European Journal of Cancer 93 (2018) 1–9

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Integrative expression quantitative trait locus–based analysis of colorectal cancer identified a functional polymorphism regulating *SLC22A5* expression

Danyi Zou, Jiao Lou, Juntao Ke, Shufang Mei, Jiaoyuan Li, Yajie Gong, Yang Yang, Ying Zhu, Jianbo Tian, Jiang Chang, Rong Zhong, Jing Gong **, Xiaoping Miao*

State Key Laboratory of Environment Health (Incubation), MOE (Ministry of Education) Key Laboratory of Environment & Health, Ministry of Environmental Protection Key Laboratory of Environment and Health (Wuhan), Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Received 13 September 2017; received in revised form 3 January 2018; accepted 6 January 2018

KEYWORDS
Colorectal cancer;
SLC22A5;
eQTL;
Genetic variants

Abstract Multiple single nucleotide polymorphisms (SNPs) have been found to be highly correlated with colorectal cancer (CRC) risk. However, the variants identified thus far only explain a small part of the cases, suggesting the existence of many uncharacterised genetic determinants. In this study, using the multilevel 'omics' data provided in The Cancer Genome Atlas, we systematically performed expression quantitative trait locus (eQTL) analysis for CRC and identified nine SNPs with significant effects on mRNA expression (correlation $|r| > 0.3$ and FDR < 0.01). Then we conducted a two-stage case–control study consisting of 1528 cases and 1528 controls to examine the associations between candidate SNPs and CRC risk. We found that rs27437 in *SLC22A5* was significantly correlated with CRC risk in both stages and the combined study (additive model, OR = 1.31, 95%CI = 1.17–1.47, $P = 1.97 \times 10^{-6}$). eQTL analysis showed that rs27437 GG and GA genotypes were associated with lower expression of *SLC22A5* compared with the AA genotype. Dual-luciferase reporter assays confirmed that the *G* risk allele could decrease the expression of luciferase. *SLC22A5* was significantly decreased in CRC tumour tissues compared with adjacent normal tissues,

* *Corresponding author*: Fax: +86 27 8369 3673.
** *Corresponding author*: Fax: +86 27 8369 3673.
E-mail addresses: gongj@hust.edu.cn (J. Gong), miaoxp@mail.hust.edu.cn (X. Miao).

https://doi.org/10.1016/j.ejca.2018.01.065

indicating that *SLC22A5* may play important roles in CRC, and rs27437 in *SLC22A5* might serve as a novel biomarker for early detection and prevention of CRC.

1. Introduction

Colorectal cancer (CRC) is the third most common cancer in males and the second most common cancer in females [1]. There has been a trend of increasing incidence in Asia including China in the past few decades [2]. Since CRC has become an enormous burden to society as well as individual family, unravelling the pathogenesis of CRC and identifying novel disease biomarkers is of great importance to help in CRC prevention and early detection.

It is well established that factors of environmental and genetic variants can lead to CRC risk. Of the genetic variants, single nucleotide polymorphisms (SNPs) are considered to play a critical role in the progression of the disease [3]. Genome-wide association studies (GWASs) have identified SNPs located in more than 30 loci associated with CRC susceptibility [4–7]; however, the majority of the SNPs associated with CRC reside in non-coding gene regions [8], indicating that risk SNPs more likely exert their roles by regulating gene expression. Expression quantitative trait loci (eQTL) are reported to have key roles in tumourigenesis by regulating the expression of their target genes [9–11]. Recently, a few of eQTL studies have been performed in colonic mucosa and significant loci have been found to correlate with CRC [12–14]. However, the studies that systematically analyse eQTLs and validate their roles in population at the same time are limited. In addition, although variants may affect gene expression, somatic copy number variants (CNVs) and DNA methylation are also widely involved in gene expression [15]. Therefore, performing eQTL analysis and adjusting the effect of CNVs and DNA methylation is of great importance [16,17].

In this study, we systematically explored eQTLs in CRC adjusted for both somatic CNVs and DNA methylation using 'omics' data from The Cancer Genome Atlas (TCGA) [18] and validated the most likely functional SNPs in a two-stage case-control study in a Chinese population. We found that rs27437 in *SLC22A5* was significantly correlated with CRC risk. The effect of the SNP on gene expression was further confirmed by luciferase reporter assays.

2. Materials and methods

2.1. Data sets and preprocessing

The germline genotype, somatic copy number and CpG methylation as well as mRNA expression of colon adenocarcinoma and rectum adenocarcinoma were downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). We chose the data set of the germline genotypes and the somatic copy number from blood-derived DNA samples inferred from Affymetrix Genome-Wide SNP 6.0 Array, whereas the data set of mRNA expression and methylation were obtained and inferred from matched tumour samples from HumanMethylation 450K array and Hiseq_RNAV2 by Illumina, respectively. Samples with all four kinds of data available were studied further. In each sample, for each transcript, we calculated the average of the segmented copy number scores of the genetic interval between the transcription start and end sites as the gene-based somatic copy number value. The values of DNA methylation and the mRNA expression were directly obtained from TCGA level 3 data. Transcript abundances were log-transformed, and HLA loci were excluded because those genes might give a high level of polymorphism.

2.2. Verification of ancestry

To minimise the influence of the population difference, we first conducted the verification of ancestry. We downloaded 210 unrelated individual genotypes from the HapMap as a reference panel (http://www.affymetrix.com/estore/). We combined the profiles of SNP data provided by both TCGA and HapMap and calculated the top two genotype principal components across the samples using SNPs with a minor allele frequency (MAF) above 0.05 using the smartpca package implemented in the EIGENSOFT6.0 software [19].

2.3. Association analysis

For each germline SNP, the correlations between the SNP and nearby genes (±1 Mb) were evaluated using a two-stage multivariate linear model. The first multivariate linear regression was performed to remove the influences of the somatic CNVs and DNA methylation on transcript abundance. The second linear regression was performed to evaluate the correlation between residual gene expression and the germline genotypes. The raw *P*-values were corrected using the Benjamini-Hochberg method, and a threshold of false discovery rate (FDR) of 0.1 was set to define significant eQTLs. The details of the algorithms were reported by Li *et al.* [20].