BS-virus-finder

Gao, Shengjie; Hu, Xuesong; Xu, Fengping; Gao, Changduo; Xiong, Kai; Zhao, Xiao; Chen, Haixiao; Zhao, Shancen; Wang, Mengyao; Fu, Dongke; Zhao, Xiaohui; Bai, Jie; Mao, Likai; Li, Bo; Wu, Song; Wang, Jian; Li, Shengbin; Yang, Huangming; Bolund, Lars; Pedersen, Christian N. S.

Published in:
GigaScience

DOI:
10.1093/gigascience/gix123

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
TECHNICAL NOTE

BS-virus-finder: virus integration calling using bisulfite sequencing data

Shengjie Gao1,2,3,7,8,9,†, Xuesong Hu2,3,†, Fengping Xu3,10,13,†, Changduo Gao4, Kai Xiong5, Xiao Zhao2,11, Haixiao Chen3,13, Shancen Zhao3,7, Mengyao Wang3, Dongke Fu2, Xiaohui Zhao6, Jie Bai3, Likai Mao3, Bo Li2,3, Song Wu8, Jian Wang3, Shengbin Li2,12,14, Huangming Yang3,7,11, Lars Bolund9,∗ and Christian N. S. Pedersen1,∗

1Bioinformatics Research Center, Aarhus University, C. F. Møllers Allé 8, DK-8000, Aarhus C, Denmark, 2Forensics Genomics International (FGI), BGI-Shenzhen, BeiShan Industrial Zone, Yantian District, Shenzhen, Guangdong 518083, China, 3BGI-Shenzhen, BeiShan Industrial Zone, Yantian District, Shenzhen, Guangdong 518083, China, 4College of Computer Science and Technology, Qingdao University, Qingdao 266071, China, 5Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Grønnegårdvej 15, DK-1870 Frederiksberg C, Denmark, 6College of Mathematics & Statistics, Changsha University of Science and Technology, Changsha 410114, China, 7James D. Watson Institute of Genome Sciences, Hangzhou 310058, China, 8The Affiliated Luohu Hospital of Shenzhen University, Shenzhen University, Shenzhen 518000, China, 9Department of Biomedicine, Aarhus University, Vennelyst Boulevard 4, DK-8000 Aarhus C, Denmark, 10Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark, 11BGI Education Center, University of Chinese Academy of Sciences, Beijing 100049, China, 12Shenzhen Key Laboratory of Forensics, BGI-Shenzhen, Shenzhen 518083, China, 13China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China and 14College of Medicine and Forensics, Xi’an Jiaotong University, Xi’an 710049, China

∗Correspondence address. Lars Bolund, Department of Biomedicine, Aarhus University, Bartholins Allé 6, 8000 Aarhus C, Denmark. Tel +4587167771, E-mail: bolund@biomed.au.dk; Christian N.S. Pedersen, Bioinformatics Research Centre, Aarhus University, C. F. Møllers Allé 8, 8000 Aarhus C, Denmark. Tel +4589155559, E-mail: cstorm@birc.au.dk

†Equal contribution

Abstract

Background: DNA methylation plays a key role in the regulation of gene expression and carcinogenesis. Bisulfite sequencing studies mainly focus on calling single nucleotide polymorphism, different methylation region, and find allele-specific DNA methylation. Until now, only a few software tools have focused on virus integration using bisulfite sequencing data. Findings: We have developed a new and easy-to-use software tool, named BS-virus-finder (BSVF,
Introduction

DNA methylation plays a crucial role in many areas including development [1, 2] and X chromosome inactivation [3] by regulating genetic imprinting and epigenetic modification without altering DNA sequences. Previous studies have shown a strong association of DNA methylation with cancer. The methylation status altering related to carcinogenesis [4], cancer recurrence [5], and metastasis [6] has already been revealed by emerging bisulfite sequencing (BS) technology. BS technology can investigate DNA methylation changes with single-base accuracy. Treatment of DNA with bisulfite converts unmethylated cytosine residues to uracil, but leaves 5-methylcytosine residues unmodified [7]. Thus, bisulfite treatment introduces specific changes in the DNA sequence that depend on the methylation status of individual cytosine residues, yielding single-nucleotide resolution information about the methylation status of a segment of DNA (Fig. 1). Various analyses can be performed on the altered sequences to retrieve this information. BS technology can reveal differences between cytosines and thymidines and sequence change resulting from bisulfite conversion. For the bases without methylation, all Cs will change to Ts on both strands. After directional library preparation, we have 2 different conversions: the Watson and the Crick strands, as shown in Fig. 1. On the Watson strand, methylated C remains C, and unmethylated C changes to T. On the Crick strand, the reverse complement happens; i.e., methylated C remains C, but in sequenced reads it is reverse-complemented to G, and unmethylated C changes to T, leading to the reverse-complement base A in sequenced reads. As base C can either be methylated or unmethylated, we can use International Union of Pure and Applied Chemistry (IUPAC) nucleotide codes “Y” and “R” to represent C/T and G/A, respectively. So, after bisulfite treatment, C changes to Y on the Watson strand, and base G changes to R on the Crick strand.

Whole-genome-based bisulfite sequencing (WGBS) has been developed to detect DNA methylation. Recent clinical studies showed that DNA methylation is associated with viral integration [8, 9]. Whole-genome BS data can be analyzed to investigate the sequence alignment and QDMR [13], DMAP [14], and SMAP [15], to identify single nucleotide polymorphisms (SNPs) via BS-SNP [16] and Bis-SNP [17], and to find allele-specific DNA methylation via SMAP [15] and Methy-Pipe [18]. However, none of them can be used for virus integration loci calling, and no software tool is currently available to detect virus integration loci by analyzing BS data. Therefore, we have developed a software tool to detect the virus integration loci by genome-wide BS analysis.

**Figure 1**: The illustration of bisulfite-altered sequence to the original.
Description of in silico and real data

Different types of paired-end (PE) reads (50 base pairs [bp], 90 bp, 150 bp) that include 700 breakpoints in chromosome 1 (chr 1) of GRCh38 were simulated in our study. Input fragments of 50 to 400 bp were randomly selected from chr 1 in the GRCh38 assembly of the human genome. The hepatitis B virus (HBV) genome (GenBank: X04615.1) was used in our simulation. Its integration length was between 45 bp and 180 bp. We cut HBV-containing segments with given PE insert size at all possible positions on every integration event. After alignment, mapping accuracy of each of the 17 different types of read mappings was calculated (Fig. 2). Mapping accuracy varied among the 17 types of read mappings in our simulation (Figs S1, S2, S3). In summary, the accuracies of several kinds of the read mappings were low (Tables S1, S2, S3), which may raise the false-negative rate. Generally, however, bwa-meth [12] performed very well.

Bisulfite sequencing is a sophisticated technique to study DNA cytosine methylation. Bisulfite treatment followed by polymerase chain reaction (PCR) amplification specifically converts unmethylated cytosine to thymine. By cooperating with next-generation sequencing technology, it is able to detect the methylation status of every cytosine in the whole genome. Moreover, longer reads make it possible to achieve higher accuracy. Besides simulated data, the PLC/PRF/5 hepatocellular carcinoma cell lines (from American Type Culture Collection [ATCC], Manassas, VA, USA) were cultured as previously described [19]. The cell line was validated by STR makers (Fig. S4). We performed whole-genome sequencing (WGS) and WGBS sequencing of this cell line (the results are shown in Table S4). Table 1 shows the analysis result for WGS data, which were compared with the output results analyzed by Vy-per [20], virus-clip [21], and Virus Finder2 [22].

Methods
Sample preparation

PLC/PRF/5 hepatocellular carcinoma cell line was obtained from ATCC (Manassas, VA, USA) and was cultured as previously described [19] and validated by STR makers (Fig. S4). In total, 15 μg of DNA was extracted to perform WGS and WGBS sequencing. Sample concentration was detected by fluorometer (Qubit-Fluorometer, Invitrogen). Sample integrity and purification was determined by Agarose Gel Electrophoresis.

Whole-genome sequencing

About 1.5 μg of gDNA was sonicated to 100–300-bp fragment genome DNA by Sonication (Covaris) and purified with QIAquick PCR Purification Kit (Qiagen). Adapter ligation and target insert size fragment recovering and quantifying library by real-time
Table 1: The comparison of BS-virus-finder with other software using real data

<table>
<thead>
<tr>
<th>Chr</th>
<th>BSVF</th>
<th>Vy-per</th>
<th>Virus-clip</th>
<th>Virus Finder2</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>143 272 758</td>
<td>2945</td>
<td>3102</td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>-</td>
<td>-</td>
<td>52 018 758</td>
<td>207 281</td>
</tr>
<tr>
<td>chr3*</td>
<td>131 451 702</td>
<td>1212</td>
<td>1322</td>
<td>52 018 758</td>
</tr>
<tr>
<td>chr3*</td>
<td>131 453 124</td>
<td>1416</td>
<td>1515</td>
<td>-</td>
</tr>
<tr>
<td>chr4*</td>
<td>180 586 417</td>
<td>136</td>
<td>378</td>
<td>-</td>
</tr>
<tr>
<td>chr4*</td>
<td>180 587 608</td>
<td>394</td>
<td>594</td>
<td>180 586 607</td>
</tr>
<tr>
<td>chr5*</td>
<td>1 297 478</td>
<td>1174</td>
<td>1315</td>
<td>-</td>
</tr>
<tr>
<td>chr7</td>
<td>110 894 616</td>
<td>2739</td>
<td>2748</td>
<td>-</td>
</tr>
<tr>
<td>chr8*</td>
<td>35 446 380</td>
<td>2389</td>
<td>2459</td>
<td>35 446 214</td>
</tr>
<tr>
<td>chr8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>chr11*</td>
<td>52 018 943</td>
<td>2631</td>
<td>2767</td>
<td>-</td>
</tr>
<tr>
<td>chr12*</td>
<td>109 573 899</td>
<td>721</td>
<td>815</td>
<td>109 573 677</td>
</tr>
<tr>
<td>chr13</td>
<td>33 088 123</td>
<td>1521</td>
<td>1603</td>
<td>-</td>
</tr>
<tr>
<td>chr13*</td>
<td>33 088 561</td>
<td>1917</td>
<td>2066</td>
<td>-</td>
</tr>
<tr>
<td>chr16*</td>
<td>69 947 046</td>
<td>2055</td>
<td>2826</td>
<td>-</td>
</tr>
<tr>
<td>chr16*</td>
<td>70 169 959</td>
<td>2055</td>
<td>2735</td>
<td>-</td>
</tr>
<tr>
<td>chr16</td>
<td>74 425 602</td>
<td>2062</td>
<td>2665</td>
<td>-</td>
</tr>
<tr>
<td>chr17*</td>
<td>82 106 786</td>
<td>407</td>
<td>489</td>
<td>82 106 984</td>
</tr>
<tr>
<td>chr17*</td>
<td>82 107 626</td>
<td>2177</td>
<td>2321</td>
<td>-</td>
</tr>
<tr>
<td>chr19</td>
<td>41 783 064</td>
<td>687</td>
<td>804</td>
<td>-</td>
</tr>
<tr>
<td>chr20</td>
<td>20 473 566</td>
<td>2415</td>
<td>2565</td>
<td>-</td>
</tr>
</tbody>
</table>

BSVF used WGBS data, and other software used WGS data.
*Supported by previous fluorescence in situ hybridization (FISH) experiments [8].
Abbreviations: HB: host breakpoint; VB: virus begin is the revealed leftmost position on virus; VE: virus end is the rightmost position on virus.

Data analysis

The read coverage situation for 1 integration is shown in Fig. 3. Four steps were implemented to detect virus integration:

1. Alignment: We use bwa-meth [12] to align bisulfite-treated sequencing reads to a hybrid reference that contains both human genome and virus sequences. For chimeric reads from the junction parts, BWA-MEM [23] will align it to 1 organism and mark the unmapped part as soft clipping, which is in fact from the other organism. This enables us to find breakpoints directly from the alignment.

2. Clustering: After alignment, the result was filtered. We select read pairs with 1 read match by the following criterion: The Phred-scaled mapping quality is larger than 30 (≥ 30), and at least 1 soft clipping is longer than 5 bp (≥ 5). The mapped parts of reads, which is marked as “M” by its CIGAR string, cover the human reference genome. For paired reads, we also add the gap between 2 mapped reads to their covered region, making read 1 and read 2 continuously covered on the human reference. Each continuous region with at least 1 bp of overlap is defined as a cluster. All reads involved are selected to form the cluster. The remaining soft clippings are viral junction candidates. Read pairs with 1 read mapped on the virus also indicate a potential virus junction between the read pairs.

3. Assembling: Within 1 cluster, all soft clipping start sites are collected. The position with the most abundance of start sites is identified as the most likely candidate breakpoint. All clipping sequences in the cluster are extracted and aligned together. A restore algorithm was used to calculate the most possible base in each position based on the aligned bases and their sequencing quality. The algorithm is based on a Bayesian model, where we compute the posterior probability estimation for A, C, G, T as:

\[
P(T_i|D) = \frac{P(T_{W_i})P(D|T_{W_i})}{\sum_{k=1}^{S} P(T_{W_k})P(D|T_{W_k})} \times \frac{P(T_{C_i})P(D|T_{C_i})}{\sum_{k=1}^{S} P(T_{C_k})P(D|T_{C_k})} = C_0 \times P(D|T_{W_i}) \times P(D|T_{C_i})
\]

\[
C_0 = \frac{\sum_{k=1}^{S} P(T_{W_k})P(D|T_{W_k})}{\sum_{k=1}^{S} P(T_{C_k})P(D|T_{C_k})}
\]
Here, D is the observation of the next-generation sequencing (NGS) reads on given position. P(Ti|D) is the likelihood component, which can be interpreted as the probability of observing D when the true genotype is Ti. Dv is a realization (or observation) of the NGS reads in the Watson strand. DC is a realization (or observation) of the NGS reads in the Crick strand. P(TWi|D) is the likelihood component, which can be interpreted as the probability of observing D when the true genotype is TWi. P ( TCi | D ) is the likelihood component, which can be interpreted as the probability of observing D when the true genotype is TCi. At each virus location, prior probability P(Ti) of each genotype Ti was set according to Table S5. The likelihood P(D | Ti) for the assumed genotype Ti was calculated from the observed allele types in the sequencing reads in formula 2. Thus, on the Watson strand, it is P(Dw|Ti), and on the Crick strand it is P(Dc|Ti). We defined the likelihood of observing allele dk in a read for a possible haploid genotype T as P(dk|T), on the Watson strand it is P(dwk|T), and on the Crick strand it is P(dc k(T)). So, for a set of n observed alleles at a locus, D = {d1, d2, , dn} on each strand, these probabilities are computed as shown by formulas (3) and (4), where Q stands for the base quality from the fastaq file.

\[
P(Dw|Ti) = \prod_{k=1}^{m} P(dwk|T), \quad P(Dc|Ti) = \prod_{k=1}^{n} P(dc k|T).
\] (2)

\[
P(dk|T) = \begin{cases} 
1 - 10^{-\frac{Q}{10}} & (T \in \{C, G, T\}) \\
1 - 10^{-\frac{Q}{10}} & (T \in \{A\}) \\
\end{cases}.
\] (3)

\[
P(Dc k(T)) = \begin{cases} 
1 - 10^{-\frac{Q}{2}} & (T \in \{C, G, T\}) \\
1 - 10^{-\frac{Q}{2}} & (T \in \{A\}) \\
\end{cases}.
\] (4)

We used “Y” and “R” to represent C/T and G/A, respectively (IUPAC nucleotide code). If a region is covered by both the Watson strand and the Crick strand, we were able to deduce the original base from Y or R by calculation.

4. Detection of viral integrations: The assembled clipping regions above were mapped to the given virus reference sequence with a Smith-Waterman local alignment tool from the EMBOSS package [24], which supports IUPAC DNA codes Y and R. Virus fragment location is extracted from the alignment results.

Discussion

In summary, we have implemented the first software tool to detect virus integration using BS data. Our software is based on bwa-meth, and by assembling and aligning soft-clip regions, it can find the virus breakpoints. However, accuracy of reads surrounding the breakpoints needs to be further improved. A virus usually integrates into regions that are homologous to both human and virus (micro-homologous) [25]. Therefore, we consider breakpoints predicted by our software tool correctly identified if they are within 10 bp of a real breakpoint (Fig. S2). With this definition, the accuracy of our predicted breakpoints can reach over 70%. Our results will be useful for analyzing BS data and related applications. Some of the results come with only a location on human genome, and the virus location missing. This may be due to the shortage of virus fragments. We simulated 3 kinds of

Figure 3: A demo plot of 1 viral integration cluster in its pre-insertion form.
reads, PE50, 90, and 150 with various lengths, and further simulated virus-inserted fragment with different lengths as well (Table S6); thus all cases described in Fig. 2 are mimicked here. All simulations sampled all possible reads, base by base with fixed insert sizes. As the result in Table S6 showed, the longer the reads, the more accurate a prediction can be achieved. In particular, for read lengths around 50 bp, BS-virus-finder is capable of finding the virus integration with an accuracy of more than 70%; for the read lengths between 90 bp and 150 bp, BS-virus-finder is capable of finding the virus integration with an accuracy of more than 90%. Apart from simulated data, we have performed WGS and WGBS sequencing of the PLC/PRF/5 hepatocellular carcinoma cell line (Table S4). As the results show, when the length of input is larger than 150 bp, the analysis result of WGBS is similar to the one of WGS. Additionally, BS-virus-finder is able to find breakpoints in 8 out of 9 regions identified by FISH [8]. Based on these experimental results, we believe that BS-virus-finder is a powerful software tool to analyze virus integration using BS data.

Additional Table S1: Alignment accuracy rate around the breakpoint region using PE50 data.
Additional Table S2: Alignment accuracy rate around the breakpoint region using PE90 data.
Additional Table S3: Alignment accuracy rate around the breakpoint region using PE150 data.
Additional Table S4: Mapping statistics of cell line sequencing data.
Additional Table S5: The prior probability of the Bayesian model used in the restoring process for bisulfite sequencing of integrated virus.
Additional Table S6: The performance of BS-virus-finder in silico with different read lengths and insert sizes.
Additional Figure S1: The performance of BS-virus-finder in various lengths of virus integration using PE50.
Additional Figure S2: The performance of BS-virus-finder in various lengths of virus integration using PE90.

Additional Figure S3: The performance of BS-virus-finder in various lengths of virus integration using PE150.

Additional Figure S4: The diagram of STR for the PLC/PRF/5 cell line.

Abbreviations


Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the National Natural Science Foundation of China (81602477) and Shenzhen Municipal Government of China (ZDSYS201507301424148).

Authors contributions


Acknowledgements

We appreciate the support of Xiaolin Liang and Hengtong Li in the College of Mathematics and Statistics, Changsha University of Science and Technology, for contributing advice to our research.

References