Multi-modal meta-analysis of 1494 hepatocellular carcinoma samples reveals vast impacts of consensus driver genes on phenotypes

Kumardeep Chaudhary[1,$], Liangqun Lu[1,2,$], Olivier B Poirion[1], Sijia Huang[1,2], Travers Ching[1,2] and 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

[$]These authors contributed equally to this work

[*]Corresponding author

E-mail: lgarmire@cc.hawaii.edu

Abstract

To characterize the phenotypic associations of driver genes in hepatocellular carcinoma (HCC), we 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 are associated with 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 driver genes are significantly linked to HCC gender and age disparities. *CTNNB1*, *ALB* and *TP53* have higher relative risks (RR=1.2, 1.1 and 1.1) in males. Oppositely, *AXIN1*, which encodes axin-1, a component of the beta-catenin (encoded by *CTNNB1*) destruction complex has higher RR (1.6) in females. *RB1* mutations are more frequent in younger patients (RR=1.3). This study consolidates a group of consensus driver genes in HCC, which collectively show vast impacts on the phenotypes. These driver genes may warrant as valuable therapeutic targets of HCC.

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

Introduction

Liver cancer is the second leading cause of cancer death worldwide, with more than 700,000 incidences and deaths in recent years (Ferlay et al., 2015). Globally, this cancer is ranked second for cancer-related mortality among men (Jemal et al., 2011). In the US, it is one of the few cancers with increased rate of ~3% per year, for both incidence and mortality (Rebecca L. Siegel, Miller, & Jemal, 2016). Hepatocellular carcinoma (HCC) is the prominent histological type of liver cancer and accounts for approximately 75%-90% of all the liver cancer cases (London & Mcglynn, 2006). 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 (Llovet et al., 2016). 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 (El-Serag & Rudolph, 2007; R L Siegel, Miller, & Jemal, 2015). 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 (European Association For The Study Of The Liver & European Organisation For Research And Treatment Of Cancer, 2012). 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 (Laursen, 2014).

The initiation and advancement of cancer are thought to occur after continuous accumulations of somatic genomic alterations, followed by the widespread manifestation of gene products

(Alizadeh et al., 2015; Greaves & Maley, 2012; Hanahan & Weinberg, 2011; Sidow & Spies, 2015). 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 (Ahn et al., 2014; Fujimoto et al., 2012; Kan et al., 2013; Li et al., 2011; Schulze et al., 2015; Totoki et al., 2011; Vogelstein et al., 2013). *TP53* and *CTNNB1* are reported as the two most frequently mutated genes in HCC (Shibata & Aburatani, 2014). Other putative driver genes include those related to genome stability, such as *ARID1A*, *ARID2*, and *MLL1-4* (Cleary et al., 2013; Fujimoto et al., 2012; Guichard et al., 2012; J. Huang et al., 2012; Li et al., 2011), *RB1* in cell cycle pathway (Schulze et al., 2015), *AXIN1* in Wnt signaling pathway (Totoki et al., 2014), *NFE2L2* in oxidative stress (Guichard et al., 2012), and *TSC1/TSC2* in MAPK signaling pathway (Guichard et al., 2012; Schulze et al., 2015). 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 (Ally et al., 2017). 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 of 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 and age. 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) (Zhu, Qiu, & Ji, 2014) 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 (Hastie T, Tibshirani R, 2017) 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 (Fujimoto et al., 2016; Guichard et al., 2012; Schulze et al., 2015). Besides ICGC, we obtained another WES dataset (KOREAN (n=231) from the early-stage HCCs (patients with surgical resection) with clinical information of patients published earlier (Ahn et al., 2014).

Consensus driver genes detection: To achieve the pool of consensus driver genes among six cohorts, we implemented the IntOGen platform (v3.0.6) (Gonzalez-Perez et al., 2013), a comprehensive standalone pipeline for the identification of driver genes. The mutation profiles, from six cohorts, were subjected to MutSigCV (v1.4) (Lawrence et al., 2013) and OncodriveFM (Gonzalez-Perez & Lopez-Bigas, 2012), 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 of q-value<0.1 for both MutSigCV and OncodriveFM was based on earlier studies (Cancer Genome Atlas Network, 2015; Lawrence et al., 2013). 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 "co-occurred", or else "exclusive". To detect the mutational exclusivity among gene sets (i.e. more than two genes), we applied the Dendrix algorithm (Vandin, Upfal, & Raphael, 2012) 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 (Jacomy, Venturini, Heymann, &

Bastian, 2014), to spatialize the graphs for mutual exclusivity. To compute the distances of the different cohorts the approach used was 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 (Altman & Tennenholtz, 2005), 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 table of CNV data similarly. We used voom function (*limma* package in R) to transform RSEM data prior to the linear modeling (Law, Chen, Shi, & Smyth, 2014), then fit the linear models by minimizing generalized least squares similar to others (Ritchie et al., 2015). These linear models consider the effects of mutations 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 (Yu, Wang, Han, & He, 2012). We used BH adjusted p-value=0.05 as threshold to select the over-represented KEGG pathways. We used Gephi (Jacomy et al., 2014) 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 (Wong & Wang, 2015), which houses the miR-target interactions predicted by MirTarget (Wang, 2016) based on CLIP-Ligation experiments.

Survival analysis of driver mutations: We used the Cox proportional hazards (Cox-PH) model (Cox, 1972) 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, grade etc.). For this, we used R *glmnet* package (Friedman et al., 2010), 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 (van Houwelingen, Bruinsma, Hart, van't Veer, & Wessels, 2006), we calculated the concordance index (CI) using function *concordance.index* in R *survcomp* package (Schroder, Culhane, Quackenbush, & Haibe-Kains, 2011), based on Harrell's C-statistics (Harrell Jr., Lee, & Mark, 1996). 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 (S. Huang et al., 2016; S. Huang, Yee, Ching, Yu, & Garmire, 2014; Wei et al., 2016). 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 provided (Supp. Table S1). We assessed mutation significance and functional impact of protein coding genes using MutSigCV and OncodriveFM modules implemented in the IntOGen pipeline (see Methods) (Figure 1A). We identified the driver genes among the individual cohorts with the stringent

threshold i.e. q-value <0.1 for both MutSigCV and OncodriveFM. Among these cohorts, TCGA contains the maximum number of drivers (20), while LICA-CN has 3 drivers only. LINC-JP, LIRI-JP, LICA-FR and KOREAN cohorts comprise 13, 14, 12 and 7 driver genes, respectively. *TP53* and *AXIN1* are two the driver genes shared by all the 6 cohorts.

Next, we set out to define the "consensus driver gene", which satisfy these criteria: (1) q-values are less than the threshold cut-off (q<0.1) in at least 3 of 6 cohorts, and (2) the mean q-value is less than the threshold cut-off (q<0.1) across the cohorts. As a result, we identified 11 out of total 30 genes as "consensus driver genes" (Figure 1B). Among these 11 genes, *TP53* and *CTNNB1* are most significantly mutated and functionally impactful genes based on q-values (Figure 1C and D), consistent with the observations earlier (Ahn et al., 2014; Shibata & Aburatani, 2014). However, some low-frequency mutation genes also have significant rankings per MutSigCV (Figure 1C). 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 1E). 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 (Figure 2A (i), (ii), (iii), (iv) and (v)). Worth mentioning, LICA-CN cohort (n=163) is most distinct from others, and has the lowest *CTNNB1* mutation rate among all

(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 (Guichard et al., 2012; J. Huang et al., 2012).

Mutual exclusivity is apparent among some drivers (Figure 2A). For example, *CTNNB1* and *TP53* mutations are mutually exclusive in three of six cohorts, with significant Fisher's exact test p-values in TCGA (P=0.0303), LICA-FR (P= 0.0166) and KOREAN (P=0.006). The mutual exclusivity between them was documented earlier (Ahn et al., 2014). To detect mutual exclusivity beyond two genes, we used the Dendrix tool (Vandin et al., 2012). Again we observed significant mutational exclusivities (p-value=<0.05) for up to five genes in all 6 cohorts (Supp. Figure S1). *TP53, CTNNB1, RB1* and *AXIN1* and another cohort-specific genes are mutually exclusive in all five cohorts except LICA-CN. The other cohort-specific driver is *HNF1A* (TCGA and LICA-FR) and *CDKN2A* (LINC-JP, LIRI-JP and KOREAN). Compared to the other five cohorts, LICA-CN cohort has most different five mutually exclusive drivers: *TP53*, *ACVR2A*, *ALB*, *CDKN2A*, and *RPS6KA3*.

We further visualized the relationships among patients, driver genes, and their topologies, using bipartite graphs (Figure 2B). The blue nodes and the labeled nodes represent patients and driver genes, respectively, and the edges between them indicate the existence of certain drivers in a particular patient. Based on the PageRank score that measures the connectivity and topologies of the graphs (see Methods), the similarity between TCGA and the other cohort descends in the following order: LINC-JP > LIRI-JP > LICA-FR > KOREAN > LICA-CN (Supp. 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 in *CTNNB1* or *ALB* (Figure 2B (vi) and Supp. Figure S2).

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

To assess the impact of the consensus drivers on the transcriptome level, we built generalized 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 earlier study (Gerstung et al., 2015; Vervier & Michaelson, 2016). 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 (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 <10%.

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, namely: cellular processes, environmental information processing, genetic information processing, metabolism, human diseases, and organismal systems (Kanehisa & Goto, 2000). 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 further support 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. Heatmap of driver genes and six pathways classes (Figure 3C, insert) shows that *TP53* has the impact on maximum number of pathways related to Metabolism and Diseases, followed by *CTNNB1*, *ARID1A* and *HNF1A*.

We extended the linear modeling approach described earlier to examine the association between consensus driver genes and miRNA (miR) expression. Contrary to the vast prevalence of correlations between mRNAs and consensus drivers, we only found 164 miRs that are significantly associated with these 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 (Supp. Figure S3). This suggests that the major impact of driver mutation is manifested in protein coding genes, rather than non-coding regulatory elements miRs. The detailed association analysis between miR expression and consensus driver gene mutation/CNV is described in the supporting information (Supp. File S1).

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 prognostic index (PI) score generated from the Cox-PH model as the threshold (S. Huang et al., 2014), and divided samples into high and low risk groups accordingly. The Kaplan-Meier survival curves of the two groups are generated (Figure 4). For all the cohorts with survival data, the log-rank P-values between the Kaplan-Meier curves are significant (TCGA: P=6.3e-03, C-index-0.58; LINC-JP: P=1.3e-02, C-index=0.67; LIRI-JP: P=4e-03, C-index=0.64 and LICA-FR: P=6.8e-03, C-index=0.60). 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 or almost significantly different survival groups (TCGA: P=1.02e-04, LINC-JP: P=2.2e-02, LIRI-JP: P=6.5e-02 and LICA-FR: P=1.9e-02) (Supp. Figure S4). All together, these results show that the driver gene mutational status is associated with HCC patients' overall survival.

Associations of consensus driver genes with gender and age disparities

Previous studies have revealed aspects of disparities in HCC, such as preferable incidents in males (El-Serag & Rudolph, 2007; McGlynn & London, 2011). To reveal the possible link between these driver genes and gender/age, we conducted Fisher's exact tests for gender, and Mann-Whitney-Wilcoxon tests for the continuous age variable. We found some significant associations of driver genes with gender and age (Figure 5). To directly illustrate differences between categories, we calculated the relative risk of each category for the mutated genes vs.

wild type genes. RR for age was calculated after dichotomizing the samples based on mean age in respective cohorts.

With regard to gender, *CTNNB1*, a proto-oncogene, shows the most consistent evidence of preferred mutations in males for all 5 cohorts (Figure 5A-E). Its strongest association comes from the TCGA cohort, based on significance level (P-value=1.5e-05) and relative risk (RR=1.4) (Figure 5A). Interestingly, *AXIN1* shows opposite and higher relative risks in females in 2 cohorts LIRI-JP (RR=2.2) and KOREAN (RR=2.2) (Figure 5C and E). Other drivers, such as *ALB* and *TP53*, are also preferred in males from 3 and 2 cohorts, respectively. For age, again *CTNNB1* is the driver gene with the strongest positive associations, for both relative risks (average RR=1.2) and the number of cohorts (4 out of 6) (Figure 5G-J). Interestingly, *RB1* is the driver gene significantly and preferably prevalent in younger patients (3 out of 6 cohorts) (Figure 5F, H and J). However, *AXIN1* shows controversial associations with age between LINC-JP and LICA-FR cohorts, which may have to do with the different ethnicities between the two.

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 impacts, phenotypic associations, and prognostic values at systematic level. Detailed molecular mechanisms for each consensus drivers, although of interest to follow up, are not the focus of this multi-modal meta-analysis report.

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 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.

Our analysis reveals some unusual findings on genes with low mutation frequencies. One of them is that the CNV of *ARID1A* is one of the most "effective" events in the 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 (Wu & Roberts, 2013). Previously, this gene is reported to be frequently deleted in HCC (Kan et al., 2013; Zhao et al., 2016). *ARID1A*, a tumor suppressor gene, is depleted in advanced HCC and hence promotes angiogenesis via angiopoietin-2 (Ang2). *ARID1A*-deficient HCCs have been suggested as a good target for anti-angiogenesis therapies e.g. using sorafenib (Hu et al., 2017). *ARID1A* mutations in HCC have been reported to be associated HCC progression and metastasis in HBV- and alcohol-related HCC (J. Huang et al., 2012; Nahon & Nault, 2017). Other infrequently mutated genes *HFN1A* and *ACVR2A* have also been reported in individual studies (Fujimoto et al., 2016; Schulze et al., 2015). 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.

Most interestingly, we have found evidence that some driver mutations are associated with gender and age disparities among HCC patients. *CTNNB1* is mostly strongly preferred in males, and it increases with age. Additionally, *TP53* and *ALB* are also more frequently mutated in males. Oppositely, *AXIN1* shows preferred mutations in females. *AXIN1* encodes tumor suppressor gene axin-1, which is part of the beta-catenin destruction complex required for regulating *CTNNB1* levels through phosphorylation and ubiquitination (Nakamura et al., 1998). The opposite trend of gender association between *AXIN1* and *CTNNB1* can be explained by their antagonism relationship. Unexpectedly, we found that driver gene *RB1* is 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 is often diagnosed in neonates.

In summary, we have identified a consensus list of 11 driver genes in HCC, as well as their associations with downstream transcriptome and patient outcomes, such as survival, age and gender. Albeit the heterogeneity and complexity of HCC, the driver genes have broad and significant impacts on global gene expression and molecular pathway functions, suggesting that HCC are genetically dominated diseases. Thus, this study provides an important and refined reference list for driver genes, which may serve as candidates for targeted therapies currently severely lacking in HCC.

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 portal (https://dcc.icgc.org/).

Acknowledgements

This research was supported by grants K01 ES025434 awarded by National Institute of Environmental Health Sciences (NIEHS) (URL: https://www.niehs.nih.gov/) through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by National Institute of General Medical Sciences (NIH/NIGMS) (URL: https://www.nigms.nih.gov/Pages/default.aspx), R01 LM012373 awarded by National Library of Medicine (NLM) (URL: https://www.nlm.nih.gov/), R01 HD084633 awarded by Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) (URL: https://www.nichd.nih.gov/Pages/index.aspx) and 14ADVC-64566 awarded by Hawaii Community Foundation Medical Research Grant (URL: https://www.hawaiicommunityfoundation.org/grants/medical-research-grants) to LXG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. 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.

Authors' Contributions

LG conceived the project. KC and LL performed data analysis, with the help of OP, SH and TC.

KC, LL and LG wrote the manuscript. All authors have read and revised the manuscript.

Conflict of Interest

Authors declare that there is no conflict of interest.

References

Ahn, S. M., Jang, S. J., Shim, J. H., Kim, D., Hong, S. M., Sung, C. O., … Kong, G. (2014). Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. *Hepatology*, *60*(6), 1972–1982. https://doi.org/10.1002/hep.27198

Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., … Zucman-Rossi, J. (2015). Toward understanding and exploiting tumor heterogeneity. *Nat Med*, *21*(8), 846–853. https://doi.org/10.1038/nm.3915

Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., … Laird, P. W. (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, *169*(7), 1327–1341.e23. https://doi.org/10.1016/j.cell.2017.05.046

Altman, A., & Tennenholtz, M. (2005). Ranking systems. In *Proceedings of the 6th ACM conference on Electronic commerce - EC '05* (pp. 1–8). New York, New York, USA: ACM Press. https://doi.org/10.1145/1064009.1064010

Cancer Genome Atlas Network. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, *517*(7536), 576–582. https://doi.org/10.1038/nature14129

Cleary, S. P., Jeck, W. R., Zhao, X., Chen, K., Selitsky, S. R., Savich, G. L., … Chiang, D. Y. (2013). Identification of driver genes in hepatocellular carcinoma by exome sequencing. *Hepatology*, *58*(5), 1693–1702. https://doi.org/10.1002/hep.26540

Cox, D. R. (1972). Regression Models and Life-Tables. *J. R. Stat. Soc. Series B Stat. Methodol.*, *34*(2), 187–220.

El-Serag, H. B., & Rudolph, K. L. (2007). Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology*, *132*(7), 2557–2576. https://doi.org/10.1053/j.gastro.2007.04.061

European Association For The Study Of The Liver, & European Organisation For Research And Treatment Of Cancer. (2012). EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. *J Hepatol*, *56*(4), 908–943. https://doi.org/10.1016/j.jhep.2011.12.001

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., … Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, *136*(5), E359–E386. https://doi.org/10.1002/ijc.29210

Friedman, J., Hastie, T., Tibshirani, R., Calhoun, V. D., Deng, H.-W., & Wang, Y.-P. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., … Nakagawa, H. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet*, *48*(5), 500–509. https://doi.org/10.1038/ng.3547

Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K. A., Hosoda, F., Nguyen, H. H., … Nakagawa, H. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet*, *44*(7), 760–764. https://doi.org/10.1038/ng.2291

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M. G., Jadersten, M., … Boultwood, J. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun*, *6*, 5901. https://doi.org/10.1038/ncomms6901

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Res*, *40*(21), e169. https://doi.org/10.1093/nar/gks743

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-

Sanz, A., … Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods*, *10*(11), 1081–1082. https://doi.org/10.1038/nmeth.2642

Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, *481*(7381), 306–313. https://doi.org/10.1038/nature10762

Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. B., … Zucman-Rossi, J. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet*, *44*(6), 694–698. https://doi.org/10.1038/ng.2256

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*.

Harrell Jr., F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, *15*(4), 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hastie T, Tibshirani R, N. B. and C. G. (2017). impute: impute: Imputation for microarray data.

Hu, C., Li, W., Tian, F., Jiang, K., Liu, X., Cen, J., … Hui, L. (2017). Arid1a regulates response to anti-angiogenic therapy in advanced hepatocellular carcinoma. *Journal of Hepatology*, *0*(0). https://doi.org/10.1016/j.jhep.2017.10.028

Huang, J., Deng, Q., Wang, Q., Li, K. Y., Dai, J. H., Li, N., … Han, Z. G. (2012). Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. *Nat Genet*, *44*(10), 1117–1121. https://doi.org/10.1038/ng.2391

Huang, S., Chong, N., Lewis, N. E., Jia, W., Xie, G., & Garmire, L. X. (2016). Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome Medicine*, *8*(1), 34. https://doi.org/10.1186/s13073-016-0289-9

Huang, S., Yee, C., Ching, T., Yu, H., & Garmire, L. X. (2014). A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of

Breast Cancer. *PLoS Computational Biology*, *10*(9), e1003851. https://doi.org/10.1371/journal.pcbi.1003851

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, *9*(6), e98679. https://doi.org/10.1371/journal.pone.0098679

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians*, *61*(2), 69–90. https://doi.org/10.3322/caac.20107

Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T. D., Gong, Z., … Mao, M. (2013). Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res*, *23*(9), 1422–1433. https://doi.org/10.1101/gr.154492.113

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, *28*(1), 27–30.

Laursen, L. (2014). A preventable cancer. *Nature*, *516*(7529), S2-3. https://doi.org/10.1038/516S2a

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218. https://doi.org/10.1038/nature12213

Li, M., Zhao, H., Zhang, X., Wood, L. D., Anders, R. A., Choti, M. A., … Kinzler, K. W. (2011). Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nat Genet*, *43*(9), 828–829. https://doi.org/10.1038/ng.903

Llovet, J. M., Zucman-Rossi, J., Pikarsky, E., Sangro, B., Schwartz, M., Sherman, M., & Gores,

G. (2016). Hepatocellular carcinoma. *Nat Rev Dis Primers*, *2*, 16018. https://doi.org/10.1038/nrdp.2016.18

London, W. T., & Mcglynn, K. A. (2006). Liver Cancer. In *Cancer Epidemiology and Prevention* (pp. 763–786). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195149616.003.0039

McGlynn, K. A., & London, W. T. (2011). The global epidemiology of hepatocellular carcinoma: present and future. *Clinics in Liver Disease*, *15*(2), 223–43, vii–x. https://doi.org/10.1016/j.cld.2011.03.006

Nahon, P., & Nault, J.-C. (2017). Constitutional and functional genetics of human alcohol-related hepatocellular carcinoma. *Liver International*, *0*, 1–11. https://doi.org/10.1111/liv.13419

Nakamura, T., Hamada, F., Ishidate, T., Anai, K., Kawahara, K., Toyoshima, K., & Akiyama, T. (1998). Axin, an inhibitor of the Wnt signalling pathway, interacts with beta-catenin, GSK-3beta and APC and reduces the beta-catenin level. *Genes to Cells*, *3*(6), 395–403. https://doi.org/10.1046/j.1365-2443.1998.00198.x

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Schroder, M. S., Culhane, A. C., Quackenbush, J., & Haibe-Kains, B. (2011). survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, *27*(22), 3206–3208. https://doi.org/10.1093/bioinformatics/btr511

Schulze, K., Imbeaud, S., Letouze, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., … Zucman-Rossi, J. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet*, *47*(5), 505–511. https://doi.org/10.1038/ng.3252

Shibata, T., & Aburatani, H. (2014). Exploration of liver cancer genomes. *Nat Rev Gastroenterol Hepatol*, *11*(6), 340–349. https://doi.org/10.1038/nrgastro.2014.6

Sidow, A., & Spies, N. (2015). Concepts in solid tumor evolution. *Trends Genet*, *31*(4), 208–214. https://doi.org/10.1016/j.tig.2015.02.001

Siegel, R. L., Miller, K. D., & Jemal, A. (2015). Cancer statistics, 2015. *CA Cancer J Clin*, *65*(1), 5–29. https://doi.org/10.3322/caac.21254

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, *66*(1), 7–30. https://doi.org/10.3322/caac.21332

Totoki, Y., Tatsuno, K., Covington, K. R., Ueda, H., Creighton, C. J., Kato, M., … Shibata, T. (2014). Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat Genet*, *46*(12), 1267–1273. https://doi.org/10.1038/ng.3126

Totoki, Y., Tatsuno, K., Yamamoto, S., Arai, Y., Hosoda, F., Ishikawa, S., … Shibata, T. (2011). High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet*, *43*(5), 464–469. https://doi.org/10.1038/ng.804

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J., & Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, *25*(18), 3201–3216. https://doi.org/10.1002/sim.2353

Vandin, F., Upfal, E., & Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Res*, *22*(2), 375–385. https://doi.org/10.1101/gr.120477.111

Vervier, K., & Michaelson, J. J. (2016). SLINGER: large-scale learning for predicting gene expression. *Scientific Reports*, *6*(1), 39360. https://doi.org/10.1038/srep39360

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr., L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Wang, X. (2016). Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics (Oxford, England)*, *32*(9), 1316–22. https://doi.org/10.1093/bioinformatics/btw002

Wei, R., De Vivo, I., Huang, S., Zhu, X., Risch, H., Moore, J. H., … Garmire, L. X. (2016). Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. *Oncotarget*, *7*(34), 55249–55263. https://doi.org/10.18632/oncotarget.10509

Wong, N., & Wang, X. (2015). miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Research*, *43*(Database issue), D146-52. https://doi.org/10.1093/nar/gku1104

Wu, J. N., & Roberts, C. W. M. (2013). ARID1A mutations in cancer: another epigenetic tumor suppressor? *Cancer Discovery*, *3*(1), 35–43. https://doi.org/10.1158/2159-8290.CD-12-0361

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhao, J., Chen, J., Lin, H., Jin, R., Liu, J., Liu, X., … Cai, X. (2016). The Clinicopathologic Significance of BAF250a (ARID1A) Expression in Hepatocellular Carcinoma. *Pathology & Oncology Research*, *22*(3), 453–459. https://doi.org/10.1007/s12253-015-0022-9

Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods*, *11*(6), 599–600. https://doi.org/10.1038/nmeth.2956

Figure Legends

Figure 1. Consensus driver genes in 6 HCC cohorts. (A) IntOGen pipeline to identify consensus driver genes. (B) Driver genes from individual cohorts. Genes with asterisks represent consensus drivers. (C) Final 11 genes with mean q-value <0.1 from MutSigCV module. (D) Same 11 genes with mean q-value <0.1 from OncodriveFM module. (E) 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. 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 5. Associations of gender and age with driver genes. Shown are subsets of consensus driver genes with significant associations with gender (Fisher's exact test with p-value<0.05) and age (Mann-

Whitney-Wilcoxon test with p-value<0.05). Age is divided to two groups by the median value (in parentheses) in each cohort.

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. 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 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.

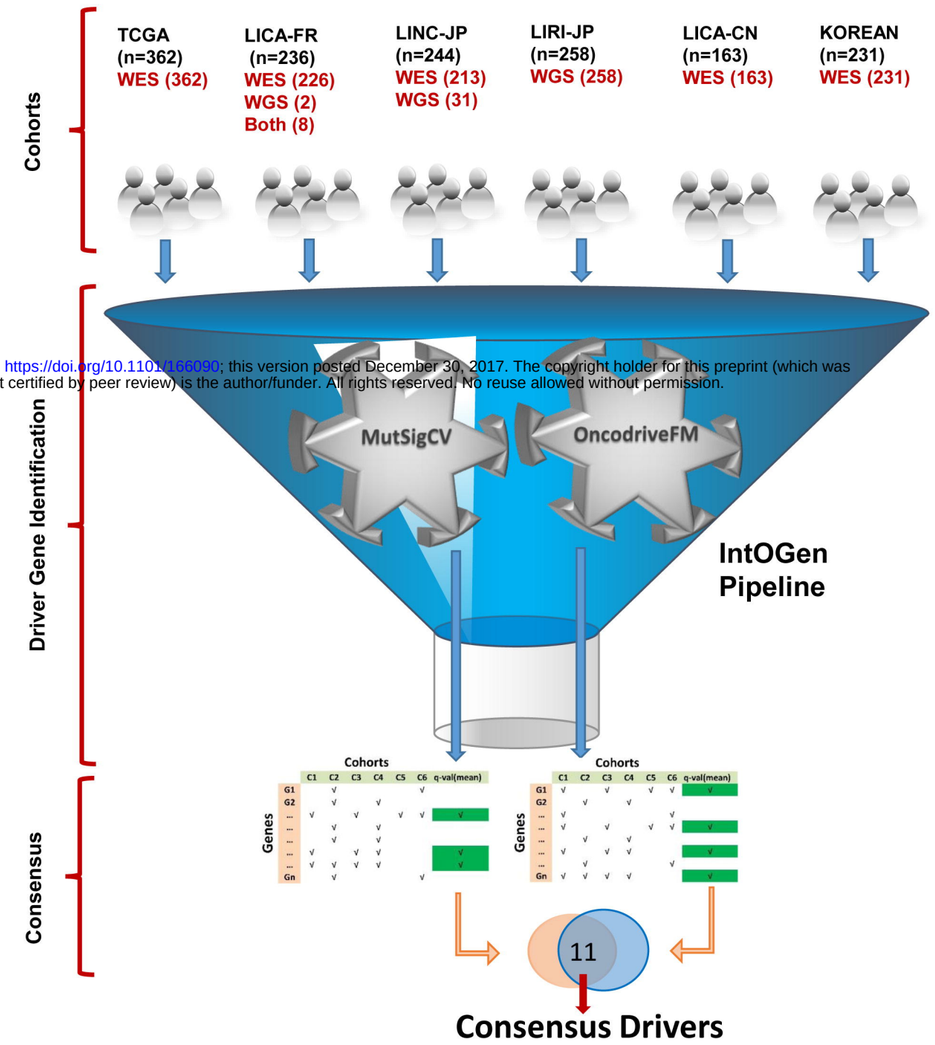
Supplementary Table

Table S1. Clinical summary of 6 HCC cohorts.

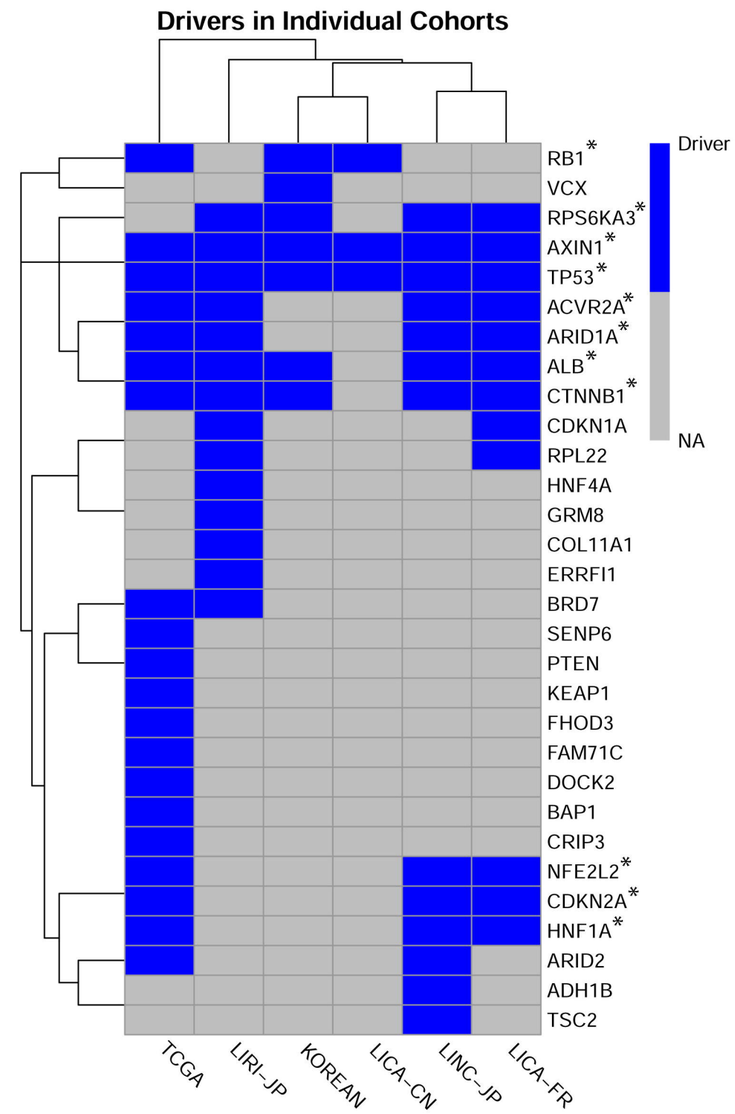
Supplementary File

File S1. Associations between miR expression and consensus driver gene mutation/CNV.

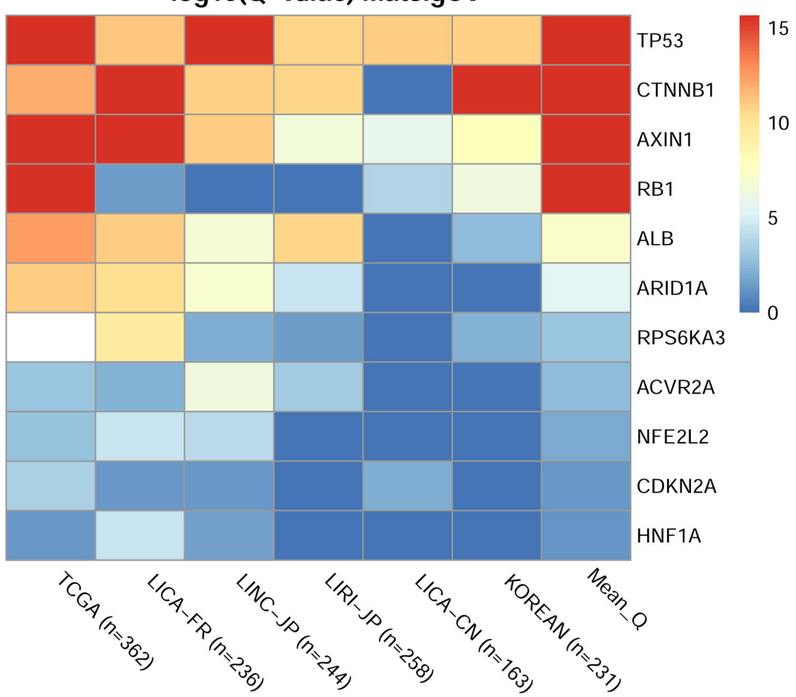
(A)



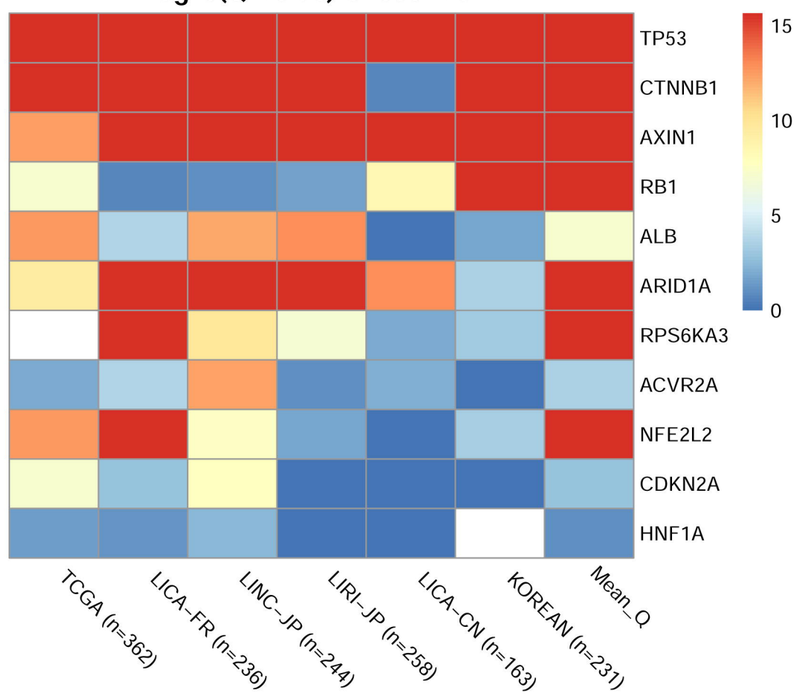
(B)



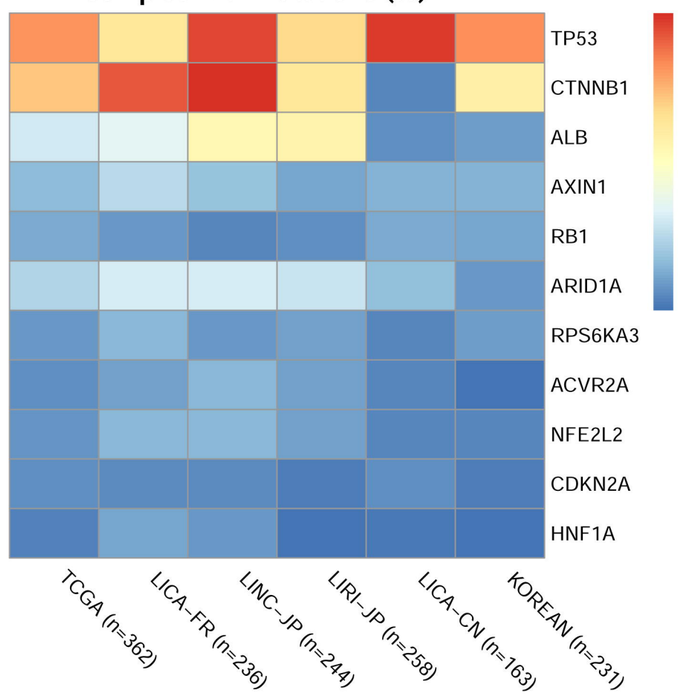
(C)



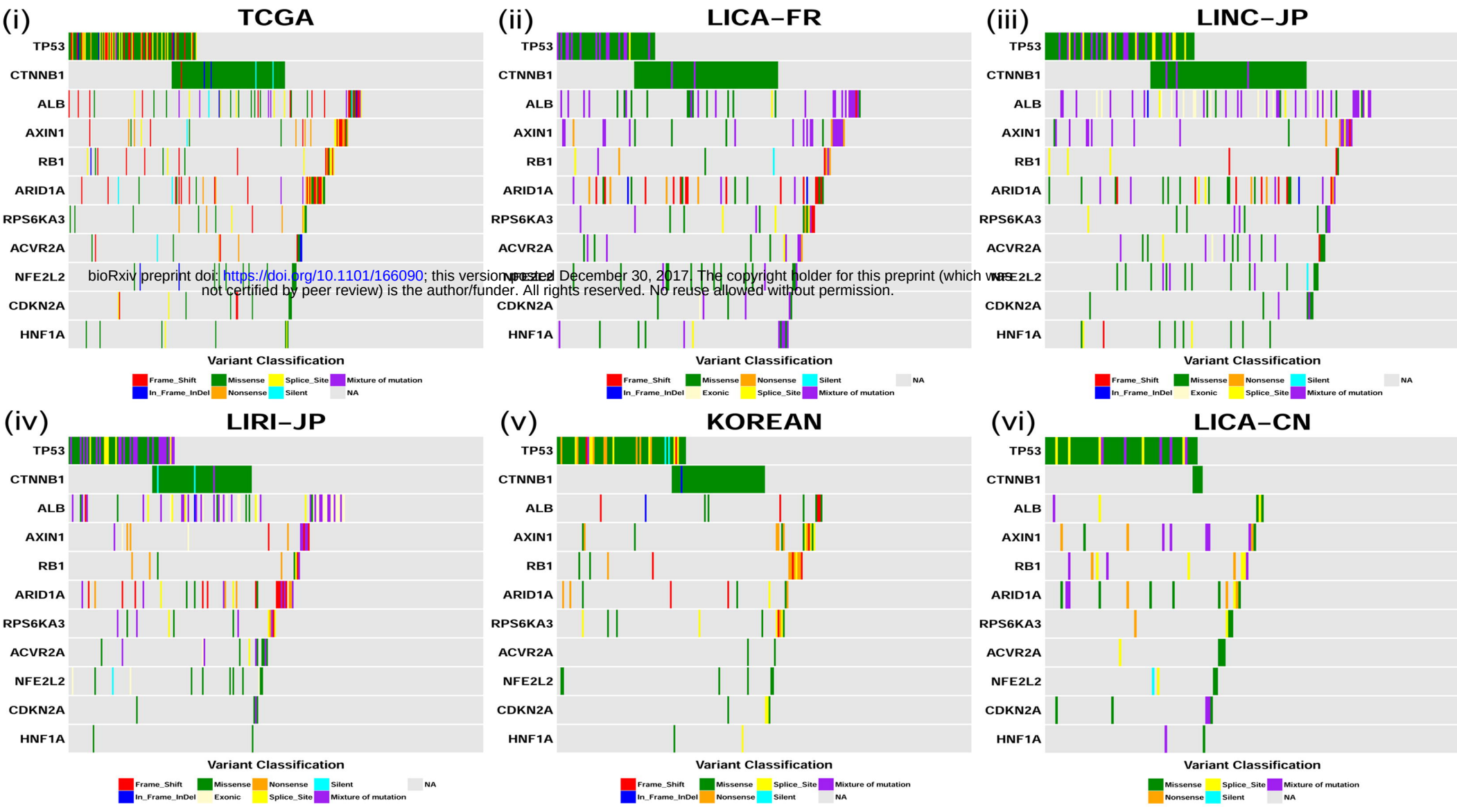
(D)



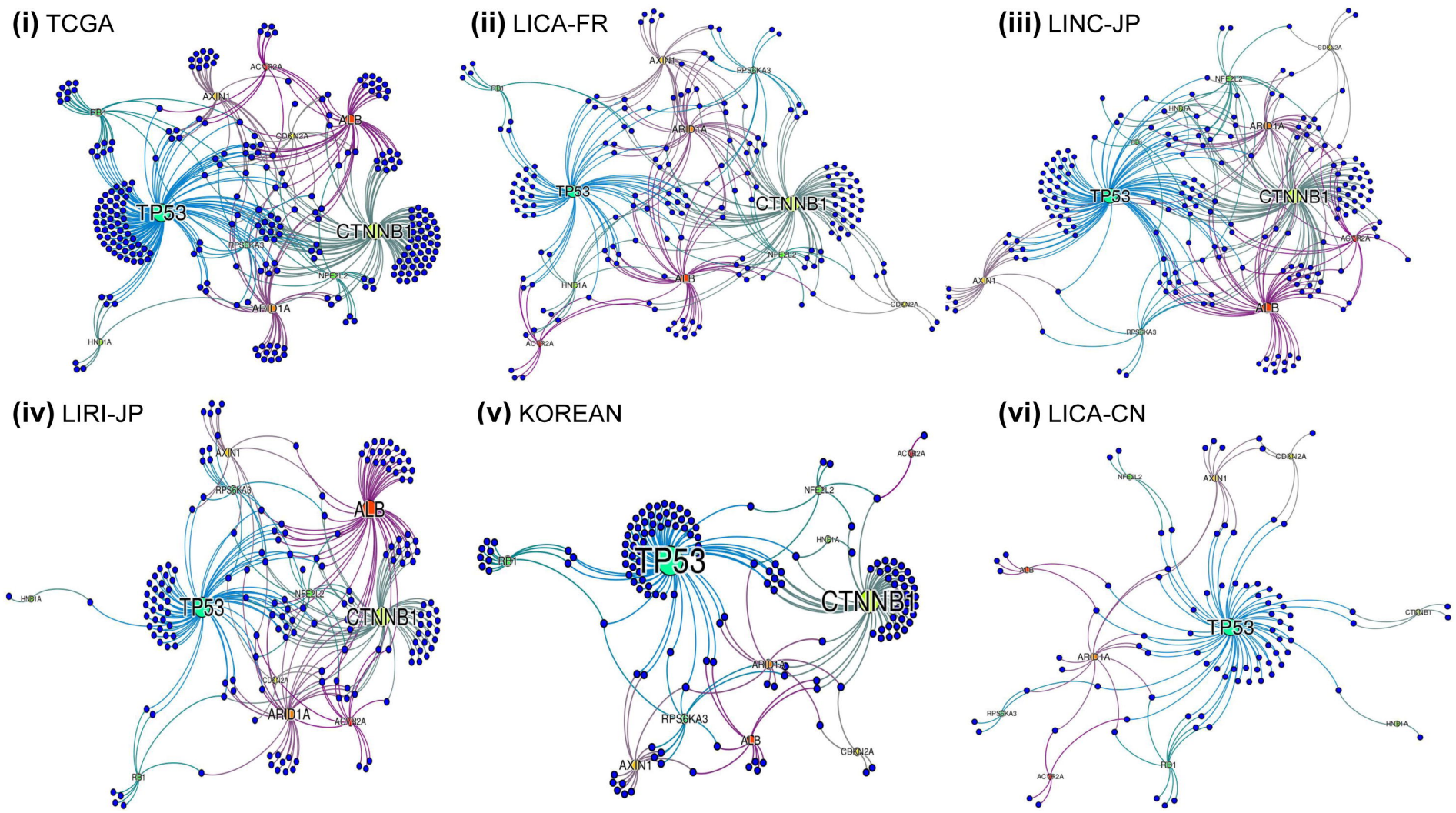
(E)



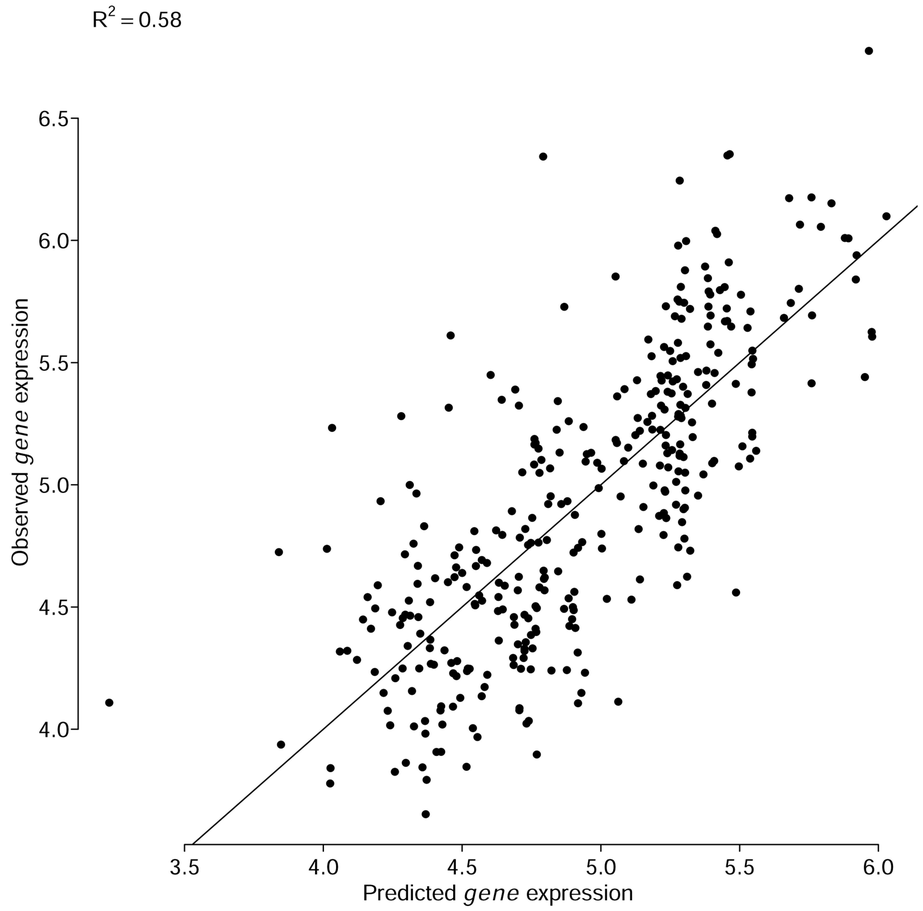
(A)



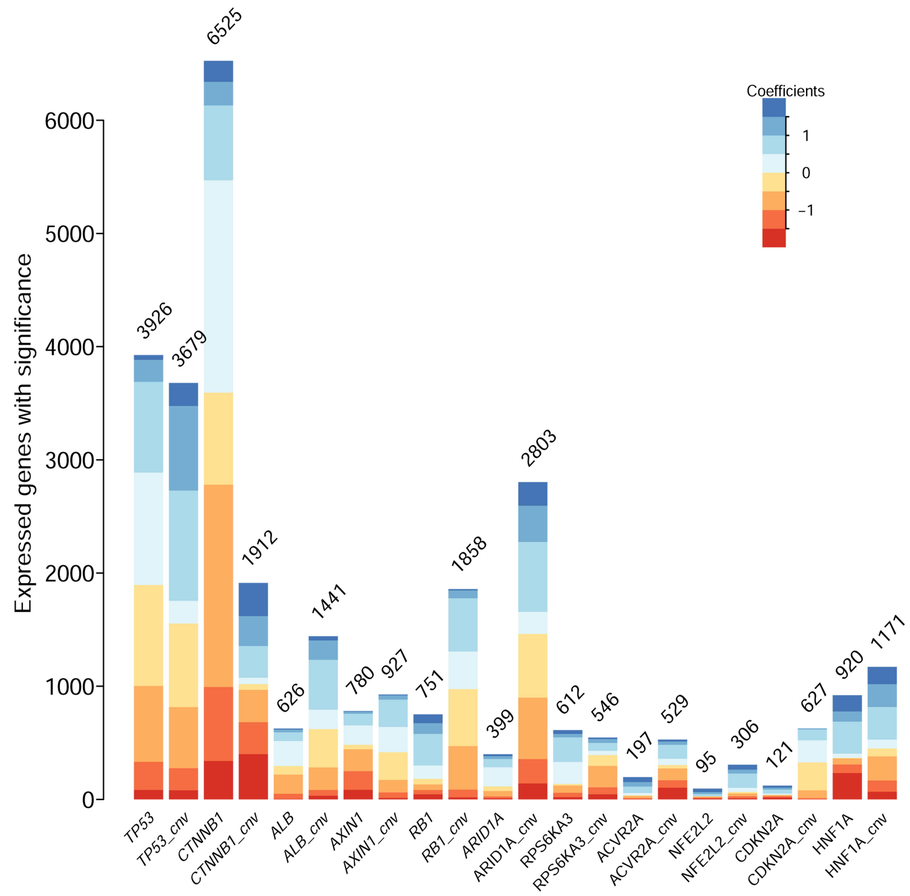
(B)



(A)

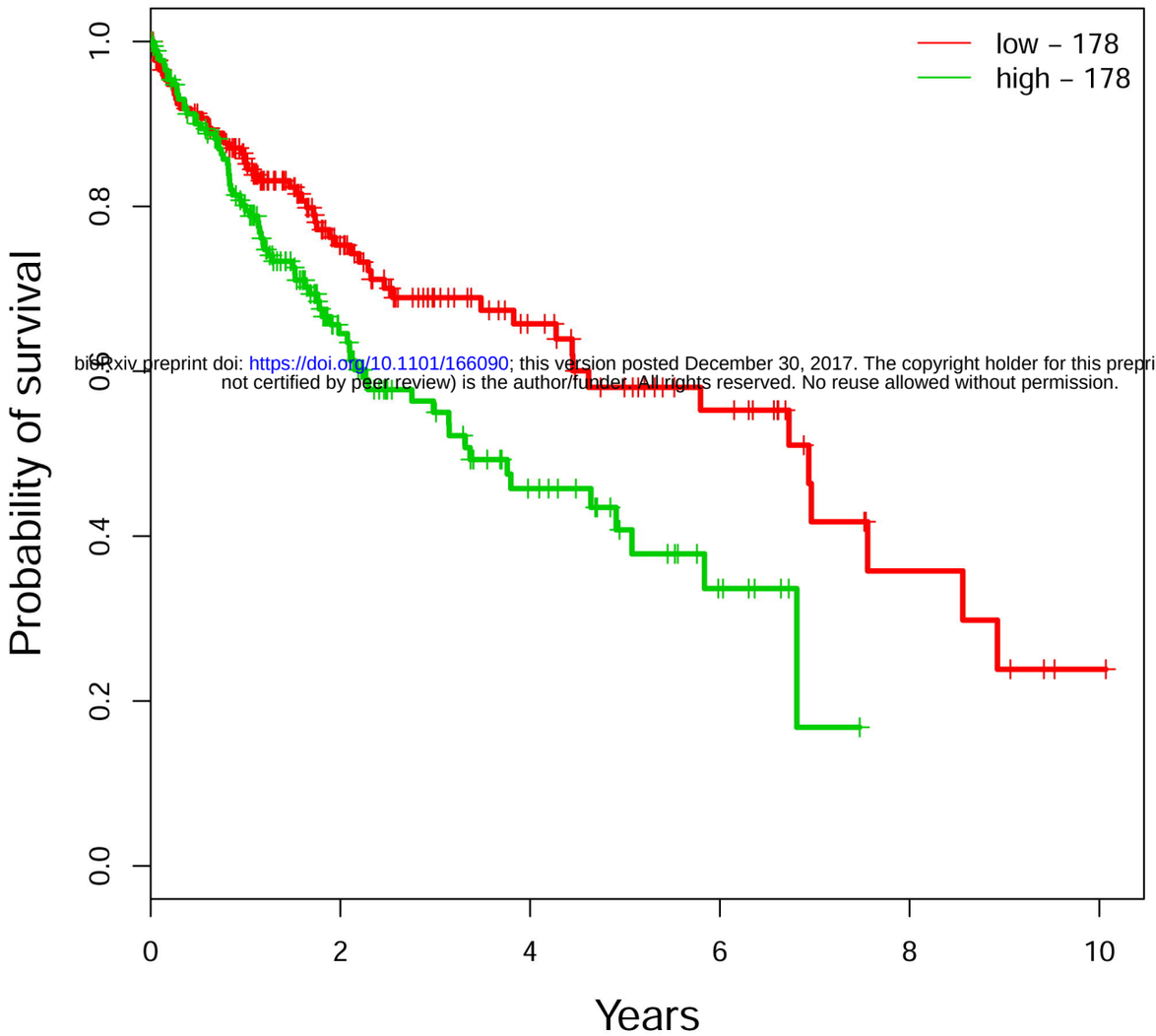


(B)

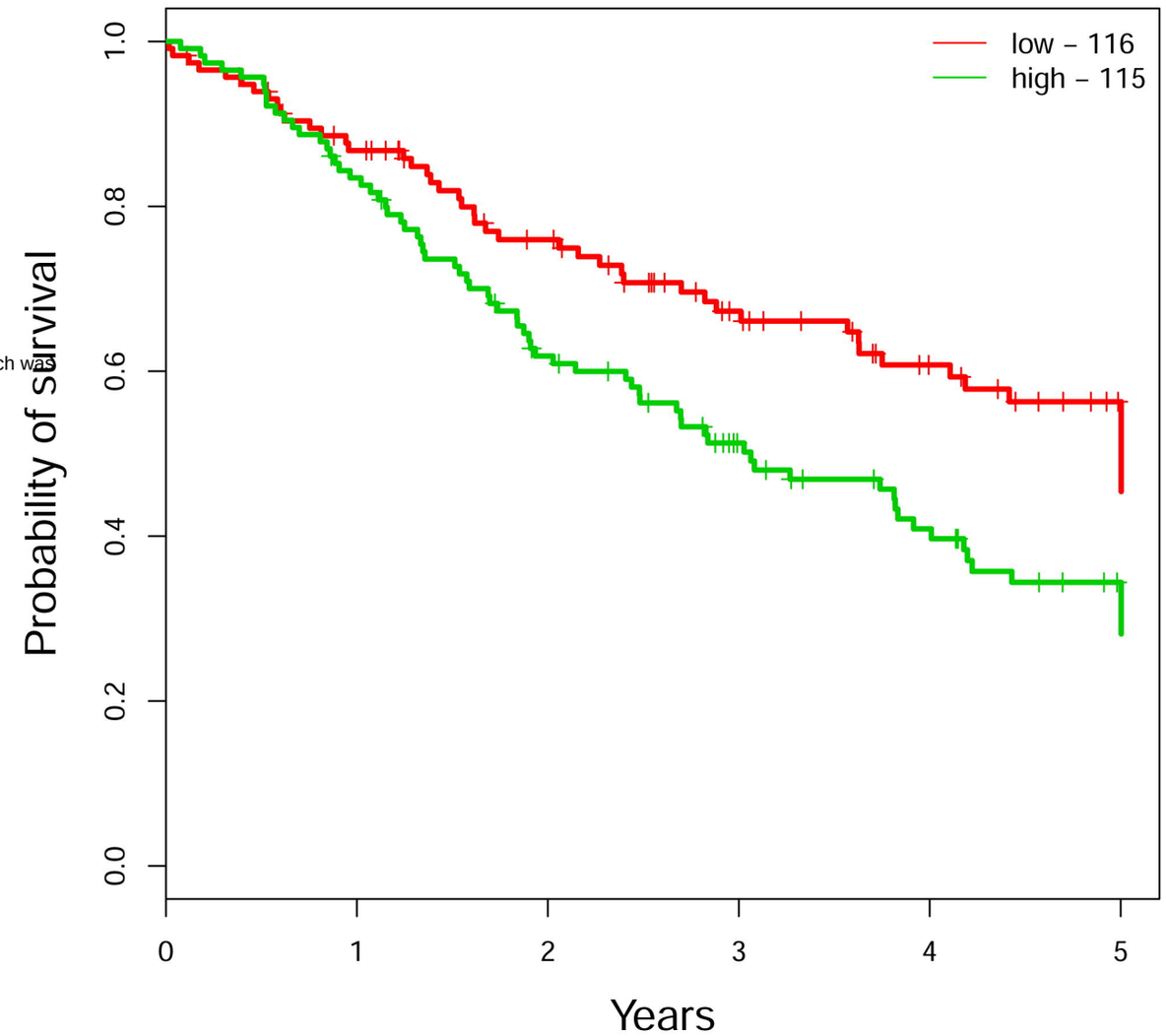


(C)

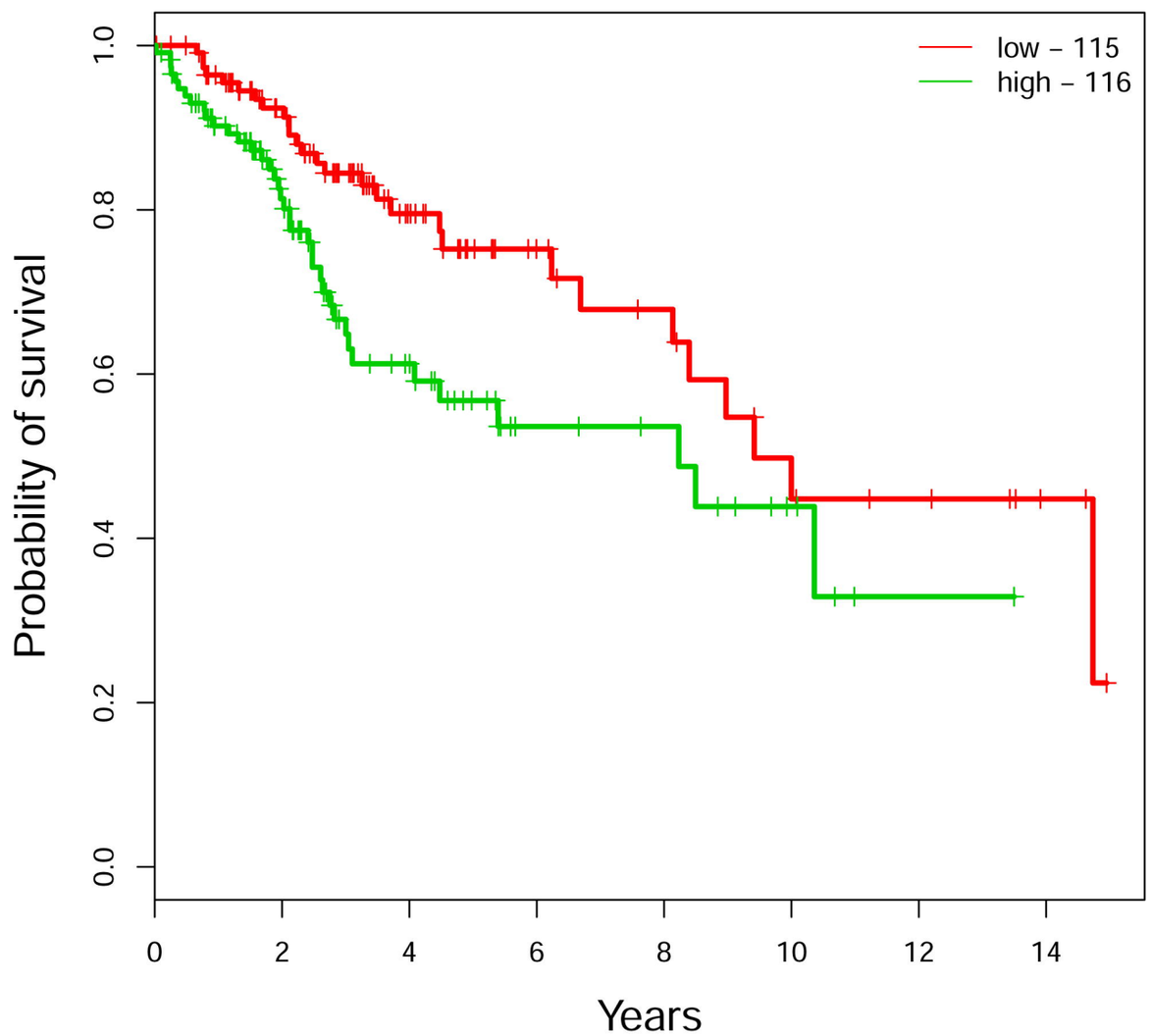
(A)



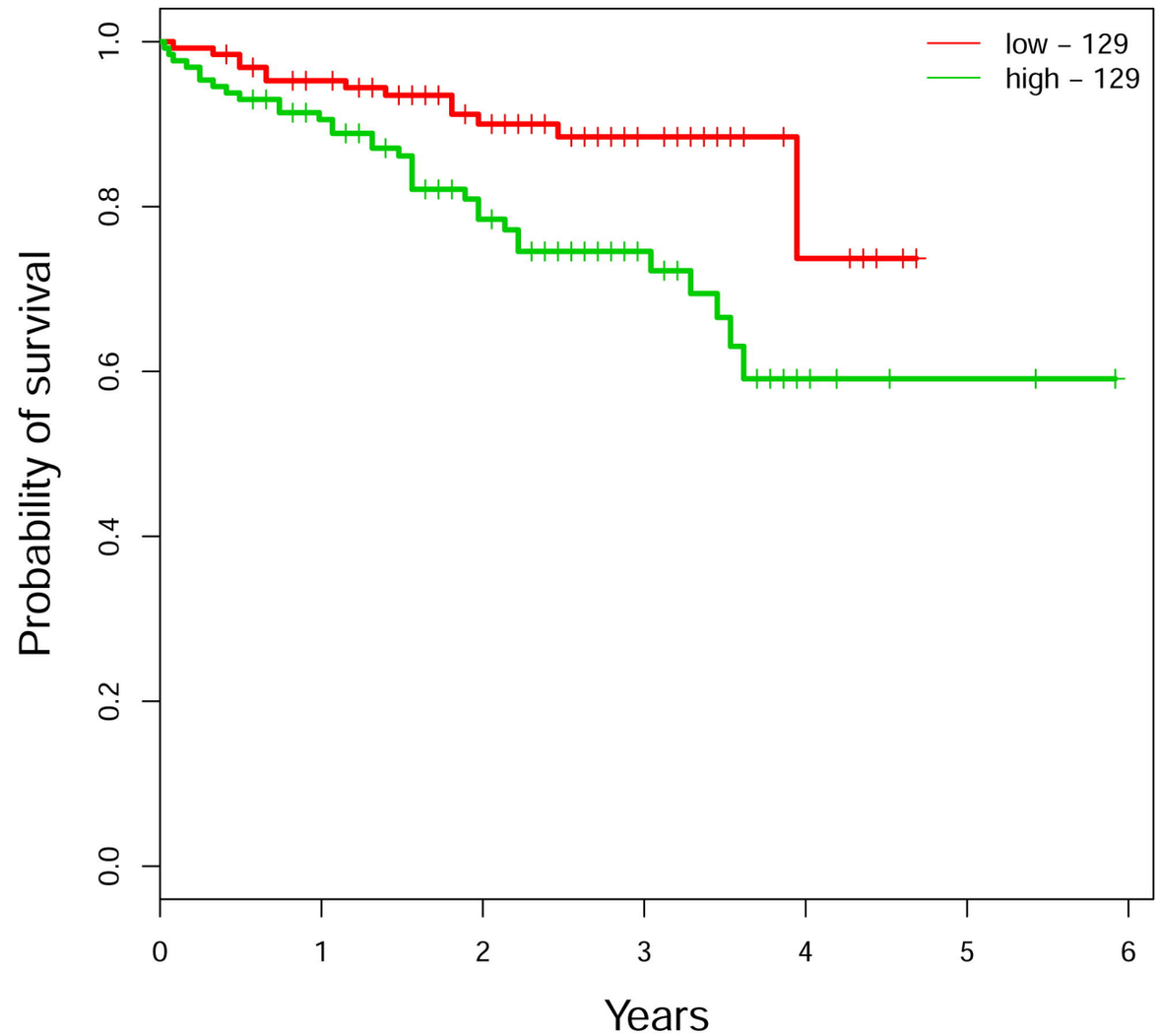
(B)

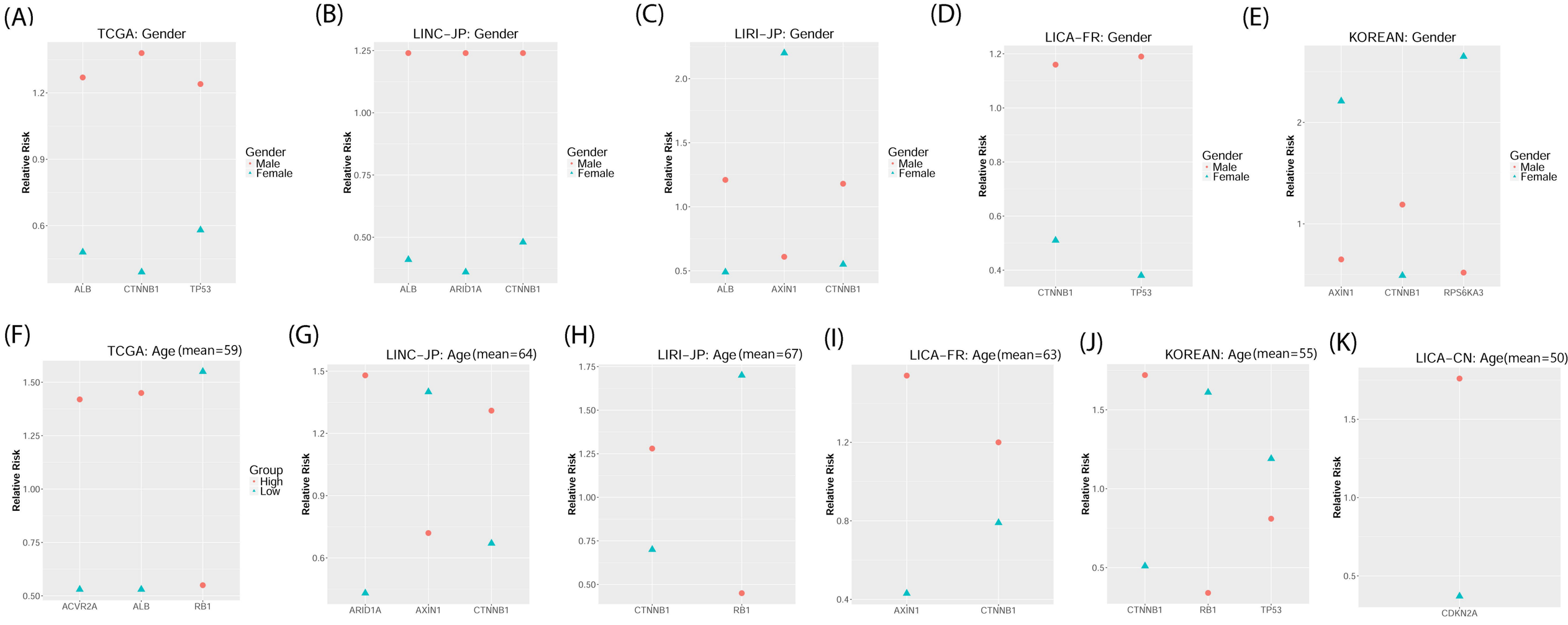


(C)



(D)




(A)
TCGA: Gender
Relative Risk
1.2
0.9
0.6
ALB
CTNNB1
TP53
Gender
Male
Female
(B)
LINC-JP: Gender
Relative Risk
1.25
1.00
0.75
0.50
ALB
ARID1A
CTNNB1
Gender
Male
Female
(C)
LIRI-JP: Gender
Relative Risk
2.0
1.5
1.0
0.5
ALB
AXIN1
CTNNB1
Gender
Male
Female
(D)
LICA-FR: Gender
Relative Risk
1.2
1.0
0.8
0.6
0.4
CTNNB1
TP53
Gender
Male
Female
(E)
KOREAN: Gender
Relative Risk
2
1
AXIN1
CTNNB1
RPS6KA3
Gender
Male
Female
(F)
TCGA: Age (mean=59)
Relative Risk
1.50
1.25
1.00
0.75
0.50
ACVR2A
ALB
RB1
Group
High
Low
(G)
LINC-JP: Age (mean=64)
Relative Risk
1.5
1.2
0.9
0.6
ARID1A
AXIN1
CTNNB1
(H)
LIRI-JP: Age (mean=67)
Relative Risk
1.75
1.50
1.25
1.00
0.75
0.50
CTNNB1
RB1
(I)
LICA-FR: Age (mean=63)
Relative Risk
1.2
0.8
0.4
AXIN1
CTNNB1
(J)
KOREAN: Age (mean=55)
Relative Risk
1.5
1.0
0.5
CTNNB1
RB1
TP53
(K)
LICA-CN: Age (mean=50)
Relative Risk
1.5
1.0
0.5
CDKN2A