Supplemental Material for
The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients

Zhaoshi Jiang[1*], Suchit Jhunjhunwala[1*], Jinfeng Liu[1], Peter M. Haverty[1], Michael I. Kennemer[2], Yinghui Guan[3], William Lee[1], Paolo Carnevali[2], Jeremy Stinson[3], Stephanie Johnson[4], Jingyu Diao[5], Stacy Yeung[3], Adrian Jubb[4], Weilan Ye[3], Thomas D. Wu[1], Sharookh B. Kapadia[5], Frederic J. de Sauvage[3], Robert C. Gentleman[1], Howard M. Stern[4], Somasekar Seshagiri[3], Krishna P. Pant[2], Zora Modrusan[3], Dennis G. Ballinger[2] and Zemin Zhang[1]

1.Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA 94080, USA
2.Complete Genomics Inc., Mountain View, CA 94043, USA
3.Department of Molecular Biology, Genentech Inc., South San Francisco, CA 94080, USA
4.Department of Pathology, Genentech Inc., South San Francisco, CA 94080, USA
5.Department of Microbial Pathogenesis, Genentech Inc., South San Francisco, CA 94080, USA

* These authors contributed equally to this study

Contents :

Section 1: Sample description and preparation

Frozen tissues and peripheral blood mononuclear cell (PBMC) samples were obtained from commercial sources (Seracare LifeSciences, ProteoGenex and Indivumed) and based on a representation and warranty from the vendors, appropriate informed consent and IRB approval was obtained. Four-micron thick frozen sections were obtained from both primary hepatocellular carcinoma and the matched non-neoplastic liver tissue for histopathological evaluation by standard H&E stain. All four samples had a tumor percentage greater than 80% (Supplemental Table 1). Chronic active portal inflammation was observed with evidence of periportal fibrosis in non-neoplastic liver tissue (Supplemental Fig. 1). The Hepatitis B virus (HBV) infection status was confirmed by polymer chain reaction (PCR). Three patients (patient identifier: H442, 31107 and 31656) are HBV positive and one patient (patient identifier: H384) is HBV negative. The DNA and RNA were extracted from the frozen tissue by a standard protocol using DNA/RNA extraction kit (Qiagen).

Section 2: Whole genome sequencing

Whole genome DNA sequencing (DNA-seq) was performed by "unchained combinatorial probe anchor ligation sequencing", as described previously (Drmanac et al. 2010). The resulting mate-paired reads with an expected intervening distance (~400 bp) were mapped to the human reference genome (NCBI Build 37). First, both paired-end reads were aligned to the reference genome, resulting in an average of 237 billion base pairs of mapped sequences per sample. The average coverage was greater than 80X (Supplemental Table 1). For locations with any evidence of differences from the reference genome, mapped reads were assembled into the best-fit diploid genome. This process results in single nucleotide variation, insertion and deletion calls with associated variant quality scores (Drmanac et al. 2010). Overall, 92~96% of the human reference genome were fully called. Furthermore, for patient 31656, we performed additional whole genome sequencing for both the tumor and non-tumor liver tissue using the same DNA libraries and brought the total coverage for both tumor and non-tumor samples for patient 31656 to ~240X. The purpose of this ultra-depth coverage sequencing was to obtain a more comprehensive collection of HBV viral integration sites from one individual.

Section 3: Transcriptome sequencing

Total RNA was subject to oligo (dT) capture and enrichment, and the resulting mRNA fraction was used to construct complementary DNA libraries. Transcriptome sequencing (RNA-seq) was

performed on the Illumina HiSeq Platform using the standard paired-end protocol. In total, 25-35 million 75 base pairs (bp) reads were generated per sample. For the purpose of mapping, we constructed a custom genome containing the sequences from the human genome (NCBI Build37), HBV reference sequences (n=73), HCV reference sequences (n=42) and HDV reference sequences (n=10). The RNA-seq reads were first aligned to ribosomal RNA sequences to remove potential ribosomal reads. The remaining reads were aligned to the custom genome using GSNAP (Wu and Nacu 2010), allowing maximum of 5 mismatches per 75bp sequencing end. The number of reads mapped to the HBV sequences varied greatly among the samples, ranging from 90 to 2200 reads per million non-ribosomal reads in the three HBV-positive patient samples. In the HBV-negative patient samples, we only detected negligible numbers of reads (fewer than 1 per million) mapped to the HBV reference genome. This result not only reassures that this patient is indeed HBV negative but also suggests that the large number of viral reads detected in the three HBV positive patient samples are not false positives due to mapping artifacts or other reasons. In addition, we did not find any evidence of HCV or HDV co-infection for all the four patients after mapping the reads against both HCV and HDV reference sequences. Eighty-three percent of reads were uniquely mapped to the human reference genome. To quantify the gene expression level, the number of reads mapped to the exons of each RefSeq gene was calculated, and the corresponding RPKM value (reads mapping to the genome per kilobase of transcript per million reads sequenced) (Mortazavi et al. 2008) was also derived.

To quantify the viral transcriptome coverage, we considered all reads with at least one end mapped to the viral genome. From the BAM files generated by GSNAP (Wu and Nacu 2010), we created a pileup at each base using Samtools v0.1.12a (Li et al. 2009), and we then calculated the coverage at each position by counting the total number of A/T/G/C base calls at each position, ignoring gaps. All three patients showed a loss in expression of the viral core protein (Supplemental Fig. 6, right panels). This was likely due to predominant integration via the DR1 site, which lies immediately upstream of the gene encoding the core protein, resulting in disruption of its promoter.

Section 4: HBV integration detection

We first aligned all reads from whole genome sequencing against a comprehensive list of Hepatitis B virus reference sequences (n= 73, Supplemental Table 2). The strain with the highest depth coverage of viral reads and the best match for discovered variants was selected as the reference strain for a given patient. We found that three patients were infected by three distinct

strains of HBV (B, C, D, genotypes) and thus the final reference genomes used in this study are AY033073.1, NC_003977 and V01460.

We then utilized the paired-end nature of our reads and searched for human-virus chimeric reads, where one end of reads mapped to human genome and the other end of reads mapped to the viral reference genome, an indication of HBV integration in the human genome. All uniquely mapped chimeric reads were retained. Adjacent or overlapping chimeric reads (within 500 bp) aligning to the human and viral genomes in the same orientation were merged to make clusters. Next, adjacent clusters, aligning within 500 bp of each other on the human genome were considered to indicate the same integration event and only one of them, with the higher number of supporting reads, was retained as a representative for that event.

The clusters with at least two chimeric reads were retained (n=48, Supplemental Table 3A). The integration sites were then compared to RefSeq gene boundaries to find genes that were directly disrupted by HBV integration (overlapping) or potentially affected by integration (within 15 kb of integration sites). The same approach was applied to transcriptome sequencing data wherein we identified 114 distinct human-viral fusion transcripts (Supplemental Table 3B). PCR of genomic DNA followed by Sanger sequencing across human-viral junctions was used to confirm these viral integration events. Six out of 8 high-frequency integration sites (with more than 10 chimeric reads support) were readily detectable by 30 cycles of PCR amplification. However, only by increasing PCR cycle number to 35, we were able to detect three low frequency cases (with 2 reads) and confirm the viral-human junction of these three cases by Sanger sequencing. This suggests that deep genome sequencing technology could pinpoint rare viral integration events that were only detectable by PCR with more sensitive conditions.

Section 5: Calculating chimeric fraction at integration loci

To estimate the fraction of alleles that contain HBV integration compared to the normal alleles at integration loci, we counted the number of normal pairs of reads mapping to the integration loci. The integration locus was defined as the interval we obtained after clustering overlapping human arms of chimeric reads or human arms mapping within 500 bp of each other. We then scanned the BAM files that contain alignments for all the reads sequenced, to select those pairs that would indicate normal alleles at the same locus. The criteria to choose the normal reads were (all of these need to be true):

- One of the reads in the pair is mapped completely within the cluster

- The mapped read is aligned to the human locus in the same direction as the human arm in the chimeric reads belonging to the cluster
- Alignment is primary, which means that a mapping record in considered only if it is the best mapping for the pair.

The chimeric fraction was then defined as the fraction of chimeric reads in the total reads (chimeric reads + normal reads) at the locus.

Section 6: Sequence features near the integration sites

We attempted to determine whether there is any positional bias for viral integration sites in the human genome. For each human-viral chimeric read, we plotted the human genomic location against the viral mapping position (Fig. 4A). We found that viral integration occurs randomly across the human genome, with no obvious positional preference, unlike in the viral genome where there is an obvious preference for the breakpoints to lay near the 3' end of the HBx gene (Fig. 4B). We then checked for any bias for proximity to sequence features near the integration sites on the human genome, including several repeat families, recombination hot spots, gene boundaries and segmental duplications (Bailey et al. 2002). For this purpose, we compared the observed distribution of distances of integration sites to these features with the distance of these features to a set of randomly selected genomic positions (1000 positions selected from each chromosome after removing assembly gap positions in the genome). For the genomic positions of repeat families, we used the track 'repeatmasker' from UCSC genomic build hg19 (http://genome.ucsc.edu). Similarly, for the positions of transcripts from known genes we used the 'refgene' track. We obtained local recombination rates at viral integration sites and at randomly chosen positions using the track 'Recomb rate', as measured by sex-averaged rates of recombination based on deCODE (Kong et al. 2002). Recombination rates were measured at integration sites so that any bias for viral integration sites to be in proximity of recombination hot spots can be discerned. To determine the proximity of an integration site from a feature, we chose the shortest distance of the site from the feature interval boundaries. If the site overlapped with a feature interval, the distance was forced to be negative. If the site was outside the genomic interval spanned by the feature, the distance was taken to be positive. This comparison did not reveal any bias for viral integration sites to be in proximity of any of these genomic features (Supplemental Fig. 7) when compared with a randomly selected set of genomic positions. The distribution of the distances of the integration sites from these features was fairly similar between the tumor samples, the normal samples and the randomly selected genomic positions. Therefore, our data support a random viral integration model on the human genome.

In contrast, a region between 1500 and 2000 bp on the HBV genome was the predominant integration site on the viral genome, since the viral arms of multiple chimeric reads mapped to this region (Fig. 4A). Based on the short-read structure, we can estimate the viral position closest to the viral-human breakpoint, and in cases where a read crosses the human-viral boundary, we can identify the exact breakpoint. Since the same insertion event can be sampled by multiple chimeric reads, in order to get the best estimate of the viral breakpoint, we clustered all reads mapped within 500 bp of each other on the human genome and then, for each cluster, we defined the viral position closest to the putative breakpoint as the viral 'junction' supported by this cluster. In Figure 4B, we represent these clusters, where the vertical bar indicates the viral junction, and the horizontal segment indicates the rest of the viral sequence supported by the chimeric reads cluster. For the sake of clarity, we do not show the full length of the viral sequence supported by the clusters, since they span variable lengths depending on how many reads constitute the cluster and their locations.

Section 7: Transcriptional consequence of HBV integration

Chimeric RNA-seq reads within 500 bp were clustered to obtain human junctions. We define 'junction' as the closest base to the human breakpoint based on these reads. Junction for a read is the base on the read that is closest to the breakpoint, and junction for a cluster is the base pair in a clustered region that is closest to the breakpoint. If two cluster junctions from the same patient lie within 2 Kbp of each other, they were assumed to correspond to the same insertion event, and the junction with the higher number of chimeric reads was retained, to obtain a set of junctions representing non-redundant insertion events. To determine the transcriptional effect on the local neighborhood of insertions, the human genome flanking each junction (10 Kbp on each side of the junctions) was divided into intervals of 200 bp. A RPKM value was assigned to each interval based on RNA-Seq reads mapped to the human genome, for the genome (tumor or non-tumor) carrying the insert, as well as the unaltered genome (non-tumor or tumor). Change in RPKM between the altered and unaltered genome was calculated as a generalized logarithm with a pseudo count of 1 (using the R-package "LMGene").

$$\Delta RPKM = log(RPKM_{altered} + \sqrt{RPKM^2_{altered} + \lambda}) - log(RPKM_{unaltered} + \sqrt{RPKM^2_{unaltered} + \lambda})$$

where

$\lambda = 1$

altered : Genome with viral insertion

unaltered : Matched genome without viral insertion

Section 8: *De novo* assembly of chimeric reads

In order to obtain precise human-viral junctions in fusion transcripts, we attempted to identify paired-end reads where a single arm crosses the junction. For this purpose, we recruited all paired-end reads from RNA-Seq where one arm mapped to the viral genome, and the other arm was either mapped to the human genome or was unmappable. We then conducted *de novo* assembly of these reads using SOAPdenovo (Li et al. 2010) (Release V1.05). Contigs that were greater than the individual read-length (75 bp) were then aligned to the viral genome using Smith-Waterman local sequence alignment (Smith and Waterman 1981). We chose contigs that partially aligned to the viral genome and one of the terminals (5' or 3') were included in the alignment, and then aligned the unmapped parts of these contigs to the human reference genome (NCBI build 37.1), using GMAP (Wu and Watanabe 2005). Contigs that aligned to both the viral and human genomes were then used to determine precise viral-human fusion junctions.

Section 9: Random sampling of reads from ultra-deep sequencing data to depict lower coverage

Patient 31656 was sequenced at ultra-deep coverage, with a mapping yield of 679 Gb for the tumor sample and 706 Gb for the non-tumor sample. Using a genome length of 2.89 Gb, based on the gap-free sequence length of NCBI genome build 37.1, this translates to a fold-coverage of ~234X and ~243X for the tumor and non-tumor samples respectively. In order to simulate lower coverage data, we divided the total reads into two categories, viral reads, where at least one arm maps to the viral genome, and non-viral reads, which may include reads mapped to the human genome or unmapped reads. Since the ratio of viral reads to total reads was of the order of 10^{-6}, we modeled the viral read counts from a draw of fewer total reads (and thus lower coverage) as a random variable following a binomial distribution, with the probability of success of given by:

$p = m_c / n_c$

where

m_c = total viral read count at c-fold coverage.

n_c = viral read count + non-viral read count at c-fold coverage.

c = total coverage of tumor (234X) or non-tumor (243X) sample.

A random binomial generator was used to generate viral read counts at a lower coverage 'x':

m_x = rbinom(size = x/c * n_c , probability = p)

From the pool of viral reads identified at full coverage, m_x reads were the chosen using a uniform distribution random generator function. Chimeric reads among these were then chosen to represent redundant integration events. Chimeric read that aligned in the same orientation to the viral/human genome and that were within 500 bp of each other were assumed to represent the same human-viral junction and were clustered together to obtain non-redundant junctions. Junctions were then clustered together if they were within 500 bp of each other on the human genome, irrespective of their alignment orientation, in order to obtain non-redundant integration events. This procedure was repeated 100 times for each value of x (25, 50, 75, 100, 125, 150, 175, 200 and 225) to obtain a mean value and standard deviation of integration events.

Section 10: Mutation detection, filtering and validation

Variant calls for each sample genome with respect to the human reference genome were made as described previously (Drmanac et al. 2010). Somatic mutations between the tumor and normal genomes were obtained by comparing their variant calls using the tool calldiff-1.3 (http://cgatools.sourceforge.net). Loci that were called as variant in the tumor and reference in the normal genome were considered as somatic mutations. Somatic scores were assigned to the mutation calls using calldiff-1.3, where a higher score indicates lower likelihood that the called variation in the tumor genome is false-positive and the reference call in the normal genome is false negative. Mutations that were present in dbSNP v131 were filtered out to obtain novel mutation calls. We further filtered out any variations that were found in 1000 Genomes (Nov 2010 release), variations present in 60 normal genomes release by Complete Genomics Inc. (The data and description of the genomes is available at http://www.completegenomics.com/sequence-data/download-data/)

Mutations were annotated for their effect on transcripts using the variant effect predictor tool (McLaren et al. 2010). An in-house version of the RefSeq database was used within the ENSEMBL framework as the data source for variant effect predictor. The different types of consequences predicted are intergenic, regulatory region, upstream (within 5 Kb), 5' UTR, complex indel (spans intron/exon border), splice site (1-3 bp into exon, 3-8 bp into intron), synonymous coding, non-synonymous coding, intronic, frameshift coding, stop gained, stop lost, 3' UTR and downstream (within 5 Kb).

We experimentally tested a selective subset of somatic single base substitutions for tumor-normal comparisons from all four patients using Sequenom (291 mutations from H442, 283 from 31107, 453 from H384 and 292 from 31656). For the purpose of validation, we filtered out mutations that were annotated as intergenic, intronic or downstream. We also filtered out mutations in pseudogenes or hypothetical genes, based on their description in the Entrez Gene database. For patients H384 and H442, which also had blood samples, somatic mutations between tumor and blood were also included. The score performed well in all the patients (Supplemental Fig. 8, AUC ranging from 0.9055 to 0.9741). After pooling the data from all four patients, a threshold score of 0.1 was used to obtain high confidence mutations at 83.8 % sensitivity and 89.5 % accuracy (i.e. FDR of 10.5%).

Section 11: Structural variations (SV) detection and validation

All uniquely mapped reads from whole genome sequencing were used for the estimation of normal pair-span. The set of mate-pairs was further refined by aligning each read within normal range of pair-span with penalties for mismatches and indels. Each read with less than four penalty units was retained. The orientation of the reads that were mapped to the forward strand of the reference genome was designated as plus ("+"); otherwise it was assigned minus ("-"). The discordant mate-pairs were defined as either (1) mate-span beyond normal range (500 bp), or (2) with discordant orientation. Adjacent discordant reads (within 500bp) with the same orientation were then merged to make a discordant reads cluster. The clusters that contained too few discordant mate pairs (<3) after merging were discarded.

We then applied a filtering process to define high confidence somatic structural variations as: (1) SVs supported by sufficient number of discordant mate-pairs (pair_count ≥ 10) and (2) with large intra-chromosomal span (≥ 5kb) in the tumor samples. We then excluded those SVs that were also present in matched non-neoplastic sample or other normal samples that we sequenced. A subset (n=51) of putative somatic SVs that overlapped with RefSeq genes were subject to experimental validation.

We designed PCR primers that flank putative somatic SV breakpoints. PCR amplification was performed on tumor and non-tumor liver tissue and/or blood samples. PCR conditions were as follows: 50ng of genomic DNA was amplified in a 25uL reaction with each primer at 400nM, each deoxynucleoside triphosphate at 300uM and 2.5 units of LongAmp™ Taq DNA polymerase (New England Biolabs, Ipswich, MA). PCR was performed with an initial denaturation at 95°C

for 3 minutes followed by 30 cycles at 95°C for 10 seconds, 56°C for 1 minute and 68°C for 1 minute and a final extension step at 68°C for 10 minutes. Specific PCR bands were purified with QIAquick Gel Extraction Kit (Qiagen, Valencia, CA). The gel-purified DNA was either sequenced directly with specific primers or cloned into TOPO cloning vectors pCR®2.1 (Invitrogen) and sequenced. A somatic breakpoint was considered validated only when the result met the following criteria, (1) A PCR band was specifically amplified in a tumor sample but not in its corresponding non-neoplastic sample(s); (2) The DNA sequence of a PCR band was unambiguously mapped to the distinct genomic region predicted by the mate-paired sequence reads. About 70% or 37/51 breakpoints were confirmed at basepair resolution.

Section 12: Copy number variations (CNV) and loss of heterozygosity (LOH) detection

DNA copy number variation (CNV) and allele-imbalance (AIB/LOH) was defined by read depth coverage and B-allele frequency analysis. For each sample, DNA sequencing reads were binned at 50 Kbp intervals along the genome and counted. The ratio of counts per bin in the tumor and its matched normal sample, $\log_2$ transformed, was calculated as the raw measure of copy number ($\log_2$(tumor/normal)). This value was corrected for GC content bias using GC content information from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/gc5Base/hg19.gc5Base.txt.gz). These GC content data were averaged over 500bp windows and the smoothed over a 1Mb window using a running mean. The residuals of the regression of $\log_2$ ratio on GC content were taken as the raw measure of copy number. These raw $\log_2$ ratios were then shifted to have a mode of 0. These values were segmented into discrete blocks of uniform copy number using the CBS algorithm from the Bioconductor package DNAcopy (Venkatraman and Olshen 2007). The parameters for CBS were smooth.region=2, outlier.SD.scale=4, smooth.SD.scale=2, and trim=0.025. Segments with a $\log_2$ ratio $\leq$-0.15 were considered as regions of copy loss whereas segments with a $\log_2$ ratio $\geq$ 0.15 were defined as copy gain regions.

Genome-wide Allelic Imbalance (AI) was assessed using the counts of A, C, G, and T nucleotides in the tumor at positions called heterozygous in the matched normal sample. The most common nucleotide was called “Allele B” and the sum of the counts for the other three nucleotides was taken as the frequency of the “Allele A”. The raw B-Allele Frequency (BAF) was calculated as BAF = 2/pi * atan(B counts / A counts) (Peiffer et al. 2006). BAF was converted to modified BAF (mBAF) by reflecting it around the value 0.5 (Diskin et al. 2008). mBAF values were averaged in the same 50kb bins used for copy number above. These binned mBAF values were segmented using CBS and the same parameters used for copy number (Diskin et al. 2008).

Segments with a mBAF value ≥ 0.75 were considered as Allelic Imbalance. Segments with AI and without copy gain were said to have Loss Of Heterozygosity (LOH).

The pattern of CNVs observed in these samples is consistent with those observed by others (Farazi and DePinho 2006), Among the most frequently reported copy number changes in HCC patient (Farazi and DePinho 2006), we observed recurrent copy number loss of chromosome 4q13, 8p23,17p12-13, 19p13 as well as gain of 8q in all four patients. Notably, TP53 copy number loss was found among all four patients and the copy number gain of CCND1 gene was found in the HBV negative sample. In addition, the copy number loss in 4q21, 4q318p12, 8p21, 14q, 16q, as well as copy number gain of 1q were only found in HBV positive samples. The allele-imbalance regions are largely overlap with regions of gain and loss in addition to broad regions of copy-neutral LOH (Fig. 6).

Section 13: Association of MIS with copy number boundaries

Out of the 9 MIS from the three patients, 6 integration sites were within 21 Kb of copy number boundaries and the rest were more than 900 Kb away (Supplemental Table 9). The boundaries themselves were assigned based on copy number segmentation, and had a resolution of 50 Kb. Thus the MIS coincided with copy number boundaries within the segmentation resolution limits. To test the significance of this association, we ran simulations, where we randomly selected 9 genomic positions mimicking integration sites, and measured their distance from the closest copy number boundary. To select 9 random loci, we first sampled 9 random chromosomes based on a multinomial distribution with probability of each chromosome proportional to its length. Then from each selected chromosome, we randomly selected a position, using the uniform distribution, from a set of positions that did not contain any assembly gaps. In each simulation, we counted how many positions were within 21 Kb of copy number boundaries (from a single patient). Out of 10,000 simulations, we observed 8226, 1620, 147 and 7 simulations that contained 0, 1, 2 and 3 positions within 21 Kb respectively. None of the simulations showed more than 3 positions within 21 Kb of the copy number boundaries. In another simulation of 100,000 runs, we observed 82611, 15970, 1353, 63 and 3 cases that contained 0, 1, 2, 3 and 4 positions within 21 Kb respectively. These results show that the probability of obtaining 6 or more integration sites within 21 Kb of copy number boundaries, as observed in our data, is statistically significant ($p < 10\text{-}5$).

Section 14: Differential gene expression from transcriptome sequencing

Statistical analyses of differentially expressed genes were performed using a method based on the negative binomial distribution as implemented in the DESeq package from Bioconductor (Anders and Huber 2010). Briefly, read counts for all Entrez genes from all samples were summarized into a count table, size factor was estimated for each sample to account for different numbers of total reads, the variance for each gene within the tumor or normal class was estimated based on the negative binomial distribution, and the nominal p-values and Benjamini-Hochberg adjusted p-values were derived. In total, we found 396 genes showing differential expression between tumor and the normal liver tissue (Supplemental Table 10) with a very stringent threshold of Benjamini-Hochberg adjusted p-values ≤ 0.01 and absolute log2 fold change ≥ 5. Functional enrichment analysis using Ingenuity Pathway Analysis showed that the list of differentially expressed genes was consistent with genes previously identified as related to liver cancer (data not shown).

Supplemental Figures Legend

Supplemental Figure 1. Histopathological features of sequenced samples.

Samples from all four patients are represented including non-neoplastic tissue (*A, C, E* and *G*) and hepatocellular carcinoma (*B, D, F* and *H*). In the non-neoplastic samples, chronic active portal inflammation is observed (asterisks) with some evidence of periportal fibrosis. The hepatocellular carcinoma samples exhibit cells of hepatocytic morphology but with growth in cords more than 3 cells thick and with evidence of pleomorphic nuclei. The scale is the same for all panels and the scale bar is shown in the panel *A*.

Supplemental Figure 2. Viral integration detection at variable depth coverage using ultra-depth whole genome sequencing.

Patient 31656 was sequenced at ultra depth coverage (>240-fold) for both tumor and non-tumor tissue. (*A*) Non-redundant insertion events in the patient were identified after clustering of reads representing the same event. Two highly abundant integration sites can be seen in the tumor against a background of widespread integration events in both the tumor (red) and non-tumor tissue (blue). (*B*) We randomly sampled reads from this data at various depth coverage levels, and quantified the number of unique viral integration events. The mean value of 100 simulations at each coverage level is shown with the error bars representing the sample standard deviation of these 100 simulated integration events. The number of events detected is roughly linear and not substantially saturated at even 240X coverage.

Supplemental Figure 3. Distinct pattern of the fraction of chimeric reads between tumor and non-tumor samples.

At HBV integration loci supported by multiple reads, fraction of alleles supporting the integration event are compared between tumor and normal samples. The chimeric fraction (chimeric reads / (chimeric reads + normal reads)) is plotted against total number of reads (chimeric reads + normal reads). Tumor samples are shown in red, tumor-matched adjacent tissue is shown in blue. Data from three HBV positive patients is shown. A cluster of major integration sites (MIS) is highlighted on the top-right corner of the plot.

Supplemental Figure 4. HBV integration within the *MLL4* gene.

HBV integration within the third exon of *MLL4* gene was evident both by DNA and RNA sequencing. (*A*) Expression level of *MLL4* for both tumor (red) and normal (blue) samples for three HBV positive patients is shown based on RNA-seq data. The top track is the tumor sample (patient 31107) with HBV integration, which shows over 20-fold up-regulation of overall expression level of *MLL4* gene compared with the samples without HBV integration. Detailed sequence analysis revealed there are two copies of HBV integrated at this locus. Fusion transcripts were detected at both human-viral junctions. The 5' fusion transcript is on the reverse strand of *MLL4*, and only covers part of its third exon. It shows more significant upregulation compared to the 3' fusion transcript. (*B*) The 3' fusion transcript is in the forward orientation in-frame fusion that covers 4th to 37th exons of *MLL4*.

Supplemental Figure 5. Novel transcripts due to HBV integration.

The two most abundant HBV integration sites in the tumor sample from patient 31656 mapped to non-genic regions. Novel fusion transcripts were detected by RNA-seq (**A** and **B**). In case (*A*), the viral insertion site also precisely co-localizes with the breakpoint of a large deletion on chr11q22, which leads to the copy number loss of a cluster of Caspases and Card-domain containing genes (Fig. 3A).

Supplemental Figure 6. Whole genome and transcriptome read depth coverage of the HBV genome.

Normalized read depth coverage at each base position (Number of reads mapping to this position / total number of viral bases sequenced) was plotted along the viral genome based on both DNA-seq (left panels) and RNA-seq (right panels). Tumor data is shown in red and matched liver samples are shown in blue. The transcriptome coverage (right panels) shows that the viral core

gene is not expressed in any of the samples we sequenced. Differences in gene expression can be observed between tumor and normal sample from the same patient at the HBx gene, and the pre-S region of HBsAg.

Supplemental Figure 7. Proximity of integration sites to human genomic features.
All HBV integration sites based on whole genome sequencing were chosen (255 sites), and their shortest distance from the boundaries of human genomic features was measured. The distribution of these distances for the tumor (T) and non-tumor (N) samples was then compared to a set of randomly selected genomic positions (R). Proximity to the following features is shown: Genomic repeats (Panels titled 'Alu', 'L1' and 'L2') and known transcripts (Panel titled 'Proximity to transcripts'). Also shown is the distribution of recombination rates at these positions (insertion sites and randomly selected genomic positions) as measured by sex-averaged rates of recombination based on deCODE (Kong et al. 2002) (From the Recomb rate track in the UCSC genome browser, build hg19)**.**

Supplemental Figure 8. Performance of somatic score.
Performance of the SomaticScore used to determine the confidence in a somatic mutation call for single nucleotide substitutions is shown as Receiver-Operating Characteristic (ROC) curves. Novel somatic mutations were chosen for each patient (291 mutations from H442, 283 from 31107, 453 from H384 and 292 from 31656) for validation using the Sequenom technology. The ROC curves are shown with the SomaticScore shown as a color gradient. For patients H384 and H442 the tumor versus blood somatic calls are also validated (T vs N: Tumor versus non-tumor liver tissue, T vs B: Tumor versus Blood).

Supplemental Figure 9. Recurrently mutated genes in HCC patients.
Genes that had any novel tumor-specific insertions, deletions or high confidence substitutions leading to non-synonymous changes, stop gains, splice site changes, or frameshifts were overlapped between different patients to get recurrently mutated genes. *TP53* was mutated in all four patients, whereas *LAMA2* and *GOLGA6L2* were the only other genes that were mutated in more than one patient. The HBV negative patient (H384) specifically had mutations in genes related to telomere maintenance and DNA recombination, like *PARP1*, *BLM* and *MLH3*.

Supplemental Figure 10. Validated fusion between *AXIN1* and *LUC7L*.

A 60 Kbp deletion on chr16p13 (patient H442) leads to a fusion transcript between *AXIN1* and *LUC7L*, and deletes *ITFG3*, *RGS11*, *ARHGDIG*, *PDIA2* and last two exons of the *AXIN1* gene. The depth coverage of RNA-seq reads from both tumor (top panel in red) and matched liver tissue (middle panel in blue) is shown. The expression level of *ITFG3*, *RGS11*, *ARHGDIG*, *PDIA2* and the last two exons of *AXIN1was* significantly reduced in the tumor sample when compared to the matched liver tissue.

Supplemental Figure 11. Genome-wide copy number variation in HCC patients.

Somatic copy number change was indicated by Log2 ratio of read depth coverage of the tumor versus matched adjacent liver tissue. Copy number loss of 4q13, 8p23, 17p12-13, 19p13 as well as gain of 8q is a shared feature between all four patients. However, the copy number loss in 4q21, 4q31, 8p12, 8p21, 14q, 16q, as well as a copy number gain of 1q were only found in the three HBV positive tumor samples.

Supplemental Figure 12. *ANGPT1* is upregulated specifically in liver cancer.

ANGPT1 is significantly upregulated in liver cancer, based on two independent data sets: Gene Logic (Gaithersburg, MD) (A, Affymetrix HG-U133 platform, representing 3,600 normal and 1,701 neoplastic samples from different human tissues. Probe set 205609_at was chosen to represent ANGPT1 expression) and public GEO data (*B*, GSE25097 data, probe 100129666_TGI_at). *ANGPT1* is significantly downregulated in other solid tumors, based on Gene Logic data (*C*). Numbers in red are -log10-transformed p-values based on a t-test. Signal intensity was measured based on RMA normalization. Boxplots are shown, with individual data points shown as dots, binned into intervals of 0.1.

Supplemental Figure 13. Copy number loss at *CASP1* in HCC patients.

Copy number of *CASP1*, which is a part of the chr11 Caspase cluster, is shown from two independent data sets (data set GSE34957, *A* and data set GSE9829, *B*). Samples below the green line show copy number loss, while samples above the red line show copy number gain. No copy number gain was observed in the GSE9829 data. Copy number loss was observed in 13.6% of HCC patients in (*A*) and 8.7% patients in (*B*).

Supplemental Figure 14. Transcript fusion breakpoints at nucleotide level.

Contigs assembled using reads putatively containing breakpoints, from each patient, were aligned to the viral strain for that patient. The alignment of the contigs to the viral sequences is shown,

with the human part of the contig in lower case and colored grey. The alignments near the DR1 region are shown, since most of the contigs aligned in this region (1717-1884 bp on the viral sequence). The transcription start positions of the pre-C mRNA are shown. The TATA-box like sequences of the promoters are underlined. The direct repeat-1 (DR1) sequence is marked with a box. The patient and the tumor status of the sample is indicated on the left (T:tumor, N:non-tumor).

Supplemental Material References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**(10): R106.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**(5583): 1003-1007.

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res* **36**(19): e126.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**(5961): 78-81.

Farazi PA, DePinho RA. 2006. Hepatocellular carcinoma pathogenesis: from genes to environment. *Nat Rev Cancer* **6**(9): 674-687.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31**(3): 241-247.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**(2): 265-272.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**(16): 2069-2070.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J et al. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* **16**(9): 1136-1148.

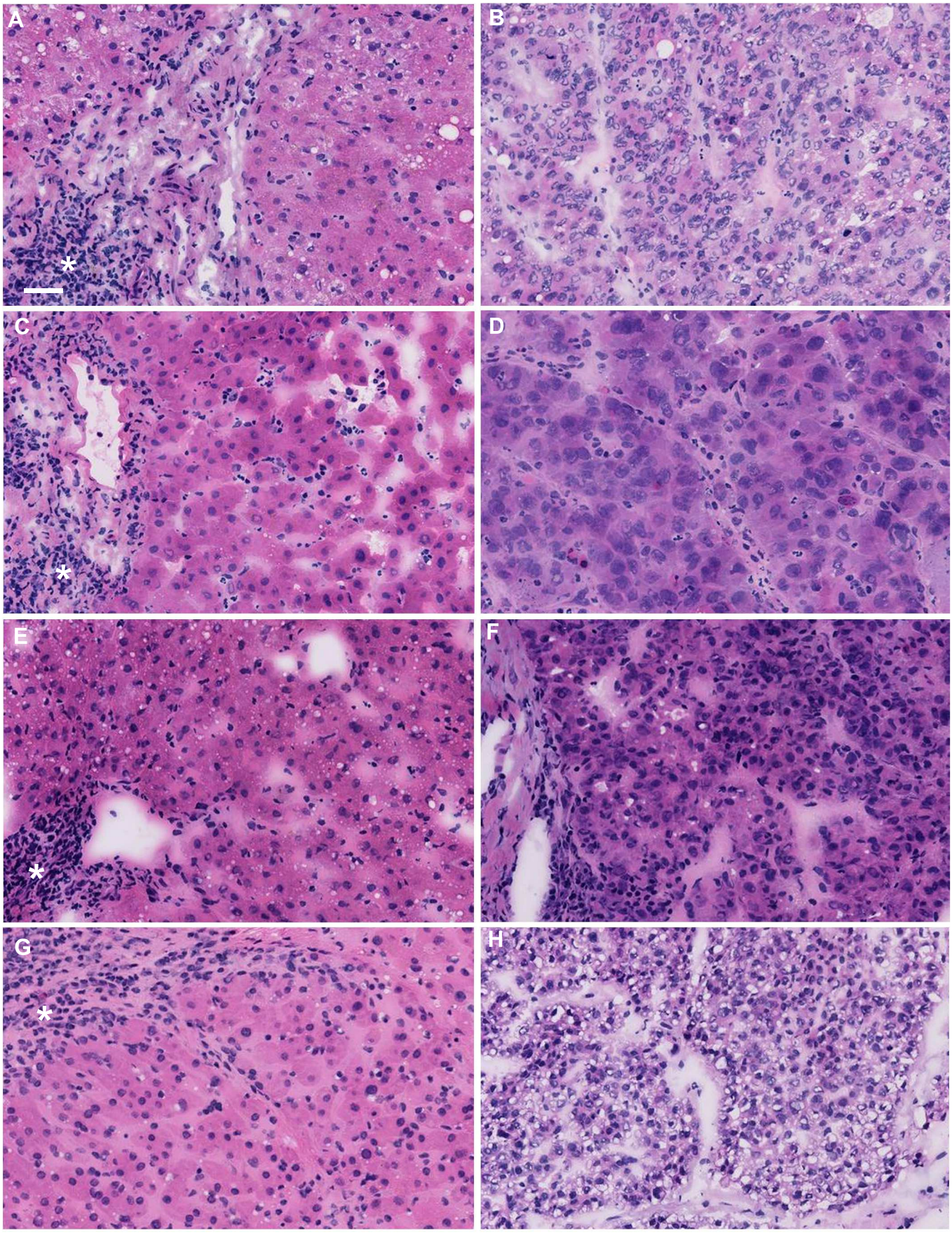
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol* **147**(1): 195-197.

Venkatraman ES, Olshen AB. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**(6): 657-663.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**(7): 873-881.

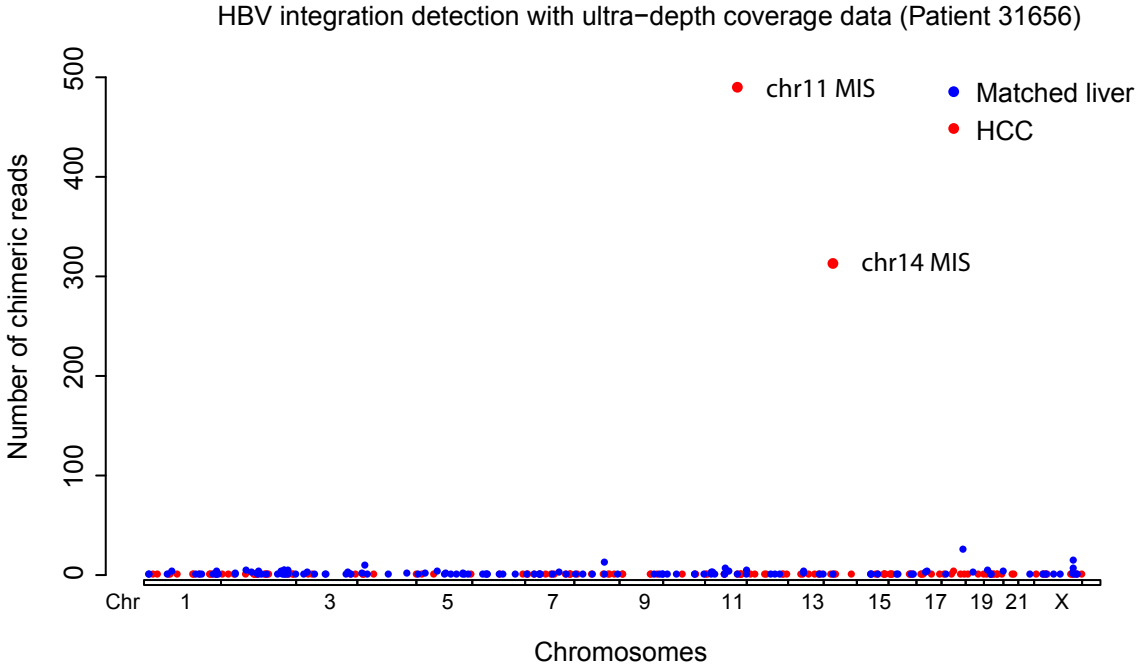
Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**(9): 1859-1875.

Supplemental Figure 1

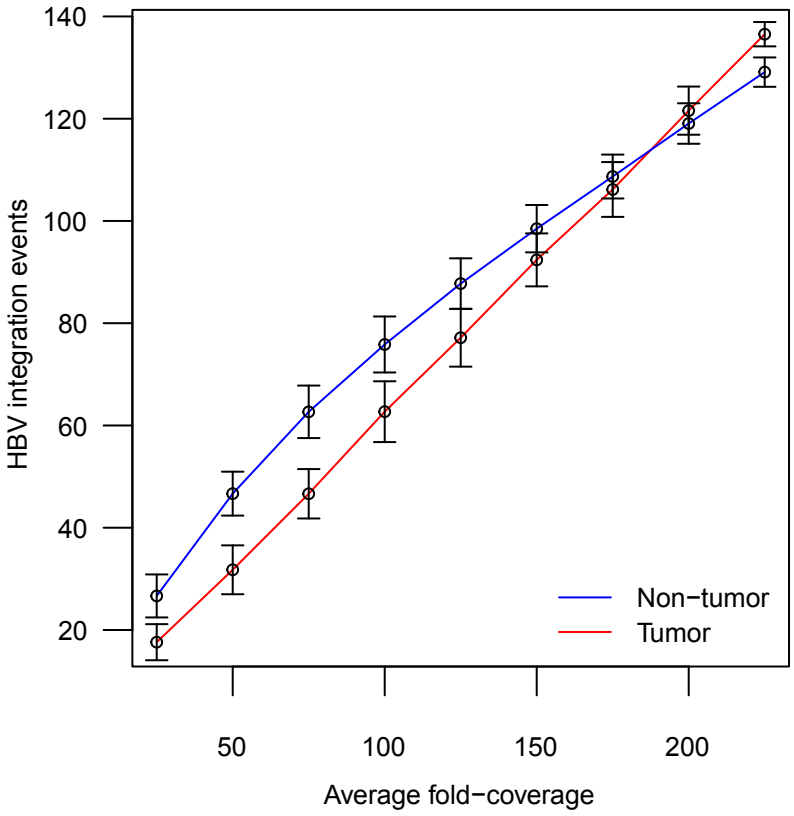


Supplemental Figure 2

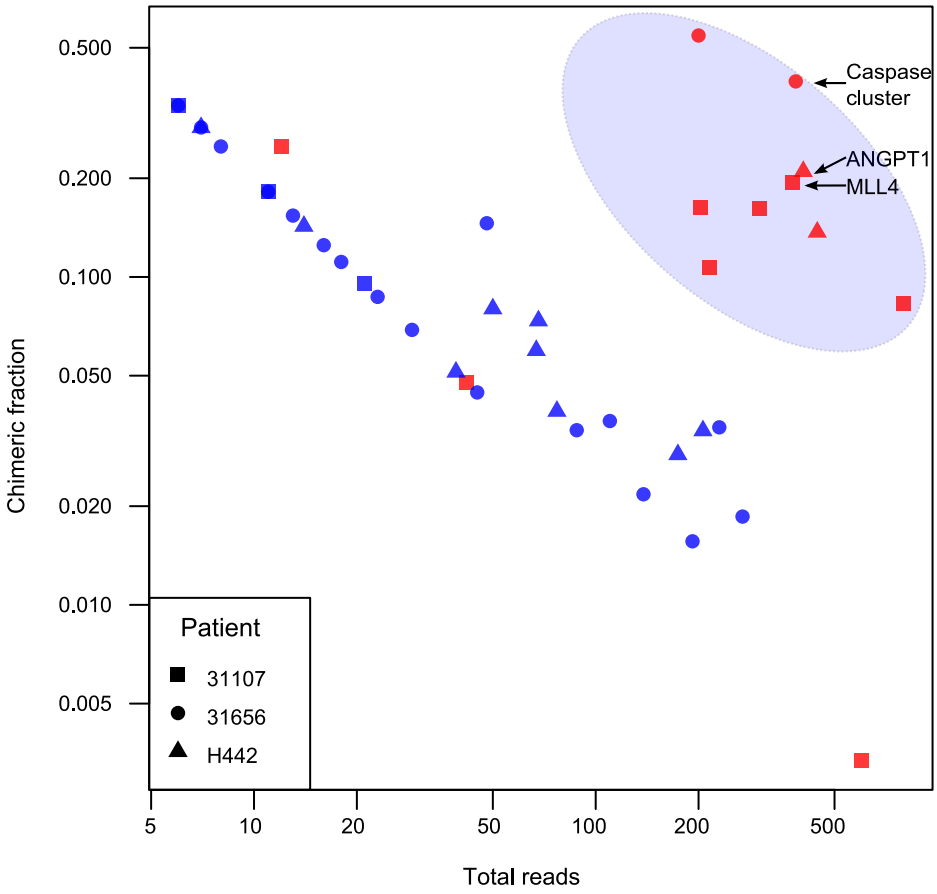
A



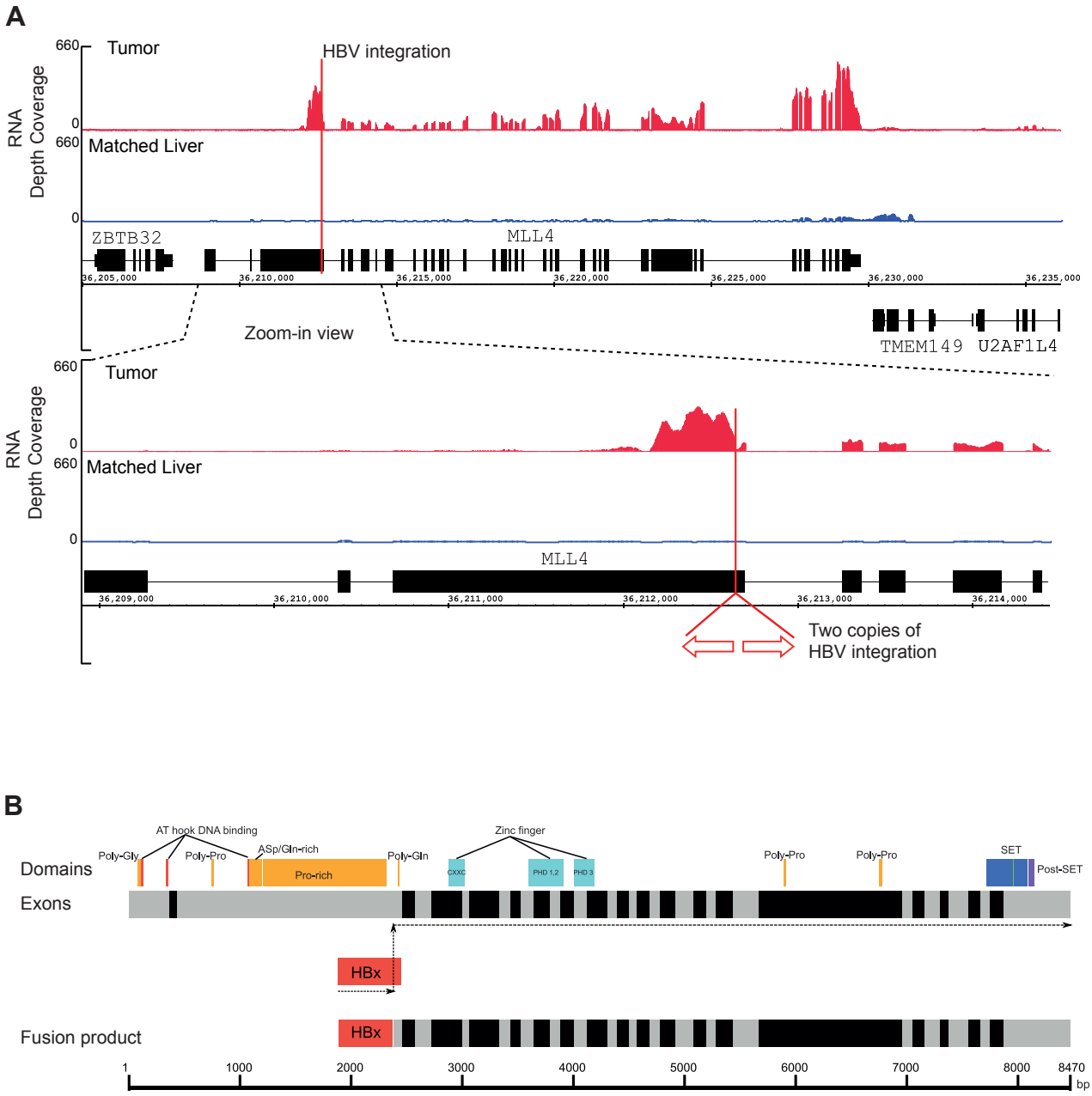
B



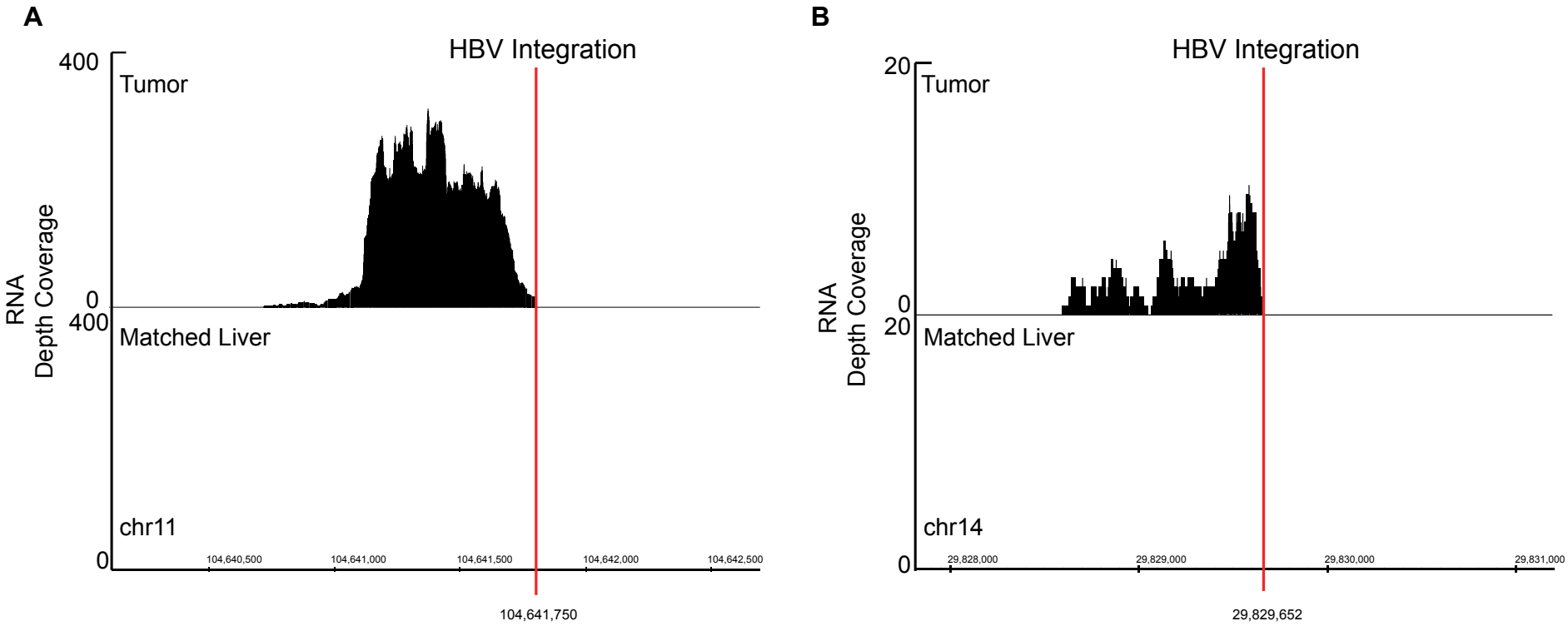
Supplemental Figure 3



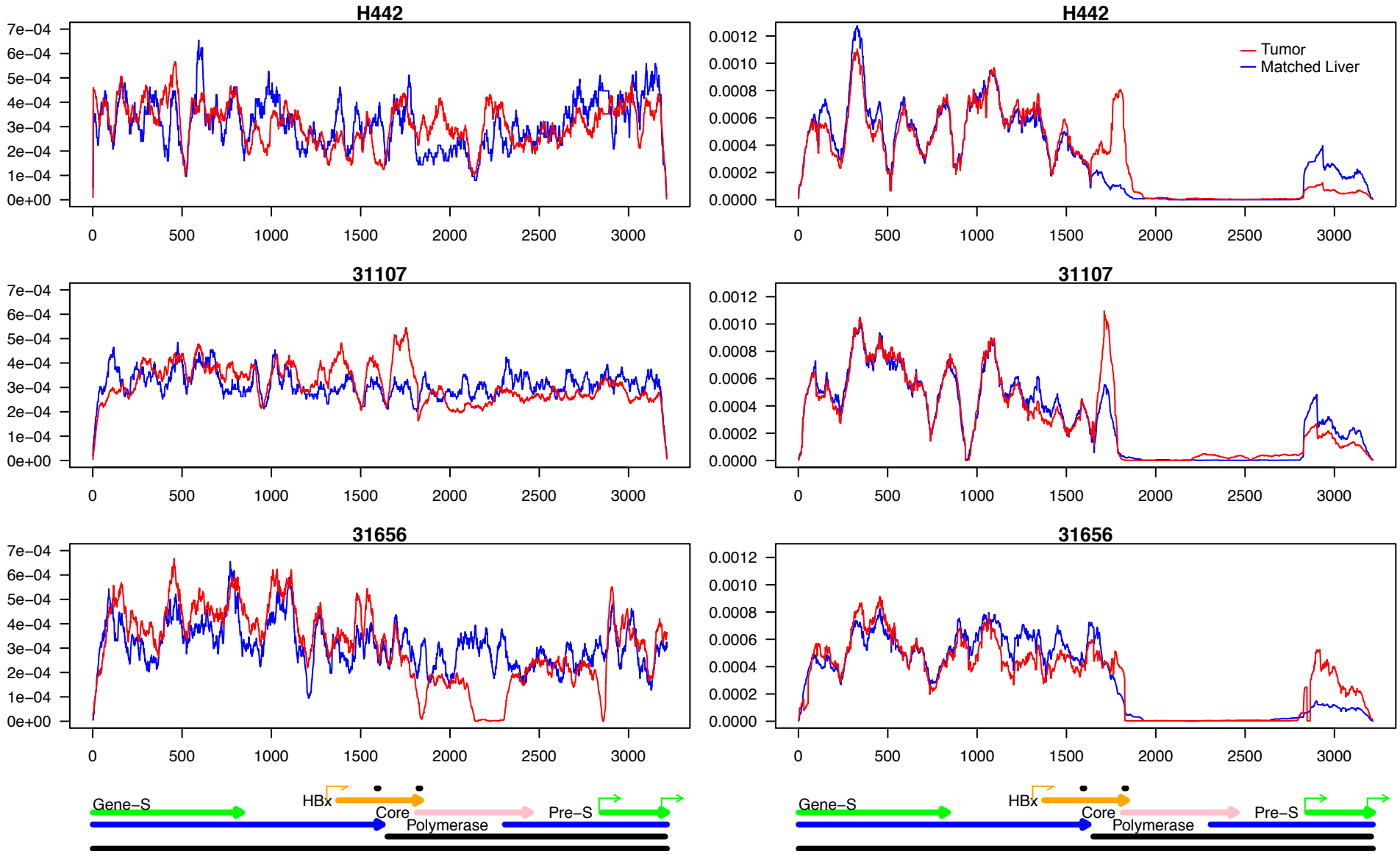
Supplemental Figure 4



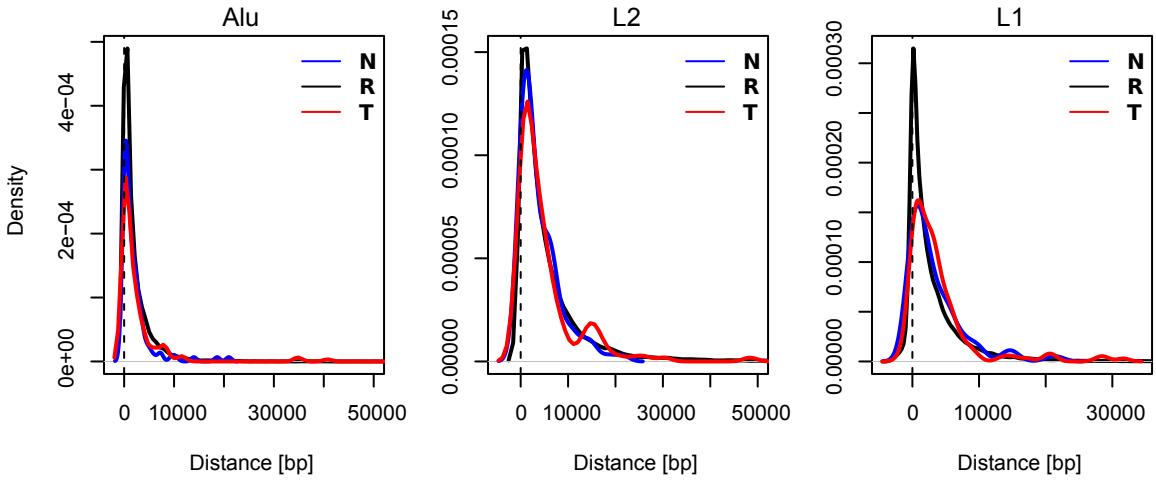
Supplemental Figure 5

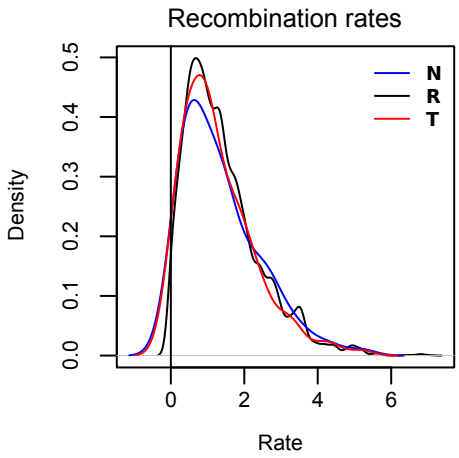


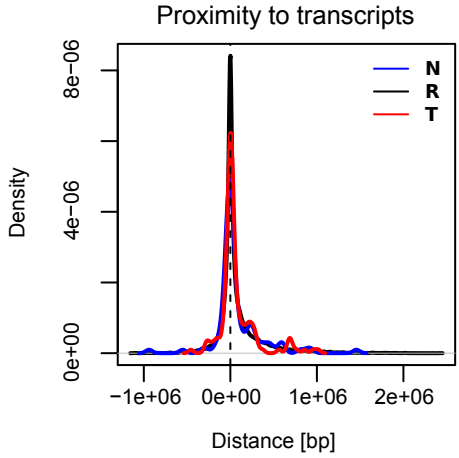
Supplemental Figure 6



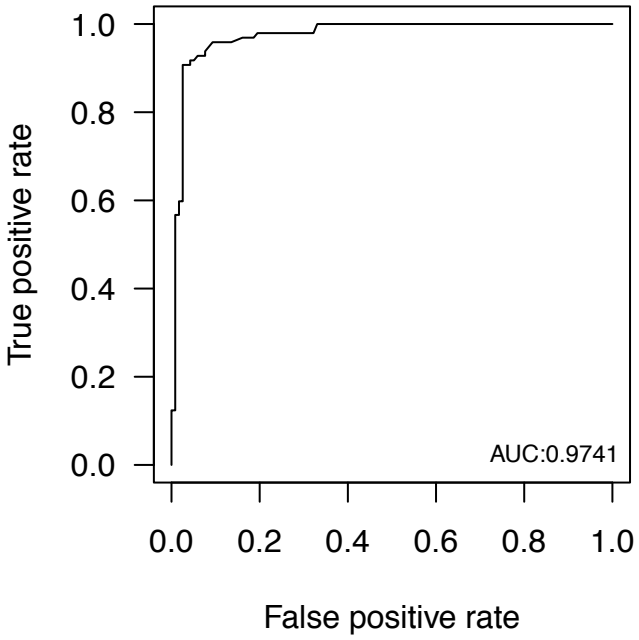
Supplemental Figure 7

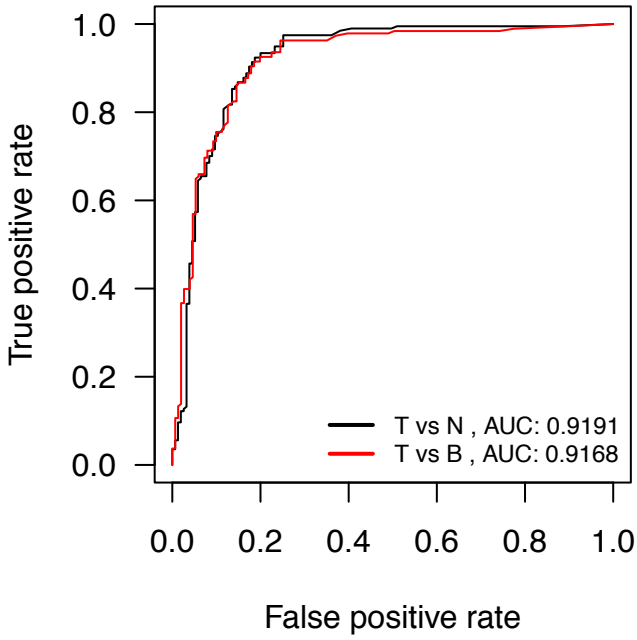


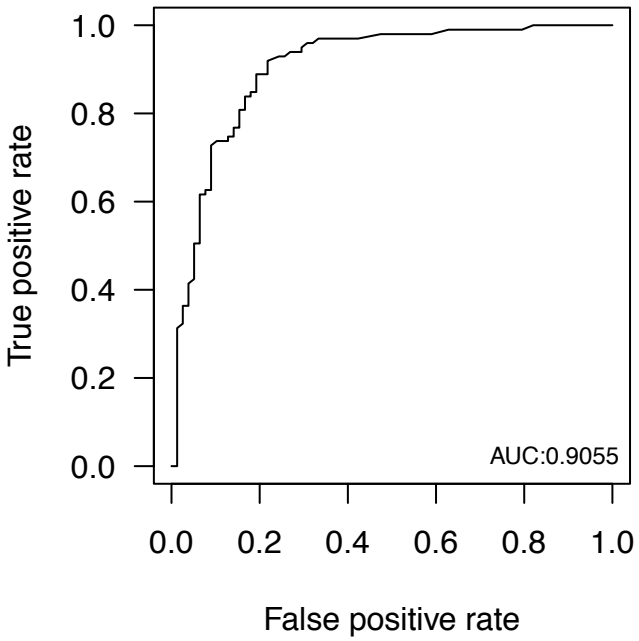


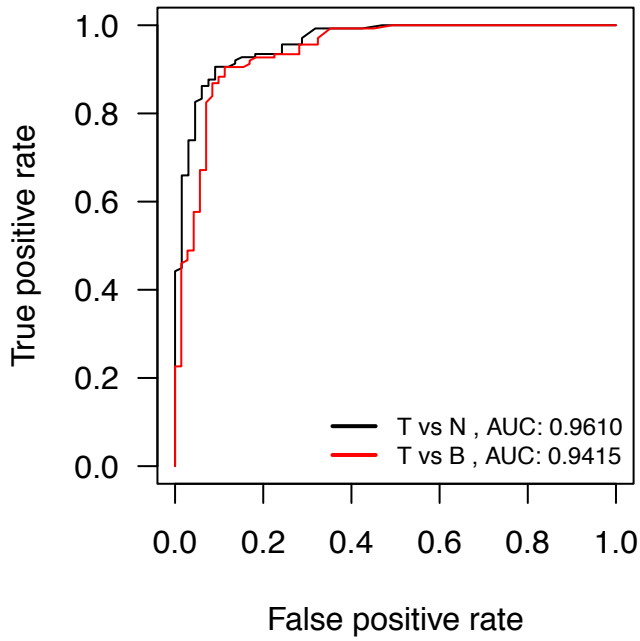


Supplemental Figure 8

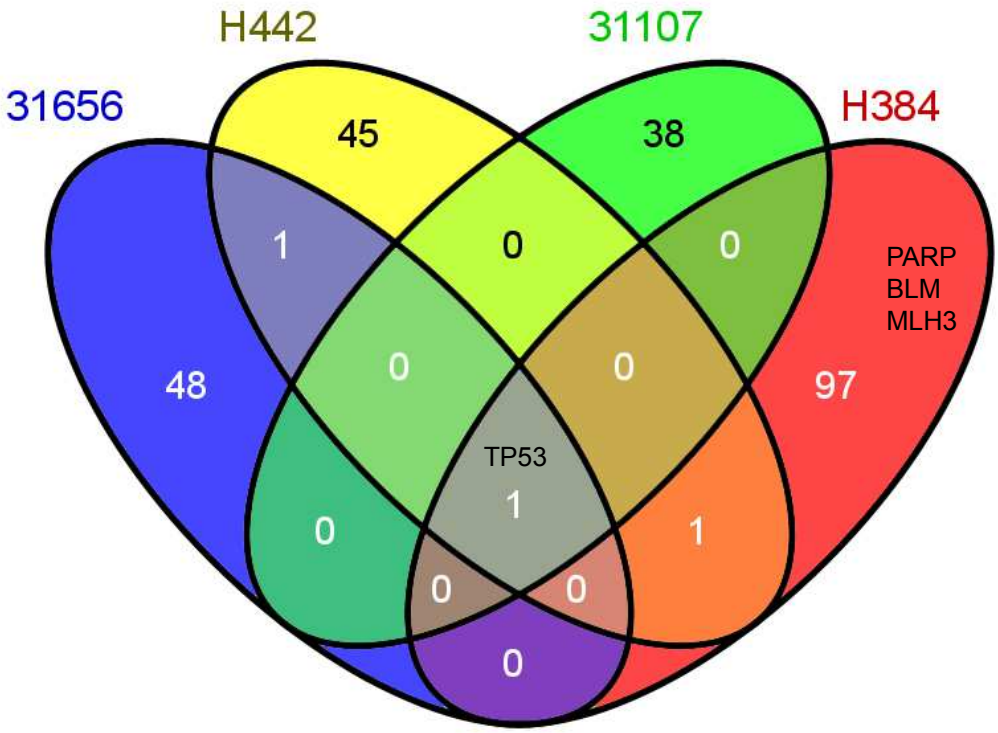




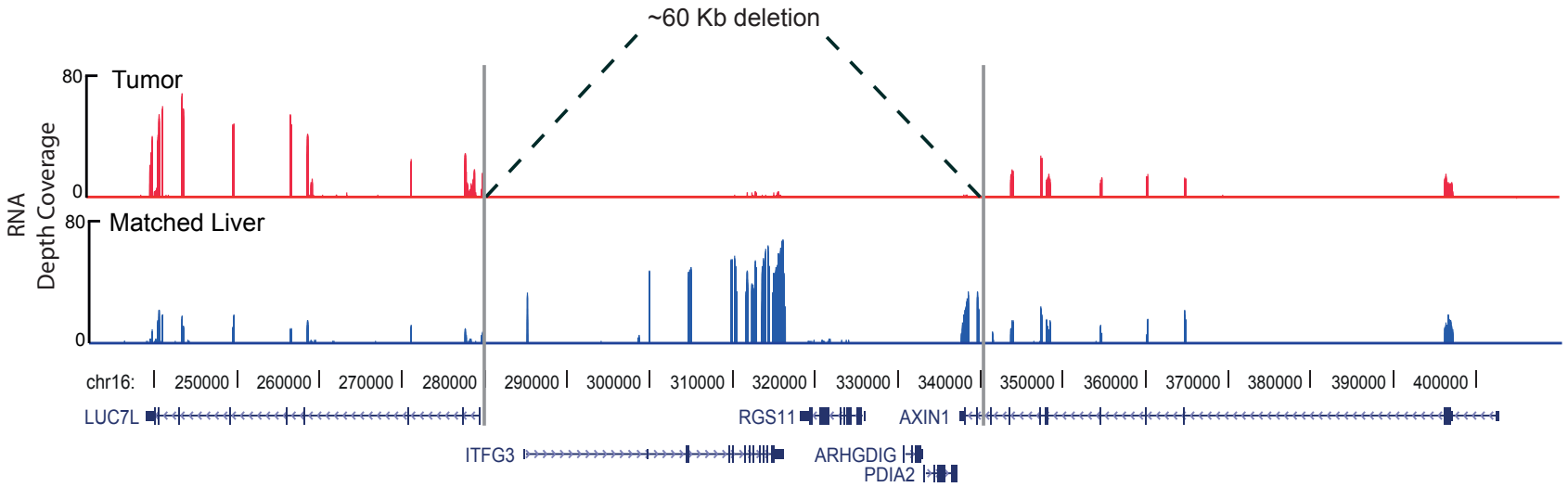




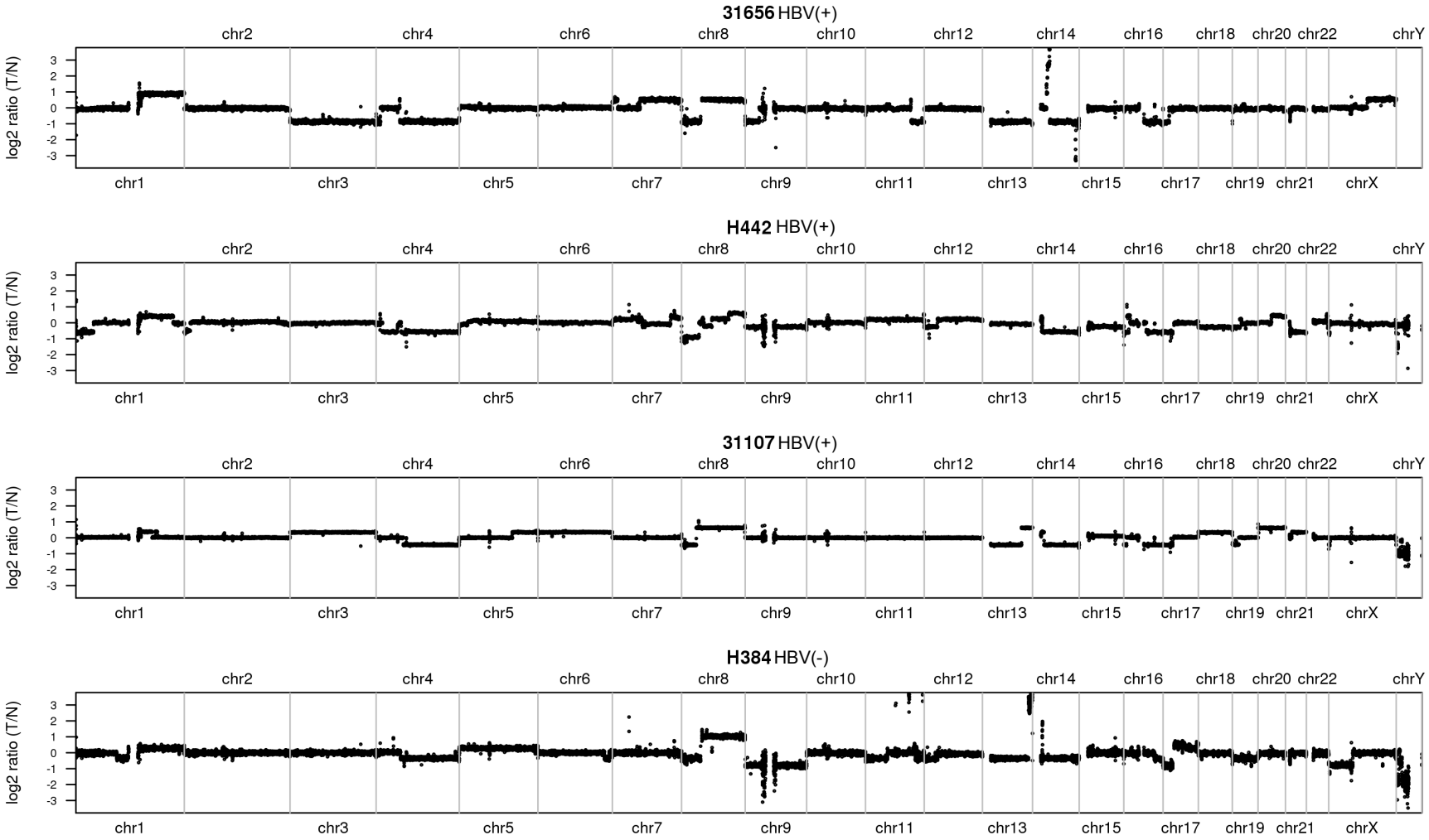
Supplemental Figure 9



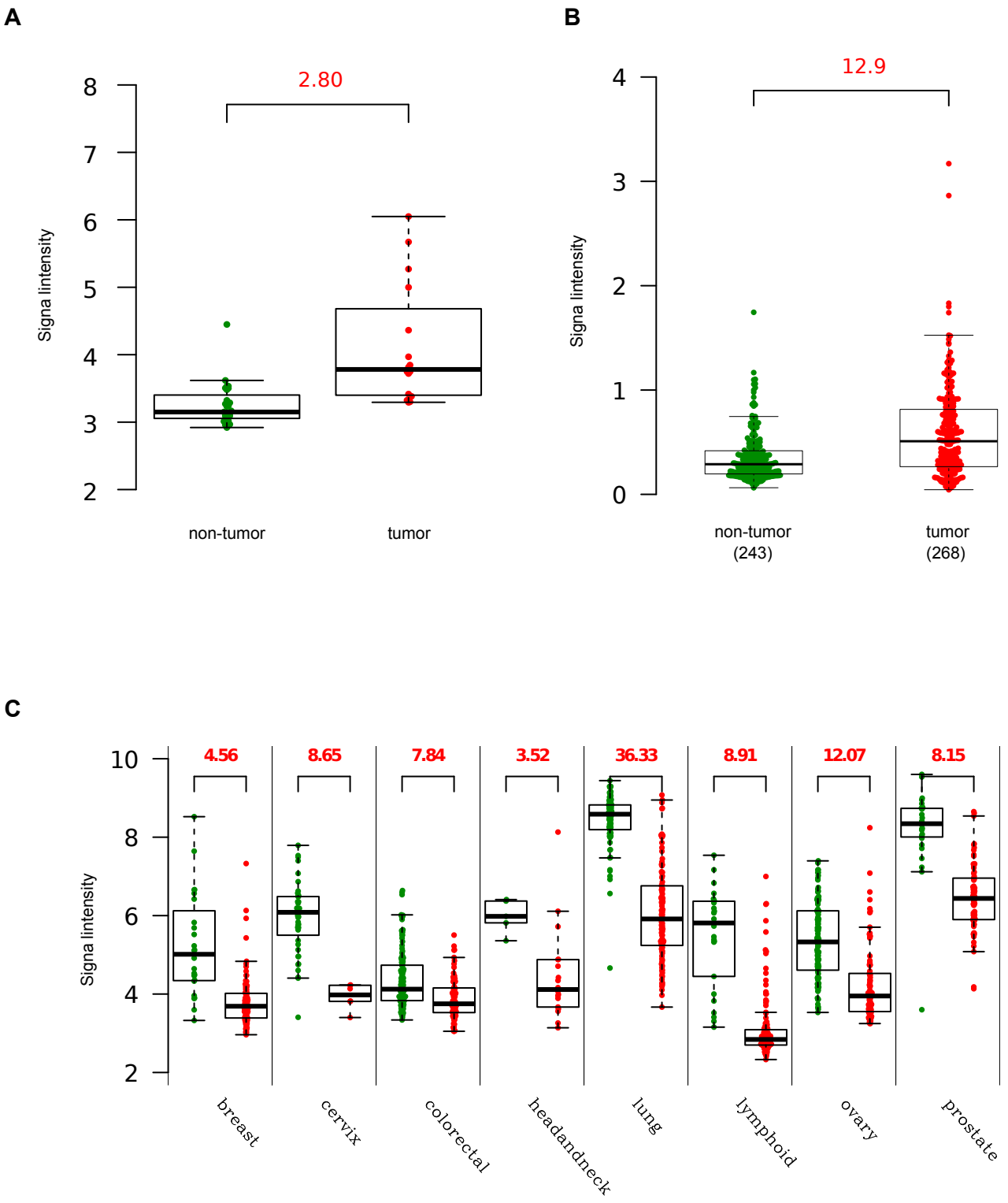
Supplemental Figure 10



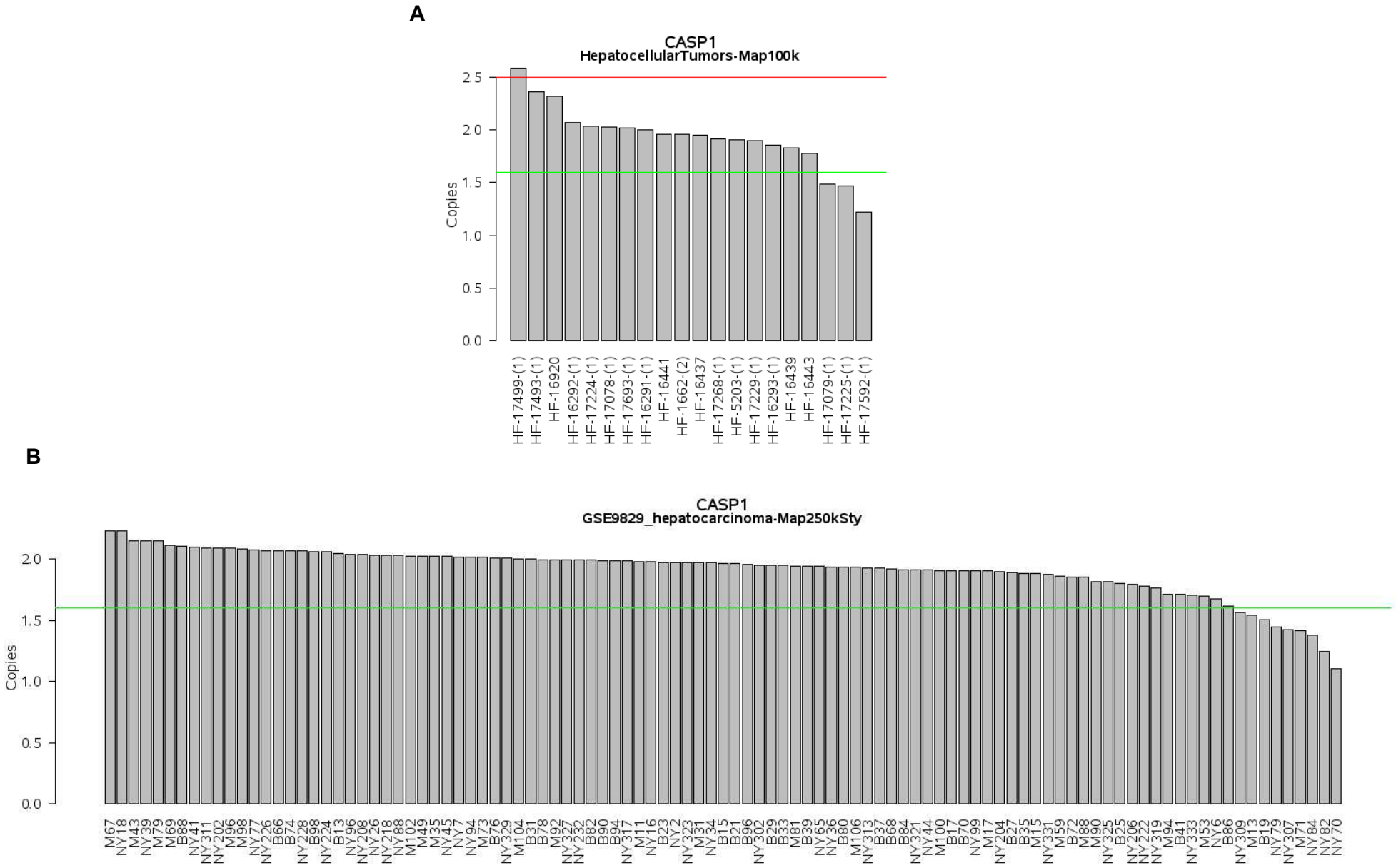
Supplemental Figure 11



Supplemental Figure 12



Supplemental Figure 13



Supplemental Figure 14