Long read sequencing and *de novo* assembly of Hepatitis B Virus identifies 5mCpG in CpG islands

Chloe Goldsmith[1,2*], Damien Cohen[2], Anaelle Duboi[2], Maria-Guadaloupe Martinez[2], Kilian Petitjean[3], Anne Corlu[3], Hector Hernandez-Vargas[4], Isabelle Chemin[2*]

1: TGFB and immune evasion, Lyon Cancer Research Center (CRCL), Lyon France

2: Hepatocarcinogenesis and Viral infection, Lyon Cancer Research Center (CRCL), Lyon France

3: Nutrition, Metabolism and Cancer (NuMeCan), Université de Rennes 1, Rennes France

4: Centre Leon Berard (CLB), Lyon Cancer Research Center (CRCL), Lyon France

* Co-corresponding Authors

Chloe.Goldsmith@lyon.unicancer.fr

Lyon Cancer Research Center (CRCL), INSERM U1052

Centre Léon Bérard, 4th floor Cheney B,

28 Rue Laennec, 69373 Lyon Cedex 08, France

Abstract

Methylation of viral DNA in a CpG context (5mCpG) can alter the expression patterns of viral genes related to infection and cellular transformation. Moreover, it may also provide clues to why certain infections are cleared, or persist with or without progression to cancer. The detection 5mCpG often requires techniques that damage DNA or introduce bias through a myriad of limitations. Therefore, we developed a method for the detection of 5mCpG on the HBV genome that does not rely on bisulfite conversion or PCR. We used cas9 guided RNPs to specifically target and enrich in HBV DNA from infected PHH, prior to sequencing with nanopores. This method is a novel approach that enables the enrichment of viral DNA in a mixture of nucleic acid material from different species. Moreover, using the developed technique, we have provided the first *de novo* assembly of naive HBV DNA, as well as the first landscape of 5mCpG from naive HBV sequences.

Introduction

HBV contains a relaxed circular partially double-stranded DNA genome (3.2kb) (Summers, O'Connell, and Millman 1975). The HBV replication cycle is mechanistically distinct compared to other viruses. One characteristic of HBV DNA replication is the protein-primed reverse transcription of an RNA intermediate called pregenomic RNA (pgRNA) that occurs after packaging in the viral nucleocapsid (Wang and Seeger 1993). Another obvious difference is that the integration of viral genomic DNA into host cellular chromosomes is not an obligatory step for HBV replication. Instead, upon the entry into hepatocytes, viral genomic DNA in the nucleocapsid is partially double stranded DNA, also called relaxed circular DNA (rcDNA); it is transported to the nucleus and converted into covalently closed circular DNA (cccDNA), which is organized as a minichromosome, and serves as a template for the transcription of viral RNAs.

The overall degree of viral replication has also been strongly linked to carcinogenesis. Therefore, understanding factors that regulate HBV replication may provide insights into preventing HCC. In this regard, recent studies have identified epigenetic modifications of HBV DNA, including methylated cytosines in CpG context (5mCpG) as a novel mechanism for the control of viral gene expression (Pollicino et al. 2006). However, the currently used tools to detect modified bases have a number of limitations.

Bisulfite sequencing has been the method of choice for the detection of 5mCpG for over a decade. This technique converts unmodified cytosines into uracil, and then the 5mCpG

levels are deduced by difference using PCR and/or sequencing (Li and Tollefsbol 2011). This techniques leads to extensive DNA damage and introduces bias through incomplete conversion. Moreover, it is not able to distinguish the difference between 5mCpG and other modified bases that can occur in the same location (eg 5hmCpG). Thus, whilst this technique has been incredible useful, there is a demand for the development of more direct measures of 5mCpG.

Nanopore sequencing is a unique, scalable technology that enables direct, real-time analysis of long DNA or RNA fragments (Madoui et al. 2015). It works by monitoring changes to an electrical current as nucleic acids are passed through a protein nanopore. The resulting signal is decoded to provide the specific DNA or RNA sequence. Moreover, this technology allows for the simultaneous detection of the nucleotide sequence as well as DNA and RNA base modifications on naive DNA (Jain et al. 2016); hence, removing introduced bias from sodium bisulfite treatment and PCR amplification. However, there is still a need to enrich in target loci or species prior to sequencing. Traditionally this would be done by PCR amplification, however, doing so would lose modified bases like 5mCpG, thus, there is a demand for the development of enrichment techniques that do not degrade DNA or result in a loss of target bases in order to fully take advantage of Nanopores ability to delineate 5mCpG levels.

The main aim of this work was to develop a novel method for the detection of modified bases in the HBV genome and determine the methylation landscape of HBV *in vitro*. We present here our work determining the HBV 5mCpG levels in infection models based, for the first time, on directly sequenced naive HBV DNA. Moreover, these methods represents a valuable and highly novel tool for the detection of modified bases on viral genomes.

Results and Discussion

Enrichment

We utilized Cas9 guided Ribonucleic proteins (RNPs) to linearise and enrich in HBV DNA from the total DNA extracted from HBV infected Primary Human Hepatocytes (PHH). This method has been described previously to enrich in target loci in the human genome (Gilpatrick et al. 2020), however, to our knowledge, we are the first to utilize this approach to sequence viral or circular DNA (Figure 1). Since the HBV genome was enriched without PCR amplification, we were able to sequence naive DNA and take full

advantage of the ability of Nanopores to detect modified bases in this setting. Briefly, starting material consisted of 3µg of total DNA extracted from HBV infected PHH. Available DNA ends were blocked by dephosphorylation with calf intestinal phosphotase, this step is integral to prevent the nuclear DNA from being available for the ligation of adapters in later steps. We then used 2 sgRNAs to target the HBV genome, one for the positive strand and an additional guide for the negative strand (Figure 1). This was important, since the RNP complex remains attached to the strand where it makes the cut, making the other strand available for the ligation of adapters; hence to ensure the sequencing of both the positive and negative strands, it was essential to use sgRNAs in this manner. After the adapters and motor proteins were ligated to the newly available DNA ends, the libraries were loaded onto the MinION devise and DNA sequencing occurred in a 5’ to 3’ directionality on a MinION R9.4.1 flow cell.

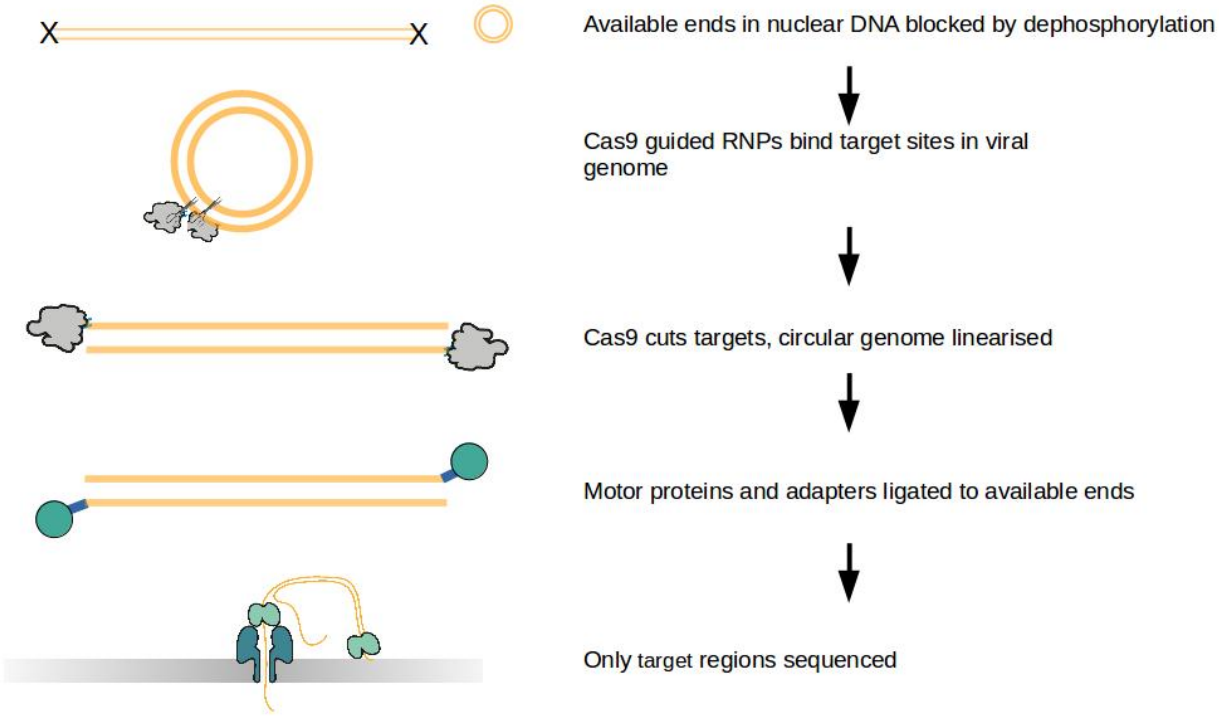

Figure 1. Overview of cas9 targeted sequencing protocol adapted for circular viral genomes. Briefly, all available DNA ends were dephosphorylated prior to liberation of target sites by cutting with cas9 guided RNPs. The circular viral genome was then linearized and prepared for the ligation of adapters and motor proteins. Libraries were then loaded onto a MinION to be sequenced with Nanopores.

Yield and coverage

Starting with 3µg of DNA extracted from HBV infected PHH, we utilized our cas9 enrichment protocol to linearise and sequence the HBV genome. The total yield ranged

between 150-200K reads collecting in the range of 1-2 GB. Raw reads were basecalled with Guppy (Version 3.3.4), a draft assembly was generated with Canu and polished with Medaka. The resulting consensus sequence was used to align the basecalled reads and calculate coverage (Figure 2) and enrichment . We obtained a clear enrichment of HBV with up to 15000x coverage at certain loci on the positive strand of HBV. We calculated approximately 5% of all reads were on target, which is the clearest enrichment using this technique that has been observed, to our knowledge. Interestingly, a gap in the coverage of the positive strand was observed. Since HBV exists as rcDNA, it is not surprising that we identified this gap in the coverage of the positive strand of this sequence. Moreover, there was a range in the length of the gap between different reads, with a clear tapering off of read length (Figure 2A). This is likely due to the variability in the stage of transcription of the HBV pgRNA occurring in the viral nucleocapsids. These PHH were highly infected with over 500 copies of total HBV detected per cell (data not shown). Thus, it was not surprising that the gap in coverage was found, since it was likely due to the sequencing of viral intermediates.

This same phenomenon is likely the reason for the lower coverage on the negative strand (Figure 2B). While we still obtained almost 200x, that was a far cry from the 15000x observed for the positive strand. However, this is likely due to a combination of the efficacy of the sgRNA guide for the negative strand as well as the high concentration of replication intermediates.

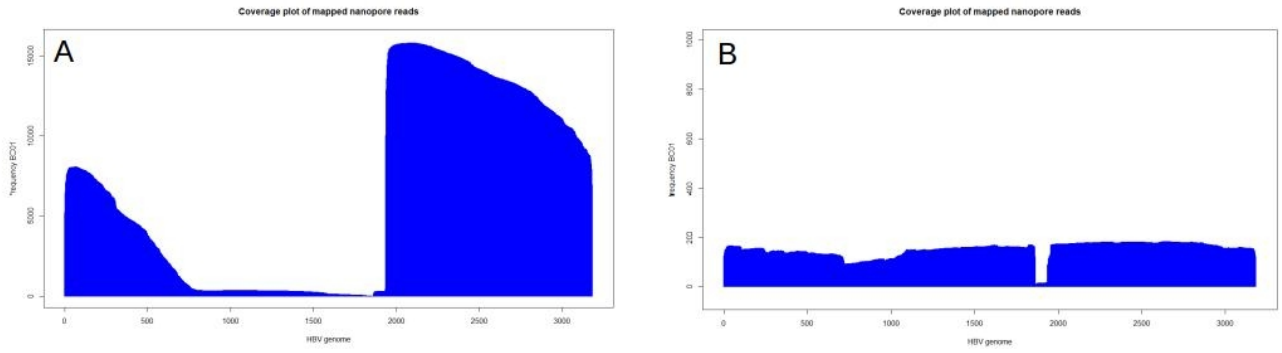

Figure 2. Coverage of HBV genome after enrichment via cas9 sequencing technique (Left = positive strand, Right = negative strand).

Nanopore sequencing detects HBV 5mCpG

We then sought to determine the validity of using Nanopores to detect modified bases on the HBV genome. To do so, we prepared a negative control by nested PCR amplification, and a positive control by using the amplified HBV sample, and methylating the CpG sites using methyltransferase enzyme M.Sssl. Beta values for each CpG site on the positive and negative strands were calculated and plotted in Figure 3 (A-B). Because nanopore sequencing directly evaluates methylation patterns on naive DNA strands, we are able to observe long-range methylation information on each DNA molecule (Figure 3C-D). Single reads were also converted for visualisation in IGV, previously described (Gilpatrick et al. 2020). As anticipated, very low levels of methylation were observed in the negative control with some residual methylation detected. Interestingly, we identified some single reads that were fully methylated in the negative control, likely a result of the presence of residual un-amplified DNA; thus, we determined that the low levels of methylation observed were attributed to this contaminating starting material. In the positive control we detected high levels of DNA methylation. However, there was high variability in the levels of 5mCpG, which, was attributed to the efficiency of the Methyltransferase in the preperation step of the fully methylated control. Moreover, we identified a number of reads that were not fully methylated that was likely contributing to the lower average levels of 5mCpG at certain loci. Nevertheless, the clear extremes observed in the different controls indicate the efficacy of Nanopore technology for the detection of 5mCpG on HBV DNA. We were therefore highly confident in this tool for the detection of 5mCpG levels on the HBV genome.

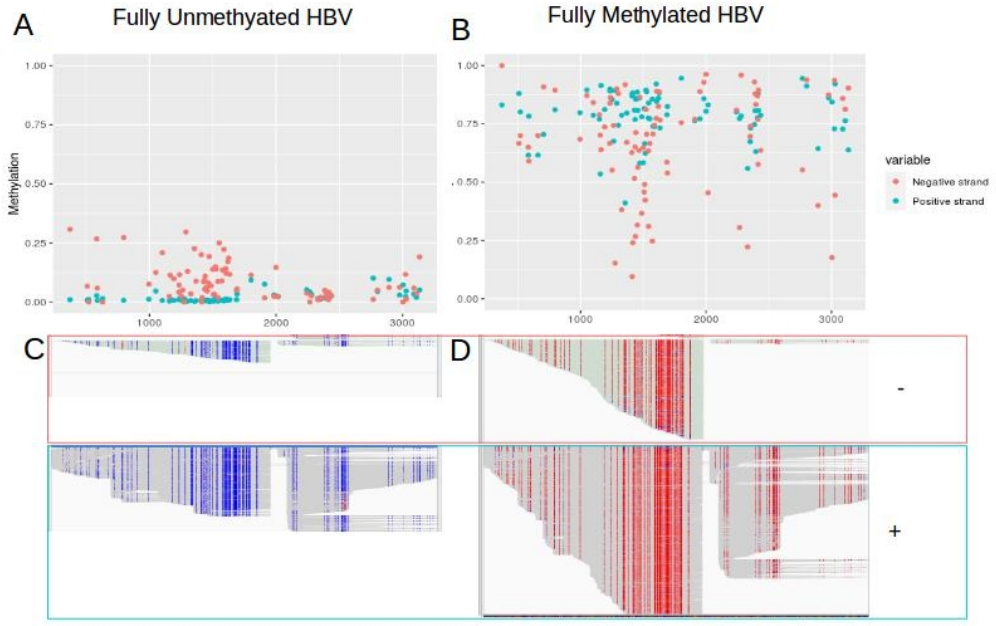

Figure 3. 5mCpG detected in HBV controls. Negative control of fully unmethylated DNA was prepared by nested PCR amplification of HBV from Infected PHH. CpG sites in HBV were Fully methylated to prepare a positive control by treatment with CpG methyltransferase enzyme (M.SssI). A-B: Strand specific 5mCpG levels for CpG sites with >10x in FM and FU controls (pink = Negative strand, green = Positive strand). C-D: Single molecule methylation of all reads aligning to HBV genome converted for visualisation in IGV (Blue = Unmethylted CpG site, Red = Methylated CpG site).

Absence of DNA methylation observed in HBV virions

While we had determined the 5mCpG levels in positive and negative controls, we next sought to determine the 5mCpG levels in an expected biologically negative control, HBV Virions. HBV DNA is reverse transcribed after being packaged in the viral capsid and likely out of touch with DNA methyltransferase enzymes (Wang and Seeger 1993). Thus, we expected that there would be an absence of 5mCpG in these samples. Interestingly, there was some very low levels of methylation observed in CpG island 1 on the positive strand (Figure 4 A-C). These levels were not higher than those detected in the amplified HBV control, and as such, were likely a result of contaminating cccDNA from dead cells also collected during the HBV viral particle purification process; alternatively, these positive reads could be due to methylation calling errors. However, the levels were extremely low and, therefore, continued to increase our confidence in Nanopore sequencing to accurately detect 5mCpG in the HBV genome.

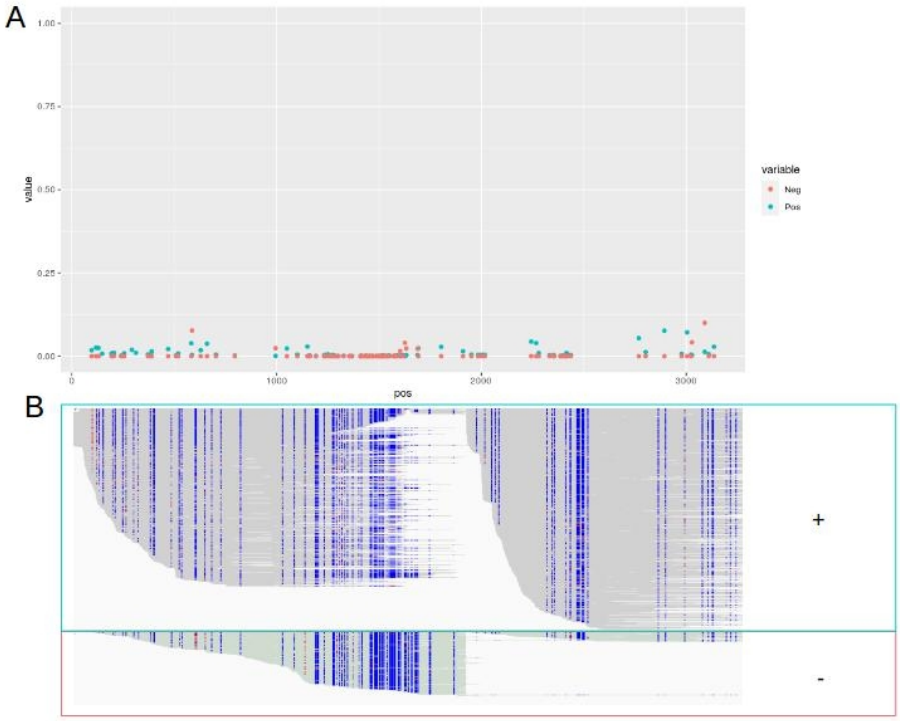

Figure 4. Single molecule methylation of HBV DNA extracted from HBV Virions (Red = Methylated, Blue = Unmethylated CpG sites).A: average 5mC for each CpG site (pink = Negative strand, green = Positive strand: B-C = Single molecule visualisation for individual DNA reads aligned to HBV genome, B=Negative strand: C = Positive strand.

Basal levels of HBV 5mCpG in infected PHH

We next sought to determine the basal levels of 5mCpG in HBV infection models *in vitro.* To do so we employed a PHH infection model where we exposed the PHH to HBV to achieve a 'natural infection'. Infection efficacy was determined by Expression of Hbs and Hbe in the supernatant by ELISA, as well as total HBV copies per cell by qPCR (data not shown). After sequencing and downstream analysis, we could ascertain the 5mCpG levels in the HBV genome.

Overall, low levels of 5mCpG were detected in the HBV genome (Figure 5). The few regions where 5mCpG was observed include in the CpG Islands. A clear difference between in methylation levels was detected in the different CpG islands, with CpG island 1 having the highest and most variable levels and CpG island 3 the least (Figure 5). While 5mCpG in host DNA has been thoroughly explored, there is very little known about the role of 5mCpG in the regulation of HBV RNA expression. Early studies were able to show that there were low levels of this modified base in HBV from patients (Fernandez et al. 2009). Furthermore, very recently, high levels of 5mCpG have been detected in integrated HBV DNA in HuH-1 cells using bisulfite converted and amplified HBV DNA (Liu

et al. 2020). Interestingly, these previous studies also identified the highest 5mCpG levels in CpG island 1. However, there had not, to our knowledge, been any investigation into levels in infected PHH or using a technique that employed sequencing of naive DNA. Therefore, the present study represents the first map of HBV 5mCpG without the use of bisulfite conversion or PCR amplification.

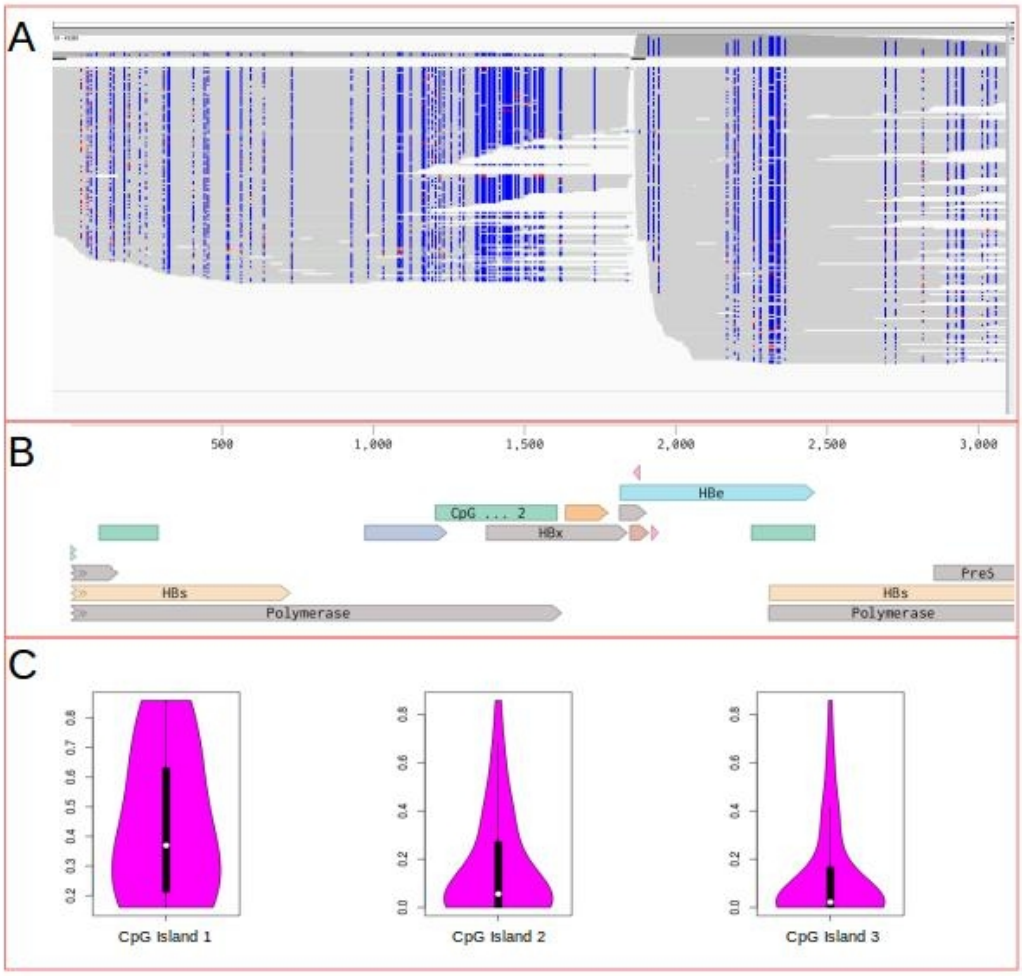

Figure 5. 5mCpG levels in HBV extracted from infected PHH. A: Single molecule visualisation of HBV reads (Red = Methylated, Blue = Unmethylated). B: Aligned sequences to HBV genome generated by de novo assembly (Green = CpG Islands, Pink = sgRNA guides). C: Methylation distribution of HBV from infected PHH in predicted CpG islands.

Conclusions

Methylation of viral DNA in infected cells can alter the expression patterns of viral genes related to infection and cellular transformation (Fernandez et al. 2009) and may also understand why certain infections are cleared, or persist with or without progression to cancer (Mirabello et al. 2012). Furthermore, the clear detection of viral methylation patterns could potentially serve as biomarkers for diseases that are currently lacking,

including occult HBV infection. However, in order to do so, the development of more sensitive high throughput techniques translatable to the clinic, is essential.

We developed a sensitive and high throughput method for the detection of 5mCpG on the HBV genome. This method is a novel approach that achieved a clear enrichment of viral DNA in a mixture of virus and host DNA. Moreover, using the developed technique, we have provided the first *de novo* assembly of naive HBV DNA, as well as the first landscape of 5mCpG from Naive HBV DNA.

Methods

Cultivation of PHH

Primary Human Hepatocytes (PHH) were extracted and maintained as previously described (Ancey et al. 2015).

HBV cultivation and infections

HBV inocula was generated as previously described (Ancey et al. 2015; Lucifora et al. 2011). PHH were naturally infected with HBV genotype D for 24h (MOI 500). A stable infection was achieved after 3 days, cells were maintained for up to 14 days. Infection efficacy was determined by quantification of Hbs and Hbe concentration in supernatant by ELISA and calculation of HBV copies/ nuclei by qPCR as previously described (Ancey et al. 2015).

DNA extraction

DNA was extracted using the MasterPure™ Complete DNA and RNA Purification Kit (Epicentre) according to manufacturers instructions.

Fully unmethylated and fully methylated controls

HBV DNA was amplified by nested PCR as previously described to prepare a negative (fully unmethylated) control. After amplification, a positive control for methylation (fully methylated) was prepared by methylating CpG dinucleotides; by incubating 1µg of DNA with S-Adenosyl methionine (SAM) (32µM) with CpG Methyltransferase (M.SssI) (4-25 units) (New England BioLabs) at 37°C for 1h before heating to 65°C for 20mins.

Nanopore library prep and sequencing

DNA (3µg) from each sample or control was enriched in HBV and linearized using cas9 guides RNPs (sequences and details available upon request). Samples were barcoded and multiplexed using the Nanopore Ligation Sequencing kit (SQK-LSK109) and Native barcode expansion kit according to manufacturers instructions (Oxford Nanopore Technology, Oxford UK). Sequencing was conducted with a Minion sequencer on ONT 1D flow cells (FLO-min106) with protein pore R9.4 1D chemistry for 48h. Reads were basecalled with GUPPY (version 3.4.4). *De novo* assembly was performed with canu (Koren et al. 2017) and assembly was polished with Medaka.

Denovo assembly

Basecalled fasQ files were used to assemble the HBV genome with canu (Koren et al. 2017), that was polished using Medaka; a tool to create a consensus sequence from nanopore sequencing data using neural networks applied from a pileup of individual sequencing reads against a draft assembly. Basecalled FastQs were then aligned to the generated consensus sequence using Minimap2 (Li 2018).

Methylation calling

We first determined the methylation status of each CpG site on every read by using the widely used tool, nanopolish (Simpson et al. 2017) used recently by (Gigante et al. 2019). For validation, we also called DNA methylation using novel tool, Medaka. CpG Islands were predicted using MethPrimer.

Converting reads for visualization in IGV

In order to take advantage of the single molecule sequencing with Nanopore, we converted the reads for visualisation in IGV as previously outlined (Gilpatrick et al. 2020)

Data availability

All detailed methods, sequencing data (basecalled fastQs and raw fast5 files) as well as assembled sequences will be made publicly available upon request and publicly available upon acceptance to a peer reviewed journal.

Conflict of interest

Chloe Goldsmith and Hector Hernandez have received travel and accommodation support to attend conferences for Oxford Nanopore Technology.

Acknowledgements

The Authors would like to acknowledge the whole U1052 team for any help along the way.

Funding

This work was supported by grants from “La ligue contre de cancer” and the “Agence nationale de recherche sur le sida et les hépatites virales (ANRS)”.

Author contributions

CG completed almost all experiments, performed analysis and wrote the manuscript: DC and AB completed experiments: GM assisted in concept generation and manuscript preparation: KP assisted in manuscript preparation: AC obtained funding and assisted in manuscript preparation: HH performed analysis and obtained funding: IC assisted in manuscript preparation and obtained funding.

References

Ancey, Pierre-Benoit, Barbara Testoni, Marion Gruffaz, Marie-Pierre Cros, Geoffroy Durand, Florence Le Calvez-Kelm, David Durantel, Zdenko Herceg, and Hector Hernandez-Vargas. 2015. “Genomic Responses to Hepatitis B Virus (HBV) Infection in Primary Human Hepatocytes.” *Oncotarget* 6(42):44877–91.

Fernandez, Agustin F., Cecilia Rosales, Pilar Lopez-Nieva, Osvaldo Graña, Esteban Ballestar, Santiago Ropero, Jesus Espada, Sonia A. Melo, Amaia Lujambio, Mario F. Fraga, Irene Pino, Biola Javierre, Francisco J. Carmona, Francesco Acquadro, Renske D. M. Steenbergen, Peter J. F. Snijders, Chris J. Meijer, Pascal Pineau, Anne Dejean, Belen Lloveras, Gabriel Capella, Josep Quer, Maria Buti, Juan-Ignacio Esteban, Helena Allende, Francisco Rodriguez-Frias, Xavier Castellsague, Janos Minarovits, Jordi Ponce, Daniela Capello, Gianluca Gaidano, Juan Cruz Cigudosa, Gonzalo Gomez-Lopez, David G. Pisano, Alfonso Valencia, Miguel Angel Piris, Francesc X. Bosch, Ellen Cahir-McFarland, Elliott Kieff, and Manel Esteller. 2009. “The Dynamic DNA Methylomes of Double-Stranded DNA Viruses Associated with Human Cancer.” *Genome Research* 19(3):438–51.

Gigante, Scott, Quentin Gouil, Alexis Lucattini, Andrew Keniry, Tamara Beck, Matthew Tinning, Lavinia Gordon, Chris Woodruff, Terence P. Speed, Marnie E. Blewitt, and Matthew E. Ritchie. 2019. “Using Long-Read Sequencing to Detect Imprinted DNA Methylation.” *Nucleic Acids Research* 47(8):e46–e46.

Gilpatrick, Timothy, Isac Lee, James E. Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Bradley Downs, Saraswati Sukumar, Fritz J. Sedlazeck, and Winston Timp. 2020. “Targeted Nanopore Sequencing with Cas9-Guided Adapter Ligation.” *Nature Biotechnology* 38(4):433–38.

Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. “The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community.” *Genome Biology* 17(1):239.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation.” *Genome Research* 27(5):722–36.

Li, Yuanyuan, and Trygve O. Tollefsbol. 2011. “DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis.” *Methods in Molecular Biology (Clifton, N.J.)* 791:11–21.

Liu, Yibin, Jingfei Cheng, Paulina Siejka-Zielińska, Carika Weldon, Hannah Roberts, Maria Lopopolo, Andrea Magri, Valentina D’Arienzo, James M. Harris, Jane A. McKeating, and Chun-Xiao Song. 2020. “Accurate Targeted Long-Read DNA Methylation and Hydroxymethylation Sequencing with TAPS.” *Genome Biology* 21(1):54.

Madoui, Mohammed-Amin, Stefan Engelen, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker, and Jean-Marc Aury. 2015. “Genome Assembly Using Nanopore-Guided Long and Error-Free DNA Reads.” *BMC Genomics* 16:327.

Mirabello, Lisa, Chang Sun, Arpita Ghosh, Ana C. Rodriguez, Mark Schiffman, Nicolas Wentzensen, Allan Hildesheim, Rolando Herrero, Sholom Wacholder, Attila Lorincz, and Robert D. Burk. 2012. “Methylation of Human Papillomavirus Type 16 Genome and Risk of Cervical Precancer in a Costa Rican Population.” *Journal of the National Cancer Institute* 104(7):556–65.

Pollicino, Teresa, Laura Belloni, Giuseppina Raffa, Natalia Pediconi, Giovanni Squadrito, Giovanni Raimondo, and Massimo Levrero. 2006. “Hepatitis B Virus Replication Is Regulated by the Acetylation Status of Hepatitis B Virus CccDNA-Bound H3 and H4 Histones.” *Gastroenterology* 130(3):823–37.

Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. “Detecting DNA Cytosine Methylation Using Nanopore Sequencing.” *Nature Methods* 14(4):407–10.

Summers, J., A. O’Connell, and I. Millman. 1975. “Genome of Hepatitis B Virus: Restriction Enzyme Cleavage and Structure of DNA Extracted from Dane Particles.” *Proceedings of the National Academy of Sciences* 72(11):4597–4601.

Vivekanandan, P., D. Thomas, and M. Torbenson. 2008. “Hepatitis B Viral DNA Is Methylated in Liver Tissues.” *Journal of Viral Hepatitis* 15(2):103–7.

Wang, G. H., and C. Seeger. 1993. “Novel Mechanism for Reverse Transcription in Hepatitis B Viruses.” *Journal of Virology* 67(11):6507–12.