A Novel and Integrative Approach to Identify Cancer Drivers for Hepatitis B Related Hepatocellular Carcinoma

Executive Summary

Liver cancer (hepatocellular carcinoma or HCC) is the third leading cause of cancer deaths worldwide. Over 50% of HCC cases are caused by Hepatitis B Virus (HBV) infection, which attacks the liver by integrating its DNA into the human genome. To explore the underlying mechanisms of HBV induced HCC, we aimed to identify HCC "driver" genes - genes whose alterations, such as by HBV integration, can "drive" a cell to a cancerous state. We hypothesized that study of recurrent HBV integration sites in HBV-HCC liver tissue using Next Generation Sequencing (NGS) would allow us to identify HCC driver genes. However, HBV integration sites are extremely difficult to detect because HBV integration is such a low frequency event. To overcome this obstacle, we applied an enrichment method to capture HBV-containing DNA in infected liver tissue before sequencing. The NGS data generated was then used to develop a novel data analysis software, *HccDriverFinder*, to identify HCC driver genes. Through analysis of 146 NGS datasets, *HccDriverFinder* was able to construct a database of 45 potential driver genes, many of which are supported by existing literature. Our software holds promise for precision medicine, as it can be used for designing a cancer treatment plan tailored to a person's genetic profile. Our final product is envisioned as an HCC driver identification kit that can stratify patients by risk for HCC. Work is also in progress to develop *HccDriverFinder* to be applied to other cancers.

Abstract

This study tested two hypotheses: i) HCC-associated HBV integration sites are the most dominant, or "major", junctions (MJs) detected in HBV-HCC due to uncontrolled clonal expansion of cancerous hepatocytes and ii) study of HCC-associated junctions will assist in identifying driver genes and unraveling mechanisms of hepatocarcinogenesis. We first prepared an NGS library from 6 pairs of HCC and adjacent non-HCC tissue DNAs and applied a primer extension capture of HBV-containing DNAs to enrich HBV-on-target reads for detection of MJs. Using the NGS data generated, we developed a novel software, "*HccDriverFinder*", to i) identify MJs (demonstrating that MJs can be detected from as little as 1 million reads), ii) incorporate PubMed data mining for driver identification, iii) detect HBV mutations (revealing a striking 98.8% HCC-linked mutations identified in cancer tissues), and iv) visualize integration events. Through analysis of 140 additional in-house HBV-HCC NGS datasets, *HccDriverFinder* constructed a database of 45 potential HCC driver genes, confirming known targets such as TERT, CCNE1, and FN1 and uncovering new recurrently integrated genes such as CSAD and ABCC13, both of which are reportedly linked with carcinogenesis. This study holds promise for the building of a driver identification kit for HCC drug development and precision medicine.

Introduction

Hepatocellular carcinoma (HCC) is the third leading cause of cancer deaths worldwide and ranks fifth in global cancer incidence, affecting approximately 800,000 people every year (1-2). The prognosis of patients with HCC is poor with a 5-year survival rate of less than 14% (3-4), due to lack of effective therapies. To improve prognosis and patient survival, a better understanding of underlying mechanisms of each HCC is essential for patient selection for drug development and precision medicine when more treatment options become available.

Hepatitis B virus (HBV) infection is a major etiology of HCC, associated with over 50% of the cases worldwide and up to 70–80% of cases in sub-Saharan or Asian countries (5-6). Chronic HBV carriers have greater than 100-fold increased relative risk of developing liver cancer (7). HBV can integrate into the human genome and progressively contribute to genomic instability, likely through rearrangement of chromosomes or insertional mutagenesis, and could thus lead to hepatocarcinogenesis (8-10). Although integration events occur in only 1 out of 1000 infected liver cells, integrated HBV DNA has been found in 85-90% of HBV related HCCs (10-11), suggesting a significant role of HBV integration in hepatocarcinogenesis.

HBV integration is known to occur randomly throughout the genome (1, 10, 12). However, with the recent application of next-generation sequencing (NGS), recurrent HBV integration sites have been identified in HCC tissue (1, 13-14). Most interestingly, these recurrent integration sites are located in genes that play causal roles in oncogenesis, also known as “driver” genes (1, 15). The most reported integration event occurs at the telomerase reverse transcriptase gene (TERT) (1, 16), a gene widely known to be abnormally active in many cancer cells by causing enhanced expression of telomerase (17-18). The occurrence of recurrent integration sites suggests that activation of oncogene(s) by HBV integration can cause

uncontrolled clonal expansion during HCC carcinogenesis. Consequently, these HCC-associated integrations will become disproportionately overrepresented and emerge as "major" junctions (MJs), as illustrated in Figure 1A. The appearance of MJs can therefore be used as an indicator of uncontrolled clonal expansion (a hallmark of cancer) and thus as a potential marker for HBV-HCC.

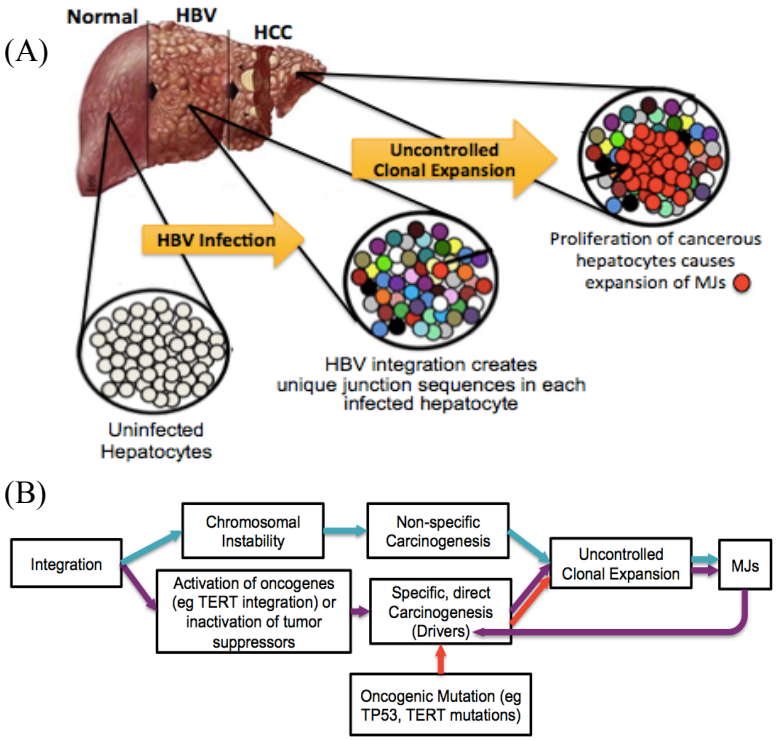

Figure 1: (A) Schematic presentation of the detection of MJs as a marker for uncontrolled clonal expansion. HBV randomly integrates into the host genome, creating a unique integration junction sequence in each hepatocyte it integrates into, as represented by the different colored circles. During tumorigenesis, cancerous cells undergo clonal expansion in which particular junction(s) will dominate by emerging as expanded MJs, as represented by the red circles. Revised from diagram by Lin (19) (B) Overview of HBV integration in HCC carcinogenesis and driver identification through MJs.

We hypothesized that i) HCC-associated junctions are the most dominant, or "major", junctions detected in HBV-HCC and that ii) through the detection of MJs in a broad sample of HBV-HCC NGS data, we can systematically identify driver genes for HCC to build an HCC driver database and outline possible mechanisms of hepatocarcinogenesis, as illustrated in Figure 1B, for precision medicine. To test our hypotheses, as outlined in Figure 2, we first constructed NGS libraries from 6 pairs of HCC and adjacent non-HCC tissue DNAs. In order to reduce the cost and computational burden of the analysis of millions of NGS reads for identifying HBV-host junctions, we applied a primer extension capture of HBV-containing DNAs to increase HBV on-target reads. We then used the ChimericSeq software (20) to identify chimeric reads from the NGS data. To facilitate efficient analysis of the vast amount of chimeric reads

identified, we developed a novel program, *HccDriverFinder*, to i) identify MJs, ii) perform PubMed (21) data mining to determine carcinogenic relevance for the building of an HCC driver database, iii) detect HCC-associated HBV mutations, and lastly iv) visualize breakpoints in the HBV genome for analysis of the potential impact of integration events. In the future, we can use the analysis performed by *HccDriverFinder* to generate a risk score for HCC screening if MJs can be detected in the peripheral from HBV infected patients. After identifying MJs from our 6 pairs of HCC and non-HCC tissue samples, we validated the NGS identified MJs by PCR-Sanger sequencing and confirmed our computational findings with published literature. Our final product is envisioned as an HCC driver identification kit, made up of our HBV enriched NGS platform and *HccDriverFinder* software, which can provide cancer genetics for drug development and precision medicine and potentially inform chronic HBV-infected patients if they have or are at risk for HCC if our kit has the sensitivity to detect MJs from the peripheral.

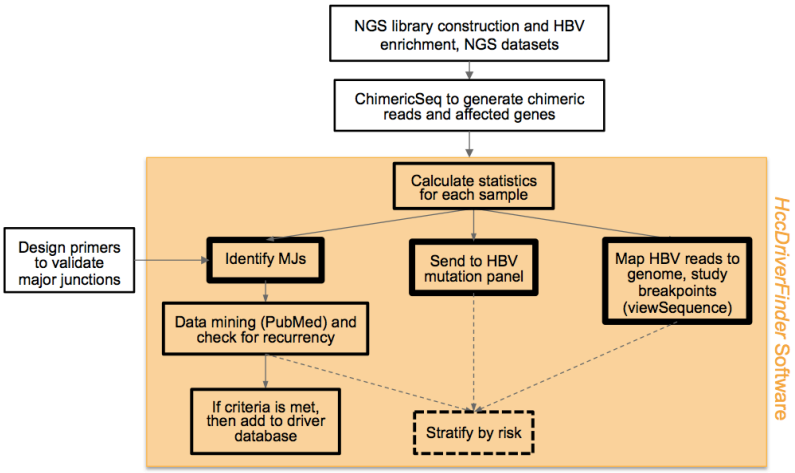

Figure 2: Schematic overview of workflow. Work is in progress to use analysis performed by DriverFinder to stratify patients by risk for developing HCC (as indicated by the dashed lines)

Materials and Methods

Study Subjects and Specimens

Archived, non-identifiable DNA samples of six paired and archived HCC and adjacent, non-HCC tissue, from five males and one female, ages ranging from 43-69 with HBV chronic

infection, collected from another previous study, were obtained from JBS Science Inc. (Doylestown, PA) and used in this study.

Preparation of HBV-Enriched NGS Data from 6 Pairs of HCC and Non-HCC Tumor Tissue

Outline of the process is shown in Figure 3. The DNA was fragmented by sonication to less than 500bp using Misonix Sonicator® XL2020 (Misonix Incorporated, Farmingdale, NY) for a total elapsed time of 20 minutes. NGS library DNA preparation was performed as described by Ding et al (13). Modifications were made including the use of high-fidelity Herculase II fusion DNA polymerase (Agilent Technologies, Santa Clara, CA) for PCR amplification. To enrich for HBV-containing DNA, a biotin HBV primer extension capture approach was conducted using the amplified library DNA in a reaction containing 25mM dNTP mix, 1x Herculase II Buffer, and 20 pmol of 43 biotinylated HBV primers that span the entire HBV genome as shown in Figure 3C. The DNA was collected as described by Gnirke (22) and subjected to two cycles of enrichment processes. In order to verify that our libraries were not lost in enrichment, we amplified about 5% of each sample for an extra 30 cycles to be seen on a 2.2% Lonza gel with a quantification-ladder (Lonza Group Ltd, Basel, Switzerland). The samples were then quantified with real-time quantitative PCR (qPCR) assays previously developed in the laboratory (HBV 1633-1680, HBV

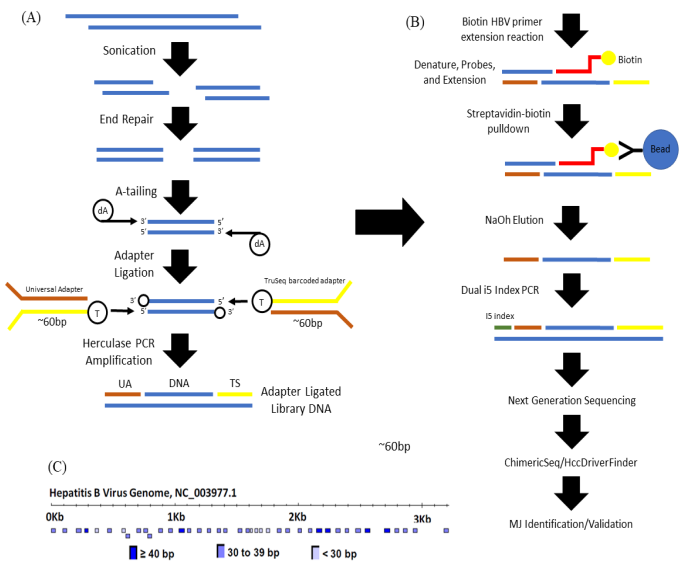

Figure 3: (A) Illustration for the preparation of an adapter ligated DNA library. (B) The enrichment process followed by data analysis and junction verification. (C) 43 HBV-specific capture primers spanning the entire genome. Revised from diagram by Lin (19).

1741-1791, Chromosome 1, and *TP53)* before enrichment and after the second enrichment in order to assess the efficiency. Index PCR was performed to quantify the enriched library for even pooling of each sample for one NGS run (150 bp paired-end reads on the Illumina MiSeq platform (Cornell Institute of Biotechnology, Ithaca, NY)). The LightCycler 480 (Roche, Germany) was used and conditions were 95°C for 5 min (95°C 10s, annealing (Table 1) for 10s, 72°C 10s) for 45 cycles, followed by (95°C 5s, 65°C 60s, 97°C continuous). NGS was performed and data was analyzed by ChimericSeq (19).

Table 1: Primer sequences and annealing temperatures for qPCR assays used in this study.

Assay	Forward Primer	Reverse Primer	Anneal Temp
HBV 1633-1680	AGGTCTTGCCCAAGGTCTTAC	TTGCTGAGAGTCCAAGAGTCC	60°C
HBV 1741-1791	TRGGGGAGGAGATAAGGTTAAAGGTC	ATGCCTACAGCCTCCTAGTACAA	65°C
Chromosome1	AGAGCAGACTTGAAAACACTCTTTTTG	TACCATTGACCTCAAAGCGG	54°C
p53	CTGCATGGGCGGCATG	TGAGGATGGGCCTCCGG	58°C
Index	AATGATACGGCGACCACC*G*A	CAAGCAGAAGACGGCATACGA	63°C
*phosphorothioate bond			

Generation of Chimeric Reads from HBV-HCC NGS Data via ChimericSeq

To detect chimeric reads, the mate paired NGS reads (Fastq format) were input into ChimericSeq, a free software that serves a wrapper function for Bowtie2 (23) in its identification of chimeric reads from NGS data. Each read within the input file was first aligned to the HBV genome and then to the human genome (hg38) using Bowtie2's Burrows-Wheeler Transform alignment algorithm. Using local alignment mode, reads that aligned to the HBV reference genome and contained a mapped portion above a threshold length (30 bp) were extracted, and all other reads were discarded. In completion, ChimericSeq generated a set of comma-separated values (csv) files with annotated information corresponding to each identified chimeric read, SAM files (tab-delimited text files) with alignment information from Bowtie2, and log text files with recorded timestamped documentation of events in the run.

Software "HccDriverFinder" development

HccDriverFinder was designed to be used in conjunction with ChimericSeq to identify and annotate MJs, potential drivers, and HCC-associated HBV mutations to report the risk score (in progress) for HCC.

The *HccDriverFinder* program was written in Python. The Pandas (24) and OpenPyxl (25) packages were used to create and manipulate Microsoft Excel documents (.xlsx). Biopython (26), an open-source collection of Python tools for computational biology, was utilized to parse Fasta files and run NCBI Blast (27). The Pysam module (28) was used for reading and writing SAM-formatted alignment files (29). Tkinter, Python's most commonly used GUI package (30), was used to create and implement the user-friendly interface. To create html generated graphic displays of our data, we modified an existing in-house software from our laboratory called viewSequence for sequence alignment and visualization.

Validation of MJ NGS reads by PCR-Sanger Sequencing

Figure 4: Schematic presentation of primer design. Human DNA shown in yellow and HBV DNA shown in red.

Major HBV junction reads identified from NGS were validated by designing specific primers to amplify the region covering the integration site as illustrated in Figure 4. Four primers (two forward for host and two reverse for HBV) were designed by Primer 3 (31) and optimized for each of the 12 junctions selected. The original tissue DNA was amplified by PCR (95°C 5 min, 40 cycles of [95°C 30s, annealing (Table 4) for 30s, 72°C 30s], and 72°C 5 min) and PCR products with the desired size were purified and Sanger sequenced by the Children's Hospital of Philadelphia as validation of the junction site.

Results

1.) Preparation of HBV-Enriched NGS Library

NGS library was prepared from 6 pairs of HCC and adjacent non-HCC tissue samples and enriched for HBV-containing DNAs as outlined in Figure 3. To ensure that our samples were

not lost in the many steps of library preparation and enrichment, we periodically examined the reactions from our prepared library (Figure 5A), first enrichment (Figure 5B), and second enrichment (Figure 5C) as described in the Materials and Methods.

Overall the samples had evenly distributed smears in sizes as expected (200-500bp template + 120bp (2 adapters) = 320-620bp) in all three gels, verifying the success of adapter ligation. Initially sample #4 had an adapter-adapter dimers at 120bp likely due to little DNA input. After two enrichments, the smear of sample #4 became more evident indicating that the adapters were not enriched.

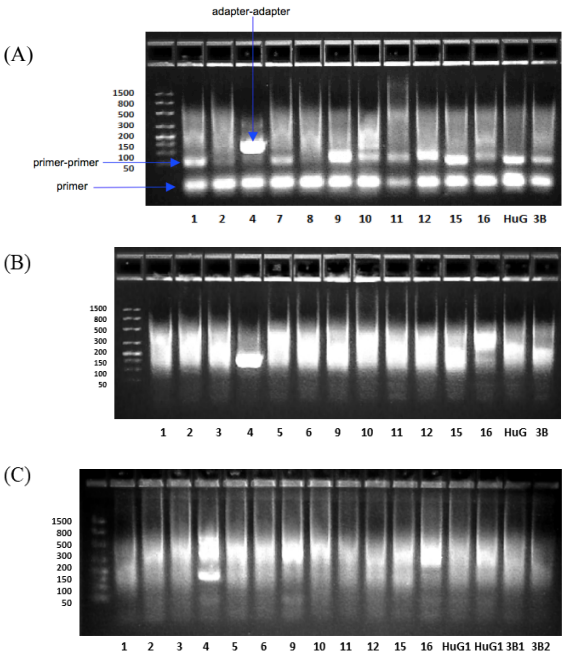

Figure 5: (A) Verification of existing libraries in each sample and control. Gel shows library constructed DNA after first enrichment (B) and second enrichment (C).

In order to feasibly detect junctions, we estimated that we would need at least 1% HBV reads, given one million NGS reads per sample. Based on the difference in size between the HBV (3.2 kb) and human ($3x10^6$ kb) genomes, there would theoretically be ~0.0001% HBV if each infected hepatocyte contained one copy of HBV. To obtain 100 HBV reads for every sample of one million reads, we therefore need a 100-fold HBV enrichment. Enrichment fold analysis was done by comparing the ratio of the two HBV regions (A: 1633-1680 and B: 1741-1791) to the two human regions (Chromosome 1 and *TP53*)

Table 2: Total enrichment folds

Patient	Tissue	HBV A/Chr 1	HBV A/p53	HBV B/Chr 1	HBV B/p53
S5	K	27	72	23	62
	N	781	10	394	5
S9	K	658	17	635	17
	N	5	1,012	0	11
S17	K	31,677	801	64	2
	N	8,352	11,312	265	359
S37	K	56	213	499	1,904
	N	2,230	57	3,636	93
S42	K	3,195	33	1,969	20
	N	7	2	952	299
S48	K	239	67	109	31
	N	43	4	28	3

K: HCC, N: Adj, non-HCC; HBV A = 1633-1680, HBV B = 1741-1791
Enrichment folds > 100 are highlighted.

$$\text{Fold enriched} = \frac{(\text{HBV DNA/Host DNA}) \text{ after enrichment}}{(HBVDNA/HostDNA) before\ enrichment}$$

after and before each enrichment (Table 2). Since 10 of the 12 samples reached the 100-fold threshold, we proceeded to the index PCR to quantify the amount of each enriched library and pooled for one paired-end NGS sequencing on Illumina's Miseq platform.

2.) Development and Implementation of "*HccDriverFinder*" Software (overview)

The chimeric reads generated from 12 HBV-HCC and non-HCC NGS datasets through ChimericSeq (as described in Materials and Methods) were used for the development of *HccDriverFinder*, a program for i) identifying MJs and drivers (as detailed in sections 2.1 to 2.3), ii) detecting HBV mutations (section 2.4), and iii) mapping/visualizing reads in the HBV genome to study HBV breakpoints for possible biological impact on interrupted genes (section 2.5). The overall workflow of the pipeline is shown in Figure 2.

2.1) MJ Identification and Validation

HccDriverFinder uses NGS data from ChimericSeq output files to identify major junctions (as delineated in Figure 6 and described below) and build an interactive data summary table (Table 3). *HccDriverFinder* first calculates the number of NGS, HBV mapped, and chimeric reads for each sample. To remove potential PCR duplicates, reads that share similar external reference coordinates (+-5 bp) are discarded, to obtain 'considered' reads. Next, *HccDriverFinder* distinguishes junctions by both their coordinates (the human and HBV reference coordinates at the site of the junction) and by their sequences (the 10 nucleotides flanking the junction, 5 on each side). After testing different lengths of junction sequences (with 10, 20, 30, 40, 50 bp surrounding the junction), we found

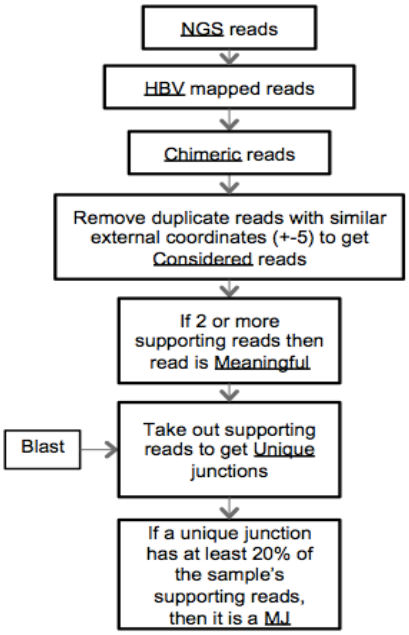

Figure 6: Flowchart for identification of MJs. The underlined terms correspond to the columns in the data summary table.

that a junction sequence of 10 bp yielded the most valid supporting reads. Reads with the same junction coordinates or same junction sequences are considered supporting reads for that particular 'unique' junction. Junctions with at least two supporting reads are categorized as 'meaningful'. As "Bowtie 2 does not guarantee that the alignment reported is the best possible in terms of alignment score (23)", *HccDriverFinder* uses NCBI Blast (27) to perform sequence similarity searches for each unique junction to take all possible alignments into account. Lastly, we currently define MJs as making up at least 20% of the meaningful reads from each sample and having at least two supporting reads. Analysis of the NGS data from our 12 samples resulted in the identification of 12 MJs (Table 3). Surprisingly, S9N had a great number of HBV reads and chimeric reads but no MJs. One explanation is that S9N may support viral replication and the free virus could have attached to all probes.

Table 3: Interactive data summary table listing the statistics for our 12 samples.

Sample	NGS Reads	HBV Reads	% HBV	Chimeric	Considered	Meaningful	Unique	MJs
S17K	982620	386	0.0393	13	13	2	1	1
S17N	1650429	2603	0.1577	89	87	28	5	2
S37K	1363080	608982	44.677	225	219	32	12	1
S37N	1407070	2661	0.1891	18	17	4	2	2
S42K	1149286	507	0.0441	17	17	0	0	0
S42N	1281123	1654	0.1291	42	42	2	1	1
S48K	1571087	526	0.0335	29	29	2	1	1
S48N	1112421	1562	0.1404	91	91	0	0	0
S5K	516861	8393	1.6238	30	30	9	4	4
S5N	636726	4554	0.7152	62	62	20	10	0
S9K	1986277	499425	25.144	73	73	0	0	0
S9N	1227874	321	0.0261	23	23	0	0	0
Total	14884854	1131574	6.0766	712	703	99	36	12

K = tumor, N = non-tumor
The cells with blue, underlined text contain hyperlinks to viewSequence html files (Figure 7).

For each sample, the summary table (Table 3) lists statistics at each step of the MJ identification process and provides links to graphic displays of the reads (Figure 7).

Figure 7: Sample viewSequence File showing reads aligned on an html page. HBV DNA is shown with red lettering and human DNA shown in yellow highlight. Unique junctions and their supporting reads (boxed in black) can be easily distinguished in this visualization format.

2.2) Validation of MJs by PCR-Sanger Sequencing

False positives from NGS reads are often generated experimentally through library preparation. To validate our NGS platform and confirm that the MJs identified were not derived from experimental artifacts, we designed primers based on the NGS sequences to verify the integration sites in the original tissue DNAs. Four primers (two forward for host and two reverse for HBV) were designed by Primer 3 (31) for each junction. We initially selected 12 junctions and tested 48 primer pairs for specificity and sensitivity against the pooled library (positive control), human genomic DNA (negative control), and water (no template control) at 5 different annealing temperatures and with various concentrations of $MgCl_2$ (if needed) (data not shown). 10 junctions had at least 1 primer pair that produced the correct product; more primer selection is in progress for the remaining 2 junctions. Of these 10 junctions, 8 had at least 10 ng of unprocessed tissue DNAs for validation (more tissue DNAs are in preparation). For each, the best primer pair was selected (as listed in Table 4) and four yielded products at expected sizes in both the positive control and tissue DNAs (S17K, S37K(1), S37K(2), and S48K) as shown in Figure 8A. Interestingly, the four junctions that worked were all from tumor tissue. The amplicons of these junctions were purified and subjected to Sanger sequencing. Three junctions have been validated with the correct junction sequence from the tissue DNAs as shown in Figure 8B, the rest of the validations are ongoing.

Table 4: Primer sequences and annealing temperatures for MJ validation PCR.

Sample	Primer	Sequence	Coordinates	Tm (°C)	Annealing (°C)
S5K	F1	AATGACTCAGAACACATGAAAATTATACT	Chr 17 (70825203, 70825232)	61	58
	R1	GATGCTGGGTCTTCCAAATTACTAC	HBV nt. (2115, 2140)	64	
S5K	F2	TCACATTACCTGACTTCAAATTATACCAT	Chr 3 (155476978, 155477007)	62	59
	R2	ATTAACGTTGACATAGCTGACTACTAATT	HBV nt. (2144, 2173)	62	
S17K	FJ2	CAAAGAAATTGCTTATACCATAAGGTGG	Chr 2 (215422687, 215442713)	63	59
	R1	CAGTTAGGATTAAAGACAGGTACAGTAG	HBV nt. (2478, 2505)	62	
S17N	F2	GCAGCTGCCGAGGGAGGGGACCGTC	Chr 21 (8205158, 8205183)	78	74
	R2	CCAGCCAGTGGGGGTTGCGTCAGCAAACT	HBV nt. (1183, 1212)	77	
S37K	F2	GAGCGCACGGCTCGGCAGCGGGGAG	Chr 5 (1295005, 1295030)	81	59
	R2	TTAAAGGTCTTTGTACTAGGAGGCTGTAG	HBV nt. (1776, 1805)	65	
S37K	F2	CACTGGGCAGAAATCACATCGCGTCAACA	Chr Un_GL000220v1 (160973, 161002)	72	63
	R1	CTGACTTCTTTCCTTCTGTTCGAGATCTC	HBV nt. (2115, 2144)	66	
S37N	F1	GAACTCCCTGACCCCTGGCGCTTCCCAAG	Chr 8 (89535388, 89535417)	77	69
	R1	CATGTGACGTGCAGAGGTGAAGCGAAGTG	HBV nt. (1575, 1604)	72	
S48K	F2	ATAGAGCAGGTTTGAAACACTCTTTCTGTAGTATC	Chr 9 (60679309, 60679344)	67	62
	R2	CTTTATAAGGATCAATGTCCATGTCCTAAAGCCAC	HBV nt. (1870, 1905)	68	

F = forward primer, R = reverse primer

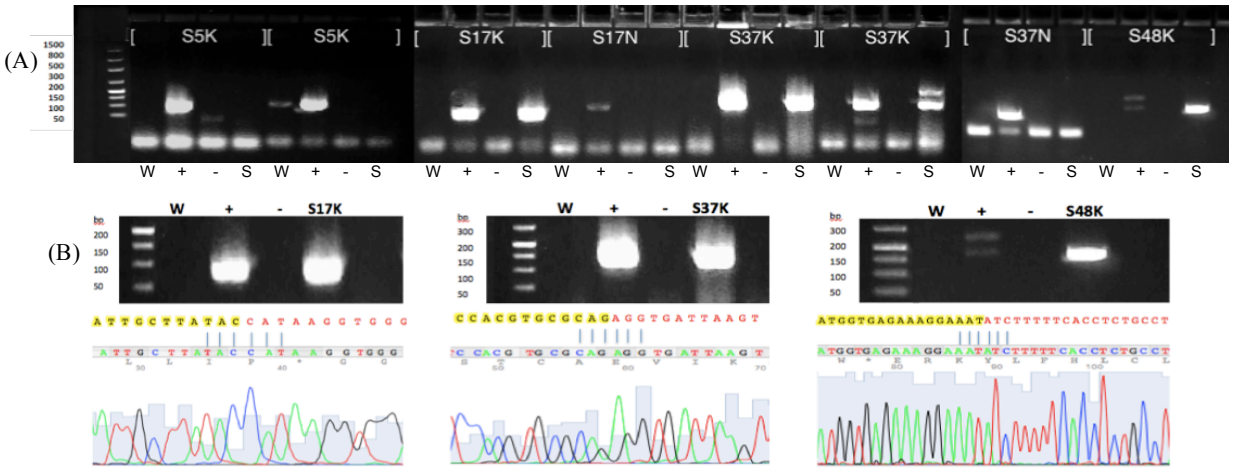

Figure 8: HBV junction validation. (A) PCR products are loaded in the order: no template water control (W), positive control (+), negative control (-), and sample DNA (S). Four of the eight junctions tested were present sample DNA (S17K, S37K(1), S37K(2), and S48K). (B) Sanger Sequencing Chromatogram vs NGS sequence. The sequences acquired from CHOP after Sanger sequencing align with the NGS sequences at the junction sites, successfully validating them as real integration sites.

2.3) Identification of Potential HCC Drivers

We used the NGS data generated from our 12 tissue samples to develop an algorithm for the automated detection of HCC driver genes from MJs. The steps *HccDriverFinder* takes to identify potential driver genes are shown in Figure 2. The HCC driver database is built from a starting list of 4 known HCC driver genes recurrently targeted by HBV integration (TERT, CCNE1, MLL4, FN1) (1, 10, 13, 14, 32). After identification of MJs, *HccDriverFinder* first checks each MJ for recurrency. If the MJ is not from a known driver gene, *HccDriverFinder* verifies if the junction gene is recurrent by searching for it in samples from other patients. Secondly, the program will search each gene in PubMed to check if it is associated with carcinogenesis. Each gene is searched in PubMed with the keywords “cancer” or “tumor” and links to the search results are documented in the output excel sheet. All genes of MJs that meet at least one of the above two criteria are added to the HCC driver gene database.

Through analysis of NGS data from 140 in-house HBV-HCC samples in addition to our 12 samples, we were able to compile a list of 45 potential driver genes, 43 of which are reportedly linked to cancer, and 2 of which were discovered to be recurrent within our own

datasets, as shown in Table 5 (full list in ReadMe in Github). As expected, TERT, was found to be the most recurrent integration site in the 152 HBV-HCC NGS datasets analyzed, followed by other known target genes such as CCNE1 and FN1, as previously reported (1, 10, 13, 14, 32). In addition to confirming known sites, we also discovered new recurrent HBV integrations in genes such as CSAD and ABCC13, both of which are reportedly linked with carcinogenesis (33-35). Furthermore, additional analyses of other published HBV-HCC datasets are ongoing. Thus, our driver database is a continually growing ranked list of potential HCC driver genes.

Table 5: Screenshot of first 18 genes in HCC Driver Gene Database

Gene	Supporting Reads	Recurrent	RecurrencyCount	PubMed Link	Samples
TERT	16	YES	4	https://www.ncbi.nlm.	A23_k, A5_k, S37K, 6
CCNE1	22	YES	2	https://www.ncbi.nlm.	A2_k, 9_1
ABCC13	9	YES	2	https://www.ncbi.nlm.	A9_k, 9_1
RP11-556G	6	YES	2		A9_k, 9_1
ARL3	8		1	https://www.ncbi.nlm.	11b_1
CSAD	6		1	https://www.ncbi.nlm.	A2_k
TRPM2	4		1	https://www.ncbi.nlm.	A9_k
COL14A1	3		1	https://www.ncbi.nlm.	A9_k
KLF3	3		1	https://www.ncbi.nlm.	S37K
VPS35	3		1	https://www.ncbi.nlm.	4_1
FN1	2	YES	1	https://www.ncbi.nlm.	S17K
ABL2	2		1	https://www.ncbi.nlm.	A6_k
AGO3	2		1	https://www.ncbi.nlm.	6
BICC1	2		1	https://www.ncbi.nlm.	A23_k
BLZF1	2		1	https://www.ncbi.nlm.	A2_k
CCSER2	2		1	https://www.ncbi.nlm.	4_1
CDIP1	2		1	https://www.ncbi.nlm.	S37K
CUBN	2		1	https://www.ncbi.nlm.	A9_k

2.4) HBV Mutation Panel:

The flowchart delineating how *HccDriverFinder* detects and processes HBV mutations is shown in Figure 9. We initially compiled a list of 29 HBV mutations known to be associated with liver cancer (36-45) from which our software would reference (full list in ReadMe). *HccDriverFinder* then opens the viral alignment SAM files and sets a new naming convention so that paired reads share the same name. The reads are then sorted by query name and Samtools rmdup (46) is run to remove PCR duplicates. Next, all HBV mapped reads are extracted and parsed one

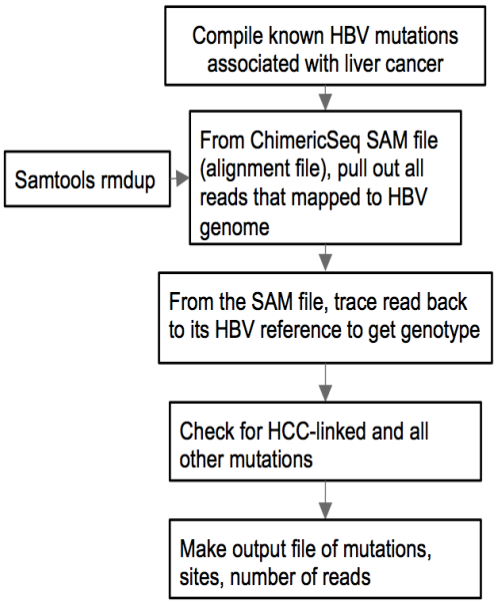

Figure 9: Flowchart for identification of HBV mutations.

read at a time. Each read is checked for known mutations (from the previously compiled list) and for any other mismatches between the query and its reference sequence. The program writes the results to two excel sheets, the first contains the positions of all the known mutations found and the second lists all other mutations sorted by frequency (as example shown in ReadMe).

Of the 1,131,574 HBV mapped reads from our 12 samples, a total of 161,454 mutations known to be associated with HCC were found. These consisted of 23 out of the 29 HCC associated HBV mutations we had compiled. The most frequently seen mutation was G1899A, which was found in 87,597 reads. G1899A is located in the pre-core region of the HBV genome and is correlated with development of cirrhosis and HCC (37-39, 40).

In addition to detecting known mutations, our program also checks for any other mismatches between the read and its reference. HBV replicates by reverse transcriptase, which lacks proofreading, and is thus extremely prone to mutations (37, 47-48). As expected, 5,188,083 mismatches (an average of 4.6 per read) were detected at 3,223 different locations, practically at every nucleotide in the HBV genome.

Table 6: Distribution of HCC-linked and all other HBV mutations across tumor and non-tumor tissue and genotype C.

	Total Mutations	% From Tumor	% From Nontumor	% Genotype C
HCC linked	161,454	98.8	1.2	98.5
Non-HCC linked	5,188,083	47.32	52.68	86.7

Strikingly, 98.8% of the HCC-linked HBV mutations detected were found in tumor tissues, revealing the high enrichment of the HCC-linked mutations in cancer (Table 6). In addition, a large proportion of the mutated reads were of a pure or recombinant form of genotype C, which is the HBV genotype most associated with a higher risk of HCC (49-51).

2.5) HBV Breakpoint Visualization:

In order to further understand integration mechanisms of HBV, we also gave *HccDriverFinder* capabilities for the interactive visualization of HBV read breakpoints. This

function was developed from viewSequence, a Python script that uses ChimericSeq's output csv file to display color-coded reads with a differentiation between host and viral sequences. We first ran our patient files in ChimericSeq with only one HBV genome. A number line was superimposed in a separate window with the length based on the largest stop point. The chimeric reads from the original viewSequence output were aligned to the number line with the smallest HBV start point on the top. In order to display HBV reads, human portions were removed from the display. This visualization has the capability to switch between viewing panels (chimeric reads, HBV reads, and MJs) and "zoom in" to investigate individual sequences as shown in Figure 10. Consistent with results from previous studies, a large proportion of breakpoints were observed between the HBV nt. 1700 and 2200, where the viral enhancer, X gene, and core gene are located (1, 14). Integration of HBV enhancer DNA could drive the overexpression of interrupted genes, and consequently the initiation or progression of HCC (10, 52-53).

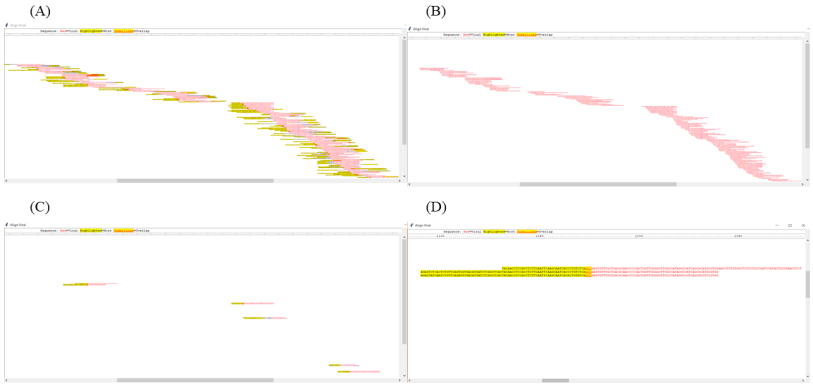

Figure 10: Breakpoint visualization from sample S37K (A) Chimeric reads (B) HBV reads (C) Chimeric reads with MJs (D) Zoomed in view with sequence

Discussion

In this study, we performed NGS DNA library preparation and a primer extension capture to demonstrate the identification of MJs in HCC tissue and generate NGS data for the development of a novel, user-friendly software, *HccDriverFinder*. Our software facilitates the analysis of millions of NGS reads for identifying and displaying integration and mutational data and subsequently assembles an HBV-HCC driver library for precision medicine. By using

HccDriverFinder, we identified HCC-associated MJs from 12 experimental NGS datasets and validated these NGS identified MJs by PCR-Sanger sequencing. Furthermore, additional analyses of 140 in-house HBV-HCC NGS datasets enabled us to assemble an HCC driver database of 45 potential driver genes. A majority of our findings are in line with published literature, thus validating the design of our computational algorithm.

Recent advances in NGS and sequence alignment have enabled cost-effective high throughput analyses of genomes at a single nucleotide resolution, and have become powerful tools to identify genetic modifications (54-56). Application of these technologies to many samples of the same cancer type enables the identification of novel recurrent modifications and presents new targets for cancer diagnostics and treatment (57). Two key challenges in cancer sequencing that we worked to overcome were i) the processing and analyzing of massive amounts of sequencing and alignment data (millions of reads), and ii) the distinguishing of the relatively small number of driver events that are responsible for the development and progression of cancer from the large number of irrelevant passenger events (55, 57-61). As a result, *HccDriverFinder* summarizes NGS and alignment data in a readable and meaningful manner. In addition, *HccDriverFinder's* improved accessibility through a user-friendly GUI interface, as shown in Figure 11, has the potential to expand NGS analytical support to a broader spectrum of users.

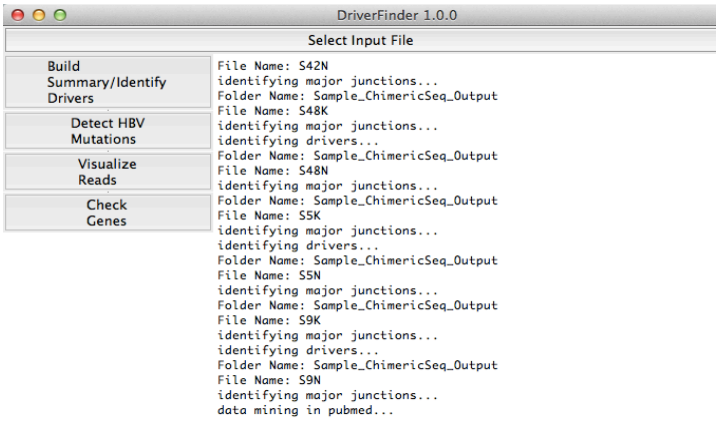

Figure 11: *HccDriverFinder* GUI.

To identify driver genes for HBV-HCC, we present a novel, automated approach to distinguish driver genes from identified MJs. With the application of this approach, we were able

to demonstrate profound analysis of HBV-DNA integration in HCC and identify 45 potential driver genes, 43 of which have been suggested to be associated with cancer by previous literature, and 2 of which were discovered to be recurrent within our datasets. In addition to confirming frequent integrations in known targets such as TERT, CCNE1, and FN1 (1, 10, 13, 14, 32), we discovered new recurrent HBV integrations in genes such as CSAD and ABCC13, both of which are reportedly linked with carcinogenesis. For instance, expression of the CSAD (Cysteine Sulfinic Acid Decarboxylas) was found to be stimulated and maintained in the precancerous fetal liver during hepatocarcinogenesis in rats (33). In addition, studies have demonstrated that ABCC13 (ATP-Binding-CassettC13) is associated with acquired resistance towards Sorafenib treatment in HCC patients (34-35). Further work is warranted to delineate the complex network of these HBV targeted genes and their overall impact on cellular phenotype.

Although MJ validation is still ongoing, we have successfully validated 3 of the MJs identified from the NGS data of our 6 paired samples by PCR-Sanger sequencing. Several junctions may not have been confirmed because they were not as dominant and therefore had insufficient input tissue DNA. Interestingly all 4 confirmed junctions were from tumor tissue, suggesting that they are in fact HCC-associated MJs.

Various mutations in the HBV genome are reportedly linked to the development of HCC (35-44). HBV has a high mutation rate (estimated 4.57×10^{-5} nt. substitutions per site per year) due to its use of reverse transcriptase that lacks proofreading activity, resulting in the rise of HBV quasi-species (37, 47-48). As expected, *HccDriverFinder's* mutation panel detected mutations at every nucleotide of the HBV reference genome. Strikingly, 98.8% of the HCC-linked HBV mutations detected were found in tumor tissues, while only 47.32% of non-HCC-linked mutations were from tumor tissue. A vast majority of the mutated reads detected were of a

pure or recombinant form of genotype C. Studies have shown that genotype C HBV, the most prevalent form, has been found to take a more aggressive disease course and is associated with a higher risk of HCC compared to other genotypes (49-51).

To survey the breakpoints on the HBV genome, we expanded the functionality of the viewSequence program for the interactive visualization and alignment of genomic data. We saw breakpoints occur most frequently between 1700-2200 bp region of the HBV genome, where the viral enhancer, X gene and core gene are located (1, 14). It has been suggested that the X protein encoded by the X gene hastens the development of hepatoma (62-64). Integration of enhancers by HBV could result in substantial activation of oncogenic genes (10, 52-53).

Gene fusions are a common feature found in cancer, such as the TMPRSS2 and ERG gene fusion, the predominant molecular subtype of prostate cancer (65-66). To explore the application of chimeric read identification in other cancers, we simulated a synthetic NGS dataset *in silico* with artificial TMPRSS2-ERG chimeric reads. ChimericSeq was able to identify these gene fusion reads with 78% sensitivity. Work is ongoing to use NGS data from the VCAP (prostate cell-line positive control) to define the parameters for better identification of gene fusion chimeric reads. *HccDriverFinder* could be implemented to stratify gene fusions in a manner similar to its categorization of HBV integration junctions.

Conclusion and Future Direction

From study of NGS DNA generated from 6 pairs of HCC and adjacent non-HCC tissue, we conclude that the primer extension capture enriched NGS approach effectively enabled identification of MJs from 1 million NGS reads. Through analysis of an additional 140 HBV-HCC NGS datasets, we conclude that the majority of MJs identified by our novel software *HccDriverFinder* are associated with HCC. *HccDriverFinder* was also used for the identification

of recurrent junction genes, detection of HCC associated HBV mutations, and analysis of HBV breakpoints to assemble an HCC driver gene database and outline mechanisms of hepatocarcinogenesis. Collectively, our work helps piece together an HBV integration map in HCC, revealing preferential regions of HBV integration within the human genome and providing deeper insight into the elaborate network of driver genes in HCC.

In the future, acquisition of more HBV-HCC NGS datasets will allow us to perform large-scale analysis of HBV-integration through the *HccDriverFinder* software, refine our algorithms for identifying MJs and drivers, and develop new criteria for risk stratification of HBV patients. Our software can be used to determine which driver genes in specific contributed a patient's cancer and thus design a treatment plan tailored to a person's genetic profile. Pharmacologic inhibition of driver gene function has been found to be highly effective (67) and linking our list of newfound HCC driver genes to an emerging toolkit of targeted therapies is likely to improve patient outcomes. Furthermore, our software for identifying driver genes can be used to select patients for targeted therapy trials in which coupling effective drugs to underlying driver events will be critical to improve patient prognosis (68). Further development to apply *HccDriverFinder* to other cancers, such as prostate cancer, is in progress. The current version of *HccDriverFinder* runs on Mac OS only, newer versions compatible with other operating systems, such as Windows and Linux, are underway.

Our laboratory has shown that HBV-host junction sequences can be detected in urine of patients with HBV-HCC (personal communication). In the future, we hope to improve detection sensitivity to identify MJs in the peripheral and employ the *HccDriverFinder* application to analyze urine sample data to stratify patients by risk for HCC in a noninvasive screening method.

References

Access the *HccDriverFinder* software at: https://github.com/Competition-Entrant-2017/DriverFinder
Username: Competition-Entrant-2017
Password: siemens17

1. Zhao LH, Liu X, Wang HY, et al. Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nat. Commun.* 7, 12992 doi: 10.1038/ncomms12992 (2016).

2. "About Hepatocellular Carcinoma." *The Mount Sinai Hospital*, 2017 Available from: http://www.mountsinai.org/patient-care/service-areas/cancer/cancer-services/liver-cancer/hepatocellular-carcinoma

3. Ferlay J, S.I., Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F., GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr

4. Bruix J, Sherman M. Management of hepatocellular carcinoma: an update. *Hepatology* 2011;53(3):1020-1022.

5. Nguyen VT, Law MG, Dore GJ. Hepatitis B-related hepatocellular carcinoma: epidemiological characteristics and disease burden. *J Viral Hepat*. 2009;16(7):453-63.

6. El-Serag, HB, Epidemiology of Viral Hepatitis and Hepatocellular Carcinoma. *Gastroenterology*, 2012;142(6):1264-1273.

7. Song IH, Kim SM, Choo YK. Risk prediction of hepatitis B virus-related hepatocellular carcinoma in the era of antiviral therapy. *World J Gastroenterol*. 2013;19(47):8867–8872.

8. Dhanasekaran R, Bandoh S, Robert LR. Molecular pathogenesis of hepatocellular carcinoma and impact of therapeutic advances. F1000Res. 2016;5

9. Ayub A, Ashfaq UA, Haque A. HBV Induced HCC: Major Risk Factors from Genetic to Molecular Level. *BioMed Research International.* vol. 2013, Article ID 810461, 14 pages, 2013. doi:10.1155/2013/810461

10. Hai H, Tamori A, Kawada N. Role of hepatitis B virus DNA integration in human hepatocarcinogenesis. *World J Gastroenterol.* 2014;20(20);6236-6243.

11. Seeger C, Mason WS. Molecular Biology of Hepatitis B Virus Infection. *Virology*. 2015;0:672-686.

12. Jiang Z, Jhunjhunwala S, Liu J, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* 2012;22(4):593-601.

13. Ding D, Lou X, Hua D, et al. Recurrent Targeted Genes of Hepatitis B Virus in the Liver Cancer Genomes Identified by a Next-Generation Sequencing–Based Approach. *PLOS Genetics*, 2012;8(12):e1003065.

14. Sung WK, Zheng H, Li S, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet.* 2012;44(7):765-9.

15. Cleary SP, Jeck WR, Zhao X, et al. Identification of Driver Genes in Hepatocellular Carcinoma by Exome Sequencing. *Hepatology*. 2013;58(5):1693-1702.

16. Kawi-Kitahata F, Asahina Y, Tanaka S, et al. Comprehensive analyses of mutations and hepatitis B virus integration in hepatocellular carcinoma with clinicopathological features. *J Gastroenterol*. 2016;51(5);473-86.

17. Chiba K, Johnson JZ, Vogan JM, et al. Cancer-associated TERT promoter mutations abrogate telomerase silencing. *eLife*. 2015;4:e07918.

18. Vinagre J, Almeida A, Pópulo H, et al. Frequency of TERT promoter mutations in human cancers. *Nat Commun.* 2013;4:2185.

19. Selena Lin, "Analysis of the complexity of HBV-host junction sequences in patients with HBV-related hepatocellular carcinoma" PhD diss., Drexel University, 2016, ProQuest Dissertations Publishing (10154105).

20. Shieh, F.-S., et al., ChimericSeq: An open-source, user-friendly interface for analyzing NGS data to identify and characterize viral-host chimeric sequences. *PLOS ONE*, 2017;12(8):e0182843.

21. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [2017 Sep 19]. Available from: https://www.ncbi.nlm.nih.gov/

22. Gnirke, A., et al., Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotech*. 2009;27(2):182-189.

23. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

24. Get pandas! September 18, 2017]; Available from: http://pandas.pydata.org/.

25. openpyxl - A Python library to read/write Excel 2010 xlsx/xlsm files. September 18, 2017]; Available from: https://openpyxl.readthedocs.io/en/default/.

26. Cock, P.J.A., et al., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423.

27. Basic Local Alignment Search Tool. September 18, 2017]; Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.

28. pysam - An interface for reading and writing SAM files. September 18, 2017]; Available from: https://github.com/pysam-developers/pysam.

29. Li, H., et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009. 25(16): p. 2078-2079.

30. Tkinter. September 18, 2017]; Available from: https://wiki.python.org/moin/TkInter.

31. Untergasser, A., et al., Primer3—new capabilities and interfaces. *Nucleic Acids Research*. 2012;40(15):e115-e115.

32. Li X, Zhang J, Yang Z, et al. The function of targeted host genes determines the oncogenicity of HBV integration in hepatocellular carcinoma. *Journal of Hepatology*. 2014;60(5):975-984.

33. Kishimoto T, Kokura K, Nakadai T, et al. Overexpression of Cysteine Sulfinic Acid Decarboxylase Stimulated by Hepatocarcinogenesis Results in Autoantibody Production in Rats. *AACR*. 1996;56(22):5230-5237.

34. Annilo T, Dean M. Degeneration of an ATP-binding cassette transporter gene, ABCC13, in different mammalian lineages. *Elsevier Genomics*. 2004;84(1):34-46.

35. Marzac C, Garrido E, Tang R, et al. ATP Binding Cassette transporters associated with chemoresistance: transcriptional profiling in extreme cohorts and their prognostic impact in a cohort of 281 acute myeloid leukemia patients. *Haematologica.* 2011;96(9):1293-1301.

36. Fang ZL, Sabin CA, Dong BQ, et al. HBV A1762T, G1764A mutations are a valuable biomarker for identifying a subset of male HBsAg carriers at extremely high risk of hepatocellular carcinoma: A prospective study. *Am J Gastroenterol*. 2008;103(9):2254-2262.

37. Caligiuri P, Cerruti R, Icardi G, et al. Overview of hepatitis B virus mutations and their implications in the management of infection. *World J Gastroeneterol*. 2016;22(1):145-154.

38. Park YM, Jang JW, Yoo SH, et al. Combinations of eight key mutations in the X/preC region and genomic activity of hepatitis B virus are associated with hepatocellular carcinoma. *J Viral Hepat.* 2014;21(3):171-7.

39. Yin J, Xie J, Liu S, et al. Association between the various mutations in the viral core promoter region to different stages of hepatitis B, ranging of asymptomatic carrier state to hepatocellular carcinoma. *Am J Gastroenterol*. 2011;106(1):81-92.

40. Liao Y, Hu X, Chen J, et al. Precore Mutation of Hepatitis B Virus May Contribute to Hepatocellular Carcinoma Risk: Evidence from an Updated Meta-Analysis. *PLoS One.* 2012;7(6):e38394.

41. Tatsukawa M, Takaki A, Shiraha H, et al. Hepatitis B virus core promoter mutations G1613A and C1653T are significantly associated with hepatocellular carcinoma in genotype C HBV-infected patients. *BMC Cancer*. 2011;11:458.

42. Shen T, Yan XM. Hepatitis B virus genetic mutations and evolution in liver diseases. *World J Gastroenterol.* 2014;20(18):5435-5441.

43. Choi CS, Cho EY, Park R, et al. X gene mutations in hepatitis B patients with cirrhosis, with and without hepatocellular carcinoma. *J Med Virol*. 2009;81(10):1721-5.

44. Khan A, Al Balwi MA, Tanaka Y, et al. Novel point mutations and mutational complexes in the enhancer II, core promoter and precore regions of hepatitis B virus genotype D1 associated with hepatocellular carcinoma in Saudi Arabia. *Int J Cancer.* 2013;133(12):2864-71.

45. Qu LS, Liu TT, Jin F, et al. Combined pre-S deletion and core promoter mutations related to hepatocellular carcinoma: A nested case-control study in China. *Hepatol Res.* 2011;41(1):54-63.

46. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 201;27(21):2987-93.

47. Wen J, Song C, Jiang D, et al. Hepatitis B virus genotype, mutations, human leukocyte antigen polymorphisms and their interactions in hepatocellular carcinoma: a multi-centre case-control study. *Sci Rep.* 2015;5:16489.

48. Wu Y, Gan Y, Gao F, et al. Novel Natural Mutations in the Hepatitis B Virus Reverse Transcriptase Domain Associated with Hepatocellular Carcinoma. *PloS One.* 2014;9:e94864.

49. Chan HL, Hui AY, Wong ML, et al. Genotype C hepatitis B virus infection is associated with an increased risk of hepatocellular carcinoma. *Gut.* 2004;53(10):1494-1498.

50. Chan HL, Wong ML, Hui AY, et al. Hepatitis B virus genotype C takes a more aggressive disease course than hepatitis B virus genotype B in hepatitis B e antigen-positive patients. *J Clin Microbiol.* 2003;41(3):1277-1279.

51. Wong GL, Chan HL, Yiu KK, et al. Meta-analysis: the association of hepatitis B virus genotypes and hepatocellular carcinoma. *Aliment Pharmacol Ther.* 2013;37(5):517-526.

52. Shamay M, Agami R, Shaul Y. HBV integrants of hepatocellular carcinoma cell lines contain an active enhancer. *Oncogene*. 2001;20(47):6811-9.

53. Tu T, Budzinska MA, Shackel NA, et al. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. *Viruses*. 2017;9(4):75.

54. Shendure J, Ji H. Next-generation DNA sequencing. *Nature Biotechnology*. 2008;26:1135-1145.

55. Nakagawa H, Shibata T. Comprehensive genome sequencing of the liver cancer genome. *Cancer Lett.* 2013;340(2):234-40.

56. Rizzo JM, Buck MJ. Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer Prev Res (Phila)*. 2012;5(7):887-900.

57. Raphael BJ, Dobson JR, Oesper L, et al. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* 2014;6(1):5.

58. Porta-Pardo E, Kamburov A, Tamborero D, et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nature Methods.* 2017;14:782-788.

59. Zhang J, Liu J, Sun J, et al. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. *Briefings in Bioinformatics*. 2014;15(2):244-255.

60. Tokheim CJ, Papadopoulos N, Kinzler KW, et al. Evaluating the evaluation of cancer driver genes. *PNAS*. 2016;113(50);14330-14335.

61. Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*. 2011;27(2):175-181.

62. Geng M, Xin X, Bi LQ, et al. Molecular mechanism of hepatitis B virus X protein function in hepatocarcinogenesis. *World J Gastroenterol*. 2015;21(38):10732-10738.

63. Arbuthnot P, Kew M. Hepatitis B virus and hepatocellular carcinoma. *Int J Exp Pathol.* 2001;82(2):77-100.

64. Guerrero RB, Roberts LR. The role of hepatitis B virus integrations in the pathogenesis of human hepatocellular carcinoma. *Journal of Hepatology*. 2005;42:760-777.

65. Mertens F, Johansson B, Fioretos T, et al. The emerging complexity of gene fusions in cancer. *Nature Rev Cancer.* 2015;15(6):371-381.

66. Tomlins SA, Laxman, B, Varambally S, et al. Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer. *Neoplasia*. 2008;10(2):177-188.

67. Mitsudomi T. Driver gene mutation and targeted therapy of lung cancer. *Gan To Kagaku Ryoho.* 2013;40(3):285-290.

68. Mills GB. An emerging toolkit for targeted cancer therapies. *Genome Res.* 2012;22(2):177-182.