Full-coverage sequencing of HIV-1 provirus from a reference plasmid

Alejandro R. Gener[1,2,3,4]

[1]Baylor College of Medicine, Integrative Molecular and Biomedical Sciences Program, Houston, Texas, USA
[2]Department of Medicine, Nephrology, Baylor College of Medicine, Houston, Texas, USA
[3]MD Anderson Cancer Center, Department of Genetics, Houston, Texas, USA
[4]Universidad Central del Caribe, School of Medicine, Bayamón, Puerto Rico, USA

Corresponding author
Alejandro R. Gener: gener@bcm.edu ; itspronouncedhenner@gmail.com

Conflict(s) of interest:
None.

Key words: nanopore, DNA sequencing, HIV-1, HIV sequence variability, retrovirus, viral heterogeneity

Funding: This work was funded in part by institutional support from Baylor College of Medicine; the Human Genome Sequencing Center, Baylor College of Medicine; private funding by East Coast Oils, Inc., Jacksonville, Florida, and my own private funding, including Student Genomics (manuscripts in prep). I have also received the PFLAG of Jacksonville scholarship for multiple years.

ABSTRACT

Objective(s): To evaluate nanopore DNA sequencing for sequencing full-length HIV-1 provirus.

Design: I used nanopore sequencing to sequence full-length HIV-1 from a plasmid (pHXB2).

Methods: pHXB2 plasmid was processed with the Rapid PCR-Barcoding library kit and sequenced on the MinION sequencer (Oxford Nanopore Technologies, Oxford., UK). Raw fast5 reads were converted into fastq (base called) with Albacore, Guppy, and FlipFlop base callers. Reads were first aligned to the reference with BWA-MEM to evaluate sample coverage manually. Reads were then assembled with Canu into contigs, and contigs manually finished in SnapGene.

Results: I sequenced full-length HXB2 HIV-1 from 5' to 3' LTR (100%), with median per-base coverage of over 9000x in one 12-barcoded 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 on either side. At least 20 variants were discovered in pHXB2 compared to reference.

Conclusions: The MinION sequencer performed as-expected, covering full-length HIV. The discovery of variants in a dogmatic reference plasmid demonstrates the need for single-molecule sequence verification moving forward. These results illustrate the utility of long read sequencing to advance the study of HIV at single integration site resolution.

INTRODUCTION

Researchers have yet to observe integrated HIV-1 in one take, hindering locus-specific studies. Complicating matters is the cycling that all orterviruses [1] undergo between two major states (as infectious virion RNA and integrated proviral DNA), repetitive viral sequences like long terminal repeats (LTRs), non-integrated forms [2], and alternative splicing of viral mRNAs [3]. Short read DNA sequencing (less than 150 base pairs (bp) in most reported experiments, but up to 500 bp) provides some information, but analyses require high coverage and extensive effort (non-exhaustive examples [4,5]). Repetitive sequences in HIV, rarity of integration events *in* vivo, and non-integrated HIV limit the ability to assign DNA variants to specific loci within each provirus, as well as the proviral integration sites (partially reviewed in [6]). Others have studied the products of HIV transcription by analyzing cellular HIV mRNAs with long read sequencing, but with reverse transcription of these viral RNAs [7], which are well known to diverge from provirus. Despite progress, researchers have yet to observe the genome of HIV-1 as complete provirus (integrated DNA), hindering locus-specific studies. To this end, current long read DNA sequencing clearly surpasses the limitations of read length of leading next-generation sequencing platforms. Here I used the MinION sequencer (Oxford Nanopore Technologies, Oxford UK) to sequence HIV-1 as plasmid (pHXB2), with the goal of evaluating the technology for future applications.

METHODS

HIV-1 plasmid

A plasmid containing the HIV-1 reference strain HXB2 was acquired from the AIDS Reagent Program. HXB2 is one of the earliest clinical isolates, but it was unknown whether this plasmid (a molecular clone of the original isolate) was ever sequence-verified after the reference sequence (below) was deposited.

HIV-1 sequences

The reference sequence of HXB2 is from NCBI, Accession number K03455.1.

DNA sequencing

Sequence of a plasmid containing HXB2 was sequence-verified with long-read nanopore sequencing on the MinION (Oxford Nanopore Technologies, Oxford UK). While competing long read sequencing platforms exist, they were not evaluated in this Student Genomics pilot and will not be discussed further. Unless otherwise noted, reagents and software were purchased from Oxford Nanopore Technologies. Briefly, stock plasmid was diluted to 5 ng final amount in milliQ water and processed with Rapid PCR Barcoding kit SQK-RPB004 along with 10 other barcoded samples (not discussed further in this manuscript) following ONT protocol RPB_9059_V1_REVA_08MAR2018. Two DNA polymerases were evaluated (barcode 10 used LA Taq (Takara); barcode 11 Taq (Sigma-Aldrich). Libraries were loaded onto a MinION flow cell version R9.4.1 and a 48-hour sequencing run was completed with MinKNOW (version 1.10.11) on a mid 2010 MacBook Pro (Apple Inc. Cupertino CA USA) with 16 Gb RAM and 2.7 GHz dual-core i7 processor (Intel Corp., Santa Clara CA USA). Residual reads from subsequent runs were pooled for final analyses.

Raw data was base called (converted from fast5 to fastQ format) with Albacore version 2.3.4 (older base caller), Guppy version 2.3.1 (current official), and FlipFlop (development Guppy) on a compute cluster. At the time, only albacore allowed for direct demultiplexing. Guppy (guppy_barcoder) provided similar functionality by providing a summary file with read ID and barcode, and grep was used to pull read length and mean read quality (**Supplemental Figure 1**) from guppy_basecaller summary file.

Demultiplexed basecalled reads were fed into Canu version 8 implemented on a cluster [8]. Genome size was estimated to be 16 Kb from agarose gel of undigested, but naturally degraded pHXB2 (data not shown). Alignments to reference were made with BWA-MEM [9], implemented in Galaxy (usegalaxy.org) [10]. To down sample data from barcode 10/LA Taq sample, I used seqtk_sample, seed 1, (Heng Li, Broad Institute) to take the amount of data (file size) in barcode 11 and divided by the amount of data (file size) in barcode 10. Seqtk was implemented in Galaxy. SAMTOOLS Flagstat [11] was used for mapping stats from BAM files from BWA-MEM, implemented in Galaxy. SnapGene version 4.3.4 was used to manually annotate contigs from Canu. Blastn (NCBI) was used to identify unknown regions of pHXB2.

Statistics

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

RESULTS

In one of the two barcoded experiments, I achieved extremely high coverage of pHXB2 (median per-base coverage of over 9000x) (**Figure 1**). In addition to kilobase-long reads, at least one read spanned the entire HIV genome (**Figure 2**).

I assembled the previously undefined plasmid pHXB2 relative to the previously published HXB2 reference from libraries made with either high fidelity or standard DNA polymerases (**Figure 3**). Genome assembly was done using Canu (a command-line utility recommended by most in the long DNA field). Canu's final output is a set of contiguous DNA sequences (contigs) as FASTA files, but a consequence of assembling plasmid sequences with this tool is partial redundancy at contig ends. Manual end-trimming of contigs was performed in SnapGene based on an estimated length of 16 kilobases. Top blastn hits from barcode 10/LA Taq pHXB2 base called with FlipFlop were as follows: for the main backbone (with origin of replication and antibiotic selection cassette for cloning), shuttle vector pTB101-CM DNA, complete sequence (based on pBR322), from 4352-8340; for the upstream element (relative to 5' LTR), Homo sapiens chromosome 3 clone RP11-83E7 map 3p, complete sequence from 58,052 to 59,165; for the downstream element, cloning vector pNHG-CapNM from 10,204 to 11,666. Other identified elements included Enterobacteria phage SP6, complete sequence from 39,683 to 39,966. Identities of query to HXB2 and hits were all approximately 99%.

On closer examination of the two pHXB2 libraries prepared in parallel and sequenced simultaneously, I counted 20 single nucleotide polymorphisms (SNPs) in this reference clone of HXB2, the reference clinical isolate for HIV-1 (**Figure 4** and **Table 1**). These mismatches were seen in all Canu assemblies and verified in IGV and/or SnapGene. This represents a ~0.2% divergence (20/9717) from the reference, which was assumed to

have perfect identity (0% divergence). Transitions were more common (14/20), and curiously falls in line with a previous report of increased transitions over transversions in infection models, because transversions are more likely to be deleterious to viral replication (i.e.: to cause protein-coding changes) [12]. Indeed, less than half (9/20) of the observed SNPs occurred in protein-coding regions, even though 92% of HXB2 is coding (per reference, 791/9719). Of those 9 SNPs in protein-coding regions, 4 caused non-synonymous mutations. Note that one of those occurs in a region overlapping both gag and pol regions, however only pol exhibited a non-synonymous change from valine to isoleucine (in p6, at position 2259 relative to HXB2). Other non-synonymous variants occurred at 4609 (in p31 integrase, arginine to lysine), 7823 (in ASP antisense protein, glycine to arginine), and 9253 (in nef, isoleucine to valine).

As previously reported [13], per-read variability was higher near homopolymers (runs of the same base) (**Supplemental Figures 2).** At the level of consensus (made from sequences mapped to reference HXB2), homopolymers contributed few, if any, obvious errors. While dips occurred at certain points near homopolymers, the consensus did not change at the sequencing depth used in this study for either barcoded pHXB2 samples, or for down sampled high-coverage data (**Figure 1**). Another interpretation is that homopolymers tend to seem truncated with ONT, with more reads in support of shorter homopolymers. Canu assemblies showed base caller-variability (**Supplemental Table 1**). That said, newer base callers tend to produce fewer gaps. The ratio of per-read deletions to per-read insertions (DEL/INS) was much higher for SNPs occurring at homopolymers and near the same base, and this difference was maintained between all base callers used (**Supplemental Figure 3**).

DISCUSSION

Nanopore sequencing surpassed the read length limitations of other sequencing modalities such as Sanger sequencing and sequencing by synthesis by at least two orders of magnitude. This work represents the first instance of complete sequencing of the HIV provirus as plasmid and contributed to the identification of single nucleotide variants which may not have been easily determined using other sequencing modalities, illustrating the importance of validating molecular clone reagents in their entirety.

First complete pass over all HIV information in reference plasmid

HIV provirus is believed to occur naturally as one or a few reverse-transcribed DNA forms integrated into the host nuclear genome. Depending on where integration occurs, local GC or AT content might cause problems for detecting integrants. HIV also has conserved transitions from areas of higher GC content (~50%) to content approximating average human GC content (~40%). To avoid sequencing bias from PCR and to accommodate for the potential heterogeneity of HIV sequences, I fractionated whole sample directly (as opposed to PCR-barcoding select amplicons) with tagmentation provided in the Rapid PCR-Barcoding kit (ONT). Tagmentation uses transposon-mediated cleavage and ligation of barcode adapters for later PCR amplification. A consequence of this fractionation was a distribution of reads (see **Supplemental Figure 1A**) shorter than longer reads reported elsewhere for ONT experiments [14]. Based on this distribution and the level of coverage, it was expected that HIV might be covered from end to end, but this would have been exceptional. That said, an example is presented here (see **Figure 2**).

Errors in a reference plasmid of HIV-1

HIV is the most studied human pathogen. By leveraging long reads with the MinION, I was able to find mutations in highly repetitive LTRs, in addition to mutations in protein-coding regions. Two independent experiments (PCR-barcoded libraries made with Takara Long Amplicon Taq and Sigma-Aldrich Taq master mixes) yielded at least 20 single nucleotide variants which differed from the reference sequence. These 20 SNPs were concordant between both experiments, and across the three base callers used. Other differences between the PCR-barcoded samples may be due to PCR amplification errors, which may yield asymmetric base differences (a site with one predominant allele in one experiment). Another possibility may be plasmid admixture in the reference stock. The latter may be more likely in the event of symmetric differential calling (example, a site which differs from the reference, but is called as two or three different bases at equal ratios). Such sites represent a challenge for conventional (i.e. non-long read) sequencing modalities. The pHXB2 sequenced was a reference plasmid, with unknown divergence between the published reference for HXB2. In this study I was also able to confirm that the backbone of this plasmid is from pBR322, aiding in the continued use of this important reagent, and illustrating the need of full-length reagent validation in all clonal viral lines moving forward. I suggest that all sequences tagged for use in clinical settings likewise be sequence-verified at the level of single-molecule sequencing, lest preventable errors go unchecked.

Improvement in ONT base callers within the past year

Albacore, Guppy, and FlipFlop basecallers were compared. Each produced reads of similar length distributions (relative to polymerase used), while Guppy and FlipFlop produced improved and best performance relative to quality score distributions (see **Supplemental Figures 1B**). Interestingly, while read length distributions were affected by

fidelity of polymerases evaluated in this work, mean quality distributions were not. This is important because of the differences in cost between higher fidelity enzymes and classic Taq. In consideration of library prep, choice of enzyme used should be based on the desired read length distribution. Regarding read mapping, the increase in mean quality score between these base callers didn't change overall mapping much given longer reads (data not shown), but did facilitate demultiplexing, resulting in approximately ~10% increases in barcoded libraries (shift in reads from unclassified to a given barcode). FlipFlop tended to handle homopolymers better than previous base callers. Homopolymers in HXB2 tended to exhibit apparent deletions near 5' ends of homopolymers (upstream due to technical artifact in BWA-MEM), but because consensus is conserved (example, at least 80% of base in called read set is identical to reference), it is unlikely that any of the homopolymer deletions are real in these experiments.

Current limitations of the approaches used in the present study are 1.) the cost of long-read sequencing, regardless of platform, compared to the cheaper short reads from sequencing by synthesis, 2.) long DNA extraction methods, and 3.) the lower per-base accuracy (low-mid 90's vs. 98-99%). As the price of long-read sequencing decreases, as the hardware and software used in base calling improve, and as library protocols improve, the cost of obtaining usable data from long reads will become negligible compared to the ability to answer historically intractable questions. Future work will include optimizing DNA extraction protocols with the goal of capturing high-coverage fuller glimpses of each HIV proviral integration site in *in vivo* HIV models and patient samples. This work has broad implications for all cells infected by both integrating and non-integrating viruses, and for the characterization of targeted regions in the genome which may be recalcitrant to previous sequencing methods. Long read sequencing is an

important emerging tool defining the post-scaffold genomics era, allowing for the characterization of anatomical landmarks at the genomic scale.

ACKNOWLEDGEMETS

ARG conceived of this project, performed experiments, analyzed results, and wrote the manuscript.

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

I would like to thank Drs. Steven Richards, Qingchang Meng and the staff of HGSC Research and Development team for their earlier support in nanopore adoption. I would like to thank the team at Oxford Nanopore Technologies for their timely improvements and continued R&D. I would also like to thank other members of the Paul E. Klotman lab, including Dr. Deborah P. Hyink (who thoughtfully helped to edit the manuscript), Taneasha Washington, and former members Dr. Gokul C. Das and Alexander Batista. I would like to thank Drs. Alana Canupp and Jim Maruniak for their early interest in my scientific development, and for the passion that they show in everything that they do.

REFERENCES

1 Krupovic M, Blomberg J, Coffin JM, Dasgupta I, Fan H, Geering AD, *et al.* Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. *J Virol* 2018; **92**:e00515-18.

2 Graf EH, Mexas AM, Yu JJ, Shaheen F, Liszewski MK, Di Mascio M, *et al.* Elite suppressors harbor low levels of integrated HIV DNA and high levels of 2-LTR circular HIV DNA compared to HIV+ patients on and off HAART. *PLoS Pathog* 2011; **7**. doi:10.1371/journal.ppat.1001300

3 Cuesta I, Mari A, Ocampo A, Miralles C, Pérez-castro S, Thomson MM. Sequence Analysis of In Vivo -Expressed HIV-1 Spliced RNAs Reveals the Usage of New and Unusual Splice Sites by Viruses of Different Subtypes. 2016; :1–24.

4 Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D, *et al.* Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evol* 2018; **4**:1–13.

5 Bruner KM, Wang Z, Simonetti FR, Bender AM, Kwon KJ, Sengupta S, *et al.* A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* 2019; **566**:120–125.

6 Pinzone MR, O'Doherty U. Measuring integrated HIV DNA ex vivo and in vitro provides insights about how reservoirs are formed and maintained. *Retrovirology* 2018; **15**:1–12.

7 Ocwieja KE, Sherrill-Mix S, Mukherjee R, Custers-Allen R, David P, Brown M, *et al.* Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res* 2012; **40**:10345–10355.

8 Walenz BP, Koren S, Bergman NH, Phillippy AM, Miller JR, Berlin K. Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Res* 2017; **27**:722–736.

9 Li H, Durbin R. Fast and accurate long-read alignment with Burrows – Wheeler transform. 2010; **26**:589–595.

10 Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016; **44**:W3–W10.

11 Wysoker A, Fennell T, Marth G, Abecasis G, Ruan J, Li H, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**:2078–2079.

12 Lyons DM, Lauring AS. Evidence for the Selective Basis of Transition-to-Transversion Substitution Bias in Two RNA Viruses. *Mol Biol Evol* 2017; **34**:3205–3215.

13 Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015; **12**:733–735.

14 Payne A, Holmes N, Rakyan V, Loose M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv* 2018; :312256.

15 Nicol JW, Helt GA, Steven G. Blanchard J, Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *BMC Bioinformatics* 2009; **25**:2730–2731.

FIGURES

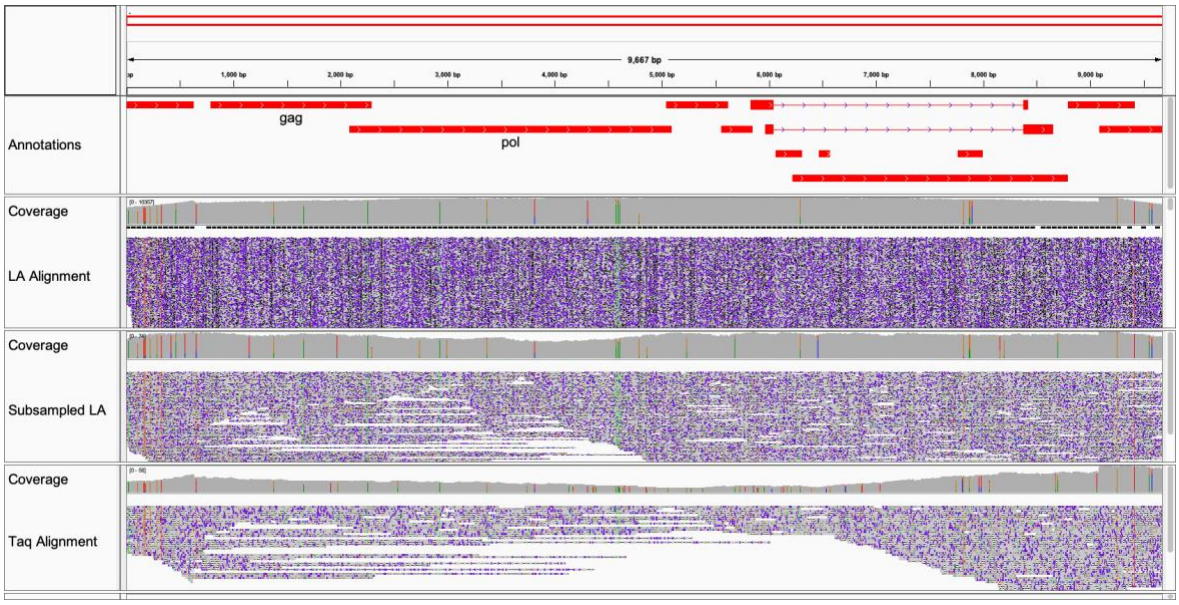


Figure 1: Nanopore reads carpet HXB2. Top 2 Coverage and Alignment panels from barcoded library 10 (bar10 = LA Taq). Bottom from Barcode 11 (bar11 =Taq). BWA-MEM was used to map reads called with FlipFlop to HXB2. Color-coding of alignment panes: Purple is an insertion in a given read relative to reference. White is either a deletion in a given read or space between two aligned reads. Note that per-read “insertion” and “deletion” do not necessarily represent true insertions or deletions actually present in the sample. Automated assembly followed by manual consensus building converts these overlapping reads into approximations of the ground truth. Visualized in Integrated Genome Browser [15].

```
@6fbf0205-5195-460e-8e28-930db50e5d79 runid=0b284792282af9a6d7275cfca845556bb4b8ac7f sampleid=Student_Genomics_Run_1 read=87223 ch=460 start_time=2018-05-09T20:57:492 barcode=barcode10
TGTTGTACTTCGTTCAGTTACGTATTGGGTGTTTAGCGGTTCTTTACCGTTGACTAGCAAGCGTTGAAAACGCTGTCCCTCTCCGTTTTCGTGCGCCGCTTCAATGTGTATAATATACATATGTATATGTATATGTATATGTATATGTATATGTATATACTTCGTTGTATATATCTTCGTTGTATATATCTTCGTTGTATATATC
```

nr/nt returned 100% coverage and 93.02% identity to HIV-1 HXB2. This read was 11,487 bases long, with mean quality score 11.984396. Base called using Guppy 2.3.1 with FlipFlop config.

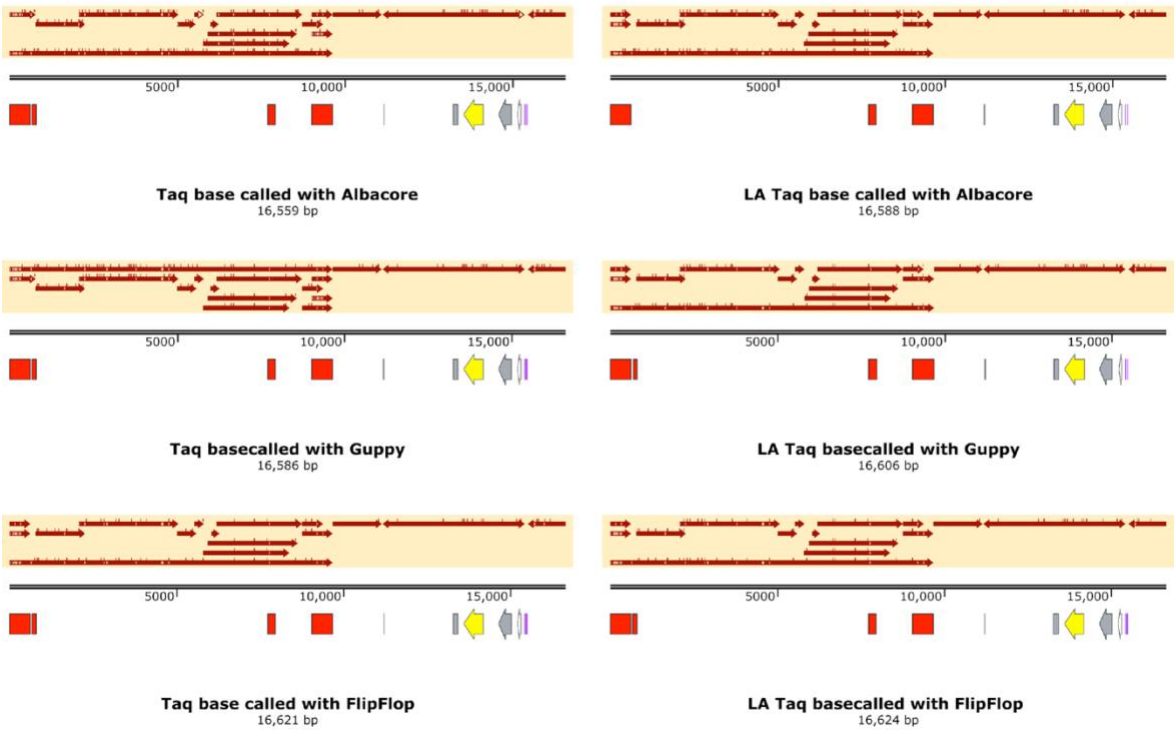


Figure 3: Assembled HXB2 plasmid, varying base caller and polymerase. Each pane (n=6) summarizes the results of contig curation. Tan boxes contain alignments of reference/blastn hits to contig (Black line). Divergence from reference decreases with newer base callers, and with long amplicon DNA polymerase (Sigma-Aldrich Taq vs. LA Taq by Takara). Annotations below the black lines were automatically imported with SnapGene and represent sequences with minimal/no divergence from common cloning features. The regions corresponding to HXB2 are between red boxes (LTRs). Viewed in SnapGene.

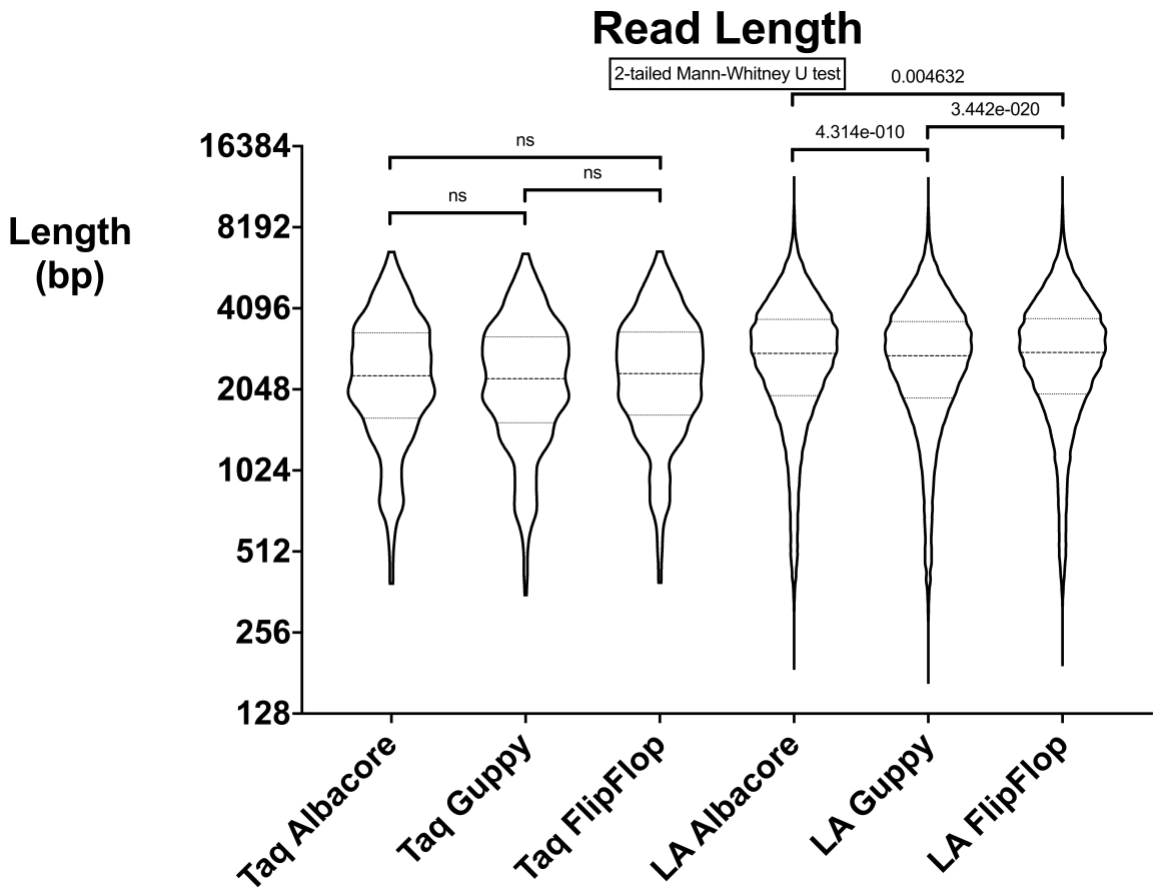
Figure 4: Single nucleotide variants from HXB2 plasmid. Base called with FlipFlop. Visualized in Integrated Genome Browser [15]. Gray indicates per-base consensus accuracy ≥ 80%.

Site	Position	Coverage	Change	Substitution Class	Change	Mutation Class (Syn/Non/Stop)	Homopolymer-adjacent?	Same as neighbor?	LANL Feature	Subfeature	Frame
1	24	5669	C>A	transversion	NA	NA	yes	yes	5'LTR	U3	NA
2	108	5028	A>G	transition	NA	NA	yes	yes	5'LTR	U3	NA
3	164	6674	G>T	transversion	NA	NA	yes	no	5'LTR	U3	NA
4	168	6722	T>G	transversion	NA	NA	yes	yes	5'LTR	U3	NA
5	176	6817	A>G	transition	NA	NA	yes	yes	5'LTR	U3	NA
6	182	6427	C>T	transition	NA	NA	yes	no	5'LTR	U3	NA
7	227	6488	A>G	transition	NA	NA	yes	yes	5'LTR	U3	NA
8	291	7122	A>G	transition	NA	NA	no	no	5'LTR	U3	NA
9	333	7521	C>T	transition	NA	NA	no	no	5'LTR	U3	NA
10	654	7927	C>T	transition	NA	NA	no	no	None	None	NA
11	1659	7378	aaG>aaA	transition	None	Syn	yes	yes	gag	p24, p55	gag frame 1
12	2259	9070	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	9126	aaG>aaA	transition	None	Syn	yes	yes	pol	p51 RT	pol frame 3
14	3812	9628	ccC>ccT	transition	None	Syn	yes	yes	pol	p51 RT	pol frame 3
15	4574	9638	acT>acA	transversion	None	Syn	no	no	pol	p31 IN	pol frame 3
16	4596	9736	Ggt>Agt	transition	None	Syn	yes	no	pol	p31 IN	pol frame 3
17	4609	9272	aGg>aAg	transition	Arg>Lys	Non	yes	yes	pol	p31 IN	pol frame 3
18	7823	9304	gcC>gcG Ggc>Cgc	transversion	ASP:Gly>Arg	Syn/Non	no	no	gp41	RRE, also ASP	gp41 frame 3, ASP -2
19	9253	9408	Ata>Gta	transition	Ile>val	Non	no	yes	nef/3'LTR	also U3	nef frame 1
20	9418	8786	C>T	transition	NA	NA	no	no	3'LTR	U3	NA

Table 1: Summary of pHXB2 sample divergence from reference HXB2. Coverage numbers are with Albacore basecalled reads, BWA-MEM in Galaxy alignment, viewed in IGV. Base-1 (first base is numbered 1, 2nd 2, etc..), relative to HXB2, Accession number K03455.1. Changed base represented as upper-case. Annotated as codon if in protein-coding region. No deletions or insertions were predicted from manual inspection. Abbreviations, ASP: antisense protein, RRE: rev-response element, NA: not applicable.

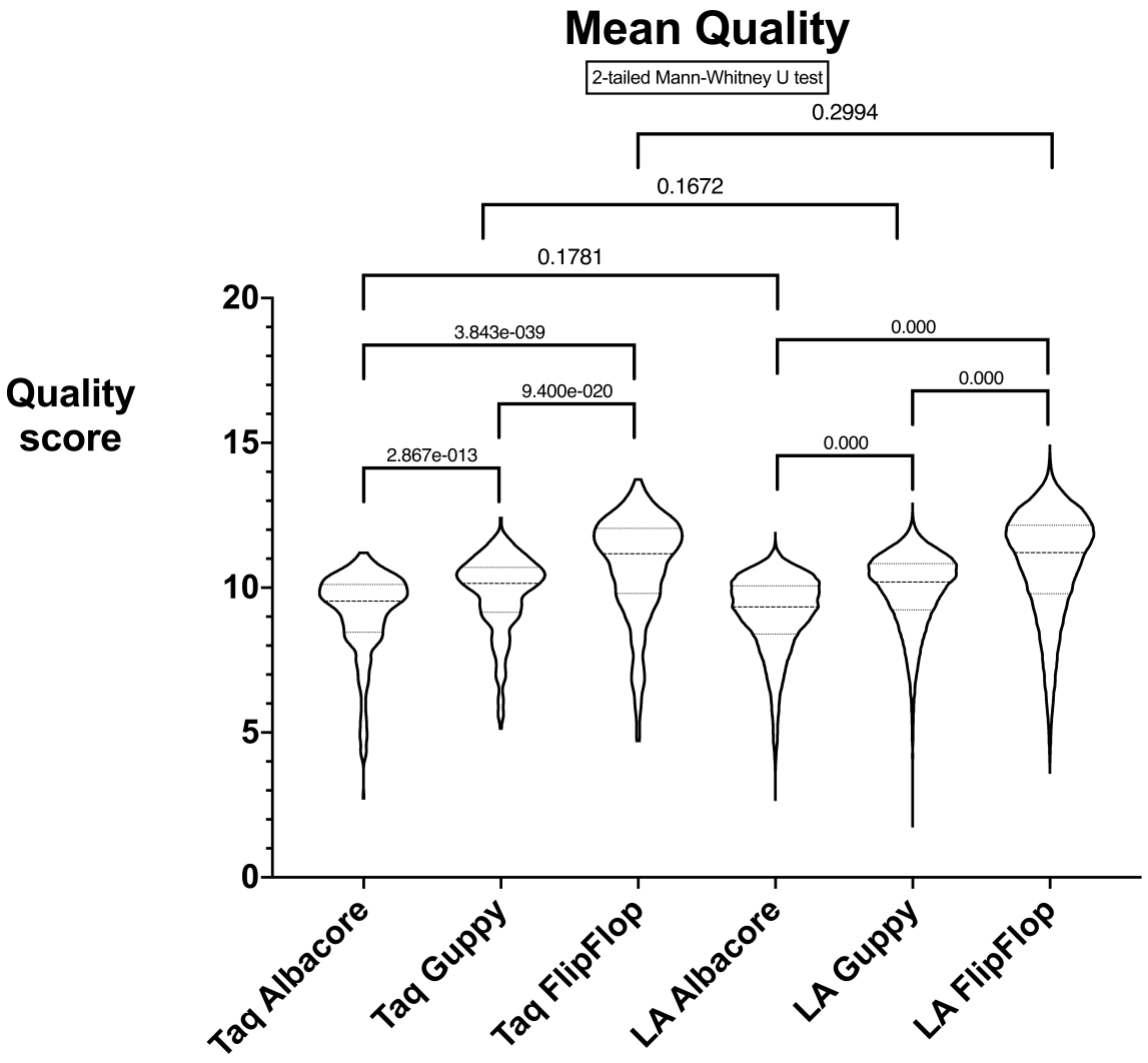
Syn: synonymous mutation. Non: non-synonymous mutation. Stop: stop codon/non-sense mutation. LTR: long terminal repeat. RT: reverse transcriptase. IN: integrase. LANL: Los Alamos National Laboratory HIV Database. Data from two separate sequencing experiments on the same plasmid sample.

SUPPLEMENTAL DATA



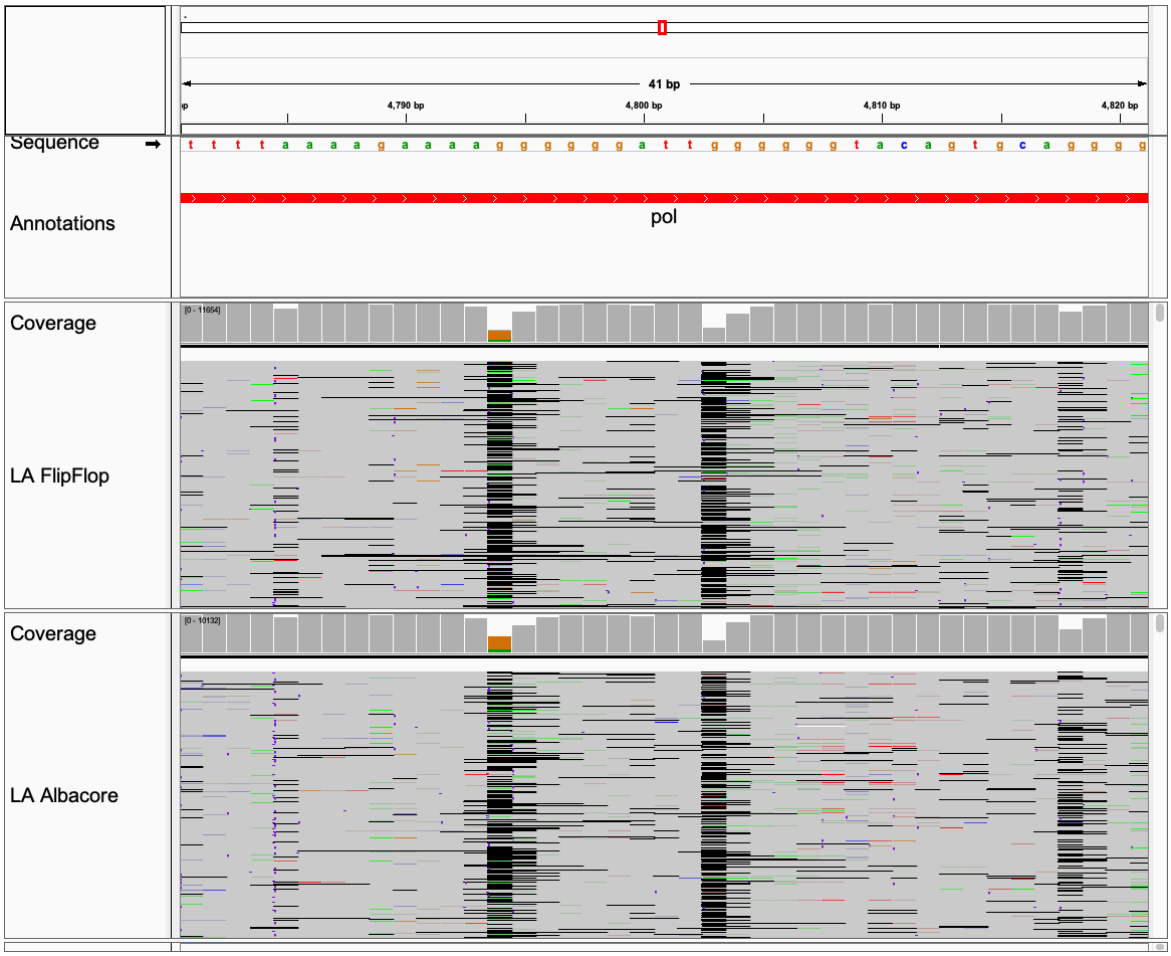
Supplemental Figure 1A: Read stats with different callers/aligners.

Median (big dash) and quartiles (little dash). Read lengths increase with higher fidelity Taq.

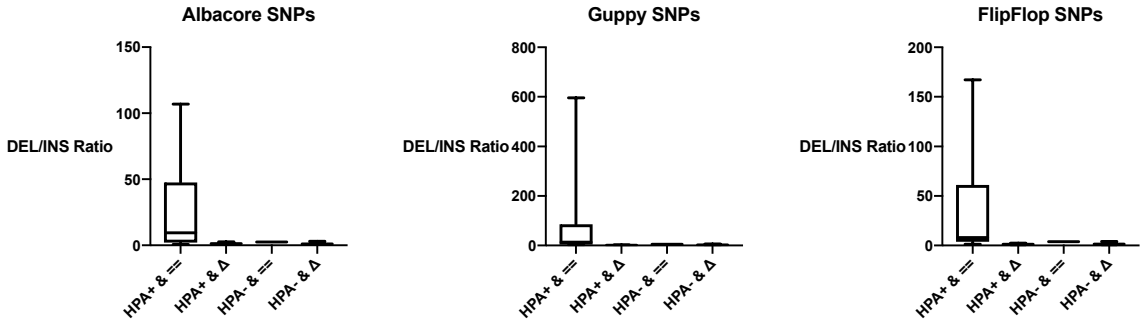


Supplemental Figure 1B: Newer base callers increase read mean quality.

Median (big dash) and quartiles (little dash). Effect of enzyme version was not statistically significant.



Supplemental Figure 2: A set of homopolymer tracks from HXB2 plasmid. Visualized in Integrated Genome Browser [15]. Alignments with BWA-MEM shown from FlipFlop (top) and Albacore (bottom) base called reads.



Supplemental Figure 3: The ratio of deletions to insertions is higher at mismatches both adjacent to homopolymers and similar to neighbor bases. Box plot shows median ("x" is mean) and quartile ranges. Y-axis is ratio. HPA: homopolymer-adjacent. ==: same as neighbor base. Δ: different than neighbor base. Higher coverage (above ~10) usually makes up for current error profile. Above true for Albacore, Guppy, and FlipFlop.

	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
ownstream 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

Supplemental Table 1: Variation in assemblies at the feature level. Assembled with Canu. NA denotes features which may not have matched exactly, but which were collapsed with adjacent features for ease of counting. Mapped with SnapGene.

LIST OF DATA (To be deposited in SRA in near future)

Albacore base called barcode 10 files
Guppy base called barcode 10 files
FlipFlop base called barcode 10 files
Albacore base called barcode 11 files
Guppy base called barcode 11 files
FlipFlop base called barcode 11 files
.dna files of contigs (n=6)