viroCapt: A Bioinformatics Pipeline for Identifying Viral Insertion in Human Host Genome

Maxime Wack, David Veyer, Camille Peneau, Sonia Lameiras, William Digan, Alain Nicolas, Jessica Zucman-Rossi, Sandrine Imbeaud, Anita Burgun, Hélène Péré, et al.

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220602

viroCapt: A Bioinformatics Pipeline for Identifying Viral Insertion in Human Host Genome

Maxime WACK[a,b,1], David VEYER[c,d], Camille PENEAU[e,f], Sonia LAMEIRAS[g], William DIGAN[a,b], Alain NICOLAS[g], Jessica ZUCMAN-ROSSI[e,f], Sandrine IMBEAUD[f], Anita BURGUN[a,b,h], Hélène PÉRÉ[c,d] and Bastien RANCE[a,b,h]

[a] *INSERM, UMRS 1138, Centre de Recherche des Cordeliers. Université de Paris, France*
[b] *Département d'Informatique Médicale, HEGP, AP-HP, France*
[c] *INSERM U970, PARCC, HEGP, Faculté de Médecine, Université de Paris, France*
[d] *Service de Microbiologie, HEGP, AP-HP, France*
[e] *Centre de Recherche des Cordeliers, INSERM, Université de Paris, France*
[f] *Functional Genomics of Solid Tumors laboratory, Labex OncoImmunology, Paris, France*
[g] *ICGex NGS platform, Institut Curie, PSL Research University, Paris, France*
[h] *Faculté de Médecine, Université de Paris, France*

Abstract. *Introduction*. The implication of viruses in human cancers, as well as the emergence of next generation sequencing has permitted to investigate further their role and pathophysiology in the development of this disease. One such mechanism is the integration of portions of viral genomes in the human genome, as well as the specific action of viral oncogenes.inding integration sites and preserved oncogenes is still relying on heavy manual intervention. *Methods*. We developed an analysis and interpretation pipeline to determine viral insertions. Using data from directed viral capture, the pipeline conducts a crude genotyping phase to select reference viral genomes, identifies chimeric reads, extracts the putative human sequences to locate in the human reference genome, scores and ranks candidate junctions, and exports tabular and visual results. *Results*. We leverage common bioinformatics tools (bowtie2, samtools, blat), and a dedicated filtering and ranking algorithm, implemented in R, to infer candidate junctions and insertions. Static results (tables, figures) are produced, as well as an interactive interpretation tool developed as a shiny web app. *Discussion*. We validated this pipeline against published results of HPV, HBV, and AAV2 insertions and show good information retrieval.

Keywords. bioinformatics, HPV, HBV, AAV2, virus, integration, cancer

[1] Corresponding Author, Maxime Wack, Département d'Informatique Médicale, HEGP, AP-HP, Paris, France; E-mail: maxime.wack@aphp.fr

1. Introduction

High-throughput data provided by next generation sequencing (NGS) have revolutionized the medical care of cancer patients. Together with treatment responses, the fine characterization of mutations in cancer cells has permitted the realization of precision medicine, and can now lead to personalized treatments in cancers. [1]

Recently, virus-induced cancers have received a great deal of attention. [2] Many viruses are implicated in the development of cancers, including HBV for hepatocellular carcinoma, [3] HPV for cervix, anal, and oropharyngeal squamous cell carcinoma, [4] EBV for nasopharynx carcinoma and Burkitt and Hodgkin lymphomaes, HTLV-1 for adult T-cell lymphoma, HIV and HHV8 for kaposi sarcoma.

The exact physiological mechanisms explaining the carcinogenic role of these viruses are not still clearly known, but recent advances in viral capture techniques followed by next generation sequencing have helped uncovering mechanisms underlying this carcinogenic role. [5,6]

Due to its high-throughput nature, NGS produces large amounts of data for which the bottleneck now lies in the interpretation of the results.We introduce `viroCapt`, a software package designed to help researcher in the analysis and interpretation of such data.

`viroCapt` manages an analytic and interpretation pipeline, as well as a visualization tool to assist in reading and interpreting the results, with a ranking method using quality and interpretability criteria.

2. Methods

The pipeline uses short read sequencing data from NGS viral capture, leveraging common bioinformatics tools (bowtie2 for viral alignment, [7] samtools for janitorial purposes, [8] and blat for human alignment [9]), and public resources (viral and human reference genomes) to produce candidate insertion sites in the human genome and splicing sites on the viral genome. Those results are ranked using custom-developed filters to assess their quality and interpretability.

The pipeline goes through the following steps (Figure 1): Quality Control, crude genotyping, local alignment to the candidate viruses, and alignment of the partially mapped reads to human, in a method described as Strategy A-II in Chen *et al.* 2019 [10]. Furthermore, we apply dedicated filters to tag specific human alignment (H), concordance of results between multiple reads (C), and presence of breakpoints on the same chromosome (T).

After the execution of the pipeline, the results can be browsed through a web app (Figure 2) showing the sequencing profile with the candidate insertions as an overlay, as well as sortable and filterable tabular view of all the results. Multiple options and filters (chromosome, quality, number of reads, length of the human sequence found) can be used to explore and narrow the results.

We validated the pipeline output on HeLa cell samples with an HPV integration and samples from hepato-cellular carcinomas with HBV and AAV2 integrations (data available upon request). [11,12,13]

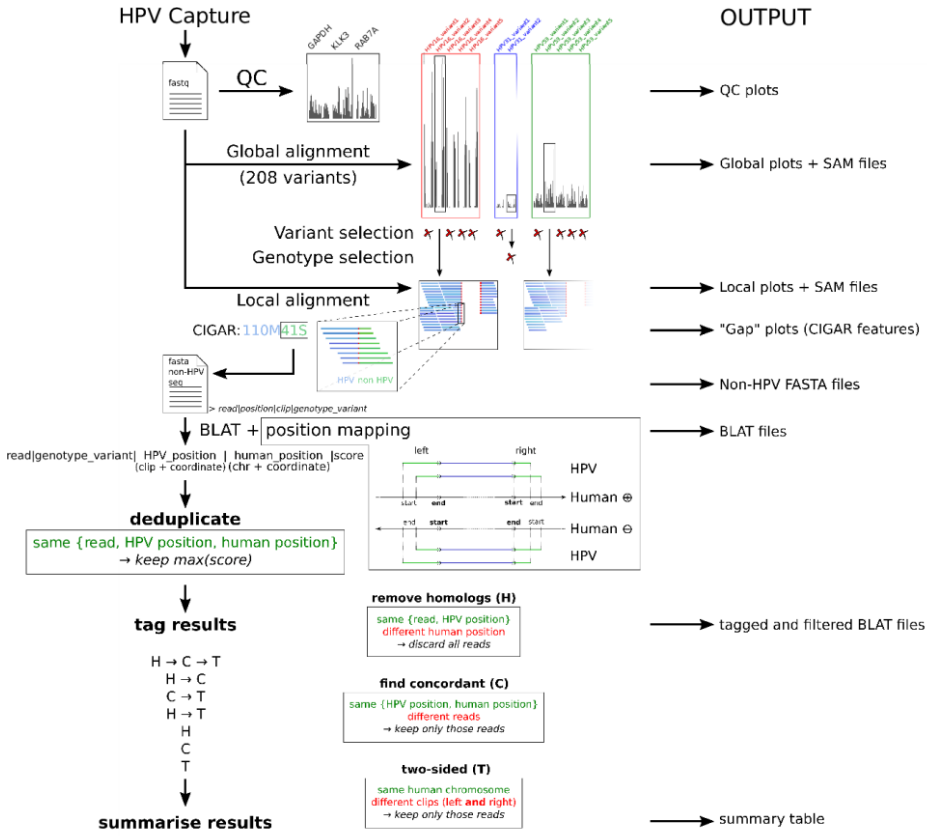

Figure 1. Overall organisation of the pipeline.

3. Results

The pipeline is implemented as a makefile orchestrating multiple bioinformatics tools, as well as R functions to process the intermediary data. Everything is packaged as an R package containing the analysis functions and the visualization tool as a shiny web application. The makefile, reference sequences for HPV and the HeLa example dataset are installed along the R package. The makefile allows for batch processing of multiple samples, and multithreading is enabled for bowtie2, samtools, blat, and most of the R logic.

Viral capture data are usually small. For example, the HeLa sample dataset is 13Mb per paired FASTQ file, and executes in 12 minutes using 4 threads on a standard laptop computer (Intel i7 CPU, 16GB of RAM).

The pipeline code can be obtained from https://github.com/maximewack/viroCapt. A demo instance showing the HeLa results can be found here: https://shiny.maximewack.com/viroCapt_MIE

The validation on samples with HPV, HBV, and AAV2 integrations yielded concordant, well-ranked results (Table 1), showing good retrieval performances of the algorithm and assorted filters.

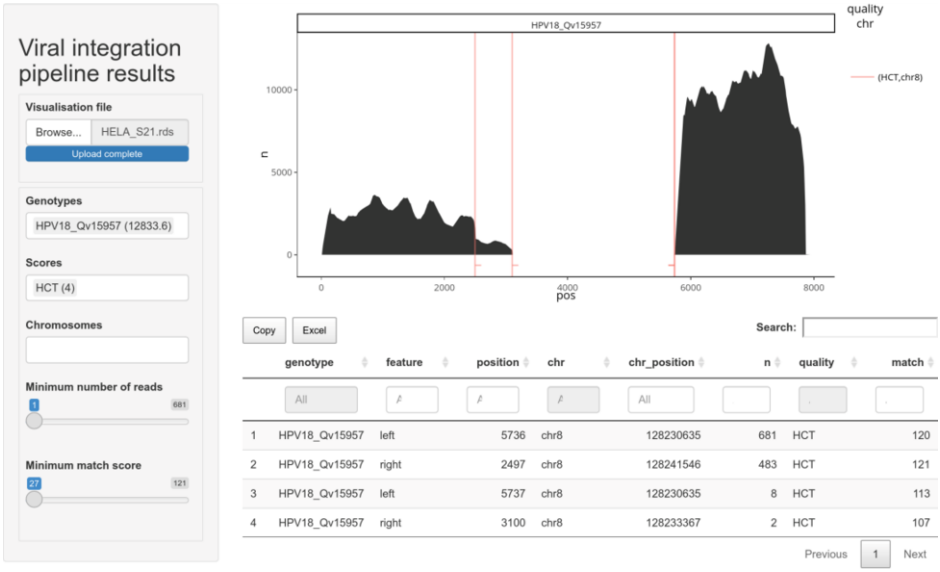

Figure 2. Screenshot of the interactive analysis tool displaying the HeLa sample results.

Table 1. Comparison of viral integration identified in viroCapt and the references

		viroCapt		**references**		
virus	**sample**	virus	human (chr:position)	virus	human	**rank**
HBV	1	1671 - 2844	19:30291247 - 30291291	1671 - 2849	30291247 - 30291287	1 - 2
		1827 - 2846	5:1295275 - 1295209	1830 - 2843	1295275 - 1295207	23 - 24
		1794 -	12:79363501 -	1789 -	79363502 -	25
		192 - 2224	6:49526654 - 49411818	189 - 2226	49526654 - 49411817	26 - 726
	2	1738 - 1824	14:90044537 - 90132716	1736 - 1828	90044536 - 90132717	1 - 3
		1496 - 587	5:1295211 - 1026640	1496 - 587	1295211 - 1026641	4 - 7
		2985 - 1864	19:30300745 - 30301945	2981 - 1853	30300749 - 30301935	14 - 15
AAV	1	4390 - 4632	5:1295307 - 1295291	4390 - 4597	1295308 - 1295291	1 - 3
	2	4571 - 4270	3:172224024 - 172224026	4270 - 4571	17222402 7 - 172224026	1 - 2
	3	*no integration*		*no integration*		
	4	*no integration*		*no integration*		
	5	4597 - 4386	3:172302190 - 172224151	4388 - 4597	172224150 - 172302191	1 - 2
HPV	HeLa	2497 - 5736	8:128241546 - 128230635	2497 - 5736	128241494 - 128230632	1 - 2
		3100 -	8:128233367 -	3088 -	128233367 -	4

4. Discussion

We compared the results obtained using our analytic pipeline to those obtained by published results on HeLa cells [11] and liver cancer samples with HPV or AAV2 integrations, [12] and most results are confirmed in the top results.

While the carcinogenic mechanisms specific to each viral insertion are still largely unknown, high throughput approaches allow the genotyping and identification of genome insertions. We already used this pipeline in clinical research to elaborate on those mechanisms [14,15].

This is not the first pipeline to implement a viral insertion finding logic, some implementing a similar strategy [16,17], but this is the first to include such a dedicated analysis tool.

The addition of an interactive web tool for the exploration of results and assisting in interpretability is a valuable asset for researchers and clinicians alike to make use of these data in an efficient manner.

References

[1] Jackson SE, Chester JD. Personalised cancer medicine. International Journal of Cancer. 2015 Jul;137(2):262-6.
[2] Parkin DM. The global health burden of infection-associated cancers in the year 2002. International Journal of Cancer. 2006 Jun;118(12):3030-44.
[3] Levrero M, Zucman-Rossi J. Mechanisms of HBV-induced hepatocellular carcinoma. Journal of Hepatology. 2016 Apr;64(1 Suppl):S84-S101.
[4] Munger K, Baldwin A, Edwards KM, Hayakawa H, Nguyen CL, Owens M, et al. Mechanisms of human papillomavirus-induced oncogenesis. Journal of Virology. 2004 Nov;78(21):11451-60.
[5] Duncavage EJ, Magrini V, Becker N, Armstrong JR, Demeter RT, Wylie T, et al. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. The Journal of molecular diagnostics: JMD. 2011 May;13(3):325-33.
[6] Holmes A, Lameiras S, Jeannot E, Marie Y, Castera L, Sastre-Garau X, et al. Mechanistic signatures of HPV insertions in cervical carcinomas. NPJ Genomic Medicine. 2016;1:16004. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5685317/.
[7] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar;9(4):357-9.
[8] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078-9.
[9] Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002 Apr;12(4):656-64.
[10] Chen X, Kost J, Li D. Comprehensive comparative analysis of methods and software for identifying viral integrations. Briefings in Bioinformatics. 2019 Nov;20(6):2088-97. Available from: https://doi.org/10.1093/bib/bby070.
[11] Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, et al. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature. 2013 Aug;500(7461):207-11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3740412/.
[12] Peneau C, Imbeaud S, La Bella T, Hirsch TZ, Caruso S, Calderaro J, et al. Hepatitis B virus integrations promote local and distant oncogenic driver alterations in hepatocellular carcinoma. Gut. 2021 Feb.
[13] La Bella T, Imbeaud S, Peneau C, Mami I, Datta S, Bayard Q, et al. Adeno-associated virus in the liver: natural history and consequences in tumour development. Gut. 2020 Apr;69(4):737-47.
[14] Morel A, Neuzillet C, Wack M, Lameiras S, Vacher S, Deloger M, et al. Mechanistic Signatures of Human Papillomavirus Insertions in Anal Squamous Cell Carcinomas. Cancers. 2019 Nov;11(12).[15]
[15] Pere H, Vernet R, Pernot S, Pavie J, Robillard N, Puech J, et al. Episomal HPV16 responsible for aggressive and deadly metastatic anal squamous cell carcinoma evidenced in peripheral blood. Scientific Reports. 2021 Feb;11(1):4633.
[16] 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 May;8(5):e64465. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064465.
[17] 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 Jan;29(2):266-7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/