




## Intra-host quasispecies reconstructions resemble inter-host variability of transmitted chronic hepatitis B virus strains

Le Clercq, L.S. <sup>1,2,✉</sup>, Bowyer, S.M. <sup>3,4</sup>, and Mayaphi, S.H. <sup>3,4</sup>

<sup>1</sup>South African National Biodiversity Institute, Pretoria, 0001, South Africa

<sup>2</sup>Department of Genetics, University of the Free State, Bloemfontein, 9300, South Africa.

<sup>3</sup>Department of Medical Virology, University of Pretoria, Pretoria, 0001, South Africa.

<sup>4</sup>Tshwane academic division, National Health Laboratory Services, Pretoria, 0001, South Africa.

The hepatitis B virus is a partially double stranded DNA virus in the Hepadnaviridae family of viruses that infect the liver cells of vertebrates including humans. The virus replicates through the reverse transcription of an RNA intermediate by a viral polymerase, akin to retroviruses. The viral polymerase has high replication capacity but low fidelity and no proofreading activity resulting in a high mutation rate. This contributes to the emergence of a cloud of mutants or quasispecies within host systems during infection. Several host and viral factors have been identified that contribute to mutations and mutation frequency in shaping viral evolution, however, because the dynamics of viral evolution cannot be understood from the fittest strain alone, the need exists to sequence and reconstruct intra-host diversity, recently made possible through next generation sequencing. Due to the extensive pipeline of bioinformatic analyses associated with next generation sequencing studies are needed to ascertain if quasispecies reconstruction methods and diversity measures accurately model known diversity. Here, next generation sequencing and various quasispecies reconstruction methods are used to model the natural evolution of viral populations across the full genome of hepatitis B virus strains from South Africa. This study illustrates that (i) different methods of quasispecies reconstruction reconstruct the same amount of diversity, (ii) intra-host diversity derived from full quasispecies analyses resembles diversity measures obtained from previous methods, (iii) inter-host diversity resembles the diversity between closely related quasispecies variants, (iv) diversity is increased in HIV-negative individuals, and (v) corroborate that seroconversion of HBV biomarkers increases mutation rates.

Quasispecies | Reconstruction, Hepatitis B Virus | Variation | Next Generation Sequencing | Intra-host | Inter-host | QuRe | QuasiRecomb | k-GEM

Correspondence: ✉ [leclercq.l.s@gmail.com](mailto:leclercq.l.s@gmail.com)

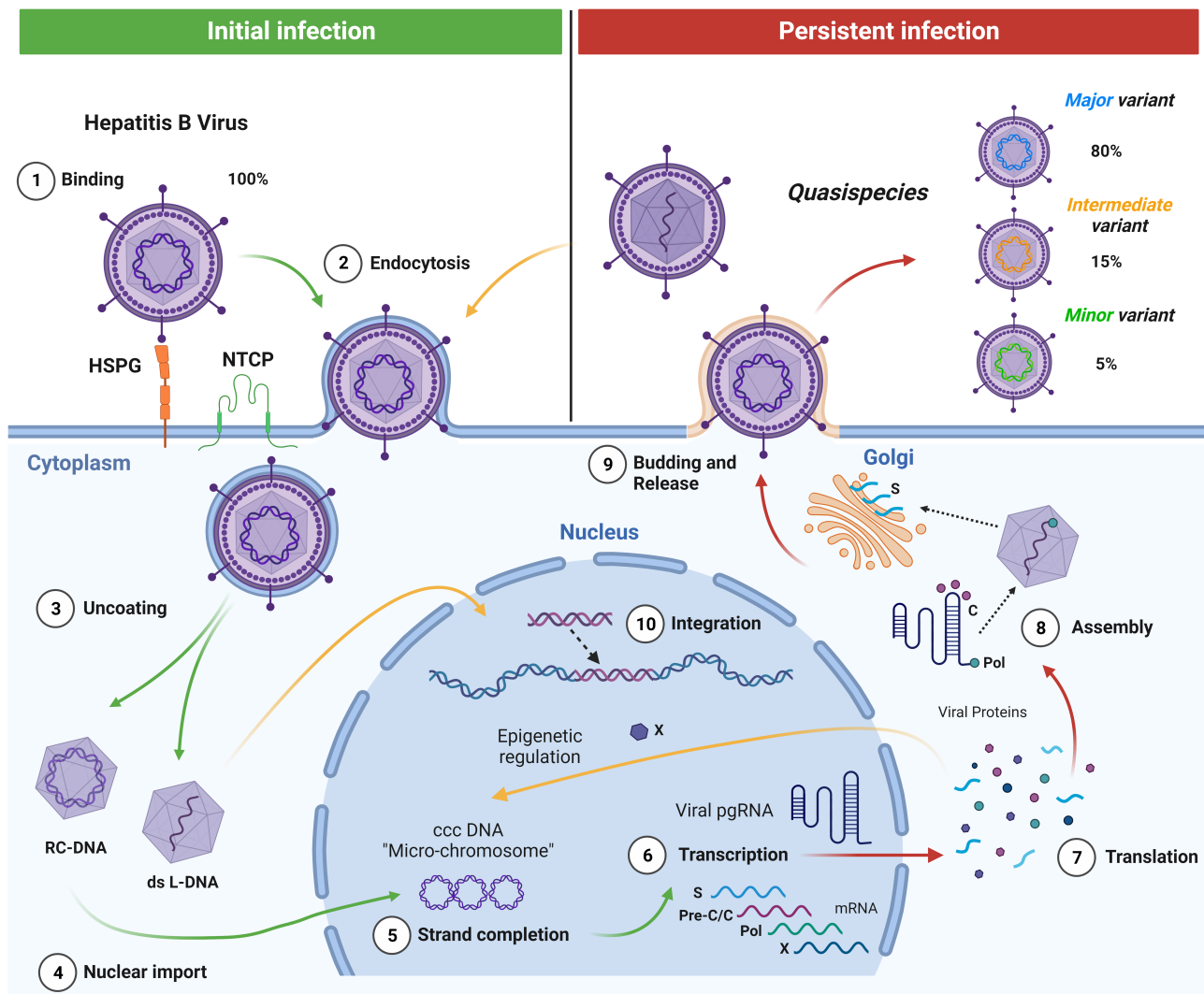
First Published: 15 May 2023 | Revised: 27 May 2023

### Introduction

The Hepatitis B Virus (HBV), Orthohepadnavirus homonensis, is a deoxyribonucleic acid (DNA) virus that infects humans, first discovered in 1965 by Blumberg [5, 6]. It is one of at least 18 species within the Hepadnaviridae family

which has five genera, two of which are well characterised (Ortho- and Avihepadnaviridae) and three (Meta-, Herpeto-, and Parahepadnaviridae) which have only been recognised recently: with new members still being discovered [44, 32, 50]. These viruses infect the hepatocytes of the liver in the five most common classes of vertebrates: mammals, birds, fish, reptiles, and amphibians. HBV is specifically classified in the Orthohepadnavirus genus of DNA viruses infecting the liver of mammals, including the related viruses that infect many non-human primates as well as those infecting squirrels, woodchucks, bats, and equines. Studies have subclassified HBV into nine widely accepted genotypes or viral subspecies, denoted A to J, based on pair-wise differences of more than eight and less than seventeen percent across the full genome. These genotypes are further partitioned—with the exception of genotypes E, G and H—into subgenotypes or molecular subtypes, based on pairwise differences larger than four percent [42, 53]. HBV genotypes seem to follow a remarkably conserved geographic distribution, reviewed by Kramvis et al. [28].

HBV is primarily a blood-borne virus that is transmitted between individuals through direct contact of bodily fluids such as blood but can also be transmitted sexually [23] through other body fluids. Exposure can either result in an acute infection, where the host system eventually clears the infection and recovers, or may result in a chronic and persistent infection [55]. Upon exposure, the virus infiltrates host liver cells and initiates replication, protein transcription, and new virion formation. This occurs through the epigenetic regulation [8], transcription and translation of viral open reading frames from covalently closed circular HBV genome copies (“mini-chromosomes”) in the nucleus; including the core gene, surface gene, X gene, and viral polymerase. The HBV genome is replicated by the reverse transcription of a viral ribonucleic acid (RNA) intermediate, the pre-genomic RNA, following encapsidation in the cytoplasm [35, 26]. Several viral components are detectable in the blood of infected individuals and have been developed into biomarkers



**Fig. 1.** Life cycle and quasispecies formation of the hepatitis B virus. (1) Upon initial infection the virus proteins interact with membrane receptors on hepatocytes such as heparan sulphate proteoglycan (HSPG) and Na<sup>+</sup>-taurocholate co-transporting polypeptide (NTCP) to fuse with the membrane (2) and enter the host cell as a coated virion. After uncoating (3), the virus replicates by nuclear import of the relaxed circular DNA (RC-DNA) viral genome prior (4) which undergoes strand completion (5) to form covalently closed circular DNA (cccDNA) mini-chromosomes. Viral mRNA is then transcribed (6) to form the viral pre-genomic RNA (pgRNA) and several primary transcripts for the translation of viral particles and proteins (7). These proteins and particles are triggered to assemble (8) the virion through the encapsidation signal of the pgRNA. The X protein acts as an epigenetic regulator of transcription by interfacing with the micro-chromosome in the nucleus. In the assembled virion, pgRNA is reverse transcribed by viral polymerase to form the partially double-stranded DNA genome. Newly formed virions, inside specialised vesicles from the Golgi, then bud (9) through the cell membrane and are released into the blood stream. Mutations, caused by the low replication fidelity and lack of proofreading from viral polymerases, results in the accumulation of a viral quasispecies or mutant cloud in the host with several major, intermediate, and minor variants in circulation. Apical reinfection with virions that have double-stranded linear DNA (ds L-DNA) instead of RC-DNA may result in the integration (10) of viral DNA into the host genome. (image created in [BioRender.com](https://www.biorender.com))

of infection and disease progression; reviewed by Kramvis et al. [29]. Copies of HBV DNA, linearised during the reverse transcription of pre-genomic RNA, can also be incorporated into the host genome [53]. The integration of hepatitis viruses into host genomes may not maintain a chronic infection, however, it was recently illustrated in several passerine bird species, including finches, that parts of the viral DNA may become endogenous to the organism it infects and remain in their genomes long after their species are no longer common hosts [17].

RNA viruses such as the hepatitis C virus and influenza virus, and reverse transcriptase dependent viruses such as HBV and the human immunodeficiency virus (HIV), show a high degree of intra-host variability. This is likely due to the high replication capacity yet low fidelity and lack of

proofreading activity of viral polymerases; typically introducing between  $1.0 \times 10^{-5}$  and  $1.0 \times 10^{-3}$  substitutions per site per cycle in retroviruses [10]. The substitution rate, calculated per site per year, for HBV has been approximated to be between  $1.4 \times 10^{-5}$  and  $5 \times 10^{-5}$ : comparable to the rate in retroviruses but nearly 104 times higher than that of other DNA viral genomes [28]. Thus, much like RNA and retroviruses, an intra-host virus population, referred to as the viral quasispecies, arises during HBV infections. This quasispecies consists of major, intermediate and minor variants which occur at frequencies of >20, 5-20, and <5 %, respectively [13].

As selection occurs on the entire population [3], and progresses over the course of infection [16], the population dynamics cannot be understood from the fittest strain alone;

**Table 1.** Summary of serology for individuals (n = 15).

ID	Primary markers			Secondary markers		Liver ALT (U/L)	Molecular		Genotype
	HBsAg	Anti-HBs	Anti-HBc	HBeAg	Anti-HBe		Viral load (IU/mL)		
3269*	+	-	+	-	-	21	9.5 x 10 <sup>1</sup>	A1	
3274*	+	-	+	+	-	60	10.0 x 10 <sup>6</sup>	A1	
3319*	-	-	+	-	-	35	12.7 x 10 <sup>1</sup>	A1	
3358*	+	-	+	-	-	NA	NA	A1	
3658*	+	-	+	-	+	29	17.0 x 10 <sup>6</sup>	A1	
3768*	-	-	+	-	-	NA	NA	A1	
3791*	+	-	+	+	-	58	>110.0 x 10 <sup>6</sup>	A1	
4070*	+	-	+	+	-	46	>110.0 x 10 <sup>6</sup>	A1	
4312*	+	-	+	+	-	61	7.0 x 10 <sup>6</sup>	A1	
LA05*	+	-	+	-	+	NA	NA	A1	
N005	+	-	+	-	+	21	2.0 x 10 <sup>3</sup>	A1	
N011	+	-	+	+	-	28	>110.0 x 10 <sup>6</sup>	D4	
N060	+	-	+	-	+	44	15.7 x 10 <sup>1</sup>	A1	
N199	-	-	+	-	-	30	9.9 x 10 <sup>1</sup>	A1	
PO04	-	-	+	-	-	NA	NA	E	

\* HIV positive samples

necessitating the study of viral populations at the intra-host level, in addition to the inter-host level. This has previously been limited by the inability to directly sequence the full quasispecies using conventional methods, however, the advent of next generation sequencing (NGS), which allows for the mass parallel sequencing of mixed samples, has largely circumvented this. Consequently a cornucopia of tools now exist to measure quasispecies attributes [18], such as quasitools [37] or QSutils [19], or directly reconstruct a viral quasispecies e.g., k-GEM [36], ShoRAH [62], Vispa [2], ViQuas [24], QuasiRecomb [59], and QuRe [48, 47]. Of these algorithms, k-GEM, QuasiRecomb, and QuRe are implemented in Java and take similar input, allowing cross-platform scaling of quasispecies reconstruction and a direct comparison of algorithm efficacy. Furthermore, QuRe was first tested and benchmarked on actual data from HBV genotypes A and D and accurately reconstructed the viral quasispecies for HBV with improved efficacy as compared to ShoRAH [46].

Quasispecies dynamics have been studied in HBV in a myriad of ways that evaluated quasispecies and single nucleotide polymorphisms (SNPs) in relation to key treatment and disease progression outcomes such as therapeutic response or resistance to antivirals [11, 52] and seroconversion of viral proteins [12, 33, 34], among others. Several key concerns do, however, remain to be addressed. These include whether NGS can detect more SNPs than conventional methods, if quasispecies reconstructions accurately reflect known within and between host diversity, and if quasispecies reconstruction and analyses provide comparable output regardless of the specific methods used.

The aim of the present study is to characterise quasispecies dynamics across the full genomes of HBV from patients with chronic infections. Specific objectives are to (i) compare different but similar quasispecies reconstruction algorithms that are implemented in Java with regards to their fidelity in constructing similar haplotypes, (ii) compare the resolution of detecting SNPs through NGS and quasispecies reconstruction

to those typically detected using traditional methods, (iii) compare the diversity of quasispecies in relation to HIV coinfection and HBV serology markers, and (iv) provide clarity on the phylogenetic relatedness of these strains.

## Methods

### Specimen selection, serology, and ethics approval.

The specimens used in the present study (n = 15) were collected as part of a larger urban hospital cohort of HBV patients from 2007-2011 by Mayaphi et al. [38, 39] and were selected for cross-sectional full genome quasispecies analyses. This follows after presenting as atypical outliers in a phylogenetic analysis of typical HBV genotypes A1 and D from South Africa using the pre-core/core and surface genes [39]. Further samples were included of HBV genotype E from a paediatric outbreak from the same hospital [45]. Sample details were deposited to the National Centre for Biotechnology Information (NCBI) BioSample database with links to the relevant BioProject (see data availability section). These specimens represented both HIV positive (n = 10) as well as HIV negative (n = 5) patients for which other serological data related to HBV infection, including antigen and antibody tests conducted by enzyme linked immunosorbent assays (ELISA), were available (Table 1). Protocol approval was obtained from the Research Protocol Committee and ethics approval from the Research Ethics Committee of the Faculty of Health Sciences, University of Pretoria (numbers: UP35/2007 and S137/2012).

**Full genome PCR amplification and NGS.** Total viral genomic DNA was extracted from 100 µL stored plasma samples with the QIAamp MinElute Virus Spin Kit (Qiagen, Hilden, Germany), according to manufacturers' instructions, and the extracted DNA eluted to a final volume of 20 µL. DNA, of the full genomes of HBV (±3221bp), was amplified by long-range polymerase chain reaction (PCR) as previously

described by Günther et al. [21, 22] with minor modifications [15]. The PCR primers that were used were as follows: Forward, P1 (1821-1841), 5'-CTT TTT CAC CTC TGC CTA ATC A-3'; Reverse, P2 (1825-1806), 5'-AAA AAG TTG CAT GGT GCT GG-3', or P2A1 (1825-1806), 5'-AAA AAG TTG CAT GAT GAT GG-3'. Success of amplification was monitored by separating reactions by electrophoresis on a 2 percent TBE-agarose gel. PCR products were purified by means of the DNA Clean and Concentrator-25 spin column kit (Zymo Research Corp., Irvine, California, USA) according to manufacturers' instructions, prior to sequencing. Amplified and purified DNA was sent to Inqaba Biotech Inc. (Hatfield, Pretoria, South Africa) for sequencing on the Illumina MiSeq sequencing platform (Illumina (Pty) Ltd., San Diego, California, USA).

### NGS quality control, filtering, and read assembly.

The FASTQ sequence files were analysed for quality (Fig. S1) and assembled to a reference in Geneious Prime 2022 ([www.geneious.com](http://www.geneious.com)). Read quality was assessed for each file to determine quality score distributions, read length distribution, overrepresented sequences and k-MERs. Paired read files were set and interleaved into a single file per specimen and were directly used for mapping and SNP calling within Geneious. Coverage was calculated for individual concatenated paired read files using the Lander-Waterman equation where coverage (C) is calculated by multiplying read length (L) by the total read number (N) and divided by the size in base-pairs of the haploid genome (G).

$$Coverage(C) = \frac{L \times N}{G}$$

For purposes of quasispecies reconstruction the files were down-sampled to account for between specimen variation in coverage by concatenating the paired end reads using BB-Merge [9] in Geneious under standard parameter settings, allowing for the inclusion of reads with Phred-scale quality scores higher than twenty and set to remove unpaired reads. Phred-scale quality scores (Q) are computed with the following equation by taking the logarithm of the base-call error probabilities (P).

$$Phred(Q) = -10 \log P$$

Duplicate reads were subsequently removed from the concatenated read files with Dedupe V.38.37 function implemented in Geneious with a k-MER threshold of thirty. Hereafter the remaining reads were filtered by size by extracting those with a minimum length of 100 bp. The interleaved read files were mapped to references AB048703 (genotype D) for specimen N011, DQ060823 (genotype E) for PO04 and AY233283 (genotype A) for the remainder.

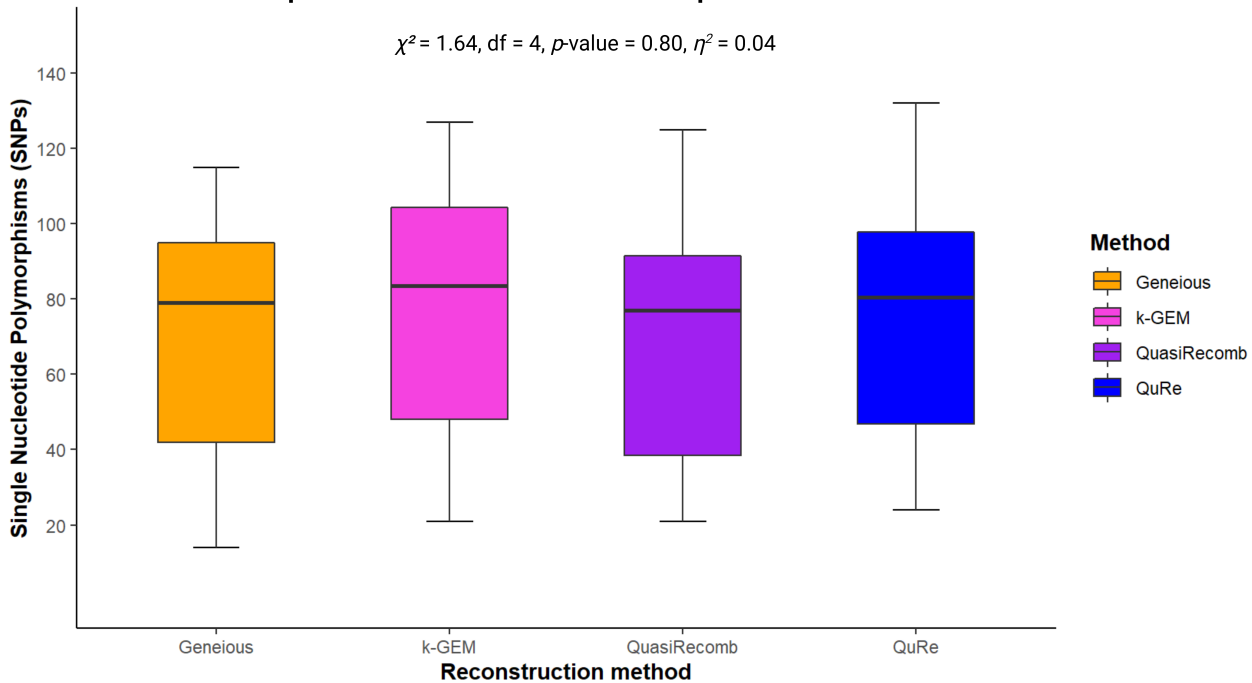
**Full genome quasispecies reconstruction.** The viral quasispecies was reconstructed for each specimen (n = 10) where sufficient read-depth or coverage was available by the three tested Java executable algorithms: QuRe V.0.99971 [48, 47], QuasiRecomb V.1.2 [3] and k-GEM V.0.3.1 [36].

Each algorithm was executed from the command prompt interface with the Java developers' kit V.1.7.0-25. The merged, de-duplicated and filtered read files in FASTA format were used as input data for QuRe, along with the appropriate references used in prior mapping. In the case of QuRe, a dictionary was built from the parsed reference and read files from which a quasi-random alignment score distribution was calculated. This was followed by an internal alignment with Jaligner and the removal of reads with an alignment p-value > 0.01. The reconstruction algorithm of QuRe was run in three phases: a fixed-size sliding window overlap, random overlaps, and assessment of the best a-posteriori overlap set. Thereafter the core reconstruction was run followed by the final clustering of the initial variants based on a random search and Bayesian Information Criterion (BIC) determination. For QuasiRecomb and k-GEM, the reduced reads mapped to references (in SAM format) were used as input data. In each instance, the input files were first parsed prior to model training and reconstruction based on the best model. QuasiRecomb was first run under default settings with automatic model selection and a range of one to five generators. This output was executed a second time with the refinement parameter. In k-GEM the reconstructions were run with a model selection between two and five generators. All results were standardized by down-sampling and using the same references across all experiments and algorithms.

**Full genome phylogenetic analysis.** Appropriate reference sequences, retrieved from GenBank as a population model, were imported into the same file as the full quasispecies for all study specimens from the three reconstruction methods. Multiple- and pair-wise alignment of the sequence file was done using MAFFT V.7.273 [27], with automatic model detection, and the resulting aligned sequences saved in PHYLIP format. A model test was conducted in MEGA X [30] to establish which model best describes the phylogeny of the data based on the BIC (Table S1). Phylogenetic analysis was done by means of the maximum likelihood method, as implemented in PhyML V.3.1 [20]. The analysis was performed under the general time reversible (GTR + I + R) model, with gamma-distributed rate variation across sites and a proportion of invariable sites and executed a second time for bootstrap analysis with 1000 replicates. The resulting tree was viewed in Figtree V.1.6.6 [49] from which it was exported in NEWICK format. Hereafter, the maximum likelihood tree was imported into MEGA for final editing.

**Quasispecies diversity estimates and statistical analysis.** Measures of quasispecies diversity [18] were measured using quasitools V.0.7.0 [37] implemented in PYTHON V.3.9, for the pooled reconstructed quasispecies, as FASTA files (n = 10), and for aligned BAM files from Geneious; for samples with incomplete coverage (n = 5). Quasispecies metrics were calculated at three levels including incidence metrics (entity level), abundance metrics (molecular level), and functional metrics (incidence level). Incidence metrics included the number of haplotypes and the total number of intra-host SNPs. Abundance metrics were computed in terms

## Box plots of inter-host SNPs detected per reconstruction method



**Fig. 2.** Box plots indicating the distributions of detected SNPs from assembled read files in Geneious or the quasispecies constructed through the three algorithms: QuRe, QuasiRecomb and k-GEM. The resulting SNPs were found to be similar to the typical pair-wise difference obtained for between individual variation in strains on GenBank and did not differ significantly between methods used ( $p\text{-value} > 0.05$ , 95% Confidence Interval (CI)).

of Shannon entropy and Gini-Simpson statistics. Three functional metrics, the mutation frequency (Mfe), functional attribute diversity (FAD), and sample nucleotide diversity ( $\pi$ ), were also computed. Statistical analyses were done using R V.4.1.3 implemented in RStudio V.2022.07.0 [57]. To ascertain if NGS increases SNP detection, a model population of HBV genomes from sequence databases was created by retrieving randomly sampled full genome sequences ( $n = 70$ ) of subgenotype A1 from the National Centre for Biotechnology Information (NCBI) GenBank resource [1], to be analysed for diversity in parallel to the study specimens. The median number of SNPs called from the mapped reads in Geneious, as well as the reconstructed quasispecies, were compared to the diversity observed for the population model from GenBank with non-parametric Kruskal-Wallis tests. For the quasispecies, SNPs were computed in comparison to the reference sequence used in mapping/reconstruction, in comparison to the closest matching reference from a basic local alignment search tool (BLAST) [61] search, and internally among variants of the same sample. The number of SNPs and number of reconstructed variants were also compared to categorical variables including HIV status and HBV serology attributes, with non-parametric Kruskal-Wallis tests.

## Results

### Reconstructed quasispecies and intra-host diversity.

Quasispecies were reconstructed for a total of ten samples with adequate coverage using each of the three reconstruction algorithms. A summary of the quasispecies and their

relative frequencies are available in Table S2. Reconstructed quasispecies from QuasiRecomb resulted in the most individual variants ( $n = 55$ ), followed by QuRe ( $n = 28$ ) and k-GEM ( $n = 22$ ). From k-GEM at least two variants, one major and one intermediate, were reconstructed whereas QuasiRecomb and QuRe only reconstructed the main variant for half of the specimens (online supplementary data) but more than five variants for other samples. From QuRe, the quasispecies of N005 was comprised of 8 variants, 3658 and N011 had five each and 3274 and N199 each had three. From specimen N005 one major variant, two intermediate and five extremely low frequency minor variants were reconstructed. The variants of specimen 3658 had a major variant, several low intermediate variants and two minor variants. Within the quasispecies of N011 there was one major variant and four minor variants. The two samples, 3274 and N199, had three variants each. Of the quasispecies reconstructed by QuasiRecomb six specimens comprised of more than one major variant. Two samples, N005 and N199 had the most diverse quasispecies with 14 and 23 variants each. Specimen 3274 had a quasispecies comprising four variants, 3791 comprised of six, and N011 and PO04 had two each. There was an overlap between QuasiRecomb and QuRe with regards to which specimens gave multiple variants and the relative number of SNPs between them. k-GEM generally gave two variants as output, one major variant which occurred at a frequency between 70 and 85 percent and one intermediate variant. There was, however, a lower degree of variability within the quasispecies from k-GEM than by other methods.

**Table 2.** Summary of NGS and quaspecies metrics (within host diversity) of study specimens

Sample ID	Coverage (C)	Phred (Q)	Incidence (Entity) Haplotypes (H)	SNPs	Abundance (Molecular) Shannon entropy (HS)	Gini-Simpson	Functional (Incidence) Mutation frequency (Mfe)	FAD	Diversity ( $\pi$ )
<i>Average:</i>			11.53	68.13	1.78	0.9	$9.6 \times 10^{-3}$	0.96	0.014
<b>Reconstructed quaspecies:</b>									
3274	81,590	27	8	20	1.91	0.84	$9.0 \times 10^{-4}$	0.066	0.0016
3319	78,190	26	4	34	1.39	0.75	$2.6 \times 10^{-3}$	0.047	0.0039
3658	90,714	24	8	34	1.91	0.84	$1.5 \times 10^{-3}$	0.112	0.0027
3791	80,614	30	9	27	2.04	0.86	$1.0 \times 10^{-3}$	0.105	0.0019
4070	67,603	28	4	15	1.39	0.75	$1.9 \times 10^{-2}$	0.458	0.0382
4312	76,646	28	4	10	1.39	0.75	$1.4 \times 10^{-2}$	0.335	0.0279
N005	69,552	26	24	184	3.18	0.96	$2.5 \times 10^{-3}$	1.958	0.0035
N011	30,455	25	9	54	2.04	0.86	$2.0 \times 10^{-3}$	0.201	0.0036
N199	46,738	22	31	248	3.43	0.97	$2.5 \times 10^{-3}$	3.942	0.0042
PO04	3,572	27	4	9	1.39	0.75	$7.0 \times 10^{-4}$	0.017	0.0014
<b>Geneious assembly:</b>									
3269	32,196	24	7	48	0.72	0.85	$7.5 \times 10^{-3}$	0.06	0.0088
3358	11,499	28	13	83	1.56	0.76	$3.1 \times 10^{-2}$	1.29	0.0404
3768	11,564	27	10	82	0.69	0.5	$1.8 \times 10^{-2}$	0.09	0.0475
LA05	6,573	23	10	65	1.09	0.66	$3.3 \times 10^{-2}$	0.39	0.056
N060	4,729	23	28	109	2.09	0.83	$2.6 \times 10^{-2}$	5.99	0.0323

Coverage (C) and quaspecies metrics, as determined for the reconstructed quaspecies and for the Geneious assemblies for samples with incomplete coverage. The anticipated coverage (C) per base, based on read files, was calculated with the Lander-Waterman equation while the average Phred-score (Q) was determined in Geneious. Quaspecies metrics we computed using quasitools and are reported at three levels: (1) incidence metrics at the entity level, (2) abundance metrics at the molecular level, and (3) functional metrics of incidence. For incidence at the entity level two metrics are reported, the number of reconstructed or detected haplotypes (H) and the total number of intra-host SNPs. Abundance metrics are given by the standard Shannon entropy (HS) and Gini-Simpson statistics, while functional metrics are given as the mutation frequency (Mfe), functional attribute diversity (FAD), and sample nucleotide diversity ( $\pi$ ).

**Table 3.** Summary statistics for variables used in statistical analysis.

<b>Quantitative variables:</b>						
<b>Variable (x)</b>	<b>Total</b>	<b>Min (xmin)</b>	<b>Max (xmax)</b>	<b>Mean</b>	<b>STD (<math>\sigma</math>)</b>	<b>SEM (<math>\sigma_M</math>)</b>
SNPs (Inter-host):						
<i>GenBank</i>	1187	29	108	79.13	21.07	5.44
<i>Geneious</i>	1053	14	115	70.2	34.1	8.81
<i>QuRe</i>	755	24	132	75.5	35.55	11.24
<i>QuasiRecomb</i>	702	21	125	70.2	35.91	11.35
<i>k-GEM</i>	771	21	127	77.1	36.31	11.48
SNPs (Intra-host):						
<i>All samples</i>	1022	9	248	68.13	67.88	17.52
Quasispecies (Haplotypes):						
<i>QuRe</i>	28	1	8	2.8	2.44	0.77
<i>QuasiRecomb</i>	54	1	23	5.4	7.38	2.33
<i>k-GEM</i>	22	1	5	2.2	1.03	0.32
<i>All samples</i>	173	4	31	11.53	8.85	2.28
<b>Qualitative variables:</b>						
<b>Variable</b>	<b>Total</b>	<b>Positive</b>	<b>Negative</b>	<b>NA</b>		
Serology:						
<i>Human Immunodeficiency Virus (HIV)</i>	60	28 (0.47)	17 (0.28)	15 (0.25)		
<i>HBV Surface Antigen (HBsAg)</i>	60	32 (0.53)	13 (0.22)	15 (0.25)		
<i>HBV Surface Antigen Antibody (Anti-HBs)</i>	60	0 (0.00)	45 (0.75)	15 (0.25)		
<i>HBV Core protein Antibody (Anti-HBc)</i>	60	45 (0.75)	0 (0.00)	15 (0.25)		
<i>HBV e protein Antigen (HBeAg)</i>	60	20 (0.33)	25 (0.42)	15 (0.25)		
<i>HBV e protein Antibody (Anti-HBe)</i>	60	10 (0.17)	35 (0.58)	15 (0.25)		
HBV Genotype:						
<i>Genotype A</i>	52					
<i>Genotype D</i>	4					
<i>Genotype E</i>	4					

Basic attributes of the variables (x) used in subsequent statistical analyses are summarized by ways of descriptive statistics. This included the quantitative variables, such as SNP and quasispecies numbers from each method, as well as qualitative or categorical variables such as HIV status or HBV related serological markers. For quantitative variables the sample size (n), total, minimum (xmin), maximum (xmax), Mean (x), standard deviation ( $\sigma$ ), and standard error of the mean ( $\sigma_M$ ) is reported. For qualitative variables the actual numbers as well as the frequencies are reported.

**Table 4.** Summary of statistical tests for which a significant relationship was found.

<b>Comparison</b>	<b>df</b>	<b><math>\chi^2</math></b>	<b>p-value</b>	<b><math>\eta^2</math></b>	<b>Effect size</b>
Quasispecies (H) vs. HIV status	1	4.61	0.03*	0.06	Moderate
SNPs (inter) vs. HBsAg	1	11.36	<0.01**	0.18	Large
SNPs (inter) vs. Anti-HBe	1	14.84	<0.01**	0.23	Large

Results from statistical analyses comparing the association between specific variables that rendered significant results. Three comparisons rendered significant results, the first between quasispecies number and the latter between the number of SNPs and two of the measured serological attributes (HBsAg and Anti-HBe). In each instance the degrees of freedom (df), Chi-squared ( $\chi^2$ ) test statistic, p-value, and Eta-squared ( $\eta^2$ ) is reported; along with the interpretation for the effect size based on the  $\eta^2$ . (Significance: \*p<0.05 and \*\*p<0.01)

Pooled quasispecies incidence metrics of intra-host variability (Table 2) showed “within specimen” haplotypes (H) ranging from 4 to 28 variants and a median number of SNPs of  $48.00 \pm 17.52$  (SEM). For the abundance metrics, a Shannon entropy (Hs) ranging from 0.69 to 3.43 was detected with a corresponding Gini-Simpson statistic of approximately  $0.90 \pm 0.04$ . The three measures of functional diversity, mutation frequency (Mfe), functional attribute diversity (FAD), and sample nucleotide diversity ( $\pi$ ), had mean values of  $0.0096 \pm 0.0029$ ,  $9.6 \times 10^{-3}$ , and  $0.014 \pm 0.005$  respectively.

**Inter-host diversity.** From the full model population file ( $n = 70$ ) of sequences retrieved from GenBank, an average of  $76 \pm 25$  SNPs were detected between individuals. Using a smaller population model equal in size to the study samples ( $n = 15$ ), a similar average of  $79 \pm 21$  SNPs was detected. Summary statistics of the quantitative variables used in statistic comparisons are provided in Table 3. The average number of SNPs between the individual quasispecies and the reference used for mapping and reconstruction ranged from  $70.20 \pm 11.35$  for QuasiRecomb to  $77.10 \pm 11.48$  for k-GEM. A statistical comparison showed no significant increase in the number of SNPs detected through mapping of read files in Geneious or through quasispecies reconstruction, as compared to the SNPs detected with conventional methods, nor a significant difference between the different reconstruction methods used (Fig. 2,  $\chi^2 = 1.64$ ,  $df = 4$ ,  $p$ -value = 0.80,  $\eta^2 = 0.04$ ). SNPs called for the quasispecies in comparison to the closest BLAST match was generally lower ( $45.53 \pm 12.76$ ) and did not significantly deviate from the median of the intra-host diversity ( $\chi^2 = 0.50$ ,  $df = 1$ ,  $p$ -value = 0.50,  $\eta^2 = 0.01$ ).

**Diversity in relation to serology.** Quasispecies diversity attributes including haplotype number and inter-host SNPs were compared to six serology attributes (summarised in Table 3) including HIV co-infection status and serology markers for HBV from ELISA assays. All samples were positive for HBV core protein antibody (Anti-HBc) and negative for HBV surface protein antibody (Anti-HBs). For the four remaining markers a significant relationship was detected between HIV status and haplotype number while a significant relationship was detected between SNPs and both HBV surface antigen (HBsAg) detection and HBV e-antigen antibody (Anti-HBe) detection. The number of reconstructed haplotypes were moderately increased in HIV negative samples ( $\chi^2 = 4.61$ ,  $df = 1$ ,  $p$ -value < 0.05,  $\eta^2 = 0.06$ ) whereas a large increase in SNP's were observed for either HBV marker (Table 4, HBsAg:  $\chi^2 = 11.36$ ,  $df = 1$ ,  $p$ -value < 0.01,  $\eta^2 = 0.18$ ; Anti-HBe:  $\chi^2 = 14.84$ ,  $df = 1$ ,  $p$ -value < 0.01,  $\eta^2 = 0.23$ ). These observations were evident even at a lower sampling rate due to high effect sizes, as measured by Eta-squared ( $\eta^2$ ) and the adequately powered ( $\beta$ -value > 0.8) non-parametric statistical analyses, tested for in post-hoc analyses.

**Quasispecies phylogenetic relatedness.** The phylogenetic relatedness of the quasispecies generated by different methods in relation to each other, as well as relevant references from the GenBank population model, is illustrated in

Fig. 3. The Quasispecies generated by each method clustered together as a single clade for each individual specimen; apart from 3319.1 QuRe which branched within the clade of 3791. Specimens 4070 and 4312 partitioned with references of Asian origin within an Asian A1 branch. Closely, but separate, to that clade was the quasispecies for N199. Two samples, 3791 and 3319, partitioned with typical African A1 sequences derived from South Africa. The remaining A1 specimens partitioned in clades between the Asian and African genotype A1 branches. N005 partitioned with reference AY233290 from South Africa while specimen 3274 partitioned to form a clade with AY233287, and 3658 branched with AY233284. Specimens N011 and PO04 grouped with the genotype D and E references and were used to root the tree.

## Discussion

Quasispecies were successfully reconstructed for several HBV specimens based on NGS sequence data with full genome coverage. Most samples reconstructed a single majority variant while many other samples had a spectrum of viral quasispecies occurring at frequencies from major variants (>20% abundance) to very low minority variants (<2% abundance). The number of detected SNPs for the quasispecies, in relation to a mapping reference, did not differ significantly from the number observed between full genome isolates of the same genotype as retrieved from online databases. Furthermore, the number of SNPs detected from each algorithm were similar to the differences between closely related isolates, such as subgenotypes, identified through BLAST searches and did not differ significantly between reconstruction methods. Phylogenetic analysis revealed that most of the isolates partition in quasispecies specific clades within HBV genotype A1. The observation that one sample partitioned within the clade of another was likely due to similarities between locally transmitted strains [31], co-infection with two strains of subgenotype A1 [4], or similar evolutionary pressures resulting in convergent evolution [58]. The outliers identified by Mayaphi et al. [39] were more closely related to Asian isolates of this genotype than characteristic South African isolates, similar to previous observations in related isolates [7]. This indicates, for the first time, that quasispecies reconstructions faithfully model known or anticipated variation between full genomes of HBV and reflect the natural evolutionary constraints of this virus.

In comparing diversity measures to HIV status, increased quasispecies diversity, as estimated by haplotype number, was observed for HIV negative specimens. Thus, the hypothesis that HIV co-infection would result in a more diverse quasispecies [43], due to a lack of immune pressure, was incorrect in this case. This would indicate that immune pressures may directly drive or enhance diversity in viral quasispecies. A positive correlation was also found for an increase in SNPs and the seroconversion, or the appearance of specific antigens and antibodies during infection, of two key HBV markers, HBsAg and Anti-HBe. This supports the observation that seroconversion is highly correlated to an increase in viral qua-



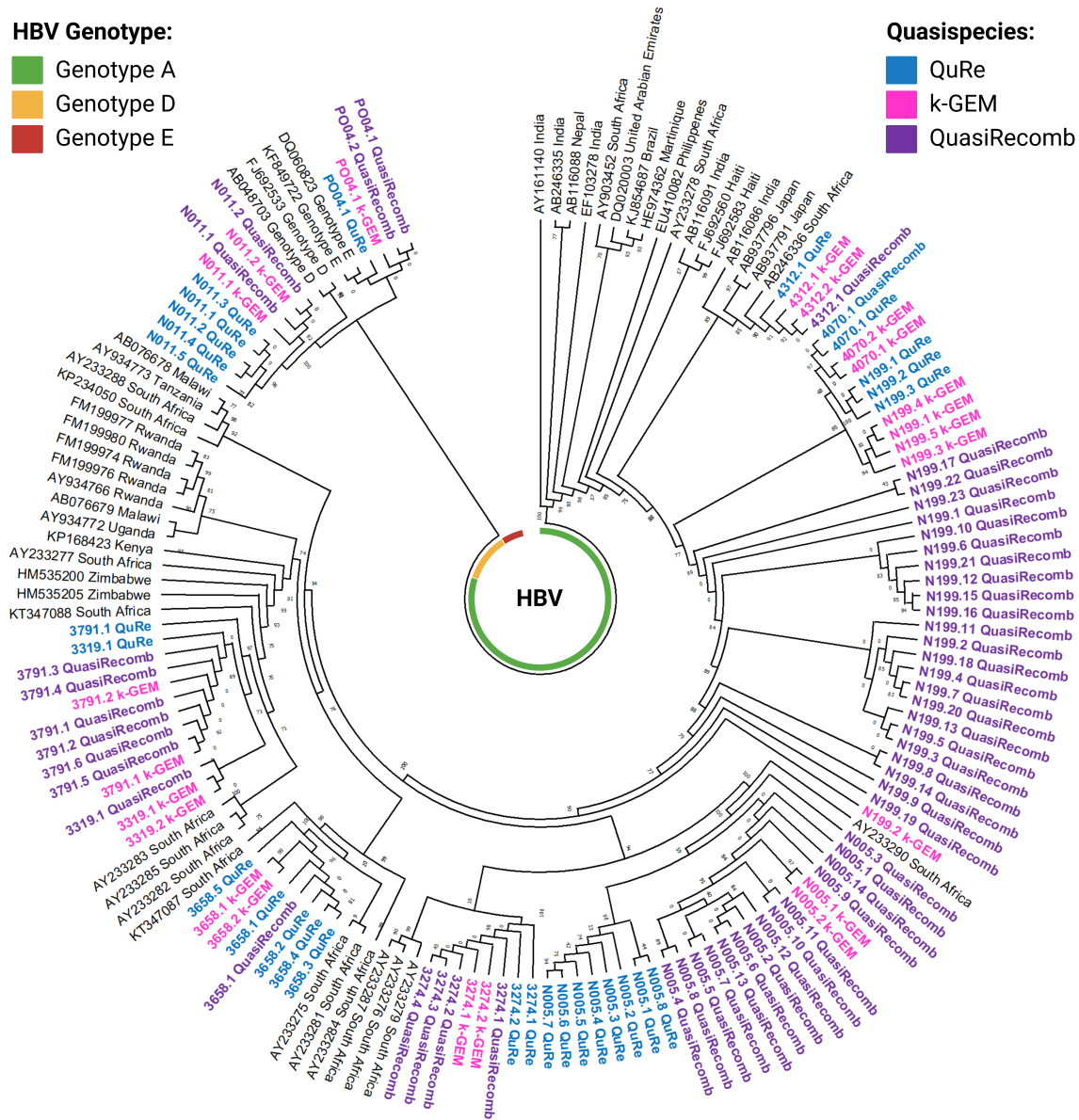
## Maximum Likelihood Phylogenetic tree

**HBV Genotype:**

- Genotype A
- Genotype D
- Genotype E

**Quasispecies:**

- QuRe
- k-GEM
- QuasiRecomb



**Fig. 3.** Phylogenetic tree, based on full genomes, constructed using the maximum likelihood (ML) method showing the relatedness of different reconstructed variants to established genotypes as well as between hosts differences. Quasispecies generated for the same sample by QuRe (blue), QuasiRecomb (purple), or k-GEM (pink) tended to cluster together with high bootstrap support. Most samples partitioned with genotype A references (green) while two samples, N011 and PO04, partitioned with genotypes D (yellow) and E (red) respectively.

species diversity [12, 33, 34]. Conversely, because these markers change predictably over time since initial infection [29], the increase in diversity may also be related to the time period since exposure and the establishment of a chronic infection given the cross-sectional nature of this study. Other measures of quasispecies diversity were consistent with previous findings for HBV [16, 52], including a mutation frequency similar to those observed for RNA viruses but significantly higher than those typically observed for DNA viruses [28].

At present, when studying viral quasispecies, scientists are inundated with an abundance of tools for the reconstruction

of quasispecies with little information available on the accuracy and applicability of a specific method to their study species. This is further confounded by the fact that many tools were initially benchmarked using simulated datasets rather than empirical data. Our comparison of the effect of different methods on the outcome revealed that, although the number of variants and their frequencies may differ between methods, the resolution in terms of SNPs detected remains constant between methods applied. Additionally, the consistency between SNPs detected from read assemblies versus reconstructed quasispecies, indicates that the actual diversity remains conserved even when more rigorous or strin-

gent filtering methods are applied in reconstructing individual genomes. This does, however, indicate that—beyond having individual genomes for phylogenetic analyses—little additional information is to be gained from reconstructing the individual variants as many quasispecies metrics can also be estimated from alignments alone [37].

Next generation sequencing promises to be of great use in the study of viral evolution as applied to both foundational as well as clinical research. Clinical research has benefitted from the more sensitive detection of pathogens, initial drug resistance screening, and therapeutic monitoring [13], as has been done for several viruses including HBV [41], usually through studies of the polymerase gene. For foundational and applied virology, this study illustrates that NGS enables the detection and sequencing of the entire viral quasispecies of HBV. This included minor variant populations for which the full genome was assembled as opposed to only the polymerase gene, similar to studies of HIV [60], influenza A [25], human rhinovirus [56], and herpes simplex virus 1 [54]. This allows the detailed study of differential mutation rates and evolutionary processes across the genome and in different genomic regions through the real-time detection of novel variants as they occur within infected hosts as several host factors and virus dynamics shape the diversity and evolution of this virus [51]. This study also illustrates that NGS and quasispecies reconstruction not only greatly enhances the ability to model viral evolution but also has the potential to possibly predict future isolates. Future studies are needed to further elucidate differences in quasispecies dynamics between acute and chronic infections, as well as quasispecies dynamics in non-human hosts of HBV with different host immune responses.

## Conclusions

1. Different methods of quasispecies reconstruction, using similar approaches in their algorithms, reconstruct the same amount of diversity.
2. Intra-host diversity derived from full quasispecies analyses resembles and does not exceed diversity measures obtained from previous methods such as Sanger, matching known inter-host diversity.
3. Intra-host diversity between closely related quasispecies variants resembles the diversity observed within subgenotypes.
4. Diversity is increased in HIV-negative individuals possibly due to immune pressure.
5. Seroconversion of HBV biomarkers correlates with increases in diversity, possibly due to immune pressure.

**Acknowledgements** Images were created in BioRender.com. Parts of this research was presented at the Faculty Research day (symposium), Faculty of Health Science, University of Pretoria. This work is based on the research supported wholly/in part by the National Research Foundation of South

Africa (Grant Number: 82831) and the Poliomyelitis Research Foundation (Grant number: 12/41 MSc).

**Data availability** Information on individual samples used in the study, as well as raw read files and a subset of reconstructed quasispecies were submitted to NCBI BioSample, Sequence Read Archive (SRA) and the Nucleotide collections with links to the BioProject PRJNA737147 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA737147>). Accession numbers are listed in Table S3. The remainder of the reconstructed quasispecies and other data, have been made available on the Zenodo depository at: <https://doi.org/10.5281/zenodo.7155147>. The protocol for full genome amplification is available on Protocols.io.



## References

- [1] Richa Agarwala et al. “Database resources of the National Center for Biotechnology Information”. In: *Nucleic Acids Research* 46.D1 (Nov. 2017), pp. D8–D13. DOI: [10.1093/nar/gkx1095](https://doi.org/10.1093/nar/gkx1095).
- [2] Irina Astrovskaya et al. “Inferring viral quasispecies spectra from 454 pyrosequencing reads”. In: *BMC Bioinformatics* 12.S6 (Dec. 2011), S1. ISSN: 1471-2105. DOI: [10.1186/1471-2105-12-S6-S1](https://doi.org/10.1186/1471-2105-12-S6-S1).
- [3] Niko Beerenwinkel et al. “Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data”. In: *Frontiers in Microbiology* 3.SEP (2012). ISSN: 1664302X. DOI: [10.3389/fmicb.2012.00329](https://doi.org/10.3389/fmicb.2012.00329).
- [4] Bastian Beggel et al. “Genotyping hepatitis B virus dual infections using population-based sequence data”. In: *Journal of General Virology* 93.PART 9 (Sept. 2012), pp. 1899–1907. ISSN: 00221317. DOI: [10.1099/VIR.0.043042-0/CITE/REFWORKS](https://doi.org/10.1099/VIR.0.043042-0/CITE/REFWORKS).
- [5] Baruch S. Blumberg, Harvey J. Alter, and Sam Visnich. “A New Antigen in Leukemia Sera”. In: *JAMA* 191.7 (Feb. 1965), pp. 541–546. ISSN: 0098-7484. DOI: [10.1001/JAMA.1965.03080070025007](https://doi.org/10.1001/JAMA.1965.03080070025007).
- [6] Baruch S. Blumberg, Alton I. Sutnick, and W. Thomas London. “Australia Antigen and Hepatitis”. In: *JAMA* 207.10 (Mar. 1969), pp. 1895–1896. ISSN: 0098-7484. DOI: [10.1001/JAMA.1969.03150230109015](https://doi.org/10.1001/JAMA.1969.03150230109015).
- [7] Sheila M. Bowyer et al. “A unique segment of the hepatitis B virus group A genotype identified in isolates from South Africa.” In: *Journal of General Virology* 78.7 (July 1997), pp. 1719–1729. DOI: [10.1099/0022-1317-78-7-1719](https://doi.org/10.1099/0022-1317-78-7-1719).

- [8] Sheila M. Bowyer et al. “Expression of the hepatitis B virus genome in chronic hepatitis B carriers and patients with hepatocellular carcinoma.” In: *Proceedings of the National Academy of Sciences* 84.3 (Feb. 1987), pp. 847–850. DOI: [10.1073/pnas.84.3.847](https://doi.org/10.1073/pnas.84.3.847).
- [9] Brian Bushnell, Jonathan Rood, and Esther Singer. “BBMerge – Accurate paired shotgun read merging via overlap”. In: *PLoS ONE* 12.10 (Oct. 2017), e0185056. ISSN: 19326203. DOI: [10.1371/journal.pone.0185056](https://doi.org/10.1371/journal.pone.0185056).
- [10] Maria R Capobianchi, E Giombini, and G Rozera. “Next-generation sequencing technology in clinical virology”. In: *Clinical Microbiology and Infection* 19.1 (2013), pp. 15–22. ISSN: 14690691. DOI: [10.1111/1469-0691.12056](https://doi.org/10.1111/1469-0691.12056).
- [11] Li Chen et al. “Early changes of hepatitis B virus quasispecies during lamivudine treatment and the correlation with antiviral efficacy”. In: *Journal of Hepatology* 50.5 (May 2009), pp. 895–905. ISSN: 0168-8278. DOI: [10.1016/j.jhep.2008.12.018](https://doi.org/10.1016/j.jhep.2008.12.018).
- [12] Yan Cheng et al. “Increased viral quasispecies evolution in HBeAg seroconverter patients treated with oral nucleoside therapy”. In: *Journal of Hepatology* 58.2 (Feb. 2013), pp. 217–224. ISSN: 01688278. DOI: [10.1016/j.jhep.2012.09.017](https://doi.org/10.1016/j.jhep.2012.09.017).
- [13] Stéphane Chevaliez, Christophe Rodriguez, and Jean Michel Pawlotsky. “New virologic tools for management of chronic hepatitis B and C”. In: *Gastroenterology* 142.6 (2012), 1303–1313. e1. ISSN: 15280012. DOI: [10.1053/j.gastro.2012.02.027](https://doi.org/10.1053/j.gastro.2012.02.027).
- [14] Louis-Stéphane Le Clercq. “Molecular characterization of full genome hepatitis B virus sequences from an urban hospital cohort in Pretoria, South Africa”. MA thesis. University of Pretoria, 2014, p. 189. ISBN: 9781321919547. DOI: [10.13140/RG.2.2.33619.71204](https://doi.org/10.13140/RG.2.2.33619.71204). URL: <https://repository.up.ac.za/handle/2263/43142>.
- [15] Louis-Stéphane Le Clercq, Sheila Mary Bowyer, and Simnikiwe Mayaphi. “Full genome PCR amplification of all African Hepatitis B Virus genotypes v1”. In: (June 2021). DOI: [10.17504/protocols.io.bvykn7uw](https://doi.org/10.17504/protocols.io.bvykn7uw).
- [16] Esteban Domingo and Jordi Gomez. “Quasispecies and its impact on viral hepatitis”. In: *Virus Research* 127.2 (Aug. 2007), pp. 131–150. ISSN: 01681702. DOI: [10.1016/j.virusres.2007.02.001](https://doi.org/10.1016/j.virusres.2007.02.001).
- [17] Clément Gilbert and Cédric Feschotte. “Genomic Fossils Calibrate the Long-Term Evolution of Hepadnaviruses”. In: *PLoS Biology* 8.9 (Sept. 2010). Ed. by Michael Emerman, e1000495. ISSN: 1545-7885. DOI: [10.1371/journal.pbio.1000495](https://doi.org/10.1371/journal.pbio.1000495).
- [18] Josep Gregori et al. “Viral quasispecies complexity measures”. In: *Virology* 493 (June 2016), pp. 227–237. ISSN: 0042-6822. DOI: [10.1016/j.virol.2016.03.017](https://doi.org/10.1016/j.virol.2016.03.017).
- [19] Mercedes Guerrero-Murillo and Josep Gregori i Font. *QSutils: Quasispecies Diversity*. 2021. URL: <https://rdr.io/bioc/QSutils/>.
- [20] Stéphane Guindon et al. “Estimating maximum likelihood phylogenies with PhyML”. In: *Methods in Molecular Biology* 537 (2009), pp. 113–137. ISSN: 10643745. DOI: [10.1007/978-1-59745-251-9\\_6/COVER](https://doi.org/10.1007/978-1-59745-251-9_6/COVER).
- [21] Stephan Günther 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”. In: *Journal of Virology* 69.9 (Sept. 1995), pp. 5437–5444. DOI: [10.1128/jvi.69.9.5437-5444.1995](https://doi.org/10.1128/jvi.69.9.5437-5444.1995).
- [22] Stephan Günther et al. “Amplification of Full-Length Hepatitis B Virus Genomes from Samples from Patients with Low Levels of Viremia: Frequency and Functional Consequences of PCR-Introduced Mutations”. In: *Journal of Clinical Microbiology* 36.2 (Feb. 1998), pp. 531–538. DOI: [10.1128/jcm.36.2.531-538.1998](https://doi.org/10.1128/jcm.36.2.531-538.1998).
- [23] Takako Inoue and Yasuhito Tanaka. “Hepatitis B virus and its sexually transmitted infection – an update”. In: *Microbial Cell* 3.9 (Sept. 2016), pp. 419–436. DOI: [10.15698/mic2016.09.527](https://doi.org/10.15698/mic2016.09.527).
- [24] Duleepa Jayasundara et al. “ViQuaS: An improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing”. In: *Bioinformatics* 31.6 (Mar. 2015), pp. 886–896. ISSN: 14602059. DOI: [10.1093/bioinformatics/btu754](https://doi.org/10.1093/bioinformatics/btu754).
- [25] Marie-Louise Kampmann et al. “A simple method for the parallel deep sequencing of full influenza A genomes”. In: *Journal of Virological Methods* 178.1–2 (2011), pp. 243–248. ISSN: 01660934. DOI: [10.1016/j.jviromet.2011.09.001](https://doi.org/10.1016/j.jviromet.2011.09.001).
- [26] P. Karayiannis. “Hepatitis B virus: virology, molecular biology, life cycle and intrahepatic spread”. In: *Hepatology International* 11.6 (Nov. 2017), pp. 500–508. DOI: [10.1007/s12072-017-9829-7](https://doi.org/10.1007/s12072-017-9829-7).
- [27] Kazutaka Katoh and Daron M. Standley. “MAFFT multiple sequence alignment software version 7: Improvements in performance and usability”. In: *Molecular Biology and Evolution* 30.4 (Apr. 2013), pp. 772–780. ISSN: 07374038. DOI: [10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010).
- [28] Anna Kramvis, Michael Kew, and Guido François. “Hepatitis B virus genotypes”. In: *Vaccine* 23.19 (Mar. 2005), pp. 2409–2423. ISSN: 0264410X. DOI: [10.1016/j.vaccine.2004.10.045](https://doi.org/10.1016/j.vaccine.2004.10.045).
- [29] Anna Kramvis et al. “A roadmap for serum biomarkers for hepatitis B virus: current status and future outlook”. In: *Nature Reviews Gastroenterology & Hepatology* 19.11 (July 2022), pp. 727–745. DOI: [10.1038/s41575-022-00649-z](https://doi.org/10.1038/s41575-022-00649-z).
- [30] Sudhir Kumar et al. “MEGA X: Molecular evolutionary genetics analysis across computing platforms”. In:

- Molecular Biology and Evolution* 35.6 (June 2018), pp. 1547–1549. ISSN: 15371719. DOI: [10.1093/molbev/msy096](https://doi.org/10.1093/molbev/msy096).
- [31] Barbara V. Lago et al. “Genetic Diversity of the Hepatitis B Virus Subgenotypes in Brazil”. In: *Viruses* 11.9 (Sept. 2019), p. 860. DOI: [10.3390/v11090860](https://doi.org/10.3390/v11090860).
- [32] Gianvito Lanave et al. “Identification of hepadnavirus in the sera of cats”. In: *Scientific Reports* 9.1 (July 2019). DOI: [10.1038/s41598-019-47175-8](https://doi.org/10.1038/s41598-019-47175-8).
- [33] Seng Gee Lim et al. “Viral Quasi-Species Evolution During Hepatitis Be Antigen Seroconversion”. In: *Gastroenterology* 133.3 (Sept. 2007), pp. 951–958. DOI: [10.1053/j.gastro.2007.06.011](https://doi.org/10.1053/j.gastro.2007.06.011).
- [34] Su-Ru Lin et al. “Whole genome deep sequencing analysis of viral quasispecies diversity and evolution in HBeAg seroconverters”. In: *JHEP Reports* 3.3 (June 2021), p. 100254. DOI: [10.1016/j.jhepr.2021.100254](https://doi.org/10.1016/j.jhepr.2021.100254).
- [35] Julie Lucifora and Fabien Zoulim. “The life cycle of hepatitis B virus and antiviral targets”. In: *Future Virology* 6.5 (May 2011), pp. 599–614. DOI: [10.2217/fv1.11.29](https://doi.org/10.2217/fv1.11.29).
- [36] Santiago Marco-Sola et al. “The GEM mapper: Fast, accurate and versatile alignment by filtration”. In: *Nature Methods* 9.12 (2012), pp. 1185–1188. ISSN: 15487091. DOI: [10.1038/nmeth.2221](https://doi.org/10.1038/nmeth.2221).
- [37] Eric Marinier et al. “quasitools: A Collection of Tools for Viral Quasispecies Analysis”. In: (Aug. 2019). DOI: [10.1101/733238](https://doi.org/10.1101/733238).
- [38] Simnikiwe H Mayaphi et al. “HBV/HIV co-infection: The dynamics of HBV in South African patients with AIDS”. In: *South African Medical Journal* 102.3 (Feb. 2012), p. 157. DOI: [10.7196/samj.4944](https://doi.org/10.7196/samj.4944).
- [39] Simnikiwe H. Mayaphi et al. “Variability of the preC/C region of hepatitis B virus genotype A from a South African cohort predominantly infected with HIV”. In: *Journal of Medical Virology* 85.11 (Nov. 2013), pp. 1883–1892. ISSN: 01466615. DOI: [10.1002/jmv.23695](https://doi.org/10.1002/jmv.23695).
- [40] Masatoshi Nei, Sudhir Kumar, et al. *Molecular evolution and phylogenetics*. Oxford University Press, USA, 2000.
- [41] Norihiro Nishijima et al. “Dynamics of Hepatitis B Virus Quasispecies in Association with Nucleos(t)ide Analogue Treatment Determined by Ultra-Deep Sequencing”. In: *PLoS ONE* 7.4 (Apr. 2012). Ed. by Antonio Bertolotti, e35052. DOI: [10.1371/journal.pone.0035052](https://doi.org/10.1371/journal.pone.0035052).
- [42] Helene Norder et al. “Genetic diversity of hepatitis B virus strains derived worldwide: Genotypes, subgenotypes, and HBsAg subtypes”. In: *Intervirology* 47.6 (Jan. 2004), pp. 289–309. ISSN: 03005526. DOI: [10.1159/000080872](https://doi.org/10.1159/000080872).
- [43] Ananya Pal et al. “Influence of HIV-associated degree of immune suppression on molecular heterogeneity of hepatitis B virus among HIV co-infected patients”. In: *Virology* 436.1 (Feb. 2013), pp. 134–142. ISSN: 00426822. DOI: [10.1016/j.virol.2012.11.003](https://doi.org/10.1016/j.virol.2012.11.003).
- [44] Patricia A. Pesavento et al. “A novel hepadnavirus is associated with chronic hepatitis and hepatocellular carcinoma in cats”. In: *Viruses* 11.10 (2019). ISSN: 19994915. DOI: [10.3390/v11100969](https://doi.org/10.3390/v11100969).
- [45] N. Du Plessis et al. “A nosocomial outbreak of hepatitis B virus infection in a paediatric haematology and oncology unit in South Africa”. In: *International Journal of Infectious Diseases* 21 (Apr. 2014), p. 407. DOI: [10.1016/j.ijid.2014.03.1259](https://doi.org/10.1016/j.ijid.2014.03.1259).
- [46] Mattia C F Prosperi et al. “Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges”. In: *Scientific Reports* 3.1 (Dec. 2013), p. 2837. ISSN: 2045-2322. DOI: [10.1038/srep02837](https://doi.org/10.1038/srep02837).
- [47] Mattia C.F. Prosperi and Marco Salemi. “QuRe: software for viral quasispecies reconstruction from next-generation sequencing data”. In: *Bioinformatics* 28.1 (Jan. 2012), pp. 132–133. ISSN: 1367-4803. DOI: [10.1093/BIOINFORMATICS/BTR627](https://doi.org/10.1093/BIOINFORMATICS/BTR627).
- [48] Mattia CF Prosperi et al. “Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing”. In: *BMC Bioinformatics* 12.1 (Jan. 2011). DOI: [10.1186/1471-2105-12-5](https://doi.org/10.1186/1471-2105-12-5).
- [49] A Rambaut. “FigTree-version 1.4.3, a graphical viewer of phylogenetic trees”. In: (2017). URL: <http://tree.bio.ed.ac.uk/software/figtree>.
- [50] Andrea Rasche et al. “A hepatitis B virus causes chronic infections in equids worldwide”. In: *Proceedings of the National Academy of Sciences* 118.13 (Mar. 2021). DOI: [10.1073/pnas.2013982118](https://doi.org/10.1073/pnas.2013982118).
- [51] Peter A. Revill et al. “The evolution and clinical impact of hepatitis B virus genome diversity”. In: *Nature Reviews Gastroenterology Hepatology* 2020 17:10 17.10 (May 2020), pp. 618–634. ISSN: 1759-5053. DOI: [10.1038/s41575-020-0296-6](https://doi.org/10.1038/s41575-020-0296-6).
- [52] Christophe Rodriguez et al. “Characterization of the dynamics of hepatitis B virus resistance to adefovir by ultra-deep pyrosequencing”. In: *Hepatology* 58.3 (Sept. 2013), pp. 890–901. ISSN: 02709139. DOI: [10.1002/hep.26383](https://doi.org/10.1002/hep.26383).
- [53] C Seeger, F Zoulim, and W Mason. “Hepadnaviruses”. In: ed. by M Knipe and P M Howley. 4th Ed. Vol. 2. Field’s virology. Philadelphia, PA: Lippincott Williams Wilkins, 2007, pp. 2977–3029. ISBN: 9781469830650.
- [54] Moriah L Szpara, Lance Parsons, and L W Enquist. “Sequence Variability in Clinical and Laboratory Isolates of Herpes Simplex Virus 1 Reveals New Mutations”. In: *Journal of Virology* 84.10 (2010), pp. 5303–5313. ISSN: 0022-538X. DOI: [10.1128/jvi.00312-10](https://doi.org/10.1128/jvi.00312-10).
- [55] Lydia S. Y. Tang et al. “Chronic Hepatitis B Infection”. In: *JAMA* 319.17 (May 2018), p. 1802. ISSN: 0098-7484. DOI: [10.1001/jama.2018.3795](https://doi.org/10.1001/jama.2018.3795).

- [56] Caroline Tapparel et al. “Rhinovirus Genome Variation during Chronic Upper and Lower Respiratory Tract Infections”. In: *PLoS ONE* 6.6 (2011), e21163. ISSN: 19326203. DOI: [10.1371/journal.pone.0021163](https://doi.org/10.1371/journal.pone.0021163).
- [57] RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2021. URL: <http://www.rstudio.com/>.
- [58] Hong Thai et al. “Convergence and coevolution of Hepatitis B virus drug resistance”. In: *Nature Communications* 2012 3:1 3.1 (Apr. 2012), pp. 1–8. ISSN: 2041-1723. DOI: [10.1038/ncomms1794](https://doi.org/10.1038/ncomms1794).
- [59] Armin Töpfer et al. “Probabilistic inference of viral quasispecies subject to recombination”. In: *Journal of Computational Biology* 20.2 (Feb. 2013), pp. 113–123. ISSN: 10665277. DOI: [10.1089/cmb.2012.0232](https://doi.org/10.1089/cmb.2012.0232).
- [60] Bram Vrancken et al. “Covering all bases in HIV research: Unveiling a hidden world of viral evolution”. In: *AIDS Reviews* 12.2 (2010), pp. 89–102. ISSN: 11396121.
- [61] Jian Ye, Scott McGinnis, and Thomas L. Madden. “BLAST: improvements for better sequence analysis”. In: *Nucleic Acids Research* 34.suppl\_2 (July 2006), W6–W9. ISSN: 0305-1048. DOI: [10.1093/nar/gkl1164](https://doi.org/10.1093/nar/gkl1164).
- [62] Osvaldo Zagordi et al. “ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data”. In: *BMC Bioinformatics* 12.1 (Apr. 2011). DOI: [10.1186/1471-2105-12-119](https://doi.org/10.1186/1471-2105-12-119).

## Word Counts

This section is *not* included in the word count.

### Notes on Manuscript.

- Abstract: 22 sentences, 251 words.
- Main text: 13 pages, 2 figures, 4 tables, 6500-7000 words.
- Supplementary material: 4 pages, 1 figure, 3 tables.

Table S1. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	lnL	(+I)	(+G)	R	f(A)	f(T)	f(C)	f(G)
<b>GTR+G+I</b>	<b>147</b>	<b>55242.528</b>	<b>53738.094</b>	<b>-26721.941</b>	<b>0.37</b>	<b>0.54</b>	<b>1.58</b>	<b>0.225</b>	<b>0.284</b>	<b>0.271</b>	<b>0.22</b>
GTR+G	146	55283.798	53789.597	-26748.694	n/a	0.25	1.58	0.225	0.284	0.271	0.22
TN93+G+I	144	55474.621	54000.886	-26856.342	0.37	0.53	1.56	0.225	0.284	0.271	0.22
HKY+G+I	143	55475.343	54011.841	-26862.821	0.37	0.53	1.57	0.225	0.284	0.271	0.22
T92+G+I	141	55490.603	54047.566	-26882.686	0.37	0.53	1.56	0.254	0.254	0.246	0.246
K2+G+I	140	55494.11	54061.307	-26890.557	0.37	0.53	1.56	0.25	0.25	0.25	0.25
TN93+G	143	55512.738	54049.236	-26881.518	n/a	0.25	1.55	0.225	0.284	0.271	0.22
HKY+G	142	55515.466	54062.197	-26889	n/a	0.25	1.56	0.225	0.284	0.271	0.22
T92+G	140	55529.825	54097.022	-26908.415	n/a	0.25	1.56	0.254	0.254	0.246	0.246
K2+G	139	55533.392	54110.821	-26916.316	n/a	0.25	1.55	0.25	0.25	0.25	0.25
GTR+I	146	56336.702	54842.501	-27275.146	0.6	n/a	1.54	0.225	0.284	0.271	0.22
HKY+I	142	56579.427	55126.158	-27420.98	0.6	n/a	1.52	0.225	0.284	0.271	0.22
TN93+I	143	56582.777	55119.275	-27416.537	0.6	n/a	1.51	0.225	0.284	0.271	0.22
K2+I	139	56592.994	55170.423	-27446.117	0.6	n/a	1.51	0.25	0.25	0.25	0.25
T92+I	140	56594.708	55161.904	-27440.856	0.6	n/a	1.51	0.254	0.254	0.246	0.246
JC+G+I	139	56675.141	55252.57	-27487.191	0.36	0.53	0.5	0.25	0.25	0.25	0.25
JC+G	138	56711.011	55298.673	-27511.243	n/a	0.25	0.5	0.25	0.25	0.25	0.25
GTR	145	60471.23	58987.262	-29348.528	n/a	n/a	1.45	0.225	0.284	0.271	0.22
TN93	142	60794.056	59340.787	-29528.295	n/a	n/a	1.46	0.225	0.284	0.271	0.22
HKY	141	60799.445	59356.409	-29537.107	n/a	n/a	1.46	0.225	0.284	0.271	0.22
K2	138	60805.288	59392.95	-29558.382	n/a	n/a	1.46	0.25	0.25	0.25	0.25
T92	139	60812.433	59389.863	-29555.837	n/a	n/a	1.46	0.254	0.254	0.246	0.246
JC	137	61923.887	60521.782	-30123.799	n/a	n/a	0.5	0.25	0.25	0.25	0.25
JC+I	138	61936.031	60523.694	-30123.754	0	n/a	0.5	0.25	0.25	0.25	0.25

Models with the lowest BIC scores (Bayesian Information Criterion) are considered to describe the substitution pattern the best. For each model, AICc value (Akaike Information Criterion, corrected), Maximum Likelihood value (lnL), and the number of parameters (including branch lengths) are also presented [40]. Non-uniformity of evolutionary rates among sites may be modeled by using a discrete Gamma distribution (+G) with 5 rate categories and by assuming that a certain fraction of sites are evolutionary invariable (+I). Whenever applicable, estimates of gamma shape parameter and/or the estimated fraction of invariant sites are shown. Assumed or estimated values of transition/transversion bias (R) are shown for each model, as well. They are followed by nucleotide frequencies (f) for each nucleotide pair. For estimating ML values, a tree topology was automatically computed. The analysis involved 70 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 2943 positions in the final dataset. Evolutionary analyses were conducted in MEGA [30].

**Table S2.** Summary of the number of variants/haplotypes reconstructed from each method along with their computed frequency.

Sample ID	N	Frequencies	Vaccine Escape	Immune Escape	Drug Resistance
<i>QuRe:</i>					
3274	2	78.64, 21.35	None	None	None
3319	1	100	None	None	None
3658	5	68.78, 7.89, 7.20, 6.87, 3.66	None	None	None
3791	1	100	None	None	None
4070	1	100	None	None	None
4312	1	100	None	None	None
N005	8	57.39, 19.96, 12.25, 7.16, 0.75, 2.10, 0.23, 0.16	None	None	None
N011	5	94.59, 2.19, 1.08, 1.07, 1.07	None	None	None
N199	3	66.00, 33.00, 1.00	None	None	None
PO04	1	100	None	None	None
<i>QuasiRecomb:</i>					
3274	4	85.00, 11.00, 3.00, 1.00	None	None	None
3319	1	100	None	None	None
3658	1	100	None	None	None
3791	6	63.00, 17.30, 12.40, 4.90, 1.75, 1.53	None	None	None
4070	1	100	None	None	None
4312	1	100	None	None	None
N005	14	57.39, 19.96, 10.26, 7.16, 2.22, 1.96, 0.24, 0.22, 0.18, 0.15, 0.14, 0.05, 0.05, 0.02	None	None	None
N011	2	72.39, 7.61	None	None	None
N199	23	9.40, 6.97, 2.87, 2.54, 2.02, 1.43, 0.98, 0.46, 0.24, 0.19, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17	None	None	None
PO04	2	78.00, 22.00	None	None	None
<i>k-GEM:</i>					
3274	2	79.00, 21.00	None	None	None
3319	2	83.60, 16.40	None	None	None
3658	2	71.45, 28.55	None	None	None
3791	2	86.54, 13.46	None	None	None
4070	2	81.10, 18.90	None	None	None
4312	2	80.73, 19.27	None	None	None
N005	2	75.51, 24.49	None	None	None
N011	5	23.13, 28.39, 18.09, 11.66, 6.20	None	None	None
N199	2	83.41, 16.59	None	None	None
PO04	1	100	None	None	None

The number of variants/haplotypes reconstructed for each specimen are indicated along with their calculated frequency as a measure of abundance for each variant, grouped by the reconstruction algorithm used. The table includes results for testing the quasispecies for the three most pertinent types of mutations, vaccine escape mutations, immune escape mutations, and drug resistance mutations, are also indicated. None of the reconstructed quasispecies harbored significant mutations in terms of measured outcomes.



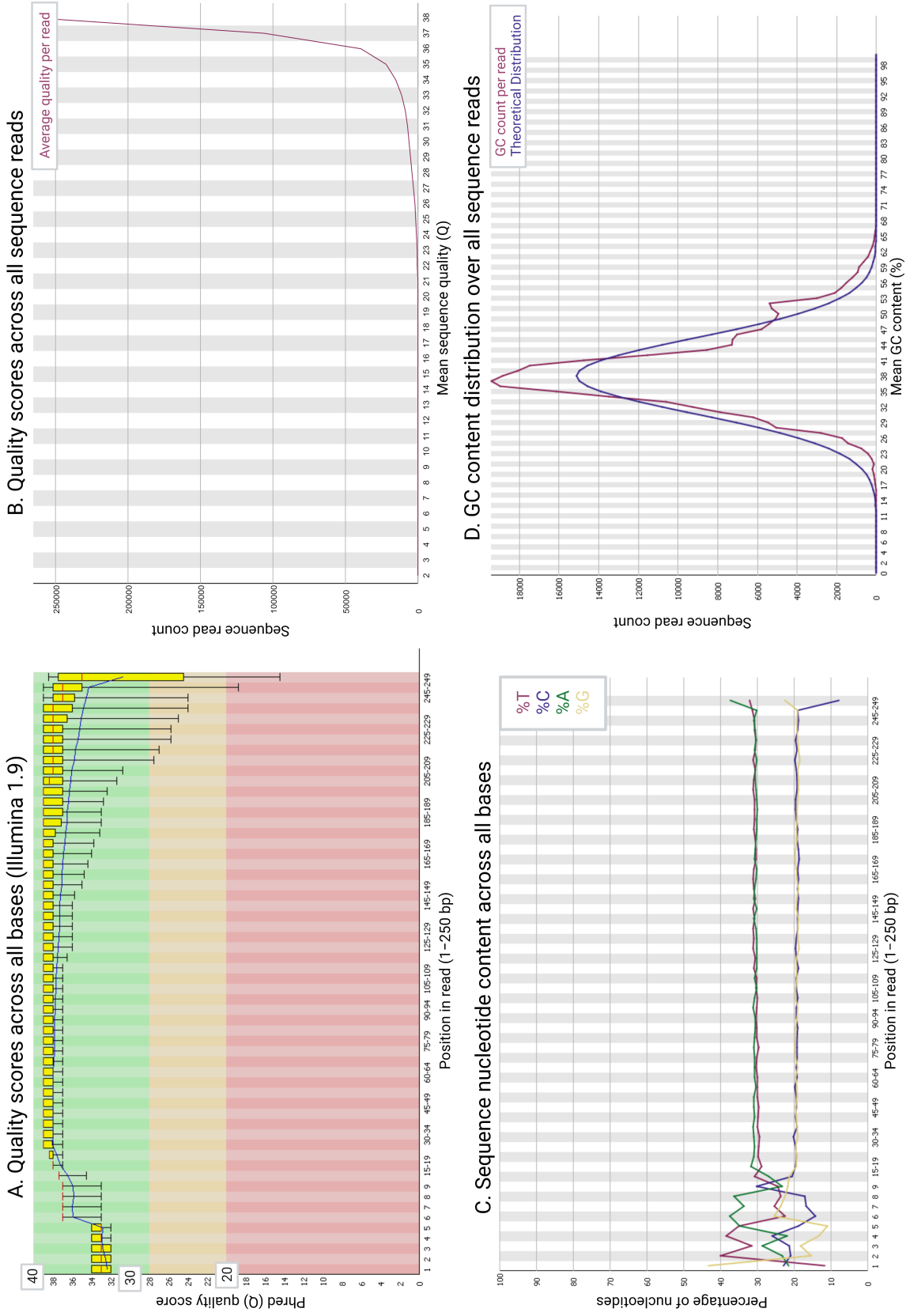
**Table S3.** Summary of NCBI accession numbers for each study sample and associated data.

Study ID	BioSample	Sequence Read Archive (SRA)	GenBank (Nucleotide)*	BioProject
3269	<a href="#">SAMN19694634</a>	<a href="#">SRX11143873</a>	NA	<a href="#">PRJNA737147</a>
3274	<a href="#">SAMN19694639</a>	<a href="#">SRX11143872</a>	<a href="#">KF922429-KF922430</a>	<a href="#">PRJNA737147</a>
3319	<a href="#">SAMN19694636</a>	<a href="#">SRX11143879</a>	<a href="#">KF922422-KF922423</a>	<a href="#">PRJNA737147</a>
3358	<a href="#">SAMN19696888</a>	<a href="#">SRX11143880</a>	NA	<a href="#">PRJNA737147</a>
3658	<a href="#">SAMN19694641</a>	<a href="#">SRX11143881</a>	<a href="#">KF922433-KF922434</a>	<a href="#">PRJNA737147</a>
3768	<a href="#">SAMN19696889</a>	<a href="#">SRX11143882</a>	NA	<a href="#">PRJNA737147</a>
3791	<a href="#">SAMN19694632</a>	<a href="#">SRX11143883</a>	<a href="#">KF922406-KF922409</a>	<a href="#">PRJNA737147</a>
4070	<a href="#">SAMN19694637</a>	<a href="#">SRX11143884</a>	<a href="#">KF922424-KF922425</a>	<a href="#">PRJNA737147</a>
4312	<a href="#">SAMN19694638</a>	<a href="#">SRX11143885</a>	<a href="#">KF922426-KF922428</a>	<a href="#">PRJNA737147</a>
LA05	<a href="#">SAMN19694648</a>	<a href="#">SRX11143886</a>	NA	<a href="#">PRJNA737147</a>
N005	<a href="#">SAMN19694635</a>	<a href="#">SRX11143874</a>	<a href="#">KF922414-KF922421</a>	<a href="#">PRJNA737147</a>
N011	<a href="#">SAMN19694640</a>	<a href="#">SRX11143875</a>	<a href="#">KF922432</a>	<a href="#">PRJNA737147</a>
N060	<a href="#">SAMN19694646</a>	<a href="#">SRX11143876</a>	NA	<a href="#">PRJNA737147</a>
N199	<a href="#">SAMN19694633</a>	<a href="#">SRX11143877</a>	<a href="#">KF922410-KF922413</a>	<a href="#">PRJNA737147</a>
PO04	<a href="#">SAMN19696890</a>	<a href="#">SRX11143878</a>	<a href="#">KF922438-KF922439</a>	<a href="#">PRJNA737147</a>

\* Based on original quasispecies reconstructed with QuRe in Le Clercq [14]

Information on the origin and biographic details for individual samples were submitted to the BioSample database. The associated MiSeq sequencing data was submitted to the SRA database and linked to the relevant samples. Preliminary quasispecies reconstructed with QuRe, reported in Le Clercq [14], were submitted to the core collection of Genbank Nucleotide database. All project data is linked to the registered BioProject.

## Graphical display of key quality metrics assessed during QC of raw reads



**Fig. S1.** Graphical display of key metrics assessed during the quality control (QC) of read files. A) Plot of average quality scores per base from position 1-250 bp for individual reads. B) Plot of quality scores for reads based on average quality score per read. C) Plot of positional nucleotide composition (%) for reads from position 1-250 bp. D) Comparison of actual versus theoretical distribution of GC content (%).