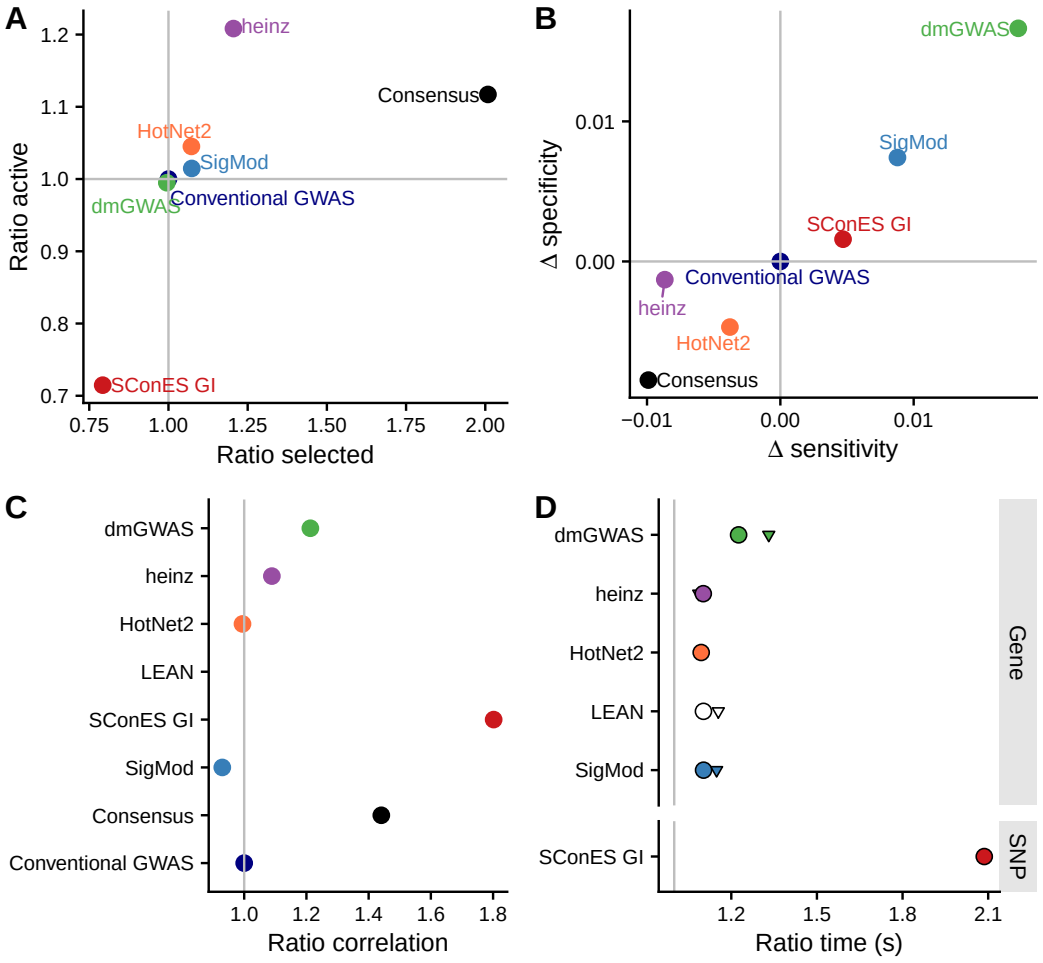

Fig. 7: Comparison of benchmark on high-throughput (HT) interactions to benchmark on both high-throughput and literature curated interactions (HT+LC). Grey lines represent no change in the statistic between the benchmarks (1 for ratios mean(HT) / mean(HT + LC), 0 for differences mean(HT) - mean(HT + LC)). **(A)** Ratios of the selected features between both benchmarks and of the active set. **(B)** Shifts in sensitivity and specificity. **(C)** Shift in Pearson correlation between benchmarks. **(D)** Ratio between the runtimes of the benchmarks. For gene network-based methods, inverted triangles represent the ratio of runtimes of the algorithms themselves, and circles the total time, which includes the algorithm themselves and the additional 119 980 seconds (1 day and 9.33 hours) that VEGAS2 took on average to compute the gene scores from SNP summary statistics. In general, adding additional interactions slightly improves the stability of the solution, but increases the solution size, has mixed effects on the sensitivity and specificity, and impacts negatively the required runtime of the algorithms.

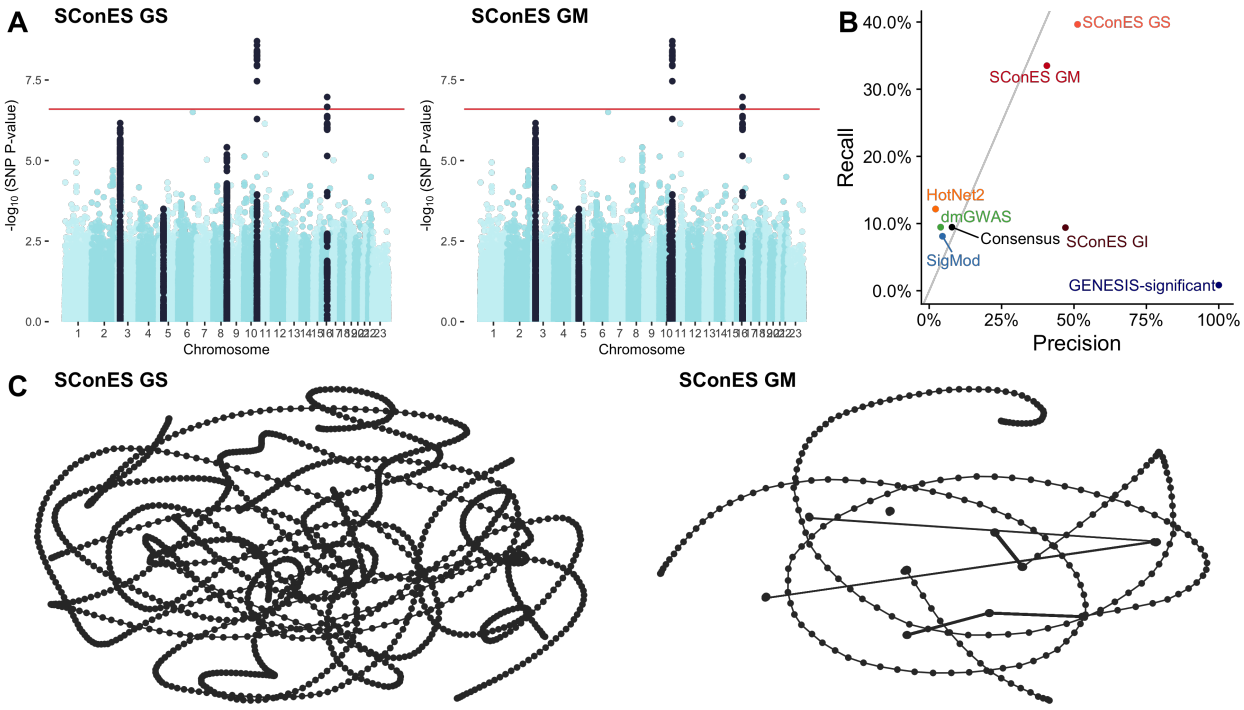

Fig. 8: Overview of the solutions produced by the SConES on the GS and GM networks (Section 4.3.3) on the GENESIS dataset. **(A)** Manhattan plots of SNPs; in black,the method's solution. The Bonferroni threshold (2.54×10^{-7}) is indicated by a red line. **(B)** Precision and recall of the evaluated methods with respect to Bonferroni-significant SNPs (SConES) or genes (other methods) in BCAC. For reference, we added a gray line with a slope of 1. **(C)** Solution networks.

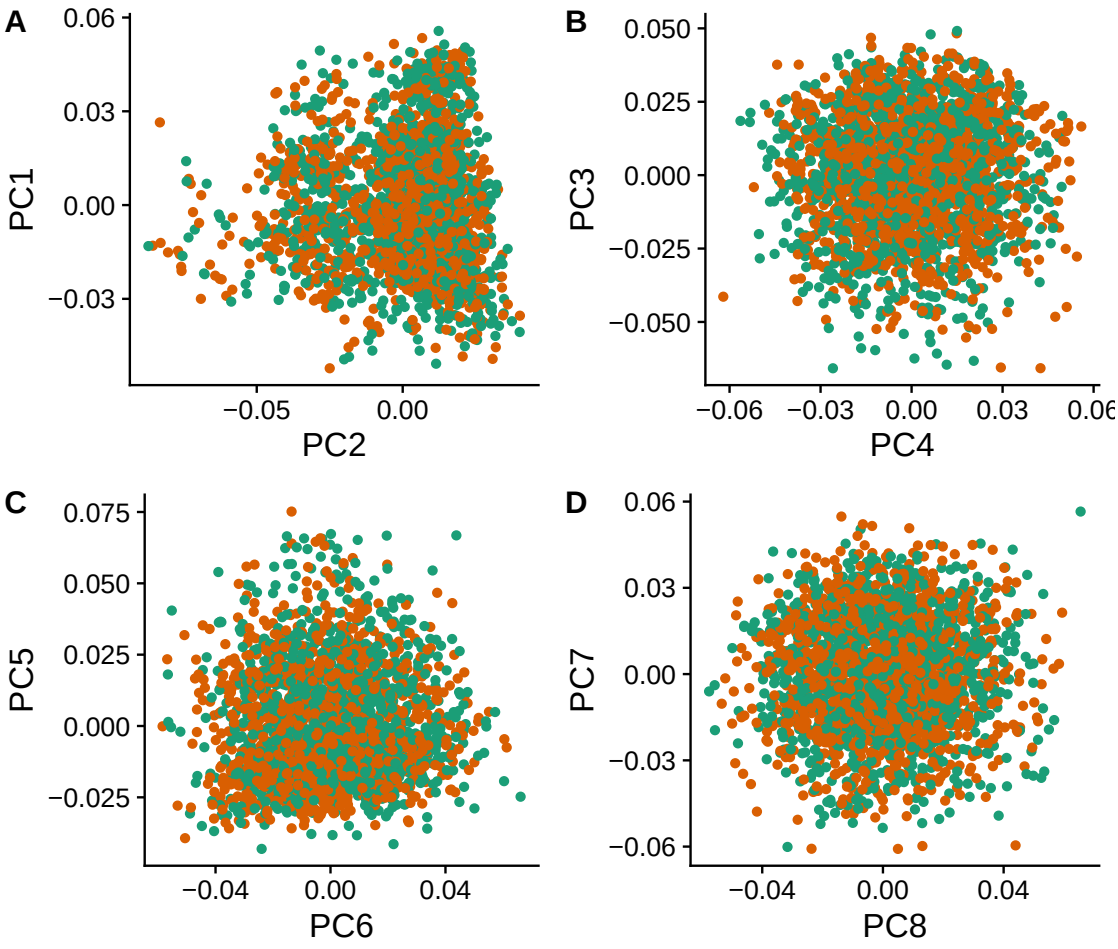

Fig. 9: GENESIS shows no differential population structure between cases and controls. **(A,B,C,D)** Eight main principal components computed on the genotypes of GENESIS. Cases are colored in green, controls in orange.