Xia *et al. BMC Medical Genomics* 2019, **12**(Suppl 1):19
https://doi.org/10.1186/s12920-018-0461-8

BMC Medical Genomics

METHODOLOGY Open Access

Detecting virus integration sites based on multiple related sequencing data by VirTect

Yuchao Xia[1†], Yun Liu[1†], Minghua Deng[1,2] and Ruibin Xi[1,3,4*]

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2018
Los Angeles, CA, USA. 10-12 June 2018

Abstract

Background: Since tumor often has a high level of intra-tumor heterogeneity, multiple tumor samples from the same patient at different locations or different time points are often sequenced to study tumor intra-heterogeneity or tumor evolution. In virus-related tumors such as human papillomavirus- and Hepatitis B Virus-related tumors, virus genome integrations can be critical driving events. It is thus important to investigate the integration sites of the virus genomes. Currently, a few algorithms for detecting virus integration sites based on high-throughput sequencing have been developed, but their insufficient performance in their sensitivity, specificity and computational complexity hinders their applications in multiple related tumor sequencing.

Results: We develop VirTect for detecting virus integration sites simultaneously from multiple related-sample data. This algorithm is mainly based on the joint analysis of short reads spanning breakpoints of integration sites from multiple samples. To achieve high specificity and breakpoint accuracy, a local precise sandwich alignment algorithm is used. Simulation and real data analyses show that, compared with other algorithms, VirTect is significantly more sensitive and has a similar or lower false discovery rate.

Conclusions: VirTect can provide more accurate breakpoint position and is computationally much more efficient in terms both memory requirement and computational time.

Keywords: Hidden Markov model, Split reads, Paired-end reads, HBV, HPV

Background

Cancer is a very heterogeneous disease. Tumor genomes can be vastly different between and, probably more importantly, within cancer patients. The intra-tumor heterogeneity poses great challenges for tumor treatments. In recent years, multi-regional high-throughput sequencing (HTS) has been widely used for studying intra-tumor heterogeneity [1–4], where tumor cells from different tumor regions are sequenced and somatic variations are profiled. These studies revealed that the level of intra-tumor heterogeneity varies greatly between different types of tumors and between different tumor patients. Because of intra-tumor heterogeneity, drug responses of tumor cells from different regions can also be significantly different [5]. Recently, joint somatic mutation detection algorithms based on multi-regional sequencing were developed [6]. These algorithms can greatly increase the sensitivity of somatic mutation detection and hence can provide more accurate mutation data for tumor heterogeneity and tumor evolution studies.

Many human cancers (~ 10–15%) are caused by viruses [7] such as human papillomavirus (HPV) [8] and Hepatitis B Virus (HBV) [9]. Viruses such as HPV and HBV can integrate their genomes to their host genomes and this genome integration is believed to be the major mechanism for their carcinogenic effects [10, 11]. Accurate detection of virus integration sites can provide invaluable information for studying molecular mechanisms of virus-related cancers, cancer genome evolution and even for developing cancer treatments. Recently, a number of algorithms have been developed for detecting viruses based on cancer

* Correspondence: ruibinxi@math.pku.edu.cn
†Yuchao Xia and Yun Liu contributed equally to this work.
[1]School of Mathematical Sciences, Peking University, Beijing 100871, China
[3]Center for Statistical Science, Peking University, Beijing 100871, China
Full list of author information is available at the end of the article

sequencing data [12–16]. Several of them are designed to detect virus integration sites. For example, Virana [13] and VirusSeq [12] can detect virus integration sites based on whole transcriptome sequencing (RNA-Seq) data. ViralFusionSeq [17] and VirusFinder [18, 19] can be used for whole-genome sequencing (WGS), whole exome sequencing (WES) data as well as for RNA-Seq data. However, the sensitivity of these methods is still low. When applying these methods to multi-regional sequencing data, because of their low sensitivity, common integration sites can very likely be detected in only a few regions but not in all regions. These false negatives can lead to over-estimation of tumor heterogeneity and incorrect inference of tumor evolution. Although increasing sequencing coverage could ameliorate this problem, it will also significantly increase the experimental costs. In addition, current methods are computationally very expensive in terms of memory requirement and computational time, making it very difficult to apply them to high coverage whole genome sequencing data.

In this paper, we introduce VirTect for sensitive and accurate detection of virus integration sites from multi-related-sample HTS data. VirTect makes full use of HTS data from multiple samples without pooling the data together and performs integrated analysis to detect virus integration sites. Compared with available virus detection methods, VirTect is significantly more sensitive and can provide more accurate breakpoint position with similar or lower false discovery rate (FDR). VirTect is computationally much more efficient than other algorithms in terms of both computational time and memory requirement—it only needs around one fifth of computational time of other methods. Furthermore, since VirTect performs joint analysis of multiple sample data, VirTect will give exactly the same breakpoint estimate for shared integration sites among different samples, and thus subsequent analysis such as tumor heterogeneity analysis and tumor evolution analysis would be more convenient.

Methods

The overall workflow of VirTect is shown in Fig. 1. VirTect uses fastq files or bam files of paired-end reads data as input. Fastq or bam files from different samples do not need to be merged as a single file and VirTect automatically extract necessary information. VirTect first aligns (for data in fastq files) or realigns (for data in bam files) short reads to human and virus reference genomes. Since the samples are related tumor samples, a portion of the virus integrations should be shared among some or all samples. Therefore, if all samples are pooled together, on average, the shared integrations would have more supporting reads than private integrations. Hence, the detection power for shared integration should be higher than private integrations. However, physically pooling all data together would be computationally not efficient. Instead, after mapping for each individual sample, VirTect extracts all reads from all samples that might contain virus integration information and use these reads jointly to detect integrations. These potential supporting reads are paired-end reads partially mapped to virus genomes or soft-clippedly mapped. Then, VirTect performs joint clustering analysis and joint precise local realignment of the extracted reads. Each cluster corresponds to one candidate virus integration site. A local precise hidden Markov Model (HMM) realignment procedure is applied to the reads in each cluster to get accurate integration sites. Details of VirTect are described below.

Data preprocessing

VirTect can use either raw data in fastq format as input or BWA-aligned [20] data in bam format as input. If users choose to use raw data as input, VirTect first maps the paired-end read to the host reference genome (e.g. human reference genome) and virus genomes using BWA. Samtools [21] are used to sort the bam files and remove duplicate reads. If the data are in bam format, we assume that short reads are only mapped to the host reference genome. In this case, VirTect first extracts all partially unaligned short reads and realign those reads to the host reference genome and the virus genome using BWA. We consider the bam format input because researchers often have bam files available for detecting other types of genomic variations. Starting from bam files, VirTect can save a significant amount of computational time. The virus genome usually can be obtained from databases such as (1) the NCBI virus database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/), (2) Genome Information Broker for Virus database (http://www.insdc.org/) [12], or (3) the virus database used by RINS [15].

VirTect extracts three sets of paired-end reads V_V, V_H and H from all related samples (Fig. 1). A paired-end read belong to V_V if its both ends are mapped to the virus genomes but at least one end is soft-clipped. If one end of a read pair is mapped to the host genome and the other end to the virus genomes, we denote this read pair belonging to V_H. The read set H consists of paired-end reads whose both ends are mapped to the host genome, but whose one ends are soft-clipped. Note that the short read sets V_V, V_H and H do not overlap and all the following analyses are based on these three sets of reads.

Virus integration sites detection

Given all reads in V_V, V_H and H , VirTect first clusters reads in V_H to get potential integration sites, and then applies a precise local realignment procedure to reads in V_V and H to get the exact breakpoint of the integration sites. In the clustering step, all reads in V_H are first sorted by

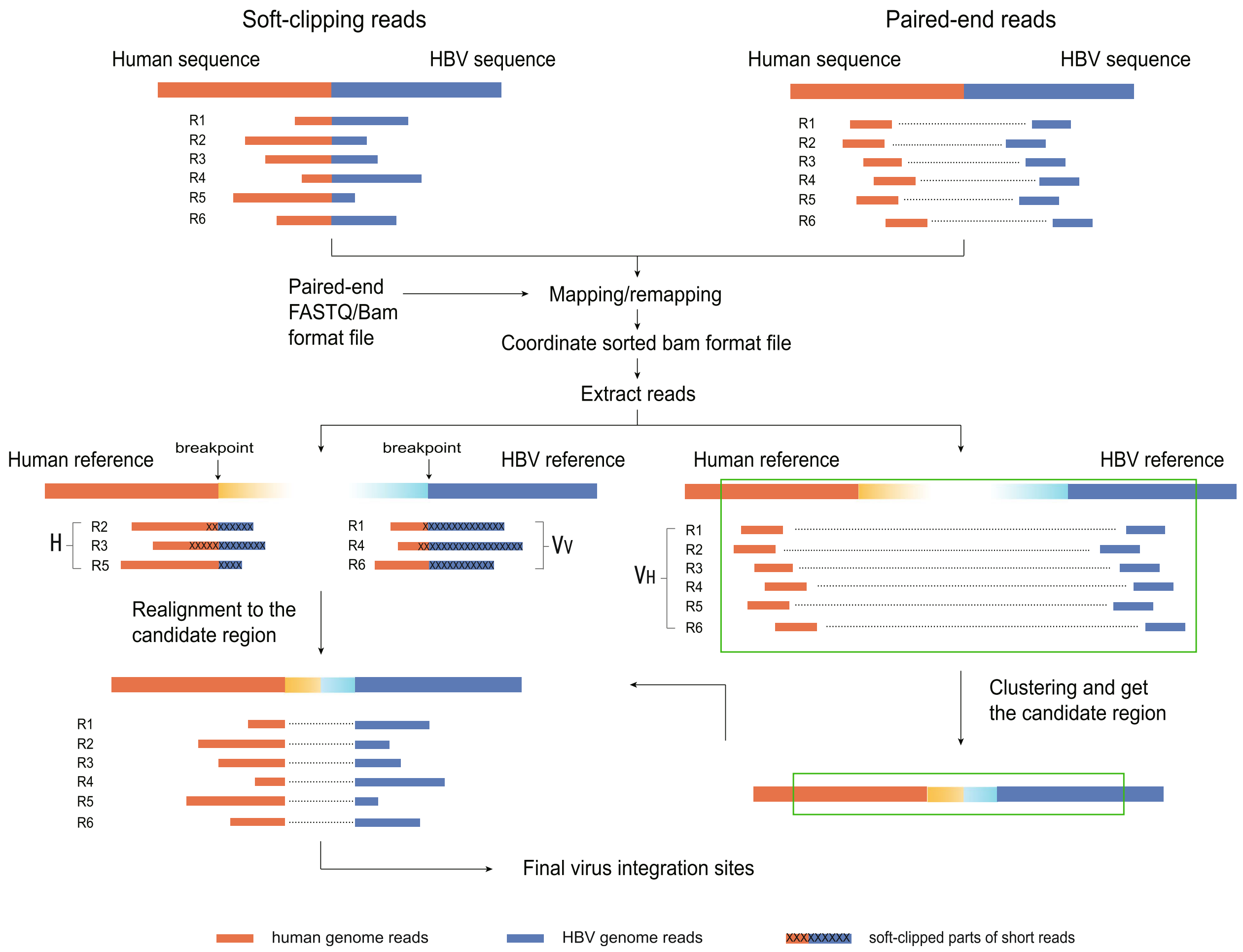

Fig. 1 The overall workflow of VirTect. After mapping or remapping of short reads to the human and virus genome, VirTect extracts short reads that might contain virus integration information for further analysis. The reads whose one ends were mapped to the human genome and other ends were mapped to the virus genome are clustered to get the candidate integration regions (bottom right). The reads that are soft-clippedly mapped to the human or the virus genome are realigned to candidate regions to get the exact breakpoints (bottom left). The soft-clipped parts of the short reads are marked with X's

their mapping coordinates on the host genome. VirTect employs an iterative procedure to cluster these reads and each cluster corresponds to a potential integration site. VirTect performs clustering for each chromosome of the host genome separately. Below we use one read to refer to one paired-end read (including both of its ends). Given a chromosome A in the host genome, suppose that $V_H^A \subset V_H$ is the set of reads whose one end is mapped to chromosome A and the other end is mapped to the virus genome. For any two pairs of short reads in V_H^A, we define their distance as the absolute difference of their mapping coordinates on chromosome A. The distance between two sets of short reads C_1 and C_2 is defined as the minimum distance between pairs of short reads in C_1 and C_2. We cluster short reads in V_H^A to M clusters $C_1, \cdots, C_M$ such that the distance between any two clusters is larger than a threshold T and each cluster cannot be further divided into two clusters with their distance greater than T. This is achieved by first sorting short reads in V_H^A increasingly by their mapping positions on chromosome A. Then, VirTect puts the first short read (i.e. the short read with the smallest mapping coordinate in A) as the first cluster C_1. Suppose that after k steps, we obtain m clusters $C_1, \cdots, C_m$. At the k + 1th step, VirTect takes the k + 1th read pair R_{k+1} and calculates its distances d_m to the cluster C_m. If d_m is less than the threshold T, we assign this read R_{k+1} to the cluster C_m; Otherwise, we assign the read R_{k+1} to a new cluster C_{m+1}. Note that since all paired-end reads are sorted increasingly by their mapping coordinates, the clusters C_i $(i = 1, \cdots, m)$ are also sorted increasingly. The distance between R_{k+1} and the cluster C_m will be the smallest among the distances between R_{k+1} and the clusters C_i ($i = 1, \cdots, m$). Therefore, d_m>T implies that all distances between R_{k+1} and C_i ($i = 1, \cdots, m$) is greater than T. By default, the threshold T is chosen as the mean insert size plus three standard deviations of the insert sizes.

Given a cluster C obtained from the above step, suppose that M_1 and M_2 are the minimum and maximum of the mapped coordinate positions in the host genome of the reads in the cluster C. Suppose that S_h (the subscript h refers to the host genome) is the subsequence of the host reference genome from $M_1 - N$ to $M_2 + N$ (N is taken as 500 by default). Then, S_h is our candidate virus integration region. Assume for now that each read in the cluster C have one end mapped to the host genome and the other end mapped to the same virus genome. Using the same procedure, we could also get a subsequence S_v (the subscript v refers to the virus genome) from the virus reference genome. VirTect concatenates the sequence S_h and S_v and get a new sequence S. This new sequence S is constructed for our precise local realignment. Then, VirTect selects reads in H that are soft-clippedly mapped to S_h and all reads from V_V for precise local HMM realignment [22]. Denote this set of reads as U. Since there is a large gap in S, a sandwich realignment [23] is used to realign the soft-clipping reads to the reference sequences based on the HMM realignment algorithm described in Xia et al. 2017 [24]. In the sandwich realignment, short reads are realigned both from the 5` end and the 3` end, which thus allows VirTect utilizing both the non-clipped and the clipped part of the short reads for precise localization of the breakpoints. The HMM model used in the sandwich realignment has an end-of-mapping state. With this state, the HMM realignment will not try to map every base pairs of short reads. Instead, it will terminate the mapping at or near the breakpoint, because after the breakpoints the mapping will contain many mismatches and/or gaps. Another advantage of this sandwich realignment is that if there are micro-insertions or micro-insertions at breakpoints, the mapping will automatically terminate and the split positions of short reads are naturally determined. The sandwich realignment does not need to explicitly search for the best split position. After realignment, VirTect first filters short reads whose either of its two directional mappings has less than 10 consecutive matches. VirTect also filters reads whose both directional mappings are mapped to the same side of junction point of S (i.e. the junction between S_h and S_v), because these reads do not span the breakpoint and cannot be used for breakpoint estimation. For the mappings to the left of the junction point of S (i.e. the mappings are to S_h), we take the median of the ending positions of the mappings as the estimate of the integration site on the host genome. The breakpoint on the virus genome is estimated by the median of the ending positions of the mappings to the right of the junction point of S (Fig. 1.). Lastly, VirTect will call this candidate region as an integration site if there are at least two reads whose sandwich alignments support the integration. All samples having paired-end alignments or sandwich alignments supporting the integration event are predicted to harbor this integration event.

If the reads in the cluster are mapped to a few different virus genomes (one end of the read always map to the host genome, but the other end map to different virus genomes), we create a concatenated sequence S for each mapped virus genome. For all soft-clippedly mapped to S_h, we perform sandwich realignment to each concatenated sequence S. We then can obtain a realignment likelihood for each concatenated sequence S and choose the virus genome with the largest likelihood for further analysis. If two likelihoods are the same, we choose the virus genome having the most number of reads in the cluster for further analysis.

Results

Simulation study

We first compare the performance of VirTect with other three methods ViralFusionSeq, VirusFinder2 and Virus-Clip [25] using simulation. We randomly select 160 viral sequences (sizes ranging from 500 bp to 1000 bp) from genotype C [26] of the HBV genome and insert them to chromosome 1, 2, 3 and 4 of the human reference genome (hg19, GRCh37). The GenBank ID of HBV Genotype C is AB014381.1. In this way, we generate five related genomes with virus integrations. Each genome has 40 virus integration sites and 25 of them are common in all five genomes. In the simulation, we also randomly put SNVs and Indels near the integration site (50 bp neighborhood). Given the four simulated genomes, we use ART [27] to simulate the Illumina paired-end reads with a read length of 100 bp and an insert size of 300 bp (standard deviation 50 bp). For each genome, we simulate six datasets at coverage 3X, 5X, 10X, 20X, 30X and 40X. For VirTect, we test its performance starting from fastq files (VirTect:fastq) and from bam files (VirTect:bam). For VirTect:fastq, we use BWA to map all paired-end reads to the human reference genome (hg19) and the HBV genomes (genotype A-H) simultaneously. For VirTect:bam, the short reads are first mapped to the human reference genome. VirTect then uses BWA to realign the partially unaligned reads to the HBV genomes (genotype A-H). For the other two algorithms, we use the default parameter settings. All algorithms are tested on a Linux sever (32-core Intel Xeon 2.40 GHz CPU and 256Gb memory).

We first apply the other three algorithms to each data set individually and compare their performances with VirTect. Figure 2 shows the sensitivities and false discovery rates (FDR) of these algorithms on each genome separately.

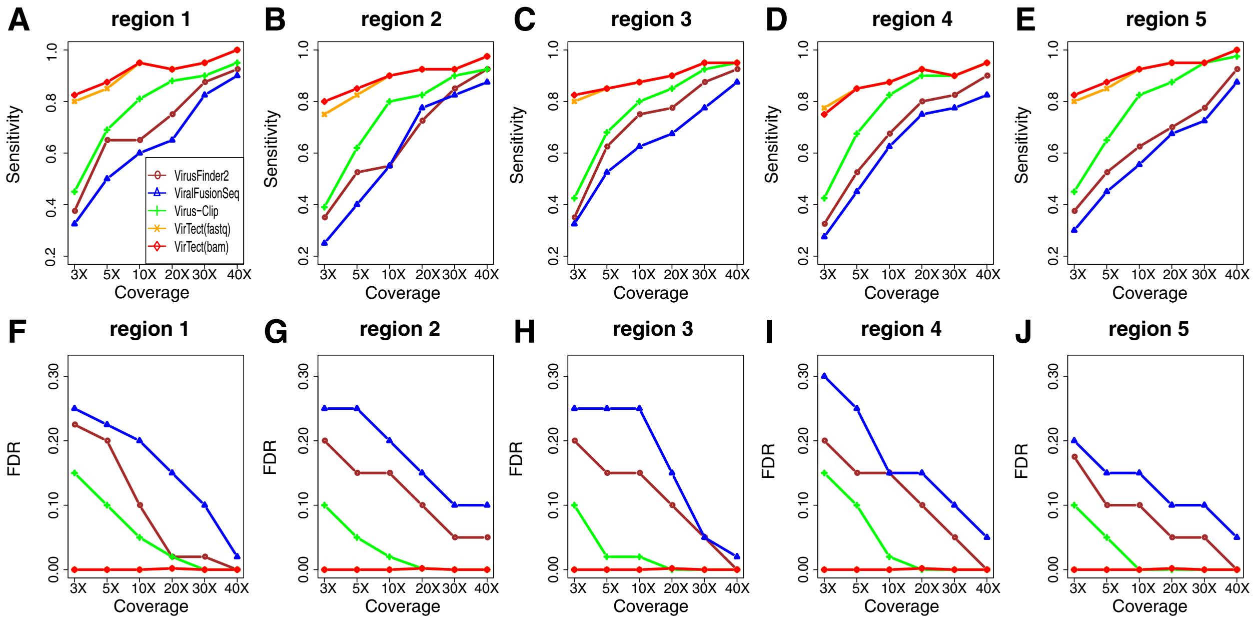

Fig. 2 The sensitivity (**a**-**e**) and FDR (**f**-**j**) of the four algorithms on the simulation data at different sequencing coverages

We define an integration prediction as a true positive if the distance between the predicted integration site and the real integration site is less than 350 bp. We find that VirTect achieves the highest sensitivities and the lowest FDR across all five genomes. Especially, at low coverage depth (3X, 5X and 10X), the sensitivities of VirTect are much higher than the other three algorithms and its FDRs are 0. VirTect:fastq is a little more sensitive than VirTect:bam at 3X and 5X coverage, but overall their performances are very similar. The other three algorithms had a higher FDR at low coverage because a number of predicted integration sites are far from the true integration sites.

The above comparison is a bit unfair for other algorithms because the other three algorithms do not use all data to detect common integration sites. We then merge sequencing data of the five genomes as one data set and

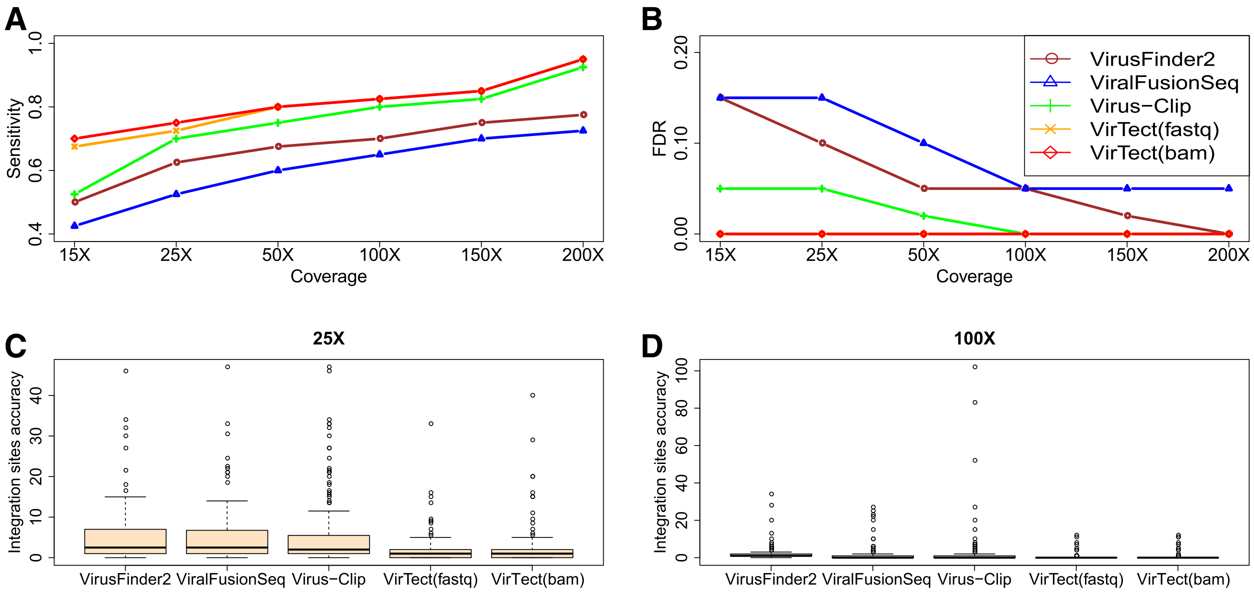

Fig. 3 The Sensitivity (**a**) and FDR (**b**) on the merged data. (**c**, **d**) Boxplots of breakpoint estimation accuracy on merged data at 25X and 100X coverage

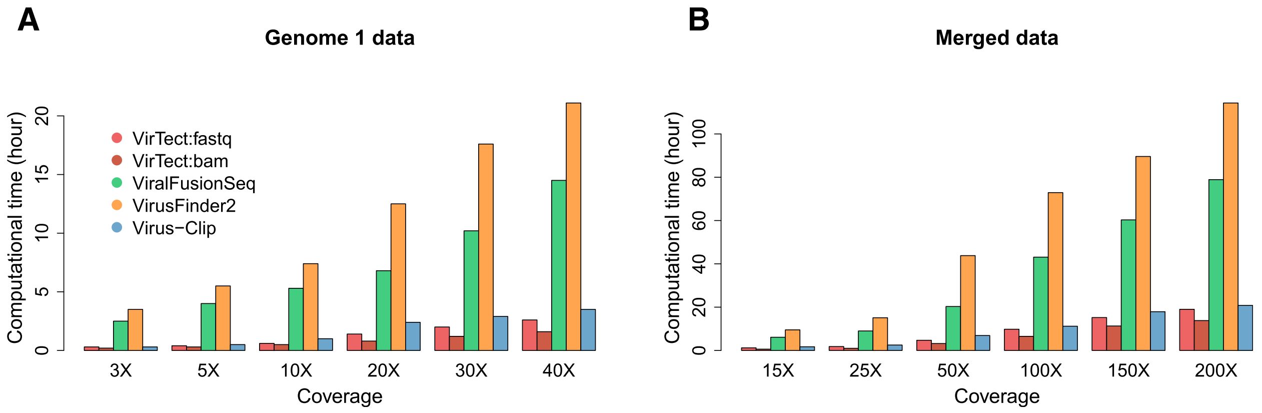

Fig. 4 a The computational time (in hour) on the simulated Genome 1 data at different coverages. **b** The computational time (in hour) on merged data at different coverages

apply the other two algorithms to the merged data and compare their performances. Note that for VirTect, we do not need to physically merge the data and this is more convenient to analyze multiple related-samples. Figure 3a-b shows that VirTect also has the highest sensitivity and the lowest FDR across all coverages among the three algorithms. Figure 3c and d shows the distance between the detected integration sites and the true integration sites at 25X and 100X coverage. Compared with the other algorithms, the integration sites predicted by VirTect are closest to the true integration sites and the predicted integration sites of VirTect are only up to a few bp away from the true integration sites in most cases. We also compare the computational time of different algorithms. Figure 4. shows the running time using eight cores on the simulation dataset of Genome 1 and the merged dataset, respectively. We see that VirTect only takes around 1 fifth of the computational time of ViralFusionSeq and VirusFinder2 and a little faster than Virus-Clip.

Real data analysis

In this section, we compare the performance of VirTect with the other two algorithms on real data sets. We consider two real data sets in this study. One is a multi-regional whole exome sequencing (WES) data from an HBV-related hepatocellular carcinoma (HCC) patient [28]. The patient's ID is 213 and tumors from five regions are sequenced by

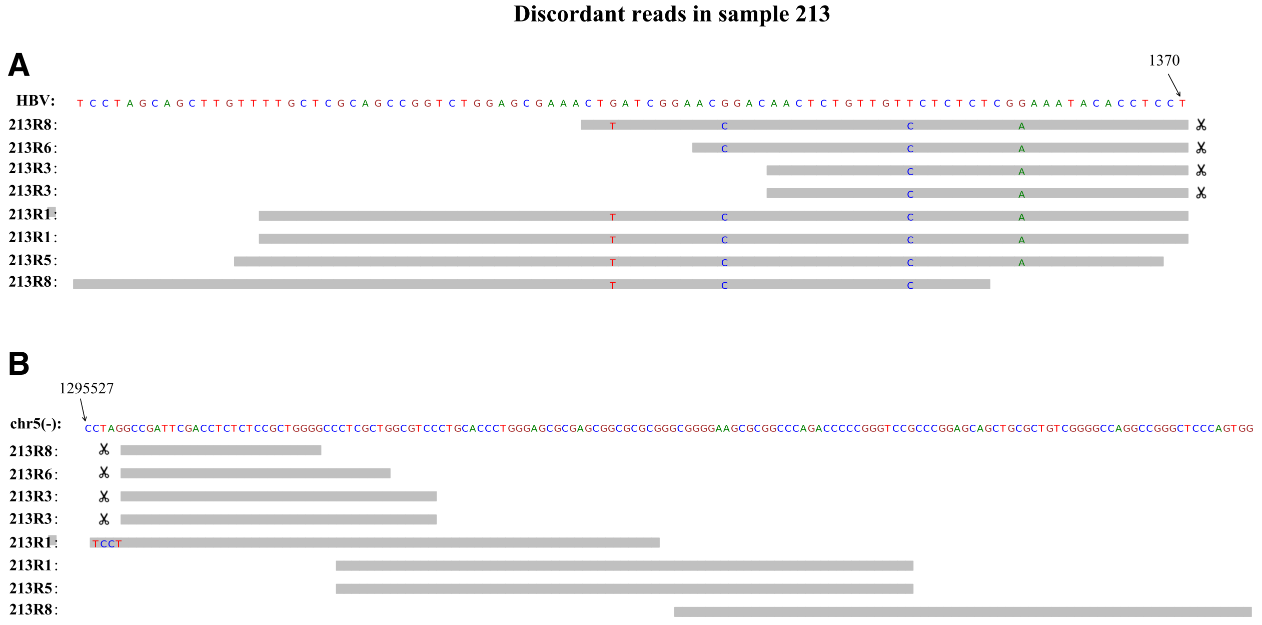
Discordant reads in sample 213
A
1370
HBV: TCCTAGCAGCTTGTTTTGCTCGCAGCCGGTCTGGAGCGAAACTGATCGGAACGGACAACTCTGTTGTTCTCTCTCGGAAATACACCTCCT
213R8: T C C A
213R6: C C A
213R3: C A
213R3: C A
213R1: T C C A
213R1: T C C A
213R5: T C C A
213R8: T C C
B
1295527
chr5(-): CCTAGGCCGATTCGACCTCTCTCCGCTGGGGCCCTCGCTGGCGTCCCTGCACCCTGGGAGCGCGAGCGGCGCGCGGGCGGGGAAGCGCGGCCCAGACCCCCGGGTCCGCCCGGAGCAGCTGCGCTGTCGGGGCCAGGCCGGGCTCCCAGTGG
213R8:
213R6:
213R3:
213R3:
213R1: TCCT
213R1:
213R5:
213R8:

Fig. 5 VirTect identifies an HBV integration site at the *TERT* promoter region in patient 213. All tumors from different regions have this integration event. The discordant and sandwich-mapped reads to the HBV genome (**a**) and human genome (**b**) are shown

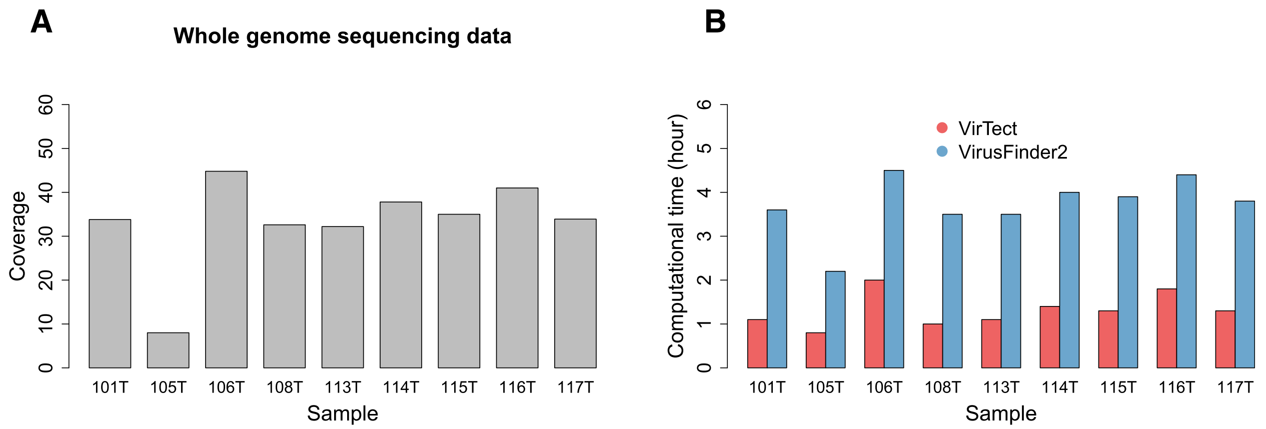

Fig. 6 a The mean coverage of the 9 WGS data. **b** The computational time (in day) of VirTect and VirusFinder2 on these 9 WGS data

Illumina platform with a read length of 75 bp. The mean insert size is 200 bp with a standard deviation of 50 bp. The other data consists of nine whole genome sequencing (WGS) data of HBV-related HCC patients [29]. The read length of this data is 90 bp and the coverage is around 30X.

For the multi-region WES data, VirTect is able to detect one HBV integration sites. The integration site is at chromosome 5:1295527 (Fig. 5). The integration sites are located at promoter region of the telomerase reverse transcriptase (TERT). Previous research showed that TERT is the most prevalent gene integrated by HBV in HCC [30]. Moreover, all tumor regions have this integration event, implying that this event might be an early carcinogenesis event. When we apply the other two algorithms to data of each region, they fail to detect any integration site. When we merge the multi-regional data together, they also fail to detect any event.

For the WGS data, we downloaded hepatocellular carcinoma samples, 101 T, 105 T, 106 T, 108 T, 113 T,114 T,115 T,116 T and 117 T reported by Sung et al. 2012 [29]. Here, we only report the results for VirTect and VirusFinder2 because ViralFusionSeq failed due to insufficient memory and Virus-Clip did not finish computation after a week. The running time of VirTect and VirusFinder2 is shown in Fig. 6. VirTect and VirusFinder2 detected all integration sites reported by Sung et al. 2012 [29]. Some of these integration sites interrupt important cancer genes such as CCNE1 (sample 106 T, chr19:30304177) and NTRK3 (sample 108 T, chr15:88688212). Details about these integration sites are in Additional file 1: Table S1. Figure 7a shows one integration cite at chr1:151503388 at the gene CGN. VirusFinder2 costs long time (> 3 days) and a large amount of memory (about 70 Gb) to finish the

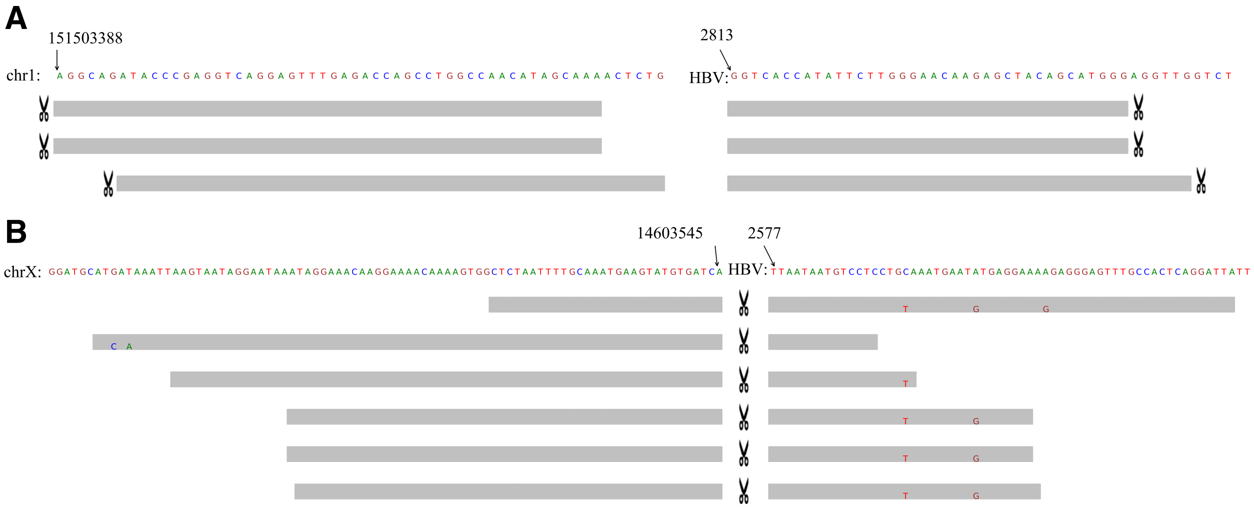
A
151503388
chr1: AGGCAGATACCCGAGGTCAGGAGTTTGAGACCAGCCTGGCCAACATAGCAAAACTCTG
2813
HBV: GGTCACCATATTCTTGGGAACAAGAGCTACAGCATGGGAGGTTGGTCT
B
14603545
chrX: GGATGCATGATAAATTAAGTAATAGGAATAAATAGGAAACAAGGAAAACAAAAGTGGCTCTAATTTTGCAAATGAAGTATGTGATCA
2577
HBV: TTAATAATGTCCTCCTGCAAATGAATATGAGGAAAAGAGGGAGTTTGCCACTCAGGATTATT
T G G
C A
T
T G
T G
T G

Fig. 7 a A known HBV integration site detected by both VirTect and VirusFinder2 in sample 101 T. The mappings of the supporting reads to the human genome (left panel) and to the virus genome (right panel) are shown. The split position of each read is marked by a scissor icon. **b** A new integration site detected by VirTect computation. In comparison, VirTect uses 1.5 days and no more than 30Gb memory. In addition to the reported integration sites, VirTect detects a new integration site at chromosome X:14603545 (Fig. 7b) overlapping with the gene GLRA3.

Discussion

To our knowledge, VirTect is the first algorithm capable of detecting virus integration from multiple related tumor samples. Intra-tumor heterogeneity and tumor cell evolution have received great research attention recently. Virus integration site can provide valuable information for inferring the evolution history of tumor cells. For example, if all tumor cells have exact the same integration site, we can confidently infer that these tumor cells evolve from the same ancestor. VirTect could also be used to detect virus integration in RNA-seq data, especially multiple-related RNA-seq data. One drawback of VirTect is that it requires known viruses. Hence, VirTect is not suitable to detect integration sites of unknown viruses.

Conclusion

In this paper, we develop a computational tool called VirTect for virus integration site detection. Simulation and real data analysis show that VirTect performs considerable better than other available algorithms in terms of sensitivity, false discovery rate, breakpoint position, computational time and memory requirement. With its high integration detection accuracy, we expect that VirTect can be widely applied to virus integration genomics studies.

Availability and requirements

Project name: VirTect.

Project home page: https://github.com/xyc0813/VirTect/

Operating system(s): Windows,Unix-like (Linux, Mac OSX).

Programming language: python(> = 2.7),Cython.

Any restrictions to use by non-academics: None.

Additional File

Additional file 1: Table S1. The integration cites detected by VirTect and VirusFinder2 in the nine whole genome sequencing data of HBV-related HCC patients. (XLS 30 kb)

Abbreviations

HBV: Hepatitis B Virus; HPV: Human papillomavirus; HTS: High-throughput sequencing; WES: Whole exome sequencing; WGS: Whole genome sequencing

Acknowledgements

Part of the analysis was performed on the Computing Platform of the Center for Life Science.

Funding

This work has been supported by the National Key Basic Research Project of China [2015CB856000, 2016YFC0207705], the National Natural Science Foundation of China [11471022, 71532001] and the Recruitment Program of Global Youth Experts of China. Publication of this article was sponsored by 11471022 grant.

Availability of data and materials

Whole-genome sequence data used in this paper can be download from the European Genome-phenome Archive (EGA) under the accession ERP001196.

About this supplement

This article has been published as part of *BMC Medical Genomics Volume 12 Supplement 1, 2019: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM) 2018: medical genomics*. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-1.

Author's contributions

YX implemented the package. YX and RX wrote the paper. YX and YL performed the analysis. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Mathematical Sciences, Peking University, Beijing 100871, China.
[2]Center for Quantitative Biology, Peking University, Beijing 100871, China.
[3]Center for Statistical Science, Peking University, Beijing 100871, China.
[4]Center for Data Science, Peking University, Beijing 100871, China.

Published: 31 January 2019

References

1. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Loo PV, Aas T, Alexandrov LB, Larsimont D, Davies H. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015;21(7):751.
2. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883–92.
3. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow CW, Cao Y, Gumbs C. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014; 346(6206):256.
4. Shi JY, Xing Q, Duan M, Wang ZC, Yang LX, Zhao YJ, Wang XY, Liu Y, Deng M, Ding ZB. Inferring the progression of multifocal liver cancer from spatial and temporal genomic heterogeneity. Oncotarget. 2015;7(3):2867–77.
5. Housman G, Byler S, Heerboth S, Lapinska K, Longacre M, Snyder N, Sarkar S. Drug resistance in Cancer: an overview. Cancers. 2014;6(3):1769–92.
6. Josephidou M, Lynch AG. TavarÃ© S: multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. Nucleic Acids Res. 2015;43(9):e61.
7. Morissette G, Flamand L. Herpesviruses and chromosomal integration. J Virol. 2010;84(23):12100–9.
8. Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, Snijders PJ, Peto J, Meijer C. MuÃ±oz N: human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol. 1999;189(1):12–9.

9. Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, Kennemer MI, Guan Y, Lee W, Carnevali P, Stinson J, Johnson S. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res. 2012; 22(4):593–601.
10. Chen Y, Williams V, Filippova M, Filippov V, Duerksenhughes P. Viral carcinogenesis: factors inducing DNA damage and virus integration. Cancers. 2014;6(4):2155.
11. Hai H, Tamori A, Kawada N. Role of hepatitis B virus DNA integration in human hepatocarcinogenesis. World Journal of Gastroenterology Wjg. 2014; 20(20):6236–43.
12. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics. 2013;29(2):266–7.
13. Schelhorn SE, Fischer M, Tolosi L, Altmuller J, Nurnberg P, Pfister H, Lengauer T, Berthold F. Sensitive detection of viral transcripts in human tumor transcriptomes. PLoS Comput Biol. 2013;9(10):e1003228.
14. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol. 2011;29(5):393–6.
15. Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics. 2012;28(8):1174–5.
16. Naeem R, Rashid M, Pain A. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. Bioinformatics. 2013;29(3):391–2.
17. Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics. 2013;29(5):649–51.
18. Wang Q, Jia P, Zhao Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS One. 2013;8(5):e64465.
19. Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome medicine. 2015;7(1):2.
20. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010;26(5):589–95.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome project data processing S: the sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
22. Durbin BR, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids; 1998.
23. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.
24. Xia Y, Liu Y, Deng M, Xi R. SVmine improves structural variation detection by integrative mining of predictions from multiple algorithms. Bioinformatics. 2017;33(21):3348–54.
25. Ho DWH, Sze KMF, Ng IOL. Virus-clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget. 2015;6(25):20959–63.
26. Kao C. HBV genotypes: epidemiology and implications regarding natural history. Current Hepatitis Reports. 2006;5(1):5–13.
27. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4.
28. Gao Q, Wang ZC, Duan M, Lin YH, Zhou XY, Worthley DL, Wang XY, Niu G, Xia Y, Deng M, et al. Cell culture system for analysis of genetic heterogeneity within hepatocellular carcinomas and response to pharmacologic agents. Gastroenterology. 2016;152(1):232–42.
29. Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet. 2012;44(7):765–9.
30. Lucifora J, Arzberger S, Durantel D, Belloni L, Strubin M, Levrero M, Zoulim F, Hantz O, Protzer U. Hepatitis B virus X protein is essential to initiate and maintain virus replication after infection. J Hepatol. 2011;55(5):996–1003.