sciendo

Research article

DOI:10.2478/rrlm-2020-0008

Identification of Susceptibility Genes in Hepatic Cancer Using Whole Exome Sequencing and Risk Prediction Model Construction

Jinghui Zheng[1,2*], Youming Tang[3], Encun Hou[4], Guangde Bai[4], Zuping Lian[4], Peisheng Xie[4], Weizhi Tang[4]

1. Ruikang Hospital Affiliated to Guangxi University of Chinese Medicine, China
2. Academician Workstation, Ruikang Hospital Affiliated to Guangxi University of Chinese Medicine,China
3. Department of Gastroenterology, Ruikang Hospital Affiliated to Guangxi University of Chinese Medicine,China
4. Department of Oncology, Ruikang Hospital Affiliated to Guangxi University of Chinese Medicine,China

Abstract

***Objective**: To identify the susceptible single nucleotide polymorphisms (SNPs) loci in HCC patients in Guangxi Region, screen biomarkers from differential SNPs loci by using predictors, and establish risk prediction models for HCC, to provide a basis of screening high-risk individuals of HCC. **Methods**: Blood sample and clinical data of 50 normal participants and 50 hepatic cancer (HCC) patients in Rui Kang Hospital affiliated to Guangxi University of Traditional Chinese Medicine were collected. Normal participants and HCC patients were assigned to training set and testing set, respectively. Whole Exome Sequencing (WES) technique was employed to compare the exon sequence of the normal participants and HCC patients. Five predictors were used to screen the biomarkers and construct HCC prediction models. The prediction models were validated with both training and testing set. **Results**: Two-hundred seventy SNPs were identified to be significantly different from HCC, among which 100 SNPs were selected as biomarkers for prediction models. Five prediction models constructed with the 100 SNPs showed good sensitivity and specificity for HCC prediction among the training set and testing set. **Conclusion**: A series of SNPs were identified as susceptible genes for HCC. Some of these SNPs including CNN2, CD177, KMT2C, and HLA-DQB1 were consistent with the previously identified polymorphisms by targeted genes examination. The prediction models constructed with part of those SNPs could accurately predict HCC development.*

***Keywords:** SNP, biomarkers, HCC*

Received: 14th June 2019; Accepted: 1st December 2019; Published: 12th January 2020

***Corresponding author**: Jinghui Zheng, Ruikang Hospital Affiliated to Guangxi University of Chinese Medicine nanning, China. E-mail: jinghui3131@sina.com

Introduction

Primary hepatocellular carcinoma (HCC) is the fifth most common malignant tumor in the world. It is one of the tumors with the fastest rising incidence and highest mortality in recent years (1-3). HCC is a complex disease. One of the high-risk factors of HCC is the family history of HCC (4). Therefore, genetic factors have a great influence on the occurrence and development of HCC. It is necessary to make comprehensive studies on them, which may provide a basis of realizing early diagnosis and early detection in HCC high-risk population.

Frequency of Single Nucleotide Polymorphisms (SNPs) in the population was more than 1%. Forms of SNPs include single base conversion, transmutation, and insertion/deletion of single base, accounting for more than 90% of all known polymorphisms (5). A comparative study of SNPs in a malignant tumor population and a normal control population can determine the relationship between SNPs loci and/or their adjacent variants and the risk of cancer. Therefore, it has been widely applied to the study of genetic susceptibility to cancer.

Large sample genome-wide association studies can provide a global understanding of the susceptibility spectrum of hepatocellular carcinoma. At present, high-throughput second-generation sequence technology has been used to complete genome sequence of a variety of animals, plants and microorganisms. However, due to the high cost of whole genome sequence, in recent years Whole Exome Sequencing (WES) technology has been used to identify pathogenic genes of diseases induced by single gene and susceptible genes spectrum of complex diseases (6). The Exome is part of eukaryotic genes, accounting for 1% of the human genome (7). During Exome sequences, targeted enrichment technology is used to capture the exon region of genome, and then coding region sequences are obtained by high-throughput sequence. Mutation gene list is obtained by data analysis. By comparison between dbSNP database, common mutations are excluded, and then genetic characteristics of diseases are found (8,9). With the progress of sequence technology and the decrease of high-throughput sequences cost, WES technology has made great contributions to the study of single-gene diseases, and has gradually been used to screen susceptible genes for complex diseases, making it possible to use a relatively small sample size to identify susceptible genes (10,11).

Though WES technology is an efficient, fast and cost-effective research method, to date there are very few reports about the application of WES technology to study the genetic susceptibility of HCC. The purpose of this study is to identify the susceptible SNPs loci in HCC patients in Guangxi Region, screen biomarkers from differential SNPs loci by using predictors, and to establish risk prediction models for HCC, to provide a basis of screening high-risk individuals of HCC.

Methods

Participants, clinical data, and sample collection

Blood samples and clinical data of 50 normal participants, 50 liver cancer patients, and 20 participants with high risk factors of HCC in Rui Kang Hospital affiliated to Guangxi University of Traditional Chinese Medicine were collected. The criteria for the selection of patients with HCC are as follows: aged between 20-70 years; diagnosed as primary HCC according to the Guidelines of Diagnosis and Treatment of Common Malignant Tumors compiled by the Chinese Anti-Cancer Association; those with other serious diseases such as malignant tumors, chronic liver and kidney diseases were excluded; those with psychiatric diseases were excluded.

Normal participants were selected according to the following criteria: aged between 20-70 years; sex ratios matched with that of the HCC group;

no risk factors of HCC other than age and sex; no abnormality was found after medical history inquiry, physical examination, electrocardiogram examination, and other blood biochemical examinations.

Inclusion criteria for HCC high-risk participants were as follows: aged over 40 years; having more than two following risk factors of primary liver cancer: family history; HBV infection; cirrhosis; those with primary liver cancer were excluded by medical history inquiry, physical examination, and so forth.

Blood samples were collected from HCC patients, normal participants, and high-risk participants. The clinical data included: demographic information: age, sex, nationality, origin, etc.; lifestyle information: diet, smoking, drinking, and exercise; physical examination information: height, weight, abdominal circumference, blood pressure, heart rate, etc.; medical history: current medical history, past medical history, family history, etc.; Laboratory examination information; drug use and efficacy: drug name, dose, use time; imaging information.

The project received the ethical approval of the committees of Rui Kang Hospital and all patients signed informed consent regarding the use of the samples.

Genome DNA extraction and capture sequencing

Three ml anticoagulant blood sample was obtained from each participant and stored under -80°C. The genome of nucleated cells in peripheral blood was extracted with TIANamp Blood DNA kit. One ug genome DNA was randomly interrupted into 250-300 bp DNA fragments by DNA ultrasound high performance sample processing system (Covaris). The end of DNA fragments was then repaired. “A” base was added at the 3’ end and library adaptor was ligated at both ends. The hybridized libraries were prepared by linear amplification. An appropriate amount of hybridization libraries and exon chips was taken for enrichment capture. The un-enriched fragments were eluted and then amplification was performed. Quality control on the amplification products was made with Agilent 2100 Bio Analysis Instrument (Agilent DNA 1000 Reagents) and qPCR. High throughput sequence was performed using BGI-seq500 System (BGI). The original image data obtained by sequences was transformed into raw reading by Base Recognition software (Base Calling) and stored as FASTQ file.

Bioinformatics analysis on high throughput sequencing data

Raw data was preprocessed to obtain clean data. BWA software was used to evaluate the clean data to ensure the reliability of the data. Then, the data were compared with the reference genome (UCSChg19). The comparison results were format-transformed, sorted and duplicate-labeled to obtain the comparison files to be analyzed. GATK Haplotype Caller was used to make SNPs&InDeLs analysis on the comparison result. Then the high-quality SNP variation results were annotated by snpEFF.

Screening of annotation results and differences analysis

The results of the annotations were screened based on the following conditions: Loci on exon and splicing were retained; variant loci affecting amino acid code were retained. Fisher Test method was used to screen the loci with significant difference between HCC and normal participant groups. The screening criteria were existence of significant difference ($p < 0.05$).

Biomarker screening & construction and validation of risk prediction models

The susceptible genes between HCC patients and normal group were assigned to training set. The Genome data onto the HCC high risk participants were assigned to test sets. Five prediction models, which include compound covariant predictor-diagonal linear discriminant analysis, Nearest Neighbor predictor, Nearest Centroid Predictor and support vector machine predictor were adopted to screen biomarkers based on the categorization of training set into experiment group and control group (12-14). For stability and accuracy of the result, randomized sampling methods of leave one out cross validation were adopted. Then the training set was used to validate the prediction accuracy of the biomarkers. In order to validate the stability of the screened biomarkers and prevent the over-fitting problem of markers, the test set was used to validate the marker.

Results

The clinical characteristics comparison between included HCC patients and normal participants

The comparison results of clinical characteristics between the experimental group (HCC patients) and control group (normal participants) is shown in table 1. There is no significant difference between the two groups in regard to each clinical parameter (all p>0.05).

Quality assessment of peripheral blood DNA

After genomic DNA was extracted from all blood samples, the DNA concentration was determined with NanoDrop2000 spectrophotometer. DNA integrity was measured by agarose gel electrophoresis. DNA fluorescence band is single. Sample concentration is within 60-150 ng/ul. A260/280 ratio is between 1.8-2.0. These results met the requirements of subsequent exon sequence.

Annotation screening results, SNPs loci with difference and screened biomarkers

The annotations screening identified 4321 SNP loci. The difference analysis among these SNP loci showed that 270 SNP loci were significantly different from the normal group and HCC group (p < 0.05) (as shown in table 2). Five classification prediction models were used to screen specific markers according to the two groups of experimental groups (HCC patients) and con-

Table 1. Clinical characteristics of included HCC patients and normal participants

Items	HCC patients	Normal participants	P value
Age	45.7±10.2	42.9±11.8	0.103
Sex (Male/ female)	22/28	26/24	0.423
BMI (Kg/ m2)	23.2±3.5	22.8±2.9	0.268
Waist circumstance (cm)	92.7±8.8	89.9±9.8	0.068

Table 2. Top 10 SNPs with significant difference between the experimental group and the control group.

Position	Gene	P value
chr19:1037811-1037811: G-C	CNN2	<0.001
chr19:1037829-1037829: A-G	CNN2	<0.001
chr19:1037808-1037808: G-A	CNN2	<0.001
chr19:1037671-1037671: A-G	CNN2	<0.001
chr19:1037845-1037845: A-G	CNN2	<0.001
chr19:43356103-43356103: T-G	CD177	<0.001
chr19:1037802-1037802: A-C	CNN2	<0.001
chr7:152248119-152248119: G-A	KMT2C	<0.001
chr6:32666525-32666525: A-G	HLA-DQB1	<0.001
chr6:32666536-32666536:C-G	HLA-DQB1	<0.001

trol group (normal participants) in the training set, and 100 biomarkers were finally obtained (as shown in table 3). The 100 markers include ASCC3, CNN2, GOLGA, HLA, MTCH, MUC, NPIPB, OR2T, SLC, WDR89, and ZNF etc.

Validation of training set and testing set on prediction model

The validation results of training set on the prediction model are shown in table 4, 5. Nearest Neighbor Predictor and support vector machine

Table 3. Part of 100 biomarkers selected by prediction models

Position	Function	Gene
chr19:1037811-1037811: G-C	Exonic	CNN2
chr6:31356732- 31356732: T-A	Exonic	HLA-B
chr19:1037808-1037808: G-A	Exonic	CNN2
chr19:1037671-1037671: A-G	Exonic	CNN2
chr6:100516271-100516271: G-C	Exonic	ASCC3
chr7:100995922-100995922: A-C	Exonic	MUC12
chr1:248573866-248573866: G-A	Exonic	OR2T34
chr7:151086824-151086824: A-G	Exonic	AGAP3
chr22:42555052-42555052: G-A	Exonic	SERHL2
chr16:2114807-2114807:C-T	Exonic	PKD1

Table 4. Accuracy of prediction models for training set

Prediction models	Accuracy
Compound covariate predictor	92%
Diagonal Linear Discriminant Analysis (DLDA) predictor	92%
1-Nearest Neighbor Predictor	97%
3-Nearest Neighbor Predictor	100%
Nearest Centroid predictor	95%
Support vector machines predictor	100%

Table 5. Sensitivity and specificity of prediction models on training set

Prediction models	Class	Sensitivity	Specificity	PPV	NPV
Compound covariate predictor	Con	0.882	1	1	0.870
	Exp	1	0.882	0.870	1
Diagonal Linear Discriminant Analysis (DLDA) predictor	Con	0.824	1	1	0.870
	Exp	1	0.824	0.870	1
1-Nearest Neighbor Predictor	Con	0.941	1	1	0.952
	Exp	1	0.941	0.952	1
3-Nearest Neighbor Predictor	Con	1	1	1	1
	Exp	1	1	1	1
Nearest Centroid predictor	Con	0.882	1	1	0.909
	Exp	1	0.882	0.909	1
Support vector machines predictor	Con	1	1	1	1
	Exp	1	1	1	1

had the highest accuracy (100%). The validation result of the testing set on the prediction model is shown in table 6. Similarly, the accuracy of Nearest Neighbor Predictor and support vector machine was highest at 100%.

Discussion

HCC is a complex disease affected significantly by genetic factors. Comparative study on genome of HCC patients with that of normal population may provide a global understanding of the susceptibility gene spectrum of hepatocellular carcinoma. Due to the advantage of WES, we employed these methods to identify the susceptible SNPs for HCC and construct prediction model based on the identified SNPs loci.

After WES analysis and prediction model screening, 100 SNPs was obtained as biomarkers. Particularly, a group of HLA, HLA-DRB1, HLA-DQB1, HLA-B, HLA-A were identified. Gene encoding Human leukocyte antigen (HLA) is located on Chromosome 6p21. The antigen is mainly expressed on the surface of immune B lymphocytes, macrophages, dendritic cells, vascular endothelial cells, and epithelial cells to cooperate with immune cells to recognize exogenous antigens. An immune-genetics study demonstrated that the diseases associated with HLA were featured by immune-function abnormality, chronic process, recurrence, and inflammation (15). Furthermore, HLA was found to be associated with HCC. According to previous reports, HLA-B*5701 was associated with drug-induced liver impairment induced by flucloxacillin (16); polymorphisms of HLA-DQB1 could predict survival outcome in HCC patients receiving surgery (17); HLA-G and classical HLA class I expression were associated with liver metastases in colorectal cancer patients (18); HLA-DQA1&DQB1 variants were found to be associated with HCC (19). According to our results together with previous findings, some genes of the HLA family may be closely correlated with HCC.

Some SNPs in CNN2 were also identified. CNN2 encoded calponin-2 is capable of binding to actin and calmodulin, etc., and it may be involved in the regulation and modulation of smooth muscle contraction. Importantly, a recent study demonstrated that expression of calponin is upregulated in hepatocellular carcinoma tissues and silencing the CNN2 could inhibit hepatocarcinoma in vitro and in vivo (20). Our result confirmed that CNN2 may be involved in HCC occurrence and development process.

Furthermore, SNPs in ASCC3 encoding Activating signal cointegration 1 complex subunit 3 protein, was found to be associated with sustained HBV infection (21). This evidence showed the selected biomarkers were relevant with HCC.

In this study, we constructed prediction models which demonstrated good sensitivity and specificity for HCC prediction in HCC patients, normal participants, HCC high risk participants. This result suggested that the screened biomarkers are specific to HCC, and the developed prediction models could be used for early diagnosis and detection in HCC high risk population.

Table 6. Accuracy of prediction models for testing set

Prediction models	Accuracy
Compound covariate predictor	94%
Diagonal Linear Discriminant Analysis (DLDA) predictor	94%
1-Nearest Neighbor Predictor	97%
3-Nearest Neighbor Predictor	100%
Nearest Centroid predictor	97%
Support vector machines predictor	100%

Conclusion

A series of SNPs were identified as susceptible genes for HCC. Some of these SNPs including CNN2, CD177, KMT2C, HLA-DQB1 were consistent with the previously identified polymorphisms by targeted genes examination. The prediction models constructed with part of those SNPs could accurately predict HCC development.

Abbreviations

HCC - hepatocellular carcinoma
WES - Whole Exome Sequencing
SNPs - Single Nucleotide Polymorphisms
BGI - BGI-seq500 System
HLA - Human leukocyte antigen

Acknowledgement

This study was approved by Guangxi scientific Research and Technology Development Plan (Gui Ke Gong 1598013-5)

Authors' contribution

JZ (Conceptualization; Writing – original draft; Writing – review & editing)
YT (Formal analysis; Methodology)
EH (Formal analysis; Software)
GB (Conceptualization; Data curation)
ZL (Resources; Supervision)
PX (Methodology; Writing – original draft)
WT (Data curation)

Conflict of interest

The authors declare that there is not any conflict of interest.

Reference

1. European Association for The Study Of The Liver, European Organization For Research And Treatment Of Cancer. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol. 2012;56(4):908-43. DOI: 10.1016/j.jhep.2011.12.001
2. Poon RT, Fan ST. Hepatectomy for hepatocellular carcinoma: patient selection and postoperative outcome. Liver Transpl. 2004;10(2 Suppl 1):S39-45. DOI: 10.1002/lt.20040
3. Ruan SM, Gong SP, Lin SQ. Analysis of Malignant Tumor Incidence in Jinan City in 2012. Chinese Journal of Cancer prevention and treatment. 2014;14:1063-7.
4. Wu XQ, Mi HF, Rong B, Chen GW, Dai L, Chen ZL, et al. Analysis of Hepatic cancer death among residents in Xiamen city in 2002-2011. Chinese Journal of Disease Control &Prevention. 2014;07:613-6.
5. Zou YP, Ge S. A novel molecular marker-SNPs and its application. Biodiversity Science. 2003;11(5):370-82.
6. Schuster SC. Next generation sequencing transforms today's biology. Nat Methods. 2008;5(1):16-8. DOI: 10.1038/nmeth1156
7. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272-6. DOI: 10.1038/nature08250
8. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010;42(1): 30-5. DOI: 10.1038/ng.499
9. Summerer D, Schracke N, Wu H, Cheng Y, Bau S, Stahler CF, et al. Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform. Genomics. 2010;95(4):241-6. DOI: 10.1016/j.ygeno.2010.01.006
10. Glessner JT, Bick AG, Ito K, Homsy J, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circ Res. 2014;115(10):884-96. DOI: 10.1161/CIRCRESAHA.115.304458
11. Brastianos PK, Taylor-Weiner A, Manley PE, Jones RT, Dias-Santagata D, Thorner AR, et al. Exome sequencing identifies BRAF mutations in papillary craniopharyngiomas. Nat Genet. 2014;46(2):161-5. DOI: 10.1038/ng.2868
12. Radmacher MD, Mcshane LM, Simon R. A Paradigm for Class Prediction Using Gene Expression

Profiles. J Comput Biol, 2002;9(3):505-11. DOI: 10.1089/106652702760138592

13. Dudoit S, Fridlyand J, Speed TP. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. J Am Stat Assoc, 2002; 97(457):77-87. DOI: 10.1198/016214502753479248
14. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proc Natl Acad Sci U S A. 2003;100(17):9991-6. DOI: 10.1073/pnas.1732008100
15. Sun SG, Tong ET. Advances in etiology and pathogenesis of multiple sclerosis. J Clin Neurol. 1995;8:5.
16. Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, et al. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat Genet. 2009 Jul;41(7):816-9. DOI: 10.1038/ng.379
17. Liu X, Yu L, Han C, Lu S, Zhu G, Su H, et al. Polymorphisms otablef HLA-DQB1 predict survival of hepatitis B virus-related hepatocellular carcinoma patients receiving hepatic resection. Clin Res Hepatol Gastroenterol. 2016 Dec;40(6): 739-47. DOI: 10.1016/j.clinre.2016.04.005
18. Swets M, König MH, Zaalberg A, Dekker-Ensink NG, Gelderblom H, van de Velde CJ, et al. HLA-G and classical HLA class I expression in primary colorectal cancer and associated liver metastases. Hum Immunol. 2016 Sep;77(9): 773-9. DOI: 10.1016/j.humimm.2016.03.001
19. Karra VK, Chowdhury SJ, Ruttala R, Gumma PK, Polipalli SK, Chakravarti A, et al. HLA-DQA1 & DQB1 variants associated with hepatitis B virus-related chronic hepatitis, cirrhosis & hepatocellular carcinoma. Indian J Med Res. 2018 Jun;147 (6):573-80. DOI: 10.4103/ijmr.IJMR_1644_15
20. Kang X, Wang F, Lan X, Li X, Zheng S, Lv Z, et al. Lentivirus-mediated shRNA Targeting CNN2 Inhibits Hepatocarcinoma in Vitro and in Vivo. Int J Med Sci. 2018 Jan 1;15(1):69-76. DOI: 10.7150/ijms.21113
21. Liu L, Zhang J, Lu Y, Fang C, Li S, Lin J. Correlation between ASCC3 gene polymorphisms and Chronic Hepatitis B in a Chinese Han population. PLoS One. 2015 Nov 4;10(11):e0141861. DOI: 10.1371/journal.pone.0141861