Classification of clear cell renal cell carcinoma based on PKM alternative splicing

Downloaded from: https://research.chalmers.se, 2020-04-25 07:33 UTC

Citation for the original published paper (version of record):
Classification of clear cell renal cell carcinoma based on PKM alternative splicing
Heliyon, 6(2)
http://dx.doi.org/10.1016/j.heliyon.2020.e03440

N.B. When citing this work, cite the original published paper.
Classification of clear cell renal cell carcinoma based on PKM alternative splicing

Xiangyu Li, Beste Turanli, Kajetan Juszczak, Woonghee Kim, Muhammad Arif, Yusuke Sato, Seishi Ogawa, Hasan Turkez, Jens Nielsen, Jan Boren, Mathias Uhlen, Cheng Zhang, Adil Mardinoglu

Keywords: Bioinformatics, Cancer research, Systems biology, PKM, Alternative splicing, Transcriptomics, Biomarker, Drug repositioning

ABSTRACT

Clear cell renal cell carcinoma (ccRCC) accounts for 70–80% of kidney cancer diagnoses and displays high molecular and histologic heterogeneity. Hence, it is necessary to reveal the underlying molecular mechanisms involved in progression of ccRCC to better stratify the patients and design effective treatment strategies. Here, we analyzed the survival outcome of ccRCC patients as a consequence of the differential expression of four transcript isoforms of the pyruvate kinase muscle type (PKM). We first extracted a classification biomarker consisting of eight gene pairs whose within-sample relative expression orderings (REOs) could be used to robustly classify the patients into two groups with distinct molecular characteristics and survival outcomes. Next, we validated our findings in a validation cohort and an independent Japanese ccRCC cohort. We finally performed drug repositioning analysis based on transcriptomic expression profiles of drug-perturbed cancer cell lines and proposed that paracetamol, nizatidine, dimethadione and conessine can be repurposed to treat the patients in one of the subtype of ccRCC whereas chenodeoxycholic acid, fenoterol and hexylcaine can be repurposed to treat the patients in the other subtype.

1. Introduction

Clear cell renal cell carcinoma (ccRCC) is the most common subtype among renal cancer (Motzer et al., 2017) and ccRCC shows high inter individual heterogeneity (Ricketts et al., 2018). Thus, it is difficult to predict survival outcomes of patients in clinical practice and design effective therapeutic strategies. Previous studies have already proposed strategies for stratification of ccRCC patients into subgroups based on different genetic and/or transcriptomic characteristics and prognoses of the patients (Brannon et al., 2010; Cancer Genome Atlas Research, 2013; Kosari et al., 2005; Takahashi et al., 2001). However, these studies failed to identify a clinically practical biomarker for classification of the patients at the personalized level or recommend targeted chemotherapy regimens for these patients.

Pyruvate kinase muscle type (PKM) is an enzyme that is involved in the final step of glycolysis and catalyzes the formation of ATP from ADP as phosphoenolpyruvate undergoes dephosphorylation to pyruvate. It is one of the key regulators of Warburg effect in tumors. Many studies have reported that different alternative splicing isoforms of PKM promote tumorigenesis (Christofk et al., 2008a; Clower et al., 2010; Yang et al., 2012), cell proliferation (Christofk et al., 2008b; Huang et al., 2017; Li et al., 2014; Lunt et al., 2015; Morita et al., 2018; Yang et al., 2011),
metastasis (Xie et al., 2019; Yang et al., 2014) and poor prognoses of patients (Ahmed et al., 2007; Uhlen et al., 2017) in different cancers. In a previous study, PKM isoform switch mediated by PTBP1 was found in ccRCC (Jiang et al., 2017). Interestingly, we also found four different isoforms of PKM play a very important role in controlling metabolism of ccRCC (Li et al., 2019b). We have observed that the expression level of four protein-coding transcripts of PKM, including ENST00000335181, which is encoding PKM2 and the most studied isoform of PKM, as well as ENST00000561609, ENST00000389093 and ENST00000568883 are highly associated with patients’ prognoses (Li et al., 2019b). Among them, high expression of ENST00000335181 and ENST00000561609 indicated a favorable survival while high expression of ENST00000389093 and ENST00000568883 indicated an unfavorable survival. Moreover, a number of conserved biological functions associated with the progression of ccRCC were oppositely dysregulated by these transcripts. Here, we hypothesized that different molecular subtypes among ccRCC patients can be characterized by the different expression patterns induced by these four prognostic transcripts and biomarkers that may be used in clinical practice can be identified.

Systems biology based approaches have been used in stratification of patients with different metabolic diseases and certain type of cancers for characterization of the patient subtypes (Altay et al., 2019; Benfeitas et al., 2019; Bidkhori et al., 2018; Bjornson et al., 2015; Mardinoglu et al., 2018). Previous studies have also proposed transcriptomics-based biomarkers for classification of tumors based on the quantitative measurement of one or multiple signature genes (Fujita et al., 2012; Jones et al., 2005; Klatte et al., 2009; Kosari et al., 2005; Zhao et al., 2006). However, this kind of transcriptional signatures are rarely used in clinical practice due to technological and translational barriers (Winslow et al., 2012). Besides problems in tissue sampling and quality control, an important factor is experimental batch effect which brings high variation of gene expression induced by the different laboratory conditions and personnel (Guan et al., 2018). To solve these problems, the use of biomarkers based on the within-sample relative expression orderings (REOs) of gene pairs has been proposed (Guo et al., 2018; Qi et al., 2016a, 2016b). REOs-based biomarkers are robust against batch effects, invariant to monotone data normalization (Iddry et al., 2010; Wang et al., 2013) and poor sample preparation (Chen et al., 2017b; Cheng et al., 2017; Liu et al., 2017).

In this study, we used the genes dysregulated by the prognostic transcripts of PKM to extract classification biomarker instead of using themselves. Since different transcripts of PKM share similar sequence, it may be difficult to design distinct primers to detect their relative abundance when using real-time PCR. Thus, gene pair biomarker is more feasible and practical in clinical diagnosis. We applied REOs-based method to identify biomarkers for classification of ccRCC by extracting the expression profiles of genes which were consistently negatively dysregulated by the four favorable and unfavorable prognostic transcripts of PKM. We developed a REOs-based biomarker using the global gene expression profiling of ccRCC in The Cancer Genome Atlas (TCGA) database and stratified the patients into two subtypes exhibiting different transcriptomic expression patterns and different patient prognosis. We also validated our findings in TCGA database as well as in another independent Japanese cohort. We finally proposed several candidate drugs that can be used in treatment of each subtype based on transcriptomic expression profiles of drug-perturbed cancer cell lines from Connectivity Map 2.0 (CMap2).

2. Result

2.1. Identification of signature gene set associated with four prognostic transcripts of PKM

In a recent study, we have found that there are molecular subtypes that could be characterized by the expression of the four different prognostic PKM transcripts (Li et al., 2019b). In order to develop a REOs-based biomarker, we identified a signature gene set associated with these four transcripts based on the gene expression profiles of TCGA ccRCC samples. We performed differential expression analysis between the tumor samples from patients with high (top 25%) and low (bottom 25%) expression of each favorable transcript, and identified 2,010 consistently significantly (FDR <1.0e-5) differentially expressed genes (DEGs) for the two favorable transcripts (Figure 1). Similarly, we identified 5,469 DEGs consistently significantly (FDR <1.0e-5) DEGs for the two unfavorable transcripts. We found that the two sets of DEGs has a significant overlap (n = 1,135; hypergeometric distribution test, p < 1.1e-16). We also observed that the concordance score of these overlapped genes is 100%, which means the up-regulated genes associated with high expression of favorable transcripts within these 1,135 genes are all down-regulated when the unfavorable transcripts exhibit high expression; and vice versa.

We followed-up survival information from the corresponding patients and found that 539 of the 1,135 genes (of which 205 and 234 are favorable and unfavorable, respectively) are significantly (univariate Cox model, FDR <0.01) associated with patients’ overall survival (OS). To identify the associated biological functions with these 539 genes, we performed GO term enrichment analysis and observed that these genes are significantly enriched in RNA splicing, RNA catalytic process and nuclear transport pathways (FDR<0.05; Table S1). Therefore, we concluded that these 539 genes may be used as the core signature genes that are associated with the differential alternative splicing of PKM among ccRCC patients and may be used for classification of tumor samples.

We calculated the co-expression coefficients between the expression of the 539 signature genes and found two major clusters in which all favorable genes are positively co-expressed while all unfavorable genes are negatively co-expressed in the opposite cluster using the hierarchical clustering (Figure 2A). Based on the expression profiles of these 539 signature genes, we employed consensus clustering to classify TCGA ccRCC samples into distinct stable sub-groups through repeated subsampling and clustering (Wilkerson and Hayes, 2010). As shown in Figure 2B, we determined an optimum number of two clusters, cluster 1 and 2, based on the lowest proportion of ambiguous clustering (Senba-baoglu et al., 2014). Using survival analysis, we observed the patients whose tumor samples classified in cluster 1 (N = 231) had significantly shorter OS than those classified in cluster 2 (N = 297) with statistical significance (log-rank test, P = 6.73e-07; Figure 2C). The results demonstrated that there are two different molecular subtypes in ccRCC with significantly different survival outcomes which are strongly associated with the function of the two favorable and two unfavorable transcripts.

2.2. Development of the REOs-based classification biomarker

To identify a biomarker that can be used in the clinical practice, we next focused on development of a REOs-based classification biomarker based on the gene expression profiles of the 539 signature genes. In brief, REOs-based biomarkers employs gene pairs with consistently reversed expression orders between the two molecular groups as indicators, and screens for a minimum combination of these gene pairs that serves as risk indicators for classification. In order to obtain a robust biomarker, we generated 100 training and 100 validation datasets by randomly selecting from TCGA ccRCC cohort and randomly separated the samples into two respective groups with 70% and 30% samples. We identified 171 gene pairs that exhibited consistent reverse gene pairs in all training datasets. We next generated 17,100 reverse gene pair combinations with a forward selection procedure and selected a final REOs-based biomarker consisted of eight reverse gene pairs with an optimal mean F-score of 0.9725 in all training datasets (Figure 2D). The full screening process are shown in Figure 1 and detailed in Method section.

Within this eight gene pairs, if more than four gene pairs exhibited reversal REOs in a sample, this sample would be classified as having a higher expression pattern of the four favorable transcripts. This kind of discrimination may be sufficient to be used as a classification biomarker instead of using the entire gene expression profile. We followed-up survival information from the corresponding patients and found that 539 of the 1,135 genes (of which 205 and 234 are favorable and unfavorable, respectively) are significantly (univariate Cox model, FDR <0.01) associated with patients’ overall survival (OS). To identify the associated biological functions with these 539 genes, we performed GO term enrichment analysis and observed that these genes are significantly enriched in RNA splicing, RNA catalytic process and nuclear transport pathways (FDR<0.05; Table S1). Therefore, we concluded that these 539 genes may be used as the core signature genes that are associated with the differential alternative splicing of PKM among ccRCC patients and may be used for classification of tumor samples.

We calculated the co-expression coefficients between the expression of the 539 signature genes and found two major clusters in which all favorable genes are positively co-expressed while all unfavorable genes are negatively co-expressed in the opposite cluster using the hierarchical clustering (Figure 2A). Based on the expression profiles of these 539 signature genes, we employed consensus clustering to classify TCGA ccRCC samples into distinct stable sub-groups through repeated subsampling and clustering (Wilkerson and Hayes, 2010). As shown in Figure 2B, we determined an optimum number of two clusters, cluster 1 and 2, based on the lowest proportion of ambiguous clustering (Senba-baoglu et al., 2014). Using survival analysis, we observed the patients whose tumor samples classified in cluster 1 (N = 231) had significantly shorter OS than those classified in cluster 2 (N = 297) with statistical significance (log-rank test, P = 6.73e-07; Figure 2C). The results demonstrated that there are two different molecular subtypes in ccRCC with significantly different survival outcomes which are strongly associated with the function of the two favorable and two unfavorable transcripts.
risk group; otherwise, this sample would be classified into the low-risk group (Figure 2D). We tested these gene pairs in the 100 validation datasets, and found that these gene pairs also showed a good classification accuracy with a mean F-score 0.9742. We also tested these gene pairs using the complete TCGA cohort, and this biomarker classified 231 samples into high-risk group and 297 samples into low-risk group. Notably, these two groups showed significantly different OS (Figure 2E; log rank test, \( P = 1.69 \times 10^{-7} \)).

2.3. Validation of the REOs-based classification biomarker

To validate if these gene pairs can be used as a biomarker for classification of ccRCC samples, we tested these gene pairs in 100 ccRCC samples obtained from an independent Japanese cohort. The biomarker classified 35 samples as high-risk group and 65 samples as low-risk group, and these two groups showed significantly different OS (Figure 2F, \( P = 7.46 \times 10^{-5} \)). We next investigated whether the high and low-risk groups identified in both TCGA and Japanese cohorts exhibited similar biological differences. We extracted the top 20% most significant DEGs (\( n = 2,694 \)) between high and low-risk groups in both the TCGA and Japanese cohorts, and observed a significant overlap between them (\( n = 1,463 \); hypergeometric distribution test, \( p < 1.11 \times 10^{-16} \)) with a concordance score 100%. In addition, we identified 66 and 80 GO terms that are significantly enriched with upregulated genes (FDR < 1.0e-05) in the high-risk group of the TCGA and Japanese cohorts, respectively, and found that 55 of them are common in both cohorts (Figure 3).
Specifically, the high-risk group was characterized by upregulated genes involved in ATP synthesis, mitochondrial respiratory process, oxidative phosphorylation, ribonucleotide and purine nucleotide metabolic process, RNA catabolic process, protein targeting to ER and membrane pathways. And the low-risk group was characterized by upregulated genes involved in histone modification and covalent chromatin modification pathways. The results suggested the molecular subtypes identified by our analysis also have consistent biological differences. Moreover, these 55 GO terms included all 27 GO terms that we recently reported to be associated with the four different prognostic transcripts in pan cancer analysis (Li et al., 2019b). We also identified three GO terms that are significantly enriched with downregulated genes in the high-risk group for both cohorts, and two of them, which are the histone modification and covalent chromatin modification pathways, are common in both groups. These results further indicated that the molecular subtypes stratified by the gene pairs are functionally related to the four prognostic transcripts of PKM.

Further, we compared our REOs-based classification with previously reported TCGA (m1 to m4) and ccA/ccB classification schemes (Brannon et al., 2010; Cancer Genome Atlas Research, 2013) (Figure 4). In TCGA cohort, approximately 96% of TCGA m1 tumors were involved in our low-risk group, and m1 group was also reported with the best prognoses in TCGA classification scheme. In addition, 73% of TCGA tumors in both m2 and m4 subtypes were involved in our high-risk groups, and they
were also shown to be with the poorest prognoses in TCGA classification. These results demonstrated that the high and low-risk groups classified by our biomarker are reinforced by the previously observed survival outcomes m1, m2 and m4. Notably, the high and low-risk groups respectively accounted for 42% and 58% of tumors previously reported as unclassified (m3) in the TCGA classification scheme. In Japanese cohort, 71% of ccA and 80% of ccB were observed in the low and high-risk groups, respectively. We found that the favorable survival for
ccA cases again reinforced the low-risk group classification based on our gene pairs used as a biomarker (Figure 4).

2.4. Drug repositioning for treatment of ccRCC subtypes

In addition to classification of the tumors, we also performed drug repositioning analysis to identify drug candidates that can be used in treatment of each subtype (Figure 5). We assumed that if a drug could reverse the dysregulated gene expression pattern from a tumor subtype to normal pattern, it could be potentially useful for treating the specific tumor subtype. We used a method developed in our previous study for drug repurposing (Turanli et al., 2018, 2019a, 2019b) and found several drugs that could be used for treatment of the high and low-risk groups (see Method). We found that four different drugs including paracetamol, nizatidine, dimethadione and conessine may be used to reverse the gene expression in samples from high-risk group, since over 80% drug-perturbed genes were mapped to the DEGs between these samples and normal samples. Paracetamol is an analgesic and antipyretic drug with the effect of anti-inflammatory and inhibits the production of multiple oxidants (Graham et al., 2013). Interestingly, it has been reported that paracetamol inhibits the cell proliferation and induces cell apoptosis in pancreatic cancer (Maisy et al., 2017), ovarian cancer and lung cancer cells (Lian et al., 2018). Nizatidine is an antagonist of histamine H2 receptors and these receptors are associated with cell proliferation, embryonic development and tumor growth (Fernandez-Nogueira et al., 2018). It has been shown that nizatidine was recommended to be added into the combination therapy for cancer treatment (Barton-Burke, 1996; Ben-Sasson, 2007; Feitelberg et al.,

![Diagram of drug repositioning](image)

Figure 5. The flowchart for drug repositioning.
Conessine, a steroid alkaloid, decreases basal NF-xB activity, regulates autophagy, has anti-inflammatory effect (Kim et al., 2016, 2018) and is proposed as a potential anti-cancer drug. Therefore, the anti-cancer effects of three of the proposed drugs have been reported in previous studies.

Similarly, we found that three different drugs including chenodexoxycholic acid, fenoterol and hexylecaine, may be used to reverse the gene expression in samples of low-risk group towards normal samples. It has been reported that natural as well as synthetic bile acids can modulate cell cycle and induce apoptosis in cancer cells (Horowitz et al., 2007; Kim et al., 2006). Chenodexoxycholic acid, a bile acid, shows anti-proliferative activity in human cancer cells (Faustino et al., 2016). Beta-adrenergic signaling pathway has been found to promote the initiation and progression of cancer including inflammation, angiogenesis, apoptosis, cell motility, DNA damage repair and epithelial-mesenchymal transition (Coe and Sood, 2012). Fenoterol, a beta-adrenoceptor agonist, has been shown to inhibit the proliferation of glioblastomas and astrocytomass cells (Bernier et al., 2013; Toll et al., 2011). Hexylecaine, a short-acting local anesthetic, has also been used to treat cancer (Gleich, 2000) even its anti-cancer activity is still unclear. In this context, these three drugs may also be potentially used for treatment of the subtype of ccRCC patients.

Moreover, we identified the gene targets for each of the drugs using DrugBank database (Wisheart et al., 2006). HRH2, encoding the histamine H2 receptor, has been reported as the gene target of nizatidine (Meredith et al., 1985). We observed that it is significantly up-regulated in the samples of high-risk group compared to normal samples. It has been demonstrated that in vitro and in vivo histamine-induced tumor cell proliferation can be blocked by H2 antagonists (Deva and Jameson, 2012; Natori et al., 2005; Tomita et al., 2003). Thus, nizatidine may be used as a promising drug for the patients classified in the high-risk group. On the other hand, GPBAR1, encoding an enzyme of the G protein-coupled receptor superfamily, has been reported as a target of chenodexoxycholic acid. It has been shown that GPBAR1 antagonizes kidney cancer cell proliferation and migration (Su et al., 2017). Based on our analysis, we have observed that GPBAR1 is significantly down-regulated in the low-risk group compared to normal samples. Thus, chenodexoxycholic acid, as an activator of GPBAR1, may be used as a promising drug for treating the patients classified in the low-risk group.

3. Discussion

PKM is one of the important regulators of Warburg effect in different human cancers (Dayton et al., 2016). Our recent study showed that four different transcripts of PKM mediate opposite survival outcomes for ccRCC patients. In this study, we identified the core signature genes which were consistently dysregulated by these four prognostic transcripts of PKM. Using these signature genes, we identified eight gene pairs whose within-samples REOs could be used to classify patients into two groups with significantly different OS. Although RNA sequencing data was used for the biomarker classification in this study, much cheaper techniques could be used once the biomarker is used in clinical practice. For instance, we could use real-time PCR, which is much cheaper compared to the sequencing approach, to determine the relative abundance of the genes involved in these 8 gene pairs to classify a ccRCC tumor sample since we only need to detect their REOs.

Real-time PCR has been used for commercial mRNA-based gene signatures. One FDA-approved signature is AlleloMap®, consisting of 20 genes, for estimating the risk of acute cellular rejection in heart transplant recipients (Starling et al., 2006). Another FDA-approved signature is Oncotype DX®, consisting 21 genes, for estimating the risk of distant recurrence in tamoxifen-treated patients with node-negative, estrogen-receptor-positive breast cancer (Paik et al., 2004). Both of the signatures used in classification of patients into different risk groups by comparing their risk scores summarized from expression levels of the signature genes with preset risk score thresholds. However, the preset thresholds from the training dataset cannot be directly applied to independent dataset since the gene expression levels are sensitive to batch effects and platform differences (Qi et al., 2016a). To solve this problem, the tissue samples must be sent to the specified laboratories for measurement with uniform quality control and data normalization, which brings great limitations for wide application of these signatures. In contrast, the relative abundance of two genes in a gene pair is much easier to be estimated by real-time PCR with appropriate performance since REOs-based biomarkers are relatively insensitive to both experimental and bioinformatics variations, which could greatly facilitate the use of REOs-based biomarker in clinical practice.

The genes involved in our classification of ccRCC tumor samples also showed closed relationship with tumor development. For instance, RPP9, one of the genes in the REOs gene pairs, plays an important role in pre-mRNA splicing and could interact with well-known oncogene PIM-1 (Maita et al., 2000). MYO9A, encodes an unconventional myosin, which is required for collective migration of epithelial cells and thus serves as a potent promotor in progression and metastasis of cancers of epithelial origin (Friedl and Gilmour, 2009; Ouderkerk and Kendell, 2014). TAZ encodes tafazzin whose overexpression promotes tumorigenicity in many cancers and its inhibition also induces tumor cell apoptosis (Chen et al., 2017a; Li et al., 2011a; Pathak et al., 2014). GADD45GIP1, also known as CRF1, encodes a nuclear-localized protein that may be induced by TP53 and regulates cell cycle. CRF1 is a novel transcriptional coactivator of STAT3 which is a well-known oncogene in multiple human cancers (Kwon et al., 2008). However, CRF1 was recruited to the upstream promoter region of TP53 to suppress cell cycle progression in HCT116 cells (Yan et al., 2017). Thus, the role of CRF1 is controversial in cancer. SNRNP200 encodes a component of U5 snRNP-specific proteins of spliceosome. It has been shown that overexpression of SNRNP200 was associated with tumor aggressiveness in prostate cancer (Jimenez-Vacas et al., 2020). P1K3CB is involved in the PI3K signaling pathway which is frequently dysregulated in cancer. Inhibition of P1K3CB remarkably suppressed cell growth, migration and enhanced apoptosis in glioblastoma (Cen et al., 2018; Pridham et al., 2018). NOLC1 functions as a chaperone for shuttling between the nucleolus and cytoplasm (Meier and Blobel, 1992). It has been reported that enhancement of NOLC1 promotes cell senescence and represses hepatocellular carcinoma cell proliferation by disturbing the organization of nucleolus (Yuan et al., 2017).

The immunotherapy is widely used in ccRCC since the disturbance of different immunological pathways were observed in ccRCC (Lopez-Beltran et al., 2018). In this study, we found some of our biomarker genes are potential targets for immunotherapy. TAZ also acts as a regulator in immune evasion of cancer by positively modulating the expression of PD-L1 in mRNA and protein level (Janse van Rensburg et al., 2018), which is a potential maker for immunotherapy. In our analysis, both TAZ and PD-L1 are significantly highly expressed in high- or low-risk samples compared with normal samples. Especially, the expression of TAZ is much higher in high-risk patients than low-risk patients, suggesting that targeting TAZ is a more promising treatment strategy for high-risk patients besides PD-L1 immunotherapy. MYO9A is highly expressed in immune cells and plays an important role in regulation of macrophage shape and motility (Bahrer et al., 2011). In our analysis, MYO9A was significantly down-regulated in high-risk samples compared with low-risk and normal samples, suggesting the innate immune response may be weak in high-risk ccRCC patients. IBTK encodes the inhibitor of Bruton tyrosine kinase which is required for B cell survival, differentiation and activation (Fiume et al., 2009). In our analysis, the expression of IBTK was increasingly down-regulated in normal, low-risk and high-risk samples, suggesting that B cell immune system is active in ccRCC. NBR1 encodes a specific autophagy receptor. It has been reported that NBR1-deficient mice have defective T helper 2 cell differentiation (Yang et al., 2010). The histone methyltransferase ASHIL is involved in the regulation of regulatory T cell polarization (Xia et al., 2017). In our analysis, both the expression of NBR1 and ASHIL were increasingly down-regulated in normal, low-risk and high-risk samples, suggesting
that T cell immune system may be impaired in ccRCC. PIK3CB is critical in neutrophil activation by immune complexes (Kulkarni et al., 2011), which is significantly down-regulated in the high-risk samples compared to low-risk and normal samples.

In conclusion, we identified two molecular subtypes of ccRCC patients with high and low-risk of mortality, and developed a REOs-based classification biomarker which could be used to classify the ccRCC patients into these subtypes at personalized level. In addition, we suggested specific treatment strategies for each subtype based on their global gene expression patterns and drug repositioning. Therefore, it is worthwhile to further explore the potential clinical use of the here identified biomarker in assisting clinical diagnosis and treatment of ccRCC patients.

4. Materials and methods

4.1. Data and preprocessing

The TCGA transcript-expression level profiles (TPM and count values) of ccRCC and matched normal kidney samples was downloaded from https://cancer günümüzse.2016 on November 27, 2018, which was quantified by Kallisto (Bray et al., 2016) based on the GENCODE reference transcriptome (version 24). The clinical information of TCGA samples was downloaded through R package TCGAbiolinks (Colaprico et al., 2016). The whole-exome sequence data of 100 ccRCC samples of patients from Japanese cohort (Sato et al., 2013) was downloaded from European Genome-phenome Archive (accession number: EGAS000001000509). BEDTools (Quinlan and Hall, 2010) was used for converting BAM to FASTQ file. Kallisto was used for estimating the count and TPM values of transcripts based on the same reference transcriptome of TCGA data. The sum value of the multiple transcripts of a gene was used as the expression value of this gene. The genes with average TPM values > 1 in ccRCC patients were analyzed.

4.2. Differential expression analysis

DESeq2 (Love et al., 2014) was used to identify DEGs between two groups. The raw count values of genes were used as input of DESeq2. The Benjamini-Hochberg (BH) procedure was used to estimate FDR.

4.3. Overlapping of two lists of DEGs

If DEG list 1 with L1 genes and DEG list 2 with L2 genes have k overlapped genes and s of these genes shows the same directions which means high expression of these genes indicates favorable/unfavorable survival or group 1/2 in both lists, the probability of observing at least s consistent genes by chance can be calculated according to the following cumulative hypergeometric distribution model:

\[ P = 1 - \sum_{i=0}^{s-1} \frac{\binom{L1}{i} \binom{L2}{L1-i}}{\binom{L}{L1}} \]  

(1)

where L represents the number of the background genes commonly detected in the datasets from which the DEGs are extracted. The two DEG lists were considered to be significantly overlapping if P < 0.05.

The concordance score s/k is used to evaluate the consistency of DEGs between the two lists. Obviously, the score ranges from 0 to 1, and the higher concordance score suggests the better consistency of two lists of DEGs.

4.4. Consensus clustering

Consensus clustering (Wilkinson and Hayes, 2010) was used for tumor classification based on the normalized expression profiles of signature genes by Z-score transformation. To achieve robust clusters, the data was resampled for 1000 times by considering 80% samples and signature genes resampling. The resampled data was transformed into a similarity matrix, termed as consensus matrix. K-means clustering was used to stratify samples based on the consensus matrix. The number of optimum cluster was determined by the lowest proportion of ambiguous clustering.

4.5. Survival analysis

The univariate Cox regression model was used to evaluate the correlation of gene expression levels with OS. Survival curves were estimated by the Kaplan-Meier method and compared with the log-rank test.

4.6. Functional enrichment analysis

GO enrichment was performed by the enrichGo function in R package ClusterProfiler (Yu et al., 2012), in which the hypergeometric distribution was used to calculate the statistical significance of biological pathways enriched with DEGs of interest.

4.7. Development of the REOs-based biomarker

In each sample, the RFO of every two signature genes (i and j) is denoted as either Gi > Gj or Gj < Gi exclusively, where Gi and Gj represent the expression values of gene i and j, respectively. For a given gene pair (Gi and Gj), we used Fisher’s exact test to evaluate whether the frequency of group 1 samples with a specific RFO pattern (Gi > Gj or Gj < Gi) was significantly different from that in group 2 samples in each training dataset. The P values are adjusted using BH procedure. The gene pairs detected with 0.05 FDR control and over 70% difference of the frequency of their REOs between two groups were denoted as reversed gene pairs. The overlapped reversed gene pairs consistently identified from all the training datasets were selected as the candidate signature gene pairs. Totally, we found 171 signature gene pairs. For each signature gene pair, according to their within-sample RFO, we classified the samples of each training dataset into high or low-risk groups and then evaluated the sensitivity and specificity of this gene pair. Here, the sensitivity is defined as the ratio of correctly identified high risk samples to all high-risk samples and the specificity is defined as the ratio of correctly identified low-risk samples to all low-risk samples. Then, from these signature gene pairs, we performed a forward selection procedure in each training dataset to search a set of gene pairs that achieved the highest F-score value, a harmonic mean of sensitivity and specificity, which is calculated as follows:

\[ F - \text{score} = \frac{2 \times (\text{sensitivity} \times \text{specificity})}{(\text{sensitivity} + \text{specificity})} \]  

(2)

Taking each of the 171 gene pairs as a seed, we added another gene pair to the biomarker at a time until the F-score did not increase. The classification rule is that a sample is classified into high or low-risk group if the majority of the REOs of the set of gene pairs within this sample vote for high or low-risk. We got 171 biomarkers based on each training dataset. Totally, we got 17,100 candidate biomarkers for all the 100 training datasets. Finally, we selected the biomarker with the lowest \((1-F_n)^2+(1-F_1)^2+\cdots+(1-F_{171})^2\) as the final biomarker, in which \(F_n\) is the F-score value in the \(n^{th}\) training dataset.

4.8. Application of CMap2 data to drug discovery

The pre- procession of CMap2 data was described in our previous study (Turani et al., 2019b). In brief, the gene expression profiles of three cell lines, HL60, MCF2 and PC3, were downloaded from https://portals.broadinstitute.org/cmap/(CMap Build 02). As shown in Figure 5, a single treatment cell line and control were selected to calculate Log2FC of genes by comparison between the two cases for each
cell line. We normalized the log2FC matrix with 12,228 genes and 683 drugs to z-score matrix and further transformed it to P value matrix. Then, the confidence score was calculated per each drug-gene interaction using the P values from three cell lines. An approximation confidence score to 1 was assumed as the higher confidence level. The drug-gene pairs with CS > 0.5 were used in further analysis. If over 80% drug-perturbed genes with CS > 0.5 were mapped to the DEGs between one subtype of cancer samples and normal samples, this drug is considered to have reverse effect on the dysregulated gene expression in this subtype of cancer samples.

Declarations

Author contribution statement

Xiangyu Li, Chen Zhang, Adil Mardinoglu: Conceived and designed the analysis; Analyzed and interpreted the data; Wrote the paper.

Beste Turanli, Kajetan Juszczak, Woonghee Kim, Muhammad Arif, Yusuke Satô, Seishi Ogawa, Hasan Turkez, Jens Nielsen, Jan Boren, Mathias Uhlen: Contributed analysis tools or data; Wrote the paper.

Funding statement

This work was supported by The Knut and Alice Wallenberg Foundation.

Competing interest 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.2020.e03440.

References
