


8-2019

Association of copy number variations with chronic hepatitis B in Chinese population

Fang Niu
University of Nebraska Medical Center

Tell us how you used this information in this [short survey](#).

Follow this and additional works at: https://digitalcommons.unmc.edu/coph_slce

 Part of the [Biostatistics Commons](#), [Categorical Data Analysis Commons](#), [Computational Biology Commons](#), [Genomics Commons](#), and the [Public Health Commons](#)

Recommended Citation

Niu, Fang, "Association of copy number variations with chronic hepatitis B in Chinese population" (2019). *Capstone Experience*. 80.

https://digitalcommons.unmc.edu/coph_slce/80

This Capstone Experience is brought to you for free and open access by the Master of Public Health at DigitalCommons@UNMC. It has been accepted for inclusion in Capstone Experience by an authorized administrator of DigitalCommons@UNMC. For more information, please contact digitalcommons@unmc.edu.

Association of copy number variations with chronic hepatitis B in Chinese population

Student: Fang Niu, MS, Master of Public Health concentration in Biostatistics

Chair: Jiangtao Luo, ph.D

Faculty: Hongmei Wang, ph.D

Preceptor: Haitao Chen, ph.D

March 16th, 2019

Abstract

With one third of the Hepatitis B virus (HBV) infection population of the world, chronic Hepatitis B (CHB) has become a top burden in China. CHB is a lifelong infection with HBV which can cause serious health problems, like cirrhosis, liver cancer or even death. HBV infection is known to result in various clinical conditions, including asymptomatic HBV carriers to chronic hepatitis and primary hepatocellular carcinoma. Several studies have shown that host genetic susceptibility could be an important factor that determines these various outcomes of HBV infection. Many Single Nucleotide Polymorphisms (SNPs) and Copy Number Variations (CNVs) have been associated with genetic susceptibility for many diseases including the HBV infection. SNPs and CNVs of the host could determine the CHB outcomes and disease progression. In this project, we conducted SNP-based and copy number polymorphic region (CNPR) - based CNV analysis of the genotyping data generated from 2,689 CHB patients and 1,200 healthy controls in Chinese population by Illumina Human OmniExpress BeadChip and OmniZhonghua BeadChip. Based on the analysis results, we found 8 deletion CNPRs, as well as 3 duplication CNPRs were significantly changed between CHB patients and healthy controls. Moreover, there were nine genes revealed the copy number loss, including FGFR3 ($p=1.49 \times 10^{-7}$), FGR-3 ($p=1.49 \times 10^{-7}$), LETM1 ($p=1.49 \times 10^{-7}$), TACC3 ($p=1.49 \times 10^{-7}$), TMEM129 ($p=1.49 \times 10^{-7}$), PANK4 ($p=4.55 \times 10^{-4}$), PLCH2 ($p=4.55 \times 10^{-4}$), CED-6 ($p=2.04 \times 10^{-4}$), DIRC1 ($p=2.04 \times 10^{-4}$), as well as three genes revealed the copy number gain, including FLJ43080 ($p=2.50 \times 10^{-5}$), CSMD3 ($p=6.288 \times 10^{-5}$), MGAT4C ($p=1.52 \times 10^{-4}$) in CHB patients compared with healthy controls. It is important to understand the functions of these genes and the mechanisms through which these genes are associated with HBV

infection and CHB development. Through the CNVs analysis, we provided potential therapeutic targets and novel diagnosis markers for HBV infection and CHB development.

Introduction

Southern Medical University Nanfang Hospital is a large-scale comprehensive tertiary hospital with medical, teaching, scientific research and preventive health care. The comprehensive ranking of the Nanfang Hospital is stable at around 15 in the domestic authority list, and 15 specialties have entered the top ten rankings of China's best specialists list (Fudan Edition).

The liver disease center in Nanfang Hospital is the “State Key Laboratory of Organ Failure Research” and the “Guangdong Provincial Key Laboratory of Viral Hepatitis Research” and the member of the Chinese Medical Association Infectious Diseases. The center consists of two parts: the clinical department for patients with various liver diseases and the laboratory for basic research on liver diseases. In the past ten years, the research group has been focusing on the genomic variation and molecular mechanism of liver cancer development, and has achieved a series of research results. They published many articles in the leading journals in the field of genetics and liver diseases, like Nature Genetics and Hepatology. During these studies they have accumulated abundant experience in liver disease research and collected a large number of samples. They have a high-quality sample library of up to 3,000 liver cancer samples and more than 4,000 chronic hepatitis B control samples closely matched with liver cancer cases, including the Qidong area in Jiangsu Province, covering many high-risk areas of the country. In

summary, Nanfang Hospital is an outstanding teaching hospital and great place for me completing my capstone project.

Hepatitis B is a liver infection disease caused by hepatitis B virus (HBV) and spreads via blood or body fluids among people (CDC, 2018a). HBV infection can result in various outcomes, from a short-term illness (Acute hepatitis B) to a lifelong infection (Chronic hepatitis B) (CDC, 2018b). The risk of developing a chronic hepatitis B infection depends on the age of who first exposure to HBV. Approximately 90% of infants who infected with HBV will develop a chronic hepatitis B (CHB). CHB can cause serious health issues, like cirrhosis, liver cancer or even death (CDC, 2018b). Globally, hepatitis B resulted in 887,000 deaths due to complications, including cirrhosis and hepatocellular carcinoma, in 2015 (WHO, 2018). An estimated 257 million people are living with chronic hepatitis B virus infection worldwide, of which around 90 million reside China (WHO, 2016). With over one third of the HBV infected population of the world, China has been identified as the area with highest hepatitis burden (WHO, 2016). Although CHB has been found that was associated with approximately 90% of neonatal HBV infection and 30% of childhood infection (Yan et al., 2014). The mechanisms of HBV infection and CHB development still remains largely unknown. With rapid advances in genome-wide association studies (GWAS), genetic variations have received more attention and the identification of disease-related genetic variations has improved our understanding in the pathway of disease development and provided tons of potential targets for the therapeutic treatment. Host genetic susceptibility has also been demonstrated as an important factor which determines the outcome of HBV infection, such as single nucleotide polymorphism (SNPs)

and gene copy number variations (CNVs) (He et al., 2006; Li et al., 2017a; Qiu et al., 2017), but is not well documented.

SNP, the most common type of genetic variation among people, is a variation in a single nucleotide that occurs within a gene or in a regulatory region near a gene (NIH, 2018). As most popular class of genetic variation among people, SNPs were demonstrated affect disease progression of HBV infection in many studies (Chang et al., 2014; Komatsu et al., 2014; Li et al., 2017b; Tai et al., 2017). CNVs, another important type of genetic variation, received more attention recently. CNVs range from 1 Kbp to several Mbp, are defined as deletions or duplications of genome segments. Recently, CNVs were believed to be the causes of many human diseases and were associated with the susceptibility of various liver diseases, including HCV infection and Hepatocellular carcinoma (Budzko et al., 2016; Zhou et al., 2017). To date, the association of CNV with chronic Hepatitis B infection are still not systematically documented.

To identify the CNVs associated with susceptibility to chronic HBV infection and outcomes, we recruited 2,689 CHB patients and 1,200 healthy controls from Qidong Liver Cancer Institute in Qidong County, China. By using the computational tool PennCNV and ParseCNV, we analyzed the genotyping data generated by Illumina Human OmniExpress BeadChip and OmniZhonghua BeadChip and evaluated the association of SNPs and CNVs with CHB susceptibility. We identified novel CHB associated CNVs in genes, which advanced the understanding of genetic factors influencing HBV infection and outcome and provided the potential targets related with HBV infection which have potential implication in prevention and clinical care of HBV infection and CHB patients.

Research Methods

Populations

The 2,689 CHB patients and 1,200 healthy controls used in the initial GWAS scan were recruited by Qidong Liver Cancer Institute in Qidong County, Jiangsu Province in eastern China, during the period from May 2006 to December 2012. CHB patients were positive for both hepatitis B surface antigen (HBsAg) and antibody immunoglobulin G to hepatitis B core antigen for at least 6 months. The controls were collected during the same period time as the CHB cases were enrolled. They were randomly selected from a pool of healthy volunteers who visited the Qidong Liver Cancer Institute for their routinely scheduled physical examinations. Both the cases and controls were residents in Qidong County, which is one of the highest endemic regions for chronic hepatitis B virus (HBV) infection and HBV-related hepatocellular carcinoma (HCC) in China. For each subject, 3 ml whole blood was obtained for extraction of genomic DNA and 3 ml serum sample was obtained for detection of HBsAg and antibodies to hepatitis C virus (HCV), and human immunodeficiency virus (HIV) using an enzyme linked immunosorbent assay (ELISA). The characteristic information for each subject, including age, gender, ethnicity, HBV infection status and history of other liver diseases were also collected with a standard interviewer-administered questionnaire and/or from medical records. The exclusion criteria for the cases were 1) positive for antibodies to HCV, or HIV, 2) with a history of any other types of liver disease, such as autoimmune hepatitis, toxic hepatitis, and primary biliary cirrhosis, 3) not residents in Qidong County. The exclusion criteria for the controls, in addition to the three items of exclusion criteria for the cases, were 1) positive

for HBsAg, 2) with a history of HBV vaccination, 3) with a history of CHB, 4) with a history of HCC, 5) less than 30 years old.

Genotyping and quality control in GWAS

The genome-wide genotyping analysis was conducted using Illumina Human OmniExpress BeadChip (delivering superior power for GWAS, providing high sample throughput with comprehensive genomic content and including 733,202 single-nucleotide polymorphisms [SNPs] across the genome) for cases and OmniZhonghua BeadChip (delivering exceptional coverage of common, intermediate, and rare variation found within Chinese populations for GWAS and including 900,015 SNPs, containing ~80% of the SNPs in HumanOmniExpress BeadChip) for controls at Genergy Biotechnology Inc. (Shanghai, China). Genotyping was performed according to the Infinium HD protocol from Illumina (San Diego, CA).

During individual quality control of genotyping data, samples were removed if they (i) had an overall genotyping rate of <95%; or (ii) were duplicates or showed familial relationships ($PI_HAT > 0.025$). SNPs will be excluded if they had (i) a call rate of <95%; (ii) a minor allele frequency (MAF) of <0.01; or (iii) $P < 1 \times 10^{-3}$ in a Hardy-Weinberg equilibrium test among controls.

During individual quality control analysis of CNV calling, samples with highly variable signal intensity were removed, including (i) LRR_SD (standard deviation of Log R Ratio) >0.3; (ii) BAF_drift (measuring departure of the B Allele Frequency from the

expected values) > 0.01 ; (iii) WF(waviness factor, the amount of dispersion in signal intensity) >0.05 or $WF < -0.05$; (iv) Number of CNVs >50 (number of called-CNVs).

CNV calling

CNV Calling Raw signal intensity files were first generated by the export function provided in Illumina GenomeStudio. Then we used UCSC Genome Browser's liftOver tool to map SNPs to the newer reference genome assembly (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>). PennCNV is a software tool for Copy Number Variation (CNV) detection from SNP genotyping arrays. By removing probes that can not be uniquely mapped to the genome and choosing proper population frequency of B allele (PFB) and GCmodel files, we used PennCNV to identify CNVs and generated a quality control summary for each sample (Lin CF et al., 2013).

Quality Controls in CNV calling

During CNV calling, CNVs were removed if (i) number of SNPs spanning less than 10; (ii) length of CNV <50 kb. We also removed the CNVs calls located in certain regions, including HLA regions, immunoglobulin regions, telomere regions, centromere regions. Previous studies show that these regions were especially likely to harbor spurious CNV calls (Wang et al., 2007).

Statistical Analysis

Predicted CNVs from the sample genomic locus can have various start and end points across individuals. To make the analysis simpler, we divided the genome into copy number polymorphic regions (CNPRs). Each individual was then assigned a copy number (CN) state for each CNPR according to the CNV predicted in that region, with CN = 2 if no CNV was predicted. In order to distinguish the CNV with a single duplication of one allele and single deletion of the other (CN=2) from non-CNV status (CN=2), we further encoded CN state with deletion and duplication variables. Plink 1.09 was used to perform association analysis (Purcell S et al., 2007). The outcome variable was a binary variable indicating HBV infection. Explanatory variables were age, gender, and CNV type. Logistic regression analysis was used to adjust for age, gender as well as CNV type. Fisher exact tests were used if 25% of cells with expected count were less than 5. If all expected counts were greater or equal to 5, Chi-squared tests were used to compare categorical variables. Bonferroni tests were used to adjust for multiple comparison.

Results

Patient characteristics

The characteristics of the recruited patients were summarized in Table 1. There were 2,689 CHB patients (69.14%) and 1,200 healthy controls (30.86%). However, 70 healthy controls and 181 CHB patients were found missing the age and gender information and were excluded for analysis. The total number of healthy controls used in the analysis was 1,130, while the number of CHB patients was 2,508. The median age for normal people

was 52 years (range, 30-80 years), the median age for CHB patients was 50 years (range, 15-87 years). There were 722 males (63.89%) and 408 females (36.11%) among the healthy controls, while 1,818 males (72.49%) and 690 females (27.51%) among the CHB patients.

Table 1. Patient characteristics

Group	Variable	No. (%)
Control (n=1,200)	Median age, years (range)	52 (30 to 80)
	Male	722 (63.89)
	Female	408 (36.11)
	Missing	70 (5.83)
	Final Num	1,130
CHB patients (n=2,689)	Median age, years (range)	50 (15 to 87)
	Male	1,818 (72.49)
	Female	690 (27.51)
	Missing	181 (6.73)
	Final Num	2,508

The distributions of age and gender of controls and CHB patients were shown in the Figure. 1.

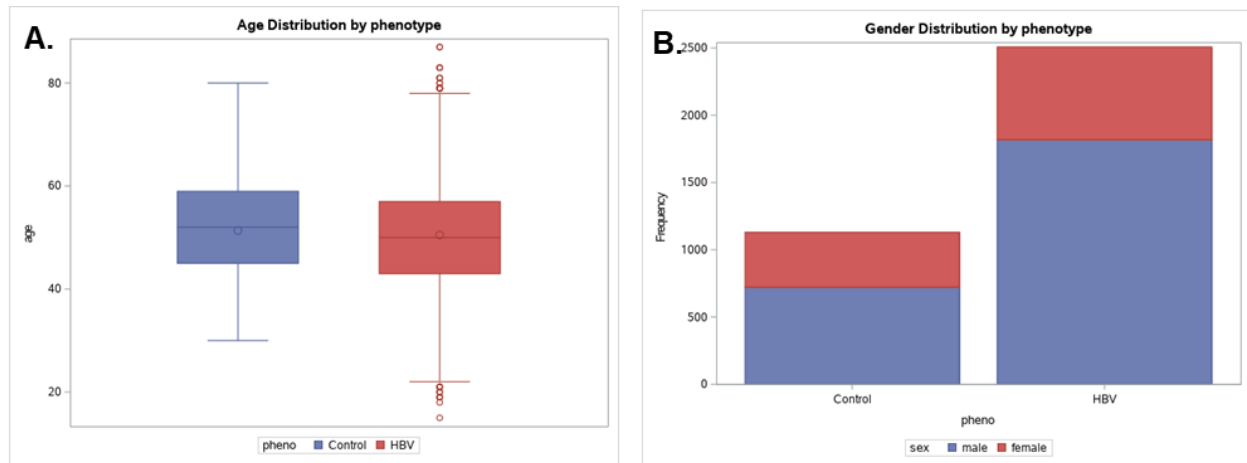


Figure 1. Age and Gender distribution of normal controls and CHB patients. (A) The boxplot of age distribution of normal controls and CHB patients. (Blue- normal controls; Red- CHB patients). (B) Gender distribution of normal controls and CHB patients. (Blue- male; Red- female).

SNP-based CNV Analysis

We implemented the case-control association analysis following the PennCNV introduction (PennCNV) which were used to identify the stretch of SNPs that tend to have copy number changes in CHB cases versus controls by Fisher's Exact Test. The P-values derived from the Fisher Exact Test comparing CHB cases and controls, were generally shown by Manhattan plot. The Manhattan plot was produced by scattering the P-values

in $-\log_{10}$ scale in the vertical axis and the physical position of SNP along chromosomes in the horizontal axis. Different chromosomes were generally distinguished with colors. Using $-\log_{10}$ scale was to highlight the small P-values, which suggested potential disease-related SNPs (Zeng et al., 2015). The horizontal solid red line marked a significance cut off of 10^{-8} , and the horizontal dashed blue line marked a significance cut off of 10^{-6} . As shown in figure 2, the Manhattan plot displayed the $-\log_{10}$ deviance P value for CNV losses in the SNP-based CNV test. We identified 110 loci ($p < 10^{-8}$) with deletion enrichment in the CHB cases. While the Figure. 3 displayed the $-\log_{10}$ deviance P value for CNV gains in the SNP-based CNV test. There were 124 loci ($p < 10^{-8}$) with duplication enrichment in the CHB cases were found here.

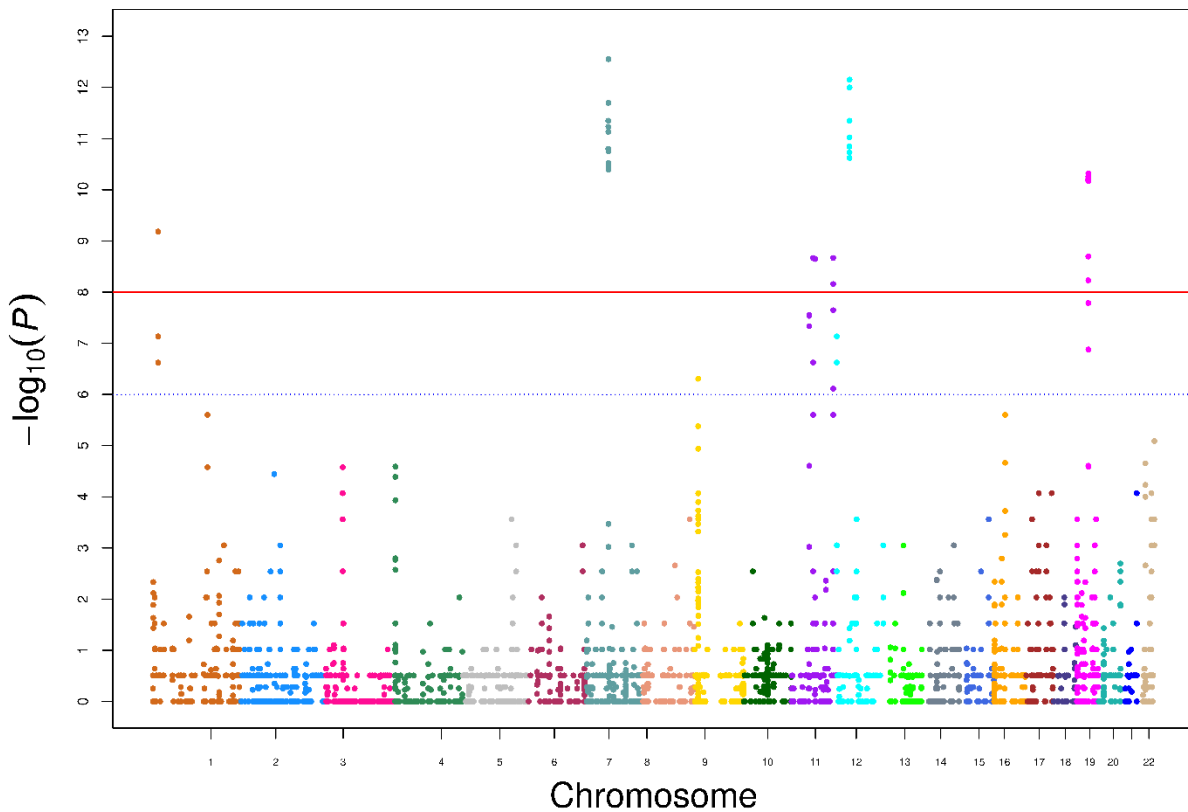


Figure 2. SNP-based CNV Manhattan plot. Manhattan plot displayed the $-\log_{10}$ deviance P value for CNV losses in the SNP-based CNV test. The y-axis showed the distribution of $-\log_{10}(p)$ where p was the Fisher's Exact Test P-value for copy number changes in CHB cases versus controls. P value cutoffs corresponding to 10^{-6} and 10^{-8} were highlighted in red solid line and blue dashed line, respectively. The x-axis showed chromosomes numbered from 1 (left) to X (right).

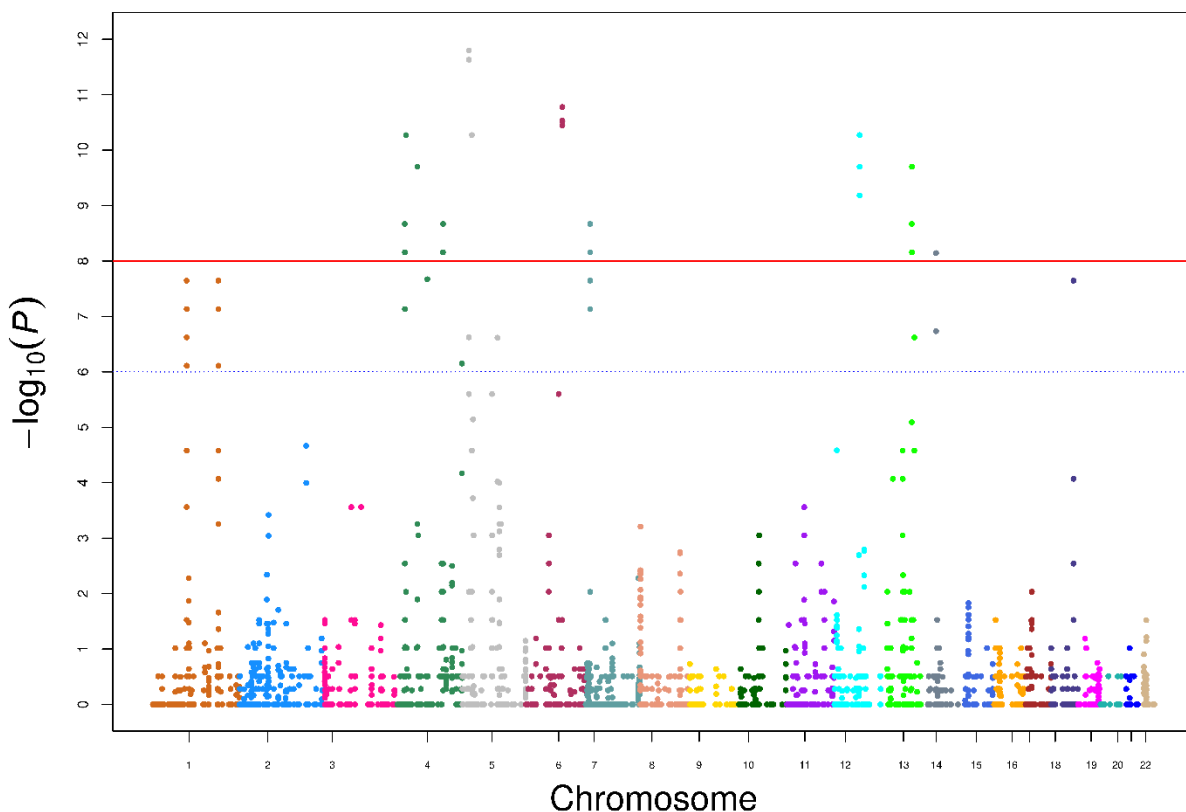


Figure 3. SNP-based CNV Manhattan plot. Manhattan plot displayed the $-\log_{10}$ deviance P value for CNV gains in the SNP-based test. The y-axis showed the distribution of $-\log_{10}(p)$ where p was the Fisher's Exact Test P-value for copy number changes in CHB cases versus controls. P value cutoffs corresponding to 10^{-6} and 10^{-8} were highlighted in red

solid line and blue dashed line, respectively. The x-axis showed chromosomes numbered from 1 (left) to X (right).

CNPR-based CNV Analysis

To further identify the specific loci that confer risk of HBV infection, we performed the CNPR-based CNV analysis and found 8 deletion CNPRs, as well as 3 duplication CNPRs were significantly changed between CHB patients and healthy controls. Moreover, there were nine genes revealed the copy number loss, including *FGFR3* ($p=1.49 \times 10^{-7}$), *FGR-3* ($p=1.49 \times 10^{-7}$), *LETM1* ($p=1.49 \times 10^{-7}$), *TACC3* ($p=1.49 \times 10^{-7}$), *TMEM129* ($p=1.49 \times 10^{-7}$), *PANK4* ($p=4.55 \times 10^{-4}$), *PLCH2* ($p=4.55 \times 10^{-4}$), *CED-6* ($p=2.04 \times 10^{-4}$), *DIRC1* ($p=2.04 \times 10^{-4}$), as well as three genes revealed the copy number gain, including *FLJ43080* ($p=2.50 \times 10^{-5}$), *CSMD3* ($p=6.288 \times 10^{-5}$), *MGAT4C* ($p=1.52 \times 10^{-4}$) in CHB patients compared with healthy controls. Besides the CNPRs in genes, the intergenic regions chr19:32757923-32761177 ($p=8.97 \times 10^{-54}$), chr11:50432844-50586426 ($p=7.76 \times 10^{-12}$), chr9:12003711-12010873 ($p=5.79 \times 10^{-8}$), chr19:24244157-24354405 ($p=4.55 \times 10^{-4}$), chr5:18673077-18728095 ($p=9.02 \times 10^{-38}$) were also associated with CHB. Among these genes, the function of *FGFR3* and *CSMD3* have been demonstrated by other researchers related with the HBV infection (Lai et al., 2016; Van et al., 2016); while, the other genes were novel and not reported. The detailed information was listed in the Table 2.

Table 2. Gene-based CNV Analysis

CNPR*	N in Cases	N in Control	Nearby Gene	Distance from Nearby Gene, bp	P**
Deletion					
chr19:32757923-32761177	243	0	AK075337	60054	8.97E-54
chr11:50432844-50586426	52	0	AB231715, AB231716, AB231717	96465	7.76E-12
chr9:12003711-12010873	77	11	TYRP1	672513	5.79E-08
			FGFR3 , FGR-3, LETM1, TACC3,		
chr4:1690555-1785787	32	0	TMEM129	0	1.49E-07
chr1:2369108-2444569	16	0	PANK4, PLCH2	0	4.55E-04
chr19:24244157-24354405	16	0	LOC100101266	106068	4.55E-04
chr5:18673077-18728095	1	94	BC028204	706723	9.02E-38
chr2:189165214-189310804	2	13	CED-6, DIRC1	0	2.04E-04
Duplication					
chr5:109856282-109991005	22	0	FLJ43080	0	2.50E-05
chr8:114024371-114181033	36	4	CSMD3	0	6.288E-05
chr12:85169592-85337570	19	0	MGAT4C	0	1.52E-04

*CNPR, copy number polymorphic region (hg18)

**Fisher exact test was used and used 5×10^{-4} as the significant cutoff. 5×10^{-4} is a conservative bar for CNV genome-wide significance surviving multiple testing correction based on analysis of Illumina and Affymetrix genome-wide SNP arrays.

Red highlight: Genes have been linked with HBV previously

Discussion

Accounting for one third of chronic hepatitis B virus infection occurrence in the world, the chronic hepatitis B infection and HBV-related complications has become large disease burden in China. There are more than 90 million chronic carriers of HBV in China and most people with HBV infection in China are unaware that they carry the disease (Chen, 2018). HBV is a member of small DNA viruses and infects one major cell of liver, hepatocytes. Using a RNA proviral intermediate, HBV replicates by reverse transcription (Beck and Nassal, 2007). It is known that HBV infection can result in various clinical conditions and approximately 90% of neonatal HBV infection and 30% of childhood infection will finally develop to CHB (Yan et al., 2014). The main risk factors related with CHB include familial spread, infant infection, infection due to immunologic inadequacy, and a history of other liver disease (Qiu et al., 2017). The causes of the CHB are intricated and remain largely unknown.

It has been demonstrated that host genetic susceptibility is an important factor which determines the outcome of HBV infection(He et al., 2006; Li et al., 2017a; Qiu et al., 2017). In 2013, Chiao-Feng et al proposed a method which make us possible to call CNVs directly from the GWAS genotyping data (Lin et al., 2013). And recently, by using this method, some CNVs were firstly identified to be associated with schizophrenia and hepatocellular carcinoma (Purcell et al., 2007; Sakai et al., 2015; Zhou et al., 2017). But CHB related CNVs are rarely reported. In the current study, we analyzed the genotyping data generated from CHB patients and normal controls and found 8 deletion CNPRs, as well as 3 duplication CNPRs are significantly changed between CHB patients and healthy controls. Moreover, there were nine genes revealed the copy number loss, including

FGFR3 ($p=1.49 \times 10^{-7}$), FGR-3 ($p=1.49 \times 10^{-7}$), LETM1 ($p=1.49 \times 10^{-7}$), TACC3 ($p=1.49 \times 10^{-7}$), TMEM129 ($p=1.49 \times 10^{-7}$), PANK4 ($p=4.55 \times 10^{-4}$), PLCH2 ($p=4.55 \times 10^{-4}$), CED-6 ($p=2.04 \times 10^{-4}$), DIRC1 ($p=2.04 \times 10^{-4}$), as well as three genes revealed the copy number gain, including FLJ43080 ($p=2.50 \times 10^{-5}$), CSMD3 ($p=6.288 \times 10^{-5}$), MGAT4C ($p=1.52 \times 10^{-4}$) in CHB patients compared with healthy controls. Besides the CNPRs in genes, the intergenic regions chr19:32757923-32761177 ($p=8.97 \times 10^{-54}$), chr11:50432844-50586426 ($p=7.76 \times 10^{-12}$), chr9:12003711-12010873 ($p=5.79 \times 10^{-8}$), chr19:24244157-24354405 ($p=4.55 \times 10^{-4}$), chr5:18673077-18728095 ($p=9.02 \times 10^{-38}$) were also associated with CHB. Among these genes, the function of FGFR3, LETM1, CED-6 and CSMD3 have been demonstrated by other researchers related with the HBV infection. Although other genes' functions in the HBV infection and CHB development haven't been demonstrated, they could be the potential targets for CHB prevention and therapy.

Interestingly, FGFR3, a protein coding gene, makes a protein called Fibroblast Growth Factor Receptor 3. The FGFR3 protein plays an important role in cellular processes, including regulation of cell growth and proliferation, determination of cell type, angiogenesis, wound healing and so on. Related studies found over-expressed FGFR3 levels in human hepatocellular carcinoma (Qiu et al., 2005). Moreover, FGFR3 has been well documented in various cancers and suggested as a therapeutic target, including bladder cancer (Gust et al., 2013), lung cancer (Yin et al., 2016) and urothelial carcinoma (Kim et al., 2018). In this study, we found significant deletion of copy number in FGFR3 in CHB cases compared with healthy controls which indicated the important role of FGFR3 in CHB disease progress and could be a potential risk factor for the hepatocellular carcinoma development.

CSMD3 (CUB and Sushi multiple domains 3) is a gene encodes a protein and functions as a tumor suppressor. Lai et.al created a transgenic mouse models by transferring the HBV pre-S/S gene and its promoter into C57B6 mice and found decreased CSMD3 expression in the mice models which indicated that downregulation of CSMD3 may contribute to the hepatocarcinogenesis. We found significant duplication of copy number in CSMD3 in CHB cases compared with healthy controls which confirmed the role of CSMD3 in HBV related hepatocarcinogenesis.

Although, there were 8 genes revealed copy number deletion and 2 genes revealed copy number duplications that were not linked with HBV infection or CHB previously. Our findings also provided rational targets for the disease prevention. For example, LETM1 (Leucine Zipper And EF-Hand Containing Transmembrane Protein 1), involved in the transport of charged calcium atoms (calcium ions) across membranes within mitochondria, was found been upregulated in the breast cancer (Li et al., 2015) Besides the LETM1, FGR, DIRC1, TACC3 and MGAT4C were also found related with various type of cancer. For example, FGR was over-expressed in the primary tumor samples (Stransky et al., 2014); overexpression of DIRC1 was associated with tumor progression in gastric cancer(Li et al., 2018); TACC3 was found involved in proliferation and differentiation of tumor cells, cancer progression and metastasis (Zhao et al., 2018). Moreover, TMEM129 were found related with the virus infection. TMEM129, as a novel E3 ligase, is hijacked by the human cytomegalovirus and play role in averting immune recognition of the infected cell (van den Boomen et al., 2014). Our results manifested these genes might be potential targets of HBV infection and related disease. The impact of copy number variations on gene expression complex and haven't been fully elucidated. But the CHB-

related CNVs identified by us could be potential targets. The further confirmation and identification of these genes are needed which help elucidate underlying mechanisms of HBV infection and CHB progression.

There are also several limitations of our study. Firstly, we do not have access to information about the medical information, education, risk behaviors of subjects in the clinical cohort. It is possible that these risk factors affect our results and the associations we saw in our study are due to population stratification. Secondly, our population-based study was performed exclusively in Chinese individuals and thus might not be generalizable to other ancestries. Thirdly, our findings in this study may need further validation by using other independent samples.

Conclusions

We analyzed the genotyping data generated from CHB patients and normal controls and identified 110 loci ($p < 10^{-8}$) with deletion enrichment and 124 loci ($p < 10^{-8}$) with duplication enrichment in the CHB cases. We also found nine genes revealed the copy number loss, including FGFR3 ($p = 1.49 \times 10^{-7}$), FGR-3 ($p = 1.49 \times 10^{-7}$), LETM1 ($p = 1.49 \times 10^{-7}$), TACC3 ($p = 1.49 \times 10^{-7}$), TMEM129 ($p = 1.49 \times 10^{-7}$), PANK4 ($p = 4.55 \times 10^{-4}$), PLCH2 ($p = 4.55 \times 10^{-4}$), CED-6 ($p = 2.04 \times 10^{-4}$), DIRC1 ($p = 2.04 \times 10^{-4}$), as well as three genes revealed the copy number gain, including FLJ43080 ($p = 2.50 \times 10^{-5}$), CSMD3 ($p = 6.288 \times 10^{-5}$), MGAT4C ($p = 1.52 \times 10^{-4}$) in CHB patients compared with healthy controls, which providing novel diagnosis markers and targets in HBV infection and CHB progression.

Ethics

Informed consent was obtained from all subjects before their participation in the study. The study was approved by the ethics committee of Qidong Liver Cancer Institute and conducted in accord with the Declaration of Helsinki principles.

References

- Beck, J., and M. Nassal. 2007. Hepatitis B virus replication. *World J Gastroenterol.* 13:48-64.
- Budzko, L., M. Marcinkowska-Swojak, P. Jackowiak, P. Kozlowski, and M. Figlerowicz. 2016. Copy number variation of genes involved in the hepatitis C virus-human interactome. *Sci Rep.* 6:31340.
- CDC. 2018a. Hepatitis B Questions and Answers for Health Professionals. Vol. 2018. CDC, CDC.
- CDC. 2018b. Hepatitis B Questions and Answers for the Public.
- Chang, S.W., C.S. Fann, W.H. Su, Y.C. Wang, C.C. Weng, C.J. Yu, C.L. Hsu, A.R. Hsieh, R.N. Chien, C.M. Chu, and D.I. Tai. 2014. A genome-wide association study on chronic HBV infection and its clinical progression in male Han-Taiwanese. *PLoS One.* 9:e99724.
- Chen, S.L., J; Wang, D; Fung, H; Wong, L; Zhao, L. 2018. The hepatitis B epidemic in China should receive more attention. *The Lancet.* 391:P1572.

- Gust, K.M., D.J. McConkey, S. Awrey, P.K. Hegarty, J. Qing, J. Bondaruk, A. Ashkenazi, B. Czerniak, C.P. Dinney, and P.C. Black. 2013. Fibroblast growth factor receptor 3 is a rational therapeutic target in bladder cancer. *Mol Cancer Ther.* 12:1245-1254.
- He, Y.L., Y.R. Zhao, S.L. Zhang, and S.M. Lin. 2006. Host susceptibility to persistent hepatitis B virus infection. *World J Gastroenterol.* 12:4788-4793.
- Kim, Y.S., K. Kim, G.Y. Kwon, S.J. Lee, and S.H. Park. 2018. Fibroblast growth factor receptor 3 (FGFR3) aberrations in muscle-invasive urothelial carcinoma. *BMC Urol.* 18:68.
- Komatsu, H., J. Murakami, A. Inui, T. Tsunoda, T. Sogo, and T. Fujisawa. 2014. Association between single-nucleotide polymorphisms and early spontaneous hepatitis B virus e antigen seroconversion in children. *BMC Res Notes.* 7:789.
- Lai, M.W., K.H. Liang, W.R. Lin, Y.H. Huang, S.F. Huang, T.C. Chen, and C.T. Yeh. 2016. Hepatocarcinogenesis in transgenic mice carrying hepatitis B virus pre-S/S gene with the sW172* mutation. *Oncogenesis.* 5:e273.
- Li, F., X. Li, G.Z. Zou, Y.F. Gao, and J. Ye. 2017a. Association between TLR7 copy number variations and hepatitis B virus infection outcome in Chinese. *World J Gastroenterol.* 23:1602-1607.
- Li, H., J. Chen, R. Zhang, R. Xu, Z. Zhang, L. Ren, Q. Yang, Y. Tian, and D. Li. 2017b. Single nucleotide polymorphisms in ZNF208 are associated with increased risk for HBV in Chinese people. *Oncotarget.* 8:112451-112459.

- Li, N., Y. Zheng, C. Xuan, Z. Lin, L. Piao, and S. Liu. 2015. LETM1 overexpression is correlated with the clinical features and survival outcome of breast cancer. *Int J Clin Exp Pathol.* 8:12893-12900.
- Li, Z., A.J. Yang, F.M. Wei, X.H. Zhao, and Z.Y. Shao. 2018. Significant association of DIRC1 overexpression with tumor progression and poor prognosis in gastric cancer. *Eur Rev Med Pharmacol Sci.* 22:8682-8689.
- Lin, C.F., A.C. Naj, and L.S. Wang. 2013. Analyzing copy number variation using SNP array data: protocols for calling CNV and association tests. *Curr Protoc Hum Genet.* 79:Unit 1 27.
- NIH. 2018. What are single nucleotide polymorphisms (SNPs)? Vol. 2018.
- PennCNV. CNV case-control comparison. Vol. 2019.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81:559-575.
- Qiu, B., W. Jiang, M. Olyaei, K. Shimura, A. Miyakawa, H. Hu, Y. Zhu, and L. Tang. 2017. Advances in the genome-wide association study of chronic hepatitis B susceptibility in Asian population. *Eur J Med Res.* 22:55.
- Qiu, W.H., B.S. Zhou, P.G. Chu, W.G. Chen, C. Chung, J. Shih, P. Hwu, C. Yeh, R. Lopez, and Y. Yen. 2005. Over-expression of fibroblast growth factor receptor 3 in human hepatocellular carcinoma. *World J Gastroenterol.* 11:5266-5272.
- Sakai, M., Y. Watanabe, T. Someya, K. Araki, M. Shibuya, K. Niizato, K. Oshima, Y. Kunii, H. Yabe, J. Matsumoto, A. Wada, M. Hino, T. Hashimoto, A. Hishimoto, N.

- Kitamura, S. Iritani, O. Shirakawa, K. Maeda, A. Miyashita, S. Niwa, H. Takahashi, A. Kakita, R. Kuwano, and H. Nawa. 2015. Assessment of copy number variations in the brain genome of schizophrenia patients. *Mol Cytogenet.* 8:46.
- Stransky, N., E. Cerami, S. Schalm, J.L. Kim, and C. Lengauer. 2014. The landscape of kinase fusions in cancer. *Nat Commun.* 5:4846.
- Tai, D.I., W.J. Jeng, and C.Y. Lin. 2017. A global perspective on hepatitis B-related single nucleotide polymorphisms and evolution during human migration. *Hepatology Commun.* 1:1005-1013.
- van den Boomen, D.J., R.T. Timms, G.L. Grice, H.R. Stagg, K. Skodt, G. Dougan, J.A. Nathan, and P.J. Lehner. 2014. TMEM129 is a Derlin-1 associated ERAD E3 ligase essential for virus-induced degradation of MHC-I. *Proc Natl Acad Sci U S A.* 111:11425-11430.
- Van, N.D., C.S. Falk, L. Sandmann, F.W. Vondran, F. Helfritz, H. Wedemeyer, M.P. Manns, S. Ciesek, and T. von Hahn. 2016. Modulation of HCV reinfection after orthotopic liver transplantation by fibroblast growth factor-2 and other non-interferon mediators. *Gut.* 65:1015-1023.
- Wang, K., M. Li, D. Hadley, R. Liu, J. Glessner, S.F. Grant, H. Hakonarson, and M. Bucan. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17:1665-1674.

- WHO. 2016. Up to 10 million people in China could die from chronic hepatitis by 2030 – Urgent action needed to bring an end to the ‘silent epidemic’. Vol. 2018. WHO, WHO.
- WHO. 2018. Hepatitis B. Vol. 2018. WHO, WHO.
- Yan, Y.P., H.X. Su, Z.H. Ji, Z.J. Shao, and Z.S. Pu. 2014. Epidemiology of Hepatitis B Virus Infection in China: Current Status and Challenges. *J Clin Transl Hepatol.* 2:15-22.
- Yin, Y., X. Ren, C. Smith, Q. Guo, M. Malabunga, I. Guernah, Y. Zhang, J. Shen, H. Sun, N. Chehab, N. Loizos, D.L. Ludwig, and D.M. Ornitz. 2016. Inhibition of fibroblast growth factor receptor 3-dependent lung adenocarcinoma with a human monoclonal antibody. *Dis Model Mech.* 9:563-571.
- Zeng, P., Y. Zhao, C. Qian, L. Zhang, R. Zhang, J. Gou, J. Liu, L. Liu, and F. Chen. 2015. Statistical analysis for genome-wide association study. *J Biomed Res.* 29:285-297.
- Zhao, C., X. He, H. Li, J. Zhou, X. Han, D. Wang, G. Tian, and F. Sui. 2018. Downregulation of TACC3 inhibits tumor growth and migration in osteosarcoma cells through regulation of the NF-kappaB signaling pathway. *Oncol Lett.* 15:6881-6886.
- Zhou, C., W. Zhang, W. Chen, Y. Yin, M. Atyah, S. Liu, L. Guo, Y. Shi, Q. Ye, Q. Dong, and N. Ren. 2017. Integrated Analysis of Copy Number Variations and Gene Expression Profiling in Hepatocellular carcinoma. *Sci Rep.* 7:10570.

Capstone Experience Reflection

My experience with the placement site

It's my great honor to work in the Southern Medical University Nanfang Hospital and thank my preceptor gave me this opportunity to work with him. I believe what I learnt during this period will benefit me a lot in my future works.

As mentioned previously, chronic hepatitis B virus infection is a serious disease burden and public health issue in China. I am very interested in CHB related research and public health activities. Nanfang Hospital is a large-scale comprehensive tertiary hospital with medical, teaching, scientific research and preventive health care. They did a lot of efforts on the HBV infection prevention and therapy. As I have known, the Nanfang hospital provides three years' free therapy to the HBV patients and arranges HBV prevention-related activities every year. They also give strong support to the HBV-related researches and made outstanding contributions in understanding the mechanisms of HBV infection and CHB progression.

I started the work of identifying the CHB associated CNVs from genotyping data with my preceptor on August, 2018 and it has been half year. During this period, I learnt the genome-wide association study (GWAS) for my first time. GWAS studies are very complex for me and I have no idea about its principle, application and method at the beginning. After the learning during this period, I realized that GWAS, an observational study of a genome-wide set of genetic variants in different individuals, is very powerful and useful tool in detecting common genetic variants associated with complex disorders. The study design and statistical method for GWAS are very interesting for me. Although I didn't have chance to be involved in the data collection process (i.e., recruiting

participants, collecting samples, and generating data), I still went over thousands of human data, which was not expected by me. It's my first time to handle such a large size of data. Moreover, I can't physically work in the Nanfang hospital as my full-time work in the United States, the distance and jet lag were all the difficult part for moving this project. I very appreciate the help from my preceptor who work in the late night to help me understand this project and solve the problems. I also learnt new software this period which I find is very useful, including Xshell and PennCNV. Xshell is a powerful terminal emulator from South Korea, supports SSH1, SSH2, FTP, SFTP, Telnet, rlogin, and Serial protocols. It enables users to directly connect to the remote hosting via Internet, so as to easily control Linux servers on Windows. By using the Xshell, I can access the data in Nanfang Hospital using my laptop and work on the analysis easily. PennCNV is an excellent tool for Copy Number Variation (CNV) detection from SNP genotyping arrays. It can be used to identify a stretch of SNPs that tend to have copy number changes in cases versus controls using Fisher's Exact Test. By using these two softwares, we found interesting CNVs associated with the susceptibility of CHB.

Greatest contributions/accomplishments

The greatest accomplishments I made from my capstone project are helping in the data cleaning, SNP-based CNV analysis, gene-based CNV analysis and give suggestions in the statistical method. We analyzed the genotyping data generated from CHB patients and normal controls and identified 110 loci ($p < 10^{-8}$) with deletion enrichment and 124 loci ($p < 10^{-8}$) with duplication enrichment in the CHB cases. We also found nine genes revealed the copy number loss, including FGFR3 ($p = 1.49 \times 10^{-7}$), FGR-3 ($p = 1.49 \times 10^{-7}$),

LETM1 ($p=1.49 \times 10^{-7}$), TACC3 ($p=1.49 \times 10^{-7}$), TMEM129 ($p=1.49 \times 10^{-7}$), PANK4 ($p=4.55 \times 10^{-4}$), PLCH2 ($p=4.55 \times 10^{-4}$), CED-6 ($p=2.04 \times 10^{-4}$), DIRC1 ($p=2.04 \times 10^{-4}$), as well as three genes revealed the copy number gain, including FLJ43080 ($p=2.50 \times 10^{-5}$), CSMD3 ($p=6.288 \times 10^{-5}$), MGAT4C ($p=1.52 \times 10^{-4}$) in CHB patients compared with healthy controls, which providing novel diagnosis markers and targets in HBV infection and CHB progression, which providing novel diagnosis markers and targets in HBV infection and CHB progression.

I also worked on the Patient characteristics. We previously had 2,689 CHB patients and 1,200 healthy controls. We didn't perform the all the patient's data for this study, because some of patients' information was missing. I worked on clean the data to narrow the population in this study to 2,508 CHB patients and 1,130 healthy controls.

As I have the pharmacology background previously, I contributed in validating the pharmacology functions of associated genes identified from our analysis. The GWAS data is useless if we don't know the identified genes' biology functions. My participation helped make the findings more meaningful.

Greatest challenges

The greatest challenges we met is the distance and jet leg during our communication and work. I couldn't physically work in Nanfang hospital which do bring inconvenience during this period. At the beginning, we couldn't find proper time zone for the discussion and I have the difficult to access the database in Nanfang hospital. In the other hand, the GWAS is a brand new field for me. I need spend lots of time on reading literatures to help me understand the purpose, principle, methods in this study. I worried a lot at the

beginning as the project didn't move smoothly. I appreciate the tolerance and patience from my preceptor. My preceptor gave me lots of encourage and helped me on every small issue I met. We communicated every week and found best time work for both of us. With the help from my preceptor, I learnt to use the Xshell and finally accessed the data base in Nanfang hospital successfully. Although I had a hard time at the beginning, the long-distance work experiences make me stronger and push me learn more skills, including time planning, communications, terminal emulator tools and so on.

Views of public health practice

Public health practice can't ignore the impact of genomics in the 21st century. I didn't realize the importance of genomics for public health practice before. After my capstone project, I have understand better how the genomics impact the public health practice. With the rapid advance of genomics technologies, predication of individual risks by genomics studies could affect clinical therapy strategies. Developments in public health practice will be necessary to ensure rapid and effective implementation of genomics studies.

Public health education

My public health education help me understand better the ethical issues in the GWAS. GWAS data is derived from DNA information which is a powerful personal identifier and provide information not just on the individual, but also on the individual's relatives, related groups, and population. The GWAS data contains large amount of individual-specific digital information and can be easily share across the internet. To understand better the

ethical issues in GWAS, I read lots of literatures and found there are three important aspects of the context of genome research methods that may affect the ethical challenges, including one data using in many different projects; data sharing; unauthorized use of digital information. In our study, we don't have access of patients' medical history, family information, personal information. So all of these factors are not used in the analysis.

Acknowledgements

We are indebted to the patients for their participation and to the physicians involved in this study. This work was supported by Nanfang hospital. We thank Nanfang hospital providing the data and software for analysis.