Acta Tropica 161 (2016) 79–85

Contents lists available at ScienceDirect

Acta Tropica

journal homepage: www.elsevier.com/locate/actatropica

Genetic variation between *Schistosoma japonicum* lineages from lake and mountainous regions in China revealed by resequencing whole genomes

Mingbo Yin[a,b,1], Xiao Liu[b,1], Bin Xu[a], Jian Huang[c], Qi Zheng[b], Zhong Yang[b], Zheng Feng[a], Ze-Guang Han[c,d,**], Wei Hu[a,b,*]

[a] *Key Laboratory of Parasite and Vector Biology Ministry of Public Health, National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, 207 Rui Jin Er Road, Shanghai 200025, China*
[b] *School of Life Sciences, Fudan University, 2005 Song Hu Road, Shanghai 200438, China*
[c] *Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China*
[d] *Shanghai-MOST Key Laboratory for Disease and Health Genomics, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China*

ARTICLE INFO

Article history:
Received 28 December 2015
Received in revised form 25 April 2016
Accepted 16 May 2016
Available online 17 May 2016

Keywords:
Schistosoma japonicum
Whole genome
Resequencing
SNPs
Variation
Divergence
Adaptation

ABSTRACT

Schistosoma infection is a major cause of morbidity and mortality worldwide. Schistosomiasis japonica is endemic in mainland China along the Yangtze River, typically distributed in two geographical categories of lake and mountainous regions. Study on schistosome genetic diversity is of interest in respect of understanding parasite biology and transmission, and formulating control strategy. Certain genetic variations may be associated with adaptations to different ecological habitats. The aim of this study is to gain insight into *Schistosoma japonicum* genetic variation, evolutionary origin and associated causes of different geographic lineages through examining homozygous Single Nucleotide Polymorphisms (SNPs) based on resequenced genome data. We collected *S. japonicum* samples from four sites, three in the lake regions (LR) of mid-east (Guichi and Tonglin in Anhui province, Laogang in Hunan province) and one in mountainous region (MR) (Xichang in Sichuan province) of south-west of China, resequenced their genomes using Next Generation Sequencing (NGS) technology, and made use of the available database of *S. japonicum* draft genomic sequence as a reference in genome mapping. A total of 14,575 SNPs from 2059 genes were identified in the four lineages. Phylogenetic analysis confirmed significant genetic variation exhibited between the different geographical lineages, and further revealed that the MR Xichang lineage is phylogenetically closer to LR Guich lineage than to other two LR lineages, and the MR lineage might be evolved from LR lineages. More than two thirds of detected SNPs were nonsynonymous; functional annotation of the SNP-containing genes showed that they are involved mainly in biological processes such as signaling and response to stimuli. Notably, unique nonsynonymous SNP variations were detected in 66 genes of MR lineage, inferring possible genetic adaption to mountainous ecological condition.

1. Introduction

Schistosomiasis is one of the most prevalent neglected diseases in tropical and subtropical regions of the world (Rollinson et al., 2013). There are three principal species infecting humans: *Schistosoma mansoni, Schistosoma japonicum* and *Schistosoma haematobium*. *S. japonicum* is distributed mainly in China, Indonesia and the Philippines. In mainland China, it is typically present in two categories of ecological habitat: the lake regions (LR) of Hubei, Hunan, Anhui, Jiangxi and Jiangsu provinces, and the mountainous region (MR) of Sichuan and Yunnan provinces along the Yangtze River (Zhao et al., 2012). Tremendous control efforts have been made for schistosomiasis in recent decades in China, achieving a 90% reduction in infection rate (Zhou et al., 2007). However, most recent studies suggest that schistosomiasis has re-emerged over the last decade, and might once again pose a severe threat to human

* Corresponding author at: School of Life Sciences, Fudan University, 2005 Song Hu Road, Shanghai 200438, China.
** Corresponding author at: Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai, 200240, China.
E-mail addresses: hanzg@sjtu.edu.cn (Z.-G. Han), huw@fudan.edu.cn (W. Hu).
[1] Contributed equally to this work.

http://dx.doi.org/10.1016/j.actatropica.2016.05.008
0001-706X/

health in Asia (Zhou et al., 2010). To understand genetic variation of the parasites is of importance in expanding the knowledge of its transmission and epidemiological features.

A variety of molecular markers have been used to investigate the genetic variability of *S. japonicum*. It was reported, for example, that random amplified polymorphic DNA (RAPD) (Gasser et al., 1996) and mitochondrial NADH dehydrogenase I gene subunit (Sorensen et al., 1998) have been used to study population variability, indicative of significant genetic differences between the populations from different geographic sites in China. A report suggested a set of microsatellite markers used to study the genetic variation of *S. japonicum* populations from eight locations covering lake and mountainous endemic regions in China (Shrivastava et al., 2005), and found high level of polymorphism within and among populations. Derived from the findings, it was argued that there may be different strains of the parasite circulating in mainland China (Shrivastava et al., 2005). Most recently, by using partial mitochondrial DNA markers, geographical separation/isolation was reported as the major factor accounting for population genetic variation of *S. japonicum* between lake and mountainous regions (Zhao et al., 2012).

Furthermore, distinct genetic divergence was detected between populations of *Oncomelania hupensis* (the sole intermediate host of *S. japonicum*) from mountainous and lake regions (Zhao et al., 2010), and the snails from different regions differed in their susceptibility to the same strain of *S. japonicum* (Cross et al., 1984). Given the variances in parasite and intermediate host snail, it might be expected that resequencing the genomes of geographical lineages of *S. japonicum* may reveal genome-wide information to further understand the parasite genetic variation, evolutionary origin of geographical clusters and environmental causes of the genetic verification. In the present study, we sampled *S. japonicum* from four locations in LR and MR of China, and resequenced their whole genomes using next generation sequencing (NGS) technology to investigate the variation of homozygous SNPs between different geographic lineages, and gain insight into the genetic variation and possible impact of ecological conditions in evolutionary adaptation.

2. Materials and methods

2.1. Ethics statement

The procedure involving laboratory rabbits was carried out based on the guidelines of the Care and Use of Laboratory Animals of the Ministry of Science and Technology of People's Republic of China ([2006]398). The animal study protocol complied with the institutional ethical guidelines, and was reviewed and approved by the ethics committee at the National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention (Permit No: IPD2008-4).

2.2. Sampling and DNA extractions

Snails (*O. hupensis*) naturally infected with *S. japonicum* were collected from four field sites in the endemic areas of Guichi and Tonglin (Anhui province), Laogang (Hunan province), and Xichang (Sichuan province). The geographic information of sample collection sites is seen in Fig. 1. These snails were collected in 2003 as part of field surveys for our previous study on genetic structure of *S. japonicum* populations in China (Yin et al., 2008). The collected snails were used for releasing cercariae to infect rabbits, from which adult worms were harvested (Yin et al., 2008). The four sampling sites are categorized into two main habitat types: low-lying lake region (LR) and mountainous region (MR) (Jiang et al., 2002; Zhao et al., 2012). Guichi, Laogang and Tonglin are characterized as LR, and Xichang as MR. We randomly selected 10 adult worms (paired gender) from each site, and then extracted genomic DNA from pooled 10 worms/site respectively by using the DNeasy Blood & Tissue Kit (Qiagen, Germany).

2.3. NGS genome resequencing and quality control

In this study, we developed a pipeline to identify effective SNPs, with an average coverage of 22.94. An overview of the pipeline is shown in Fig. S1. The genomes were resequenced by pair-end sequencing using an Illumina Genome Analyzer II in the National Engineering Center for Biochip at Shanghai, China. FastQC (Andrews, 2011) was used for quality control of raw data. Adapters (10 bases from the 5′ ends) and low quality reads were removed. Low quality reads were defined using the following criteria: 1) reads with more than 10% Ns, or 2) more than 50% of bases with a read quality score of less than 5.

2.4. De novo assembly

De novo assembly of genomes was carried out in both SOAP (Luo et al., 2012) and *Velvet* (Zerbino and Birney, 2008). Both methods were based on a *de Bruijn* graph (Ye et al., 2012), which are currently the most popular tools for genome assembly. In this study, SOAP and *Velvet* assemblers generated very short N50, indicating a low overall coverage of the query data. The outputs from *Velvet* were used for all further analyses as it achieved better performance (length of N50) than SOAP.

2.5. Reference genome mapping

The *S. japonicum* scaffold (version 2) was obtained from the Chinese National Human Genome Center at Shanghai, and used as the reference genome (generated from Anhui Guichi isolate). Resequenced genomes were mapped using BWA (Li and Durbin, 2009) with a maximum edit distance of 5 and 2% uniform base error rate.

2.6. SNP calling

Overlapping loci of all samples from the genome alignment was extracted for SNP analyses. An overview of data generation, processing and analyses is shown in Fig. S1. A multiple sequence alignment was constructed using ClustalW (Tamura et al., 2007) for each 100% overlapped locus. SNPs were then called for each ecological site linage, respectively.

2.7. Phylogenetic analysis

Phylogenetic analysis is the most commonly used tool for estimating the evolutionary relationships among organisms or lineages based on observed evolutionary characters, such as DNA or protein sequences (Brocchieri, 2001). In the current study, the character sequences for inferring phylogeny were provided by SNPs and mutated amino acids in the collected population samples. By grouping the SNPs as synonymous and non-synonymous type and according to codon mapping, mutated amino acids were translated and used to infer functional evolution. A polygenetic tree was constructed using FastTree (Price et al., 2009) with a maximum likelihood estimator. GTR (Tavaré, 1986) and WAG models (Whelan and Goldman, 2001) were used for nucleotide and amino acid sequences, respectively. The Forester package version 1.030 (Zmasek and Eddy, 2001) was applied for tree visualization with the mid-point as root.