Zhu et al. BMC Genomics (2017) 18:829
DOI 10.1186/s12864-017-4217-1

BMC Genomics

RESEARCH ARTICLE Open Access

Single-virion sequencing of lamivudine-treated HBV populations reveal population evolution dynamics and demographic history

Yuan O. Zhu[1*], Pauline P. K. Aw[1], Paola Florez de Sessions[1], Shuzhen Hong[1], Lee Xian See[2], Lewis Z. Hong[2], Andreas Wilm[1], Chen Hao Li[1], Stephane Hue[3], Seng Gee Lim[4], Niranjan Nagarajan[1], William F. Burkholder[2] and Martin Hibberd[1,3]

Abstract

Background: Viral populations are complex, dynamic, and fast evolving. The evolution of groups of closely related viruses in a competitive environment is termed quasispecies. To fully understand the role that quasispecies play in viral evolution, characterizing the trajectories of viral genotypes in an evolving population is the key. In particular, long-range haplotype information for thousands of individual viruses is critical; yet generating this information is non-trivial. Popular deep sequencing methods generate relatively short reads that do not preserve linkage information, while third generation sequencing methods have higher error rates that make detection of low frequency mutations a bioinformatics challenge. Here we applied BAsE-Seq, an Illumina-based single-virion sequencing technology, to eight samples from four chronic hepatitis B (CHB) patients – once before antiviral treatment and once after viral rebound due to resistance.

Results: With single-virion sequencing, we obtained 248–8796 single-virion sequences per sample, which allowed us to find evidence for both hard and soft selective sweeps. We were able to reconstruct population demographic history that was independently verified by clinically collected data. We further verified four of the samples independently through PacBio SMRT and Illumina Pooled deep sequencing.

Conclusions: Overall, we showed that single-virion sequencing yields insight into viral evolution and population dynamics in an efficient and high throughput manner. We believe that single-virion sequencing is widely applicable to the study of viral evolution in the context of drug resistance and host adaptation, allows differentiation between soft or hard selective sweeps, and may be useful in the reconstruction of intra-host viral population demographic history.

Keywords: Single-virion sequencing, Viral evolution, Adaptation regime, Drug resistance, Chronic hepatitis B, Population demographic history, Bayesian MCMC

* Correspondence: yuanzhu26@gmail.com
[1]Genome Institute of Singapore, Singapore 138672, Singapore
Full list of author information is available at the end of the article

Background

Viral intra-host evolution is a critical obstacle in the treatment of chronic infectious diseases. It is the root cause of viral host immune escape and drug resistance, and consequently a major impediment in disease cure and eradication [1, 2]. The hepatitis B virus (HBV) is a prime example. HBV is a small, circular DNA virus. The HBV polymerase is error prone, with an estimated error every 10^5 to 10^7 bases [3]. When coupled with a large viral load (often between ~10^3 copies/ml to 10^7 copies/ml of serum), this can give rise to substantial viral diversity in active infections [4]. In other words, a sufficiently large viral population can potentially carry, or produce within a short period of time, all possible mutations, thus providing a genetic reservoir for rapid viral response and adaptation [5–10]. Practically, the accumulation of viral mutations is indicative of chronic disease progression and severity [11–13]. Mutations that quickly become predominant in the population are also indicators for how the viruses might be circumventing host response and treatment that enable fresh approaches for drug development research [14].

HBV viral populations, specifically those in chronic infections, can be extremely diverse genetically. Part of the reason is due to high mutation rates leading to the presence of quasispecies [15–20]. Another contributing factor may also stem from genetic repositories in the form of stable covalently closed circular DNA (cccDNA) in infected hepatocytes. Identifying medically important mutations in such populations can become complicated. First, consensus sequence changes occur relatively slowly. For HBV, the mean number of nucleotide substitutions is only estimated at between 1.5×10^{-5} to 7.9×10^{-5} nucleotide substitutions per site per year [19]. The study of consensus sequences alone may not reveal underlying quasispecies dynamics, which may be much more rapid as the population constantly explores possible genotypes [20]. Second, these hidden quasispecies dynamics may be important in understanding the key indicators of viral fitness. Human host immune response, host genetics, treatment regimes, and finally the viral genotype itself likely interact in a complex fashion that exerts multiple, possibly contradictory selective forces on the virus that ultimately culminates in clinical outcome. Identifying the relevant subpopulation of viruses that are reacting to selective pressures of interest, whether it is nucleoside analogues, interferon treatments, or a change in host immune response can reveal important viral indicators for disease progression.

In order to leverage the recent advancements in next generation sequencing (NGS) technology, we explored single-virion sequencing as an option for characterizing quasispecies diversity in active infections. Deep population sequencing is routinely used to identify polymorphisms, including extremely rare alleles [21–26]. However, without linkage information, it remains difficult to describe quasispecies based on allele frequencies alone. A large number of complete genomes from a single viral population must be sequenced to be confident of full quasispecies diversity. Traditionally, such studies require viruses to be individually cloned and sequenced – a rather tedious process requiring a large amount of work and precious source material [19, 27]. However, the complexity and importance of quasispecies has never been clearer [28], and there are two recent next NGS technologies that can be applied to single-virion sequencing in a high-throughput manner, promising up to thousands of viral sequences from every chronic hepatitis B (CHB) patient sample. BAsE-Seq is an Illumina-based method that makes use of random 20mer barcodes to tag every single viral genome with a unique sequence. The barcoded genomes are then amplified as a single amplicon for library construction [29]. Reads from BAsE-Seq libraries can be reassembled into individual viral genomes in silico post sequencing, effectively constructing thousands of viral genomes with full haplotype information. An alternative approach uses single molecule real time sequencing technology (SMRT) on the Pacific Biosciences platform (PacBio) to produce long reads for individual molecules (up to 60 kb). While single pass sequencing error rates are high, the relatively small 3 kb HBV genomes can be read up to dozens of times by the same polymerase, sharply lowering error rates and yielding highly accurate genome sequences, with the additional benefit of not requiring a reference genome [30].

We aimed to apply these single-virion sequencing methods in a manner tailored to characterizing viral population diversity, quasispecies structure, and population evolution. More specifically, we aimed to discover additional information on viral evolutionary dynamics not visible to regular deep sequencing. We picked a relatively well-understood model – that of HBV resistance to the antiviral drug Lamivudine - where the most common resistance alleles are well characterized [31], and obtained two serum samples from each of four CHB patients who were treated with and subsequently developed resistance to Lamivudine. We searched for resistance mutations in each of the patients and tried to reveal additional quasispecies dynamics using single-virion sequencing. We found that single-virion sequencing reveals vital information about viral population heterogeneity and fluctuations in population composition during viral evolution.

Methods

Sample identification and collection in the clinic

Both clonal lab strains and patient samples were used in this study. Plasmids with clonal HBV sequences (referred to as Clone-1 and Clone-2 in the text) were constructed and processed as previously detailed in [28] and sequenced as controls. Patient samples were recruited to

test single-virion sequencing on biological populations as well as to describe any additional information that can be gained through haplotype sequencing. Only patients who gained resistance to Lamivudine with serum samples of suitably high viral load (>10^3 viral copy number/ml) both pre-treatment and post-resistance were considered. As per standard clinical practice, patients who stop responding to anti-viral treatment were tested for resistance mutations through capillary sequencing. For these patients, viral DNA was extracted from 200 μl of serum using the Qiagen Blood mini kit, and the extracted HBV genome was PCR amplified using the Dynazyme DNA polymerase and the primers [Fwd (5′-G[T/C]GTAGACTCGTGGTGGACTTCTCTC-3′).

Rev. (5′-TGACA[T/A/G/C]ACTTTCCAATCA AT-3′)]. The amplified 650 bp fragment was purified by gel electrophoresis and extraction, followed by direct sequencing on an ABI 3730XL DNA Analyzers (SI Table 1). 4 patients with gains of resistance mutations totaling 8 time points had serum samples pulled from the database for sequencing.

Barcode-directed assembly for extra-long sequences (BAsE-Seq)

BAsE-Seq was carried out on all samples. They were: Clone-1, Clone-2, Patient 1 timepoints 1.1 and 1.2, Patient 2 timepoints 2.1 and 2.2, Patient 7 timepoints 7.1 and 7.2, and Patient 11 timepoints 11.1 and 11.2. Library preparation was carried out according to the protocol as described in [28]. Briefly, a total of 10^6 HBV genomes were subjected to a 2-cycle PCR that assigned unique barcodes to each strand of the HBV genome. Two rounds of PCR were carried out to amplify the product, using HBV specific primers (5′-GCTCTTCTTTTTCACCTCTGCCTAA TCA-3′ and 5′-GCTCTTCAAAAAGTTGCATGGTGC TGG-3′), taking care to stay within the exponential amplification regime during each round of PCR to minimize the generation of chimeric PCR products [28]. Specifically, a two-stage PCR protocol was employed such that reactions were stopped in the log-linear phase. The final amplicon spans 3175 bp. Samples were exonuclease- digested to generate a pool of nested deletions fragments, which were end-repaired and circularized. Circular products were fragmented and tagged with the Illumina adaptors followed by 14 cycles of PCR to incorporate primers for sequencing. The resulting 2 × 101 bp reads were trimmed for adaptor sequences and base quality with Trimmomatic [32]. A subset of 10,000 read pairs was first BWA-MEM (v 0.7.10) mapped to 8 known HBV genotypes A-H one at a time [33, 34]. All reads were then BWA-MEM mapped to the genotype reference with the lowest number of mismatches. At this point, mapped reads with identical barcodes, signifying their origin from the same viral molecule, are sorted into individual folders for further processing. For each barcode, aligned reads were duplicate-marked, realigned, and recalibrated with GATK v2.7 [35]. Finally, SNVs were called with LoFreq v2.1.2 [36] and incorporated into the final sequences for each barcode (Additional file 1: Figure S4, S5, S8). Full-length viral sequences that passed all quality filters went on to be part of the population analysis [SI]. Maximum Likelihood phylogenies were built from the top 100 sequences with the highest coverage for easier visual interpretation using FastTree v2.1.8 [37, 38]. PHYLIP Neighbor Joining trees were constructed [39] from the full set of viral sequences obtained [SI]. All trees presented were drawn with iTOL [40, 41]. For further details about the pipeline including all processing and error filters refer to [SI].

Pooled deep sequencing (Illumina)

Pooled deep sequencing was carried out on Clone-1, Clone-2, Patient 1 timepoints 1.1 and 1.2, and Patient 2 timepoints 2.1 and 2.2. Insufficient DNA remained from Patients 7 and patient 11 after BAsE-Seq sequencing, and these four samples had to be excluded from Pooled deep sequencing. For a detailed protocol of sample library preparation, refer to [31]. Briefly, 10^6 HBV viral genomes were PCR amplified using the same primers as

Table 1 Patient sample nomenclature and viral copy number

Sample ID	Patient	Date	Viral Copies/ul	Single-virion Sequences	Nucleotide Diversity Π
1.1	1	15th Nov 94	378,500	1717†, 1635*	0.0012/base
1.2	1	3rd Jul 99	52,270	3331†, 2514*	0.0024/base
2.1	2	22nd May 95	239,750	391†, 1330*	0.0032/base
2.2	2	29th Aug 97	44,687	3747†, 2504*	0.0013/base
7.1	7	19th May 95	69,180	2647†	0.0022/base
7.2	7	11th Apr 00	48,083	789†	0.0016/base
11.1	11	19th Jun 95	208,275	2754†	0.0211/base
11.2	11	3rd Nov 98	466,200	248†	0.0052/base

†-total number of single-virion sequences obtained from a BAsE-Seq library. *-total number of single-virion sequences obtained from a PacBio library. Nucleotide diversity Π, the arithmetic mean between all pairwise differences between viral sequences within each viral population, were calculated from BAsE-Seq single-virion sequences for the entire amplified genomic sequence of 3175 bases (3215 minus the 40 bases that were not amplified)

in BAsE-Seq that cover all but the first 40 bp of the 3215 bp genome. 2–3 μg of PCR product for each viral DNA sample was sheared to achieve a fragment size range between 100 and 300 bp. Library preparation was performed using the Qiagen GeneRead DNA Library I Kit according to manufacturer instructions. After end-repair, A-tailing, and adapter ligation, ligated products in the 200–400 bp range were gel-extracted, and subjected to 14 PCR cycles to incorporate multiplexing indices. The final product was quantified and run on a Illumina HiSeq 2000 instrument. Resulting Illumina 2 × 101 bp reads were trimmed by base quality with Trimmomatic and mapped to the concatenated HBV pan-genome consisting of all 8 major genotypes A-H with BWA-MEM (Additional file 1: Figure S1, Table S2). All concordantly mapped read pairs were duplicate-marked, realigned, and recalibrated with GATK 2.7. SNVs present in the pool were called based on comparison with the best match genotype sequence using LoFreq 2.1.2.

PacBio library construction and analysis

PacBio SMRT sequencing was later carried out on Clone-1, Clone-2, Patient 1 timepoints 1.1 and 1.2, and Patient 2 timepoints 2.1 and 2,2. Insufficient DNA remained from Patients 7 and patient 11 after BAsE-Seq sequencing, and these four samples had to be excluded from PacBio library construction. 10^6 HBV viral genomes were PCR amplified using the same primers as mentioned above under Pooled Deep Sequencing and BAsE-Seq. 2–3 μg of PCR product was used for PacBio library construction following the 2 kb Template Preparation and Sequencing protocol. Library products were quantified on Agilent 2100 Bioanalyzer, and run on a PacBio instrument with V6 chemistry. PacBio raw reads were first processed with the SMRT Portal analysis programs. To focus on full length functional viruses, circular consensus sequences (CCS) from each library were called with a cutoff of at least 10× subreads within a polymerase read and a minimum subread length of 2500 bp using the RS_ReadsOfInsert application (Additional file 1: Figure 2a, b). CCSs were multiple-sequence aligned against all 8 genotypes with MUSCLE [42]. Bases within the CCS reads with quality scores <75 were masked as Ns to filter out false positives (Additional file 1: Figure. 2c), and the resulting (nearly) full-length viral sequences were BWA-SW mapped as extremely long reads to the concatenated HBV pan-genome consisting of all 8 major genotypes A-H (Additional file 1: Figure 3). (Although a reference panel is not necessary for PacBio long reads, it was included in the analysis here for direct comparison between outputs from the platforms.) Segregating sites within the viral populations were called with LoFreq 2.1.2 with primer regions masked. Maximum Likelihood phylogenies were built from the highest quality 100 CCs sequences of the correct length (3175 bp) using FastTree v2.1.8. Neighbor Joining trees were constructed from the full set of viral sequences obtained using PHYLIP [SI]. All trees presented were drawn with iTOL. For a detailed protocol regarding PacBio read processing and error filters refer to [SI].

Reconstruction of demographic history by BEAST

A Bayesian Markov Chain Monte Carlo (MCMC) approach was implemented using BEAST v1.8.4 [43] on all sets of 4 patient samples in order to estimate demographic and evolutionary parameters, using the Bayesian skyline plot as a coalescent prior. Unique single-virion sequences constructed from BAsE-Seq libraries often carried missing information due to uneven coverage. Because an excess of 'N's can overwhelm the true signal, only the top 100 sequences with the highest overall coverage were used for BEAST analysis. A final fragment of 3134 coding bases was used for demographic history reconstruction. Samples prior to Lamivudine treatment were defined as sequences collected on day 0 and samples post drug resistance annotated as sequences collected n days after. We employed the GTR + Γ_4 unlinked codon model of nucleotide substitution and a strict molecular clock. The MCMC chain length was set to 1E9 to 2E9 generations, depending on the patient sample in question, with sampling of every $1E4^{th}$. Convergence of the estimates was considered satisfactory when the effective sample size (ESS), calculated in Tracer v1.6, was >200 for all parameters. The first 10% of the estimates was discarded as burn-in. Where necessary, multiple runs were merged using LogCombiner as part of the BEAST package. Run results were analyzed and skyline plots, showing changes in effective population time over time, generated with Tracer v1.6 [44].

Results

Single-virion sequencing platform error-rates

The three platforms - Pooled Deep sequencing, BAsE-Seq, and PacBio - were tested on two HBV clones with known sequences [28] for pipeline construction and optimization. The data available from each platform is represented in (Fig. 1). Pipelines tailored to each platform were then applied to viral populations from two patients (P1 and P2) to gauge single-virion sequencing performance and throughput on clinical samples [SI]. We began the study with BAsE-Seq, and added PacBio sequencing libraries as the technology became more accessible. We sequenced lab clones with know sequences and estimated the error rates of both methods by summarizing the frequency of base differences in sequenced haplotypes. The base error rates of BAsE-Seq and PacBio libraries were between 0.02–0.3 and 0.2–1.3 per kb of

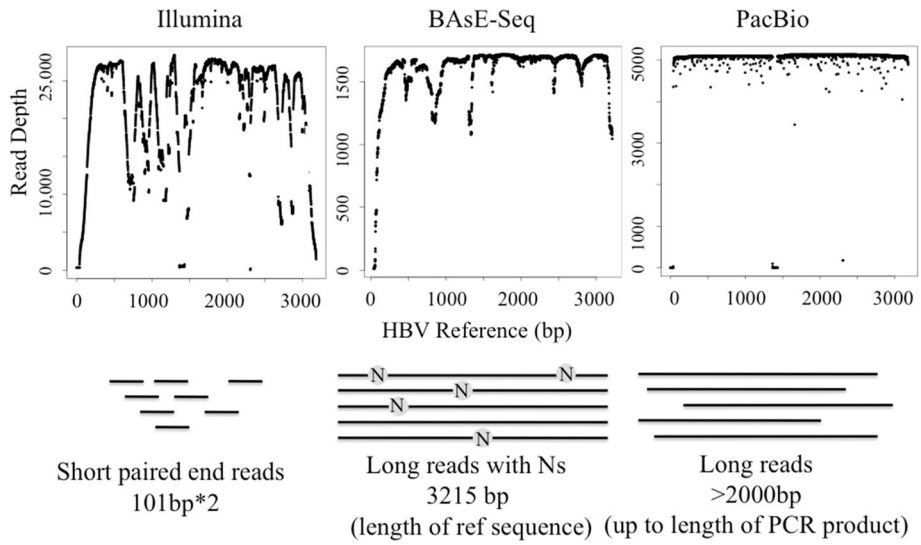

Fig. 1 Graphical representation of sequence reads obtained from BAsE-Seq, PacBio, and Pooled Deep sequencing. X-axis: Genome position. Y-axis: Relative coverage. Each dot represents coverage at a single base position. From left to right: Pooled deep sequencing short reads give uneven genomic coverage. BAsE-Seq reconstitutes single-virion sequences by mapping barcoded short reads to a reference, thus matching exact reference length but may have missing information spread throughout depending on local coverage as shown by dips in the coverage plot. PacBio circular consensus reads (CCSs) can vary in length, but will not necessarily require a reference for construction. A library of 3 kb amplicons will cover the entire genome evenly as shown. Dip in coverage around 1.5 kb is due to a deletion in the sample

single-virion sequence respectively, and the error rates for small indels in BAsE-Seq and PacBio reads were <0.02/kb and 2.9–3.4/kb respectively. PacBio sequencing is known to have higher error rates in homopolymer runs [45]. We tried to further reduce PacBio error rates through careful selection of PCR polymerase [Additional file 1: Figure S2a-b] and CCS quality filters [Additional file 1: Figure S2c], but still faced multiple sequence alignment issues that gave false positive single nucleotide variants (SNVs) [Additional file 1: Figure S3]. As PacBio errors are random and thus low in frequency when consolidated across all reads, we bypassed this issue by mapping CCs reads to a reference sequence [Additional file 1 Figure S6–S7, S9–S13], and only considering these positions in our population analysis. Because PacBio is a reference independent sequencing technology, it is also possible to map

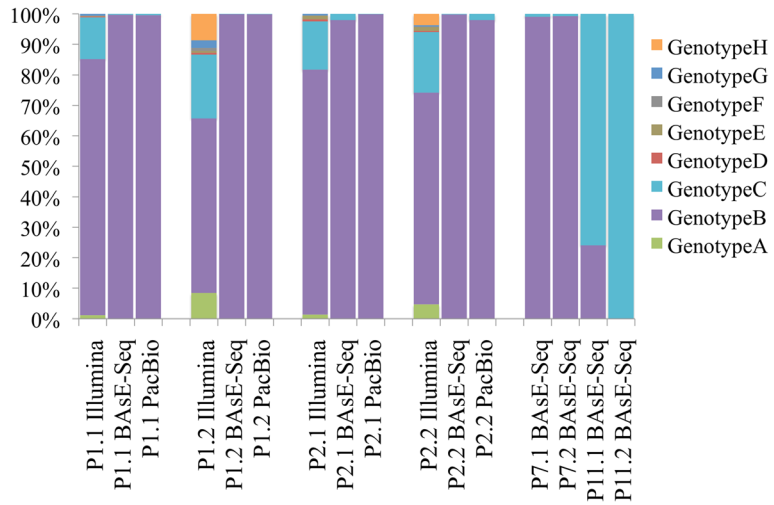

Fig. 2 Genotype composition (Y-axis) of samples sequenced (X-axis) as reported by BWA-SW coverage across 8 reference genotype sequences. Samples are organized from left to right in order of P1.1, P1.2, P2.1, P2.2, P7.1, P7.2, P11.1, and P11.2, where P1.1 and P1.2 are two time points from the same patient P1, before and after drug resistance respectively. Wherever data from multiple technologies are available, order follows Pooled deep sequencing, BAsE-Seq, and lastly PacBio

Table 2 The frequencies of detected resistance alleles in each of the 4 patients after drug resistance. None of these mutations were observed (below detection limit) in the populations prior to development of resistance

#	Mutation	Pooled Deep Seq	BAsE-Seq	PacBio
P1	M204 V	0.525	0.711	0.567
	L180 M	0.560	0.752	0.611
	M204I	0.457	0.251	0.358
P2	M204I	0.978	0.882	0.990
P7	M204 V		0.870	
	L180 M		0.946	
P11	M204I		0.948	
	L180 M		0.679	

reads to a de novo reference sequence generated from the run. Here, we elected for a common reference sequence across all samples for ease of comparison across BAsE-Seq, PacBio, and Pooled deep sequencing data. A more detailed explanation of all work conducted in this comparison exercise is available in Additional file 1.

Classic resistance mutations observed

Serum samples were taken from each patient twice for viral DNA library construction - once before they were treated with Lamivudine (labeled as P1.1, P2.1, P7.1, P11.1) and once after viral loads rebounded to detectable levels (labeled as P1.2, P2.2, P7.2, P11.2) (Table 1).

The four libraries from patients P1 and P2 were sequenced on Pooled Deep sequencing, BAsE-Seq, and PacBio platforms. The remaining four libraries from patients P7 and P11 were sequenced and analyzed only by BAsE-Seq due to limited patient serum availability.

Viral genotype composition in each sample was estimated from the percentage of reads mapping to each genotype reference in the pan-genome panel. Three out of four patients carried Genotype B viruses. The only exception was P11, who carried a mixed Genotype B and Genotype C infection prior to drug treatment, but only Genotype C viruses post Lamivudine resistance (Fig. 2). Illumina short reads tend to mis-map in regions where sequence divergence is ~3% between the references

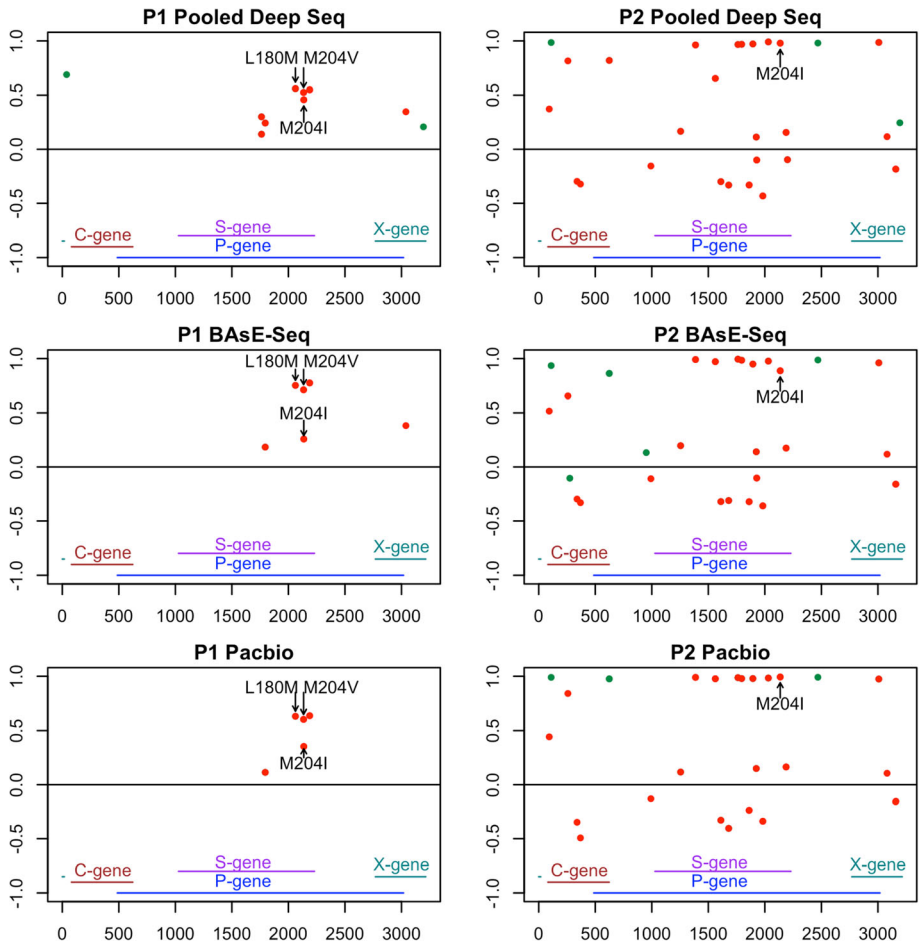

Fig. 3 Change in allele frequencies of variable sites in P1 and P2 after drug treatment (detection limit >0.01). Left panels shows the delta change in allele frequencies of segregating sites (Y-axis) across the genome (X-axis) in patient 1 on all 3 sequencing platforms. Right panels show the corresponding data for patient 2. All non-synonymous changes were colored in red, synonymous changes in green

used, an issue absent in BAsE-Seq and PacBio long reads [Additional file 1 Figure S4]. Therefore, genotype identification in Illumina libraries must take into account evenness of coverage across the references, or number of mismatches in mapped reads, in addition to absolute percentage of reads mapped.

Lamivudine resistance is achieved through mutations in the reverse transcriptase (RT) domain of the polymerase gene in HBV [46–48]. Two resistance phenotypes made up of three amino acid changes, M204I and L180 M + M204 V, are the most commonly observed. They confer similarly high resistance and only require one to two nucleotide changes [49]. Both of these resistant genotypes were found, and together explained resistance in all four patients (Table 2). The discrepancy in allele frequencies between the platforms may have been due to sampling error of low viral load samples.

While P2 (M204I) and P7 (M204 V+ L180 M) carried single resistance phenotypes, P1 carried both M204I and M204 V+ L180 M. The genotypes are not mutually exclusive and all three of the point mutations were found in the same patient at significant frequencies (Figs. 3, 4). P11 was nearly fixed for M204I, but also carried L180 M at a high frequency.

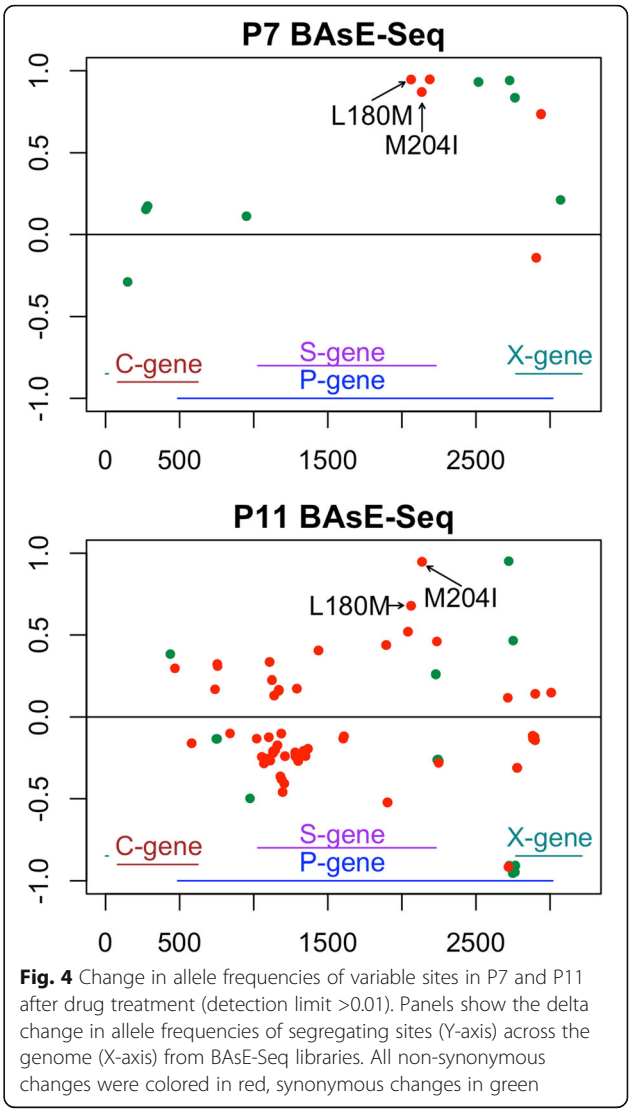

Fig. 4 Change in allele frequencies of variable sites in P7 and P11 after drug treatment (detection limit >0.01). Panels show the delta change in allele frequencies of segregating sites (Y-axis) across the genome (X-axis) from BAsE-Seq libraries. All non-synonymous changes were colored in red, synonymous changes in green

Discussion

Viral Quasispecies reveal both hard and soft selective sweeps

Single-virion haplotypes should yield deeper insights into how these resistance genotypes evolved. Here, we asked if we could identify whether resistance mutations came from a single source and quickly swept to high frequency (hard sweep), or multiple sources that then grew in frequency independently (soft sweep) [50–53]. Note that this question is extremely difficult to address without long-range haplotype information. Hard sweeps are likely to happen with lower mutation rate or extremely strong selection, where adaptive mutations occur one at a time and immediately outcompete other genotypes within the population. Soft sweeps tend to dominate if mutation rates are higher, selection is milder, population is large [54], and multiple lineages carrying advantageous mutations may be present at a time, all increasing in frequency due to the consequent selective advantage [55, 56]. While HBV is a DNA virus that mutates relatively slowly as compared to RNA viruses, it is also true that Lamivudine exerts a strong selective pressure against viral replication. We also asked if we could identify whether resistance alleles were from de novo mutations or from existing low frequency variants. Adaptation from de novo mutation is usually defined as serial fixation of novel alleles, with just one adaptive allele rising to fixation at a time, whereas adaptation from standing variation often also carries with it multiple pre-existing mutations linked to the advantageous allele [57–60]. Which model is more relevant is partly determined by population diversity and the presence of pre-existing drug resistant strains. A clonal viral infection, such as a recent or mono-strain infection seeded by very few drug-naïve virions is less likely to carry pre-existing resistance alleles as compared to a mixed infection or a long-term infection that has had time to diversify within the patient. There is also the possibility that a large, highly mutable viral population could theoretically carry all possible mutations in its quasispecies mutant pool at any point in time. Determining the correct model for HBV population evolution will be important for describing and modeling adaptation.

We made use of nearly full genome haplotype information from BAsE-Seq to characterize viral population quasi-species composition before and after drug treatment (Additional file 2). PacBio trees for P1 and P2 are

available in SI. Phylogenetic trees built from viral haplotypes revealed three different patterns in how these patients gained viral resistance (Additional file 3).

Two patients, P2 and P7, had trees that showed clear mono-clonal gains of resistance, suggestive of hard sweeps (Fig. 5a,b, Additional file 1 Figure S14–S15). Allele frequency changes showed clusters of SNVs that increased in allele frequency together (11 SNVs in P2 spanning the entire 3.2 kb sequenced region [Fig. 3] and 6 SNVs in P7 spanning 2 kb–2.8 kb [Fig. 4]). Haplotype information confirmed that these were linked SNVs on the same haplotype. 9/11 SNVs in the P2 cluster were within the RT domain of the polymerase gene, and 7/11 SNVs were non-synonymous mutations within a 750 bp window. All six SNVs in the P7 cluster were within the RT domain of the polymerase gene, and three were non-

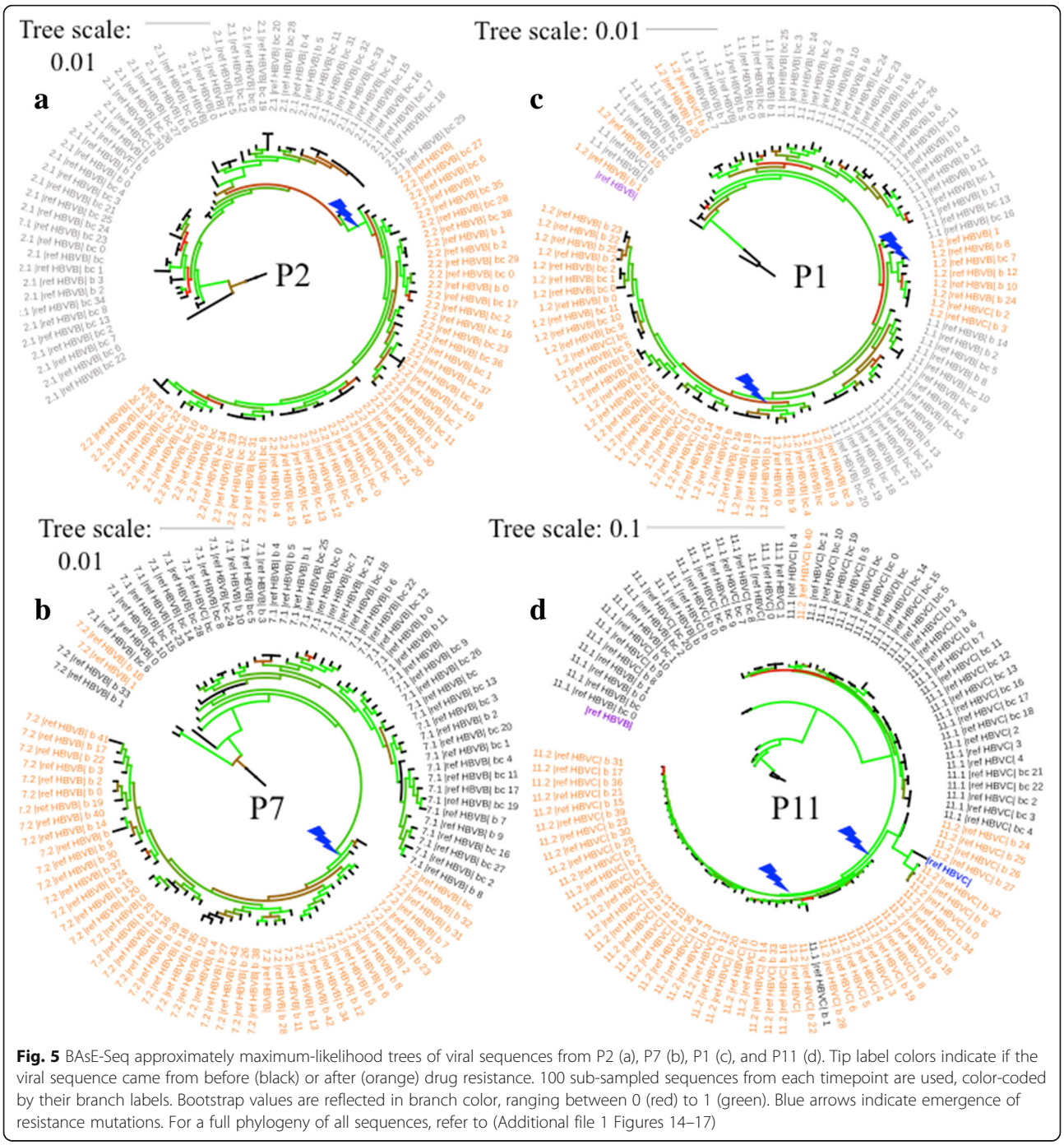

Fig. 5 BAsE-Seq approximately maximum-likelihood trees of viral sequences from P2 (a), P7 (b), P1 (c), and P11 (d). Tip label colors indicate if the viral sequence came from before (black) or after (orange) drug resistance. 100 sub-sampled sequences from each timepoint are used, color-coded by their branch labels. Bootstrap values are reflected in branch color, ranging between 0 (red) to 1 (green). Blue arrows indicate emergence of resistance mutations. For a full phylogeny of all sequences, refer to (Additional file 1 Figures 14–17)

synonymous mutations within a 150 bp window. These two sweeps with numerous SNVs linked to the resistance allele would support a model of evolution from standing variation. However, these exact combinations of SNVs were not found in the treatment naïve timepoints for either patient. The closest haplotypes found pre-treatment shared just 6/11 SNVs for P2 and 2/6 SNVs for P7. Haplotype analysis of linked SNVs in PacBio sequences for P2 showed the same pattern (Fig. 3).

We suggest three possible explanations for this linkage. First, there could be a detection limit for extremely rare haplotypes in the pre-treatment timepoints. We may simply have failed to sequence them. Second, because our samples came from patient blood samples, latent viral reservoirs outside of the blood stream could be contributing to the viral population. Perhaps even reshuffling viral haplotypes through recombination. Again, they would be missed by serum samples. Finally, the resistance mutation could have occurred later during the treatment regime by chance, and happened to rise on the background of a viral sequence that already accumulated multiple nucleotide differences from the population consensus. There was in fact a two to three fold difference in nucleotide diversity across the eight patient samples (Table 1), suggesting a range of quasispecies complexity across patients, which may affect observed evolutionary dynamics.

The two remaining patients, P1 and P11, had trees showing at least two independent instances of gain of resistance, in other words soft sweeps (Fig. 5c,d, SI Additional file 1 Figure S16–S17). P1 was highly clonal with just 6 sites shifting in frequency over time (Fig. 3), and independently gained M204I and L180 M + M204 V on two haplotypes. Again, this was seen in the PacBio library as well (Fig. 3). P11 carried the most diverse population out of all four patients, starting as a mixed population of 26% Genotype B and 74% Genotype C (Fig. 2). The same resistance allele M204I evolved twice on Genotype C sequences but none on Genotype B sequences, resulting in a resistant viral population that was 100% Genotype C. One lineage further gained the L180 M mutation, although it is unclear whether that conferred additional resistance on a M204I background.

Reconstructing demographic history The BEAST analysis for P11 showed effective samples sizes (ESS) of >500, indicating convergence. Sequences came from day −1233 (sample before Lamivudine treatment) and day 0 (sample taken after viral resistance and consequently viral load rebound) [SI Table 3]. Reconstructed demographic history showed an initial exponential growth phase post infection, followed by a plateau. A sharp dip in effective viral population size (N_e) then occurred sometime between −1350 to −750 days prior to day 0.

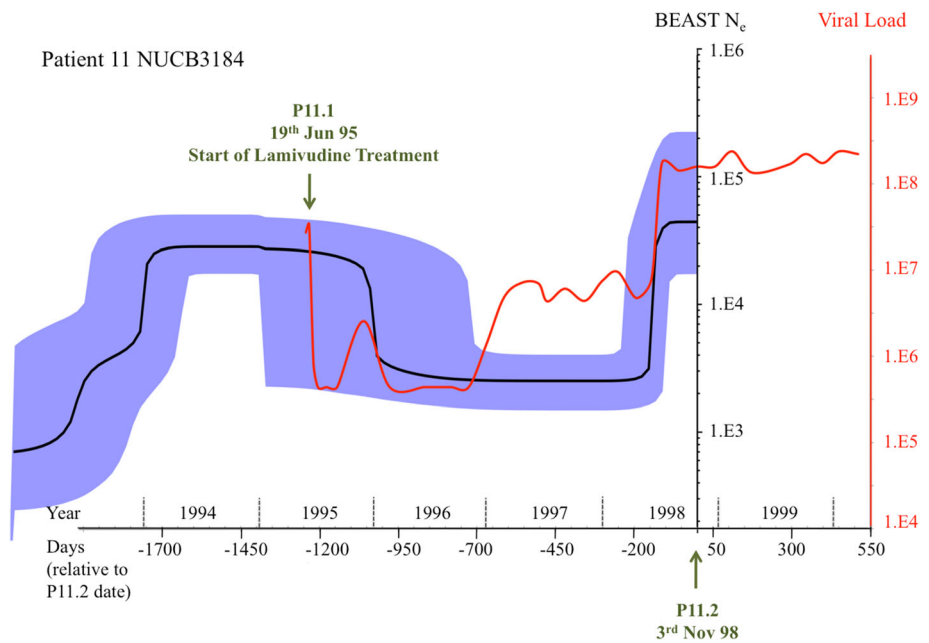

Fig. 6 Overlay of BEAST reconstructed demographic history and clinical records of patient viral load (serum) for patient 11. X-axis – Timeline represented in days, going forward in time from left to right. Green arrows point out the two timepoints that were used for single-virion sequencing. Y-axis labeled "Viral load" – corresponding to red line tracing all available records of patient viral load (log10 scale) over time. Y-axis labeled "BEAST N_e" - Effective population size over time as simulated by BEAST (log10 scale) shown by black line (plotted values are local medians, with 95% highest probability density (HPD) confidence interval colored in blue)

From actual patient viral load information, the population crash post drug treatment occurred right after the sample was take on day −1233. There was also a sharp increase in median N_e about 150 days prior to day 0 in the skyline plot, which matched clinical data almost perfectly (Fig. 6). Although the two major changes in population size were well identified, a smaller increase in the viral population size around day −700 was not, possibly indicating some limitations when reconstructing smaller scale changes in population demography, or that these viruses did not contribute to the effective population size. Due to the smaller nucleotide diversity present in the other patient samples (Table 1), runs for P1, P2 and P7 did not coalesce at 1E9 replications (Additional file 1 Figure S18–S20).

Conclusions

Haplotype information is vital for revealing hidden population dynamics invisible in standard deep sequencing data. While single-virion sequencing remains technically challenging, we employed two complementary single-virion sequencing platforms to reveal – and cross-validate - such information. We can tell, from nucleotide diversity calculations, the heterogeneity of a population. We can estimate, using up to thousands of single-virion haplotypes, the relative proportions of genotypes and quasispecies present in an infection. We can determine if resistance evolved from a single source, or multiple times independently. Using samples taken at different timepoints, we can begin to explore whether evolution occurs from standing variation or de novo mutations, and how that is linked to quasispecies complexity. While lamivudine resistance is a relatively simple adaptive process with very specific alleles conferring fitness gains, this work shows the potential of applying single-virion sequencing to complex events such as viral response to immune enhancement or viral dynamics during an active HBV flare. It may also be valuable in the study of difficult topics such as cccDNA stability, viral recombination, and viral reservoirs. Single-virion sequencing is therefore a powerful tool for understanding the role of viruses across disease stages of clinical importance.

Additional files

Additional file 1: A supplementary materials file provides additional technical details and figures deemed unnecessary for the main text, including BEAST results for all patient samples. (PDF 2754 kb)

Additional file 2: Supplementary_genomes.fasta. High quality single virion sequences Single-virion sequences that were reconstructed with BAsE-Seq and used in FastTree phylogeny analysis for all 4 patients. (FASTA 1289 kb)

Additional file 3: Supplementary_FastTrees.txt. Newick format phylogenetic trees. FastTree output in Newick format for all 4 patients (TXT 16 kb)

Abbreviations

BAsE-Seq: Barcode-directed Assembly for Extra-long Sequences; cccDNA: covalently close circular DNA; CCs: circular consensus sequence; CHB: chronic hepatitis B; ESS: effective sample size; HBV: hepatitis B virus; MCMC: Markov chain Monte Carlo; NGS: next generation sequencing; SMRT: Single Molecule, Real-Time; SNV: single nucleotide variant.

Acknowledgements

We thank Wendy Soon, Gary Chen and the entire Next Generation Sequencing Platform team at the Genome Institute of Singapore for their support and expertise. We thank Siddarth Signh and the Pacific Biosciences support team for their expertise and invaluable advice and feedback.

Consent for publication

(Not applicable).

Funding

The work presented here was funded by the Agency of Science, Technology and Research (A*STAR) and the grant JCO CDA 13302FG059. YOZ, PA, PFS, SGL, and MH were also funded by the grant titled Eradication of HBV TCR Program: NMRC/TCR/014-NUHS/2015. NN and WFB were also funded by the grant JCO DP 1334 k00082.

Availability of data and materials

Raw sequencing reads from all libraries were deposited and publicly available on the NCBI SRA database under BioProject PRJNA407696. Scripts used in the BAsE-Seq pipeline are available on Github at https://github.com/OliviaZhu26/Single_virion_seq. 100 high quality single virion sequences from each patient, and the phylogenetic trees built from them using FastTree, are included in Additional file 1.

Authors' contributions

LZH, NN, WFB and MH conceived and designed the experiments. SGL procured the samples. PPKA, PFS, SZH, and LXS constructed the libraries. YOZ, AW, CHL, SH, and MH conducted the analyses. YOZ, PFS, LZH, AW, SH, NN, WFB, and MH drafted and edited the manuscript. All authors have read and approved the final version of this manuscript.

Authors' information

SZH is currently affiliated with Chugai Pharmabody Research Pte Ltd., Singapore 138,623. LZH is currently affiliated with Translational Biomarkers, Merck Research Laboratories, MSD, Singapore 138,665.

Ethics approval and consent to participate

All patients provided written informed consent according to the Declaration of Helsinki, and all study protocols were approved by the Domain Specific Review Board (DSRB), National Healthcare Group (NHG), Singapore.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Genome Institute of Singapore, Singapore 138672, Singapore. [2]Institute of Molecular and Cell Biology, Singapore 138673, Singapore. [3]London School of Hygiene and Tropical Medicine, London, UK. [4]National University Hospital, Singapore 119074, Singapore.

Received: 8 May 2017 Accepted: 16 October 2017
Published online: 27 October 2017

References

1. Hughes D, Andersson DI. Evolutionary consequences of drug resistance: shared principles across diverse targets and organisms. Nat Rev Genet. 2015 Aug;16(8):459–71.
2. Presloid JB, Novella IS. RNA viruses and RNAi: Quasispecies implications for viral escape. Viruses. 2015 Jun 19;7(6):3226–40. https://doi.org/10.3390/v7062768.

3. Caligiuri P, Cerruti R, Icardi G, Bruzzone B. Overview of hepatitis B virus mutations and their implications in the management of infection. World J Gastroenterol. 2016 Jan 7;22(1):145–54. https://doi.org/10.3748/wjg.v22.i1.145.
4. Osiowy C, Giles E, Tanaka Y, Mizokami M, Minuk GY. Molecular evolution of hepatitis B virus over 25 years. J Virol. 2006 Nov;80(21):10307–14. https://doi.org/10.1128/JVI.00996-06.
5. Yim HJ, Hussain M, Liu Y, Wong SN, Fung SK, Lok AS. Evolution of multi-drμg resistant hepatitis B virus during sequential therapy. Hepatology. 2006; 44:703–12.
6. Margeridon-Thermet S, Shulman N, Ahmed A, Shahriar R, Liu T, Wang C, Holmes SP, Babrzadeh F, Gharizadeh B, Hanczaruk B, Simen BB, Egholm M, Shafer RW. Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naive patients. J Infect Dis. 2009;199:1275–85.
7. Chen L, Zhang Q, Yu DM, Wan MB, Zhang XX. Early changes of hepatitis B virus quasispecies during lamivudine treatment and the correlation with antiviral efficacy. J Hepatol. 2009;50:895–905.
8. Tang YZ, Liu L, Pan MM, Wang YM, Deng GH. Evolutionary pattern of full hepatitis B virus genome during sequential nucleos(t)ide analog therapy. Antivir Res. 2011;90:116–25.
9. Cheng Y, Guindon S, Rodrigo A, Lim SG. Increased viral quasispecies evolution in HBeAg seroconverter patients treated with oral nucleoside therapy. J Hepatol. 2013 B Feb;58(2):217–24. doi: https://doi.org/10.1016/j.jhep.2012.09.017. Epub 2012 Sep 27.
10. Cheng Y, Guindon S, Rodrigo A, Wee LY, Inoue M, Thompson AJ, Locarnini S, Lim SG. Cumulative viral evolutionary changes in chronic hepatitis B virus infection precedes hepatitis B e antigen seroconversion. Gut. 2013 A Sep;62(9):1347–55. doi: https://doi.org/10.1136/gutjnl-2012-302408. Epub 2012 Dec 15.
11. Sterneck M, Gunther S, Gerlach J, Naoumov NV, Santantonio T, Fischer L, Rogiers X, et al. Hepatitis B virus sequence changes evolving in liver transplant recipients with fulminant hepatitis. J Hepatol. 1997;26:754–64.
12. Gunther S, Fischer L, Pult I, Sterneck M. Will H. Naturally occurring variants of hepatitis B virus. Adv Virus Res. 1999;52:25–137.
13. Hannoun C, Horal P, Lindh M. Long-term mutation rates in the hepatitis B virus genome. J Gen Virol. 2000 Jan;81(Pt 1):75–83.
14. Chisari F, Ferrari C, Hepatitis B. Virus immunopathogenesis. Ann Rev Immunol. 1995;13:29–60.
15. Eigen M. Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften. 1971;58(10):465–523.
16. Eigen M. On the nature of virus quasispecies. Trends Microbiol. 1996; 4(6):216–8.
17. Eigen M, McCaskill J, Schuster P. Molecular quasi-species. J Phys Chem. 1988;92:6881–91.
18. Simmonds P, Midgley S. Recombination in the genesis and evolution of hepatitis B virus genotypes. J Virol. 2005;79(24):15467–76.
19. Domingo E, Sheldon J, Perales C. Viral Quasispecies Evolution. Microbiol Mol Biol Rev. June 2012;76(2):159–216.
20. Widasari DI, Yano Y, Heriyanto DS, Utsumi T, Yamani LN, Rinonce HT, Wasityastuti W, Lusida MI. Soetjipto, Okada R, Murakami Y, Tanahashi T, Azuma T, Hayashi Y. A deep-sequencing method detects drμg-resistant mutations in the hepatitis B virus in Indonesians. Intervirology. 2014;57(6):384–92.
21. Yamani LN, Yano Y, Utsumi T, Juniastuti, Wandono H, Widjanarko D, Triantanoe A, Wasityastuti W, Liang Y, Okada R, Tanahashi T, Murakami Y, Azuma T, Soetjipto, Lusida MI, Hayashi Y. Ultradeep sequencing for detection of Quasispecies variants in the major hydrophilic region of hepatitis B virus in Indonesian patients. J Clin Microbiol 2015 Oct;53(10): 3165–3175. doi: https://doi.org/10.1128/JCM.00602-15. Epub 2015 Jul 22.
22. Linlin Yan, Henghui Zhang, Hui Ma, Di Liu, Wei Li, Yulin Kang, Ruifeng Yang, Jianghua Wang, Gaixia He, Xingwang Xie, Hao Wang, Lai Wei, Zuhong Lu, Qixiang Shao & Hongsong Chen. Deep sequencing of hepatitis B virus basal core promoter and precore mutants in HBeAg-positive chronic hepatitis B patients. Scientific reports 5, Article number: 17950 (2015) doi:https://doi.org/10.1038/srep17950.
23. Lil F, Zhang D, Li Y, Jiang D, Luo S, Du N, Chen W, Deng L, Zeng C. Whole genome characterization of hepatitis B virus quasispecies with massively parallel pyrosequencing. Clin Microbiol Infect. March 2015;21(3):280–7.
24. Ode H, Matsuda M, Matsuoka K, Hachiya A, Hattori J, Kito Y, Yokomaku Y, Iwatani Y, Sμgiura W. eCollection 2015. Quasispecies analyses of the HIV-1 near-full-length genome with Illumina MiSeq. Front Microbiol. 2015 Nov 12; 6:1258. https://doi.org/10.3389/fmicb.2015.01258.
25. Chen S, Wu J, Gu E, Shen Y, Wang F, Zhang W. Evaluation of the dynamic pattern of viral evolution in patients with virological breakthroμgh during treatments with nucleoside/nucleotie analogs by ultra-deep pyrosequencing. Mol Med Rep. 2016;13(1):651–60.
26. Lim SG, Cheng Y, Guindon S, Seet BL, Lee LY, Hu P, Wasser S, Peter FJ, Tan T, Goode M, Rodrigo AG. Viral quasi-species evolution during hepatitis be antigen seroconversion. Gastroenterology. 2007;133(3):951–8.
27. Andino R, Domingo E. Viral quasispecies. Virology. 2015 May;479-480:46–51.
28. Hong LZ, Hong S, Wong HT, Aw PPK, Cheng Y, Wilm a, de sessions PF, Lim SG, Nagarajan N, Hibberd ML, quake SR, Burkholder WF. BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. Genome Biology2014 15:517 doi: https://doi.org/10.1186/s13059-014-0517-9.
29. Dilernia DA, Chien JT, Monaco DC, Brown MPS, Ende Z, Deymier MJ, Yue L, Paxinos EE, Allen S, Tirado-Ramos A, Hunter E, Multiplexed highly-accurate DNA. Sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. Nucl. Acids res. 16. November 2015;43(20):e129. https://doi.org/10.1093/nar/gkv630.
30. Pallier C, Castéra L, Soulier A, Hézode C, Nordmann P, Dhumeaux D, Pawlotsky JM. Dynamics of hepatitis B virus resistance to lamivudine. J Virol. 2006 Jan;80(2):643–53. https://doi.org/10.1128/JVI.80.2.643-653.
31. Aw PP, de Sessions PF, Wilm A, Hoang LT, Nagarajan N, Sessions OM, Hibberd ML. Next-generation whole genome sequencing of dengue virus. Methods Mol Biol 2014;1138:175–195.
32. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
33. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25(14):1754–60.
34. Li H, Durbin R. Fast and accurate long-read alignment with burrows wheeler transform. Bioinformatics. 2010;26(5):589–95.
35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010 Sep; 20(9):1297–1303. doi: https://doi.org/10.1101/gr.107524.110. Epub 2010 Jul 19.
36. Wilm A. Aw PP, Bertrand D, yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagaranjan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell population heterogeneity from high-throμghput sequencing datasets. Nucleic Acids Res. 2012 Dec;40(22):11189–201.
37. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26:1641–50. https://doi.org/10.1093/molbev/msp077.
38. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490. https://doi.org/10.1371/journal.pone.0009490.
39. Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*
40. Letunic and Bork. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127–8.
41. Letunic and Bork. Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 2011; https://doi.org/10.1093/nar/gkr201.
42. Edgar RCMUSCLE. Multiple sequence alignment with high accuracy and high throμghput. Nucleic Acids Res. 2004;32(5):1792–7.
43. Drummond AJ, Rambaut A. BEAST: bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214.
44. Rambaut A, Suchard MA, Xie D, Drummond AJ. Trace. 2014:v1.6. Available from http://beast.community/tracer.
45. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14:R51. https://doi.org/10.1186/gb-2013-14-5-r51
46. He X, Wang F, Huang B, Chen P, Zhong L. Detection and analysis of resitance mutations of hepatitis B virus. Int J Clin Exp Med. 2015;8(6):9630–9.
47. Libbrecht E, Doutreloigne J, Van De Velde H, Yuen MF, Lai CL, Shapiro F, Sablon E. Evolution of primary and compensatory lamivudine resistance mutations in chronic hepatitis B virus-infected patients during long-term lamivudine treatment, assessed by a line probe assay. J Clin Microbiol. December 2007;45(12):3935–41.
48. Ko SY, Oh HB, Park CW, Lee HC, Lee JE. Analysis of hepatitis B virus drug-resistant mutant haplotypes by ultra-deep pyrosequencing. Clin. Microbiol and Infect vol 18. 10:E404–11.

49. Fischer KP, Gutfreund KS, Tyrrell DL. Lamivudine resistance in hepatits B: mechanisms and clinical implications. Drug Resist Updat. 2001;4:118–27.
50. Hermisson J, Pennings PS. Soft sweeps. Genetics. 2005 April;169(4):2335–52.
51. Barton N. Understanding adaptation in large populations. PLoS Genet. 2010; 6:e1000987.
52. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013 Nov. 28(11):https://doi.org/10.1016/j,tree.2013.08.003.
53. Pennings PS, Hermisson J. Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol. 2006;23:1076–84.
54. Orr HA, Betancourt AJ. Haldane's sieve and adaptation from the standing genetic variation. Genetics. 2001;157:875–84.
55. Innan H, Kim Y. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A. 2004;101:10667–72.
56. Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991;129:555–62.
57. Hudson RR, et al. Evidence for positive selection in the superoxide dismutase (sod) region of Drosophila Melanogaster. Genetics. 1994;136: 1329–40.
58. Barton NH. The effect of hitch-hiking on neutral genealogies. Genet Res. 1998;72:123–33.
59. Fay JC, Hitchhiking WCI. Under positive Darwinian selection. Genetics. 2000; 155:1405–13.
60. Durrett R, Schweinsberg J. Approximating selective sweeps. Theor Popul Biol. 2004;66:129–38.