A hepatitis B virus (HBV) sequence variation graph improves sequence alignment and sample-specific consensus sequence construction for genetic analysis of HBV

Dylan Duchen,[1] Steven Clipman,[2] Candelaria Vergara,[1] Chloe L. Thio,[2] David L. Thomas,[2] Priya Duggal,[1] Genevieve L. Wojcik[1]

[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA

[2] Division of Infectious Diseases, Johns Hopkins School of Medicine, Baltimore, MD, 21205, USA

ABSTRACT

Hepatitis B virus (HBV) remains a global public health concern, with over 250 million individuals living with chronic HBV infection (CHB) and no curative therapy currently available. Viral diversity is associated with CHB pathogenesis and immunological control of infection. Improved methods to characterize the viral genome at both the population and intra-host level could aid drug development efforts. Conventionally, HBV sequencing data are aligned to a linear reference genome and only sequences capable of aligning to the reference are captured for analysis. Reference selection has additional consequences, including sample-specific 'consensus' sequence construction. It remains unclear how to select a reference from available sequences and whether a single reference is sufficient for genetic analyses. Using simulated short-read sequencing data generated from full-length publicly available HBV genome sequences and HBV sequencing data from a longitudinally sampled individual with CHB, we investigate alternative graph-based alignment approaches. We demonstrate that using a phylogenetically representative 'genome graph' for alignment, rather than linear reference sequences, avoids issues of reference ambiguity, improves alignment, and facilitates the construction of sample-specific consensus sequences genetically similar to an individual's infection. Graph-based methods can therefore improve efforts to characterize the genetics of viral pathogens, including HBV, and may have broad implications in host pathogen research.

INTRODUCTION

Approximately one-third of the world's population has been exposed to the hepatitis B virus (HBV), a major cause of hepatocellular carcinoma and end-stage liver disease.[1] With over 250 million individuals suffering from chronic HBV infection (CHB), novel drugs are needed as no effective curative therapies currently exist.[2] While spontaneous recovery occurs, the biological mechanisms underlying the immunological control of HBV remain unclear. In addition to age, clinical and environmental factors, and host genetic variation,[3,4] viral genetic diversity contributes to the pathogenesis and the severity of CHB.[5–10]

HBV has a small (3.2kb) partially double-stranded circular genome with four overlapping gene-encoding regions and a higher mutation rate than most DNA viruses.[11] With 10 known genetically and geographically distinct HBV genotypes and >30 subgenotypes, CHB is also caused by recombinants or mixtures of genotypes.[12–19] Additionally, like other chronic viral infections, intra-host CHB diversity involves multiple viral strains which evolve, mutate, and change in frequency over time, termed a viral quasispecies.[20–24] This intra-host diversity has also been shown to influence disease progression,[25–27] treatment outcome,[28,29] and confound molecular epidemiology or surveillance efforts.[30]

Characterizing this extensive genetic variation is therefore important for advancing our understanding of the natural history of disease and potential treatment targets. Sequencing-based analyses of HBV and other microbial pathogens usually involve an initial alignment of sequencing data to a representative reference genome. Choosing the right reference is critical, as only data sufficiently similar to the reference can be aligned and retained within subsequent analyses.[31,32] However, it can be difficult to select the most appropriate reference sequence when analyzing clinical CHB samples of unknown HBV genotype or subgenotype. One option is to assess all potential HBV reference sequences to identify the most appropriate reference sequence but this is both computationally expensive and can still fail within the context of recombinant or mixed infections.

Nevertheless, the use of unrepresentative reference sequences can interfere with characterizing pathogen diversity, resulting in false or missing mutations and biased phylogenetic relationships.[33–35] Reference selection can also affect the ability to accurately derive sample-specific ‘consensus’ sequences, which provide an approximation of the genome sequence causing an infection. This issue of reference ambiguity is especially problematic for CHB, as a set of phylogenetically representative HBV reference sequences has only recently been proposed.[36] Furthermore, given the extreme diversity of HBV, the use of a single reference sequence, even of the correct HBV genotype, may be insufficient.[32,33]

Aligning to a phylogenetically representative ‘genome graph’ constructed from many different HBV genome sequences rather than a single HBV genome sequence could potentially avoid these reference ambiguity issues. Genome graphs are comprised of ‘nodes’ which reflect stretches of genetic sequence connected by ‘edges’ which determine the path a genome sequence traverses across a subset of nodes within the graph.[31,34,37,38] Graph-based structures containing genetic variation from multiple genomes have been shown to improve sequence alignment and variant calling for highly variable regions of the human genome and microbial organisms like *Escherichia coli*.[31,34,39] A graph-based reference containing a representative sampling of the genetic variation observed across all known HBV genotypes/subgenotypes might also improve sequence alignment and variant calling, as well as enable the generation of accurate sample-specific consensus sequences for HBV-related genetic analyses. Consensus sequences reflect the most commonly observed nucleotide at each site across the genome, inferred from aligning sequencing data against a specific reference sequence. The construction of consensus sequences is a typical objective of viral-focused genetic analyses,[40,41] including HBV.[42,43] However, whether a graph-based approach can improve viral sequence alignment and sample-specific consensus sequence construction has, to our knowledge, yet to be demonstrated.

In this study, we leverage 2,837 publicly available full-length HBV genomes, simulated high-throughput sequencing data from these HBV sequences, and real-world longitudinally collected sequencing data from an individual with CHB to identify the optimal alignment method as determined by the proportion of successfully aligned HBV sequencing data. Additionally, by comparing alignment derived sample-specific consensus sequences, the accuracy of graph vs. linear reference-based alignment methods will be evaluated.

MATERIALS AND METHODS

Source of genetic sequence data

Full-length HBV genome sequences: A set of non-redundant full-length HBV genomes (N=2,837) was obtained from the publicly available resource provided by McNaughton et al.[36] Briefly, 7,108 full-length HBV genomes were obtained from the HBVdb database (https://hbvdb.lyon.inserm.fr/HBVdb/) and recombinant or highly similar full-length HBV genome sequences were removed. A set of 44 sequences representative of all phylogenetically identified genotypes, subgenotypes, and genetically distinct clades was then identified for use as reference sequences in downstream analyses.

High-throughput CHB sequencing data: HBV-targeted sequencing data from an individual included in a longitudinal cohort study of treatment naïve individuals with CHB was obtained via the NCBI Sequence Read Archive (BioProject ID: 479693).[44] Sample-level clinical and demographic data were obtained through communication with study authors. The high-throughput CHB sequencing data included in this study reflects five longitudinally sampled visits from a single individual (identifier ‘C4’). C4 is a male who was 36 years old at the time the first sample was obtained in 1991. The final sample was obtained in 1996. Individual C4 remained chronically infected and HBeAg positive for the entirety of study follow-up. Sequencing was performed using an Illumina HiSeq 2500, as previously described.[44] Cutadapt was used to trim adapters, poor-quality bases, and reads <36bp long.[45] FastQC was used to ensure the post-QC data passed Illumina sequencing-related QC checks.[46]

Patients with HBV infection, including individual C4, were recruited with fully informed written consent from the Division of Gastroenterology and Hepatology at the National University Health System, Singapore.[44]

Simulated high-throughput sequencing data: Realistic high throughput HBV sequencing data were simulated using InSilicoSeq, which enables the generation of error-prone Illumina-like sequencing data with pre-specified abundance/coverages.[47] Two datasets of paired-end sequences/reads were generated using an Illumina HiSeq error model, the first set (N=50,000 reads/genome) was simulated from each of the recommended HBV reference sequences (N=44). The second set (N=500,000 reads total) was simulated from a randomly selected HBV genome sequence from each of the 9 HBV genotypes (excluding genotype J, as only a single isolate remains available) and 50 additional randomly selected HBV genomes not included within the HBV reference graph (N=59).

Sequence-to-graph alignment

HBV reference graph construction and alignment: A sequence variation graph, termed the HBV reference graph, was created using the full set of phylogenetically representative reference sequences (N=44) (**Supplementary materials**).[36] The HBV reference graph was created using the pangenome graph builder (PGGB) pipeline (https://github.com/pangenome/pggb), which performs pairwise whole-genome alignment using wfmash and graph induction using the seqwish software.[48,49] PGGB can then sort and order the graph via partial order alignment using smoothxg (https://github.com/pangenome/smoothxg).

The variation graph toolkit (VG, v1.39) was used to perform all graph-related format conversions, indexing, sequence-to-graph alignment, and collating of mapping/alignment statistics, as described in the VG documentation.[31,50] Fast short-read alignment via the VG giraffe mapper was accomplished by creating a haplotype-aware graph index where each reference genome was indexed as a unique haplotype.[51] Highly accurate but more computationally intensive graph-based mapping was performed using the VG map mapper.

Establishing internal validity of the HBV reference graph: All reads simulated from the graph-embedded HBV genomes were concatenated and then randomly subsampled to 20,000X coverage seven times. Coverage-based subsampling was performed using rasusa.[52] These subsampled HBV sequencing datasets were aligned to the HBV reference graph using the haplotype-aware VG giraffe mapper. For the graph to be internally valid, we required >99% of the reads simulated from HBV genomes embedded within the graph to successfully align.

To assess whether each path within the graph was utilized during sequence alignment, and to test whether aligned sequences had the highest alignment scores to graph-embedded HBV genomes which were more genetically similar to the aligned sequences, each full-length HBV sequence (N=2,837) was aligned to the graph using VG map. A 'correct' alignment was observed if the reference path with the highest alignment score was of the same HBV genotype as the query sequence. Path-specific alignment scores were also derived for alignments made using the set of simulated high throughput sequencing data from HBV genomes not used in graph construction (N=59). Briefly, by identifying the graph nodes for each path with successful alignments, the genome path with the most alignments was able to be identified (**Figure S1**). Path-specific alignment scores were derived using the sum of weights estimated for each node involved in a successful alignment. Weights reflect the path depth of each node (i.e. the number of genome sequences containing/traversing through the node), with nodes traversed by a single HBV genome weighted heavily and nodes traversed by all genomes weighted least (**Figure S2**). A ‘correct’ alignment was observed if the path with the highest weighted alignment score was of the same HBV genotype as the genome sequence used to simulate the HBV sequencing data.

Alignment of HBV sequencing datasets – graph vs. linear references

Simulated HBV sequencing data: To determine whether a graph-based reference improves sequence alignment compared to linear reference-based approaches for HBV sequencing datasets, we aligned the combined simulated high-throughput sequencing data (generated from 59 HBV genomes not included within the graph) to the graph using VG giraffe and to each linear reference sequence (N=44) using BWA-MEM. The proportion of successfully aligned reads was obtained using ‘VG stats’ and ‘SAMtools flagstat’,[53] respectively. While comparisons of the computational time and resources required for variation graph and linear-reference-based aligners have been performed previously,[50] ‘/usr/bin/time’ estimates for the alignments using BWA-MEM, VG giraffe, VG giraffe in fast-mode, and VG map can be found in **Table S1**. To approximate a more realistic scenario, in which the observed genetic diversity spans a subset of HBV genotypes known to circulate within a geographic region rather than all currently known HBV genotypes/subgenotypes, linear reference and graph-based alignment comparisons were performed using simulated sequencing data from randomly selected HBV genotype B (N=6) and C (N=12) sequences, the primary genotypes endemic in East and Southeast Asia.[54,55]

We estimated the depth of coverage across the HBV genome for the alignment of all simulated high-throughput sequencing data to each linear reference using ‘SAMtools depth’. Genotype-specific depth estimates were obtained by estimating the mean alignment depth across alignments made using references of the same genotype via a sliding-window approach (50bp wide) in R. Local minima in depth were estimated using the ggmisc package in R. To approximate site-specific depth of coverage across the HBV genome from the graph-based alignment, the start site of each successfully aligned read was used to infer coverage by estimating a rolling sum of the median number of reads within a sliding window the length of the simulated reads (125bp).

To facilitate the comparison of whether regions of poor coverage corresponded to loci of increased pairwise diversity, the local nucleotide sequence diversity across the set of reference sequences (N=44) was estimated using a sliding window approach (150bp wide) in R using the pegas package.[56]

Alignment of real CHB sequencing data: To determine the approximate sequencing depth for each CHB sample (N=5), raw sequencing data were aligned to each linear reference sequence (N=44) using BWA-MEM.[57] Alignment quality was assessed using Qualimap (v2.2.1).[58] The proportion of successfully aligned reads were estimated using ‘SAMtools flagstat’. For each sample, the linear reference with the highest proportion of successfully aligned reads was an HBV subgenotype B2 sequence (GenBank ID: GU815637). For alignments to this reference, mean depth of coverage ranged from 82,930X-334,157X. To reduce computational time and resources required for our analyses, QC-passed reads were down-sampled to obtain an average coverage of

20,000X. To test whether subsampling altered the proportion of successfully aligned reads, subsampled reads were also aligned to each linear reference sequence and the proportion of successful alignments was compared to the alignments involving all QC-passed sequencing data using a binomial generalized linear mixed model (GLMM) with random intercepts in R. The GLMM treated each alignment as a binomial outcome (successful alignment vs. not successful alignment), with the total number of reference-specific alignments used as weights. Whether alignments of these subsampled reads to an extended linear reference, obtained by concatenating the first 120bp of each reference to the end of each sequence, altered alignment statistics were also assessed using the same GLMM performed in R. To identify whether reads which failed to align to sub-optimal linear references (non-HBV subgenotype B2) were uniformly distributed across the genome, unaligned reads from each linear reference-based alignment were re-aligned to the best performing linear reference. The genome-wide distribution of these ‘rescued’ reads was visually assessed in R.

Graph-based alignment of subsampled CHB sequencing data was performed using the VG map mapper. For samples with higher alignment proportions to the graph than any linear reference, unmapped reads from the best performing linear reference for each sample were re-mapped to the graph and the genome-wide distribution of the reads ‘rescued’ via graph alignment was visualized using R. To identify and visualize the loci where HBV sequence was rescued via graph alignment, the rescued reads were queried via BLAST against a compacted de Bruijn Graph comprised of the reference sequences and *de novo* (reference-free) assembled HBV haplotypes from each sample created using Bifrost and visualized with Bandage.[59,60] We also performed BLAST in Bandage using these successfully re-mapped reads against the HBV reference graph only to confirm that rescued reads mapped to regions of increased graph complexity.

Derived consensus sequences – graph vs. linear reference sequences

Simulated HBV sequencing data: Consensus sequences were obtained from linear reference-based alignments of simulated non-graph derived sequencing data using iVar.[61] iVar was developed to analyze amplicon-based viral sequencing data and leverages SAMtools to call variants and derive a consensus from the most common nucleotide across each position in an alignment file. We used a minimum base-level quality score of 20 and depth threshold of 10 while accounting for ambiguous nucleotides. For graph-based alignments, we used a wholly graph-based variant calling approach leveraging the alignments across all paths using VG (https://github.com/vgteam/vg), followed by consensus generation via bcftools.[62]

Longitudinal CHB sample consensus sequences: Prior to performing alignment and variant calling for the real CHB samples, QC-passed paired-end reads were merged using PEAR and filtered to retain the highest quality reads >150bp long for analysis via bbmap.[63,64] Reads were aligned to each linear reference or the HBV reference graph, followed by iVar-based consensus sequence identification. We also performed variant calling using the LoFreq software, a variant calling tool able to identify even low-frequency variants from high-coverage data across diverse genetic sequencing datasets,[65] for each linear reference-based alignment followed by consensus generation using a majority allele rule for each site (i.e. alleles with frequency >50% were integrated into the consensus sequence) via bcftools. For LoFreq-derived consensuses, we used the same depth threshold (10) used in iVar and estimated insertion/deletion qualities which were used in addition to LoFreq's method of combining base-level, mapping, and alignment quality information to determine variant quality and identify the majority nucleotide at each position, accounting for insertions/deletions. Graph-based alignment was performed using VG giraffe. VG-based variant calling using the graph alignments and consensus generation were obtained via bcftools. For these CHB samples, we also derived consensus sequences after re-aligning the successful graph-aligned reads to a single path within the graph via VG, termed surjection, followed by iVar consensus construction. Graph-aligned reads were surjected into the path corresponding to the best-performing linear reference.

Consensus sequence comparisons

Consensus sequence comparisons from simulated HBV sequencing data: Comparisons between each simulation-based consensus sequence and the full set of HBV genomes from which reads were simulated were performed using Mash,[66] which estimates a genetic distance metric, the Mash distance, based on the estimated mutation rate between two sets of sequences and the jaccard index (the fraction of k-mers shared between the comparison sequences). The Mash distance also approximates average nucleotide identity (ANI) estimates, with ANI equivalent to one minus the distance estimate, while also having the benefit of facilitating comparisons between sequences/sequencing datasets of variable lengths/sizes.[66] Given the short length of the HBV genome (3.2kb), a k-mer sequence length of 7 was used for Mash distance estimations.[67,68] The consensus sequence with the lowest estimated genetic distance with the set of full-length HBV genome sequences can be inferred to be the most accurate or genetically representative consensus sequence.

Identifying accurate consensus sequences from real CHB sequencing data: To facilitate comparisons between CHB-derived consensus sequences and to identify the most genetically similar consensus to the HBV quasispecies of each sample, we estimated the Mash distance between each consensus and the subsampled HBV sequencing data which aligned to the best performing reference (linear or graph-based) for each sample. We also performed *de novo* HBV strain-level assembly using SAVAGE and VG-Flow to identify the viral haplotypes comprising each CHB infection.[69,70] For each sample, the best-performing linear reference was added to the SAVAGE output for VG-Flow to improve strain-level contiguity and assembly. The set of sample-specific viral haplotypes with frequencies >1% were included in all pairwise genetic distance comparisons. The consensus sequence with the lowest estimated genetic distance with the HBV-specific high throughput sequencing data can be inferred to be the most accurate and genetically representative consensus sequence for each sample.

RESULTS

Simulations to assess HBV sequence-to-graph alignment and coverage

Short Illumina-like reads were simulated from 59 genetically diverse HBV genomes encompassing 9 HBV genotypes and aligned to a non-overlapping set of 44 phylogenetically representative HBV genome sequences reflecting the known breadth of HBV diversity. Despite all reads being of HBV origin (N=500,002 reads), only 84.3% to 96.6% of sequences successfully aligned to these 44 linear references (**Figure 1**). In contrast, >99.9% of this diverse simulated HBV sequencing data successfully aligned to an HBV reference graph constructed using the same set of 44 phylogenetically representative HBV genome sequences. To ensure that loci from across all HBV genomes used to create the graph are adequately represented by the reference graph, seven randomly subsampled sets of simulated high-throughput HBV sequencing data (N=507,938 reads) generated from these 44 genomes were aligned to the HBV reference graph, with 100% of reads always aligning to the graph.

To reflect a more realistic scenario utilizing simulated HBV sequencing data, we limited the simulated data used in our alignment comparisons to those generated from HBV sequences of genotypes B or C. While >99.9% of simulated reads from genotypes B and C successfully aligned to the reference graph, a high proportion of these reads also successfully aligned to linear reference sequences of genotype B (97.9%-98.4%) and C (97.6%-98.8%) (**Figure S3**).

The simulated HBV sequences that failed to align to the linear references (3.4%-15.7%) were not uniformly distributed across the genome, with loci observed to have precipitous drops in coverage corresponding to loci of increased genetic diversity (**Figure 1**). Within these loci, drops in coverage were highly heterogeneous across reference sequences of different HBV genotypes, with the lowest proportion of successfully aligned HBV sequencing data and the most precipitous drops in coverage in regions of increased nucleotide diversity occurring for HBV genotypes H and G reference sequences. As >99.9% of sequencing data successfully aligned to the HBV reference graph, no major coverage differences were observed.

To determine whether graph-aligned HBV sequences aligned best to the path/reference sequence embedded within the graph of the same HBV genotype as the query sequence, all full-length HBV genome sequences (N=2,837) and each set of simulated short-read HBV sequencing data generated from the HBV genomes used in the simulations (N=59) were aligned to the HBV reference graph. For each alignment, query sequences always resulted in paths of the same HBV genotype having the highest alignment score (**Table 1**), demonstrating the importance of representing each phylogenetically distinct HBV genotype within the reference graph. These results also demonstrate that the path-specificity of sequence-to-graph alignment can enable HBV genotype prediction using either the alignment score directly or a metric based on the path-depth of nodes with successful alignments for genome-length and high-throughput HBV sequences, respectively.

Alignment of real CHB sequencing data to an HBV reference graph

Unlike our simulated HBV sequencing datasets, real CHB-derived HBV sequencing data can reflect extensive genetic variation due to both host and pathogen-derived evolutionary pressures in addition to any sample processing or sequencing-related errors. Additionally, real CHB sequencing data can have highly variable quality and coverage distributions across the genome. For the analyses of real longitudinally collected CHB samples, graph-based sequence alignment consistently achieved higher proportions of successfully aligned HBV sequence data compared to any single linear reference (N=44), with 98.6%, 98.7%, 93.5%, 99.4%, and 98.7% of successfully aligned sequence for samples SRR747499, SRR747500, SRR747501, SRR747502, SRR747513 (**Figure 2**), respectively (**Figure S4**). The choice of linear reference had a significant effect on the proportion of aligned sequences across the five samples, with a per-sample difference between the best and worst performing linear reference ranging from 32.8% to 38.1%. The best performing linear reference (HBV subgenotype B2, GenBank ID: GU815637) resulted in 98.5%, 98.6%, 92.4%, 99.3%, and 98.6% of aligned sequences for each sample (SRR747499, SRR747500, SRR747501, SRR747502, SRR747513, respectively), which were all lower than the proportion of successful graph-based alignments. Notably, differences were also observed between references of the ‘correct’ HBV genotype (B), with a per-sample difference between the best and worst-performing reference ranging between 7.2% (85.3% vs. 92.4% for SRR7471501) and 7.8% (90.8% vs. 98.6% for SRR7471513). Thus, compared to graph-based alignment across these samples, up to 8.8% of HBV sequencing data can be missed due to the use of a linear reference sequence of the *correct* HBV genotype compared to the HBV reference graph.

HBV-derived sequences that failed to align to the non-subgenotype B2 reference were also not uniformly distributed across the genome (**Figure S5**). The distribution of where rescued reads aligned to the B2 reference is informative as more reads from non-HBV genotype B references were rescued across the HBV genome except in the pre-core/core region. At this locus, the average distribution of rescued reads from genotype C was always the lowest (**Figure S5**), which is unsurprising as the pre-core/core region within HBV subgenotype B2 reflects a known recombination event between genotypes B and C.[55] For reads which still failed to align to the best performing linear B2 reference sequence across each sample, 30.5%-63.2% were rescued via graph-based alignment (N=130,154, 71,526, 83,348, 61,692, 52,071 reads rescued, respectively). The distribution in the start sites of all rescued reads was similarly non-uniformly distributed. Interestingly, the loci in which graph-based alignment rescued the most reads also corresponded to loci with increased pairwise nucleotide diversity estimated across the 44 phylogenetically representative proposed HBV reference sequences (**Figure S6**), suggesting regions of increased genetic diversity globally may correspond to loci of increased intra-host sequence variation in real CHB samples.

There was no significant difference in the observed proportion of successfully aligned reads when using all or a subset of the QC-passed HBV sequencing data across the real CHB samples (P>0.99). Additionally, there was no significant difference in the proportion of aligned reads when the full-length linear reference sequences or extended linear reference sequences were used across the linear reference-based alignments (P=0.76) (**Figure S7**).

Graph-derived consensus sequences are more genetically similar to HBV sequencing datasets

While no single full genome-length HBV sequence could realistically capture the sequence variation observed across our simulated high throughput HBV sequencing data, the graph-based variant calling performed using the variation graph toolkit (VG) provided a consensus sequence with the lowest genetic distance to the full set of HBV genomes used in the simulations (**Figure S8**) mash distances ranging between 7.50×10^{-2} and 7.82×10^{-2}. Using the Mash distance as an approximation of average nucleotide identity (ANI), consensus sequences had ANIs ranging between 92.2%-92.5%, with the consensus inferred from VG-based variant calling having the highest ANI (92.5%).

For consensus sequences derived using the subset of HBV sequencing data generated from HBV genotypes B/C only, VG-based variant calling also resulted in the sequence with the lowest Mash distance and highest ANI compared to the full HBV genotype B/C sequences. However, all genotype B and C specific consensus sequences had similar Mash distances, ranging between 6.01×10^{-2} and 6.17×10^{-2}, and ANIs ranged between 93.8% and 94.0% (**Figure S9**).

For analyses of the real CHB sequencing data, the *de novo* assembled viral haplotypes always had the lowest Mash distance compared to the HBV sequencing data for each sample. This suggests viral haplotypes comprising an individual's CHB quasispecies better approximate the overall sequence diversity of an infection than any derived consensus sequence.

For consensus sequence comparisons, a graph-based variant calling approach resulted in consensus sequences with the lowest average Mash distance and highest ANI compared to each sample-specific set of HBV sequencing data across the longitudinal CHB samples. Our graph-based consensus sequence construction method provided improvements (i.e., a reduction in genetic distance) over attempts involving linear reference sequences when variants were identified via LoFreq and every sample other than SRR7471499 when consensus sequences were generated using iVar (**Figure 3**, **Figure S10**). For this sample, graph-based variant calling and consensus sequence generation resulted in the same genetic similarity estimate (ANI=89.3%) as a consensus derived using a subgenotype C1 reference sequence (GenBank ID: DQ089781). While both iVar and LoFreq can be used to identify variants across a diverse set of viral pathogens,[71,72] LoFreq has repeatedly been used to identify variants from real CHB-derived HBV sequencing data.[73,74] Additionally, while both consensus identification methods used the same site-specific depth threshold, our LoFreq-based approach accounted for insertions and deletions, potentially explaining its consistently lower Mash distance compared to the consensus sequences obtained via iVar (**Figure 3**, **Figure S10**).

For linear reference iVar-based variant calling across all samples, Mash distances ranged between 1.07×10^{-1} and 1.27×10^{-1}, for SRR7471499, SRR7471500, and SRR7471502. For SRR7471501, distances ranged between 1.07×10^{-1} and 1.30×10^{-1} and for SRR7471513 ranged between 1.07×10^{-1} and 1.28×10^{-1}. Mash distances from the LoFreq-derived consensus sequences ranged between 1.07×10^{-1} and 1.10×10^{-1} for SRR7471499, SRR7471501, SRR7471502, and SRR7471513 and 1.07×10^{-1} to 1.09×10^{-1} for SRR7471500. VG-based variant calling derived consensus sequences each had Mash distances of 1.07×10^{-1}. Additionally, we observed no differences in the Mash distances between consensus sequences derived using reads surjected into the HBV subgenotype B2 path (B2, GenBank ID: GU815637) and the linear reference-based alignment derived consensus using the same B2 reference.

Discussion

In this study, we confirm that the choice of reference plays a critical role in the alignment of high throughput HBV sequencing data and can influence the construction of sample-specific consensus sequences in genetic studies of CHB. We also demonstrate that sequence variation graphs can improve upon widely accepted methodologies used for sequence alignment of highly diverse pathogens such as HBV. Using both real CHB and simulated high diversity HBV sequencing datasets, we show that alignment to a phylogenetically representative reference graph results in a higher proportion of successful sequence alignment and facilitates the generation of accurate sample-specific consensus sequences.

As the benefits of sequence-to-graph alignment are greatest for highly diverse sequencing datasets, the utility of graph-based sequence alignment is dependent upon the research question of interest. For example, sequence-to-graph alignment recovers only marginally more simulated sequencing data generated from a subset of HBV genotype B and C sequences (**Figure S3**) compared to linear reference-based approaches using genotype B or C reference sequences. Furthermore, linear reference-based sequence alignment is highly successful at capturing HBV sequences from regions across the HBV genome with non-extreme global sequence diversity (**Figure 1**). While our results demonstrate that for regions of increased diversity any single linear reference is likely insufficient to capture the genetic variation observed across all HBV genotypes/subgenotypes or mixed CHB infections, many CHB infections are comprised of a single HBV genotype, and thus linear reference-based alignment using a correct genotype/subgenotype sequence would not be expected to omit important information. This is, however, not guaranteed as we find >8% of viral sequence can fail aligning to a phylogenetically representative linear reference sequence. For more genetically diverse infections, a hybrid approach in which a linear reference-based alignment is followed by graph alignment of unmapped reads could also solve issues related to reference ambiguity while limiting the computational burden associated with graph-based sequence alignment. Notably, we find reads rescued via graph alignment largely originate from regions across the HBV genome of increased global sequence diversity (**Figure S6**), suggesting these loci could also correspond to regions of increased intra-host genetic variation.

The generation of consensus sequences is an important product of microbial-focused genomic analyses, and novel software and workflows devoted to generating pathogen consensus sequences, including for SARS-CoV-2,[40] continue to be developed. In addition to providing an accurate characterization of the genome comprising a clinical infection, publicly available consensus sequences enable molecular epidemiology-focused research, allow for large-scale phylogenetic analysis, and can aid disease surveillance efforts.[75–78] Consensus sequences can also serve as the ‘reference’ sequence in subsequent bioinformatic analyses, reducing the number of spuriously identified variants for HBV.[32] Thus, care should be taken to ensure the most accurate and genetically representative sequences are obtained from clinical CHB or other HBV-infected samples. We show that graph-based alignment and variant calling can often improve upon linear reference-based approaches to derive sample-specific consensus sequences, even when such efforts utilize reference sequences of the correct HBV subgenotype, which produces consensus sequences less genetically similar to an individual’s CHB quasispecies than sequences derived via graph alignment. However, given the minute differences in average nucleotide identity between consensus sequences obtained from any of the best performing linear references and our graph-based approach, the deleterious consequences of linear reference-based alignment are likely minimal when effort is made to first identify a genetically representative reference sequence for use in alignment.

However, when reference selection is not carefully considered, unrepresentative reference sequences can impact the fidelity of consensus sequences and other downstream phylogenetic-focused analyses.[33,79] It is therefore possible that some publicly available full-length HBV genome sequences are not the most accurate HBV-related sample-specific consensus sequences. While alternative approaches to infer sample-specific consensus sequences exist,[5,32] our approach using the Mash distance to compare and identify the consensus sequence that best approximates the set of HBV sequencing data or *de novo* assembled haplotypes could provide a mechanism by which sample-specific consensus sequences are compared and selected for use as ideal reference sequences.

While sequence-to-graph alignment requires more computational resources than linear alignment-based approaches, especially for the VG map mapper (**Table S1**), if the goal is to capture and retain as much HBV-related sequencing data as possible for analysis, we show that graph-based methods outperform traditional linear reference-based alignment for HBV. We should note that effective tools enabling sequence-to-graph alignment and the subsequent identification of graph-derived genetic variation are a relatively recent development. Improvements in computational performance have already been demonstrated through graph-simplification and the development of more advanced mapping and variant identification models,[80–82] with further improvements expected.[83]

While alternatives to graph-based alignment which leverage multiple reference sequences have also been developed, such as alignment using multiple linear references in tandem,[84,85] their performance has not been assessed using HBV sequencing data. Furthermore, an added benefit of graph-based approaches is that differences observed between the embedded paths/reference sequences can be utilized during variant calling to identify loci of genetic variation, in addition to mutations inferred from the alignment of sequencing data directly. The ability to leverage graph topology was demonstrated in our use of path depth to infer the genotype of HBV sequences aligned to the reference graph. Alternative graph-based prediction methods, including models for HBV subgenotype prediction or recombination detection, are worth further exploration. Whether metrics linked to graph topology or complexity, including path depth, can be used to better characterize the viral genetics of CHB quasispecies, either within specific regions or across the genome,[86,87] or the genetic diversity of HBV generally, remains unexplored. For example, we observe the distribution of path depth within the phylogenetically representative HBV reference graph approximates a universal gene frequency distribution typical of many bacterial species (**Figure S2**),[34] despite there being no distinction between a core and accessory genome for HBV. Future efforts should investigate the utility and potential clinical importance of these graph-derived measures of genetic complexity for HBV and other microbial pathogens of public health importance. For example, graph-to-graph comparisons could enable the analysis of genetic sequence data in ways that Euclidean data structures cannot.

Graph-based approaches are being increasingly used to investigate highly genetically diverse microbial pathogens and regions of the human genome. In this study, we demonstrate the limitations of using linear HBV reference sequences to derive consensus sequences for CHB samples. Furthermore, we hope to mitigate issues of HBV reference ambiguity by making this HBV reference graph publicly available, which will also promote the use of graph-based advances in genetic analyses to improve our understanding of CHB genetics.

Declaration of interests

None

Acknowledgments

Funding for this work was provided via 2R01AI148049 along with a COVID-19 supplement under the same grant number (D.L.T., P.D., G.L.W) and Burroughs-Wellcome Fund, MD-GEM training grant (D.D.). Support was also provided by the National Human Genome Research Institute (NHGRI) grant R35HG011944 (G.L.W.).

Data and code availability

The longitudinal HBV sequencing data utilized in this study is available as an NCBI BioProject under the accession PRJNA479693. Simulated HBV sequencing data and the HBV reference graph have been deposited on Zenodo and can be accessed using the following doi: 10.5281/zenodo.6646207.

Code used to construct and index the HBV reference graph, align sequencing data to the graph, and infer a consensus sequence can be accessed at https://github.com/dduchen/HBV_reference_graph_manuscript.

REFERENCES

1. WHO. Preventing Perinatal Hepatitis B Virus Transmission□: A Guide for Introducing and Strengthening Hepatitis B Birth Dose Vaccination. (2015).
2. Asselah, T., Loureiro, D., Boyer, N. & Mansouri, A. Targets and future direct-acting antiviral approaches to achieve hepatitis B virus cure. *Lancet Gastroenterol. Hepatol.* **4**, 883–892 (2019).
3. Zhang, Z. *et al.* Host Genetic Determinants of Hepatitis B Virus Infection. *Front. Genet.* **10**, 1–24 (2019).
4. Trépo, C., Chan, H. L. Y. & Lok, A. Hepatitis B virus infection. *Lancet* **384**, 2053–2063 (2014).
5. Podlaha, O. *et al.* Large-scale viral genome analysis identifies novel clinical associations between hepatitis B virus and chronically infected patients. *Sci. Rep.* **9**, 10529 (2019).
6. Akahane, Y. *et al.* Chronic active hepatitis with hepatitis B virus DNA and antibody against e antigen in the serum. Disturbed synthesis and secretion of e antigen from hepatocytes due to a point mutation in the precore region. *Gastroenterology* (1990) doi:10.1016/0016-5085(90)90632-B.
7. Günther, S. *et al.* Type, prevalence, and significance of core promoter/enhancer II mutations in hepatitis B viruses from immunosuppressed patients with severe liver disease. *J. Virol.* (1996).
8. Günther, S., Piwon, N. & Will, H. Wild-type levels of pregenomic RNA and replication but reduced pre-C RNA and e-antigen synthesis of hepatitis B virus with C(1653)→T, A(1762)→T and G(1764)→A mutations in the core promoter. *J. Gen. Virol.* (1998) doi:10.1099/0022-1317-79-2-375.
9. Nguyen, M. H. & Keeffe, E. B. Are hepatitis B e antigen (HBeAg)-positive chronic hepatitis B and HBeAg-negative chronic hepatitis B distinct diseases? *Clin. Infect. Dis.* **47**, 1312–4 (2008).
10. Cao, G.-W. Clinical relevance and public health significance of hepatitis B virus genomic variations. *World J. Gastroenterol.* **15**, 5761–9 (2009).
11. McNaughton, A. L. *et al.* Insights From Deep Sequencing of the HBV Genome—Unique, Tiny, and Misunderstood. *Gastroenterology* **156**, 384–399 (2018).
12. Toan, N. L. *et al.* Impact of the hepatitis B virus genotype and genotype mixtures on the course of liver disease in Vietnam. *Hepatology* **43**, 1375–1384 (2006).
13. Lin, C.-L. *et al.* High prevalence of occult hepatitis B virus infection in Taiwanese intravenous drug users. *J. Med. Virol.* **79**, 1674–1678 (2007).
14. Shen, L. *et al.* Molecular epidemiological study of hepatitis B virus genotypes in Southwest, China. *J. Med. Virol.* **86**, 1307–1313 (2014).
15. Liu, B., Yang, J., Yan, L., Zhuang, H. & Li, T. Novel HBV recombinants between genotypes B and C in 3′-terminal reverse transcriptase (RT) sequences are associated with enhanced viral DNA load, higher RT point mutation rates and place of birth among Chinese patients. *Infect. Genet. Evol.* **57**, 26–35 (2018).
16. Huy, T. T. T., Ngoc, T. T. & Abe, K. New Complex Recombinant Genotype of Hepatitis B Virus Identified in Vietnam. *J. Virol.* **82**, 5657–5663 (2008).
17. Tatematsu, K. *et al.* A Genetic Variant of Hepatitis B Virus Divergent from Known Human and Ape Genotypes Isolated from a Japanese Patient and Provisionally Assigned to New Genotype J. *J. Virol.* **83**, 10538–10547 (2009).
18. Guirgis, B. S. S., Abbas, R. O. & Azzazy, H. M. E. Hepatitis B virus genotyping: Current methods and clinical implications. *Int. J. Infect. Dis.* **14**, e941–e953 (2010).
19. Shi, W. *et al.* Hepatitis B virus subgenotyping: History, effects of recombination,

misclassifications, and corrections. *Infect. Genet. Evol.* **16**, 355–361 (2013).

20. Domingo, E., Sheldon, J. & Perales, C. Viral Quasispecies Evolution. *Microbiol. Mol. Biol. Rev.* **76**, 159–216 (2012).
21. Poirier, E. Z. & Vignuzzi, M. Virus population dynamics during infection. *Curr. Opin. Virol.* **23**, 82–87 (2017).
22. Zhou, T.-C. *et al.* Evolution of full-length genomes of HBV quasispecies in sera of patients with a coexistence of HBsAg and anti-HBs antibodies. *Sci. Rep.* **7**, 661 (2017).
23. Yang, Z.-T. *et al.* Characterization of Full-Length Genomes of Hepatitis B Virus Quasispecies in Sera of Patients at Different Phases of Infection. *J. Clin. Microbiol.* **53**, 2203–2214 (2015).
24. Domingo, E. & Perales, C. Viral quasispecies. *PLOS Genet.* **15**, e1008271 (2019).
25. Cao, L. *et al.* Coexistence of Hepatitis B Virus Quasispecies Enhances Viral Replication and the Ability To Induce Host Antibody and Cellular Immune Responses. *J. Virol.* **88**, 8656–8666 (2014).
26. Zhang, A. Y. *et al.* Deep sequencing analysis of quasispecies in the HBV pre-S region and its association with hepatocellular carcinoma. *J. Gastroenterol.* **52**, 1064–1074 (2017).
27. Cheng, Y. *et al.* Cumulative viral evolutionary changes in chronic hepatitis B virus infection precedes hepatitis B e antigen seroconversion. *Gut* **62**, 1347–1355 (2013).
28. Chen, L. *et al.* Increased intrahepatic quasispecies heterogeneity correlates with off-treatment sustained response to nucleos(t)ide analogues in e antigen-positive chronic hepatitis B patients. *Clin. Microbiol. Infect.* **22**, 201–207 (2016).
29. Liu, F. *et al.* Evolutionary patterns of hepatitis B virus quasispecies under different selective pressures: correlation with antiviral efficacy. *Gut* **60**, 1269–1277 (2011).
30. Ngui, S. L. & Teo, C. G. Hepatitis B virus genomic heterogeneity: Variation between quasispecies may confound molecular epidemiological analyses of transmission incidents. *J. Viral Hepat.* **4**, 309–315 (1997).
31. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–881 (2018).
32. Liu, W.-C. *et al.* Aligning to the sample-specific reference sequence to optimize the accuracy of next-generation sequencing analysis for hepatitis B virus. *Hepatol. Int.* **10**, 147–157 (2016).
33. Valiente-Mullor, C. *et al.* One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLOS Comput. Biol.* **17**, e1008678 (2021).
34. Colquhoun, R. M. *et al.* Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. *Genome Biol.* **22**, 267 (2021).
35. Rick, J. A., Brock, C. D., Lewanski, A. L., Golcher-Benavides, J. & Wagner, C. E. Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. *bioRxiv* 2022.03.10.483737 (2022) doi:10.1101/2022.03.10.483737.
36. McNaughton, A. L., Revill, P. A., Littlejohn, M., Matthews, P. C. & Ansari, M. A. Analysis of genomic-length HBV sequences to determine genotype and subgenotype reference sequences. *J. Gen. Virol.* **101**, 271–283 (2020).
37. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, annurev-genom-120219-080406 (2020).
38. Eizenga, J. M. *et al.* Succinct dynamic variation graphs. 1–6 (2020).
39. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).

40. Moshiri, N. *et al.* The ViReflow pipeline enables user friendly large scale viral consensus genome reconstruction. *Sci. Rep.* **12**, 5077 (2022).
41. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).
42. Bui, T. T. T. *et al.* Molecular characterization of hepatitis B virus in Vietnam. *BMC Infect. Dis.* **17**, 601 (2017).
43. McNaughton, A. L. *et al.* Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). *Sci. Rep.* **9**, 7081 (2019).
44. Cheng, Y. *et al.* Multifactorial heterogeneity of virus-specific T cells and association with the progression of human chronic hepatitis B infection. *Sci. Immunol.* **4**, eaau6905 (2019).
45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
46. Andrews, S. FastQC. *Babraham Bioinforma.* (2010).
47. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J. & Bongcam-Rudloff, E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics* **35**, 521–522 (2019).
48. Marco-Sola, S., Moure, J. C., Moreto, M. & Espinosa, A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**, 456–463 (2021).
49. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *bioRxiv* 14–19 (2022) doi:10.1101/2022.02.14.4804.
50. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science (80-.).* **374**, (2021).
51. Sirén, J., Garrison, E., Novak, A. M., Paten, B. & Durbin, R. Haplotype-aware graph indexes. *Bioinformatics* **36**, 400–407 (2018).
52. Hall, M. Rasusa: Randomly subsample sequencing reads to a specified coverage. *J. Open Source Softw.* **7**, 3941 (2022).
53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
54. Velkov, S., Ott, J., Protzer, U. & Michler, T. The Global Hepatitis B Virus Genotype Distribution Approximated from Available Genotyping Data. *Genes (Basel).* **9**, 495 (2018).
55. Sugauchi, F. *et al.* Hepatitis B Virus of Genotype B with or without Recombination with Genotype C over the Precore Region plus the Core Gene. *J. Virol.* **76**, 5985–5992 (2002).
56. Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–20 (2010).
57. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).
58. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–4 (2016).
59. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies: Fig. 1. *Bioinformatics* **31**, 3350–3352 (2015).
60. Holley, G. & Melsted, P. Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biol.* **21**, 249 (2020).
61. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
62. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4

(2021).

63. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).
64. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
65. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
66. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
67. Solis-Reyes, S., Avino, M., Poon, A. & Kari, L. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS One* **13**, 1–21 (2018).
68. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
69. Baaijens, J. ., El Aabidine, A. Z., Rivals, E. & Schönhuth, A. De novo assembly of viral quasispecies using overlap graphs. *Genome Res.* **27**, 835–848 (2017).
70. Baaijens, J. A., Stougie, L. & Schönhuth, A. Strain-aware assembly of genomes from mixed samples using flow variation graphs. *bioRxiv* 645721 (2020) doi:10.1101/645721.
71. Dezordi, F. Z. *et al.* ViralFlow: A Versatile Automated Workflow for SARS-CoV-2 Genome Assembly, Lineage Assignment, Mutations and Intrahost Variant Detection. *Viruses* **14**, 217 (2022).
72. Liu, Y. *et al.* Rescuing low frequency variants within intra-host viral populations directly from Oxford Nanopore sequencing data. *Nat. Commun.* **13**, 1321 (2022).
73. Zhu, Y. O. *et al.* Single-virion sequencing of lamivudine-treated HBV populations reveal population evolution dynamics and demographic history. *BMC Genomics* **18**, 1–12 (2017).
74. Betz-Stablein, B. D. *et al.* Single-Molecule Sequencing Reveals Complex Genome Variation of Hepatitis B Virus during 15 Years of Chronic Infection following Liver Transplantation. *J. Virol.* **90**, 7171–7183 (2016).
75. Saravanan, K. A. *et al.* Role of genomics in combating COVID-19 pandemic. *Gene* **823**, 146387 (2022).
76. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 2–4 (2017).
77. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
78. Armstrong, G. L. *et al.* Pathogen Genomics in Public Health. *N. Engl. J. Med.* **381**, 2569–2580 (2019).
79. Rick, J. A., Brock, C. D., Lewanski, A. L., Golcher-Benavides, J. & Wagner, C. E. Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. *bioRxiv* 2022.03.10.483737 (2022) doi:10.1101/2022.03.10.483737.
80. Jain, C., Tavakoli, N. & Aluru, S. A variant selection framework for genome graphs. *bioRxiv* 1–8 (2021) doi:10.1101/2021.02.02.429378.
81. Pritt, J., Chen, N.-C. & Langmead, B. FORGe: prioritizing variants for graph genomes. *Genome Biol.* **19**, 220 (2018).
82. Monsu, M. & Comin, M. Fast alignment of reads to a variation graph with application to SNP detection. *J. Integr. Bioinform.* **18**, (2021).
83. Baaijens, J. A. *et al.* Computational graph pangenomics: a tutorial on data structures and their applications. *Nat. Comput.* **6**, (2022).
84. Chen, N., Paulin, L. F., Sedlazeck, F. J., Koren, S. & Adam, M. Improved

sequence mapping using a complete reference genome and lift-over. *bioRxiv* (2022) doi:10.1101/2022.04.27.489683.

85. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* **22**, 8 (2021).

86. Ibragimov, R., Malek, M., Guo, J. & Baumbach, J. GEDEVO: An evolutionary graph edit distance algorithm for biological network alignment. *OpenAccess Ser. Informatics* **34**, 68–79 (2013).

87. Qiu, Y. & Kingsford, C. The Effect of Genome Graph Expressiveness on the Discrepancy Between Genome Graph Distance and String Set Distance. *bioRxiv* (2022) doi:10.1101/2022.02.18.481102.

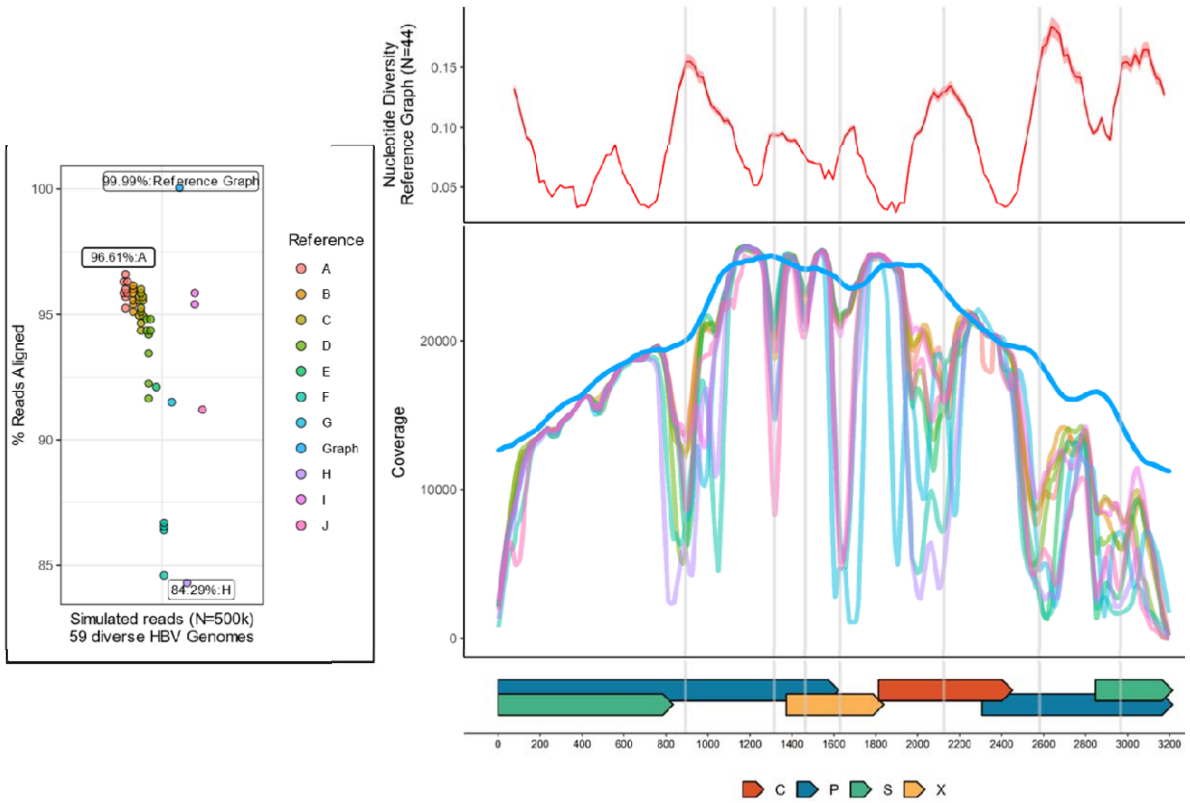

Figure 1: Alignment and depth of coverage across the HBV genome for simulated HBV sequencing data. On the left, points reflect the proportion of successfully aligned simulated reads, colored by either the genotype of the reference or whether graph-based alignment was performed. The Y axis reflects the proportion of successfully aligned reads. Labels indicate the reference sequence genotype or graph used in the alignment and the proportion of reads aligned. On the right, the X axis reflects genome position, with gene regions provided as colored bars along the base of the figure. C=core, P=polymerase, S=Surface, X=X. The top right panel reflects the average nucleotide diversity across the genome, with the Y axis reflecting the average pairwise nucleotide diversity (0-1). For the bottom right panel, the Y axis reflects depth of coverage across the HBV genome obtained when using reference sequences of different HBV genotypes, using the same genotype-specific color scheme as the left panel. Significant drops in coverage are indicated with the grey vertical lines.

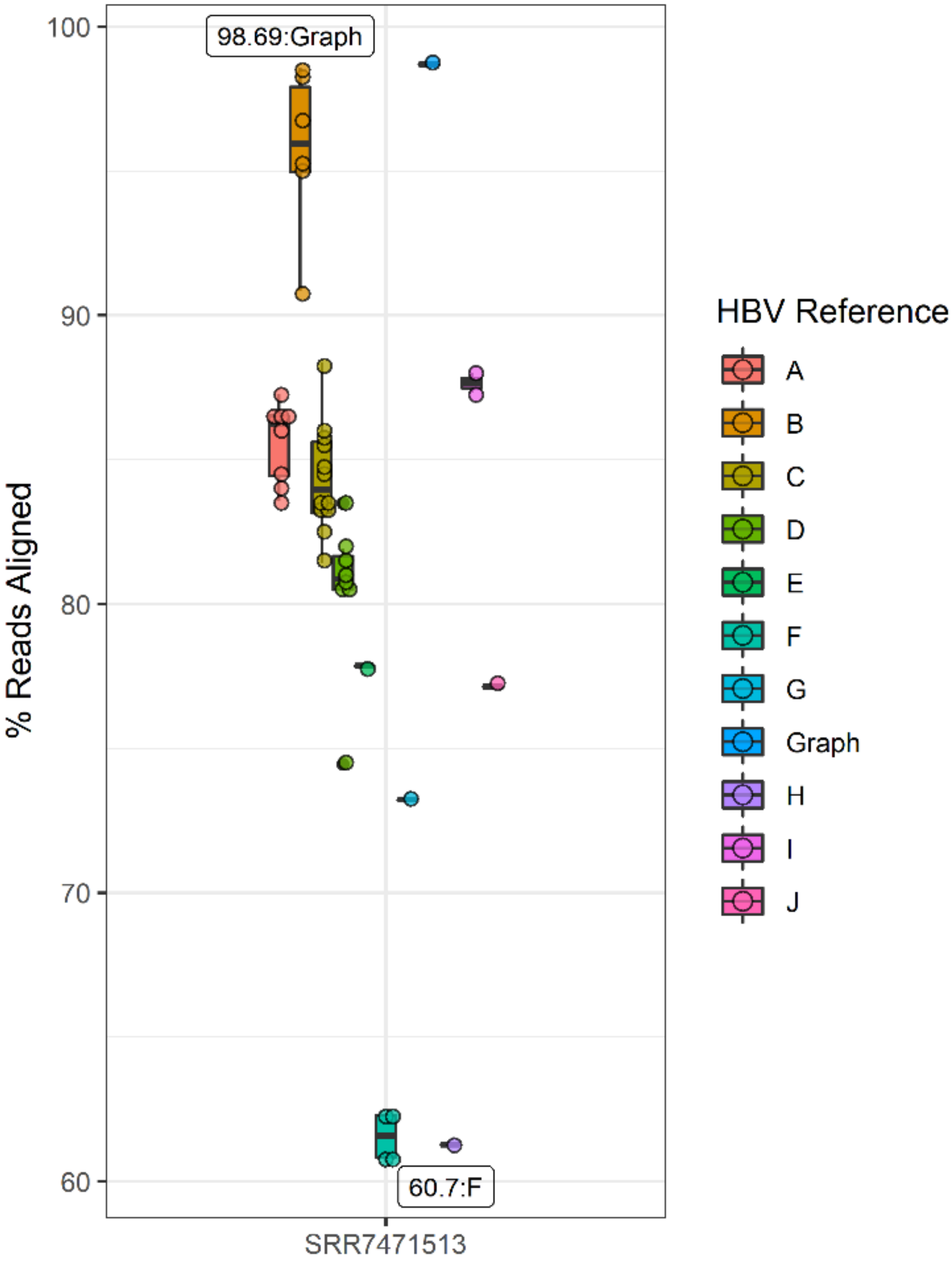

Figure 2: Proportion of successfully aligned CHB sequencing data for sample SRR7471513. Points reflect the proportion of successfully aligned sequences, colored by either the genotype for linear reference-based alignment or if sequences were aligned to the HBV reference graph. The Y axis reflects the proportion of successfully aligned reads. The X axis indicates which longitudinally collected CHB sample was used. Labels reflect the genotype of the reference or whether the reference graph was used for the highest and lowest observed aligned proportions.

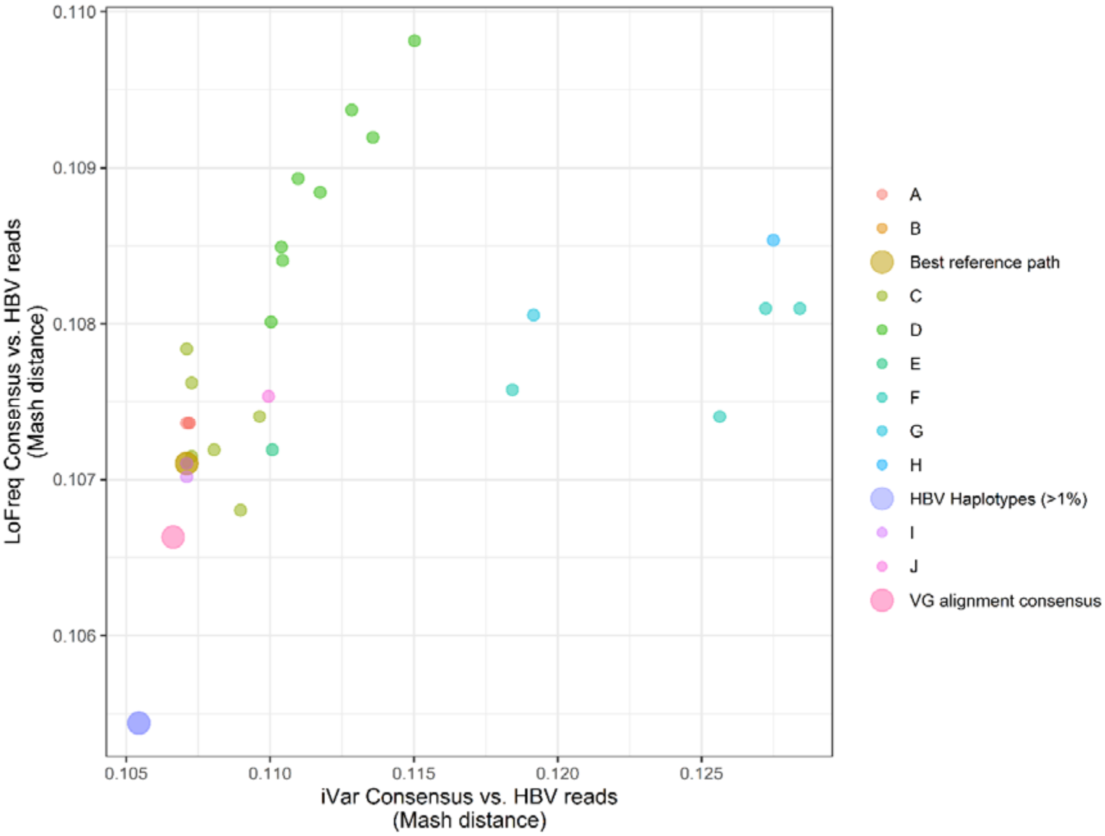

Figure 3: Genetic distance comparisons of consensus sequences and *de novo* assembled HBV haplotypes with CHB sequencing data from sample SRR7471513. Points reflect the Mash distance estimated between each consensus sequence generated from the 44 HBV reference sequences or the HBV reference graph. The Y axis reflects the Mash distance estimated between each LoFreq-derived consensus sequence and the X axis reflects the Mash distance estimated between each iVar-derived consensus sequence. The color of each point reflects the genotype of the reference used to generate a consensus, or if the consensus was derived via graph-based alignment or reflects sample-specific HBV haplotypes. Points for graph-derived consensus sequences, including VG-based variant calling ('VG alignment consensus'), graph-based surjection ('Best reference path'), and the *de novo* assembled viral strains ('HBV Haplotypes (>1%)') are enlarged.

Genotype	Whole-genome alignment (N)	Short-read alignment (N)
A	259	7
B	687	17
C	1094	19
D	549	9
E	145	2
F	80	2
G	3	1
H	11	1
I	9	1

Table 1: Sequence-to-graph alignment genotype prediction. Whole genome sequences or simulated high throughput (i.e., short-read) sequencing data consistently aligned best to graph-embedded paths of the correct HBV genotype. The number of each aligned sequences or datasets from the respective HBV genotype are listed.