SUPPLEMENTARY MATERIAL

Supplementary tables

Table S1: Separate Excel file: Genes enriched at different stages of liver development

Table S2: Separate Excel file: Liver epithelial-differentiation genes suppressed in HCC

Table S3: Separate Excel file: Genes differentially expressed between normal liver and HCC

Table S4 Separate Excel file: Mutations identified by targeted next generation sequencing of *GATA4, HNF1A, ARID1A, ARID2, SMARCA4, SMARCAD1 CTNNB1* and *TP53.*

Table S5 Separate Excel file: Peptides/proteins identified by LCMS/MS in GATA4 and GATA4-V267M co-immunoprecipitates

Table S6 Separate Excel file: Primers used for murine genotyping, QRT-PCR, infusion cloning, Sanger sequencing and targeted deep sequencing

Supplementary figures

Figure S1: GATA4 expression in HCC *vs.* normal liver in three publically available gene expression datasets

Figure S2: Body weights, liver weights, H&E and KI67 staining with liver-conditional *Gata4* haploinsufficiency

Figure S3: Chromatin state at baseline in ESC of hepatocyte commitment and late-differentiation transcription factor genes with high and low expression respectively in *Gata4* haploinsufficient livers

Figure S4: Gene expression of hepatocyte epithelial-differentiation genes in normal liver versus HCC (TCGA series) classified by histological grade

Figure S5: Expression pattern in HCC of transcription factors that drive different stages of hepatocyte maturation

Figure S6: Copy number variance (CNV) analysis of HCC cell lines PLC and HepG2, and sequencing of *ARID1A* in HepG2

Figure S7 Next generation and conventional Sanger sequencing data of *GATA4*, showing germ-line mutation *GATA4* V267M in two cases.

Figure S8: Cytoplasmic and nuclear localization of GATA4 wild-type and GATA4 V267M

Figure S9: SDS-PAGE of Flag-GATA4 and Flag-GATA4 V267M immunoprecipitates

Figure S10: Relative enrichment of major peptides in the GATA4 *vs.* GATA4 V267M protein interactomes

Figure S11: Immunoprecipitation-Western blots of GATA4 and GATA4 V267M (Western blots for Flag-GATA4 or Flag-GATA4 V267M, MED12 and SMARCA5)

Figure S12: Mutation of GATA4 coactivators in HCC.

Figure S13: Expression of key hepatocyte differentiation-driving transcription factors in HCC grouped by genetic inactivating alterations in genes for GATA4 or GATA4 coactivators.

Figure S14: Targeted deep sequencing of *CTNNB1* and *TP53* in the Singapore HCC series

SUPPLEMENTARY FIGURES

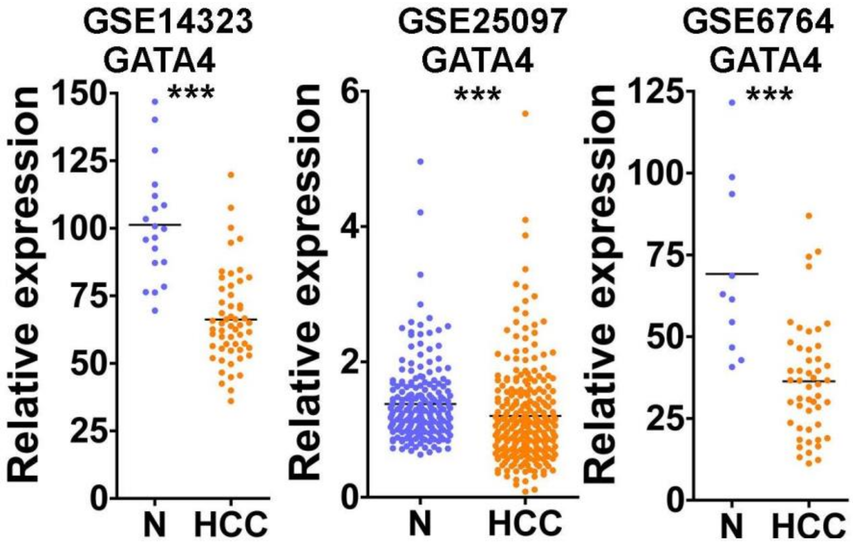

Figure S1. GATA4 is significantly less expressed in HCC vs. adjacent non-malignant liver in multiple gene expression datasets. Gene expression data was downloaded from GEO database and expression levels were analyzed in normal versus HCC cases in three independent studies. Data with the accession number GSE14323 - HCV associated HCC, GSE25097 - HCC with varying stages from early to advanced HCC, and GSE6764 - HCV-induced HCC at vaious stages were analyzed. In all three studies GATA4 was significantly less expressed in HCC relative to normal ***p<0.0001 Wilcoxon rank sum test.

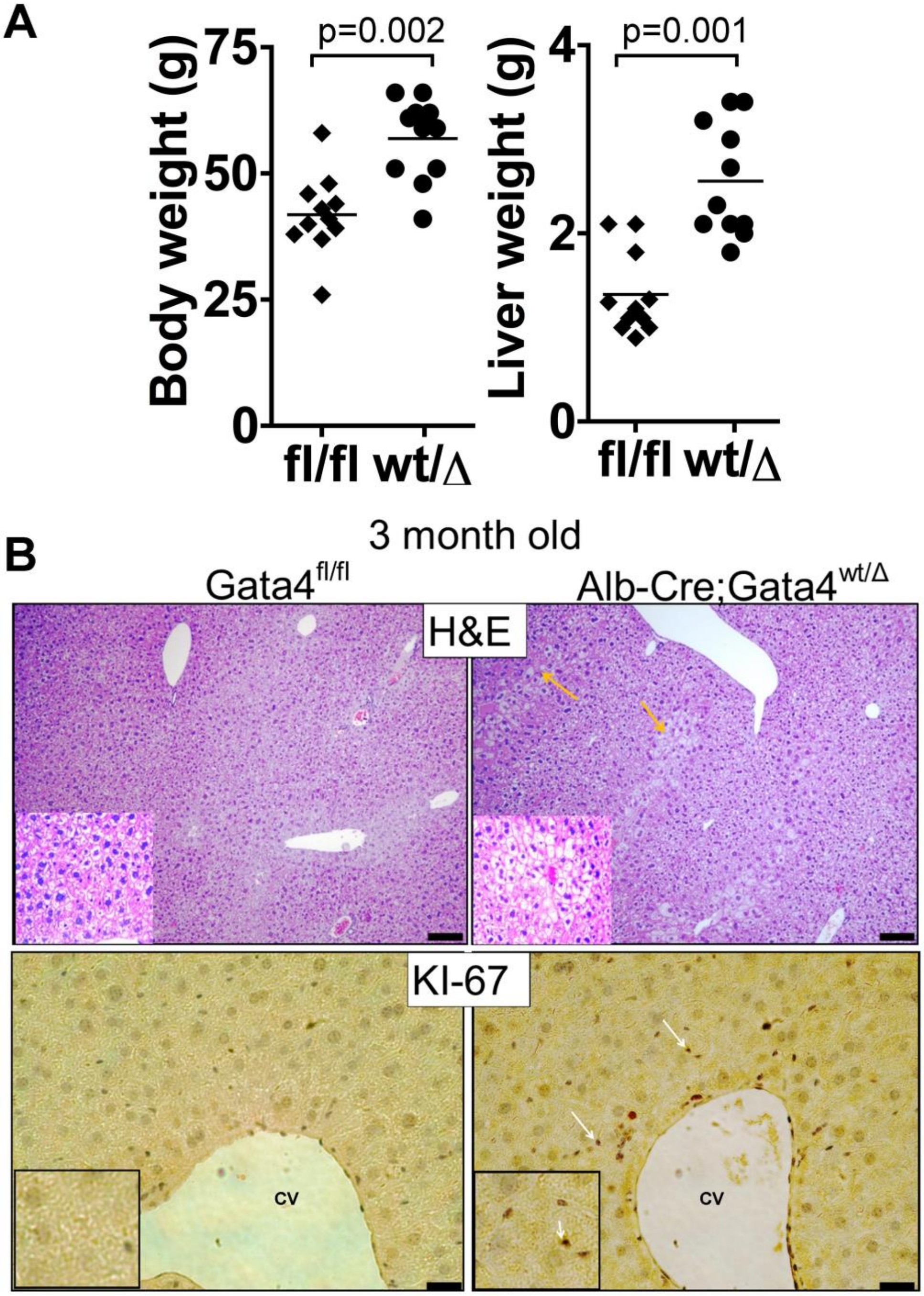

Figure S2. Phenotype analyses of liver-conditional *Gata4* haploinsufficient mice. **A)** Overall bodyweight and liver weights of *Gata4*$^{fl/fl}$ versus *Gata4*$^{wt/\Delta}$ measured at 8 months. **B)** Hematoxylin and eosin staining (H&E) and proliferation marker KI67 analysis by immunohistochemistry in *Gata4*$^{fl/fl}$ versus *Gata4*$^{wt/\Delta}$ livers at 3 months. Yellow arrows = lipid accumulation, White arrows = KI67 positive nuclei. CV= central vein

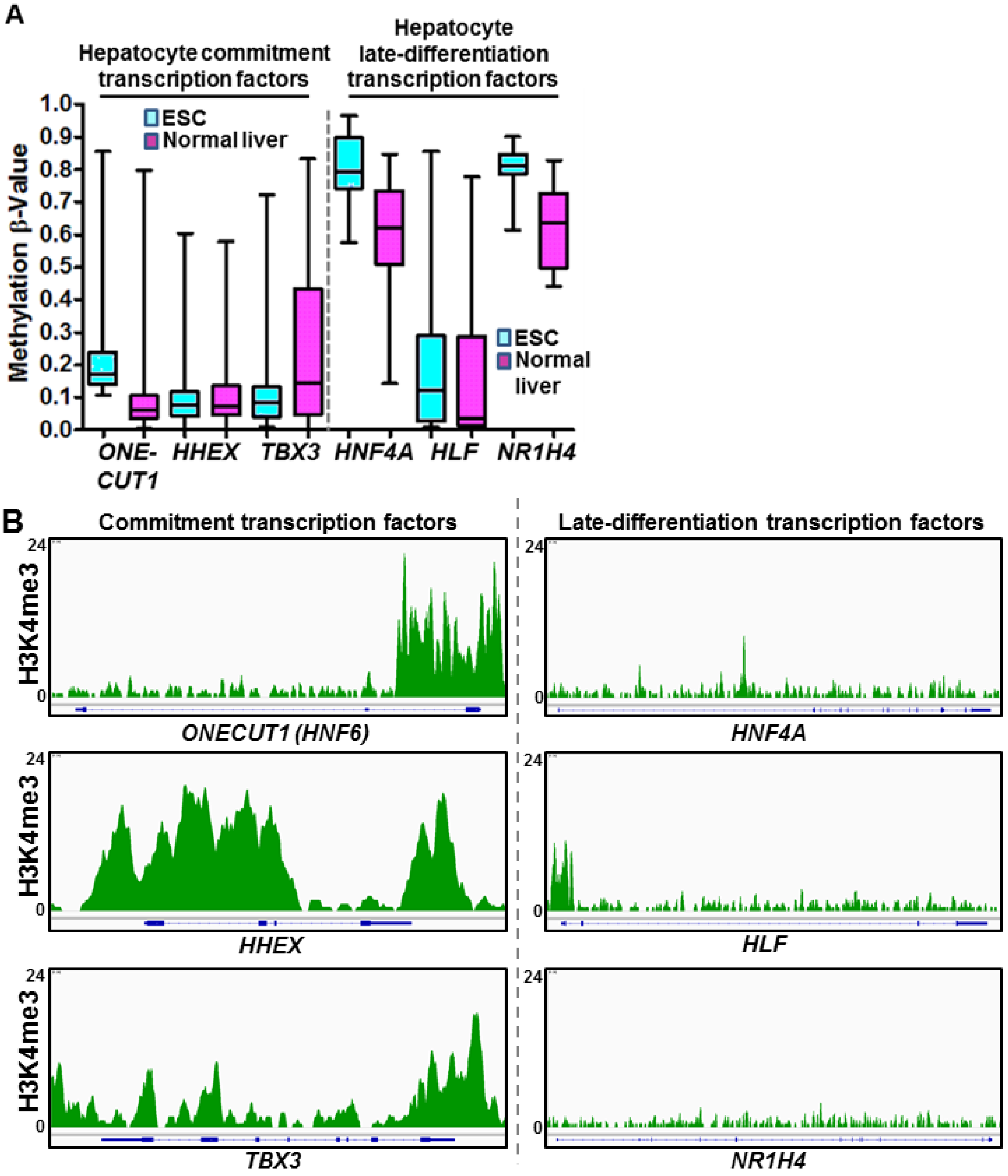

Figure S3. A) Hepatocyte commitment/early-differentiation transcription factor genes that are highly expressed in *Gata4* haploinsufficient livers have chromatin that is poised for gene activation, that is DNA CpG hypomethylated, even in the earliest tissue precursors embryonic stem cells (ESC), while late-differentiation transcription factor genes that are suppressed in *Gata4* haploinsufficient livers have repressed chromatin with high CpG methylation levels at this baseline. Plotted are medians and interquartile range of methylation values (β-values) by Illumina 450k CpG array for the CpG linked with these genes. β-values ESC (n=19) and normal liver (n=4) from GSE31848 (Ref.1). *ONECUT1* 33 CpG, *HHEX* 19 CpG, *TBX3* 31 CpG, *HNF4A* 27 CpG, *HLF* 18 CpG, *NR1H4* 11 CpG. **B) The poised character of over-expressed hepatocyte commitment transcription factor *vs.* suppressed hepatocyte late-differentiation transcription factor genes in ESC was also seen by chromatin-immunoprecipitation sequencing (ChIP-Seq) analysis for the epigenetic activation mark H3K4me3**. ChIP-Seq data in H1 ESC from Encode. *Reference:* 1.Nazor, K.L., Altun, G., Lynch, C., Tran, H., Harness, J.V., Slavin, I., Garitaonandia, I., Muller, F.J., Wang, Y.C., Boscolo, F.S., et al. 2012. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. Cell Stem Cell 10:620-634.

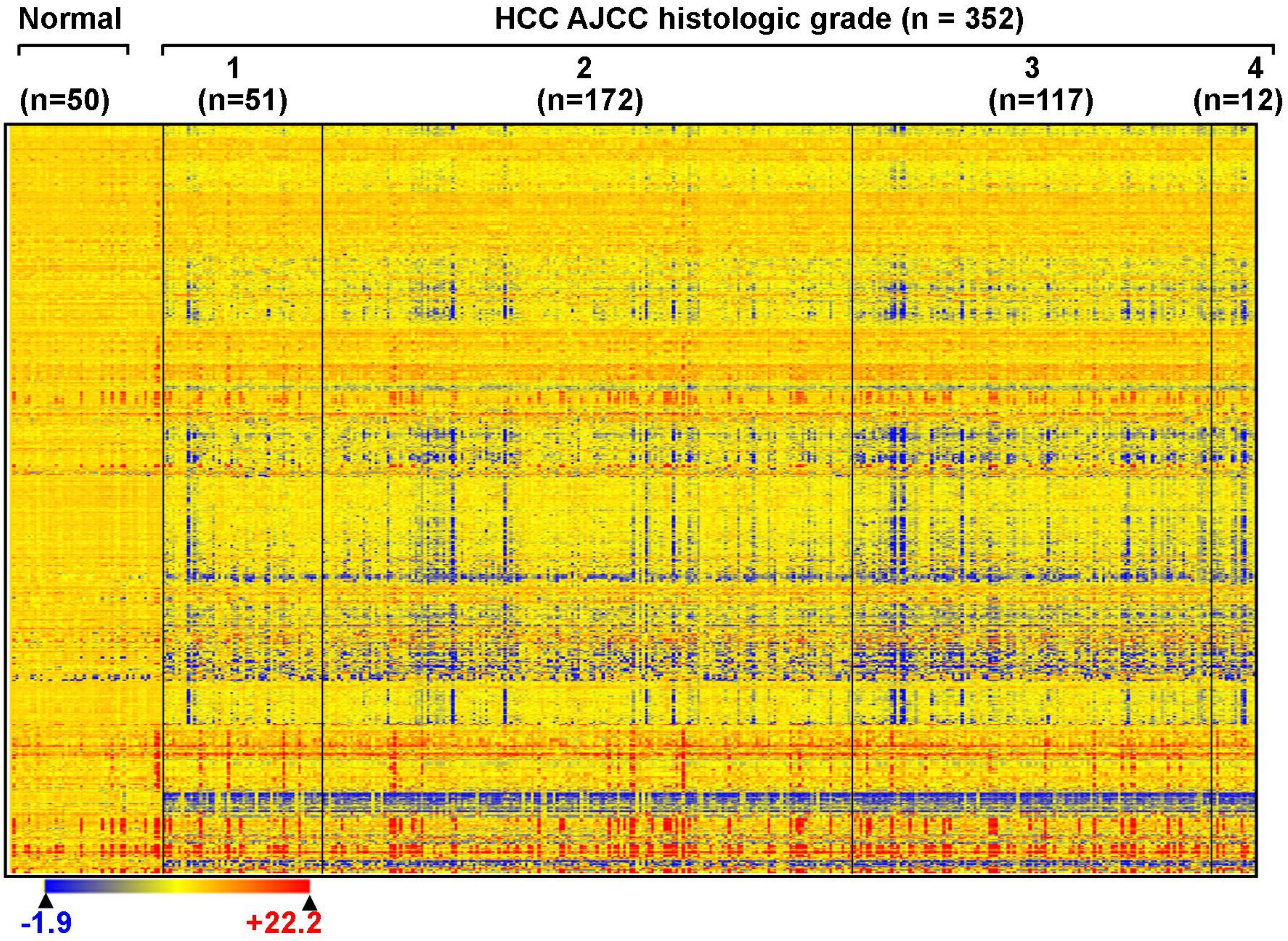

Figure S4. Suppression of hepatocyte epithelial-differentiation genes in all histological grades of HCC. Hepatocyte epithelial-differentiation genes identified by DAVID gene ontology analysis and suppressed in HCC vs. normal liver in the Singapore HCC series (600 genes) were analyzed for expression levels by RNA sequencing in the TCGA HCC series stratified by histological grade (TCGA LIHC)(normal liver n=50, HCC n=352). Histological grade = American joint committee (AJCC) pathological staging.

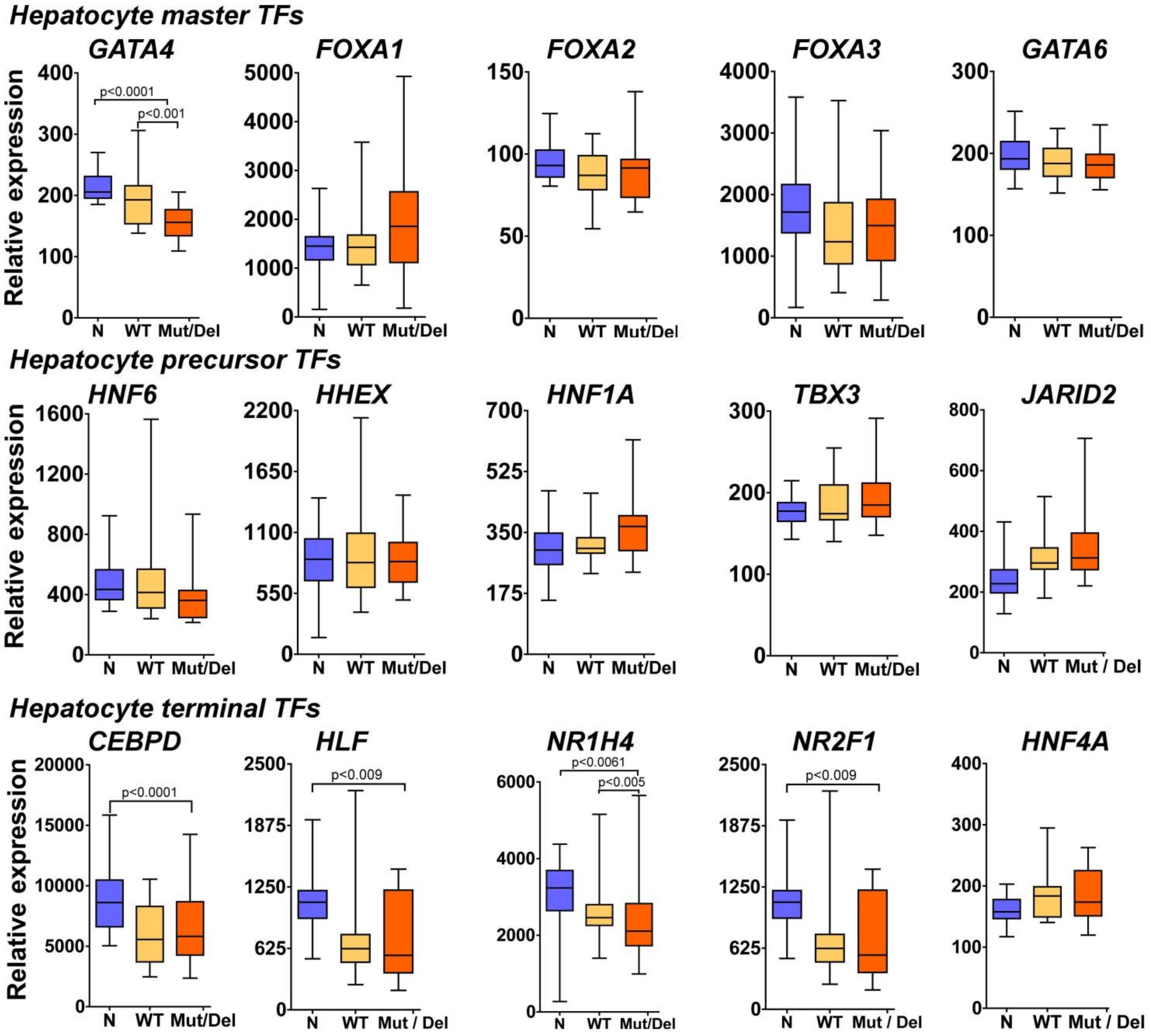

Figure S5. Expression levels of hepatocyte transcription factors in normal liver *vs.* HCC without *GATA4* haploinsufficiency *vs.* HCC with *GATA4* haploinsufficiency. Of the early master transcription factors essential for generating the hepatocyte lineage, only *GATA4* was significantly less expressed in HCC versus normal livers. *GATA4* expression was lowest in HCC with 8p-loss (*GATA4* haploinsufficiency). Hepatocyte precursor transcription factors (*HNF1A*, *HNF6 TBX3 HHEX JARID2*) had preserved expression in HCC, however, hepatocyte late-differentiation transcription factors *CEBPD*, *HLF*, *NRIH4* and *NR2F1* were significantly less expressed in HCC *vs.* normal liver.

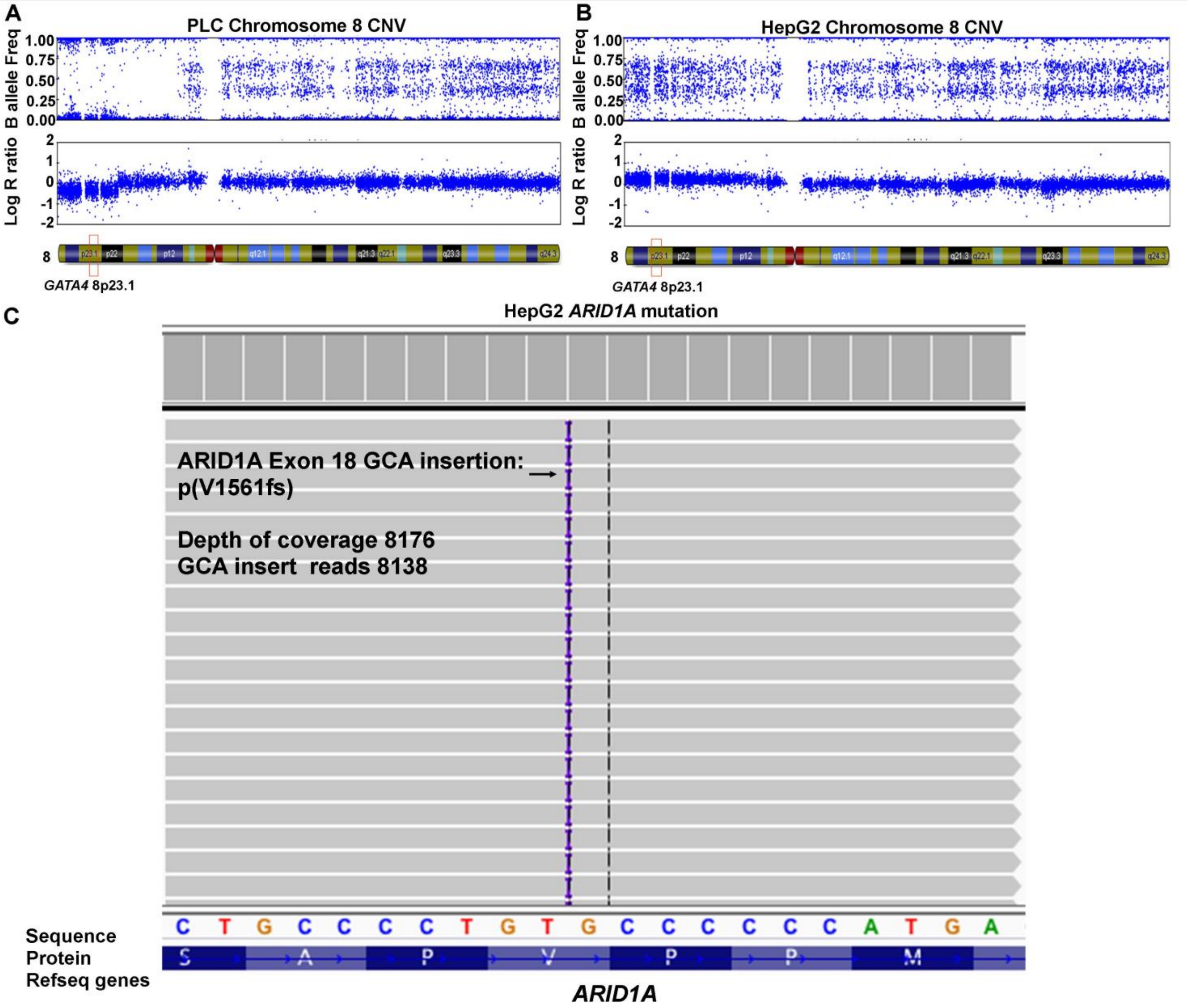

Figure S6. Copy number variance (CNV) analysis of HCC cell lines PLC and Sk-HepG2. DNA was isolated from PLC and Sk-HepG2 cells and analyzed by SNPA array. B allele frequency = probability to observe one parental allele, LogR ratio = logarithm of observed over expected. **A)** Chromosome 8p deletion incorporating the *GATA4* locus in PLC cells **B)** No chromosome 8p deletion in HepG2. **C)** All coding regions of *ARID1A* were sequenced by next generation sequencing, identifying an insertion frameshift mutation in exon 18 altering the amino acid Valine1561 by frameshift (pV1564*fs). This mutation is also noted in COSMIC with accession number COSM211769.

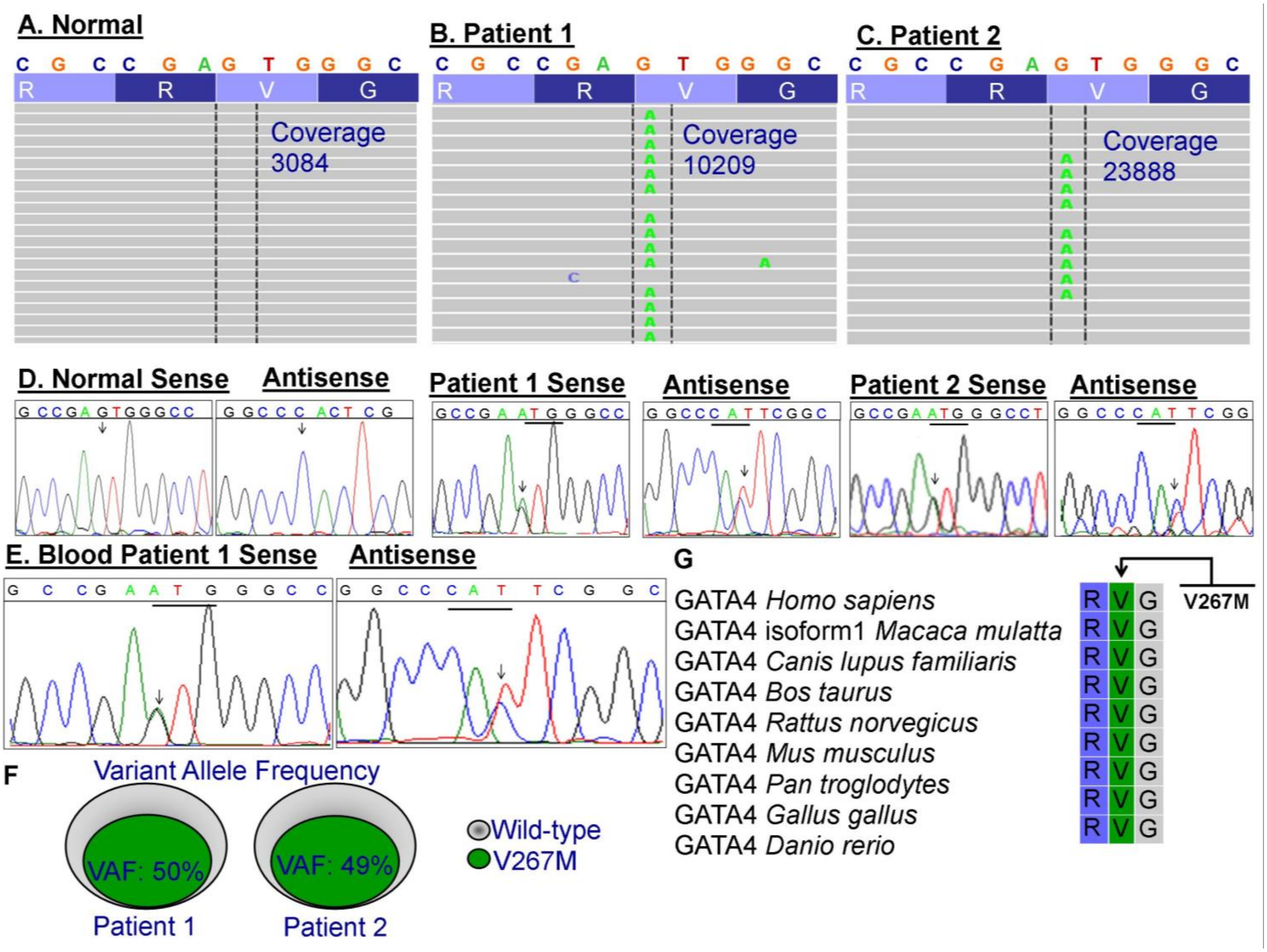

Figure S7. The same missense germ-line mutation of *GATA4* was identified in two cases of atypical HCC. **A)** Targeted next generation deep sequencing of *GATA4* exon4 in normal wild-type DNA coverage = 3084. **B)** *GATA4* exon4 mutation of HCC Patient 1 that alters the amino acid Valine 267 to Methionine (V267M), coverage = 10209. **C)** Same mutation in HCC Patient 2, coverage = 23888. **D)** Sanger sequencing results showing the same *GATA4* V267M mutation; normal wild-type DNA used as control, black arrowheads indicate mutated allele (sequence show both sense and antisense strands) black line indicates mutated codon (GTG>ATG). **E)** Sanger sequencing analysis of DNA from peripheral blood mononuclear cells, confirming germline origin of the mutation (results for Patient 1 shown). **F)** Variant allele frequency of the mutation (variant reads over total number of reads). **G)** Amino acid V267 altered by the mutation is conserved in multiple species.

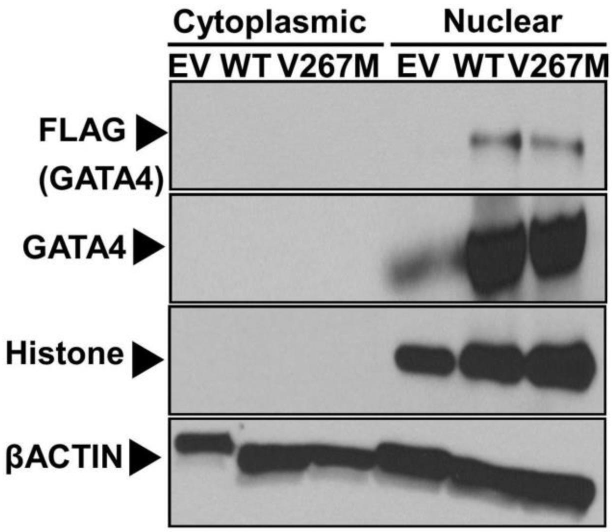

Figure S8. GATA4 and GATA4 V267M protein both localize to the nucleus. HCC cells (PLC) were transfected with expression vectors for Flag-tagged GATA4 or GATA4 V267M. Cytoplasmic and nuclear protein lysates were generated and analyzed by Western blot. Flag antibody only detects transfected GATA4, GATA4 antibody detects both transfected and endogenous GATA4. Histone antibody was used as control for nuclear protein lysates. Actin antibody was used as loading control for both nuclear and cytoplasmic fractions.

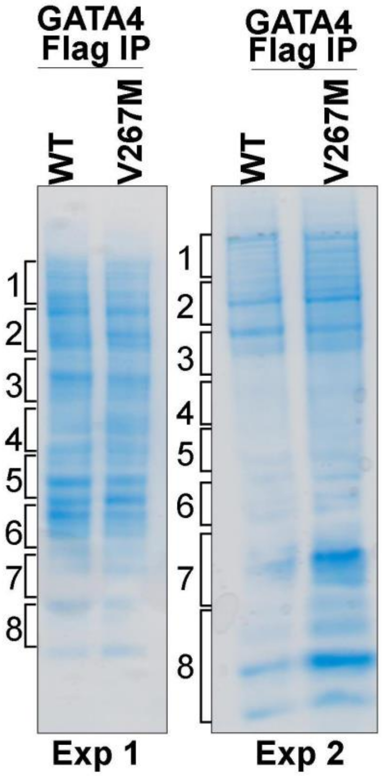

Figure S9: SDS-PAGE separation of Flag-GATA4 and Flag-GATA4 V267M co-immunoprecipitate from HCC cells (PLC) transfected with expression vectors for Flag-GATA4 and Flag-GATA4 V267M. Gels were stained with Coomasie Blue dye. Regions indicated by 1-8 were excised and trypsin-digested for extraction of proteins. Purified proteins were then analyzed by LCMS/MS.

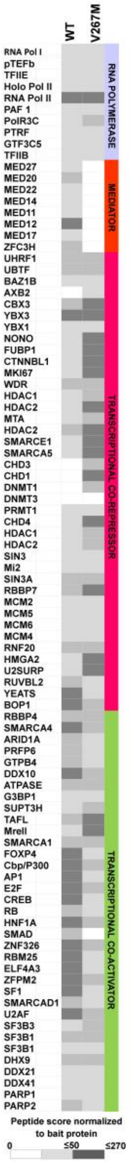

Figure S10. Major Peptides identified in GATA4 *vs.* GATA4 V267M interactomes. Liquid chromatography tandem mass spectrometry (LCMS/MS) analyses. *GATA4* deficient HCC cell line PLC was transfected with expression vectors for flag-GATA4 or flag-GATA4 V267M. Nuclear protein was subject to immunoprecipitation by anti-flag antibody and the protein interactome was analyzed by LCMS/MS. Identified peptides were normalized to bait protein (GATA4) from two biological replicates.

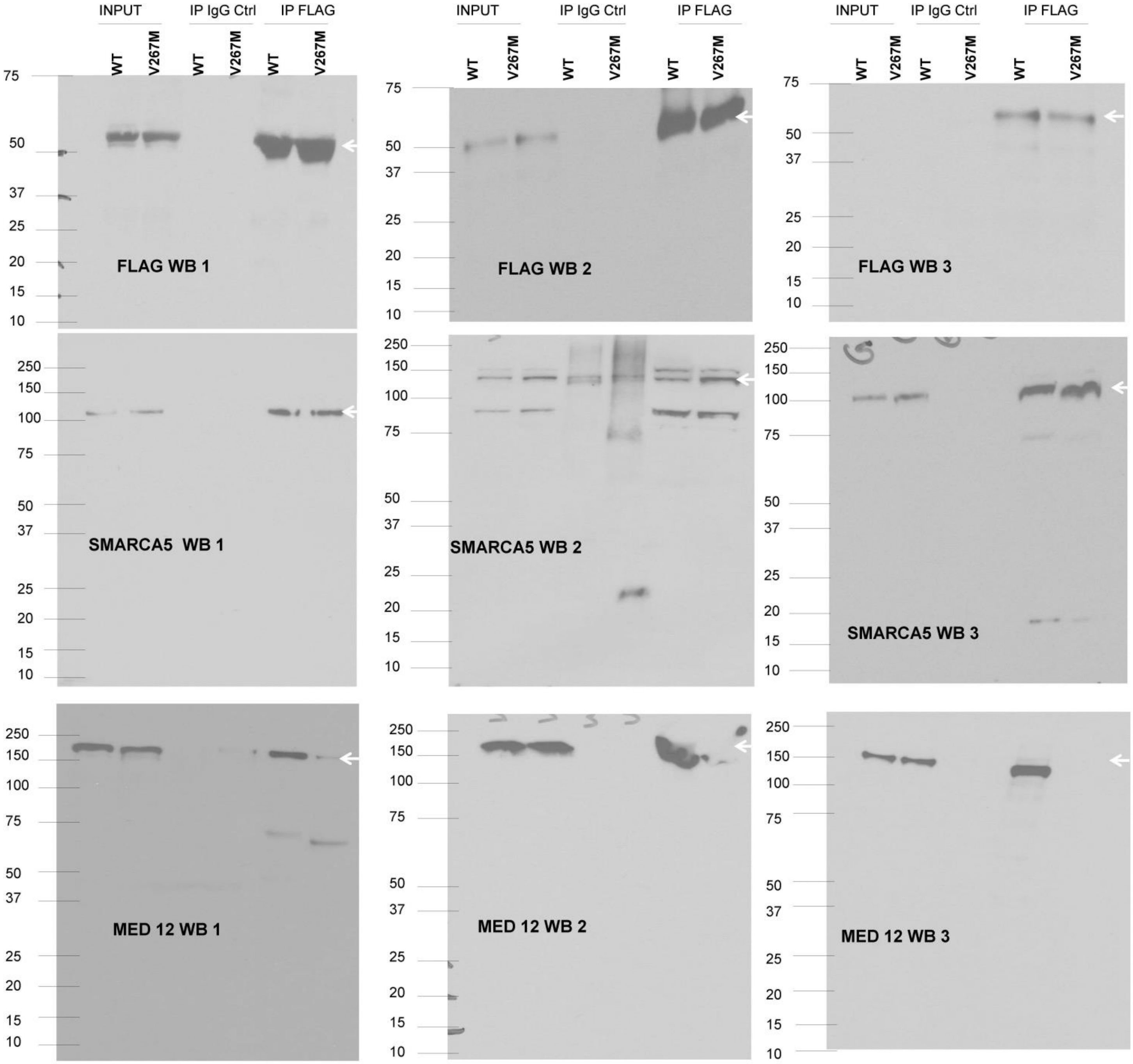

Figure S11. GATA4 V267M does not interact with MED12. *GATA4* deficient HCC cell line PLC was transfected with expression vectors for flag-GATA4 or flag-GATA4 V267M. Nuclear protein was subject to immunoprecipitation by control IgG and anti-flag antibody. Western blot for anti-flag, SMARCA5, and MED12. Three biological replicates.

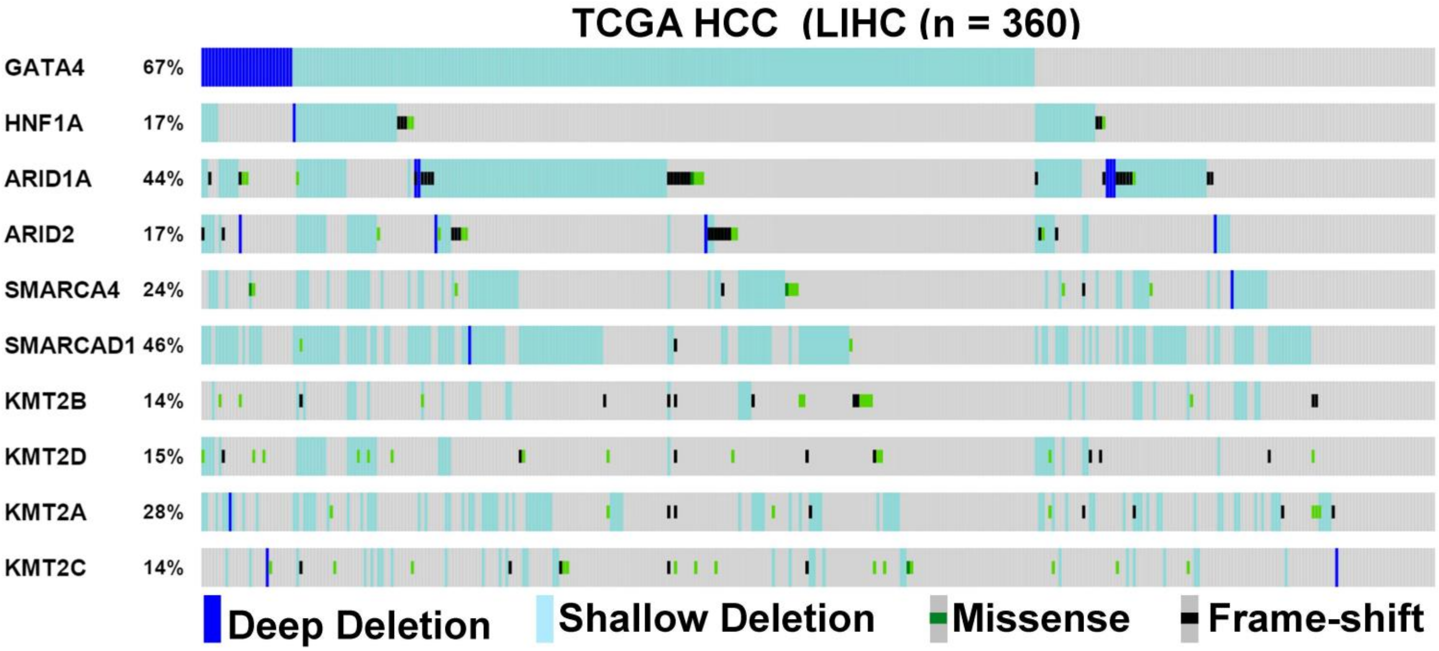

Figure S12. Genetic alterations in the genes for *GATA4* or its coactivators. GATA4 coactivators were identified by LCMS/MS analysis of its protein interactome. TCGA LIHC copy number and mutation data was analyzed using Cbioportal.

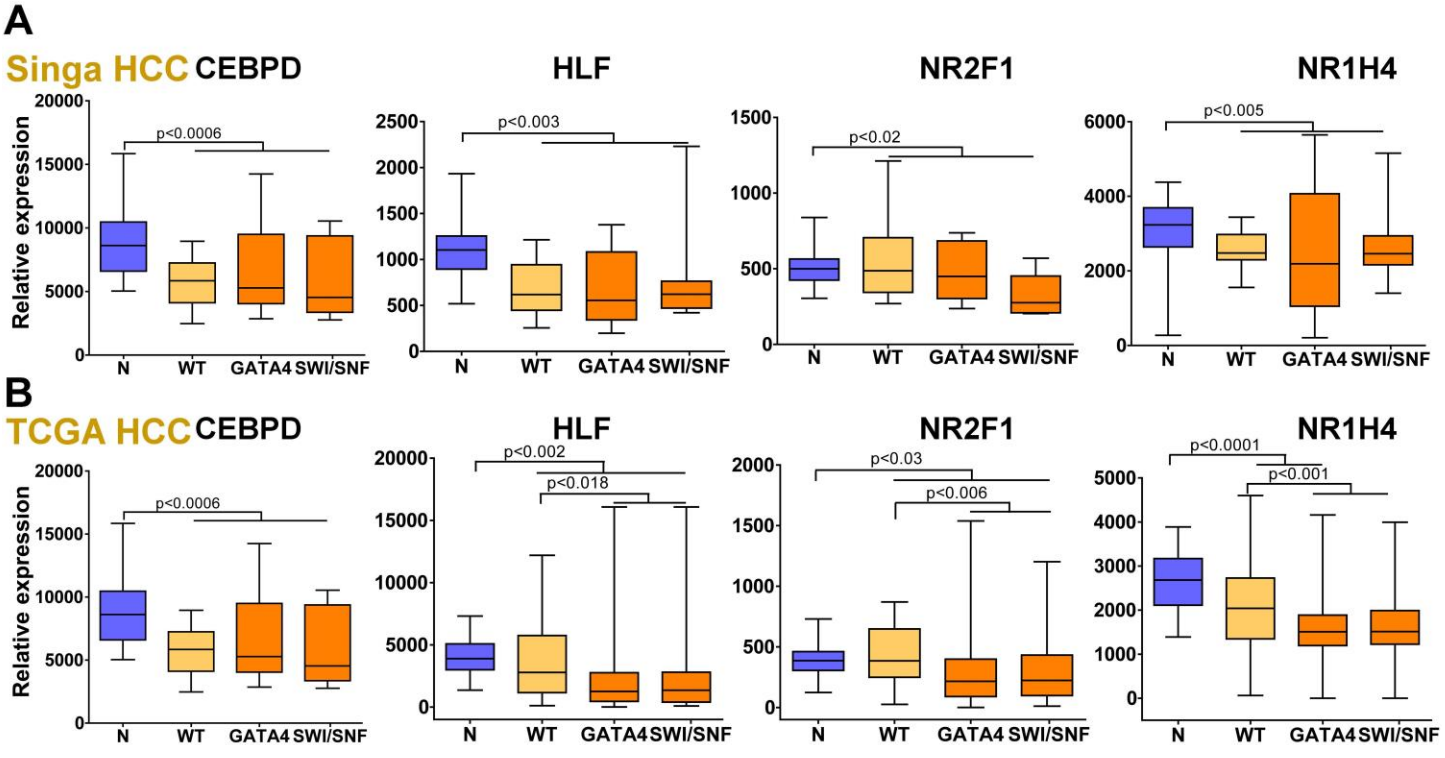

Figure S13: Hepatocyte late-differentiation transcription factors are less expressed in HCC with deletion/mutation of *GATA4* or in HCC with deletion/mutation of GATA4 coactivators (*ARID1A* and/or *SMARCAD1* and/or *ARID2* and/or *SMARCA4*). A) Expression in HCC of the Singapore series. Gene expression microarray. **B**) **Expression in HCC of the TCGA series**. RNA sequencing.

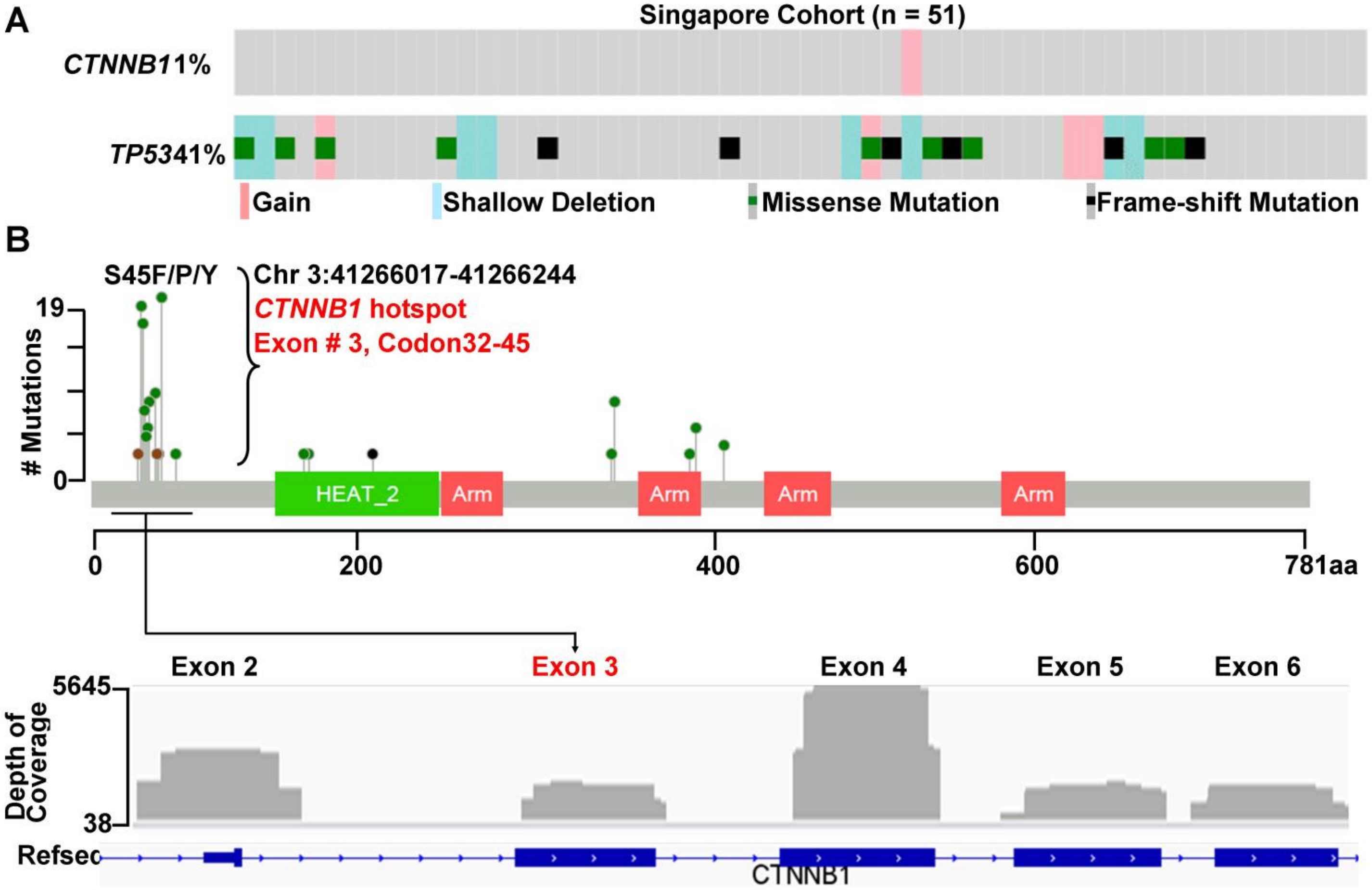

Figure S14. ***CTNNB1*** **and** ***TP53*** **mutations in Singapore HCC. A)** ***CTNNB1*** **mutations were not detected in this series of HCC while** ***TP53*** **was found frequently mutated. B) Analyses of** ***CTNNB1*** **mutations using TCGA LIHC data demonstrated that hotspot mutations accumulated at codon 32-45 corresponding to exon 3. This region and all the coding regions sequenced in the Singapore HCC cohort had good depth of coverage (range 38-5645) but no mutations were identified.**