Heliyon 8 (2022) e09616

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Multi-omics analysis to screen potential therapeutic biomarkers for anti-cancer compounds

Ruxue Li [a,*,1], Wuai Zhou [b,1]

[a] School of Nursing, Beijing University of Chinese Medicine, Beijing, China
[b] Department of Automation, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
Multi-omics analysis
Traditional Chinese medicine
Biomarker

ABSTRACT

Discover potential biomarkers of the response for anti-cancer therapies, including traditional Chinese medicine (TCM), is a critical but much different task in the field of cancer research. Based on accumulated data and sophisticated methods, multi-omics analysis provides a feasible strategy for the discovery of potential therapeutic biomarkers. Here, we screened the potential therapeutic biomarkers for anti-cancer compounds in TCM through multi-omics data analysis. Firstly, compounds in TCM were collected from the public databases. Then, the molecules that those compounds can intervene on cell lines were carefully filtered out from existing drug bioactivity datasets. Finally, multi-omics analysis including gene mutation analysis, differential expression gene analysis, copy number variation analysis and clinical survival analysis for pan-cancer were conducted to screen potential therapeutic biomarkers for compounds in TCM. 13 molecules of compounds in TCM namely ERBB2, MYC, FLT4, TEK, GLI1, TOP2A, PDE10A, SLC6A3, GPR55, TERT, EGFR, KCNA3 and HDAC4 are differentially expressed, high frequently mutated, obtain high copy number variation rate and also significant in survival, are considered as the potential therapeutic biomarkers.

1. Introduction

Cancer is a multifactorial disease which leads to approximately 1,918,030 new cases and 609,360 deaths in the United States alone in 2022 [1]. The treatment of cancer developed greatly, such as surgery, chemotherapy, radiotherapy and immunotherapy. A broad range of complementary and alternative medicine interventions are often favored and are an appropriate option along with or even potentially instead of standard anti-cancer therapy [2]. Safe and beneficial complementary therapies should be integrated into regular cancer care to improve patient quality of life and outcome [3]. According to the nationwide survey of coverage of the urban basic medical insurance for health service in China from 2008 to 2010, 42.4% cancer patients have used the anti-cancer Chinese patent medicines [4]. With increasing scientific evidence in biological, chemical, and medical research, as well as clinical trials, the use of traditional Chinese medicine (TCM) in cancer treatment is gradually being recognized as a complementary and alternative therapy all over the world [5]. Actually, TCM for softening hardness to dissipate stagnation has achieved much progress in treatment of malignant tumors [6]. However, the biomarkers of anti-cancer compounds in TCM are still unclear, which limits the innovation and development of complementary and alternative medicine including TCM. Therefore, it is necessary to dissect cancer-associated genes and identify potential therapeutic biomarkers for anti-cancer compounds in TCM.

With the development of sequencing technology, omics data, e.g., gene mutation data, copy number variation data, gene expression data and clinical data, accumulated greatly, which strongly promotes the discovery of therapeutic biomarkers. The Cancer Genome Atlas (TCGA) is a typical public funded project that aims to discover cancer-causing genome alterations in over 30 human tumors through integrated multi-dimensional analyses [7]. Meanwhile, the Gene Expression Omnibus (GEO) provides a friendly academic community to share high-throughput gene expression data generated mostly by microarray technology [8, 9]. Compared with TCGA, GEO can provide time series data reflecting different stages of disease, so as to dynamically reflect the occurrence and development of disease.

Meanwhile, anti-cancer drug sensitivity data also provides a basis for the discovery of drug therapeutic biomarkers. The Cancer Cell Line Encyclopedia 1 (CCLE) [10] and Cancer Genome Project (CGP) [11] characterize genomes, mRNA expression, and anti-cancer drug

* Corresponding author.
E-mail address: liruxue_tp@163.com (R. Li).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2022.e09616
Received 7 December 2021; Received in revised form 26 April 2022; Accepted 26 May 2022

dose–responses across large numbers of cell lines, promoting the relationship cellular biochemical context to drug sensitivity. Especially, ChEMBL offers a large-scale bioactivity for drug discovery [12], including the compounds in TCM. Actually, public omics data analysis has contributed to the comprehensive and integrative genomic and molecular characterization of various cancers [13, 14, 15, 16, 17, 18], as well as drug target discovery and biomarkers [19, 20]. These progresses provide a new sight for us to find potential therapeutic biomarkers of anti-cancer compounds in TCM from the perspective of multi-omics analysis and many studies have demonstrated the advantages of this strategy. Li et al. performed multi-platform omics analysis of serial plasma and urine samples collected from patients during the course of COVID-19 [19]. By analyzing these omics data, they revealed several potential therapeutic targets. Further, they chose 25 important molecular signatures as potential biomarkers for the prediction of disease severity. They demonstrated that omics data proposed not only potential therapeutic targets, but also biomarkers for understanding the pathogenesis of severe COVID-19. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification [21]. Yang et al. analyzed 117 primary glioblastoma patients' data that contained SNP, DNA copy, DNA methylation, mRNA expression, and clinical information [22]. They finally divided patients into HX-1 and HX-2 according the molecular feature. Compared to HX-1 subtype, the HX-2 subtype was identified with higher gene co-occurring events, tumor mutation burden, and poor median overall survival. Thus, HX-1 and HX-2 subtypes may make sense as the potential prognostic biomarkers for patients with glioblastoma. The LinkedOmics database, which is the first multi-omics database that integrates mass spectrometry (MS)-based global proteomics data generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) on selected TCGA tumor samples, contains multi-omics data and clinical data for 32 cancer types and a total of 11,158 patients from TCGA [23]. Amjad et al. employed three well-known biomarker identification methods (i.e., ClusterOne, MCODE, and BioDiscML) to identify the potential breast cancer biomarkers using omics data. They finally concluded that the descriptive values of gene biomarkers in terms of their biological aspects that have been determined by a given methodology and the predictive power of the models developed based on the identified gene biomarkers should be considered simultaneously while validating the biomarker identification approaches [24].

In this study, we screened the potential therapeutic biomarkers for anti-cancer compounds in TCM through multi-omics data analysis. Compounds in TCM were firstly collected from the public database, and then the targets that those compounds can intervene on cell lines were obtained from ChEMBL, GDSC [25] and DrugBank [26]. Finally, gene mutation analysis, differential expression gene (DEG) analysis, copy number variation (CNV) analysis and clinical survival analysis for TCGA pan-cancer were conducted to screen potential therapeutic biomarkers for those compounds. The workflow is shown in Figure 1.

2. Results

2.1. Collection of compounds in TCM and corresponding targets

The compounds in TCM and their corresponding targets were collected from public databases. Total 14522 compounds were collected

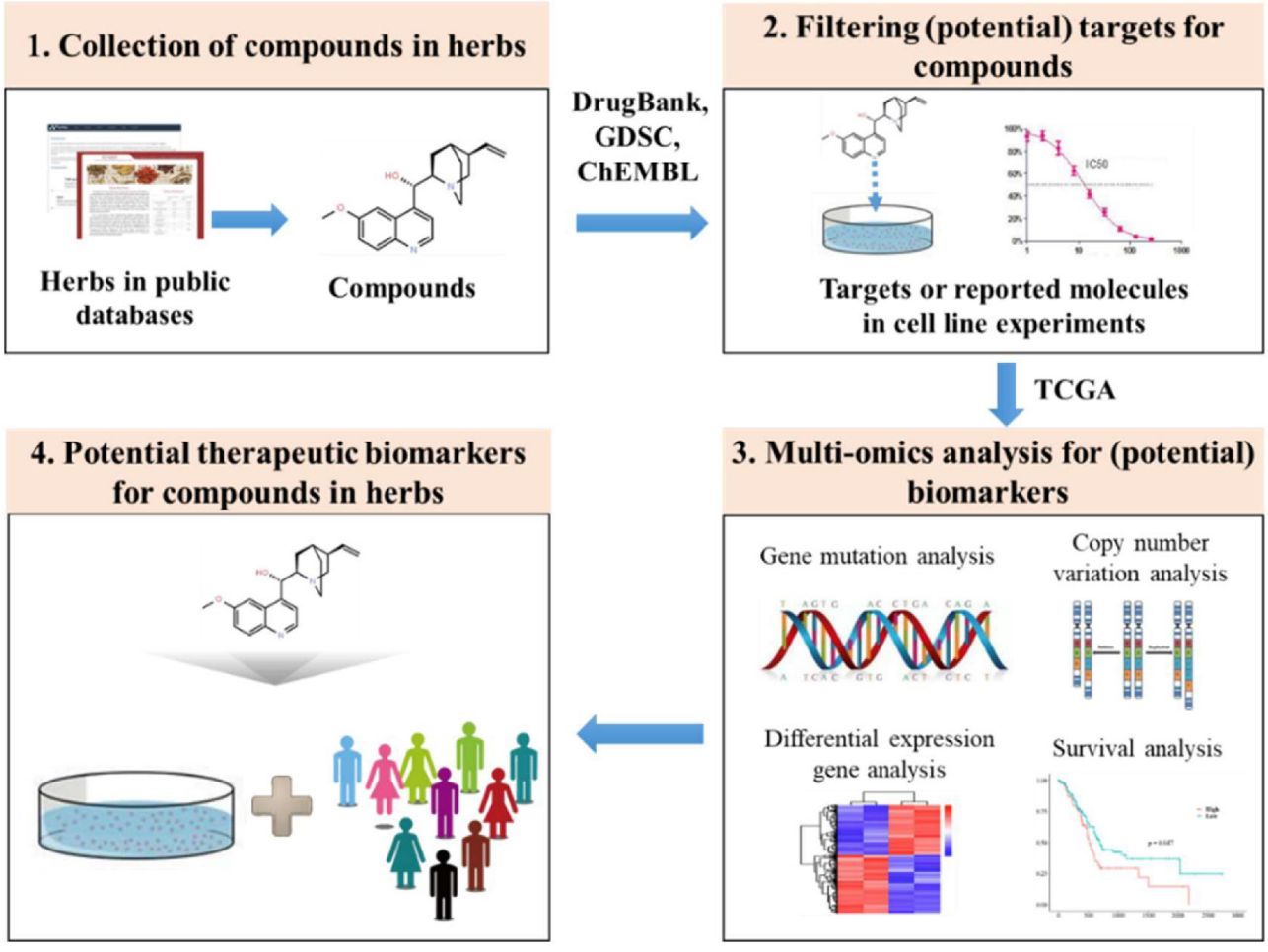

Figure 1. Workflow of this study.

from TCMID [27], SymMap [28] and HIT [29] (Figure 2A). Among those compounds, 792 bioactivity records for 103 compounds were filtered out from ChEMBL, resulting 322 molecules for those 103 compounds. The IC_{50} of these 792 bioactivity records mainly range from 0.018nM to 54.6nM (Figure 2B). The molecular weight of the 103 compounds mainly range from 200 g/mol to 600 g/mol (Figure 2C). The cells mainly come from embryonic kidney fibroblasts, cervical adenocarcinoma cells and ovarian (Figure 2D). The top 3 cells are HEK293, HeLa and Sf9 (Figure 2E). More than 180 targets only relate to only 1 compound, approximately 100 targets relate to 2 compounds, and approximately 25 targets relate to 3 compounds. The results above show that a large amount of bioactivity data of compounds in TCM have been accumulated, and these studies involve a variety of diseases and cells. However, that most of the compounds tend to be associated with only a few targets, which follows the characteristics of power law distribution (Figure 2F). EGFR, ERBB, ERBB1 and HER1 get the most number of compounds in the 792 bioactivity records (Figure 2G), while compound ellagic acid gets the most number of targets (Figure 2H). EGFR is also the target of ellagic acid, and the expression of EGFR is very sensitive to ellagic acid (Figure 2I).

2.2. Enrichment analysis of targets of compounds in TCM

We conducted enrichment analysis of targets of compounds in TCM and the significant enriched KEGG, GO biological processes (BP), cellular components (CC) and molecular functions (MF) terms were reserved (Figure 3). It seems that EGFR related terms are closely related to targets of compounds in TCM. For example, EGFR tyrosine kinase inhibitor resistance is significant enriched in KEGG (adj.P = 3.62e-07) and the many KEGG signaling pathways crosstalk with EGFR related terms are also enriched, such as MAPK signaling, PI3K-Akt signaling and etc (Figure 3A), ERK1 and ERK2 cascade (adj.P = 4.88e-08), and positive regulation of ERK1 and ERK2 cascade (adj.P = 6.90e-08) (Figure 3B).

Meanwhile, as we are concerning the anti-cancer effect of TCM, we especially analyzed the cancer related terms in the enrichment result. KEGG Apoptosis (adj.P = 2.57e-07), Cell cycle (adj.P = 1.63e-05) (Figure 3A), GO BP regulation of epithelial cell proliferation (adj.P = 6.90e-08) (Figure 3B) and GO CC cyclin-dependent protein kinase holoenzyme complex (adj.P = 1.54e-04) (Figure 3C) are all cancer related. The results demonstrated that there is potential to screen therapeutic biomarkers from those targets recorded in ChEMBL.

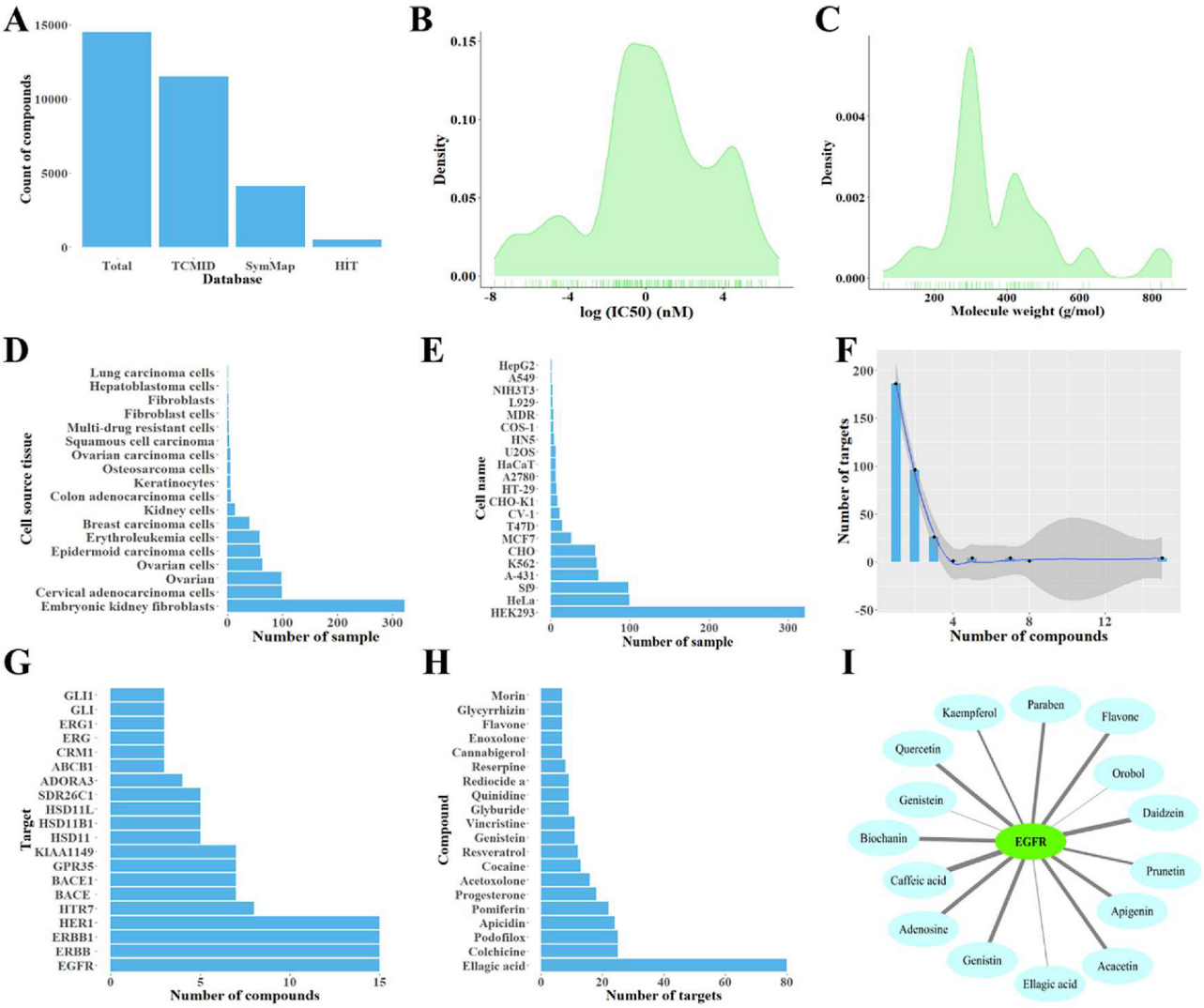

Figure 2. Collection of compounds in TCM and the corresponding targets. (A) The number of total compounds and the number of compounds in each database; (B) Distribution of the IC_{50} of the 792 bioactivity records documented in ChEMBL; (C) Distribution of the molecule weight (g/mol) of the 103 compounds in TCM; (D) Cell source tissue of the 792 bioactivity records documented in ChEMBL; (E) Cell of the 792 bioactivity records documented in ChEMBL; (F) Distribution of the number of targets of compounds documented in ChEMBL; (G) The number of compounds for targets in the 792 bioactivity records; (H) The number of targets for compounds in the 792 bioactivity records; (I) Compound-target network for EGFR. The width of the edge represents the IC_{50}, the larger width of the edge represents the larger IC_{50}, and vice versa.

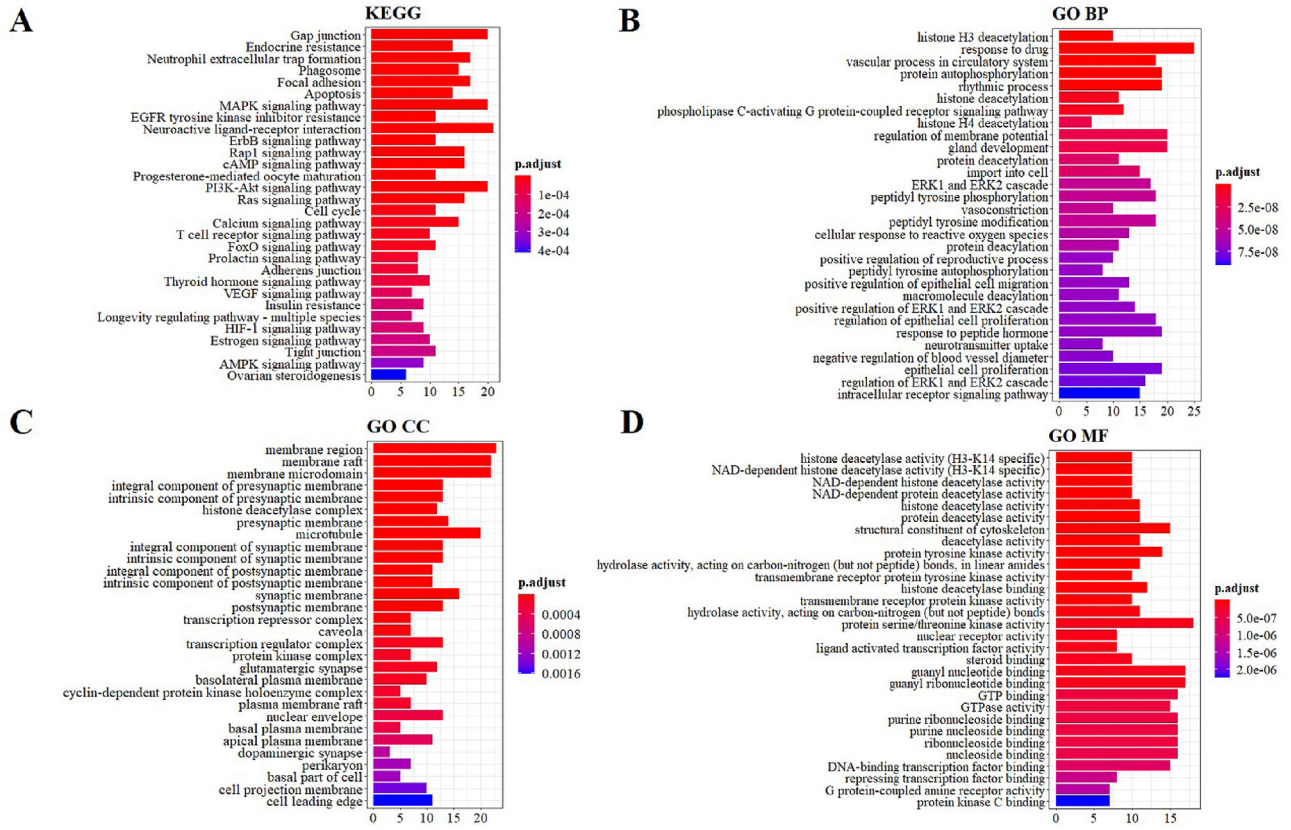

Figure 3. Enrichment analysis of targets of compounds in TCM. (A) Enriched KEGG terms; (B) Enriched GO biological processes terms; (C) Enriched GO cellular components terms; (D) Enriched GO molecular functions terms.

2.3. Compounds in TCM tend to target on genes that weakly differentially expressed

DEG analysis was performed using the R package TCGAbiolinks [30, 31, 32]. The analysis was performed for 23 TCGA cancers LUSC, CHOL, GBM, KICH, UCEC, COAD, KIRP, KIRC, READ, LUAD, BRCA, LIHC, HNSC, BLCA, STAD, PCPG, THCA, PRAD, CESC, ESCA, SARC, PAAD and THYM. The genes were considered differential expressed at a false discovery rate (FDR) < 0.01, and abs $(\log_2^{FC}) \geq 1$ as a cut-off.

The DEGs in each cancer are shown in Figure 4. Many cancers are observed to both have upregulated and downregulated genes (Figure 4A). Based on the statistic of the number of DEGs, LUSC and GBM have the most downregulated genes, whereas LUSC have the most upregulated genes. The upregulated and downregulated genes are helpful for us to study the effect of TCM and its compounds on the expression of them in tumor models. The top DEGs found in no less than 18 cancers are listed in Figure 4B. The most upregulated genes include MCM10, PRC1, BUB1, KIF20A, IQGAP3, KIF4A, DTL, SPC24, UHRF1, FAM111B, CDKN3, CDC45, CKAP2L, CDC25C, CLSPN, NEIL3, RDM1, GPR19, PPEF1, TOP2A and TERT. On the other hand, the most downregulated genes include CRHBP, DPT and TNXB. We also note that many genes are only upregulated or only downregulated across many cancers, such as MCM10, PRC1, BUB1, KIF20A, IQGAP3, KIF4A, DTL, SPC24, UHRF1, FAM111B, CDKN3, CDC45, CKAP2L, CDC25C, CLSPN, NEIL3, RDM1, GPR19, TOP2A and TERT are only upregulated in many cancers, while CRHBP, DPT and TNXB are only downregulated in many cancers. We speculate that genes with this characteristic may contribute to the classification of cancer, and deserve our attention in the further analysis.

In all cancers, the number of upregulated DEGs is more than that of the downregulated DEGs. Meanwhile, we sorted the DEGs according to the sum of upregulated and downregulated cancers. Compared with the compounds in TCM, targets of western anti-cancer drugs cover more DEGs before the top 2050. But after top 2050, the targets of compounds in TCM cover more DEGs than western anti-cancer drug. However, the coverage of compounds in TCM and western anti-cancer drugs are all lower than the that of the known cancer genes (KCGs) (Figure 4C and D). This indicates that compounds in TCM are more likely to target genes weakly differential expressed. In TCM theory, a TCM formulae may regulate the disease-related network by "tiny and multiple effects", and thus lead to a "emerging" effect [33]. This attribute of TCM is consistent with the that the TCM are likely to target DEGs after 2050.

2.4. CNV analysis reveals that targets of compounds in TCM play similar role as targets of western anti-cancer drug and KCGs

Copy number variation (CNV) is a major contribution to the genome variability among individuals, which alter the diploid status of DNA. CNV includes deletions and duplications. CNV analysis was performed for all TCGA cancers. The genes were sorted by the CNV rate in descending order. It is estimated that 4.8–9.5% of the human genome contributes to CNV depending on the level of stringency of the human genome CNV map [34]. As a reference, the top 1% genes for each cancer type were reserved as the genes with high CNV rate and involved in the further analysis. In all TCGA cancers, we find that the frequency of deletion is significantly higher than that of amplification (Figure 5A). However, the KCGs and targets of anti-cancer western drugs are amplified more often, although not significantly (Figures 5B and 5C). While, the targets of compounds in TCM obtain more amplification significantly than deletion (Figure 5D).

27 targets of compounds in TCM MYC, PTK2, CCND1, CNR1, FLT4, DRD1, ADRA1B, KCNA3, EGFR, TEK, CDK4, GLI1, TERT, GPR35, SLC6A3, KCNH2, TUBB2A, TUBB2B, GPR55, CNR2, HDAC4, TUBA4A, ERBB2, TUBB4B, TOP2A, PDE10A and NFKBIA are found to have high CNV rate in 16 cancers PAAD, ESCA, PRAD, KIRC, PCPG, HNSC, GBM,

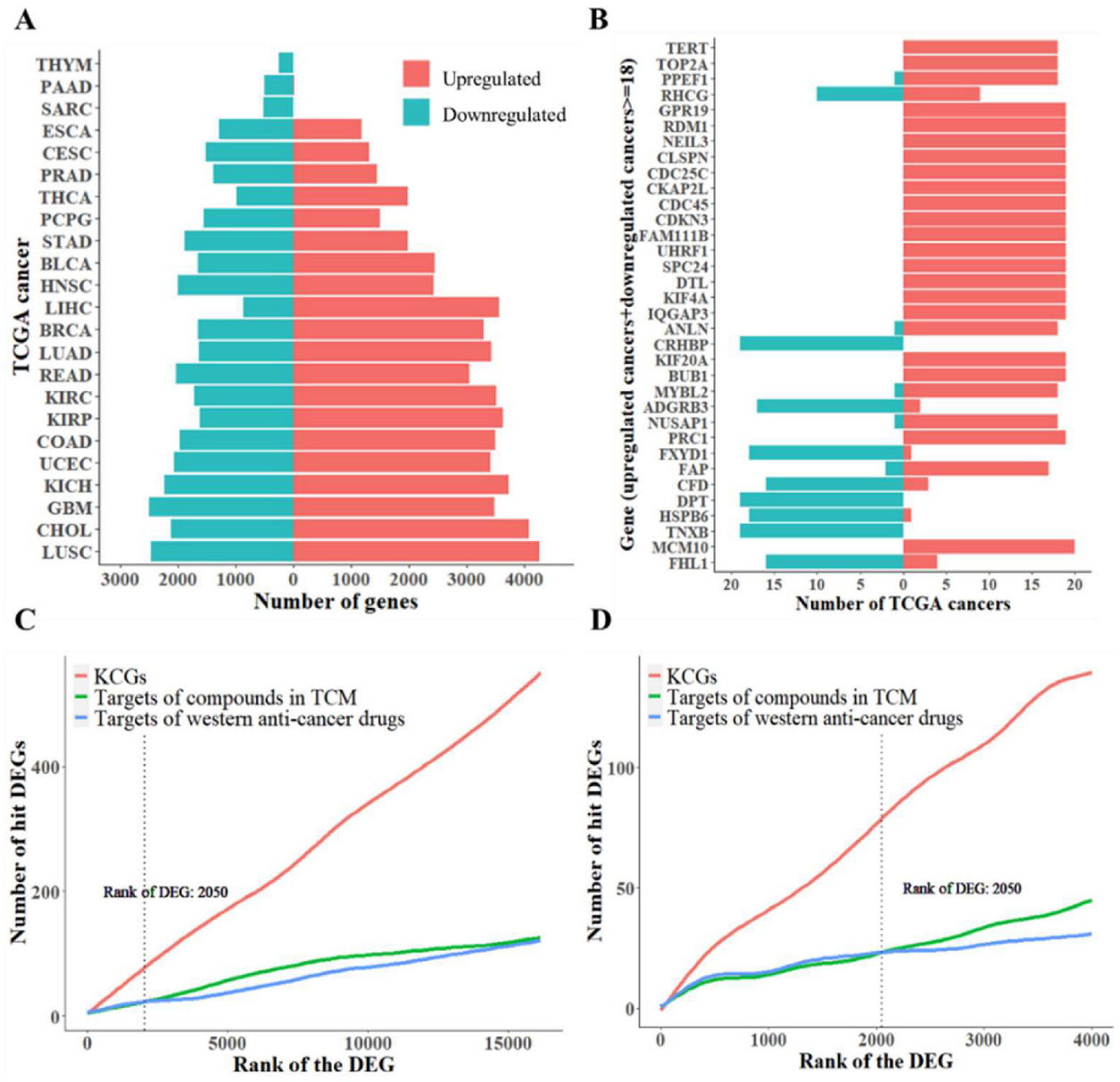

Figure 4. DEG analysis across all TCGA cancer and comparison of effect on the DEGs of compounds in TCM and western anti-cancer drugs. (A) The number of upregulated genes and downregulated genes observed in cancers; (B) The number of cancers observed in upregulated genes and downregulated genes; (C) Comparison of the coverage of DEGs; (D) Comparison of the coverage of DEGs.

KICH, THYM, BLCA, BRCA, COAD, CESC, STAD, SKCM and LUAD (Figure 5E). Some genes are mainly amplified in various cancers, such as CCND1 in BLCA, ESCA, BRCA and HNSC, MYC in BRCA, ESCA, PAAD and STAD, EGFR in GBM and TOP2A in STAD. While, some genes are mainly deleted in various cancers, such as TERT in KICH, GPR35 in BLCA, CESC and KICH, GPR55, HDAC4 and TUBA4A in CESC. When considered the targets of compounds in TCM, 9 cancers are mainly amplified, while the remaining 7 cancers are mainly deleted.

2.5. Gene mutation analysis reveals that targets of compounds in TCM have comparable coverage with targets of western anti-cancer drugs

The mutation frequency for each gene in each cancer type was defined as the times of mutation divided by the sample size of the annotation file. The genes were sorted by the mutation frequency in descending order. The top 5% genes for each cancer type were reserved as the genes with high mutation frequency and involved in the further analysis. For each gene, we counted the cancers that it appears to have high mutation frequency and reserved the top 30 genes that have the most cancers. Finally, the cancers UCEC, STAD, SKCM, SARC, LUSC, LIHC, HNSC, ESCA, CESC, BRCA and BLCA get the most frequently mutated genes (Figure 6A), and genes TTN, MUC16 and LRP1B get the most frequently mutated cancers (Figure 6B).

We sorted the genes according to the number of cancers that they were frequently mutated. The targets of compounds in TCM have a comparable coverage with the targets of western anti-cancer drugs. The coverage of targets of compounds in TCM and targets of western anti-cancer drugs are all lower than the that of KCGs (Figure 6C). Targets of compounds in TCM are most frequent significantly mutated in SKCM, BRCA, LUAD and HNSC (Figure 6D). These results show that there is no significant difference in the level of gene mutation between targets of compounds in TCM and targets of western anti-cancer drugs.

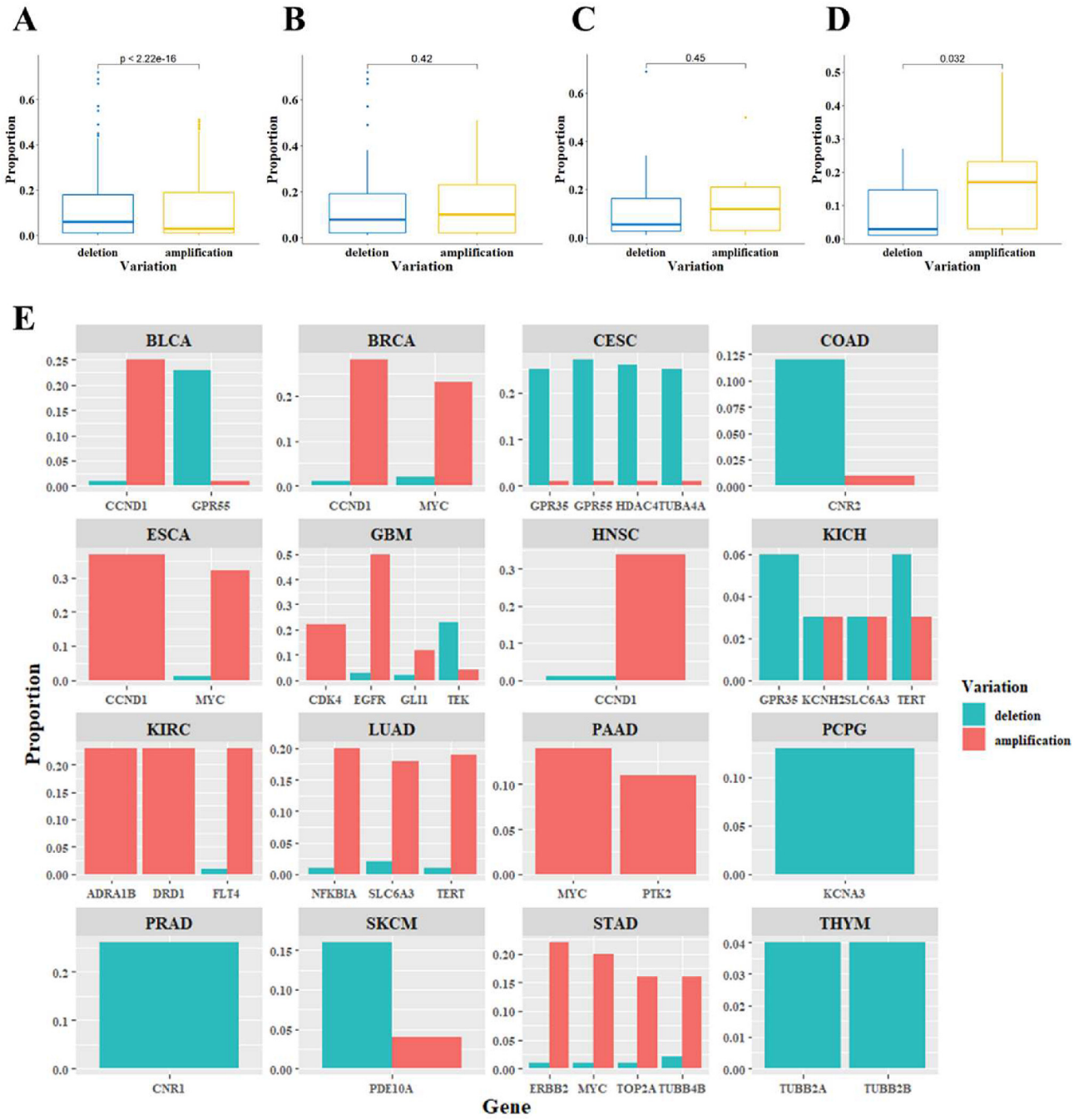

Figure 5. CNV analysis across all TCGA cancer. (A) Variation type statistics of the genes in all cancer; (B) Variation type statistics of the KCGs in all cancer; (C) Variation type statistics of the targets of western anti-cancer drug; (D) Variation type statistics of targets of compounds in TCM; (E) CNV analysis result for the targets of compounds in TCM in all related cancers.

2.6. Survival analysis to screen potential therapeutic biomarkers for compounds in TCM

Survival analysis was performed for all genes in all cancers, and a p-value $\leq$0.05 was used as the cut-off to select significant genes. ACC, KICH, MESO, LGG and KIRC have the most number of genes that are significant in survival (Figure 7A). While, TGCT, DLBC and UCS have the least genes that are significant in survival (Figure 7A). 49 genes are significant in survival in more than 13 TCGA cancers. HOXC5, VGF and ERCC6L are top 3 genes that have the most cancers in which they are significant in survival (Figure 7B).

We sorted the genes according to the number of cancers that they are significant in survival. The targets of compounds in TCM also have a comparable coverage with the targets of western anti-cancer drugs. The coverage of targets of compounds in TCM and targets of western anti-cancer drugs are all also lower than the that of KCGs (Figure 7C). 13 targets of compounds in TCM, namely ERBB2, MYC, FLT4, TEK, GLI1, TOP2A, PDE10A, SLC6A3, GPR55, TERT, EGFR, KCNA3 and HDAC4, differentially expressed, high frequently mutated, obtain high CNV rate and also significant in survival (Figure 7D). As shown in Figure 7E, EGFR is the common target of 15 compounds, GLI1 is the target of 3 compounds, HDAC4 is the target of 2 compounds, and the others are targets of only 1 compound. These targets are expected to be therapeutic biomarkers of the corresponding compounds.

We also investigated the role of several commonly genes in patient survival (Figure 8). In BLCA and HNSC, the overexpression of EGFR is

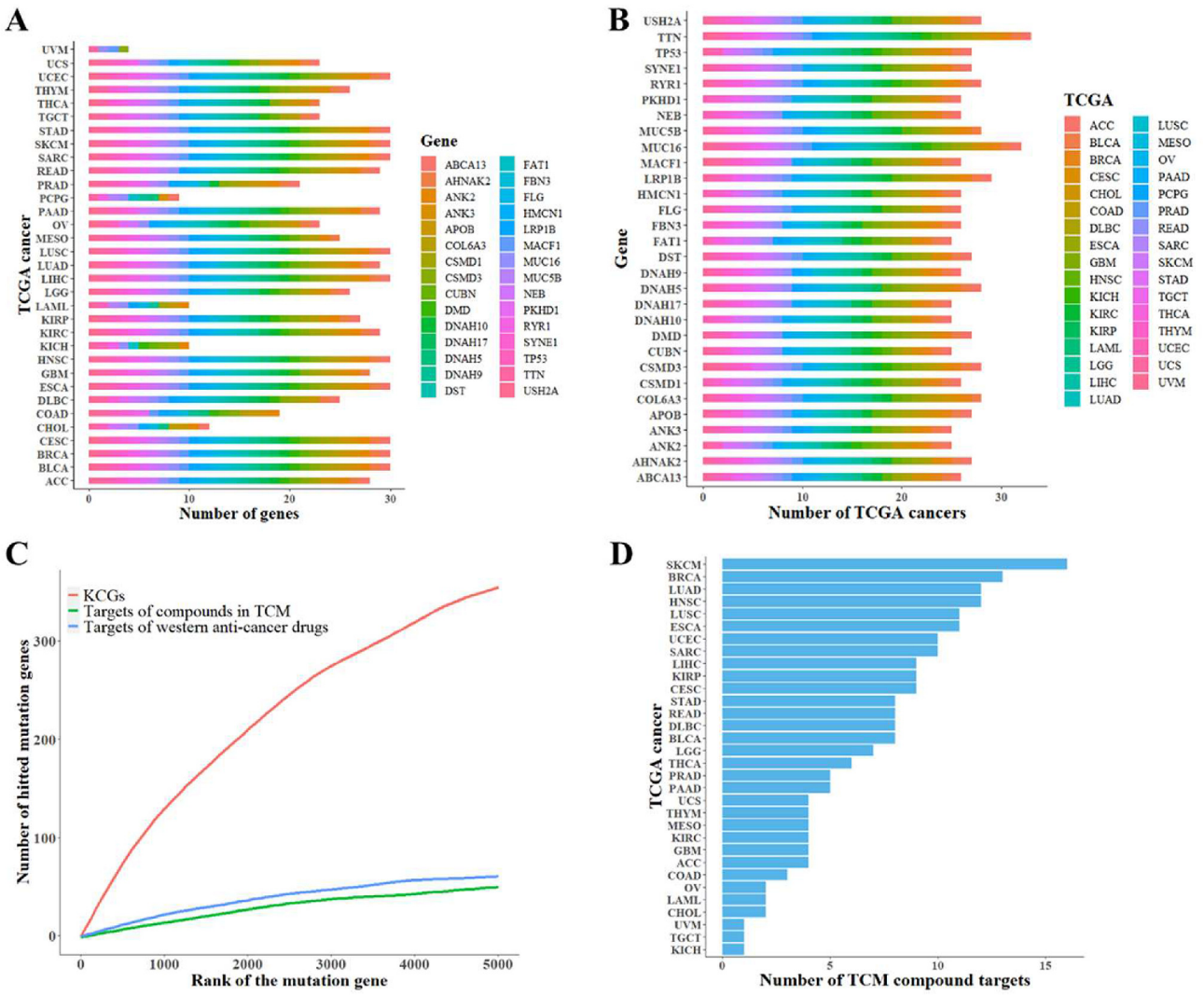

Figure 6. Gene mutation analysis across all TCGA cancer. (A) The number of significant mutated genes in all cancers; (B) The number of cancers corresponding to significant mutated genes; (C) Comparison of the coverage for the genes that are significant mutated for KCGs, targets of western anti-cancer drugs and targets of compounds in TCM; (D) The number of the targets of compounds in TCM that are significantly mutated for each cancer.

found to be significant in lower overall survival, suggesting its use as a therapeutic biomarker in BLCA and HNSC patients (Figure 8A and B). TERT was found to be upregulated in 18 TCGA cancers (Figure 4B). The overexpression of TERT in HNSC is also found to be significantly in higher overall survival (Figure 8C). While in MESO, the overexpression of TERT is found to be significant in lower overall survival, suggesting its use as a therapeutic biomarker in MESO patients (Figure 8D).

3. Discussion

TCM plays an important role in health maintenance for the people [35]. Scientific studies have showed that TCM could prevent and treat various diseases, such as cancer, cardiovascular diseases, infection and so on [36, 37, 38, 39, 40]. During the pandemic of COVID-19, TCM also plays an essential role in the prevention and treatment of pneumonia caused by SARS-CoV2 [41,42]. Even with a large number of microscopic experimental detection techniques [43, 44], the usage of TCM is still mainly limited to the macro level, and there is a huge demand to quantify the efficacy of TCM in the treatment of diseases, such as cancer. TCM is gradually being recognized as a complementary and alternative therapy all over the world in cancer treatment [5]. However, how to find therapeutic molecular biomarkers for anti-cancer compounds in TCM is always a huge challenge.

The discovery of a targeted therapeutic compound along with its companion predictive biomarker is a major goal of clinical development for a personalized anti-cancer therapy to date. As a widely used platform, GDSC provides a unique resource incorporating large drug sensitivity and genomic datasets to facilitate the discovery of new therapeutic biomarkers for cancer therapies [25]. GDSC currently contains drug sensitivity data for almost 75,000 experiments, describing response to 138 anti-cancer drugs across almost 700 cancer cell lines. As a case study, previous study has showed that dsRNA mediates its therapeutic effect through TLR3 expressed on tumor cells, and could therefore represent an effective targeted treatment in patients with TLR3-positive cancers [45]. Thus, TLR3 as a biomarker for the therapeutic efficacy of dsRNA in breast cancer. Analysis of the biological basis also promote the discovery of therapeutic biomarkers of TCM for cancer treatment [43], as well as potential therapeutic biomarkers of TCM for the treatment of COVID-19 related cytokine storm [46, 47]. Therefore, the discovery of therapeutic biomarkers of anti-cancer compounds in TCM has a well scientific basis and exactly has obtained the attraction of researchers around the world.

Here we presented the evidence of the predictive value of therapeutic biomarkers, e.g., EGFR for BLCA and HNSC, TERT for HNSC and MESO. Although few conformed in clinical trials, several EGFR-pathway inhibitor biomarkers still been researched for HNSC, and the predictive value is obvious to all [48]. We also note that EGFR can be used as a prognostic

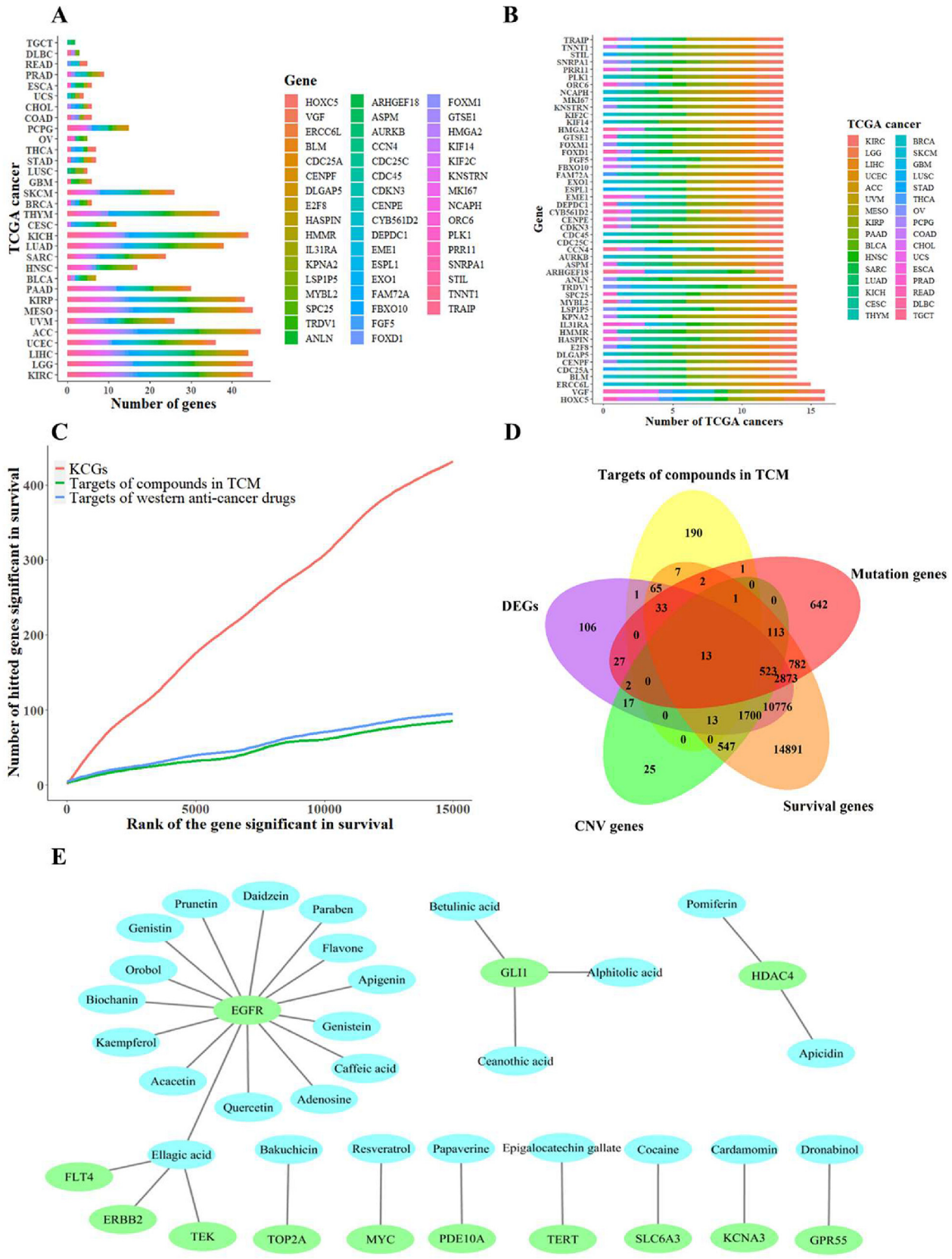

Figure 7. Survival analysis across all TCGA cancer. (A) The number of genes significant in survival in all cancers; (B) The number of cancers corresponding to genes that are significant in survival; (C) Comparison of the coverage of genes significant in survival for KCGs, targets of western anti-cancer drugs and targets of compounds in TCM; (D) Intersection of the targets of compounds in TCM, DEGs, CNV genes, mutation genes and survival genes; (E) Compound-target network for the 13 selected therapeutic biomarkers.

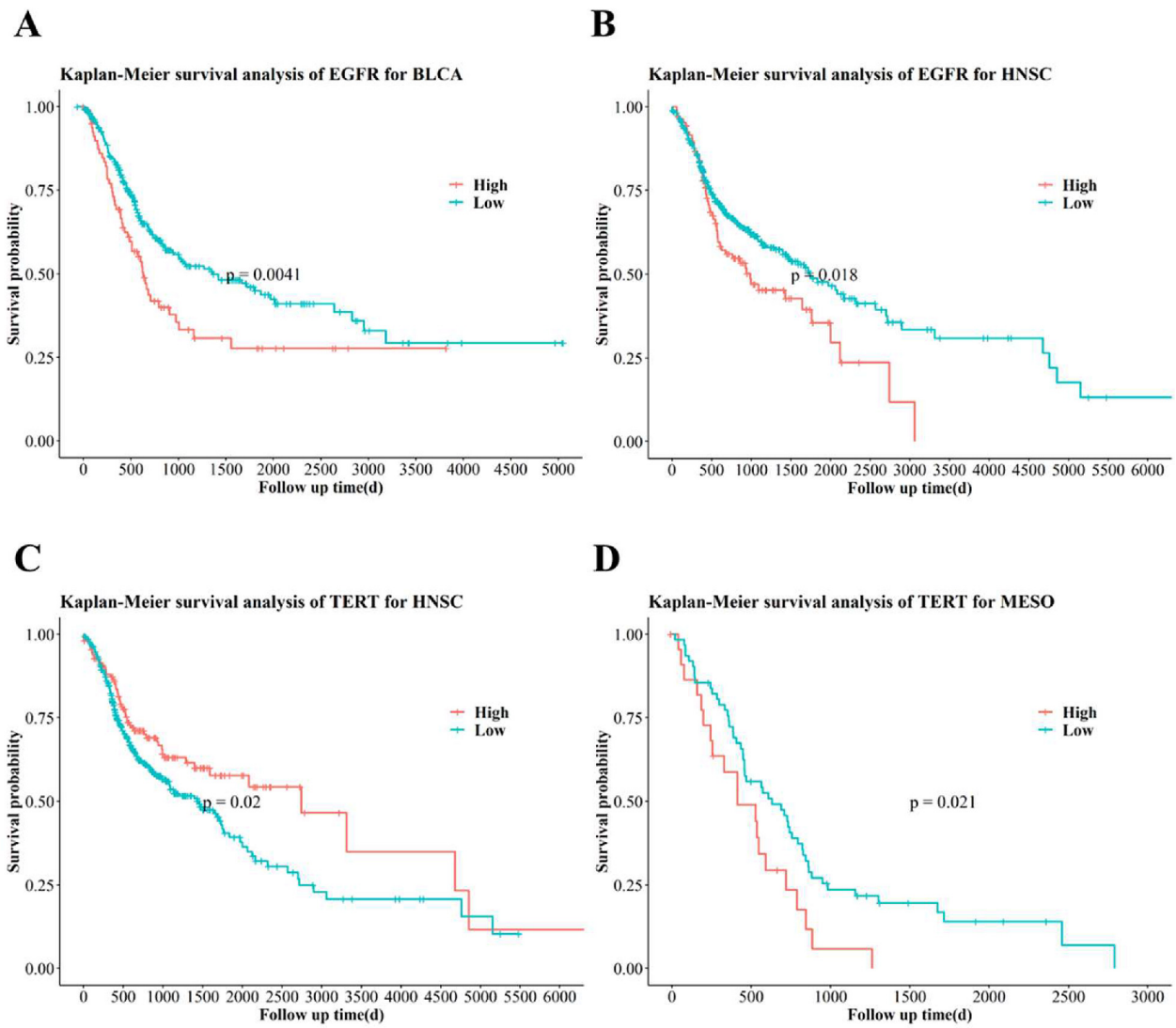

Figure 8. Role of EGFR and TERT in therapeutics. (A) Survival plot of EGFR high vs. low gene expression for BLCA; (B) Survival plot of EGFR high vs. low gene expression for HNSC; (C) Survival plot of TERT high vs. low gene expression for HNSC; (D) Survival plot of TERT high vs. low gene expression for MESO.

biomarker in BLCA due to the significant association of EGFR overexpression with tumor grade, muscularis propria invasion and recurrence [49]. TCM compound Quercetin treatment suppressed cell growth by inducing G2 arrest and apoptosis in EGFR-overexpressing HNSC cancer cells [50]. Meanwhile, we found that TERT may be a potential biomarker for HNSC by multi-omics analysis, case-control study also discovered that HNSC cases, especially oral cancer cases, had shorter telomere length than controls, and rs2736100 (TERT SNP) related to relative telomere length (RTL) in European was associated with both telomere length and HNSC risk in this southeast Chinese population [51].

In the analysis of multi-omics data, we sorted the genes in descending order according to the number of cancers. However, the genes with high rank in DEGs, genes differentially mutated, genes significant in survival and genes copy number variated get lower overlap. For example, only one gene ASPM appears in all the top 100 genes of DEGs, genes differentially mutated, genes significant in survival and genes copy number variated. This is contrary to our expectation that the genes screened out will rank higher in all multi-omics data. Therefore, it is not comprehensive to only consider the number of cancers in the process of biomarker screening. Although there is no well consistency in the rank of gene in multi-omics data, it does indicate that some outlier signals can be found by integrating different levels of omics data, which may help us to find new therapeutic biomarkers.

In the future, we will improve our study in several aspects. First, we will discuss the relationship between different levels of omics data, e.g., the correlation analysis of different levels of omics data. This may help us to filter out results with more stringent conditions. Second, the scope of external verification should be expanded, such as using gene expression data in GEO for external verification. Third, more analysis, e.g., single nucleotide polymorphism analysis and infiltrated immune cells estimate need to be conducted for more comprehensively analysis. Finally, experiments should be conducted to verify our findings.

In summary, we conducted multi-omics analysis including gene mutation analysis, differential expression gene analysis, copy number variation analysis and clinical survival analysis for pan-cancers to screen potential therapeutic biomarkers for compounds in TCM. Finally, 13 molecules of compounds in TCM are considered as the potential therapeutic biomarkers. This strategy may be a potential way to screen therapeutic biomarkers for anti-cancer therapy including complementary and alternative therapy.

4. Materials and methods

4.1. Collection of compounds in TCM

The compounds of TCM were obtained from public database TCMID [27], SymMap [28] and HIT [29]. The number of compounds in those databases were 11525, 4103 and 489 separately. After removing duplicated ones, 14522 compounds were finally reserved.

4.2. Molecules related to the compounds in TCM

We collected the molecules related to the compounds in TCM from databases ChEMBL [12] and GDSC [25]. For the 14522 compounds, 103 have activity data in ChEMBL database, filtering by the following rules: (1) assay conducted in Homo sapiens cell line; (2) assay with IC_{50}; (3) the investigational agent has CID and ChEMBL ID; (4) the investigational gene has gene symbol. 322 molecules related to the 103 compounds were reserved. In addition, TOP1, XIAP, AKT1, AKT2, AKT3 and AMPK are documented as the target of some compounds in GDSC database. Finally, 326 molecules were reserved as the (potential) targets of compounds in TCM.

4.3. KCGs and targets of western anti-cancer drugs

In this study, we compared the KCGs with targets of compounds in TCM and targets of western anti-cancer drugs from the perspective of multi-omics analysis. Targets already known to be related to certain cancer are referred as golden standard. 711 expert-curated KCGs are obtained from NCG [52]. The western anti-cancer drugs and the 314 corresponding targets are obtained from GDSC [25].

4.4. DEG analysis

DEG analysis was performed using the Bioconductor tool TCGAbiolinks [30, 31, 32]. We firstly downloaded gene expression quantification data for 33 cancers from TCGA. Then, we filtered the samples with low correlation according to the spearman correlation coefficient with cut-off 0.6. The mRNA transcripts were normalized using EDASeq package. We removed the genes with low counts according the quantile cut-off 0.25. Finally, FDR cut-off of 0.01, and an absolute $\log_2$ fold change cut off 1 were used to obtain the list of DEGs.

4.5. Survival analysis

Survival analysis was carried out using R tools, survival and survminer in the background, for the gene expression data. Patients were segregated into high and low expression groups based on the expression mean value for each cancer. Kaplan-Meier (KM) analysis was performed, and a p-value ≤ 0.05 was used as the cut-off to select significant genes.

4.6. CNV analysis

For analyzing CNV, gene level copy number scores data for 24 types of cancers were downloaded from TCGA using TCGAbiolinks. Both amplified and deleted genes were collected, and we defined the CNV rate of a certain gene as the ratio of sample size the gene was amplified or deleted to the total sample size in a specific cancer. Only the tumor samples were analyzed. The genes were sorted by the CNV rate in descending order. The top 1% genes for each cancer type were reserved as the genes with high CNV rate and involved in the further analysis.

4.7. Gene mutation analysis

We processed mutation annotation files downloaded from TCGA using TCGAbiolinks. The mutation frequency for each gene in each cancer type was defined as the times of mutation divided by the sample size of the annotation file. The genes were sorted by the mutation frequency. The top 5% genes for each cancer type were reserved as the genes with high mutation frequency and involved in the further analysis.

4.8. Enrichment analysis

Enrichment analysis was conducted using R tool clusterProfiler [53]. Significant enriched terms of KEGG, biological processes (BP), molecular functions (MF) and cellular components (CC) related to given genes were calculated in the background of org.Hs.eg.db. The p-value cut-off was set as 0.01.

Declarations

Author contribution statement

Wuai Zhou: Conceived and designed the experiments; Wrote the paper.

Ruxue Li: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data included in article/supp. material/referenced in article.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2022.e09616.

Acknowledgements

Not applicable.

References

[1] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, CA A Cancer J. Clin. 72 (2022) 7–33.
[2] West, H. Complementary and alternative medicine in cancer care. JAMA Oncol. 4, 139.
[3] G. Deng, B. Cassileth, Complementary or alternative medicine in cancer care-myths and realities, Nat. Rev. Clin. Oncol. 10 (2013) 656–664.
[4] Wu, M., Lu, P., Shi, L. & Shao, L. Traditional Chinese patent medicines for cancer treatment in China: a nationwide medical insurance data analysis. Oncotarget 6, 38283–38295.
[5] Wang, C. Y., Bai, X. Y. & Wang, C. H. Traditional Chinese medicine: a treasured natural resource of anticancer drug research and development. Am. J. Chin. Med. 42, 543–559.
[6] X, W. X D. Wang, Y. Gao, J. Zhang, X.Y. Zhu, Research progress of Chinese herbal medicine for softening hardness to dissipate stagnation in treatment of malignant tumors, Chinese Journal of Experimental Traditional Medical Formulae 26 (2020) 219–225.
[7] Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. 19, A68–A77.
[8] Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210.
[9] Barrett, T. *et al.* NCBI GEO: archive for functional genomics datasets-10 years on. Nucleic Acids Res. 39, D1005–D1010.
[10] Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607.
[11] Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–575.

[12] Bento, A. P. c., Gaulton, A., Hersey, A., Bellis, L. J. & Chambers, J. The ChEMBL bioactivity database: an update. Nucleic Acids Res. 42, D1083–D1090.
[13] Cancer Genome Atlas Research, N. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 169, 1327–1341.
[14] Cancer Genome Atlas Research, N. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513, 202–209.
[15] Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550.
[16] Cancer Genome Atlas Research, N. et al. Integrated genomic and molecular characterization of cervical cancer. Nature 543, 378–384.
[17] Cancer Genome Atlas Research, N. et al. Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73.
[18] Cancer Genome Atlas Research, N. et al. Integrated genomic characterization of oesophageal carcinoma. Nature 541, 169–175.
[19] Li, Y. *et al.* Multi-platform omics analysis reveals molecular signature for COVID-19 pathogenesis, prognosis and drug target discovery. Signal Transduct. Targeted Ther. 6, 155.
[20] Montaner, J. *et al.* Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. Nat. Rev. Neurol. 16, 247–264.
[21] Wang, T. X. *et al.* MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat. Commun. 12.
[22] Yuan, Y. *et al.* Multi-omics analysis reveals novel subtypes and driver genes in glioblastoma. Front. Genet. 11, 565341.
[23] Vasaikar, S. V., Straub, P., Wang, J. & Zhang, B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 46, D956–D963.
[24] Amjad, E., Asnaashari, S., Sokouti, B. & Dastmalchi, S. Impact of gene biomarker discovery tools based on protein-protein interaction and machine learning on performance of artificial intelligence models in predicting clinical stages of breast cancer. Interdiscipl. Sci. Comput. Life Sci. 12, 476–486.
[25] Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 41, D955–D961.
[26] S, W. D. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–D1082.
[27] Lin, H. *et al.* TCMID 2.0: a comprehensive resource for TCM. Nucleic Acids Res. 46, D1117–D1120.
[28] Wu, Y. *et al.* SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Res. 47, D1110–D1117.
[29] Hao, Y. *et al.* HIT: linking herbal active ingredients to targets. Nucleic Acids Res. 39, D1055–D1059.
[30] Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C. & Papaleo, E. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS Comput. Biol. 15, e1006701.
[31] Silva, T. C., Colaprico, A., Olsen, C., Angelo, F. D. & Noushmehr, H. TCGA Workflow: analyze cancer genomics and epigenomics data using Bioconductor packages. F1000research 5, 1542.
[32] Colaprico, A., Silva, T. C., Olsen, C., Garofano, L. & Noushmehr, H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 44, e71.
[33] S. Li, Y.Y. Wang, J. Liang, L.Y. D, A discussion and case study of complexities in traditional Chinese medicine, J. Syst. Simul. 14 (2002) 1429–1432.
[34] Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. Nat. Rev. Genet. 16, 172–183.
[35] Cheung, F. TCM made in China. Nature 480.
[36] Wang, C. *et al.* Oseltamivir compared with the Chinese traditional therapy maxingshigan-yinqiaosan in the treatment of H1N1 influenza: a randomized trial. Ann. Intern. Med. 155, 217–225.
[37] Li, X. *et al.* A multicenter, randomized, double-blind, parallel-group, placebo-controlled study of the effects of qili qiangxin capsules in patients with chronic heart failure. J. Am. Coll. Cardiol. 62, 1065–1072.
[38] Chen, Q. *et al.* Effect of Huaier granule on recurrence after curative resection of HCC: a multicentre, randomised clinical trial. Gut 67, 2006–2016.
[39] Zhong, L. L. D. *et al.* Efficacy of MaZiRenWan, a Chinese herbal medicine, in patients with functional constipation in a randomized controlled trial. Clin. Gastroenterol. Hepatol. 17, 1303–1310.
[40] Zhang, D. Y. *et al.* Treatment of masked hypertension with a Chinese herbal formula: a randomized, placebo-controlled trial. Circulation 142, 1821–1830.
[41] Z. Zhao, et al., Prevention and Treatment of COVID-19 Using Traditional Chinese Medicine: A Review, Phytomedicine, 2021, p. 153308.
[42] Runfeng, L. *et al.* Lianhuaqingwen exerts anti-viral and anti-inflammatory activity against novel coronavirus (SARS-CoV-2). Pharmacol. Res. 156, 104761.
[43] Zheng, J. *et al.* Network pharmacology to unveil the biological basis of health-strengthening herbal medicine in cancer treatment. Cancers 10, 461.
[44] Guo, Y. *et al.* Network-based combinatorial CRISPR-Cas9 screens identify synergistic modules in human cells. ACS Synth. Biol. 8, 482–490.
[45] Salaun, B. *et al.* TLR3 as a biomarker for the therapeutic efficacy of double-stranded RNA in breast cancer. Cancer Res. 71, 1607–1614.
[46] Dai, Y. *et al.* A large-scale transcriptional study reveals inhibition of COVID-19 related cytokine storm by traditional Chinese medicines. Sci. Bull. 66, 884–888.
[47] Chen, R. *et al.* HMGB1 as a potential biomarker and therapeutic target for severe COVID-19. Heliyon 6, e05672.
[48] de Kort, W. W. B., Spelier, S., Devriese, L. A., van Es, R. J. J. & Willems, S. M. Predictive value of EGFR-PI3K-AKT-mTOR-Pathway inhibitor biomarkers for head and neck squamous cell carcinoma: a systematic review. Mol. Diagn. Ther. 25, 123–136.
[49] Hashmi, A. A. *et al.* Prognostic significance of epidermal growth factor receptor (EGFR) over expression in urothelial carcinoma of urinary bladder. BMC Urol. 18, 59.
[50] Huang, C. Y. *et al.* Quercetin induces growth arrest through activation of FOXO1 transcription factor in EGFR-overexpressing oral cancer cells. JNB (J. Nutr. Biochem.) 24, 1596–1603.
[51] Gu, Y. *et al.* Telomere length, genetic variants and risk of squamous cell carcinoma of the head and neck in Southeast Chinese. Sci Rep-Uk 6, 20675.
[52] Repana, D. *et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 20, 1–12.
[53] Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics-a Journal of Integrative Biology 16, 284–287.