Methods 129 (2017) 24–32

Contents lists available at ScienceDirect

Methods

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

Detecting exact breakpoints of deletions with diversity in hepatitis B viral genomic DNA from next-generation sequencing data

Ji-Hong Cheng [b], Wen-Chun Liu [c,d], Ting-Tsung Chang [c,d], Sun-Yuan Hsieh [b], Vincent S. Tseng [a,*]

[a] *Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan*
[b] *Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan*
[c] *Department of Internal Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan 701, Taiwan*
[d] *Infectious Disease and Signaling Research Center, National Cheng Kung University, Tainan, Taiwan*

ARTICLE INFO

Article history:
Received 16 February 2017
Received in revised form 8 June 2017
Accepted 7 August 2017
Available online 10 August 2017

Keywords:
Data mining
Machine learning
Breakpoint
Deletion detection
Hepatitis B virus
Next-generation sequencing

ABSTRACT

Many studies have suggested that deletions of Hepatitis B Viral (HBV) are associated with the development of progressive liver diseases, even ultimately resulting in hepatocellular carcinoma (HCC). Among the methods for detecting deletions from next-generation sequencing (NGS) data, few methods considered the characteristics of virus, such as high evolution rates and high divergence among the different HBV genomes. Sequencing high divergence HBV genome sequences using the NGS technology outputs millions of reads. Thus, detecting exact breakpoints of deletions from these big and complex data incurs very high computational cost. We proposed a novel analytical method named VirDelect (Virus Deletion Detect), which uses split read alignment base to detect exact breakpoint and diversity variable to consider high divergence in single-end reads data, such that the computational cost can be reduced without losing accuracy. We use four simulated reads datasets and two real pair-end reads datasets of HBV genome sequence to verify VirDelect accuracy by score functions. The experimental results show that VirDelect outperforms the state-of-the-art method Pindel in terms of accuracy score for all simulated datasets and VirDelect had only two base errors even in real datasets. VirDelect is also shown to deliver high accuracy in analyzing the single-end read data as well as pair-end data. VirDelect can serve as an effective and efficient bioinformatics tool for physiologists with high accuracy and efficient performance and applicable to further analysis with characteristics similar to HBV on genome length and high divergence. The software program of VirDelect can be downloaded at https://sourceforge.net/projects/virdelect/.

1. Introduction

Patients with hepatitis B virus (HBV) infection are at risk of developing liver cirrhosis and hepatocellular carcinoma (HCC) [1–3]. Human HBV is the prototype member of the family Hepadnaviridae, which includes a variety of avian and mammalian viruses that share similar genomic organization, organtrophisms, and a unique strategy of genome replication [4]. The HBV genome comprises a partially double stranded 3.2 kb DNA organized into four open-reading frames (ORFs) (Fig. 1) and multiple regulatory elements [5]. Because HBV reverse transcriptase lacks proofreading activity, the composition of the viral quasispecies evolves over time depending on the selective pressure, including the host immune response. The characteristics of viral quasispecies have been implicated in the exacerbation of chronic hepatitis B (CHB) and the development of liver cancer [6]. Many studies have suggested that HBV pre-S deletions are associated with the development of progressive liver diseases. Three different yet structurally related HBV viral surface proteins are translate from a single open reading frame, as follows: large (L), middle (M), and small (S) proteins (Fig. 1). The S protein consists of 226 amino acids (aa). The M protein is an extension of the S protein, with an additional 55 aa (i.e., pre-S2 region). The L protein is an extension of the M protein, with an additional 108–119 aa depending on the genotype. Some in vitro studies have shown that pre-S deletion mutants can cause the accumulation of L surface proteins in the endoplasmic reticulum (ER), resulting in ER stress. Other related studies have suggested that ER stress results in the generation of large amounts of reactive oxygen species, which can cause oxidative DNA damage, inducing mutagenesis in the genome and ultimately resulting in HCC.

[*] Corresponding author.
E-mail address: vtseng@cs.nctu.edu.tw (V.S. Tseng).

http://dx.doi.org/10.1016/j.ymeth.2017.08.005
1046-2023/© 2017 Published by Elsevier Inc.

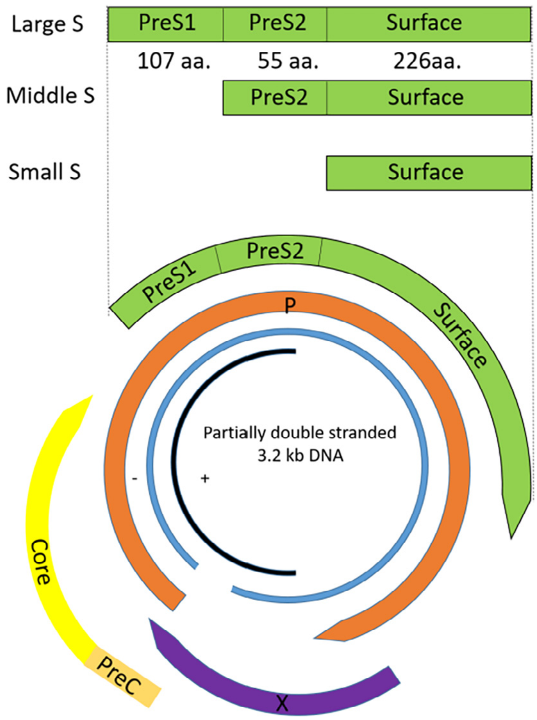

Fig. 1. Schematic representation of the hepatitis B virus genome.

In the sequence analysis of HBV, it is necessary to consider the characteristics of HBV such as a higher evolution rate and high divergence among the different HBV genomes. Hepadnaviridae show higher evolution rates than other DNA viridae and human genomes but lower rates than RNA viridae. The evolution rates of HBV have been calculated to be in the ranges of 1–5 × 10^{-5} and 5–8 × 10^{-5} per site per year in immuno-competent patients [7], 6–18 × 10^{-3} in liver transplant recipients with fulminant reinfection [8], and 0.3–1.3 × 10^{-3} for the HBV core gene in a case of perinatally acquired chronic HBV infection [9]. More than 1011 HBV virions can be produced daily in a single patient [10]. Because of this heterogeneity, there is a genetic classification system for HBV, in which different HBV genotypes, subgenotypes, and subtypes (serotypes) have been defined. HBV was classified into ten genotypes, HBV/A–J, with an inter-genotypic diversity of at least 8% in the complete genome sequence [11], with genotypes B and C occurring mainly in Asia and genotypes A and D occurring mainly in Africa and Europe [12]. Multiple subgenotypes have been recognized within genotypes (almost 40 subgenotypes), with more than 4% and less than 7.5% divergence over the full-length genome [13]. HBV "clades" are used as a subdivision within subgenotypes, presenting less than 4% nucleotide diversity over the complete genome sequences [14]. Based on the high divergence among different HBV genomes, a general reference sequence is not applicable.

Among the methods for analyzing next-generation sequencing (NGS) data, few methods for detecting breakpoints of deletions or other structural variations (SVs) are currently designed for sequences with high evolution rates and high divergence biometrics. However, the biological sequence data for the genomic sequence of HBV have such characteristics. In addition, it has been challenging to develop a method to find the breakpoints of long deletions in high-variant sequences in pair-end reads and single-end reads. NGS is currently a very popular and important sequencing technology, and it has the advantages of high throughput and high speed [15]. Sequencing HBV genome sequences using NGS technology typically outputs millions of reads. These are large and complex data, with the characteristics of big data. Sequence analysis of these data often requires high computational costs, especially when analyzing structural variations in the genome.

Scholars have proposed several different methods for the sequence analysis of structural variations (SVs), such as BreakDancer [16], GASV [17], PEMer [18], and Breakpointer [19], to detect breakpoints, but these methods can only provide the approximate positions of breakpoints. The GASVPro method for detecting exact breakpoints [20] uses the probabilistic model, and PRISM [21], SVseq2 [22], and Pindel [23] use the split read alignment method [25]. These methods use anchored reads mapping or a linking signature [24] to reduce the computational cost. However, sequencing technologies such as Roche 454 that generate single reads data are not suitable for use with these two methods to reduce the cost of computing. The split read alignment method takes only one mate read of a pair of reads from a mapped reads file (such as a SAM file), splits the read into a prefix and suffix [24], and then uses them to align a reference. Anchored reads mapping [24] uses the location of only one mate read of a pair of reads to estimate the SVs of the region and then takes the other read of this pair to perform a split read alignment. Thus, the search space required for alignment can be reduced. A linking signature involves the analysis of concordant mate pairs with several features, and these features of the discordant pairs are used to estimate the location of SVs. In addition, there are other useful analysis tools and methods for detecting SVs in NGS data, but most of these tools are focused on the human genome [25], and there are few deletion analysis tools designed for virus characteristics.

In this study, we used the split read alignment method to obtain the exact breakpoints of deletions. We developed an effective method, named VirDelect (Virus Deletion Detect), to reduce the computational cost of split read alignment without reducing accuracy. We designed a diversity formula for HBV in split read alignment and used this formula to consider the characteristics of HBV with higher evolution rates and high divergence. VirDelect is currently used to analyze single-end data, but with pair-end data, it can also detect the breakpoints of deletions using one end of the pair. We used VirDelect to detect the breakpoints of deletions in simulated data with different deletion lengths and different virus diversities and analyzed two real pair-end reads datasets. The experimental results indicated that VirDelect can identify more exact breakpoints of deletions than Pindel [23].

2. Materials and methods

2.1. Hepatitis B viral whole genome cloning

The HBV full genomes of Clone_N66 and Clone_H44 (KJ790200) were extracted from chronic hepatitis B patients and sequenced using a direct Sanger sequencer (Applied Biosystems, Life Technologies, Taipei, Taiwan). Clone_N66 and Clone_H44 (KJ790200) [26] were genotype B HBV strains. These full-length HBV genome sequences of AB602818 (genotype B, Asia) and KJ790200 (Clone_H44; genotype B, Taiwan) from the GenBank database and Clone_N66 (Taiwan) were used as the reference. The HBV genome has a circular form, and traditionally, most researchers number the HBV genome from the middle point of the EcoRI restriction site, namely the first nucleotide of the full HBV genome. However, the known deletion fragments are usually located in the HBV preS region (nt2848-nt154). In order to keep this region in the middle of the reference sequence, we set nt1600 as the starting position and nt1599 as the final position.