CSCE Extended Abstract/Poster Paper

# Detection of Virus Integration Sites in Tumor Genomes Using Deep Convolutional Neural Networks

Vibhuti Gupta[1§], Qingguo Wang[1§], Emirrah Sanders[2]
[1]School of Applied Computational Sciences, Meharry Medical College, Nashville, TN, 37208, USA
[2]School of Graduate Studies and Research, Meharry Medical College, Nashville, TN, 37208, USA
[§]Corresponding Email: vgupta@mmc.edu, qiwang@mmc.edu

*Abstract*—**Pathogenic viruses are estimated to be responsible for 15% of all human cancers globally and pose significant threats to public health. Viruses integrate their genetic material to host genome and increases the risk of cancer promoting changes in it. To understand the molecular mechanisms of viral mediated cancers, a necessary step is to detect virus insertion sites in cancer genomes, however the challenges posed by exponentially increasing volume of tumor sequencing data and by accurate analysis of data needs to be addressed. In this paper, we propose a deep convolutional neural network (CNN) based framework for detecting virus integration sites in tumor genomes. Our contributions are twofold: (i) we design a novel approach by constructing image pairs from NGS data including reads from host and virus genome; (ii) we utilize one-hot encoded images with lower computational complexity to represent viral integration sites and leverage the power of twin Deep CNN networks to perform the detection; (iii) we integrate above techniques into an end-to-end framework to produce next-generation tool for NGS-based detection of virus integration. Finally, we point out related datasets for experiments and future research directions.**

***Keywords— Virus; cancer; Deep CNN; NGS; genomes***

## I. Introduction

Pathogenic viruses pose significant threats to public health throughout the world and are estimated to be responsible for 15% of all human cancers globally [1][2]. For example, human papillomavirus (HPV) causes 91 percent of cases of cervical cancer, the fourth most common cancer in women globally [3]. In cervical cancer and some other viral mediated cancers, viruses can integrate their genetic material into host cell genome [4]. The process of viral integration causes damages to the host cell DNA and increases the risk of cancer-promoting changes occurring in the host genome [5][6]. Therefore, to understand the molecular mechanisms of viral mediated cancers, a necessary step is to detect viruses and their insertion sites in cancer genomes.

With the rapid advances in next generation sequencing (NGS) technologies over the past two decades and their increasingly widespread applications in hospitals, many NGS-based tools [7-9] were developed to detect viruses and their insertion sites, however due to the challenge of accurate detection, the sensitivity of today's tools is still not satisfactory [10,11] not matching that of the established quantitative technologies. Virus insertions in human genomes cause instability of the human genome, leading to elevated mutation rates. The fusion-caused mutations make the alignment of short reads to reference genome difficult and, consequently, the detection of virus integration sites challenging. Moreover, high virus mutation rates, lead to viral sequence divergence which adversely impact the detection and makes the NGS reads sampled from the real virus genomes less likely to align to the commonly used virus reference sequences.

Existing NGS tools use statistical models to identify viral integration events [12 – 17] , however due to the noise in the sequencing data and uncertainty in the alignment of reads, these tools only retain the reads for analysis that pass various quality filters, the thresholds of which are largely empirically determined, in order to control false positive rates. As a consequence of such filtering, their ability of detecting cryptic viral insertions is compromised. Some of the existing tools such as VirusFinder [7][8] played a critical role in characterizing integration sites of undiagnosed viruses of arbitrary types through the sequencing data. Due to the uniqueness and accuracy of VirusFinder, it has been widely used in investigating various types of cancer [11][18], however due to exponentially increasing volume of genomics data and its inability to detect cryptic viral insertion events, further research is required to tackle these challenges.

Due to advances in machine learning nowadays, deep neural networks are widely used for various applications such as image recognition, genomic analysis, COVID-19 detection etc. Deep CNNs are very powerful in visual recognition tasks due to their efficiency in capturing the spatial and temporal dependencies of the input [19]. The sequencing reads from a sample form an image in essence after being aligned to reference genome (or transcriptome). Comparing with traditional methods, deep CNNs are composed of stacks of processing layers, enabling them to learn complex features hierarchically from imaging data and, hence, making them suitable to tackle the complexity of virus integration detection.

In this paper, we propose a deep convolutional neural network (CNN) based approach to detect virus integration sites

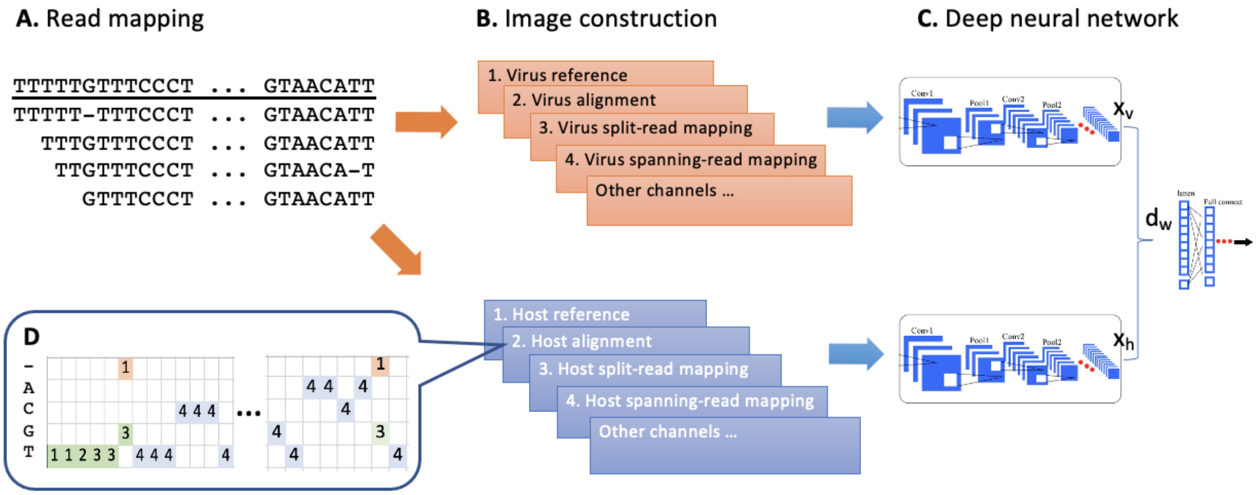

**Fig. 1** Proposed Deep CNN based approach for detecting Virus integration sites in tumor genomes

in tumor genomes to improve NGS-based detection of virus integration. To the best of our knowledge, this paper represents the first effort to use deep neural network to characterize virus integrations from NGS data that challenge current methods for accurate detection. This will not only help cancer researchers to investigate the etiologic association of viruses with cancer but will also produce a state-of-the-art tool for the scientific community

The rest of the paper is organized as follows: Section II discusses related work; Section III describes the proposed approach followed by conclusion and future works in Section IV.

## II. RELATED WORK

Recent advances in artificial intelligence have made deep CNN the primary model for virtually every image related problem. Deep CNN, as a class of deep learning algorithms, is composed of stacks of processing layers, allowing it to learn complex features hierarchically from imaging data. The CNN networks typically utilize multiple convolution-pooling modules that are built on top of each other to learn from input-output pairs. Input images will be fed into the first convolution-pooling module of the CNN networks to perform a series of convolution operations followed by rectified linear activation (ReLU) and a max-pooling to extract linear features from the input image. The output of the final convolution-pooling module of the CNNs will be fed to a fully connected module, which will be trained to perform predictions. Model training is fully automated, thereby removing the need of feature engineering and human intervention. This makes deep CNNs suitable for handling the complexity of virus insertion site characterization, and thus, effectively avoid the limitations of today's tools.

There are some works using Deep CNNs for virus integration sites detection. DeepHINT [20] uses CNN with an attention module to learn the contextual sequence features of HIV integration for the prediction of HIV integration sites from primary DNA sequence. Besides HIV, similar frameworks have been utilized to study local genomic environment of integration sites of other types of viruses, e.g., HBV [21] and HPV [22]. Deep CNN have also gained significant traction for calling genomic variants from NGS data. DeepVariant, a tool that pioneered this approach, transforms aligned reads characteristic of candidate variants into images and then applies CNNs to call small variants [23]. Another tool, NeuSomatic, detects somatic variants instead of germline variants using CNN [24]. Deep CNNs have achieved good performance for the detection of complex structural variations (SVs) as well [25]. Although with these successful applications and great potential to transform NGS-based methods, deep CNNs have not been utilized to detect virus integrations from NGS data.

## III. PROPOSED APPROACH

Figure 1 shows an overview of our proposed Deep CNN approach for detecting virus integration sites in tumor genomes. Since each virus integration involves two distinct breakpoints: one in host genome and another in virus genome, the aligned reads centered on a candidate virus integration site can be used to construct an image pair accordingly. With image pairs as input, we propose to use Siamese CNN for virus integration detection. Figure 1 illustrates the proposed model, in which the features of the two images will be extracted using twin deep CNN networks (Figure 1C). Then, the features extracted from the image pairs will be combined and provided to a fully connected layer for virus integration prediction.

The input (Figure 1A) to our model are genomic regions that potentially harbor virus insertion events, together with the NGS reads aligned to them. Here, the sequence above the solid line represents the host reference genome and the sequences under are reads mapped to it. Two images are constructed from the aligned reads, one for reads from host genome and another for reads from virus genome. We used one-hot encoded images to represent viral integration sites, because they lead to a simplified CNN architecture with lower computational complexity for both training and prediction. One-hot encoded images are 3-dimensional images with many channels, each recording certain alignment signal (Panel B in Figure 1). The Panel D in Figure 1 illustrates how to construct a channel in a one-hot encoded image: for the four reads mapped to the reference in the alignment file in Panel 1A, the frequency of A/C/G/T/- characters in each column is recorded in the corresponding column in the channel in the Panel D. Here, the character '**-**' represents a gap, which is added to the reference sequence whenever there is an insertion in reads.

Both the twin CNN networks in the proposed model will utilize multiple convolution-pooling modules that are built on top of each other to learn from input-output pairs (Panel C in Figure 1). Input images will be fed into the first convolution-pooling module of the twin CNN networks to perform a series of convolution operations followed by rectified linear activation (ReLU) and a max-pooling to extract linear features from the input image. The output of the final convolution-pooling module of the twin CNNs will be combined and then fed to a fully connected module, which will be trained to perform predictions using a softmax classifier.

## IV. CONCLUSION AND FUTURE WORK

We have presented our proposed Deep CNN based approach in this paper for detecting the virus integration sites in tumor genomes. This approach has potential to improve the detection accuracy of virus integration sites which will further help in understanding the etiologic association of viruses with cancer and other diseases. We are applying the proposed approach in the human tumor samples of various publicly available datasets. We are also interested in applying the technique to real clinical samples for further validation. We will provide experimental results in the upcoming full paper.

## V. ACKNOWLEDGMENT

This work was supported by American cancer society Diversity in cancer research institutional development (ACS-DICRIDG) grant award number ACS DICRIDG-21-071-DICRIDGT.

## REFERENCES

[1] Parkin DM. The global health burden of infection-associated cancers in the year 2002. Int J Cancer. 2006;118(12):3030-3044.

[2] zur Hausen H. Viruses in human cancers. Science. 1991;254(5035):1167-1173.

[3] Human papillomavirus (HPV) and cervical cancer | WHO. Accessed January 16, 2021. https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer

[4] Nguyen N-PD, Deshpande V, Luebeck J, Mischel PS, Bafna V. ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res*. 2018;46(7):3309-3325.

[5] Jiang Z, Jhunjhunwala S, Liu J, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res*. 2012;22(4):593-601.

[6] Akagi K, Li J, Broutian TR, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res*. 2014;24(2):185-199

[7] 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

[8] Wang Q, Jia P, Zhao Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med*. 2015;7(1):2.

[9] 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. *Bioinforma Oxf Engl*. 2013;29(2):266-267.

[10] Chen X, Kost J, Li D. Comprehensive comparative analysis of methods and software for identifying viral integrations. *Brief Bioinform*.

[11] Zapatka M, Borozan I, Brewer DS, et al. The landscape of viral associations in human cancers. *Nat Genet*. 2020;52(3):320-330.

[12] Afzal S, Wilkening S, von Kalle C, Schmidt M, Fronza R. GENE-IS: Time-Efficient and Accurate Analysis of Viral Integration Events in Large-Scale Gene Therapy Data. *Mol Ther Nucleic Acids*. 2017;6:133-139.

[13] Xia Y, Liu Y, Deng M, Xi R. Detecting virus integration sites based on multiple related sequencing data by VirTect. *BMC Med Genomics*. 2019;12(Suppl 1):19.

[14] Baheti S, Tang X, O'Brien DR, et al. HGT-ID: an efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data. *BMC Bioinformatics*. 2018;19(1):271.

[15] VIRUSBreakend: Viral Integration Recognition Using Single Breakends |bioRxiv. Accessed January 26, 2021. https://www.biorxiv.org/content/10.1101/2020.12.09.418731v1

[16] Rajaby R, Zhou Y, Meng Y, et al. SurVirus: a repeat-aware virus integration caller. *Nucleic Acids Res*. 2021;(gkaa1237).

[17] Zeng X, Zhao L, Shen C, Zhou Y, Li G, Sung W-K. HIVID2: an accurate tool to detect virus integrations in the host genome. *Bioinformatics*. 2021;(btab031).

[18] Bratman SV, Bruce JP, O'Sullivan B, et al. Human Papillomavirus Genotype Association With Survival in Head and Neck Squamous Cell Carcinoma. JAMA Oncol. 2016;2(6):823-826.

[19] Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., & Crossa, J. (2021). A review of deep learning applications for genomic selection. BMC genomics, 22(1), 1-23.

[20] Hu H, Xiao A, Zhang S, et al. DeepHINT: understanding HIV-1 integration via deep learning with attention. Bioinformatics. 2019;35(10):1660-1667

[21] Tian R, Zhou P, Li M, et al. DeepHPV: a deep learning model to predict human papillomavirus integration sites. *Brief Bioinform*. 2020;(bbaa242).

[22] Wu C, Guo X, Li M, et al. DeepHBV: A deep learning model to predict hepatitis B virus (HBV) integration sites. *bioRxiv*. Published online January 8, 2021:2021.01.08.425855.

[23] Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983-987.

[24] Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep convolutional neural networks for accurate somatic mutation detection. *Nat Commun*. 2019;10. doi:10.1038/s41467-019-09027-x

[25] Cai L, Wu Y, Gao J. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. *BMC Bioinformatics*. 2019;20(1):665.