bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Biological networks and GWAS: comparing and combining network methods to understand the genetics of familial breast cancer susceptibility in 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], GENESIS study group[¶]

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
¶Membership list can be found in the Acknowledgments section.
* hector.climente(at)riken.jp

Abstract

Network approaches to disease use biological networks, which model functional relationships between the molecules in a cell, to generate hypotheses about the genetics of complex diseases. Several among them jointly consider gene scores, representing the association between each gene and the disease, and the biological context of each gene, modeled by a network. Here, we study six such network methods using gene scores from GENESIS, a genome-wide association study (GWAS) on French women with non-BRCA familial breast cancer. We provide a critical comparison of these six methods, discussing the impact of their mathematical formulation and parameters. Using a biological network yields more compelling results than standard GWAS analyses. Indeed, we find significant overlaps between our solutions and the genes identified in the largest GWAS on breast cancer susceptibility. We further propose to combine these solutions into a consensus network, which brings further insights. The consensus network contains *COPS5*, a gene related to multiple hallmarks of cancer, and 14 of its neighbors. The main drawback of network methods is that they are not robust to small perturbations in their inputs. Therefore, we propose a stable consensus solution, formed by the most consistently selected genes in multiple subsamples of the data. In GENESIS, it is composed of 68 genes, enriched in known breast cancer susceptibility genes (*BLM*, *CASP8*, *CASP10*, *DNAJC1*, *FGFR2*, *MRPS30*, and *SLC4A7*, P-value $= 3 \times 10^{-4}$) and occupying more central positions in the network than most genes. The network is organized around *CUL3*, which is involved in the regulation of several genes linked to cancer progression. In conclusion, we showed how network methods help overcome the lack of statistical power of GWAS and improve their interpretation. Project-agnostic implementations of all methods are available at `https://github.com/hclimente/gwas-tools`.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Author summary

Genome-wide association studies (GWAS) scan thousands of genomes 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. However, 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. The omnigenic model offers an appealing explanation: only a limited number of *core* genes are directly involved in the disease, but gene functions are deeply interrelated, and so many other genes can 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. This study applies and compares 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. We provide ready-to-use implementations of all the examined methods.

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 from the disease of interest, and the controls, taken from the general population. Then, a statistical test of association (e.g., based on logistic regression) is conducted between each SNP and the phenotype. 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 deepening our understanding of the disease.

GWAS have successfully identified thousands of variants underlying many common diseases [2]. However, this experimental setting also presents inherent 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 it can only detect variants with larger effects [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, most GWAS apply a conservative multiple test correction, typically the previously mentioned Bonferroni correction. However, Bonferroni correction is overly conservative when the statistical tests correlate, as happens in GWAS [4]. Another open issue is the interpretation of the results, as the functional consequences of most common variants are unknown. 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 strongly interrelated and influence each other's function, 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 that includes the structure of biological data might help alleviate the issues above.

For this reason, many authors turn to network biology to handle the complex interplay of biomolecules that lead to disease [7, 8]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

function; for instance, their expressions are 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 [9]. In fact, an examination of biological networks shows that disease genes have differential properties [9, 10]: they tend to occupy central positions in the network (although not the most central ones); disease genes for the same pathology tend to cluster in modules; and often they are bottlenecks that interconnect modules.

Network-based discovery methods exploit the differential properties described above to identify disease genes on GWAS data [11, 12]. In essence, each gene receives a score of association with the disease, computed from the GWAS data, and a set of biological relationships, given by a network built on prior knowledge. Then, the problem becomes finding a functionally-related set of highly-scoring genes. Multiple solutions have been proposed to this problem, often stemming from different mathematical frameworks and considerations of what the optimal solution looks like. For example, some methods restrict the problem to specific types of subnetworks. Such is the case of LEAN [13], which focuses on "star" subnetworks, i.e., instances where both a gene and its direct interactors are associated with the disease. Other algorithms, like dmGWAS [14] and heinz [15], do not impose such strong constraints and search for 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 [16], SConES [17], and SigMod [18].

In this work, we studied the relevance of network-based approaches to genetics by applying these six network methods to GWAS data. They use different interpretations of the omnigenic model and provide a representative view of the field. We worked on the GENESIS dataset [19], a study on familial breast cancer conducted in the French population. After a classical GWAS approach, we used these network methods to identify additional breast cancer susceptibility genes. Lastly, we compared the solutions obtained by the different methods and aggregated them to obtain consensus solutions of predisposition to familial breast cancer that addressed their shortcomings.

2 Materials and methods

2.1 GENESIS dataset, preprocessing, and quality control

The GENE Sisters (GENESIS) study investigated risk factors for familial breast cancer in the French population [19]. Index cases were patients with infiltrating mammary or ductal adenocarcinoma, who had a sister with breast cancer, and tested negative for *BRCA1* and *BRCA2* pathogenic variants. Controls were unaffected colleagues 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 were controls, and 1 298 were cases. The genotyping platform was the iCOGS array, a custom Illumina array designed to study the genetic susceptibility to hormone-related cancers [20]. It contained 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.

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. We also removed a subset of 20 duplicated SNPs in *FGFR2*. We excluded the samples with more than 10% missing

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genotypes. After controlling for relatedness, we excluded 17 additional samples (6 for sample identity error, 6 controls related to other samples, 2 cases 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 (S1 Fig). Independently, we did not find evidence of genomic inflation ($\lambda = 1.05$) either, further confirming the absence of confounding population structure.

2.2 SNP- and gene-based GWAS

To measure the association between genotype and susceptibility to breast cancer, we performed a per-SNP 1 d.f. χ^2 allelic test using PLINK v1.90 [21]. To obtain significant SNPs, we performed a Bonferroni correction to keep the family-wise error rate below 5%. The threshold used was $\frac{0.05}{197083} = 2.54 \times 10^{-7}$.

Then, we used VEGAS2 [22] to compute the gene-level association score from the P-values of the SNPs mapped to them. More specifically, 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 computed VEGAS2 scores for each gene using only the 10% of SNPs with the lowest P-values among all those mapped to it. We used the 62 193 genes described in GENCODE 31 [23], although only 54 612 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 mapped to at least one of these genes. We also performed a Bonferroni correction to obtain significant genes; in this case, the threshold of significance was $\frac{0.05}{32767} = 1.53 \times 10^{-6}$.

2.3 Network methods

2.3.1 Mathematical notations

In this article, we used 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.

Nodes can be described by properties provided by the topology of the graph. We focused on two of those: degree centrality and betweenness centrality. The degree centrality, or simply emphdegree, is the number of edges that a node has. The betweenness centrality, or *betweenness*, is the number of times a node participates in shortest paths between two other nodes.

We also used two matrices that describe two different properties of a graph. Both matrices are square and have as many rows and columns as nodes are in the network. The element (i, j) hence represents a relationship between v_i and v_j. The *adjacency matrix* W_G contains a 1 when the corresponding nodes are connected, and 0 otherwise; its diagonal is zero. The *degree matrix* D_G is a diagonal matrix that contains the degree of the different nodes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.3.2 Networks

Gene network The mathematical formulations of the different network methods are compatible with any type of biological network (e.g., from protein interactions or gene co-expression). Here, we used protein-protein interaction networks (PPIN) for all network methods except SConES, as PPINs are interpretable, well-characterized, and the methods were designed to run efficiently on them. We built our PPIN from both binary and co-complex interactions stored in the HINT database (release April 2019) [24]. Unless otherwise specified, we used only interactions coming from high-throughput experiments, leaving out targeted studies that might bias the topology of the PPIN. Out of the 146 722 interactions from high-throughput experiments that HINT stores, we could 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 2.3.3).

Additionally, we compared the results obtained on this PPIN with those obtained on a 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 [17] was the only network method designed to handle SNP networks. As in gene networks, two SNPs were connected in a SNP network when there was evidence of shared functionality between them. Azencott et al. [17] proposed three ways of building such 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 using the PPIN described above, as it fitted the scope of this work better. However, at different stages, we also compared GI to GS and GM to understand how including the PPIN affects SConES' output. For the GM network, we used the mapping described in Section 2.3.5. In all three, we scored the nodes using the 1 d.f. χ^2 statistic of association. The properties of these three subnetworks are available in S1 Table.

2.3.3 High-score subnetwork search algorithms

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.). Several aspects have to be considered 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 open-source methods that had an implementation available and accessible documentation. We summarize their main differences in Table 1. We scored both SNPs and genes with the P-values (or transformations) computed in Section 2.2.

dmGWAS dmGWAS seeks the subgraph with the highest local density in low P-values [14]. To that end, it searches candidate solutions using a greedy, "seed and extend", heuristic:

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Summary of the differences between the network methods.

Method	Field	Nodes	Exhaustive	Solution	Comp.	Input	Scoring	Ref.
dmGWAS	GWAS	Genes	No	-	1	Summary	$-\log_{10}(P)$	[14]
heinz	Omics	Genes	Yes	-	1	Summary	BUM	[15]
HotNet2	Omics	Genes	Yes	Module	≥ 1	Summary	Local FDR	[16]
LEAN	Omics	Genes	Yes	Star	≥ 1	Summary	$-\log_{10}(P)$	[13]
SConES	GWAS	SNPs	Yes	Module	≥ 1	Genotypes	1 d.f. χ^2	[17]
SigMod	GWAS	Genes	Yes	Module	≥ 1	Summary	$\Phi^{-1}(1 - P)$	[18]

Field: field in which the algorithm was developed. ***Nodes***: the type of nodes in the network, either genes (PPIN) or SNPs. ***Exhaustive***: whether the method explores all the possible solutions given the selected parameters. ***Solution***: additional properties enforced on the solution, 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 were transformed into node scores. In the case of heinz, BUM stands for beta-uniform model; 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 Z_m for S_i as

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

   where k is the number of genes in S_i, z_j is the Z score of gene j, computed as $\phi^{-1}(1 - \text{P-value}_j)$, 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 Z_{m+1} by more than a threshold $Z_m \times (1 + \mathtt{r})$.
5. Repeat 2-4 until no further enlargement is possible.
6. Add S_i to the list of subnetworks to return. Normalize its Z-score as

$$Z_N = \frac{Z_m - \text{mean}\left(Z_m(\pi)\right)}{\text{SD}\left(Z_m(\pi)\right)}, \tag{2}$$

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

DmGWAS carries out this process on every gene in the PPIN. We used the implementation of dmGWAS in the dmGWAS 3.0 R package [25]. Unless otherwise specified, we used the suggested parameters $\mathtt{d} = 2$ and $\mathtt{r} = 0.1$. We used the function `simpleChoose` to select the solution, which aggregates the top 1% subnetworks.

heinz The goal of heinz is to identify the highest-scored connected subnetwork [15]. The authors proposed a transformation of the genes' P-value into a score that is negative under weak association with the phenotype, and positive under a strong one. This transformation is achieved by modeling 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, which seeks to select the connected subnetwork S that maximizes the *profit* p(S), defined as:

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$p(S) = \sum_{v \in V_S} p(v) - \sum_{e \in E_S} c(e). \tag{3}$$

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. [26] which, in practice, is often fast and optimal, although neither is guaranteed. We used BioNet's implementation of heinz [27, 28].

HotNet2 HotNet2 was developed to find connected subgraphs of genes frequently mutated in cancer [16]. To that end, it considers both the local topology of the PPIN and the nodes' scores. An insulated heat diffusion process captures the former: 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 a stationary state is reached, in which the temperature of the nodes does not change anymore, and results in a diffusion matrix F. F is used to compute the similarity matrix E that models exchanged heat as

$$E = F \operatorname{diag}(w(V)), \tag{4}$$

where $\operatorname{diag}(w(V))$ 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. Hence, E_{ij} can be interpreted as a (non-symmetric) similarity between those two nodes. To obtain densely connected solutions, HotNet2 prunes E, only preserving edges such that $w(E) > \delta$. Lastly, HotNet2 evaluates the statistical significance of the solutions by comparing their size to the size of PPINs obtained by permuting the node scores. We assigned the initial node scores as in Nakka et al. [29], giving a 0 to the genes with a VEGAS2 P-values of association with the disease, and -$\log_{10}$(P-value) to those likely to be. In the GENESIS dataset, the threshold separating both was a P-value of 0.125, which we obtained using a local FDR approach [30]. 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 [16]. HotNet2 is implemented in Python [31].

LEAN LEAN searches altered "star" subnetworks, that is, subnetworks composed of one central node and all its interactors [13]. By imposing this restriction, LEAN can exhaustively test all such subnetworks (one per node). For a particular star subnetwork of size m LEAN performs three steps:

1. Rank the P-values of the involved nodes as $\mathrm{p}_1 \leq \ldots \leq \mathrm{p}_\mathrm{m}$.
2. Conduct k binomial tests to compute the probability of having k out of m P-values lower or equal to p_k under the null hypothesis. The minimum of these k P-values is the score of the subnetwork.
3. Transform this score 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.

We adjust these P-values for multiple testing through a Benjamini-Hochberg correction. We used the implementation of LEAN from the `LEANR` R package [32].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SConES SConES searches the minimal, modular, and maximally associated subnetwork in a SNP graph [17]. Specifically, it solves the problem

$$\arg\max_{S \subseteq G} \underbrace{\sum_{v \in V_S} w(v)}_{\text{association}} - \underbrace{\lambda \sum_{v \in V_S} \sum_{u \notin V_S} W_{vu}}_{\text{connectivity}} - \underbrace{\eta |V_S|}_{\text{sparsity}} , \tag{5}$$

where λ and η are parameters that control the sparsity and the connectivity of the model. The connectivity term penalizes disconnected solutions, with many edges between selected and unselected nodes. Given a λ and an η, Eq 5 has a unique solution that SConES finds using a graph min-cut procedure. As in Azencott et al. [17], we selected λ and η by cross-validation, choosing the values that produce the most stable solution across folds. In this case, the selected parameters 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` [33].

SigMod SigMod searches the highest-scoring, most densely connected subnetwork [18]. It addresses an optimization problem similar to that of SConES (Eq 5), but with a different connectivity term that favors solutions containing many edges:

$$\arg\max_{S \in G} \underbrace{\sum_{v \in V_S} w(v)}_{\text{association}} + \underbrace{\lambda \sum_{v \in V_S} \sum_{u \in V_S} W_{vu}}_{\text{connectivity}} - \underbrace{\eta |V_S|}_{\text{sparsity}} . \tag{6}$$

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

SigMod presents three important differences with SConES. First, it was designed for PPINs. Second, it favors solutions containing many edges between the selected nodes. SConES, instead, penalizes connections between selected and unselected nodes. Third, it explores the grid of parameters differently, and processes their respective solutions. Specifically, for the range of $\lambda = \lambda_{\text{min}}, \ldots, \lambda_{\text{max}}$ for the same η, it prioritizes the solution with the largest change in size from λ_{n} to $\lambda_{\text{n+1}}$. Additionally, that change needs to be larger than a user-specified threshold `maxjump`. Such a large change implies that the network is densely interconnected. This results in one candidate solution for each η, which is 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. Finally, SigMod chooses the candidate solution with the highest standardized score, and that is not larger than a user-specified threshold (`nmax`). We used the default parameters `maxjump` $= 10$ and `nmax` $= 300$. SigMod is implemented in an R package [34].

Consensus We built a consensus solution by retaining the genes selected by at least two of the six methods (using SConES GI for SConES). It includes any edge between the selected genes in the PPIN.

We performed all the computations in the cluster described in Section 2.8.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.3.4 Parameter space

We used the network methods with the parameters recommended by their authors, or with the default parameters in their absence. Additionally, we explored the parameter space of the different methods to study how they alter the output.

dmGWAS We tested multiple values for `r` (0.0001, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, and 1) and `d` (1, 2, and 3).

heinz We tested multiple FDR thresholds (0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1).

HotNet2 We tested different thresholds to decide which genes would receive a score of 0 and which ones a score of $-log_{10}$(P-value): 0.001, 0.01, 0.05, 0.125, 0.25, and 0.5.

LEAN We used the following significance cutoffs for LEAN's P-values (0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, and 1).

SConES We used the values of λ and η that `martini` explores by default (35.54, 5.40, 0.82, 0.12, 0.02, 0.01, 4.39e-4, 6.68e-5, 1.02e-5, and 1.55e-6 in both cases)

SigMod We tested multiple values for the parameters `nmax` (10, 50, 100, 300, 700, 1000, and 10 000) and `maxjump` (5, 10, 20, 30, and 50).

2.3.5 Comparing SNP-methods to gene-methods and vice versa

In multiple steps of this article, we compared the outcome of a method that works on genes with the outcome of one that works on SNPs. For this purpose, we used the SNP-gene correspondence described in Section 2.2. To convert a list of SNPs into a list of genes, we included all the genes mapped to any of those SNPs. Conversely, to convert a list of genes into a list of SNPs, we included all the SNPs mapped to any of those genes.

2.4 Pathway enrichment analysis

We searched for pathways enriched in the gene solutions produced by the above methods. We conducted a hypergeometric test on pathways from Reactome [35] using the function `enrichPathway` from the `ReactomePA` R package [36]. The universe of genes included any gene that we could map to a SNP in the iCOGS array (Section 2.2). We adjusted the P-values for multiple testing as in Benjamini and Hochberg [37] (BH): pathways with a BH adjusted P-value < 0.05 were deemed significant.

2.5 Benchmark of methods

We evaluated multiple properties of the different methods (described in Sections 2.5.1 and 2.5.2) through a 5-fold subsampling setting. We applied each method to 5 random subsets of the original GENESIS 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. Every method and repetition ran in the same computational settings (Section 2.8).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.5.1 Properties of the solution

We compared the runtime, the number of selected features (genes or SNPs), and the stability (sensitivity to the choice of train set) of the different network methods. Nogueira and Brown [38] proposed quantifying a method's stability using the Pearson correlation between the genes selected on different subsamples. This correlation was calculated between vectors with the length of the total number of features, containing a 0 at position i if feature i was not selected and a 1 if it was.

2.5.2 Classification accuracy of selected SNPs

A desirable solution offers good predictive power on the 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, which searches for a small subset of SNPs that provides good classification accuracy at predicting 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*). The active set was a plausible, more sparse solution with comparable predictive power to the original solution. The L1 penalty was set by cross-validation, choosing the value that minimized misclassification error.

We applied each network method to each train set. Then, we trained the classifier on the same train set using only the selected SNPs. When the method retrieved a list of genes, we proceeded as explained in Section 2.3.5. Lastly we evaluated the sensitivity and the specificity of the classifier on the test set. To obtain a baseline, we also trained the classifier on all the SNPs of the train set.

We did not expect a linear model on selected SNPs to separate cases from controls well. Indeed, the lifetime cumulative 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 [39]. However, classification accuracy may be one additional informative criterion on which to evaluate solutions.

2.6 Comparison to state-of-the-art

An alternative way to evaluate the methods is by comparing their solutions to an external dataset. For that purpose, we used the 153 genes associated to familial breast cancer on DisGeNET [40]. 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. BCAC included 13 250 641 SNPs genotyped or imputed on 228 951 women of European ancestry, mostly from the general population [41]. Through imputation, BCAC includes more SNPs than the iCOGS array used for GENESIS (Section 2.1). However, in all the comparisons in this paper we focused on the SNPs that passed quality control in GENESIS. Hence, we used the same Bonferroni threshold as in Section 2.2 to determine the significant SNPs in BCAC. We also computed gene-scores in the BCAC data using VEGAS2, as in Section 2.1. In this case, we did use the summary statistics of all 13 250 641 available SNPs and the genotypes from European samples from the 1000 Genomes Project [42] to compute the LD patterns. Since these genotypes did not include chromosome X, we excluded it from this analysis. All comparisons included only the genes in common between GENESIS and BCAC, so we used a different Bonferroni threshold (1.66×10^{-6}) to call gene significance.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.7 Network rewirings

Rewiring the PPIN while preserving the number of edges of each gene allowed to study the impact of the topology on the output of network methods. Indeed, the edges lose their biological meaning while the topology of the network is conserved. We produced 100 such rewirings by randomly swapping edges in the PPIN. We still scored the genes as described in Section 2.3.3. We only applied only four methods on the rewirings: heinz, dmGWAS, LEAN, and SigMod. We excluded HotNet2 and SConES since they took notably longer to run.

2.8 Computational resources

We ran all the computations on a Slurm cluster, running Ubuntu 16.04.2 on the nodes. The CPU models on the nodes were Intel Xeon CPU E5-2450 v2 at 2.50GHz and Intel Xeon E5-2440 at 2.40GHz. The nodes running heinz and HotNet2 had 20GB of memory; the ones running dmGWAS, LEAN, SConES, and SigMod, 60GB. For the benchmark (Section 2.5), we ran each of the methods on the same Ubuntu 16.04.2 node, with a CPU Intel Xeon E5-2450 v2 at 2.50GHz, and 60GB of memory.

2.9 Code and data availability

We developed computational pipelines for several steps of GWAS analyses, such as physically mapping SNPs to genes, computing gene scores, and running six different network methods. We created a pipeline with a clear interface that should work on any GWAS dataset for each of those processes. They are compiled in `https://github.com/hclimente/gwas-tools`. The code that applies them to GENESIS, 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 solutions on NDEx (`http://www.ndexbio.org`), under the UUID e9b0e22a-e9b0-11e9-bb65-0ac135e8bacf.

Summary statistics for SNPs and genes are available at `https://github.com/hclimente/genewa`. We cannot share genotype data publicly for confidentiality reasons, but are available from GENESIS. Interested researchers can contact nadine.andrieu(at)curie.fr.

3 Results

3.1 Conventional SNP- and gene-based analyses retrieve the *FGFR2* locus in the GENESIS dataset

We conducted association analyses in the GENESIS dataset (Section 2.1) at both SNP and gene levels (Section 2.2). At the SNP level, two genomic regions had a P-value lower than the Bonferroni threshold on chromosomes 10 and 16 (S2 FigA). The former overlaps with the gene *FGFR2*, the latter with *CASC16* and the protein-coding gene *TOX3*. Variants in both *FGFR2* and *TOX3* have been repeatedly associated with breast cancer susceptibility in other case-control studies [41], *BRCA1* and *BRCA2* carrier studies [43], and hereditary breast and ovarian cancer families negative for mutations in *BRCA1* and *BRCA2* [44]. At the gene level, only *FGFR2* was significantly associated with breast cancer (S2 FigB).

Closer examination revealed two other regions (3p24 and 8q24) having low, albeit not genome-wide significant, P-values. Both of them have been associated with breast cancer susceptibility in the past [45, 46]. We applied an L1-penalized logistic regression using all GENESIS genotypes as input and the phenotype (cancer/healthy) as the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

outcome (Section 2.5.2). The algorithm selected 100 SNPs, both from all regions mentioned above and new ones (S2 FigC). However, it was 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 (Section 2.3.5) were not interconnected in the protein-protein interaction network (PPIN, Section 2.3.2). Moreover, the classification performance of the model was low (sensitivity = 55%, specificity = 55%, Section 2.5). Together, these issues motivated exploring network methods, which consider not only statistical association but also the location of each gene in a PPIN to find susceptibility genes.

3.2 Network methods successfully identify genes associated with breast cancer

We applied six network methods to the GENESIS dataset (Section 2.3.3). As none of the networks examined by LEAN was significant (Benjamini-Hochberg [BH] correction adjusted P-value < 0.05), we obtained five solutions (Fig 1): one for each of the remaining four gene-based methods, and one for SConES GI (which works at the SNP level).

These solutions differed in many aspects, making it hard to draw joint conclusions. For starters, the overlap between the genes featured in each solution was relatively small (Fig 1A). However, the methods tended to agree on the genes with the strongest signal: genes selected by more methods tended to have lower P-value of association (Fig 1B).

Another major difference was the solution size: the largest solution, produced by HotNet2, contained 440 genes, while heinz's contained only 4 genes. While SConES GI did not recover any protein coding gene, working with SNP networks rather than gene networks allowed it to retrieve four subnetworks in intergenic regions and another subnetwork overlapping an RNA gene (*RNU6-420P*).

The topologies of the five solutions differed as well (Fig 1C), as measured by the median centrality and the number of connected components (Table 2). Three methods yielded more than one connected component: SConES, as described above, SigMod, and HotNet2. HotNet2 produced 135 subnetworks, 115 of which have fewer than five genes. The second largest subnetwork (13 nodes) contained the two breast cancer susceptibility genes *CASP8* and *BLM* (Section 2.6).

Table 2. Summary statistics on the solutions of multiple network methods on the PPIN. The first row contains the summary statistics on the whole PPIN.

Network	# genes	# edges	# components	$\overline{\text{Betweenness}}$	$\hat{\text{P}}_{\text{gene}}$	# genes in consensus
HINT HT	13 619	142 541	15	16 706	0.46	93/93
dmGWAS	194	450	1	49 115	0.19	55/93
heinz	4	3	1	113 633	0.001	4/93
HotNet2	440	374	130	7 739	0.048	63/93
LEAN	0	0	0	-	-	0/93
SConES GI	0 (1)	0	0	-	-	0/93
SigMod	142	249	11	92 603	0.008	84/93
Consensus	93	186	21	50 737	0.006	93/93
Stable consensus	68	49	32	94 854	0.005	43/93

 # 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. **# components**: number of connected components. $\overline{\textbf{Betweenness}}$: median betweenness of the selected genes in the PPIN. $\hat{\textbf{P}}_{\textbf{gene}}$: median VEGAS2 P-value of the selected genes. **# genes in consensus**: number of genes in common between the method's solution and the 93 genes in the consensus solution.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

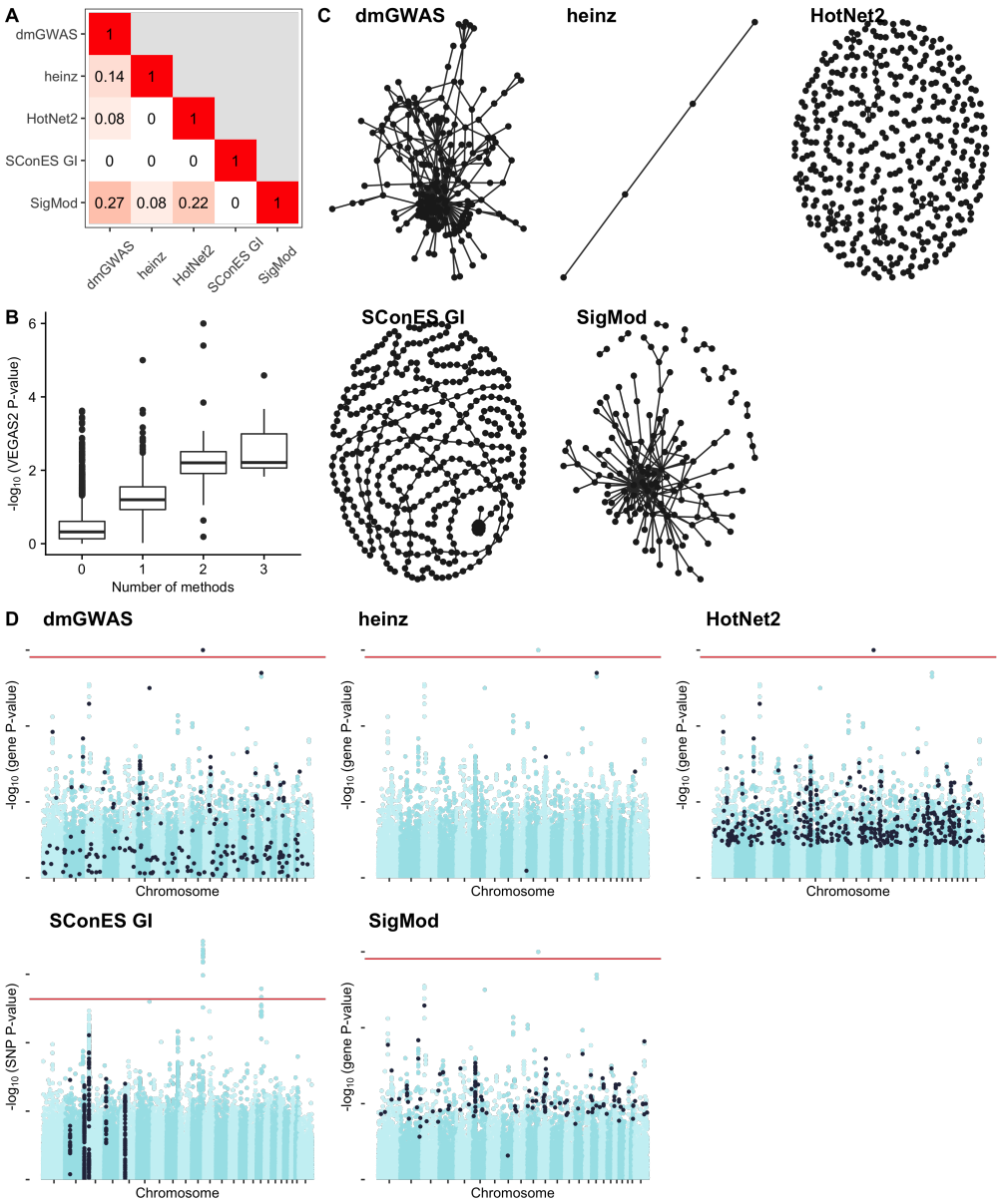

Fig 1. Overview of the solutions produced by the different network methods (Section 2.3.3) on the GENESIS dataset. As LEAN did not produce any significant solution (BH adjusted P-value < 0.05), it is not shown. Unless indicated otherwise, results refer to SNPs for SConES GI, and to genes for the other methods. **(A)** Overlap between the genes selected by each method, measured by Pearson correlation between indicator vectors (Sections 2.5.1 and 2.3.5). **(B)** Distribution of VEGAS2 P-values of the genes in the PPIN not selected by any network method (12 213), and of those selected by 1 (575), 2 (73), or 3 (20) methods. **(C)** Solution networks produced by the different methods. **(D)** Manhattan plots of SNPs/genes; in black, the method's solution. The red line indicates the Bonferroni threshold (2.54×10^{-7} for SNPs, 1.53×10^{-6} for genes).

Lastly, a pathway enrichment analysis (Section 2.4) also showed similarities and differences between the methods' solutions. It linked different parts of SigMod's

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

solution to four processes (S2 Table): protein translation (including mitochondrial), mRNA splicing, protein misfolding, and keratinization (BH adjusted P-values < 0.03). Interestingly, the dmGWAS solution (S3 Table) was also related to protein misfolding (*attenuation phase*, BH adjusted P-value $= 0.01$). However, it additionally included 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 [47]. As with SigMod, the genes in HotNet2's solution (S4 Table) were 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 were additional common themes. All obtained solutions had lower association P-values than the whole PPIN (median VEGAS2 P-value $\ll 0.46$, Table 2), despite containing genes with higher P-values as well (Fig 1D). This illustrates the trade-off between controlling for type I error 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 VEGAS2 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 solutions presented other desirable properties. First, four of them were enriched in known breast cancer susceptibility genes (dmGWAS, heinz, HotNet2, and SigMod, Fisher's exact test one-sided P-value < 0.03). Second, the genes in three solutions displayed, on average, a significantly higher betweenness centrality than the rest of the genes (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 [48], 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, the SNPs in SConES' solution were also more central than the average SNP (S1 Table), suggesting that causal SNPs are also more central than non-associated SNPs.

3.3 A case study: the consensus solution

Despite their shared properties, the differences between the solutions suggested that each of them captured different aspects of cancer susceptibility. Indeed, out of the 668 genes that were selected by at least one method, only 93 were selected by at least two, 20 by three, and none by four or more. Encouragingly, the more methods selected a gene, the higher its association score to the phenotype (Fig 1B), a relationship that plateaued at 2. Hence, to leverage their strengths and compensate for their respective weaknesses, we built a consensus solution using the genes shared among at least two solutions (Section 2.3.3). This solution (Fig 2) contained 93 genes and exhibited 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 solution also showed similar pathways as the individual solutions (S5 Table). 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 [49]. Interestingly, increased mitochondrial translation has been found in cancer cells [50], and its inhibition was proposed as a therapeutic target. With regards to the attenuation phase of heat shock response, it involved 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 mapped to any of these three genes, 9 mapped to all three, and 4 to two, which made it hard to disentangle their effects.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

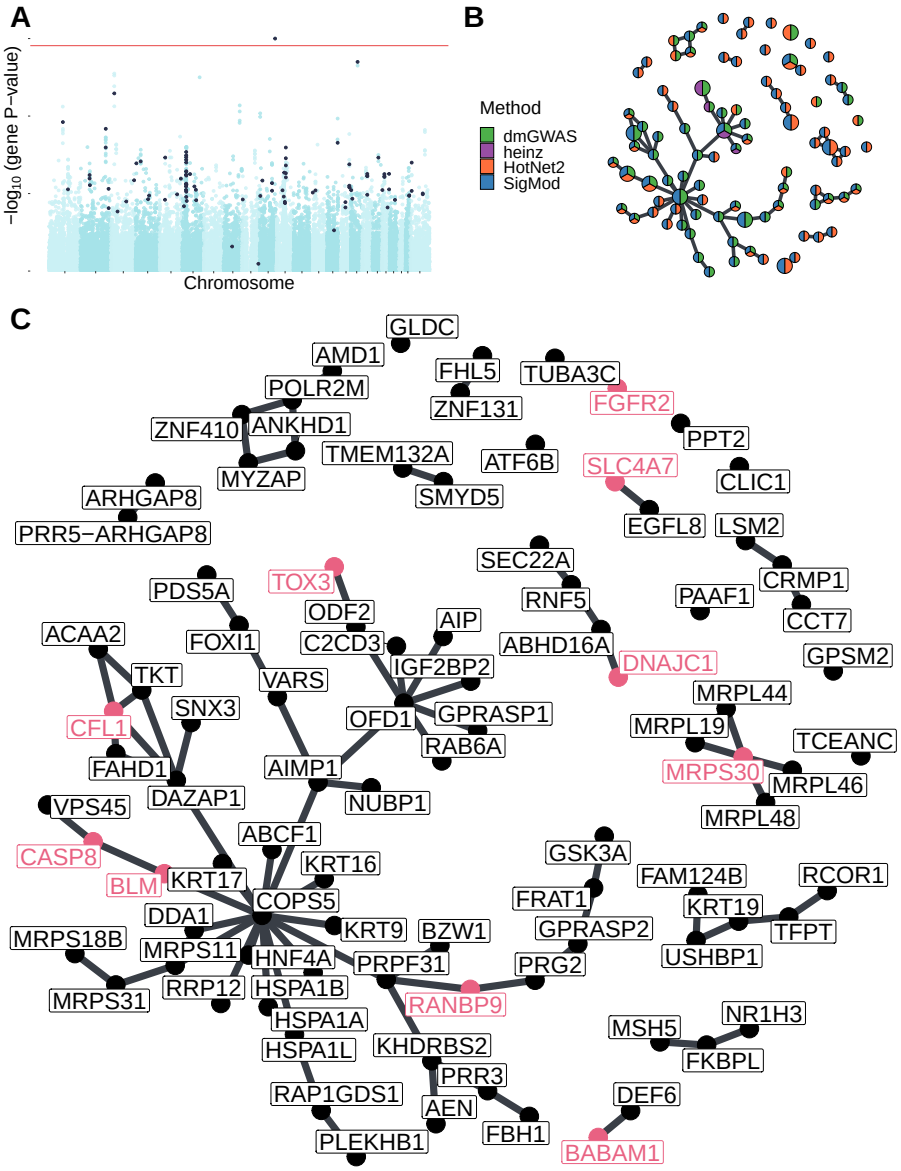

Fig 2. Consensus solution on GENESIS (Section 2.3.3). (A) Manhattan plot of genes; in black, the ones in the consensus solution. The red line indicates the Bonferroni threshold (1.53×10^{-6} for genes). **(B)** Consensus network. Each gene is represented by a pie chart, which shows the methods that selected it. We enlarged the two most central genes (*COPS5* and *OFD1*), the known breast cancer susceptibility genes, and the BCAC-significant genes (Section 2.6). **(C)** The nodes are in the same disposition as in panel B, but we indicated every gene name. We colored in pink the names of known breast cancer susceptibility genes and BCAC-significant genes.

HSPA1A was the most strongly associated gene (VEGAS2 P-value = 8.37×10^{-4}).

Topologically, the consensus consisted of a connected component composed of 49 genes and multiple smaller subnetworks (Fig 2B and C). Among the latter, 19 genes were in subnetworks containing a single gene or two connected nodes. This implied that they did not have a consistently altered neighborhood but were strongly associated

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

themselves and hence picked by at least two methods. The large connected component contained genes that are highly central in the PPIN. This property weakly anticorrelated with the P-value of association to the disease (Pearson correlation coefficient = -0.26, S3 Fig). This anticorrelation suggested that these genes were selected because they were on the shortest path between two high scoring genes. Because of this, we hypothesize that highly central genes might contribute to the heritability through alterations of their neighborhood, consistent with the omnigenic model of disease [6]. For instance, the most central node in the consensus solution was *COPS5*, a component of the COP9 signalosome that regulates multiple signaling pathways. *COPS5* is related to multiple hallmarks of cancer and is overexpressed in multiple tumors, including breast and ovarian cancer [51]. Despite its lack of association in GENESIS or in studies conducted by the Breast Cancer Association Consortium (BCAC) [41] (VEGAS2 P-value of 0.22 and 0.14, respectively), its neighbors in the consensus solution had consistently low P-values (median VEGAS2 P-value = 0.006).

3.4 Network methods boost discovery

We compared the results obtained with different network methods to the European sample of BCAC, the largest GWAS to date on breast cancer (Section 2.6). Although BCAC case-control studies do not necessarily target cases with a familial history of breast cancer like GENESIS does, this comparison is pertinent since we expect a shared genetic architecture at the gene level, at which most network methods operate. Together with BCAC's scale (90 times more samples than GENESIS), this shared genetic architecture provided a reasonable counterfactual of what we would expect if GENESIS had a larger sample size. We computed a gene association score on BCAC (Section 2.6). The solutions provided by the different network methods overlapped significantly with BCAC hits (Fisher's exact test P-value < 0.019). The gene-based methods achieved comparable precision (2%-25%) and recall (1.3-12.1%) at recovering BCAC-significant genes (S4 FigA). Interestingly, while SConES GI achieved a similar recall at the SNP-level (8.6%), it showed a much higher precision (47.3%).

3.5 Network methods share limitations

We compared the six network methods in a 5-fold subsampling setting (Section 2.5). In this comparison we measured properties (Fig 3 and S4 Fig): the size of the solution; the sensitivity and the specificity of an L1-penalized logistic regression classifier on the selected SNPs; the stability of the methods; and their computational runtime. The solution size varies greatly between the different methods (Fig 3A). Heinz produced the smallest solutions, with an average of 182 selected SNPs (Section 2.3.5) while the largest ones came from SConES GI (6 256.6 SNPs) and dmGWAS (4 255.0 SNPs). LEAN did not produce any solution in any of the subsamples.

To determine whether the selected SNPs could predict cancer susceptibility, we computed the classifiers' performances on *test sets* (S4 FigB). The different classifiers displayed similarly low 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 to minimize prediction error. Of course, although these performances were low, we did not expect to separate cases from controls well using exclusively genetic data [52].

Another desirable quality of a selection algorithm is the stability of the solution with respect to small changes in the input (Section 2.5.1). Heinz was highly stable in our benchmark, while the other methods displayed similarly low stabilities (Fig 3B).

In terms of computational runtime, the fastest method was heinz (Fig 3C), which returned a solution in a few seconds. HotNet2 was the slowest (3 days and 14 hours on

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

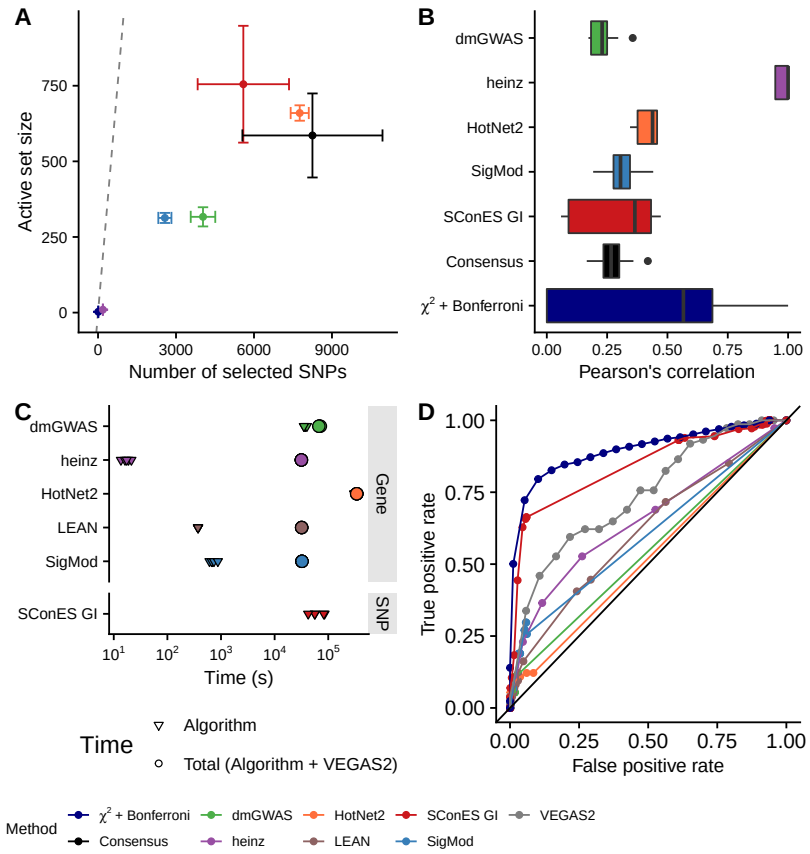

Fig 3. Comparison of network methods on GENESIS. Each method was run 5 times on a random subset containing 80% of the samples and tested on the remaining samples (Section 2.5). As LEAN did not select any gene, we excluded it from panels A and B. **(A)** Number of SNPs selected by each method and number of SNPs in the active set (i.e., the number of SNPs selected by the classifier, Section 2.5.2). Points are the average over the 5 runs; the error bars represent the standard error of the mean. A grey diagonal line with slope 1 is added for comparison, indicating the upper bound of the active set. For reference, the active set of the classifier using all the SNPs as input included, on average, 154 117.4 SNPs. **(B)** Pairwise Pearson correlations of the solutions produced by different methods. A Pearson correlation of 1 means the two solutions are the same. **(C)** Runtime of the evaluated methods, by type of network used (PPIN or SNP). For gene-based methods, inverted triangles represent the runtime of the algorithm alone, 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. **(D)** True positive rate and true negative rate of the methods, obtained using different parameter combinations (Section 3.7). We used as true positives BCAC-significant SNPs (for SConES and χ^2 + Bonferroni) and genes (for the remaining methods, Section 2.6). We used the whole dataset in this panel.

average). Including the time required to compute the gene scores, however, slowed down considerably gene-based methods; on this benchmark, that step took on average 1 day and 9.33 hours. Including this first step, it took 5 days on average for HotNet2 to produce a result.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Using different combinations of parameters (Section 2.3.4), we computed how good each of the methods was at recovering the results of a conventional GWAS on BCAC (Section 2.6, Fig 3D). SConES exhibits the largest area under the curve since, when $\lambda = 0$ (i.e., network topology is disregarded), it is equivalent to a Bonferroni correction. The remaining network methods have similar areas under the curve, with heinz having the largest one.

3.6 Network topology and association scores matter and might lead to ambiguous results

As shown above, and despite their similarities, the different ways of modeling the problem led to remarkably different solutions. Importantly, understanding which assumptions the methods made allowed us to understand the results more in depth. For instance, the fact that LEAN did not return any gene implied that there was no gene such that both itself and its environment were, on average, strongly associated with the disease.

In the GENESIS dataset, heinz's solution was very conservative, providing a small solution with the lowest median P-value (Table 2). By repeatedly selecting this compact solution, heinz was the most stable method (Fig 3B). Its conservativeness stemmed from its preprocessing step, which modeled 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 (S2 FigB), only 36 genes retained a positive score after that transformation. However, this small solution did not provide much insight into the susceptibility mechanisms to cancer. Importantly, it ignored genes that were associated with cancer in this dataset, like *FGFR2*.

On the other end of the spectrum, dmGWAS, HotNet2, and SigMod produced large solutions. DmGWAS' solution was the lowest scoring solution on average because of its greedy framework, which is biased towards larger solutions [53]. It considered all nodes at distance 2 of the examined subnetwork and accepted a weakly associated gene if it was linked to another, high scoring one. Aggregating the results of successive greedy searches exacerbates this bias, leading to a large, tightly connected cluster of unassociated genes (Fig 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 solution was even harder to interpret, being composed of 440 genes divided into 135 subnetworks. Lastly, SigMod missed some of the highest scoring breast cancer susceptibility genes in the dataset, like *FGFR2* and *TOX3*.

Another peculiarity of network methods was their relationship to degree centrality. We studied random rewirings of the PPIN that preserved node centrality (Section 2.7). In this setting, network methods favored central genes (Fig 4B) even though highly central genes often had no association to breast cancer susceptibility (Fig 4C). We found this bias especially in SigMod (S6 Fig), which selected three highly central, unassociated genes in both the PPIN and in many of the random rewirings: *COPS5*, *CUL3*, and *FN1*. However, as we showed in Section 3.3 and will show in 3.8, 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 [54], and it anticorrelates with poor prognosis in other cancer types [55, 56].

By using a SNP subnetwork, SConES analyzed each SNP in its functional context. Therefore, it could select SNPs located in genes not included in the PPIN and in non-coding regions. We compared the solution of SConES in the GI network (using PPIN information), to the one using only positional information (GS network) and to the one using positional and gene annotations (GM network). Importantly, SConES

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

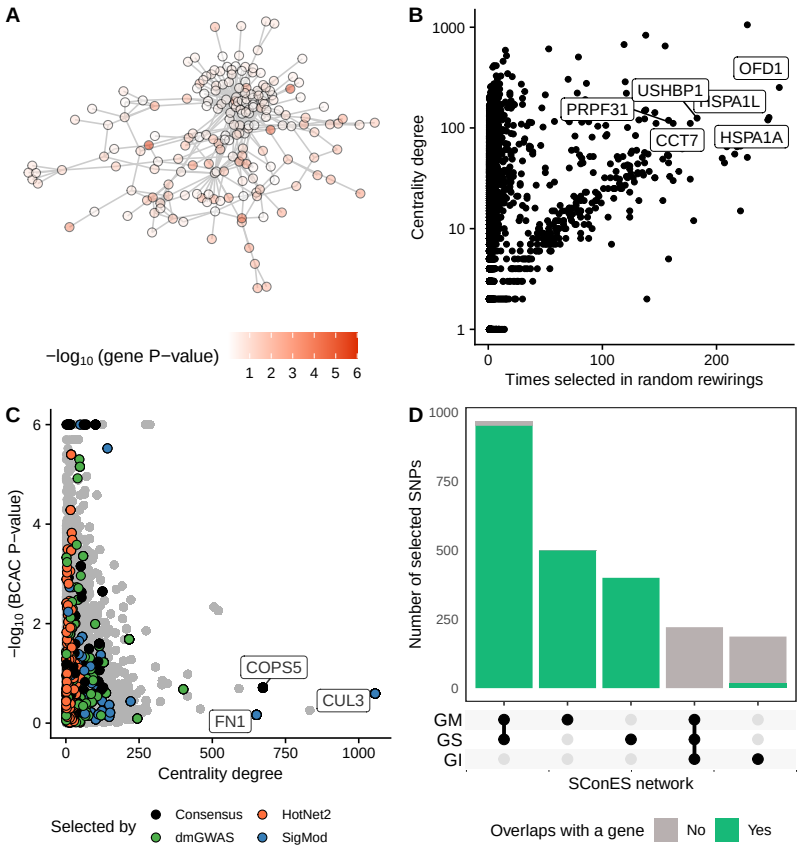

Fig 4. Drawbacks encountered when using network methods. (A) DmGWAS solution, with the genes colored according to the -$\log_{10}$ of their P-value. **(B)** Number of times a gene was selected by either dmGWAS, heinz, LEAN, or SigMod in 100 rewirings of the PPIN (Section 2.4) and its centrality degree. **(C)** 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. **(D)** Overlap between the solutions of SConES GS, GM, or GI. Barplots are colored based on whether the SNPs map to a gene or not (Section 2.3.5).

produced similar results on the GS and GM networks (S5 Fig). While the solutions on those two considerably overlap with SConES GI's, they contained additional gene-coding segments (Fig 4C). In fact, both SConES GS and GM selected chromosome regions related to breast cancer, like 3p24 (*SLC4A7/NEK10* [57]), 5p12 (*FGF10*, *MRPS30* [49]), 10q26 (*FGFR2*), and 16q12 (*TOX3*). In addition to those, SConES GS selected region 8q24, also linked to breast cancer (*POU5F1B* [58]).

3.7 Different parameters produce similarly-sized solutions

We explored methods' parameter space by running them under different combinations of parameters (Section 2.3.4). In agreement with their formulations (Section 2.3.3), larger values of specific parameters produced less stringent solutions (S7 FigA): for HotNet2 and heinz, this is the threshold above which genes receive a positive score; for dmGWAS, it is the `d` parameter, which controls how far neighbors could be added; for SigMod, it is `nmax`, which specifies the maximum size of the solution; and for LEAN, it

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is the P-value threshold to consider a solution significant. Two parameters had the opposite effect (the larger, the more stringent): SigMod's `maxjump`, which sets the threshold to consider an increment in λ "large enough"; and SConES' η, where higher values produce smaller solutions. However, two of the parameters did not have the expected effect: dmGWAS' `r`, which controls the minimum increment in the score required to add a gene; and SigMod's `maxjump`, which sets the threshold to consider an increment in λ "large enough". In both cases, the size of the solution was very similar across the different values. Despite the differences in size, the solutions' size was relatively robust to the choice of parameters (S7 FigB).

We computed the Pearson correlation between the different solutions as in Section 2.5.1 to study how the parameters affected which genes and SNPs were selected (S8 Fig). This analysis showed that dmGWAS and SigMod were robust to two parameters: the parameter `d` determined dmGWAS' output more than `r`; for SigMod, it was $nmax$ rather than $maxjump$.

SConES presented an interesting case in terms of feature selection: most of the explored combinations of parameters led to trivial solutions (they included either all the SNPs or none of them) (S8 Fig). To explore a more meaningful parameter space, we selected the parameters in two rounds in our experiments. First, we explored the whole sample space. Then, we focused on a range of η and λ 1.5 orders of magnitude above and below the best parameters, respectively. This second parameter space was more diverse, which allowed to find more interesting solutions.

3.8 Building a stable consensus network preserves global network properties

Most network methods, including the consensus, were highly unstable (Fig 3B), raising questions about the results' reliability. We built a new, *stable consensus* solution using the genes selected most often across the 30 solutions obtained by running the 6 methods on 5 different splits of the data (Section 2.5). Such a network should capture the subnetworks more often found altered, and hence should be more resistant to noise. We used only genes selected in at least 7 solutions, which corresponded to 1% of all genes selected at least once. The resulting stability-based consensus was composed of 68 genes (Fig 5). This network shared most of the properties of the consensus: breast cancer susceptibility genes were overrepresented (P-value $= 3 \times 10^{-4}$), 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 were more central than average (P-value $= 1.1 \times 10^{-14}$); and a considerable number of nodes (19) were isolated (Fig 5B and C).

Despite these similarities, the consensus and the stable consensus included different genes. In the stable consensus network, the most central gene was *CUL3*, which was absent from the previous consensus solution and had a low association score in both GENESIS and BCAC (P-values of 0.04 and 0.26, respectively). This gene was a component of Cullin-RING ubiquitin ligases. Encouragingly, it impacts the protein levels of multiple genes relevant for cancer progression [59], and its overexpression was also linked to increased sensitivity to carcinogens [60].

4 Discussion

In recent years, the GWAS' ability to unravel the mechanisms leading to complex diseases has been called into question [6]. First, the omnigenic model proposes that gene functions are interwoven in a dense co-function network. The practical consequence is that larger and larger GWAS will lead to discovering an uninformative wide-spread pleiotropy. Second, its conservative statistical framework hinders GWAS discovery.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

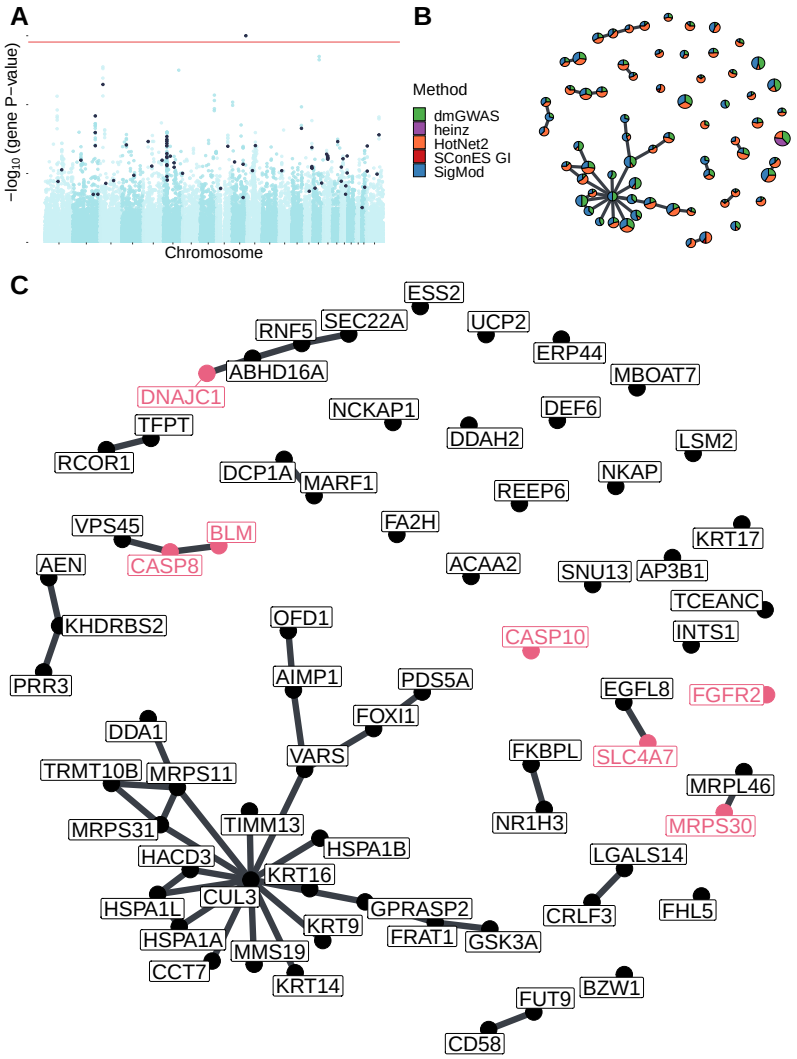

Fig 5. Stable consensus solution on GENESIS (Section 3.8). **(A)** Manhattan plot of genes; in black, the ones in the stable consensus solution. The red line indicates the Bonferroni threshold (1.53×10^{-6} for genes). **(B)** Stable consensus network. Each gene is represented by a pie chart, which shows the methods that selected it. We enlarged the most central gene (*CUL3*), the known breast cancer susceptibility genes, and the BCAC-significant genes (Section 2.6). **(C)** The nodes are in the same disposition as in panel B, but we indicated every gene name. We colored in pink the names of known breast cancer susceptibility genes and BCAC-significant genes.

Network methods elegantly address these two issues by using both association scores and an interaction network to consider 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 discovery have been proposed in recent years. In this article we evaluated six of these methods (Section 2.3.3) by applying them to the GENESIS study, a GWAS dataset on familial breast cancer (Section 2.1).

DmGWAS, Heinz, HotNet2, SConES, and SigMod all yielded compelling solutions, which include (but are not limited to) known breast cancer susceptibility genes (Section 3.2). In general, the selected genes and SNPs were more central than most

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

other genes and SNPs, agreeing with the observation that disease genes are more central [48]. 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 network methods (Section 3.6). However, we found support in the literature for the involvement of the selected highly central genes (*COPS5*, *FN1*, and *CUL3*) in breast cancer susceptibility (Sections 3.3, 3.6, and 3.8). Despite these similarities, the methods' solutions were notably different. At one end of the spectrum, SConES and heinz preferred high scoring solutions, which were also small and hence did not shed much light on the disease's etiology. On the other end, dmGWAS, HotNet2 and SigMod gravitated towards lower scoring but larger solutions, which provided a wide overview of the biological context. While this deepened our understanding of breast cancer susceptibility and provided biological hypotheses, they required further analyses. For instance, we examined the centrality of the selected genes to understand how much that property was driving their selection (Section 3.6). However, all solutions shared two drawbacks. First, they were all equally bad at discriminating cases from controls. However, the classification accuracy of network methods was similar to that of a classifier trained on the entire genome (Section 3.5), which suggests that cases and controls are difficult to separate in the GENESIS dataset. This may be due to limited statistical power, which reduces the ability to identify relevant SNPs. However, in any event, we do not expect to separate people who have or will develop cancer from others on the sole basis of their genomes, ignoring all environmental factors and chance events. Hence, network methods were preferable to the logistic regression classifier since they did "no worse" at classification while providing an interpretable solution. Second, all methods were remarkably unstable, yielding different solutions for slightly different inputs. This might partly have been caused by the instability of the P-values themselves in low statistical power settings [61]. Hence, heinz's conservative transformation of P-values, which favored only the most extreme ones, led to improved stability. Another source of instability might have been the redundancy inherent to biological networks, a consequence of an evolutionary pressure to avoid single points of failure [62]. Hence, biological networks will often have multiple paths connecting two high-scoring nodes.

To overcome these limitations while exploiting the each method's strengths, we proposed combining them into a consensus solution. We used the straightforward strategy of including any node that was recovered by at least two methods. We thus proposed two solutions (Sections 3.3 and 3.8): a consensus solution, which addressed the heterogeneity of the solutions, and a stable consensus solution, which addresseded the instability of the methods. They both included the majority of the strongly associated smaller solutions and captured genes and broader mechanisms related to cancer, thus synthesizing the mechanisms altered in breast cancer cases. Thanks to their smaller size and network structure, they provided compelling hypotheses on genes like *COPS5* and *CUL3*, which lack genome-wide association with the disease but are related to cancer at the expression level and consistently interact with high scoring genes. Notably, while the consensus approach was as unstable as the individual network-guided methods, the stable consensus network retained the ability to provide compelling hypotheses and had better stability. This supported that redundant but equivalent mechanisms might cause instability and supported the conclusions obtained on the individual solutions.

In this work, we have compared our results to significant genes and SNPs in the BCAC study [41]. Network methods showed modest precision but much higher recall at recovering BCAC hits (Section 3.4). While precision might be desirable when a subset of useful markers is required (for instance, for diagnosis), higher recall is desirable in exploratory settings. Nonetheless, BCAC was not an ideal ground truth. First, the studied populations are not entirely overlapping: BCAC focused on a pan-European cohort, while GENESIS targeted the French population. Second, the study designs

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

differed: a high proportion of breast cancer cases investigated in BCAC were sporadic (not selected according to family history), while GENESIS was a homogeneous dataset not included in BCAC focused on the French high-risk population attending the family cancer clinics. Finally, and this is indeed the motivation for this study, GWAS are unlikely to identify all genes relevant for the disease: some might only show up in rare-variant studies; others might have too small effect sizes. Network methods account for this by including genes with low association scores but with relevant topological properties. Hence, network methods and GWAS, even when well-powered, are unlikely to capture exactly the same sets of genes. This might partly excuse the low precisions displayed in Section 3.4 and the low AUC displayed in Section 3.5.

As not all PPIN databases compiled the same interactions, the choice of the PPIN determines the final output. In this work, we used only interactions from HINT from high-throughput experiments (Section 2.3.2). This responded to concerns about adding interactions identified in targeted studies and falling into "rich getting richer" problems: since popular genes have a higher proportion of their interactions described [10, 24], they might bias discovery towards themselves by reducing the average shortest path length between two random nodes. On the other hand, Huang et al. [12] found that larger networks were more useful than smaller networks to identify disease genes. This would support using the largest networks in our experiments. However, when we compared the impact of using a larger PPIN containing interactions from both high-throughput experiments and the literature (Section 2.3.2), for most of the methods it did not change much the size or the stability of the solution, the classification accuracy, or the runtime (S10 Fig). This supports using only interactions from high-throughput experiments, which produced similar solutions and avoided falling into "circular reasonings", where the best-known genes were artificially pushed into the solutions, as we observed in Section 3.6.

The strength of network-based analyses comes from leveraging prior knowledge to boost discovery. In consequence, they show their shortcomings on understudied genes, especially those not in the network. Out of the 32 767 genes to which we mapped the genotyped SNPs, 60.7% (19 887) were not in the PPIN. The majority of those (14 660) are non-coding genes, mainly lncRNA, miRNA, and snRNA (S9 Fig). Nevertheless, RNA genes like *CASC16* were associated to breast cancer (Section 3.1), reminding us of the importance of using networks beyond coding genes. Besides, even protein-coding genes linked to breast cancer susceptibility [57], like *NEK10* (P-value 1.6×10^{-5}, overlapping with *SLC4A7*) or *POU5F1B*, were absent from the PPIN. However, on average protein-coding genes absent from the PPIN were less associated with breast cancer susceptibility (Wilcoxon rank-sum P-value $= 2.79 \times 10^{-8}$, median P-values of 0.43 and 0.47). This could not be due to well-known genes having more known interactions because we only used interactions from high-throughput experiments. As disease genes tend to be more central [48], we hypothesize that it was due to interactions between central genes being more likely. It is worth noting that network methods that do not use PPIs, like SConES GS and GM, recovered SNPs in *NEK10* and *CASC16*. Moreover, both SConES GM and GI recovered intergenic regions, which might contain key regulatory elements [63], but 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, and other strategies (e.g., eQTLs, SNPs associated to gene expression) might lead to different results.

A crucial step for the gene-based methods is the computation of gene scores. In this work, we used VEGAS2 [22] due to the flexibility it offers to use user-specified gene annotations. However, it presents known problems: selection of an appropriate

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

percentage of top SNPs, long runtimes and P-value precision limited to the number of permutations [29]). Additionally, other algorithms like PEGASUS [29], SKAT [64] or COMBAT [65] might have more statistical power.

How to handle linkage disequilibrium (LD) is often a concern among GWAS practitioners. Often, the question is whether an LD-based pruning of the genotypes will improve the results. 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. In Section 3.3 we highlighted ambiguities that appear when genes overlap or are in LD. The presented case is paradigmatic since all three genes are in the HLA region, the most gene-dense region of the genome [66]. Network methods are prone to selecting such genes when they are functionally related, and hence interconnected in the PPIN. But the opposite case is also true: when genes are not functionally related (and hence disconnected in the PPIN), network methods might disregard them even if they have high association scores. With regards to SConES, fewer SNPs would lead to simpler SNP networks and, possibly, shorter runtimes. However, LD patterns also affect SConES' in other ways, since its formulation penalizes selecting a SNP and not its neighbors, via a nonzero parameter η in Eq 5. Due to LD, nearby SNPs' P-values correlate; since positional information determines SNP networks, nearby SNPs are likely to be connected. Hence, SConES tends to select LD-blocks formed by low P-value SNPs. This might explain why SConES produced similar results on the GS and GM networks, heavily affected by LD (Section 3.6). However, this same behavior raises the burden of proof required to select 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. This could be caused by the absence of joint association of a gene and most of its neighbors, a hypothesis supported by LEAN's lack of results. Yet, a different combination of parameters could lead to a more informative SConES' solution (e.g., a lower λ in Eq 5), although it is unclear how to find it. In addition, due to the design of the iCOGS array (Section 2.1), 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 understudied — hindering the accuracy with which the GS network captures the genome structure. A strong LD pruning might address such problems.

To produce the two consensus solutions, we faced practical challenges due to the differences in interfaces, preprocessing steps, and unexpected behaviors of the various methods. To make it easier for others to apply these methods to new datasets and aggregate their solutions, we built six `nextflow` pipelines [67] with a consistent interface and, whenever possible, parallelized computation. They are available on GitHub: hclimente/gwas-tools (Section 2.9). Importantly, we compiled those methods with a permissive license into a Docker image for easier use, available on Docker Hub hclimente/gwas-tools.

5 Supporting information

S1 Table. Summary statistics on the results of SConES on the three SNP networks (Section 2.3.2). The first row within each block contains the summary statistics on the whole network.

S2 Table. Pathway enrichment analyses of the genes in SigMod's solution.

S3 Table. Pathway enrichment analyses of the genes in dmGWAS' solution.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S4 Table. Pathway enrichment analyses of the genes in HotNet2's solution.

S5 Table. Pathway enrichment analyses of the genes in the consensus' solution.

S1 Fig. 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.

S2 Fig. Association in GENESIS. The red lines represent the Bonferroni thresholds. (A) SNP association, measured from the outcome of a 1 d.f. χ^2 allelic test (Section 2.2). Significant SNPs 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 [22] using the 10% of SNPs with the lowest P-values (Section 2.2). The Bonferroni threshold is 1.53×10^{-6}. **(C)** SNP association as in panel (A). The SNPs in black were selected by an L1-penalized logistic regression (Section 2.5.2, $\lambda = 0.03$).

S3 Fig. Relationship between the $\log_{10}$ of the betweenness centrality and the -$\log_{10}$ of the VEGAS2 P-value of the genes in the consensus solution. The blue line represents a fitted generalized linear model.

S4 Fig. Additional benchmarks of the network methods. (A) 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. This panel is identical to Fig 2. **(B)** Sensitivity and specificity on the test set of the L1-penalized logistic regression trained on the features selected by each of the methods. The performance of the classifier trained on all SNPs is also displayed. Points are the average over the 5 runs; the error bars represent the standard error of the mean.

S5 Fig. Pearson correlation between the different solutions. (A) Correlation between selected SNPs. **(B)** Correlation between selected genes. In general, the solutions display a very low overlap.

S6 Fig. Number of times a gene was selected by either dmGWAS, heinz, LEAN, or SigMod in 100 rewirings of the PPIN (Section 2.7) and its centrality degree. This figure is equivalent to Fig 4B, split by method.

S7 Fig. Size of the solutions obtained under different parameters. (A) Size of the solution produced by different parameter values, expressed as a percentage of the maximum solution size for the method, or the highest tested value for the parameter, respectively. The size of the solution is the median among all the solution sizes for the same parameter. **(B)** Boxplot of the solution sizes of the methods under the explored parameters (Section 2.3.4).

S8 Fig. Pearson correlation between the solutions obtained under different parameters, computed as in Section 2.5.1. Grey tiles represent the cases where we could not compute the Pearson correlation because the two vectors were either all ones (all genes/SNPs were selected) or zeros (no genes/SNPs were selected).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S9 Fig. Biotypes of genes from the annotation that are not present in the HINT PPIN.

S10 Fig. Comparison of the benchmark on high-throughput (HT) interactions to the 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 (Section 2.5.2). **(B)** Shifts in sensitivity and specificity. **(C)** Shift in Pearson correlation between benchmarks. **(D)** Ratio between the runtimes of the benchmarks. For gene-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 improved the stability of the solution. However, it increased the solution size and the required runtime, and had mixed effects on the sensitivity and specificity.

S11 Fig. Overview of the solutions produced by the SConES on the GS and GM networks (Section 2.3.2) on the GENESIS dataset. (A) Manhattan plots of SNPs (Section 2.2); in black, the method's solution. The red line indicates the Bonferroni threshold (2.54×10^{-7}). **(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.

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: S. Eon-Marchais, M. Marcou, D. Le Gal, L. Toulemonde, J. Beauvallet, N. Mebirouk, E. Cavaciuti), the biological resource center (S. Mazoyer, F. Damiola, L. Barjhoux, C. Verny-Pierre, V. Sornin). We wish to pay tribute to Olga M. Sinilnikova, one of the initiators and principal investigators of GENESIS, and who died prematurely on June 30, 2014.

We thank all the GENESIS collaborating cancer clinics clinics (Clinique Sainte Catherine, Avignon: H. Dreyfus; Hôpital Saint Jacques, Besançon: M-A. Collonge-Rame; Institut Bergonié, Bordeaux: M.Longy, A. Floquet, E. Barouk-Simonet; CHU, Brest: S. Audebert; Centre François Baclesse, Caen: P. Berthet; Hôpital Dieu, Chambéry: S. Fert-Ferrer; Centre Jean Perrin, Clermont-Ferrand: Y-J. Bignon; Hôpital Pasteur, Colmar: J-M. Limacher; Hôpital d'Enfants CHU – Centre Georges François Leclerc, Dijon: L. Faivre-Olivier; CHU, Fort de France: O. Bera; CHU Albert Michallon, Grenoble: D. Leroux; Hôpital Flaubert, Le Havre: V. Layet; Centre Oscar Lambret, Lille: P. Vennin, C. Adenis; Hôpital Jeanne de Flandre, Lille: S. Lejeune-Dumoulin, S. Manouvier-Hanu; CHRU Dupuytren, Limoges: L. Venat-Bouvet; Centre Léon Bérard, Lyon: C. Lasset, V. Bonadona; Hôpital Edouard Herriot, Lyon: S. Giraud; Institut Paoli-Calmettes, Marseille: F. Eisinger, L. Huiart; Centre Val d'Aurelle – Paul Lamarque, Montpellier: I. Coupier; CHU Arnaud de Villeneuve, Montpellier: I. Coupier, P. Pujol; Centre René Gauducheau, Nantes: C. Delnatte; Centre Catherine de Sienne, Nantes: A. Lortholary; Centre Antoine Lacassagne, Nice: M. Frénay, V. Mari; Hôpital Caremeau, Nîmes: J. Chiesa; Réseau Oncogénétique Poitou Charente, Niort: P. Gesta; Institut Curie, Paris: D. Stoppa-Lyonnet, M. Gauthier-Villars, B. Buecher, A. de Pauw, C. Abadie, M. Belotti; Hôpital Saint-Louis, Paris: O. Cohen-Haguenauer; Centre

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Viggo-Petersen, Paris: F. Cornélis; Hôpital Tenon, Paris: A. Fajac; GH Pitié Salpétrière et Hôpital Beaujon, Paris: C. Colas, F. Soubrier, P. Hammel, A. Fajac; Institut Jean Godinot, Reims: C. Penet, T. D. Nguyen; Polyclinique Courlancy, Reims: L. Demange*, C. Penet; Centre Eugène Marquis, Rennes: C. Dugast*; Centre Henri Becquerel, Rouen: A. Chevrier, T. Frebourg, J. Tinat, I. Tennevet, A. Rossi; Hôpital René Huguenin/Institut Curie, Saint Cloud: C. Noguès, L. Demange*, E. Mouret-Fourme; CHU, Saint-Etienne: F. Prieur; Centre Paul Strauss, Strasbourg: J-P. Fricker, H. Schuster; Hôpital Civil, Strasbourg: O. Caron, C. Maugard; Institut Claudius Regaud, Toulouse: L. Gladieff, V. Feillel; Hôpital Bretonneau, Tours: I. Mortemousque; Centre Alexis Vautrin, Vandoeuvre-les-Nancy: E. Luporsi; Hôpital de Bravois, Vandoeuvre-les-Nancy: P. Jonveaux; Gustave Roussy, Villejuif: A. Chompret*, O. Caron). *Deceased prematurely

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.

References

1. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Computational Biology. 2012;8(12):e1002822. doi: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;47(D1):D1005–D1012. doi:10.1093/nar/gky1120.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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;101(1):5–22. doi:10.1016/j.ajhg.2017.06.005.

4. Wang MH, Cordell HJ, Van Steen K. Statistical methods for genome-wide association studies. Seminars in Cancer Biology. 2018;doi:10.1016/j.semcancer.2018.04.008.

5. Barton NH, Etheridge AM, Véber A. The infinitesimal model: Definition, derivation, and implications. Theoretical Population Biology. 2017;118:50–73. doi:10.1016/j.tpb.2017.06.001.

6. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169(7):1177–1186. doi:10.1016/j.cell.2017.05.038.

7. Furlong LI. Human diseases through the lens of network biology. Trends in Genetics. 2013;29(3):150–159. doi:10.1016/j.tig.2012.11.004.

8. Leiserson MD, Eldridge, Jonathan V, Ramachandran, Sohini, Raphael, Benjamin J. Network analysis of GWAS data. Current Opinion in Genetics & Development. 2013;23(6):602–610.

9. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011;12(1):56–68. doi:10.1038/nrg2918.

10. Cai JJ, Borenstein E, Petrov DA. Broker Genes in Human Disease. Genome Biology and Evolution. 2010;2:815–825. doi:10.1093/gbe/evq064.

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

12. 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;6(4):484–495.e5. doi:10.1016/j.cels.2018.03.001.

13. 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; p. btw676. doi:10.1093/bioinformatics/btw676.

14. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. Bioinformatics. 2011;27(1):95–102. doi:10.1093/bioinformatics/btq615.

15. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24(13):i223–i231. doi:10.1093/bioinformatics/btn161.

16. 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;47(2):106–114. doi:10.1038/ng.3168.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

17. Azencott CA, Grimm D, Sugiyama M, Kawahara Y, Borgwardt KM. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics. 2013;29(13):i171–i179. doi:10.1093/bioinformatics/btt238.

18. 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; p. btx004. doi:10.1093/bioinformatics/btx004.

19. 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;16(1):13. doi:10.1186/s12885-015-2028-9.

20. Sakoda LC, Jorgenson E, Witte JS. Turning of COGS moves forward findings for hormonally mediated cancers. Nature Genetics. 2013;45(4):345–348. doi:10.1038/ng.2587.

21. 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;4(1):7. doi:10.1186/s13742-015-0047-8.

22. Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. Twin Research and Human Genetics. 2015;18(1):86–91. doi:10.1017/thg.2014.79.

23. 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;47(D1):D766–D773. doi:10.1093/nar/gky955.

24. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Systems Biology. 2012;6(1):92. doi:10.1186/1752-0509-6-92.

25. Wang Q, Jia P. dmGWAS 3.0; 2014. `https://bioinfo.uth.edu/dmGWAS/`.

26. 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;105(2-3):427–449. doi:10.1007/s10107-005-0660-x.

27. Beisser D, Klau GW, Dandekar T, Muller T, Dittrich MT. BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics. 2010;26(8):1129–1130. doi:10.1093/bioinformatics/btq089.

28. Dittrich M, Beisser D. BioNet; 2008. `https://bioconductor.org/packages/BioNet/`.

29. Nakka P, Raphael BJ, Ramachandran S. Gene and Network Analysis of Common Variants Reveals Novel Associations in Multiple Complex Diseases. Genetics. 2016;204(2):783–798. doi:10.1534/genetics.116.188391.

30. Scheid S, Spang R. twilight; a Bioconductor package for estimating the local false discovery rate. Bioinformatics. 2005;21(12):2921–2922. doi:10.1093/bioinformatics/bti436.

31. Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al.. HotNet2; 2018. `https://github.com/raphael-group/hotnet2`.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

32. Gwinner F. LEANR; 2016. https://cran.r-project.org/web/packages/LEANR/.

33. Climente-González H, Azencott CA. martini; 2019. https://www.bioconductor.org/packages/martini/.

34. Liu Y. SigMod v2; 2018. https://github.com/YuanlongLiu/SigMod.

35. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2019;doi:10.1093/nar/gkz1031.

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

37. 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;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

38. 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. Available from: http://link.springer.com/10.1007/978-3-319-46227-1_28.

39. 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;100(2):421–425. doi:10.1038/sj.bjc.6604830.

40. 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;45(D1):D833–D839. doi:10.1093/nar/gkw943.

41. 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;551(7678):92–94. doi:10.1038/nature24284.

42. 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;526(7571):68–74. doi:10.1038/nature15393.

43. 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;13(6). doi:10.1186/bcr3052.

44. 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;132(5):523–536. doi:10.1007/s00439-013-1269-4.

45. 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;12(1):156. doi:10.1186/1471-2350-12-156.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

46. 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;41(5):585–590. doi:10.1038/ng.354.

47. Nielsen FC, van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. Nature Reviews Cancer. 2016;16(9):599–612. doi:10.1038/nrc.2016.72.

48. 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;6(1):24570. doi:10.1038/srep24570.

49. 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;8(2):273–284. doi:10.1016/j.molonc.2013.11.008.

50. 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;471(4):639–645. doi:10.1016/j.bbrc.2016.02.072.

51. 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;9:135. doi:10.3389/fphar.2018.00135.

52. de los Campos G, Vazquez AI, Hsu S, Lello L. Complex-Trait Prediction in the Era of Big Data. Trends in Genetics. 2018;34(10):746–754. doi:10.1016/j.tig.2018.07.004.

53. Nikolayeva I, Guitart Pla O, Schwikowski B. Network module identification—A widespread theoretical bias and best practices. Methods. 2018;132:19–25. doi:10.1016/j.ymeth.2017.08.008.

54. 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;38(18):2362–2370. doi:10.1016/s0959-8049(02)00210-1.

55. 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;36(6):3145–3153. doi:10.3892/or.2016.5177.

56. 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;431:123–132. doi:10.1016/j.mce.2016.05.007.

57. 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;41(5):585–590. doi:10.1038/ng.354.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.04.076661; this version posted October 27, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

58. 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;94(3):395–404. doi:10.1016/j.ajhg.2014.01.019.

59. Chen HY, Chen RH. Cullin 3 Ubiquitin Ligases in Cancer Biology: Functions and Therapeutic Implications. Frontiers in Oncology. 2016;6. doi:10.3389/fonc.2016.00113.

60. 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;8(8):2432–2440. doi:10.1158/1535-7163.mct-08-1186.

61. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nature Methods. 2015;12(3):179–185. doi:10.1038/nmeth.3288.

62. Wagner A, Wright J. Alternative routes and mutational robustness in complex regulatory networks. Biosystems. 2007;88(1-2):163–172. doi:10.1016/j.biosystems.2006.06.002.

63. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. The American Journal of Human Genetics. 2018;102(5):717–730. doi:10.1016/j.ajhg.2018.04.002.

64. 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;92(6):841–853. doi:10.1016/j.ajhg.2013.04.015.

65. Wang M, Huang J, Liu Y, Ma L, Potash JB, Han S. COMBAT: A Combined Association Test for Genes Using Summary Statistics. Genetics. 2017;207(3):883–891. doi:10.1534/genetics.117.300257.

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

67. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature Biotechnology. 2017;35(4):316–319. doi:10.1038/nbt.3820.