

A network approach to prioritizing susceptibility genes for genome-wide association studies

Somayeh Kafaie  | Yuanzhu Chen | Ting Hu 

Department of Computer Science,
Memorial University, St. John's, NL,
Canada

Correspondence

Ting Hu, Department of Computer
Science, Memorial University, St. John's,
NL A1B 3X5, Canada.
Email: ting.hu@mun.ca

Funding information

Natural Sciences and Engineering
Research Council of Canada, Grant/
Award Numbers: Discovery Grant
RGPIN-2016-04699, Discovery Grant
RGPIN-2017-05201

Abstract

The heritability of complex diseases including cancer is often attributed to multiple interacting genetic alterations. Such a non-linear, non-additive gene-gene interaction effect, that is, *epistasis*, renders univariable analysis methods ineffective for genome-wide association studies. In recent years, network science has seen increasing applications in modeling epistasis to characterize the complex relationships between a large number of genetic variations and the phenotypic outcome. In this study, by constructing a statistical epistasis network of colorectal cancer (CRC), we proposed to use multiple network measures to prioritize genes that influence the disease risk of CRC through synergistic interaction effects. We computed and analyzed several global and local properties of the large CRC epistasis network. We utilized topological properties of network vertices such as the edge strength, vertex centrality, and occurrence at different graphlets to identify genes that may be of potential biological relevance to CRC. We found 512 top-ranked single-nucleotide polymorphisms, among which *COL22A1*, *RGS7*, *WWOX*, and *CELF2* were the four susceptibility genes prioritized by all described metrics as the most influential on CRC.

KEYWORDS

centrality, complex networks, epistasis, gene-gene interaction, GWAS

1 | INTRODUCTION

Studies show that susceptibility to complex diseases such as cancer is multifactorial, and both environmental and genetic factors play an important role (Bookman et al., 2011; Thomas, 2010). In recent years, genome-wide association studies (GWAS) (Hirschhorn & Daly, 2005; Wang, Barratt, Clayton, & Todd, 2005) have identified many genetic attributes and their association to complex human disease (Hunter & Kraft, 2007). GWAS have granted access to a significant number of single-nucleotide polymorphisms (SNPs) and, usually by designing case-control studies, aim to highlight a set of SNPs statistically associated with a disease.

Many studies confirm that the association of individual SNPs with complex phenotypes is usually of small effect size, and the heritability of diseases cannot be explained by any single SNP variants (Cordell, 2009; Thomas, 2010). Therefore, the estimates of the causal effects of genes or any other somatic variants should be studied carefully because of epistasis or non-linear interaction effect among multiple genetic attributes (Cordell, 2002; Moore, 2003; Phillips, 2008; Verma et al., 2018; Wilkins, Cannataro, Shuch, & Townsend, 2018). Epistasis has been realized to contribute significantly to the complex relationship between genetic and phenotypic variations, especially in cancer (Im et al., 2018; Park & Lehner, 2015). Studying epistasis in GWAS context, SNP interactions associated with a particular

complex trait, is significantly compute-intensive and challenging due to the large search space in GWAS (Manduchi, Chesi, Hall, Grant, & Moore, 2018).

To model the complex relationship among this massive number of attributes and their epistasis, the modern network science (Barabási, 2016; Newman, 2002), emerged recently, proposes a promising framework. Network science can be applied to a vast variety of topics in biology, such as protein–protein interaction networks (Rual et al., 2005; Stelzl, 2005), genetic regulatory networks (Carninci, 2005), metabolic networks (Duarte et al., 2007; Hu et al., 2018; Jeong, Tombor, Albert, Oltvai, & Barabasi, 2000), RNA networks (Lewis, Burge, & Bartel, 2005), gene coexpression networks (Stuart, Segal, Koller, & Kim, 2003), and gene–gene interaction networks (Beltrao, Cagney, & Krogan, 2010; Boone, Bussey, & Andrews, 2007). Network science provides required mathematical models and tools to predict the performance of such complex networks with a significant number of interactions involving genes, proteins, and metabolites (Barabási, 2016), and their properties and behaviors can be analyzed by studying certain attributes of networks such as hub size, degree distribution, node centrality, connectedness, and clustering coefficient. In fact, it facilitates the structured representation of pairwise interactions of genetic attributes and modeling their association with complex phenotypic traits (Hu & Moore, 2013).

In this study, we proposed to use network analysis to prioritize disease susceptibility genes. We constructed a colorectal cancer (CRC) epistasis network based on information gain (Cover & Thomas, 2006; Fan et al., 2011). By treating genetic factors and phenotypic traits as random variables, information-theoretic measures can be used to quantify the synergistic effect among genetic attributes. In our network, vertices indicate the SNPs while an edge between two of them represents an interaction relationship significantly ($p \leq 0.01$) stronger than a threshold. We studied several global and local properties of our CRC epistasis network such as the degree distribution, centrality measures, assortativity coefficient, and clustering coefficient, and found local structures of the network (e.g., graphlets and motifs), to rank SNPs based on their topological importance in the epistasis network. Through functional annotation and literature search, we identified a set of genes that might be of potential biological relevance to CRC.

2 | METHODS

2.1 | CRC GWAS data set and preprocessing

CRC is the third most common cancer globally, and more than 1.2 million new cases of CRCs are diagnosed

worldwide each year. This type of cancer caused more than 600,000 deaths around the world in 2008 (Ahmedin et al., 2011), and more than 52,000 deaths in the United States in 2015 (Siegel, Miller, & Jemal, 2018). CRC is a disease of genome instability, and the accumulation of genetic and epigenetic alterations transforming colonic cells into adenocarcinoma cells is its driving force (Cao et al., 2015). Therefore, studying the genetic component of CRC can guide us to its better understanding and identifying its predictive and influential genes leading to more efficient diagnoses, treatment, and even prevention (Dorani, Hu, Woods, & Zhai, 2018).

The data set used in this study was obtained from Newfoundland Familial Colorectal Cancer Registries (NFCCR), and included 265,195 SNPs. The study participants included 656 cases and 496 controls, where the CRC cases, diagnosed during 1999–2003, were identified through the population tumor registry maintained by the Newfoundland Cancer Registry. Both case and control participants were from 20 to 74 years old, although cases were slightly older than controls (mean, 62.7 for cases and 60.5 for controls; Sun et al., 2011). Controls were selected through random digit dialing (Wang et al., 2009). For genotyping a custom Affymetrix genome-wide platform, called the Axiom CORECT Set, was used including about 1.3 million SNPs with insertions and deletions (indels) on two physical genotyping chips (pegs) (Schumacher et al., 2015).

We conducted quality control on both “per-individual” and “per-marker” basis to maximize the number of markers remaining in the study (Anderson et al., 2010). In per-individual quality control, individuals with discordant sex information were identified, the sex chromosome was removed, and individuals with elevated missing data rates or outlying heterozygosity rate were identified. Per-marker quality control consisted of removing substandard SNPs and identifying SNPs with a significantly different ($p < 0.00001$) missing data rate between cases and controls. Finally, dependent SNPs were identified and pruned (i.e., linkage disequilibrium pruning with correlation coefficient $r^2 > 0.6$, linkage disequilibrium window size 2,000, and step size 200), and we removed all SNPs that their genotype for at least 1% of the individuals is missing. Regarding the rest of missing genotypes, a frequency-based method was used, in which any missing value of an individual was filled with the most common genotype of the corresponding SNP in the population (Hu et al., 2011). After quality control and linkage disequilibrium pruning, the pre-processed data set included 190,142 SNPs from 626 cases and 472 controls. Then, by removing SNPs with at least 1% missing individuals, 185,180 SNPs left from 1,098 individuals.

2.2 | Filtering

Because analyzing such a large data set with many thousands SNPs for interactions is impractical in terms of computational cost, we decided to apply feature selection or filtering techniques to reduce the number of SNPs to about 10,000. Because in filtering we are interested in selecting SNPs with strong interactions to construct the network, we decided to choose Relief-based algorithms which are also fast and flexible. In general, Relief-based algorithms are filtering methods to reduce the number of features in the data set by assigning scores to individual features based on their discriminant power. They are only filter methods known with the ability to detect feature dependencies without an exhaustive examination of all feature combinations (Urbanowicz, Meeker, Cava, Olson, & Moore, 2018; Urbanowicz, Olson, Schmitt, Meeker, & Moore, 2018).

ReliefF (Kira & Rendell, 1992) is a well-know filtering method that assigns a weight to each SNP representing the relevance of the SNP to the disease status, where weights are estimated using genetically similar individuals. In fact, for any individual r , the nearest neighbors are found and, depending on their disease status (i.e., case or control), are separated in two groups of *hit* (i.e., the same status as r) and *miss* (i.e., different status than r). Then, the weight of SNPs is increased with respect to the distance between r and miss members (i.e., such SNPs are more predictive of the disease), and is decreased in terms of the distance between r and hit members (i.e., they are less predictive of the disease).

The difference of two individuals R_i and R_j regarding SNP a is measured as

$$\text{diff}(a, R_i, R_j) = \begin{cases} 0 & \text{genotype}(a, R_i) = \text{genotype}(a, R_j) \\ 1 & \text{otherwise,} \end{cases}$$

and the distance between two individuals can be calculated as

$$\text{dist}(R_i, R_j) = \sum_{\forall a \in A} \text{diff}(a, R_i, R_j),$$

where A denotes the set of all SNPs. Thus, the nearest neighbor of each individual is an individual with the most number of SNPs of the same genotype.

Spatially uniform ReliefF (SURF) (Greene, Penrod, Kiralis, & Moore, 2009) is an extension of ReliefF that, instead of choosing a constant number of nearest neighbors, selects all neighbors within a fixed distance T of the individual. Furthermore, it has been recommended to apply an iterative Relief approach such as

tuned ReliefF (TURF) in large feature space with more than 10,000 features (Urbanowicz et al., 2018). TURF algorithm (Moore & White, 2007) improves the performance of ReliefF by running it several times. To re-estimate the weight and relevance of remaining SNPs more accurately, each time TURF removes noisy SNPs that have the lowest weight values. The combination of TURF with both SURF and ReliefF has shown promising results regarding finding the SNPs with the strongest pairwise interactions (Dorani & Hu, 2018).

To choose between SURF and ReliefF, we implemented the combination of TURF and each of them, scored SNPs by them, selected about top 10,000 SNPs and calculated the distribution of pairwise information gain (Cover & Thomas, 2006) for SNPs selected by each. Because we were looking for SNPs with stronger interactions, and the result showed that the pairwise interaction in the SNPs selected by TURF + SURF were stronger than those selected by TURF + ReliefF, the combination of TURF and SURF, with the pseudocode shown in Figure 1, was applied to choose the top SNPs. After filtering, the data set used in network construction and our analysis consisted of 9,996 SNPs from 626 CRC cases and 472 controls.

2.3 | Pairwise interaction quantification

We measured the strength of the interaction between any pair of SNPs in terms of information gain (IG) (Andrew et al., 2012; Fan et al., 2011; Hu et al., 2013b, 2013b;

```

P_r = 0.01
Snp_th = 10000
while (|A| > snp_th) do
  for i=1 to |R| do
    H = {r ∈ R | ph(r)=ph(R_i) and dist(r,R_i)<T}
    M = {r ∈ R | ph(r)≠ph(R_i) and dist(r,R_i)<T}
    foreach a in A do
      foreach h in H do
        W[a] = W[a] - diff(a,R_i,h)/( |R| * |H| )
      endfor
      foreach m in M do
        W[a] = W[a] + diff(a,R_i,m)/( |R| * |M| )
      endfor
    endfor
  endfor
  sort attributes in terms of their weight W
  n_remove = |A| * p_r
  remove n_remove SNPs with the least weight from A
endwhile

```

FIGURE 1 The implementation of TURF + SURF. A and R denote the set of SNPs (i.e., attributes) and individuals, respectively. snp_th represents the maximum number of SNPs supposed to be kept after filtering, T is the mean distance between all individuals considered as the threshold to find nearby individuals, $\text{ph}(x)$ returns the phenotype of individual x (i.e., case or control), and p_r denotes the percentage of SNPs removed at each iteration of TURF. SNP: single-nucleotide polymorphism; SURF: spatially uniform ReliefF; TURF: tuned ReliefF

Moore & Hu, 2015; Pan et al., 2014). This metric quantifies the amount of the disease status that the epistatic effect of the corresponding genotype pair represents. The IG between SNPs A and B with class variable C , denoting status case or control, can be measured as:

$$IG(A, B; C) = I(A, B; C) - I(A; C) - I(B; C), \quad (1)$$

where $I(A; C)$ denotes the mutual information of SNP A 's genotype and the disease status C , that is, the main effect of A on C , and $I(A, B; C)$ measures the explanation of C by combining A and B . Thus, $IG(A, B; C)$ captures the synergistic, non-additive effect between A and B on explaining C . Mutual information $I(A; C)$ can be calculated as follows.

$$I(A; C) = H(C) - H(C|A), \quad (2)$$

where $H(C) = \sum_c p(c) \times \log_{\frac{1}{p(c)}}$ is the entropy of C , $H(C|A) = \sum_{a,c} p(a, c) \times \log_{\frac{p(a)}{p(a,c)}}$ is the conditional entropy of C given knowledge of SNP A , $p(c)$ is the probability that an individual has class c , $p(a)$ is the frequency of individuals with genotype a , and $p(a, c)$ is the frequency of individuals in either the case or the control group that carry genotype a .

2.4 | Epistasis network construction

A network or graph G is defined as a set of vertices $V(G) = \{v_1, v_2, \dots, v_n\}$, where $|V(G)|$ denotes the number of vertices in the network, and a set of edges $E(G)$, that its members are two-element subsets of V (West, 2001). In our epistasis network, each SNP is represented by a vertex, and the edges demonstrate the interactions among pairs of SNPs measured in terms of pairwise IG. Because IG measures the amount of the disease status explained by the corresponding genotype pair interaction for each edge, a stronger effect from the interaction between two SNPs translates into a higher weight of their corresponding edge into the network.

To assess the significance of the strength of these SNP pairwise interactions (i.e., IG values), we generated 1,000 permutations of the CRC GWAS data set by randomly shuffling the disease status and assigning them to the samples to remove the association between genotypes and phenotype. Then, the pairwise IG values for the SNPs in all permuted datasets were similarly measured, and the significance p -value of IG of SNP pairs in real data set was evaluated using the null hypothesis of no association between the genotypes and the phenotype.

In the final constructed epistasis network, we defined an edge between two vertices if (a) the calculated pairwise IG between the corresponding SNPs was greater than a threshold (IG-cut-off), and (b) its p -value, drawn from permutation test, was less than or equal to 0.01. To decide on the optimal threshold of IG, we used several IG-cut-off values (from the highest observed in the data set decremented by 0.001) to construct a network and observed the change of different features like the number of vertices, the number of edges, and the size of the largest connected component. Obviously, a network constructed based on a smaller threshold always includes all vertices and edges of the one based on a larger threshold. We decreased IG-cut-off gradually until a dominant connected component (i.e., a connected component with the majority of vertices) appeared in the network. Therefore, with this cutoff, we included the majority of the SNPs and their strongest and most significant pairwise interactions for the subsequent network analyses.

2.5 | Ranking SNPs in the epistasis network

After constructing our epistasis network, represented as a weighted and undirected graph, we analyzed it based on several global and local network properties. The definition of all these properties has been presented in Section S1. We also found *motifs* (Milo et al., 2002; Shen-Orr, Milo, Mangan, & Alon, 2002) representing particular patterns at which SNPs interact with each other in a more frequent and meaningful way in comparison to a random network. Furthermore, we found small connected non-isomorphic induced subgraphs of our network, called graphlets (Pržulj, Corneil, & Jurisica, 2004; Pržulj, 2006), and their orbits representing different distinct positions of vertices in a graphlet. Based on these analytical findings, we used the following three important local properties of the vertices to rank SNPs based on their topological importance in our epistasis network and chose top SNPs with potential high disease association: (a) SNP pairwise interaction strength, that is, IG, (b) vertex centrality measures, and (c) appearances of SNPs at different orbits of graphlets.

3 | RESULTS

3.1 | Statistical epistasis network of CRC

As discussed in Section 2.4, by decreasing IG-cut-off value from 0.02 with a decrement of 0.001, we were looking for the threshold at which the dominant

connected component appears. As shown in Table 1, from IG-cut-off = 0.02 to IG-cut-off = 0.015, the size of the largest connected component is very small in comparison to the total number of vertices, and SNPs are connected in small groups. However, at IG-cut-off = 0.014, the majority of the SNPs (~57%) are connected to each other (3,244 out of 5,683 SNPs) with 5,006 edges while more than half of SNPs are non-isolated vertices (5,683 out of 9,996 SNPs). Then, at IG-cut-off = 0.013, the number of vertices and edges grow to 7,950 and 10,720, respectively with more than 94% of the nodes in the largest connected component (7,504 SNPs). At IG-cut-off = 0.011 and after that the network is fully connected (i.e., the largest connected component includes all vertices and the network consists of one single connected component), and IG-cut-off = 0.009 is the point that finally all 9,996 SNPs are attached to the network. Based on these results, we chose IG-cut-off = 0.014 as the threshold to pick the edges and construct our epistasis network to include the majority of the SNPs with a minimal set of the strongest and the most significant pairwise interactions.

Because IG measured between two SNPs represents how much this SNP pair's interaction effect is associated with the disease, we chose IG as a metric to find SNP pairs with the strongest interaction. In fact, we calculated the mean (μ) and the standard deviation (σ) for all pairwise IG values, and selected SNP pairs with their IG values greater than $\mu + 3\sigma$. Table S3 presents the list of top SNP pairs with the highest IG values along with their associated genes extracted from NCBI¹.

3.2 | Global properties of the network

With the derived thresholds of $IG > 0.014$ and $p \leq 0.01$, our CRC epistasis network consisted of 5,683 vertices and 5,006 edges. Each vertex denotes a SNP and each edge represents a strong and significant interaction between the corresponding endpoint SNPs in terms of IG. Table 2 summarizes the basic parameters of the network.

- *Degree distribution and connected components:* In our network, each vertex has a degree between 1 and 8 with average 1.762 neighbors per vertex. Based on the degree distribution shown in Figure 2, the majority of SNPs interact only with one other SNP while there are a small number of hubs interacting with eight other SNPs directly. In addition, the largest connected component in our network included 3,244 vertices and other

vertices grouped in significantly small connected components of 2–36 vertices, as shown in Figure S1.

- *Shortest path and diameter:* Based on small-world effect, the average path length in networks typically scale as $\log|V(G)|$ (Newman, 2010), and because of the slow-growing of logarithm function in terms of its argument, the value of average path length usually remains small even for large networks. However, for our epistasis network, the average path length equals 20.996, which is significantly larger than $\log(5681) = 3.75$. Figure S2 presents the distribution of the shortest path length in the network. In addition, the diameter was measured as 58 for our network.
- *Assortativity coefficient:* The calculated assortativity coefficient for our network is about -0.044 , which classify it as a “disassortative mixing” network. It means that in our network, there is a tendency for high-degree SNPs to be attached to low-degree ones. As explained in (Newman, 2002), the technological and biological networks studied like Internet, World Wide Web, protein interactions, neural network, and food web all have disassortative mixing, high-degree vertices preferentially connect with low-degree ones and vice versa. In general, it has been proven that disassortative networks percolate harder and the giant connected component will appear slower at them. Also, they are more vulnerable in case of attack to the high-degree nodes (i.e., hubs) because these hubs are distributed

TABLE 1 The comparison of networks constructed based on various IG-cut-off values. Note that while at larger IG-cut-off values most of vertices were either isolated or in small groups, by decreasing the IG-cut-off value the network grew, more vertices were connected, and finally, at IG-cut-off = 0.014 the dominant connected component emerged

IG-cut-off	Number of vertices	Number of edges	Size of the LCC
0.020	118	60	3
0.019	276	140	3
0.018	536	277	3
0.017	1050	568	5
0.016	1975	1152	8
0.015	3518	2392	25
0.014	5683	5006	3244
0.013	7951	10724	7505
0.012	9319	22532	9273
0.011	9827	47329	9827
0.010	9979	99668	9979
0.009	9996	197300	9996
0.008	9996	327353	9996

Note. IG: information gain; LCC: largest connected component.

¹<https://www.ncbi.nlm.nih.gov/snp>

TABLE 2 Basic parameters of our epistasis network, where each edge represents an IG value >0.014 between its endpoints (and $p \leq 0.01$)

Parameter	Value
Number of vertices	5683
Number of Edges	5006
Clustering coefficient	0
Number of connected components	761
Network diameter	58
Average path length	20.9957433
Average number of neighbors	1.761
Assortativity coefficient	-0.043694176

Note. IG: information gain.

broadly across the network linking low-degree nodes and parts of the network (Barabási, 2016; Newman, 2002). To measure the significance of the calculated value, we performed a set of significance tests in which the assortativity coefficient for 1,000 permuted networks was calculated. These permuted networks had the same number of vertices and edges as our network; however, their neighborhood structure is permuted by swapping edges $10 \times |E(G)|$ times (Hu, Andrew, Karagas, & Moore, 2013a). Every time, two edges, $e_{i,j}$ and $e_{h,k}$, are chosen randomly and given that there is no edge between the endpoints of these two edges (i.e., $e_{i,k}, e_{h,j} \notin E(G)$), the endpoints are swapped such that the edges $e_{i,j}$ and $e_{h,k}$ in the network are replaced with the edges $e_{i,k}$ and $e_{h,j}$. The results showed that only one

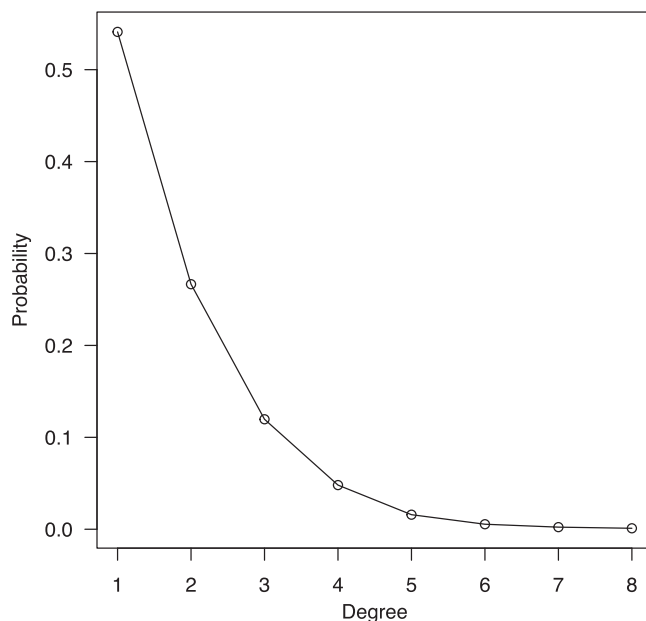


FIGURE 2 The degree distribution of our epistasis network. The degree distribution of our network does not show a heavy-tail degree distribution as in scale-free networks

random network (out of 1,000 random networks) had assortativity coefficient equal to or less than that of our network, that is, $p = 0.001$, which shows the strength of disassortative mixing in our epistasis network. Table S1 summarizes assortativity coefficient measured for the networks constructed by different IG-cut-off values and p -values measured by generating 1,000 random networks as explained here.

- *Clustering coefficient*: Surprisingly the value of clustering coefficient for our network is 0; meaning that there is no triangle in the network or the neighbors of a vertex are never neighbor themselves. As discussed in Newman, 2010, the expected clustering coefficient for a network with given number of vertices and degree distribution but with random connections between vertices can be calculated as

$$E(\text{clustering-coefficient}) = \frac{1}{|V(G)|} \frac{[\langle k^2 \rangle - \langle k \rangle]^2}{\langle k \rangle^3}, \quad (3)$$

where $\langle k \rangle$ denotes the mean of the degree distribution and $\langle k^2 \rangle$ is the mean for the square of the degrees (i.e., second moment). While the clustering coefficient of our network is 0, its expected value equals 0.0002. Table S2 summarizes the clustering coefficient measured for the networks constructed by different IG-cut-off values, their expected clustering coefficient using (3) and p -value measured by generating 1,000 random networks as explained for assortativity coefficient.

- *K-core components*: While $k = 1$ represents all connected components in the network that are 761 components, for $k = 2$ we found only one component consisted of 658 vertices that each is connected to at least two other vertices of the component. However, our network does not show a dense core because the k -core does not go beyond $k = 2$, meaning that there exists no subset of vertices in which each of them is connected to at least three other vertices.

3.3 | Local properties of the network

Apart from global properties, local properties concern individual vertices and their nearby neighborhood.

- *Motifs and graphlets*: We used FANMOD (Wernicke & Rasche, 2006) to find the frequent patterns of our network with up to eight vertices, and the extracted motifs are shown in Figure 3. Furthermore, the graphlets found in our network with 3–5 vertices are shown in Figure 4. The results prove that in our network there is no loop with the length of 5 or less. We

used ORCA (Hočevár & Demšar, 2014) to count the number of times that each SNP appears in different orbits across all graphlets, and generated table *snp_orbit* (number of SNPs \times number of orbits), where *snp_orbit*(i, j) denotes the number of occurrences of SNP i at orbit j . Then, for each orbit, we selected 50 SNPs with the highest occurrences at the given orbit. We noticed that from 73 possible orbits described in (Pržulj, 2006), only 16 orbits are available in the structure of our network. In addition, most selected SNPs are common among the list of top SNPs of several orbits. The list of these 146 SNPs, as well as their associated genes and the list of orbits for which the SNP was chosen (i.e., was ranked among 50 top SNPs), have been summarized in Table S5.

- **Centrality measures:** To find the vertices that are more important or more “central,” we measured several centrality metrics such as degree, closeness, and betweenness centrality and page rank for all vertices of the network. Degree centrality is one of the features to identify the most influential nodes, and it is usually reasonable to assume that SNPs with more and stronger connections (i.e., hubs) may have more influence than those who have fewer and weaker connections (Albert & Barabási, 2002; Bonacich, 1972). Furthermore, the vertices with high betweenness are usually connecting two parts of the network and are seen as the bottlenecks of information flow (Barabasi, Gulbahce, & Loscalzo, 2011; De, Hu, Moore, & Gilbert-Diamond, 2015). To take into account

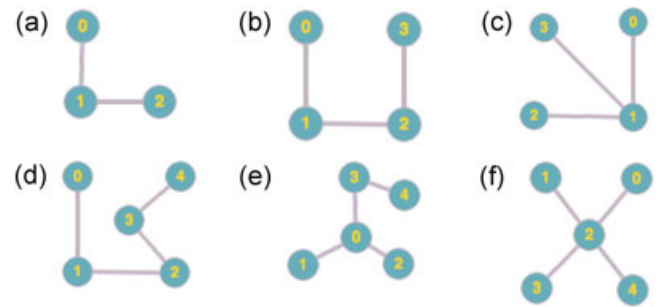


FIGURE 4 The graphlets extracted by ORCA with (a) three, (b-c) four, and (d-f) five vertices. There are 29 possible graphlets, with three to five vertices, and our network contains only six

the effect of the pairwise IG calculated for SNPs, we here only discuss the *weighted* centrality metrics. The histograms of all centrality metrics calculated for our network are shown in Figure 5. We calculated the pairwise correlation between centrality metrics measured for vertices of our network. As expected and shown in Figure 5, it represents a strong correlation between weighted page rank and weighted degree centrality. Then, we found outlier SNPs that have high ranks by multiple centrality measures. In fact, we ranked SNPs based on each of four centrality measures, separately. Then, for each SNP, we calculated the sum of its ranks for all four metrics (called aggregated rank), and then sorted SNPs in ascending order of their aggregated rank. We calculated the mean (μ) and the standard deviation (σ) of aggregated

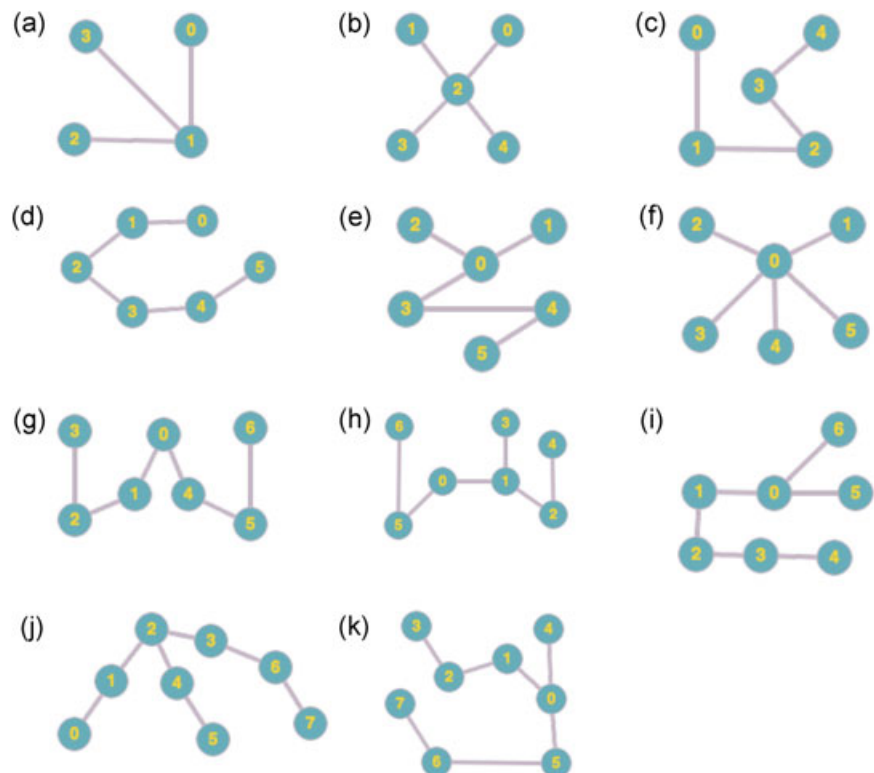


FIGURE 3 The motifs extracted by FANMOD with (a) four, (b-c) five, (d-f) six, (g-i) seven, and (j-k) eight vertices. The tool did not find any motif with three vertices

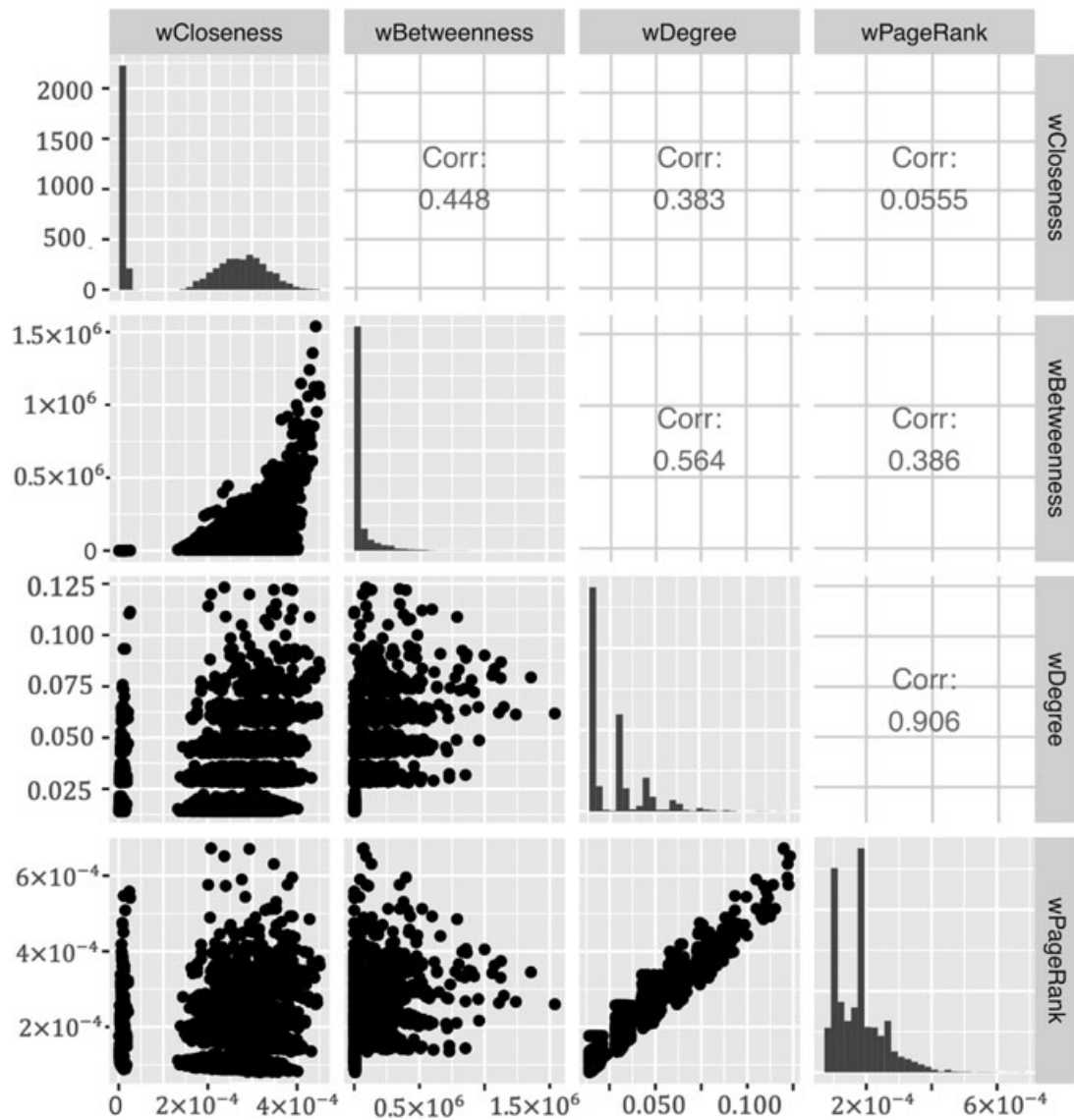


FIGURE 5 The pairwise correlation between different weighted degree centrality, weighted closeness centrality, weighted betweenness centrality, and weighted page rank as well as their histogram

rank across all SNPs, and selected the SNPs that their aggregated rank was $< \mu - 2\sigma$. Table S7 represents these top SNPs and their corresponding genes. We noticed that while the majority of the best SNPs are in the largest connected component, a few are not. All top SNPs that are not in the largest connected component have high weighted page rank and weighted degree centrality but the low closeness and betweenness centralities. In fact, due to the way that we measured closeness and betweenness centrality, these metrics will be larger for the nodes in the largest connected component in comparison to those which are in other (very small) components. For example for closeness centrality, the existence of a path to any other node can increase the value of the metric; hence more vertices available in the largest connected component can contribute to a higher closeness centrality value for the

vertices of the largest connected component. Figure S3 shows the topology of the components of the top SNPs not in the largest connected component. We used Cytoscape (Shannon et al., 2003) to visualize these structures, and in each topology, the yellow vertex represents the SNP which was in the list of high-rank SNPs.

3.4 | Top SNPs based on different metrics

As discussed before, we selected top SNPs based on three different criteria. (a) The highest pairwise IG, (b) most occurrences at different orbits of graphlets, and (c) the highest value of centrality measures. We also found the list of associated genes for the SNPs selected

by each criterion; 153, 155, and 80 genes for criterion 1, 2, and 3, respectively. Figure 6 shows the number of top SNPs extracted from coding regions and noncoding regions based on the three metrics as well as the Venn diagrams representing the number of common coding SNPs, noncoding SNPs and genes selected based on the metrics. The list of top SNPs and their corresponding genes, chosen by at least two (of the three) criteria, is summarized in Table 3. Also, the lists of top SNPs extracted from noncoding regions which have been enriched as Expression quantitative trait loci (eQTLs).² for IG, graphlet and centrality metrics have been provided in Tables S4, S6, and S8, respectively. Furthermore, Figure 7 presents a network of the two-hop neighborhood of these 25 top SNPs. While in our epistasis network, all these top SNPs are in the largest connected component, as can be seen in Figure 7, the top SNPs here are separated into 12 different connected components. In fact, there exist nine components with only one top SNP, meaning that these top SNPs are not connected to any other top SNP in up to 5-hop neighborhood.

3.4.1 | Validating the significance of selected top SNPs

To further validate the significance of these selected SNPs regarding the disease, we developed machine learning models to measure the accuracy of disease prediction using the top SNPs. In fact, we trained four models based on logistic regression with stochastic gradient descent, K -nearest neighbors, support vector machine, and random forests using 50 top SNPs as the explanatory features. These top SNPs were selected based on each of our suggested metrics individually as well as their combination, called Union. To compare the results, we also selected 50 top SNPs based on TURF + SURF filtering approach, explained in Section 2.2, and allelic odds ratio (Clarke et al., 2011), explained in Section S1. Each classification model was trained using 90% of the data set, while a k -fold cross-validation with random search was used to select hyperparameters of the model. Then, we measured the accuracy of predicting disease in the test set (the other 10% of the data set). This process was repeated 10 times for each metric and model, and the average accuracy calculated for different metrics in the models has been shown in Figure 8.

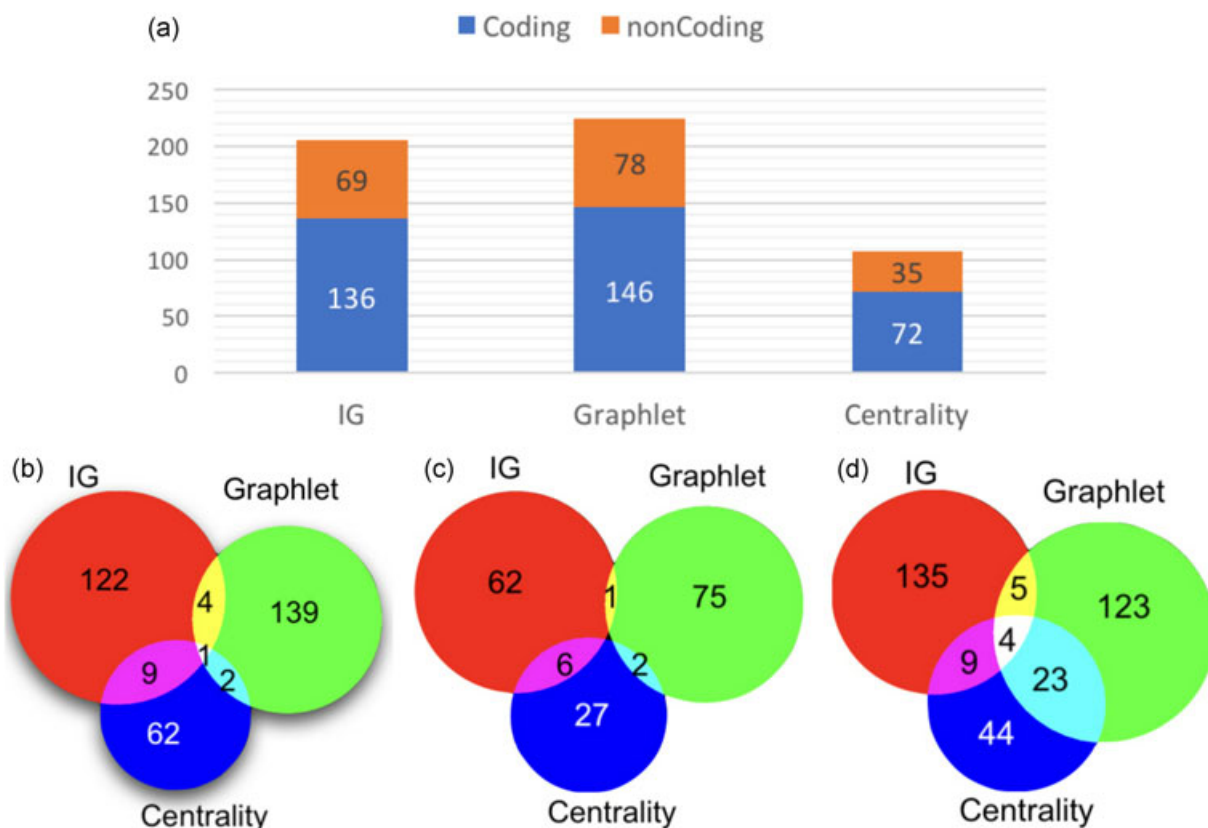


FIGURE 6 (a) The number of top SNPs extracted from coding and noncoding regions based on IG, graphlets and centrality measures. (b-d) Venn diagram for three different criteria (IG, graphlet and centrality) representing the number of (b) SNPs in coding region, (c) SNPs from noncoding region, and (d) genes

TABLE 3 Top SNPs chosen based on at least two of the following three criteria: (a) The highest pairwise IG, (b) most occurrences at different orbits of graphlets, and (c) the highest value of vertex centrality measures

SNP	Gene	Criteria
rs6983378	<i>COL22A1</i>	IG, graphlet, centrality
rs7012042	<i>COL22A1</i>	IG, centrality
rs1160595	<i>NRXN1</i>	IG, graphlet
rs12493550	<i>HTR3D</i>	IG, graphlet
rs3793695	<i>CRTAC1</i>	IG, graphlet
rs6854489	<i>LOC105377448</i>	IG, graphlet
rs11757878	-	IG, graphlet
rs10071657	<i>LOC101929710</i>	IG, centrality
rs751150	<i>MYT1L</i>	IG, centrality
rs16841104	<i>RGS7</i>	IG, centrality
rs3851997	<i>CPNE4</i>	IG, centrality
rs3826616	<i>SERPINB8</i>	IG, centrality
rs9430004	<i>CAMK1G</i>	IG, centrality
rs835484	<i>CHST11</i>	IG, centrality
rs6716943	<i>LOC105373962</i>	IG, centrality
rs13107574	-	IG, centrality
rs7160402	-	IG, centrality
rs1890629	-	IG, centrality
rs2907639	-	IG, centrality
rs7628760	-	IG, centrality
rs1381574	-	IG, centrality
rs1978153	<i>ABCC3</i>	Graphlet, centrality
rs9310213	<i>FOXP1</i>	Graphlet, centrality
rs6743932	-	Graphlet, centrality
rs10829973	-	Graphlet, centrality

Note. IG: information gain.

As shown in Figure 8, the SNPs selected based on the combination of our three suggested metrics (i.e., Union) predict the disease with higher accuracy than the SNPs selected by other metrics in almost all models. Furthermore, while in SVM model the performance of TURF + SURF and Odds ratio is very close to Union, in other models, Union outperforms them significantly. Interestingly, in KNN model, the performance of both TURF+SURF and Odds ratio is even worse than that of each of our suggested metrics individually. Also, these results show that Centrality is more promising in identifying key SNPs than IG and graphlet, and in SVM its performance is even close to Union.

3.4.2 | Main versus interaction effects of the top SNPs

One of the advantages of using IG to calculate pairwise interactions among SNPs is that it is able to exclude individual main effect of SNPs in calculation of higher-order interaction effects. To verify that, we have calculated the main effect of all top SNPs chosen by our three criteria and found their rank among all SNPs selected for network construction ($n = 9996$). This ranking can be found as column “Main effect rank” in Tables S3, S5, and S7. Furthermore, we have selected 100 top SNPs chosen based on our three metrics and found their main effect ranking among all SNPs selected for network construction after filtering ($n = 9,996$). Table 4 presents the number of SNPs (of these top 100 SNPs) selected based on our three different criteria (i.e., IG, graphlet, and centrality) that are among top 1%, 5%, 10%, and 50% SNPs ranked based on the main effect. As can be seen in this table, only a small fraction of our top SNPs are among SNPs with high main effect, which can indicate that IG, used here to construct the network and create the edges, is able to exclude the main effect in measuring pairwise interactions (as shown in Equation (1)), and capture the cases with only high pairwise interactions.

4 | DISCUSSION

To identify the synergistic effect of multiple genes on phenotypic status, we studied the interaction of SNPs in CRC. Our data set included 265,195 SNPs from 656 cases and 496 controls. After applying quality control, data imputation and filtering, by calculating the pairwise IG between remained 9,996 SNPs, we constructed our epistasis network, where an edge links two SNPs if IG associated with its endpoints is significantly ($p \leq 0.01$) greater than a given threshold. We started from the highest measured IG as the threshold, and kept decreasing the value until a giant connected component appeared in the network at IG-cut-off = 0.014. The final epistasis network consisted of 5,683 SNPs as the vertices and 5,006 edges, and we analyzed several global and local properties of the network.

In our network, the vertex degree does not go beyond 8, and most SNPs interact only with at most a couple of other SNPs. Also, the average path length is significantly greater than the expected value based on small-world effect, and with high significance ($p = 0.001$), it is a disassortative mixing network in which high-degree SNPs tend to attach to low-degree ones. All these findings as well as the facts that our network’s clustering coefficient

²eQTL is a locus that explains a fraction of the genetic variance of a gene expression phenotype (Nica & Dermitzakis, 2013).

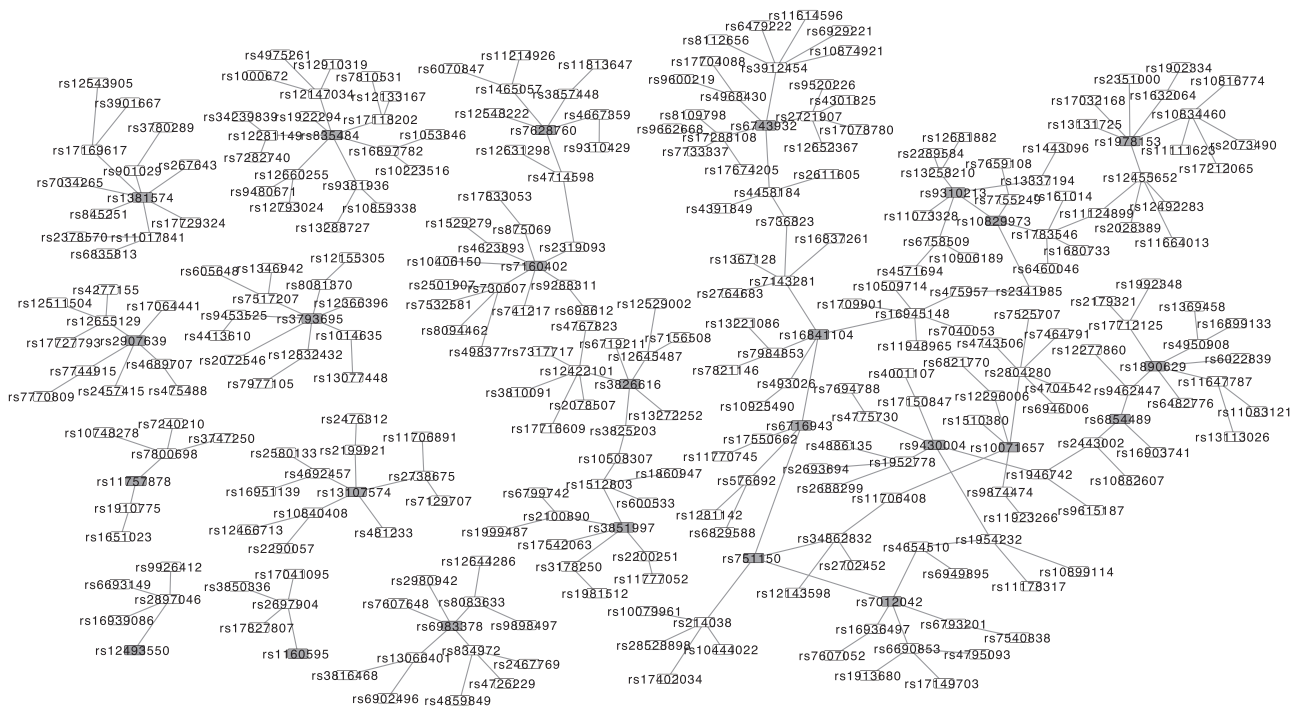


FIGURE 7 The network of all top SNPs chosen based on at least two of the three criteria and their two-hop neighborhood. Top SNPs are highlighted in gray. This network consists of 299 vertices connected by 287 edges in 12 separate connected components

is 0, and there is no k -core component with $k > 2$, confirms that it has a sparse and tree-like structure in which most SNPs interact with at most 2 other SNPs to create chain or star shape.

In addition, to identify important vertices, we computed several centrality measures for the vertices in the network. It is found that hubs (i.e., vertices with high-degree centrality), as well as bottlenecks (i.e., the vertices

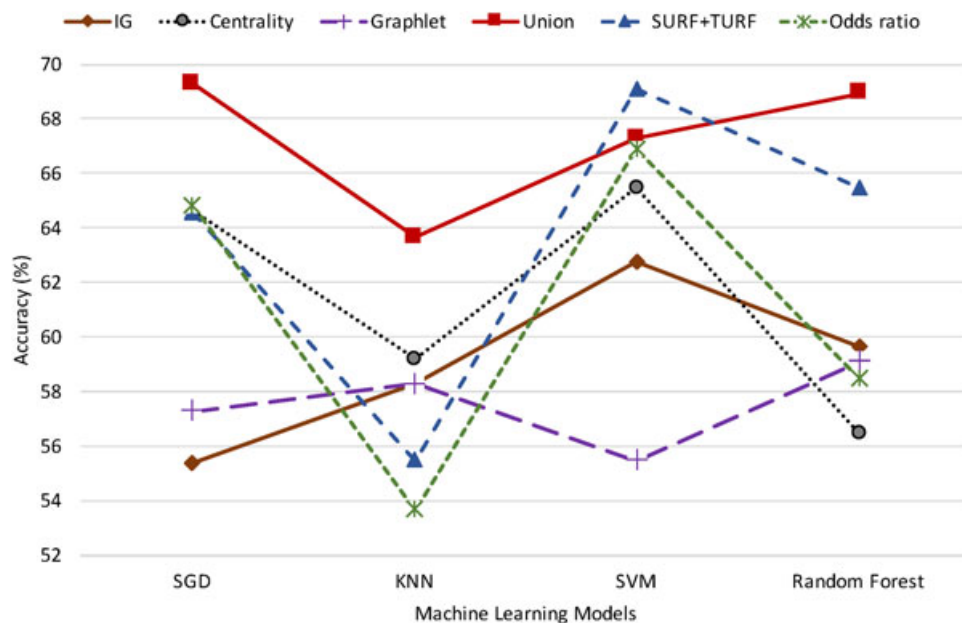


FIGURE 8 The accuracy of disease prediction using four machine learning models, namely logistic regression with SGD, KNN, SVM, and random forest, using 50 top SNPs as features chosen based on IG, centrality, graphlet, Union (IG + centrality + graphlet), SURF+TURF, and allelic odds ratio. IG: information gain; KNN: K -nearest neighbors; SGD: stochastic gradient descent; SNP: single-nucleotide polymorphism; SURF: spatially uniform ReliefF; SVM: support vector machine; TURF: tuned ReliefF

TABLE 4 The number of SNPs (out of top 100 SNPs) selected based on our three different criteria (i.e., IG, graphlet, and centrality) that are among top 1%, 5%, 10%, and 50% SNPs ranked based on the main effect where the total number of SNPs is 9,996

Criterion	Top 1% (rank < 100)	Top 5% (rank < 500)	Top 10% (rank < 1000)	Top 50% (rank < 5000)
IG	1	5	11	57
Graphlet	1	2	5	46
Centrality	4	13	24	68

Note. IG: information gain; SNP: single-nucleotide polymorphism.

with high betweenness centrality) within a protein interaction network, are often encoded by essential genes in model organisms (De et al., 2015; Jeong, Mason, Barabási, & Oltvai, 2001; Yu, Kim, Sprecher, Trifonov, & Gerstein, 2007). Furthermore, in networks such as metabolic networks and statistical epistasis networks, closeness centrality has been used to highlight central vertices (Ma & Zeng, 2003; De et al., 2015). We also found motifs with less than nine vertices and graphlets with up to five vertices in our epistasis network. Based on calculated clustering coefficient and found k -core components, motifs, and graphlets, we could not identify any cycle in the network. We are not sure about the biological mechanism and reasoning behind this long path length and the tree-like structure of our epistasis network. However, this is an interesting observation, which can be further investigated in future studies.

Moreover, as explained in Section 3.4, we selected top SNPs based on IG value, graphlets, and centrality measures, and found their corresponding gene lists. *COL22A1*, *RGS7*, *WWOX*, and *CELF2* were four genes selected based on all three metrics. *RGS7* has been identified as a tumor-suppressor gene resulting in the invasion of human cancer cells (Aissani, Wiener, & Zhang, 2013; Qutob et al., 2018). Żelazowski et al. (2011) studied the correlation of *WWOX* gene expression in CRC patients and proved the tumor-suppressive role of *WWOX* gene expression in the colon. Also, it has been observed that *CELF2* levels are reduced in colon tumor tissue compared to normal ones suggesting *CELF2* as a potential tumor-suppressor gene that Ramalingam, Ramamoorthy, Subramaniam, and Anant (2012) believe it might play a crucial role in tumor initiation and progression. Furthermore, among other top genes chosen based on our metrics, studies showed that *MTHFD1L* is over-expressed in colorectal and breast cancers (Jain et al., 2012; Sugiura, Nagano, Inoue, & Hirotsani, 2004), and plays an essential role in support of cancer growth (Lee et al., 2017). In addition, *NRXN1* has been identified as one of the genes associated with CRC filtered out by

the KEGG (Kyoto Encyclopedia of Genes and Genomes) analysis (Yang, Feng, Ma, Li, & Xie, 2017) and one of significantly down-regulated genes in cancer stroma (Nishida et al., 2012), and its frequent genetic, epigenetic, and transcriptional alterations were identified in Laterally spreading tumors (LSTs; Hesson et al., 2016).



In summary, we constructed a network consisting of a significant number of SNPs, and utilized several network properties to highlight a few key SNPs and genes with potential high disease/phenotype association. While the influence of some of these selected genes on CRC has already been proved in literature, the effect of the rest can be validated by further biological studies. Annotation of these important genes can help classify diseases more accurately and develop more efficient drugs. It can also contribute to identifying people with high cancer risk and providing more effective and timely diagnosis, treatment, and even prevention for the diseases.

For future studies, we expect (a) to analyze other CRC GWAS data and see if our results can be replicated, although considering that the replication might be challenging given the reason of the geographical isolation in Newfoundland, and the resulting low diversity and unique genetic background captured in the data used in this study; and (b) to apply our analysis framework to GWAS data on other diseases.

ACKNOWLEDGMENTS

This study was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grants RGPIN-2016-04699 to T.H. and RGPIN-2017-05201 to Y.C.

ORCID

Somayeh Kafaie  <http://orcid.org/0000-0002-5685-6487>
Ting Hu  <http://orcid.org/0000-0001-6382-0602>

REFERENCES

- Ahmedin, J., Freddie, B., Melissa, M. C., Jacques, F., Elizabeth, W., & David, F. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61(2), 69–90.
- Aissani, B., Wiener, H., & Zhang, K. (2013). Multiple hits for the association of uterine fibroids on human chromosome 1q43. *PLoS One*, 8(3), e58399.
- Albert, R., & Barabási, A. -L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47–97.
- Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5, 1564–1583.
- Andrew, A. S., Hu, T., Gu, J., Gui, J., Ye, Y., Marsit, C. J., & Karagas, M. R. (2012). HSD3B and gene–gene interactions in a

- pathway-based analysis of genetic susceptibility to bladder cancer. *PLoS One*, 7(12), 1–8.
- Barabási, A. -L. (2016). *Network science*. Cambridge: Cambridge University Press.
- Barabasi, A. -L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Reviews Genetics*, 12(1), 56–68.
- Beltrao, P., Cagney, G., & Krogan, N. J. (2010). Quantitative genetic interactions reveal biological modularity. *Cell*, 141(5), 739–745.
- Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2, 113–120.
- Bookman, E. B., McAllister, K., Gillanders, E., Wanke, K., Balshaw, D., Rutter, J., & NIH GxE Interplay Workshop Participants, N. G. I. W. (2011). Gene-environment interplay in common complex diseases: Forging an integrative model—recommendations from an nih workshop. *Genetic Epidemiology*, 35(4), 217–225.
- Boone, C., Bussey, H., & Andrews, B. J. (2007). Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6), 437–449.
- Cao, W., Wu, W., Yan, M., Tian, F., Ma, C., Zhang, Q., & Biddle, F. G. (2015). Multiple region whole-exome sequencing reveals dramatically evolving intratumor genomic heterogeneity in esophageal squamous cell carcinoma. *Oncogenesis*, 4(11), e175.
- Carninci, P., et al. (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740), 1559–1563.
- Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6, 121–133.
- Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20), 2463–2468.
- Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10, 392–404.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd eda). Wiley.
- De, R., Hu, T., Moore, J. H., & Gilbert-Diamond, D. (2015). Characterizing gene–gene interactions in a statistical epistasis network of twelve candidate genes for obesity. *BioData Mining*, 8, 45.
- Dorani, F., & Hu, T. (2018). Feature selection for detecting gene–gene interactions in genome-wide association studies. In *Proceedings of 21st European Conference on the Applications of Evolutionary Computation (Evoapplications)*. Vol. 10784, Springer, 33–46.
- Dorani, F., Hu, T., Woods, M. O., & Zhai, G. (2018). Ensemble learning for detecting gene–gene interactions in colorectal cancer. *PeerJ*, 6, e5854.
- Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., & Palsson, B. Ø (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6), 1777–1782.
- Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M., & Moore, J. (2011). Entropy-based information gain approaches to detect and to characterize gene–gene and gene-environment interactions/correlations of complex diseases. *Genetic Epidemiology*, 35(7), 706–721.
- Greene, C. S., Penrod, N. M., Kiralis, J., & Moore, J. H. (2009). Spatially uniform ReliefF (SURF) for computationally-efficient filtering of gene–gene interactions. *BioData Mining*, 2(1), 5.
- Hesson, L. B., Ng, B., Zarzour, P., Srivastava, S., Kwok, C. -T., Packham, D., & Ward, R. L. (2016). Integrated genetic, epigenetic, and transcriptional profiling identifies molecular pathways in the development of laterally spreading tumors. *Molecular Cancer Research*, 14(12), 1217–1228.
- Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6, 95–108.
- Hočevar, T., & Demšar, J. (2014). A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4), 559–565.
- Hu, T., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2013a). Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models. *Pacific Symposium on Biocomputing*, 18, 397–408.
- Hu, T., Chen, Y., Kiralis, J. W., Collins, R. L., Wejse, C., Sirugo, G., & Moore, J. H. (2013b). An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*, 20(4), 630–636.
- Hu, T., & Moore, J. H. (2013). Network modeling of statistical epistasis. *Biological knowledge discovery handbook* (175–190). Wiley-Blackwell.
- Hu, T., Oksanen, K., Zhang, W., Randell, E., Furey, A., Sun, G., & Zhai, G. (2018). An evolutionary learning and network approach to identifying key metabolites for osteoarthritis. *PLoS Computational Biology*, 14(3), e1005986.
- Hu, T., Sinnott-Armstrong, N. A., Kiralis, J. W., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2011). Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, 12(1), 1–13.
- Hunter, D. J., & Kraft, P. (2007). Drinking from the fire hose - statistical issues in genomewide association studies. *New England Journal of Medicine*, 357(5), 436–439.
- Im, C., Ness, K. K., Kaste, S. C., Chemitilly, W., Moon, W., Sapkota, Y., & Wilson, C. L. (2018). Genome-wide search for higher order epistasis as modifiers of treatment effects on bone mineral density in childhood cancer survivors. *European Journal of Human Genetics*, 26(2), 275–286.
- Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A. L., & Mootha, V. K. (2012). Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science*, 336(6084), 1040–1044.
- Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411, 41–42.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804), 651–654.
- Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In D. Sleeman, & P. Edwards (Eds.), *Machine learning proceedings of the AAAI'92* (pp. 249–256). Morgan Kaufmann.
- Lee, D., Xu, I. M. -J., Chiu, D. K. -C., Lai, R. K. -H., Tse, A. P. -W., Li, L. L., & Wong, C. C. -L. (2017). Folate cycle enzyme MTHFD1L confers metabolic advantages in hepatocellular

- carcinoma. *The Journal of Clinical Investigation*, 127(5), 1856–1872.
- Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1), 15–20.
- Ma, H., & Zeng, A. -P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11), 1423–30.
- Manduchi, E., Chesi, A., Hall, M. A., Grant, S. F. A., & Moore, J. H. (2018). Leveraging putative enhancer-promoter interactions to investigate two-way epistasis in Type 2 Diabetes GWAS. Pacific symposium on biocomputing. Vol. 23, 548–558.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, 298(5594), 824–827.
- Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56(1-3), 73–82.
- Moore, J. H., & Hu, T. (2015). J. Moore, & S. Williams (Eds.), *Epistasis analysis using information theory*. New York, NY: Humana Press.
- Moore, J. H., & White, B. C. (2007). Tuning ReliefF for genome-wide genetic analysis. Proceedings of the 5th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Vol. 4447, Springer-Verlag, 166–175.
- Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20), 1–4.
- Newman, M. E. J. (2010). *Networks: an introduction* (1st eda). Oxford University Press.
- Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620), 1–6.
- Nishida, N., Nagahara, M., Sato, T., Mimori, K., Sudo, T., Tanaka, F., & Mori, M. (2012). Microarray analysis of colorectal cancer stromal tissue reveals upregulation of two oncogenic mirna clusters. *Clinical Cancer Research*, 18(11), 3054–3070.
- Pan, Q., Hu, T., Malley, J. D., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2014). A system-level pathway-phenotype association analysis using synthetic feature random forest. *Genetic Epidemiology*, 38(3), 209–219.
- Park, S., & Lehner, B. (2015). Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types. *Molecular Systems Biology*, 11(7), 824.
- Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9, 855–867.
- Pržulj, N., Corneil, D. G., & Jurisica, I. (2004). Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18), 3508–3515.
- Pržulj, N. (2006). Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2), e177–e183.
- Qutob, N., Masuho, I., Alon, M., Emmanuel, R., Cohen, I., DiPizio, A., & Samuels, Y. (2018). RGS7 is recurrently mutated in melanoma and promotes migration and invasion of human cancer cells. *Scientific Reports*, 8(1), 1–10.
- Ramalingam, S., Ramamoorthy, P., Subramaniam, D., & Anant, S. (2012). Reduced expression of RNA binding protein CELF2, a putative tumor suppressor gene in colon cancer. *Immunogastroenterology*, 1(1), 27–33.
- Rual, J. -F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062), 1173–1178.
- Schumacher, F. R., Schmit, S. L., Jiao, S., Edlund, C. K., Wang, H., Zhang, B., & Peters, U. (2015). Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature Communications*, 6(1), 7138.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
- Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1), 64–68.
- Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, 68(1), 7–30.
- Stelzl, U., et al. (2005). A human protein–protein interaction network: A resource for annotating the proteome. *Cell*, 122(6), 957–968.
- Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643), 249–255.
- Sugiura, T., Nagano, Y., Inoue, T., & Hirotsani, K. (2004). A novel mitochondrial C1-tetrahydrofolate synthetase is upregulated in human colon adenocarcinoma. *Biochemical and Biophysical Research Communications*, 315(1), 204–211.
- Sun, Z., Wang, P., Roebathan, B., Cotterchio, M., Green, R., Buehler, S., & Parfrey, P. (2011). Calcium and vitamin D and risk of colorectal cancer: Results from a large population-based case-control study in Newfoundland and Labrador and Ontario. *Canadian Journal of Public Health*, 102(5), 382–389.
- Thomas, D. (2010). Gene-environment-wide association studies: Emerging approaches. *Nature Reviews Genetics*, 11(4), 259–272.
- Urbanowicz, R. J., Meeker, M., Cava, W. L., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85, 189–203.
- Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., & Moore, J. H. (2018). Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, 85, 168–188.
- Verma, S. S., Lucas, A., Zhang, X., Veturi, Y., Dudek, S., Li, B., & Ritchie, M. D. (2018). Collective feature selection to identify crucial epistatic variants. *BioData Mining*, 11(1), 5.
- Wang, P. P., Dicks, E., Gong, X., Buehler, S., Zhao, J., Squires, J., & Parfrey, P. S. (2009). Validity of random-digit-dialing in recruiting controls in a case-control study. *Am J Health Behav.*, 33(5), 513–520.
- Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, 6(2), 109–118.
- Wernicke, S., & Rasche, F. (2006). FANMOD: A tool for fast network motif detection. *Bioinformatics*, 22(9), 1152–1153.
- West, D. B. (2001). *Introduction to graph theory* (1st ed.). Upper Saddle River, New Jersey: Prentice Hall.

- Wilkins, J. F., Cannataro, V. L., Shuch, B., & Townsend, J. P. (2018). Analysis of mutation, selection, and epistasis: An informed approach to cancer clinical trials. *Oncotarget*, *9*(32), 22243–22253.
- Yang, Q., Feng, M., Ma, X., Li, H., & Xie, W. (2017). Gene expression profile comparison between colorectal cancer and adjacent normal tissues. *Oncology Letters*, *14*(5), 6071–6078.
- Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., & Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, *3*(4), e59.
- Żelazowski, M. J., Płuciennik, E., Pasz-Walczak, G., Potemski, P., Kordek, R., & Bednarek, A. K. (2011). WWOX expression in colorectal cancer - a real-time quantitative RT-PCR study. *Tumour Biology*, *32*(3), 551–560.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kafaie S, Chen Y, Hu T. A network approach to prioritizing susceptibility genes for genome-wide association studies. *Genet. Epidemiol.* 2019;1–15.
<https://doi.org/10.1002/gepi.22198>