國立臺灣大學生命科學院生態學與演化生物學研究所

碩士論文

Institute of Ecology and Evolutionary Biology
College of Life Science
National Taiwan University

Master Thesis

在單鹼基解析度下探索 B 型肝炎病毒複製錯誤率

Replication Error Rate of Hepatitis B Virus at Single Base Pair Resolution

劉宇恳

Yu Ken Low

指導教授：王弘毅 博士

吳慧琳 博士

Advisors: Hurng-Yi Wang, Ph.D.

Hui-Lin Wu, Ph.D.

中華民國 112 年 7 月

July, 2023

doi:10.6342/NTU202302511

國立臺灣大學碩士學位論文
口試委員會審定書

在單鹼基解析度下探索B型肝炎病毒複製錯誤率

Replication Error Rate of Hepatitis B Virus at Single Base Pair Resolution

本論文係劉宇恳君（學號 R09B44021）在國立臺灣大學生態學與演化生物學研究所完成之碩士學位論文，於民國112年7月17日承下列考試委員審查通過及口試及格，特此證明

口試委員：

臺灣大學醫學院臨床醫學研究所　陳培哲 博士 陳培哲

臺灣大學醫學院臨床醫學研究所　楊宏志 博士 楊宏志

臺灣大學醫學院臨床醫學研究所（指導教授）　吳慧琳 博士 吳慧琳

臺灣大學生態學與演化生物學研究所（指導教授）　王弘毅 博士 王弘毅

所長 楊月鈴 胡哲明(心) 8/4 （簽名）

doi:10.6342/NTU202302511

誌謝

首先非常感謝我的指導教授-王弘毅老師，他盡心盡力地給予指導與資源，讓我可以順利完成這論文。過程中實驗不順利，老師都會不厭其煩跟我一起思考可以改善的地方，並提供想法與建議。課業外，老師還會關心我們的狀況，必要時還會幫忙解決遇到的問題，時不時還請吃飯，帶大家出遊，帶我們釣魚，跟大家談笑風生。還有要感謝老師在這研究遇到瓶頸且毫無進展時，有大量的耐心，不責備，讓我有毅力堅持做完研究。

除了王老師，我的共同指導教授-吳慧琳老師也給了我很多的鼓勵、資源及珍貴的建議。當王老師和我在實驗上毫無頭緒時，吳老師都能指引我們正確的方向，並提供解決方法。這研究不少的問題都無法用一般的方法解決，但是吳老師都可以想到其他的方法來處理，然後都非常有耐心地教導我。

我畢生感激兩位指導教授這三年來努力栽培我。

感謝遠在馬來西亞的父母給我精神與經濟上的支助，讓我可以好好地讀碩班做研究。雖然三年的疫情無法回國，我在這還可以感受到你們的鼓勵與耐心。

再來要感謝出現在 R638 的夥伴們：戴睿紘、陳冠蓁、王芷琳、曾奕承、呂昊鈞、陳毓蓉、江允芃、游宗翰、張永朋、邱宇晨、彭翊倫、Carina Terry。大家一起外出遊玩、吃飯、看電影、打遊戲，讓枯燥的研究生生活增添許多樂趣。當實驗量有點大時，感謝奕承及冠蓁提供協助，減輕我的工作量。

特別感謝博班學長戴睿紘在分析上給了非常多的幫助。沒有他的幫忙，這研究的分析會加倍困難。進來這實驗室前，我對程式語言及 Linux 一竅不通，所以他教了我很多的基礎知識，讓我可以順利自我學習下去。人生能遇到幾多個願意花精力來幫忙學弟妹的人呢。

此外，要感謝吳慧琳老師實驗室 R634 的李泳璁和黃博鋐教我養細胞、養病毒、純化病毒及給我許多實驗上的教導。

感謝中研院生多中心的王子元老師及趙淑妙老師在 DNA 定序的部分提供了許多幫助與資源。

最後，感謝兩位口委-陳培哲老師及楊宏志老師針對我的研究給了很多珍貴的意見及建議，讓這研究更加好。

doi:10.6342/NTU202302511

摘要

基因多樣性源自於突變，並且在演化裡扮演着重要的角色。儘管 B 型肝炎病毒是一種 DNA 病毒，但是它在複製過程中會使用到自己的反轉錄酶，使其演化速率趨近於 RNA 病毒。雖然目前有許多 B 型肝炎病毒的相關研究，可是我們對其突變特徵仍然不明，局限了我們對於 B 型肝炎病毒基因多樣性的認知。本研究透過在每個分子上標記獨特辨識碼，達成單一分子解析度，同時利用第三代定序技術，如 Oxford Nanopore Technology (ONT) 與 Pacific Biosciences (PacBio)，對在 *in vitro* 中只有複製過一次的病毒進行接近全基因的定序。本研究在橫跨 B 型肝炎病毒 98%的基因組進行定序，ONT 與 PacBio 個別發掘了 8,500 和 6,300 個突變，並達到大於 80,000 的定序深度。從定序結果裡，根據兩個不同的定序平台裡估算出 B 型肝炎病毒在每個位點每次複製之錯誤率為 2.28×10^{-5} (PacBio)及 3.28×10^{-5} (ONT) 個核苷酸，與 C 型肝炎病毒相似。兩個定序平台算出的突變差異分別為 162 倍(ONT)及 29 倍(PacBio)。在 N/S 比例裡發現，實際人類族群裡的 B 型肝炎病毒個別基因的 N/S 比例為 0.18 至 1.15，但是本研究的 *in vitro* 裡發現個別基因的 N/S 比例為 2.65 至 3.04，姑且推估在人體裡有 52 到 94%的非同義突變是有害並且會被移除。除此之外，本研究找到十九個 B 型肝炎病毒的剪接變異體，其中五個為新型剪接變異體。以我們所知，本研究是第一個成功估算 B 型肝炎病毒之單次複製錯誤率、在一個無篩選壓力的環境裡觀察此病毒的突變模式及透過單一分子解析度定序 98%基因組探索此病毒之突變特徵。本研究可作為 B 型肝炎病毒在自然選汰、基因負荷及可演化性裡的基礎狀態，從此可以對其抗藥性及免疫逃避風險提供資料。

關鍵詞：B 型肝炎病毒、複製錯誤率、自然選汰、單鹼基解析度、第三代定序

doi:10.6342/NTU202302511

Abstract

Spontaneous mutations, serving as the ultimate source of genetic variation, play a prominent role in evolution. Hepatitis B virus (HBV), while being a DNA virus, utilises its own error-prone reverse transcriptase for replication, giving it an evolutionary rate closer to that of an RNA virus. Although HBV is a well-studied virus, its nucleotide substitution profile remains poorly understood, which limits our comprehension of how HBV generates diversity. This study achieved single-molecule resolution by tagging each molecule with unique molecular identifiers and utilised third-generation sequencing techniques, including Oxford Nanopore Technology (ONT) and Pacific Biosciences (PacBio), to sequence HBV genome that had only undergone a single round of replication *in vitro*. We identified 8,500 and 6,300 mutations from ONT and PacBio, respectively, spanning across 98% of the HBV genome, with an average depth of >80,000. The estimated replication error rates of HBV from PacBio and ONT are 2.28×10^{-5} and 3.28×10^{-5} nucleotide/site/replication, respectively, which is the same as HCV. The differences in mutability was found to be 162-fold and 29-fold for ONT and PacBio, respectively. In contrast to inter-hosts studies, where the N/S ratio for different genes ranged between 0.18 and 1.15, the *in vitro* investigation from this study shows a ratio ranging from 2.65 to 3.04, suggesting that approximately 52 to 94% of nonsynonymous mutations generated within hosts were deleterious and subsequently removed. Furthermore, we have identified nineteen HBV splice variants, five of which are novel. To our knowledge, this is the first study to estimate the error rate of HBV on a per replication basis, demonstrate the mutation pattern of HBV in a selectively neutral environment, and explore the nucleotide substitution profile of nearly full-length HBV with single-molecule resolution. These findings establish a site-specific reference for scrutinizing the aspects of natural

doi:10.6342/NTU202302511

selection, genetic load, and HBV's evolutionary potential. Such insights are instrumental in evaluating the imminent risks tied to drug resistance and immune evasion.

Keywords: Hepatitis B virus, replication error rate, natural selection, single base pair resolution, third-generation sequencing

doi:10.6342/NTU202302511

Contents

doi:10.6342/NTU202302511

doi:10.6342/NTU202302511

List of tables

doi:10.6342/NTU202302511

List of figures

doi:10.6342/NTU202302511

Chapter 1: Introduction

Hepatitis B virus (HBV), a species of partially double-stranded DNA (dsDNA) virus (genus *Orthohepadnavirus*; family *Hepadnaviridae*; order *Blubervirales*; class *Revtraviricetes*; phylum *Artververicota*; kingdom *Pararnavirae*; realm *Riboviria*) [1] with a genome size of approximately 3,200 base pairs (bp), is responsible for the disease hepatitis B. Currently, about 300 million people are suffering from chronic HBV infection, with an estimation of 1.5 million new cases each year. Those who were infected are living with an elevated risk of life threatening cirrhosis and hepatocellular carcinoma, resulted more than 800,000 deaths in year 2019 [2]. As of 2023, no known cure has been approved for hepatitis B, but fortunately, an effective vaccine is available and has drastically reduced the number of new infections and related deaths over the past decades all over the world [3].

Genetic mutations, as the main driving force of genetic variation, introduce changes in organisms, whether advantageous or disadvantageous. Any being carrying mutation that is favourable in the environment it resides is able to outcompetes its competitors and greatly increases its chance of propagation, i.e increased fitness. On the other hand, any disadvantageous mutation can be detrimental for the being carrying it and potentially prevents its lineage from progressing further, i.e decreased fitness. All organisms and viruses are naturally bound to this rule. Apart from selection force, time also play an important role, as time progress, mutations accumulate and genetic variations occur. Mutations were mainly introduced during its replication cycle, thus as time goes on, more mutations appear. Different viruses have different mutation rates, and the general pattern is that DNA viruses have a lower mutation rate than RNA viruses [4]. Although HBV is a DNA virus and does not

doi:10.6342/NTU202302511

belong to the retrovirus family, it utilises an error-prone RNA-dependent DNA polymerase, i.e. reverse transcriptase, in its replication cycle.

The life cycle of HBV [5], in brief, starts when a Dane particle enters a cell, where it uncoats its nucleocapsid and releases its relaxed-circular DNA (rcDNA) into the nucleoplasm. After rcDNA entered the nucleus, host cell machineries 'repair' rcDNA into covalently closed circular DNA (cccDNA) [6] [7], which initiates transcription. Four main RNA products are transcribed, one of which is the pregenomic RNA (pgRNA). After exiting the nucleus, pgRNA is packaged into nucleocapsid, along with translated HBV polymerase, which reverse transcribe the pgRNA into rcDNA, i.e. genomic DNA. The DNA-containing nucleocapsid is either to be re-imported back into the nucleus or ready to be enveloped and secreted from the host. The resulting rcDNA is a partially double-stranded circular DNA, which has a longer-than-genome minus strand and a shorter plus strand.

Hence, during the replication cycle of HBV, mutations can occur at two stages: (i) when viral RNA is transcribed from the cccDNA by host RNA polymerase II (Pol II), and (ii) when the single-stranded pgRNA is reverse transcribed into rcDNA by viral reverse transcriptase. In the natural infection system, nevertheless, the rcDNA can be delivered into nucleus and converted into cccDNA from which the next run of transcription begins. Therefore, the pgRNA may actually contain mutations accumulated from many runs of transcription and reverse transcription by host Pol II and viral polymerase, respectively. The situation can become complicated when different mutants actually have different replication abilities. This variation in replication dynamics will cause mutations that have higher replicative ability to increase their copy numbers and to outcompete those with lower replicative efficiencies. In addition, mutations capable of escaping from or adapting to host

doi:10.6342/NTU202302511

immune environment may increase their frequencies as well. Consequently, the genetic variation from natural infection is resulted from the complex interaction governed by mutation and selection. Under this circumstance, the inherent mutation profile is masked.

Here an *in vitro* transfection model is proposed to estimate mutation profile across HBV genome by transfecting HBV plasmid into human cell lines. The transfection model provides a couple of advantages. First, this model includes all necessary components for HBV replication, and thus can directly measure mutation profile in different parts of the genome. Second, because human cell lines lack the appropriate receptors for HBV, the newly generated virions cannot enter into the subsequent life cycle, thus, rules out the concern of selection among virus variants.

Therefore, a newly developed method which can generate high-accuracy long-read sequences using unique molecular identifiers was applied. By analysing the recovered sequences, the inherent mutation rate and profile of HBV can be estimated. Furthermore, by study the mutation profile with the genetic variation observed within and among HBV carriers, the effect of selection can be revealed.

High throughput sequencing is a method necessary for studying low-abundant and heterogeneous variants within populations, such as the quasi-species of HBV. Short-read Illumina sequencing has demonstrated to be an effective way due to its ultra-high throughput and relative low error rate. Nevertheless, the maximum sequencing size of ~500 bp precludes its application on assays required long-range information [8].

Unique molecular identifiers (UMIs) have been applied to sequence long fragments with short-reads via assembly [9] [10]. However, the method requires a substantial amount of DNA template (100 ng － 1 μg) which limits its application.

doi:10.6342/NTU202302511

The high error rates of Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) impedes their applications to confidently identify true UMI tag sequences necessary to assign reads to their template molecules. This study aims to apply a newly developed workflow that combines UMIs with sequencing of long DNA fragment on the ONT and PacBio platforms to produce single molecule consensus sequences with high accuracy [11]. The UMIs contain a special internal pattern that avoids error-prone homopolymer stretches and improves recognition of UMI-tags. In combined with filtering based on UMI length and pattern allows for a robust determination of true UMI sequences in error-prone ONT and PacBio data.

doi:10.6342/NTU202302511

Chapter 2: Materials and methods

Cell culture and virus

Huh7 cells (kindly provided by Dr. Hui-Lin Wu) were cultured in Dulbecco's Modified Eagle Medium (DMEM) (Corning, USA, cat. no.: 10-013-CM), containing 10% fetal bovine serum (Corning, USA, cat. no.: 35-010-CV), 2 mM glutagro (Corning, USA, cat. no.: 25-015-CI), 1% MEM Nonessential Amino Acids (Corning, USA, cat. no.: 25-025-CI), and 10 mM HEPES (Corning, USA, cat. no.: 25-060-CI), and were incubated in 37℃, 5% CO_2. pAAV/HBV1.2 plasmid was kindly provided by Professor Pei-Jer Chen. Prior to the transfection of pAAV/HBV1.2 plasmid into Huh7 cells, approximately 4×10^6 cells were passaged to a new 100 mm tissue culture dish (Iwaki AGC, Japan, cat. no.: 3020-100) and incubated overnight. For each dish, 16 μg of pAAV/HBV1.2 plasmid [12] was transfected into Huh7 cell with Lipofectamine 3000 (Thermo Fisher, USA, cat. no.: L3000001), according to the manufacturer's protocol, and was incubated overnight. Transfected cells were washed with PBS on day 1 and have their medium replenished. Cell culture supernatants were collected on days 3, 5, and 7 post-transfection. HBV viral titre was determined with cobas e 411 analyzer (Roche, Switzerland) to ensure the success of transfection.

Extraction of HBV DNA

Collected cell culture supernatants were filtered with 0.22 μm Millex-GS filter unit (Merck, USA, cat. no.: SLGSV255F) and subjected to ultracentrifugation with SW 32 Ti rotor (Beckman Coulter, USA, cat. no.: 369650) at 25,000 rpm 4℃ for 16 hours. The supernatant was discarded, the pellet was resuspended with 200 μL of PBS, and the viral DNA was extracted using QIAamp DNA Blood mini kit (Qiagen, Germany, cat. no.: 51106) by following the manufacturer's protocol.

doi:10.6342/NTU202302511

Cleaving plasmid DNA

After DNA extraction, any carried-over plasmid was cleaved with DpnI (New England Biolabs, USA, cat. no.: R0176S) in a 50 μL reaction, which contained 44 μL of DNA, 5 μL of 10× rCutSmart Buffer (New England Biolabs, USA, cat. no.: B6004S), and 1 μL of DpnI, at 37℃ for 3 hours and heat inactivated at 80℃ for 20 minutes. Plasmid contamination was checked using PCR with vector-specific primers in a 25 μL reaction, containing 15.75 μL of nuclease-free water, 5 μL of 5× Platinum SuperFi buffer (Invitrogen, USA, cat. no.: 12351010), 0.5 μL of 10 mM dNTP (Bioman Scientific Co., cat. no.: D1001), 1.25 μL of 10 μM pAAV/HBV1.2_Backbone_F primer (Genomics, Taiwan), 1.25 μL of 10 μM pAAV/HBV1.2_Backbone_R primer (Genomics, Taiwan), 0.25 μL of Platinum SuperFi Polymerase (Invitrogen, USA, cat. no.: 12351010), and 1 μL of DNA. Plasmid-checking PCR was done with the following program: initial denaturation of 95℃ for 30 seconds, followed by 40 cycles of 95℃ 30 seconds denaturation, annealing at 49℃ for 30 seconds, and 72℃ extension for 2 minutes. PCR products were visually analysed with 1.5% agarose gel electrophoresis, as shown in Figure 2B.

Quantifying HBV DNA

HBV DNA quantification was done by using agarose gel-based quantitative PCR in a 25 μL reaction, which contained 15.75 μL of nuclease-free water, 5 μL of 5× Platinum SuperFi buffer (Invitrogen, USA, cat. no.: 12351010), 0.5 μL of 10 mM dNTP (Bioman Scientific Co., cat. no.: D1001), 1.25 μL of 10 μM HBV_3kb_Tag_F (Integrated DNA Technologies, USA), 1.25 μL of 10 μM HBV_3kb_Tag_R2 (Integrated DNA Technologies, USA), 0.25 μL of Platinum SuperFi Polymerase (Invitrogen, USA, cat. no.: 12351010), and 1 μL of DNA. The PCR program was as

doi:10.6342/NTU202302511

follow: 95℃ initial denaturation for 30 seconds, followed by 25 cycles of 95℃ denaturation for 30 seconds, 65℃ of annealing for 30 seconds, and extension of 72℃ for 3 minutes. Serial diluted of previously quantified AD38 derived HBV DNA was used as standard (7×10^5 copies, 7×10^4 copies, 7×10^3 copies, and 7×10^2 copies). After PCR, 2.5 μL of PCR products were mixed with 0.5 μL of FluoroDye (SMOBIO Technology, Taiwan, cat. no.: DL5000) and electrophoresed at 100 V for 35 minute in a 1.5% agarose gel in TAE buffer. The Huh7 derived HBV DNA was approximately quantified through visual means. Example is shown in Figure 2C.

Primer design

For both plasmid and HBV DNA, amplification of the full-length HBV using PCR was unachievable for this application because the sensitivity of previously designed primer for full-length HBV, primer P1 and P2 [13], were insufficient, thereby requiring much more DNA than the system utilised here can handle. Primers with UMI were based on previously designed primer [11] and has been tweaked for targeting HBV sequence (Figure 1B). See Table 1 for a list of all primer used and which primers were paired together for which dataset.

Tagging and amplifying HBV DNA for long-read sequencing

The protocols for PCR and purification were largely based on previously described method [11], with some tweaks to better suit this application. Approximately 1×10^5 copies of pAAV/HBV1.2 plasmid or 1×10^6 copies of Huh7 derived HBV DNA were used in the first PCR in a 45 μL reaction, which contained 10 μL of 5× Platinum SuperFi buffer (Invitrogen, USA, cat. no.: 12351010), 1 μL of 10 mM dNTP (Bioman Scientific Co., cat. no.: D1001), 0.5 μL of Platinum SuperFi

doi:10.6342/NTU202302511

Polymerase (Invitrogen, USA, cat. no.: 12351010), 3.3 μL of HBV DNA, and 30.2 μL of nuclease-free water (Thermo Scientific, USA, cat. no.: R0582). The PCR program was 1 minute of denaturation at 95℃, 40 seconds of annealing at 65℃, and 5 minutes of repair at 72℃. After repairing, 2.5 μL of 10 μM forward primer (see Table 1B for specific primer pair) and 2.5 μL of 10 μM reverse primer (see Table 1B for specific primer pair) of primers were added into the first PCR mix and DNA was tagged with UMI, which was done with 2 cycles of 95℃ denaturation for 30 seconds, 65℃ anneal for 30 seconds, and 72℃ extension for 4 minutes. The PCR product was purified with 22.5 μL (0.45×) of KAPA Pure Beads (Roche, Switzerland, cat. no.: 07983271001), following the manufacturer's protocol, and eluted with 20 μL of nuclease-free water. Purified amplicon was then amplified with a second round of PCR in a 100 μL reaction, containing 20 μL of 5× Platinum SuperFi buffer, 2 μL of 10 mM dNTP, 5 μL of HBV_Amp_F 10 μM primer (Integrated DNA Technologies, USA), 5 μL of 10 μM HBV_Amp_R primer (Integrated DNA Technologies, USA), 1 μL of Platinum SuperFi Polymerase, 20 μL of purified DNA, and 47 μL of nuclease-free water. The PCR program was an initial denaturation of 95℃ for 30 seconds, followed by 25 cycles of 95℃ denaturation for 30 seconds, 55℃ annealing for 30 seconds, and 72℃ extension for 4 minutes. A 2 minutes of final extension at 72℃ concluded the second round of PCR. The amplicon was then purified with 45 μL (0.45×) of KAPA Pure Beads and eluted with 60 μL of nuclease-free water. After purification, amplicon was aliquoted into 3 × 20 μL and amplified again with a third round of PCR in 3 × 100 μL reactions. The PCR condition for each reaction was the same as the second PCR but the PCR program was changed to an initial denaturation of 95℃ for 30 seconds, followed by 6 cycles of 95℃ denaturation for 30 seconds, 55℃ annealing for 30 seconds, and 72℃ extension for 4 minutes, ended with a final extension at 72℃ for 2

doi:10.6342/NTU202302511

minutes. Lastly, PCR products were pooled together (300 μL), purified with 135 μL (0.45×) of KAPA Pure Beads, and eluted with 50 μL of nuclease-free water. The concentration and quality of the tagged-amplified-purified amplicons were checked by Nanodrop and agarose gel electrophoresis, respectively, before sequencing by MinION or PacBio. A total of 3.2 μg and 7.6 μg of amplicons were generated for MinION and PacBio Sequel IIe sequencing, respectively. Schematics for PCR are shown in Figure 1C and example of the amplicons are shown in Figure 2D.

Data generation

For Oxford Nanopore Technologies, MinION R.9.4.1 flowcells (FLO-MIN106) were used with the software MinKNOW Core version 19.12.5. High-accuracy basecalling was done with Guppy version 3.4.4 in GPU mode (NVIDIA GeForce GTX 1060 6GB) with this command: *guppy_basecaller -i $INPUT_FOLDER -s $OUTPUT_FOLDER -c dna_r9.4.1_450bps_hac.cfg -x auto --chunks_per_runner 1024 --num_callers 3*. As for PacBio, Sequel I and Sequel IIe were used for plasmid sequencing and Huh7 derived HBV DNA, respectively. Both sequencing techniques were carried out by Academia Sinica Biodiversity Research Center, with Dr. Tzi-Yuan Wang (MinION) and NGS High Throughput Genomics Core facility (PacBio).

Identifying valid unique molecular identifiers (UMI)

All basecalled fastq files from sequencing were first have their UMI identified, corrected, sorted, and listed out by using *longread_umi* pipeline [11]. Sequences from MinION used the command: *longread_umi nanopore_pipeline -d $INPUT_FASTQ -v 15 -o $OUTPUT_FOLDER -s 100 -e 100 -m 1400 -M 3300 -f*

doi:10.6342/NTU202302511

CAAGCAGAAGACGGCATACGAGAT -F AAAAAGTTGCATGGTGCTGG -r ATGATACGGCGACCACCGAGATC -R GTCCTACTGTTCAAGCCTCCA -c 4 -p 2 -q r941_min_high_g344 -t 10 -T 4. Although the recommended UMI depth requirement for ONT data was ≥ 25, this study set the UMI depth threshold of ≥ 15. This decision was made in consideration of increased data recovery while does not compromise on mismatch error rate [11].

The basecalled reads from PacBio used the following command: *longread_umi pacbio_pipeline -d $INPUT_FASTQ -o $OUTPUT_FOLDER -v 3 -m 1400 -M 3300 -s 100 -e 100 -f CAAGCAGAAGACGGCATACGAGAT -F AAAAAGTTGCATGGTGCTGG -r AATGATACGGCGACCACCGAGATC -R GTCCTACTGTTCAAGCCTCCA -c 2 -t 10*. A fasta file with a list of all the identified valid UMI can be found in the output folder of *longread_umi* with the name *consensus_raconx$NUM1_medakax$NUM2_$NUM3.fa*. The *$NUM1*, *$NUM2*, and *$NUM3*, signifying rounds of polished by Racon, Medaka, and minimum UMI depth, respectively. Henceforth, this file shall be called *$UMI.fa*.

Search for inversion and recombination

To search for any inversion or recombination event between HBV DNA and human or plasmid DNA, a local BLAST [14] was set up, with pAAV/HBV1.2 reference as database, and executed the following command: *blastn -task blastn -query $UMI.fa -db $DATABASE.fa -evalue 1e-20 -num_threads 4 -out $OUTPUT.txt -outfmt "6 qseqid qcovus"*. It output a list with two columns, the first column was the name of each UMI (*qseqid*) whereas the second column was the query coverage (*qcovus*). After that, a custom python code was utilised for searching and averaging out the query coverage for each UMI, then it sorted by the averaged query coverage

doi:10.6342/NTU202302511

by ascending value. Any UMI with an average query coverage of 98% or below was individually checked for by using MEGAX version 10.2 [15]. Any sequence that was found to be partially inverted or recombined with either human or plasmid DNA was removed from further analysis.

Alignment

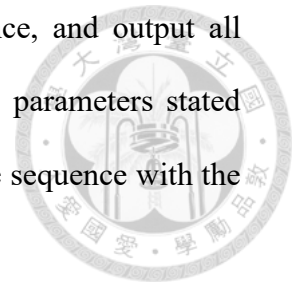
All valid UMI were aligned to the reference sequence, pAAV/HBV1.2, by MAFFT version 7.310 [16] with the command *mafft --6merpair --thread 10 --keeplength --adjustdirection --addfragments $UMI.fa $REFERENCE.fa > $UMI_ALIGNED.fa*. The purpose of *--6merpair* was to increase alignment speed, *--keeplength --adjustdirection* for keeping the length and direction, respectively, of the output the same as the reference sequence, and *--addfragments* was to add and align every sequences in *$UMI.fa* to *$REFERENCE.fa*, which is useful when 'fragmented' sequences were expected, such as splice variants. Two 'different' reference sequences were used, both have the exact same sequence, except one has EcoRI recognition site as the starting site (HBV defined starting site) and the other started on site 1,873 (amplicon's actual starting site). The latter was used for aligning UMI at first, while the former was used after correcting the position of all sequences.

Correcting sequence position

The starting site of HBV is defined as the middle of EcoRI recognition site (GAA|TTC), but *de facto* start site of the amplicon was not, so a custom python script was used to tackle this. First, this script finds the position of 'GAATTC' in the given reference sequence, then it counts the position which the middle of 'GAATTC' is in. After that, on the same position for each and every sequence, it prints the second half

doi:10.6342/NTU202302511

of the sequence first, follow by printing first half of the sequence, and output all sequences into a fasta file. By using MAFFT, applying the same parameters stated above, the position-corrected sequences were aligned to a reference sequence with the HBV defined start site.



Searching and counting mutations

After the alignment from the previous step, all sequences were in the correct position and length. A custom python code was wrote for finding mutation on each site when compared to pAAV/HBV1.2 reference. First, this code compared the base on each site with the same site in reference sequence, if it was the same base, it was ignored and proceed to the next site, but if a mismatch was found, then it would be written into a CSV file with three columns: the substitution type (e.g cytosine to thymine was expressed as 'C>T'), the name of the UMI with this mutation, and the position of that mutation. All deletions, '-', were ignored. At the same time, the occurrence of all substitution types were summarised in another CSV file, e.g cytosine to thymine occurred 100 times throughout the whole dataset was expressed as 'C>T,100'.

Calculating replication error rate

All mutations found from the previous step, except C→A and G→T in the plasmid data and C→A in the Huh7 derived HBV DNA data, were summed and divided by the summation of sequenced bases, which was done by counting the length of each valid UMI and sum them all up.

$$\text{Replication error rate} = \frac{\text{Sum of all mutations}}{\text{Sum of all sequenced bases}}$$

doi:10.6342/NTU202302511

Identifying splice variants

From the output of *longread_umi*, a file (*variants.fa)* was generated, containing consensus sequences supported by three or more UMI. By aligning all the consensus sequences with MAFFT to the reference sequence, all aligned sequences were inspected with MEGAX and visually identified. Identified novel splice variants must meet all three criteria to be considered as real splice variants: 'GG' or 'GT' or 'GC' at the splice donor site, 'AG' at the splice acceptor site, and not within the pre-S2 deletion region [17] [18]. Novel splice variants were named following previously described nomenclature [19].

Mapping out mutation distribution across genome

To visualise mutation distribution, a custom python script was written that calculate the nucleotide frequency of each nucleotides in each site, then another python script was used for adding up the nucleotide frequencies of alternative nucleotides, i.e the summation of nucleotide frequencies of the other three nucleotides that were not the reference nucleotide. With the alternative nucleotide frequency for each site in hand, a custom sliding window python script was used to count the average of alternative nucleotide frequency in a specified window size and slid along the HBV genome with a specified step size. A window size and step size of 100 bp and 10 bp, respectively, was selected. Due to a 69 bp gap between site 1,803 to 1,873, the last window prior to site 1,803 stopped at site 1,800 and restarted the window from site 1,873. The mutation distribution across the HBV genome was plotted out with the sliding window python script output. For the inter-host genotype A HBV, 970 sequences were downloaded from the nucleotide dataset of HBVdb [20] and

doi:10.6342/NTU202302511

subjected to alignment and has their mutations counted as described in this study prior to plotting out the mutation distribution.

Nonsynonymous/synonymous ratio (N/S ratio)

The expected number of nonsynonymous and synonymous mutations were calculated with a custom python code, because all C→A (and C→T later) mutations have to be removed from the expected N/S ratio for a fair comparison with observed N/S ratio. In brief, this python code counts the possible number of nonsynonymous and synonymous mutations a codon can experience if only a single mutation happened to a single nucleotide, e.g first codon position: CTT can turn into TTT, GTT, and ATT, which gives one nonsynonymous mutation because CTT to ATT and TTT were ignored; second and third codon positions were too processed the same way; stop codon was considered as a separate category, not nonsynonymous mutation. After processing all the codons, the sequences of every gene and interested open reading frame were fed into the python code, which counted the amount of nonsynonymous and synonymous mutations for each codon and summed the number up at the end. As for the observed number of nonsynonymous and synonymous mutations, the calculation was done by another custom python code and the C→A/T mutations were simply removed from the input CSV file, which was generated from '***Searching and counting mutations***'. By giving the starting site and ending site, on the reference genome, of a desired reading frame, this python code retrieved the position and the mutation type from the input CSV file, then it applied the mutation to one position at a time and check whether that mutation was nonsynonymous or synonymous. N/S ratio was generated from dividing the number of nonsynonymous mutation by the number of synonymous mutation. To count the amount of

doi:10.6342/NTU202302511

nonsynonymous mutation lost due to selection in HBV within human population, the amount of nonsynonymous and synonymous mutations in were first counted as stated above. Then, for each gene, the number of synonymous mutation *in vivo* was divided by that of *in vitro*, the resulting number was multiplied by the number of nonsynonymous mutation *in vitro*. Lastly, the number from the previous step was to divide the number of nonsynonymous mutation *in vivo*. This showed the percentage of mutations remained *in vivo* and the percentage of nonsynonymous mutations lost due to selection *in vivo* can be calculated by simply minus one.

To estimate the strength of natural selection, all synonymous substitutions are assumed to be neutral. The expected number of nonsynonymous substitutions should equal to the expected N/S ratios times the observed number of synonymous substitutions. The differences between observed and expected number of nonsynonymous substitutions is thus caused by selection. The expected N/S ratios were calculated from Huh7 derived HBV and the observed number of synonymous and nonsynonymous substitutions are from HBVdb.

Types of mutation in every mutated site

Every site that had experienced mutation was fed into a custom python code which finds the type of mutations that occurred in those given sites, then the code summarised how many of sites show one or two or three types of mutation. Another custom python code was written to find the nucleotides on sites exhibiting three types of mutation.

doi:10.6342/NTU202302511

Chapter 3: Results

Background error rate of this methodology

In order to know the limit of our methodology, a plasmid control was needed. Plasmid has a reported error rate of 5.4×10^{-10} nucleotide/site/replication [21], which is several orders of magnitude lower than any virus, and thus a suitable option as our control. The plasmid used was pAAV/HBV1.2, a plasmid that can generate Dane particle when transfected into hepatocyte derived cells [12], thus an ideal candidate for both plasmid control and *in vitro* transfection. As a prove of concept, a HBV-specific primer pair that can produce a 2,010 bp PCR product was used (Table 1B). Varying amounts of initial DNA in the UMI-tagging PCR were tried, which were 5×10^7, 5×10^6, 5×10^5, 1×10^5, and 5×10^4 copies of DNA (Figure 2A). During the preliminary test with 5×10^5 initial copies of plasmid and sequenced with one MinION cell, no usable data can be recovered due to insufficient read depth. The tool that analyses UMI (*longread_umi*) requires every UMI to reach a certain read depth (ONT: ≥ 15, PacBio: ≥ 3) in order to be considered valid, thus using copious amount of DNA will in turn requires more sequencing power. In the end, 1×10^5 copies of initial plasmid DNA was tried and the read depth was found to be enough for analysis. It is undeniable that by increasing the amount of initial DNA, the quantity and quality of the obtained data can be increased too, but the cost in sequencing would quickly become prohibitively expensive.

From the MinION sequencing result for plasmid, 14 Gb of raw data and 5,866,550 reads were generated. 20,012 (20%) of DNA molecules were tagged with valid UMI, 2,075,261 reads were analysed, and the average read depth was 20,011 (standard deviation (SD): 24) (Table 2A). A total of 823 mutations were found. The

doi:10.6342/NTU202302511

estimated error rate was 2.05×10^{-5} nucleotide/site. The nucleotide substitution profile shows a strong mutation composition bias towards C→A and G→T (C→A: 322, G→T: 306, T→C: 63, G→A: 44, C→T: 27, A→G: 18, A→C: 17, T→A: 12, T→G: 8, A→T: 3, C→G: 2, G→C: 1). Given that C→A is just the complementary of G→T, their ratio is about 1:1, and distributed almost evenly across the genome, we hypothesised that these errors were introduced during the first round of PCR, i.e. UMI-tagging PCR. These biases generated from PCR were not unheard of, even from high fidelity PCR polymerases, and the PCR polymerase used in this study was unable to correct mismatch involving changing any nucleotide to thymine [22], while another study [23] shows that G→T error can be introduced during the first round of PCR. Within all the 525 cytosine sites sequenced, 34.10% (179) of them mutated to adenosine, while 39.81% (166) of the 417 guanine sites mutated to thymine.

To ensure that this bias was not stemmed from the sequencing platform, the plasmid was treated with the same procedure, but now with a pair of primer which generate 3,115 bp amplicons and sequenced with Sequel I (Table 1B). The generated HiFi reads has a mean read passes of seventeen and input into *longread_umi* to extract all valid UMI and call all variants. Although the sequencing result from Sequel I has far fewer reads (232,781), less valid UMI (3,455), and lower average read depth (3,455; SD: 0.03), yet 818 mutations were found, which is close to the 823 mutations found in MinION (Table 2A). The nucleotide substitution profile is similar to that of MinION, with an overwhelming C→A and G→T mutations: C→A: 532, G→T: 271, G→A: 7, C→T: 5, A→G: 2, C→G: 1. As shown, Sequel I also has this strong C→A and G→T bias, albeit the ratio between them was more skewered towards C→A than the nucleotide substitution profile from MinION. All the cytosine sites sequenced in this dataset, 22.04% (184 out of 835) mutated into adenosine and 26.60% (183 out of

doi:10.6342/NTU202302511

688) of guanine sites mutated into tyrosine. In terms of the exact number of cytosine and guanine sites that had experienced this C→A and G→T bias, they were actually quite close within the same sequencing platform (MinION: 179 C vs. 166 G; Sequel I: 184 C vs. 183 G). Combining the fact that this bias was occupying the vast majority of mutations (MinION: 76.31%, Sequel I: 98.17%) in both sequencing platforms, plasmid is unlikely to have such high mutation rate, and the PCR polymerase used in this study is unable to correct a specific kind of error (any nucleotide to thymine) [22] [23], we assume that this was a PCR artefact, which will be removed from further analyses. After removing this PCR artefact, the amount of mutations left in ONT MinION and PacBio Sequel I were 195 and 15, respectively, with respective error rates of 4.85×10^{-6} and 1.39×10^{-6} nucleotide/site.

After the removal of C→A and G→T, 144 sequences contain a single mutation, 22 sequences has two, and one sequence each has three and four mutations, respectively, from MinION sequencing of plasmid (Figure 3A). Whereas in Sequel I, all fifteen mutations were individually came from different sequences (Figure 3B).

Transfection model

To generate Dane particle, pAAV/HBV1.2 plasmid was transfected into Huh7 cells by using lipofection. The lack of appropriate receptors on Huh7 cell disallowed any generated Dane particle to re-infect [24], ensuring any HBV particle collected from the cell culture supernatant had only replicated once and removed the competition among HBV particles, i.e. no selection pressure. After transfecting pAAV/HBV1.2 plasmid into Huh7 cells, cell culture supernatants were collected and DNA was extracted. More details can be found in Materials and Methods. Concerned of carry-over plasmid from the transfection, DpnI, a methylation-specific restriction

doi:10.6342/NTU202302511

enzyme, was used to cleave plasmid. To ensure the full digestion of plasmid, a PCR test was conducted with vector-specific primers. Any plasmid contamination would generate a 2,604 bp PCR product and no plasmid was detected from the DpnI-treated Huh7 derived HBV DNA (Figure 2B).

Quantification of HBV DNA

Real-time/quantitative PCR (qPCR) was attempted for quantifying the amount of extracted HBV DNA, but the amount of Huh7 derived HBV DNA quantified by qPCR was markedly more than expectation. We hypothesised that because qPCR synthesised short amplicons (about 150 bp) and DpnI is an endonuclease with a specific restriction site, the remaining fragments of plasmid might be long enough for amplification.

To test this hypothesis, a HBV-specific primer pair ('SP-5 HBV +2413' and 'HBV -2551' in Table 1A) were picked, with 'SP-5 HBV +2413' containing the cutting site of DpnI, and ran a qPCR with two groups of plasmid, one had been treated by DpnI and the other had not. The qPCR result showed that the DpnI-treated group was practically indistinguishable from the untreated group (Figure 4), thus showing that, at this scale, qPCR is unreliable in quantifying HBV DNA when plasmid was involved.

Furthermore, as amplifying full-length HBV was the target of this study, we opted for normal PCR with HBV_3kb_Tag_F and HBV_3kb_Tag_R2 as primers and quantified with agarose gel. Although gel-based quantification method is much more inaccurate and has larger margin of error compared to qPCR, it was sufficient enough for an approximation within an order of magnitude. Besides, this quantification method was quantifying full-length HBV DNA instead of fragmented HBV DNA,

doi:10.6342/NTU202302511

thus making it a fitter choice for our application. As AD38 cell line does not require plasmid transfection to generate Dane particles, quantification of AD38 derived HBV by qPCR is accurate, thus a serial diluted AD38 derived HBV DNA of known quantity was used as copy number standard and compared with Huh7 derived HBV DNA with the normal full-length PCR. This quantifying method proved to be suitable and sufficient enough for this application (Figure 2C).

Tagging, amplifying, and sequencing Huh7 derived HBV DNA

After trials and errors, I was not able to generate any amplicon with previously designed primers for amplifying full-length HBV (forward: 5'-TTT TTC ACC TCT GCC TAA TCA and reverse: 5'-AAA AAG TTG CAT GGT GCT GG) [13]. A hypothesis was put forward that as the HBV DNA contains a short repeat sequence in both the terminal ends on the minus strand, both sites are opened to annealing by the forward primer, allowing two forward primers to be annealed onto the same DNA molecule. Given that the PCR polymerase used in this study has no helicase activity, the extension from one end was blocked by another primer during the first cycle of UMI-tagging PCR, thereby creating an incomplete amplicon without the binding site for the reverse primer to anneal to during the second PCR cycle (Figure 5). A new primer was designed and this subject will be delved deeper in Discussion.

By using the newly designed primer (HBV_3kb_Tag_F) that skips site 1,804 to 1,872 and spans across 98% of the entire HBV genome (3,152/3,221), usable amplicon could be generated. To identify more mutations and increase sequencing power, the initial copies of Huh7 derived HBV DNA were increased to approximately 1×10^6 and two MinION flow cells were used for sequencing. Forty-five Gb of raw data and 14,005,936 reads were generated. After analysing the sequencing result from

doi:10.6342/NTU202302511

MinION, a total of 84,427 valid UMIs were identified, 1,758,783 reads were used, 260,268,297 bases were sequenced with an average depth of 82,573 (SD: 2,282) (Table 2B). A total of 11,905 mutations were found with the following nucleotide substitution profile: C→A: 2,275, C→T: 2,015, A→G: 1,961, G→A: 1,929, T→C: 1,125, T→G: 1,038, A→T: 449, T→A: 325, G→C: 279, A→C: 228, G→T: 194, C→G: 87.

Similar to the plasmid sequencing results, the PCR artefact remains the most abundant, but unlike plasmid's results, only C→A substitutions are dominating the nucleotide substitution profile while G→T substitutions do not. This can be explained by the fact that only the minus strand of HBV DNA is a complete genome and only that particular strand was tagged during PCR. As stated before, the PCR polymerase used in this study has the tendency to change any nucleotide to thymine, thus any G→T substitution on the minus strand will be C→A when read from the direction of plus strand, as they are the complement of each other. This phenomenon becomes particularly evident when only a single strand of DNA is available for PCR [23]. So, for all the sequencing results from Huh7 derived HBV, C→A mutations are excluded. After excluding C→A mutations, 9,630 mutations remained.

Another observation was noticed when the mutation distribution was plotted out on a per site basis. There were a several sites with incredibly high mutation number that was also observed from the plasmid MinION data, which can be seen in Figure 6A. Unsurprisingly, most of these sites were either flanked by homopolymer or the edge of homopolymer, which remains a major obstacle faced by Oxford Nanopore Technology (ONT) [25]. These mutation hotspots were the result of ONT sequencing platform rather than plasmid itself, for these hotspots were not found in PacBio sequencing platform in both the plasmid and the Huh7 derived HBV data (show later).

doi:10.6342/NTU202302511

By listing out sites with the most abundant mutations from plasmid MinION data, any sites that were occupying more than 0.3% of all mutations in data were removed from the Huh7 derived HBV MinION data, namely site 198, 209, 849, 2,625, 2,626, 2,627, 3,059, 3,060, and 3,218. This brought the total number of mutations down to 8,549, which gave a mutation rate of 3.28×10^{-5} nucleotide/site/replication and a nucleotide substitution profile of C→T: 1,984, A→G: 1,845, G→A: 1,750, T→C: 917, T→G: 566, A→T: 447, T→A: 323, G→C: 277, G→T: 194, A→C: 159, C→G: 87 (Table 3B). The replication error rate of Huh7 derived HBV estimated from MinION sequencing is an order of magnitude above the systematic error rate.

By subjecting Huh7 derived HBV DNA with the same treatment as before, the amplicon was sent for Sequel IIe sequencing. 8.8 Gb of raw data (2,711,133 reads; mean passes: 21) were retrieved and analysed by *longread_umi*. A total of 752,044 reads were used, which provided 280,482,121 bases, and 89,973 UMIs were identified (Table 2B). With an average depth of 88,986 (SD: 1,084) across the genome, 8,452 mutations were found with the following nucleotide substitution profile: C→A: 2,067, C→T: 1,573, G→A: 1,391, A→G: 1,021, T→C: 779, T→G: 464, A→T: 380, G→C: 332, T→A: 260, G→T: 102, C→G: 47, A→C: 36. Similar to prior sequencing results, the PCR artefact 'C→A' remained dominant and was subsequently removed from further analysis, which left 6,385 mutations. Unlike the MinION sequencing results, no mutation hotspot was found in the Sequel IIe data. This result gave a mutation rate of 2.28×10^{-5} nucleotide/site/replication, near the estimated mutation rate from ONT MinION. The nucleotide substitution profile shows a highly diverse mutations, very much unlike the PacBio Sequel I sequencing result of plasmid, suggesting what has been obtained are the results of HBV spontaneous mutations.

doi:10.6342/NTU202302511

Similar to the Huh7 derived HBV MinION data, all four transitions are the most abundant, followed by rest of the transversions. Between both sequencing platforms, the nucleotide substitution profiles of Huh7 derived HBV are highly similar (Pearson's r = 0.97) (Table 4B) and the transition to transversion ratios are approximate to each other (MinION: 3.16, Sequel IIe: 2.94). As shown in Figure 3C and 3D, the vast majority of mutated sequences have only a single mutation (MinION: 59%, Sequel IIe: 75.16%) while every sequence has approximately 5% (Sequel IIe) to 6% (MinION) chance of gaining a single mutation.

In conclusion, the inherent mutation rate, i.e single replication error rate, of Huh7 derived HBV is between 2.28×10^{-5} to 3.28×10^{-5} nucleotide/site/replication.

Mutation distribution of HBV genome

Figure 6A show the mutation distribution across the HBV genome. A notable pattern can be seen is that several mutation hotspots can be found in ONT MinION, regardless of DNA type, and they are slightly correlated to each other (Spearman's ρ = 0.27) Table 4A. The mutation distributions of plasmid's MinION and Sequel I are negatively correlated to each other (Spearman's ρ = -0.40), shown in Table 4A. After investigating the top ten mutation hotspots (site 17, 714, 787, 853, 854, 1,747, 1,755, 2,085, 2,090, and 3,121), seven of them are either homopolymer site or neighbouring it (site 714, 787, 853, 854, 1,747, 2,085, and 2,090). This shows that MinION has rather strong sequencing bias and this is not an accurate representation of mutation distribution in Huh7 derived HBV. On the contrary, PacBio sequencing results do not exhibit such bias and have an evenly distributed mutations. The correlation in mutation distribution between MinION and Sequel IIe for Huh7 derived HBV is

doi:10.6342/NTU202302511

comparatively low (Spearman's ρ = 0.32) (Table 4A), suggesting that the mutations that had occurred during replication were spontaneous.

To compare the mutations obtained in this study with those within human population, 970 sequences of genotype A inter-host HBV were downloaded from HBVdb [20]. There is no correlation in mutation distribution between the *in vivo* and Sequel IIe *in vitro* (Spearman's ρ = 0.11) as shown in Table 4A. These suggest that what was obtained from sequencing Huh7 derived HBV with PacBio Sequel IIe represents the inherent mutation distribution of HBV. By comparing the normalised mutation distribution of Sequel IIe Huh7 derived HBV (*in vitro*) with genotype A inter-host HBV (*in vivo*), the fluctuation magnitude of inter-host HBV was 4.6 × more than that of Huh7 derived HBV (Sequel IIe *in vitro* normalised average: 1.0, standard deviation: 0.13; *in vivo* normalised average: 1.0, standard deviation: 0.58) (Figure 6B), demonstrating the effects of natural selection.

The nucleotide substitution profile of inter-host HBV sequences is as follow: C→T: 13,047, A→G: 12,538, G→A: 11,103, T→C: 10,626, C→A: 5,485, A→C: 4,921, A→T: 4,471, T→A: 3,712, G→T: 3,328, T→G: 2,278, G→C: 1,591, C→G: 559 (Table 3C). After removing C→A mutations from the dataset, the nucleotide substitution profile of genotype A inter-host HBV is highly correlated with that of Huh7 derived HBV (Pearson's *r*: MinION = 0.92; Sequel IIe = 0.88) Table 4B. Therefore, while the mutation distributions between *in vivo* and *in vitro* data showed great inconsistency, the nucleotide substitution profiles are similar, suggesting that the discrepancy in mutation distributions between two datasets are caused by natural selection.

A noticeable similarity between the *in vivo* and *in vitro* was that C→T mutations were always the most dominant transition mutation, especially *in vitro*

doi:10.6342/NTU202302511

where C→T were about twice as frequent as T→C. This was theorised by a study [26] that an RNA editing enzyme, activation-induced cytidine deaminase (AID), turns C→U in HBV RNA. Although the study also pointed out that AID was capable of turning C→T in HBV minus strand DNA, thus a G→A mutation from the perspective of plus strand, but this substitution was not as dominant as C→T. Another study [27] also pointed out that APOBEC3B is capable of introduce hypermutations in HBV rcDNA in the form of C→T and G→A mutations. Yet, no study was found that can explain the abundance of A→G mutations.

Selection intensity of genotype A HBV in human population

The selection pressure on HBV, be it competition among Dane particles or resisting clearance by hosts, within human population can be reflected by ratio of nonsynonymous mutation to synonymous mutation (N/S ratio). As nonsynonymous mutation changes amino acid, the resulting mutation can be either advantageous or disadvantageous, which subsequently changes the N/S ratio. Under a completely neutral situation, the observed N/S ratio of a gene should be the ratio of number of nonsynonymous sites divided by number of synonymous sites, as shown in the 'Expected' panel in Table 5. For Huh7 derived HBV, the N/S ratio of polymerase and surface genes are significantly smaller than the expected values (Table 5A), indicating that there are fewer nonsynonymous mutations than expected. The deviation from neutral expectation is perplexing, as Huh7 derived HBV only experienced one run of replication with no noticeable selection forces.

Concerned of activation-induced cytidine deaminase (AID) and apolipoprotein B mRNA editing catalytic polypeptide 3B (APOBEC3B) played a role in affecting N/S ratio, since AID and APOBEC3B causes C to U substitutions in RNA [26] [27]

doi:10.6342/NTU202302511

which only results in synonymous changes at two-fold degenerate sites (transitions such as C↔T and A↔G only cause synonymous changes, while transversion cause nonsynonymous changes at two-fold degenerate sites), it can decrease the observed N/S ratio which, in turn, can cause deviations from neutral expectations. To get around this potential problem, C→T mutations were removed from both the expected and observed mutations for further analysis. As shown in Table 5B, every N/S ratio does increase slightly and only the polymerase gene remained significantly lower than the expected N/S ratio. The difference became insignificant after correction for multiple tests. Further splitting the polymerase gene into four functional domains, namely terminal protein, spacer, reverse transcriptase, and RNase H, show that all of them are consistent with neutral expectation.

Next, comparing the N/S ratios derived from Huh7 derived HBV vis-à-vis those from inter-host HBV sequences. As Table 5B demonstrates that the nucleotide substitution profile and distribution of the former are neutral, the deviation from this neutral expectation serves as a measure of the strength of selection. The N/S ratios of all ORFs derived from inter-host HBV sequences (Table 6B) are significantly smaller than those from Huh7 derived HBV (Table 5B). For example, the N/S ratio of core from the former and the latter are 0.18 and 2.93, respectively. Assuming the synonymous substitutions are neutral, the difference suggests that approximately only 6% (see Materials and Methods) of amino acid changes were maintained and 94% of the rest were removed from the population. The latter is defined as the strength of negative selection which ranges from 94% (core ORF) to 52% (surface ORF) (Table 7). Considering that HBV has a very compact genome and overlapping ORFs covered the majority of the genome, HBV faced a considerable amount of selection pressure *in*

doi:10.6342/NTU202302511

vivo, causing the majority of nonsynonymous substitutions in HBV to be deleterious and subsequently removed, thus showing a strong negative selection signature.

After splitting polymerase into four domains, the terminal protein, reverse transcriptase, and RNase H are under strong negative selection, as expected. Interestingly, the spacer region shows a significant 204% increase in nonsynonymous mutations (Table 7), but surface ORF of the same region (site 2,860 to 163) shows a strong negative selection where 57% of nonsynonymous mutations gets removed. This can be explained by the positions of reading frames, where the first, second, and third codon position of surface ORF is the second, third, and first position, respectively, of spacer ORF. Thus synonymous mutations, usually occur at the third codon position, in spacer region can cause nonsynonymous mutations in surface ORF, which will be eliminated by selection. As a result, synonymous mutations in spacer region are reduced, which in turn cause high N/S ratio (Table 6B). This suggests that the significant increase in nonsynonymous mutations in spacer region is not a result of positive selection.

Different non-overlapping regions were analysed separately. As these regions contain only one reading frame, it is suggested by previous study that they are under less functional constrains [28]. Nevertheless, the strength of negative selection ranges from 65% to 95%. Therefore, in contrast to previous belief, the majority of non-overlapping regions are under strong selective constrain.

Mutation hotspots **in vivo** ***and*** **in vitro**

In both *in vitro* datasets, MinION dataset contains 103 highly variable sites (mutation rate $\geq 1 \times 10^{-4}$) while Sequel IIe dataset contains 25 sites; there are fifteen

doi:10.6342/NTU202302511

highly variable sites shared between them (Table 8A). Note that of all the seventeen types of substitution found, half of them are A→G.

By listing out the top 5% sites with the highest mutation rate from the *in vitro* (MinION and Sequel IIe) and *in vivo* datasets, only two sites were found sharing among all three datasets, site 85 and 3,073. When individually compare each *in vitro* dataset with *in vivo* dataset, MinION and Sequel IIe shared eleven and ten high mutation rate sites, respectively, with the *in vivo* dataset (Table 8B and 8C).

There are 1,244 highly conserved sites (mutation rate = 0) found *in vivo*, and finding whether any of them overlap with high mutation sites *in vitro* (mutation rate ≥ 1×10^{-4}) in both MinION and Sequel IIe datasets, four sites remained, site 740, 1,725, 1,920, and 2,619. All the mutations found *in vitro* for these four sites were nonsynonymous mutations (Table 9).

To identify whether there is any correlation between *in vitro* mutation frequency and *in vivo* nucleotide diversity, every site *in vitro* was sorted according to mutation frequency and split into per site, percentiles (a hundred groups), deciles (ten groups), or quartiles (four groups). Then, every site *in vivo* was sorted corresponding to the *in vitro*. On a per site basis, no correlation was found between these two datasets (Spearman's ρ: 0.02, p-value: 0.33; Figure 7A). Division by percentile, too, yielded a low correlation (Spearman's ρ: 0.25, p-value: 0.25; Figure 7B), and division by decile also has a rather low correlation (Spearman's ρ: 0.36, p-value: 0.39; Figure 7C). Further grouping into five groups, one unchanged and four quantiles (quartile), also results a low correlation (Spearman's ρ: 0.60, p-value: 0.35; Figure 7D). The correlations between *in vitro* and *in vivo* are low regardless of grouping size and is in contrast to a previous study on HCV where the correlation between *in vitro* mutation frequency and *in vivo* nucleotide diversity is found to be higher in both the percentile

doi:10.6342/NTU202302511

(Spearman's ρ: 0.42) and the quartile (Spearman's ρ: 1) [9]. This suggests that the nucleotide substitution pattern found *in vivo* HBV bears no resemblance to those *in vitro* and are the result of natural selection.

Mutation types found in each mutated site

Of all the 2,536 mutated sites in MinION Huh7 derived HBV, 1,830 sites contain only a single type of mutation, while 622 sites and 84 sites contain two and three types of mutation, respectively. This pattern is extremely similar to that of Sequel IIe Huh7 derived HBV with 2,533 mutated site, which coincidentally also has 1,830 sites with a single type of mutation, follow by two types of mutation in 615 sites, and 88 sites with three. In terms of nucleotide composition in sites that show three types of mutation, the pattern is too quite similar between MinION and Sequel IIe (MinION: T: 42, A: 18, G: 24, C: 0; Sequel IIe: T: 45, A: 10, G: 33, C: 0), and ten sites with three types of mutation were found sharing between MinION and Sequel IIe. The lack of cytosine site with three types of mutation was due to the removal of C→A mutation in both Huh7 derived HBV datasets.

Splice variants

The role that HBV splice variants play in the life cycle of HBV remains largely unknown, as they do not seems to be essential in transfection-based system [29], but clinical studies have shown that the proportion between circulating splice variants and wild type HBV has a positive correlation with disease progression and severity [17] [30] [31], albeit the proportion could varies wildly [32]. Given that this study utilises long-read sequencing with single-molecule resolution, the proportion of

doi:10.6342/NTU202302511

splice variants found can be precisely counted. Furthermore, as the *in vitro* transfection model used in this study minimised selection pressure, the proportion found here should represents the fundamental occurrence of each splice variant in genotype A HBV. Within the whole HBV 'population' sequenced in this study, 19 splice variants were found, which comprised 2.43% in the whole 'population', with sp1 as the most dominant splice variant, standing at 1.62% in the whole HBV 'population' or 66.65% in terms of splice variants' 'population'. The rest of the splice variants found were comparatively minuscule, which none of them single-handedly occupied more than 0.21%. 14 splice variants out of the 19 found were previously described splice variants and the other five are novel splice variants and shall be named sp23, sp24, sp25, sp26, and sp27, following previously described nomenclature [19]. All the positions of the splicing sites and their proportion within the splice variants 'population' discovered in this study are shown in Figure 8 and Table 10.

doi:10.6342/NTU202302511

Chapter 4: Discussion

Previous studies have estimated HBV evolutionary rate with varying orders of magnitude [33] [34] and these estimations were the results of selection. Besides, previous studies on full-length HBV lack the single-molecule and single-nucleotide resolution that this study possesses, which allows for a comprehensive analysis on HBV's mutations, heterogeneity, and splice variants with an even read depth across the genome.

This is the first study that directly detects and estimates the replication error rate of HBV in an environment without any selection pressure, as the *in vitro* model used in this study prevented HBV particles from reinfecting the cells, ensuring that every single Dane particle had only replicated once. Unlike previous study [9] that have estimated the replication error rate of HCV (3.5×10^{-5} nucleotide/replication), albeit done it indirectly, which utilised infection model and estimated the replication error rate by dividing the duration of HCV replication cycle. Comparing this study and the previous study [9], our average read depth was much deeper (> 80,000 versus 23,000) and has substantially more reads (> 750,000 versus 68,700). The HBV replication error rate estimated in this study (2.28×10^{-5} to 3.28×10^{-5} nucleotide/site/replication) shows that the virus has a replication error rate similar as retroviruses [35] [36] and RNA virus [9], and is several orders of magnitude above that of DNA viruses [4].

Although there was little correlation in genomic variation between the HBVdb inter-host HBV and both the Huh7 derived HBV datasets (Table 4, Figure 7), the nucleotide substitution profiles show strong correlations between them. Furthermore, no correlation was found in nucleotide variability between *in vivo* and *in vitro*. These

doi:10.6342/NTU202302511

evidences suggest that the nucleotide substitution profile of HBV is controlled by its polymerase and the genetic variation of the virus is governed by natural selection.

An obvious weakness in this study is the bias generated during PCR, which resulted in removing C→A transversion from the analysis, but this was an limitation imposed by the current technology as described by previous study [22]. This phenomenon was too described in other studies which sequenced PCR amplicons with Illumina platforms [23] [37], showing that this bias is not platform-specific. A silver lining from this is that we found out only the minus strand was used during UMI-tagging PCR.

Although the high amount of C→T and G→A substitution can be attributed to AID [26] and APOBEC3B [27], three transitions C→T, G→A, and A→G are showing similar level of mutation count (Table 11), suggesting something might have reduced the amount of T→C substitutions. Retroviruses have a slightly different nucleotide substitution profile, with G→A and A→G substitutions dominate the rest of the transitions and this was attributed to RNA-editing enzymes such as the APOBEC and the adenosine deaminase acting on RNA (ADAR) [37] [38] [39]. Curiously, in both HBV and HIV, T→C substitutions are the least frequent transition mutation type (Table 11).

Ultra-deep sequencing (average depth > 80,000) of near full-length HBV has allowed us to identify the splice variants at a highly precise manner and paved a way for us to identify new splice variants. Despite the fact that splice variants play no part in HBV replication in transfection-based system [29], their population proportion found in this study can be used as a benchmark of their occurrences in a neutral environment and compare with other genotypes in future studies. Another observation was made in this study, which two recombination events were found, one was

doi:10.6342/NTU202302511

recombined with a section of plasmid (in the plasmid data) and the other was found to had recombined with a section of human genome (in the *in vitro* data). As far as we know, these cannot be PCR chimaera, for *longread_umi* removes PCR chimaera. Their occurrences were immensely low (1 in > 80,000) and future research may be needed on the mechanism of their occurrence.

No study came without hurdle and this one brought plenty. During investigation, several methodological obstacles were faced and overcame. Most of the issues faced were stemmed from the fact that only a minuscule amount of HBV DNA can be used in the first PCR, i.e. UMI-tagging PCR, and was further exacerbated by the unusual structure of HBV DNA. The methodological obstacles faced in this study are written down here to explain the rationale of any specific method taken in this study instead of a 'better' conventional method that did not work.

Position for primer to anneal

The most commonly used primers to target and amplify full-length HBV are 5'-TTT TTC ACC TCT GCC TAA TCA and 5'-AAA AAG TTG CAT GGT GCT GG [13], but due to the overlapping design of this pair of primer, the PCR efficiency was found to be inadequate for this study. Previous studies used this primer to amplified HBV DNA with dozens of PCR cycles, but the nature of this study restricted the PCR cycle using HBV-specific primer down to two, effectively rendered this pair of primer useless. Several options were tried, such as changing annealing temperature, adding DMSO into PCR buffer, changing Taq polymerase, and using more than one type of Taq polymerases in PCR. None of these worked. The only effective method was to redesign the primer. By leaving a gap of 69 bp long, the tagging of UMI can be successfully performed with the amount of HBV DNA limited

doi:10.6342/NTU202302511

by the design of this study (10^5 to 10^6 copies). I hypothesised that this was due to the short repeat sequence in both terminal end of minus strand HBV DNA and the lack of helicase activity in Taq polymerase. Thus if a single minus strand DNA was to anneal by two primers at both terminal ends, the extension would be stopped when the Taq polymerase reached the position of another primer and blocked by it (Figure 5). The UMI-tagging primers are ~60 bp long, with a ~40 bp long tail dangling at the 5'-end that does not anneal onto HBV DNA. In addition, the plus strand of HBV DNA is incomplete and practically turning HBV DNA into a single-stranded DNA. When designing primer for HBV DNA, visualise the DNA as a linear DNA instead of a circular one and using the direction of the minus strand as template made the job easier. As shown in this study, simply avoiding the repeat sequence on one primer allowed us to tag UMI onto HBV DNA without a hitch, but the major drawback was losing information in the skipped region. Necessary sacrifice was needed in order to obtain the desired outcome.

Removing plasmid DNA

Due to the usage of lipofection, the original idea was that a simple detergent and nuclease were good enough to digest plasmid. Thus, the following treatment was used prior to the DNA extraction: 100 μL cell culture supernatant + 1 μL NP-40 + 0.1 μL micrococcal nuclease + 1 μL DNase I, incubate at 37°C overnight. But a disparity was noticed from the qPCR results (at this stage, the inadequacy of qPCR had not occurred) that there were several orders of magnitude more HBV DNA, per mL, extracted from the transfected cell culture than usual. This was too good to be true. Another hypothesis was raised that the detergent used, NP-40, was not strong enough to break open Lipofectamine 3000. To test this, pAAV/HBV1.2 plasmid was

doi:10.6342/NTU202302511

packaged into Lipofectamine 3000, as per manufacturer's protocol, and spilt to two groups, one was treated with NP-40 plus nuclease and the other was only treated with NP-40. After incubating overnight, the plasmid was extracted and a qPCR test was performed targeting ampicillin-resistance gene on pAAV/HBV1.2 plasmid and the result showed that both groups had pretty much the same amount of plasmid. This suggests that NP-40 was not strong enough of a detergent and the plasmid was protected inside the liposome. Stronger detergent was not considered because it might just denature the HBV nucleocapsid and has all the HBV DNA digested. In the end, a methylation-specific endonuclease DpnI was utilised for plasmid digestion. As DpnI only recognises and cleaves specific sequence, this generates another problem where fragments of plasmid were messing up the accuracy of qPCR.

Failure to quantify HBV DNA with qPCR

qPCR, in general, generates short segments of amplicon (~150 bp) for quantification which posed a problem as the plasmid fragments from DpnI were much longer than that, thus when quantifying HBV DNA with qPCR, fragments of plasmid inflated the amount of quantified DNA by a hundred fold. This was noticed when about 10^7 copies of HBV DNA were needed to performed a successful PCR, while previously only about 10^5 copies were required. Even by changing to a primer with DpnI cut site did not change anything. In the end, we opted for the less accurate but reliable full-length normal PCR and gel-based quantification. The detection of plasmid contamination was met with similar problem in qPCR and was resolved with the same way.

doi:10.6342/NTU202302511

In conclusion, this study shows that the replication error rate of HBV lies between 2.28×10^{-5} and 3.28×10^{-5} nucleotide/site/replication and has provided a genome-wide picture regarding the mutational landscape of HBV in a selectively neutral environment, which can be used as a baseline for uncovering the intensity of natural selection experienced by HBV in other environment.

doi:10.6342/NTU202302511

References

1. Walker, P.J., et al., *Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020).* Arch Virol, 2020. **165**(11): p. 2737-2748.
2. Organization, W.H. *Hepatitis B*. 2022 4 March 2023]; Available from: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b.
3. Collaborators, G.B.D.H.B., *Global, regional, and national burden of hepatitis B, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet Gastroenterol Hepatol, 2022. **7**(9): p. 796-829.
4. Sanjuan, R., et al., *Viral mutation rates.* J Virol, 2010. **84**(19): p. 9733-48.
5. Urban, S., et al., *The replication cycle of hepatitis B virus.* J Hepatol, 2010. **52**(2): p. 282-4.
6. Wei, L. and A. Ploss, *Core components of DNA lagging strand synthesis machinery are essential for hepatitis B virus cccDNA formation.* Nat Microbiol, 2020. **5**(5): p. 715-726.
7. Wei, L. and A. Ploss, *Hepatitis B virus cccDNA is formed through distinct repair processes of each strand.* Nat Commun, 2021. **12**(1): p. 1591.
8. Johnson, J.S., et al., *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.* Nat Commun, 2019. **10**(1): p. 5029.
9. Geller, R., et al., *Highly heterogeneous mutation rates in the hepatitis C virus genome.* Nat Microbiol, 2016. **1**(7): p. 16045.
10. Kennedy, S.R., et al., *Detecting ultralow-frequency mutations by Duplex Sequencing.* Nat Protoc, 2014. **9**(11): p. 2586-606.
11. Karst, S.M., et al., *High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing.* Nat Methods, 2021. **18**(2).
12. Huang, L.-R., et al., *An immunocompetent mouse model for the tolerance of human chronic hepatitis B virus infection.* Proceedings of the National Academy of Sciences, 2006. **103**(47): p. 17862-17867.
13. Gunther, S., et al., *A novel method for efficient amplification of whole hepatitis B virus genomes permits rapid functional analysis and reveals deletion mutants in immunosuppressed patients.* J Virol, 1995. **69**(9): p. 5437-44.
14. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
15. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.
16. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.
17. Lee, G.H., S. Wasser, and S.G. Lim, *Hepatitis B pregenomic RNA splicing--the products, the regulatory mechanisms and its biological significance.* Virus Res, 2008. **136**(1-2): p. 1-7.
18. Suzuki, Y., et al., *HBV preS deletion mapping using deep sequencing demonstrates a unique association with viral markers.* PLoS One, 2019. **14**(2): p. e0212559.
19. Kremsdorf, D., et al., *Alternative splicing of viral transcripts: the dark side of HBV.* Gut, 2021. **70**(12): p. 2373-2382.

doi:10.6342/NTU202302511

20. Hayer, J., et al., *HBVdb: a knowledge database for Hepatitis B Virus.* Nucleic Acids Res, 2013. **41**(Database issue): p. D566-70.
21. Drake, J.W., et al., *Rates of spontaneous mutation.* Genetics, 1998. **148**(4): p. 1667-86.
22. Filges, S., et al., *Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes.* Sci Rep, 2019. **9**(1): p. 3503.
23. Schmitt, M.W., et al., *Detection of ultra-rare mutations by next-generation sequencing.* Proc Natl Acad Sci U S A, 2012. **109**(36): p. 14508-13.
24. Tong, S. and J. Li, *Identification of NTCP as an HBV Receptor: The Beginning of the End or the End of the Beginning?* Gastroenterology, 2014. **146**(4): p. 902-905.
25. Delahaye, C. and J. Nicolas, *Sequencing DNA with nanopores: Troubles and biases.* PLoS One, 2021. **16**(10): p. e0257521.
26. Liang, G., et al., *RNA editing of hepatitis B virus transcripts by activation-induced cytidine deaminase.* Proc Natl Acad Sci U S A, 2013. **110**(6): p. 2246-51.
27. Chen, Y., et al., *APOBEC3B edits HBV DNA and inhibits HBV replication during reverse transcription.* Antiviral Res, 2018. **149**: p. 16-25.
28. Mizokami, M., et al., *Constrained evolution with respect to gene overlap of hepatitis B virus.* J Mol Evol, 1997. **44 Suppl 1**: p. S83-90.
29. Su, T.S., et al., *Hepatitis B virus transcript produced by RNA splicing.* J Virol, 1989. **63**(9): p. 4011-8.
30. Soussan, P., et al., *The expression of hepatitis B spliced protein (HBSP) encoded by a spliced hepatitis B virus RNA is associated with viral replication and liver fibrosis.* J Hepatol, 2003. **38**(3): p. 343-8.
31. Duriez, M., et al., *Alternative splicing of hepatitis B virus: A novel virus/host interaction altering liver immunity.* J Hepatol, 2017. **67**(4): p. 687-699.
32. Chen, J., et al., *Hepatitis B virus spliced variants are associated with an impaired response to interferon therapy.* Sci Rep, 2015. **5**: p. 16459.
33. Wang, H.Y., et al., *Distinct hepatitis B virus dynamics in the immunotolerant and early immunoclearance phases.* J Virol, 2010. **84**(7): p. 3454-63.
34. Zhou, Y. and E.C. Holmes, *Bayesian estimates of the evolutionary rate and age of hepatitis B virus.* J Mol Evol, 2007. **65**(2): p. 197-205.
35. Mansky, L.M. and H.M. Temin, *Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase.* J Virol, 1995. **69**(8): p. 5087-94.
36. Mansky, L.M., *Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line.* AIDS Res Hum Retroviruses, 1996. **12**(4): p. 307-14.
37. Rawson, J.M., et al., *HIV-1 and HIV-2 exhibit similar mutation frequencies and spectra in the absence of G-to-A hypermutation.* Retrovirology, 2015. **12**: p. 60.
38. Abram, M.E., et al., *Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication.* J Virol, 2010. **84**(19): p. 9864-78.
39. Sun, N. and S.S. Yau, *In-depth investigation of the point mutation pattern of HIV-1.* Front Cell Infect Microbiol, 2022. **12**: p. 1033481.
40. Cromer, D., et al., *HIV-1 Mutation and Recombination Rates Are Different in Macrophages and T-cells.* Viruses, 2016. **8**(4): p. 118.

doi:10.6342/NTU202302511

Tables

Table 1 Primers' sequences and their pairing. (A) Primers used in this study and their sequences. Primer binding start and end sites are numbered according to genotype A HBV. (B) The different combination of primer pairs for each dataset with their respective amplicon size.

A

Primer name	Primer sequence	Start	End
HBV_2kb_Tag_F	CAAGCAGAAGACGGCATACGAGAT NNNYRNNNYRNNNYRNNN ATTCGCACTCCTCCAGCTTA	2,276	2,295
HBV_2kb_Tag_R	AATGATACGGCGACCACCGAGATC NNNYRNNNYRNNNYRNNN GTTGGCGAGAAAGTGAAAGC	1,087	1,106
HBV_3kb_Tag_F	AATGATACGGCGACCACCGAGATC NNNYRNNNYRNNNYRNNN GTCCTACTGTTCAAGCCTCCA	1,854	1,874
HBV_3kb_Tag_R1	CAAGCAGAAGACGGCATACGAGAT NNNYRNNNYRNNNYRNNN TGCCTACAGCCTCCTAGTACA	1,769	1,789
HBV_3kb_Tag_R2	CAAGCAGAAGACGGCATACGAGAT NNNYRNNNYRNNNYRNNN AAAAAGTTGCATGGTGCTGG	1,806	1,825
HBV_Amp_F	CAAGCAGAAGACGGCATACGAGAT	N/A	N/A
HBV_Amp_R	AATGATACGGCGACCACCGAGATC	N/A	N/A
pAAV/HBV1.2_Backbone_F	TCCTTGAGAGTTTTCGCCCC	N/A	N/A
pAAV/HBV1.2_Backbone_R	GCCCACTACGTGAACCATCA	N/A	N/A
SP-5 (HBV +2413)	CCGCGTCGCAGAAGATCT	2,417	2,434
HBV -2551	GGAAADGADGGRGTTTKCCA	2,532	2,551
	N: A/T/G/C; Y:C/T; R: A/G		

B

Sequencing platforms & DNA	Amplicon size (bp)	Forward	Reverse
MinION plasmid	2,010	HBV_2kb_Tag_F	HBV_2kb_Tag_R
Sequel I plasmid	3,115	HBV_3kb_Tag_F1	HBV_3kb_Tag_R
MinION Huh7 derived HBV	3,152	HBV_3kb_Tag_F2	HBV_3kb_Tag_R
Sequel IIe Huh7 derived HBV	3,152	HBV_3kb_Tag_F2	HBV_3kb_Tag_R

doi:10.6342/NTU202302511

Table 2 Detailed information of every dataset and their original nucleotide substitution profile. Each DNA was sequenced with two different sequencing platforms. (A) Details from sequencing pAAV/HBV1.2. (B) Details from sequencing Huh7 derived HBV. The comparatively higher standard deviation (SD) for the average read depth was due to the existence of splice variants in Huh7 derived HBV.

	A				B			
	pAAV/HBV1.2 plasmid				Huh7 derived HBV			
Sequencing platforms	Nanopore MinION		PacBio Sequel I		Nanopore MinION		PacBio Sequel IIe	
Number of cell (s)	1		1		2		1	
Raw data size	14 Gb		0.7 Gb		45 Gb		8.8 Gb	
Total reads	5,866,550		232,781		14,005,936		2,711,133	
Reads used	2,075,261		32,252		1,758,783		752,044	
Total bases	40,224,120		10,762,325		260,268,297		280,482,121	
Total UMI	20,012		3,455		84,427		89,973	
Average depth	20,011 (SD: 24)		3,455 (SD: 0.03)		82,573 (SD: 2,282)		88,986 (SD: 1,084)	
Number of mutations (all)	823		818		11,905		8,452	
Mutations and count (all)	C→A	322	C→A	532	C→A	2,275	C→A	2,067
	G→T	306	G→T	271	C→T	2,015	C→T	1,573
	T→C	63	G→A	7	A→G	1,961	G→A	1,391
	G→A	44	C→T	5	G→A	1,929	A→G	1,021
	C→T	27	A→G	2	T→C	1,125	T→C	779
	A→G	18	C→G	1	T→G	1,038	T→G	464
	A→C	17			A→T	449	A→T	380
	T→A	12			T→A	325	G→C	332
	T→G	8			G→C	279	T→A	260
	A→T	3			A→C	228	G→T	102
	C→G	2			G→T	194	C→G	47
	G→C	1			C→G	87	A→C	36

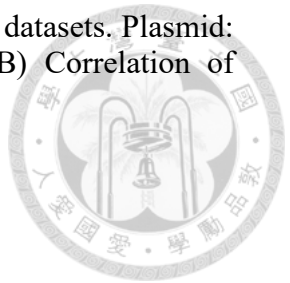
doi:10.6342/NTU202302511

Table 3 Mutations finally used for analyses in this study. (A) In both pAAV/HBV1.2 plasmid datasets, C→A and G→T mutations were removed. (B) In Huh7 derived HBV, C→A and 'plasmid hotspots' mutations were removed from MinION dataset and only C→A mutations were removed from Sequel IIe dataset. (C) The genotype A inter-host HBV dataset was downloaded from HBVdb and has C→A mutations removed.

	A				**B**				**C**	
	pAAV/HBV1.2 plasmid				**Huh7 derived HBV**				**HBVdb genotype A inter-host HBV (n=970)**	
Sequencing platforms	Nanopore MinION		PacBio Sequel I		Nanopore MinION		PacBio Sequel IIe		N/A	
Number of mutations (used)	195		15		8,549		6,385		68,174	
Transition:transversion	3.53		14.00		3.16		2.94		2.27	
Error rate	4.85×10^{-6}		1.39×10^{-6}		3.28×10^{-5}		2.28×10^{-5}		N/A	
Mutations and count (used)	T→C	63	G→A	7	C→T	1,984	C→T	1,573	C→T	13,047
	G→A	44	C→T	5	A→G	1,845	G→A	1,391	A→G	12,538
	C→T	27	A→G	2	G→A	1,750	A→G	1,021	G→A	11,103
	A→G	18	C→G	1	T→C	917	T→C	779	T→C	10,626
	A→C	17			T→G	566	T→G	464	A→C	4,921
	T→A	12			A→T	447	A→T	380	A→T	4,471
	T→G	8			T→A	323	G→C	332	T→A	3,712
	A→T	3			G→C	277	T→A	260	G→T	3,328
	C→G	2			G→T	194	G→T	102	T→G	2,278
	G→C	1			A→C	159	C→G	47	G→C	1,591
					C→G	87	A→C	36	C→G	559

doi:10.6342/NTU202302511

Table 4 Correlation between different datasets. (A) Shows the correlation of mutation distribution between different datasets. Plasmid: pAAV/HBV1.2, Huh7: Huh7 derived HBV, HBVdb: inter-host genotype A HBV downloaded from HBVdb. (B) Correlation of nucleotide substitution profile between different datasets. Asterisk (*) indicates significance (p-value < 0.05).



A

Mutation distributions correlation (Spearman's ρ)				
	ONT plasmid	**PacBio plasmid**	**ONT Huh7**	**PacBio Huh7**
PacBio plasmid	-0.40 *			
ONT Huh7	0.27 *	-0.01		
PacBio Huh7	0.15 *	-0.17 *	0.32 *	
HBVdb	0.08	-0.04	0.26 *	0.11

B

Nucleotide substitution profiles correlation (Pearson's r)		
	ONT Huh7	**PacBio Huh7**
PacBio Huh7	0.97 *	
HBVdb	0.92 *	0.88 *

doi:10.6342/NTU202302511

Table 5 Nonsynonymous to synonymous ratio (N/S ratio) of Huh7 derived HBV. The expected N/S was calculated from reference genome, while the observed N/S ratio came from Huh7 derived HBV PacBio Sequel IIe. Panel A shows N/S ratio with C→A mutations removed and Panel B shows C→A/T mutations removed. In Panel A, the observed N/S ratios of polymerase (P gene) and core (C gene) are significantly lower than the expected, but removing C→T (Panel B) shows that only P gene remain significantly lower. By further splitting P gene into functional domains, all are consistent with neutral expectation. Non-overlap: non-overlapping reading frame, numbers indicate start and end of said reading frame. Overlap: overlapping reading frame, the first alphabet indicates the reading frame of the gene, numbers indicate start and end of said reading frame. Asterisk (*) indicates significance (p-value < 0.05) calculated from Fisher's exact test.

A

	C→A mutations removed						
	Reference (expected)			Huh7 derived HBV (observed)			
Genes	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
C gene	376.33	132.67	2.84	960	316	3.04	0.55
P gene	1658.67	551.67	3.01	3548	1434	2.47	8×10^{-4} *
S gene	772.67	268.67	2.88	1674	737	2.27	4.9×10^{-3} *
X gene	274.33	97.33	2.82	595	238	2.50	0.4038
Polymerase	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
Terminal protein	368.33	114.33	3.22	809	316	2.56	0.07
Spacer	342.00	109.00	3.14	759	303	2.50	0.09
Reverse transcriptase	667.33	213	3.13	1457	552	2.64	0.07
RNase H	281	115.33	2.44	523	263	1.99	0.13
Non-overlap	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
835_1374	358	117	3.06	727	296	2.46	0.09
1623_1802	123.67	35	3.53	312	103	3.03	0.51
1838_2305	286.33	94.33	3.04	732	248	2.95	0.89
2460_2852	263.33	80	3.29	601	216	2.78	0.30
Overlap	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
X_P_1374_1622	150.67	62.33	2.42	283	135	2.10	0.47
P_X_1375_1623	145.67	64.33	2.26	267	151	1.77	0.18
P_C_2307_2456	98	32.67	3.00	198	94	2.11	0.17
C_P_2306_2458	90	38.33	2.35	228	68	3.35	0.15

B

	C→A/T mutations removed						
	Reference (expected)			Huh7 derived HBV (observed)			
Genes	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
C gene	345.67	118	2.93	742	244	3.04	0.79
P gene	1523	466.33	3.27	2686	959	2.80	0.018 *
S gene	704.67	226.33	3.11	1291	488	2.65	0.08
X gene	249.33	79.67	3.13	450	151	2.98	0.81
Polymerase	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
Terminal protein	342.33	100.67	3.40	626	227	2.76	0.14
Spacer	313.00	87.00	3.60	572	188	3.04	0.28
Reverse transcriptase	619.67	182.67	3.39	1144	395	2.90	0.13
RNase H	248	96	2.58	344	149	2.31	0.49
Non-overlap	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
835_1374	358	117	3.06	331.67	102	3.25	0.13
1623_1802	123.67	35	3.53	118	28.67	4.12	0.72
1838_2305	286.33	94.33	3.04	266.33	82.67	3.22	0.82
2460_2852	263.33	80	3.29	245.33	73.67	3.33	0.21
Overlap	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
X_P_1374_1622	176.67	72.33	2.44	164	73	2.25	0.59
P_X_1375_1623	145.67	64.33	2.26	155	82	1.89	0.29
P_C_2307_2456	112	38	2.95	141	45	3.13	0.68
C_P_2306_2458	90	38.33	2.35	141	49	2.88	0.36

doi:10.6342/NTU202302511

Table 6 N/S ratio of inter-host genotype A HBV downloaded from HBVdb. Every single gene, open reading frame, and domain show a significant difference from the expected, demonstrating the effect of natural selection. All types of mutation were kept and used in this specific dataset. Asterisk (*) indicates significance (p-value < 0.05) calculated from Fisher's exact test. Non-overlap: non-overlapping reading frame, numbers indicate start and end of said open reading frame. Overlap: overlapping reading frame, the first alphabet indicates the open reading frame of the gene, numbers indicate start and end of said open reading frame.

	A			B			
	Reference (expected)			HBVdb genotype A inter-host HBV (observed)			
Genes	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
C gene	432.83	149.17	2.90	1858	10095	0.18	$< 1\times10^{-5}$ *
P gene	1904.67	630.33	3.02	27159	33495	0.81	$< 1\times10^{-5}$ *
S gene	898.84	298.16	3.01	12780	11737	1.09	$< 1\times10^{-5}$ *
X gene	317.17	111.83	2.84	3286	2859	1.15	$< 1\times10^{-5}$ *
Polymerase	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
Terminal protein	422.67	129.33	3.27	4349	11707	0.37	$< 1\times10^{-5}$ *
Spacer	396.83	128.17	3.10	14829	1949	7.61	$< 1\times10^{-5}$ *
Reverse transcriptase	757.5	241.5	3.14	5938	12877	0.46	$< 1\times10^{-5}$ *
RNase H	327.67	131.33	2.50	2043	6962	0.29	$< 1\times10^{-5}$ *
Non-overlap	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
835_1374	407.17	132.83	3.07	1815	14212	0.13	$< 1\times10^{-5}$ *
1623_1802	140.5	39.5	3.56	1185	1117	1.06	$< 1\times10^{-5}$ *
1838_2305	327	105	3.11	1518	9163	0.17	$< 1\times10^{-5}$ *
2460_2852	303.33	89.67	3.38	3309	11472	0.29	$< 1\times10^{-5}$ *
Overlap	**Non-syn**	**Syn**	**N/S**	**Non-syn**	**Syn**	**N/S**	***p*-value**
X_P_1374_1622	176.67	72.33	2.44	2101	1742	1.21	$< 1\times10^{-5}$ *
P_X_1375_1623	173.33	72.67	2.39	1802	2040	0.88	$< 1\times10^{-5}$ *
P_C_2307_2456	112	38	2.95	1034	233	4.44	0.04*
C_P_2306_2458	105.83	44.17	2.40	340	932	0.36	$< 1\times10^{-5}$ *
S_P_2860_163	392	130	3.02	9145	7541	1.21	$< 1\times10^{-5}$ *

doi:10.6342/NTU202302511

Table 7 The percentage of nonsynonymous mutations lost *in vivo* due to selection. Apart from two regions, every other region shows a strong purifying selection. The spacer in polymerase functions as a bridge and does not play any role in HBV DNA replication. The overlapping region between 2,307 and 2,456 contains immune epitope for the HBV polymerase, thus showing a positive selection signature. Non-overlap: non-overlapping reading frame, numbers indicate start and end of said open reading frame. Overlap: overlapping reading frame, the first alphabet indicates the open reading frame of the gene, numbers indicate start and end of said open reading frame.

Genes	**Mutations lost *in vivo***
C gene	-93.94%
P gene	-67.23%
S gene	-52.06%
X gene	-54.03%

Polymerase	**Mutations lost *in vivo***
Terminal protein	-85.49%
Spacer	+203.74%
Reverse transcriptase	-53.89%
RNase H	-85.24%

Non-overlap	**Mutations lost *in vivo***
835_1374	-94.80%
1623_1802	-64.98%
1838_2305	-94.39%
2460_2852	-89.63%

Overlap	**Mutations lost *in vivo***
X_P_1374_1622	-42.47%
P_X_1375_1623	-54.82%
P_C_2307_2456	+110.68%
C_P_2306_2458	-89.12%
S_P_2860_163	-56.96%

doi:10.6342/NTU202302511

Table 8 Highly variable sites found *in vivo* and *in vitro*. (A) In both *in vitro* datasets, fifteen highly variable sites (mutation rate ≥ 1×10^{-4}) are shared between them. (B) The highly variable sites (top 5%) shared between *in vivo* and ONT MinION datasets. (C) The highly variable sites (top 5%) shared between *in vivo* and PacBio Sequel IIe datasets. Site 85 and 3,073 were the only two sites that are shared among these three datasets.

A

Sites	Huh7 derived HBV	
	MinION	Sequel IIe
357	T→A	T→A
710	C→T	C→T
740	T→G	T→G
1,657	A→G	A→G
1,725	A→G	A→G
1,755	A→G	A→G
1,760	A→G	A→G
1,892	C→T	C→T
1,917	A→G	A→G
1,919	A→G	A→G
1,920	A→T, A→G	A→T, A→G
1,940	G→A	G→A
2,619	G→A, G→T	G→A
2,774	T→G	T→G
3,190	A→G	A→G

B

Sites	HBVdb	Huh7 derived HBV MinION
46	T→A, T→C	T→G, T→A, T→C
85	A→G, A→C	A→T, A→G
128	G→A, G→T, G→C	G→A
286	A→T, A→G, A→C	A→T
287	T→G, T→A, T→C	T→G, T→A, T→C
1,464	T→G, T→A, T→C	T→G, T→A, T→C
2,486	C→T	C→T
2,600	G→A, G→T, G→C	G→A
3,073	A→T, A→G, A→C	A→T, A→G
3,121	A→T, A→G, A→C	A→G, A→C
3,124	A→G	A→T, A→G

C

Sites	HBVdb	Huh7 derived HBV Sequel IIe
85	A→C, A→G	A→T, A→G
97	G→A, G→C	G→A
1,023	T→A, T→G, T→C	T→C
1,512	G→A, G→C, G→T	G→A, G→C
2,239	G→A, G→C, G→T	G→A, G→C
2,242	T→A, T→G, T→C	T→A, T→G, T→C
2,741	A→G	A→T, A→G
2,913	C→T	C→T
3,069	C→T, C→G	C→T, C→G
3,073	A→T, A→C, A→G	A→T, A→G

doi:10.6342/NTU202302511

Table 9 Sites that are highly conserved *in vivo* but highly variable *in vitro*. aa: amino acid. P: polymerase, S: surface, X: X protein, C: core.

Genes		P		S		X		C	
Genome position	***in vitro* mutation**	**aa position**	**aa change**	**aa position**	**aa change**	**aa position**	**aa change**	**aa position**	**aa change**
740	T→G	552	Met→Arg	196	Trp→Gly				
1,725	A→G					118	Lys→Glu		
1,920	A→T							36	Lys→Ile
	A→G							36	Lys→Arg
2,619	G→A	105	Glu→Lys						
	G→T	105	Glu→Stop						

doi:10.6342/NTU202302511

Table 10 Splice variants and their respective 'population size' percentages in the whole HBV 'population' and within splice variants 'population'. Dagger (†) indicates novel splice variants.

Splice variants	Whole population (%)	Splice variants population (%)
Wild type	97.5658	N/A
sp1	1.6224	66.6512
sp2	0.1317	5.4117
sp3	0.2015	8.2794
sp6	0.0856	3.5153
sp7	0.0597	2.4514
sp8	0.0315	1.2951
sp9	0.1407	5.7817
sp10	0.0304	1.2488
sp11	0.0034	0.1388
sp13	0.0203	0.8326
sp14	0.0225	0.9251
sp15	0.0281	1.1563
sp16	0.0135	0.5550
sp19	0.0045	0.1850
sp23†	0.0214	0.8788
sp24†	0.0034	0.1388
sp25†	0.0056	0.2313
sp26†	0.0034	0.1388
sp27†	0.0045	0.1850

doi:10.6342/NTU202302511

Table 11 Comparison of nucleotide substitution profiles between HBV from this study and HIV-1 from other studies. (A) Percentage of transitions in its respective dataset. In HBV, C→T substitutions are the most abundant substitution type, unlike HIV-1, which is G→A. In HIV-1, the G→A and A→G hypermutations are attributed to RNA-editing enzyme, such as ADAR and APOBEC. T→C substitution remains the least frequent substitution type in both viruses. (B) Percentage of transversions in its respective dataset. Parentheses indicate number of mutations.

A

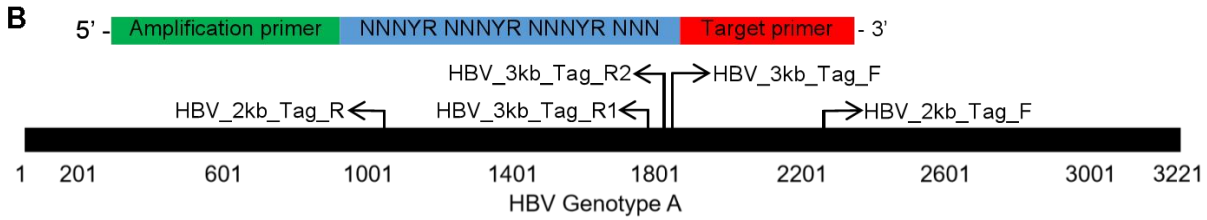
Virus	G→A	A→G	C→T	T→C	Reference
HBV	20.5% (1,750)	21.6% (1,845)	23.2% (1,984)	10.7% (917)	This study (MinION)
HBV	21.8% (1,391)	16.0% (1,021)	24.6% (1,573)	12.2% (779)	This study (Sequel IIe)
HBV	15.1% (11,103)	17.0% (12,538)	17.7% (13,047)	14.4% (10,626)	This study (HBVdb)
HIV-1	21.1% (2,122,321)	19.8% (1,993,745)	11.1% (1,113,518)	10.9% (1,096,809)	[39]
HIV-1	36.1% (499)	16.6% (230)	16.2% (223)	10.0% (138)	[40]

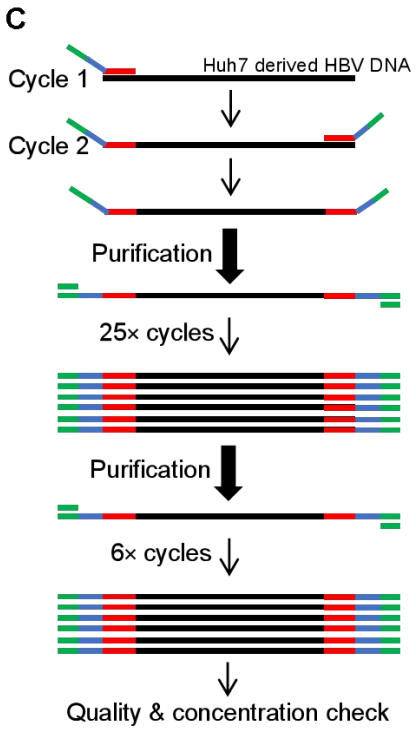
B

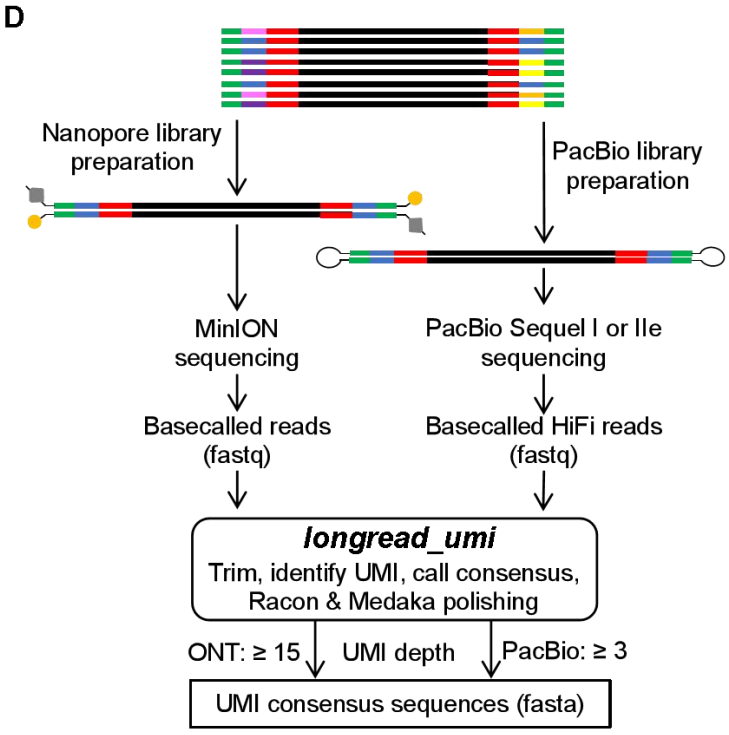
Virus	C→A	A→C	T→A	A→T	T→G	G→T	G→C	C→G	Reference
HBV	N/A	1.9% (159)	3.8% (323)	5.2% (447)	6.6% (566)	2.3% (194)	3.2% (277)	1.0% (87)	This study (MinION)
HBV	N/A	0.6% (36)	4.1% (260)	6.0% (380)	7.3% (464)	1.6% (102)	5.2% (332)	0.7% (47)	This study (Sequel IIe)
HBV	7.4% (5,485)	6.7% (4,921)	5.0% (3,712)	6.1% (4,471)	3.1% (2,278)	4.5% (3,328)	2.2% (1,591)	0.8% (559)	This study (HBVdb)
HIV-1	7.4% (746,010)	6.6% (659,021)	5.3% (532,914)	5.1% (512,784)	3.5% (351,074)	3.2% (317,821)	3.1% (314,199)	2.9% (294,147)	[39]
HIV-1	6.2% (85)	2.1% (28)	1.2% (17)	2.7% (37)	2.5% (35	4.3% (59)	1.1% (15)	1.3% (17)	[40]

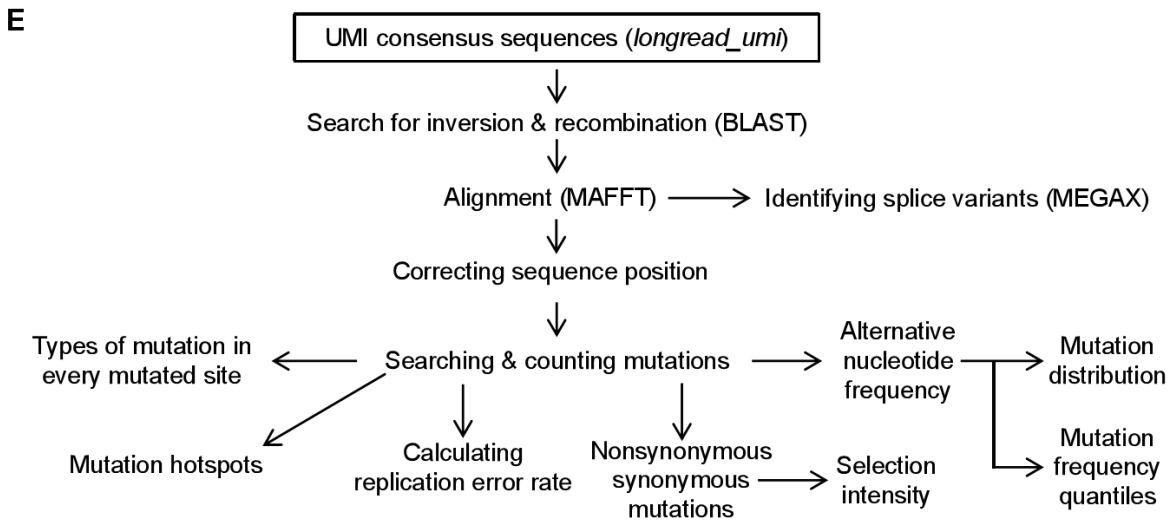
doi:10.6342/NTU202302511

Figures









doi:10.6342/NTU202302511

Figure 1 Overview of the system utilised in this study.
(A) Dane particles were generated by transfecting pAAV/HBV1.2 plasmid into Huh7 cells. After DNA was extracted from cell culture supernatant, plasmid was cleaved with a methylation-specific restriction enzyme, DpnI.
(B) Design of the unique molecular identifier (UMI) tagging primers and their binding sites in genotype A HBV genome. Green: amplification primer sequence that ensures only DNA molecules tagged with UMI can be amplified. Blue: semi-random UMI sequence, N: A/T/C/G, Y: T/C, R: A/G. Red: HBV-specific target sequence.
(C) Schematics of PCR reaction. Two PCR cycles for UMI-tagging PCR, follow by purification and 25 cycles of amplification PCR. Another round of purification was performed and the amplicons were amplified with PCR (six cycles). Finally, amplicons were purified again and have their quality and concentration were checked by 1.5% agarose gel (TAE) electrophoresis and Nanodrop, respectively.
(D) Purified amplicons were subjected to library preparation prior to long-read sequencing. Basecalled reads were input into a pipeline called *longread_umi* [11]. The output was a file containing all the consensus sequence of every valid UMI (read depth ≥ 15 for ONT; ≥ 3 for PacBio).
(E) Flowchart of the data analysis performed in this study. Tools that were used are stated in parentheses and the rest were done by custom python scripts.

doi:10.6342/NTU202302511

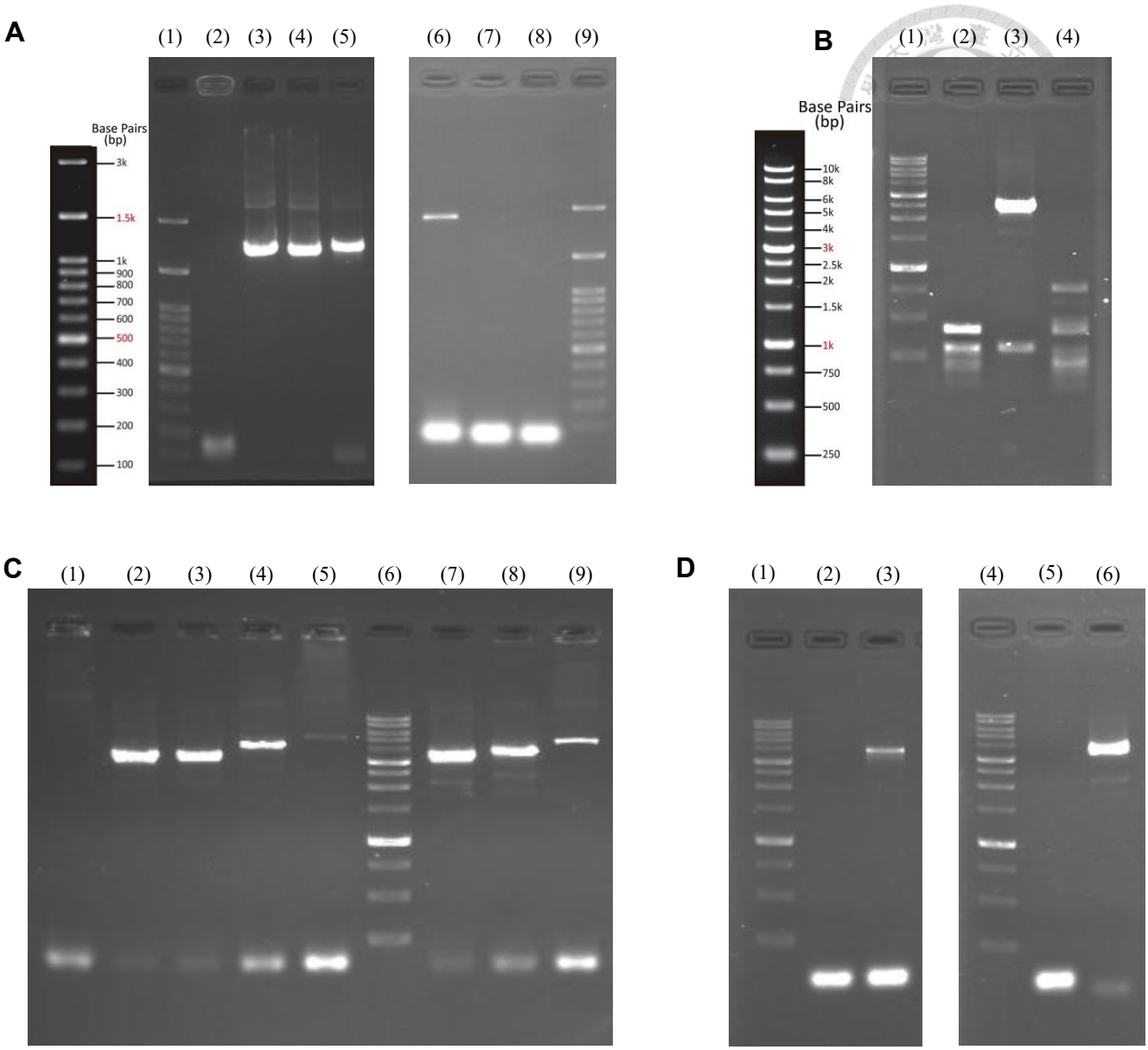


Figure 2 1.5% agarose gel (TAE) electrophoresis of amplicons.
(A) Testing the lowest possible required DNA copy number of initial DNA for PCR. (1) marker, (2) negative control, (3) 5×10^7, (4) 5×10^6, (5) 5×10^5, (6) 1×10^5, (7) 5×10^4, (8) negative control, (9) marker.
(B) Check the possible contamination of plasmid DNA by PCR with vector-specific primers. If plasmid is present, a 2,604 bp amplicon will be generated: (1) marker, (2) negative control, (3) positive control (5×10^5 copies of plasmid), (4) DpnI-treated Huh7 derived HBV DNA.
(C) Gel-based quantification: AD38 derived HBV DNA standard (copy number): (1) negative control (2) 7×10^5, (3) 7×10^4, (4) 7×10^3, (5) 7×10^2; (6) marker; serial diluted Huh7 derived HBV DNA: (7) no dilution, (8) 10× dilution, (9) 100× dilution. The apparent difference in amplicon size is due to the DNA dye used in this study affects the DNA's overall electrical charge and subsequently affects the migration of DNA if the concentration varies wildly.
(D) Huh7 derived HBV DNA tagged amplicon: (1) marker, (2) negative control, (3) 1st amplification, (4) marker, (5) negative control, (6) 2nd amplification.

doi:10.6342/NTU202302511

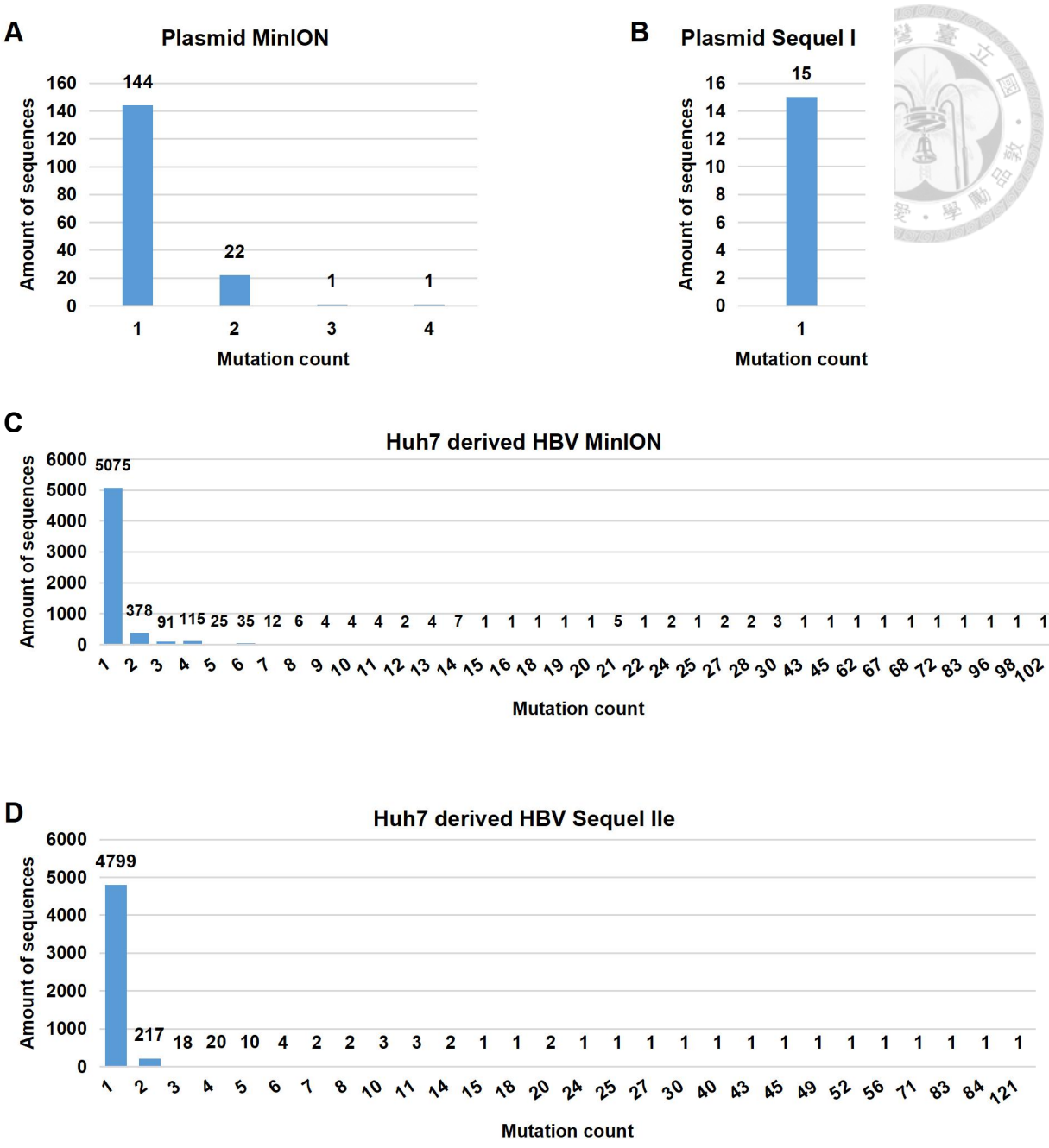


Figure 3 The amount of mutation in every molecule.

(A) In MinION plasmid sequencing dataset, 144 sequences exhibit one mutation, 22 sequences show two mutations, while one sequence each has three and four mutations, respectively.

(B) All fifteen mutations from Sequel I plasmid dataset are from unique sequence.

(C) The sequencing of Huh7 derived HBV with MinION shows that the majority (59%) of mutated sequences have only experienced a single mutation.

(D) The Huh7 derived HBV sequenced with Sequel IIe dataset shows that the three-quarters (75.16%) of mutated sequences contain one mutation.

doi:10.6342/NTU202302511

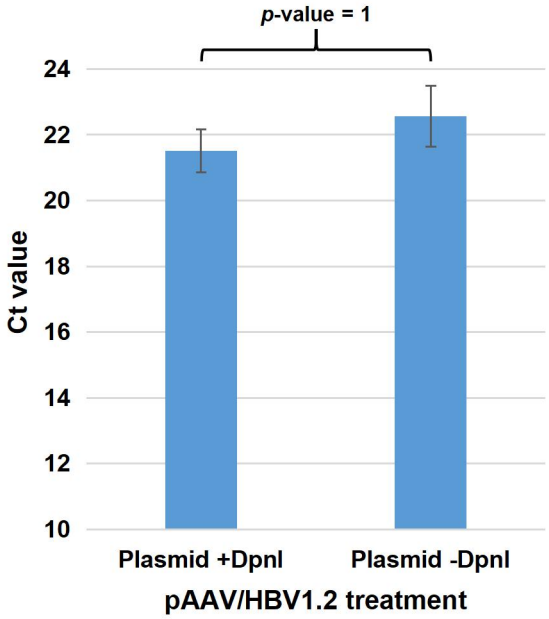


Figure 4 Ct value from qPCR of DpnI-treated and untreated pAAV/HBV1.2.
Triplicates qPCR results show that DpnI-digested plasmid cannot be meaningfully identified from untreated plasmid (Fisher's exact test *p*-value = 1). DpnI-treated plasmid (Plasmid +DpnI), average: 21.51, SD: 0.65. Untreated plasmid (Plasmid -DpnI), average: 22.57, SD: 0.92. The primers used in the qPCR were SP-5 (HBV +2413) and HBV -2551.

doi:10.6342/NTU202302511

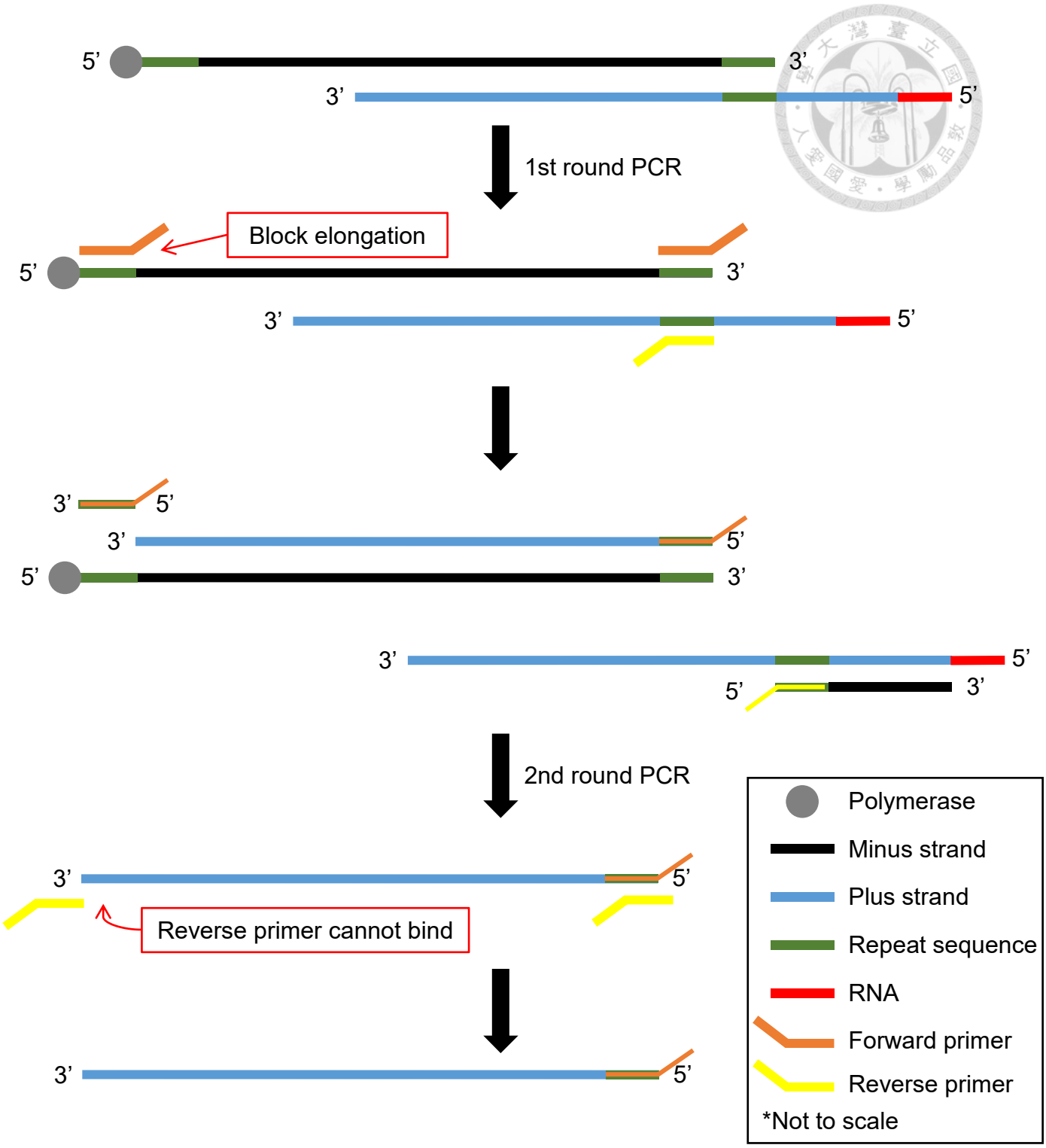


Figure 5 Schematics showing the problem with overlapping primers.
During UMI tagging PCR, the binding sites of previously described full-length HBV primers (forward: 5’-TTT TTC ACC TCT GCC TAA TCA and reverse: 5’-AAA AAG TTG CAT GGT GCT GG) [14] overlapped and blocked the synthesis of plus strand, thus creating an amplicon without any binding site for the reverse primer in the second round of PCR and ultimately an amplicon with only a single UMI. Two UMIs, one at each terminal end, were required to be qualified as valid by *longread_umi*. The original minus strand was the only usable template, due to the original plus strand was not a complete genome. By moving the binding site of the reverse primer several base pair to the 5’ direction has allowed us to tag two UMIs onto a single DNA.

doi:10.6342/NTU202302511

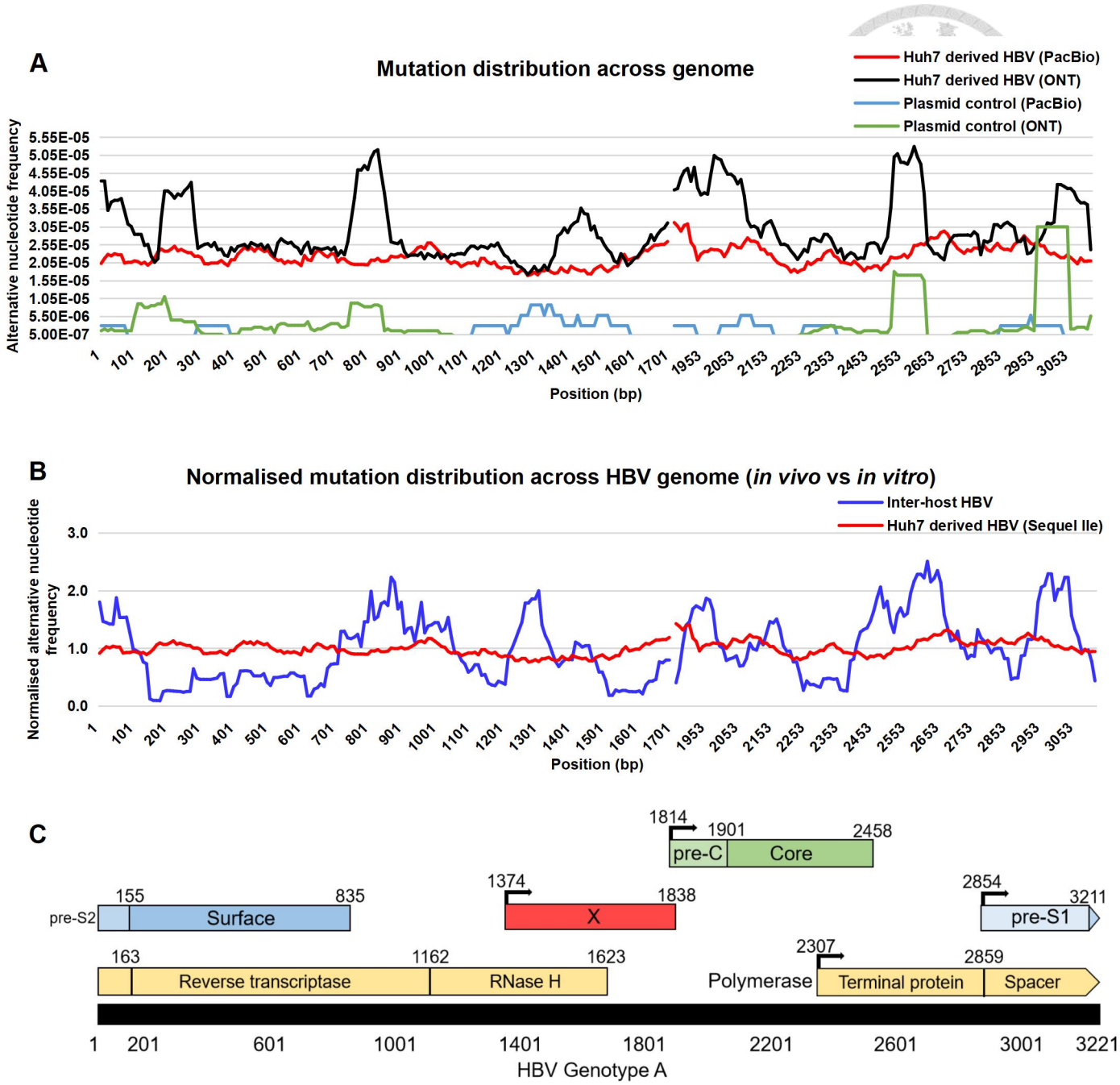


Figure 6 Genetic variation across genotype A HBV genome.

(A) The mutation distributions of all four datasets showing together with sliding window (window size = 100 bp, step size = 10 bp). The PacBio datasets do not exhibit mutational hotspots as observed in both MinION datasets (ONT).

(B) A normalised mutation distribution of inter-host genotype A HBV downloaded from HBVdb (blue) and PacBio Sequel IIe Huh7 derived HBV (red). Sliding window size = 100 bp, step size = 10 bp.

(C) The genome structure of genotype A HBV with all the genes/open reading frames and their respective positions. Polymerase gene (coloured in beige) are further split into four functional domains.

doi:10.6342/NTU202302511

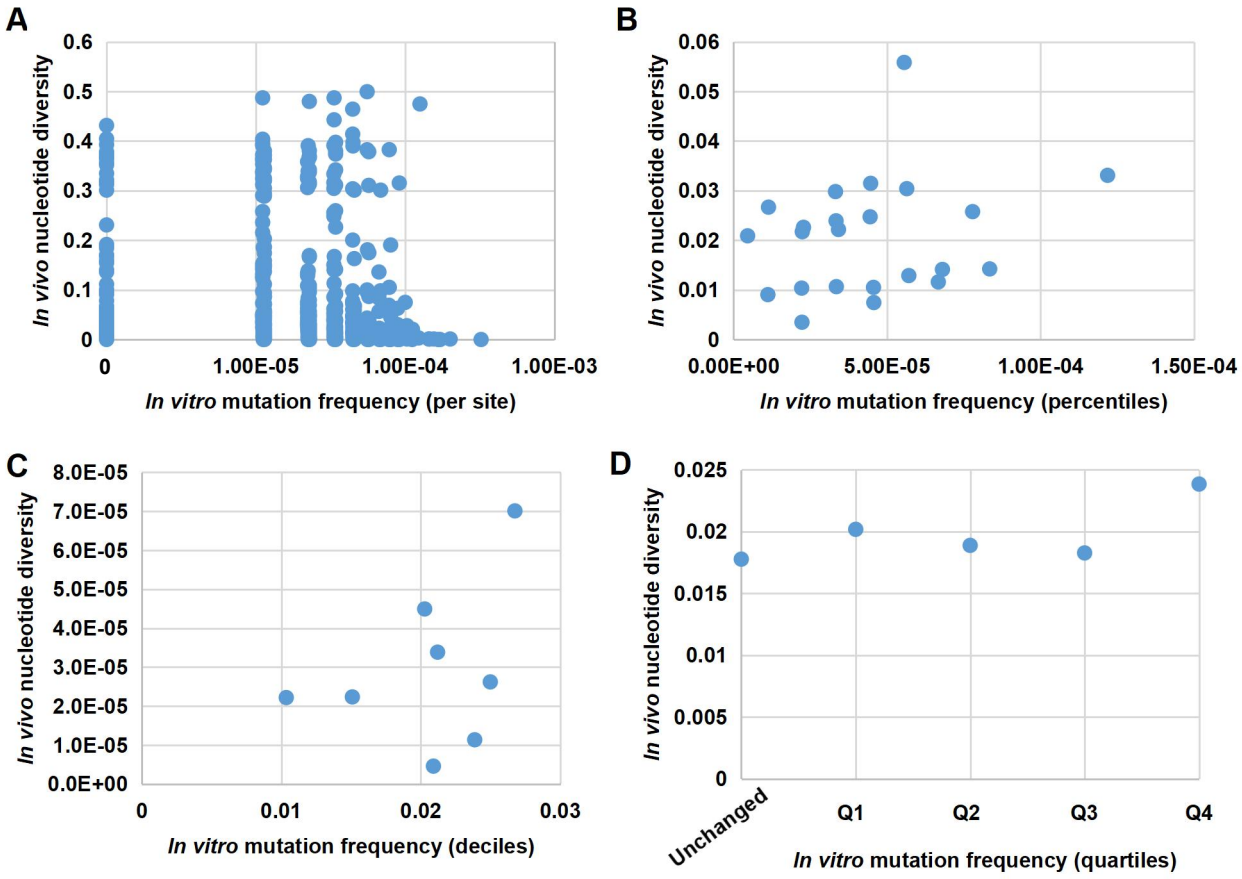


Figure 7 Scatter plots of *in vitro* mutation frequency versus *in vivo* nucleotide diversity for each genome site.

(A) Sort *in vitro* mutation frequency on a per site basis and compare the nucleotide of the same site *in vivo*. No correlation in nucleotide variability was found between these two datasets (Spearman's ρ: 0.02, p-value: 0.33). X-axis is shown in log scale.

(B) Sort *in vitro* mutation frequency into 100 groups (percentiles) and compare the nucleotide diversity of same site *in vivo*. There is little correlation between them (Spearman's ρ: 0.25, p-value: 0.25).

(C) The same plot as (B) but split into 10 groups (deciles) shows that the correlation remains low (Spearman's ρ: 0.36, p-value: 0.39).

(D) The same plot as (B) but *in vitro* mutation frequency were grouped into five categories, four quantiles (quartiles) and an unchanged (no mutation *in vitro*). The correlation among them are not significant (Spearman's ρ: 0.60, p-value: 0.35), showing that high mutation frequency sites *in vitro* do not correlates with high *in vitro* nucleotide diversity sites.

doi:10.6342/NTU202302511

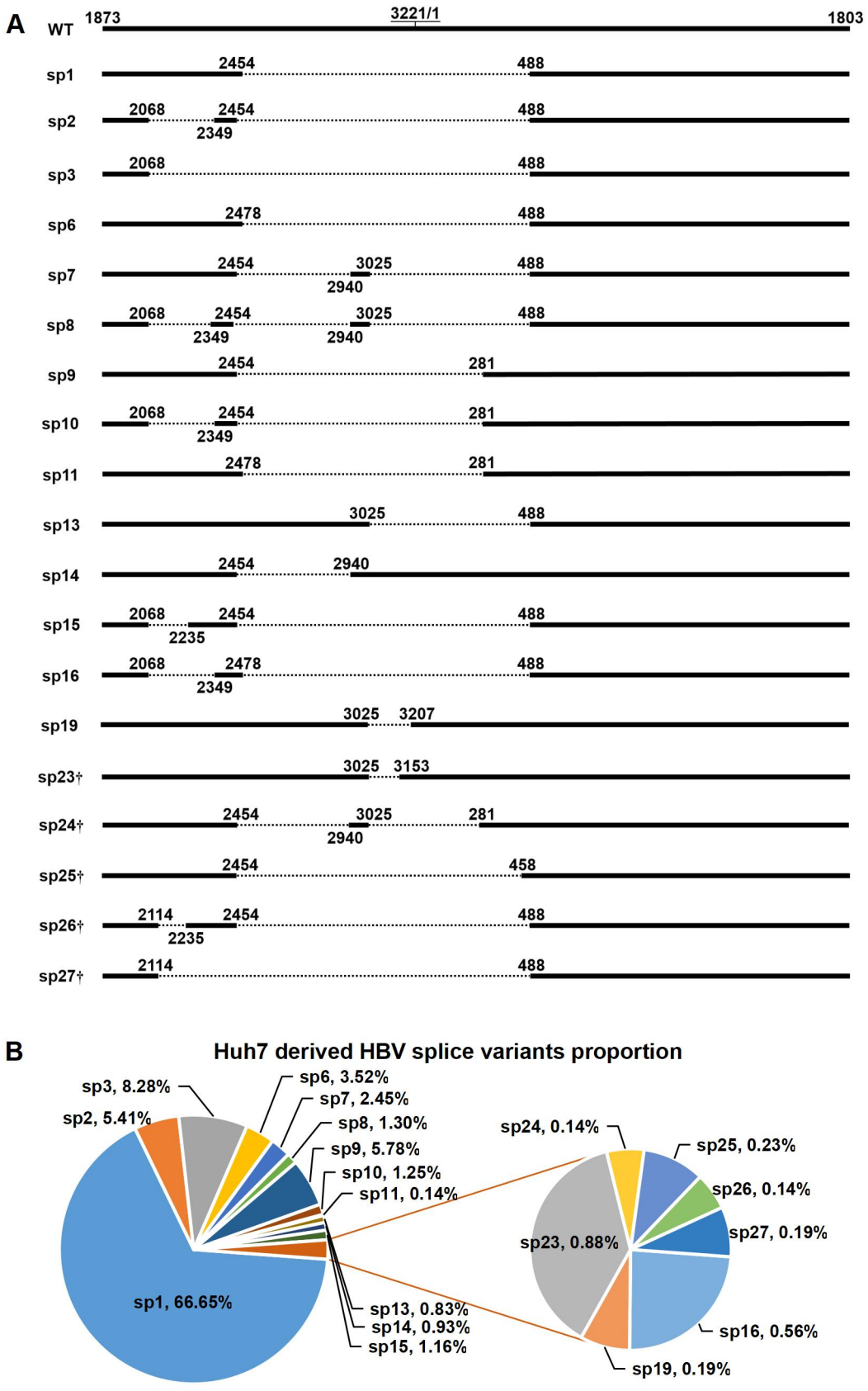


Figure 8 Splice variants and the proportion of each splice variant within their population.
(A) Nineteen splice variants found in this study with their respective splice donor and acceptor sites. Positions are in relation to genotype A HBV. WT: wild type HBV, bold lines: HBV DNA, dotted lines: spliced regions, and dagger (†): novel splice variants.
(B) Population proportion of each splice variant within the splice variant population found in Huh7 derived HBV.

doi:10.6342/NTU202302511