G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2014;12(4):203-207
http://dx.doi.org/10.5808/GI.2014.12.4.203

ORIGINAL ARTICLE

Genome-Wide Association Study of Hepatitis in Korean Populations

Youngbok Hong, Sejong Oh*

Department of Nanobiomedical Science, Dankook University, Cheonan 330-714, Korea

Hepatitis is a common and serious disease for the Korean population. It is caused by a virus, the A and B types of which are plentiful in Koreans. In this study, we tried to find genetic factors for hepatitis through genome-wide association studies. We took 368 cases and 1,500 controls from Anseong and Ansan cohort data. About 300,000 single-nucleotide polymorphisms and 20 epidemiological variables were analyzed. We did not find any meaningful significant single nucleotide polymorphisms, but we confirmed the influence of major epidemiological variables on hepatitis.

Keywords: cohort analysis, genome-wide association study, hepatitis

Introduction

Hepatitis is inflammation of the liver, most commonly caused by a viral infection [1]. Five main hepatitis viruses are known, referred to as types A, B, C, D, and E. We are concerned with these main types because of the burden of illness and death; they also have the potential for outbreaks and epidemic spread. In particular, types B and C lead to chronic disease in hundreds of millions of people and together constitute the most common cause of liver cirrhosis and cancer [1]. The Korean Health Insurance Review and Assessment Service [2] reported on C type hepatitis patients in Korea, described in Table 1. Table 1 shows that the number of patients has been increased steadily. The prevalence rate of hepatitis type C in Koreans is 1%–1.5%. From the prevalence rate, 500,000–600,000 patients are estimated. Only 10% of them are treated. If we find risk factors for hepatitis, it is helpful for the prevention and treatment of hepatitis.

In this study, we performed a genome-wide association study of hepatitis in Korean populations. We tried to find significant single-nucleotide polymorphisms (SNPs) and epidemiological traits related to hepatitis.

Methods

Phenotype and genotype data

The study subjects are based on the Anseong and Ansan cohort data, part of the Korea Association Resource (KARE) projects. The genotypes and phenotypes of the cohort population are described in Cho *et al*. [3]. Subjects with genotype accuracies below 98%, high missing genotype call rates (≥4%), high heterozygosity (>30%), or inconsistency in sex were excluded from subsequent analyses. Individuals who had a tumor were excluded, as were related individuals whose estimated identity-by-state values were high (>0.80). After these quality control steps, 352,000 SNP genotypes for 8,842 individuals were selected [4]. The epidemiological trait data for these individuals were also from the KARE project. Among the total of 8,842 individual cases, 368 had hepatitis with age over 30, and 1,500 controls were randomly selected from non-hepatitis individuals with age over 30. Table 2 summarizes the clinical characteristics of the phenotypes in this study.

The chosen dataset was imbalanced; the number of cases was smaller than controls. A dataset is imbalanced if it contains many more samples from one class than from the rest of the classes [5]. In this case, the classification analysis

Received July 8, 2014; Revised September 23, 2014; Accepted September 24, 2014

***Corresponding author:** Tel: +82-41-529-6288, Fax: +82-41-559-4899, E-mail: sejongoh@dankook.ac.kr

This is 2014 KNIH KARE best paper awarded.

Table 1. Statistics of hepatitis type C

Year	No. of patients
2008	40,683
2009	42,365
2010	41,525
2011	43,879
2012	45,890

showed good accuracy in the majority class but very poor accuracy in the minority class. Therefore, we needed to transform the imbalanced dataset to a balanced dataset. We applied an 'oversampling' scheme [5] to overcome the imbalance problem. The final dataset contained 1,500 controls and 1,500 cases.

Statistical analysis

To find significant SNPs, we used PLINK, version 1.07 [6].

Table 2. Clinical characteristics of variables in this study

Variable		Control	Case	p-value
No. of population		1,500 (17)	368 (4.2)	-
AS1_Sex	Male (%)	931 (62)	238 (64.7)	3.91×10^{-9}
AS1_Age	Age	51.7 ± 8.9	50.2 ± 8	0.000405
AS1_Height	Height	160 ± 8.8	163 ± 8.9	2×10^{-10}
AS1_BMI	Body mass index	24.5 ± 3.1	24.9 ± 3.2	1.02×10^{-5}
AS1_SBP	SBP (mm Hg)	121.2 ± 18.7	120.2 ± 17.8	0.446807
AS1_DBP	DBP (mm Hg)	80 ± 11.4	80 ± 11.5	0.884027
AS1_PdDm	Diagnosis of diabetes	51 (3.4)	46 (12.5)	2×10^{-16}
AS1_PdUl	Diagnosis of gastritis	334 (22.2)	112 (30.4)	1.68×10^{-6}
AS1_PdAl	Diagnosis of allergy	81 (5.4)	35 (9.5)	1.77×10^{-5}
AS1_PdHn	Diagnosis of external head injury	3 (0.2)	5 (1.4)	0.008527
AS1_DrugAr	Taking arthritis drug	50 (3.33)	17 (4.6)	0.000566
AS1_Albumin	Degree of albumin	4.3 ± 0.33	4.2 ± 0.35	2.89×10^{-16}

Values are presented as number (%) or mean ± SD.
SBP, systolic blood pressure; DBP, diastolic blood pressure.

Table 3. Top-ranked SNPs of genome-wide association analysis

CHR	RSID	BP	Gene	Minor allele	CHISQ	p-value	OR
11	rs11025185	19550382	*NAV2*	A	23.67	1.45×10^{-9}	1.15
16	rs4467099	11450395		A	19.71	5.72×10^{-13}	1.09
15	rs1432133	24811092		T	19.47	6.84×10^{-9}	1.02
12	rs6582709	46104168		T	19.39	0.00293	2.07
5	rs17568725	171103246		T	19.38	1.27×10^{-8}	0.90
12	rs2097726	46105143		T	19.05	-	0.62
6	rs6569628	130137425		T	18.3	0.99228	1.00
14	rs8014067	61623010	*SYT16*	T	18.09	2.52×10^{-9}	0.59
6	rs9375664	130134371		T	17.95	-	1.77
6	rs2326864	130136091		A	17.73	0.32261	0.61
8	rs2607612	24662484		G	17.72	0.68156	1.04
12	rs6582710	46104230		C	17.5	-	0.83
15	rs2174866	51251512		T	17.41	8.78×10^{-7}	1.19
6	rs10484389	22183241	*CASC15*	T	17.13	1.55×10^{-8}	1.18
8	rs7814301	24645945		C	16.93	-	1.05
8	rs4368986	24641948		A	16.89	0.57515	0.92
13	rs9522267	110994368		T	16.85	4.48×10^{-10}	0.92
1	rs7518687	166899607		A	16.78	1.45×10^{-10}	1.19
8	rs6985699	24658799		G	16.76	0.06382	1.09
10	rs4474337	10819693		T	16.7	1.76×10^{-5}	0.95

RSID, reference SNP ID obtained from dbSNP database; BP, base pair based on the human reference genome, ver. 36 (NCBI); CHISQ, chi-square value; OR, odds ratio.

Other statistical analyses were performed using R, version 3.1. We used logistic regression to find major factors related with hepatitis. Receiver operating characteristic (ROC)/area under the curve (AUC) analysis was performed to confirm the prediction power of the major factors that were found.

Results

Genome-wide association studies

Table 3 summarizes 20 SNPs that were top-ranked by p-value in the genome-wide association analysis. Unfortunately, there were no significant SNPs that met $p < 5 \times 10^{-8}$. We performed logistic regression test on the top-ranked SNPs, and we took 10 SNPs in Table 4. Logistic regression measures the relationship between a categorical dependent variable (phenotype) and one or more independent variables (SNPs). Fig. 1 shows the ROC plot for the classification test using the 10 SNPs. The AUC value from the ROC plot is 0.700; it is not enough as a biomarker.

Table 4. Logistic regression test for SNP data

SNP	Coefficient value	p-value
rs4467099	0.090529	5.72×10^{-13}
rs9522267	−0.081189	4.48×10^{-10}
rs7518687	0.178053	1.45×10^{-10}
rs1432133	0.096962	6.84×10^{-9}
rs8014067	−0.115841	2.52×10^{-9}
rs11025185	0.140190	1.45×10^{-9}
rs10484389	0.161836	1.55×10^{-8}
rs17568725	−0.104275	1.27×10^{-8}
rs2174866	0.170567	8.78×10^{-7}
rs4474337	−0.055057	1.76×10^{-5}
rs6582709	0.728803	0.00293

SNP, single nucleotide polymorphism.

Epidemiological studies

Using the traits in Table 2, we performed a logistic regression test. Table 5 summarizes the results. As we can see, diabetes, gastritis, allergy, external head injury, taking arthritis drug, and degree of albumin were highly correlated with hepatitis. In the case of diabetes, the probability that a diabetes patient had hepatitis was 4 times higher than a diabetes-free person. In general, the hepatitis C virus is often associated with diabetes, and some diabetics may even develop chronic hepatitis [7]. Gastritis is influenced by hepatitis. If we have hepatitis, the probability of getting gastritis is increased 1.5 times. Especially, chronic gastritis develops by chronic hepatitis [8]. Sometimes, allergy causes

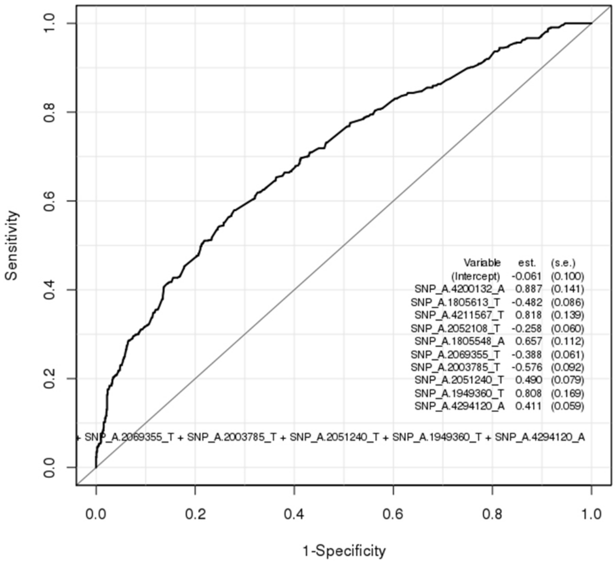

Fig. 1. Receiver operating characteristic plot for 4 single nucleotide polymorphisms derived from logistic regression (area under the curve, 0.700).

Table 5. Logistic regression test for epidemiological data

ID	Variable	Coefficient value	p-value	OR
T1	Sex	−0.1688894	3.91×10^{-9}***	0.84
T2	Age	−0.0041674	0.000405***	0.99
T3	Diagnosis of diabetes	0.2717830	2×10^{-16}***	1.31
T4	Diagnosis of gastritis	0.0960720	1.68×10^{-6}***	1.1
T5	Diagnosis of allergy	0.1461045	1.77×10^{-5}***	1.16
T6	Diagnosis of external head injury	0.2739906	0.008527**	1.31
T7	Taking arthritis drug	0.1577881	0.000566***	1.17
T8	Degree of albumin	−0.2201075	2.89×10^{-16}***	0.8
T9	Height	0.0058235	0.000304	1.0
T10	BMI	0.0126750	1.02×10^{-5}***	1.01

Significant codes: '***', 0.001; '**', 0.01.
OR, odds ratio; BMI, body mass index.

Table 6. Area under the curve (AUC) values

Variable	AUC
T1 + T2 + T3 + T10	0.647
T1 + T2 + T3 + T10 + T4	0.657
T1 + T2 + T3 + T10 + T4 + T5	0.662
T1 + T2 + T3 + T10 + T4 + T5 + T6	0.666
T1 + T2 + T3 + T10 + T4 + T5 + T6 + T7	0.670
T1 + T2 + T3 + T10 + T4 + T5 + T6 + T7 + T8	0.690
T1 + T2 + T3 + T10 + T4 + T5 + T6 + T7 + T8 + T9	0.693

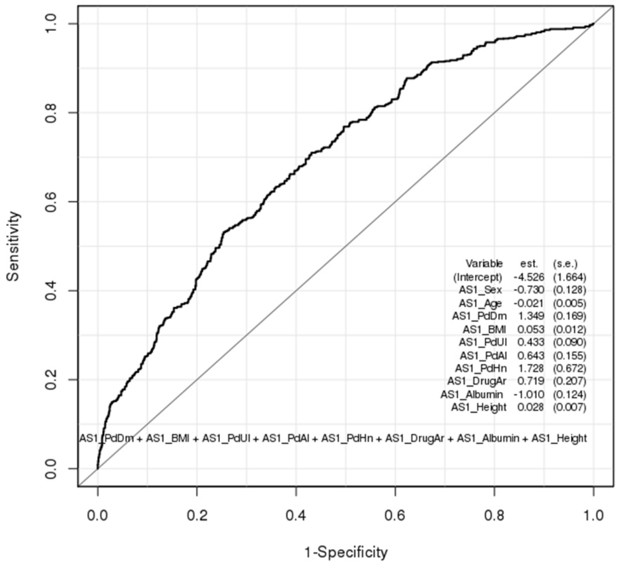

Fig. 2. Receiver operating characteristic plot for 8 epidemiological variables derived from logistic regression (area under the curve, 0.693).

hepatitis. It increases the probability of hepatitis by 1.8 times. Hepatitis virus can cause arthritis [9]; it increases the probability of arthritis by 1.9 times. The degree of albumin is inversely proportional to hepatitis (odds ratio [OR], 0.8), because the liver makes albumin, and hepatitis enervates the process. The OR between external head injury and hepatitis is very high (OR, 1.31). We cannot explain the medical relationship between them. It needs more analysis.

Table 6 shows AUC values from the variables in Table 5. As we can see, four variables can explain 64.7% of the cause of hepatitis, and 8 variables can explain 69.3%. Fig. 2 shows the ROC plot for eight variables in Table 3.

Discussion

From the epidemiological analysis, we found relevant variables with hepatitis. We confirmed that hepatitis has a wide relation with other diseases. If we make a disease network in which the node is a disease and the edge is a correlation coefficient between two nodes, we can understand the relationship among diseases more clearly. Current known disease networks [10, 11] do not show detailed relationships between hepatitis and other diseases. This is a future research topic.

KARE data are the result of a cohort study. It contains a small number of samples for specific diseases, whereas the whole population is very big. It induces an imbalanced dataset for statistical analysis. Our study implies a basic limitation, even though we tried to complement the problem. We also did not find any significant SNPs related with hepatitis. If we combine the knowledge of other biological databases, we may get a more meaningful interpretation for the results of our experiment.

Acknowledgments

This work was supported by grants from the Korea Centers for Disease Control and Prevention, Republic of Korea (4845-301, 4851-302, 4851-307).

References

1. World Health Organization (WHO). What is hepatitis? Geneva: World Health Organization, 2014. Accessed 2014 Nov 1. Available from: http://www.who.int/features/qa/76/en/.
2. Korean Health Insurance Review & Assessment Service. Seoul: Korean Health Insurance Review & Assessment Service, 2014. Accessed 2014 Nov 1. Available from: http://www. hira.or.kr/main.do.
3. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41:527-534.
4. Hong KW, Kim SS, Kim Y. Genome-wide association study of orthostatic hypotension and supine-standing blood pressure changes in two korean populations. *Genomics Inform* 2013;11:129-134.
5. Ganganwar V. An overview of classification algorithms for imbalanced datasets. *Int J Emerg Technol Adv Eng* 2012;2:42-7.
6. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA,

Bender D, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559-575.

7. Negro F, Alaei M. Hepatitis C virus and type 2 diabetes. *World J Gastroenterol* 2009;15:1537-1547.
8. The Free Dictionary. Gastritis. Huntingdon Valley: Farlex Inc., c2003-2014. Accessed 2014 Nov 1. Available from: http://encyclopedia2.thefreedictionary.com/.
9. American College of Rheumatology. HCV and Rheumatic Disease. Lake Boulevard: American College of Rheumatology, 2014. Accessed 2014 Nov 1. Available from: https://www.rheumatology.org/Practice/ Clinical/Patients/.
10. Diseasome. Human disease network. Dieseasome, 2014. Accessed 2014 Nov 1. Available from: http://diseasome.eu/map.html.
11. HuDiNe. Human disease network. HuDiNe, 2014. Accessed 2014 Nov 1. Available from: http://hudine.neu.edu/.