Hindawi
Complexity
Volume 2017, Article ID 4826206, 10 pages
https://doi.org/10.1155/2017/4826206

Research Article

DriverFinder: A Gene Length-Based Network Method to Identify Cancer Driver Genes

Pi-Jing Wei,[1] Di Zhang,[1] Hai-Tao Li,[2] Junfeng Xia,[3] and Chun-Hou Zheng[1]

[1] *College of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China*
[2] *State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu 210018, China*
[3] *Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China*

Correspondence should be addressed to Chun-Hou Zheng; zhengch99@126.com

Received 7 March 2017; Revised 8 June 2017; Accepted 3 July 2017; Published 10 August 2017

Academic Editor: Haiying Wang

Integration of multi-omics data of cancer can help people to explore cancers comprehensively. However, with a large volume of different omics and functional data being generated, there is a major challenge to distinguish functional driver genes from a sea of inconsequential passenger genes that accrue stochastically but do not contribute to cancer development. In this paper, we present a gene length-based network method, named DriverFinder, to identify driver genes by integrating somatic mutations, copy number variations, gene-gene interaction network, tumor expression, and normal expression data. To illustrate the performance of DriverFinder, it is applied to four cancer types from The Cancer Genome Atlas including breast cancer, head and neck squamous cell carcinoma, thyroid carcinoma, and kidney renal clear cell carcinoma. Compared with some conventional methods, the results demonstrate that the proposed method is effective. Moreover, it can decrease the influence of gene length in identifying driver genes and identify some rare mutated driver genes.

1. Background

At present, understanding the mechanisms of cancer development and uncovering actionable target genes for cancer treatment are still difficult challenges. With rapid advances in high-throughput sequencing technologies, some large-scale cancer genomics projects, such as The Cancer Genome Atlas (TCGA) [1] and International Cancer Genome Consortium (ICGC) [2], have produced different omics data including a rich dataset of whole-exome and RNA sequence data [3, 4], which provides chances to allow us to accurately infer tumor-specific alterations [5] and help in precision medicine in cancers treatment [6, 7]. However, many of genetic changes represent neutral variations that do not contribute to cancer development which are called passenger mutations [6, 8]. Only a few alterations are causally implicated in the process of oncogenesis and provide a selection growth advantage which are referred to as driver mutations [8, 9]. Hence, it is a major challenge to distinguish pathogenic driver mutations from the so-called random mutated passenger mutations [10].

Previously, there were multiple computational methods to identify driver genes based on gene mutational frequency (termed as frequency-based method) in a large cohort of cancer patients [11–13]. However, the infrequently mutated drivers are inclined to be ignored by frequency-based methods. Also mutational heterogeneity in cancer genomes is an important factor affecting the performance of frequency-based methods [14]. In addition, further studies have realized that driver mutations or genes disrupt some cellular signaling or regulatory pathways which promote the progression of cancer [15, 16]. In fact, genes affect various biological processes by related complex networks instead of acting in isolation in cancer [17]. In addition, the cancer is a result of interplay of various types of genetic changes which form complex and dynamic networks [18]. Thus, many network-based and pathway-based approaches have been proposed to prioritize driver mutations and genes. For instance, Dendrix was a pathway-based algorithm for discovery of mutated driver pathways in cancer using somatic mutation data [19]. After that, Multi-Dendrix algorithm was proposed to extend

Dendrix method in order to guarantee yielding the optimal set of pathways [20]. MDPFinder was also a pathway-based method to solve the so-called maximum weight submatrix problem proposed in Dendrix method [19] which was aimed at identifying mutated driver pathways from mutation data in cancer [21]. And Zhang et al. proposed CoMDP method which focused on cooccurring driver pathways rather than single pathway [22]. In addition, iMCMC was a network-based method by integrating somatic mutation, CNVs, and gene expressions without any prior information [6]. Another method, DawnRank, was also a network-based algorithm to discover personalized causal driver mutations by ranking mutated genes according to their potential to be drivers based on PageRank algorithm [23]. Bashashati et al. developed a method called DriverNet which comprehensively analyzed genomes and transcriptomes datasets to identify likely driver genes in population-level by virtue of their effect on mRNA expression networks and also reveal the infrequent but important genes and patterns of pathway [10]. VarWalker was a personalized network-assisted approach to prioritize well-known, infrequently mutated genes and interpret mutation data in NGS studies [24].

Although some proposed methods can determine potential drivers, most of them do not consider the influence of gene length to the results; in other words, they may identify some likely false positive driver genes according to known driver genes datasets. And it has been indicated that driver genes are related to not only mutation frequency, but also mutation context or gene length [25] and variants tend to arise more frequently in long genes [26]. For example, *TTN*, the longest gene in human genome, accumulates many variants just due to its length [24, 26]. *TTN* may be selected in many computational methods; however, it usually serves as passenger gene [27]. This phenomenon indicates that many current methods have a strong preference towards identifying long genes [24]. So it is essential to filter those frequently mutated genes due to long length. VarWalker takes into account the gene length; however, it does not consider the influence of mutation to expression. In addition, some genomic variations in a gene may lead to extreme changes in some outlying genes expression level which are associated with the mutated gene through gene-gene interaction network or pathways and these outlying genes are often called outliers [10]. And, it has been proved that cancer-associated genes are more effectively detected by interindividual variation analysis rather than only calculating differences in the mean expression across different samples [28]. That is, the outliers are related to not only tumor expression distribution but also the corresponding normal expression distribution. Moreover, various cellular processes are often affected by genes in complex networks rather than genes acting in isolation [17] and cancer is also related to a set of genes interacting together in a molecular network [29]. So networks usually provide a convenient way to explore the context within which single gene operates [30]. It should be noted that prior knowledge such as protein-protein interaction (PPI) network can provide some useful information; however, prior knowledge is limited and may lead to discarding some important information in some instances [22]. In our previous work [31], we only consider the prior information of gene-gene interaction network. So it is essential to extend gene-gene interaction network.

In this study, we proposed an integrated framework named DriverFinder to identify driver genes by integrating somatic mutation data, copy number variations (CNVs), tumor and normal expression data, and gene-gene interaction network. Firstly, the gene length is taken into consideration to filter some frequently genes because of long length. Moreover, in this method, we integrated tumor expression and normal expression to construct outlier matrix rather than only using tumor expression. Furthermore, to increase accuracy of identifying drivers, we calculated Pearson correlation coefficients (PCC) of genes and combined them with PPI network to construct a new dynamic interaction network for each cancer type. In order to estimate the performance of DriverFinder method, we applied it to four different large-scale TCGA datasets, including breast cancer (BRCA), head and neck squamous cell carcinoma (HNSC), thyroid carcinoma (THCA), and kidney renal clear cell carcinoma (KIRC), and compared it with MUFFINN [32], DriverNet [10], and frequency-based method. The results demonstrated that DriverFinder can identify drivers effectively and decrease false positive, that is, filtering some long and frequently mutated but functionally neutral genes.

2. Materials and Methods

We proposed DriverFinder to identify cancer driver genes by integrating multiomics data (Figure 1). The detailed description of it is shown in Figure 1.

As is shown in Figure 1, the first step is to estimate the occurrence of mutation events in the genome by fitting them into a generalized additive model [24]. Then a weighted resample-based test is used to filter long passenger genes according to approximated probabilities based on benchmark of coding gene length. Secondly, for gene expression, we compare expression of tumors with normal samples to determine the outlying genes. Then a gene-gene interaction network combined with prior knowledge and PCC among expression data is used to relate mutations to their consequent effect on gene expression. The associations between mutated and outlying genes are formulated using a bipartite graph where the left nodes indicate the genes mutation status and the right nodes indicate the outlying expression status in each patient. For each patient, there is an edge between g_i and g_j if the left partition gene g_i is mutated and the right gene g_j is an outlier in some samples and they also have high correlations in gene-gene interaction network. Secondly, greedy algorithm is used to prioritize mutated genes based on the coverage. In each iteration of the greedy algorithm, the mutated gene on the left partition of the bipartite graph which relates to the most outlying expression genes is chosen. Until all the outlying expression genes are covered by the least mutated genes on the left, iteration is stopped. Then the mutated genes are ranked based on their coverage. So genes with the most outlying expression are appointed as candidate driver genes. Finally, the statistical significance test based on null distribution is applied to these putative driver genes.

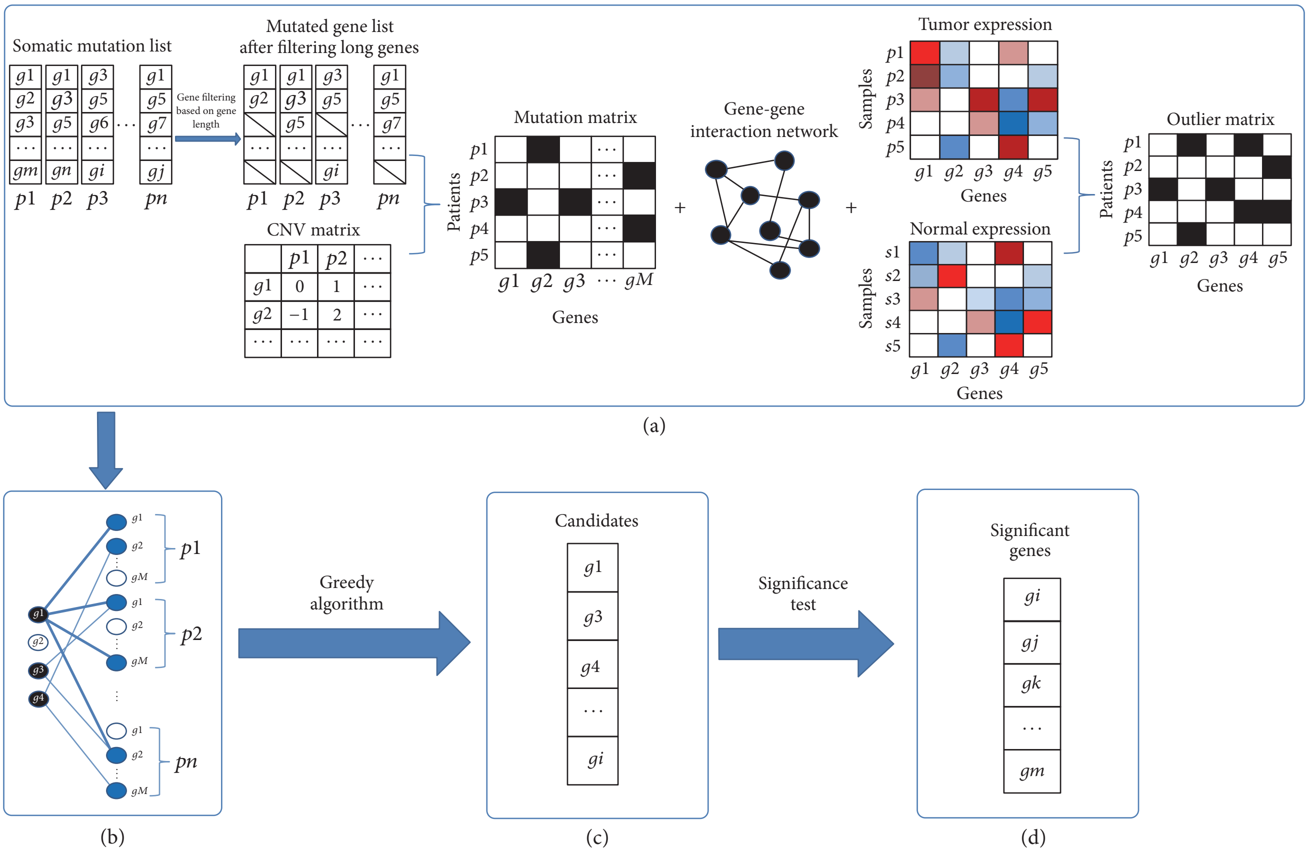


Figure 1: *The flowchart of DriverFinder method.* (a) Input datasets consist of somatic mutation, CNV, normal and tumor expression data, and gene-gene interaction network. A generalized additive model is performed on somatic mutation data to filter mutated genes which occurred at random due to long length. After that, the residual significant genes are combined with CNV to construct mutation data. The gene-gene interaction network is constructed by integrating prior gene-gene interaction network and Pearson correlated coefficient network. And the outlying matrix is constructed by analyzing interindividual variation in tumor and normal expression. (b) Given mutation data, outlying data, and gene-gene interaction network, the bipartite graph is obtained. The black nodes on the left indicate mutated genes and the blue nodes on the right represent outlying expression genes. (c) Candidate genes are obtained by greedy algorithm. The more outlying expression events the gene overlaps, the higher the gene ranks are. (d) Statistical test is performed on candidate genes to select important putative drivers by p value < 0.05.

2.1. Construction of Mutation Matrix M. In terms of somatic mutation, we downloaded it from TCGA data portal (https://cancergenome.nih.gov/) and only considered the data of level 2. As a matter of fact, Jia et al. explored long genes in two datasets and examined gene length effects by plotting the proportion of mutated genes versus their complementary DNA (cDNA) length [24]. They discovered that two sets of mutated genes were positively correlated with cDNA length, and longer genes were more likely to be mutated genes [24]. Hence, frequency-based methods may be inclined to select long genes as drivers. So, it is necessary to perform gene length-based filtration. In this study, in order to accurately estimate the mutation rate for each gene, generalized additive model was adopted to compute a probability weight vector (PWV) for the mutation genes of each sample [24].

Here, only somatic mutated genes mapped onto the benchmark of consensus coding sequences (CCDS) dataset [33] which contains a core set of consistently annotated and high quality human and mouse protein coding regions are reserved. And those mapped genes in this study have been allocated cDNA length based on their coding sequences [24]. Assuming the vector X as the cDNA gene length and the following model is used to assess the probability of a gene to be mutated,

$$\operatorname{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \sim S(X), \tag{1}$$

where $S(\cdot)$ represents an unspecified smooth function and π = num of mutant genes/total num of CCDS genes represents the proportion of mutant genes in the specific samples [24]. Each gene then would be assigned a PWV value. Afterwards, a resampling test based on the probability of each gene is performed 1000 times in each sample and the null distribution is that in genes mutations occur at random. Then we define mutation frequency as

$$f = \frac{\text{num of selecting the genes in 1000 times}}{1000}, \tag{2}$$

where f represents mutation frequency. Next we filter out genes with frequencies ≥ 5% in random datasets unless they are Cancer Gene Census (CGC) genes. Then a list of significant mutant genes S is obtained.

As for CNVs, which have been processed by GISTIC 2.0, they are acquired from http://gdac.broadinstitute.org/runs/ (v2014_10_17). There are five types of copy number including amplification, gain, diploid, heterozygous deletion, and homozygous deletion in this dataset. Here, we only screen out amplifications and homozygous deletions to construct CNV matrix C. Finally, the significant mutated gene list S and CNV matrix C are combined to generate the patient-mutation binary matrix M, in which $M_{ij} = 1$ indicates there is genetic alteration, that is, mutation, amplification, or homozygous deletion, in the jth gene of the ith sample. Otherwise, $M_{ij} = 0$.

2.2. Construction of Expression Outlier Matrix E. Gene expression data (level 3) including tumor and normal expression data is also downloaded from TCGA data portal. Moreover, some studies have shown that assessment of interindividual variation of gene expression performs well in predicting cancer-associated genes [28]. So in this study, the outlying matrix is determined based on analysis of interindividual variation in tumor and normal expression rather than differences in mean expression levels or only tumor expression distribution [28]. For each type of cancers, there are two expression datasets $T(i, j)$ and $N(i, j)$ which indicate the real-valued expression measure of gene i in sample j of tumor and normal datasets, respectively. For each gene, the outliers in this study are defined as tumors whose expression levels are outside the four-standard deviation range of the expression values of the gene across all the normal samples [28]. It is formularized as

$$\begin{gathered} \text{tumor expression} < m(N) - 4 * \text{sd}(N) \\ \text{or tumor expression} > m(N) + 4 * \text{sd}(N) \end{gathered} \tag{3}$$

in which $m(N)$ is the mean expression and $\text{sd}(N)$ indicates the standard deviation of gene expression in normal samples. Then the binary patient-outlier matrix E is constructed and the value of $E(i, j)$ indicated whether gene i in patient j is an outlier among the population-level distribution for that gene. If the expression of gene i is an outlier in patient j, $E(i, j) = 1$; otherwise, $E(i, j) = 0$.

2.3. Gene-Gene Interaction Network. It is noteworthy that most prior knowledge such as PPI network or pathways is incomplete and a great deal of knowledge about biological pathways remains unclear [22]. In our previous study [31], we relied on prior knowledge about gene influence graph integrated from known gene networks [10] which often leaved out some likely important nodes. So in this study, we constructed a new dynamic gene-gene interaction network by incorporating gene-gene correlation coefficients with prior knowledge. That is, firstly, PCCs between pairwise genes are obtained by normalized tumor expression. Acceptable correlations with PCC > 0.75 are often considered high correlated and selected [34]. Here, we choose 0.8 as threshold to ensure selected pairwise genes being higher correlated and increase reliability. The edges with PCC > 0.8 are selected and set to 1, otherwise 0. In order to retrieve some important prior knowledge simultaneously, the known gene network (termed the influence graph in DriverNet [10]) is mapped onto the binary matrix obtained from correlation coefficients matrix. So a new and dynamic gene-gene interaction network (termed as G after) included prior knowledge and deduced knowledge is established. When there is a correlation, that is, PCC > 0.8 or 1 in influence graph between gene i and gene j, $G_{ij} = 1$; otherwise, $G_{ij} = 0$.

2.4. Significance Estimation. With the aim of testing the statistical significance of the driver candidates, we apply a randomization framework. The algorithm is run on the random N permuted original datasets (mutation data and outlier data). Then we assess the significance by seeing if the results on real data are significantly different from the results on random datasets and obtain the p value of each candidate drivers. The statistical significance of g is defined as follows [10]:

$$p \text{ value}(g) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(\text{COV}_{gij} > \text{COV}_g\right)}{\sum_{i=1}^{N} M_i}, \tag{4}$$

where N is permutation times and M_i is the number of candidate drivers in the ith run of the approach. COV_g is the coverage of g calculated from our method. Here we choose $N = 50$. The statistical significance of g means that the times of the observed driver genes with coverage are more than COV_g. Finally, genes with p value less than 0.05 are nominated as candidate drivers.

3. Results

3.1. Datasets. In this work, four TCGA datasets, BRCA, HNSC, KIRC, and THCA, were applied to our method. For each cancer type, four different omics data consisting of somatic mutation, tumor expression, normal expression, and CNV were used. The BRCA dataset includes copy number, 531 tumor samples' and 62 normal samples' expression data, and 962 samples' somatic mutation data. KIRC dataset contains copy number variations, 417 samples' somatic mutations data, and accompanying 534 tumor and 72 normal samples' expression data. The HNSC dataset contains 509 patients' somatic mutation data, 522 tumor and 44 normal samples' expression data, and copy number data. For THCA, it includes copy number variations, 435 patients' somatic mutation, and 513 tumor patients' and 59 normal samples' expression data. For each type of cancer, we only consider the samples which are common in tumor expression dataset and somatic mutation dataset.

3.2. Performance Evaluation. To evaluate the performance of our method on identifying known driver genes, we used annotated cancer-related genes datasets CGC database (15/7/2015) [35] and 20/20 rule [25] as approximate benchmarks. CGC is a database which catalogues 571 genes whose mutations have been causally involved in cancer [35]. 20/20 rule contains 138 driver genes in which 125 genes are affected by subtle mutations and 13 are affected by amplification or homozygous deletion [25]. We compared our method with frequency-based method, DriverNet [10], and MUFFINN [32] based on these two benchmarks.

In this work, we first counted the number of known drivers according to CGC genes of these four methods. For comparison, three measures based on top N genes including precision, recall, and $F1$ score are used which are defined as follows:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{(\#\text{ genes found in DriverFinder}) \cap (\#\text{ mutated genes in CGC})}{(\#\text{ genes found in DriverFinder})}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{(\#\text{ genes found in DriverFinder}) \cap (\#\text{ mutated genes in CGC})}{(\#\text{ genes in CGC})}, \\ F1\text{ score} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \end{aligned} \tag{5}$$

where TP indicates the number of overlapping genes found in our method and annotated genes associated with cancers in CGC. FP means the number of genes identified in our method, however not cataloged by CGC. FN is the number of genes in CGC but not contained in our method.

In general, DriverFinder almost outperforms other three methods in the top ranking genes of all the four cancer datasets (Figure 2; results of DriverFinder are shown in Supplementary File 1, in Supplementary Material available online at https://doi.org/10.1155/2017/4826206). Although, after approximately ranking top 30 genes in KIRC, MUFFINN performs comparably with DriverFinder, it has poorer performance in the top 30 genes. And the same phenomenon arises after top 60 genes in THCA. Analogous to it, the cumulative number of retrieved cancer genes annotated by 20/20 rule in KIRC by MUFFINN (Figure 3(c)) is also more than DriverFinder. A potential explanation of them may be that the total numbers of mutations in KIRC (10359 genes with 21089 mutations) and THCA (8899 genes with 16497 mutations) are remarkably less than in BRCA (16717 genes with 118098 mutations) and HNSC (14830 genes with 57164 mutations). So the gene-gene interaction network in KIRC and THCA may be simpler than in BRCA and HNSC; that is, genes may easily correlate directly with each other. And MUFFINN consider mutations only in direct neighbors [32]. On the one hand, this difference may help MUFFINN retrieve more genes; on the other hand, the number of mutations indicates that there may be more passengers (i.e., noise) in BRCA and HNSC and DriverFinder is more stable with noises.

Moreover, DriverFinder outperforms other three methods in BRCA, HNSC, and THCA by the cumulative numbers with 20/20 rule (Figure 3). In KIRC, it also has a better performance than DriverNet and frequency-based method within the top 100 genes.

3.3. DriverFinder Decreases the Effect of Gene Length. It is worth noting that, from the results of the DriverFinder, we can find that it can decrease the false positive because it has the good performance on filtering randomly mutated genes due to long length. The longest gene across human genome is *TTN* and it has been proven that higher mutation rate in it is likely to be artifacts [23, 36]. For example, in BRCA, *TTN* ranked 4 and 6 in frequency-based method and in MUFFINN, respectively, due to high mutation rate. Also it ranked 51 (p value = 0.031) as a candidate driver in DriverNet algorithm; however, it was filtered out with DriverFinder. In KIRC and HNSC, it ranked 4 and 3, respectively, based on mutation frequency and 22 and 18 in DriverNet, respectively. In addition, it is also at top 4 and 3 with MUFFINN in KIRC and HNSC, respectively, but it does not rank among the top 160 or 790 separately according to DriverFinder. Furthermore, for THCA, frequency-based method and MUFFINN ranked *TTN* as top 4 and 5, respectively. However, it is not identified by DriverFinder. These results proved the better performance of DriverFinder on filtering randomly mutated genes with long length than DriverNet, MUFFINN, and frequency-based method.

3.4. Pathway Enrichment Analysis. In order to investigate cancer-related pathways among the significant candidate drivers, Kyoto Encyclopedia of Genes and Genomes (KEGG) [37] pathway enrichment analysis was performed by statistics significantly candidate driver genes with p values less than 0.05 (see Supplementary File 2). The top 20 significant pathways are shown in Figure 4. We observed that the most enriched terms are cancer-related pathways in four cancer datasets. Moreover, ErbB signaling pathway, which is significantly enriched in BRCA (p value = $4.79E-07$) and KIRC (p value = $6.23E-07$), has been reported to play important roles in many tumors and *ErbB2/ErbB3* heterodimer functioned as an oncogenic unit in breast cancer [38]. Also, the significantly enriched VEGF signaling pathway ($4.94E-05$) in HNSC plays a pivotal role in tumor angiogenesis [39].

3.5. Discovering Rare Driver Genes. In this subsection, we exhibited that DriverFinder can identify rarely mutated but significant candidate driver genes which are defined as genes whose mutation frequency < 2% across all patients cohort. Here, we only selected highly ranked (top 30) rare genes for further analysis.

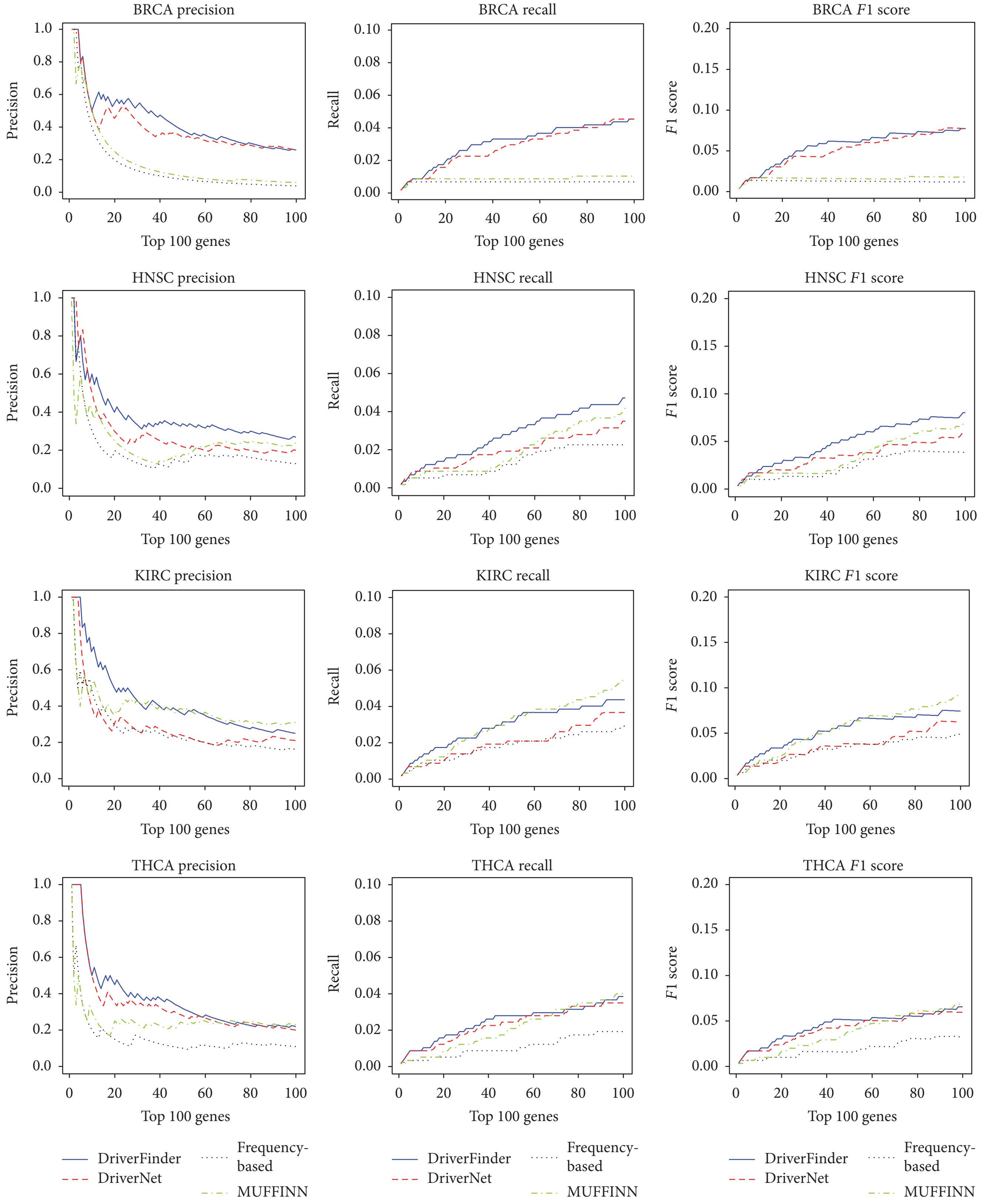


Figure 2: Performance comparison with CGC (precision, recall, and $F1$ score) of DriverFinder, DriverNet, MUFFINN, and frequency-based method on BRCA, HNSC, KIRC, and THCA datasets.

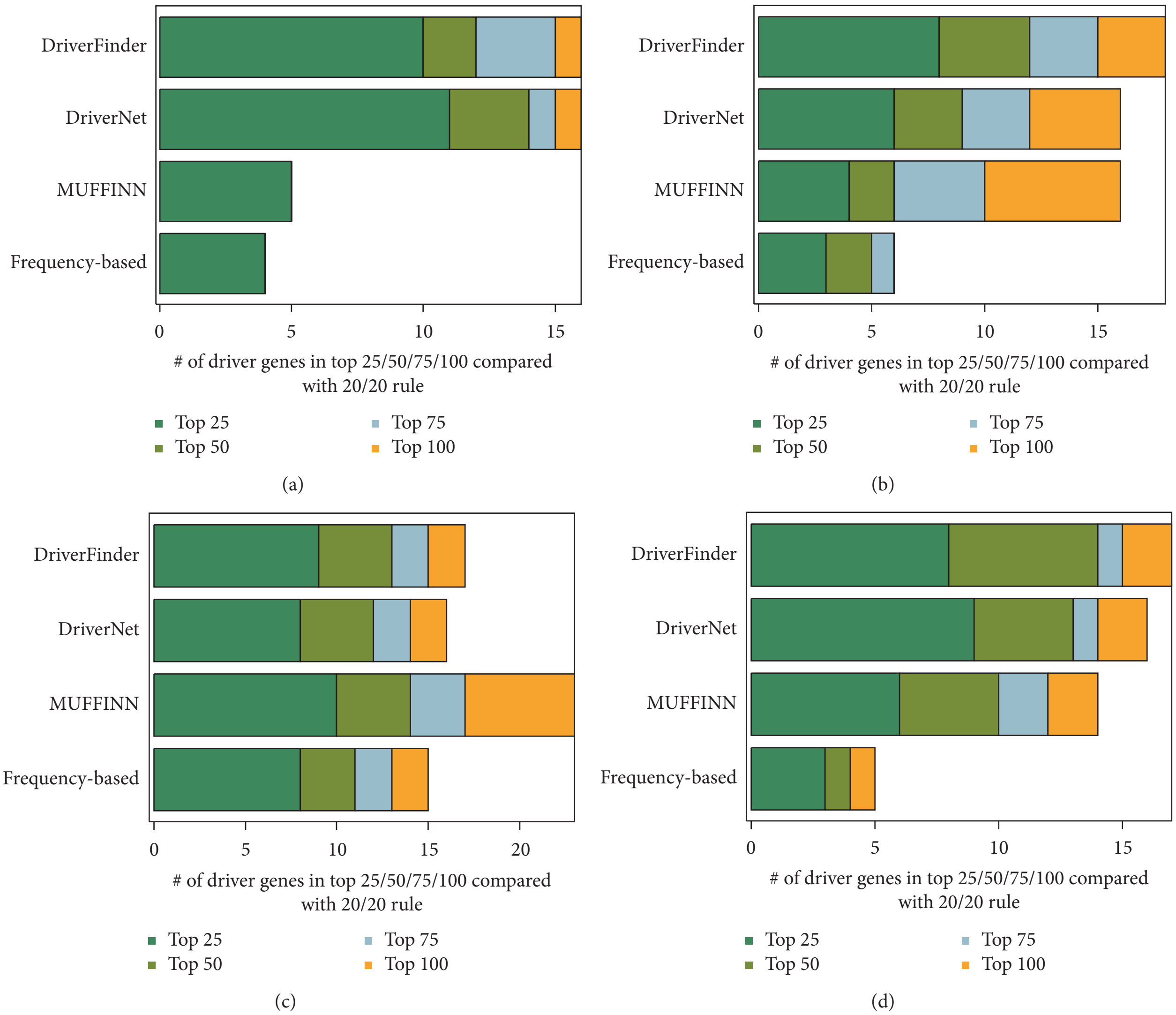


FIGURE 3: Cumulative numbers of retrieved cancer genes annotated by 20/20 rule within top 25, 50, 75, and 100 of (a) BRCA, (b) HNSC, (c) KIRC, and (d) THCA using four different methods.

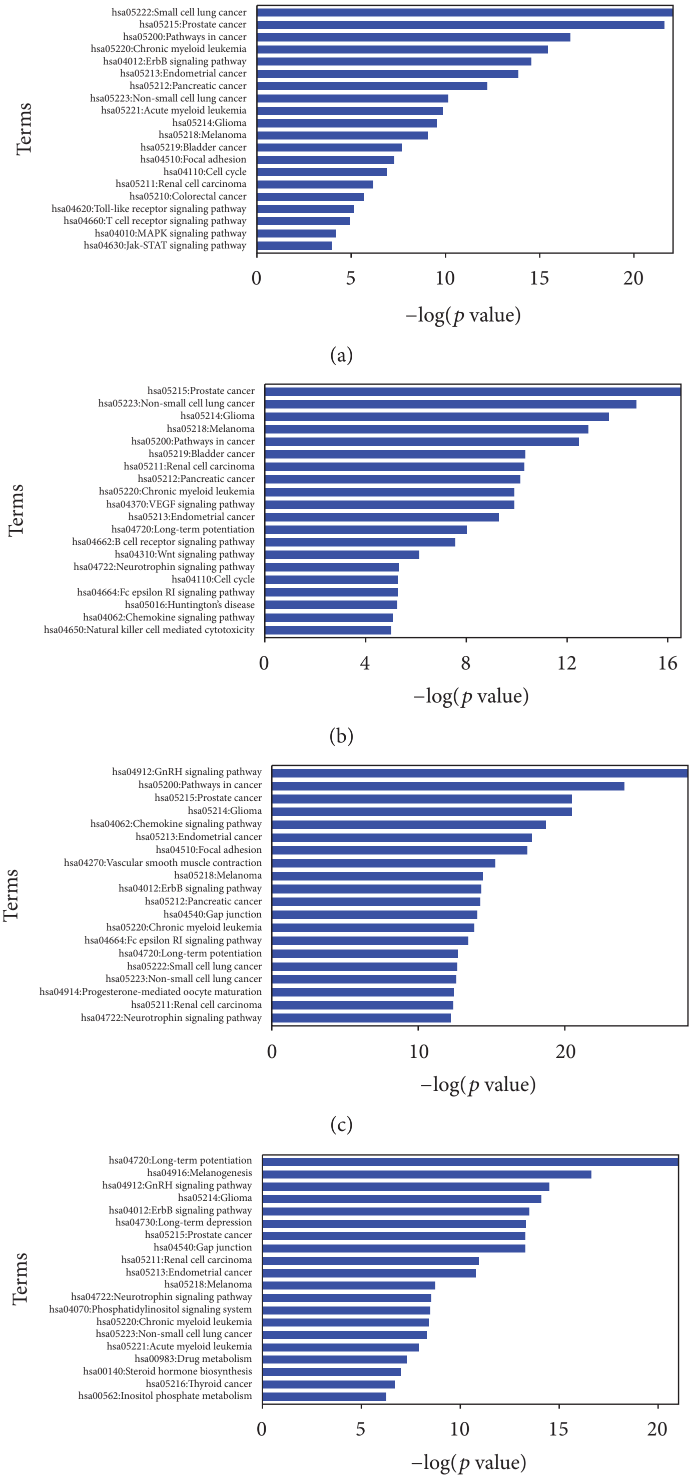
In BRCA, 3 rare genes (*PIK3R1*, *CREBBP*, and *PRKACB*) are ranked in top 30. In them, for *PIK3R1* (ranked 23) which is not ranked top 30 in the other three methods, underexpression might possibly lead to PI3K pathway activation and confer tumor development and progression in humans and it is a clinically useful independent prognostic marker in breast cancer [40]. Due to its low frequency mutations, any further statistical analyses concerning a possible association between *PIK3R1* mutations and clinical parameters are not allowed [40] and it is easily ignored by frequency-based methods. Moreover, for *CREBBP* (ranked 25) which also was not ranked top 30 in the other three methods, it has been occasionally reported in breast cancer [41]. *PRKACB* downregulates in non-small cell lung cancer and the effect of its upregulation on cell proliferation, apoptosis, and invasion also has been investigated [42].

In HNSC, also 3 rare genes are selected and one of them (*UGT2B4,* ranked 30) has shown potential to be a novel driver. *UGT2B4* genotypes associated with decreased enzyme activities are found to increase the risk of esophageal squamous cell carcinoma [43].

For KIRC and THCA, there are some important rare genes not identified by other methods. For example, *CLTC* in KIRC, which is contained in CGC and ranked 11 with low mutated frequency 1.68% in DriverFinder, encodes a major subunit of clathrin and is a fusion partner of *TFE3*. And CLTC-TFE3 is the fifth gene fusion involving *TFE3* in pediatric renal cell carcinomas [44]. Another example is *AKT1* (0.69% of cases) in THCA, which is identified by DriverFinder (ranked 30) and contained in CGC; it is a serine/threonine protein kinase and its downstream proteins have been reported to be frequently activated in human cancers [45].

4. Discussion

Cancer is a complex disease and difficult to treat, and distinguishing driver genes from a mass of neutral passenger genes



(a)

(b)

(c)

(d)

Figure 4: The KEGG pathway enrichment of (a) BRCA, (b) HNSC, (c) KIRC, and (d) THCA by significant genes with p values < 0.05 in DriverFinder.

is extremely important to understand the mechanism of the cancer and design targeted treatments. In this study, we introduced a comprehensive framework DriverFinder to identify driver genes by incorporating genomes, transcriptomes, and gene-gene interaction information. We implemented gene-based filtering model to exclude genes that were mutated largely due to random events. The method was applied to four independent cancer datasets from TCGA and the results demonstrated that the power of it across multiple tumor types was mainly better than DriverNet, MUFFINN, and frequency-based methods. In summary, this method has advantages in both filtering random mutated genes and identifying driver genes regardless of their mutation frequencies. We expect that it can also be applied to other complex cancer types.

However, in this work, we only explained the changes of the expression by somatic mutations, although other molecular or genetic changes such as transcription factors, methylations, and microRNAs also affect expression of other genes and play important roles in the development of cancer [46]. Therefore, it is necessary to extend the method so that driver genes can be determined by not only somatic alterations but also other different types of molecular changes. Also, we can extend our aim of identifying drivers by some methods such as machine learning methods [47–51].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (no. 61672037), the Key Project of Anhui Provincial Education Department (no. KJ2017ZD01), and the Anhui Provincial Natural Science Foundation (nos. 1508085QF135 and 1608085MF136).

References

[1] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, and G. M. Mastrogianakis, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

[2] T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker, C. Bell, and R. R. Bernabé, "International network of cancer genome projects," *Nature*, vol. 464, pp. 993–998, 2010.

[3] J. X. Liu, Y. Xu, Y. L. Gao, C. H. Zheng, D. Wang, and Q. Zhu, "A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-seq data," *IEEE/ACM Transactions on Computational Biology Bioinformatics*, vol. 13, no. 2, pp. 392–398, 2016.

[4] Y. X. Wang, J. X. Liu, Y. L. Gao, C. H. Zheng, and J. L. Shang, "Differentially expressed genes selection via Laplacian regularized low-rank representation method," *Computational Biology Chemistry*, 2016.

[5] C. Suo, O. Hrydziuszko, D. Lee et al., "Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival," *Bioinformatics*, vol. 31, no. 16, pp. 2607–2613, 2015.

[6] J. Zhang, S. Zhang, Y. Wang, and X.-S. Zhang, "Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data," *BMC systems biology*, vol. 7, p. S4, 2013.

[7] J. Zhang and S. Zhang, "Discovery of cancer common and specific driver gene sets," *Nucleic Acids Research*, vol. 45, no. 10, article e86, 2017.

[8] J. Foo, L. L. Liu, K. Leder et al., "An evolutionary approach for identifying driver mutations in colorectal cancer," *PLoS Computational Biology*, vol. 11, no. 9, Article ID e1004350, 2015.

[9] R. Tian, M. K. Basu, and E. Capriotti, "Computational methods and resources for the interpretation of genomic variants in cancer," *BMC Genomics*, vol. 16, supplement 8, article S7, 2015.

[10] A. Bashashati, G. Haffari, J. Ding et al., "DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer," *Genome Biology*, vol. 13, article R124, 2012.

[11] Y. Ping, Y. Deng, L. Wang et al., "Identifying core gene modules in glioblastoma based on multilayer factor-mediated dysfunctional regulatory networks through integrating multi-dimensional genomic data," *Nucleic Acids Research*, vol. 43, no. 4, pp. 1997–2007, 2015.

[12] A. Youn and R. Simon, "Identifying cancer driver genes in tumor genome sequencing studies," *Bioinformatics*, vol. 27, no. 2, Article ID btq630, pp. 175–181, 2011.

[13] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, and K. Cibulskis, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, pp. 1069–1075, 2008.

[14] J. Zhang and S. Zhang, "The discovery of mutated driver pathways in cancer: models and algorithms," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. PP, no. 99, 2016.

[15] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature Medicine*, vol. 10, no. 8, pp. 789–799, 2004.

[16] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[17] M. Grechkin, B. A. Logsdon, A. J. Gentles, and S.-I. Lee, "Identifying network perturbation in cancer," *PLoS Computational Biology*, vol. 12, no. 5, Article ID e1004888, 2016.

[18] F. Cheng, J. Zhao, and Z. Zhao, "Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes," *Briefings in Bioinformatics*, vol. 17, no. 4, Article ID bbv068, pp. 642–656, 2016.

[19] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," *Genome Research*, vol. 22, no. 2, pp. 375–385, 2012.

[20] M. D. M. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, "Simultaneous identification of multiple driver pathways in cancer," *PLoS Computational Biology*, vol. 9, no. 5, Article ID e1003054, 2013.

[21] J. Zhao, S. Zhang, L.-Y. Wu, and X.-S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," *Bioinformatics*, vol. 28, no. 22, pp. 2940–2947, 2012.

[22] J. Zhang, L.-Y. Wu, X.-S. Zhang, and S. Zhang, "Discovery of co-occurring driver pathways in cancer," *BMC Bioinformatics*, vol. 15, article 271, 2014.

[23] J. P. Hou and J. Ma, "DawnRank: discovering personalized driver genes in cancer," *Genome Medicine*, vol. 6, article 56, 2014.

[24] P. Jia and Z. Zhao, "VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data," *PLoS Computational Biology*, vol. 10, no. 2, Article ID e1003460, 2014.

[25] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr., and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 340, no. 6127, pp. 1546–1558, 2013.

[26] C. Shyr, M. Tarailo-Graovac, M. Gottlieb, J. J. Y. Lee, C. Van Karnebeek, and W. W. Wasserman, "FLAGS, frequently mutated genes in public exomes," *BMC Medical Genomics*, vol. 7, article 64, 2014.

[27] J. J. Waterfall, E. Arons, R. L. Walker et al., "High prevalence of MAP2K1 mutations in variant and IGHV4-34-expressing hairy-cell leukemias," *Nature Genetics*, vol. 46, pp. 8–10, 2014.

[28] I. P. Gorlov, J.-Y. Yang, J. Byun et al., "How to get the most from microarray data: advice from reverse genomics," *BMC Genomics*, vol. 15, article 223, 2014.

[29] C. Ma, Y. Chen, D. Wilkins, X. Chen, and J. Zhang, "An unsupervised learning approach to find ovarian cancer genes through integration of biological data," *BMC Genomics*, vol. 16, supplement 9, article S3, 2015.

[30] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature*, vol. 461, no. 7261, pp. 218–223, 2009.

[31] P.-J. Wei, D. Zhang, C.-H. Zheng, and J. Xia, "Cancer genes discovery based on integtating transcriptomic data and the impact of gene length," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '16)*, pp. 1265–1268, Shenzhen, China, December 2016.

[32] A. Cho, J. E. Shim, E. Kim, F. Supek, B. Lehner, and I. Lee, "MUFFINN: cancer gene discovery via network analysis of somatic mutation data," *Genome Biology*, vol. 17, article 129, 2016.

[33] K. D. Pruitt, J. Harrow, R. A. Harte et al., "The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes," *Genome Research*, vol. 19, no. 7, pp. 1316–1323, 2009.

[34] T. Bammler, R. P. Beyer, and S. Bhattacharya, "Standardizing global gene expression analysis between laboratories and across platforms," *Nature Methods*, vol. 2, no. 5, pp. 351–356, 2005.

[35] P. A. Futreal, L. Coin, M. Marshall et al., "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[36] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature Protocols*, vol. 4, no. 7, pp. 1073–1082, 2009.

[37] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[38] N. E. Hynes and G. MacDonald, "ErbB receptors and signaling pathways in cancer," *Current Opinion in Cell Biology*, vol. 21, no. 2, pp. 177–184, 2009.

[39] J. Ma, H. Sawai, N. Ochi et al., "PTEN regulate angiogenesis through PI3K/Akt/VEGF signaling pathway in human pancreatic cancer cells," *Molecular and Cellular Biochemistry*, vol. 331, no. 1-2, pp. 161–171, 2009.

[40] M. Cizkova, S. Vacher, D. Meseure et al., "PIK3R1 underexpression is an independent prognostic marker in breast cancer," *BMC Cancer*, vol. 13, article 545, 2013.

[41] M. P. H. M. Jansen, J. W. M. Martens, J. C. A. Helmijr et al., "Cell-free DNA mutations as biomarkers in breast cancer patients receiving tamoxifen," *Oncotarget*, vol. 7, no. 28, pp. 43412–43418, 2016.

[42] Y. Chen, Y. Gao, Y. Tian, and D.-L. Tian, "PRKACB is Downregulated in non-small cell lung cancer and exogenous PRKACB inhibits proliferation and invasion of LTEP-A2 cells," *Oncology Letters*, vol. 5, no. 6, pp. 1803–1808, 2013.

[43] P. Dura, J. Salomon, R. H. M. Te Morsche et al., "High enzyme activity UGT1A1 or low activity UGT1A8 and UGT2B4 genotypes increase esophageal cancer risk," *International Journal of Oncology*, vol. 40, no. 6, pp. 1789–1796, 2012.

[44] P. Argani, M. Y. Lui, J. Couturier, R. Bouvier, J.-C. Fournet, and M. Ladanyi, "A novel CLTC-TFE3 gene fusion in pediatric renal adenocarcinoma with t(X;17) (p11.2;q23)," *Oncogene*, vol. 22, no. 34, pp. 5374–5378, 2003.

[45] D. Lee, I. G. Do, K. Choi, O. S. Chang, K. T. Jang, and D. Choi, "The expression of phospho-AKT1 and phospho-MTOR is associated with a favorable prognosis independent of PTEN expression in intrahepatic cholangiocarcinomas," *Modern Pathology An Official Journal of the United States & Canadian Academy of Pathology Inc*, vol. 25, p. 131, 2011.

[46] B. Amgalan and H. Lee, "DEOD: uncovering dominant effects of cancer-driver genes based on a partial covariance selection method," *Bioinformatics*, vol. 31, no. 15, pp. 2452–2460, 2015.

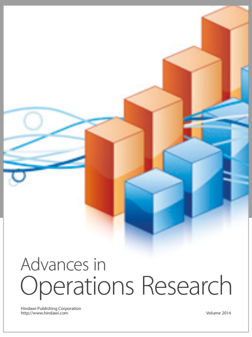
[47] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental learning for ν-support vector regression," *Neural Networks the Official Journal of the International Neural Network Society*, vol. 67, pp. 140–150, 2015.

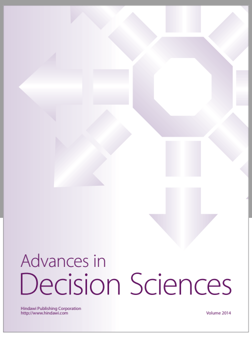
[48] Y. Zheng, J. Byeungwoo, D. Xu, Q. M. J. Wu, and Z. Hui, "Image segmentation by generalized hierarchical fuzzy C-means algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 28, pp. 4024–4028, 2015.

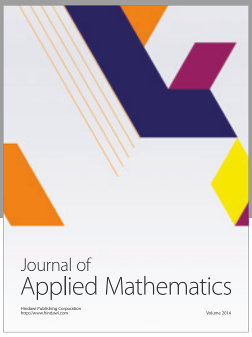
[49] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," *Pattern Recognition*, vol. 40, no. 3, pp. 825–838, 2007.

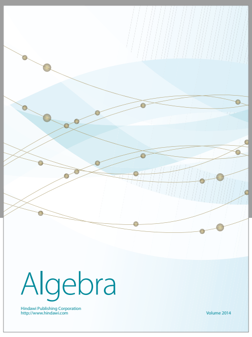
[50] X. Wen, L. Shao, Y. Xue, and W. Fang, "A rapid learning algorithm for vehicle classification," *Information Sciences*, vol. 295, pp. 395–406, 2015.

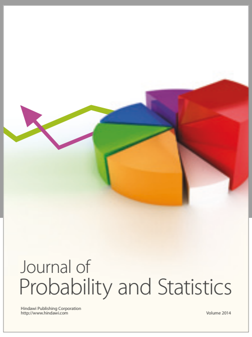
[51] S. G. Ge, J. Xia, W. Sha, and C. H. Zheng, "Cancer subtype discovery based on integrative model of multigenomic data," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. PP, no. 99, p. 1, 2016.


Advances in
Operations Research

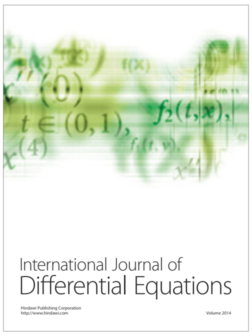

Advances in
Decision Sciences

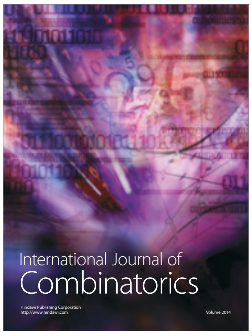

Journal of
Applied Mathematics

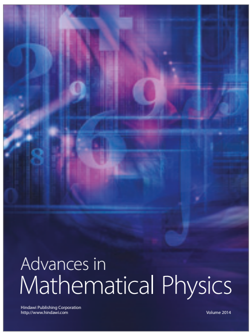

Algebra

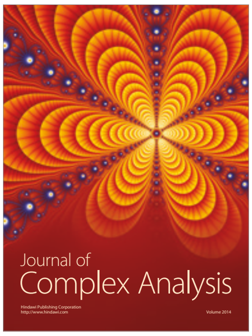

Journal of
Probability and Statistics

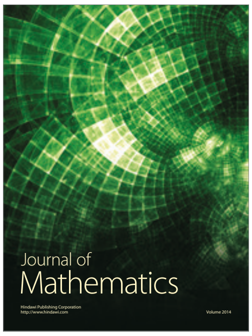
The Scientific
World Journal

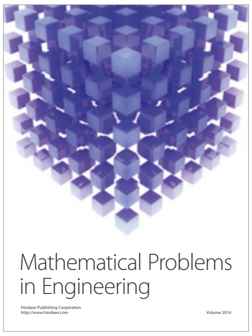

International Journal of
Differential Equations

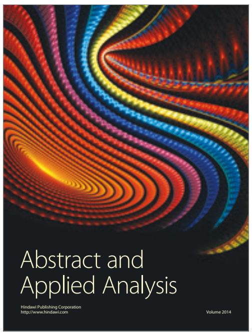

International Journal of
Combinatorics

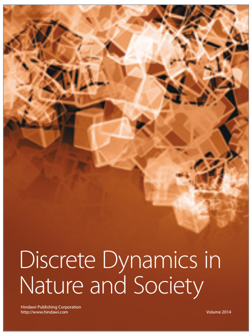

Advances in
Mathematical Physics

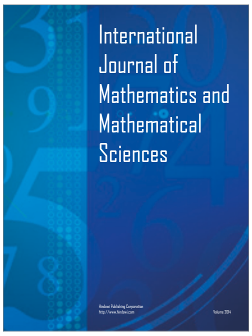

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

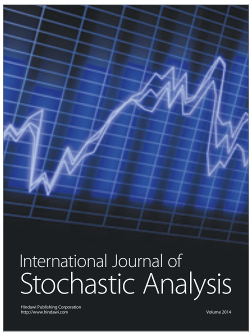

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization