A genome-wide association study identified new variants associated with the risk of chronic hepatitis B

Yoon Jun Kim, Hwi Young Kim, Jeong-Hoon Lee, Su Jong Yu, Jung-Hwan Yoon, Hyo-Suk Lee, Chung Yong Kim, Jae Youn Cheong, Sung Won Cho, Neung Hwa Park, Byung Lae Park, Seok Namgoong, Lyoung Hyo Kim, Hyun Sub Cheong and Hyoung Doo Shin

1Department of Internal Medicine and Liver Research Institute, Seoul National University, Seoul, Republic of Korea
2Department of Internal Medicine, Seoul National University Boramae Medical Center, Seoul, Republic of Korea
3Department of Gastroenterology, Ajou University School of Medicine, Suwon, Republic of Korea
4Department of Internal Medicine, Ulsan University Hospital, Ulsan, Republic of Korea
5Department of Genetic Epidemiology, SNP Genetics, Inc., Sogang University, Seoul, Republic of Korea
6Department of Life Science, Sogang University, Seoul, Republic of Korea

Received February 4, 2013; Revised May 23, 2013; Accepted June 4, 2013

Hepatitis B virus (HBV) infection is the predominant risk factor for chronic hepatitis B (CHB), liver cirrhosis (LC) and hepatocellular carcinoma (HCC). Recently, several genome-wide association studies (GWASs) of CHB identified human leukocyte antigen (HLA) loci, including HLA-DP and HLA-DQ in Asian populations, as being associated with the risk of CHB. To confirm and identify the host genetic factors related to CHB infection, we performed another GWAS using a higher-density chip in Korean CHB carriers. We analyzed 1400 samples from Korean population (400 CHB cases and 1000 population controls) using a higher-density GWAS chip [1140419 single nucleotide polymorphisms (SNPs)]. In subsequent replication analysis, we further analyzed in an independent study of a Korean CHB cohort consisting of 2909 Korean samples (971 cases and 1938 controls). Logistic regression methods were used for statistical analysis adjusting for age and sex as covariates. This study identified two new risk-associated loci for CHB on the HLA region of chromosome 6, e.g. rs652888 on euchromatic histone-lysine-methyltransferase 2 (EHMT2, $P = 7.07 \times 10^{-13}$) and rs1419881 on transcription factor 19 (TCF19, $P = 1.26 \times 10^{-18}$). Conditional analysis with nearby HLA CHB loci that were previously known, confirmed the independent genetic effects of these two loci on CHB.

Conclusion: The GWAS and the subsequent validation study identified new variants associated with the risk of CHB. These findings may advance the understanding of genetic susceptibility to CHB.

INTRODUCTION

Hepatitis B virus (HBV) infection is one of the major infectious diseases. It is also the predominant risk factor for liver cirrhosis (LC) and hepatocellular carcinoma (HCC) (1). More than two billion people alive today have been infected with HBV at some time in their life. Of these, ~360 million people remain infected and become carriers of the virus (2). The association between CHB and HCC is now well established. The risk of developing HCC among carriers of the HBV surface antigen (HBsAg) has been proven to be more than 100 times greater than that of people without the antigen (3). More than 90% of perinatal infection by vertical transmission from HBsAg-positive mothers results in persistent HBV infection. In contrast, the majority of horizontal infections in adulthood are cleared, and only 5–10% of such cases become persistent carriers of HBV. Among persistently infected individuals, the outcomes are highly variable: some achieve clearance of HBV infection; however, 10–30% will eventually develop LC and HCC (4).
Recently, several genome-wide association studies (GWASs) of CHB identified human leukocyte antigen (HLA) loci, including HLA-DP and HLA-DQ in Asian populations (5–9), as associated with the risk of CHB. In this study, we performed another GWAS using CHB carriers and a higher-density chip [1 140 419 single nucleotide polymorphisms (SNPs)] to confirm and identify the host genetic factors related to CHB. In the subsequent replication analysis, we validated the novel associated SNPs in an independent set of Korean individuals.

RESULTS

In the initial GWAS stage, 1400 DNA samples were genotyped using the Illumina HumanM-1 Quad BeadChip® (Illumina, San Diego, CA, USA) in 400 CHB patients and 1000 population controls (Supplementary Material, Table S1). After quality control (QC) was applied (call rate ≥0.98, P-value of the Hardy–Weinberg equilibrium ≥0.001 and MAF ≥0.02), we obtained a final dataset of 719 265 genotypes (719 265/1 140 419; 63.1%). The most of QC failed SNPs were from a low frequency. No significant stratifications were detected using both all GWAS SNPs and HLA region SNPs. (Supplementary Material, Table S2) were identified. The top signals (rs2856718 and rs7453920) on 6p21.32 (5,7).

Among 181 SNPs which satisfied the genome-wide significance level, interestingly, additional sets of association signals on euchromatic histone-lysine-methyltransferase 2 (EHMT2) and transcription factor 19 (TCF19) around 6p21.33 (Fig. 1; Supplementary Material, Table S2) were identified. The top signals were observed on rs652888 of EHMT2 (P = 1.96 × 10⁻⁸) and rs1419881 of TCF19 (P = 4.20 × 10⁻⁸), respectively (Table 1).

Because these newly identified loci are located near the known CHB HLA loci on the same chromosome (6), we investigated whether the new association signals might simply be from the tracking of known genetic effects of the HLA loci. First, to test the independence of association from HLA loci (HLA-DP and HLA-DQ), LDs among these markers including known SNPs on HLA DQ and DP were analyzed (Supplementary Material, Table S3). The results indicated that no tight LDs were observed between known CHB HLA loci and newly identified SNPs (pairwise r² < 0.05). Next, we performed conditional analysis with known HLA loci (rs2977535 and rs3077 in HLA-DP and rs2856718 and rs7453920 in HLA-DQ). Even after stratification of known HLA loci in conditional analysis, both SNPs on EHMT2 and TCF19 were strongly associated with the risk of CHB. In the subsequent replication analysis, we validated the novel associated SNPs in an independent set of Korean individuals.

### Table 1. Association results of CHB loci in the GWAS and replication study (n = 4399)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Position (bp)</th>
<th>NC (n = 1000)</th>
<th>OR(95%CI)</th>
<th>P-value</th>
<th>OR(95%CI)</th>
<th>P-value</th>
<th>Replication Combined Analysis (NC, n = 971)</th>
<th>OR(95%CI)</th>
<th>P-value</th>
<th>OR(95%CI)</th>
<th>P-value</th>
<th>OR(95%CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-DP</td>
<td>6</td>
<td>rs277535</td>
<td>3'UTR</td>
<td>0.348</td>
<td>0.48</td>
<td>1.28</td>
<td>0.24</td>
<td>1.00 (0.75–1.32)</td>
<td>0.02</td>
<td>1.00</td>
<td>1.00 (0.75–1.32)</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>rs7453920</td>
<td>6 Intron</td>
<td>0.320</td>
<td>0.42</td>
<td>1.00</td>
<td>0.32</td>
<td>0.32 (0.21–0.48)</td>
<td>4.17E-05</td>
<td>0.32</td>
<td>0.32 (0.21–0.48)</td>
<td>4.17E-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>rs92206766</td>
<td>5'UTR</td>
<td>0.304</td>
<td>0.46</td>
<td>1.00</td>
<td>0.32</td>
<td>0.30 (0.20–0.43)</td>
<td>3.74E-05</td>
<td>0.32</td>
<td>0.30 (0.20–0.43)</td>
<td>3.74E-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>rs3095239</td>
<td>3'UTR</td>
<td>0.310</td>
<td>0.48</td>
<td>1.00</td>
<td>0.32</td>
<td>0.31 (0.21–0.46)</td>
<td>2.74E-05</td>
<td>0.32</td>
<td>0.31 (0.21–0.46)</td>
<td>2.74E-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>rs1419881</td>
<td>3'UTR</td>
<td>0.329</td>
<td>0.44</td>
<td>1.00</td>
<td>0.32</td>
<td>0.32 (0.22–0.46)</td>
<td>2.60E-05</td>
<td>0.32</td>
<td>0.32 (0.22–0.46)</td>
<td>2.60E-05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Previously identified HLA CHB loci.

SNPs with P-values < 1 × 10⁻⁸ are shown.

P-value of association analyses with top 10 PCs as covariates.
associated with CHB, indicating their independent effects on CHB susceptibility (Table 2).

To validate the genetic effects of these two SNPs on CHB, we further analyzed them in an independent study of a Korean CHB cohort consisting of 971 cases and 1938 controls from Ulsan University Hospital (Supplementary Material, Table S1). Validation analysis also showed strong associations with CHB. Combined analysis of a fixed-effects model using the Mantel–Haenszel method showed strong significant associations of these two SNPs \( P = 7.07 \times 10^{-11} \), odds ratio (OR) = 1.38 for rs652888 of EHMT2 and \( P = 1.26 \times 10^{-18} \), OR = 0.73 for rs1419881 of TCF9 (Table 1).

We estimated the proportion of variance explained by all common variants captured on the SNP arrays in a polygenic model by using a genome-wide complex trait analysis (GCTA). Assuming a population prevalence of 5% in Korean, heritability of CHB susceptibility due to the SNP array genotypes was estimated to be 20.9% of the variance (\( P \)-value = \( 1 \times 10^{-11} \), se = 0.06).

In pathway analysis using i-GSEA4GWAS, 303,148 SNPs of 719,265 SNPs were mapped to 16,887 genes that were assigned to 257 pathways. When mapping SNPs were limited within 5 kb around genes, seven pathways were significantly enriched with association signals with false discovery rate <0.05 and nominal \( P \)-value <0.01 (Supplementary Material, Table S4). Among the seven pathways identified using i-GSEA4GWAS, the top four significant pathways (cell adhesion molecules, type I diabetes mellitus, antigen processing and presentation and T-cell signal transduction) are involved in the regulation of immune function, consisting of subgroups of human immune-system genes sets. EHMT2 and TCF9 were not included in significantly enriched gene sets in this pathway analysis.

**DISCUSSION**

It is well known that the major mode of infection in HBV-endemic areas, including Korea, is perinatal transmission (1,10). The mechanisms underlying resolution of acute HBV infection or its progression to chronicity remain largely undetermined. When determining the chronicity of HBV infection within a group of patients who are presumed to have been infected at the same age (i.e. perinatally in Korea), it is apparent that the outcome of the infection does not appear to be determined by variations in the virulence of viral strains (11,12) or environmental factors; instead, host factors are more likely to influence the disease outcome (13,14). A strong genetic component determining the outcomes of HBV infection has been established through twin studies (14).

Recently, two GWASs identified several variants in the HLA region (rs9277535, rs3077, rs2856718 and rs7453920) and were
Table 2. Conditional analysis of top SNPs in GWAS (n = 1400) and replication study (n = 2909)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Gene</th>
<th>rs#</th>
<th>(P)-value conditioned by</th>
<th>(P)-value conditioned by</th>
<th>(P)-value conditioned by</th>
<th>(P)-value conditioned by</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs9277535 (HLA-DP)</td>
<td>rs3077 (HLA-DP)</td>
<td>rs2856718 (HLA-DQ)</td>
<td>rs7453920 (HLA-DQ)</td>
</tr>
<tr>
<td>GWAS</td>
<td>HLA-DP</td>
<td>rs9277535 endangering</td>
<td>2.60E–11</td>
<td>0.02</td>
<td>2.06E–09</td>
<td>7.81E–09</td>
</tr>
<tr>
<td></td>
<td>HLA-DP</td>
<td>rs3077 endangering</td>
<td>2.71E–11</td>
<td>0.02</td>
<td>9.99E–09</td>
<td>3.67E–08</td>
</tr>
<tr>
<td></td>
<td>HLA-DQ</td>
<td>rs2856718</td>
<td>5.10E–09</td>
<td>2.50E–07</td>
<td>1.09E–06</td>
<td>4.37E–05</td>
</tr>
<tr>
<td></td>
<td>HLA-DQ</td>
<td>rs7453920</td>
<td>4.31E–09</td>
<td>2.04E–06</td>
<td>7.34E–06</td>
<td>5.54E–05</td>
</tr>
<tr>
<td></td>
<td>EHMT2</td>
<td>rs652888</td>
<td>1.96E–09</td>
<td>7.10E–09</td>
<td>3.24E–08</td>
<td>1.41E–06</td>
</tr>
<tr>
<td></td>
<td>TCF19</td>
<td>rs1419881</td>
<td>4.20E–09</td>
<td>5.05E–07</td>
<td>7.70E–07</td>
<td>1.04E–06</td>
</tr>
<tr>
<td>Replication</td>
<td>HLA-DP</td>
<td>rs9277535</td>
<td>2.92E–30</td>
<td>-</td>
<td>1.12E–27</td>
<td>3.34E–24</td>
</tr>
<tr>
<td></td>
<td>HLA-DP</td>
<td>rs3077</td>
<td>4.38E–29</td>
<td>0.0001</td>
<td>-</td>
<td>2.81E–24</td>
</tr>
<tr>
<td></td>
<td>HLA-DQ</td>
<td>rs2856718</td>
<td>3.89E–17</td>
<td>1.43E–14</td>
<td>2.36E–12</td>
<td>9.29E–09</td>
</tr>
<tr>
<td></td>
<td>HLA-DQ</td>
<td>rs7453920</td>
<td>2.40E–18</td>
<td>1.52E–12</td>
<td>7.37E–11</td>
<td>2.30E–10</td>
</tr>
<tr>
<td></td>
<td>EHMT2</td>
<td>rs652888</td>
<td>2.78E–06</td>
<td>4.00E–06</td>
<td>2.47E–05</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>TCF19</td>
<td>rs1419881</td>
<td>4.51E–11</td>
<td>1.48E–07</td>
<td>5.22E–07</td>
<td>2.78E–08</td>
</tr>
<tr>
<td>Combined</td>
<td>HLA-DP</td>
<td>rs9277535</td>
<td>3.74E–40</td>
<td>-</td>
<td>5.61E–07</td>
<td>5.72E–36</td>
</tr>
<tr>
<td></td>
<td>HLA-DP</td>
<td>rs3077</td>
<td>5.25E–39</td>
<td>8.87E–06</td>
<td>-</td>
<td>5.56E–32</td>
</tr>
<tr>
<td></td>
<td>HLA-DQ</td>
<td>rs2856718</td>
<td>1.78E–24</td>
<td>2.60E–20</td>
<td>1.69E–17</td>
<td>1.84E–12</td>
</tr>
<tr>
<td></td>
<td>HLA-DQ</td>
<td>rs7453920</td>
<td>6.71E–26</td>
<td>5.84E–18</td>
<td>1.16E–15</td>
<td>3.20E–14</td>
</tr>
<tr>
<td></td>
<td>EHMT2</td>
<td>rs652888</td>
<td>7.07E–13</td>
<td>3.01E–12</td>
<td>8.29E–11</td>
<td>8.42E–07</td>
</tr>
<tr>
<td></td>
<td>TCF19</td>
<td>rs1419881</td>
<td>1.26E–18</td>
<td>3.38E–13</td>
<td>2.00E–12</td>
<td>1.02E–13</td>
</tr>
</tbody>
</table>

Conditional \(P\)-values were estimated using Plink software.

HBV, hepatitis B virus; CHB, chronic hepatitis B; LC, liver cirrhosis; HCC, hepatocellular carcinoma; GWAS, genome-wide association study; HBsAg, HBV surface antigen; HLA, human leukocyte antigen; EHMT2, euchromatic histone-lysine-methyltransferase 2; TCF19, transcription factor 19.

\(^{\ast}\)Previously identified CHB loci.
associated with CHB infection in Japanese populations (5,7). In this study, we could replicate those genetic effects in Korean population (Table 1). Another GWAS reported HLA variants (rs3135363, rs9267665 and rs9277535) associated with hepatitis B vaccine response (15). For those variants, significant associations were also detected (P = 0.0001 and 7.26 × 10^{-7} for rs3135363 and rs9267665, respectively; Supplementary Material, Table S5), although the significance of associations did not overcome Bonferroni’s correction (P < 7.0 × 10^{-8}) in this study.

TCF19 (transcription factor 19, previously known as SC1, MIM 600912) is a late growth-regulated gene, like the histone genes and the genes coding the proteins of the DNA synthesis apparatus. It has characteristics of a trans-activating factor, which exerts its role in the regulation of expression of other genes necessary for the later stages of cell cycle progression (16). Recently, TCF19 was suggested as a potential causal gene for the risk of type 1 diabetes and psoriasis vulgaris (17,18). In addition, a non-synonymous SNP in TCF19, rs7750641, was reported to be associated with mean cell hemoglobin and blood cell counts, including lymphocyte count, white blood cell count, hematocrit count and eosinophil count (19). However, due to extremely low frequency of rs7750641 (freq. = 0.00125, data not shown), this had no power in this study. Although the top signal SNP, rs1419881, was detected on 3′ UTR of exon 4, further fine mapping of causal genes as well as studies of the mechanism(s) underlying susceptibility of CHB infection are needed.

EHMT2 (MIM 604599, also known as G9A) is a histone lysine methyltransferase localized in euchromatin regions. It catalyzes the monomethylation and dimethylation of histone H3 at lysines 9 and 27 (H3K9 and K27). In addition, it is endowed with methyltransferase activity to concomitantly repress the downstream effector, the cell adhesion molecule Ep-CAM, thereby promoting the invasion step of the invasion-metastasis cascade (20). Although the role of EHMT2 in transcriptional regulation has been well described, the pathologic consequences of its dysfunction in human disease have not been well understood. It has been reported that enhanced expression of EHMT2 is involved in the proliferation of cancer cells through negative regulation of SIAH1 (seven in absentia homolog) (21). Interestingly, EHMT2 was reported to be responsible for transcriptional quiescence of the latent HIV-1 provirus by promoting repressive dimethylation and for the maintenance of viral latency (22). This might be explained by the report that G9a-deficient helper T (Th) cells were specifically impaired in their induction of Th2 lineage-specific cytokines and failed to protect against infection from the intestinal helminth Trichuris muris. A synonymous SNP [EHMT2-S237S (rs535586)] in exon 7 has been found to be associated with colorectal cancers and breast cancer risk (23). It has been suggested that the presence of the rare allele of SNP S237S might modify the mRNA folding, stability and translation of the EHMT2 protein and consequently lead to altered protein activity. In this study, EHMT2-S237S (rs535586) also showed association with CHB (P = 0.0014, OR = 1.56; data not shown). Although the most significant association was detected at rs652888 in exon 22 (1.26 × 10^{-18}), further mapping and functional studies are needed.

In this study, we have several limitations for sampling, e.g. the use of population controls whose responses to an HBV infection are unknown, and no information on infection age in cases of patients, by which we could not distinguish chronic carriers infected in infancy from those infected in adulthood. These limitations may have caused deduction of power to detect the prominent associations of HLA and/or non-HLA variation in the risk of chronic HBV infection. In addition, although we focused on most highly associated SNPs in the genes, we could not pinpoint the causal SNP and rule out the possibility of existence of other causal SNPs around this gene rich region due to the extensive LD among the genes.

In this study, we demonstrate that two novel variants in EHMT2 and TCF19 are involved in the susceptibility of CHB infection through an initial GWAS and a subsequent validation study. Further functional characterizations of these genes are needed to fully evaluate their contribution to CHB infection.

MATERIALS AND METHODS

Study subjects
A total of 4309 subjects (1371 cases and 2938 controls) were analyzed. For the initial GWAS, 400 DNA samples from chronic hepatitis B (CHB) patients were obtained from the outpatient clinic of the Liver Unit and the Center for Health Promotion at Seoul National University Hospital and Ajou University Medical Center. A thousand of the population control samples were provided by Korea BioBank, Center for Genome Science, National Institute of Health, Korea Centers for Disease Control and Prevention. For the replication study, 971 CHB patients and 1938 population control samples were collected from Ulsan University Hospital (Seoul, Korea) and Korean BioBank, respectively. The diagnosis of chronic carriers was established based on seropositivity of the HBsAg (Enzygnost1 HBsAg 5.0; Dade Behring, Marburg, Germany) over a 6-month period (Supplementary Material, Table S1). All the participants provided written informed consent. This project was approved by the ethics committees at each institution.

SNP genotyping
In the GWAS stage, 1400 DNA samples were genotyped using the Illumina HumanOmni1-Quad BeadChip® (1 140 419 SNPs, Illumina). Samples were processed according to the Illumina Infinium-II assay manual. SNP QC was applied as follows: an SNP call rate of ≥0.98 in both cases and controls and a P-value of the Hardy–Weinberg equilibrium test of ≥0.001 in controls. SNPs with a minor allele frequency of ≤0.02 in both case and control samples were excluded from further analysis.

For the replication study, we only selected two top signal SNPs located on gene regions (rs652888 and rs1419881 on TCF19 and EHMT2, respectively) which had shown multiple signals in Figure 1. We analyzed 971 cases and 1938 population controls using the TaqMan® genotyping system (Applied Biosystems, Foster City, CA, USA). TaqMan® genotyping QC was performed in 10% of the samples by duplicate checking (rate of concordance in duplicates = 100%).

Statistical analysis
In the GWAS stage and replication analyses, statistical significance (P-value) of the associations with each SNP and ORs were assessed with logistic regression for the additive model
using HelixTree™ software (Golden Helix). Haploview software was employed to draw the LD structure (Broad Institute, http://www.broadinstitute.org/mpg/haploview). The possible population stratification in this study using principal components analysis (PCA) was examined using HelixTree™. In addition, to examine possible stratification among our study populations, we incorporated the top 10 PCs as covariates in association analyses. To investigate whether the new association signals might simply be from the tracking of known genetic effects of the HLA loci (rs9277535 and rs3077 in HLA-DP and rs2856718 and rs7453920 in HLA-DQ), conditional logistic regression analyses were performed using PLINK software (Table 2). We implemented GCTA (http://www.complextraitgenomics.com/software/gcta/) to estimate the proportion of variance explained by variants on the SNP array (24). Pathway-Express (http://vortex.cs.wayne.edu/ontoexpress/) was used to explore the most biologically relevant pathways impacted by a list of input genes (25). To perform pathway analysis, we used the improved gene-set-enrichment analysis approach (i-GSEA4GWAS) that estimates the pathway-level interactions between genetic variants and CHB risks using F-values obtained at the SNP level of GWAS dataset (26).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the Korea Science and Engineering Foundation (KOSEF) funded by the Korean government (MEST) (No. 2011-0004453); a Sogang University Research Grant for 2011 (SRF-201114006.01); the SK Telecom Research Fund (3420130070); the Basic Research Laboratory Program through the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology (2010-0001200) and a grant from the Korea Healthcare technology R&D Project, Ministry of Health & Welfare, Republic of Korea (No. A101023).

REFERENCES