Knowledge-based genetic association study of hepatitis B virus related hepatocellular carcinoma

Deke Jiang (dekejiang17@smu.edu.cn)
Southern Medical University Nanfang Hospital https://orcid.org/0000-0002-7888-2344

Jiaen Deng
Department of Psychiatry, the University of HONG KONG

Changzheng Dong
Ningbo University School of Medicine

Xiaopin Ma
State Key Labaratory of Genetic Engineering, School of Life Sciences, Fudan University

Qianyi Xiao
Center for Genomic Translational Medicine and Prevention, School of Public Health, Fudan University

Bin Zhou
Hepatology Unit, Nanfang Hospital, Southern Medical University

Chou Yang
Southern Medical University Nanfang Hospital

Lin Wei
University of Chicago

Carly Conran
University of Chicago

S. Lilly Zheng
University of Chicago

Irene Oi-lin Ng
University of Hong Kong

Long Yu
Fudan University

Jianfeng Xu
University of Chicago

Pak C. Sham
University of Hong Kong

Xiaolong Qi
Southern Medical University Nanfang Hospital

Jinlin Hou
Southern Medical University Nanfang Hospital

Yuan Ji
University of Chicago

Guangwen Cao
Second Military Medical University

Miaoxin Li
University of Hong Kong

Research article

Keywords: knowledge-based genetic association, susceptibility, hepatitis B virus, hepatocellular carcinoma

Posted Date: September 18th, 2019

DOI: https://doi.org/10.21203/rs.2.14627/v1

Version of Record: A version of this preprint was published at BMC Cancer on May 11th, 2020. See the published version at https://doi.org/10.1186/s12885-020-06842-0.

Abstract

Background : Recent genome-wide association studies (GWASs) have suggested several susceptibility loci of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC) by statistical analysis at individual single-nucleotide polymorphisms (SNPs). However, these loci only explain a small fraction of HBV-related HCC heritability. In the present study, we aimed to identify additional susceptibility loci of HBV-related HCC using advanced gene- and gene-set-based association tests.

Methods: We performed a meta-analysis of two existing GWASs of HBV-related HCC, based on which a series of association analyses at genes and multiple gene sets curated according to current knowledge were carried out for prioritizing potential risk genes. A series of prioritized SNPs were selected to replicate genetic associations in an independent sample of 965 cases and 923 controls.

Results: The gene-based association analysis suggested that five genes are significantly associated with HBV-related HCC risk: RNY4, GOLGA8M, LINC01207, WHAMMP2 and SLC39A8. Through gene-set-based association analysis, we found that the genes in systemic lupus erythematosus pathway may be relevant to development of HBV-related HCC. Three previously reported genes, NAT2, GSTA1 and GSTA2, were also highlighted to be susceptibility genes of HBV-related HCC when genes were stratified in a liver-specific expression set. However, probably due to small sample size, none of the genes prioritized by knowledge-based association analyses are successfully replicated in an independent sample.

Conclusions: This comprehensive knowledge-based association mining study suggested several promising genes significantly associated with HBV-related HCC risk. More experiments or larger samples are needed to validate their contribution to the pathogenic mechanism of HCC.

Background

Hepatocellular carcinoma (HCC) is one of the most common cancers worldwide. With 750,000 new HCC cases diagnosed each year, it is the third leading cause of cancer mortality.(1) As many as 30% of patients diagnosed with hepatitis, fibrosis or cirrhosis ultimately develop HCC. In high endemic areas such as Africa and Asia, at least 60% of HCC is associated with hepatitis B virus (HBV).(2) However, only a minority of HBV carriers develops HCC. HBV carriers with a family history of HCC were estimated to have over two-fold risk for HCC compared with those without a family history of HCC.(3) Furthermore, genetic complex segregation analysis suggested that major genes may be involved in the genetic predisposition to develop HCC at an earlier age.(4)

Genome-wide association study (GWAS) is a widely used strategy for identifying risk loci of complex diseases. Recently, several GWASs on risk of HBV-related HCC were conducted using single-nucleotide polymorphisms (SNPs)-based statistical association tests. Multiple susceptibility loci were identified, including rs17401966 in intron 24 of *KIF1B* at 1p36.22, rs7574865 in intron 3 of *STAT4* at 2q32.2–32.3, rs9275319 between *HLA-DQB1* and *HLA-DQA2* at 6p21.3, rs9272105 between *HLA-DQA1* and *HLA-DRB1* at 6p21.3, and rs455804 in intron 1 of *GRIK1* at 21q21.3.(5–7) However, these susceptibility loci account for only a small fraction of the contribution of genetics to HBV-related HCC. Identifying additional genetic alterations associated with HBV-related HCC may be difficult due to the relatively weak effects of many individual risk SNPs, which may be unidentifiable with the currently available, relatively small sample sizes.(8) SNP-based statistical association tests alone in GWAS do not have enough power to discover most risk loci for human complex diseases. Gene- and biological pathway-based association analysis has been proposed to have superior statistical power compared with conventional statistical tests, as it relieves multiple testing and enriches signals.(9) Moreover, gene- and biological pathway-based analysis also lends itself to introducing more disease-specific knowledge into the analysis.

In the present study, we performed a gene-based association analysis with meta-analysis *p*-values from two independent HBV-related HCC GWASs. The gene-based *p*-values were further evaluated within multiple gene-sets defined according to

knowledge of HCC. SNPs within prioritized genes were selected for replication in two independent HBV-related HCC case/control populations.

Methods

Two existing GWASs on HBV-related HCC

The association *p*-values were obtained from two previous GWASs on HBV-related HCC in Chinese populations for meta-analysis and knowledge-based association analysis. One study(7) contained 2,689 chronic HBV carriers (1,212 HBV-related HCC cases and 1,477 controls) recruited from May 2006 to December 2012 by the Qidong Liver Cancer Institute in Jiangsu Province of Mainland China. The other study(10) consisted of 95 HBV-infected HCC patients (cases) and 97 HBV-infected patients without HCC (controls) recruited at Queen Mary Hospital, Hong Kong. The sample inclusion and exclusion criteria were described in the original papers.(7, 10)

Subjects in replication studies

The subjects in replication, including 965 chronic HBV carriers with HCC as cases and 923 chronic HBV carriers without HCC as controls, were recruited from the affiliated hospitals of the Second Military Medical University, Shanghai, China. All the samples are of Han Chinese descent and have participated in previously published studies.(7, 11)

The study was performed in accordance with guidelines approved by the local ethical committees from all participating centers involved in both the GWAS stage and the replication stage. An informed consent to participate in the study was obtained from each subject in accordance with the declaration of Helsinki principles. All study participants approved the storage of their frozen DNA specimens, for research purposes, in our laboratory.

Genotyping and quality control in replication

Genomic DNA from the peripheral blood of all participants in replication was extracted using the QIAamp DNA Blood Mini Kit (QIAGEN GmbH, Hilden, Germany). Genotyping analyses for replication samples were conducted using the Sequenom MassArray system (Sequenom) according to the manufacturer's instructions. Genotyping quality was examined by a detailed QC procedure consisting of a 95% successful call rate, duplicate calling of genotypes, and internal positive control samples and two water samples (PCR negative controls) included in each 96-well plate. Genotype analysis was performed by technicians in a blind fashion.

Meta-analysis of variants

The association *p*-values of untyped SNPs were imputed directly by the tool FAPI (http://grass.cgs.hku.hk/limx/fapi/) (12)with default settings. The *p*-values of the two GWASs were then combined by Stouffer's Z-score method for meta-analysis on FAPI as well:

$$Z_{meta} = \frac{\sum_{i=1}^{N}(w_i * z_i)}{\sqrt{\sum_{i=1}^{N} w^2}} \text{ where } w_i = \sqrt{n_i}$$

in which *N* is the number of GWASs, z_i is the individual z-score of the i_{th} GWAS study, and n_i is the sample size of the i_{th} study.

Gene-based and gene-set-based analysis

The knowledge-based secondary analysis platform KGG (http://grass.cgs.hku.hk/limx/kgg/) was used to map the SNPs onto reference genes (UCSC RefGene hg19), and to perform gene-based and gene-set-based association analysis with default settings. The phased genotypes of Eastern Asian samples in the 1000 Genomes Project(13) were used to account for linkage disequilibrium of SNPs through KGG. The Benjamini-Hochberg approach was used to control false discovery rate (FDR) of genome-wide genes at a level, which is a more powerful multiple testing approach than Bonferroni correction when there are multiple susceptibility genes.

Variants functional annotation

The genomic annotation tools, HaploReg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php)(14)and RegulomeDB Version 1.1 (http://regulomedb.org/)(15), were used to annotate SNPs with epigenomic markers and potential regulatory elements, including regions of DNase I hypersensitivity, binding sites for transcription factors (TFs), promoter regions that have been biochemically characterized to regulate transcription, chromatin states as well as DNase foot printing, PWMs, and DNA Methylation. KGGSeq (Version 1.0)(16, 17) was used to annotate selected SNP with four regulatory or functional prediction scores (including CADD.CScore(18), SuRFR(19), FunSeq2(20) and cepip(21)).

Results

We first combined the association *p*-values of variants by meta-analysis from two independent GWASs. Association analyses at genes and multiple knowledge-based gene-sets were carried to prioritize potential HBV-related HCC susceptibility genes. A series of prioritized variants were selected to replicate their genetic associations in a group of independent case-control samples. The overall workflow is shown in Figure 1.

Genome-wide meta-analysis of two HBV-related HCC GWASs in Chinese populations

Association *p*-values were imputed based on the linkage disequilibrium (LD) pattern in the Eastern Asian Panel from the 1000 Genomes Project. A genome-wide meta-analysis was then performed with SNP *p*-values from two existing Chinese HCC GWASs using the tool FAPI.(12) After quality control (QC), 5,375,073 meta-analysis *p*-values of SNPs were obtained. The Manhattan plot and QQ plots of *p*-values are shown in Supplementary Figure 1 and Supplementary Figure 2, respectively. At the upper tail of the QQ plot, there is a deviation from the 95% confidence level of the non-hypothesis line, suggesting the existence of association signals at some SNPs.

Gene-based association analysis

We then used the meta-analysis *p*-values for gene-based association analysis by GATES(22) on a tool called KGG (version 3.5).(23) In addition to SNPs within the untranslated regions, introns and exons, the meta-analysis *p*-values of SNPs within 5kb upstream and downstream of a gene were also included in the gene-based association test. SNPs in overlapping regions of multiple genes were assigned to all involved genes. The QQ plots of gene-based *p*-values are shown in Figure 2.

According to the gene-based *p*-values, three genes, *RNY4, GOLGA8M* and *LINC01207* passed the multiple testing correction by FDR, 0.05 (Table 1). Interestingly, the *RNY4* and *LINC01207* are non-coding RNA genes, which have not been previously well studied. In addition, two genes, *WHAMMP2* and *SLC39A8,* have nearly significant *p*-values on the genome, corrected $p = 0.054$ (Table 1). We further annotated the two RNA genes (*RNY4* and *LINC01207)* and the pseudogene

WHAMMP2 with known regulatory elements and epigenomic markers by the UCSC genome browser (http://genome.ucsc.edu). While the *RNY4* has no known regulatory factors (See Supplementary Figure 3), the *LINC01207* and *WHAMMP2* genes have many regulatory factors and epigenomic markers (See Supplementary Figure 4 and Supplementary Figure 5). These annotations imply that the latter two genes are functionally active despite not encoding proteins.

Prioritization of genes in different gene-sets

To select more promising genes for replication in independent samples, we resorted to a series of gene-set resources to prioritize genes with suggestive association *p*-values. We first examined the association with HCC in 1,057 canonical pathways curated in the Molecular Signatures Database (MSigDB V 4.0), after removing the pathways containing too few (<5) or too many (>300) genes. According to the gene-set-based association *p*-value by the Wilcoxon test on KGG, one pathway, the systemic lupus erythematosus (SLE) pathway, passed the significance level (nominal $p = 1.63\times10^{-5}$, corrected $p = 0.017$). Seven genes have a gene-based *p*-value below 0.05 in the pathway.

Then, we investigated whether the genes highly and specifically expressed in human liver were associated with HCC. In the database, Tissue-specific Gene Expression and Regulation (TiGER, http://bioinfo.wilmer.jhu.edu/tiger/), 309 genes preferentially expressed in liver were retrieved. In the human proteome atlas (http://www.proteinatlas.org/humanproteome), 433 genes showing elevated expression of proteins in liver compared to other tissue types were retrieved as well. To reduce potential false positives, we only used overlapping genes in the two sets. As a result, a total of 189 genes were obtained. The gene *NAT2* had the lowest gene-based *p*-value of 0.01, the genes *GSTA1* and *GSTA2* had the second and third smallest *p*-values (See the genes and *p*-values in Table 2 and Supplementary Table 1).

We also examined the association of recurrent integrated genes by HBV reported in previous studies,(24–27) the genes reported to be genetically associated with HBV-related HCC risk in previous studies, and HCC risk genes defined by COSMIC database (http://cancer.sanger.ac.uk/cosmic). However, none of the genes had a promising association *p*-value with HCC in our samples (see the genes and *p*-values in Supplementary Table 2–4).

Replication study in independent samples

We replicated genetic association at genes prioritized by the above gene-based and gene-set-based associations in a group of independent HBV-related HCC case-control samples. In total, 21 SNPs of the prioritized genes were selected according to the stability of their allele sequences in ancestry matched reference panel in the 1000 Genomes Project and/or their predicted functional importance by RegulomeDB (http://regulomedb.org/) with regulatory elements. After the genotype quality assessment, two SNPs were excluded because they failed to pass the Hardy-Weinberg equilibrium test ($p<0.001$).

Three genetic models (additive, dominant and recessive) were considered under a logistic regression framework in which the HCC status was adjusted for sex and age. Generally, the independent sample failed to replicate a significant association in the discovery sample after correcting multiple testing. Only two SNPs, rs389883 and rs17343667, had an association *p*-value below 0.05. The rs389883, which is in intron region of *STK19,* had *p*-values of 0.026 and 0.032 for HCC association under additive and recessive models, respectively, with a protective effect at the minor allele G. However, in the original Qidong GWAS sample and Hong Kong GWAS sample, G was estimated to have a risk effect. The other SNP, rs17343667, which is located in the first intron of *EIF2AK1,* had an association *p*-value equal to 0.02 under the additive model with an odds ratio of 1.27 for the minor allele, which was found to have a risk effect in both original Qidong and Hong Kong GWAS samples (Table 3). In addition, the regulator potential of rs17343667 was supported by expression

quantitative trait locus (eQTL) and TF binding/ DNase peak (scored 1f) in RegulomeDB (See details in Supplementary Figure 6).

Discussion

This study utilized knowledge-based approaches to mine new susceptibility loci of HBV-related HCC in existing HBV-related HCC GWAS data sets. The gene-based association analysis suggested five statistically significant genes including *RNY4, GOLGA8M, LINC01207, WHAMMP2* and *SLC39A8.* The gene-set-based association analysis implied that genes in the SLE pathway may be relevant to the development of HCC. In addition, three genes, *NAT2, GSTA1* and *GSTA2,* were also highlighted when genes were stratified in some functional sets. Furthermore, our analysis also suggested that the germline susceptibility loci of HBV-related HCC are unlikely to be enriched in recurrent targeted genes of HBV infection, or HCC risk genes with many somatic mutations. However, probably due to small sizes in our replication samples, no associations prioritized by the knowledge-based association analysis are successfully replicated in an independent sample. The rs17343667 of *EIF2AK1* is the only one with suggestive significance.

Our study is the first to indicate that these five genes (*RNY4, GOLGA8M, LINC01207, WHAMMP2* and *SLC39A8,* which were discovered by gene-based association analysis) are relevant to the development of HBV-related HCC. For *RNY4, GOLGA8M* and *WHAMMP2,* there are no publications, to our knowledge, about their roles in risk of HCC or other cancers until this study. *LINC01207* has been implicated as a biomarker for survival of colorectal adenocarcinoma(28) and promoting proliferation of lung adenocarcinoma.(29) *SLC39A8* has been reported to regulate IFN-γ level in T cells(30) and influence trace element homeostasis in liver,(31, 32) which may be relevant to the development of HCC. Functional studies are warranted to explore the mechanisms of the potential roles of these genes in risk of HBV-related HCC.

Interestingly, our finding that the SLE pathway-related genes may be relevant to the development of HBV-related HCC is supported by a recent meta-analysis involving 59,662 SLE patients, which suggested that SLE had a relative risk of 3.21 (95% CI, 1.70–6.05) for liver cancer.(33) In addition, studies have found that a number of risk genes are shared by SLE and HBV-related HCC, such as *STAT4* and genes in the *HLA* region.(7, 34) Our results may further explain the comorbidity of the two diseases from a genetics aspect.

The three genes, *NAT2, GSTA1* and *GSTA2,* that are highly expressed in liver have been previously suggested to be relevant to HCC risk. Both Gelatti et al.(35) and Yu et al.(36) observed a significant association between *NAT2* genetic polymorphisms and HCC susceptibility among chronic HBV carriers who were smokers. Huang et al.,(37) found that the *NAT2* gene polymorphisms may confer different susceptibilities to the effect of red meat intake on HCC. *GSTA1* polymorphism was suggested to be associated with an increased risk of occurrence of HCC, and decreased expression of *GSTA1* was considered as a marker of advanced and highly aggressive HCC.(38) *GSTA1* polymorphism was also reported to correlate with both *GSTA1* and *GSTA2* expression in the liver, which is expected to be of significance for individual risk of cancer or individual response to chemotherapeutic agents.(39)

The negative findings in all curated gene sets were unexpected. Particularly, three gene sets (recurrent targeted genes of HBV infection, HCC risk genes with many somatic mutations and genes highly and specifically expressed in human liver) appeared to be very biologically relevant to the development of HCC. In the analyses, there were no trends that genes with smaller HCC association *p*-values were enriched in the gene sets. These results suggest that the biological context or connection of underlying susceptibility genes is elusive, and that it is difficult to use our current knowledge to identify the unknown susceptibility genes of HCC. Using larger sample sizes for hypothesis-free GWASs is likely the only reliable way for identification of HCC risk genes at present.

The SNP rs17343667 in the *EIF2AK1* is a promising candidate susceptibility variant although it only has a suggestively significant *p*-value in the small replication samples. In RegulomeDB, this SNP is a *cis* eQTL of lymphoblastoid and is

located within the DNase peak and histone modifications of multiple tissues and cell types. In the HaploReg (v4.1) database, this SNP is located within multiple regulatory elements, such as histone marks, DNAse and transcription Motifs. *EIF2AK1* encodes a kinase protein for translation initiation to downregulate protein synthesis in response to stress. Previous studies suggested that *EIF2AK1* mRNA and protein were overexpressed and the kinase activity was enhanced in HCC.(40, 41)

Conclusion

We performed the first systematic gene- and gene-set-based association study of HCC. Our study suggested several promising genes significantly associated with HCC risk, which may shed insights into pathogenic mechanisms of this fatal disorder. However, the negative associations in multiple curated gene sets also imply that it is difficult to infer gene associations using our current biological knowledge. More hypothesis-free genetic studies with larger sample sizes are needed to elucidate the susceptibility genes and mechanisms of HCC.

Abbreviations

eQTL, expression quantitative trait locus; FDR, false discovery rate; GWAS, genome-wide associated studies; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; LD, linkage disequilibrium; QC, quality control; SLE, systemic lupus erythematosus; SNP, single nucleotide polymorphism; TF, transcription factor.

Declarations

Ethics approval and consent to participate

The study was performed in accordance with guidelines approved by the local ethical committees from all participating centers (The Ethics Committee of Qidong Liver Cancer Institute; the Ethics Committee of the Second Military Medical University; and the Institutional Review Board of Queen Mary Hospital, University of Hong Kong) involved in both the GWAS stage and the replication stage. An informed consent to participate in the study was obtained from each subject in accordance with the declaration of Helsinki principles. All study participants approved the storage of their frozen DNA specimens, for research purposes, in our laboratory.

Consent for publication

Not applicable.

Availability of data and materials

Please contact author for data requests.

Competing interests

The authors declare that they have no conflict of interest.

Funding

This study was supported by Hong Kong Health and the Medical Research Fund (01121436 and 02132236 to M. X. L.), the National Natural Science Foundation of China (31100895, 81472618 and 81670535 to D. K. J., 81402297 to Q. L. X.), the National Science and Technology Major Project (No. 2018ZX10301202 to D. K. J. and J. H.), the Local Innovative and Research Teams Project of Guangdong Pearl River Talents Program (No. 2017BT01S131 to D. K. J. and J. H.), the Innovative Research Team Project of Guangxi Province (2017GXNSFGA198002 to D. K. J.), the Grant for Recruited Talents to Start Scientific Research from Nanfang Hospital (to D. K. J.), and the Outstanding Youths Development Scheme of Nanfang Hospital, Southern Medical University (No. 2017J001 to D. K. J.). The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author Contributions

D. K. J.: study concept and design, material support, obtained funding, analysis and interpretation of the data, and drafting of the manuscript; J. D.: analysis and interpretation of the data, and drafting of the manuscript; C. D.: analysis and interpretation of the data; X. M.: material support; Q. X.: material support; B. Z.: material support; C. Y.: revision of the manuscript; L. W.: analysis and interpretation of the data; C. C.: critical revision of the manuscript; S. L. Z.: technical, acquisition of data; I. O. N.: study concept and material support; L. Y.: material support; J. X.: material support; P. C. S.: study concept and design; X. Q.: critical revision of the manuscript; J. H.: material support; Y. J.: analysis and interpretation of the data; G. C.: material support; M. X.L: study supervision, study concept and design, obtained funding, analysis and interpretation of data, drafting of the manuscript. All authors have read and approved the manuscript.

Acknowledgements

Not applicable.

References

1.Pinyol R, Llovet JM. Hepatocellular carcinoma: genome-scale metabolic models for hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol. 2014;11(6):336–7.

2.Arzumanyan A, Reis HM, Feitelson MA. Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. Nat Rev Cancer. 2013;13(2):123–35.

3.Yu MW, Chang HC, Liaw YF, Lin SM, Lee SD, Liu CJ, et al. Familial risk of hepatocellular carcinoma among chronic hepatitis B carriers and their relatives. J Natl Cancer Inst. 2000;92(14):1159–64.

4.Cai RL, Meng W, Lu HY, Lin WY, Jiang F, Shen FM. Segregation analysis of hepatocellular carcinoma in a moderately high-incidence area of East China. World J Gastroenterol. 2003;9(11):2428–32.

5.Zhang H, Zhai Y, Hu Z, Wu C, Qian J, Jia W, et al. Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. Nature genetics. 2010;42(9):755–8.

6.Li S, Qian J, Yang Y, Zhao W, Dai J, Bei JX, et al. GWAS identifies novel susceptibility loci on 6p21.32 and 21q21.3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. PLoS Genet. 2012;8(7):e1002791.

7.Jiang DK, Sun J, Cao G, Liu Y, Lin D, Gao YZ, et al. Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. Nature genetics. 2013;45(1):72–5.

8.Manolio TA. Bringing genome-wide association findings into clinical use. Nature reviews Genetics. 2013;14(8):549–58.

9.Kwak IY, Pan W. Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. Bioinformatics. 2017;33(1):64–71.

10.Chan KY, Wong CM, Kwan JS, Lee JM, Cheung KW, Yuen MF, et al. Genome-wide association study of hepatocellular carcinoma in Southern Chinese patients with chronic hepatitis B virus infection. PloS one. 2011;6(12):e28798.

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

12.Kwan JS, Li MX, Deng JE, Sham PC. FAPI: Fast and accurate P-value Imputation for genome-wide association study. Eur J Hum Genet. 2016;24(5):761–6.

13.Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75–81.

14.Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40(Database issue):D930–4.

15.Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M. Dynamic trans-acting factor colocalization in human cells. Cell. 2013;155(3):713–24.

16.Li M, Li J, Li MJ, Pan Z, Hsu JS, Liu DJ, et al. Robust and rapid algorithms facilitate large-scale whole genome sequencing downstream analysis in an integrative framework. Nucleic acids research. 2017;45(9):e75.

17.Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic acids research. 2012;40(7):e53.

18.Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics. 2014;46(3):310–5.

19.Ryan NM, Morris SW, Porteous DJ, Taylor MS, Evans KL. SuRFing the genomics wave: an R package for prioritising SNPs by functionality. Genome medicine. 2014;6(10):79.

20.Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome biology. 2014;15(10):480.

21.Li MJ, Li M, Liu Z, Yan B, Pan Z, Huang D, et al. cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. Genome biology. 2017;18(1):52.

22.Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet. 2011;88(3):283–93.

23.Li MX, Sham PC, Cherny SS, Song YQ. A knowledge-based weighting framework to boost the power of genome-wide association studies. PloS one. 2010;5(12):e14480.

24.Paterlini-Brechot P, Saigo K, Murakami Y, Chami M, Gozuacik D, Mugnier C, et al. Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. Oncogene. 2003;22(25):3911–6.

25.Ding D, Lou X, Hua D, Yu W, Li L, Wang J, et al. Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach. PLoS Genet. 2012;8(12):e1003065.

26.Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nature genetics. 2012;44(7):765–9.

27.Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome research. 2012;22(4):593–601.

28.Zeng JH, Liang L, He RQ, Tang RX, Cai XY, Chen JQ, et al. Comprehensive investigation of a novel differentially expressed lncRNA expression profile signature to assess the survival of patients with colorectal adenocarcinoma. Oncotarget. 2017;8(10):16811–28.

29.Wang G, Chen H, Liu J. The long noncoding RNA LINC01207 promotes proliferation of lung adenocarcinoma. American journal of cancer research. 2015;5(10):3162–73.

30.Aydemir TB, Liuzzi JP, McClellan S, Cousins RJ. Zinc transporter ZIP8 (SLC39A8) and zinc influence IFN-gamma expression in activated human T cells. Journal of leukocyte biology. 2009;86(2):337–48.

31.Lin W, Vann DR, Doulias PT, Wang T, Landesberg G, Li X, et al. Hepatic metal ion transporter ZIP8 regulates manganese homeostasis and manganese-dependent enzyme activity. The Journal of clinical investigation. 2017;127(6):2407–17.

32.Engelken J, Espadas G, Mancuso FM, Bonet N, Scherr AL, Jimenez-Alvarez V, et al. Signatures of Evolutionary Adaptation in Quantitative Trait Loci Influencing Trace Element Homeostasis in Liver. Molecular biology and evolution. 2016;33(3):738–54.

33.Cao L, Tong H, Xu G, Liu P, Meng H, Wang J, et al. Systemic lupus erythematous and malignancy risk: a meta-analysis. PloS one. 2015;10(4):e0122964.

34.Teruel M, Alarcon-Riquelme ME. The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned. Journal of autoimmunity. 2016;74:161–75.

35.Gelatti U, Covolo L, Talamini R, Tagger A, Barbone F, Martelli C, et al. N-Acetyltransferase–2, glutathione S-transferase M1 and T1 genetic polymorphisms, cigarette smoking and hepatocellular carcinoma: a case-control study. Int J Cancer. 2005;115(2):301–6.

36.Yu MW, Pai CI, Yang SY, Hsiao TJ, Chang HC, Lin SM, et al. Role of N-acetyltransferase polymorphisms in hepatitis B related hepatocellular carcinoma: impact of smoking on risk. Gut. 2000;47(5):703–9.

37.Huang YS, Chern HD, Wu JC, Chao Y, Huang YH, Chang FY, et al. Polymorphism of the N-acetyltransferase 2 gene, red meat intake, and the susceptibility of hepatocellular carcinoma. The American journal of gastroenterology. 2003;98(6):1417–22.

38.Akhdar H, El Shamieh S, Musso O, Desert R, Joumaa W, Guyader D, et al. The rs3957357C>T SNP in GSTA1 Is Associated with a Higher Risk of Occurrence of Hepatocellular Carcinoma in European Individuals. PloS one. 2016;11(12):e0167543.

39.Coles BF, Morel F, Rauch C, Huber WW, Yang M, Teitel CH, et al. Effect of polymorphism in the human glutathione S-transferase A1 promoter on hepatic GSTA1 and GSTA2 expression. Pharmacogenetics. 2001;11(8):663–9.

40.Alisi A, Mele R, Spaziani A, Tavolaro S, Palescandolo E, Balsano C. Thr 446 phosphorylation of PKR by HCV core protein deregulates G2/M phase in HCC cells. J Cell Physiol. 2005;205(1):25–31.

41.Hiasa Y, Kamegaya Y, Nuriya H, Onji M, Kohara M, Schmidt EV, et al. Protein kinase R is increased and is functional in hepatitis C virus-related hepatocellular carcinoma. The American journal of gastroenterology. 2003;98(11):2528–34.

Tables

Table 1. The top 5 genes according to gene-based *p*-values

Gene	Locus	Type	Nominal *p*	Corrected *p* [a]
RNY4	7q36	non-coding RNA	1.20×10^{-6}	0.030
GOLGA8M	15q13.1	protein-coding gene	3.00×10^{-6}	0.037
LINC01207	4q32	non-coding RNA	5.73×10^{-6}	0.047
WHAMMP2	15q13.1	pseudogene	8.90×10^{-6}	0.054
SLC39A8	4q24	protein-coding gene	1.08×10^{-5}	0.054

[a] The *p*-values are corrected by the Benjamini-Hochberg FDR approach.

Table 2. Genetic association *p*-values of genes preferentially expressed in liver

Gene Symbol[a]	*p*	CHR	Start Position	Length (BP)	Number of SNPs
NAT2	0.010	8	18248754	9970	69
GSTA1	0.011	6	52656177	12589	50
GSTA2	0.013	6	52614884	13478	36
UGT2B10	0.017	4	69870294	-172558	116
UROC1	0.027	3	126200007	36610	89
AQP9	0.032	15	58430394	47714	182
HAO1	0.033	20	7863630	57464	117
TF	0.036	3	133464976	32875	93
SAA2	0.037	11	18266774	3448	51
C3	0.041	19	6677845	42849	138

Note. CHR: chromosome; BP: base pairs.

[a] Only the genes with a *p*-value less than 0.05 are listed in this table. The whole gene list is shown in Supplementary Table 1.

Table 3. Summary of genetic association results in the replication

Batch ID	CHR	SNP	BP	CADD.CScore	SuRFR	FunSeq2	HCCCell_Prob	RegulomeDB	A1	A2	Additive[a]		Dominant[a]		Recessive[a]	
											OR (95% CI)	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
1	1	rs3813948	207269858	-0.039	14.356	0.7635	0.796	5	C	T	1.04 (0.87-1.24)	0.668	1.02 (0.83-1.25)	0.845	1.27 (0.72-2.23)	0.412
1	2	rs60325402	16077873	0.144	17.3	0.1852	0.370	5	A	T	0.85 (0.59-1.21)	0.360	0.83 (0.58-1.19)	0.319	-[b]	0.999
1	3	rs7612684	178984575	-0.163	19.334	0.8109	0.370	4	G	A	0.79 (0.59-1.05)	0.105	0.76 (0.56-1.03)	0.080	1.53 (0.23-10.01)	0.657
1	3	rs76863563	178987536	-0.498	15.493	0.1881	0.370	5	C	T	0.91 (0.66-1.26)	0.567	0.89 (0.64-1.24)	0.506	2.00 (0.17-24.06)	0.587
1	5	rs116966235	57794613	-0.636	.	0.1852	0.370	3a	G	A	1.07 (0.79-1.46)	0.670	1.06 (0.78-1.45)	0.705	-[b]	0.999
1	5	rs12514619	1783655	1.741	7.556	2.705	0.370	2b	C	T	1.10 (0.94-1.28)	0.252	1.06 (0.87-1.28)	0.563	1.45 (0.96-2.20)	0.078
1	6	rs389883	31947460	0.142	14.213	1.623	0.370	1f	G	T	0.86 (0.75-0.98)	0.026	0.86 (0.71-1.03)	0.108	0.73 (0.55-0.97)	0.032
1	6	rs615672	32574171	-0.162	4.627	0.7972	0.370	6	G	C	0.93 (0.81-1.07)	0.293	0.98 (0.81-1.17)	0.795	0.74 (0.54-1.01)	0.056
1	7	rs17343667	6065194	0.392	15.543	0.8898	0.370	1f	A	G	1.11 (0.96-1.27)	0.151	1.27 (1.04-1.55)	0.020	0.97 (0.76-1.24)	0.792
1	7	rs55744175	18332396	2.275	17.195	0.6909	0.370	5	A	G	1.05 (0.90-1.24)	0.524	1.07 (0.89-1.30)	0.474	1.02 (0.65-1.62)	0.924
1	8	rs16898013	124138891	0.780	17.314	0	0.370	3a	A	G	0.85 (0.63-1.16)	0.306	0.85 (0.61-1.16)	0.304	0.82 (0.11-6.23)	0.847
1	8	rs2275959	37455059	0.245	6.377	0.3114	0.863	4	A	G	0.98 (0.86-1.12)	0.791	1.02 (0.83-1.25)	0.854	0.93 (0.74-1.16)	0.503
1	8	rs2736020	15714529	-0.002	3.977	9.418E-161	0.370	7	C	T	1.09 (0.94-1.25)	0.255	1.13 (0.93-1.36)	0.209	1.06 (0.79-1.44)	0.687
1	10	rs3001719	10409365	-0.113	3.277	0.1852	0.370	5	G	T	1.08 (0.94-1.25)	0.288	1.11 (0.92-1.34)	0.261	1.08 (0.76-1.52)	0.674
1	11	rs10897243	62043174	-0.497	15.511	4.535E-33	0.370	6	G	C	0.92 (0.79-1.08)	0.311	0.93 (0.77-1.13)	0.468	0.81 (0.55-1.20)	0.296
1	12	rs79475045	39083557	-0.264	15.822	0.1881	0.370	5	T	G	0.88 (0.73-1.06)	0.189	0.91 (0.74-1.12)	0.377	0.55 (0.29-1.06)	0.072
1	12	rs979722	118217304	0.014	15.899	0.4365	0.370	7	C	T	1.05 (0.91-1.20)	0.512	1.05 (0.87-1.27)	0.597	1.09 (0.81-1.46)	0.577
1	16	rs12918376	56558181	-0.025	12.043	4.562E-74	0.370	6	T	G	1.11 (0.96-1.27)	0.153	1.11 (0.91-1.35)	0.303	1.19 (0.92-1.54)	0.182
1	20	rs2425046	33871661	0.090	17.787	1.78	0.918	2b	C	T	0.98 (0.77-1.24)	0.848	0.92 (0.72-1.19)	0.540	2.20 (0.79-6.14)	0.134

Note. CHR: chromosome; BP: base pairs; OR: odd ratio; CI: confidence interval; A1: minor allele; A2: major allele; CADD.CScore, SuRFR and FunSeq2 scores are annotated by KGGSeq (V1.0). HCCCell_Prob: Probability of cell type-specific regulation in GENCODE liver cancer cells (HepG2).

[a] This model was tested under Logistic regression model with adjustment for age and sex.

[b] The value is not available.

Figures

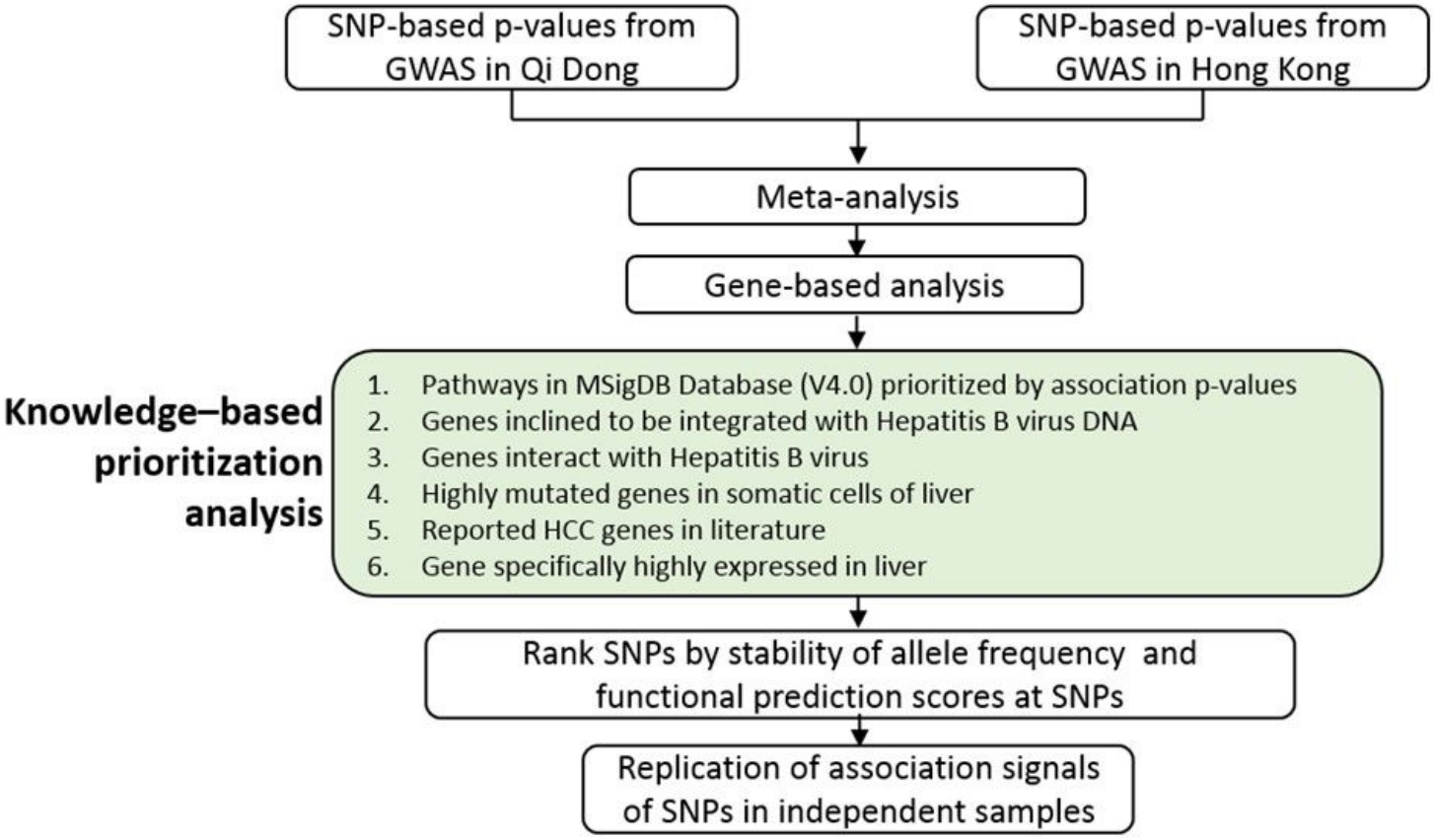

Figure 1

Knowledge-based prioritization framework of SNPs' statistical p-values for association with HCC

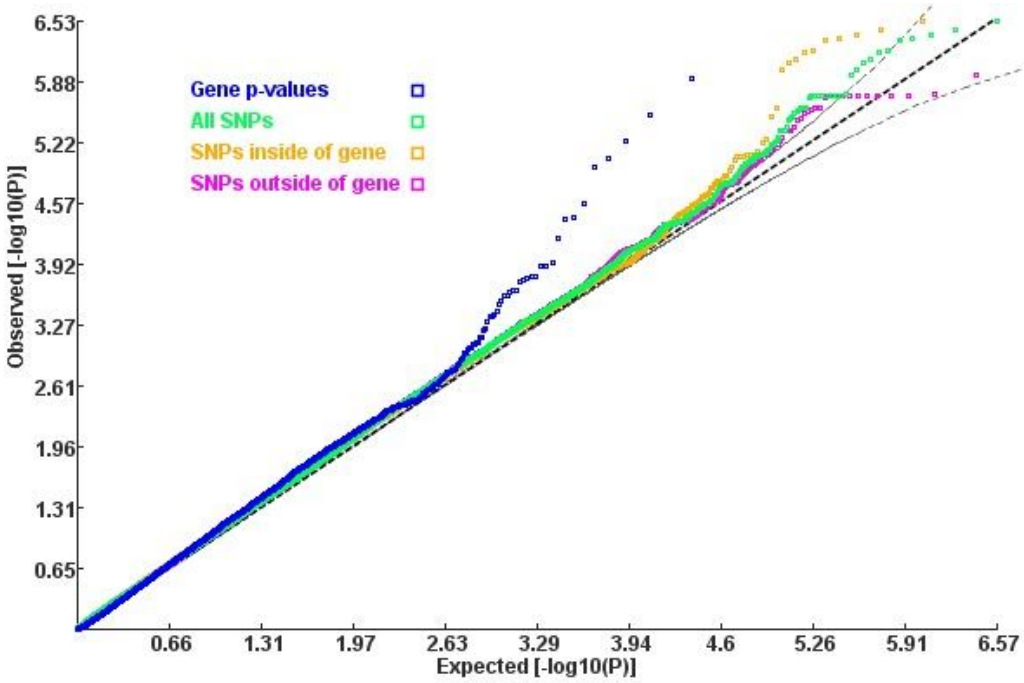

Figure 2

Quantile-quantile plot of gene-based p-values and SNP-based p-values

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx