Long-read single-cell sequencing reveals expressions of hypermutation clusters of isoforms in human liver cancer cells

Silvia Liu[1,2,3,+], Yan-Ping Yu[1,2,3], Bao-Guo Ren[1,2,3], Tuval Ben-Yehezkel[4], Caroline Obert[4], Mat Smith[4], Wenjia Wang[5], Alina Ostrowska[1,3], Alejandro Soto-Gutierrez[1,3], and Jian-Hua Luo[1,2,3,+]

Department of Pathology[1], High Throughput Genome Center[2], Pittsburgh Liver Research Center[3], Biostatistics[5], University of Pittsburgh, 3550 Terrace Street, Pittsburgh, PA 15261; Element Biosciences[4], Inc, 10055 Barnes Canyon Road, Suite 100, San Diego, CA 92121.

+ - Co-corresponding authors: Jian-Hua Luo, 3550 Terrace Street, Scaife S-728, Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261; E-mail: luoj@upmc.edu; Silvia Liu, S-406 Biomedical Science Tower, 203 Lothrop Street, Pittsburgh, PA 15261; E-mail: shl96@pitt.edu.

Abstract:

The protein diversity of mammalian cells is determined by arrays of isoforms from genes. Protein mutation is essential in species evolution and cancer development. Accurate Long-read transcriptome sequencing at single-cell level is required to decipher the spectrum of protein expressions in mammalian organisms. In this report, we developed a synthetic long-read single-cell sequencing technology based on LOOPseq technique. We applied this technology to analyze 447 transcriptomes of hepatocellular carcinoma (HCC) and benign liver from an individual. Through Uniform Manifold Approximation and Projection (UMAP) analysis, we identified a panel of mutation mRNA isoforms highly specific to HCC cells. The evolution pathways that led to the hyper-mutation clusters in single human leukocyte antigen (HLA) molecules were identified. Novel fusion transcripts were detected. The combination of gene expressions, fusion gene transcripts, and mutation gene expressions significantly improved the classification of liver cancer cells versus benign hepatocytes. In conclusion, LOOPseq single-cell technology may hold promise to provide a new level of precision analysis on the mammalian transcriptome.

Introduction:

Mammalian organisms are composed of numerous cells with multiple different roles. Individual cells are supported by a broad array of proteins with a variety of functions. While protein expression is dictated by the level of gene expression, the structure and function of the protein is largely determined by the isoforms of the mRNA of a given gene and are impacted by mutations or other structural alterations to the amino acids (Faustino and Cooper, 2003). To understand the role of each cell in an organism, broad spectrum mRNA isoform and mutational gene expression analyses at the single-cell level is necessary.

In the last 10 years, great strides in the field of long-read sequencing have enabled the quantification of mRNA isoforms in mammalian samples (Logsdon et al., 2020; Nakano et al., 2017). These sequencing technologies have been successfully employed to quantify mRNA isoforms from the bulk samples (Athanasopoulou et al., 2022). However, little progress has been made in developing a technology to analyze mutated mRNA expressions at the single-cell level. Among the long-read sequencing solutions, LoopSeq synthetic long-read sequencing technology has been shown to produce the lowest error rate (Liu et al., 2021) and thus may be most suited for the mutational isoform expression analyses. In this report, we developed a strategy to integrate Element Biosciences' LoopSeq intramolecular barcoding technique with 10x Genomics' cell barcoding scheme to create a single-cell long-read isoform analysis vehicle. To demonstrate the utility of this methodology, we analyzed the isoforms of over 440 transcriptomes from the cells originating from a hepatocellular carcinoma (HCC) patient. The results showed evolutionary patterns of single molecule mutational gene expression from benign hepatocytes to liver cancer cells.

Results:

Single-cell LoopSeq strategy:

The strategy of incorporating LoopSeq long-read technology with single-cell sequencing starts with utilizing the output of 10x Genomics' 3' single-cell assay. Approximately 200-300 cells from samples of both benign liver or HCC from a patient were encapsulated and unique molecular barcoded using a Gel Beads-in emulsion (GEM) system. The Gel Beads were dissolved, and any respectively associated cells were lysed prior to reverse transcription, template switching and transcript extension. The resulting 10x Adapter sequence contains a 16-base pair barcode, followed by a 12-base pair unique molecular identifier (UMI) and a 30-base pair poly(dT) sequence. Full length cDNA was then amplified from purified first-strand cDNA (Figure 1, blue box). A Loop adapter (containing 12-base pair unique molecular identifiers [LOOP UMI] and a 6 base pair sample index [SI]) was ligated to the 10x Genomics barcoded cDNA and subsequently enriched by exome probe sets that represent 19433 genes in the human genome. This was followed by amplification and the random intramolecular distribution of the LOOP UMIs throughout their respective cDNA molecules. The LOOP UMI distributed fragments were then subjected to short-read library preparation for sequencing (Figure 1, green box). Binned cDNA short-reads from individual LOOP UMIs were de novo assembled to generate consensus sequences of long mRNA transcripts.

To produce sufficient long-reads for single-cell analysis, 2.985 billion short-reads were sequenced across the benign liver cells, while 3.814 billion short-reads were evaluated from the HCC cells. The assembly of these short-read sequences resulted in 5.8 million long-read transcripts for the benign liver sample and 6 million for the HCC sample. The mapping of 10x Genomics cell barcodes resulted 447 valid single cell transcriptomes (162 from the benign liver sample and 285 from the HCC sample). There were an average number of 1186 unique genes per cell and 1331 unique isoforms per cell in the benign liver sample and 1266 unique genes and 1416 unique isoforms in the HCC sample.

Mutational gene and isoform expressions in the cells from benign liver and HCC samples

To identify the mutational gene expressions in each cell, exome sequencing was performed on benign liver, HCC, and gallbladder samples from the same individual. The sequencing results from the gallbladder were used as a germline reference to establish whether the structural variants found in the benign hepatocytes or the HCC samples were somatic mutations. The expression data were then limited to including only non-synonymous mutations detected in the exomes of the benign liver or the HCC samples. The expressions of these mutations were further filtered by the requiring of detecting the same mutation in a minimum of three different cells. Based on these criteria, a total of 2939 mutations were found to be expressed in the HCC and benign hepatocyte samples.

To investigate the role of mutational gene expression in HCC development, mutated gene expression levels were normalized to 'share' of the mutated transcripts relative to all the transcripts of a given gene, while mutational isoform expressions were normalized to 'share' of the mutation isoforms relative to all the transcripts of a given isoform. When mutational gene expressions were compared across all the transcriptomes, variations of mutated gene expressions were found (Supplemental Figures 1A and 1B). To remove potential non-contributing mutational gene expressions, genes with a normalized mutated gene expression standard deviation (SD) <0.4 across all the cells were removed. This resulted in 282 genes with mutated gene expression standard deviations ≥ 0.4. Uniform Manifold Approximation and Projection (UMAP) analysis was then applied to 447 transcriptomes from the HCC and benign liver samples based on these genes. As shown in Figure 2A, Supplemental Table 1, and Supplemental Figure 2A), many cells from the HCC sample clustered to a pole position relative to the cells from the benign liver, while other cells from the HCC sample moved in proximity to the cells from the benign liver. To investigate whether mutational isoform expressions contributed to the development of HCC, similar

removal of isoforms with mutated isoform expression standard deviations <0.4 was performed. The resulting 205 mutational isoforms were then assessed in cells from the benign liver and HCC samples for UMAP clustering. As shown in Figure 2B, Supplemental Table 2, and Supplemental Figure 2B, three distinct clusters emerged: one cluster was entirely composed of cells from the HCC sample, while the other two were mixtures of cells from the benign liver and the HCC samples, suggesting that some of these cells were in the transitional stage.

Mutations of genes involving antigen presentation dominated the mutation expression landscape

To examine the mutational characteristics of isoform expression in these transcriptomes, the clusters of the mutational isoform expressions in the UMAP were relabeled as A, B, and C groups (Figure 2C). A total of 3335 mutation isoforms were found in cluster A, while 2175 and 1783 mutation isoforms were found in clusters B and C, respectively. The overlapping of the mutations from these three groups, as pictured in a Venn diagram (Figure 2D), indicated that 1523 mutation isoforms were uniquely present in cluster A, while only 442 and 288 mutations were present in clusters B and C, respectively. These unique mutational isoforms were then combined and applied to the cluster analysis of 447 transcriptomes. The UMAP clustering generated four distinct clusters (Figure 2E), with 3 of the clusters composed entirely of cells from the HCC samples, distinctly separated from the 4th cluster, which was a mixture of cells from the benign liver and the HCC. We then limited the mutational isoforms to those that were expressed in at least 5 single-cell transcriptomes. This uncovered 113 mutations which met the established criteria (Supplemental Table 3). When UMAP analysis was performed based on these 113 mutation isoforms, eight distinctive clusters were resolved. All but one clusters contained cells co-migrated with cells of their sources (Figure 2F).

To examine the mutational isoform expressions in these clusters, 8 clusters were relabeled as A through H (Figure 3A). Among 113 mutation isoforms, the major histocompatibility complex (HLA) was the most prominent with 68 iterations (60.2%) (Supplemental Table 3, Figure 3B). Specifically, HLA-B NM_005514_2 mutations G283A and C44G were mostly present in cluster A. Cells in cluster B had mutations G572C, G539T, A527T, C463T, and G283A of HLA-B NM_005514_2, and cells in cluster C had mutations G379C and A167T of the same molecule. Cells in cluster D had up to 25 different mutations in HLA-DQB1 NM_002123. Cells in cluster E had partial mutations overlapped with those of clusters A and B. Surprisingly, cells in cluster F, which were from the benign liver, contained unique mutations in HLA-C NM_002117 molecule (T539G, C419T, G176A), while cells in cluster G, another cluster from the benign liver, had mutation G176A in the same molecule in addition to a mutation in ribosomal protein S9 (G525C, RPS9 NM_001013_5). Cluster H was a collection of cells with few mutations in the list. When the clusters were relabeled with mutations in HLA, all cells from clusters A through G were positive for some HLA mutations (Figure 3C). On the other hand, only 10 cells from group H were positive for the HLA mutation, suggesting these mutations in HLA molecules are highly cancer-specific (2.9×10^{-30}).

Evolution of mutations in HLA molecules

Long-read sequencing enabled us to identify multiple mutations in the same molecule. Indeed, most HLA molecules contained multiple mutations. A salient example of a multi-mutation molecule is HLA-DQB1, where up to 25 missense mutations were identified in a single molecule of NM_002123 (Figure 4A). We hypothesize that the collection of these mutations started from sporadic isolated mutations and accumulated over time in the development of HCC. To look for the origin of the mutation clusters, we searched for isolated mutation(s) that were the common denominators amongst the larger mutation clusters. As shown in Figure 4B, the largest cluster (49 cells) of mutations occurred in the HLA-DQB1 NM_002123 molecule. The mutation cluster contained 25 single nucleotide variants that caused 24

amino acids changes within the single molecule. There were several possible nucleotide mutation accumulation pathways that could have led to the formation of this hypermutation cluster. One of the pathways appears to have started at aa252 with the modification of Arginine to Histidine. The spread of the mutations would have been in one direction from 3' end to 5' end in a mostly contiguous fashion. However, the main pathway of accumulation of mutations is likely to have come from the mid-segment of the molecule since many cells containing subsets of mutations in this segment were detected, albeit they have larger hops in the accumulation process. Some isolated mutations, such as R252H, S214N+R199H, occurred in cells from the benign liver sample. They were associated with malignancy when more mutations were accumulated.

The stepwise accumulation of mutations in single molecules also occurred in HLA-B, HLA-C, and HLA-DRB1. In the HLA-B NM_005514_2 molecule, a total of 11 mutations were identified. The hypermutation cluster in the single protein started from 9 different isolated mutations. The main pathway of mutation accumulation appeared to start from the isolated mutations of W191S or A15G (Supplemental Figures 3A and 3B). These mutations expanded in a contiguous fashion and reached the peak at 8 mutations, as evidenced in 149 cells. One cell continued to expand its mutation repertoire up to 10 (supplemental figure 3B). For HLA-C NM_002117, 14 different missense mutations were identified (Supplemental Figures 3C and 3D). The major cascade of the mutation accumulation appeared to start from the isolated mutations of L180R or S140F. The expressions of the combination of L180R and S140F mutations accounted for most cells (n=222) that contained HLA-C mutants, followed by the combination of L180, S140F and R59Q (n=147). One cell accumulated 8 mutations in the single molecule (supplemental figure 3D). The accumulation of these mutations appeared non-contiguous. For HLA-DRB1 NM_002124, up to 5 different mutations in a single molecule were identified (Supplemental Figures 3E and 3F). All five isolated mutations were identified. The peak mutation accumulation (as seen in 81 cells) is the

combination of S133A, A103P, A102G, and T80R. Multiple pathways were detected that might lead to this pattern of mutation accumulation.

Fusion gene expression in single-cell level

Gene fusion is one of the hallmarks of human cancers. To identify fusion gene transcripts in the sample, we applied SQANTI (Tardaguila et al., 2018) annotation to the long-reads in order to identify transcripts that mapped to two different genes using the criteria described previously (Liu et al., 2021; Yu et al., 2019a; Yu et al., 2014a) and in the methods. To rule out potential artificial chimera, the fusion gene must be corroborated by at least two different cells. After multilayer screening, 21 fusion genes were identified, and 3 fusion genes were selected to validate experimentally. Among these fusion genes, ACTR2-EML4 was detected only in the cancer sample (Figure 5 and Supplemental Figure 4). ACTR2 is a major component of ARP2/3 complex and is responsible for cell shape and motility, while EML4 contains WD repeats that are essential for protein-protein interaction. The fusion retains most of the WD repeat domain from EML4 while removing most of the amino acid sequence from ACTR2. The fusion protein likely functions as a decoy interference protein that negatively impacts the microtubule organization activity of EML4. PDCD6 is an EF hand domain-containing protein and has calcium and magnesium binding activity. CCDC127 is coiled-coil domain containing transcription repressor. The PDCD6-CCDC127 fusion retained most of the coiled-coil domain from CCDC127 and a single EF hand domain from PDCD6. The signaling response of the fusion protein may be altered because of the new calcium-binding motif in the molecule. Finally, the FLG-FGG fusion is a unique chromosomal translocation product where the chromosome breakpoint is located in the exons. The fusion is a truncation of plasminogen. The removal of the C-terminus from plasminogen may lead to constitutive activation of its protease and to enhance blood coagulation and other cell signaling activities of plasminogen.

To investigate whether fusion transcripts had an impact on transcriptome clusters, we added these fusion genes to the mix of 113 mutational isoforms to perform UMAP analysis. As shown in Supplemental Figures 5 A-C, the cancer cell clusters A through D appeared to shift significantly to the left and underwent major reshuffling among the groups. On the other hand, clusters F through H remained in similar positions, while cluster E moved to the right, indicating that fusions impacted mostly the characteristics of cancer cells but had a very limited impact on benign hepatocytes.

Mutational gene expression and fusion transcript enhanced transcriptome clustering of benign hepatocytes and HCC.

Cell clustering and segregation can be determined by the differential expression of transcripts. Our mutational gene expression analyses suggested that some benign hepatocytes harbored mutations that resembled those of malignant cells. To reduce the complexity amongst the transcriptomes, we removed genes or isoforms that across all samples had with expression standard deviations less than 0.5, 0.8, 1.0, and 1.4, respectively. As shown in Supplemental Figures 6-7 and Supplemental Tables 4-11, the segregation of two groups of cells occurred when genes or isoforms had standard deviations >0.5. The segregation became more pronounced when the standard deviations were larger, with mostly malignant cells in one group and a mixture of malignant and benign hepatocytes in the other. Such cluster segregations were similarly found in either genes or isoform analyses. To examine the relationship between the isoforms and genes, the lists of isoforms and genes at each standard deviation were overlapped through Venn diagrams (Supplemental Figures 8A-D). Interestingly, gene lists included all the isoforms within the same range of standard deviation. To investigate the roles of gene expression alterations that were not accompanied with isoform expression changes, UMAP analyses were performed based on the non-overlapped genes. The results indicated a dramatic reduction of segregation of cells between benign liver and HCC. In contrast, gene-based clustering using genes that

showed both gene and isoform level changes had segregations between benign hepatocytes and HCC cells similar to those performed with the full lists, suggesting that the isoform alterations were the underlying causes that separated the cells between these two samples. Examination of the gene list (182, Supplemental Table 6) with standard deviations ≥1.0 showed a consistent down-expression of genes of apolipoprotein family, up-expression of genes of ribosomal protein family and HLA family in cells from the HCC sample, indicating that cancer cells were less hepatic differentiated, but more active in protein synthesis and immune evasion.

To investigate whether the mutation analysis improved the segregation between cells from the benign liver and HCC, UMAP analysis was performed using genes with standard deviations ≥ 1.0 (182 wild-type genes) and standard deviations ≥0.4 (282 mutated genes). The results showed that the combination of gene and mutational gene expressions generated three clusters: with 2 clusters comprised mostly cells from the cancer sample and 1 cluster of cells mostly from the benign liver (Supplemental Figures 9A-C). When the clusters were relabeled as A, B, and C. Cluster A (mostly benign hepatocyte group) had a gain of 7 cells from the benign liver sample and loss of 27 cells from the HCC sample in comparison with that of gene expression analysis alone (Supplemental Figure 7C), suggesting that the mutation analysis helped to reclassify some of the cells misassigned by gene expression analysis. To investigate whether fusion gene analysis would add value to the clustering of cells from HCC and benign liver, fusion genes were added to the UMAP analysis. The results showed that cluster B moved closer to the cancer cell cluster (cluster C, Figures 6A-C and Supplemental Table 12). Cluster A gained one cell from the benign liver sample and four cells from the HCC sample. Five cells from the benign liver were consistently classified as cancer by the cluster analysis. Further analysis showed that these cells had significant down-expression of genes of apolipoprotein family, up-expression of genes of ribosomal protein and HLA

families, and extensive mutations in HLA molecules, suggesting that they were probably the cancer cells embedded in the benign liver sample.

Discussion:

Long-read sequencing is essential to detect isoform expressions from genes. Synthetic long-read sequencing offers a valuable solution to analyze isoform transcripts of a gene because of its high accuracy, low-error rate, and quantification suitability. However, due to the low-yield nature of most synthetic long-read sequencing methodologies for transcriptome analysis, analyses are mostly limited to a few targeted genes (Gupta et al., 2018). To our knowledge, this is the first study to analyze broad-spectrum mutational isoform expression at the single-cell level using synthetic long-read sequencing technology. The technology described in this study may have broad utility in biology and medicine: it can be applied to quantify the diversity of isoform expression, resolve mutational gene expression and be used to discover novel fusion genes and new isoforms in any mammalian biological system. For medical research, the technology may help to determine which specific protein structure should be targeted by making the specific mutational isoform expression information available.

Currently, there is a lack of studies on multiple mutations in a single molecule or mutational gene expression at the single-cell level due to the absence of a reliable method. Our study suggested that mutations were mainly expressed in a specific isoform of a gene for a given cell. The lack of diversity of mutational isoform expression in a given cell may be due to the preferred splicing process of cells of different lineage or of different differentiation stages. Alternatively, mutations may have an impact on the RNA splicing process. Interestingly, mutational gene expression of antigen-presenting genes dominated the expressed mutation list from HCC cells. Most mutations occurred in the extracellular domain of the HLA molecules. For HLA-B and HLA-C, all three α-domains were mutated, and for HLA-

DQB1 and HLA-DRB1, both β-domains and the peptide binding motifs were impacted. These mutations may alter the interaction with T lymphocytes (Chan et al., 2018; Kondo et al., 2004). There was a broad spectrum of somatic mutations that affected the HLA gene, since both cytosolic and endocytic pathways of antigen presentation may be blocked (Arnaiz-Villena et al., 2022; Manoury et al., 2022). Interestingly, the expression levels of the mutated HLA molecules also increased in comparison with wild-type alleles from the benign hepatocytes. The hypermutations of these HLA molecules may shield cancer cells from being recognized and targeted by T lymphocytes and allow the cancer cells to evade the host immune surveillance.

The hypermutations in several HLA molecules are of interest because they probably did not happen overnight. Several isolated mutations were also detected in cells from the benign liver samples, suggesting that these mutations accumulated through a clonal progression fashion from a relatively benign background. In the process of malignant transformation, additional mutations in the HLA molecules were acquired due to the pressure from the cellular immune response. Malignant cells with few mutations in the HLA molecules may be destroyed by T lymphocytes, while those with newer mutations evaded the attack. Presumably, the cellular immune system adapted to the new mutations of these HLA molecules and resumed the response to the cancer cells. These cycles may continue to the extent that the mutations overwhelmed the cellular immune system. However, such hyper-mutation clusters may make cancer cells highly vulnerable to artificial immune intervention such as drug conjugated (Thomas et al., 2016) or radio-isotope (Bush, 2002; Guleria et al., 2017) labeled humanized antibody specific for these mutations or cancer vaccine since almost all of these mutations are in the extracellular domains. CRISPR-cas9 mediated genome targeting (Chen et al., 2017) at these mutation sites could be an option.

Methods:

Single-cell sample preparation: HCC samples and benign liver samples were freshly dissected from a patient who underwent liver transplantation. The procurement procedure was approved by the institution review board of University of Pittsburgh. The procedure was compliant with all regulations related to the protocol. The dissected tissues were minced by scalpel and digested with collagenase/protease solution (VitaCyte, 007-1010) until the tissue was fully digested. The digestion time for each preparation was in a range 45-60 min. The digested tissue was removed and immediately cooled with ice-cold Leibovitz's L-15 Medium (Invitrogen, 11415114) supplemented with 10% fetal bovine serum (Sigma, F4135). The single-cell suspension was verified under the microscope. The number of live cells was estimated by trypan blue staining using a hemacytometer.

10x Genomic single-cell and unique molecular index (UMI) barcoding: Approximately 200-300 cells from both HCC or benign liver samples were loaded onto the Chromium next GEM chip G, where the cells were encapsulated with oligo-dT coated Gel Beads and partitioning oil. The Gel Beads are subsequently dissolved, and the individual cells are lysed. Using the Chromium Next GEM Single Cell 3'Reagent Kit v3.1 from 10x Genomics, Inc., first strand synthesis was performed using the following thermal cycler parameters: With the lid set at 53 ^{o}C, incubate at 53 ^{o}C for 45 minutes, followed by 85 ^{o}C for 5 minutes. The first strand cDNA was then purified using the kit provided Dynabead clean-up mix. cDNA was then amplified using provided primers using the following program: with the lid set at 105 ^{o}C, 98°C for 3 min, then 11 cycles of 98°C for 15 seconds, 63°C for 20 seconds, 72°C for 1 minute, ended with 72°C for 1 min. Samples were pooled by group prior to long-read library preparation.

LoopSeq UMI ligation and transcriptome enrichment: The 10X Genomics barcoded cDNA were appended with a LoopSeq-specific adapter (containing the LoopSeq UMI) using a one-step barcoding

method. Four microliters of water, 11 µL of Barcoding Master Mix, and 5 µL of 10 ng of single cell cDNA were combined. The 20 µL reaction is incubated with a 100 °C heated lid at 95 °C for 3 minutes, 95 °C for 30 seconds, 60 °C for 45 seconds and 72 °C for 10 minutes. The LoopSeq adapted cDNA was then purified using 0.6x SPRI and resuspended in 20 µL of pre-warmed Hybridization Mix. The bead slurry was then enriched by a human core exome capture procedure (Twist Bioscience, CA). In brief, 5 µL of Buffer EB, 5 µL blocker solution, 6 µL LoopSeq adapter blocker, 4 µL biotinylated exome probe solution, and 30 µL hybridization enhancer were added to the bead slurry and incubated at 95 °C for 5 minutes followed by 70 °C for 16 hours. The hybridized cDNA was then captured by streptavidin beads following the protocol recommended by the manufacturer. Ten microliters of the probe captured, LoopSeq adapted cDNA was then combined with 5 µL of barcode oligo primers, and 35 µL of a PCR Barcoding cocktail for amplification using the following parameters: 95°C for 3 minutes, 12 cycles of 95°C for 30 seconds, 68°C for 45 seconds, 72°C for 2 minutes. The amplified captured cDNA underwent a 0.6x SPRI purification and was eluted in 30 µL of Buffer EB. This product is diluted with Buffer EB to adjust for a desired long read barcode complexity. Two microliters of each diluted product were independently combined with 18 µL of an Amp Mix S and an Amplification Additive Master mix and underwent thermocycling using the following parameters: with a 100 °C heated lid, amplify samples at 95 °C for three minutes, followed by 22 cycles of 95 °C for 30 seconds, 60 °C for 45 seconds, and 72 °C for 2 minutes. Ten microliters of each amplification reaction was then pooled underwent a 0.6x SPRI purification before elution in 40 µL of Buffer EB.

LOOP UMI distribution and library construction: Thirty microliters of the eluate was then combined with 10 µL distribution mix and 4 µL distribution enzyme and incubated at 20°C for 15 min. The reaction was then terminated by heating to 75°C for 5 minutes. The distributed UMIs were activated by incubating the reaction with 56 µL of activation mixture cocktail at 20°C for 2 hours and neutralized with

the addition of 6 µL of neutralization enzyme and heating at 37 °C for 15 minutes. The samples were then 0.8x SPRI purified to remove small undistributed UMI DNA. Thirty-five microliters of the LOOP UMI distributed cDNA was then fragmented with 15 µL of fragmentation enzyme master mix at 32°C for 5 minutes, followed by 65°C for 30 minutes. The fragmented LOOP UMI distributed cDNA was ligated with 40 µL of Ligation master and 10 µL of Ligation Enzyme at 20°C for 15 minutes. The ligated DNA was 0.6x SPRI purified, eluted in 20 µL of Buffer EB and amplified using the 25 µL of Index Master Mix and 5 µL of index primers in the following condition: 95°C for 3 minutes, then 12 cycles of 95°C for 30 seconds, 65°C for 45 seconds, and 72°C for 30 seconds. The amplified product undergoes a final 0.6x SPRI purification and 20 µL elution in Buffer EB. After the final short read library was quantified via qPCR and assessed for quality using a Agilent bioanalyzer 2100, the library cocktail was sequenced on an Illumina NovaSeq.

Taqman qRT-PCR assay for fusion genes: Total RNA was extracted using TRIzol (Invitrogen, CA) (Chen et al., 2015; Yu et al., 2019a; Yu et al., 2019b; Yu et al., 2014b; Zuo et al., 2017). Two micrograms of RNA were used to synthesize the first-strand cDNA with random hexamer primers and Superscript II™ (Invitrogen, CA). One microliter of each cDNA sample was used for TaqMan PCR (Eppendorf RealPlex Mastercycler and Applied Biosystems QuantStudio 3) with 50 heating cycles at 94°C for 30 seconds, 61°C for 30 seconds, and 72°C for 30 seconds using the following primer sequences: GAGTGATATCAGACACCGAGC/TTTCTGGGACTCCCTAGACCA and the following TaqMan probe: 5'-/56-FAM/AA GCTCTCT/ZEN/CCAACGGTTGGA/3IABkFQ/-3' for PDCD6-CCDC127, AGGAAGGTGGTGGTGTGCGA/TTGGGTGAAACTCCACAGCCA and the following TaqMan probe: 5'-/56-FAM/AACGGCACC/ZEN/GGGACAACAAAT/3IABkFQ/-3' for ACTR2-EML4, CCACAGGAAAGAAGTGTCAGTC/GTTATGGAGTTTTCAACATGGGG and the following TaqMan probe: 5'-/56-FAM/AAGCTCTCT/ZEN/CCAACGGTTGGA/3IABkFQ/-3' for PLG-FGG in an Eppendorf RealPlex™ thermocycler.

De novo assembly of long-read transcripts, short-read trimming and long-read alignment. The long-read transcripts were assembled using SPADES (Bankevich et al., 2012) from a python script with the following parameters:

command = spades.py -k 21,33,55,77,99,127 -t 1 --careful --sc --pe1-1 left.fq --pe1-2 right.fq --pe1-s unpaired.fq -o spades_output --disable-gzip-output. Short-read trimming was performed using Trimmomatics (Bolger et al., 2014) with the following parameters: command = ['java -jar ' + pipeline.prog_path + '/Trimmomatic-0.36/trimmomatic-0.36.jar PE -threads 32 -trimlog ' + trim_log_file + ' ','../' + pipeline.input_params['raw_file_R1'], '../' + pipeline.input_params['raw_file_R2'],' '.join(trim_output_files), 'ILLUMINACLIP:' + pipeline.prog_path + '/JAStrim.fa:2:40:14:3:true TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:36']. Long-read alignment through BLAST was performed from a python script with default parameters: blastn -db <reference database> -query contig_list.fa -perc_identity=<pct_id_threshold> -qcov_hsp_perc=<qcov_threshold> -max_target_seqs=<max_seqs> -num_threads=16 -outfmt=6 > mapping.blst.

Gene and isoform expression analysis on Loop-seq single-cell data: Paired HCC sample (HCC and benign liver) were compared by LoopSeq single-cell transcriptome sequencing. In total, six runs were performed on each library and were pooled together for analysis. LoopSeq long-reads were analyzed using SQANTI (Tardaguila et al., 2018) for gene and isoform annotation (human reference hg38). Based on the cell (10X) barcodes and molecule (Loop) barcodes from both long-reads and short reads, long-read molecules were able to be assigned to cells and unique molecules. UMIs were quantified at both gene and isoform levels based on the SQANTI annotation and cell/molecule assignment. Valid cells were defined as cells with more than 1000 long-read molecules. In total, 162 normal cells and 285 tumor cells

were used for the downstream analysis, with the highest number of long-read transcripts reaching 56745/cell for HCC, and 49476/cell for benign liver.

Single-cell expression data were integrated by R/Bioconductor package *Seurat (Hao et al., 2021)*. Expression standard deviations per gene and isoforms across all the cells were calculated. Genes or isoforms (SDGs and SDIs, respectively) with standard deviations above a certain threshold were defined. Cell clustering was performed based on the expression profile of these selected genes/isoforms and was visualized by UMAP algorithm (McInnes et al., 2018). Markers identifying each cluster were detected by comparing the cells in a specific cluster and all the other cells not in that cluster.

Mutation calling on whole exome sequencing data: Whole exome sequencing (WES) was performed on the same HCC patient with three libraries: HCC cells, benign liver cells, and normal gallbladder tissue. For each library, low-quality reads and adapter sequences were trimmed from the raw sequencing reads by the tool Trimmomatic (Bolger et al., 2014). After pre-processing, the surviving reads were aligned to human reference genome hg38 by Burrows-Wheeler Aligner. Picard (http://broadinstitute.github.io/picard) was employed to sort the aligned files and mark duplicates. Alternative (single nucleotide variants or SNPs) calling was then performed by SAMtools *mpileup* function (Li, 2011; Li et al., 2009), and somatic mutations on paired samples (normal gallbladder vs. benign liver pair, or normal gallbladder vs. HCC liver pair) were called by GATK *MuTect2* function (McKenna et al., 2010). Amino acids were annotated to those alternatives by SnfEff (Cingolani et al., 2012). The mutations of interest were selected by the following criteria: (1) The mutation must be non-synonymous or stop gain; (2) The mutation must be present in either HCC or benign liver samples but not in the normal gallbladder tissue. These mutations will serve as a validation set for the long-read single-cell transcriptome data analysis. All the pipelines were run by default parameter settings.

Mutation isoform analysis on LoopSeq single-cell data: Single-cell transcriptome long reads were aligned to human reference genome hg38 by long-read aligner Minimap2 (Li, 2018). Alternative (single nucleotide variants or SNPs) calling was then performed by SAMtools *mpileup* function (Li, 2011; Li et al., 2009). To avoid sequencing errors, RNA editing events, and non-tumor specific mutations, only mutations validated by the whole exome sequencing method were used. Based on the long-read cell barcode, the number of reference reads and alternative reads per mutation position and per valid single cell were quantified for the downstream analysis. The mutation rate was calculated by the number of alternative reads over the total reads (sum of reference and alternative reads). Based on the SQANTI annotation (Tardaguila et al., 2018) of the long-read, mutations were quantified both at gene and isoform levels.

The standard deviation of the mutation rate (per gene or isoform mutation) was calculated across all the valid cells. High variable mutations were defined as those with SDs greater than 0.4. These SD mutations were then used as features to perform cell clustering and UMAP visualization. Isoform-level mutation analysis resolved three clusters based on SD>=0.4 mutations. Unique mutations per cluster were defined by the mutations that exist in at least five cells of that cluster, but not in any of the cells in the other two clusters. A total of 113 unique mutations were found among the three clusters. Based on these unique mutations, additional clustering was performed, and 8 sparse clusters were detected and used to group the cells. HLA-related mutations were specifically examined and quantified across the 8 clusters. Evolution flowcharts were generated based on the progression of the mutation sites.

Fusion transcript detection on Loop-seq single-cell data

Fusion transcripts were called by two pipelines: (1) SQANTI (Tardaguila et al., 2018) performs the fusion annotation on the long-read sequencing data. (2) Based on the Minimap2 (Li, 2018) alignment and hg38 UCSC annotation file, fusions were called from the long-reads that were aligned to two genes. Based on all the fusion calling, the following filtering criteria were applied: (1) Eliminate the fusions where the head and tail genes were in cis-direction and were less than 40 kb apart; (2) Eliminate the fusions whose head genes have more than 2 tail partners in all the fusion callings; (3) Eliminate the fusions whose tail genes have more than 2 head gene partners in all the fusion callings; (4) Only keep those fusions whose joining points are located at the edge of the exons; (5) Fusions must be detected in at least 2 cells. These selected fusions and experimentally validated fusions were subsequently used for downstream analysis.

Integrative analysis to combine expression, mutation, and fusion data

High SD expression genes (or isoforms), high SD mutation genes (or isoforms), and selected fusion transcripts were integrated. UMAP (McInnes et al., 2018) cell visualization was applied, combining all three feature sets to perform the cell clustering. Data visualization was performed by R/Bioconductor package *ComplexHeatmap (Gu et al., 2016)* and *ggplot2 (Wickham, 2016)*.

Data availability

LOOP Single-cell transcriptome sequencing data has been deposited to Gene Expression Omnibus (GEO) database with access ID: GSE223743. Data can be accessed via the link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE223743. To review GEO accession GSE223743:

Go to

https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.ncbi.nlm.nih.gov%2Fgeo%2Fquery%2Facc.cgi%3Facc%3DGSE223743&data=05%7C01%7Cshl96%40pitt.edu%7Ca6bb7daab19b45e

650ba08daffa85a48%7C9ef9f489e0a04eeb87cc3a526112fd0d%7C1%7C0%7C638103395780381805%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=8URuG5ndTGy77iCwriHTaVuqL7OPcNJMdALaYdejN6A%3D&reserved=0, then enter token utoliguehxgfhin into the box.

Acknowledgement: We thank Songyang Zheng for technical support. This work is in part supported by grants from National Cancer institute (1R56CA229262-01 to JHL), National institute of digestive diseases and Kidney (P30- DK120531-01), National Institute of Health (UL1TR001857 and S10OD028483) and The University of Pittsburgh Clinical and Translational Science Institute.

Statement of conflict of interest: Tuval Ben-Yehezkel, Caroline Obert, and Mat Smith are employees of Element Biosciences, Inc. Silvia Liu, Yan-Ping Yu, Bao-Guo Ren, Wenjia Wang, Alina Ostrowska, Alejandro Soto-Gutierrez, and Jian-Hua Luo declare no conflict of interest.

References

Arnaiz-Villena, A., Suarez-Trujillo, F., Juarez, I., Rodríguez-Sainz, C., Palacio-Gruber, J., Vaquero-Yuste, C., Molina-Alejandre, M., Fernández-Cruz, E., and Martin-Villa, J.M. (2022). Evolution and molecular interactions of major histocompatibility complex (MHC)-G,-E and-F genes. Cellular and Molecular Life Sciences *79*, 464.
Athanasopoulou, K., Boti, M.A., Adamopoulos, P.G., Skourou, P.C., and Scorilas, A. (2022). Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. Life *12*, 30.
Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., and Prjibelski, A.D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology *19*, 455-477.
Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.
Bush, S. (2002). Monoclonal antibodies conjugated with radioisotopes for the treatment of non-Hodgkin's lymphoma. Paper presented at: Seminars in oncology nursing (Elsevier).
Chan, K.F., Gully, B.S., Gras, S., Beringer, D.X., Kjer-Nielsen, L., Cebon, J., McCluskey, J., Chen, W., and Rossjohn, J. (2018). Divergent T-cell receptor recognition modes of a HLA-I restricted extended tumour-associated peptide. Nature communications *9*, 1-13.
Chen, Z.-H., Yan, P.Y., Michalopoulos, G., Nelson, J., and Luo, J.-H. (2015). The DNA replication licensing factor miniature chromosome maintenance 7 is essential for RNA splicing of epidermal growth factor receptor, c-Met, and platelet-derived growth factor receptor. Journal of Biological Chemistry *290*, 1404-1411.
Chen, Z.-H., Yu, Y.P., Zuo, Z.-H., Nelson, J.B., Michalopoulos, G.K., Monga, S., Liu, S., Tseng, G., and Luo, J.-H. (2017). Targeting genomic rearrangements in tumor cells through Cas9-mediated insertion of a suicide gene. Nature biotechnology *35*, 543-550.
Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly *6*, 80-92.
Faustino, N.A., and Cooper, T.A. (2003). Pre-mRNA splicing and human disease. Genes & development *17*, 419-437.
Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847-2849.
Guleria, M., Das, T., Kumar, C., Amirdhanayagam, J., Sarma, H.D., and Banerjee, S. (2017). Preparation of clinical-scale 177 L u-R ituximab: O ptimization of protocols for conjugation, radiolabeling, and freeze-dried kit formulation. Journal of Labelled Compounds and Radiopharmaceuticals *60*, 234-241.
Gupta, I., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A.B., and Sloan, S.A. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nature biotechnology *36*, 1197-1202.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., and Zager, M. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587. e3529.
Kondo, E., Akatsuka, Y., Kuzushima, K., Tsujimura, K., Asakura, S., Tajima, K., Kagami, Y., Kodera, Y., Tanimoto, M., and Morishima, Y. (2004). Identification of novel CTL epitopes of CMV-pp65 presented by a variety of HLA alleles. Blood *103*, 630-638.
Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987-2993.
Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078-2079.
Liu, S., Wu, I., Yu, Y.-P., Balamotis, M., Ren, B., Ben Yehezkel, T., and Luo, J.-H. (2021). Targeted transcriptome analysis using synthetic long read sequencing uncovers isoform reprograming in the progression of colon cancer. Communications Biology *4*, 506.
Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nature Reviews Genetics *21*, 597-614.
Manoury, B., Maisonneuve, L., and Podsypanina, K. (2022). The role of endoplasmic reticulum stress in the MHC class I antigen presentation pathway of dendritic cells. Molecular Immunology *144*, 44-48.
McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.
Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., Shinzato, M., Minami, M., Nakanishi, T., and Teruya, K. (2017). Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Human Cell *30*, 149-161.
Tardaguila, M., De La Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., and Verheggen, K. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome research *28*, 396-411.
Thomas, A., Teicher, B.A., and Hassan, R. (2016). Antibody–drug conjugates for cancer therapy. The Lancet Oncology *17*, e254-e262.
Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).
Yu, Y.-P., Liu, P., Nelson, J., Hamilton, R.L., Bhargava, R., Michalopoulos, G., Chen, Q., Zhang, J., Ma, D., and Pennathur, A. (2019a). Identification of recurrent fusion genes across multiple cancer types. Scientific reports *9*, 1-9.
Yu, Y.-P., Tsung, A., Liu, S., Nalesnick, M., Geller, D., Michalopoulos, G., and Luo, J.-H. (2019b). Detection of fusion transcripts in the serum samples of patients with hepatocellular carcinoma. Oncotarget *10*, 3352.

Yu, Y.P., Ding, Y., Chen, Z., Liu, S., Michalopoulos, A., Chen, R., Gulzar, Z.G., Yang, B., Cieply, K.M., and Luvison, A. (2014a). Novel fusion transcripts associate with progressive prostate cancer. Am J Pathol *184*, 2840-2849.
Yu, Y.P., Michalopoulos, A., Ding, Y., Tseng, G., and Luo, J.-H. (2014b). High fidelity copy number analysis of formalin-fixed and paraffin-embedded tissues using Affymetrix Cytoscan HD chip. PloS one *9*, e92820.
Zuo, Z.H., Yu, Y.P., Martin, A., and Luo, J.H. (2017). Cellular stress response 1 down-regulates the expression of epidermal growth factor receptor and platelet-derived growth factor receptor through inactivation of splicing factor 3A3. Molecular carcinogenesis *56*, 315-324.

Figure 1: Schema of the workflow for the single-cell LoopSeq assay. 200-300 live cells per sample were co-partitioned with Gel Beads and subsequently lysed. The captured mRNAs were reverse-transcribed and barcoded using Chromium Next GEM 3' reagent 3.1 kit (10x Genomics). The cellular barcoded cDNAs were ligated with a LoopSeq Adapter (Element Biosciences) and enriched by human core exome capturing (Twist Biosciences). This was followed by amplification and intramolecular distribution of the LOOP UMI located on the LoopSeq Adapter. The LOOP UMI barcoded cDNAs were then fragmented and ligated with an adaptor to generate a short-read sequencing library before sequencing.

Figure 2. Mutation expression clustering of cells from HCC and its benign liver counterpart. (A) UMAP clustering of cells from the HCC and benign liver, based on mutational gene expressions shares with standard deviations ≥0.4. Red-cells are from HCC; Blue-cells are from benign liver. (B) UMAP clustering of cells from the HCC and benign liver based on mutational isoform expression shares with standard deviations ≥0.4. (C) Relabeling of clusters from (B) as 'A', 'B', and 'C'. (D) Venn diagram of mutational isoform expressions in cells from clusters A, B, and C. (E) UMAP clustering of cells from the HCC and benign liver based on the unique mutational isoform expressions from clusters A, B, and C. (F) UMAP clustering of cells from the HCC and benign liver based on the unique mutational isoform expression in at least 5 cells from clusters A, B, and C.

Figure 3. Mutations in HLA molecules dominated the landscape of HCC mutational isoform expressions. (A) The UMAP clusters from Figure 2F were relabeled as A through H groups as indicated. Cells from HCC and benign liver in each cluster are indicated. (B) Heat map of 113 mutational isoform expressions in the HCC and benign liver and clusters A through H. (C) Relabeling of UMAP clusters from (A) with cells expressing mutation HLA isoforms in triangles. Cells expressing mutation HLA isoforms in each cluster are indicated.

Figure 4. Evolution of mutations in HLA-DQB1 molecule. (A) Somatic mutations in single molecules of HLA-DQB1 NM_002123. The position of mutation is indicated at the bottom of the graph. The mutation is numerically numbered from C-terminus to N-terminus. The numbers of cells expressing these mutation isoforms from each cluster or sample are indicated in the right panel. Close circle-mutation codon; Open circle-wild type codon. Open rectangle-double single-nucleotide mutation in the same codon. (B) Pathway flow chart of mutation accumulation in single molecules of HLA-DQB1. The area of the circle is proportional to the accumulated number of mutations in a molecule. The scale on the left indicates the number of mutations in a single molecule but is not mathematically scaled. The arrow indicates the potential pathways of mutation accumulation in these molecules. The number of white text indicates specific mutation(s) in a molecule. The number of red text indicates the number of cells expressing the mutation(s).

Figure 5. Fusion gene expression validation in HCC sample. Left panel: ACTR2-EML4 fusion. Top: Chromosome organization of EML4 and ACTR2 exons. The directions of transcriptions are indicated. 2nd from the top: Exon representations in ACTR2-EML4 fusion transcript, EML4 NM001145076.3, and ACTR2 NM001005386.3. Middle: Chromogram of Sanger sequencing. The segments for ACTR2 and EML4 are indicated. Bottom: Protein domain and motif organizations of EML4, ACTR2, and ACTR2-EML4 fusion. **Middle panel:** PDCD6-CCDC127 fusion. Top: Chromosome organization of CCDC127 and PDCD6 exons. The directions of transcriptions are indicated. 2nd from the top: Exon representations in PDCD6-CCDC127 fusion transcript, CCDC127 NM145265.3, and PDCD6 NM013232.4. Middle: Chromogram of Sanger sequencing. The segments for PDCD6 and CCDC127 are indicated. Bottom: Protein domain and motif organizations of CCDC127, PDCD6, and PDCD6-CCDC127 fusion. **Right panel:** PLG-FGG fusion. Top: Chromosome organization of PLG and FGG exons. The directions of transcriptions are indicated. 2nd from

the top: Exon representations in PLG-FGG fusion transcript, PLG NM000301.5, and FGG NM000509.6. Middle: Chromogram of Sanger sequencing. The segments for PLG and FGG are indicated. Bottom: Protein domain and motif organizations of PLG, FGG, and PLG-FGG fusion. The open-reading frame of FGG was eliminated due to frameshift in PLG-FGG fusion. Unrelated four additional amino acids were added to the truncated N-terminus of PLG.

Figure 6. UMAP clustering of cells from the HCC and benign liver based on the combination of normal gene expression, mutational gene expression share, and fusion gene expression share. (A) UMAP clustering of cells from HCC and benign liver samples based on 182 gene expressions with a standard deviation ≥1, 282 mutational gene expression shares with standard deviations ≥0.4, and 20 fusion gene expression shares of any standard deviation. (B) Relabeling of clusters from (A) as clusters 'A', 'B', and 'C'. The number of cells from HCC and benign liver in each cluster is indicated. (C) Heat map of 182 gene expressions, 282 mutational gene expression shares, and 20 fusion genes expression shares for cells from clusters 'A', 'B', and 'C'. Cells from the HCC and benign liver are indicated.

Supplemental Figure 1. Mutation expression standard deviations. (A) Mutational gene expressions share standard deviation across all transcriptomes. (B) Mutational isoform expression share standard deviation across all transcriptomes.

Supplemental Figure 2. Heat maps of mutational gene expression. (A) Heat map of mutational gene expression share with standard deviations ≥0.4. Cells from the HCC and benign liver are indicated. (B) Heat map of mutational isoform expression shares with standard deviation ≥0.4. Cells from the HCC and benign liver are indicated.

Supplemental Figure 3. Evolution of mutations in HLA-B, HLA-C, and HLA-DRB1 molecules. (A) Somatic mutations in single molecules of HLA-B NM_005514_2. The position of the mutations is indicated at the bottom of the graph. Mutations are numerically numbered from C-terminus to N-terminus. The numbers of cells expressing these mutation transcripts from each cluster or sample are indicated in the right panel. Close circle-mutation codon; Open circle-wild type codon. (B) Pathway flow chart of mutation accumulation in single molecules of HLA-B. The area of the circle is proportional to the accumulated number of mutations in a molecule. The scale on the left indicates the number of mutations in a single molecule but is not mathematically scaled. The arrow indicates the potential pathways of mutation accumulation in these molecules. The number in white text indicates the specific mutation(s) in a molecule. The number in red text indicates the number of cells expressing the mutation(s). (C) Somatic mutations in single molecules of HLA-C NM_002117. The position of the mutation is indicated at the bottom of the graph. The mutation is numerically numbered from C-terminus to N-terminus. The numbers of cells expressing these mutation transcripts from each cluster or sample are indicated in the right panel. Close circle-mutation codon; Open circle-wild type codon; Open rectangle-double single-nucleotide mutation in the same codon. (D) Pathway flow chart of mutation

accumulation in single molecules of HLA-C. The area of the circle is proportional to the accumulated number of mutations in a molecule. The scale on the left indicates the number of mutations in a single molecule but is not mathematically scaled. The arrow indicates the potential pathways of mutation accumulation in these molecules. The number in white text indicates specific mutation(s) in a molecule. The number in red text indicates the number of cells expressing the mutation(s). (E) Somatic mutations in single molecules of HLA-DRB1 NM_002124. The position of mutation is indicated at the bottom of the graph. The mutation is numerically numbered from C-terminus to N-terminus. The numbers of cells expressing these mutation transcripts from each cluster or sample are indicated in the right panel. Close circle-mutation codon; Open circle-wild type codon. (F) Pathway flow chart of mutation accumulation in single molecules of HLA-DRB1. The area of the circle is proportional to the accumulated number of mutations in a molecule. The scale on the left indicated the number of mutations in a single molecule but is not mathematically scaled. The arrow indicates the potential pathways of mutation accumulation in these molecules. The number in white text indicates specific mutation(s) in a molecule. The number in red text indicates the number of cells expressing the mutation(s).

Supplemental Figure 4. Taqman RT-PCR of fusion transcripts in HCC and benign liver samples. Top panel: Taqman RT-PCR results of PDCD6-CCDC127, ACTR2-EML4, PLG-FGG, and β-actin from the benign liver sample. **Bottom panel:** Taqman RT-PCR results of PDCD6-CCDC127, ACTR2-EML4, PLG-FGG, and β-actin from the HCC sample.

Supplemental Figure 5. The impact of fusion gene expressions on the cell clustering generated by mutational isoform expressions. (A) UMAP cluster analysis of cells from HCC and benign liver based on 113 mutational isoform expressions and 20 fusion gene expressions. (B) Relabeling of clusters from (A)

as clusters A through H. (C) Heat map of mutational isoform expressions and fusion gene expression shares for clusters A through H. The cells from HCC and benign liver were indicated.

Supplemental Figure 6. UMAP cluster analyses of cells from the HCC and benign liver based on gene expressions with different standard deviation cutoffs. (A) UMAP clustering of cells based on gene expression with standard deviations at least 0.5, 0.8, 1.0, or 1.4. The numbers of genes employed in the UMAP analysis are indicated. Blue dot-cell from the benign liver; Red dot-cell from HCC. (B) UMAP clustering of cells based on isoform expressions with standard deviations at least 0.5, 0.8, 1.0, or 1.4. The numbers of genes employed in the UMAP analysis are indicated. Blue dot-cell from the benign liver; Red dot-cell from HCC.

Supplemental Figure 7. Segregation of cells between HCC and benign liver samples. (A) Relabeling of clusters as 'A' and 'B' based on gene expressions with a standard deviation ≥0.5. Left panel: UMAP clustering. The numbers of cells from the HCC and benign liver samples in each cluster are indicated. Right panel: Heatmap of clusters A and B. Cells from HCC and benign liver are indicated. (B) Relabeling of clusters as 'A' and 'B' based on gene expressions with a standard deviation ≥0.8. Left panel: UMAP clustering. The numbers of cells from HCC and benign liver in each cluster are indicated. Right panel: Heatmap of clusters A and B. Cells from HCC and benign liver are indicated. (C) Relabeling of clusters as 'A' and 'B' based on gene expressions with a standard deviation of ≥1.0. Left panel: UMAP clustering. The numbers of cells from HCC and benign liver in each cluster are indicated. Right panel: Heatmap of clusters A and B. Cells from HCC and benign liver are indicated. (D) Relabeling of clusters as 'A', 'B', and 'C' based on gene expressions with a standard deviation ≥1.4. Left panel: UMAP clustering. The numbers of cells from HCC and benign liver in each cluster are indicated. Right panel: Heatmap of clusters A, B, and C. Cells from HCC and benign liver are indicated. (E) Relabeling of clusters as 'A' and 'B' based on

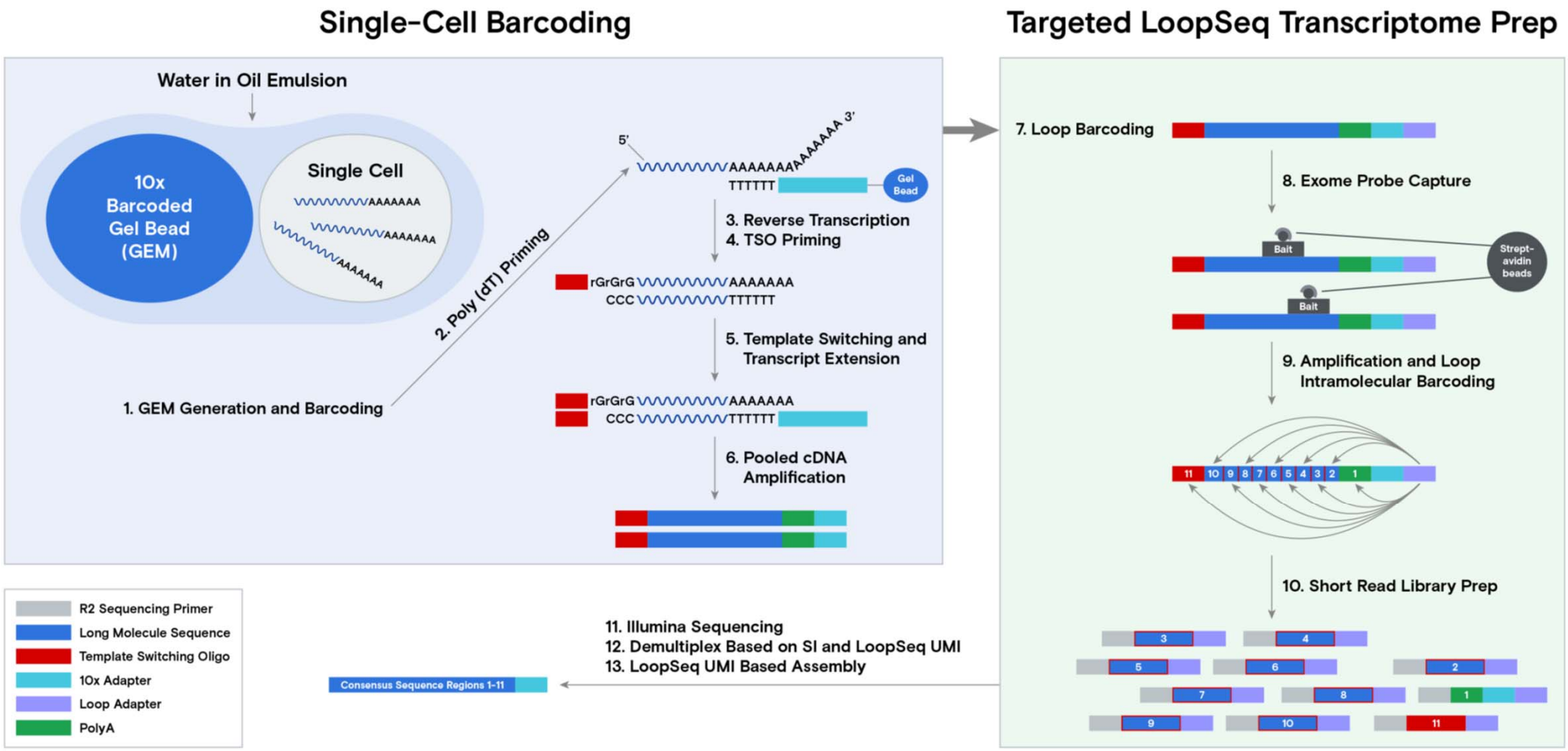
isoform expressions with a standard deviation ≥0.5. Left panel: UMAP clustering. The numbers of cells from HCC and benign liver in each cluster are indicated. Right panel: Heatmap of clusters A and B. Cells from HCC and benign liver are indicated. (F) Relabeling of clusters as 'A' and 'B' based on isoform expressions with a standard deviation ≥0.8. Left panel: UMAP clustering. The numbers of cells from HCC and benign liver in each cluster are indicated. Right panel: Heatmap of clusters A and B. Cells from HCC and benign liver are indicated. (G) Relabeling of clusters as 'A' and 'B' based on isoform expressions with a standard deviation ≥1.0. Left panel: UMAP clustering. The numbers of cells from HCC and benign liver in each cluster are indicated. Right panel: Heatmap of clusters A and B. Cells from HCC and benign liver are indicated. (H) Relabeling of clusters as 'A', 'B', and 'C' based on isoform expressions with a standard deviation ≥1.4. Left panel: UMAP clustering. The numbers of cells from HCC and benign liver in each cluster were indicated. Right panel: Heatmap of clusters A, B, and C. Cells from HCC and benign liver are indicated.

Supplemental Figure 8. Genes with isoform expression alterations played key roles in segregating cells between the HCC and benign liver samples. (A) The role of isoform expressions in segregating cells between the HCC and benign liver when the standard deviation was ≥0.5. Left panel: Venn diagram between gene expressions and isoform expressions with standard deviations ≥0.5. Middle panel: UMAP clustering with genes not overlapping with isoforms. Right panel: UMAP clustering with genes overlapping with isoforms. (B) The role of isoform expressions in segregating cells between the HCC and benign liver when the standard deviation was ≥ 0.8. Left panel: Venn diagram between gene expressions and isoform expressions with standard deviations ≥0.8. Middle panel: UMAP clustering with genes not overlapping with isoforms. Right panel: UMAP clustering with genes overlapping with isoforms. (C) The role of isoform expression in segregating cells between the HCC and benign liver when the standard deviation ≥1.0. Left panel: Venn diagram between gene expressions and isoform expressions with

standard deviations ≥ 1.0. Middle panel: UMAP clustering with genes not overlapping with isoforms. Right panel: UMAP clustering with genes overlapping with isoforms. (D) The role of isoform expression in segregating cells between the HCC and benign liver when the standard deviation was ≥1.4. Left panel: Venn diagram between gene expressions and isoform expressions with a standard deviation ≥1.4. Middle panel: UMAP clustering with genes not overlapping with isoforms. Right panel: UMAP clustering with genes overlapping with isoforms.

Supplemental Figure 9: UMAP clustering of cells from HCC and benign liver based on the combination of normal gene expressions and mutational gene expression shares. (A) UMAP clustering of cells from the HCC and benign liver samples based on 182 gene expressions with standard deviations ≥1.0 and 282 mutational gene expression shares with standard deviations ≥0.4, (B) Relabeling of clusters from (A) as clusters A, B, and C. The number of cells from the HCC and benign liver in each cluster are indicated. (C) Heat map of 182 gene expressions and 282 shares for cells from clusters A, B, and C. Cells from the HCC and benign liver mutational gene expressions are indicated.

Figure 1


Single-Cell Barcoding
Targeted LoopSeq Transcriptome Prep
Water in Oil Emulsion
10x Barcoded Gel Bead (GEM)
Single Cell
AAAAAAA
AAAAAAA
AAAAAAA
1. GEM Generation and Barcoding
2. Poly (dT) Priming
5'
3'
AAAAAAAAAAAAAAA
TTTTTT
Gel Bead
3. Reverse Transcription
4. TSO Priming
rGrGrG
CCC
AAAAAAA
TTTTTT
5. Template Switching and Transcript Extension
6. Pooled cDNA Amplification
7. Loop Barcoding
8. Exome Probe Capture
Bait
Bait
Strept-avidin beads
9. Amplification and Loop Intramolecular Barcoding
11 10 9 8 7 6 5 4 3 2 1
10. Short Read Library Prep
3 4 5 6 2 7 8 1 9 10 11
11. Illumina Sequencing
12. Demultiplex Based on SI and LoopSeq UMI
13. LoopSeq UMI Based Assembly
Consensus Sequence Regions 1-11
R2 Sequencing Primer
Long Molecule Sequence
Template Switching Oligo
10x Adapter
Loop Adapter
PolyA

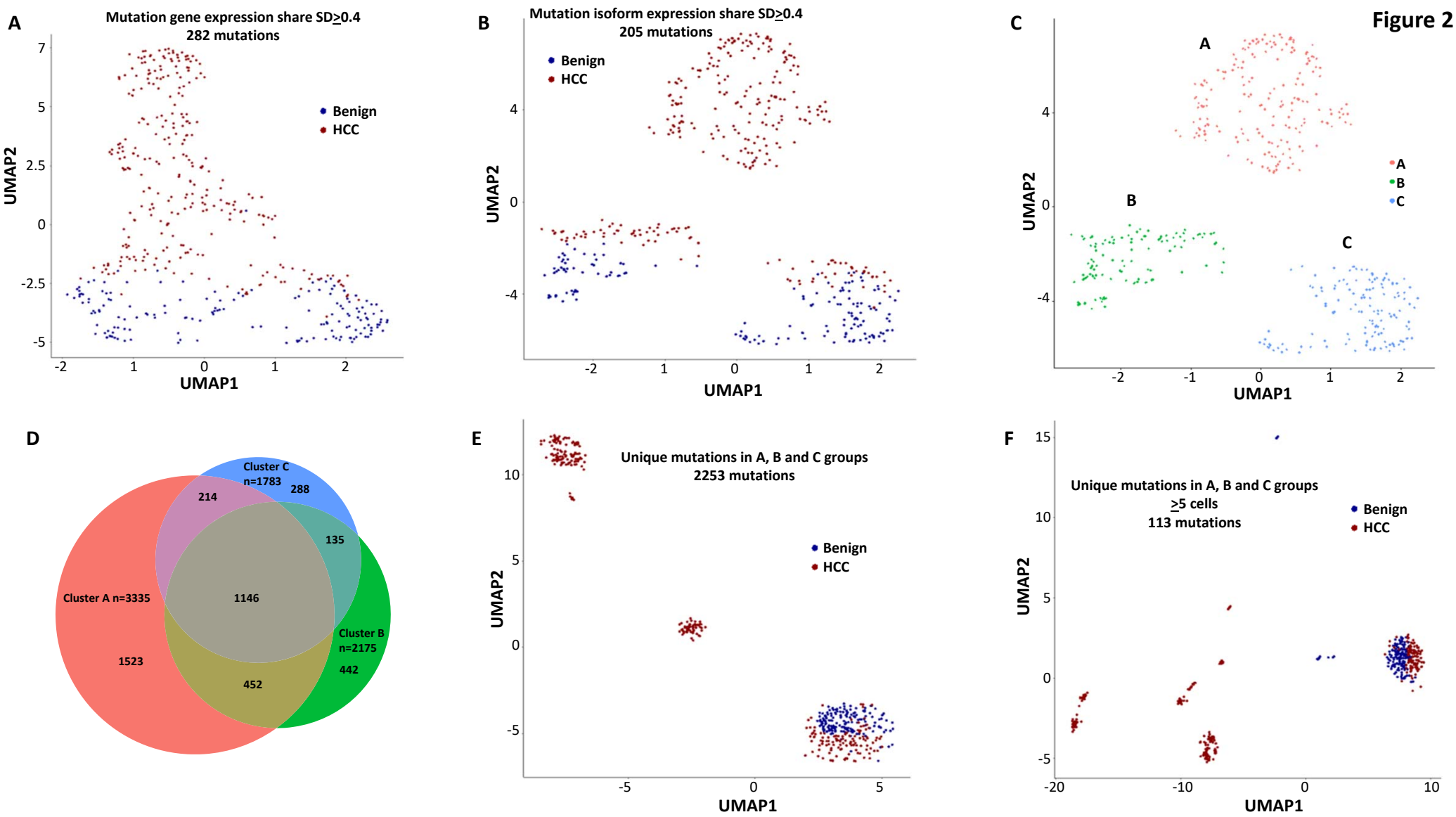
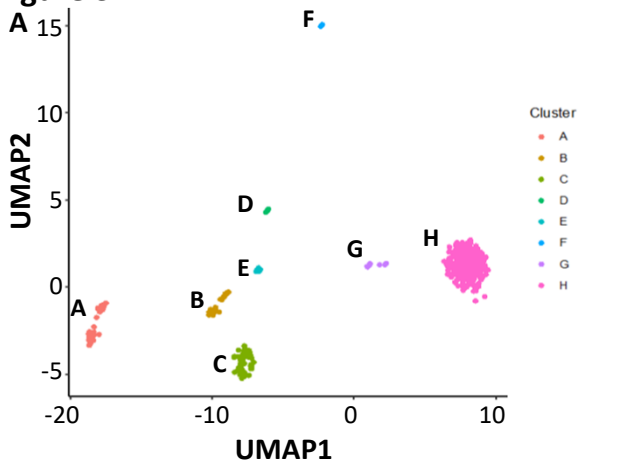
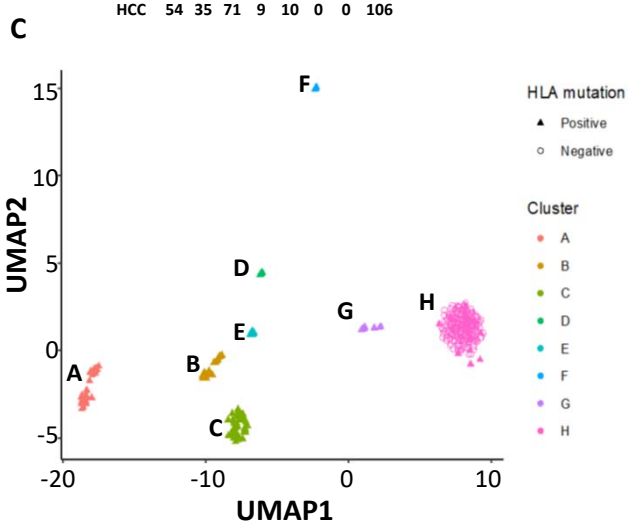


Figure 2
A
Mutation gene expression share SD≥0.4
282 mutations
Benign
HCC
UMAP2
UMAP1
B
Mutation isoform expression share SD≥0.4
205 mutations
Benign
HCC
UMAP2
UMAP1
C
A
B
C
UMAP2
UMAP1
D
Cluster C
n=1783
288
214
135
Cluster A n=3335
1146
Cluster B
n=2175
1523
442
452
E
Unique mutations in A, B and C groups
2253 mutations
Benign
HCC
UMAP2
UMAP1
F
Unique mutations in A, B and C groups
≥5 cells
113 mutations
Benign
HCC
UMAP2
UMAP1

Figure 3



	A	B	C	D	E	F	G	H
Benign	0	0	0	0	0	5	15	142
HCC	54	35	71	9	10	0	0	106



	A	B	C	D	E	F	G	H
Negative	0	0	0	0	0	0	0	238
Positive	54	35	71	9	10	5	15	10

	Benign	HCC
Negative	142	96
Positive	20	189

p-value = 2.9×10^{-30}

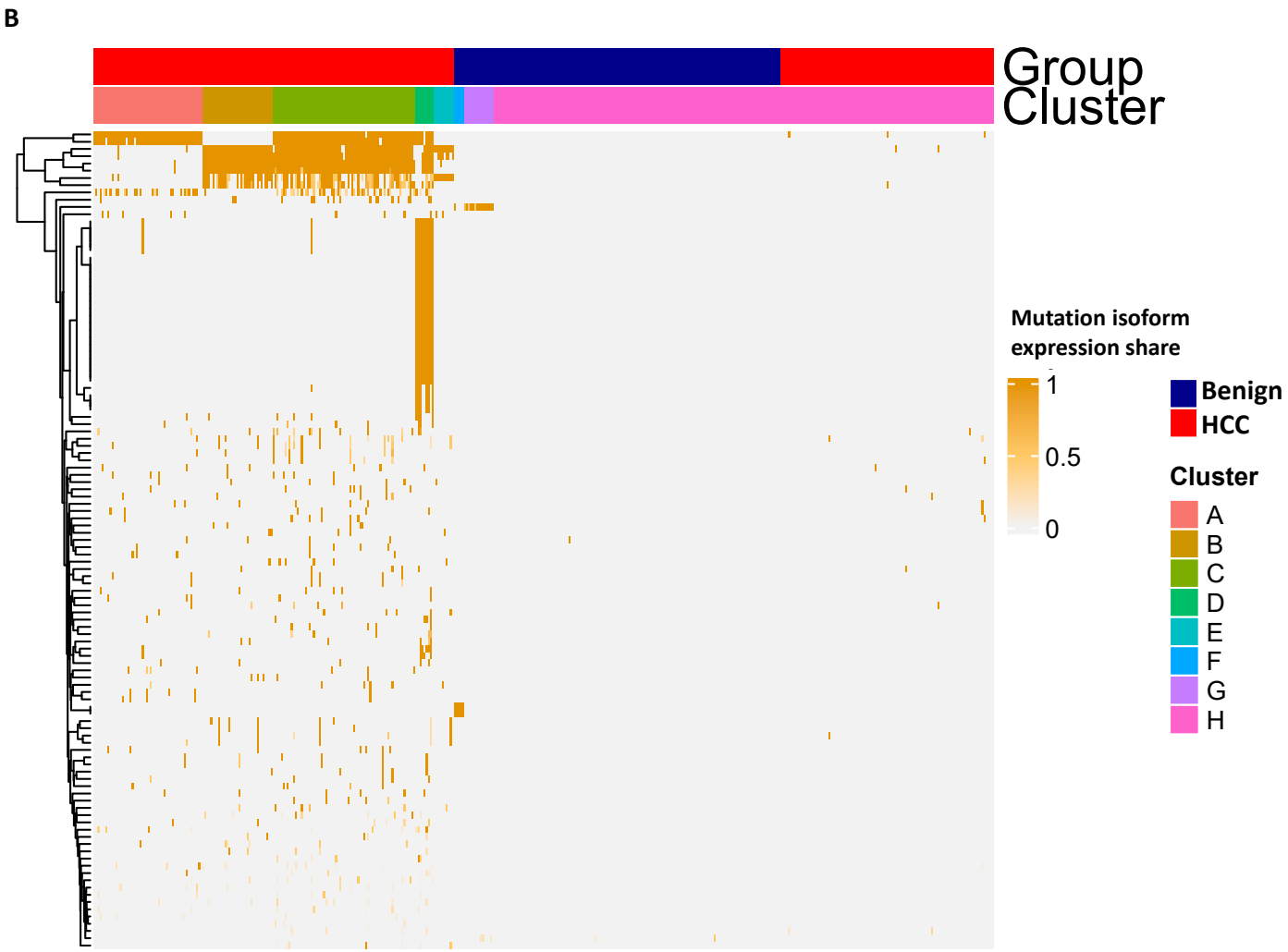
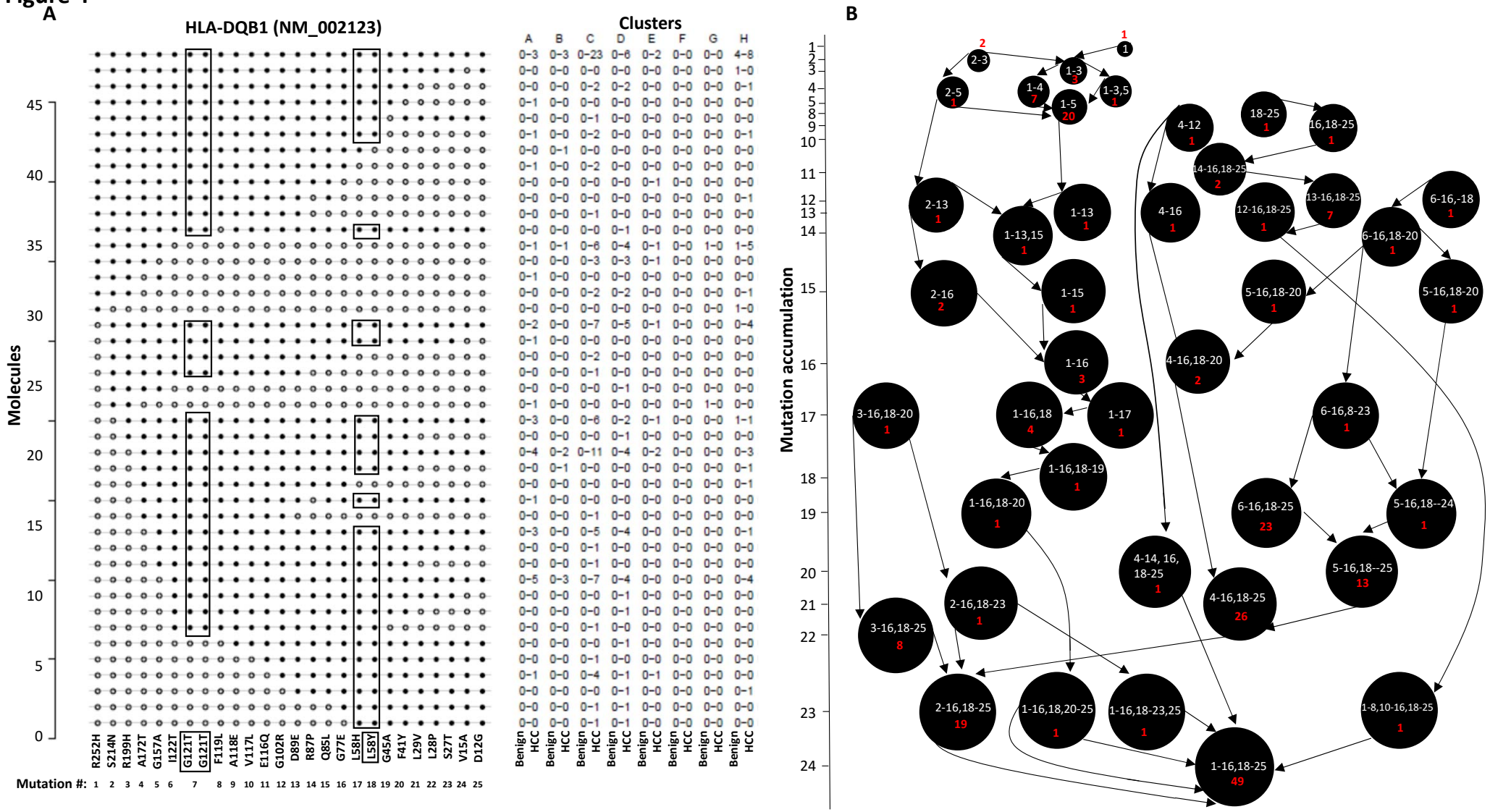


Figure 4


A
HLA-DQB1 (NM_002123)
Clusters
A B C D E F G H
Molecules
Mutation #:
R252H S214N R199H A172T G157A I122T G121T G121T F119L A118E V117L E116Q G102R D89E R87P Q85L G77E L58H L58Y G45A F41Y L29V L28P S27T V15A D12G
Benign HCC
B
Mutation accumulation
1
2-3
1-3
1-3,5
2-5
1-4
1-5
4-12
18-25
16,18-25
14-16,18-25
2-13
1-13
4-16
12-16,18-25
13-16,18-25
6-16,-18
1-13,15
6-16,18-20
2-16
1-15
5-16,18-20
5-16,18-20
1-16
4-16,18-20
3-16,18-20
1-16,18
1-17
6-16,8-23
1-16,18-19
1-16,18-20
6-16,18-25
5-16,18--24
4-14, 16, 18-25
5-16,18--25
2-16,18-23
4-16,18-25
3-16,18-25
2-16,18-25
1-16,18,20-25
1-16,18-23,25
1-8,10-16,18-25
1-16,18-25

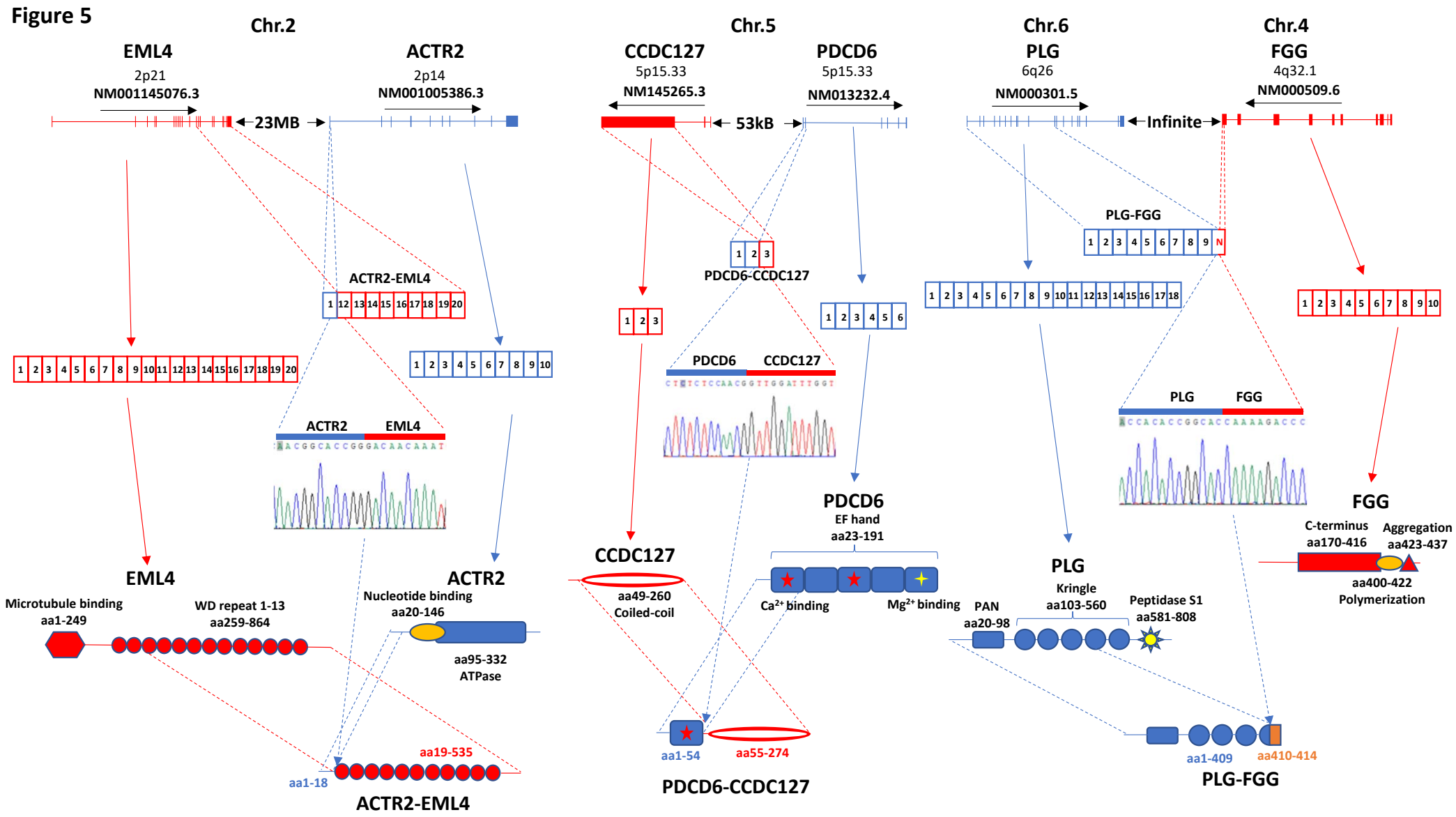
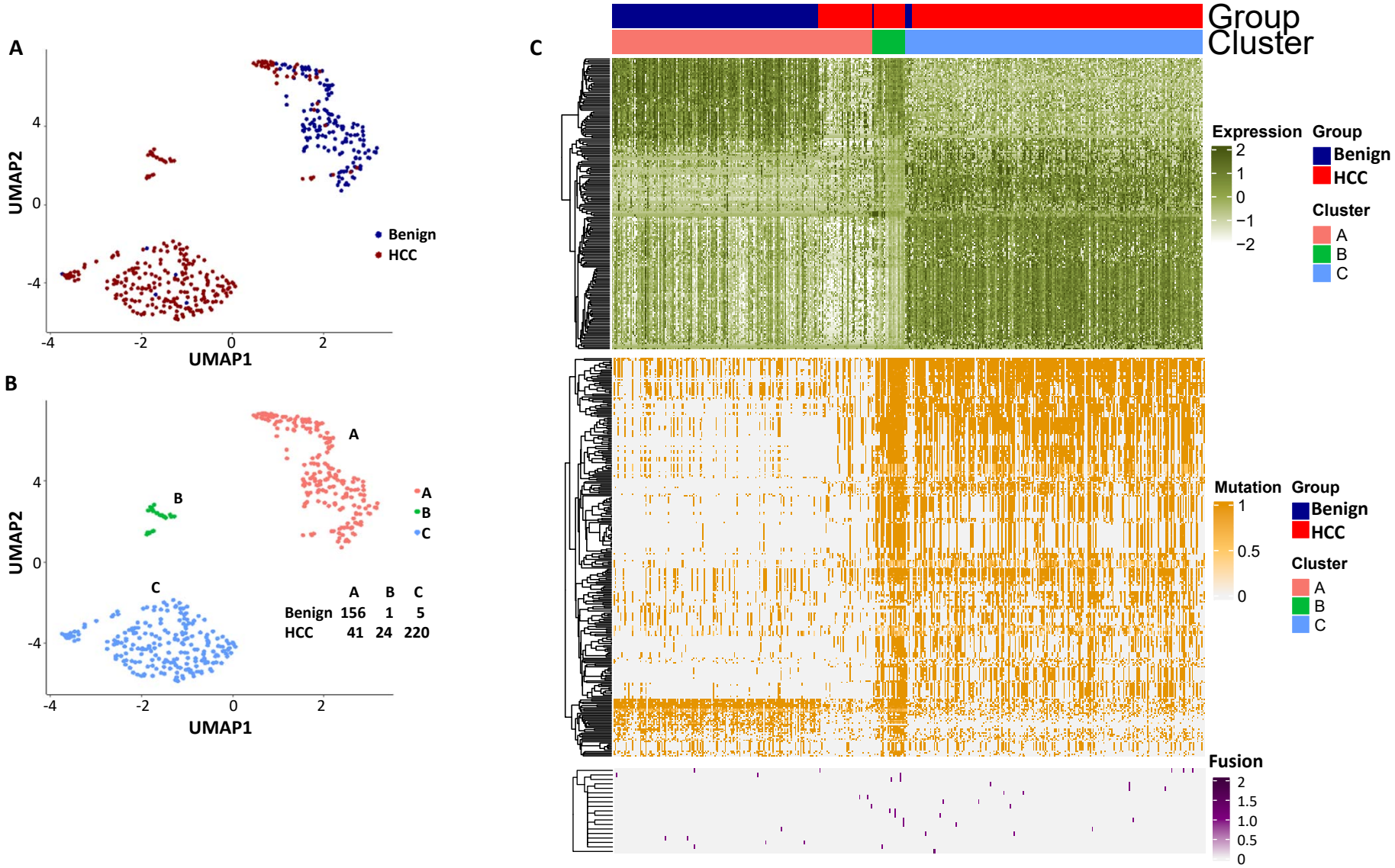


Figure 5
Chr.2
EML4
2p21
NM001145076.3
ACTR2
2p14
NM001005386.3
23MB
ACTR2-EML4
ACTR2
EML4
Chr.5
CCDC127
5p15.33
NM145265.3
PDCD6
5p15.33
NM013232.4
53kB
PDCD6-CCDC127
PDCD6
CCDC127
Chr.6
PLG
6q26
NM000301.5
Chr.4
FGG
4q32.1
NM000509.6
Infinite
PLG-FGG
PLG
FGG
EML4
Microtubule binding
aa1-249
WD repeat 1-13
aa259-864
ACTR2
Nucleotide binding
aa20-146
aa95-332
ATPase
aa19-535
aa1-18
ACTR2-EML4
CCDC127
aa49-260
Coiled-coil
PDCD6
EF hand
aa23-191
Ca2+ binding
Mg2+ binding
aa1-54
aa55-274
PDCD6-CCDC127
PLG
Kringle
aa103-560
PAN
aa20-98
Peptidase S1
aa581-808
aa1-409
aa410-414
PLG-FGG
FGG
C-terminus
aa170-416
Aggregation
aa423-437
aa400-422
Polymerization

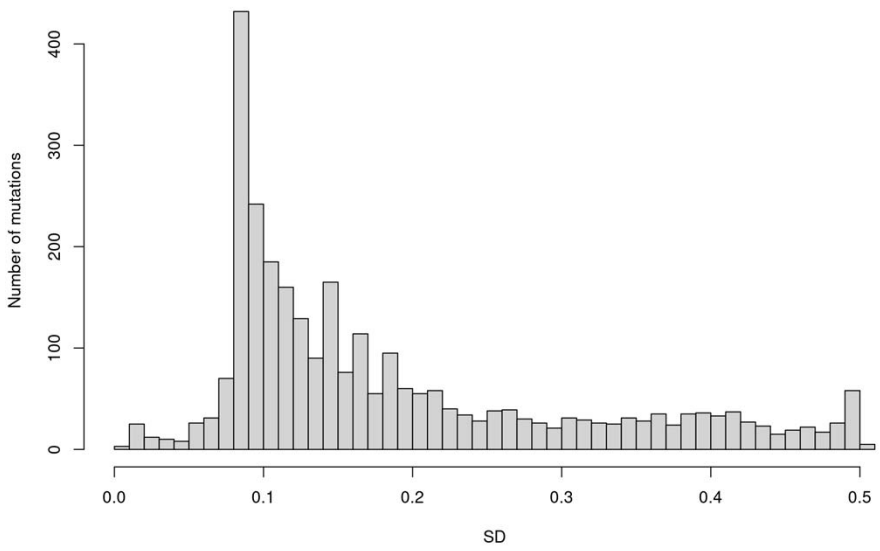
Figure 6


A
UMAP2
UMAP1
Benign
HCC
B
A
B
C
A B C
Benign 156 1 5
HCC 41 24 220
C
Group
Cluster
Expression
Group
Benign
HCC
Cluster
A
B
C
Mutation
Fusion

Supplemental figure 1

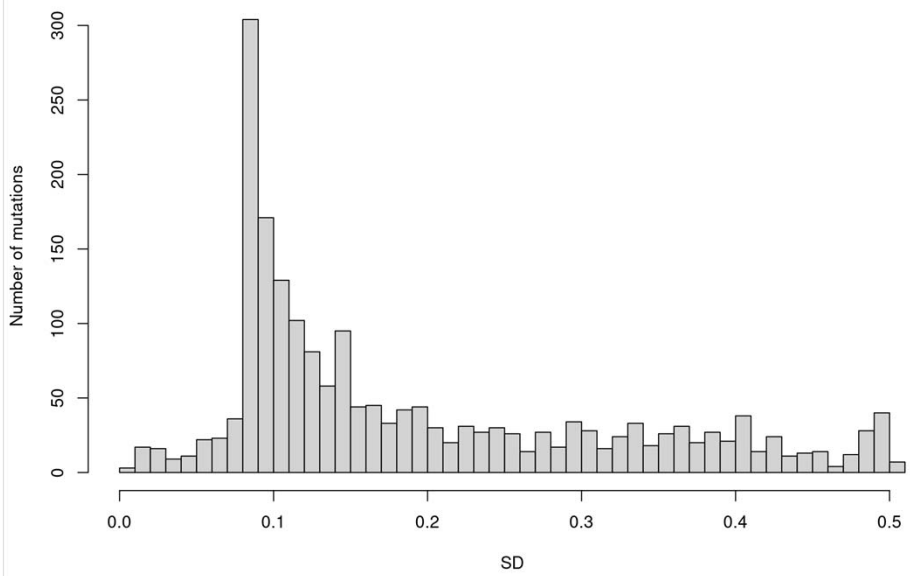
A

Mutation gene expression share SD distribution



B

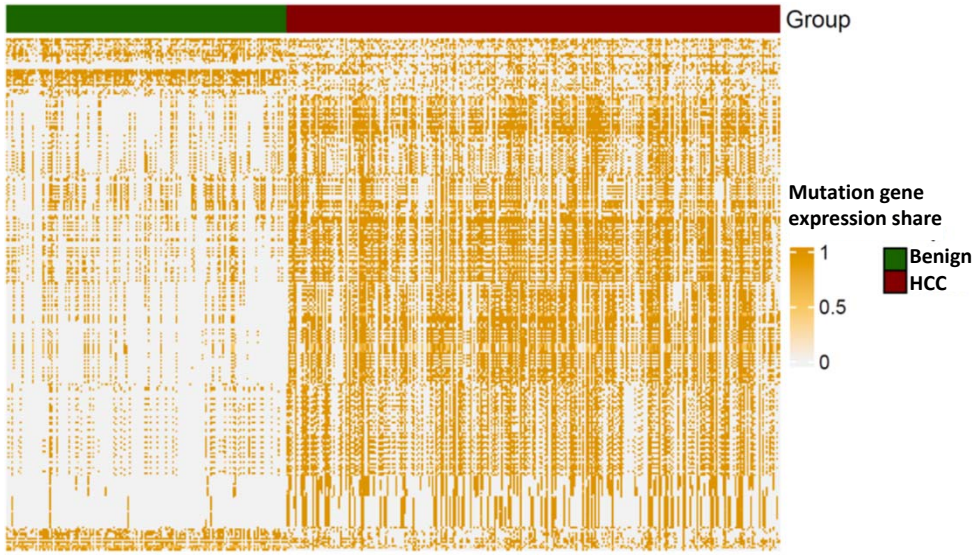
Mutation isoform expression share SD distribution



Supplemental figure 2

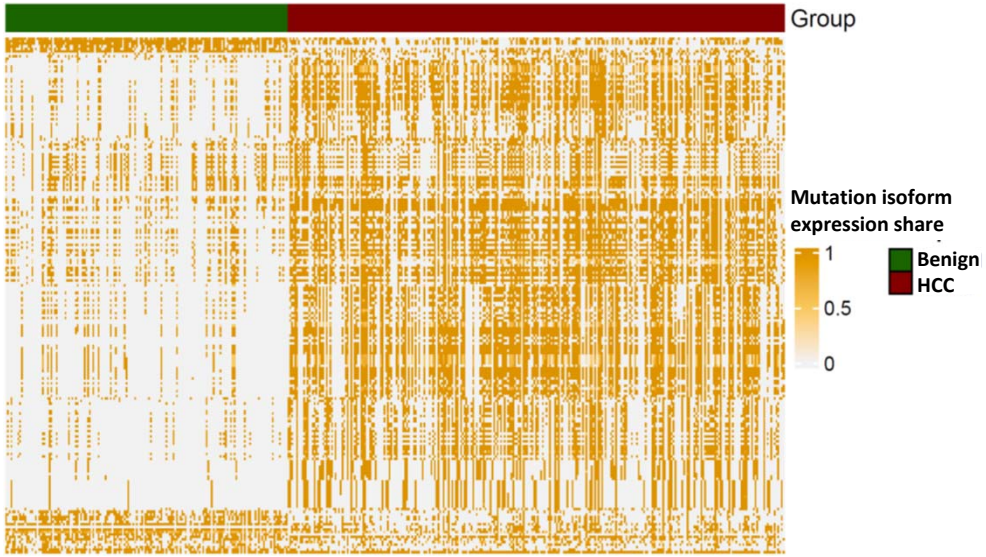
A

Mutation gene expression share SD≥0.4

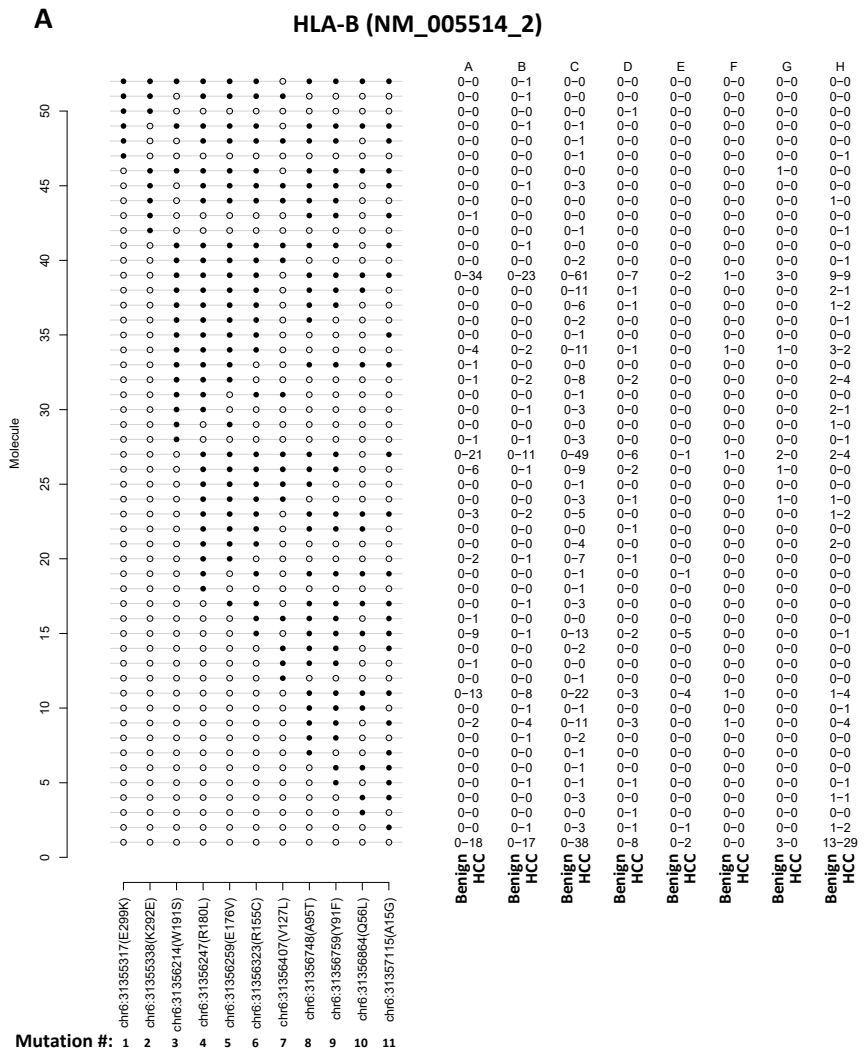


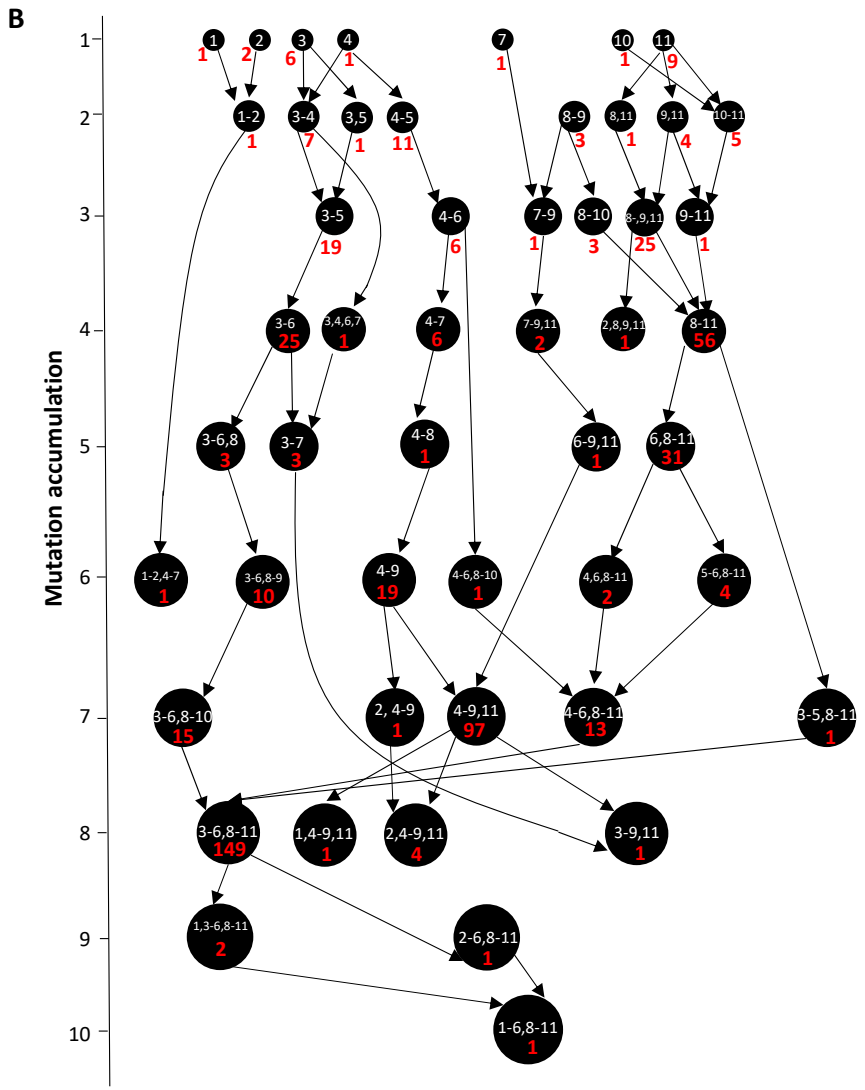
B

Mutation isoform expression share SD≥0.4



Supplemental figure 3

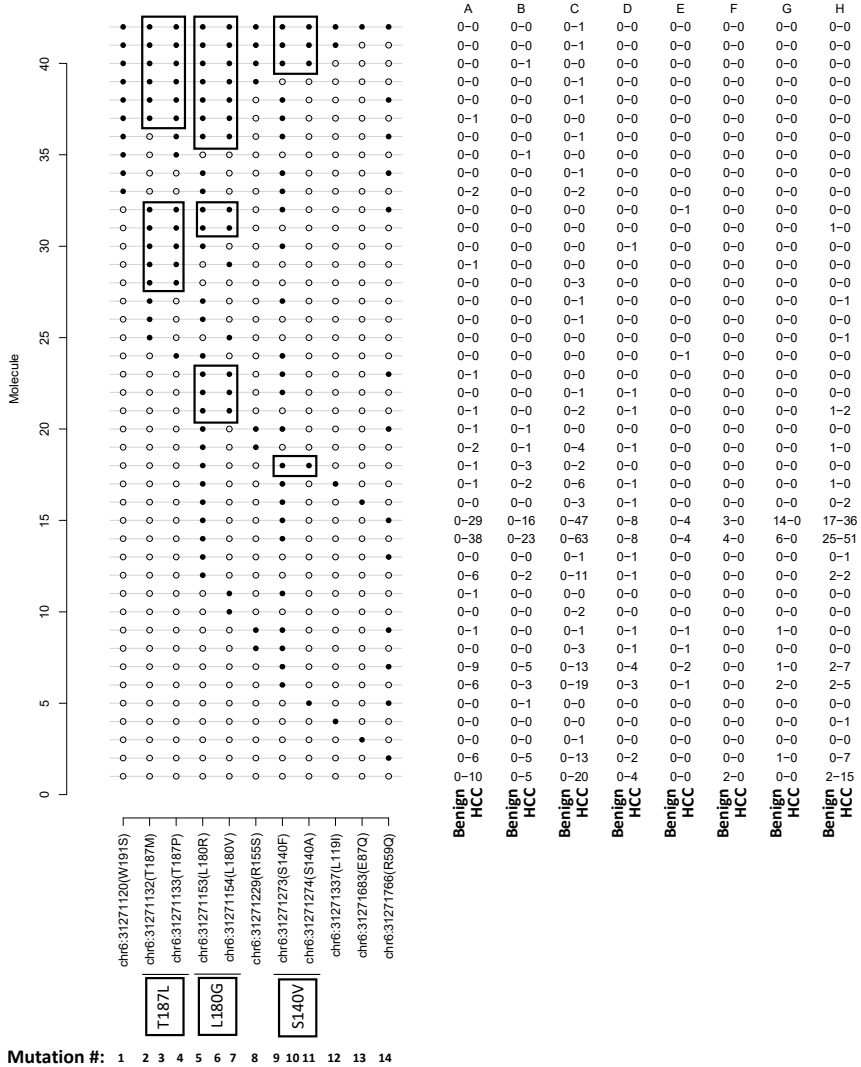




Supplemental figure 3

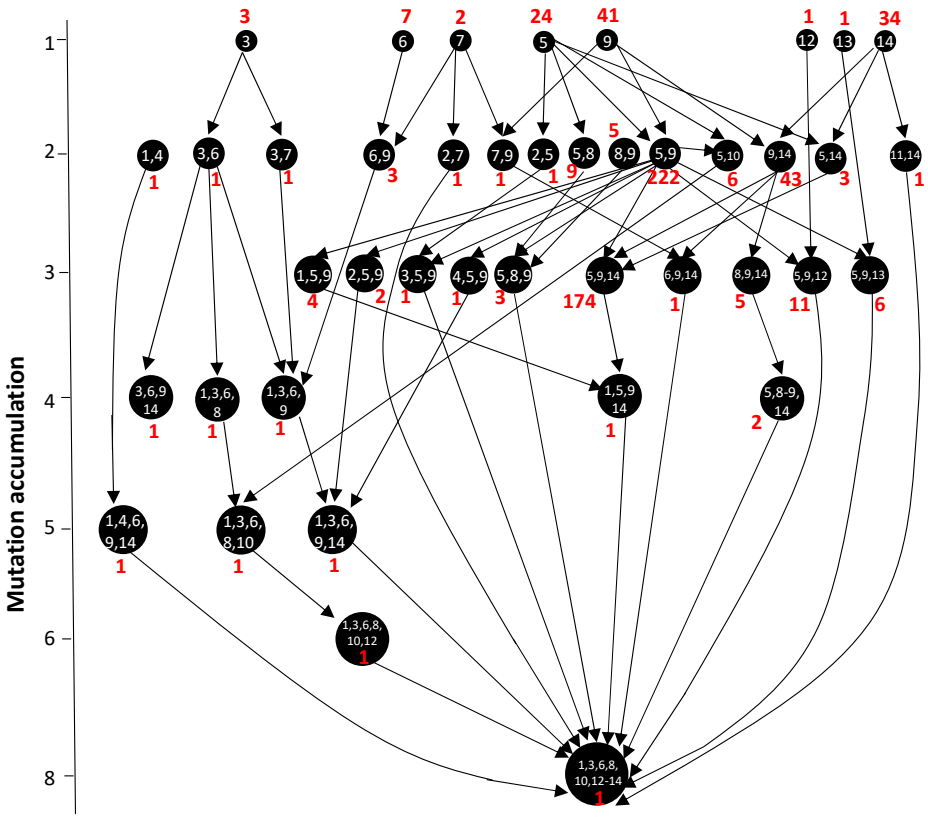
HLA-C (NM_002117)

C



	A	B	C	D	E	F	G	H
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–0	0–1	0–0	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–1	0–0	0–0	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–0	0–1	0–0	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–2	0–0	0–2	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–0	0–0	0–1	0–0	0–0	0–0
	0–0	0–0	0–0	0–0	0–0	0–0	0–0	1–0
	0–0	0–0	0–0	0–1	0–0	0–0	0–0	0–0
	0–1	0–0	0–0	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–3	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–1
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–0	0–0	0–0	0–0	0–0	0–1
	0–0	0–0	0–0	0–0	0–1	0–0	0–0	0–0
	0–1	0–0	0–0	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–1	0–1	0–0	0–0	0–0	0–0
	0–1	0–0	0–2	0–1	0–0	0–0	0–0	1–2
	0–1	0–1	0–0	0–0	0–0	0–0	0–0	0–0
	0–2	0–1	0–4	0–1	0–0	0–0	0–0	1–0
	0–1	0–3	0–2	0–0	0–0	0–0	0–0	0–0
	0–1	0–2	0–6	0–1	0–0	0–0	0–0	1–0
	0–0	0–0	0–3	0–1	0–0	0–0	0–0	0–2
	0–29	0–16	0–47	0–8	0–4	3–0	14–0	17–36
	0–38	0–23	0–63	0–8	0–4	4–0	6–0	25–51
	0–0	0–0	0–1	0–1	0–0	0–0	0–0	0–1
	0–6	0–2	0–11	0–1	0–0	0–0	0–0	2–2
	0–1	0–0	0–0	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–2	0–0	0–0	0–0	0–0	0–0
	0–1	0–0	0–1	0–1	0–1	0–0	1–0	0–0
	0–0	0–0	0–3	0–1	0–1	0–0	0–0	0–0
	0–9	0–5	0–13	0–4	0–2	0–0	1–0	2–7
	0–6	0–3	0–19	0–3	0–1	0–0	2–0	2–5
	0–0	0–1	0–0	0–0	0–0	0–0	0–0	0–0
	0–0	0–0	0–0	0–0	0–0	0–0	0–0	0–1
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–6	0–5	0–13	0–2	0–0	0–0	1–0	0–7
	0–10	0–5	0–20	0–4	0–0	2–0	0–0	2–15
	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC

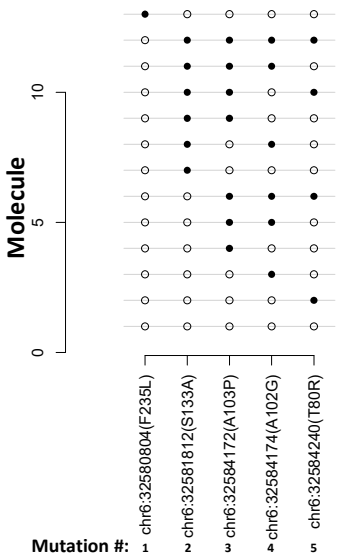
D



Supplemental figure 3

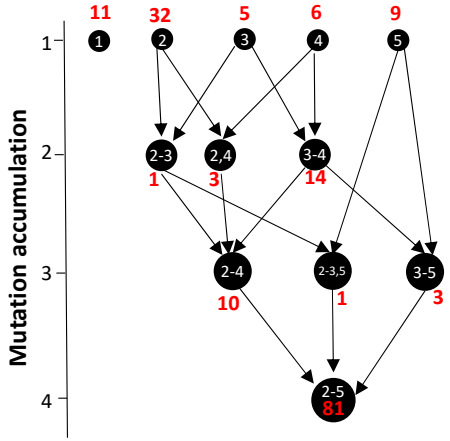
HLA-DRB1 (NM_002124)

E



	A	B	C	D	E	F	G	H
	0–1	0–0	0–5	0–3	0–2	0–0	0–0	0–0
	0–15	0–9	0–28	0–8	0–0	0–0	1–0	6–14
	0–2	0–0	0–6	0–1	0–0	0–0	0–0	0–1
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	0–0
	0–1	0–0	0–0	0–0	0–0	0–0	0–0	0–0
	0–1	0–0	0–2	0–0	0–0	0–0	0–0	0–0
	0–6	0–4	0–13	0–5	0–0	0–0	1–0	1–2
	0–0	0–0	0–1	0–1	0–0	0–0	0–0	0–1
	0–2	0–1	0–6	0–1	0–0	0–0	1–0	0–3
	0–0	0–0	0–1	0–0	0–0	0–0	0–0	1–3
	0–1	0–0	0–2	0–1	0–0	0–0	0–0	2–0
	0–0	0–0	0–5	0–2	0–1	0–0	0–0	1–0
	0–40	0–24	0–58	0–9	0–5	2–0	3–0	29–54
	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC	Benign HCC

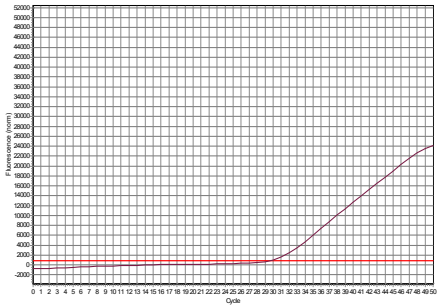
F



Supplemental figure 4

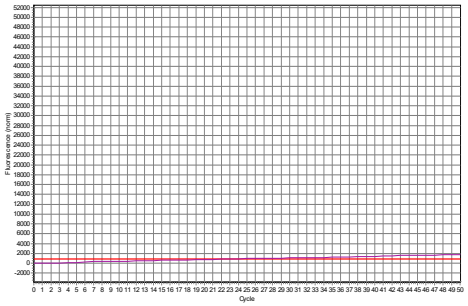
Benign liver

PDCD6-CCDC127



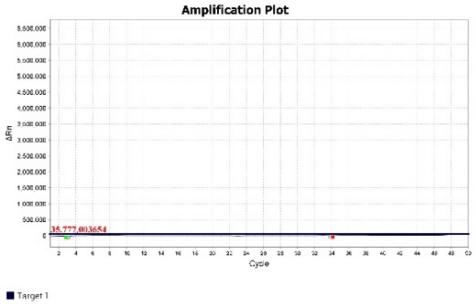
Threshold: 848 (Noiseband)
Baseline settings: automatic, Drift correction OFF

ACTR2-EML4

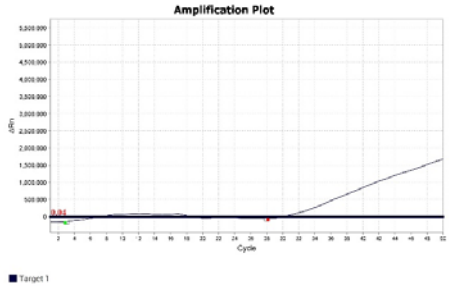


Threshold: 848 (Noiseband)
Baseline settings: automatic, Drift correction OFF

PLG-FGG

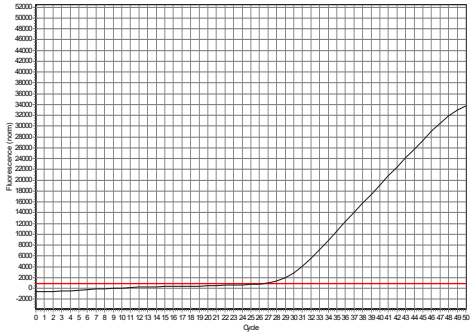


β-actin



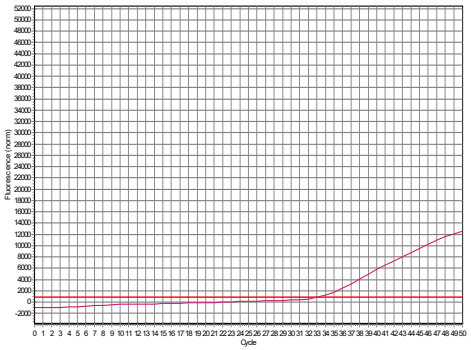
Hepatocellular carcinoma

PDCD6-CCDC127



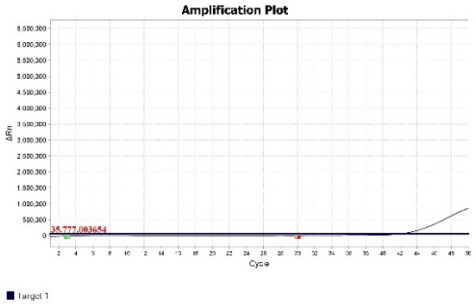
Threshold: 848 (Noiseband)
Baseline settings: automatic, Drift correction OFF

ACTR2-EML4

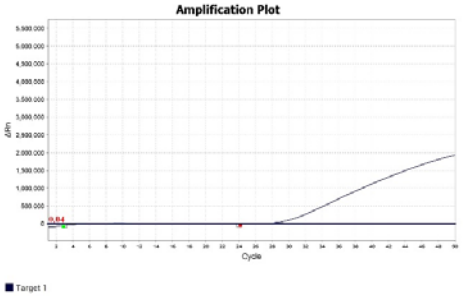


Threshold: 848 (Noiseband)
Baseline settings: automatic, Drift correction OFF

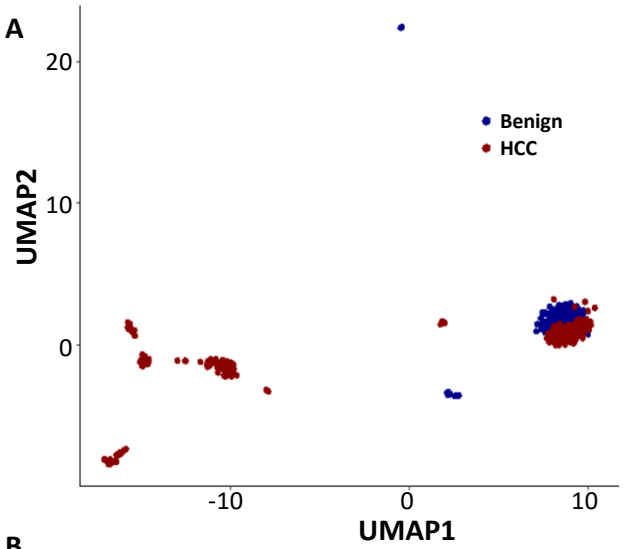
PLG-FGG

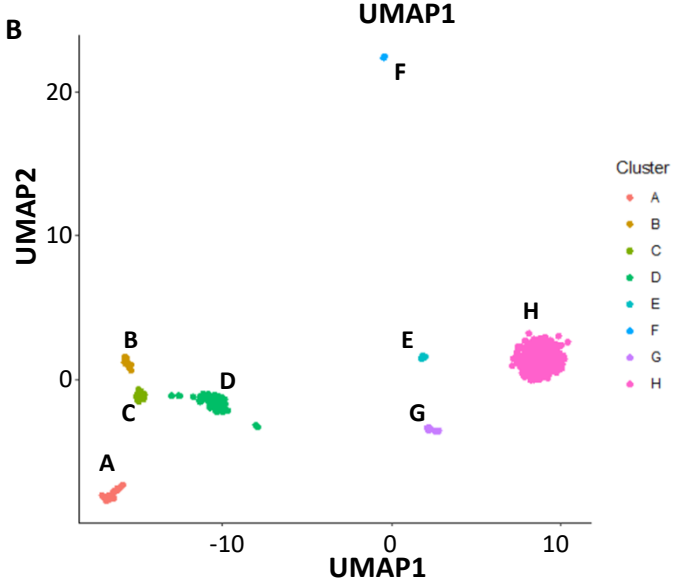


β-actin

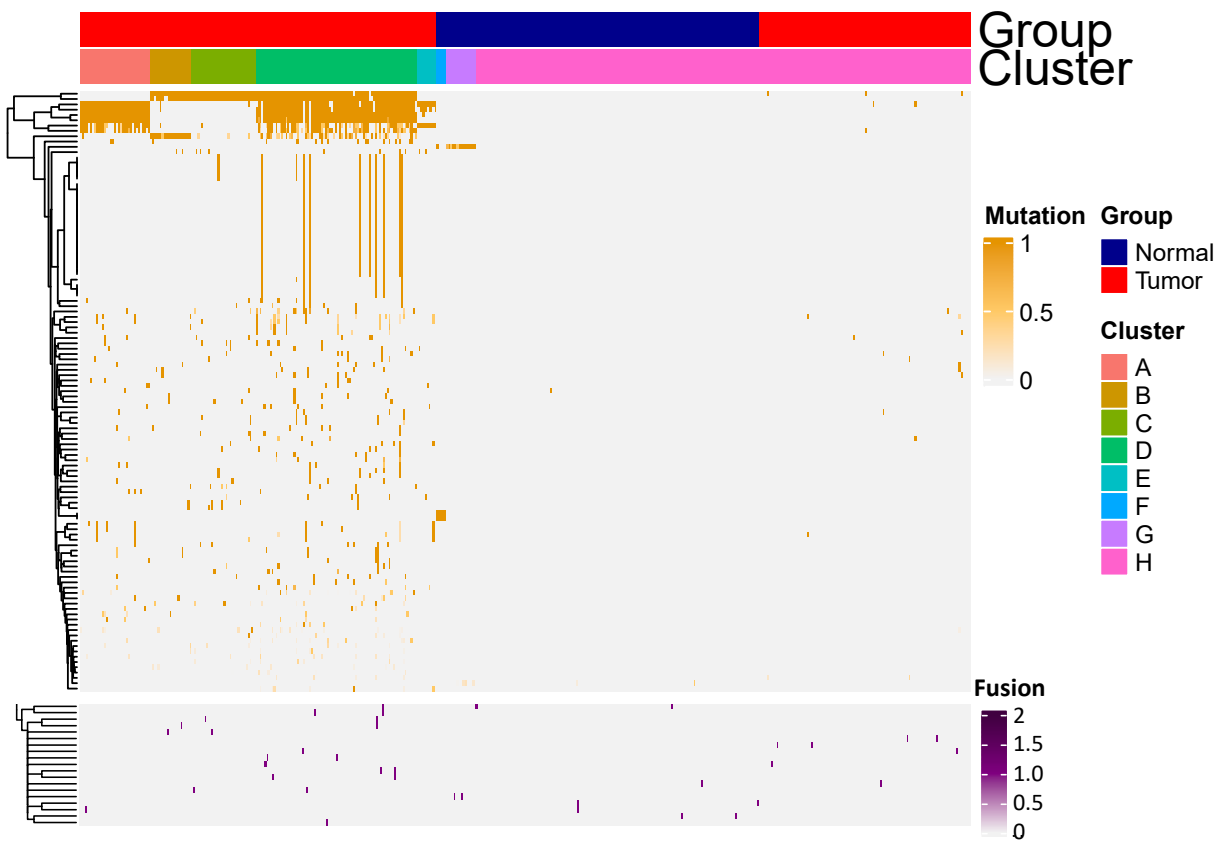


Supplemental figure 5

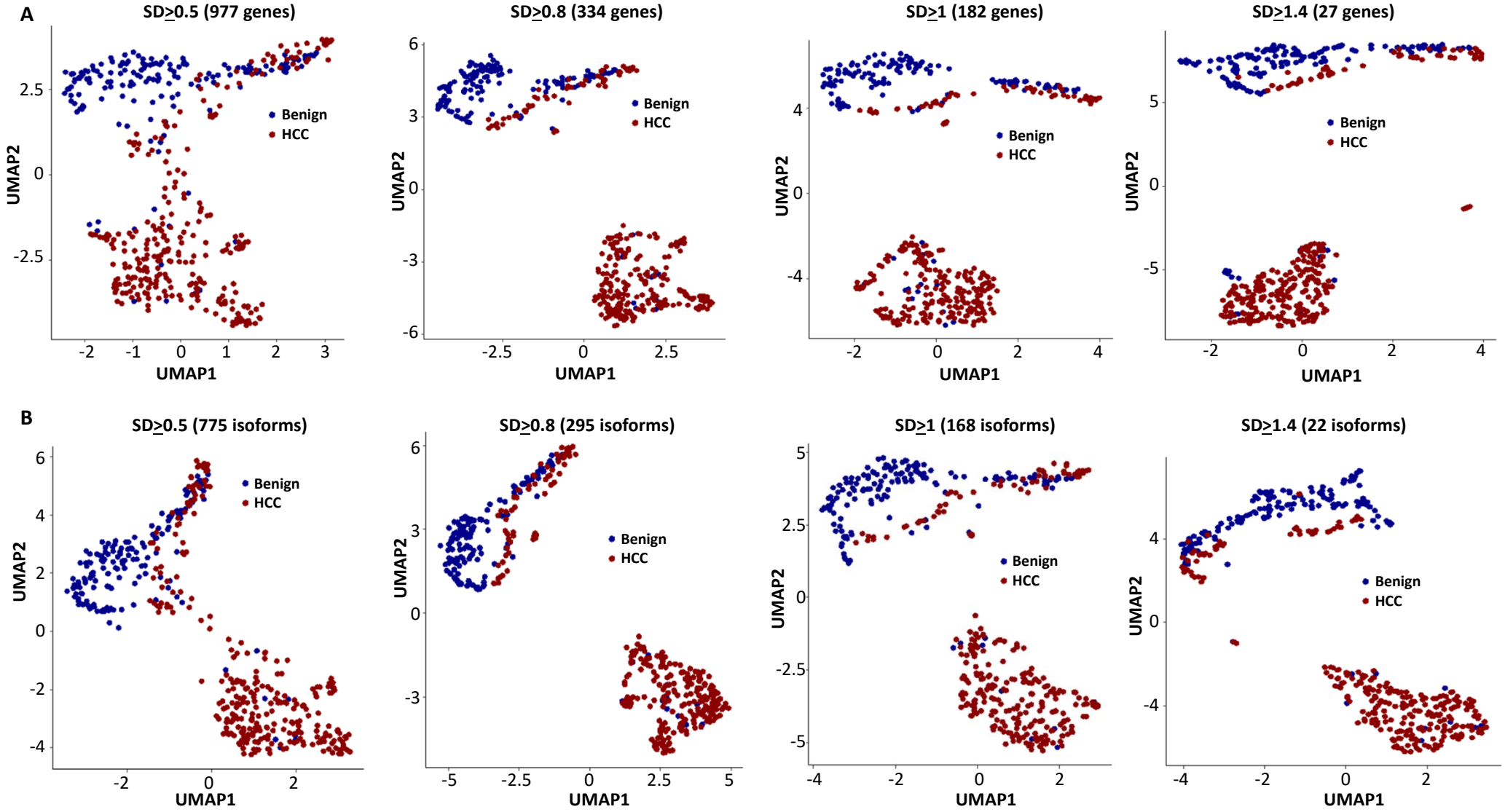




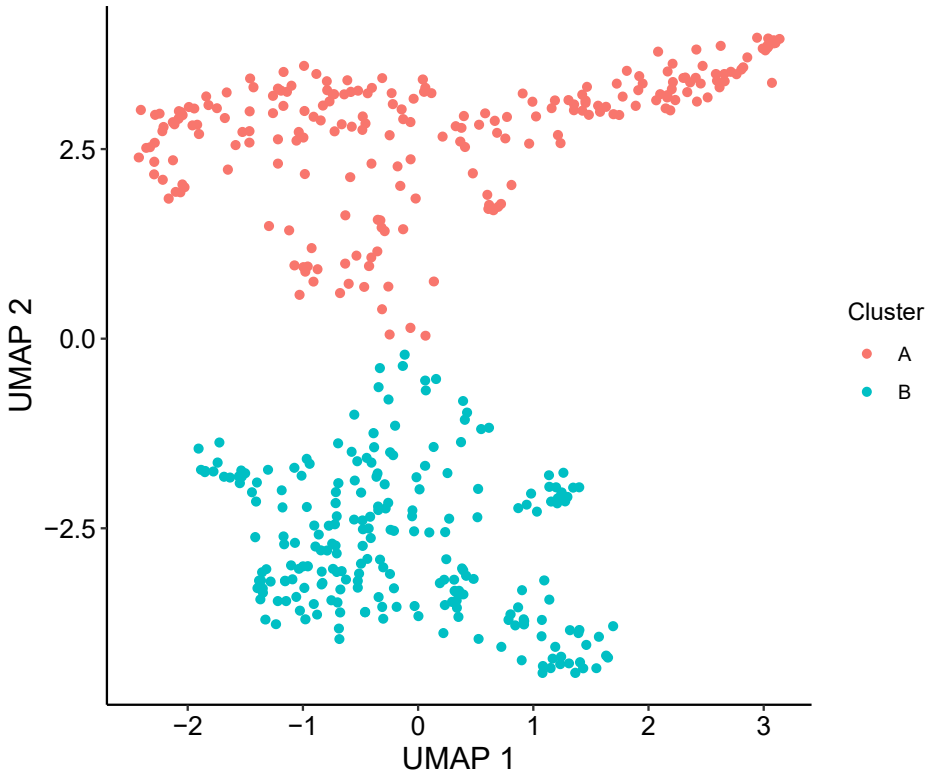
	A	B	C	D	E	F	G	H
Benign	0	0	0	0	0	5	15	142
HCC	35	21	32	81	10	0	0	106



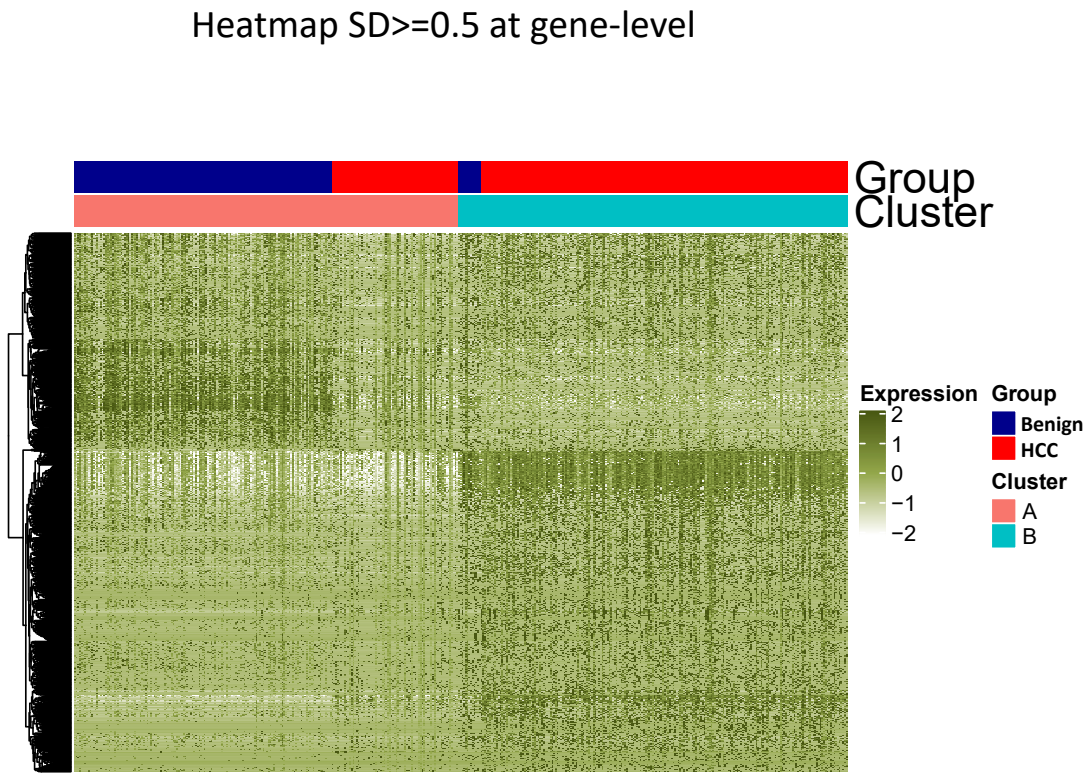
Supplemental figure 6



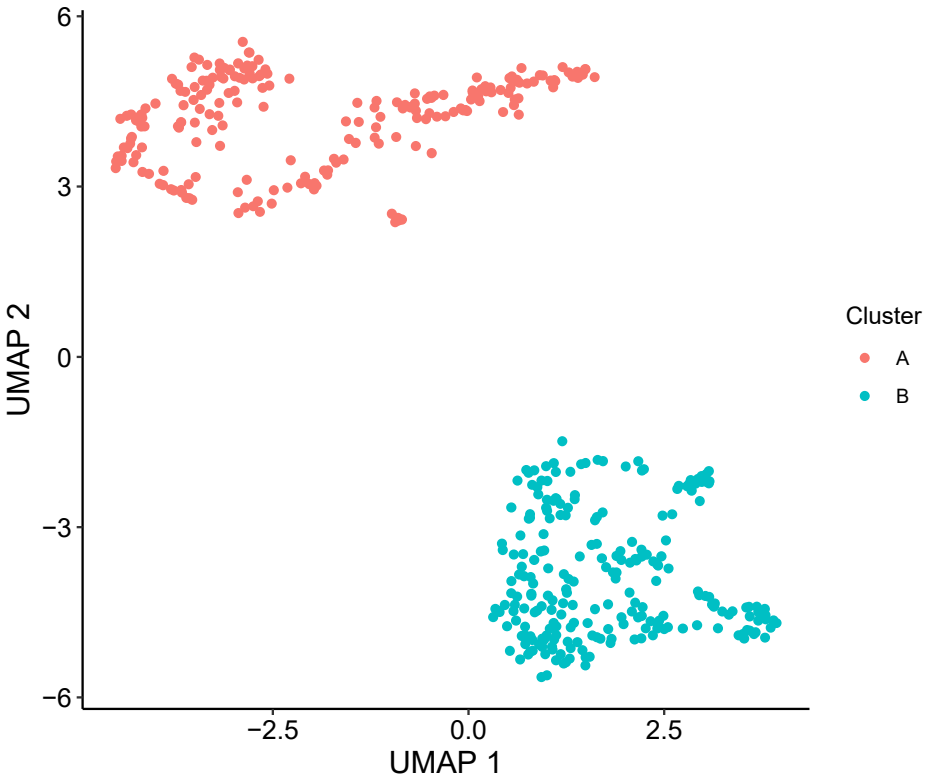
Supplemental figure 7A



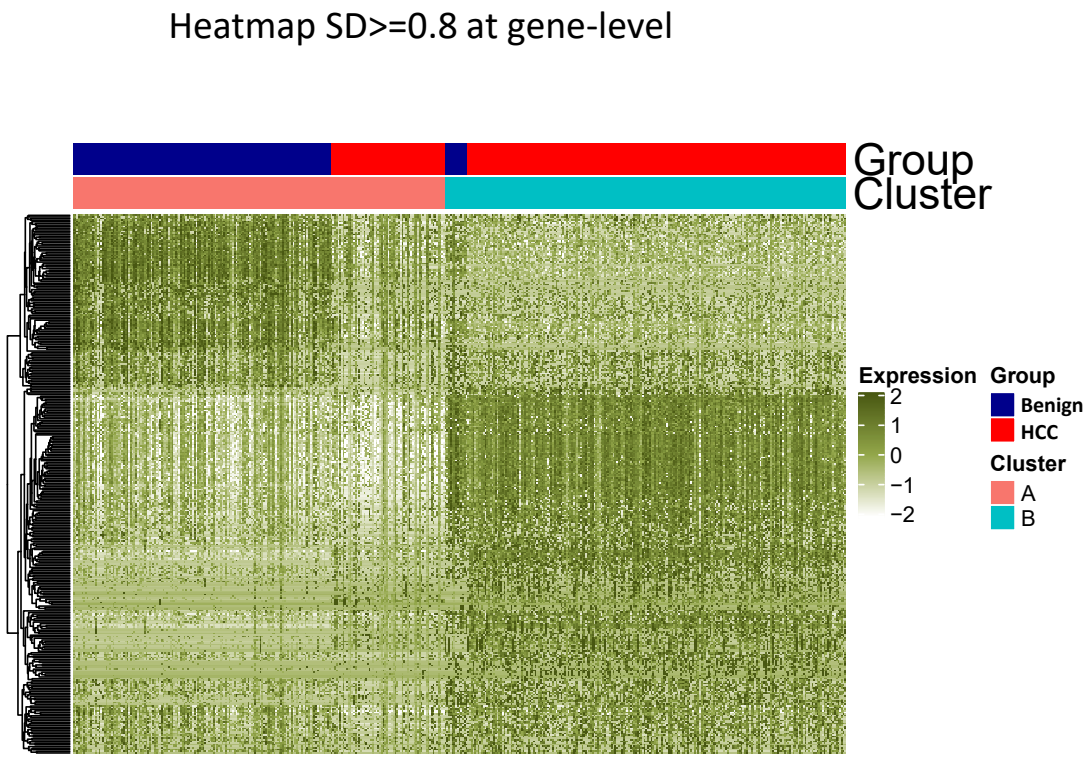
	A	B
Benign liver	149	13
HCC	73	212



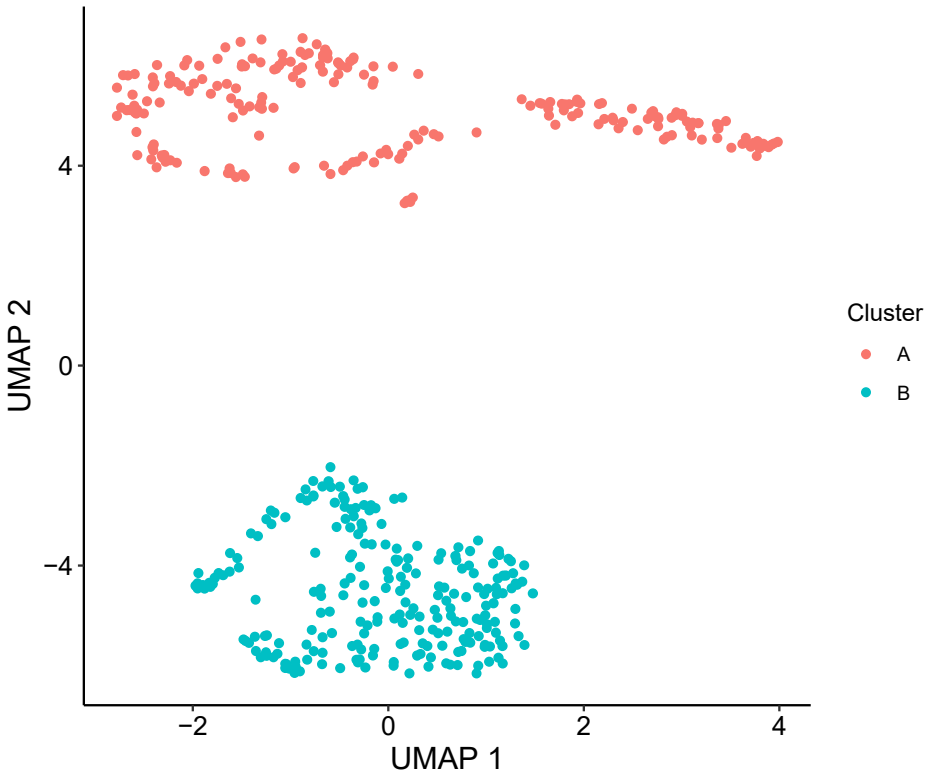
Supplemental figure 7B



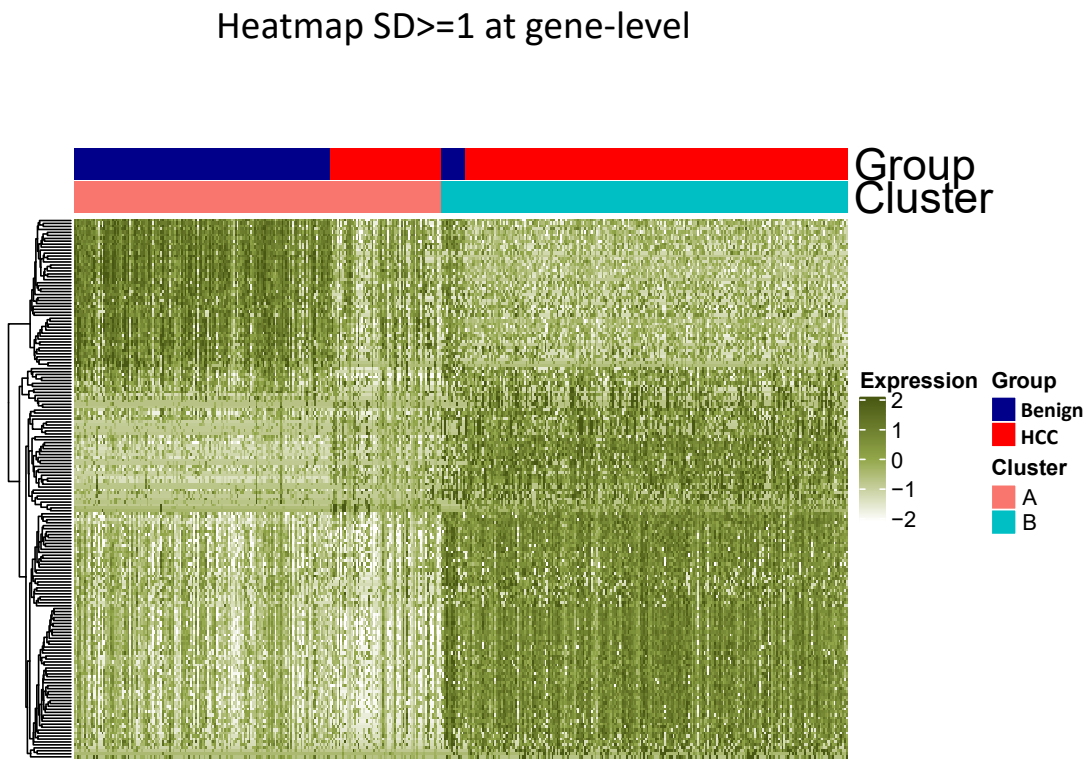
	A	B
Benign liver	149	13
HCC	66	219



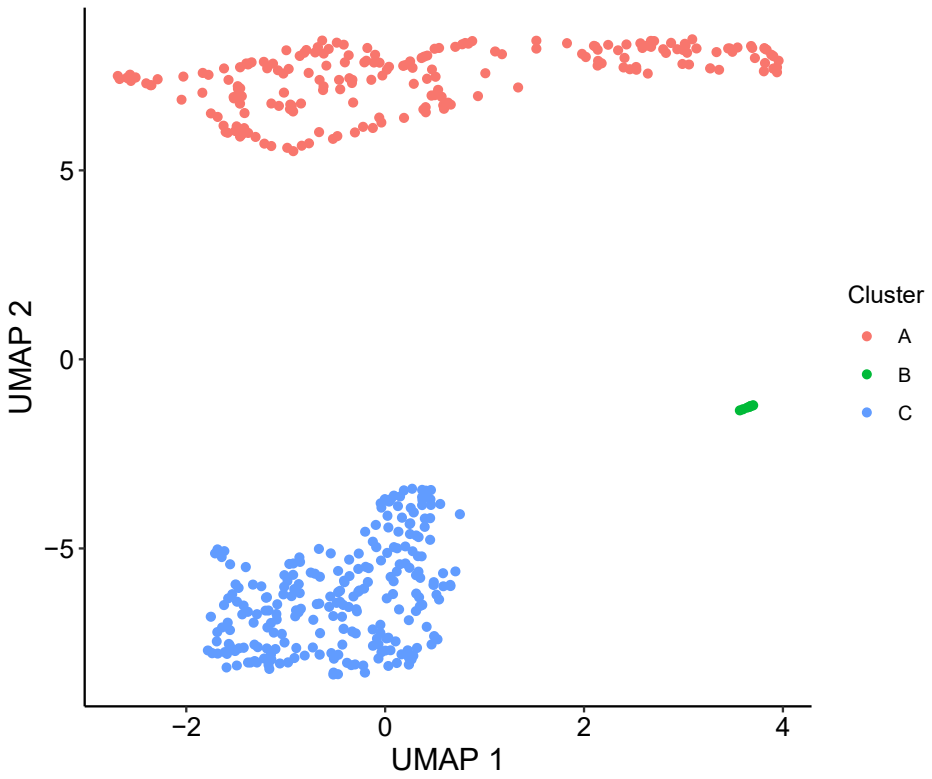
Supplemental figure 7C



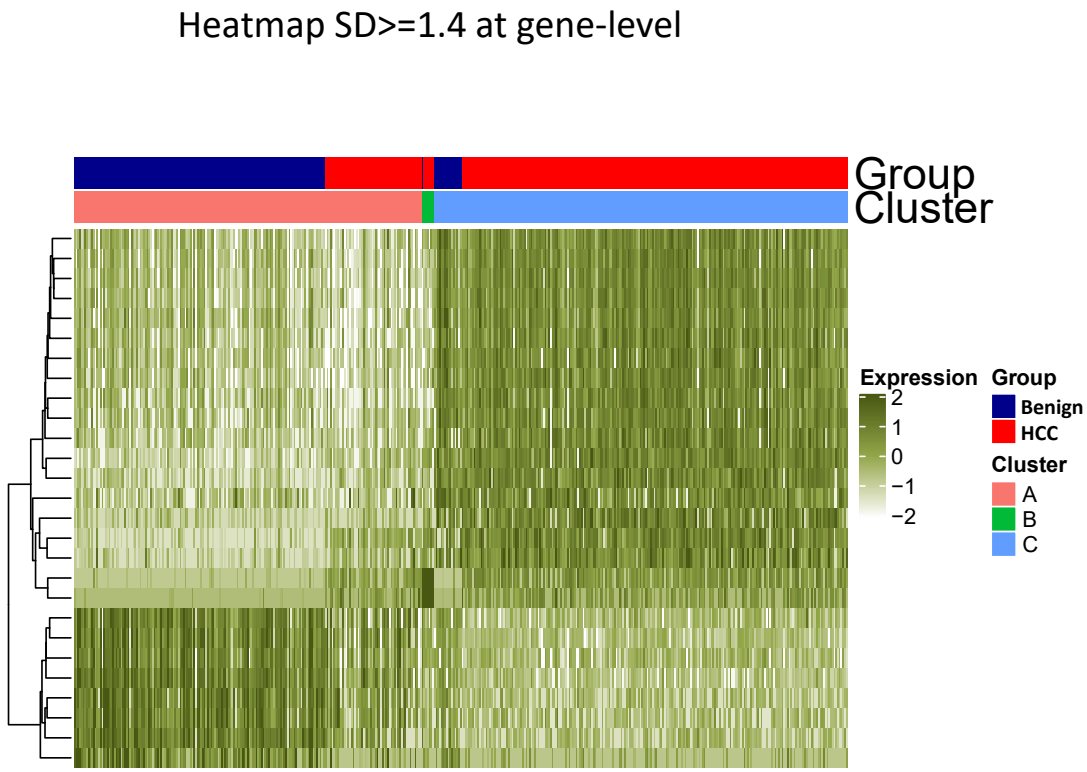
	A	B
Benign liver	148	14
HCC	64	221



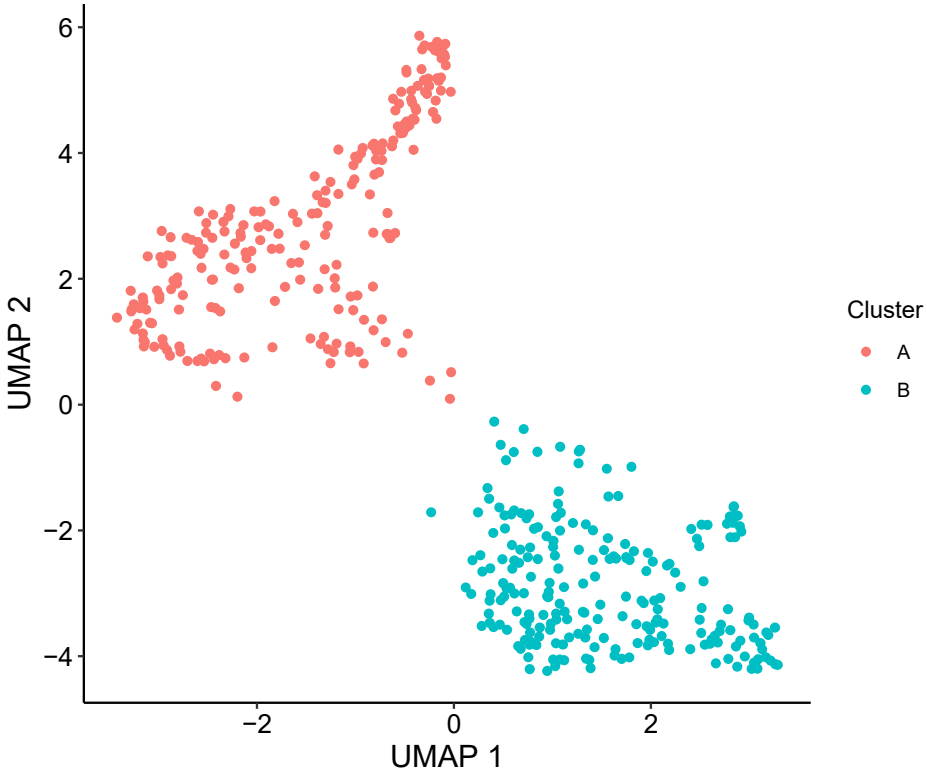
Supplemental figure 7D



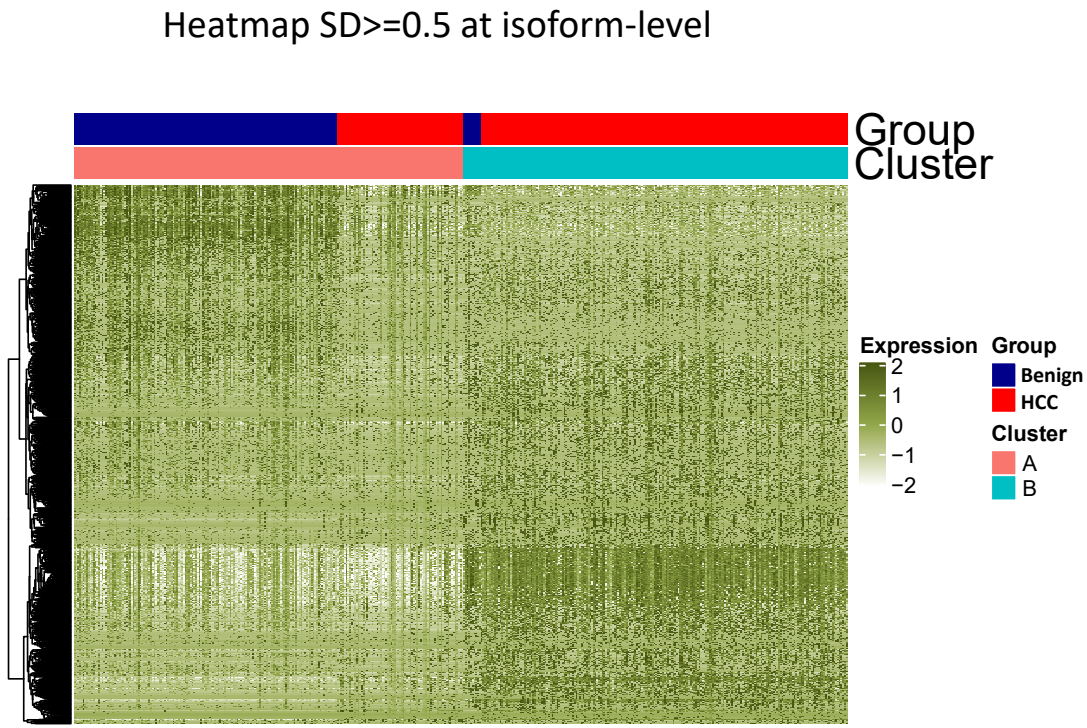
	A	B	C
Benign liver	145	1	16
HCC	56	6	223



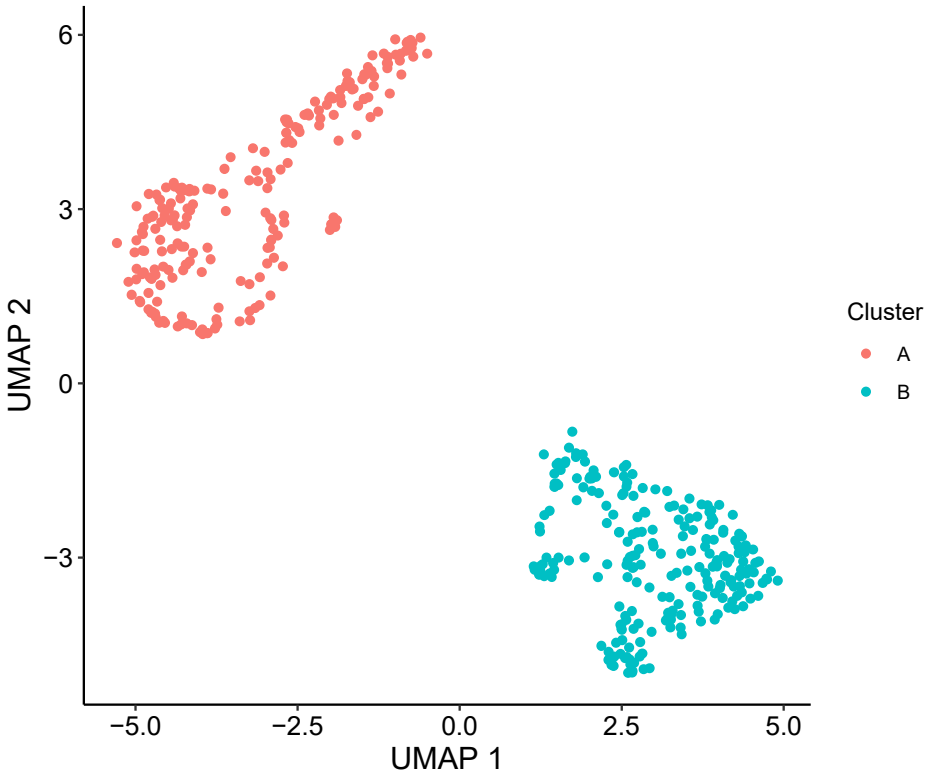
Supplemental figure 7E



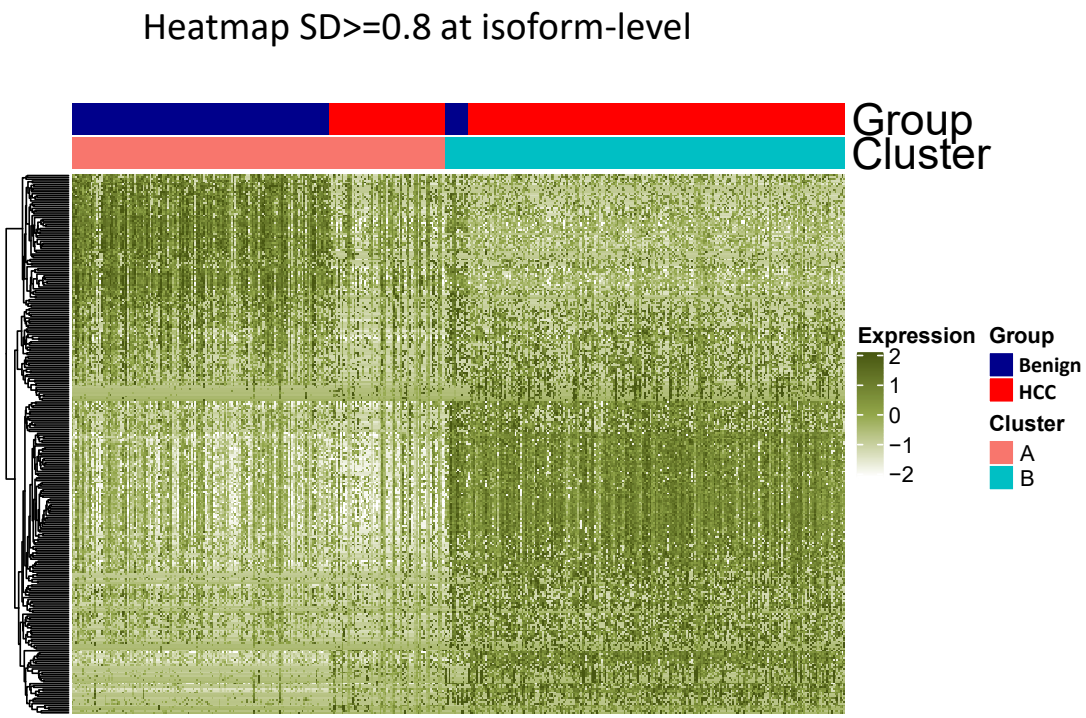
	A	B
Benign liver	152	10
HCC	73	212



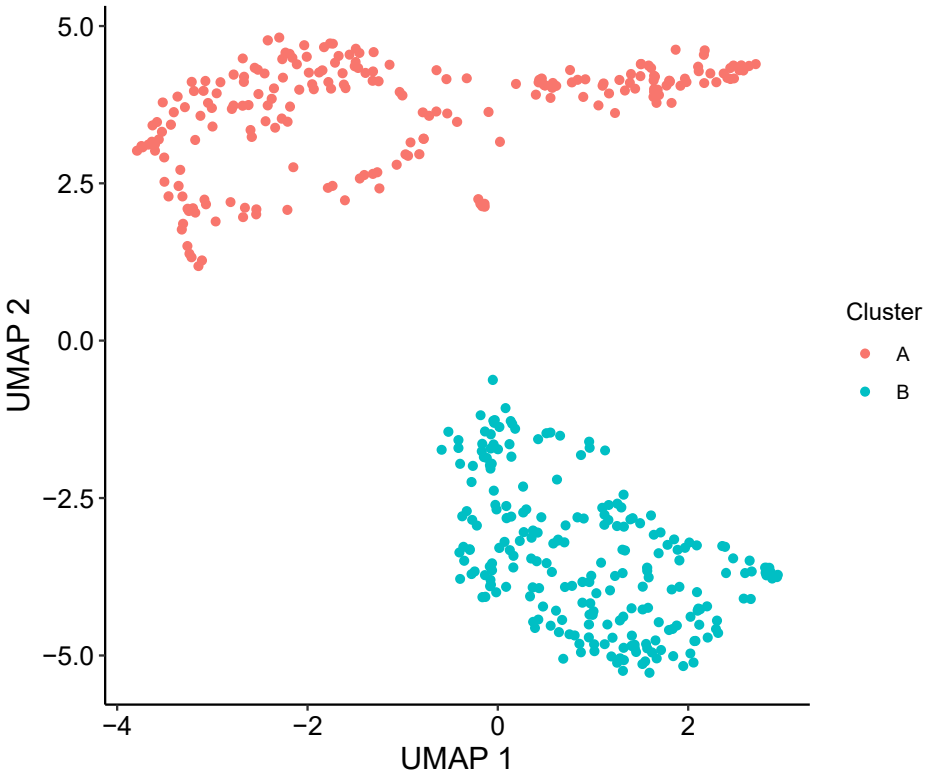
Supplemental figure 7F



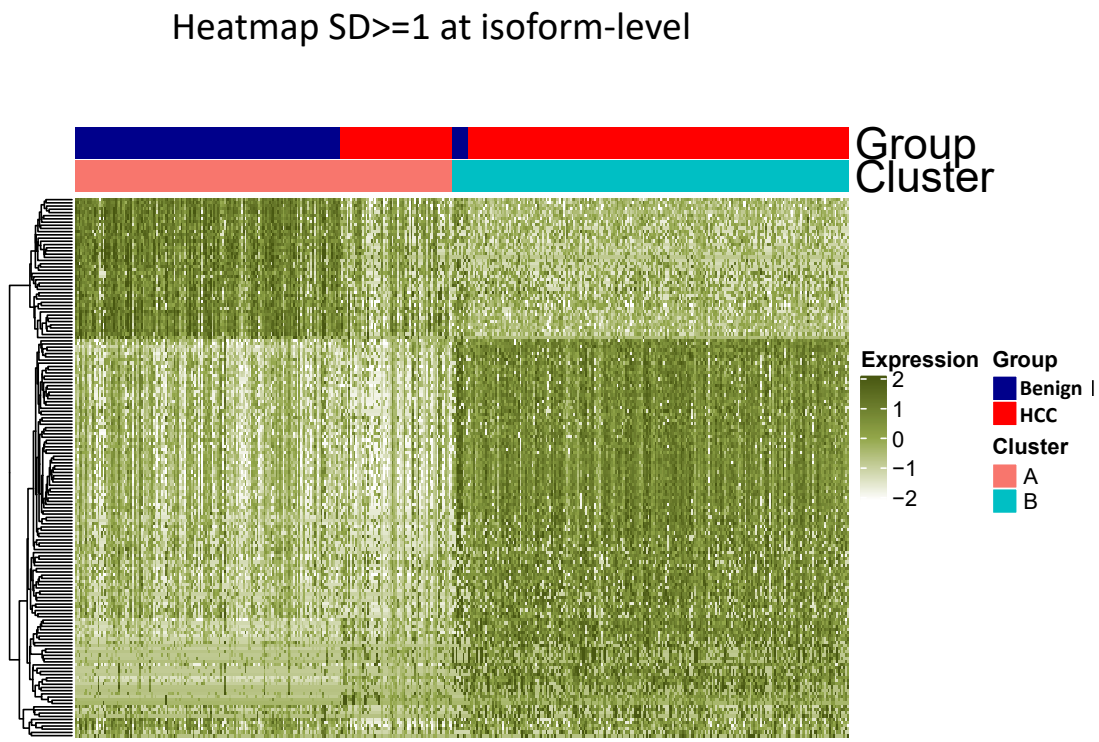
	A	B
Benign liver	149	13
HCC	67	218



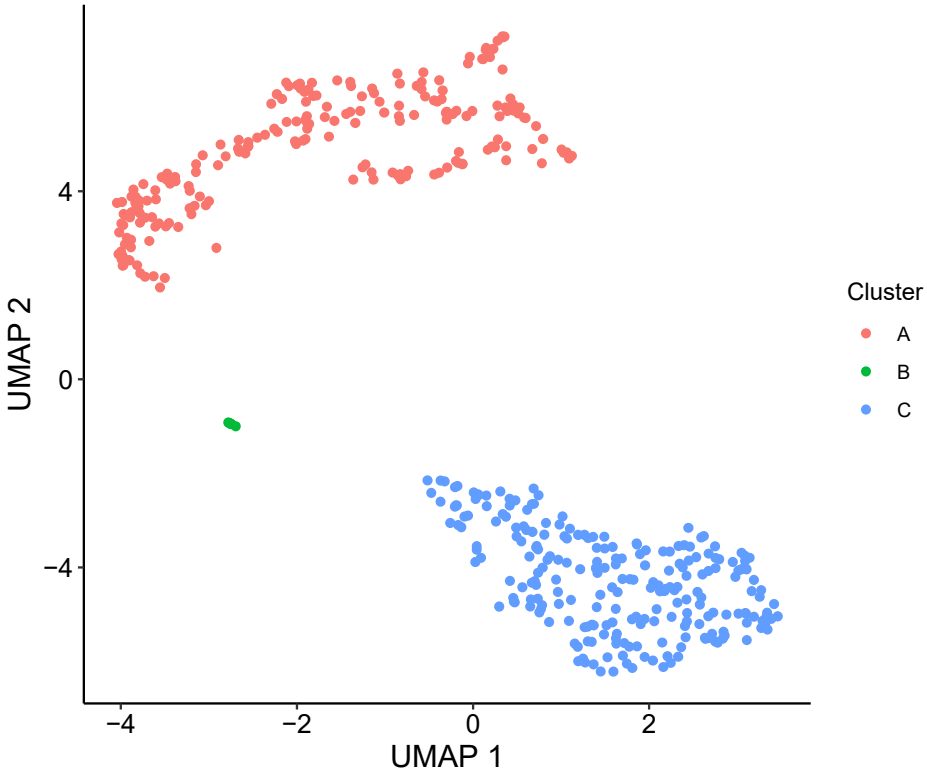
Supplemental figure 7G



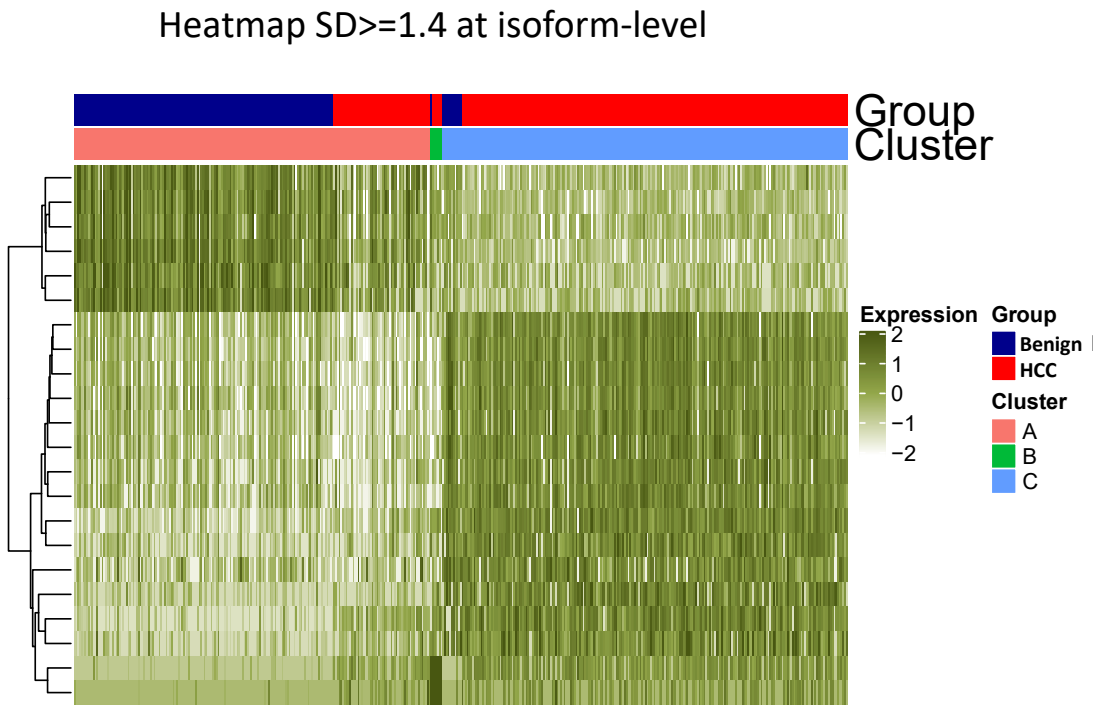
	A	B
Benign liver	153	9
HCC	65	220



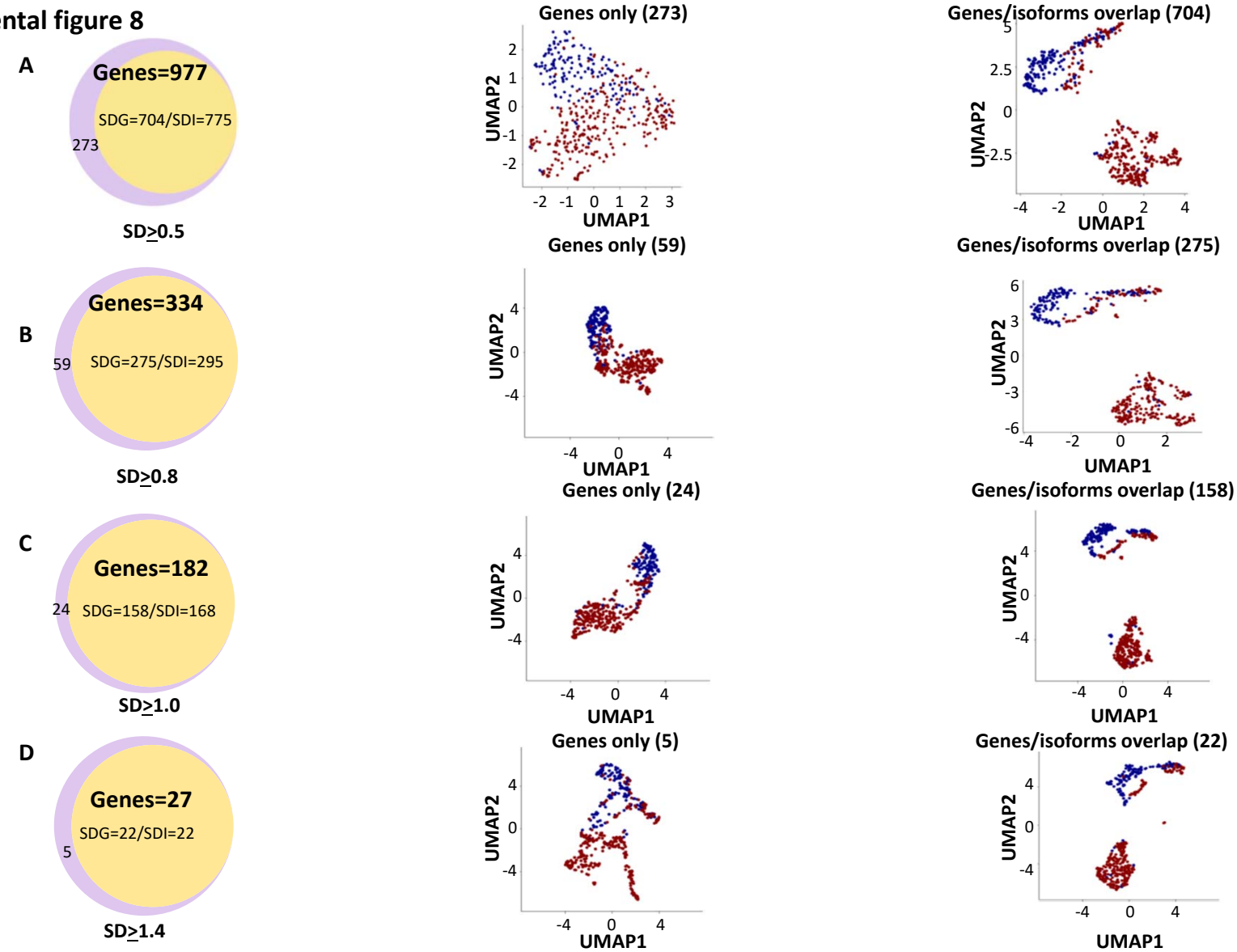
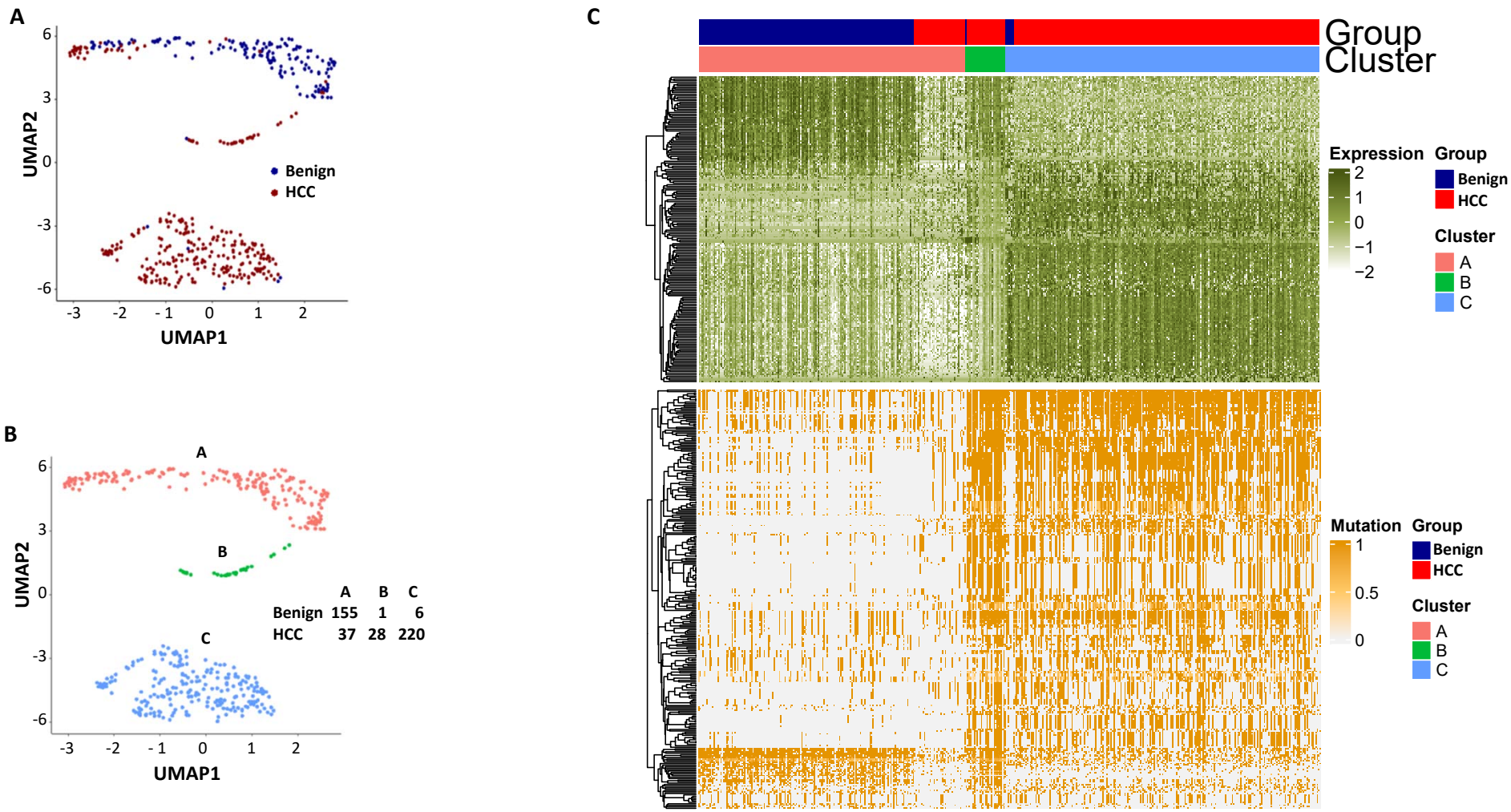
Supplemental figure 7H



	A	B	C
Benign liver	150	1	11
HCC	56	6	223



Supplemental figure 8



Supplemental figure 9