



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/198933/>

Version: Published Version

---

**Article:**

Boswell, M.T., Nazziwa, J., Kuroki, K. et al. (2022) Intrahost evolution of the HIV-2 capsid correlates with progression to AIDS. *Virus Evolution*, 8 (2). veac075. ISSN: 2057-1577

<https://doi.org/10.1093/ve/veac075>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. 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](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# Intrahost evolution of the HIV-2 capsid correlates with progression to AIDS

M. T. Boswell,<sup>1,†</sup> J. Nazziwa,<sup>2</sup> K. Kuroki,<sup>3</sup> A. Palm,<sup>2</sup> S. Karlson,<sup>2</sup> F. Månsson,<sup>2</sup> A. Biague,<sup>4</sup> Z. J. da Silva,<sup>4</sup> C. O. Onyango,<sup>5</sup> T. I. de Silva,<sup>6,7</sup> A. Jaye,<sup>7</sup> H. Norrgren,<sup>8</sup> P. Medstrand,<sup>2</sup> M. Jansson,<sup>9,‡</sup> K. Maenaka,<sup>3</sup> S. L. Rowland-Jones,<sup>1,7</sup> and J. Esbjörnsson,<sup>1,2,\*§</sup> the SWEGUB CORE group

<sup>1</sup>Nuffield Department of Medicine, University of Oxford, Roosevelt Drive, OX3 7FZ, Oxford, UK, <sup>2</sup>Department of Translational Medicine, Lund University, Sölvegatan 17, 223 62, Lund, Sweden, <sup>3</sup>Faculty of Pharmaceutical Sciences and Global Station for Biosurfaces and Drug Discovery, Hokkaido University, Kita-12, Nishi-6, Kita-ku, Sapporo 060-0812, Japan, <sup>4</sup>National Public Health Laboratory, V94M+HM4, Bissau, Guinea-Bissau, <sup>5</sup>US Centres for Disease Control, KEMRI Complex, Mbagathi Road off Mbagathi Way PO Box 606-00621, Kenya, <sup>6</sup>Department of Infection, Immunity and Cardiovascular Disease, The Medical School, University of Sheffield, Beech Hill Rd, S10 2RX, Sheffield, UK, <sup>7</sup>Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Atlantic Boulevard, Fajara P. O. Box 273, Banjul, The Gambia, <sup>8</sup>Department of Clinical Sciences Lund, Lund University, Sölvegatan 19, 221 84 Lund, Sweden and <sup>9</sup>Department of Laboratory Medicine, Lund University, Sölvegatan 19, Sweden

<sup>†</sup><https://orcid.org/0000-0003-2152-1617>

<sup>‡</sup><https://orcid.org/0000-0001-6536-8146>

<sup>§</sup><https://orcid.org/0000-0002-6088-7796>

\*Corresponding authors: E-mail: [boswell.michaelt@gmail.com](mailto:boswell.michaelt@gmail.com); [joakim.esbjornsson@med.lu.se](mailto:joakim.esbjornsson@med.lu.se)

## Abstract

HIV-2 infection will progress to AIDS in most patients without treatment, albeit at approximately half the rate of HIV-1 infection. HIV-2 capsid (p26) amino acid polymorphisms are associated with lower viral loads and enhanced processing of T cell epitopes, which may lead to protective Gag-specific T cell responses common in slower progressors. Lower virus evolutionary rates, and positive selection on conserved residues in HIV-2 *env* have been associated with slower progression to AIDS. In this study we analysed 369 heterochronous HIV-2 p26 sequences from 12 participants with a median age of 30 years at enrolment. CD4% change over time was used to stratify participants into relative faster and slower progressor groups. We analysed p26 sequence diversity evolution, measured site-specific selection pressures and evolutionary rates, and determined if these evolutionary parameters were associated with progression status. Faster progressors had lower CD4% and faster CD4% decline rates. Median pairwise sequence diversity was higher in faster progressors ( $5.7 \times 10^{-3}$  versus  $1.4 \times 10^{-3}$  base substitutions per site,  $P < 0.001$ ). p26 evolved under negative selection in both groups ( $dN/dS = 0.12$ ). Median virus evolutionary rates were higher in faster than slower progressors – synonymous rates:  $4.6 \times 10^{-3}$  vs.  $2.3 \times 10^{-3}$ ; and nonsynonymous rates:  $6.9 \times 10^{-4}$  vs.  $2.7 \times 10^{-4}$  substitutions/site/year, respectively. Virus evolutionary rates correlated negatively with CD4% change rates ( $\rho = -0.8$ ,  $P = 0.02$ ), but not CD4% level. The signature amino acid at p26 positions 6, 12 and 119 differed between faster (6A, 12I, 119A) and slower (6G, 12V, 119P) progressors. These amino acid positions clustered near to the TRIM5 $\alpha$ /p26 hexamer interface surface. p26 evolutionary rates were associated with progression to AIDS and were mostly driven by synonymous substitutions. Nonsynonymous evolutionary rates were an order of magnitude lower than synonymous rates, with limited amino acid sequence evolution over time within hosts. These results indicate HIV-2 p26 may be an attractive therapeutic target.

**Key words:** HIV-2; evolution; capsid; p26

## Introduction

HIV-1 and HIV-2 are retroviruses that transmitted from non-human primates to humans in the 20th century. HIV-1 has four main groups, of which M accounts for 98% of global infections. HIV-1 group M descended from SIVcpz, which circulates in chimpanzees (Sharp and Hahn 2011). HIV-2 is descended from SIVsm, which circulates in sooty mangabeys, and has nine groups (A–I), of which A and B account almost all human infections (Sharp and Hahn 2011). HIV-1 has caused a global pandemic, whereas HIV-2 has remained an endemic infectious disease in West Africa, with limited spread outside the region (Visseaux

et al. 2016). In addition to stark differences in transmissibility, the two viruses differ in disease progression rates (Kanki et al. 1994; Esbjörnsson et al. 2019). In the absence of treatment, HIV-1 will progress to AIDS twice as fast as HIV-2. However, HIV-2 disease progression is highly variable, with some patients developing AIDS in similar timeframes as HIV-1 and others progressing far more slowly.

An extensive body of literature has connected intrahost evolution of HIV-1 to disease progression and phenotypic trait development (Williamson 2003; Bello et al. 2007; Lemey et al. 2007; Mild et al. 2010; Mild et al. 2013; Garcia-Knight et al. 2016;

Theys et al. 2018). Synonymous evolutionary rates (nucleotide substitutions that do not change amino acid sequences) in HIV-1 *env* correlate with time to AIDS, CD4+ count decline rates, and viral load, while nonsynonymous evolutionary rates are not associated with disease progression. Similar associations have been reported for intrahost evolution HIV-1 *gag* (Norström et al. 2014). In HIV-1 infection, higher replicative capacity of founder viruses explains much of the variation in disease progression rates and is strongly associated with increasing T cell activation and exhaustion (Claiborne et al. 2015). Together, these results suggest that faster virus replication drives HIV-1 disease progression due to heightened immune activation, rather than evolution being driven by immune escape (Lemey et al. 2007).

HIV-2 *env* evolutionary rates in faster progressors are approximately double that of slower progressors (Palm et al. 2019). In HIV-2 infection, the relationship between immune activation and progressive CD4+ T cell loss is similar to that in HIV-1, at equivalent degrees of immune suppression (Sousa et al. 2002). This suggests that evolutionary rates, immune activation, and subsequent disease progression are similarly linked in HIV-1 and HIV-2 infection.

The HIV-1 capsid is a fullerene cone structure made up of 216 p24 hexamers and 12 p24 pentamers (Zhao et al. 2013). In HIV-2, the homologous capsid protein is p26. The capsid is critical in the viral replication cycle, with several indispensable functions including nucleotide supply to the replicating virus, nuclear import to facilitate integration, cyclophilin-A (CyPA) binding, and immune evasion (Matreyek and Engelman 2011; Schaller et al. 2011; Jacques et al. 2016; Lahaye et al. 2018). HIV-1 and HIV-2 capsids bind to the innate sensor NANO in the nucleus, which triggers the cyclic GMP-AMP synthase and stimulator of interferon genes (cGAS-STING) pathway to activate innate immune responses in macrophages and dendritic cells (Lahaye et al. 2018). HIV-2 binds NANO with higher affinity than HIV-1. HIV-2 capsids are also more sensitive to restriction factor TRIM5 $\alpha$  than HIV-1 (Mamede et al. 2017). Polymorphisms in HIV-2 p26 have been linked to disease progression. Specifically, prolines at p26 positions 119, 159, and 178 are associated with enhanced proteasomal processing of T cell epitopes often targeted in those participants with lower viral loads and slower disease progression (Onyango et al. 2010; Jallow et al. 2015; de Silva et al. 2018a). Additionally, proline at position 119 in p26 alters the conformation of the capsid structure and is associated with increased sensitivity to TRIM5 $\alpha$  (Song et al. 2007; Miyamoto et al. 2011).

To investigate the hypothesis that disease progression in HIV-2 infection correlates with p26 sequence evolution, we tested whether CD4% kinetics is associated with (1) pairwise sequence diversity evolution, (2) site-specific selection pressures, (3) synonymous and nonsynonymous evolutionary rates, and (4) amino acid sequence variation in the virus quasispecies.

## Methods

This section provides a summary of the main laboratory and statistical methods used in this study. Further details are available in the [supplementary materials](#).

### Study participants

Study participants were recruited to the Guinea-Bissau Police cohort between 1990 and 2009. Informed consent was obtained from the participants, and the study was approved by the research ethic committees of the Ministry of Health in Guinea-Bissau, Lund

University, and the Karolinska Institute, Sweden. Twelve HIV-2-positive participants were included in this analysis based on availability of plasma samples. Selection criteria included that the participants were HIV-2-mono-infected and antiretroviral therapy (ART)-naïve at the time of plasma sample collection and had longitudinal CD4+ T cell measurements available, which allowed for estimation of disease progression (Table S1). Clinical and immunological staging of HIV disease was performed according to the World Health Organisation (WHO) criteria (Organization WH 2007). A total of seven participants had an estimated date of HIV-2 seroconversion ('seroincident') and five participants were HIV-2-positive at enrolment ('seroprevalent'). T cell data were analysed from the first time point after HIV-2 detection; in seroincident participants, this time point was the first one after their documented seroconversion, and in seroprevalent participants, at enrolment into the cohort.

### Analysis of disease progression markers in HIV-2

Both absolute CD4+ T cell counts and CD4+ T cell percentage (CD4%) are reliable immunological markers of HIV disease progression. In resource-limited settings, CD4% measurements are less sensitive to specimen handling, participant age, or time of sampling when compared to absolute CD4 counts (Anglaret et al. 1997; Norrgren et al. 2003; van der Loeff et al. 2010; Esbjörnsson et al. 2012). We therefore chose to analyse disease progression using CD4% change over time. Briefly, participants' longitudinal CD4% were analysed in per-participant linear regression models. The CD4% level at the midpoint in follow-up time after HIV-2 detection and the CD4% change rate (slope of the regression line) were extracted from these models and analysed as markers of disease progression.

To create disease progression groups, participants were ranked and classified according to three approaches as previously described (Palm et al. 2019): (1) from highest midpoint CD4% to the lowest (those above the mean were classified as slower progressors and those below the mean as faster progressors); (2) from highest positive change rate to the lowest negative change rate; and (3) the midpoint CD4% and CD4% change rate were then transformed into proportional values, added together, and averaged for each participant. This gave a combined coefficient for each participant, which was weighted equally according to midpoint CD4%, and the rate of change, which accounts for differences in disease stage at enrolment. The combined coefficient was used to rank participants and stratify them into relative progression with distinct disease phenotypes—faster and slower disease progression (Table S2). All analyses that refer to progression groups used this combined coefficient for stratification.

### RNA extraction and PCR amplification

Briefly, plasma samples had previously been collected from participants and stored at  $-8^{\circ}\text{C}$ ; we extracted RNA using the RNeasy Lipid Tissue Mini Kit (Qiagen, Venlo, Netherlands) with minor modifications to the manufacturer's instructions. Following RNA extraction, a nested PCR was performed with 5  $\mu\text{L}$  of extracted RNA. The first step involved a one-step reverse transcription PCR (RT-PCR) using the SuperScript IV One-Step RT-PCR System with Platinum Taq DNA Polymerase ([supplementary materials \[Tables S3–6\]](#)). Immediately following the RT-PCR, a second nested reaction using an inner set of HIV-2 p26 primers was performed using the Dream Taq PCR kit (Thermo Fisher Scientific, Waltham, MA (Fig. S1). Primer sequences are listed in [Table S4](#).

## Cloning and sequencing

Amplicons of 846 nucleotides from positions 1411–2257 on HIV-2 BEN.M30502 were cloned into pCR<sup>®</sup>4TOPO<sup>®</sup> vector using the TOPO-TA cloning kit (Invitrogen, Carlsbad, CA, USA); 23 white colonies were randomly picked and amplified by colony PCR using the Advantage 2 PCR kit (Takara, Kusatsu, Japan) (Fig. S2) to confirm the presence of the insert. The plasmids containing inserts were sequenced by Sanger sequencing using the inner PCR primers (Table S4) (Macrogen Europe, Amsterdam, the Netherlands).

## Sequence analysis

Raw sequence data were analysed in Geneious Prime v. 2019.2 (Geneious, Biomatters, Auckland, New Zealand), mapped to the HIV-2 BEN.M30502 reference sequence, and trimmed for quality (Kearse et al. 2012). Contigs were constructed for each participant from single forward and reverse reads. Poor-quality reads and contigs with high-quality scores of less than 80% were not included in further downstream analysis. This score indicates that 80% of bases are of high quality and that the likelihood of a false-positive base reading is 1:10 000 (Biomatters). Mixed peaks in electropherograms were resolved in favour of the consensus sequence for each participant's alignment—thus ensuring that mixed bases were not resolved across participants. Contigs were assembled into participant-specific alignments and mapped to HIV-2 BEN.M30502 to ensure that all participant sequences mapped to the same coding region of *gag*. Recombination between sequences grouped in an alignment can violate the assumptions made in phylogenetic analysis (Posada and Crandall 2002; Martin et al. 2015). We therefore screened each participant-specific data set for recombination in RDP4 using a combination of the following methods: RDP, GENECONV, BootScan, MaxChi, and Chimaera (Smith 1992; Padidam 1999; Martin and Rybicki 2000; Posada and Crandall 2001; Martin et al. 2005; Martin et al. 2015). All sequences showing evidence of recombination were removed from further analysis.

## Bayesian phylogenetic analysis

All Bayesian phylogenetic parameters were specified in BEAUti v. 1.10.4 and run in BEAST v1.10.4 with parallel processing by BEAGLE (Ayres et al. 2012; Drummond et al. 2012). Model output logs were analysed in Tracer v1.7 (Rambaut et al. 2018) and assessed for convergence by visual inspection of the posterior distribution chains and if effective sample sizes (ESSs) were >100, after 10% burn-in. All analyses were run in duplicate to assess reproducible convergence, which would include those with ESS between 100 and 200. Runs with ESS values <100 were not used. Bayesian phylogenetic models were used to (1) perform viral subtyping, (2) reconstruct the most recent common ancestor (MRCA) for the participants' combined sequence alignment, (3) measure pairwise sequence diversity within participant sequence alignments, (4) estimate site-specific dN/dS ratios in p26 for each participant, (5) estimate relaxed molecular clock estimates for individual participants, and (6) estimate participant-specific molecular clock rates via a hierarchical phylogenetic model (HPM).

i) Viral subtyping was performed by sampling HIV-2 *gag* sequences from the clonal sequences. *Gag* sequences for HIV-2 groups A–G as well as SIVsmm and HIV-1 were downloaded from the Los Alamos National Laboratory sequence database, and an alignment of the participant HIV-2 sequences with these reference sequences was created (Table S7). Viral subtyping was performed using a

Bayesian phylogenetic approach. A strict molecular clock model, Hasegawa-Kishino-Yano (HKY) substitution model, and constant population size were specified. The Markov chain was run for  $2 \times 10^6$  iterations.

- ii) The following parameters were used for the MRCA reconstruction—strict molecular clock model, General Time-Reversible (GTR) nucleotide substitution model, no site heterogeneity or codon partition, and constant population size. The Markov chain was run for  $2.5 \times 10^6$  iterations. The MRCA was used as a reference sequence for substitution identification and mapping of amino acid substitutions over time (discussed below).
- iii) To measure pairwise sequence diversity, the sequences from the participant samples are labelled according to the time point of collection, and these are then aligned using CLUSTAL-W. Using GARLI V2.01, 200 maximum likelihood bootstrap trees are generated for each data set. The diversity estimates in base substitution per base site are obtained from each of the 200 trees using the BIOTREE:IO function of the BioPerl package. These estimates are then summarised in R to get the mean diversity and the 95% confidence intervals.
- iv) To estimate selection pressure, the dN/dS ratio is calculated by dividing the nonsynonymous evolutionary (dN) rate by the synonymous (dS) evolutionary rate (a ratio of less than 1.0 will indicate negative selection; a ratio equal to 1.0 neutral selection, and a ratio greater than 1.0 positive selection at a site). We measured site-specific selection pressures using Renaissance counting procedures performed with the following model parameters; a strict molecular clock model, GTR nucleotide substitution model, no site heterogeneity, a 1,2,3 codon partition, constant population size, and  $4 \times 10^7$  Markov chain Monte Carlo (MCMC) chains (Lemey et al. 2012).
- v) To determine combined, synonymous, and nonsynonymous evolutionary rates derived from the participant-specific relaxed molecular clock models, we used software packages developed by Lemey et al. (Lemey et al. 2007). Briefly, for each participant nucleotide alignment, a relaxed molecular clock model was created in BEAUti and run in BEAST. Model parameters included an HKY nucleotide substitution model, gamma site heterogeneity, a constant population size, and  $2 \times 10^8$  MCMC chains. For each participant, 10,000 trees were simulated as a posterior sample distribution. From these 10,000 trees, 200 were selected randomly after a 10% burn-in and separated into nucleotide substitution unit-denoted and time unit-denoted trees. The substitution trees were then separated further using HyPhy into expected synonymous and nonsynonymous substitution trees. Next, these substitution trees were analysed separately in conjunction with their respective time unit trees to generate estimates of the combined, synonymous, and nonsynonymous evolutionary rates. Finally, the substitution trees were used to estimate divergence as a function of time from the first time point to the last time point in the alignment. This allowed us to plot divergence over time.
- vi) A significant limitation of the relaxed molecular clock estimates described above are interparticipant variability in number of sequences, number of time points, and total follow-up time. The HPM allows for feedback across participant average estimates to improve participant specific estimates, thereby making use of a larger data set (all sequences linked in the HPM) to inform smaller partitions in the data (participant-specific estimates). Another main difference in

the HPM estimates versus those above are that HPMs assume a strict molecular clock per partition, which does not vary for each taxon, while the previous estimates assumed an uncorrelated relaxed molecular clock per participant. We partitioned the HPM by using fixed factor terms for the combined coefficient stratification, as well as the log values of the midpoint CD4 % and CD4 % change rate. Model parameters included a strict molecular clock model, HKY nucleotide substitution model, gamma site heterogeneity, a 1, 2, and 3 codon partition, and constant population size. The Markov chain was run for  $6 \times 10^8$  iterations. We calculated Bayes Factors to test whether the midpoint CD4 % and CD4 % change rate were significant explanatory variables in evolutionary rate estimates. We repeated the HPM analysis with a narrower range of prior values, which could be assessed by the model. Our initial HPM used a hyperprior scale parameter of 1000, and we subsequently adjusted these to scales of 100 and 10—this allows for a narrower distribution of values to be assessed for prior parameters (Raghwani et al. 2018).

### Amino acid sequence analysis

We translated and aligned all p26 sequences to the MRCA to identify single amino acid polymorphisms (SAAPs). We categorised SAAPs as majority substitutions if they occurred in more than 50 % of sequences and minority substitutions if the prevalence was below 50 %. Polymorphisms found in a single sequence were designated as rare substitutions. Rare single nucleotide polymorphisms identified in virus quasispecies using cloning and Sanger sequencing are often not found using next-generation sequencing platforms and may represent a combination of PCR and sequencing errors (Iyer et al. 2015).

To test whether the frequency of substitutions significantly differed between progression groups, we divided the sequence alignment by progression status (faster vs. slower progressors). We then created new alignments of 1,000 sequences per progression group (randomly selected via bootstrap sampling with replacement). We then compared the amino acid frequencies by the progressor group using the Viral Epidemiology Signature Pattern Analysis (VESPA) tool (Korber and Myers 1992). VESPA identifies signature patterns, which differ between alignments and reports frequencies of the specific amino acids in each alignment. HIV-2 p26 structural models were generated in Pymol v. 1.8, using protein sequence PDB ID: 2WLTV (Schrödinger 2015). Residues that have been linked to important functional regions/structures on the HIV-1 p24 protein were used to infer sites on HIV-2 p26, which may play a similar role. This was performed by aligning reference sequences HIV-1.NL4-3 and HIV-2 BEN.M30502.

### Caio cohort sequence analysis

We investigated whether the signature amino acid substitutions identified by the VESPA analysis were associated with disease progression markers in a larger external cohort. This was performed by testing the association of these p26 amino acid substitutions with CD4 % and HIV-2 plasma viral loads in 86 HIV-2-positive participants from the Caio cohort (Onyango et al. 2010; de Silva et al. 2018b).

### Statistical analysis

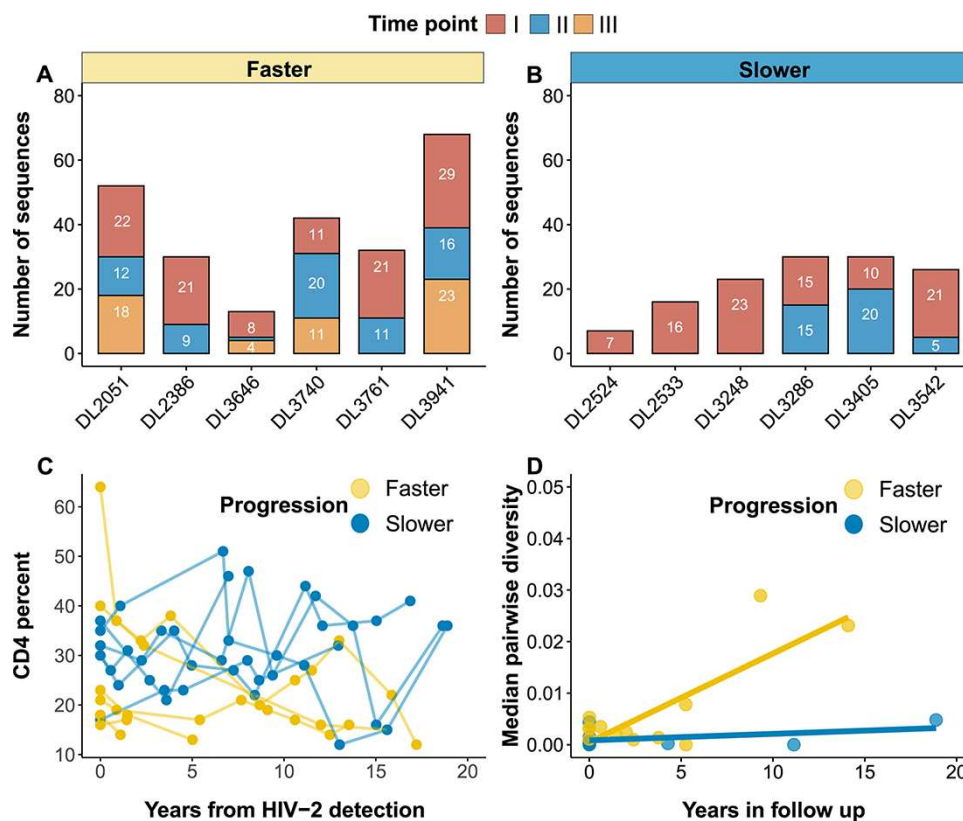
All statistical analyses were two-sided, and data visualisations were performed in R studio v. 4.0.3, unless specified otherwise (Kassambara 2018; RStudio Team 2021). Baseline characteristics of participants were summarised as means and standard deviations

(SD), medians, interquartile ranges (IQR), and counts with percentages. Pairwise testing was performed using the Mann–Whitney (MW) *U* test for non-normally distributed data and the Student's *t*-test for normally distributed data. To compare distributions of repeated measurements between the progression groups, we used Friedman tests (FT) with effect sizes reported by Kendall's *W* value (small effect:  $W = 0.1$ – $0.3$ ; moderate effect:  $0.3$ – $0.5$  and large effect  $>0.5$ ). The frequency of categorical variables was assessed by chi squared tests. Proportional differences were also assessed using Fisher's exact test and reported as odds ratios (ORs). Correlation statistics were calculated using Pearson's (parametric) or Spearman's (non-parametric) correlation coefficients dependent on the variables' distribution. Pairwise sequence diversity evolution was quantified using linear mixed effects models with model fit assessed by the likelihood ratio test (LRT). A false discovery rate was used to correct for multiple comparisons. Statistical significance was determined as  $P < 0.05$ .

## Results

Our analysis included 12 HIV-2-positive participants from the Guinea-Bissau Police cohort. All participants were male with a median age at enrolment of 30 (IQR: 28–37) years. The cohort median midpoint CD4 % was 27.5 %, and the median CD4 % change rate was  $-0.05$  % per year. After calculation of the combined coefficient and assignment to progression groups, there were six faster progressors and six slower progressors. Faster progressors had a significantly lower midpoint CD4 % than slower progressors (20.7 % vs. 30.6 %,  $P = 0.02$ , MW). In addition, the faster progressors' CD4 % decreased by 1.4 % per year, whereas the slower progressors' CD4 % increased by 0.6 % per year ( $P = 0.02$ , MW). One of the slower progressors developed severe immunosuppression (CD4 % below 15), whereas 5/6 faster progressors reached a CD4 % below 15 % during follow-up (Fig. 1). In seroincident participants, the median time from HIV-2 detection to the first sample was 4.5 (IQR: 3.9–6.7) years, and in seroprevalent participants, the median time from HIV-2 detection (enrolment) to sample was 14.4 (IQR: 2.4–15.6) years. The mean follow-up time from the first to last sequencing sample was 5.9 years, 6.1 years for faster progressors, and 5.7 for slower progressors.

After aligning the sequences and trimming for quality, the analysed *gag* sequences spanned 735 nucleotides in a single open reading frame, which mapped to nucleotide 1460–2194 on HIV-2 BEN.M30502. These sequences are available from GenBank with accession numbers OL872372-872739 and OM146012. This sequence included the first 687 nucleotides, from 5' to 3', of the p26 region of *gag*. Twenty-five plasma samples yielded PCR products from the twelve participants, with PCR products from longitudinal sample time points generated from nine participants (Table 1). In total, 575 clones were generated and sequenced, whereof 173 were removed due to poor sequence quality or presence of stop codons. Of the remaining 402 sequences, thirty three were possible recombinant sequences and were removed from further analysis. This resulted in a final data set of 369 sequences from twelve study participants (median: 30 sequences/participant; Fig. 1). Participant sequences formed monophyletic clusters with high posterior support values, suggesting that contamination or labelling errors did not occur during sample handling. No clear clustering pattern by progression status was observed (Fig. S3). Moreover, the subtype analysis indicated that all sequences clustered with HIV-2 group A reference sequences (Table S7 and Fig. S4).



**Figure 1.** Summary of number of sequences, CD4% kinetics, and pairwise diversity by progression status. (A and B) Bar plots summarise the number of sequences per participant and time point. Panels are stratified by progression status as determined by the combined coefficient. There are six faster and slower progressors. Slower progressors had fewer sequences (numbers inside the bars) and time points for analysis than faster progressors. (C) CD4% kinetics shown for faster and slower progressors as defined by the combined coefficient. Faster progressors had a significantly lower CD4% and faster CD4% decline rates than slower progressors ( $P < 0.05$ , MW). (D) Scatter plots with fitted linear regression lines showing pairwise diversity increasing over time, presented as nucleotide substitutions per site (y-axis) and time in years between samples (x-axis). Pairwise sequence diversity was higher in faster progressors ( $P < 0.05$ , MW).

### Pairwise sequence diversity is associated with disease progression status

Sequence diversity increased over time (Spearman correlation:  $\rho = 0.26$ ,  $P < 0.001$ , Fig. 1). The pooled median sequence diversity was significantly higher among faster than slower progressors ( $5.7 \times 10^{-3}$  versus  $1.4 \times 10^{-3}$  base substitutions per site, MW  $P < 0.001$ ). Analysing linear regression models of diversity over time, the average sequence diversity increased by  $1.7 \times 10^{-3}$  (95% CI:  $1.7\text{--}1.8 \times 10^{-3}$ ,  $P < 0.001$ ) substitutions per site, per year (s/s/y), for the faster progressors, and increased by  $1.3 \times 10^{-4}$  s/s/y (95% CI:  $1.1\text{--}1.4 \times 10^{-4}$ ,  $P < 0.001$ ) for the slower progressors. To account for the unequal contribution of sequences, time points, and inter-participant variability, we analysed mixed effects models. A random slope and intercept model (time (i.e. slope of diversity) allowed to vary by participant (random effect) provided the best model fit ( $P < 0.001$ , LRT). In this model, diversity change over time was similar in progression groups (faster progression diversity increased by  $1.2 \times 10^{-3}$  relative to slower progression, 95% CI =  $-9.2 \times 10^{-4}\text{--}3.5 \times 10^{-3}$ ,  $P = 0.2$ ).

### Progression status is associated with synonymous and nonsynonymous evolutionary rates

The results for the relaxed clock estimates are reported for nine participants (six faster and three slower) who had sequences from multiple time points available. The median evolutionary rate in

p26 for all participants was  $4.0 \times 10^{-3}$  (IQR:  $2.6\text{--}6.9 \times 10^{-3}$ ) s/s/y. Evolutionary rates correlated negatively with CD4% change rates, but not midpoint CD4% (Spearman correlation:  $\rho = -0.8$  and  $-0.5$ , respectively,  $P = 0.02$  and  $0.17$ ). Faster progressors had a significantly higher evolutionary rate than slower progressors ( $5.4 \times 10^{-3}$  vs.  $2.5 \times 10^{-3}$  s/s/y,  $W = 0.4$ , FT  $P < 0.001$ , Fig. 2). The median synonymous evolutionary rate for all participants was  $3.5 \times 10^{-3}$  (IQR:  $2.3\text{--}6.1 \times 10^{-3}$ ) s/s/y and was also significantly higher in faster than slower progressors ( $4.6 \times 10^{-3}$  vs.  $2.3 \times 10^{-3}$  s/s/y,  $W = 0.4$ , FT  $P < 0.001$ , Fig. 2). The median nonsynonymous evolutionary rate for all participants was  $4.1 \times 10^{-4}$  (IQR:  $2.4\text{--}8.8 \times 10^{-4}$ ) s/s/y and was significantly higher in faster than slower progressors ( $6.9 \times 10^{-4}$  vs.  $2.7 \times 10^{-4}$  s/s/y,  $W = 0.4$ , FT  $P < 0.001$ , Fig. 2). The dN/dS ratio for p26 was 0.12 (IQR:  $0.08\text{--}0.21$ ), and the difference between faster and slower progressors was small but significant ( $0.13$  vs.  $0.11$ ,  $W = 0.2$ , FT  $P < 0.001$ ). Relaxed evolutionary rate estimates are summarised for each participant in the supplementary materials (Table S8). The effect sizes for progression groups on evolutionary rate comparisons were moderate ( $W = 0.4$ ), and for dN/dS ratios, small ( $W = 0.2$ ). These comparisons were repeated with evolutionary rates in log space, and the results of the analysis were unchanged (Fig. S5). Synonymous and nonsynonymous divergence in p26 increased linearly with time in both faster and slower progressors (Fig. 2). Linear regression models were fitted to each participant's divergence over time ( $r^2$  for synonymous and nonsynonymous divergence was 0.94 and 0.82, respectively).

**Table 1.** Summary results for the study participants.

Study ID	Age <sup>a</sup>	HIV-2 serostatus <sup>b</sup>	Sequencing time point: year	Years HIV-2-positive <sup>c</sup>	Midpoint CD4% <sup>d</sup>	CD4% change rate <sup>e</sup>	Progression <sup>f</sup>
DL2051	28	Prevalent	I: 2004 II: 2006 III: 2009	14 17 20	30.7	-1.8	Faster
DL2386 <sup>g</sup>	30	Incident	I: 2004 II: 2013	8 18	25.3	-1.6	Faster
DL2524	36	Prevalent	I: 2008	17	32.7	+0.7	Slower
DL2533	25	Prevalent	I: 2006	16	25.1	+1.1	Slower
DL3248	19	Incident	I: 2004	3	29.4	+1.0	Slower
DL3286	39	Incident	I: 1997 II: 2008	4 16	32.7	-0.1	Slower
DL3405	28	Incident	I: 2004 II: 2008	5 9	33.1	+1.5	Slower
DL3542	29	Prevalent	I: 1994 II: 2013	1 20	30.8	-1.0	Slower
DL3646	38	Prevalent	I: 1996 II: 2009 III: 2010	2 16 17	18.2	-0.3	Faster
DL3740	39	Incident	I: 2007 II: 2009 III: 2009	3 5 5	16.5	+0.7	Faster
DL3761	32	Incident	I: 2009 II: 2010	9 9	16.0	-3.7	Faster
DL3941	27	Incident	I: 2004 II: 2008 III: 2010	4 8 10	17.4	-1.9	Faster

<sup>a</sup>Age at enrolment to the Police Cohort.

<sup>b</sup>HIV-2 detected includes seroincident (where an estimated date of seroconversion is available) and seroprevalent participants (where a participant was HIV-2-positive at time of enrolment).

<sup>c</sup>Years of follow-up from when HIV-2 was detected, up to the sequencing time point.

<sup>d</sup>Midpoint CD4% calculated from per-participant regression models.

<sup>e</sup>CD4% change rate per year is the linear regression coefficient of CD4% over time.

<sup>f</sup>Progression status as determined by the combined coefficient.

<sup>g</sup>Sample I for participant DL2086 did not have a time stamp. We therefore used the midpoint in follow-up time for this participant with an uncertainty correction of  $\pm 9$  years in all phylogenetic models, which required a date for the sample.

We next analysed a series of HPMs to generate estimates of evolutionary rates in *p26* for each participant with multiple time points and tested the correlation between evolutionary rates and midpoint CD4% and CD4% change rate. The median HPM evolutionary rate for all participants was  $3.1 \times 10^{-3}$  s/s/y (95% Highest Posterior Density interval (HPD):  $1.9-4.5 \times 10^{-3}$ ) and correlated well with the relaxed clock estimates (Spearman correlation:  $\rho = 0.8$ ,  $P = 0.008$ ). We analysed the midpoint CD4%, CD4% change rates, and the combined coefficient stratification as fixed effects to determine if the disease progression markers explained variation in evolutionary rates. There was weak evidence that disease progression markers were associated with evolutionary rate variation. The Bayes Factor for midpoint CD4%, CD4% change rate, and combined coefficient stratifications ranged from 0.4 to 0.5. Median HPM evolutionary rates did not correlate with the midpoint CD4% or CD4% change rate (Spearman correlation:  $\rho = -0.3$  and  $-0.7$  respectively,  $P > 0.05$ ). The median evolutionary rate estimates were similar across hyperprior scale values. The median clock rate across all participants for a scale of 10, 100, and 1,000 was  $2.9 \times 10^{-3}$  (95% HPD:  $1.6-4.9 \times 10^{-3}$ ),  $3.1 \times 10^{-3}$  (95% HPD:  $1.8-4.9 \times 10^{-3}$ ), and  $3.1 \times 10^{-3}$  (95% HPD:  $1.9-4.5 \times 10^{-3}$ ) s/s/y, respectively.

### Estimation of site-specific selection pressures in *p26*

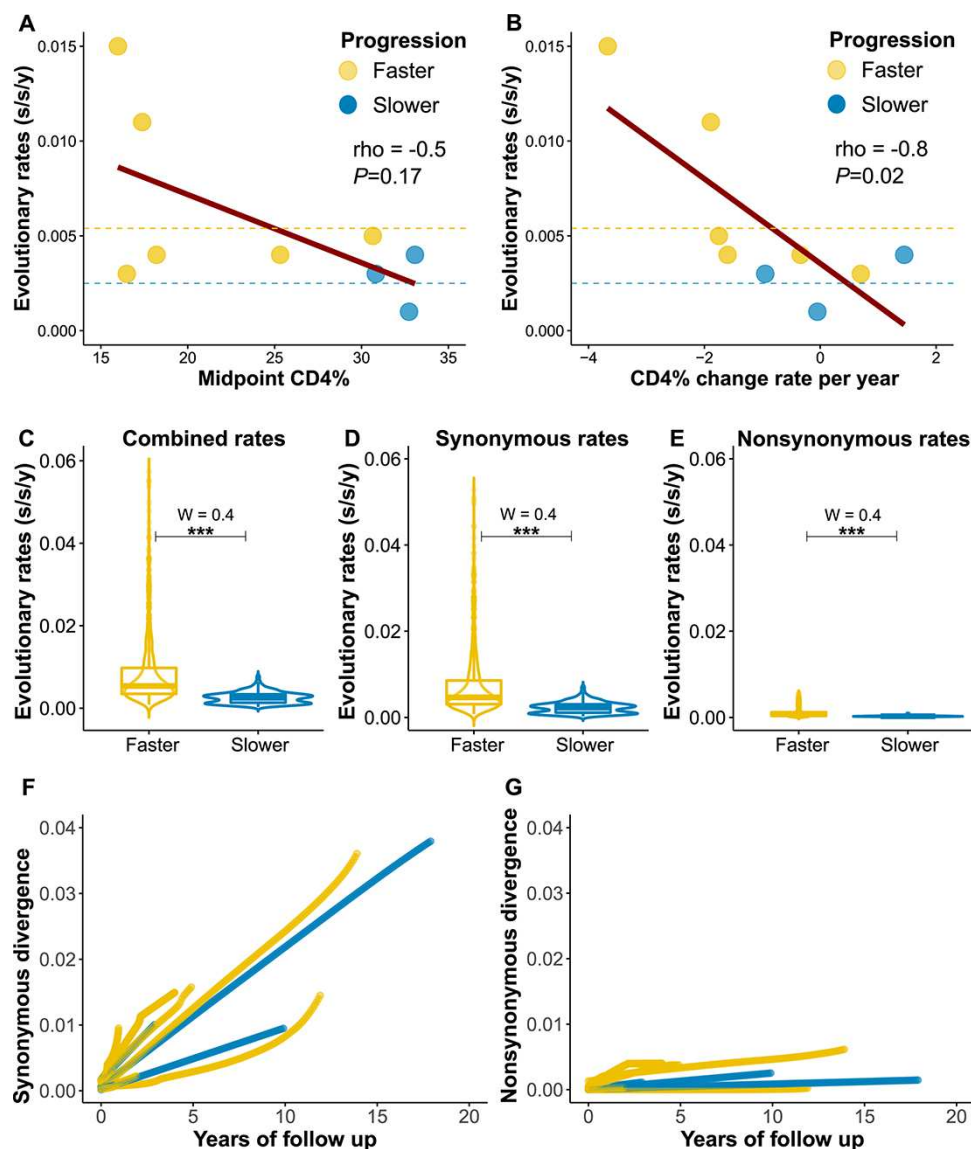
Next, we used Renaissance counting procedures to quantify selection pressures in *p26*, per participant, by estimating site-specific

dN/dS ratios (Lemey et al. 2012). We included participants with two or more time points for this analysis ( $n = 9$ ). Negative selection predominated for all participants, with only one faster progressor showing a signature of positive selection at one site (DL2051 at position five).

### HIV-2 *p26* amino acid signatures differ by progression status

All sequences were used to reconstruct the MRCA sequence at the root of the maximum clade credibility tree (Fig. S3). The MRCA sequence aligned well with HIV-2 BEN.M30502, differing at 10 amino acid positions, and was used for amino acid substitution identification. In total, sixty-one positions among the participants' amino acid sequences differed from the MRCA sequence. After excluding rare substitutions (found in only one sequence), faster progressor amino acid sequences were more likely to differ from the MRCA than slower progressor sequences (35 vs. 19 positions differed, OR = 2.0, 95% CI = 1.1-3.8,  $P = 0.03$ ).

Sixty-nine unique amino acid substitutions were identified, and twenty eight of these were rare substitutions. Of the forty-one remaining substitutions, two were major substitutions (present in more than 50% of all sequences) and thirty nine were minor substitutions (present in less than 50% of all sequences) Two positions on *p26*, 85, and 96 in the CyPA binding loop showed variation from the MRCA (Fig. 3). Based on the results from the VESPA analysis, the amino acid at three positions differed between faster and slower progressors. At positions 6, 12, and 119, in slower

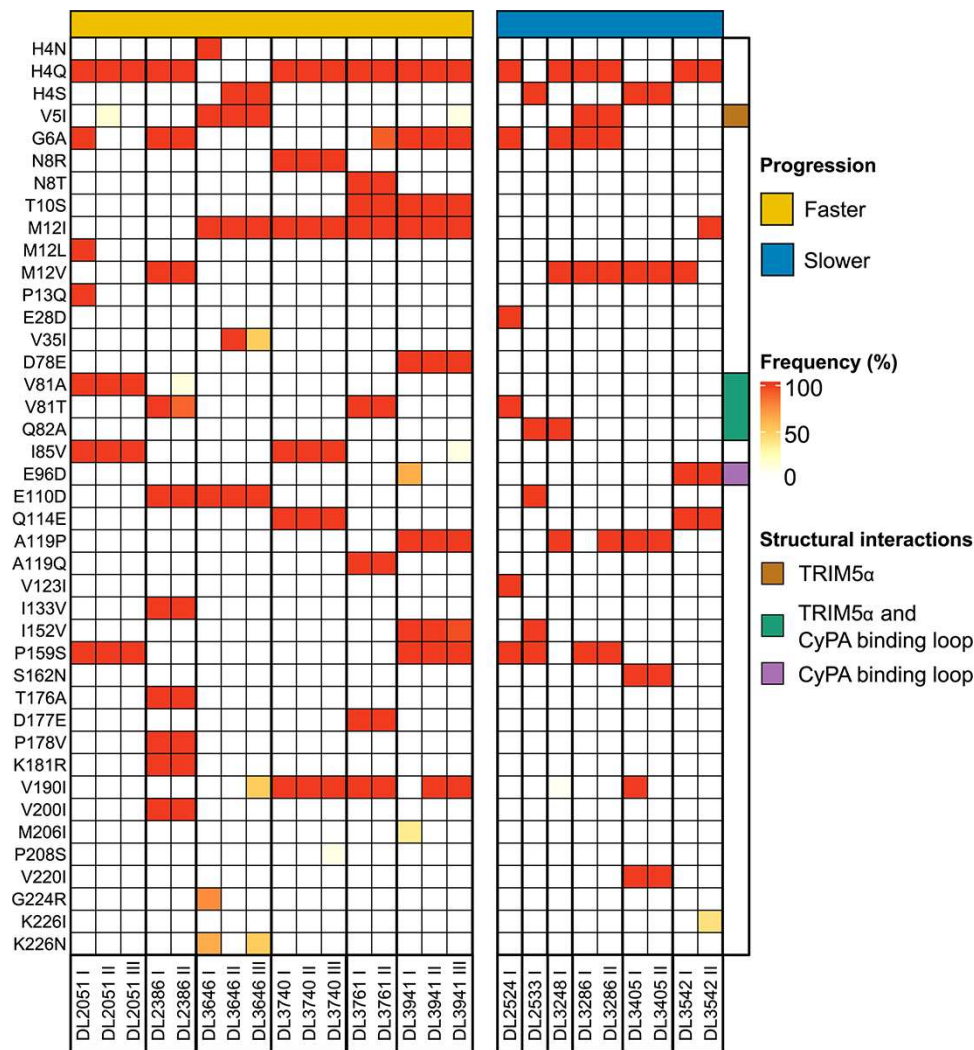


**Figure 2.** Evolutionary rates in relation to disease progression markers. Results are shown for the nine participants with multiple sequence time points A - B) Median relaxed clock evolutionary rates (y-axis) correlated significantly with CD4% change rate, but not CD4% (x-axes). Dashed lines show the median evolutionary rates for faster and slower progressor groups. C - E) Violin plots showing evolutionary rate distributions from 200 randomly sampled trees per participant (y-axis) by their respective progression group (x-axis). Combined, synonymous and nonsynonymous evolutionary rates were significantly higher in faster progressors (FT  $P < 0.001$ ). F - G) Line plots of synonymous and nonsynonymous divergence (measured in substitutions per site) over time by progressor group. The accumulated divergence (y-axis) from the first analysed sample for each participant indicated by time = 0 on the x-axis. \*\*\* =  $P < 0.001$ .

progressors, the most common amino acids were glycine, valine, and proline, and in faster progressors, the most common were alanine, isoleucine, and alanine (Fig. 4). These amino acid positions localised to the N-terminal domain of HIV-2 p26 (Fig. 4) (Price et al. 2009). In addition, positions 6 and 12 are located at the p26 hexamer interface surface, while 119 is next to the CyPA binding loop (Price et al. 2009; Skorupka et al. 2019; Yu et al. 2020: 5). Many of the substitutions were stable in follow-up, and there was little evidence of specific substitutions being consistently selected for—agreeing with the Renaissance counting results (Fig. 3).

### Amino acid residues at p26 positions 6, 12, and 119 in the Caio cohort

To test whether the association between p26 amino acids and disease progression held in a larger cohort, we assessed their association with CD4% and HIV-2 viral loads in the Caio cohort. The proportions of amino acids at positions 6, 12, and 119 in the Caio cohort were similar to those in the Police cohort (Fig. S6). p26 amino acids were not associated with CD4% in the Caio cohort, and only proline at position 119 was associated with lower HIV-2 plasma viral loads (Fig. S6).



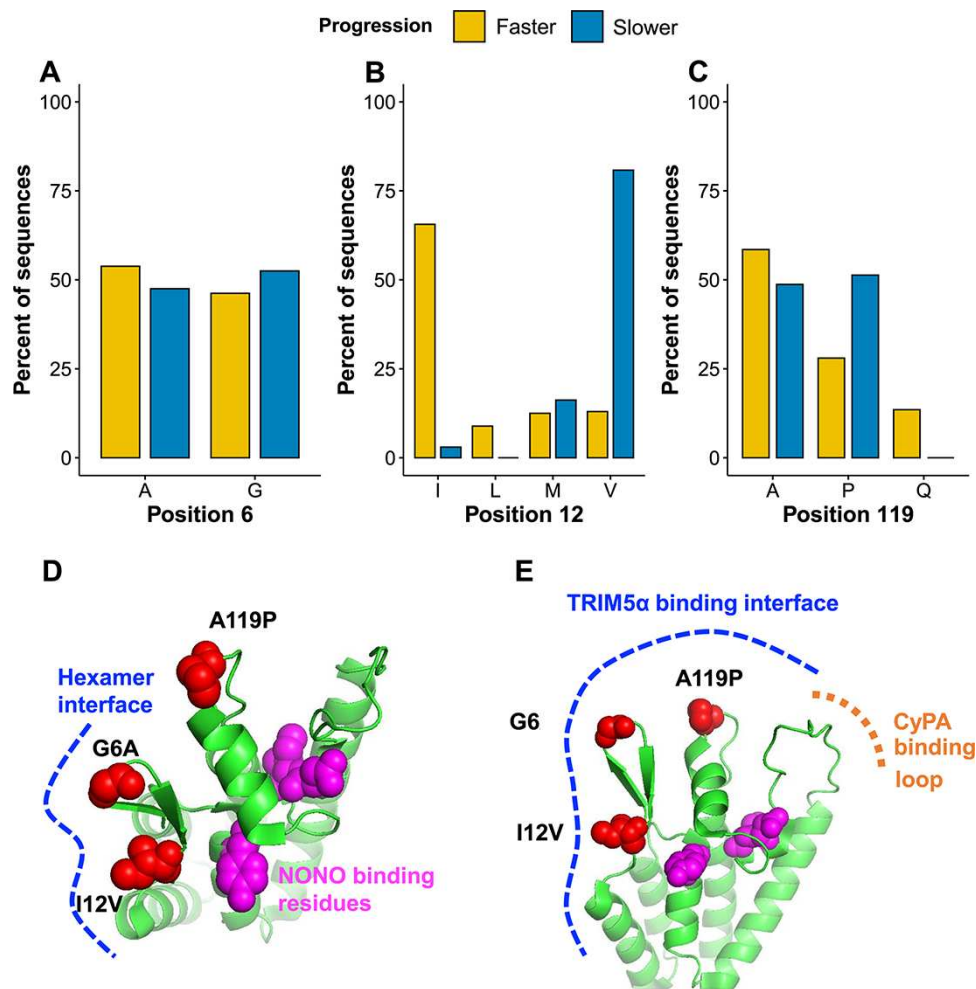
**Figure 3.** Heatmap of HIV-2 p26 amino acid substitution frequencies over time. The y-axis shows the amino acid substitutions' frequencies, with the MRCA amino acid sequence used as a reference for selected positions. Each column is a participant-specific time point. Participants with their respective time points (I – III) are shown on the x-axis, split by progression status. Amino acid frequencies are shown as a percentage of the sequenced quasispecies, with the scale on the right.

## Discussion

Synonymous and nonsynonymous evolutionary rates in p26 were significantly higher in faster progressors and correlated negatively with CD4% change rates. This broadly agrees with a previous study of partial HIV-2 *env* from the same cohort, which found that faster progressors (determined via the midpoint CD4% or the combined coefficient stratification) had higher evolutionary rates (Palm et al. 2019). In HIV-1, p24 synonymous and nonsynonymous evolutionary rates have shown comparable estimates ( $1.6 \pm 1.3 \times 10^{-3}$  s/s/y and  $2.5 \pm 4.8 \times 10^{-3}$  s/s/y, respectively) (Raghwani et al. 2018). In our analysis, HIV-2 nonsynonymous evolutionary rates are an order of magnitude lower than the synonymous rates. Findings from a study of interhost evolution of HIV-2 p26 showed little evidence of positive selection in HIV-2 p26 evolution (de Silva et al. 2018b). This is surprising as strong Gag-specific cytotoxic lymphocyte (CTL) responses are common in HIV-2 long-term non-progressors (de Silva et al. 2013a). It is possible that the HIV-2 capsid is not able to readily adapt to host immune responses without the virus losing ability to successfully replicate, although

this would need further experimental confirmation. If true, this may explain why protection from Gag-specific CTL responses is sustained in HIV-2 long-term non-progressors. This adds further evidence that immune escape and subsequent positive selection in the HIV-2 capsid are not associated with disease progression.

Hierarchical phylogenetic model estimates had narrower 95% HPD intervals than the individual relaxed clock models, which is an expected effect of using HPM (Suchard et al. 2003). Besides the larger HPD intervals in the non-HPM estimates, we had good agreement in results between relaxed clock models and the HPM. In addition, hyperprior scales did not significantly affect evolutionary rate estimates, which suggests that our estimates reflected the sequence data and not model parameters' priors. As expected, synonymous substitutions were the main mode of evolution in HIV-2 p26, which suggests that a process that is common to virus replication (either replicative capacity, generation time, or immune activation) is driving both intrahost evolution and disease progression (Lemey et al. 2007; Claiborne et al. 2015; Asowata et al. 2021).



**Figure 4.** Amino acid sequence diversity in HIV-2 p26 by progression status. A - C) The amino acid signature was different at three positions on p26 between faster and slower progressor alignments. The y-axis shows the percentage of sequences from the pooled sequences per progressor group. D) This model of HIV-2 p26 NTD (PDB ID: 2WLV) with space-filling models of the three residues and those involved in NONO binding. The p26 hexamer interface is shown by a dashed line. E) The same model as in panel D, but at a different angle, indicating the larger TRIM5 $\alpha$  binding interface and CyPA binding loop.

Amino acid variation from the MRCA sequence was greater in faster compared with slower progressors. Pairwise diversity and nonsynonymous divergence were also significantly higher in faster progressors. With a sustained higher nonsynonymous evolutionary rate, we would expect a greater accumulation of amino acid diversity in faster progressors compared to slower progressors. Sequence diversity correlates positively with advancing disease progression in HIV-1, and our results suggest a similar finding for HIV-2 (Troyer et al. 2005).

Inferred sites on HIV-2 p26 with structural/functional motifs, which have previously been linked to the HIV-1 capsid nucleotide pore, nuclear import of the capsid, and NONO binding, were completely conserved across all sequences (Matreyek and Engelman 2011; Schaller et al. 2011; Jacques et al. 2016; Lahaye et al. 2018). This is in line with the conserved nature of retrovirus capsids (Rihn et al. 2013; Mamede et al. 2017; Lahaye et al. 2018). At three sites on p26, the most common amino acid differed between faster and slower progressors—6, 12, and 119. Position 12 flanks a histidine at position 11, which forms part of the nucleotide channel on the capsid (Jacques et al. 2016). The M12I substitution characterised faster progressor sequences, while slower progressors'

most common residue was valine. At position 119, slower progressors often had a proline and faster progressors an alanine. P119 has been associated with lower viral loads and enhanced sensitivity to TRIM5 $\alpha$  (Song et al. 2007; Onyango et al. 2010; Miyamoto et al. 2011). However, the latter finding has been contested (Takeuchi et al. 2013). Position 119 is adjacent to the CyPA binding loop and may interact with TRIM5 $\alpha$ . Positions 6 and 12 are located close to the hexamer interface, as well as the CyPA binding loop. This raises the possibility that amino acid variation at these residues can influence the hexamer formation and/or TRIM5 $\alpha$ -mediated immune activation. Further investigation is needed, and *in vitro* experiments would be valuable in determining whether any of the substitutions or, haplotypes thereof, affect virus replicative capacity or HIV-2-specific immune responses. An alternative (and perhaps more plausible) interpretation of the amino acid diversity near to the p26 hexamer/TRIM5 $\alpha$  interface surfaces is that this region has accumulated diversity from ancestral sequences which adapted to primate hosts and that the mutations identified in this study do not change the HIV-2 capsid's function in human infection (Meyerson and Sawyer 2011; Sauter and Kirchoff 2019). In support of this interpretation, we have shown

that in a larger cohort, amino acids at positions 6, 12, and 119 were not associated with CD4% and only A119P was associated with lower HIV-2 plasma viral loads.

Most p26 amino acid substitutions were stable over time, again indicating that negative selection was dominant in this region's evolution. We offer two explanations for this: first, the capsid is limited in its adaptive capacity, although this does not explain why we observed significant differences in amino acid sequences between participants. Second, the median time from HIV-2 detection to plasma sample collection was 9.3 years. In p26, the primary driver of adaptive mutations is CTL responses (Leigdowicz et al. 2007; de Silva et al. 2013b). Cytotoxic lymphocyte responses drive HIV adaptation in the early stages after infection, and therefore, these changes might not have been captured by our sampling timeframe (Ganusov et al. 2011; Leviyang and Ganusov 2015).

Limitations to our study include that we analysed a small sample of HIV-2-positive participants ( $n = 12$ ); however, this is a relatively large number given the logistical difficulties in obtaining HIV-2 RNA from plasma longitudinally. Within this cohort, we were more likely to successfully sequence HIV-2 p26 in faster progressors, which is probably due to the typically higher viral loads in these participants, and this also generated a bias in our sample of sequences. Furthermore, all participants were men, who are more likely to have faster disease progression (Peterson et al. 2011; Jespersen et al. 2016). In addition, the samples used in this study were stored for prolonged periods of time in resource-limited settings where storage at  $-80^{\circ}\text{C}$  was not always feasible. It is therefore possible that storage and handling conditions over the past few decades since collection might have decreased the levels of viral RNA in the samples we used, introducing added bias towards higher viral load samples. This means that although our intrahost diversity estimates within a single time point might have been affected by long-term storage and repeated freeze-thaw cycles, it is unlikely that the intrahost divergence and evolutionary rate estimates between time points will have been severely affected. Reassuringly, our results agreed with pre-existing evidence that evolutionary rates correlate with disease progression markers in HIV-2 infection, in that we observed significant levels of sequence diversity within time points, indicating that the sequence diversity within samples was preserved.

Our study highlights what may be a fundamental difference in intrahost evolution between HIV-1 and HIV-2, in that HIV-1 p24 is able to adapt to host immune responses, whereas HIV-2 p26 is more limited in this regard (Buggert et al. 2014; Norström et al. 2014). There are fewer antiretrovirals available to treat HIV-2 than HIV-1; our results indicate that the new generation of direct capsid inhibitors may be attractive options for ART in HIV-2-positive individuals (Yant et al. 2019; HIV-2 Infection | NIH).

## Data availability

All HIV-2 sequences used in this analysis are available on GenBank with accession numbers OL872372-872739 and OM146012. All R files and XML files used for this analysis are available from the authors on request.

## Supplementary data

Supplementary data are available at *Virus Evolution* online.

## Acknowledgements

The listed authors and the members of the Sweden Guinea-Bissau Cohort Research (SWEGUB CORE) group, including Babetida

N'Buna, Antonio Biague, Ansu Biai, Cidia Camara, Zacarias Jose da Silva, Joakim Esbjörnsson, Marianne Jansson, Sara Karlson, Jacob Lopatko Lindman, Patrik Medstrand, Fredrik Månsson, Hans Norrgren, Angelica A. Palm, Gülsen Özkaya Sahin, and Sten Wilhelmson are indebted to the staff of the Police Clinics and the National Public Health Laboratory in Bissau, Guinea-Bissau. We thank Matthew Cotten for providing critical feedback on this manuscript.

## Funding

This work was supported by funding from the Swedish Research Council (grant No. 2016-01417) and the Swedish Society for Medical Research (grant No. SA-2016). M.T.B. was supported by a Commonwealth Scholarship (ZACS-2016-943).

**Conflict of interest:** The authors declare no competing interests.

## Author contributions

M.T.B., S.R.J., and J.E. conceptualized and designed the study. M.T.B., S.R.J., and J.E. provided funding for the study. The SWEGUB CORE group provided samples from which new sequences used in the study were generated. M.T.B. performed laboratory work and inferential analyses and produced all figures and tables. J.N. assisted with phylogenetic modelling and data analysis. A.P. and S.K. processed samples and performed laboratory work. K.K. and K.M. performed the structural p26 modelling and interpreted the results. C.O., T.D.S., and A.J. collected the Caio p26 sequence and clinical data. M.T.B. and J.E. wrote the manuscript, and all the authors reviewed, edited, and approved the manuscript for submission.

## References

- Anglaret, X. et al. (1997) 'CD4+ T-lymphocyte Counts in HIV Infection: Are European Standards Applicable to African Patients?', *Journal of Acquired Immune Deficiency Syndromes & Human Retrovirology*, 14: 361–7.
- Asowata, O. E. et al. (2021) 'Irreversible Depletion of Intestinal CD4+ T-cells Is Associated with T-cell Activation during Chronic HIV Infection', *JCI Insight*, 6: e146162.
- Ayres, D. L. et al. (2012) 'BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics', *Systematic Biology*, 61: 170–3.
- Bello, G. et al. (2007) 'Plasma Viral Load Threshold for Sustaining Intrahost HIV Type 1 Evolution', *AIDS Research and Human Retroviruses*, 23: 1242–50.
- Biomatters. Geneious Prime User Manual. <<https://assets.geneious.com/documentation/geneious/GeneiousPrimeManual.pdf>>.
- Buggert, M. et al. (2014) 'Functional Avidity and IL-2/perforin Production Is Linked to the Emergence of Mutations within HLA-B\*5701-restricted Epitopes and HIV-1 Disease Progression', *The Journal of Immunology*, 192: 4685–96.
- Claiborne, D. T. et al. (2015) 'Replicative Fitness of Transmitted HIV-1 Drives Acute Immune Activation, Proviral Load in Memory CD4+ T Cells, and Disease Progression', *Proceedings of the National Academy of Sciences of the United States of America*, 112: E1480–9.
- de Silva, T. I. et al. (2018a) 'HLA-associated Polymorphisms in the HIV-2 Capsid Highlight Key Differences between HIV-1 and HIV-2 Immune Adaptation', *AIDS*, 32: 709–14.
- et al. (2018b) 'HLA-associated Polymorphisms in the HIV-2 Capsid Highlight Key Differences between HIV-1 and HIV-2 Immune Adaptation', *AIDS (London, England)*, 32: 709–14.

- et al. (2013a) 'Correlates of T-cell-mediated Viral Control and Phenotype of CD8+ T Cells in HIV-2, a Naturally Contained Human Retroviral Infection', *Blood*, 27: 4330–9.
- et al. (2013b) 'Population Dynamics of HIV-2 in Rural West Africa: Comparison with HIV-1 and Ongoing Transmission at the Heart of the Epidemic', *AIDS (London, England)*, 27: 125–34.
- Drummond, A. J. et al. (2012) 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29: 1969–73.
- Esbjörnsson, J. et al. (2012) 'Inhibition of HIV-1 Disease Progression by Contemporaneous HIV-2 Infection', *New England Journal of Medicine*, 367: 224–32.
- et al. (2019) 'Long-term Follow-up of HIV-2-related AIDS and Mortality in Guinea-Bissau: A Prospective Open Cohort Study', *The Lancet HIV*, 6: e25–31.
- Ganusov, V. V. et al. (2011) 'Fitness Costs and Diversity of the Cytotoxic T Lymphocyte (CTL) Response Determine the Rate of CTL Escape during Acute and Chronic Phases of HIV Infection', *Journal of Virology*, 85: 10518–28.
- Garcia-Knight, M. A. et al. (2016) 'Viral Evolution and Cytotoxic T Cell Restricted Selection in Acute Infant HIV-1 Infection', *Scientific Reports*, 6: 29536.
- HIV-2 Infection | NIH. <<https://clinicalinfo.hiv.gov/en/guidelines/adult-and-adolescent-arv/hiv-2-infection>> accessed, July 27 2021.
- Iyer, S. et al. (2015) 'Comparison of Major and Minor Viral SNPs Identified through Single Template Sequencing and Pyrosequencing in Acute HIV-1 Infection', *PLoS ONE*, 10: e0135903.
- Jacques, D. A. et al. (2016) 'HIV-1 Uses Dynamic Capsid Pores to Import Nucleotides and Fuel Encapsidated DNA Synthesis', *Nature*, 536: 349–53.
- Jallow, S. et al. (2015) 'The Presence of Prolines in the Flanking Region of an Immunodominant HIV-2 Gag Epitope Influences the Quality and Quantity of the Epitope Generated', *European Journal of Immunology*, 45: 2232–42.
- Jespersen, S. et al. (2016) 'Differential Effects of Sex in a West African Cohort of HIV-1, HIV-2 and HIV-1/2 Dually Infected Patients: Men are Worse Off', *Tropical Medicine & International Health*, 21: 253–62.
- Kanki, P. J. et al. (1994) 'Slower Heterosexual Spread of HIV-2 than HIV-1', *The Lancet*, 343: 943–6.
- Kassambara, A. 'Ggpubr: "Ggplot2" Based Publication Ready Plots'. R Package 2018, R package version 0.1.8.
- Kearse, M. et al. (2012) 'Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data', *Bioinformatics*, 28: 1647–9.
- Korber, B., and Myers, G. (1992) 'Signature Pattern Analysis: A Method for Assessing Viral Sequence Relatedness', *AIDS Research and Human Retroviruses*, 8: 1549–60.
- Lahaye, X. et al. (2018) 'NONO Detects the Nuclear HIV Capsid to Promote cGAS-Mediated Innate Immune Activation', *Cell*, 175: 488–501.e22.
- Leligdowicz, A. et al. (2007) 'Robust Gag-specific T Cell Responses Characterize Viremia Control in HIV-2 Infection', *The Journal of Clinical Investigation*, 117: 3067–74.
- Lemey, P. et al. (2007) 'Synonymous Substitution Rates Predict HIV Disease Progression as a Result of Underlying Replication Dynamics', *PLoS Computational Biology*, 3: e29.
- et al. (2012) 'A Counting Renaissance: Combining Stochastic Mapping and Empirical Bayes to Quickly Detect Amino Acid Sites under Positive Selection', *Bioinformatics (Oxford, England)*, 28: 3248–56.
- Leviyang, S., and Ganusov, V. V. (2015) 'Broad CTL Response in Early HIV Infection Drives Multiple Concurrent CTL Escapes', *PLoS Computational Biology*, 11: e1004492.
- Mamede, J. I. et al. (2017) 'Cyclophilins and Nucleoporins are Required for Infection Mediated by Capsids from Circulating HIV-2 Primary Isolates', *Scientific Reports*, 7: 45214.
- Martin, D., and Rybicki, E. (2000) 'RDP: Detection of Recombination Amongst Aligned Sequences', *Bioinformatics*, 16: 562–3.
- Martin, D. P. et al. (2015) 'RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes', *Virus Evolution*, 1: vev003.
- et al. (2005) 'A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints', *AIDS Research and Human Retroviruses*, 21: 98–102.
- Matreyek, K. A., and Engelman, A. (2011) 'The Requirement for Nucleoporin NUP153 during Human Immunodeficiency Virus Type 1 Infection Is Determined by the Viral Capsid', *Journal of Virology*, 85: 7818–27.
- Meyerson, N. R., and Sawyer, S. L. (2011) 'Two-stepping through Time: Mammals and Viruses', *Trends in Microbiology*, 19: 286–94.
- Mild, M. et al. (2013) 'High Inpatient HIV-1 Evolutionary Rate Is Associated with CCR5-to-CXCR4 Coreceptor Switch', *Infection, Genetics and Evolution*, 19: 369–77.
- et al. (2010) 'Differences in Molecular Evolution between Switch (R5 to R5X4/X4-tropic) and Non-switch (R5-tropic Only) HIV-1 Populations during Infection', *Infection, Genetics and Evolution*, 10: 356–64.
- Miyamoto, T. et al. (2011) 'A Single Amino Acid of Human Immunodeficiency Virus Type 2 Capsid Protein Affects Conformation of Two External Loops and Viral Sensitivity to TRIM5 $\alpha$ . Lee Y-M (Ed.)', *PLoS ONE*, 6: e22779.
- Norrgrén, H. et al. (2003) 'Clinical Progression in Early and Late Stages of Disease in a Cohort of Individuals Infected with Human Immunodeficiency Virus-2 in Guinea-Bissau', *Scandinavian Journal of Infectious Diseases*, 35: 265–72.
- Norström, M. M. et al. (2014) 'Baseline CD4+ T Cell Counts Correlates with HIV-1 Synonymous Rate in HLA-B\*5701 Subjects with Different Risk of Disease Progression', *PLOS Computational Biology*, 10: e1003830.
- Onyango, C. O. et al. (2010) 'HIV-2 Capsids Distinguish High and Low Virus Load Patients in a West African Community Cohort', *Vaccine*, 28: B60–7.
- Organization WH. (2007) *WHO Case Definitions of HIV for Surveillance and Revised Clinical Staging and Immunological Classification of HIV-Related Disease in Adults and Children*. Geneva, World Health Organization.
- Padidam, M., Sawyer, S., and Fauquet, C. M. (1999) 'Possible Emergence of New Geminiviruses by Frequent Recombination', *Virology*, 265: 218–25.
- Palm, A. A. et al. (2019) 'Low Postseroconversion CD4+ T-cell Level Is Associated with Faster Disease Progression and Higher Viral Evolutionary Rate in HIV-2 Infection', *mBio*, 10: e01245–18.
- Peterson, I. et al. (2011) 'Mortality and Immunovirological Outcomes on Antiretroviral Therapy in HIV-1 and HIV-2-infected Individuals in the Gambia', *AIDS*, 25: 2167–75.
- Posada, D., and Crandall, K. A. (2001) 'Evaluation of Methods for Detecting Recombination from DNA Sequences: Computer Simulations', *Proceedings of the National Academy of Sciences*, 98: 13757–62.
- (2002) 'The Effect of Recombination on the Accuracy of Phylogeny Estimation', *Journal of Molecular Evolution*, 54: 396–402.
- Price, A. J. et al. (2009) 'Active Site Remodeling Switches HIV Specificity of Antiretroviral TRIMCyp', *Nature Structural & Molecular Biology*, 16: 1036–42.
- Raghvani, J. et al. (2018) 'Evolution of HIV-1 within Untreated Individuals and at the Population Scale in Uganda', *PLoS Pathogens*, 14: e1007167.

- Rambaut, A. et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Susko E (Ed.)', *Systematic Biology*, 67: 901–4.
- Rihn, S. J. et al. (2013) 'Extreme Genetic Fragility of the HIV-1 Capsid', *PLoS Pathogens*, 9: e1003461.
- RStudio Team. (2021) *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC.
- Sauter, D., and Kirchhoff, F. (2019) 'Key Viral Adaptations Preceding the AIDS Pandemic', *Cell Host & Microbe*, 25: 27–38.
- Schaller, T. et al. (2011) 'HIV-1 Capsid-Cyclophilin Interactions Determine Nuclear Import Pathway, Integration Targeting and Replication Efficiency', *PLoS Pathogens*, 7: e1002439.
- Schrödinger, L. L. C. (2015) 'The PyMOL Molecular Graphics System, Version 1.8'.
- Sharp, P. M., and Hahn, B. H. (2011) 'Origins of HIV and the AIDS Pandemic', *Cold Spring Harbor Perspectives in Medicine*, 1: a006841.
- Skorupka, K. A. et al. (2019) 'Hierarchical Assembly Governs TRIM5 $\alpha$  Recognition of HIV-1 and Retroviral Capsids', *Science Advances*, 5: eaaw3631.
- Smith, J. M. (1992) 'Analyzing the Mosaic Structure of Genes', *Journal of Molecular Evolution*, 34: 126–9.
- Song, H. et al. (2007) 'A Single Amino Acid of the Human Immunodeficiency Virus Type 2 Capsid Affects Its Replication in the Presence of Cynomolgus Monkey and Human TRIM5 S', *Journal of Virology*, 81: 7280–5.
- Sousa, A. E. et al. (2002) 'CD4 T Cell Depletion Is Linked Directly to Immune Activation in the Pathogenesis of HIV-1 and HIV-2 but Only Indirectly to the Viral Load', *The Journal of Immunology*, 169: 3400–6.
- Suchard, M. A. et al. (2003) 'Hierarchical Phylogenetic Models for Analyzing Multipartite Sequence Data', *Systematic Biology*, 52: 649–64.
- Takeuchi, J. S. et al. (2013) 'High Level of Susceptibility to Human TRIM5 $\alpha$  Conferred by HIV-2 Capsid Sequences', *Retrovirology*, 10: 1742–4690.
- Theys, K. et al. (2018) 'The Impact of HIV-1 Within-host Evolution on Transmission Dynamics', *Current Opinion in Virology*, 28: 92–101.
- Troyer, R. M. et al. (2005) 'Changes in Human Immunodeficiency Virus Type 1 Fitness and Genetic Diversity during Disease Progression', *Journal of Virology*, 79: 9006–18.
- van der Loeff, M. F. S. et al. (2010) 'Undetectable Plasma Viral Load Predicts Normal Survival in HIV-2-infected People in a West African Village', *Retrovirology*, 7: 1742–4690.
- Visseaux, B. et al. (2016) 'Hiv-2 Molecular Epidemiology', *Infection, Genetics and Evolution*, 46: 233–40.
- Williamson, S. (2003) 'Adaptation in the Env Gene of HIV-1 and Evolutionary Theories of Disease Progression', *Molecular Biology and Evolution*, 20: 1318–25.
- Yant, S. R. et al. (2019) 'A Highly Potent Long-acting Small-molecule HIV-1 Capsid Inhibitor with Efficacy in A Humanized Mouse Model', *Nature Medicine*, 25: 1377–84.
- Yu, A. et al. (2020) 'TRIM5 $\alpha$  Self-assembly and Compartmentalization of the HIV-1 Viral Capsid', *Nature Communications*, 11: 1307.
- Zhao, G. et al. (2013) 'Mature HIV-1 Capsid Structure by Cryo-electron Microscopy and All-atom Molecular Dynamics', *Nature*, 497: 643–6.