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§], Wei Zou[5], Brian T. Foley[6], Deborah P. Hyink*[2], Paul E. Klotman*[1,2]

[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]Division of Infectious Diseases, the 1st Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China
[6]Theoretical Biology and Biophysics Group T-6, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

*Equal contributions.

[§]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, LTR phasing

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 and a control plasmid pNL4-3_gag-pol(Δ1443-4553)_EGFP 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 and pNL4-3_gag-pol(Δ1443-4553)_EGFP 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 and pNL4-3_gag-pol(Δ1443-4553)_EGFP 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. There were no variants detected in the HIV-1 segments of pNL4-3_gag-pol(Δ1443-4553)_EGFP.

Conclusions: Nanopore sequencing performed as-expected, phasing LTRs, and even covering full-length HIV. The discovery of variants in a reference plasmid demonstrates the need for

24 sequence verification moving forward, in line with calls from funding agencies for reagent
25 verification. These results illustrate the utility of long-read DNA-seq to advance the study of
26 HIV at single integration site resolution.
27

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 while sparing HIV-1 sequence, 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

51 infectious virion RNA and integrated proviral DNA **Figure 1B**), repetitive viral sequences like
52 long terminal repeats (LTRs), non-integrated forms [7], rarity of integration events *in vivo*
53 (reviewed in [8]), and alternative splicing of viral mRNAs [9]. Short-read DNA sequencing
54 (<150 base pairs (bp) in most reported experiments, but up to 500 bp for either Illumina
55 sequencing-by-synthesis or <1,000 bp for chain termination sequencing) provides some
56 information, but analyses require high coverage and/or extensive effort (non-exhaustive
57 examples [10], [11]). These factors limit the ability to assign variants to specific loci within each
58 provirus, as well as at the proviral integration site(s) (reviewed in [12]). Despite progress (HIV
59 DNA) [13], (HIV RNA) [14], [15], [16], researchers have yet to observe the genome of HIV-1 as
60 complete provirus (integrated DNA) in a single read, hindering locus-specific studies. To this
61 end, current long-read DNA sequencing clearly surpasses the limitations of read length of
62 leading next-generation/short-read sequencing platforms. Here we used the MinION sequencer
63 to sequence HIV-1 plasmid pHXB2_D in a pilot study focusing on coverage acquisition (as
64 opposed to full-length sequencing), with the goal of evaluating the technology for future
65 applications.
66
67

68 **Methods**

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

74 **HIV-1 plasmids**

75 A plasmid, "pHXB2_D" (alternate names pHXB2, pHXB-2D), believed to contain the
76 HIV-1 reference strain HXB2 [17] was acquired from the NIH AIDS Reagent and Reference
77 Program (ARP) via BioServe. pHXB2_D was believed to be a molecular clone (likely a
78 restriction product of HXB2 proviral DNA inserted into an unknown cloning plasmid backbone)
79 from one of the earliest clinical "HXB2" HIV-1 isolates. At the time of this work, it was
80 unknown whether this plasmid was ever sequence-verified before or after the reference sequence
81 for HXB2 was deposited.

82 The provenance of pNL4-3_gag-pol(Δ1443-4553)_EGFP, a reporter construct of pNL4-3
83 with a gag-pol deletion between base 1443 and 4553 is known. HIV-1 NL4-3 (pNL4-3) was a
84 fusion of NY5 and LAV/HXB2 plasmids [18] that to our knowledge are not readily available.
85 pEVd1443 [19] was a deletion construct made from pNL4-3 used to make several HIV-1
86 transgenic animals, including the FVB/N-Tg(HIV)26Aln/PkltJ (The Jackson Laboratory stock
87 No: 022354) "Tg26" mouse. The deletion in pEVd1443 was made by SphI cutting between
88 d1443 and 1444 with binding site 1443-1448, and cutting at a BalI site at 4551-4556 with blunt
89 cutting between 4553 and 4554. The EGFP cassette includes additional sequence upstream and
90 downstream of EGFP coding sequence. SphI and BalI may still be used to excise EGFP cassette.

91 A reporter construct was designed mimicking the pEVd1443 deletion: pNL4-3: ΔG/P-EGFP
92 [20]. Dr. Wei Zou rederived pNL4-3: ΔG/P-EGFP at BCM [21]. Both constructs (plasmid and
93 mouse) retained parts of gag and pol, with limited effects on protein-coding capacity, such as
94 expression of p17 [22]. Based on Addgene naming conventions, we suggest pNL4-3_gag-
95 pol(Δ1443-4553)_EGFP to replace the previous name pNL4-3: ΔG/P-EGFP for clarity.
96

97 **HIV-1 reference sequences**

98 The reference sequence of HXB2 is from the National Center for Biotechnology
99 Information (NCBI), Genbank accession number K03455.1. It runs from the beginning of the 5'
100 LTR to the end of the 3' LTR, and is 9,719 bp. This is similar to another HIV-1 reference that
101 NCBI uses, AF033819.3. This is a 9,181 base HXB2-like sequence that starts at the 96 bp repeat
102 in the 5'LTR, continues with the 5'UTR (U5), extends past the 3'UTR (U3) to the end of the 96
103 bp repeat in 3'LTR, with one SNV at the *vpu* start codon aTg to aCg at position AF033819.3:560
104 or K03455.1:6063. The reference sequence of NL4-3 is as a plasmid with accession number
105 AF324493.1. It runs from the beginning of the 5' LTR to the end of the 3' LTR, spanning 9,709
106 bp, and includes plasmid backbone with total length 14,825 bp.
107

108 **Long-read DNA sequencing**

109 A plasmid containing HXB2 was sequence-verified with long-read nanopore sequencing
110 on a MinION Mk1B (Oxford Nanopore Technologies, Oxford, UK). Unless otherwise noted,
111 reagents (and software) were purchased (or acquired) from Oxford Nanopore. Briefly, stock
112 plasmid was diluted to 5 ng final amount in ultrapure water (as two samples) and processed with
113 Rapid PCR Barcoding kit SQK-RPB004 along with 10 other barcoded samples (not discussed

114 further in this manuscript) following ONT protocol RPB_9059_V1_REVA_08MAR2018
115 (**Figure 1C**), a public description of which is here: https://store.nanoporetech.com/us/sample-
116 prep/rapid-pcr-barcoding-kit.html. Two DNA polymerases were evaluated (barcode 10 used
117 high-fidelity LA (for "long amplicon") Taq (Takara); barcode 11 Taq (Sigma-Aldrich). Libraries
118 were loaded onto a MinION flow cell version R9.4.1 and a 48-hour sequencing run was
119 completed with MinKNOW (version 1.10.11). Residual reads from subsequent runs were pooled
120 for final analyses. Long read data for pNL4-3_gag-pol(Δ1443-4553)_EGFP was generated in
121 other barcoded experiments (not shown).

122 Raw data was basecalled (converted from FAST5 to FASTQ format) with Albacore
123 version 2.3.4 (older basecaller), Guppy version 2.3.1 (current official at time of work), and
124 FlipFlop (Guppy development config). Mapping to reference was done with Minimap2 [23] and
125 BWA-MEM [24], implemented in Galaxy (usegalaxy.org) [25]. Alignments (.bam and .bai files)
126 were visualized in the Integrative Genomics Viewer [26] unless otherwise noted. For *de novo*
127 assembly, demultiplexed basecalled reads were fed into Canu version 1.8 [27]. Genome size was
128 estimated to be 16 Kb from agarose gel of undigested, but naturally degraded linearized
129 pHXB2_D (data not shown). SnapGene version 4.3.4 was used to manually annotate contigs
130 from Canu. Blastn (NCBI) was used to identify unknown regions of pHXB2_D. Polishing was
131 performed on ONT-only assemblies with Medaka (https://github.com/nanoporetech/medaka), in
132 Galaxy. Medaka models: r941_min_fast_g303, r941_min_high_g303, r941_min_high_g330.
133 Inference batch size (-b) = 100. The final pHXB2_D assembly and other full-length HIV clones
134 from the ARP were aligned to the most recent human reference genome (hg38) with Minimap2
135 in Galaxy with the following parameters: Long assembly to reference mapping (-k19 -w19 -A1 -
136 B19 -O39,81 -E3,1 -s200 -z200 --min-occ-floor=100).

137 **Statistics**

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

141 **Short-read DNA sequencing**

142 pHXB2_D and control pNL4-3_gag-pol(Δ1443-4553)_EGFP were provided as 35 ul at
143 ~63 ng/ul to the Center for Computational & Integrative Biology DNA Core at Massachusetts
144 General Hospital, an external DNA sequencing core specializing in high-throughput next
145 generation (short-read) plasmid sequencing and assembly. Neither HXB2/pNL4-3 reference
146 sequences nor pHXB2_D/pNL4-3_gag-pol(Δ1443-4553)_EGFP draft assemblies (from this
147 work) were provided to core staff at the time of sequencing so that testing would remain masked.
148 While the core's exact library prep is proprietary, multiplexed library prep and 150 single-end
149 Illumina (ILMN) sequencing were most likely performed on a MiSeq with platform-specific
150 reagents (V2 chemistry, per their website) and barcoding. Data was returned as FASTQ.
151 FASTQC [28] was used in Galaxy for in-house data quality control, and read lengths were all
152 142 bp per this tool. Mapping as above.

153 **Sequence comparisons**

154 We used MAFFT v7.475 [29], [30] to compare the LTR sequences of pHXB2_D and
155 HXB2, and pNL4-3 and pNL4-3_gag-pol(Δ1443-4553)_EGFP. For cladistics, we used BLAST
156 at HIV-DB (https://www.hiv.lanl.gov/content/sequence/BASIC_BLAST/basic_blast.html) to
157 find other HXB2-like genomes. The top 50 BLAST hits included many sequences pNL43 clones.
158 pNL4-3 is an artificial recombinant of the NY5 clone with LAV and/or the HXB2 clone [18].
159 The recombination point is marked by an EcoRI restriction site. We then made a multi-sequence

160 alignment with the final pHXB2_D assembly, the top BAST hits, and the HIV-1 M group
161 subtype reference set using GeneCutter
162 (https://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html), and built the
163 maximum likelihood tree using IQ-tree
164 (https://www.hiv.lanl.gov/content/sequence/IQTREE/iqtree.html). pNL4-3_gag-pol(Δ1443-
165 4553)_EGFP was not included in the above trees because of absence of divergence from pNL4-3
166 sequences outside of the EGFP cassette.
167

168 **Results**

169 Viewing mapped data in IGV, the long reads (median read length >2000 bp, **Figure 1E**)
170 from both pHXB2_D ONT experiments clearly covered each LTR (**Figure 1F, Supplemental**
171 **Figures 1**, **3**), while shorter reads collapsed into one of either LTR (**Figure 1F, Supplemental**
172 **Figures 3D,3E**). This was also seen when long reads were shorter than LTRs (<600 bp).
173 Mappers BWA-MEM and Minimap2 were chosen based on their ability to handle long and short
174 reads. Other mappers were not evaluated. BWA-MEM mapped more ambiguously, piling
175 partially mapped reads between each LTR; Minimap2 mapped with higher fidelity to reference
176 without splitting reads. Coverage as sequencing depth was higher and more even from the
177 higher-fidelity LA Taq library (**Supplemental Figure 1**). pNL4-3 was known to have distinct
178 LTRs because it was a synthetic recombinant. The higher variant density in NL4-3 LTRs enabled
179 mapping and phasing from short-read data only (**Supplemental Figure 2**).

180 We counted 20 single nucleotide variants (SNVs) in this reference clone of HXB2 (**Table**
181 **1, Supplemental Table 3, Supplemental Figure 3E**). These mismatches were seen in all Canu
182 assemblies (**Supplemental Figures 4A,4B**), verified in IGV and/or SnapGene, and were
183 orthogonally verified by short-read sequencing performed by the external core given masked
184 samples (**Supplemental Figure 3E**). These mismatches represent a ~0.21% divergence from
185 reference HXB2 K03455.1 (20/9719), which was assumed to have perfect identity (0%
186 divergence). Transitions were more common (14/20) (**Table 1**), coinciding with a previous
187 report of increased transitions over transversions in infection models, because transversions are
188 more likely to be deleterious to viral replication (i.e.: to cause protein-coding changes) [31].
189 Indeed, almost half (9/20) of the observed SNVs occurred in protein-coding regions, even though
190 92% of HXB2 is coding (791/9719). Of those 9 SNVs in protein-coding regions, 4 caused non-

191 synonymous mutations. One of those occurs in a region overlapping both gag and pol regions,
192 however only pol exhibited a non-synonymous change from valine to isoleucine in p6, at
193 position 2259 relative to HXB2. Other non-synonymous variants occurred at 4609 (in p31
194 integrase, arginine to lysine), 7823 (in ASP antisense protein, glycine to arginine), and 9253 (in
195 nef, isoleucine to valine). 11/20 SNVs were in LTRs (see **Supplemental Figure 3** for counting
196 based on mapping); 8/20 of these would have been missed with mapping-only variant calling or
197 consensus. The longest HIV-mapping read (**Figure 2**) phased 16/20 SNVs (failed at sites
198 2,8,10,12, **Table 1**). pNL4-3_gag-pol(Δ1443-4553)_EGFP did not have HIV-1 or plasmid
199 backbone variants supported by long and short reads outside of the EGFP cassette.

200 We assembled the previously undefined plasmid pHXB2_D (**Supplemental Figures**
201 **4A,4B**). Canu's final output was a set of contiguous DNA sequences (contigs) as FASTA files. A
202 consequence of assembling plasmid sequences with this tool was partial redundancy at contig
203 ends (**Supplemental Figure 4C)**. Manual end-trimming of contigs was performed in SnapGene
204 based on an estimated length of 16 kilobases. Top blastn hits from barcode 10/LA Taq pHXB2
205 basecalled with FlipFlop were as follows: for the main backbone (with origin of replication and
206 antibiotic selection cassette for cloning), shuttle vector pTB101-CM DNA, complete sequence
207 (based on pBR322), from 4352-8340; for the upstream element (relative to 5' LTR), Homo
208 sapiens chromosome 3 clone RP11-83E7 map 3p, complete sequence from 58,052 to 59,165; for
209 the downstream element, cloning vector pNHG-CapNM from 10,204 to 11,666. Other identified
210 elements included Enterobacteria phage SP6 (the SP6 promoter, per SnapGene's "Detect
211 common features"), complete sequence from 39,683 to 39,966. Identities of query to HXB2 and
212 hits were all approximately 99%. The MGH CCIB DNA Core's proprietary *de novo* UltraCycler
213 v1.0 assembler (Brian Seed and Huajun Wang, unpublished) was able to assemble both 5' and 3'

214 LTRs with short-read data only but may have collapsed SNVs into an artificial single consensus.
215 Long-read mapping and assembly (and polished assemblies) orthogonally validated LTRs, and
216 supported a single HIV-1 HXB2_D haplotype (**Supplemental Figure 4,6**). A final LTR-phased
217 and annotated assembly leveraging short and long reads is provided as pHXB2_D
218 Genbank:MW079479 (embargoed until publication). Importantly, for pHXB2_D, each LTR was
219 identical, which is distinct from the current HXB2 (K03455.1) (**Figure 3A**). Compared to pNL4-
220 3_gag-pol(Δ1443-4553)_EGFP (ACCESSION_TBD) , each LTR was distinct, but identical to
221 pNL4-3's distinct 5' and 3' LTRs (AF324493.1) (**Figures 3B,6**).

222 To determine whether pHXB2_D was an isolated provirus (as opposed to a cDNA clone),
223 the pHXB2_D assembly was aligned to the current human reference hg38, returning a single
224 complete insertion site on 3p24.3 (**Figure 4A, Supplemental Table 2**). As expected, our pNL4-
225 3_gag-pol(Δ1443-4553)_EGFP had homology arms from two chromosomes (**Figures 4B,6,**
226 **Supplemental Table 2**). We sought to put our pHXB2_D assembly into context of other HXB2-
227 like references available (**Figure 5**). pHXB2_D (red) clusters closely with HXB2 reference
228 (K03455) and related clone sequences (green). pNL4-3 clones in blue. The LTR-masked HIV-
229 spanning segment of pHXB2_D is most homologous to B.FR.1983.DM461230 and
230 B.FR.1983.CS793683, which are identical except for areas in nef and a GFP insertion (verified
231 by blastn). This finding suggests they were from the same stock. HIV-1 M group subtype
232 reference set (HIV Sequence Database) was added to put HXB2s and pNL4-3 clones into
233 perspective. HXB2 (believed to be a complete isolate) and NL4-3 (synthetic clone based on two
234 early isolates [18]) are examples of HIV type 1 (HIV-1), group M, subgroup B.

235 As previously reported [32], per-read variability in ONT data was higher near
236 homopolymers (runs of the same base) (**Supplemental Figure 5A**). For the datasets generated in

237 the present study, homopolymers were counted and classified as continuous (unbroken run of a
238 given nucleobase) vs. discontinuous (broken run of a given nucleobase) (**Supplemental Figures**
239 **5B,5D,5F,5H**). A/T (2 hydrogen bonds; 2H) and G/C (3 hydrogen bonds; 3H) were evaluated.
240 Because runs longer than 4 or 5 were rare in these datasets, it was impossible to evaluate longer
241 homopolymers. A simple calculation Abs(Δ)=Abs(#homopolymers$_{reference}$ -
242 #homopolymers$_{assembly}$) helped to evaluate the performance of basecallers, such that better
243 basecallers had smaller Abs(Δ) (**Supplemental Figures 5C,5E,5G,5I,5K**). At the level of
244 consensus (made from sequences mapped to reference HXB2), homopolymers contributed few,
245 if any, obvious errors. A special case of homopolymer, dimer runs, was noted to cause persistent
246 errors regardless of ONT basecaller (**Supplemental Figures 5J,5K**). While dips occurred at
247 certain points near homopolymers, the consensus did not change much at the sequencing depth
248 used in this study for either barcoded pHXB2_D samples (**Supplemental Figures 1,3,4**).
249 Another interpretation is that homopolymers tend to seem truncated with ONT, with more reads
250 in support of shorter homopolymers. Canu assemblies showed basecaller-dependent variability
251 (**Supplemental Table 3**). That said, newer basecallers tended to produce fewer and smaller per-
252 read truncations. Assemblies without polishing did not correct all homopolymer truncations
253 (**Supplemental Figure 4A**). Polishing assemblies tended to correct these toward the final
254 pHXB2_D assembly (**Supplemental Figures 4B,6**). Data from polished ONT-only assemblies
255 and short-read sequencing do not support the truncations (gaps relative to reference) suggested
256 by unpolished ONT-only assemblies, representing a known current limitation of ONT. These are
257 not the same as the 20 SNVs supported by BOTH long- and short-read sequencing performed in
258 this study. The ratio of per-read deletions to per-read insertions (DEL/INS) was much higher for
259 SNVs occurring at homopolymers and near the same base, and this difference was maintained

260 between all basecallers used (**Supplemental Figure 5L**). These changes created more

261 problematic (longer) homopolymers.

262

263

264 ## Discussion

265 This work represents the first instance of complete and unambiguous sequencing of HIV-
266 1 provirus as plasmid and contributed to the identification of single nucleotide variants which
267 may not have been easily determined using other sequencing modalities, illustrating the
268 importance of validating molecular reagents in their entirety, and with complementary
269 approaches. Nanopore sequencing surpassed the read length limitations of traditional sequencing
270 modalities used for HIV such as Sanger sequencing and sequencing-by-synthesis by at least two
271 orders of magnitude. Other long-read DNA sequencing technologies such as PacBio's zero-mode
272 waveguide DNA sequencing were not evaluated in this work, but in principle would be
273 interchangeable for nanopore sequencing. Paired-end sequencing (as either DNA-seq or RNA-
274 seq) was not evaluated in this work, but has shown promise phasing LTRs in our hands [33]–
275 [35].

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

277 HIV provirus is believed to occur naturally as one or a few copies of reverse-transcribed
278 DNA forms integrated into the host nuclear genome. Depending on where integration occurs,
279 local GC or AT content might cause problems for detecting integrants with PCR. HIV also has
280 conserved transitions from areas of higher GC content (~60%) to content approximating average
281 human GC content (~40%). To limit PCR sequencing bias and to accommodate for the potential
282 heterogeneity of HIV sequences, we fractionated whole sample directly (as opposed to PCR-
283 barcoding select amplicons) with tagmentation provided in the Rapid PCR-Barcoding kit (ONT).
284 Tagmentation in this workd used transposon-mediated cleavage and ligation of barcode adapters
285 for later PCR amplification. A consequence of this fractionation was a distribution of reads
286 (**Figure 1E**) shorter than longer reads reported elsewhere for ONT experiments [36]. Based on

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

291 **Long reads enable LTR phasing and HIV haplotype definition**

292 We created 6 assemblies for pHXB2_D from ONT-only data (**Supplemental Figure 4**),
293 each with a common set of 20 SNVs (11 in LTRs), and final assemblies (a single HIV-1
294 HXB2_D haplotype; a single HIV-1 NL4-3_gag-pol(Δ1443-4553)_EGFP haplotype) leveraging
295 long- and short-read data. The external core's *de novo* assembly pipeline identified the same 20
296 SNVs, and variants in the LTRs were supported by ONT unambiguously. That the core's
297 assembler was able to phase LTR variants in these samples may have been because the samples
298 had high amounts of the same upstream and downstream sequences because of coming from one
299 plasmid. The core's assembler thus may have had additional sequencing information at the edges
300 of HXB2, helping it to map deeper into each LTR. This approach would likely fail in samples
301 with multiple integrations (as in various animal models of HIV disease [37]), which have
302 unknown upstream and downstream sequences, or in samples from natural human infection,
303 which is well known to exhibit multiple pseudo-random integration sites between cells [38],
304 [39], but with mostly single integration events per cell [8]. Inverse PCR (iPCR) is an alternative
305 method [40] with its own issues (e.g., PCR biases, HIV concatemers, host repeats). While current
306 PCR reagents have extended the range of what can be seen with iPCR, current approaches are
307 likewise limited by long DNA extraction methods, sample amount, and remain to be optimized.
308 If coverage is sufficient (≥10 reads in non-homopolymers and non-dimer runs), long-read
309 sequencing can provide linked variant information to individual integration sites. Identical 5' and

310 ‘3 LTRs (**Figure 3**) in the context of a single integration event (**Figure 4A**) support this integrant
311 being a *bona fide* provirus [41]. Other proviruses also had identical LTR pairs (**Supplemental**
312 **Table 2**). Technical limitations such as PCR errors before earlier sequencing may explain the
313 variability in the HXB2 reference LTRs. These were sequenced at a time before paired-end 150
314 or long-read DNA-seq were available to phase LTRs, raising the possibility that these LTRs
315 were incorrectly annotated by depositors assuming identity and copy-and-pasting the sequence of
316 one LTR for both without being able to unambiguously resolve each LTR.

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

318 Up until 2020, HIV had been the most studied human pathogen, but HIV reagents are not
319 routinely re(verified). The pHXB2_D sequenced was allegedly a reference plasmid, with
320 unknown divergence between the published reference HXB2. Three independent experiments
321 (two long-read with PCR-barcoded libraries made with regular and long-amplicon Taq master
322 mixes, one short-read) yielded at least 20 single nucleotide variants in pHXB2_D which differed
323 from the HXB2 reference sequence (**Table 1, Supplemental Figure 3**), which were also
324 concordant across the three basecallers used (**Supplemental Table 3**) and are therefore not PCR
325 errors. By leveraging long reads with the MinION, we were able to find mutations in highly
326 repetitive LTRs relative to HXB2 Genbank:K03455.1 which are often assumed (but until now
327 never proven) to be identical (**Table 1**, **Figure 1**, **Supplemental Figures 1, 3E**), as well as
328 mutations in protein-coding regions (**Table 1**). We were also able to confirm that the backbone
329 of this plasmid is from pSP62 [17], a pBR322 derivative with the SP6 promoter [42], aiding in
330 the continued use of this important reagent, and illustrating the need of full-length reagent
331 validation moving forward. We suggest that all clinical reagents (e.g., vectors) be sequence-

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

334 **Improvement in ONT basecallers over time**

335 Albacore, Guppy, and FlipFlop basecallers were compared. Each produced reads of
336 similar length distributions (relative to polymerase used), while Guppy and FlipFlop produced
337 improved and best performance relative to quality score distributions (**Figure 1D**). Interestingly,
338 while read length distributions were affected by fidelity of polymerases evaluated in this work,
339 mean quality distributions were not. This is important because of the differences in cost between
340 higher fidelity Taq and classic Taq enzymes. That said, higher fidelity LA Taq produced much
341 higher coverage compared to Taq (**Supplemental Figure 1**). In consideration of library prep,
342 choice of enzyme used should be based on the desired read-length distribution and coverage.
343 Regarding read mapping, the increase in mean quality score between these basecallers improved
344 overall mapping, in part by facilitating demultiplexing, resulting in approximately ~10%
345 increases number of reads in barcoded libraries before mapping (shift in reads from unclassified
346 to a given barcode). FlipFlop tended to handle homopolymers better than previous basecallers
347 (**Supplemental Figures 5,6**). Homopolymers in HXB2 tended to exhibit apparent deletions near
348 5' ends of homopolymers (upstream due to technical artifact from mapping), but because
349 consensus is conserved (example, at least 80% of base in called read set is identical to reference),
350 and because short-read data lacks INDELS at these sites, it is unlikely that any of these
351 homopolymer deletions are real in these experiments. Dimer runs – stretches of repeating 2-mers
352 (pronounced "two-mers") – proved challenging regardless of basecaller. Mapping as above may
353 be used to aid in manually calling these when they occur. Albacore is currently deprecated, and
354 current versions of Guppy now incorporate a version of FlipFlop called Guppy High-ACcuracy

355 (HAC). Guppy HAC and subsequence versions were not evaluated in this work. Polishing is
356 becoming standard practice for processing assemblies from ONT data because it redresses most
357 homopolymer errors propagated into long-read-only assemblies. The best manually finished and
358 polished contig had 1 error out of 16,722 bases, illustrating the utility of ONT hardware when
359 paired with burgeoning software.
360

361 **Conclusions**

362 HIV informatics, the study of HIV sequence information, has been limited by the
363 common assumption that sequence fidelity exists between reference genomes available in
364 sequence databases and similarly named HIV clones. Modern DNA sequencing methods, such as
365 long- and short-read sequencing, are available to redress this issue. Long-read sequencing fills in
366 gaps left behind by short-read interrogation of HIV-1. Current limitations of the approaches used
367 in the present work to study HIV are 1.) the cost of long-read sequencing, regardless of platform,
368 compared to the cheaper short reads from sequencing-by-synthesis, 2.) long DNA extraction
369 methods in diseased tissue (Gener, unpublished), and 3.) the lower per-base accuracy (low-mid
370 90's with ONT vs. 98-99% with ILMN or newer PacBio HiFi), including difficulty near
371 homopolymers and dimer runs (**Supplemental Figure 5**). A nontrivial but redressable limitation
372 is availability of personnel trained to prepare sequencing libraries, to run sequencing, and to
373 analyze results. As the price of long-read sequencing decreases, hardware and software used in
374 basecalling and library protocols improve, and with the advent of more user-friendly tools, the
375 cost of obtaining usable data from long reads will become negligible compared to the ability to
376 answer historically intractable questions. This work raises the possibility of being able to detect
377 at least some recombination events, in a reference-free manner requiring only the comparison of
378 LTRs from the same integrants (**Figure 6**). We suggest that pHXB2_D and pNL4-3 constructs
379 may be used as negative and positive controls for the development of such screens. While other
380 HIV reference proviral clones were reported to have identical LTR pairs, this remains to be
381 tested in other clones, since other clones were generated with shorter sequencing methods. For
382 example, pNL4-3_gag-pol(Δ1443-4553)_EGFP had distinct LTRs as a plasmid. However, if an
383 NL4-3 virus is made from pNL4-3, the LTR sequences would homogenize to pNL4-3's 3' LTR

384 sequence. Future work will include optimizing DNA extraction protocols with the goal of
385 capturing higher-coverage fuller glimpses of each HIV proviral integration site in *in vivo* HIV
386 models and patient samples. This work has broad implications for all cells infected by both
387 integrating and non-integrating viruses, and for the characterization of targeted regions in the
388 genome which may be recalcitrant to previous sequencing methods. Long-read sequencing is an
389 important emerging tool defining the post-scaffold genomic era, allowing for the characterization
390 of anatomical landmarks of hosts and pathogens at the genomic scale.

391 **Disclaimer**

392 Erratum: Preprint version 1 of this work [43] incorrectly cited the Integrated Genome
393 Browser for work that was completed with the Integrative Genomics Viewer. Apologies for the
394 mistake.

395 **Funding**

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

404 **Competing interests**

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

407 **Authors' contributions**

408 ARG conceived of this project, performed experiments, analyzed results, and drafted the
409 manuscript. WZ rederived pNL4-3_gag-pol(Δ1443-4553)_EGFP. All authors discussed data and
410 edited the manuscript. ARG and PK provided funding.

411 **Acknowledgements**

412 As part of a summer bioinformatics internship in the Paul E. Klotman Laboratory at
413 Baylor College of Medicine, Akash Naik supervised by ARG performed *in silico* mapping

414 analyses/experiments, generated and/or aided in the synthesis of **Supplemental Figure 4**, and
415 assisted in writing relevant portions, discussing, and editing this manuscript. During a second
416 summer internship with American Physician Scientists Association Virtual Summer Research
417 Program, the following students were supervised by ARG helped to create **Figure 1A** and
418 **Supplemental Table 1**: Yini Liang, Kirk Niekamp, Maliha Jeba, Delmarie M. Rivera
419 Rodríguez. Orthogonal sequence verification was performed as a service by staff at the Center
420 for Computational & Integrative Biology DNA Core at Massachusetts General Hospital, Boston,
421 MA, USA.

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

431 **Available additional files**

432 Albacore basecalled barcode 10

433 Guppy basecalled barcode 10

434 FlipFlop basecalled barcode 10

435 Albacore basecalled barcode 11

436 Guppy basecalled barcode 11

437 FlipFlop basecalled barcode 11

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

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

440 .dna files of contigs (n=6)

441 MGH data (raw + contig)

442 Supplemental Tables

443 Supplemental Figures

444

445

446

447 **References**

448 [1] F. Barré-Sinoussi *et al.*, "Isolation of a T-lymphotropic retrovirus from a patient at risk for
449 acquired immune deficiency syndrome (AIDS).," *Science*, vol. 220, no. 4599, pp. 868–
450 871, May 1983, doi: 10.1126/science.6189183.
451 [2] S. Wain-Hobson *et al.*, "LAV revisited: origins of the early HIV-1 isolates from Institut
452 Pasteur.," *Science*, vol. 252, no. 5008, pp. 961–965, May 1991, doi:
453 10.1126/science.2035026.
454 [3] L. Ratner *et al.*, "Complete nucleotide sequence of the AIDS virus, HTLV-III.," *Nature*,
455 vol. 313, no. 6000, pp. 277–284, Jan. 1985, doi: 10.1038/313277a0.
456 [4] A. R. Gener and J. T. Kimata, "Full-coverage native RNA sequencing of HIV-1 viruses,"
457 *bioRxiv*, p. 845610, Jan. 2019, doi: 10.1101/845610.
458 [5] G. M. Shaw, B. H. Hahn, S. K. Arya, J. E. Groopman, R. C. Gallo, and F. Wong-Staal,
459 "Molecular characterization of human T-cell leukemia (lymphotropic) virus type III in the
460 acquired immune deficiency syndrome.," *Science*, vol. 226, no. 4679, pp. 1165–1171,
461 Dec. 1984, doi: 10.1126/science.6095449.
462 [6] M. Krupovic *et al.*, "Ortervirales: New Virus Order Unifying Five Families of Reverse-
463 Transcribing Viruses," *J. Virol.*, vol. 92, no. 12, pp. e00515-18, May 2018, doi:
464 10.1128/JVI.00515-18.
465 [7] E. H. Graf *et al.*, "Elite suppressors harbor low levels of integrated HIV DNA and high
466 levels of 2-LTR circular HIV DNA compared to HIV+ patients on and off HAART,"
467 *PLoS Pathog.*, vol. 7, no. 2, 2011, doi: 10.1371/journal.ppat.1001300.
468 [8] Y. Ito *et al.*, "Number of infection events per cell during HIV-1 cell-free infection," *Sci.*
469 *Rep.*, vol. 7, no. 1, p. 6559, 2017, doi: 10.1038/s41598-017-03954-9.

470 [9] I. Cuesta, A. Mari, A. Ocampo, C. Miralles, S. Pérez-castro, and M. M. Thomson,
471 "Sequence Analysis of In Vivo -Expressed HIV-1 Spliced RNAs Reveals the Usage of
472 New and Unusual Splice Sites by Viruses of Different Subtypes," pp. 1–24, 2016, doi:
473 10.1371/journal.pone.0158525.
474 [10] C. Wymant *et al.*, "Easy and accurate reconstruction of whole HIV genomes from short-
475 read sequence data with shiver," *Virus Evol.*, vol. 4, no. 1, pp. 1–13, 2018, doi:
476 10.1093/ve/vey007.
477 [11] K. M. Bruner *et al.*, "A quantitative approach for measuring the reservoir of latent HIV-1
478 proviruses," *Nature*, vol. 566, no. 7742, pp. 120–125, 2019, doi: 10.1038/s41586-019-
479 0898-8.
480 [12] M. R. Pinzone and U. O'Doherty, "Measuring integrated HIV DNA ex vivo and in vitro
481 provides insights about how reservoirs are formed and maintained," *Retrovirology*, vol.
482 15, no. 1, pp. 1–12, 2018, doi: 10.1186/s12977-018-0396-3.
483 [13] K. B. Einkauf *et al.*, "Intact HIV-1 proviruses accumulate at distinct chromosomal
484 positions during prolonged antiretroviral therapy Find the latest version : Intact HIV-1
485 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral
486 therapy," vol. 129, no. 3, pp. 988–998, 2019.
487 [14] D. Bonsall *et al.*, "THAA0101 - HIV genotyping and phylogenetics in the HPTN 071
488 (PopART) study: Validation of a high-throughput sequencing assay for viral load
489 quantification, genotyping, resistance testing and high-resolution transmission
490 networking," in *22nd International AIDS Conference (AIDS2018)*, 2018, p. Oral Abstract.
491 [15] A. N. Banin *et al.*, "Development of a Versatile, Near Full Genome Amplification and
492 Sequencing Approach for a Broad Variety of HIV-1 Group M Variants," *Viruses*, vol. 11,

493 no. 4, p. 317, Apr. 2019, doi: 10.3390/v11040317.

494 [16] N. Nguyen Quang *et al.*, "Dynamic nanopore long-read sequencing analysis of HIV-1
495 splicing events during the early steps of infection," *Retrovirology*, vol. 17, no. 1, p. 25,
496 2020, doi: 10.1186/s12977-020-00533-1.

497 [17] A. G. Fisher, E. Collalti, L. Ratner, R. C. Gallo, and F. Wong-Staal, "A molecular clone of
498 HTLV-III with biological activity," *Nature*, vol. 316, no. 6025, pp. 262–265, 1985, doi:
499 10.1038/316262a0.

500 [18] A. Adachi *et al.*, "Production of acquired immunodeficiency syndrome-associated
501 retrovirus in human and nonhuman cells transfected with an infectious molecular clone.,"
502 *J. Virol.*, vol. 59, no. 2, pp. 284–91, 1986.

503 [19] P. Dickie *et al.*, "HIV-associated nephropathy in transgenic mice expressing HIV-1
504 genes," *Virology*, 1991. [Online]. Available: http://ac.els-cdn.com/0042682291907595/1-
505 s2.0-0042682291907595-main.pdf?_tid=8f811f10-d10c-11e5-82e8-
506 00000aacb35e&acdnat=1455228938_33d4226549c6410971ced1c4c3573a44. [Accessed:
507 11-Feb-2016].

508 [20] M. Husain, "HIV-1 Nef Induces Proliferation and Anchorage-Independent Growth in
509 Podocytes," *J. Am. Soc. Nephrol.*, vol. 13, no. 7, pp. 1806–1815, 2002, doi:
510 10.1097/01.ASN.0000019642.55998.69.

511 [21] H. Li *et al.*, "Epigenetic regulation of RCAN1 expression in kidney disease and its role in
512 podocyte injury," *Kidney Int.*, vol. 94, no. 6, pp. 1160–1176, 2018, doi:
513 10.1016/j.kint.2018.07.023.

514 [22] S. Curreli *et al.*, "B cell lymphoma in HIV transgenic mice.," *Retrovirology*, vol. 10, p.
515 92, Jan. 2013, doi: 10.1186/1742-4690-10-92.

516 [23] H. Li, “Minimap2: pairwise alignment for nucleotide sequences,” *Bioinformatics*, vol. 34,
517 no. 18, pp. 3094–3100, May 2018, doi: 10.1093/bioinformatics/bty191.
518 [24] H. Li and R. Durbin, “Fast and accurate long-read alignment with Burrows – Wheeler
519 transform,” vol. 26, no. 5, pp. 589–595, 2010, doi: 10.1093/bioinformatics/btp698.
520 [25] E. Afgan *et al.*, “The Galaxy platform for accessible, reproducible and collaborative
521 biomedical analyses: 2016 update,” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W3–W10,
522 2016, doi: 10.1093/nar/gkw343.
523 [26] J. T. Robinson *et al.*, “Integrative genomics viewer,” *Nat Biotechnol*, vol. 29, no. 1, pp.
524 24–26, 2011, doi: 10.1038/nbt0111-24.
525 [27] B. P. Walenz, S. Koren, N. H. Bergman, A. M. Phillippy, J. R. Miller, and K. Berlin,
526 “Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat
527 separation,” *Genome Res.*, vol. 27, no. 5, pp. 722–736, 2017, doi: 10.1101/gr.215087.116.
528 [28] S. Andrews, “FastQC A Quality Control tool for High Throughput Sequence Data.”
529 [29] K. Katoh and D. M. Standley, “MAFFT multiple sequence alignment software version 7:
530 Improvements in performance and usability,” *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780,
531 2013, doi: 10.1093/molbev/mst010.
532 [30] K. Katoh, J. Rozewicki, and K. D. Yamada, “MAFFT online service: multiple sequence
533 alignment, interactive sequence choice and visualization,” *Brief. Bioinform.*, vol. 20, no. 4,
534 pp. 1160–1166, Sep. 2017, doi: 10.1093/bib/bbx108.
535 [31] D. M. Lyons and A. S. Lauring, “Evidence for the Selective Basis of Transition-to-
536 Transversion Substitution Bias in Two RNA Viruses,” *Mol. Biol. Evol.*, vol. 34, no. 12,
537 pp. 3205–3215, 2017, doi: 10.1093/molbev/msx251.
538 [32] N. J. Loman, J. Quick, and J. T. Simpson, “A complete bacterial genome assembled de

539 novo using only nanopore sequencing data," *Nat. Methods*, vol. 12, no. 8, pp. 733–735,
540 2015, doi: 10.1038/nmeth.3444.
541 [33] A. Gener *et al.*, "PEA0011 - Insights from HIV-1 transgene insertions in the murine
542 model of HIV-associated nephropathy," in *23rd International AIDS Conference*
543 *(AIDS2020)*, 2020, vol. ePoster.
544 [34] A. R. Gener *et al.*, "P39 - Insights from comprehensive transcript models of HIV-1," in
545 *Genome Informatics 2020*, 2020, p. ePoster.
546 [35] A. R. Gener, T. Washington, D. Hyink, and P. Klotman, "3264 - The Multiple HIV-1
547 Transgenes in the Murine Model of HIV-Associated Nephropathy Fail to Segregate as
548 Expected," in *American Society of Human Genetics Annual Meeting*, 2020, p. ePoster.
549 [36] A. Payne, N. Holmes, V. Rakyan, and M. Loose, "Whale watching with BulkVis: A
550 graphical viewer for Oxford Nanopore bulk fast5 files," *bioRxiv*, p. 312256, Jan. 2018,
551 doi: 10.1101/312256.
552 [37] P. Rosenstiel, A. Gharavi, V. D'Agati, and P. Klotman, "Transgenic and infectious animal
553 models of HIV-associated nephropathy.," *J. Am. Soc. Nephrol.*, vol. 20, no. 11, pp. 2296–
554 304, 2009, doi: 10.1681/ASN.2008121230.
555 [38] M. Kvaratskhelia, A. Sharma, R. C. Larue, E. Serrao, and A. Engelman, "Molecular
556 mechanisms of retroviral integration site selection.," *Nucleic Acids Res.*, vol. 42, no. 16,
557 pp. gku769-, 2014, doi: 10.1093/nar/gku769.
558 [39] B. Marini *et al.*, "Nuclear architecture dictates HIV-1 integration site selection," *Nature*,
559 vol. 521, pp. 227–233, 2015, doi: 10.1038/nature14226.
560 [40] H. Ochman, A. S. Gerber, and D. L. Hartl, "Genetic applications of an inverse polymerase
561 chain reaction," *Genetics*, vol. 120, no. 3, pp. 621–623, 1988.

562 [41] W.-S. Hu and S. H. Hughes, "HIV-1 Reverse Transcription," *Cold Spring Harb. Perspect.*
563 *Med.* , vol. 2, no. 10, Oct. 2012, doi: 10.1101/cshperspect.a006882.
564 [42] M. R. Green, T. Maniatis, and D. A. Melton, "Human beta-globin pre-mRNA synthesized
565 in vitro is accurately spliced in Xenopus oocyte nuclei.," *Cell*, vol. 32, no. 3, pp. 681–
566 694, Mar. 1983, doi: 10.1016/0092-8674(83)90054-5.
567 [43] A. R. Gener, "Full-coverage sequencing of HIV-1 provirus from a reference plasmid,"
568 *bioRxiv*, p. 611848, Jan. 2019, doi: 10.1101/611848.
569 [44] B. Lucic *et al.*, "Spatially clustered loci with multiple enhancers are frequent targets of
570 HIV-1," *bioRxiv*, 2018.
571 [45] W. J. Kent *et al.*, "The Human Genome Browser at UCSC," *Genome Res.* , vol. 12, no. 6,
572 pp. 996–1006, Jun. 2002, doi: 10.1101/gr.229102.
573 [46] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "IDBA – A Practical Iterative de
574 Bruijn Graph De Novo Assembler BT - Research in Computational Molecular Biology,"
575 2010, pp. 426–440.
576 [47] P. J. A. Cock, B. A. Grüning, K. Paszkiewicz, and L. Pritchard, "Galaxy tools and
577 workflows for sequence analysis with applications in molecular plant pathology," *PeerJ*,
578 vol. 1, p. e167, 2013, doi: 10.7717/peerj.167.
579 [48] C. B, W. T, and S. S, "Genome sequence assembly using trace signals and additional
580 sequence information.," *Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma.*, vol. 99, pp. 45–
581 56.
582 [49] A. Bankevich *et al.*, "SPAdes: A New Genome Assembly Algorithm and Its Applications
583 to Single-Cell Sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, Apr. 2012, doi:
584 10.1089/cmb.2012.0021.

585 [50] G. Cuccuru *et al.*, "Orione, a web-based framework for NGS analysis in microbiology,"
586 *Bioinformatics*, vol. 30, no. 13, pp. 1928–1929, Jul. 2014, doi:
587 10.1093/bioinformatics/btu135.
588 [51] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt, "Assembling millions of short
589 DNA sequences using SSAKE," *Bioinformatics*, vol. 23, no. 4, pp. 500–501, Feb. 2007,
590 doi: 10.1093/bioinformatics/btl629.
591
592

Tables

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

598 Content. Base-1 (first base is numbered 1, 2nd 2, etc.), relative to HXB2, Genbank:K03455.1.
599 Changed base represented as upper-case. Annotated as codon if in protein-coding region. No
600 deletions or insertions were predicted from manual inspection or supported by short-read
601 sequencing. Abbreviations, ASP: antisense protein, RRE: rev-response element, NA: not
602 applicable. Syn: synonymous mutation. Non: non-synonymous mutation. Stop: stop codon/non-
603 sense mutation. LTR: long terminal repeat. RT: reverse transcriptase. IN: integrase. LANL: Los
604 Alamos National Laboratory HIV Sequence Database. Data from three separate sequencing
605 experiments on the same plasmid sample support these 20 sites. Note site 1-8 variants in 5'LTR
606 have been previously reported (LANL), albethey ambiguously. These may also be incorrectly
607 annotated as variants in nef.

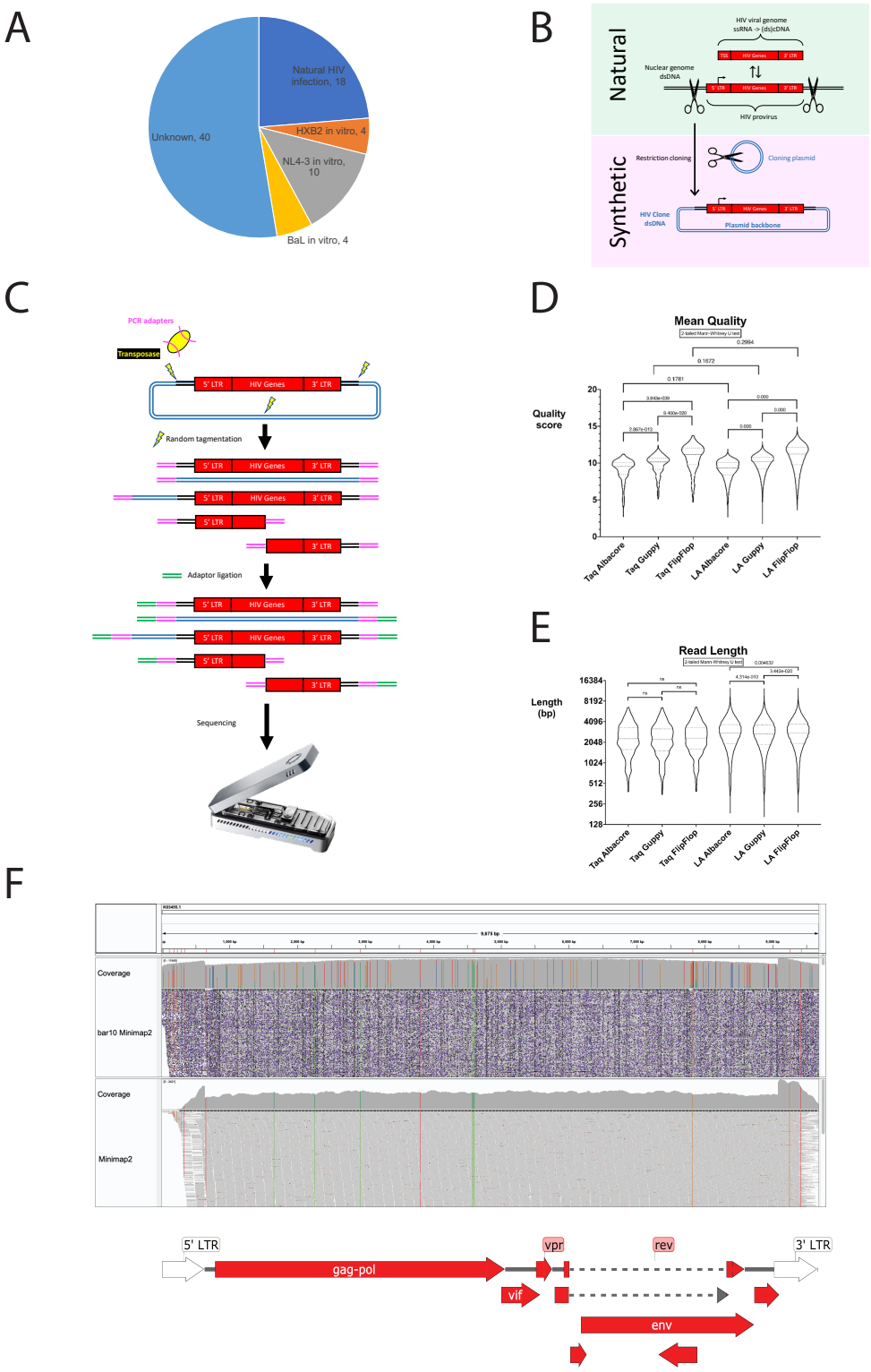
608 **Figures**

609

610

611 **Figure 1: HIV information in pHXB2_D is recovered by long-read sequencing and**

612 **mapping.**



613

614 Figure 1A: HXB2 is still a commonly used resource. It is the reference HIV-1 genome, derived
615 from one of the earliest clinical isolates. While older HIV samples are occasionally rediscovered,
616 they are not made routinely available to researchers. All public HIV-1 RNA-seq datasets were
617 obtained from the NCBI SRA with the following search phrase: "HIV-1" AND "RNA-seq".
618 Metadata from these 2527 runs (number current as of 7/21/2020) were used to make a pie chart
619 summary.
620 Figure 1B: HIV information comes from three main sources: proviruses (HIV sandwiched
621 between two assumedly identical full-length long terminal repeats (LTRs)), unspliced HIV
622 mRNAs (also known as viral genomes) starting from the transcription start site and ending in the
623 3' LTR [4], and engineered proviruses recovered in their entirety or stitched together from
624 multiple isolates like NL4-3 [18].
625 Figure 1C: ONT library prep pipeline. Tagmentation cleaves double-stranded DNA, ligating
626 barcoded PCR adapters (magenta). PCR-adapted DNA may be amplified. After amplification
627 and cleanup, ONT sequencing adapters (green) are ligated. Barcoded samples may be pooled and
628 sequenced.
629 Figure 1D: Newer basecallers increase read mean quality. Median (big dash) and quartiles (little
630 dash). Effect of enzyme version was not statistically significant.
631 Figure 1E: Read stats with different callers/aligners. Median (big dash) and quartiles (little dash).
632 Read lengths increase with higher fidelity Taq.
633 Figure 1F: Sequencing coverage with long- vs. short-read single-end 150 bp (trimmed to 142 bp)
634 DNA sequencing. Long-read sequencing covers ambiguously mappable areas missed by short-
635 read in HXB2 reference Genbank:K03455.1 (**Supplemental Figures 3D,3E**), but at the expense
636 of accuracy near homopolymers longer than about 4 nucleobases (**Supplemental Figure 5**).

Short-read mapping fails at repetitive elements longer than their read lengths (**Supplemental Figures 3D,3E**). 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).

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


```
@6fbf0205-5195-460e-8e28-930db50e5d79 runid=0b284792282af9a6d7275cfca845556b4b8ac7f sampleid=Student_Genomics_Run_1 read=87223 ch=460 start_time=2018-05-09T20:57:49Z barcode=barcode10
TGTTGTACTCTGGTTGACCTCGTTGTCGTTGTCGTTGTGTGTGTGTTTGCCGCTGCTGCGCGCGCGCAGGAGGAGGTGGGGGAGGATCATCCTGCTGCTAATCCTACATATTTGTGATGGAGGGGGTGGAGATTTTTGTGAGAAAAAAACAGAAAATATGAGAGTTTATAAGAA
CTAACACTTCTTCACATAACAAAACGAAATGTTCGAACTACGTAACTAAATGATGAAAAATATTCCCGATATCACTGGCTGTTTGGATGTGGCTATCAGGGTTTATTTTCCCCTTTCCTTCTTCTTGTTTGCTCTTTCGCTCTGACTCTGGCTCACCGCTGGCCCCTTCATTT
ACATTTGATAACTGTAGAGAGATTATTATAATCCTGCTCATTAGACAGATCAATCTGAAGTTGGCAAGTTTTTAAAATATAACTACCTATTTTTAAAAAGGGATGCCTTTAGTTTAGTTAACAATATATACTGCACATTTTGGTTTAAAAGGCCTGTTTACTACCACTGATTAACTATATGCCCACTGAGGCAACTCCTT
CTTTTGTTTTTATTCAAATATTTACTAATTACCAGGACTCCTGTGTGCTAATACAATGGTGCTCTACTTTCCACCTATATACTAGGGGAGACCAAGCACTATCACCCATACCTCTGAGCTTCCTAACTGGGTTACCTGTAGTTAACTAAGTAACTAAGTAACTCAAGCTAGCCAAAATCATCCCCACCCCATACCTA
TTACCACTGCCAATTACCTGTGGTTTCTTTACTCTAAACCTGTGATTCCTCTGAATTATTTCATTTTAAAGAAATTGTATTTGTTAAATATGTACTACAAACTTAGTAGTTGGAAAGACTAATTCACTCCCAAAGAAGACAAGATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTAGCAGAA
CTACACACCAGGGCCAGAGATTCGGATTACCACTGACCCCTTTCGATGGTGCTACAGAGCTAGTACCAGTTGGACCTAGATAAGTGTAAAGAGGCCAATAAAGGAGGGGAGCACCAGCTTGTTACACCCTGTGAGCCTGCATGGGATGATGTCTGATAGTGTGGTTGATATCTGCAGCCGCCTAGCAT
TTCATCGTGGCCGAGCTGCATCCGGGTACTTCAAGAACTGCTGATATGCAGAGTTCCTGCTGGGGACTTTCCGCTGGGGCGGCTGGGCGGGCTGGGCGGGCTGGGCGGCTGGGCGGCCTCAGATCCTGCATATAAGCAGCTGCTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGG
CTCTCTGGCTAACTAGGGGGGGCACTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGGTGTGTACTCTGGTAACTAGAGATCCCTCAGACCCTTTGGTCAGTGTGGAAAATCTCTAGCAGTGGCGCCCGAACAGGGACTTGGCGAAAGGGAAACCAGAGGAGCTCTCGGCGCAGGAC
TCGGCTTGCTGAAGCGCGCACGGCAAGAGCGAGGGGCGGCGACTGTTGAAGTACGCCAAAAATTTTGACTAGCGGAAGGCTAGAAAAGGAGAGATAGGTGCGGAAAGCGTCAGTATTAAGCGGGGAAAATTAGATCGATGGGAAAAATTAATAAAACATATGGTATGGGCA
AGCAGGGAGCTAGAACGATTCGCAATTAATCCCTGGCCTGTTAGAAACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCATCCCTTCAGACAGGATCAGAAAGAACTTAGATCATTATATATAATACAGTAGCAACCCTCTATTGTGTGCATCAAAGGATGAGATAAAAGACACCAAGGAAGCTTTAGACA
AGATAGAGGAGAGAGCAAAACAAAAAGTAAGAAAAAAGCACAGCAAGCAGCAGCTGACACAGGACACAACAATCAGGTCAGCCAAAATTACCTATAATTGCAGAACATCCAGGGCAAATGGTACATCAGGCCATATCACCTAGAACTTTAAATGCATGGGTAAAAGTAGTAGAAGAGGCTTTCAGCCCAGGAGTG
ATACCCATGTTTTCAGCATTATCAGAAGGAGCCACCACACACTTCTTCTACGCCCATCGCTCGCCGGCCGGGCAACGGTGCGCCCTCGCCGCCGCCGTGCGCGCCGATGACTACGAGGGCGTGGTGTGGTTGTCTGTTGTGTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGTTACTT
ACATAGCAGGAACTACTAGTACCCTTCAGGAACAAATAGGATGATGACAAAATAATCCACCTATCCCAGTAGGAGAAATTTATAAAAGATGGATAATCCTGGGATTAAATAAAATAGTAAGAATGTATAGCCCTACCAGCATTCGCCGGTGACGTACAGGAAACAGCTATGACCATGATTACGCCAAGCGCGCAGA
ATAAAACTCTAAGAGCCGAGCAAGCTTCACAGGAGGTAAAAATTGGATGACAGAAACCTTGTTGGTCCAAATGCGAACCCAGATTGTAAGACTATTTTTAAAAGCATTGGGACCAGGCCTGGTCCAATGAGTTTGTTAATACCCTCCTTTTAGTGAAATTGGGACAGCGTGCGAATGATGATGCTGATTATGATGGGGCAG
GCCAAGTAATACAAATTTACCATAATGATGCAGAGAGCAATTTTAGGAACCAAAGAAAGATTGTTAAGAGGTGTTTCAATTGTGGCAAGAAGGGCACACAACCAGAAATTGCAGGGCCCCTAGGAAAAAGGGCTGTTGGAAATGTGGAAAGGAAGGACACCAAATGAAAGATGTGTACTGAGAGACAGGCTAATTTTTA
GGGAAGATCTGGCCTTCCTACAAGGAGGCCAGGGATTTATCCAGAGCAGACCAGAGCCAACAGCCCCACCAGAAGAGAGCTTCAGGTCTGGGGTACAGTACCAACTCCTCAGGAAGCGGGAGACAAGGACTGTATCCTTAACTTCCTCGGGTACTCTTTGGGTACTCTTTGGGTACTCTTTGGGCGACCCCTCGTCACAATA
TAGGGGGCAACTAAAGGAAGCTCTATTAGATACAGGAGCAGATGATATTGAAAAAATTACACCAAGCTTCTCTGCCTATCACCTTCCCCCAAAAAGAAAGGGGGGACTGGAAGGGCTAATTCACTCCCAAAGAAGACAAGATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTAGCAGAAC
TACAATCGACATTAATTGGAAGAAATCTGTTGACTCAGATTGGTTGCACTTTAAATTTTCCCATTAGCCCTATTGAGACTGTACAGTAAAATTAAAGCCAGGAATGGATGGCCCAAAAGAAAGTTAAACAATGGCCATTGACAGAAGAAAAACAAAAGCATTAGTAAAAATTTGTACAGAGGTGGAAAAGGAAGGGAAA
ATTTCAAAAATTGGGCCTAAAATCACTTCTTACTTAGTATTTTACATAAAGAAAAAGACAGTACTAAATGGAGAAAATAAGTAGATTTCAGAGAACTTAAGAGTCCTGGGAAGTTCAGTGTGAATACCACATCCTTGGGGATTAAAAGAAAAATCAGTACAGTACTGGATGTGGGTGATATA
TTTTTCCAGTTCCCTTAGATGAAGACTTCAGGAAATATACTGCATTTACCATACCTAGTATAAACAATGAGACACCAGGGATTAGATATCAGTACAATGTTAACTACACAGGGATGGAAAGGATCACCAGCAATATTCCAAAGTAGCATGACAAAAATCTTGAAACCTTTTAGAAAACAAAATCAGACATAGTTATCTA
TCAATACATGGATGATTGTATGTAGGATCTGACTTAGAAATAGGGCAGCATAGACAAAAATAGAGGAGCTGAGACAACATCTGTTGGGGGTGGGGCTTACCACACCAGACAAAAACATCAGAAAGAACGCACTCCTTCCTTTGGATGGGTTATGAACTCCGTACGATAAATGGACGAGGTACAGCCTATG
GTGCTGTGCGCGAGAAGAAAGACAGCTGGCTGTGTGCAGAAAAAGAAGAGTGGAAGAGCTGTTTATCCATTTTCAAATTTTCAATTTTCTGCACATGCAGAAAGATAGATGGGGATGGGGCACCACTAAACAGTGTGTTAGATCTGCTGAAGCTTACGTACGCGCAGGAACACTGACGCTCTCGCACCCATCTCT
TAAAGAACCAGTACATGGGTGTATTATGACCCATCAAAAGACTTAATAGCAGAAATACAGAAGCAGGGGCAAGGCCAAAATGGACACATATCAAATTTATCAAGAGCCATTTAAAAATCTGAAACAGGAATTCAGAGAATGAGGGTGCCCACACTAATGATGTAAACCAATTAACAGAGGCAGTGCAAAAGTAGCCAC
AGAAGCATGGTAATTGGAAGACTCCTAAATTTAAACTGCCCATACAAAAGGAAACATGGGAACATGGTGGACAGGTTGGTAAGCCTCAGATTGAGTTTGTTAATACCCTCCTTTTAGTGAAATTGGGACAGCGTGCGAATGATGATGCTGATTATGATGGGGCAG
CCAACAGATAAATTAAGGAAAAGCAGGATATGTTACTAATAGAGGAAGACAAAAAGTTGTCCACCTAACTGACACAACAAATCAGAAGACTGAGTTACAAGCAATTTTATCTAGCTTTGCAGGATTCGGGATTAAGTAAACATAGTAAAAGCAGACTCCACATATGCGTAGATGGAGACATTCAAGCACAACCAGATCAAAGTG
AATCAGGGTTAGTCAATCAAATAATAGAGCAGTTAATAAAAATAGGAAAGGTCTATCTGGCATGGGTACCAGCACACAAAAGGAATTGGGGGAATGAACAAGTAGATAAATTAGTCAGTGCTGGAATCAGGAAGTACTATTTTTAGATGGAATAGATAGACCCCAAGGTGAACATGAGAAATGACAGCAGTAATTGGAGA
GCAATAACAGTGAGTTACAACCTGCCACCTGTGGTAGCAAAGAAATAGTAGCCAGCTGTGATAAATGTCAGCTAAAAGAAAGGAAGAAATACATCATGGACAAGTAGACTGTAGTCAGGAATATGGCAACTAGATTGTACACATTTAAGGAAGGAAAAGTTATCCTGGTAGCAGTTCATGTGGCAGTGGATCGTATAGAAG
CAGAAGCTTATTCCAGCAGAACAGGCAGGAAGTATGCAGCAATATTTCTTGGATTTTAAATTACGACGAAGATGGCCATGTCAACAGTTCACCAGTGCTACGGTTAAGGCCCGCTGTTGGTGGAGGCGGGAATCAAGCAGAATTCCCTCACAATCCCCAAAGTCAAGGAGTATGTAGAATCTATG
AATAAAGGATTAAAAGAAATTATAGGACAGGTAAGAGATCGAGGCTGAACATCTTAAGACAGCAGTACAAAATGGGGGATTGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCACAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAATCAATTA
ATTTTGGGTTTATTACAGGGACAGAATCCACTTTGGAAAGGACCCGTAAGCTCCTCTGGAAGGTGAAGGGCAGTAGTAATACAAGATAATAGTGACATAAAATAAGTAGTGCCAGAGAAAAGCAAAAGCCTATTGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTTGGCAAGTAGACAGGATGAGGATTAGAACATG
GAAAAGTTTAGTAAAACACCATATGTATGTCCAGGGAGAAGCTAGGGGATGGTTTTATAGACATCACTATGAAAGCCCTCATCCAGGAATAAGTTCAGAAGTACACATCCCACTAGGGGGGATCTGCATTTAAATAACAACATTTGGGGTCTGCATACAGGAGAAAAGAGACTGACACATTTGGGTCCAGGGGTCTCC
ATAGAATGGAGGAAAAGAGATTATAGCACACAAGTAGACCCTGAACTTAGCTAGCAGACCAACTAATTCATCTGTATTGCTGACTCTGTTGAACTTCGCTCATTAAGAAAGGCCTTATTAGGACACATAGTTTAGCCCCTAGGTCTGAATATCAAGCTGGGACATAACAAGGTAGGATCTCTACAATACTTTGCCATAGCAGCAATTA
TAACACCAAAAGATAAAGCCACCTTCCCTACTGGTTACAGAAACTTCCTGGAGGATGAGTGGAACAAGCCCCAAGAAGACAAGAGACCTGAGGAGCCACACAATGATGGACACTAGAGCTTTAGAGCTTTAAGAGCTTATGAGACATTTCTAGGATTTGGCTCCATGGCAAGACAACATATCTATAAACT
TATGGGGATACTTGGGCACAGGAGTGGAAGCCATAATAGAATCTGCAACAACTGCTGTTTATCCATTTTCAAATTTTCAATTTTCGCACATGAGCAGGACCAGCAAGAATGGACCAGTAGATCCTGAGCTAGACCCGTGGAAGCATCAGGAAGTCAGCTTAAACTGTTGTACCAATTGCT
ATTGTAAAAGTGTTGCTTTCATTGCCCAAGTTTCATAACAAAAGCCTTAGGCATCTCCTATGGCAGGAAGAAGCGGGAGACAGCGACGAAGAGCTCATCAGAACAGTCGACTCATCAAGCTTCTCTATCAAAGCAGTAAGTAGTACATGTAACGCAACCTATACCAATAGCGCAATAGTAGCATTAGTAAGTAACAATA
ATAATAGACGAACTAATTGTGTCTGTCATGAAAAGTAATCATAGAATATAAGGAAGGAATTATAAGGAAGCAGAAAATACAGAGGTTAAATTGATAGACTAAGGGCAGAAGACAGTGGCAATGAGAGTGAAGGAGAAATATCAGCACTTTGGTGCTTGGGATGTTGATGATC
TGTAGTGCTTTTACTTAGAAAAATTGTGGATACAGTCTATTATGGGGTACCTGTGTGTGGAAGGAAGCAACCTACTCTATTTTGTGCATCAGATGCTAAAGCATATGATACAGAGGTACACATCAATCATTTTAGGAGCCACACATGCCTGTGTGACCCTGGGAACCCCAACCCACAAGAAGTAGTATTGGTAAATGTGACAGA
AAATTTTAACATGTGGAAAAATGACATGGTAGACACAGATGCATTGAGGATATATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAATTAACTTACTCTGTGTTAGTTAAGTGCACATCATTTGAAGAATGATACTATACCAATAGTAGTAGCGGAGAATGATACAAAACTGTCTCTTTC
AATATCAGCACAAGCAAAAAATTGGAGTGCAGAAAGAATATGCATTTTTAAAACTTGATATAAATACCAATAGATAATGATACTACCAGCTATAAGTTGCAACAGGGCTGTCCTCAAGGTATCCTTTGAGCCAATTCCATACATTATTGTGCCCGGCTGGTTTGCGATTCTAA
AATGTAATAATAAGGCACGTTCAATGGAACAGGACCATGTACAAATAAATCAGCACAGTACAATGTACACACATGGAATTAGCCAGTAGATCAACTCAACTGCTGTTAAATGGCAGTCTACGCAGAGAAGAGGGTAGTAATTAGATCTGTCAATTTCACGGACAATGCTAAAACCATAATAACAGCTGAACACATCT
GTAGAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAAGAATCCGTATCCACAGAGGGACCAGGAGGCATTTGTTACAATAGGAAAAATAGGAAGTATGAACATTAATAATAGTGAGCAAATGGAATAACACTTAAACAAGATAGCTAGCAAATCTTGGAAGCCAACCAAA
ACAAATAATCTTTACTTAATCCTCAGGAGGGACCCGAAATTGTAACACTTACAGTTTTGGTGTGGAGGGAATTTTTTCTACTGTAATTCAACACAACTGTTTAATAGTACTGGTTTAATAGTACTTGGAGTACTGAAGGGTCAAATAACACTGAAGGAAGTGACACAATCACCCTCCCATGGAGATAAAACAAATTATA
AACATGTGGCGAGAGTAGGAAAAGCAATGTATGCCCTCCCTCACGTGGCAAATTATGTGTTCATCAAATATTACAGGGCTGTATTAACAAGGAGATGGTGGGTAATAGCAACAATGAGTCCGGAGATCTCCAGACCTGGGGGAGGAGATGTGAGGGAGTAGATATATAAATATAATGAGTAGTAAAATT
GAACCATTAGGAGTAGCACCCACCAAGGCAAAGAGAAGAGTGGGTGCGAGAGCGTCAGTATTAAGCGGGGGAGAATTAGATCGATGGGAAAAATTCGGTTACTGAACCAGACAACCATGGGCCACATCAATGCCCGCGCGGTCGAGGCCAGACGACAATATTCTGGTTCGGTGCAACACTCACAGCAAGCTGCTGAGCAGCAT
CTGTTGCAACTCACAGTCTGAAGATCAAGCAGCTCAGGCAAGAATCCTGGCTGTGGAGATACCTAAAGGATCAAATGCTGAGTTTGGGGTTGCTCTGGAGAAACTCATTGCACCACTGCTTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTCTGGAACAGATTTGATCACACGACCTGGATGGAGTGGGACCA
GAGAAAGTAACAATTACACAAGCTTAATACACTCCTTGGTGAAGAATCGCAAAACCAGCAAGAAAAAGAATGAACAAAAATTATTGGAATTAGATAAATGGGCAAGTTTCTTGTGGTCTTGTAGTAGTTTGGGTGCTGACTTCTTTTACCGTAGACCAGATCTGGTAGGTTTAAGAATAGTTTT
TGCTGTACTTTCTTATAGTGAATAGAGTTAGGCAGGGATATTCACCATTATCGTTTCAGACCCACCTCAACCCCGAGGGGACCCGACAGGCCCGAAGAATTAGAAGAAGGTGGAGAGAGAGACAGATCCATTCGATTAGTGAACGGATCCTTGGCACTTATCTGGGACGATCTGCGGAGCGCATCGTCTTCAGCTACCACC
GCTTGAGACACTCTTGATTGTAACGAGATTGTGGAACTTCTGGGACGCAGGGGGTGGAAGCCCTCAAATATTGGTGGAATCTCCTACAGTATTGGAGTCAGGAGCTAAAGAATAGTGCTGTTAGCTCATTCCACCATGACAGTACGCTGAGGGAGCGGGCGATGTAGGGTTATAGAAGTATTGTAGAAGTGTTGTAGAAGTG
CCATGAGCTATTCGCCACATACCTGAAAAAGGAGATATAAGCAGGGCTTGGAAGGATTTTTGGGGTGCAGTCAGGGGAAAATGGTCAAATGGGTCAAAAGTAGTGTGATTGGCGCCTACTGTGTAAGAGAAAGATGAGACAAGGCTGAGCCAGCAGCAGCAGATAGGGTGGGGAGCAGCATCTCGGAAAACATGGAG
CAATCACAAGTAGCAATACAGCAGCTACCAATGCTGCTTGTGCCTGGCTGAGAGCACAAGAGGAGGAGGAGGTGGGTTTTCAGTCACACCTCCGCCTTTAAGACCAATGACTTACAAGGCAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGACTGGGAAGGCTAATTCACTCCCAAAGAAGACAAGATATCCT
TGATCTGTGATCCTCACCACACAAGAGCTACTTCCCTGATTAGCAGACTACACACGGGGCCAGGGGTCGAATGTCCACTGACCTTTGGATGGTGCTACAAGCTAGTACCAGTTGGGAGGCGAGATAAAGGTAGAAGGAGAGAACACCAGCTTGTTACACCCTGTGAGCCTGCATGGGATGGATGACCTGCTAGCA
CCGGAGAGAAAGTGTTAGAGTGGAGGTTTGACAACCGCCTAGCATTTCATCCACGTGGCCGGGTGCATCCTCGGAGTACTTCAGAAACTGCTGATATCGGAGCTTGCTAGGGACTTTCCGCTGGGGGACTTTCCAGGGAGGCGTGGCCTGGGCGGGACTGGGGTGGCGAGCCCTCAGATCCTGTAAGCAGCTGCTTTGCCTG
TACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGGGTCTCTGGCTAACTAGGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGGTGTGTACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGCGCCCGAACAGGG
TTATTCAGTATTTATAACTTGCAAAGGAATGAATATCAGGGAGTGAAGAAGACCTTGACATTTGAATGAATTTGATCAGAGAATTGGAGAGAGCTAGACTAGACACAGACAGGGCAAAGCTGCAGAAATTTGGAGAGAGCCAGCAGAGATGCTGACATACCTTGGCTAATCCCAGATCCTAGCAGTCACAGTCTT
ATAATGATTTATAGTTTAAAGTATAAACTTATCTAATTTACTATTCTGACAGATATTAAATTAATCTCAAATATCATAAGAGATGTATCCCATTAACCTGAAACTGAGAGGGAAAGATGTTGAAGTAATTTTCCCACAATTACAGCATCCGGTTAGTTACGACTCTATGATCTTCTGACACAAATTCCATTTACTCCTC
ACCCTATGACTCAGTCGAATATATCAAAGTTATGGACATTATGCTAAGTAACAAATTATCCCTTTTATATAGTAAATACTGAGTAGATTGAGAGAAGGAAAAATTGTTTCGCAAACCTGAAATAACTTCAAGAAGAAGAGGTGAGGATAAAATTAACAAGTAGCTTGGTTAGAAAAGG
GATTCAAATGCATAAAGCAAGGGATAAATTTTCTGGCAACAGAACTATACAATATAACCTTAAATATGACTTCAAATAATTGTTGGAACTTGATAAAACTATTAAATATTATTGAAGATATCAATATTATAAATGTATTTACTTTTAAAGAATGTGTATCATTGAGAGTAGAAAACAATCCTTTATTATCACATTTTGTC
AAACAAGTTTGTTATTTTTAACACAAGTAAGAATACTGCATTCAATTAAGTTGACTATGCAGATTTTGTGTTTTTGTTAAATTTAAGAAAGAGATAACTGAAGCGGCGCACGAAAAACGGGAGAGGACAAGGTTTCAACGCTTGTCTGC
+
)''.+62124.89860,2)'%%#$,+))**((((#%%+)***$(*236<6221/,/2+/&&(16/&&:;>+6&:;98BCD282;:12>27))41?=9*,+,:890-3255:84;/*(O/3))&%((O$&003:69'+25+/-
2+)+/,/79/+,&2246778'.%%3AF==:4,01=7BB<?:;6108550()=:=8:555=:53%%*2+,0-6;@@14)
%%(?<<>6:;<=;B=56::3:9?<?>7:<612354,5#*19872:-<;9:;A4::>8667)'#.%196?@?<@:.;/;<@CF?BADA76:;:AC?B=<:;>$S.$%(.:9<<9>;<7(+-@>?H?98A5B<@@&9>:<6:59>8:=4,/-
/58:'&'(+,+4/+,,,((0<;>==79-@A872:53254400&245;=<6+1;;88:=:=857*$4+,,,11:9:8=:=:<<<:;>:;:;:;
```

646 The 5th longest read in the barcode 10 set (read ID 6fbf0205-5195-460e-8e28-930db50e5d79)
647 contained full-length HIV-1. Query (full read) blastn against HIV (taxid:11676) returned 92.95%
648 identity to HIV-1, complete genome (Genbank:AF033819.3). Limiting query to HXB2 (red)
649 blastn against Nucleotide collection nr/nt returned 100% coverage and 93.02% identity to HIV-1
650 HXB2. This read was 11,487 bases long, with mean quality score 11.984396. Basecalled using
651 Guppy 2.3.1 with FlipFlop config.
652

Figure 3A: 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 3B: pNL4-3_gag-pol(Δ1443-4553)_EGFP (ACCESSION_TBD) has distinct LTRs, consistent with pNL4-3 (AF324493.1)

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


AF324493.1_5LTR tggaagggctaatttggtcccaaaaaagacaagagatccttgatctgtggatctaccaca
ACCESSION_TBD_5 tggaagggctaatttggtcccaaaaaagacaagagatccttgatctgtggatctaccaca
AF324493.1_3LTR tggaagggctaattcactcccaaagaagacaagatatccttgatctgtggatctaccaca
ACCESSION_TBD_3 tggaagggctaattcactcccaaagaagacaagatatccttgatctgtggatctaccaca
                **************.. *******.********* *************************

AF324493.1_5LTR cacaaggctacttccctgattggcagaactacacaccagggccagggatcagatatccac
ACCESSION_TBD_5 cacaaggctacttccctgattggcagaactacacaccagggccagggatcagatatccac
AF324493.1_3LTR cacaaggctacttccctgattggcagaactacacaccagggccaggggtcagatatccac
ACCESSION_TBD_3 cacaaggctacttccctgattggcagaactacacaccagggccaggggtcagatatccac
                ************************************************.************

AF324493.1_5LTR tgacctttggatggtgcttcaagttagtaccagttgaaccagagcaagtagaagaggcca
ACCESSION_TBD_5 tgacctttggatggtgcttcaagttagtaccagttgaaccagagcaagtagaagaggcca
AF324493.1_3LTR tgacctttggatggtgctacaagctagtaccagttgagccagataaggtagaagaggcca
ACCESSION_TBD_3 tgacctttggatggtgctacaagctagtaccagttgagccagataaggtagaagaggcca
                ****************** ****.*************.*****  *.*************

AF324493.1_5LTR atgaaggagagaacaacagcttgttacaccctatgagccagcatgggatggaggacccgg
ACCESSION_TBD_5 atgaaggagagaacaacagcttgttacaccctatgagccagcatgggatggaggacccgg
AF324493.1_3LTR ataaaggagagaacaccagcttgttacaccctgtgagcctgcatggaatggatgaccctg
ACCESSION_TBD_3 ataaaggagagaacaccagcttgttacaccctgtgagcctgcatggaatggatgaccctg
                **.************ ****************.****** ******.***** ***** *

AF324493.1_5LTR agggagaagtattagtgtggaagtttgacagcctcctagcatttcgtcacatggcccgag
ACCESSION_TBD_5 agggagaagtattagtgtggaagtttgacagcctcctagcatttcgtcacatggcccgag
AF324493.1_3LTR agagagaagtgttagagtggaggtttgacagccgcctagcatttcatcacgtggcccgag
ACCESSION_TBD_3 agagagaagtgttagagtggaggtttgacagccgcctagcatttcatcacgtggcccgag
                **.*******.**** *****.*********** ***********.****.*********

AF324493.1_5LTR agctgcatccggagtactacaaagactgctgacatcgagctttctacaagggactttccg
ACCESSION_TBD_5 agctgcatccggagtactacaaagactgctgacatcgagctttctacaagggactttccg
AF324493.1_3LTR agctgcatccggagtacttcaagaactgctgacatcgagcttgctacaagggactttccg
ACCESSION_TBD_3 agctgcatccggagtacttcaagaactgctgacatcgagcttgctacaagggactttccg
                ****************** ***..****************** *****************
```

```
AF324493.1_5LTR ctggggactttccagggaggtgtggcctgggcgggactggggagtggcgagccctcagat
ACCESSION_TBD_5 ctggggactttccagggaggtgtggcctgggcgggactggggagtggcgagccctcagat
AF324493.1_3LTR ctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagat
ACCESSION_TBD_3 ctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagat
                ********************.***************************************

AF324493.1_5LTR gctacatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
ACCESSION_TBD_5 gctacatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
AF324493.1_3LTR gctgcatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
ACCESSION_TBD_3 gctgcatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctga
                ***.*******************************************************

AF324493.1_5LTR gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
ACCESSION_TBD_5 gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
AF324493.1_3LTR gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
ACCESSION_TBD_3 gcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgcct
                ************************************************************

AF324493.1_5LTR tgagtgctcaaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
ACCESSION_TBD_5 tgagtgctcaaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
AF324493.1_3LTR tgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
ACCESSION_TBD_3 tgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctc
                ********. ***********************************************

AF324493.1_5LTR agaccctttagtcagtgtggaaaatctctagca
ACCESSION_TBD_5 agaccctttagtcagtgtggaaaatctctagca
AF324493.1_3LTR agaccctttagtcagtgtggaaaatctctagca
ACCESSION_TBD_3 agaccctttagtcagtgtggaaaatctctagca
                **********************************
```

Figure 4A: HXB2 integration site

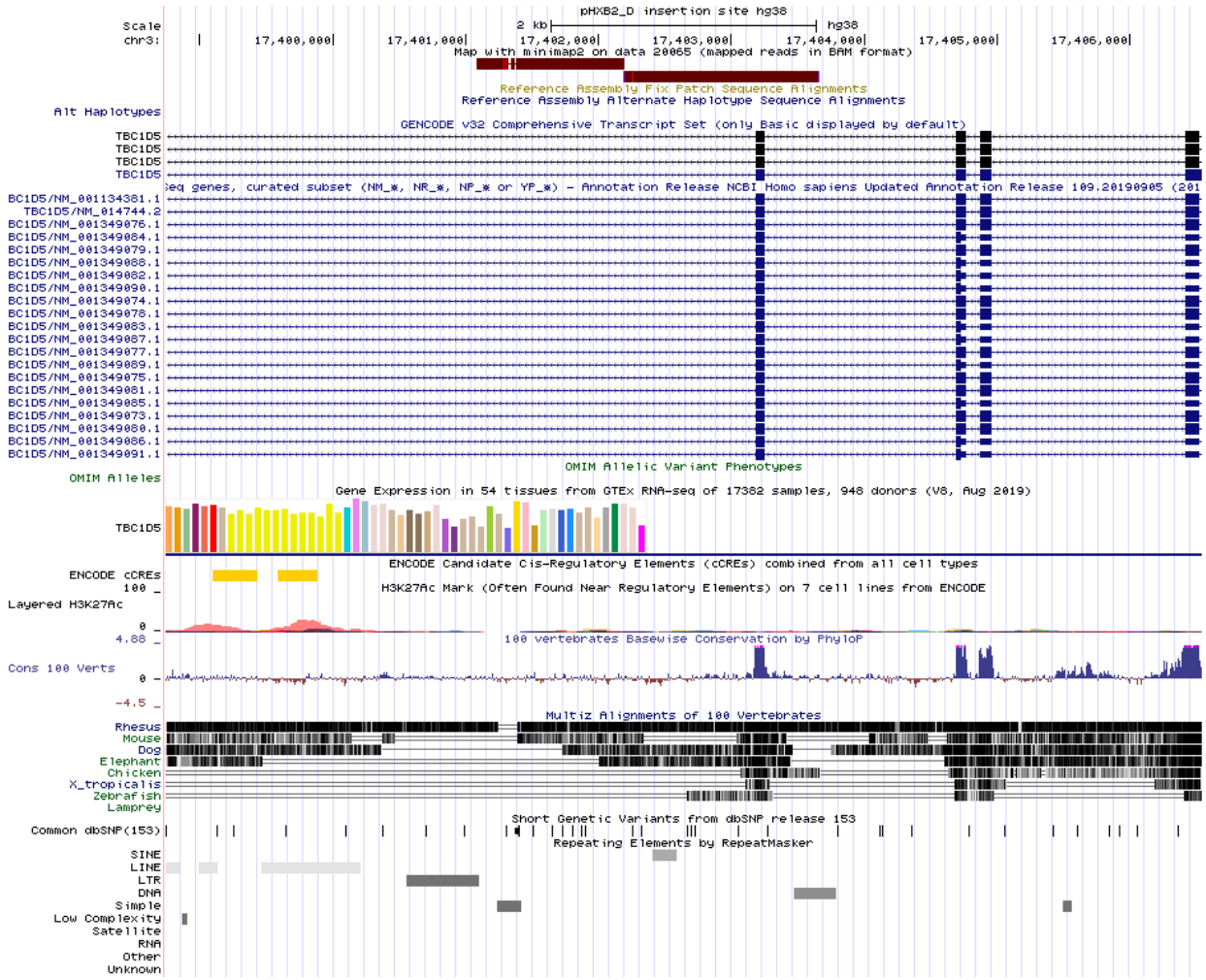
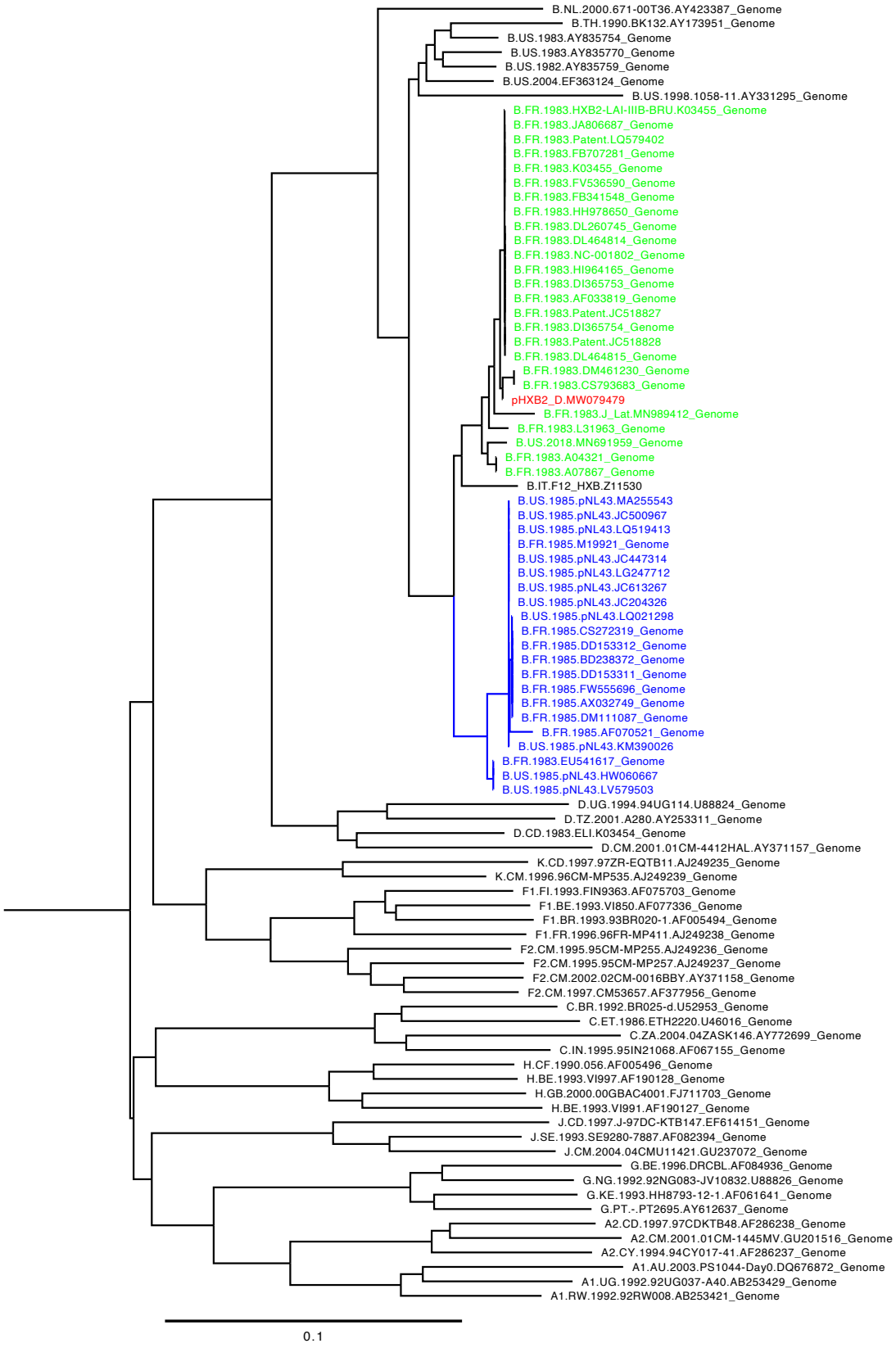


Figure 4B: NL4-3 integration sites

Figure 4A: 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 study are consistent with HIV-1 integration behavior [44]. Visualized in UCSC Genome

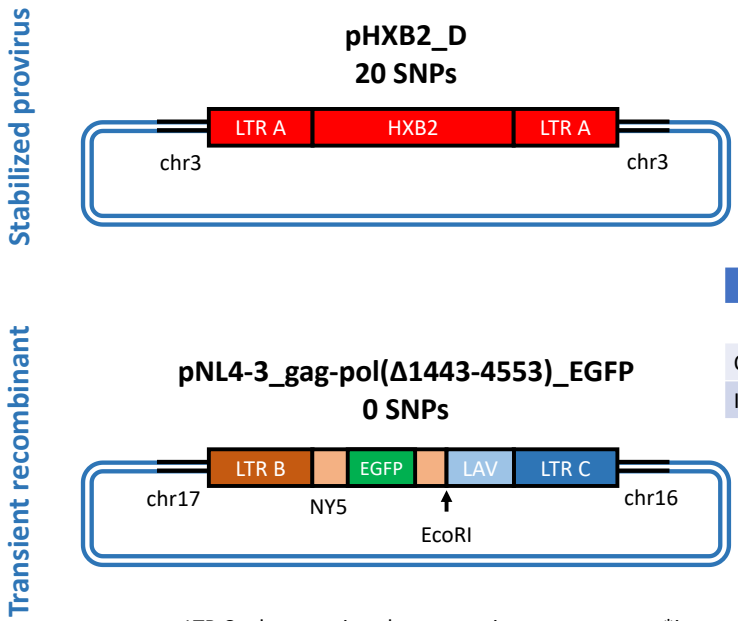
813 Browser [45]. Figure 4B: pNL4-3_gag-pol(Δ1443-4553)_EGFP's, and therefore NL4-3's,
814 integration sites fall on annotated repeats, the longer reads help to locate both sites. Alignment
815 quality is 60 for both homology arms (**Supplemental Table 2**). These integration sites would
816 likely be missed by any method leveraging reads shorter than the homology arms.
817

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



819

820 Figure 6: Summary of long- vs. short-read mapping by ability to phase LTRs



LTR phasing success		
	pHXB2_D	pNL4-3Δ
ONT PCR-barcode	Yes	Yes
ILMN SE 150	Yes*	Yes

LTR C takes over in subsequent viruses.

*intermittent success with ambiguous mapping at LTR.

821

822 **Supplemental Information**

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

824 To assemble pHXB2_D, we tried the following short read assemblers on short-read data
825 from the external core: IDBA [46], MIRA [47], [48], SPAdes [49], and SSAKE [50], [51]. These
826 were chosen as a convenience because they were already stably implemented in Galaxy
827 (specifically usegalaxy.eu). Of these, SSAKE produced discontinuous assemblies with default
828 parameters. The discontinuous contigs did however map to the core's assembly (not shown).

829 **Enabling STEM outreach**

830 This work was performed as two control experiments with identically prepared libraries
831 for a STEM outreach initiative, Student Genomics (Gener, et al., manuscript in prep). Given the
832 constraints of the Student Genomics pilot, a rapid sequencing kit with tagmentation (explained
833 below) with PCR barcoding was used to pool samples for ONT sequencing, with the
834 consequence of fragmenting plasmid DNA more than what would have been ideal for capturing
835 full-length HIV. That said, these controls could have been just as easily replaced by any
836 samples/experiments benefiting from long-read sequencing at moderate-to-high coverage.
837

838 **Supplemental Tables**

839

840

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

842 See **Supplemental Digital Content.**

843 See also **Figure 1A**.

844

845

846

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

848 See **Supplemental Digital Content.**

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

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

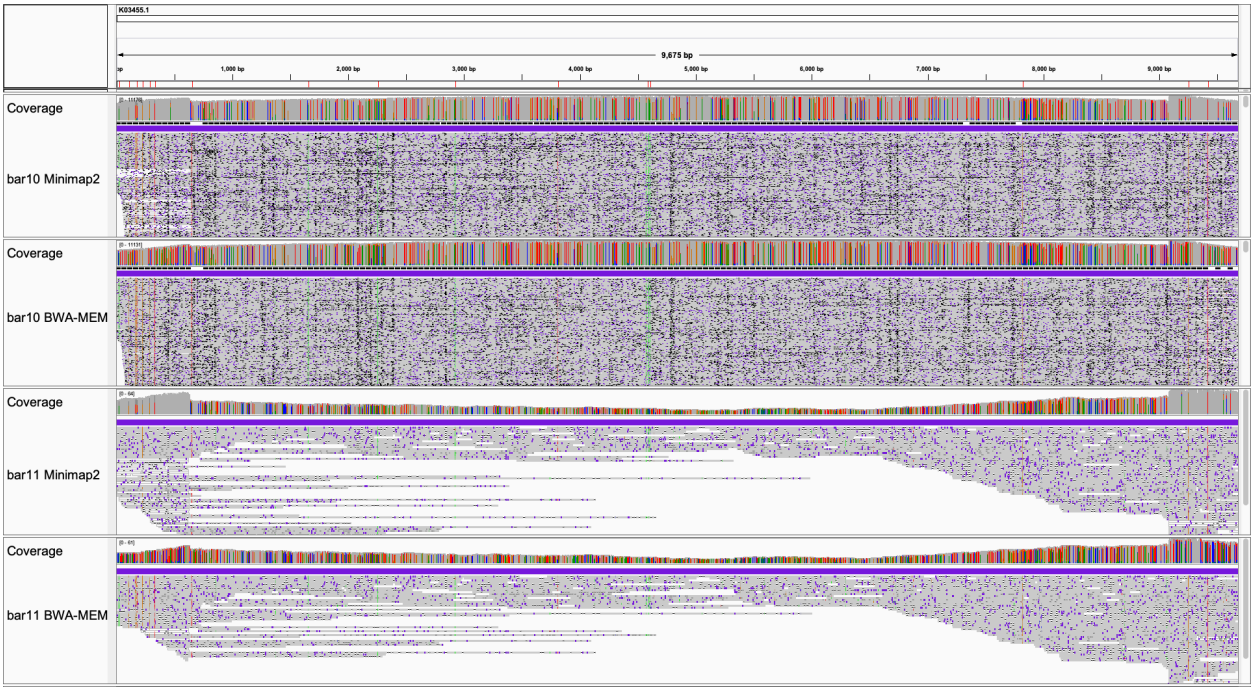
Assembled with Canu. NA denotes features which may not have matched exactly, but which were collapsed with adjacent features to facilitate counting. Variants called manually by mapping assemblies over HXB2 features with SnapGene.

863 **Supplemental Figures**

864

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

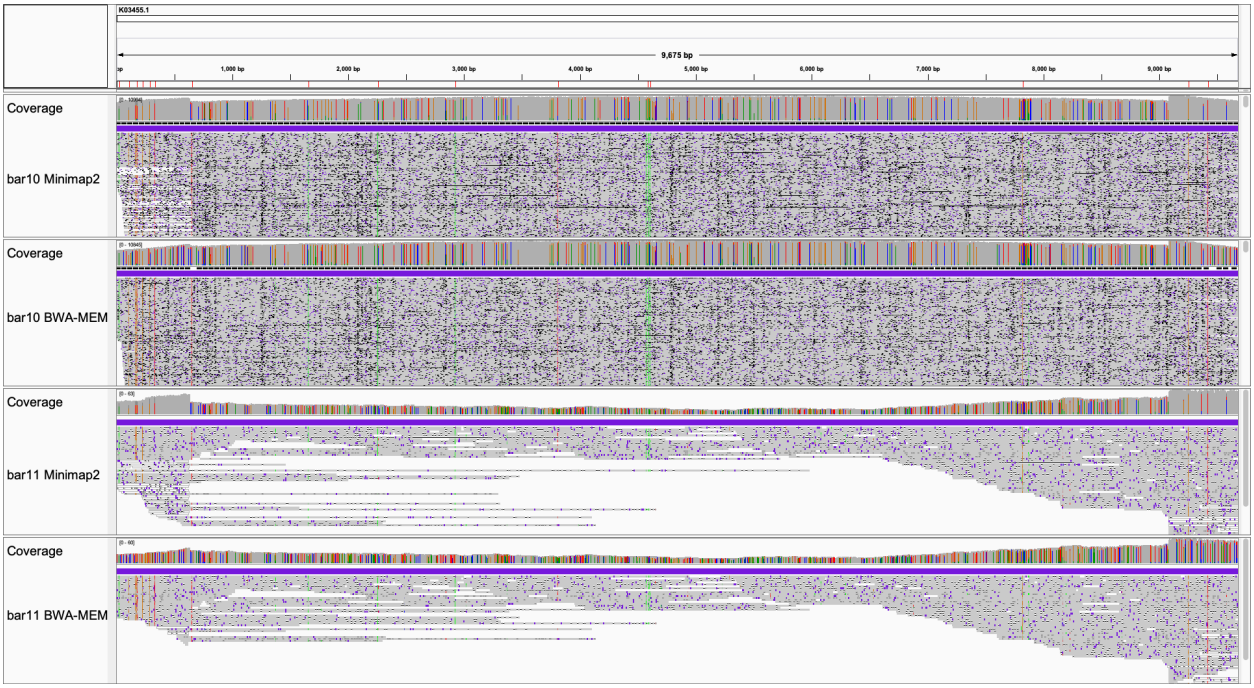
866 **depends on DNA polymerase and mapper. ONT basecaller = Albacore (worst).**



867
868

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

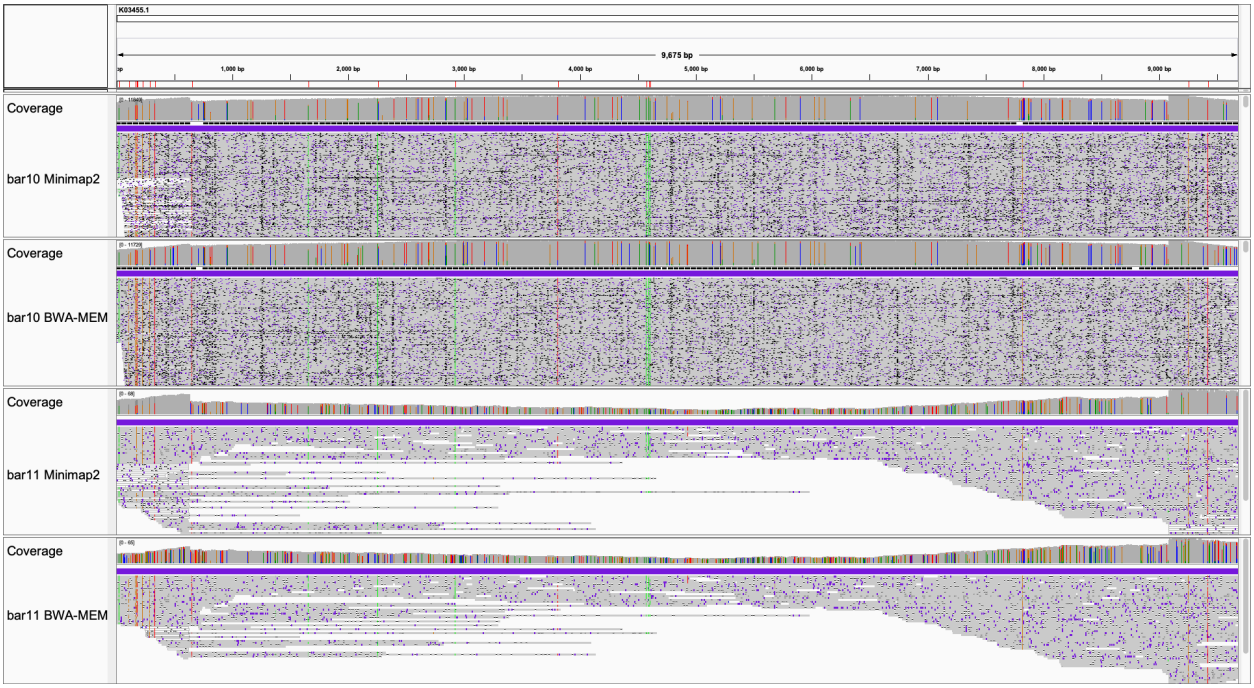
870 **depends on DNA polymerase and mapper. ONT basecaller = Guppy.**



871
872

873 **Supplemental Figure 1C: Unbiased nanopore DNA sequencing coverage over HXB2**

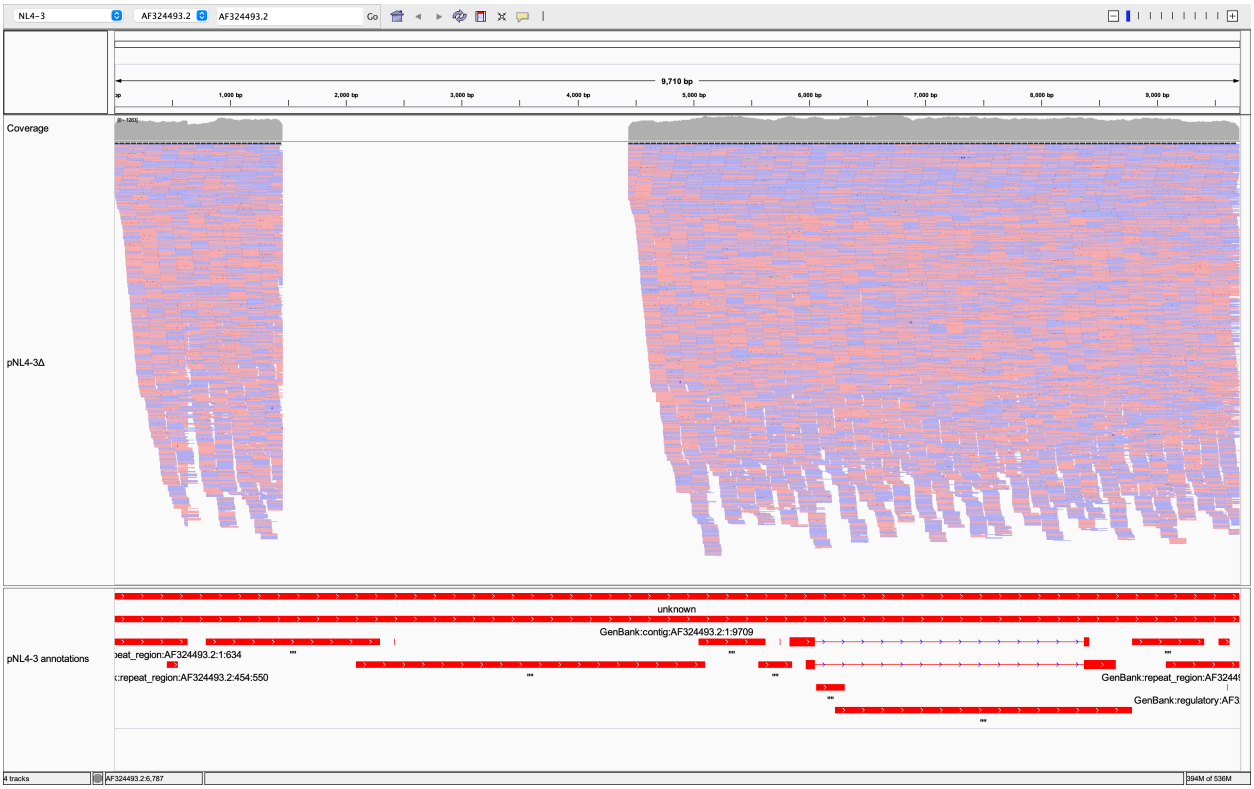
874 **depends on DNA polymerase and mapper. ONT basecaller = FlipFlop (best).**



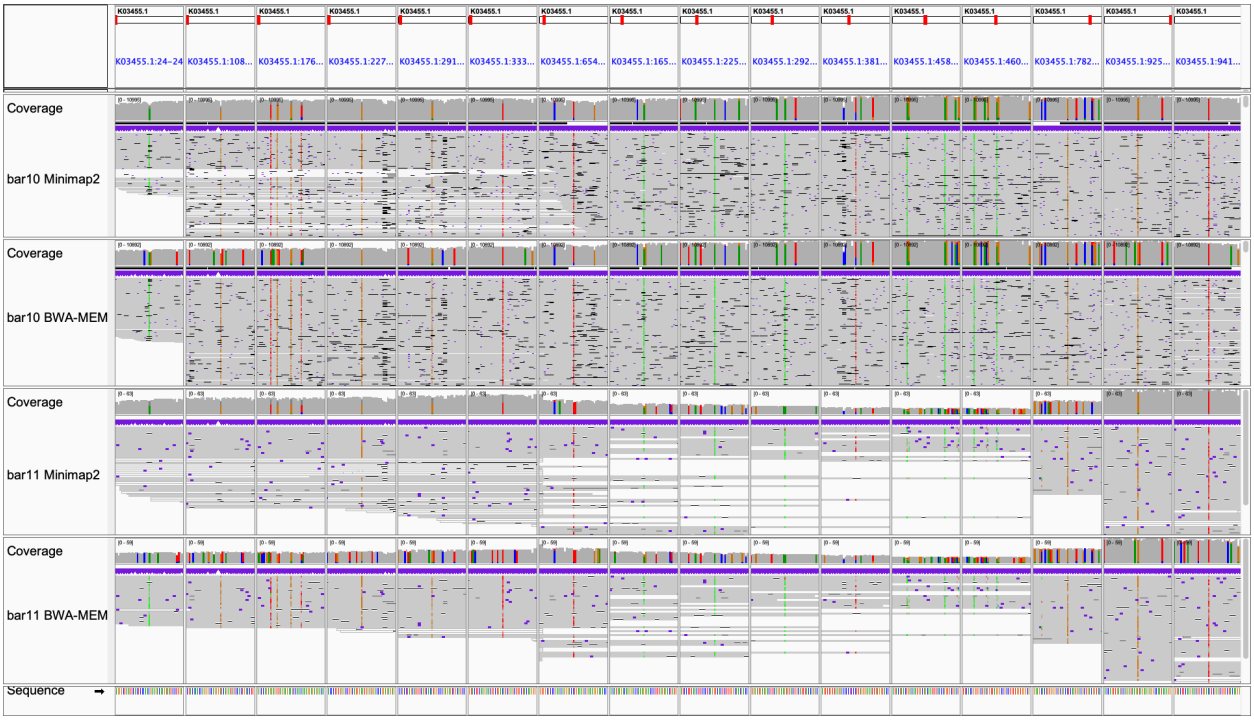
875
876

877 Top two Coverage and Alignment panels from barcoded library 10 (bar10 = LA Taq). Bottom
878 two from Barcode 11 (bar11 = Taq). Minimap2 and BWA-MEM were used to map reads
879 basecalled with Albacore (worst), Guppy, or FlipFlop (best) to HXB2. Color-coding: Red below
880 genome scale marks 20 SNVs across the HIV segment of pHXB2_D. Purple is an insertion in a
881 given read relative to reference. White is either a deletion in a given read or space between two
882 aligned reads. Gray in alignment field means base same as reference, and in coverage field
883 means major allele is at least 95% the same as reference. Per-read "insertions" and "deletions"
884 do not necessarily represent true insertions or deletions actually present in the sample, because
885 each read is likely an imperfect independent observation. Automated assembly followed by
886 manual consensus building converts these overlapping reads into approximations of the ground
887 truth. "Unbiased" refers to not amplifying a given region (e.g., pol) before ligating ONT
888 sequencing adapters. In the present approach, the tagmentation process randomly cuts DNA,
889 creating ~2000 bp pieces. Tagmented DNA is then amplified based on tagmentation adapters.
890

Supplemental Figure 2: Reads map well to HIV-1 NL4-3 segment of pNL4-3 assembly because NL4-3 LTRs are distinct.

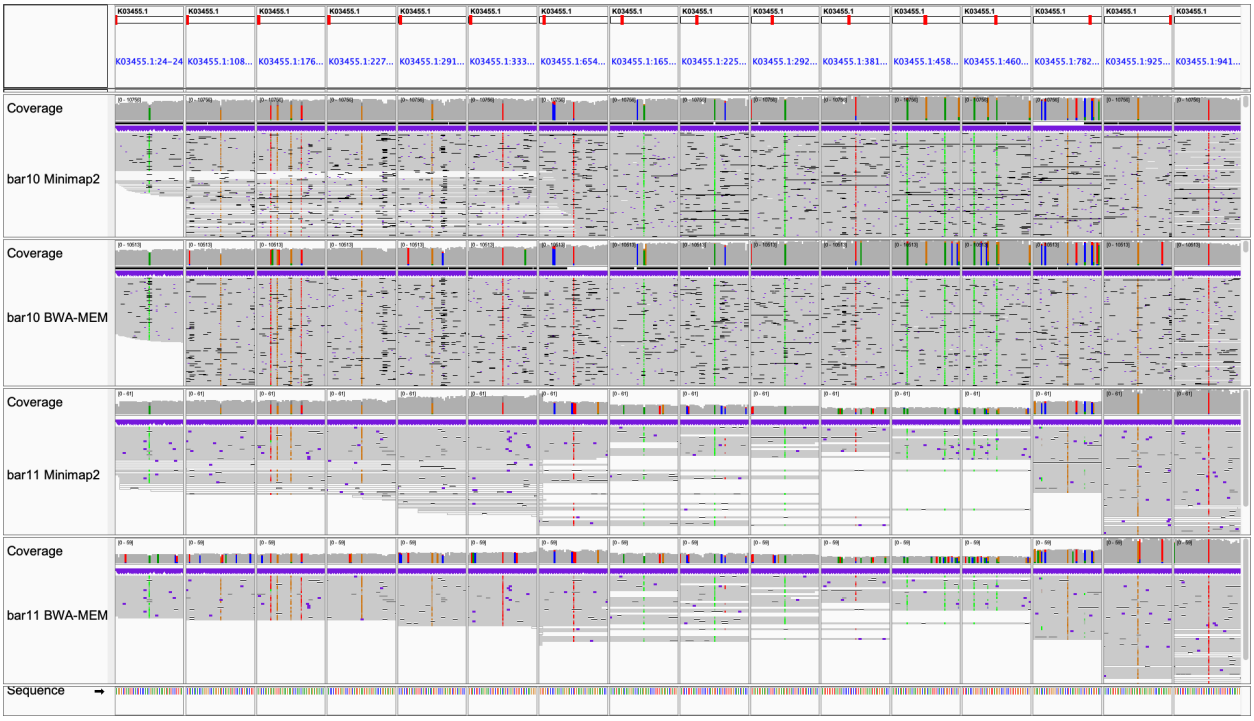


Supplemental Figure 3A: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller = Albacore (worst).



900 **Supplemental Figure 3B: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT**

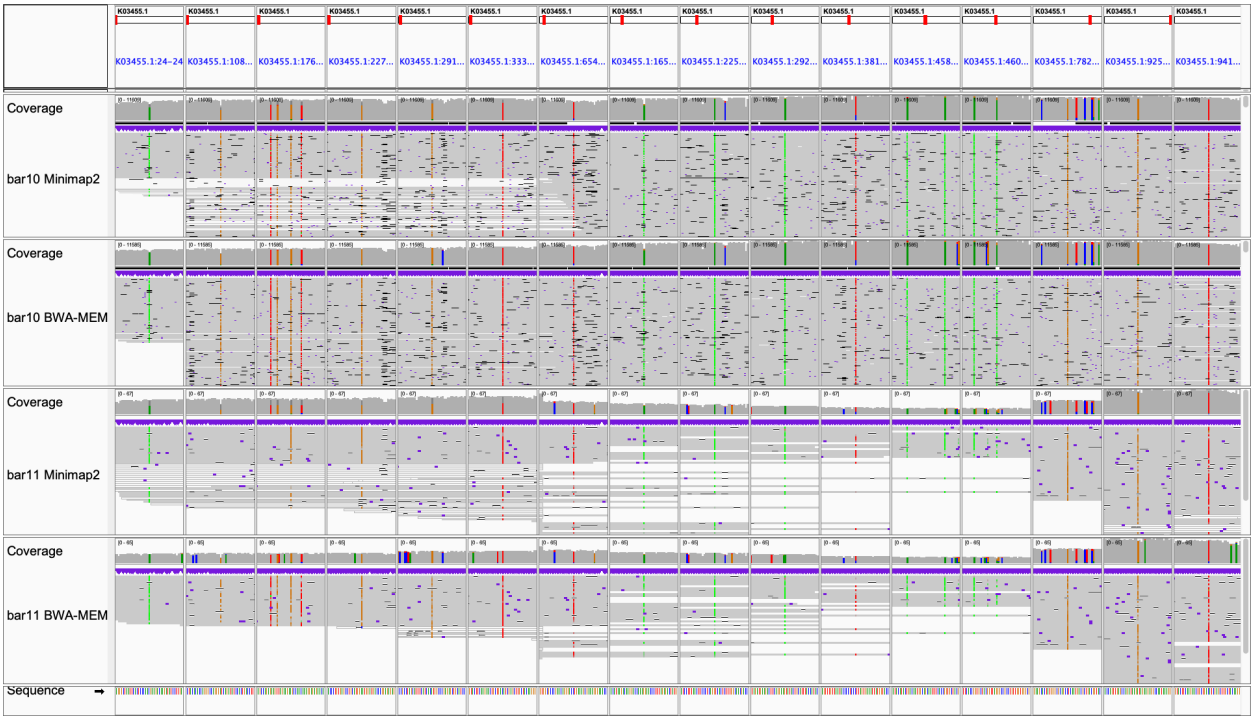
901 **basecaller = Guppy.**



902
903

904 **Supplemental Figure 3C: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT**

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



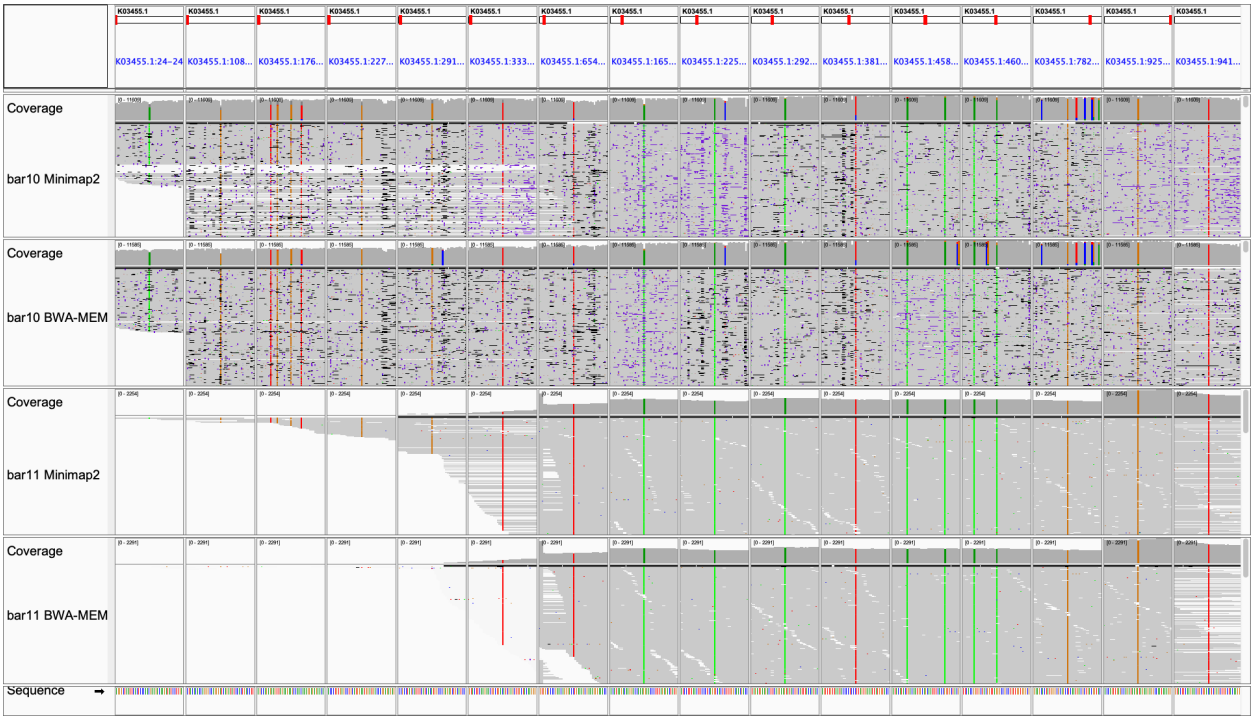
906
907

908 **Supplemental Figure 3D: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs.**

909 **short reads (HIV genome).**

910
911

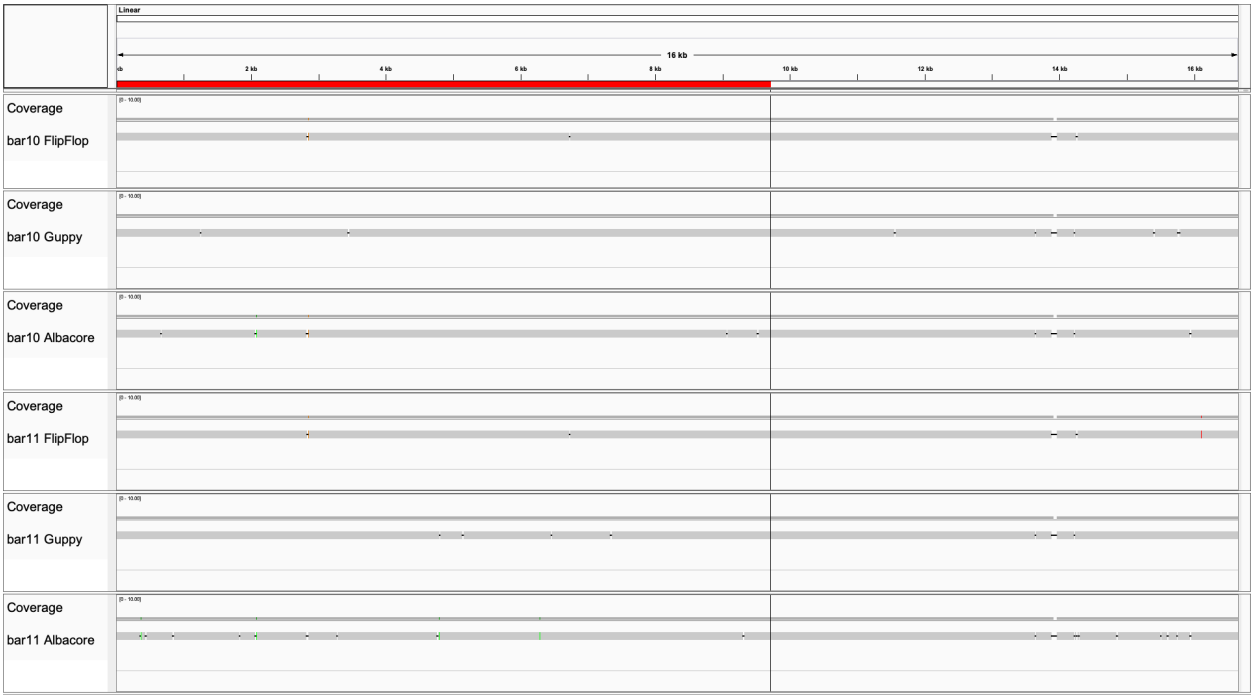
Supplemental Figure 3E: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs. short reads (20 SNV-focused).



916 Supplemental Figure 3A: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller
917 = Albacore (worst). Gray indicates per-base consensus accuracy ≥ 80%. These alignments are
918 the noisiest (less gray and most divergent from reference) between Supplemental Figures 3A,
919 3B, and 3C.
920 Supplemental Figure 3B: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller
921 = Guppy.
922 Supplemental Figure 3C: HIV single nucleotide variants (SNVs) in pHXB2_D. ONT basecaller
923 = FlipFlop (best). These alignments are the least noisy (most gray and like reference) between
924 Supplemental Figures 3A, 3B, and 3C.
925 Supplemental Figure 3D: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs. short
926 reads (HIV genome). Long reads outperform short reads at HIV-1 LTRs. ONT basecaller =
927 FlipFlop. Short read as single-end 150, clipped to 142, provided by external core. Mappers =
928 Minimap2 (better), BWA-MEM.
929 Supplemental Figure 3E: HIV single nucleotide variants (SNVs) in pHXB2_D, long vs. short
930 reads (SNV-focused).
931

932 **Supplemental Figure 4A: Assembling pHXB2_D from long reads only, varying basecaller**

933 **and polymerase.**



934

935 Each pane (n=6) summarizes the results of contig curation. Divergence from reference decreases

936 with newer basecallers, and with long amplicon DNA polymerase (Sigma-Aldrich Taq vs. LA

937 Taq by Takara). Errors in assembly occurred at homopolymers (most often deletions not visible

938 at this resolution; see **Supplemental Figure 6**), dimer or trimer runs. bar10 = LA Taq library.

939 bar11 = Taq library. pHXB2_D Genbank:MW079479. Best contigs presented, manually curated

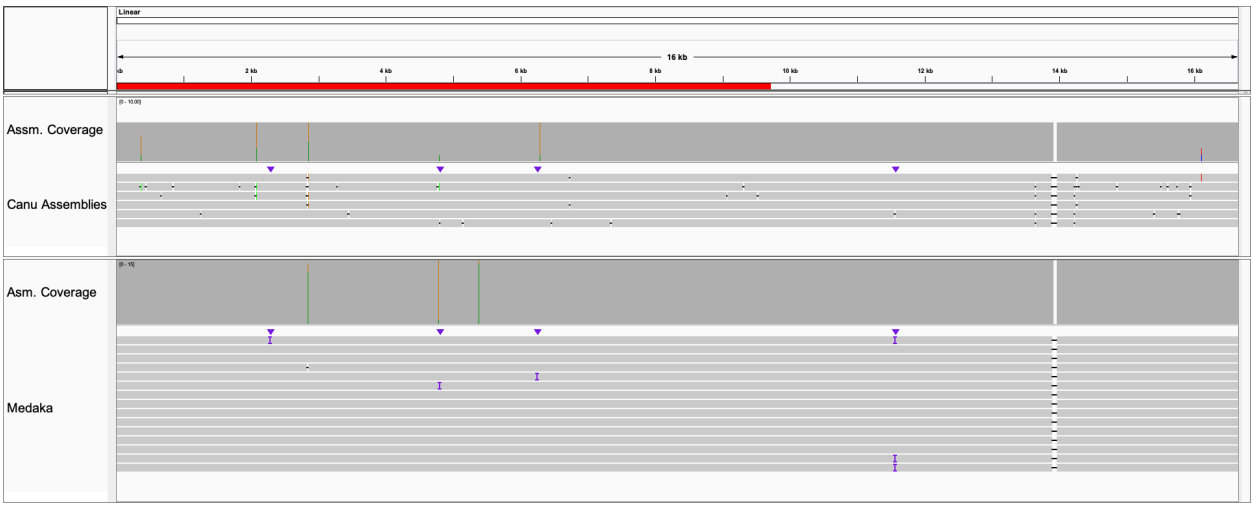
940 to match pHXB2_D coordinates. Note LTRs (beginning and terminal 634 bp of red bar) are

941 resolved in almost all assemblies. See **Supplemental Table 3** for differences between assemblies

942 and the reference (left red). Plasmid backbone (right) differences are not reported.

943

Supplemental Figure 4B: ONT errors corrected by polishing ONT-only assemblies.

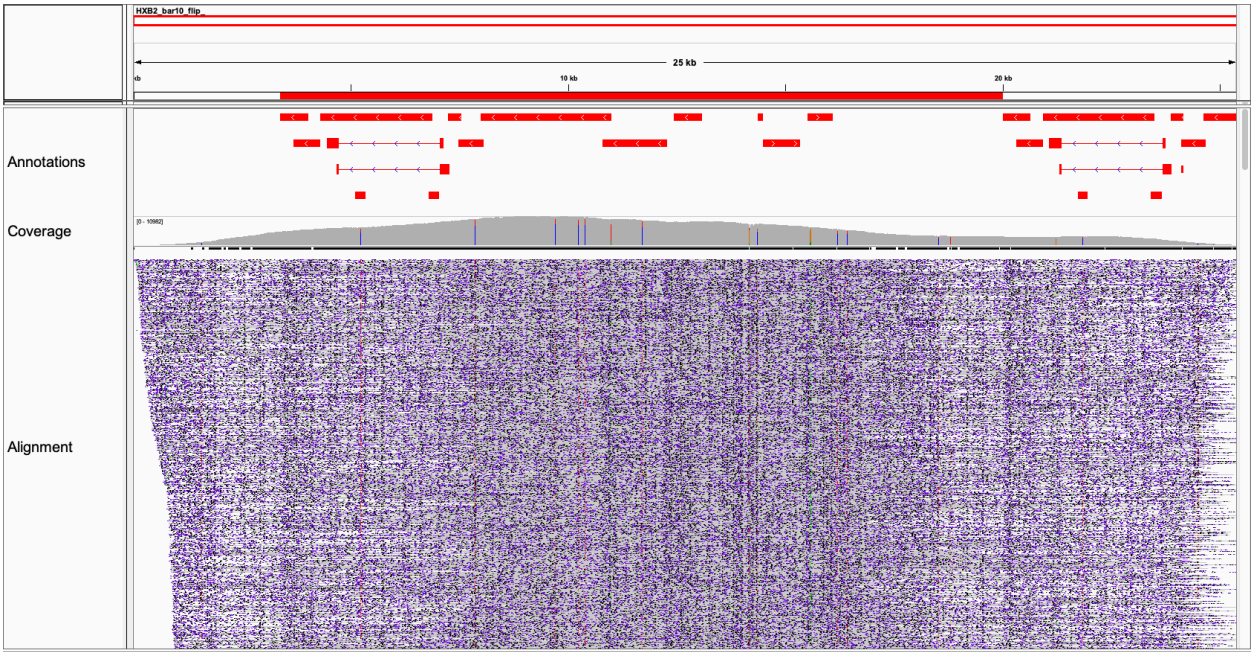


Assemblies polished with Medaka (ONT). Top: pHXB2_D genome. Bottom: HIV-only segment. The best polished assembly had one error in the entire plasmid (1 error out of 16,722 bases), with a corresponding consensus accuracy of 99.99402%. This happened to be in HIV segment (HIV-1 between position 1 and 9719; 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-

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

955 assemblies.

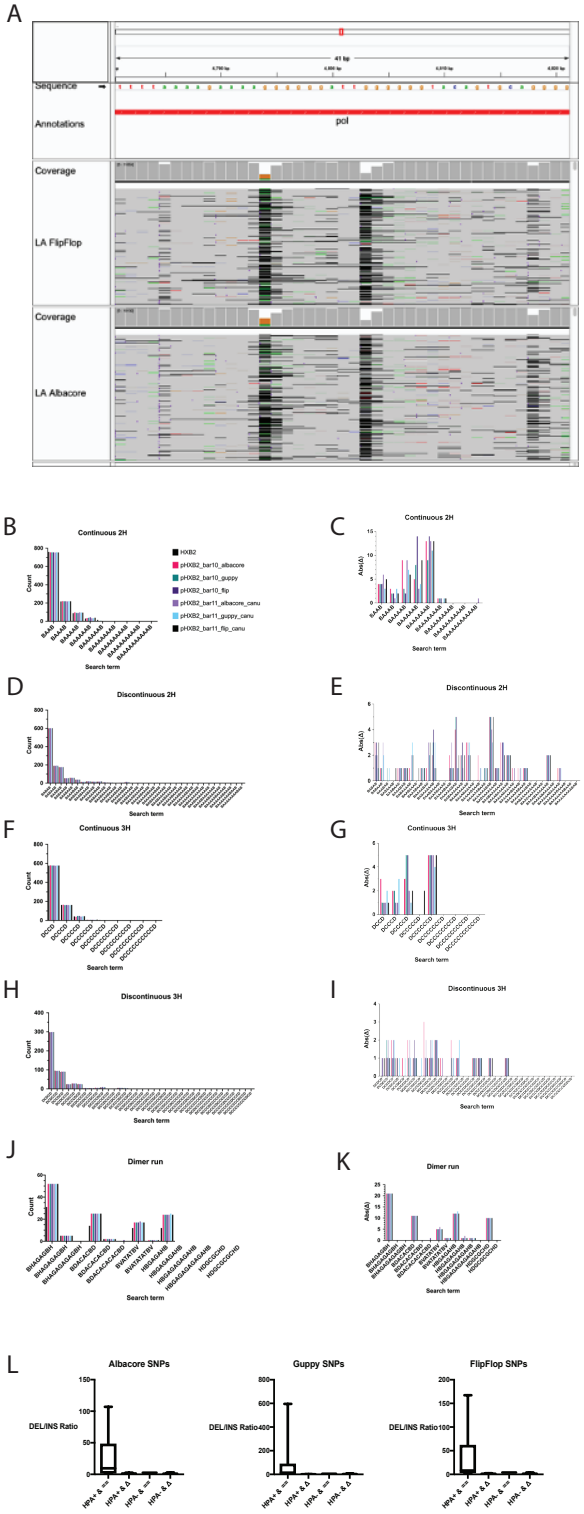
Supplemental Figure 4C: Mappability of long reads over contigs during assembly quality control.



Coverage depends on context. Abrupt changes in coverage from terminal regions of HXB2 (**Figure 1F, Supplemental Figures 1 and 3)** were artifacts from supplying mappers with an HIV reference without a plasmid backbone. Long reads from barcode 10 (LA Taq) mapped with minimap2 [23] 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 4A, 4B**.

970 **Supplemental Figure 5: Homopolymers and dimer runs are ONT artifacts in unpolished**

971 **assemblies.**

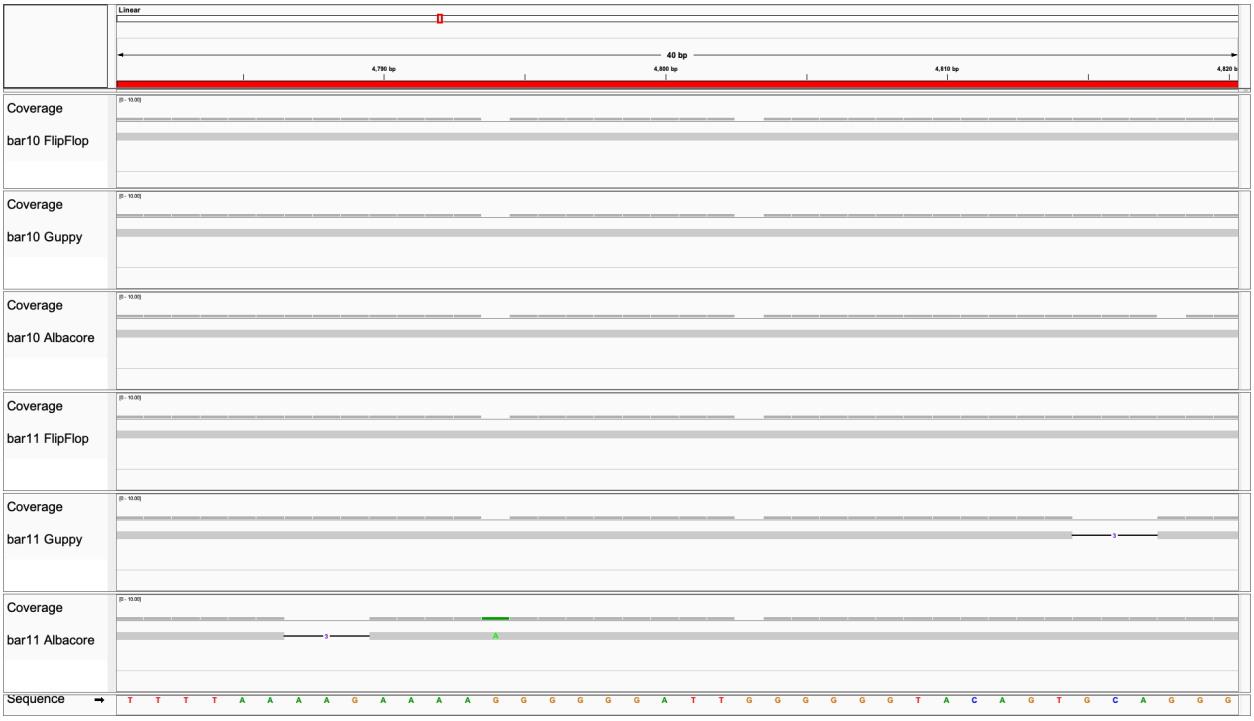


972

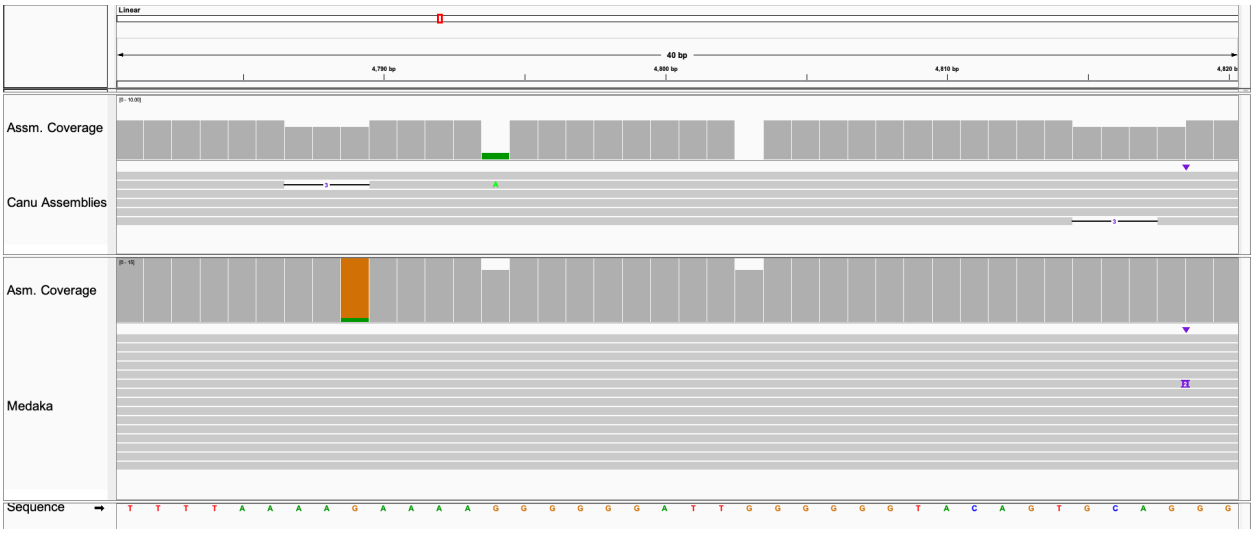
973 Supplemental Figure 5A: A set of homopolymer tracks from HXB2 plasmid. Alignments with
974 BWA-MEM shown from FlipFlop (top) and Albacore (bottom) basecalled reads. Mapping is
975 pre-assembly.
976 Supplemental Figure 5B: Continuous 2H counts in unpolished assemblies. 2H = A or T
977 homodimers.
978 Supplemental Figure 5C: Continuous 2H Absolute Difference.
979 Supplemental Figure 5D: Discontinuous 2H counts in unpolished assemblies.
980 Supplemental Figure 5E: Discontinuous 2H Absolute Difference.
981 Supplemental Figure 5F: Continuous 3H counts in unpolished assemblies. 3H = C or G
982 homodimers.
983 Supplemental Figure 5G: Continuous 3H Absolute Difference.
984 Supplemental Figure 5H: Discontinuous 3H counts in unpolished assemblies.
985 Supplemental Figure 5I: Discontinuous 3H Absolute Difference.
986 Supplemental Figure 5J: Dimer run counts in unpolished assemblies.
987 Supplemental Figure 5K: Dimer run Absolute Difference. Dimer runs as pairs are the most
988 problematic, with runs as triplets being resolvable by ONT.
989 Supplemental Figure 5L: The ratio of deletions to insertions is higher at mismatches both
990 adjacent to homopolymers and similar to neighbor bases. Box plot shows median ("x" is mean)
991 and quartile ranges. Y-axis is ratio. HPA: homopolymer-adjacent. ==: same as neighbor base. Δ:
992 different than neighbor base. Higher coverage (above ~10) usually makes up for current error
993 profile. Above true for Albacore, Guppy, and FlipFlop.

994 **Supplemental Figure 6: Assembly partially resolved homopolymers, which are improved**

995 **by polishing**



996



997

998 Top: Six ONT-only assemblies. Bottom: polished ONT-only assemblies, varying Medaka

999 models. Deletions at 5’ of G homopolymers were not corrected, regardless of basecaller or Taq

1000 isoform. Note that polishing was not performed. IGV window is Linear:4,781-4,820. Bottom:

1001 polishing canu assemblies with medaka abrogated most ONT artifacts. Best medaka setting

1002 tested: r941_min_high_g330.

1003