Reference plasmid pHXB2_D is an HIV-1 molecular clone that exhibits identical LTRs and a single integration site indicative of an HIV provirus

Alejandro R. Gener[1,2,3,4§], Brian T. Foley[5], Deborah P. Hyink[2], Paul E. Klotman[2]

[1]Integrative Molecular and Biomedical Sciences Program, Baylor College of Medicine, Houston, Texas, USA
[2]Margaret M. and Albert B. Alkek Department of Medicine, Nephrology, Baylor College of Medicine, Houston, Texas, USA
[3]Department of Genetics, MD Anderson Cancer Center, Houston, Texas, USA
[4]School of Medicine, Universidad Central del Caribe, Bayamón, Puerto Rico, USA
[5]Theoretical Biology and Biophysics Group T-6, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

[§]Corresponding author: Alejandro R. Gener

One Baylor Plaza
Mail Stop 710
Houston, Texas, 77030, USA
9045715562
gener@bcm.edu ; itspronouncedhenner@gmail.com

Keywords: HIV-1, reagent verification, nanopore DNA sequencing, provirus, plasmid, sequence variability, resequencing

Abstract

Objective: To compare long-read nanopore DNA sequencing (DNA-seq) with short-read sequencing-by-synthesis for sequencing a full-length (e.g., non-deletion, nor reporter) HIV-1 model provirus in plasmid pHXB2_D.

Design: We sequenced pHXB2_D with long- and short-read DNA-seq, evaluating sample variability with resequencing (sequencing and mapping to reference HXB2) and *de novo* viral genome assembly.

Methods: We prepared pHXB2_D for long-read nanopore DNA-seq, varying DNA polymerases Taq (Sigma-Aldrich) and Long Amplicon (LA) Taq (Takara). Nanopore basecallers were compared. After aligning reads to the reference HXB2 to evaluate sample coverage, we looked for variants. We next assembled reads into contigs, followed by finishing and polishing. We hired an external core to sequence-verify pHXB2_D with single-end 150 base-long Illumina reads, after masking sample identity.

Results: We achieved full-coverage (100%) of HXB2 HIV-1 from 5' to 3' long terminal repeats (LTRs), with median per-base coverage of over 9000x in one experiment on a single MinION flow cell. The longest HIV-spanning read to-date was generated, at a length of 11,487 bases, which included full-length HIV-1 and plasmid backbone with flanking host sequences supporting a single HXB2 integration event. We discovered 20 single nucleotide variants in pHXB2_D compared to reference, verified by short-read DNA sequencing.

Conclusions: Nanopore sequencing performed as-expected, even covering full-length HIV. The discovery of variants in a reference plasmid demonstrates the need for sequence verification moving forward, in line with calls from funding agencies for reagent verification. These results

23 illustrate the utility of long-read DNA-seq to advance the study of HIV at single integration site

24 resolution.

25

Introduction

Much of what we know about human acquired immunodeficiency syndrome (AIDS) came after isolating the causative agent – the human immunodeficiency virus type 1 (HIV-1) – and describing the viral genome information content. The HIV-1 isolate HXB2 (also known as HTLV-III and HIV-1_{LAI} or LAV/BRU [1], [2]) was the first full-length replication-competent HIV genome sequenced [3]. Derivative clones commonly called "HXB2" are still used for *in vitro* infection assays, including RNA (almost always cDNA [4]) sequencing (**Figure 1A** and **Supplemental Table 1**). Despite the availability of the HXB2 HIV-1 reference sequence [3], no sequence is available for any complete and readily available HXB2 clone.

HIV clones were originally made by choosing non-cutter restriction enzymes to digest intact proviral sequences upstream and downstream of unknown integration sites from infected host cells, followed by ligation into an *E. coli* cloning vector (plasmid) (**Figure 1B**), allowing for low-error (but not error-free) propagation [5]. These clones became available before tractable sequencing methods permitted routine sequence verification. As such, it was uncommon to sequence them. While funding agencies now require investigators to include in their proposals plans to validate their key reagents, these funders tend to leave the process up to investigators and may not always follow up on whether a given reagent is ever actually validated (or revalidated between changes of hand). Investigators do not regularly validate their clones, in part because there is no universally accepted standard. Instead, a common practice is to assume a given clone, often kindly gifted from a colleague, is as reported. As such, we often do not truly know what we have been working with for 35+ years.

Making sense of the information from HIV sequencing experiments is complicated by many factors, including the cycling that all orterviruses [6] undergo between two major states (as

49 infectious virion RNA and integrated proviral DNA **Figure 1B**), repetitive viral sequences like
50 long terminal repeats (LTRs), non-integrated forms [7], rarity of integration events *in vivo*
51 (reviewed in [8]), and alternative splicing of viral mRNAs [9]. Short-read DNA sequencing
52 (<150 base pairs (bp) in most reported experiments, but up to 500 bp for either Illumina
53 sequencing-by-synthesis or <1,000 bp for chain termination sequencing) provides some
54 information, but analyses require high coverage and/or extensive effort (non-exhaustive
55 examples [10], [11]). These factors limit the ability to assign variants to specific loci within each
56 provirus, as well as at the proviral integration site(s) (reviewed in [12]). Despite progress (HIV
57 DNA) [13], (HIV RNA) [14], [15], researchers have yet to observe the genome of HIV-1 as
58 complete provirus (integrated DNA) in a single read, hindering locus-specific studies. To this
59 end, current long-read DNA sequencing clearly surpasses the limitations of read length of
60 leading next-generation/short-read sequencing platforms. Here we used the MinION sequencer
61 to sequence HIV-1 plasmid pHXB2_D in a pilot study focusing on coverage acquisition (as
62 opposed to full-length sequencing), with the goal of evaluating the technology for future
63 applications.

64

65

66 # Methods

67 This work did not include human or animal subjects. Nanopore libraries for this work
68 were prepared in their entirety by ARG in a Biosafety Level 2 laboratory on main campus at
69 Baylor College of Medicine. Nanopore sequencing was completed between April and May of
70 2018 as two of several control experiments included in the Student Genomics pilot run
71 (**Supplemental Information**). Short-read sequencing was completed in April 2019.

72 **HIV-1 plasmid**

73 A plasmid, "pHXB2_D" (alternate names pHXB2, pHXB-2D), believed to contain the
74 HIV-1 reference strain HXB2 [16] was acquired from the NIH AIDS Reagent and Reference
75 Program (ARP) via BioServe. pHXB2_D is believed to be a molecular clone (likely a restriction
76 product of HXB2 proviral DNA inserted into an unknown cloning plasmid backbone) from one
77 of the earliest clinical "HXB2" HIV-1 isolates. At the time of this work, it was unknown whether
78 this plasmid was ever sequence-verified before or after the reference sequence for HXB2 was
79 deposited.

80 **HIV-1 reference sequence**

81 The reference sequence of HXB2 is from the National Center for Biotechnology
82 Information (NCBI), Genbank accession number K03455.1. It runs from the beginning of the 5'
83 LTR to the end of the 3' LTR, and is 9,719 bp.

84 **Long-read DNA sequencing**

85 A plasmid containing HXB2 was sequence-verified with long-read nanopore sequencing
86 on a MinION Mk1B (Oxford Nanopore Technologies, Oxford, UK). Unless otherwise noted,
87 reagents (and software) were purchased (or acquired) from Oxford Nanopore. Briefly, stock

88 plasmid was diluted to 5 ng final amount in ultrapure water (as two samples) and processed with
89 Rapid PCR Barcoding kit SQK-RPB004 along with 10 other barcoded samples (not discussed
90 further in this manuscript) following ONT protocol RPB_9059_V1_REVA_08MAR2018
91 (**Figure 1C**), a public description of which is here: https://store.nanoporetech.com/us/sample-
92 prep/rapid-pcr-barcoding-kit.html. Two DNA polymerases were evaluated (barcode 10 used
93 high-fidelity LA (for "long amplicon") Taq (Takara); barcode 11 Taq (Sigma-Aldrich). Libraries
94 were loaded onto a MinION flow cell version R9.4.1 and a 48-hour sequencing run was
95 completed with MinKNOW (version 1.10.11). Residual reads from subsequent runs were pooled
96 for final analyses.

97 Raw data was basecalled (converted from FAST5 to FASTQ format) with Albacore
98 version 2.3.4 (older basecaller), Guppy version 2.3.1 (current official at time of work), and
99 FlipFlop (Guppy development config). Mapping to reference was done with Minimap2 [17] and
100 BWA-MEM [18], implemented in Galaxy (usegalaxy.org) [19]. Alignments (.bam and .bai files)
101 were visualized in the Integrative Genomics Viewer [20] unless otherwise noted. For *de novo*
102 assembly, demultiplexed basecalled reads were fed into Canu version 1.8 [21]. Genome size was
103 estimated to be 16 Kb from agarose gel of undigested, but naturally degraded linearized
104 pHXB2_D (data not shown). SnapGene version 4.3.4 was used to manually annotate contigs
105 from Canu. Blastn (NCBI) was used to identify unknown regions of pHXB2_D. Polishing was
106 performed on ONT-only assemblies with Medaka (https://github.com/nanoporetech/medaka), in
107 Galaxy. Medaka models: r941_min_fast_g303, r941_min_high_g303, r941_min_high_g330.
108 Inference batch size (-b) = 100. The final pHXB2_D assembly and other full-length HIV clones
109 from the ARP were aligned to the most recent human reference genome (hg38) with Minimap2

110 in Galaxy with the following parameters: Long assembly to reference mapping (-k19 -w19 -A1 -
111 B19 -O39,81 -E3,1 -s200 -z200 --min-occ-floor=100).

112 **Statistics**

113 Two-tailed Mann-Whitney U tests were used to compare distributions in long-read data.
114 P-values are reported over brackets delineating relevant comparisons. Calculations and graphing
115 were done with GraphPad Prism for macOS version 8.0.2.

116 **Short-read DNA sequencing**

117 pHXB2_D was provided as 35 ul at 63 ng/ul to the Center for Computational &
118 Integrative Biology DNA Core at Massachusetts General Hospital, an external DNA sequencing
119 core specializing in high-throughput next generation (short-read) plasmid sequencing and
120 assembly. Neither HXB2 reference sequence nor pHXB2_D draft assemblies (from this work)
121 were provided to core staff at the time of sequencing so that testing would remain masked. While
122 the core's exact library prep is proprietary, multiplexed library prep and 150 single-end Illumina
123 (ILMN) sequencing were most likely performed on a MiSeq with platform-specific reagents (V2
124 chemistry, per their website) and barcoding. Data was returned as FASTQ. FASTQC [22] was
125 used in Galaxy for in-house data quality control, and read lengths were all 142 bp per this tool.
126 Mapping as above.

127 **Sequence comparisons**

128 We used MAFFT v7.475 [23], [24] to compare the LTR sequences of pHXB2_D and
129 HXB2. For cladistics, we used BLAST at HIV-DB
130 (https://www.hiv.lanl.gov/content/sequence/BASIC_BLAST/basic_blast.html) to find other
131 HXB2-like genomes. The top 50 BLAST hits included many sequences pNL43 clones. pNL4-3

132 is an artificial recombinant of the NY5 clone with the HXB2 clone. We then made a multi-
133 sequence alignment with the final pHXB2_D assembly, the top BAST hits, and the HIV-1 M
134 group subtype reference set using GeneCutter
135 (https://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html), and built the
136 maximum likelihood tree using IQ-tree
137 (https://www.hiv.lanl.gov/content/sequence/IQTREE/iqtree.html).

138

Results

Viewing mapped data in IGV, the long reads (median read length >2000 bp, **Figures 1E**) from both ONT experiments clearly covered each LTR (**Figure 1F, Supplemental Figures 1**, **2**), while shorter reads collapsed into one of either LTR (**Figures 1F, Supplemental Figures 2D,2E**). This was also seen when long reads were shorter than LTRs (<600 bp). Mappers BWA-MEM and Minimap2 were chosen based on their ability to handle long and short reads. Other mappers were not evaluated. BWA-MEM mapped more ambiguously, piling partially mapped reads between each LTR; Minimap2 mapped with higher fidelity to reference without splitting reads. Coverage as sequencing depth was higher and more even from the higher-fidelity LA Taq library (**Supplemental Figure 1**).

We counted 20 single nucleotide variants (SNVs) in this reference clone of HXB2 (**Table 1, Supplemental Table 3, Supplemental Figure 2E**). These mismatches were seen in all Canu assemblies (**Supplemental Figure 3A,3B**), verified in IGV and/or SnapGene, and were orthogonally verified by short-read sequencing performed by the external core given masked samples (**Supplemental Figure 2E**). These mismatches represent a ~0.21% divergence from reference HXB2 K03455.1 (20/9719), which was assumed to have perfect identity (0% divergence). Transitions were more common (14/20) (**Table 1**), coinciding with a previous report of increased transitions over transversions in infection models, because transversions are more likely to be deleterious to viral replication (i.e.: to cause protein-coding changes) [25]. Indeed, almost half (9/20) of the observed SNVs occurred in protein-coding regions, even though 92% of HXB2 is coding (791/9719). Of those 9 SNVs in protein-coding regions, 4 caused non-synonymous mutations. One of those occurs in a region overlapping both gag and pol regions, however only pol exhibited a non-synonymous change from valine to isoleucine in p6, at

162 position 2259 relative to HXB2. Other non-synonymous variants occurred at 4609 (in p31
163 integrase, arginine to lysine), 7823 (in ASP antisense protein, glycine to arginine), and 9253 (in
164 nef, isoleucine to valine). 11/20 SNVs were in LTRs (see **Supplemental Figure 2** for counting
165 based on mapping); 8/20 of these would have been missed with mapping-only variant calling or
166 consensus. The longest HIV-mapping read (**Figure 2**) phased 16/20 SNVs (failed at sites
167 2,8,10,12, **Table 1**).

168 We assembled the previously undefined plasmid pHXB2_D (**Supplemental Figure**
169 **3A,3B**). Canu's final output was a set of contiguous DNA sequences (contigs) as FASTA files. A
170 consequence of assembling plasmid sequences with this tool was partial redundancy at contig
171 ends (**Supplemental Figure 3C)**. Manual end-trimming of contigs was performed in SnapGene
172 based on an estimated length of 16 kilobases. Top blastn hits from barcode 10/LA Taq pHXB2
173 basecalled with FlipFlop were as follows: for the main backbone (with origin of replication and
174 antibiotic selection cassette for cloning), shuttle vector pTB101-CM DNA, complete sequence
175 (based on pBR322), from 4352-8340; for the upstream element (relative to 5' LTR), Homo
176 sapiens chromosome 3 clone RP11-83E7 map 3p, complete sequence from 58,052 to 59,165; for
177 the downstream element, cloning vector pNHG-CapNM from 10,204 to 11,666. Other identified
178 elements included Enterobacteria phage SP6 (the SP6 promoter, per SnapGene's "Detect
179 common features"), complete sequence from 39,683 to 39,966. Identities of query to HXB2 and
180 hits were all approximately 99%. The MGH CCIB DNA Core's proprietary *de novo* UltraCycler
181 v1.0 assembler (Brian Seed and Huajun Wang, unpublished) was able to assemble both 5' and 3'
182 LTRs with short-read data only but may have collapsed SNVs into an artificial single consensus.
183 Long-read mapping and assembly (and polished assemblies) supported a single HIV-1 HXB2_D
184 haplotype (**Supplemental Figure 3,5**); a final LTR-phased and annotated assembly leveraging

185 short and long reads is provided as pHXB2_D Genbank:MW079479 (embargoed until
186 publication). Importantly, for pHXB2_D, each LTR was identical, which is distinct from the
187 current HXB2 (K03455.1) (**Figure 3**). To determine whether pHXB2_D was an isolated provirus
188 (as opposed to a cDNA clone), the pHXB2_D assembly was aligned to the current human
189 reference hg38, returning a single complete insertion site on 3p24.3 (**Figure 4, Supplemental**
190 **Table 2**). We sought to put our pHXB2_D assembly into context of other HXB2-like references
191 available (**Figure 5**). pHXB2_D (red) clusters closely with HXB2 reference (K03455) and
192 related clone sequences (green). pNL4-3 clones in blue. The LTR-masked HIV-spanning
193 segment of pHXB2_D is most homologous to B.FR.1983.DM461230 and B.FR.1983.CS793683,
194 which are identical except for areas in nef and a GFP insertion (verified by blastn). This finding
195 suggests they were from the same stock. HIV-1 M group subtype reference set (HIV Sequence
196 Database) was added to put HXB2s and pNL4-3 clones into perspective. HXB2 (believed to be a
197 complete isolate) and NL4-3 (synthetic clone based on two early isolates) are examples of HIV
198 type 1 (HIV-1), group M, subgroup B.

199 As previously reported [26], per-read variability in ONT data was higher near
200 homopolymers (runs of the same base) (**Supplemental Figure 4A**). For the datasets generated in
201 the present study, homopolymers were counted and classified as continuous (unbroken run of a
202 given nucleobase) vs. discontinuous (broken run of a given nucleobase) (**Supplemental Figure**
203 **4B,4D,4F,4H**). A/T (2 hydrogen bonds; 2H) and G/C (3 hydrogen bonds; 3H) were evaluated.
204 Because runs longer than 4 or 5 were rare in these datasets, it was impossible to evaluate longer
205 homopolymers. A simple calculation $Abs(\Delta)=Abs(\#homopolymers_{reference} -$
206 $\#homopolymers_{assembly})$ helped to evaluate the performance of basecallers, such that better
207 basecallers had smaller Abs(Δ) (**Supplemental Figure 4C,4E,4G,4I,4K**). At the level of

consensus (made from sequences mapped to reference HXB2), homopolymers contributed few, if any, obvious errors. A special case of homopolymer, dimer runs, was noted to cause persistent errors regardless of ONT basecaller (**Supplemental Figure 4J,4K**). While dips occurred at certain points near homopolymers, the consensus did not change much at the sequencing depth used in this study for either barcoded pHXB2_D samples (**Supplemental Figures 1,2,3**). Another interpretation is that homopolymers tend to seem truncated with ONT, with more reads in support of shorter homopolymers. Canu assemblies showed basecaller-dependent variability (**Supplemental Table 3**). That said, newer basecallers tended to produce fewer and smaller per-read truncations. Assemblies without polishing did not correct all homopolymer truncations (**Supplemental Figure 3A**). Polishing assemblies tended to correct these toward the final pHXB2_D assembly (**Supplemental Figure 3B,5**). Data from polished ONT-only assemblies and short-read sequencing do not support the truncations (gaps relative to reference) suggested by unpolished ONT-only assemblies, representing a known current limitation of ONT. These are not the same as the 20 SNVs supported by BOTH long- and short-read sequencing performed in this study. The ratio of per-read deletions to per-read insertions (DEL/INS) was much higher for SNVs occurring at homopolymers and near the same base, and this difference was maintained between all basecallers used (**Supplemental Figure 4L**). These changes created more problematic (longer) homopolymers.

228 ## Discussion

229 This work represents the first instance of complete and unambiguous sequencing of HIV-
230 1 provirus as plasmid and contributed to the identification of single nucleotide variants which
231 may not have been easily determined using other sequencing modalities, illustrating the
232 importance of validating molecular reagents in their entirety, and with complementary
233 approaches. Nanopore sequencing surpassed the read length limitations of traditional sequencing
234 modalities used for HIV such as Sanger sequencing and sequencing-by-synthesis by at least two
235 orders of magnitude. Other long-read DNA sequencing technologies such as PacBio's zero-mode
236 waveguide DNA sequencing were not evaluated in this work, but in principle would be
237 interchangeable for nanopore sequencing. Paired-end sequencing (as either DNA-seq or RNA-
238 seq) was not evaluated in this work, but has shown promise phasing LTRs in our hands [27]–
239 [29].

240 **First complete pass over all HIV information in reference plasmid pHXB2_D**

241 HIV provirus is believed to occur naturally as one or a few copies of reverse-transcribed
242 DNA forms integrated into the host nuclear genome. Depending on where integration occurs,
243 local GC or AT content might cause problems for detecting integrants with PCR. HIV also has
244 conserved transitions from areas of higher GC content (~60%) to content approximating average
245 human GC content (~40%). To limit PCR sequencing bias and to accommodate for the potential
246 heterogeneity of HIV sequences, we fractionated whole sample directly (as opposed to PCR-
247 barcoding select amplicons) with tagmentation provided in the Rapid PCR-Barcoding kit (ONT).
248 Tagmentation uses transposon-mediated cleavage and ligation of barcode adapters for later PCR
249 amplification. A consequence of this fractionation was a distribution of reads (**Figure 1E**)
250 shorter than longer reads reported elsewhere for ONT experiments [30]. Based on this

251 distribution and the level of coverage, it was expected that HIV might be covered from end to
252 end, but this would have been exceptional. That said, an example is presented here (**Figure 2**).
253 The provirus status of pHXB2_D is supported by recovery of both upstream and downstream
254 homology arms which map to a single human integration site.

255 **Long reads enable LTR discrimination and HIV haplotype definition**

256 We created 6 assemblies for pHXB2_D from ONT-only data (**Supplemental Figure 3**),
257 each with a common set of 20 SNVs (11 in LTRs), and 1 final assembly (a single HIV-1
258 HXB2_D haplotype) leveraging long- and short-read data. The external core's *de novo* assembly
259 pipeline identified the same 20 SNVs, and variants in the LTRs were supported by ONT
260 unambiguously. That the core's assembler was able to phase LTR variants in these samples may
261 have been because the samples had high amounts of the same upstream and downstream
262 sequences because of coming from one plasmid. The core's assembler thus may have had
263 additional sequencing information at the edges of HXB2, helping it to map deeper into each
264 LTR. This approach would likely fail in samples with multiple integrations (as in various animal
265 models of HIV disease [31]), which have unknown upstream and downstream sequences, or in
266 samples from natural human infection, which is well known to exhibit multiple pseudo-random
267 integration sites between cells [32], [33], but with mostly single integration events per cell [8].
268 Inverse PCR (iPCR) is an alternative method [34] with its own issues (e.g., PCR biases, HIV
269 concatemers, host repeats). While current PCR reagents have extended the range of what can be
270 seen with iPCR, current approaches are likewise limited by long DNA extraction methods,
271 sample amount, and remain to be optimized. If coverage is sufficient (≥10 reads in non-
272 homopolymers and non-dimer runs), long-read sequencing can provide linked variant
273 information to individual integration sites. Identical 5' and '3 LTRs (**Figure 3**) in the context of

274 a single integration event (**Figure 4**) support this integrant being a *bona fide* provirus [35]. Other
275 proviruses also had identical LTR pairs (**Supplemental Table 2**). Technical limitations such as
276 PCR errors before earlier sequencing may explain the variability in the HXB2 reference LTRs.
277 These were sequenced at a time before paired-end 150 or long-read DNA-seq were available to
278 phase LTRs, raising the possibility that these LTRs were incorrectly annotated by depositors
279 assuming identity and copy-and-pasting the sequence of one LTR for both without being able to
280 unambiguously resolve each LTR.

281 **Mutations in a reference HIV-1 plasmid illustrate the need for reagent verification**

282 Up until 2020, HIV had been the most studied human pathogen, but reagents are not
283 routinely re(verified). The pHXB2_D sequenced was allegedly a reference plasmid, with
284 unknown divergence between the published reference HXB2. Three independent experiments
285 (two long-read with PCR-barcoded libraries made with regular and long-amplicon Taq master
286 mixes, one short-read) yielded at least 20 single nucleotide variants in pHXB2_D which differed
287 from the HXB2 reference sequence (**Table 1, Supplemental Figure 2**), which were also
288 concordant across the three basecallers used (**Supplemental Table 3**), and are therefore not PCR
289 errors. By leveraging long reads with the MinION, we were able to find mutations in highly
290 repetitive LTRs relative to HXB2 Genbank:K03455.1 which are often assumed (but until now
291 never proven) to be identical (**Table 1**, **Figure 1**, **Supplemental Figures 1, 2E**), as well as
292 mutations in protein-coding regions (**Table 1**). We were also able to confirm that the backbone
293 of this plasmid is from pSP62 [16], a pBR322 derivative with the SP6 promoter [36], aiding in
294 the continued use of this important reagent, and illustrating the need of full-length reagent
295 validation moving forward. We suggest that all clinical reagents (e.g., vectors) be sequence-

296 verified at the level of single-molecule sequencing as standard quality control to protect against
297 sample heterogeneity.

298 **Improvement in ONT basecallers over time**

299 Albacore, Guppy, and FlipFlop basecallers were compared. Each produced reads of
300 similar length distributions (relative to polymerase used), while Guppy and FlipFlop produced
301 improved and best performance relative to quality score distributions (**Figure 1D**). Interestingly,
302 while read length distributions were affected by fidelity of polymerases evaluated in this work,
303 mean quality distributions were not. This is important because of the differences in cost between
304 higher fidelity Taq and classic Taq enzymes. That said, higher fidelity LA Taq produced much
305 higher coverage compared to Taq (**Supplemental Figure 1**). In consideration of library prep,
306 choice of enzyme used should be based on the desired read-length distribution and coverage.
307 Regarding read mapping, the increase in mean quality score between these basecallers improved
308 overall mapping, in part by facilitating demultiplexing, resulting in approximately ~10%
309 increases number of reads in barcoded libraries before mapping (shift in reads from unclassified
310 to a given barcode). FlipFlop tended to handle homopolymers better than previous basecallers
311 (**Supplemental Figures 4,5**). Homopolymers in HXB2 tended to exhibit apparent deletions near
312 5' ends of homopolymers (upstream due to technical artifact from mapping), but because
313 consensus is near conserved (example, at least 80% of base in called read set is identical to
314 reference), and because short-read data lacks INDELS at these sites, it is unlikely that any of
315 these homopolymer deletions are real in these experiments. Dimer runs – stretches of repeating
316 2-mers (pronounced "two-mers") – proved challenging regardless of basecaller. Mapping as
317 above may be used to aid in manually calling these when they occur. Albacore is currently
318 deprecated, and current versions of Guppy now incorporate a version of FlipFlop called Guppy

High-ACcuracy (HAC). Guppy HAC and subsequence versions were not evaluated in this work. Polishing is becoming standard practice for processing assemblies from ONT data because it redresses most homopolymer errors propagated into long-read-only assemblies. The best manually finished and polished contig had 1 error out of 16,722 bases, illustrating the utility of ONT hardware when paired with burgeoning software.

Conclusions

HIV informatics, the study of HIV sequence information, has been limited by the common assumption that sequence fidelity exists between reference genomes available in sequence databases and similarly named HIV clones. Modern DNA sequencing methods, such as long- and short-read sequencing, are available to redress this issue. Long-read sequencing fills in gaps left behind by short-read interrogation of HIV-1. Current limitations of the approaches used in the present work to study HIV are 1.) the cost of long-read sequencing, regardless of platform, compared to the cheaper short reads from sequencing-by-synthesis, 2.) long DNA extraction methods in diseased tissue (Gener, unpublished), and 3.) the lower per-base accuracy (low-mid 90's with ONT vs. 98-99% with ILMN or newer PacBio HiFi), including difficulty near homopolymers and dimer runs (**Supplemental Figure 4**). A nontrivial but redressable limitation is availability of personnel trained to prepare sequencing libraries, to run sequencing, and to analyze results. As the price of long-read sequencing decreases, hardware and software used in basecalling and library protocols improve, and with the advent of more user-friendly tools, the cost of obtaining usable data from long reads will become negligible compared to the ability to answer historically intractable questions. This work raises the possibility of being able to detect at least some recombination events, in a reference-free manner requiring only the comparison of LTRs from the same integrants. While pHXB2_D and other HIV reference clones had identical LTR pairs, this remains to be tested in other clones, since other clones were generated with shorter sequencing methods. Future work will include optimizing DNA extraction protocols with the goal of capturing higher-coverage fuller glimpses of each HIV proviral integration site in *in vivo* HIV models and patient samples. This work has broad implications for all cells infected by both integrating and non-integrating viruses, and for the characterization of targeted regions in

348 the genome which may be recalcitrant to previous sequencing methods. Long-read sequencing is
349 an important emerging tool defining the post-scaffold genomic era, allowing for the
350 characterization of anatomical landmarks of hosts and pathogens at the genomic scale.

351 **Disclaimer**

352 Erratum: Preprint version 1 of this work [37] incorrectly cited the Integrated Genome
353 Browser for work that was completed with the Integrative Genomics Viewer. Apologies for the
354 mistake.

355 **Funding**

356 This work was funded in part by institutional support from Baylor College of Medicine;
357 the Human Genome Sequencing Center at Baylor College of Medicine; private funding by Bob
358 Ostendorf, CEO of East Coast Oils, Inc., Jacksonville, Florida; ARG's own private funding,
359 including Student Genomics (manuscripts in prep). Compute resources from the Computational
360 and Integrative Biomedical Research Center at BCM ("sphere" cluster managed by Dr. Steven
361 Ludtke) and the Department of Molecular and Human Genetics at BCM ("taco" cluster managed
362 by Mr. Tanner Beck and Dr. Charles Lin) greatly facilitated the completion of this work. ARG
363 has also received the PFLAG of Jacksonville scholarship for multiple years.

364 **Competing interests**

365 ARG received travel bursaries from Oxford Nanopore Technologies (ONT). The present
366 work was completed independently of ONT. Other authors declare no conflicts of interest.

367 **Authors' contributions**

368 ARG conceived of this project, performed experiments, analyzed results, and drafted the
369 manuscript. BTF, DPH and PK discussed data and edited the manuscript. ARG and PK provided
370 funding.

371 **Acknowledgements**

As part of a summer bioinformatics internship in the Paul E. Klotman Laboratory at Baylor College of Medicine, Akash Naik supervised by ARG performed *in silico* mapping analyses/experiments, generated and/or aided in the synthesis of **Supplemental Figure 3**, and assisted in writing relevant portions, discussing, and editing this manuscript. During a second summer internship with American Physician Scientists Association Virtual Summer Research Program, the following students were supervised by ARG helped to create **Figure 1A** and S**upplemental Table 1**: Yini Liang, Kirk Niekamp, Maliha Jeba, Delmarie M. Rivera Rodríguez. Orthogonal sequence verification was performed as a service by staff at the Center for Computational & Integrative Biology DNA Core at Massachusetts General Hospital, Boston, MA, USA.

We would like to thank the staff at the DNA Core for their exceptional services, including expert analyses and rapid turnaround time. We would like to thank Drs. Steven Richards, Qingchang Meng and the staff of the Human Genome Sequencing Center Research (HGSC) and Development (R&D) team for their earlier support in nanopore adoption. We would like to thank the team at Oxford Nanopore Technologies for their timely improvements and continued R&D. I would also like to thank Ms. Taneasha Monique Washington (current) and former members of the Paul E. Klotman lab, Dr. Gokul C. Das and Alexander Batista. I would like to thank Dr. Alana Canupp and the late Dr. Jim Maruniak for their early interest in my scientific development, and for the passion that they show in everything that they do.

Available additional files

Albacore basecalled barcode 10

Guppy basecalled barcode 10

394 FlipFlop basecalled barcode 10

395 Albacore basecalled barcode 11

396 Guppy basecalled barcode 11

397 FlipFlop basecalled barcode 11

398 Minimap2 and BWA-MEM alignments (.bam and .bai)

399 Clipboards from points of interest (verified SNVs; n=20)

400 .dna files of contigs (n=6)

401 MGH data (raw + contig)

402 Supplemental Tables

403 Supplemental Figures

404

405

406

407 **References**

408 [1] F. Barré-Sinoussi *et al.*, "Isolation of a T-lymphotropic retrovirus from a patient at risk for
409 acquired immune deficiency syndrome (AIDS).," *Science*, vol. 220, no. 4599, pp. 868–
410 871, May 1983, doi: 10.1126/science.6189183.

411 [2] S. Wain-Hobson *et al.*, "LAV revisited: origins of the early HIV-1 isolates from Institut
412 Pasteur.," *Science*, vol. 252, no. 5008, pp. 961–965, May 1991, doi:
413 10.1126/science.2035026.

414 [3] L. Ratner *et al.*, "Complete nucleotide sequence of the AIDS virus, HTLV-III.," *Nature*,
415 vol. 313, no. 6000, pp. 277–284, Jan. 1985, doi: 10.1038/313277a0.

416 [4] A. R. Gener and J. T. Kimata, "Full-coverage native RNA sequencing of HIV-1 viruses,"
417 *bioRxiv*, p. 845610, Jan. 2019, doi: 10.1101/845610.

418 [5] G. M. Shaw, B. H. Hahn, S. K. Arya, J. E. Groopman, R. C. Gallo, and F. Wong-Staal,
419 "Molecular characterization of human T-cell leukemia (lymphotropic) virus type III in the
420 acquired immune deficiency syndrome.," *Science*, vol. 226, no. 4679, pp. 1165–1171,
421 Dec. 1984, doi: 10.1126/science.6095449.

422 [6] M. Krupovic *et al.*, "Ortervirales: New Virus Order Unifying Five Families of Reverse-
423 Transcribing Viruses," *J. Virol.*, vol. 92, no. 12, pp. e00515-18, May 2018, doi:
424 10.1128/JVI.00515-18.

425 [7] E. H. Graf *et al.*, "Elite suppressors harbor low levels of integrated HIV DNA and high
426 levels of 2-LTR circular HIV DNA compared to HIV+ patients on and off HAART,"
427 *PLoS Pathog.*, vol. 7, no. 2, 2011, doi: 10.1371/journal.ppat.1001300.

428 [8] Y. Ito *et al.*, "Number of infection events per cell during HIV-1 cell-free infection," *Sci.*
429 *Rep.*, vol. 7, no. 1, p. 6559, 2017, doi: 10.1038/s41598-017-03954-9.

430 [9] I. Cuesta, A. Mari, A. Ocampo, C. Miralles, S. Pérez-castro, and M. M. Thomson,
431 "Sequence Analysis of In Vivo -Expressed HIV-1 Spliced RNAs Reveals the Usage of
432 New and Unusual Splice Sites by Viruses of Different Subtypes," pp. 1–24, 2016, doi:
433 10.1371/journal.pone.0158525.

434 [10] C. Wymant *et al.*, "Easy and accurate reconstruction of whole HIV genomes from short-
435 read sequence data with shiver," *Virus Evol.*, vol. 4, no. 1, pp. 1–13, 2018, doi:
436 10.1093/ve/vey007.

437 [11] K. M. Bruner *et al.*, "A quantitative approach for measuring the reservoir of latent HIV-1
438 proviruses," *Nature*, vol. 566, no. 7742, pp. 120–125, 2019, doi: 10.1038/s41586-019-
439 0898-8.

440 [12] M. R. Pinzone and U. O'Doherty, "Measuring integrated HIV DNA ex vivo and in vitro
441 provides insights about how reservoirs are formed and maintained," *Retrovirology*, vol.
442 15, no. 1, pp. 1–12, 2018, doi: 10.1186/s12977-018-0396-3.

443 [13] K. B. Einkauf *et al.*, "Intact HIV-1 proviruses accumulate at distinct chromosomal
444 positions during prolonged antiretroviral therapy Find the latest version : Intact HIV-1
445 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral
446 therapy," vol. 129, no. 3, pp. 988–998, 2019.

447 [14] D. Bonsall *et al.*, "THAA0101 - HIV genotyping and phylogenetics in the HPTN 071
448 (PopART) study: Validation of a high-throughput sequencing assay for viral load

449 quantification, genotyping, resistance testing and high-resolution transmission
450 networking," in *22nd International AIDS Conference (AIDS2018)*, 2018, p. Oral Abstract.

451 [15] A. N. Banin *et al.*, "Development of a Versatile, Near Full Genome Amplification and
452 Sequencing Approach for a Broad Variety of HIV-1 Group M Variants," *Viruses*, vol. 11,
453 no. 4, p. 317, Apr. 2019, doi: 10.3390/v11040317.

454 [16] A. G. Fisher, E. Collalti, L. Ratner, R. C. Gallo, and F. Wong-Staal, "A molecular clone of
455 HTLV-III with biological activity," *Nature*, vol. 316, no. 6025, pp. 262–265, 1985, doi:
456 10.1038/316262a0.

457 [17] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34,
458 no. 18, pp. 3094–3100, May 2018, doi: 10.1093/bioinformatics/bty191.

459 [18] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows – Wheeler
460 transform," vol. 26, no. 5, pp. 589–595, 2010, doi: 10.1093/bioinformatics/btp698.

461 [19] E. Afgan *et al.*, "The Galaxy platform for accessible, reproducible and collaborative
462 biomedical analyses: 2016 update," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W3–W10,
463 2016, doi: 10.1093/nar/gkw343.

464 [20] J. T. Robinson *et al.*, "Integrative genomics viewer," *Nat Biotechnol*, vol. 29, no. 1, pp.
465 24–26, 2011, doi: 10.1038/nbt0111-24.

466 [21] B. P. Walenz, S. Koren, N. H. Bergman, A. M. Phillippy, J. R. Miller, and K. Berlin,
467 "Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat
468 separation," *Genome Res.*, vol. 27, no. 5, pp. 722–736, 2017, doi: 10.1101/gr.215087.116.

469 [22] S. Andrews, "FastQC A Quality Control tool for High Throughput Sequence Data."

470 [23] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7:
471 Improvements in performance and usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780,
472 2013, doi: 10.1093/molbev/mst010.

473 [24] K. Katoh, J. Rozewicki, and K. D. Yamada, "MAFFT online service: multiple sequence
474 alignment, interactive sequence choice and visualization," *Brief. Bioinform.*, vol. 20, no. 4,
475 pp. 1160–1166, Sep. 2017, doi: 10.1093/bib/bbx108.

476 [25] D. M. Lyons and A. S. Lauring, "Evidence for the Selective Basis of Transition-to-
477 Transversion Substitution Bias in Two RNA Viruses," *Mol. Biol. Evol.*, vol. 34, no. 12,
478 pp. 3205–3215, 2017, doi: 10.1093/molbev/msx251.

479 [26] N. J. Loman, J. Quick, and J. T. Simpson, "A complete bacterial genome assembled de
480 novo using only nanopore sequencing data," *Nat. Methods*, vol. 12, no. 8, pp. 733–735,
481 2015, doi: 10.1038/nmeth.3444.

482 [27] A. Gener *et al.*, "PEA0011 - Insights from HIV-1 transgene insertions in the murine
483 model of HIV-associated nephropathy," in *23rd International AIDS Conference*
484 *(AIDS2020)*, 2020, vol. ePoster.

485 [28] A. R. Gener *et al.*, "P39 - Insights from comprehensive transcript models of HIV-1," in
486 *Genome Informatics 2020*, 2020, p. ePoster.

487 [29] A. R. Gener, T. Washington, D. Hyink, and P. Klotman, "3264 - The Multiple HIV-1
488 Transgenes in the Murine Model of HIV-Associated Nephropathy Fail to Segregate as
489 Expected," in *American Society of Human Genetics Annual Meeting*, 2020, p. ePoster.

490 [30] A. Payne, N. Holmes, V. Rakyan, and M. Loose, "Whale watching with BulkVis: A

491 graphical viewer for Oxford Nanopore bulk fast5 files," *bioRxiv*, p. 312256, Jan. 2018,
492 doi: 10.1101/312256.

493 [31] P. Rosenstiel, A. Gharavi, V. D'Agati, and P. Klotman, "Transgenic and infectious animal
494 models of HIV-associated nephropathy.," *J. Am. Soc. Nephrol.*, vol. 20, no. 11, pp. 2296–
495 304, 2009, doi: 10.1681/ASN.2008121230.

496 [32] M. Kvaratskhelia, A. Sharma, R. C. Larue, E. Serrao, and A. Engelman, "Molecular
497 mechanisms of retroviral integration site selection.," *Nucleic Acids Res.*, vol. 42, no. 16,
498 pp. gku769-, 2014, doi: 10.1093/nar/gku769.

499 [33] B. Marini *et al.*, "Nuclear architecture dictates HIV-1 integration site selection," *Nature*,
500 vol. 521, pp. 227–233, 2015, doi: 10.1038/nature14226.

501 [34] H. Ochman, A. S. Gerber, and D. L. Hartl, "Genetic applications of an inverse polymerase
502 chain reaction," *Genetics*, vol. 120, no. 3, pp. 621–623, 1988.

503 [35] W.-S. Hu and S. H. Hughes, "HIV-1 Reverse Transcription," *Cold Spring Harb. Perspect.*
504 *Med.* , vol. 2, no. 10, Oct. 2012, doi: 10.1101/cshperspect.a006882.

505 [36] M. R. Green, T. Maniatis, and D. A. Melton, "Human beta-globin pre-mRNA synthesized
506 in vitro is accurately spliced in Xenopus oocyte nuclei.," *Cell*, vol. 32, no. 3, pp. 681–
507 694, Mar. 1983, doi: 10.1016/0092-8674(83)90054-5.

508 [37] A. R. Gener, "Full-coverage sequencing of HIV-1 provirus from a reference plasmid,"
509 *bioRxiv*, p. 611848, Jan. 2019, doi: 10.1101/611848.

510 [38] A. Adachi *et al.*, "Production of acquired immunodeficiency syndrome-associated
511 retrovirus in human and nonhuman cells transfected with an infectious molecular clone.,"

512 *J. Virol.*, vol. 59, no. 2, pp. 284–91, 1986.

513 [39] B. Lucic *et al.*, "Spatially clustered loci with multiple enhancers are frequent targets of
514 HIV-1," *bioRxiv*, 2018.

515 [40] W. J. Kent *et al.*, "The Human Genome Browser at UCSC," *Genome Res.* , vol. 12, no. 6,
516 pp. 996–1006, Jun. 2002, doi: 10.1101/gr.229102.

517 [41] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "IDBA – A Practical Iterative de
518 Bruijn Graph De Novo Assembler BT - Research in Computational Molecular Biology,"
519 2010, pp. 426–440.

520 [42] P. J. A. Cock, B. A. Grüning, K. Paszkiewicz, and L. Pritchard, "Galaxy tools and
521 workflows for sequence analysis with applications in molecular plant pathology," *PeerJ*,
522 vol. 1, p. e167, 2013, doi: 10.7717/peerj.167.

523 [43] C. B, W. T, and S. S, "Genome sequence assembly using trace signals and additional
524 sequence information.," *Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma.*, vol. 99, pp. 45–
525 56.

526 [44] A. Bankevich *et al.*, "SPAdes: A New Genome Assembly Algorithm and Its Applications
527 to Single-Cell Sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, Apr. 2012, doi:
528 10.1089/cmb.2012.0021.

529 [45] G. Cuccuru *et al.*, "Orione, a web-based framework for NGS analysis in microbiology,"
530 *Bioinformatics*, vol. 30, no. 13, pp. 1928–1929, Jul. 2014, doi:
531 10.1093/bioinformatics/btu135.

532 [46] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt, "Assembling millions of short

533 DNA sequences using SSAKE," *Bioinformatics*, vol. 23, no. 4, pp. 500–501, Feb. 2007,
534 doi: 10.1093/bioinformatics/btl629.

535

536

Tables

539 **Table 1: Summary of pHXB2 sample divergence from reference HXB2.**

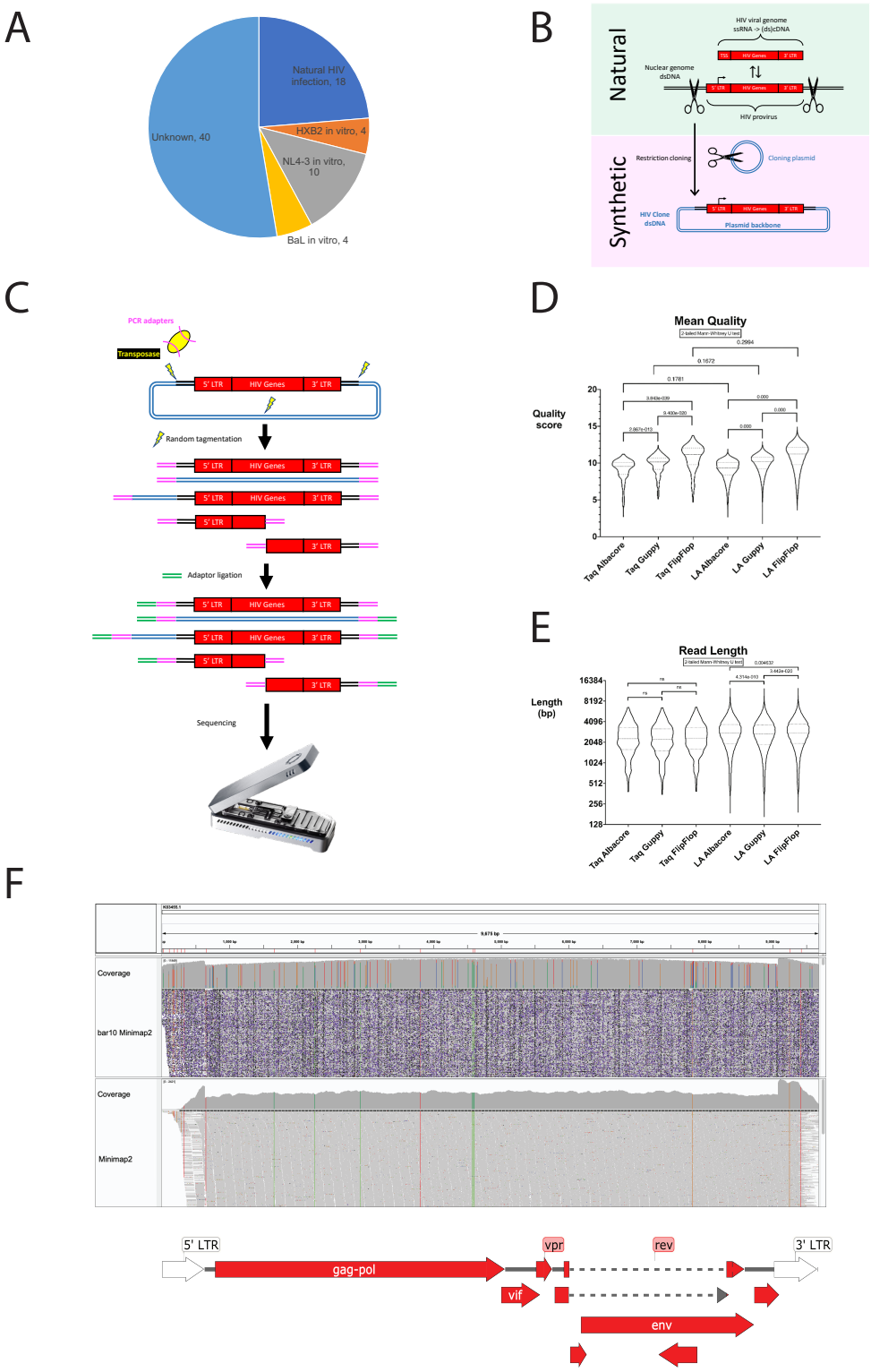
Site	Position	Change	Substitution Class	Change	Mutation Class (Syn/Non /Stop)	Homopolymer-adjacent?	Same as neighbor?	LANL Feature	Subfeature	Frame
1	24	C>A	transversion	NA	NA	yes	yes	5'LTR	U3	NA
2	108	A>G	transition	NA	NA	yes	yes	5'LTR	U3	NA
3	164	G>T	transversion	NA	NA	yes	no	5'LTR	U3	NA
4	168	T>G	transversion	NA	NA	yes	yes	5'LTR	U3	NA
5	176	A>G	transition	NA	NA	yes	yes	5'LTR	U3	NA
6	182	C>T	transition	NA	NA	yes	no	5'LTR	U3	NA
7	227	A>G	transition	NA	NA	yes	yes	5'LTR	U3	NA
8	291	A>G	transition	NA	NA	no	no	5'LTR	U3	NA
9	333	C>T	transition	NA	NA	no	no	5'LTR	U3	NA
10	654	C>T	transition	NA	NA	no	no	None	None	NA
11	1659	aaG>aaA	transition	None	Syn	yes	yes	gag	p24, p55	gag frame 1
12	2259	gag:agG>agA pol:Gtc>Atc	transition	gag:Arg>Arg pol:Val>Ile	Syn/Non	no	no	gagpol	p6	gag frame 1 pol frame 3
13	2927	aaG>aaA	transition	None	Syn	yes	yes	pol	p51 RT	pol frame 3
14	3812	ccC>ccT	transition	None	Syn	yes	yes	pol	p51 RT	pol frame 3
15	4574	acT>acA	transversion	None	Syn	no	no	pol	p31 IN	pol frame 3
16	4596	Ggt>Agt	transition	None	Syn	yes	no	pol	p31 IN	pol frame 3

17	4609	aGg>aAg	transition	Arg>Lys	Non	yes	yes	pol	p31 IN	pol frame 3
18	7823	gcC>gcG Ggc>Cgc	transversion	ASP:Gly>Arg	Syn/Non	no	no	gp41	RRE, also ASP	gp41 frame 3, ASP -2
19	9253	Ata>Gta	transition	Ile>val	Non	no	yes	nef/3'LTR	also U3	nef frame 1
20	9418	C>T	transition	NA	NA	no	no	3'LTR	U3	NA

Coverage numbers vary by input (albacore, guppy, FlipFlop basecalled FASTQ) and mapping method (Minimap2 vs. BWA-MEM). This information is provided as Supplemental Digital Content. Base-1 (first base is numbered 1, 2nd 2, etc.), relative to HXB2, Genbank:K03455.1. Changed base represented as upper-case. Annotated as codon if in protein-coding region. No deletions or insertions were predicted from manual inspection or supported by short-read sequencing. Abbreviations, ASP: antisense protein, RRE: rev-response element, NA: not applicable. Syn: synonymous mutation. Non: non-synonymous mutation. Stop: stop codon/non-sense mutation. LTR: long terminal repeat. RT: reverse transcriptase. IN: integrase. LANL: Los Alamos National Laboratory HIV Sequence Database. Data from three separate sequencing experiments on the same plasmid sample support these 20 sites. Note site 1-8 variants in 5’LTR have been previously reported (LANL), albethey ambiguously. These may also be incorrectly annotated as variants in nef.

Figures

Figure 1: HIV information in pHXB2_D is recovered by long-read sequencing and mapping.



558 Figure 1A: HXB2 is still a commonly used resource. It is the reference HIV-1 genome, derived
559 from one of the earliest clinical isolates. While older HIV samples are occasionally rediscovered,
560 they are not made routinely available to researchers. All public HIV-1 RNA-seq datasets were
561 obtained from the NCBI SRA with the following search phrase: "HIV-1" AND "RNA-seq".
562 Metadata from these 2527 runs (number current as of 7/21/2020) were used to make a pie chart
563 summary.

564 Figure 1B: HIV information comes from three main sources: proviruses (HIV sandwiched
565 between two assumedly identical full-length long terminal repeats (LTRs)), unspliced HIV
566 mRNAs (also known as viral genomes) starting from the transcription start site and ending in the
567 3' LTR [4], and engineered proviruses recovered in their entirety or stitched together from
568 multiple isolates like NL4-3 [38].

569 Figure 1C: ONT library prep pipeline. Tagmentation cleaves double-stranded DNA, ligating
570 barcoded PCR adapters (magenta). PCR-adapted DNA may be amplified. After amplification
571 and cleanup, ONT sequencing adapters (green) are ligated. Barcoded samples may be pooled and
572 sequenced.

573 Figure 1D: Newer basecallers increase read mean quality. Median (big dash) and quartiles (little
574 dash). Effect of enzyme version was not statistically significant.

575 Figure 1E: Read stats with different callers/aligners. Median (big dash) and quartiles (little dash).
576 Read lengths increase with higher fidelity Taq.

577 Figure 1F: Sequencing coverage with long- vs. short-read single-end 150 bp (trimmed to 142 bp)
578 DNA sequencing. Long-read sequencing covers ambiguously mappable areas missed by short-
579 read in HXB2 reference Genbank:K03455.1 (**Supplemental Figure 2D,2E**), but at the expense

of accuracy near homopolymers longer than about 4 nucleobases (**Supplemental Figure 4**). Short-read mapping fails at repetitive elements longer than their read lengths (**Supplemental Figure 2D,2E**). Long read Minimap2 settings: map-ont -k15. Short read Minimap2 settings: Short reads without splicing (-k21 -w11 --sr -F800 -A2 -B8 -O12,32 -E2,1 -r50 -p.5 -N20 -f1000,5000 -n2 -m20 -s40 -g200 -2K50m --heap-sort=yes --secondary=no) (sr).

588 **Figure 2: Longest read containing complete full-length HIV-1 reference HXB2**


```
@6fbf0205-5195-460e-8e28-930db50e5d79 runid=0b284792282af9a6d7275cfca845556bb4b8ac7f sampleid=Student_Genomics_Run_1 read=87223 ch=460 start_time=2018-05-09T20:57:49Z barcode=barcode10
TGTTGTACTTCGTTGTCGTTGTCGTTGTCGTTGTCGTTGTCGTTGTCGTTGTCGTTGTCGCCGCTGCTGCT
```

590 The 5th longest read in the barcode 10 set (read ID 6fbf0205-5195-460e-8e28-930db50e5d79)
591 contained full-length HIV-1. Query (full read) blastn against HIV (taxid:11676) returned 92.95%
592 identity to HIV-1, complete genome (Genbank:AF033819.3). Limiting query to HXB2 (red)
593 blastn against Nucleotide collection nr/nt returned 100% coverage and 93.02% identity to HIV-1
594 HXB2. This read was 11,487 bases long, with mean quality score 11.984396. Basecalled using
595 Guppy 2.3.1 with FlipFlop config.

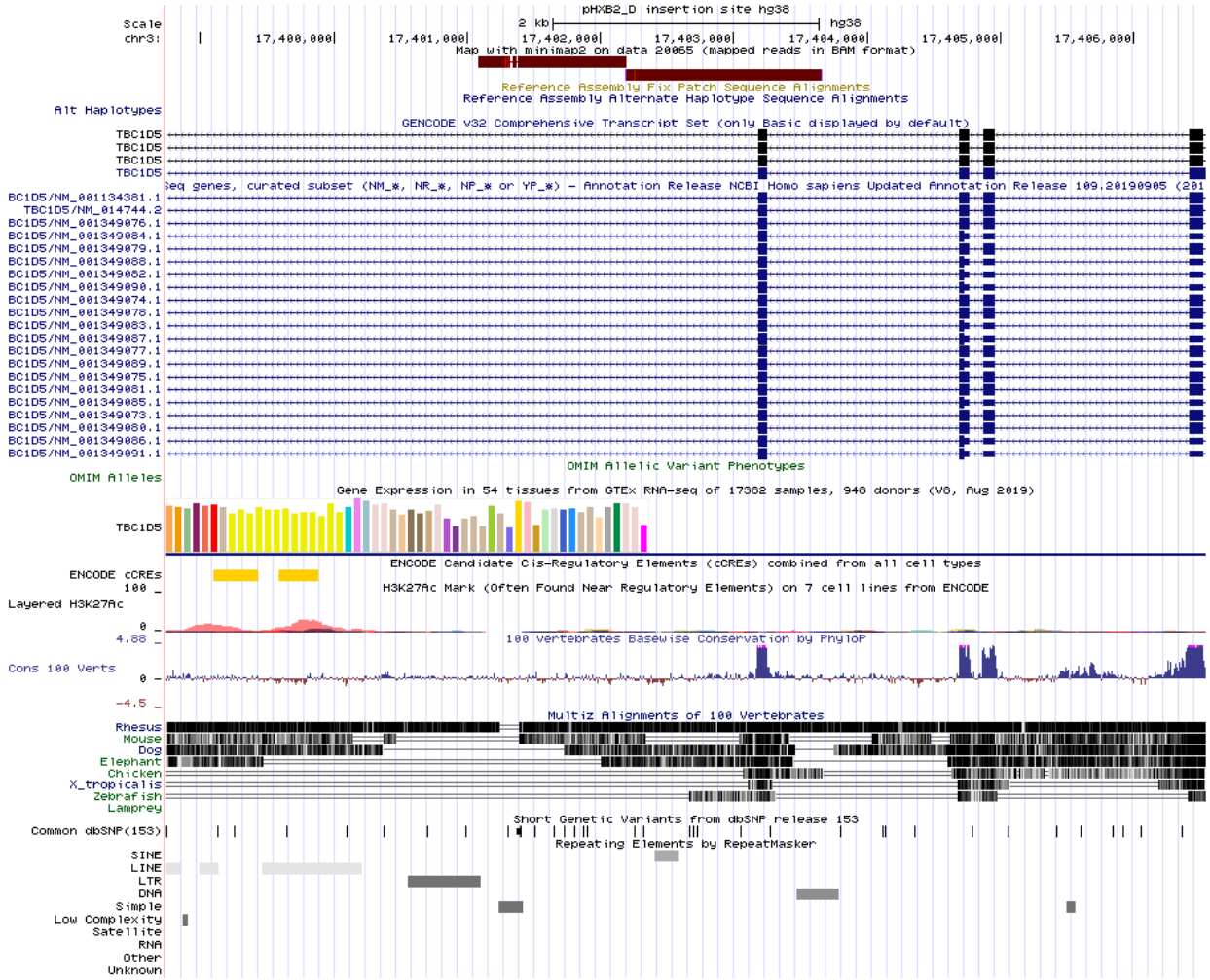
596

Figure 3: pHXB2_D has identical LTRs, resolving likely errors in HXB2 (K03455.1)

```
CLUSTAL format alignment by MAFFT (v7.475)


K03455.1_5'LTR  tggaagggctaattcactcccaacgaagacaagatatccttgatctgtggatctaccaca
pHXB2_D_5'LTR   tggaagggctaattcactcccaaagaagacaagatatccttgatctgtggatctaccaca
pHXB2_D_3'LTR   tggaagggctaattcactcccaaagaagacaagatatccttgatctgtggatctaccaca
K03455.1_3'LTR  tggaagggctaattcactcccaaagaagacaagatatccttgatctgtggatctaccaca
                ********************** **********************************

K03455.1_5'LTR  cacaaggctacttccctgattagcagaactacacaccagggccagggatcagatatccac
pHXB2_D_5'LTR   cacaaggctacttccctgattagcagaactacacaccagggccaggggtcagatatccac
pHXB2_D_3'LTR   cacaaggctacttccctgattagcagaactacacaccagggccaggggtcagatatccac
K03455.1_3'LTR  cacaaggctacttccctgattagcagaactacacaccagggccaggggtcagatatccac
                ************************************************************

K03455.1_5'LTR  tgacctttggatggtgctacaagctagtaccagttgagccagagaagttagaagaagcca
pHXB2_D_5'LTR   tgacctttggatggtgctacaagctagtaccagttgagccagataaggtagaagaggcca
pHXB2_D_3'LTR   tgacctttggatggtgctacaagctagtaccagttgagccagataaggtagaagaggcca
K03455.1_3'LTR  tgacctttggatggtgctacaagctagtaccagttgagccagataagatagaagaggcca
                ****************************************** *** *******.****

K03455.1_5'LTR  acaaaggagagaacaccagcttgttacaccctgtgagcctgcatggaatggatgacccgg
pHXB2_D_5'LTR   ataaaggagagaacaccagcttgttacaccctgtgagcctgcatgggatggatgacccgg
pHXB2_D_3'LTR   ataaaggagagaacaccagcttgttacaccctgtgagcctgcatgggatggatgacccgg
K03455.1_3'LTR  ataaaggagagaacaccagcttgttacaccctgtgagcctgcatgggatggatgacccgg
                *.*********************************************.*************

K03455.1_5'LTR  agagagaagtgttagagtggaggtttgacagccgcctagcatttcatcacatggcccgag
pHXB2_D_5'LTR   agagagaagtgttagagtggaggtttgacagccgcctagcatttcatcacgtggcccgag
pHXB2_D_3'LTR   agagagaagtgttagagtggaggtttgacagccgcctagcatttcatcacgtggcccgag
K03455.1_3'LTR  agagagaagtgttagagtggaggtttgacagccgcctagcatttcatcacgtggcccgag
                ****************************************************.*********

K03455.1_5'LTR  agctgcatccggagtacttcaagaactgctgacatcgagcttgctacaagggactttccg
pHXB2_D_5'LTR   agctgcatccggagtacttcaagaactgctgatatcgagcttgctacaagggactttccg
pHXB2_D_3'LTR   agctgcatccggagtacttcaagaactgctgatatcgagcttgctacaagggactttccg
K03455.1_3'LTR  agctgcatccggagtacttcaagaactgctgacatcgagcttgctacaagggactttccg
                ********************************.***************************

K03455.1_5'LTR  ctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagat
pHXB2_D_5'LTR   ctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagat
```

```
pHXB2_D_3'LTR   ctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagat
K03455.1_3'LTR  ctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagat
                ************************************************************

K03455.1_5'LTR  cctgcatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
pHXB2_D_5'LTR   cctgcatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
pHXB2_D_3'LTR   cctgcatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
K03455.1_3'LTR  cctgcatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
                ************************************************************

K03455.1_5'LTR  gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
pHXB2_D_5'LTR   gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
pHXB2_D_3'LTR   gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
K03455.1_3'LTR  gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
                ************************************************************

K03455.1_5'LTR  tgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
pHXB2_D_5'LTR   tgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
pHXB2_D_3'LTR   tgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
K03455.1_3'LTR  tgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
                ************************************************************

K03455.1_5'LTR  agaccctttagtcagtgtggaaaatctctagca
pHXB2_D_5'LTR   agaccctttagtcagtgtggaaaatctctagca
pHXB2_D_3'LTR   agaccctttagtcagtgtggaaaatctctagca
K03455.1_3'LTR  agaccctttagtcagtgtggaaaatctctagca
                **********************************
```

Figure 4: HXB2 integration site

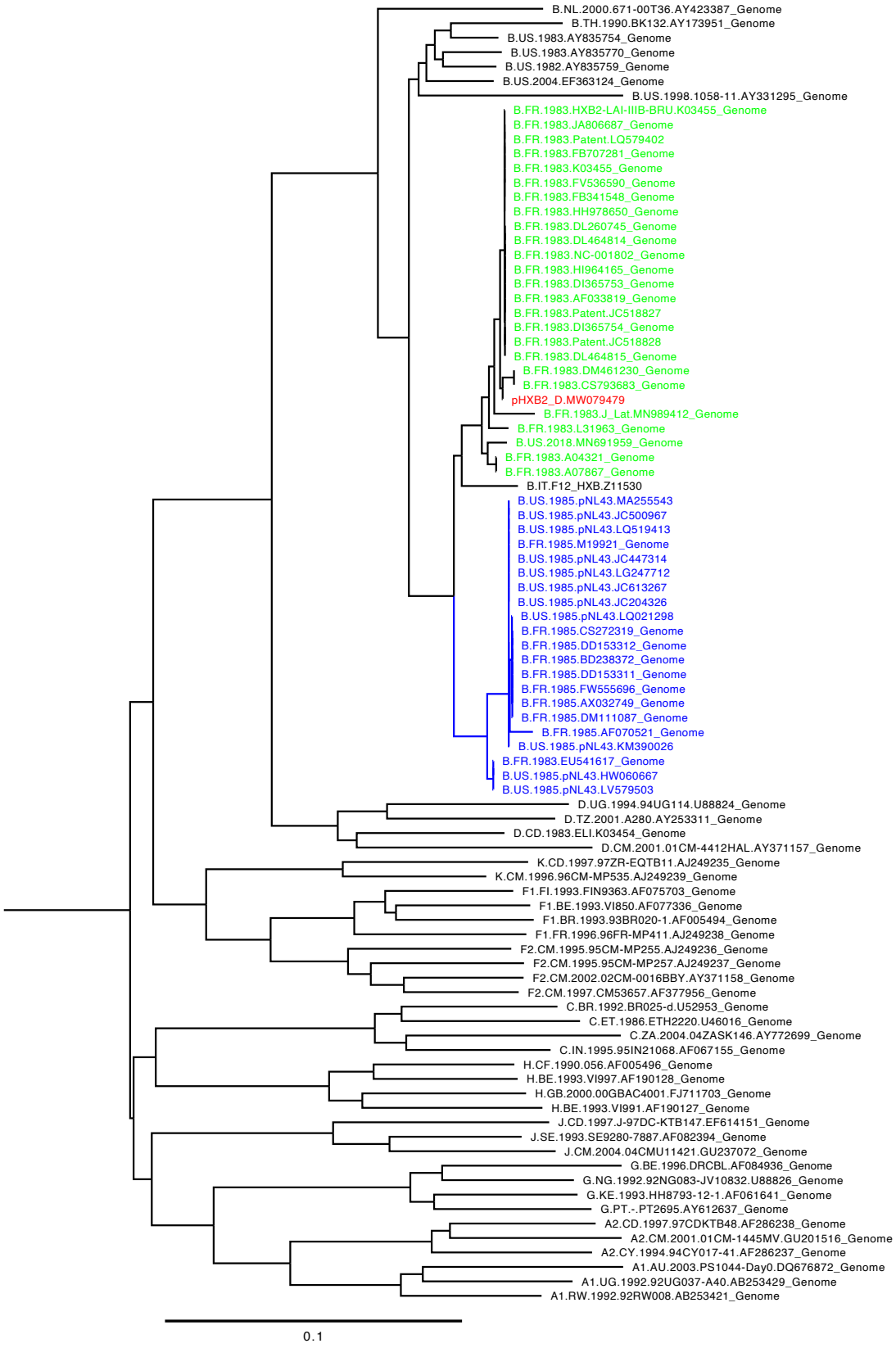


pHXB2_D's, and therefore HXB2's, integration site is unambiguously singular (falls outside of annotated repeat), and in the same orientation (minus strand relative to hg38) as target gene TBC1D5. Alignment quality is 60 for both homology arms (**Supplemental Table 2**). Features captured by homology arms in pHXB2_D and other clones verified as proviruses in the present

677 study are consistent with HIV-1 integration behavior [39]. Visualized in UCSC Genome Browser
678 [40].

679

680 **Figure 5: pHXB2_D provenance and top 50 neighbors**



681

682 **Supplemental Information**

683 **Data exploration with long- and short-read mapping**

684 To assemble pHXB2_D, we used tried the following short read assemblers on short-read
685 data from the external core: IDBA [41], MIRA [42], [43], SPAdes [44], and SSAKE [45], [46].
686 These were chosen as a convenience because they were already stably implemented in Galaxy
687 (specifically usegalaxy.eu). Of these, SSAKE produced discontinuous assemblies with default
688 parameters. The discontinuous contigs did however map to the core's assembly (not shown).

689 **Enabling STEM outreach**

690 This work was performed as two control experiments with identically prepared libraries
691 for a STEM outreach initiative, Student Genomics (Gener, et al., manuscript in prep). Given the
692 constraints of the Student Genomics pilot, a rapid sequencing kit with tagmentation (explained
693 below) with PCR barcoding was used to pool samples for ONT sequencing, with the
694 consequence of fragmenting plasmid DNA more than what would have been ideal for capturing
695 full-length HIV. That said, these controls could have been just as easily replaced by any
696 samples/experiments benefiting from long-read sequencing at moderate-to-high coverage.

697

Supplemental Tables

701 **Supplemental Table 1: HXB2 is still a common HIV clone.**

702 See **Supplemental Digital Content.**

703 See also **Figure 1A**.

704

705

706

707 **Supplemental Table 2: HIV provirus clones**

708 See **Supplemental Digital Content.**

709 Of the HIV clones available through ARP, the table represents the only validated proviruses with
710 both upstream and downstream homology arms mapping to the same integration sites. pNL4-3 is
711 included as a known chimera with two integration half-sites. Other clones were made with
712 cDNA cloning, usually TA cloning (per ARP entries). Note: Reference hg38. Aligner: minimap2
713 with "Long Assembly" mapping settings. All homology arms had Alignment quality = 60.
714 Upstream = host plus strand; independent of integration orientation. Coordinates reported from
715 UCSC. ARP = NIH AIDS Reagent and Reference Program. IS = integration site.

716 **Supplemental Table 3: Variation in assemblies at the feature level.**

	Mismatches						Gaps (INDEL)					
	Taq			LA Taq			Taq			LA Taq		
	Albacore	Guppy	FlipFlop	Albacore	Guppy	FlipFlop	Albacore	Guppy	FlipFlop	Albacore	Guppy	FlipFlop
5' LTR	NA	9	9	9	9	9	NA	NA	0	2	0	0
gag	2	2	2	2	2	2	12	10	9	9	8	8
5' LTR+ψ	10	10	NA	NA	NA	NA	5	2	NA	NA	NA	NA
pol	7	6	6	6	6	6	26	22	9	18	11	10
vif	0	0	0	0	0	0	3	3	2	3	1	1
vpr	0	0	0	0	0	0	1	1	1	0	0	0
tat	2	1	1	1	1	1	10	6	3	7	4	5
rev	2	1	1	1	1	1	10	7	4	7	4	5
vpu	1	0	0	0	0	0	0	0	0	0	0	0
gp160	2	1	1	1	1	1	11	7	4	8	4	5
nef	1	1	1	1	1	1	3	2	2	3	2	1
3' LTR	2	2	NA	NA	NA	NA	2	0	NA	NA	NA	NA
nef+3' LTR	NA	2	2	2	2	2	NA	2	2	5	2	1
HXB2	22	20	20	20	20	20	61	46	28	47	27	25
Downstream bridge	0	0	0	0	0	0	4	5	1	2	1	1
pBR322-related	0	0	0	0	0	0	19	19	13	18	19	15
Upstream bridge	2	2	3	2	2	2	8	6	5	7	7	3

717

718 Assembled with Canu. NA denotes features which may not have matched exactly, but which
719 were collapsed with adjacent features to facilitate counting. Variants called manually by
720 mapping assemblies over HXB2 features with SnapGene.

721

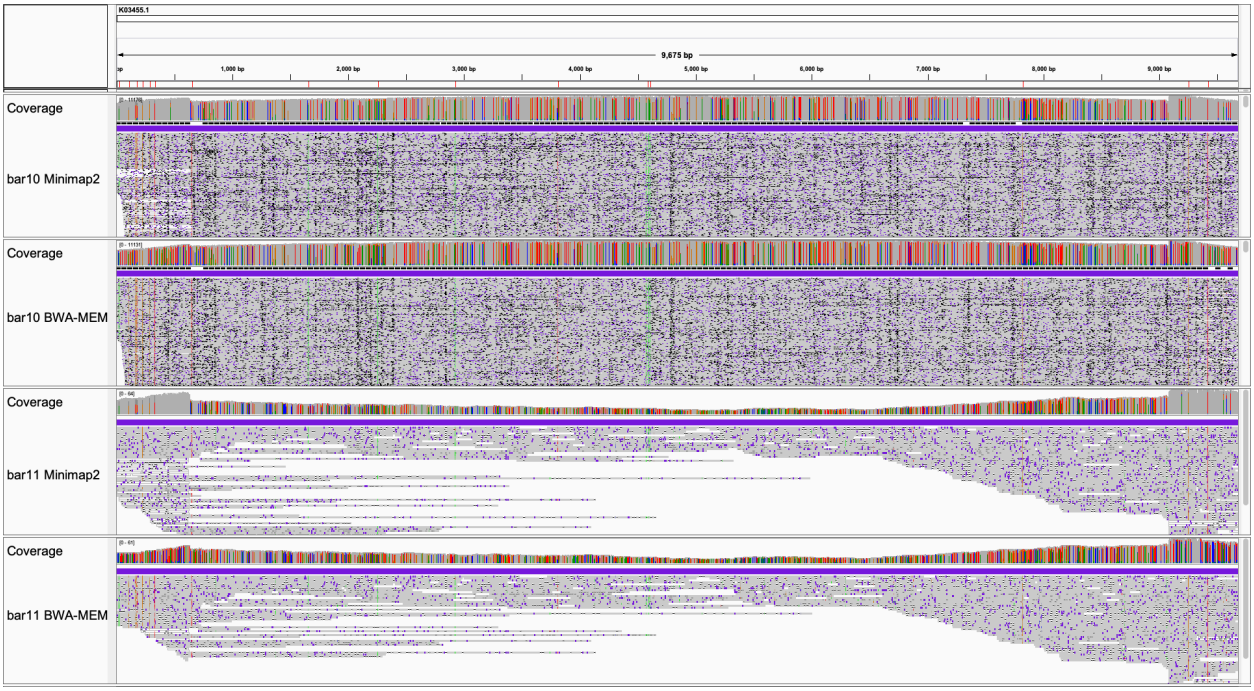
722

723 **Supplemental Figures**

724

725 **Supplemental Figure 1A: Unbiased nanopore DNA sequencing coverage over HXB2**

726 **depends on DNA polymerase and mapper. ONT basecaller = Albacore (worst).**

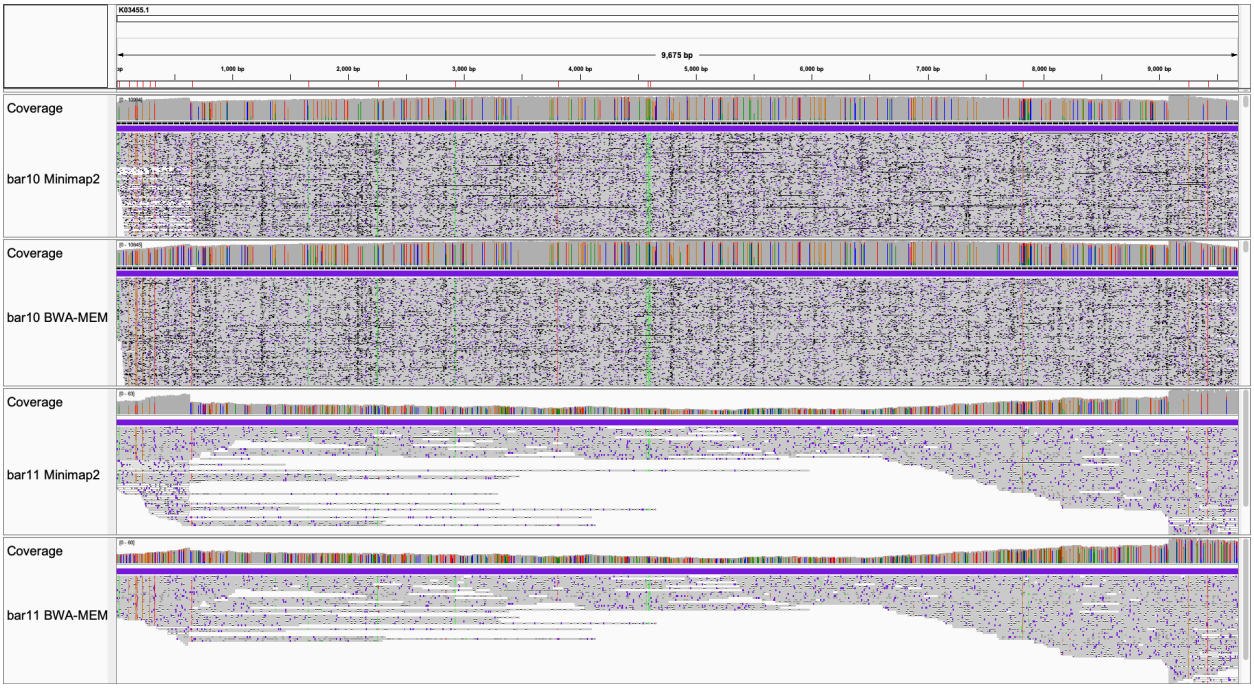


727

728

729 **Supplemental Figure 1B: Unbiased nanopore DNA sequencing coverage over HXB2**

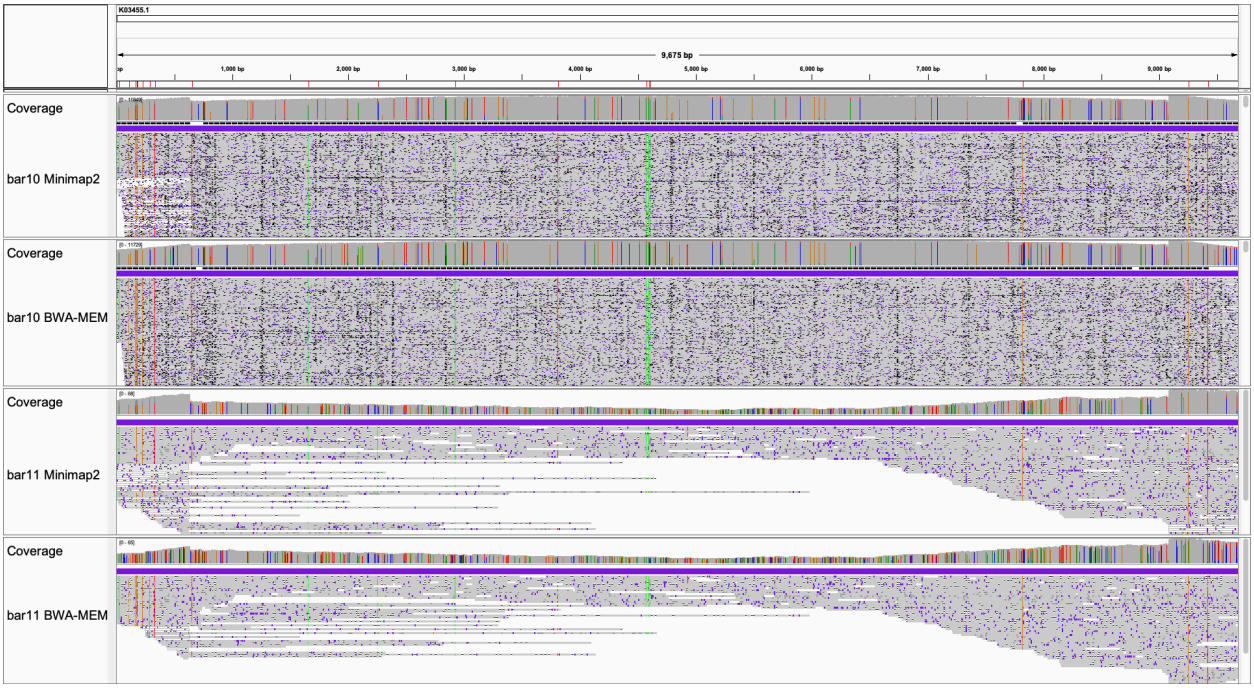
730 **depends on DNA polymerase and mapper. ONT basecaller = Guppy.**



731

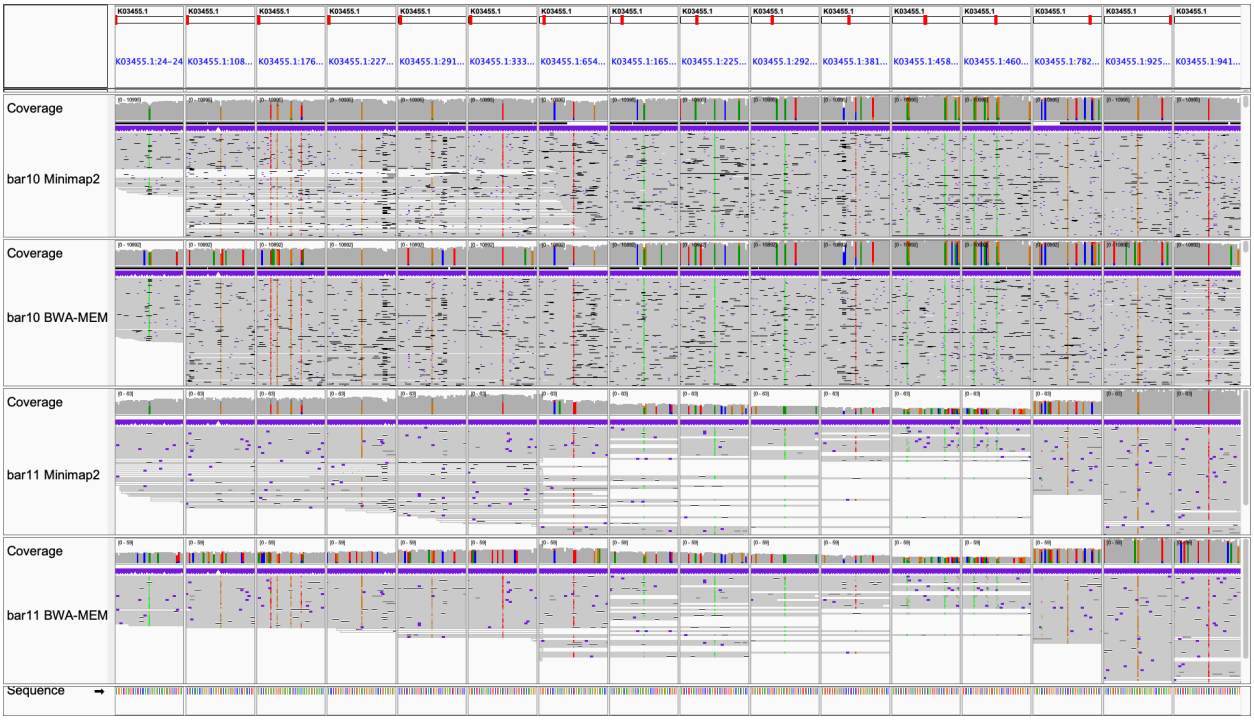
732

Supplemental Figure 1C: Unbiased nanopore DNA sequencing coverage over HXB2 depends on DNA polymerase and mapper. ONT basecaller = FlipFlop (best).



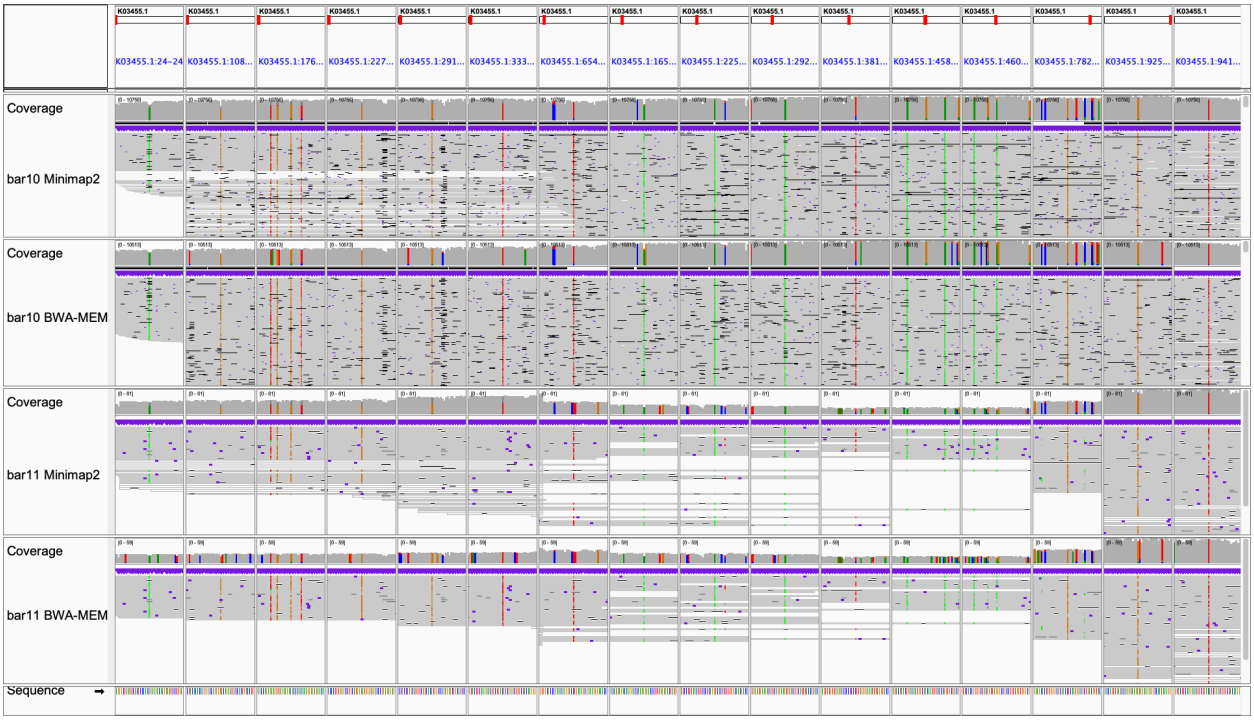
Top two Coverage and Alignment panels from barcoded library 10 (bar10 = LA Taq). Bottom two from Barcode 11 (bar11 = Taq). Minimap2 and BWA-MEM were used to map reads basecalled with Albacore (worst), Guppy, or FlipFlop (best) to HXB2. Color-coding: Red below genome scale marks 20 SNVs across the HIV segment of pHXB2_D. Purple is an insertion in a given read relative to reference. White is either a deletion in a given read or space between two aligned reads. Gray in alignment field means base same as reference, and in coverage field means major allele is at least 95% the same as reference. Per-read “insertions” and “deletions” do not necessarily represent true insertions or deletions actually present in the sample, because each read is likely an imperfect independent observation. Automated assembly followed by manual consensus building converts these overlapping reads into approximations of the ground truth. “Unbiased” refers to not amplifying a given region (e.g., pol) before ligating ONT sequencing adapters. In the present approach, the tagmentation process randomly cuts DNA, creating ~2000 bp pieces. Tagmented DNA is then amplified based on tagmentation adapters.

Supplemental Figure 2A: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller = Albacore (worst).



755 **Supplemental Figure 2B: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT**

756 **basecaller = Guppy.**

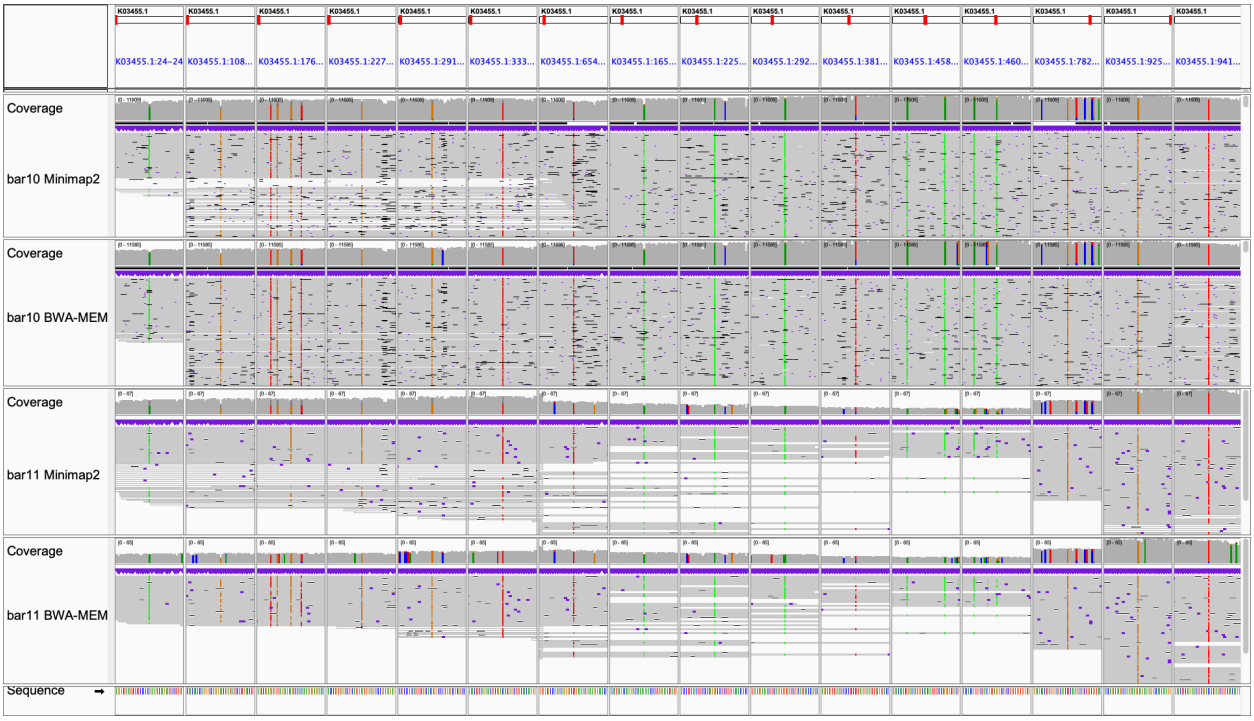


757

758

759 **Supplemental Figure 2C: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT**

760 **basecaller = FlipFlop (best).**



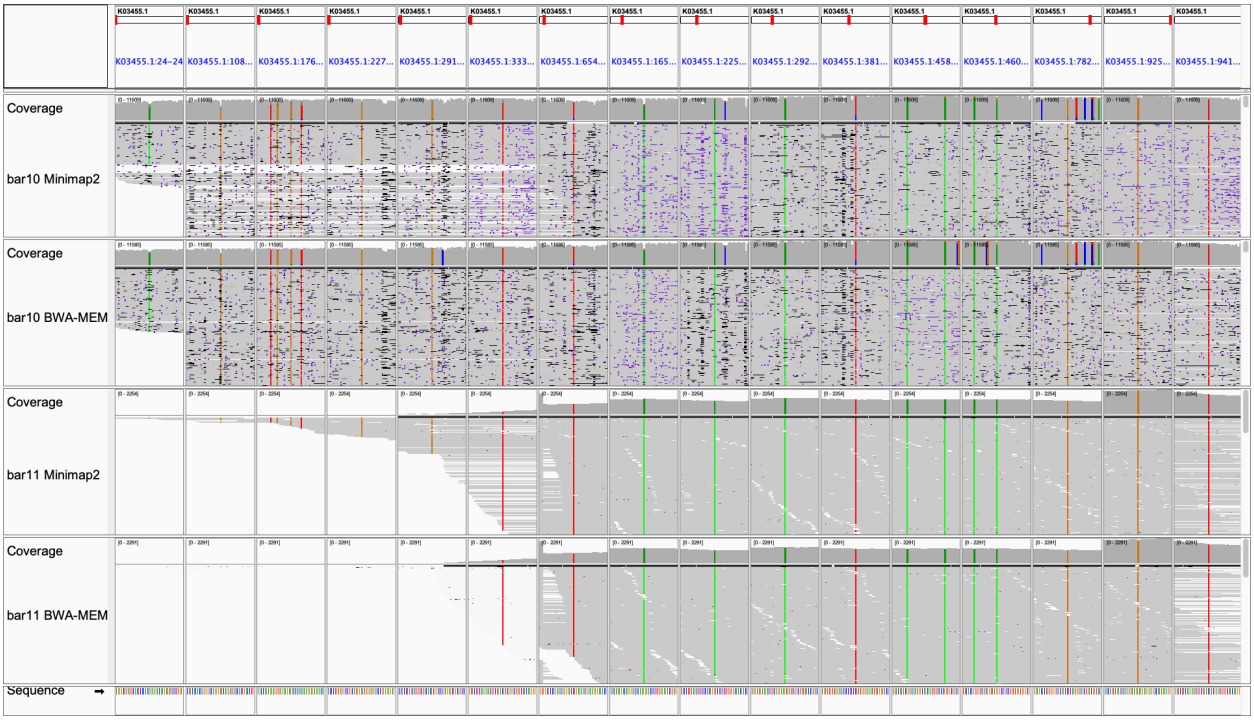
761

762

Supplemental Figure 2D: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs. short reads (HIV genome)

767 **Supplemental Figure 2E: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs.**

768 **short reads (20 SNV-focused)**



769

770

771 Supplemental Figure 2A: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller
772 = Albacore (worst). Gray indicates per-base consensus accuracy ≥ 80%. These alignments are
773 the noisiest (less gray and most divergent from reference) between Supplemental Figures 2A,
774 2B, and 2C.

775 Supplemental Figure 2B: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller
776 = Guppy.

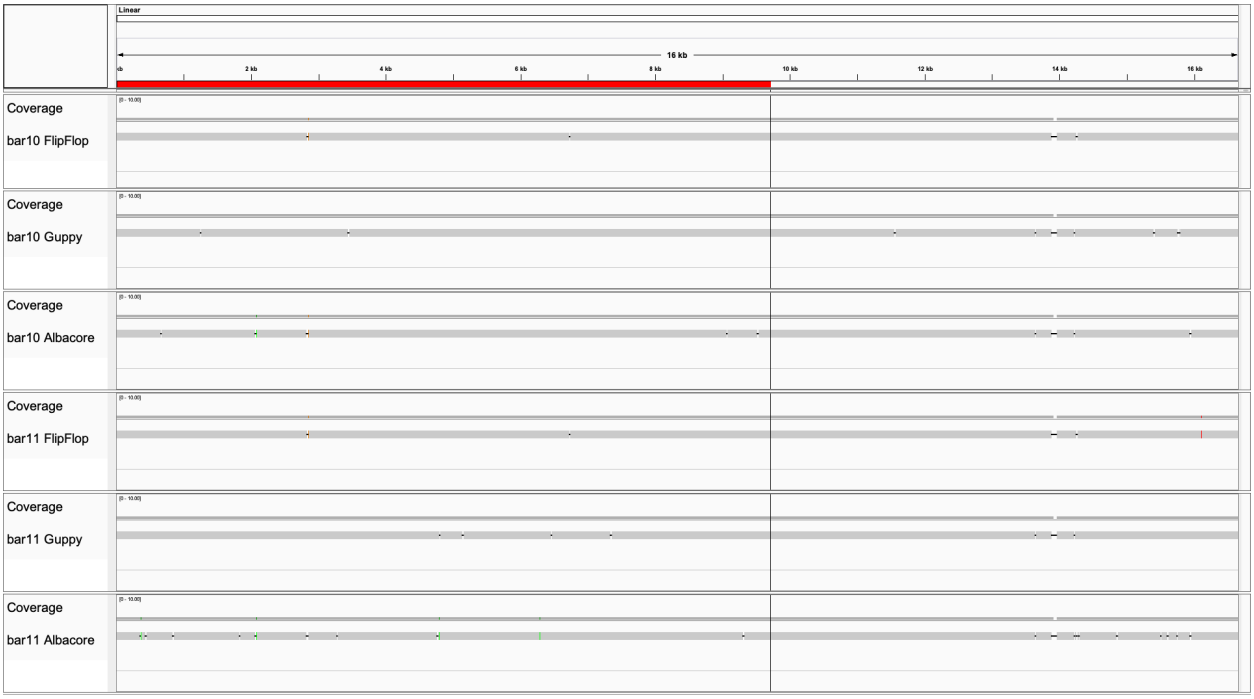
777 Supplemental Figure 2C: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller
778 = FlipFlop (best). These alignments are the least noisy (most gray and like reference) between
779 Supplemental Figures 2A, 2B, and 2C.

780 Supplemental Figure 2D: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs. short
781 reads (HIV genome). Long reads outperform short reads at HIV-1 LTRs. ONT basecaller =
782 FlipFlop. Short read as single-end 150, clipped to 142, provided by external core. Mappers =
783 Minimap2 (better), BWA-MEM.

784 Supplemental Figure 2E: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs. short
785 reads (SNV-focused).

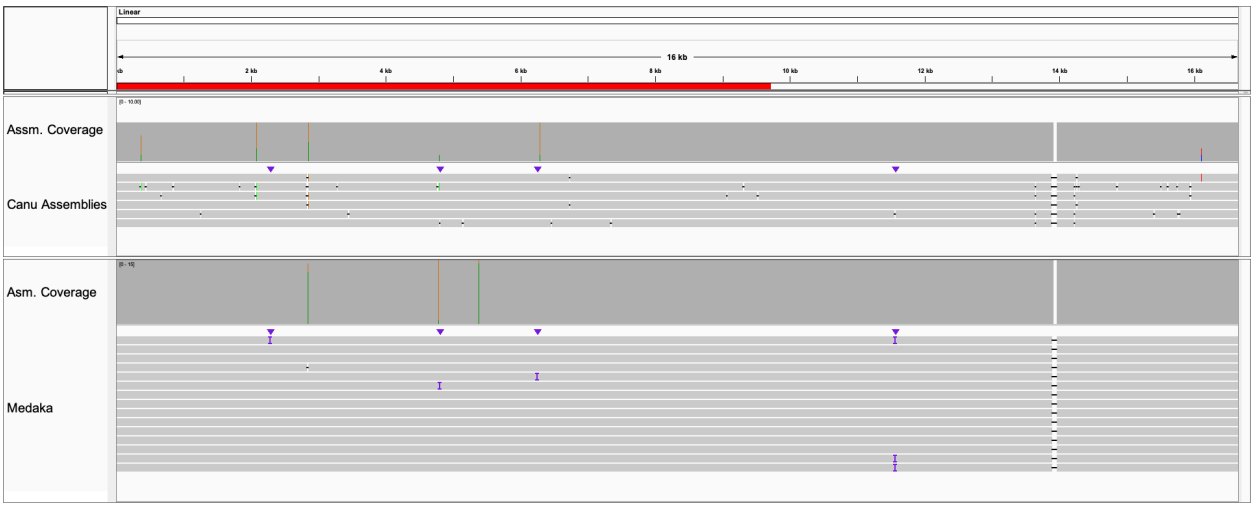
786

Supplemental Figure 3A: Assembling pHXB2_D from long reads only, varying basecaller and polymerase.



Each pane (n=6) summarizes the results of contig curation. Divergence from reference decreases with newer basecallers, and with long amplicon DNA polymerase (Sigma-Aldrich Taq vs. LA Taq by Takara). Errors in assembly occurred at homopolymers (most often deletions not visible at this resolution; see **Supplemental Figure 5**), dimer or trimer runs. bar10 = LA Taq library. bar11 = Taq library. pHXB2_D Genbank:MW079479. Best contigs presented, manually curated to match pHXB2_D coordinates. Note LTRs (beginning and terminal 634 bp of red bar) are resolved in almost all assemblies. See **Supplemental Table 3** for differences between assemblies and the reference (left red). Plasmid backbone (right) differences are not reported.

Supplemental Figure 3B: ONT errors corrected by polishing ONT-only assemblies.

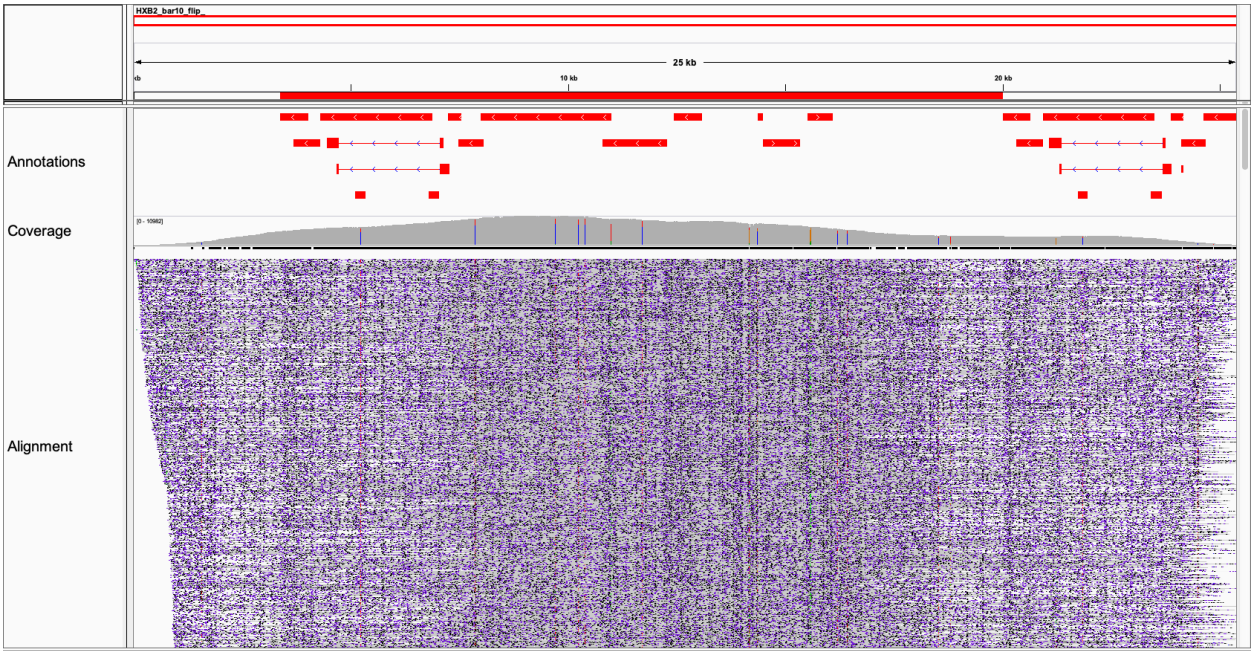


Assemblies polished with Medaka (ONT). Top: pHXB2_D genome. Bottom: HIV-only segment. The best polished assembly had one error in the entire plasmid (1 error out of 16,722 bases), with a corresponding consensus accuracy of 99.99402%. This happened to be in HIV segment (HIV-1 between position 1 and 9719; 1 error out of 9719 bases), with corresponding accuracy of 99.989711%. Note the conserved 52 bp gap in the backbone of pHXB2_D was redundant sequence included in the short-read assembly from the core. It was not supported by long-read data, and therefor was validated as a technical artifact from the core's pipeline. Reference: short-

809 read assembly. LTRs (beginning and terminal 634 bp of red bar) are resolved in polished

810 assemblies.

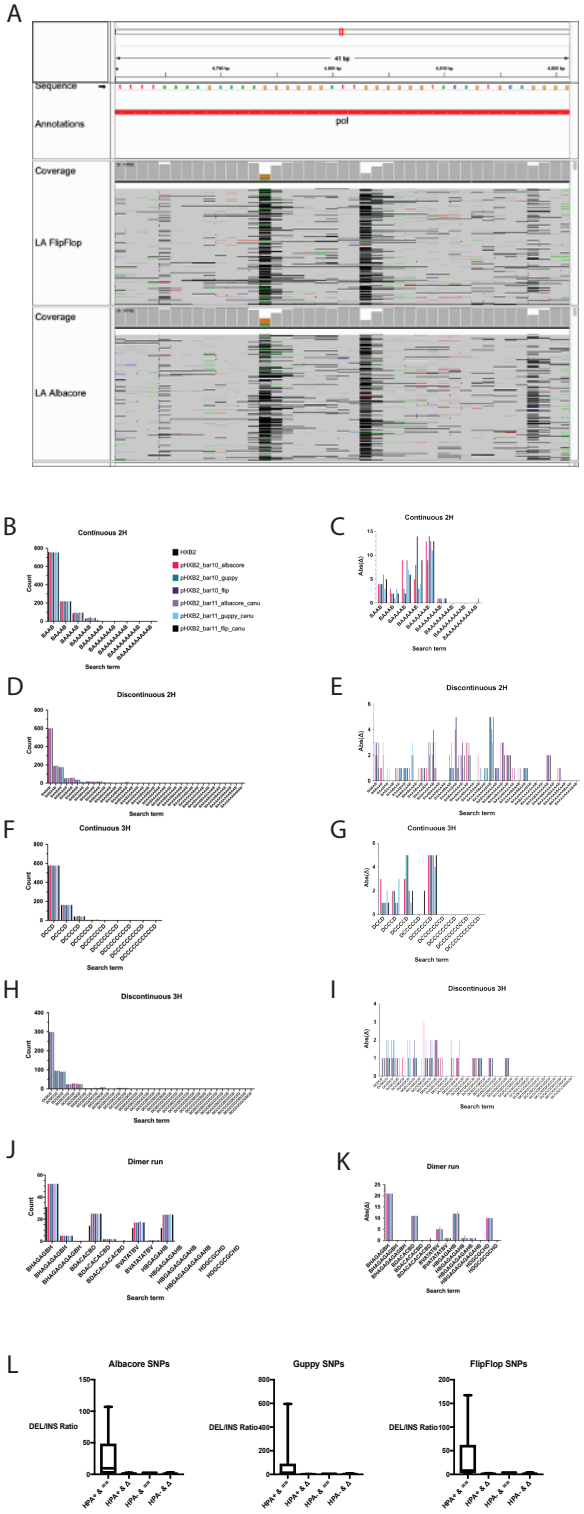
Supplemental Figure 3C: Mappability of long reads over contigs during assembly quality control



Coverage depends on context. Abrupt changes in coverage from terminal regions of HXB2 (**Figure 1F, Supplemental Figures 1 and 2)** were artifacts from supplying mappers with an HIV reference without a plasmid backbone. Long reads from barcode 10 (LA Taq) mapped with minimap2 [17] and "reference" contig from assembly (Canu v1.8) with basecalled data (FlipFlop). Stripes in this figure are not SNVs. They represent technical variability at homopolymers. Assemblies were manually curated to start with 5' LTR in the sense orientation, leaving the plasmid backbone on the left. Because there were not real insertions in the HIV segment, the HIV coordinates are the same as HXB2 (both 9719 bp long). Compare with 52 bp technical artifact from the core's short read assembly in **Supplemental Figure 3A, 3B**.

825 **Supplemental Figure 4: Homopolymers and dimer runs are ONT artifacts in unpolished**

826 **assemblies.**



827

828 Supplemental Figure 4A: A set of homopolymer tracks from HXB2 plasmid. Alignments with
829 BWA-MEM shown from FlipFlop (top) and Albacore (bottom) basecalled reads. Mapping is
830 pre-assembly.

831 Supplemental Figure 4B: Continuous 2H counts in unpolished assemblies. 2H = A or T
832 homodimers.

833 Supplemental Figure 4C: Continuous 2H Absolute Difference.

834 Supplemental Figure 4D: Discontinuous 2H counts in unpolished assemblies.

835 Supplemental Figure 4E: Discontinuous 2H Absolute Difference.

836 Supplemental Figure 4F: Continuous 3H counts in unpolished assemblies. 3H = C or G
837 homodimers.

838 Supplemental Figure 4G: Continuous 3H Absolute Difference.

839 Supplemental Figure 4H: Discontinuous 3H counts in unpolished assemblies.

840 Supplemental Figure 4I: Discontinuous 3H Absolute Difference.

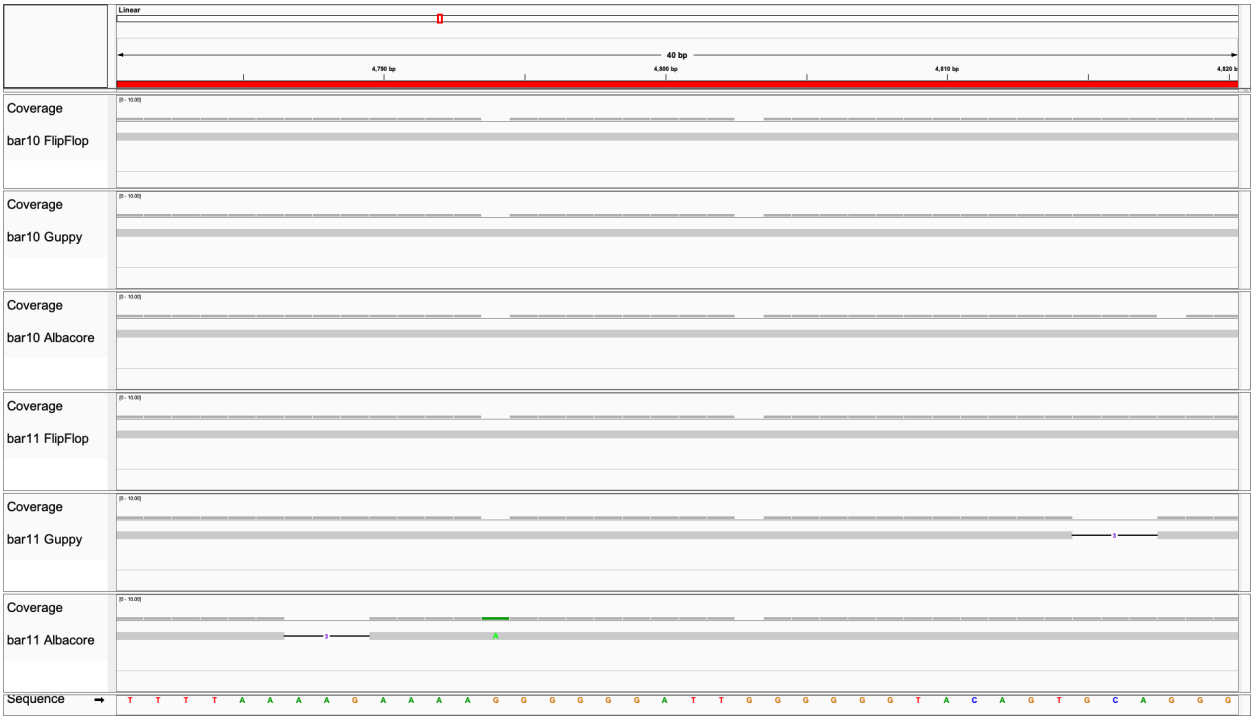
841 Supplemental Figure 4J: Dimer run counts in unpolished assemblies.

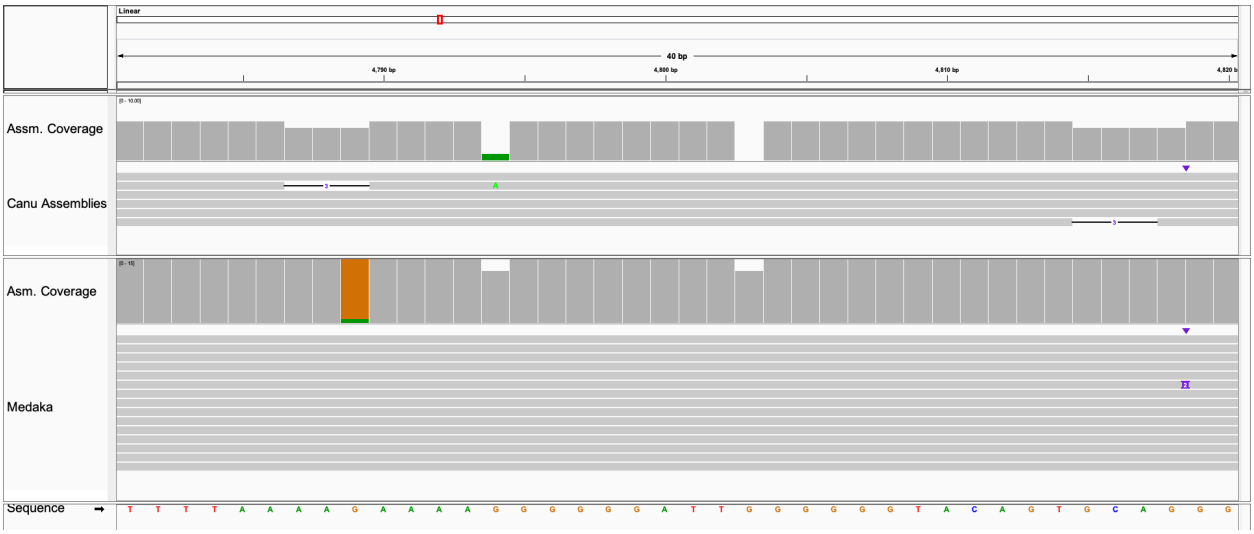
842 Supplemental Figure 4K: Dimer run Absolute Difference. Dimer runs as pairs are the most
843 problematic, with runs as triplets being resolvable by ONT.

844 Supplemental Figure 4L: The ratio of deletions to insertions is higher at mismatches both
845 adjacent to homopolymers and similar to neighbor bases. Box plot shows median ("x" is mean)
846 and quartile ranges. Y-axis is ratio. HPA: homopolymer-adjacent. ==: same as neighbor base. Δ:

different than neighbor base. Higher coverage (above ~10) usually makes up for current error profile. Above true for Albacore, Guppy, and FlipFlop.

Supplemental Figure 5: Assembly partially resolved homopolymers, which are improved by polishing





Top: Six ONT-only assemblies. Bottom: polished ONT-only assemblies, varying Medaka models. Deletions at 5’ of G homopolymers were not corrected, regardless of basecaller or Taq isoform. Note that polishing was not performed. IGV window is Linear:4,781-4,820. Bottom:

856 polishing canu assemblies with medaka abrogated most ONT artifacts. Best medaka setting
857 tested: r941_min_high_g330.

858

Word count:

Abstract: **245**/250

96 brackets – 46 references = 50 citations

Main text: 3935-57 (subheadings) – 50 citations = **3828**

Tables: 1

Figures: 5