genes

MDPI

Review

Network-Based Approaches to Explore Complex Biological Systems towards Network Medicine

Giulia Fiscon [1,2], Federica Conte [1,2], Lorenzo Farina [3] and Paola Paci [1,2,*]

[1] Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council, via dei Taurini 19, 00185 Rome, Italy; giulia.fiscon@iasi.cnr.it (G.F.); federica.conte@iasi.cnr.it (F.C.)

[2] SysBio Centre of Systems Biology, Piazza della Scienza, 3, 20126 Milan, Italy

[3] Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Viale Ariosto 25, 00185 Rome, Italy; farina@diag.uniroma1.it

* Correspondence: paola.paci@iasi.cnr.it

Received: 3 August 2018; Accepted: 30 August 2018; Published: 31 August 2018

Abstract: Network medicine relies on different types of networks: from the molecular level of protein–protein interactions to gene regulatory network and correlation studies of gene expression. Among network approaches based on the analysis of the topological properties of protein–protein interaction (PPI) networks, we discuss the widespread DIAMOnD (disease module detection) algorithm. Starting from the assumption that PPI networks can be viewed as maps where diseases can be identified with localized perturbation within a specific neighborhood (i.e., disease modules), DIAMOnD performs a systematic analysis of the human PPI network to uncover new disease-associated genes by exploiting the connectivity significance instead of connection density. The past few years have witnessed the increasing interest in understanding the molecular mechanism of post-transcriptional regulation with a special emphasis on non-coding RNAs since they are emerging as key regulators of many cellular processes in both physiological and pathological states. Recent findings show that coding genes are not the only targets that microRNAs interact with. In fact, there is a pool of different RNAs—including long non-coding RNAs (lncRNAs) —competing with each other to attract microRNAs for interactions, thus acting as competing endogenous RNAs (ceRNAs). The framework of regulatory networks provides a powerful tool to gather new insights into ceRNA regulatory mechanisms. Here, we describe a data-driven model recently developed to explore the lncRNA-associated ceRNA activity in breast invasive carcinoma. On the other hand, a very promising example of the co-expression network is the one implemented by the software SWIM (switch miner), which combines topological properties of correlation networks with gene expression data in order to identify a small pool of genes—called switch genes—critically associated with drastic changes in cell phenotype. Here, we describe SWIM tool along with its applications to cancer research and compare its predictions with DIAMOnD disease genes.

Keywords: bioinformatics; network medicine; gene co-expression network; regulatory network; ceRNA; PPI network

1. Network Medicine: An Emergent Paradigm in Medicine

The exploitation of the emerging network-based approaches to medicine enables multiple potential biological and clinical applications by offering an intuitive and reliable way to explore systematically the molecular complexity of a particular disease and thus leading to the identification of disease genes as potential drug targets and biomarkers. A disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex network of intracellular and intercellular interactions. This entirely new perspective, in which critical biological factors are nearly always the result of multiple pathobiological pathways that interact through an interconnected network to control disease pathobiology, has been able to kick-start a new medical paradigm called "Network Medicine" [1]. The fundamental tenet of network medicine is to look at diseases as perturbations within the interactome, i.e., the comprehensive network map of molecular components and their interactions [1,2]. The overall ambition is both to develop a global understanding of how interactome perturbations result in disease traits, and to translate computational insights into concrete clinical applications, such as new drugs and therapies or diagnostic tools. Furthermore, the representation of biological complex systems as networks enables the visualization of the interactome underlying structure, revealing new functional roles, and proposing new and fresh interpretations of data.

A network is a set of nodes and edges, where nodes are linked together if a kind of interaction occurs between them. As witnessed by the last-years increasing number of publications (Figure 1), most attention has been recently directed towards molecular networks, including:

1. Protein–protein interaction (PPI) networks, whose nodes are proteins that are linked to each other by physical interactions [3,4];
2. Regulatory networks, whose directed links represent regulatory relationships between a transcription factor and a gene [5];
3. Co-expression networks, in which genes with similar co-expression patterns are linked [6].

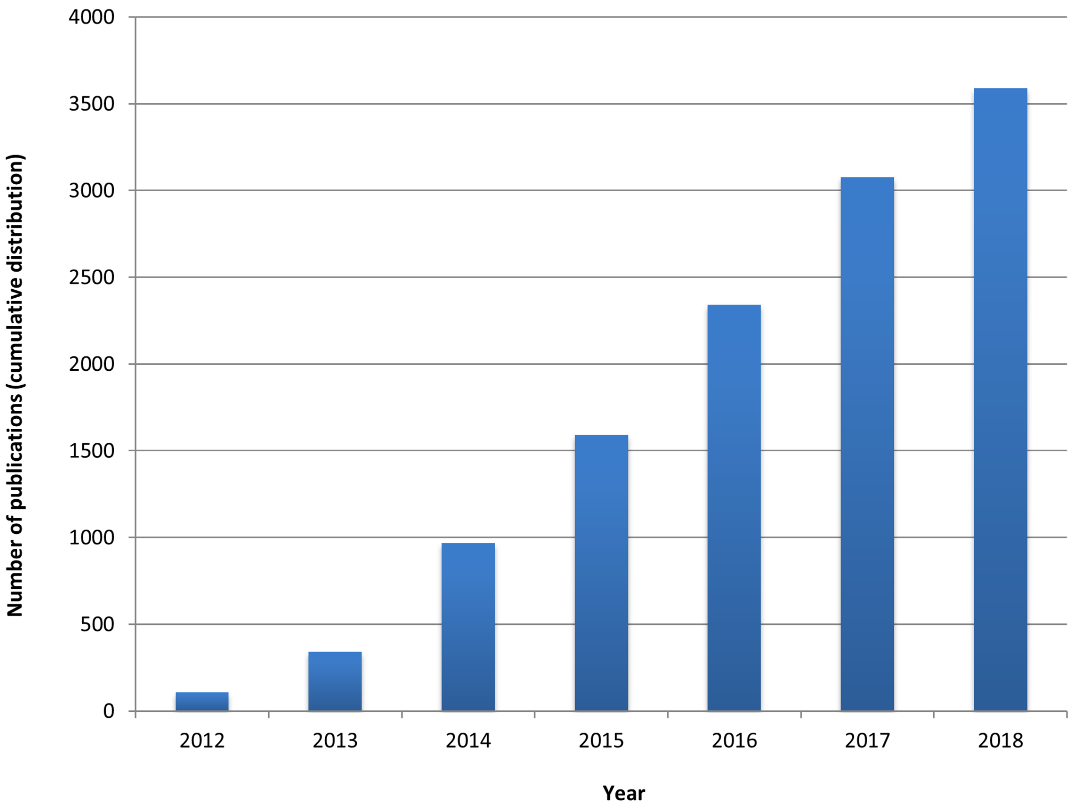

Figure 1. Number of 2012–2018 publications related to network-based approaches to medicine. The figure shows the number of articles published by year obtained by searching the following specific keywords in Pubmed: network-based approach, network medicine, biological network.

Table 1 summarizes some of the widespread disease-gene prediction methods/tools based on the analysis of the topological properties of the above-mentioned molecular networks.

The PPI-network based approaches can be mainly classified into: (i) local methods, based on searching for direct interactions between candidate genes and known disease genes (e.g., Oti et al. [7], GenePANDA [8]); (ii) global methods, that model how the information flow in the cell to assess the proximity and connectivity between known disease genes and candidate genes (e.g., DADA [9], DIAMOnD [10], PRINCE [11], ProDiGe [12]). In this review, we provide a detailed description of DIAMOnD [10], the most used and newest approach at the state of the art.

In the past, a great effort has been devoted to understanding the molecular mechanism of post-transcriptional regulation with a special emphasis on non-coding RNAs (ncRNAs) since they are emerging as key regulators of many cellular processes in both physiological and pathological states [13–16]. This class of RNA species appears really heterogeneous, including the intensively studied microRNAs (miRNAs)—small non-coding RNAs of 20–22 nucleotides long—as well as the most recent acknowledged long non-coding RNAs (lncRNAs)—non-protein coding RNAs greater than 200 nucleotides in length and lacking of extended open reading frames [17–19]. Recent findings show that coding genes are not the only targets that miRNAs interact with. In fact, there is a pool of different RNAs competing with each other to attract miRNAs for interactions, thus acting as competing endogenous RNAs (ceRNAs). This intriguing mechanism, also known as "target mimicry" process, was first discovered in plants [20]. Crucial triggers of this new layer of post-transcriptional regulation are "decoys"—or miRNA "sponges"—including both coding and non-coding RNAs, such as pseudogenes, lncRNAs, large intergenic ncRNAs, and circular RNAs [21–24]. Sponges exert their decoy activity by recruiting miRNA molecules via base-pairing with miRNA-recognition elements (MREs), which they share with a target, subsequently causing release of the target from miRNA control. The framework of regulatory networks provides a powerful tool to gather new insights into ceRNA regulatory mechanisms. Among computational methods describing miRNA-sponge interactions by exploiting gene regulatory networks approach, the ceRNA model of Paci et al. [25] stood out as the best method in terms of the percentage of discovered miRNA-sponge interactions associated with breast cancer, according to a thorough and comparison study proposed in [26]. In this review, we provide a detailed description of ceRNA model implemented by Paci et al. [25] along with its application to human breast invasive carcinoma [27].

The framework of co-expression networks provides a powerful tool to gather biologically relevant information from the interaction structure underpinning patterns of gene expression, thus accelerating the interpretation of molecular mechanisms at the root of significant biological processes. A new promising approach based on the co-expression network is the one implemented by SWIM (SWItch Miner) [28], a software able to identify key genes in a network of interactions of various sorts by defining appropriate roles of genes according to their local/global positioning in the overall network. The latter property being of crucial importance, given that, recently, it has been shown that genes associated with a disease are localized in specific neighborhoods, or disease modules, within the interactome [1]. SWIM builds the co-expression network by using the Pearson correlation coefficient between the expression profiles of any pair of genes. Nodes in this network are RNA transcripts and a link occurs between two nodes if their expression profiles are highly correlated or highly anti-correlated. In order to identify disease modules, SWIM makes use of network clustering algorithms like k-means and, then, assigns a role to each node in the network according to its inter- and intra-cluster interaction. Understanding diseases in the context of these networks allows to address some fundamental properties of the disease-specific genes, which are called "switch genes" by SWIM. In this review, we provide a detailed description of SWIM along with its applications in viticulture [29] and oncology [28,30].

Notably, another tool based on the co-expression network is WGCNA (Weighted Correlation Network Analysis) [31], a comprehensive collection of R functions for performing various aspects of gene co-expression network analysis on high-dimensional data, including functions for network construction, module detection, gene selection, calculations of topological properties, data simulation and visualization.

This review is organized as follows: in section 2, we detail DIAMOnD algorithm developed by Barabasi and co-authors [10] to discover new candidate disease genes associated with a particular phenotype; in section 3, we detail the ceRNA model proposed by Paci et al. [25,27] and its application to human breast cancer; in section 4, we describe the software SWIM (SWItch Miner) [28] and its applications in viticulture [29] and oncology [28,30].

Table 1. Overview of network-based approaches to medicine.

Method/Tool	**Brief Description**	**Availability**	**Reference**
Protein–Protein Interaction Network			
Oti et al.	It identifies new candidate disease genes by searching for disease proteins having interaction partners located within loci associated with the same disease	Prediction results available	[7]
GenePANDA (Gene Prioritizing Approach using Network Distance Analysis)	It identifies new candidate disease genes based on their relative distance to known disease genes in a functional association network	Prediction results available	[8]
DADA (Degree-Aware Disease Gene Prioritization)	It prioritizes candidate disease genes with respect to a disease of interest based on network proximity measure, calculated by using Random Walk with Restarts algorithm [32] with some statistical adjustment	MATLAB software package	[9]
DIAMOnD (DIseAse MOdule Detection)	It identifies full disease modules around a set of known disease proteins by performing a systematic analysis of the PPI-network that exploits the “connectivity significance” instead of local connection density	Python software package	[10]
PRINCE (PRIoritizatioN and Complex Elucidation)	It prioritizes genes related to a query disease based on their closeness, in the PPI-network, to genes causing phenotypically similar disorders to the query disease	Cytoscape Plug-in	[11,33]
ProDiGe (Prioritization Of Disease Genes)	It implements a novel machine learning strategy for gene prioritization based on learning from a set of positive examples (e.g., known disease genes) and unlabeled examples (e.g., candidate genes), allowing heterogeneous data integration	MATLAB software package	[12]
Regulatory Network			
MMI-network (MiRNA-Mediated Interactions network)	ceRNA model based on partial association to investigate the role of lncRNAs as miRNA sponges in human breast cancer. It computes for each triplet (lncRNA, miRNA, messenger RNA (mRNA)) the difference between Pearson correlation of (lncRNA, mRNA) and partial correlation (lncRNA, mRNA | miRNA) to examine the contribution of the miRNA into the lncRNA/mRNA relationship	Prediction results available	[25,27]
PANDA (Passing Attributes between Networks for Data Assimilation)	It implements a message-passing model using multiple sources of information to predict regulatory relationships, and used it to integrate protein–protein interaction, gene expression, and sequence motif data to reconstruct genome-wide, condition-specific regulatory networks in yeast as a model	MATLAB/ R/Python software packages	[34]
Sonawane et al.	It uses PANDA to infer gene regulatory networks for 38 different tissues by integrating GTEx RNA-sequencing (RNA-seq) data with a canonical set of transcription factors to target gene edges and protein–protein interactions	Prediction results available	[35]
Co-Expression Network			
SWIM (Switch Miner)	Wizard-like software that integrates gene expression data with network topological properties for identifying a small pool of genes (i.e., switch genes) critically associated with drastic changes in cell phenotype	MATLAB software package	[28–30]
WGCNA (Weighted Correlation Network Analysis)	Collection of R functions for performing weighted correlation network analysis of large data sets, including functions for network construction, module identification, topological properties calculation, data manipulation and visualization	R software package	[31]

2. DIseAse MOdule Detection (DIAMOnD)

Recently, several studies have shown how cellular components associated with a specific disease are not randomly scattered within the human interactome, but agglomerate in specific regions, suggesting the existence of specific “disease modules” for each disease. Barabasi et al. [1] proposed a research pipeline for the identification and validation of disease modules. In particular, for any specific disease the pipeline consists of the following steps:

- Interactome reconstruction merges the most up-to-date information on protein–protein interactions, co-complex memberships, regulatory interactions and metabolic network maps in the tissue and cell line of interest.
- Disease gene (seed) identification collects the known disease-associated genes obtained from linkage analysis, genome-wide association studies or other sources, which serve as the seed of the disease module.
- In disease module identification, the seed genes are placed on the interactome, with the aim of identifying a subnetwork that contains most of the disease-associated components, exploiting both the functional and topological modularity of the network.
- Pathway identification can be used in instances in which the number of components contained in the ascertained disease module is so large that it cannot serve as a tractable starting point for further experimental work.
- Disease modules are tested for their functional and dynamic homogeneity.

More recently, Barabasi and co-authors proposed a novel algorithm—DIAMOnD (DIseAse MOdule Detection) [10]—to uncover disease modules associated with a particular phenotype. By systematically analyzing the protein–protein interactions of 70 diseases, they showed that disease modules do not coincide with topological communities of densely interconnected proteins and instead identified the interaction significance as the key quantity to characterize the connection patterns among disease proteins. To extract disease modules, DIAMOnD algorithm starts from proteins known to be associated with a particular disease (i.e., seed proteins) and prioritizes the other proteins of the interactome having a significant fraction of their interactions with seed proteins. In particular, DIAMOnD encompasses the following steps:

1. For all proteins with at least one connection to any of the seed proteins, it calculates the "connectivity significance". Specifically, DIAMOnD uses the hypergeometric distribution to calculate the statistical significance of having drawn k_s seed proteins (out of k total draws) from a population of N proteins including s_0 seed proteins. The hypergeometric distribution is:

$$p(k,k_s) = \frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}} \quad (1)$$

   and then, the "connectivity significance" is obtained as:

$$p\text{-value}(k,k_s) = \sum_{k_i=k_s}^{k} p(k,k_i) \quad (2)$$

   In a network view, the population of N proteins corresponds to the nodes of the PPI-network and k are the nearest neighbors of a certain protein in the network. This set of nearest neighbors must include k_s seed proteins. Thus, $p(k,k_s)$ is the probability that a protein with a total of k links has exactly k_s links to seed proteins and p-value(k,k_s) is the probability that a protein with a total of k links has more connections to seed proteins than expected (Figure 2).
2. It ranks the proteins according to their respective p-values. The protein with the highest rank (i.e., lowest p-value) is called "candidate protein".
3. It adds the candidate protein to the set of seed proteins, increasing their number from s_0 to $s_1 = s_0 + 1$.
4. It iterates steps 1–3 with the expanded set of seed proteins, pulling one protein at a time into the growing disease module.

The procedure 1–4 can be repeated until the module spans across the entire network. The order in which the proteins are being added to the module provide a ranking of all proteins reflecting relevance association to the disease.

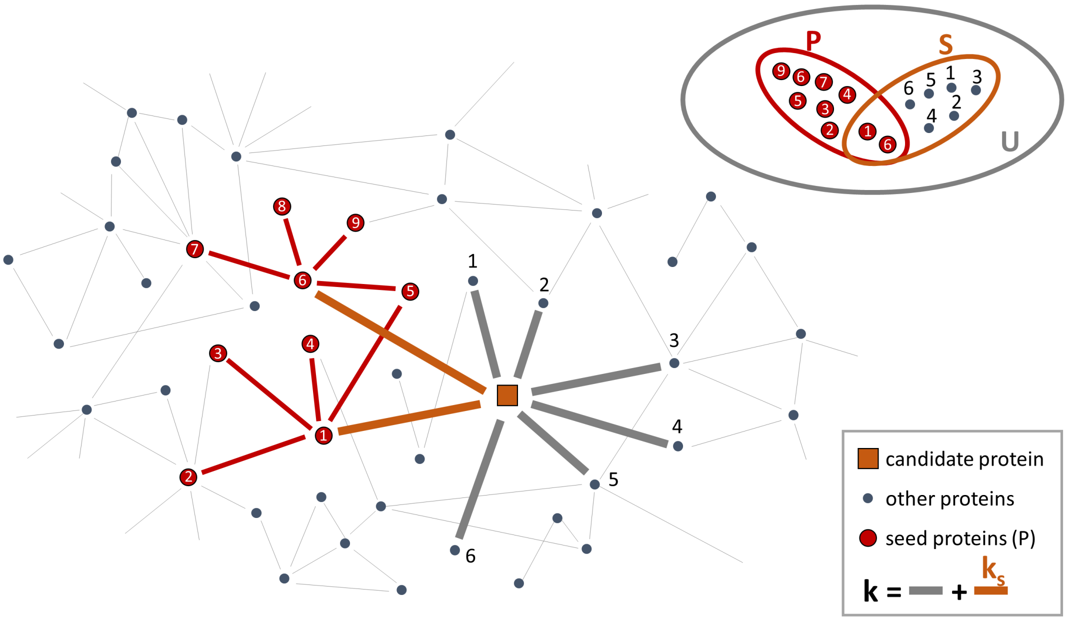

Figure 2. Sketch of step 1 of DIAMOnD. The network corresponds to the interactome where the red balls are the seed proteins, the orange square is the protein to test with k connections (orange and grey thick links) including k_s links to seed proteins (orange thick links), the grey balls refer to other proteins in the PPI-network. The sets at top-right correspond to: U is the ensemble of the total number of nodes in the PPI-network, S is the ensemble of the draw of k proteins, including k_s seed proteins ($k_s = 2$ in this example), P is the ensemble of the seed proteins.

3. miRNA-Mediated Interactions Network: A Competing Endogenous RNA Model Exploiting the Topological Properties of Regulatory Networks

Recent findings have identified ceRNAs as the drivers in many disease conditions, including cancers [22,36–39]. They indirectly regulate each other by reducing the amount of miRNAs available to target messenger RNAs via the binding of MREs [40]. Recently, a computational method [25] was developed for identifying putative lncRNAs acting as miRNAs sponges in human breast cancer. In this study [25], the authors used normalized level three RNA- and miRNA- sequencing expression data of breast cancer adenocarcinoma (brca) from IlluminaHiSeq platform that were retrieved from TCGA (The Cancer Genome Atlas) [41,42]. The study concerned 72 samples for which the complete sets of tumor and matched normal profiles (for both RNA-seq and miRNA-seq data) were available. The computational ceRNA model is based on three hypotheses:

i. RNAs competing for the same miRNA are marked by a highly positive correlation.
ii. Interaction between the RNAs competing for the same miRNA is indirect, i.e., mediated by miRNA.
iii. RNAs competing for the same miRNA harbor one or more MREs for the miRNA they sponge.

For what concerns the first hypothesis, the top-correlated messenger RNA (mRNA)/lncRNA pairs in normal and cancer data sets were selected by setting in both cases the correlation threshold to the 99th percentile of the corresponding overall correlation distribution (Figure 3A right).

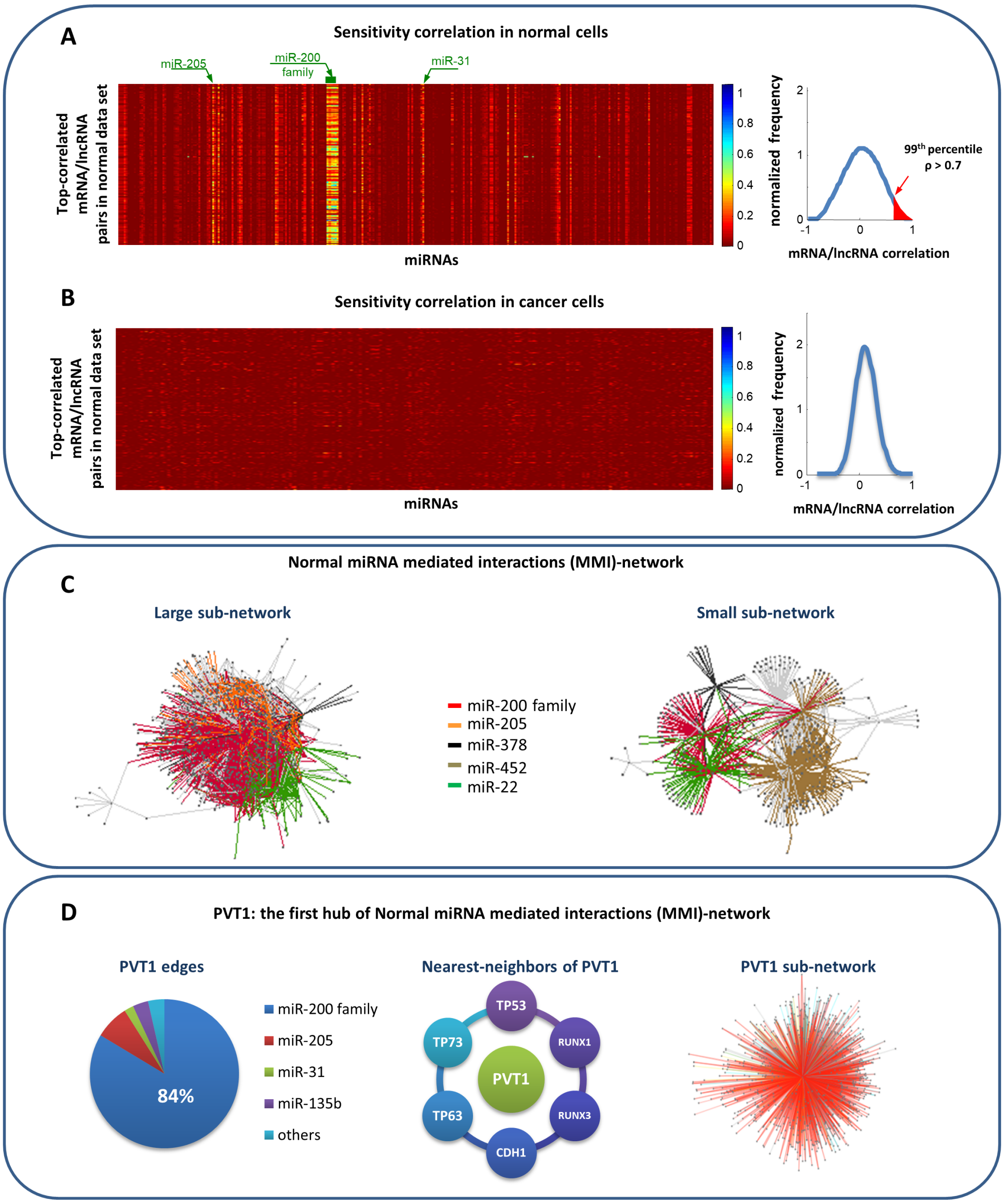

Figure 3. Results of Paci et al. model to predict miRNA sponge interactions in breast invasive carcinoma [25]. (**A**) Heatmap showing the sensitivity correlation for the top-correlated mRNA/lncRNA pairs (e.g., pairs for which the Pearson correlation between their expression profiles exceeds the 99th percentile of the overall correlation distribution) in normal breast tissues. Red color corresponds to zero sensitivity correlation, meaning that the interaction between the selected RNA pairs is direct and not mediated by miRNAs. Light vertical stripes point to miRNAs that are mediating the interaction, suggesting putative competing endogenous RNAs. (**B**) The same as in panel (**A**) but using data from breast cancer tissues. (**C**) The normal MMI-network (1738 nodes and 32,375 edges) built starting from the expression data of normal breast tissues. Nodes represent both mRNAs and lncRNAs; edges represent miRNAs that are mediating their interactions. Each pair of linked nodes fulfills two requirements: (i) sensitivity correlation > 0.3 and (ii) one or more shared MREs, for each miRNA linking them. Colors correspond to different miRNAs. (**D**) PVT1 subnetwork analysis. From left to right: the percentage of the miRNAs sponged by PVT1 with respect to all of its; some nearest neighbors of PVT1 that are well-known cancer genes; the sponge interactions sub-network of PVT1 (753 nodes and 2169 edges).

For what concerns the second hypothesis, to investigate the scenario in which specific miRNAs may mediate the interactions of the top-correlated mRNA/lncRNA pairs, a well-established tool of multivariate analysis (i.e., the partial correlation) was applied to each selected mRNA/lncRNA pair with respect to each miRNA in their dataset. In general, the partial correlation ($\rho_{XY|Z}$) measures the extent to which an observed correlation between two variables X and Y (here, the expression profiles of a mRNA and a lncRNA) relies on the presence of a third controlling variable Z (here, the expression profile of a miRNA) and it is computed as:

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}} \tag{3}$$

where ρ_{XY} is the Pearson correlation. Then, for each triplet mRNA/lncRNA/miRNA the *sensitivity correlation S*, defined as:

$$S = \rho_{XY} - \rho_{XY|Z} \tag{4}$$

was computed. The XYZ triplets with S > 0.3, corresponding to a drop of about the 30% in the correlation between XY when Z is removed, were selected. The sensitivity distribution of the top-correlated mRNA/lncRNA pairs (XY) is plotted removing one miRNA (Z) molecule at time (Figure 3A left).

For what concerns the third hypothesis, the triplets mRNA/lncRNA/miRNA that are enriched in binding sites of the shared miRNA (hypergeometric test *p*-value < 0.01) were selected by performing a seed match analysis. The minimal pairing requirement to predict a miRNA target recognition is a perfect match to positions 2–7 (6-mer miRNA seed) at the 5′-end of the mature miRNA sequence [43].

Integrating the results of multivariate statistical analysis and seed match analysis, the so-called miRNA-mediated interactions (MMI) network was built both in normal (Figure 3C) and cancer tissues [25]. Nodes in the networks represent mRNAs and lncRNAs with highly correlated expression profiles while edges represent miRNAs mediating their interactions. Normal MMI-network accounted for 1738 nodes and 32,375 edges (Figure 3C). Linked nodes are required to fulfill the above-mentioned hypothesis of the ceRNA model, which corresponds to the following mathematical constraints:

i. Matching high values of the Pearson correlation between their expression profiles ($\rho > 0.7$);
ii. Matching high values of the sensitivity correlation (S > 0.3);
iii. Sharing binding sites for miRNAs (6-mer miRNA seed match).

This study revealed the existence of a complex regulatory network in normal samples that appears to be missing in tumor samples (and vice-versa), highlighting a marked rewiring in the ceRNA program between normal and pathological breast tissues (Figure 3B). At the heart of this phenomenon, there was the recently and widely studied oncogene PVT1 [44–63] that switched from being the first of the hubs in the normal MMI network to fall outside the list of nodes of the cancer network. In normal network, PVT1 revealed a net binding preference towards the miR-200 family (Figure 3D), which antagonized to regulate the expression of hundreds of mRNAs that are known to be related to the cancer development and progression (e.g., GATA3, CDH1, TP53, TP63, TP73, RUNX1, and RUNX3). Despite its up-regulation in breast cancer tissues, mimicked by the miR-200 family members, PVT1 stopped working as ceRNA in the cancerous state.

The specific conditions required for a ceRNA landscape to occur are still far from being determined. However, in a recent study [27], the authors emphasized the importance of the relative concentration of the ceRNAs, and their related miRNAs. In particular, they focused on the withdrawal of the PVT1 ceRNA activity functioning as miR-200 sponge in breast cancer tissues, betting on a titration mechanism as the main culprit (i.e., large changes in the ceRNA expression levels either overcome, or relieve, the miRNA repression on competing RNAs; similarly, a very large miRNA over-expression may abolish competition). Firstly, they performed a gene expression and sequence analysis of PVT1 genomic locus, which revealed the existence of multiple isoforms representing all the possible configurations (Figure 4A): missing the binding site (e.g., Iso11 and Iso12 in Figure 4A); hosting the binding site for all (e.g., Iso1 in Figure 4A) or some members of the miR-200 family (e.g., Iso6 or Iso7 in Figure 4A). By performing the principal component analysis (PCA) using the feature abundance levels of all the PVT1 isoforms across normal and cancer samples, the authors found that two principal components (PCs) were able to explain more than the 80% of the variance of the data (Figure 4B left). In particular, the first PC–explaining about the 60% of the total variance of the analyzed data—referred to the variation of the TCONS_147501 isoform that, missing the binding site, did not interact with the miR-200 family; while the second PC—explaining by alone about the 20% of the total variance—corresponded to the variation of the TCONS_147426 isoform that, hosting the binding site for the miR-200b/200c/429 cluster, could be act as competitor of the targets of these miRNAs. By drawing the score plot (Figure 4B right), it emerged that the first PC and the second PC were able to to separate the contribution of the isoform missing the binding site for any members of the miR-200 family (i.e., TCONS_147501, blue isoform in Figure 4B) and of the isoform hosting the binding site for the miR-200b/200c/429 cluster (i.e., TCONS_147426, red isoform in Figure 4B) from all the others, respectively. Then, since the isoform harboring the binding site (i.e., TCONS_147426) and the isoform missing the binding site for the miR-200 family members (i.e., TCONS_147501) were the only isoforms that changed, the author evaluated the ratio between the abundance of each one with respect to one representative member of the miR-200b/200c/429 cluster (i.e., miR-200b) in both normal and cancer tissues (Figure 4C). From this analysis, they found that only the PVT1 isoform harboring the binding site for miR-200b showed a drastic decrease in its relative concentration with respect to the miRNA abundance from normal to cancer tissues, providing a plausibility argument to the breakdown of the sponge program orchestrated by the oncogene PVT1 (Figure 4C).

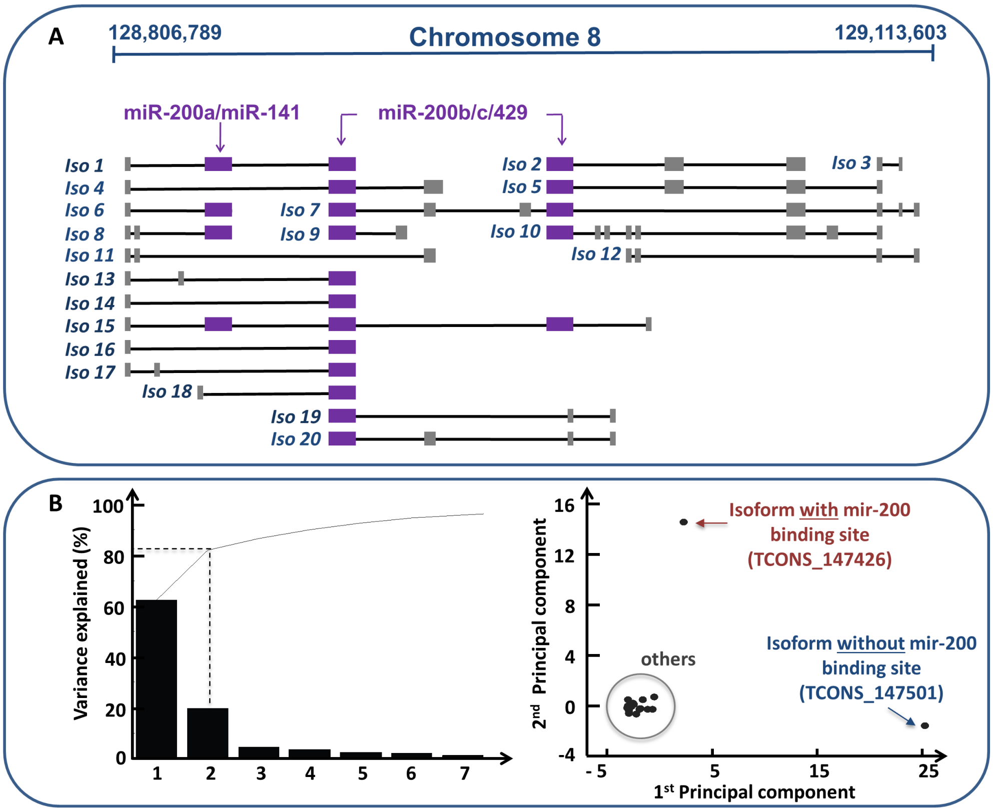

Figure 4. *Cont.*

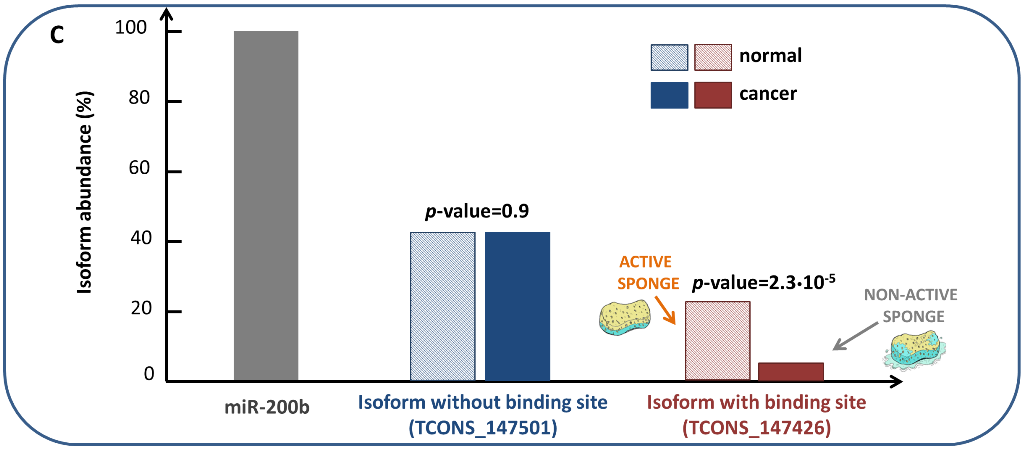

Figure 4. Analysis of PVT1 isoforms [27]. (**A**) Sketch of PVT1 genomic locus as reconstructed by Cufflinks spans across a genome interval of over 300 kb (i.e., 128,806,789-129,113,603 bases within the February 2009 human genome build GRCh37/hg19) on the forward strand of chromosome 8. PVT1 locus gives rise to 91 different variants according to raw RNA-seq data of TCGA (The Cancer Genome Atlas) for breast invasive carcinoma. The isoform names correspond to an increasing symbolic numbering and not to the actual nomenclature of the PVT1 variants. Lines represent introns and boxes (violet and grey) represent exons. Violet boxes correspond to the binding sites for the miR-200 family members. Note that some isoforms lack such binding sites (e.g., Iso11 and Iso12). (**B**) (Left) The percent variability explained by each principal component (PC) shown by the Pareto chart. This chart contains both bars and a line graph, where individual values are represented in descending order by bars, and the line represents the cumulative total value. The y-axis represents the percentage of the data variance explained by each PC, whereas the x-axis represents the principal components that are able to explain the first 100% of the cumulative distribution. PCA is performed using the variations of all the isoforms between normal and cancer tissues. (Right) The scatter plot of the projection of the original data (i.e., the variations of all the isoforms between normal and cancer tissues) onto the first two PCs; the x-axis contains the first PC while the y-axis contains the second PC. In this plot, it is possible to group isoforms in three classes: the isoform missing the binding site for the miR-200 family members (blue isoform, TCONS_147501), the isoform with the seed match for the miR-200b/200c/429 cluster (red isoform, TCONS_147426), and all the others. The first PC is able to separate the variation of the blue isoform from the others; the second PC is able to separate the variation of the red isoform from the others. (**C**) The ratio between the abundance of the red isoform (TCONS_147426, with the binding site for the miR-200b/200c/429 cluster) and blue isoform (TCONS_147501, without the binding site) with respect to the miR-200b in both normal (striped rectangle) and cancer tissues (full boxes). In the normal tissues only the isoform of PVT1 gene harboring the binding site for the miR-200b/200c/429 cluster acts as a sponge regulator of the miR-200 family members. In cancer tissues, it stops working as a sponge since its concentration is much lower than the concentration of the miR-200 family members.

4. SWItchMiner (SWIM): A Tool Exploiting the Topological Properties of Gene Co-Expression Networks

SWItchMiner (SWIM) [28] is a wizard-like software implementation of a network-based model. By combining topological properties of co-expression networks with genome-wide analysis, SWIM is able to identify a small pool of genes, called switch genes that are critically associated with drastic changes in cell phenotype.

4.1. SWIM Algorithm

The algorithm implemented by SWIM [28] is composed of the steps depicted in Figure 5.

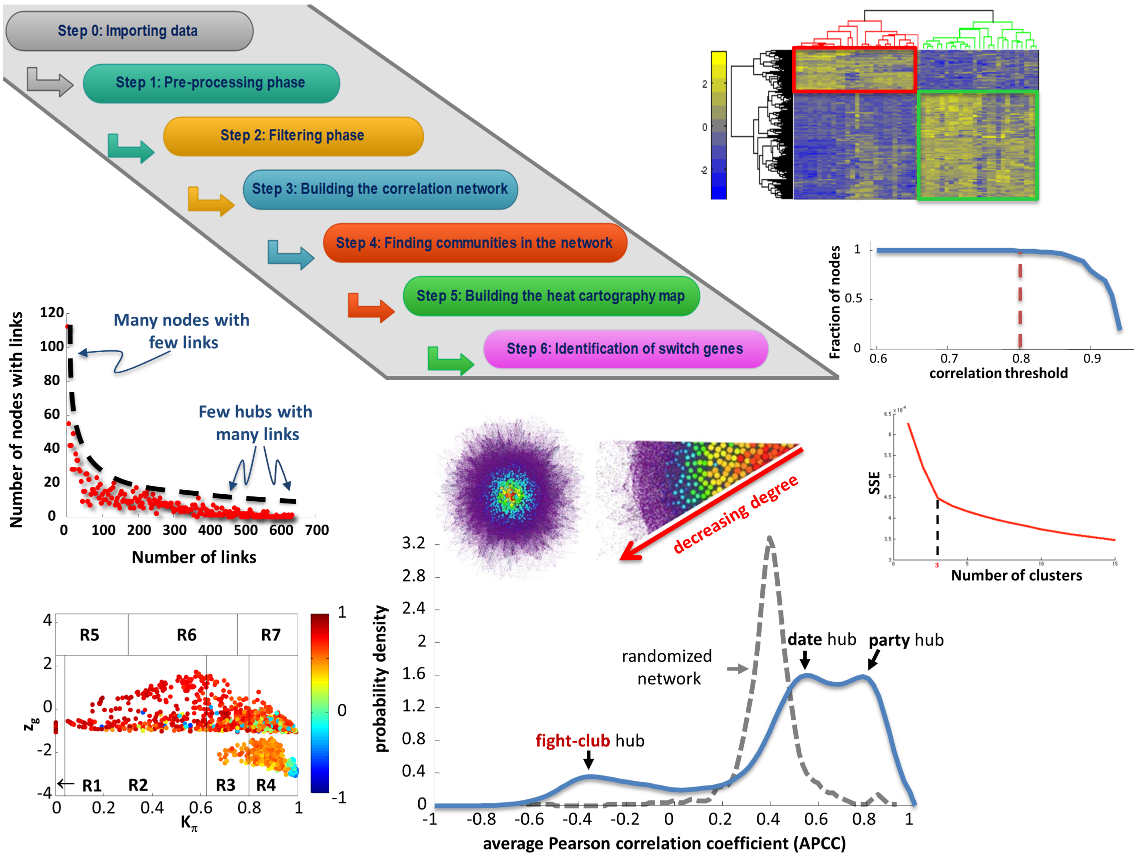
Step 0: Importing data
Step 1: Pre-processing phase
Step 2: Filtering phase
Step 3: Building the correlation network
Step 4: Finding communities in the network
Step 5: Building the heat cartography map
Step 6: Identification of switch genes
Fraction of nodes
correlation threshold
SSE
Number of clusters
Number of nodes with links
Number of links
Many nodes with few links
Few hubs with many links
decreasing degree
R5
R6
R7
R1
R2
R3
R4
z_g
K_π
probability density
randomized network
date hub
party hub
fight-club hub
average Pearson correlation coefficient (APCC)

Figure 5. SWIM flowchart. The figure depicts the steps performed by SWIM [28] and shows some examples of outputs obtained by running SWIM on the grapevine dataset [29].

4.1.1. Differential Gene Expression Analysis

Pre-processing phase. SWIM includes a pre-processing phase (Step 1 in Figure 5) for removing genes whose expression is mostly zero or change very little. This step requires the selection of two specific thresholds: the first threshold regards the maximum number of zeros allowed for the expression values of each gene across samples; the second one concerns the minimum variation—measured by the Inter Quartile Range (IQR) percentile—allowed for each gene across samples.

Filtering phase. SWIM implements a filtering phase (Step 2 in Figure 5) that allows to remove genes whose expression between two given conditions (A and B) does not change enough or does it without statistical significance. This step requires the selection of other two specific thresholds. Considering the logarithm of the ratio between the average expression of samples in condition A and the average expression of samples in condition B (log fold-change), the first threshold allows to remove the genes falling behind, in absolute value, a fixed cutoff on the log fold-change. The second threshold concerns the smallest probability (p-value) for which the data allow to reject the null hypothesis (i.e., the means of the two distributions are identical) of the Student's t-test. Actually, since this statistical test will be repeated multiple times (as much as the genes under testing), the obtained p-values must be adjusted. To correct multiple tests, SWIM makes use of False Discovery Rate (FDR) method [64] and thus the threshold refers to the FDR values. At the end of this phase, the differentially expressed genes between conditions A and B have been identified.

4.1.2. Network Analysis

Building the correlation network. SWIM builds a co-expression network of differentially expressed RNAs based on the Pearson correlation between the expression profiles of gene pairs (Step 3 in Figure 5). In this network, two nodes are connected if the absolute value of the Pearson correlation for their expression profiles is greater than a given threshold. The choice of this threshold should reflect a right balance between the number of edges and the number of connected components of the network: the number of edges should be as small as possible in order to have a manageable network (pointing towards a higher threshold) and the number of connected components should be as small as possible in order to preserve the integrity of the network (pointing towards a smaller threshold).

Finding communities in the network. To find communities in the network, SWIM makes use of the k-means algorithm [65], a method of cluster decomposition whose aim is to partition n objects (i.e., the nodes of the co-expression network) into N clusters (Step 4 in Figure 5). The quality of clustering was evaluated by minimizing the sum of the squared error (SSE), depending on the distance of each object to its closest centroid. A reasonable choice of the number of clusters is suggested by the position of an elbow in the SSE plot computed as a function of N. As distance measure, it is used $dist(x, y) = 1 - \rho(x, y)$, where $\rho(x, y)$ is the Pearson correlation between expression profiles of nodes x and y. The k-means algorithm, despite being the most widely used clustering algorithm, has some intrinsic limitations. Firstly, the number of clusters must be set in advance; secondly, it guarantees convergence only to a local minimum of SSE; thirdly, the initial position of the centroids is randomly chosen to cause a dependence of the partitioning on initialization. However, some reasonable assumptions can be done and are described in the following. There is no strict method to determine the "correct" number of clusters. Among others, SWIM uses an approach—named "Scree plot"—that evaluates the behavior of the SSE function to vary the number of clusters. Then, the position of an elbow in the scree plot—i.e., where the "cliff" reaches a bottom plateau—determines an appropriate number of clusters. Since finding the global optimum of SSE is theoretically NP-hard [66], it is commonly assumed that is sufficient to carry out a number of random initialization followed by a selection of the best separated solution, measured by the lowest SSE [67]. Moreover, the partition with the lowest SSE is commonly assumed to be reproducible under repeated initializations [67]. Thus, for a given number of clusters, SWIM allows repeating the clustering many times (replicates), each with a new set of initial cluster centroid positions. For each replicate, the k-means algorithm performs iterative partitioning (iterations) until the minimum of the SSE function is reached. Then, the cluster configuration with the lowest SSE values among all replicates will be chosen, for that number of clusters.

4.1.3. Role assignment to network nodes

Building the heat cartography map. Once the modular structure of the complex network has been found, roles have to be assigned to each node. This is done by dividing the plan according to two parameters, the clusterphobic coefficient K_π and the global within-module degree z_g. The clusterphobic coefficient K_π measures the "fear" of being confined in a cluster, in analogy with the claustrophobic disorder. A high value of K_π denotes nodes having much more external than internal links. The global within-module degree z_g measures how "well-connected" each node is to other nodes in its own community. In the following, the formal definitions of these parameters for a generic node i [28]:

$$K_\pi^i = 1 - \left(\frac{k_i^{in}}{k_i}\right)^2 \tag{5}$$

$$z_g^i = \frac{k_i^{in} - \bar{k}_{C_i}}{\sigma_{C_i}} \tag{6}$$

where k_i^{in} is the number of links of node i to nodes in its module C_i, k_i is the total degree of node i, $\bar{k}_{C_i}$ and σ_{C_i} are the average and standard deviation of the total degree distribution of the nodes in the module C_i. This definition of z_g quantifies how much a node is a hub (i.e., degree exceeding 5 [68]) in its community and thus represents a measure of local connectivity. On the contrary, the parameter K_π evaluating the ratio of internal to external connections of a node represents a measure of global connectivity. Note that $K_\pi = 0$ when a node has only links within its module, i.e., it does not communicate with the other modules ($k_i^{in} = k_i$); while K_π is close to 1 when the majority of its links are external to its own module. According to the global within-module degree z_g and the clusterphobic coefficient K_π values, the plane is divided into seven regions (R1–R7), each defines a specific node role [69]. High z_g values correspond to nodes that are hubs within their module (local hubs), while high values of K_π identify nodes that interact mainly outside their community. Then, SWIM colors nodes in the cartography according to the average Pearson correlation coefficient (APCC) between the expression profiles of each node and its nearest neighbors [68]. This representation of the network is defined as “heat cartography map” (Step 5 in Figure 5). By computing the APCC of expression over all interaction partners of each hub in PPI networks in yeast, the authors in [68] concluded that hubs fall into two distinct categories: date hubs that display low co-expression with their partners (low APCC) and party hubs that have high co-expression (high APCC). In the gene expression networks, the distribution of APCCs appears to be trimodal [28,29] where, similar to PPI networks, two peaks represent low (date hubs) and high (party hubs) positive APCC values, but with the addition of a new third peak which is characteristic of gene expression networks and represents negative APCC values. Nodes populating this peak are called “fight-club hubs” [29].

Identification of switch genes. Looking at the heat cartography map, SWIM identifies the so-called “switch genes” (Step 6 in Figure 5): the subset of the fight-club hubs that mainly interact outside their community (region R4). In particular, they satisfy the following topological and expression features:

i. Being not a hub in their own cluster ($z_g < 2.5$);
ii. Having many links outside their own cluster ($K_\pi > 0.8$);
iii. Having a negative average weight of their incident links (APCC < 0).

At the end of Step 6, SWIM gives the opportunity to perform further analyses regarding the evaluation of network robustness, which is the resilience to errors, by studying the effect on the network connectivity of removing nodes by decreasing degree. In particular, SWIM evaluates the effect on the average shortest path (the shortest path between two nodes is the minimum number of edges connecting them and the average shortest path of a network is the average of the shortest paths for all possible pairs of network nodes) of removing randomly chosen nodes, switch genes, fight-club hubs, date and party hubs. Since scale-free networks have few hubs and many non-hub nodes, they are amazingly resistant to a random removal of nodes, while the removal of hubs causes an effect known as “vulnerability to attack” to allude to the fact that the integrity of the network is destroyed.

4.2. SWIM Applications

SWIM was amenable to detect switch genes in different organisms and cell conditions, leading to the identification of key players in biologically relevant scenarios, including but not limited to human cancer [28–30].

4.2.1. Grapevine Analysis

SWIM was successfully applied in plants [29] for studying the transition between mature to immature phase of the developmental program of of *Vitis Vinifera*. In this study, switch genes resulted to be master regulators of the transcriptome remodeling that marks the developmental shift of grapevine from immature to mature growth. Specifically, the authors found about one hundred switch genes in grapevine that appears to be anti-correlated with about 1000 genes (more than 50% of the entire co-expression network). All switch genes, expressed at low levels in vegetative/green tissues, showed a significant increase in mature/woody organs, suggesting a potential regulatory role in the immature-mature transition. Among switch genes, they found many transcription factors like NAC-domain genes and many targets of tissue-specific miRNAs, such as the NAC33 switch gene and miRNA-164 whose interaction was experimentally validated. The authors propose a transcriptional regulatory network in which tissue-specific and stage-specific miRNAs regulate the expression of several switch genes (Figure 6A). Switch genes appeared to be over-expressed in mature organs and they are presumably regulated by tissue specific microRNAs with an opposite trend. They resulted also anti-correlated with thousands of genes that were down-regulated in mature tissue, meaning that the transition to mature growth in grapevine was mainly due to the suppression of vegetative pathways such as photosynthesis and cell proliferation rather than the activation of maturation-specific pathways. The immature fruits grow and make photosynthesis, then stop growing and start ripening and activate the secondary metabolism like color, aroma and flavor of the fruit. The results of this study are very important for the winemaking industries because the identification of which genes are responsible for the ripening can help to face climate changes and thus improve the quality of wine.

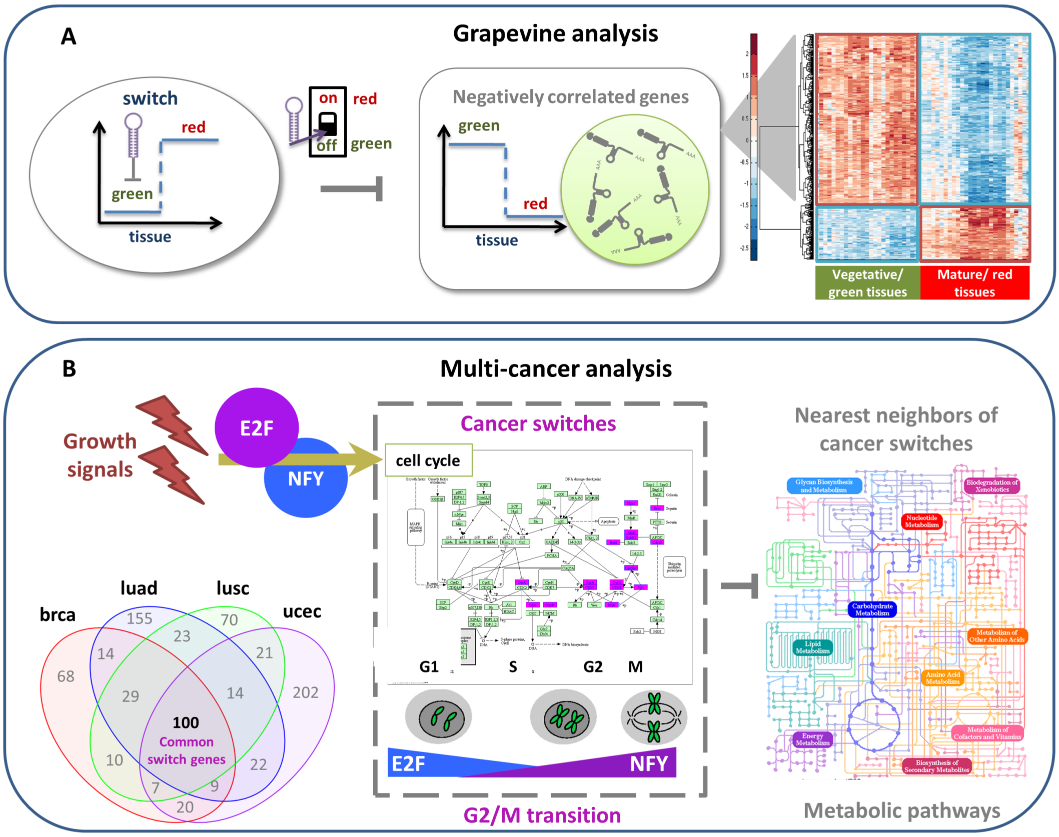

Figure 6. *Cont.*

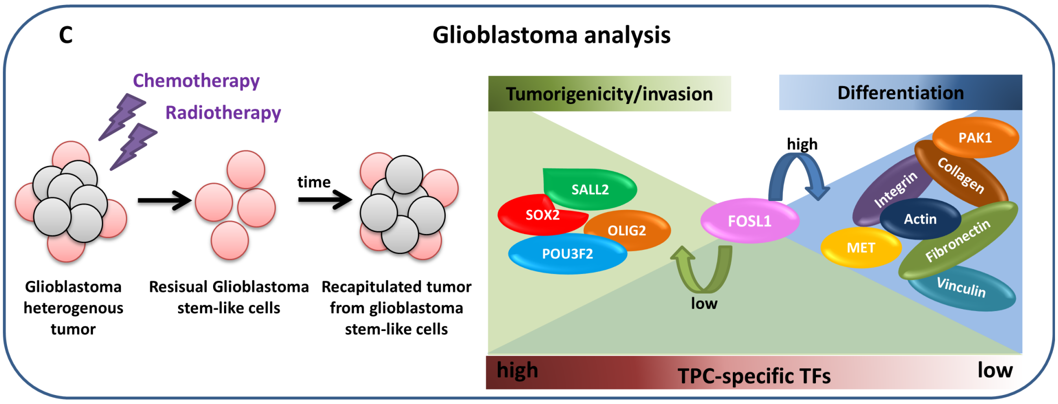

Figure 6. SWIM applications in (**A**) grapevine analysis [29], (**B**) multi-cancers analysis [28], and (**C**) glioblastoma analysis [30]. (**A**) A sketch of the switch gene regulation mechanism in grapevine. During the vegetative/green phase of organ development, switch genes are repressed by miRNAs and vegetative genes are expressed. In the transition to the mature/red phase, these miRNAs are deactivated, the switch genes are expressed and their anti-correlated vegetative genes are turned off. The heat map shows the transcription level of positively and negatively correlated genes with a typical switch, where expression values increase from blue to red. (**B**) A sketch of the switch gene regulation mechanism in human cancers. SWIM extracted a set of 100 cancer-recurrent switch genes across four tumors—breast invasive carcinoma (brca), lung squamous cell carcinoma (lusc), lung adenocarcinoma (luad), uterine corpus endometrial carcinoma (ucec)—that showed a marked functional enrichment in cell cycle and specifically on the regulation of the G2-to-M transition. The promoter motif analysis suggested that two major transcription factors (namely E2F and NFY) lead to the activation of the switch gene layer of gene regulation. Activation of switch genes in these cancers seems to predominantly repress several metabolic pathways, possibly leading to the well-known metabolic rewiring characterizing cancer cells. (**C**) A sketch of the switch gene regulation mechanism in human glioblastoma. Glioblastoma subpopulation of self-renewing, stem-like cells has been shown to be responsible for tumor initiation, progression, resistance to treatment, and relapse. Among switch genes identified by SWIM involved in the transition from a stem-like to a differentiated phenotype of glioblastoma cells, FOSL1 stands out as a promising candidate to trigger the differentiation. On one hand, it has been found positively correlated with genes encoding proteins linked to the focal adhesion complex and extracellular matrix (ECM) receptor interaction (e.g., integrins, collagen, and signaling proteins). Conversely, it is negatively regulated with well-known neurodevelopmental transcription factors (TFs) specific of stem-like identity, including the core set of OLIG2, POU3F2, SALL2, SOX2 [70]. Thus, it could be considered as putative controller of stem-like cell differentiation process by repressing the core set of neurodevelopmental TFs and by modulating the equilibrium between cell adhesion and migration.

4.2.2. Multi-Cancer Analysis

SWIM was applied to a large panel of cancer datasets obtained from TCGA [41,42] to identify switch genes that could be critically associated with the drastic changes in the physiological state of tissues induced by the cancer development [28]. In this study, the authors found disease-specific switch genes, as well as common switch genes that were shared among different tumors. The list of common switch genes encompassed 100 RNA transcripts that appeared all up-regulated in cancer and enriched in cell cycle, in particular at the transition between the G2 phase and M phase. Moreover, their promoter regions appeared enriched in three known regulatory motifs: the cell cycle gene homology (CHR) element, the nuclear transcription factor Y (NFY) binding motif the E2F transcription factor (E2F) binding motif, that are known to participate in the regulation of progression through the cell cycle. Finally, the analysis of the functional annotation of their negative nearest neighbor highlighted a general association of crucial nodes of metabolic process. The authors propose a regulatory network to describe the switch gene mechanism in human cancers (Figure 6B). In summary, the list of 100 common switch genes appear to be over-expressed in cancer tissues, suggesting a their potential role in the malignant transformation; they are involved in cell cycle, at the G2/M transition; they are anti-correlated to some metabolic pathways that in turn appear to be switched-off in cancer and they are activated by growth signals. The activation of switch genes by growth transcription factors could accelerate the late phase of the cell cycle with a consequent increase of cancer progression and promote the rewiring of some metabolic pathways, hallmarks of the malignant transformation [71].

The list of disease-specific switch genes identified by SWIM encompassed protein coding genes, long non-coding, and miRNAs, recovering many known key cancer players, but also many new potential biomarkers not yet characterized in cancer context. Motivated by the growing interest in lncRNAs, which appears to hold strong promise as novel biomarkers and therapeutic targets for cancer [72,73], and by the widespread recognized role of miRNAs as key negative regulators in many intracellular processes as well as in carcinogenesis [74], in the following we discuss miRNAs, lncRNAs, and mRNAs acting as switch genes in the multi-cancer analysis across TCGA cancer datasets, separately.

miRNA-diseasome network. We built a diseasome bipartite network consisting of two disjoint sets of nodes: one set corresponds to the human cancer types under study (diseases); the other set corresponds to all miRNAs acting as switch genes in each disease (Figure 7A). A disease and a miRNA are then connected by a link if that miRNA acts as switch gene in that disease. This representation allows to highlight which are miRNAs involved in multiple diseases (i.e., grey nodes in Figure 7A) and which are disease-specific miRNAs for each tumor (i.e., nodes that are colored according to each tumor type in Figure 7A). We called this network "miRNA-diseasome", in analogy with the human diseasome network of [2], where the two disjoints set of nodes corresponds to disorders and disease genes. In this diseasome bipartite network a link is placed between a disorders and a disease gene if mutations in that gene lead to the specific disorder. Then, starting from the miRNA-diseasome bipartite network, we generated a biologically relevant network projection that we called "miRNA-disease network" (MDN) (Figure 7B). In the MDN nodes represent diseases, and two diseases are connected to each other if they share at least one miRNA acting as switch gene in both diseases. This representation provides a disease-centered view of the miRNA-diseasome and allows us to highlight tumors with the highest number of miRNAs acting as switch genes (i.e., blca that results as a major hub in the network), as well as tumors with the highest number of shared miRNAs (i.e., blca and ucec).

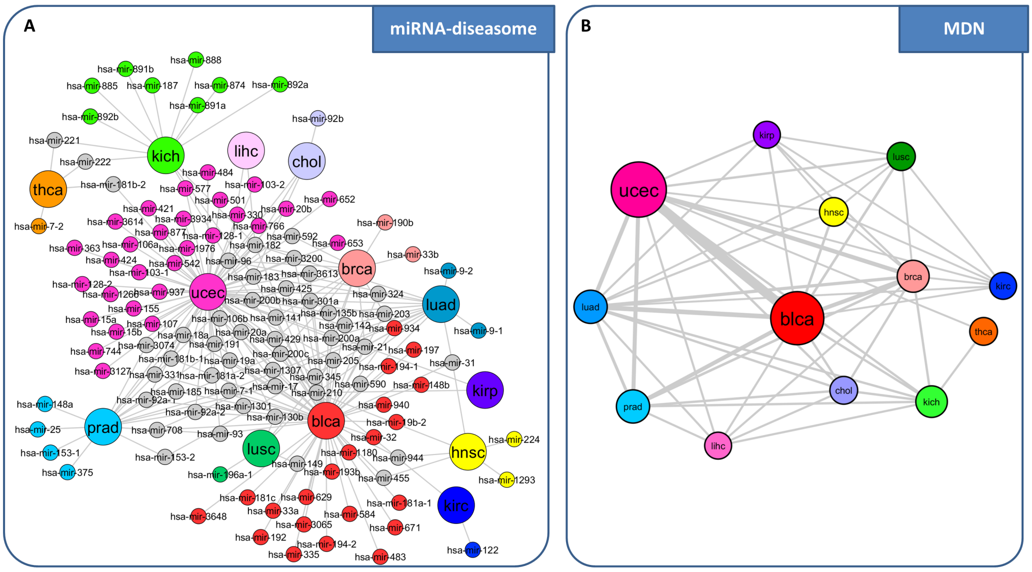

Figure 7. Comparative analysis of miRNAs acting as switch genes in the large panel of TCGA

cancer datasets. (**A**) miRNA-diseasome. The bipartite network is composed of two disjoint sets of nodes with different size: the larger ones correspond to the analyzed human cancer types from TCGA, whereas the smaller ones correspond to all miRNAs acting as switch genes. A link occurs between a tumor type and a miRNA if the miRNA acts as switch gene for that tumor. Different colors are associated to different tumor types. miRNAs are colored based on the tumor type to which they belong. Nodes are light gray if the corresponding miRNAs are associated with more than one tumor type. (**B**) miRNA-disease network (MDN). The MDN is the projection of the miRNA-diseasome bipartite network, in which nodes correspond to tumor types (diseases) and two diseases are connected if there is at least one miRNA that acts as switch gene in both. The width of a link is proportional to the number of miRNAs that are acting as switch genes in both diseases. The size of a node is proportional to the number of microRNAs acting as switch genes for that disease. Different node colors are associated with different diseases. blca: bladder urothelial carcinoma, chol: cholangiocarcinoma, hnsc: head and neck squamous cell carcinoma, kich: kidney chromophobe, kirc: kidney renal clear cell carcinoma, kirp: kidney renal papillary cell carcinoma, lihc: liver hepatocellular carcinoma, prad: prostate adenocarcinoma, thca: thyroid carcinoma.

lncRNA-diseasome network. We performed the same analysis as for miRNAs in order to build a lncRNA-diseasome bipartite network, where one set of nodes corresponds to the human cancer types analyzed by SWIM in [28] and the other set corresponds to all lncRNAs acting as switch genes in each disease (Figure 8A). A disease and a lncRNA are then connected by a link if that lncRNA acts as switch gene in that disease. Then, starting from the lncRNA-diseasome bipartite network, we generated the "lncRNA-disease network" (LDN) (Figure 8B), where nodes represent diseases, and two diseases are connected to each other if they share at least one lncRNA acting as switch genes in both diseases.

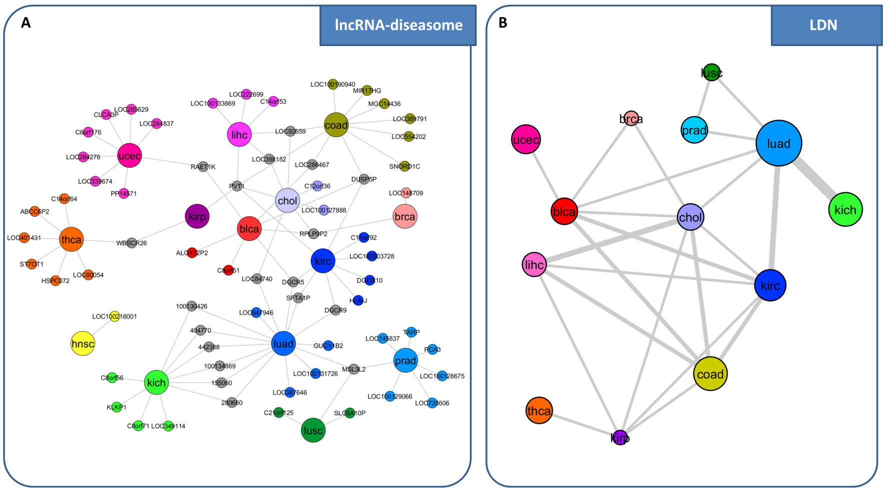

Figure 8. Comparative analysis of lncRNAs acting as switch genes in the large panel of TCGA cancer datasets. (**A**) lncRNA-diseasome. The bipartite network is composed of two disjoint sets of nodes with different size: the larger ones correspond to the analyzed human cancer types from TCGA, whereas the smaller ones correspond to all lncRNAs acting as switch genes. A link occurs between a tumor type and a lncRNA if the lncRNA acts as switch gene for that tumor. Different colors are associated to different tumor types. lncRNAs are colored based on the tumor type to which they belong. Nodes are light gray if the corresponding lncRNAs are associated with more than one tumor type. (**B**) lncRNA-disease network (LDN). The LDN is the projection of the lncRNA-diseasome bipartite network, in which nodes correspond to tumor types (diseases) and two diseases are connected if there is at least one lncRNA that acts as switch gene in both. The width of a link is proportional to the number of lncRNAs that are acting as switch genes in both diseases. The size of a node is proportional to the number of lncRNAs acting as switch genes for that disease. Different node colors are associated with different diseases.

mRNA-diseasome network. Finally, we built a mRNA-diseasome bipartite network, where one set of nodes corresponds to the human cancer types analyzed by SWIM in [28] and the other set corresponds to all mRNAs acting as switch genes in each disease (Figure 9A). Then, as before, starting from the mRNA-diseasome bipartite network, we generated the "mRNA-disease network" (MRDN) (Figure 9B), where nodes represent diseases, and two diseases are connected to each other if they share at least one mRNA acting as switch genes in both diseases.

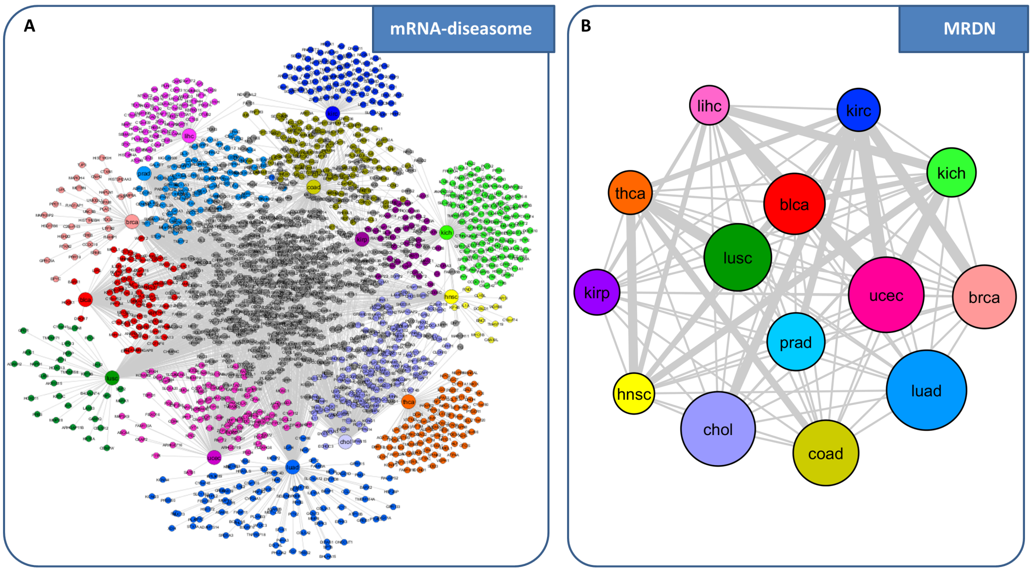

Figure 9. Comparative analysis of protein-coding genes (mRNAs) acting as switch genes in the large panel of TCGA cancer datasets. (**A**) mRNA-diseasome. The bipartite network is composed of two disjoint sets of nodes with different size: the larger ones correspond to the analyzed human cancer types from TCGA, whereas the smaller ones correspond to all protein coding genes acting as switch genes. A link occurs between a tumor type and a mRNA if the mRNA acts as switch gene for that tumor. Different colors are associated to different tumor types. mRNAs are colored based on the tumor type to which they belong. Nodes are light gray if the corresponding miRNAs are associated with more than one tumor type. (**B**) mRNA-disease network (MRDN). The MRDN is the projection of the mRNA-diseasome bipartite network, in which nodes correspond to tumor types (diseases) and two diseases are connected if there is at least one mRNA that acts as switch gene in both. The width of a link is proportional to the number of mRNAs that are acting as switch genes in both diseases. The size of a node is proportional to the number of mRNAs acting as switch genes for that disease. Different node colors are associated with different diseases.

4.2.3. Glioblastoma analysis

Glioblastoma multiforme (GBM) is the most frequently diagnosed and aggressive brain tumor with the 5-years survival rate achieved for only 5% of patients. Several studies identified a subpopulation of GBM cells with radio/chemotherapy-resistant properties that have a role in driving tumor initiation, progression, resistance to treatment, and relapse [75]. Due to their abilities of self-renewal, proliferation, and differentiation into multiple lineages, these cells are named cancer stem-like cells and are held responsible for carcinogenesis (Figure 6C). The identification of genes responsible of the stem-like phenotype are going to dominate cancer research scene as effective and long-lasting therapeutic strategy. In this context, a recent study [70] identified a 4-core of neurodevelopmental TFs (transcription factors) (i.e., OLIG2, POU3F2, SALL2, SOX2), which are selectively expressed in glioblastoma stem-like cells and have been shown to be sufficient to fully reprogram differentiated cells into glioblastoma stem-like cells. In order to computationally identify genes controlling cancer stem-like cells differentiation and invasion, SWIM was applied to gene expression profiles from two independent GBM datasets [30], publicly available on the Gene Expression

Omnibus (GEO) repository: RNA-seq data obtained from stem-like tumor-propagating cells and differentiated glioblastoma cells (i.e., GSE54792 [70]); Affymetrix HG-U133 Plus 2.0 microrarrays expression data from glioblastoma stem-like cell lines, the corresponding primary tumors, and conventional glioma cell lines (i.e., GSE23806 [76])).

SWIM identified the FOS like transcription factor FOSL1 as the most promising switch gene (Figure 6C) shared from both datasets [30]. Indeed, FOSL1 fulfills very interesting features that made it eligible as new potential therapeutic target: it is down-regulated in stem-like cells; it is highly negatively correlated with the 4-core TFs (OLIG2, POU3F2, SALL2, SOX2); the promoter regions of the 4-core TFs were found to harbor a consensus binding motif for FOSL1; it was found to act as repressor transcription factor [77]; it is positively correlated with genes encoding proteins crucial for cell-matrix adhesion and cell motility (e.g., actin, collagen, fribonectin, and several integrins), which can influence cell adhesion dynamics and migration, and thus the cancer invasiveness.

Taken together these considerations prompted the authors to bet on FOSL1, which could promote the differentiation process of GBM stem-like cells by repressing the 4-core TFs and consequently halted cancer growth and invasion. This should allow for anticipation of care as well as the reduction of the social impact of diseases and the restraint of health costs.

4.3. SWIM Switch Genes towards DIAMOnD Disease Genes

Next we explore the performance of SWIM on human breast invasive carcinoma. Since the full set of disease proteins is unknown, we cannot assess the performance directly in terms of true positives/negatives. We therefore compare the results obtained by SWIM with the known disease-associated genes and with the DIAMOnD disease genes. SWIM was applied to breast invasive carcinoma expression data from high-throughput RNA-seq downloaded from TCGA data portal. Data correspond to normalized level three data from RNASeq Version 2 created by using MapSplice [78] to do the alignment and RSEM [79] to perform the quantification and normalization. The study concerned 103 samples for which the complete sets of tumor and matched normal profiles were available. By running SWIM, we obtained 257 switch genes and we selected for this comparison only the 195 switch genes that are coding RNAs. The known disease-associated genes for breast neoplasms disease (in total $s_0 = 40$ seed proteins) were provided by DIAMOnD paper [10] that integrated data from OMIM (Online Mendelian Inheritance in Man) [80] and GWAS (Genome-Wide Association Studies). Finally, the DIAMOnD disease genes were retrieved by running DIAMOnD algorithm for breast neoplasms disease and retaining the first 500 new disease genes. In total, we considered 540 seed genes, including the 40 breast neoplasm associated genes and the first 500 DIAMOnD disease genes.

We compared the performance of SWIM switch genes to seed genes as well as to random expectation for the same number of genes drawn randomly from the network. The performance is based on the number of switch genes that are considered true positives. To quantify the statistical significance of a given number of true positives at a given iteration step i, we used a sliding window approach: at each iteration step i the same number of seed genes was considered. We used genes in the interval $[s_0 + (i - 1)]$ until the sliding window spans across the entire set of 540 seed genes and count the number of true positives among switch genes. The statistical significance of an observed number is then determined using the hypergeometric distribution. For breast neoplasms disease, we found that switch genes are significantly enriched (p-value <0.05) in seed genes until 175 iterations (Figure 10A), significantly higher than random expectation (Figure 10B). This results is extremely encouraging since the authors of DIAMOnD claim in the paper [10] that the first ∼200 DIAMOnD genes are found to participate in important seed pathways at a rate similar to the one within the seed proteins themselves.

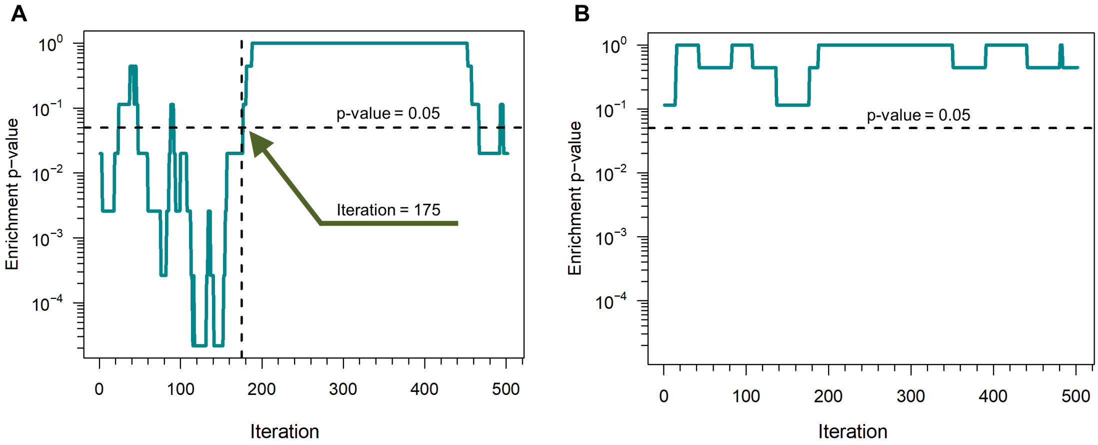

Figure 10. Performance of SWIM in human breast invasive carcinoma. (**A**) Enrichment *p*-values of switch genes in seed genes obtained by running SWIM for breast invasive carcinoma. (**B**) Enrichment *p*-values for a subset of genes drawn randomly from the network in seed genes.

5. Conclusions

In the last decade, the great advances in high-throughput technologies have led to massive amounts of genomic, transcriptomic, proteomic and metabolomic data capable to provide new opportunities for identifying potential biomarkers and developing effective treatments for human diseases. The availability of this huge amount of data has revolutionized biomedical science and, in particular, cancer genomics, but only the amount is not enough. If in the past there was a difficulty in collecting genetic data, today the challenge is to give them meaning and it is therefore essential to use effective informatics solutions capable of managing, analyzing and integrating these biological "Big Data". The need to take on this new challenge has paved the way for a paradigm shift towards the development of temporal and spatial multi-level models, from molecular machineries to single cells, whole organism and individuals, including the environment, to reveal the underlying links between components. This new type of medical paradigm is called "Network Medicine". Rather than trying to understand pathogenesis into a reductionist framework, network medicine entwines the many facets of disease in many different types of networks: from the physical interactions acting in a cell to the information flow through biological components. The representation of complex systems as networks is of paramount importance for visualizing the interactome underlying structure, revealing new functional roles, and proposing new and fresh interpretations of data. Networks can be obtained from any sort of information: known protein–protein interactions, gene expression profiles, functional annotation, etc. In this review, we focused on three different classes of approaches that use different types of interaction networks to infer novel cancer genes: methods using PPI networks, methods using regulatory networks and methods using co-expression networks.

One important bias in the methods that predict cancer genes is the direct or indirect incorporation of prior knowledge. Methods using PPI networks suffer more from this problem compared with methods using co-expression and/or regulatory networks. In fact, many proteins or genes have been extensively studied and hence have a higher number of connections in the protein networks. Moreover, since network biology is still far from completing the human interactome, PPI networks are suffering from false negatives (i.e., missing interactions) and false positives (i.e., false interactions) [1]. In the future, much computational effort is needed to complete the human interactome and increase its confidence.

Finally, network-based approaches using PPI networks are restricted to mutations that affect protein coding regions of the genome, thus they cannot be used to predict novel cancer genes among ncRNAs. Although the other two classes of approaches discussed in this review do not suffer of this bias, they lack the capability to integrate different data collections. In the future, we expect that the development of more accurate computational tools will be able to overcome this limitation.

This review provides a limited view of the landscape of the existing approaches that using network theory to deal with issues related to medicine, but it has the potential to stimulate the growth of new methods as well as the improvement of the existing ones. This will fuel advances in network medicine supporting the planning of disease prevention and treatment.

Author Contributions: P.P. conceived and designed the research; G.F. analyzed the data described in this review; G.F. made figures; F.C. performed the bioinformatics analysis described in this review; P.P. and L.F. developed the computational models described in this review; all authors wrote the paper.

Funding: This research was funded by SysBioNet, Italian Roadmap Research Infrastructures 2012.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barabási, A.L.; Gulbahce, N.; Loscalzo, J. Network Medicine: A Network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56. [CrossRef] [PubMed]
2. Goh, K.I.; Cusick, M.E.; Valle, D.; Childs, B.; Vidal, M.; Barabási, A.L. The human disease network. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8685–8690. [CrossRef] [PubMed]
3. Rual, J.F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G.F.; Gibbons, F.D.; Dreze, M.; Ayivi-Guedehoussou, N.; et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **2005**, *437*, 1173–1178. [CrossRef] [PubMed]
4. Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F.H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; et al. A human protein–protein interaction network: A resource for annotating the proteome. *Cell* **2005**, *122*, 957–968. [CrossRef] [PubMed]
5. Carninci, P.; Kasukawa, T.; Katayama, S.; Gough, J.; Frith, M.C.; Maeda, N.; Oyama, R.; Ravasi, T.; Lenhard, B.; Wells, C.; et al. The transcriptional landscape of the mammalian genome. *Science* **2005**, *309*, 1559–1563. [CrossRef] [PubMed]
6. Stuart, J.M.; Segal, E.; Koller, D.; Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **2003**, *302*, 249–255. [CrossRef] [PubMed]
7. Oti, M.; Snel, B.; Huynen, M.A.; Brunner, H.G. Predicting disease genes using protein–protein interactions. *Am. J. Med. Genet.* **2006**, *43*, 691–698. [CrossRef] [PubMed]
8. Yin, T.; Chen, S.; Wu, X.; Tian, W. GenePANDA—A novel network-based gene prioritizing tool for complex diseases. *Sci. Rep.* **2017**, *7*, 43258. [CrossRef] [PubMed]
9. Erten, S.; Bebek, G.; Ewing, R.M.; Koyutürk, M. DADA: Degree-aware algorithms for network-based disease gene prioritization. *BioData Min.* **2011**, *4*, 19. [CrossRef] [PubMed]
10. Ghiassian, S.D.; Menche, J.; Barabási, A.L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* **2015**, *11*, e1004120. [CrossRef] [PubMed]
11. Vanunu, O.; Magger, O.; Ruppin, E.; Shlomi, T.; Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **2010**, *6*, e1000641. [CrossRef] [PubMed]
12. Mordelet, F.; Vert, J.P. ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinform.* **2011**, *12*, 389. [CrossRef] [PubMed]
13. Esteller, M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* **2011**, *12*, 861–874. [CrossRef] [PubMed]
14. Mattick, J.S. The central role of RNA in human development and cognition. *FEBS Lett.* **2011**, *585*, 1600–1616. [CrossRef] [PubMed]
15. Birney, E.; Stamatoyannopoulos, J.A.; Dutta, A.; Guigó, R.; Gingeras, T.R.; Margulies, E.H.; Weng, Z.; Snyder, M.; Dermitzakis, E.T.; Stamatoyannopoulos, J.A.; et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007**, *447*, 799–816. [CrossRef] [PubMed]
16. Knowling, S.; Morris, K.V. Non-coding RNA and antisense RNA. Nature's trash or treasure? *Biochimie* **2011**, *93*, 1922–1927. [CrossRef] [PubMed]

17. Mercer, T.R.; Dinger, M.E.; Mattick, J.S. Long non-coding RNAs: Insights into functions. *Nat. Rev. Genet.* **2009**, *10*, 155–159. [CrossRef] [PubMed]
18. Ponting, C.P.; Oliver, P.L.; Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **2009**, *136*, 629–641. [CrossRef] [PubMed]
19. Chang, H.Y. Genome Regulation by Long Non-Coding RNAs. *Blood* **2013**, *122*, SCI–29.
20. Franco-Zorrilla, J.M.; Valli, A.; Todesco, M.; Mateos, I.; Puga, M.I.; Rubio-Somoza, I.; Leyva, A.; Weigel, D.; García, J.A.; Paz-Ares, J. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* **2007**, *39*, 1033. [CrossRef] [PubMed]
21. Ebert, M.S.; Neilson, J.R.; Sharp, P.A. MicroRNA sponges: Competitive inhibitors of small RNAs in mammalian cells. *Nat. Methods* **2007**, *4*, 721–726. [CrossRef] [PubMed]
22. Poliseno, L.; Salmena, L.; Zhang, J.; Carver, B.; Haveman, W.J.; Pandolfi, P.P. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **2010**, *465*, 1033–1038. [CrossRef] [PubMed]
23. Hansen, T.B.; Jensen, T.I.; Clausen, B.H.; Bramsen, J.B.; Finsen, B.; Damgaard, C.K.; Kjems, J. Natural RNA circles function as efficient microRNA sponges. *Nature* **2013**, *495*, 384. [CrossRef] [PubMed]
24. Memczak, S.; Jens, M.; Elefsinioti, A.; Torti, F.; Krueger, J.; Rybak, A.; Maier, L.; Mackowiak, S.D.; Gregersen, L.H.; Munschauer, M.; et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **2013**, *495*, 333. [CrossRef] [PubMed]
25. Paci, P.; Colombo, T.; Farina, L. Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Syst. Biol.* **2014**, *8*, 83. [CrossRef] [PubMed]
26. Le, T.D.; Zhang, J.; Liu, L.; Li, J. Computational methods for identifying miRNA sponge interactions. *Brief. Bioinform.* **2016**, bbw042. [CrossRef] [PubMed]
27. Conte, F.; Fiscon, G.; Chiara, M.; Colombo, T.; Farina, L.; Paci, P. Role of the long non-coding RNA PVT1 in the dysregulation of the ceRNA-ceRNA network in human breast cancer. *PLoS ONE* **2017**, *12*, e0171661. [CrossRef] [PubMed]
28. Paci, P.; Colombo, T.; Fiscon, G.; Gurtner, A.; Pavesi, G.; Farina, L. SWIM: A computational tool to unveiling crucial nodes in complex biological networks. *Sci. Rep.* **2017**, *7*, 44797. [CrossRef] [PubMed]
29. Palumbo, M.C.; Zenoni, S.; Fasoli, M.; Massonnet, M.; Farina, L.; Castiglione, F.; Pezzotti, M.; Paci, P. Integrated network analysis identifies fight-club nodes as a class of hubs encompassing key putative switch genes that induce major transcriptome reprogramming during grapevine development. *Plant Cell* **2014**, *26*, 4617–4635. [CrossRef] [PubMed]
30. Fiscon, G.; Conte, F.; Licursi, V.; Nasi, S.; Paci, P. Computational identification of specific genes for glioblastoma stem-like cells identity. *Sci. Rep.* **2018**, *8*, 7769. [CrossRef] [PubMed]
31. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef] [PubMed]
32. Köhler, S.; Bauer, S.; Horn, D.; Robinson, P.N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **2008**, *82*, 949–958. [CrossRef] [PubMed]
33. Gottlieb, A.; Magger, O.; Berman, I.; Ruppin, E.; Sharan, R. PRINCIPLE: A tool for associating genes with diseases via network propagation. *Bioinformatics* **2011**, *27*, 3325–3326. [CrossRef] [PubMed]
34. Glass, K.; Huttenhower, C.; Quackenbush, J.; Yuan, G.C. Passing messages between biological networks to refine predicted interactions. *PLoS ONE* **2013**, *8*, e64832. [CrossRef] [PubMed]
35. Sonawane, A.R.; Platig, J.; Fagny, M.; Chen, C.Y.; Paulson, J.N.; Lopes-Ramos, C.M.; DeMeo, D.L.; Quackenbush, J.; Glass, K.; Kuijjer, M.L. Understanding tissue-specific gene regulation. *Cell Rep.* **2017**, *21*, 1077–1088. [CrossRef] [PubMed]
36. Poliseno, L.; Pandolfi, P. PTEN ceRNA networks in human cancer. *Methods* **2015**, *77*, 41–50, doi:10.1016/j.ymeth.2015.01.013. [CrossRef] [PubMed]
37. Ergun, S.; Oztuzcu, S. Oncocers: ceRNA-mediated cross-talk by sponging miRNAs in oncogenic pathways. *Tumor Biol.* **2015**, *36*, 3129–3136. [CrossRef] [PubMed]
38. Qi, X.; Zhang, D.H.; Wu, N.; Xiao, J.H.; Wang, X.; Ma, W. ceRNA in cancer: Possible functions and clinical implications. *Am. J. Med. Genet.* **2015**, *52*, 710–718. [CrossRef] [PubMed]

39. Yang, C.; Wu, D.; Gao, L.; Liu, X.; Jin, Y.; Wang, D.; Wang, T.; Li, X. Competing endogenous RNA networks in human cancer: Hypothesis, validation, and perspectives. *Oncotarget* **2016**, *7*, 13479–13490. [CrossRef] [PubMed]
40. Salmena, L.; Poliseno, L.; Tay, Y.; Kats, L.; Pandolfi, P.P. A ceRNA hypothesis: The Rosetta Stone of a hidden RNA language? *Cell* **2011**, *146*, 353–358. [CrossRef] [PubMed]
41. Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
42. Tomczak, K.; Czerwinska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68–A77. [CrossRef] [PubMed]
43. Lewis, B.P.; Burge, C.B.; Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **2005**, *120*, 15–20. [CrossRef] [PubMed]
44. Tseng, Y.Y.; Moriarity, B.S.; Gong, W.; Akiyama, R.; Tiwari, A.; Kawakami, H.; Ronning, P.; Reuland, B.; Guenther, K.; Beadnell, T.C.; et al. PVT1 dependence in cancer with MYC copy-number increase. *Nature* **2014**, *512*, 82. [CrossRef] [PubMed]
45. Iden, M.; Fye, S.; Li, K.; Chowdhury, T.; Ramchandran, R.; Rader, J. The lncRNA PVT1 contributes to the cervical cancer phenotype and associates with poor patient prognosis. *PLoS ONE* **2016**, 11, e0156274. [CrossRef] [PubMed]
46. Colombo, T.; Farina, L.; Macino, G.; Paci, P. PVT1: A rising star among oncogenic long noncoding RNAs. *Biomed Res. Int.* **2015**, *2015*, 304208. [CrossRef] [PubMed]
47. Huppi, K.; Siwarski, D.; Skurla, R.; Klinman, D.; Mushinski, J. Pvt-1 transcripts are found in normal tissues and are altered by reciprocal (6; 15) translocations in mouse plasmacytomas. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 6964–6968. [CrossRef] [PubMed]
48. Huppi, K.; Siwarski, D. Chimeric transcripts with an open reading frame are generated as a result of translocation to the Pvt-1 region in mouse B-cell tumors. *Int. J. Cancer* **1994**, *59*, 848–851. [CrossRef] [PubMed]
49. Guan, Y.; Kuo, W.L.; Stilwell, J.L.; Takano, H.; Lapuk, A.V.; Fridlyand, J.; Mao, J.H.; Yu, M.; Miller, M.A.; Santos, J.L.; et al. Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin. Cancer Res.* **2007**, *13*, 5745–5755. [CrossRef] [PubMed]
50. Graham, M.; Adams, J.M. Chromosome 8 breakpoint far 3′of the c-*myc* oncogene in a Burkitt's lymphoma 2; 8 variant translocation is equivalent to the murine *pvt*-1 locus. *EMBO J.* **1986**, *5*, 2845. [PubMed]
51. Hodgson, G.; Hager, J.H.; Volik, S.; Hariono, S.; Wernick, M.; Moore, D.; Albertson, D.G.; Pinkel, D.; Collins, C.; Hanahan, D.; et al. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet.* **2001**, *29*, 459–464. [CrossRef] [PubMed]
52. Meyer, K.B.; Maia, A.T.; O'Reilly, M.; Ghoussaini, M.; Prathalingam, R.; Porter-Gill, P.; Ambs, S.; Prokunina-Olsson, L.; Carroll, J.; Ponder, B.A. A functional variant at a prostate cancer predisposition locus at 8q24 is associated with *PVT1* expression. *PLoS Genet.* **2011**, *7*, e1002165. [CrossRef] [PubMed]
53. Chapman, M.H.; Tidswell, R.; Dooley, J.S.; Sandanayake, N.S.; Cerec, V.; Deheragoda, M.; Lee, A.J.; Swanton, C.; Andreola, F.; Pereira, S.P. Whole genome RNA expression profiling of endoscopic biliary brushings provides data suitable for biomarker discovery in cholangiocarcinoma. *J. Hepatol.* **2012**, *56*, 877–885. [CrossRef] [PubMed]
54. Wang, F.; Yuan, J.H.; Wang, S.B.; Yang, F.; Yuan, S.X.; Ye, C.; Yang, N.; Zhou, W.P.; Li, W.L.; Li, W.; et al. Oncofetal long noncoding RNA PVT1 promotes proliferation and stem cell-like property of hepatocellular carcinoma cells by stabilizing NOP2. *Hepatology* **2014**, *60*, 1278–1290. [CrossRef] [PubMed]
55. Zhuang, C.; Li, J.; Liu, Y.; Chen, M.; Yuan, J.; Fu, X.; Zhan, Y.; Liu, L.; Lin, J.; Zhou, Q.; Xu, W.; Zhao, G.; Cai, Z.; Huang, W. Tetracycline-inducible shRNA targeting long non-coding RNA PVT1 inhibits cell growth and induces apoptosis in bladder cancer cells. *Oncotarget* **2015**, *6*, 41194–41203. [CrossRef] [PubMed]
56. Zhou, Q.; Chen, J.; Feng, J.; Wang, J. Long noncoding RNA PVT1 modulates thyroid cancer cell proliferation by recruiting EZH2 and regulating thyroid-stimulating hormone receptor (TSHR). *Tumor Biol.* **2016**, *37*, 3105–3113. [CrossRef] [PubMed]
57. Cui, D.; Yu, C.H.; Liu, M.; Xia, Q.Q.; Zhang, Y.F.; Jiang, W.L. Long non-coding RNA PVT1 as a novel biomarker for diagnosis and prognosis of non-small cell lung cancer. *Tumor Biol.* **2016**, *37*, 4127–4134. [CrossRef] [PubMed]

58. Yang, T.; Zhou, H.; Liu, P.; Yan, L.; Yao, W.; Chen, K.; Zeng, J.; Li, H.; Hu, J.; Xu, H.; et al. lncRNA PVT1 and its splicing variant function as competing endogenous RNA to regulate clear cell renal cell carcinoma progression. *Oncotarget* **2017**, *8*, 85353. [CrossRef] [PubMed]
59. Chen, W.; Zhu, H.; Yin, L.; Wang, T.; Wu, J.; Xu, J.; Tao, H.; Liu, J.; He, X. lncRNA-PVT1 facilitates invasion through upregulation of MMP9 in nonsmall cell lung cancer cell. *DNA Cell Biol.* **2017**, *36*, 787–793. [CrossRef] [PubMed]
60. Zheng, J.; Hu, L.; Cheng, J.; Xu, J.; Zhong, Z.; Yang, Y.; Yuan, Z. lncRNA PVT1 promotes the angiogenesis of vascular endothelial cell by targeting miR-26b to activate CTGF/ANGPT2. *Int. J. Mol. Med.* **2018**, *42*, 489–496. [CrossRef] [PubMed]
61. He, Y.; Jing, Y.; Wei, F.; Tang, Y.; Yang, L.; Luo, J.; Yang, P.; Ni, Q.; Pang, J.; Liao, Q.; et al. Long non-coding RNA PVT1 predicts poor prognosis and induces radioresistance by regulating DNA repair and cell apoptosis in nasopharyngeal carcinoma. *Cell Death Dis.* **2018**, *9*, 235. [CrossRef] [PubMed]
62. Houshmand, M.; Yazdi, N.; Kazemi, A.; Atashi, A.; Hamidieh, A.A.; Najemdini, A.A.; Pour, M.M.; Zarif, M.N. Long non-coding RNA PVT1 as a novel candidate for targeted therapy in hematologic malignancies. *Int. J. Biochem. Cell Biol.* **2018**, *98*, 54–64. [CrossRef] [PubMed]
63. Chen, L.; Ma, D.; Li, Y.; Li, X.; Zhao, L.; Zhang, J.; Song, Y. Effect of long non-coding RNA PVT1 on cell proliferation and migration in melanoma. *Int. J. Mol. Med.* **2018**, *41*, 1275–1282. [CrossRef] [PubMed]
64. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, 289–300.
65. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1979**, *28*, 100–108. [CrossRef]
66. Meilă, M. The uniqueness of a good optimum for k-means. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 625–632.
67. Lisboa, P.J.; Etchells, T.A.; Jarman, I.H.; Chambers, S.J. Finding reproducible cluster partitions for the k-means algorithm. *BMC Bioinform.* **2013**, *14*, S8. [CrossRef] [PubMed]
68. Han, J.D.J.; Bertin, N.; Hao, T.; Goldberg, D.S.; Berriz, G.F.; Zhang, L.V.; Dupuy, D.; Walhout, A.J.; Cusick, M.E.; Roth, F.P.; et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **2004**, *430*, 88–93. [CrossRef] [PubMed]
69. Guimera, R.; Amaral, L.A.N. Functional cartography of complex metabolic networks. *Nature* **2005**, *433*, 895–900. [CrossRef] [PubMed]
70. Suva, M.L.; Rheinbay, E.; Gillespie, S.M.; Patel, A.P.; Wakimoto, H.; Rabkin, S.D.; Riggi, N.; Chi, A.S.; Cahill, D.P.; Nahed, B.V.; et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **2014**, *157*, 580–594. [CrossRef] [PubMed]
71. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
72. Bhan, A.; Soleimani, M.; Mandal, S.S. Long noncoding RNA and cancer: A new paradigm. *Cancer Res.* **2017**, *77*, 3965–3981. [CrossRef] [PubMed]
73. Hu, G.; Niu, F.; Humburg, B.A.; Liao, K.; Bendi, S.; Callen, S.; Fox, H.S.; Buch, S. Molecular mechanisms of long noncoding RNAs and their role in disease pathogenesis. *Oncotarget* **2018**, *9*, 18648. [CrossRef] [PubMed]
74. Peng, Y.; Croce, C.M. The role of microRNAs in human cancer. *Signal Transduct. Targeted Ther.* **2016**, *1*, 15004. [CrossRef] [PubMed]
75. Tabatabai, G.; Weller, M. Glioblastoma stem cells. *Cell Tissue Res.* **2011**, *343*, 459–465. [CrossRef] [PubMed]
76. Schulte, A.; Günther, H.S.; Phillips, H.S.; Kemming, D.; Martens, T.; Kharbanda, S.; Soriano, R.H.; Modrusan, Z.; Zapf, S.; Westphal, M.; et al. A distinct subset of glioma cell lines with stem cell-like properties reflects the transcriptional phenotype of glioblastomas and overexpresses CXCR4 as therapeutic target. *Glia* **2011**, *59*, 590–602. [CrossRef] [PubMed]
77. Galvagni, F.; Orlandini, M.; Oliviero, S. Role of the AP-1 transcription factor FOSL1 in endothelial cells adhesion and migration. *Cell Adhes. Migr.* **2013**, *7*, 408–411. [CrossRef] [PubMed]
78. Wang, K.; Singh, D.; Zeng, Z.; Coleman, S.J.; Huang, Y.; Savich, G.L.; He, X.; Mieczkowski, P.; Grimm, S.A.; Perou, C.M.; et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucl. Acids Res.* **2010**, *8*, e178. [CrossRef] [PubMed]

79. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
80. Hamosh, A.; Scott, A.F.; Amberger, J.S.; Bocchini, C.A.; McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl. Acids Res.* **2011**, *33*, 514–517. [CrossRef] [PubMed]