viruses

MDPI

Article

Validation of Variant Assembly Using HAPHPIPE with Next-Generation Sequence Data from Viruses

Keylie M. Gibson [1,*,†], Margaret C. Steiner [1,†], Uzma Rentia [1], Matthew L. Bendall [1], Marcos Pérez-Losada [1,2,3] and Keith A. Crandall [1,2]

1 Computational Biology Institute, Milken Institute School of Public Health, The George Washington University, Washington, DC 20052, USA; steinerm@gwmail.gwu.edu (M.C.S.); uzma_rentia@gwmail.gwu.edu (U.R.); mlb4001@med.cornell.edu (M.L.B.); mlosada@gwu.edu (M.P.-L.); kcrandall@gwu.edu (K.A.C.)

2 Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, Washington, DC 20052, USA

3 CIBIO-InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4169-007 Vairão, Portugal

* Correspondence: kmgibson@gwu.edu

† These authors contributed equally to this work.

Received: 28 May 2020; Accepted: 6 July 2020; Published: 14 July 2020

Abstract: Next-generation sequencing (NGS) offers a powerful opportunity to identify low-abundance, intra-host viral sequence variants, yet the focus of many bioinformatic tools on consensus sequence construction has precluded a thorough analysis of intra-host diversity. To take full advantage of the resolution of NGS data, we developed HAplotype PHylodynamics PIPEline (HAPHPIPE), an open-source tool for the de novo and reference-based assembly of viral NGS data, with both consensus sequence assembly and a focus on the quantification of intra-host variation through haplotype reconstruction. We validate and compare the consensus sequence assembly methods of HAPHPIPE to those of two alternative software packages, HyDRA and Geneious, using simulated HIV and empirical HIV, HCV, and SARS-CoV-2 datasets. Our validation methods included read mapping, genetic distance, and genetic diversity metrics. In simulated NGS data, HAPHPIPE generated *pol* consensus sequences significantly closer to the true consensus sequence than those produced by HyDRA and Geneious and performed comparably to Geneious for HIV *gp120* sequences. Furthermore, using empirical data from multiple viruses, we demonstrate that HAPHPIPE can analyze larger sequence datasets due to its greater computational speed. Therefore, we contend that HAPHPIPE provides a more user-friendly platform for users with and without bioinformatics experience to implement current best practices for viral NGS assembly than other currently available options.

Keywords: bioinformatics; validation; simulation; viruses; consensus; haplotypes; HIV; HCV; SARS-CoV-2

1. Introduction

Next-generation sequence (NGS) data provide a new opportunity to more efficiently study viral diversity, especially within-host sequence variation, which is key to understanding the evolutionary dynamics of viral populations both within and amongst hosts. NGS provides an opportunity to better explore viral sequence evolution over time [1] and evolution among hosts, including the direction of cross-species transmission [2], or elucidate the origin of viral epidemics [3]. While some studies capitalize on the ability of NGS data to identify intra-host sequence variants, the majority rely on consensus sequence estimation. This results in a loss of resolution in intra-patient viral diversity,

which has nontrivial implications for downstream evolutionary inferences [4]. Thus, improving consensus sequence estimation methods is of great interest to the virology community.

There are two general approaches when constructing a consensus sequence from NGS data: de novo assembly and reference-based assembly (for reviews, see [5,6]). For reference-based assembly, sequencing reads are aligned (or mapped) to a reference sequence and a consensus sequence is then generated, often using majority rule, where the most frequently encountered nucleotide at each aligned position is chosen to be the nucleotide in the consensus sequence at that same position. Alternatively, consensus sequences can be generated using specified percentage cutoffs or by inserting ambiguity codes at sites with incongruities. De novo assembly does not require a reference sequence, but instead attempts to reconstruct the full sequence (or region of interest, such as an amplicon) by identifying overlapping nucleotides among the sequence reads. While reference-based assembly requires less memory, computational effort and sequencing depth, the generated consensus sequence may reflect the nucleotide composition of the reference sequence (i.e., bias towards the reference sequence) [7–9], thus potentially impacting the accuracy of downstream analyses. Issues in de novo assembly commonly arise from the large amount of computing effort required and the computational complexity of identifying overlapping regions in short reads. This issue is further compounded by highly variable short reads seen in quickly evolving retroviruses due to high genetic diversity, although small genomes still require relatively minimal computational power for assembly.

Recently, Ji et al. [10] recommended best practices for processing HIV NGS data, which include reference-based assembly using Bowtie2 [11] as the short read aligner and HXB2 (NCBI accession: K03455; [12]) as the reference sequence for constructing a consensus sequence. Many studies implement reference-based assembly [10,13,14] with tools such as CLC Main Workbench (Qiagen, Hilden, Germany) [15–18], Geneious (https://www.geneious.com) [19–24], HyDRA [25–28], SmartGene (Switzerland) [21,29,30], PAseq [31,32] and Amplicon Variant Analyzer (AVA; pyrosequencing-based platform) [21,33–37]. Other studies complete de novo assembly with tools such as Geneious [38], CLC Main Workbench (Qiagen) [16,39,40], and Iterative Virus Assembler (IVA) [22,41–47]. HCV studies follow similar patterns to those of HIV-1 [48–59]. A combination of both assembly approaches has been implemented to construct a consensus sequence by first mapping the reads to a reference sequence and then completing de novo assembly of those mapped reads [53,60].

Our software, HAplotype PHylodynamics PIPEline (HAPHPIPE), was designed to make viral NGS analyses more accessible and versatile for researchers and to provide the opportunity for identifying within-host variation by assembling variants from NGS data [61]. HAPHPIPE implements de novo or reference-based assembly followed by iterative refinement steps to assemble a better-representative consensus sequence for the entire viral population being surveyed. HAPHPIPE also implements haplotype reconstruction tools to facilitate the use of haplotypes in downstream analyses. The inclusion of haplotype data helps researchers to quantify within-host variation and, thereby, make improved inferences about associations with phenotypic traits.

A fundamental component of good software development is testing and validation [62]. Accordingly, we aim to test and validate HAPHPIPE using simulated HIV-1 data and empirical HIV-1 [63], HCV [64], and SARS-CoV-2 data. We hypothesize that in our simulation study: (i) due to the high genetic diversity of HIV-1, the de novo assembly strategy, regardless of platform, will produce a consensus sequence that is more genetically similar to the true sample isolate sequence than the reference-based assembly strategy; and (ii) that HAPHPIPE will perform as well as or better than HyDRA and Geneious, based on the metrics of read mapping, genetic distance, and genetic diversity. We also evaluate genetic distance to address potential bias in reference-based assembly. For the purposes of this study, we focus on the composition of the consensus sequences, but additionally report haplotype data for the HIV and HCV empirical datasets as validation of HAPHPIPE's intra-host analytical capabilities. See Eliseev et al. [65] for an evaluation of haplotype reconstruction tools.

One aspect of note in NGS assembly specific to HIV is that often, these analyses are performed by users whose primary work is not in bioinformatics, such as clinicians, due to the often highly

translational goals of HIV research. Because viral sequence analyses are rapidly becoming a staple of public health surveillance efforts, one of the primary goals of HAPHPIPE was to make intricate, command-based software accessible to introductory level bioinformatics users. Hence, we have included significant documentation and pre-scripted pipelines along with our software to facilitate its use by users of all backgrounds.

2. Materials and Methods

Below, we introduce the three tools that were used in generating consensus sequences, as well as the simulated and empirical datasets used in this study. Finally, we discuss the analyses completed for the pipelines and consensus sequences, along with true and reference sequences used (Figure 1). We compared HAPHPIPE to two other software programs, Geneious and HyDRA, based on their frequent use in viral studies [19–28,38,51–53,56,57], particularly among clinicians and those new to bioinformatics analysis. In particular, we chose HyDRA over similar web-based platforms such as PASeq [66] due to its popularity in the HIV research community; we chose Geneious as a representative of commercial software frequently used in genomics analysis. In selecting these methods, we target this validation study on performance for clinical and public health applications—which are especially pertinent for the empirical viral data included: HIV, HCV, and SARS-CoV-2. We suggest HAPHPIPE as a viable alternative to commercial and closed-source platforms for these applications.

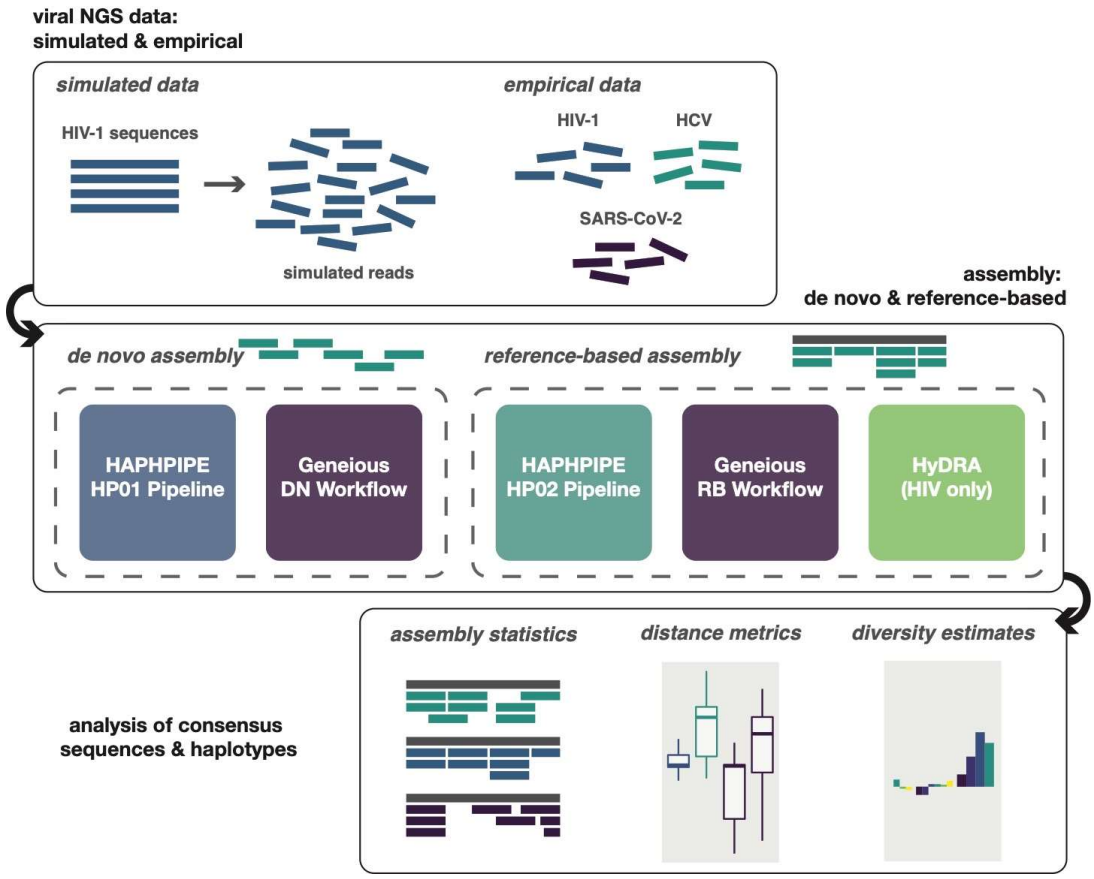

Figure 1. Methods overview. Sequencing reads were simulated for each simulation sample, while reads for empirical data were gathered from NCBI SRA database. Reads for each sample were assembled and a consensus sequence generated through the de novo pipelines for HAPHPIPE (HP01) and Geneious and the reference-based pipelines for HAPHPIPE (HP02), Geneious, and HyDRA. Only HIV samples were analyzed through HyDRA, because HyDRA is HIV-specific. All resulting consensus sequences were analyzed using a variety of metrics including assembly statistics, genetic distance from reference or true sequence metrics, and diversity estimates, such as nucleotide diversity.

2.1. HAPHPIPE

HAPHPIPE is a user-friendly tool designed for the customizable processing and analysis of viral NGS data [61]. HAPHPIPE is available for installation via Bioconda, a popular open-source, bioinformatics-specific software distribution system for Python packages. Installation of the HAPHPIPE suite requires only one command. HAPHPIPE is constructed in a modular format that has five main components: manipulating reads (*reads* stage), assembling consensus sequences (*assemble* stage), haplotype reconstruction (*haplotype* stage), post-analysis steps such as summary statistics or region extraction (*description* stage) and phylogenetics (*phylo* stage) [61]. There are two example pipelines included in the tool: haphpipe_assemble_01 and haphpipe_assemble_02 (Table 1). Haphpipe_assemble_01 is a de novo assembly pipeline that takes raw Illumina sequencing data, quality trims and error corrects the reads with Trimmomatic [67] and SPAdes [68], respectively, completes de novo assembly to form contigs with SPAdes, and forms scaffolds from the contigs with MUMMER 3+ [69]. Finally, the corrected reads are mapped back to the de novo assembled sequence with Bowtie2 [11]. The initial consensus sequence is updated through iterative refinement steps, where the corrected reads are repeatedly mapped back to the newly formed consensus sequence and the consensus sequence is updated with the new majority nucleotides at a position using a modified majority rule (taking quality scores and read depth into account). By default, this continues until the consensus sequence shows no improvement or changes to base composition or five refinement steps are completed. Alternatively, haphpipe_assemble_02 is a reference-based assembly approach, in which error corrected reads are mapped against a reference genome with Bowtie2 instead of being assembled de novo. All other steps are the same as in haphpipe_assemble_01.

Table 1. Characteristics of the compared programs.

Pipeline	Haphpipe_Assemble_01	Haphpipe_Assemble_02	HyRDA	Geneious	Geneious
Assembly approach	De novo	Reference	Reference	Reference	De novo
Reference	None	Flexible	HXB2	Flexible	None
Consensus refinement	Available	Available	No	Available	No
Genes	All	All	*pol*	All	All
Use	Free, Command line	Free, Command line	Free, Web-based	Paid GUI	Paid GUI
Read trimming	Available, Trimmomatic	Available, Trimmomatic	Yes, parameters set by user	Available, modified-Mott algorithm	Available, modified-Mott algorithm
Error correction	Available, SPAdes	Available, SPAdes	Set sequencing platform error rate	Available, BBNorm *	Available, BBNorm *

* BBNorm is part of BBTools package and found at http://seqanswers.com/forums/showthread.php?t=49763.

HAPHPIPE also implements PredictHaplo [70] and CliqueSNV [71], two haplotype reconstruction tools. For purposes of this study, we present results generated by PredictHaplo, which was chosen for the HAPHPIPE suite because it was determined to have the best performance for capturing intra-host viral variation compared to eleven other haplotype reconstruction tools in a recent study [21] of diversity levels observed in viral intra-patient data. For a more thorough explanation of HAPHPIPE and detailed user instructions, see Bendall et al. [61] or https://gwcbi.github.io/haphpipe_docs/.

2.2. HyDRA

HyDRA is a freely available, web-based tool that utilizes a wrapper for Bowtie2 for reference-based mapping of Illumina MiSeq reads to HXB2, similar to the HAPHPIPE reference-based pipeline (Table 1). Briefly, the default parameters included: a minimum of 20% frequency for a base to be included in the consensus sequence, the default mutation database—which is the Stanford SDRM 2009 list of mutations—a target coverage of 10,000 reads, a minimum read length of 100 bp, a minimum average read quality score of 30, a sequencing platform error rate of 0.0021, a minimum variant quality of 30, a minimum read depth of 100 for a variant call, a minimum allele count of five to be considered a variant, and a minimum amino acid frequency of 0.01 for a mutation to be considered in the drug-resistant report.

Compared to HAPHPIPE, HyDRA requires a target read coverage (10,000 reads) whereas HAPHPIPE does not require a target. For quality trimming of the reads, the defaults for HAPHPIPE include trimming base pairs from the 3′ and 5′ ends of the reads that fall below a quality of 3 or contain 'N', removing any leftover adapter sequences, a sliding window of 4:15, which clips the read once the average quality of the window is below 15, and requires a minimum read length of 36 bp, which is less than HyDRA, which requires a minimum average quality score of 20 and a minimum read length of 100 bp. Rather than a set error rate, HAPHPIPE utilizes a more accurate correction tool with the built in error correction module of SPAdes, BayesHammer [72]. For variant calling, the default minimum variant quality is 15 in HAPHPIPE, again lower than HyDRA. Furthermore, HyDRA assembles data from each read pair separately, as opposed to pairing reads during assembly, and constructs one reconstructed *pol* region as opposed to multiple regions across the entire genome (i.e., *PRRT*, *int*, and *gp120*). HyDRA also does not allow for the assembly of envelope proteins and is restricted to only HIV-1 sequence analyses.

2.3. Geneious

Geneious is a commercial, desktop software that hosts a suite of bioinformatic tools to analyze sequence data (Table 1). For the purposes of this study, we followed the protocol detailed by Dudley et al. [19], which is a reference-based assembly. We paired the raw reads and then trimmed on both the 5′ and 3′ ends with an error probability limit of 0.001 using the modified-Mott algorithm (see Geneious documentation). We then mapped the trimmed reads against the reference sequence HXB2 with the Geneious mapper. Parameters were as follows: a maximum of 15 gaps were allowed per read, a maximum gap size of 15, a minimum overlap identify of 80%, a minimum word length of 14, an index word length of 12, a maximum of 15% mismatches per read, and a maximum ambiguity of 16, and searched more thoroughly for poor match reads. A consensus sequence was saved after mapping with a default base threshold of the highest quality, and for reads without a quality score at a particular base, a default threshold of 65% was used for the consensus sequence. This reference-based workflow in Geneious is similar to the HAPHPIPE reference-based pipeline but sets limits for gaps, does not include an error-correction step, and includes ambiguity codes in the consensus sequence. The default in HAPHPIPE uses the 'fast-sensitive' option in Bowtie2, which allows no mismatches in seed alignment (considerably lower compared to Geneious 15% mismatches per read), requires a seed length of 20 (larger than Geneious value of 12), and requires a seed interval of 1 + 0.50 using the square-root of the read length. For more detailed explanations of Bowtie2 parameters, see the user manual at http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml.

We additionally adapted this reference-based workflow for de novo assembly. The same trimming parameters for quality control steps and the Geneious assembler were used for de novo assembly of contigs using the following parameters: variants with coverage over six were not merged, the merging of homopolymer variants, the production of scaffolds, a maximum of 15 gaps were allowed per read, a maximum gap size of 15, a minimum overlap identify of 80%, a minimum word length of 14, ignored words repeated more than 200 times, an index word length of 12, a reanalysis quality threshold of eight, a maximum of 15% mismatches per read, a maximum ambiguity of 16, and more thorough searching for poor match reads. Additionally, the de novo workflow required additional computational resources, so 32 GB of memory was allocated. The de novo assembly workflow, which is similar to the HAPHPIPE de novo pipeline, uses Geneious's own proprietary de novo assembler to construct contigs, while HAPHPIPE uses SPAdes, which has been shown to produce longer and more accurately assembled contigs [68,73]. It is also fast and does not require as many computational resources as Geneious—requiring only 5 GB of memory for assembly and 8 GB of disk space. This is an important consideration for researchers wishing to conduct large-scale studies with many samples, as HAPHPIPE can also be run efficiently in parallel on an HPC cluster, whereas Geneious must be run in its GUI form. Lastly, in the Geneious de novo workflow, contigs had to be manually mapped back to reference

sequences in order for amplicons to be identified, whereas in HAPHPIPE contigs are automatically scaffolded and labeled, thus reducing manual effort.

2.4. Data Simulation

A total of 100 HIV-1 subtype B genomes and 50 HIV-1 non-subtype B genomes were randomly pulled (with no duplicates) from the Los Alamos HIV Database (LANL; hiv.lanl.gov) reference genome 2017 list (Table S1; [16,28,40,74–136]). For each sequence, all ambiguity codes (M, R, W, etc.) were replaced randomly with one of the corresponding nucleotides (M = A or C, R = A or G, etc.). We then extracted the protease and reverse transcriptase (*PRRT*, HXB2 numbering: 2252–3869), integrase (*int*, HXB2 numbering: 4230–5093), and *gp120* (HXB2 numbering: 6225–7757) gene regions (i.e., amplicons). Gaps were then removed from the each sequence, which was labeled as the "truesequence.fasta" for each sequence. We then simulated reads for each sequence based on the respective "truesequence.fasta" with ART v. MountRainier [137], a simulation tool that generates NGS reads from a consensus sequence. We simulated 150 bp paired-end reads with a 2000x fold coverage, a mean fragment size of 215 bp, and a standard deviation of 120 bp. These reads were also error-prone, implemented with the integrated Illumina MiSeq platform error profile, which means that the reads contained errors known to be caused by the sequencing platform itself, creating a realistic representation of a standard NGS dataset. This process resulted in a total of 25,000 paired-end reads per sequence, covering each of the three targeted gene regions. This procedure was repeated for all 100 subtype B and 50 non-subtype B sequences.

2.5. Analyses and Testing

The simulated FASTQ read files were used as the inputs for both assembly pipelines in HAPHPIPE (haphpipe_assemble_01 and haphpipe_assemble_02) [61]. HyDRA Web v. 1.5.1 and Geneious v. 10.2.6 were implemented using the workflows described previously, each of which resulted in a consensus sequence for each respective true sequence. For Geneious and HyDRA workflows, we additionally extracted amplicon regions by aligning to HXB2. We then aligned all consensus sequences generated from each pipeline and workflow using MAFFT v. 7.309 [138] and estimated genetic diversity with DnaSP v. 6 [139], specifically recording nucleotide diversity (π), Watterson's theta (θ), variable sites, and estimated number of haplotypes. We did the same for the true sequences for purposes of comparison. The genetic distance for each generated consensus sequence with respect to the true sequence was calculated using proportional (p)-distances [140], the number of nucleotide differences per site over the length of the alignment. While consensus sequences constructed with HAPHPIPE did not contain ambiguity codes, those constructed by HyDRA and Geneious workflows contained many. As we wanted to account for these fairly, we also calculated an adjusted p-distance, which gave differences with ambiguity codes fractional weight. P-distances and adjusted p-distances were also calculated between each of the generated consensus sequences and the HXB2 reference sequence.

Results were visualized with R v. 3.6.0 [141] in R Studio v. 1.2.1335 [142] using the ggplot2 v. 3.1.1 [143] package. Hypothesis testing for genetic p-distance and adjusted p-distance results were performed using the non-parametric Kruskal–Wallis test [144,145] with the R package stats v 3.6.0 and the Dunn post-hoc test [146] v 1.3.5 for multiple comparisons in R. For the Dunn test, we report p-values adjusted with the Holm method [147]. Hypothesis testing for distance results between the initial and final consensus sequence generated by each HAPHPIPE pipeline was also performed using the Wilcoxon signed-rank test [148].

2.6. Empirical Data Applications

All SRA accessions, each representing empirical HIV-1 NGS amplicon data from two populations sampled repeatedly over one year, were selected from the BioProject accession PRJNA506879 [63]. The average read count per sample was 583,828 reads. We put the raw NGS reads for each of the 36 HIV samples through both HAPHPIPE pipelines, with the same amplicons used in the simulated HIV-1

data above and the HXB2 reference sequence. We generated haplotypes with PredictHaplo. We also processed the raw reads through HyDRA and Geneious in the same manner as the simulated HIV-1 data. We estimated the genetic diversity of the consensus sequences of the HAPHPIPE pipelines, HyDRA and Geneious with DnaSP. Finally, we calculated p-distances and adjusted genetic p-distances between a sequence (predicted haplotypes from both HAPHPIPE pipelines and the consensus sequences from both HAPHPIPE pipelines, HyDRA, and Geneious) and HXB2 reference sequence, using the same analysis steps as described above with the simulated data. Subtyping of all consensus sequences and haplotypes was performed using the REGA HIV subtyping tool [149].

A total of 23 SRA accessions of HCV sequence data were selected from Babcock et al. [64] (accession numbers: SRR1170557, SRR1170560-SRR1170568, SRR1170576-SRR1170579, SRR1170671-SRR1170679), each representing HCV viral variants from a patient cohort at several time points spanning two months. The average read count was 7.9 million paired-end reads per sample, and we analyzed these samples through the same pipelines as the empirical HIV-1 data above, except HyDRA, which is HIV-1-specific. For reference-based pipelines, we used the H77 HCV reference sequence (accession: NC_004102), which is of HCV subtype 1a [150]. The following amplicons were included: *core* (H77 numbering: 342–914), *E1* (H77 numbering: 915–1490), and *E2* (H77 numbering: 1491–2579). We also reconstructed haplotypes for this dataset, estimated genetic diversity for consensus sequences from each pipeline, and calculated the corresponding genetic distances (p-distance and adjusted p-distance) between each of the resulting pipeline consensus sequences and H77. Subtyping of all consensus sequences and haplotypes was performed using the Genome Detective HCV subtyping tool (www.genomedetective.com/app/typingtool/hcv/).

At the time of this study, four high-quality SARS-CoV-2 SRA samples were available (accession numbers: SRR11140744, SRR11140746, SRR11140748, SRR11140750), and we analyzed each through both HAPHPIPE pipelines and Geneious workflows. The average read count was 395,407 paired-end reads for the first three samples, while the last sample was smaller with only 17,208 reads. We used the severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 (accession NC_045512) as the reference sequence [151]. We sought to test whole genome assembly, so the entire genome was used (i.e., no amplicons were assembled for this dataset); this meant that, for the de novo assembly pipeline for HAPHPIPE (haphpipe_assemble_01), we used the numbering 0 to 29,902 for the reference GTF file. Again, we analyzed assembly statistics, genetic distance from reference metrics, and diversity estimates, as with the previously described datasets.

3. Results

For simplicity, we present and discuss results from adjusted p-distances, and when results from non-adjusted p-distances differ significantly, we addressed these inconsistencies within each section. All genetic p-distance results and the associated figures can be found in Supplemental Materials.

3.1. Simulated Data

In the simulated data, the HAPHPIPE de novo pipeline did not proceed to a third refinement step (Table 2), which indicates that refining a second time did not improve the assembly. The reference-based pipeline, 93% and 88% of the simulated subtype B and non-subtype B samples, respectively, terminated at a third refinement step and only 2% and 4%, respectively, required a fourth refinement step (Table 2). The HAPHPIPE de novo pipeline produced >96% alignment rates, while refinement during the reference-based pipeline improved alignment rates further by an average of 6.72–21.1% (Table 2). Geneious produced lower mapping rates—82.63% and 63.05% for reference-based, and 64.20% and 64.30% for de novo workflows for subtype B and non-subtype B samples, respectively (Table 2). In subtype B samples, there was no significant difference in genetic distance from the true sequence between initial and final steps of the de novo pipeline (p value in the range [0.371, 1]) and distance from the true sequence decreased significantly after refinement for all genes except *int* in the reference-based pipeline ($p < 0.001$; Figure 2, Table A1). For non-subtype B samples, genetic distance from the true

sequence between the initial and final steps decreased significantly for all genes except *int* (for which sequences were already extremely close to the true sequence) in the de novo pipeline ($p < 0.001$), and increased significantly in the reference-based pipeline for all genes ($p < 0.001$; Figure 2, Table A1).

Table 2. Comparison of the de novo and reference-based assembly pipelines in HAPHPIPE for the simulated dataset.

Tool	N	Avg Bowtie2 Alignment First Step	Avg Bowtie2 Alignment Last Step	Number of Samples Refine.02	Number of Samples Refine.03	Number of Samples Refine.04	Number of Samples Refine.05
Simulated HIV data subtype B							
HP01	100	99.87%	99.64%	3	0	0	0
HP02	100	88.95%	95.67%	100	93	2	0
GDN	100	NA	64.20%	NA	NA	NA	NA
GRB	100	NA	82.63%	NA	NA	NA	NA
Simulated HIV data non-subtype B							
HP01	50	98.80%	96.92%	50	0	0	0
HP02	50	59.37%	80.47%	50	44	2	0
GDN	50	NA	64.30%	NA	NA	NA	NA
GRB	50	NA	63.05%	NA	NA	NA	NA

Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, N = number of sequences.

The differences in the genetic distance among true and consensus sequences generated from all platforms were significant ($p < 0.001$, Table S2), and post-hoc analysis indicated many significant pairwise comparisons (Figure 2, Table A2). Overall, for the simulation data, HAPHPIPE pipelines generated consensus sequences significantly closer to the true sequence than those produced by HyDRA ($p < 0.001$; Table A2) for all amplicons (*PRRT*, *int* and *pol*) in both subtype B and non-B sequences. Similarly, both HAPHPIPE pipelines generated consensus sequences significantly closer to the true sequence than Geneious reference-based workflow (all $p < 0.001$, subtype B for *gp120*: $p < 0.05$; Table A2). For subtype B sequences, HAPHPIPE pipelines generated consensus sequences significantly closer to the true sequence than those produced by both workflows in Geneious for *pol*, *PRRT*, and *int* ($p < 0.001$; Table A2). For non-subtype B sequences, HAPHPIPE generated consensus sequences significantly closer to the true sequence than those produced by the Geneious de novo workflow for *pol*, *PRRT*, and *int* ($p < 0.001$, except for Geneious de novo workflow vs. HAPHPIPE reference-based pipeline in *PRRT*: $p < 0.05$; Table A2). For both subtype B and non-subtype B sequences, in *gp120* for de novo assembly, there was no significant difference in distance from the true sequence between HAPHPIPE and Geneious consensus sequences ($p = 0.427$, 0.121; Figure 2, Table A2).

For subtype B sequences, there were no significant differences between distance to HXB2 before and after refinement in the HAPHPIPE de novo pipeline for all genes and in the reference-based pipeline for *pol* genes (p value in the range [0.167, 1]; Figure 3, Table A3). However, *gp120* reference-based consensus sequences were significantly closer to HXB2 post-refinement ($p < 0.001$; Table A3). In non-subtype B data, HAPHPIPE de novo consensus sequences were significantly further from HXB2 post-refinement across all genes ($p < 0.01$), however, reference-based *gp120* consensus sequences showed the opposite and were closer to HXB2 post-refinement ($p < 0.001$; Table A3). For *pol* genes, there were no significant differences in distance to HXB2 post-refinement in the reference-based consensus sequences (p value in the range [0.161, 1]; Table A3). Similar trends as those for the true sequence were seen in genetic distance to the HIV reference sequence, HXB2, for subtype B *pol* data; however, not for subtype B *gp120* or in general for non-subtype B data, which are less similar to HXB2 because of high variability and difference in subtype, respectively. Median distance to HXB2 was notably greater than distance to the true sequence in all genes for *gp120* sequences in the simulation dataset (Figure 3). Moreover, there were no significant differences between the genetic distance to HXB2 for consensus sequences from the HAPHPIPE pipelines and Geneious workflows for subtype B *gp120* sequences ($p = 0.0799$;

Table A4, Table S2). In non-subtype B data, there was no significant difference in the distance of *gp120* HAPHPIPE de novo and reference-based consensus sequences to HXB2 ($p = 0.191$; Table A4), while the same de novo consensus sequences were shown to be significantly closer to the true sequence (Table A2). Additionally, there was no significant difference in *PRRT* sequences between distance of the two HAPHPIPE consensus sequences ($p = 0.774$; Table A4) or between Geneious and HAPHPIPE reference-based sequences ($p = 0.090$) to HXB2 as there was in distance to the true sequence (Table A2). There was also no significant difference in *pol* in distance to HXB2 between the two HAPHPIPE pipelines ($p = 0.879$; Table A4), as was also seen in distance to the true sequence (Table A2).

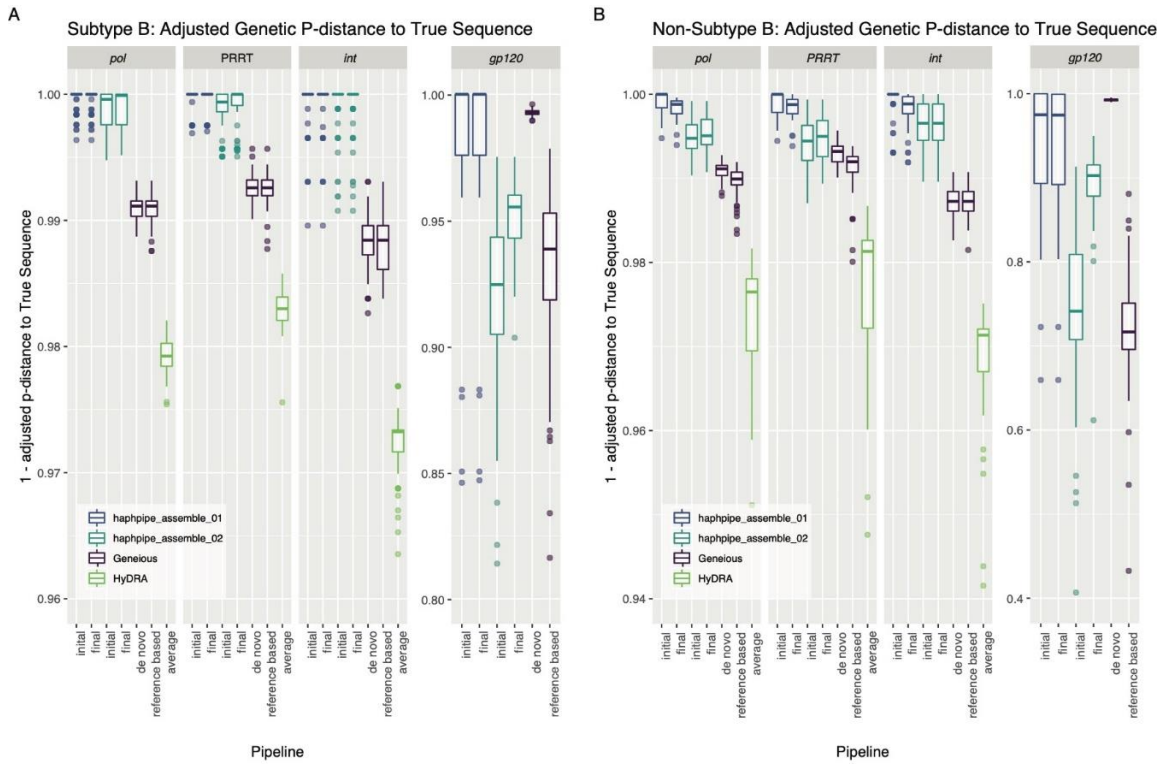

Figure 2. Adjusted genetic p-distance (displayed as a difference from 1) between consensus sequence and true sequence for all pipelines for the simulated HIV (**A**) subtype B dataset and (**B**) non-subtype B dataset. Ambiguous nucleotides were accounted for by giving fractional weight in alignment. A value closer to 1.00 indicates the consensus sequence is more genetically similar to the true sequence. The *x*-axis order from left to right for an individual panel: adjusted genetic p-distance between the true sequence and (i) the initial assembled sequence followed by (ii) the final assemble sequence for haphpipe_assemble_01 pipeline (de novo assembly); (iii) the initial assembled sequence followed by (iv) the final assemble sequence for haphpipe_assemble_02 pipeline (reference-based assembly); the final consensus sequence for the Geneious (v) de novo workflow and the (vi) reference-based workflow; and finally, the (vi) average between the final two sequences (one for each read file) for HyDRA. The three amplicons are shown, as well as a combination of *PRRT* and *int* amplicons into *pol*. There are no results for HyDRA in the *gp120* gene because HyDRA only analyzes the *pol* gene.

As for genetic diversity, consensus sequences from HAPHPIPE (either pipeline) and Geneious reference-based workflow resulted in the greatest underestimations of nucleotide diversity (π) in *gp120*, which represents current diversity estimates [152], compared to estimates from the true sequence for both subtype B and non-subtype B sequences (Figure 4, Table A5). Reference-based assembly, in general, resulted in greater underestimations of π for non-subtype B sequences (Figure 4, Table A5). Similarly, all Geneious and HAPHPIPE consensus sequences resulted in underestimations of Watterson's theta (θ), which represents historical diversity [152], for *gp120* in the non-subtype B sequences (Figure 4,

Table A5). However, all of these differences were quite small, with a magnitude less than 0.08 for π and less than 0.04 for θ.

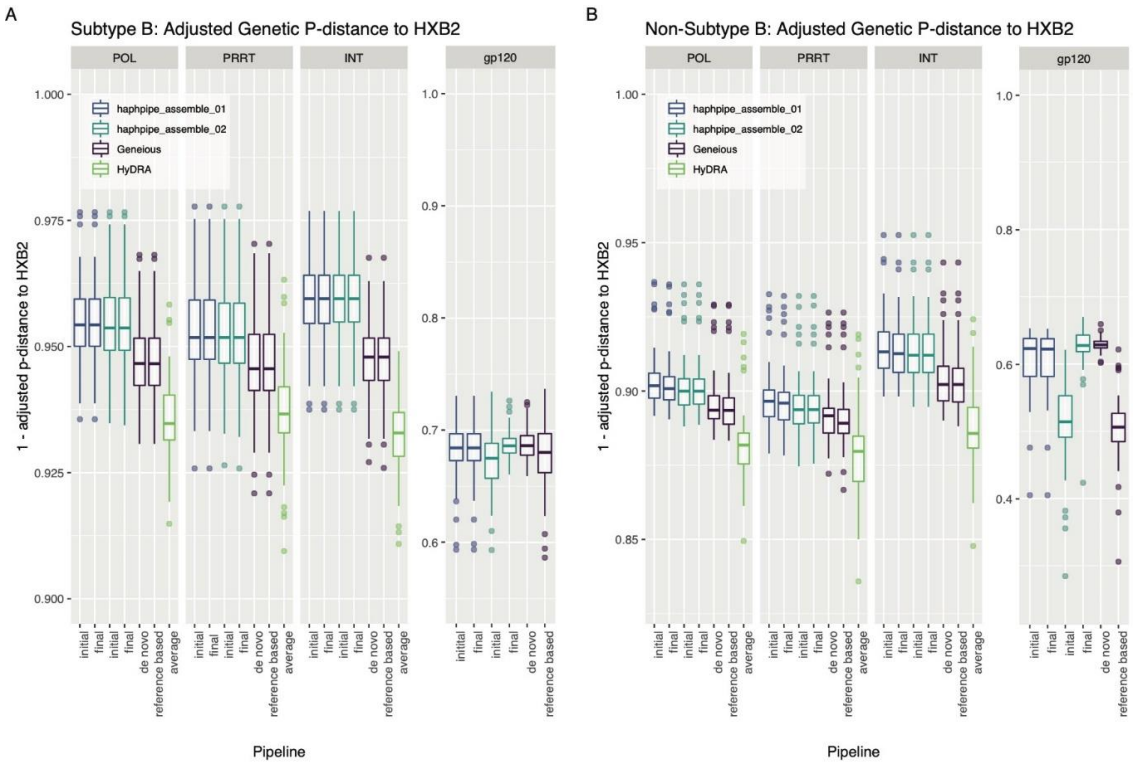

Figure 3. Adjusted genetic p-distance (displayed as a difference from 1) between consensus sequence and HXB2, the reference sequence for HIV, for all pipelines for the simulated HIV (**A**) subtype B dataset and (**B**) non-subtype B dataset. Ambiguous nucleotides were accounted for by giving fractional weight in alignment. A value closer to 1.00 indicates that the consensus sequence is more genetically similar to the reference sequence. The *x*-axis order from left to right for an individual panel: adjusted genetic p-distance between the reference sequence and (i) the initial assembled sequence followed by (ii) the final assemble sequence for haphpipe_assemble_01 pipeline (de novo assembly); (iii) the initial assembled sequence followed by (iv) the final assemble sequence for haphpipe_assemble_02 pipeline (reference-based assembly); the final consensus sequence for the Geneious (v) de novo workflow and the (vi) reference-based workflow; and finally, the (vi) average between the final two sequences (one for each read file) for HyDRA. The three amplicons are shown, as well as a combination of *PRRT* and *int* amplicons into *pol*. There are no results for HyDRA in the *gp120* gene because HyDRA only analyzes the *pol* gene.

The results for adjusted p-distance differed from those of non-adjusted p-distance in some comparisons (Tables S2–S10). Namely, using non-adjusted p-distance to the true sequence, we showed no significant difference between HAPHPIPE and Geneious reference-based sequences in *gp120* for both subtype B and non-B (Figure S1, Table S4). In distance from HXB2 for subtype B sequences, we did find significant differences between methods ($p < 0.001$; Figure S2, Table S6), in particular that HAPHPIPE and both Geneious workflows were closer to HXB2 than HAPHPIPE reference-based sequences and that HAPHPIPE and Geneious de novo sequences were closer to HXB2 than Geneious reference-based sequences ($p < 0.05$; Table S6). In non-subtype B data, non-adjusted p-distance indicated that, in *gp120*, HAPHPIPE and Geneious de novo sequences were closer to HXB2 than HAPHPIPE reference-based sequences ($p < 0.001$), and that there was no difference between distance in Geneious and HAPHPIPE reference-based sequences ($p = 0.541$; Table S6).

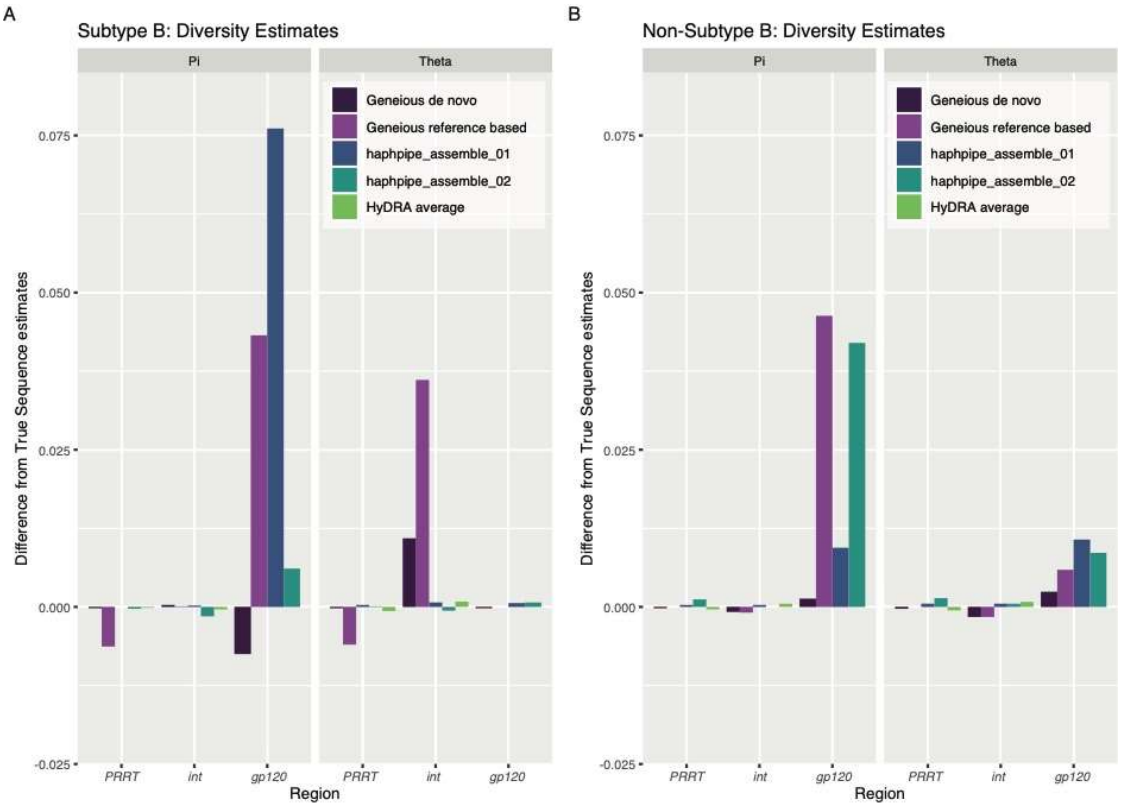

Figure 4. Difference between the estimated genetic diversity from the true sequence and each pipeline (calculated as estimate of true sequences—estimate of pipeline consensus sequences) for the simulated HIV (**A**) subtype B dataset and (**B**) non-subtype B dataset. Positive value indicates an underestimation of the genetic diversity with the consensus sequences from the pipeline, and a negative value indicates an overestimation of the genetic diversity with the consensus sequences from the pipeline. *PRRT* = protease and reverse transcriptase, *int* = integrase, *gp120* = gene within envelope gene region, Pi = nucleotide diversity, Theta = Watterson's genetic diversity.

3.2. Empirical Data

Empirical HCV data were subsampled to 3 million reads per FASTQ file prior to assembly on all platforms due to memory limitations. For the Geneious de novo workflow, data were additionally subsampled to 100,000 reads per FASTQ file because at larger file sizes the assembly step failed to complete. Similarly, the empirical SARS-CoV-2 dataset was subsampled to 100,000 reads per FASTQ file for both Geneious workflows. However, both HAPHPIPE pipelines were able to run on the full dataset. No subsampling was necessary for the empirical HIV data on any platform.

In the HAPHPIPE pipelines, the majority of samples in all empirical data ceased at three refinement steps (Table 3). While the Geneious de novo alignment rates were much higher than those of both HAPHPIPE pipelines and the Geneious reference-based workflow for HIV and HCV data, these values reflect the percentage of reads mapped to contigs, which were then scaffolded to the reference sequence (Table 3). Only 27.78%, 29.07%, and 13.11% of these contigs mapped back to the reference for HIV, HCV, and Sars-CoV-2 data, respectively, and the number of reads included in the final scaffolded sequence is not reported. Contig mapping rates are not available for the HAPHPIPE pipelines because contigs are only used to build a scaffold—further refinement steps utilize reads directly instead of contigs. For the empirical HIV data, sequencing covered the entire genome as a set of five amplicons [63], while our assembly targeted only three genes as distinct amplicons. In the following subsections, we compare the effects of assembly methods among all platforms, as well as the implications of these assembly-related data, on the final consensus sequence.

Table 3. Comparison of the effect of consensus generation on estimated genetic diversity across the empirical datasets.

Tool	Num of Seqs	Avg Bowtie2 Alignment First Step	Avg Bowtie2 Alignment Last Step	Number of Samples Refine.02	Number of Samples Refine.03	Number of Samples Refine.04	Number of Samples Refine.05
			Empirical HIV Data				
HP01	36	53.67%	44.23%	36	10	0	0
HP02	36	54.69%	47.58%	36	5	0	0
GDN	36	NA	89.98%	NA	NA	NA	NA
GRB	36	NA	27.68%	NA	NA	NA	NA
			Empirical HCV Data				
HP01	23	90.46%	72.40%	23	19	7	1
HP02	23	78.13%	75.27%	23	22	4	1
GDN *^	23	NA	97.27%	NA	NA	NA	NA
GRB	23	NA	63.31%	NA	NA	NA	NA
			Empirical SARS-CoV-2 Data				
HP01	4	100%	94.32%	4	3	0	0
HP02	4	100%	94.36%	4	4	0	0
GDN *^	4	NA	80.55%	NA	NA	NA	NA
GRB *	4	NA	94.23%	NA	NA	NA	NA

* Total reads had to be subsampled to 100,000 reads per FASTQ file for Geneious to produce results. ^ Alignment rates are for reads mapped to contigs. Not all contigs were scaffolded to the reference: 27.78% of contigs for HIV, 29.07% of contigs for HCV, and 13.11% of contigs for SARS-CoV-2 data were used after scaffolding. Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, N = number of sequences.

3.2.1. Empirical HIV Dataset

All consensus sequences and reconstructed haplotypes generated from the empirical HIV data were confirmed as Subtype B by the REGA subtyping tool. HAPHPIPE de novo sequences were significantly closer to HXB2 after refinement in all genes ($p < 0.001$; Figure 5, Table A6); while for HAPHPIPE reference-based sequences, only *gp120* was significantly closer to the reference genome ($p = 0.003$; Table A6). Consensus sequences produced for each gene in the empirical HIV dataset showed significant differences in genetic distance to the HIV subtype B reference sequence, HXB2 ($p < 0.001$; Table A7). Notably, HAPHPIPE reference-based consensus sequences were significantly closer to HXB2 than de novo consensus sequences across all gene regions ($p < 0.001$; Figure 5, Table A7); unlike in the simulated data, which showed similar results between both HAPHPIPE pipelines (Figure 3, Table A3). In fact, HAPHPIPE de novo sequences were significantly farther from HXB2 than any other pipeline in all three genes, despite the sample data also being subtype B. For Geneious, no significant differences were found between de novo and reference-based sequences in any gene (p value in the range [0.630, 1.00]; Table A7). There were also no significant differences found between any of the three reference-based pipelines in *pol* genes (*PRRT* and *int* (p value in the range [.251, 1.00]), yet in *gp120* HAPHPIPE, reference-based sequences were significantly closer to HXB2 than those from Geneious (Table A7). In HAPHPIPE, reference-based sequences for *PRRT* and *gp120*, reconstructed haplotypes were significantly farther from HXB2 than the HAPHPIPE reference-based consensus sequences ($p < 0.05$), as were de novo *gp120* sequences ($p < 0.05$). Haplotypes for de novo *PRRT* sequences and *int* sequences from both pipelines showed no significant difference in distance from HXB2 (Table A7). In all three gene regions, haplotypes had considerably more variable sites than consensus sequences (Table 4). Compared to consensus sequences, haplotypes also had higher values of Watterson's theta in all genes and higher values of pi in *PRRT* and *gp120* (Table 4). HyDRA sequences had the lowest values of both pi and theta for both *pol* genes (Table 4). Both diversity metrics were also higher for HAPHPIPE pipelines than their Geneious counterparts, in both de novo and reference-based workflows (Table 4).

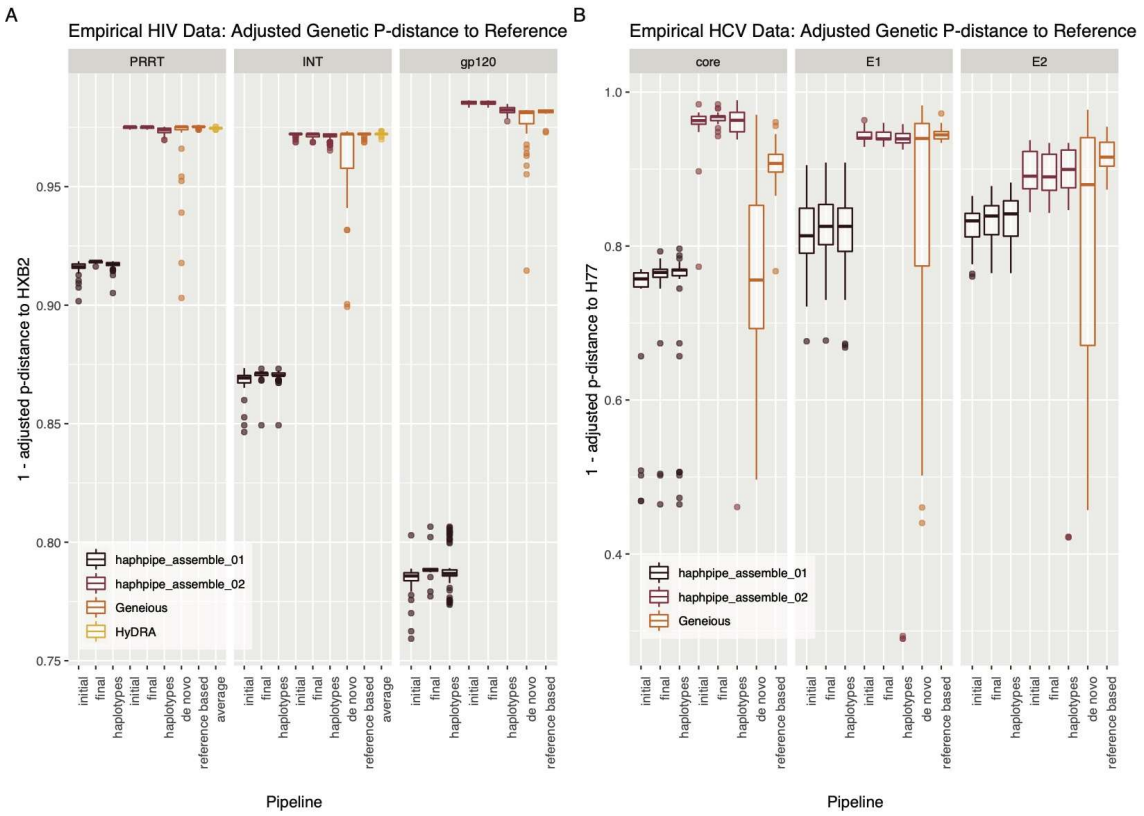

Figure 5. Adjusted genetic p-distance (displayed as a difference from 1) between consensus sequence and HXB2, the reference sequence for HIV, for all pipelines for the empirical (**A**) HIV dataset and (**B**) HCV dataset. Ambiguous nucleotides were accounted for by giving fractional weight in alignment. A value closer to 1.00 indicates that the consensus sequence is more genetically similar to the reference sequence. The y-axes are different for each HIV and HCV, with HCV showing greater variance between samples. The *x*-axis order from left to right for an individual panel: adjusted genetic p-distance between the reference sequence and (i) the initial assembled sequence, (ii) the final assemble sequence and (iii) the reconstructed haplotypes for haphpipe_assemble_01 pipeline (de novo assembly); (iv) the initial assembled sequence, (v) the final assemble sequence, and (vi) the reconstructed haplotypes for haphpipe_assemble_02 pipeline (reference-based assembly); the final consensus sequence for the Geneious (vii) de novo workflow and (vi) reference-based workflow; and finally, the (viii) average between the final two sequences (one for each read file) for HyDRA. The three amplicons are shown for both empirical datasets (HIV: *PRRT*, *int*, *gp120* and HCV: *core*, *E1*, *E2*). There are no results for HyDRA in the *gp120* gene for HIV or for any HCV genes because HyDRA only analyzes the *pol* gene region of HIV.

Table 4. Comparison of the effect of consensus generation on estimated genetic diversity across the empirical HIV dataset.

	PRRT					***int***					***gp120***				
Pipeline	**N**	**H**	**S**	**π**	**θ**	**N**	**H**	**S**	**π**	**θ**	**N**	**H**	**S**	**π**	**θ**
HP01	36	7	6	0.0005	0.0008	36	5	5	0.0009	0.0013	36	9	10	0.0012	0.0015
HP02	36	6	6	0.0005	0.0009	36	5	5	0.0009	0.0014	36	10	13	0.0016	0.0016
HP01 haplotypes	132	78	119	0.0028	0.0127	162	52	35	0.0021	0.0064	224	167	149	0.0049	0.0154
HP02 haplotypes	139	85	110	0.0034	0.0123	140	46	34	0.0021	0.0071	70	40	61	0.0026	0.0067
HyDRA	36	1	2	0.0003	0.0003	36	1	2	0.0003	0.0006	NA	NA	NA	NA	NA
GRB	36	1	3	0.0004	0.0005	36	1	3	0.0004	0.0009	36	1	5	0.0006	0.0007
GDN	36	2	4	0.0005	0.0006	36	2	4	0.0007	0.0011	36	3	9	0.0009	0.0011

HyDRA is an average between read1 and read2. Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, N = number of sequences, H = number of haplotypes, S = number of polymorphic sites, π = nucleotide diversity, θ = Watterson's genetic diversity, *PRRT* = protease and reverse transcriptase, *int* = integrase.

3.2.2. Empirical HCV Dataset

Notably, subtyping results from the Genome Detective HCV subtyping tool indicate that 42.2% of consensus sequences could not be assigned to a subtype, including sequences from each sample and from each platform and assembly method. The remaining sequences were assigned to subtype 1a, the same as the H77 reference sequence. This indicates that the data are likely not purely type 1a and may include either significant mutations or recombination with other HCV subtypes. HAPHPIPE de novo sequences were significantly closer to H77 after refinement for all genes ($p < 0.001$), as were HAPHPIPE reference-based sequences for *core* and *E2* ($p < 0.01$), but no significant difference was seen in *E1* ($p = 0.058$; Figure 5, Table A8). For all three genes (*core*, *E1*, and *E2*), HAPHPIPE reference-based consensus sequences were closer to H77 than HAPHPIPE de novo consensus sequences ($p < 0.001$; Figure 5, Table A8). HAPHPIPE de novo sequences were significantly closer to H77 than Geneious de novo sequences in *core* and farther in *E1* ($p < 0.01$), though this difference was not significant in *E2* ($p = 0.069$; Figure 5, Table A8). Geneious de novo sequences were significantly farther from H77 than Geneious reference-based sequences in both *core* and *E2* ($p < 0.05$), but not in *E1* ($p = 0.201$, Figure 5, Table A8). HAPHPIPE reference-based consensus sequences were significantly closer to H77 than all other pipelines in *core* ($p < 0.01$) and showed no significant differences compared to either Geneious workflow in the envelope genes (*E1*, *E2*; *p* value in the range [0.369, 1]; Figure 5, Table A8). No significant differences in distance from H77 were seen between consensus sequences and their respective haplotypes for either HAPHPIPE pipeline in all genes (*p* value in the range [0.702,1.00], Table A8). The number of variable sites in haplotypes versus consensus sequences was highly variable, with no consistent trends seen across genes (Table 5). Overall, diversity estimates for Geneious de novo sequences were higher than the other constructed consensus sequences, which could be related to the extreme subsampling for this dataset (Table 5).

Table 5. Comparison of the effect of consensus generation on estimated genetic diversity across the empirical HCV dataset.

	core					*E1*					*E2*				
Pipeline	**N**	**H**	**S**	**π**	**θ**	**N**	**H**	**S**	**π**	**θ**	**N**	**H**	**S**	**π**	**θ**
HP01	23	16	33	0.0360	0.0379	23	21	188	0.0814	0.0793	23	20	406	0.1047	0.0960
HP02	23	16	26	0.0373	0.0366	23	20	151	0.0703	0.0707	23	23	380	0.1028	0.0960
HP01 haplotypes	63	32	46	0.0328	0.0407	48	36	216	0.0792	0.0739	61	53	462	0.1021	0.0845
HP02 haplotypes	67	30	114	0.0619	0.1241	49	32	228	0.0812	0.0842	56	44	615	0.1213	0.1209
GRB	23	2	377	0.4561	0.3493	23	5	167	0.0736	0.0939	23	3	482	0.1236	0.1484
GDN	23	14	55	0.0826	0.0746	23	16	146	0.0678	0.0691	23	22	389	0.1051	0.0991

Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, N = number of sequences, H = number of haplotypes, S = number of polymorphic sites, π = nucleotide diversity, θ = Watterson's genetic diversity.

3.2.3. Empirical SARS-CoV-2 Dataset

Genetic distance from the reference sequence, Wuhan-Hu-1, was low across all four samples and pipelines, ranging from 0.00097 to 0.10384 (Table 6). Geneious de novo consensus sequences, though, showed the highest average genetic distance from the Wuhan-Hu-1 reference sequence, as well as the highest standard deviation in distance, indicating that these results are uniquely far from the reference sequence—even when compared to HAPHPIPE de novo consensus sequences—and that the results are most variable, likely as a result of the extreme subsampling of reads for the Geneious platform (Table 6).

Table 6. Comparison of the effect of consensus generation on estimated genetic diversity and adjusted genetic p-distance to the reference sequence across the empirical SARS-CoV-2 dataset.

	Diversity Estimates					Adjusted Genetic p-Distance		
Pipeline	**N**	**H**	**S**	**π**	**θ**		**Average**	**STDEV**
HP01	4	1	0	0	0	**HP01 Initial**	0.0033	0.0035
HP02	4	1	0	0	0	**HP01 Final**	0.0035	0.0034
GDN	4	4	371	0.0081	0.0083	**HP02 Initial**	0.0019	0.0015
GRB	4	1	13	0.0004	0.0004	**HP02 Final**	0.0021	0.0019
						GDN	0.0417	0.0423
						GRB	0.0031	0.0008

Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, N = number of sequences, H = number of haplotypes, S = number of polymorphic sites, π = nucleotide diversity, θ = Watterson's genetic diversity, STDEV = standard deviation.

3.2.4. Differences between Genetic Distance Measurements

There were few differences in results based on non-adjusted p-distance in the empirical data. Non-adjusted p-distance results indicate that for *PRRT*, both HAPHPIPE and Geneious reference-based consensus sequences were closer to HXB2 than those from HyDRA ($p < 0.05$); HAPHPIPE reference-based sequences were also closer to HXB2 than Geneious de novo sequences (Figure S3, Table S7). Additionally, the difference between de novo *gp120* sequences and haplotypes was not significant ($p = 0.057$; Table S7). In the HCV *core*, HAPHPIPE de novo sequences were closer to H77 than Geneious de novo ($p < 0.05$), while in *E1* both HAPHPIPE and Geneious reference-based sequences were closer to H77 than Geneious de novo ($p < 0.001$), and there was no difference between HAPHPIPE and Geneious de novo ($p = 0.573$). In *E2,* HAPHPIPE reference-based sequences were closer to H77 than Geneious de novo ($p < 0.001$; Table S8).

4. Discussion

In this study, we benchmarked the performance of HAPHPIPE consensus sequence assembly with two commonly used alternatives, HyDRA and Geneious. We found that in the simulation study, HAPHPIPE performed better than HyDRA, and while it performed equally well or better than Geneious, it can, in addition, handle larger datasets with greater speed. We also validated the performance on HAPHPIPE with empirical data. In all analyses, we addressed genetic distance from true sequences and/or reference sequences, genetic diversity, and assembly statistics, and in the real data we additionally constructed haplotypes using HAPHPIPE's PredictHaplo implementation. When we discuss genetic distance, we refer to adjusted genetic p-distance, which takes into account ambiguity codes, unless otherwise specified. Lastly, we place our software into context with other NGS assembly software.

4.1. Simulated Data

For the simulated dataset, mapping rates were higher in both HAPHPIPE pipelines than both Geneious workflows. We also found that often two refinement steps were efficient in creating a better, more representative consensus sequence. In subtype B simulated data, the iterative refinement step in HAPHPIPE resulted in final sequences that were closer to the true sequence than the initial consensus for most genes in reference-based assembly. In de novo assembly, initial consensus sequences were already very close to the true sequence before refinement, and so no significant difference in distance was seen after this step. In non-subtype B data, sequences in reference-based assembly were significantly closer to the true sequence after refinement, yet sequences in de novo assembly were significantly farther; this is likely due to the difference in subtype between the consensus and the

reference. The use of subtype-specific references would be necessary to evaluate the extent that these differences impact performance.

De novo assembly with HAPHPIPE produced a consensus sequence that is significantly more genetically similar to the true sample isolate sequence compared to reference-based assembly with HyDRA or Geneious. This was especially notable in non-subtype B HIV-1 *pol* sequences, which are most commonly targeted for clinical applications. There was no significant difference in genetic distance for the simulated HIV datasets between the HAPHPIPE pipelines, except for the *gp120* amplicon, in which de novo assembly constructed a consensus sequence genetically closer to the true sequence. De novo assembly for *gp120* with Geneious and HAPHPIPE constructed similar consensus sequences and provided a more accurate representation of the true sequence overall compared to reference-based assembly pipelines. This suggests that for more diverse genes such as envelope proteins, using either tool's de novo option would result in similarly constructed consensus sequences that are more likely to be accurate than reference-based counterparts. Adjusted and non-adjusted genetic p-distance results to the true sequence were not identical, with non-adjusted genetic p-distance having shown no significant difference between HAPHPIPE and Geneious reference-based sequences in *gp120* for both subtypes. We discuss genetic distance to HXB2 below in Section 4.3 (bias towards reference sequence).

4.2. Empirical Data

4.2.1. Assembly Statistics and Performance

Two main aspects of assembly, which varied across platforms for both de novo and reference-based assembly, were feasible input data size and read mapping rates. Geneious workflows required extensive subsampling that HAPHPIPE pipelines did not. These subsampling steps undoubtedly affected assembly statistics and the resulting consensus sequences, and the inability of Geneious to complete assembly with higher than 100,000 reads limits its comparison to HAPHPIPE. For example, the higher read mapping rate of Geneious de novo assembly in the empirical HCV data may be artificially high compared to HAPHPIPE, as far fewer reads were available to begin with and less than a third of assembled contigs successfully scaffolded to the reference sequence.

Additionally, the empirical HIV data covered the entire genome as a set of five amplicons, while here we constructed only three. Thus, the relatively low mapping rates reflect that not all sequencing reads mapped to the three genes (*PRRT*, *int*, and *gp120*), and that the high mapping rate of reads in Geneious de novo assembly could indicate incorrect mapping of reads from elsewhere in the genome to these three gene regions. Similar to HCV, the mapping rates across platforms for HIV reflected high rates for Geneious compared to HAPHPIPE. While not an effect of subsampling, this result is limited by less than a third of assembled contigs scaffolding to the reference and may also indicate incorrect mapping of reads into amplicons. In SARS-CoV-2 assembly, where the entire genome was used and sequences differed much less from the reference, HAPHPIPE mapping rates were higher, supporting that these factors may have contributed to the opposite pattern seen in the other datasets.

It is also notable that mapping rates were lower after the refinement step for all empirical datasets. This effect is most pronounced in the HAPHPIPE de novo pipeline. However, in the simulation dataset, mapping rates either remained the same post-refinement or increased, particularly in the reference-based pipeline. It is possible that the decrease in mapping rates post-refinement in the empirical data is due to the initial incorrect mapping of reads to genome regions, in which case removing these reads at the refinement step helped to improve the specificity of the alignment. This effect would not be seen in the simulated data, as all reads were derived from the specific amplicons used. Further, the increase in the mapping rates of simulated data for the reference-based pipeline may be due to incorrect initial mapping to the reference sequence during assembly, which was then corrected in the refinement step, leading to a higher overall mapping rate.

4.2.2. HIV

In the empirical HIV data, HAPHPIPE de novo consensus sequences were consistently farther from the reference sequence, HXB2, than all others, including HyDRA. In HAPHPIPE, the refinement of de novo sequences decreased distance from HXB2, while in reference-based sequences this was only seen in *gp120*. In general, the reconstructed haplotypes were significantly farther from HXB2 than their respective HAPHPIPE consensus sequences only in *gp120* for the de novo pipeline, and in both *PRRT* and *gp120* for the reference-based pipeline. Haplotypes from *PRRT* and *gp120* in both pipelines also showed increased genetic diversity in three metrics: variable sites, pi, and Watterson's theta. The observation that the haplotypes exhibited more diversity and greater distance from HXB2 than reference-based consensus sequences suggests that these consensus sequences may have been biased towards HXB2. While all consensus sequences were confirmed to be of subtype B, divergence from HXB2 could likely be a result of drug resistance mutations. As such, the greater distance of HAPHPIPE de novo sequences from HXB2 as compared to consensus sequences from other methods may reflect higher accuracy to the "true" sequence, although this hypothesis cannot be directly tested, highlighting a key limitation in the analysis of empirical data for this purpose.

4.2.3. HCV

Results from the empirical HCV data were considerably more variable than those from the empirical HIV data. Pairwise differences in distance from H77 across platforms yielded inconsistent results, particularly in envelope genes *(E1* and *E2*). In contrast to the empirical HIV dataset, haplotypes were not significantly more or less distant from H77 than the respective HAPHPIPE consensus sequences in any gene. While haplotypes again showed slightly higher metrics of genetic diversity, overall these metrics were highly variable. Additionally, the Geneious reference-based workflow showed the highest pi and Watterson's theta for the *core* gene by an order of magnitude larger compared to the others, including haplotypes. Two main confounders in these data are subsampling of Geneious de novo data to 100,000 reads and inconsistency in subtype, with over 40% of consensus sequences obtained—from each sample and each method—being of indeterminate type. This limits the use of H77 as an appropriate standard of comparison. Thus, more definitive conclusions from the empirical HCV data cannot be made.

4.2.4. SARS-CoV-2

For both HAPHPIPE and Geneious platforms, the distance of assembled SARS-CoV-2 sequences from the reference, Wuhan-Hu-1, was quite low. Geneious de novo consensus sequences were the most variable in distance from the reference and, on average, were farther from the reference. This effect was not seen in HAPHPIPE de novo sequences, implying that it may be due to the subsampling of SARS-CoV-2 data to 100,000 reads for Geneious. A greater number of samples would be necessary to investigate this claim, however at the time of this study limited NGS data were publicly available, thus limiting our analysis and potentially contributing to the low distance to reference.

4.2.5. Effects of Ambiguity Codes

Adjusted and non-adjusted genetic p-distance results, with respect to the reference sequences for the empirical datasets, were not identical. In particular, in HIV data, HAPHPIPE and Geneious showed significant differences when using non-adjusted p-distance, but these differences were not significant when using adjusted p-distance. In HCV data, most notable differences occurred in pairwise comparisons involving the Geneious de novo workflow, in which these consensus sequences were further from the reference sequence with non-adjusted p-distance. This observation is not surprising given the increase in ambiguity codes in Geneious de novo workflow, again likely due to the extreme subsampling. In SARS-CoV-2 data, there were no differences in non-adjusted p-distance and adjusted p-distance for HAPHPIPE consensus sequences, while in both Geneious workflows, adjusted p-distance

was higher, again due to the inclusion of ambiguity codes in Geneious consensus sequences, likely compounded with the subsampling as well.

4.3. Bias towards Reference Sequence

In this study, we used HXB2 consistently for all HIV subtypes for purposes of comparison with HyDRA, which only uses HXB2 and does not allow the user to change the reference sequence. The refinement of HAPHPIPE de novo consensus sequences resulted in greater distance from HXB2 in the non-subtype B simulated dataset. Reference-based sequences in *gp120* were closer to HXB2 post-refinement in both datasets. These results suggest potential bias of non-subtype B *pol* sequences and *gp120* sequences—regardless of subtype—to HXB2 in reference-based assembly. Moreover, distances from HXB2 were significantly different to distances from the true sequence in the *gp120* region of subtype B sequences and in all genes except *int* in non-subtype B sequences. This observation suggests that the use of HXB2 as the reference introduced bias to the consensus sequences in variable regions of subtype B sequences (i.e., *gp120*) and in multiple regions of non-subtype B sequences. Placed in this context, it is likely that bias towards HXB2 would be most present in variable regions, regardless of subtype, and that the inconclusive subtype of the empirical HCV data may have introduced nontrivial bias to results. These results further suggest that in practice, subtype-specific reference sequences should be used whenever possible. Although this change in reference is possible with some viral assembly pipelines, it is a key feature that cannot be changed in HyDRA, which is often used in clinical practice to identify drug-resistant mutations. Thus, we support the conclusion that de novo assembly may be better than reference-based assembly for variable regions of the genome (e.g., envelope genes) and when subtype is either mixed or uncertain.

4.4. Limitations

Our study has a few notable limitations. First, we were primarily constrained by the lack of knowledge of "true" sequence compositions for the empirical data. Therefore, we were unable to determine how genetically similar the constructed consensus sequences were to the "true" sequence(s) for each pipeline. However, consensus sequences do not occur in the viral population thus distinguishing them from ancestral sequences [119,153]; therefore, obtaining a "true" consensus sequence is not entirely attainable. Still, consensus sequences are often used in a case of best representative sequence for computational efficiency or, often in clinical applications, for the identification of drug-resistant mutations to guide medical treatment. As such, even if presented with a "true" consensus sequence for comparison, as we have done in the data simulation section, we still compare to an estimated sequence that may or may not reflect true intra-patient diversity, thus limiting the generalization of resulting conclusions. NGS circumvents this limitation by facilitating the reconstruction of viral haplotypes, which have been shown to improve the resolution of phylodynamic inferences [154]. Thus, the implementation of haplotype reconstruction methods in HAPHPIPE facilitates more accurate, informative analyses of both intra- and inter-patient viral diversity and evolution.

Another notable limitation we encountered was the inability of Geneious workflows to be completed on the full set of reads for each of the HCV and SARS-CoV-2 empirical data. Therefore, for both of these datasets, the resulting Geneious consensus sequences were likely skewed compared to the HAPHPIPE pipelines, which did not require as extensive subsampling for HCV data and required no subsampling for HIV or SARS-CoV-2 data. Furthermore, it took each empirical sample at least two days to complete the de novo assembly pipeline on Geneious. We hypothesize that Geneious had trouble with viral NGS data because viruses are fast-evolving and contain many variants, thus it was difficult for the program to orient the reads together for de novo assembly, as in HCV, or even against a reference, as in the case of SARS-CoV-2. We further hypothesize that the large size of both datasets may have posed memory issues during the run, as likely indicated by the increased time to completion, which could impact functionality. This caveat in Geneious further emphasizes the applicability of

HAPHPIPE for assembling viral sequences from NGS data. At the time of this study, limited NGS data were available for SARS-CoV-2 sequences. In particular, only four samples of good quality were publicly available, thus limiting this aspect of our analysis.

4.5. Utility of Software

Geneious and HyDRA both present downsides in practical use. Although Geneious is an interactive software that is easy to use with its graphical user interface (GUI), it does require payment. It is also cumbersome and time consuming to complete assemblies, especially de novo assemblies, on large-scale projects with many NGS samples. We streamlined the process by making workflows, but in doing so, it was time-intensive to rename output files efficiently and distinguish output files. Furthermore, we had trouble, despite allocating ample memory and subsampling reads, with assembly workflows for the empirical datasets. Moreover, each sample processed through the de novo assembly workflow took at least two days to finish and produce a consensus sequence, with some empirical HIV samples taking upwards of seven days. While Geneious does include or have available extensions to phylogenetic tools such as multiple sequence alignment and building trees, the full capabilities of such software may not be available or feasible to be run locally, thus necessitating the use of an additional tool for phylodynamic steps.

Although HyDRA is a free, online-based software, it produced the least accurate results here for HIV-1 in our simulation study. This drop in accuracy could be due to HyDRA's inability to analyze paired-end data together. Moreover, HyDRA only allows for the assembly of polymerase genes for HIV, due to the emphasis on drug-resistant mutations, and does not allow the user to change the reference sequence. Both Geneious and HyDRA constructed consensus sequences with many ambiguity codes, which could be due to intra-host variation or drug-resistant mutations. HyDRA does not include options for phylogenetics steps.

The main advantage of Geneious and HyDRA as compared to HAPHPIPE, particularly for those unfamiliar with UNIX-based command line and bash, is that the former two include an easy-to-use GUI. However, for more advanced users and those analyzing large-scale datasets, this becomes a hindrance to efficiency. Additionally, storing large-scale NGS datasets locally on the user's computer as opposed to remotely on a high-performance cluster, may pose limitations for larger studies in using GUIs. While HAPHPIPE does require knowledge of the command line interface, the example pipelines given (haphpipe_assemble_01 and haphpipe_assemble_02), as well as the thorough documentation and beginner-focused user guide [61], simplify this process for non-bioinformatic users. We have also simplified the installation process for HAPHPIPE by adding it to Bioconda, a popular bioinformatics software repository. The command line interface of HAPHPIPE presents several clear advantages over the GUI-based programs used in this study, namely the ability to run on a high-performance cluster using parallelization techniques, compatibility with bash scripting to automate assembly of many samples at once, and extensibility both for custom pipelines and for additional modules. One further limitation of HAPHPIPE, specifically in clinical and public health applications, is the lack of a module for HIV drug-resistance identification. However, HAPHPIPE is applicable for a variety of viruses and does include general variant calling as well as consensus sequences in formats compatible with existing DRM identification tools, such as Stanford HIVdb [155,156]. HAPHPIPE provides an additional step in analyzing NGS data from intra-host populations by implementing wrappers for estimating haplotypes. By performing haplotype reconstruction, in addition to consensus assembly, our approach can output a more detailed representation of the haplotype diversity in a sample from NGS data. Therefore, HAPHPIPE can be used both to create a more accurate consensus sequence and to capture viral variants within the data by presenting reconstructed haplotype sequences.

4.6. Comparison to other Viral Pipelines

In this study, we compare the performance of HAPHPIPE to that of Geneious and HyDRA, both GUI-based tools that are most commonly used among clinicians and others whose primary

background is not necessarily bioinformatics. However, several additional viral assembly pipelines on the command line exist and may be compared to HAPHPIPE. While a full comparison of the performance of these methods with respect to HAPHPIPE is outside the scope of this validation study, here we discuss three in this context: viral-ngs, MiCall, and V-pipe.

Viral-ngs is an open source, command-line package (https://github.com/broadinstitute/viral-ngs) that utilizes Trinity [157]. De novo assembly with SPAdes is offered as an alternative option. Following de novo assembly, Gap2Seq [158] is used to fill the gaps in the generated scaffold. The remaining steps in the assembly portion of viral-ngs uses reference-based assembly improvements to generate the final consensus sequence. Like HAPHPIPE, specifically the de novo pipeline, viral-ngs uses MUMMER to orient and merge contigs with the assistance of a reference FASTA sequence, however, HAPHPIPE additionally includes an option to construct amplicon-specific sequences at this stage. In the final, major step of viral-ngs assembly, viral-ngs uses Novoalign (http://www.novocraft.com)—which requires a commercial license—to call back reads and align them to the crude de novo assembly, which is then iteratively improved. In both HAPHPIPE and viral-ngs, this step is named refine_assembly and has essentially the same function, although HAPHPIPE allows the user to define the number of desired iterations (with default settings being set at five iterations), while the analogous stage in viral-ngs is set at two iterations. Viral-ngs also implements an additional imputation step as a part of refinement. Lastly, HAPHPIPE and viral-ngs include distinct focuses for downstream analysis, namely phylodynamics versus metagenomics, respectively, and thus may appeal to different user bases.

The reference-based assembly pipeline MiCall (https://github.com/cfe-lab/MiCall) is based on Bowtie2—as are viral-ngs and HAPHPIPE—and has been noted in the literature to be interchangeable with platforms such as HyDRA and PASeq. Like HAPHPIPE, MiCall includes both assembly options. Following quality control measures, the reads are mapped to a reference amplicon. From there, reads are either assembled in relation to a reference or used to create contigs de novo. Then in a stage akin to HAPHPIPE refine_assembly, reads are mapped onto the previously created consensus and improved. Unlike HAPHPIPE, however, this stage is only executed iteratively for reference-based assemblies. As a pipeline tailored to HIV, like HyDRA, MiCall also has a stage, 'resistance', that determines the reads' resistance to antiretroviral therapies (ART). MiCall is more HIV and HCV specific, whereas HAPHPIPE has more applicability across many other viral species.

V-pipe is a publicly available, command line tool (https://cbg-ethz.github.io/V-pipe/) in which reads are mapped to an initial reference to generate a crude alignment. The initial reference can be provided or created de novo from the software VICUNA [159]. The pipeline then uses a Hidden-Markov Model (HMM)-based aligner designed for NGS reads of small genomes that are prone to indels, such as HCV and HIV, (ngshmmalign; https://github.com/cbg-ethz/ngshmmalign). The original reads are then mapped against the profile HMM. The creation of the HMM profile is unique to V-pipe, as is its utilization of VICUNA and ngshmmalign over the established BWA [160] and Bowtie2 wrappers. Like HAPHPIPE, but unlike viral-ngs and MiCall, V-pipe calls haplotypes. Rather than using PredictHaplo, V-pipe implements HaploClique and Savage for global haplotype reconstruction and ShoRAH for local reconstruction, all of which performed poorly relative to PredictHaplo in a recent comparison of haplotype reconstruction tools [65].

5. Conclusions

We found that NGS viral analysis is improved with the use of HAPHPIPE, particularly in conserved regions. Furthermore, we demonstrated that de novo assembly performs better than reference-based assembly at generating a consensus sequence that is closer to the true sequence. Additionally, we further validated the performance of HAPHPIPE across multiple viruses of varying genome lengths, as well as both amplicon and whole genome viral assembly from NGS data. We found that HAPHPIPE facilitated the use of a greater quantity of empirical data and completed assemblies more quickly than other methods, in particular for datasets of viruses with greater genomes (e.g., SARS-CoV-2 whole-genome assembly) and with greater sequencing depth (e.g., the empirical HCV data used here). While in this

study we mainly focused on two commonly used GUI-based tools, we also compared the functionality of our software to other available command-line platforms. A thorough comparative study of the performance of the many terminal-based viral assembly tools and pipelines would be of great value to the research community.

Based on the conclusions of this validation study, we believe that HAPHPIPE provides a more efficient and informative pipeline for the analysis of NGS viral data, particularly for translational clinical and public health research. HAPHPIPE is a single, open-source tool that allows for customization of the analyses, generates a more accurate viral consensus sequence, and produces properly formatted outputs for further phylodynamic analyses, as well as integrates these methods into a unified framework. By including user-friendly wrappers for complex bioinformatics programs, detailed documentation, and a beginner-level User Guide and protocol publication, as well as by maintaining our program as open-source, freely available software, we expect HAPHPIPE to make sophisticated genomics analysis more accessible to researchers across many biomedical fields.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/12/7/758/s1, Figure S1: Genetic p-distance (displayed as a difference from 1) between consensus sequence and true sequence for all pipelines for the simulated HIV A) subtype B dataset and B) non-subtype B dataset, Figure S2: Genetic p-distance (displayed as a difference from 1) between consensus sequence and HXB2, the reference sequence for HIV, for all pipelines for the simulated HIV A) subtype B dataset and B) non-subtype B dataset, Figure S3: Genetic p-distance (displayed as a difference from 1) between consensus sequence and HXB2, the reference sequence for HIV, for all pipelines for the empirical A) HIV dataset and B) HCV dataset, Table S1: Accessions used in validation study, Table S2: Kruskal-Wallis rank-sum test of the genetic p-distance and adjusted genetic p-distance of the pipeline consensus sequence from the true sequence or reference sequence for all datasets, Table S3: Wilcoxon signed-rank comparisons of the genetic p-distance from the true sequence between the initial and final consensus sequences constructed in the HAPHPIPE pipelines for the simulation dataset, Table S4: Kruskal-Wallis multiple comparisons (Dunn test) of the genetic p-distance of the pipeline consensus sequence from the true sequence for the simulation dataset, Table S5: Wilcoxon signed-rank comparisons of the genetic p-distance from HXB2, the HIV reference sequence, between the initial and final consensus sequences constructed in the HAPHPIPE pipelines for the simulation dataset, Table S6: Kruskal-Wallis multiple comparisons (Dunn test) of the genetic p-distance of the pipeline consensus sequence from HXB2, the HIV reference sequence, for the simulation dataset, Table S7: Kruskal-Wallis multiple comparisons (Dunn test) of the genetic p-distance of the pipeline consensus sequence from HXB2, the HIV reference sequence, for the HIV empirical dataset, Table S8: Kruskal-Wallis multiple comparisons (Dunn test) of the genetic p-distance of the pipeline consensus sequence from H77, the HCV reference sequence, for the HCV empirical dataset, Table S9: Wilcoxon signed-rank comparisons of the genetic p-distance from the reference sequence between the initial and final consensus sequences constructed in the HAPHPIPE pipelines for the empirical HIV and HCV datasets, Table S10: Genetic p-distance to the reference sequence across the empirical SARS-CoV-2 dataset.

Author Contributions: Conceptualization, K.M.G. and M.L.B.; data curation and methodology, K.M.G., M.C.S. and M.L.B.; software, investigation and formal analysis, K.M.G., M.C.S. and U.R.; validation, K.M.G., M.P.-L. and K.A.C.; writing—original draft preparation, K.M.G.; visualization and writing—review and editing, K.M.G. and M.C.S.; project administration, K.M.G., M.L.B. and K.A.C.; supervision and funding acquisition, M.L.-P. and K.A.C.; resources, K.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by a DC D-CFAR pilot award, a 2015 HIV Phylodynamics Supplement award from the District of Columbia for AIDS Research, an NIH funded program (AI117970), and NIH grants AI076059 and UL1TR001876. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Wilcoxon signed-rank comparisons of the adjusted genetic p-distance from the true sequence between the initial and final consensus sequences constructed in the HAPHPIPE pipelines for the simulation dataset.

	HP01				HP02			
Sub B	**Pseudomedian**	**CI Low**	**CI high**	***p* Value**	**Pseudomedian**	**CI Low**	**CI High**	***p* Value**
pol	−0.0004	NA [1]	NA [1]	1	0.0004	0.0004	0.0004	2.3×10^{-06} ***
PRRT	−0.0004	NA [1]	NA [1]	0.3711	0.0006	0.0006	0.0006	3.6×10^{-06} ***
int	−0.0003	NA [1]	NA [1]	1	0.0009	NA [1]	NA [1]	1
gp120	0.0008	NA [1]	NA [1]	0.3710	0.0298	0.0259	0.0333	$<2.2 \times 10^{-16}$ ***
Non-B	**Pseudomedian**	**CI Low**	**CI High**	***p* Value**	**Pseudomedian**	**CI Low**	**CI high**	***p* value**
pol	−0.0008	−0.0010	−0.0008	2.5×10^{-09} ***	0.0006	0.0005	0.0008	8.3×10^{-06} ***
PRRT	−0.0009	−0.0010	−0.0008	3.5×10^{-09} ***	0.0009	0.0006	0.0011	8.4×10^{-06} ***
int	−0.0012	−0.0017	−0.0011	1.2×10^{-06} ***	0.0017	NA [1]	NA [1]	1
gp120	−0.0007	−0.0009	−0.0006	4.7×10^{-06} ***	0.1475	0.1296	0.1657	7.8×10^{-10} ***

[1] No confidence intervals were constructed because too many differences were zero. A positive value indicates that the refined sequences are more genetically similar to the true sequence, while a negative value indicates that the refined sequences are less genetically similar to the true sequence. Abbreviations: Non-B: non-subtype B sequences, Sub B: subtype B sequences, HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, *pol* = polymerase, combination of *PRRT* and *int*, *PRRT* = protease and reverse transcriptase, *int* = integrase, CI: confidence interval, *** indicates $p < 0.001$.

Table A2. Kruskal–Wallis multiple comparisons (Dunn test) of the adjusted genetic p-distance of the pipeline consensus sequence from the true sequence for the simulation dataset. Adjusted *p*-values are reported (Holm adjustment).

		Subtype B Simulation Data					Non-Subtype B Simulation Data				
Gene	**Pipeline**	**HP01**	**HP02**	**GDN**	**GRB**	**HyDRA**	**HP01**	**HP02**	**GDN**	**GRB**	**HyDRA**
pol	**HP01**		0.1168 ***	1.1×10^{-25} ***	3.4×10^{-27} ***	2.7×10^{-73} ***		0.0079 **	5.2×10^{-13} ***	6.4×10^{-19} ***	1.3×10^{-40} ***
	HP02	0.1168		9.2×10^{-18} ***	5.5×10^{-19} ***	4.5×10^{-59} ***	0.0079 **		1.7×10^{-05} ***	2.7×10^{-09} ***	1.9×10^{-25} ***
	GDN	1.1×10^{-25} ***	9.2×10^{-18} ***		0.7396	1.4×10^{-13} ***	5.2×10^{-13} ***	1.7×10^{-05} ***		0.1021	7.9×10^{-09} ***
	GRN	3.4×10^{-27} ***	5.5×10^{-19} ***	0.7396		1.3×10^{-12} ***	6.4×10^{-19} ***	2.7×10^{-09} ***	0.1021		3.2×10^{-05} ***
	HyDRA	2.7×10^{-73} ***	4.5×10^{-59} ***	1.4×10^{-13} ***	1.3×10^{-12} ***		1.3×10^{-40} ***	1.9×10^{-25} ***	7.9×10^{-09} ***	3.2×10^{-05} ***	
PRRT	**HP01**		0.1053 ***	4.2×10^{-26} ***	2.7×10^{-27} ***	5.0×10^{-74} ***		8.5×10^{-05} ***	7.7×10^{-10} ***	3.1×10^{-16} ***	1.8×10^{-39} ***
	HP02	0.1053		6.2×10^{-18} ***	6.9×10^{-19} ***	2.0×10^{-59} ***	8.5×10^{-05} ***		0.0499 *	9.4×10^{-05} ***	6.1×10^{-19} ***
	GDN	4.2×10^{-26} ***	6.2×10^{-18} ***		0.7918	1.4×10^{-13} ***	7.7×10^{-10} ***	0.0499 *		0.0470 *	3.9×10^{-11} ***
	GRN	2.7×10^{-27} ***	6.9×10^{-19} ***	0.7918		7.7×10^{-13} ***	3.1×10^{-16} ***	9.4×10^{-05} ***	0.0470 *		4.8×10^{-06} ***
	HyDRA	5.0×10^{-74} ***	2.0×10^{-59} ***	1.4×10^{-13} ***	7.7×10^{-13} ***		1.8×10^{-39} ***	6.1×10^{-19} ***	3.9×10^{-11} ***	4.8×10^{-06} ***	
int	**HP01**		1	8.7×10^{-23} ***	1.1×10^{-23} ***	9.4×10^{-69} ***		0.5092	2.0×10^{-12} ***	1.2×10^{-13} ***	6.0×10^{-36} ***
	HP02	1		1.4×10^{-21} ***	2.0×10^{-22} ***	2.1×10^{-66} ***	0.5092		3.7×10^{-09} ***	3.5×10^{-10} ***	5.9×10^{-30} ***
	GDN	8.7×10^{-23} ***	1.4×10^{-21} ***		0.8290	9.1×10^{-14} ***	2.0×10^{-12} ***	3.7×10^{-09} ***		0.6963	2.7×10^{-07} ***
	GRN	1.1×10^{-23} ***	2.0×10^{-22} ***	0.8290		3.6×10^{-13} ***	1.2×10^{-13} ***	3.5×10^{-10} ***	0.6963		1.6×10^{-06} ***
	HyDRA	9.4×10^{-69} ***	2.1×10^{-66} ***	9.1×10^{-14} ***	3.6×10^{-13} ***		6.0×10^{-36} ***	5.9×10^{-30} ***	2.7×10^{-07} ***	1.6×10^{-06} ***	
gp120	**HP01**		2.8×10^{-26} ***	0.4269	2.0×10^{-40} ***	NA		6.5×10^{-05} ***	0.1212	2.5×10^{-19} ***	NA
	HP02	2.8×10^{-26} ***		8.3×10^{-23} ***	0.0134 *	NA	6.5×10^{-05} ***		4.7×10^{-08} ***	1.6×10^{-06} ***	NA
	GDN	0.4269	8.3×10^{-23} ***		5.5×10^{-36} ***	NA	0.1212	4.7×10^{-08} ***		5.3×10^{-26} ***	NA
	GRN	2.0×10^{-40} ***	0.0134 *	5.5×10^{-36} ***		NA	2.5×10^{-19} ***	1.6×10^{-06} ***	5.3×10^{-26} ***		NA

Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, *pol* = polymerase, combination of *PRRT* and *int*, *PRRT* = protease and reverse transcriptase, *int* = integrase, *** indicates $p < 0.001$, * indicates $p < 0.05$.

Table A3. Wilcoxon signed-rank comparisons of the adjusted genetic p-distance from HXB2, the HIV reference sequence, between the initial and final consensus sequences constructed in the HAPHPIPE pipelines for the simulation dataset.

	HP01				HP02			
Sub B	**pseudomedian**	**CI low**	**CI high**	***p* value**	**pseudomedian**	**CI low**	**CI high**	***p* value**
pol	0.0008	NA [1]	NA [1]	1	8.9×10^{-06}	-7.2×10^{-05}	0.0001	0.2074
PRRT	0.0002	NA [1]	NA [1]	1	-2.6×10^{-05}	−0.0002	0.0002	0.1670
int	0.0020	NA [1]	NA [1]	1	0.0009	NA [1]	NA [1]	1
gp120	0.0006	0.0006	0.0006	0.3711	0.0135	0.0105 *	0.0169	2.1×10^{-13} ***
Non-B	**pseudomedian**	**CI low**	**CI high**	***p* value**	**pseudomedian**	**CI low**	**CI high**	***p* value**
pol	−0.0009	−0.0010	−0.0008	1.8×10^{-09} ***	0.0001	-2.8×10^{-05}	0.0003	0.1692
PRRT	−0.0009	−0.0010	−0.0008	4.6×10^{-09} ***	0.0002	-3.7×10^{-05}	0.0004	0.1614
int	−0.0012	−0.0017	−0.0012	2.1×10^{-07} ***	−0.0006	NA [1]	NA [1]	1
gp120	−0.0006	−0.0007	−0.0005	5.9×10^{-05} ***	0.1070	0.0949	0.1209	7.8×10^{-10} ***

[1] No confidence intervals were constructed because too many differences were zero. A positive value indicates that the refined sequences are more genetically similar to the reference sequence (HXB2), while a negative value indicates that the refined sequences are less genetically similar to the reference sequence (HXB2). Abbreviations: Sub B: HIV subtype B sequences, Non-B: HIV non-subtype B sequences, HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, *pol* = polymerase, combination of *PRRT* and *int*, *PRRT* = protease and reverse transcriptase, *int* = integrase, CI: confidence interval, *** indicates $p < 0.001$, * indicates $p < 0.05$.

Table A4. Kruskal–Wallis multiple comparisons (Dunn test) of the adjusted genetic p-distance of the pipeline consensus sequence from HXB2, the HIV reference sequence, for the simulation dataset. Adjusted *p*-values are reported (Holm adjustment).

		Subtype B Simulation Data					Non-Subtype B Simulation Data				
Gene	**Pipeline**	**HP01**	**HP02**	**GDN**	**GRB**	**HyDRA**	**HP01**	**HP02**	**GDN**	**GRB**	**HyDRA**
pol	**HP01**		1	2.3×10^{-08} ***	1.4×10^{-08} ***	2.7×10^{-38} ***		0.8790	0.0013 **	0.0002 ***	1.2×10^{-17} ***
	HP02	1		5.8×10^{-08} ***	4.0×10^{-08} ***	5.6×10^{-37} ***	0.8790		0.0120 *	0.0029 ***	7.9×10^{-15} ***
	GDN	2.3×10^{-08} ***	5.8×10^{-08} ***		0.9104	3.2×10^{-12} ***	0.0013	0.0120 *		0.6142 ***	1.9×10^{-06} ***
	GRN	1.4×10^{-08} ***	4.0×10^{-08} ***	0.9104		6.3×10^{-12} ***	0.0002 ***	0.0029 ***	0.6142		2.2×10^{-05} ***
	HyDRA	2.7×10^{-38} ***	5.6×10^{-37} ***	3.2×10^{-12} ***	6.3×10^{-12} ***		1.2×10^{-17} ***	7.9×10^{-15} ***	1.9×10^{-06} ***	2.2×10^{-05} ***	
PRRT	**HP01**		1	0.0001 ***	7.8×10^{-05} ***	5.9×10^{-24} ***		0.7740	0.0387 *	0.0099 **	6.8×10^{-13} ***
	HP02	1		0.0004 ***	0.0003 ***	2.9×10^{-22} ***	0.7740		0.2163	0.0897	3.1×10^{-10} ***
	GDN	0.0001 ***	0.0004 ***		0.9049	9.8×10^{-09} ***	0.0387 *	0.2163		0.6279	1.1×10^{-05} ***
	GRN	7.8×10^{-05} ***	0.0003 ***	0.9049		1.8×10^{-08} ***	0.0099 **	0.0897	0.6279		9.8×10^{-05} ***
	HyDRA	5.9×10^{-24} ***	2.9×10^{-22} ***	9.8×10^{-09} ***	1.80×10^{-08} ***		6.8×10^{-13} ***	3.1×10^{-10} ***	1.1×10^{-05} ***	9.8×10^{-05} ***	
int	**HP01**		0.9649	6.0×10^{-12} ***	4.7×10^{-12} ***	7.5×10^{-49} ***		1	0.0002 ***	0.0001 ***	3.2×10^{-19} ***
	HP02	0.9649		5.8×10^{-12} ***	4.1×10^{-12} ***	4.3×10^{-49} ***	1		0.0005 ***	0.0003 ***	5.0×10^{-18} ***
	GDN	6.0×10^{-12} ***	5.8×10^{-12} ***		1	4.9×10^{-14} ***	0.0002 ***	0.0005 ***		0.8497	2.4×10^{-06} ***
	GRN	4.7×10^{-12} ***	4.1×10^{-12} ***	1		9.6×10^{-14} ***	0.0001 ***	0.0003 ***	0.8497		5.6×10^{-06} ***
	HyDRA	7.5×10^{-49} ***	4.3×10^{-49} ***	4.9×10^{-14} ***	9.6×10^{-14} ***		3.2×10^{-19} ***	5.0×10^{-18} ***	2.4×10^{-06} ***	5.6×10^{-06} ***	
gp120	**HP01**		NA [1]	NA [1]	NA [1]	NA [1]		0.1905	0.1731	3.4×10^{-11} ***	NA
	HP02	NA [1]		NA [1]	NA [1]	NA [1]	0.1905		0.8873	2.2×10^{-17} ***	NA
	GDN	NA [1]	NA [1]		NA [1]	NA [1]	0.1731	0.8873		6.4×10^{-17} ***	NA
	GRN	NA [1]	NA [1]	NA [1]		NA [1]	3.4×10^{-11} ***	2.2×10^{-17} ***	6.4×10^{-17} ***		NA

[1] The Kruskal–Wallis rank-sum test was not significant ($p = 0.0799$, Table S2). Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, *pol* = polymerase, combination of *PRRT* and *int*, *PRRT* = protease and reverse transcriptase, *int* = integrase, CI: confidence interval, *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$.

Table A5. Comparison of the effect of consensus generation on estimated genetic diversity across the simulation dataset.

Simulated Data Subtype B													
		PRRT				***int***				***gp120***			
Pipeline	**N**	**H**	**S**	**π**	**θ**	**H**	**S**	**π**	**θ**	**H**	**S**	**π**	**θ**
True sequences	100	100	826	0.0654	0.0987	100	389	0.0510	0.0870	100	1407	0.2219	0.1913
HP01	100	100	824	0.0654	0.0984	100	386	0.0508	0.0863	100	1378	0.1458	0.1831
HP02	100	100	823	0.0657	0.0986	100	388	0.0525	0.0876	100	1362	0.2158	0.1877
HyDRA	100	100	820	0.0656	0.0994	100	377	0.0514	0.0862	NA	NA	NA	NA
GRB	100	100	835	0.0717	0.1047	100	384	0.0509	0.0509	100	1234	0.1787	0.1657
GDN	100	100	823	0.0656	0.0989	100	384	0.0507	0.0761	100	1421	0.2294	0.1930
Simulated Data non-Subtype B													
Pipeline	**N**	**H**	**S**	**π**	**θ**	**H**	**S**	**π**	**θ**	**H**	**S**	**π**	**θ**
True sequences	50	50	753	0.1016	0.1039	50	339	0.0749	0.0876	50	1185	0.2419	0.1809
HP01	50	50	749	0.1013	0.1034	50	337	0.0746	0.0871	50	1106	0.2325	0.1702
HP02	50	50	746	0.1004	0.1025	50	335	0.0749	0.0871	50	957	0.1999	0.1723
HyDRA	50	50	745	0.1020	0.1045	50	328	0.0749	0.0868	NA	NA	NA	NA
GRB	50	50	748	0.1016	0.1039	50	338	0.0758	0.0892	50	916	0.1956	0.1750
GDN	50	50	751	0.1018	0.1042	50	338	0.0757	0.0892	50	1180	0.2406	0.1785

HyDRA is an average of read1 and read2. Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, *PRRT* = protease and reverse transcriptase, *int* = integrase, N = number of sequences, H = number of haplotypes, S = number of polymorphic sites, π = nucleotide diversity, θ = Watterson's genetic diversity.

Table A6. Wilcoxon signed-rank comparisons of the adjusted genetic p-distance from the reference sequence between the initial and final consensus sequences constructed in the HAPHPIPE pipelines for the empirical HIV and HCV datasets. The reference sequences for HIV and HCV were HXB2 and H77, respectively.

	HP01				HP02			
Empirical HIV dataset								
	pseudomedian	**CI low**	**CI high**	***p* value**	**pseudomedian**	**CI low**	**CI high**	***p* value**
PRRT	0.0022	0.0017	0.0035	1.5×10^{-06} ***	NA [1]	NA [1]	NA [1]	NA [1]
int	0.0021	0.0016	0.0031	2.2×10^{-05} ***	−0.0012	NA [1]	NA [1]	0.1489
gp120	0.0026	0.0020	0.0038	2.6×10^{-07} ***	−0.0005	−0.0006	−0.0005	0.0029 **
Empirical HCV dataset								
	pseudomedian	**CI low**	**CI high**	***p* value**	**pseudomedian**	**CI low**	**CI high**	***p* value**
core	0.0083	0.0021	0.0125	0.0012 **	0.0101	0.0025	0.1027	0.0412 *
E1	0.0086	0.0054	0.0121	0.0001 ***	−0.0026	−0.0069	−0.0017	0.0579
E2	0.0079	0.0048	0.0144	0.0001 ***	−0.0027	−0.0040	−0.0013	0.0001 ***

[1] No confidence intervals were constructed because too many differences were zero. A positive value indicates that the refined sequences are more genetically similar to the reference sequence, while a negative value indicates that the refined sequences are less genetically similar to the reference sequence. The reference sequence for the empirical HIV and HCV datasets were HXB2 and H77, respectively. Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, *PRRT* = protease and reverse transcriptase, *int* = integrase, CI: confidence interval, *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$.

Table A7. Kruskal–Wallis multiple comparisons (Dunn test) of the adjusted genetic p-distance of the pipeline consensus sequence from HXB2, the HIV reference sequence, for the HIV empirical dataset. Adjusted *p*-values are reported (Holm adjustment).

Gene	Pipeline	HP01	HP02	HP01 Haps	HP02 Haps	GDN	GRB	HyDRA
Pol	**HP01**		4.6×10^{-14} ***	0.4447	1.0×10^{-09} ***	2.1×10^{-10} ***	2.0×10^{-15} ***	1.5×10^{-08} ***
	HP02	4.6×10^{-14} ***		3.2×10^{-30} ***	0.0058 **	0.7591	0.6986	0.4900
	HP01 haps	0.4447	3.2×10^{-30} ***		5.1×10^{-36} ***	2.3×10^{-23} ***	1.0×10^{-32} ***	8.9×10^{-20} ***
	HP02 haps	1.0×10^{-09} ***	0.0058 **	5.1×10^{-36} ***		0.3641	0.0009 ***	0.9894
	GDN	2.1×10^{-10} ***	0.7591	2.3×10^{-23} ***	0.3641		0.6298	1
	GRB	2.0×10^{-15} ***	0.6986	1.0×10^{-32} ***	0.0009 ***	0.6298		0.2508
	HyDRA	1.5×10^{-08} ***	0.4900	8.9×10^{-20} ***	0.9894	1	0.2508	
Int	**HP01**		4.9×10^{-12} ***	1	1.6×10^{-14} ***	2.4×10^{-09} ***	2.2×10^{-13} ***	6.0×10^{-16} ***
	HP02	4.9×10^{-12} ***		1.0×10^{-23} ***	1	1	1	1
	HP01 haps	1	1.0×10^{-23} ***		3.7×10^{-47} ***	8.1×10^{-19} ***	3.5×10^{-26} ***	7.7×10^{-31} ***
	HP02 haps	1.6×10^{-14} ***	1	3.7×10^{-47} ***		0.9955	0.8553	0.1024
	GDN	2.4×10^{-09} ***	1	8.1×10^{-19} ***	0.9955		1	0.3966
	GRB	2.2×10^{-13} ***	1	3.5×10^{-26} ***	0.8553	1		1
	HyDRA	6.0×10^{-16} ***	1	7.7×10^{-31} ***	0.1024	0.3966	1	
gp120	**HP01**		1.8×10^{-14} ***	0.0455 *	2.6×10^{-09} ***	0.0001 ***	3.2×10^{-06} ***	NA
	HP02	1.8×10^{-14} ***		2.1×10^{-37} ***	0.0228 *	0.0023 **	0.0241 *	NA
	HP01 haps	0.0455 *	2.1×10^{-37} ***		2.8×10^{-36} ***	1.5×10^{-15} ***	4.1×10^{-19} ***	NA
	HP02 haps	2.6×10^{-09} ***	0.0228 *	2.8×10^{-36} ***		0.5908	0.6382	NA
	GDN	0.0001 ***	0.0023 *	1.5×10^{-15} ***	0.5908		0.9509	NA
	GRB	3.2×10^{-06} ***	0.0241 *	4.1×10^{-19} ***	0.6382	0.9509		NA

Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, haps = haplotypes, *pol* = polymerase, combination of *PRRT* and *int*, *PRRT* = protease and reverse transcriptase, *int* = integrase, *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$.

Table A8. Kruskal–Wallis multiple comparisons (Dunn test) of the adjusted genetic p-distance of the pipeline consensus sequence from H77, the HCV reference sequence, for the HCV empirical dataset. Adjusted p-values are reported (Holm adjustment).

Gene	Pipeline	HP01	HP02	HP01 Haps	HP02 Haps	GDN	GRB	HyDRA
Core	**HP01**		1.0×10^{-10} ***	0.7016	6.8×10^{-14} ***	1	0.0027 **	
	HP02	1.0×10^{-10} ***		4.6×10^{-14} ***	1	3.0×10^{-08} ***	0.0071 **	1.0×10^{-10} ***
	HP01 haps	0.7016	4.6×10^{-14} ***		2.0×10^{-23} ***	1	0.0007 ***	0.7016
	HP02 haps	6.8×10^{-14} ***	1	2.0×10^{-23} ***		1.9×10^{-10} ***	0.0039 **	6.8×10^{-14} ***
	GDN	1	3.0×10^{-08} ***	1	1.9×10^{-10} ***		0.0359 *	1
	GRB	0.0027 **	0.0071 **	0.0007 ***	0.0039 **	0.0359 *		0.0027 **
E1	**HP01**		9.0×10^{-07} ***	0.9367	1.0×10^{-06} ***	0.0014 **	3.5×10^{-08} ***	
	HP02	9.0×10^{-07} ***		3.5×10^{-09} ***	1	0.6392	1	9.0×10^{-07} ***
	HP01 haps	0.9367	3.5×10^{-09} ***		2.4×10^{-10} ***	8.2×10^{-05} ***	4.3×10^{-11} ***	0.9367
	HP02 haps	1.0×10^{-06} ***	1	2.4×10^{-10} ***		1	0.5477	1.0×10^{-06} ***
	GDN	0.0014 **	0.6392	8.2×10^{-05} ***	1		0.2005	0.0014 **
	GRB	3.5×10^{-08} ***	1	4.3×10^{-11} ***	0.5477	0.2005		3.5×10^{-08} ***
E2	**HP01**		2.8×10^{-04} ***	1	3.6×10^{-06} ***	0.0685	5.5×10^{-08} ***	
	HP02	2.8×10^{-04} ***		1.7×10^{-05} ***	0.8929	0.3236	0.3692	2.8×10^{-04} ***
	HP01 haps	1	1.7×10^{-05} ***		8.7×10^{-10} ***	0.0327 *	1.2×10^{-10} ***	1
	HP02 haps	3.6×10^{-06} ***	0.8929	8.7×10^{-10} ***		0.2427	0.3076	3.6×10^{-06} ***
	GDN	0.0685	0.3236	0.0327 *	0.2427		0.0090 **	0.0685
	GRB	5.5×10^{-08} ***	0.3692	1.2×10^{-10} ***	0.3076	0.0090 **		5.5×10^{-08} ***

Abbreviations: HP01 = haphpipe_assemble_01 (de novo assembly), HP02 = haphpipe_assemble_02 (reference-based assembly), GDN = Geneious de novo assembly, GRB = Geneious reference-based assembly, haps = haplotypes, *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$.

References

1. Zanini, F.; Brodin, J.; Thebo, L.; Lanz, C.; Bratt, G.; Albert, J.; Neher, R.A. Population genomics of intrapatient HIV-1 evolution. *Elife* **2015**, *4*, 1–26. [CrossRef] [PubMed]
2. Bonnaud, E.M.; Troupin, C.; Dacheux, L.; Holmes, E.C.; Monchatre-Leroy, E.; Tanguy, M.; Bouchier, C.; Cliquet, F.; Barrat, J.; Bourhy, H. Comparison of intra- and inter-host genetic diversity in rabies virus during experimental cross-species transmission. *PLOS Pathog.* **2019**, *15*, e1007799. [CrossRef]
3. Pagán, I. The diversity, evolution and epidemiology of plant viruses: A phylogenetic view. *Infect. Genet. Evol.* **2018**, *65*, 187–199. [CrossRef]
4. Pérez-Losada, M.; Arenas, M.; Galán, J.C.; Bracho, M.A.; Hillung, J.; García-González, N.; González-Candelas, F. High-throughput sequencing (HTS) for the analysis of viral populations. *Infect. Genet. Evol.* **2020**, *80*, 104208. [CrossRef] [PubMed]
5. Simpson, J.T.; Pop, M. The theory and practice of genome sequence assembly. *Annu. Rev. Genom. Hum. Genet.* **2015**, *16*, 153–172. [CrossRef] [PubMed]
6. Nagarajan, N.; Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **2013**, *14*, 157–167. [CrossRef] [PubMed]
7. Günther, T.; Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **2019**, *15*, 1–20. [CrossRef]
8. Posada-Cespedes, S.; Seifert, D.; Beerenwinkel, N. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Res.* **2017**, *239*, 17–32. [CrossRef]
9. Archer, J.; Rambaut, A.; Taillon, B.E.; Richard Harrigan, P.; Lewis, M.; Robertson, D.L. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time-an ultra-deep approach. *PLoS Comput. Biol.* **2010**, *6*. [CrossRef]
10. Ji, H.; Enns, E.; Brumme, C.J.; Parkin, N.; Howison, M.; Lee, E.R.; Capina, R.; Marinier, E.; Avila-Rios, S.; Sandstrom, P.; et al. Bioinformatic data processing pipelines in support of next-generation sequencing-based HIV drug resistance testing: The Winnipeg Consensus. *J. Int. AIDS Soc.* **2018**, *21*, 1–14. [CrossRef]
11. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
12. Korber, B.; Foley, B.T.; Kuiken, C.; Pillai, S.K.; Sodroski, J.G. Others Numbering positions in HIV relative to HXB2CG. *AIDS Res. Hum. Retrovir.* **1998**, *3*, 102–111.
13. Banin, A.N.; Tuen, M.; Bimela, J.S.; Tongo, M.; Zappile, P.; Khodadadi-Jamayran, A.; Nanfack, A.J.; Meli, J.; Wang, X.; Mbanya, D.; et al. Development of a versatile, near full genome amplification and sequencing approach for a broad variety of HIV-1 group M variants. *Viruses* **2019**, *11*, 317. [CrossRef]
14. Kijak, G.H.; Sanders-Buell, E.; Pham, P.; Harbolick, E.A.; Oropeza, C.; O'Sullivan, A.M.; Bose, M.; Beckett, C.G.; Milazzo, M.; Robb, M.L.; et al. Next-generation sequencing of HIV-1 single genome amplicons. *Biomol. Detect. Quantif.* **2019**, *17*, 100080. [CrossRef] [PubMed]
15. Yamaguchi, J.; Olivo, A.; Laeyendecker, O.; Forberg, K.; Ndembi, N.; Mbanya, D.; Kaptue, L.; Quinn, T.C.; Cloherty, G.A.; Rodgers, M.A.; et al. Universal target capture of HIV sequences from NGS libraries. *Front. Microbiol.* **2018**, *9*, 1–13. [CrossRef] [PubMed]
16. Berg, M.G.; Yamaguchi, J.; Alessandri-Gradt, E.; Tell, R.W.; Plantier, J.C.; Brennan, C.A. A pan-HIV strategy for complete genome sequencing. *J. Clin. Microbiol.* **2016**, *54*, 868–882. [CrossRef]
17. Fonager, J.; Larsson, J.T.; Hussing, C.; Engsig, F.N.; Nielsen, C.; Fischer, T.K. Identification of minority resistance mutations in the HIV-1 integrase coding region using next generation sequencing. *J. Clin. Virol.* **2015**, *73*, 95–100. [CrossRef]
18. Pessoa, R.; Sanabani, S.S. High prevalence of HIV-1 transmitted drug- resistance mutations from proviral DNA massively parallel sequencing data of therapy- na ï ve chronically infected Brazilian blood donors. *PLoS ONE* **2017**, *12*, e0785559. [CrossRef]
19. Dudley, D.M.; Bailey, A.L.; Mehta, S.H.; Hughes, A.L.; Kirk, G.D.; Westergaard, R.P.; O'Connor, D.H. Cross-clade simultaneous HIV drug resistance genotyping for reverse transcriptase, protease, and integrase inhibitor mutations by Illumina MiSeq. *Retrovirology* **2014**, *11*, 1–15. [CrossRef] [PubMed]
20. Matume, N.D.; Tebit, D.M.; Gray, L.R.; Hammarskjold, M.L.; Rekosh, D.; Bessong, P.O. Next generation sequencing reveals a high frequency of CXCR4 utilizing viruses in HIV-1 chronically infected drug experienced individuals in South Africa. *J. Clin. Virol.* **2018**, *103*, 81–87. [CrossRef] [PubMed]

21. Perrier, M.; Désiré, N.; Storto, A.; Todesco, E.; Rodriguez, C.; Bertine, M.; Le Hingrat, Q.; Visseaux, B.; Calvez, V.; Descamps, D.; et al. Evaluation of different analysis pipelines for the detection of HIV-1 minority resistant variants. *PLoS ONE* **2018**, *13*, e0198334. [CrossRef]
22. Alves, B.M.; Siqueira, J.D.; Prellwitz, I.M.; Botelho, O.M.; Da Hora, V.P.; Sanabani, S.; Recordon-Pinson, P.; Fleury, H.; Soares, E.A.; Soares, M.A. Estimating HIV-1 genetic diversity in Brazil through next-generation sequencing. *Front. Microbiol.* **2019**, *10*, 1–11. [CrossRef] [PubMed]
23. Alves, B.M.; Siqueira, J.D.; Garrido, M.M.; Botelho, O.M.; Prellwitz, I.M.; Ribeiro, S.R.; Soares, E.A.; Soares, M.A. Genome Quasispecies from Patients with Undetectable Viral Load Undergoing First-Line HAART Therapy. *Viruses* **2017**, *9*, 392. [CrossRef] [PubMed]
24. Derache, A.; Iwuji, C.C.; Danaviah, S.; Giandhari, J.; Marcelin, A.-G.; Calvez, V.; de Oliveira, T.; Dabis, F.; Pillay, D.; Gupta, R.K. Predicted antiviral activity of tenofovir versus abacavir in combination with a cytosine analogue and the integrase inhibitor dolutegravir in HIV-1-infected South African patients initiating or failing first-line ART. *J. Anitmicrobial Chemother.* **2019**, *74*, 473–479. [CrossRef] [PubMed]
25. Ávila-Ríos, S.; García-Morales, C.; Matías-Florentino, M.; Romero-Mora, K.A.; Tapia-Trejo, D.; Quiroz-Morales, V.S.; Reyes-Gopar, H.; Ji, H.; Sandstrom, P.; Casillas-Rodríguez, J.; et al. Pretreatment HIV-drug resistance in Mexico and its impact on the effectiveness of first-line antiretroviral therapy: A nationally representative 2015 WHO survey. *Lancet HIV* **2016**, *3*, e579–e591. [CrossRef]
26. Taylor, T.; Lee, E.R.; Nykoluk, M.; Enns, E.; Liang, B.; Capina, R.; Gauthier, M.K.; Van Domselaar, G.; Sandstrom, P.; Brooks, J.; et al. A MiSeq-HyDRA platform for enhanced HIV drug resistance genotyping and surveillance. *Sci. Rep.* **2019**, *9*, 1–11. [CrossRef]
27. Arias, A.; López, P.; Sánchez, R.; Yamamura, Y.; Rivera-Amill, V. Sanger and next generation sequencing approaches to evaluate HIV-1 virus in blood compartments. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1697. [CrossRef]
28. Hora, B.; Keating, S.M.; Chen, Y.; Sanchez, A.M.; Sabino, E.; Hunt, G.; Ledwaba, J.; Hackett, J.; Swanson, P.; Hewlett, I.; et al. Genetic characterization of a panel of diverse HIV-1 isolates at seven international sites. *PLoS ONE* **2016**, *11*, e0157340. [CrossRef]
29. Nguyen, T.; Fofana, D.B.; Lê, M.P.; Charpentier, C.; Peytavin, G.; Wirden, M.; Lambert-Niclot, S.; Desire, N.; Grude, M.; Morand-Joubert, L.; et al. Prevalence and clinical impact of minority resistant variants in patients failing an integrase inhibitor-based regimen by ultra-deep sequencing. *J. Antimicrob. Chemother.* **2018**, *73*, 2485–2492. [CrossRef]
30. Charpentier, C.; Montes, B.; Perrier, M.; Meftah, N.; Reynes, J. HIV-1 DNA ultra-deep sequencing analysis at initiation of the dual therapy dolutegravir ! lamivudine in the maintenance DOLULAM pilot study. *J. Anitmicrobial Chemother.* **2017**, *72*, 2831–2836. [CrossRef]
31. Inzaule, S.C.; Hamers, R.L.; Noguera-julian, M.; Casadella, M.; Parera, M.; De Wit, T.F.R.; Paredes, R. Primary resistance to integrase strand transfer inhibitors in patients infected with diverse HIV-1 subtypes in sub-Saharan Africa. *J. Anitmicrobial Chemother.* **2018**, *73*, 1167–1172. [CrossRef] [PubMed]
32. Dalmat, R.R.; Makhsous, N.; Pepper, G.G.; Magaret, A.; Jerome, K.R.; Wald, A. Limited Marginal Utility of Deep Sequencing for HIV Drug Resistance Testing in the Age of Integrase Inhibitors. *J. Clin. Microbiol.* **2018**, *56*, 1–13. [CrossRef] [PubMed]
33. Redd, A.D.; Mullis, C.E.; Serwadda, D.; Kong, X.; Martens, C.; Ricklefs, S.M.; Tobian, A.A.R.; Xiao, C.; Grabowski, M.K.; Nalugoda, F.; et al. The Rates of HIV Superinfection and Primary HIV Incidence in a General Population in Rakai, Uganda. *J. Infect. Dis.* **2012**, *206*, 267–274. [CrossRef]
34. Eshleman, S.H.; Hudelson, S.E.; Redd, A.D.; Wang, L.; Debes, R.; Chen, Y.Q.; Martens, C.A.; Ricklefs, S.M.; Selig, E.J.; Porcella, S.F.; et al. Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial. *J. Infect. Dis* **2011**, *204*, 1918–1926. [CrossRef] [PubMed]
35. Tumiotto, C.; Riviere, L.; Bellecave, P.; Recordon-pinson, P.; Latitude, P.; Vilain-parce, A.; Guidicelli, G. Sanger and Next-Generation Sequencing data for characterization of CTL epitopes in archived HIV-1 proviral DNA. *PLoS ONE* **2017**, *12*, e0185211. [CrossRef] [PubMed]
36. Raymond, H.F.; Al-Tayyib, A.; Neaigus, A.; Reilly, K.H.; Braunstein, S.; Brady, K.A.; Sey, E.; Risser, J.; Padget, P.; Lalota, M.; et al. HIV Among MSM and Heterosexual Women in the United States: An Ecologic Analysis. *J. Acquir. Immune Defic. Syndr.* **2017**, *75*, S276–S280. [CrossRef] [PubMed]

37. Saliou, A.; Delobel, P.; Dubois, M.; Nicot, F.; Calvez, V.; Masquelier, B.; Izopet, J.; the ANRS AC11 Resostamce Study Group. Concordance between Two Phenotypic Assays and Ultradeep Pyrosequencing for Determining HIV-1 Tropism. *Antimicrob. Agents Chemother.* **2011**, *55*, 2831–2836. [CrossRef]
38. Prosperi, M.C.F.; Yin, L.; Nolan, D.J.; Lowe, A.D.; Goodenow, M.M.; Salemi, M. Empirical validation of viral quasispecies assembly algorithms: State-of-the-art and challenges. *Sci. Rep.* **2013**, *3*, 2837. [CrossRef]
39. Mbondji-wonje, C.; Dong, M.; Wang, X.; Zhao, J.; Ragupathy, V.; Sanchez, A.M.; Denny, T.N.; Hewlett, I. Distinctive variation in the U3R region of the 5 ′ Long Terminal Repeat from diverse HIV-1 strains. *PLoS ONE* **2018**, *13*, e0195661. [CrossRef]
40. Hiener, B.; Eden, J.; Horsburgh, B.A.; Palmer, S. Amplification of Near Full-length HIV-1 Proviruses for Next-Generation Sequencing. *J. Vis. Exp.* **2018**, *140*, e58016. [CrossRef]
41. Hunt, M.; Gall, A.; Ong, S.H.; Brener, J.; Ferns, B.; Goulder, P.; Nastouli, E.; Keane, J.A.; Kellam, P.; Otto, T.D. IVA: Accurate de novo assembly of RNA virus genomes. *Bioinformatics* **2015**, *31*, 2374–2376. [CrossRef] [PubMed]
42. Aralaguppe, S.G.; Siddik, A.B.; Manickam, A.; Ambikan, A.T.; Kumar, M.M.; Fernandes, S.J.; Amogne, W.; Bangaruswamy, D.K.; Hanna, L.E.; Sonnerborg, A.; et al. Multiplexed next-generation sequencing and de novo assembly to obtain near full-length HIV-1 genome from plasma virus. *J. Virol Methods* **2016**, *236*, 98–104. [CrossRef] [PubMed]
43. Alampalli, S.V.; Thomson, M.M.; Sampathkumar, R.; Sivaraman, K.; Anto Jesuraj, U.K.J.; Dhar, C.; Souza, G.D.; Berry, N.; Vyakarnam, A. Deep sequencing of near full-length HIV-1 genomes from plasma identifies circulating subtype C and infrequent occurrence of AC recombinant form in Southern India. *PLoS ONE* **2017**, *12*, e0188603. [CrossRef]
44. Cornelissen, M.; Gall, A.; Vink, M.; Zorgdrager, F.; Edwards, S.; Jurriaans, S.; Bakker, M.; Ong, S.H.; Gras, L.; Van Sighem, A.; et al. From clinical sample to complete genome: Comparing methods for the extraction of HIV-1 RNA for high-throughput deep sequencing. *Virus Res.* **2017**, *239*, 10–16. [CrossRef]
45. Ratmann, O.; Hodcroft, E.B.; Pickles, M.; Cori, A.; Hall, M.; Lycett, S.; Colijn, C.; Dearlove, B.; Didelot, X.; Frost, S.; et al. Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Mol. Biol Evol.* **2017**, *34*, 185–203. [CrossRef] [PubMed]
46. Ratmann, O.; Grabowski, M.K.; Hall, M.; Golubchik, T.; Wymant, C.; Abeler-dörner, L.; Bonsall, D.; Hoppe, A.; Brown, A.L.; De Oliveira, T.; et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nat. Commun.* **2019**, *10*, 1411. [CrossRef] [PubMed]
47. Kafando, A.; Fournier, E.; Serhir, B.; Martineau, C.; Doualla-, F.; Sylla, M.; Chamberland, A.; El-, M.; Sangare, M.N.; Charest, H.; et al. HIV-1 envelope sequence-based diversity measures for identifying recent infections. *PLoS ONE* **2017**, *12*, e0189999. [CrossRef] [PubMed]
48. Thomson, E.; Ip, C.L.C.; Badhan, A.; Christiansen, M.T.; Adamson, W.; Ansari, M.A.; Bibby, D.; Breuer, J.; Brown, A.; Bowden, R.; et al. Comparison of Next-Generation Sequencing Technologies for Comprehensive Assessment of Full-Length Hepatitis C Viral Genomes. *J. Clin. Microbiol.* **2016**, *54*, 2470–2484. [CrossRef] [PubMed]
49. Welzel, T.M.; Bhardwaj, N.; Hedskog, C.; Chodavarapu, K.; Camus, G.; McNally, J.; Brainard, D.; Miller, M.D.; Mo, H.; Svarovskaia, E.; et al. Global epidemiology of HCV subtypes and resistance-associated substitutions evaluated by sequencing-based subtype analyses. *J. Hepatol.* **2017**, *67*, 224–236. [CrossRef]
50. Ishida, Y.; Hayashida, T.; Sugiyama, M.; Tsuchiya, K.; Kikuchi, Y.; Mizokami, M.; Oka, S.; Gatanaga, H. Full-Genome Analysis of Hepatitis C Virus in Japanese and Non-Japanese Patients Coinfected with HIV-1 in Tokyo. *J. Acquir. Immune Defic. Syndr.* **2019**, *80*, 350–357. [CrossRef]
51. Nguyen, T.; Delaugerre, C.; Valantin, M.-A.; Amiel, C.; Netzer, E.; L'yavanc, T.; Ohayon, M.; Valin, N.; Day, N.; Kreplak, G.; et al. Shared HCV Transmission Networks Among HIV-1–Positive and HIV-1–Negative Men Having Sex With Men by Ultradeep Sequencing. *JAIDS J. Acquir. Immune Defic. Syndr.* **2019**, *82*, 105–110. [CrossRef] [PubMed]
52. Isaacs, S.R.; Kim, K.W.; Cheng, J.X.; Bull, R.A.; Stelzer-Braid, S.; Luciani, F.; Rawlinson, W.D.; Craig, M.E. Amplification and next generation sequencing of near full-length human enteroviruses for identification and characterisation from clinical samples. *Sci. Rep.* **2018**, *8*, 1–9. [CrossRef] [PubMed]
53. Majumdar, M.; Martin, J. Detection by direct next generation sequencing analysis of emerging enterovirus D68 and C109 strains in an environmental sample from Scotland. *Front. Microbiol.* **2018**, *9*, 1–11. [CrossRef] [PubMed]

54. Gaspareto, K.V.; Ribeiro, R.M.; de Mello Malta, F.; Gomes-Gouvêa, M.S.; Muto, N.H.; Mendes-Correa, M.C.; Rozanski, A.; Carrilho, F.J.; Sabino, E.C.; Pinho, J.R.R. HCV inter-subtype 1a/1b recombinant detected by complete-genome next-generation sequencing. *Arch. Virol.* **2016**, *161*, 2161–2168. [CrossRef] [PubMed]
55. Stelzl, E.; Haas, B.; Bauer, B.; Zhang, S.; Fiss, E.H.; Hillman, G.; Hamilton, A.T.; Mehta, R.; Heil, M.L.; Marins, E.G.; et al. First identification of a recombinant form of hepatitis C virus in Austrian patients by full-genome next generation sequencing. *PLoS ONE* **2017**, *12*, e0181273. [CrossRef]
56. Ogishi, M.; Yotsuyanagi, H.; Tsutsumi, T.; Gatanaga, H.; Ode, H.; Sugiura, W.; Moriya, K.; Oka, S.; Kimura, S.; Koike, K. Deconvoluting the composition of low-frequency hepatitis C viral quasispecies: Comparison of genotypes and NS3 resistance-associated variants between HCV/HIV coinfected hemophiliacs and HCV monoinfected patients in Japan. *PLoS ONE* **2015**, *10*, e0119145. [CrossRef]
57. Rodrigo, C.; Leung, P.; Lloyd, A.R.; Bull, R.A.; Luciani, F.; Grebely, J.; Dore, G.J.; Applegate, T.; Page, K.; Bruneau, J.; et al. Genomic variability of within-host hepatitis C variants in acute infection. *J. Viral Hepat.* **2019**, *26*, 476–484. [CrossRef]
58. Abayasingam, A.; Leung, P.; Eltahla, A.; Bull, R.A.; Luciani, F.; Grebely, J.; Dore, G.J.; Applegate, T.; Page, K.; Bruneau, J.; et al. Genomic characterization of hepatitis C virus transmitted founder variants with deep sequencing. *Infect. Genet. Evol.* **2019**, *71*, 36–41. [CrossRef]
59. Eltahla, A.A.; Leung, P.; Pirozyan, M.R.; Rodrigo, C.; Grebely, J.; Applegate, T.; Maher, L.; Luciani, F.; Lloyd, A.R.; Bull, R.A. Dynamic evolution of hepatitis C virus resistance-associated substitutions in the absence of antiviral treatment. *Sci. Rep.* **2017**, *7*, 1–6. [CrossRef] [PubMed]
60. Iles, J.C.; Njouom, R.; Foupouapouognigni, Y.; Bonsall, D.; Bowden, R.; Trebes, A.; Piazza, P.; Barnes, E.; Pépin, J.; Klenerman, P.; et al. Characterization of hepatitis c virus recombination in Cameroon by use of nonspecific next-generation sequencing. *J. Clin. Microbiol.* **2015**, *53*, 3155–3164. [CrossRef] [PubMed]
61. Bendall, M.L.; Gibson, K.M.; Steiner, M.C.; Pérez-Losada, M.; Keith, A. Crandall HAPHPIPE: Haplotype reconstruction and real-time phylodynamics for deep sequencing of intra-host viral populations. *Mol. Biol. Evol.* **2020**, (in press).
62. Ambler, S.W. *Agile Modeling*; John Wiley & Sons: Hoboken, NJ, USA, 2002.
63. Bertels, F.; Leemann, C.; Metzner, K.J.; Regoes, R.R. Parallel Evolution of HIV-1 in a Long-Term Experiment. *Mol. Biol. Evol.* **2019**, *36*, 2400–2414. [CrossRef] [PubMed]
64. Babcock, G.J.; Iyer, S.; Smith, H.L.; Wang, Y.; Rowley, K.; Ambrosino, D.M.; Zamore, P.D.; Pierce, B.G.; Molrine, D.C.; Weng, Z. High-throughput sequencing analysis of post-liver transplantation HCV E2 glycoprotein evolution in the presence and absence of neutralizing monoclonal antibody. *PLoS ONE* **2014**, *9*, e0100325. [CrossRef] [PubMed]
65. Eliseev, A.; Gibson, K.M.; Avdeyev, P.; Novik, D.; Bendall, M.L.; Pérez-Losada, M.; Alexeev, N.; Crandall, K.A. Evaluation of haplotype callers for next-generation sequencing of viruses. *Infect. Genet. Evol.* **2020**, *82*, 104277. [CrossRef] [PubMed]
66. Noguera-Julian, M.; Edgil, D.; Harrigan, P.R.; Sandstrom, P.; Godfrey, C.; Paredes, R. Next-Generation Human Immunodeficiency Virus Sequencing for Patient Management and Drug Resistance Surveillance. *J. Infect. Dis.* **2017**, *216*, S829–S833. [CrossRef]
67. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
68. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
69. Alnafee, K. Versatile and open software for comparing large genomes. *Genome Biol.* **2016**, *5*, 12–23. [CrossRef]
70. Prabhakaran, S.; Rey, M.; Zagordi, O.; Beerenwinkel, N.; Roth, V. HIV haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 182–191. [CrossRef]
71. Knyazev, S.; Tsyvina, V.; Melnyk, A.; Malygina, T.; Porozov, Y.B.; Campbell, E.; Switzer, W.M.; Skums, P.; Zelikovsky, A. CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads. *bioRxiv* **2018**, 1–8. [CrossRef]
72. Nikolenko, S.I.; Korobeynikov, A.I.; Alekseyev, M.A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genom.* **2013**, *14*. [CrossRef]

73. Nurk, S.; Bankevich, A.; Antipov, D.; Gurevich, A.A.; Korobeynikov, A.; Lapidus, A.; Prjibelski, A.D.; Pyshkin, A.; Sirotkin, A.; Sirotkin, Y.; et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **2013**, *20*, 714–737. [CrossRef] [PubMed]
74. Leal, E.; Villanova, F.E. Diversity of HIV-1 Subtype B: Implications to the Origin of BF Recombinants. *PLoS ONE* **2010**, *5*, e011833. [CrossRef] [PubMed]
75. Ibe, S.; Shigemi, U.; Sawaki, K.; Fujisaki, S.; Hattori, J.; Yokomaku, Y.; Mamiya, N.; Hamaguchi, M.; Kaneda, T. Analysis of near full-length genomic sequences of drug-resistant HIV-1 spreading among therapy-naive individuals in Nagoya, Japan: Amino acid mutations associated with viral replication activity. *AIDS Res. Hum. Retrovir.* **2008**, *24*, 1121–1125. [CrossRef] [PubMed]
76. Oelrichs, R.; Tsykin, A.; Rhodes, D.; Solomon, A.; Ellett, A.; McPhee, D.; Deacon, N. Genomic sequence of HIV type 1 from four members of the Sydney Blood Bank Cohort of long-term nonprogressors. *AIDS Res. Hum. Retrovir.* **1998**, *14*, 811–814. [CrossRef]
77. Novelli, P.; Vella, C.; Oxford, J.; Daniels, R.S. Construction and biological characterization of an infectious molecular clone of HIV type 1GB8. *AIDS Res. Hum. Retrovir.* **2000**, *16*, 1175–1178. [CrossRef]
78. Hierholzer, J.; Montano, S.; Hoelscher, M.; Negrete, M.; Hierholzer, M.; Avila, M.M.; Carrillo, M.G.; Russi, J.C.; Vinoles, J.; Alava, A.; et al. Molecular Epidemiology of HIV Type 1 in Ecuador, Peru, Bolivia, Uruguay, and Argentina. *AIDS Res. Hum. Retrovir.* **2002**, *18*, 1339–1350. [CrossRef]
79. Viñoles, J.; Serra, M.; Russi, J.C.; Ruchansky, D.; Sosa-Estani, S.; Montano, S.M.; Carrion, G.; Eyzaguirre, L.M.; Carr, J.K.; Olson, J.G.; et al. Seroincidence and phylogeny of human immunodeficiency virus infections in a cohort of commercial sex workers in Montevideo, Uruguay. *Am. J. Trop. Med. Hyg.* **2005**, *72*, 495–500. [CrossRef] [PubMed]
80. Saad, M.D.; Al-Jaufy, A.; Grahan, R.R.; Nadai, Y.; Earhart, K.C.; Sanchez, J.L.; Carr, J.K. HIV type 1 strains common in Europe, Africa, and Asia cocirculate in Yemen. *AIDS Res. Hum. Retrovir.* **2005**, *21*, 644–648. [CrossRef] [PubMed]
81. Mikhail, M.; Wang, B.; Lemey, P.; Beckthold, B.; Vandamme, A.M.; Gill, M.J.; Saksena, N.K. Role of viral evolutionary rate in HIV-1 disease progression in a linked cohort. *Retrovirology* **2005**, *2*, 1–10. [CrossRef] [PubMed]
82. Allen, T.M.; Altfeld, M.; Geer, S.C.; Kalife, E.T.; Moore, C.; Desouza, I.; Feeney, M.E.; Eldridge, R.L.; Maier, E.L.; Kaufmann, D.E.; et al. Selective Escape from CD8+ T-Cell Responses Represents a Major Driving Force of Human Immunodeficiency Virus Type 1 (HIV-1) Sequence Diversity and Reveals Constraints on HIV-1 Evolution. *J. Virol.* **2005**, *79*, 13239–13249. [CrossRef]
83. Li, B.; Gladden, A.D.; Altfeld, M.; Kaldor, J.M.; Cooper, D.A.; Kelleher, A.D.; Allen, T.M. Rapid Reversion of Sequence Polymorphisms Dominates Early Human Immunodeficiency Virus Type 1 Evolution. *J. Virol.* **2007**, *81*, 193–201. [CrossRef]
84. Frahm, N.; Kaufmann, D.E.; Yusim, K.; Muldoon, M.; Kesmir, C.; Linde, C.H.; Fischer, W.; Allen, T.M.; Li, B.; McMahon, B.H.; et al. Increased Sequence Diversity Coverage Improves Detection of HIV-Specific T Cell Responses. *J. Immunol.* **2007**, *179*, 6638–6650. [CrossRef]
85. Andresen, B.S.; Vinner, L.; Tang, S.; Bragstad, K.; Kronborg, G.; Gerstoft, J.; Corbet, S.; Fomsgaard, A. Characterization of near full-length genomes of HIV type 1 strains in Denmark: Basis for a universal therapeutic vaccine. *AIDS Res. Hum. Retrovir.* **2007**, *23*, 1442–1448. [CrossRef]
86. Nadai, Y.; Eyzaguirre, L.M.; Sill, A.; Cleghorn, F.; Nolte, C.; Charurat, M.; Collado-Chastel, S.; Jack, N.; Bartholomew, C.; Pape, J.W.; et al. HIV-1 epidemic in the Caribbean is dominated by subtype B. *PLoS ONE* **2009**, *4*, e04814. [CrossRef] [PubMed]
87. Diaz, R.S.; Leal, É.; Sanabani, S.; Sucupira, M.C.A.; Tanuri, A.; Sabino, E.C.; Janini, L.M. Selective regimes and evolutionary rates of HIV-1 subtype B V3 variants in the Brazilian epidemic. *Virology* **2008**, *381*, 184–193. [CrossRef] [PubMed]
88. Kousiappa, I.; Van De Vijver, D.A.M.C.; Kostrikis, L.G. Near full-length genetic analysis of HIV sequences derived from Cyprus: Evidence of a highly polyphyletic and evolving infection. *AIDS Res. Hum. Retrovir.* **2009**, *25*, 727–740. [CrossRef] [PubMed]
89. Wang, Y.E.; Li, B.; Carlson, J.M.; Streeck, H.; Gladden, A.D.; Goodman, R.; Schneidewind, A.; Power, K.A.; Toth, I.; Frahm, N.; et al. Protective HLA Class I Alleles That Restrict Acute-Phase CD8+ T-Cell Responses Are Associated with Viral Escape Mutations Located in Highly Conserved Regions of Human Immunodeficiency Virus Type 1. *J. Virol.* **2009**, *83*, 1845–1855. [CrossRef] [PubMed]

90. Salazar-Gonzalez, J.F.; Salazar, M.G.; Keele, B.F.; Learn, G.H.; Giorgi, E.E.; Li, H.; Decker, J.M.; Wang, S.; Baalwa, J.; Kraus, M.H.; et al. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J. Exp. Med.* **2009**, *206*, 1273–1289. [CrossRef]
91. Cuevas, M.T.; Fernandez-Garcia, A.; Pinilla, M.; Garcia-Alvarez, V.; Thomson, M.; Delgado, E.; Gonzalez-Galeano, M.; Miralles, C.; Serrano-Bengoechea, E.; Ojea de Castro, R.; et al. Short communication: Biological and genetic characterization of HIV type 1 subtype B and nonsubtype B transmitted viruses: Usefulness for vaccine candidate assessment. *AIDS Res. Hum. Retrovir.* **2010**, *26*, 1019–1025. [CrossRef] [PubMed]
92. Turnbull, E.L.; Wong, M.; Wang, S.; Wei, X.; Jones, N.A.; Conrod, K.E.; Aldam, D.; Turner, J.; Pellegrino, P.; Keele, B.F.; et al. Kinetics of Expansion of Epitope-Specific T Cell Responses during Primary HIV-1 Infection. *J. Immunol.* **2009**, *182*, 7131–7145. [CrossRef]
93. Rolland, M.; Tovanabutra, S.; Allan, C.; Frahm, N.; Peter, B.; Sanders-buell, E.; Heath, L.; Magaret, C.A.; Bose, M.; Sullivan, A.O.; et al. Genetic impact of vaccination on breakthrough HIV-1 sequences from the Step trial. *Nat. Med.* **2011**, *17*, 366–371. [CrossRef] [PubMed]
94. Kousiappa, I.; Achilleos, C.; Hezka, J.; Lazarou, Y.; Othonos, K.; Demetriades, I.; Kostrikis, L.G. Molecular characterization of HIV type 1 strains from newly diagnosed patients in Cyprus (2007–2009) recovers multiple clades including unique recombinant strains and lack of transmitted drug resistance. *AIDS Res. Hum. Retrovir.* **2011**, *27*, 1183–1199. [CrossRef]
95. Eyzaguirrea, L.M.; Charurata, M.; Redfielda, R.R.; Blattnera, W.A.; Carra, J.K.; Sajadi, M.M. Elevated hypermutation levels in HIV-1 natural viral suppressors. *Virology* **2013**, *443*, 306–312. [CrossRef] [PubMed]
96. Li, Z.; He, X.; Wang, Z.; Xing, H.; Li, F.; Yang, Y.; Wang, Q.; Takebe, Y.; Shao, Y. Tracing the origin and history of HIV-1 subtype B′ epidemic by near full-length genome analyses. *Aids* **2012**, *26*, 877–884. [CrossRef] [PubMed]
97. Kijak, G.H.; Tovanabutra, S.; Rerks-Ngarm, S.; Nitayaphan, S.; Eamsila, C.; Kunasol, P.; Khamboonruang, C.; Thongcharoen, P.; Namwat, C.; Premsri, N.; et al. Molecular Evolution of the HIV-1 Thai Epidemic between the Time of RV144 Immunogen Selection to the Execution of the Vaccine Efficacy Trial. *J. Virol.* **2013**, *87*, 7265–7281. [CrossRef] [PubMed]
98. Sanabani, S.S.; de Pastena, É.R.S.; da Costa, A.C.; Martinez, V.P.; Kleine-Neto, W.; de Oliveira, A.C.S.; Sauer, M.M.; Bassichetto, K.C.; Oliveira, S.M.S.; Tomiyama, H.T.I.; et al. Characterization of partial and near full-length genomes of HIV-1 strains sampled from recently infected individuals in São Paulo, Brazil. *PLoS ONE* **2011**, *6*, e025869. [CrossRef] [PubMed]
99. Henn, M.R.; Boutwell, C.L.; Charlebois, P.; Lennon, N.J.; Power, K.A.; Macalalad, A.R.; Berlin, A.M.; Malboeuf, C.M.; Ryan, E.M.; Gnerre, S.; et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* **2012**, *8*, e1002529. [CrossRef]
100. Edlefsen, P.T.; Rolland, M.; Hertz, T.; Tovanabutra, S.; Gartland, A.J.; deCamp, A.C.; Magaret, C.A.; Ahmed, H.; Gottardo, R.; Juraska, M.; et al. Comprehensive Sieve Analysis of Breakthrough HIV-1 Sequences in the RV144 Vaccine Efficacy Trial. *PLoS Comput. Biol.* **2015**, *11*, 1–37. [CrossRef]
101. An, M.; Han, X.; Xu, J.; Chu, Z.; Jia, M.; Wu, H.; Lu, L.; Takebe, Y.; Shang, H. Reconstituting the Epidemic History of HIV Strain CRF01_AE among Men Who Have Sex with Men (MSM) in Liaoning, Northeastern China: Implications for the Expanding Epidemic among MSM in China. *J. Virol.* **2012**, *86*, 12402–12406. [CrossRef]
102. Sanchez-Pescador, R.; Power, M.D.; Barr, P.J.; Steimer, K.S.; Stempien, M.M.; Brown-Shimer, S.L.; Gee, W.W.; Renard, A.; Randolph, A.; Levy, J.A. Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). *Science* **1985**, *227*, 484–492. [CrossRef] [PubMed]
103. Han, X.; An, M.; Zhao, B.; Duan, S.; Yang, S.; Xu, J.; Zhang, M.; McGoogan, J.M.; Takebe, Y.; Shang, H. High Prevalence of HIV-1 Intersubtype B′/C Recombinants among Injecting Drug Users in Dehong, China. *PLoS ONE* **2013**, *8*, e065337. [CrossRef] [PubMed]
104. Salgado, M.; Swanson, M.D.; Pohlmeyer, C.W.; Buckheit, R.W.; Wu, J.; Archin, N.M.; Williams, T.M.; Margolis, D.M.; Siliciano, R.F.; Garcia, J.V.; et al. HLA-B*57 Elite Suppressor and Chronic Progressor HIV-1 Isolates Replicate Vigorously and Cause CD4+ T Cell Depletion in Humanized BLT Mice. *J. Virol.* **2014**, *88*, 3340–3352. [CrossRef]

105. Cho, Y.K.; Kim, J.E.; Foley, B.T. Phylogenetic analysis of near full-length HIV type 1 genomic sequences from 21 korean individuals. *AIDS Res. Hum. Retrovir.* **2013**, *29*, 738–743. [CrossRef] [PubMed]
106. Pessôa, R.; Watanabe, J.T.; Calabria, P.; Alencar, C.S.; Loureiro, P.; Lopes, M.E.; Proetti, A.B.; Félix, A.C.; Ester, C. International Component of the NHLBI Epidemiology and Donor Evaluation Study-III (REDS-III) Enhanced detection of viral diversity using partial and near full- length genomes of HIV-1 provirus deep sequencing data from recently infected donors at four blood centers in Brazil. *Transfusion* **2015**, *55*, 980–990. [CrossRef] [PubMed]
107. Grossmann, S.; Nowak, P.; Neogi, U. Subtype-independent near full-length HIV-1 genome sequencing and assembly to be used in large molecular epidemiological studies and clinical management. *J. Int. AIDS Soc.* **2015**, *18*, 1–8. [CrossRef] [PubMed]
108. Blanco, M.; Machado, L.Y.; Díaz, H.; Ruiz, N.; Romay, D.; Silva, E. HIV-1 genetic variability in Cuba and implications for transmission and clinical progression. *MEDICC Rev.* **2015**, *17*, 25–31.
109. Tully, D.C.; Ogilvie, C.B.; Batorsky, R.E.; Bean, D.J.; Power, K.A.; Ghebremichael, M.; Bedard, H.E.; Gladden, A.D.; Seese, A.M.; Amero, M.A.; et al. Differences in the Selection Bottleneck between Modes of Sexual Transmission Influence the Genetic Composition of the HIV-1 Founder Virus. *PLoS Pathog.* **2016**, *12*, 1–29. [CrossRef]
110. Pessôa, R.; Loureiro, P.; Lopes, M.E.; Carneiro-Proietti, A.B.F.; Sabino, E.C.; Busch, M.P.; Sanabani, S.S. Ultra-deep sequencing of HIV-1 near full-length and partial proviral genomes reveals high genetic diversity among Brazilian blood donors. *PLoS ONE* **2016**, *11*, e0152499. [CrossRef]
111. Bruner, K.M.; Murray, A.J.; Pollack, R.A.; Soliman, M.G.; Sarah, B.; Capoferri, A.A.; Lai, J.; Strain, M.C.; Lada, S.M.; Ho, Y.; et al. Defective proviruses rapidly accumulate during acute HIV-1 infection Katherine. *Nat. Med.* **2016**, *22*, 1043–1049. [CrossRef]
112. Cevallos, C.G.; Jones, L.R.; Pando, M.A.; Carr, J.K.; Avila, M.M.; Quarleri, J. Genomic characterization and molecular evolution analysis of subtype B and BF recombinant HIV-1 strains among Argentinean men who have sex with men reveal a complex scenario. *PLoS ONE* **2017**, *12*, e0189705. [CrossRef] [PubMed]
113. Ghosh, S.K.; Fultz, P.N.; Keddie, E.; Saag, M.S.; Sharp, P.M.; Hahn, B.H.; Shaw, G.M. A Molecular Clone of HIV-1 Tropic and Cytopathic for Human and Chimpanzee Lymphocytes. *Virology* **1993**, *194*, 858–864. [CrossRef]
114. Neogi, U.; Siddik, A.B.; Kalaghatgi, P.; Gisslén, M.; Bratt, G.; Marrone, G.; Sönnerborg, A. Recent increased identification and transmission of HIV-1 unique recombinant forms in Sweden. *Sci. Rep.* **2017**, *7*, 1–9. [CrossRef] [PubMed]
115. Kreutz, R.; Dietrich, U.; Kuhnel, H.; Nieselt-Struwe, K.; Eigen, M.; Rubsamen-Waigmann, H. Analysis of the envelope region of the highly divergent HIV-2ALT isolate extends the known range of variability within the primate immunodeficiency viruses. *AIDS Res. Hum. Retrovir.* **1992**, *8*, 1619–1629. [CrossRef] [PubMed]
116. Carr, J.K.; Laukkanen, T.; Salminen, M.O.; Albert, J.; Alaeus, A.; Kim, B.; Sanders-Buell, E.; Birx, D.L.; McCutchan, F.E. Characterization of subtype A HIV-1 from Africa by full genome sequencing. *Aids* **1999**, *13*, 1819–1826. [CrossRef] [PubMed]
117. Laukkanen, T.; Albert, J.; Liitsola, K.; Green, S.D.; Carr, J.K.; Leitner, T.; McCutchan, F.E.; Salminen, M.O. Virtually full-length sequences of HIV type 1 subtype J reference strains. *AIDS Res. Hum. Retrovir.* **1999**, *15*, 293–297. [CrossRef] [PubMed]
118. Hoelscher, M.; Kim, B.; Maboko, L.; Mhalu, F.; Von Sonnenburg, F.; Birx, D.L.; McCutchan, F.E. High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. *Aids* **2001**, *15*, 1461–1470. [CrossRef]
119. Novitsky, V.; Smith, U.R.; Gilbert, P.; McLane, M.F.; Chigwedere, P.; Williamson, C.; Ndung'u, T.; Klein, I.; Chang, S.Y.; Peter, T.; et al. Human Immunodeficiency Virus Type 1 Subtype C Molecular Phylogeny: Consensus Sequence for an AIDS Vaccine Design? *J. Virol.* **2002**, *76*, 5435–5451. [CrossRef]
120. Harris, M.E.; Serwadda, D.; Sewankambo, N.; Kim, B.; Kigozi, G.; Kiwanuka, N.; Phillips, J.B.; Wabwire, F.; Meehen, M.; Lutalo, T.; et al. Among 46 near full length HIV type 1 genome sequences from Rakai District, Uganda, subtype D and AD recombinants predominate. *AIDS Res. Hum. Retrovir.* **2002**, *18*, 1281–1290. [CrossRef]
121. Triques, K.; Bourgeois, A.; Vidal, N.; Mpoudi-Ngole, E.; Mulanga-Kabeya, C.; Nzilambi, N.; Torimiro, N.; Saman, E.; Delaporte, E.; Peeters, M. Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res. Hum. Retrovir.* **2000**, *16*, 139–151. [CrossRef]

122. Arroyo, M.A.; Hoelscher, M.; Sanders-Buell, E.; Herbinger, K.-H.; Samky, E.; Maboko, L.; Hoffmann, O.; Robb, M.R.; Birx, D.L.; McCutchan, F.E. HIV type 1 subtypes among blood donors in the Mbeya region of southwest Tanzania. *AIDS Res. Hum. Retrovir.* **2004**, *20*, 895–901. [CrossRef] [PubMed]
123. Kijak, G.H.; Sanders-Buell, E.; Wolfe, N.D.; Mpoudi-Ngole, E.; Kim, B.; Brown, B.; Robb, M.L.; Birx, D.L.; Burke, D.S.; Carr, J.K.; et al. Development and application of a high-throughput HIV type 1 genotyping assay to identify CRF02_AG in West/West Central Africa. *AIDS Res. Hum. Retrovir.* **2004**, *20*, 521–530. [CrossRef] [PubMed]
124. Carrion, G.; Eyzaguirre, L.; Montano, S.M.; Laguna-Torres, V.; Serra, M.; Aguayo, N.; Avila, M.M.; Ruchansky, D.; Pando, M.A.; Vinoles, J.; et al. Documentation of subtype C HIV Type 1 strains in Argentina, Paraguay, and Uruguay. *AIDS Res. Hum. Retrovir.* **2004**, *20*, 1022–1025. [CrossRef] [PubMed]
125. Qiu, Z.; Xing, H.; Wei, M.; Duan, Y.; Zhao, Q.; Xu, J.; Shao, Y. Characterization of five nearly full-length genomes of early HIV type 1 strains in Ruili city: Implications for the genesis of CRF07_BC and CRF08_BC circulating in China. *AIDS Res. Hum. Retrovir.* **2005**, *21*, 1051–1056. [CrossRef] [PubMed]
126. Rousseau, C.M.; Birditt, B.A.; McKay, A.R.; Stoddard, J.N.; Lee, T.C.; McLaughlin, S.; Moore, S.W.; Shindo, N.; Learn, G.H.; Korber, B.T.; et al. Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. *J. Virol. Methods* **2006**, *136*, 118–125. [CrossRef] [PubMed]
127. Abecasis, A.B.; Lemey, P.; Vidal, N.; de Oliveira, T.; Peeters, M.; Camacho, R.; Shapiro, B.; Rambaut, A.; Vandamme, A.-M. Recombination Confounds the Early Evolutionary History of Human Immunodeficiency Virus Type 1: Subtype G Is a Circulating Recombinant Form. *J. Virol.* **2007**, *81*, 8543–8551. [CrossRef] [PubMed]
128. Tovanabutra, S.; Sanders, E.J.; Graham, S.M.; Mwangome, M.; Peshu, N.; McClelland, R.S.; Muhaari, A.; Crossler, J.; Price, M.A.; Gilmour, J.; et al. Evaluation of HIV type 1 strains in men having sex with men and in female sex workers in Mombasa, Kenya. *AIDS Res. Hum. Retrovir.* **2010**, *26*, 123–131. [CrossRef]
129. Treurnicht, F.K.; Seoighe, C.; Martin, D.P.; Wood, N.; Abrahams, M.-R.; Rosa, D.d.A.; Bredell, H.; Woodman, Z.; Hide, W.; Mlisana, K.; et al. Adaptive changes in HIV-1 subtype C proteins during early infection are driven by changes in HLA-associated immune pressure. *Virology* **2010**, *396*, 213–225. [CrossRef]
130. Wilkinson, E.; Holzmayer, V.; Jacobs, G.B.; De Oliveira, T.; Brennan, C.A.; Hackett, J.; Van Rensburg, E.J.; Engelbrecht, S. Sequencing and phylogenetic analysis of near full-length HIV-1 subtypes A, B, G and unique recombinant AC and AD viral strains identified in South Africa. *AIDS Res. Hum. Retrovir.* **2015**, *31*, 412–420. [CrossRef]
131. Tongo, M.; Dorfman, J.R.; Abrahams, M.R.; Mpoudi-Ngole, E.; Burgers, W.A.; Martin, D.P. Near full-length HIV type 1M genomic sequences from Cameroon. *Evol. Med. Public Heal.* **2015**, *2015*, 254–265. [CrossRef]
132. Billings, E.; Sanders-Buell, E.; Bose, M.; Bradfield, A.; Lei, E.; Kijak, G.H.; Arroyo, M.A.; Kibaya, R.M.; Scott, P.T.; Wasunna, M.K.; et al. The number and complexity of pure and recombinant HIV-1 strains observed within incident infections during the HIV and malaria cohort study conducted in Kericho, Kenya, from 2003 to 2006. *PLoS ONE* **2015**, *10*, e0135124. [CrossRef] [PubMed]
133. Amogne, W.; Bontell, I.; Grossmann, S.; Aderaye, G.; Lindquist, L.; Sonnerborg, A.; Neogi, U. Phylogenetic Analysis of Ethiopian HIV-1 Subtype C Near Full-Length Genomes Reveals High Intrasubtype Diversity and a Strong Geographical Cluster. *AIDS Res. Hum. Retrovir.* **2016**, *32*, 471–474. [CrossRef] [PubMed]
134. Chen, Y.; Hora, B.; DeMarco, T.; Shah, S.A.; Ahmed, M.; Sanchez, A.M.; Su, C.; Carter, M.; Stone, M.; Hasan, R.; et al. Fast dissemination of new HIV-1 CRF02/A1 recombinants in Pakistan. *PLoS ONE* **2016**, *11*, e0167839. [CrossRef]
135. Billings, E.; Sanders-Buell, E.; Bose, M.; Kijak, G.H.; Bradfield, A.; Crossler, J.; Arroyo, M.A.; Maboko, L.; Hoffmann, O.; Geis, S.; et al. HIV-1 Genetic Diversity Among Incident Infections in Mbeya, Tanzania. *AIDS Res. Hum. Retrovir.* **2017**, *33*, 373–381. [CrossRef]
136. Rodgers, M.A.; Wilkinson, E.; Vallari, A.; McArthur, C.; Sthreshley, L.; Brennan, C.A.; Cloherty, G.; de Oliveira, T. Sensitive Next-Generation Sequencing Method Reveals Deep Genetic Diversity of HIV-1 in the Democratic Republic of the Congo. *J. Virol.* **2017**, *91*, 1–18. [CrossRef] [PubMed]
137. Huang, W.; Li, L.; Myers, J.R.; Marth, G.T. ART: A next-generation sequencing read simulator. *Bioinformatics* **2012**, *28*, 593–594. [CrossRef]
138. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]

139. Rozas, J.; Ferrer-Mata, A.; Sanchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sanchez-Gracia, A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [CrossRef]
140. Strimmer, K.; von Haeseler, A.; Salemi, M. Genetic distances and nucleotide substitution models. *Phylogenetic Handb.* **2012**, 111–141. [CrossRef]
141. R Core Team. *R: A Language and Environment for Statistical Computing*; RC Team: Vienna, Austria, 2013.
142. RStudio Team. *RStudio: Integrated Development for R.*; RStudio Team: Boston, MA, USA, 2015.
143. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
144. McDonald, J. *Handbook of Biological Statistics*, 3rd ed.; Sparky House Publishing: Baltimore, Maryland, 2014.
145. Kruskal, W.H.; Wallis, A.W. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
146. Dinno, A. Dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums. *R Package Version* **2017**, *1*.
147. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J. Stat.* **1979**, *6*, 65–70.
148. Woolson, R.F. Wilcoxon signed-rank test. In *Wiley Encyclopedia of Clincal Trials*; D'Agostino, R.B., Sullivan, L., Massaro, J., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2008; pp. 7–9.
149. Pineda-Peña, A.C.; Faria, N.R.; Imbrechts, S.; Libin, P.; Abecasis, A.B.; Deforche, K.; Gómez-López, A.; Camacho, R.J.; De Oliveira, T.; Vandamme, A.M. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* **2013**, *19*, 337–348. [CrossRef] [PubMed]
150. Kuiken, C.; Combet, C.; Bukh, J.; Shin-I, T.; Deleage, G.; Mizokami, M.; Richardson, R.; Sablon, E.; Yusim, K.; Pawlotsky, J.M.; et al. A comprehensive system for consistent numbering of HCV sequences, proteins and epitopes. *Hepatology* **2006**, *44*, 1355–1361. [CrossRef]
151. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265–269. [CrossRef]
152. Crandall, K.A.; Posada, D.; Vasco, D. Effective population sizes: Missing measures and missing concepts. *Anim. Conserv.* **1999**, *2*, 317–319. [CrossRef]
153. Nickle, D.C.; Jensen, M.A.; Shriner, G.S.G.D.; Learn, G.H.; Rodrigo, A.G.; Mullins, J.I. Consensus and Ancestral State HIV Vaccines. *Science* **2003**, *299*, 1515–1519. [CrossRef] [PubMed]
154. Gibson, K.; Jair, K.; Castel, A.D.; Bendall, M.L.; Wilbourn, B.; Jordan, J.A.; Crandall, K.A.; Pérez-Losada, M.; the DC Cohort Executive Committee. A cross-sectional study to characterize local HIV-1 dynamics in Washington, DC using next-generation sequencing. *Sci. Rep.* **2020**, *10*, 1–18. [CrossRef] [PubMed]
155. Liu, T.; Shafer, R. Web Resources for HIV type 1 Genotypic-Resistance Test Interpretation. *Clin. Infect. Dis* **2006**, *42*, 1608–1618. [CrossRef]
156. Rhee, S.Y. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **2003**, *31*, 298–303. [CrossRef]
157. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef] [PubMed]
158. Salmela, L.; Sahlin, K.; Mäkinen, V.; Tomescu, A.I. Gap filling as exact path length problem. *J. Comput. Biol.* **2016**, *23*, 347–361. [CrossRef] [PubMed]
159. Yang, X.; Charlebois, P.; Gnerre, S.; Coole, M.G.; Lennon, N.J.; Levin, J.Z.; Qu, J.; Ryan, E.M.; Zody, M.C.; Henn, M.R. De novo assembly of highly diverse viral populations. *BMC Genom.* **2012**, *13*. [CrossRef] [PubMed]
160. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]