TITLE:

# Comprehensive transcriptome analysis of HCT116 cell line with different p53 and p73 status along with further identification and validation of candidate biomarkers using TCGA and GEO databases along with prognostic analysis in Colorectal Cancer through multi-Omics approach

## Introduction:

The p53 tumor suppressor family includes the transcription factor p73, which shares significant structural and functional similarities with p53. Similar to p53, p73 is present in very small amounts but is quickly elevated in response to genotoxic stress. In addition to mediating the genotoxic stress response, p73 can bind to the p53 response regions and transactivate p53 target genes implicated in cell cycle arrest and apoptotic cell death. Additionally, p73 restoration causes a tumor-suppressing impact similar to that of p53 and is known to activate target genes independent of p53.

Global Cancer Statistics 2018 indicates that colorectal cancer (CRC) accounts for ~10% of all diagnosed cancers and cancer-related deaths in the world each year (Bray et al., 2018). According to the data from Cancer Registry Annual Report, the incidence and mortality of CRCs have been increasing in the past 10 years (Zheng et al., 2019). With the improvement of surgical methods and the launch of early tumor diagnosis and treatment, the current levels of diagnosis and treatment of CRC have been greatly improved. However, the prognosis of clinical CRC is still not optimistic. Many researches have shown that the occurrence and development of CRCs may be related to genetic, lifestyle, obesity, and environmental factors, while the exact etiology and the mechanism are still unclear (Bray et al., 2018). To further clarify the pathogenesis of CRCs and to improve the precision of treatment of CRCs, genetic research, study of tumor signaling pathways, and biological target therapy are continuing to deepen, which are gradually being applied in clinic.

As we all know, the occurrence, development, overall survival time, and recurrence and non-recurrence of tumors are not only related to the pathological type and clinical stage of the tumor but also closely related to the expression and pathway of tumor genes (Bogaert and Prenen, 2014). More and more studies

suggested that there are many abnormally expressed genes in the gene expression of CRCs, relative to normal tissues, which are closely related to the proliferation, differentiation, apoptosis, metastasis, recurrence, and survival time of CRC (Lu et al., 2012; Liu et al., 2013; Gan et al., 2018; Branchi et al., 2019). The analysis of abnormally expressed genes has very important clinical significance for the targeted therapy, prognosis analysis, and recurrence risk prediction of CRC. Currently, there have been a lot of clinical researches on tumor recurrence genes and signaling pathways, and the gene recurrent model (GRM) has been established to make up for the traditional tumor classification and staging recurrence prediction, providing more genetic information and more accurate prediction data (Chen et al., 2019; Yang et al., 2020). For example, Sun et al. (2019) found that exosomal CPNE3 showed potential implications in CRC diagnosis and prognosis. Carcinoembryonic antigen (CEA) was a recommended prognostic marker in CRC for tumor diagnosis and monitoring response to therapy (Campos-da-Paz et al., 2018). Ahluwalia et al. (2019) identified a novel 4-gene prognostic signature that had clinical utility in colorectal cancer. However, there are few clinical studies about biomarkers and gene pathways which have no risk of recurrence and tumor survival time.

p73 belongs to p53 tumour suppressor family, involved in the similar line of function being tumor suppressor in nature, like cell cycle arrest and anti apoptotic function. Because of its anti proliferative function, it is present at a very low levels and gets induced upon genotoxic stress. The former shares substantial amount of structural and functional homology with the latter. P73 is known to be tumor suppressor as p73 knock out mice is extremely vulnerable to tumor development. The transcription factor p73 is a member of the p53 tumor suppressor family and shows substantial structural and functional homology with p531 . Analogous to p53, p73 is present at very low levels but it gets swiftly induced upon genotoxic stress2 . p73 can bind to the p53 response elements and transactivate p53 target genes involved in cell cycle arrest and apoptotic cell death as well as mediate genotoxic stress response2,3 . In addition, p73 is known to activate p53-independent target genes4 and p73 restoration elicits a p53-like tumor suppressive effect5 . An extensive search of the p73 status in human primary tumors revealed that p73 mutations are detected in fewer than 0.5% of human cancers, whereas over 50% of cancers carry p53 mutations6 , making p73 an attractive target for therapeutic intervention. Furthermore, unlike p53 gene, which shows only little alternative splicing, p73 gives rise to multiple isoforms, due to alternative promoter usage and differential mRNA splicing7 . Functionally, the TAp73 isoforms closely mimic p53 in the ability to stimulate transcription of death genes and to trigger programmed cell death and has been shown to be a

bona fide tumor suppressor whereas the DNp73 isoforms are strong inhibitors of transcriptionally active p73 and p538,9 . p73 has been known to execute its tumor suppressive function by guarding the genomic stability and promoting cell cycle arrest, replicative senescence or apoptosis

In this study, we obtained transcriptome profile for HCT 116 p53-/- p73+/+ and p53-/- p73 kd cells. Further the differential gene expression lists were corroborated with gene expression profiles of colorectal cancer patients from gene expression omnibus namely gse . subsequently the differentially expressed genes were looked for gene ontology, kegg pathway enrichment analysis, co expression data and protein-protein network analysis. and candidate genes were looked for overall survival of colorectal cancer patients.

In this study, we attempted to evaluate and forecast potential CLM biomarkers. The Gene Expression Omnibus (GEO) database's gene expression profiles were first used to identify important DEGs. Through GO (Gene ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis, we specifically discovered the biological functions and signal transduction pathways of the chosen DEGs. Additionally, we created protein-protein interaction (PPI) networks and used data mining to assess the predictive values of candidate genes in CRC. Finally, we examined the gene expression profile data sets, RNA-Seq data sets from the GEO and TCGA databases, and used CNV (copy number variation), GESA (Gene Set Enrichment Analysis), IHC (immunohistochemistry), and other methods for verification to assess the reliability of the chosen candidate biomarkers. Additionally, the verification outcomes are essentially

Finally, using the Cancer Genome Atlas [TCGA] dataset, the GEPIA and UALCAN online tools were used to associate candidate genes with colorectal cancer overall survival (OS), disease-free survival (RFS), and pathological staging analysis. It was discovered that PRC1,HNRNPM DTL FANCIEXO1

UBE2,IDICER,1PTEN.PRPF19,CDC45,MKI67,EFTUD2,BLM,POLE2,PLK1 EGFRmay be significant factors in colorectal cancer overall survival and disease-free survival and may represent potential treatment targets for clinical application in the future.

We report that 6 hub genes that are found to be consistently upregulated in the double loss of tumor suppressor function i.e. p53 and p73. Transcriptomics analysis of HCT116 cells which are p53-/- p73+/+ and p53-/- p73kd reveals enrichment of around nnnnn of significantly upregulated and downregulated genes which are found to be enriched in metastasis biological pathway. Further this data was corroborated with expression data of colorectal cancer patients which are having p53 mutants and p73 being downregulated in colorectal cancer patient. Taken together this study provides us clues of hub genes that can be good biomarker for disease detection because these genes were found to be associated with disease free and overall survival of the patient.

Material and Methods:

3.1 **Isolation, Qualitative and quantitative analysis of RNA**:

RNA was isolated from samples by trizol method. The quality of the RNA was checked on 1% Formaldehyde Denaturing Agarose gel and quantified using Nanodrop 8000 spectrophotometer.

3.2 **Illumina 2 x 150 PE library preparation**:

Library was prepared using Illumina TruSeq stranded mRNA Library Preparation Kit and as per its described protocol. Briefly, mRNA was enriched from total RNA followed by fragmentation. The fragmented mRNA was converted into first-strand cDNA, followed by second-strand generation, A-tailing, adapter ligation and finally ended by limited number of PCR amplification of the adaptor-ligated libraries. Library quantification and qualification was performed using DNA High Sensitivity Assay Kit.

3.3 **Quantity and quality check (QC) of library on Bioanalyzer 2100**:

The amplified library was analyzed on Bioanalyzer 2100 (Agilent Technologies) using High Sensitivity (HS) DNA chip as per manufacturer's instructions.

3.4 **Cluster Generation and Sequencing**:

After obtaining the Qubit concentration for the library and the mean peak size from Bioanalyser profile, library was loaded into Illumina platform for cluster generation and sequencing. Paired End sequencing allows the template fragments to be sequenced in both the forward and reverse directions. The library molecules bind to complementary adapter oligos on paired-end flow cell. The adapters were designed to allow selective cleavage of the forward strands after re-synthesis of

the reverse strand during sequencing. The copied reverse strand was then used to sequence from the opposite end of the fragment.

## Bioinformatics Analysis

Data Generation on Illumina Platform

The next generation sequencing for Sample_1_Control and Sample_3_KD samples was performed on the Illumina platform which resulted in generation of high quality data.

## Reference Genome Information

The Homo sapiens genome, was downloaded from NCBI site: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.38_GRCh38.p12/) was used for reference based mapping

## Mapping to Reference Genome

The high quality reads of Sample_1_Control and Sample_3_KD samples were separately aligned to the Homo sapiens genome using HISAT2 (version-hisat2-2.0.5) at default parameters. The software package SAMtools (version-0.1.18, http://samtools.sourceforge.net/) was used to convert sequence alignment/map (SAM) file to sorted binary alignment/map (BAM) file. Mapped reads ratio (MRR) to the reference in each dataset was calculated by applying flagstat command of SAMtools software to the BAM file.

## Transcript assembly:

In a RNA-seq experiment, it is crucial to have accurate reconstructions of all the isoforms expressed from each gene, as well as estimates of the relative abundance of those isoforms. StringTie assembles transcripts from RNA-seq reads that have been aligned to the genome, first grouping the reads into distinct gene loci and then assembling each locus into as many isoforms as are needed to explain the data. Following this, StringTie simultaneously assembles and quantify the final transcripts by using network flow algorithm and starting from most highly abundant transcripts. The GTF annotation files, containing exon structures of "known" genes, are then used to annotate the assembled transcripts and quantify the expression of known genes as well as derive clues if a novel transcript has been found in the sample. After assembling each sample, the full set of assemblies is passed to StringTie's merge function, which merges together all the gene structures found in any of the samples. This step is required because transcripts

in some of the samples might only be partially covered by reads, and as a consequence only partial versions of them will be assembled in the initial StringTie run. The merge step creates a set of transcripts that is consistent across all samples, so that the transcripts can be compared in subsequent steps.

Identification of differentially expressed genes for Sample_1_Control and Sample_3_KD samples The aligned reads were assembled by Stringtie version-1.3.3b and then the differentially expressed genes were detected and quantified by Cuffdiff version- 2.2.1 using a rigorous sophisticated statistical analysis. The expression of the genes were calculated in terms of FPKM (Fragment per kilobase per million mapped reads). Differential gene expression analysis has been carried out between Sample_1_Control vs Sample_3_KD samples.

Heatmap of Sample_1_Control and Sample_3_KD Samples:

Top 50 significantly expressed genes (i.e. highly up and highly downregulated genes) were represented in form of heatmap using pheatmap package from R software. The color coding ranges from red to green where shades of red represent highly expressed genes and shades of green represents low expressed genes.

GO Sequence Distribution of Genes

The Gene Ontology project provides controlled vocabularies of defined terms representing gene product properties. These cover three domains: Cellular Component, the parts of a cell or its extracellular environment; Molecular Function, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and Biological Process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Functional and pathway enrichment analysis. The GO (http://www.geneontology.org) database is a widely used annotation tool which provides a functional classification for genomic data, including biological process (BP), cellular component (CC), and molecular function (MF) [32].

The “Kyoto Encyclopedia of Genes and Genomes” (KEGG, http://www.genome.ad.jp/kegg/) database is used for analysis, and visualization of gene networks and functions [33]. Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) is a

comprehensive functional analysis tool used for systematic and integrative analysis of large gene sets [34]. In the present study, GO enrichment and KEGG pathway analysis were performed using the DAVID website to study the function of DEGs. Values with $p<0.05$ were considered statistically signifcant.

Results

Gene Ontology analysis was performed using the 10,874 differentially expressed genes. Blast2GO command line version 4.1 was used for the same.

Pathway analysis Ortholog assignment and mapping of the genes to the biological pathways were performed using KEGG automatic annotation server (KAAS).

Results

KEGG analysis

All the 10,874 differentially expressed genes were compared against the KEGG database using BLASTX with threshold bit-score value of 60 (default). The mapped genes represented metabolic pathways of major biomolecules such as carbohydrates, lipids, nucleotides, amino acids, glycans, cofactors, vitamins, terpenoids, polyketides, etc. The mapped genes also represented the genes involved in metabolism, genetic information processing, environmental information processing and cellular processes.

PPI network construction and analysis

The putative molecular pathways based on physical and functional interactions are examined using the Search Tool for the Retrieval of Interacting Genes, or STRING (version: 11.0, https://string-db.org). For the purpose of building a PPI network and identifying important genes, protein pairings with a total score of >0.4 were chosen in this investigation.

A total of ~15-22 GB high quality data was generated from Sample_1_Control and Sample_3_KD samples using Illumina platform.

- The high quality reads of Sample_1_Control and Sample_3_KD samples were mapped to Homo sapiens genome using HISAT2 tool. In total, ~93% of reads were mapped to the reference genome in each sample.

- StringTie assembly resulted in 12419 and 12334 transcripts in Sample_1_Control and Sample_3_KD samples respectively. Merging the transcripts from both samples resulted in 24452 transcripts.

- The Cuffdiff package was used to detect significantly DE genes between Sample_1_Control vs Sample_3_KD samples. A total of 10874 genes are differentially expressed in Sample_1_Control vs Sample_3_KD samples. In total, 454 up- regulated and 156 down-regulated genes were found to be significantly differentially expressed (P-value < 0.05).

- The GO distribution was carried out using Blast2GO command line tool. 2478 genes were assigned to Biological Process, 2387 to Cellular Component and 2600 genes assigned to Molecular Function category. • The pathway analysis was carried out using KAAS server (KEGG Automatic Annotation Server). The genes were enriched in different functional pathway categories which were predominantly categorized into Metabolism, Genetic Information Processing, Environmental Information Processing and Cellular Processes. A total of 738 genes contributed to the activity of Signal transduction pathway followed by 443 genes were involved in Cancer pathway.

Results:

Transcriptome Heat Map:

HeatMap obtained using pheatmap package from R software reveals Top 50 significantly expressed genes (i.e. highly up and highly downregulated genes). The color coding ranges from red to green where shades of red represents highly expressed genes and shades of green represents low expressed genes. Table 1 shows the list of top50 significantly expressed gens along with their fold change.

GO Sequence Distribution of Genes

Gene Ontology (GO) enrichment analysis reveals a total of ….. enriched GO categories using a p value cut off< 0.05 for Gene Ontology (GO) enrichment analysis revealed a total of 54 enriched GO categories using a p value cutof ≤ 0.05 for 3,099 diferentially-exported transcripts strong association with regulation of ion transport, GTPase activity and axonogenesis among the major BPs; regulation of transporter and tyrosine phosphatase receptor activity among the major MFs; and asymmetric synapse, synaptic membrane and glutamatergic

synapse among the major CCs. Whereas, among 10,200 diferentially-exported transcripts showed enrichment of 56 GO, out of which the transcripts belonged to 28 BPs, 12 MFs and 16 CCs categories (Fig. 5B). Regulation of GTPase activity, post synaptic transmission, sodium ion transmembrane transport were the major BPs, cation channel complex, cell-cell adherens junctions, post synaptic density, emerged as top CCs; and GTPase activity, regulation of neurotransmitter activity were the major MFs that carried the highest z-score. KEGG pathway analysis of diferentially-exported transcripts identifed 9 pathways in HPV-positive exosomal transcript sets. Te exported transcripts were mainly associated with the regulation of the calcium signaling, cAMP signaling, axon guidance, leukocyte transendothelial migration, circadian entrainment, longterm potentiation, glutamatergic synapse, GnRH secretion and morphine addiction.

## Pathway analysis

Ortholog assignment and mapping of the genes to the biological pathways were performed using KEGG automatic annotation server (KAAS). All the 10,874 differentially expressed genes were compared against the KEGG database using BLASTX with threshold bit-score value of 60 (default). The mapped genes represented metabolic pathways of major biomolecules such as carbohydrates, lipids, nucleotides, amino acids, glycans, cofactors, vitamins, terpenoids, polyketides, etc. The mapped genes also represented the genes involved in metabolism, genetic information processing, environmental information processing and cellular processes. A total of 738 genes contributed to the activity of Signal transduction pathway followed by 443 genes were involved in Cancer pathway.

## Identification of DEGs in transcriptome analysis, gene expression omnibus and TCGA datsets