Supplementary Information for

Functional Genomic Complexity Defines Intratumor Heterogeneity and Tumor Aggressiveness in Liver Cancer

So Mee Kwon[1,2], Anuradha Budhu[1], Hyun Goo Woo[1,3], Jittiporn Chaisaingmongkol[1,4,5], Hien Dang[1], Marshonna Forgues[1], Curtis C. Harris[1], Gao Zhang[6], Noam Auslander[7], Eytan Ruppin[7], Chulabhorn Mahidol[4], Mathuros Ruchirawat[4,5], Xin Wei Wang[1,*]

[1]Laboratory of Human Carcinogenesis and Liver Cancer Program, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, USA; [2]Department of Physiology, Ajou University School of Medicine, Suwon, 16499, Republic of Korea; [3]Department of Biomedical Science, Graduate School, Ajou University, Suwon, 16499, Republic of Korea; [4]Laboratory of Chemical Carcinogenesis, Chulabhorn Research Institute, Bangkok 10210, Thailand; [5]Center of Excellence on Environmental Health and Toxicology, Office of the Higher Education Commission, Ministry of Education, Bangkok 10400, Thailand; [6]Molecular and Cellular Oncogenesis Program and Melanoma Research Center, The Wistar Institute, Philadelphia, PA 19104, USA; [7]Cancer Data Science Lab, National Cancer Institute, National Institute of health, MD 20892, USA

*Corresponding author: xw3u@nih.gov

This PDF file includes:

Supplementary Materials and Methods

Data preprocessing

Transcriptome data and copy number data from TIGER-LC cohort were processed as follows. For Affymetrix Human Transcriptome Array 2.0 data, expression level of individual 914,585 exons was extracted and normalized based on the Robust Multi-array Average (RMA) method and sketch-quantile normalization method using the Transcriptome Analysis Console (TAC) Software 4.0. For transcripts with more than one exon probe sets, the mean expression was calculated and total 64,597 transcripts were used further analysis. For profiling of copy number for tumors and paired non-tumor tissues generated based on Affymetrix Genome-Wide Human SNP Nsp/Sty 6.0, we applied the ***crlmm*** R package into the raw CEL files to estimate copy number based on the CRLMM algorithm[1]. Briefly, the ***crlmm*** package adapts the robust multichip average (RMA) to genotyping platforms based on the SNP-RMA algorithm[2]. For probes for polymorphic loci, the raw intensities for each allele are quantile normalized[3] to a target reference distribution obtained from the HapMap phase 2 samples. The Affymetrix 6.0 platform contains 3 or 4 identical probes for each allele. The normalized intensities for a set of identical probes are summarized by the median. For nonpolymorphic loci, only one probe per loci is available and the intensities are quantile normalized without a subsequent summarization step. Additional details regarding the preprocessing of Affymetrix CEL files are described elsewhere[2]. Somatic copy number variations were inferred by CBS (Circular binary segmentation) algorithm[4]. The genomic locations of segmented regions were converted from hg19 to hg38 by applying the UCSC *liftOver* R package. The copy number value of segmented regions was merged or separated for corresponding transcriptome probes, resulting in the allocation of copy number value for each segment corresponding 64,597 transcripts. For the validation cohort, HCC cohort of 247 Chinese patients from LCI[5] and TCGA LIHC cohort with 377 HCC patients were used. Transcriptome data and aCGH data for LCI cohort were processed as described previously[6]. Copy number value for each segmented region was allocated into each corresponding gene probe of the transcriptome data located in the segmented region resulting in 10,127 features. The level 3 RNA-seq v2.0 data and Affymetrix SNP 6.0 data were downloaded from TCGA Research Network (http://cancergenome.nih.gov/ ; release 1.0). Gene-level annotated transcriptome data segmented data were used for further analysis. All processing was conducted using R packages of Bioconductor 3.5 (https://cran.r-project.org/doc/FAQ/R-FAQ.html).

Calculation of global correlation

The global correlation coefficients and global correlation p-value based on the total transcriptome probes and corresponding genomic segments were calculated. For this, SCNA value for genomic segments corresponding to the 64,597 transcript probes was assigned using the GenomicRanges R packages. Hereafter, CN denotes copy number value and EXP denotes mRNA expression value.

$$\begin{array}{ccc} \text{CNV} & & \text{EXP} \\ \begin{pmatrix} c_{11} & \cdots & c_{1m} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nm} \end{pmatrix} & \times & \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & \ddots & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix} \end{array}$$

Where c_{nm} represents the SCNA value of nth sample corresponding mth feature. Matched features are expressed below.

$$\begin{array}{cc} & \begin{array}{ccccc} F_1 & & F_k & & F_m \end{array} \\ \begin{array}{c} S_1 \\ \vdots \\ S_i \\ \vdots \\ S_n \end{array} & \begin{pmatrix} c_{11}e_{11} & \cdots & c_{1k}e_{1k} & \cdots & c_{1m}e_{1m} \\ \vdots & & \vdots & & \vdots \\ c_{i1}e_{i1} & & c_{ik}e_{ik} & & c_{im}e_{im} \\ \vdots & & \vdots & & \vdots \\ c_{n1}e_{n1} & \cdots & c_{nk}e_{nk} & \cdots & c_{nm}e_{nm} \end{pmatrix} \end{array}$$

Where F_k indicates kth feature and S_i indicates for ith sample. Global correlation of M number of from n number of the tumor (T) and non-tumor (NT) of HCC and iCCA sample was calculated. Permutated correlation coefficient and p-value were used to compare between T and NT. Significantly correlated features of transcriptome probes and corresponding segmented regions were selected (p-value < 0.05 & median absolute deviation (MAD) >20% of the overall distribution) for further analysis.

Calculation of SCNA frequency and Inference of Arm-level SCNA

To define amplified or deleted region, we applied a threshold, 0.2 or -0.2, respectively, to the log2 transformed copy number value for individual 64,597 features. The fraction of patients who showed amplification or deletion was calculated for each feature. We calculated the frequency of arm-level amplifications and deletions based on the GISTIC (Genomic Identification of Significant Targets in Cancer) algorithm[7] in the GISTIC_2.0 module of GenePattern[8]. The segments with $\log_2$ ratio > 0.2 and < -0.2 were defined as chromosomal amplifications and deletions following the default value of the algorithm, respectively.

LOH and allelic specific copy number

LOH (Loss of Heterozygosity) for each sample was inferred using Genotyping Console 4.0 and the output CHP file was used as an input file to calculate allele-specific copy numbers using the Partek Genomics Suite 7.5. By merging the copy number of the segmented region defined by an algorithm in Partek and LOH data, the allele-specific copy number was estimated. For further analysis, we calculated the proportion of sample with allele-specific copy number change in each segmented region.

The biological relevance of PCC or tFA associated genes

To examine if PCC or tFA was associated with the biological process, we performed a correlation analysis between PCC and all the transcriptome features. Positively or negatively associated genes were selected based on the correlation estimate and p-value (above top 5% or below the bottom 5% of correlation estimate and p-value < 0.05). Gene ontology enrichment analysis was performed using R package gProfileR based on GO: BP.

Gene Set Enrichment Analysis (GSEA) and single-sample GSEA (ssGSEA)

GSEA was implemented in GenePattern[8] based on the C5 GO gene set of biological process, C2 curated gene sets of KEGG pathway, and C6 Oncogenic gene sets in Molecular Signatures Database (MSigDB database v5.2). Expression data of individual samples were transformed into the gene set enrichment score P-value from the Kolmogorov-Smirnov (ks) test was used a single sample enrichment score.

Measurement of chromosomal instability

To infer chromosomal instability, we devised two indicators, one is based on the SCNA proportion of individual sample level and the other is based on the summation of the length of segments with SCNA. For comparison of chromosomal instability of individual sample level, we defined the segments with log2 ratio > 0.2 and < -0.2 as chromosomal amplifications and deletions by applying noise cutoff of 0.2, respectively and proportion of amplified (CIN_{ampl}) or deleted features (CIN_{del}) over total features were calculated. CIN_{ampl} and CIN_{del} for the individual patient were calculated based on the copy number value for 64,597 features. The summation of CIN_{ampl} and CIN_{del} was used as CIN score for further analysis.

$$CIN_{ampl} = \frac{\sum No.of\ amplified\ segment}{\sum No.of\ \ segment} * 100$$

$$CIN_{del=} \frac{\sum No.of\ deleted\ segment}{\sum No.of\ \ segment} * 100$$

$$CIN=CIN_{ampl} + CIN_{del}$$

As another aspect of the chromosome instability indicator, we also calculated the total length for total amplified (GIN_{gain}) or deleted regions (GIN_{loss}) and used the summation of the GIN_{gain} and GIN_{loss} for total SCNA length as a genomic instability (GIN) score as follows.

$$GIN_{gain} = \sum \text{length of amplified segments}$$

$$GIN_{loss} = \sum \text{length of deleted segments}$$

$$GIN= GIN_{gain\,+}\,GIN_{loss}$$

Total functional aneuploidy (tFA)

We calculated total functional aneuploidy (tFA) in each sample based on coordinated aberrations in the expression of genes localized to each chromosomal region using the adapted computational method from the previously published paper[9]. Briefly, it is a computational method to characterize aneuploidy in tumor samples based on coordinated aberrations in the expression of genes localized to each chromosomal region. For a given data set, all of the normalized gene expression measurements present on the microarray and mapping to a given chromosomal cytoband region were grouped into a set designated "B"(short for band). The rest of the genes, localized elsewhere in the genome, were grouped into a set "G" (short for genome). The functional aneuploidy measure for the given cytoband is the value of student's t statistic comparing sets B and G. Sum of all functional aneuploidy magnitudes (the absolute t statistics) in a given tumor sample. Therefore, the tFA is a total summarized level of chromosomal aberration in a given tumor in a univariate measure.

Differentially Expressed Genes (DEG) and Gene Ontology analysis

By comparing the expression between HFGC and LFGC of each tumor type, we selected differentially expressed genes in each subtype based on the fold change and permutation p-value from the permutation t-test with 1,000 resamplings (FC >0.5 or -0.5 & perm p-value <0.005). Gene ontology enrichment analysis was performed based on the DAVID 6.7[10].

Immune score

An estimation of the relative fractions of immune/inflammatory cell subsets from tissue expression profiles of Thai HCC, iCCA or TCGA HCC was conducted using CIBERSORT[11]. The gene expression data was converted by quantile normalization of the log2 scaled expression matrix and relative fractions of leukocytes were quantified according to the website (https://cibersort.stanford.edu/index.php) with implemented analyses using the built-in LM22 signature matrix (LM22). The immune score of individual tumor or non-tumor tissue was calculated as a summation of the 22 types of tumor-infiltrating lymphocytes (TILs) fraction based on CIBERSORT output. Since the output value was ranged from 0 to 1, for calculation convenience, we transformed the output value by multiplying by 100 and added one before the log2 transformation. The summation of transformed value for each TILs was used as the estimate of the immune score. Considering the difference of clinical outcome between LFGC and HFGC, TILs enriched in LFGC than HFGC were defined as favorable or adverse, vice versa. The summation of adverse or favorable TILs fractions was used as immune score of adverse or favorable TILs.

Mutation Map

MutationMapper (version 1.0) in the cBioPortal (http://www.cbioportal.org/tools.jsp) was used to plot the lollipop mutation diagram view with genomic coordinates to annotate *TP53* variants[12,13].

Validation with melanoma dataset

We used transcriptome data from skin cutaneous melanoma datasets derived from TCGA_SKCM[14] study (n=472) and metastatic melanoma from Hugo[15] study (n=28) to validate the association between FGC and immunotherapy with immune checkpoint blockade (ICB). We calculated tFAs in the individual sample and used them as a surrogate of PCC on the assumption that tFA were strongly associated with PCC based on our findings on liver cancer. Among TCGA_SKCM, 13 samples, which were pretreated with anti-CTLA-4 therapy, were included or excluded to perform KM survival analysis to examine whether the tFA level predicts responsiveness to ICB treatment. To compare high and low group, patients were stratified based on the tFA level into high (above 3rd quartile) and low group (below 1st quartile) in the TCGA_SKCM. Among the patients with anti-CTLA-4 pretreatment, tFA levels were compared between responders and non-responders based on the Welch's two-sample t-test. As another independent cohort, metastatic melanoma samples from Hugo study[15], where 28 patients were pre-treated with anti-PD-1 therapy, were used to validate the association between tFA and ICB responsiveness. To perform KM survival analysis, we divided patients into high and low groups based on the median level of tFA. We classified patients into "Responder" and "Non-Responder" as followed; "Responder" indicates those who marked as "Complete Response" or "Partial Response", while "Non-Responder" indicates those who marked as "Clinical Progressive Disease" or "Stable Disease" according to the response column of the clinical data.

Statistical Analyses

Kaplan-Meier (KM) Survival Analysis was performed based on the survival R package and p-value from the log-rank test based on the Cox Proportional-Hazards Regression model was used to compare overall survival. The permutation t-test was calculated based on the perm R package by 1,000 resamplings. The correlation coefficient and p-value were calculated based on the Pearson's product-moment correlation. After filtering based on the global correlation p-value (p-value<0.05) and MAD of copy number value (MAD > the value of 20% of MAD percentile), correlation coefficient was calculated in the individual subject using the corresponding correlated segment and transcriptome sets. All statistical tests were performed using R.

To perform permutation student's t-test, we applied R function, ***perm.ttest***, as follows.

```
perm.ttest=function(eset, g.st,  level=NULL, t.test=F, permp=T, permp.exact=NULL,
ordered=T, mc.cores=1,...){
   if(inherits(eset, "ExpressionSet"))   expr=exprs(eset) else    expr=eset
   if(ncol(expr)!=length(g.st))  cat("class labels has a different length")
  if(!inherits(g.st, "factor")) g.st=factor(g.st)
  if(!is.null(level)) g.st=factor(g.st, level=level)
  res=NULL
  if(t.test){
    message("Calculating T test p-values")
    if(mc.cores>1){
      if(Sys.info()[['sysname']]=="Windows") {
        res=mclapply(1:nrow(expr), function(a) try(t.test(as.numeric(expr[a,])~g.st),
silent=T),mc.cores=mc.cores,expr=expr,g.st=g.st, packageToLoad=c("stat","perm"))
      }else{
        res=mclapply(1:nrow(expr), function(a) try(t.test(as.numeric(expr[a,])~g.st),
silent=T),mc.cores=mc.cores,...=...)
      }
    }
    if(mc.cores==1) res=lapply(1:nrow(expr), function(a)
try(t.test(as.numeric(expr[a,])~g.st), silent=T))
    tval=as.numeric(sapply(res, function(a) try(a$stat, silent=T)))
    test.p=as.numeric(sapply(res, function(a) try(a$p.val, silent=T)))
    res=data.frame(t.stat=(tval), ttest.p=test.p)
    rownames(res)=rownames(expr)
  }
  if(permp)  {
    message("Calculating permuted T test p-values")
    if(mc.cores>1){
      if(Sys.info()[['sysname']]=="Windows") {
        res$perm.p=as.numeric(mclapply(1:nrow(expr), function(a)
try(permTS(as.numeric(expr[a,]) ~ g.st)$p.value,
silent=T),mc.cores=mc.cores,cluster.export=F, expr=expr,g.st=g.st,
packageToLoad=c("stat","perm")))
      }else{
        res$perm.p=as.numeric(parallel::mclapply(1:nrow(expr), function(a)
try(permTS(as.numeric(expr[a,]) ~ g.st)$p.value, silent=T),mc.cores=mc.cores))
      }
    }
    if(mc.cores==1) res$perm.p=as.numeric(lapply(1:nrow(expr), function(a)
try(permTS(as.numeric(expr[a,]) ~ g.st)$p.value, silent=T)))
    res$FDR = p.adjust(res$perm.p, "BH")
  }
  class.mean=sapply(levels(g.st), function(a) rowMeans(expr[,which(g.st==a)],
na.rm=T))
  colnames(class.mean) = paste(colnames(class.mean), "(mean)")
  fc=as.matrix(class.mean[,1]-class.mean[,2])
  res=cbind(as.data.frame(res), class.mean, fc)
  if(ordered)  res=res[order(-fc),]
  return(res)
}
```

Arguments

# eset : expression set

# g.st= group

# t.test=if set to F, permutation test will be performed

# permp= if set to T, permutation test will be performed

# permp.exact=NULL

# ordered=if set to T, features will be ordered with the decreasing order of fold difference

# mc.cores=the number of multi-core

To perform the permutation correlation test, we applied R function, ***cor.perm***, as follows.

```
cor.perm = function (x, y, nperm = nperm){
 cor.r = cor (x = x, y = y)
 cor.p = cor.test (x = x, y = y)$p.value
 perm.r = sapply (1:nperm, FUN = function (i) cor (x = x, y = sample (y)))
 perm.r = c(perm.r, cor.r)

## one-tailed. probability
  #perm.p = sum (perm.r>= cor.r)/(nperm + 1)
  ## two-tailed. probability
 perm.p = sum (abs(perm.r)>= abs(cor.r))/(nperm + 1)
 return (list(perm.p =perm.p))
}
```

Supplementary Figures

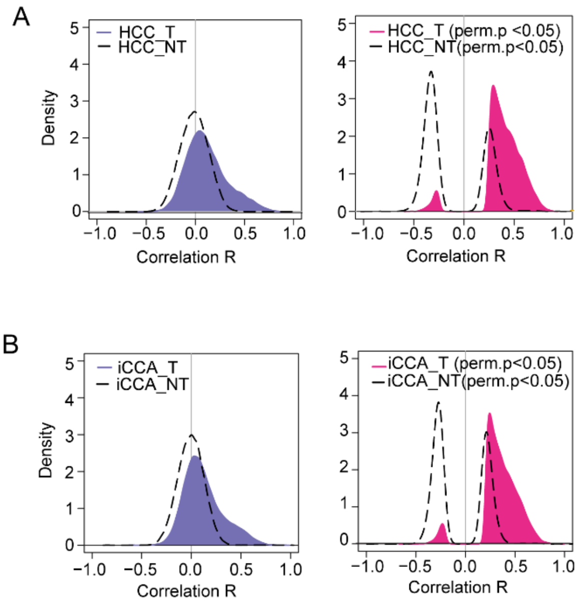


Figure S1. Distribution of global correlation coefficient in PLC (A-B) The density histogram shows the distribution of global correlation coefficient based on the permutated Pearson's correlation of DNA copy number (CN) and mRNA expression (EXP) from Tumor tissues and corresponding non-tumor tissue (A: HCC; n=64, HCC_NT; n=59, B: iCCA ; n=90, iCCA_NT; n=90). The distribution of correlation R is shown before (left panel) and after applying cut-off based on permutation p-value (perm.p <0.05) (right panel).

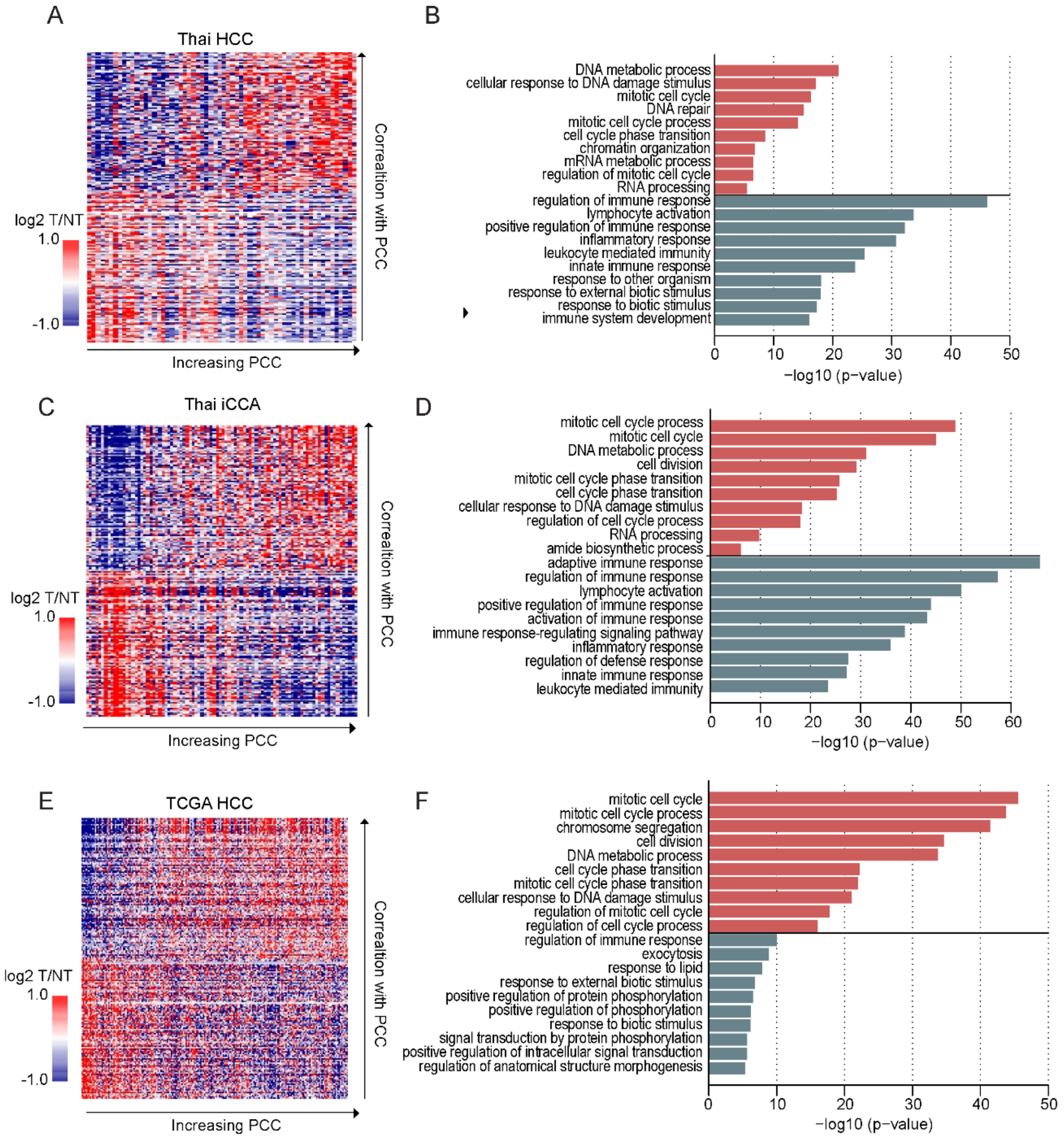


Figure S2. Gene Ontology (GO) of PCC associated genes (A, C, and E) Positively or negatively PCC associated genes were selected based on the correlation coefficient and p-value (more than 95% or less than 5% of estimate and p-value < 0.01). Heatmap shows the expression level of selected genes in Thai HCC, iCCA, and TCGA HCC cohorts (A, C, and E, respectively). Samples are represented in columns according to the PCC increasing order. Selected genes were represented in the row according to the decreasing of correlation coefficient with PCC. (B, D, and F) GO Enrichment Analysis of selected genes in Thai HCC, iCCA, and TCGA HCC cohorts were performed (B, D, and F,

respectively). Top10 ranked process based on the precision rank was shown. Orange and green color indicates positively and negatively correlated gene sets, respectively.

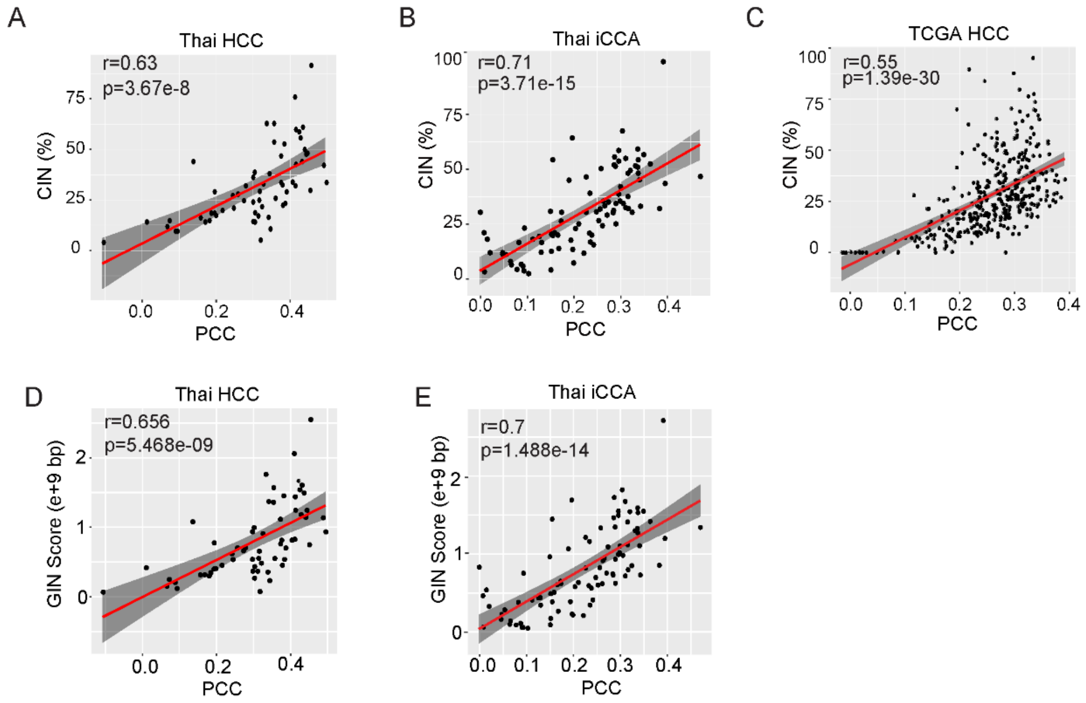


Figure S3. Association of PCC with CIN and GIN (A-C) PCC shows strong association with CIN in Thai HCC, Thai iCCA and TCGA HCC, respectively. (D-E) Genomic instability (GIN) length regarding the copy number gain or copy number loss (Methods) was calculated in the individual sample and the summation of the total SCNA length was calculated as GIN score. PCC shows strong association with GIN score of individual Thai HCC (D) and iCCA (E) samples

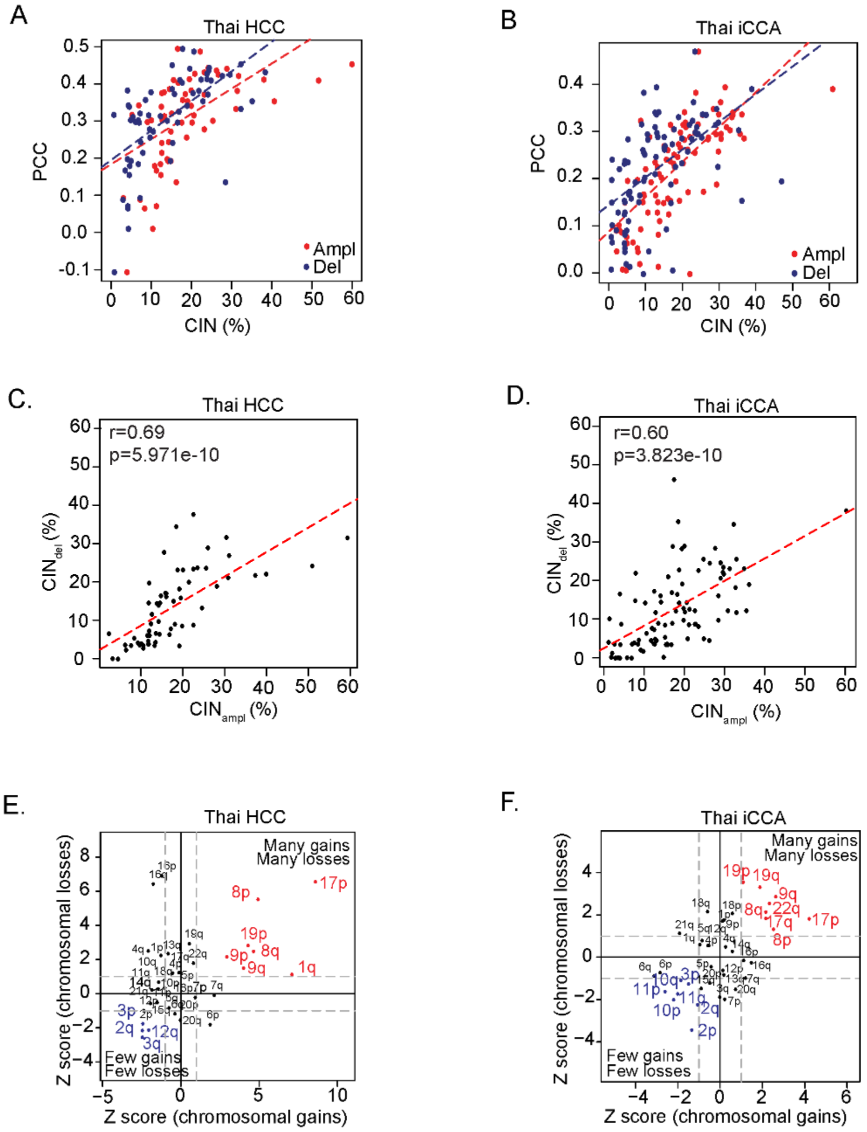


Figure S4. Association of amplified or deleted CIN (CIN_{ampl} or CIN_{del}) with PCC (A-B) Strong associations of CIN_{ampl} or CIN_{del} with PCC in Thai HCC (A) and Thai iCCA (B) are shown. Red or blue dots indicate CIN_{ampl} or CIN_{del}, respectively. Red or blue dots indicate CIN_{ampl} or CIN_{del}, respectively. (C-D) A strong linear association between CIN_{ampl} and CIN_{del} was shown in Thai HCC (C) and Thai iCCA (D). Coefficient estimates and p-value based on Pearson's correlation were depicted. (E-F) The frequency of recurrent arm-level SCNA of Thai HCC (E) and Thai iCCA (F) are shown. Chromosomal arms are shown with respect to the frequency of arm-level gain (x-axis) and loss (y-axis), respectively. As a frequency measure, Z score from GISTIC output was used. Vertical dotted blue lines indicate Z score of the arm-level gain frequency is 1 and

horizontal dotted blue lines indicate Z score of the arm-level loss frequency is 1. The arms with many gains and many losses or with few gains or few losses were highlighted in red or blue colors, respectively.

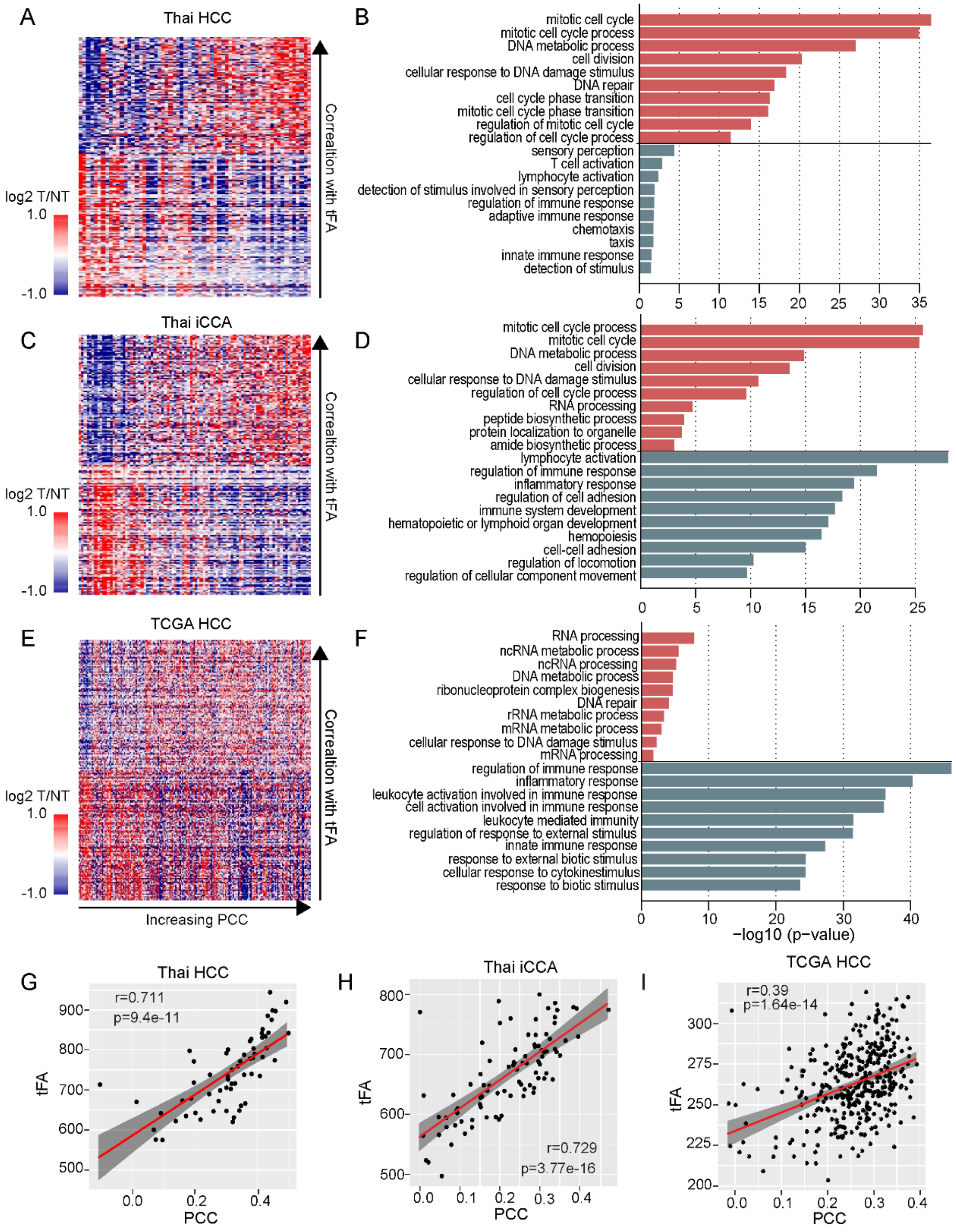


Figure S5. Gene Ontology (GO) of tFA associated genes (A-F) Functional relevance of PCC with tFA were examined among Thai HCC, Thai iCCA, and TCGA HCC, respectively. Positively or negatively tFA associated genes were selected based on the correlation coefficient and p-value (more than 95% or less than 5% of estimate and p-value <0.01). Heatmap shows the expression level of selected genes in Thai HCC, iCCA,

and TCGA HCC cohorts. Samples were represented in columns according to the FGC increasing order and selected genes were represented in the row according to the decreasing of correlation coefficient with tFA (A, C, and D, respectively). GO Enrichment Analysis of selected genes in Thai HCC, iCCA, and TCGA HCC cohorts were performed (B, D, and F, respectively). Top10 ranked process based on the precision rank was shown. Orange and green color indicates positively and negatively correlated gene sets, respectively. (G-I) PCC shows a strong association with tFA in Thai HCC, Thai iCCA and TCGA HCC, respectively. Coefficient estimates and p-value based on Pearson's correlation were depicted.

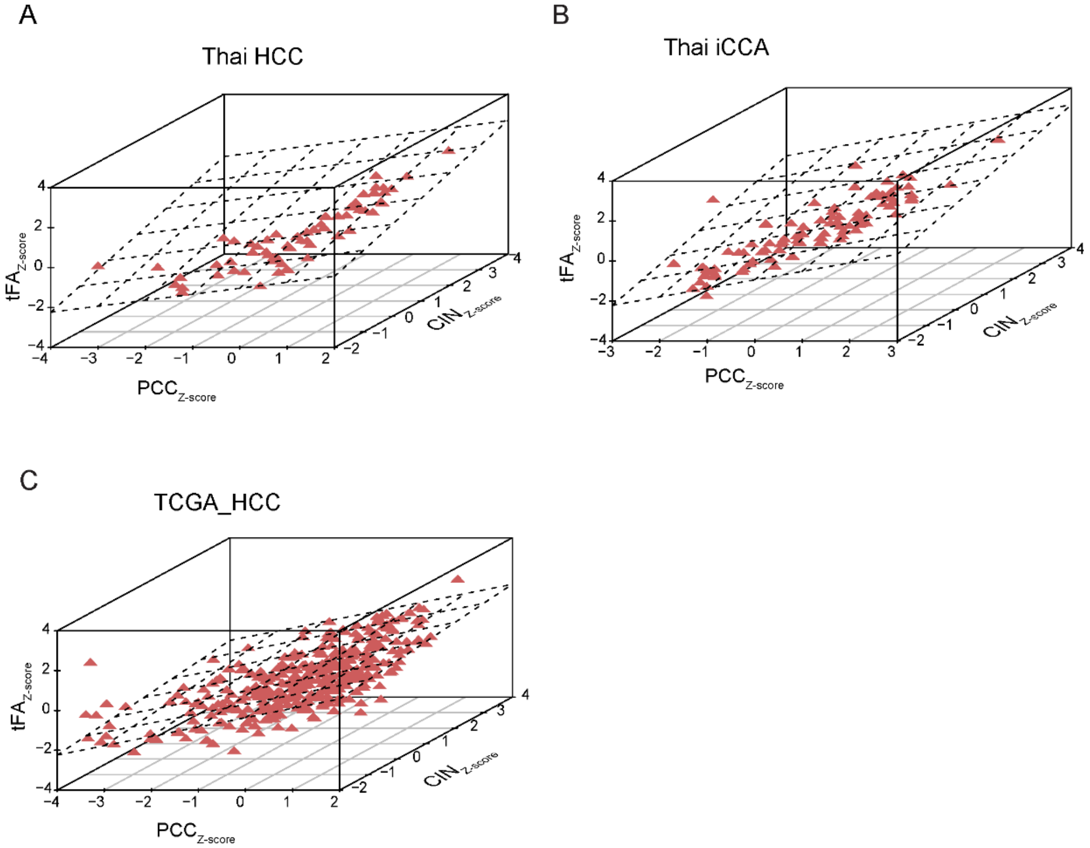


Figure S6. The collective association among PCC, CIN, and tFA (A-C) Collective association among the PCC (x-axis), CIN (y-axis), and tFA (z-axis) are shown in Thai HCC, iCCA, and TCGA HCC, respectively.

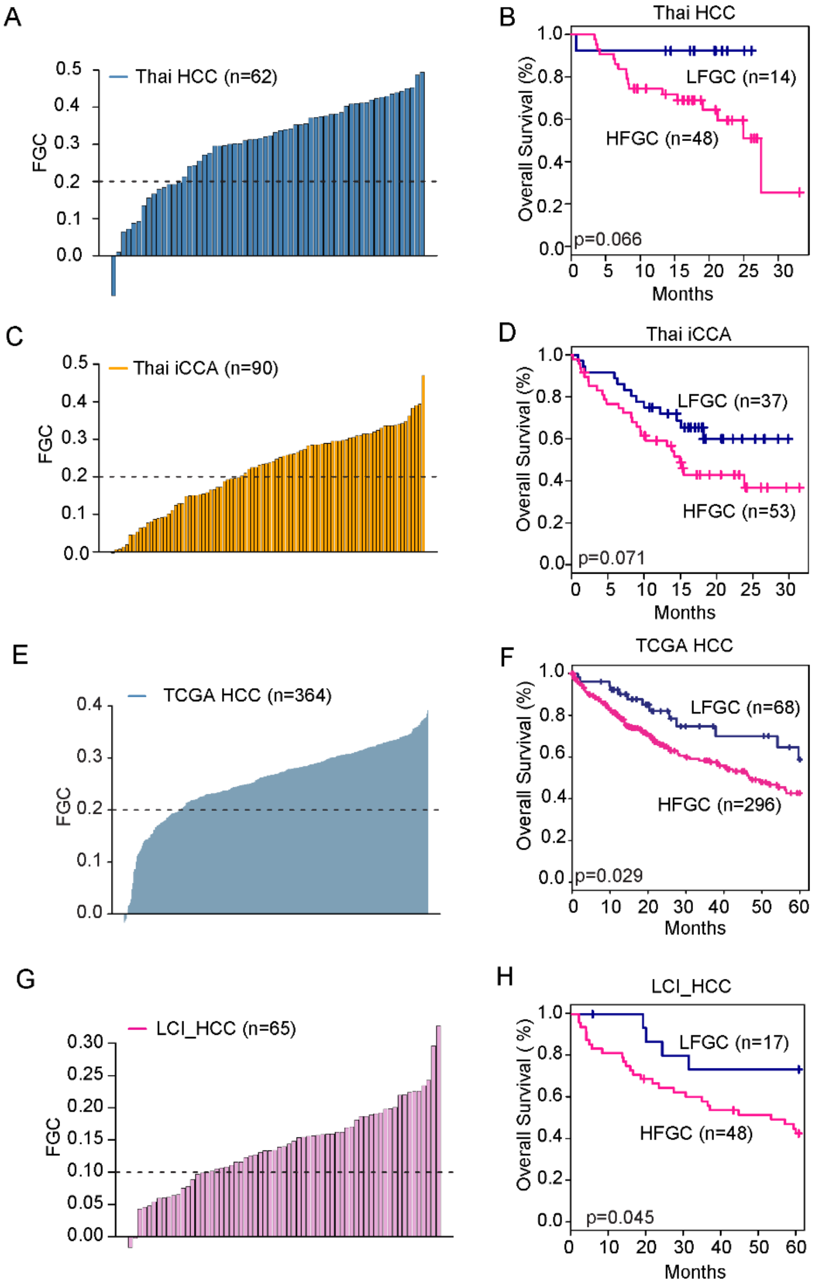


Figure S7. Validation of FGC in independent cohorts (A, C, E, and G) (A) FGC values among the Thai HCC, Thai iCCA, TCGA HCC, and LCI HCC are plotted in rank order, respectively. The dotted line indicates the cut-off FGC value, 0.2, applied to separate into FGC high (HFGC) and FGC low (LFGC) group in each tumor type, except for LCI HCC. (B, D, F, and H) Kaplan-Meier survival analysis performed based on LFGC and HFGC among the Thai HCC, Thai iCCA, TCGA HCC, and LCI HCC shows a

significant difference in the overall survival, respectively. The statistical P value was generated by the Cox-Mantel log-rank test.

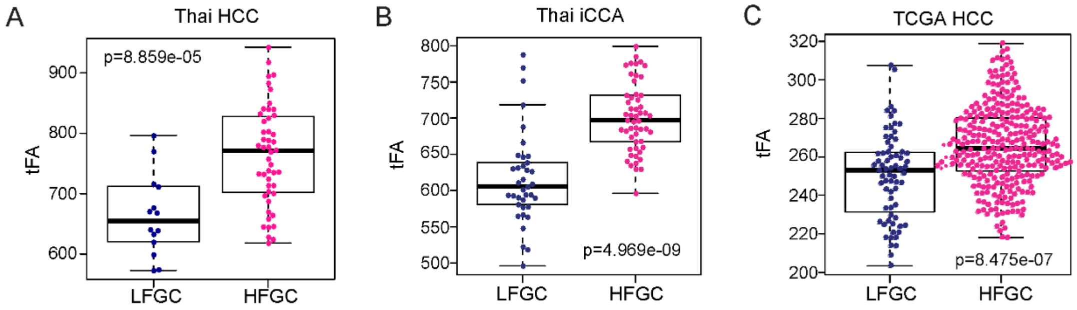


Figure S8. Comparison of tFA between HFGC and LFGC (A-C) HFGC shows the higher value of tFA in Thai HCC, Thai iCCA, and TCGA HCC, respectively. P-value based on Welch's two-sample t-test was depicted.

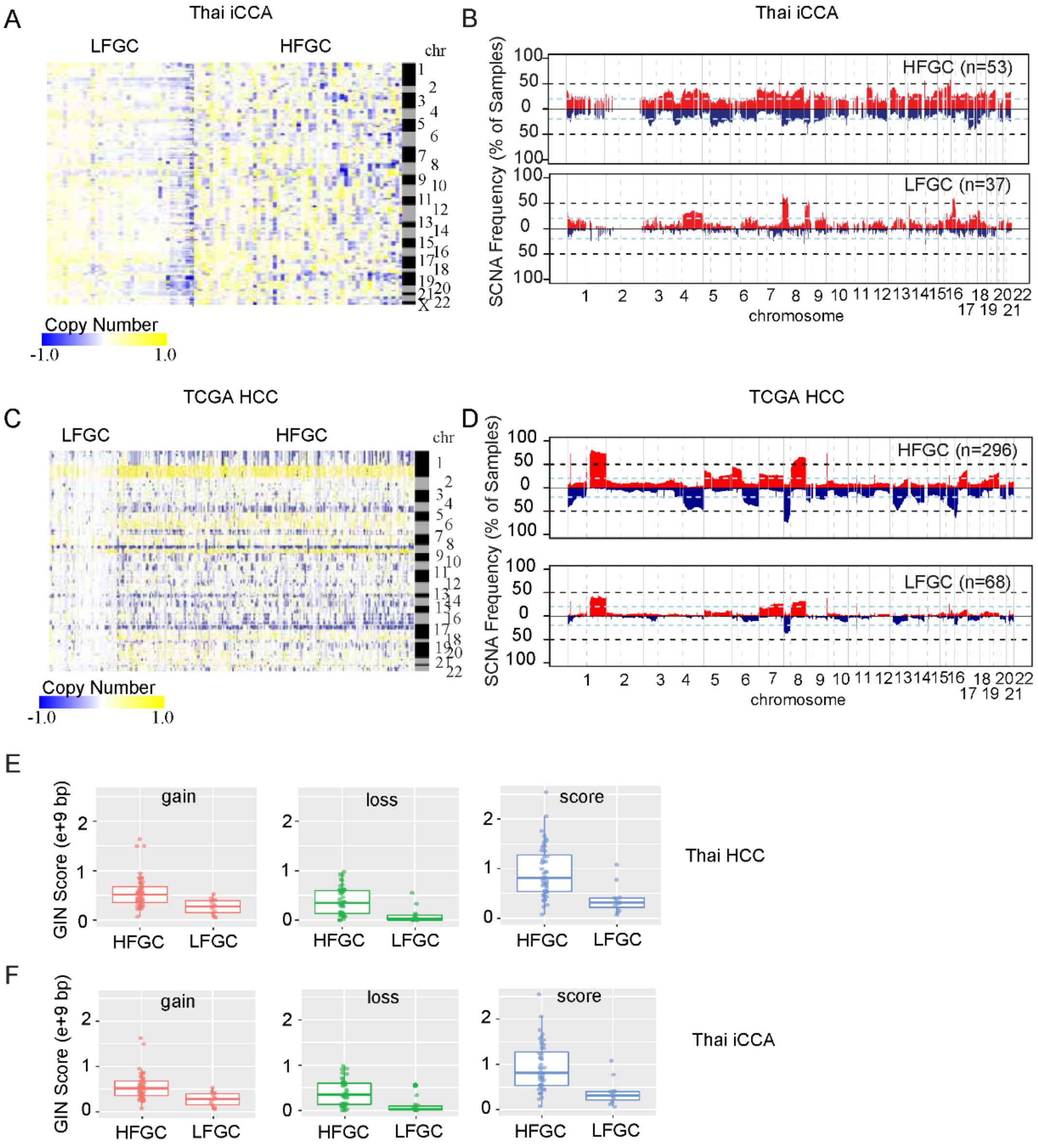


Figure S9. Comparison of SCNA in HFGC and LFGC (A and C) Heatmap shows copy number value of individual samples of Thai iCCA (A) and TCGA HCC (C) corresponding to the correlated segments regions, respectively. Samples are represented in columns, grouped by the HFGC and LFGC and segment regions are represented in rows according to the chromosomal location. (B and D) The frequency of SCNA among HFGC and the LFGC subtype of Thai iCCA (B) and TCGA HCC (D) are plotted corresponding to the correlated segmented region, respectively. The sample frequencies with copy number gain and loss (log2 (copy number) >0.2 or log2 (copy number) < -0.2)

are shown in red and blue, respectively. The upper panel is the SCNA frequency plot for HFGC subtype and lower panel is the SCNA frequency plot for LFGC subtype. Chromosome boundaries and centromere positions are indicated by vertical solid and dashed lines, respectively. Horizontal dashed blue lines indicate frequency of 50%. Horizontal dotted black lines indicate frequency of 20%. (E-F) Genomic instability (GIN) scores were compared between HFGC and LFGC. Boxplots for GIN length regarding the gain (top), loss (middle), and score (bottom) for HFGC and LFGC subtype of Thai HCC (E) and Thai iCCA (F) are shown. GIN length regarding the copy number gain or copy number loss (Methods) was calculated in the individual sample and the summation of the total SCNA length was calculated as score.

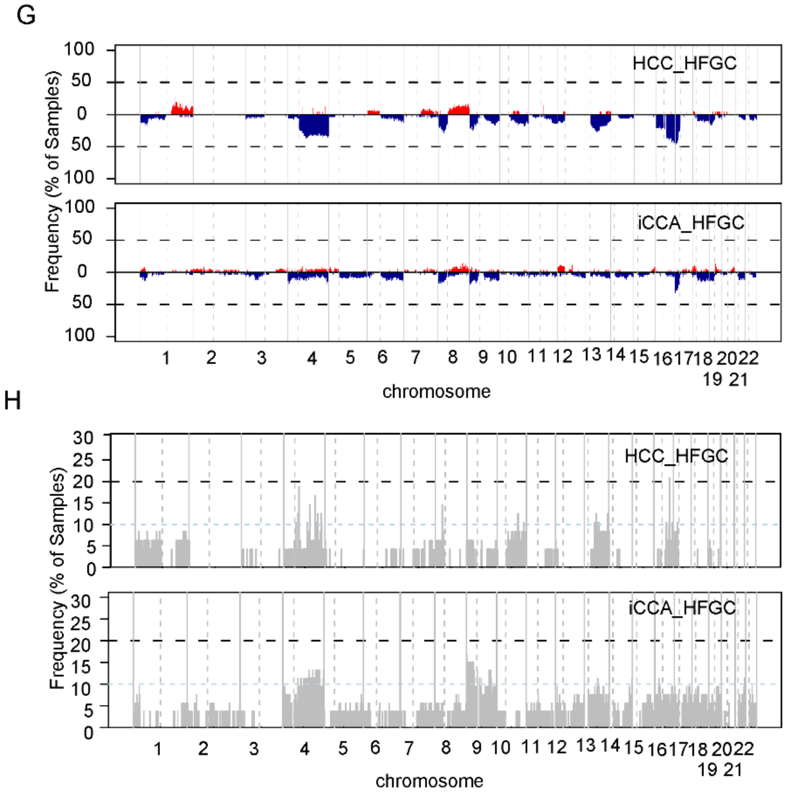


Figure S9. Comparison of SCNA in HFGC and (G-H) Allelic imbalance frequency between HFGC and LFGC was compared. (G) Frequencies of samples with amplified (AMP W/ LOH) or deletion region with LOH (DEL W/ LOH) among HCC_HFGC (upper panel) and iCCA_HFGC (lower panel) are plotted according to the chromosome location. AMP W/ LOH or DEL W/ LOH are shown in red or blue, respectively. Chromosome boundaries and centromere positions are indicated by vertical solid and dashed lines, respectively. Horizontal dashed blue lines indicate the frequency of 20%. (H) Frequencies of samples with segment region with CN LOH among HCC_HFGC (upper panel) and iCCA_HFGC (lower panel) are plotted according to the chromosome location. Chromosome boundaries and centromere positions are indicated by vertical solid and dashed lines, respectively. Horizontal dashed blue lines indicate the frequency of 10%.

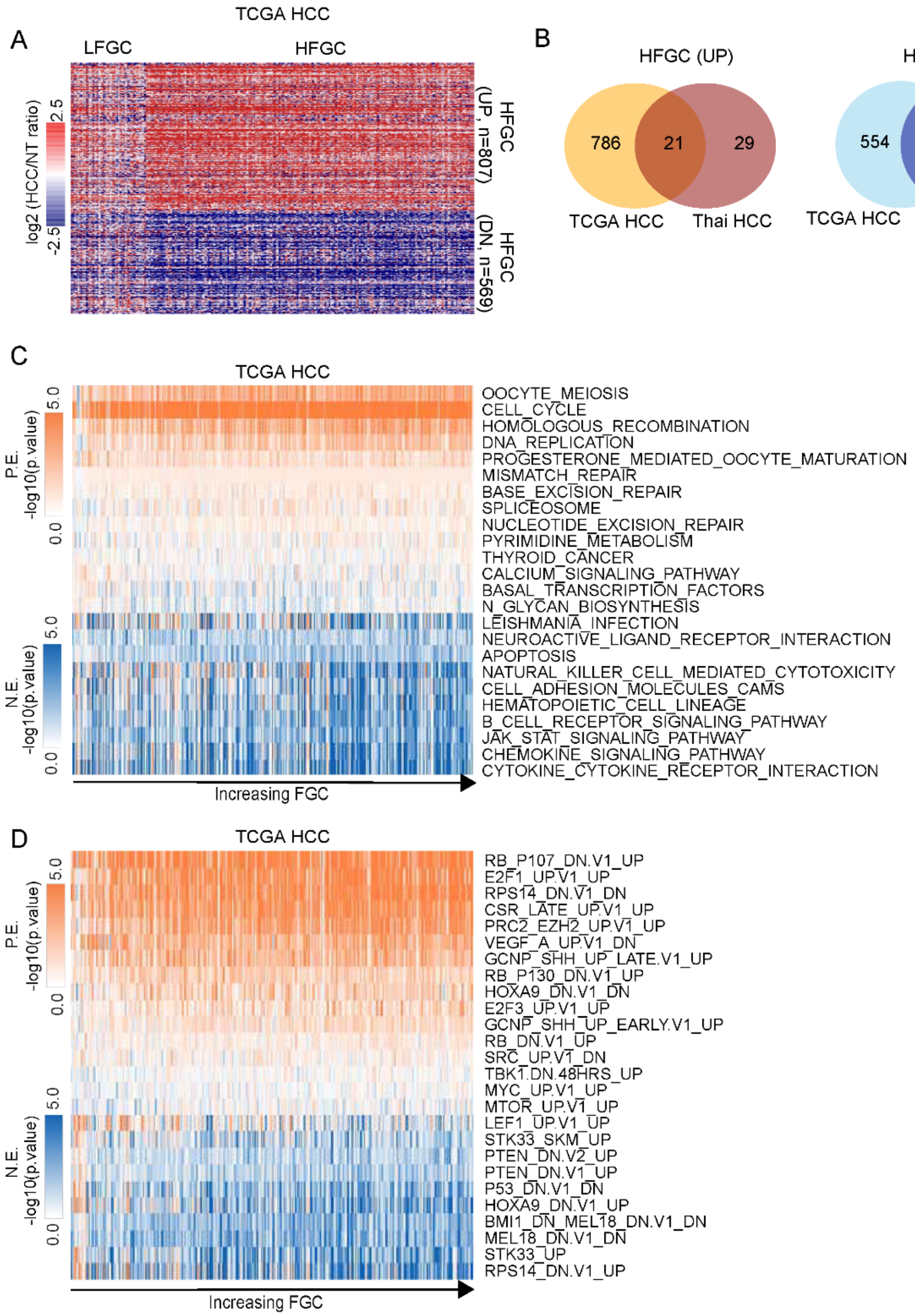


Figure S10. Differentially expressed genes (DEG) between HFGC and LFGC of TCGA (A) Heatmap shows the expression of DEG between HFGC and LFGC of TCGA HCC. 807 Up-regulated genes and 569 Down-regulated genes were selected based on the permutation t-test (p-value < 0.005 and log2 fold change >0.5 or <-0.5, respectively). Each gene expression value was normalized based on the mean of non-tumor tissue.

Samples are represented in columns, grouped by the HFGC and LFGC and genes are represented in rows. (B) Venn diagrams show the overlapped genes between DEG of HCC and of TCGA HCC. Up- and down-regulated genes are analyzed separately. (C and D) Gene Set Enrichment Analysis was performed with mRNA expression data from TCGA HCC based on the gene sets derived from the KEGG pathway gene sets (C) and oncogenic signature (D) in Molecular Signatures Database (MSigDB database v5.2). P-value from the Kolmogorov-Smirnov (ks) test was transformed in -log scaled and used in the plot. Samples were represented in columns according to the FGC increasing order and log-transformed p-value for each gene set was represented in rows in the rank-order. Shown are the gene sets selected based on the significant difference between HFGC and LFGC subtype. P.E. p-value and N.E. p-value denotes the p-value for positively and negatively enriched gene sets, respectively.

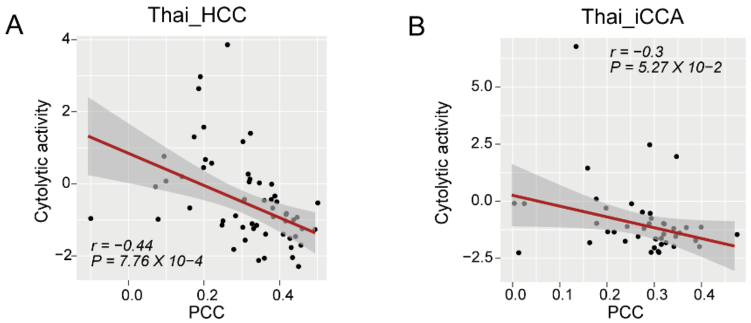


Figure S11. Association of PCC with immune cytolytic activity in Thai PLC (A and B) The association between PCC and immune cytolytic activity, defined as log-average of *GZMA* and *PRF1* expression, derived from a tumor with high tumor purity. Three different estimates for tumor purity of Thai PLC were calculated based on the IHC, ESTIMATES, and ABSOLUTE methods. Samples with high tumor purity (>0.8 of tumor purity more than 1 method) were selected and examined the association between PCC and cytolytic activity in Thai_HCC (A, n=56) and Thai_CCA(B, n=43), respectively.

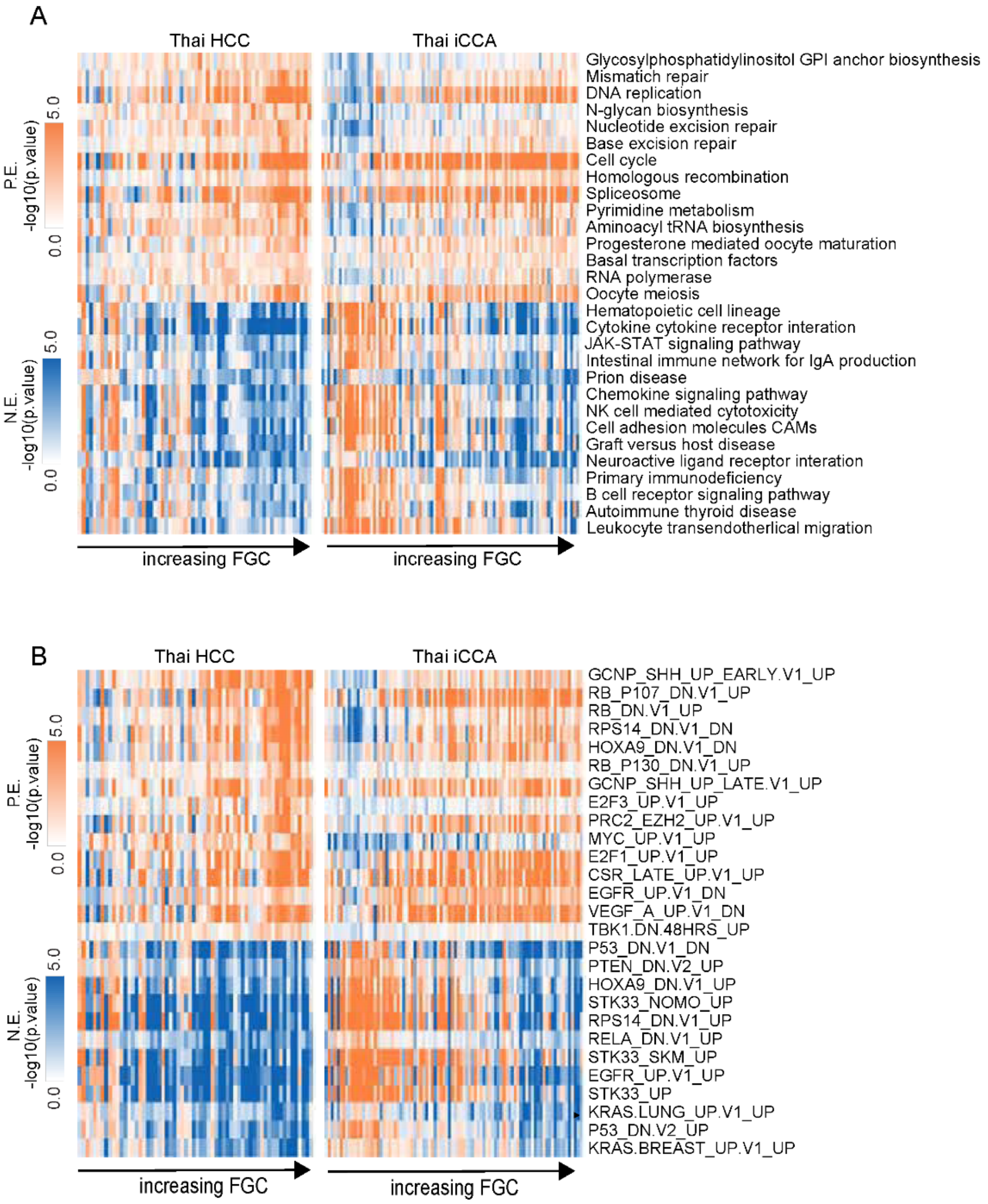


Figure S12. Gene Set Enrichment Analysis of HFGC and LFGC of Thai PLC (A and B) Single sample Gene Set Enrichment Analysis (ssGSEA) was performed with mRNA expression data from Thai HCC and Thai iCCA, respectively, based on the gene sets derived from the KEGG pathway gene sets (A) and oncogenic signature (B) (MSigDB database v5.2). P-value from the Kolmogorov-Smirnov (ks) test was transformed in -log scaled and used in the plot. Samples are represented in columns according to the rank order of FGC value and log-transformed p-value for each gene set was represented in rows. Shown are the overlapped gene sets significantly enriched both Thai HCC and

iCCA. P.E. p-value and N.E. p-value denote the p-value for positively and negatively enriched gene sets, respectively.

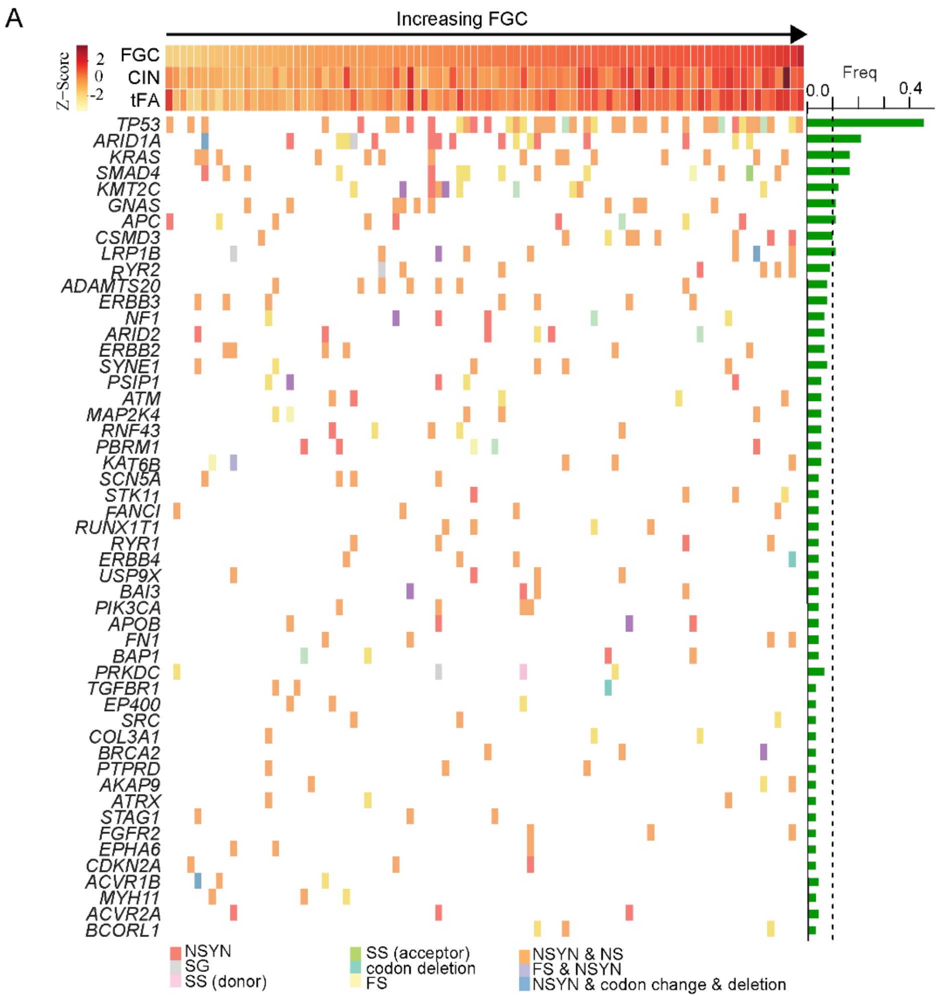


Figure S13. Integrative analysis based on PCC showed *TP53* as a Cancer functional genomic complexity (FGCs) driver. (A) (Top panel) Association between CIN, FGC, and tFA is shown in the barplot. Z-scores for FGC, CIN, and tFA in each Thai iCCA sample were plotted in each barplot in the FGC ranked order. (Bottom panel) Shown were 51 genes with mutations of more than 3 samples in Thai iCCA. The right plot shows the mutation frequency for each gene in the frequency order. The dotted line indicates the mutation frequency of 0.1. The left plot shows the occurrence of mutation regarding gene in each sample. Each bar plot represents each gene. Different color indicates different mutation type. Thai iCCA samples were represented in columns in the same order of top panel. NSYN, non-synonymous mutation; FS, frameshift mutation; SS, splice site mutation; NS, nonsense mutation.

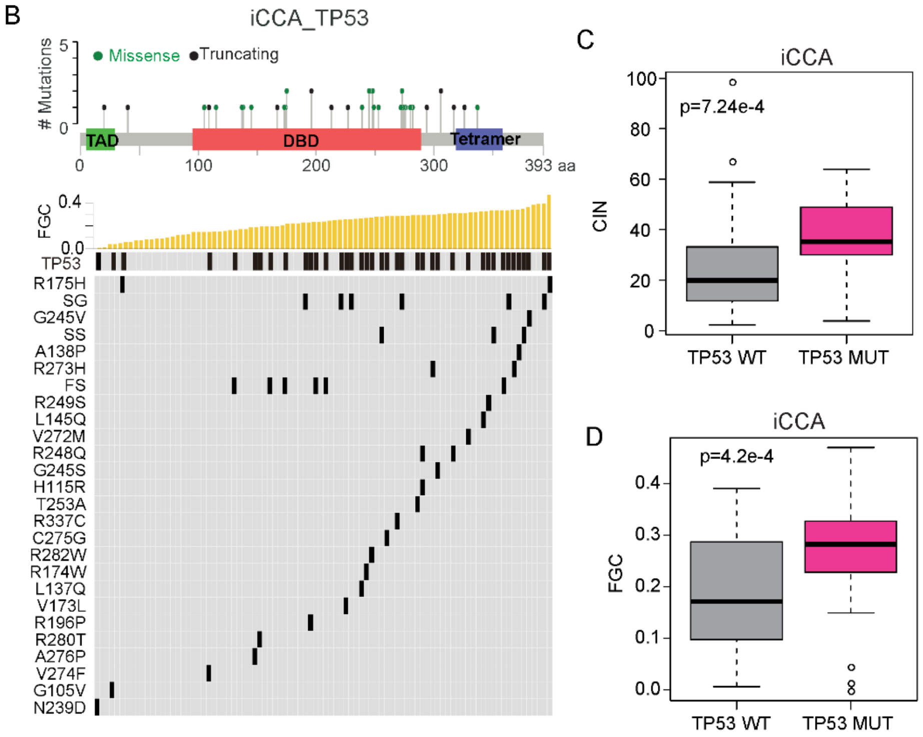


Figure S13. Integrative analysis based on PCC showed *TP53* as a Cancer functional genomic complexity (FGCs) driver. (B) (Top panel) Mutation mapper indicates the site where the *TP53* mutation occurred. Transactivation motif (TAD; 6-29), DNA binding motif (DBD; 95-288), and tetramerization motif (Tetramer; 318-358) were depicted in the different colored box; green, orange, and navy, respectively. Green or black dots indicate missense or truncating mutation, respectively. (Bottom panel) The top plot indicates the FGC score of each sample in the rank order. The incidence of TP53 mutation in each sample plotted in black in the bottom plot according to the mutation sites. (C-D) The CIN(C) and PCC (D) between *TP53* WT and *TP53* mutation among Thai iCCA. P-values based on the Welch two-sample t-test were depicted

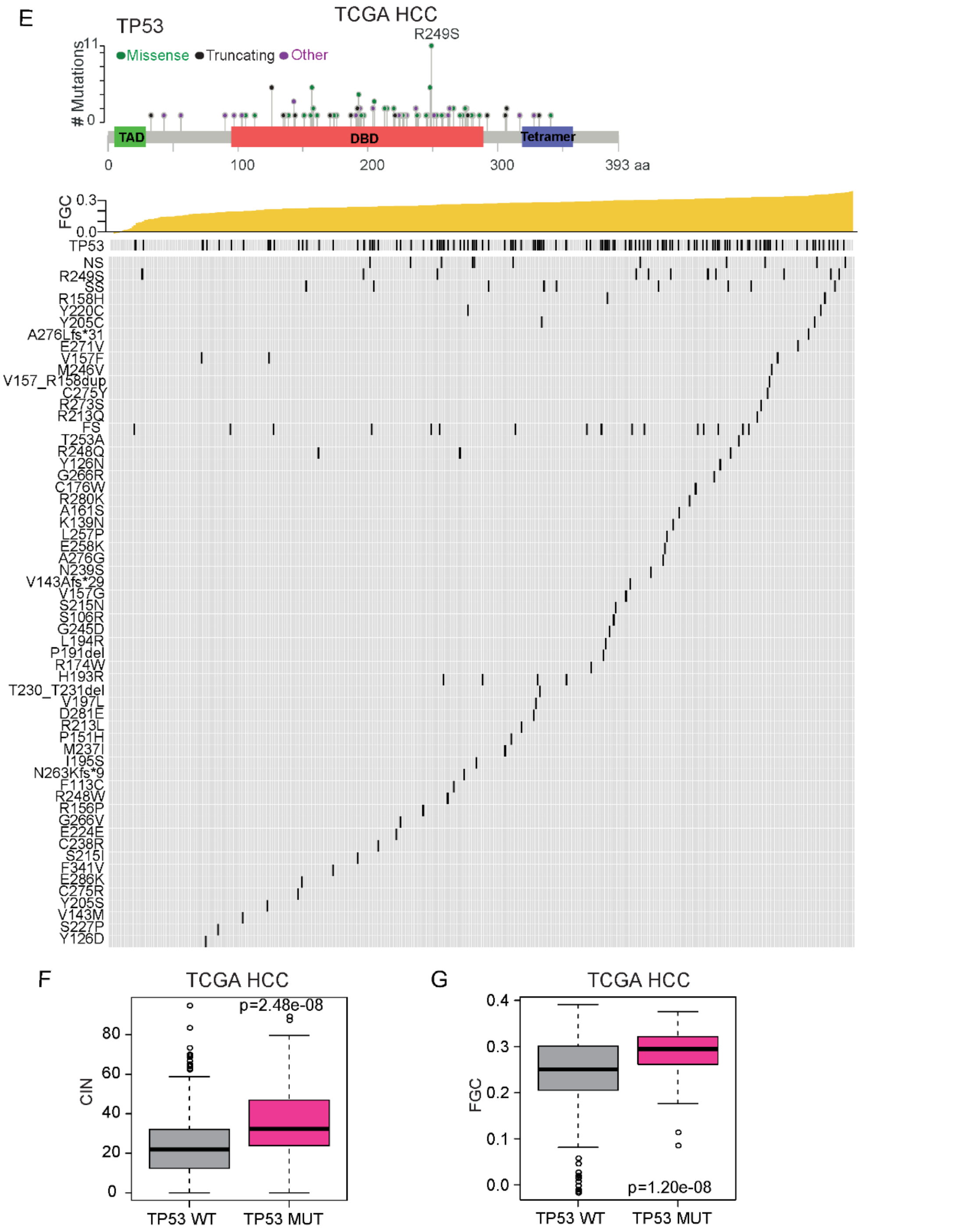
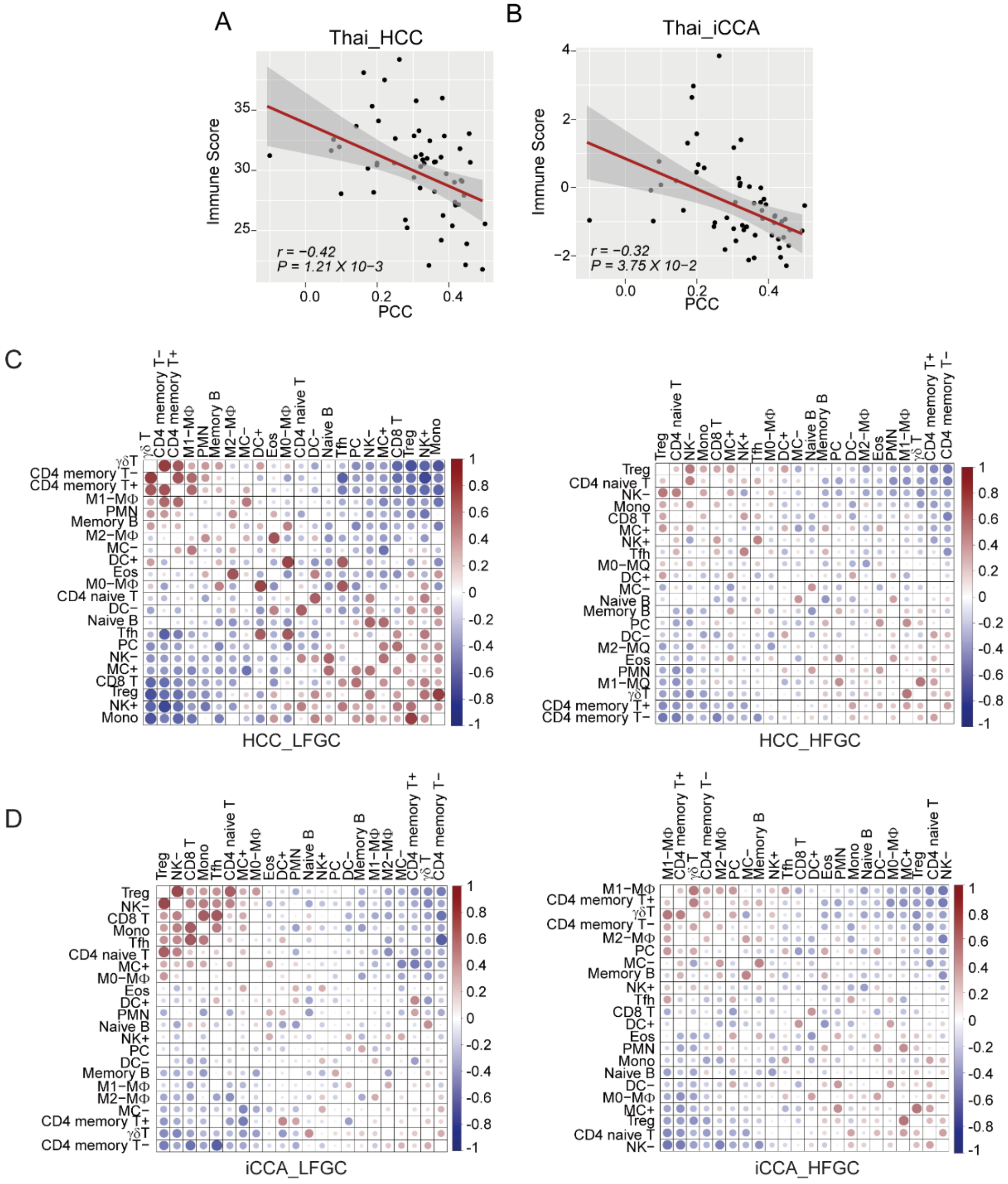


Figure S13. Integrative analysis based on PCC showed *TP53* as a Cancer functional genomic complexity (FGCs) driver. (E) (Top panel) Mutation mapper indicates the site where the *TP53* mutation occurred among TCGA HCC. Transactivation motif (TAD) (6-29), DNA binding motif (DBD) (95-288), and tetramerization motif (Tetramer) (318-358) were depicted in the different colored box; green, orange, and navy, respectively. Green

or black dots indicate missense or truncating mutation, respectively. (Bottom panel) The top plot indicates the FGC score of each sample in the rank order. The incidence of *TP53* mutation in each sample plotted in black in the bottom plot according to the mutation sites. (F-G) The CIN (F) and FGC (G) level between *TP53* WT and *TP53* mutation among TCGA HCC. P-values based on the Welch two-sample t-test were depicted. NS, SS, and FS stand for non-sense mutation, splice site mutation, and frameshift mutation, respectively


A
Thai_HCC
Immune Score
35
30
25
r = −0.42
P = 1.21 X 10−3
0.0
0.2
0.4
PCC
B
Thai_iCCA
Immune Score
4
2
0
−2
r = −0.32
P = 3.75 X 10−2
0.0
0.2
0.4
PCC
C
γδ T
CD4 memory T−
CD4 memory T+
M1−MΦ
PMN
Memory B
M2−MΦ
MC−
DC+
Eos
M0−MΦ
CD4 naive T
DC−
Naive B
Tfh
PC
NK−
MC+
CD8 T
Treg
NK+
Mono
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
HCC_LFGC
Treg
CD4 naive T
NK−
Mono
CD8 T
MC+
NK+
Tfh
M0−MΦ
DC+
MC−
Naive B
Memory B
PC
DC−
M2−MΦ
Eos
PMN
M1−MΦ
γδ T
CD4 memory T+
CD4 memory T−
M0−MQ
M2−MQ
M1−MQ
HCC_HFGC
D
Treg
NK−
CD8 T
Mono
Tfh
CD4 naive T
MC+
M0−MΦ
Eos
DC+
PMN
Naive B
NK+
PC
DC−
Memory B
M1−MΦ
M2−MΦ
MC−
CD4 memory T+
γδ T
CD4 memory T−
iCCA_LFGC
M1−MΦ
CD4 memory T+
γδ T
CD4 memory T−
M2−MΦ
PC
MC−
Memory B
NK+
Tfh
CD8 T
DC+
Eos
PMN
Mono
Naive B
DC−
M0−MΦ
MC+
Treg
CD4 naive T
NK−
iCCA_HFGC

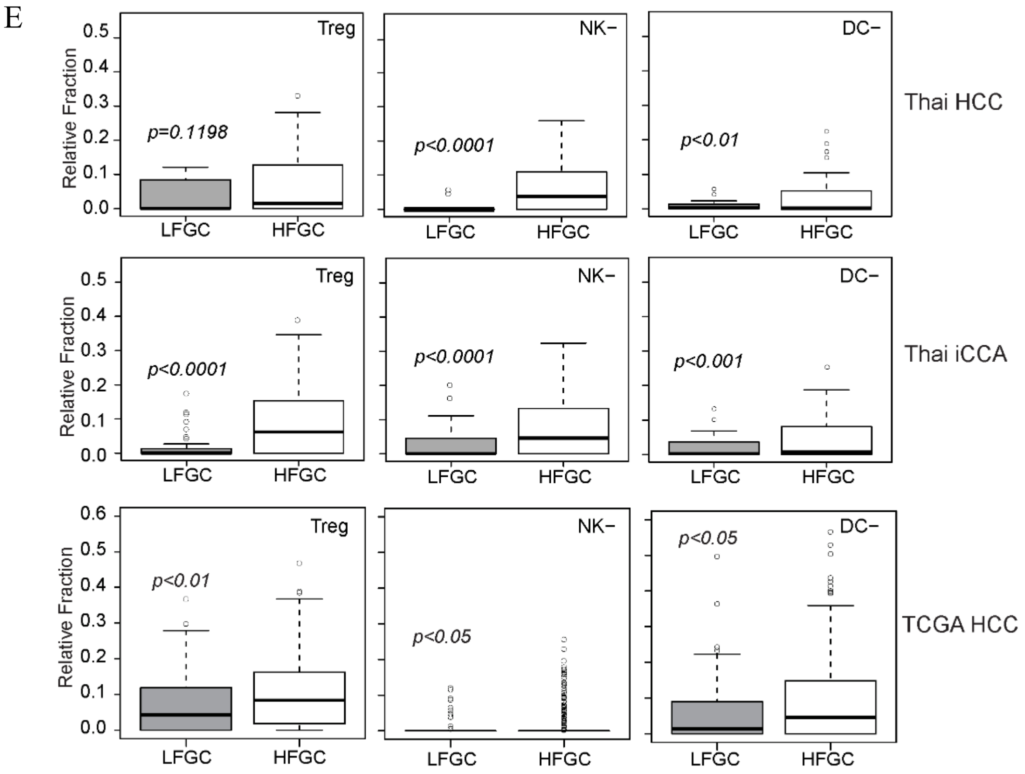
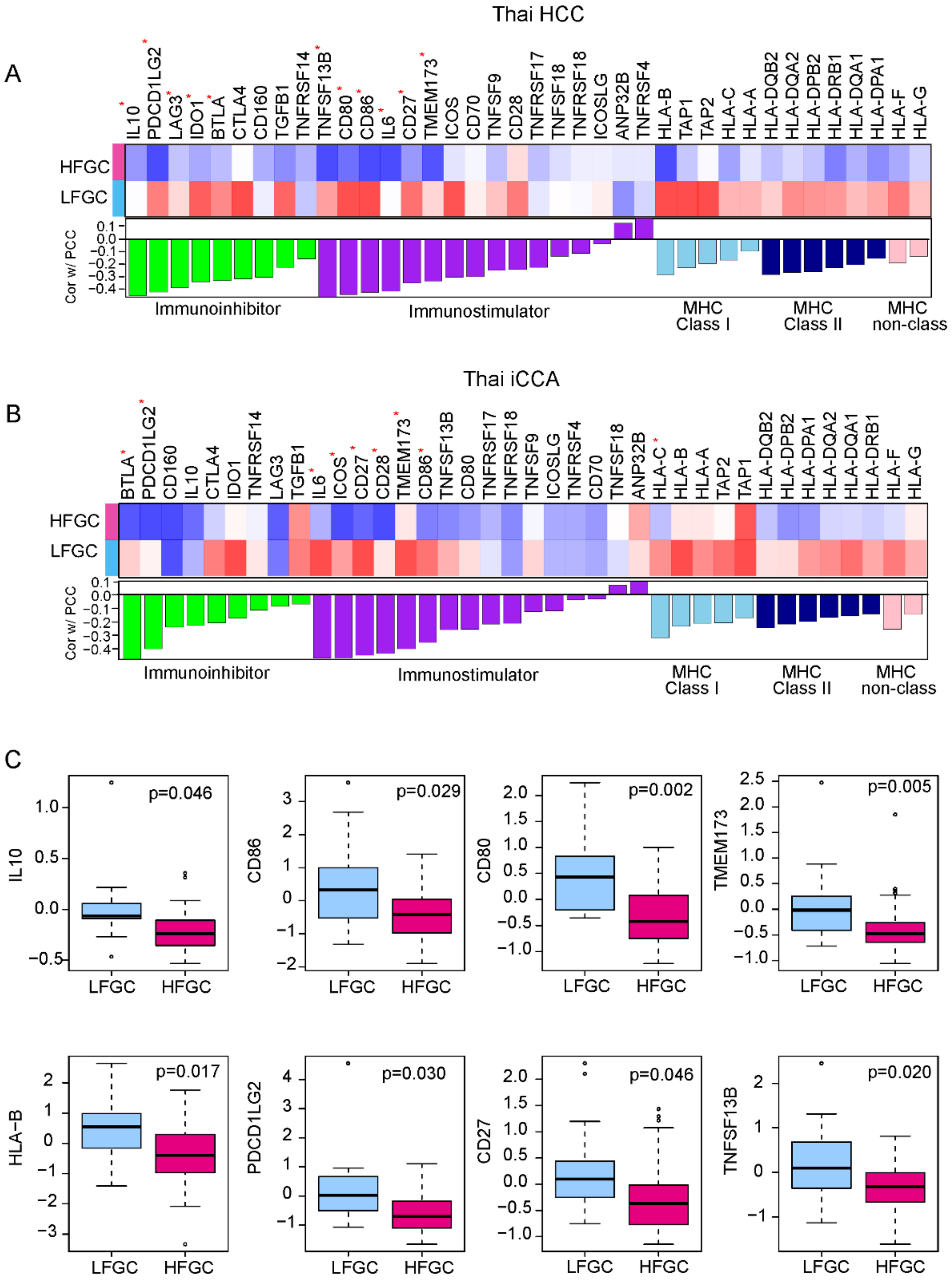


Figure S14. Association of PCC with cancer immunity in Thai PLC (A-B) Three different estimates for tumor purity of Thai PLC were calculated based on the IHC, ESTIMATES, and ABSOLUTE methods. Samples with high tumor purity (>0.8 of tumor purity more than 1 method) were selected and examined the association between PCC and immune score in Thai_HCC (A, n=56) and Thai_CCA(B, n=43), respectively. (C-D) Associations between 22 types of TIL subpopulations in the LFGC (left panel of each) and HFGC (right panel of each) of Thai HCC (C) and iCCA (D) are shown on a scale from red to blue (1 to -1). The color intensity and the size of the circle are proportional to the correlation coefficients. The proportion of 22 types of TILs based on the CIBERSORT analysis output in the LFGC and HFGC of HCC and iCCA are used. (E) Comparison of TIL subpopulations between HFGC and LFGC of Thai HCC, Thai iCCA, and TCGA HCC. Each boxplot shows the relative abundance of the TIL subpopulation between HFGC and LFGC. From left to right, representative TILs, regulatory T cell (Treg), NK- cell, dendritic cells (DC) in LFGC and HFGC of Thai HCC (top), Thai iCCA (middle), and TCGA HCC (bottom) are compared. P-values by Welch two-sample t-test are depicted in the plot.

A
Thai HCC
IL10
PDCD1LG2
LAG3
IDO1
BTLA
CTLA4
CD160
TGFB1
TNFRSF14
TNFSF13B
CD80
CD86
IL6
CD27
TMEM173
ICOS
CD70
TNFSF9
CD28
TNFRSF17
TNFSF18
TNFRSF18
ICOSLG
ANP32B
TNFRSF4
HLA-B
TAP1
TAP2
HLA-C
HLA-A
HLA-DQB2
HLA-DQA2
HLA-DPB2
HLA-DRB1
HLA-DQA1
HLA-DPA1
HLA-F
HLA-G
HFGC
LFGC
Cor w/ PCC
0.1
0.0
−0.1
−0.2
−0.3
−0.4
Immunoinhibitor
Immunostimulator
MHC Class I
MHC Class II
MHC non-class
B
Thai iCCA
BTLA
PDCD1LG2
CD160
IL10
CTLA4
IDO1
TNFRSF14
LAG3
TGFB1
IL6
ICOS
CD27
CD28
TMEM173
CD86
TNFSF13B
CD80
TNFRSF17
TNFRSF18
TNFSF9
ICOSLG
TNFRSF4
CD70
TNFSF18
ANP32B
HLA-C
HLA-B
HLA-A
TAP2
TAP1
HLA-DQB2
HLA-DPB2
HLA-DPA1
HLA-DQA2
HLA-DQA1
HLA-DRB1
HLA-F
HLA-G
HFGC
LFGC
Cor w/ PCC
0.1
0.0
−0.1
−0.2
−0.3
−0.4
Immunoinhibitor
Immunostimulator
MHC Class I
MHC Class II
MHC non-class
C
IL10
p=0.046
1.0
0.5
0.0
−0.5
LFGC
HFGC
CD86
p=0.029
3
2
1
0
−1
−2
LFGC
HFGC
CD80
p=0.002
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
LFGC
HFGC
TMEM173
p=0.005
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
LFGC
HFGC
HLA−B
p=0.017
2
1
0
−1
−2
−3
LFGC
HFGC
PDCD1LG2
p=0.030
4
3
2
1
0
−1
LFGC
HFGC
CD27
p=0.046
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
LFGC
HFGC
TNFSF13B
p=0.020
2
1
0
−1
LFGC
HFGC

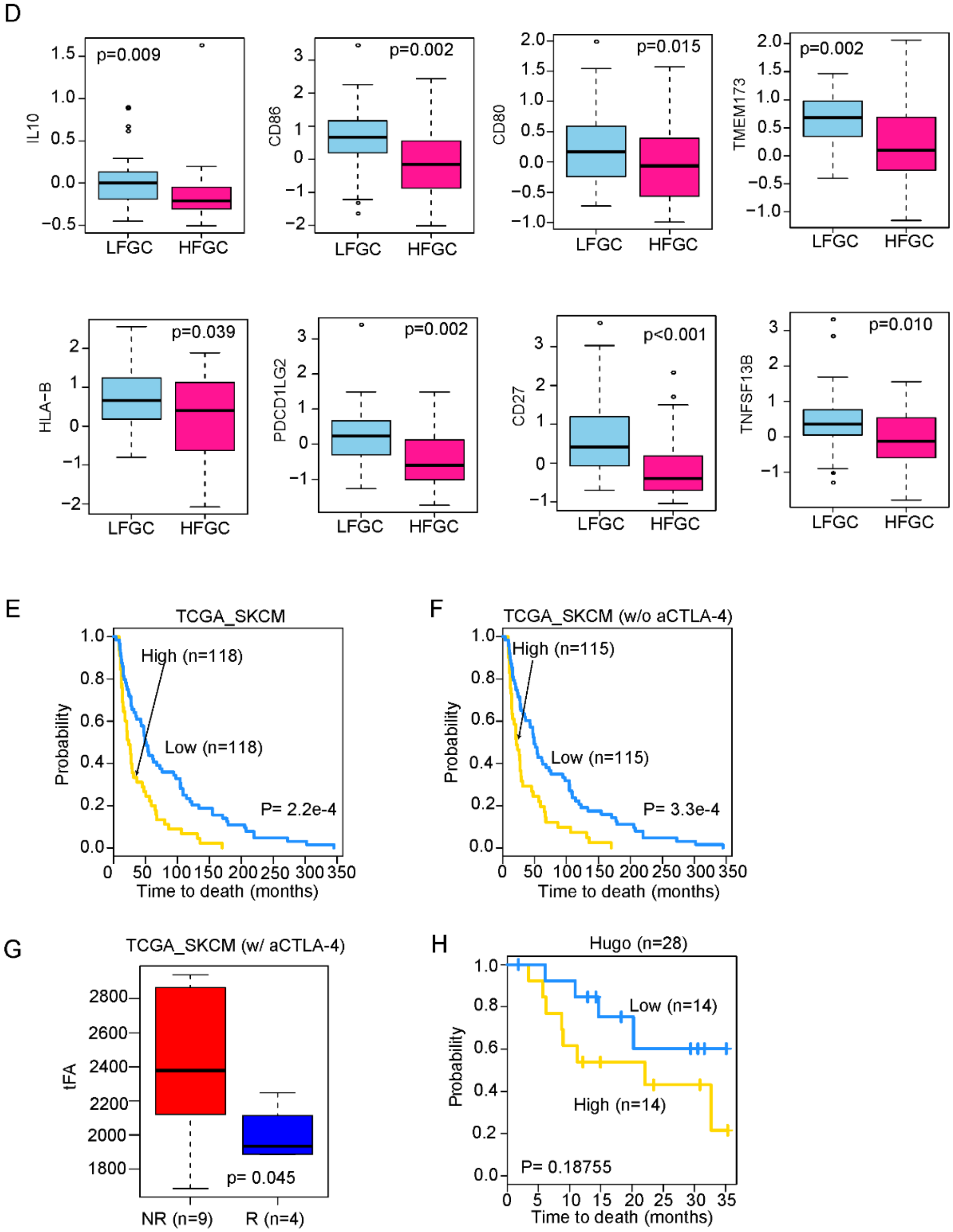


Figure S14. Association of FGC with immunomodulators (A-B) (Top panel) Heatmaps show the expression level of genes regarding selected inhibitors, stimulators of immune response, MHC class I, II, and non-class among HFGC and LFGC of Thai HCC (A) and iCCA (B), respectively. (Bottom panel) Associations of FGC with selected genes were shown in bar among Thai HCC and iCCA, respectively. Coefficient estimates and

p-value based on Pearson’s correlation were estimated. Significantly FGC associated genes were marked with red star (p-value <0.01). (C-D) Comparison of selected genes between HFGC and LFGC of Thai HCC (C) Thai iCCA (D). P-values by Welch two-sample t-test are depicted in the plot. (C-D) Comparison of selected genes between HFGC and LFGC of Thai HCC (C) Thai iCCA (D). P-values by Welch two-sample t-test are depicted in the plot. (E-G) Skin cutaneous melanoma data from TCGA (TCGA_SKCM, n=472) was used to examine the association between tFA and immunotherapy. KM survival analysis was performed including or excluding the patients' groups who received pre-treatment of anti-CTLA-4 (E and F, respectively). Patients were stratified into high and low groups based on the tFA level. Patients with tFA levels above 3rd quartile or below 1st quartile were assigned into high and low groups, respectively. (G) In the TCGA_SKCM anti-CTLA-4 pre-treatment subset, tFA levels between responders (R) and non-responders (NR) were compared. (H) Metastatic melanoma patients with pre-treatment of anti-PD-1 therapy were used. Patients were stratified into high and low groups based on the median value of tFA. KM survival analysis was performed between high and low tFA group (H). The number of patients in each group is shown.

References

1 Scharpf, R. B., Irizarry, R. A., Ritchie, M. E., Carvalho, B. & Ruczinski, I. Using the R Package crlmm for Genotyping and Copy Number Estimation. *J Stat Softw* **40**, 1-32 (2011).

2 Carvalho, B., Bengtsson, H., Speed, T. P. & Irizarry, R. A. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* **8**, 485-499, doi:10.1093/biostatistics/kxl042 (2007).

3 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

4 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572, doi:10.1093/biostatistics/kxh008 (2004).

5 Roessler, S. *et al.* A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* **70**, 10202-10212, doi:10.1158/0008-5472.CAN-10-2607 (2010).

6 Roessler, S. *et al.* Integrative genomic identification of genes on 8p associated with hepatocellular carcinoma progression and patient survival. *Gastroenterology* **142**, 957-966 (2012).

7 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

8 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

9 Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043-1048, doi:10.1038/ng1861 (2006).

10 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

11 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).

12 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1, doi:10.1126/scisignal.2004088 (2013).

13 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401-404, doi:10.1158/2159-8290.CD-12-0095 (2012).

14 Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).

15 Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **168**, 542, doi:10.1016/j.cell.2017.01.010 (2017).