Applied Network Science

RESEARCH Open Access

TOPDRIVER: the novel identifier of cancer driver genes in Gastric cancer and Melanoma

Seyed Mohammad Razavi[1,2], Farzaneh Rami[2], Seyede Houri Razavi[1,2] and Changiz Eslahchi[1,2*]

*Correspondence:
ch-eslahchi@sbu.ac.ir
[1]Computer Science Department, Mathematical sciences faculty, Shahid Beheshti University, Tehran, Iran
[2]School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Abstract
Nowadays, research has found a strong relationship between genomic status and occurrence of disease. Cancer is one of the most common diseases that leads to a high annual mortality rate worldwide, and the disease's genetic content remains challenging. Detecting driver genes of different cancers could help in early diagnosis and treatment. In this paper, we proposed TOPDRIVER, a network-based algorithm, to detect cancer driver genes in cancers. An initial network was constructed by integrating four different omic datasets: HPRD, NCBI, KEGG, and GTEx. This integration created a gene similarity profile that provided a comprehensive perspective of gene interaction in each subtype of cancer and allocated weights to the edges of the network. The vertex scores were calculated using a gene-disease association dataset (DisGeNet) and a molecular functional disease similarity. In this step, the genes network was jagged and faced with a zero-one gap problem. A diffusion kernel was implemented to smooth the vertex scores to overcome this problem. Finally, potential driver genes were extracted according to the topology of the network, genes overall biological functions, and their involvement in cancer pathways. TOPDRIVER has been applied to two subtypes of gastric cancer and one subtype of melanoma. The method could nominate a considerable number of well-known driver genes of these cancers and also introduce novel driver genes. NKX3-1, KIDINS220, and RIPK4 have introduced for gastrointestinal cancer, UBA3, UBE2M, and RRAGA for hereditary gastric cancer and CIT for invasive melanoma. Biological evidences represents TOPDRIVER's efficiency in a subtype-specific manner.

Keywords: Network-based methods, Cancer driver genes, Gastric cancer, Melanoma

Introduction

Cancer is one of the most common multifactorial diseases (Ji et al. 2010), which could be caused by environmental factors and/or genomic aberrations (e.g., sequence mutations, aberrant promoter methylation, and structural lesions such as gains/losses, inversions, and translocations) (Greenman et al. 2007; Haverty et al. 2008; Beroukhim et al. 2007; Charames and Bapat 2003). These aberrations play critical roles in both initiation and progression of cancer. Some of these aberrations disturb normal cellular processes and contribute to cancer initiation and progression called driver mutation, while others emerge simply as victims of genomic instability resulting from cancer progression called passenger mutation (Pole et al. 2006). Genes that contain driver mutations are generally called driver genes. A driver gene is causally implicated in the process of oncogenesis, while a passenger gene makes no contribution to cancer development itself, but is simply a by-product of the genomic instability observed in cancer genomes. Driver and passenger genes can be differentiated by the functional roles they play in cells. Different genomic data that measure gene functions at different dimensions would be highly informative to separate potential driver from passenger genes. Recently, several methods have been proposed to identify potential driver genes based on systematic integration of genome scale data of CNA, gene expression profiles, protein-protein interaction, epigenetic, metabolism pathways, sequence similarity and Gene Ontology. Distinguishing driver genes from passenger genes has thus been considered an important goal of cancer genome analysis and has a major impact on cancer prediction, diagnosis, and treatment, especially in the field of personalized medicine and therapy (Santarius et al. 2010; Futreal et al. 2004).

On the other hand, functional differences between driver and passenger genes lead to their distinct roles in initiation and progression of cancer. In this case, different omic data measuring gene functions would be highly informative in identifying candidate driver genes. The critical challenge is to analyze and integrate such information using the most efficient and meaningful way. Literature brings different methods of detecting driver genes to the table.

Cells consist of various complex, plastic, and dynamic molecular structures that networks could meaningfully implement. Under the molecular network analysis, a genetic aberration may cause network architectural change by affecting or removing a node or its neighbors within the network. Network-based approaches usually study the entire molecular structure of cells as a network and abundance of cancer genomic data gives them a tremendous opportunity to understand cancer initiation and progression. So, it seems that they could be beneficial methods in driver genes detection.

Over the past few years, researchers have presented various approaches to detect cancer driver genes. Generally, these methods divide into three major categories: statistical-based, machine learning-based and network-based methods. Active driver (Reimand and Bader 2013), eDriver (Porta-Pardo and Godzik 2014), iDriver (Yang et al. 2016) are some examples of statistical-based methods. Emidio Capriotti (2011), Ryslik (2013), Luo (2017), Yuan (2018) and their colleagues also represented several machine learning methods to detect cancer driver genes.

From the third category, MEMo (Ciriello et al. 2011) was developed to distinguish cliques in a pathway or network by using mutually exclusive mutation using The Cancer Genome Atlas (TCGA dataset). Fabio Vandin et al. (2011) derived two algorithms that use somatic mutations to find new driver pathways. Their method finds two sets of genes with high coverage and high exclusivity and defines a measurement to quantify the degree of exhibiting a set in both criteria. Finding a set of genes to optimize this measure is a challenge. They introduce an optimal greedy solution to find this set of genes (Vandin et al. 2011). PRADIGM-SHIFT (Ng et al. 2012) was another network-based method that proposed to detect the mutation impact on pathways and defined a gain or loss of function for mutations using TCGA dataset. It focuses on small pathways with specific genes and attempts to estimate the pathway effects of a mutation in a tumor sample by defining a score for each gene. Hou and his colleagues (Hou and Ma 2014) proposed a method based on PageRank algorithm to rank altered genes for each individual by application of dynamic damping factors and providing a specific model for network's directionality. It uses TCGA, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Mutual Exclusivity Modules in Cancer (MEMo), Pathway Interaction Database (PID), The Cancer Genomics Cloud (CGC), and Pan-Cancer datasets to prioritize genes according to their influence on the perturbation of downstream genes. It means that a gene will rank higher if it directly or indirectly causes more changes in the expression of downstream target genes. MAXDRIVER (Chen et al. 2013) is another network-based method that used several optimization strategies to identify potential cancer driver genes. This method builds a network by integration of a fused gene functional similarity network, gene-disease associations and a disease phenotypic similarity network. Then a maximum information flow strategy was employed to prioritize candidate genes that were located in CNA regions and finally top one-ranked genes were considered to be driver genes, while others were considered passenger genes. Ramsahai and his colleagues presented a pathway-oriented method for detecting cancer driver genes (Ramsahai et al. 2017). Their strategy benefits from combining data from three different sources on the prediction outcome of cancer driver genes by DriverNet and DawnRank. Finally, 33 new candidate drivers were identified.

Generally, among all methods, those that integrate different data sets to build their systemic knowledge could be more successful in detection of driver genes, since they study the problem from different aspects. Network-based methods have been widely applied in different fields. These methods are beneficial tools when we are faced with high-throughput biological data (Sanchez-Garcia et al. 2014). They are also beneficial for analyzing a problem in different omic datasets based on gene interactions.

In this paper, to identify potential driver genes we integrate four publicly available human genomic data. We have presented a novel network-based method to find new cancer driver genes. By this method, we detect driver genes as the one that either have a high burden of driver mutations or include mutations that have a higher probability to initiate a cancer. To do this, we introduce TOPDRIVER as a computational method to identify candidate driver genes. First, we systematically integrate several omic data to build a network of cancer genes. In this network vertex scores come from disease-disease similarity profile (Cheng et al. 2014) and edge weights come from a systematic integration of these datasets. This network usually faces a zero-one gap problem (Wang et al. 2008) in vertex scores, which refers to the potential pitfall that assigns biased ranks to some of the network nodes. A jagged network will emerge as a result of this zero-one problem. This jagged network has been defined as a network containing pairs of vertices connected to a high weighted edge, whereas one vertex has a high score and the other has a low score. To overcome this problem, a diffusion kernel method is used to adjust the vertices score to modify jagged network. Diffusion kernel method propagates the vertex scores according to their edge weights. We built an induced subgraph after network modification by considering a few numbers of genes related to the intended cancer subtype and prioritizing genes with permutation test. Subsequently, we chose some of the unknown candidate genes as probable driver genes for aforementioned cancer subtypes and provided biological evidence of their contribution to the development of cancer. To test the capability of TOPDRIVER in detecting novel driver genes we employed it to gastrointestinal cancer, hereditary gastric cancer, and melanoma. To evaluate the set of identified candidate driver genes in different cancers, the list of our prioritized candidate driver genes compared with the set of driver genes introduced by IntOGen and TCGA datasets. The Cancer Genome

Atlas (TCGA) is a project that catalogues genetic mutations responsible for cancer, using genome sequencing and bioinformatics. TCGA applies high-throughput genome analysis techniques to improve the ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease. IntOGen is a web platform used to identify cancer drivers across tumor types and to present the results of the systematic analysis of the latest available large data sets of tumor somatic mutations. The results reveal the effective role of smoothing scores in PPI networks and the importance of efficient and systematic integration of several omic data. So, TOPDRIVER would be a good choice for identifying candidate driver genes in different cancers.

Materials and methods

In this part, the method and datasets are fully described. TOPDRIVER is designed to build a network and detect driver genes by integrating four types of omic datasets. We used Python 2.7 and numpy package for data analysis and other implementations.

Datasets:

1. HPRD: It is a database of proteomic information related to human proteins (https://hprd.org). We extracted protein-protein interactions (PPI) from version 9.0 (released on Apr 13, 2011) of the HPRD database consisted of 30,047 proteins and 41,327 interactions.
2. NCBI RefSeq protein sequences: It represents a collection of genome, transcript and protein sequences (Pruitt et al. 2006). We obtained the protein sequences of all genes in FASTA format from this database.
3. KEGG: A dataset for biological interpretation of genome sequences and other high-throughput data (Kanehisa et al. 2015). A total of 200 human pathways were obtained from the KEGG database.
4. GTEx: It evaluates the connection between genetic variation and gene expression in normal human tissues and ultimately determines how this connection leads to disease susceptibility and development (GTEx Consortium and et al. 2015). The gene expression data for 53 human tissues were downloaded from GTEx (GTEx Analysis v7).
5. Gene Ontology (GO): Classifies gene product function using structured and controlled vocabulary (Consortium GO 2012). The Gene Ontology (GO) biological process domain and the corresponding annotations for human genes have been downloaded from this collection. We focused on genes annotated with at least fifteen informative GO terms that appeared at or below the fifth level of the GO hierarchy and had annotated at least five genes. We acquired a total of 3,842 well-annotated genes and named them "GOSet."
6. DisGeNET: It is a flexible platform that includes the molecular foundation of specific human diseases and their comorbidities, disease gene properties, etc. We downloaded all the gene-disease associations from DisGeNET. DisGeNET provides weighted relations between genes and diseases that was a requirement for our present work (Piñero et al. 2016).

Method

In this part, in five steps, the TOPDRIVER method is explained. In the beginning, the four datasets were stored in matrix form. Integration of these four datasets constructed a coherent gene functional similarity matrix (GSP). The initial network of genes, G_N, obtained through the integration of GSP matrix, disease molecular functional similarity matrix, and gene-disease associations matrix. Subsequently, we applied a diffusion kernel on the graph to overcome zero-one gap problem to balanced vertex scores of the jagged network. In the next step, an induced graph, $N(d)$, was obtained for each disease. Then, a permutation test applied on $N(d)$ vertex scores and candidate genes prioritized by using their p-value. Finally, due to biological evidence, some of these top-ranked genes with the lowest p-values were considered as driver. In the following, we describe each part of the method in details.

Step1: Gene Similarity Profile (GSP) construction

In this step, first four similarities were defined between each pair of genes based on the information of HPRD, RefSeq, KEGG and GTEx datasets. Given a pair of genes i and j, we designated their similarity $r_{ij}^{(1)}$ as the unit weight 1, if these two genes have interaction in HPRD, and 0 otherwise. In fact, this similarity presented the adjacency matrix of the protein-protein interaction network obtained from HPRD. Since the HPRD interactions are defined among proteins and we need interactions among genes, we mapped each protein to it's corresponded gene using NCBI RefSeq protein sequences dataset. So, while there was an interaction between two proteins we set an interaction between its corresponded genes. After doing so, we had 37,364 interaction among 9,515 genes. The similarity of the corresponding protein sequences of genes i and j is denoted by $r_{ij}^{(2)}$, which was obtained by the following formula:

$$r_{ij}^{(2)} = \begin{cases} -\log\left(e_{ij}\right)/\max_{ij}\left\{-\log\left(e_{ij}\right)\right\}, & e_{ij} \neq 0 \\ 1, & e_{ij} = 0 \end{cases} \tag{1}$$

where e_{ij} was the e-value that was presented by NCBI BLASTP. In other words, this similarity represents normalized BLAST e-value between two protein sequences.

For each gene i, using KEGG data set, we define a 200-dimensional binary vector p_i by assigning 1 to an element if the gene was presented in the corresponding pathway and assigning 0 otherwise. Let $p_i.p_j$ denoted the dot product of these two vectors corresponding to genes i and j. Then, the correlation between two genes defines by the cosine of angles in a space of dimentionality. A correlation of -1 indicates that the pair of genes in the opposite direction, a correlation of 1 indicates coincident vectors, and correlation 0 indicates orthogonality.The gene coincidence relation of two genes i and j is described as follows:

$$r_{ij}^{(3)} = \frac{p_i.p_j}{|p_i||p_j|} \tag{2}$$

Now for each of the 9515 genes a 53-dimensional vector was built using GTEx data set. Let e_i denote a vector of size 53 which its element corresponding to the expression level of gene i for each specific tissue. Finally, the correlation coefficient of expression of the genes i and j using covariance and variance of these vectors were computed by $r_{ij}^{(4)}$ as follow:

$$r_{ij}^{(4)} = \left| \frac{cov(e_i, e_j)}{\sigma\left(e_i\right)\sigma\left(e_j\right)} \right| \tag{3}$$

The correlation coefficient is commonly used as a measure of divergence of gene expression profiles between different tissues. None of the above similarities reflected a thorough explanation of gene functional similarity, we needed to integrate these similarities in order to present a more comprehensive similarity revealing relationships between gene properties and gene functions. To acquire this aim, we constructed a new Gene Similarity Profile

(GSP). We used biological process domain of the Gene Ontology and the related annotations for human genes to define the similarity between genes by selecting genes with at least fifteen informative GO terms that presented at or under the fifth level of the GO hierarchy and had at least five genes annotated.

The functional similarities r_{ij} between genes i and j is acquired by Resnik method (Chen et al. 2013), which can be obtained from the software package GOSemSim (Sudhakar 2009). Multiple regression is a powerful method for merging different attributes, so we used a multiple regression for merging the four databases in order to build an informative matrix. This method has also successfully applied in Chen et al. (2013).We adjusted the following regression model:

$$\log \frac{R_{ij}}{1 - R_{ij}} = \alpha_0 + \sum_{k=1}^{4} \alpha_k . r_{ij}^{(k)} \tag{4}$$

where α_k, $k = 0, \ldots, 4$ were regression coefficients. Since GOSemSim just included 3842 of our genes, 3842 by 3842 equations were considered in Eq. 4. Due to using this approach, we computed $\alpha_0 = -1.1304$ for the regression intercept, $\alpha_1 = 6.4162$ for the gene sequence similarity, $\alpha_2 = 0.2408$ for the gene co-expression pattern, $\alpha_3 = 1.0651$ for the pathway co-occurrence relationship and $\alpha_4 = 0.2071$ for protein-protein interactions. With these acquired parameters, GSP was calculated between each pair of genes as follows:

$$GSP_{ij} = \alpha_0 + \sum_{k=1}^{4} \alpha_k . r_{ij}^{(k)} \tag{5}$$

Using well studied genes with more GO annotations helps us to properly learn the coefficients of the above regression model. While the regression coefficients defined for informative 3842 genes, we apply them for integrating four previously similarities among all pair of 9515 genes of HPRD.

Step2: Weighted network construction

In this step, we constructed a weighted network, G_N, where N indicates vertices which are representatives of the 9515 genes. There is an edge between genes i and j with weight GSP_{ij} if the GSP_{ij} was not zero. Let L_i be the set of disease related to gene i, and $W_m(i)$ denoted the influence of gene i in disease m, which is calculated using gene-disease association dataset (DisGeNET) (Piñero et al. 2016). For two disease m and n let $S(m, n)$ denote the molecular functional similarity of these disease which is obtained from (SemFunSim) (Cheng et al. 2014). Now we define a vertex score for gene i as:

$$S_i = \frac{\sum\limits_{m,n \in L_i} s(m, n) w_m(i) w_n(i)}{\sum\limits_{m,n \in L_i} w_m(i) w_n(i)} \tag{6}$$

Step3: Applying diffusion

Since the network G_N is a jagged network, we utilized the diffusion kernel and time evolution operator as a machine learning strategy to overcome zero-one gap problem (Wang et al. 2008) via smoothing the vertex scores of this jagged network. Diffusion kernel helps to smooth the vertex scores by the weights of their neighbors (Babaei et al. 2013). To do this, diffusion kernel applied to the Laplacian matrix of graph G_N (Chung 1997). The Laplacian matrix $H = [H_{ij}]$ defines as:

$$H_{ij} = \begin{cases} -GSP_{ij}, & for\ i \sim j \\ \sum_{i \neq k} GSP_{ik}, & for\ i = j \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The diffusion kernel, which is applied on G_N is:

$$K_\beta = e^{\beta H} \tag{8}$$

and time evolution operator considered as:

$$Z_\beta = K_\beta Z_0 \tag{9}$$

where Z_0 denoted the initial vertex scores of G_N. Note that diffusion kernels form continuous families and indexed by the real parameter β. Z_β is the representation of vertex scores in the weighted network after a desirable modification using diffusion kernel-based method. In this work, we set β as 0.01. More details are available in Lafferty and Kondor (2002). By this method, the score of the vertices of G_N is smoothed and G'_N is considered as modified scored network.

Step4: Introducing candidate driver genes for an intended disease d

Let A_d considered as the set of genes related to disease d, which is obtained from DisGeNET gene-disease association dataset. Let B_d denote the set of vertices, which are adjacent to at least one of the genes in A_d. The disease corresponding graph G'_d obtained from G'_N by the elimination of vertices that are not members of $A_d \cup B_d$. In this Step, we are going to candidate some genes as driver based on high confident scoring value. To do this a permutation procedure is used to determine the genes with significant scores. The subset of genes whose scores have significant empirical p-value is considered as the set of candidate driver genes. The empirical p-value is computed based on the following permutation procedure. To generate the null distribution, we considered 1000 random shuffle of Z_0, $Z^a, 1 \leq a \leq 1000$, on the score of vertices of G'_d. The empirical p-value of gene i is computed as follows:

$$p - value(i) = \frac{\left|\{a \mid Z_{0.01}(i) \leq Z^a_{0.01}(i)\}\right|}{1000} \tag{10}$$

where $Z_{0.01}(i)$ and $Z^a_{0.01}(i)$ denoted the initial diffusion score of i and diffusion score of i in a-th permutation test. Subsequently, was considered the set C_d of genes with p-values less than 0.05 as candidate driver genes of disease d.

It seems that mutation related data sets may impress algorithms to focus on some biased regions of the genome to predict potential driver genes (Chen et al. 2013). We applied non-mutational data sets like KEGG, HPRD, GTEx, NCBI RefSeq, and GO to introduce candidate cancer driver genes. KEGG data let us focus on cancer-related pathways and genes. Cancer is associated with a plethora of gene expression differentiations and interaction between them to loss of control over vital cellular functions. So, applying HPRD and GTEx helps to consider both of gene expression data and interaction between genes. The algorithm used NCBIRefSeq dataset to generate gene sequence similarity by PBLAST e-value. If PBLAST returns a low amount of e-value for two protein sequences, it can be concluded that any cancer-related alteration in one of protein's corresponding gene could have the same effect in the other one. So, if one of these genes is driver the other one could be considered as a potential driver gene too. Since these data are not enough for introducing potential driver genes, it is necessary to use biological evidence which demonstrated in step 5.

Step5: Exploration of potential driver genes

In this step, the genes in C_d analyzed based on four criteria: the genes overall function, protein neighbors in N'_d, protein involvement in cancer-related pathways, and the reputation of genes in cancer literature. A gene is nominated as a potential driver gene:

- if its overall function is cancer-related
- if it is the neighbor of a strong tumor suppressor gene or oncogene in the obtained network
- if it is involved in cancer-related pathways
- if it has an altered status in cancer (whether altered expression and function or gene mutation)

The entire procedure of identifying potential cancer driver genes described in Fig. 1.

Results

At the end of TOPDRIVER step 4, a set of candidate driver genes are introduced. To evaluate this set, we apply the method on two subtypes of gastric cancer (hereditary and gastrointestinal) and an invasive melanoma subtype. In the next section, we evaluated these sets by 5-fold cross validation method.

5-fold cross validation

We performed 5-fold cross validation (5-CV) for each disease to evaluate the performance of candidate driver genes. We extracted A_d, the set of related genes to each cancer from the DisGeNet. A_d, for Gastrointestinal cancer contained 11 genes as TP73, ARHGAP24, ERBB2, GAST, TP53, EZH2, SLC12A9, CDKN1A, TGFA, IL1B, and TLR4. For hereditary gastric cancer, A_d contained 13 genes as CDH1, JUP, MRC1, TP53, APC, SMUG1, CTNNA1, DNAJB4, MAP3K6, TSC1, CTNND1, CTNNB1, FZR1. Finally for Melanoma, A_d contained the following 23 genes, EPHB2, PCNA, AKT1, CD82, MAPK1, BRAF, ADAM15, ORAI1, CDK4, BLM, RND3, NRAS, EZH2, SRF, ADAM9, SPP1, CDKN2A, STIM1, MMP2, MMP1, DDX11, ACTN4, and VEGFA. Now for cross validation, A_d was partitioned randomly to 5 approximately equal sets. Method of cross-validation executed in 5 iterations. In each iteration, we consider one set as test and union of the four other sets as a training set. The algorithm executed with training set as A_d in step 4. A new set of candidate genes was obtained by running the diffusion kernel and permutation test. The existence of test set members in this set were examined. For each disease, 5-fold cross validation repeated 30 times and the results were averaged to ensure unbiased evaluations.

In order to control for type I errors as a result of testing associations with three cancer subtypes, Gastrointestinal, Hereditari, and Melanoma, respectively, the Benjamini-Hochberg false discovery rate (FDR) method was used (Benjamini and Hochberg 1995) for q=0.05. In each repeat of 5-fold cross validation the P-value corresponded to $q = 0.05$ has changed. But they approximately are around 0.003, 0.009 and 0.001 for Gastrointestinal, Hereditary, and Melanoma, respectively.

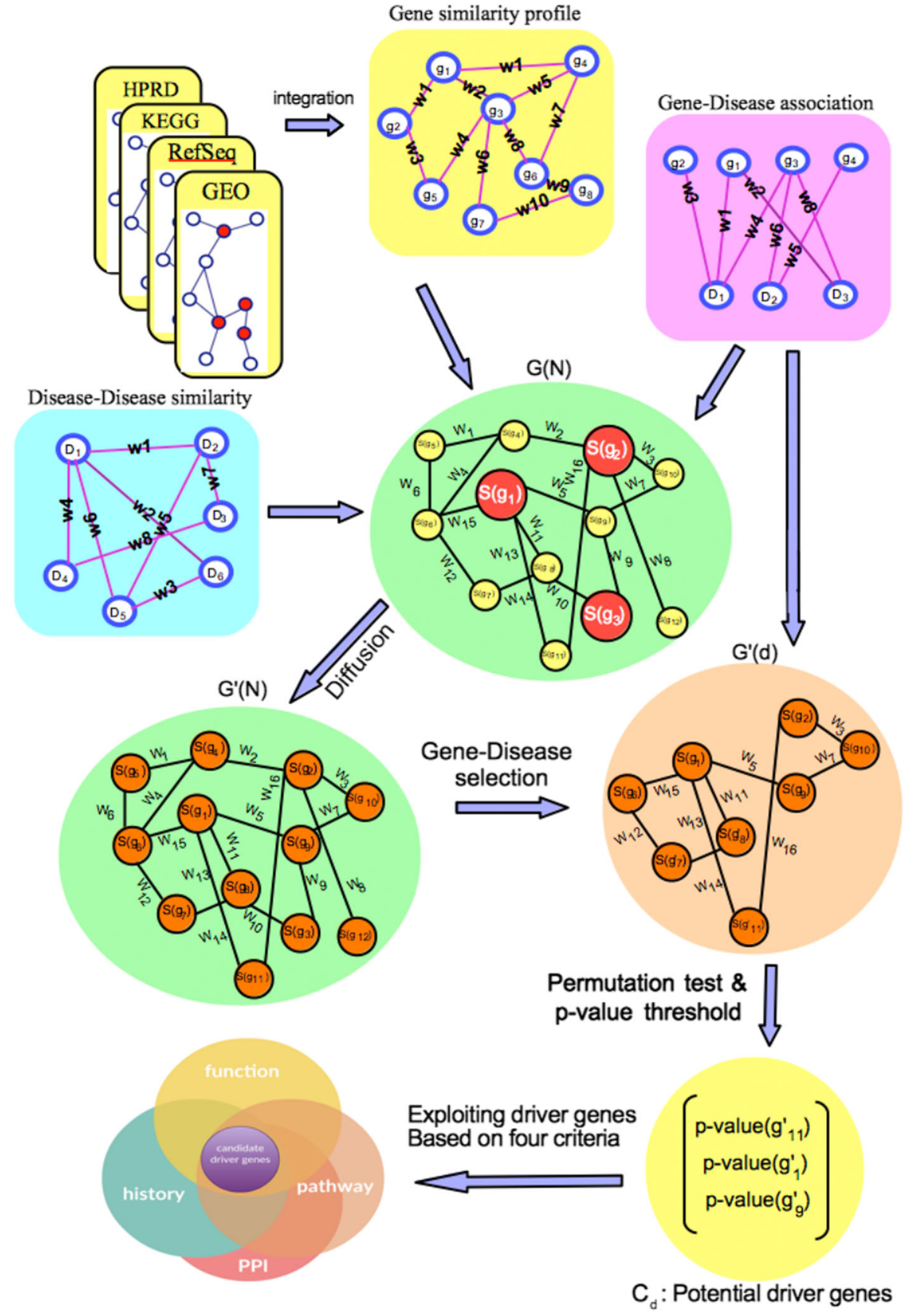

Fig. 1 Workflow of TOPDRIVER. The figure depicts genes by g $_i$, edge weights by w $_i$, disease by D $_i$, and vertex score by S(g $_i$). Gene similarity matrix were constructed by integration of four datasets that were depicted in upper left corner. The G(N) were built with application of gene similarity matrix, disease-disease similarity matrix, and gene-disease association. The G(N) depicts the discrimination among gene scores by coloring high-scored vertices with red and the other low-scored vertices by yellow. As G′(N) shows, the jagged network was amended using diffusion kernel. Afterward, we select an induce graph using gene-disease association by finding a set of genes related to an intended disease. Finally, we applied a permutation test and *p*-value cut-off to choose related candidate genes

In this method to obtain a null distribution we considered two steps, permutation approach and diffusion kernel step. The false discovery rate (FDR) is obtained for each of the test sets. After the null distribution has been built, the p-value of each gene is calculated according to Eq. (10). The p-values of all genes are sorted in the ascending order. Then for FDR, $q = 0.05$, we can determine how many genes are accepted based on their p-values according to the Benjamini-Hochberg (BH) method. The true FDR is calculated as the ratio of the number of genes which are not belonging to the test set and the total number of genes identifications. In fact, the detected genes which are not in the test set are considered as false positive. The average of FDR for 30 iteration of 5-fold cross validation was 0.90, 0.94, 0.95, for Gastrointestinal cancer, Hereditary cancer and Invasive Melanoma respectively. We believe that FDR could not be a good measure of evaluation for our method as there was biological evidence that in each repetition of 5-fold cross validation, false positive nodes (nodes not included in the test set) are actually related to the intended disease. To prove this fact, we used MSigDB (http://software.broadinstitute.org/gsea/msigdb/annotate.jsp). MSigDB categorize members of a gene set by gene families. For example, in one iteration for gastrointestinal cancer, we obtained a set of 24 genes which contains all test set genes $\{TLR4, GAST, TP53\}$, and 6 genes of the train set. In this case, the algorithm extracted 3 known tumor suppressor genes (TP35, TP73, SYK), 3 oncogenes (ARAF, BRAF, KIT), 5 protein Kinases, 2 translocated cancer genes and the other cancer-related genes (see Table 1). On the other hand, the algorithm detected 7 structural genes which most of them involved in function of gastric tissue (such as pepsin). So, it seems that the method has the capability to detect tissue specific genes. Furthermore, the method has beneficially separated three inflammatory, structural, and cell-cycle gene kingdoms.

This shows the approximate differentiation of gastric functional genes as a semi-separate entity that could be another indicator of algorithm's ability to place genes in their proper pathway. Therefore, if we choose a new unknown gene that is placed in the cell cycle control entity, we could be relatively certain that the new gene is involved in the cell cycle control. This is a valuable point for future studies in the discovery of gene function.

As an another example, the algorithm obtained 38 genes for one iteration of hereditary gastric cancer which contained all of the genes in test set except one of them, MRC1, CTNNA1, DNAJB4 and 8 genes from training set. 31 of these 38 genes were cell-cycle genes (81%) and the other 7 genes were structural genes. These genes are mainly members of the WNT and MAPK pathway, both of which are common pathways involving in gastric cancer. Furthermore, we saw some valuable genes that trigger hereditary gastric cancer such as TP53 and CDH. We also see 3 proved gastric cancer driver genes (including TP53, CTNNB1, CDH1) among these 38 genes. In this case, the algorithm identified 4 known tumor suppressor genes, 1 oncogene and other cancer-related genes (see Table 1).

Table 1 Protein family of false discovery genes

Cancer	C and GF	TF	HP	CDM	PK	TCG	OG	TS
Gastrointestinal	2	2	0	1	5	2	3	3
Hereditary	0	4	0	1	0	1	1	4
Melanoma	2	2	0	2	9	3	3	4

C and GF: cytokines and growth factors, TF: transcription factors, HP: homeodomain proteins, CDF: cell differentiation markers, PK: protein kinases, TCG:translocated cancer genes, OG:oncogenes,TS:tumor suppressors

In the case of Melanoma, structural and metastatic genes weren't separated from each other, since the selected subtype of melanoma was invasive. These structural genes may be altered in the cancer situation or be involved in the cancer invasion procedure. However, we made a very strict selection of genes that were not directly involved in the cancer invasion as structural genes. According to this, for one of the iteration algorithm detected 54 genes contained all the test genes, AKT1, VEGFA, SRF, MMP2, ADAM9, and 12 genes from training set. 34 genes out of them (63%) are cell-cycle genes, 11 genes are metastatic (20%), and 9 genes are structural (17%). Among these 54 genes, there exist 5 proven melanoma driver genes including SYK, RAC1, MAP2K1, BRAF, and BLM. Table 1 shows the gene family of 25 of these 54 genes.

According to these investigations, one can conclude that the false negative genes which are obtained from the step 4 of the algorithm, are mostly related to desired cancer type.

Performance of GPS with eliminating each type of similarities

In this part, we omitted each similarity matrix in step one of TOPDRIVER method, in order to study the effect of each similarity matrices on the final results of 5-fold cross validation after step 4. Then, these results compared with the results of integrating four similarity matrix. Based on previous section, to compare the results, we considered the precision of detected genes after elimination of a similarity matrix with the precision of the detected genes by considering all the similarity matrices. Table 2, represents the average precisions for 30 iterations of 5-fold cross validation with the elimination of each similarity matrix and without it. According to this table the precisions has changed significantly by each similarity matrix elimination. Due to the results of Table 2, It is clear that using cancer-related pathways from KEGG could significantly impress the results of TOPDRIVER. So, new versions of KEGG will also improve our results.

In the next section, the potential driver genes of two subtypes of gastric cancer (hereditary and gastrointestinal) and an invasive melanoma subtype which are detected from step 5 of TOPDRIVER method are introduced.

Predicted driver genes of gastrointestinal cancer

The candidate driver genes for this disease, after the end of TOPDRIVER step 4, are 88 genes. The full list of these genes and their *p*-values can be found in the Additional file 1: Table S1. According to Intogen (www.intogen.org) database, we found that 11 out of 88 genes (12.5%) were known as potential gastric driver genes. We used the STRING database by 0.95 threshold to depict a graphical representation of these 88 genes. Figure 2 obtained after elimination of vertices with degree zero or one iteratively until no vertices with a degrees below 2 existed. Based on Step 5, our method nominated three potential driver genes for gastrointestinal cancer: NKX3-1, KIDINS220, and RIPK4. NKX3-1 is

Table 2 Precision of driver gene detection after elimination of each similarity matrix

Similarity	Gastrointestinal	Hereditary	Melanoma
Removing HPRD	0.6166	0.4266	0.6050
Removing NCBI RefSeq protein sequences	0.6500	0.3666	0.5730
Removing KEGG	0.4266	0.3300	0.5900
Removing GTEx	0.5548	0.4506	0.5504
Integrating four similarity	0.7438	0.7896	0.8312

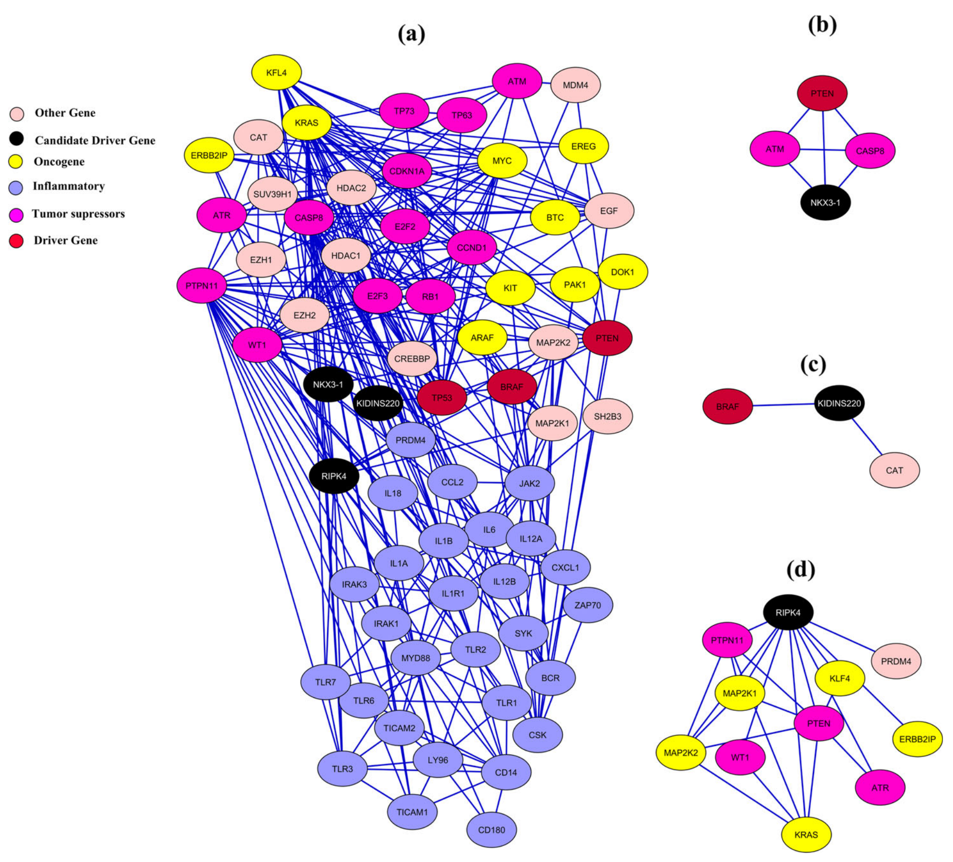

Fig. 2 Gastrointestinal network. The figures depict graph C_{d} using 0.95 threshold in STRING

a homeobox-containing transcription factor, which has mostly known as a strong negative regulator of epithelial cell growth and several studies indicated its down-regulation as well as gene deletion in prostatic cancer (Camacho et al. 2017; Nodouzi et al. 2015). Moreover, it has been introduced as a strong tumor suppressor and regulator of proliferation in the prostatic epithelium (Martinez et al. 2014). This function had also seen in other cancers such as hepatocellular carcinoma (Jiang et al. 2017). Based on TCGA, the highest mutation rate for the gene was detected in the endometrial and colorectal carcinomas (Tomczak et al. 2015). The most frequent mutations in NKX3-1 are missense substitution and deletion (cancer instute N 2018d). As mentioned before, NKX3-1 is considered as a potent tumor suppressor in some cancers such as prostatic cancer (Camacho et al. 2017; Nodouzi et al. 2015). It also has an interaction with strong TSGs such as PTEN and ATM and is a member of apoptosis signaling pathway via interaction with CASP8 (Fig. 2b). So, we have predicted NKX3-1 decreased expression in gastrointestinal cancer. The second potential driver gene for gastrointestinal cancer is KIDINS220. KIDINS220 is a multi-functional transmembrane scaffold protein, which is involved in the cell signaling regulation, cancer development, and metastasis. It is also involved in the several other tumor formation processes such as apoptosis and vascular development (Cai et al. 2017). Based on TCGA, the highest mutation rate for KIDINS220 was observed in the endometrial carcinoma, colorectal adenocarcinoma, and melanoma (cancer instute N 2018c); however, the mutation rate in the stomach cancer is also approximately high

(5.91%) (cancer instute N 2018b). Besides, different kinds of gene alteration were seen for the gene in the different cancer types, most of them result in the gene overexpression (Carvalho et al. 2014; Sakamoto et al. 2017). This gene is also a member of BRAF signaling pathway, which is known as an intra-cellular oncogene. KIDINS220 affects on two strong mediators of RAS signaling pathway: BRAF and CAT. It is also introduced as a target for MAPK signaling pathway inhibition during melanoma targeted therapy (Cai et al. 2017). KIDINS220 overexpression has been reported in several cancers. It also is a member of RAS/MAPK signaling pathway, which is usually activated in most of cancers (Cai et al. 2017). According to these points, we have predicted the increase of KIDINS220 expression in gastric cancer. The third and last potential driver gene for gastrointestinal cancer is RIPK4. RIPK4 is a member of RIP family and affects on the regulation of MAPK oncogene signaling pathway (via interaction with MAP2K1/2, and KRAS). Moreover, its alteration in the DNA and RNA levels has been seen in several cancer types such as hepatocellular carcinoma and cervical cancer (Liu et al. 2015; Heim et al. 2015). Furthermore, RIPK4 involves in tumor development, and invasion processes (via interaction with PTPN11, KLF4, WT1, ATR, PTEN, and ERBB2IP) (Qi et al. 2018). However, there is a contradiction in the RIPK4 gene expression reports. Some studies (Azizmohammadi et al. 2017) introduced RIPK4 as an activator of MAPK and declared that RIPK4 expression increases during cancer progression (Azizmohammadi et al. 2017), while others introduced it as a tumor suppressor that indicates down-regulation (Wang et al. 2014). According to these notes, RIPK4 seems to have a tissue-specific behavior. Based on TCGA, the highest mutation rate for RIPK4 has benn seen in the endometrial carcinoma, and B cell lymphoma (cancer instute N 2018b); however, the mutation rate in the stomach cancer is also approximately high (4.09%). Despite the copy number variation that is detected in RIPK4 gene in colorectal cancer, the most frequent mutation in the gene is missense substitution (cancer instute N 2018b). Unfortunately, we couldn't make any prediction for RIPK4 because of huge controversy in its alteration pattern in different cancers. This controversy has also seen in the cancers from the same embryonic germ layer (such as other digestive system organs). Thus, alteration of RIPK4 gene expression in gastric cancer is obvious; however, the alteration direction is not clear for us.

Predicted driver genes of hereditary gastric cancer

To detect potential hereditary cancer driver genes, we extracted 13 genes related to this disease from DisGeNet dataset. In this case, a list of 181 candidate driver genes is introduced by algorithm (Additional file 2). Similar to Fig. 2, we used the STRING database to draw a graphical representation of these candidate driver genes (Fig. 3). According to Intogen, we found 10 out of 181 genes (approximately 6% for hereditary gastric cancer), approved as gastric driver genes.

Our method nominated three potential driver genes for hereditary gastric cancer: UBA3, UBE2M, RRAGA. Two potential driver genes nominated by our method for hereditary gastric cancer are members of a big protein family relating to the ubiquitination process and proteolysis, named "ubiquitin-like modifier activating enzyme 3 (UBA3)" and "ubiquitin-conjugating enzyme E2M (UBE2M)". Although these two members have not been introduced as a driver gene in the oncogenesis until now, their up-regulation was detected in different types of cancer (Additional file 5: Table S5). UBA3 is a member of activator (E1s) enzymes, which participate in the regulation of NEDD8 and is an

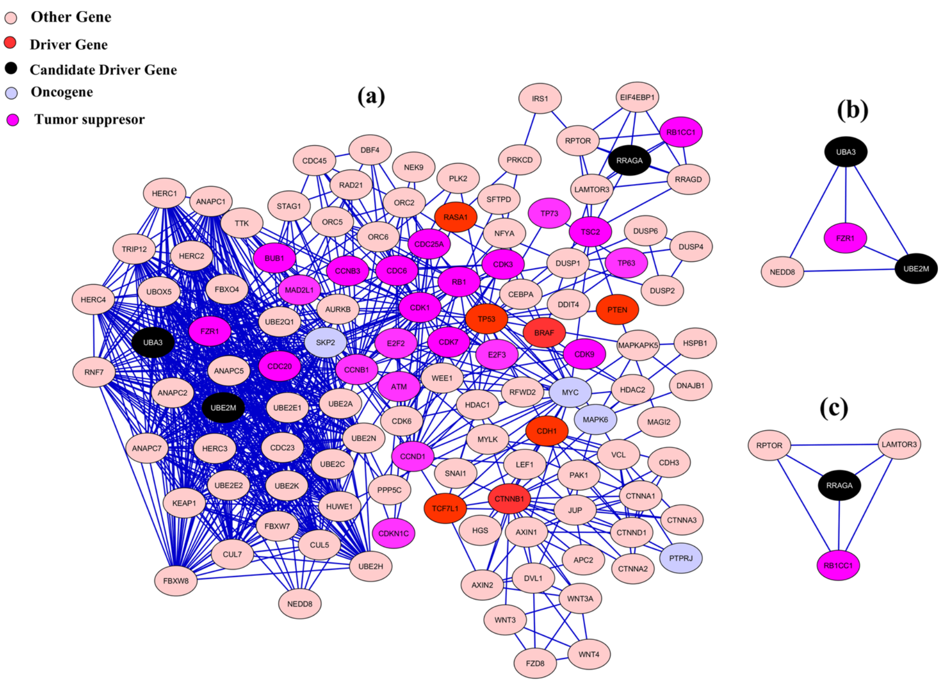

Fig. 3 Hereditary gastric cancer network. The STRING graph for hereditary Subtype of Gastric cancer based on the genes archived by our algorithm

important factor for cell cycle control and embryogenesis. Then, the activated NEDD8 binds to UBE2M, a conjugating (E2s) class enzyme, to regulate cell proliferation and cell cycle progression (Ying et al. 2018). Besides, UBA3 and UBE2M seems to involved in the inhibition of FZR1 (CDH1) that is a known tumor suppressor gene for hereditary gastric cancer (Zylberberg et al. 2018), via activation of its proteolysis. As depicted in Fig. 3b, FZR1 and NEDD8 are these genes neighbors.

In addition to the expression pattern alterations in UBA3 and UBE2M genes, several DNA mutations were indicated for them. The highest mutation rate for UBA3 has detected in the uterus and colorectal carcinomas, and the most frequent mutation for this gene is a small deletion at the 3' UTR (cancer instute N 2018c). The highest mutation rate for UBE2M also observed in the Uterus carcinoma, however, the B-cell lymphoma is in the second place, and interestingly, the most frequent mutation for it is also a small deletion in at the 3'UTR (cancer instute N 2018d). Since digestive system is mainly originated in the endodermal embryonic layer, it is probable that digestive system related cancers show relatively similar expression patterns. Thus, the observed up-regulation of UBE2M and UBA3 in esophageal squamous cell carcinomas and intrahepatic cholangiocarcinoma could be repeated in gastric cancer as another organ of the digestive system. On the other hand, recently researchers proved the estrogen receptors (ERα, ERβ) involvement in the progression of gastric cancer by affecting P21, P27, P53, and E-Cadherin expressions (Tang et al. 2017). Furthermore, several studies reported a decrease in the expression of these receptors related to poor prognosis of gastric cancer (Tang et al. 2017). UBE2M and UBA3 are factors that their overexpression downregulate the estrogen receptors expression (Fan et al. 2003). So, we predict an increased expression pattern for UBE2M and UBA3 in hereditary gastric cancer, while ER expression is down-regulated.

The third potential driver gene for hereditary gastric cancer is RRAGA. RRAGA is a member of Ras-related GTPase enzymes, which plays an important activator role in the mTOR pathway (a major regulator of cell growth and cell metabolic processes balance) (Petit et al. 2013). RRAGA has shown overexpression and hyper-activation in cancer cells as a result of removing its negative regulator (GATOR1 complex) (Bar-Peled et al. 2013). Thus, we predicted an increased pattern of expression for it in gastric cancer. This alteration makes the tumor cells capable of handling their changing metabolic needs (e.g., levels of energy and amino acids) (Petit et al. 2013; Porta-Pardo and Godzik 2014). Among potential driver genes (C_d) RRAGA's most important neighbors are RB1, LAMTOR3, and RPTOR (see Fig. 3c). Several gene mutations are also indicated for RRAGA. Based on TCGA, the highest RRAGA mutation rate has been detected in endometrial carcinomas, and the most frequent mutation in this gene is inappropriate substitution (cancer instute N 2018).

Melanoma predicted driver genes

In case of Melanoma, the set of related genes contained 23 member, which were extracted from the DisGeNet database. Similar to previous sections, we obtained a list of 148 candidate driver genes (Fig. 4a). The full list of these genes and their *p*-values can be found in the Additional file 3: Table S3. We found 8 of 148 genes (approximately 5.5 percent) to be proved melanoma-driver genes compared to intogen. Our method has nominated

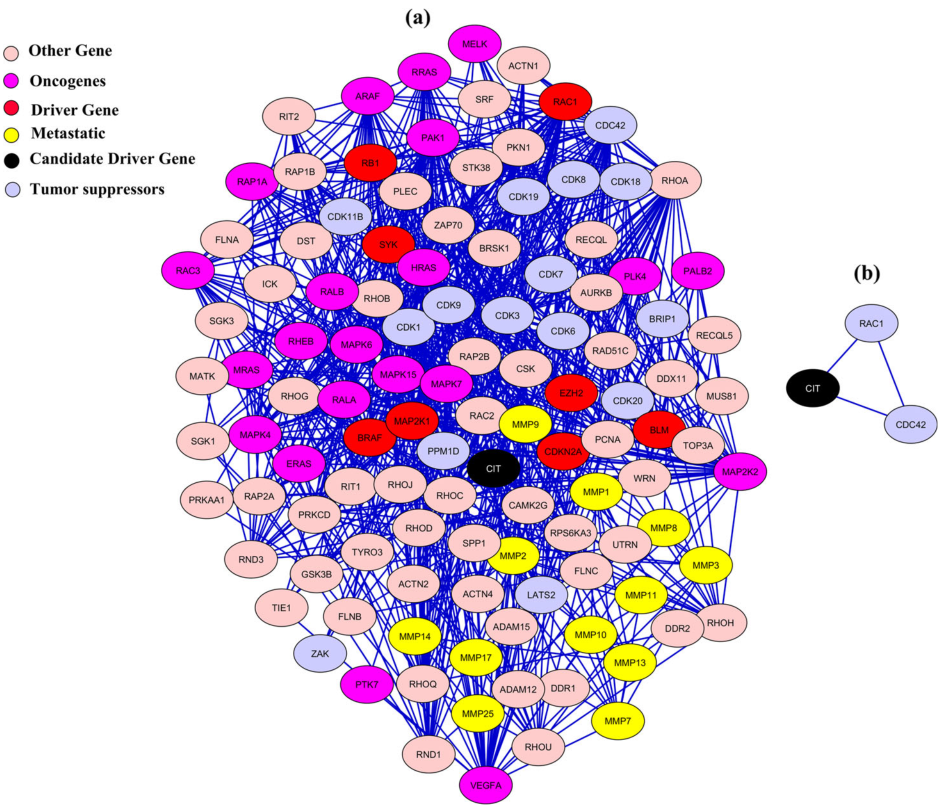

Fig. 4 Melanoma Network. The STRING graph for invasive subtype of melanoma based on the genes archived by our algorithm

one potential driver gene for melanoma, CIT, based on the four criteria. CIT is involved in the cell cycle via regulation of cytokinesis. Although the main function of CIT in cancer is not clear, its overexpression has been observed in some cancers (Wu et al. 2017). On the other hand, the signaling pathway in which CIT is involved was also undetected. According to RAC1 (CIT neighbor) we suggest that it should be a member of Wnt pathway. Furthermore, according to CDC42 (the other detected neighbor) it seems that CIT has a role in cell cycle regulation. Regarding these points and based on the behavior of CIT in the colon cancer (a cancer with the same embryonic germ lines as melanoma), we have predicted an increase of CIT expression whether due to the genomic changes or the post-transcriptional alterations. Based on TCGA, the highest mutation rate for CIT has seen in the uterus endometrial carcinoma (cancer instute N 2018a); however, melanoma is at the second place with an approximately high mutation rate (9.38%) and the most common mutation in this gene is a deletion at the gene's 3'UTR location (cancer instute N 2018a). Additional file 4: Table S4 and Additional file 5: Table S5 summarize the value of each criterion for the potential driver genes.

Discussion

Recently, by the easier generation of high-dimensional molecular profiling data sets, the challenge has shifted toward better analysis of this information. Especially, the integration of different omic datasets is a beneficial idea for detecting hidden rules of biological systems. In this paper, we presented TOPDRIVER, a network-based approach to nominate cancer driver genes in Gastric cancer and Melanoma. A machine learning method used in this approach to integrate multiple omic data to create a weighted network. Then, combinational and biological strategies selected potential driver genes. Previous studies indicated that gastric cancer can occur by activation of both intra- and extracellular reactions. The intracellular reactions depend on the function of cell cycle and cell growth regulators such as oncogenes and tumor suppressor genes. The extracellular pathway, however, is activated via infectious agents such as Helicobacter Pylori (H.Pylori) that may also activate the inflammatory pathway (Karimi et al. 2014). Extracellular reactions are not a common cause of hereditary gastric cancer, although in non-hereditary cases they play an important role. As shown in Figs. 2 and 3, two different inflammatory and intracellular pathways have been significantly separated by our method. It seems, therefore, that another ability of our algorithm is to distinguish the importance of inflammatory genes in the occurrence of non-hereditary gastric cancer, while there is no separation between inflammatory and intracellular pathways in the hereditary graph. On the other hand, there is a significant difference between gene lists obtained for hereditary gastric cancer and gastrointestinal cancer; though some similar genes are scored with different p-values in each subtype of gastric cancer by the algorithm. Our further investigations indicated that genes of high importance in the incidence of hereditary gastric cancer (such as CDH1 (Zylberberg et al. 2018)) have been placed on the top gene list of subtype-related genes, whereas these genes do not exist on the non-hereditary top gene list or do not have a suitable p-value. Our method could differentiate the invasive subtype of the disease from non-invasive ones and find the top genes independently for each subtype; as a large number of genes detected in an invasive subtype of melanoma were genes involved in the processes of metastasis and cell migration.

One of the strong points of this algorithm is that it has built the initial network with a very few genes and then expanded the network. More precisely, it consisted of 13 basic genes for hereditary subtype (including CDH1, TP53, etc), and 11 basic genes for gastrointestinal subtype of gastric cancer (including TP53, TGFA, etc), and 23 basic genes for Melanoma. TOPDRIVER nominated seven genes as potential cancer driver gene (3 for hereditary, and 3 for gastrointestinal cancer subtypes, and 1 for melanoma), which are involved in tumorgenesis process and predicted their probable alteration. Moreover, we found 17 genes (CCND1, TP73, E2F2, HDAC1, HDAC2, ATM, TP53, MYC, BRAF, RB1, TLR1, PTEN, PAK1, TLR6, E2F3, TP63, TP73), which were common among our potential driver genes of gastric cancer subtypes and were approved for cancer involvement. Of these 17 genes, three (BRAF, PAK1, and RB1) were also common among subtypes of gastric cancer and melanoma. Taking together, TOPDRIVER has introduced a set of potential, subtype specific driver genes for each cancer. However, TOPDRIVER potential genes should be experimentally validated that is our future program. Investigation of cancer in its early stages, for example using gene expression techniques and its comparison with normal samples may be beneficial in the diagnosis of cancer. Biotechnological techniques such as over-expressing the potential gene in a normal cell and evaluating the oncogenesis process or knocking down the potential gene using siRNA could be effective in determining the exact value of the genes as a marker in both oncogenesis and early cancer detection.

Conclusion

Oncogenomics attempts to characterize genes that drive cancer. These genes are oncogene or tumor suppressor and affect the initiation and progression of cancer. Identifying and targeting an individual patient's driver genes can lead to increased efficacy of treatment. In this paper, we presented a new network-based algorithm to detect cancer driver genes. TOPDRIVER integrated four omic datasets and constructed the initial network. The network weighted by application of a gene-disease association dataset and the molecular functional disease similarity. Then, the network weighs balanced with the diffusion kernel method. Finally, the induced graph of the desired cancer subtype investigated according to four biological criteria to introduce the potential driver genes.

Additional files

Additional file 1: TOPDRIVER gastrointestinal genes and their *p*-values. (PDF 27 kb)

Additional file 2: TOPDRIVER hereditary genes and their *p*-values. (PDF 40 kb)

Additional file 3: TOPDRIVER melanoma genes and their *p*-values. (PDF 35 kb)

Additional file 4: The Main characteristics of potential driver genes. The main characteristics of our potential genes based on the function and gene neighborhood criteria. The last two column indicate the gene mutation rate in the gastric cancer and the KEGG pathways which are related to our selected genes. (PDF 231 kb)

Additional file 5: The status of alteration of potential driver genes in the cancers. (PDF 267 kb)

Abbreviations

CNA: Copy number aberrations; GO: Gene ontology; GSP: Gene similarity profile; GTEx: The genotype-tissue expression; HPRD: Human protein reference database; KEGG: Kyoto encyclopedia of genes and genomes; NCBI: National center for biotechnology information; PPI: Protein-protein interaction; RefSeq: Reference sequence; TCGA: The cancer genome atlas; TSG: Tumor suppressor gene

Authors' contributions

SMR, FR, SHR and CE participated in the design of the study. SMR carried out data acquisition and computational analysis. FR carried out biological interpretation results of manuscript. SHR and FR wrote the manuscript. All authors read and approved the final manuscript.

Funding
No funding

Availability of data and materials
Model design and parameterization are fully detailed in the main text and supplementary information.

Competing interests
The authors declare that they have no competing interests.

Received: 29 November 2018 Accepted: 30 August 2019
Published online: 18 October 2019

References

Azizmohammadi S, Azizmohammadi S, Safari A, Kaghazian M, Sadrkhanlo M, Behnod V, et al. (2017) High-level expression of RIPK4 and EZH2 contributes to lymph node metastasis and predicts favorable prognosis in patients with cervical cancer. Oncol Res Featuring Preclinical Clin Can Ther 25(4):495–501

Babaei S, Hulsman M, Reinders RM, Jeroen d (2013) Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. BMC Bioinformatics 14(1):29

Bar-Peled L, Chantranupong L, Cherniack AD, Chen WW, Ottina KA, Grabiner BC, et al. (2013) A Tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. Science 340(6136):1100–1106

Benjamini Y, Hochberg Y (1995) ontrolling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing NJournal of The Royal Statistical Society Series. J R Stat Soc B 57:289–300

Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci 104(50):20007–20012

Cai S, Cai J, Jiang WG, Ye L (2017) Kidins220 and tumour development: Insights into a complexity of cross-talk among signalling pathways. Int J Mol Med 40(4):965–971

Camacho N, Van Loo P, Edwards S, Kay JD, Matthews L, Haase K, et al. (2017) Appraising the relevance of DNA copy number loss and gain in prostate cancer using whole genome DNA sequence data. PLoS Genet 13(9):e1007001

cancer instute N (2018) CIT in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000122966

cancer instute N (2018) KDINS220 in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000134313

cancer instute N (2018) KIDINS220 in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000134313

cancer instute N (2018) NKX3-1 in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000167034

cancer instute N (2018) RRAGA in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000155876

cancer instute N (2018) RRAGA in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000122966

cancer instute N (2018) RIPK4 in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000183421

cancer instute N (2018) UBA30 in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000144744

cancer instute N (2018) UBE2M in GDC data protal. https://portal.gdc.cancer.gov/genes/ENSG00000130725

Capriotti E, Altman RB (2011) A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics 98(4):310–317

Carvalho D, Mackay A, Bjerke L, Grundy RG, Lopes C, Reis RM, et al. (2014) The prognostic role of intragenic copy number breakpoints and identification of novel fusion genes in paediatric high grade glioma. Acta Neuropathol Commun 2(1):23

Charames GS, Bapat B (2003) Genomic instability and cancer. Curr Mol Med 3(7):589–596

Chen Y, Hao J, Jiang W, He T, Zhang X, Jiang T, et al. (2013) Identifying potential cancer driver genes by genomic data integration. Sci Rep 3:3538

Cheng L, Li J, Ju P, Peng J, Wang Y (2014) SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. PloS ONE 9(6):e99415

Chung F (1997) Spectral graph theory. Number 92 in regional conference series in mathematics. Am Math Soc. https://doi.org/10.1090/cbms/092

Ciriello G, Cerami E, Sander C, Schultz N (2011) Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. https://doi.org/10.1101/gr.125567.111

Consortium GO (2012) Gene Ontology annotations and resources. Nucleic Acids Res 41(D1):D530–D535

Fan M, Bigsby RM, Nephew KP (2003) The NEDD8 pathway is required for proteasome-mediated degradation of human estrogen receptor (ER)-α and essential for the antiproliferative activity of ICI 182,780 in ERα-positive breast cancer cells. Mol Endocrinol 17(3):356–365

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. (2004) A census of human cancer genes. Nat Rev Can 4(3):177

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446(7132):153

GTEx Consortium, et al. (2015) The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science 348(6235):648–660

Haverty PM, Fridlyand J, Li L, Getz G, Beroukhim R, Lohr S, et al. (2008) High-resolution genomic and expression analyses of copy number alterations in breast tumors. Genes Chromosomes Can 47(6):530–542

Heim D, Cornils K, Schulze K, Fehse B, Lohse A, Brümmendorf T, et al. (2015) Retroviral insertional mutagenesis in telomerase-immortalized hepatocytes identifies RIPK4 as novel tumor suppressor in human hepatocarcinogenesis. Oncogene 34(3):364

Hou JP, Ma J (2014) DawnRank: discovering personalized driver genes in cancer. Genome Med 6(7):56

Ji X, Tang J, Halberg R, Busam D, Ferriera S, Peña MMO, et al. (2010) Distinguishing between cancer driver and passenger gene alteration candidates via cross-species comparison: a pilot study. BMC Can 10(1):426

Jiang J, Liu Z, Ge C, Chen C, Zhao F, Li H, et al. (2017) NK3 homeobox 1 (NKX3. 1) up-regulates forkhead box O1 expression in hepatocellular carcinoma and thereby suppresses tumor proliferation and invasion. J Biol Chem 292(47):19146–19159

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2015) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44(D1):D457–D462
Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F (2014) Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. Can Epidemiol Prev Biomark 23(5):700–713
Lafferty R, Kondor J (2002) Diffusion kernels on graphs and other discrete structures. In: Proceedings of the Nineteenth International Conference on Machine Learning (11). pp 315–322
Liu DQ, Li FF, Zhang JB, Zhou TJ, Xue WQ, Zheng XH, et al. (2015) Increased RIPK4 expression is associated with progression and poor prognosis in cervical squamous cell carcinoma patients. Sci Rep 5:11955
Luo P, Tian LP, Ruan J, Wu F (2017) Disease gene prediction by integrating PPI networks, clinical RNA-Seq data and OMIM data. IEEE/ACM Trans Comput Biol Bioinformatics. https://doi.org/10.1109/tcbb.2017.2770120
Martinez EE, Darke AK, Tangen CM, Goodman PJ, Fowke JH, Klein EA, et al. (2014) A functional variant in NKX3. 1 associated with prostate cancer risk in the Selenium and Vitamin E Cancer Prevention Trial (SELECT). Can Prev Res:0075. https://doi.org/10.1158/1940-6207.capr-14-0075
Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, et al. (2012) PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics 28(18):i640–i646
Nodouzi V, Nowroozi M, Hashemi M, Javadi G, Mahdian R (2015) Concurrent down-regulation of PTEN and NKX3. 1 expression in Iranian patients with prostate cancer. Int Braz J Urol 41(5):898–905
Petit CS, Roczniak-Ferguson A, Ferguson SM (2013) Recruitment of folliculin to lysosomes supports the amino acid–dependent activation of Rag GTPases. J Cell Biol 202(7):1107–1122
Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. (2016) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res:gkw943. https://doi.org/10.1093/nar/gkw943
Pole J, Courtay-Cahen C, Garcia M, Blood K, Cooke S, Alsop A, et al. (2006) High-resolution analysis of chromosome rearrangements on 8p in breast, colon and pancreatic cancer reveals a complex pattern of loss, gain and translocation. Oncogene 25(41):5693
Porta-Pardo E, Godzik A (2014) e-Driver: a novel method to identify protein regions driving cancer. Bioinformatics 30(21):3109–3114
Pruitt KD, Tatusova T, Maglott DR (2006) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35(suppl_1):D61–D65
Qi ZH, Xu HX, Zhang SR, Xu JZ, Li S, Gao HL, et al. (2018) RIPK4/PEBP1 axis promotes pancreatic cancer cell migration and invasion by activating RAF1/MEK/ERK signaling. Int J Oncol 52(4):1105–1116
Ramsahai E, Walkins K, Tripathi V, John M (2017) The use of gene interaction networks to improve the identification of cancer driver genes. PeerJ 5:e2568
Reimand J, Bader GD (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol 9(1):637
Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H (2013) Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics 14(1):190
Sakamoto K, Imamura T, Kanayama T, Yano M, Asai D, Deguchi T, et al. (2017) Ph-like acute lymphoblastic leukemia with a novel PAX5-KIDINS220 fusion transcript. Genes Chromosomes Can 56(4):278–284
Sanchez-Garcia F, Villagrasa P, Matsui J, Kotliar D, Castro V, Akavia UD, et al. (2014) Integration of genomic data enables selective discovery of breast cancer drivers. Cell 159(6):1461–1475
Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS (2010) A census of amplified and overexpressed human cancer genes. Nat Rev Can 10(1):59
Sudhakar A (2009) History of cancer, ancient and modern treatment methods. J Can Sci Ther 1(2):1
Tang W, Liu R, Yan Y, Pan X, Wang M, Han X, et al. (2017) Expression of estrogen receptors and androgen receptor and their clinical significance in gastric cancer. Oncotarget 8(25):40765
Tomczak K, Czerwińska P, Wiznerowicz M (2015) The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemp Oncol 19(1A):A68
Vandin F, Upfal E, Raphael BJ (2011) De novo discovery of mutated driver pathways in cancer. Genome Res. https://doi.org/10.1007/978-3-642-20036-6_44
Wang X, Tao T, Sun JT, Shakery A, Zhai C (2008) Dirichletrank: Solving the zero-one gap problem of pagerank. ACM Trans Inf Syst (TOIS) 26(2):10
Wang X, Zhu W, Zhou Y, Xu W, Wang H (2014) RIPK4 is downregulated in poorly differentiated tongue cancer and is associated with migration/invasion and cisplatin-induced apoptosis. Int J Biol Mark 29(2):150–159
Wu Z, Zhu X, Xu W, Zhang Y, Chen L, Qiu F, et al. (2017) Up-regulation of CIT promotes the growth of colon cancer cells. Oncotarget 8(42):71954
Yang H, Wei Q, Zhong X, Yang H, Li B (2016) Cancer driver gene discovery through an integrative genomics approach in a non-parametric Bayesian framework. Bioinformatics 33(4):483–490
Ying J, Zhang M, Qiu X, Lu Y (2018) Targeting the neddylation pathway in cells as a potential therapeutic approach for diseases. Can Chemother Pharmacol:1–12. https://doi.org/https://doi.org/10.1007/s00280-018-3541-8
Yuan F, Lu W (2018) Prediction of potential drivers connecting different dysfunctional levels in lung adenocarcinoma via a protein–protein interaction network. Biochim Biophys Acta (BBA) Mol Basis Dis 1864(6):2284–2293
Zylberberg HM, Sultan K, Rubin S (2018) Hereditary diffuse gastric cancer: One family's story. World J Clin Cases 6(1):1

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.