



This is a repository copy of *Low-bias RNA sequencing of the HIV-2 genome from blood plasma*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/141767/>

Version: Published Version

Article:

James, K.L., de Silva, T.I. orcid.org/0000-0002-6498-9212, Brown, K. et al. (5 more authors) (2018) Low-bias RNA sequencing of the HIV-2 genome from blood plasma. *Journal of Virology*, 93. e00677-18. ISSN 0022-538X

<https://doi.org/10.1128/JVI.00677-18>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>



Low-Bias RNA Sequencing of the HIV-2 Genome from Blood Plasma

Katherine L. James,^{a,b} Thushan I. de Silva,^{c,e} Katherine Brown,^d Hilton Whittle,^e Stephen Taylor,^f Gilean McVean,^b Joakim Esbjörnsson,^{a,g} Sarah L. Rowland-Jones^a

^aNuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

^bWellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

^cImperial College London, Department of Medicine, London, United Kingdom

^dCGAT, University of Oxford, Oxford, United Kingdom

^eMRC Unit The Gambia at the London School of Hygiene and Tropical Medicine, Banjul, the Gambia

^fWeatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom

^gDepartment of Laboratory Medicine, Lund University, Lund, Sweden

ABSTRACT Accurate determination of the genetic diversity present in the HIV quasi-species is critical for the development of a preventative vaccine: in particular, little is known about viral genetic diversity for the second type of HIV, HIV-2. A better understanding of HIV-2 biology is relevant to the HIV vaccine field because a substantial proportion of infected people experience long-term viral control, and prior HIV-2 infection has been associated with slower HIV-1 disease progression in coinfecting subjects. The majority of traditional and next-generation sequencing methods have relied on target amplification prior to sequencing, introducing biases that may obscure the true signals of diversity in the viral population. Additionally, target enrichment through PCR requires *a priori* sequence knowledge, which is lacking for HIV-2. Therefore, a target enrichment free method of library preparation would be valuable for the field. We applied an RNA shotgun sequencing (RNA-Seq) method without PCR amplification to cultured viral stocks and patient plasma samples from HIV-2-infected individuals. Libraries generated from total plasma RNA were analyzed with a two-step pipeline: (i) *de novo* genome assembly, followed by (ii) read remapping. By this approach, whole-genome sequences were generated with a 28× to 67× mean depth of coverage. Assembled reads showed a low level of GC bias, and comparison of the genome diversities at the intrahost level showed low diversity in the accessory gene *vpx* in all patients. Our study demonstrates that RNA-Seq is a feasible full-genome *de novo* sequencing method for blood plasma samples collected from HIV-2-infected individuals.

IMPORTANCE An accurate picture of viral genetic diversity is critical for the development of a globally effective HIV vaccine. However, sequencing strategies are often complicated by target enrichment prior to sequencing, introducing biases that can distort variant frequencies, which are not easily corrected for in downstream analyses. Additionally, detailed *a priori* sequence knowledge is needed to inform robust primer design when employing PCR amplification, a factor that is often lacking when working with tropical diseases localized in developing countries. Previous work has demonstrated that direct RNA shotgun sequencing (RNA-Seq) can be used to circumvent these issues for hepatitis C virus (HCV) and norovirus. We applied RNA-Seq to total RNA extracted from HIV-2 blood plasma samples, demonstrating the applicability of this technique to HIV-2 and allowing us to generate a dynamic picture of genetic diversity over the whole genome of HIV-2 in the context of low-bias sequencing.

Citation James KL, de Silva TI, Brown K, Whittle H, Taylor S, McVean G, Esbjörnsson J, Rowland-Jones SL. 2019. Low-bias RNA sequencing of the HIV-2 genome from blood plasma. *J Virol* 93:e00677-18. <https://doi.org/10.1128/JVI.00677-18>.

Editor Frank Kirchhoff, Ulm University Medical Center

Copyright © 2018 James et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Joakim Esbjörnsson, joakim.esbjornsson@med.lu.se, or Sarah L. Rowland-Jones, sarah.rowland-jones@ndm.ox.ac.uk. G.M., J.E., and S.L.R.-J. contributed equally to this article.

Received 9 May 2018

Accepted 14 September 2018

Accepted manuscript posted online 17 October 2018

Published 10 December 2018

KEYWORDS HIV-2, RNA sequencing, next-generation sequencing, *vpx*, whole genome

Human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2), the two causative agents of AIDS, are human pathogens of high importance (1). Following the introduction of HIV-1 and HIV-2 into human populations through zoonotic transmission of simian immunodeficiency viruses (SIVs) infecting several species of apes and non-human primates, HIV-1 and HIV-2 are estimated to have infected more than 75 million people worldwide, resulting in over 40 million deaths (2).

While HIV-1 and HIV-2 share some common features, a major difference between the two viruses is the typical viral load (VL) associated with chronic infection. In patients infected with HIV-2, viral load is strongly correlated with disease progression, and a large proportion (~37% in the Caió cohort) maintained undetectable viral loads and high CD4 counts in the absence of treatment during follow-up (sometimes for more than two decades) (3). In addition, based on previous data from this cohort, subjects with VLs of >10,000 were regarded as having a high likelihood of subsequent progression (4). Indeed, patients with a viral load of more than 10,000 copies/ml have been suggested to be HIV-2 progressors with a reduced survival probability that is similar to that seen in HIV-1-infected individuals in the absence of treatment (5). Furthermore, lack of HIV-2 control is associated with lower viral loads than for HIV-1 in patients matched by disease stage (4, 6–8).

HIV-1 disease progression has also been correlated with viral coreceptor use or molecular properties like glycosylation patterns, charge, and length of the envelope gene (9–12). Although cytopathic CXCR4-using virions have been isolated from HIV-2-infected individuals in late-stage disease (13, 14), less is known about correlations between molecular properties and disease stage in HIV-2 infection, particularly outside the envelope gene (15, 16). One of the main barriers to a globally protective HIV vaccine is the ability of HIV to evolve rapidly, introducing mutations that abrogate the binding of neutralizing antibodies, rendering vaccine responses ineffective (17). Therefore, a major focus of HIV research has been to understand the factors affecting viral evolution and to identify viral epitopes of high conservation as potential vaccine targets (18).

Due to the relatively low copy number of the small-sized HIV single-stranded RNA (ssRNA) genome (~10,000 bases), target enrichment is normally required prior to sequencing in order to generate sufficient DNA for downstream sequencing applications (19). The most common method of target enrichment is PCR amplification (20). This method has two major drawbacks. The first is the requirement for detailed *a priori* sequence knowledge to inform robust primer design that ensures that the majority of variants in the viral quasispecies are captured (21). Different amplification strategies have shown sensitivities down to 3,000 copies/ml, demonstrating the difficulty of generating robust and high-depth sequence data from patients without detectable plasma viremia (22, 23). However, the sequence database for HIV-2 is significantly smaller than for HIV-1, and a robust and sensitive pan-HIV-2 primer set has yet to be defined and thoroughly evaluated. Mutations in primer binding sites can also reduce binding efficiency and therefore alter the proportion of specific variants in the final pool of amplicons or, in extreme cases, abrogate primer binding completely, resulting in the loss of that variant in the final analysis (24). The second drawback is that PCR is stochastically biased by amplicons from previous cycles acting as templates in the subsequent amplification cycles with the potential to further distort the picture of the viral diversity (25).

Several methods have been proposed to circumvent these problems and reduce the biases introduced into sequencing data through target enrichment. For example, primer identification (ID) allows identification of reads derived from the same viral template through incorporation of a unique 8-mer tag during the reverse transcription of viral RNA (26). Downstream reads can be pooled according to template, and multiple reads from the same template can be used for error correction. A study using primer

TABLE 1 Clinical data related to the analyzed samples and controls^a

Sample	Sampling year	Sex ^b	Country	CD4 (cells/ μ l)	Viral load (cp/ml) ^c	Clinical status at sample date
TD003	2010	F	Guinea-Bissau	560	82,005	Asymptomatic
TD006	2010	F	Guinea-Bissau	1,176	<50	Asymptomatic
TD013	2010	M	Guinea-Bissau	509	1,632	Asymptomatic
TD024	2010	F	Guinea-Bissau	191	10,560	AIDS
TD031	2010	F	Guinea-Bissau	407	107,183	Asymptomatic
TD062	2010	M	Guinea-Bissau	497	139,519	Asymptomatic
CBL20	1988	M	The Gambia	18	NA	AIDS
ROD	1985	M	Cape Verde	100	NA	AIDS

^aPatients samples included TD003 to TD062. Controls included CBL20 and ROD.

^bF, female; M, male.

^cAs determined by an in-house RT-PCR assay (66). cp, copies.

ID observed biased diversity estimates between 2- and 100-fold compared to a library generated without any PCR bias correction, highlighting the importance of considering this factor when sequencing a highly diverse population, such as HIV (26). However, primer ID still relies on sufficient *a priori* sequence knowledge to allow robust primer design, and the incorporation of the barcode into the 3' end of the cDNA molecule means that it is not applicable to library preparation techniques involving random fragmentation of the target, such as those employed when using Illumina platforms.

Shotgun RNA sequencing (RNA-Seq) has been demonstrated as a powerful tool for the study of RNA viruses (27). Library preparation is performed using random-hexamer priming of the total RNA in a sample, negating the need for sequence-specific target enrichment (28). This is particularly desirable for HIV-2, for which the sequence data available are significantly limited compared with those for HIV-1. Few studies have applied RNA-Seq to human RNA viruses. For example, Ninomiya et al. applied RNA-Seq to plasma samples taken from two chronically hepatitis C virus (HCV)-infected patients and demonstrated nearly full-length genome sequences with a mean depth of coverage between 50 \times to 70 \times for the two patients (29). In another study, Batty et al. further expanded this method, presenting a high-throughput method for norovirus sequencing allowing 77 fecal samples to be sequenced, with a mean depth of coverage of 100 \times and a success rate of more than 99% (30). The authors compared this with a PCR amplification strategy and found that the success rate for whole-genome amplification using PCR was 29%. This represents a significant decrease in the performance compared to that of RNA-Seq. RNA-Seq has also been used for the discovery of two novel SIVs, demonstrating the power of this method of sequencing without prior sequence information in viral discovery (31).

In the present study, we applied RNA-Seq library preparation methods to both patient plasma samples taken from a rural West African community cohort and cultured lab-adapted HIV-2 reference strains. We show that RNA-Seq followed by *de novo* assembly is a feasible and powerful approach when applied to HIV-2 samples with viral loads of at least 5,280 copies/ml. In addition, we demonstrate that RNA-Seq represents a novel, low-bias method of HIV-2 sequencing. Finally, we computed estimates of nucleotide diversity for each gene of HIV-2 on both the intra- and interhost levels. These analyses indicated consistently low estimates of diversity in the accessory gene *vpx* within hosts, highlighting the importance of this HIV-2-specific gene in successful HIV-2 infection.

RESULTS

Patient and sample characteristics. Samples from a panel of six members of the Caió HIV-2 community cohort (TD003, TD006, TD013, TD024, TD031, and TD062), whose plasma viral loads represented the broad spectrum seen in natural HIV-2 infection, as well as cultures of two lab-adapted HIV-2 strains (HIV-2 ROD and HIV-2 CBL20) were subjected to standard RNA-Seq library preparation (Table 1).

Assessment of RNA-Seq using HIV-2 ROD. First, we assessed the performance of RNA-Seq using the well-characterized reference strain HIV-2 ROD. Following initial

TABLE 2 Summary of read mapping to sample-specific reference sequences

Sample ID	Bowtie2		BWA-SW		GSNAP		NovoAlign	
	Mean depth	Reads aligning	Mean depth	Reads aligning	Mean depth	Reads aligning	Mean depth	Reads aligning
TD024	28.53×	3,709	31.90×	3,988	27.69×	3,426	27.79×	3,463
TD031	62.33×	7,658	67.23×	8,044	60.33×	7,172	60.50×	7,267
TD062	50.01×	6,617	59.61×	7,468	45.92×	5,658	46.64×	5,751
CBL20	5,502×	539,906	4,734×	412,557	6,451×	618,464	5,156×	432,538
ROD	1,924×	165,506	1,794×	152,105	2,146×	176,885	1,862×	155,696

quality control and removal of low-quality reads and adaptor contamination, reads were assessed for the presence of biased random-hexamer priming. The random-hexamer analysis indicated that a random-hexamer bias affected the first 13 bp of the read. The remaining high-quality reads were assembled to the HIV-2 ROD reference genome sequence (accession number [BD413542](#)). The mean depth of coverage over the whole genome was around 2,000× for all alignment tools, with GSNAP having the highest mean depth (Table 2). All four alignment tools produced a slightly positive GC bias, and more GC-rich regions tended to have higher coverage. The slopes were very similar (0.79 to 0.91), implying that the assembly algorithm used does not affect the GC bias. In order to assess how divergent the HIV-2 ROD that was propagated for the present study was from the published reference sequence, polymorphisms that were fixed at a frequency of >95% in the sample population were annotated as single nucleotide polymorphisms (SNPs) using VarScan (Fig. 1). The BWA-SW build was used for this analysis, as it agreed with the majority consensus at each site of conflict. All genes except *vif* had SNPs (in total 70 SNPs), and the majority of SNPs were seen in *gag*, *pol*, and *nef*. However, when corrected for gene length, *nef* showed the greatest contribution to divergence from the reference genome.

De novo genome assembly and factors influencing RNA-Seq success rate. After showing that our RNA-Seq approach could be used for whole-genome sequencing of a high-copy number and lab-adapted HIV-2 strain, we assessed the feasibility of using RNA-Seq to generate whole-genome sequences directly from primary patient blood plasma samples (Table 1). Clinical blood plasma samples often contain significant amounts of human RNA, making it challenging to perform *de novo* assembly of minority species (such as HIV-2). VICUNA is designed to target populations with high mutation rates and map minority variants into a single consensus sequence and is therefore particularly suitable for HIV-2, considering the few publicly available HIV-2 whole-genome sequences. In addition, since HIV-2 blood plasma samples usually have significantly lower viral copy numbers than propagated virus isolates, we included a lab-adapted HIV-2 strain derived from a Gambian subject (CBL20), for which the whole-genome sequence is unknown, as a high-viremia control (Table 1).

In total, whole-genome assembly was successful for three of the six patient samples (TD024, TD031, and TD062) and the control (CBL20). Successful patient samples showed complete capture of the protein-coding region of HIV-2 and merged contigs ranged from 9,397 to 9776 bp in length. A merged contig spanning the complete coding region was also assembled for CBL20, demonstrating the applicability of RNA-Seq to both *in vitro* and *ex vivo* samples. Although we are unable to provide a formal cutoff value, these results suggest that a viral load of at least 5,000 copies/ml is needed for successful sequencing, with an expectation of at least 0.001% HIV-derived RNA (Table 3). When these limits are considered, the success rate was 75%.

In order to assess how well *de novo* assembly using VICUNA had captured the HIV-2 genome, consensus sequences were aligned to the commonly used HIV-2 group A reference sequence UC2 (accession number [U38293](#)) and annotated according to homology (Table S1). In patient sample TD031, the first 177 bp of the *gag* leader sequence was missing, whereas the long terminal repeat (LTR) region lacked coverage for patient samples TD024 and TD062 and the reference strain CBL20.

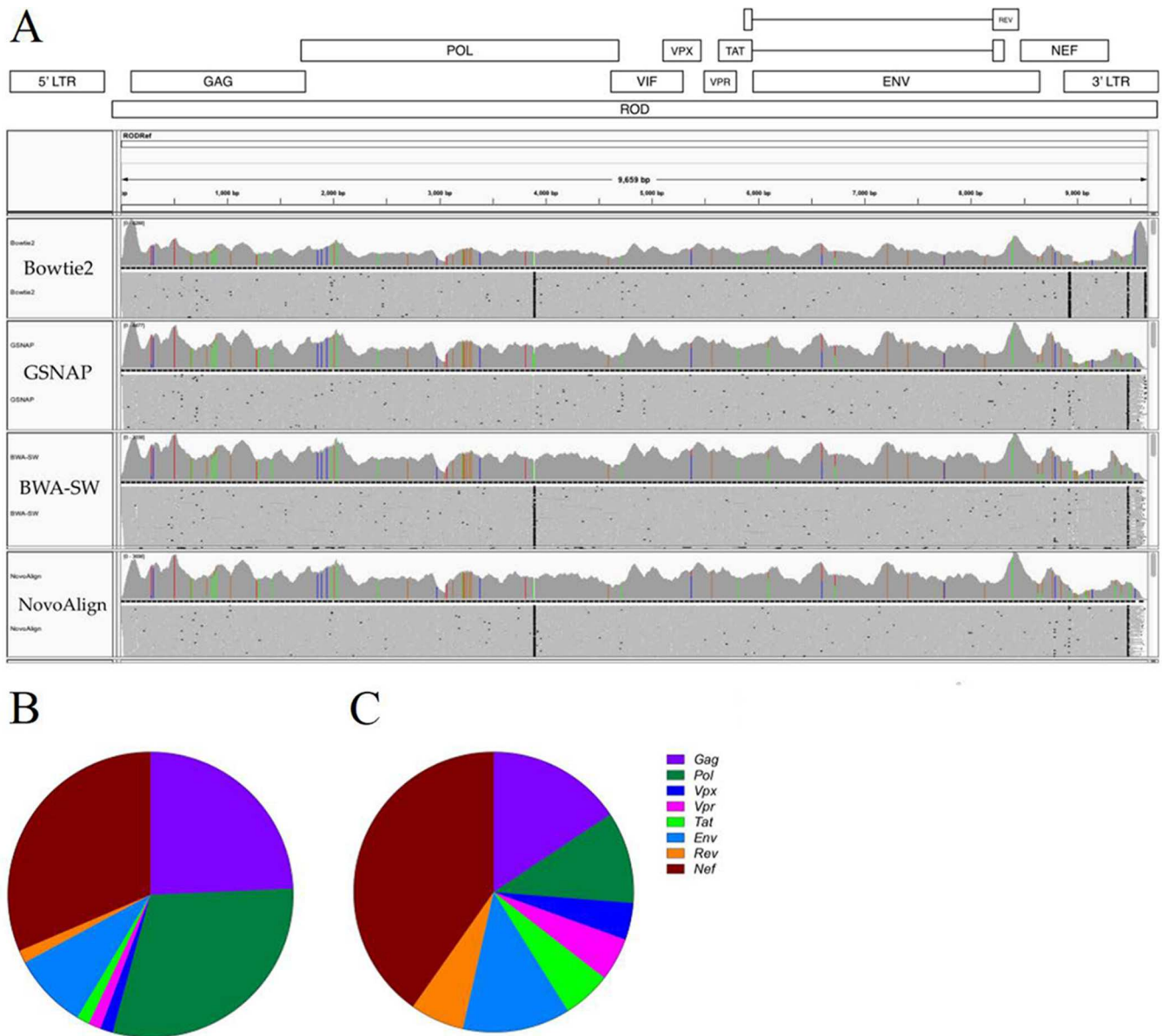


FIG 1 Divergence *in vitro* of the lab-adapted HIV-2 isolate HIV-2 ROD. Assembled reads were visualized in Integrative Genomics Viewer (54) and mismatched sites were colored (A). Sites of conservation with the published reference sequence are shown in gray. Single nucleotide polymorphisms (SNPs) were defined as fixed at a frequency of >95%, and the total number of SNPs in each gene was calculated (B). In order to allow for varying gene lengths, the frequency of SNPs in each gene was also calculated (C).

Phylogenetic analysis of *de novo* genome sequences. A BLAST analysis of the newly generated sequences indicated highest similarity with HIV-2 group A sequences. Bayesian phylogenetic analysis with the 20 publicly available HIV-2 group A near full-genome sequences confirmed this (Fig. 2; Table S1). This analysis also showed that the newly generated sequences were clearly distinguishable from existing reference sequences. As expected, the HIV-2 ROD sequence generated in the present study and the published reference sequence were closely related and clustered together, with a posterior probability of 1.

Read remapping to the patient-specific consensus whole-genome sequences. In contrast to resequencing projects, in which a high depth of coverage is required for error correction, deep sequencing of pathogen populations uses high depth of coverage to gain a picture of the diversity in the population as a whole (27). Following *de novo* assembly of a patient-specific consensus genome sequence, we assessed the

TABLE 3 Samples included in the present study and *de novo* assembly statistics

Sample ID	No. of viral copies ^a	Total RNA (ng) ^b	Predicted HIV RNA (%) ^c	No. of reads aligning to viral reference	Genome covered by all contigs (%)	No. of genes intact	Merged contig length (bp)
TD003	41,002	8.70	0.0023	0	0	0	0
TD006	<50	7.65	0.0000033	0	0	0	0
TD013	816	34.00	0.000012	0	0	0	0
TD024	5,280	2.30	0.0011	4,998	93	9	9,531
TD031	53,591	3.10	0.0087	9,065	90	9	9,397
TD062	69,759	2.85	0.012	13,304	87	9	9,776
CBL20	>10,000,000	9.65	>20	930,072	87	9	9,885

^aAbsolute viral input estimated from viral load.

^bTotal RNA input used for library preparation.

^cEstimated using a viral genome length of 10 kb and absolute viral input.

performance of four commonly used alignment tools when remapping reads to the patient-specific consensus (Table 2). Read remapping was performed using the total reads without prior HIV-2 enrichment or digital subtraction of human sequences to allow an assessment of how these tools perform in the context of a high level of contamination. This is likely to be a factor of all pathogen sequencing strategies employing RNA-Seq. Mean depth and range of coverage were compared for each aligner (Fig. 3). These results show consistent performance of the four aligners, with mean depths of coverage ranging from 28× to 67× for the three patient samples. This range is in line with previous RNA-Seq studies, showing that RNA-Seq is a feasible tool for generating HIV-2 whole-genome sequences. Additionally, the high similarity indicates that read mapping is robust and repeatable irrespective of which alignment

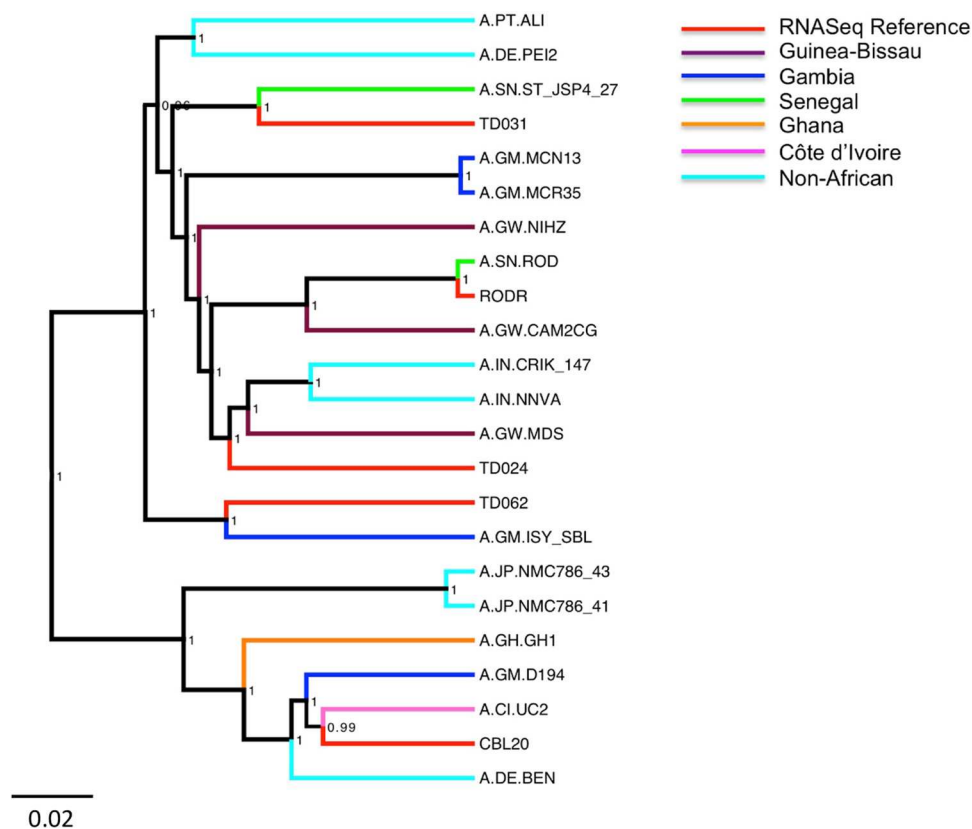


FIG 2 Bayesian phylogeny of HIV-2 genome sequences generated in the present study. Eighteen whole-genome HIV-2 group A sequences were included as a reference set (Table S1). Reference sequences are colored according to country of origin, and sequences generated in the present study are shown in red. Bayesian posterior probabilities are included on the corresponding nodes, and the scale bar represents the number of nucleotide substitutions per site.

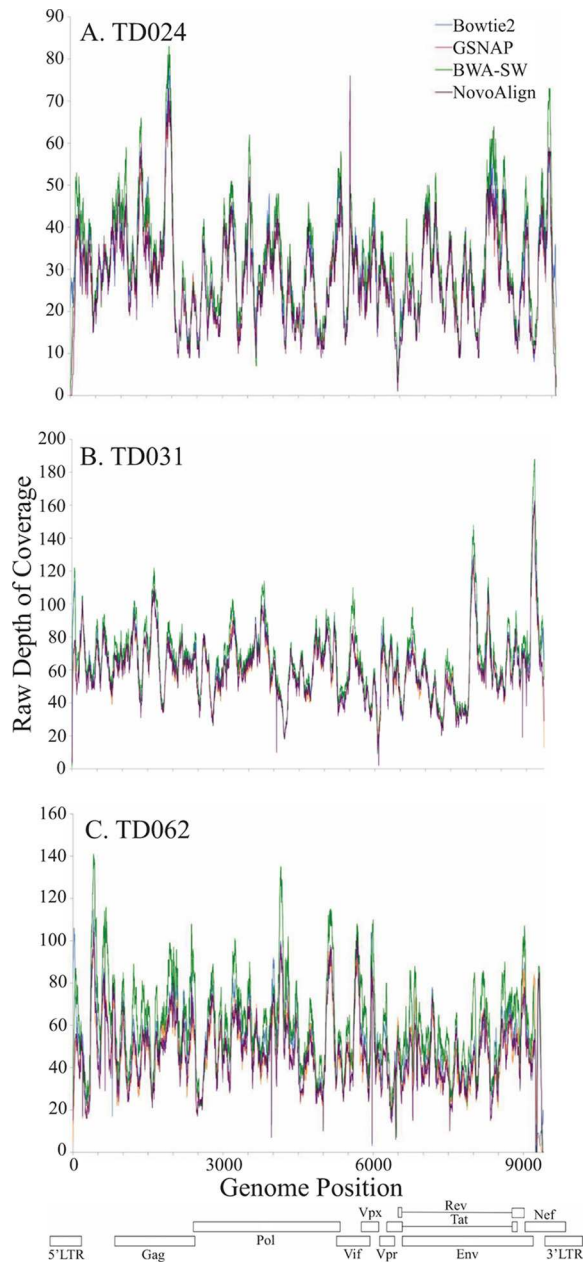


FIG 3 Depth of coverage with the four different aligners. Depth of coverage for each locus was plotted for TD024 (A), TD031 (B), and TD062 (C). Open rectangles represent the locations of HIV-2 genes, and the position of the longest merged contig is also shown for each sample. Coverage plots are shown for each of the four aligners, Bowtie2, GSNAP, BWA-SW, and NovoAlign. Coverage was plotted as raw depth, showing the number of reads mapping to each locus.

algorithm is employed following *de novo* assembly of patient-specific consensus sequences. Error rates were consistent over cycles, and there was no evidence of a drop in accuracy over the length of the read (Fig. 4).

Assessment of the random-hexamer bias. A commonly recognized bias that is specific to RNA-Seq protocols is the random-hexamer bias (32, 33). Hypothetical differential binding affinities between different random hexamers result in biased nucleotide composition at the 3' ends of the reads, normally spanning 7 to 13 bp. In line with previous studies and our initial assessment of RNA-Seq using HIV-2 ROD, the random-hexamer analysis indicated that a random hexamer bias affected the first 13 bp of the read (Fig. S1). The pattern of the bias was similar in all three patient samples and

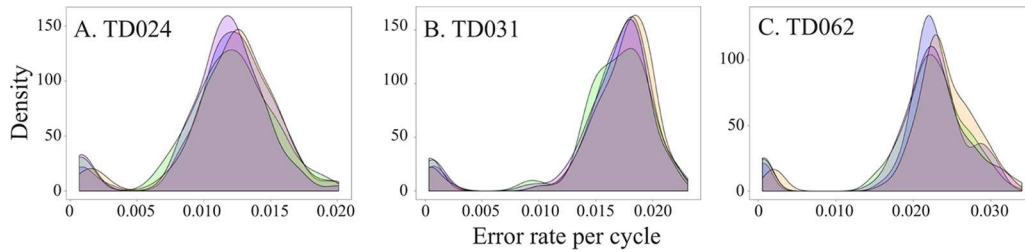


FIG 4 Error rate per sequencing cycle. Error rates were estimated for each sample using the GATK package using the number of mismatches seen during each cycle and the recalibrated quality score of each position. This gave an estimation of error rate per cycle for each of the 100 sequencing cycles for the first read set of each pair. Error rates per cycle were plotted for samples TD024 (A), TD031 (B), and TD062 (C) for each aligner. Mean error rate per cycle was plotted against the probability density for all 100 cycles. Generally, error rates were consistent over cycles, and there was no evidence of a drop in accuracy over the length of the read. Estimations of mean error rates were a little on the high side (TD024, 1.3%; TD031, 1.7%; TD062, 2.1%), likely caused by the underlying variability of HIV-2. GATK was not able to distinguish between low-frequency variants in the viral population and sequencing errors. However, predicted mean error rates are in line with what would be expected and are informative when choosing a cutoff frequency for reliable SNP calling.

the control (CBL20), suggesting that there may be preferential binding to the same motifs in all samples. This biased read composition can be attributed to random-hexamer bias rather than low-quality sequencing at the end of the reads, as the median Q-score was constant over the length of the read. The effect of the bias did not extend past the first 13 bp of each read, and the nucleotide composition stabilized after this point. A correction was not applied to account for the biased nucleotide composition of the first 13 bp, as removal of these positions does not remove the effects of this bias seen in downstream analyses.

Quantification of the GC bias and depth of coverage as a function of genomic context. Depth of coverage in samples sequenced using Illumina short-read chemistry can be affected by the local GC content of the genome (34). We assessed the effect of local GC content on depth of coverage using a custom script which took a sliding window of 50 bp, with a step size of 20 bp, and calculated percent GC and mean depth of coverage in each window. The extent of the GC bias was quantified using the slope of the linear regression line, and the bias was assessed for each aligner individually (Table 4). To further compare the different aligners, the mean depth of coverage was normalized in each window using the genome-wide mean depth of coverage (Fig. 5). All assemblies showed a slight, positive GC bias, suggesting that GC-rich regions had a depth of coverage that was higher than the mean. For patient samples TD024 and TD031, the magnitudes of the slope were similar for all four aligners, suggesting a constant effect when different assembly algorithms were employed. In contrast, sample TD062 showed more fluctuation between aligners. However, the magnitude of the bias was lowest for this patient, suggesting that the overall effect of the GC bias would be reduced, in spite of the fluctuations. Hence, a positive GC bias in HIV-2 samples sequenced using RNA-Seq may confer variability in depth of coverage over the genome. However, the magnitude of the bias was in line with previous studies and did not show a loss of coverage of any genomic regions due to GC bias (34).

In order to assess whether genomic context could affect depth of coverage, the HIV-2 genome was partitioned according to gene and mean depth of coverage was compared for each gene individually. The effect of genomic context on depth of

TABLE 4 Summary statistics for the GC bias present in assembled reads^a

Sample ID	Bowtie2		BWA-SW		GSNAP		NovoAlign	
	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept
TD024	1.80	0.17	1.95	0.10	1.74	0.19	1.79	0.15
TD031	1.04	0.54	1.17	0.47	0.97	0.56	1.02	0.54
TD062	0.66	0.70	0.88	0.57	0.35	0.81	0.56	0.71

^aEstimated by fitting a linear regression to the mean values in each sliding window.

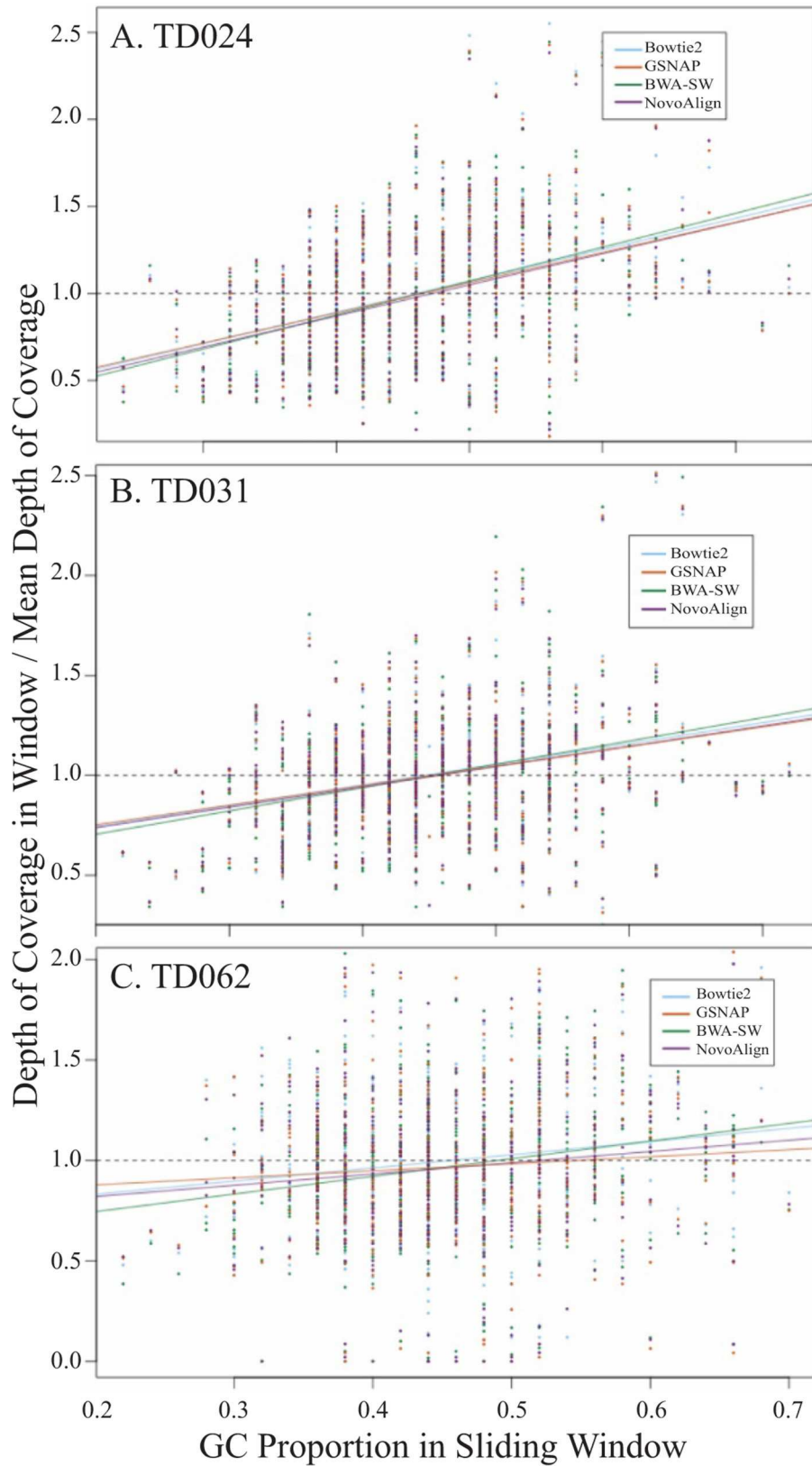


FIG 5 Scatter plots showing the GC bias in assembled reads. GC proportion and normalized depth of coverage in each window were plotted for each aligner individually and then grouped by patient sample. Plots are shown for patients TD024 (A), TD031 (B), and TD062 (C). A linear regression was fitted to assess the magnitude and direction of the bias. Regression lines are colored by aligner. The dashed line indicates the expected regression in the absence of any positive or negative GC bias.

coverage was visualized by plotting mean depth of coverage as a function of GC content for each gene (Fig. 6). All aligners showed a similar pattern of coverage and no consistent loss of coverage in any genomic region.

No general trends in molecular properties between the analyzed HIV-2 strains or correlations with clinical stage. To characterize the molecular properties of the newly generated sequences and to put them in a broader perspective, we performed an in-depth analysis of these and the 20 selected and publicly available HIV-2 group A sequences. Associations between molecular and biological properties were assessed by available clinical and epidemiological data (Table S1; Fig. S2). All analyses were performed per HIV-2 gene. In the data set there were two occasions of duplicate origin, i.e., two sequences that had been generated from the same original patient sample (Table S1; Fig. S2, RODR and A.SN.ROD plus A.JP.NMC786_41 and A.JP.NMC786_41). These were counted only once when assessing associations between molecular and biological properties between sequences collected during the asymptomatic versus AIDS stage of disease. No significant differences in sequence length, net charge, total charge, or number of potential N-linked glycosylation sites (PNGS), in any of the nine HIV-2 genes, were found between sequences from asymptomatic patients ($n = 6$) and AIDS stage patients ($n = 12$) (Table S1; Fig. S2). Prediction of coreceptor tropism based on the *env* gp120 V3 region indicated that 50% (3 of 6) and 58% (7 of 12) of the participants had CXCR4-using viruses in the asymptomatic and AIDS stages, respectively ($P = 1.00$, two-tailed Fisher's exact test) (Table 5). Furthermore, we found no diagnostic motifs or amino acids between asymptomatic and AIDS stage patients in any of the nine HIV-2 genes (Table S1; Fig. S2).

Genome-wide estimation of genetic diversity in HIV-2 in the context of low-bias sequencing. To determine how the diversity varies over the HIV-2 genome, we estimated nucleotide pairwise diversity from assembled reads (Bowtie2 assembly) using a custom script. Raw estimates of diversity of the whole genome were 0.0010 substitution/site for TD024, 0.0007 substitution/site for TD031, and 0.0014 substitution/site for TD062. In comparison, the raw estimates for the *env* gene were 0.0013 substitution/site for TD024, 0.0008 substitution/site for TD031, and 0.0020 substitution/site for TD062. To compare the relative genetic diversity between the patients, we normalized the raw estimates using the genome average (Fig. 7). Overall, our analysis showed similar results between patients, with the highest level of within-host diversity seen in *env* for all three patients, whereas the lowest diversity was seen in the *vpx* and *rev* genes. Interestingly, the diversity in *pol* seemed to be higher than for the genes *gag*, *vpx*, *tat*, *rev*, and *vif* for all three patients.

To compare the above-described results of intrahost viral diversity in different HIV-2 genes with viral diversity in different HIV-2 genes between hosts, we performed a phylogenetic bootstrap analysis of our newly generated whole-genome sequences and the reference sequences. This analysis showed that, similar to the intrahost viral diversity above, the *env* gene was the most diverse gene, followed by the *nef* gene. However, in contrast to the intrahost analysis, this analysis indicated that *pol* was the second least diverse gene between hosts (only *vif* was less diverse [Fig. 8]).

DISCUSSION

Deep sequencing of HIV offers unparalleled opportunities to gain a high-resolution picture of the nature and diversity of the viral quasispecies in a single patient. Our study presents a novel pan-HIV-2 whole-genome amplification strategy using RNA-Seq, allowing the entire protein-coding region of HIV-2 to be sequenced without the need for detailed *a priori* sequence knowledge. We show a broad applicability of this method, presenting data from both lab-adapted isolates and patient plasma samples. To our knowledge, only one previous study has used a next-generation sequencing approach to determine the full genome of HIV-2 (35). However, we used HIV-2 isolates propagated in cell culture prior to library preparation and aligned the generated sequence reads to a common reference strain (HIV-2 BEN). We analyzed patient samples and were able to successfully sequence samples of viral loads down to 5,280 copies/ml and an

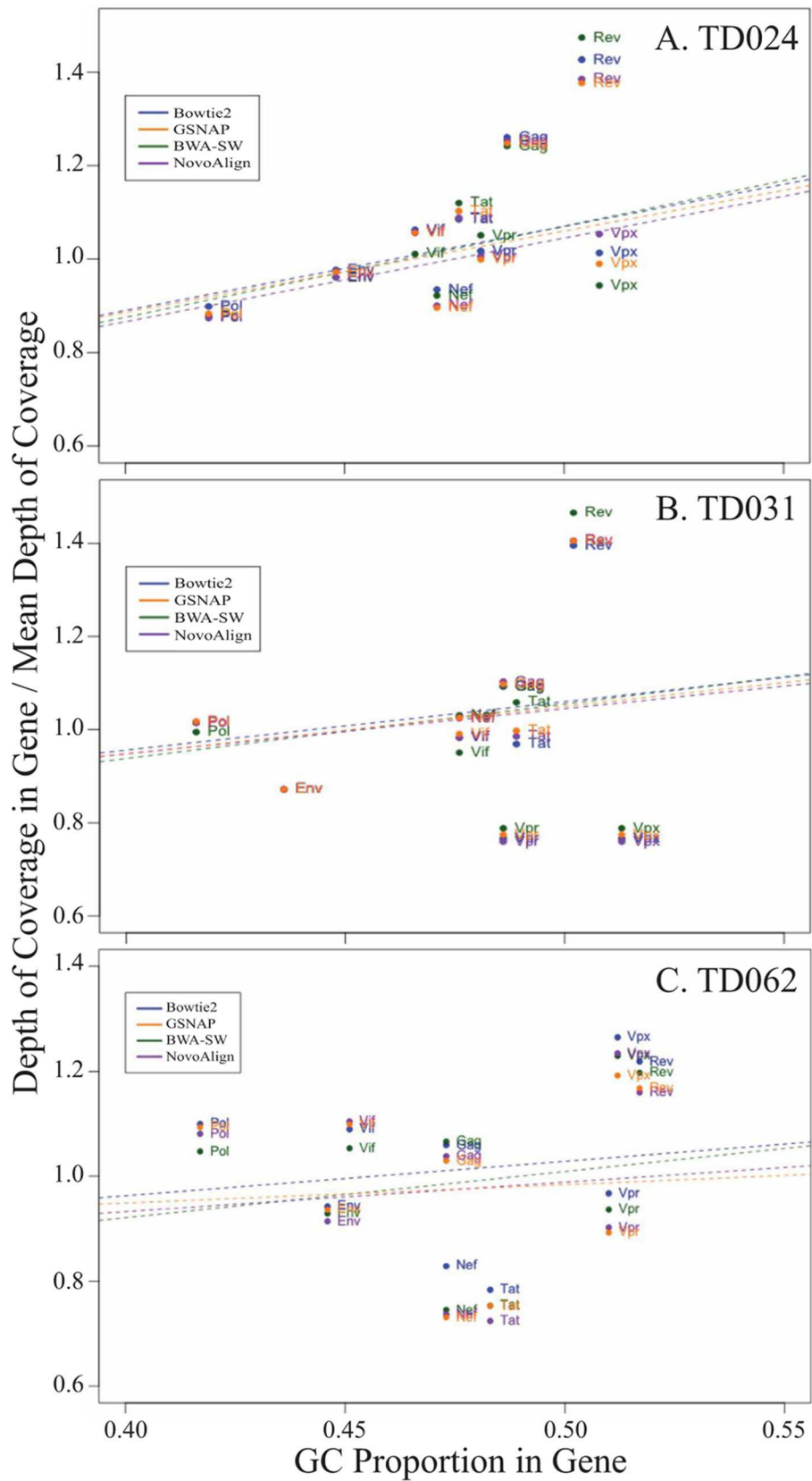


FIG 6 Depth of coverage as a function of genomic context. The depth of coverage was calculated individually for each gene of the HIV-2 genome for TD024 (A), TD031 (B), and TD062 (C). Depth of coverage was colored by aligner and plotted against the mean GC content of the gene. The predicted GC bias is represented by the dashed line.

TABLE 5 V3 characteristics and coreceptor tropism^a

Sequence name	Length (aa)	Net charge ^b	Total charge ^c	PNGS ^d	L18Z ^e	V19K/R ^e	Net> +6 ^e	Ins24 ^e	Predicted tropism ^f
TD024	34	5	5	1	N	N	N	N	R5
TD031	34	5	5	1	N	N	N	N	R5
TD062	34	5	5	1	N	N	N	N	R5
CBL20	34	6	6	1	N	N	N	N	R5
RODR	36	7	7	1	N	N	Y	N	R5X4
A.CI.UC2	35	7	7	1	N	N	Y	Y	R5X4
A.DE.BEN	34	5	5	1	N	N	N	N	R5
A.DE.PE12	35	7	9	1	N	Y	Y	Y	R5X4
A.GH.GH1	34	5	5	1	N	N	N	N	R5
A.GM.D194	35	8	8	1	N	N	Y	Y	R5X4
A.GM.ISY_SBL	35	8	10	1	N	N	Y	Y	R5X4
A.GM.MCN13	35	7	7	1	N	N	Y	Y	R5X4
A.GM.MCR35	35	7	7	1	N	N	Y	Y	R5X4
A.GW.CAM2CG	34	6	6	1	N	N	N	N	R5
A.GW.MDS	34	6	6	1	N	N	N	N	R5
A.GW.NIHZ	35	7	7	1	N	N	Y	Y	R5X4
A.IN.CR1K_147	35	7	7	1	N	N	Y	Y	R5X4
A.IN.NNVA	34	5	5	1	Y	N	N	N	R5X4
A.JP.NMC786_41	35	6	6	1	N	N	N	Y	R5X4
A.JP.NMC786_43	35	6	6	1	N	N	N	Y	R5X4
A.PT.ALI	34	6	6	1	N	N	N	N	R5
A.SN.ROD	36	7	7	1	N	N	Y	N	R5X4
A.SN.ST_JSP4_27	34	6	6	1	N	N	N	N	R5

^aCoreceptor tropism was assessed as described in reference 59.

^bNet charge of sequences was determined based on each lysine and arginine contributing +1 and each aspartic acid and glutamic acid contributing -1.

^cTotal counts of amino acids were also assessed as described previously (58).

^dNumber of potential N-linked glycosylation sites (PNGS) as defined in N-GLYCOSITE (37).

^eN, no; Y, yes.

^fR5, CCR5-using virus; R5X4, CCR5- and CXCR4-using virus.

expectation of at least 0.001% HIV-2 RNA in the sample. When these conditions were fulfilled, the success rate was 75%, which is lower than previously reported by Batty et al. when applying RNA-Seq to norovirus. However, the lower HIV-2 plasma viral loads of the patient samples used in the present study readily explain this reduced success rate. A putative cutoff of approximately 5,000 copies/ml restricts this method to viremic HIV-2 patients, and it is possible that an alternative approach would be needed for

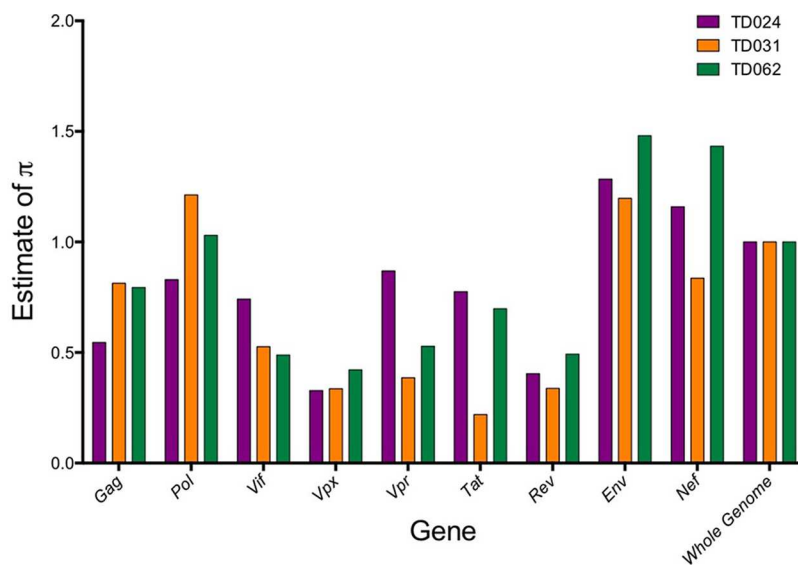


FIG 7 Nucleotide diversity plot. Nucleotide site diversity estimates per gene, normalized to the whole-genome estimate. Diversity was estimated for samples TD024, TD031, and TD062. Calculation of the diversity relative to the whole-genome estimate was performed to allow a comparison between patients.

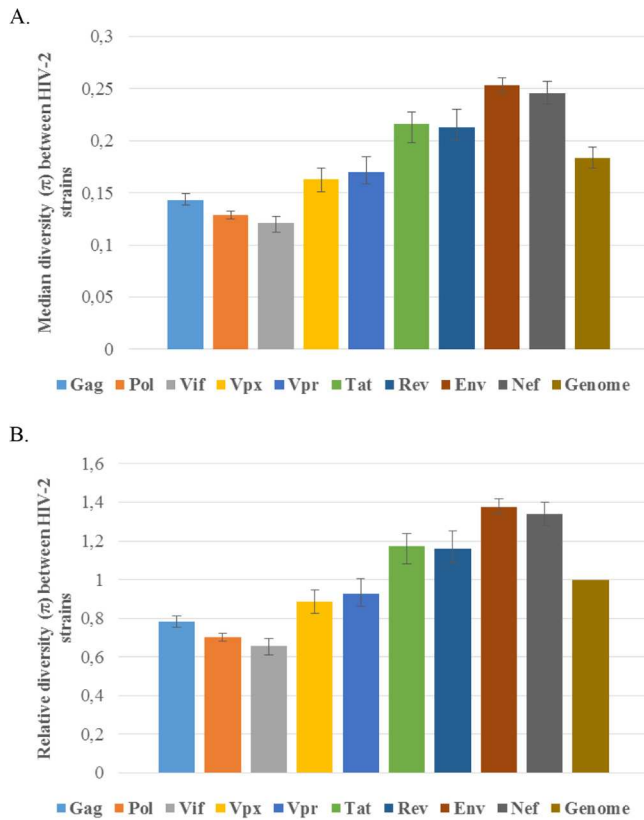


FIG 8 Median diversity between the analyzed HIV-2 strains. The median diversity was determined phylogenetically by analysis of 200 gene-specific maximum likelihood bootstrap replicates. The error bars show the interquartile ranges of the diversity estimates. (A) Diversity estimates in substitutions per site. (B) Diversity estimates relative to the average median estimate of all analyzed genes (i.e., genome).

samples with lower viral loads. However, we anticipate that RNA-Seq could also be successfully applied to samples taken from untreated HIV-1 patients, where the typical viral load is 10 to 1,000 times higher than for HIV-2.

While RNA-Seq allows whole-genome sequencing of HIV-2 without the need for detailed prior sequence knowledge, the lack of sequence-specific target amplification also leads to a reduction in the use of PCR amplification and its resulting biases, thereby generating sequence data that are more representative of the true population frequencies. In this study, we aimed to quantify the other biases known to be associated with RNA-Seq. We found evidence of a moderate positive GC bias which varied between samples but was consistent when different aligners were used. We also found evidence of a biased nucleotide composition in the first 13 bases of the reads, suggesting the presence of non-random random hexamer priming. Although these biases could be responsible for the fluctuations in coverage over the genome, we observed no correlation between genomic location and depth of coverage. This suggests that these fluctuations were randomly distributed and not due to the various diversities seen in different functional genomic sites.

While patient consensus sequences contained all nine genes of HIV-2 in intact reading frames, there was some variability in the assembly of the 3' and 5' LTR and the *gag* leader sequence. In patient sample TD031, the loss of the first 177 bases of the *gag* leader sequence could be attributed to the failure of the RNA-Seq library preparation method to capture this region. The initial fragmentation step in library preparation can lead to the loss of distal regions of the RNA molecule, and this is the most probable cause of the lack of coverage in this genomic region. For patient samples TD024 and TD062 and the reference strain CBL20, the lack of coverage was probably due to the nature of the LTRs in HIV-2. The 5' and 3' LTR regions only exist as true 990-base

repeats in the proviral form of the virus, whereas in the RNA genome, the 5' LTR comprises the R and U5 regions and the 3' LTR is composed of the R and U3 regions (36). The sequence alignment used during assembly contained HIV-2 sequences from both cDNA and RNA HIV-2 genomes, so assembly was conducted using "complete" LTRs, both containing U5, R, and U3 (37). Ambiguous read mapping is normally resolved by using the location of the read mate to provide information on the most likely coordinates. In the present study, the insert size (250 to 350 bases) and the nature of the LTRs made mapping reads for the R region problematic, as the read mate will also fall in the LTR. Therefore, it was not possible to resolve the correct orientation of the reads, resulting in the loss of coverage of one LTR.

The ability to sequence the whole HIV-2 genome in a single experiment allowed us to compare pairwise nucleotide diversities between the different HIV-2 genes. A study comparing outcomes between HIV-1- and HIV-2-coinfected and HIV-1-monoinfected subjects followed from early infection showed that the extent of HIV-1 genetic diversity strongly correlated with time to AIDS (38). Much less is known about genetic diversity in HIV-2 infection. In our small study of three subjects, all of whom presented with raised viral load, the patterns of within-host diversity were similar in all three. The highest diversity was seen in *env*, which is in keeping with previous observations in HIV-1 infection. Diversity in partial fragments of HIV-2 *env* has previously been estimated through different approaches, and although on the lower side, our estimated intrahost *env* diversities were in the same range as those from previous studies, which used molecular cloning for sequence generation (15, 39, 40). Similarly, a high level of *nef* diversity was seen in all three patients. In HIV-1 infection, *pol* is thought to be highly conserved for functional reasons and therefore typically shows a relatively lower diversity than, for example, the *env* gene (41). In contrast, we observed a high level of within-host HIV-2 *pol* diversity in all three subjects studied in this investigation. Interestingly, a recent study has shown a high level of within-host diversity in *pol* following vertical HIV-1 transmission (42), although vertical transmission of HIV-2 is uncommon and is unlikely to be implicated in our study subjects. There are some potential caveats of our intrahost diversity analysis. (i) HIV-2 diversity has been reported to increase over the course of infection (15, 39). It is possible that parameters such as the duration of infection or the mode of transmission influenced the diversity levels in our study, but these parameters are not known for the three study subjects. (ii) For some single nucleotides over the genome, the coverage was less than 20 sequence reads, and from a sample perspective, the depth of coverage was positively correlated with the viral copy number. On the one hand, low coverage may underestimate the true genetic diversity. On the other hand, some regions of the genome are evolutionarily conserved, and only a limited number of virus variants can theoretically coexist in a sample taken from a subject with a low viral load.

vpx is an HIV-2/SIVmm specific accessory gene that is entirely absent from the HIV-1/SIVcpz lineage. The main role of *vpx* is antagonism of the host restriction factor SAMHD1, which blocks reverse transcription of viral RNA in slowly dividing cells, such as macrophages and resting CD4⁺ T cells (43). Our observation of a consistently low level of *vpx* diversity may indicate a high level of conservation in *vpx*, suggesting that *vpx* has a critical role in HIV-2 pathogenesis. The implications of *vpx* diversity in HIV-2 infection are not yet clearly defined, but a recent study by Yu et al. identified an SNP in a *vpx* allele derived from a viremic patient that totally abrogated the ability of *vpx* to promote SAMHD1 degradation *in vitro* (44).

In conclusion, we show that RNA-Seq library preparation methods can be applied to HIV-2 blood plasma samples. Resulting *de novo* genome assemblies captured the entire coding region of HIV-2 in intact open reading frames and read remapping allowed us to demonstrate the importance of a two-step analysis pipeline. In the context of a highly diverse retrovirus, such as HIV-2, the selection or generation of an appropriate reference sequence is a critical first step, allowing robust and repeatable downstream read mapping. We also demonstrated a low level of GC and random-hexamer bias and, in the absence of sequence-specific target amplification, showed that RNA-Seq offers a

method of whole-genome HIV-2 sequencing in a low bias context. However, some challenges in RNA-Seq remain. For example, although the sequencing costs have fallen dramatically in recent years, RNA-Seq is still expensive and costs continue to be a barrier to an even more widespread adoption. In the present study, we multiplexed six patient samples using the Illumina HiSeq in order to reach a mean depth of coverage of up to $67\times$. Although this coverage is more than sufficient for consensus sequence calling, it may be too low if the primary goal is to determine minority variants (at least in samples with high viral loads). Hence, the importance of developing novel and low-bias HIV sequencing protocols cannot be overstated, as the ability to gain a complete and accurate picture of HIV genetic diversity is critical to the development of globally effective and preventative HIV vaccines.

MATERIALS AND METHODS

Patient sample collection. All patient samples used in the present study were collected from members of the Caió community cohort who had provided written and informed consent. Samples were collected prior to the start of the present study. Plasma was separated from whole blood through centrifugation ($5,000 \times g$, 5 min, and 4°C) and filtration ($0.45\text{-}\mu\text{m}$ filter; Millipore, Billerica, MA). Plasma samples were stored at -80°C before being transported to Oxford, United Kingdom, in a liquid nitrogen dry shipper. None of the samples used had any record of previous freeze-thaw cycles.

In vitro culture of lab-adapted HIV-2 reference strains. The lab-adapted HIV-2 strains HIV-2 ROD and HIV-2 CBL20 were propagated *in vitro* in the lymphocyte cell line H9, a single cell clone derived from a HUT 78 cell line. Infection of 5×10^6 cells was carried out with $200 \mu\text{l}$ of 9×10^3 50% tissue culture infective doses (TCID_{50})/ml of viral stock. Cells were removed through centrifugation at $250 \times g$ for 10 min, and supernatant was collected on days 3, 5, 7, 9, 11, 13, and 15. HIV-2 concentrations were assayed using a colorimetric reverse transcriptase assay (Roche). For each isolate, the supernatant sample with the highest reverse transcriptase concentration was selected for RNA-Seq.

RNA extraction, RNA quantification, and DNase treatment. Total nucleic acid was extracted directly from $500 \mu\text{l}$ of patient plasma or purified supernatant using the QIAamp UltraSens viral kit (Qiagen). Extraction was performed according to the manufacturer's protocol, with the substitution of carrier RNA with linear acrylamide (Ambion) as the nucleic acid coprecipitant. Final elution was performed in $12 \mu\text{l}$ of H_2O . DNA was removed from the samples through treatment with DNase I (Turbo DNase; Ambion) according to the manufacturer's protocol. RNA concentration was estimated using the QuBit RNA assay (Invitrogen).

Library preparation and sequencing. Sequencing libraries were prepared from $5 \mu\text{l}$ of the eluted RNA using the NEBNext Ultra RNA library prep kit for Illumina (New England BioLabs) according to the manufacturer's protocol. Sequencing libraries were multiplexed and sequenced using the Illumina HiSeq or MiSeq platform (Illumina). Patient samples were multiplexed at 6/lane (HiSeq), generating 2×100 nucleotide paired-end reads, and lab-adapted strains were multiplexed at 2/lane (MiSeq), generating 2×150 nucleotide paired-end reads.

De novo genome assembly and read remapping. Sequence data were analyzed using a custom pipeline. Reads were trimmed using Sickle, stipulating a median Q-score of >30 and a read length of >40 bp (45). *De novo* genome assembly was performed using VICUNA (46), with the addition of the optional contamination removal step. During contamination removal, HIV-2-derived reads were identified through similarity to a multiple-sequence alignment containing a set of 18 publicly available HIV-2 group A sequence data (Table S1). Overlapping contiguous sequences generated by VICUNA were assembled into whole-genome sequences using the map-to-reference feature in Geneious v6.1.6 (47) and manually inspected to derive a whole-genome consensus sequence. Consensus genome sequences were manually inspected to ensure that they contained intact open reading frames. Reads were remapped to the consensus genome sequence using Bowtie2 (48), BWA-SW (49), GSNAP (50), and NovoAlign (51) for each sample. Files containing assembled reads were manipulated using the SAMtools package (52), and downstream statistical analyses and data visualizations were performed using R (53) and the Interactive Genome Viewer (54). Error rates were estimated using the ErrorRatePerCycle feature of GATK (55).

Quantification of biases. Random-hexamer bias was assessed through visualization of the base composition of reads using FASTQC (56). GC bias was quantified using a custom Python script that scanned the genome using a 50-bp sliding window with a step size of 20 bp. Mean GC content and mean depth of coverage were computed for each window, and GC bias was assessed by fitting a linear regression in R (57).

Analysis of molecular properties. Analyses of molecular properties were performed using an in-house Perl script with potential N-linked glycosylation sites (PNGS) as defined in N-GLYCOSITE (37). Net charge of sequences was determined based on each lysine and arginine contributing +1 and each aspartic acid and glutamic acid contributing -1 . Total counts of amino acids were assessed as described previously (58). Coreceptor tropism was predicted using four major determinants of dual/CXCR4 coreceptor use (L18Z, V19K/R, V3 net charge of $>+6$, and insertions at position 24) (59). CXCR4 use was considered when at least one of the criteria was fulfilled. Sample donors were classified as having been sampled during either the asymptomatic or AIDS stage (as defined by clinical assessment at the sample time point).

Phylogenetic analysis. A reference set of 20 HIV-2 group A whole-genome sequences were obtained from the Los Alamos HIV database (Table S1) (37). Reference sequences were aligned with consensus whole-genome sequences using Muscle (60), and the alignment was manually inspected using Geneious v6.1.6. A Bayesian phylogeny was inferred using BEAST v1.8.0 (61), under the general time-reversible model of nucleotide substitution with a proportion of invariant sites and gamma-distributed rate heterogeneity, as determined by jModelTest2 (57). The Markov Chain Monte Carlo algorithm was run using 100,000,000 iterations with samples taken from the posterior distribution every 10,000 generations. Following a burn-in corresponding to 10% of the samples, the resulting maximum clade credibility (MCC) tree was visualized using FigTree v1.4.1 (62).

Estimation of genetic diversity. Mpileup files were generated from assembled reads using the SAMtools package, and variants were called using VarScan (63) with a cutoff frequency of 0.05. Nucleotide pairwise diversity (π) was estimated using the Nei and Li method (64) through a custom Python script, taking depth at each position as a proxy for population size and the product of frequency of alternative variants and depth as the number of pairwise differences between sequences. Estimates of diversity were generated for each individual gene and over the whole genome, and estimates were normalized using the whole-genome average to allow comparison between patients. For comparison, we also calculated diversity at the population level by averaging pairwise phylogenetic tree distances in Garli v2.0 (65). This was done for each gene separately based on 200 maximum likelihood bootstrap replicates as described previously (58).

Statistics. Two-tailed Fisher's exact test was used to assess data (IBM SPSS Statistics for Windows, v23.0; IBM Corp., Armonk, NY).

Ethics. Ethical approval was granted by the Gambian government/MRC joint ethics committee (SCC1204) and the Oxford tropical research ethics committee (170-12).

Accession number(s). Nucleotide sequences were deposited in GenBank under the following accession numbers: [MH681607](https://doi.org/10.1093/ncbi/acc1111) to [MH681611](https://doi.org/10.1093/ncbi/acc1111).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/JVI.00677-18>.

SUPPLEMENTAL FILE 1, PDF file, 1.6 MB.

ACKNOWLEDGMENTS

K.J. was supported by a Wellcome Trust 4-year Ph.D. studentship (grant no. H5RSZMO). J.E. was supported by the Swedish Research Council (350-2012-6628 and 2016-01417) and the Swedish Society of Medical Research (SA-2016).

We thank Shokouh Makvandi-Nejad and Lorna Witty for their input into the project design and Takayuki Chikata and Masafumi Takiguchi for their sequencing support.

We declare no conflicts of interest.

REFERENCES

- de Silva TI, Cotten M, Rowland-Jones SL. 2008. HIV-2: the forgotten AIDS virus. *Trends Microbiol* 16:588–595. <https://doi.org/10.1016/j.tim.2008.09.003>.
- Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pepin J, Posada D, Peeters M, Pybus OG, Lemey P. 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346:56–61. <https://doi.org/10.1126/science.1256739>.
- Berry N, Jaffar S, Schim van der Loeff M, Ariyoshi K, Harding E, N'Gom PT, Dias F, Wilkins A, Ricard D, Aaby P, Tedder R, Whittle H. 2002. Low level viremia and high CD4% predict normal survival in a cohort of HIV type-2-infected villagers. *AIDS Res Hum Retroviruses* 18:1167–1173. <https://doi.org/10.1089/08892220260387904>.
- van der Loeff MF, Larke N, Kaye S, Berry N, Ariyoshi K, Alabi A, van Tienen C, Leligidowicz A, Sarge-Njie R, da Silva Z, Jaye A, Ricard D, Vincent T, Jones SR, Aaby P, Jaffar S, Whittle H. 2010. Undetectable plasma viral load predicts normal survival in HIV-2-infected people in a West African village. *Retrovirology* 7:46. <https://doi.org/10.1186/1742-4690-7-46>.
- Hansmann A, Schim van der Loeff MF, Kaye S, Awasana AA, Sarge-Njie R, O'Donovan D, Ariyoshi K, Alabi A, Milligan P, Whittle HC. 2005. Baseline plasma viral load and CD4 cell percentage predict survival in HIV-1- and HIV-2-infected women in a community-based cohort in The Gambia. *J Acquir Immune Defic Syndr* 38:335–341.
- Popper SJ, Sarr AD, Travers KU, Gueye-Ndiaye A, Mboup S, Essex ME, Kanki PJ. 1999. Lower human immunodeficiency virus (HIV) type 2 viral load reflects the difference in pathogenicity of HIV-1 and HIV-2. *J Infect Dis* 180:1116–1121. <https://doi.org/10.1086/315010>.
- Andersson S, Norrgren H, da Silva Z, Biague A, Bamba S, Kwok S, Christopherson C, Biberfeld G, Albert J. 2000. Plasma viral load in HIV-1 and HIV-2 singly and dually infected individuals in Guinea-Bissau, West Africa: significantly lower plasma virus set point in HIV-2 infection than in HIV-1 infection. *Arch Intern Med* 160:3286–3293. <https://doi.org/10.1001/archinte.160.21.3286>.
- Gottlieb GS, Sow PS, Hawes SE, Ndoeye I, Redman M, Coll-Seck AM, Faye-Niang MA, Diop A, Kuypers JM, Critchlow CW, Respass R, Mullins JI, Kiviat NB. 2002. Equal plasma viral loads predict a similar rate of CD4+ T cell decline in human immunodeficiency virus (HIV) type 1- and HIV-2-infected individuals from Senegal, West Africa. *J Infect Dis* 185:905–914. <https://doi.org/10.1086/339295>.
- Cheng-Mayer C, Seto D, Tateno M, Levy JA. 1988. Biologic features of HIV-1 that correlate with virulence in the host. *Science* 240:80–82. <https://doi.org/10.1126/science.2832945>.
- Fenyo EM, Esbjornsson J, Medstrand P, Jansson M. 2011. Human immunodeficiency virus type 1 biological variation and coreceptor use: from concept to clinical significance. *J Intern Med* 270:520–531. <https://doi.org/10.1111/j.1365-2796.2011.02455.x>.
- Mild M, Gray RR, Kvist A, Lemey P, Goodenow MM, Fenyo EM, Albert J, Salemi M, Esbjornsson J, Medstrand P. 2013. High inpatient HIV-1 evolutionary rate is associated with CCR5-to-CXCR4 coreceptor switch. *Infect Genet Evol* 19:369–377. <https://doi.org/10.1016/j.meegid.2013.05.004>.
- Mild M, Kvist A, Esbjornsson J, Karlsson I, Fenyo EM, Medstrand P. 2010. Differences in molecular evolution between switch (R5 to R5X4/X4-tropic) and non-switch (R5-tropic only) HIV-1 populations during infec-

- tion. *Infect Genet Evol* 10:356–364. <https://doi.org/10.1016/j.meegid.2009.05.003>.
13. Visseaux B, Charpentier C, Rouard C, Fagard C, Glohi D, Tubiana R, Damond F, Brun-Vezinet F, Matheron S, Descamps D, French HIVACCO. 2014. HIV-2 X4 tropism is associated with lower CD4+ cell count in treatment-experienced patients. *AIDS* 28:2160–2162. <https://doi.org/10.1097/QAD.0000000000000373>.
 14. Morner A, Bjorndal A, Albert J, Kewalramani VN, Littman DR, Inoue R, Thorstensson R, Fenyo EM, Bjorling E. 1999. Primary human immunodeficiency virus type 2 (HIV-2) isolates, like HIV-1 isolates, frequently use CCR5 but show promiscuity in coreceptor usage. *J Virol* 73:2343–2349.
 15. MacNeil A, Sankale JL, Meloni ST, Sarr AD, Mboup S, Kanki P. 2007. Long-term inpatient viral evolution during HIV-2 infection. *J Infect Dis* 195:726–733. <https://doi.org/10.1086/511308>.
 16. Shi Y, Brandin E, Vincic E, Jansson M, Blaxhult A, Gyllensten K, Moberg L, Brostrom C, Fenyo EM, Albert J. 2005. Evolution of human immunodeficiency virus type 2 coreceptor usage, autologous neutralization, envelope sequence and glycosylation. *J Gen Virol* 86:3385–3396. <https://doi.org/10.1099/vir.0.81259-0>.
 17. Kawashima Y, Pfaffertott K, Frater J, Matthews P, Payne R, Addo M, Gatanaga H, Fujiwara M, Hachiya A, Koizumi H, Kuse N, Oka S, Duda A, Prendergast A, Crawford H, Leslie A, Brumme Z, Brumme C, Allen T, Brander C, Kaslow R, Tang J, Hunter E, Allen S, Mulenga J, Branch S, Roach T, John M, Mallal S, Ogwu A, Shapiro R, Prado JG, Fidler S, Weber J, Pybus OG, Klenerman P, Ndung'u T, Phillips R, Heckerman D, Harrigan PR, Walker BD, Takiguchi M, Goulder P. 2009. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458:641–645. <https://doi.org/10.1038/nature07746>.
 18. Barouch DH. 2008. Challenges in the development of an HIV-1 vaccine. *Nature* 455:613–619. <https://doi.org/10.1038/nature07352>.
 19. Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A, Robertson DL. 2012. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics* 13:47. <https://doi.org/10.1186/1471-2105-13-47>.
 20. Gao F. 2005. Amplification and cloning of near full-length HIV-2 genomes. *Methods Mol Biol* 304:399–407. <https://doi.org/10.1385/1-59259-907-9:399>.
 21. Pan W, Byrne-Steele M, Wang C, Lu S, Clemmons S, Zahorchak RJ, Han J. 2014. DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnol* 14:10. <https://doi.org/10.1186/1472-6750-14-10>.
 22. Berg MG, Yamaguchi J, Alessandri-Gradt E, Tell RW, Plantier JC, Brennan CA. 2016. A pan-HIV strategy for complete genome sequencing. *J Clin Microbiol* 54:868–882. <https://doi.org/10.1128/JCM.02479-15>.
 23. Gall A, Ferns B, Morris C, Watson S, Cotten M, Robinson M, Berry N, Pillay D, Kellam P. 2012. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J Clin Microbiol* 50:3838–3844. <https://doi.org/10.1128/JCM.01516-12>.
 24. Pinto AJ, Raskin L. 2012. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One* 7:e43093. <https://doi.org/10.1371/journal.pone.0043093>.
 25. Smith EN, Jepsen K, Khosroheidari M, Rassenti LZ, D'Antonio M, Ghia EM, Carson DA, Jamieson CH, Kippis TJ, Frazer KA. 2014. Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. *Genome Biol* 15:420. <https://doi.org/10.1186/PREACCEPT-1251182501124451>.
 26. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. *Proc Natl Acad Sci U S A* 108:20166–20171. <https://doi.org/10.1073/pnas.1110064108>.
 27. McElroy K, Thomas T, Luciani F. 2014. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb Inform Exp* 4:1. <https://doi.org/10.1186/2042-5783-4-1>.
 28. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628. <https://doi.org/10.1038/nmeth.1226>.
 29. Ninomiya M, Ueno Y, Funayama R, Nagashima Y, Nishida Y, Kondo Y, Inoue J, Kakazu E, Kimura O, Nakayama K, Shimosegawa T. 2012. Use of Illumina deep sequencing technology to differentiate hepatitis C virus variants. *J Clin Microbiol* 50:857–866. <https://doi.org/10.1128/JCM.05715-11>.
 30. Batty EM, Wong TH, Trebes A, Argoud K, Attar M, Buck D, Ip CL, Golubchik T, Cule M, Bowden R, Manginis C, Klenerman P, Barnes E, Walker AS, Wyllie DH, Wilson DJ, Dingle KE, Peto TE, Crook DW, Piazza P. 2013. A modified RNA-Seq approach for whole genome sequencing of RNA viruses from faecal and blood samples. *PLoS One* 8:e66129. <https://doi.org/10.1371/journal.pone.0066129>.
 31. Lauck M, Switzer WM, Sibley SD, Hyeroba D, Tumukunde A, Weny G, Taylor B, Shankar A, Ting N, Chapman CA, Friedrich TC, Goldberg TL, O'Connor DH. 2013. Discovery and full genome characterization of two highly divergent simian immunodeficiency viruses infecting black-and-white colobus monkeys (Colobus guereza) in Kibale National Park, Uganda. *Retrovirology* 10:107. <https://doi.org/10.1186/1742-4690-10-107>.
 32. Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 38:e131. <https://doi.org/10.1093/nar/gkp1195>.
 33. van Gurp TP, McIntyre LM, Verhoeven KJ. 2013. Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS One* 8:e85583. <https://doi.org/10.1371/journal.pone.0085583>.
 34. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. 2013. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* 8:e62856. <https://doi.org/10.1371/journal.pone.0062856>.
 35. Yamaguchi J, Brennan CA, Alessandri-Gradt E, Plantier JC, Cloherty GA, Berg MG. 2017. HIV-2 surveillance with next-generation sequencing reveals mutations in a cytotoxic lymphocyte-restricted epitope involved in long-term nonprogression. *AIDS Res Hum Retroviruses* 33:347–352. <https://doi.org/10.1089/aid.2016.0229>.
 36. Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE (ed). 2007. *Fields virology*, 5th ed. Lippincott Williams & Wilkins, Philadelphia, PA.
 37. LANL. HIV sequence database. <https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html>.
 38. Esbjornsson J, Mansson F, Kvist A, Isberg PE, Nowroozaladeh S, Biague AJ, da Silva ZJ, Jansson M, Fenyo EM, Norrgren H, Medstrand P. 2012. Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. *N Engl J Med* 367:224–232. <https://doi.org/10.1056/NEJMoa1113244>.
 39. Borrego P, Marcelino JM, Rocha C, Doroana M, Antunes F, Maltez F, Gomes P, Novo C, Barroso H, Taveira N. 2008. The role of the humoral immune response in the molecular evolution of the envelope C2, V3 and C3 regions in chronically HIV-2 infected patients. *Retrovirology* 5:78. <https://doi.org/10.1186/1742-4690-5-78>.
 40. de Silva TI, Aasa-Chapman M, Cotten M, Hue S, Robinson J, Bibollet-Ruche F, Sarge-Njie R, Berry N, Jaye A, Aaby P, Whittle H, Rowland-Jones S, Weiss R. 2012. Potent autologous and heterologous neutralizing antibody responses occur in HIV-2 infection across a broad range of infection outcomes. *J Virol* 86:930–946. <https://doi.org/10.1128/JVI.06126-11>.
 41. Maldarelli F, Kearney M, Palmer S, Stephens R, Mican J, Polis MA, Davey RT, Kovacs J, Shao W, Rock-Kress D, Metcalf JA, Rehm C, Greer SE, Lucey DL, Danley K, Alter H, Mellors JW, Coffin JM. 2013. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *J Virol* 87:10313–10323. <https://doi.org/10.1128/JVI.01225-12>.
 42. Lipscomb JT, Switzer WM, Li JF, Masciotra S, Owen SM, Johnson JA. 2014. HIV reverse-transcriptase drug resistance mutations during early infection reveal greater transmission diversity than in envelope sequences. *J Infect Dis* 210:1827–1837. <https://doi.org/10.1093/infdis/jiu333>.
 43. Fujita M, Nomaguchi M, Adachi A, Otsuka M. 2012. SAMHD1-dependent and -independent functions of HIV-2/SIV Vpx protein. *Front Microbiol* 3:297. <https://doi.org/10.3389/fmicb.2012.00297>.
 44. Yu H, Usmani SM, Borch A, Kramer J, Sturzel CM, Khalid M, Li X, Krnavek D, van der Ende ME, Osterhaus AD, Gruters RA, Kirchhoff F. 2013. The efficiency of Vpx-mediated SAMHD1 antagonism does not correlate with the potency of viral control in HIV-2-infected individuals. *Retrovirology* 10:27. <https://doi.org/10.1186/1742-4690-10-27>.
 45. Joshi NA, Fass JN. 2011. Sickie: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). <https://github.com/najoshi/sickle>.
 46. Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ, Qu J, Ryan EM, Zody MC, Henn MR. 2012. De novo assembly of highly diverse viral populations. *BMC Genomics* 13:475. <https://doi.org/10.1186/1471-2164-13-475>.
 47. Kearse S, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>.

48. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>.
49. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595. <https://doi.org/10.1093/bioinformatics/btp698>.
50. Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859–1875. <https://doi.org/10.1093/bioinformatics/bti310>.
51. Novocraft. Novoalign short read mapper. <http://www.novocraft.com>.
52. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
53. R. The R Project for Statistical Computing. <https://www.r-project.org/>.
54. Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14:178–192. <https://doi.org/10.1093/bib/bbs017>.
55. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303. <https://doi.org/10.1101/gr.107524.110>.
56. Andrews S. FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
57. Darrriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9:772. <https://doi.org/10.1038/nmeth.2109>.
58. Esbjörnsson J, Månsson F, Martínez-Arias W, Vincic E, Biague AJ, da Silva ZJ, Fenyö EM, Norrgren H, Medstrand P. 2010. Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease—indication of an evolving epidemic in West Africa. *Retrovirology* 7:23. <https://doi.org/10.1186/1742-4690-7-23>.
59. Visseaux B, Hurtado-Nedelec M, Charpentier C, Collin G, Storto A, Matheron S, Larrouy L, Damond F, Brun-Vezinet F, Descamps D, ANRS CO 05 HIV-2 Cohort. 2012. Molecular determinants of HIV-2 R5-X4 tropism in the V3 loop: development of a new genotypic tool. *J Infect Dis* 205: 111–120. <https://doi.org/10.1093/infdis/jir698>.
60. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. <https://doi.org/10.1093/nar/gkh340>.
61. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214. <https://doi.org/10.1186/1471-2148-7-214>.
62. Rambaut A. FigTree. <http://tree.bio.ed.ac.uk/software/figtree/>.
63. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25:2283–2285. <https://doi.org/10.1093/bioinformatics/btp373>.
64. Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* 76: 5269–5273. <https://doi.org/10.1073/pnas.76.10.5269>.
65. Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD dissertation. University of Texas at Austin, Austin, TX.
66. Nyamweya S, Townend J, Zaman A, Steele SJ, Jeffries D, Rowland-Jones S, Whittle H, Flanagan KL, Jaye A. 2012. Are plasma biomarkers of immune activation predictive of HIV progression: a longitudinal comparison and analyses in HIV-1 and HIV-2 infections? *PLoS One* 7:e44411. <https://doi.org/10.1371/journal.pone.0044411>.