Viral diversity from next-generation sequencing of HIV-1 samples provides precise estimates of infection recency and time since infection

Louisa A. Carlisle[1,2], Teja Turk[1,2*], Katharina Kusejko[1,2], Karin J. Metzner[1,2], Christine Leemann[1,2], Corinne Schenkel[1,2], Nadine Bachmann[1,2], Susana Posada[3,4], Niko Beerenwinkel[3,4], Jürg Böni[2,5], Sabine Yerly[6], Thomas Klimkait[7], Matthieu Perreau[8], Dominique L. Braun[1,2], Andri Rauch[9], Alexandra Calmy[6], Matthias Cavassini[10], Manuel Battegay[11], Pietro Vernazza[12], Enos Bernasconi[13], Huldrych F. Günthard[1,2]*, Roger D. Kouyos[1,2]*, and the Swiss HIV Cohort Study*

1 Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, 8091 Zurich, Switzerland

2 Institute of Medical Virology, University of Zurich, 8057 Zurich, Switzerland

3 Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland

4 SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland

5 Swiss National Center for Retroviruses, University of Zurich, 8057 Zurich, Switzerland

6 Laboratory of Virology and Division of Infectious Diseases, Geneva University Hospital, 1205 Geneva, Switzerland

7 Molecular Virology, Department of Biomedicine–Petersplatz, University of Basel, 4051 Basel, Switzerland

8 Division of Immunology and Allergy, Lausanne University Hospital, 1011 Lausanne, Switzerland

9 Department of Infectious Diseases, Bern University Hospital, University of Bern, 3010 Bern, Switzerland

10 Division of Infectious Diseases, Lausanne University Hospital, 1011 Lausanne, Switzerland

11 Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, 4031 Basel, Switzerland

12 Division of Infectious Diseases, Cantonal Hospital St Gallen, 9007 St. Gallen, Switzerland

13 Division of Infectious Diseases, Regional Hospital Lugano, 6900 Lugano, Switzerland

**These authors contributed equally to the manuscript*

Corresponding author:

Prof. Dr. Roger Kouyos

Division of Infectious Diseases and Hospital Epidemiology

University Hospital Zurich, Rämistrasse 100, CH-8091 Zürich

+41 44 255 36 10, roger.kouyos@usz.ch

Alternate corresponding author:

Prof. Dr. Huldrych Günthard

Division of Infectious Diseases and Hospital Epidemiology

University Hospital Zurich, Rämistrasse 100, CH-8091 Zürich

+41 44 255 34 50, huldrych.guenthard@usz.ch

Short summary:

With mean absolute error below one year HIV genetic diversity derived from NGS sequencing is both superior estimator of time since infection and superior classifier of infection recency compared to the genetic diversity calculated from Sanger sequencing.

Abstract

Background

HIV-1 genetic diversity increases over the course of infection, and can be used to infer time since infection (TSI) and consequently also infection recency, crucial quantities for HIV-1 surveillance and the understanding of viral pathogenesis.

Methods

We considered 313 HIV-infected individuals for whom reliable estimates of infection dates and next-generation sequencing (NGS)-derived nucleotide frequency data were available. Fraction of ambiguous nucleotides (FAN) obtained by population sequencing were available for 207 samples. We assessed whether average pairwise diversity (APD) calculated using NGS sequences provided a more exact prediction of TSI and classification of infection recency (<1 year post-infection) compared to FAN.

Results

NGS-derived APD classifies an infection as recent with a sensitivity of 88% and specificity of 85%. When considering only the 207 samples for which FAN were available, NGS-derived APD exhibited a higher sensitivity (90% vs 78%) and specificity (95% vs 67%) than FAN. Additionally, APD can estimate TSI with a mean absolute error of 0.84 years, compared to 1.03 years for FAN.

Conclusions

Viral diversity from NGS data is more precise than that from population sequencing in its ability to predict infection recency, and provides an estimated TSI with a mean absolute error of below one year.

Keywords: HIV-1, next-generation sequencing, diversity, infection recency, time since infection

Introduction

The time since infection (TSI) of HIV-positive patients is of key importance for the study of viral pathogenesis and epidemiology as well as for clinical purposes, yet is often unknown. At least 10% of people infected with HIV-1 do not experience clear symptoms during primary infection [1], and symptoms are not specific to HIV-1 so can be misidentified [2–4]. For chronically-infected patients presenting later in infection, identifying the likely transmission event is mostly infeasible. A related unknown is whether a patient has a recent infection, defined as a TSI below one year. These patients are important to identify for both research and public health purposes, as they show increased transmission rates [5–7], and recent infections inform incidence assays. Identifying recent infections may also be useful for targeting key groups in prevention strategies [8], specific populations of patients for cure research [9], and for treatment simplification [10].

It is well established that HIV-1 viral diversity increases over time within an infected individual [11–14], and thus it should be possible to use diversity as a measure for TSI. We previously studied data from Sanger population sequencing [15], which is typically used for genotypic drug resistance testing. The fraction of ambiguous nucleotide calls in these sequences was used as a measure for diversity and established as a predictor of recent infection, and has been subsequently validated [16,17]. However, population sequencing can only detect minor variants at frequencies above 20% [18], limiting its precision. Additionally, defining a nucleotide call as ambiguous depends on the interpretation of the semi-quantitative chromatogram data and individual laboratory set-up, which may introduce biases or inconsistencies [19].

Next-generation sequencing (NGS) is steadily replacing Sanger sequencing for genotypic resistance testing, and as such we can expect an increase in the availability of HIV NGS sequences from many patients in the coming years. NGS can detect variants down to a frequency of around 1% [20], and hence potentially provide more precise information on infection dates. Variants are reported along with their frequency in the sample, rather than a position simply being marked as ambiguous, which increases the amount of quantitative information returned from NGS sequencing compared to Sanger population sequencing. Puller et al. [21] present a simple method based on NGS sequences for predicting TSI from viral diversity, using average pairwise diversity (APD) as a predictor, and showed a good correlation of NGS-diversity with TSI.

Here, we analyse 331 NGS-sequenced samples from two Swiss HIV cohorts and assess the utility of the APD to estimate infection recency and TSI. In particular, we aim to compare the accuracy of these estimates with those derived from population sequencing data.

Materials and Methods

Patients

We considered samples from the Swiss HIV Cohort Study (SHCS) [22] and Zurich Primary HIV Infection Study (ZPHI). The SHCS is highly representative of the HIV-1 epidemic in Switzerland and includes broad, in-depth and high quality genetic, biological, clinical and demographic data. The ZPHI is a largely overlapping, smaller cohort comprising patients diagnosed during primary infection, providing us with well-characterised dates of infection. Blood samples are collected at HIV diagnosis and on occasion at later time points. Our full sample set consisted of 331 samples from 313 HIV-positive individuals from the SHCS and ZPHI cohorts, who fulfilled the following criteria: had at least one ART-naïve NGS sequenced sample, with coverage above 100 reads per base over at least 50% of both *gag* and *pol* third codon positions; and had a precise date of infection, defined as being one of the following (see also table 1 and supplementary materials 1):

- Patients enlisted in the ZPHI predominantly have an estimated date of infection recorded, which has a high degree of certainty. Those with some uncertainty were required to have no more than one year between the recorded date of infection and the earliest or latest possible infection dates, as estimated by the physician.
- Within the SHCS, a recorded primary infection indicates that a patient became infected within three months prior to the recorded date of diagnosis. For these patients, we therefore estimated the date of infection as the date of diagnosis minus 45 days.
- The remaining patients were required to have maximally 2 years between the last negative and first positive HIV-1 test, and the date of infection was taken as the midpoint between these dates, as used by [21].

We identified a subset of 207 samples from 206 individuals for which ambiguous-nucleotide scores [15] were available from the same sample times, to allow the two methods to be compared.

Sequencing

Next-generation sequencing using Illumina technology

Whole genome sequencing was performed in the context of previous studies, using the following protocol: HIV-1 RNA from patient plasma was isolated, reverse transcribed, amplified, and sequenced as described previously [20,23]. When the first pan-PCR was unsuccessful, semi-/nested PCRs were performed using the primers listed in supplementary table 1. Samples were sequenced using the MiSeq Reagent Kit v2 (500 cycles) (Illumina).

Frequencies of minority variants were obtained from V-pipe [24], which filtered and aligned the reads against the HIV-1 HXB2 genome (GenBank accession number K03455) using the BWA-MEM aligner. Default options were used except for running the additional rule "minor_variants". This returns the frequency of any minority variants detected in the sample, at all positions with a minimum coverage of 100 reads, along with the coverage at each position.

Population sequencing using the Sanger technology

We used previously calculated ambiguous nucleotide scores, which had been derived from routine genotypic HIV-1 drug resistance testing performed by population sequencing, as previously described [15]. Sequences covered the partial *pol* region, namely, protease and a minimum of codons 28-225 of reverse transcriptase.

Diversity score calculation

We calculated the APD over the third codon positions of the *gag* or *pol* regions using equation 1 (from [21]). We focussed on these regions based on the findings from [21] and preliminary analyses, which indicated that the steadiest accumulation of mutations occurs over *gag* and *pol* third codon positions.

$$\mathrm{APD} = \frac{1}{L}\sum_{i=1}^{L}\Theta\left(1 - x_i^m - x_c\right)\left[\sum_{\alpha} x_{i\alpha}\left(1 - x_{i\alpha}\right)\right] \qquad (1)$$

Equation 1 first determines whether sequence position i has any diversity, namely if the sum of the frequency of minor variants $(1 - x_i^m)$ is above the cut-off x_c. If this is the case, it sums the diversity contribution $x_{i\alpha}(1 - x_{i\alpha})$ of each variant $\alpha \in \{A, C, G, T, deletion\}$ at that position, where $x_{i\alpha}$ denotes the frequency of variant α at position i. Otherwise, i.e., if the sum of minority variants is below the cut-off x_c, position i is assigned zero diversity by function Θ. Finally, the diversity over all positions from 1 to the sequence length L is averaged. The cut-off x_c is necessary to remove sequencing errors from the calculation, so was set to 1% as this is the approximate detection limit of Illumina. This calculation is functionally equivalent to the average fraction of positions at which two randomly drawn sequences differ.

In total, we assessed the following three diversity scores:

APD *gag* Average pairwise diversity calculated over third codon positions in *gag* using Illumina sequence data.

APD *pol* Average pairwise diversity calculated over third codon positions in *pol* using Illumina sequence data.

FAN Fraction of ambiguous nucleotides calculated from Sanger population sequencing data of partial *pol* regions [15].

Data analysis

Analysis of the data was performed in R 3.3.2 [25], using the packages data.table [26], pROC [27], readstata13 [28], RColorBrewer [29], inctools [30], and DescTools [31]. We evaluated the predictability of an infection as being recent (infected for <1 year) or chronic based on viral diversity, using receiver operator characteristics (ROC) analyses and mean duration of recent infection (MDRI) [32] with recent infection as the positive outcome. For this recency analysis, we restricted our sample set to the 317 samples from 299 patients that could be clearly classified as recent or chronic due to their window of uncertainty being entirely below or above the 1 year definition of recency. We compared the classification abilities of our two NGS-derived diversity scores, and that of the ambiguous nucleotides score.

We estimated TSI using a linear model (equation 2), with the model coefficients β and α calculated via linear regression.

$$\text{Estimated TSI} = \beta\text{D} + \alpha \qquad (2)$$
$$where\ \text{D} = diversity\ score$$

We used leave-one-out cross-validation to assess the validity of this model, with the mean absolute error (MAE) as the primary outcome. Specifically, we cycled through all the samples, assigning one at a time as the 'test' sample and calculated the model coefficients using all the remaining samples. For each test sample, we then took the absolute value of the difference between the estimated and the actual TSI, and calculated the average of this across all samples. This provided a simple summary statistic that we used to compare different diversity measures and models. We also compared the performance of our optimal model coefficients to those suggested by Puller et al. [21].

We conducted an outlier analysis by fitting an asymptotic curve to the data and used it to define potential outlier samples with unusually high APD scores. An asymptotic fit was chosen to reflect the eventual saturation of viral diversity over time [14], and to provide a clear cut-off above which samples could be considered as possible outliers exhibiting a diversity too high to be consistent with within-host evolution. Specifically, we fitted the following model:

$$\text{APD} = \gamma - \eta\lambda^{\text{TSI}} \qquad (3)$$

where γ, η, and λ are free parameters. Samples with APD scores above the asymptote were taken as potential outliers.

Ethics approval and consent to participate

The SHCS was approved by the ethics committees of the participating institutions (Kantonale Ethikkommission Bern, Ethikkommission des Kantons St. Gallen, Comité Départemental d'Éthique des Spécialités Médicales et de Médicine Communataire et de Premier Recours, Kantonale Ethikkommission Zürich, Repubblica et Cantone Ticino–Comitato Ethico Cantonale, Commission Cantonale d'Éthique de la Recherche sur l'Être Humain, Ethikkommission beider Basel for the SHCS and Kantonale Ethikkommission Zürich for the ZPHI). Written informed consent was obtained from all participants.

Results

We performed ROC analyses to quantify the ability of APD to correctly classify infections as being recent or chronic (figure 1). The sensitivity and specificity were evaluated using an a priori-defined APD cut-off of <0.01 to classify samples as recent. Both APD *gag* and APD *pol* resulted in high ROC areas under the curve (AUC), at 0.92 and 0.93 respectively (figure 1a). Our cut-off of 0.01 gave high sensitivities and specificities over both regions: 88% and 85% respectively over *gag*, and 87% and 85% over *pol*. Note that by setting a more conservative diversity cut-off, a specificity of ≥99% with sensitivity >78% can be achieved for APD calculated over *pol* (figure 1a). To mitigate the effect of the TSI distribution on ROC analysis, we moreover derived the MDRI against false recency rate (FRR) profile. Whereas the pre-defined cut-off of 0.01 yielded MDRI and FRR 0.81 years and 13% respectively over *gag*, and 0.82 years and 15% over *pol*, a cut-off of 0.006 for APD over *pol* reduces the FRR below 2% while keeping the MDRI above the targeted 0.5 years (figure 1c). The NGS-based diversity is therefore a strong classifier of infection recency.

We compared the classification ability of APD to the fraction of ambiguous nucleotides (FAN), using the subset of 197 samples from patients with NGS and population sequencing data available from the same time-points. Consistent with the complete dataset, we found that the AUCs for *gag* and *pol* were very high, 0.95 in both cases, whilst the AUC for population sequencing was clearly lower at 0.77 (figure 1b). NGS data also provided higher specificities and sensitivities than population sequencing (figure 1b; the FAN cut-off of 0.005 ambiguous positions was taken from [15]). Results were similar when comparing FAN to APD calculated over the same partial *pol* region (supplementary figure 1), showing that the length of the sequences was not a major confounding factor. Additionally, the NGS-derived diversity based recency assays outperformed the FAN-based classification with respect to area under the MDRI versus FRR curve (figure 1d). Diversity measured from NGS data is thus a superior classifier to diversity measured from population sequencing.

We used linear regression to determine the association of APD and TSI and to find the optimal coefficients for estimating TSI (table 2). R^2 was 0.38 for APD *gag*, and 0.27 for APD *pol*. The same trends were seen in the subset of samples that we could compare to the ambiguous nucleotides method, with R^2 values being 0.31 for

APD gag, 0.26 for APD *pol*, and lower (0.18) for FAN. We proceeded to study the predictive power of APD in a linear model. Leave-one-out cross validation was used to calculate the MAE across all samples, and so compare different diversity measures and models. We found that the MAE in the estimated TSI was less than one year, being 0.84 years for APD *gag* and 0.92 years for APD *pol*. The solid lines in figure 2 show the resulting linear models. We also compared our model coefficients to those suggested by [21] (figure 2, dashed lines), which resulted in higher errors than our models did (MAE of 1.31 years for APD *gag* and 1.40 years for APD *pol*).

Prediction strength of NGS data can be contrasted with that of population sequencing data. In doing so, we found that APD predicts TSI more precisely than the fraction of ambiguous nucleotides. We repeated our regression analyses on the ambiguous nucleotide comparison sample subset (figure 3). The MAE remained strongly below one year for the NGS-derived diversity scores (0.85 years for APD *gag* and 0.91 years for APD *pol*), whilst the population sequencing-derived scores gave a MAE of 1.03 years, showing a marginally lower ability to predict TSI.

We observed a few potential outliers that have very high diversity scores, which is inconsistent with them being the result of purely within-host diversification, and is suggestive of superinfection [33]. We identified samples that could be considered as suspected outliers by fitting the asymptotic curve from Equation 3 to the data. We found that for APD *gag* and *pol* respectively, six and 14 samples lay above the diversity saturation point as defined by the asymptote γ of the fitted curve (figure 4). We then defined the intersecting five samples that lay above the asymptotes for both *gag* and *pol* as the outlier samples. Removing them increased the R^2 for APD *gag* to 0.44, but the R^2 of APD *pol* remained at 0.27. The MAEs for both *gag* and *pol* were reduced, to 0.76 years and 0.85 years respectively. Samples with unusually high diversity scores therefore have a negative effect on the accuracy of this method.

Discussion

We studied the predictability of an HIV-1 infection as being recent or chronic based on viral genetic diversity using NGS sequences, and compared this to the predictability based on the fraction of ambiguous nucleotides based on Sanger population sequencing. We find that APD over either *gag* or *pol* third codon positions yields a higher precision for predicting recent infections. We primarily attribute this to the increased information available from NGS over population sequencing, due to more sensitive detection and quantification of single minority variants.

We applied a linear model to estimate the TSI from diversity scores, comparing the performance of APD calculated over *gag* and *pol*, and fraction of ambiguous nucleotides. We found the smallest MAE in TSI estimation for APD calculated over *gag*. Estimation of TSI from the fraction of ambiguous nucleotides showed a lower precision. We also compared our linear models to those identified by [21]. Although the latter do not perform as well as ours, they yielded good estimates of TSI. Thus, given that our samples extend to almost 19 years post-infection, this is an impressive external validation of [21] on a larger and broader independent sample set. The ability to infer a date of infection long beyond the first year makes this technique stand out among the multiple studies and methods that primarily look at inferring infection recency, and is information that is applicable to many research and epidemiological methods beyond incidence assays.

Noting that a few samples had very high diversity scores, we applied an outlier analysis to investigate the effect of removing these samples. This was based on the premise that such high scores were biologically implausible for a single viral population undergoing within-host diversification, and as such, these samples may reflect superinfection occurrences [33]. Moreover, although we lack detailed information for four of these five samples, one of them originates from a patient who had superinfection confirmed by other methods [34], supporting this hypothesis. The MAE decreased by almost 0.1 years for both APD *gag* and APD *pol*. These findings therefore suggest that samples with unusually high diversity, i.e. APD scores above approximately 0.04, should be considered for removal from such analyses as they may reflect a superinfection and reduce the accuracy of this method. Further work should include an analysis of these outliers to verify potential superinfection or identify alternative confounding factors.

One shortcoming of the APD method is that a small subset of samples from early infections have high APD scores, and so their TSI is overestimated. These samples likely come from infections with multiple founders, which increases the mean population diversity score despite the individual sub-populations arising from each founder having low diversities. Constructing haplotypes from the NGS data and considering the distribution of diversity within samples may help, as Park et al. [35,36] showed using the Hamming distance. However, constructing full-length haplotypes from HIV-1 Illumina sequences is highly challenging [37] and not validated as a tool to reconstruct HIV-1 diversity within an infected patient. It is therefore not currently clear whether such an approach would provide advantages in estimating TSI. Many other approaches that are capable of distinguishing multiply-founded infection require specialised or more complicated sequencing processes such as single genome amplification [38,39], or samples from multiple time-points [40,41]. As such, they could not be applied to NGS sequences generated for routine resistance testing, limiting their wider applicability particularly retrospectively on routinely generated clinical data. Conversely, the approach presented here provides infection dates and recency estimates as a free by-product of current or near-future routine clinical care. We therefore believe that the ability of this method to be applied to such sequences will outweigh the limitations of the technique in many settings.

An alternate approach to address this shortcoming would be combining the APD with other serological, biomarker and/or epidemiological information into a multi-assay algorithm [42–44]. These are typically constructed to provide a binary classification (recently or chronically infected). However, one can imagine an approach for estimating TSI either by combining multiple factors into a single equation, or using one factor to provide a window of likely TSI and subsequent factors to narrow down the possible TSI further within that window. Such assays would have to be evaluated to see not only if they can improve on the accuracy of predictions provided, but whether they provide a significant enough increase to justify the additional information and calculations required.

APD calculated from standard NGS sequences additionally contains some inherent error from the PCR amplification process [45]. Whilst techniques such as the use of primer-IDs have been developed to overcome this [40], this requires an additional step which is unlikely to be conducted in the context of a routine clinical setting. Noteworthy, the superior accuracy of the NGS-based diversity estimator over the Sanger sequencing-based approach indicates that the additional information from NGS sequences outweighs the higher error rate at the level of individual reads.

A limitation of this study is that whilst we included a large sample size obtained from well-documented cohorts, our true TSI dates are still estimates based on the available data, and so have some inherent error that we did not quantify or include in our analyses. However, we found the same trends when restricting the sample set to patients who only had a maximum of six months of uncertainty for the date of infection (supplementary figures 2 and 3), suggesting that this did not impact our results greatly. A further limitation arises from our samples originating from Switzerland-based cohorts, and therefore having fairly homogenous patient and viral characteristics. Further studies should therefore be conducted to assess the performance of this measure in other populations and other viral subtypes, before we can declare the broader applicability of APD as a proxy for TSI.

In conclusion, we have shown the utility of APD as a measure of diversity and tool for estimating TSI. We have provided an external validation of the TSI-estimator from Puller et al. [21]; a novelty in over basic recency categorisation methods. Additionally, we could show that APD provides accurate estimates of infection recency with specificity and sensitivity above 85%, and demonstrated the superiority of this NGS-derived diversity measure over the fraction of ambiguous nucleotides detected by Sanger population sequencing. With increasing ease of sequencing and decreasing costs NGS sequencing is becoming more commonplace for resistance testing to monitor transmitted resistance due to increasing migration (especially from the resource-limited settings with considerable prevalence of HIV resistance), and thus the data will become readily available for uses such as the method presented here.

Acknowledgements

We thank the patients who participate in the Zurich Primary HIV Infection Study and in the Swiss HIV Cohort Study; the physicians and study nurses, for excellent patient care; the resistance laboratories, for high-quality genotyping drug resistance testing; SmartGene (Zug, Switzerland), for technical support; Alexandra Scherrer, Susanne Wild, Anna Traytel from the SHCS data center for data management, Danièle Perraudin, Mirjam Minichiello and Marianne Amstutz for administration. The members of the Swiss HIV Cohort Study include the following:

Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Günthard HF (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert C, Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Rudin C (Chairman of the Mother & Child Substudy), Scherrer AU (Head of Data Centre), Schmid P, Speck R, Stöckle M, Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S.

Funding

This work was supported by the Swiss National Science Foundation [grant number BSSGI0_155851]. HFG was supported by SNF grant 179571. Furthermore, this study has been financed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation [grant number 148522], by the SHCS research foundation by the Yvonne Jacob Foundation (to HFG), by the clinical research priority program of the University of Zurich "Viral infectious diseases, ZPHI" (to HFG), and by the SystemsX.ch grant HIVX (to HFG). HFG has received an unrestricted research grant from Gilead to the SHCS Research Foundation. V-pipe is funded by the SIB Swiss Institute of Bioinformatics as an SIB Competitive Resource.

Conflicts of interests

HFG has received unrestricted research grants from Gilead Sciences and Roche; fees for data and safety monitoring board membership from Merck; and consulting/advisory board membership fees from Gilead

Sciences, Sandoz and Mepha. MB has received research or educational grants by Abb Vie AG, Gilead Sciences Switzerland Sàrl, Janssen-Cilag AG, MSD Merck Sharp & Dohme AG and ViiV Healthcare GmbH. EB has received fees for his institution for participation to advisory board from MSD, Gilead Sciences, ViiV Healthcare, Abbvie and Janssen. MC has received research and travel grants for his institution from ViiV and Gilead. AC has received unrestricted educational and research grants from MSD, Gilead and ViiV. AR reports support to his institution for advisory boards and/or travel grants from Janssen-Cilag, MSD, Gilead Sciences, Abbvie, and Bristol-Myers Squibb, and an unrestricted research grant from Gilead Sciences. All remuneration went to his home institution and not to AR personally, and all remuneration was provided outside the submitted work. KJM has received travel grants and honoraria from Gilead Sciences, Roche Diagnostics, GlaxoSmithKline, Merck Sharp & Dohme, Bristol-Myers Squibb, ViiV and Abbott; and the University of Zurich received research grants from Gilead Science, Roche, and Merck Sharp & Dohme for studies that Dr Metzner serves as principal investigator, and advisory board honoraria from Gilead Sciences. RDK has received speaker honoraria and travel grants from Gilead Sciences, none of which are in relation with the submitted manuscript.

Presented in part: AIDS, July 2018, Amsterdam, The Netherlands (abstract number THPEC197)

Corresponding author:

Prof. Dr. Roger Kouyos

Division of Infectious Diseases and Hospital Epidemiology

University Hospital Zurich, Rämistrasse 100, CH-8091 Zürich

+41 44 255 36 10, roger.kouyos@usz

1. Schacker T, Collier AC, Hughes J, Shea T, Corey L. Clinical and epidemiologic features of primary HIV infection. Annals of Internal Medicine. **1996**; 125(4):257–264.

2. Robb ML, Eller LA, Kibuuka H, et al. Prospective Study of Acute HIV-1 Infection in Adults in East Africa and Thailand. N Engl J Med. **2016**; 374(22):2120–2130.

3. Hecht FM, Busch MP, Rawal B, et al. Use of laboratory tests and clinical symptoms for identification of primary HIV infection. AIDS. **2002**; 16(8):1119–1129.

4. Braun DL, Kouyos RD, Balmer B, Grube C, Weber R, Günthard HF. Frequency and spectrum of unexpected clinical manifestations of primary HIV-1 infection. Clinical Infectious Diseases. **2015**; 61(6):1013–1021.

5. Marzel A, Shilaih M, Yang W-L, et al. HIV-1 Transmission During Recent Infection and During Treatment Interruptions as Major Drivers of New Infections in the Swiss HIV Cohort Study. Clin Infect Dis. **2016**; 62(1):115–122.

6. Volz EM, Ionides E, Romero-Severson EO, Brandt M-G, Mokotoff E, Koopman JS. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. PLOS Medicine. **2013**; 10(12):e1001568.

7. Braun DL, Kouyos R, Oberle C, et al. A novel Acute Retroviral Syndrome Severity Score predicts the key surrogate markers for HIV-1 disease progression. PLoS ONE. **2014**; 9(12):e114111.

8. Turk T, Bachmann N, Kadelka C, et al. Assessing the danger of self-sustained HIV epidemics in heterosexuals by population based phylogenetic cluster analysis. Elife. **2017**; 6.

9. Schmid A, Gianella S, Wyl V von, et al. Profound depletion of HIV-1 transcription in patients initiating antiretroviral therapy during acute infection. PLoS ONE. **2010**; 5(10):e13310.

10. Braun DL, Turk T, Tschumi F, et al. Non-inferiority of simplified dolutegravir monotherapy compared to continued combination

antiretroviral therapy that was initiated during primary HIV infection: a randomized, controlled, multi-site, open-label, non-inferiority trial. Clinical Infectious Diseases [Internet]. **2019**; . Available from: https://dx.doi.org/10.1093/cid/ciy1131

11. Troyer RM, Collins KR, Abraha A, et al. Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression. J Virol. **2005**; 79(14):9006–9018.

12. Domingo E, Holland J. RNA virus mutations and fitness for survival. Annual Reviews in Microbiology. **1997**; 51(1):151–178.

13. Tebit DM, Nankya I, Arts EJ, Gao Y. HIV diversity, recombination and disease progression: how does fitness "fit" into the puzzle. AIDS Rev. **2007**; 9(2):75–87.

14. Shankarappa R, Margolick JB, Gange SJ, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol. **1999**; 73(12):10489–10502.

15. Kouyos RD, Wyl V von, Yerly S, et al. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clin Infect Dis. **2011**; 52(4):532–539.

16. Ragonnet-Cronin M, Aris-Brosou S, Joanisse I, et al. Genetic diversity as a marker for timing infection in HIV-infected patients: evaluation of a 6-month window and comparison with BED. J Infect Dis. **2012**; 206(5):756–764.

17. Andersson E, Shao W, Bontell I, et al. Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. Infect Genet Evol. **2013**; 18:125–131.

18. Günthard HF, Wong JK, Ignacio CC, Havlir DV, Richman DD. Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide sequencing of HIV type 1 pol from clinical samples. AIDS Research and Human Retroviruses. **1998**; 14(10):869–876.

19. Huang DD, Eshleman SH, Brambilla DJ, Palumbo PE, Bremer JW. Evaluation of the editing process in human immunodeficiency virus type 1 genotyping. J Clin Microbiol. **2003**; 41(7):3265–3272.

20. Giallonardo FD, Töpfer A, Rey M, et al. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. **2014**; 42(14):e115.

21. Puller V, Neher R, Albert J. Estimating time of HIV-1 infection from next-generation sequence diversity. PLOS Computational Biology. **2017**; 13(10):e1005775.

22. Swiss HIV Cohort Study, Schoeni-Affolter F, Ledergerber B, et al. Cohort profile: the Swiss HIV Cohort study. Int J Epidemiol. **2010**; 39(5):1179–1189.

23. Gall A, Ferns B, Morris C, et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol. **2012**; 50(12):3838–3844.

24. Seifert D, Posada C'espedes S, Beerenwinkel N. V-pipe [Internet]. GitHub; 2017. Available from: https://github.com/cbg-ethz/V-pipe

25. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: http://www.R-project.org/

26. Dowle M, Srinivasan A. data.table: Extension of `data.frame` [Internet]. 2017. Available from: https://CRAN.R-project.org/package=data.table

27. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. **2011**; 12:77.

28. Garbuszus JM, Jeworutzki S. readstata13: Import "Stata" Data Files [Internet]. 2017. Available from: https://CRAN.R-project.org/package=readstata13

29. Neuwirth E. RColorBrewer: ColorBrewer Palettes [Internet]. 2014. Available from: https://CRAN.R-project.org/package=RColorBrewer

30. Welte A, Grebe E, McIntosh A, et al. inctools: Incidence Estimation Tools [Internet]. 2018. Available from: https://CRAN.R-project.org/package=inctools

31. Signorell A. DescTools: Tools for Descriptive Statistics [Internet]. 2019. Available from: https://CRAN.R-project.org/package=DescTools

32. Kassanjee R, McWalter TA, Bärnighausen T, Welte A. A new general biomarker-based incidence estimator. Epidemiology. **2012**; 23(5):721–728.

33. Cornelissen M, Jurriaans S, Kozaczynska K, et al. Routine HIV-1 genotyping as a tool to identify dual infections. AIDS. **2007**; 21(7):807–811.

34. Chaudron S, Metzner K, Marzel A, et al. HIV-1 superinfection in the Swiss HIV Cohort Study: a large scale screen [abstract]. Proceedings of CROI. International Antiviral Society-USA; 2018.

35. Park SY, Love TM, Nelson J, Thurston SW, Perelson AS, Lee HY. Designing a genome-based HIV incidence assay with high sensitivity and specificity. AIDS (London, England). **2011**; 25(16):F13.

36. Park SY, Goeken N, Lee HJ, Bolan R, Dubé MP, Lee HY. Developing high-throughput HIV incidence assay with pyrosequencing platform. Journal of Virology. **2013**; :JVI–03128.

37. Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Frontiers in Microbiology. **2012**; 3:329.

38. Novitsky V, Moyo S, Wang R, Gaseitsiwe S, Essex M. Deciphering Multiplicity of HIV-1C Infection: Transmission of Closely Related Multiple Viral Lineages. PLoS ONE. **2016**; 11(11):e0166746.

39. Xia X-Y, Ge M, Hsi JH, others. High-accuracy identification of incident HIV-1 infections using a sequence clustering based diversity measure. PloS one. **2014**; 9(6):e100081.

40. Dennis AM, Zhou S, Sellers CJ, et al. Using Primer-ID Deep Sequencing to Detect Recent Human Immunodeficiency Virus Type 1 Infection. J Infect Dis. **2018**; 218(11):1777–1782.

41. Park SY, Love TMT, Kapoor S, Lee HY. HIITE: HIV-1 incidence and infection time estimator. Bioinformatics. **2018**; 34(12):2046–2052.

42. Brookmeyer R, Konikoff J, Laeyendecker O, Eshleman SH. Estimation of HIV incidence using multiple biomarkers. Am J Epidemiol. **2013**; 177(3):264–272.

43. Laeyendecker O, Brookmeyer R, Cousins MM, et al. HIV incidence determination in the United States: a multiassay approach. The Journal of Infectious Diseases. **2012**; 207(2):232–239.

44. Cousins MM, Konikoff J, Laeyendecker O, et al. HIV diversity as a biomarker for HIV incidence estimation: including a high resolution melting diversity assay in a multi-assay algorithm. Journal of Clinical Microbiology. **2013**; :JCM–02040.

45. Di Giallonardo F, Zagordi O, Duport Y, et al. Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. PLoS ONE. **2013**; 8(9):e74249.

Tables

Table 1: Sample and patient characteristics

Measure		Samples[a]		Patients	
		No.	%	No.	%
Total		331		313	
Gender of patient	Female	27	8	27	9
	Male	304	92	286	91
Ethnicity of patient	Asian	4	1	4	1
	Black	13	4	13	4
	Hispano-American	24	7	24	8
	White	284	86	267	85
	Other/unknown	6	2	5	2
Risk group of patient	HET	57	17	56	18
	IDU	12	4	12	4
	MSM	240	73	224	72
	Other/unknown	22	7	21	7
Age of patient when sample taken	Minimum	19	-	19	-
	1st quartile	30	-	30	-
	Median	35	-	35	-
	3rd quartile	42	-	43	-
	Maximum	79	-	79	-
Subtype of virus	A	8	2	8	3
	B	210	63	195	62
	C	6	2	6	2
	D	1	0	1	0
	F	3	1	3	1
	G	4	1	4	1
	01_AE	17	5	17	5
	02_AG	5	2	5	2
	Other/unknown	77	23	74	24
Time since infection at sample date	< 6 months	257	78	-	-
	6—12 months	15	5	-	-
	12—18 months	10	3	-	-
	18—24 months	5	2	-	-
	24—48 months	22	7	-	-
	≥ 48 months	22	7	-	-
Source of TSI information	ZPHI	256	77	238	76
	of which, no upper or	*134*	*52*	127	41

	Primary infection in SHCS	16	5	16	5
	Midpoint positive negative	59	18	59	19
Clearly recent or chronic	Yes	317	96	299	96
	No	14	4	14	4

[a] Twelve patients had two longitudinal samples available, and three patients had three longitudinal samples available

Table 2: Optimal coefficients for estimating TSI from APD using equation 2, as found in our analyses

	Full sample set		**Subset of samples**			**Full set minus outliers**	
	n = 331		n = 207			n = 326	
	APD *gag*	APD *pol*	APD *gag*	APD *pol*	FAN	APD *gag*	APD *pol*
β	142.8	117.8	136.9	124.8	159.2	188.6	122.9
α	-0.0662	0.0809	-0.0433	-0.0123	0.179	-0.308	0.0554

Figure legends

Figure 1: **Average pairwise diversity (APD) provides a good classifier of infection recency.** Receiver operator characteristics (ROC) curves (upper row) and mean duration of recent infection (MDRI) against false recency rate (FRR) curve (lower two panels), with a recent infection defined as being less than one year post infection and being taken as the positive outcome. Each line corresponds to the indicated diversity measure. Black dots on the curves show the diversity score cut-off for that curve, with corresponding specificities and sensitivities or FRR and MDRI in brackets. AUC = area under the curve, FAN = fraction of ambiguous nucleotides. a&c) Classification abilities of APD over *gag* and *pol*, with all 317 samples included for which time since infection could be clearly defined as recent or chronic. b&d) Comparison of the classification ability of NGS-derived diversity score with ambiguous nucleotides from population sequencing. Sample size was restricted to the 197 NGS-sequenced samples that had a corresponding ambiguous nucleotide score from the same time point.

Figure 2: **APD correlates well with time since infection (TSI).** TSI against APD scores over third codon positions in *gag*, and *pol (upper panels) and the same data with log-transformed TSI (lower panels)*. Upper panels: The calculated linear regression models are shown as the solid lines; Puller et al.'s [21] linear models are shown as the dashed lines. These models were then used to predict TSI from the diversity score. All 331 samples were included. A further qualitative observation is that a large proportion of samples that don't follow the general trend are those that have relatively high APD scores despite being sampled early during infection. These are suspected to be infections founded by multiple virions; a phenomenon known to cause problems for inferring TSI from average viral diversity measures [30] (see discussion). Another smaller group of potential outlier samples are those with very high APD scores (and varying TSI), which prompted the further outlier analysis.

Figure 3: **APD correlates more strongly with TSI than the fraction of ambiguous nucleotides (FAN).** Upper panels: TSI against APD scores over third codon positions in *gag* and *pol*, and fraction of ambiguous nucleotide scores in partial *pol* reads. Linear regression models are shown, which were then used to predict TSI from the diversity score. The sample size is restricted to the 207 samples that have corresponding ambiguous nucleotide scores, to allow for a more direct comparison of the two methods. Lower panels: The same data and models from the upper panels with log-transformed TSI axis.

Figure 4: **Outliers with exceptionally high APD scores can be identified and removed.** Top: APD *gag* and APD *pol* against TSI, with an asymptotic curve fitted to define outliers. The pale dashed lines show the asymptote itself (at 0.0408 for APD

gag and 0.0336 for APD *pol*), above which samples may be potential outliers. Samples that are above this line for both APD *gag* and APD *pol* are highlighted in blue. Note that the axes have been switched, as the asymptotic curve was fitted by taking APD as the dependent variable against TSI. This was done because the presence of a singularity makes fitting an asymptotic curve to the data with TSI as the dependent variable very challenging.

Bottom: TSI against APD *gag* and APD *pol*, with five outliers removed. Lines show the linear regression models, recalculated without the outliers.

Figure 1

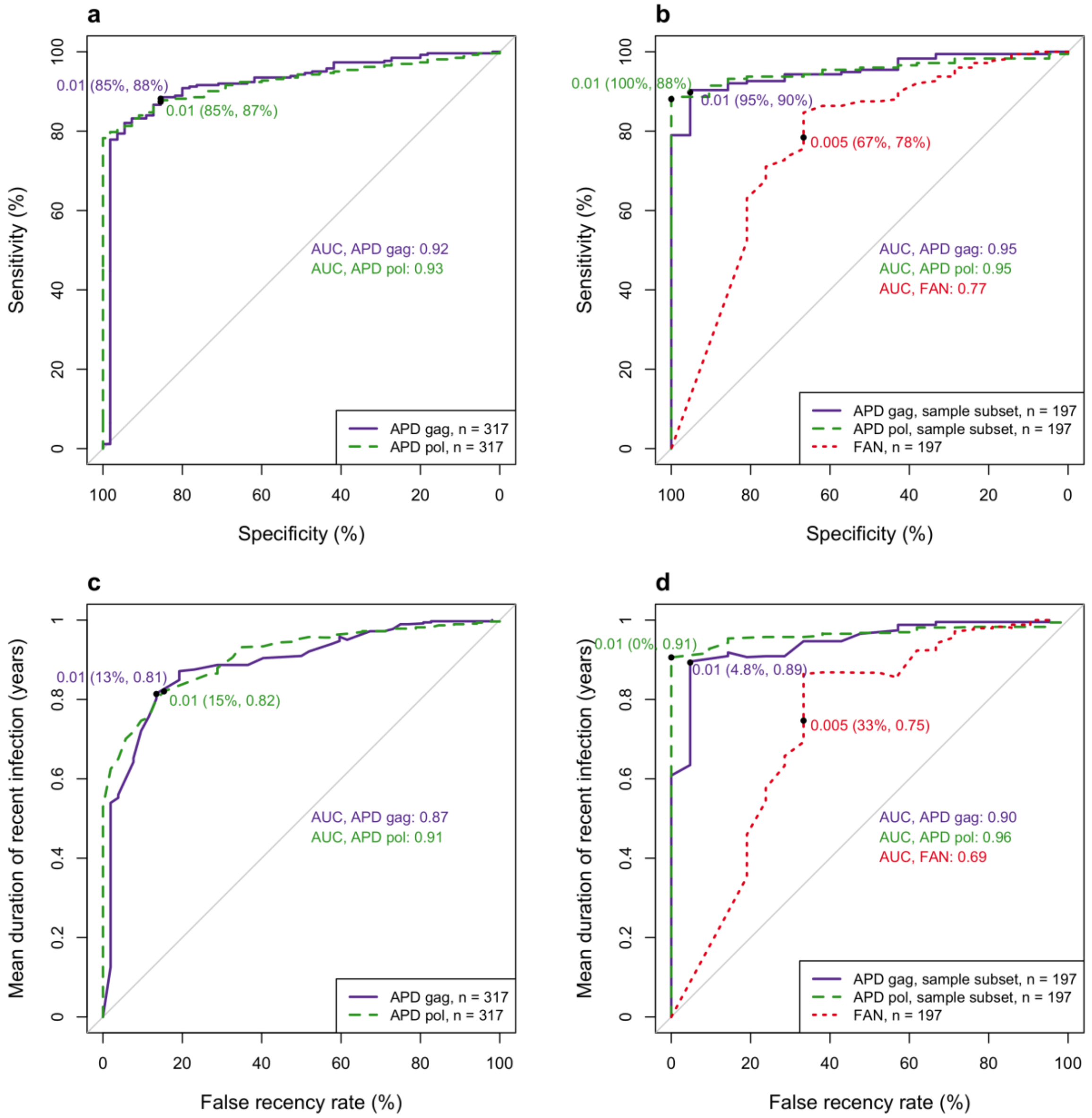
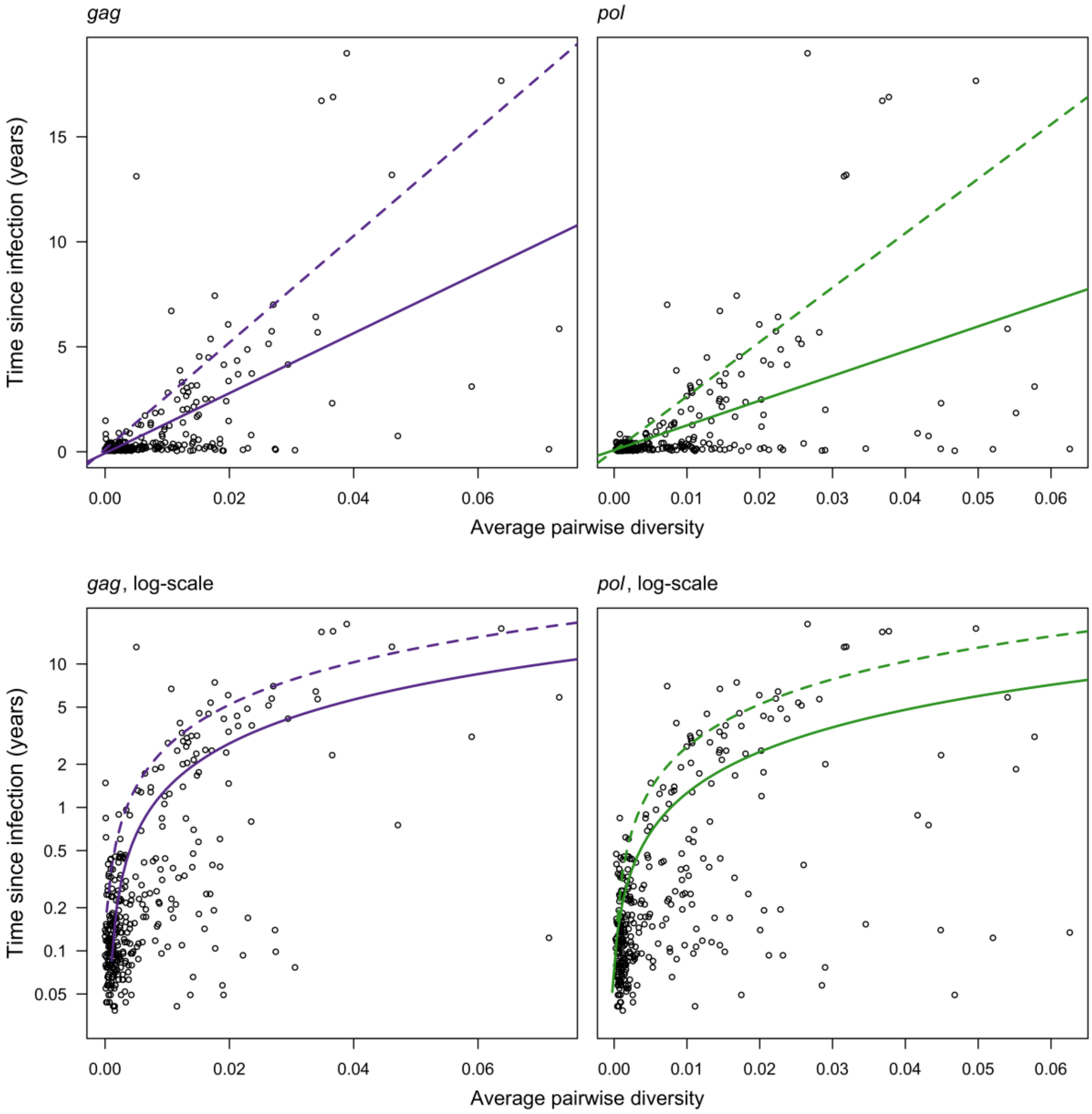
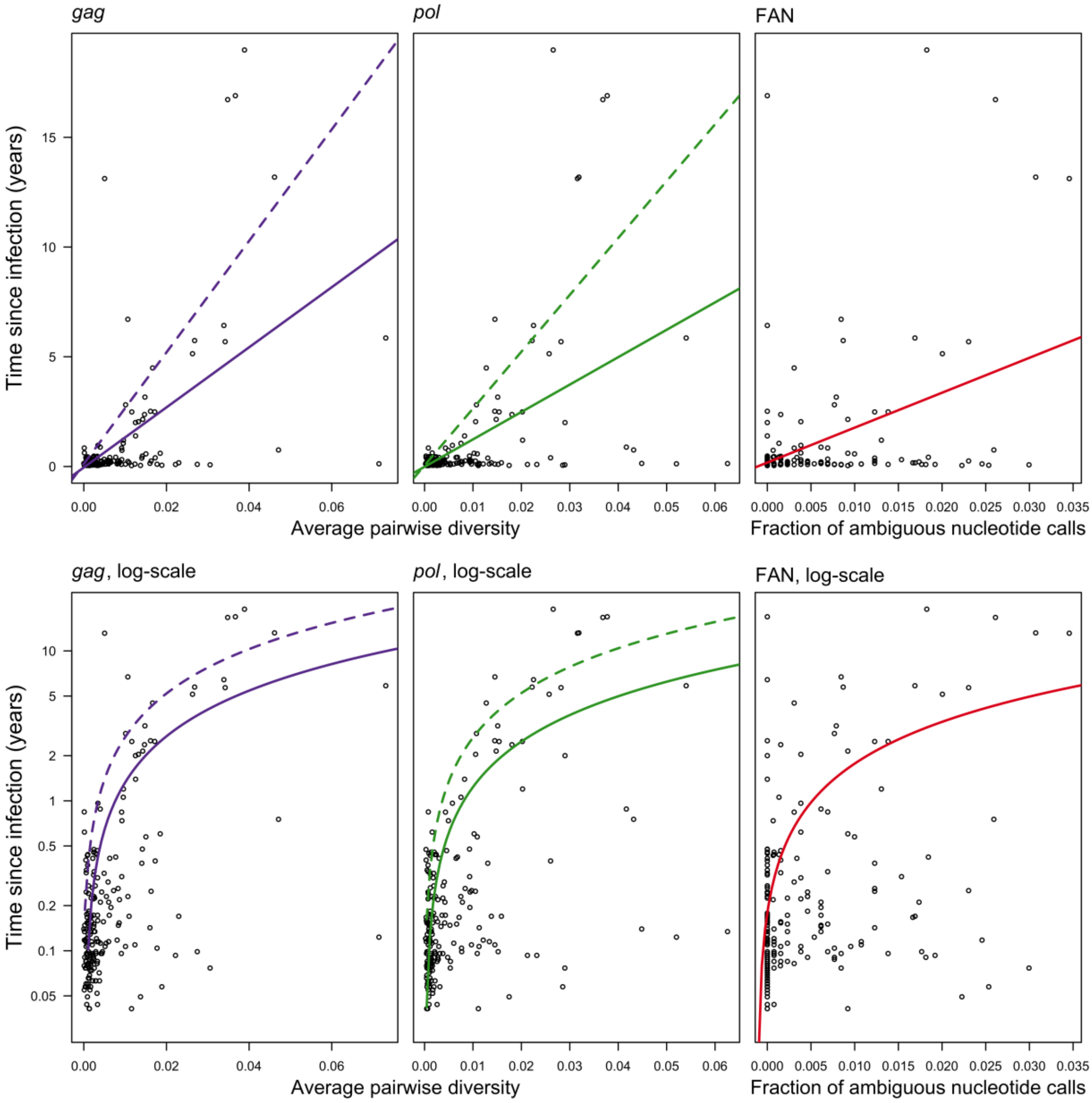

a
Sensitivity (%)
100
80
60
40
20
0
0.01 (85%, 88%)
0.01 (85%, 87%)
AUC, APD gag: 0.92
AUC, APD pol: 0.93
APD gag, n = 317
APD pol, n = 317
100
80
60
40
20
0
Specificity (%)
b
Sensitivity (%)
100
80
60
40
20
0
0.01 (100%, 88%)
0.01 (95%, 90%)
0.005 (67%, 78%)
AUC, APD gag: 0.95
AUC, APD pol: 0.95
AUC, FAN: 0.77
APD gag, sample subset, n = 197
APD pol, sample subset, n = 197
FAN, n = 197
100
80
60
40
20
0
Specificity (%)
c
Mean duration of recent infection (years)
1
0.8
0.6
0.4
0.2
0
0.01 (13%, 0.81)
0.01 (15%, 0.82)
AUC, APD gag: 0.87
AUC, APD pol: 0.91
APD gag, n = 317
APD pol, n = 317
0
20
40
60
80
100
False recency rate (%)
d
Mean duration of recent infection (years)
1
0.8
0.6
0.4
0.2
0
0.01 (0%, 0.91)
0.01 (4.8%, 0.89)
0.005 (33%, 0.75)
AUC, APD gag: 0.90
AUC, APD pol: 0.96
AUC, FAN: 0.69
APD gag, sample subset, n = 197
APD pol, sample subset, n = 197
FAN, n = 197
0
20
40
60
80
100
False recency rate (%)

Figure 2



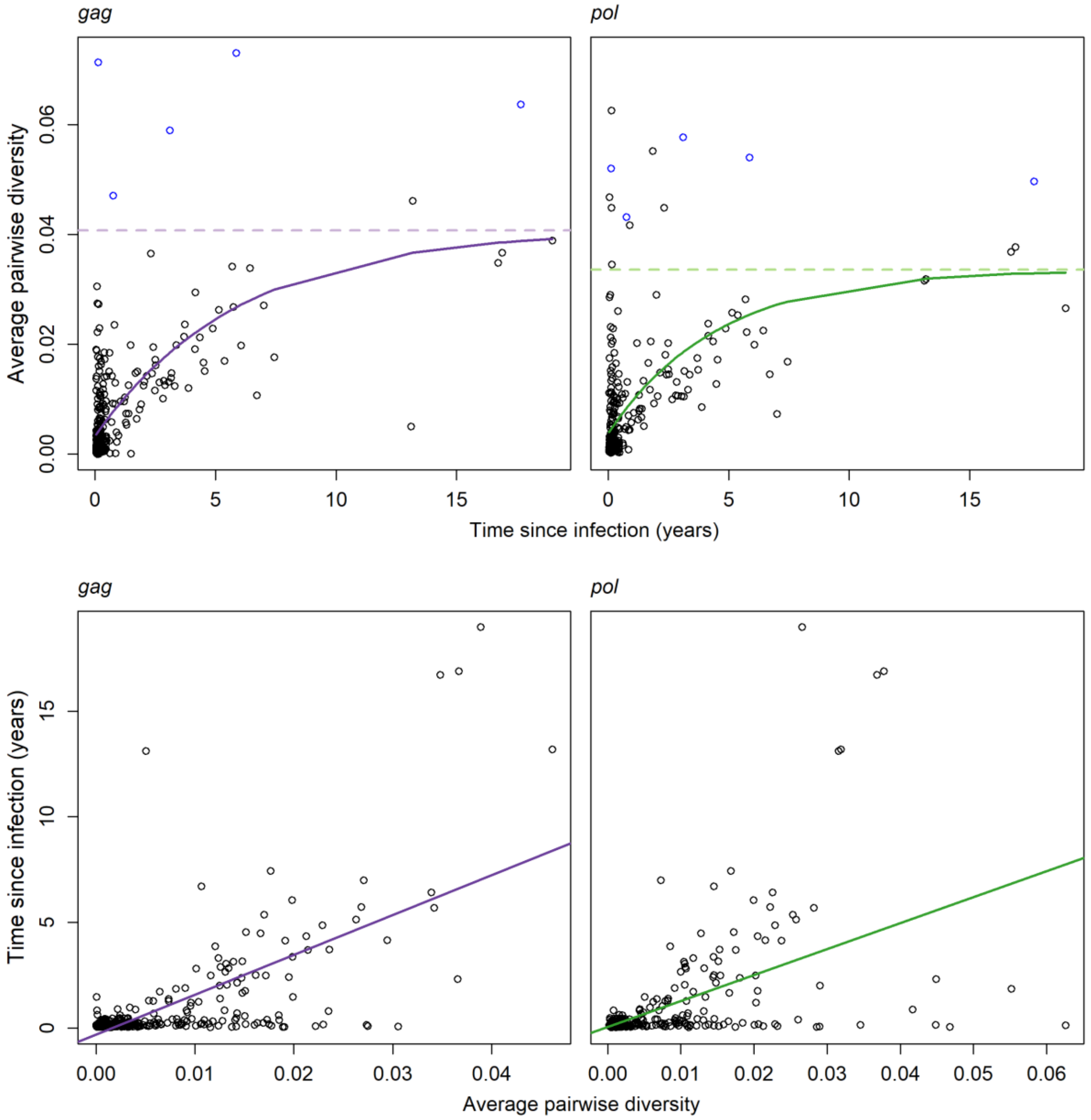
gag
pol
Time since infection (years)
15
10
5
0
0.00
0.02
0.04
0.06
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Average pairwise diversity
gag, log-scale
pol, log-scale
Time since infection (years)
10
5
2
1
0.5
0.2
0.1
0.05
0.00
0.02
0.04
0.06
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Average pairwise diversity

Figure 3



gag
pol
FAN
Time since infection (years)
0
5
10
15
0.00
0.02
0.04
0.06
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Average pairwise diversity
Fraction of ambiguous nucleotide calls
gag, log-scale
pol, log-scale
FAN, log-scale
0.05
0.1
0.2
0.5
1
2
5
10

Figure 4


gag
pol
Average pairwise diversity
0.00
0.02
0.04
0.06
0
5
10
15
Time since infection (years)
gag
pol
Time since infection (years)
0
5
10
15
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Average pairwise diversity