Bioinformatics Advance Access published November 17, 2012

Category: Genome analysis

VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue

Yunxin Chen[1], Hui Yao[2], Erika J Thompson[3], Nizar M Tannir[1], John N Weinstein[2] and Xiaoping Su[2*]
Departments of [1]Genitourinary Medical Oncology, [2]Bioinformatics and Computational Biology, and [3]Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

Associate Editor: Dr. Michael Brudno

ABSTRACT

Summary: We developed a new algorithmic method, VirusSeq, for detecting known viruses and their integration sites in the human genome using next-generation sequencing data. We evaluated VirusSeq on whole-transcriptome sequencing (RNA-Seq) data of 256 human cancer samples from The Cancer Genome Atlas (TCGA). Using these data, we showed that VirusSeq accurately detects the known viruses and their integration sites with high sensitivity and specificity. VirusSeq can also perform this function using whole-genome sequencing data of human tissue.

Availability: VirusSeq has been implemented in PERL and is available at http://odin.mdacc.tmc.edu/~xsu1/VirusSeq.html

Contact: xsu1@mdanderson.org

1 INTRODUCTION

About 12% of all human cancers are known to be caused by viruses (Hausen, 2009), thus, the detection of viruses in human cancer tissue has significant clinical implications in oncology. The advent of next-generation sequencing (NGS) technologies using paired-end reads allows for the detection of viruses in human cancer tissue at unprecedented levels of efficiency and precision. Several groups have developed computational tools for pathogen/virus discovery by exploiting the great amount of NGS data obtained from human tissue (Kostic et al., 2011, Isakov et al., 2011). These groups have implemented a computational subtraction analysis, which has also been used to discover a new polyomavirus associated with most cases of Merkel cell carcinoma (Feng, et al., 2008).

Although detecting viruses in human tissue is important in clinical oncology, investigating virus integration sites in host cell chromosomes is equally valuable since insertional mutagenesis is one of the most critical steps in the pathogenesis of hepatitis B virus (HBV)-mediated hepatocellular carcinoma (HCC; Paterlini-Brechot, et al., 2003). NGS data have been used to map the HBV integration sites in HCC samples (Sung et al., 2012, Jiang et al., 2012). However, no software tool is currently available for detecting viral integration sites by NGS data. We present VirusSeq, which starts with computational subtraction on NGS data, and subsequently identifies viruses and their potential integration sites in the human genome with high specificity and sensitivity.

2 METHODS

(1) Mapping/Alignment. The paired-end (PE) reads in FASTQ format are used as input. VirusSeq works with both whole-genome and whole-transcriptome sequencing data. The raw PE reads are aligned to the reference genome using MOSAIK (Hiller et al., 2008) alignment software, which implements both a hashing scheme and the Smith-Waterman algorithm to produce gapped optimal alignments.

(2) Virus detection from NGS data: VirusSeq starts with computational subtraction of human sequences by aligning raw PE reads from whole-genome/transcriptome sequencing to the human genome reference. Thus, a set of nonhuman sequences is effectively generated by subtracting the human sequences. In the second step, VirusSeq aligns the nonhuman sequences against a comprehensive database that includes all known viral sequences from Genome Information Broker for Viruses (GIB-V, http://gib-v.genes.nig.ac.jp/), and quantifies the virus representation by the overall count of mapped reads within a virus genome to determine the existence of viruses in human samples with an empirical cutoff. Any virus with an overall count of mapped reads below the cutoff is treated as nonexistent. We used 1000 as the cutoff for the overall count of mapped reads within a virus genome; this cutoff should be applicable for both RNA-Seq data and whole-genome sequencing data with 30X coverage. This cutoff should be reduced by half or more for low-pass whole-genome sequencing data.

(3) Identification of virus integration sites: The genome sequences of viruses, which are well known in terms of cancer association and were detected in the previous step in our TCGA dataset, were concatenated into a single chromosome named chrVirus, with related annotation of each viral gene in refFlat format. A new hybrid reference genome named hg19Virus is built by combining hg19 and chrVirus (designated as chr25 in hg19Virus). All PE reads without computational subtraction are mapped to this reference (hg19Virus). If the PE reads are uniquely mapped with one end to one human chromosome and the other to chr25, the read pair is reported as a discordant read pair. All discordant reads are then annotated with human and viral genes defined in the curated refFlat file. VirusSeq then clusters the discordant read pairs that support the same integration (fusion) event (e.g., HBV-MLL4). VirusSeq implements a dynamic clustering procedure (details in Supplementary Notes) to accurately determine the boundary of the cluster, whose size is constrained by the insert size (fragment length) distribution. In order to remove outliers within a cluster, VirusSeq implements the robust "extreme studentized deviate (ESD)" multiple-outlier detection procedure (Rosner, 1983). Once outliers are detected within a cluster, the cluster boundary is reset by excluding the outlier reads. VirusSeq reports the fusion candidates by using both supporting pairs (at least four) and junction spanning reads (at least one) as the cutoffs. Meanwhile, an *in silico* sequence is generated using the consensus of reads within discord-

[*]To whom correspondence should be addressed.

ant read clusters for each fusion candidate to help the PCR primer design, which facilitates quick PCR validation.

3 RESULTS AND DISCUSSION

To test the accuracy of VirusSeq, we analyzed RNA-Seq data of 17 HCC cancers available in the TCGA database and detected HBV transcripts in 4 cases (Table 1A), two of which are from patients with serologic evidence of HBV infection and one from a patient who is seronegative for HBV (and hepatitis C virus). Serology data were not available for the remaining case. Viral integration loci identified in our analysis included known genes MLL4 (2 cancers; both from the two HBV-seropositive patients) and TERT, ITGAD, TEAD1, TECRL, C19orf55, and MIR548D2. Interestingly, the cancer with TERT-associated HBV sequences came from the patient who was reportedly seronegative for HBV. Our findings validate other reports that have demonstrated HBV insertion in TERT and MLL4 (Ferber et al., 2003, Saigo et al., 2008).

Table 1A. Characterization of genes with hepatitis B virus (HBV) integration breakpoints in hepatocellular carcinoma.

SampleID	*Support pairs*	*Viral transcripts*	*Host genes*	*Integration locations*
L526401A	14	HBVgp2_S protein	ITGAD	5prime
	131	HBVgp4_core/e-antigen	ITGAD	intron5
	29	HBVgp4_precore/core protein	ITGAD	intron5
	6	HBVgp2_S protein	MIR548D2	intron2
	281	HBVgp3_X protein	TEAD1	intron2
	5	HBVgp1_polymerase	TECRL	3prime
	18	HBVgp2_S protein	TECRL	3prime
	21	HBVgp3_X protein	TECRL	3prime
LA11601A	20	HBVgp2_S protein	C19orf55	intron6
	8	HBVgp2_S protein	MLL4	intron5
LA11901A	21	HBVgp4_core/e-antigen	MLL4	exon3
LA1HT01A	5	HBVgp2_S protein	TERT	intron1

Table 1B. Estimation of sensitivity and specificity for HPV16 detection in head and neck squamous cell carcinoma (HNSCC) samples. P-values were calculated by Fisher's exact test.

		HPV+	*HPV-*	*All*	*P-value*
HPV *in situ* hybridization	-	0	35	35	
	+	6	0	6	
	All	6	35	41	<0.0001
HPV by p16 immunohistochemistry	-	0	36	36	
	+	7	0	7	
	All	7	36	43	<0.0001

We also analyzed RNA-Seq data of 239 cases of head and neck squamous cell carcinoma (HNSCC) available in the TCGA database. We detected human papillomavirus (HPV) transcripts in 37 cancers, as follows: 30 cancers with HPV16, 5 cancers with HPV33, 1 cancer with HPV35, and 1 cancer with Epstein-Barr virus (EBV). In 24 cancers, HPV transcripts encoding for key viral proteins/oncoproteins (E7 in 22 cases; E6 in 20 cases; E1 in 17 cases and E4 in 8 cases) were integrated in the cancer genome, the majority in association with known genes. We used the HPV16 status from colorimetric *in situ* hybridization and the p16 immunohistochemistry data (clinical data) to estimate the sensitivity and specificity for HPV16 detection. We found that a total of 8 samples were HPV16-positive from colorimetric *in situ* hybridization (6 HPV16-positive) and/or p16 immunohistochemistry (7 HPV16-positive), and 36 samples were HPV16-negative. The HPV16 status was not available for all the remaining samples. The confidence intervals (CIs) were estimated using the Wilson score method by taking the sample size into consideration. For this HNSCC dataset, the sensitivity was 100% (8/8) with a 95% CI of 67.6% to 100%, and specificity was 100% (36/36) with a 95% CI of 90.4% to 100% (Table 1B).

We have developed a new algorithmic method called VirusSeq for detecting the known viruses and their integration sites in the human genome using NGS data. We evaluated VirusSeq on RNA-Seq data of 17 HCC and 239 HNSCC samples, and showed that VirusSeq accurately detects the known viruses and their integration sites. VirusSeq can also perform this function using whole-genome sequencing data obtained from human tissue. The main limitation of VirusSeq is the requirement of the known virus database to nominate candidate viruses in human cancer tissue. This will certainly miss novel viruses that are not in the virus database. We expect VirusSeq to be an effective solution for detecting viruses and their integration sites in cancer studies. We invite users to test our software.

ACKNOWLEDGEMENTS

This work was supported in part by the National Cancer Institute, U.S. National Institutes of Health, through grant U24CA143883 for TCGA, and grant P30 CA016672, the Cancer Center Support Grant to The University of Texas MD Anderson Cancer Center; also by the National Center for Research Resources through grant UL1TR000371; and the H. A. and Mary K. Chapman Foundation and the Michael & Susan Dell Foundation.

REFERENCES

Feng, H. *et al*. (2008) Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science*, **319**(5866), 1096-1100.

Ferber, M.J. et al. (2003) Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. Oncogene, 22, 3813-3820.

Hausen, Z. (2009) The search for infectious causes of human cancers: where and why. Virology, 329, 1-10.

Hiller, L.W. et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. Nature Methods, 5, 183-188.

Isakov, O. et al. (2011) Pathogen detection using short-RNA deep sequencing subtraction and assembly. Bioinformatics, 27(15), 2027-2030.

Jiang, Z. et al. (2012) The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Research, 22, 593-601.

Kostic, A.D. et al. (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature Biotech., 29(5), 393-396.

Paterlini-Brechot, P. et al. (2003) Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. Oncogene, 22(25), 3911-6.

Rosner, B. (1983) Percentage points for a generalized ESD many outlier procedure. Technometrics, 25(2), 165-172.

Saigo, A. et al. (2008) Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (MLL4) gene and rearrangements of MLL4 in human hepatocellular carcinoma. Hum Mutat., 29, 703-708.

Sung, W. et al. (2012). Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nature Genetics, 44(7), 765-769.