OPEN ACCESS Freely available online

PLOS ONE

Analysis of Ultra-Deep Pyrosequencing and Cloning Based Sequencing of the Basic Core Promoter/Precore/Core Region of Hepatitis B Virus Using Newly Developed Bioinformatics Tools

Mukhlid Yousif[1☯], Trevor G. Bell[1☯], Hatim Mudawi[2], Dieter Glebe[3], Anna Kramvis[1]*

1 Hepatitis Virus Diversity Research Programme, Department of Internal Medicine, University of the Witwatersrand, Johannesburg, Gauteng, South Africa, **2** Department of Medicine, Faculty of Medicine, University of Khartoum, Khartoum, Khartoum State, Sudan, **3** Institute of Medical Virology, National Reference Centre of Hepatitis B and D, Justus, Liebig-University of Giessen, Giessen, Hesse, Germany

Abstract

Aims: The aims of this study were to develop bioinformatics tools to explore ultra-deep pyrosequencing (UDPS) data, to test these tools, and to use them to determine the optimum error threshold, and to compare results from UDPS and cloning based sequencing (CBS).

Methods: Four serum samples, infected with either genotype D or E, from HBeAg-positive and HBeAg-negative patients were randomly selected. UDPS and CBS were used to sequence the basic core promoter/precore region of HBV. Two online bioinformatics tools, the ''Deep Threshold Tool'' and the ''Rosetta Tool'' (http://hvdr.bioinf.wits.ac.za/tools/), were built to test and analyze the generated data.

Results: A total of 10952 reads were generated by UDPS on the 454 GS Junior platform. In the four samples, substitutions, detected at 0.5% threshold or above, were identified at 39 unique positions, 25 of which were non-synonymous mutations. Sample #2 (HBeAg-negative, genotype D) had substitutions in 26 positions, followed by sample #1 (HBeAg-negative, genotype E) in 12 positions, sample #3 (HBeAg-positive, genotype D) in 7 positions and sample #4 (HBeAg-positive, genotype E) in only four positions. The ratio of nucleotide substitutions between isolates from HBeAg-negative and HBeAg-positive patients was 3.5:1. Compared to genotype E isolates, genotype D isolates showed greater variation in the X, basic core promoter/precore and core regions. Only 18 of the 39 positions identified by UDPS were detected by CBS, which detected 14 of the 25 non-synonymous mutations detected by UDPS.

Conclusion: UDPS data should be approached with caution. Appropriate curation of read data is required prior to analysis, in order to clean the data and eliminate artefacts. CBS detected fewer than 50% of the substitutions detected by UDPS. Furthermore it is important that the appropriate consensus (reference) sequence is used in order to identify variants correctly.

Citation: Yousif M, Bell TG, Mudawi H, Glebe D, Kramvis A (2014) Analysis of Ultra-Deep Pyrosequencing and Cloning Based Sequencing of the Basic Core Promoter/Precore/Core Region of Hepatitis B Virus Using Newly Developed Bioinformatics Tools. PLoS ONE 9(4): e95377. doi:10.1371/journal.pone.0095377

Editor: John E. Tavis, Saint Louis University, United States of America

Received December 21, 2013; **Accepted** March 26, 2014; **Published** April 16, 2014

Funding: The authors acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) of the special African-German collaborative programme, "Africa Initiative", (Grant GL595/3-1 to AK, HM and DG), Faculty of Health Sciences, University of the Witwatersrand. TB received post-doctoral funding from the University of the Witwatersrand (SPARC) and the National Research Foundation of South Africa (GUN #86215). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Anna Kramvis is a PLOS ONE Editorial Board member. This does not alter the author's adherence to PLOS ONE Editorial policies and criteria.

* E-mail: Anna.Kramvis@wits.ac.za

☯ These authors contributed equally to this work.

Introduction

The continued improvement of DNA sequencing technologies has led to the development of next generation sequencing (NGS) methods, including ultra-deep pyrosequencing (UDPS), which are capable of sequencing many thousands of nucleotides, quickly and at a low cost per nucleotide. These technologies have overcome the disadvantages of the traditional dye-terminating DNA sequencing technology developed by Frederick Sanger [1]. These disadvantages include the relatively high cost per nucleotide, in terms of money and time, and the fact that Sanger sequencing is only capable of detecting sequence variants, which are present in 20% or more of a quasispecies population [2,3]. Moreover, NGS methods also overcome several of the drawbacks of cloning based sequencing (CBS), such as the time, money and expertise required to prepare samples, especially when a large number of clones is required [4]. NGS methods are used primarily for *de novo* or ''shot-gun'' sequencing of new or known genomes. This produces a very

large number of short reads, which are then assembled to produce a complete sequence. Several algorithms and tools exist to process these short reads [5]. In addition to producing short reads, the pyrosequencing platform can be used for amplicon re-sequencing (UDPS). These longer reads are typically an amplicon covering a genomic region of interest. At present, the GS Titanium UDPS chemistry produces reads of approximately 400 bases in length.

Few bioinformatic tools, which are affordable and accessible to resource-constrained environments, are currently available to assist with the processing and analysis of amplicon re-sequencing data. The Roche AVA software (http://www.454.com/products/analysis-software/#amplicon-tabbing), although free of charge, can only be installed on a computer running a particular GNU/Linux distribution, and a number of commercial software packages cost several thousand US dollars for a single license. Alignment and visualization tools, which are used routinely for smaller datasets, are not suitable for datasets containing hundreds or thousands of reads. Additionally, many of these software solutions require a level of technical expertise, which many biological researchers may not possess.

Pyrosequencing is an error-prone technique [6]. Distinguishing between a true biological variant and an error (artefact) is a vital step in analysing pyrosequencing data. Although a number of studies discuss error correction in pyrosequencing data [6,7], there is currently no consensus regarding the error threshold, which should be applied. Knowledge of well-characterized regions of a genome is important in order to develop tools to examine pyrosequencing data and to distinguish between artefacts and true variations.

Hepatitis B virus (HBV) displays remarkable sequence heterogeneity, with 9 genotypes (named A to I) currently recognized [8,9]. The precore/core (PC/C) open reading frame (ORF) of HBV encodes for both the hepatitis B e antigen (HBeAg) and the core protein (HBcAg). This region is preceded by the basic core promoter (BCP) region, which controls transcription of both the PC/C mRNA and the pregenomic RNA (pgRNA) during the replication cycle [10]. The BCP/PC ORF overlaps the X ORF. HBeAg is a soluble, non-particulate protein that is secreted in the serum or expressed on the surface of the hepatocyte [11,12]. Conventionally, HBeAg expression is an indicator of active HBV infection and on-going viral replication [12]. However, HBeAg expression may be reduced or completely suppressed by various viral mutations, even in the presence of viral replication. Mutations in two regions may affect HBeAg expression: precore mutations (for example, G1896A) [13] and BCP mutations (for example, A1762T/G1764A) [14]. The viral capsid is composed of HBcAg [15]. Mutations may occur more frequently in N-terminal or central region of the core protein, which does not overlap other reading frames [16].

Using a segment of this well-characterized BCP/PC/C region of HBV as a model, the objectives of this study were to:

- develop bioinformatics tools to explore UDPS data,
- test and use them to determine the optimum error threshold, and
- compare results between UDPS and CBS using HBeAg-positive and –negative sera infected, with either genotype D or E.

Materials and Methods

Sample Selection

Written informed consent was obtained from all participants and the consent was approved by the Sudanese Ministry of Health, who gave permission for the sera to be used for research purposes. The Human Ethics Committees of the University of the Witwatersrand and the University of Khartoum approved the study. Four serum samples were selected from our previous study on HBV from monoinfected individuals, where the HBV genotype was determined using phylogenetic analysis [17]. Sample #1 was HBeAg-negative and infected with genotype E of HBV (GenBank, KF170783), sample #2 was HBeAg-negative, genotype D (KF170739), sample #3 was HBeAg-positive, genotype D (KF170740) and sample #4 was HBeAg-positive, genotype E (KF170788).

Wet Laboratory Work

Ultra-Deep Pyrosequencing (UDPS). A region of the HBV genome (1653–1959 from *Eco*R1 restriction site) was amplified using a slight modification of a previously described method [18]. Primers 1606 (+) and 1974 (−) were used for the first round PCR, and 1653 (+) and 1959 (−) for the second round PCR. The first round PCR was followed by gel-purification using Zymoclean Gel DNA Recovery Kit (Zymo Research Corp, Irvine, CA, USA). For the second round PCR, modified primers, which were ligated to adaptors and tags, were used (Table 1). Following second round PCR, the amplicons were gel-purified and subjected to UDPS in the forward direction on the Roche 454 GS Junior platform (454 Life Sciences, Roche Company, Switzerland), which provided reads covering the region of interest (coordinates 1653–1959). The UDPS sequencing data has been submitted to the GenBank SRA database, as BioProject accession: PRJNA239442 and the following are the BioSample accessions: SAMN02664575, SAMN02664576, SAMN02664577, SAMN02664578.

Cloning Based Sequencing (CBS). After nested PCR, the 307 nucleotide amplicon (1653–1959 from *EcoR*1 site) was gel-purified and cloned into pTZ57R/T vector (55 ng/μl) using Instaclone PCR Cloning Kit (Fermentas, Waltham, MA, USA), and transformed into TOP10 *Escherichia coli* (Invitrogen, Carlsbad, CA, USA). The transformants were grown on Ampicillin plates. Positive clones were identified by restriction fragment length polymorphism (RFLP) assay. At least 20 clones per sample were sequenced by direct sequencing, using a BigDye Terminator v3.0 Cycle Sequencing Ready Reaction Kit (Applied Biosystems., Foster City, USA) on an ABI 3130XL Genetic Analyzer (Applied Biosystems). The sequencing primer used was M13 forward (5′-GTAAAACGACGGCCAGT-3′). A phylogenetic tree was generated as described previously [17]. Clone sequences have been deposited in GenBank: Genotype D: KJ496256-KJ496297, Genotype E: KJ496206-KJ496255.

Dry Laboratory Work

Data pre-processing. UDPS data for three sequencing runs, for each of the four samples, was processed and analyzed as shown in the flow diagram (Figure 1). The data from each run, for each sample, was processed individually. Separate binary standard flowgram format (SFF) files were opened in the R statistical programming language [19], using the "raw" clip-mode parameter (which does not perform any clipping or trimming) of the "rSFFreader" library [20]. Sequence data were searched for the forward and reverse primer sequences and the adaptor sequence for verification. Sequence lengths in each file were plotted and examined statistically (*data not shown*).

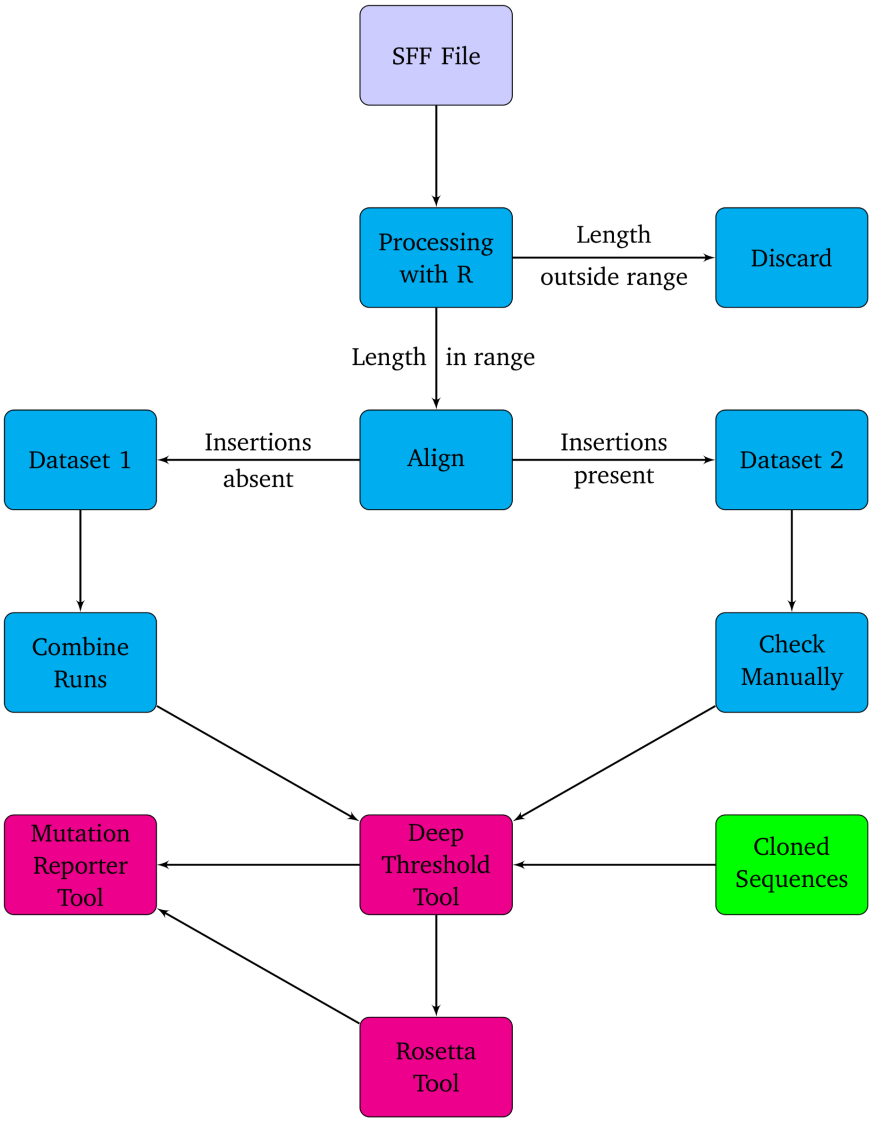


Figure 1. Flow diagram of the analysis procedure used for the UDPS data.
doi:10.1371/journal.pone.0095377.g001

Table 1. Sequences of HBV-specific primers, tags and adaptors used for ultra-deep pyrosequencing.

	Position from *Eco*RI site	HBV specific target sequence	Tag sequences[¶]	Adaptor sequences[€]
PCR 1	1606 (+) [1606–1625]	5′-GCATGGAGACCACCGTGAAC-3′	No tag	No adaptor
	1974 (−) [1974–1955]	5′GGAAAGAAGTCAGAAGGCAA-3′	No tag	No adaptor
PCR 2	1653 (+) [1653–1672]	5′-CGTATCGCCTCCCTCGCGCCATCAG-3′	ACACGACGACT[1]	CAT AAG AGG ACT CTT GGA CT
			ACACGTAGTAT[2]	
			ACACTACTCGT[3]	
			ACGACACGTAT[4]	
	1959 (−) [1959–1940]	5′-CTATGCGCCTTGCCAGCCCGCTCAG-3′	No tag	GGC AAA AAC GAG AGT AAC TC

[¶]Short specific sequences used to label the different samples: [1:] sample #1; [2:] sample #2; [3:] sample #3; [4:] sample #4.
[€]Short specific sequence used to ligated onto the ends of the fragments. These adaptors provide priming sequences for both amplification and sequencing of the sample-library fragments.
doi:10.1371/journal.pone.0095377.t001

Figure 2. The input pages of the bioinformatics tools (A) ''Deep Threshold Tool''. The first field specifies the input FASTA file. Fields are available for the user to specify the nucleotide offset mapping of the first position in the input file, the number of nucleotides (length) to process, the starting and ending probabilities of error to examine, and the probability of error increment (step) to use. (**B**) ''Rosetta Tool''. The first field specifies the input FASTA file. Fields are available for the user to specify the nucleotide offset mapping of the first position in the input file, the position of the first in-frame nucleotide of the coding region of interest, the last in-frame nucleotide of the coding region of interest, the amino acid offset of the first amino acid in the coding region of interest, and the probability of error to use.
doi:10.1371/journal.pone.0095377.g002

The distribution of all sequence lengths was examined and a length range was selected, which excluded reads with very low counts. Several Linux command-line BASH scripts and Python programming language scripts (*available on request*) were written to include only reads within a specified length range (between 330 to 360 nucleotides) for further processing. A genotype D reference sequence (GU456684) was then added to each dataset, and the file was aligned with the Muscle program [21]. Each alignment was then processed by a Python script, which scanned the reference sequence in the alignment and removed any reads from the alignment with an insertion (a residue aligned with a gap in the reference sequence). In the remaining alignment (excluding reads with insertions), positions (columns), containing only gaps, were collapsed and this alignment was ''Dataset 1''. The repeated runs for all ''Dataset 1'' sequences for each sample were then combined into one dataset, the final ''Dataset 1''. The file containing reads with insertions was ''Dataset 2'' for each run and these were processed individually because of variable read lengths, as a result of insertions at different positions in the reads.

Development of deep threshold tool. For pyrosequencing data of human immunodeficiency virus (HIV), a probability of error, ranging from 0.5% to 1%, has been used [6]. In the present study, using HBV data, a web-based tool (the ''Deep Threshold Tool'') (http://hvdr.bioinf.wits.ac.za/tools/) was developed to examine the number of errors in each position (column) in an alignment, depending on the probability of error value. In order to examine the number of errors, the tool requires an input alignment in FASTA format, the lower and upper bounds of the probability of error, and an increment value (Figure 2A). A nucleotide mapping offset can be specified, so that the resulting output coordinates reflect the correct position of the sequence in the entire genome. Potentially untidy ends of reads (such as the reverse primer region) can be excluded from the analysis by specifying a length shorter than the sequence length.

Statistical calculation of the threshold. A nucleotide was considered an ''error'' if its frequency in a column in the alignment was less than the threshold, which was determined as follows. An expected frequency of $\boldsymbol{E = probability\ of\ error \times read\ depth\ (R)}$ was used. A Pearson's χ^2 test statistic was calculated as follows:

$$M = \frac{(O-E)^2}{E} + \frac{((R-O)-(R-E))^2}{R-E}$$

with ***O*** being the observed value, starting at 1. If ***M*** was less than the χ^2 distribution (with $\alpha = 0.05$ and one degree of freedom), then ***O*** was incremented by a value of one and the test was repeated. The value for O at which the χ^2 distribution was exceeded, was considered the threshold value (count). This threshold was calculated for each position in the alignment. Any nucleotide with a frequency below this threshold was considered an error or artefact.

Development of rosetta tool. Amino acid data were examined using the newly-developed "Rosetta Tool". This tool requires the same input file as the "Deep Threshold Tool ". It also requires a nucleotide offset mapping and the start and end positions of a protein region. This does not have to include the position of the start or stop codon; any region of a protein can be processed, as long as the number of nucleotides specified by the range is a multiple of three. The probability of error at which the data must be analyzed is also required (Figure 2B).

Results

A total of 10952 reads were generated on the 454 GS Junior platform for the three runs for all four samples. Of these, 9738 reads (88.9%) were included in the study (2002, 3049, 1955 and 2732 reads for samples 1, 2, 3 and 4, respectively) and 1214 reads (11.1%), which were considered either too short or too long, were excluded. These 9738 reads were split into Dataset 1 (8967 reads, 92.1%) and Dataset 2 (771 reads, 7.9%) (Figure 1). Ninety-two clones were generated for all four samples: 23 clones for sample #1, 22 for sample #2, 20 for sample #3 and 27 for sample #4.

Deep Threshold Tool Output

The output page generated by the Deep threshold Tool includes a table for each increment of the probability of error (Figure 3), which shows the distribution of nucleotides at all columns at which at least one base can be considered an "error". Because a nucleotide was considered an "error", if its frequency in a column in the alignment was less than the threshold, any variation above the threshold was considered a legitimate variant for that probability of error (Figure 4). Figure 5 summarizes the results graphically. This summary table was consulted and the lowest probability of error, at which established, well-characterized variants are still detected, was selected. In the present study, the

Input file	Sample2.aa
Length	287
Probability of Error	0.0100
Expected	29
Threshold	40
Number of Interesting Columns	10
Interesting columns	[1727, 1739, 1742, 1748, 1753, 1772, 1775, 1909, 1912, 1913]

base/locus	1727	1739	1742	1748	1753	1772	1775	1909	1912	1913
A	63	16	13	41	6	2548	282	1	0	0
C	0	25	0	0	15	13	0	248	395	2698
G	2787	2746	2792	2805	40	1	2565	0	0	151
T	0	63	0	3	2789	274	1	2601	2455	1
M	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0
-	0	0	45	1	0	14	2	0	0	0

Figure 3. An example section of the output from the ''Deep Threshold Tool'', showing the two tables of output provided for each probability of error examined. The "expected" and "threshold" counts are shown in the top table, as well as the number of interesting columns (those columns containing at least one mutation at above-threshold frequency), and a list of the interesting columns. The bottom table provides detailed output, showing the number of each residue occurring in each interesting column.
doi:10.1371/journal.pone.0095377.g003

Probability of Error	Expected	Threshold	IC (#)	Interesting Columns
0.0050 (0.5%)	14	22	25	[1678, 1680, 1706, 1724, 1725, 1727, 1728, 1730, 1736, 1739, 1741, 1742, 1745, 1748, 1751, 1753, 1761, 1772, 1773, 1775, 1842, 1896, 1909, 1912, 1913]
0.0060 (0.6%)	17	26	21	[1678, 1680, 1724, 1725, 1727, 1728, 1736, 1739, 1741, 1742, 1745, 1748, 1753, 1761, 1772, 1775, 1842, 1896, 1909, 1912, 1913]
0.0070 (0.7%)	20	29	17	[1678, 1724, 1725, 1727, 1728, 1739, 1741, 1742, 1745, 1748, 1753, 1772, 1775, 1842, 1909, 1912, 1913]
0.0080 (0.8%)	23	33	15	[1678, 1724, 1727, 1728, 1739, 1741, 1742, 1748, 1753, 1772, 1775, 1842, 1909, 1912, 1913]
0.0090 (0.9%)	26	37	11	[1724, 1727, 1739, 1742, 1748, 1753, 1772, 1775, 1909, 1912, 1913]
0.0100 (1.0%)	29	40	10	[1727, 1739, 1742, 1748, 1753, 1772, 1775, 1909, 1912, 1913]

Figure 4. The first of two summary output tables provided by the ''Deep Threshold Tool''. For each probability of error in the range specified, the expected and threshold values are shown, the number of interesting columns (IC) and the list of these interesting columns. This table provides a summary of the output provided in the previous tables (Figure 3 *top*).
doi:10.1371/journal.pone.0095377.g004

lowest probability of error at which substitutions at positions 1753, 1773 and 1896 are still evident was 0.5%, and this was taken to be our probability of error value.

Rosetta Tool Output

Alignments generated from direct sequencing, UDPS or CBS can also be submitted to the Rosetta Tool. This would typically be done in order to make use of the nucleotide/amino acid alignment viewer component of the tool. The tool produces a number of output tables (Figures 6–8). Figure 6 is an alignment showing each codon followed by the amino acid. Amino acids have been colour-coded according six different categories: Aliphatic (Glycine, Alanine, Valine, Leucine and Isoleucine), Hydroxyl (Serine, Cysteine, Threonine and Methionine), Cyclic (Proline), Aromatic (Phenylalanine, Tyrosine and Tryptophan), Basic (Histidine, Lysine and Arginin) and Acidic (Aspartate, Glutamate, Asparagine and Glutamine). The display of nucleotides or amino acids can be toggled on or off for ease of reference. Figure 7 shows the distribution of each residue at each position at which at least one residue is considered an error. Such error residue counts are highlighted with a black background for reference. Figure 8 contains separate tables for each codon at which at least one residue is an ''error'', and shows the distribution of codons and amino acids at this position. Synonymous and non-synonymous mutations can be differentiated. Rows containing substitutions occurring below the threshold, ''error'' nucleotides are highlighted with a black background.

In order to analyze the data downstream, the Rosetta Tool produces a ''masked'' data file, which is generated by replacing all ''error'' residues in the nucleotide alignment, with an ''X'' character. This alignment is then be translated into amino acids, with an amino acid of ''X'' used whenever at least one ''X'' character per codon occurs. Both the nucleotide and amino acid masked files can be downloaded in FASTA format.

Using the selected probability of error of 0.5%, masked files were generated and the UDPS data were then analyzed using the two newly developed tools and the Mutation Reporter Tool [22].

Analysis of Pyrosequencing Reads

Each sample in Dataset 1 was then analysed using the newly developed ''Deep Threshold Tool'' and a probability of error of 0.5% was selected, because this was the lowest probability of error at which all three well characterized mutations (T1753G/C, T1773C and G1896A) were present. The resulting threshold (count) value will differ depending on the number of reads (depth) in each file, for a given probability of error. For each sample, output of the ''Deep Threshold Tool'' lists the loci detected at above threshold value and these were then analyzed using the Mutation Reporter Tool, with a reference motif being the corresponding consensus sequences for each genotype or sub-genotype. The distribution of substitutions at the nucleotide level in the BCP/PC/C region varied between samples, depending on the HBV genotype and HBeAg status (Figure 9). At 0.5% probability of error or above, substitutions were identified at 39 unique positions in the four samples:31 in the X region (1674 to 1838 from the *Eco*R1 site; 165 nucleotides), three in the PC region (1814 to 1900; 87 nucleotides) and five in the core region (1901 to 1939; 39 nucleotides) (Figure 9). Ten of the 39 positions were present in at least two samples.

Based on the fact that direct sequencing is capable of detecting substitutions occuring in ≥20%, of the quasispecies population substitutions were classified as high frequency (≥20%) and low frequency substitutions (<20%). High frequency substitutions were found at 11 positions and low frequency at 28 positions.

Probability of Error	1678	1680	1706	1724	1725	1727	1728	1730	1736	1739	1741	1742	1745	1748	1751	1753	1761	1772	1773	1775	1842	1896	1909	1912	1913
0.0100 (1.0%)						•				•		•		•		•		•		•			•	•	•
0.0090 (0.9%)				•		•				•		•		•		•		•		•			•	•	•
0.0080 (0.8%)	•			•		•	•			•	•	•		•		•		•		•	•		•	•	•
0.0070 (0.7%)	•			•	•	•	•			•	•	•	•	•		•		•		•	•		•	•	•
0.0060 (0.6%)	•	•		•	•	•	•		•	•	•	•	•	•		•	•	•		•	•	•	•	•	•
0.0050 (0.5%)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Figure 5. The second of two summary output tables provided by the ''Deep Threshold Tool''. For each probability of error in the range specified (shown in reverse order in this table), a bullet is shown in the corresponding column of the table for each interesting column at which at least one mutation occurred at above-threshold frequency. This table can be consulted to determine the probability of error, which should be used on a given dataset. In this example, the well-characterized positions 1753, 1773 and 1896 are examined, and a probability of error of 0.005 selected, as this is the highest probability of error at which above-threshold mutations at the three positions are detected.
doi:10.1371/journal.pone.0095377.g005

Show/Hide Nucleotides | Show/Hide Amino Acids

ID	1814->	1	1817->	2	1820->	3	1823->	4	1826->	5	1829->	6
H6E7F4R01BUX0T	ATG	M	CAA	Q	CTT	L	TTT	F	CAC	H	CTC	L
H6E7F4R01AXWWO	ATG	M	CAA	Q	CTT	L	GTT	V	CAC	H	CTC	L
H6E7F4R01A9N1P	ATG	M	CAA	Q	CTT	L	TTT	F	CAC	H	CTC	L
H6E7F4R01CW0K1	ATG	M	CAA	Q	CTT	L	TTT	F	CAC	H	CTC	L
H6E7F4R01AKB00	ATG	M	CAA	Q	CTT	L	TTT	F	CAC	H	CTC	L
H6E7F4R01BNNCJ	ATG	M	CAA	Q	CTT	L	TTT	F	CAC	H	CTC	L
H6E7F4R01CBND9	ATG	M	CAA	Q	CTT	L	TTT	F	CAC	H	CTC	L
H6E7F4R01C3746	ATG	M	CAA	Q	CTT	L	TTT	F	CAC	H	CTC	L

Figure 6. The first output table of the ''Rosetta Tool'', showing codon (triplets), followed by single-letter translated amino acids, for each read in the input FASTA alignment. This alignment can be used to easily locate mutations of interest and to locate synonymous and non-synonymous mutations. The visibility of the nucleotide and/or amino acids columns can be toggled on or off, as required.
doi:10.1371/journal.pone.0095377.g006

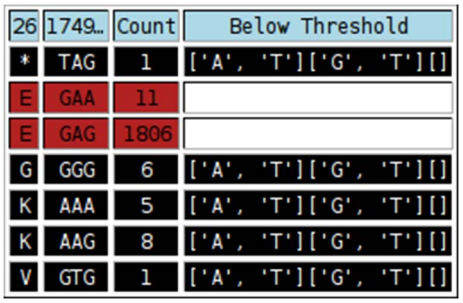
A consensus of genotype E was used to identify substitutions in genotype E (samples #1 and #4). The T1741C substitution was detected in both samples at a high frequency, regardless of the HBeAg status, while the following substitutions: A1757G, A1762T, G1764A, G1896A, G1937A/T and A1938C, were found at a high frequency in HBV from a HBeAg-negative patient (sample #1) (Figure 9). Substitutions A1735G, G1742A, A1747C, T1753C and T1909C were found at a low frequency in sample #1, and T1707C was found at a low frequency in sample #4 (Figure 9).

Similarly, when the genotype D sequences (samples #2 and #3) were compared to their corresponding consensus sequence, 1678T was found in sample #2 and 1678C in sample #3. The consensus of genotype D had 1678T. From phylogenetic analysis carried out in our previous study, HBV from sample #2 belongs to subgenotype D1 and from sample #3 to subgenotype D6 [17]. The consensus of subgenotype D1 has T at 1678 and that of

Number of reads: 2850
Threshold count: 167

AA	Codon	1814	1817	1820	1823	1826	1829	1832	1835	1838	1841	1844	1847	1850	1853
*	TAA		5												
*	TAG														
*	TGA												1		
-	-AG														
-	-GG														
-	-TC									1					
-	C-T			1											
-	CC-														
-	CT-								1						
-	T-G														
-	TC-														
-	TG-												1		
A	GCA													1	
A	GCC														
A	GCG														
A	GCT														
C	TGC							2824					14		5
C	TGT							7					2825		2830

Figure 7. The second output table of the ''Rosetta Tool'' (truncated), showing the frequency of each possible codon (triplet) and where it occurs in the alignment. Frequencies shown with a black background occurred at below-threshold levels and therefore can be disregarded.
doi:10.1371/journal.pone.0095377.g007



26	1749…	Count	Below Threshold
*	TAG	1	['A', 'T']['G', 'T'][]
E	GAA	11	
E	GAG	1806	
G	GGG	6	['A', 'T']['G', 'T'][]
K	AAA	5	['A', 'T']['G', 'T'][]
K	AAG	8	['A', 'T']['G', 'T'][]
V	GTG	1	['A', 'T']['G', 'T'][]

Figure 8. Examples of the final series of tables output by the "Rosetta Tool", showing details of the codons (triplets) and amino acids occurring at each position in the alignment. Cells with black backgrounds indicate where at least one nucleotide in the triplet occurred at below-threshold levels. These rows can be disregarded. The "Below Threshold" column lists the residues, for each position of the codon (indicated by the square brackets), which were below the. threshold.
doi:10.1371/journal.pone.0095377.g008

subgenotype D6 has C. Therefore, when sample #3 was compared to the consensus of subgenotype D6, only low frequency substitutions were detected (T1696C, G1733A, G1745A, G1748, G1751A, G1756A and T1765C) (Figure 9). When the reference sequence was changed from the D to D1, the mutation pattern of sample #2 (subgenotype D1), changed (Figures 9). Using either reference sequence D or D1, the following substitutions were detected with high frequency: A1727G, C1730A, A1761C, G1764A, A1775G and G1896A, whereas the frequency of 1773T and 1912T decreased when using D1 instead of D as the reference sequence (Figure 9). The following substitutions relative to D1, occurred in sample #2 at low frequency: T1678C, A1680C, C1706T, T1724C, A1725C, G1728A, G1736A, G1739C/T, T1741C, G1745A, G1748A, G1751, T1753G, A1772T, T1773C, T1842C, T1909C, T1912C and C1913G.

Summarizing the above, in the four samples substitutions were identified at 39 unique positions. Sample #2 (HBeAg-negative, genotype D) had substitutions in 26 positions, followed by sample #1 (HBeAg-negative, genotype E) in 12 positions, sample #3 (HBeAg-positive, genotype D) in 7 positions and sample #4 (HBeAg-positive, genotype E) in only four positions. The ratio of nucleotide substitutions between isolates from HBeAg-negative and HBeAg-positive patients was 3.5:1. Moreover, genotype D isolates showed greater variation in the X, PC and core regions, compared to genotype E isolates, with the two genotype D samples having 33 substitutions compared to the 16 detected in the genotype E samples.

The "Rosetta Tool", which was developed as part of this study, was used to analyze sequence data at the amino acid level. Substitutions identified at the nucleotide level were translated into amino acids and classified as synonymous or non-synonymous. Fourteen substitutions, 12 in the X region and 2 in the C region, were synonymous. Twenty-five, 19 in the X region, three each in the PC and C regions, were non-synonymous mutations. All non-synonymous mutations occurred within single, non-overlapping reading frames (1653 to 1814, and 1839 to 1939 from the *Eco*R1 restriction site), and the region between the start of the PC and the end of the X (1814 to 1838) was completely conserved in all ultradeep pyrosequences.

Most of the 21 insertions found in Dataset 2 occurred within homopolymeric regions and were therefore considered to be PCR or pyrosequencing artefacts [23].

Analysis of CBS and Comparison to UDPS

At least 20 clones were generated per sample. The BCP/PC region sequenced is relatively short and does not differentiate genotypes D and E following phylogetic analysis. Both identical and multiple clones were generated, with HBV from HBeAg-negative sera showing greater divergence (Figure 10).

CBS data was analyzed at the 39 loci, previously recognized by UDPS, using the Mutation Reporter Tool and a consensus sequence for each genotype/subgenotype as the reference sequence. In the four samples, substitutions at 18 of the 39 positions (46.2%) were detected by CBS (Table 2) (Figure 11). CBS detected all high frequency substitutions but only 25% (7/28) of the low frequency substitutions (Table 2). Moreover, the following nucleotide substitutions were detected in different samples by either UDPS or CBS, at position:

1. 1707 by CBS in sample #2 and by UDPS in sample #4,
2. 1775 by CBS in sample #1 and by UDPS in sample #2 and
3. 1842 by CBS in sample #3 and by UDPS in sample #2.

For the four samples, CBS detected only 14 of the 25 non-synonymous mutations, detected by UDPS (58.3%).

Discussion

The aims of this study were to build bioinformatics tools to assist in determining the threshold at which pyrosequencing data should be analyzed, and to compare quasispecies distributions obtained using UDPS and CBS, and compare results between UDPS and CBS using HBeAg-positive and –negative sera infected, with either genotype D or E.

Direct (Sanger) sequencing produces a single "read" for each sample. After curating the sequence and resolving ambiguous bases, the sequence is ready for further downstream processing. Whilst UDPS, which generates several thousand reads per sample, is a powerful technology, the analysis of the read data before downstream processing is critical. The depth of coverage provided by UDPS is also one of its shortcomings, as the data needs to be carefully curated for errors (artefacts), which may have been introduced by the PCR amplification and/or the sequencing process [2,23]. The increased sensitivity of the platform to detect thousands of reads also means that it may generate such artefacts.

A probability of error of between 0.5% and 1% for UDPS has been used previously for HIV samples [6]. Subsequent studies on HBV sequence data have either used the same probability of error, or have not reported details of this component of the analysis [2,3,24]. The probability of error, which is used, will influence the downstream detection of variants. As such, selecting an appropriate probability of error is an essential step in the analysis. In response to the lack of consensus in selecting a probability of error and determining a threshold, we developed an online bioinformatics tool to explore this aspect of the analysis. The "Deep Threshold Tool" provides the researcher with detailed output of variation at different probabilities of error. The analysis is objective and repeatable, and the selected probability of error can be reported and defended. Data for a project can be processed by the tool, so that a probability of error can be selected for that specific project, organism or assay. Using a fixed, predetermined probability of error for the UDPS platform as a whole is overly-broad and too general, as it is not possible to indicate how a particular probability of error would be applicable to a different organism, genomic region or investigation. Using the "Deep Threshold Tool" developed in the present study, a probability of

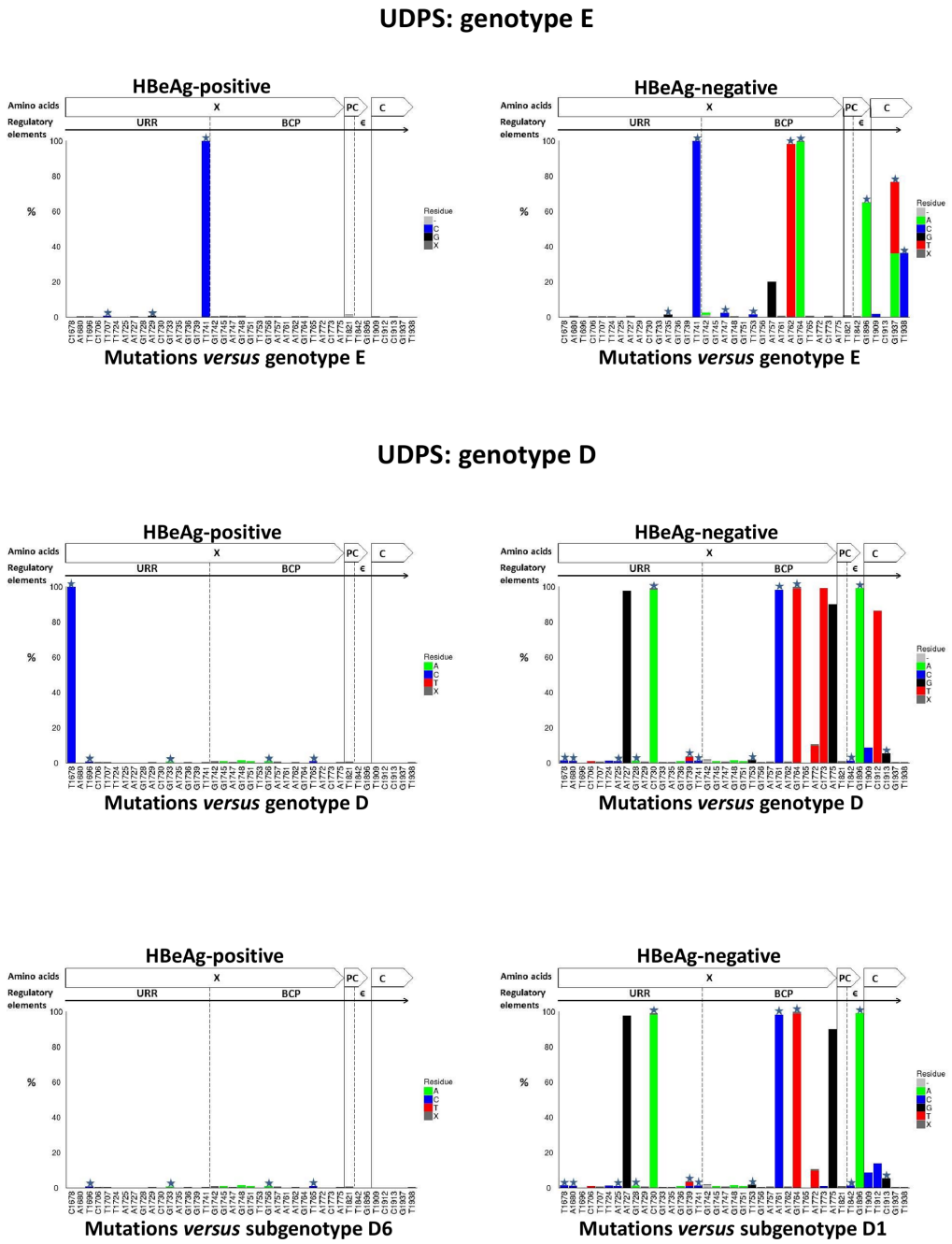


Figure 9. Graphs showing mutation distribution of the UDPS data at the nucleotide level using either genotype E or D consensus sequence as the reference. A star indicates a non-synonymous mutation. The graphs were built using the Mutation Reporter Tool [22].
doi:10.1371/journal.pone.0095377.g009

error of 0.5% was selected for the BCP/PC/C region of HBV, which agrees with previous reports for HIV [6].

The output must be interpreted in light of existing biological knowledge of the variation known to occur in the sequenced region. The tool is objective and outputs results for different probabilities of error "blindly". There is no "right answer" or absolute correct threshold, as we cannot possess complete knowledge of all the stochastic processes, from the sample to the PCR to the sequencing platform to the sequence results. Variation may be introduced at the various PCR stages, rather than by the sequencing hardware itself [23]. What we can do, however, is to interrogate these data at different probabilities of error, and make an informed decision on which value to select. It is important that the method used to process and curate the UDPS data, as well as any numerical values used (such as probability of error or threshold), be reported in all UDPS studies. Failure to provide this level of detail makes it difficult to accurately assess and relate any results reported.

The emergence of G1896A mutation in the PC region is known to be associated with HBeAg seroconverion [13]. The presence of wild-type (G) at 1896 in sample #1 and sample #2,which were isolated from HBeAg-negative patients, confirms the ability of UDPS to detect minor populations, which may not be detected by Sanger sequencing [25,26]. Similar results have been reported in

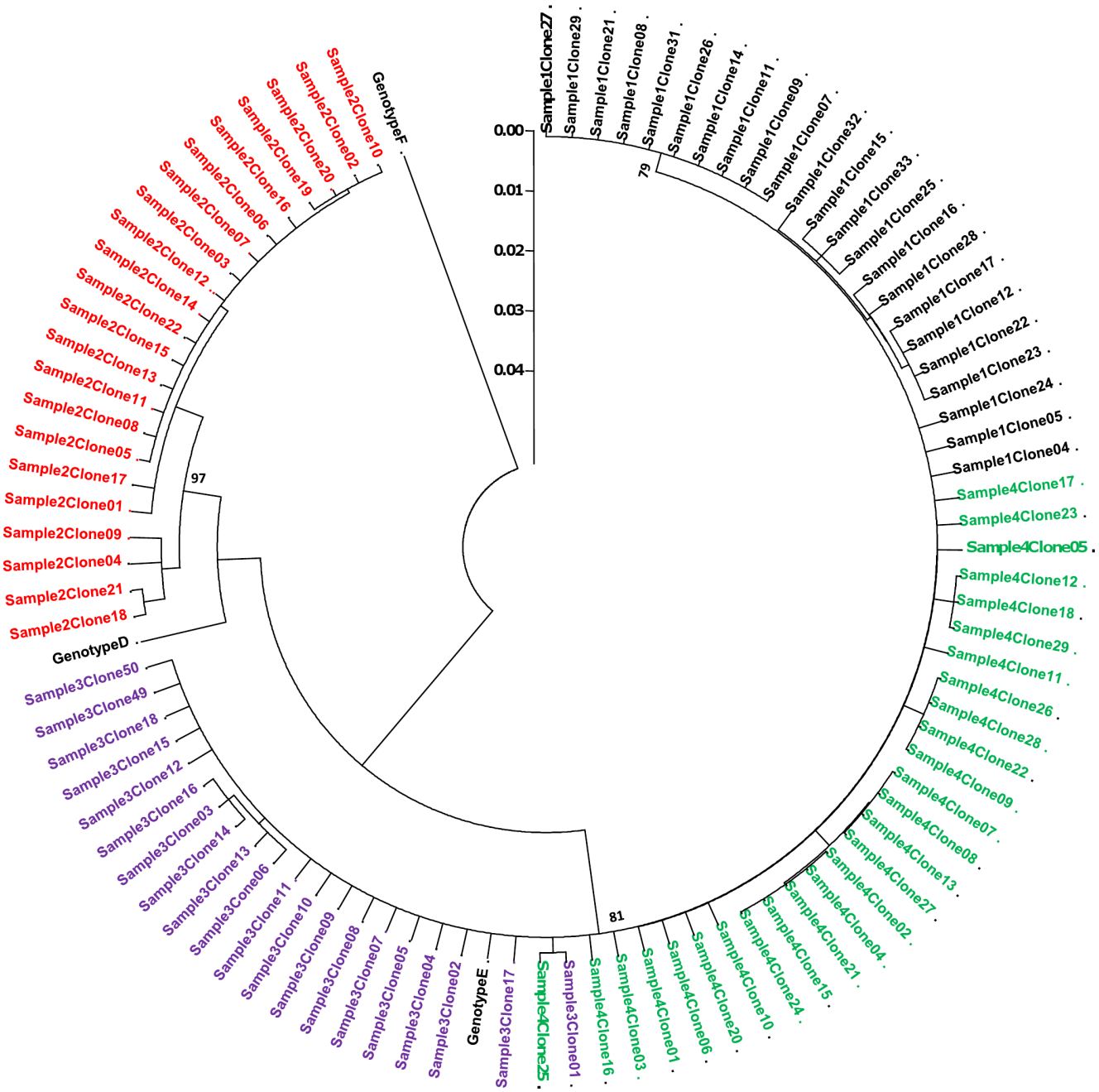


Figure 10. A rooted phylogenetic tree of 92 cloned BCP/PC sequences (position 1653 to 1939 from *EcoR*I site) from four serum samples. Sample #1 was HBeAg-negative and infected with genotype E of HBV, sample #2 was HBeAg-negative, genotype D, sample #3 was HBeAg-positive, genotype D and sample #4 was HBeAg-positive, genotype E. Bootstrap statistical analysis was performed using 1000 datasets, indicated as percentages on the nodes. The letters, D and E, represent the genotypes.
doi:10.1371/journal.pone.0095377.g010

more recent HBV studies. The HBV population from HBeAg-positive sera showed a high percentage of stop codon mutations in the precore region, while isolates from HBeAg-negative carriers had a low percentage of wild-type residues at codon 28 [24].

Although the selection of genotype D samples was random, we later discovered that sample #3 belonged to subgenotype D6, while sample #2 belonged to subgenotype D1. As illustrated in Figures 9 and 10, knowledge of the genotype and subgenotype of HBV is important when determining the presence of mutations. Depending on the reference or consensus sequence used, the variant at a particular position, may either represent the signature of a particular subgenotype or be a legitimate mutation. Therefore, where possible, a consensus sequence of the genotype or subgenotype should be used, to ensure that variants are examined in the appropriate context.

Six mutations (A1757G, A1762T, G1764A G1896A, G1937A/T and A1938C) were found in high frequency (>20%) in sample #1, genotype E isolated from a HBeAg-negative patient. The G1896A mutation is known to create the stop codon at amino acid 28 and to abrogate HBeAg expression [13], while the double

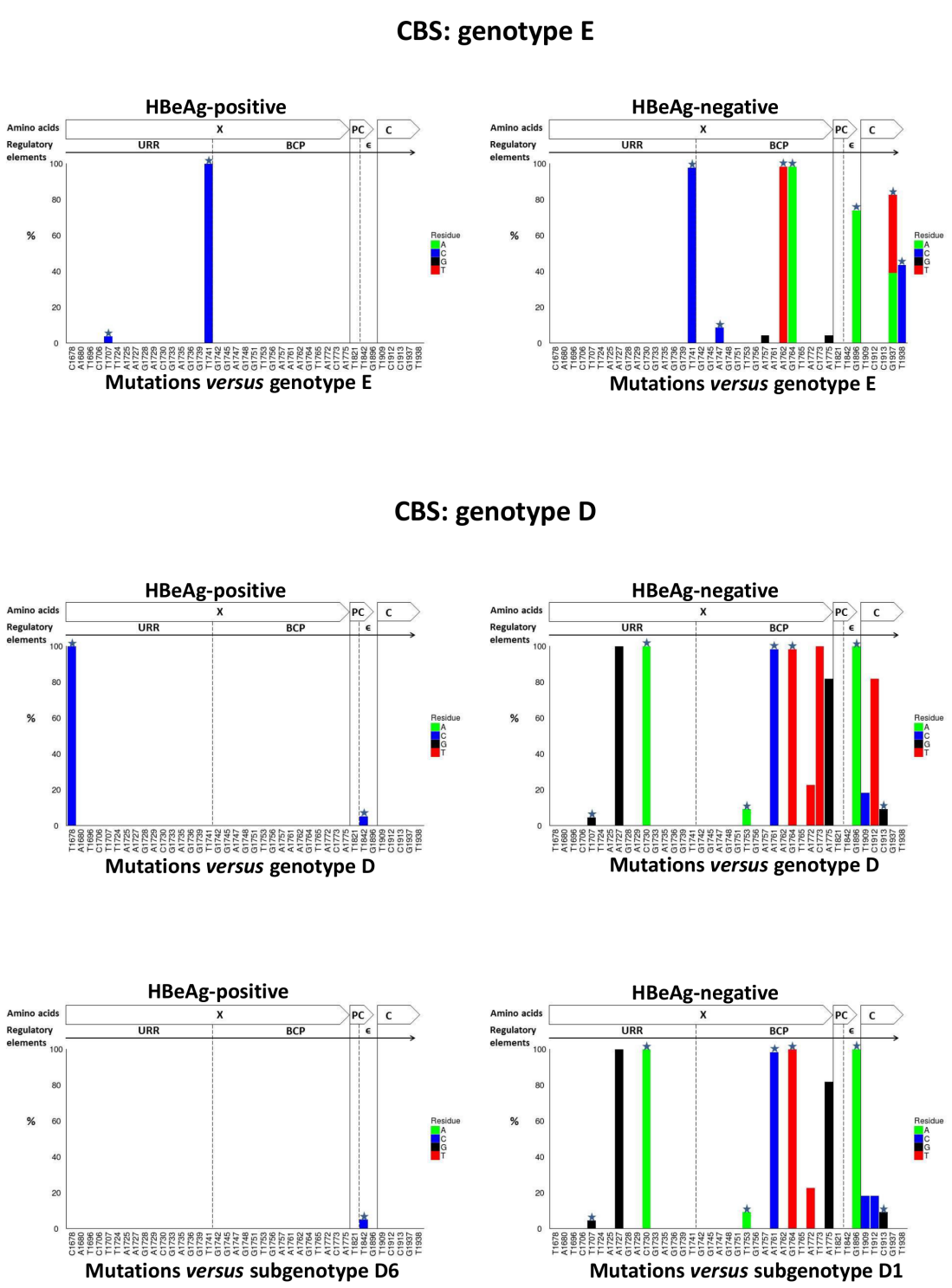


Figure 11. Graphs showing mutation distribution of the CBS data at the nucleotide level using either genotype E or D consensus sequence as the reference. A star indicates a non-synonymous mutation. The graphs were built using the Mutation Reporter Tool [22].
doi:10.1371/journal.pone.0095377.g011

mutation A1762T/G1764A is known to down-regulate the transcription of precore mRNA that is translated into HBeAg [14]. Although A1757G is a synonymous mutation and thus has no effect on the protein sequence, it overlaps *cis*-regulatory elements within the basic core promoter. In the present study, 1757G was found to be associated with A1762T/G1764A. This association has also been shown by others, who found that chronic hepatitis patients infected with HBV with 1757G/1762A1764A had higher HBV DNA levels compared to patients infected with the wild-type 1757A/1762T1764A [27]. Moreover, A1757G was found to in HCC patients infected by genotype C [28]. Non-synonymous mutations G1937A/T and T1938C within the core region occurred at a high frequency (Figure 9). These mutations are located within a T-cell epitope, which is an important

Table 2. Mutations detected by CBS and UDPS.

Frequency	Detected by	Mutations
High (≥20%)	UDPS and CBS	A1727G,C1730A,T1741C,A1757G,A1761C,A1762T,G1764A/T,A1775G,G1896A,G1937A/T,T1938C
Low (<20%)	UDPS and CBS	T1707C, A1747C, T1753G, A1772T, T1909C,T1912C, C1913G
Low (<20%)	UDPS	T1678C, A1680C, T1696C, C1706T, T1724C, A1725C, G1728A, A1729G, G1733A, A1735G, G1736A, G1739C/T, T1741C, G1742A, G1745A, G1748A, G1751A, T1753C, G1756A,T1765C, T1733C, T1842C

doi:10.1371/journal.pone.0095377.t002

component of the host's immune response to HBV infection [29]. These two mutations have recently been reported in strains of HBV genotype B isolated from Taiwanese patients [30]. Other substitutions (T1707C, A1735G, A1747C and T1909C) were found at low frequencies (<20%) and have not been reported in previous studies.

In sample #2 (genotype D isolated from HBeAg-negative), mutations A1727G, C1730A, A1761C, G1764A, A1775G and G1896A were detected at high frequency. A1727G and C1730A are located in the Enhancer II region and have been detected in cirrhotic patients [28] and are associated with reduced HBcAg expression and HBV DNA levels in the liver [31]. A1761C has previously been detected within a mutational motif (1761–1766) in isolates from patients with cirrhosis and chronic hepatitis [32]. The A1775G is associated with loss of HBeAg in Taiwanese children [33]. T1678C, G1753A and T1773C, which were found in the minority of the quasispecies population, have previously been associated with severity of HBV infection and progression to HCC [28,34].

The following substitutions were found as minor populations and have not previously been documented. In HBV from HBeAg-negative samples: A1735G, G1742A, A1747C and T1909C in genotype E and A1680C, C1706T, T1724C, A1725C, G1728A, G1736A, G1739C/T, G1751A, A1772T, T1842, T1909C, T1912C and C1913G in genotype D (Figure 9) and in HBeAg-positive samples: T1696C, G1733A and G1751A in genotype D and T1707C in genotype E. Mutations G1745A and G1748A were found in both HBeAg-negative and HBeAg-positive genotype D samples. It is possible that these have not previously been detected because direct (Sanger) sequencing can only detect variation that occurs in 20% or more of the population. More extensive studies may reveal the relevance of these minor variants.

The genotype E isolates were found to harbour fewer mutations in the X, PC and core regions compared to genotype D, which is in agreement with previous studies showing low genetic diversity of genotype E [35,36]. Furthermore, a greater number of mutations were found in HBeAg-negative samples of both genotype D and E compared to HBeAg-positive samples. It was reported that the frequency of HBV mutations is higher in HBeAg-negative patients, this is as a result the immune response of the host against the virus before the loss of HBeAg [37]. However, because only four samples, belonging to the two genotypes from HBeAg-positive and HBeAg-negative samples, were analyzed, additional samples would be required before any firm conclusions can be reached about the differences in nucleotide divergences between these genotypes from HBeAg-positive and –negative sera.

In this study, where 9738 sequence reads were generated by UDPS, 39 unique positions were detected by UDPS, while only 18 (46.2%) of these position were detected by CBS. High frequency substitutions were found in 11 positions and were all detected by CBS, whereas only 6/28 (25%) low frequency substitutions were detected by CBS ($p<0.05$) (Figures 9 and 10).

Although the testing of the tools was done on a small sample set and the findings cannot be generalized, it is evident that the data generated by the increased read-depth provided by UDPS should be approached with caution. Appropriate curation and examination of the reads are required to ensure that artefacts are not interpreted as variants. Moreover, identification of variants must be performed against a suitable reference or consensus sequence, as a "mutation" of interest may simply be a known signature or variant when examined in the correct genotypic or subgenotypic context. UDPS detected a greater number of substitutions than CBS. Relative to CBS, UDPS is cheaper to undertake, both in terms of time and expense. However, without rigorous and careful examination and interpretation of read data, the results generated by UDPS may be misleading. As illlustrated in the present study, a thorough knowledge of the genome of interest and its known variants is essential in order to accurately and reliably interpret the high resolution read data generated by UDPS.

Acknowledgments

We thank Justin Southey for guidance with some aspects of the statistical calculations.

Author Contributions

Conceived and designed the experiments: AK MY. Performed the experiments: MY. Analyzed the data: MY TB AK. Contributed reagents/materials/analysis tools: AK TB HM DG. Wrote the paper: MY TB AK. Read and approved the paper: MY TB HM DG AK. Designed the analysis algorithms: MY TB. Developed and wrote the tools and maintained the server: TB. Analysis the data: MY TB AK.

References

1. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74: 5463–5467.
2. Solmone M, Vincenti D, Prosperi MC, Bruselles A, Ippolito G, et al. (2009) Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. J Virol 83: 1718–1726.
3. Margeridon-Thermet S, Shulman NS, Ahmed A, Shahriar R, Liu T, et al. (2009) Ultra-Deep Pyrosequencing of Hepatitis B Virus Quasispecies from Nucleoside and Nucleotide Reverse-Transcriptase Inhibitor (NRTI)–Treated Patients and NRTI-Naive Patients. Journal of Infectious Diseases 199: 1275–1285.
4. Ramírez C, Gregori J, Buti M, Tabernero D, Camós S, et al. (2013) A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. Antiviral Research.
5. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature biotechnology 26: 1135–1145.
6. Eriksson N, Pachter L, Mitsuya Y, Rhee S-Y, Wang C, et al. (2008) Viral population estimation using pyrosequencing. PLoS Computational Biology 4: e1000074.

7. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8: R143.
8. Kramvis A, Arakawa K, Yu MC, Nogueira R, Stram DO, et al. (2008) Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. J Med Virol 80: 27–46.
9. Kurbanov F, Tanaka Y, Mizokami M (2010) Geographical and genetic diversity of the human hepatitis B virus. Hepatology Research 40: 14–30.
10. Kramvis A, Kew MC (1999) The core promoter of hepatitis B virus. J Viral Hepat 6: 415–427.
11. Revill P, Yuen L, Walsh R, Perrault M, Locarnini S, et al. (2010) Bioinformatic analysis of the hepadnavirus e-antigen and its precursor identifies remarkable sequence conservation in all orthohepadnaviruses. J Med Virol 82: 104–115.
12. Jean-Jean O, Levrero M, Hans W, Perricaudet M, Rossignol J-M (1989) Expression mechanism of the hepatitis B virus (HBV) C gene and biosynthesis of HBe antigen. Virology 170: 99–106.
13. Carman W, Hadziyannis S, McGarvey M, Jacyna M, Karayiannis P, et al. (1989) Mutation preventing formation of hepatitis B e antigen in patients with chronic hepatitis B infection. The Lancet 334: 588–591.
14. Okamoto H, Tsuda F, Akahane Y, Sugai Y, Yoshiba M, et al. (1994) Hepatitis B virus with mutations in the core promoter for an e antigen-negative phenotype in carriers with antibody to e antigen. Journal of Virology 68: 8102–8110.
15. Seeger C, Mason WS (2000) Hepatitis B virus biology. Microbiology and Molecular Biology Reviews 64: 51-+.
16. Mizokami M, Orito E, Ohba K, Ikeo K, Lau JY, et al. (1997) Constrained evolution with respect to gene overlap of hepatitis B virus. J Mol Evol 44 Suppl 1: S83–90.
17. Yousif M, Mudawi H, Bakhiet S, Glebe D, Kramvis A (2013) Molecular characterization of hepatitis B virus in liver disease patients and asymptomatic carriers of the virus in Sudan. BMC Infectious Diseases 13: 328.
18. Takahashi K, Aoyama K, Ohno N, Iwata K, Akahane Y, et al. (1995) The precore/core promoter mutant (T1762A1764) of hepatitis B virus: clinical significance and an easy method for detection. J Gen Virol 76 (Pt 12): 3159–3164.
19. R Development Core Team (2005) R: A language and environment for statistical computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. Available: http://www.R-project. org.
20. Settles M, Hunter S, Sarver B, Zhbannikov IaC, Kyu-Chul (2011) rSFFreader: rSFFreader reads in sff files generated by Roche 454 and Life Sciences Ion Torrent sequencers. R package version 080.
21. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32: 1792–1797.
22. Bell TG, Kramvis A (2013) Mutation Reporter Tool: An online tool to interrogate loci of interest, with its utility demonstrated using hepatitis B virus. Virology journal 10: 62.
23. Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, et al. (2013) PCR-Induced Transitions Are the Major Source of Error in Cleaned Ultra-Deep Pyrosequencing Data. Plos One 8: e70388.
24. Homs M, Buti M, Quer J, Jardí R, Schaper M, et al. (2011) Ultra-deep pyrosequencing analysis of the hepatitis B virus preCore region and main catalytic motif of the viral polymerase in the same viral genome. Nucleic acids research 39: 8457–8471.
25. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, et al. (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic acids research 35: e91.
26. Ijaz S, Arnold C, Dervisevic S, Mechurova J, Tatman N, et al. (2008) Dynamics of lamivudine-resistant hepatitis B virus during adefovir monotherapy versus lamivudine plus adefovir combination therapy. J Med Virol 80: 1160–1170.
27. Sendi H, Mehrab-Mohseni M, Zali MR, Norder H, Magnius LO (2005) T1764G1766 core promoter double mutants are restricted to Hepatitis B virus strains with an A1757 and are common in genotype D. Journal of General Virology 86: 2451–2458.
28. Yin J, Xie J, Liu S, Zhang H, Han L, et al. (2010) Association between the various mutations in viral core promoter region to different stages of hepatitis B, ranging of asymptomatic carrier state to hepatocellular carcinoma. The American journal of gastroenterology 106: 81–92.
29. Radecke K, Protzer U, Trippler M, Meyer zum Büschenfelde KH, Gerken G (2000) Selection of hepatitis B virus variants with aminoacid substitutions inside the core antigen during interferon-α therapy. Journal of medical virology 62: 479–486.
30. Wu J-F, Ni Y-H, Chen H-L, Hsu H-Y, Chang M-H (2013) The Impact of Hepatitis B Virus Precore/Core Gene Carboxyl Terminal Mutations on Viral Biosynthesis and the Host Immune Response. Journal of Infectious Diseases: jit638.
31. Zhu R, Zhang H-P, Yu H, Li H, Ling Y-Q, et al. (2008) Hepatitis B virus mutations associated with in situ expression of hepatitis B core antigen, viral load and prognosis in chronic hepatitis B patients. Pathology - Research and Practice 204: 731–742.
32. Veazjalali M, Norder H, Magnius L, Jazayeri S, Alavian S, et al. (2009) A new core promoter mutation and premature stop codon in the S gene in HBV strains from Iranian patients with cirrhosis. Journal of Viral Hepatitis 16: 259–264.
33. Ni Y-H, Chang M-H, Hsu H-Y, Tsuei D-J (2004) Longitudinal study on mutation profiles of core promoter and precore regions of the hepatitis B virus genome in children. Pediatric research 56: 396–399.
34. Ouneissa R, Bahri O, Alaya-Bouafif NB, Chouaieb S, Yahia AB, et al. (2012) Frequency and clinical significance of core promoter and precore region mutations in Tunisian patients infected chronically with hepatitis B. Journal of Medical Virology 84: 1719–1726.
35. Hübschen JM, Andernach IE, Muller CP (2008) Hepatitis B virus genotype E variability in Africa. Journal of Clinical Virology 43: 376–380.
36. Kramvis A, Restorp K, Norder H, Botha JF, Magnius LO, et al. (2005) Full genome analysis of hepatitis B virus genotype E strains from South-Western Africa and madagascar reveals low genetic variability. Journal of Medical Virology 77: 47–52.
37. Desmond CP, Gaudieri S, James IR, Pfafferott K, Chopra A, et al. (2012) Viral adaptation to host immune responses occurs in chronic hepatitis B virus (HBV) infection, and adaptation is greatest in HBV e antigen-negative disease. Journal of Virology 86: 1181–1192.