Phenotypic associations of consensus driver genes in hepatocellular carcinoma

Kumardeep Chaudhary,[1,3] Liangqun Lu,[1,2,3] Olivier B Poirion,[1] Travers Ching,[1,2] Lana X Garmire[1,2*]

[1]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA

[2]Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI 96822, USA

[3]These authors contributed equally to this work

Authors' email addresses:

LL: liangqunlu@gmail.com

KC: kchaudhary@cc.hawaii.edu

OBP: OPoirion@cc.hawaii.edu

TC: TChing@cc.hawaii.edu

LXG: lgarmire@cc.hawaii.edu (*Corresponding author)

Abstract

Hepatocellular carcinoma (HCC) is the main form of malignant liver cancer with poor survival. Although some critical driver aberrations were identified in HCC, neither the transcriptomic nor the clinical associations of many of these driver genes have been well-characterized. We used the state of art driver detection methods MutSigCV and OncodriveFM to identify 11 consensus driver genes across six HCC cohorts and 1,494 samples in total. The consensus driver genes include *TP53*, *CTNNB1*, *ALB*, *AXIN1*, *RB1*, *ARID1A*, *RPS6KA3*, *ACVR2A*, *NFE2L2*, *CDKN2A* and *HNF1A*. Integrative analysis of driver mutations, copy number variations and transcriptomic data reveals that these driver genes are associated with the majority (63%) of the mRNA transcriptome, but only a small fraction (9%) of miRNAs. Genes associated with *TP53*, *CTNNB1*, *ARID1A* and *HNF1A* mutations contribute to four most densely connected clusters of biological pathways. Phenotypically, these driver genes are significantly associated with patients' overall survival. Some of these driver genes are significantly associated with gender (*TP53*, *CTNNB1*, *ALB*), grade (*TP53*, *ALB*, *RB1*) and age (*CTNNB1*, *AXIN1*, *RB1*) in multiple cohorts. In summary, this study reveals the vast impacts of driver gene mutations (genotype) in HCC phenotypes, which may be valuable therapeutic targets of HCC.

Keywords: Driver genes, hepatocellular carcinoma, genotype to phenotype, transcriptomics

Background

Liver cancer is the second leading cause of cancer death worldwide, with more than 700,000 incidences and deaths in recent years [1]. Globally, this cancer is ranked second for cancer-related mortality among men [2]. In the US, it is one of the few cancers with increased rate of ~3% per year, for both incidence and mortality [3]. Hepatocellular carcinoma (HCC) is the prominent histological type of liver cancer and accounts for approximately 75%-90% of all the liver cancer cases [4]. The incidence rates of HCC vary by factors such as race, gender, age as well as demographic regions. East Asians are twice likely to develop liver cancer compared to Caucasian or African American populations [5]. Additionally, males have 2 to 4 times higher incidence rates than females. The incidence rates peak around 60-65 years for males and 65-70 for females [6,7]. Various other risk factors for the HCC development have been well-determined, such as cirrhosis, hepatitis B (HBV) infection, hepatitis C (HCV) infection, alcohol abuse, obesity and environmental toxic intake [8]. While HBV infection is the major risk for HCC cases in East Asian countries, HCV and alcohol abuse are the leading causes of HCC in North America and Europe [9].

The initiation and advancement of cancer are thought to occur after continuous accumulations of somatic genomic alterations, followed by the widespread manifestation on gene products [10–13]. Using the whole genome sequencing (WGS) or whole exome-sequencing (WES) technology, many studies have aimed to determine candidate driver gene mutations in HCC, the type of mutations that confer a selective growth advantage to the cell [14–20]. *TP53* and *CTNNB1* are reported as the two most frequently mutated genes in HCC [21]. Other putative driver genes include those related to genome stability, such as *ARID1A*, *ARID2*, and *MLL1-4*

[15,17,22–24], *RB1* in cell cycle pathway [16], *AXIN1* in Wnt signaling pathway [25], *NFE2L2* in oxidative stress [22], and *TSC1/TSC2* in MAPK signaling pathway [16,22]. A recent analysis of Hepatocellular Carcinoma from The Cancer Genome Atlas (TCGA) reported the significant mutation of *LZTR1* (encoding an adaptor of CUL3-containing E3 ligase complexes) and *EEF1A1* (encoding eukaryotic translation elongation factor), apart from previously reported *CTNNB1*, *TP53* and *ALB* genes [26]. However, given the high heterogeneity of HCC populations due to race, risk factors etc., a consensus list of driver genes among different HCC cohorts are yet to be identified. Moreover, the impact of driver mutations on HCC phenotypes, such as gene expression, have not been adequately investigated.

To address these issues, we have collectively analyzed six HCC cohorts to derive 11 most significant consensus driver genes with significant functional impacts. To examine the impact of driver mutations on gene expression, we performed comprehensive analysis on driver mutation, copy number variation (CNV), gene expression and miRNA (miR) expression. Subsequent KEGG pathways and network analysis for these genes identified alterations in a broad spectrum of functions ranging from metabolic pathways, cell cycle to signaling pathways, as well as functional differences among the mutually exclusive driver genes. At the phenotypic level, we observed that consensus putative driver genes are predictive of survival differences among patients from cohorts with survival data. Some putative driver genes are significantly associated with physiological and clinical characteristics such as gender, age, and tumor grade. In summary, we present the comprehensive picture of the functional relevance of driver genes in HCC, from molecular to phenotypic levels.

Methods

Dataset and processing: We used public domain HCC data from The Cancer Genome Atlas (TCGA) available at Genomic Data Commons (GDC) data portal, as of March 2017. In total, RNA-seq, CNV and miR-seq data comprise 371, 371 and 369 tumor samples, respectively. We used the R package TCGA-Assembler (v2.0) [27] to download the TCGA data. The mRNA-seq data are represented as the normalized gene expression RSEM (RNA-Seq by Expectation Maximization) quantification values obtained from Illumina HiSeq assay platform, while miR-seq data include ‘reads per million miR mapped’ (RPM) quantification values from Illumina HiSeq assay platform. CNV data represent gene-level copy number values obtained by taking the average copy number of genomic regions of a gene from the Affymetrix SNP Array 6.0 assay platform. To handle the missing values, we performed three steps. First, we removed the biological features (i.e. genes/miRs) if they were missing in more than 20% of the samples. Similarly, we removed the samples if they were missing for more than 20% of the features. Second, we used K nearest neighbor based imputation using R *impute* package [28] to fill out the missing values. Last, we removed the genes with very low expression values (i.e. with RSEM/RPM<=10 in the remaining samples). For TCGA mutation profile, the comprehensive Mutation Annotation File (LIHC-TP.final_analysis_set.maf) was downloaded from the FireBrowse portal of the Broad institute. We retrieved 362 samples (with HCC histology) having paired tumor and normal adjacent tissue WES data. Additionally, we obtained WES data from Liver Cancer (France): LICA-FR (n=236), Liver Cancer (NCC, Japan): LINC-JP (n=244) and Liver Cancer (China): LICA-CN (n=163) cohorts, and WGS data from Liver Cancer (RIKEN, Japan): LIRI-JP (n=258), all available as simple somatic mutation files from the International

Cancer Genome Consortium (ICGC) web portal. These data from ICGC liver cohorts were published in the previous studies [16,22,29]. Besides ICGC, we obtained another WES dataset (KOREAN (n=231) form the early-stage HCCs (patients with surgical resection) with clinical information of patients published earlier [18].

Consensus driver genes detection: To achieve the pool of consensus driver genes among six cohorts, we implemented the IntOGen platform (v3.0.6) [30], a comprehensive standalone pipeline for the identification of driver genes. The mutation profiles, from six cohorts, were subjected to MutSigCV (v1.4) [31] and OncodriveFM [32], both incorporated in the IntOGen pipeline. MutSigCV represents an advanced version of MutSig tool, which seeks to identify genes with significant positive selection during tumorigenesis. It calculates the personalized and gene-specific background random mutation rates, along with the implementation of expression levels and replication times as covariate factors. Complementarily, OncodriveFM uncovers the significant mutation space by applying the functional impact based positive selection to identify the driver genes. From each module (i.e. MutSigCV and OncodriveFM) separately, we identified the genes which satisfied: (i) q-values less than the threshold cut-off ($q<0.1$) in at least 3 of 6 cohorts, and (ii) mean q-value less than the threshold cut-off ($q<0.1$), across the cohorts. The final set of drivers were obtained by the intersection of the genes found in two modules. The threshold value of q-value<0.1 for both MutSigCV and OncodriveFM was based on earlier studies [31,33]. For downstream analyses, we excluded intergenic and intronic mutations.

Determination of mutual exclusivity and co-occurrence: For each pair of consensus driver genes, we determined their association based on Fisher's exact test with a p-value <0.05. For significant associations, if the log odds ratio was more than 0 for a pair of genes, the pair was called "cooccurred", or else "exclusive". To detect the mutational exclusivity among gene sets (i.e. more than two genes), we applied the Dendrix algorithm [34] which is specialized to fish out gene sets with high coverage and exclusivity across the samples. We used gene set numbers k=4, k=5 and calculated their maximum weight with consideration of mutated genes and samples. We ran 100,000 iterations using Markov chain Monte Carlo approach to calculate empirical p-values for the top gene sets with the maximum weight.

For each cohort, we also used a bipartite graph to represent the mutations in the driver genes for each patient, using the patients and the driver genes as the distinct set of nodes. We used ForceAtlas2, a graph layout algorithm implemented in Gephi [35], to spatialize the graphs for mutual exclusivity. To compute the distances of the different cohorts the approach used is as follows: using the bipartite graph of each cohort, we computed the PageRank scores, a measure reflecting the connectivity of a node in a network [36], of the 11 driver genes. We used these scores as features representing cohorts. We then used Ward's minimum variance method to cluster both the genes and the PageRank scores.

Modeling relationships between consensus driver and gene expression: We made a binary (1, 0) matrix to indicate the mutation status of consensus driver genes in all samples. A value of 1 means the existence of at least one variant within the gene body, in the categories of nonsense, missense, inframe indel, frameshift, silent, splice site, transcription starting site and nonstop mutation. Otherwise, 0 was assigned to the gene. We made another count table of CNV data similarly. We used function voom (*limma* package in R) to transform RSEM data prior to the linear modeling [37], then fit the linear models by minimizing generalized least squares similar to others [38]. These linear models consider the effects of multiple consensus driver genes (predictors) and their CNVs on expression values of individual genes (responses) as follows:

$$y_g = \beta_{0g} + \sum_{i=0}^{n} \beta_{1i} X_{1i} + \beta_{2i} X_{2i} + \in \qquad (1)$$

Where y_g is the vector representing expression value of gene g across all the n samples, β_{0g} is that baseline value of g, X_1 and X_2 are the mutation status and CNV of the consensus driver gene g in sample i, β_1 and β_2 are coefficients for the same associated with the mutation status and CNV, respectively. We performed multiple hypothesis tests on the significance values of the coefficients across all the genes using Benjamin–Hochberg (BH) adjustment, to determine the significant association between the driver genes and expression of all the genes (BH adjusted p-value <0.05).

Pathway enrichment and network analysis: We conducted pathway enrichment analysis of the genes impacted by somatic mutations and CNVs, using R package clusterProfiler [39]. We used BH adjusted p-value=0.05 as threshold to select the over-represented KEGG pathways. We used Gephi [35] bipartite graph to visualize driver gene-enriched pathways network.

Modeling relationships between consensus drivers and miR expression: To find the relationship between driver genes (mutation and CNV) and impact of miR expression, we implemented the linear model similar to that of equation (1). Here driver genes' mutation and CNV status are treated as independent variables and miR expression as the response variable. To narrow down miRs that directly target these 11 drivers, we mined miRDB resource [40], which houses the miR-target interactions predicted by MirTarget [41] based on CLIP-Ligation experiments.

Survival analysis of driver mutations: We used the Cox proportional hazards (Cox-PH) model [42] implemented in R *survival* package for the overall survival (OS) analysis of consensus driver genes. We also conducted Cox-PH model to fit the overall effect of all 11 driver genes on OS, with or without adjustments of clinical and physiological parameters (e.g. age, gender, and grade). For this, we used R *glmnet* package [43], since it enables penalization through ridge regression. We performed cross-validation to obtain the optimal regularization hyperparameter. The hyperparameter was selected by minimizing the mean cross-validated partial likelihood. To evaluate the performance of the survival models [44]. We calculated the concordance index (CI) using function *concordance.index* in R *survcomp* package [45], based on Harrell's C-statistics [46]. We dichotomized the samples into high- and low-risk groups based on the median prognosis index (PI) score, the fitted survival values of the Cox-PH model [47–49]. In the case of ties for the median PI, we shuffled the samples and randomly assigned them to either risk groups. We plotted the Kaplan-Meier survival curves for the two risk groups and calculated the log-rank p-value of the survival difference between them. We performed the similar survival analysis by adjusting the Cox-PH model with different physiological and clinical factors (e.g. age, gender, grade, race, tumor stage and risk).

Results

Detection of consensus driver genes

To identify the consensus pool of drive genes among multiple cohorts of diverse populations, we used paired tumor-normal tissue of HCC WES data from TCGA as well as five other cohorts (WES/WGS). The clinical summary of patients in these 6 cohorts is listed in **Table S1**. We assessed mutation significance and functional impact of protein coding genes using MutSigCV and OncodriveFM modules implemented in the IntOGen pipeline (see Materials and Methods). From the two integrative modules of this pipeline, we obtained 11 and 21 genes with mean q-value <0.1 from MutSigCV and OncodriveFM, respectively. By intersecting results in these two gene sets, we identified 11 common genes as “consensus driver genes” (see Methods) (**Figure 1A**). Among these 11 genes, *TP53* and *CTNNB1* are most significantly mutated and functionally impactful genes based on q-values (**Figures 1B & 1C**), consistent with the observations earlier [18,21]. However, some low-frequency mutation genes also have significant rankings per MutSigCV (**Figure 1B**). For examples, *HNF1A*, *CDKN2A*, *NFE2L2* and *ACVR2A* are all significant (mean q-values: 0.03, 0.02, 0.005 and 0.0009 respectively), although their average mutation frequencies are less than 5% (**Figure 1D**). Thus, this workflow efficiently detects less frequent but consistently important driver genes.

Meta-analysis of consensus driver genes among cohorts

Next, we explored the mutation exclusivity status among these 11 driver genes across different populations (**Figure 2A**). We used colored tiles in the plot to represent the specific type of mutation (e.g. missense, silent, frame shift etc.). A similar trend of mutation distribution exists in TCGA, three ICGC cohorts with large sample size (i.e. LINC-JP, LIRI-JP and LICA-FR) and the KOREAN cohort (**Figures 2A (i), (ii), (iii), (iv) and (v)**). Worth mentioning, LICA-CN (n=163) data have the lowest *CTNNB1* mutation rate from the other cohorts (**Figure 2A (vi)**). This exception may be attributable to HBV infection in LICA-CN cohort, as previous studies of HBV patients also reported rare existence of *CTNNB1* mutations [22,23]. The top mutated genes in the six cohorts show clear evidence of mutual exclusivity (**Figure 2A**). For example, *CTNNB1* and *TP53* mutations are mutually exclusive in four of six cohorts, with significant Fisher's exact test p-values in TCGA (P=0.0303), LICA-FR (P= 0.0166), KOREAN (P=0.006) and LINC-JP cohort (P=0.057). To detect mutual exclusivity beyond two genes, we used the tool Dendrix [34]. Again we observed significant mutational exclusivities (p-value=<0.05) for up to five genes in all 6 cohorts (**Figure S1**). *TP53, CTNNB1, RB1* and *AXIN1* and another cohort-specific gene are mutually exclusive in all five cohorts except LICA-CN. This other driver is *HNF1A* for TCGA and LICA-FR cohorts and *CDKN2A* for LINC-JP, LIRI-JP and KOREA cohorts. Compared to the five cohorts, LICA-CN cohort has most different five mutual exclusive drivers: *TP53*, *ACVR2A*, *ALB*, *CDKN2A*, and *RPS6KA3*.

We further visualized the mutual exclusivity using bipartite graphs, where patients, driver genes, and their topologies are highlighted in a connective graph (**Figure 2B**). The blue nodes and the labeled nodes represent patients and driver genes, respectively, and the edge between them indicates the existence of certain driver in a particular patient. Based on the PageRank score that measures the connectivity and topologies of the graphs (see Materials and Methods), the similarity between TCGA and the other cohort descends in the following order: LINC-JP > LIRI-JP> LICA-FR > KOREAN > LICA-CN (**Figure S2**). KOREAN and LICA-CN cohorts are most distinct from other cohorts, with much fewer patients showing mutations in at least two driver genes. While KOREAN cohort mostly mutates in *TP53* and *CTNNB1* (however lacking *ALB* mutations like the other three cohorts), LICA-CN most dominantly mutates in *TP53* but not *CTNNB1* or *ALB* (**Figure 2B (vi) and S2**).

The associations between gene expression and consensus driver gene mutation/CNV

To obtain the lists of genes impacted by the consensus drivers at the transcriptional level, we built linear models using these driver genes' mutation profile and their CNVs as the predictors, whereas gene expression values as the response variables, similar to another study [50]. These genetics based models decently predict gene expression values (R^2=0.58) (**Figure 3A**), indicating that albeit the complex genetics and epigenetics regulatory mechanisms of gene expression, HCC driver gene mutations still convey important functional impacts on gene expression. Overall, our results show that around 63% (13,036) of genes are significantly associated (BH adjusted p-value <0.05) with these consensus driver genes. We list the number of genes significantly associated to each consensus driver gene in these linear models in **Figure 3B**. The top two mutated genes are *TP53* and *CTNNB1* as expected, both affecting over four thousand genes. Strikingly, the CNV of *ARID1A* is ranked 4th and linked to expression changes in over 2,800 genes, despite its relatively low mutation rate of ~5%.

To investigate the biological processes that these 13,036 genes are involved in, we conducted KEGG pathway enrichment analysis and detected 108 significantly (BH adjusted p-values <0.05) associated pathways (**Figure 3C**). We further categorized these pathways into 6 super groups according to the KEGG pathway organization [51], namely: cellular processes, environmental information processing, genetic information processing, metabolism, human diseases, and organismal systems [51]. It is not surprising that the pathway super-group affected most by the consensus driver genes belongs to metabolic pathways. Among the driver genes, *TP53*, *CTNNB1*, *ARID1A* and *HNF1A* are most densely connected to enriched pathways, due to the associations with gene expression changes. Some signaling pathways in the environmental information processing group are significantly influenced by driver genes, especially *CTNNB1*, which is associated with PI3K-Akt pathway, Wnt pathway and CGMP-PKG signaling pathway. The association network between driver genes and pathways provide functional explanations for mutual exclusivities observed earlier, at least partially, in that certain pathways are commonly associated by two mutually exclusive drivers. Between the well-known mutually exclusive *TP53* and *CTNNB1*, multiple pathways in amino acid and glucose metabolism are shared. *TP53* and *ARID1A* are both involved in amino acid and fatty acid metabolism pathways. *CTNNB1-ARID1A-HNF1A* share CYP450 metabolism pathway.

Associations between miR expression and consensus driver gene mutation/CNV

We extended the linear modeling approach described earlier to examine the association between consensus driver genes and miR expression. As a result, we found 164 miRs significantly associated with the consensus drivers. Among them, 120 miRs are associated with driver gene CNV-level changes, 88 miRs are associated with the driver mutations, and 44 miRs are associated with both of them. **Figures 4A and B** show the bipartite graphs of associations between miRs and driver gene mutations and CNVs, respectively. Overall, *CTNNB1* and *TP53* are most influential drivers, as expected, affecting 77 and 52 miRs respectively. From the mutation perspective, *CTNNB1* is dominantly associated with the most number (60) of miRs (**Figure 4A**), whereas from the driver gene CNV perspective, *TP53* is associated with the most number (44) of miRs (**Figure 4B**). From the miR-centric perspective, *hsa-mir-616* is associated with CNVs of the most number (4) of genes *TP53*, *RB1*, *ARID1A* and *CDKN2A*. This miR was overexpressed in HCC and promoted tumorigensis [52]. On the other hand, *hsa-mir-203a* is associated with the most number (3) of genes (*TP53*, *CTNNB1* and *HNF1A*) at the mutation level.

Since the miRs discovered above are based on linear models, they may directly or indirectly target the consensus drivers. To help narrow down the miRs that may directly target these drivers, we searched miRDB resource [40], in which miR and targets are predicted by correlational analysis of thousands of miRNA-target interactions from CLIP-Seq experiments. As results, we obtained 15 miRs that are predicted to target one or more consensus driver genes (**Figure S3**). *CTNNB1* is the direct target of 6 miRs: *hsa-mir-214, hsa-mir-330, hsa-mir-885*, *hsa-mir-3591*, *hsa-mir-5586* and *hsa-mir-6715b*, where *hsa-miR-214* and *hsa-miR-885-5p* have been reported to be targeting *CTNNB1* [53,54]. *TP53* is the target of *hsa-mir-149*, which was shown to be directly up regulated by P53 [55] (**Figure S3A**). *ARID1A* is the potential target of *hsa-mir-181b-1*, *hsa-mir-1976*, *hsa-mir-92a-2* and *hsa-mir-511* (**Figure S3B**). Future experimental studies are warranted to decipher if these miRs are indeed associated with driver genes at primary target level, or secondary, tertiary levels.

Associations between consensus driver genes and survival outcome

In order to test survival associations from all the driver mutations, we built multivariate Cox-PH models on overall survival in each of the four cohorts that have survival data (TCGA, LINC-JP, LIRI-JP and LICA-FR). We used the median PI score generated from the Cox-PH model as the threshold [47], and divided samples into high and low risk groups accordingly. The Kaplan-Meier survival curves of the two groups are shown in **Figure 5**. For all the cohorts with survival data, the log-rank P-values between the Kaplan-Meier curves are significant (TCGA: P=0.0063, C-index-0.58; LICA-FR: P=0.0068, C-index=0.60; LINC-JP: P=0.013, C-index=0.67 and LIRI-

JP: P=0.004, C-index=0.64). To avoid potential confounding from age, gender, grade, stage in all 4 cohorts, as well as risk factor and race in TCGA cohort, we adjusted the Cox-PH model by these variables accordingly. Still, we identified significantly different survival groups (TCGA: P=4.3e-05, LINC-JP: P=3.1e-02, LIRI-JP: P=6e-06 and LICA-FR: P=7.6e-04) (**Figure S4**). All together, these results show that the driver gene mutational status is associated with HCC patients' overall survival.

Associations between consensus driver genes and other physiological and clinical characteristics

To reveal the possible associations of these driver genes with physiological and clinical characteristics of patients, such as gender, age, and grade, risk factors etc., we conducted Fisher's exact tests for the categorical variables, and Mann-Whitney-Wilcoxon test for the continuous age variable. With regard to gender, *CTNNB1*, *ALB* and *TP53* are statistically more frequently mutated in males than females, in 5, 3 and 2 out of 6 cohorts, respectively (**Figure 6A**). The statistical significance of bias towards male is especially striking for *CTNNB1*, a proto-oncogene, with the strongest association from the TCGA cohort (P-value=1.5e-05). For associations with grade, *TP53* mutation frequency increases with higher tumor grades in 4 out of 6 cohorts, where the most significant association comes from LINC-JP cohort (P-value=1e-05) (**Figure 6B**). Four other genes, *CTNNB1*, *ALB*, *RB1*, and *CDKN2A* also have significant associations with grade in one or two cohorts. For age, again *CTNNB1* is the driver gene with the strongest associations, for both significance level and the number of cohorts (4 out of 6) (**Figure 6C**). It has the strongest association in the KOREAN cohort (P-value=1.3e-04). Interestingly, *RB1* is another driver gene significantly but inversely related to age (3 out of 6 cohorts).

Discussion

In this study, we have pushed forward our understanding of the molecular and clinical associations of HCC drivers using multiple cohorts. Despite the heterogeneity among the datasets, we identified eleven consensus driver genes derived from HCC WES/WGS data. Anchoring on these consensus driver genes, we investigated in-depth their transcriptomic (mRNA and miRs) impacts, physiological and clinical associations, and prognosis values.

A major contribution of this study is to associate the drivers with transcriptomic changes, which was previously unknown. The consensus driver genes are correlated to around 63% mRNA transcriptome. These influenced genes are involved in various pathways in cell cycle and DNA repair, metabolism, and signaling transduction. Interestingly, network analysis results show that the mutually exclusively mutated genes have effects on some common biological processes, which may explain why mutations in both genes do not usually co-occur within the same patient.

Surprisingly, only 9% of miRs are associated with the consensus drivers globally, suggesting the major and direct role of driver mutations is on protein coding genes rather than regulatory components such as miRs. For the miRs that do link with consensus drivers, the relationship is more significant at the CNV level of drivers, rather than mutation level of drivers. This makes sense, as the change of copy numbers of drivers are integers, supposed to be greater than mutations that are modulating.

Our analysis reveals some unusual findings on genes with low mutation frequencies. One of them is that *ARID1A* CNV is one of the most “effective” events in driver genes, prevalently associated with transcriptomic changes of 2,803 genes. *ARID1A* is a chromatin remodeller which is involved in transcriptional activation and considered as tumor suppressor [56]. Previously, this gene is reported to be frequently deleted in HCC [20,57]. *ARID1A* mutations in HCC have been reported to be associated HCC progression and metastasis in HBV- and alcohol-related HCC [23,58]. Other infrequently mutated genes *HFN1A* and *ACVR2A* have also been reported in individual studies much less frequently [16,29]. Our stringent criteria for selection of consensus driver genes among 6 HCC cohorts highlights these low mutated genes with consensus, reflecting that these may play a crucial role in HCC etiology. Along with *TP53* and *CTNNB1*, two genes *HNF1A* and *ARID1A* stand out with densely connected sub networks.

We highlight some low-frequency mutation genes that were less studied in HCC before, such as *ARID1A*, *HNF1A*, *ACVR2A* and *NFE2L2*. These mutations all together are significantly associated with HCC patients' survival in all cohorts that have survival records, providing the promise that therapeutics targeting these mutations may improve patient survival. Moreover, we have found extensive evidence that these driver mutations are associated with gender bias, age and tumor grade, thus warranty more patient-specific treatment plans. *CTNNB1* is more frequent in males, and it increases with age, and tumor grades. *TP53* is more frequent in males, and also increases with age. Interestingly, we also found that *ALB* mutations are biased in males and increases with grade, to a less degree than *CTNNB1*. Another interesting driver gene *RB1* is significantly associated with grade, however, reversely related to ages of HCC patients. We do not know the etiology of such reversal age dependency of *RB1* in HCC. However, it has been well known that mutations in both alleles of the *RB1* gene are essential for retinoblastoma, which often diagnosed in neonatals.

In summary, we have identified a consensus list of 11 driver genes in HCC and their associations with molecular and clinical phenotypes. Albeit the heterogeneity and complexity of HCC, the driver genes have broad and significant impacts on global gene expression and molecular pathway functions. Future mechanistic studies will be needed to understand the causality between driver mutations and impacted genes and miRs. Nevertheless, this study provides an important reference for driver mutations in HCC patients, which may be valuable for HCC prognosis prediction. It may even provide insights on combinatorial targeted therapies, which are severely lacking in HCC.

Conclusions

We have identified 11 consensus driver genes in HCC from six international cohorts, representing the largest endeavor on genotype-phenotype analysis. These driver genes and physiological and clinical factors show significant associations, particularly on patient survival, gender, age and risk factors. Moreover, the 11 driver genes exert impacts on the majority of mRNA transcriptome. Altogether, this study provides a list of driver genes in HCC that may be good therapeutic targets.

Abbreviations

BH: Benjamini-Hochberg; CI: Concordance Index; CNV: Copy Number Variation; Cox-PH: Cox-Proportional Hazards; GDC: Genomic Data Commons; HBV: Hepatitis B Virus; HCC: Hepatocellular carcinoma; HCV: Hepatitis C Virus; ICGC: International Cancer Genome Consortium; LICA-CN: Liver Cancer (China); LICA-FR: Liver Cancer (France); LINC-JP: Liver Cancer (NCC, Japan); LIRI-JP: Liver Cancer (RIKEN, Japan); miR: micro RNA; miRDB: MicroRNA Target Prediction And Functional Study Database; OS: Overall Survival; RPM:

Reads per million miR mapped; RSEM: RNA-Seq by Expectation Maximization; TCGA: The Cancer Genome Atlas; WES: Whole Exome-Sequencing; WGS: Whole Genome Sequencing

Declarations

Ethics approval and consent to participate

This study is secondary-data analysis from all de-identified and publicly available data obtained from TCGA consortium, thus the informed consent and IRB approval was not required.

Consent for publication

Not required

Availability of data and material

All HCC data are downloaded from the TCGA portal (https://tcga-data.nci.nih.gov/tcga/ and https://portal.gdc.cancer.gov/), FireBrowse portal from Broad institute (http://firebrowse.org/) and ICGC (https://dcc.icgc.org/).

Competing interests

The authors declare no competing interests.

Funding

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20

COBRE GM103457 awarded by NIH/NIGMS, R01 LM012373 awarded by NLM, R01 HD084633 awarded by NICHD and Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to L.X. Garmire.

Authors' contributions

LG conceived the project. LL performed data analysis, with the help of KC and TC. LL, KC and LG wrote the manuscript. OP made the bipartite graphs. All authors have read and revised the manuscript.

Acknowledgements

The authors would also like to thank Dr. Herbert Yu, Dr. Maarit Tiirikainen as well as all the other group members in Garmire lab for helpful discussions and suggestions.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer. 2015;136:E359–86.

2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA. Cancer J. Clin. 2011;61:69–90.

3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA. Cancer J. Clin. 2016;66:7–30.

4. London WT, Mcglynn KA. Liver Cancer. Cancer Epidemiol. Prev. Oxford University Press; 2006. p. 763–86.

5. Llovet JM, Zucman-Rossi J, Pikarsky E, Sangro B, Schwartz M, Sherman M, et al. Hepatocellular carcinoma. Nat Rev Dis Prim. 2016;2:16018.

6. El-Serag HB, Rudolph KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology. 2007;132:2557–76.

7. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015;65:5–29.

8. European Association For The Study Of The Liver, European Organisation For Research And Treatment Of Cancer. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol. 2012;56:908–43.

9. Laursen L. A preventable cancer. Nature. 2014;516:S2-3.

10. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481:306–13.

11. Sidow A, Spies N. Concepts in solid tumor evolution. Trends Genet. 2015;31:208–14.

12. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015;21:846–53.

13. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011. p. 646–74.

14. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW. Cancer genome landscapes. Science (80-.). 2013;339:1546–58.

15. Li M, Zhao H, Zhang X, Wood LD, Anders RA, Choti MA, et al. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet. 2011;43:828–9.

16. Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. 2015;47:505–11.

17. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet. 2012;44:760–4.

18. Ahn SM, Jang SJ, Shim JH, Kim D, Hong SM, Sung CO, et al. Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. Hepatology. 2014;60:1972–82.

19. Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, et al. High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet. 2011;43:464–9.

20. Kan Z, Zheng H, Liu X, Li S, Barber TD, Gong Z, et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res. 2013;23:1422–33.

21. Shibata T, Aburatani H. Exploration of liver cancer genomes. Nat Rev Gastroenterol Hepatol. 2014;11:340–9.

22. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet. 2012;44:694–8.

23. Huang J, Deng Q, Wang Q, Li KY, Dai JH, Li N, et al. Exome sequencing of hepatitis B

virus-associated hepatocellular carcinoma. Nat Genet. 2012;44:1117–21.

24. Cleary SP, Jeck WR, Zhao X, Chen K, Selitsky SR, Savich GL, et al. Identification of driver genes in hepatocellular carcinoma by exome sequencing. Hepatology. 2013;58:1693–702.

25. Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat Genet. 2014;46:1267–73.

26. Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. Elsevier; 2017;169:1327–1341.e23.

27. Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods. 2014;11:599–600.

28. Hastie T, Tibshirani R NB and CG. impute: impute: Imputation for microarray data. 2017.

29. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet. 2016;48:500–9.

30. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013;10:1081–2.

31. Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.

32. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012;40:e169.

33. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576–82.

34. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2012;22:375–85.

35. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One. 2014;9:e98679.

36. Altman A, Tennenholtz M. Ranking systems. Proc. 6th ACM Conf. Electron. Commer. - EC ’05. New York, New York, USA: ACM Press; 2005. p. 1–8.

37. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.

38. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

39. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.

40. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res. Oxford University Press; 2015;43:D146-52.

41. Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. Bioinformatics. 2016;32:1316–22.

42. Cox DR. Regression Models and Life-Tables. J. R. Stat. Soc. Ser. B Stat. Methodol. 1972;34:187–220.

43. Friedman J, Hastie T, Tibshirani R, Calhoun VD, Deng H-W, Wang Y-P. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. BioMed Central; 2010;33:1–22.

44. van Houwelingen HC, Bruinsma T, Hart AAM, van’t Veer LJ, Wessels LFA. Cross-

validated Cox regression on microarray gene expression data. Stat. Med. 2006;25:3201–16.

45. Schroder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011;27:3206–8.

46. Harrell Jr. FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.

47. Huang S, Yee C, Ching T, Yu H, Garmire LX. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. Rzhetsky A, editor. PLoS Comput. Biol. 2014;10:e1003851.

48. Huang S, Chong N, Lewis NE, Jia W, Xie G, Garmire LX. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. Genome Med. 2016;8:34.

49. Wei R, De Vivo I, Huang S, Zhu X, Risch H, Moore JH, et al. Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. Oncotarget. 2016;7:55249–63.

50. Gerstung M, Pellagatti A, Malcovati L, Giagounidis A, Porta MG, Jadersten M, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nat Commun. 2015;6:5901.

51. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

52. Zhang D, Zhou P, Wang W, Wang X, Li J, Sun X, et al. MicroRNA-616 promotes the migration, invasion and epithelial-mesenchymal transition of HCC by targeting PTEN. Oncol. Rep. 2016;35:366–74.

53. Xia H, Ooi LLPJ, Hui KM, Patel B, Majumdar A. MiR-214 Targets β-Catenin Pathway to

Suppress Invasion, Stem-Like Traits and Recurrence of Human Hepatocellular Carcinoma. Gottardi C, editor. PLoS One. Public Library of Science; 2012;7:e44206.

54. Zhang Z, Yin J, Yang J, Shen W, Zhang C, Mou W, et al. miR-885-5p suppresses hepatocellular carcinoma metastasis and inhibits Wnt/β-catenin signaling pathway. Oncotarget. Impact Journals; 2016;7:75038–51.

55. Jin L, Hu WL, Jiang CC, Wang JX, Han CC, Chu P, et al. MicroRNA-149*, a p53-responsive microRNA, functions as an oncogenic regulator in human melanoma. Proc. Natl. Acad. Sci. U. S. A. 2011;108:15840–5.

56. Wu JN, Roberts CWM. ARID1A mutations in cancer: another epigenetic tumor suppressor? Cancer Discov. NIH Public Access; 2013;3:35–43.

57. Zhao J, Chen J, Lin H, Jin R, Liu J, Liu X, et al. The Clinicopathologic Significance of BAF250a (ARID1A) Expression in Hepatocellular Carcinoma. Pathol. Oncol. Res. 2016;22:453–9.

58. Nahon P, Nault J-C. Constitutional and functional genetics of human alcohol-related hepatocellular carcinoma. Liver Int. 2017;0:1–11.

Figure Legends

Figure 1. Consensus driver genes in 6 HCC cohorts. (A) IntOGen pipeline to identify consensus driver genes. (B) Final 11 genes with mean q-value <0.1 from MutSigCV module. (C) Same 11 genes with mean q-value <0.1 from OncodriveFM module. (D) Percentage of sample coverage of driver gene mutations.

Figure 2. Mutual exclusivity among different driver genes in 6 HCC cohorts. (A) Co-mutation plots for the 6 HCC cohorts, where each colored tile represents one of the mutation types (i.e. frame shift, in-frame indel, missense, exonic, nonsense, splice site, silent or mixture of mutations) (i) TCGA (ii) LINC-JP (iii) LIRI-JP (iv) LICA-FR (v) KOREAN (vi) LICA-CN cohorts. (B) Bipartite graphs for mutual exclusivity of the same cohorts in (A). Blue nodes represent the patients and the other labeled nodes represent consensus driver genes, whose size is proportional to their degree.

Figure 3. Associations of consensus driver genes with mRNA expression. (A) Correlation between observed and predicted gene expression. (B) The number of genes whose expression values are significantly associated with the driver gene mutation/CNV statuses. (C) Enriched KEGG pathways network among significant genes as shown in (B). The thickness of edges is proportional to the -log10 p-adjusted value.

Figure 4. Associations of consensus driver genes with miR expression. Bipartite graphs representing (A) correlations between miR expression and mutations of the genes. (B) Correlations between miR expression and CNV. Green color represents positive correlation and red for anti-correlation.

Figure 5. Kaplan-Meier estimates of overall survival (OS) in 4 HCC cohorts. A Cox-PH regression was used to build the overall survival model featuring the driver genes mutation profile. The samples were dichotomized into high and low risk groups by the median Prognostic Index (PI). (A) TCGA (B) LINC-JP (C) LIRI-JP (D) LICA-FR cohorts.

Figure 6. Associations of clinical factors with driver genes. Shown are genes with significant associations to physiological and clinical factors, including gender, grade (Fisher's exact test with p-value< 0.05) and age (Mann-Whitney-Wilcoxon test with p-value<0.05).

Supplementary Material

Supplementary Figures

Figure S1. Bipartite graph showing the distribution of mutually exclusive genes identified with Dendrix across the 6 cohorts. Purple and yellow nodes represent genes and cohorts, respectively where the size of the node is proportional to the connectivity.

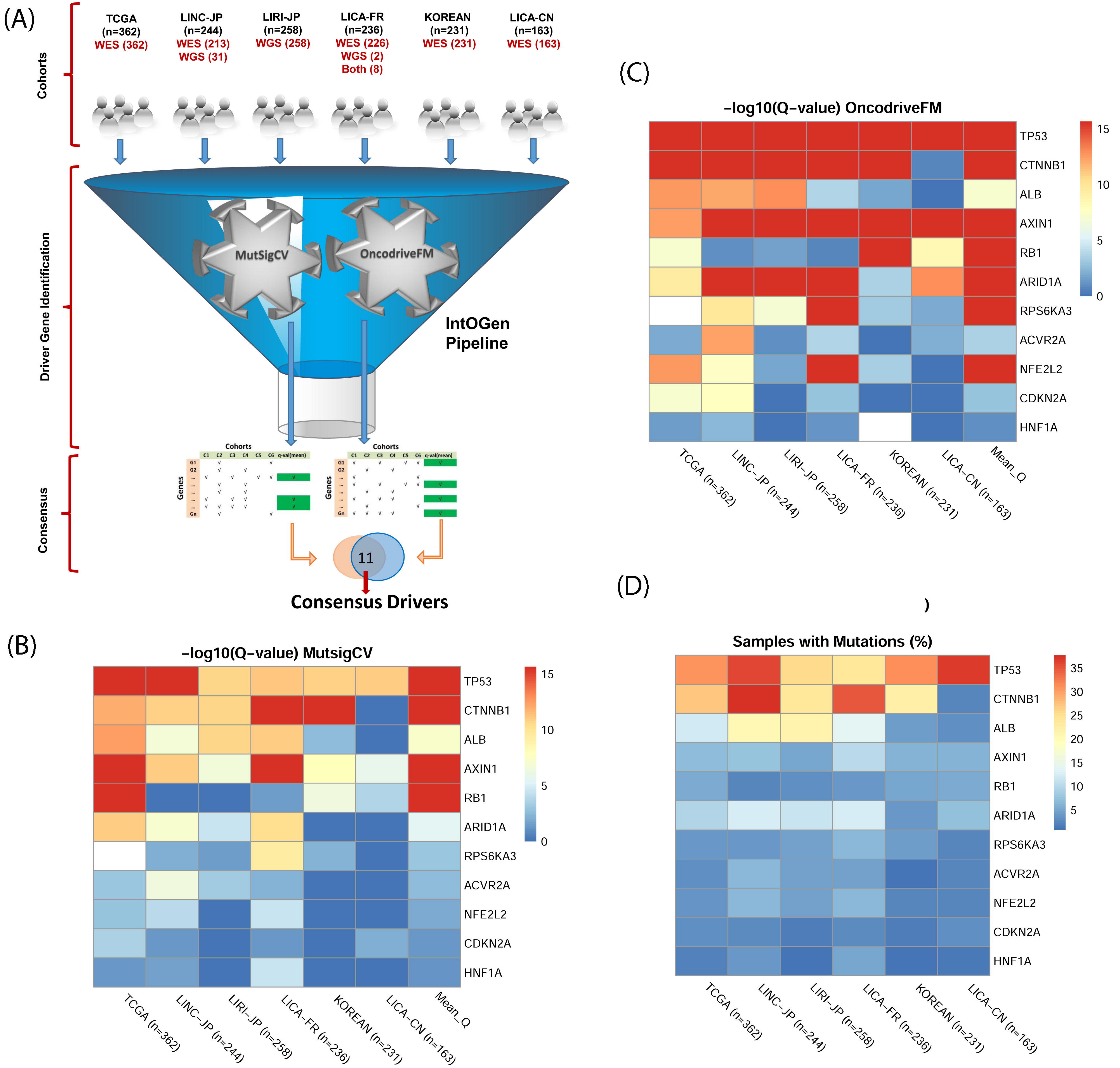
Figure S2. Similarity of the bipartite graphs among 6 HCC cohorts based on pageRank scores. The darker color represents more predominant mutations.

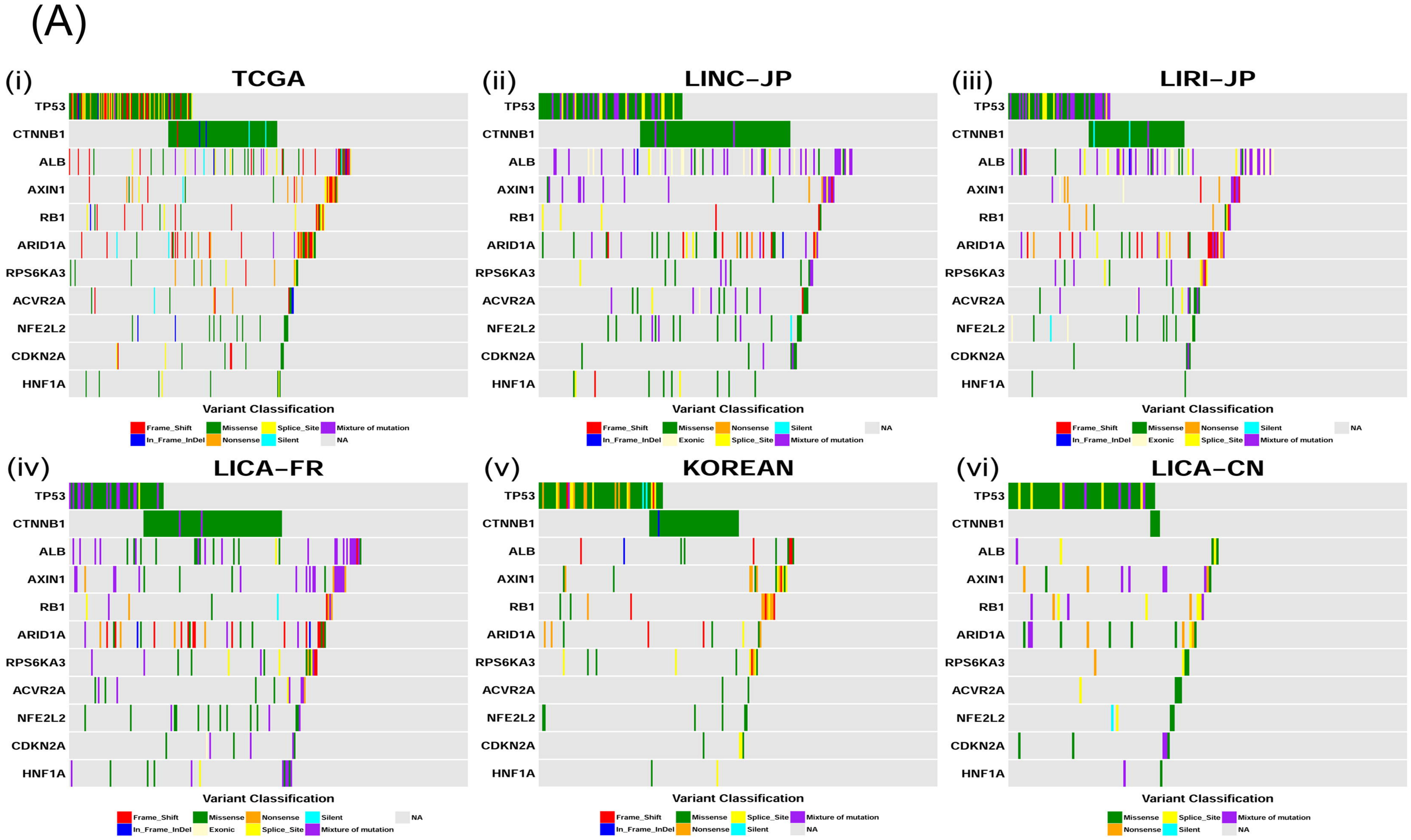
Figure S3. miRs significantly associated with consensus drivers as well as being predicted to target them directly. (A) association between the miRs and predicted targets among 11 driver genes mapped on the background bipartite graph of miRs and driver genes' mutations and (B) association between the miRs and predicted targets among 11 driver genes mapped on the background bipartite graph of miRs and driver genes' CNV. Red colored edges represent the miRs with a non-null correlation with their predicted targets. The edge thicknesses are proportional to the absolute value of the correlations.

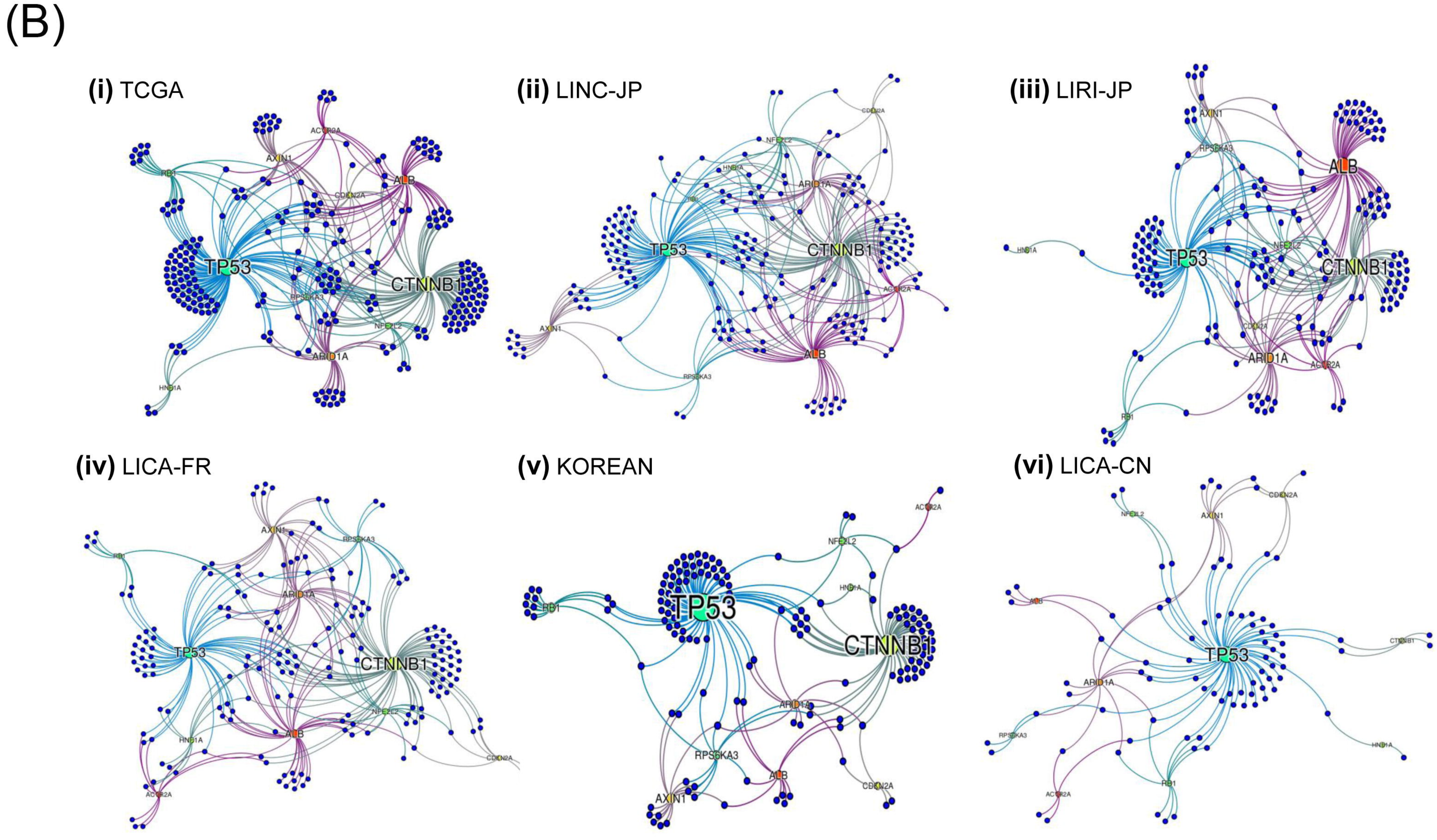
Figure S4. Kaplan-Meier of overall survival (OS) after adjustment for gender, age, stage, grade, race, and risk factors. (A) TCGA (B) LINC-JP (C) LIRI-JP (D) LICA-FR cohorts.

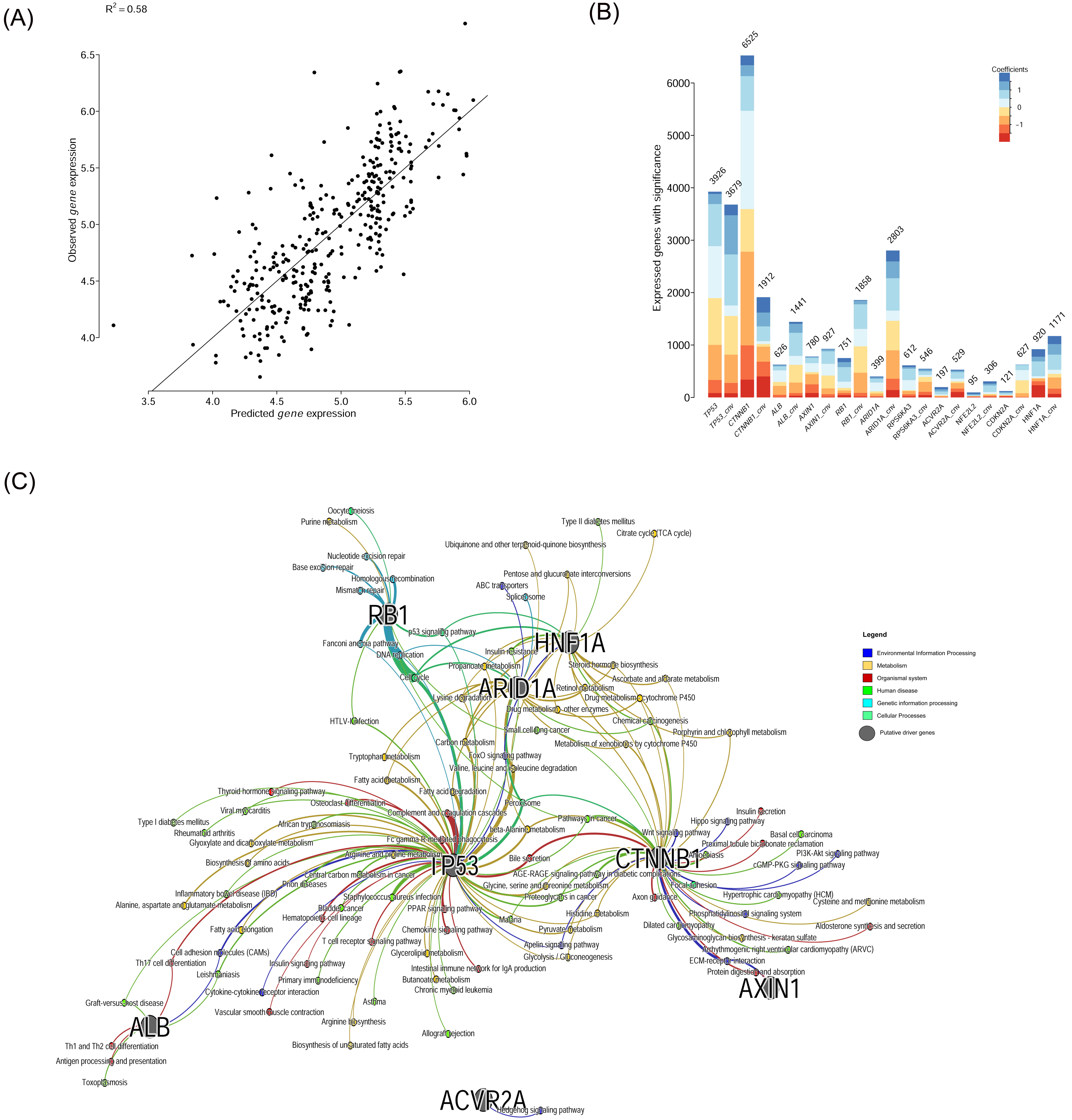
Supplementary Table

Table S1. Clinical summary of 6 HCC cohorts.

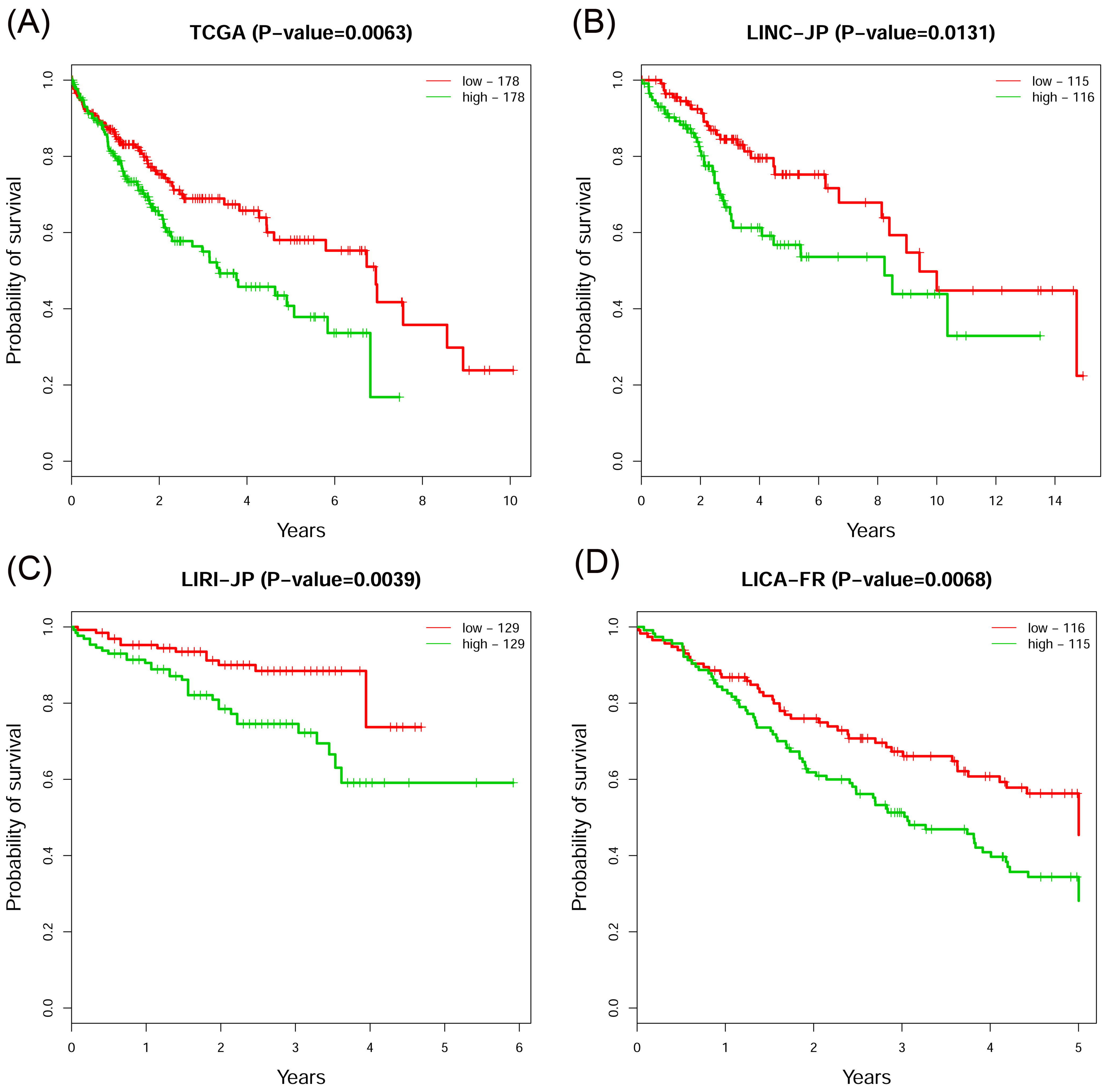

(A)
Cohorts
Driver Gene Identification
Consensus
TCGA (n=362) WES (362)
LINC-JP (n=244) WES (213) WGS (31)
LIRI-JP (n=258) WGS (258)
LICA-FR (n=236) WES (226) WGS (2) Both (8)
KOREAN (n=231) WES (231)
LICA-CN (n=163) WES (163)
MutSigCV
OncodriveFM
IntOGen Pipeline
Cohorts
C1 C2 C3 C4 C5 C6 q-val(mean)
Genes
G1
G2
Gn
11
Consensus Drivers
(B)
-log10(Q-value) MutsigCV
TP53
CTNNB1
ALB
AXIN1
RB1
ARID1A
RPS6KA3
ACVR2A
NFE2L2
CDKN2A
HNF1A
15
10
5
0
TCGA (n=362)
LINC-JP (n=244)
LIRI-JP (n=258)
LICA-FR (n=236)
KOREAN (n=231)
LICA-CN (n=163)
Mean_Q
(C)
-log10(Q-value) OncodriveFM
TP53
CTNNB1
ALB
AXIN1
RB1
ARID1A
RPS6KA3
ACVR2A
NFE2L2
CDKN2A
HNF1A
15
10
5
0
TCGA (n=362)
LINC-JP (n=244)
LIRI-JP (n=258)
LICA-FR (n=236)
KOREAN (n=231)
LICA-CN (n=163)
Mean_Q
(D)
)
Samples with Mutations (%)
TP53
CTNNB1
ALB
AXIN1
RB1
ARID1A
RPS6KA3
ACVR2A
NFE2L2
CDKN2A
HNF1A
35
30
25
20
15
10
5
TCGA (n=362)
LINC-JP (n=244)
LIRI-JP (n=258)
LICA-FR (n=236)
KOREAN (n=231)
LICA-CN (n=163)


(A)
(i) TCGA
(ii) LINC–JP
(iii) LIRI–JP
(iv) LICA–FR
(v) KOREAN
(vi) LICA–CN
TP53
CTNNB1
ALB
AXIN1
RB1
ARID1A
RPS6KA3
ACVR2A
NFE2L2
CDKN2A
HNF1A
Variant Classification
Frame_Shift
In_Frame_InDel
Missense
Nonsense
Splice_Site
Silent
Exonic
Mixture of mutation
NA


(B)
(i) TCGA
(ii) LINC-JP
(iii) LIRI-JP
(iv) LICA-FR
(v) KOREAN
(vi) LICA-CN
TP53
CTNNB1
ALB
AXIN1
RB1
ARID1A
RPS6KA3
ACVR2A
NFE2L2
CDKN2A
HNF1A


(A)
$R^2 = 0.58$
Observed *gene* expression
Predicted *gene* expression
6.5
6.0
5.5
5.0
4.5
4.0
3.5
4.0
4.5
5.0
5.5
6.0
(B)
Expressed genes with significance
Coefficients
1
0
−1
6000
5000
4000
3000
2000
1000
0
3926
3679
6525
1912
626
1441
780
927
751
1858
399
2803
612
546
197
529
95
306
121
627
920
1171
TP53
TP53_cnv
CTNNB1
CTNNB1_cnv
ALB
ALB_cnv
AXIN1
AXIN1_cnv
RB1
RB1_cnv
ARID1A
ARID1A_cnv
RPS6KA3
RPS6KA3_cnv
ACVR2A
ACVR2A_cnv
NFE2L2
NFE2L2_cnv
CDKN2A
CDKN2A_cnv
HNF1A
HNF1A_cnv
(C)
Legend
Environmental Information Processing
Metabolism
Organismal system
Human disease
Genetic information processing
Cellular Processes
Putative driver genes
RB1
HNF1A
ARID1A
TP53
CTNNB1
AXIN1
ALB
ACVR2A
Oocyte meiosis
Purine metabolism
Type II diabetes mellitus
Citrate cycle (TCA cycle)
Ubiquinone and other terpenoid-quinone biosynthesis
Nucleotide excision repair
Base excision repair
Homologous recombination
Mismatch repair
Pentose and glucuronate interconversions
ABC transporters
Spliceosome
p53 signaling pathway
Fanconi anemia pathway
DNA replication
Insulin resistance
Propanoate metabolism
Steroid hormone biosynthesis
Cell cycle
Ascorbate and aldarate metabolism
Retinol metabolism
Lysine degradation
Drug metabolism - cytochrome P450
Drug metabolism - other enzymes
HTLV-I infection
Chemical carcinogenesis
Small cell lung cancer
Porphyrin and chlorophyll metabolism
Metabolism of xenobiotics by cytochrome P450
Carbon metabolism
Tryptophan metabolism
FoxO signaling pathway
Valine, leucine and isoleucine degradation
Fatty acid metabolism
Fatty acid degradation
Thyroid hormone signaling pathway
Peroxisome
Osteoclast differentiation
Insulin secretion
Viral myocarditis
Complement and coagulation cascades
Pathways in cancer
Hippo signaling pathway
Type I diabetes mellitus
African trypanosomiasis
Wnt signaling pathway
Basal cell carcinoma
Rheumatoid arthritis
beta-Alanine metabolism
Proximal tubule bicarbonate reclamation
Glyoxylate and dicarboxylate metabolism
Fc gamma R-mediated phagocytosis
PI3K-Akt signaling pathway
Amoebiasis
Arginine and proline metabolism
Bile secretion
cGMP-PKG signaling pathway
Biosynthesis of amino acids
AGE-RAGE signaling pathway in diabetic complications
Central carbon metabolism in cancer
Prion diseases
Focal adhesion
Glycine, serine and threonine metabolism
Inflammatory bowel disease (IBD)
Hypertrophic cardiomyopathy (HCM)
Staphylococcus aureus infection
Axon guidance
Proteoglycans in cancer
Cysteine and methionine metabolism
Alanine, aspartate and glutamate metabolism
Bladder cancer
PPAR signaling pathway
Hematopoietic cell lineage
Histidine metabolism
Phosphatidylinositol signaling system
Malaria
Dilated cardiomyopathy
Aldosterone synthesis and secretion
Fatty acid elongation
Chemokine signaling pathway
Pyruvate metabolism
Glycosaminoglycan biosynthesis - keratan sulfate
T cell receptor signaling pathway
Apelin signaling pathway
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Cell adhesion molecules (CAMs)
Glycerolipid metabolism
Glycolysis / Gluconeogenesis
ECM-receptor interaction
Th17 cell differentiation
Insulin signaling pathway
Intestinal immune network for IgA production
Protein digestion and absorption
Leishmaniasis
Primary immunodeficiency
Butanoate metabolism
Chronic myeloid leukemia
Cytokine-cytokine receptor interaction
Asthma
Graft-versus-host disease
Vascular smooth muscle contraction
Arginine biosynthesis
Allograft rejection
Th1 and Th2 cell differentiation
Biosynthesis of unsaturated fatty acids
Antigen processing and presentation
Toxoplasmosis
Hedgehog signaling pathway

(A)
CDKN2A
mir-144
mir-486
mir-451a
ACVR2A
mir-320b-2
AXIN1
mir-29a
mir-132
ARID1A
ALB
mir-4433
mir-1295b
mir-361
mir-374b
mir-151a
mir-542
mir-107
mir-330
mir-450b
mir-222
mir-27a
mir-181a-1
mir-221
mir-103a-2
mir-10a
mir-676
NFE2L2
mir-181b-1
mir-4661
mir-766
mir-450a-1
mir-944
mir-204
mir-199b
mir-218-2
mir-195
mir-552
mir-373
mir-497
mir-375
mir-32
mir-199a-2
mir-181a-2
mir-548b
mir-216b
CTNNB1
mir-2114
mir-217
mir-499a
mir-624
mir-371a
mir-550a-1
mir-372
RPS6KA3
mir-466
mir-885
mir-6715a
mir-551b
mir-708
mir-214
mir-122
mir-199a-1
mir-148a
mir-3591
mir-92b
mir-34a
mir-6718
mir-216a
mir-181c
mir-1269a
mir-5586
mir-577
RB1
mir-181d
mir-767
mir-105-2
mir-3144
mir-23a
mir-193a
mir-30b
mir-200b
mir-429
mir-4636
mir-30a
mir-200a
HNF1A
mir-6715b
mir-203a
mir-2355
mir-194-1
mir-26b
mir-192
TP53
mir-1269b
mir-194-2
mir-215
mir-5589
mir-149
🟩 Correlated
🟥 Anti-correlated
(B)
mir-16-2
mir-340
mir-3682
mir-107
mir-145
mir-423
mir-4772
mir-4636
mir-15b
mir-577
mir-4652
mir-3074
mir-377
mir-154
mir-210
RB1
mir-10a
mir-487a
mir-758
NFE2L2
mir-26b
RPS6KA3
mir-505
mir-576
mir-4662a
mir-140
mir-629
mir-548x
mir-942
mir-335
mir-548b
let-7i
mir-181a-1
mir-34a
mir-4661
mir-3144
mir-374b
mir-27a
mir-186
mir-181a-2
mir-23a
mir-181b-1
mir-375
mir-616
mir-150
mir-1976
ARID1A
mir-378c
mir-574
mir-92a-2
mir-181c
mir-181d
mir-934
mir-99b
mir-598
mir-511
mir-200a
mir-429
mir-1269b
mir-193a
CTNNB1
mir-21
ALB
mir-30c-1
mir-92b
mir-378a
mir-191
mir-194-2
mir-200b
mir-1266
mir-29b-1
mir-216b
mir-125a
mir-425
mir-483
mir-29b-2
mir-155
mir-222
mir-216a
mir-122
mir-142
mir-1180
mir-194-1
mir-148a
mir-744
mir-146a
CDKN2A
mir-192
TP53
mir-101-1
mir-30a
mir-217
mir-592
mir-342
mir-1269a
mir-5589
HNF1A
mir-4433
mir-551b
mir-1295b
mir-135b
mir-324
mir-29c
mir-215
mir-29a
mir-223
mir-625
mir-455
mir-28
mir-5586
mir-1270-1
mir-500a
mir-3130-1
mir-4524a
mir-125b-2
mir-501
mir-651
mir-219a-1
mir-30e
mir-450a-2
mir-221
AXIN1
ACVR2A
mir-196a-1
mir-3127
mir-3200
mir-1226
mir-193b
mir-3937
mir-195
mir-103a-1


(A)
TCGA (P−value=0.0063)
low − 178
high − 178
Probability of survival
Years
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
(B)
LINC−JP (P−value=0.0131)
low − 115
high − 116
Probability of survival
Years
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
(C)
LIRI−JP (P−value=0.0039)
low − 129
high − 129
Probability of survival
Years
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
(D)
LICA−FR (P−value=0.0068)
low − 116
high − 115
Probability of survival
Years
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5

(A)
Gender
-log10(P-value)
5
4
3
2
TP53
CTNNB1
ALB
AXIN1
ARID1A
RPS6KA3
Cohort
TCGA
LINC-JP
LIRI-JP
LICA-FR
KOREAN
LICA-CN
Gender
Male
Female
(B)
Grade
-log10(P-value)
5
4
3
2
TP53
CTNNB1
ALB
RB1
CDKN2A
Grade
G2
G2/3
G3
(C)
Age
-log10(P-value)
3
2
TP53
CTNNB1
ALB
AXIN1
RB1
ARID1A
ACVR2A
CDKN2A
Group
Higher
Lower