COMPUTATIONAL CHARACTERIZATION FOR GENOME INTEGRATION SITES OF HEPATITIS B VIRUS (HBV) AND GENE TARGETS OF HBV X PROTEIN

LIU LIZHEN

NATIONAL UNIVERSITY OF SINGAPORE

2012

COMPUTATIONAL CHARACTERIZATION FOR GENOME INTEGRATION SITES OF HEPATITIS B VIRUS (HBV) AND GENE TARGETS OF HBV X PROTEIN

LIU LIZHEN
(B.Sc. (Hons.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF BIOCHEMISTRY

YONG LOO LIN SCHOOL OF MEDICINE

NATIONAL UNIVERSITY OF SINGAPORE

2012

ACKNOWLEDGEMENT

First, I wish to express my sincere gratitude to my supervisor, Associate Professor Caroline Lee Guat Lay, Principal Investigator in National Cancer Centre of Singapore and Associate Prof. under Department of Biochemistry, Yong Loo Lin School of Medicine in National University of Singapore, for providing me such a good opportunity to do this project and for her valuable guidance and enlightenment.

I would also like to extend my gratitude to my project collaborators Miss Toh Soo Ting, Miss Chan Siew Choo Cheryl, and Dr. Wang Yu for providing me the high-throughput Roche 454 pyrosequencing data, ChIP-Seq Illumina sequencing data, microarray expression profiles, and HCC patient clinical data for analysis. I would like to thank them for their unfailing guidance and help throughout the project. In addition, I would also like to show my appreciation to all the other members of our research laboratory, especially Mr. Mah Way Champ, Dr. Cao Yi, Miss Jin Yu and Mr. Wang Jingbo, for their invaluable help, advice and suggestions during the course of this project. I would also like to convey my heart-felt thanks to Associate Prof. Henry Yang He, Associate Prof. Ken Wing Kin Sung, and Associate Prof. Tan Tin Wee for their support to complete this thesis.

Last but not least, I wish to thank my beloved parents, my husband, and my good friends for their understanding, love, support and encouragement.

TABLE OF CONTENTS

ACKNOWLEDGEMENT i

TABLE OF CONTENTS ii

SUMMARY vi

LIST OF FIGURES ix

LIST OF TABLES xi

ABBREVIATIONS xii

CHAPTER 1: Literature Review and Introduction 1

1.1 HBV-Host Genome Integration 1

1.2 Limitation of PCR-based Methods to identify HBV-Host Genome Integrations 2

1.3 Application of Targeted Deep Sequencing Techniques to Identify Viral-host Integration Boundaries 3

1.4 Analysis of Targeted Deep Sequencing Data to Identify Viral-host Integration Boundaries 7

1.5 HBx-Interacting Transcription Factors 11

1.6 Limitation of ChIP-Chip Methods to Profile Protein-DNA Interactions 12

1.7 Application of ChIP-Seq Methods to Profile Protein-DNA Interactions 16

1.8 Analysis of ChIP-Seq Data to Identify DNA-binding Sites of Proteins 18

1.9 Motif Enrichment Analysis to Identify Co-Factors of Proteins 27

1.10 Project Objectives 30

1.10.1 Computational Analysis for Characterization of HBV-Host Genome Integration Sites 31

1.10.2 Computational Analysis for Identification of Putative Deregulated Direct Gene Targets of HBx 35

1.10.3 Summary of Project Objectives 42

CHAPTER 2: Computational Characterization of HBV-Host Genome Integration Sites 45

2.1 Materials and Methods 45

2.1.1 Data Collection: HBV-containing DNA Fragments Enrichment and FLX Sequencing Library Construction 45

2.1.2 Computational Identification of HBV-Host Junction Sites from FLX Sequencing Data 46

2.2 Results 50

2.2.1 Sequence Identities of the FLX Sequencing Reads 50

2.2.2 Sequence Capture Coverage of HBV Genome from FLX data 53

2.2.3 Identification of Modified HBV and HBV-Human Genome Junctions 54

2.2.4 Analysis of HBV-Host Junctions with Junction Points on HBx gene 59

2.3 Discussion and Future Work 63

CHAPTER 3: Computational Identification of Putative Direct Gene Targets of HBx 67

3.1 Materials and Methods 67

3.1.1 Data Collection: ChIP-Seq Libraries, Expression Profiles & 100 HCC Patients Clinical Data 70

3.1.2 Computational Identification of DNA Binding Sites of HBx 72

3.1.3 Annotation of Genome-wide Potential HBx Binding Sites 74

3.1.4 Motif Enrichment Analysis for Potential HBx-interacting Transcription Factors 74

3.1.5 Analysis of THLE3 Microarray Expression Profiles to Predict Deregulated Direct Gene Targets of HBx 76

3.1.6 Gene Ontology Analysis for Deregulated Gene Targets of HBx 76

3.1.7 Analysis of Microarray Expression Profiles of 100 HCC Patients 77

3.1.8 HCC Patients Clinical Data Analysis to Identify Clinically Associated Deregulated Gene Targets of HBx 77

3.1.9 Correlation of Expressions of HBx and HBx Deregulated Gene Targets in 100 HCC Patients Tumor and Adjacent Non-Tumor Tissues 81

3.2 Results 81

3.2.1 Analysis of ChIP-Seq Data and Identification of Potential DNA Binding Sites of HBx 81

3.2.2 Genome-Wide Distribution of Potential DNA Binding Sites of HBx 84

3.2.3 Potential HBx-Interacting Transcription Factors Predicted from HepG2 ChIP-chip and THLE3 ChIP-Seq Data 86

3.2.4 Potential HBx Deregulated Direct Gene Targets in THLE3 Cells 90

3.2.5 Clinically Associated Potential HBx Deregulated Gene Targets 91

3.2.5.1 Expression of Potential HBx Deregulated Gene Targets in HCC Patients 91

3.2.5.2 Association of Potential HBx Deregulated Gene Targets with HCC Patient Survival Time 92

3.2.5.3 Association of Potential HBx Deregulated Gene Targets with HCC Patients’ Categorical Clinical Features 97

3.2.5.4 Summary of Associations of Potential HBx Deregulated Gene Targets with HCC Patient Clinical Features 101

3.2.6 Correlation of Clinically Associated HBx Deregulated Gene Targets with HBx Protein Expression in the 100 HCC Patients 105

3.3 Discussion and Future Work 106

CHAPTER 4: Conclusion and Future Work 113

4.1 Characterization of HBV-Host Genome Integration Sites in HCC Patients 114

4.2 Future Work on the Computational Analysis Pipeline in Identifying Virus-Host Genome Integration Sites 115

4.3 Identification of HBx Genomic Binding Sites, HBx-interacting Transcription Factors, and Clinically Associated Deregulated Direct Gene Targets of HBx 117

4.4 Future Work on the Clinically Associated Gene Targets of HBx 122

4.5 Conclusion 123

CHAPTER 5: Supplementary Tables 124

References 130

Author’s Publications 144

SUMMARY

Chronic Hepatitis B viral (HBV) infection has been epidemiologically linked to the development of Hepatocellular Carcinoma (HCC) in patients. A significant characterization of chronic HBV infection is the integration of HBV DNA into multiple locations within the host DNA. This integration of viral DNA into host genome has been implicated to contribute to hepatocarcinogenesis through either insertional-mutagenesis or the retention/expression of the original/modified HBV proteins. One viral protein, HBx, has been strongly suggested to play important roles in oncogenicity through the deregulation of host genes. However, the association between chronic HBV infection and HCC remains poorly understood.

Our laboratory had enriched for HBV sequences in 48 HBV-associated HCC patients and employed the FLX Genome Sequencer to characterize variations in the HBV DNA as well as HBV integration events in these patients. In this thesis, I employed a computational workflow to analyze the high-throughput sequencing data, and identified 60 contigs/reads with altered HBV DNA and 63 contigs/reads carrying both HBV and human DNA within the same read from which the HBV-HG junction sites were inferred. Various variations such as insertions, deletions, duplications and inversions were observed from the 60 altered HBV sequences. Interestingly, the HBV-HG integrations were found to preferentially occur at the HBx gene locus (27/63=42.9%) and the 3’ C-terminal of HBx carrying p53 binding domain was often deleted to fuse with the human genome. Deletion of p53 binding domain of HBx may potentially promote carcinogenesis in HCC patients, as p53 is a well-known tumor suppressor. The N-terminal two third of HBx gene carrying transactivation domains were often retained in the integrated form. In

addition, most of the genome integrations were found to occur at the non-coding regions of human genome, such as, gene promoters (4/63), introns (21/63) and intergenic regions (30/63). Nevertheless, computational scanning of the integrated sequences for open reading frames have shown that the genome integration may either lead to early termination of HBV genes or expression of potential chimeric transcripts fusing HBV and human DNA. Significantly, our laboratory has successfully experimentally validated a subset of the integrated sequences and the expression of chimeric transcripts. By characterization of HBV genome integration sites using high throughput targeted genome sequencing, we are now better positioned to gain improved insights on how HBV genome integration may contribute to hepatocarcinogenesis in HCC patients.

To further elucidate the role of the HBx gene in HCC, our laboratory employed chromatin immunoprecipitation and sequencing using the Solexa Genome Sequencer (ChIP-Seq) on immortalized liver cell line, THLE3 using HBx antibodies. I employed a computational workflow to integrate the high throughput ChIP-Seq data, microarray expression profiles for both cell lines (THLE3) and 100 HBV-associated HCC patients, and the clinical data of the 100 HCC patients. A total of 2860 potential HBx binding sites were identified and were found to be significantly enriched in exons and promoter regions of genes ($p<0.00001$). Interestingly, almost half of the predicted binding sites within exons/introns were localized in the first and last exons/introns, indicating the potential regulatory effect of HBx on gene expressions. 195 potential HBx-interacting transcription factors were predicted, of which 129 were commonly predicted from our previous ChIP-chip data on HepG2 cells. 143 potential HBx deregulated direct gene targets were identified in THLE3 cells, indicating the pleiotropic nature of HBx: interact

with a variety of transcription factors and deregulate a large set of genes. 18 of these 143 HBx-associated deregulated genes were also consistently differentially deregulated in the 100 HCC patients. Seven of these 18 genes were found significantly associated with various patients' clinical features including survival, tumor grade, tumor invasion, liver cirrhosis, tumor capsulation and multifocality. By identification of clinically associated potential HBx deregulated direct gene targets, we are now in a better position to explore the role of HBx in hepatocarcinogenesis in HCC patients.

LIST OF FIGURES

Figure 1.1: Hybrid capture to capture viral-host integration sites 7

Figure 1.2: ChIP-chip and ChIP-Seq workflow. 14

Figure 1.3: Analysis of ChIP-Seq sequencing data 19

Figure 1.4: Bimodal enrichment pattern of ChIP-Seq sequencing data 25

Figure 1.5: Aims of the project. 44

Figure 2.1: Flowchart of the HBV enrichment strategy applied in our laboratory. 46

Figure 2.2: Analysis pipeline for identification of HBV-human junctions from 454 FLX sequencing data 47

Figure 2.3: Typical patterns of HBV-containing sequence identities 50

Figure 2.4: Summary of sequence identities for all 1,902,755 raw FLX sequencing reads 51

Figure 2.5: Coverage of the HBV genome (3215bp) by the 378 HBV-containing sequences including 220 assembled contigs and 158 unassembled reads in patients 54

Figure 2.6: Enrichment of HBV-HG junctions with integration sites on HBx gene 57

Figure 2.7: Location plot of the 27 predicted HBV-HG junctions where the junction points fall on HBx gene (Supplementary Table S2) 62

Figure 3.1: Workflow of computational analysis to identify genomic binding sites of HBx and putative clinically associated direct target genes of HBx 68

Figure 3.2: Flowchart of experimental design for generation of ChIP-Seq data and gene expression profiles performed in the laboratory 71

Figure 3.3: Flowchart of the statistical hypothesis testing on the association of HBx deregulated gene targets with patient clinical data 79

Figure 3.4: Processing of ChIP-Seq raw reads to identify potential HBx binding sites... 82

Figure 3.5: Genome-wide distribution of the 2860 potential DNA binding sites of HBx predicted from ChIP-Seq data in THLE3 cells 85

Figure 3.6: Summary of the computational analysis results for identification of HBx binding sites, potential HBx-interacting transcription factors and HBx deregulated direct gene targets 89

Figure 3.7: Hierarchical clustering of the 23 potential HBx deregulated gene targets that are significantly differentially expressed between tumor and adjacent non-tumor tissues of the 100 HCC patients with average fold change above 2.. 92

Figure 3.8: Survival plots for the four survival-associated potential HBx deregulated gene targets... 96

Figure 3.9: Plots for the six potential HBx deregulated gene targets that showed significant associations with the 100 HCC patients' categorical clinical features 99

Figure 3.10: Scatter plots for the expressions of the seven potential HBx deregulated gene targets and expressions of HBx protein in 100 HCC patients 106

LIST OF TABLES

Table 1.1: Comparison of metrics and performances of three next-generation DNA sequencing platforms and two third generation sequencing technologies.......................... 5

Table 1.2: Comparison of metrics and performances of ChIP-chip and ChIP-Seq technologies 17

Table 1.3: Comparisons of various peak-calling algorithms for ChIP-Seq data 23

Table 2.1: Summary of the identities of the 378 HBV-containing sequences. 56

Table 2.2: Summary of the locations of the 63 junction points on human genome. 59

Table 3.1: Comparisons of ChIP-chip data on HepG2 cells (Sung et al., 2009) and ChIP-Seq data on THLE3 cells. 87

Table 3.2: Summary of corrected two-sided significance values from the clinical statistical tests on the 18 potential HBx deregulated gene targets. 94

Table 3.3: Summary of the clinical associations for the seven potential HBx deregulated gene targets 103

Table 3.4: Functional annotations of the seven clinically associated potential HBx deregulated gene targets.. 104

Supplementary Table S1: Number of sequences (assembled contigs and unassembled reads) that were classified into the five major groups in each patient sample............... 124

Supplementary Table S2: Information for 56 HBV-HG junctions and seven modified HBV-HG junctions predicted in different patient samples. 125

Supplementary Table S3: List of the 195 enriched transcription factors from ChIP-Seq data in THLE3 cells, among which 129 were common with the transcription factors predicted from ChIP-chip data in HepG2 cells.. 127

ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
CCAT	Control-based ChIP-Seq Analysis Tool
ChIP-	Chromatin Immunoprecipitation
DAVID	Database for Annotation, Visualization and Integrated Discovery
FDR	False Discovery Rate
HBV	Hepatitis B Virus
HBx	Hepatitis B virus X gene
HCC	Hepatocellular Carcinoma
HOMER	Hypergeometric Optimization of Motif Enrichment
MACS	Model-based Analysis for ChIP-Seq
NGS	Next Generation Sequencing
IP	Immunoprecipitation
UTR	Un-Translated Region
SPSS	Statistical Package for the Social Sciences
TSS	Transcription Start Site

CHAPTER 1: Literature Review and Introduction

1.1 HBV-Host Genome Integration

Hepatocellular carcinoma (HCC) is the fifth most common subtype of liver cancer and is found to be the third leading cause of cancer death in the world due to late diagnosis and limited treatment options (Blum, 2005; Lupberger and Hildt, 2007). There are many risk factors that may cause the development of HCC, including chronic infections of hepatitis B or C virus (HBV/HCV), aflatoxin exposure and excessive alcohol consumption. However, the most epidemiologically associated risk factor is HBV infection, as it has been estimated that chronic HBV infection accounts for 50-55% of all HCC cases in the world (Arbuthnot and Kew, 2001; Chang, 2003; Lupberger and Hildt, 2007; Parkin et al., 2001). As HBV infection precedes the development of HCC by several years, the time gap could allow multiple cellular events such as genetic or chromosomal changes to occur which eventually lead to HCC. One of the key mechanisms in hepatocarcinogenesis involves the integration of HBV genome into the host genome, which is observed in 85-90% of HCC cases and has been reported by many isolated studies to play important roles in HCC development (Bonilla Guerrero and Roberts, 2005; Buendia, 1992; Robinson, 1994). HBV genome integration occurs at early stage after HBV infection and is reported to contribute to host chromosomal instability by various complex genome alteration events which may result in large inverted duplications, deletions, and chromosomal translocations (Tan, 2011). Studies also have shown that frequent HBV genome integrations and variations may disrupt host genes that are essential for cell signalling, proliferation and apoptosis (Boyault et al., 2007; Kuang et al., 2004; Murakami et al., 2005; Paterlini-Brechot et al., 2003; Saigo et al., 2008; Tan, 2011). Therefore, HBV-host genome integrations and alterations may play a crucial role in HBV-induced development of HCC. However, the detailed mechanism of how HBV genome integration may gradually lead to hepatocarcinogenesis in HCC patients remains unclear (Ng and Lee, 2011; Tan, 2011).

1.2 Limitation of PCR-based Methods to identify HBV-Host Genome Integrations

Previously, many research groups have characterized HBV integrations using PCR-based (Polymerase Chain Reaction) methods such as HBV-Alu-PCR which designed one primer specific to HBV sequence and another primer directed to the most abundant mobile Alu elements/repeats of human genome to amplify the virus/cellular DNA junctions (Tu et al., 2006), cassette-ligation mediated PCR which used cassette-ligated human genome DNA fragments adjacent to the integrated HBV DNA as a template for nested PCR with the cassette- and HBV-specific primers to identify HBV integration sites from the HBV DNA amplified from HCC patient liver tissues (Saigo et al., 2008; Tamori et al., 2003; Tamori et al., 2005), and low resolution southern blot which hybridized the HBV DNA extracted from HCC patient tumor tissues with the HBV DNA regions as probes to identify integrated HBV DNA sequences (Tamori et al., 2003; Urashima et al., 1997) etc. These methods combining PCR and capillary sequencing have shown that HBV integration sites might not be entirely random as generally believed, and more importantly, HBV was observed to be mutated or truncated in the integrated form.

As a result, due to the lack of knowledge on the virus sequences retained in the integrated form and the extremely high sensitivity of PCR, a potential problem associated with PCR methods is that primers designed may reside at truncated or mutated or polymorphic regions of the virus genome, resulting in failure in amplification and thus leading to potential increased false negative rates in discovering virus-host integration sites. In addition, without prior knowledge of the virus integrated sequences, PCR primers and reactions covering the whole virus genome may be required in order to fully characterize the virus integration sites, and this would be extremely labour intensive to carry out. More efficient and higher resolution techniques are needed for detection of virus integration sites in a genome-wide scale in order to overcome the limited prior knowledge of integration sites. In recent years, targeted genome sequencing-based approaches have rapidly replaced PCR-based methods (combination of PCR and capillary sequencing) to discover genome structural variants including virus-host integrations (Ansorge, 2009; Mardis, 2009).

1.3 Application of Targeted Deep Sequencing Techniques to Identify Viral-host Integration Boundaries

The low throughput and high cost of the traditional Sanger capillary-based sequencing has been a key limiting factor for full-sequencing-based approaches.

There has been increased demand for the development of low-cost and high-throughput sequencing technologies. In recent years with the emergence of "Next Generation Sequencing" (NGS) technologies such as Roche/454 Life Sciences™**,** Illumina/Solexa™ Sequencing and Applied Biosystems SOLiD™, sequencing costs have been brought down by several orders of magnitude and throughput has been raised by hundreds of folds (Shendure and Ji, 2008). In addition, third generation sequencing techniques, such as Ion Torrent™ semiconductor sequencing, Complete Genomics™ DNA Nanoball (DNB) sequencing (Drmanac et al., 2010; Porreca, 2010), etc. are providing another big boost to this approach with ever higher throughput and lower cost. These deep sequencing techniques enable parallelization of sequencing processes, producing millions of sequence reads at once. Table 1.1 compares the performances of three next-generation sequencing platforms and two third generation sequencing technologies.

In particular, Roche 454 Life Sciences has the ability to sequence whole genomes in days, with 99% accuracy and at a cost of 100x less than using the capillary-based sequencing methods. Besides, the Roche FLX 454 pyrosequencing technology can even achieve average read length of 400bp which has drastically increased the sequencing depth and capacity. Development of these high-accuracy, high-throughput and low cost sequencing techniques has improved the applications of sequence-based methods to a whole genome scale with fine-tuned resolution to single base precision (Mardis, 2008a, b; Schuster, 2008; Stephens et al., 2009).

Table 1.1: Comparison of metrics and performances of three next-generation DNA sequencing platforms and two third generation sequencing technologies: 454 pyrosequencing, Illumina Solexa sequencing, Applied Biosystems SOLiD sequencing, Ion Torrent semiconductor sequencing and Complete Genomics DNA Nanoball sequencing (DNB). (Table updated from Shendure and Ji 2008)

Deep-sequencing platforms	Next Generation Sequencing			Third Generation Sequencing	
	454 pyrosequencing	**Illumina (Solexa) sequencing**	**Applied Biosystems SOLiD sequencing**	**Ion Torrent Semiconductor sequencing**	**Complete Genomics DNA Nanoball sequencing**
URL	http://www.454.com	http://www.illumina.com/pages.ilmn?ID=204	http://www.appliedbiosystems.com.sg/	http://www.iontorrent.com/	http://www.completegenomics.com/services/technology/details/
Sequencing Chemistry	Pyrosequencing	Polymerase-based sequence-by-synthesis	Ligation-based sequencing	Semiconductor sequencing	Unchained ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR	Emulsion PCR	rolling circle replication
Mb per run	100 Mb	600,000 Mb	170,000 Mb	100 Mb	180,000 Mb
Time per run	7 hours	9 days	9 days	1.5 hours	12 days
Read length	400 bp	2x100 bp	35x75 bp	200 bp	35 bp (mate-pair)
Cost per run	$8,438 USD	$20,000 USD	$4,000 USD	$350 USD	$20,000 USD per genome
Cost per Mb	$84.39 USD	$0.03 USD	$0.04 USD	$5.00 USD	
Cost per instrument	$500,000 USD	$600,000 USD	$595,000 USD	$50,000 USD	N.A

Nowadays, a variety of techniques that specifically capture genomic genes or regions of interest from genomic samples coupled with ultra-high throughput NGS sequencers, has been increasingly adapted for and applied in cancer research for the detection of larger genome structural variants, including insertions/deletions, translocations and viral insertions (Abel et al., 2010; Duncavage et al., 2011; Hernandez et al., 2011; Mardis, 2009; Stephens et al., 2009). Such targeted deep sequencing narrows down the sequencing to important genes or regions of interest instead of the entire genome. It allows analysis of interesting genomic sequence variants more efficiently and at even lower cost, especially in the context that NGS has the capacity to sequence multiple experimental samples in a single run by using “barcodes” or indexed labels for individual samples (Mardis, 2008; Abel et al., 2010). To reduce costs, it is often necessary to select regions of interest before sequencing. There are several target enrichment methods, including standard PCR, ligation-based PCR or hybrid capture (Mamanova et al., 2010; Summerer, 2009). In the context of viral-human genome integration, hybrid capture enrichment adopts a basic principle that uses viral-specific probes to hybridize with DNA fragments containing viral sequences or viral-human integration boundaries. The un-hybridized DNA fragments containing only human sequences are washed away, and the captured DNA sequences of interest are then eluted for deep sequencing (See Fig 1.1). Analysis of the deep sequencing data can identify chimeric sequences which contains the viral-host integration boundaries. Hybrid capture is advantageous over PCR-based enrichment approaches by allowing identification of novel viral integration sites or translocation breakpoints (Abel et al., 2010; Mamanova et al., 2010).

With limited knowledge of HBV-human integration sites and to fully characterise HBV-human integrations over whole genome, our laboratory has proposed to apply the hybrid capture enrichment strategy to capture DNA fragments containing HBV sequences or HBV-host integration sequences from the complex HCC patient genomic DNA samples, coupled with ultra-high throughput FLX 454-pyrosequencer, to identify the chimeric sequences representing HBV-human integration sites. As part of a larger research project in our HBV research laboratory, I have proposed an analysis pipeline for the ultra-high throughput FLX sequencing data to characterize HBV-human genome integration sites.

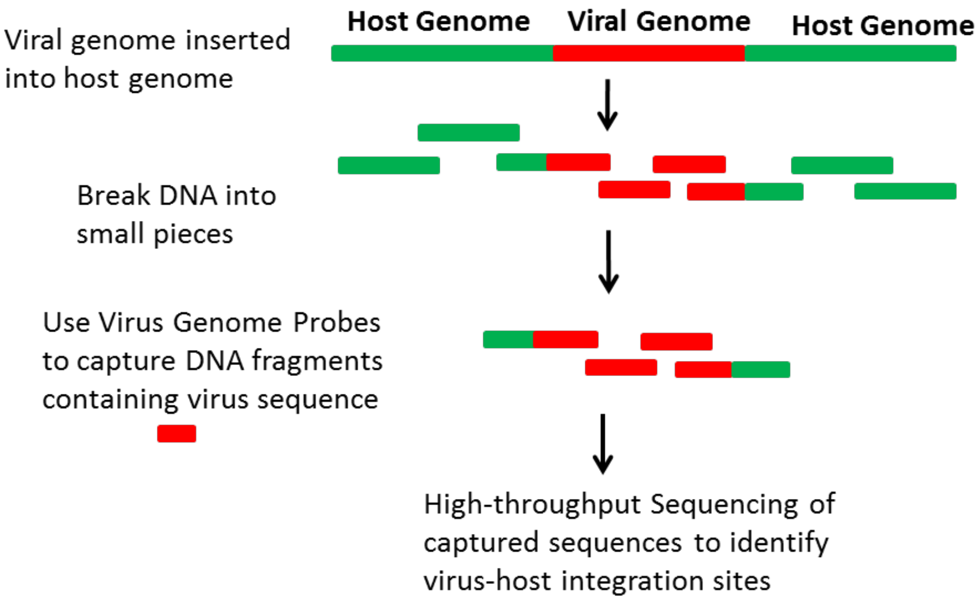

Host Genome
Viral Genome
Host Genome
Viral genome inserted into host genome
Break DNA into small pieces
Use Virus Genome Probes to capture DNA fragments containing virus sequence
High-throughput Sequencing of captured sequences to identify virus-host integration sites

Figure 1.1: Hybrid capture to capture viral-host integration sites. Human genomic DNA (green) containing inserted viral genome (red) is first fragmented to certain size. The fragmented DNAs are then hybridized with capture probes that are specific and complementary to viral DNA sequences, and subsequently fragments not containing virus DNA are washed away. The captured DNA containing viral sequences are then eluted for deep sequencing. By analyzing the sequencing data, chimeric reads consisting of the human/virus integration boundaries can be identified.

1.4 Analysis of Targeted Deep Sequencing Data to Identify Viral-host Integration Boundaries

Targeted genomic regions of interest can be sequenced at great depth using next generation sequencing technologies. There have been several programs developed to date that analyse deep sequencing data for locating sequence variants, such as, Pindel (Ye et al., 2009), BreakDancer (Chen et al., 2009), MoDIL (Lee et al., 2009), PEMer (Korbel et al., 2009), VariationHunter (Hormozdiari et al., 2010), and SLOPE (Abel et al., 2010) etc. Pindel, BreakDancer, MoDIL, PEMer and VariationHunter are specifically designed to analyse sequence data generated from whole genomes while SLOPE is developed to analyse targeted sequence data. Pindel identifies insertions/deletions with single-base resolution but is not designed to detect virus insertion boundaries or sequence breakpoints. BreakDancer, MoDIL, PEMer, VariantionHunter and SLOPE rely on discordant mapping of paired-end diTag sequencing data to detect genome structural variants. Paired-end diTag sequencing of targeted DNA fragments is one of the popular strategies used to discover genome-wide sequence structural variations, based on the principle that the paired-end tags generated from high-throughput sequencer can be aligned back to the host reference genome sequences and abnormal separations or locations between the two reads of a pair suggest a potential genome structural variation, like insertion, deletion, rearrangements and translocation (Bashir et al., 2008; Korbel et al., 2007; Ng et al., 2006; Ruan et al., 2007; Tuzun et al., 2005; Volik et al., 2003). However, the problem associated with these discordant paired-end strategies in characterizing virus-host integration boundaries is that they generally cannot achieve single-base resolution and might have relatively high false positive rates because of limited prior knowledge of the virus insertion size (Mardis et al., 2009). Hence, due to limited prior knowledge of the HBV virus insertion size in human genome and in order to comprehensively characterize the integration sites precisely in single-base resolution, we embarked single-end sequencing with FLX 454 pyrosequencer which is capable of generating significantly longer reads.

Analysis of the single-end high-throughput sequencing data usually begins with the alignment of the sequencing reads back to the entire host reference genome, and this mostly is the key limiting and time-consuming step in the analysis process. Currently, there are many available sequence assembly software designed for aligning deep sequencing data, including:

a) *de novo* assemblers that merge sequences based on overlaps between sequence reads, such as, ABySS (Simpson et al., 2009), SSAKE (Warren et al., 2007), VCAKE (http://vcake.sourceforge.net/), EULER-SR (Chaisson and Pevzner, 2008), Velvet (Zerbino and Birney, 2008), MIRA (http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html), and NextGENe (http://www.softgenetics.com/NextGENe_9.html);

b) reference-guided assemblers that map sequence reads to a known reference genome, such as, RMAP (Smith et al., 2009; Smith et al., 2008), SeqMap (Jiang and Wong, 2008), SHRiMP (http://compbio.cs.toronto.edu/shrimp/), ZOOM (Lin et al., 2008), MAQ (http://maq.sourceforge.net/), NovoAlign (http://biowulf.nih.gov/apps/novocraft.html), GenomeMapper (http://1001genomes.org/downloads/genomemapper.html), MOSAIK (http://bioinformatics.bc.edu/marthlab/Mosaik#News), BWA (Li and Durbin, 2009, 2010) and Bowtie (Langmead et al., 2009);

c) assemblers that can do both *de novo* and reference-guided assembly, including SOAP (Li et al., 2008), CLC Genomics Workbench (www.clcbio.com/genomics/), and DNASTAR SeqMan NGen (http://www.dnastar.com/t-nextgen-seqman-ngen.aspx).

A common feature of these sequence assemblers is that they are computationally intensive requiring large computational power and processing memory. Besides being time-consuming, most of these assemblers are either only suitable for small genomes, or are restrained to a limited number of input sequences in each assembly run, or restricted to certain sequence read length. FLX 454 pyrosequencer generates sequence reads of variable lengths ranging from thirty to thousands base pairs. A commercial assembler, the SeqMan NGen which is developed by the company DNASTAR, is fast, accurate and specifically designed for 454 pyrosequencing reads with no restrictions on the number of input sequences and sequence lengths. SeqMan NGen was found to be ideal for *de novo* assembly of 454 pyrosequencing data, permitting closely examination of the quality and reliability of the assembled sequences for post-assembly analysis.

Although SeqMan NGen was designed for 454 sequencing data, it is less suitable for identifying virus-host integration boundaries compared to the standalone BLAST (Basic Local Alignment Search Tool) program (Altschul et al., 1997). This is because SeqMan NGen can only detect the alignments where the full length of the sequencing reads match to the reference genome when doing reference-guided assembly, while BLAST searches for local alignments between the reads and reference genomes allowing identification of reads with one part mapped to virus genome and the other part aligned to host genome, thereby leading to the identification of virus-host integration sites. BLAST is the most commonly used tool to search against large genome sequence databases, and is perfectly suitable for sequencing reads of variable lengths. Also BLAST provides additional options for users to set the mapping thresholds to adjust the stringency of alignments, such as matching identities, E-values and low complexity filter etc. In this study, I implemented an analysis workflow utilizing BLAST to map the targeted high-throughput single-end deep sequencing reads of variable lengths to both human and virus reference genome sequences, in order to identify the virus-human integration boundaries.

1.5 HBx-Interacting Transcription Factors

Due to unresponsiveness to treatment and late symptom recognition, HCC is one of the most common and lethal cancer in the world (Blum, 2005; Lupberger and Hildt, 2007). It is estimated that 50-55% of HCC cases in the world are associated with chronic infection of HBV (Parkin et al., 2001). The viral X-gene (HBx) of HBV is conserved among all mammalian hepadnaviruses and the HBx protein has been implicated to play a major role in the development of HCC in chronic HBV-infected patients.

HBx is a multifunctional protein of length 154-amino acids. It acts as a promiscuous transactivator that disrupts host cellular gene expressions and subsequent cellular pathways, such as, signalling pathways, DNA repair mechanisms, proliferation, and apoptotic cell death (Becker et al., 1998; Groisman et al., 1999; Lee and Lee, 2007; Matsuda and Ichida, 2009), which ultimately may lead to tumorigenesis. HBx is implicated to modulate aberrant host gene expressions not by binding to DNA directly but through its interactions with transcription factors (Andrisani and Barnabas, 1999; Ganem, 2001; Wu et al., 2001). Currently, various transcription factors (e.g. NF-kappa B, NF-AT, AP1, P53, E2F1, CREB, STAT3), as well as several general transcription machinery complexes in the cell (e.g. TATA-binding protein, TFIIB, TFIIH, RPB5), have been reported to interact with HBx (Benn et al., 1996; Cheong et al., 1995; Maguire et al., 1991; Qadri et al., 1995; Waris et al., 2001; Williams and Andrisani, 1995). Deregulating host gene expression through interaction with transcription factors has been known to be one of the major underlying mechanisms that HBx plays in carcinogenesis. Systematically identifying the list of transcription factors that interacts with HBx and the direct gene targets of HBx-transcription factor complex could provide further insights into HBx functions in HCC. To address this, our laboratory had been the very first to generate antibodies against HBx protein and utilize chip-based chromatin immunoprecipitation technology (ChIP-chip) to identify genomic binding sites and candidate gene targets of HBx (Sung et al., 2009).

1.6 Limitation of ChIP-Chip Methods to Profile Protein-DNA Interactions

ChIP-chip, which is the coupling of chromatin immunoprecipitation with microarray chip technology, was initially described in 1999 and has been widely used in past few years to investigate protein-DNA interactions and determine the binding sites of proteins in genome (Aparicio et al., 2004; Blat and Kleckner, 1999; Buck and Lieb, 2004). Most ChIP-chip protocols first fragment the genomic DNA into small pieces, and then employ specific antibodies against DNA-binding proteins of interest to immunoprecipitate chromatin cross-linked with proteins of interest, before hybridizing the immunoprecipitated DNA fragments onto primarily promoter-sequence microarray chip (see Fig 1.2). ChIP-chip is powerful enough to determine the binding sites of DNA-binding proteins at high resolution and on a genome-wide basis. Several studies have applied ChIP-chip using antibodies against specific transcription factors to identify binding sites and candidate gene targets of those transcription factors (e.g. E2F, c-Myc, P53, and P65 etc) (Li et al., 2003; Lim et al., 2007; Wei et al., 2006; Weinmann et al., 2001; Zeller et al., 2006). Similarly, to characterize DNA binding sites of HBx directly on a genome-wide basis, our laboratory has generated antibodies specifically against HBx protein which is useful for chromatin immunoprecipitation, and successfully predicted a list of DNA binding sites from ChIP-chip technique, direct gene targets of HBx and a list of potential HBx-interacting transcription factors obtained from motif enrichment analysis (Sung et al., 2009).

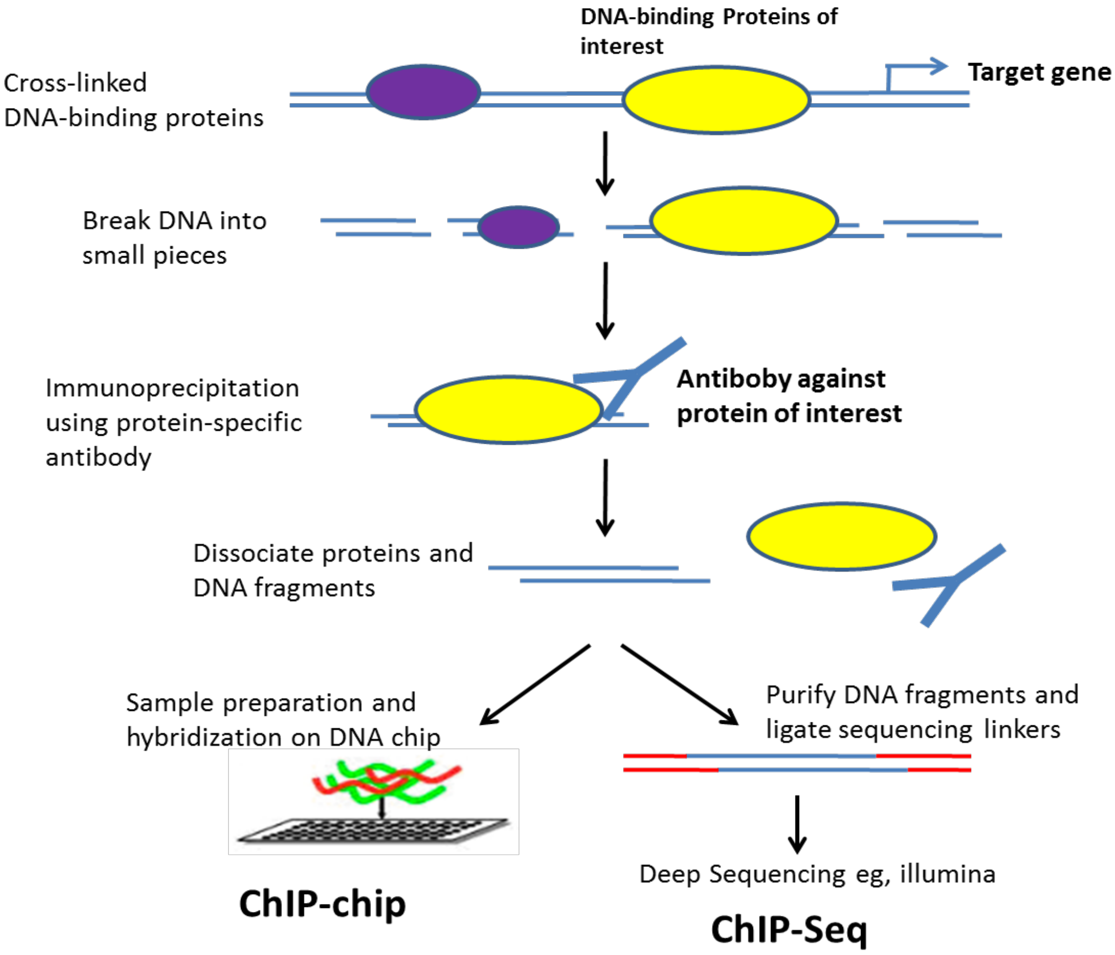

DNA-binding Proteins of interest
Target gene
Cross-linked DNA-binding proteins
Break DNA into small pieces
Immunoprecipitation using protein-specific antibody
Antiboby against protein of interest
Dissociate proteins and DNA fragments
Sample preparation and hybridization on DNA chip
Purify DNA fragments and ligate sequencing linkers
Deep Sequencing eg, illumina
ChIP-chip
ChIP-Seq

Figure 1.2: ChIP-chip and ChIP-Seq workflow. DNA-binding proteins are first cross-linked to double-stranded genomic DNA, including protein of interest (yellow) and other uninteresting proteins (purple). The protein-bounded DNA strands are then broken up into small pieces, using methods like sonication. Antibodies specifically against the protein of interest are added in to immunoprecipitate chromatin bound with the proteins of interest. After dissociation with the bound proteins, the immunoprecipitated DNA fragments are prepared either for hybridization on microarray DNA chip (ChIP-chip) or high-throughput deep sequencing (ChIP-Seq). Both ChIP-chip and ChIP-Seq are designed to detect binding sites of DNA-binding proteins in high resolution on a genome-wide basis. However, ChIP-Seq is advantageous over ChIP-chip since ChIP-Seq can determine binding sites over the whole genome while ChIP-chip is limited to the genome regions tiled on microarray chip.

However, one problem associated with ChIP-chip-based methods is that, these array-based methods are restricted to the genome regions tiled on the microarray chip, for example, tiled array of 1.5kb promoter regions of human genes (Sung et al., 2009), and this would probably lead to increased false negative rates as true binding sites of HBx on the un-tiled regions of the genome will not be interrogated. As a consequence, the high false negative rates of ChIP-chip may cause bias in downstream analysis when predicting potential HBx-interacting transcription factor motifs based on the identified list of binding sites. Additionally, due to the existence of hybridization noise, spatial variation, dye bias, technical bias, dynamic intensity signal measurements, and lack of reproducibility associated with DNA microarray chip experiments, most published studies using ChIP-chip methods repeated their experiments at least three times (technical replicates) to maintain experimental accuracy, technical precision and biological significance (Dombkowski et al., 2004; Eklund and Szallasi, 2008; Febbo and Kantoff, 2006; Rosenzweig et al., 2004; Steger et al., 2011). Though there are currently many software packages available aiming to minimize array background noises and artefacts, statistical analysis of the large amount of raw data with multiple technical replicates generated from arrays is always facing a challenge to extract biologically meaningful information. Therefore, with the need to reduce false negative rates and improve analysing accuracy for ChIP-chip method, a recent advance that couples chromatin immunoprecipitation with ultra high-throughput deep DNA sequencing technology (ChIP-Seq) was employed to investigate protein-DNA interactions on a genome-wide basis (Barski et al., 2007; Johnson et al., 2007; Robertson et al., 2007).

1.7 Application of ChIP-Seq Methods to Profile Protein-DNA Interactions

As shown previously in Fig 1.2, ChIP-Seq is a technique that consists of ChIP method that uses antibody specific to the protein of interest to immunoprecipitate and enrich for the DNA fragments bound by protein of interest, followed by size selection and ultra high-throughput deep sequencing of the enriched DNA fragments associated with the protein of interest (Johnson et al., 2007). Both ChIP-chip and ChIP-Seq require highly specific antibodies that could specifically recognize and immunoprecipitate chromatin crossed-linked with protein of interest. Nevertheless, with the advent of ultra high-throughput deep sequencing technique, ChIP-Seq offers many advantages over ChIP-chip with higher base-pair resolution, greater genome coverage, increased sensitivity and specificity, no hybridization noise and dye bias generated from the cross-hybridization step in ChIP-chip (Park, 2009; Robertson et al., 2007). A review paper published by Park (2009) provides a detailed comparison of ChIP-chip and ChIP-Seq methods including experimental protocols and computational data analysis. Table 1.2 briefly summarizes a comparison for ChIP-chip and ChIP-Seq technologies.

Table 1.2: Comparison of metrics and performances of ChIP-chip and ChIP-Seq technologies. (Table updated from Park, 2009)

Properties	**ChIP-chip**	**ChIP-Seq**
Cost	$400-800 USD per array; multiple array needed for large genome	$1,000-2,000 USD per sample lane
Genome coverage	only on promoters, specific genes or certain chromosomal regions	entire genome
Genomic repeats	can avoid repeats from array	repeats are sequenced
Platform noises	die bias & hybridization noise	possible GC bias
Multiplexing	no	yes by using library index or barcode
Amount of input IP DNA	more	less
Peak detection	fewer peaks with broader width	larger number of more localized peaks
Resolution	array-specific (30-100 bp)	single nucleotide
Reproducibility	microarray lower reproducibility (at least three technical replicates)	high
Signal-to-noise ratio	lower	better
Bioinformatics analysis	harder (multiple replicates)	easier

ChIP-Seq, first described in 2007, was one of the very early applications of next generation sequencing technologies (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007). With the decreasing cost of ultra high-throughput sequencing, there has been an increasing trend nowadays to apply ChIP-Seq methods to systematically profile protein-DNA interactions and assess putative genome-wide binding sites of important proteins, including polymerases, transcription factors and tumor suppressor proteins, in the areas of cancer research, transcriptional regulatory networks studies and immune function studies (Botcheva et al., 2011; Hawkins et al., 2010; Northrup and Zhao, 2011; Park, 2009; Scisciani et al., 2011; White, 2011; Xie et al., 2011). For example, Botcheva et al., (2011) was the first that successfully profiled genome-wide *de novo* mapping of the putative genomic binding sites of the tumor suppressor p53 in normal and cancer-derived human cells, by applying ChIP-Seq experiments and computationally analysing ChIP-Seq data for high-confidence ChIP-Seq peaks. It has been shown that ChIP-Seq is sufficiently powerful enough to identify genomic binding sites of DNA-binding proteins with large genome coverage. This gives our laboratory the incentive to apply ChIP-Seq methods instead of ChIP-chip using antibodies against HBx to profile genomic binding sites of HBx over the whole human genome.

1.8 Analysis of ChIP-Seq Data to Identify DNA-binding Sites of Proteins

ChIP-Seq experiments generate large quantities of high-throughput sequencing data. All profiling technologies would produce noise artefacts, and ChIP–Seq is also of no exception (Park, 2009). Thus, effective computational analysis of ChIP-Seq data will be crucial to generate biologically meaningful results. The purified DNA fragments from ChIP experiments can be sequenced by any of the next-generation platforms, such as Illumina Solexa Genome Analyzer, Roche 454 platform, and Applied Biosystems (ABI) SOLiD platforms (Shendure and Ji, 2008). The image data generated from the sequencing platforms are converted by the base caller software into sequence tags, which are referred as ChIP-Seq sequencing data. Preliminary analysis of the ChIP-Seq data consists of two major steps: a) mapping the sequence tags into reference genome; and b) peak-calling to find enriched regions as potential binding sites of the protein of interest, as shown in Fig 1.3.

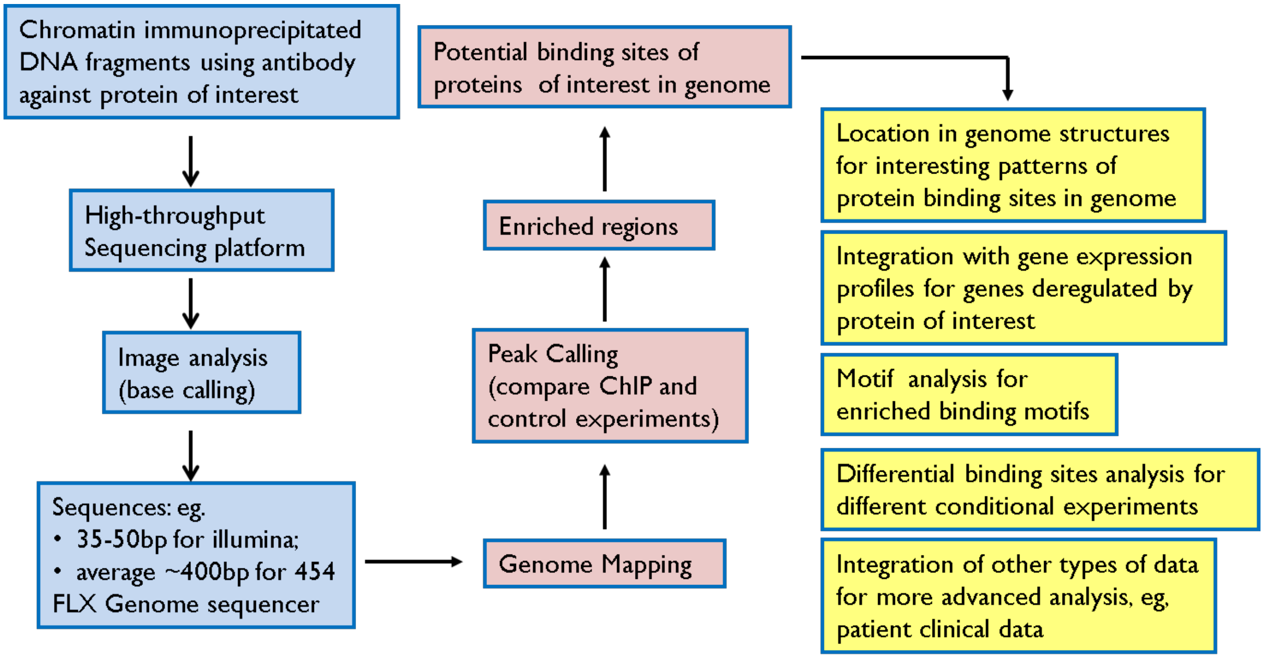

Chromatin immunoprecipitated DNA fragments using antibody against protein of interest
High-throughput Sequencing platform
Image analysis (base calling)
Sequences: eg.
• 35-50bp for illumina;
• average ~400bp for 454 FLX Genome sequencer
Genome Mapping
Peak Calling (compare ChIP and control experiments)
Enriched regions
Potential binding sites of proteins of interest in genome
Location in genome structures for interesting patterns of protein binding sites in genome
Integration with gene expression profiles for genes deregulated by protein of interest
Motif analysis for enriched binding motifs
Differential binding sites analysis for different conditional experiments
Integration of other types of data for more advanced analysis, eg, patient clinical data

Figure 1.3: Analysis of ChIP-Seq sequencing data. The images from the next-generation sequencing platform for chromatin immunoprecipitated DNA fragments using antibodies against protein of interest are first converted using base caller software into sequence tags, which will then be mapped to the reference genome. A step of peak calling comparing the ChIP-Seq profile with control sample profile will generate of list of enriched peak regions ranked by statistical significance measures representing the potential binding sites of the protein of interest in reference genome. Subsequently, the profiles of enriched regions can be further analyzed for more information, such as the binding motifs enriched, location of the binding sites in genome structures, integration of gene expression data, differential binding profile analysis, and so on. Processes for generation of sequencing data are highlighted in blue, while computational identification of genomic binding sites of proteins is highlighted in pink and post identification analysis is highlighted in yellow.

Mapping of sequence reads into the reference genome will give the intensities or counts of reads mapped to genome regions, and analysing the read intensities over the genome will produce a list of regions with enriched mapped reads (“peak-calling”), as the potential genome-wide binding sites of the protein of interest (Hoffman and Jones, 2009). With the profile of potential binding sites, further analysis can be done, such as, transcription factor binding motifs enrichment analysis, location of the binding regions over the genome relative to genome structures, correlation of gene expressions, differential binding sites between different cellular conditions, and so on (Park, 2009).

However, there are various potential sources of artefacts in ChIP-Seq experiments, which may result in the detection of insignificant peaks. For instance, shearing of DNA strands into fragments with a commonly used method like “sonication”, usually does not result in uniform fragmentation of the genome and thus leads to the uneven distribution of sequence tags across the genome, since some genome regions, such as open chromatin regions, are more easily fragmented than other genome regions, such as closed regions (Park, 2009). Therefore, in order to avoid such bias, control experiments are designed to pair up control profiles with the ChIP-Seq profiles so as to measure the significance of the peaks. These control samples used for sequencing are either input DNA which is a portion of the sheared DNA sample without immunoprecipitation, or mock DNA with DNA obtained from immunoprecipitation without antibodies, or DNA from nonspecific immunoprecipitation using an antibody against a protein that is not known to be involved in DNA binding. Input DNA has been used widely as the control sample in ChIP–Seq studies to remove the artefacts and bias from the ChIP-Seq experiments, such as the variable solubility of different regions, DNA shearing and amplification (Park, 2009). By comparing the read intensities in ChIP-Seq profile to the control sample profile at the paired-up genome regions, one can measure the significance of the peaks. Thus, the peak-calling step of the ChIP-Seq data analysis compares the ChIP sample profile to the control sample profile, and detects the potential enriched regions ranked by statistical significance measurements.

As for the very first step of analysis of the ChIP-Seq sequencing data, there are many different reference-guided short read mapping tools available, described earlier in Section 1.4, such as, Eland (part of the Illumina suite), GenomeMapper, RMAP, SeqMap, SHRiMP, ZOOM, NovoAlign, MOSAIK, MAQ, BWA and Bowtie. In particular, compared the other aligners, Bowtie is ultra-fast for Illumina short sequencing reads of uniform length of about 36bp, allowing multiple-core parallel processing and memory-efficient for large genomes, while maintaining a comparable mapping accuracy (Langmead et al., 2009). The Illumina sequencer produces single-end sequencing reads of short length, e.g. 36bp, and when aligning millions of reads of such short length to a large reference genome, e.g. human genome, a remarkable portion of the reads would probably match to multiple positions in the genome. In order to maintain the mapping accuracy, thresholds usually are set to remove the sequence reads that match ambiguously to the reference genome. Bowtie provides options for users to set the mapping thresholds, such as gapped or un-gapped alignment, number of mismatches allowed in the alignment, number of hits to output for users, and so on (Langmead et al., 2009). By setting the various options, one can decide the thresholds for the alignments between the sequence reads and reference genome, and achieve a balance between the mapping accuracy and the number of sequencing reads remained for peak-calling and advanced analysis.

Mapping of the sequencing reads generate the read intensities or counts within genome regions, and comparing the read intensities over genome regions in ChIP sample to the control sample can produce a list of peak regions where the reads are significantly enriched in ChIP sample over control sample (Hoffman and Jones, 2009). There are many different peak-calling software packages that utilize control sample profile, such as, E-RANGE (Mortazavi et al., 2008), spp package (Kharchenko et al., 2008), MACS (Zhang et al., 2008), QuEST (Valouev et al., 2008), SISSRs (Jothi et al., 2008), GLITR (Tuteja et al., 2009), PeakSeq (Rozowsky et al., 2009), CisGenome (Ji et al., 2011; Jiang et al., 2010), Sole-Search (Blahnik et al., 2010), and CCAT (Xu et al., 2010). A detailed comparison of various available peak-calling software algorithms is summarized in Table 1.3.

Table 1.3: Comparisons of various peak-calling algorithms for ChIP-Seq data, including E-RANGE, spp package, MACS, QuEST, SiSSRs, GLITR, PeakSeq, CisGenome, Sole-Search and CCAT. (Table adjusted and updated from Pepke et al., 2009)

Peak-caller	Signal-Profile	Tag shift or extension	Control Data	Peak Criteria	Peak ranked by	FDR (false discovery rate)	reference
E-RANGE	shift tag, aggregation	peak estimate or user-input	fold enrichment of ChIP over control, calculate *p* value	height cutoff, local peak estimate	*p* value	$\frac{\#\text{ peaks in control}}{\#\text{ peaks in ChIP}}$	Mortazavi, Williams et al. 2008
spp package	shift tag, window scan	estimate tag shift to maximize strand correlation	subtract control from ChIP before peak-calling	Poisson *p* value for paired peaks	*p* value	$\frac{\#\text{ peaks in control}}{\#\text{ peaks in ChIP}}$	Kharchenko, Tolstorukov et al. 2008
MACS	shift tag, window scan	estimate from peak-pairs	swap ChIP & control datasets to calculate FDR	local region Poisson *p* value	*p* value	$\frac{\#\text{ peaks in control}}{\#\text{ peaks in ChIP}}$	Zhang, Liu et al. 2008
QuEST	shift tag, kernel density estimation	estimate local shift to maximize strand correlation	fold enrichment, control data split into pseudo-ChIP to compute FDR	height cutoff, background ratio	*q* value	$\frac{\#\text{ peaks in pseudo-ChIP}}{\#\text{ peaks in ChIP}}$	Valouev, Johnson et al. 2008
SiSSRs	shift tag, window-scan	average distance of nearest tag pairs	compute fold enrichment of ChIP over control	N_+-N_- sign change, N_++N_- threshold in region	*p* value	control distribution	Jothi, Cuddapah et al. 2008
GLITR	extend tag, aggregation	user-input	control data split into pseudo-ChIP to calculate FDR	peak height cutoff & fold enrichment	peak height & fold enrichment	$\frac{\#\text{ peaks in pseudo-ChIP}}{\#\text{ peaks in ChIP}}$	(Tuteja, White et al. 2009
PeakSeq	extend tag, aggregation	user-input	significance of ChIP enrichment over control	local region binomial *p* value	*q* value	binomial for ChIP & control	Rozowsky, Euskirchen et al. 2009
CisGenome	shift tag, window-scan	average distance of peak-pairs	conditional binomial distribution to estimate FDR	number of reads in window, number of ChIP reads minus control reads	number of reads under peak	conditional binomial distribution of ChIP over control	Jiang, Wang et al. 2010; Ji, Jiang et al. 2011
Sole-Search	extend tag, window scan	user-input	determine peak height cutoff & calculate fold enrichment	peak height cutoff & enrichment significance cutoff (one sample t-test)	peak height & enrichment significance	$\frac{\#\text{ peaks in control}}{\#\text{ peaks in ChIP}}$	Blahnik, Dou et al. 2010
CCAT	shift tag, window scan	estimate from peak-pairs	swap ChIP & control datasets to calculate FDR	local region Poisson *p* value	*p* value	$\frac{\#\text{ peaks in control}}{\#\text{ peaks in ChIP}}$	Xu, Handoko et al. 2010

The peak-calling step in these software packages generally can be summarized into three basic sub-components: (i) generate signal profiles along each chromosome based on read/tag counts, (ii) find enriched peak regions in ChIP data relative to background control data (peak-calling) and (iii) assign statistical significance to filter out false positives and rank high-confidence peak calls (Pepke et al., 2009). Most algorithms generate smooth signal distributions/profiles using a fixed-width sliding window centered at each genome position and replacing the read/tag count in that genome position with the summed read counts within the window or modified signal values based on some assumptions of the distributions. Since the immunoprecipitated DNA fragments are double-stranded with the two strands equally likely to be sequenced from 5’ to 3’, the single-ended sequencing reads/tags are expected to come from both strands and form two density distributions (one for forward strand, and the other for reverse complement strand), which occur upstream and downstream with true DNA-protein crosslinking or binding sites in-between, as illustrated in Fig 1.4. Based on this bimodal enrichment pattern, programs like MACS, SiSSRs, spp package, QuEST, FindPeaks, E-RANGE, GLITR, and CCAT first shift the reads by half of the DNA fragment length (either user-defined or estimated from ChIP data) in a strand-specific manner and then build the signal profile based on the shifted read positions, such that, the corresponding distributions of two strands will overlay giving rise to a “summit” that has the local maximum and most likely represent the true DNA-protein binding sites. Some other programs may alternatively extend the genome location of the reads to accomplish the same goals. This strand-specific read shifting could considerably improve “summit” resolution and better locate the precise binding sites if the shifted distance is accurate (Pepke et al., 2009).

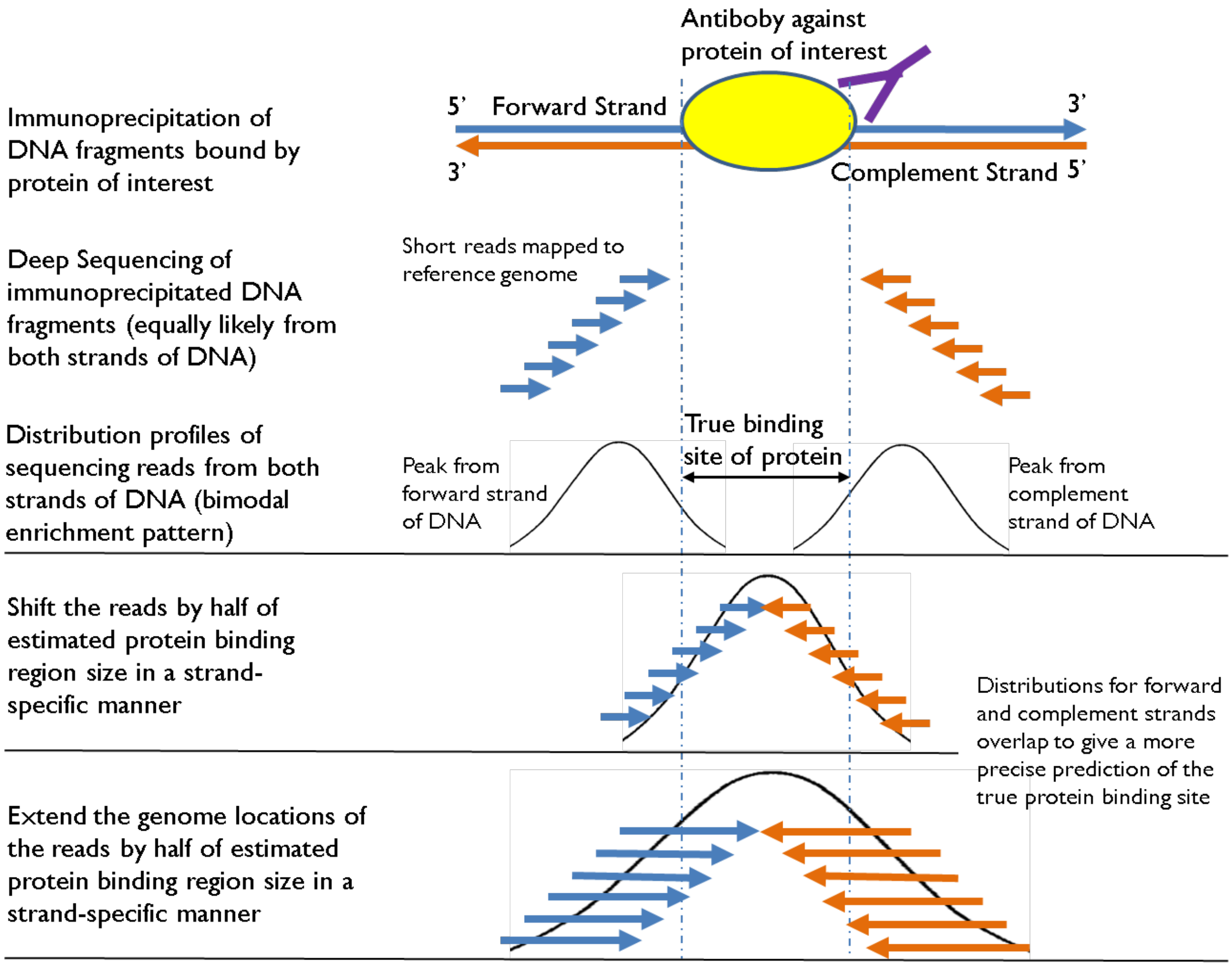


Figure 1.4: Bimodal enrichment pattern of ChIP-Seq sequencing data. Since the immunoprecipitated DNA fragments are double-stranded with the two strands equally likely to be sequenced from 5' to 3', the single-ended sequencing reads/tags are expected to come from both strands and form two density distributions (one for forward strand, and the other for reverse complement strand), which occur upstream and downstream with true DNA-protein crosslinking or binding sites in-between. In order to improve binding site detection resolution, some peak-calling algorithms first either shift the reads by half of the DNA fragment length or alternatively extend the genome location of the reads to the expected DNA fragment length in a strand-specific manner, and then build the signal profile based on the shifted or extended read positions, such that, the corresponding distributions of two strands will overlay giving rise to a "summit" that has the local maximum and most likely represent the true DNA-protein binding sites. This could significantly improve the precise detection of binding site location. (Figure redrawn from Park, 2009)

When comparing the ChIP profile to control sample profile, most peak-calling programs calculate fold enrichment of reads in ChIP over control sample along genome regions, and assign statistical significance to each enriched peak in ChIP data. Different programs employed different methods to compute the significance to filter out false positives and rank for high-confidence peaks. For example, some built sophisticated statistical models from control data to assess the significance of ChIP peaks (Blahnik et al., 2010; Boyle et al., 2008; Ji et al., 2008; Mortazavi et al., 2008; Nix et al., 2008; Qin et al., 2010; Rozowsky et al., 2009; Spyrou et al., 2009; Valouev et al., 2008; Xu et al., 2010; Zhang et al., 2008), some calculate empirical false discovery rate by swapping ChIP and control data to identify enriched peaks in control data (False Discovery Rate (FDR) = number of peaks in control / number of peaks in ChIP) (Kharchenko et al., 2008; Lun et al., 2009; Xu et al., 2010; Zhang et al., 2008), and some calculate FDR by partitioning control data to generate pseudo-ChIP data if control data is large enough (Tuteja et al., 2009; Valouev et al., 2008). Among these peak-calling algorithms, MACS has been evaluated to be superior over others with good sensitivity and specificity that gives higher true positive rates, higher ranking accuracy, better peak positional accuracy and precision (spatial resolution) (Wilbanks and Facciotti, 2010). MACS algorithm will (1) first remove duplicate reads in the datasets that may arise from ChIP-DNA amplification and sequencing library preparation, (2) linearly scale the total number of reads in control data to be the same with that in ChIP data, (3) empirically model the size of the true protein binding site based on the bimodal enrichment pattern, (4) shift the genome locations of the reads in a strand-specific manner by half of the estimated size of the protein binding site, (5) scan the genome using sliding windows of user-defined width to identify candidate peaks with significant read enrichment based on *p*-values calculated from dynamic Poisson distribution of reads, (6) swap ChIP and control datasets and call peaks in control data, and (7) calculate FDR for each detected peak in ChIP data and rank them using the *p*-values. CCAT applies similar algorithm with MACS but is superior to MACS as it estimates noise rate and resample the datasets to balance ChIP and control sample sizes instead of using linear scaling as in MACS (Xu et al., 2010).

1.9 Motif Enrichment Analysis to Identify Co-Factors of Proteins

With the profile of potential protein binding sites identified from peak-calling tools, further analysis can be done, such as, binding motifs enriched, location of the binding regions over the genome, correlation with gene expressions, differential binding sites between different cellular conditions, and so on (Park, 2009). For example, motif enrichment analysis can predict the DNA-binding motifs for the protein of interest by first extracting the genomic sequences of the identified DNA-binding sites from the reference genome, and then scanning against known DNA-binding motifs, e.g. TRANSFAC database (Wingender et al., 1996), or predicting novel binding motifs using *de novo* motif finding algorithms. Proteins such as transcription factors, generally do not work alone and usually function with other transcription factors (co-factors) in a combinatory fashion to regulate target gene expressions precisely. Particularly, in situations of where factors co-associated with the protein of interest are present, ChIP-Seq protocol using antibodies against the protein of interest could immunoprecipitate the DNA fragments bound by the protein of interest that is meanwhile co-localized with its interacting factors (Liu et al., 2010). In such cases, motif enrichment analysis of the predicted DNA-binding sites against known motif databases will help identify the interacting DNA-binding co-factors for the protein of interest globally.

Sequence motif discovery algorithms play an important part in order to better understand the protein-DNA interaction mechanisms, and the structures and functions of proteins (Bailey, 2008). There are various motif finding software tools available for ChIP-Seq data, including MDScan (Liu et al., 2002), Weeder (Pavesi et al., 2004), MEME (Bailey et al., 2009; Bailey et al., 2006), GALF-G (Chan et al., 2009), Tmod (Sun et al., 2010), HOMER (Heinz et al., 2010), HMS (Hu et al., 2010), recently published CENTDIST (Zhang et al., 2011) sand DREME (Bailey, 2011), etc. Most motif finding algorithms scan the genomic sequences of the DNA-binding regions identified from ChIP-Seq data, and search for either statistically overrepresented word-based oligonucleotide (motifs) with high occurrence frequency, or probabilistic sequence models with model parameters estimated from maximum-likelihood or by Bayesian inference (Das and Dai, 2007). The word-based algorithm guarantees global optimality and is suitable to identify short motifs in eukaryotic genomes, while the probabilistic approach involves representation of motif models by position weight matrix (Bucher, 1990) and is appropriate for longer motifs. All these algorithms mentioned have been reported to be able to correctly detect the motifs that were previously detected by laboratory experimental approaches, and some *de novo* motif finding algorithms can find novel motifs. In particular, for eukaryotic genomes, HOMER (Hypergeometric Optimization of Motif Enrichment) motif enrichment algorithm against known motifs database comprises of the following steps: (i) first, the program randomly selects a set of background genomic sequences with similar length and GC content to the target sequences (potential binding sites), (ii) it then assigns weights to each background sequence to minimize the imbalance in sequence contents with the target sequences, (iii) it further calculates the occurrence of each known motif in the target and background sequences and (iv) then computes a significance value (e.g. *p*-value) for the enrichment of the motif in the target sequences over background sequences (Heinz et al., 2010). The motifs identified are ranked by their enrichment significance values produced by the motif finder algorithm, and usually significantly enriched motifs with significance measurement values passing defined threshold (e.g. p-value < 0.05) are selected as the potential co-factors that may interact with the protein of interest when binding to genomic DNA.

In addition to identifying co-factors for protein of interest, ChIP-Seq data also has the capacity to permit further analysis to uncover protein-DNA interaction patterns and gene regulation mechanisms, such as correlation with gene expression profiles, relationship of DNA-binding sites to genome structures, differential binding sites of proteins in response to different cellular conditions, and so on. In summary, our laboratory has employed ChIP-Seq Illumina ultra high-throughput sequencing technique with antibodies specifically against HBx protein for immunoprecipitation, and in this study, I implemented a computational workflow to analyse the ultra high-throughput ChIP-Seq sequencing data, identify potential genomic binding site of HBx in a genome-wide scale, discover transcription factors that potentially interact with HBx form motif enrichment analysis, and identify potential deregulated direct gene targets of HBx from integration of gene expression profiles, for a better understanding of the underlying mechanisms of HBV-induced hepatocarcinogenesis.

1.10 Project Objectives

Chronic HBV infections may gradually lead to the development of HCC in patients (Arbuthnot and Kew, 2001; Bonilla Guerrero and Roberts, 2005; Buendia, 1992; Chang, 2003; Lupberger and Hildt, 2007; Parkin et al., 2001; Robinson, 1994). However, the association between chronic HBV infection and HCC hepatocarcinogenesis remains incompletely understood (Ng and Lee, 2011), though a few underlying mechanisms have been proposed by numerous studies, such as, HBV genome integration into human genome (Bill and Summers, 2004; Bonilla Guerrero and Roberts, 2005; Buendia, 1992; Goto et al., 1993; Jiang et al., 2012; Murakami et al., 2005; Pineau et al., 1998; Robinson, 1994; Saigo et al., 2008; Tan, 2011; Tu et al., 2006), HBx deregulation of host genes expression through interactions with transcription factors (Andrisani and Barnabas, 1999; Ganem, 2001; Sung et al., 2009; Wu et al., 2001) or through deregulation of regulatory microRNA expressions (Kong et al., 2011; Shan et al., 2011; Wang et al., 2010; Wang et al., 2012; Wu et al., 2011; Yip et al., 2011; Yuan et al., 2012) or through epigenetic modifications (e.g. DNA methylation status of genes) (Arzumanyan et al., 2012; Huang et al., 2010; Jung et al., 2010; Kim et al., 2010; Madzima et al., 2011; Park et al., 2011; Su et al., 2008; Um et al., 2011; Zhu et al., 2010). In this project, we focused on two essential underlying mechanisms: HBV genome integration as described in Chapter 2, and HBx deregulation of host gene expression through interactions with transcription factors, as described in Chapter 3 of this MSc thesis.

1.10.1 Computational Analysis for Characterization of HBV-Host Genome Integration Sites

HBV-host genome integration is very commonly observed in HBV-associated HCC cases, and is believed to be one of the key mechanisms involved in hepatocarcinogenesis. HBV integration into the host genome could result in possible disruption of host gene expressions, expression of modified viral proteins or viral-host chimeric proteins that are potentially oncogenic and contribute to hepatocarcinogenesis. Therefore studying where HBV genome integrates into the host genome will promote understanding of the underlying mechanisms of how HBV infection gradually leads to development of HCC. HBV genome integration into human genome has been reported for many years, however, the details of how HBV genome integration may contribute to hepatocarcinogenesis is still incompletely understood (Ng and Lee, 2011). To address it, the very first step is to know the locations where HBV DNA is inserted to human DNA. To comprehensively characterize HBV genome integration sites and study the variations of HBV DNA in HCC patients, with limited prior knowledge on how HBV DNA fuses with human genome, our laboratory employed targeted high throughput FLX sequencing techniques on 48 HBV-associated HCC patients' samples (tumor and adjacent non-tumor tissues) to enrich for HBV-containing DNA fragments. To maximize the targeted sequencing capacity, a set of 6bp "barcodes" was used to label for individual patient samples, allowing sequencing of multiple samples in a single run. Millions of sequencing reads of variable lengths were generated. In this project, I implemented a computational workflow to analyse the high-throughput sequencing data and identify sequences carrying both HBV and human DNA within the same sequence where the HBV-HG integration sites can be inferred. The analysis pipeline aimed to accomplish the following specific objectives at various analysis steps:

Specific Objective 1: Remove noise and insignificant reads from the ultra high-throughput sequencing data

Hybrid capture strategy to enrich for HBV sequences has limited enrichment efficiency. Due to the sequence similarities between HBV and human genomes, non-specific hybridization of HBV probes with human genome sequences may potentially capture insignificant DNA fragments that do not contain HBV sequences or HBV-human integration boundaries. In addition, the last elution step of hybrid capture method may also cause noise DNA fragments to be sequenced. Thus the pool of millions of raw sequencing reads from ultra high-throughput FLX sequencer may contain lots of insignificant and noise reads, which purely belong to human genome and neither contain HBV sequences nor HBV-human integration boundaries. To identify HBV-containing reads, the very first step is to filter out the pure human sequences and the sequences that do not contain HBV sequences, from the large raw sequence library. Mapping the raw sequencing reads against human genome and HBV genome respectively may help to achieve this objective.

Specific Objective 2: Identify HBV-host integration boundaries by aligning sequence reads against both HBV and human genomes

Removal of noise and insignificant reads will largely reduce the size of the sequencing library. FLX 454 pyrosequencer is reported to have an average read length of 400 bases, and thus can largely reduce ambiguities when searching for read identities against genome databases. More importantly, this particular long read length feature of FLX sequencer could allow us to identify HBV-host integration boundaries directly from the raw sequence reads that can be long enough to accommodate the junction sites. A sequence read containing a HBV-host junction site is hypothesized to have at least one region of the sequence aligned to human genome and the other region aligned to HBV genome. Therefore, after removal of the noise sequence reads from the FLX sequencing library, potential HBV-host junction sites could be directly identified from the rest of the raw sequence reads by aligning the sequence reads against both human and HBV genomes and examining the hits of each sequence read to the two reference genomes.

Nevertheless, one should not neglect the possibility that DNA shearing or fragmentation in the initial ChIP-Seq experimental step may happen to occur at the HBV-host junction sites, in which case, those HBV-host junction sites might be disrupted and missed out in the sequencing mapping analysis. Therefore, in this study and in my analysis pipeline, a step of *de novo* assembling of sequence reads into longer sequences which we called "contigs" was also incorporated, with the purpose of recovering any possible disrupted HBV-host junction sites. Furthermore, assembly provides representative or consensus sequences ("contigs") by merging overlapping reads, and this could help us to reconstruct the original long DNA sequences from the fragmented DNA sequence reads. Thereby, the "contigs" and the remaining unassembled sequence reads could be mapped against both human and HBV genomes respectively in order to identify HBV-host integration boundaries.

Specific Objective 3: Perform post-identification analysis to get more detailed information on HBV-host integrations

After identification of HBV-host integration boundaries in HCC patients, more detailed post-identification analysis may be performed to understand HBV-host integration mechanisms. Questions to be answered may include: whether the integration of HBV genome into human genome is a random process or do they follow any conserved patterns; whether there are any integration sites conserved among different patient tissues; whether there is any obvious difference in the integration sites between tumor and adjacent non-tumor tissues of HCC patients;

whether there is any HBV gene that is more preferred to be integrated into the human genome; whether there is any functional domain of HBV genes that is often conserved and integrated in the human genome; whether there is any host gene that is disrupted by integration; what is the functional importance of disrupted host genes; and so on. Efforts trying to answer these biological questions may help us identify genes or factors of HBV that are important for hepatocarcinogenesis and further understand the details of HBV-host genome integrations in HCC patients with chronic HBV infections.

1.10.2 Computational Analysis for Identification of Putative Deregulated Direct Gene Targets of HBx

The HBV genome of length 3215bp consists of 4 major genes encoding for core protein, surface antigen protein, polymerase, and viral X-gene. Among the four genes, the viral X-gene (HBx) of HBV is conserved among all mammalian hepadnaviruses and the small protein (154 amino acids) encoded has been strongly implicated to play a major role in hepatocarcinogenesis and the development of HCC in chronic HBV-infected patients. HBx protein acts as a transactivator that disrupts host cellular gene expressions and subsequent cellular pathways which may lead to cancer. HBx has been reported to deregulate host genes expression through various mechanisms, such as, through interactions with transcription factors (Andrisani and Barnabas, 1999; Ganem, 2001; Sung et al., 2009; Wu et al., 2001), through deregulation of regulatory microRNA expressions (Kong et al., 2011; Shan et al., 2011; Wang et al., 2010; Wang et al., 2012; Wu et al., 2011; Yip et al., 2011; Yuan et al., 2012), or through epigenetic modifications (e.g. DNA methylation status of genes) (Arzumanyan et al., 2012; Huang et al., 2010; Jung et al., 2010; Kim et al., 2010; Madzima et al., 2011; Park et al., 2011; Su et al., 2008; Um et al., 2011; Zhu et al., 2010). In this project, we focused on HBx deregulation of host gene expressions through interactions with transcription factors.

Deregulating host gene expressions through interactions with transcription factors has been known to be one of the major underlying mechanisms that HBx plays in hepatocarcinogenesis. HBx protein does not bind to DNA directly, but through interacting with transcription factors. It regulates gene expression by changing the DNA binding affinities of transcription factors. Systematically identifying the list of transcription factors that HBx interacts and the direct target genes of HBx-transcription factor complex could provide further insights into HBx functions in the development of HCC. Our laboratory has previously systematically profiled HBx genomic binding sites and HBx-interacting transcription factors using ChIP-chip method on 1.5kb promoter regions of human genes (Sung et al., 2009). However, there are various limitations and bias associated with ChIP-chip, as discussed in Section 1.6. Therefore, our laboratory has turned to apply ChIP-Seq technology coupled with Illumina high-throughput sequencing technique on primary liver cell line THLE3 transfected with HBx-expressing adenoviruses, to identify a more comprehensive and unbiased list of HBx genomic binding sites. ChIP-Seq technique uses antibody specific to the protein of interest to immunoprecipitate DNA fragments bound by protein of interest, followed by size selection and sequencing of the enriched DNA fragments. The sequencing step in ChIP-Seq enables identification of genome-wide DNA-protein binding sites, and has become the major trend in the field of studying protein-DNA interactions. The Illumina sequencing approach applied by our laboratory produced millions of short sequence reads of 36bp for both control and HBx-expressing THLE3 cell samples. In addition, previous studies on HBx deregulation of host genes were mainly carried out in cell lines due to lack of patients data. In this study, with the availability of HCC patients data in our laboratory, I implemented a computational workflow to analyse the ChIP-Seq sequencing data and integrate the microarray expression profiles and clinical data of 100 HBV-associated HCC patients to identify a more comprehensive list of potential genomic binding sites of HBx, HBx-interacting transcription factors and potential HBx deregulated direct gene targets with clinical inferences in HCC patients. The analysis process aims to accomplish the following objectives:

Specific Objective 4: Align ChIP-Seq Illumina sequencing reads to human genome and remove reads mapped ambiguously to human genome

The Illumina Solexa sequencer produced millions of single-end sequence reads of 36bp, out of which, a significant portion might match to multiple positions in the human genome, because of the short read length (36bp) and the large human genome size. In order to maintain the mapping accuracy, thresholds must be set to remove sequence reads that match ambiguously to human genome. An ultrafast reference-guided short read aligner Bowtie (Langmead et al., 2009) provides options for users to set the thresholds, such as gapped or un-gapped alignment, number of mismatches allowed in the alignment, number of matches to output for users, and so on. Mapping criteria/parameters need to be carefully selected in this project to achieve a balance between the mapping accuracy and the amount of sequencing reads remained usable for downstream analysis.

Specific Objective 5: Identify significantly enriched peak regions as potential DNA binding sites of HBx using peak-calling algorithms

Following aligning sequence reads to human genome and filtering out reads with ambiguous matches in human genome, the next step is to analyse the intensities (counts) of reads mapped on genome regions and identify regions, where the reads are significantly enriched, as the potential HBx binding sites. This peak-calling step will compare the read intensities in HBx-expressing THLE3 cells against the read intensities in control THLE3 cells, and then identify regions (peaks) with reads significantly enriched in HBx-expressing THLE3 cells. Our control THLE3 ChIP-Seq data was important because it served as background noise model to help filter out false positive regions that might come from DNA shearing biases, antibody immunoprecipitation biases or sequencing artefacts. The peak-calling step generally can be summarized into three basic sub-components: (i) generate signal profiles along each chromosome based on read/tag counts, (ii) find enriched peak regions in ChIP data relative to background control data (peak-calling) and (iii) assign statistical significance to filter out false positives and rank high-confidence peak calls. Since the immunoprecipitated DNA fragments are double-stranded with the two strands equally likely to be sequenced from 5’ to 3’, the single-ended reads/tags are expected to come from both strands and form two density distributions (one for forward strand, and the other for reverse complement strand), which occur upstream and downstream with true DNA-protein crosslinking or binding sites in-between. Thus, the sequencing reads are expected to show a bimodal enrichment pattern for a true binding site, and therefore, strand-specific shifting or extending of the reads could yield more precise prediction of DNA binding sites. Two peak-calling tools MACS (http://liulab.dfci.harvard.edu/MACS/) and CCAT (http://cmb.gis.a-star.edu.sg/ChIPSeq/paperCCAT.html) were utilized in this project to first shift the genomic positions of the reads by half of the estimated DNA fragment length in a strand-specific manner and then call peaks with significant read enrichments in ChIP data relative to background control data. Peaks commonly predicted from the two peak-calling algorithms (MACS and CCAT) can be selected as the potential genomic binding sites of HBx, and this application of multiple peak-calling algorithms on the same dataset will give us more confidence on the predictions.

Specific Objective 6: Identify potential HBx-interacting transcription factors from motif enrichment analysis

From the list of enriched peaks representing potential binding sites of HBx over human genome, motif enrichment analysis was performed to predict the transcription factor binding motifs enriched within the predicted candidate binding sites of HBx, to obtain a list of predicted transcription factors that potentially interact with HBx to bind to DNA. This analysis consists of two major steps: a) extraction of the genomic DNA sequences for the candidate binding sites of HBx, and 2) scanning of these peak sequences against the known human transcription factor motifs using known motif finder algorithms. For example, HOMER (http://biowhat.ucsd.edu/homer/chipseq/), developed in the Glass lab of UCSD (Heinz et al., 2010), first randomly selects a set of background genome sequences of similar length and GC content to the target sequences (potential binding sites), assesses the occurrences of each known motif in the background and target sequences, and calculates significance values for each known motif enriched in target sequences relative to background sequences. Thereafter, the enriched motifs could then be ranked based on the significance values produced by HOMER and those motifs with significance values above the threshold were considered as the transcription factors that may potentially interact with HBx and bind to genome DNA.

Specific Objective 7: Identify potential direct gene targets of HBx by integrating microarray expression profiles for THLE3 cell line

A list of differentially expressed genes was identified using a two-colour expression profiling array of HBx-expressing and control THLE3 cells. These differentially expressed genes were hypothesized to be deregulated (either up- or down-regulated) upon the presence of HBx protein in THLE3 cells. In this study, through the integration of the microarray gene expression profiles and ChIP-Seq data, a list of putative direct gene targets of HBx deregulated through HBx-transcription factor interactions could be identified.

Specific Objective 8: Evaluate the clinical relevance of the potential deregulated direct gene targets of HBx

Since these potential deregulated direct gene targets of HBx were predicted from primary liver cell line (THLE3), we evaluated if these deregulated gene targets of HBx are clinically relevant. To address this, the microarray gene expression profiles and the clinical data collected from 100 HBV-associated HCC patients were integrated. The expression values for the potential HBx deregulated direct gene targets in the tumor and adjacent non-tumor tissues of the 100 HCC patients were first examined. As HBx is reported to have oncogenic potential, gene targets, which were appropriately differentially expressed in tumor over adjacent non-tumor tissues in HCC patients similar to what was observed in THLE3 cells upon the presence/expression of HBx protein, were selected for further analysis as these are likely to be related to hepatocarcinogenesis. With the availability of the clinical data of the 100 HCC patients, associations between these genes and the patients' clinical characteristics and survival potential can then be evaluated by performing various statistical tests, such as T-test, One way ANOVA, non-parametric tests (median test, Mann-Whitney U test and Kruskal-Wallis test), and Kaplan-Meier survival test. Gene targets with significant clinical associations were considered to be highly likely to have clinical inferences of HCC patients.

In summary, in this section of this project, I integrated the ChIP-Seq data and microarray expression profiles from both cell line and HCC patients, as well as clinical data from 100 HBV-associated HCC patients, and performed a series of computational analysis as described above, to identify genome-wide potential HBx-binding sites, potential HBx-interacting transcription factors, and putative clinically associated HBx direct gene targets that were deregulated indirectly by HBx through interactions with transcription factors. Previous studies on HBx deregulation of host genes were mainly carried out in cell lines due to lack of patients data. In this study, we are the very first to integrate the clinical data of a large series of HCC patients and identify potential HBx deregulated gene targets with significant clinical inferences. Identification of clinically significant direct gene targets of HBx may help us to further understand the underlying mechanisms of HBV-induced hepatocarcinogenesis, and facilitate future discovery of potential drug targets and novel drug therapies for HCC.

1.10.3 Summary of Project Objectives

Figure 1.5 briefly summarizes the various specific objectives of this project. By characterization of HBV genome integration sites and identification of clinically associated deregulated gene targets of HBx, we aim to get more insights of the two essential underlying mechanisms that may potentially contribute to HBV-induced hepatocarcinogenesis in HCC patients: HBV genome integration into human genome and HBx deregulation of host gene expressions through interactions with transcription factors. In the following sections of this MSc thesis, Chapter 2 will describe the experimental design, data analysis pipeline, and interpretation of results on HBV-host junction sites obtained by analysing the FLX high-throughput sequencing data of enriched HBV-containing DNA fragments extracted from 48 HBV-associated HCC patients. Chapter 3 will describe the ChIP-Seq experimental design, data analysis pipeline and interpretation of results on HBx genomic binding sites, HBx-interacting transcription factors, and clinically associated deregulated direct gene targets of HBx obtained by integrating the ChIP-Seq data, microarray expression profiles for both liver cell line (THLE3) and HCC patients, as well as the clinical data of 100 HCC patients. Chapter 4 will give a summary and conclusion of this study.

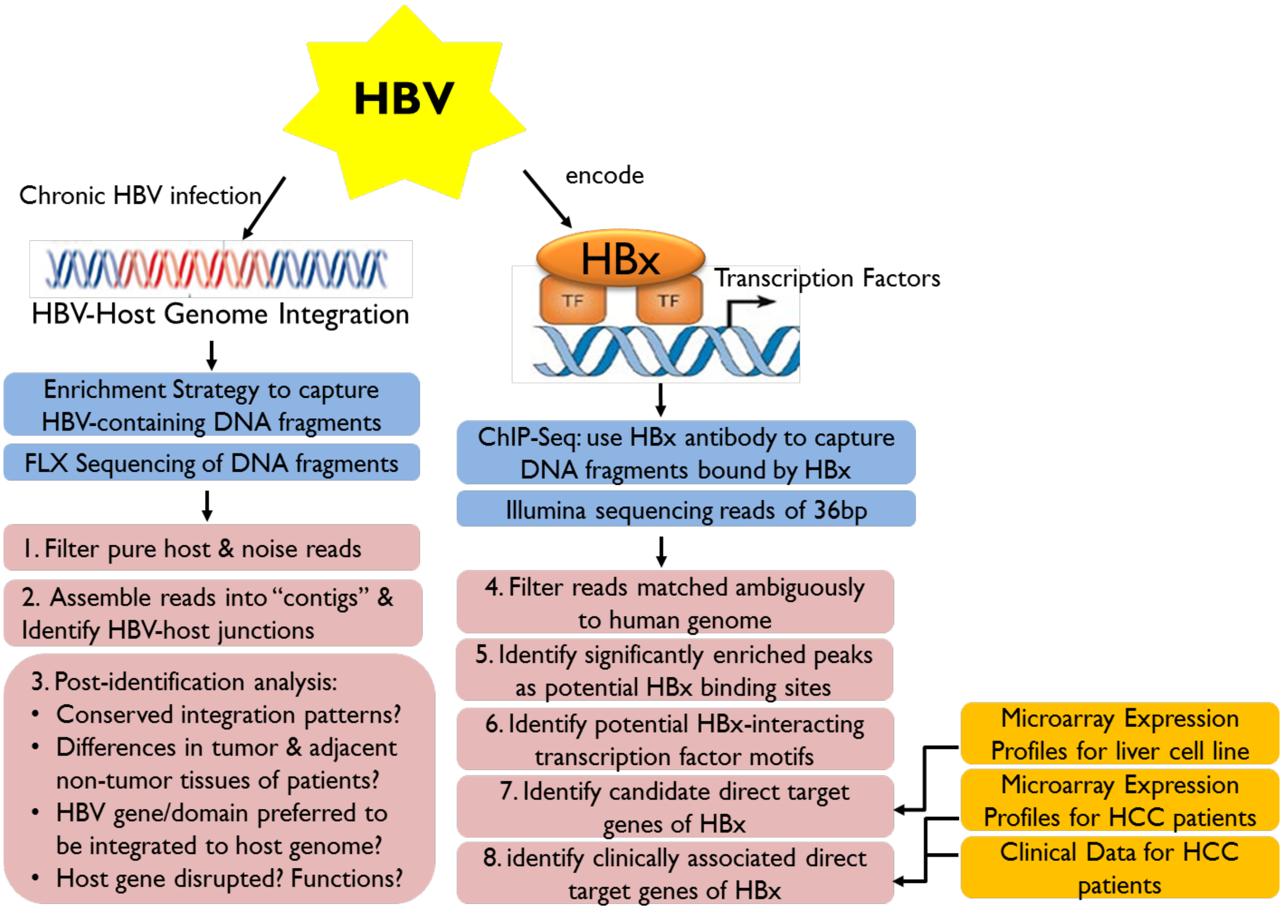

HBV
Chronic HBV infection
encode
HBx
TF
TF
Transcription Factors
HBV-Host Genome Integration
Enrichment Strategy to capture HBV-containing DNA fragments
FLX Sequencing of DNA fragments
1. Filter pure host & noise reads
2. Assemble reads into "contigs" & Identify HBV-host junctions
3. Post-identification analysis:
• Conserved integration patterns?
• Differences in tumor & adjacent non-tumor tissues of patients?
• HBV gene/domain preferred to be integrated to host genome?
• Host gene disrupted? Functions?
ChIP-Seq: use HBx antibody to capture DNA fragments bound by HBx
Illumina sequencing reads of 36bp
4. Filter reads matched ambiguously to human genome
5. Identify significantly enriched peaks as potential HBx binding sites
6. Identify potential HBx-interacting transcription factor motifs
7. Identify candidate direct target genes of HBx
8. identify clinically associated direct target genes of HBx
Microarray Expression Profiles for liver cell line
Microarray Expression Profiles for HCC patients
Clinical Data for HCC patients

Figure 1.5: Aims of the project. Our laboratory has applied a targeted NGS technique to identify HBV-human integration sites. I implemented a workflow to analyze the high-throughput FLX sequencing data to: 1) first remove pure host and noise reads by aligning reads to human and HBV genomes, 2) *de novo* assemble the reads into longer "contigs" to reconstruct original DNA sequences and recover junction sites disrupted from DNA shearing; then identify HBV-host junction sites by searching "contigs" against human and HBV genomes, and last 3) perform post-identification analysis to get more detailed information on HBV-host genome integrations. Our laboratory also applied ChIP-Seq technique with HBx antibodies to profile genomic binding sites of HBx. I integrated the ChIP-Seq sequencing data, microarray expression profiles and HCC patient clinical data to: 4) first remove ChIP-Seq sequencing reads that matched ambiguously to the human genome, 5) identify significantly enriched peak regions as potential DNA binding sites of HBx using peak-calling algorithms, 6) identify potential HBx-interacting transcription factors from motif enrichment analysis, 7) identify potential deregulated direct gene targets of HBx by integrating microarray expression profiles, and last 8) identify clinically significant deregulated direct gene targets of HBx by integrating clinical features and survival time data of 100 HBV-associated HCC patients. The high-throughput sequencing data used in this project are highlighted in blue, while the microarray expression profiles and clinical data used are highlighted in orange and the computational analysis processes are highlighted in pink.

CHAPTER 2: Computational Characterization of HBV-Host Genome Integration Sites

2.1 Materials and Methods

2.1.1 Data Collection: HBV-containing DNA Fragments Enrichment and FLX Sequencing Library Construction

DNA samples were extracted from tumor and adjacent non-tumor tissues of 48 HBV-positive HCC patients, and a total of 96 FLX sequencing libraries were then constructed in our laboratory following the protocol briefly shown in Fig 2.1. The key step of the HBV sequence enrichment strategy was to use the specifically designed HBV probes to pull down HBV-containing DNA fragments from extracted patient DNA samples. To ensure maximum coverage of the whole HBV genome with minimum bias, 26 3' biotinylated HBV probes which are 70-mer long were specifically designed based primarily on conserved regions. Nevertheless, probes within some less conserved regions of 96 genotype B and genotype C HBV genome sequences downloaded from NCBI data repository had to be designed as well. In order to make full use of the sequencing capacity, pooled FLX sequencing of all the 96 tissue samples was done, where each sample library had a unique 6bp library barcode attached at the 5' of the DNA fragments for identification. This enrichment strategy facilitated the capture of novel HBV-containing sequences from the 48 pairs of patient DNA samples, without prior knowledge on which part of HBV genome integrated to human genome.

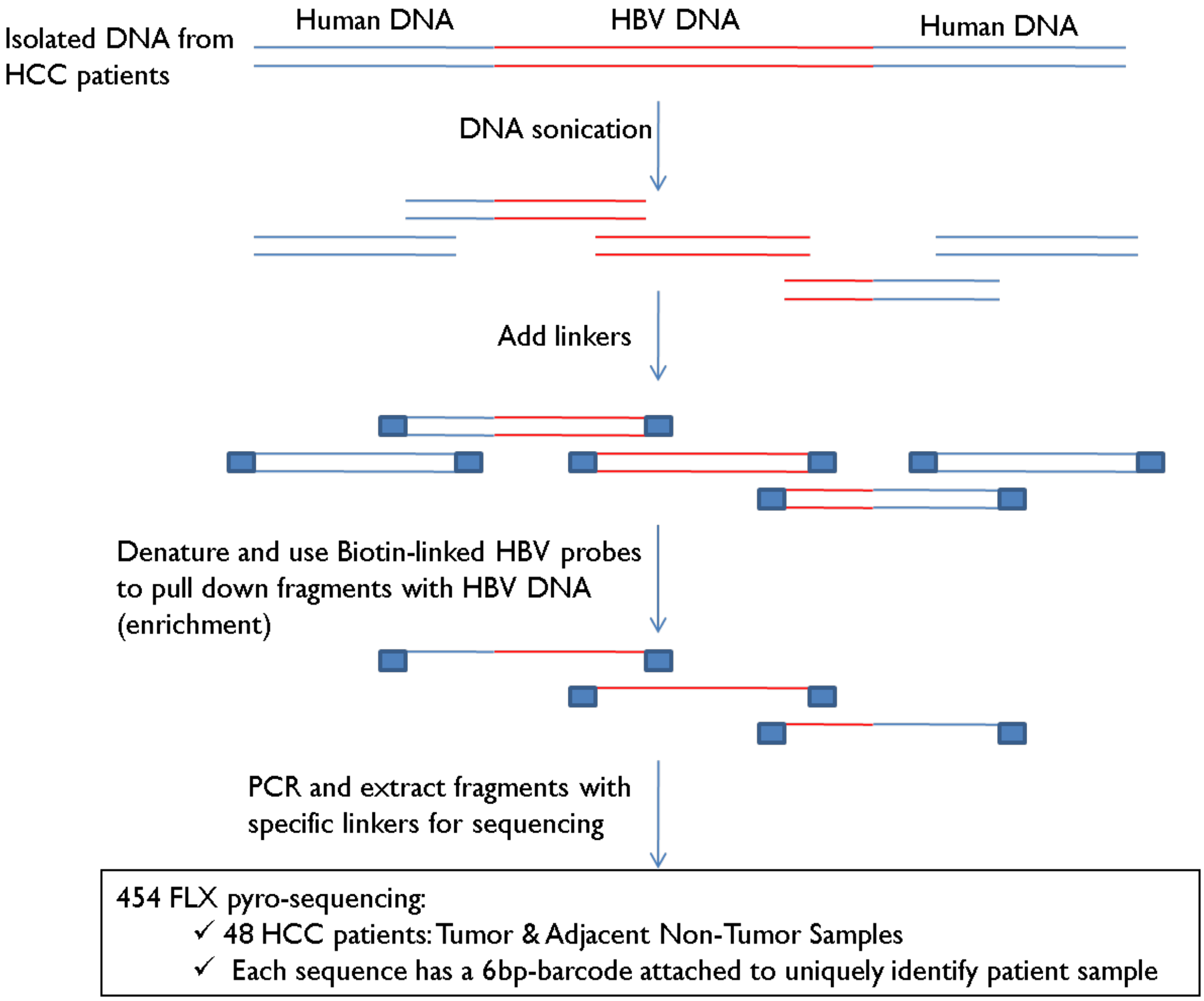

Isolated DNA from HCC patients
Human DNA
HBV DNA
Human DNA
DNA sonication
Add linkers
Denature and use Biotin-linked HBV probes to pull down fragments with HBV DNA (enrichment)
PCR and extract fragments with specific linkers for sequencing
454 FLX pyro-sequencing:
✓ 48 HCC patients: Tumor & Adjacent Non-Tumor Samples
✓ Each sequence has a 6bp-barcode attached to uniquely identify patient sample

Figure 2.1: Flowchart of the HBV enrichment strategy applied in our laboratory. Each extracted DNA sample was first sonicated into small fragments of 300 to 800bp, and linkers were added for sequencing purpose. The double-stranded DNA fragments were then denatured into single-stranded fragments and twenty-six 3' biotinylated HBV probes of length 70-mer were then used to pull down the fragments that contain HBV genome sequences. The HBV-containing DNA fragments were then undergone pooled FLX sequencing, where each sample library had a unique 6bp library barcode added to the 5' of the sequence reads for identification.

2.1.2 Computational Identification of HBV-Host Junction Sites from FLX Sequencing Data

A specific analysis pipeline was developed to identify the HBV-human junctions from the 454 FLX sequencing reads (Fig.2.2).

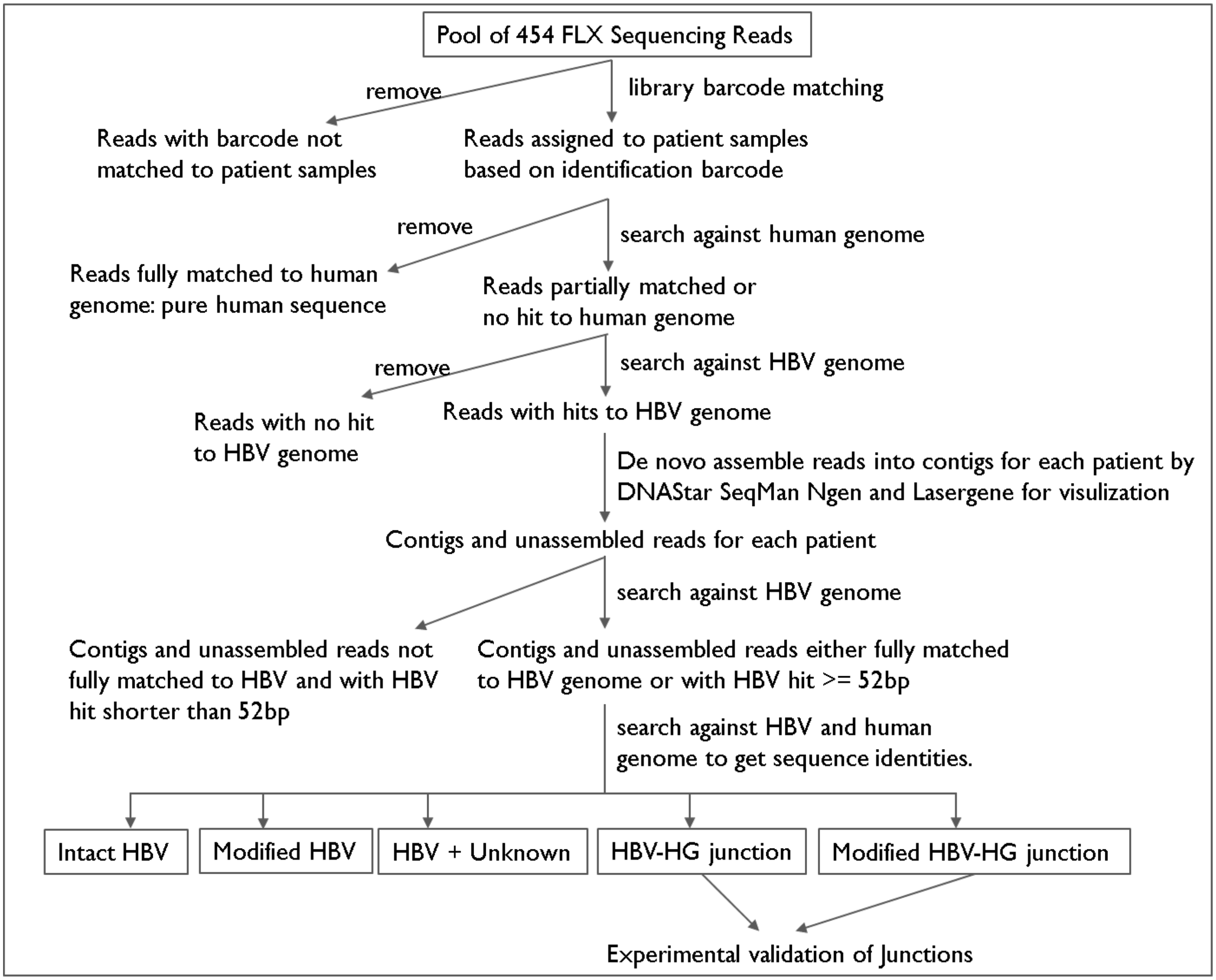

Pool of 454 FLX Sequencing Reads
remove
library barcode matching
Reads with barcode not matched to patient samples
Reads assigned to patient samples based on identification barcode
remove
search against human genome
Reads fully matched to human genome: pure human sequence
Reads partially matched or no hit to human genome
remove
search against HBV genome
Reads with no hit to HBV genome
Reads with hits to HBV genome
De novo assemble reads into contigs for each patient by DNAStar SeqMan Ngen and Lasergene for visulization
Contigs and unassembled reads for each patient
search against HBV genome
Contigs and unassembled reads not fully matched to HBV and with HBV hit shorter than 52bp
Contigs and unassembled reads either fully matched to HBV genome or with HBV hit >= 52bp
search against HBV and human genome to get sequence identities.
Intact HBV
Modified HBV
HBV + Unknown
HBV-HG junction
Modified HBV-HG junction
Experimental validation of Junctions

Figure 2.2: Analysis pipeline for identification of HBV-human junctions from 454 FLX sequencing data. NCBI standalone soft tool BLASTN 2.2.23+ was used to search sequence reads against human genome (NCBI reference contig assembly of Build37) and HBV genome (human HBV strain genome sequences downloaded from NCBI genome database). Reads not assigned to any patient samples, or fully matched to human genome, or with no hit to HBV genome were filtered out. Then *de novo* assembly algorithm from DNASTAR SeqMan NGen was used to assemble the rest reads into contigs for each patient. Contigs and unassembled reads, which were either fully aligned to HBV genome or with at least 52 consecutive bases aligned to HBV genome, were selected for further identifications. The sequence identities were classified into five groups where the two groups “HBV-HG junction” and “Modified HBV-HG junction” contained the predicted junction points where HBV sequences insert into human genome. A graphic representation of the typical sequences identities is shown in Fig 2.3.

From the pool of millions of raw sequence reads, those reads whereby the 6bp barcodes were not matched to any patient were first removed, and the remaining reads were assigned to different patients based on the unique identification barcodes. The assigned reads were then searched against human genome (NCBI reference contig assembly Build37) using NCBI standalone soft tool BLASTN 2.2.23+ (Basic Local Alignment Search Tool) (Altschul et al., 1997). Reads that were fully matched to human genome were further removed as “pure” human sequences. Full match to human genome was defined based on two criteria: 1) the matching identity of the local alignment between sequence read and human genome was above 80%; 2) the 3’ and 5’ ends of the sequence read that were not covered by the local alignment must be shorter than 12bp. The remaining reads were further searched against human HBV genome strains sequences downloaded from NCBI Genome database, and those with no hit to HBV genome were further removed as insignificant reads. Reads that remained after the above filtering process were used for downstream identification of HBV-HG junctions.

To recover possible HBV-HG junction sites that might be disrupted during DNA shearing process as well as to reconstruct the original DNA sequences from fragmented sequencing reads, the remaining reads were assembled into longer sequences known as “contigs” for each patient sample using *de novo* assembly algorithm DNASTAR SeqMan NGen and LaserGene for visualization of assembly. For each patient, the assembled contigs and unassembled reads left were then searched against HBV genome, and those either fully matched to HBV genome or having HBV hit longer than 52bp were selected. The reason 52bp was used as a threshold is that the longest identical region between human genome and most HBV strains genome is 52bp. Thus sequences that had at least 52 consecutive bases aligned to HBV genome would carry more confidence that the aligned part of the sequence were truly derived from HBV genome rather than human genome.

Subsequently, the selected contigs and unassembled reads either fully aligned to HBV genome or with at least 52 consecutive bases aligned to HBV genome, were searched against both HBV and human genomes for sequence identities. In this study, sequence identities were classified into five groups: one group is named "intact HBV" referring to sequences that are fully matched to HBV genome; second group is "modified HBV" including sequences with one region aligned to HBV genome and the other region aligned to a different region of HBV genome; the third group is "HBV+Unknown" containing sequences with one region aligned to HBV genome but the other region comprising only a few bps that is too short to be accurately mapped to either HBV or human genome ("unknown"); the fourth group is "HBV-HG junction" representing sequences with one region aligned to HBV genome but the other region mapped to human genome; and the last group is "modified HBV-HG junction" referring to sequences with one region identified as "modified HBV" and the other region aligned to human genome. Contigs and unassembled reads grouped into "HBV-HG junction" and "modified HBV-HG junction" would contain the predicted junction points where HBV sequences insert into human genome. The HBV-HG and modified HBV-HG junctions were then further classified into two types for ease of post-identification analysis: Type I has the 5’ end of the HBV genome sequence deleted at the integration site, and Type II has the 3’ end of the HBV genome deleted at the integration site (Fig 2.3). This classification will facilitate the identification of genes or functional domains of HBV genome that are deleted or conserved after the integration events.

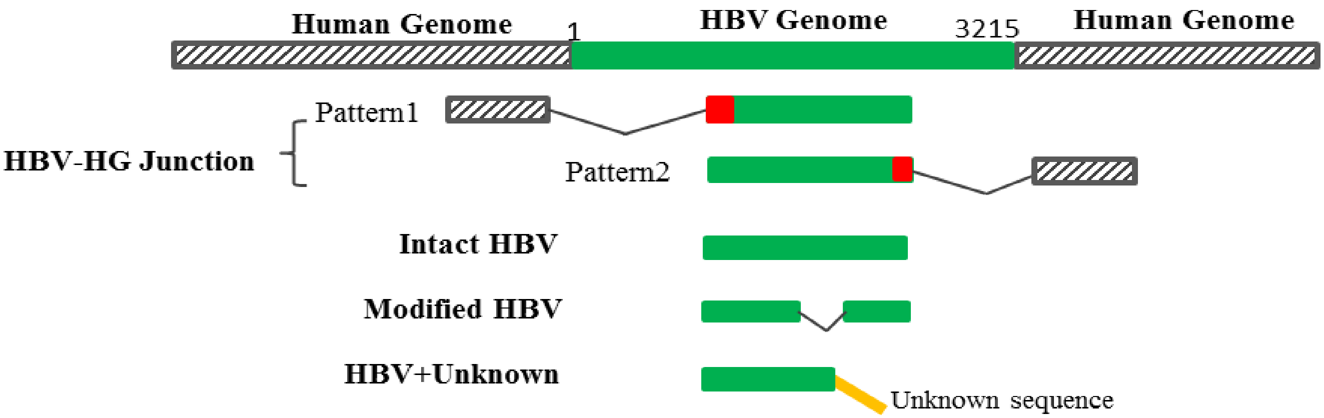
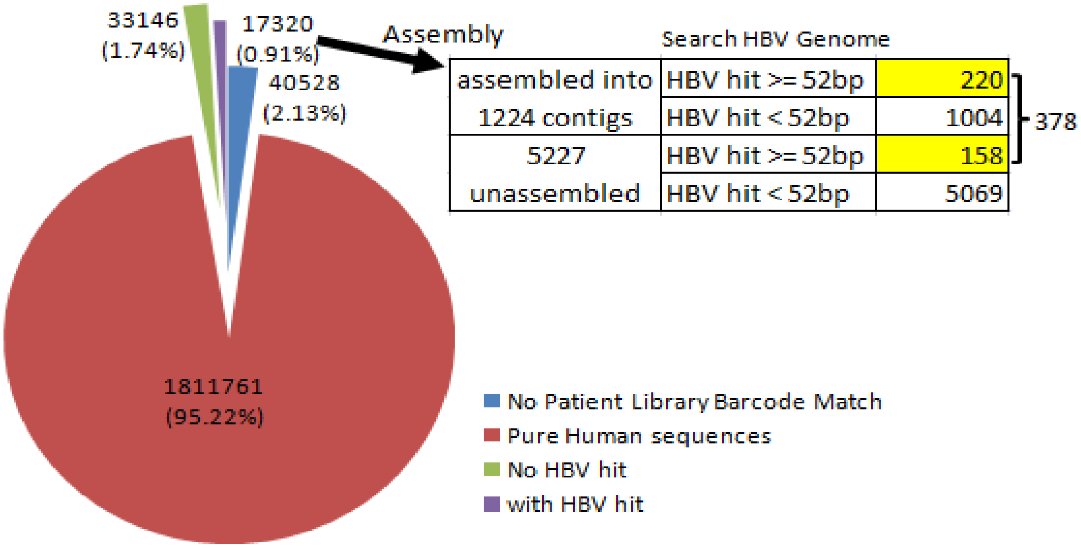


Figure 2.3: Typical patterns of HBV-containing sequence identities: intact HBV that is fully aligned to HBV genome, modified HBV with one region of the sequence aligned to HBV genome and the other region aligned to human genome, HBV+Unknown with one region aligned to HBV genome and the other region not known to both HBV and human genomes, and HBV-HG junctions with one region aligned to HBV genome and the other region aligned to human genome. In this study, the HBV-HG junctions were further grouped into two types: Type I with the 5’ end of HBV genome deleted (pattern1) and Type II with 3’ end of HBV genome deleted (pattern2) in the integrations. The red color highlighted the junction points on HBV genome.

2.2 Results

2.2.1 Sequence Identities of the FLX Sequencing Reads

A total of 1,902,755 raw sequence reads were obtained from 454 FLX pyrosequencing of our 48 pairs of DNA samples extracted from tumor and adjacent non-tumor tissues of HBV-positive HCC patients. Each sequence read had a 6bp library barcode at the 5' to facilitate the unique identification of the specific patient DNA sample. The raw sequence reads were then analysed following the pipeline shown in Fig 2.2 and a summary of the sequence identities was shown in the Fig 2.4.



assembled into 1224 contigs	HBV hit >= 52bp	220
	HBV hit < 52bp	1004
5227 unassembled	HBV hit >= 52bp	158
	HBV hit < 52bp	5069

Figure 2.4: Summary of sequence identities for all 1,902,755 raw FLX sequencing reads. About 2.13% of the raw sequence reads were not assigned to any patient sample because of no match of unique 6bp library barcodes. A large portion of the raw reads, about 95.22%, were fully matched to human genome and another 1.74% had no hit to HBV genome. These sequence reads were removed from assembly analysis. The rest 0.91% reads with at least one hit to HBV genome were then uploaded to SeqMan NGen *de novo* assembler for assembly into longer sequences which we called "contigs". As a result, 1224 contigs were formed and 5227 raw reads remain unassembled. These contigs and unassembled reads were then searched against HBV genome, and 220 contigs plus 158 unassembled reads were found either fully matched to HBV genome or with at least 52 consecutive bases aligned to HBV genome. These 378 sequences (220+158) were considered confidently containing HBV sequences and would be used for downstream analysis to identify HBV-host integrations.

About 2.13% of the raw sequences did not match to the unique 6bp barcodes for the patient samples and were thus removed at first place. The remaining reads were then assigned to the 48 patients each with paired tumor and adjacent non-tumor tissue samples based on the barcode identification. The assigned raw reads had average length of 254bp and ranged from 34bp to 1121bp. These reads were then searched against human and HBV genomes using BLAST, and a large portion of the raw reads (95.22%) were found to be fully aligned to human genome and another small proportion of 1.74% did not align to the HBV genome. “Fully matched" to human genome was defined as: there are at least 80% identity between the query sequence and the reference genome sequence and not more than 12bp of the 5’ and 3’ ends of the sequence does not align to the reference sequence. Sequences that were either purely human or did not align to HBV genome are probably results of non-specific enrichment, and were then removed from downstream assembly analysis. The remaining 0.91% of the reads that had hits to HBV genome were assembled into longer sequences ("contig") using SeqMan NGen *de novo* assembler.

The criteria for the *de novo* assembly are as follows: for two sequence reads to assemble into one longer sequence, the overlapping fragment between the two reads must be longer than 19bp and the overlapping region should be above 85% identical. In addition, the assembler algorithm also calculates the probability that an observed overlapping fragment is also observed amongst other input sequences. Hence, the longer the overlapping fragment, the less likely that fragment is observed in other sequences, thus the more confidence one can have that the overlapping fragments are not merely due to a chance event. The assembly was done for every single patient sample, and the assembled contigs and unassembled reads were then searched against the HBV genome. As a result, a total of 378 sequences (220 contigs and 158 unassembled reads) were found to either align completely with the HBV genome or have greater than 52 consecutive bases identical with the HBV genome. 52bp was chosen as a threshold is because that the longest identical region between the HBV and human genome sequences is 52bp. Thus a sequence, which had greater than 52 consecutive bases locally aligned to HBV genome, was most likely derived from the HBV genome rather than human genome.

Hence, a total of 378 sequences (220+158) were found highly likely to be derived from the HBV genome, with some aligning to both human and HBV genome sequences where potential HBV-host integration sites can be identified. These 378 HBV-containing sequences were then used for downstream analysis including the identification of HBV-host integration sites.

2.2.2 Sequence Capture Coverage of HBV Genome from FLX data

To determine if the entire HBV genome can be enriched by the specifically designed HBV probes in our laboratory and also to determine if there are preferred enrichment on specific regions of the HBV genome, among the 378 identified HBV-containing sequences, the number of sequences that aligned to each position of the circular HBV genome (1 to 3215bp) was counted and then plotted as shown in Fig 2.5. As evident in Fig 2.5, the designed HBV probes were capable of capturing the entire HBV genome, and in particular, the HBx gene (position: 1374-1838) was relatively more abundant than other regions of the HBV genome in these patients. Interestingly, the 3' end (near position 1838) of the HBx gene had relatively lower abundance than the 5' end (near position 1374).

This difference in the abundances of the 5’ and 3’ ends of HBx gene in HCC patient liver tissues may indicate the differences of their functions.

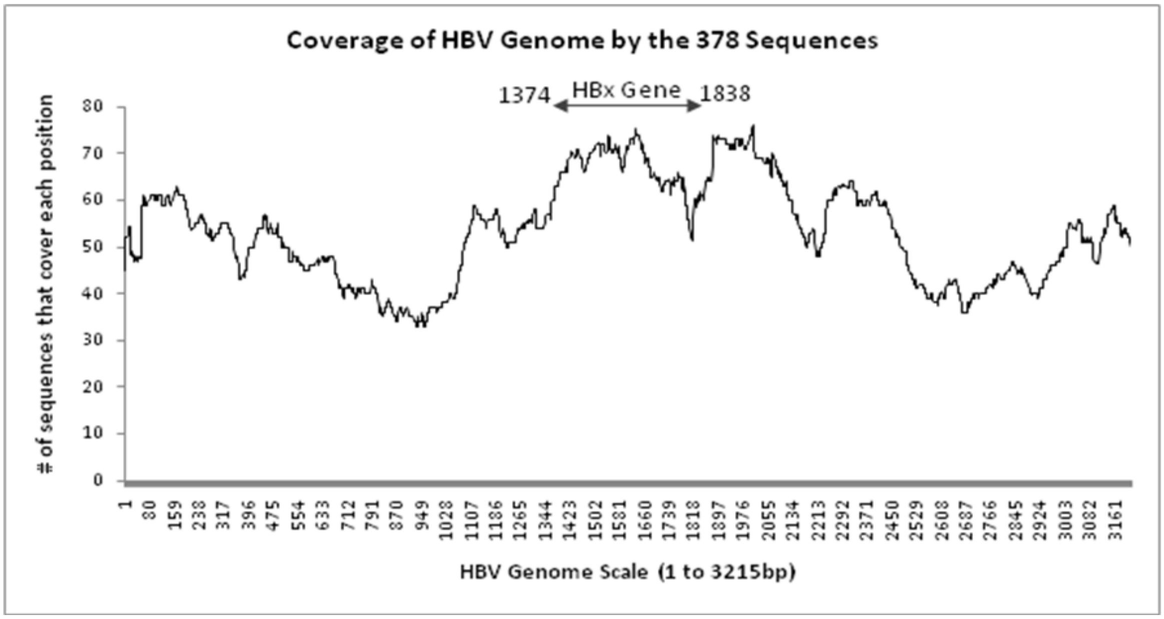
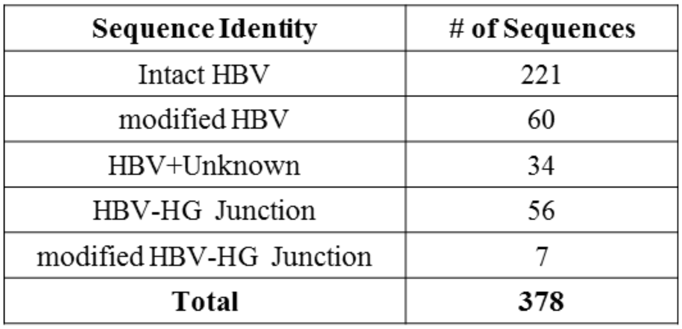


Figure 2.5: Coverage of the HBV genome (3215bp) by the 378 HBV-containing sequences including 220 assembled contigs and 158 unassembled reads in patients. For each position of the HBV genome of length 3215bp, the number of sequences that covered the position was counted and then plotted as a distribution over the entire HBV genome. It turned out that the entire HBV genome could be captured by the HBV probes designed in our laboratory, and the HBx gene (1374-1838) may exist more abundantly relative to other parts of the HBV genome in patients. And the 3’ end of HBx gene (near position 1838) may have lower abundance than the 5’ end (near position 1374) in patient liver tissues.

2.2.3 Identification of Modified HBV and HBV-Human Genome Junctions

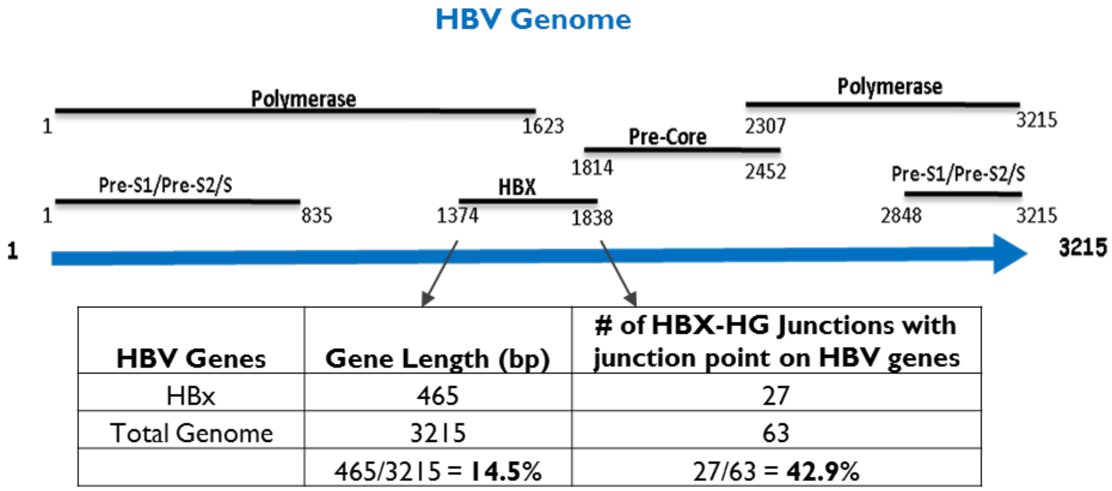
These 378 HBV-containing sequences were searched against HBV and human genomes, and then classified them into the five groups, as shown in Table 2.1 and Supplementary Table S1. Of the 378 sequences, 221 were fully matched to HBV genome ("intact HBV"); 60 were chimeric where one region of the sequences aligned to HBV genome and the other region aligned to a different region of HBV genome ("modified HBV"); 34 had one region aligned to HBV genome but the other region containing only a few bases that makes it difficult to map uniquely to human or HBV reference genomes ("HBV+Unknown"); 56 had one region aligned to the HBV genome and the other region aligned to human genome ("HBV-HG junction"); and 7 had one region identified as "modified HBV" and the other region aligning to the human genome sequences ("modified HBV-HG junction"). The 221 "intact HBV" sequences were not long enough for us to determine whether they come from free HBV species or integrated HBV with human genome in patients. The 60 "modified HBV" sequences had various alterations and were probably consequences of the complex events (deletions, insertions, duplications, inversions and rearrangements) that occurred after the HBV genome is integrated into human genome. These 378 HBV-containing contigs and unassembled reads were distributed amongst 42/48 (~87.5%) patients. Among these 42 patients with HBV-containing sequences, ~35.7% (15/42) had various alterations/modifications in their HBV sequences including insertion, deletions, duplications and inversions (“modified HBV” and “modified HBV-HG junctions”), while ~52.4% (22/42) carried both HBV and human sequences within the same sequence from which integration sites can be inferred (“HBV-HG junction” and “modified HBV-HG junctions”). Our laboratory has experimentally successfully validated a subset of "intact HBV", "modified HBV" and “HBV-HG junction” sequences. However, there seems to be no conserved patterns of the alterations/modifications of the HBV sequences (“modified HBV”) among patients, during integration events. This may suggest that the alterations of HBV genome sequences after insertion into human genome were either very complex or relative random as implicated by previous studies.

Table 2.1: Summary of the identities of the 378 HBV-containing sequences. The sequence identities were classified into five categories: intact HBV, modified/altered HBV, HBV+unknown, HBV-Human junction, and modified HBV-human junction.



Sequence Identity	**# of Sequences**
Intact HBV	221
modified HBV	60
HBV+Unknown	34
HBV-HG Junction	56
modified HBV-HG Junction	7
Total	**378**

Nevertheless, some patterns were apparent when “HBV-HG junction” sequences where the HBV and human genomic sequences integrated, were examined. The 56 “HBV-HG junctions” and 7 “modified HBV-HG junctions” comprised junction points that fused HBV sequence to the human sequence. The detailed information of these 63 HBV-Human junctions was listed in Supplementary Table S2, which illustrated the integration positions of HBV and human genomes, junction points, and the HBV and human genes where the junction points resided. Twenty seven of the 63 junctions (~42.9%) were predicted to have the junction points on HBx gene. Although the HBx gene is only 465bp of the HBV genome of 3215bp (~14.5%) as seen in Fig 2.6, the junction sites were almost three times more enriched in the vicinity of HBx gene (42.9% ~= 3*14.5%) with significant Chi-Square two-sided *p*-value of 0.0008, compared to other genes of HBV genome. Thus, we may conclude that HBV genome integrations may preferentially occur at the HBx gene, indicating potential functions of HBx gene in HBV-induced hepatocarcinogenesis. Our laboratory has successfully experimentally validated a selected subset of 23 HBV-HG junctions, of which 20 had the junction points on HBx gene, as indicated in the 8th column of Supplementary Table S2. This indicated that our analysis pipeline (Fig 2.2) is robust for the FLX sequencing data in identifying novel HBV-human integration sites, and at least a subset of 23 novel HBV-human integration sites have been experimentally validated to exist in HCC patient liver tumor or adjacent non-tumor tissues.



HBV Genes	Gene Length (bp)	# of HBX-HG Junctions with junction point on HBV genes
HBx	465	27
Total Genome	3215	63
	465/3215 = **14.5%**	27/63 = **42.9%**

Figure 2.6: Enrichment of HBV-HG junctions with integration sites on HBx gene. The circular HBV genome consists of four major coding genes: polymerase (2307-3215 & 1-1623), pre-core (1814-2452), surface protein (2848-3215 & 1-835), and HBx (1374-1838). Out of the 63 HBV-human junctions predicted, 27 (~42.9%) were predicted to occur on HBx gene. HBx gene is a small open reading frame of length 465bp which accounts only ~14.5% of the entire HBV genome. We could see that the HBV genome integrations were almost three times enriched in HBx gene (42.9% ~= 3*14.5%), with significant Chi-Square two-sided *p*-value of 0.0008, compared to other genes of HBV genome. Therefore it could be concluded that HBV genome integrations may preferentially occur on HBx gene, indicating potential functions of HBx for HBV-associated HCC.

Integration of HBV genome into human genome is implicated to have two basic functions: disrupting the host gene expressions or expressing chimeric transcripts and proteins. Within human genome, most of the junction points were located at the non-coding regions, such as, promoter, introns and intergenic regions (Table 2.2), suggesting that the HBV insertions may potentially disrupt the regulatory elements on the human genome of host gene expressions. To check whether these junctions potentially change host gene expressions, the expression values of the nearest genes for each junction site were examined in patient tumor and non-tumor tissues from patient cDNA expression microarray profiles. Among the 63 junctions, there were 13 nearest genes that were differentially expressed between tumor and non-tumor tissues by at least 2 fold (Supplementary Table S2). Nine of the 13 genes had the HBV junction points on introns, while three genes had junction points at downstream of gene regions and one at promoter region. The true cellular events leading to the final change of these 13 host gene expressions remain unknown, but it might potentially be due to the viral genome integration. Nevertheless, most of the 63 viral integrations occur on non-coding regions of host genome, and by computationally scanning for opening reading frames in the integrated genome, we hypothesized that these viral integrations may result in early termination of viral genes (ie, expression of modified viral proteins) or expression of viral-host chimeric transcripts that might be potentially oncogenic. Our laboratory has experimentally validated the existence of viral-host chimeric transcripts and functional evaluation of these chimeric transcripts is still in progress in the laboratory.

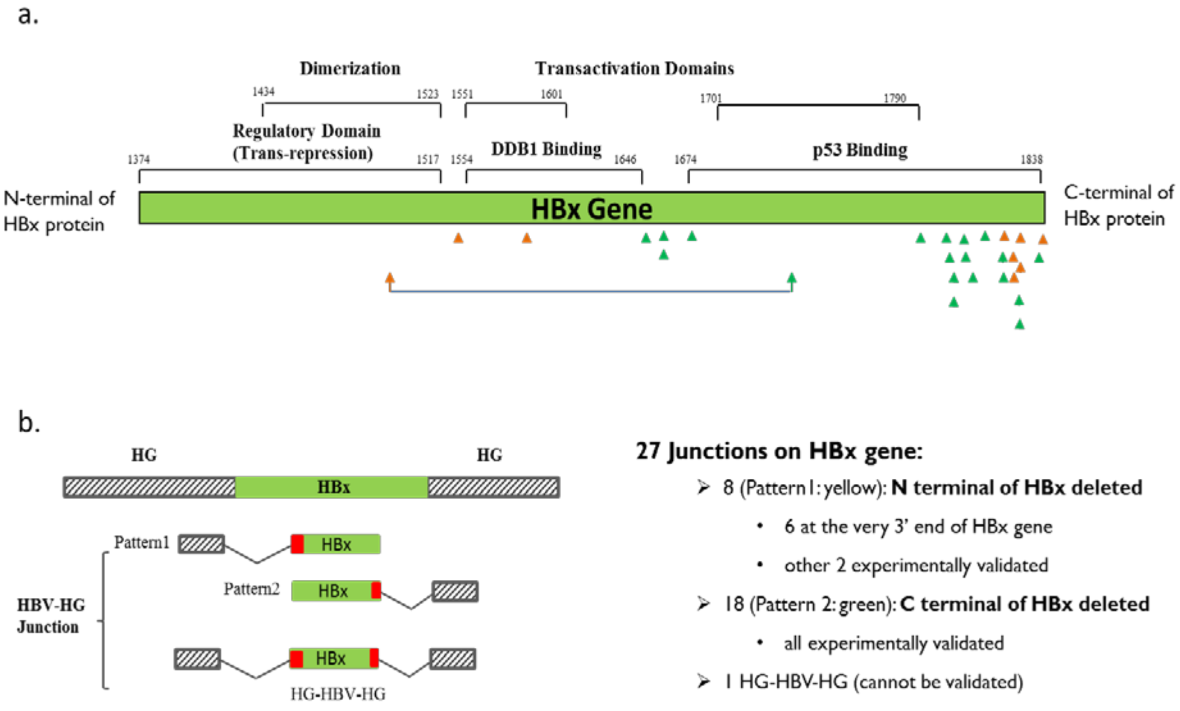
Table 2.2: Summary of the locations of the 63 junction points on human genome.

Location of HBV-HG Junctions	# of Junctions
Promoter	4
Intron	21
Exon	1
Hypothetical pseudo-genes	5
Intergenic	30
Not Annotated	1
Non-coding RNA	1
Total	63

2.2.4 Analysis of HBV-Host Junctions with Junction Points on HBx gene

Since HBx gene has been implicated to play a major role in the development of HCC in chronic HBV-infected patients, the 27 HBx-HG chimeric junctions were further examined. HBx gene of length 465bp codes for HBx protein of 154 amino acids which is a multifunctional protein with trans-repression regulatory domain, dimerization domain, and DDB1 binding domain at N-terminal, and transactivation domain and p53 binding domain at C-terminal (Fig 2.7). To see which functional domains of HBx were affected by the integrations, the locations of the 27 HBx-HG junction points were plotted on HBx gene as shown in Fig 2.7. Since the coding domains of HBx protein were from the positive strand of HBx gene, the HBx-HG junctions were grouped into two junction patterns (Fig 2.3 & Fig 2.7): pattern 1 with the 5' end of HBx gene (N-terminal of HBx protein) deleted at the junction point; and pattern 2 with the 3' end of HBx gene (C-terminal of HBx protein) deleted at the junction point. Of the 27 junctions with fusion point on HBx gene, eight had the N-terminal of HBx deleted in the integrations (junction pattern 1), 18 had the C-terminal of HBx deleted in the integrations (junction pattern 2), and 1 had a special case of HG-HBV-HG junction pattern which had the first and last part of the sequence aligned to human genome while the middle part aligned to HBV genome. Six out of the eight junctions with N-terminal of HBx deleted located at the very 3' end of HBx gene (e.g. position 1820), with only a few base pairs remained in the junctions, indicating that almost the entire HBx gene was deleted in the integrations. In fact, it could be regarded that the HBx gene was not involved in these six junctions. Nevertheless, the other two pattern 1 junctions with the trans-repression domain removed were experimentally validated by our laboratory to exist in patients. Furthermore, the 18 junctions with C-terminal of HBx deleted mostly located within the p53 binding domain, suggesting that the p53 binding domain of HBx was deleted or partially deleted in the integrations. These 18 junctions have all been experimentally validated in our laboratory. The single HG-HBV-HG junction cannot be experimentally validated and is probably due to assembly errors. In summary, the two junctions with N-terminal trans-repression domain deleted and the 18 junctions with C-terminal p53 binding domain of HBx deleted in the integrations were all experimentally validated and may lead to expression of potential chimeric transcripts fusing HBx gene and human sequences.

Interestingly, the 18 junctions with C-terminal of HBx deleted all had the p53 binding domain deleted or partially deleted in the integrations. This may implicate that the N-terminal two thirds of HBx (amino acids 1 to 100 or nucleotide 1374 to 1673) preferentially remain intact in the integrations. N-terminal two thirds of HBx comprise of trans-repression domain, dimerization domain, DDB1 binding domain and transactivation domains, which may be important for HBx functions in patients. On the other hand, deletion or partial deletion of p53 binding domain in HBx may potentially abolish its interaction with p53. P53 is a well-known tumor suppressor gene regulating apoptosis and thus elimination of HBx interaction with p53 may promote hepatocarcinogenesis and development of HCC. In addition, this observation of more frequent deletion of C-terminal than N-terminal of HBx in integrations is consistent with our earlier result of sequence coverage over whole HBV genome, shown in Fig 2.5, where the 3' end of HBx gene had relatively lower abundance than the 5' end of HBx gene.



27 Junctions on HBx gene:

- 8 (Pattern1: yellow): **N terminal of HBx deleted**
  - 6 at the very 3' end of HBx gene
  - other 2 experimentally validated
- 18 (Pattern 2: green): **C terminal of HBx deleted**
  - all experimentally validated
- 1 HG-HBV-HG (cannot be validated)

Figure 2.7: Location plot of the 27 predicted HBV-HG junctions where the junction points fall on HBx gene (Supplementary Table S2). a) The reported functional domains of HBx protein included trans-repression regulatory domain, dimerization domain, DDB1 binding domain, transactivation domain, and p53 binding domain. Of the 27 junctions, eight had the N-terminal of HBx deleted (yellow arrow), 18 had the C-terminal of HBx deleted (green arrow) and one had a special case of HG-HBV-HG junction pattern, in which the first and last part of the sequence aligned to human genome while the middle part aligned to HBV genome. Six out of the eight junctions with N-terminal of HBx deleted located at the very 3' end of HBx gene with only a few base pairs of HBx gene remained in the junction, and thus could be regarded not involving HBx gene. The other two junctions with N-terminal of HBx deleted had the trans-repression domain deleted, and were experimentally validated by our laboratory colleagues. The 18 junctions with C-terminal of HBx deleted were all experimentally validated and mostly had the p53 binding domain deleted or partially deleted in the integrations. The one HG-HBV-HG junction cannot be experimentally validated and could possibly be due to assembly errors. In total, the two junctions with N-terminal trans-repression domain deleted and the 18 junctions with C-terminal p53 binding domain of HBx deleted may potentially lead to expression of chimeric transcripts fusing HBx and human sequences. b) Part B illustrates the HBV-HG junction patterns in genome integrations: pattern 1 with N-terminal of HBx deleted; pattern 2 with C-terminal of HBx deleted; and HG-HBV-HG junction with both N-terminal and C-terminal of HBx deleted. Red color highlights the integration sites of HBV genome with human genome.

2.3 Discussion and Future Work

Currently most of the approaches used to study the integration sites of HBV in the host genome were all PCR-based methods which required prior information on the integrated HBV sequences, or assumption of integration of certain parts of HBV. To comprehensively characterize HBV genome integration boundaries in HCC patients, our laboratory developed an unbiased HBV enrichment strategy followed by next generation sequencing (454 life science FLX sequencer) to capture HBV related DNA fragments from the complex genomic DNA samples extracted from the tumor and adjacent non-tumor tissues of 48 HBV-positive HCC patients. In this study, I implemented a pipeline to analyse the ultra high-throughput sequencing data, and was able to identify various novel modified/altered HBV sequences as well as novel HBV-host junctions, without much prior knowledge or assumption of which part of HBV genome is integrated into host genome. A total of 378 sequences including assembled contigs and unassembled reads were found to contain HBV sequences, out of which, 60 were altered HBV sequences (e.g. insertion, deletion, duplication and inversion) and 63 comprised of HBV-HG junctions. These 378 HBV-containing sequences were distributed amongst 42/48 (87.5%) patients. Of the 42 patients, 35.7% (15/42) had various alterations in their HBV sequences, including insertions, deletions, duplications and inversions, while 52.4% (22/42) carried both HBV and human sequences within the same sequence from which integration sites can be inferred (HBV-HG junctions).

Particularly, our laboratory has successfully validated a batch of altered HBV sequences and HBV-HG junctions. Presence of the altered HBV sequences in patients confirmed that after HBV genome inserted into host genome, complex manipulations events may have occurred, such as insertions, deletions and duplications, inversions and rearrangements. However, there seemed to be no conserved patterns of these alteration events across the 48 patients, which implicate that the alteration process might be either very complex or relatively random. Nevertheless, we do observe that the junction/fusion points of HBV genome with human genome were enriched on HBx gene (27/63 junctions), and more interestingly the C-terminal of HBx was often deleted at the integration sites. HBx protein of length 154 amino acids has been implicated to play a major role in HBV-induced hepatocarcinogenesis, and our observation suggested that integration of C-terminal deleted HBx with human genome may be potentially functionally important in hepatocarcinogenesis. P53 has been reported to interact with HBx at C-terminal amino acids 100 to 154. Deletion or partial deletion of C-terminal p53 binding domain of HBx may potentially abolish its interaction with p53. P53 is a well-known tumor suppressor gene regulating apoptosis and thus abortion of HBx interaction with p53 may promote hepatocarcinogenesis and development of HCC. More importantly, the N-terminal two thirds of HBx (amino acids 1 to 100) were observed to be preferentially retained in the integrated form. The N-terminal two thirds of HBx comprise trans-repression domain, dimerization domain, DDB1 binding domain and transactivation domains, and may be important for HBx functions in HBV-induced hepatocarcinogenesis.

Integration of HBV into human genome is implicated to either disrupt host gene expressions or express chimeric transcripts and proteins to functionally participate in hepatocarcinogenesis. We have observed that most of the 63 HBV-HG junctions occurred at the non-coding regions of human genome (Table 2.2), therefore, the integration of HBV genome may interrupt the regulatory elements of host genes, such as promoters, introns, and intergenic regions. Examining the nearest genes for the 63 junctions, we found 13 were differentially expressed in tumor and adjacent non-tumor tissues of HCC patients with at least 2 fold changes. Though it might be due to the viral genome integration, the true cellular events leading to the change of these 13 host gene expressions are still unknown. Nevertheless, by computationally scanning for opening reading frames, we hypothesized that in addition to interrupting host gene expressions, these viral integrations may also result in early termination of viral gene expression (i.e. expression of modified viral proteins) or expression of viral-host chimeric transcripts that may be potentially oncogenic. Our laboratory has experimentally observed the existence of novel viral-host chimeric transcripts. The functional evaluation of these chimeric transcripts/proteins is still in progress in the laboratory.

Identification of novel HBV-human genome integration sites, modified viral protein expression and potential viral-host chimeric transcripts could facilitate the understanding of the underlying mechanisms of HBV genome integrations into human genome, and may give us more knowledge on how HBV infection gradually leads to hepatocarcinogenesis. The major benefit of utilizing hybrid capture method coupled with single-end high-throughput 454 FLX pyrosequencing is that the reads could be long enough to permit the identification of precise virus insertion sites in human genome at single-base resolution with one region of the read aligned to human genome and the other region aligned to virus genome. Last but not least, this has been among the very first to comprehensively characterize HBV integrations utilizing high throughput sequencing techniques in a large series of samples from 48 HCC patients on a genome-wide basis without much prior knowledge of the integration sites. By the identification and characterization of these genome integration sites, we are now better positioned to understand the underlying mechanism of how HBV genome integrations may contribute to HBV-induced hepatocarcinogenesis in HBV-associated HCC patients.

CHAPTER 3: Computational Identification of Putative Direct Gene Targets of HBx

3.1 Materials and Methods

HBx protein has been implicated to play an important role in HBV-induced hepatocarcinogenesis, and one of the reported underlying mechanisms is that HBx binds to DNA indirectly through interactions with transcription factors and deregulate host gene expressions by changing transcription factor binding affinities to DNA. In order to profile the genome binding sites of HBx-transcription factor complex on a genome-wide basis with single-base resolution, our laboratory has utilized antibodies specifically against HBx protein to immunoprecipitate DNA fragments potentially bound by HBx-transcription factor complex in primary immortalized liver cell line (THLE3) transfected with HBx-expressing adenoviruses, and then applied high-throughput Illumina sequencer to sequence the immunoprecipitated DNA fragments. This ChIP-Seq technology produced millions of sequence reads of uniform length 36 bp for HBx-expressing THLE3 cells (AdHBx) and control THLE3 cells (AdEasy). In this project, I implemented an analysis pipeline to integrate the THLE3 ChIP-Seq sequencing data, microarray expression profiles for THLE3 cells and 100 HCC patients, and 100 HCC patient clinical data to identify global genome binding sites of HBx and predict putative clinically associated direct gene targets of HBx, as shown in Fig 3.1.

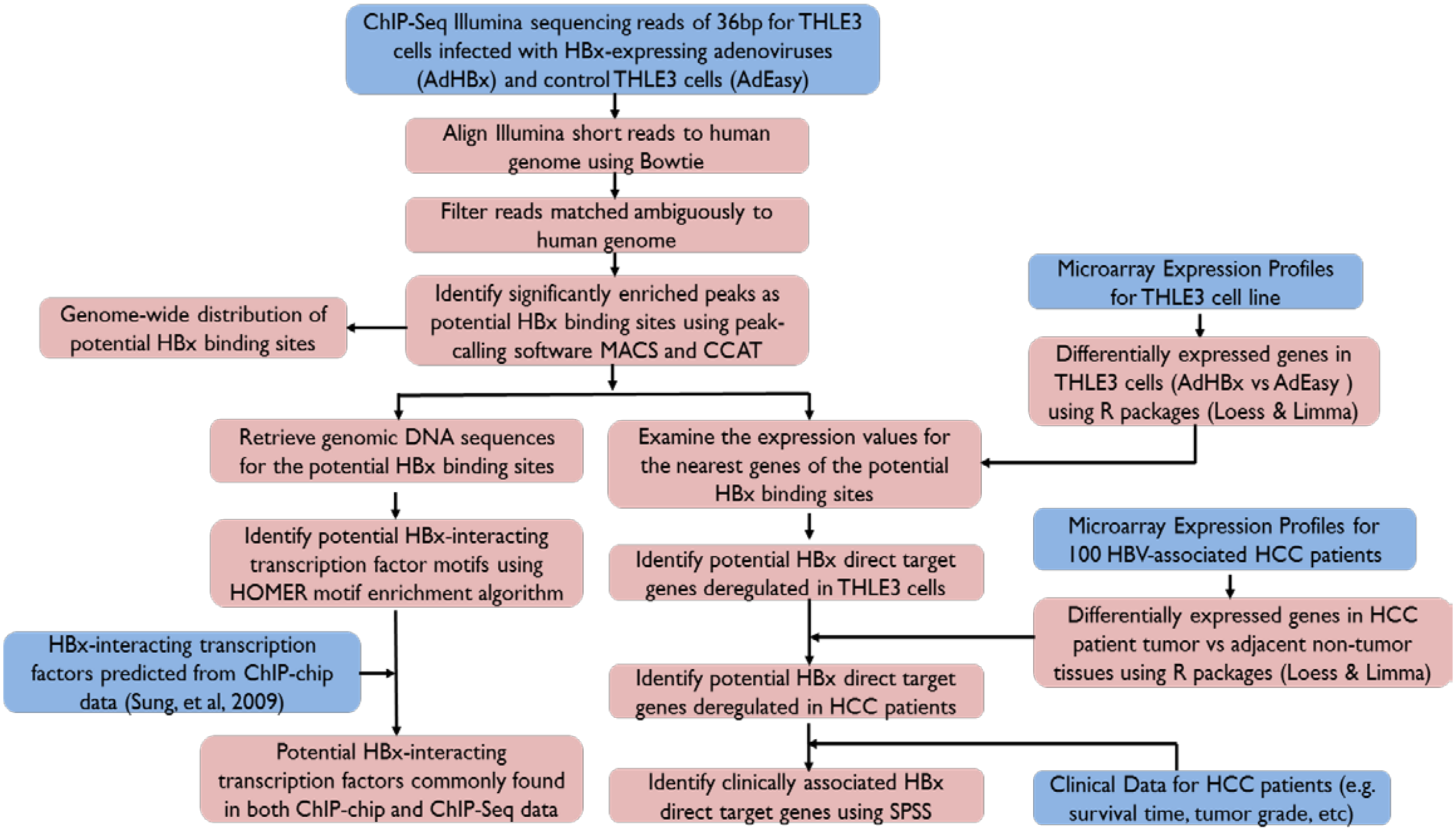

ChIP-Seq Illumina sequencing reads of 36bp for THLE3 cells infected with HBx-expressing adenoviruses (AdHBx) and control THLE3 cells (AdEasy)
Align Illumina short reads to human genome using Bowtie
Filter reads matched ambiguously to human genome
Identify significantly enriched peaks as potential HBx binding sites using peak-calling software MACS and CCAT
Genome-wide distribution of potential HBx binding sites
Microarray Expression Profiles for THLE3 cell line
Differentially expressed genes in THLE3 cells (AdHBx vs AdEasy) using R packages (Loess & Limma)
Retrieve genomic DNA sequences for the potential HBx binding sites
Examine the expression values for the nearest genes of the potential HBx binding sites
Identify potential HBx-interacting transcription factor motifs using HOMER motif enrichment algorithm
Identify potential HBx direct target genes deregulated in THLE3 cells
Microarray Expression Profiles for 100 HBV-associated HCC patients
Differentially expressed genes in HCC patient tumor vs adjacent non-tumor tissues using R packages (Loess & Limma)
HBx-interacting transcription factors predicted from ChIP-chip data (Sung, et al, 2009)
Identify potential HBx direct target genes deregulated in HCC patients
Potential HBx-interacting transcription factors commonly found in both ChIP-chip and ChIP-Seq data
Identify clinically associated HBx direct target genes using SPSS
Clinical Data for HCC patients (e.g. survival time, tumor grade, etc)

Figure 3.1: Workflow of computational analysis to identify genomic binding sites of HBx and putative clinically associated direct target genes of HBx. With the Illumina sequencing reads of 36bp generated from ChIP-Seq technique in THLE3 cells transfected with HBx-expression adenoviruses (AdHBx) and control THLE3 cells (AdEasy), I first aligned the short reads to human genome using Bowtie and removed those reads matched ambiguously to human genome. The remaining reads were then analyzed using peak-calling tools MACS and CCAT to identify significantly enriched peaks as potential HBx binding sites. Location of the predicted HBx binding sites relative to the genome structures can be plotted and examined. The genomic sequences of the predicted HBx binding sites were retrieved and scanned against TRANSFAC motif database using HOMER motif enrichment algorithm for potential HBx-interacting transcription factors, which were then compared with the previously predicted transcription factors from the ChIP-chip data (Sung, Lu et al. 2009). Additionally, the microarray expression profiles for the nearest genes of the potential HBx binding sites were analyzed using R packages “Loess” for normalization and “Limma” for differential expressions to identify potential HBx deregulated direct gene targets in THLE3 cells. Microarray expression profiles for 100 HBV-associated HCC patients were also analyzed using “Loess” and “Limma” and integrated to get a list of potential HBx direct gene targets that were also differentially expressed in HCC patient tumor and adjacent non-tumor tissues. Integration and analysis of the HCC patient clinical data using statistical analysis package SPSS further narrowed down a list of clinically significant potential direct gene targets of HBx.

I analyzed the ChIP-Seq Illumina sequencing data, microarray expression profiles, and HCC patient clinical data obtained in our laboratory (highlighted in blue in Fig 3.1). For the Illumina high-throughput sequencing reads of 36bp generated from the ChIP-Seq technique on THLE3 cells transfected with HBx-expressing adenoviruses (AdHBx) and control THLE3 cells (AdEasy), the short reads were first mapped into human genome using reference-guided aligner Bowtie (Langmead et al., 2009), and those reads that matched ambiguously to human genome were removed. Peak-calling software tool MACS (Model-based Analysis for ChIP-Seq) (Zhang et al., 2008) was then used to scan the remaining reads with unique best match to human genome. Significantly enriched peaks identified from MACS were then re-confirmed using another peak-calling tool CCAT (Control-based ChIP-Seq Analysis Tool) (Xu et al., 2010). Common enriched peaks predicted from both MACS and CCAT were identified as the potential HBx binding sites in human genome. Genomic locations of the potential HBx binding sites were then examined for possible patterns of HBx binding sites relative to genome structures (e.g. promoters, introns, exons, and intergenic regions). The genomic DNA sequences of the potential HBx binding sites were also retrieved and scanned against TRANSFAC transcription factor known motif database (Wingender et al., 1996) using HOMER motif enrichment algorithm (Hypergeometric Optimization of Motif Enrichment) (Heinz et al., 2010) to identify significantly over-represented motifs as the potential HBx-interacting transcription factors. Our laboratory has previously utilized ChIP-chip technique on UV-treated liver cell line HepG2 cells transfected with HBx-expressing adenoviruses, and published a list of predicted HBx-interacting transcription factors (Sung et al., 2009).

Therefore, the potential HBx-interacting transcription factors predicted from the ChIP-Seq data were also compared to those from ChIP-chip data for commonly predicted HBx-interacting transcription factors. Further, microarray expression profiles for THLE3 cells (AdHBx vs AdEasy) was analyzed using R packages "Loess" for microarray normalization and "Limma" for identification of differentially expressed genes in THLE3 cells. Expression values for the nearest genes of the potential HBx binding sites were then examined to identify potential HBx deregulated direct gene targets in THLE3 cells. Microarray expression profiles for the 100 HBV-associated HCC patients (tumor vs adjacent non-tumor tissues) were also analyzed and integrated to identify potential HBx direct gene targets that were differentially expressed in HCC patients. To check whether these potential HBx direct target genes identified from THLE3 cells are truly related to HCC, clinical data of 100 HCC patients including survival profile etc., were also integrated to identify significant clinically associated gene targets of HBx using statistical analytic package SPSS (Statistical Package for the Social Sciences) (Mather and Austin, 1983). The following sub-sections under Materials and Methods will elaborate in details of the computational analysis shown in Fig 3.2

3.1.1 Data Collection: ChIP-Seq Libraries, Expression Profiles & 100 HCC Patients Clinical Data

As shown in Fig 3.2, Chromatin immunoprecipitation (ChIP) was performed on DNA samples extracted from control immortalized primary normal liver cell line THLE3 cells (AdEasy) and HBx-expressing adenoviruses transfected THLE3 cells (AdHBx), using antibodies specifically against HBx protein to immunoprecipitate sheared DNA fragments that are bound indirectly by HBx. High throughput Illumina sequencing of the immunoprecipitated DNA fragments were then performed to construct AdEasy and AdHBx THLE3 ChIP-Seq libraries, which can be further analysed to detect the potential global genomic binding sites of HBx and predict HBx-interacting transcription factors. The sequence reads are of length 36bp.

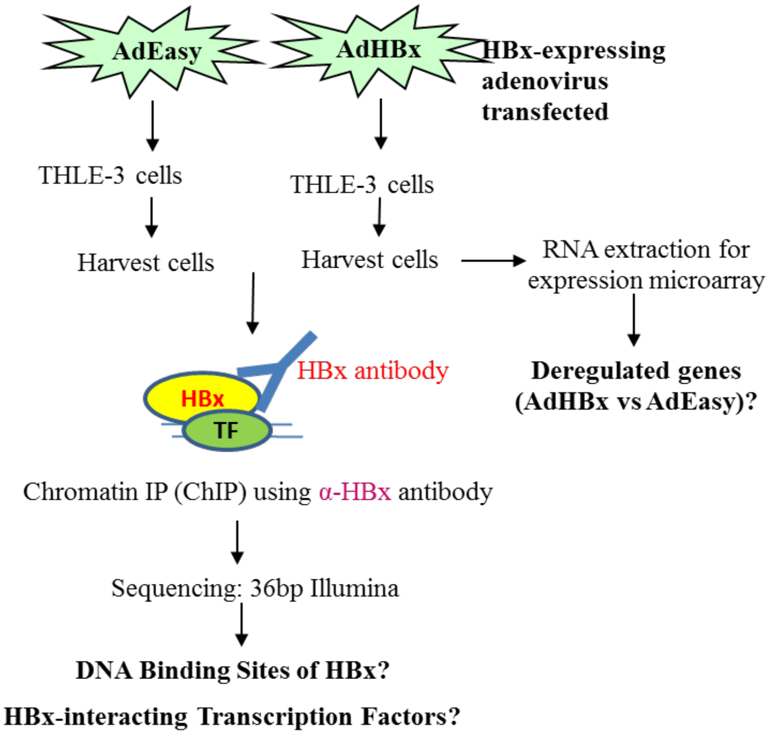

AdEasy
AdHBx
HBx-expressing adenovirus transfected
THLE-3 cells
THLE-3 cells
Harvest cells
Harvest cells
RNA extraction for expression microarray
Deregulated genes (AdHBx vs AdEasy)?
HBx antibody
HBx
TF
Chromatin IP (ChIP) using α-HBx antibody
Sequencing: 36bp Illumina
DNA Binding Sites of HBx?
HBx-interacting Transcription Factors?

Figure 3.2: Flowchart of experimental design for generation of ChIP-Seq data and gene expression profiles performed in the laboratory. DNA samples were first extracted from control immortalized primary normal liver cell line THLE3 cells (AdEasy) and HBx-expressing adenoviruses transfected THLE3 cells (AdHBx), then sonicated into small fragments, followed by chromatin immunoprecipitation using anti-HBx antibodies. After size selection, the immunoprecipitated DNA fragments were then sent for high throughput Illumina sequencing (36bp) to construct AdEasy and AdHBx THLE3 ChIP-Seq libraries. RNA samples were also extracted in our laboratory from the same sets of AdEasy and AdHBx THLE3 cells for Agilent two-color expression microarray in order to examine the change of gene expressions upon the presence of HBx protein in THLE3 cells.

To examine the change of gene expressions upon expression of HBx protein in THLE3 cells, RNA samples were also extracted in our laboratory from the same sets of AdEasy and AdHBx THLE3 cells for Agilent two-colour expression microarray, where control AdEasy sample was labelled as Cy3 (green) and AdHBx sample was labelled as Cy5 (red). Similarly, Agilent two-colour microarray of 100 HCC patients were performed by labelling the DNA sample extracted from the tumor tissue of a patient as Cy3 (green) and that from the adjacent non-tumor tissue in the same patient as Cy5 (red). Clinical data of the 100 HCC patients were also available for analysis. The patient clinical data include tumor grade (1, 2, 3 & 4), tumor encapsulation (Yes/No), tumor necrosis (Yes/No), vascular invasion (Yes/No), multifocality (Yes/No), local tumor extension (confined tumor: Yes/No), normal liver cirrhosis (Yes/No), normal liver steatosis (Yes/No), hepatic dysplasia (Yes/No), and survival time. Seventy five out of the 100 HCC patients have survival time data: 16 patients died from HCC and the other 59 patients were considered "censored cases" (48 alive at the time of recording, 6 dead but not due to HCC, and 5 lost of follow up). The HBx protein expression levels in both tumor and adjacent non-tumor tissues of the 100 HCC patients had also been determined in the laboratory.

3.1.2 Computational Identification of DNA Binding Sites of HBx

The short sequence reads of length 36bp from ChIP-Seq Illumina high-throughput sequencing platform for THLE3 AdEasy and AdHBx cells were first aligned to human genome (HG19) with no gaps permitted and a maximum of 2 mismatches allowed using an ultrafast and memory-efficient short read aligner Bowtie. Reads that aligned ambiguously to human genome were removed (i.e. reads matched to multiple regions of human genome with the same best alignment score and significance), and reads with unique best match to human genome were selected for downstream peak finding process. An existing ChIP-Seq peak calling algorithm MACS was applied taking AdEasy sample as negative control and AdHBx sample as ChIP. MACS algorithm will (1) first remove duplicate reads in the datasets that may arise from ChIP-DNA amplification and sequencing library preparation, (2) linearly scale the total number of reads in control data to be the same with that in ChIP data, (3) empirically model the size of the true protein binding site based on the bimodal enrichment pattern, (4) shift the genome locations of the reads in a strand-specific manner by half of the estimated size of the protein binding site, (5) scan the genome using sliding windows of user-defined width to identify candidate peaks with significant read enrichment in AdHBx sample based on *p*-values calculated from dynamic Poisson distribution of reads, (6) swap ChIP and control datasets and call peaks in control data, and (7) calculate FDR for each detected peak in ChIP data and rank them using *p*-values and FDR. In this study, MACS was applied using the threshold of at least 10 folds enrichment in AdHBx sample over control AdEasy sample and enrichment significance *p*-value less than 0.00005 for finding candidate enriched peaks. Another peak calling soft tool CCAT was used simultaneously to re-confirm the list of potential enriched peaks. CCAT adopts similar algorithm with MACS but is superior over MACS by estimating noise rate and resampling datasets to balance ChIP and control sample size instead of linear scaling as in MACS. Only enriched peaks predicted from MACS with their peak summits covered also by regions predicted from CCAT were considered as the potential DNA binding sites of HBx.

3.1.3 Annotation of Genome-wide Potential HBx Binding Sites

The potential HBx binding sites identified from ChIP-Seq data were mapped to the in-house human reference genome annotation database (HG19) from HOMER package (Hypergeometric Optimization of Motif Enrichment) developed in the Glass lab of UCSD (Heinz et al., 2010). This annotation database includes detailed information on human gene promoters, transcription start sites (TSS), introns, exons, gene 5’ and 3’ un-translated regions (UTRs). Promoter region was defined as 5kb upstream to the TSS of reference genes, and intergenic regions were defined as genomic regions other than promoters and gene body regions. Peaks located within multiple annotation categories (e.g. peaks fall in promoter of one gene but exon of another gene) were classified based on the precedence order that promoter comes first followed by 5' UTR and 3' UTR which then precede introns and exons. Mapping of peaks to annotation database was done using Microsoft SQL server 2005 software which stores relational databases and provides comprehensive functions for users to search and manipulate the cross-linked data tables.

3.1.4 Motif Enrichment Analysis for Potential HBx-interacting Transcription Factors

Genomic DNA sequences of the potential HBx binding sites were extracted from human genome (HG19) for transcription factor motif enrichment analysis. HOMER known motif enrichment algorithm scripts (Heinz et al., 2010) were downloaded (http://biowhat.ucsd.edu/homer/) and applied using transcription factor position weight matrices from TRANSFAC motif database version 11.3 (Wingender et al., 1996) which covered 601 vertebrate motif matrices for 389 transcription factor families. As mentioned earlier, HOMER motif enrichment algorithm comprises of the following steps: (i) first, the program randomly selects a set of background genomic sequences with similar length and GC contents to the potential binding site sequences named as "target sequences", (ii) it then assigns weights to each background sequence to minimize the imbalance in sequence contents with the target sequences, (iii) it further calculates the occurrence of each known motif in the target and background sequences and (iv) then computes a significance value (e.g. *p*-value) for the enrichment of the motif in the target sequences over background sequences (Heinz et al., 2010). Transcription factor motifs were ranked according to their enrichment *p*-values reported by HOMER and those with *p*-value below 0.05 were considered as significantly enriched/over-represented within the potential HBx binding sites. These significantly over-represented ones were known as transcription factors that may potentially interact with HBx to form a complex and bind to DNA.

The list of significantly enriched HBx-interacting transcription factors predicted in this study was also compared to the list HBx-interacting transcription factors discovered previously in our laboratory from ChIP-chip method (Sung et al., 2009). This is to see whether there is any HBx-interacting transcription factor that is commonly predicted in different liver cell lines under different experimental conditions (UV-treated HepG2 vs THLE3) using different experimental methods (ChIP-chip vs ChIP-Seq).

3.1.5 Analysis of THLE3 Microarray Expression Profiles to Predict Deregulated Direct Gene Targets of HBx

The Agilent two-colour microarray data with four biological replicates for THLE3 AdHBx (labelled as Cy5) and AdEasy (labelled as Cy3) cells were analysed using R packages "Loess" for array normalization and "Limma" for detection of differentially expressed genes in AdHBx over AdEasy cells. Genes with at least 1.5-fold expression change in AdHBx relative to AdEasy cells were selected as differentially expressed genes. These differentially expressed genes were hypothesized to be deregulated (either up- or down-regulated) either directly by HBx or due to downstream regulation effect of HBx. Expression values for the corresponding nearest genes of the potential HBx binding sites were then examined, and those differentially expressed with at least 1.5-fold change were selected as candidate deregulated direct gene targets of HBx.

3.1.6 Gene Ontology Analysis for Deregulated Gene Targets of HBx

The list of candidate HBx deregulated direct gene targets were uploaded to a web-based gene ontology analysis application DAVID (Database for Annotation, Visualization and Integrated Discovery) (http://david.abcc.ncifcrf.gov/) for enriched biological processes, molecular functions, KEGG pathways and so on.

Gene ontology terms with Benjamini-corrected *p*-values below 0.05 were considered as significantly enriched among the deregulated direct gene targets of HBx, and might be potentially deregulated by HBx.

3.1.7 Analysis of Microarray Expression Profiles of 100 HCC Patients

The Agilent two-colour microarray data for the 100 HCC patients' tumor (labelled as Cy5) and adjacent non-tumor tissues (labelled as Cy3) were analysed using R packages "Loess" for array normalization and "Limma" for detection of differentially expressed genes. Genes with at least 2-fold average expression change in tumor relative to adjacent non-tumor tissues of the 100 HCC patients with multiple test corrected *p*-values below 0.05 were selected as significantly differentially expressed genes in HCC patients.

3.1.8 HCC Patients Clinical Data Analysis to Identify Clinically Associated Deregulated Gene Targets of HBx

From the list of candidate HBx deregulated gene targets identified from THLE3 ChIP-Seq data, those genes that were also significantly differentially expressed in HCC patients' tumor over adjacent non-tumor tissues were first selected. HBx is reported to have oncogenic potential, thus, the gene targets displaying the same deregulation direction in THLE3 cells (AdHBx over AdEasy) and in HCC patients (tumor over adjacent non-tumor tissues) were further selected. That is, if the gene targets were down-regulated when HBx is expressed in THLE3 cells and also down-regulated in tumor compared to adjacent non-tumor tissues of HCC patients, they were considered likely to be related to HCC; and similarly for up-regulation. Thereby those gene targets deregulated consistently in THLE3 cells (AdHBx over AdEasy) and in HCC patients (tumor over adjacent non-tumor tissues) were selected for downstream clinical statistical tests to investigate whether these gene targets had any clinical inferences in HBV-associated HCC.

Statistical tests were performed on these selected deregulated gene targets, using IBM statistical analytical soft tool SPSS 19 (Statistical Package for the Social Sciences) (Mather and Austin, 1983) to find the gene targets significantly associated with HCC patient clinical features. The categorical clinical data of the 100 HCC patients include tumor grade (1, 2, 3 & 4), tumor encapsulation (Yes/No), tumor necrosis (Yes/No), vascular invasion (Yes/No), multifocality (Yes/No), local tumor extension (confined tumor: Yes/No), normal liver cirrhosis (Yes/No), normal liver steatosis (Yes/No), and hepatic dysplasia (Yes/No). Each of the selected HBx deregulated gene targets was tested against each of these clinical features. For a statistical hypothesis testing on the association between a deregulated gene target and a clinical categorical feature, the 100 HCC patients were first divided into two or more than two groups based on the number of different factor groups (e.g. Yes/No) in that clinical feature, and the gene expression values (log2 fold change in tumor over adjacent non-tumor tissues) were then compared across these patient groups, with the null hypothesis that the different patient groups have similar gene expressions.

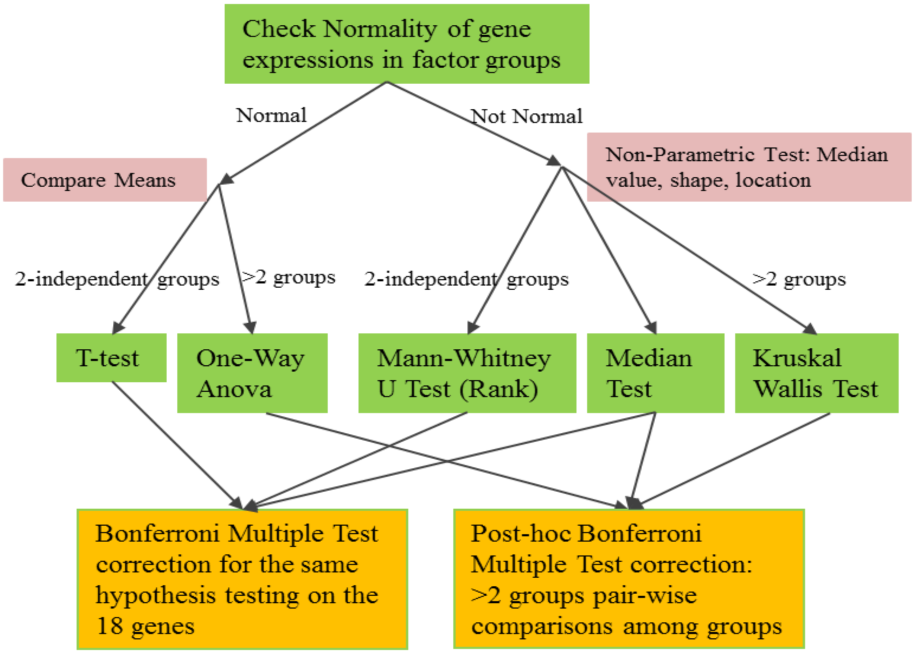

Check Normality of gene expressions in factor groups
Normal
Not Normal
Compare Means
Non-Parametric Test: Median value, shape, location
2-independent groups
>2 groups
2-independent groups
>2 groups
T-test
One-Way Anova
Mann-Whitney U Test (Rank)
Median Test
Kruskal Wallis Test
Bonferroni Multiple Test correction for the same hypothesis testing on the 18 genes
Post-hoc Bonferroni Multiple Test correction: >2 groups pair-wise comparisons among groups

Figure 3.3: Flowchart of the statistical hypothesis testing on the association of HBx deregulated gene targets with patient clinical data. Each of the gene targets was tested on each clinical feature. In a hypothesis testing on a clinical feature, the normality of the gene expression distributions in patient groups (e.g. Yes/No) was first checked: if the normality was valid, T-test or one-way ANOVA was applied depending on the number of patient groups (2 or >2 groups); if the normality not valid, non-parametric tests, like Median test, Mann-Whitney U test and Kruskal-Wallis test, were performed. For clinical features with more than 2 patient groups, Bonferroni multiple test correction was applied to adjust the two-sided *p*-values obtained from post-hoc pair-wise comparisons between the patient groups. Bonferroni multiple test correction was also conducted when the same hypothesis testing on the same clinical feature was repeatedly performed on a set of gene targets. A significant association between a gene target and a clinical feature was obtained if the corrected two-sided *p*-value was less than 0.05.

As shown in Fig 3.3, for a hypothesis testing of the association between a gene target and a clinical feature, the normality of the distribution of the target gene expression in each patient group (e.g. Yes/No) was first checked: if the gene expression for all patient groups were normally distributed, T-test or one-way ANOVA was applied depending on the number of patient groups (T-test for two groups and one-way ANOVA for more than two groups); if normality not valid, non-parametric tests, like Median test, Mann-Whitney U test for two patient groups and Kruskal-Wallis test for more than two patient groups, were performed. For clinical features with more than two groups (e.g. tumor grade: 1, 2, 3 &4), Bonferroni multiple test correction was applied to adjust the two-sided *p*-values obtained from post-hoc pair-wise comparisons among patient groups. Bonferroni multiple test correction of two-sided *p*-values was also conducted when the same hypothesis testing on the same clinical feature was repeatedly performed on a set of HBx deregulated gene targets. A significant association between a gene target and a clinical feature was established if the corrected two-sided *p*-value from the statistical tests were less than 0.05.

In addition to clinical categorical data, 75 out of the 100 HCC patients had the survival profiles: 16 died from HCC and the other 59 patients were considered "censored cases" (48 alive at the time of recording, 6 dead but not due to HCC, and 5 lost of follow up). These "censored cases" were also included in survival time analysis. For survival time analysis on each HBx deregulated gene target, the 75 patients were first divided into 2 groups based on the gene expression in patients: one group with higher expression of that gene in tumor than adjacent non-tumor tissues, and the other group with lower expression of that gene in tumor than adjacent non-tumor tissues. The survival time between the two patients groups were then compared using Kaplan-Meier mean and median survival time tests, with the null hypothesis that the two patient groups have similar survival profiles. Genes with two-sided *p*-value less than 0.05 were regarded highly likely to be associated with patients’ survival time.

3.1.9 Correlation of Expressions of HBx and HBx Deregulated Gene Targets in 100 HCC Patients Tumor and Adjacent Non-Tumor Tissues

With HBx protein expression values measured by our laboratory in tumor and adjacent non-tumor tissues of the 100 HCC patients, a linear regression model for each clinically significant HBx deregulated gene target was built to investigate whether there is any linear correlation/relationship between the expressions of the gene target and HBx protein in HCC patients. Pearson correlations were calculated to measure the strength of the linear relationships between HBx and HBx gene targets expressions, and two-sided *p*-values measuring the significance of the correlations were also computed using SPSS tool.

3.2 Results

3.2.1 Analysis of ChIP-Seq Data and Identification of Potential DNA Binding Sites of HBx

A previous study (Sung et al., 2009) done by our laboratory using ChIP-chip technique on gene promoter regions for UV-treated liver cell line HepG2 cells has suggested that HBx deregulated host gene expressions not by binding directly to gene promoter but through interactions with transcription factors. ChIP-Seq technique is advantageous over ChIP-chip primarily in terms of the capability of predicting the DNA binding sites of HBx globally with larger genome coverage, since ChIP-chip is restricted to the limited number of probes used in array chip. Therefore our laboratory has utilized ChIP-Seq technique on primary liver cell line THLE3 cells (AdHBx vs AdEasy) to comprehensively profile HBx binding sites globally with single-base resolution.

As shown in Fig 3.4, after chromatin immunoprecipitation of DNA fragments bound by HBx-transcription factor complexes and Illumina sequencing of the enriched DNA fragments (ChIP-Seq), we got 14,270,900 and 11,885,806 raw short reads of length 36bp for THLE3 AdEasy and AdHBx cells respectively.

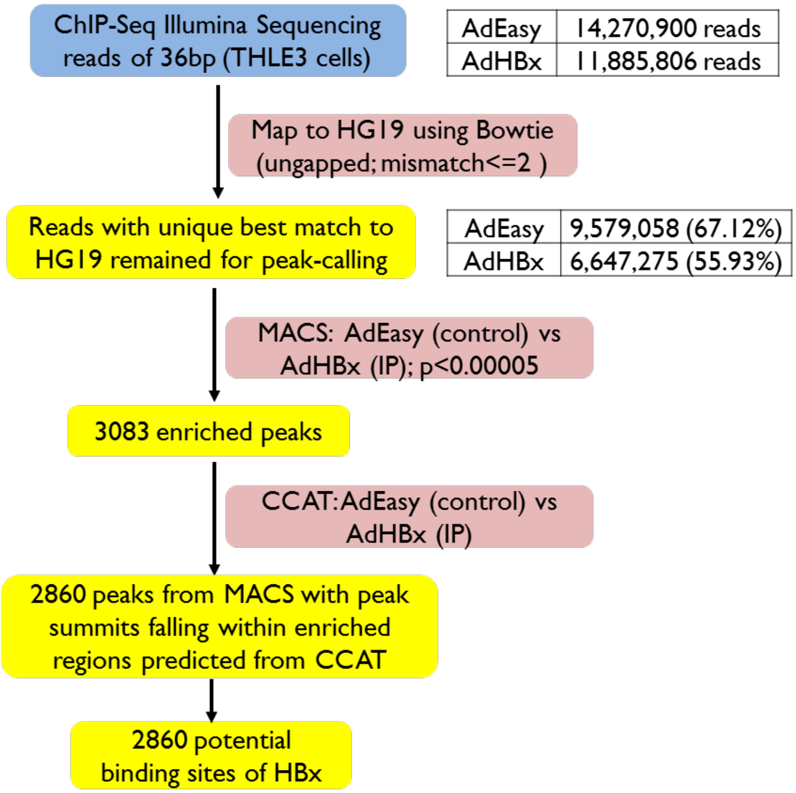

ChIP-Seq Illumina Sequencing reads of 36bp (THLE3 cells)
AdEasy 14,270,900 reads
AdHBx 11,885,806 reads
Map to HG19 using Bowtie (ungapped; mismatch<=2)
Reads with unique best match to HG19 remained for peak-calling
AdEasy 9,579,058 (67.12%)
AdHBx 6,647,275 (55.93%)
MACS: AdEasy (control) vs AdHBx (IP); p<0.00005
3083 enriched peaks
CCAT: AdEasy (control) vs AdHBx (IP)
2860 peaks from MACS with peak summits falling within enriched regions predicted from CCAT
2860 potential binding sites of HBx

Figure 3.4: Processing of ChIP-Seq raw reads to identify potential HBx binding sites. In total our laboratory got 14,270,900 and 11,885,806 raw reads of length 36bp for THLE3 AdEasy and AdHBx cells respectively from Illumina sequencing. The raw reads were first searched against human genome (HG19) using Bowtie with criteria of no gaps permitted and a maximum of 2 mismatches allowed, and reads with unique best match to human genome were then selected. In the end, 67.12% of raw reads for AdEasy and 55.93% for AdHBx cells were used for downstream peak finding process. While treating AdEasy as control, MACS was applied searching for enriched peak regions, and 3083 peaks were found enriched in AdHBx over AdEasy cells. To re-confirm the enrichment of peaks, CCAT was also applied, and 2860 out of the 3083 peaks from MACS were finally obtained with their MACS peak summits falling within enriched regions predicted by CCAT. These 2860 peak regions were the final list of potential genomic binding sites of HBx and used for further analysis.

Following the workflow pipeline illustrated in Fig 3.1, an ultra-fast aligner Bowtie with comparable accuracy was first used to map these short reads to human reference genome, with no gaps permitted and a maximum of two mismatches allowed in the alignments. Because of the short read length (36bp) and the large human genome size, a remarkable portion of these reads could match to multiple regions of human genome. To maintain the balance of mapping accuracy and the number of reads retained for downstream analysis, those reads that matched to multiple regions of human genome with the same best alignment score and significance were removed. Reads with unique best match to human genome were remained for downstream analysis. As a result, 67.12% and 55.93% of the raw reads for AdEasy and AdHBx cells were selected respectively for peak-calling step.

After removal of reads that matched ambiguously to human genome, the peak-calling software tool MACS (Zhang et al., 2008) was applied treating AdEasy sample as negative control, with the criteria of 600bp sliding window, at least 10 folds enrichment in ChIP over control samples and enrichment significance of *p*-value less than 0.00005. In the end, a list of 3083 peaks was obtained with enriched reads in AdHBx relative to AdEasy cells. To re-confirm the enriched peaks, another peak-calling tool CCAT (Xu et al., 2010), which adopts similar algorithm with MACS but is superior in its way to balance the ChIP and control sample size, was simultaneously applied. Of the 3083 peaks, 2860 had their peak summits predicted from MACS falling within the enriched regions predicted from

CCAT, and these 2860 enriched peaks were considered as the potential DNA binding sites of HBx in human genome.

3.2.2 Genome-Wide Distribution of Potential DNA Binding Sites of HBx

Since ChIP-Seq is capable of detecting genome-wide protein binding sites, we may wonder how these predicted potential HBx binding sites distribute over human genome. The 2860 potential HBx binding sites were aligned to the in-house human genome annotation database of HOMER developed by the Glass lab of UCSD (Heinz et al., 2010), which include detailed position information on gene promoters, introns, exons, un-translation regions (5’ or 3’ UTR), and transcription start sites (TSS). The promoter regions were defined as 5kb upstream to the TSS of reference genes, and the intergenic regions were defined as genomic regions other than promoters and gene body regions. Potential HBx binding sites located within multiple genome annotation categories (e.g. peaks fall in the promoter of one gene but exon of another gene) were classified based on the precedence order that promoter comes first followed by 5' UTR and 3' UTR which then precede introns and exons. Distributions of the 2860 potential HBx binding sites on each annotation category were shown in Fig 3.5A.

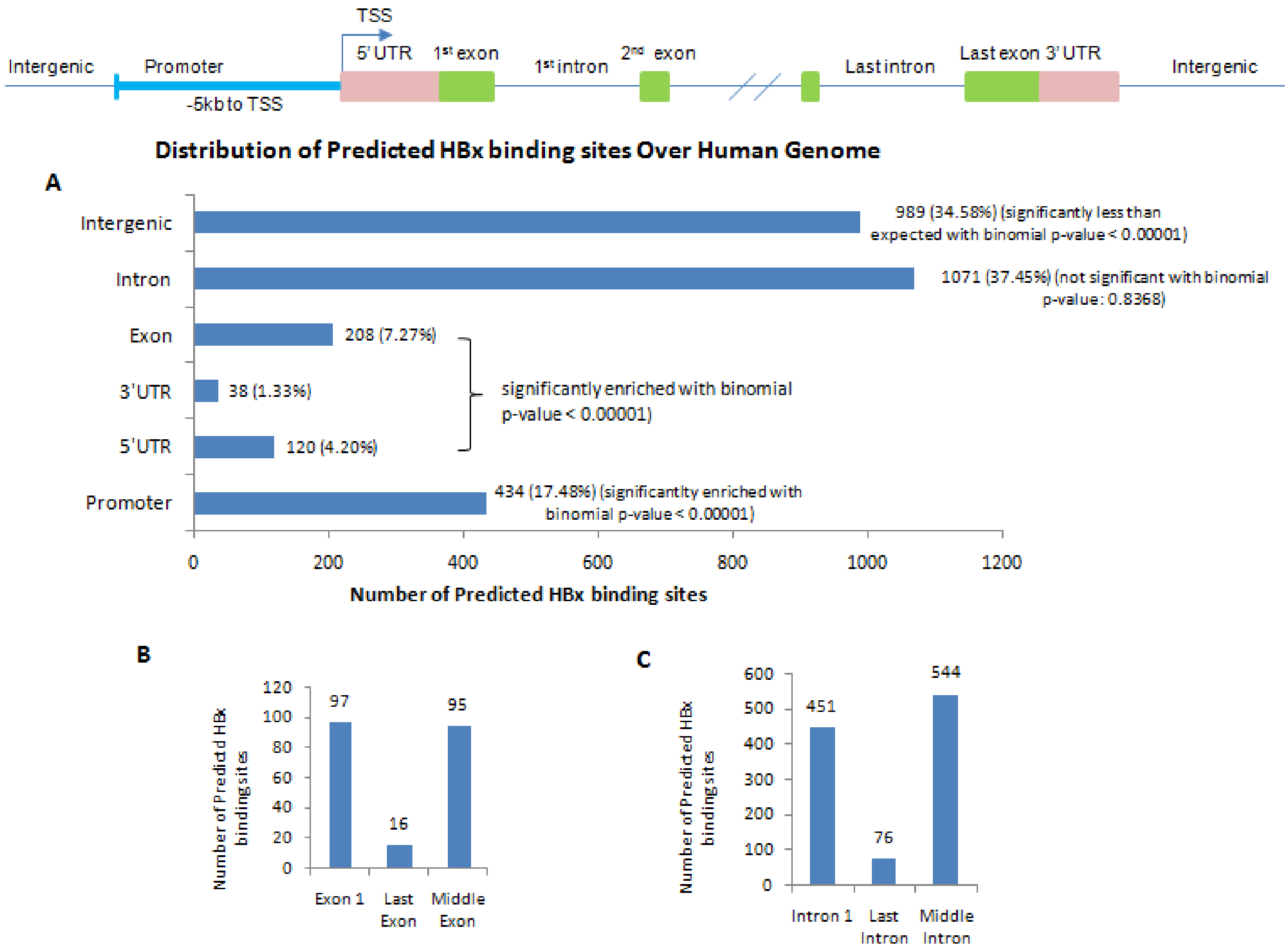


Figure 3.5: Genome-wide distribution of the 2860 potential DNA binding sites of HBx predicted from ChIP-Seq data in THLE3 cells. Promoter is defined as 5kb upstream to transcription start sites (TSS) of human genes. A) Distribution of the 2860 potential HBx binding sites on different regions of genome: promoters, introns, exons, 5'UTR, 3'UTR, and intergenic regions. Since introns and intergenic regions are the two longest categories in human genome, most potential binding sites located within introns of genes (37.45%) and intergenic regions (34.58%). Next abundant is the promoter, followed by exons, 5'UTR and 3'UTR. Importantly, these 2860 predicted HBx binding sites were found to be significantly enriched in exons and promoter regions of genes and significantly less distributed in intergenic regions with binomial two-tailed p-values less than 0.00001. B) Almost half of the binding sites within introns located in the first and last introns. C) Almost half of the binding sites within exons located in the first and last exons. Enrichment on the first and last introns/exons relative to other middle introns/exons indicates the potential regulatory effect of HBx on gene expressions.

We could see that most of the potential HBx binding sites located within introns of genes (37.45%) and intergenic regions (34.58%), which is not surprising since introns and intergenic regions are the two longest categories in human genome. Next abundant are the promoters, exons, 5'UTR and 3'UTR in descending order. However, these 2860 predicted HBx binding sites were found to be significantly enriched in exons and promoter regions of genes and significantly less distributed in intergenic regions with binomial two-tailed p-values less than 0.00001. Among the potential binding sites within introns and exons (Fig 3.5B and 3.5C), we found almost half were located within the first and last introns or exons of genes. Enrichment of the potential binding sites of HBx in the first and last introns and exons of genes relative to other middle introns and exons might suggest the potential regulatory effect of HBx on gene expression.

3.2.3 Potential HBx-Interacting Transcription Factors Predicted from HepG2 ChIP-chip and THLE3 ChIP-Seq Data

To identify the transcription factors that may potentially interact with HBx in THLE3 cells, the genomic DNA sequences for the 2860 global potential HBx binding sites were extracted and scanned against the known transcription factor binding motifs in TRANSFAC database (Wingender et al., 1996) using motif finding scripts from HOMER (Heinz et al., 2010). Of the 601 vertebrate motif matrices tested, 195 transcription factor motifs were found significantly enriched/over-represented within the 2860 potential HBx binding sites in THLE3 cells (with HOMER *p*-value < 0.05). These 195 transcription factor motifs were ranked according to the *p*-values, shown in Supplementary Table S2.

As mentioned earlier, a previous ChIP-chip study done by our laboratory (Sung, Lu et al. 2009) has identified 144 potential HBx-interacting transcription factors in UV-treated and HBx-expressing adenoviruses transfected HepG2 cells. Among these 195 potential HBx-interacting transcription factors predicted from THLE3 cells, 129 were commonly predicted from the previous ChIP-chip study on HepG2 cells (Sung et al., 2009) (Table 3.1).

Table 3.1: Comparisons of ChIP-chip data on HepG2 cells (Sung et al., 2009) and ChIP-Seq data on THLE3 cells.

	ChIP-chip	ChIP-Seq	Overlap
Liver cell line	HepG2 (UV-treated)	immortalized THLE3	N.A
Differential gene expressions (AdHBx over AdEasy)	10,145 (Fold change >=2)	3,876 (Fold change >=1.5)	1213 (646 same deregulation direction; 567 opposite)
Genome coverage	1.5kb gene promoters	whole genome	N.A
Potential HBx binding sites	971	2860	7
Potential HBx deregulated direct gene targets	184 (Fold change >=2)	143 (Fold change >=1.5)	2 (1 same deregulation direction; 1 opposite)
Potential HBx-interacting transcription factors	144	195	129

The 129 potential HBx-interacting transcription factors commonly predicted from THLE3 ChIP-Seq data and HepG2 ChIP-chip data include previously reported ones that either interact with HBx or are activated by HBx, such as, SP1 (Lee et al., 1998), AP1 (Benn et al., 1996), AP2 (Kim and Rho, 2002), E2F (Weinmann et al., 2001), E2F1 (Choi et al., 2002; Sung et al., 2009), CREB (Maguire et al., 1991), SMAD4 (Sung et al., 2009), YY1 (Sung et al., 2009), NFKAPPAB50 (Su and Schneider, 1996), STAT3 (Waris et al., 2001), and so on. The 195 enriched transcription factors from THLE3 ChIP-Seq data also include C-Myc (Li et al., 2003; Zeller et al., 2006) and P53 (Wei et al., 2006), that were previously reported to interact with HBx but not predicted by HepG2 ChIP-chip data. This could be an evidence of the improvement of ChIP-Seq over ChIP-chip in reducing false negative rates, since the ChIP-chip data only covered 1.5kb promoter regions of human genes while ChIP-Seq was able to detect binding sites over entire genome.

Though around 90% of the potential HBx-interacting transcription factors (129/144) in HepG2 cells were also found in THLE3 cells, only 7 out of the 971 HBx binding sites predicted in HepG2 cells overlapped with the 2860 HBx binding sites predicted in THLE3 cells, as shown in Table 3.1. There is also very little overlap on HBx deregulated direct gene targets predicted in the two datasets (Table 3.1). This indicates that, even though THLE3 ChIP-Seq and HepG2 ChIP-chip data predicted most similar sets of potential HBx-interacting transcription factors, the deregulated gene targets of HBx were very different between the two cell lines. Possible explanations for this gene targets difference may be that: a) HepG2 and THLE3 are two different primary liver cell lines and they might be physiologically very different; and b) HepG2 cells were UV-treated before chromatin immunoprecipitation while THLE3 cells were not. From the microarray gene expression profiles shown in Table 3.1, it was also observed that gene expressions changed much more drastically in UV-treated HepG2 cells than genes in THLE3 cells: 10,145 genes with above 2 fold change in UV-treated HepG2 AdHBx over AdEasy cells (Sung et al., 2009), while only 3,876 genes with above 1.5 fold change in THLE3 AdHBx over AdEasy cells. This indicated that UV-treatment might have enforced the regulation effect of HBx on gene expressions in HepG2 cells and this may possibly contribute to the differences in HBx gene targets between the two liver cell lines (HepG2 and THLE3).

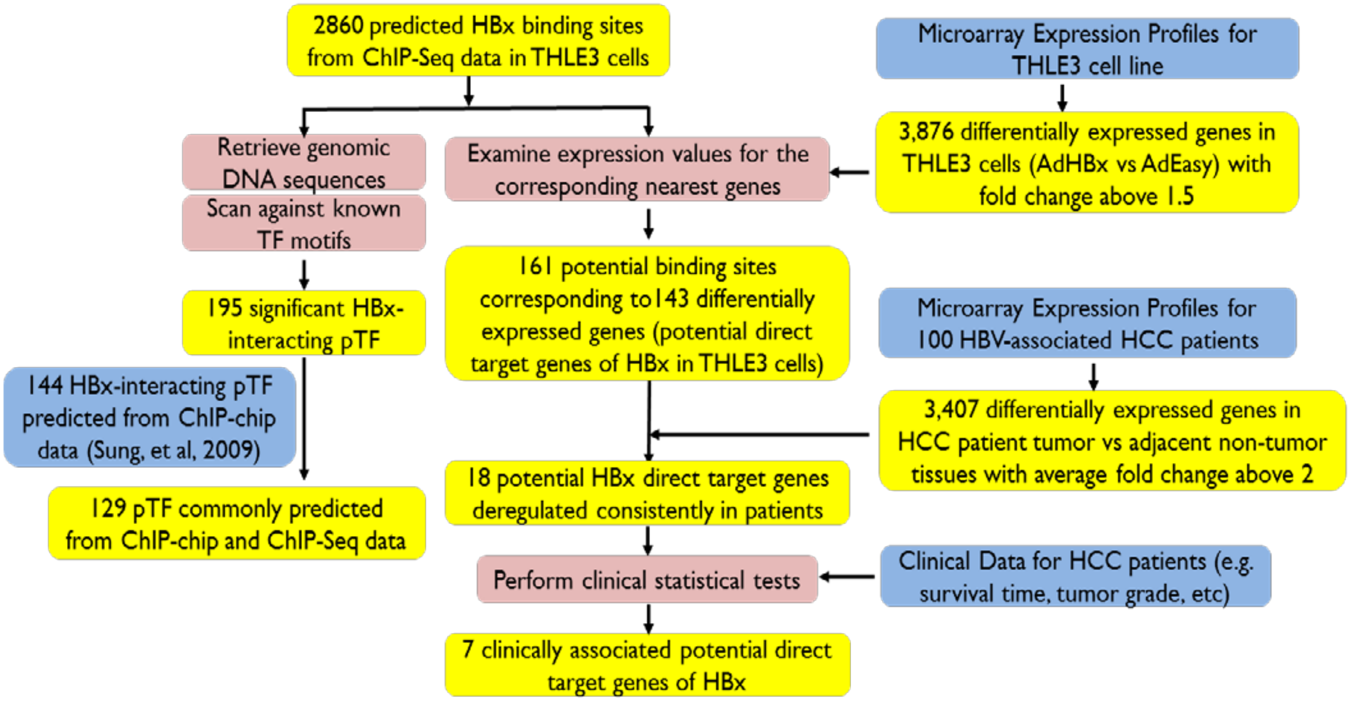

2860 predicted HBx binding sites from ChIP-Seq data in THLE3 cells
Microarray Expression Profiles for THLE3 cell line
Retrieve genomic DNA sequences
Scan against known TF motifs
Examine expression values for the corresponding nearest genes
3,876 differentially expressed genes in THLE3 cells (AdHBx vs AdEasy) with fold change above 1.5
195 significant HBx-interacting pTF
161 potential binding sites corresponding to143 differentially expressed genes (potential direct target genes of HBx in THLE3 cells)
Microarray Expression Profiles for 100 HBV-associated HCC patients
144 HBx-interacting pTF predicted from ChIP-chip data (Sung, et al, 2009)
3,407 differentially expressed genes in HCC patient tumor vs adjacent non-tumor tissues with average fold change above 2
18 potential HBx direct target genes deregulated consistently in patients
129 pTF commonly predicted from ChIP-chip and ChIP-Seq data
Perform clinical statistical tests
Clinical Data for HCC patients (e.g. survival time, tumor grade, etc)
7 clinically associated potential direct target genes of HBx

Figure 3.6: Summary of the computational analysis results for identification of HBx binding sites, potential HBx-interacting transcription factors and HBx deregulated direct gene targets. Genomic sequences for the 2860 potential HBx binding sites were extracted and scanned against TRANSFAC transcription factor binding motif database, and 195 transcription factors were found significantly enriched. Compared with the 144 transcription factors predicted from HepG2 ChIP-chip data (Sung et al., 2009), there were 129 transcription factors commonly found in ChIP-chip and ChIP-Seq data. Analysis of microarray expression profiles identified 3,876 genes differentially expressed in THLE3 cell AdHBx over AdEasy with fold change above 1.5. Integration of the expression values for the nearest genes of the 2860 potential HBx binding sites allowed identification of 161 potential binding sites corresponding to 143 differential genes in THLE3 cells. Analysis of microarray expression profiles in HCC patients predicted 3,407 genes differentially expressed with average fold change above 2 in patient tumor over adjacent non-tumor tissues. Of the 143 genes, 18 were differentially expressed in patients with the same deregulation direction in THLE3 cells. HCC patients' clinical data were also integrated and statistical tests were performed on the 18 genes, of which, 7 were found clinically associated. These 7 clinically associated genes were considered as the candidate HBx deregulated direct gene targets that may potentially be related to hepatocarcinogenesis.

3.2.4 Potential HBx Deregulated Direct Gene Targets in THLE3 Cells

HBx has been implicated to interact with transcription factors, change the DNA binding affinity of the transcription factors, and consequently regulate gene transcription and expression. To identify the deregulated direct gene targets of HBx, expression profiles for the nearest genes of the 2860 potential HBx binding sites were examined in THLE3 cells. As summarized in Fig 3.6, of the 2860 potential HBx binding sites, 161 sites corresponding to 143 genes displayed differential gene expressions (fold change above 1.5) in THLE3 cells (AdHBx over AdEasy). These 143 differentially expressed genes with potential HBx binding sites nearby were identified as the potential deregulated direct gene targets of HBx. Gene ontology analysis of these 143 potential deregulated direct gene targets of HBx showed that the top two significantly enriched biological processes are developmental process (Benjamini-corrected *p*-value: 4.14E-06) and multicellular organismal process (Benjamini-corrected *p*-value: 1.08E-04). These two biological processes were also found significantly enriched in the 184 potential deregulated gene targets of HBx predicted from the ChIP-chip data on UV-treated HepG2 cells (Sung et al., 2009). The top significantly enriched molecular function of the 143 potential deregulated gene targets of HBx in THLE3 cells is transcriptional factor activity (Benjamini-corrected *p*-value: 0.045388), which, however, was not significantly enriched in the 184 potential deregulated direct gene targets of HBx in HepG2 cells.

3.2.5 Clinically Associated Potential HBx Deregulated Gene Targets

3.2.5.1 Expression of Potential HBx Deregulated Gene Targets in HCC Patients

These 143 potential deregulated gene targets of HBx were identified in primary liver THLE3 cell line. Analysis of the microarray expression profiles in 100 HBV-associated HCC patients found 3,407 genes differentially expressed in patients' tumor over adjacent non-tumor tissues with average fold change above 2 and adjusted *p*-values less than 0.05. To further examine whether these 143 potential gene targets were truly related to HCC, their expression values in the 100 HCC patients were investigated, and 23 out of the 143 genes were found significantly differentially in HCC patients' tumor and adjacent non-tumor tissues with average fold change above 2.

As shown in the hierarchical clustering graph of the 23 genes' expressions in the 4 biological replicates of THLE3 cells and the 100 HCC patients (Fig 3.7), 18 out of the 23 genes had the same deregulation direction in THLE3 AdHBx over AdEasy cells and in HCC patients' tumor over adjacent non-tumor tissues. That is, these 18 genes were deregulated consistently in THLE3 cells (AdHBx over AdEasy) and in 100 HCC patients (tumor over adjacent non-tumor tissues). Of these 18 genes, 15 were consistently down-regulated in THLE3 cells and HCC patients; while the other 3 genes were consistently up-regulated. HBx is of oncogenic potential, so having the same deregulation direction in tumor over adjacent non-tumor tissues of HCC patients and in THLE3 cells when HBx is expressed indicates that these 18 genes might be potentially associated with hepatocellular carcinogenesis and development of HBV-associated HCC.

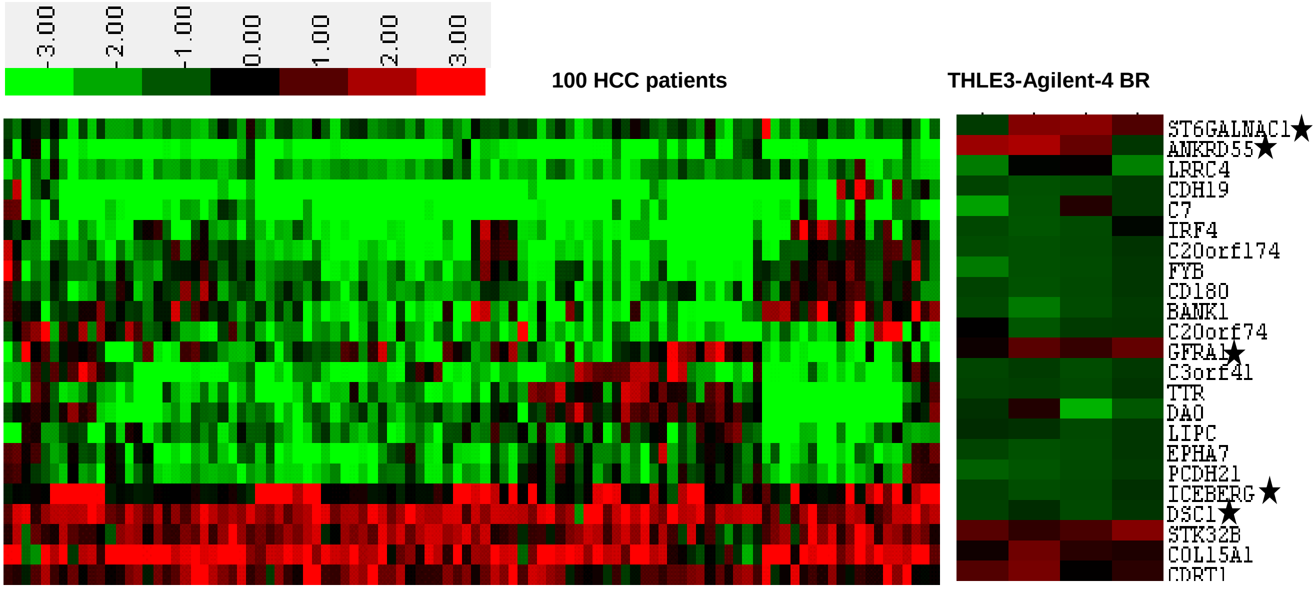


Figure 3.7: Hierarchical clustering of the 23 potential HBx deregulated gene targets that are significantly differentially expressed between tumor and adjacent non-tumor tissues of the 100 HCC patients with average fold change above 2. The log2 fold change of tumor over adjacent non-tumor tissues on each patient, and the log2 fold change of THLE3 AdHBx over AdEasy cells (four biological replicates), were used for clustering (average linkage). Green represents down-regulation in HCC patients' tumor over adjacent non-tumor tissues or in THLE3 AdHBx over AdEasy cells, while red represents up-regulation. The 5 genes marked with black star are the ones with opposite deregulation directions in THLE3 cells and HCC patients, while the remaining 18 genes are consistently deregulated in THLE3 cells and HCC patients indicating potential clinical inferences in HCC patients.

3.2.5.2 Association of Potential HBx Deregulated Gene Targets with HCC Patient Survival Time

As shown in Fig 3.6, to further evaluate the association of the 18 HBx deregulated gene targets with HCC, the gene expression profiles and clinical data including survival profiles of the 100 HCC patients were integrated, and various statistical tests were conducted searching for gene targets with significant associations with patient clinical features. Of these 100 patients, 75 had survival time data consisting of 16 patients dead from HCC cancer and the other 59 patients classified as “censored cases” (48 still alive at the time of recording, 6 dead due to reasons other than HCC and 5 lost of follow-up). These “censored cases” cases were still used in survival time analysis. Kaplan-Meier survival time analysis were performed for each of the 18 potential HBx deregulated gene targets by first dividing the patients into two groups (one group with gene expression in tumor tissue higher than adjacent non-tumor tissue, and the other group with gene expression in tumor tissue lower than adjacent non-tumor tissue) and then comparing the survival time between the two patient groups. The null hypothesis was that patient group with gene expression higher in tumor than adjacent non-tumor tissue have the same survival time with the patient group with gene expression lower in tumor than adjacent non-tumor tissue. The significance *p*-values obtained from Kaplan-Meier survival time analysis for each of the 18 target genes are shown in the 2nd column of Table 3.2.

Table 3.2: Summary of corrected two-sided significance values from the clinical statistical tests on the 18 potential HBx deregulated gene targets. For a hypothesis testing of a gene target on a clinical categorical feature, the patients were first divided into different categorical groups and the gene expressions were then compared between patient groups. If the gene expressions in each patient group follow normal distribution, T-test (for features with 2 sub-groups) or one-way ANOVA (for features with >2 sub-groups) were performed; else, median test and Mann-Whitney U test (2 sub-groups) or Kruskal-Wallis test (>2 sub-groups) were performed. In this table, cells with only single number represents *p*-values from T-test or one-way ANOVA, while cells with two numbers in a bracket had the first number being the *p*-value from median test and the 2nd number from Mann-Whitney U test or Kruskal-Wallis test. For Kaplan-Meier survival time analysis, patients were first divided into two groups based on gene expressions in patient tumor and adjacent non-tumor tissues and the survival time were then compared between the two patient groups. The 2nd column shows the two-sided *p*-values for survival time analysis. For clinical features with only 2 sub-groups, Bonferroni multiple test correction was applied to adjust the two-sided *p*-values (column 3 to 10). For clinical features with more than 2 sub-groups, the *p*-values from one-way ANOVA or Kruskal-Wallis tests were first checked, and if a *p*-value was significant (less than 0.05), pair-wise comparisons and post-hoc Bonferroni corrections were then further performed. The feature in column 11 had 3 sub-groups but the *p*-values were all not significant, so pair-wise comparisons were not conducted. The category "Tumor Grade" had 4 sub-groups, and column 12 shows the *p*-values from one-way ANOVA, median test, or Kruskal-Wallis tests. Pair-wise comparisons (6 comparisons for 4 sub-groups) were then conducted for those with *p*-values less than 0.05 in column 12, and the Bonferroni corrected *p*-values for pair-wise comparisons were shown in the last column. In total, 6 genes showed significant clinical associations (shaded in yellow): BANK1, STK32B, DAO, C20orf74, FYB and CRDT1. Though DAO and TTR had survival time *p*-values not significant (shaded in grey), they did show a clear survival difference between patient groups (Fig.3.7).

18 Target Genes	Survival Time	Hepatic Capsule (TumorFree/ Tumor Present)	Hepatic Displasia (Yes/No)	Liver Invasion (Confined to liver/Tumor Invades)	Multifocal (Yes/No)	Normal Liver Cirrhosis (Yes/No)	Tumor Capsule (Yes/No)	Tumor Necrosis (Yes/No)	Vascular Invasions (Yes/No)	Viral Infection (HBV, HCV, None)	Tumor Grade (1,2,3,4)	
											>2 groups	pair-wise comparisons
TTR	0.063	1	1	1	1	0.318	1	1	1	0.324	(0.179, 0.332)	
EPHA7	0.104	1	1	1	1	1	1	1	1	0.473	0.13	
C20orf174	0.691	0.66	1	1	1	0.072	1	1	1	0.508	0.167	
CD180	0.427	1	(1, 1)	1	1	0.528	1	1	1	0.951	(0.258, 0.152)	
BANK1	0.957	1	(1, 1)	1	0.756	(0.012, <0.012)	1	0.54	(1, 1)	0.169	0.772	
STK32B	0.017	(1, 1)	1	<0.008	1	1	1	(1, 1)	1	0.778	(0.117, 0.143)	
CDH19	0.157	(0.384, 1)	(1, 1)	(1, 0.91)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(0.953, 0.591)	(0.767, 0.114)	
C7	0.542	(1, 1)	(1, 1)	(1, 0.51)	(0.187, 0.792)	(0.588, 1)	(1, 1)	(1, 1)	(1, 0.71)	(0.769, 0.731)	(0.030, 0.035)	(1;1;1;0.48;0.426;0.192, 1;1;1;1;0.288;0.114)
LRRC4	0.669	(1, 1)	1	(1, 1)	(1, 1)	(0.936, 1)	1	(1, 1)	(1, 1)	(0.670, 0.925)	(0.328, 0.807)	
C3orf41	0.948	(1, 1)	(1, 1)	1	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(0.42, 0.13)	(0.663, 0.166)	(0.561, 0.594)	
PCDH21	0.121	(1, 1)	1	(1, 1)	(1, 1)	1	(1, 1)	(1, 1)	(1, 1)	(0.464, 0.876)	(0.346, 0.381)	
LIPC	0.178	(1, 1)	1	1	(1, 1)	(1, 1)	1	1	1	(0.464, 0.481)	(0.767, 0.453)	
DAO	0.094	(1, 1)	1	(1, 1)	(1, 0.748)	(1, 0.66)	(1, 1)	(1, 1)	0.232	0.183	(0.097, 0.015)	1;1;0.827;0.07;0.059; 0.013 (grade 2&4)
C20orf74	0.003	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(0.588, 0.012)	0.572	(1, 1)	(1, 1)	(0.822, 0.826)	(0.767, 0.800)	
IRF4	0.452	1	1	(1, 1)	0.567	(1, 0.612)	(1, 0.917)	(1, 1)	(1, 1)	0.399	0.152	
FYB	0.364	(1, 1)	1	(1, 1)	(0.044, 0.638)	(0.588, 0.204)	1	(1, 1)	1	(0.743, 0.847)	(0.062, 0.040)	1;1;0.75;0.642;0.462; 0.174
CDRT1	0.955	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(1, 1)	0.022	(1, 1)	(1, 1)	(0.232, 0.234)	(0.301, 0.523)	
COL15A1	0.608	(1, 1)	(1, 1)	(1, 1)	(1, 1)	(0.588, 0.372)	(1, 1)	(1, 1)	(1, 1)	(0.351, 0.281)	(0.049, 0.189)	1;1;1;1;1;0.114

Two genes C20orf74 and STK32B were found showing significant differences in survival time between patient groups with two-sided *p*-value less than 0.05. As illustrated in Fig 3.8, HCC patients with lower C20orf74 expression in tumor than adjacent non-tumor tissues were more likely to have longer survival time than patients with higher C20orf74 expression in tumor than adjacent non-tumor tissues (two-sided *p*-value: 0.003). Patients with higher STK32B expression in tumor than adjacent non-tumor tissues were more likely to have longer survival time than patients with lower STK32B expression in tumor than adjacent non-tumor tissue (two sided *p*-value: 0.017). Another two genes DAO and TTR showed clear differences in survival time between patient groups though not statistically significant. Similar to STK32B, patients with higher DAO and TTR expressions in tumor than adjacent non-tumor tissues appear to have longer survival time than patients with lower DAO and TTR expressions in tumor than adjacent non-tumor tissues (not statistically significant).

In summary, higher expressions of STK32B, TTR, and DAO, and lower expression of C20orf74 in tumor over adjacent non-tumor tissues in HCC patients are associated with longer survival time. These four genes were also functionally important (Table 3.4): C20orf74 is a Ral GTPase activating protein involving in regulation of signal transduction; STK32B is a serine/threonine kinase which is important for protein amino acid post-translational modifications; DAO is a D-amino acid oxidase participating in cellular amino acid metabolic process; and TTR possesses transporter activity. Together with the evidence of being potentially directly deregulated by HBx in THLE3 cells and being significantly differentially expressed in tumor over adjacent non-tumor tissues of HCC patients, these four survival-associated genes, C20orf74, STK32B, DAO and TTR, might be important for HCC development and worth further investigations.

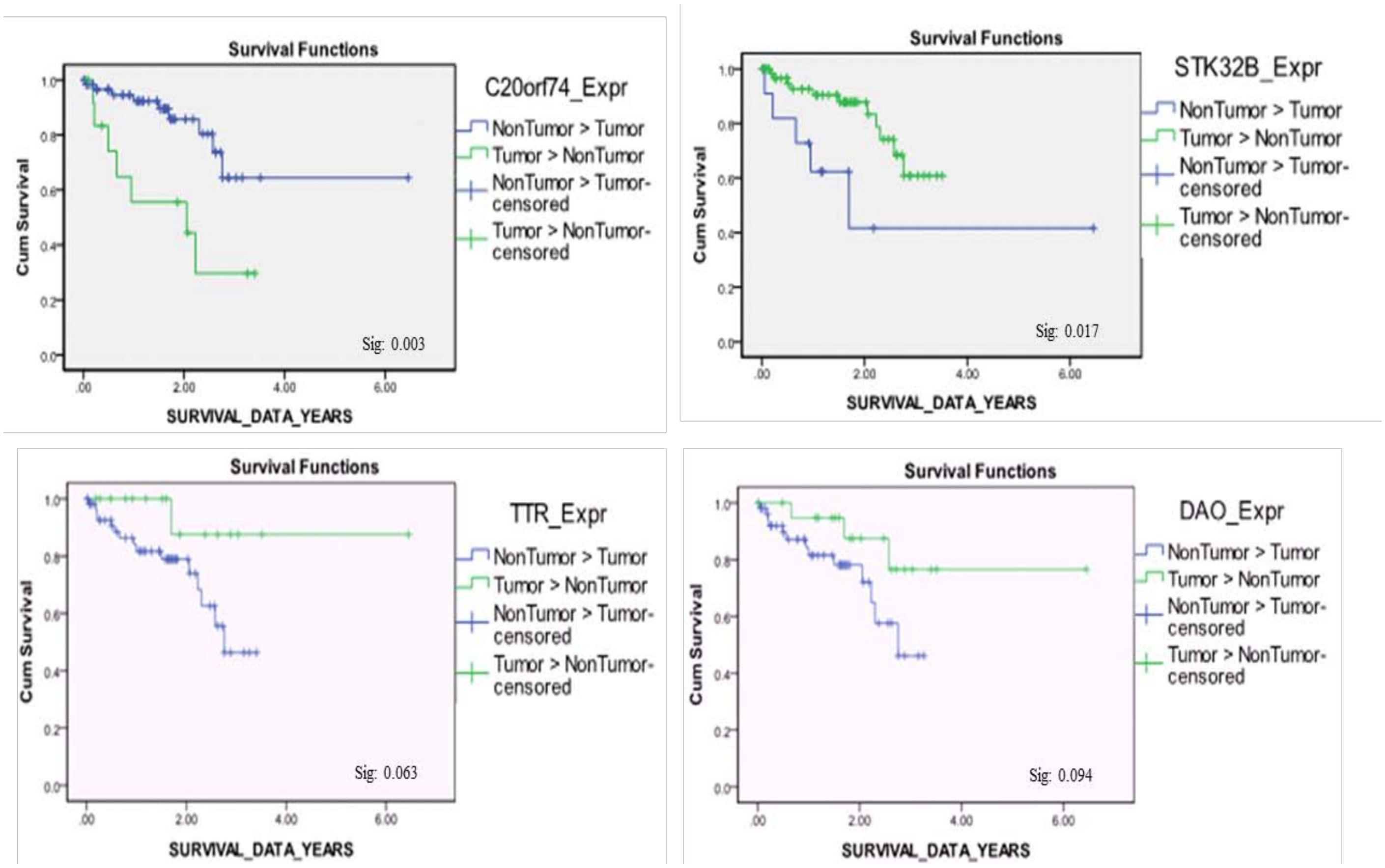


Figure 3.8: Survival plots for the four survival-associated potential HBx deregulated gene targets. For each potential HBx gene target, patients were divided into two groups: one with higher expression of that gene in tumor than adjacent non-tumor tissues; and the other with lower expression of that gene in tumor than adjacent non-tumor tissues. Kaplan-Meier survival analysis was conducted comparing the survival time between the two patient groups. Two genes C20orf74 and STK32B showed significant survival time differences between patient groups (2-sided *p*-value: 0.003 and 0.017 respectively), and another two genes TTR and DAO showed clear differences though not statistically significant. Higher expression of STK32B, TTR, and DAO, and lower expression of C20orf74 in tumor over adjacent non-tumor tissues in HCC patients are associated with longer survival time. The plus sign on the survival curve refers to “censored” cases defined as those patients either still alive at the time of recording or dead from reasons other than HCC or lost of follow-up. These “censored” cases were also used in Kaplan-Meier survival analysis.

3.2.5.3 Association of Potential HBx Deregulated Gene Targets with HCC Patients' Categorical Clinical Features

In addition to HCC patients' survival profiles, other categorical clinical data were also available in our laboratory, including tumor grade (1, 2, 3 & 4), tumor encapsulation (Yes/No), tumor necrosis (Yes/No), vascular invasion (Yes/No), multifocality (Yes/No), local tumor extension (confined tumor: Yes/No), normal liver cirrhosis (Yes/No), normal liver steatosis (Yes/No), and hepatic dysplasia (Yes/No). Statistical analytic soft tool SPSS version 19 (Mather and Austin, 1983) was applied to perform T-test, one-way ANOVA, or non-parametric tests, such as median test, Mann-Whitney U test and Kruskal-Wallis test. Each of the 18 potential HBx deregulated genes was tested on each of the clinical categorical features, with statistical tests that were well chosen based on the properties of gene expression data in patients as described in Fig 3.3 and Table 3.2. For a hypothesis testing of a gene target on a clinical feature, the 100 HCC patients were first divided into two or more than two groups based on the sub factor groups of that clinical feature (e.g. Yes/No), and the gene expressions were then compared between these patient groups. The log2 fold change in patients' tumor over adjacent non-tumor tissues were used as gene expression values. The null hypothesis was that these patient groups have similar gene expression profiles. Bonferroni multiple test correction was also applied in the end to adjust the two-sided *p*-values when the same hypothesis testing was performed repeatedly on a set of gene targets or when post-hoc pairwise comparisons were conducted for clinical features with more than two sub factor groups (e.g. tumor grade: 1, 2, 3 &

4). A significant association of a gene target with a patient clinical feature was established when the corrected *p*-value was less than 0.05. Table 3.2 listed the multiple test corrected *p*-values of all the statistical tests that have been performed for the 18 potential HBx deregulated gene targets on the various clinical features, and the significant *p*-values obtained were highlighted in yellow.

In summary, among the 18 potential HBx deregulated gene targets, 6 genes were found displaying significant associations with various HCC patients' clinical categorical features. These 6 genes were BANK1, STK32B, DAO, C20orf74, FYB and CDRT1, and were significantly associated with tumor grade, liver invasions, multifocality of patients, normal liver cirrhosis, and tumor encapsulations. Details of the significant associations were shown in Fig 3.9.

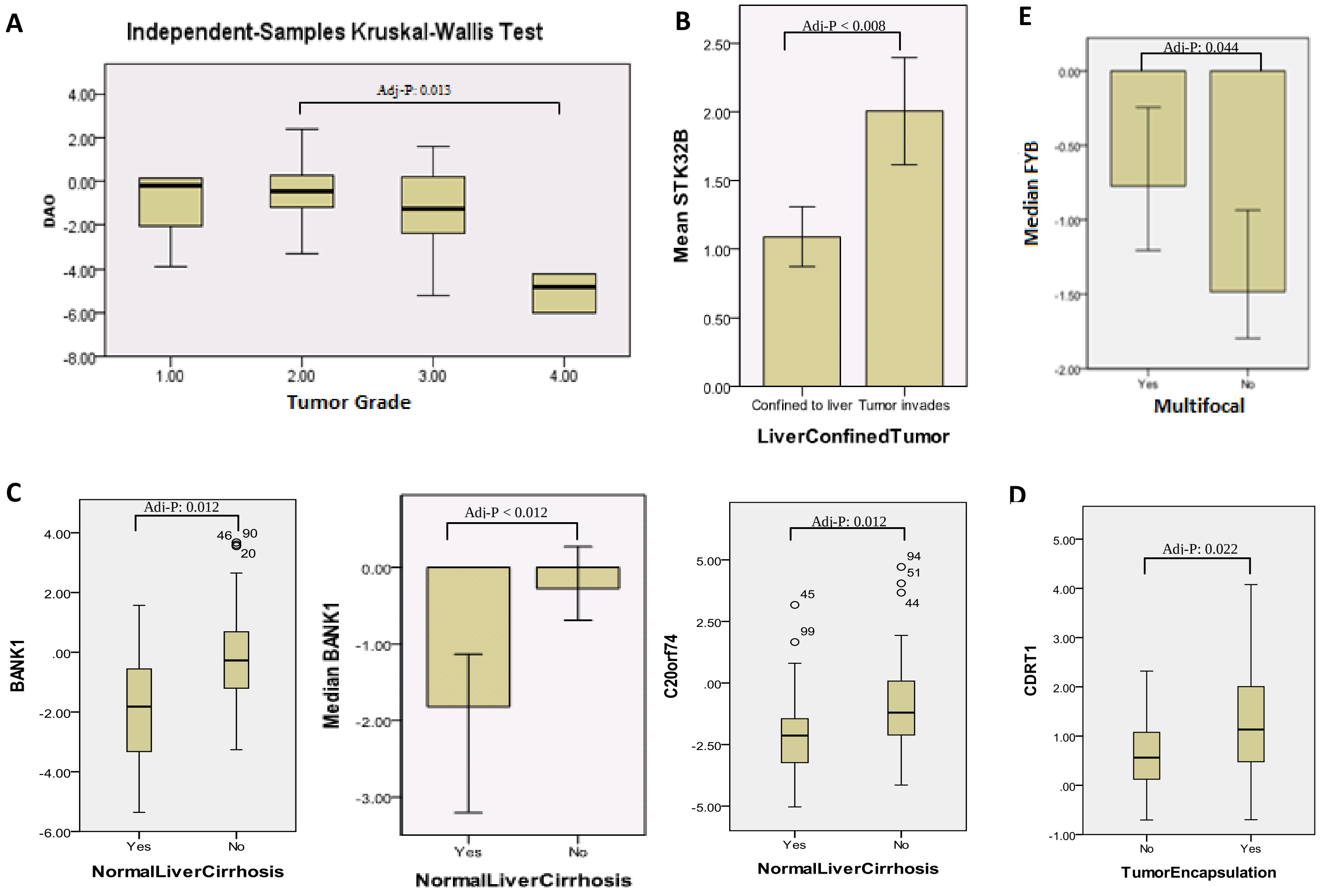


Figure 3.9: Plots for the six potential HBx deregulated gene targets that showed significant associations with the 100 HCC patients' categorical clinical features. X-axis represents the sub factor groups of the clinical features, and y-axis shows the log2 fold change expression values of genes in tumor over adjacent non-tumor tissues of HCC patients. Of the 18 potential gene targets, the following significant associations with corrected *p*-values less than 0.05 were found: A). Patients with tumor grade 4 are highly likely to have stronger down-regulation of DAO in tumor over adjacent non-tumor tissues than patients with tumor grade 2, with corrected *p*-value of 0.013 obtained from non-parametric Kruskal-Wallis test. B). Patients with local tumor invasion are highly likely to have stronger up-regulation of STK32B in tumor over adjacent non-tumor tissues than patients with tumor confined to liver, with corrected *p*-values below 0.008 obtained from 2-independent samples T-test. C). Patients with normal liver cirrhosis are highly likely to have stronger down-regulation of BANK1 and C20orf74 in tumor over adjacent non-tumor tissues than patients with no normal liver cirrhosis, with corrected *p*-values of 0.012 obtained from median test and Mann-Whitney U tests. D). Patients with tumor encapsulation are highly likely to have stronger up-regulation of CDRT1 in tumor over adjacent non-tumor tissues than patients with no tumor encapsulation, with corrected *p*-value of 0.022 obtained from Mann-Whitney U test. E). Patients without multifocality are likely to have stronger down-regulation of FYB in tumor over adjacent non-tumor tissues than multifocal patients, with corrected *p*-value of 0.044 obtained from median test.

As plotted and explained in Fig 3.9, the associations of the six potential HBx deregulated gene targets with various categorical clinical features could be interpreted and summarized as follows:

a) Patients with tumor grade 4 are highly likely to have stronger down-regulation of DAO in tumor over adjacent non-tumor tissues than patients with tumor grade 2, with Bonferroni multiple test corrected *p*-value of 0.013 obtained from post-hoc pair-wise comparisons following non-parametric Kruskal-Wallis test.

b) Patients with local tumor invasion are highly likely to have stronger up-regulation of STK32B in tumor over adjacent non-tumor tissues than patients with tumor confined to liver, with Bonferroni multiple test corrected *p*-values below 0.008 obtained from 2-independent samples T-test.

c) Patients with normal liver cirrhosis are highly likely to have stronger down-regulation of BANK1 and C20orf74 in tumor over adjacent non-tumor tissues than patients with no normal liver cirrhosis, with Bonferroni multiple test corrected *p*-values of 0.012 obtained from median test and Mann-Whitney U tests.

d) Patients with tumor encapsulation are highly likely to have stronger up-regulation of CDRT1 in tumor over adjacent non-tumor tissues than patients with no tumor encapsulation, with Bonferroni multiple test corrected *p*-value of 0.022 obtained from Mann-Whitney U test.

e) Patients without multifocality are likely to have stronger down-regulation of FYB in tumor over adjacent non-tumor tissues than multifocal patients, with Bonferroni multiple test corrected *p*-value of 0.044 obtained from median test.

3.2.5.4 Summary of Associations of Potential HBx Deregulated Gene Targets with HCC Patient Clinical Features

As shown in Fig 3.8 & 3.9 and summarized in Table 3.3, there were in total seven potential HBx deregulated direct gene targets that might be associated with HCC patients' survival profiles and various categorical clinical features including tumor grade, liver invasions, multifocality of patients, normal liver cirrhosis, and tumor encapsulations. The potential clinical associations and inferences of these 7 potential HBx gene targets could be summarized as following points:

a) HCC patients with tumor grade 4 were highly likely to have stronger down-regulation of DAO in tumor over adjacent non-tumor tissues than patients with tumor grade 2 (adjusted *p*-value: 0.013). Furthermore, DAO also displayed clear survival differences between patient groups with higher and lower DAO expressions in tumor over adjacent non-tumor tissues, though statistically not significant with two-sided *p*-value of 0.094. Higher expressions of DAO in tumor over adjacent non-tumor tissues in HCC patients are likely to be associated with longer survival time (statistically not significant). In other words, lower expressions of DAO in tumor over adjacent non-tumor tissues in HCC patients are likely to be associated with shorter survival time. Therefore, in summary, lower expression of DAO in patients' tumor over adjacent non-tumor tissues is associated with shorter survival time and larger tumor grade. These associations of DAO with shorter survival time and larger tumor grade seemed to be valid, in the sense that, HCC patients with larger tumor grade are more likely to have shorter survival time.

b) Two other significant survival-associated potential HBx gene targets STK32B and C20orf74 were also found significantly associated with tumor extension (adjusted *p*-value <0.008) and normal liver cirrhosis (adjusted *p*-value: 0.012) respectively. Higher expression of STK32B in tumor than adjacent non-tumor tissues of HCC patients was correlated with longer survival time and local tumor extension, while lower expression of C20orf74 in tumor than adjacent non-tumor tissues of HCC patients was correlated with longer survival time and normal liver cirrhosis. These established associations of STK32B and C20orf74 with longer survival time, local tumor extension and normal liver cirrhosis seemed to be valid, with the reasoning that, HCC patients who survived longer time might be more likely to have local tumor extension and develop normal liver cirrhosis.

c) Higher expression of TTR in tumor over adjacent non-tumor tissues in HCC patients is associated with longer survival time, though statistically not significant with two-sided *p*-value of 0.063.

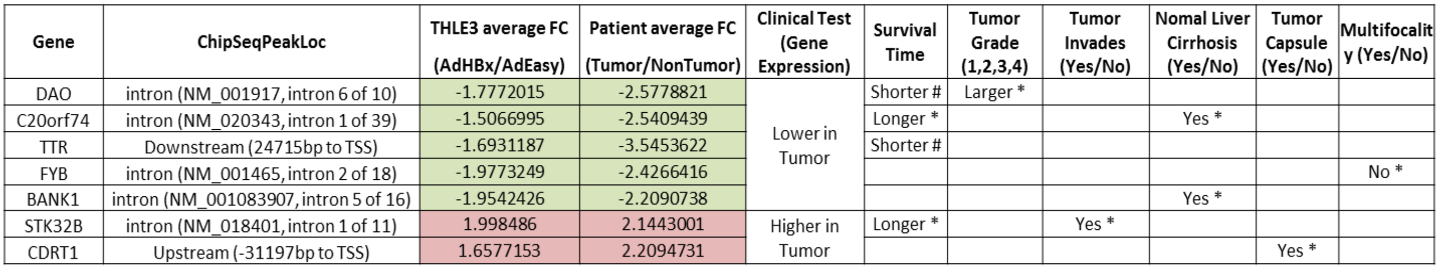
d) In addition, another three potential HBx deregulated gene targets FYB, BANK1 and CDRT1 were found significantly associated with patients' multifocality, normal liver cirrhosis and tumor encapsulation respectively.

- Patients not multifocal were likely to have stronger down-regulation of FYB expression in tumor over adjacent non-tumor tissues than multifocal patients (adjusted *p*-value: 0.044).

- Patients with normal liver cirrhosis were highly likely to have stronger down-regulation of BANK1 expression in tumor over adjacent non-tumor tissues than patients with no normal liver cirrhosis (adjusted *p*-value: 0.012).

- Patients with tumor encapsulation were highly likely to have stronger up-regulation of CDRT1 expression in tumor over adjacent non-tumor tissues than patients without tumor encapsulation (adjusted *p*-value: 0.022).

Table 3.3: Summary of the clinical associations for the seven potential HBx deregulated gene targets. Five out of the seven genes were down-regulated (green) in both THLE3 cells (AdHBx over AdEasy) and HCC patients (tumor over adjacent non-tumor tissues), while the other two were up-regulated (red). Down-regulation of DAO in patients (tumor over adjacent non-tumor tissues) was associated with shorter survival time (statistical not significant) and larger tumor grade. Down-regulation of C20orf74 in patients was related to longer survival time and normal liver cirrhosis. Down-regulation of TTR in patients was associated with shorter survival time (statistical not significant). Down-regulations of FYB and BANK1 in patients were associated with patient non-multifocality and normal liver cirrhosis respectively. Up-regulation of STK32B in patients was related to longer survival time and tumor invasion. Up-regulation of CDRT1 in patients was related to tumor encapsulation.



Gene	ChipSeqPeakLoc	THLE3 average FC (AdHBx/AdEasy)	Patient average FC (Tumor/NonTumor)	Clinical Test (Gene Expression)	Survival Time	Tumor Grade (1,2,3,4)	Tumor Invades (Yes/No)	Nomal Liver Cirrhosis (Yes/No)	Tumor Capsule (Yes/No)	Multifocality (Yes/No)
DAO	intron (NM_001917, intron 6 of 10)	-1.7772015	-2.5778821	Lower in Tumor	Shorter #	Larger *				
C20orf74	intron (NM_020343, intron 1 of 39)	-1.5066995	-2.5409439		Longer *			Yes *		
TTR	Downstream (24715bp to TSS)	-1.6931187	-3.5453622		Shorter #					
FYB	intron (NM_001465, intron 2 of 18)	-1.9773249	-2.4266416							No *
BANK1	intron (NM_001083907, intron 5 of 16)	-1.9542426	-2.2090738					Yes *		
STK32B	intron (NM_018401, intron 1 of 11)	1.998486	2.1443001	Higher in Tumor	Longer *		Yes *			
CDRT1	Upstream (-31197bp to TSS)	1.6577153	2.2094731						Yes *	

*: association observed statistically significant after Bonferroni multiple test correction (adjusted *p*-value < 0.05)
#: association statistically not significant but clear difference were observed (DAO: 0.094; TTR: 0.063)

Other than the four genes found associated with survival time (C20orf74, STK32B, DAO and TTR), the additional three genes associated with categorical clinical features (CDRT1, FYB and BANK1) were also functionally important. CDRT1 acts as protein-ubiquitin ligase; FYB involves in protein amino acid modifications and immune system process; and BANK1 involves also in immune system process. As annotated in Table 3.4, these seven putative HBx gene targets functionally involve in amino acid metabolic process, protein post translational modifications, regulation of signal transduction, protein transport, and immune responses. These seven potential HBx gene targets are likely to have clinical inferences in HCC patients and worth further investigation.

Table 3.4: Functional annotations of the seven clinically associated potential HBx deregulated gene targets. These genes functionally involve in amino acid metabolic processes, protein post translational modifications, regulation of signal transduction, protein transport, and immune responses.

Gene	Gene Name	Molecular Function	Biological Process
DAO	D-amino-acid oxidase	catalytic activity, D-amino-acid oxidase activity, oxidoreductase activity	cellular amino acid metabolic process, cellular nitrogen compound metabolic process, primary metabolic process, oxidation reduction
C20orf74	RALGAPA2 (Ral GTPase activating protein, alpha subunit 2)	GTPase activator activity, enzyme activator activity, nucleoside-triphosphatase regulator activity	regulation of signal transduction, regulation of cell communication, regulation of biological process, regulation of small GTPase mediated signal transduction
STK32B	serine/threonine kinase 32B	nucleotide & nucleoside binding, ion binding, catalytic activity, protein kinase activity, ATP binding, transferase activity, phosphotransferase activity, ribonucleotide binding	protein modification process, protein amino acid phosphorylation, phosphate metabolic process, phosphorylation, biopolymer modification, post-translational protein modification
TTR	transthyretin	receptor binding, transmembrane transporter activity, hormone binding	transport, localization
CDRT1	CMT1A duplicated region transcript 1	zinc ion binding, ion binding, cation binding, metal ion binding	Members of the F-box protein family; act as protein-ubiquitin ligases
FYB	FYN binding protein	receptor binding, protein binding,	immune system process, post-translational protein modification, protein amino acid phosphorylation, protein targeting, protein import into nucleus, phosphorus metabolic process, nucleocytoplasmic transport, signal transduction, protein kinase cascade, protein transport, biopolymer modification, response to stimulus,
BANK1	B-cell scaffold protein with ankyrin repeats 1		cell activation, immune system process,B cell activation

3.2.6 Correlation of Clinically Associated HBx Deregulated Gene Targets with HBx Protein Expression in the 100 HCC Patients

The seven clinically associated potential HBx deregulated gene targets were not only deregulated when HBx is expressed in THLE3 cells, but also deregulated in tumor over adjacent non-tumor tissues of HCC patients. To examine whether there is any direct correlation between these gene targets and HBx protein, linear regression models were built and Pearson correlations were calculated based on these genes and HBx protein expression values in the 100 HCC patients (log2 fold change in tumor over adjacent non-tumor tissues). However, it turned out that none of these seven potential gene targets showed significant correlations with HBx protein expression, and all the correlation R square values were quite small, as seen in the scatter plots of Fig 3.10. This observation of low linear correlation between potential HBx deregulated gene targets and HBx protein expressions in HCC patients may be explained that patients were physiologically very complex and many other factors, such as, medication, diet, other diseases and environmental factors, could also affect the gene expressions in patients. In that case, the deregulation effect of HBx on gene expressions could possibly be masked or disturbed by other factors in patients. As a result, the relationship between gene expressions and HBx expressions in patients was no longer linear. Thus, the low linear correlations could not prove that these genes were completely unrelated to HBx. Nevertheless, these seven potential HBx gene targets with predicted HBx binding sites nearby were shown deregulated in THLE3 cells upon expression of HBx protein, and this could be the evidence supporting the deregulation potential of HBx on these genes expressions.

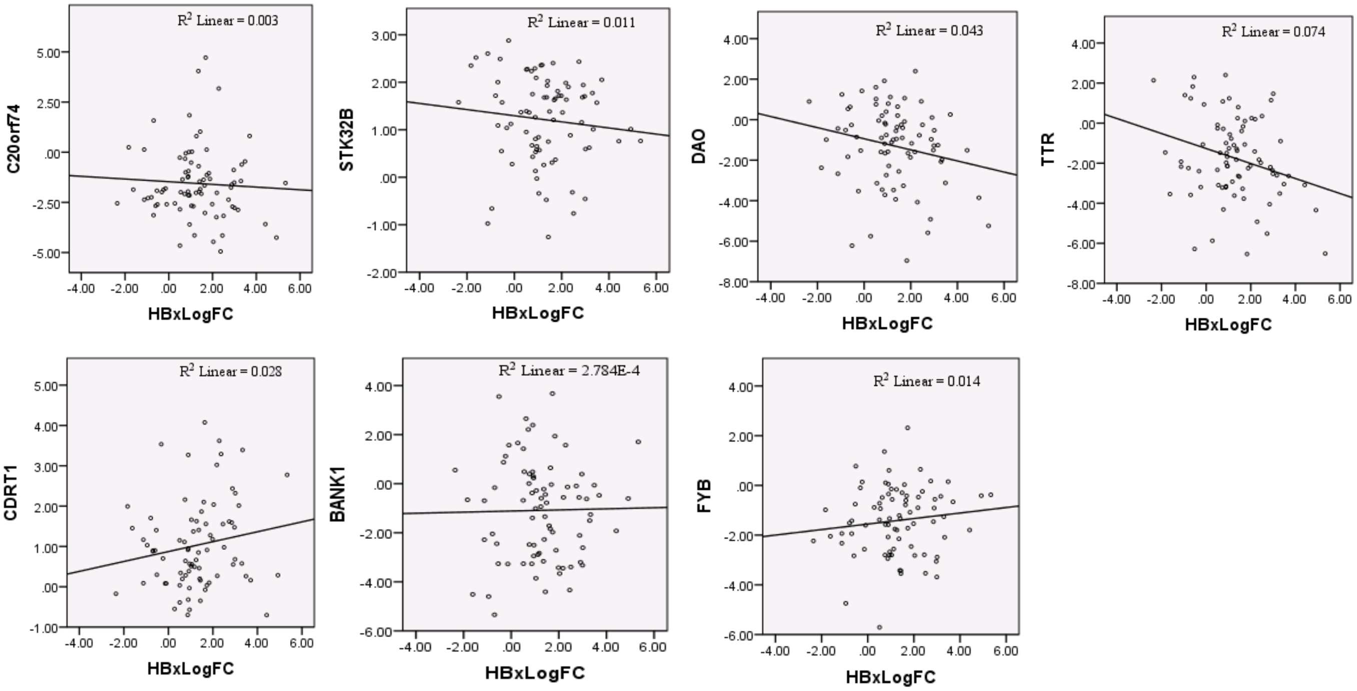


Figure 3.10: Scatter plots for the expressions of the seven potential HBx deregulated gene targets and expressions of HBx protein in 100 HCC patients. The log2 fold change of tumor over adjacent non-tumor tissues on each patient for genes (y-axis) and HBx protein (x-axis) were plotted. The Pearson correlations (R square value) calculated between the genes and HBx were quite low with highest value 0.074. HCC patients are physiologically very complex and there are many other factors that may affect gene expressions. Thus, the deregulation effects of HBx on gene expressions could be masked or disturbed, which may explain the low linear correlations between genes and HBx expressions in HCC patients.

3.3 Discussion and Future Work

HBx was reported not to bind DNA directly, but though interactions with transcription factors. HBx binds to transcription factors and regulate gene by chaning the DNA binding affinities of transcription factors. To unravel the genome-wide binding sites of HBx, our laboratory applied chromatin immunoprecipitation followed by Illumina high-throughput sequencing (ChIP-Seq) with antibodies specifically against HBx protein in primary liver cell line THLE3 cells transfected with HBx-expressing adenoviruses. Our laboratory has also previously utilized chip-based chromatin immunoprecipitation (ChIP-chip) in ultraviolet (UV)-treated primary liver cell line HepG2 cells transfected with HBx-expressing adenoviruses, to profile genomic binding sites of HBx. Compared to the ChIP-chip study, this ChIP-Seq strategy has overcome the limitations of ChIP-chip technology, such as, limited number of probes used in ChIP-chip hybridization array, hybridization noise and dye bias. It has also eliminated the possible artificial effects introduced by UV-treatment on HepG2 cells. In this study, an analysis pipeline was implemented to integrate the ChIP-Seq sequencing data, microarray expression profiles and 100 HBV-associated HCC patients' clinical data. In the end, a list of potential genomic binding sites of HBx was discovered on a single-base resolution from the ChIP-Seq data, and a list of putative HBx deregulated direct gene targets was predicted to have significant clinical inferences in HBV-associated HCC patients.

The millions of ChIP-Seq short sequencing reads of 36bp were first aligned to human genome and those with unique best match to human genome were remained for peak-calling. THLE3 cells transfected with HBx-expressing adenoviruses (AdHBx) were compared against control THLE3 cells (AdEasy), and a total of 2860 potential genomic binding sites of HBx were predicted using peak calling algorithm MACS and CCAT. Most of these 2860 potential HBx binding sites located within gene introns and intergenic regions, and this is not surprising since introns and intergenic regions are the two longest categories in human genome. Next abundant locations of the 2860 potential binding sites are promoters, exons, 5'UTR and 3'UTR in descending order. However, these 2860 predicted HBx binding sites were found to be significantly enriched in exons and promoter regions of genes ($p<0.00001$). Interestingly, among the binding sites within introns and exons, over half were in the first and last introns and exons. This may suggest the potential regulatory effect of HBx on gene expressions.

The genomic sequences of the 2860 potential HBx binding sites were then retrieved for known motif enrichment analysis using HOMER. 195 transcription factor binding motifs were found significantly over-represented within the 2860 potential HBx binding sites. These 195 transcription factors are the potential candidate co-factors that HBx may interact to bind to DNA and deregulate gene expressions in HBx-expressing THLE3 cells. Surprisingly, of the 195 potential HBx-interacting transcription factors, 129 were also found to be significantly over-represented within the HBx binding sites predicted from our previous ChIP-chip data done in our laboratory on UV-treated liver HepG2 cells (Sung et al., 2009) (Supplementary Table S3). This evidence further confirmed the list of potential transcription factors that HBx may interact in cell lines, even though the potential deregulated direct gene targets of HBx in these two liver cell lines did not overlap much which is probably due to their physiological differences and the artificial effects introduced by UV-treatment on HepG2 cells. In addition, these 129 commonly predicted HBx-interacting transcription factors include most of the transcription factors that were previously reported by other studies to interact with HBx or being activated by HBx. This could suggest that our computational workflow is robust to analyse ChIP-Seq data for identification of HBx-interacting transcription factors.

In this study, microarray gene expression profiles in THLE3 cells (AdHBx over AdEasy) were also integrated and a list of 143 potential deregulated direct gene targets of HBx were identified, indicating the pleiotropic nature of HBx: interact with a variety of transcription factors and deregulate a large set of genes. Though the potential HBx deregulated direct gene targets predicted from the two cell lines THLE3 and HepG2 were very different, they were enriched in similar biological pathways with top two being developmental process and multicellular organismal process. Nevertheless, the 143 potential HBx deregulated gene targets predicted from THLE3 ChIP-Seq data were significantly enriched in transcription factor activities, which however are not significant in HepG2 ChIP-chip data. In addition, the potential HBx-interacting transcription factors predicted from ChIP-Seq data include c-Myc and P53 that were previously reported to interact with HBx but not significantly found in ChIP-chip data. These two observations may possibly indicate that ChIP-Seq techniques could help to reduce false negative rates compared to ChIP-chip techniques.

Previous studies on HBx deregulated gene targets were limited to cell lines due to lack of patient data. In this study, to further evaluate whether the putative HBx deregulated gene targets predicted from liver cell lines are truly related to HCC, microarray expression profiles and clinical data of 100 HBV-associated HCC patients were integrated, and statistical tests were performed to identify potential HBx deregulated gene targets that have significant clinical inferences in HCC patients. 18 out of the 143 putative HBx deregulated gene targets were demonstrated to be significantly differentially expressed in tumor over adjacent non-tumor tissues in the

100 HCC patients with the same deregulation direction in THLE3 cells when HBx is expressed (AdHBx over AdEasy) (Fig 3.7). HBx is of oncogenic potential, so these 18 potential HBx deregulated gene targets having the same deregulation direction in tumor over adjacent non-tumor tissues of HCC patients and in THLE3 cells when HBx is present might be potentially associated with HCC. To further confirm the association, statistical tests were performed for each of these 18 genes on each of the clinical features of the 100 HCC patients. It turned out that seven out of the 18 genes had obvious association with patient survival time, tumor grade, tumor invasion, normal liver cirrhosis, tumor encapsulation and multifocality (Fig 3.8, Fig 3.9 and Table 3.3). Particularly, higher expression of STK32B in tumor than adjacent non-tumor tissues of HCC patients was significantly correlated with longer survival time and local tumor extension, while lower expression of C20orf74 in tumor than adjacent non-tumor tissues of HCC patients is significantly correlated with longer survival time and normal liver cirrhosis. Patients with higher DAO and TTR expressions in tumor than adjacent non-tumor tissues appeared clearly to have longer survival time than patients with lower DAO and TTR expressions in tumor than adjacent non-tumor tissues, though not statistically significant. Other three genes BANK1, CDRT1 and FYB were significantly related to normal liver cirrhosis, tumor encapsulation and multifocality respectively. Furthermore, these seven genes were functionally important involving in amino acid metabolic process, protein post-translational modifications, protein transport, regulation of signal transduction, and immune responses (Table 3.4). Thus these seven potential HBx deregulated gene

targets are highly likely to have clinical inferences in HCC patients and worth further investigations.

With the clinically associated potential HBx deregulated direct gene targets, we may wonder whether there is any linear correlation between the gene expression and HBx protein expression. To answer this question, linear regression models were constructed for each of the seven genes with HBx protein expressions in the 100 HCC patients. Pearson correlations and correlation significance were also calculated to evaluate the strength and significance of the linear relationships. Unfortunately, none of the seven genes were significantly correlated with HBx protein expressions in patients, and the correlation R square values were all very low (Fig 3.9). One possible explanation for the low correlations is that HCC patients were physiologically very complex and many other factors could also affect gene expressions in patients, such as, environmental factors, medication, other diseases, emotions, and so on. In that case, unlike in cell line, the HBx protein was no longer the only factor affecting gene expressions in patients, and the deregulation effect of HBx might be masked or disturbed by other factors. Therefore the low linear correlations do not conclude weak relationships between HBx and gene expressions in patients. More importantly, it has been shown that these potential HBx target genes with HBx binding sites nearby were deregulated upon the expression of HBx proteins in THLE3 cells, which supports the deregulation effect of HBx on gene expressions.

In summary, by analysing the THLE3 ChIP-Seq sequencing data and comparing with HepG2 ChIP-chip data, a list of potential global HBx binding sites on a single-base resolution was identified, and a more comprehensive list of potential HBx-interacting transcription factors was confirmed. In this study, the pleiotropic nature of HBx has been further concluded as a transactivator deregulating a large set of genes indirectly through interactions with a variety of transcription factors. To date, the underlying mechanism of how HBx deregulation of host gene expressions through interactions with transcription factors contributes to hepatocarcinogenesis in HCC patients was still unclear due to lack of patient data. In this study, the microarray expression profiles and clinical data of 100 HBV-associated HCC patients were first ever utilized, and seven putative HBx deregulated gene targets were identified to be significantly associated with patients' clinical features including survival profiles. These putative HBx gene targets with significant clinical inferences may potentially involve in HBx-induced hepatocarcinogenesis. However, this study may only identify a small portion of HBx gene targets that potentially play a role in hepatocarcinogenesis. Nevertheless, we gained more knowledge on the potential genomic binding sites of HBx, HBx-interacting transcription factors and putative deregulated direct gene targets of HBx with potential clinical inferences in HCC patients. By identification of clinically associated HBx deregulated direct gene targets, we are now in a better position to explore the roles of the HBx in HBx-induced carcinogenesis. Future work on examining the pathways that HBx may deregulate by targeting these clinically associated genes through interactions with transcription factors is to be done. This may facilitate future discovery of potential drug targets and development of new therapies for HCC patients.

CHAPTER 4: Conclusion and Future Work

Chronic HBV infections have been identified as a major risk factor for HCC accounting for 50-55% of all HCC cases in the world, and may gradually induce the development of HCC in patients (Arbuthnot and Kew, 2001; Bonilla Guerrero and Roberts, 2005; Buendia, 1992; Chang, 2003; Lupberger and Hildt, 2007; Parkin et al., 2001; Robinson, 1994; Tan, 2011). Various mechanisms have been proposed currently by scientists for HBV-associated development of HCC, such as, multi-locus HBV genome integrations into human genome (Bill and Summers, 2004; Bonilla Guerrero and Roberts, 2005; Buendia, 1992; Goto et al., 1993; Jiang et al., 2012; Murakami et al., 2005; Pineau et al., 1998; Robinson, 1994; Saigo et al., 2008; Tan, 2011; Tu et al., 2006), HBx deregulation of host genes expression through interactions with transcription factors (Andrisani and Barnabas, 1999; Ganem, 2001; Sung et al., 2009; Wu et al., 2001) or through epigenetic modifications of genes (e.g. alteration of DNA methylation status of genes) (Arzumanyan et al., 2012; Huang et al., 2010; Jung et al., 2010; Kim et al., 2010; Madzima et al., 2011; Park et al., 2011; Su et al., 2008; Um et al., 2011; Zhu et al., 2010) or through deregulation of regulatory microRNA expressions (Kong et al., 2011; Shan et al., 2011; Wang et al., 2010; Wang et al., 2012; Wu et al., 2011; Yip et al., 2011; Yuan et al., 2012), etc. Understanding these underlying mechanisms for HBV-induced carcinogenesis in HCC patients will help future identification of potential drug targets and development of new therapies for HCC treatment. However, the molecular pathogenesis of HBV-induced hepatocarcinogenesis in HCC patients is still unclear (Ng and Lee, 2011; Tan, 2011). In this project, we focused on two essential underlying mechanisms: HBV genome integrations and HBx deregulation of host genes expression through interactions with transcription factors.

4.1 Characterization of HBV-Host Genome Integration Sites in HCC Patients

To understand HBV genome integration events, we first need to know where the HBV genome is fused with human genome. Previous studies on HBV genome integration sites are mainly PCR-based methods, which are labour-intensive and require prior-knowledge of HBV DNA that are fused with human DNA, which however is very limited currently (Saigo et al., 2008; Tamori et al., 2003; Tamori et al., 2005; Tu et al., 2006; Urashima et al., 1997). To comprehensively characterize HBV genome integration sites with human genome and to study the variation of HBV DNA in HCC patients, our laboratory has applied targeted deep sequencing (454 FLX sequencer) techniques to enrich for HBV-containing DNA fragments extracted from 48 HBV-positive HCC patients' tumor and adjacent non-tumor tissues. In this study, I implemented a computational workflow to analyse the high throughput FLX sequencing data to identify integrated sequences carrying both HBV and human DNA within the same sequence from which the HBV-HG integration sites can be inferred. In the end, a set of 60 novel altered HBV sequences and 63 HBV-HG integrated sequences were successfully identified. Various alteration events such as insertion, deletion, duplication, and inversion were observed from the 60 altered HBV sequences. Novel HBV genome integration boundaries were also inferred from the HBV-HG integrated sequences which carried both HBV and human DNA within the same sequence. Interestingly, it was found that the HBV-HG integrations preferentially occurred on the small HBx gene (27/63=42.9%) and the C-terminal of HBx carrying p53 binding domain was often removed to fuse with human genome. Deletion of p53 binding domain of HBx may potentially promote carcinogenesis, as p53 is a well-known tumor suppressor. The N-terminal two third of HBx gene carrying transactivation domains (amino acid 1 to 100) were often retained in the integrated form, indicating the transactivator nature of HBx. Significantly, our laboratory has successfully experimentally validated the existence of altered HBV sequences and HBV-HG integration sites in HCC patients. These findings concluded the potential important role of HBx in HBV-associated carcinogenesis in HCC patients. By computational scanning of the HBV-HG integrated sequences for open reading frames, it has been observed that HBV-HG integrations may potentially lead to either early termination of HBV genes (e.g. HBx gene) or expression of chimeric transcripts. More significantly, based on my prediction results, our laboratory has experimentally proved the existence of chimeric transcripts *in vivo*, and functional evaluation on the oncogenic potential of these chimeric transcripts is still in progress in the laboratory.

4.2 Future Work on the Computational Analysis Pipeline in Identifying Virus-Host Genome Integration Sites

The major benefit of utilizing single-end high-throughput 454 FLX pyrosequencing is that the reads produced could be long enough to allow identification of precise virus insertion sites in human genome at single-base resolution with one region of the read aligned to human genome and the other region aligned to virus genome. Analysing the millions of long sequencing reads of variable lengths has always been a challenge, which requires intensive computational power to map the reads to reference genome. Even though there are many software tools available specifically designed to map high throughput sequencing reads to reference genome, it is simply far from enough in the case of identifying viral-host integration sites. To extract the most biological meaningful results from high throughput sequencing data generated from specifically designed experiments, a carefully implemented analysis workflow is always needed to best fit the data. When analysing the FLX sequencing data in this study, mapping of the sequencing reads to reference genomes was only the very first step, after which, noise and insignificant reads that do not contain HBV sequences were removed to clean the data. Before proceeding to identify HBV-HG integration sites, a step of *de novo* assembly was incorporated to recover possible integration sites that were disrupted by DNA fragmentation. When identifying HBV-HG integrated sequences, the sequences were searched against both human and HBV genomes, and each sequence must be carefully examined to identify possible integration sites based on its alignments with human and HBV genomes. Sequences containing HBV genome integration sites were hypothesized to have at least one region of the sequence aligned to human genome and the other region aligned to HBV genome. Fortunately, findings in this study have promisingly proved the robustness of my computational approach to analyse the high throughput sequencing data for comprehensive identification of viral-host genome integrations. Future work to compact the analysis workflows as an integrated standalone platform is to be done, such that other researchers can also submit their high throughput sequencing data and perform similar analysis to identify viral-host genome integration sites. Previous studies on HBV genome integration sites were mainly PCR-based methods (Saigo et al., 2008; Tamori et al., 2003; Tamori et al., 2005; Tu et al., 2006; Urashima et al., 1997). This has been among the very first to comprehensively characterize the HBV-HG integration sites in a large series of samples from 48 HCC patients, and we are now in a better position to understand how HBV genome integration may potentially contribute to hepatocarcinogenesis in HCC patients.

4.3 Identification of HBx Genomic Binding Sites, HBx-interacting Transcription Factors, and Clinically Associated Deregulated Direct Gene Targets of HBx

Other than HBV genome integration into human genome after long term HBV infection in patients, HBx deregulation of host gene expression through interactions with transcription factors was reported also to potentially contribute to hepatocarcinogenesis. HBx is a small protein of length 154 amino acids which is reported to have oncogenic potential. HBx protein has been implicated to play an important role in HBV-induced development of HCC. HBx does not have a DNA-binding domain, and is found to bind to DNA indirectly through interactions with transcription factors. Interaction of HBx with transcription factors may change the DNA binding affinities of transcription factors and consequently lead to regulation of host gene expression. Our laboratory has previously systematically profiled the HBx genomic binding sites, HBx-interacting transcription factors and HBx deregulated direct genes in a large-scale using ChIP-chip method in UV-treated liver cell line (HepG2) on 1.5kb promoter regions of human genes (Sung et al., 2009). However, the detailed mechanism of how HBx deregulation of host gene expressions through interactions with transcription factors may contribute to hepatocarcinogenesis is still incompletely understood (Ng and Lee, 2011; Sung et al., 2009).

There are various limitations and bias associated with ChIP-chip method, such as limited genome coverage (e.g. 1.5kb promoter regions of genes), hybridization noise, dye bias, and low reproducibility. To overcome the limitations associated with ChIP-chip and to obtain a more comprehensive and unbiased list of HBx genomic binding sites on single-base resolution, our laboratory has turned to apply chromatin immunoprecipitation coupled with high throughput sequencing technology (ChIP-Seq) to sequence immunoprecipitated DNA fragments bound by HBx in primary liver cell line THLE3 cells transfected with HBx-expressing adenoviruses. In this study, I implemented an analysis pipeline to integrate and analyse the ChIP-Seq sequencing data, microarray expression profiles for both THLE3 cells and 100 HCC patients, as well as the clinical data for the 100 HCC patients. In the end, a list of 2860 potential HBx binding sites, a list of 195 potential HBx-interacting transcription factors, and a list of 143 potential HBx deregulated direct gene targets were identified in THLE3 cells. Among these 143 potential HBx deregulated direct gene targets, seven were found also deregulated with significant clinical inferences in 100 HCC patients. These seven genes were associated with various patients' clinical features including survival time, tumor grade, liver invasions, patients' multifocality, normal liver cirrhosis, and tumor encapsulations.

The 2860 potential HBx genomic binding sites were mostly located within introns and intergenic regions. It is not surprising since introns and intergenic regions are the two longest categories in human genome. Next abundant locations of the potential HBx binding sites are promoters, exons, 5'UTR and 3'UTR in descending order. However, these 2860 predicted HBx binding sites were found to be significantly enriched in exons and promoter regions of genes, and significantly less located in intergenic regions ($p<0.00001$). Interestingly, among the binding sites within introns and exons, over half were in the first and last introns and exons, suggesting the potential regulatory effect of HBx on gene expressions. Among the 195 transcription factors significantly over-represented within the 2860 potential HBx binding sites identified from the THLE3 ChIP-Seq data in this project, 129 were also found significantly over-represented within the 971 potential HBx binding sites identified from the UV-treated HepG2 ChIP-chip data done previous in our laboratory. Though THLE3 ChIP-Seq and HepG2 ChIP-chip data predicted similar sets of potential HBx-interacting transcription factors with 129 motifs in common, the two datasets had very different sets of potential HBx deregulated gene targets, which might be due to the physiological differences of the two different cell lines and the artificial effects introduced by UV-treatment on HepG2 cells. Nevertheless, this has confirmed a list of 129 transcription factors that may potentially interact with HBx and bind to DNA for gene regulations. These 129 transcription factors include previously reported ones that either interact with HBx or are activated by HBx, such as, SP1 (Lee et al., 1998), AP1 (Benn et al., 1996), AP2 (Kim and Rho, 2002), E2F (Weinmann et al., 2001), E2F1 (Choi et al., 2002; Sung et al., 2009), CREB (Maguire et al., 1991), SMAD4 (Sung et al., 2009), YY1 (Sung et al., 2009), NFKAPPAB50 (Su and Schneider, 1996), STAT3 (Waris et al., 2001), and so on. The 195 enriched transcription factors predicted from THLE3 ChIP-Seq data also include C-Myc (Li et al., 2003; Zeller et al., 2006) and P53 (Wei et al., 2006), that were previously reported to interact with HBx but not predicted from the HepG2 ChIP-chip data. This may suggest ChIP-Seq is advantageous over ChIP-chip in reducing false negative rates, since the ChIP-chip data only covered 1.5kb promoter regions of human genes while ChIP-Seq was able to detect the binding sites over entire genome.

Integration of microarray expression profiles in THLE3 cells (AdHBx over AdEasy) for the corresponding nearest genes of the 2860 potential HBx binding sites identified 143 potential HBx deregulated direct gene targets in THLE3 cells. These 143 potential gene targets were significantly enriched in developmental process and multicellular organismal process, and these two biological processes were also found significantly enriched by the 184 potential gene targets of HBx predicted from the HepG2 ChIP-chip data (Sung et al., 2009). The top significantly enriched molecular function of the 143 potential HBx gene targets from THLE3 ChIP-Seq data is transcriptional factor activity, which, however, was not significantly enriched in the 184 potential HBx gene targets from HepG2

ChIP-chip data. This may further confirm the advantage of ChIP-Seq over ChIP-chip in reducing false positive rates.

Further integration of microarray expression profiles for the 100 HBV-associated HCC patients (tumor over adjacent non-tumor tissues) found 18 out of the 143 potential HBx direct gene targets also displaying differential expressions in HCC patients with consistent deregulation directions in THLE3 cells. To evaluate whether these 18 genes were truly related to HCC, statistical tests were performed on the 100 HCC patients' clinical data, and seven out of the 18 genes were found to have significant clinical inferences. These seven potential HBx deregulated gene targets (DAO, C20orf74, TTR, STK32B, FYB, BANK1 and CDRT1) are associated with various patients' clinical features including survival time, tumor grade, liver invasions, patients' multifocality, normal liver cirrhosis, and tumor encapsulations. These seven clinically associated target genes were also functionally important involving in amino acid metabolic process, protein post-translational modifications, protein transport, regulation of signal transduction, and immune responses. However, these seven genes were found to have very low linear correlations with HBx protein expressions in the 100 HCC patients, which could be explained by the physiological complexities of HCC patients. Other than HBx, there are many other factors affecting gene expressions in HCC patients, such as environmental factors, medications, diets, other diseases, emotions and so on. Therefore, the low linear correlations cannot exclude the relatedness between genes and HBx expressions in HCC patients. Thus, these seven potential gene targets of HBx with clinical inferences in HCC patients are worth further investigations.

4.4 Future Work on the Clinically Associated Gene Targets of HBx

Future work should examine the pathways that HBx may deregulate. By targeting these clinically associated genes, it may help in the future discovery of potential drug targets and development of new therapies for HCC treatment. ChIP-Seq techniques have been widely utilized by researchers to comprehensively profile the genomic binding sites of proteins of interest. By doing motif enrichment analysis, co-factors that potentially interact with the proteins of interest when binding to DNA could also be identified. Integration of the microarray expression profiles enables identification of potential deregulated direct gene targets for the proteins of interest. Previous studies profiling HBx genomic binding sites are mainly in cell lines due to lack of patients data. In this study, we have utilized ChIP-Seq techniques, and successfully predicted HBx genomic binding sites, HBx-interacting transcription factors and HBx deregulated direct gene targets in primary liver cell line THLE3 cells. With the availability of 100 HCC patients' microarray expression profiles and clinical data, we are the very first to integrate patient data and identified seven HBx gene targets with significant clinical inferences in HCC patients. Novel analysis pipeline integrating the cell line ChIP-Seq sequencing data, microarray expression profiles for both cell lines and HCC patients, and HCC patient clinical data was successfully implemented in this study. Consistency between the THLE3 ChIP-Seq data and HepG2 ChIP-chip data in terms of HBx-interacting transcription factors has proved the robustness of the analysis pipeline. More studies on the functions and pathways for the seven clinically associated deregulated gene targets of HBx need to be done, to identify the potential pathways that HBx may deregulate. In short, by identification of deregulated direct gene targets of HBx with significant clinical inferences in HCC patients, we are now better positioned to explore the roles of HBx contributing to the hepatocarcinogenesis in HCC patients.

4.5 Conclusion

In conclusion, this project has focused on two underlying mechanisms both of which may contribute to HBV-induced hepatocarcinogenesis in HCC patients: HBV multi-locus genome integrations into the human genome and HBx deregulation of host gene expressions through interactions with transcription factors. In this study, computational analysis pipelines have been implemented and successfully analysed the high throughput FLX pyrosequencing data, ChIP-Seq Illumina sequencing data, microarray expression profiles and HCC patient clinical data. Our findings provided comprehensive characterization of the HBV genome integration sites with human genome and preliminary identification of putative HBx deregulated direct gene targets with significant clinical inferences in HCC patients. We are now better positioned to explore the underlying mechanisms of HBV-induced hepatocarcinogenesis in HCC patients, which may potentially facilitate future discovery of drug targets and new therapies for HCC patients.

CHAPTER 5: Supplementary Tables

Supplementary Table S1: Number of sequences (assembled contigs and unassembled reads) that were classified into the five major groups in each patient sample. In total, our laboratory obtained 378 sequences including 221 "intact HBV", 60 "modified HBV", 34 "HBV+unknown", 56 "HBV-HG junctions", and 7 "modified HBV-HG junctions".

Patient	Intact HBV			modified HBV			HBV+Unknown			HBV-HG Junction			modified HBV-HG			Total
	Contig	Unasm Read	Total	Contig	Unasm Read	Total	Contig	Unasm Read	Total	Contig	Unasm Read	Total	Contig	Unasm Read	Total	
p2N		1	1													1
p3N	1	1	2													2
p4N	4	1	5													5
p4T	5	3	8													8
p5N	3		3	2	1	3	1		1	2	1	3				10
p5T	2		2	1	1	2		1	1		1	1				6
p6T	1	1	2								1	1				3
p7N		1	1								1	1				2
p8N	4		4					1	1	2	1	3				8
p8T	1	2	3							1		1				4
p9N		1	1	1	1	2		1	1							4
p9T		1	1													1
p10N	13	5	18	14	4	18	4	2	6	2	7	9				51
p10T	6	1	7	8		8		1	1	4	2	6	3	1	4	26
p11T	1	4	5													5
p13T	1	1	2	1		1					1	1				4
p14N	3		3													3
p14T		1	1													1
p15N		1	1													1
p15T	3	1	4								1	1				5
p16N	2	1	3		1	1		1	1		1	1				6
p16T	2		2													2
p17N	1	3	4													4
p18N	1	1	2													2
p18T	2		2	1		1										3
p19N	2	1	3				1	3	4	3	1	4				11
p19T	3	3	6							1		1				7
p20N	5	4	9													9
p20T	5	4	9	3	2	5	2		2	3		3				19
p21N	1	1	2													2
p21T	2	3	5													5
p22N		2	2													2
p22T	1		1		1	1		1	1				1		1	4
p23N	2	1	3								1	1				4
p23T	1	4	5	2	1	3	1		1							9
p24T	1		1							1		1				2
p26N	1		1													1
p28N	2	1	3							2		2				5
p28T	2	2	4							1		1				5
p29N	5	4	9		1	1										10
p29T	4	1	5							1		1				6
p30T		2	2								1	1				3
p31T	1	1	2								1	1				3
p32N	3	2	5													5
p32T	1		1													1
p33T	1	1	2													2
p34T	1	2	3				1		1							4
p35T	1		1													1
p36N		1	1													1
p36T	2	1	3	2		2							1		1	6
p37N	1	5	6					1	1		1	1				8
p37T	3		3							1	1	2				5
p38T	2	2	4													4
p39N		3	3													3
p39T	1		1	2		2	1		1							4
p40N	1	1	2				1	1	2	1	1	2				6
p40T	1	3	4				1		1							5
p42N	1	1	2													2
p42T	2	1	3													3
p43T		1	1													1
p44N		1	1													1
p44T		2	2					1	1							3
p45N	3	1	4	7		7	4	3	7	4	2	6	1		1	25
p45T	1	1	2	1		1										3
p46N	1	1	2		1	1										3
p46T	3	2	5		1	1										6
p48N		1	1								1	1				2
Total	123	98	221	45	15	60	17	17	34	29	27	56	6	1	7	378

Supplementary Table S2: Information for 56 HBV-HG junctions and seven modified HBV-HG junctions predicted in different patient samples. The 3rd and 4th column displays the positions of the sequence matched to HBV genome and human genome respectively, with the number in red color highlighting the junction point and the plus/minus sign in bracket representing the strand of HBV or human genome that the sequence matched to. The 5th column shows the type of junction patterns for each sequence (See Fig 2.3). The 6th and 7th columns specify the HBV genes and human genes where the junction points reside. There are four major HBV genes: polymerase, precore, HBx and S (surface protein). Out of the 63 junctions, 27 had the junction points on HBx gene. The 8th and 9th columns show the junctions that have been validated and the chimeric transcripts that have been detected expressed in patient samples. The 10th column displays the expression fold change in patient tumor over non-tumor tissues for the nearest genes of the junctions. Positive fold change represents up-regulation of genes in tumor over non-tumor tissues in patient and negative value represent down-regulation. Genes highlighted are the ones differentially expressed in patient tumor over non-tumor tissues with at least 2 fold change: red for up-regulation and yellow for down-regulation. The last column shows the molecular functions and pathways involved for the genes differentially expressed.

Patient	HBV-Human Junction (56) & modified HBV-Human Junction (7)									
	Total	HBV part	HG Part	Junction Pattern	HBVJunctionPoint	HGJunctionPoint	Junction Validated	Chimeric Transcript	GeneExpression: fold change Tumor/NonTumor	GeneFunction
p5N	3	1966-2059(+)	NT_005403.17:Chr2:20541731-20542059 (+)	1	precore	BBS5-Promoter(-3.4kb)			BBS5: -1.164	
		1477-1975(+)	NT_010498.1:Chr16:2993363- 2993419 (-)	2	precore	29kb upstream to C16orf78			C16orf78: -1.262	
		1853-2206(-)	NT_030059.13:Chr10:10089307-10089419 (+)	1	precore	500kb upstream to Loc100128586			pseudogene	
p5T	1	1820-2095(+)	NT_006576.16:Chr5:792307-792457 (-)	1	HBx&precore	ZDHHC11-intron 11	yes		ZDHHC11: -1.333	
p6T	1	1495-1790(+)	NT_016354.19:Chr4:60468621- 60468770 (-)	2	HBx	30kb upstream to LOC389223	yes	Expressed	pseudogene	
p7N	1	2295-2403(+)	NT_005612.16:Chr3:76198699-76198537 (-)	1	precore	SEC62-Intron 5	yes		SEC62	
p8N	3	1878-2660 (-)	NT_167187.1:Chr8:20552236-20552413 (+)	2	polymerase	50kb downstream to NRG1			NRG1: 1.161	
		1678-1836(-)	NT_167214.1: unplacedHG:128734-128832 (+)	2	HBx&precore	11kb downstream to LOC100286895			pseudogene	
		1590-1804(-)	NT_010783.15:Chr17:17240216-17240371(+)	2	HBx	60kb upstream to Loc100128713	yes		pseudogene	
p8T	1	1302-1780(-)	NT_008046.16:Chr8: 10474414-10474450 (-)	2	HBx	GDF6-intron 1	yes	Expressed	GDF6: -8.739	growth differentiation factor (formation of bones, joints, skull & axial skeleton)

p10N	9	133-331(-)*	NT_009714.17:Chr12:15724057-15724089 (+);
			NT_009714.17:Chr12:15723951-15724001(-)
				1	polymerase&S	120kb donstream to ETNK1			ETNK1: 1.119	

		1682-1795(+)	NT_011109.16:Chr19:20259097 -20259209 (-)	2	HBx	KPTN-Promoter (-4kb)	yes		KPTN: 2.146	actin binding protein (cell motion) (filament organization)
		111-1172(-)	NT_011362.10:Chr20:25100849- 25100967 (-)	1	polymerase	30kb upstream to C20orf108			C20orf108: -1.107	
		274-449 (+)	NT_007592.15:Chr6:55255483-55255635 (-)	1	polymerase&S	HMGCLL1-intron6			HMGCLL1: -2.139	hydroxymethylglutaryl-CoA lyase activity;metal ion binding (metabolic
		788-911(+)	NT_005403.17:Chr2: 63523480 -63523666 (-)	1	polymerase&S	ERBB4-intron1			ERBB4: 1.039	
		1104-1325(-)	NT_010194.17:Chr15:2204668-2204696 (-)	1	polymerase	20kb upstream to TRPM1			TRPM1: -1.491	
		1720-1806(-)	NT_026437.12:Chr14: 42942788 -42942818 (-)	2	HBx	PRKCH-intron9			PRKCH: -2.063	protein kinase C (protein modification & phosphorylation)
		1666-1811(-)	NT_024524.14:Chr13: 16958192-16958260 (-)	2	HBx	NBEA-intron	yes		NBEA: 1.624	
		967-1122(+)	NT_022135.16:Chr2:1113100-1113176 (-)	1	polymerase	LOC100507581-promoter(-4kb)			hypothetical RNA gene	
p10T	6	36- 210(+)	NT_024862.14:Chr17:582722- 582856 (-)	2	polymerase&S	44kb downstream to LOC100129683	yes		pseudogene	
		1631-1820(-)	NT_006576.16:Chr5:1287850-1287993 (+)	2	HBx	TERT-promoter(-2688bp)	yes		TERT: 1.648	
		569- 1192(+)	NT_030059.13:Chr10: 44588143 -44588334	2	polymerase	BTAF1-Intron			BTAF1: -1.256	
		1268-1316(-)	NT_167206.1:ChrY:327776-327740 (-)	2	polymerase	LOC100506481-promoter(-11bp)			pseudogene	
		1014-1316(-)	NT_167206.1:ChrY:327749- 327805 (+)	2	polymerase	LOC100506481-promoter(-20bp)			pseudogene	
		581-954(+)	NT_006576.16:Chr5:1446416-1446477 (+)	2	polymerase	11kb upstream to SLC6A3	yes		SLC6A3: -1.089	
	4	2-39(-), 3018-3215(-), 1631-1820(+)	NT_011651.17:ChrX:28754024-28754062 (-)	2	HBx&precore	5kb downstream to MUM1L1			MUM1L1: -28.148	melanoma associated antigen (mutated) 1-like 1
		1080-1316(-), 1608-2115(+)	NT_167206.1:ChrY:327761-327805 (+)	2	polymerase	LOC100506481-promoter(-15bp)			pseudogene	
		1755-2613(-), 1818-2007(+), 2068-3215(+), 2- 1316(+)	NT_167206.1:ChrY:327746- 327805 (-)	2	polymerase	LOC100506481-promoter(-32bp)			pseudogene	
		2220- 2337 (-), 1968-2122(-)	NT_010783.15:Chr17:2948900- 2948982 (+)	2	precore&polymerase	CDK12-Intron10			CDK12: -1.049	
p13T	1	688-794(-)	NT_011109.16:Chr19:10488025-10488401 (+)	1	polymerase&S	9kb upstream to ZNF607			ZNF607: 1.703	
p15T	1	536- 956(-)	NT_010718.16:Chr17:11715111-11715178 (-)	2	polymerase	65kb downstream to MAP2K4			MAP2K4: -7.013	mitogen-activated protein kinase kinase (MAPKKK cascade,protein modification)
p16N	1	618-822(-)	NT_026437.12:Chr14: 15744062-15744190 (+)	1	polymerase&S	47kb downstream to Loc100128921			pseudogene	
p19N	4	401-561(-)	NT_029419.12:Chr12:23883663- 23883778 (-)	1	polymerase&S	113kb downstream to PGBD3P1			PGBD3P1	
		1822-2215(-)	NT_167197.1:ChrX:27303621-27303669 (-)	1	HBx&precore	IL1RAPL1-inton 5			IL1RAPL1: -1.006	
		395-737(-)	NT_010966.14:Chr18:18433732- 18433844 (-)	1	polymerase&S	LOC647946 (non-coding RNA)			non-coding RNA	
		392 -569(+)	NT_167190.1:Chr12:2270409 -2270226 (-)	1	polymerase&S	3kb downstream to LRRC55; 36kb upstream to APLNR			LRRC55: -1.079; APLNR: 1.736	
p19T	1	1522-1802(-)	NT_030059.13:Chr10: 68127017-68127161 (-)	2	HBx	ATRNL1-intron 26	yes	Expressed	ATRNL1: 14.765	attractin-like 1 (cell surface receptor linked signal transduction)
p20T	3	1628-1819(-)	NT_008413.18:Chr9: 33938611-33938663 (-)	2	HBx&precore	UBAP2-intron12			UBAP2:-1.128	
		937- 1792(+)	NT_010718.16:Chr17:8745855-8745951 (+)	2	HBx	NTN1-LastIntron	yes		NTN1:7.036	netrin 1 (axon guidance and cell migration) (variation of netrin may involve in cancer
		1481-1741(-)**	NT_030059.13:Chr10:16274412-16274458 (+); NT_030059.13:Chr10: 2366039 -2366072(+)	1&2	HBx	85kb downstream to REEP3; NCOA4-promoter(-3kb)			REEP3:-1.295, NCOA4:1.115	
p22T	1	141-1819(-), 2974-3092(+)	NT_167187.1:Chr8:30951033-30951157(+)	2	HBx&precore	50kb upstream to POTEA; 26kb downstream to HGSNAT	yes	Expressed	POTEA,HGSNAT:1.475	
p23N	1	1837-2014(+)	NT_033899.8:Chr11:7899487- 7899717 (+)	1	HBx&precore	300kb upstream to PDGFD	yes		PDGFD	
p24T	1	1472-1647(-)	NT_167190.1:Chr11:12563743-12563941 (-)	2	HBx	AIP-intron 5	yes	Expressed	AIP:1.911	
p28N	2	962- 1663(-)	NT_007933.15:Chr7:41050503-41050539 (-)	2	HBx	SLC26A5-intron	yes	Expressed	SLC26A5	
		2143-2516(+)	NT_010718.16:Chr17:13143919-13144021(-)	1	precore	35kb upstream to HS3ST3A1	yes in T&N		HS3ST3A1:1.588	
p28T	1	446- 1663(-)	NT_007933.15:Chr7:41050503-41050619 (-)	2	HBx	SLC26A5-intron	yes	Expressed	SLC26A5	
p29T	1	1883-2078(+)	NT_025741.15:Chr6:28145055-28145220 (-)	1	precore	18kb upstream to TRDN	yes		TRDN:1.439	
p30T	1	1820-1897(+)	NT_010783.15:Chr17:4893254 -4893516 (-)	1	HBx&precore	KRT32-intron6 & Exon6			KRT32:41.256	
p31T	1	1575-1761(+)	NT_008183.19:Chr8: 32531628 -32531673 (+)	1	polymerase&HBx	9kb downstream to HEV1	yes	Expressed	HEV1	
p36T	1	1093-1152(-), 874-1826(-)	NT_008413.18:Chr9: 25075832- 25076007 (-)	2	polymerase	591kb upstream to TUSC1	yes	Expressed	TUSC1:-1.428	
p37N	1	1417-1678(-)	NT_011109.16:Chr19:10829043 -10829084 (-)	2	HBx	SIPA1L3-intron2			SIPA1L3:-1.649	
p37T	2	2006 -2578(-)	NT_030059.13:Chr10: 41254724-41255047 (+)	1	precore	12kb downstream to LIPF	yes		LIPF:-1.234	
		1442-1796(+)	NT_030059.13:Chr10:41577178-41577115(-)	2	HBx	FAS transcript variant 1-Intron 7	yes	Expressed	FAS:-2.847	TNF receptor superfamily (regulate programmed cell death & apoptosis) (transduce proliferation in normal cells)
p40N	2	1952-2307(-)	NT_029490.4:Chr21:370831- 370944 (-)	1	precore	10kb upstream to BAGE2&BAGE3			BAGE2, BAGE3	
		108-400(+)	NT_005612.16:Chr3:89972106- 89972186 (+)	1	polymerase&S	YEATS2-Intron 13			YEATS2:1.266	
p45N	6	1821-2552(+)	NT_029289.11:Chr5:5912940- 5913134 (+)	1	HBx&precore	140kb downstream to ASSP10			pseudogene	
		1952-2041(+)	NT_025028.14:Chr18: 10744807-10745097 (-)	1	precore	500kb upstream to CDH7			CDH7:1.011	
		394 -490(-)	NT_006713.15:Chr5:15433615-15433668 (-)	1	polymerase&S	CENPK-Intron 5			CENPK:5.836	centromere protein (mitotic cell cycle) (regulation of transcription from RNA polymerase II promoter)
		1546-1600(-)	NT_008470.19:Chr9:33306748- 33306795 (-)	1	polymerase&HBx	BAAT-intron 1			BAAT:-2.439	liver amino acid N-acyltransferase (lipid metabolic process) (bile acid metabolic & transport process)
		2574-2710(-)	NT_022459.15:Chr3: 4187591-4187654 (-)	2	polymerase	193kb upstream to LOC100128160			hypothetical	
		1819-1934(+)	NT_030059.13:Chr10: 58098120 -58098199 (-)	1	HBx&precore	268kb downstream to SORCS3			SORCS3:3.269	sortilin-related VPS10 domain containing receptor 3 (cell surface receptor linked signal transduction) (strongly expressed in the central nervous system)
	1	3010-3215(+), 53-366(+)	NT_011520.12:Chr22:5748454-5748491 (-)	1	polymerase&S	MYO18B-Intron 39			MYO18B:9.699	myosin 18B (regulate muscle-specific genes & intracellular trafficking) (Mutations associated with lung cancer)
p48N	1	152-382(+)	NW_927993.1:unplacedHG:1-35 (-)	1	polymerase&S	unplaced not annotated			unannotated	
Total	63									

Supplementary Table S3: List of the 195 enriched transcription factors from ChIP-Seq data in THLE3 cells, among which 129 were common with the transcription factors predicted from ChIP-chip data in HepG2 cells. The ranking of the enriched transcriptional factors was based on significance *p*-values from motif enrichment analysis.

Motif Name	Consensus	ChipSeq *P*-value	ChipSeq Rank	ChipChip Rank	ChipChip *p*-value	Reference
HEB	GCCAGCTG	1.68E-55	1	39	1.4E-31	-
MYOD	CNGNNNCAGGTGNCGNAG	3.96E-54	2	53	1.1E-23	(Barnabas et al., 1997)
MINI20	NNCCGGCCCCACGCAGGNGCA	7.01E-53	3	47	7.7E-26	-
AP4	CTCAGCTGGT	1.03E-49	4	37	2.7E-33	-
LRH1	CNGACCTTGNAC	2.38E-47	5			
E12	GGCAGGTGNCG	1.18E-46	6	73	3.3E-17	-
E47	NCNGCAGGTGTNCNC	1.03E-41	7	35	5E-36	-
AP2	GCCCCCAGGCGGNGNN	3.81E-39	8	4	2E-198	(Kim and Rho, 2002)
HIC1	NCCGGGTGCCCGGGG	1.73E-37	9			
AP2ALPHA	GCCNNNGGG	2.77E-37	10	7	3E-159	-
USF	CCACGTGN	2.77E-37	11	63	7E-21	-
VMAF	ANATGCTGACTCAGCACNN	4.50E-37	12	110	2.4E-05	-
HEN1	NTGGGNCNCAGCTGCGNCCCNN	1.10E-36	13	33	2.9E-38	-
ZF5	NGGGGGCGCGCTT	4.45E-34	14	14	9.4E-76	-
DR1	GGGNCAAAGGTCA	4.45E-34	15	113	6.4E-05	-
LBP1	CAGCTGC	3.74E-32	16	34	1.5E-36	-
MUSCLE	NNCCGCCNCCACCCCGGNGCC	3.19E-31	17	11	4.3E-86	-
SMAD4	GTGGGGCAGCCANCT	4.85E-31	18	30	1.9E-43	-
MZF1	AGTGGGGA	7.51E-31	19	80	3.4E-15	-
AREB6	CTGCACCTGTGC	2.68E-30	20	142	0.0321	-
MYOGENIN	GGCAGCTG	4.14E-30	21	38	2.9E-32	-
E2A	CACCTGNC	6.26E-30	22	69	5.6E-18	-
COUP	TGACCTTTGACCC	2.72E-29	23	121	0.00149	-
TAL1	TCCAGCTGCT	1.57E-28	24	107	1.1E-05	-
LMO2COM	CNNCAGGTGCNG	1.72E-28	25	60	1.4E-21	-
LMAF	GGTCAGCAG	2.56E-28	26			
HNF4	NGGNCA	1.28E-27	27	119	0.00057	-
MAZR	NGGGGGGGGGCCA	4.24E-27	28	25	2.4E-50	-
CBF	NNNNCTGCGGTTANNN	2.06E-26	29			
CP2	GCNCNACCCAG	4.51E-26	30	48	3.9E-25	-
CACCCBINDINGFACTOR	CANCCCNTGGGTGTGG	8.37E-26	31	29	6.5E-44	-
AP2GAMMA	GCCCNNGGG	3.05E-25	32	6	3E-167	-
HES1	ANGNCTCGTGGCNNG	3.05E-25	33	36	8E-35	-
ER	CAGGTCACGGT	6.61E-25	34			
ETS1	ACAGGAAGTGNNTGC	6.26E-24	35			
ZIC1	TGGGTGGTC	1.31E-23	36	54	3.3E-23	-
ATF3	CTCTGACGTCANCG	4.22E-23	37	88	4.6E-11	(Barnabas et al., 1997)
MYC	CACGTGN	1.69E-22	38	66	1E-18	-
PAX4	NANNCCCACCCN	2.39E-22	39	40	3.9E-31	-
MEIS1	NNNTGACAGGNC	3.44E-22	40			
AP2REP	CAGTGGG	4.86E-22	41	70	6.1E-18	-
AML1	TGTGGT	4.86E-22	42			
TAL1ALPHAE47	NCGAACAGATGGTNNN	9.91E-22	43	134	0.01206	-
AR	TGAGCACGN	1.42E-21	44	122	0.00181	(Zheng et al., 2007)
PAX5	TCGAGGCGCANTGATGCGTAGCCGCCCC	5.29E-21	45	23	2E-54	-
EBOX	CCACGTGNCN	1.13E-20	46	67	1.6E-18	-
TBX5	TNAGGTGTTA	1.13E-20	47			
MINI19	NNCNGNCNCCACNCAGGNGCC	3.15E-20	48	12	3.7E-82	-
LRF	NGGGCCCCC	3.15E-20	49			
TFE	TCATGTGN	3.15E-20	50			
NMYC	TNCCACGTGNCN	4.43E-20	51	72	2.2E-17	-
MOVOB	GNGGGGG	8.65E-20	52			
NRSF	GCGCTGTCCGTGGTGCTGA	2.12E-19	53	19	5.3E-58	-
KAISO	NTCCTGCTAN	2.33E-19	54			
CLOCKBMAL	ACACGTGG	2.49E-19	55			
ETS2	GACAGGAAGTANTT	9.23E-19	56			
PPAR	TGACCTTTGNCCC	1.20E-18	57	116	0.00014	-
GR	TNTGTTCT	1.20E-18	58			
ZNF219	CGCCCCCCNCCC	3.13E-18	59			
NERF	TGNCAGGAAGTAGGTNNC	5.55E-18	60	71	1.9E-17	-

HIF1	GCGTACGTGCGGNN	5.96E-18	61	15	1.3E-69	(Moon et al., 2004)
NANOG	GGGNCCATTTCC	1.92E-17	62			
MTF1	TNTGCACACGGCCC	9.71E-17	63	45	1.1E-26	-
R	NTGGCCGCGNANCGTGGTGCA	2.73E-16	64	51	7.2E-24	-
YY1	NCNCGGCCATCTTGNCTGNT	4.15E-16	65	106	1.1E-05	-
RFX	CTGTTGCCA	4.40E-16	66	137	0.01492	-
ATF	CNCTGACGTCNNCC	8.11E-16	67	62	2.5E-21	(Maguire et al., 1991)
EGR2	NTGCGTGGGCGT	1.98E-15	68			
KROX	CCCGCCCCCGCCCC	2.61E-15	69	5	8E-197	-
SREBP	GNNATCACCCCA	4.68E-15	70			
NRF2	ACCGGAAGAG	1.09E-14	71	46	2.5E-26	-
ZIC3	TGGGTGGTC	1.44E-14	72	96	2.7E-08	-
STRA13	NNGTCACGTGANNN	1.96E-14	73	101	4.6E-07	-
TAL1BETAITF2	GNNAACAGATGGTNTN	2.42E-14	74			
EGR	GTGGGGGCGAC	2.53E-14	75	10	9E-125	(Yoo and Lee, 2004)
CACD	CCACACCC	2.53E-14	76			
TAL1BETAE47	NNGAACAGATGGTCNN	2.53E-14	77			
WT1	CCCNCCCNC	2.53E-14	78			
NFE2	TGCTGAGTCAC	3.21E-14	79			
CREB	CGTCAN	3.35E-14	80	21	1.3E-57	(Maguire et al., 1991)
CMYB	CCNAANGGCNGTTGGGGG	1.01E-13	81	55	5.1E-23	-
PPARA	TNGGGTCATTGGGGTCANG	1.05E-13	82	82	6E-13	(Kim et al., 2007)
P53	AGACATGCCT	1.33E-13	83			
GABP	ACCGGAAGTGCA	1.71E-13	84			
CETS1P54	ACCGGAAGTN	2.94E-13	85	89	8.5E-11	-
SP1	GGGGCGGGGC	3.91E-13	86	3	2E-262	(Lee et al., 1998)
DEC	CCCCAAGTGAAGG	3.91E-13	87	141	0.02661	-
HAND1E47	ANNGGNGTCTGGCATT	5.08E-13	88			
MYCMAX	NGACCACGTGGTCN	6.42E-13	89	81	4.3E-13	-
VDR	GGGTNAANGGGGTGA	8.61E-13	90	58	4.4E-22	-
E2F1DP1	TTTCCCGC	1.10E-12	91	17	2.9E-63	-
RP58	NAAACATCTGGA	1.40E-12	92			
ELK1	NNNNCCGGAAGTNN	1.46E-12	93	61	1.4E-21	(Goto et al., 2003)
E2F1	NTTCGCGC	4.02E-12	94	9	8E-132	-
LFA1	GGGGTCAG	4.02E-12	95	20	8.7E-58	-
CETS168	CAGGAAGC	5.28E-12	96	65	9E-20	-
NFY	TAACCAATCAC	5.28E-12	97	75	3.1E-16	-
EBF	GTCCCTTGGGA	9.48E-12	98	114	6.5E-05	-
NFMUE1	CGGCCATCT	1.27E-11	99	78	1.5E-15	-
STAT3	NNNTTCCN	3.78E-11	100	97	3.4E-08	(Waris et al., 2001)
MAZ	GGGGAGGG	4.82E-11	101	16	9.2E-69	(Su et al., 2007)
STAF	NTTACCCANAATGCATTGCGNN	5.29E-11	102			
SMAD3	TGTCTGTCT	6.11E-11	103	140	0.02596	-
OLF1	NNCNANTCCCCAGGGAGNNTGN	8.86E-11	104	84	2.9E-12	-
CAAT	NNTAGCCAATCA	1.24E-10	105	118	0.00042	-
MIF1	NNGTTGCTAGGCAACNGG	1.25E-10	106	99	1.8E-07	(Zhang et al., 2006)
SREBP1	NATCACGTGAC	1.60E-10	107	31	3.1E-42	(Kim et al., 2007)
NF1	NTGGNNNNNTGCCAANN	1.60E-10	108			
MYB	NNNGNCAGTTN	2.01E-10	109			
ETS	ANCCACTTCCTG	2.53E-10	110	105	5.6E-06	-
TFIII	AGAGGGAGG	3.19E-10	111	42	5.3E-30	-
AML	ANGTNTGTGGTTANC	6.33E-10	112			
NRF1	CGCATGCGCA	7.29E-10	113	22	2.4E-55	-
CACBINDINGPROTEIN	GAGGGTGGG	9.85E-10	114	32	4.2E-41	-
USF2	CACGTG	9.85E-10	115	68	3.5E-18	-
AP1	GGTGACTCAGA	9.85E-10	116			
EGR1	ATGCGTGGGCGT	1.12E-09	117			
MAF	TGCTGAGTCAN	1.55E-09	118	115	8.1E-05	-
MYOGNF1	CACCTGTTNNNTTTGGCACGGNGCCAACN	1.77E-09	119	104	5E-06	-
CMYC	NACCACGTGCTC	1.84E-09	120			
SRF	GNCCATATAAGGAC	1.84E-09	121			
AHRHIF	TGCGTGCGN	1.94E-09	122	59	5.3E-22	-
E2F	TTTCGCGC	5.75E-09	123	8	3E-144	(Choi et al., 2002)
ACAAT	GATTGGTGG	1.26E-08	124	131	0.00658	-
E4F1	GCTACGTCAC	1.33E-08	125	130	0.00651	(Rui et al., 2006)
P300	NCNGGGAGTGNGNG	1.65E-08	126	98	6.6E-08	(Cougot et al., 2007)
SF1	TGACCTTG	1.65E-08	127			
SZF11	CCAGGGTATCAGCCG	2.30E-08	128			
PR	GANAGAACAN	3.74E-08	129			

DR4	TGACCTNTACTGACCCC	4.54E-08	130	132	0.00659	-
GC	NGGGGGCGGGGCTN	1.01E-07	131	13	5.9E-82	-
VMYB	NCTAACGGN	1.47E-07	132	95	1.9E-08	-
PEBP	GNTAACCACAAANNT	1.47E-07	133			
TGIF	AGCTGTCANNA	1.47E-07	134			
ROAZ	GCACCCAAGGGTGC	1.81E-07	135			
ZID	NGGCTCTATCATC	2.17E-07	136	129	0.00643	-
AP1FJ	GGTGACTCAGT	2.17E-07	137			
DELTAEF1	NNTCACCTNAN	3.84E-07	138			
DR3	GATGAACTTNCTGAACCGTTT	3.87E-07	139	109	2.3E-05	-
ALPHACP1	CAGCCAATGAG	5.45E-07	140	143	0.03746	-
ATF4	CCTGACGCAATG	5.61E-07	141	102	8.7E-07	-
ZIC2	GGGGTGGTC	7.99E-07	142	57	3.4E-22	-
CREBATF	GTGACGTCA	1.40E-06	143	90	5.8E-10	-
RREB1	CCCCAAACCACCCC	1.54E-06	144	41	4.7E-31	-
AHR	CTTGCGTGNGN	2.37E-06	145	50	2.7E-24	-
OSF2	ACCACAAA	2.37E-06	146			
TEL2	CTACTTCCTG	2.92E-06	147			
PU1	AGAGGAAG	3.31E-06	148	144	0.04264	-
SP3	AGCCTTGGGGAGGG	6.19E-06	149	43	7.5E-29	-
HMX1	CAAGTGCGTG	6.39E-06	150	108	1.2E-05	-
GATA1	CNNGATNGNN	6.56E-06	151	56	1.7E-22	-
EGR3	NTGCGTGGGCGT	8.79E-06	152			
AHRARNT	TNNGGNTTGCGTGCCC	1.07E-05	153	52	7.7E-24	-
XPF1	TCAGAAGAAC	2.36E-05	154	125	0.00379	-
RBPJK	TTCCCACG	3.24E-05	155			
COUPTF	NNNNNTGACCTTTGCCCNCTGCN	5.07E-05	156	87	1.6E-11	-
PAX9	GAGACGCAGCGAGGAGTGACCACC	5.42E-05	157	76	1.3E-15	-
E2	NNACCGNNANCGGTGC	6.85E-05	158	100	2.7E-07	-
GCM	CANACCCGCATT	6.85E-05	159	124	0.00282	-
TTF1	CNCTCAAGNGNN	6.85E-05	160			
MEF3	GGGTCAGGTTTCA	8.03E-05	161	112	3.5E-05	-
GEN_INI	CCTCANTC	9.18E-05	162			
T3R	CANTGAGGTCACGCNN	1.04E-04	163	103	4.2E-06	-
BACH2	CGTGAGTCATC	1.05E-04	164	120	0.00092	-
E2F1DP2	TTTCCCGC	1.62E-04	165	26	2.1E-49	-
SMAD	TAGNCAGACAG	1.62E-04	166	94	1.1E-08	-
NGFIC	ATGCGTGGGCGG	1.75E-04	167			
ARNT	GTTGTCACGTGNNCGN	2.14E-04	168	83	6E-13	-
MAX	NAANCACGTGNTTN	2.54E-04	169	85	9.9E-12	-
STAT6	GNCTTCCT	3.66E-04	170			
RFX1	NNGTNGCCTGGCAACNN	5.44E-04	171			
PAX6	CTGACCTGGAACTC	7.88E-04	172			
TFIIA	TATAAAAGGACC	9.74E-04	173			
NFKAPPAB50	GGGGATTCCC	1.85E-03	174	77	1.5E-15	(Su and Schneider, 1996)
CDPCR3	CACCAATANGTATNG	1.95E-03	175			
STAT1	CANTTCCG	3.22E-03	176	117	0.00025	-
CREBP1	GGTGACGTAACT	3.33E-03	177	92	1.1E-09	(Cougot et al., 2007)
GATA2	NNNGATAGNN	3.62E-03	178	93	6.2E-09	-
UF1H3BETA	GGTGGGGGAGGGGC	4.48E-03	179			
GLI	NNTGGGTGGTCC	5.59E-03	180	86	9.9E-12	-
BRCA	TTNNGTTG	6.72E-03	181			
ATF1	CTCTGACGTCA	7.40E-03	182	91	9.9E-10	-
ATF6	TGACGTGG	7.94E-03	183	64	2E-20	(Li et al., 2007)
SPZ1	GNNGGAGGGTATGGC	1.00E-02	184	44	7.3E-28	-
ZEC	CAAGGTTGGTTGC	1.04E-02	185			
DEAF1	CCGCCCTCGGGTATTTCCGGAGNNG	1.11E-02	186	24	9.1E-52	-
PPARG	AACTAGGNCAAAGGTCA	1.21E-02	187			
TCF11	GTCATNNTNNNNN	1.47E-02	188			
TEF1	GGAATG	2.11E-02	189			
HSF1	NTTCTAGAANNTTCTCC	2.36E-02	190			
E2F4DP2	TTTCCCGC	2.72E-02	191	27	6.9E-49	-
PEA3	ACATCCT	3.24E-02	192			
ERR1	NNNTCAAGGTCANA	3.77E-02	193			
E2F4DP1	TTTCGCGC	4.43E-02	194	28	2.4E-46	-
IK1	NNTTGGGAATACC	4.45E-02	195			

References

Abel, H.J., Duncavage, E.J., Becker, N., Armstrong, J.R., Magrini, V.J., Pfeifer, J.D., 2010. SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. Bioinformatics 26, 2684-2688.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

Andrisani, O.M., Barnabas, S., 1999. The transcriptional function of the hepatitis B virus X protein and its role in hepatocarcinogenesis (Review). Int J Oncol 15, 373-379.

Ansorge, W.J., 2009. Next-generation DNA sequencing techniques. N Biotechnol 25, 195-203.

Aparicio, O., Geisberg, J.V., Struhl, K., 2004. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol Chapter 17, Unit 17 17.

Arbuthnot, P., Kew, M., 2001. Hepatitis B virus and hepatocellular carcinoma. Int J Exp Pathol 82, 77-100.

Arzumanyan, A., Friedman, T., Kotei, E., Ng, I.O., Lian, Z., Feitelson, M.A., 2012. Epigenetic repression of E-cadherin expression by hepatitis B virus x antigen in liver cancer. Oncogene 31, 563-572.

Bailey, T.L., 2008. Discovering sequence motifs. Methods Mol Biol 452, 231-251.

Bailey, T.L., 2011. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27, 1653-1659.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37, W202-208.

Bailey, T.L., Williams, N., Misleh, C., Li, W.W., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34, W369-373.

Barnabas, S., Hai, T., Andrisani, O.M., 1997. The hepatitis B virus X protein enhances the DNA binding potential and transcription efficacy of bZip transcription factors. J Biol Chem 272, 20684-20690.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K., 2007. High-resolution profiling of histone methylations in the human genome. Cell 129, 823-837.

Bashir, A., Volik, S., Collins, C., Bafna, V., Raphael, B.J., 2008. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. PLoS Comput Biol 4, e1000051.

Becker, S.A., Lee, T.H., Butel, J.S., Slagle, B.L., 1998. Hepatitis B virus X protein interferes with cellular DNA repair. J Virol 72, 266-272.

Benn, J., Su, F., Doria, M., Schneider, R.J., 1996. Hepatitis B virus HBx protein induces transcription factor AP-1 by activation of extracellular signal-regulated and c-Jun N-terminal mitogen-activated protein kinases. J Virol 70, 4978-4985.

Bill, C.A., Summers, J., 2004. Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. Proc Natl Acad Sci U S A 101, 11135-11140.

Blahnik, K.R., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A.R., Iyengar, S., Nicolet, C.M., Ludascher, B., Korf, I., Farnham, P.J., 2010. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. Nucleic Acids Res 38, e13.

Blat, Y., Kleckner, N., 1999. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. Cell 98, 249-259.

Blum, H.E., 2005. Hepatocellular carcinoma: therapy and prevention. World J Gastroenterol 11, 7391-7400.

Bonilla Guerrero, R., Roberts, L.R., 2005. The role of hepatitis B virus integrations in the pathogenesis of human hepatocellular carcinoma. J Hepatol 42, 760-777.

Botcheva, K., McCorkle, S.R., McCombie, W.R., Dunn, J.J., Anderson, C.W., 2011. Distinct p53 genomic binding patterns in normal and cancer-derived human cells. Cell Cycle 10.

Boyault, S., Rickman, D.S., de Reynies, A., Balabaud, C., Rebouissou, S., Jeannot, E., Herault, A., Saric, J., Belghiti, J., Franco, D., Bioulac-Sage, P., Laurent-Puig, P., Zucman-Rossi, J., 2007. Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology 45, 42-52.

Boyle, A.P., Guinney, J., Crawford, G.E., Furey, T.S., 2008. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 24, 2537-2538.

Bucher, P., 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212, 563-578.

Buck, M.J., Lieb, J.D., 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83, 349-360.

Buendia, M.A., 1992. Hepatitis B viruses and hepatocellular carcinoma. Adv Cancer Res 59, 167-226.

Chaisson, M.J., Pevzner, P.A., 2008. Short read fragment assembly of bacterial genomes. Genome Res 18, 324-330.

Chan, T.M., Li, G., Leung, K.S., Lee, K.H., 2009. Discovering multiple realistic TFBS motifs based on a generalized model. BMC Bioinformatics 10, 321.

Chang, M.H., 2003. Decreasing incidence of hepatocellular carcinoma among children following universal hepatitis B immunization. Liver Int 23, 6.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T.J., Wilson, R.K., Ding, L., Mardis, E.R., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6, 677-681.

Cheong, J.H., Yi, M., Lin, Y., Murakami, S., 1995. Human RPB5, a subunit shared by eukaryotic nuclear RNA polymerases, binds human hepatitis B virus X protein and may play a role in X transactivation. Embo J 14, 143-150.

Choi, M., Lee, H., Rho, H.M., 2002. E2F1 activates the human p53 promoter and overcomes the repressive effect of hepatitis B viral X protein (Hbx) on the p53 promoter. IUBMB Life 53, 309-317.

Cougot, D., Wu, Y., Cairo, S., Caramel, J., Renard, C.A., Levy, L., Buendia, M.A., Neuveut, C., 2007. The hepatitis B virus X protein functionally interacts with CREB-binding protein/p300 in the regulation of CREB-mediated transcription. J Biol Chem 282, 4277-4287.

Das, M.K., Dai, H.K., 2007. A survey of DNA motif finding algorithms. BMC Bioinformatics 8 Suppl 7, S21.

Dombkowski, A.A., Thibodeau, B.J., Starcevic, S.L., Novak, R.F., 2004. Gene-specific dye bias in microarray reference designs. FEBS Lett 560, 120-124.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K.P., Baccash, J., Borcherding, A.P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J.C., Hacker, C.R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C.E., Morenzoni, M., Morey, R.E., Mutch, K., Perazich, H., Perry, K., Peters, B.A., Peterson, J., Pethiyagoda, C.L., Pothuraju, K., Richter, C., Rosenbaum, A.M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K.W., Sheppy, C.G., Sun, M., Thakuria, J.V., Tran, A., Vu, D., Zaranek, A.W., Wu, X., Drmanac, S., Oliphant, A.R., Banyai, W.C., Martin, B., Ballinger, D.G., Church, G.M., Reid, C.A., 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327, 78-81.

Duncavage, E.J., Magrini, V., Becker, N., Armstrong, J.R., Demeter, R.T., Wylie, T., Abel, H.J., Pfeifer, J.D., 2011. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J Mol Diagn 13, 325-333.

Eklund, A.C., Szallasi, Z., 2008. Correction of technical bias in clinical microarray data improves concordance with known biological information. Genome Biol 9, R26.

Febbo, P.G., Kantoff, P.W., 2006. Noise and bias in microarray analysis of tumor specimens. J Clin Oncol 24, 3719-3721.

Ganem, D., 2001. Virology. The X files--one step closer to closure. Science 294, 2299-2300.

Goto, T., Kato, N., Yoshida, H., Otsuka, M., Moriyama, M., Shiratori, Y., Koike, K., Matsumura, M., Omata, M., 2003. Synergistic activation of the serum response element-dependent pathway by hepatitis B virus x protein and large-isoform hepatitis delta antigen. J Infect Dis 187, 820-828.

Goto, Y., Yoshida, J., Kuzushima, K., Terashima, M., Morishima, T., 1993. Patterns of hepatitis B virus DNA integration in liver tissue of children with chronic infections. J Pediatr Gastroenterol Nutr 16, 70-74.

Groisman, I.J., Koshy, R., Henkler, F., Groopman, J.D., Alaoui-Jamali, M.A., 1999. Downregulation of DNA excision repair by the hepatitis B virus-x protein occurs in p53-proficient and p53-deficient cells. Carcinogenesis 20, 479-483.

Hawkins, R.D., Hon, G.C., Ren, B., 2010. Next-generation genomics: an integrative approach. Nat Rev Genet 11, 476-486.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576-589.

Hernandez, P., Martis, M., Dorado, G., Pfeifer, M., Galvez, S., Schaaf, S., Jouve, N., Simkova, H., Valarik, M., Dolezel, J., Mayer, K.F., 2011. Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. Plant J.

Hoffman, B.G., Jones, S.J., 2009. Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. J Endocrinol 201, 1-13.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E.E., Sahinalp, S.C., 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26, i350-357.

Hu, M., Yu, J., Taylor, J.M., Chinnaiyan, A.M., Qin, Z.S., 2010. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. Nucleic Acids Res 38, 2154-2167.

Huang, J., Wang, Y., Guo, Y., Sun, S., 2010. Down-regulated microRNA-152 induces aberrant DNA methylation in hepatitis B virus-related hepatocellular carcinoma by targeting DNA methyltransferase 1. Hepatology 52, 60-70.

Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M., Wong, W.H., 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26, 1293-1300.

Ji, H., Jiang, H., Ma, W., Wong, W.H., 2011. Using CisGenome to analyze ChIP-chip and ChIP-seq data. Curr Protoc Bioinformatics Chapter 2, Unit2 13.

Jiang, H., Wang, F., Dyer, N.P., Wong, W.H., 2010. CisGenome Browser: a flexible tool for genomic data visualization. Bioinformatics 26, 1781-1782.

Jiang, H., Wong, W.H., 2008. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24, 2395-2396.

Jiang, Z., Jhunjhunwala, S., Liu, J., Haverty, P.M., Kennemer, M.I., Guan, Y., Lee, W., Carnevali, P., Stinson, J., Johnson, S., Diao, J., Yeung, S., Jubb, A., Ye, W., Wu, T.D., Kapadia, S.B., de Sauvage, F.J., Gentleman, R.C., Stern, H.M., Seshagiri, S., Pant, K.P., Modrusan, Z., Ballinger, D.G., Zhang, Z., 2012. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res 22, 593-601.

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497-1502.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., Zhao, K., 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36, 5221-5231.

Jung, J.K., Park, S.H., Jang, K.L., 2010. Hepatitis B virus X protein overcomes the growth-inhibitory potential of retinoic acid by downregulating retinoic acid receptor-beta2 expression via DNA methylation. J Gen Virol 91, 493-500.

Kharchenko, P.V., Tolstorukov, M.Y., Park, P.J., 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26, 1351-1359.

Kim, J.H., Rho, H.M., 2002. Activation of the human transforming growth factor alpha (TGF-alpha) gene by the hepatitis B viral X protein (HBx) through AP-2 sites. Mol Cell Biochem 231, 155-161.

Kim, K.H., Shin, H.J., Kim, K., Choi, H.M., Rhee, S.H., Moon, H.B., Kim, H.H., Yang, U.S., Yu, D.Y., Cheong, J., 2007. Hepatitis B virus X protein induces hepatic steatosis via transcriptional activation of SREBP1 and PPARgamma. Gastroenterology 132, 1955-1967.

Kim, Y.J., Jung, J.K., Lee, S.Y., Jang, K.L., 2010. Hepatitis B virus X protein overcomes stress-induced premature senescence by repressing p16(INK4a) expression via DNA methylation. Cancer Lett 288, 226-235.

Kong, G., Zhang, J., Zhang, S., Shan, C., Ye, L., Zhang, X., 2011. Upregulated microRNA-29a by hepatitis B virus X protein enhances hepatoma cell migration by targeting PTEN in cell culture model. PLoS One 6, e19518.

Korbel, J.O., Abyzov, A., Mu, X.J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., Gerstein, M.B., 2009. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10, R23.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M., Snyder, M., 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420-426.

Kuang, S.Y., Jackson, P.E., Wang, J.B., Lu, P.X., Munoz, A., Qian, G.S., Kensler, T.W., Groopman, J.D., 2004. Specific mutations of hepatitis B virus in plasma predict liver cancer development. Proc Natl Acad Sci U S A 101, 3575-3580.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25.

Lee, A.T., Lee, C.G., 2007. Oncogenesis and transforming viruses: the hepatitis B virus and hepatocellularcarcinoma--the etiopathogenic link. Front Biosci 12, 234-245.

Lee, S., Hormozdiari, F., Alkan, C., Brudno, M., 2009. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods 6, 473-474.

Lee, Y.I., Lee, S., Lee, Y., Bong, Y.S., Hyun, S.W., Yoo, Y.D., Kim, S.J., Kim, Y.W., Poo, H.R., 1998. The human hepatitis B virus transactivator X gene product regulates Sp1 mediated transcription of an insulin-like growth factor II promoter 4. Oncogene 16, 2367-2380.

Li, B., Gao, B., Ye, L., Han, X., Wang, W., Kong, L., Fang, X., Zeng, Y., Zheng, H., Li, S., Wu, Z., 2007. Hepatitis B virus X protein (HBx) activates ATF6 and IRE1-XBP1 pathways of unfolded protein response. Virus Res 124, 44-49.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.

Li, R., Li, Y., Kristiansen, K., Wang, J., 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713-714.

Li, Z., Van Calcar, S., Qu, C., Cavenee, W.K., Zhang, M.Q., Ren, B., 2003. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. Proc Natl Acad Sci U S A 100, 8164-8169.

Lim, C.A., Yao, F., Wong, J.J., George, J., Xu, H., Chiu, K.P., Sung, W.K., Lipovich, L., Vega, V.B., Chen, J., Shahab, A., Zhao, X.D., Hibberd, M., Wei, C.L., Lim, B., Ng, H.H., Ruan, Y., Chin, K.C., 2007. Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation. Mol Cell 27, 622-635.

Lin, H., Zhang, Z., Zhang, M.Q., Ma, B., Li, M., 2008. ZOOM! Zillions of oligos mapped. Bioinformatics 24, 2431-2437.

Liu, E.T., Pott, S., Huss, M., 2010. Q&A: ChIP-seq technologies and the study of gene regulation. BMC Biol 8, 56.

Liu, X.S., Brutlag, D.L., Liu, J.S., 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 20, 835-839.

Lun, D.S., Sherrid, A., Weiner, B., Sherman, D.R., Galagan, J.E., 2009. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. Genome Biol 10, R142.

Lupberger, J., Hildt, E., 2007. Hepatitis B virus-induced oncogenesis. World J Gastroenterol 13, 74-81.

Madzima, T.F., Mills, E.S., Gardiner, J.M., McGinnis, K.M., 2011. Identification of epigenetic regulators of a transcriptionally silenced transgene in maize. G3 (Bethesda) 1, 75-83.

Maguire, H.F., Hoeffler, J.P., Siddiqui, A., 1991. HBV X protein alters the DNA binding specificity of CREB and ATF-2 by protein-protein interactions. Science 252, 842-844.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., Turner, D.J., 2010. Target-enrichment strategies for next-generation sequencing. Nat Methods 7, 111-118.

Mardis, E.R., 2008a. The impact of next-generation sequencing technology on genetics. Trends Genet 24, 133-141.

Mardis, E.R., 2008b. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9, 387-402.

Mardis, E.R., 2009. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med 1, 40.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., Fulton, L.A., Locke, D.P., Magrini, V.J., Abbott, R.M., Vickery, T.L., Reed, J.S., Robinson, J.S., Wylie, T., Smith, S.M., Carmichael, L., Eldred, J.M., Harris, C.C., Walker, J., Peck, J.B., Du, F., Dukes, A.F., Sanderson, G.E., Brummett, A.M., Clark, E., McMichael, J.F., Meyer, R.J., Schindler, J.K., Pohl, C.S., Wallis, J.W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M.E., Ivy, J.V., Kalicki, J., Elliott, G., Ries, R.E., Payton, J.E., Westervelt, P., Tomasson, M.H., Watson, M.A., Baty, J., Heath, S., Shannon, W.D., Nagarajan, R., Link, D.C., Walter, M.J., Graubert, T.A., DiPersio, J.F., Wilson, R.K., Ley, T.J., 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 361, 1058-1066.

Mather, L.E., Austin, K.L., 1983. The Statistical Package for the Social Sciences (SPSS) as an adjunct to pharmacokinetic analysis. Biopharm Drug Dispos 4, 157-172.

Matsuda, Y., Ichida, T., 2009. Impact of hepatitis B virus X protein on the DNA damage response during hepatocarcinogenesis. Med Mol Morphol 42, 138-142.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S., Bernstein, B.E., 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553-560.

Moon, E.J., Jeong, C.H., Jeong, J.W., Kim, K.R., Yu, D.Y., Murakami, S., Kim, C.W., Kim, K.W., 2004. Hepatitis B virus X protein induces angiogenesis by stabilizing hypoxia-inducible factor-1alpha. Faseb J 18, 382-384.

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

Murakami, Y., Saigo, K., Takashima, H., Minami, M., Okanoue, T., Brechot, C., Paterlini-Brechot, P., 2005. Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas. Gut 54, 1162-1168.

Ng, P., Tan, J.J., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C., Du, L., Sung, W.K., Wei, C.L., Ruan, Y., 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic Acids Res 34, e84.

Ng, S.A., Lee, C., 2011. Hepatitis B virus X gene and hepatocarcinogenesis. J Gastroenterol 46, 974-990.

Nix, D.A., Courdy, S.J., Boucher, K.M., 2008. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics 9, 523.

Northrup, D.L., Zhao, K., 2011. Application of ChIP-Seq and related techniques to the study of immune function. Immunity 34, 830-842.

Park, P.J., 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10, 669-680.

Park, S.H., Jung, J.K., Lim, J.S., Tiwari, I., Jang, K.L., 2011. Hepatitis B virus X protein overcomes all-trans retinoic acid-induced cellular senescence by downregulating levels of p16 and p21 via DNA methylation. J Gen Virol 92, 1309-1317.

Parkin, D.M., Bray, F.I., Devesa, S.S., 2001. Cancer burden in the year 2000. The global picture. Eur J Cancer 37 Suppl 8, S4-66.

Paterlini-Brechot, P., Saigo, K., Murakami, Y., Chami, M., Gozuacik, D., Mugnier, C., Lagorce, D., Brechot, C., 2003. Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. Oncogene 22, 3911-3916.

Pavesi, G., Mereghetti, P., Mauri, G., Pesole, G., 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res 32, W199-203.

Pepke, S., Wold, B., Mortazavi, A., 2009. Computation for ChIP-seq and RNA-seq studies. Nat Methods 6, S22-32.

Pineau, P., Marchio, A., Mattei, M.G., Kim, W.H., Youn, J.K., Tiollais, P., Dejean, A., 1998. Extensive analysis of duplicated-inverted hepatitis B virus integrations in human hepatocellular carcinoma. J Gen Virol 79 (Pt 3), 591-600.

Porreca, G.J., 2010. Genome sequencing on nanoballs. Nat Biotechnol 28, 43-44.

Qadri, I., Maguire, H.F., Siddiqui, A., 1995. Hepatitis B virus transactivator protein X interacts with the TATA-binding protein. Proc Natl Acad Sci U S A 92, 1003-1007.

Qin, Z.S., Yu, J., Shen, J., Maher, C.A., Hu, M., Kalyana-Sundaram, S., Chinnaiyan, A.M., 2010. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC Bioinformatics 11, 369.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O.L., He, A., Marra, M., Snyder, M., Jones, S., 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4, 651-657.

Robinson, W.S., 1994. Molecular events in the pathogenesis of hepadnavirus-associated hepatocellular carcinoma. Annu Rev Med 45, 297-323.

Rosenzweig, B.A., Pine, P.S., Domon, O.E., Morris, S.M., Chen, J.J., Sistare, F.D., 2004. Dye bias correction in dual-labeled cDNA microarray gene expression measurements. Environ Health Perspect 112, 480-487.

Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., Gerstein, M.B., 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27, 66-75.

Ruan, Y., Ooi, H.S., Choo, S.W., Chiu, K.P., Zhao, X.D., Srinivasan, K.G., Yao, F., Choo, C.Y., Liu, J., Ariyaratne, P., Bin, W.G., Kuznetsov, V.A., Shahab, A., Sung, W.K., Bourque, G., Palanisamy, N., Wei, C.L., 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res 17, 828-838.

Rui, E., Moura, P.R., Goncalves, K.A., Rooney, R.J., Kobarg, J., 2006. Interaction of the hepatitis B virus protein HBx with the human transcription regulatory protein p120E4F in vitro. Virus Res 115, 31-42.

Saigo, K., Yoshida, K., Ikeda, R., Sakamoto, Y., Murakami, Y., Urashima, T., Asano, T., Kenmochi, T., Inoue, I., 2008. Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (MLL4) gene and rearrangements of MLL4 in human hepatocellular carcinoma. Hum Mutat 29, 703-708.

Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. Nat Methods 5, 16-18.

Scisciani, C., Vossio, S., Guerrieri, F., Schinzari, V., De Iaco, R., Cervello, M., Montalto, G., Pollicino, T., Raimondo, G., Levrero, M., Pediconi, N., 2011. Transcriptional regulation of mir-224 upregulated in human HCCs by NFkB inflammatory pathways. J Hepatol.

Shan, C., Zhang, S., Cui, W., You, X., Kong, G., Du, Y., Qiu, L., Ye, L., Zhang, X., 2011. Hepatitis B virus X protein activates CD59 involving DNA binding and let-7i in protection of hepatoma and hepatic cells from complement attack. Carcinogenesis 32, 1190-1197.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat Biotechnol 26, 1135-1145.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res 19, 1117-1123.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., Zhang, M.Q., 2009. Updates to the RMAP short-read mapping software. Bioinformatics 25, 2841-2842.

Smith, A.D., Xuan, Z., Zhang, M.Q., 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9, 128.

Spyrou, C., Stark, R., Lynch, A.G., Tavare, S., 2009. BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics 10, 299.

Steger, D., Berry, D., Haider, S., Horn, M., Wagner, M., Stocker, R., Loy, A., 2011. Systematic spatial bias in DNA microarray hybridization is caused by probe spot position-dependent variability in lateral diffusion. PLoS One 6, e23727.

Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J., Greenman, C.D., Jia, M., Latimer, C., Teague, J.W., Lau, K.W., Burton, J., Quail, M.A., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A.M., Martens, J.W., Silver, D.P., Langerod, A., Russnes, H.E., Foekens, J.A., Reis-Filho, J.S., van 't Veer, L., Richardson, A.L., Borresen-Dale, A.L., Campbell, P.J., Futreal, P.A., Stratton, M.R., 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462, 1005-1010.

Su, F., Schneider, R.J., 1996. Hepatitis B virus HBx protein activates transcription factor NF-kappaB by acting on multiple cytoplasmic inhibitors of rel-related proteins. J Virol 70, 4558-4566.

Su, H., Zhao, J., Xiong, Y., Xu, T., Zhou, F., Yuan, Y., Zhang, Y., Zhuang, S.M., 2008. Large-scale analysis of the genetic and epigenetic alterations in hepatocellular carcinoma from Southeast China. Mutat Res 641, 27-35.

Su, J.M., Lai, X.M., Lan, K.H., Li, C.P., Chao, Y., Yen, S.H., Chang, F.Y., Lee, S.D., Lee, W.P., 2007. X protein of hepatitis B virus functions as a transcriptional corepressor on the human telomerase promoter. Hepatology 46, 402-413.

Sun, H., Yuan, Y., Wu, Y., Liu, H., Liu, J.S., Xie, H., 2010. Tmod: toolbox of motif discovery. Bioinformatics 26, 405-407.

Sung, W.K., Lu, Y., Lee, C.W., Zhang, D., Ronaghi, M., Lee, C.G., 2009. Deregulated direct targets of the hepatitis B virus (HBV) protein, HBx, identified through chromatin immunoprecipitation and expression microarray profiling. J Biol Chem 284, 21941-21954.

Tamori, A., Nishiguchi, S., Kubo, S., Narimatsu, T., Habu, D., Takeda, T., Hirohashi, K., Shiomi, S., 2003. HBV DNA integration and HBV-transcript expression in non-B, non-C hepatocellular carcinoma in Japan. J Med Virol 71, 492-498.

Tamori, A., Nishiguchi, S., Shiomi, S., Hayashi, T., Kobayashi, S., Habu, D., Takeda, T., Seki, S., Hirohashi, K., Tanaka, H., Kubo, S., 2005. Hepatitis B virus DNA integration in hepatocellular carcinoma after interferon-induced disappearance of hepatitis C virus. Am J Gastroenterol 100, 1748-1753.

Tan, Y.J., 2011. Hepatitis B virus infection and the risk of hepatocellular carcinoma. World J Gastroenterol 17, 4853-4857.

Tu, H., Gao, H.F., Ma, G.H., Liu, Y., 2006. [Identification of hepatitis B virus integration sites in hepatocellular carcinoma tissues from patients with chronic hepatitis B]. Zhonghua Yi Xue Za Zhi 86, 1249-1252.

Tuteja, G., White, P., Schug, J., Kaestner, K.H., 2009. Extracting transcription factor targets from ChIP-Seq data. Nucleic Acids Res 37, e113.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M.V., Eichler, E.E., 2005. Fine-scale structural variation of the human genome. Nat Genet 37, 727-732.

Um, T.H., Kim, H., Oh, B.K., Kim, M.S., Kim, K.S., Jung, G., Park, Y.N., 2011. Aberrant CpG island hypermethylation in dysplastic nodules and early HCC of hepatitis B virus-related human multistep hepatocarcinogenesis. J Hepatol 54, 939-947.

Urashima, T., Saigo, K., Kobayashi, S., Imaseki, H., Matsubara, H., Koide, Y., Asano, T., Kondo, Y., Koike, K., Isono, K., 1997. Identification of hepatitis B virus integration in hepatitis C virus-infected hepatocellular carcinoma tissues. J Hepatol 26, 771-778.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., Sidow, A., 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5, 829-834.

Volik, S., Zhao, S., Chin, K., Brebner, J.H., Herndon, D.R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W.L., Magrane, G., De Jong, P., Gray, J.W., Collins, C., 2003. End-sequence profiling: sequence-based analysis of aberrant genomes. Proc Natl Acad Sci U S A 100, 7696-7701.

Wang, Y., Lu, Y., Toh, S.T., Sung, W.K., Tan, P., Chow, P., Chung, A.Y., Jooi, L.L., Lee, C.G., 2010. Lethal-7 is down-regulated by the hepatitis B virus x protein and targets signal transducer and activator of transcription 3. J Hepatol 53, 57-66.

Wang, Y., Toh, H.C., Chow, P., Chung, A.Y., Meyers, D.J., Cole, P.A., Ooi, L.L., Lee, C.G., 2012. MicroRNA-224 is up-regulated in hepatocellular carcinoma through epigenetic mechanisms. FASEB journal : official publication of the Federation of American Societies for Experimental Biology.

Waris, G., Huh, K.W., Siddiqui, A., 2001. Mitochondrially associated hepatitis B virus X protein constitutively activates transcription factors STAT-3 and NF-kappa B via oxidative stress. Mol Cell Biol 21, 7721-7730.

Warren, R.L., Sutton, G.G., Jones, S.J., Holt, R.A., 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23, 500-501.

Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., Liu, J., Zhao, X.D., Chew, J.L., Lee, Y.L., Kuznetsov, V.A., Sung, W.K., Miller, L.D., Lim, B., Liu, E.T., Yu, Q., Ng, H.H., Ruan, Y., 2006. A global map of p53 transcription-factor binding sites in the human genome. Cell 124, 207-219.

Weinmann, A.S., Bartley, S.M., Zhang, T., Zhang, M.Q., Farnham, P.J., 2001. Use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol Cell Biol 21, 13.

White, R.J., 2011. Transcription by RNA polymerase III: more complex than we thought. Nat Rev Genet 12, 459-463.

Wilbanks, E.G., Facciotti, M.T., 2010. Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One 5, e11471.

Williams, J.S., Andrisani, O.M., 1995. The hepatitis B virus X protein targets the basic region-leucine zipper domain of CREB. Proc Natl Acad Sci U S A 92, 3819-3823.

Wingender, E., Dietze, P., Karas, H., Knuppel, R., 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24, 238-241.

Wu, C.G., Salvay, D.M., Forgues, M., Valerie, K., Farnsworth, J., Markin, R.S., Wang, X.W., 2001. Distinctive gene expression profiles associated with Hepatitis B virus x protein. Oncogene 20, 3674-3682.

Wu, G., Yu, F., Xiao, Z., Xu, K., Xu, J., Tang, W., Wang, J., Song, E., 2011. Hepatitis B virus X protein downregulates expression of the miR-16 family in malignant hepatocytes in vitro. Br J Cancer 105, 146-153.

Xie, Z., Hu, S., Qian, J., Blackshaw, S., Zhu, H., 2011. Systematic characterization of protein-DNA interactions. Cell Mol Life Sci 68, 1657-1668.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.L., Lin, F., Sung, W.K., 2010. A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics 26, 1199-1204.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z., 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865-2871.

Yip, W.K., Cheng, A.S., Zhu, R., Lung, R.W., Tsang, D.P., Lau, S.S., Chen, Y., Sung, J.G., Lai, P.B., Ng, E.K., Yu, J., Wong, N., To, K.F., Wong, V.W., Sung, J.J., Chan, H.L., 2011. Carboxyl-terminal truncated HBx regulates a distinct microRNA transcription program in hepatocellular carcinoma development. PLoS One 6, e22888.

Yoo, Y.G., Lee, M.O., 2004. Hepatitis B virus X protein induces expression of Fas ligand gene through enhancing transcriptional activity of early growth response factor. J Biol Chem 279, 36242-36249.

Yuan, K., Lian, Z., Sun, B., Clayton, M.M., Ng, I.O., Feitelson, M.A., 2012. Role of miR-148a in Hepatitis B Associated Hepatocellular Carcinoma. PLoS One 7, e35331.

Zeller, K.I., Zhao, X., Lee, C.W., Chiu, K.P., Yao, F., Yustein, J.T., Ooi, H.S., Orlov, Y.L., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., Kuznetsov, V.A., Sung, W.K., Ruan, Y., Dang, C.V., Wei, C.L., 2006. Global mapping of c-Myc binding sites and target gene networks in human B cells. Proc Natl Acad Sci U S A 103, 17834-17839.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821-829.

Zhang, S., Lin, R., Zhou, Z., Wen, S., Lin, L., Chen, S., Shan, Y., Cong, Y., Wang, S., 2006. Macrophage migration inhibitory factor interacts with HBx and inhibits its apoptotic activity. Biochem Biophys Res Commun 342, 671-679.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137.

Zhang, Z., Chang, C.W., Goh, W.L., Sung, W.K., Cheung, E., 2011. CENTDIST: discovery of co-associated factors by motif distribution. Nucleic Acids Res 39, W391-399.

Zheng, Y., Chen, W.L., Ma, W.L., Chang, C., Ou, J.H., 2007. Enhancement of gene transactivation activity of androgen receptor by hepatitis B virus X protein. Virology 363, 454-461.

Zhu, Y.Z., Zhu, R., Shi, L.G., Mao, Y., Zheng, G.J., Chen, Q., Zhu, H.G., 2010. Hepatitis B virus X protein promotes hypermethylation of p16(INK4A) promoter through upregulation of DNA methyltransferases in hepatocarcinogenesis. Exp Mol Pathol 89, 268-275.

Author’s Publications

Toh, S.T., Jin, Y., **Liu, L.**, Wang, J., Babrzadeh, F., Gharizadeh, B., Ronaghi, M., Toh, H.C., Chow, K.H., Chung, Y.F., Ooi, L.P.J., Lee, C.G., 2012. Deep Sequencing of the Hepatitis B Virus in Hepatocellular Carcinoma Patients reveals Non-Random Integration Events, Structural Alterations and Sequence Variations. (Submitted for publication consideration)