ORIGINAL RESEARCH
published: 08 January 2021
doi: 10.3389/fgene.2020.613033

An Efficient and Easy-to-Use Network-Based Integrative Method of Multi-Omics Data for Cancer Genes Discovery

*Ting Wei[1,2], Botao Fa[1,2], Chengwen Luo[1,2], Luke Johnston[2], Yue Zhang[1,2] and Zhangsheng Yu[1,2]**

[1] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2] SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Chunhou Zheng,
Anhui University, China

Reviewed by:
Junfeng Xia,
Anhui University, China
Zhu-Hong You,
Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences (CAS), China
Jin-Xing Liu,
Qufu Normal University, China

***Correspondence:**
Zhangsheng Yu
yuzhangsheng@sjtu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *01 October 2020*
Accepted: *25 November 2020*
Published: *08 January 2021*

Citation:
Wei T, Fa B, Luo C, Johnston L, Zhang Y and Yu Z (2021) An Efficient and Easy-to-Use Network-Based Integrative Method of Multi-Omics Data for Cancer Genes Discovery. Front. Genet. 11:613033. doi: 10.3389/fgene.2020.613033

Identifying personalized driver genes is essential for discovering critical biomarkers and developing effective personalized therapies of cancers. However, few methods consider weights for different types of mutations and efficiently distinguish driver genes over a larger number of passenger genes. We propose MinNetRank (Minimum used for Network-based Ranking), a new method for prioritizing cancer genes that sets weights for different types of mutations, considers the incoming and outgoing degree of interaction network simultaneously, and uses minimum strategy to integrate multi-omics data. MinNetRank prioritizes cancer genes among multi-omics data for each sample. The sample-specific rankings of genes are then integrated into a population-level ranking. When evaluating the accuracy and robustness of prioritizing driver genes, our method almost always significantly outperforms other methods in terms of precision, F1 score, and partial area under the curve (AUC) on six cancer datasets. Importantly, MinNetRank is efficient in discovering novel driver genes. SP1 is selected as a candidate driver gene only by our method (ranked top three), and SP1 RNA and protein differential expression between tumor and normal samples are statistically significant in liver hepatocellular carcinoma. The top seven genes stratify patients into two subtypes exhibiting statistically significant survival differences in five cancer types. These top seven genes are associated with overall survival, as illustrated by previous researchers. MinNetRank can be very useful for identifying cancer driver genes, and these biologically relevant marker genes are associated with clinical outcome. The R package of MinNetRank is available at https://github.com/weitinging/MinNetRank.

Keywords: multi-omics, network-based methods, cancer gene prediction, driver genes, tumor stratification

INTRODUCTION

Rapid technological advances in high-throughput sequencing have driven the development of omics field. Omics data types include genomics, transcriptomics, proteomics, epigenomics, and metabolomics (Hasin et al., 2017). However, a single type of "omics" only provides limited insights into the biological mechanisms of diseases. Additionally, the different omics data events are somewhat interdependent. An integrative study of multi-omics data contributes to a holistic understanding of the molecular function (Sun and Hu, 2016). An essential question in cancer genomics is distinguishing driver genes, which are causally implicated in oncogenesis, from biologically neutral passenger genes that are immaterial to neoplasia (Greenman et al., 2007). Passenger mutations can become driver mutations (and vice versa) under changing environmental conditions and selection pressures, increasing the complexity of intratumor heterogeneity (Yap et al., 2012). Accumulating evidence suggests that identifying personalized driver genes is essential for the development of effective personalized therapies and realizing the goals of precision medicine (Dagogo-Jack and Shaw, 2018). A critical but challenging step is to incorporate different omics data in a meaningful and efficient way to discover cancer driver genes and elucidate potential causative changes of cancer (Huang et al., 2017). The main approaches for distinguishing driver genes from passenger genes can be divided into frequency-based methods and network-based approaches.

Frequency-based methods estimate the background mutation rate (BMR) representing the rate of random passenger mutations and identify driver genes that harbor significantly more somatic mutations than BMR (Ding et al., 2008; Pon and Marra, 2015). However, accurately estimating BMR is difficult because of the variability among cancer types, among samples of the same cancer type, and between genomes (Pon and Marra, 2015). Subsequent frequency-based methods, such as MuSiC and MutSigCV, have been developed to correct for one or more of these factors (Dees et al., 2012; Lawrence et al., 2013). Somatic mutations are characterized by a small number of frequently mutated genes and many infrequently mutated genes. Moreover, more than 99.9% of the somatic mutations in tumors are passengers (Vogelstein et al., 2013). It is challenging to identify infrequent or rare driver genes by methods based only on mutation frequency.

Network-based approaches have emerged as promising and powerful methods to detect low-frequency and high-frequency mutated driver genes due to their ability to model gene interactions. For network-based approaches, nodes representing genes and edges are links between two genes if there is an interaction between them (Huang et al., 2017). Network-based methods have been successfully applied to many biomedical fields, such as the discovery of mutation subnetwork (Vandin et al., 2011), prediction of drug–target interaction, and cancer gene prioritization (Bashashati et al., 2012; Chen et al., 2012; Yu et al., 2013). HotNet2 uses a network diffusion model to simultaneously assess the frequency of somatic mutation and the local topology of the interaction network and detects significantly mutated subnetworks (Leiserson et al., 2015). Mutations for Functional Impact on Network Neighbors (MUFFIN) is a method for prioritizing cancer genes accounting for mutation frequency of genes and their direct neighbors in functional network (Cho et al., 2016). Both HotNet2 and MUFFIN use mutation data only without integrating other omics data. DawnRank is a single patient approach to rank potential driver genes based on their impact on downstream differential expression genes in the interaction network (Hou and Ma, 2014). NetICS predicts mediator genes affected by proximal upstream-located aberrant genes and proximal downstream-located differentially expressed genes (Dimitrakopoulos et al., 2018). Both DawnRank and NetICS consider only incoming degree or outgoing degree of interaction network for single omics. For example, DawnRank only considers incoming degree for expression data. It is desirable to use incoming and outgoing degree simultaneously. Driver_IRW (Driver genes discovery with Improved Random Walk method) assigns different transition probabilities for different genes of the interaction network (Wei et al., 2020). DeepDriver predicts cancer driver genes based on mutation-based features and gene similarity networks using deep convolutional neural networks (Luo et al., 2019). None of these methods consider the different weights for the different types of mutations; however, the weighting method is essential for sample-specific study. Furthermore, none of these methods investigate the relationship between the top rankings of genes and overall survival. Therefore, we develop a more meaningful and efficient method that considers different weight coefficients for the various types of mutations, simultaneously considers the incoming and outgoing degree of interaction network for single omics, and uses minimum strategy to integrate multi-omics data.

We present a new method called MinNetRank that uses minimum strategy among multi-omics data to prioritize cancer genes (**Figure 1**). The main steps of MinNetRank include (1) single-omics data analysis: calculating mutation relevance scores and expression relevance scores of genes for each sample using network diffusion based on incoming and outgoing degree. We further consider different weight coefficients for the different types of mutations and propose Weighted_MinNetRank. (2) The integration of multi-omics data: calculating the minimum value of mutation relevance score and expression relevance score as an integrated score for each gene in each sample. A higher minimum value reflects a higher mutation relevance score and expression relevance score simultaneously; (3) prioritizing driver genes: aggregating the sample-specific and integrated-score-based rankings of genes into a robust population-level gene ranking.

We apply Weighted_MinNetRank and MinNetRank to analyze five The Cancer Genome Atlas (TCGA) datasets (hepatocellular carcinoma, stomach adenocarcinoma, bladder urothelial carcinoma, lung adenocarcinoma, and skin cutaneous melanoma) and one International Cancer Genome Consortium (ICGC) dataset (hepatocellular carcinoma). We select the top 50 genes of population-level ranking as candidate driver genes. We systematically examine the performance of Weighted_MinNetRank and MinNetRank from three aspects. Firstly, Weighted_MinNetRank and MinNetRank outperform other methods [Mean, Maximum, DawnRank, NetICS, and a commonly used frequency-based method (Freq)] in terms of precision, F1 score, and partial area under the curve (AUC) value of selecting cancer driver genes. Secondly, Weighted_MinNetRank and MinNetRank detect rare and novel candidate driver genes (e.g., SP1 in hepatocellular carcinoma). Finally, the top seven genes can be used as prognostic biomarkers for risk stratification. The survival difference between two

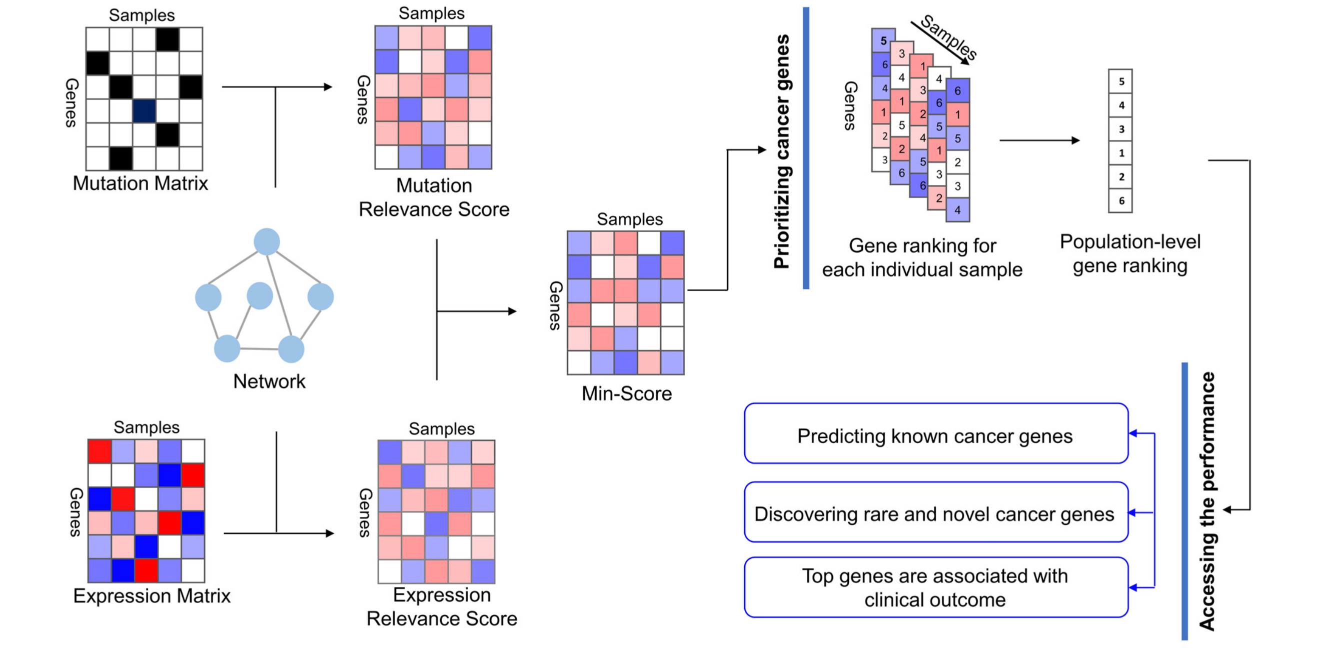
Samples
Genes
Mutation Matrix
Samples
Genes
Mutation Relevance Score
Network
Samples
Genes
Expression Matrix
Samples
Genes
Expression Relevance Score
Samples
Genes
Min-Score
Prioritizing cancer genes
Samples
Genes
5 6 4 1 2 3
3 4 1 5 2 6
1 3 2 4 5 6
4 6 5 1 3 2
6 1 5 2 3 4
Gene ranking for each individual sample
5 4 3 1 2 6
Population-level gene ranking
Accessing the performance
Predicting known cancer genes
Discovering rare and novel cancer genes
Top genes are associated with clinical outcome

FIGURE 1 | Overview of MinNetRank workflow. MinNetRank integrates mutation data and expression data into the interaction network; MinNetRank utilizes minimum strategy to select the candidate driver genes with both high mutation relevance score and high expression relevance score; the sample-specific and integrated-score-based rankings of genes are integrated into the overall rankings. We access the performance in predicting known cancer genes, discovering personalized driver genes, and survival risk stratification of tumor samples.

subtypes (low-risk and high-risk groups) is statistically significant in all six datasets.

RESULTS

We propose a new method (MinNetRank) that uses minimum strategy among multi-omics data to prioritize cancer genes. For comparison, we also add the performance of mean (Mean) and maximum (Maximum) to integrate the mutation data and expression data. All mutations have the same weight for MinNetRank. We further consider different weight coefficients for the different types of mutations (Weighted_MinNetRank). In this study, Weighted_MinNetRank and MinNetRank are compared with other five methods [Mean, Maximum, DawnRank (Hou and Ma, 2014), NetICS (Dimitrakopoulos et al., 2018), and Freq] on five types of cancer (liver hepatocellular carcinoma, stomach adenocarcinoma, lung adenocarcinoma, bladder urothelial carcinoma, and skin cutaneous melanoma). Freq is a simple and common method based only on mutation frequency, which compares the mutation frequency of genes in tumor patient (Dimitrakopoulos et al., 2018; Guo et al., 2018). Weighted_MinNetRank and MinNetRank are an efficient and easy-to-use network-based method for cancer genes discovery by integrating multi-omics data, as shown in the subsequent results.

Overview of MinNetRank

The schematic in **Figure 1** illustrates the three-step procedure of our new method MinNetRank. MinNetRank requires three input files: gene mutations, gene expression for tumor and normal samples, and the interaction network.

Step 1: calculating mutation relevance score and expression relevance score using RWR (Random Walker with Restart) algorithm. The $_{n\times m}$ matrix S^M is the gene mutation status for each sample, where $_n$ is the number of genes, and $_m$ is the number of samples. $S^M_{ik} = 1$ if gene $_i$ is mutated in sample k and $S^M_{ik} = 0$ otherwise. We further consider different weight coefficients for the different types of mutations and supplement a new method (Weighted_MinNetRank). We normalize each column of S^M by $S^M/\mathrm{colSum}(\ S^M)$. We define the $_{n\times m}$ mutation relevance score matrix W^M as multiplication between diffused matrix $_D$ and S^M:

$$W^M = DS^M. \tag{1}$$

The D_{ij} reflects the connectivity between gene i and gene j, and S^M_{ik} reflects the mutation status of gene i in sample k. The product W^M_{ik} is gene i's mutation relevance score in sample k, defined as the proximity of gene i to mutation genes.

Similarly, the $_{n\times m}$ matrix S^E is RNA differential expression score (Absolute value of Log2 Fold-Change, *ALFC*) for each sample. We define the expression relevance score matrix W^E as,

$$W^E = DS^E. \tag{2}$$

Step 2: minimum value of mutation relevance score and expression relevance score. To integrate multi-omics data (gene mutation and expression data), the mutation relevance score and expression relevance score are combined to produce a gene min-score for each sample. The min-score is the minimum value of W_{ik}^{M} and W_{ik}^{E}:

$$W = pmin\left(W^{M}, W^{E}\right). \tag{3}$$

pmin is R function and returns the minimum of the corresponding elements of the two input vectors. W_{ik} is the minimum value of W_{ik}^{M} and $W_{ik}^{E}(i \in 1 \cdots n, k \in 1 \cdots m)$, where $_n$ is the number of genes, and $_m$ is the number of samples. The high score of W_{ik} means that gene *i* is proximal to many mutation genes and differentially expressed genes for each *k*. The minimum value is a meaningful and efficient way to integrate multi-omics data for the following two reasons:

Firstly, the minimum strategy reduces extreme values that may be potential outliers in highly skewed distributions. The probability distribution of W_{*k}^{M} (the mutation relevance scores for genes in sample *k*) and W_{*k}^{E} (the expression relevance scores for genes in sample *k*) is a positively skewed distribution. This means that some genes have extremely high scores. These high scores may be due to the technical noise of high-throughput sequencing and the incomplete interaction network. For example, as shown in **Figure 2**, sample TCGA-BC-A10X has three mutated genes in TCGA-LIHC, and only one gene (*OR2C3*) of these is in the interaction network. The *OR2C3* mutation relevance score in TCGA-BC-A10X is evidently high ($W_{ik}^{M} = 0.48$, $i =$*OR2C3* and $k =$TCGA-BC-A10X) and is ranked 1st. Meanwhile, the *OR2C3* expression relevance score in TCGA-BC-A10X is 3.24-06 and is ranked 8, 221st. Henceforth, the high mutation relevance score needs to be cautiously processed. Lastly, the min-score of *OR2C3* mutation relevance score and expression relevance score is ranked 1, 943rd. *OR2C3* is an olfactory receptor protein and probably is not a potential driver gene (Malnic et al., 2004; Riessland et al., 2017).

Secondly, the minimum ("double high") strategy is necessary to prioritize cancer genes having a higher biological relevance. If one gene has a relatively high mutation relevance score but low expression relevance score (such as *OR2C3* in TCGA-BC-A10X), this gene may not be a potential driver gene since differential gene expression is the downstream events of DNA mutation (Sager, 1997). In the other case, the *SI* expression relevance score in TCGA-DD-AAE2 is ranked 8th ($W_{ik}^{E} = 0.0012$, $i =$*SI*, and $k =$TCGA-DD-AAE2), and the mutation relevance score is ranked last. Only *MGAM* interacts with *SI* in the interaction network, and TCGA-DD-AAE2 has no *SI* or *MGAM* mutation. We hope the candidate driver genes have a high mutation relevance score and high expression relevance score.

MinNetRank used a minimum strategy to integrate multi-omics data (mutation data and expression data). We further investigated which data have the greatest effect on the minimum score. We calculated the proportion of mutation relevance score and expression relevance score in minimum scores for the top 50 candidate cancer genes. The proportion of mutation relevance score was 0.657 in all six datasets, and expression relevance score was 0.347. Mutation relevance score affected the minimum score more.

Step 3: integrating sample-specific rankings of genes into a population-level ranking. We transform the min-scores into rankings, since min-scores indicate the relative importance of each sample's genes. To integrate the sample-specific rankings of genes into a robust population-level ranking, we calculate the sum of per-sample ranking. Each step of MinNetRank is based on single sample analysis, such as using the per-sample network diffusion, calculating the minimum value of mutation relevance score and expression relevance score for each gene in each sample, and transforming min-scores into rankings for each sample. We calculate the sum of per-sample ranking as the population-level ranking.

To perform a systematic comparison of seven methods (Weighted_MinNetRank, MinNetRank, Mean, Maximum, DawnRank, NetICS, and Freq), the 576 genes annotated in cancer gene census (CGC) are used as the gold standard cancer driver gene set, and the genes not in CGC are the negative set. The evaluation metrics (precision, F1 score, and partial AUC value) are based on the top 50 genes of six different datasets (five TCGA datasets and one ICGC dataset). The five TCGA datasets are regarding hepatocellular carcinoma (TCGA-LIHC), stomach adenocarcinoma (TCGA-STAD), bladder urothelial carcinoma (TCGA-BLCA), lung adenocarcinoma (TCGA-LUAD), and skin cutaneous melanoma (TCGA-SKCM), respectively. The one ICGC dataset includes hepatocellular carcinoma data from LIRI-JP (Liver Cancer–RIKEN, JP) project (LIRI-LIHC) (Fa et al., 2019). Skin cutaneous melanoma, lung adenocarcinoma, bladder urothelial carcinoma, and stomach adenocarcinoma have a high mutation burden (Martincorena and Campbell, 2015), and LIHC has two different datasets. Both are common cancer types and pose increasing public concerns. The detailed descriptions of six datasets are provided in **Table 1**. The somatic mutations include non-synonymous simple nucleotide variation (SNV) and insertions and deletions (InDels) in coding regions.

MinNetRank Accurately Predicted Cancer Gene

In general, considering the weights for the different types of mutations (Weighted_MinNetRank) had a better performance than other six methods (MinNetRank, Mean, Maximum, NetICS, DawnRank, and Freq) in all six cancer datasets (TCGA-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, TCGA-SKCM, and LIRI-LIHC). Weighting for the different types of mutations was essential for a personalized analysis. As shown in **Figure 3** (for datasets TCGA-LIHC and LIRI-LIHC), **Supplementary Figure 1** (for datasets TCGA-STAD and TCGA-BLCA), and **Supplementary Figure 2** (for datasets TCGA-LUAD and TCGA-SKCM), Weighted_MinNetRank and MinNetRank achieved a higher precision, F1 score, and AUC in all six datasets, namely, Weighted_MinNetRank and MinNetRank could rank the known gold standard cancer driver genes higher. The AUC of Freq was not calculated as the mutation frequency for some genes were the same.

MinNetRank Robustly Predicted Cancer Gene

The Weighted_MinNetRank and MinNetRank also had the advantage of obtaining robust and stable results using the subset

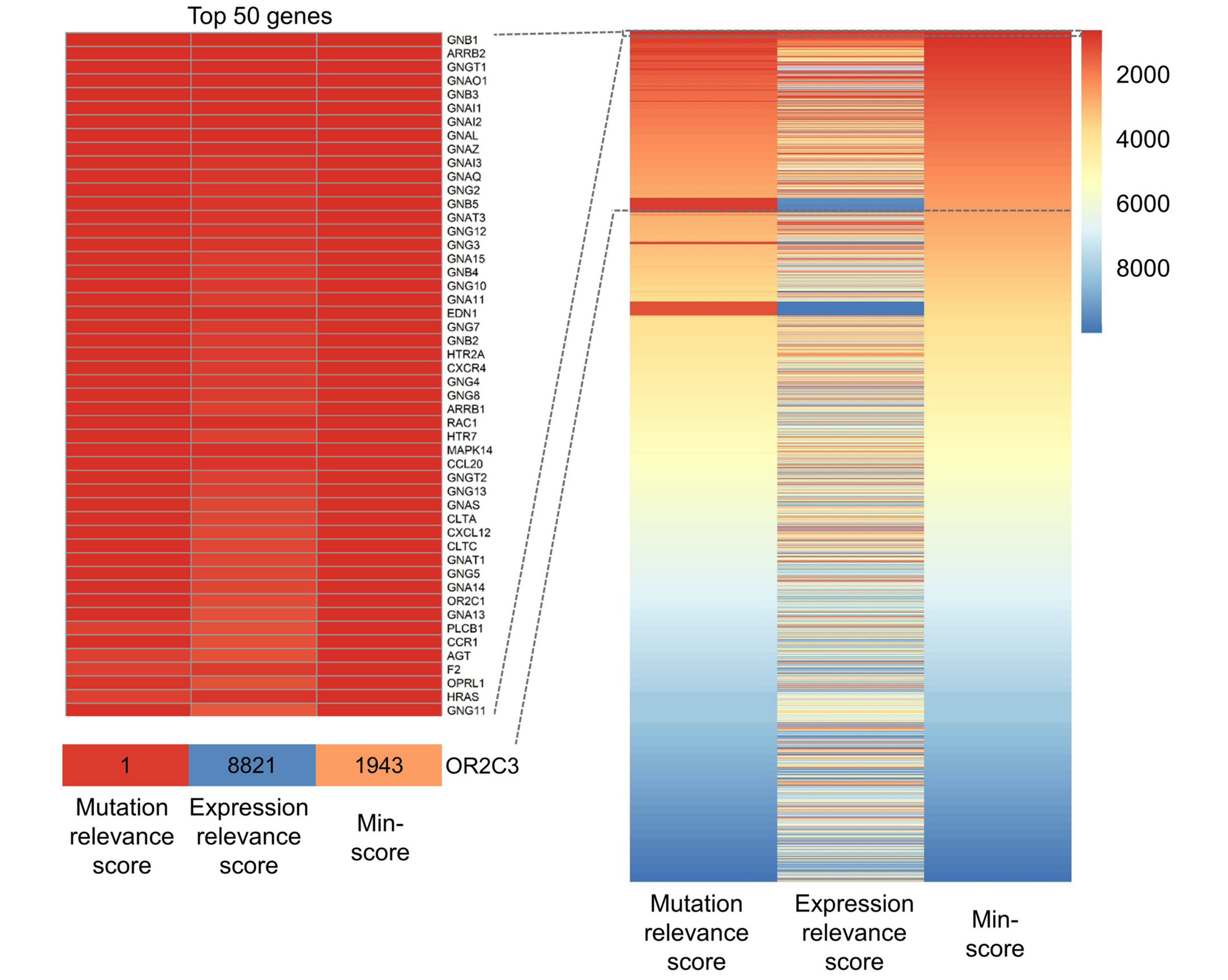

FIGURE 2 | The heatmap of rankings of mutation relevance scores, expression relevance scores, and min-scores. Each row represents a gene. Ranking differences are shown in different colors. Red means high ranking (high score), and blue means low ranking. The rankings of genes are ordered by the rankings of min-scores. The left is the enlarged drawing of the top 50 genes with both high mutation relevance score and high expression relevance score. OR2C3 has high mutation relevance score (red) and low expression relevance score (blue). The ranking of OR2C3 for mutation relevance score, expression relevance score, and min-score in TCGA-BC-A10X is 1, 8,821, and 1,943, respectively.

of samples with different sample sizes. We calculated the mean and standard deviation (SD) of the precision values P (mean precision of the top 50 genes), F1 scores, and partial AUC values after 10 runs. The precision value was proportional to the area under the precision curve (**Figure 3A**). All six methods used the same subset of samples, and the subset of samples was randomly selected from all samples by R. Using the same subset of samples, we compared the results of six methods. The mean of the precision, F1 score, and partial AUC for Weighted_MinNetRank and MinNetRank was higher than other methods, and the SD was smaller [**Figure 4** (for datasets TCGA-LIHC and LIRI-LIHC), **Supplementary Figure 3** (for datasets TCGA-STAD and TCGA-BLCA), and **Supplementary Figure 4** (for datasets TCGA-LUAD and TCGA-SKCM)]. The performance in all six datasets and different sample sizes showed the robustness of our method. Furthermore, Weighted_MinNetRank and MinNetRank still performed well, even with a smaller number of samples.

In order to evaluate the contribution of each part of Weighted_MinNetRank and MinNetRank (calculating the relevance score using both incoming and outgoing degree of the interaction network for single omics, using minimum strategy to integrate multi-omics data, and the different weighted methods), we calculated the precision, F1 score, and partial AUC value of the top 50 candidate cancer genes. We also added network metrics (degree centrality, betweenness centrality, and the mean of degree and betweenness centrality). We needed to calculate the baselines of the network only once, and the results were the same for all datasets. As shown in **Table 2**, Weighted_MinNetRank had a better performance than all other methods in terms of precision, F1 score,

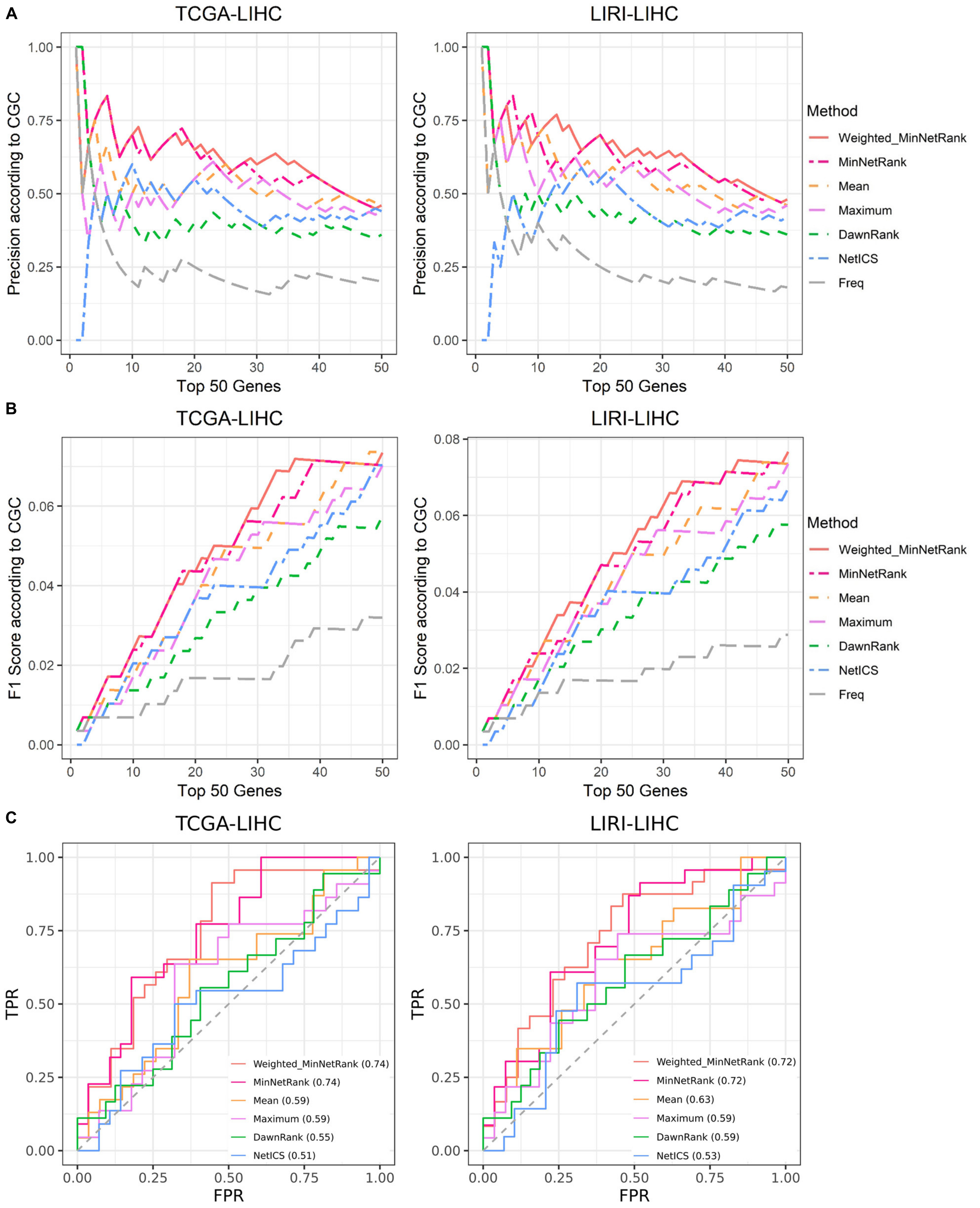

FIGURE 3 | Comparison of precision, F1 score, and AUC for different methods in TCGA-LIHC and LIRI-LIHC datasets. **(A)** The *X*-axis is the top 50 candidate cancer genes, and the *Y*-axis is the precision according to known cancer genes (in CGC). **(B)** The *X*-axis is the top 50 candidate cancer genes, and the *Y*-axis is the F1 score according to known cancer genes. **(C)** The ROC curve of the top 50 candidate cancer genes.

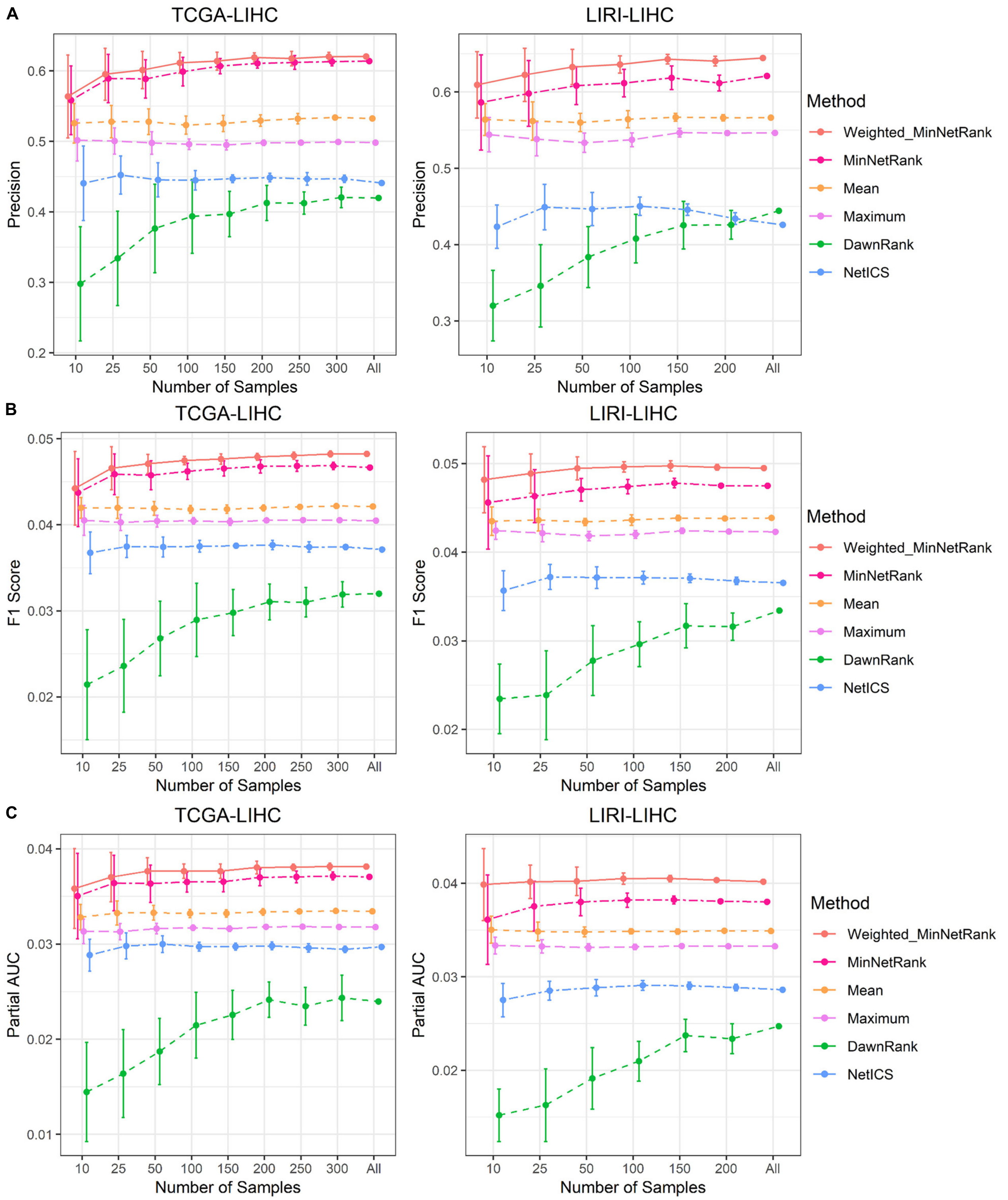

FIGURE 4 | Robustness of results using the subset of samples in TCGA-LIHC and LIRI-LIHC datasets. **(A)** The *X*-axis is the subset of samples, and the *Y*-axis is the mean and SD of the precision values after 10 runs using the subset of samples. **(B)** The *X*-axis is the subset of samples, and the *Y*-axis is the mean and SD of the F1 score after 10 runs using the subset of samples. **(C)** The *X*-axis is the subset of samples, and the *Y*-axis is the mean and SD of the partial AUC after 10 runs using the subset of samples.

TABLE 1 | Six datasets used in MinNetRank.

Datasets	Data type	Samples	Website
TCGA-LIHC	Mutation	363	https://portal.gdc.cancer.gov/projects/TCGA-LIHC
	RNA expression (tumor)	371	
	RNA expression (normal)	50	
LIRI-LIHC	Mutation	258	https://dcc.icgc.org/projects/LIRI-JP
	RNA expression (tumor)	230	
	RNA expression (normal)	197	
TCGA-STAD	Mutation	437	https://portal.gdc.cancer.gov/projects/TCGA-STAD
	RNA expression (tumor)	375	
	RNA expression (normal)	32	
TCGA-BLCA	Mutation	412	https://portal.gdc.cancer.gov/projects/TCGA-BLCA
	RNA expression (tumor)	408	
	RNA expression (normal)	19	
TCGA-LUAD	Mutation	565	https://portal.gdc.cancer.gov/projects/TCGA-LUAD
	RNA expression (tumor)	513	
	RNA expression (normal)	59	
TCGA-SKCM	Mutation	467	https://portal.gdc.cancer.gov/projects/TCGA-SKCM
	RNA expression (tumor)	468	
	RNA expression (normal)	1	

and partial AUC in all six datasets. For weighted methods, Weighted_MinNetRank_PrCID had better performance than PrDSM weighted methods (Weighted_MinNetRank_PrDSM and Weighted_MinNetRank_Filter_PrDSM) in all datasets. There was no significant difference between Weighted_MinNetRank_PrCID and Weighted_MinNetRank. There were some possible reasons for this phenomenon. Firstly, there were many synonymous mutations in all datasets (32,381 synonymous mutations on average); however, the percentage of deleterious synonymous mutations was relatively small (9.76% in the study of PrDSM) (Cheng et al., 2019). Many benign synonymous mutations increased noise. We may need to pre-process the scores of synonymous mutations (Weighted_MinNetRank_Filter_PrDSM performed better than Weighted_MinNetRank_PrDSM). Secondly, the number of missense mutations was the largest, and the number of frameshift mutations was small, so Weighted_MinNetRank weighting for missense mutations had almost the same performance as Weighted_MinNetRank_PrCID weighting for missense mutations and frameshift mutations. LIRI-LIHC dataset did not provide the position information of frameshift mutations in cDNA, so Weighted_MinNetRank_PrCID was not available for LIRI-LIHC dataset.

MinNetRank Discovered Rare and Novel Driver Genes

In addition to obtaining the accurate and robust results, one of the main advantages of MinNetRank was to discover rare and personalized cancer genes. Personalized driver genes could contribute to the development of personalized medicine.

A gene was considered as a rare gene if the gene was mutated in a small number of samples (<5%). For the top 50 candidate driver genes of MinNetRank, the numbers of rare genes in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM were 48 (96%), 48 (96%), 42 (84%), 44 (88%), 48 (96%), and 42 (84%), respectively. Among rare genes, 28 genes (58.33%), 27 genes (56.25%), 27 genes (64.28%), 27 genes (61.36%), 27 genes (56.25%), and 27 genes (64.28%) have not been classified as known cancer gene in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively. We further investigated the rare genes in CGC (gold standard cancer driver gene set), and there were 98.00, 97.95, 85.05, 90.79, 91.73, and 82.11% rare genes in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively. The proportion of rare genes in CGC was high, and the proportion of rare genes for all CGC known cancer genes was approximately the same as the proportion of rare genes for the top 50 candidate driver genes.

MinNetRank also identified novel cancer driver genes that have not been classified as drivers by other methods. Taking an example for *SP1*, *SP1* was considered as a cancer gene only by MinNetRank and was ranked 3rd, 3rd, 3rd, 2nd, 3rd, and 1st in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively (**Supplementary Table 1**). The mutation frequency of *SP1* was 8.26×10^{-3}, 1.60×10^{-2}, 2.43×10^{-2}, 8.85×10^{-3}, and 1.07×10^{-2} (ranked 2903rd, 6393rd, 1599th, 7892nd, and 10330th in terms of the mutation frequency) in TCGA-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively. *SP1* was a zinc finger transcription factor and was reported to be associated with cell differentiation, proliferation, and apoptosis (Beishline and Azizkhan-Clifford, 2015; Safe et al., 2018). Using pathway enrichment analysis, we found that *SP1* was involved in multiple pathways enriched by known cancer genes, such as the transforming growth factor (TGF)-beta signaling pathway and choline metabolism in cancer and breast cancer.

As shown in **Figure 5** (for datasets TCGA-LIHC and LIRI-LIHC), **Supplementary Figure 5** (for datasets TCGA-STAD

TABLE 2 | The performance of each part of MinNetRank according to the precision, F1 score, and partial AUC value.

Metrics	Methods	TCGA-LIHC	LIRI-LIHC	TCGA-STAD	TCGA-BLCA	TCGA-LUAD	TCGA-SKCM
Precision	Weighted_MinNetRank	0.620	0.645	0.602	0.623	0.583	0.533
	Weighted_MinNetRank_PrDSM	0.615	0.633	0.591	0.613	0.573	0.523
	Weighted_MinNetRank_Filter_PrDSM	0.621	0.629	0.599	0.621	0.575	0.528
	Weighted_MinNetRank_PrCID	0.628	–	0.594	0.630	0.580	0.533
	MinNetRank	0.614	0.621	0.585	0.608	0.576	0.515
	MinNetRank (mutation)	0.569	0.576	0.514	0.563	0.445	0.390
	MinNetRank (expression)	0.574	0.580	0.479	0.517	0.512	0.549
	DawnRank	0.420	0.444	0.473	0.586	0.405	0.404
	NetICS	0.441	0.426	0.437	0.453	0.393	0.161
	Mean	0.532	0.566	0.461	0.520	0.414	0.411
	Maximum	0.498	0.546	0.452	0.483	0.405	0.420
	Freq	0.255	0.277	0.249	0.511	0.194	0.149
	Degree centrality	0.189	0.189	0.189	0.189	0.189	0.189
	Betweenness centrality	0.521	0.521	0.521	0.521	0.521	0.521
	Mean of degree and betweenness	0.493	0.493	0.493	0.493	0.493	0.493
F1 score	Weighted_MinNetRank	0.048	0.049	0.046	0.048	0.044	0.042
	Weighted_MinNetRank_PrDSM	0.047	0.049	0.045	0.047	0.044	0.041
	Weighted_MinNetRank_Filter_PrDSM	0.048	0.048	0.046	0.047	0.044	0.041
	Weighted_MinNetRank_PrCID	0.048	–	0.045	0.047	0.044	0.042
	MinNetRank	0.047	0.047	0.045	0.046	0.043	0.041
	MinNetRank (mutation)	0.043	0.044	0.042	0.044	0.039	0.036
	MinNetRank (expression)	0.045	0.046	0.039	0.040	0.040	0.043
	DawnRank	0.032	0.033	0.039	0.043	0.029	0.027
	NetICS	0.037	0.037	0.037	0.037	0.035	0.016
	Mean	0.042	0.044	0.038	0.041	0.037	0.037
	Maximum	0.040	0.042	0.037	0.039	0.037	0.039
	Freq	0.018	0.018	0.017	0.038	0.012	0.011
	Degree centrality	0.013	0.013	0.013	0.013	0.013	0.013
	Betweenness centrality	0.044	0.044	0.044	0.044	0.044	0.044
	Mean of degree and betweenness	0.042	0.042	0.042	0.042	0.042	0.042
Partial AUC	Weighted_MinNetRank	0.038	0.040	0.035	0.038	0.034	0.033
	Weighted_MinNetRank_PrDSM	0.037	0.039	0.034	0.037	0.034	0.032
	Weighted_MinNetRank_Filter_PrDSM	0.038	0.039	0.035	0.038	0.034	0.032
	Weighted_MinNetRank_PrCID	0.038	–	0.034	0.038	0.034	0.033
	MinNetRank	0.037	0.038	0.034	0.037	0.034	0.032
	MinNetRank (mutation)	0.033	0.036	0.032	0.035	0.031	0.029
	MinNetRank (expression)	0.034	0.035	0.031	0.031	0.031	0.034
	DawnRank	0.024	0.025	0.032	0.036	0.022	0.021
	NetICS	0.030	0.029	0.030	0.029	0.029	0.011
	Mean	0.033	0.035	0.031	0.034	0.029	0.031
	Maximum	0.032	0.033	0.029	0.031	0.028	0.030
	Freq	0.011	0.011	0.010	0.026	0.007	0.006
	Degree centrality	0.007	0.007	0.007	0.007	0.007	0.007
	Betweenness centrality	0.035	0.035	0.035	0.035	0.035	0.035
	Mean of degree and betweenness	0.033	0.033	0.033	0.033	0.033	0.033

and TCGA-BLCA), and **Supplementary Figure 6** (for datasets TCGA-LUAD and TCGA-SKCM), *SP1* RNA expression of tumor samples was statistically higher than normal samples in TCGA-LIHC (Wilcoxon Rank-Sum, P = 6.85e-13), LIRI-LIHC (Wilcoxon Rank-Sum, P = 2.2e-16), and TCGA-STAD (Wilcoxon Rank-Sum, P = 5.89e-10). The differential expression was not significant in TCGA-BLCA (Wilcoxon Rank-Sum, P = 0.17), TCGA-LUAD (Wilcoxon Rank-Sum, P = 0.95), and TCGA-SKCM (Wilcoxon Rank-Sum, P = 0.21). We further validated *SP1* expression on the protein level, and the differential protein expression between tumor and normal samples was significant in LIHC (Wilcoxon Signed Rank test, P = 4.14e-13). Only LIHC had protein expression data from CPTAC (The National Cancer Institute's Clinical Proteomic Tumor Analysis

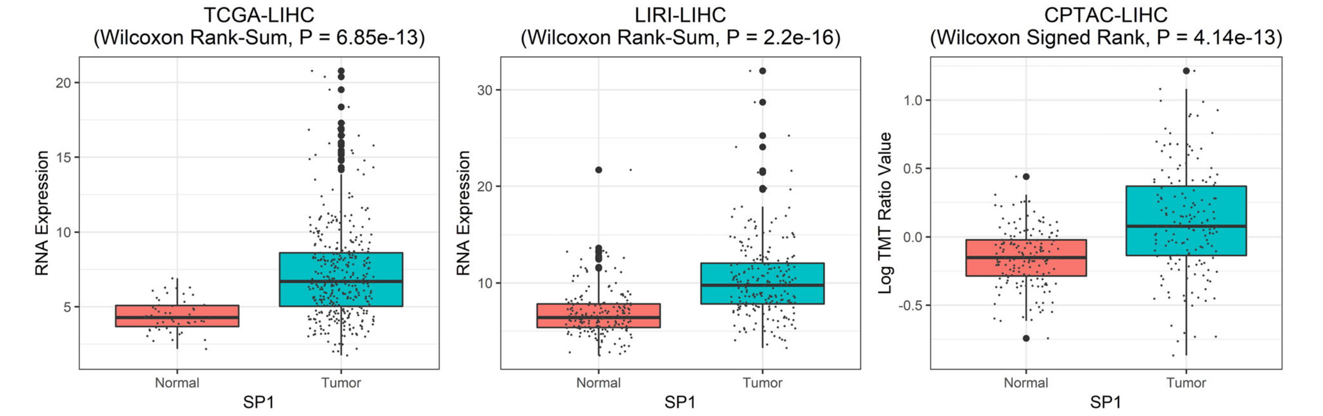

FIGURE 5 | The *SP1* differential expression between tumor and normal samples. From left to right: *SP1* RNA differential expression in TCGA-LIHC dataset, *SP1* RNA differential expression in LIRI-LIHC dataset, and protein differential expression for LIHC from CPTAC dataset.

Consortium) dataset. These results suggested that *SP1* can be the biomarker of hepatocellular carcinoma.

Top Genes of MinNetRank Were Associated With Clinical Outcome

For each dataset, we selected seven genes with top ranking and high SD as biomarkers for tumor stratification (mentioned in the section "Materials and Methods"). We performed unsupervised K-means clustering using obtained biomarkers to assign each patient into either high-risk or low-risk groups. The Kaplan–Meier survival curves of the two groups are well separated, and the log-rank P-values of the survival difference between two groups are 9.21e-04, 1.23e-05, 2.42e-03, 3.75e-03, 9.21e-04, and 4.19e-02 for TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively [**Figure 6** (for datasets TCGA-LIHC and LIRI-LIHC), **Supplementary Figure 7** (for datasets TCGA-STAD and TCGA-BLCA), and **Supplementary Figure 8** (for datasets TCGA-LUAD and TCGA-SKCM)].

In the two liver cancer datasets (TCGA-LIHC and LIRI-LIHC), there were six shared genes (*CTNNB1*, *JUN*, *PIK3R1*, *RAC1*, *SRC*, and *TP53*). All these genes used for tumor stratification are biologically relevant. *CTNNB1* regulated cell growth and adhesion and was predictive for recurrence in aggressive fibromatosis (van Broekhoven et al., 2015). *JUN* (AP-1 Transcription Factor Subunit) participated in regulating a diverse array of cellular processes, including proliferation, apoptosis, differentiation, and survival (Trop-Steinberg and Azar, 2017). *PIK3R1* was a prognostic biomarker for breast cancer (Cizkova et al., 2013). *RAC1* regulated a wide range of cellular events, including the control of cell growth and the activation of protein kinases (Lou et al., 2018). *SRC* was prognostic relevant to colon cancer and rectal cancer (Martínez-Pérez et al., 2017). *TP53* was one of the most frequent alterations and potential prognostic markers in human cancers (Olivier et al., 2010). *GRB2* was the special biomarker for TCGA-LIHC, and *MAPK14* was for LIRI-LIHC. *GRB2* was evaluated as a prognostic marker for lung adenocarcinoma (Toki et al., 2016). *MAPK14* was a member of the MAP kinase family. MAPK pathway regulated cell proliferation, differentiation, and development (Fang and Richardson, 2005). The seven biomarkers are the same in TCGA-STAD and TCGA-BLCA (*CTNNB1*, *GRB2*, *JUN*, *RAC1*, *SP1*, *SRC*, and *TP53*). These seven genes were reported to be related to prognosis (Hang et al., 2016). For TCGA-LUAD and TCGA-SKCM, there were six shared genes (*CTNNB1*, *JUN*, *RAC1*, *SRC*, *TP53*, and *GRB2*). *GNB1* was the special biomarker for TCGA-LUAD, and FYN was for TCGA-SKCM. *FYN* was tyrosine kinases and was an essential molecule in cancer pathogenesis and drug resistance (Elias and Ditzel, 2015). In summary, the top seven genes were associated with clinical outcome and were biologically relevant in all six datasets. These results suggested that MinNetRank could also be a promising method for tumor stratification.

NetICS and DawnRank did not investigate the prognostic value of top genes in cancer. To evaluate the performance of predicting the clinical outcome for different methods, we used the same criterion to choose the top seven genes for each method in six datasets. Compared with NetICS and DawnRank, only Weighted_MinNetRank and MinNetRank obtained a statistically significant survival risk difference between the high-risk and low-risk groups in all six datasets (**Supplementary Table 2**).

DISCUSSION

Extensive genetic heterogeneity exists between tumors of different tissues and between individuals with the same tumor type (Burrell et al., 2013). The personalized mutation profile is the key to advance personalized disease diagnosis and therapy in the clinic (Sheng et al., 2015; Olivier et al., 2019). However, few methods could efficiently prioritize driver genes over many passenger genes in a single patient. The critical challenge facing today is to predict rare and even personalized driver genes with higher accuracy. We develop MinNetRank, an efficient and easy-to-use method that integrates the mutation data, expression data and interaction network to prioritize each sample's driver genes.

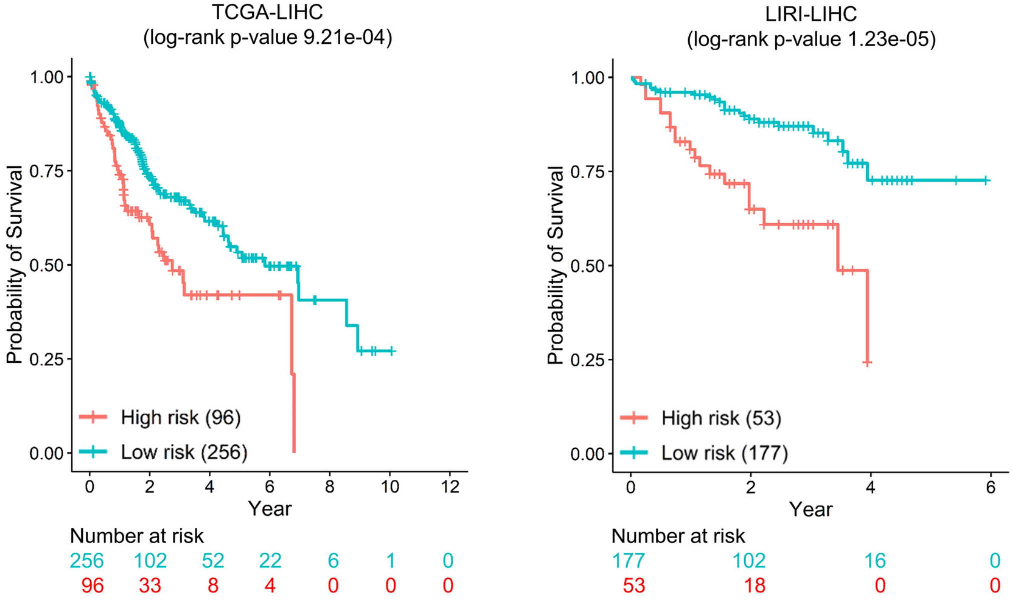

FIGURE 6 | The survival difference between the high-risk group and the low-risk group.

Weighted_MinNetRank further considers the different weights for the different types of mutations.

Weighted_MinNetRank and MinNetRank achieve a higher precision, F1 score, and partial AUC value of prioritizing cancer genes in five TCGA datasets (TCGA-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM). We also utilize an additional liver cancer cohort (LIRI-LIHC) to validate the result of TCGA-LIHC. Better performance in all datasets demonstrates the proposed approach's robustness (**Figure 3** and **Table 2**). We use top candidate driver genes for pathway enrichment analysis and find some signaling pathways previously studied in cancer, such as the Ras signaling pathway and ErbB signaling pathway. Furthermore, we first investigate the relationship between the top seven genes and clinical outcome and find the statistically significant survival difference between the low-risk and high-risk groups in all six datasets only for Weighted_MinNetRank and MinNetRank. The top seven genes are biologically relevant and could be used as biomarkers for survival risk stratification. Accurate outcome prediction is important for personalized cancer therapies in clinical practice, for instance, a low-risk patient can be advised to select a less radical therapy.

We demonstrate that MinNetRank can discover rare and novel cancer genes. Personalized driver genes could contribute to developing personalized diagnosis and therapy. *SP1* is considered a candidate driver gene only by MinNetRank and is ranked top three in all six datasets. The RNA expression of *SP1* is significantly higher in LIHC tumor samples (TCGA-LIHC and LIRI-LIHC datasets) and STAD tumor samples (TCGA-STAD dataset). The differential expression is further validated on the protein level in LIHC. *SP1* is the biomarker for tumor stratification in TCGA-STAD and TCGA-BLCA, and *SP1* RNA expression is associated with survival outcome in TCGA-STAD dataset (Cox proportional hazards model, $P = 0.02$). These results are in accordance with the reports in literatures (Shi and Zhang, 2019). Targeting *SP1* is highly promising strategy in cancer chemotherapy (Vizcaíno et al., 2015).

Using both the incoming and outgoing degree of interaction network, the minimum strategy and weighting for the different types of mutations all contribute to the accuracy and robustness of prioritizing driver genes. Known cancer genes have a higher incoming and outgoing degree, and simultaneously considering incoming and outgoing degree is rational. MinNetRank adopts a minimum strategy to prioritize cancer genes with a high mutation relevance score and high expression relevance score. These enable our method to select more relevant genes and avoid the potential outliers, which are common in high-throughput sequencing technologies due to the positively skewed distributions of mutation and expression relevance scores. Weighting for different types of mutations is essential for sample-specific study and finding personalized driver genes.

There are some limitations to MinNetRank and similar methods. Firstly, MinNetRank largely depends on the interaction network. Although many interaction sources exist, such as experiment, co-expression, and text mining, the interaction network is still incomplete. If the mutation gene or differentially expressed gene is not in the interaction network, this gene would not be used for network diffusion and not be as a candidate cancer gene. Secondly, MinNetRank uses paired tumor and normal samples to calculate *ALFC*; however, TCGA datasets have a limited number of normal samples with expression data. Thirdly, MinNetRank only integrates mutation data and expression data into the interaction network. Besides mutation data, other events, such as miRNA differential expression, epigenetic changes, copy number variation, and structure variation, could also contribute to cancer progression. Differential expression data, including

RNA expression data and protein expression data, could be combined. We may need to improve MinNetRank from two aspects in the future. On one hand, we could integrate the gene co-expression network with the interaction networks (Hou et al., 2019; Wei et al., 2020). We also need to incorporate additional types of omics data (genomics, transcriptomics, proteomics, epigenomics, and images). On the other hand, Weighted_MinNetRank only considers mutations in coding region. We may need to incorporate non-coding mutations. We also need to give weight coefficients for all mutations through multiple techniques.

Integrating different types of omics data is often used to better elucidate the molecular function. However, sound study designs and solid analytical strategies are needed to advance human disease research further. For example, the mean precision of the top 50 cancer genes is 0.61 (MinNetRank) and 0.56 (NetICS) in TCGA-LIHC and 0.61 (MinNetRank) and 0.54 (NetICS) in TCGA-BLCA. The top 50 candidate cancer genes of NetICS used here are from the published paper (Dimitrakopoulos et al., 2018). In this article, NetICS integrates different types of data that include somatic mutation, copy number variation, methylation, miRNA expression, gene expression, and protein expression. Although MinNetRank only focuses on integrating the mutation data and expression data, the mean precision of MinNetRank is still higher than that of NetICS.

CONCLUSION

This article developed a new method (denoted as MinNetRank) by setting weights for different types of mutations and using the minimum strategy to integrate multi-omics for cancer genes discovery. Minimum strategy reduced the influence of extreme scores in highly skewed distributions and was the "double high" strategy to prioritize cancer genes, having a relatively high mutation score and expression score. Different weight coefficients for the different types of mutations contributed to the better performance. We demonstrated our method's accuracy and robustness in prioritizing driver genes on five TCGA datasets and one ICGC dataset. Besides, MinNetRank has the advantage of discovering rare and personalized cancer genes. The top seven candidate driver genes stratified patients into two subtypes (high-risk and low-risk groups) exhibiting significant survival differences and could be used as prognostic biomarkers for survival. Of course, our method has room for improvement. Gene co-expression network and more types of omics data should be incorporated, and different weight coefficients should be considered.

MATERIALS AND METHODS

Dataset

The genes annotated in the CGC can be used to benchmark known cancer genes (Tate et al., 2019). This gold standard known cancer gene set includes 576 genes (July 2019)[1]. Many cancer studies use CGC genes as the benchmark for the evaluation (Bashashati et al., 2012; Hou and Ma, 2014; Bertrand et al., 2015; Wei et al., 2017; Guo et al., 2018).

Interaction Network

We used the interaction network that has been widely used in the related paper (Hou and Ma, 2014; Guo et al., 2018). The interaction network integrated a variety of resources, including the network used in MEMo as well as the up-to-date information from Reactome (Croft et al., 2011; Ciriello et al., 2012), the NCI-Nature Pathway Interaction Database (Schaefer et al., 2009), and KEGG (Kanehisa et al., 2016). The resulting interaction network consisted of 11,648 genes and 211,794 edges. The average degree centrality of interaction network was 34.20, and the average betweenness centrality was 1.58E-04.

MinNetRank

MinNetRank uses an interaction network that could discover cancer driver genes more efficiently (Leiserson et al., 2015). One of the main reasons for this is the high connectivity (high incoming degree and outgoing degree) of known cancer genes in the interaction network. For example, the mean and median of incoming degree for known cancer genes (in CGC) are 36.06 and 17, which are much higher than those of the genes that are not classified as known cancer genes (17.41 and 3, respectively). Also, the mean and median outgoing degree of known cancer genes are 30.37 and 12, which are much higher than those of the genes that are not in CGC (17.66 and 4, respectively). To a certain extent, this is expected since genes with high connectivity could exert a more significant influence on the biological system (Winter et al., 2012). RWR algorithm models how closely related the two genes are and measures both the direct and indirect neighbors of each gene in the interaction network, making it more sensitive for prioritizing cancer driver genes (Dimitrakopoulos et al., 2018). Unlike NetICS and DawnRank, we consider both incoming and outgoing degree of interaction network for single omics.

Diffused Matrix

Let A be the $n \times n$ adjacency matrix of an interaction network where n represents the number of nodes (the number of genes in the interaction network). A is a 0–1 matrix and $a_{ij} = 1$ if there is a directed edge from node j to node i. A' is the transpose of matrix A and $a_{ji} = 1$ if there is a directed edge from node i to node j. We denote $deg_j^{out} = \sum_{i=1}^{N} a_{ij}$ as the outgoing degree of node j or the number of outgoing edges. While $deg_j^{in} = \sum_{i=1}^{N} a_{ji}$ is the incoming degree of node j. MinNetRank considers both the incoming degree and outcoming degree, so we define the normalized adjacency matrix A^{norm} as,

$$A^{norm} = \begin{pmatrix} \frac{a_{11}+a_{11}}{deg_1^{out}+deg_1^{in}} & \cdots & \frac{a_{1n}+a_{n1}}{deg_n^{out}+deg_n^{in}} \\ \vdots & \ddots & \vdots \\ \frac{a_{n1}+a_{1n}}{deg_1^{out}+deg_1^{in}} & \cdots & \frac{a_{nn}+a_{nn}}{deg_n^{out}+deg_n^{in}} \end{pmatrix}. \tag{4}$$

We define the diffused matrix D as,

$$D = \beta \left[I - (1-\beta) A^{norm} \right]^{-1} \tag{5}$$

[1] https://cancer.sanger.ac.uk/census

The value of D_{ij} lies between 0 and 1 and reflects the connectivity between nodes j and i. Higher score means that two genes are more closely related. The restart probability of $\beta(0 \leq \beta \leq 1)$ determines the degree of diffusion, namely, how far the random walker can move in the network. When β=1, there is no diffusion, namely, we do not use the information of the interaction network. When β=0, gene mutation score or differential expression score (see below) diffuses to the whole network. β depends on the interaction network and is independent of any mutation data or expression data. We chose β to balance diffusion and retainment (Leiserson et al., 2015), and β is 0.48 in this study. The diffused matrix $_D$ needs to be computed only once for a given interaction network.

ALFC

For each patient $_k$, we calculate the Absolute value of Log2 Fold-Change ($_{ALFC}$) of gene $_i$ for the paired tumor and normal samples as a differential expression score. The fold change, or relative difference, is widely used to measure differential gene expression (Love et al., 2014). The absolute value of fold change is taken in order to capture both upregulation and downregulation.

$$ALFC_{ik}=\begin{cases} \left|\log_2 \frac{\text{gene } i \text{ expression of tumor sample in patient } k}{\text{gene } i \text{ expression of normal sample in patient } k}\right| & \text{paried tumor and normal samples} \\ \left|\log_2 \frac{\text{gene } i \text{ expression of tumor sample in patient } k}{\text{the mean of gene } i \text{ expression of all normal samples}}\right| & \text{unpaired} \end{cases} \tag{6}$$

Weighted_MinNetRank

Weighted_MinNetRank uses SIFT scores (between 0 and 1) as the weight coefficients for missense mutations and gives the same weight with 1 to other mutations (stop-gain, stop-loss, frameshift, and non-frameshift) (Ng and Henikoff, 2001). Although synonymous mutations do not alter amino acids, some deleterious synonymous mutations play important roles in cancer (Wen et al., 2016). We further incorporate synonymous mutations and use PrDSM scores as the weights for synonymous mutations (Weighted_MinNetRank_PrDSM). We also use PrDSM scores greater than 0.38 as the weights (Weighted_MinNetRank_Filter_PrDSM). If a PrDSM score is greater than 0.308, the corresponding synonymous mutation is considered as deleterious (Cheng et al., 2019). Besides, we use PredCID scores as the weights for frameshift mutations (Weighted_MinNetRank_PrCID) (Yue et al., 2020).

Assessing the Performance in Predicting Known Cancer Genes

In order to assess the performance in predicting known cancer genes, our method (Weighted_MinNetRank and MinNetRank) was compared with NetICS (Dimitrakopoulos et al., 2018), DawnRank (Hou and Ma, 2014), and Freq. The top 50 genes of the population-level ranking were identified as candidate driver genes and compared with the positive genes in CGC. We used the precision, F1 score, and partial AUC value to evaluate the performance. The precision was defined as expression (7) and can be viewed as the measure of exactness. The recall was the percentage of total known cancer genes correctly predicted by MinNetRank. F1 score combined recall and precision using the harmonic mean. There were many more negative genes than positive genes (positives/negatives = 0.052) and even fewer positive genes when we considered cancer type-specific known cancer genes (positives/negatives ≈ 0.0029). It was more informative to use partial AUC, which considered the number of true positives scored higher than the nth highest scoring negatives, measured for all values from 1 to n (Dimitrakopoulos et al., 2018). Precision, F1 score, and partial AUC were based on the top 50 genes.

$$\text{precison} = \frac{(\text{CGC genes}) \cap (\text{Top N predicted driver genes})}{\text{Top N predicted driver genes}}. \tag{7}$$

$$\text{recall} = \frac{(\text{CGC genes}) \cap (\text{Top N predicted driver genes})}{\text{CGC genes}}. \tag{8}$$

$$F1\ Score = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{9}$$

$$AUC_n = \frac{1}{nT}\sum_{i=1}^{n} T_i, \tag{10}$$

where $_T$ was the total number of known cancer genes in CGC, and T_i was the number of positives scored higher than the ithhighest scoring negatives.

Assessing the Robustness Using the Subset of Samples

In order to further compare these methods, we calculated the precision, F1 score, and partial AUC using the subset of samples with different sample sizes. We experimented with sample sizes of $n = 10,\ 25,\ 50{*}1,\ 50{*}2,\ \ldots,\ 50{*}\lceil N/50 \rceil$, and N was the total sample size of multi-omics data. For each sample size, we performed 10 random samples. We defined the precision value P= mean(p_i), where p_i was the precision of top $_i$ candidate cancer gene, $i = 1,\ 2,\ \ldots,\ 50$. The mean and SD of precision value, F1 score, and partial AUC value for 10 runs were used to measure the robustness.

Tumor Stratification

Some papers used gene mutation data and expression data to identify genes that were indicators for survival. Using these biomarkers, patients can be stratified into subtypes (Haider et al., 2014). We further investigated the relationship between the top genes of population-level ranking and patients' survival time. Genes whose expression with a low variation between tumors provided very limited information for tumor stratification (Winter et al., 2012). According to the genes' rankings, we selected the top seven genes with a greater SD of expression than five as biomarkers for each dataset (Winter et al., 2012). Using these seven biomarkers, K-means clustering (unsupervised learning algorithm) assigned each patient to one of the two clusters (high-risk and low-risk groups). The log-rank test was then used to compare the survival differences of the two groups (R survival package).

DATA AVAILABILITY STATEMENT

The mutation data, expression data, and clinical data of the TCGA dataset are available in the TCGA Data Portal (https://portal.gdc.cancer.gov/projects/). Those from LIRI-JP are available in ICGC Data Portal (https://dcc.icgc.org/projects/LIRI-JP). The LIHC protein expression data are from CPTAC Data Portal (https://proteomics.cancer.gov/data-portal). The detail descriptions of these data are provided in **Table 1**. The example data used to demonstrate MinNetRank are available at https://github.com/weitinging/MinNetRank.

AUTHOR CONTRIBUTIONS

ZY and TW: conceptualization, methodology, and validation. TW: software, formal analysis, investigation, and writing–original draft preparation. BF and TW: data curation. ZY, CL, LJ, and TW: writing–review and editing. ZY: supervision and project administration. ZY and YZ: funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China 11671256 and Shanghai Jiao Tong University STAR Grant.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.613033/full#supplementary-material

Supplementary Figure 1 | Comparison of precision and F1 score for different methods in TCGA-STAD and TCGA-BLCA datasets.

Supplementary Figure 2 | Comparison of precision and F1 score for different methods in TCGA-LUAD and TCGA-SKCM datasets.

Supplementary Figure 3 | Robustness of results using the subset of samples in TCGA-STAD and TCGA-BLCA datasets.

Supplementary Figure 4 | Robustness of results using the subset of samples in TCGA-LUAD and TCGA-SKCM datasets.

Supplementary Figure 5 | The SP1 differential expression between tumor and normal in TCGA-STAD and TCGA-BLCA.

Supplementary Figure 6 | The SP1 differential expression between tumor and normal in TCGA-LUAD and TCGA-SKCM.

Supplementary Figure 7 | The survival difference between high-risk group and low-risk group in TCGA-STAD and TCGA-BLCA.

Supplementary Figure 8 | The survival difference between high-risk group and low-risk group in TCGA-LUAD and TCGA-SKCM.

Supplementary Table 1 | Top 50 driver genes of six methods.

Supplementary Table 2 | The log-rank *P*-value of tumor stratification for each method in six datasets.

REFERENCES

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., et al. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* 13:R124. doi: 10.1186/gb-2012-13-12-r124

Beishline, K., and Azizkhan-Clifford, J. (2015). Sp1 and the 'hallmarks of cancer'. *FEBS J.* 282, 224–258. doi: 10.1111/febs.13148

Bertrand, D., Chng, K. R., Sherbaf, F. G., Kiesel, A., Chia, B. K., Sia, Y. Y., et al. (2015). Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.* 43:e44. doi: 10.1093/nar/gku1393

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625

Chen, X., Liu, M. X., and Yan, G. Y. (2012). Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* 8, 1970–1978. doi: 10.1039/c2mb00002d

Cheng, N., Li, M., Zhao, L., Zhang, B., Yang, Y., Zheng, C.-H., et al. (2019). Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Brief. Bioinform.* 21, 970–981. doi: 10.1093/bib/bbz047

Cho, A., Shim, J. E., Kim, E., Supek, F., Lehner, B., and Lee, I. (2016). MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.* 17:129. doi: 10.1186/s13059-016-0989-x

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111

Cizkova, M., Vacher, S., Meseure, D., Trassard, M., Susini, A., Mlcuchova, D., et al. (2013). PIK3R1 underexpression is an independent prognostic marker in breast cancer. *BMC Cancer* 13:545. doi: 10.1186/1471-2407-13-545

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39, D691–D697. doi: 10.1093/nar/gkq1018

Dagogo-Jack, I., and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* 15, 81–94. doi: 10.1038/nrclinonc.2017.166

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111

Dimitrakopoulos, C., Hindupur, S. K., Häfliger, L., Behr, J., Montazeri, H., Hall, M. N., et al. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 34, 2441–2448. doi: 10.1093/bioinformatics/bty148

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455, 1069–1075. doi: 10.1038/nature07423

Elias, D., and Ditzel, H. J. (2015). Fyn is an important molecule in cancer pathogenesis and drug resistance. *Pharmacol. Res.* 100, 250–254. doi: 10.1016/j.phrs.2015.08.010

Fa, B., Luo, C., Tang, Z., Yan, Y., Zhang, Y., and Yu, Z. (2019). Pathway-based biomarker identification with crosstalk analysis for robust prognosis prediction in hepatocellular carcinoma. *EBioMedicine* 44, 250–260. doi: 10.1016/j.ebiom.2019.05.010

Fang, J. Y., and Richardson, B. C. (2005). The MAPK signalling pathways and colorectal cancer. *Lancet Oncol.* 6, 322–327. doi: 10.1016/S1470-2045(05)70168-6

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158. doi: 10.1038/nature05610

Guo, W. F., Zhang, S. W., Liu, L. L., Liu, F., Shi, Q. Q., Zhang, L., et al. (2018). Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics* 34, 1893–1903. doi: 10.1093/bioinformatics/bty006

Haider, S., Wang, J., Nagano, A., Desai, A., Arumugam, P., Dumartin, L., et al. (2014). A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma. *Genome Med.* 6:105. doi: 10.1186/s13073-014-0105-3

Hang, J., Hu, H., Huang, J., Han, T., Zhuo, M., Zhou, Y., et al. (2016). Sp1 and COX2 expression is positively correlated with a poor prognosis in pancreatic ductal adenocarcinoma. *Oncotarget* 7, 28207–28217. doi: 10.18632/oncotarget.8593

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18:83. doi: 10.1186/s13059-017-1215-1

Hou, J. P., and Ma, J. (2014). DawnRank: discovering personalized driver genes in cancer. *Genome Med.* 6:56. doi: 10.1186/s13073-014-0056-8

Hou, M.-X., Gao, Y.-L., Liu, J.-X., Shang, J., Zhu, R., and Yuan, S.-S. (2019). A new method for mining information of co-expression network based on multi-cancers integrated data. *BMC Med. Genomics* 12(Suppl. 7):155. doi: 10.1186/s12920-019-0608-2

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8:84. doi: 10.3389/fgene.2017.00084

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Lou, S., Wang, P., Yang, J., Ma, J., Liu, C., and Zhou, M. (2018). Prognostic and clinicopathological value of Rac1 in cancer survival: evidence from a meta-analysis. *J. Cancer* 9, 2571–2579. doi: 10.7150/jca.24824

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Luo, P., Ding, Y., Lei, X., and Wu, F. X. (2019). deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet.* 10:13. doi: 10.3389/fgene.2019.00013

Malnic, B., Godfrey, P. A., and Buck, L. B. (2004). The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2584–2589. doi: 10.1073/pnas.0307882100

Martincorena, I., and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science* 349, 1483–1489. doi: 10.1126/science.aab4082

Martínez-Pérez, J., Lopez-Calderero, I., Saez, C., Benavent, M., Limon, M. L., Gonzalez-Exposito, R., et al. (2017). Prognostic relevance of Src activation in stage II-III colon cancer. *Hum. Pathol.* 67, 119–125. doi: 10.1016/j.humpath.2017.05.025

Ng, P. C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* 11, 863–874. doi: 10.1101/gr.176601

Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D., and Cox, L. A. (2019). The need for multi-omics biomarker signatures in precision medicine. *Int. J. Mol. Sci.* 20:4781. doi: 10.3390/ijms20194781

Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* 2:a001008. doi: 10.1101/cshperspect.a001008

Pon, J. R., and Marra, M. A. (2015). Driver and passenger mutations in cancer. *Annu. Rev. Pathol.* 10, 25–50. doi: 10.1146/annurev-pathol-012414-040312

Riessland, M., Kaczmarek, A., Schneider, S., Swoboda, K. J., Lohr, H., Bradler, C., et al. (2017). Neurocalcin delta suppression protects against spinal muscular atrophy in humans and across species by restoring impaired endocytosis. *Am. J. Hum. Genet.* 100, 297–315. doi: 10.1016/j.ajhg.2017.01.005

Safe, S., Abbruzzese, J., Abdelrahim, M., and Hedrick, E. (2018). Specificity protein transcription factors and cancer: opportunities for drug development. *Cancer Prev. Res. (Phila)* 11, 371–382. doi: 10.1158/1940-6207.CAPR-17-0407

Sager, R. (1997). Expression genetics in cancer: shifting the focus from DNA to RNA. *Proc. Natl. Acad. Sci. U.S.A.* 94, 952–955. doi: 10.1073/pnas.94.3.952

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, D674–D679. doi: 10.1093/nar/gkn653

Sheng, J., Li, F., and Wong, S. T. (2015). Optimal drug prediction from personal genomics profiles. *IEEE J. Biomed. Health Inform.* 19, 1264–1270. doi: 10.1109/JBHI.2015.2412522

Shi, S., and Zhang, Z. G. (2019). Role of Sp1 expression in gastric cancer: a meta-analysis and bioinformatics analysis. *Oncol. Lett.* 18, 4126–4135. doi: 10.3892/ol.2019.10775

Sun, Y. V., and Hu, Y. J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* 93, 147–190. doi: 10.1016/bs.adgen.2015.11.004

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015

Toki, M. I., Carvajal-Hausdorf, D. E., Altan, M., McLaughlin, J., Henick, B., Schalper, K. A., et al. (2016). EGFR-GRB2 protein colocalization is a prognostic factor unrelated to overall EGFR expression or EGFR mutation in lung adenocarcinoma. *J. Thorac. Oncol.* 11, 1901–1911. doi: 10.1016/j.jtho.2016.06.025

Trop-Steinberg, S., and Azar, Y. (2017). AP-1 expression and its clinical relevance in immune disorders and cancer. *Am. J. Med. Sci.* 353, 474–483. doi: 10.1016/j.amjms.2017.01.019

van Broekhoven, D. L., Verhoef, C., Grünhagen, D. J., van Gorp, J. M., den Bakker, M. A., Hinrichs, J. W., et al. (2015). Prognostic value of CTNNB1 gene mutation in primary sporadic aggressive fibromatosis. *Ann. Surg. Oncol.* 22, 1464–1470. doi: 10.1245/s10434-014-4156-x

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265

Vizcaíno, C., Mansilla, S., and Portugal, J. (2015). Sp1 transcription factor: a long-standing target in cancer chemotherapy. *Pharmacol. Ther.* 152, 111–124. doi: 10.1016/j.pharmthera.2015.05.008

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Wei, P. J., Wu, F. X., Xia, J., Su, Y., Wang, J., and Zheng, C. H. (2020). Prioritizing cancer genes based on an improved random walk method. *Front. Genet.* 11:377. doi: 10.3389/fgene.2020.00377

Wei, P.-J., Zhang, D., Li, H.-T., Xia, J., and Zheng, C.-H. (2017). DriverFinder: a gene length-based network method to identify cancer driver genes. *Complexity* 2017:4826206. doi: 10.1155/2017/4826206

Wen, P., Xiao, P., and Xia, J. (2016). dbDSM: a manually curated database for deleterious synonymous mutations. *Bioinformatics* 32, 1914–1916. doi: 10.1093/bioinformatics/btw086

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knosel, T., et al. (2012). Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol.* 8:e1002511. doi: 10.1371/journal.pcbi.1002511

Yap, T. A., Gerlinger, M., Futreal, P. A., Pusztai, L., and Swanton, C. (2012). Intratumor heterogeneity: seeing the wood for the trees. *Sci. Transl. Med.* 4:127s110. doi: 10.1126/scitranslmed.3003854

Yu, D., Kim, M., Xiao, G., and Hwang, T. H. (2013). Review of biological network data and its applications. *Genomics Inform.* 11, 200–210. doi: 10.5808/GI.2013.11.4.200

Yue, Z., Chu, X., and Xia, J. (2020). PredCID: prediction of driver frameshift indels in human cancer. *Brief. Bioinform.* bbaa119. doi: 10.1093/bib/bbaa119

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.