TPQCI: A Topology Potential-Based Method to Quantify Functional Influence of Copy Number Variations

Yusong Liu[a,b], Xiufen Ye[a], Xiaohui Zhan[d], Christina Y. Yu[b,e], Jie Zhang[b], Kun Huang[*b,c]

[a] Collage of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, Heilongjiang, 150001, China

[b] Indiana University School of Medicine, Indianapolis, Indiana, 46202, USA

[c] Regenstrief Institute, Indianapolis, Indiana, 46202, USA

[d] Department of Bioinformatics, School of Basic Medicine, Chongqing Medical University, Chongqing, 400016, China

[e] Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, 43210, USA

* Indicate corresponding author

Corresponding author:

Kun Huang

Indiana University School of Medicine, Indianapolis, Indiana, 46202, USA

Tel: +1(317) 278-7722

E-mail: kunhuang@iu.edu

This is the author's manuscript of the article published in final edited form as:

Liu, Y., Ye, X., Zhan, X., Yu, C. Y., Zhang, J., & Huang, K. (2021). TPQCI: A topology potential-based method to quantify functional influence of copy number variations. Methods, 192, 46–56. https://doi.org/10.1016/j.ymeth.2021.04.015

TPQCI: A Topology Potential-based Method to Quantify Functional Influence of Copy Number Variations

Abstract

Copy number variation (CNV) is a major type of chromosomal structural variation that play important roles in many diseases including cancers. Due to genome instability, a large number of CNV events can be detected in diseases such as cancer. Therefore, it is important to identify the functionally important CNVs in diseases, which currently still poses a challenge in genomics. One of the critical steps to solve the problem is to define the influence of CNV. In this paper, we provide a topology potential based method, TPQCI, to quantify this kind of influence by integrating statistics, gene regulatory associations, and biological function information. We used this metric to detect functionally enriched genes on genomic segments with CNV in breast cancer and multiple myeloma and discovered biological functions influenced by CNV. Our results demonstrate that, by using our proposed TPQCI metric, we can detect disease-specific genes that are influenced by CNVs. Source codes of TPQCI are provided in Github (https://github.com/usos/TPQCI).

1 Introduction

Copy number variation (CNV) is a type of chromosomal structural variation in which the number of copies of a particular genomic section varies from one individual to another [1]. Cancer is a heterogeneous disease with various genetic variations, with CNV as a major source. To date, many studies have been carried out to characterize the close relationship between CNVs and human cancers such as breast cancer [2-4], multiple myeloma [5, 6], gastric and colorectal cancers [7, 8] and many others [9-11]. Although remarkable achievements have been made to explore the relationships between CNV and human cancer, detecting how and what these variations affect still presents a difficult challenge [12-14]. Thus, it is of great interest to explore how CNVs promote cancer development in order to increase our understanding of the mechanism of cancer tumorigenesis and development.

Currently, even though numerous studies have taken into account the contribution of CNVs to cancer, there are not many related tools. Lai *et al.* have developed an R package to investigate the relationship between gene expression and CNV [15]. Liu *et al.* have built a multi-omics database for cancer driver gene research and incorporated computational tools to define CNV and methylation drivers [16]. Peng *et al.* proposed the *remMap* method to model the dependence of RNA expression levels on DNA copy numbers through multivariate linear regressions to study the influence of DNA copy number alterations on RNA transcript levels [17]. Most of these current methods only considered the influence of CNV from the angle of gene expression regulation. Actually, functional relationships at the protein level is also an important factor to be considered but have been long ignored when defining CNV influence. Therefore, the understanding of CNV contributions in human cancers are still incomplete and more comparative studies are still necessary.

The biological system is complex, with genes often working together to perform a certain biological function. Protein-protein interaction (PPI) network is an important tool that helps to identify a group of proteins contributing to a particular function. Therefore, identifying cancer related genes based on PPI networks not only considers the interaction of different genes at the protein level but also helps to explain the particular biological functions that contribute to cancer development [18-20].

To achieve the goal of defining the influence of CNV in cancers, various factors should be considered. Genes with different patterns of CNV occurrence (e.g. common recurrent events, low-frequency recurrent events, and rare events) across individuals may imply distinctive relationships among cancers [21]. In other words, the frequency distribution of CNVs in different cancer types is an important factor to characterize this kind of genetic variation [22]. Moreover, given CNVs play roles by affecting gene expression and then inducing the dysregulation of biological functions, investigating the regulatory relationships between CNVs and gene expression levels helps to reveal the potential roles of CNV in cancer. Specifically, traditional statistics such as pairwise correlations are common choices to construct the regulatory maps between CNVs and gene expression [23]. In addition, even though there may not always be direct relationships between CNVs and interaction of different genes at the protein level, interactions between proteins can also help to explain particular biological functions that contribute to cancer development. Thus, to systematically assess the influence of CNV on cancers, all the factors mentioned above should be considered. Since many current studies focus only on the relational estimation between CNVs and gene expression, in order to better assess the influence of CNVs in cancers, new methods taking account of frequency distribution across individuals, impacts of gene expression, and functional relationships between proteins should be developed.

To quantify the influence of CNVs as our goal, we need to analyze relationships across different omics levels and integrate multiple types of data. The tools of network analysis are best suited to solve this problem. Based on these considerations, topology potential is an ideal choice for gauging the downstream influence of CNVs over gene networks. Topology potential is a metric used to determine the essentiality of a node in network, which was first presented by Gan *et al*. [24]. Utilizing multiple characteristics of the nodes, the topology potential metric can integrate information and knowledge beyond topology properties in network analysis. It has been widely used to identify modules in complex networks [25-27]. In bioinformatics, researchers have applied it to find essential proteins in protein-protein interaction (PPI) networks by using this metric [28]. Our previous work also used it to detect gene co-expression modules [29].

In this paper, we proposed a novel metric TPQCI, *Topology Potential-based Quantification of CNV Influence,* to systematically measure the impact of CNVs on cancers through the integration of both molecular data including CNV and gene expression with PPI network data. The effectiveness of this measurement was confirmed by fold enrichment of disease related genes. We separately applied TPQCI to breast cancer and multiple myeloma data to investigate the CNV influence in each cancer type. Two modules containing highly CNV-influenced genes in PPI network were identified for each cancer. Cytoband enrichment and Gene Ontology (GO) enrichment analyses were performed for each module to identify strongly associated cytobands and biological processes which play critical roles in cancer. Moreover, a comprehensive analysis revealed that distinctive CNVs for breast cancer and multiple myeloma were observed and these cytogenetic aberrations were known recurrent genetic aberrations in each cancer. In summary, our method, TPQCI, was able to detect cancer-specific genes influenced by CNV that promote oncogenesis and development of cancer. TPQCI can be further applied to other cancers and diseases. The source codes of calculating TPQCI and detecting functional influenced CNV modules can be obtained from Github (https://github.com/usos/TPQCI).

2 Methods

In this work, we performed PPI network analysis on both breast cancer and multiple myeloma datasets, to establish relationships between CNVs and gene expressions in these diseases. To achieve this goal, we first designed a metric, TPQCI, by using topology potential to quantify the influence of CNVs in each disease. Then, a gene module detection process was performed in PPI network to identify genes strongly influenced by CNVs in both diseases. Enrichment analyses were then carried out to reveal the biological implications of the detected modules.

2.1 Data sources

We used a multiple myeloma and a breast cancer dataset dataset in our analysis. The multiple myeloma data was obtained from the CoMMpass study by the Multiple Myeloma Research Foundation (MMRF) (https://themmrf.org) and the breast cancer dataset was obtained from The Cancer Genome Atlas (TCGA) project distributed by UCSC XENA (https://xenabrowser.net/datapages/) [30]. We extracted RNA-seq data and CNV ratio data from both datasets. Genes related to specific disease were acquired from the DisGeNET database (https://www.disgenet.org/) [31]. The human PPI network which we used was obtained from PINA 2.0 (https://omics.bjcancer.org/pina/) [32].

2.1.1 PPI network

PPI network from PINA 2.0 contained approximately 16,000 nodes and 170,000 interactions. Data from PINA comes from six different databases: IntAct, MINT, BioGRID, DIP, HPRD, and MIPS MPact [32]. In our experiment, we removed self-loops, multiple edges, and isolated nodes inside the network. Genes that were not present in the CNV or RNA-seq datasets were also removed.

2.1.2 Multiple myeloma dataset

The multiple myeloma dataset was obtained from the IA11 version of the MMRF CoMMpass study. These data were generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org). From the README file provided by the MMRF, RNA-seq expression estimates data were extracted from Fastq files using SALMON 0.5.1 based on cDNA fasta file for Ensemble v74 transcript models by the CoMMpass study. The read estimates were normalized by transcripts-per-million (TPM). Copy number estimates were established from the long-insert sequencing results using existing TGen developed tools. CNVs were identified by an analysis of differential clone coverage.

In this research, we used data from the newly diagnosed patients, and there were 657 patients' samples with matched CNV and RNA-seq data. We conducted pre-processing on the RNA-seq data by removing genes whose TPM reads were zero in more than half of the samples. Genes with the lowest 20% of mean values and lowest 10% variance were also be removed. Finally, we logarithmically transformed the RNA-seq expression so that the data closely follow Gaussian distributions. After pre-processing, there were 13,248 protein coding genes contained in the RNA-seq and CNV data, which were used in the following analysis.

We obtained multiple myeloma related genes from the DisGeNET dataset (disease id C0026764). There were 1,311 genes inside this dataset, and 1,017 of them were detected in our processed data. We identified CNV-related genes as genes whose median copy number or interquartile range (IQR) was greater than 0.3. From the 1,017 genes, we identified that 438 genes related to myeloma and CNV. A list of these genes is provided in Supplementary File 1.

2.1.3 Breast Cancer dataset

Breast cancer data was extracted from TCGA dataset. There were 1,098 samples of breast cancer in TCGA, and we chose to use the primary tumor data, resulting in 1,078 samples. CNV ratio data was estimated by GISTIC2. RNA-seq data was normalized by TPM and pre-processed in the same manner as the multiple myeloma data as described above. After pre-processing, 13,247 protein coding genes were used in the following analysis.

We chose the gene set named Breast Carcinoma (Disease id C0678222) in the DisGeNET dataset as disease related genes of breast cancer, which contains 4,962 genes and 1,864 were present in our processed data. We identified 1,275 genes that satisfied the criterion of disease related CNV genes. A list of these genes is provided in Supplementary File 1.

2.2 Quantify the Influence of CNV

As we described in the introduction, to quantify the influence of CNV, we need to quantify three factors:

(1) Frequency of occurring CNVs,
(2) Relationships between CNV and RNA expression, and
(3) Functional relationships among proteins002E

Among these three factors, frequency of occurring CNVs can be easily quantified by statistical analysis, relationships between CNV and RNA expression can be calculated by correlation between gene level CNV log ratio with RNA expression, and functional relationships among proteins can be represented by PPI networks. Since these factors representing the influence of CNV at different levels, it is possible to construct a network to reveal the influence entirely. We call this double weighted (i.e. both edges and nodes weighted) network as *CNV Influence Network* (CIN). For CINs, network framework is PPI network, weight of edges is correlation between gene level CNV log ratio with RNA expression and weight of nodes is frequency of CNV events occurrence. In network analysis field, the centrality of a node measures the importance of a node in the network [33, 34]. Therefore, centrality of nodes in CIN can be treat as the quantification of overall influence of CNV. In this piece of work, to make a better tradeoff between the three factors, we employed topology potential as the metric of centrality in CIN and we call this metric as *Topology Potential based quantification of CNV Influence* (TPQCI). The workflow of calculating TPQCI is shown in Figure 1. The source codes of calculating TPQCI are provided in Github (https://github.com/usos/TPQCI).

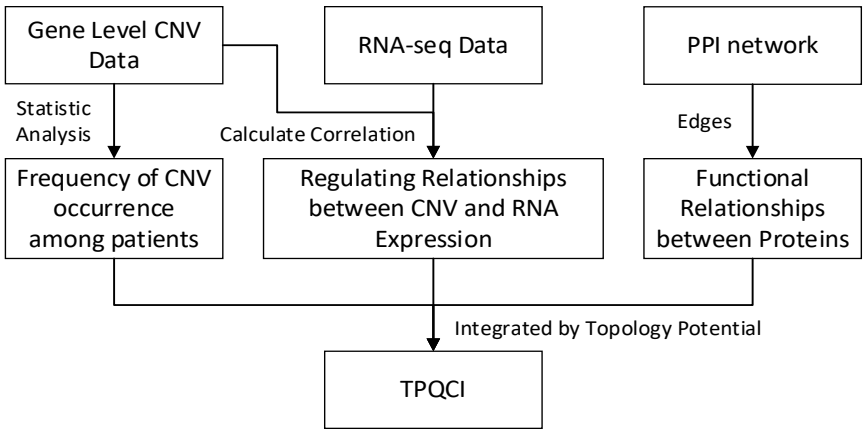

Figure 1. The workflow for calculating TPQCI.

2.2.1 Calculating TPQCI

For each CNV log ratio, if the ratio of a gene is greater than 0.3 or less than -0.3, we consider CNV to have occurred as an amplification or deletion, respectively. Based on this criterion, we can calculate the frequency of a CNV occurrence. Correlation between CNV ratio and gene expression was calculated by the Pearson correlation coefficient (PCC). Since the PPI network defined functional associations between genes, we selected the first-order neighborhood of a CNV gene as its scope of influence.

Since we use correlation between CNV ratio and gene expression to depict regulatory associations, the edges between genes is directed. We used out-degree topology potential to integrate the three factors. Topology potential is a metric of centralization, which is used to describe the interaction and association among network nodes. For directed weighted networks $G(V, E, M, W)$, where V is the node set, E is the set of directed edges, M is the set of nodes' properties, and W is the set of weights of the directed edges. Let the size of node set be $N = |V|$. Out-degree topology potential φ_{out} of any node $v_i \in V$ can be determined by the formula below [35]:

$$\varphi_{out}(v_i) = \sum_{j=1}^{N} m_i \exp\left(-\left(\frac{d_{w_{i\to j}}}{\sigma}\right)^2\right). \quad (1)$$

In the Eq. (1), $m_i, m_j \in M$ are both greater than 0 and $d_{w_{i\to j}}$ is the distance of node i to j under the influence of the weight of edge. σ is a parameter to control influence range of each node. The multiplication of the node's properties (e.g., CNV frequency) with the exponential term related to the edge weight (contained in $d_{w_{i\to j}}$) effectively integrates the two types of information without a direct trade-off. According the property of Gaussian function, if $d_{w_{i\to j}}$ is greater than $\frac{3\sigma}{\sqrt{2}}$, φ_{out} will quickly decay to 0 [36]. This property of topology potential in essence amplifies the influence of strong relationships and suppress the weak ones, which makes our proposed TPQCI metric reflect the functional influence more accurately.

For the application of this work, we can simplify and specific the definition of φ_{out}. Since we only consider the first-order neighborhood of CNV genes, we only calculate the topology potential component of the CNV gene itself and its first-order neighbors. In addition, we use PCC (i.e., $\rho(c_i, r_j)$) between copy number ratio of the CNV gene $v_i(c_i)$ and expression of its neighbor $v_j(r_j)$ as the edge weights in order to define $d_{w_{i\to j}}$:

$$d_{w_{i\to j}} = \frac{1}{|\rho(c_i, r_j)|} - 1 \quad (2)$$

Furthermore, for two variables, we consider them have linear relationship if the absolute value of PCC is greater than a specific threshold. In biological analysis, the empirical value of the threshold is often set around 0.3. Therefore, according to the property of σ we mentioned previously and Eq. (2), we can set the parameter σ to 1.10. Let r_{c_i} represent the frequency of CNV occurring on gene i, and $G(V, E)$ be PPI network, TPQCI of gene i can be calculated by:

$$TPQCI(i) = \sum_{j\in V} r_{c_i} \exp\left(-\left(\frac{1/|\rho(c_i, r_j)| - 1}{1.10}\right)^2\right) \cdot \delta(i, j) \quad (3)$$

In Eq. (3), $\delta(i,j)$ is an indicator function. Let $U_1(i) \subset V$ be the first order neighborhood of gene i in the PPI network G, $\delta(i,j)$ can be expressed as:

$$\delta(i,j) = \begin{cases} 1 & j \in U_1(i), \\ 0 & otherwise. \end{cases} \quad (4)$$

2.2.2 Verification of TPQCI

Since different diseases may be affected by different CNVs, we want to ensure that our results are disease specific. Specifically, when all the genes are sorted by the TPQCI metric, we expect to see known disease related genes ranked higher in each experiment. Therefore, we performed functional verification of TPQCI based on the two assumptions below:

(1) Correlation of the TPQCI metric between the breast cancer and multiple myeloma datasets were employed to test if TPQCI is disease specific. For samples from different disease, if their TPQCI is correlated, it will suggest that the type of disease is not the main factor affecting TPQCI. We use PCC to test the correlation between TPQCI from multiple myeloma and breast cancer.

(2) We used fold enrichment to verify that our proposed TPQCI metric can cover more disease related genes than just using the degree of the PPI network. Fold enrichment is a statistical concept describing how many folds more did something happen than we would expect by random chance [37]. Let G be the set of all nodes inside a network, $G_n \subset G$ is the set of top n nodes sorted by some metric, $V \subset G$ is the set of verified reference nodes and $V_n = V \cap G_n$. Fold enrichment f_n of top n genes is defined below according to [38]:

$$f_n = \frac{|V_n|/|V|}{|G_n|/|G|} \quad (5)$$

In the equation above, the notation $|\cdot|$ denotes the number of elements inside a set. it is easy to find that f_n will convergence to 1 while n is increasing to $|G|$. For two different ranking metrics, since we prefer the metric that rank disease related genes to be on the top, the greater f_n when n is small generally implies better performance on a verified reference set. In our work, G is the PPI network and V is the set of disease related CNV genes.

2.3 Detect Modules Influenced in PPI Networks

Once we defined the metric of CNV influence (TPQCI), we can use it to identify gene modules that are heavily affected by CNVs in a PPI network. Our module search method is based on the concept of attraction in topology potential. For any node $v \in V$, if there is a path leading to a representative node $v^* \in V$ and the topology potential of every node on the path increases in turn, then v is said to be attracted by v^*. Such a path is named as an *attraction chain* [36].

It is natural that genes with weak influence will be attracted by stronger ones. So, if they are attracted in a direct path, this pathway may contribute to some biological process. We proposed a module detection method based on this assumption. Pseudocode of our module detection method is displayed in Algorithm 1.

We first identify genes with local maximal TPQCI values in the whole network. To identify the local maximal nodes, we first randomly select node in the network as a seed and find all its neighbors. The seed node and its neighbors will be marked as visited. If the seed node has the maximum topology potential among its neighborhood, we consider this seed node a local maximal node. Otherwise, the node with the

maximum topology potential will become new seed and we repeat the above operations until we find a local maximal node. After identifying a local maximal node, we will select an unvisited node as a new seed and seek other local maximal nodes until all nodes are visited. Let N be the number of nodes of the PPI network and n be average neighborhood size of each node. The average time complexity of the local maximal nodes search process should be $O(Nn)$. Since PPI network is a scale-free network, we have $n \ll N$ and the average time complexity can be expressed as $O(N)$ approximately [39]. Line 1-14 of Algorithm 1 describe this process. Then, we find attraction chains started by all the local maximal nodes and generate modules. Breadth-first search (BFS) method is used to finish this work. BFS is a commonly used approach for traversing or searching tree or graph data structures. For graphs, it starts at some arbitrary node and explores all of the neighbor nodes at the present depth prior to moving on to the nodes at the next depth level [40]. Line 15-21 of Algorithm 1 describe our proposed module detection process.

Algorithm 1 Detecting CNV influenced gene modules in PPI network

Input: Node weighted PPI network $G(V, E, M)$, in which V is the set of gene, E is the set of edges between genes in PPI network, and M is TPQCI for each gene in PPI network;
Global TPQCI threshold τ_g;
Local TPQCI threshold τ_l;
Maximum overlap between modules β;
Minimum module size μ

Output: Merged Detected Modules U

Algorithm:
#***Find local maximum nodes lm***
1: Sort all nodes V by M in decreasing order
2: let $n = |V|$
3: initialize a length n all zero vector f_1, $lm = \emptyset$
4: **while** ($sum\ (f_1) \neq n$)
5: select a node v randomly
6: **while** ($f_1[v] == 1$)
7: select another node v randomly
8: **end while**
9: find neighborhood ne_v of node v
10: find node v_{top} in ne_v with the max TPQCI
11: **while** ($v_{top} \neq v$)
12: $f_1[v] = 1$
13: $v = v_{top}$
14: find neighborhood ne_v of node v
15: find node v_{top} in ne_v with the max TPQCI
11: **end while**
12: $lm = lm \cup \{v\}$
13: $f_1[ne_v] = 1$
14: **end while**
Generate modules U
15: get maximum TPQCI tp_{max} in lm
16: $lm = \{v \mid v \in lm \text{ and } TPQCI > tp_{max} * \tau_g\}$
17: initialize a length $|lm|$ vector $U = \emptyset$
18: **foreach** ($v \in lm$)

19: $m_v = \{\text{BFS search nodes satisfy attract chain and } TPQCI > v * \tau_l\}$
20: $U = U \cup \{m_v\}$
21: **end foreach**
Merge modules
22: $U = \{m \mid m \in U \text{ and } size > \mu\}$
23: Merge modules with highly overlap in U respect to β
24: Output U

To limit the number and size of detected modules, we set two parameters, global TPQCI threshold τ_g and local TPQCI threshold τ_l, to limit small TPQCI values of local maximum nodes and nodes inside modules. Both τ_g and τ_l are ratio thresholds, which range between 0-1. τ_g limits the TPQCI of local maximum nodes for initialization of new modules, to keep TPQCI of local maximum nodes used to generate modules greater than τ_g times of global maximum TPQCI. This parameter limits the number of modules detected directly. The other parameter, τ_l, limits module size directly. τ_l ensures that the TPQCI of each node inside a module should be greater than the τ_g times of maximum TPQCI of that module. Unless otherwise specified, we set $\tau_g = 0.05, \tau_l = 0.2$ for multiple myeloma dataset and $\tau_g = 0.15, \tau_l = 0.3$ for breast cancer dataset. In addition, modules that overlap more than a parameter β will be merged to enhance the independence between modules. For two sets M_1 and M_2, Overlap between them can be calculated by:

$$Overlap(M_1, M_2) = \frac{|M_1 \cap M_2|}{\min(|M_1|, |M_2|)} \tag{6}$$

For the convenience of downstream analyses, we can also limit the minimum size of modules by parameter μ. For both diseases, we let $\beta = 0.5$ and $\mu = 20$ in this paper. The source codes for detecting functional influenced modules are provided in Github (https://github.com/usos/TPQCI). We will discuss the influence of these parameters in detail in the Discussion section.

2.4 Biological Analysis of Influenced Modules

To explore the biological basis of various modules, we performed cytoband enrichment analysis and functional Gene Ontology (GO) enrichment analysis for module genes of each cancer type using ToppGene (https://toppgene.cchmc.org) [41]. The Fisher's exact test was used to calculate p-values for gene set enrichment and the Benjamini-Hochberg false discovery rate (BH FDR) was used to calculate q-values for multiple test compensation. Only cytobands and GO terms with q-values less than 0.05 were considered significantly enriched.

3 Results

3.1 Performance of TPQCI

We calculated TPQCI on our multiple myeloma and breast cancer datasets. All samples that meet the criteria of data preprocessing were used to evaluate TPQCI metric. In the multiple myeloma dataset, 33 genes having a TPQCI greater than 1 including well known multiple myeloma related genes *ICAM1* (aka *CD54*) and *CSNK1A1* (aka *CK1α*) [42]. In the breast cancer dataset, 48 genes having a TPQCI greater than 1 including major cancer drivers such as *MYC* and *TP53* while the top gene *COPS5* are also known to be associated with breast cancer development [43]. We list the TPQCI and CNV frequency of top 10 genes in

each disease in Table 1 and all genes in Supplementary File 2. The cumulative distribution of TPQCI in multiple myeloma and breast cancer is displayed in Table 2. The PCC between TPQCI in multiple myeloma and breast cancer was 0.123 ($p < 0.001$), indicating that there was no linear correlation between TPQCI in multiple myeloma and breast cancer.

Table 1 TPQCI and frequency of CNV happening of top 10 genes in multiple myeloma and breast cancer

	Multiple Myeloma			Breast Cancer		
	Gene Name	**CNV Freq.**	**TPQCI**	**Gene Name**	**CNV Freq.**	**TPQCI**
1	*RPL4*	0.520	2.325	*COPS5*	0.487	3.080
2	*RPS3*	0.402	1.979	*YWHAZ*	0.557	2.543
3	*RPS25*	0.419	1.964	*ZC3H18*	0.607	2.243
4	*ICAM1*	0.589	1.874	*UBC*	0.235	2.232
5	*RIOK2*	0.444	1.758	*RPL7*	0.494	2.207
6	*RPL7A*	0.558	1.694	*TERF2*	0.586	2.109
7	*RPS28*	0.579	1.656	*PABPC1*	0.555	2.080
8	*ADRBK1*	0.361	1.650	*MYC*	0.574	1.941
9	*CSNK1A1*	0.453	1.641	*TP53*	0.505	1.870
10	*RPL18A*	0.572	1.627	*IKBKE*	0.665	1.812

Table 2 Cumulative distribution of TPQCI in multiple myeloma and breast cancer.

Value of TPQCI	**Number of Genes**	
	Multiple myeloma	**Breast cancer**
≥ 1	33	48
≥ 0.1	396	3628
≥ 0.01	2313	6419
≥ 0.001	4150	7488
Total	13248	13247

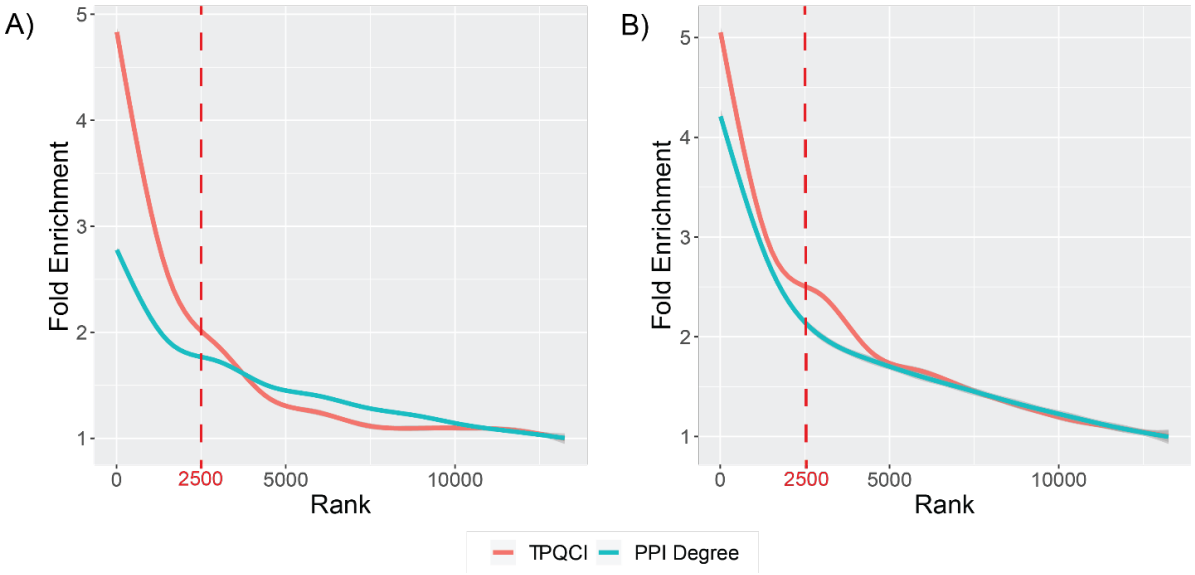

Figure 2 Comparison of fold enrichment of genes ranked by TPQCI and PPI network degree for (a)breast cancer and (b)multiple myeloma.

Fold enrichment was employed to test TPQCI's ability of cover disease related CNV genes. We used our proposed TPQCI and degree of PPI networks as a metric to rank all the genes and compared their fold enrichment in disease related CNV genes. The result is displayed in Figure 2. In both datasets, if rank is higher than 2,500, TPQCI revealed better coverage on disease related CNV genes than degree of PPI networks. 210 (47.9%) of multiple myeloma related and 485 (38.0%) of breast cancer related CNV genes were identified by the queue sorted by TPQCI. These observations confirm that TPQCI has a better performance in ranking disease-related genes to the top of the list.

3.2 Tuning τ_g and τ_l in Module Detect

There are four parameters in our proposed module detection method. Among of them, τ_g and τ_l are the ones that have a large effect on the result, and we tried to tune them in our datasets.

τ_g limits TPQCI of local maximum nodes which affects the number of modules detected directly. Figure 3 shows how τ_g affects the number of modules detected before merging modules. When $\tau_g = 0$ (i.e. no limit on TPQCI for local maximum nodes), the number of modules detected is equal to the number of nodes with local maximum TPQCI in PPI network. As the parameters increase, the number of modules detected decreases rapidly and finally turns into 1. Compared with the breast cancer dataset, the number of modules detected in the multiple myeloma dataset decreased faster, meaning τ_g should be smaller for detecting modules in the multiple myeloma dataset. Therefore, we choose τ_g as 0.05 for the multiple myeloma dataset and 0.15 for the breast cancer dataset.

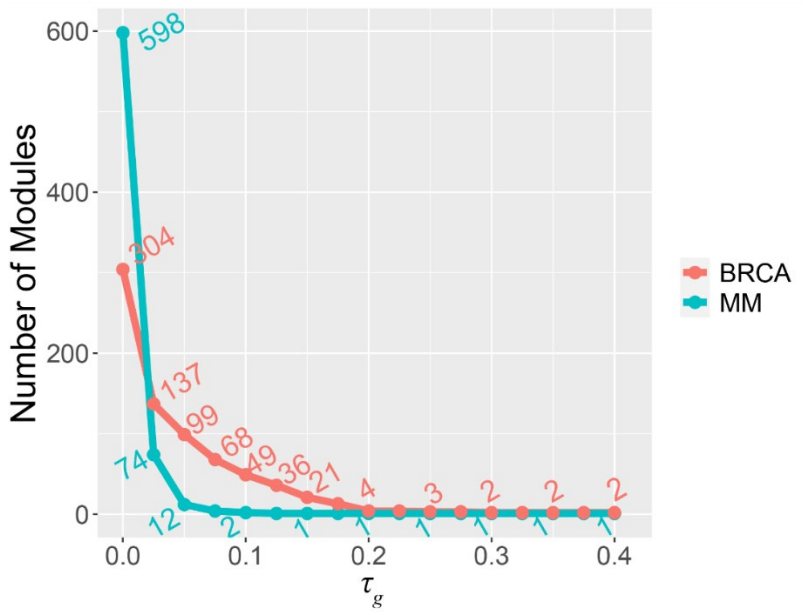

Figure 3 Number of modules detected (before module merge) on both breast cancer and multiple myeloma datasets while setting different τ_g. In this figure, BRCA denotes breast cancer and MM denotes multiple myeloma. Since the number of modules detected barely change when τ_g is greater than 0.4, we only display results for τ_g ranging from [0, 0.4].

Compared with τ_g, τ_l limits the size of modules detected before merging modules by limiting the TPQCI value of nodes inside a module. Figure 4 demonstrates the influence of τ_l on the sizes of detected

modules. While $\tau_l = 0$ (i.e. no limit on TPQCI of nodes inside modules), the size of all modules detected was nearly the whole PPI network. However, sizes of modules plummet with a small increase in τ_l. While the parameter increased further, size of modules decreased smoothly and finally became 1 when $\tau_l = 1$. Ignoring the circumstance of $\tau_l = 0$, modules detected in the breast cancer dataset was generally larger than the ones detected in the multiple myeloma dataset, which is similar to the situation of τ_g. Thus τ_l should also be set smaller while detecting modules in the multiple myeloma dataset. τ_l was set to 0.2 for multiple myeloma and 0.3 for breast cancer for our analysis.

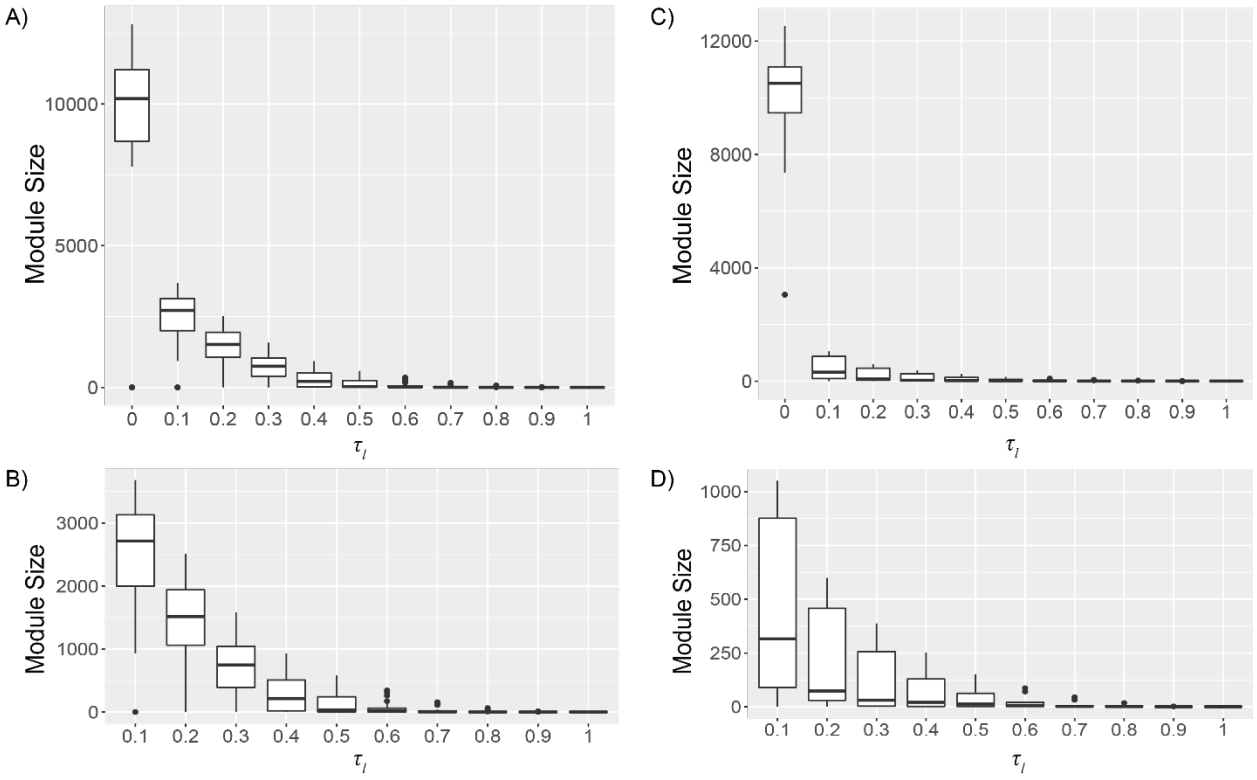

Figure 4 Boxplot of module sizes detected (before module merge) in both (a)-(b) breast cancer and (c)-(d) multiple myeloma datasets while setting different τ_l. Subfigure (b) and (d) are enlargements of (a) and (c) while τ_l was set to 0.1 to 1. In this analysis, τ_g was set to 0.15 for breast cancer dataset and 0.05 for multiple myeloma dataset.

For the convenience of downstream analysis, our module detection method also includes two other parameters, β and μ. Parameter β limits the maximum overlap between modules. Any two modules whose overlap ratio is greater than β will be merged. μ limits the minimum size of a module. Any module whose size is less than μ will be bypassed from subsequent analysis. These two parameters do not affect the module detection process directly, but they are necessary for downstream analysis. We set $\beta = 0.5$ and $\mu = 20$ in this paper. After merging modules, we obtained two modules for each dataset. Details of the modules are provided in Supplementary File 3.

3.3 Biological Analysis of the Module Detection Results

We calculated TPQCI for multiple myeloma and breast cancer datasets separately to estimate how CNVs contribute to cancers. Two copy number influence modules were identified for each cancer type. To explore the biological basis of these modules, cytoband enrichment analysis and Gene Ontology (GO) enrichment analysis were performed based on the genes of each module.

In module 1 of multiple myeloma, cytobands such as 9q34, 11p, 11q, 15q22-25, 19p13 and 19q13 were significantly enriched (Table 3). Metabolic related biological processes for protein synthesis, protein localization, and immune related biological processes were highly enriched (Table 4). In module 2, chromosome such as 1q, 5q, 9q, 11p, 11q, 15q15, 19p13, and 19q12-13 were significantly enriched (Table 3). Moreover, metabolic, cell cycle, and cell death related biological processes were highly enriched (Table 4). Taken together, we observed that chromosome 9q34, 11q13-14, 19p13 and 19q13 were significantly enriched in both module 1 and module 2, although the genes enriched in those locations were different (Supplementary File 3). Furthermore, enriched cytobands specific to each module were also identified. This suggests that our method was able to identify CNV-influenced genes that contribute to different biological functions in myeloma development.

Table 3 Significantly enriched cytobands in multiple myeloma modules

Module (Size)	Location	#Genes enriched	*p*-Value	FDR B&H
Module 1 (57)	5q31-q33	4	7.51E-03	2.15E-02
	9q34	3	1.62E-03	1.39E-02
	11p12	2	3.48E-03	1.41E-02
	11p15	3	4.31E-04	9.26E-03
	11q13-q22	6	1.01E-03	1.39E-02
	15q22-q25	4	3.23E-03	1.41E-02
	19p13	5	2.30E-04	9.26E-03
	19q13	5	3.60E-03	1.41E-02
Module 2 (667)	1q21	14	3.30E-09	5.46E-07
	1q31	5	2.49E-05	1.22E-03
	5q12-q13	2	2.19E-03	3.40E-02
	5q21-q22	2	2.19E-03	3.40E-02
	9q31	4	2.71E-04	8.00E-03
	9q34	10	7.69E-08	4.54E-06
	11p11-p12	2	3.60E-03	4.62E-02
	11p14-p15	17	2.72E-08	2.01E-06
	11q12-q14	45	1.11E-08	1.09E-06
	11q22	2	1.11E-03	2.18E-02
	13q34	6	1.06E-03	2.18E-02
	15q15	5	3.02E-03	4.24E-02
	19p13	35	3.70E-09	5.46E-07
	19q12-q13	7	6.41E-04	1.72E-02
	Xq28	13	1.15E-04	3.76E-03

Table 4 Top 10 Significantly enriched biological processes in multiple myeloma modules

Module (Size)	ID	Name	#Genes enriched	*p*-Value	FDR B&H
Module 1 (57)	GO:0006614	SRP-dependent co-translational protein targeting to membrane	23	1.08E-38	2.95E-35
	GO:0006613	Co-translational protein targeting to membrane	23	2.81E-38	3.84E-35

	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	23	3.22E-37	2.20E-34
	GO:0045047	protein targeting to ER	23	3.22E-37	2.20E-34
	GO:0072599	establishment of protein localization to endoplasmic reticulum	23	7.35E-37	4.02E-34
	GO:0006413	translational initiation	25	1.03E-35	4.69E-33
	GO:0070972	protein localization to endoplasmic reticulum	23	7.09E-35	2.77E-32
	GO:0016032	viral process	35	6.29E-34	2.15E-31
	GO:0019083	viral transcription	24	7.86E-34	2.39E-31
	GO:0019080	viral gene expression	24	4.42E-33	1.21E-30
Module 2 (667)	GO:0051130	positive regulation of cellular component organization	109	2.65E-18	1.04E-14
	GO:0051726	regulation of cell cycle	106	3.00E-18	1.04E-14
	GO:0007049	cell cycle	137	2.32E-17	5.38E-14
	GO:0034622	cellular protein-containing complex assembly	99	1.58E-15	2.74E-12
	GO:0016071	mRNA metabolic process	78	7.48E-15	1.04E-11
	GO:0051276	chromosome organization	97	1.96E-14	2.26E-11
	GO:0042981	regulation of apoptotic process	120	3.11E-14	3.08E-11
	GO:0051338	regulation of transferase activity	86	5.45E-14	4.73E-11
	GO:0043067	regulation of programmed cell death	120	1.09E-13	8.42E-11
	GO:0010941	regulation of cell death	126	2.84E-13	1.97E-10

Similar to multiple myeloma, two copy number influence modules for breast cancer were identified. In module 1, cytobands for 1q21-25, 8q21-24 and 16q (Table 5) were significantly enriched. Metabolic related biological processes were significantly enriched (Table 6). In module 2, cytobands such as 16p13, 16q, 17p13, 17q and 22q (Table 5) were significant enriched. Again, metabolic related biological processes were significantly enriched (Table 6). These observations suggest that in module 1 and module 2, different CNV-influenced genes may affect the same biological processes to drive breast cancer development.

Table 5 Significantly enriched cytobands in breast cancer modules

Module (Size)	Location	#Genes enriched	*p*-Value	FDR B&H
Module 1 (130)	1q21	10	8.56E-05	1.39E-03
	1q23	2	8.09E-03	2.73E-02
	1q25	3	6.05E-03	2.58E-02
	1q31-q42	1	7.49E-03	2.64E-02
	6q25-q27	1	3.77E-03	1.79E-02
	8q11-q13	3	3.07E-03	1.79E-02
	8q21-q24	32	2.91E-08	2.35E-06
	16q13-q21	1	1.49E-02	4.32E-02
	16q22-24	8	2.62E-07	7.06E-06
	17p13	2	8.53E-03	2.76E-02
	17q12	3	7.32E-03	2.64E-02
	22q12-q13	1	1.49E-02	4.32E-02

Module 2 (2160)	1p36	9	5.11E-04	2.05E-02
	1q21	19	3.54E-06	5.13E-04
	1q42	6	5.43E-04	2.07E-02
	3p21	11	1.03E-03	2.85E-02
	5q31	10	4.65E-05	3.74E-03
	6p21	17	9.27E-06	9.57E-04
	6q21	21	1.15E-06	2.08E-04
	8p11	6	1.09E-04	6.06E-03
	8p21	14	1.02E-03	2.85E-02
	8q11	3	2.16E-03	4.96E-02
	8q24	19	3.09E-04	1.49E-02
	11q23	13	1.71E-03	4.27E-02
	13q14	6	2.26E-03	4.96E-02
	16p13	43	2.34E-09	1.69E-06
	16q22	19	7.20E-05	4.73E-03
	16q24	10	2.10E-03	4.96E-02
	17p13	26	5.73E-04	2.07E-02
	17q11	19	7.20E-05	4.73E-03
	17q21-q23	54	1.78E-07	4.29E-05
	17q25	51	2.28E-08	8.25E-06
	20p13	17	1.69E-04	8.74E-03
	20q13	7	1.61E-03	4.16E-02
	22q12-q13	37	3.67E-04	1.56E-02

Table 6 Top 10 Significantly enriched biological processes in breast cancer modules

Module (Size)	**ID**	**Name**	**#Genes enriched**	***p*-Value**	**FDR B&H**
Module 1 (130)	GO:0044265	cellular macromolecule catabolic process	39	1.15E-17	4.22E-14
	GO:0009057	macromolecule catabolic process	41	1.88E-16	3.45E-13
	GO:0016032	viral process	30	1.78E-14	2.18E-11
	GO:0070647	protein modification by small protein conjugation or removal	34	2.64E-14	2.42E-11
	GO:0044403	symbiotic process	30	9.90E-14	7.27E-11
	GO:0072594	establishment of protein localization to organelle	25	1.39E-13	8.51E-11
	GO:0006511	ubiquitin-dependent protein catabolic process	25	3.39E-13	1.70E-10
	GO:0044419	interspecies interaction between organisms	30	3.82E-13	1.70E-10
	GO:0019941	modification-dependent protein catabolic process	25	4.17E-13	1.70E-10
	GO:0016071	mRNA metabolic process	29	4.88E-13	1.79E-10
Module 2 (2160)	GO:0070647	protein modification by small protein conjugation or removal	334	1.91E-73	1.76E-69

	GO:0016071	mRNA metabolic process	276	1.63E-64	7.53E-61
	GO:0032446	protein modification by small protein conjugation	279	1.61E-62	4.95E-59
	GO:0044265	cellular macromolecule catabolic process	321	8.79E-60	2.03E-56
	GO:0016567	protein ubiquitination	256	7.46E-58	1.38E-54
	GO:0009057	macromolecule catabolic process	355	6.55E-56	1.01E-52
	GO:0043632	modification-dependent macromolecule catabolic process	194	1.24E-42	1.63E-39
	GO:0006511	ubiquitin-dependent protein catabolic process	190	3.08E-42	3.32E-39
	GO:0007049	cell cycle	402	3.24E-42	3.32E-39
	GO:0009894	regulation of catabolic process	259	1.26E-41	1.14E-38

4 Discussion

4.1 TPQCI is an Effective Metric to Quantify CNV Influence

TPQCI is a metric to quantify CNV influence that integrates frequency of CNV occurrence, relationships between CNV and gene expression, and protein-protein interactions. The fold enrichment analysis showed that our TPQCI metric was disease specific. The results suggest that, by integrating information from various levels of biological relationships, TPQCI may be able to detect potential driver CNVs in diseases. Fold enrichment analysis reflects the ability of a metric to represent known disease-related genes. From Figure 2(a), we found that more breast cancer related CNV genes were ranked higher by our TPQCI metric. In Figure 2(b), though not as good as in breast cancer, we observed a similar result in multiple myeloma. This indicates that, compared with PPI degree, the TPQCI metric can identify more disease related genes that are influence by CNVs. Based on the cumulative distributions of TPQCI in both cancers (Table 2), we also found that there were only a few genes with high TPQCI in both diseases. This is also consistent with the fact that different cancers are driven by different genetic factors.

In addition, in both the breast cancer and multiple myeloma datasets, the top 2,500 TPQCI ranked genes cannot cover most of our selected disease related CNV genes. A potential reason is that even though these genes are disease related CNV genes, they are not drivers and therefore do not have a high TPQCI. Another possible reason is the way we define the CNV event. In this paper, we applied a simple hard threshold on log ratio (±0.3) to select amplification or deletion event to demonstrate the process without loss of generality. However, such simplified designation of CNV events may result in a tradeoff between sensitivity and precision [44]. At present, more sophisticated methods such as iCopyDAV [45], CNV_IFTV [46], and CONDEL [47] have been developed to tackle such drawbacks. It can be anticipated that incorporation of these methods instead of using the log ratio threshold have potential to improve the coverage over disease related CNV genes. Nevertheless, TPQCI is an effective framework for integrating CNV influence in the network analysis.

From the result in Table 2, we observed that the TPQCI values for most of genes in both diseases studied were less than 1 (99.64% in breast cancer and 99.75% in multiple myeloma). But there were a few genes with TPQCI values greater than 1. Some of these genes with greater TPQCI were reported to be disease related (such as *TP53* in breast cancer and *ICAM1* in multiple myeloma) and have high abnormal CNV ratio. Therefore, we perceive that TPQCI being greater than 1 is a reasonable indicator that the CNV has great

functional influence on the datasets we used. However, for different datasets and diseases, there may have different criteria to determine significant relationship between CNVs and functions. In general, a greater TPQCI implies a stronger functional influence of a CNV.

4.2 Effects of Parameters in Module Detection

While detecting modules using our proposed TPQCI metric, we introduced four parameters, τ_g, τ_l, β and μ. From Figure 3 and Figure 4, we can deduce that our module detection method is sensitive to τ_g and τ_l. When the two parameters increase from zero, the number and size of modules drastically change. This phenomenon is mainly caused by the distribution of TPQCI. Table 2 illustrated that a great number of genes only have a small TPQCI, but they can still become genes inside modules and even local maximum genes. Therefore, when we impose a threshold on TPQCI, these types of genes will not be able to generate or join a module, leading to the drastic change. In addition, genes with lower TPQCI values are unlikely to be functional CNV related genes based on our previous analysis, thus it is important to set a threshold to filter out the genes with low TPQCI values.

In Figure 4, we found that when $\tau_l = 0$, the size of almost all modules increases to cover nearly the entire PPI network. Since PPI network is a scale-free network, there are some hub genes (e.g., *UBC*) that have extremely high degree, and leads to a high possibility that these kinds of genes are located in the attraction chain of most local maximum genes. If there is no limitation (i.e. $\tau_l = 0$), the huge amount of genes connected to hub genes will also be included, and result in modules nearly covering the entire PPI network. When a τ_l threshold is imposed, the modules we detect are subsets of modules detected with no limitation. In general, for the same dataset, modules detected by greater τ_l are also subsets of ones detected by smaller τ_l. In addition, due to the existence of hub genes, modules whose local maximum genes have similar TPQCI may have high overlap as they may both connect to the hub genes. For downstream analysis, high level overlap suggests similar functions. Therefore, it is necessary to merge the highly overlapped modules.

Since we constrained TPQCI while detecting modules by τ_g and τ_l, we can find that there were some modules whose size is very small. These modules are difficult for further downstream analysis due to the lack of consensus information. Thus, we removed this kind of modules by introducing parameter μ.

4.3 Biological Analysis

Based on our TPQCI method, two modules were identified for multiple myeloma and breast cancer, respectively. Significantly enriched cytobands and biological processes for each module were identified. Distinctive CNVs for different cancer types were observed and most of these CNVs are known to be associated with their respective disease.

In breast cancer, multiple cytobands such as chromosomes 1q, 6p21, 6q21, 8p, 8q, 11q23, 16p13, 16q, 17p13, 17q and 22q were enriched. The CNVs in most of these chromosome locations are closely associated with breast cancer [48-53]. CNVs on chromosomes 1q, 8p, 8q, 16q and 17q are prominent features in breast cancer [48, 49]. Chen *et al.* have reported that karyotypic changes at 1q are very frequent in breast cancer and 1q21 could contribute to the initiation of the disease [50]. Many studies have also indicated that amplification of chromosomal region 8q21-q24 is associated with advanced tumors and poor prognosis in breast cancer [51, 52]. The loss of chromosome 16q has been addressed as a key influencing factor of breast carcinogenesis [53]. Interestingly, we observed that most of these CNVs affected the same biological process category (metabolic related biological processes) which play a crucial

role during tumorigenesis. Based on these observations, we infer that CNVs of different chromosome locations may function similarly by affecting metabolic related biological processes to drive breast cancer development.

In multiple myeloma, chromosomes 1q, 5q, 9q31, 11p, 11q, 15q ,19p13 and 19q13 were found to be highly enriched. Numerous studies have indicated that cytogenetic aberrations in these chromosomes are recurrent events that contribute to the disease and patient risk stratification [54, 55]. Specifically, the gain of odd numbered chromosomes such 5, 7, 9, 11, 15 and 19 is a major class of recurrent genetic abnormality in multiple myeloma and associated with favorable prognosis Our method was able to recapitulate the finding of cytoband 5q31, which was found to also confer a more favorable prognosis [55]. Conversely, genomic abnormalities of 1q is a high incidence event associated with very poor prognosis [56]. Though there was some overlap in cytobands between modules 1 and 2, the results from GO enrichment analyses were different (protein localization and viral response vs cell cycle and apoptosis), suggesting the CNV-influenced genes in each module contribute to a few key biological functions related to the disease.

In summary, by applying the TPQCI method, we were able to identify modules specific to each cancer type. Our proposed method was able to detect genes whose chromosomal locations were previously identified to be associated with their corresponding disease. Through our analyses, we identified CNV-influenced gene modules that behave in two ways: (1) gene modules enriched in different cytobands demonstrated similar biological functions and (2) gene modules enriched in similar cytobands demonstrated different biological functions. These observations suggest that modules detected by our proposed TPQCI metric can identify disease specific functionally enriched genes influenced by CNVs.

4.4 Possibility of module dividing

By performing enrichment analysis on modules, we can easily determine the biological significances of unsupervised detected modules. However, if the detected gene module is too large, it will become a challenge for the downstream analysis and interpretation since the giant module may include too much information. Therefore, it is necessary to further divide the large module into sub-modules in this case. In our proposed module detection algorithm, there are two parameters, τ_g and τ_l, which can be tuned to adjust the module size, giving us the possibility to further dividing the modules. Thus, when large modules appear, we can rerun the module detection algorithm on them with a smaller τ_g and greater τ_l to obtain sub-modules.

5 Conclusion

In this paper, we proposed a new metric called TPQCI to quantify the influence of CNVs on a single gene based on the functional genomics data of diseases. This metric uses topology potential to integrate CNV frequency, correlation between CNV and gene expression, and interactions provided by PPI network. We demonstrate that TPQCI can effectively measure CNV influence. By using TPQCI to detect functionally enriched genes influenced by CNVs in multiple myeloma and breast cancer, we found most of the significantly enriched cytobands were confirmed in their corresponding cancer. This reflects that our proposed TPQCI metric can effectively assess the impact that CNVs in different cancers.

Supplementary materials

Supplementary File 1 Selected CNV genes associate with multiple myeloma and breast cancer

Supplementary File 2 TPQCI and frequency of CNV happening of each gene in multiple myeloma and breast cancer

Supplementary File 3 Modules detected by our proposed method in multiple myeloma and breast cancer

Acknowledgments

We thank the MMRF for the CoMMpass study data that were generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org).

Funds

This work is partially supported by the Indiana University Precision Health Initiative (to Kun Huang and Jie Zhang), the State Key Program of National Natural Science Foundation of China (Grant No.61633004, to Xiufen Ye), the National key research and development program of China (Grant No. 2018YFC0310102 and 2017YFC0306001, to Xiufen Ye) , the Development Project of Applied Technology in Harbin (Grant No.2016RAXXJ071, to Xiufen Ye) and China Scholarship Council (No. 201806680029 to Yusong Liu).

Reference

[1] D.F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T.D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C.H. Ihm, K. Kristiansson, D.G. Macarthur, J.R. Macdonald, I. Onyiah, A.W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N.P. Carter, C. Lee, S.W. Scherer, M.E. Hurles, Origins and functional impact of copy number variation in the human genome, Nature 464(7289) (2010) 704-12.

[2] X. Lu, X. Li, P. Liu, X. Qian, Q. Miao, S. Peng, The Integrative Method Based on the Module-Network for Identifying Driver Genes in Cancer Subtypes, Molecules 23(2) (2018).

[3] S. Srihari, M. Kalimutho, S. Lal, J. Singla, D. Patel, P.T. Simpson, K.K. Khanna, M.A. Ragan, Understanding the functional impact of copy number alterations in breast cancer using a network modeling approach, Mol Biosyst 12(3) (2016) 963-72.

[4] W. Zhou, Z. Zhao, R. Wang, Y. Han, C. Wang, F. Yang, Y. Han, H. Liang, L. Qi, C. Wang, Z. Guo, Y. Gu, Identification of driver copy number alterations in diverse cancer types and application in drug repositioning, Mol Oncol 11(10) (2017) 1459-1474.

[5] B.A. Walker, E.M. Boyle, C.P. Wardell, A. Murison, D.B. Begum, N.M. Dahir, P.Z. Proszek, D.C. Johnson, M.F. Kaiser, L. Melchor, L.I. Aronson, M. Scales, C. Pawlyn, F. Mirabella, J.R. Jones, A. Brioli, A. Mikulasova, D.A. Cairns, W.M. Gregory, A. Quartilho, M.T. Drayson, N. Russell, G. Cook, G.H. Jackson, X. Leleu, F.E. Davies, G.J. Morgan, Mutational Spectrum, Copy Number Changes, and Outcome: Results of a Sequencing Study of Patients With Newly Diagnosed Myeloma, J Clin Oncol 33(33) (2015) 3911-20.

[6] C.Y. Yu, S. Xiang, Z. Huang, T.S. Johnson, X. Zhan, Z. Han, M. Abu Zaid, K. Huang, Gene Co-expression Network and Copy Number Variation Analyses Identify Transcription Factors Associated With Multiple Myeloma Progression, Front Genet 10 (2019) 468.

[7] L. Liang, J.Y. Fang, J. Xu, Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy, Oncogene 35(12) (2016) 1475-82.

[8] H. Wang, L. Liang, J.Y. Fang, J. Xu, Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers, Oncogene 35(16) (2016) 2011-9.

[9] L. Xu, Y. Zheng, J. Liu, D. Rakheja, S. Singleterry, T.W. Laetsch, J.F. Shern, J. Khan, T.J. Triche, D.S. Hawkins, J.F. Amatruda, S.X. Skapek, Integrative Bayesian Analysis Identifies Rhabdomyosarcoma Disease Genes, Cell Rep 24(1) (2018) 238-251.

[10] L. Zhang, Y. Yuan, K.H. Lu, L. Zhang, Identification of recurrent focal copy number variations and their putative targeted driver genes in ovarian cancer, BMC Bioinformatics 17(1) (2016) 222.
[11] X. Zhao, Y. Lei, G. Li, Y. Cheng, H. Yang, L. Xie, H. Long, R. Jiang, Integrative analysis of cancer driver genes in prostate adenocarcinoma, Mol Med Rep 19(4) (2019) 2707-2715.
[12] M. Fanciulli, E. Petretto, T.J. Aitman, Gene copy number variation and common human disease, Clin Genet 77(3) (2010) 201-13.
[13] C.N. Henrichsen, E. Chaignat, A. Reymond, Copy number variants, diseases and gene expression, Hum Mol Genet 18(R1) (2009) R1-8.
[14] H. Hieronymus, R. Murali, A. Tin, K. Yadav, W. Abida, H. Moller, D. Berney, H. Scher, B. Carver, P. Scardino, N. Schultz, B. Taylor, A. Vickers, J. Cuzick, C.L. Sawyers, Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death, Elife 7 (2018).
[15] Y.P. Lai, L.B. Wang, W.A. Wang, L.C. Lai, M.H. Tsai, T.P. Lu, E.Y. Chuang, iGC-an integrated analysis package of gene expression and copy number alteration, BMC Bioinformatics 18(1) (2017) 35.
[16] S.H. Liu, P.C. Shen, C.Y. Chen, A.N. Hsu, Y.C. Cho, Y.L. Lai, F.H. Chen, C.Y. Li, S.C. Wang, M. Chen, I.F. Chung, W.C. Cheng, DriverDBv3: a multi-omics database for cancer driver gene research, Nucleic Acids Res 48(D1) (2020) D863-d870.
[17] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J.R. Pollack, P. Wang, Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer, Ann Appl Stat 4(1) (2010) 53-77.
[18] L. Li, Q. Lei, S. Zhang, L. Kong, B. Qin, Screening and identification of key biomarkers in hepatocellular carcinoma: Evidence from bioinformatic analysis, Oncol Rep 38(5) (2017) 2607-2618.
[19] J. Ren, L. Shang, Q. Wang, J. Li, Ranking Cancer Proteins by Integrating PPI Network and Protein Expression Profiles, Biomed Res Int 2019 (2019) 3907195.
[20] B. Liang, C. Li, J. Zhao, Identification of key pathways and genes in colorectal cancer using bioinformatics analysis, Med Oncol 33(10) (2016) 111.
[21] T.I. Zack, S.E. Schumacher, S.L. Carter, A.D. Cherniack, G. Saksena, B. Tabak, M.S. Lawrence, C.-Z. Zhsng, J. Wala, C.H. Mermel, C. Sougnez, S.B. Gabriel, B. Hernandez, H. Shen, P.W. Laird, G. Getz, M. Meyerson, R. Beroukhim, Pan-cancer patterns of somatic copy number alteration, Nature genetics 45(10) (2013) 1134-1140.
[22] F. Zhang, W. Gu, M.E. Hurles, J.R. Lupski, Copy number variation in human health, disease, and evolution, Annu Rev Genomics Hum Genet 10 (2009) 451-81.
[23] X. Shao, N. Lv, J. Liao, J. Long, R. Xue, N. Ai, D. Xu, X. Fan, Copy number variation is highly correlated with differential gene expression: a pan-cancer study, BMC Med Genet 20(1) (2019) 175.
[24] W.-Y. Gan, N. He, D.-Y. Li, J.-M. Wang, Community discovery method in networks based on topological potential, Journal of Software 20(8) (2009) 2241-2254.
[25] Z. Kang, G. Shi, S. Huang, W. Chen, X. Pu, J.T. Zhou, Z. Xu, Multi-graph fusion for multi-view spectral clustering, Knowledge-Based Systems 189 (2020) 105102.
[26] F. Liu, D. Choi, L. Xie, K. Roeder, Global spectral clustering in dynamic networks, Proceedings of the National Academy of Sciences 115(5) (2018) 927-932.
[27] Y. Wang, L. Wu, X. Lin, J. Gao, Multiview Spectral Clustering via Structured Low-Rank Matrix Factorization, IEEE Transactions on Neural Networks and Learning Systems 29(10) (2018) 4833-4843.
[28] M. Li, Y. Lu, J. Wang, F. Wu, Y. Pan, A Topology Potential-Based Method for Identifying Essential Proteins from PPI Networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics 12(2) (2015) 372-383.
[29] Y. Liu, C.Y. Yu, W. Shao, J. Hou, W. Feng, J. Zhang, X. Ye, K. Huang, TPSC: A Module Detection Method Based on Topology Potential and Spectral Clustering in Weighted Networks and Its Application in Gene Co-expression Module Discovery, International Conference on Intelligent Biology and Medicine (ICIBM2020), Virtual, 2020.

[30] M.J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A.N. Brooks, J. Zhu, D. Haussler, Visualizing and interpreting cancer genomics data via the Xena platform, Nature Biotechnology 38(6) (2020) 675-678.
[31] J. Piñero, J.M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L.I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update, Nucleic Acids Research 48(D1) (2019) D845-D855.
[32] M.J. Cowley, M. Pinese, K.S. Kassahn, N. Waddell, J.V. Pearson, S.M. Grimmond, A.V. Biankin, S. Hautaniemi, J. Wu, PINA v2.0: mining interactome modules, Nucleic Acids Res 40(Database issue) (2012) D862-5.
[33] M.G. Everett, S.P. Borgatti, Induced, endogenous and exogenous centrality, Social Networks 32(4) (2010) 339-344.
[34] A. Singh, R.R. Singh, S.R.S. Iyengar, Node-weighted centrality: a new way of centrality hybridization, Computational Social Networks 7(1) (2020) 6.
[35] X. Yu, B. Wu, Y. Liu, Node role analysis algorithm based on directed topological potential, 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2014, pp. 681-685.
[36] Z. Wang, Z. Chen, Y. Zhao, S. Chen, A Community Detection Algorithm Based on Topology Potential and Spectral Clustering, The Scientific World Journal 2014 (2014) 329325.
[37] W. Huang da, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res 37(1) (2009) 1-13.
[38] Y. Xiang, P.R. Payne, K. Huang, Transactional database transformation and its application in prioritizing human disease genes, IEEE/ACM Trans Comput Biol Bioinform 9(1) (2012) 294-304.
[39] A.-L. Barabási, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nature Reviews Genetics 5(2) (2004) 101-113.
[40] M. Kurant, A. Markopoulou, P. Thiran, On the bias of BFS (Breadth First Search), 2010 22nd International Teletraffic Congress (lTC 22), 2010, pp. 1-8.
[41] J. Chen, B.J. Aronow, A.G. Jegga, Disease candidate gene identification and prioritization using protein interaction networks, BMC Bioinformatics 10(1) (2009) 73.
[42] S. Manni, M. Carrino, F. Piazza, Role of protein kinases CK1α and CK2 in multiple myeloma: regulation of pivotal survival and stress-managing pathways, J Hematol Oncol 10(1) (2017) 157.
[43] G. Liu, M. Yu, B. Wu, S. Guo, X. Huang, F. Zhou, F.X. Claret, Y. Pan, Jab1/Cops5 contributes to chemoresistance in breast cancer by regulating Rad51, Cell Signal 53 (2019) 39-48.
[44] X. Yuan, G. Yu, X. Hou, I.-M. Shih, R. Clarke, J. Zhang, E.P. Hoffman, R.R. Wang, Z. Zhang, Y. Wang, Genome-wide identification of significant aberrations in cancer genome, BMC Genomics 13(1) (2012) 342.
[45] P. Dharanipragada, S. Vogeti, N. Parekh, iCopyDAV: Integrated platform for copy number variations—Detection, annotation and visualization, PLOS ONE 13(4) (2018) e0195334.
[46] X. Yuan, J. Yu, J. Xi, L. Yang, J. Shang, Z. Li, J. Duan, CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2019) 1-1.
[47] X. Yuan, J. Bai, J. Zhang, L. Yang, J. Duan, Y. Li, M. Gao, CONDEL: Detecting Copy Number Variation and Genotyping Deletion Zygosity from Single Tumor Samples Using Sequence Data, IEEE/ACM Transactions on Computational Biology and Bioinformatics 17(4) (2020) 1141-1153.
[48] E.S. Hwang, S. DeVries, K.L. Chew, D.H. Moore, 2nd, K. Kerlikowske, A. Thor, B.M. Ljung, F.M. Waldman, Patterns of chromosomal alterations in breast ductal carcinoma in situ, Clin Cancer Res 10(15) (2004) 5160-7.
[49] A.C. Berger, A. Korkut, R.S. Kanchi, A.M. Hegde, W. Lenoir, W. Liu, Y. Liu, H. Fan, H. Shen, V. Ravikumar, A. Rao, A. Schultz, X. Li, P. Sumazin, C. Williams, P. Mestdagh, P.H. Gunaratne, C. Yau, R. Bowlby, A.G. Robertson, D.G. Tiezzi, C. Wang, A.D. Cherniack, A.K. Godwin, N.M. Kuderer, J.S. Rader, R.E. Zuna, A.K. Sood, A.J. Lazar, A.I. Ojesina, C. Adebamowo, S.N. Adebamowo, K.A. Baggerly, T.W. Chen, H.S. Chiu, S.

Lefever, L. Liu, K. MacKenzie, S. Orsulic, J. Roszik, C.S. Shelley, Q. Song, C.P. Vellano, N. Wentzensen, J.N. Weinstein, G.B. Mills, D.A. Levine, R. Akbani, A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers, Cancer Cell 33(4) (2018) 690-705.e9.
[50] L.C. Chen, C. Dollbaum, H.S. Smith, Loss of heterozygosity on chromosome 1q in human breast cancer, Proc Natl Acad Sci U S A 86(18) (1989) 7204-7.
[51] K. Rennstam, M. Ahlstedt-Soini, B. Baldetorp, P.O. Bendahl, A. Borg, R. Karhu, M. Tanner, M. Tirkkonen, J. Isola, Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. A study of 305 tumors by comparative genomic hybridization, Cancer Res 63(24) (2003) 8861-8.
[52] S. Weber-Mangal, H.P. Sinn, S. Popp, R. Klaes, R. Emig, M. Bentz, U. Mansmann, G. Bastert, C.R. Bartram, A. Jauch, Breast cancer in young women (< or = 35 years): Genomic aberrations detected by comparative genomic hybridization, Int J Cancer 107(4) (2003) 583-92.
[53] H. Bürger, M. de Boer, P.J. van Diest, E. Korsching, Chromosome 16q loss--a genetic key to the understanding of breast carcinogenesis, Histol Histopathol 28(3) (2013) 311-20.
[54] G.J. Morgan, B.A. Walker, F.E. Davies, The genetic architecture of multiple myeloma, Nat Rev Cancer 12(5) (2012) 335-48.
[55] H. Avet-Loiseau, C. Li, F. Magrangeas, W. Gouraud, C. Charbonnel, J.L. Harousseau, M. Attal, G. Marit, C. Mathiot, T. Facon, P. Moreau, K.C. Anderson, L. Campion, N.C. Munshi, S. Minvielle, Prognostic significance of copy-number alterations in multiple myeloma, J Clin Oncol 27(27) (2009) 4585-90.
[56] P. Liebisch, C. Wendl, A. Wellmann, A. Kröber, G. Schilling, H. Goldschmidt, H. Einsele, C. Straka, M. Bentz, S. Stilgenbauer, H. Döhner, High incidence of trisomies 1q, 9q, and 11q in multiple myeloma: results from a comprehensive molecular cytogenetic analysis, Leukemia 17(12) (2003) 2535-7.