Pathways of topological rank analysis (PoTRA): a novel method to detect pathways involved in hepatocellular carcinoma

Chaoxing Li Corresp., 1, Li Liu 2, Valentin Dinu 2

1 School of Life Sciences, Arizona State University, Tempe, Arizona, United States
2 Department of Biomedical Informatics, Arizona State University, Scottsdale, Arizona, United States

Corresponding Author: Chaoxing Li
Email address: chaoxing@asu.edu

Complex diseases such as cancer are usually the result of a combination of environmental factors and one or several biological pathways consisting of sets of genes. Each biological pathway exerts its function by delivering signaling through the gene network. Theoretically, a pathway is supposed to have a robust topological structure under normal physiological conditions. However, the pathway’s topological structure could be altered under some pathological condition. It is well known that a normal biological network includes a small number of well-connected hub nodes and a large number of nodes that are non-hubs. In addition, it is reported that the loss of connectivity is a common topological trait of cancer networks, which is an assumption of our method. Hence, from normal to cancer, the process of the network losing connectivity might be the process of disrupting the structure of the network, namely, the number of hub genes might be altered in cancer compared to that in normal or the distribution of topological ranks of genes might be altered. Based on this, we propose a new PageRank-based method called Pathways of Topological Rank Analysis (PoTRA) to detect pathways involved in cancer. We use PageRank to measure the relative topological ranks of genes in each biological pathway, then select hub genes for each pathway, and use Fisher’s exact test to test if the number of hub genes in each pathway is altered from normal to cancer. Alternatively, if the distribution of topological ranks of gene in a pathway is altered between normal and cancer, this pathway might also be involved in cancer. Hence, we use the Kolmogorov-Smirnov test to detect pathways that have an altered distribution of topological ranks of genes between two phenotypes. We apply PoTRA to study hepatocellular carcinoma (HCC) and several subtypes of HCC. Very interestingly, we discover that all significant pathways in HCC are cancer-associated generally, while several significant pathways in subtypes of HCC are HCC subtype-associated specifically. In conclusion, PoTRA is a new approach to explore and discover pathways involved in cancer. PoTRA can be used as a complement to
other existing methods to broaden our understanding of the biological mechanisms behind
cancer at the system-level.
Pathways of topological rank analysis (PoTRA): a novel method to detect pathways involved in hepatocellular carcinoma

Chaoxing Li\textsuperscript{1*}, Li Liu\textsuperscript{2}, Valentin Dinu\textsuperscript{2}

\textsuperscript{1}School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA
\textsuperscript{2}Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ 85255, USA

Corresponding Author:
Chaoxing Li

Email address: chaoxing@asu.edu

Abstract
Complex diseases such as cancer are usually the result of a combination of environmental factors and one or several biological pathways consisting of sets of genes. Each biological pathway exerts its function by delivering signaling through the gene network. Theoretically, a pathway is supposed to have a robust topological structure under normal physiological conditions. However, the pathway’s topological structure could be altered under some pathological condition. It is well known that a normal biological network includes a small number of well-connected hub nodes and a large number of nodes that are non-hubs. In addition, it is reported that the loss of connectivity is a common topological trait of cancer networks, which is an assumption of our method. Hence, from normal to cancer, the process of the network losing connectivity might be the process of disrupting the structure of the network, namely, the number of hub genes might be altered in cancer compared to that in normal or the distribution of topological ranks of genes might be altered. Based on this, we propose a new PageRank-based method called Pathways of Topological Rank Analysis (PoTRA) to detect pathways involved in cancer. We use PageRank to measure the relative topological ranks of genes in each biological pathway, then select hub genes for each pathway, and use Fisher’s exact test to test if the number of hub genes in each pathway is altered from normal to cancer. Alternatively, if the distribution of topological ranks of gene in a pathway is altered between normal and cancer, this pathway might also be involved in cancer. Hence, we use the Kolmogorov-Smirnov test to detect pathways that have an altered distribution of topological ranks of genes between two phenotypes. We apply PoTRA to study hepatocellular carcinoma (HCC) and several subtypes of HCC. Very interestingly, we discover that all significant pathways in HCC are cancer-associated generally, while several significant pathways in subtypes of HCC are HCC subtype-associated specifically. In conclusion, PoTRA is a new approach to explore and discover pathways involved in cancer. PoTRA can be used as a complement to other existing methods to broaden our understanding of the biological mechanisms behind cancer at the system-level.

Introduction

High throughput technologies, such as genomic sequencing and microarrays, allow the genome-wide analysis of molecular factors associated with disease. While the technologies have advanced and have been refined to generate an increasing amount of high quality data, challenges remain in understanding the biological processes involved in disease etiology, particularly for complex disorders.

As we know, individual genomic alterations may result in diseases. For example, Hemophilia A is an X-linked recessive bleeding disorder caused by a deficiency in the activity of coagulation factor VIII (Franchini & Mannucci, 2012). Huntington disease (HD) is an autosomal dominant progressive neurodegenerative disorder with a distinct phenotype characterized by chorea, dystonia, incoordination, cognitive decline, and behavioral difficulties, which is caused by a heterozygous expanded trinucleotide repeat \((\text{CAG})_n\), encoding glutamine, in the gene encoding huntingtin (HTT) on chromosome 4p16 (Walker, 2007; Dayalu & Albin, 2015).

In addition to monogenic diseases, many diseases are complex, such as diabetes, schizophrenia, or cancer, and are believed to be caused by a combination of genomic alterations, epigenetic, environmental and lifestyle factors (Schork, 1997; Hindorff, Gillanders & Manolio, 2011).
Genomic disease association analysis suggests that complex diseases are not caused by individual genomic alterations. First, the complex disease phenotype is associated with many genes. Second, it may be associated with interactions among many genes. Therefore, more and more literature has been focusing on analyzing sets of genes associated with some phenotype. Gene expression profiles have been used to assess the activity of biological networks. Several approaches have been developed to identify active subnetworks across different phenotypes from changes in gene expression. One of the first such studies is a general approach to searching for “active sub-networks” associated with high levels of differential expression (Ideker et al., 2002). This approach identifies a set of genes that form a subnetwork whose expression is altered across two different phenotypes. Another very well-known method, Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), is a computational method that determines whether a pre-defined set of genes shows statistically significant, concordant differences between two phenotypes, which is also based on differential expression of a set of genes between two phenotypes. These approaches, while powerful and popular, are limited by the fact that they ignore the topology of the gene networks and sets that they investigate. Indeed, differential gene expression level in biological network might influence phenotypes. However, only investigating the differential expression levels of gene may be not sufficient since the topology of biological pathway is also an important characteristic of biological pathways and the role they play in both normal and pathological processes, as described below.

It is well known that the topological structure is very important for biological networks and it determines how genes interact with each other, governing how specific genes and biological pathways operate in the promotion or inhibition of human diseases (Tavazoie et al., 1999; Goeman & Bühlmann, 2007; Tarca et al., 2009; Taylor et al., 2009; Khatri, Sirotka & Butte, 2012; Rhinn et al., 2013; Mitrea et al., 2013). Related to this, a hub gene within a biological network is an important gene which acts to influence the activity of a number of genes (Flintoft, 2004), even influence the activity and function of the entire biological network. Hence, there has been an increased interest to analyze the co-regulation and co-expression of genes in the biological network, and many approaches have been developed to identify differential co-regulation and co-expression of genes in the subnetwork (Kostka & Spang, 2004; Lai et al., 2004; Reverter et al., 2006; Watson, 2006; Choi & Kendziorski, 2009; Leonardson et al., 2010; Langfelder et al., 2011; Odibat & Reddy, 2012). It has been a trend to extend differential expression analysis to differential network analysis (de la Fuente, 2010).

Most of the approaches for topology-based network and pathway analysis are based on different correlation-based metrics to identify differential networks between two different phenotypes. Generally, there are three main ways to compare networks for differential network analysis. The first approach handles weighted networks and uses some functions of the edge-specific weight differences as edge weights to construct differential networks (Hudson, Reverter & Dalrymple, 2009; Tesson, Breitling & Jansen, 2010; Liu et al., 2010; Rhinn et al., 2013). The second approach tries to find co-expressed gene sets and identify which correlation patterns are different between sets across conditions (Watson, 2006; Rahmatallah, Emmert-Streib & Glazko, 2014). This approach formulates summary measures that represent co-expression in a biological network and compares the metric between sets. The third approach compares the topology of...
biological networks across different phenotypes by using measures such as degree of nodes or modularity (Reverter et al., 2006; Zhang et al., 2009). However, the PoTRA method uses a topology-based metric to identify differential networks between two phenotypes. In addition to using different metrics, some of the other tools are based on correlation pattern of genes and identify groups of genes whose correlation patterns behave differentially across different datasets (Watson, 2006; Hudson, Reverter & Dalrymple, 2009; Tesson, Breitling & Jansen, 2010; Liu et al., 2010; Rhinn et al., 2013). Compared to these tools, PoTRA is directly based on topological ranks of genes and aims to identify pathways where the topological ranks of genes are different across datasets, which is more biologically intuitive. In this method, not only do we use correlation networks but we also use combined networks by taking intersected networks of correlation networks and KEGG curated pathways. Hence, when KEGG curated pathway information is employed, the topological rank-based PoTRA method can apply to the combined networks, while the previous correlation-based methods cannot, which is a limitation of the previously discussed correlation-only based methods. Regarding the previous tools based on topology (Reverter et al., 2006; Zhang et al., 2009), Zhang et al. focuses on identifying genes involved in topological changes, while PoTRA focuses on identifying pathways involved in topological changes. Also, Reverter et al. focuses on identifying genes with differential connectivity between two phenotypes, which is also different from PoTRA’s application scenario.

Although the above methods for differential network analysis can deal with some important biological questions, they are still limited. In general, they are based on a basic hypothesis that some connections between genes across the groups could be thought of as “passenger” events and other connections are unique to either one of groups and thus could be “driver” events that contribute to disease progression and development. Hence, they focus on the contribution of individual differential connections to disease. This results in several limitations. First, each differential connection is regarded by these methods to have an equal contribution to disease. However, it is well understood that loss of a connection between two hub genes from normal to disease is more deleterious than loss of a connection between two non-hub genes. Second, how differential connections (“driver” connections mentioned above) between pairs of genes are associated with diseases is still not very biologically intuitive, because how the dependency between genes contributes to diseases is usually little understood.

To address these problems, we propose a new PageRank-based method called Pathways of Topological Rank Analysis (PoTRA) to detect pathways associated with cancer. PageRank is an algorithm initially used by Google Search to rank websites in their search engine results (Page et al., 1999). It is a way of measuring the importance of nodes in a network. More generally, PageRank has been applied to other networks, e.g., social networks (Pedroche et al., 2013; Wang et al., 2013). To date, there have been several studies using PageRank for gene expression and network analysis (Morrison et al., 2005; Winter et al., 2012; Kimmel & Visweswaran, 2013; Hou & Ma, 2014; Bourdakou, Athanasiadis & Spyrou, 2016; Zeng et al., 2016; Ramsahai et al., 2017; Morshed Osmani & Rahman, 2018). These studies focus on ranking genes and discovering key driver genes in disease, and do not try to detect dysregulated pathways involved in disease. Other studies (Winter et al., 2012; Zeng et al., 2016) use PageRank to select topological
important genes and simply see which pathways that these topological important genes are involved in. These PageRank-related approaches are very different from our approach.

Our approach embodied by PoTRA is motivated by the observation that the loss of connectivity is a common topological trait of cancer networks (Anglani et al., 2014), as well as the prior knowledge that a normal biological network includes a small number of well-connected hub nodes and a large number of nodes that are non-hubs (Albert, 2005; Khanin & Wit, 2006; Zhu, Gerstein & Snyder, 2007). However, from normal to cancer, the process of the network losing connectivity might be the process of disrupting the structure of the network, namely, the number of hub genes might be altered in cancer compared to that in normal or the distribution of topological ranks of genes might be altered. Thus, we hypothesize that if the number of hub genes is different in a pathway between normal and cancer, this pathway might be involved in cancer. Based on this hypothesis, we propose to detect pathways involved in cancer by testing if the number of hub genes for each pathway is different between normal and cancer samples.

Our approach embodied by PoTRA is also motivated by that the topological ranks of genes within biological pathways might be altered in cancer. Based on this, we propose to detect pathways involved in cancer by testing if the distribution of PageRank scores of genes for each pathway is altered between normal and cancer samples.

Therefore, the PoTRA approach computes topological ranks of genes in each pathway, and then detects pathways with significantly altered number of hub genes between normal and cancer, and, alternatively, detect pathways with significantly altered distributions of topological ranks of genes in corresponding pathways between two phenotypes. Namely, we first use the Google search PageRank algorithm to measure the relative topological ranks of genes in a biological pathway across different conditions. Then, we use Fisher’s exact test to estimate if the number of hub genes in each pathway is significantly different between normal and cancer, and, alternatively, use the Kolmogorov–Smirnov test to estimate if the distribution of PageRank scores of genes in each pathway is significantly different between normal and cancer. As an illustration, we apply PoTRA to study hepatocellular carcinoma (HCC) and its subtypes and identify disease-relevant pathways. In conclusion, PoTRA is a new approach to explore and discover cancer-associated pathways. PoTRA can be used as a complement to other existing methods to enrich our understanding of the biological mechanisms behind cancer at the systems-level.

**Materials and Methods**

Overview of the PoTRA method

**Figure 1. Overview of the PoTRA method.**

Below we detail the steps of the PoTRA method, as illustrated in **Figure 1**.

1. Data
To illustrate the PoTRA method, we use publicly available gene expression datasets from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) hepatocellular carcinoma (HCC) study. We analyze and contrast 50 HCC samples and 50 tumor-adjacent normal samples (“normal samples” in future sections). In addition, the datasets also include gene expression profiles for several HCC subtypes. We further analyze and contrast 22 hepatitis B-induced HCC samples and 22 tumor-adjacent normal samples, 34 hepatitis C-induced HCC samples and 34 tumor-adjacent normal samples, and 50 alcohol-induced HCC samples and 50 tumor-adjacent normal samples. There are 20,531 gene expression values for each sample. Pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa & Goto, 2000) is used. To date, there is much known about etiology of HCC (Beasley, 1988; Sanyal, Yoon & Lencioni, 2010; Wang et al., 2012; Goossens & Hoshida, 2015) and knowledge of pathways involved in HCC (Villanueva et al., 2008; Zhou et al., 2010; Wang et al., 2017), which makes it easier to illustrate and assess the PoTRA method.

2. Construction of gene co-expression network for a pathway

We apply the PoTRA method to gene expression profiles for several phenotypes, such as normal and cancer and cancer subtypes. First, we select genes for each pathway, using pathway information from KEGG. For each pathway, we determine the gene-gene interactions by using the Pearson’s correlation to test each co-expressed gene pair. The test calculates a P-value for the association between each pair of genes. A significance threshold of 0.05 is used. When the P-value of a pair of genes is below 0.05, we establish an edge between the corresponding two genes; otherwise, there is no edge between them. We implement it through a built-in function called “cor.test()” in the statistical software package R (https://www.r-project.org/). In this way, we can construct gene co-expression networks (i.e., pathways) for normal and cancer, separately. Of note, all the gene co-expression networks (i.e., pathways) used by PoTRA are undirected graphs, because co-expression networks only focus on gene pairs with a similar expression pattern across samples, in other words, the transcript levels of two co-expressed genes rise and fall together across samples.

In addition to construction of gene networks based on correlation alone, we, alternatively, also construct gene networks by combining the correlation with the pre-defined interaction from pathway databases.

3. PageRank analysis for genes within a pathway for normal and cancer

Based on the above constructed interactions within a pathway, we can observe that some genes are hub genes whereas others are non-hub genes. We apply the PageRank algorithm (Page et al., 1999) to obtain the corresponding topological importance for each gene within the pathway for normal and cancer, separately, see Figure 2.
Figure 2. The topological rank analysis for each gene within a pathway. For genes within a specified pathway, according to Step 2, we construct a corresponding gene co-expression network for normal and cancer, separately. Then we apply the PageRank method to obtain the topological importance of each gene for normal and cancer, separately. \( PR(gene\ i)_{\text{normal}} \) represents the PageRank score of the gene \( i \) for normal samples, while \( PR(gene\ i)_{\text{cancer}} \) represents the PageRank score of the gene \( i \) for cancer samples.

We implement it by using the page.rank() function from the igraph (Csárdi & Nepusz, 2006) R package. As mentioned in Step 2, all the networks that we construct are undirected graphs. Thus, the PageRank algorithm used in our approach is based on undirected graphs.

The PageRank algorithm

The PageRank algorithm is used by the Google search engine to rank the importance of web pages, which is based on the assumption that the importance of a web page is high in a network if this web page has connections with other nodes of high importance. This idea is naturally applied to analyzing biological networks, where the importance of a gene is high if this gene is connected to other genes of high importance. In our study, the gene-gene network is an undirected graph where a node represents a gene and the edges can be defined by prior knowledge (e.g., KEGG database).

The output from the PageRank algorithm is a probability distribution representing the likelihood that a person randomly clicking on links will arrive at any particular web page. A probability is a numeric value between 0 and 1. The sum of probabilities for all web pages is equal to 1. The probability of a web page is proportional with the time spent at the web page when a person surfs the web. This idea can also be intuitively extended to ranking genes in gene networks where the probability of a gene is proportional with the time a research scientist spends looking and returning at the same gene when analyzing research results. For additional details of PageRank, please refer to (Page et al., 1999).

4. Detect pathways with significant changes between normal and cancer scores

4.1. Detect pathways with significantly altered number of hub genes between normal and cancer using Fisher’s exact test

As mentioned above, PoTRA is motivated by the observation that the loss of connectivity is a common topological trait of cancer networks (Anglani et al., 2014) and the prior knowledge that a normal biological network includes a small number of well-connected hub nodes and a large number of nodes that are non-hubs (Albert, 2005; Khanin & Wit, 2006; Zhu, Gerstein & Snyder, 2007). From normal to cancer, the process of the network losing connectivity might be the process of disrupting the structure of the network, which can result in an altered number of hub genes between normal and cancer. Hence, a statistics that we compare between two phenotypes is the number of hub genes. The PageRank scores of all genes of a pathway form a distribution,
and we use the 95\textsuperscript{th} percentile of the distribution (one-tail) in normal samples as cutoff value for hub genes for both normal and cancer samples. The genes in this pathway with PageRank scores that are above the cutoff value are identified as hub genes for this pathway. Then we count the number of hub genes for normal and cancer, separately. Next, we use Fisher’s exact test to assess if the number of hub genes is significantly different between normal and cancer. For details, see Table 1.

Table 1. **Contingency table for Fisher’s exact test.** We use the 95\textsuperscript{th} percentile of the distribution (one-tail) in normal samples as cutoff value for hub genes for both normal and cancer samples. The value “a” represents the number of genes whose PageRank scores are below the cutoff value for normal samples. The value “b” represents the number of genes whose PageRank scores are above the cutoff value for normal samples. The values “c” and “d” are the corresponding values for cancer. We use Fisher’s exact test to assess if the number of hub genes is significantly different between normal and cancer.

Fisher’s exact test estimates the probability of obtaining any such set of values, given by the hypergeometric distribution:

\[
P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}
\]

where \(n=a+b+c+d\), and \(\binom{i}{j}\) is the binomial coefficient and the symbol “!” indicates the factorial operator.

**Formula 1** gives the exact hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that the number of hub genes is the same for a specified pathway between normal and cancer. If this test statistic is significant, it indicates that there is a significantly different number of hub genes between normal and cancer, thereby this pathway might be involved in cancer. By studying many pathways from the KEGG database we generate a multiple hypothesis testing problem. We address this issue by correcting the P-values calculated for each pathway using the False Discovery Rate (FDR) approach, using the Benjamini and Hochberg procedure (Benjamini & Hochberg, 1995).

4.2. Detect pathways with significantly altered distributions of PageRank scores for genes between normal and cancer using Kolmogorov-Smirnov test

Alternatively, PoTRA is implemented by testing if the distribution of PageRank scores of genes is altered between normal and cancer, using the two-sample Kolmogorov-Smirnov test. The two-sample Kolmogorov-Smirnov test is a nonparametric test of equality of continuous, one-dimensional probability distributions that can be used to compare two samples. The
Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution.

The empirical distribution function \( F_n(x) \) for \( n \) i.i.d. observation \( X_i \) is defined as:

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty, x]}(X_i) \tag{2}
\]

where \( I_{[-\infty, x]}(X_i) \) is the indicator function, equal to 1 if \( X_i \leq x \) and equal to 0 otherwise.

The two-sample Kolmogorov-Smirnov test may be used to test if two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov-Smirnov statistic is:

\[
D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{3}
\]

where \( F_{1,n} \) and \( F_{2,m} \) are the empirical distribution functions of the first and the second sample respectively, and sup is the supremum function.

If the test statistic for a specific pathway is significant, it indicates that the two underlying distributions differ between normal and cancer, thereby this pathway is involved in cancer. By studying many pathways from the KEGG database we generate a multiple hypothesis testing problem. We address this issue by correcting the P-values calculated for each pathway using the False Discovery Rate (FDR) approach, using the Benjamini and Hochberg procedure (Benjamini & Hochberg, 1995).

Software tools

All the analysis is conducted using the R statistical programming language, using the following R Biocoductor packages: graphite for pathway databases, igraph for PageRank function and graph for visualization.

Results

We apply PoTRA to analyze and contrast 50 HCC samples and 50 tumor-adjacent normal samples. All data come from The Cancer Genome Atlas (TCGA) hepatocellular carcinoma (HCC) study.

1. PoTRA for HCC vs. normal samples using correlation-based networks

To illustrate the PoTRA method, we use a cancer-associated pathway, “MAPK signaling pathway”, as an example in the following section.

1.1. Construction of a gene co-expression network for a pathway
As suggested before, “MAPK signaling pathway” might be comprised of different interactions between genes under different conditions, such as normal versus cancer conditions. First, we need to find the genes that this pathway consists of by using the KEGG database. In practice, we implement it by using an R package called graphite. Second, for the genes of this pathway, we identify the interactions between genes for normal and cancer samples, separately. There are 14,005 edges for normal and 5,170 edges for cancer. For the information of “MAPK signaling pathway” in normal and cancer samples, see Supplementary Table S1.

1.2. PageRank analysis for genes within a pathway for normal and cancer

Based on the interactions identified in the previous section, we can obtain a PageRank score for each gene in “MAPK signaling pathway” for normal and cancer, separately, which quantifies the influence of a gene on the activity of other genes in this pathway. For the results for this step, 2 vectors with 250 PageRank values, one for normal and one for cancer, separately, see Supplementary Table S1. As previously mentioned, the PageRank values in each vector add up to 1.

Figure 3 illustrates the distributions of PageRank scores for genes in “MAPK signaling pathway” for normal and cancer, separately.

Figure 3. The kernel density distribution of PageRank scores of genes in “MAPK signaling pathway”. The red line shows the kernel density distribution of PageRank scores for cancer and the black one is for normal samples. Note that the mean for the two distributions is the same, i.e., mean = 1/N=0.004, where N = 250 is the number of genes in the “MAPK signaling pathway” pathway. We use the 95th-percentile cutoff (= 0.006035) of the kernel distribution in normal samples as cutoff for hub genes for both normal and cancer samples.

The number of hub genes changes. While there are only 13, strongly-connected (with more edges, an average of 121 edges) hub genes in the normal samples, there are 40 hub genes in the cancer samples, more loosely-connected (with fewer edges, an average of 75 edges).

As mentioned before, the process of the network losing connectivity might be the process of disrupting the structure of the network which includes a small number of hub nodes and a large number of non-hub nodes, namely, the number of hub genes might be altered in cancer compared to that in normal. The altered number of hub genes might be a trait of a pathway in cancer, which suggests the pathway is involved in cancer if the change in hub gene number is statistically significant.

Importantly, we find increased variance (median change 5.2 -fold for the MAPK pathway, or 6.1 fold change for all genes) of gene expression in cancer samples compared to normal samples, which is included in Supplementary Table S1. The increased variance in cancer samples results in lower Pearson correlations between genes in cancer, which leads to lower correlations and loss of connectivity in cancer. This loss of connectivity can lead to the disruption of the structure of
the biological networks. To quantify these changes in the network topology between normal and
disease, we use Fisher’s exact test to identify changes in the number of network hub genes, and
the Kolmogorov Smirnov test to test if the distributions of PageRank scores are significantly
different between normal and cancer. These tests are described below.

1.3.1. Fisher’s exact test for comparing the number of hub genes in the pathway

We next use Fisher’s exact test to test if the number of hub genes for “MAPK signaling
pathway” is significantly different between normal and cancer. The result for the “MAPK
signaling pathway” pathway is included in Table 2.

Table 2. The “MAPK signaling pathway” pathway identified by PoTRA for HCC using
Fisher’s exact test. The P value is adjusted by False Discovery Rate (FDR).

| Pathway                  | P Value (FDR adjusted) |
|--------------------------|------------------------|
| MAPK signaling pathway   | < 0.05                 |

The low P-value in Table 2 indicates that the number of hub genes in cancer samples is
significantly different from that in normal samples, suggesting that the “MAPK signaling
pathway” pathway is involved in HCC. This example suggests that a normal biological network
includes a small number of hub nodes and a large number of non-hub nodes. Moreover, the loss
of connectivity from normal to cancer might lead to disrupting the structure of the network in
cancer, thereby resulting in the fact that the number of hub genes is altered in cancer compared to
that in normal.

Then we apply the same approach to other pathways from KEGG to compare HCC vs. normal
samples. The significant pathways are shown in Table 3.

Table 3. The significant KEGG pathways identified by PoTRA for HCC using Fisher’s
exact test. FDR adjusted P-values are below 0.05.

| Pathway                  | P Value (FDR adjusted) |
|--------------------------|------------------------|
| MAPK signaling pathway   | < 0.05                 |
| Other pathways           | > 0.05                 |

We find three significant pathways with altered number of hub genes between normal and
cancer. It is well known that these three pathways are strongly associated with cancer in general.
MAPK signaling pathway plays a role in the regulation of gene expression, cellular growth, and
survival (Knight & Irving, 2014). Abnormal MAPK signaling might lead to uncontrolled or
increased cell proliferation and resistance to apoptosis (Santarpia, Lippman & El-Naggar, 2012;
Burotto et al., 2014). Interestingly, we also find that loss of connectivity and the larger number of
hub genes for cancer are characteristics of the other two pathways as well.

The levels for parameter thresholds can be varied. Hence, we also have tried different threshold
levels for hub genes in Fisher’s exact test: 0.9, 0.85, 0.8, 0.75 and 0.7. For these threshold levels,
the p-values are larger than that when the threshold is 0.95, and we do not find any significant
pathways using Fisher’s exact test. This suggests that, for this test, using a limited number of
genes of higher “hubness” (higher PageRank values) is important when comparing normal and
cancer.
1.3.2. The Kolmogorov–Smirnov test for comparing distributions of PageRank scores of genes for pathways

We next use the Kolmogorov–Smirnov test to test if the two distributions of PageRank scores of genes for “MAPK signaling pathway” are significantly different between normal and cancer. The result for the “MAPK signaling pathway” pathway is included in Table 4.

Table 4. The “MAPK signaling pathway” pathway identified by PoTRA for HCC using the Kolmogorov–Smirnov test. The P value is adjusted by False Discovery Rate (FDR).

The low P-value in Table 4 indicates that the distribution of PageRank scores of genes in cancer samples is significantly different from that in normal samples, suggesting that the “MAPK signaling pathway” pathway is involved in HCC. MAPK signaling pathway plays a role in the regulation of gene expression, cellular growth, and survival (Knight & Irving, 2014). Abnormal MAPK signaling might lead to uncontrolled or increased cell proliferation and resistance to apoptosis (Santarpia, Lippman & El-Naggar, 2012; Burotto et al., 2014).

Then we apply the same approach to other pathways from KEGG to compare HCC with normal samples. The significant pathways are shown in Table 5.

Table 5. The significant KEGG pathways identified by PoTRA for HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.

We find three significant pathways with altered distribution of PageRank scores of genes between normal and cancer. In addition to “MAPK signaling pathway”, we find two other pathways: “RNA transport” and “mRNA surveillance pathway”. For the “RNA transport” pathway, it is well known that the nuclear export of mRNA is intrinsically linked to the co-transcriptional processing of nascent transcripts synthesized by the RNA polymerase II. This functional coupling is very important for the survival of cells allowing for nuclear export of fully processed transcripts, which could lead to genome instability and to various forms of cancer (Siddiqui & Borden, 2012; Hautbergue, 2017). The “mRNA surveillance pathway” is a quality control mechanism that detects and degrades abnormal mRNAs, which includes nonsense-mediated mRNA decay (NMD), nonstop mRNA decay (NSD), and no-go decay (NGD). NMD can degrade PTC (Premature termination codons)-containing transcripts which cause a large proportion of human cancers (Lindeboom, Supek & Lehner, 2016). Nonstop decay is the mechanism of identifying and disposing aberrant transcripts that lack in-frame stop codons. These transcripts are identified during translation when the ribosome arrives at the 3’ end of the mRNA and stalls at the end of the poly(A) tail. This surveillance mechanism protects the cells from potentially harmful truncated proteins, but it may also be involved in mediating critical cellular functions of transcripts that are prone to stop codon read-through, which have implications in human cancers (Klauer & van Hoof, 2012). No-go decay (NGD) is a eukaryotic
quality control mechanism that evolved to cope with translational arrests (Simms, Yan & Zaher, 2017). These above processes are strongly related to cancer.

1.4. PoTRA for cancer subtype analysis

1.4.1. Fisher’s exact test for cancer subtype analysis

Many complex diseases have subtypes and/or can be classified into different categories based on diagnosis, pathology, phenotype characteristics, etc. To further assess the PoTRA method, we apply it to several subtypes of the HCC TCGA data. There are several risk factors associated with HCC, such as hepatitis B, hepatitis C and alcohol (Beasley, 1988; Sanyal, Yoon & Lencioni, 2010; Hoshida et al., 2014; Goossens & Hoshida, 2015). Here, we apply PoTRA to compare these three subtypes of HCC samples with normal samples.

Table 6 illustrates the Fisher’s exact test results for comparing normal with hepatitis B-induced HCC samples.

Table 6. The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using Fisher’s exact test. FDR adjusted P-values are below 0.05.

There is one common pathway, Pathways in cancer, between Table 6 and Table 3. There are seven other new pathways, which are very interesting and associated with the hepatitis B-induced HCC. First, the “Hepatitis B” pathway is detected by our method. Hepatitis B is the most important and direct factor causing hepatitis B-induced HCC. In addition, we find two other pathways, HTLV-I (Human T-cell leukemia virus type I) infection and Epstein-Barr virus infection, which are strongly associated with virus infection and cancer. This is consistent with the viral pathology of hepatitis B-induced HCC. Besides, some studies show that hepatitis B virus infection can contribute to the impairment of insulin signaling, which is another pathway identified by PoTRA (Kim, Kim & Cheong, 2010; Barthel et al., 2016). Finally, the other three pathways, Hippo signaling pathway, Neurotrophin signaling pathway and mTOR signaling pathway are associated with cancer in general. Hippo signaling pathway is reported to be able to control organ size through regulating cell proliferation and apoptosis (Saucedo & Edgar, 2007; Pan, 2010). It is reported that neurotrophins can regulate cancer stem cells (Chopin et al., 2016), and neurotrophins contribute to pro-survival signaling in many different types of cancer (Molloy, Read & Gorman, 2011). mTOR signaling pathway is a well-known cancer-associated pathway. Alterations of mTOR signaling pathway have significant effects on cancer progression. The major components of mTOR signaling pathway are critical effectors in cell signaling pathways commonly deregulated in cancers (Guertin & Sabatini, 2007; Villanueva et al., 2008; Pópulo, Lopes & Soares, 2012).

These results suggest that PoTRA can be used to identify not only the pathways associated with cancer in general, but also those pathways associated with cancer subtypes, such as hepatitis B-induced HCC specifically.
Results of the PoTRA analysis from two other HCC subtypes, hepatitis C-induced HCC and alcohol-induced HCC, are included in Table 7 and Table 8, respectively.

Table 7. The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using Fisher’s exact test. FDR adjusted P-values are below 0.05.

In Table 5, we find two common pathways, Pathways in cancer and MAPK signaling pathway, between Table 5 and Table 3. Among the other pathways, we find several pathways related to cancer generally and hepatitis C-induced HCC specifically. First, HTLV-I infection is also listed in this table, and, as mentioned above, is associated with virus infection and cancer. Almost all other pathways are associated with cancer in general. PI3K-Akt signaling pathway is a key regulator of normal cellular processes involved in cell growth, proliferation, motility, survival, and apoptosis (Porta, Paglino & Mosca, 2014). The Proteoglycans in cancer pathway is involved in regulation of proteoglycans, heavily glycosylated proteins present especially in connective tissue in cancer. Rap1 signaling pathway is reported to be involved in cancer cell migration, invasion and metastasis (Bailey, 2009; Zhang et al., 2017). The cAMP signaling pathway regulates a number of biological processes, such as cell growth and adhesion, neuronal signaling, energy homeostasis and muscle relaxation (Fajardo, Piazza & Tinsley, 2014). The key component of Focal adhesion pathway, Focal adhesion kinase (FAK), is reported to enable activation by growth factor receptors or integrins in different types of cancers. FAK is an important mediator of cell proliferation, cell migration, cell growth (Golubovskaya, Kweh & Cancé, 2009; Tai, Chen & Shen, 2015). A large volume of literature shows Ras signaling pathway is involved in several aspects of normal cell growth and malignant transformation, and plays an important role in cancer development and progression (Vojtek & Der, 1998; Downward, 2003; Santarpia, Lippman & El-Naggar, 2012; Knight & Irving, 2014). FoxO signaling pathway is involved in the regulation of the cell cycle, apoptosis and metabolism (Schmidt et al., 2002; Fu & Tindall, 2008; Gross, van den Heuvel & Birnbaum, 2008). Besides, activity of FoxO signaling pathway also affects stem cell maintenance and lifespan (Eijkelenboom & Burgering, 2013). ErbB signaling pathway plays roles in cancer development and progression (Hynes & Lane, 2005; Seshacharyulu et al., 2012), as well as in cancer cell migration and invasion (Appert-Collin et al., 2015). ErbB signaling pathway is associated with the development of a wide variety of types of solid tumor if ErbB signaling is excessive (Cho & Leahy, 2002). The Axon guidance pathway is also reported to regulate cell migration and apoptosis, and be associated with tumorigenesis (Chédotal, Kerjan & Moreau-Fauvarque, 2005).

Table 8. The significant KEGG pathways identified by PoTRA for alcohol-induced HCC using Fisher’s exact test. FDR adjusted P-values are below 0.05.

We find two common pathways between Table 8 and Table 3, MAPK signaling pathway and Pathways in cancer. As mentioned above, PI3K-Akt signaling pathway also plays an important role in cancer (Porta, Paglino & Mosca, 2014).
1.4.2. Kolmogorov–Smirnov test for cancer subtype analysis

Table 9 illustrates the Kolmogorov–Smirnov test results for comparing normal with hepatitis C-induced HCC samples.

Table 9. The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.

In Table 9, we find two common pathways, RNA transport and MAPK signaling pathway, between Table 9 and Table 5. Among the other pathways, we find pathways related to cancer generally and hepatitis C-induced HCC specifically. First, HTLV-I infection is also listed in this table, and, as mentioned above, is associated with virus infection and cancer. PI3K-Akt signaling pathway is a key regulator of normal cellular processes involved in cell growth, proliferation, motility, survival, and apoptosis (Porta, Paglino & Mosca, 2014). Importantly, hepatitis C virus activates PI3K-Akt signaling to enhance entry and replication, and meanwhile PI3K-Akt signaling pathway also can increase HCV translation (Liu et al., 2012; Shi, Hoffman & Liu, 2016), which suggest that PI3K-Akt signaling pathway is associated with hepatitis C virus infection specifically. The “Pathways in cancer” pathway is associated with cancer in general. The “Proteoglycans in cancer” pathway is involved in regulation of proteoglycans, heavily glycosylated proteins present especially in connective tissue in cancer (Iozzo & Sanderson, 2011; Baghy et al., 2016).

Results of the PoTRA analysis from two other HCC subtypes, alcohol-induced HCC and hepatitis B-induced HCC, are included in Table 10 and Table 11, respectively.

Table 10. The significant KEGG pathways identified by PoTRA for alcohol-induced HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.

We find three common pathways between Table 10 and Table 5, RNA transport, MAPK signaling pathway and mRNA surveillance pathway. As mentioned above, the “Pathways in cancer” pathway is strongly associated with cancer in general and PI3K-Akt signaling pathway also plays an important role in cancer (Porta, Paglino & Mosca, 2014).

Table 11. The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.

There is no common pathway between Table 11 and Table 5. Very interestingly, for the association between hepatitis B-induced HCC and the “Primary bile acid biosynthesis” pathway, some studies demonstrate that hepatitis B virus infection can alter bile acid metabolism as a consequence of impaired bile acid uptake (Oehler et al., 2014; Geier, 2014). Also, some studies show that hepatitis B virus infection can contribute to the impairment of insulin signaling.
Kim & Cheong, 2010; Barthel et al., 2016). These two pathways mentioned are specifically related to hepatitis B virus infection. Arginine and proline metabolism is one of the central pathways for the biosynthesis of the amino acids arginine and proline from glutamate. Some studies have suggested that altered arginine and proline metabolism is linked to metastasis formation in cancer (Elia et al., 2017). Arginine serves as an intermediate in the urea cycle and as a precursor for protein, polyamine, creatine and nitric oxide (NO) biosynthesis. NO may influence tumor initiation, promotion, and progression, tumor-cell adhesion, apoptosis angiogenesis, differentiation, chemosensitivity, radiosensitivity, and tumor-induced immunosuppression (Lind, 2004).

2. PoTRA for HCC vs. normal samples using networks that combine correlation-based networks with curated interaction networks

The above results show that the Fisher’s exact test and the Kolmogorov–Smirnov test can identify cancer-related pathways based on correlation networks. In addition to correlation networks, in this section we apply PoTRA to gene networks constructed by intersecting correlation networks with pre-defined networks from the KEGG database. As performed in the previous section we investigate if Fisher’s exact test and Kolmogorov–Smirnov test are still able to robustly discover cancer-associated pathways and differential pathways between HCC and subtypes of HCC based on the combined networks.

2.1. Fisher’s exact test for HCC and subtypes of HCC

For HCC, we use Fisher’s exact test and identify no significant pathway with altered number of hub genes between normal and HCC.

For hepatitis C-induced HCC, we identify two significant pathways, listed in Table 12.

| Table 12. The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using the Fisher’s exact test based on combined networks. FDR adjusted P-values are below 0.05. E.comb.normal represents the number of edges in the combined network for normal samples, while E.comb.case is for cancer samples, respectively. |
| As mentioned above, Epstein-Barr virus infection is a pathway associated with virus infection, which is specifically related to hepatitis C induced HCC. P53 signaling pathway is a classical oncogenic pathway, and it can regulate the cell cycle, apoptosis and help prevent cancer. The major component of p53 signaling pathway, p53 protein, is most frequently altered in human cancer (May & May, 1999; Sherr & McCormick, 2002; Sui et al., 2011; Stegh, 2012).

For hepatitis B-induced HCC, we identify one significant pathway, listed in Table 13.
Table 13. The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using the Fisher’s exact test based on combined networks. FDR adjusted P-values are below 0.05. E.comb.normal represents the number of edges in the combined network for normal samples, while E.comb.case is for cancer samples, respectively.

As such, we identify the same significant pathway “Epstein-Barr virus infection”, which is also specific for hepatitis C-induced HCC, discussed above.

For alcohol-induced HCC, we identify no significant pathways.

First, we can observe a large loss of connectivity for combined networks vs. correlation networks in normal and in cancer. For example, Table 13 lists only 84 edges in normal and 22 edges in cancer samples for the Epstein-Barr virus infection pathway using combined correlation and curated interactions. This is in contrast to 1524 edges in normal and 435 edges in cancer shown in Table 6 for the same pathway, when using all the correlation-based edges. We can observe that the proportion of edges in normal to cancer for the correlation network and the combined network are approximately the same (about 4:1). On average, we observe a reduction by, a factor of 23.18 in the number of edges between correlation and combined networks. However, the PoTRA-Fisher’s exact test still can identify pathways specifically involved in subtypes of HCC.

Furthermore, because combined networks have a large loss of connectivity from correlation networks to combined networks, the PageRank scores of genes are more evenly distributed in combined networks than in correlation networks. Hence, fewer significant pathways with altered number of hub genes between normal and cancer are identified for combined networks than that for correlation networks. Although Fisher’s exact test identifies fewer significant pathways, this test can still discover differential pathways specific for subtypes of HCC between normal and cancer, suggesting that the Fisher’s exact test is able to robustly discover pathways involved in cancer and subtypes of cancer.

2.2. Kolmogorov–Smirnov test for HCC and subtypes of HCC

All significant pathways for HCC and each subtype of HCC (hepatitis C-induced HCC, hepatitis B-induced HCC and alcohol-induced HCC) using the Kolmogorov-Smirnov test on combined networks are included in Supplementary Table S2, Supplementary Table S3, Supplementary Table S4 and Supplementary Table S5, respectively. Here, we review the top 10 significant pathways from HCC and each subtype of HCC.

Among the results of the PoTRA analysis for HCC in Supplementary Table S2, we can find many pathways involved in HCC in the top 10 significant pathways, such as Cytokine-cytokine receptor interaction, cAMP signaling pathway, p53 signaling pathway, etc. Cytokine-cytokine receptor interaction exerts a vast array of immunoregulatory actions critical to cancers (Schreiber & Walter, 2010; Spangler et al., 2015). The cAMP signaling pathway regulates a number of biological processes, such as cell growth and adhesion, neuronal signaling, energy homeostasis and muscle relaxation (Fajardo, Piazza & Tinsley, 2014). P53 signaling pathway is a classical oncogenic pathway as mentioned above.
Among the results for hepatitis C-induced HCC (Supplementary Table S3), we find seven non-common pathways for this subtype of HCC in the top 10 pathways, such as MAPK signaling pathway, Calcium signaling pathway, cGMP-PKG signaling pathway, Adrenergic signaling in cardiomyocytes, HIF-1 signaling pathway, Toll-like receptor signaling pathway and Insulin resistance. Among them, some of them are strongly related to immune system and inflammation, such as cGMP-PKG signaling pathway, HIF-1 signaling pathway, Toll-like receptor signaling pathway and Insulin resistance. These immune- and inflammation-related pathways are also specific for hepatitis C (virus infection)-induced HCC.

Among the results for hepatitis B-induced HCC (Supplementary Table S4), we also find seven non-common pathways between this subtype of HCC and HCC in the top 10 pathways, such as Adrenergic signaling in cardiomyocytes, Breast cancer, Calcium signaling pathway, cGMP-PKG signaling pathway, MAPK signaling pathway, PI3K-Akt signaling pathway and Tuberculosis. Among them, there are some pathways related to immune system and inflammation, such as cGMP-PKG signaling pathway and PI3K-Akt signaling pathway, which are specific for hepatitis B (virus infection)-induced HCC.

Among the results for alcohol-induced HCC (Supplementary Table S5), we also find four non-common pathways between this subtype of HCC and HCC in the top 10 pathways, such as MAPK signaling pathway, Pathways in cancer, Calcium signaling pathway and Progesterone-mediated oocyte maturation. Among these four pathways, Calcium signaling pathway is shown to be specifically associated with alcohol in some studies (Gruol & Parsons, 1996; Li, Li & Guo, 2014; Bartlett et al., 2017).

As mentioned above, the combined networks, which intersect correlation networks with curated database networks, have a large loss of gene-gene interactions when compared to correlation networks. Hence, the distribution of PageRank scores of genes might be changed more from normal to cancer than that in correlation networks. Hence, more significant pathways with altered distribution of PageRank scores of genes are identified for combined networks than that for correlation networks. Although the Kolmogorov–Smirnov test identifies a large amount of significant pathways, we can use this test for combined networks to rank the pathways associated with cancer.

Discussion

We propose a PageRank-based method, Pathway of Topological Rank Analysis (PoTRA), for identifying pathways involved in cancer. PoTRA is motivated by the observation that the loss of connectivity is a common topological trait of cancer networks (Anglani et al., 2014) and the prior knowledge that a normal biological network includes a small number of hub nodes and a large number of non-hub nodes (Albert, 2005; Khanin & Wit, 2006; Zhu, Gerstein & Snyder, 2007). From normal to cancer, the process of the network losing connectivity might be the process of disrupting the structure of the network, which can result in an altered number of hub genes between normal and cancer. The PoTRA analysis is based on topological ranks of genes in
biological pathways, and PoTRA detects pathways involved in cancer by testing if the number of hub genes in pathways is altered between normal and cancer.

To illustrate the method, PoTRA is applied to several TCGA hepatocellular carcinoma datasets. The results in our study are in agreement with prior knowledge of HCC from literature. We find that a high proportion of statistically significant pathways play important roles in cancer, indicating that the altered number of hub genes for these pathways might indeed be a reflection of the underlying biological causes that lead to cancer. Moreover, in the comparison between normal and each subtype of HCC, most importantly, the “Hepatitis B” pathway and several pathways associated with virus infection dramatically become significant pathways in hepatitis B-induced HCC, suggesting that PoTRA is capable of detecting pathways associated with disease subtypes. We also find several pathways associated with HCC generally and subtype specifically in hepatitis C-induced HCC and in alcohol-induced HCC.

In our approach, the correlation method is used to construct gene co-expression networks for normal and cancer, respectively. A gene co-expression network is an undirected graph, where each node represents a gene, and each edge is established if there is a significant co-expression relationship between two genes (Stuart et al., 2003). Stuart JM, et al. (Stuart et al., 2003) found 22,163 co-expression relationships, each of which has been conserved across evolution, suggesting that the co-expressions between genes confer a selective advantage and thus these genes are functionally related. Gene co-expression networks are biologically interesting since co-expressed genes might be controlled by members of the same pathway, or the same transcriptional regulatory program or protein complex (Weirauch, 2011), and could be functionally related, suggesting that co-expression is common in the human genome. A gene co-expression network can be constructed by looking for pairs of genes with a similar expression pattern across samples, i.e., the transcript levels of two co-expressed genes rise and fall together across samples. As method of network construction, we use Pearson’s correlation in consideration of saving computing time, because the other methods are relatively computationally intensive. In addition, we also use combined networks by taking the intersection of KEGG curated networks and correlation networks, which increases the reliability of network construction. This approach also shows that the results using correlation networks and the results using combined networks are consistent.

Here, we construct gene networks using two ways. One approach is based on correlation networks, while the other approach is to combine (intersect) the correlation networks with pre-defined networks from pathway databases. Because the combined networks lose a large amount of gene-gene interactions from the correlation networks, the power is reduced and the PageRank scores of genes tend to more evenly distributed in combined networks than that in correlation networks. Hence, fewer significant pathways with altered number of hub genes between two phenotypes are identified for combined networks than that for correlation networks. Moreover, the distribution of topological ranks of genes might be changed more from normal to cancer than that in correlation networks. Thus, more significant pathways with altered distribution of PageRank scores of genes are identified for combined networks than that for correlation
networks. Although the Kolmogorov–Smirnov test identifies a number of significant pathways for combined networks, we can use the Kolmogorov–Smirnov test to rank those pathways.

We apply the Fisher’s exact test and the Kolmogorov–Smirnov test to analyze the PageRank scores for the two types of gene networks as mentioned above. The corresponding results suggest that the Fisher’s exact test and the Kolmogorov–Smirnov test can identify and rank cancer-associated pathways. This suggests that the PoTRA method is robust to choice of network building approach and to statistical analysis method for identification of cancer-related pathways.

**Future Directions**

The hypothesis of our study is based on the fact that the loss of connectivity is a common topological trait of cancer networks (Anglani et al., 2014). It is not yet well understood if this trait is a characteristic of other complex diseases. Thus, we need to be cautious about the applicability of this method to other diseases. However, this trait could be applicable to other complex diseases. Thus, although PoTRA is motivated by work on cancer, it could apply to other complex diseases as well. This area needs to be further investigated.

In this study, we apply PoTRA to pre-defined biological pathways, from the well-curated KEGG pathway database. However, the PoTRA method can also be applied to any set of genes of interest, such as functional gene subnetworks. This could be an interesting area to further explore.

In this article, we focus on the details of the PoTRA methodology. For different pathway databases, such as Reactome, Biocarta, etc., the method would apply in a similar manner, while the final results might vary slightly, depending on the data being used. This direction – changes in analysis results based on different pathway databases – would be an interesting area to investigate in the future.

**Conclusion**

In summary, PoTRA provides a new method for detecting cancer-associated pathways. PoTRA may be used to augment existing methods and provide a richer, more systematic understanding of cancer mechanisms.

**Availability**

R software to carry out the PoTRA computation is available via [http://www.dinulab.org/tools](http://www.dinulab.org/tools).

**Reference**
Albert R. 2005. Scale-free networks in cell biology. *Journal of Cell Science* 118:4947–4957. DOI: 10.1242/jcs.02714.

Anglani R., Creanza TM., Liuzzi VC., Piepoli A., Panza A., Andriulli A., Ancona N. 2014. Loss of Connectivity in Cancer Co-Expression Networks. *PLOS ONE* 9:e87075. DOI: 10.1371/journal.pone.0087075.

Appert-Collin A., Hubert P., Crémel G., Bennasroune A. 2015. Role of ErbB Receptors in Cancer Cell Migration and Invasion. *Frontiers in Pharmacology* 6. DOI: 10.3389/fphar.2015.00283.

Baghy K., Tátrai P., Regős E., Koválszky I. 2016. Proteoglycans in liver cancer. *World Journal of Gastroenterology* 22:379–393. DOI: 10.3748/wjg.v22.i1.379.

Bailey CL. 2009. The Roles of Rap1 in Cancer Metastasis and Pancreatic Islet Beta Cell Function.

Barthel SR., Medvedev R., Heinrich T., Büchner SM., Kettern N., Hildt E. 2016. Hepatitis B virus inhibits insulin receptor signaling and impairs liver regeneration via intracellular retention of the insulin receptor. *Cellular and molecular life sciences: CMLS* 73:4121–4140. DOI: 10.1007/s00018-016-2259-1.

Bartlett PJ., Antony AN., Agarwal A., Hilly M., Prince VL., Combettes L., Hoek JB., Gaspers LD. 2017. Chronic alcohol feeding potentiates hormone-induced calcium signalling in hepatocytes. *The Journal of Physiology* 595:3143–3164. DOI: 10.1113/JP273891.

Beasley RP. 1988. Hepatitis B virus. The major etiology of hepatocellular carcinoma. *Cancer* 61:1942–1956.

Benjamini Y., Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57:289–300.

Bourdakou MM., Athanasiadis EI., Spyrou GM. 2016. Discovering gene re-ranking efficiency and conserved gene-gene relationships derived from gene co-expression network analysis on breast cancer data. *Scientific Reports* 6:20518. DOI: 10.1038/srep20518.
Burotto M., Chiou VL., Lee J-M., Kohn EC. 2014. The MAPK pathway across different malignancies: A new perspective. *Cancer* 120:3446–3456. DOI: 10.1002/cncr.28864.

Chédotal A., Kerjan G., Moreau-Fauvarque C. 2005. The brain within the tumor: new roles for axon guidance molecules in cancers. *Cell Death and Differentiation* 12:1044–1056. DOI: 10.1038/sj.cdd.4401707.

Cho H-S., Leahy DJ. 2002. Structure of the extracellular region of HER3 reveals an interdomain tether. *Science (New York, N.Y.)* 297:1330–1333. DOI: 10.1126/science.1074611.

Choi Y., Kendziorski C. 2009. Statistical methods for gene set co-expression analysis. *Bioinformatics (Oxford, England)* 25:2780–2786. DOI: 10.1093/bioinformatics/btp502.

Chopin V., Lagadec C., Toillon R-A., Le Bourhis X. 2016. Neurotrophin signaling in cancer stem cells. *Cellular and molecular life sciences: CMLS* 73:1859–1870. DOI: 10.1007/s00018-016-2156-7.

Csárdi G., Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Systems*.

Dayalu P., Albin RL. 2015. Huntington Disease: Pathogenesis and Treatment. *Neurologic Clinics* 33:101–114. DOI: 10.1016/j.ncl.2014.09.003.

Downward J. 2003. Targeting RAS signalling pathways in cancer therapy. *Nature Reviews Cancer* 3:11–22. DOI: 10.1038/nrc969.

Eijkelenboom A., Burgering BMT. 2013. FOXOs: signalling integrators for homeostasis maintenance. *Nature Reviews Molecular Cell Biology* 14:83–97. DOI: 10.1038/nrm3507.

Elia I., Broekaert D., Christen S., Boon R., Radaelli E., Orth MF., Verfaillie C., Grünewald TGP., Fendt S-M. 2017. Proline metabolism supports metastasis formation and could be inhibited to selectively target metastasizing cancer cells. *Nature Communications* 8:ncomms15267. DOI: 10.1038/ncomms15267.
Fajardo AM., Piazza GA., Tinsley HN. 2014. The Role of Cyclic Nucleotide Signaling Pathways in Cancer: Targets for Prevention and Treatment. *Cancers* 6:436–458. DOI: 10.3390/cancers6010436.

Flintoft L. 2004. Rewiring the network. *Nature Reviews Genetics* 5:808–808. DOI: 10.1038/nrg1476.

Franchini M., Mannucci PM. 2012. Past, present and future of hemophilia: a narrative review. *Orphanet Journal of Rare Diseases* 7:24. DOI: 10.1186/1750-1172-7-24.

Fu Z., Tindall DJ. 2008. FOXOs, cancer and regulation of apoptosis. *Oncogene* 27:2312–2319. DOI: 10.1038/onc.2008.24.

de la Fuente A. 2010. From ‘differential expression’ to ‘differential networking’ – identification of dysfunctional regulatory networks in diseases. *Trends in Genetics* 26:326–333. DOI: 10.1016/j.tig.2010.05.001.

Geier A. 2014. Hepatitis B virus: The “metabolovirus” highjacks cholesterol and bile acid metabolism. *Hepatology* 60:1458–1460. DOI: 10.1002/hep.27224.

Goeman JJ., Bühlmann P. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)* 23:980–987. DOI: 10.1093/bioinformatics/btm051.

Golubovskaya VM., Kweh FA., Cance WG. 2009. Focal adhesion kinase and cancer. *Histology and Histopathology* 24:503–510. DOI: 10.14670/HH-24.503.

Goossens N., Hoshida Y. 2015. Hepatitis C virus-induced hepatocellular carcinoma. *Clinical and Molecular Hepatology* 21:105–114. DOI: 10.3350/cmh.2015.21.2.105.

Gross DN., van den Heuvel APJ., Birnbaum MJ. 2008. The role of FoxO in the regulation of metabolism. *Oncogene* 27:2320–2336. DOI: 10.1038/onc.2008.25.

Gruol DL., Parsons KL. 1996. Chronic alcohol reduces calcium signaling elicited by glutamate receptor stimulation in developing cerebellar neurons. *Brain Research* 728:166–174. DOI: 10.1016/0006-8993(96)00185-0.
Guertin DA., Sabatini DM. 2007. Defining the Role of mTOR in Cancer. Cancer Cell 12:9–22. DOI: 10.1016/j.ccr.2007.05.008.

Hautbergue GM. 2017. RNA Nuclear Export: From Neurological Disorders to Cancer. Advances in Experimental Medicine and Biology 1007:89–109. DOI: 10.1007/978-3-319-60733-7_6.

Hindorff LA., Gillanders EM., Manolio TA. 2011. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis 32:945–954. DOI: 10.1093/carcin/bgr056.

Hoshida Y., Fuchs BC., Bardeesy N., Baumert TF., Chung RT. 2014. Pathogenesis and prevention of hepatitis C virus-induced hepatocellular carcinoma. Journal of Hepatology 61:S79-90. DOI: 10.1016/j.jhep.2014.07.010.

Hou JP., Ma J. 2014. DawnRank: discovering personalized driver genes in cancer. Genome Medicine 6:56. DOI: 10.1186/s13073-014-0056-8.

Hudson NJ., Reverter A., Dalrymple BP. 2009. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. PLoS computational biology 5:e1000382. DOI: 10.1371/journal.pcbi.1000382.

Hynes NE., Lane HA. 2005. ERBB receptors and cancer: the complexity of targeted inhibitors. Nature Reviews Cancer 5:341–354. DOI: 10.1038/nrc1609.

Ideker T., Ozier O., Schwikowski B., Siegel AF. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics (Oxford, England) 18 Suppl 1:S233-240.

Iozzo RV., Sanderson RD. 2011. Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. Journal of Cellular and Molecular Medicine 15:1013–1031. DOI: 10.1111/j.1582-4934.2010.01236.x.

Kanehisa M., Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28:27–30.
Khanin R., Wit E. 2006. How scale-free are biological networks. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 13:810–818. DOI: 10.1089/cmb.2006.13.810.

Khatri P., Sirota M., Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* 8:e1002375. DOI: 10.1371/journal.pcbi.1002375.

Kim K., Kim KH., Cheong J. 2010. Hepatitis B Virus X Protein Impairs Hepatic Insulin Signaling Through Degradation of IRS1 and Induction of SOCS3. *PLoS ONE* 5. DOI: 10.1371/journal.pone.0008649.

Kimmel C., Visweswaran S. 2013. An algorithm for network-based gene prioritization that encodes knowledge both in nodes and in links. *PloS One* 8:e79564. DOI: 10.1371/journal.pone.0079564.

Klauer AA., van Hoof A. 2012. Degradation of mRNAs that lack a stop codon: A decade of nonstop progress. *Wiley interdisciplinary reviews. RNA* 3:649–660. DOI: 10.1002/wrna.1124.

Knight T., Irving JAE. 2014. Ras/Raf/MEK/ERK Pathway Activation in Childhood Acute Lymphoblastic Leukemia and Its Therapeutic Targeting. *Frontiers in Oncology* 4:160. DOI: 10.3389/fonc.2014.00160.

Kostka D., Spang R. 2004. Finding disease specific alterations in the co-expression of genes. *Bioinformatics (Oxford, England)* 20 Suppl 1:i194-199. DOI: 10.1093/bioinformatics/bth909.

Lai Y., Wu B., Chen L., Zhao H. 2004. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics (Oxford, England)* 20:3146–3155. DOI: 10.1093/bioinformatics/bth379.

Langfelder P., Luo R., Oldham MC., Horvath S. 2011. Is my network module preserved and reproducible? *PLoS computational biology* 7:e1001057. DOI: 10.1371/journal.pcbi.1001057.

Leonardson AS., Zhu J., Chen Y., Wang K., Lamb JR., Reitman M., Emilsson V., Schadt EE. 2010. The effect of food intake on gene expression in human peripheral blood. *Human Molecular Genetics* 19:159–169. DOI: 10.1093/hmg/ddp476.
Li D., Li J., Guo Y. 2014. Genome-wide association study knowledge-driven pathway analysis of alcohol dependence implicates the calcium signaling pathway. Chinese Medical Journal 127:2229–2235.

Lind DS. 2004. Arginine and Cancer. The Journal of Nutrition 134:2837S–2841S.

Lindeboom RGH., Supek F., Lehner B. 2016. The rules and impact of nonsense-mediated mRNA decay in human cancers. Nature genetics 48:1112–1118. DOI: 10.1038/ng.3664.

Liu Z., Tian Y., Machida K., Lai MMC., Luo G., Foung SKH., Ou JJ. 2012. Transient Activation of the PI3K-AKT Pathway by Hepatitis C Virus to Enhance Viral Entry. Journal of Biological Chemistry 287:41922–41930. DOI: 10.1074/jbc.M112.414789.

Liu B-H., Yu H., Tu K., Li C., Li Y-X., Li Y-Y. 2010. DCGL: an R package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics (Oxford, England) 26:2637–2638. DOI: 10.1093/bioinformatics/btq471.

May P., May E. 1999. Twenty years of p53 research: structural and functional aspects of the p53 protein. Oncogene 18:7621–7636. DOI: 10.1038/sj.onc.1203285.

Mitrea C., Taghavi Z., Bokanizad B., Hanoudi S., Tagett R., Donato M., Voichița C., Drăghici S. 2013. Methods and approaches in the topology-based analysis of biological pathways. Frontiers in Physiology 4:278. DOI: 10.3389/fphys.2013.00278.

Molloy NH., Read DE., Gorman AM. 2011. Nerve Growth Factor in Cancer Cell Death and Survival. Cancers 3:510–530. DOI: 10.3390/cancers3010510.

Morrison JL., Breitling R., Higham DJ., Gilbert DR. 2005. GeneRank: using search engine technology for the analysis of microarray experiments. BMC bioinformatics 6:233. DOI: 10.1186/1471-2105-6-233.

Morshed Osmani G., Rahman S. 2018. Using Google’s PageRank Algorithm to Identify Important Attributes of Genes.
O dibat O., Reddy CK. 2012. Ranking differential hubs in gene co-expression networks. *Journal of Bioinformatics and Computational Biology* 10:1240002. DOI: 10.1142/S0219720012400021.

Oehler N., Volz T., Bhadra OD., Kah J., Allweiss L., Giersch K., Bierwolf J., Riecken K., Pollok JM., Lohse AW., Fehse B., Petersen J., Urban S., Lütgehetmann M., Heeren J., Dandri M. 2014. Binding of hepatitis B virus to its cellular receptor alters the expression profile of genes of bile acid metabolism. *Hepatology (Baltimore, Md.)* 60:1483–1493. DOI: 10.1002/hep.27159.

Page L., Brin S., Motwani R., Winograd T. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*.

Pan D. 2010. The Hippo Signaling Pathway in Development and Cancer. *Developmental cell* 19:491–505. DOI: 10.1016/j.devcel.2010.09.011.

Pedroche F., Moreno F., González A., Valencia A. 2013. Leadership groups on Social Network Sites based on Personalized PageRank. *Mathematical and Computer Modelling* 57:1891–1896. DOI: 10.1016/j.mcm.2011.12.026.

Pópulo H., Lopes JM., Soares P. 2012. The mTOR signalling pathway in human cancer. *International Journal of Molecular Sciences* 13:1886–1918. DOI: 10.3390/ijms13021886.

Porta C., Paglino C., Mosca A. 2014. Targeting PI3K/Akt/mTOR Signaling in Cancer. *Frontiers in Oncology* 4:64. DOI: 10.3389/fonc.2014.00064.

Rahmatallah Y., Emmert-Streib F., Glazko G. 2014. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics (Oxford, England)* 30:360–368. DOI: 10.1093/bioinformatics/btt687.

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. DOI: 10.7717/peerj.2568.

Reverter A., Ingham A., Lehnert SA., Tan S-H., Wang Y., Ratnakumar A., Dalrymple BP. 2006. Simultaneous identification of differential gene expression and connectivity in inflammation,
921 adipogenesis and cancer. *Bioinformatics (Oxford, England)* 22:2396–2404. DOI: 10.1093/bioinformatics/btl392.

923 Rhinn H., Fujita R., Qiang L., Cheng R., Lee JH., Abeliovich A. 2013. Integrative genomics identifies APOE ε4 effectors in Alzheimer’s disease. *Nature* 500:45–50. DOI: 10.1038/nature12415.

925 Santarpia L., Lippman SM., El-Naggar AK. 2012. Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets* 16:103–119. DOI: 10.1517/14728222.2011.645805.

928 Sanyal AJ., Yoon SK., Lencioni R. 2010. The etiology of hepatocellular carcinoma and consequences for treatment. *The Oncologist* 15 Suppl 4:14–22. DOI: 10.1634/theoncologist.2010-S4-14.

930 Saucedo LJ., Edgar BA. 2007. Filling out the Hippo pathway. *Nature Reviews Molecular Cell Biology* 8:613–621. DOI: 10.1038/nrm2221.

932 Schmidt M., de Mattos SF., van der Horst A., Klompmaker R., Kops GJPL., Lam EW-F., Burgering BMT., Medema RH. 2002. Cell Cycle Inhibition by FoxO Forkhead Transcription Factors Involves Downregulation of Cyclin D. *Molecular and Cellular Biology* 22:7842–7852. DOI: 10.1128/MCB.22.22.7842-7852.2002.

936 Schork NJ. 1997. Genetics of complex disease: approaches, problems, and solutions. *American Journal of Respiratory and Critical Care Medicine* 156:S103-109. DOI: 10.1164/ajrccm.156.4.12-tac-5.

938 Schreiber G., Walter MR. 2010. Cytokine receptor interactions as drug targets. *Current opinion in chemical biology* 14:511–519. DOI: 10.1016/j.cbpa.2010.06.165.

940 Seshacharyulu P., Ponnusamy MP., Haridas D., Jain M., Ganti A., Batra SK. 2012. Targeting the EGFR signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets* 16:15–31. DOI: 10.1517/14728222.2011.648617.

943 Sherr CJ., McCormick F. 2002. The RB and p53 pathways in cancer. *Cancer Cell* 2:103–112. DOI: 10.1016/S1535-6108(02)00102-2.
Shi Q., Hoffman B., Liu Q. 2016. PI3K-Akt signaling pathway upregulates hepatitis C virus RNA translation through the activation of SREBPs. *Virology* 490:99–108. DOI: 10.1016/j.virol.2016.01.012.

Siddiqui N., Borden KLB. 2012. mRNA export and cancer. *Wiley interdisciplinary reviews. RNA* 3:13–25. DOI: 10.1002/wrna.101.

Simms CL., Yan LL., Zaher HS. 2017. Ribosome Collision Is Critical for Quality Control during No-Go Decay. *Molecular Cell* 68:361–373.e5. DOI: 10.1016/j.molcel.2017.08.019.

Spangler JB., Moraga I., Mendoza JL., Garcia KC. 2015. Insights into Cytokine–Receptor Interactions from Cytokine Engineering. *Annual review of immunology* 33:139–167. DOI: 10.1146/annurev-immunol-032713-120211.

Stegh AH. 2012. Targeting the p53 signaling pathway in cancer therapy – the promises, challenges and perils. *Expert Opinion on Therapeutic Targets* 16:67–83. DOI: 10.1517/14728222.2011.643299.

Stuart JM., Segal E., Koller D., Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)* 302:249–255. DOI: 10.1126/science.1087447.

Subramanian A., Tamayo P., Mootha VK., Mukherjee S., Ebert BL., Gillette MA., Paulovich A., Pomeroy SL., Golub TR., Lander ES., Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102:15545–15550. DOI: 10.1073/pnas.0506580102.

Sui X., Jin L., Huang X., Geng S., He C., Hu X. 2011. p53 signaling and autophagy in cancer: A revolutionary strategy could be developed for cancer treatment. *Autophagy* 7:565–571. DOI: 10.4161/auto.7.6.14073.

Tai Y-L., Chen L-C., Shen T-L. 2015. Emerging roles of focal adhesion kinase in cancer. *BioMed Research International* 2015:690690. DOI: 10.1155/2015/690690.
Tarca AL., Draghici S., Khatri P., Hassan SS., Mittal P., Kim J-S., Kim CJ., Kusanovic JP., Romero R. 2009. A novel signaling pathway impact analysis. *Bioinformatics (Oxford, England)* 25:75–82. DOI: 10.1093/bioinformatics/btn577.

Tavazoie S., Hughes JD., Campbell MJ., Cho RJ., Church GM. 1999. Systematic determination of genetic network architecture. *Nature Genetics* 22:281–285. DOI: 10.1038/10343.

Taylor IW., Linding R., Warde-Farley D., Liu Y., Pesquita C., Faria D., Bull S., Pawson T., Morris Q., Wrana JL. 2009. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology* 27:199–204. DOI: 10.1038/nbt.1522.

Tesson BM., Breitling R., Jansen RC. 2010. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* 11:497. DOI: 10.1186/1471-2105-11-497.

Villanueva A., Chiang DY., Newell P., Peix J., Thung S., Alsinet C., Tovar V., Roayaie S., Minguéz B., Sole M., Battiston C., Van Laarhoven S., Fiel MI., Di Feo A., Hoshida Y., Yea S., Toffanin S., Ramos A., Martignetti JA., Mazzaferro V., Bruix J., Waxman S., Schwartz M., Meyerson M., Friedman SL., Llovet JM. 2008. Pivotal role of mTOR signaling in hepatocellular carcinoma. *Gastroenterology* 135:1972–1983, 1983.e1–11. DOI: 10.1053/j.gastro.2008.08.008.

Vojtek AB., Der CJ. 1998. Increasing complexity of the Ras signaling pathway. *The Journal of Biological Chemistry* 273:19925–19928.

Walker FO. 2007. Huntington’s disease. *The Lancet* 369:218–228. DOI: 10.1016/S0140-6736(07)60111-1.

Wang W., Pan Q., Fuhler GM., Smits R., Peppelenbosch MP. 2017. Action and function of Wnt/β-catenin signaling in the progression from chronic hepatitis C to hepatocellular carcinoma. *Journal of Gastroenterology* 52:419–431. DOI: 10.1007/s00535-016-1299-5.

Wang C., Wang X., Gong G., Ben Q., Qiu W., Chen Y., Li G., Wang L. 2012. Increased risk of hepatocellular carcinoma in patients with diabetes mellitus: a systematic review and meta-analysis of cohort studies. *International Journal of Cancer* 130:1639–1648. DOI: 10.1002/ijc.26165.
Wang R., Zhang W., Deng H., Wang N., Miao Q., Zhao X. 2013. Discover Community Leader in Social Network with PageRank. In: Advances in Swarm Intelligence. Springer, Berlin, Heidelberg, 154–162. DOI: 10.1007/978-3-642-38715-9_19.

Watson M. 2006. CoXpress: differential co-expression in gene expression data. BMC bioinformatics 7:509. DOI: 10.1186/1471-2105-7-509.

Weirauch MT. 2011. Gene Coexpression Networks for the Analysis of DNA Microarray Data. In: Dehmer, Emmert-Streib F, Graber A, Salvador A eds. Applied Statistics for Network Biology. Wiley-VCH Verlag GmbH & Co. KGaA, 215–250. DOI: 10.1002/9783527638079.ch11.

Winter C., Kristiansen G., Kersting S., Roy J., Aust D., Knösel T., Rümmele P., Jahnke B., Henrich V., Rücker F., Niedergethmann M., Weichert W., Bahra M., Schlitt HJ., Settmacher U., Friess H., Büchler M., Saeger H-D., Schroeder M., Pilarsky C., Grützmann R. 2012. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS computational biology 8:e1002511. DOI: 10.1371/journal.pcbi.1002511.

Zeng X., Zhao J., Wu X., Shi H., Liu W., Cui B., Yang L., Ding X., Song P. 2016. PageRank analysis reveals topologically expressed genes correspond to psoriasis and their functions are associated with apoptosis resistance. Molecular Medicine Reports 13:3969–3976. DOI: 10.3892/mmr.2016.4999.

Zhang B., Li H., Riggins RB., Zhan M., Xuan J., Zhang Z., Hoffman EP., Clarke R., Wang Y. 2009. Differential dependency network analysis to identify condition-specific topological changes in biological networks. Bioinformatics (Oxford, England) 25:526–532. DOI: 10.1093/bioinformatics/btn660.

Zhang Y-L., Wang R-C., Cheng K., Ring BZ., Su L. 2017. Roles of Rap1 signaling in tumor cell migration and invasion. Cancer Biology & Medicine 14:90–99. DOI: 10.20892/j.issn.2095-3941.2016.0086.

Zhou L., Huang Y., Li J., Wang Z. 2010. The mTOR pathway is associated with the poor prognosis of human hepatocellular carcinoma. Medical Oncology (Northwood, London, England) 27:255–261. DOI: 10.1007/s12032-009-9201-4.
Zhu X., Gerstein M., Snyder M. 2007. Getting connected: analysis and principles of biological networks. 

Genes & Development 21:1010–1024. DOI: 10.1101/gad.1528707.
Figure 1

Overview of the PoTRA method.
Figure 2

The topological rank analysis for each gene within a pathway.

For genes within a specified pathway, according to Step 2, we construct a corresponding gene co-expression network for normal and cancer, separately. Then we apply the PageRank method to obtain the topological importance of each gene for normal and cancer, separately. $\text{PR}(\text{gene i})_{\text{normal}}$ represents the PageRank score of the gene i for normal samples, while $\text{PR}(\text{gene i})_{\text{cancer}}$ represents the PageRank score of the gene i for cancer samples.
Figure 3

The kernel density distribution of PageRank scores of genes in “MAPK signaling pathway”.

The red line shows the kernel density distribution of PageRank scores for cancer and the black one is for normal samples. Note that the mean for the two distributions is the same, i.e., mean = 1/N=0.004, where N = 250 is the number of genes in the “MAPK signaling pathway” pathway. We use the 95th-percentile cutoff (= 0.006035) of the kernel distribution in normal samples as cutoff for hub genes for both normal and cancer samples.
The 95th percentile cutoff from normal samples is used as a cutoff for hub genes for both normal and cancer samples.
Table 1 (on next page)

Contingency table for Fisher’s exact test.

We use the 95th percentile of the distribution (one-tail) in normal samples as cutoff value for hub genes for both normal and cancer samples. The value “a” represents the number of genes whose PageRank scores are below the cutoff value for normal samples. The value “b” represents the number of genes whose PageRank scores are above the cutoff value for normal samples. The values “c” and “d” are the corresponding values for cancer. We use Fisher’s exact test to assess if the number of hub genes is significantly different between normal and cancer.
|                  | Number of non-hub genes | Number of hub genes | Total  |
|------------------|-------------------------|---------------------|--------|
| Normal           | a                       | b                   | a+b    |
| Cancer           | c                       | d                   | c+d    |
| Total            | a+c                     | b+d                 | a+b+c+d|

**Table 1. Contingency table for Fisher’s exact test.** We use the 95th percentile of the distribution (one-tail) in normal samples as cutoff value for hub genes for both normal and cancer samples. The value “a” represents the number of genes whose PageRank scores are below the cutoff value for normal samples. The value “b” represents the number of genes whose PageRank scores are above the cutoff value for normal samples. The values “c” and “d” are the corresponding values for cancer. We use Fisher’s exact test to assess if the number of hub genes is significantly different between normal and cancer.
Table 2 (on next page)

The “MAPK signaling pathway” pathway identified by PoTRA for HCC using Fisher’s exact test.

The P value is adjusted by False Discovery Rate (FDR).
| Gene Count(L) | # of edges_normal | # of edges_cancer | # of hub genes_normal | # of hub genes_cancer | Adjusted P-value |
|---------------|-------------------|-------------------|-----------------------|-----------------------|-------------------|
| MAPK signaling pathway               | 250               | 14005             | 5170                  | 13                    | 40                | 0.0158            |

Table 2. The “MAPK signaling pathway” pathway identified by PoTRA for HCC using Fisher’s exact test. The P value is adjusted by False Discovery Rate (FDR).
Table 3 (on next page)

The significant KEGG pathways identified by PoTRA for HCC using Fisher’s exact test.

FDR adjusted P-values are below 0.05.
| Pathways in cancer | # of edges | # of hub genes | # of hub genes | Adjusted P-value |
|--------------------|------------|----------------|----------------|------------------|
| 1                   | 310        | 24924          | 9136           | 16               | 48               | 0.0081          |
| MAPK signaling      | 252        | 14005          | 5170           | 13               | 40               | 0.0158          |
| 2 pathway           | 143        | 3589           | 1175           | 8                | 29               | 0.0278          |

**Table 3.** The significant KEGG pathways identified by PoTRA for HCC using Fisher’s exact test. FDR adjusted P-values are below 0.05.
Table 4 (on next page)

The “MAPK signaling pathway” pathway identified by PoTRA for HCC using the Kolmogorov-Smirnov test.

The P value is adjusted by False Discovery Rate (FDR).
Table 4. The “MAPK signaling pathway” pathway identified by PoTRA for HCC using the Kolmogorov–Smirnov test. The P value is adjusted by False Discovery Rate (FDR).

| Pathway          | Gene Count(L) | # of edges_normal | # of edges_cancer | Adjusted P-value |
|------------------|---------------|-------------------|-------------------|------------------|
| MAPK signaling   | 250           | 14005             | 5170              | 0.0278423        |
Table 5 (on next page)

The significant KEGG pathways identified by PoTRA for HCC using the Kolmogorov-Smirnov test.

FDR adjusted P-values are below 0.05.
| Gene Count(L) | # of edges\_normal | # of edges\_cancer | Adjusted P-value |
|--------------|-------------------|-------------------|-----------------|
| 1 RNA transport | 133 | 7168 | 5343 | 6.72E-08 |
| 2 mRNA surveillance pathway | 70 | 1867 | 1252 | 0.023328272 |
| 3 MAPK signaling pathway | 250 | 14005 | 5170 | 0.027842298 |

Table 5. The significant KEGG pathways identified by PoTRA for HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.
Table 6 (on next page)

The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using Fisher’s exact test.

FDR adjusted P-values are below 0.05.
| Pathway                                      | # of genes in normal | # of edges in normal | # of genes in cancer | # of edges in cancer | # of hub genes in normal | # of hub genes in cancer | Adjusted P-value |
|----------------------------------------------|----------------------|----------------------|----------------------|----------------------|-------------------------|-------------------------|-------------------|
| 1 Insulin signaling pathway                  | 139                  | 2692                 | 958                  | 7                    | 34                      | 34                      | 0.0007            |
| 2 Pathways in cancer                         | 310                  | 11194                | 3792                 | 16                   | 52                      | 52                      | 0.0007            |
| 3 Hippo signaling pathway                    | 151                  | 2836                 | 970                  | 8                    | 31                      | 31                      | 0.0072            |
| 4 HTLV-I infection                           | 194                  | 5518                 | 2080                 | 10                   | 35                      | 35                      | 0.0072            |
| Neurotrophin signaling pathway               | 117                  | 2441                 | 895                  | 6                    | 25                      | 25                      | 0.0195            |
| 5 pathway                                    | 144                  | 3410                 | 832                  | 8                    | 28                      | 28                      | 0.0240            |
| 6 mTOR signaling pathway                     | 85                   | 1524                 | 435                  | 5                    | 21                      | 21                      | 0.0353            |
| Epstein-Barr virus                           |                      |                      |                      |                      |                         |                         |                   |
| 7 infection                                  | 134                  | 2708                 | 828                  | 7                    | 25                      | 25                      | 0.0353            |

Table 6. The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using Fisher’s exact test. FDR adjusted P-values are below 0.05.
The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using Fisher’s exact test.

FDR adjusted P-values are below 0.05.
| # | Pathways in cancer | 310 | 22253 | 7791 | 16 | 62 | 2.89E-06 |
|---|-------------------|-----|-------|------|----|----|---------|
| 2 | PI3K-Akt signaling pathway | 340 | 19901 | 6594 | 17 | 65 | 2.89E-06 |
| 3 | MAPK signaling pathway | 252 | 11986 | 4168 | 13 | 47 | 0.0003 |
| 4 | Proteoglycans in cancer | 204 | 9815 | 3642 | 11 | 38 | 0.0033 |
| 5 | Rap1 signaling pathway | 208 | 8294 | 3587 | 11 | 34 | 0.0215 |
| 6 | Adrenergic signaling in cardiomyocytes | 149 | 3594 | 1355 | 8 | 27 | 0.0372 |
| 7 | cAMP signaling pathway | 196 | 5106 | 2493 | 10 | 30 | 0.0372 |
| 8 | Focal adhesion | 203 | 10225 | 4656 | 11 | 32 | 0.0372 |
| 9 | HTLV-I infection | 194 | 9843 | 4030 | 10 | 30 | 0.0372 |
| 10 | Ras signaling pathway | 226 | 10098 | 3931 | 12 | 33 | 0.0376 |
| 11 | FoxO signaling pathway | 126 | 3391 | 1222 | 7 | 24 | 0.0380 |
| 12 | Osteoclast differentiation | 123 | 4418 | 1452 | 7 | 24 | 0.0380 |
| 13 | ErbB signaling pathway | 88 | 2128 | 814 | 5 | 20 | 0.0400 |
| 14 | Axon guidance | 167 | 6203 | 2705 | 9 | 27 | 0.0433 |

Table 7. The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using Fisher’s exact test. FDR adjusted P-values are below 0.05.
Table 8 (on next page)

The significant KEGG pathways identified by PoTRA for alcohol-induced HCC using Fisher’s exact test.

FDR adjusted P-values are below 0.05.
| Pathway                        | # of edges normal | # of edges cancer | # of hub genes normal | # of hub genes cancer | Adjusted P-value |
|-------------------------------|-------------------|-------------------|-----------------------|-----------------------|------------------|
| PI3K-Akt signaling 1 pathway  | 340               | 23928             | 8733                  | 17                    | 0.0006           |
| MAPK signaling 2 pathway      | 252               | 14005             | 5767                  | 13                    | 0.0007           |
| 3 Pathways in cancer          | 310               | 24924             | 10191                 | 16                    | 0.0043           |

Table 8. The significant KEGG pathways identified by PoTRA for alcohol-induced HCC using Fisher’s exact test. FDR adjusted P-values are below 0.05.
Table 9 (on next page)

The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using the Kolmogorov-Smirnov test.

FDR adjusted P-values are below 0.05.
| Gene Count(L) # of edges_normal | # of edges_cancer | Adjusted P-value |
|--------------------------------|-------------------|------------------|
| 1 RNA transport 133 6877 4001 | 3.33E-06          |
| 2 Pathways in cancer 310 22253 | 7791              | 0.0055638        |
| 3 Proteoglycans in cancer 204 9815 | 3642              | 0.0055638        |
| 4 MAPK signaling pathway 250 11986 | 4168              | 0.01535456       |
| PI3K-Akt signaling pathway 340 19901 | 6594              | 0.01975723       |
| 5 HTLV-I infection 194 9843 4030 | 4030              | 0.01975723       |

Table 9. The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.
The significant KEGG pathways identified by PoTRA for alcohol-induced HCC using the Kolmogorov-Smirnov test.

FDR adjusted P-values are below 0.05.
| Gene Count(L) | # of edges_normal | # of edges_cancer | Adjusted P-value |
|--------------|------------------|------------------|------------------|
| 1 RNA transport | 133 | 7168 | 4816 | 2.94E-08 |
| 2 Pathways in cancer | 310 | 24924 | 10191 | 0.0062832 |
| 3 PI3K-Akt signaling pathway | 340 | 23928 | 8733 | 0.0062832 |
| 4 MAPK signaling pathway | 250 | 14005 | 5767 | 0.01050039 |
| 5 mRNA surveillance pathway | 70 | 1867 | 1096 | 0.01166414 |

Table 10. The significant KEGG pathways identified by PoTRA for alcohol-induced HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.
Table 11 (on next page)

The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using the Kolmogorov-Smirnov test.

FDR adjusted P-values are below 0.05.
| Gene Count(L) | # of edges_normal | # of edges_cancer | Adjusted P-value |
|---------------|-------------------|------------------|------------------|
| Arginine and proline metabolism | 50 | 163 | 26 | 0.005412 |
| Glyoxylate and dicarboxylate metabolism | 26 | 216 | 19 | 0.02765948 |
| Primary bile acid biosynthesis | 17 | 62 | 6 | 0.02765948 |
| Insulin signaling pathway | 139 | 2692 | 958 | 0.04057992 |
| Vasopressin-regulated water reabsorption | 22 | 58 | 8 | 0.04057992 |

Table 11. The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using the Kolmogorov–Smirnov test. FDR adjusted P-values are below 0.05.
Table 12 (on next page)

The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using the Fisher’s exact test based on combined networks.

FDR adjusted P-values are below 0.05. E.comb.normal represents the number of edges in the combined network for normal samples, while E.comb.case is for cancer samples, respectively.
| Gene Counts                        | E.comb.normal | E.comb.case | Adjusted P value |
|-----------------------------------|---------------|-------------|------------------|
| Epstein-Barr virus infection      | 85            | 131         | 43               | 0.0103488       |
| p53 signaling pathway             | 68            | 57          | 20               | 0.0103488       |

**Table 12. The significant KEGG pathways identified by PoTRA for hepatitis C-induced HCC using the Fisher’s exact test based on combined networks.** FDR adjusted P-values are below 0.05. E.comb.normal represents the number of edges in the combined network for normal samples, while E.comb.case is for cancer samples, respectively.
**Table 13** (on next page)

The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using the Fisher’s exact test based on combined networks.

FDR adjusted P-values are below 0.05. E.comb.normal represents the number of edges in the combined network for normal samples, while E.comb.case is for cancer samples, respectively.
| Gene Counts | E.comb.normal | E.comb.case | Adjusted P value |
|-------------|---------------|-------------|-----------------|
| Epstein-Barr virus infection | 85 | 84 | 22 | 0.0103488 |

Table 13. The significant KEGG pathways identified by PoTRA for hepatitis B-induced HCC using the Fisher’s exact test based on combined networks. FDR adjusted P-values are below 0.05. E.comb.normal represents the number of edges in the combined network for normal samples, while E.comb.case is for cancer samples, respectively.