Deep sequencing of full length Hepatitis B Virus (HBV) genomes using Rolling Circle Amplification and Nanopore

Mariateresa de Cesare[1], Hannah Roberts[1], Anna McNaughton[2], David Bonsall[1,2,3], Sheila Lumley[2], Anthony Brown[2], Rory Bowden[1], Eleanor Barnes[2], Philippa C Matthews[2,4]

[1]The Wellcome Centre for Human Genetics, Old Road Campus, Roosevelt Drive, Oxford OX3 7BN UK,
[2]Nuffield Department of Medicine, Peter Medawar Building, South Parks Road, Oxford OX1 3SY UK, [3]Big Data Institute, Old Road Campus, Roosevelt Drive, Oxford OX3 7FZ UK,
[4]Department of Infectious Diseases and Microbiology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK

Introduction

The development of unbiased, whole-genome sequencing methods for viruses including HIV and hepatitis C virus (HCV) have been particularly informative, providing insight into transmission events, the emergence of drug resistance and disease-related outcomes (1). The advancement of such techniques for hepatitis B virus (HBV) will be crucial whilst working towards elimination targets set by United Nations Sustainable Development Goals (2).

Currently, the most widely used sequencing platforms are not portable, require substantial financial investment, protocols can take 1-2 days, and are only capable of sequencing relatively short reads, making haplotype reconstruction challenging.
'Third generation' sequencing approaches such as Oxford Nanopore Technologies (ONT) MinION can overcome these limitations (3).

A potentially significant drawback of Nanopore technology remains the high error rates. A previous study used rolling circle amplification (RCA) to generate concatenated sequences, allowing correction to be made on the basis of the most frequent base derived from multiple end-to-end copies of the same sequence (4). RCA-based techniques are naturally appealing for viruses with circular genomes such as HBV, and methods for the RCA of full-length HBV genomes have been previously published (5, 6). RCA followed by Nanopore sequencing has been pioneered for deriving sequences from other circular viruses (7), but this methodological pipeline has not previously been reported for deriving robust full-length deep sequence data for HBV.

RCA and Nanopore for HBV is appealing

- Concatenated genomes provide a useful method for error correction on the Nanopore platform;
- Large products (derived through RCA) are well-suited to Nanopore sequencing;
- Nanopore will enable the deep sequencing of complete viral haplotypes;
- New platforms like GridION and PromethION are making Nanopore increasingly affordable.

Materials and Methods

Samples and DNA extraction

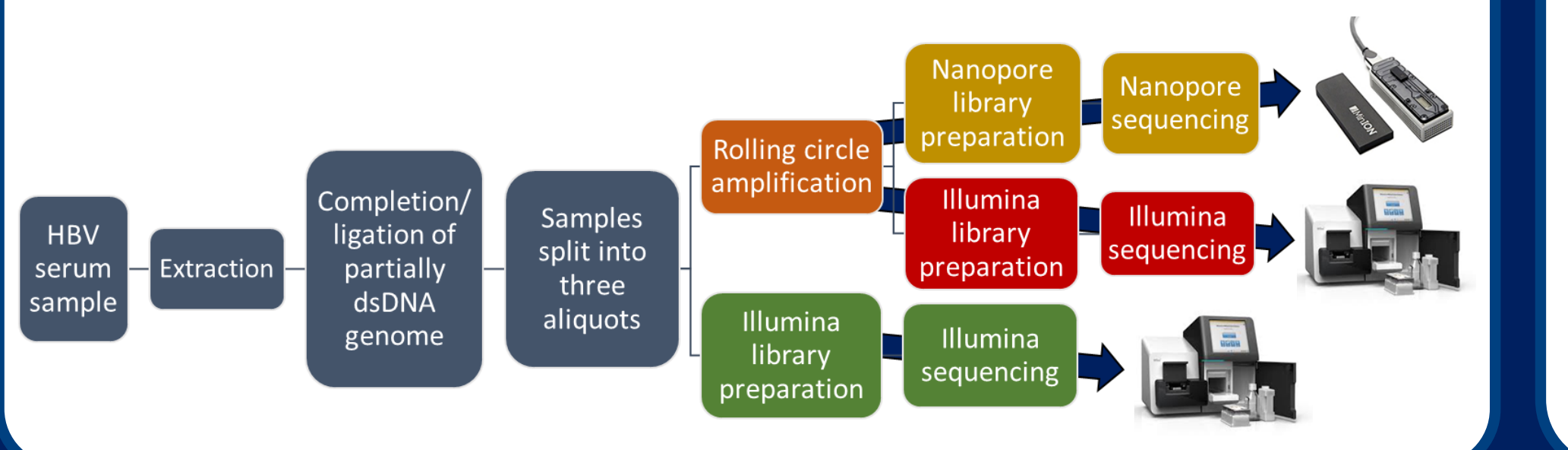
- We used serum samples from HBV-infected adults recruited in Oxford, UK. All samples had viral loads above the limit of detection (>8.23log$_{10}$ IU/ml). All participants provided valid written informed consent.
- Total DNA was extracted from 500µl plasma using the NucliSENS magnetic extraction system (bioMérieux) and eluted into 30µl of kit buffer.

Completion/ligation (C/L) and rolling circle amplification (RCA)

- Methods have previously been described by Martel et al., 2013 (6) and an overview is given in Fig 1. In brief:
- For each serum sample, C/L reactions were prepared in triplicate using 6.4µl extracted DNA added to 3.6µl reaction mix.
- Two of the triplicate reactions then underwent RCA as described in Martel et al., 2013 (6).
- RCA products were assessed by gel electrophoresis and HBV-specific qPCR.

Library preparations and sequencing

- For each sample, both the product of the C/L reaction and an RCA reaction underwent library preparation using a modified Nextera DNA library prep protocol (Illumina) and were sequenced on an Illumina MiSeq sequencer.
- The remaining RCA reaction products were prepared for Nanopore sequencing. First, potential branching generated by RCA was resolved by digesting with a T7 endonuclease. Subsequent library preparation was performed with a 1D Genomic DNA ligation protocol (ONT). The samples were sequenced on a MinION,


HBV serum sample → Extraction → Completion/ ligation of partially dsDNA genome → Samples split into three aliquots

Rolling circle amplification → Nanopore library preparation → Nanopore sequencing

Rolling circle amplification → Illumina library preparation → Illumina sequencing

Illumina library preparation → Illumina sequencing

Results

Genotype and consensus sequences

Full length sequences were derived using all methods. Good concordance was observed between the consensus sequences as generated by C/L and RCA (Illumina) and Nanopore (after error correction). Phylogenetic analysis of consensus sequences for samples 1331 and 1332 alongside reference sequences for genotypes (gt) A-J was performed (Fig 2). Samples 1331 and 1332 were genotyped as gtC and gtE, respectively.

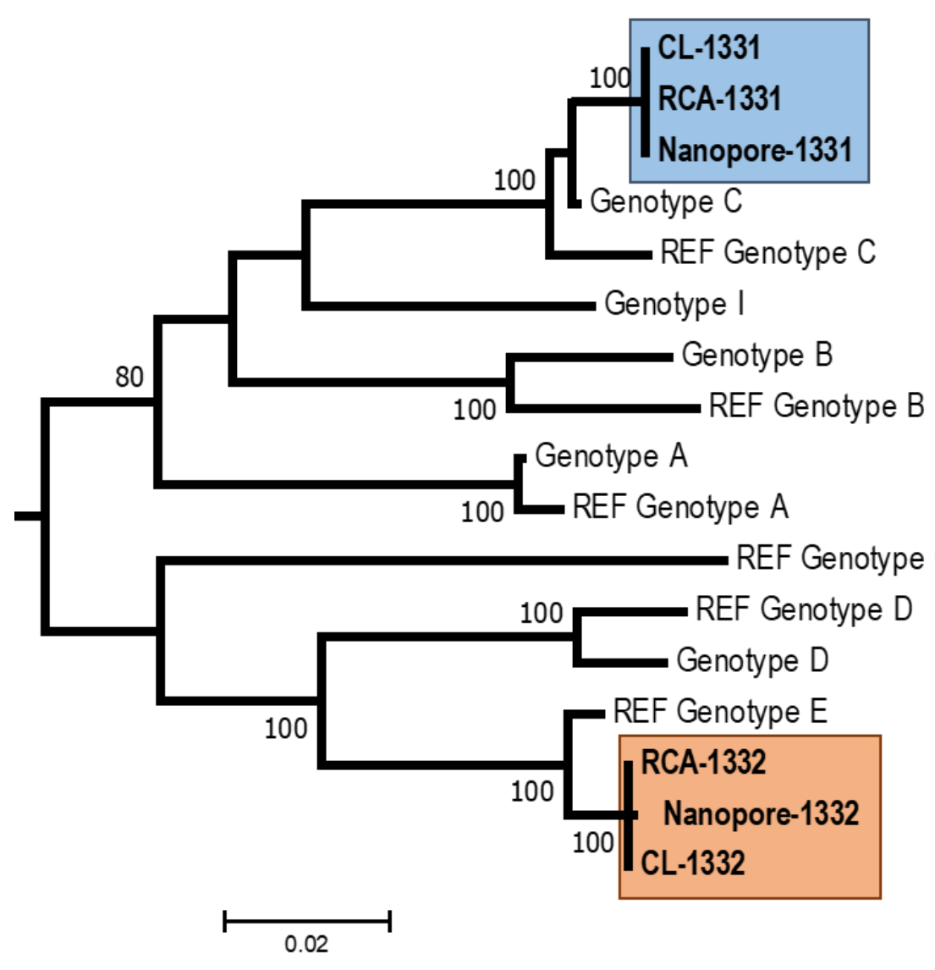


Figure 2; Phylogenetic tree of consensus sequences

Identification of HBV

We successfully sequenced individual whole HBV genomes as single reads using Nanopore. The highest yield sample produced ~1.3m reads; 87% of these were HBV (Fig 3). BLAST identified 2,738 reads that contained the complete HBV genome in a single strand.

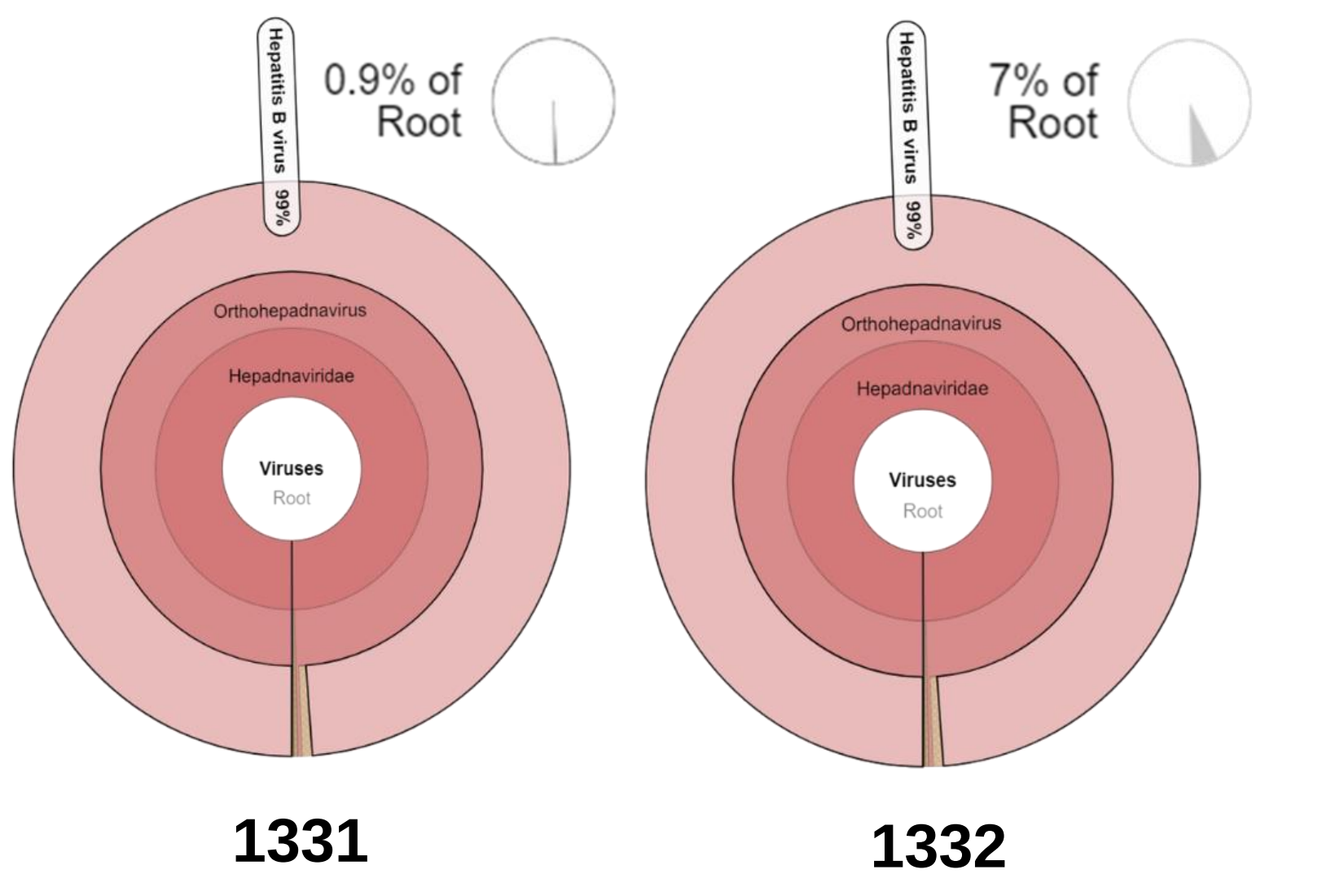


Figure 3; Kraken plots showing the identification of HBV in samples 1331 and 1332 using Nanopore, with total proportion of all reads that were found to be HBV

Coverage across the genome

An analysis of coverage depth of Illumina data generated either by C/L or RCA for 1331, 1332 and another three samples (that were not run on MinION) (Fig 4) indicated that RCA improved site coverage for every sample, particularly at nt 750-1600. A consistent dip in coverage was observed at nt 2500-2800 for all samples, corresponding to a highly variable region of the virus.

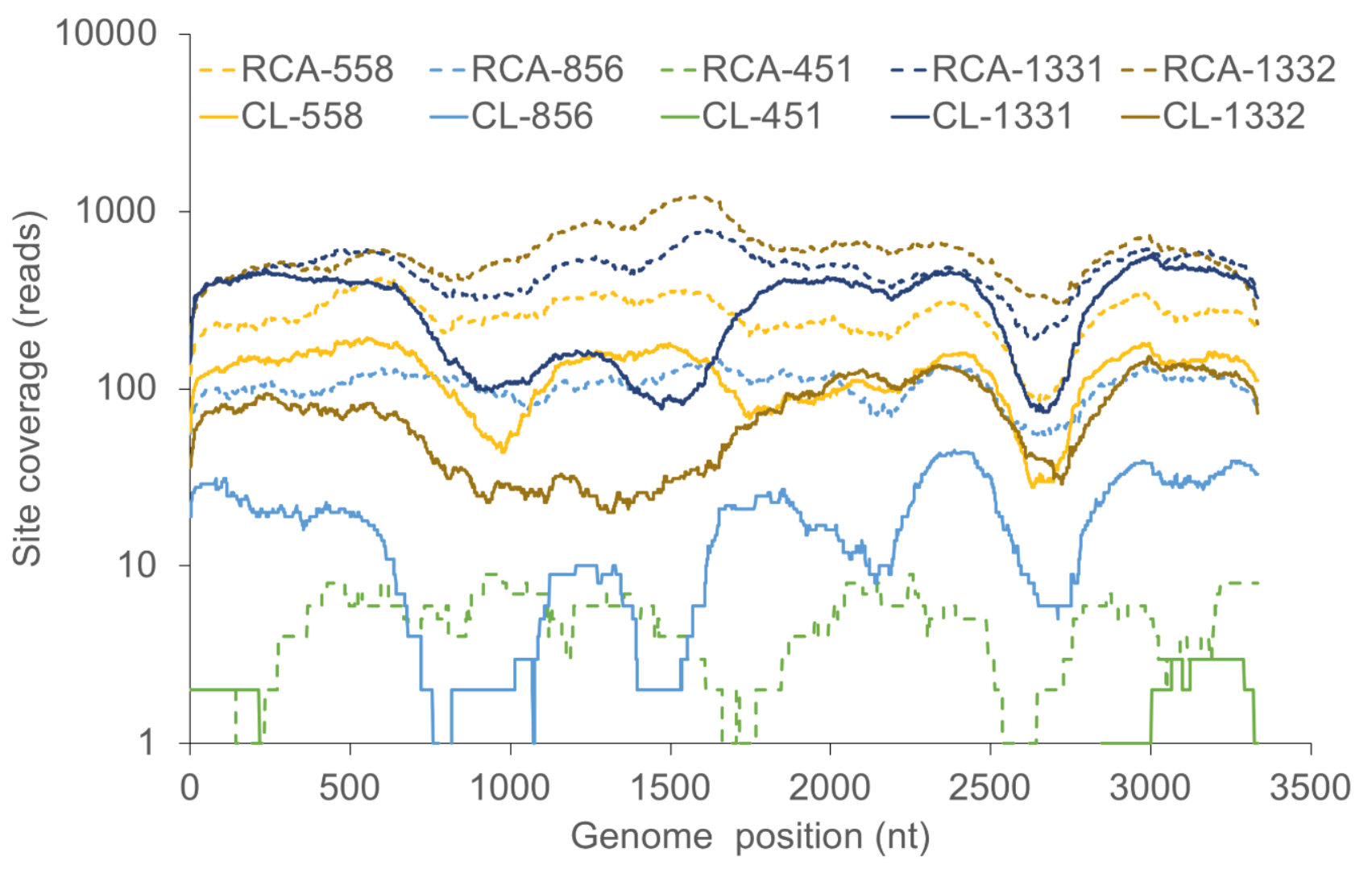


Figure 4; Genome coverage and read depth for multiple samples and methods

Results *(Nanopore data)*

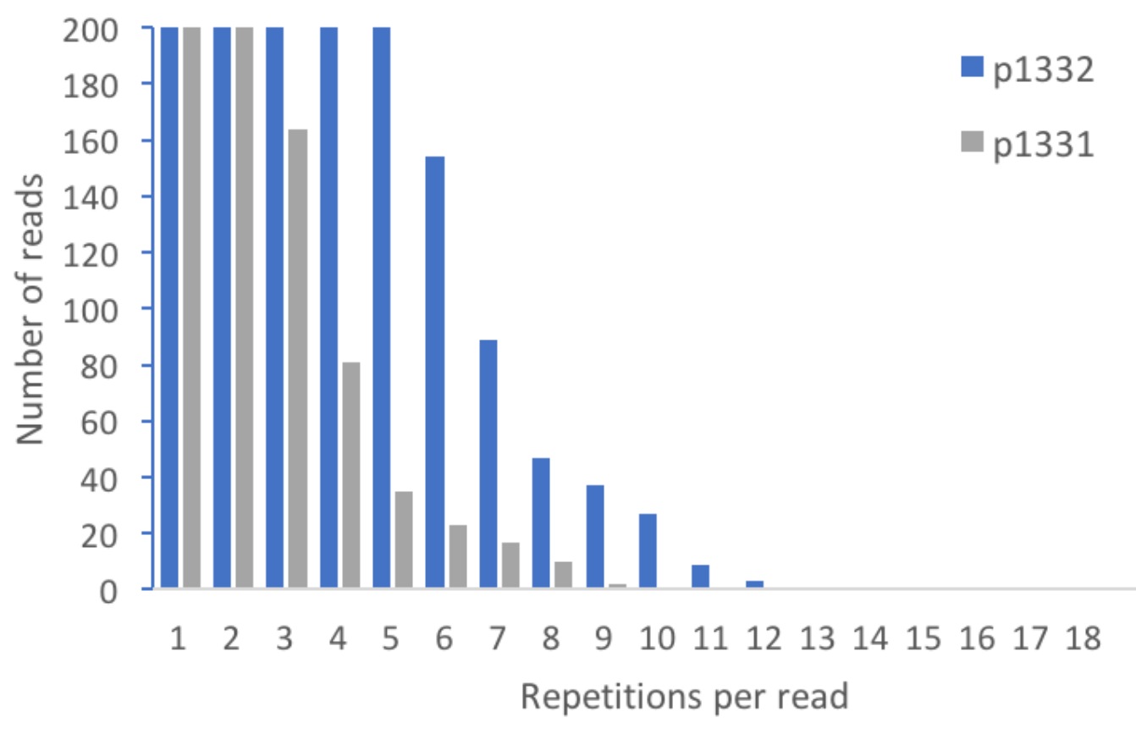
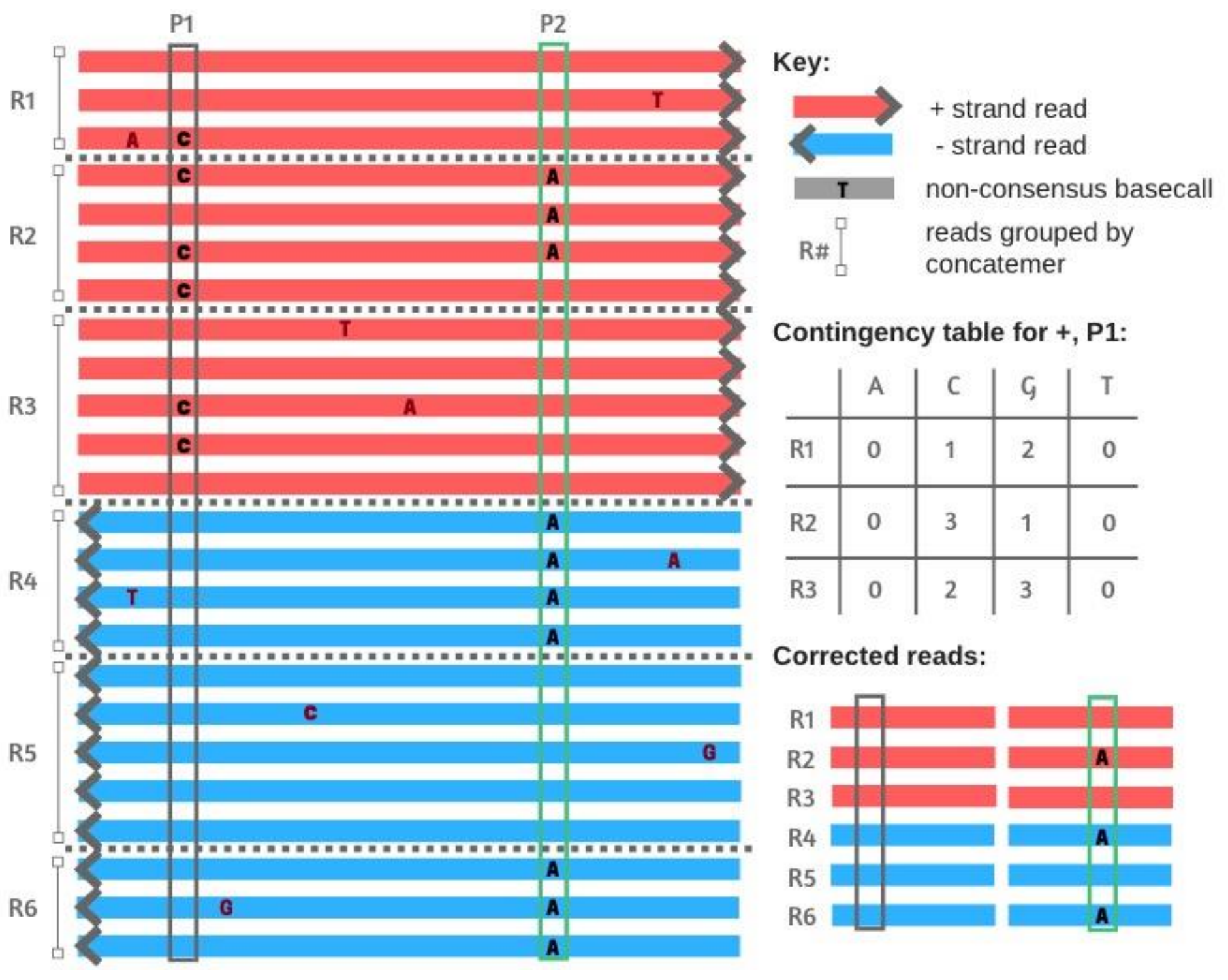


Figure 5; Number of concatenated HBV genomes ('repetitions') per read and frequency within the Nanopore data for two samples, 1331 and 1332



Contingency table for +, P1:

	A	C	G	T
R1	0	1	2	0
R2	0	3	1	0
R3	0	2	3	0

Corrected reads:

Figure 6: Identifying and removing strand-specific basecaller errors.

Read length

Reads containing at least one full-length HBV genome were selectively identified and the number concatenated genomes per read for each sample analysed. 1332 had a larger number of reads with high (>7) numbers of concatenated genomes (Fig 5).

Identification of basecaller errors

Albacore 2.0.2 was used for basecalling. Basecaller errors in Nanopore data are both randomly distributed and systematically miscalled. Repeated errors are frequently strand-specific, e.g C at position P1 (Fig 6). To correct errors, we developed a method to distinguish between repeated basecaller errors (P1) and genuine nucleotide variants (P2) (Fig 6) as follows:

- Consider positive (+, red) and negative (-, blue) strand reads separately.
- For each set of reads (+ or -), and for each site with variation, test for an association between base and concatemer using a contingency table, as illustrated.
- If no association is found, this suggests strand-specific miscalling: derive corrected base from the consensus across all concatemers.
- If an association is found, this implies genome-specific variants: derive corrected base from the consensus within each concatemer.

Identification of variable sites

After removing sequencing errors as above, the remaining nucleotide variants detected in the Nanopore data show good concordance with those detected by Illumina sequencing, as evidenced by the correlation of base frequencies within the two data sets (Fig 7). The low numbers of variable sites within the sample suggest that the sample has relatively low within-host diversity.

Error correction and branch lengths

Uncorrected Nanopore sequence contains a large number of errors, as indicated by the long braches in the phylogenetic tree in Fig 8A. Taking the concatenated reads and calling a consensus average of the full-genome in each read substantially reduces the branch lengths (Fig 8B). Using methods shown in Fig 6, taking into account strand-specific miscalling, additional error is identified and removed (Fig 8C) with the corrected tree showing low patient diversity, in keeping with the Illumina data (Fig 7).

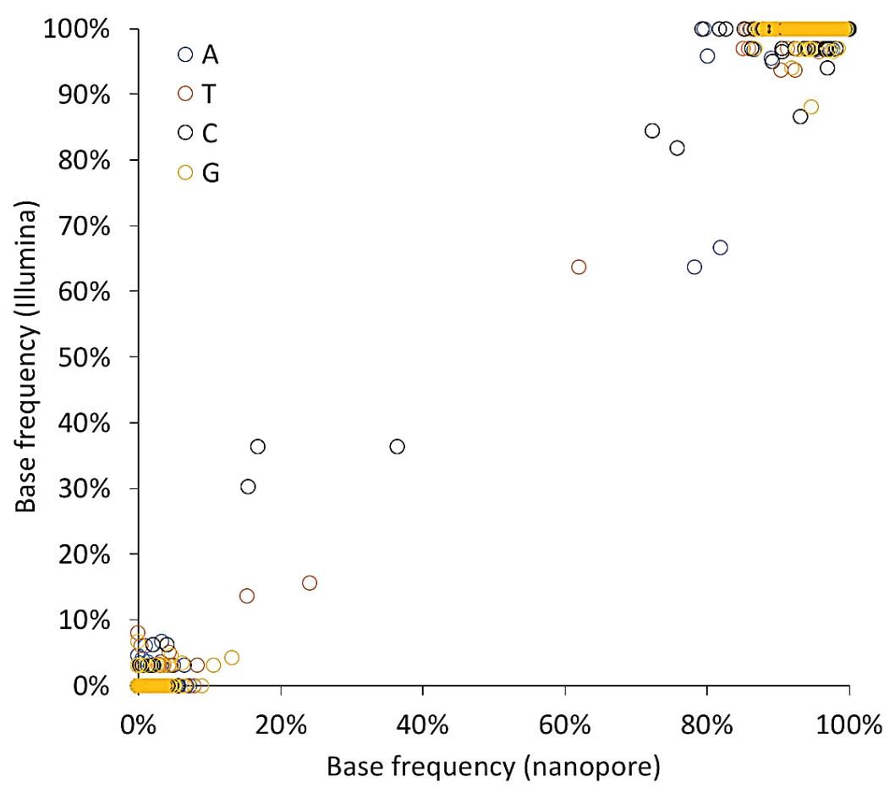


Figure 7: Base frequencies at all sites (each site is represented by four points) using Nanopore vs Illumina

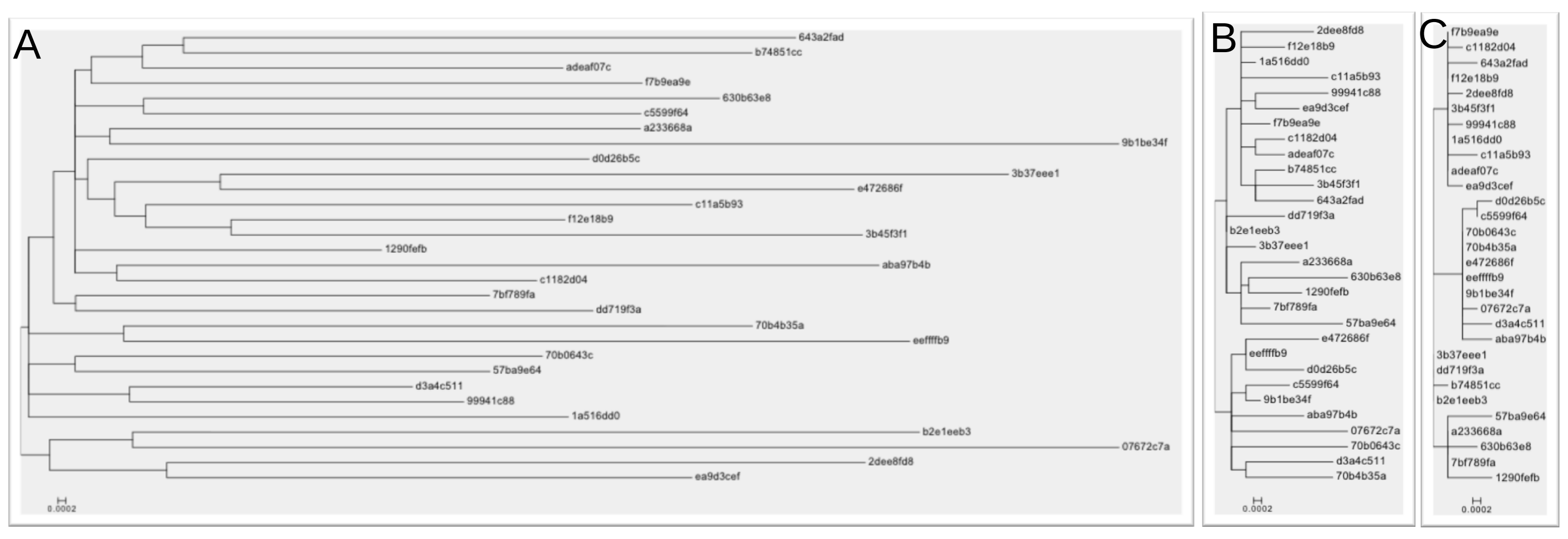


Figure 8. Maximum likelihood trees of full-length HBV sequences from sample 1331 sequenced by Nanopore. (A) Prior to correction (one full-length read per concatemer), (B) concatemers corrected by consensus averaging and (C) concatemers corrected according to our method in Figure 6.

Evaluation of RCA

Shannon entropy (SE) within the reads returned from RCA and C/L data sets (sequenced on Illumina) were analysed to examine how well viral diversity was represented using RCA-based approaches (Fig 9). Diversity (measured by SE) was consistently higher in the samples that only underwent CL prior to sequencing, suggesting that the RCA approach may be under-representing diversity within the samples.

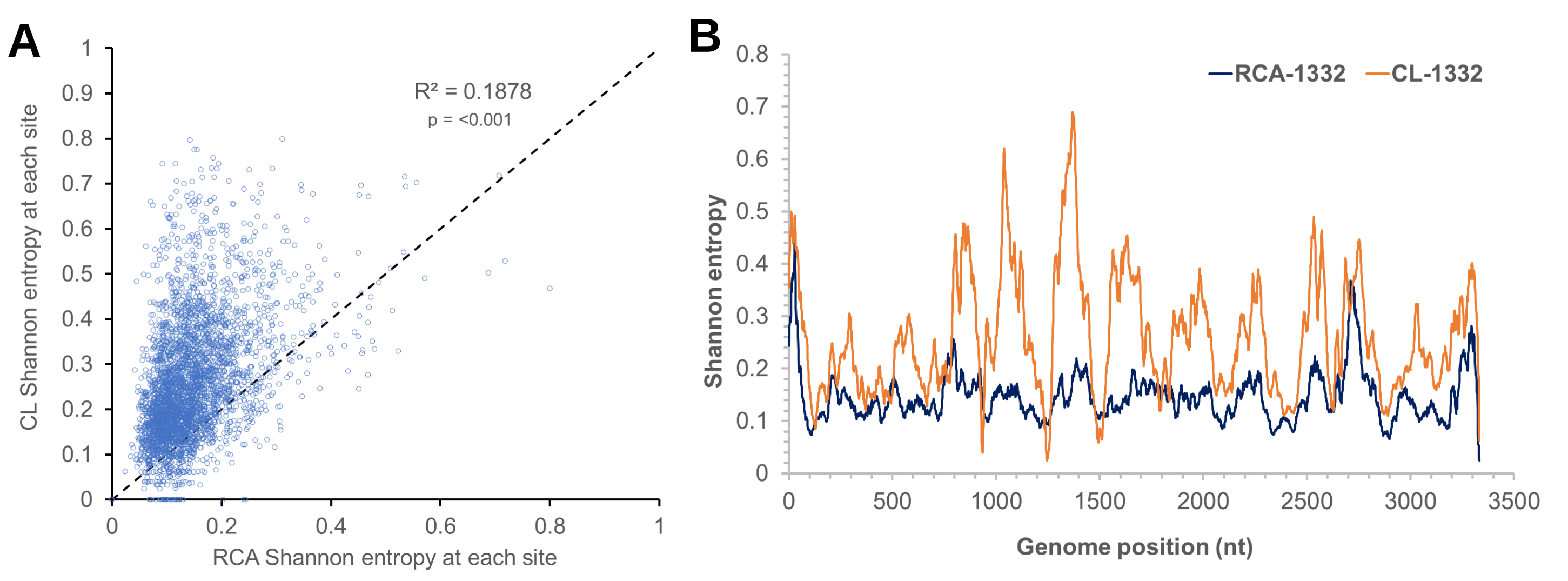


Figure 9; Concordance of Shannon entropy (SE) at each site across the HBV genome for 1332 in reads returned from Illumina. (A) Correlation between SE for C/L and RCA-generated sequences; (B) SE for C/L and RCA-generated sequences plotted along the HBV genome.

Conclusions and Future work

We have demonstrated that an isothermal RCA method functions to provide enrichment of HBV DNA and concatenation of the HBV genome. This is the first example of a deep-sequencing method for characterising viral variants that takes full advantage of Nanopore's ability to sequence whole genomes within single reads, and also facilitates correction of sequencing error. There is exciting potential for these tools to be refined for full-length, high-resolution sequencing analysis of HBV (and other viruses), impacting both clinical and research settings.

- ▶ This approach, using 8 conserved primers, can be applied to multiple HBV genotypes;
- ▶ RCA considerably improves read depth across the genome, although a consistent dip in coverage is observed at nt 2500-2800;
- ▶ After applying our error correction methods, we are able to accurately characterise within-host diversity from individual whole viral genomes
- ▶ Low levels of within-host diversity were observed within 1331 and 1332. Both samples were HBeAg positive and it is possible this lack of diversity is linked to their HBeAg status.

- ▶ Our data suggest that the RCA-based approach may be limiting the amount of diversity represented in the sequencing data. Further investigation of this finding is required.
- ▶ We will trial this approach with a wider range of HBV genotypes and HBeAg-negative samples.
- ▶ We will test the RCA-based enrichment method on samples with a wider range of viral loads to determine sensitivity

References

(1) Quiñones-Mateu *et al*; J Clin Virol. 2014 Sep; 61(1): 9–19.
(2) WHO (2016) Draft global health sector strategy on viral hepatitis, 2016–2021
(3) Pennisi E *et al*; Science. 2016;351(6275):800-1.
(4) Li C *et al*; Gigascience. 2016;5(1):34.
(5) Margeridon S *et al*; Antimicrob Agents Chemother. 2008;52(9):3068-73.
(6) Martel *et al*.; J Virol Methods. 2013 Nov;193(2):653-9.
(7) Vanmechelen B *et al*; Curr Protoc Microbiol. 2017;44:1E 12 1-1E
(8) Garson *et al*; J Virol Methods. 2005 Jun;126(1-2):207-13.

Contact: philippa.matthews@ndm.ox.ac.uk
Look for Mariateresa de Cesare and Hannah Roberts at LC2018.