Knowledge-based analyses reveal new candidate genes associated with risk 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: March 18th, 2020

DOI: https://doi.org/10.21203/rs.2.14627/v3

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 knowledge-based analysis.

Methods: We performed knowledge-based analysis (including gene- and gene-set-based association tests) on variant-level association *p*-values from two existing GWASs of HBV-related HCC. Five different types of gene-sets were collected for the association analysis. A number of SNPs within the gene prioritized by the knowledge-based association tests were selected to replicate genetic associations in an independent sample of 965 cases and 923 controls.

Results: The gene-based association analysis detected four genes significantly or suggestively associated with HBV-related HCC risk: *SLC39A8*, *GOLGA8M*, *SMIM31*, and *WHAMMP2*. The gene-set-based association analysis prioritized two promising gene set for HCC, cell cycle G1/S transition and NOTCH1 intracellular domain regulates transcription. Within the gene sets, three promising candidate genes (*CDC45*, *NCOR1* and *KAT2A*) were further prioritized for HCC. Among genes of liver-specific expression, multiple genes previously implicated in HCC were also highlighted. However, probably due to small sample size, none of the genes prioritized by the knowledge-based association analyses are successfully replicated in the independent sample.

Conclusions: This comprehensive knowledge-based association mining study suggested several promising genes and gene-sets associated with HBV-related HCC risks, which facilitate follow-up functional studies on 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 series of knowledge-based analyses (including gene- and gene-set-based association tests) on variant-level association *p*-values from two in-house GWASs of HBV-related HCC. SNPs within genes prioritized by the knowledge-based analyses 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 inclusion and exclusion criteria for all the subjects have been previously described [7, 11]. Briefly, all the subjects were negative for antibodies to hepatitis C virus, or human immunodeficiency virus; and had no other types of liver disease, such as autoimmune hepatitis, toxic hepatitis, and primary biliary cirrhosis. All the controls were chronic HBV carriers and had, by self-report, no history of HCC or other cancers. Chronic HBV carriers were defined as positive for both hepatitis B surface antigen and antibody immunoglobulin G to hepatitis B core antigen for at least 6 months. All the cases were chronic HBV carriers and diagnosed as HCC patients. The diagnosis of HCC was based on a) positive findings on cytological or pathological examination and/or b) positive images on angiogram, ultrasonography, computed tomography and/or magnetic resonance imaging, combined with an Alpha-fetoprotein level ≥ 400 ng/ml. All the cases were confirmed to not have other cancers by an initial screening. The mean (standard deviation) ages of the cases and controls were 50.8 (±12.2) years and 52.9 (±11.2) years, respectively. The male to female ratio were 5.3 in cases and 1.6 in controls, respectively.

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:

See formula 1 in the supplementary files.

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. Two types of gene-based association tests, GATES [13] and ECS [14], were employed for the analysis which combined SNP-level association signal according to the best significance and accumulated significance respectively. In addition, LDRT [15] was adopted for gene-set-based association analysis. The phased genotypes of Eastern Asian samples in the 1000 Genomes Project [16] 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 or genes within gene-sets, 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) [17] and RegulomeDB Version 1.1 (http://regulomedb.org/) [18], 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) [19, 20] was used to annotate selected SNP with four regulatory or functional prediction scores (including CADD.CScore [21], SuRFR [22], FunSeq2 [23] and cepip [24]).

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. The small proportion of significant signals was consistent with the estimated low heritability in the samples by GCTA, 0.063 (±0.028) on the underlying liability scale [25].

Gene-based association analysis

We then used the meta-analysis *p*-values for gene-based association analysis by GATES [13] and ECS [14] on a tool called KGG (version 3.5) [26]. 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 by GATES and ECS. 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 by GATES, two genes, *SLC39A8* and *GOLGA8M* passed the multiple-testing correction by FDR, 0.05 (Table 1). In addition, two genes, *SMIM31* and *WHAMMP2*, have nearly significant *q*-values (<0.06 by GATES) on the genome (Table 1). Interestingly, *SMIM31*, encoding small integral membrane protein 31, was annotated as a long noncoding RNA gene (*LINC01207*) previously. We further annotated the pseudogene, *WHAMMP2*, with known regulatory elements and epigenomic markers by the UCSC genome browser (http://genome.ucsc.edu). Although it is annotated as a pseudogene, there are multiple regulatory factors binding sites and epigenomic markers in *WHAMMP2* (See Supplementary Figure 3). These annotations imply that this gene is also functionally active despite not encoding proteins. The other gene-based test, ECS, detected no significant gene. The gene with smallest *p*-value (7.5E-06) is *RNF157-AS1*.

Prioritization of genes in different gene-sets

To select more promising candidate 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. The gene-set-based association *p*-value was performed by LDRT [15] on KGG. Although no gene-sets passed multiple testing (FDR $q<0.05$), several promising functional gene sets are prioritized. The top two gene sets according to the *p*-value are the cell cycle G1/S transition ($p=5.5E-4$) and the NOTCH1 intracellular domain regulates transcription ($p=7.1E-4$). In the G1/S transition gene set, 12 out 99 genes had gene-based association ($p<0.05$, See details in Supplementary Excel Table 1). The gene with the smallest *p*-value is *CDC45* ($p=1.1E-4$) in this gene set. In the gene set of NOTCH1 intracellular domain regulates transcription, 10 out 40 genes had gene-based association ($p<0.05$, See details in Supplementary Excel Table 1). In the set, *NCOR1* had the smallest *p*-value ($p=5.8E-3$). The second gene, *KAT2A*, had similar *p*-value (6.6E-3).

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. Three genes (*PAH*, *UGT2B10* and *UROC1*) had

the FDR *q* values <0.1 by ECS while GATES did not detect any significant gene (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 [27-30], 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. Due to budget limit, only 21 SNPs were selected for the replication. The SNPs were at prioritized genes according to consistency of their allele frequencies in ancestry matched reference panel in the 1000 Genomes Project and HapMap Project, and/or their predicted functional importance by RegulomeDB (http://regulomedb.org/) with regulatory elements (See examples in Supplementary Figures 3 and 4). 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. None of the 19 SNPs survived the multiple Bonferroni correction for family-wise error rate 0.05. Only two SNPs, rs17343667 and rs389883, had a nominal *p*-value below 0.05. The rs17343667, which is located in the first intron of *EIF2AK1*, had an association *p*-value equal to 0.02 under the dominant 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). However, its *p*-value was only 0.15 under the additive model. 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. Therefore, the SNP-level replication was generally negative.

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 four suggestively significant genes including *SLC39A8*, *GOLGA8M*, *SMIM31*and *WHAMMP2*. The gene-set-based association analysis prioritized three top genes (*CDC45*, *NCOR1* and *KAT2A*), which have been implicated with HCC previously, mainly through regulated expression. In addition, three genes, *PAH*, *UGT2B10* and *UROC1* were also highlighted when multiple-testing correction (FDR q<0.1) was performed among genes highly and specifically expressed in human liver. However, probably due to small sizes in our replication samples, no associations prioritized by the knowledge-based association analysis were successfully replicated in an independent sample. The rs17343667 of *EIF2AK1* is the only one with suggestive significance. 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.

According to our estimation, HCC has relatively low heritability (6.3%). It is unlikely that there are susceptibility genes or loci of large effect size. The association test enriched the association signals of multiple loci in multiple genes with low effect size so that the susceptibility pathways and gene sets can be prioritized. Moreover, it is easier to

prioritize potential susceptibility genes given the prioritized gene sets. In our analysis, a non-trivial fraction of genes within the gene sets achieved moderately significant *p*-values. It is likely that some of the genes may achieve genome-wide significance when sample sizes are increased. However, almost all of the genes would be ignored by the genome-wide *p*-value threshold in the present samples (1307 cases vs.1574 controls).

Our study is the first to show that genetic variations of two genes (*SLC39A8* and *GOLGA8M*) are significantly associated with the development of HBV-related HCC. *SLC39A8* encodes a member of the *SLC39* family of solute-carrier genes (Zrt/Irt-like protein 8, ZIP8), which may play an important role in autophagy during ethanol exposure in human hepatoma cells [31]. Liu et al. suggested that hepatic ZIP8 deficiency was associated with tumor formation [32]. Moreover, *SLC39A8* has been reported to regulate IFN-γ level in T cells [33] and influence trace element homeostasis in liver [34, 35], which may be relevant to the development of HCC. *GOLGA8M* encodes golgin A8 family member M. Although it has not been linked to cancer, a study suggested that palindromic *GOLGA8* core duplicons promoted chromosome microdeletion and evolutionary instability [36]. In addition, two other genes (*SMIM31* and *WHAMMP2*) also achieved suggestively significant *p*-values. *SMIM31* has been implicated as a biomarker for survival of colorectal adenocarcinoma [37] and promoting proliferation of lung adenocarcinoma [38]. *RNF157-AS1*, which was implicated by ECS, is an antisense RNA gene. Differential expression between tumor and non-tumor tissue at this gene has been founded in lung cancer [39] and ovarian cancer [40]. Anyhow, functional validation studies are needed to explore the mechanisms of the potential roles of these genes in risk of HBV-related HCC.

The successful prioritization of two gene sets that are highly relevant to cancer development also implies the power of the knowledge-based analysis. The top two functional gene-sets are cell cycle G1/S transition and NOTCH1 intracellular domain regulates transcription. There have been numerous studies linking these functional gene sets to HCC [41-44]. For example, Wang et al. recently showed that lnc-UCID promotes G1/S Transition and hepatoma growth by preventing DHX9-Mediated CDK6 down-regulation [41]. As the gene with the smallest *p*-value in the cell cycle G1/S transition gene set, *CDC45* encodes cell division control protein 45 and has been linked to many cancers according to its expression, including HCC [45] and colorectal cancer [46]. *NCOR1*, the gene with the smallest *p*-value in the gene set of NOTCH1 intracellular domain regulates transcription, encodes a protein that mediates ligand-independent transcription repression of thyroid-hormone and retinoic-acid receptors, which may regulate de novo fatty acids synthesis in liver regeneration and hepatocarcinogenesis in mice [47]. For another gene with similar *p*-value as *NCOR1* in the gene set of NOTCH1 intracellular domain regulates transcription, *KAT2A* encodes lysine acetyltransferase 2A and was linked to HCC. For instance, Majaz et al. suggested that KAT2A may promote human HCC progression by enhancing AIB1 expression [48].The highly and specifically expression in human liver is also an effective stratum for prioritization of HCC susceptibility genes. When multiple testing correction is carried out in this gene set, three genes *PAH*, *UGT2B10* and *UROC1* achieved suggestive significance level (FDR $q<0.1$). All of the three genes have been implicated with HCC by multiple studies. The most significant gene *PAH* ($p=3.5E-4$ and $q=0.064$) has the largest number of literature supports, that is, many studies have implicated this gene in development of HCC. For example, Miller et al. showed p-Chlorphenylalanine effect on phenylalanine hydroxylase in hepatoma cells in culture [49]. Gopalakrishnan and Anderson showed the epigenetic activation of phenylalanine hydroxylase in mouse erythroleukemia cells by the cytoplast of rat hepatoma cells [50]. *UGT2B10* ($p=7.9E-4$) encodes UDP-Glucuronosyltransferase 2B10. Hanioka et al. showed that expression of *UGT2B* isoforms (including *UGT2B10*) was significantly increased by AFB1 in HepG2 cells [51]. *UROC1* ($p=1.4E-3$) encodes enzyme involved in histidine catabolism, metabolizing urocanic acid to formiminoglutamic acid. Zhang et al. showed that UROC1 may play important roles in HCC development, especially alcohol-related HCC development and progression [52].

The negative findings in the curated gene sets of recurrent targeted genes of HBV infection and HCC risk genes with many somatic mutations are unexpected to some extent. Both gene sets appeared to be 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 issue of negative association at variants in replication sample is consistent with that in the discovery sample. Due to low effect size, no variants in the discovery GWAS sample of 1,307 HBV-related HCC cases and 1,574 controls had a *p*-value less than the widely-adopted genome-wide cutoff (5E-8). It was the gene-based association analysis combing the *p*-values of multiple SNPs that achieved genome-wide significant *p*-values at some genes. Because of budget limit, however, most genes only had one selected SNPs to maximize the total number of genes for replication. Therefore, we were unable to carry out the gene-based association in the replication study as we did in the GWAS sample. Unfortunately, probably due to low effect size, no variants achieved significant *p*-value in the replication sample of 965 HBV-related HCC cases and 923 controls. The SNP-level negative replication implies either more powerful gene-based association study or larger sample is needed for identifying HCC susceptibility genes.

In 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 failure in replication study also implies small effect size of the susceptibility genes. More hypothesis-free genetic studies with larger sample sizes are needed to elucidate the susceptibility genes and mechanisms of HCC.

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 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 (No. 01121436 and 02132236 to M.X.L.), the National Natural Science Foundation of China (No. 31100895, 81472618 and 81670535 to D.K.J., No. 81402297 to Q.L.X.), the National Science and Technology Major Project (No. 2017ZX10202202 and 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 Research and Development Project in Key Areas of Guangdong Province (No. 2019B020227004 to D.K.J.), the Innovative Research Team Project of Guangxi Province (No. 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.).

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.

Acknowledgements

Not applicable.

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.

References

1. Pinyol R, Llovet JM: **Hepatocellular carcinoma: genome-scale metabolic models for hepatocellular carcinoma**. *Nat Rev Gastroenterol Hepatol* 2014, **11**(6):336-337.
2. Arzumanyan A, Reis HM, Feitelson MA: **Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma**. *Nat Rev Cancer* 2013, **13**(2):123-135.
3. Yu MW, Chang HC, Liaw YF, Lin SM, Lee SD, Liu CJ, Chen PJ, Hsiao TJ, Lee PH, Chen CJ: **Familial risk of hepatocellular carcinoma among chronic hepatitis B carriers and their relatives**. *J Natl Cancer Inst* 2000, **92**(14):1159-1164.
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 journal of gastroenterology* 2003, **9**(11):2428-2432.
5. Zhang H, Zhai Y, Hu Z, Wu C, Qian J, Jia W, Ma F, Huang W, Yu L, Yue 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):755-758.

6. Li S, Qian J, Yang Y, Zhao W, Dai J, Bei JX, Foo JN, McLaren PJ, Li Z, Yang J *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, Ren WH, Long XD, Zhang H, Ma XP *et al*: **Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma**. *Nature genetics* 2013, **45**(1):72-75.
8. Manolio TA: **Bringing genome-wide association findings into clinical use**. *Nature reviews Genetics* 2013, **14**(8):549-558.
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, Lai CL, Poon RT, Sham PC, Ng IO: **Genome-wide association study of hepatocellular carcinoma in Southern Chinese patients with chronic hepatitis B virus infection**. *PloS one* 2011, **6**(12):e28798.
11. Jiang DK, Ma XP, Yu H, Cao G, Ding DL, Chen H, Huang HX, Gao YZ, Wu XP, Long XD *et al*: **Genetic variants in five novel loci including CFB and CD40 predispose to chronic hepatitis B**. *Hepatology* 2015, **62**(1):118-128.
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):761-766.
13. Li MX, Gui HS, Kwan JS, Sham PC: **GATES: a rapid and powerful gene-based association test using extended Simes procedure**. *American journal of human genetics* 2011, **88**(3):283-293.
14. Li M, Jiang L, Mak TSH, Kwan JSH, Xue C, Chen P, Leung HC, Cui L, Li T, Sham PC: **A powerful conditional gene-based association approach implicated functionally important genes for schizophrenia**. *Bioinformatics* 2019, **35**(4):628-635.
15. Gui H, Kwan JS, Sham PC, Cherny SS, Li M: **Sharing of Genes and Pathways Across Complex Phenotypes: A Multilevel Genome-Wide Analysis**. *Genetics* 2017, **206**(3):1601-1609.
16. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M *et al*: **An integrated map of structural variation in 2,504 human genomes**. *Nature* 2015, **526**(7571):75-81.
17. Ward LD, Kellis M: **HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants**. *Nucleic acids research* 2012, **40**(Database issue):D930-934.
18. Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M: **Dynamic trans-acting factor colocalization in human cells**. *Cell* 2013, **155**(3):713-724.
19. Li M, Li J, Li MJ, Pan Z, Hsu JS, Liu DJ, Zhan X, Wang J, Song Y, Sham PC: **Robust and rapid algorithms facilitate large-scale whole genome sequencing downstream analysis in an integrative framework**. *Nucleic acids research* 2017, **45**(9):e75.
20. 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.
21. 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):310-315.
22. 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.

23. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M: **FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer**. *Genome biology* 2014, **15**(10):480.
24. Li MJ, Li M, Liu Z, Yan B, Pan Z, Huang D, Liang Q, Ying D, Xu F, Yao H *et al*: **cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes**. *Genome biology* 2017, **18**(1):52.
25. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis**. *American journal of human genetics* 2011, **88**(1):76-82.
26. 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.
27. Paterlini-Brechot P, Saigo K, Murakami Y, Chami M, Gozuacik D, Mugnier C, Lagorce D, Brechot C: **Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene**. *Oncogene* 2003, **22**(25):3911-3916.
28. Ding D, Lou X, Hua D, Yu W, Li L, Wang J, Gao F, Zhao N, Ren G, Li L *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.
29. Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C *et al*: **Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma**. *Nature genetics* 2012, **44**(7):765-769.
30. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, Carnevali P, Stinson J, Johnson S *et al*: **The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients**. *Genome research* 2012, **22**(4):593-601.
31. Liuzzi JP, Yoo C: **Role of zinc in the regulation of autophagy during ethanol exposure in human hepatoma cells**. *Biological trace element research* 2013, **156**(1-3):350-356.
32. Liu L, Geng X, Cai Y, Copple B, Yoshinaga M, Shen J, Nebert DW, Wang H, Liu Z: **Hepatic ZIP8 deficiency is associated with disrupted selenium homeostasis, liver pathology, and tumor formation**. *American journal of physiology Gastrointestinal and liver physiology* 2018, **315**(4):G569-G579.
33. 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):337-348.
34. Lin W, Vann DR, Doulias PT, Wang T, Landesberg G, Li X, Ricciotti E, Scalia R, He M, Hand NJ *et al*: **Hepatic metal ion transporter ZIP8 regulates manganese homeostasis and manganese-dependent enzyme activity**. *The Journal of clinical investigation* 2017, **127**(6):2407-2417.
35. Engelken J, Espadas G, Mancuso FM, Bonet N, Scherr AL, Jimenez-Alvarez V, Codina-Sola M, Medina-Stacey D, Spataro N, Stoneking M *et al*: **Signatures of Evolutionary Adaptation in Quantitative Trait Loci Influencing Trace Element Homeostasis in Liver**. *Molecular biology and evolution* 2016, **33**(3):738-754.
36. Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA, Miroballo M, Graves TA, Vives L, Malig M *et al*: **Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability**. *Nature genetics* 2014, **46**(12):1293-1302.
37. Zeng JH, Liang L, He RQ, Tang RX, Cai XY, Chen JQ, Luo DZ, Chen G: **Comprehensive investigation of a novel differentially expressed lncRNA expression profile signature to assess the survival of patients with colorectal adenocarcinoma**. *Oncotarget* 2017, **8**(10):16811-16828.
38. Wang G, Chen H, Liu J: **The long noncoding RNA LINC01207 promotes proliferation of lung adenocarcinoma**. *American journal of cancer research* 2015, **5**(10):3162-3173.

39. Qiao F, Li N, Li W: **Integrative Bioinformatics Analysis Reveals Potential Long Non-Coding RNA Biomarkers and Analysis of Function in Non-Smoking Females with Lung Cancer**. *Medical science monitor : international medical journal of experimental and clinical research* 2018, **24**:5771-5778.
40. Zhan L, Li J, Wei B: **Long non-coding RNAs in ovarian cancer**. *Journal of experimental & clinical cancer research : CR* 2018, **37**(1):120.
41. Wang YL, Liu JY, Yang JE, Yu XM, Chen ZL, Chen YJ, Kuang M, Zhu Y, Zhuang SM: **Lnc-UCID Promotes G1/S Transition and Hepatoma Growth by Preventing DHX9-Mediated CDK6 Down-regulation**. *Hepatology* 2019, **70**(1):259-275.
42. Liu RY, Diao CF, Zhang Y, Wu N, Wan HY, Nong XY, Liu M, Tang H: **miR-371-5p down-regulates pre mRNA processing factor 4 homolog B (PRPF4B) and facilitates the G1/S transition in human hepatocellular carcinoma cells**. *Cancer letters* 2013, **335**(2):351-360.
43. Zhang L, Chen J, Yong J, Qiao L, Xu L, Liu C: **An essential role of RNF187 in Notch1 mediated metastasis of hepatocellular carcinoma**. *Journal of experimental & clinical cancer research : CR* 2019, **38**(1):384.
44. Fang S, Liu M, Li L, Zhang FF, Li Y, Yan Q, Cui YZ, Zhu YH, Yuan YF, Guan XY: **Lymphoid enhancer-binding factor-1 promotes stemness and poor differentiation of hepatocellular carcinoma by directly activating the NOTCH pathway**. *Oncogene* 2019, **38**(21):4061-4074.
45. Sang L, Wang XM, Xu DY, Zhao WJ: **Bioinformatics analysis of aberrantly methylated-differentially expressed genes and pathways in hepatocellular carcinoma**. *World journal of gastroenterology* 2018, **24**(24):2605-2616.
46. Yang S, Ren X, Liang Y, Yan Y, Zhou Y, Hu J, Wang Z, Song F, Wang F, Liao W *et al*: **KNK437 restricts the growth and metastasis of colorectal cancer via targeting DNAJA1/CDC45 axis**. *Oncogene* 2019.
47. Ou-Yang Q, Lin XM, Zhu YJ, Zheng B, Li L, Yang YC, Hou GJ, Chen X, Luo GJ, Huo F *et al*: **Distinct role of nuclear receptor corepressor 1 regulated de novo fatty acids synthesis in liver regeneration and hepatocarcinogenesis in mice**. *Hepatology* 2018, **67**(3):1071-1087.
48. Majaz S, Tong Z, Peng K, Wang W, Ren W, Li M, Liu K, Mo P, Li W, Yu C: **Histone acetyl transferase GCN5 promotes human hepatocellular carcinoma progression by enhancing AIB1 expression**. *Cell & bioscience* 2016, **6**:47.
49. Miller MR, McClure D, Shiman R: **p-Chlorphenylalanine effect on phenylalanine hydroxylase in hepatoma cells in culture**. *The Journal of biological chemistry* 1975, **250**(3):1132-1140.
50. Gopalakrishnan TV, Anderson WF: **Epigenetic activation of phenylalanine hydroxylase in mouse erythroleukemia cells by the cytoplast of rat hepatoma cells**. *Proceedings of the National Academy of Sciences of the United States of America* 1979, **76**(8):3932-3936.
51. Hanioka N, Nonaka Y, Saito K, Negishi T, Okamoto K, Kataoka H, Narimatsu S: **Effect of aflatoxin B1 on UDP-glucuronosyltransferase mRNA expression in HepG2 cells**. *Chemosphere* 2012, **89**(5):526-529.
52. Zhang X, Kang C, Li N, Liu X, Zhang J, Gao F, Dai L: **Identification of special key genes for alcohol-related hepatocellular carcinoma through bioinformatic analysis**. *PeerJ* 2019, **7**:e6375.

Tables

Table 1. The top 5 genes according to gene-based *p*-values by GATES and ECS, respectively

Gene	CHR	Type	#SNP	GATES		ECS	
				Nominal	Corrected	Nominal	Corrected
				p	*p* [a]	*p*	*p* [a]
SLC39A8	4	protein-coding gene	222	1.63E-06	0.04138	0.22495	0.83656
GOLGA8M	15	protein-coding gene	11	3.19E-06	0.04138	1.57E-05	0.20422
SMIM31	4	protein-coding gene	300	6.43E-06	0.05560	0.00424	0.52238
WHAMMP2	15	pseudogene	14	9.03E-06	0.05858	0.00031	0.32182
CLDN5	22	protein-coding gene	22	2.76E-05	0.12596	0.01665	0.62074
RNF157-AS1	17	non-coding RNA	20	3.84E-05	0.12655	7.50E-06	0.19448
LRRC9	14	other	348	0.00249	0.62765	4.46E-05	0.32182
LINC02062	5	non-coding RNA	150	0.02402	0.69927	0.00007	0.32182
TTL	2	protein-coding gene	101	0.01254	0.65610	7.64E-05	0.32182

Note. CHR: chromosome.

[a] The *p*-values are corrected by the Benjamini-Hochberg FDR approach. *SLC39A8*, *GOLGA8M*, *SMIM31*, *WHAMMP2* and *CLDN5* are the top five genes according to GATES. *RNF157-AS1*, *GOLGA8M*, *LRRC9*, *LINC02062* and *TTL* are the top five genes according to ECS.

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

Gene Symbol[a]	GATES	ECS	ECS	CHR	Start Position	Length (BP)	Number of SNPs
	p	*p*	*q*				
PAH	>0.05	0.00035	0.064	12	103230666	80356	266
UGT2B10	0.01504	0.00079	0.073	4	69870294	172553	122
UROC1	0.02728	0.00138	0.085	3	126200008	36608	92
TF	0.00293	0.01388	0.386	3	133465236	50249	288
C4A	>0.05	0.01472	0.386	6	31949833	20624	20
SLCO1B1	>0.05	0.01528	0.386	12	21284127	108603	298
C5	>0.05	0.01605	0.386	9	123761950	50603	150
GSTA2	0.04013	0.01864	0.386	6	52614884	13389	59
C4B	>0.05	0.02012	0.386	6	31982571	12113	38
HAO1	0.03276	0.02195	0.386	20	7863631	57474	118
NAT2	0.01023	0.02308	0.386	8	18248791	9934	70
GSTA1	0.01104	>0.05	0.697	6	52656170	12444	50
AQP9	0.03199	>0.05	0.733	15	58430579	47531	182
SAA2	0.04169	>0.05	0.804	11	18266786	3429	55
APOA2	0.04277	>0.05	0.733	1	161192081	1337	22

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

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	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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

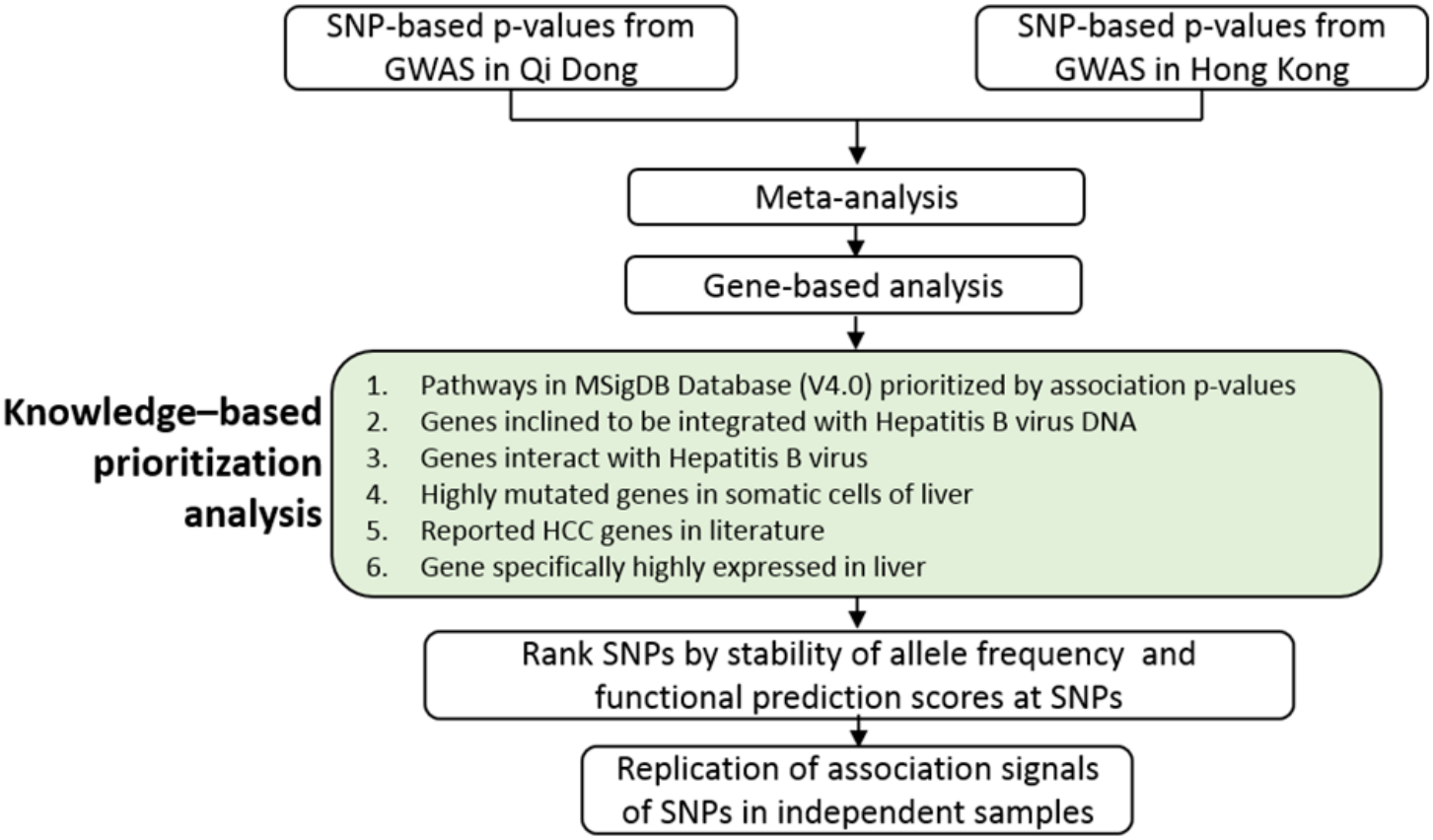

Figure 1

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

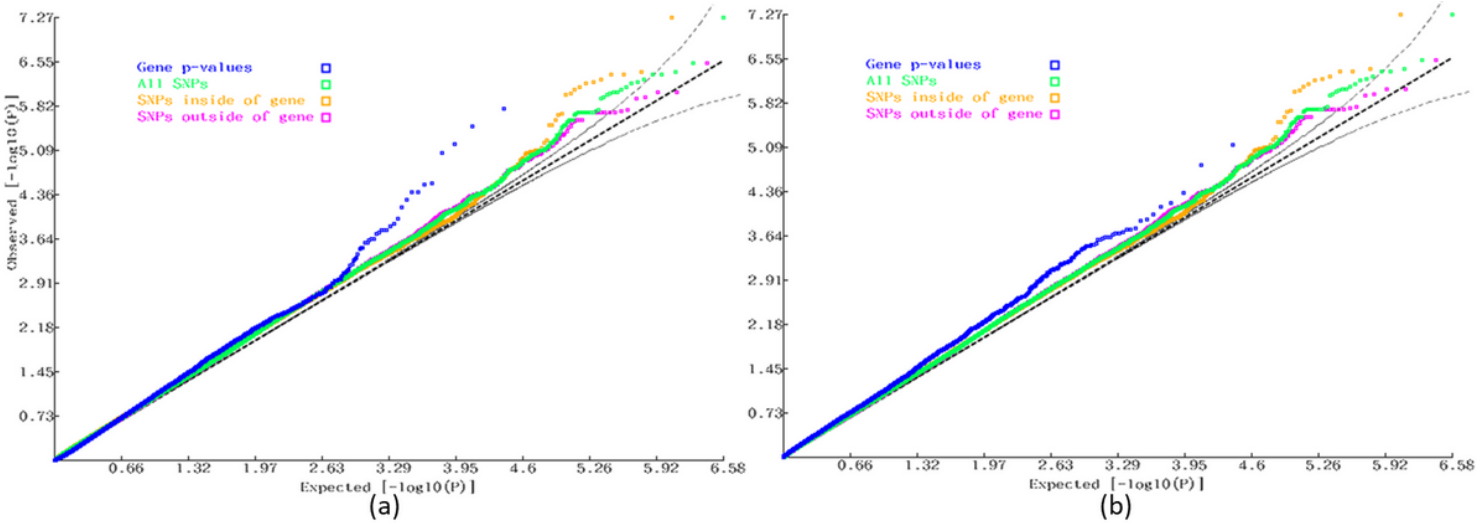

Figure 2

Quantile-quantile plot of gene-based p-values and SNP-based p-values a) the p-values produced by GATES b) the p-values produced by ECS.

Supplementary Files

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

- formulas.docx
- BCAND1902656Supplementarymaterials.docx
- BCAND1902656SupplementaryExcelTable1.xlsx