Estimating time of HIV-1 infection from next-generation sequence diversity

Vadim Puller[1,2¤*], Richard Neher[1,2¤], Jan Albert[3,4]

1 Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany
2 Biozentrum, University of Basel, Basel, Switzerland
3 Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden
4 Department of Clinical Microbiology, Karolinska University Hospital, Stockholm, Sweden

¤Current Address: Biozentrum, University of Basel, Basel, Switzerland
* vadim.puller@unibas.ch

Abstract

Estimating the time since infection (TI) in newly diagnosed HIV-1 patients is challenging, but important to understand the epidemiology of the infection. Existing biomarkers for the recent infection are relatively imprecise. Here we explore the utility of virus diversity estimated by next-generation sequencing (NGS) as novel biomarker by using a recent genome-wide longitudinal dataset obtained from 11 untreated HIV-1-infected patients with known dates of infection.

Virus diversity increased linearly with time, particularly at 3rd codon positions, with little inter-patient variation. The precision of the TI estimate improved with increasing sequencing depth, showing the superiority of NGS over counting polymorphic sites in Sanger sequences, which is one of the alternative biomarkers. The full advantage of the high sequencing resolution of NGS was utilized with continuous diversity measures, average Hamming distance or site entropy, rather than the fraction of polymorphic sites. The precision depended on the genomic region and codon position and was highest when 3rd codon positions in the entire *pol* gene was used. For these data TI estimates had a mean absolute error of around 1 year. The error increased only slightly from around 0.6 years at a TI of 6 months to around 1.1 year at 6 years. In addition, NGS diversity compared favorably with other biomarkers for binary classification of patients as being recently or long-term infected.

Our results show that virus diversity determined by NGS can be used to estimate time since HIV-1 infection with a precision that is better than most alternative biomarkers. Importantly, TI can be estimated many years after infection. We provide regression coefficients that can be used for TI estimation.

Author summary

HIV-1 establishes a chronic infection, which may last for many years before the infected person is diagnosed. The resulting uncertainty in the date of infection leads to difficulties in estimating the number of infected but undiagnosed persons as well as the number of new infections, which is necessary for developing appropriate public health policies and interventions. Such estimates would be much easier if the time since HIV-1

infection for newly diagnosed cases could be accurately estimated. Three types of biomarkers have been shown to contain information about the time since HIV-1 infection, but unfortunately they only distinguish between recent and long-term infections (concentration of HIV-1-specific antibodies) or are too imprecise (immune status as measured by levels of CD4+ T-lymphocytes and viral sequence diversity). In this paper we show that recent advances in sequencing technologies, i.e. the development of next generation sequencing, enable significantly more precise determination of the time since HIV-1 infection, even many years after the infection event. This is a significant advance which could translate into more effective HIV-1 prevention.

Introduction

At diagnosis most HIV-1 infected patients have an established HIV-1 infection of unknown duration. This uncertainty complicates inference about the epidemiology of HIV-1. Consequently, there is limited information about the true incidence of HIV-1, the number of hidden, undiagnosed infected persons, the magnitude of the problem referred to as "late presentation" and other important aspects of HIV-1 spread.

Several biomarkers that classify patients as recently or long-term infected have been used to estimate HIV-1 incidence in populations [1–7]. These biomarkers can be divided into three main categories: (i) serological incidence tests, (ii) CD4+ T-lymphocyte (CD4)-based estimates and (iii) sequence-based estimates. Importantly, these biomarkers usually do not determine the time since infection (TI), which limits their utility.

Serological incidence assays are based on knowledge about the development and maturation of HIV-1 antibody responses (reviewed in [1, 4–6, 8]). Among the serological assays, the BED assay is the most widely used [4, 8, 9]. CD4 counts are determined as part of routine clinical care a CD4 count below 350 cells/μL (or an AIDS-defining illness) at diagnosis is defined as late presentation [10, 11]. However, CD4 count is an imprecise measure of TI because its rate of decline is quite variable. [12–15]

Sequence-based methods focus on the increase in intra-patient HIV-1 sequence diversity following infection [16]. Kouyos et al [17] showed that time since infection correlated with the fraction of polymorphic nucleotides in partial HIV-1 *pol* gene sequences determined by Sanger sequencing. Others have later reported similar findings [18, 19]. This idea was expanded to other measures of sequence diversity, such as mean Hamming distance [6, 7, 20] and high-resolution melting (HRM) [21]. These studies (except HRM) used sequences generated by traditional Sanger population sequencing often performed as part of routine HIV-1 resistance testing.

Here, we have investigated the utility of estimating time since HIV-1 infection using genetic diversity in whole genome deep sequencing data generated by next-generation sequencing (NGS) on the Illumina platform [22]. We show that sequence diversity is a useful biomarker that grows approximately linearly with time during the first 8 years of infection. We found that the *pol* gene was best suited to calculate TI because diversity, mostly at third positions, accumulated more steadily in *pol* than in other genomic regions. Inclusion of intra-patient single nucleotide variants (iSNVs, also referred to as "polymorphisms") down to the detection limit of NGS (i.e. 1% or less) improved the accuracy of TI estimations as did exclusion of 1st and 2nd codon position sites. NGS provided more accurate estimates of TI than Sanger sequencing, which at best detects iSNVs down to 25% [23–25].

Table 1. Summary of patient characteristics.

Patient	Gender	Transmission	Age	Subtype[a]	CD4 count[b]	viral load[c]	HIV RNA from plasma		
							# samples	first/last since EDI[d]	
p1	F	HET	37	01_AE	537	¡500	12	0.3	8.2
p2	M	MSM	32	B	1185	31000	6	0.2	5.5
p3	M	MSM	52	B	694	17000	10	0.4	8.4
p4	M	MSM	29	B	1064	6800	8	0.3	8.4
p5	M	MSM	38	B	521	¡50	7	0.4	5.9
p6	M	HET	31	C	480	6800	7	0.2	7.0
p7	M	MSM	25	B	496	500	11	5[e]	15.9
p8	M	MSM	35	B	460	2900	7	0.2	6.0
p9	M	MSM	32	B	720	12000	8	0.3	8.1
p10	M	MSM	34	B	420	13000	9	0.09	5.2
p11	M	MSM	53	B	850	1100	7	0.6	5.6

Table notes [a]at diagnosis; [b]cells/µl; [c]HIV RNA copies/ml; [d]EDI: estimated date of infection; all times are given in years; [e]sequencing failed in earlier samples due to low plasma HIV-1 RNA levels.

Materials and methods

Patients

The study included HIV-1 whole-genome sequence data from 11 patients who were diagnosed in Sweden between 1990 and 2003 (Table 1). Details about the patients and the samples have been published [22] and are also available online at `hiv.biozentrum.unibas.ch`. The patients were selected based on the following criteria: 1) A relatively well-defined time of infection based on a negative HIV antibody test less than two years before the first positive test or a laboratory documented primary HIV infection; 2) No antiretroviral therapy (ART) during a minimum of approximately five years following diagnosis; and 3) Availability of biobank plasma samples covering this time period. As previously described 6 - 12 plasma samples per patient were retrieved from biobanks and used for full-genome HIV-1 RNA deep sequencing [22]. Time of infection was calculated using clinical and laboratory findings including Fiebig staging and BED testing [22].

The study was conducted according to the Declaration of Helsinki. Ethical approval was granted by the Regional Ethical Review board in Stockholm, Sweden (Dnr 2012/505 and 2014/646). Patients participating in the study gave written and oral informed consent to participate.

HIV-1 RNA sequences:

Whole-genome deep-sequencing of virus RNA populations in plasma samples obtained before start of therapy was performed as previously described [22, 26]. In short, total RNA in plasma was extracted using RNeasy® Lipid Tissue Mini Kit (Qiagen Cat No. 74804) and amplified using a one-step RT-PCR with outer primers for six overlapping regions and Superscript ® III One-Step RT-PCR with Platinum ® Taq High Fidelity High Enzyme Mix (Invitrogen, Carlsbad, California, US). An optimized Illumina Nextera XT library preparation protocol was used together with a kit from the same supplier to build DNA libraries, which were sequenced on the Illumina MiSeq instrument with 2 x 250bp or 2x 300bp sequencing kits (MS-102-2003/MS-10-3003).

Sequencing reads are available in the European Nucleotide Archive under accession number PRJEB9618 and processed data are available at `hiv.biozentrum.unibas.ch`.

All analyses were done in Python using the libraries numpy, biopython, and matplotlib [27–29].

Patient-specific consensus sequences were constructed using an iterative mapping procedure [22, 26]. All reads were then remapped against this reference to calculate iSNV frequencies (i.e. pile-ups or tables how often each nucleotide was observed at every position of the genome). These iSNV frequency tables were then used to calculate average Hamming distances, average alignment entropies, or the number of sites with variation above a cutoff x_c.

Statistical procedures

We have used three different diversity measures: number of polymorphic sites, Hamming distance, and site entropy. All of these measures can be straight-forwardly calculated from the frequencies of different nucleotides $x_{i\alpha}$ at site $i = 1 \ldots L$ and $\alpha \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ along the genome. Prior to calculation the nucleotide frequencies for each site were normalized to sum to unity (i.e. ignoring gaps or positions not called by the sequencer.)

For all methods, we introduce a cut-off x_c. Sites at which the sum of all minor variants is smaller than x_c contribute zero to the diversity measure. This cut-off serves to filter sequencing errors or rare variation that cannot be reproducibly measured across samples. When using the *fraction of polymorphic sites* as diversity measure, x_c serves as the values above which sites are considered "polymorphic". Specifically, the *fraction of polymorphic sites* is defined as

$$D_A = \frac{1}{L} \sum_{i=1}^{L} \Theta(x_c < 1 - x_i^m) \tag{1}$$

where $\Theta(x)$ is 1 when $x > 0$ and 0 otherwise and x_i^m is the frequency of the dominant nucleotide at position i. D_A is thus the fraction of sites at which the dominant nucleotide is less frequent than $1 - x_c$.

The *average Hamming distance* is the average pairwise distance between sequences and can be calculated from the $x_{i\alpha}$ as

$$D_H = \frac{1}{L} \sum_{i=1}^{L} \Theta(x_c < 1 - x_i^m) \left[\sum_{\alpha} x_{i\alpha}(1 - x_{i\alpha}) \right] \tag{2}$$

The average Hamming distance is hence similar to the number of polymorphic sites with the important difference that the contribution of each site is weighted by a frequency dependent factor.

The *average site entropy* is defined by

$$D_E = -\frac{1}{L} \sum_{i=1}^{L} \Theta(x_c < 1 - x_i^m) \left[\sum_{\alpha} x_{i\alpha} \log(x_{i\alpha}) \right] \tag{3}$$

and different from the average Hamming distance by the weighing function used. The entropy puts more weight on sites with rare variation.

The Hamming distance is used as the default diversity measure, unless otherwise specified.

TI was calculated using

$$\hat{\mathrm{TI}} = s \times D + t_0, \tag{4}$$

where D is the chosen diversity measure, s is the conversion factor between diversity and time, and t_0 is the intercept value intended to accommodate possible non-linearity of diversity at small times.

To estimate values of s, t_0 we minimized the average prediction error for the available data points in respect to these two parameters (see S1 Appendix for more details)

The error in estimating s, t_0 was calculated by sampling randomly from the available data points (bootstrapping) or by sampling randomly the patients.

Cross-validation

To test the accuracy of the TI inference we used ten of the eleven patients as training data (to determine the slope and the intercept) treating the eleventh patient as test data. This procedure was repeated for every patient. Leaving out one patient at a time, rather than one sample at a time, gives more accurate confidence intervals as different samples from the same patients are correlated. We included in statistical analysis only samples where more than 50% of the sites in the averaging window were successfully sequenced.

Results

Patients characteristics:

We analyzed recently published sequence data from 11 HIV-1 infected patients who were diagnosed in Sweden between 1990 and 2003 [22]. Patient characteristics are summarized in Table 1. These patients had a relatively well-defined date of infection, which was estimated using a combination of time between last HIV-1-negative and first HIV-1-positive sample, BED tests, laboratory-documented primary HIV-1 infection, and Fiebig staging as described in [22]. Here, we take this date estimate as the true time of infection and investigate how accurately the time since infection (TI) can be estimated from sequence diversity in one sample. We will refer to this estimate as the *estimated time since infection* (ETI).

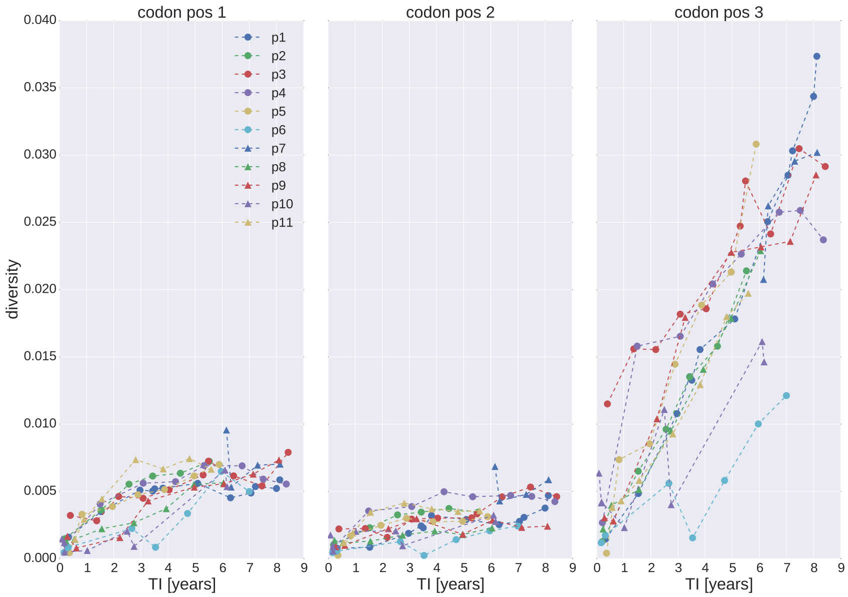
Sequence diversity as a biomarker

All three diversity measures described in Materials and Methods grew linearly with time in the eleven patients. Fig. 1 shows Hamming distance in the *pol* separately for each codon position. Most diversity in *pol* is synonymous and accumulates at 3rd codon positions, while diversity at 1st and 2nd codon positions remained low throughout. This dichotomy was less pronounced in other genes [22, 30] (see also S1 Fig and S2 Fig). In *env* in particular frequent selective sweeps limit diversity later in infection [16, 22].

We quantified the fraction of variation in diversity measures that can be explained by a linear regression by calculating the Pearson correlation coefficient between sample date and diversity, see S3 Fig and S4 Fig. In all patients for whom early samples were available, a linear regression explained between 70% and 90% of variation if rare iSNVs below 20% population frequency were included, that is the cut-off x_c was below 0.2. We observed a sharp decrease in linearity when only iSNVs between 20% and 80% were included, that is the x_c cut-off is larger than 0.2. This decrease is due to increased noise as fewer and fewer sites contribute to the diversity measures.

Furthermore, as seen from S4 Fig, the diversity of 3rd codon positions (which are mostly synonymous) exhibited even stronger linearity, whereas the trajectories at the 1st and 2nd codon positions saturated quickly after the infection, Fig. 1. Thus, in the following we limit the analysis mainly to sites in 3rd codon positions (whenever we are dealing with a whole gene, i.e. when the reading frame is known.) However, the results reported below show that inclusion of 1st and 2nd codon positions has a limited effect on the accuracy of the TI estimates.

Fig 1. Diversity in pol as a function of the time since infection (TI) and 1st, 2nd and 3rd codon positions. (Genetic region: *pol*, diversity measure: Hamming distance, $x_c = 0$.)

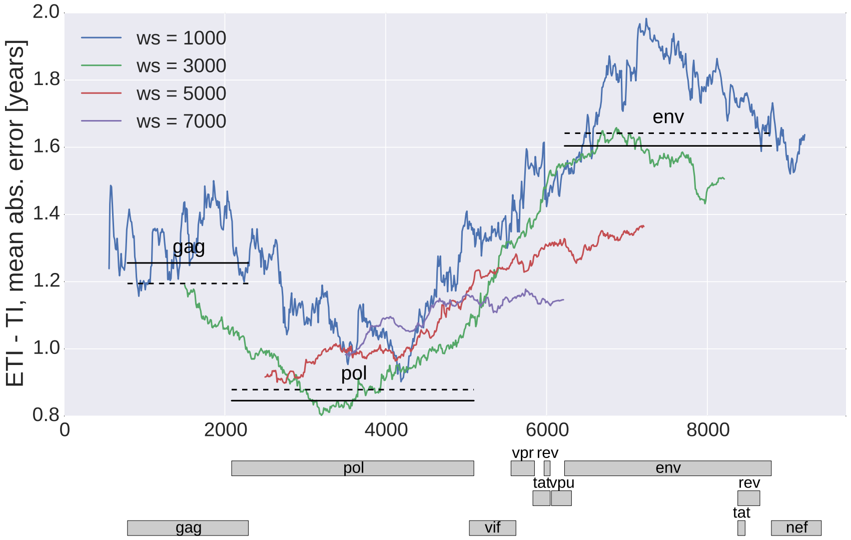


Diversity in *pol* yielded the most accurate TI estimates

Accurate estimation of TI requires averaging diversity across many sites. To investigate which regions of the genome yields the most accurate estimates and how many sites should be averaged, we calculated the average prediction error for averaging windows of different length and in different regions of the genome. In Fig. 2 the mean absolute error (MAE) for the estimated TIs is shown as a function of the genome window position for different window sizes. We found that the most precise TI estimates were obtained for windows with a length of 2000-3000bp. The precision was highest if the window covered the *pol* gene and significantly lower in *env*. Smaller windows contain fewer sites and therefore gave less precise estimates. Larger windows also gave less precise TI estimates because they necessarily include regions in which diversity saturates (such as *env*, S2 Fig) as well as regions that often were sequenced less deeply in our dataset (again *env*, S5 Fig and S6 Fig).

The precision of the TI estimates obtained for windows corresponding to particular genes are shown in Fig. 2 by lines indicating the position of the gene and the corresponding average absolute error of the TI estimate. The dashed lines correspond to estimates using only synonymous sites of the corresponding gene. Although exclusion of non-synonymous sites (i.e. two thirds of all sites) increased the statistical error, this was compensated by a better prediction due to higher linearity of the corresponding diversity trajectories. Therefore we further consider only 3rd codon positions in *pol*, which gave the most precise TI estimates.

Fig 2. Mean absolute prediction error of TI as a function of position in genome and different sizes of the genome window (ws). Straight solid lines correspond to the error when estimation is based on diversity in the genes *gag*, *pol* or *env*. The dashed lines are analogous estimates using diversity only at 3rd codon position. Diversity measure: Hamming distance, $x_c = 0$.

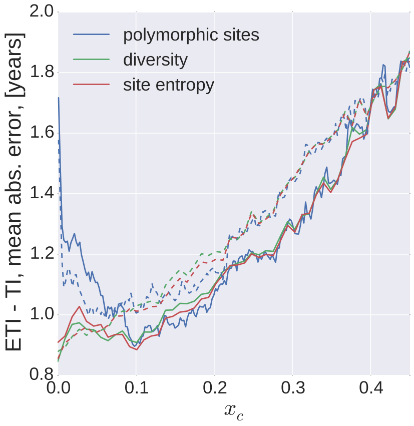


Accuracy increased with higher sequencing resolution

Genetic diversity in NGS data can be quantified by different measures and we investigated the performance of three related measures – Hamming distance, site entropy, and fraction of polymorphic sites. As discussed above, these measures put different emphasis on iSNVs at different frequency. Even though the Hamming distance and site entropy formally do not require a cutoff on minority iSNVs, in practise a cutoff is necessary to remove low-level experimental errors. We therefore introduced an iSNV cutoff in calculating the diversity based on Hamming distance and site entropy, x_c, taking into account only the data with iSNVs above x_c. For diversity measures based on polymorphic sites, the cutoff value corresponds to the iSNV level from which the site is considered polymorphic. This emulates ambiguous base calls by Sanger sequencers , which (at best) can identify minority iSNVs above a threshold of around 25% [23–25].

Given the high ability of NGS to detect low level iSNVs one can consider the dependence of the TI estimation on the cutoff value, as shown for all three diversity measures in Fig. 3. All three diversity measures performed equally well and increasingly better at cutoff values x_c down to around 10%. This indicates that diversity in NGS data allows more accurate estimates of TI than ambiguous base calls in Sanger sequences and that this primarily is due to better sensitivity and accuracy for low level iSNVs. For cutoffs below 10% the error of TI estimates based on counting polymorphic sites was greater than those based on the other two measures. Indeed, at $x_c = 0$ this measure includes all sites and becomes completely insensitive to differences in iSNV levels. The other two measures do not suffer from this problem, because they put different weights on sites with different iSNV levels. Therefore the prediction errors

Fig 3. Mean absolute error as a function of the low-frequency cutoff (x_c). Different diversity measures perform very similarly when the cut-off x_c is greater than approximately 10%. Hamming distance and entropy outperform fraction of polymorphic sites for low x_c. This graph is based on diversity in *pol*. Solid lines correspond to using all sites, dashed - only 3rd codon positions.



continue to decrease all the way down to $x_c = 0$ for estimates based 3rd codon positions and Hamming distance or site entropy.

The noticeably non-monotonous behavior of the predictions based on Hamming distance and site entropy when all sites are taken into account is due to the saturation of 1st and 2nd codon positions (i.e. non-synonymous) diversity. The 1st and 2nd codon positions tend to be conserved since they result in mostly non-synonymous mutations. Depending on the fitness cost associated with the mutation, diversity at these sites saturates [30]. As the threshold x_c is lowered, sites with higher and higher fitness costs contribute to the diversity measures and the effect of the saturation behavior becomes more pronounced. Note that this effect is not consistently reproduced in other genes (see supplementary S7 Fig.) The fact that at the cutoffs above 10% the all sites measure performs somewhat better is equally not consistently reproduced in other genes.

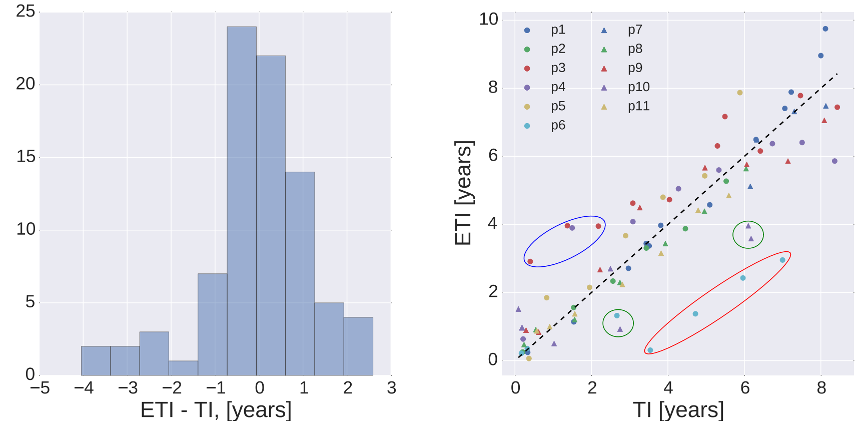
In the following analyses we opted for using the Hamming distance measure, taking into account only the diversity at the 3rd codon positions, with zero iSNV cutoff. Hamming distance was chosen because the results were virtually indistinguishable from those produced using site entropy, but easier to calculate and interpret.

Distribution of prediction errors

The results above indicated that more than 50% of the estimated TIs fell within a window of about one year centered at the actual TI. Fig. 4 (Left) shows a more direct analysis of the distribution of TI prediction errors. The distribution is tightly peaked around zero, but has a left tail corresponding to samples estimated to have been obtained shorter after infection than they actually were drawn, i.e. diversity being lower than expected. Most of these samples were from p6, who throughout infection had lower diversity than other patients. In addition biological reasons for low diversity, amplification problems and low RNA template numbers (i.e. low virus levels) can explain underestimation of diversity.

Some samples were estimated to have been drawn later after infection than the true duration of infection. In particular early samples from p10 and p3 had markedly higher diversity than expected. For both patients, we have evidence that their infections were established by more than one virion resulting in carry-over of diversity from the donor.

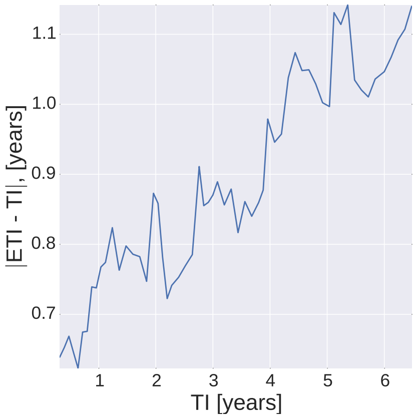
Fig 4. (Left) Distribution of the estimation error. (Right) Estimated time of infection (ETI) versus actual time of infection (TI). (Genetic region: 3rd codon positions in *pol*, diversity measure: Hamming distance. The encircled outliers are discussed in the text.)



This excess diversity gradually decreased in p10 and, somewhat slower, in p3.

Next, we investigated how the prediction error depended on the time since infection. Fig. 5 shows the average absolute error of the estimated TI versus the true TI, averaged over $n = 25$ consecutive data points. This average error (see for details S2 Appendix) was surprisingly stable over TIs, and only increased slightly from around 0.6 years to around 1.1 year as the age of infection increased from 6 months to 6 years. This increase can be attributed to bigger statistical fluctuations of diversity later after infection due to factors such as variations in the rate of diversification or differences in number and strength of selective sweeps that reduce diversity.

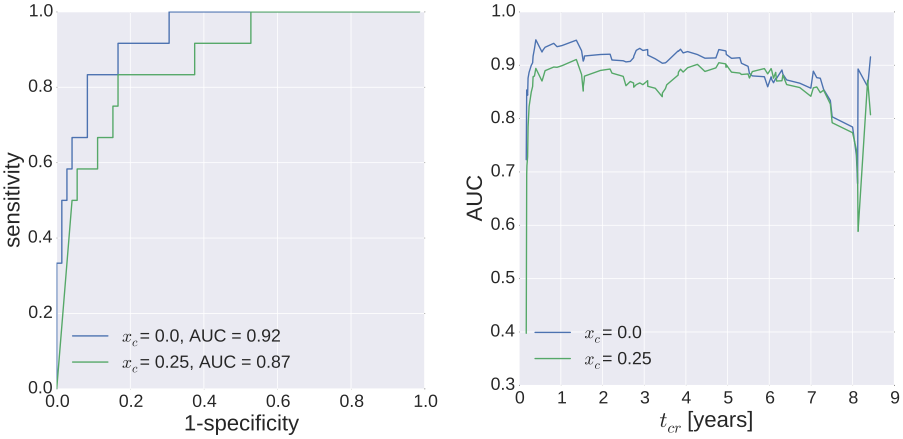
Fig 5. Estimation error dependence on time of infection (TI): TI and $|\mathrm{ETI} - \mathrm{TI}|$ averaged over $n = 25$ adjacent points. (Genetic region: 3rd codon positions in *pol*, diversity measure: Hamming distance.)



Distinguishing recent and long-term infections

Currently, most methods for cross-sectional incidence estimation rely on binary classifications into recent and long-term infections. Recent infections are defined as

Fig 6. (Left) Receiver operating characteristic (ROC) curve. The infection is considered as recent if TI is less than 6 months. (Right) Dependence of the area under the ROC curve (AUC) on the time defining infection as recent. (Genetic region: 3rd codon positions in *pol*, diversity measure: Hamming distance.)



those with the infection time less than a specified cutoff ranging from a few months to one year. Such binary classification can be done based on NGS data; if the diversity is less than a specified cutoff value D_{cr} infection is classified as recent, and otherwise as long-term. One then defines sensitivity as the probability of correctly classifying a recent infection as recent (true positive rate), and specificity as the probability of correctly identifying long-term infection (true negative rate). The results can be characterized using the receiver operating characteristic (ROC) curve [7], shown in Fig. 6 (Left). Different points on the curve correspond to different values of D_{cr}, and we assume six months as the duration of a recent infection. Overall our results confirm that sequence diversity is a good biomarker for distinguishing recent and long-term infections, as one can simultaneously achieve high sensitivity and high specificity (the area under the curve (AUC) is close to 1). Lowering the iSNV frequency cutoff increases the AUC from 0.87 for $x_c = 0.25$ to 0.92 for $x_c = 0.0$, underscoring the advantage of using NGS that can identify and quantify minority iSNVs.

The area under the curve shows itself nearly independent on the time cutoff (t_{cr}) for recent vs. long-term infections, as seen in Fig. 6 (Right). The steep rise of the curve is mainly due to the small number of samples with times smaller than t_{cr}. The same is true for the decline of the curve at large times, where it is compounded by increase in the diversity fluctuations. Otherwise the classifier works very well for t_{cr} with the cutoff ranging from four months to five years. (Of course, using five years as a cutoff for recent infection is somewhat pointless.)

Recommended regression coefficients

In Table 2 we list the values of slope and intercept that can be used to estimate the infection date from the known diversity calculated as Hamming distances for 3rd codon positions in *pol* gene. The supplementary materials contain similar tables for the two other diversity measures (S1 Table and S2 Table), as well as for the case when all codon positions are taken into account (S3 Table, S4 Table and S5 Table). As the iSNV resolution may vary between different sequencing methods and facilities, we list the values of the parameters for different cutoffs, implying that all the frequencies below the cutoff value are set to zero and the corresponding sites do not contribute to the

Table 2. Recommended slope and intercept values depending on the cutoff.

	slope and intercept			slope only	
cutoff (x_c)	slope (s)^a	intercept (t_0)^b	MAE^b	slope (s)^a	MAE^b
0.00	250.28	-0.08	0.88	247.59	0.86
0.05	297.17	0.27	0.96	321.96	0.99
0.10	350.52	0.50	1.03	386.22	1.05
0.15	407.65	0.75	1.14	490.79	1.18
0.20	484.39	0.79	1.23	556.58	1.29
0.25	551.75	1.02	1.35	701.51	1.44
0.30	692.19	1.20	1.46	850.78	1.61
0.35	780.52	1.56	1.60	1185.82	1.73
0.40	1218.07	1.54	1.69	1685.91	1.86
0.45	1394.26	2.53	1.87	2312.54	2.61

Table notes Genetic region: 3rd codon positions in *pol*, diversity measure: Hamming distance. ^{a}in years/diversity; ^{b}in years.

diversity measure. Note that the slope (and intercept) increases with increasing iSNV cut-offs because fewer and fewer sites contribute to diversity. In addition to the two parameter model, we also investigated the performance of a model with the slope as the single parameter, i.e. no intercept ($t_0 = 0$). This model has a slightly higher absolute prediction error. HOwever, for low values of the cut-off $x_c \approx 5\%$ these models agree and for cut-offs below 20% the two models perform equally well.

Discussion

Most newly diagnosed HIV-1 patients have an infection of unknown duration. This is problematic because accurate estimation of the time since infection (TI) is essential for understanding important aspects of HIV-1 epidemiology such incidence, proportion of undiagnosed patients and late presentation. Most previous methods are suboptimal because they only categorize patients as being recently or long-term infected and/or are imprecise. Here, we show that genetic diversity calculated from NGS data enables fairly accurate estimation of TI, even many years after infection. We also show that NGS is superior to Sanger sequencing because inclusion of minority iSNVs below the Sanger detection limit (around 25%) substantially improves the accuracy of the TI estimates.

Our results show that the most precise estimation of TI were obtained Hamming distances or site entropy based on 3rd codon positions in the *pol* gene. For these data viral diversity increases approximately linearly during at least 8 years after infection, which allows estimation of TI during this time period. The accuracy of the TI estimate was approximately +/- 1 year in long-term infections, and slightly lower during the first year of infection. For convenience we provide a supplement table that translates viral diversity into TI.

We also investigated the performance of NGS diversity as binary classifier for identifying recent vs. long-term infections using ROC statistics. When low-level iSNVs were included the AUC was 0.92, which is higher than published AUCs for other biomarkers, BED (0.78), Limiting Antigen assay (LAg)(0.81) and pairwise genetic distances (PwD) of single genome sequences (SGS) (0.83) [7]. NGS diversity even compared favorably with combinations of PwD and BED or LAg [7]. Furthermore, the ROC characteristics of our method was similar to those reported for Hamming distances calculated from SGS data from the *env* [6]. It should be stressed SGS is a labor-intensive and costly method whereas NGS is becoming increasingly available and affordable.

Furthermore, our NGS-based method has the advantage of being able to estimate TI, not just distinguishing recent vs. long-term infections. We found that *env* was less suitable than *pol* for estimating TI, especially if longer time had elapsed since infection. This is because frequent selective sweeps in the *env* continuously remove diversity and this effect becomes increasingly evident with increasing time since infection [16, 22].

We investigated how the TI estimates were influenced by the the length of the sequence, region of genome, nucleotide position, iSNV cutoff and type of distance measure. We found that the precision of the TI estimates was highest in the *pol* gene and lowest in the *env*. For this reason the most accurate results were obtained using 2000-3000 base pair long sequences covering *pol*, while omitting *env*. Furthermore, we found that most of temporal signal came from 3rd codon positions (which usually are synonymous sites) and that omission 1st and 2nd codon positions improved TI estimates for iSNV cutoffs (x_c) below 10%, i.e. when the full potential of NGS was utilized. Finally, Hamming distance and site entropy outperformed the distance measure based on fraction of polymorphic sites at low iSNV cutoffs. Hamming distance was preferred because it is more straightforward to calculate than site entropy. Based on our results we make the following provisional recommendations for TI estimation based on NGS data; Hamming distance on 3rd codon positions in *pol* sequences with a 0% cutoff for iSNVs.

Our study and method have several limitations. Ideally we should have studied more than 11 patients, but patients with known time of infection, long followup without therapy and suitable biobank samples are rare. Today it would be unethical to delay start of ART. Thus, our study subjects were diagnosed between 1990 and 2003 and retrospectively identified and investigated using stored biobank samples [22]. Furthermore, whole-genome deep sequencing of longitudinal plasma samples is a major undertaking and our study was based on close to 100 samples and more than 100 million sequence reads. Another limitation is that the true (i.e. ”known”) TI was estimated from laboratory and clinical data and therefore has an error that we have not factored in because it is difficult to estimate.

A potential problem with NGS data is incomplete sampling of virus diversity in samples with low virus levels. If the sequencing library is dominated by a few template molecules, the TI estimate might be erroneously short. A related problem is due to the reduced ability of NGS to correctly estimate TI in infections that were founded by multiple virions. Two of our 11 patients showed evidence of such multiple infections. Consequentially, the duration of infection of the earliest samples of these two patients was overestimated. Interestingly, the overestimation of TI diminished over time. This happens because excess diversity is removed because favored iSNVs are selected for over time [22]. Infections founded by more than virion can be partially resolved by full genomes obtained from SGS [6], but this approach is too labour-intensive for broad use. A potential way to alleviate errors in TI estimates due to low template input or multiple infections would be to combine our NGS diversity measure with other biomarkers. Such multi-assay algorithms (MAAs) can improve the precision of classification into recent vs. long-term infection [31, 32] In a separate project we currently are investigating the utility of a MAA based on CD4, BED and fraction of polymorphic sites in Sanger *pol* gene sequences for TI estimation (Giardina, in preparation). We plan to go on to investigate if incorporation of virus diversity determined by NGS in the MAA can further improve estimation of TI.

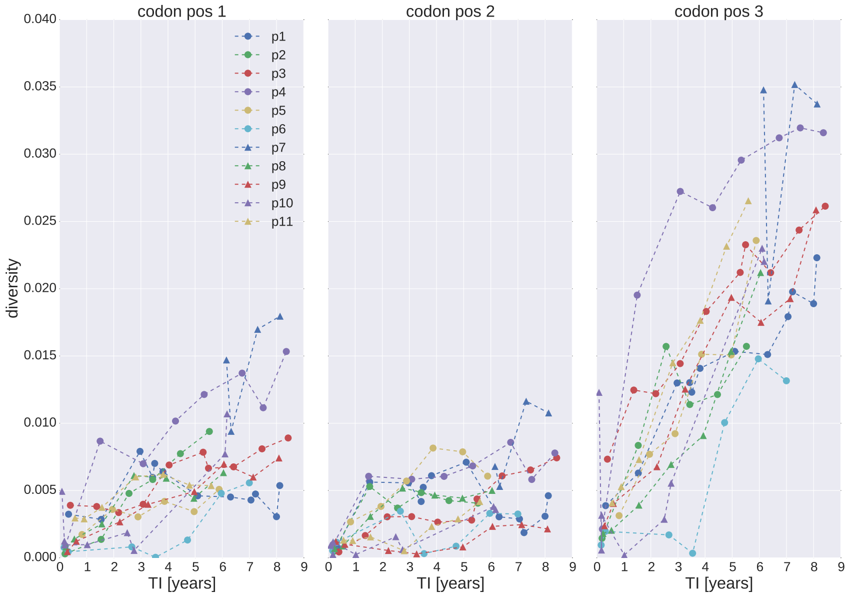
Finally, NGS is not yet part of routine diagnostics for HIV resistance. However, in the coming years NGS can be expected to replace Sanger sequencing for clinical HIV-1 resistance testing, which is recommended for all newly diagnosed patients (in developed countries). Thus, while our method for estimating TI from NGS data currently requires extra laboratory work, NGS data is likely to become increasingly available as part of routine HIV-1 care, which will increase the utility of our method.

Conclusion

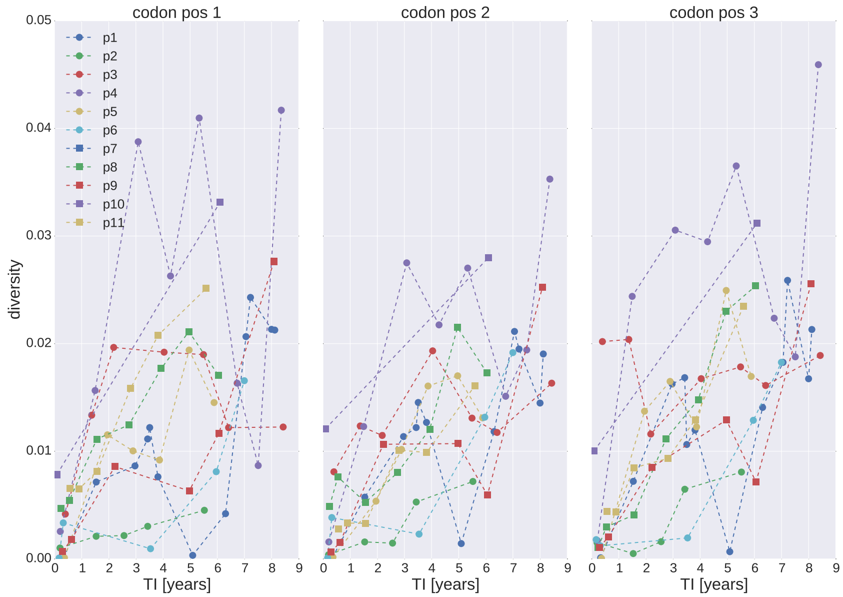
In conclusion, we show that sequence diversity determined by NGS can be used to estimate time since HIV-1 infection with a precision that is better than most alternative biomarkers. Importantly, TI can be estimated many years after infection, whereas most alternative methods only categorize patients as being recently or long-term infected or are less precise. We found that TI was most accurately estimated using 3rd codon positions in *pol* sequences with a 0% cutoff for iSNVs and that Hamming distances was the preferred distance measure. Samples with low virus levels and infections established by multiple virions can give rise to misleading levels of virus diversity. Algorithms based on NGS diversity in combinations with other biomarkers may prove to be very useful.

Supporting information

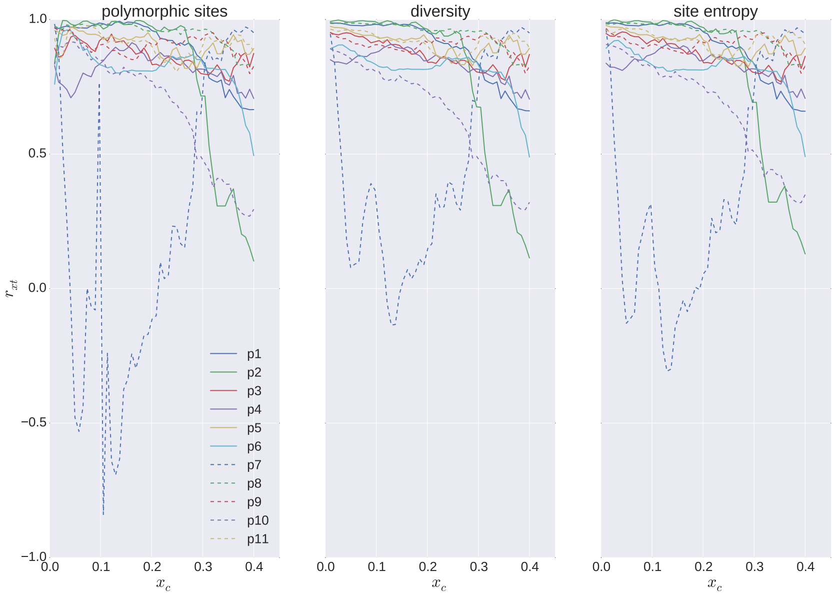
S1 Fig. Diversity in *gag* as a function of the time since infection (TI). (Diversity measure: Hamming distance, $x_c = 0$.)



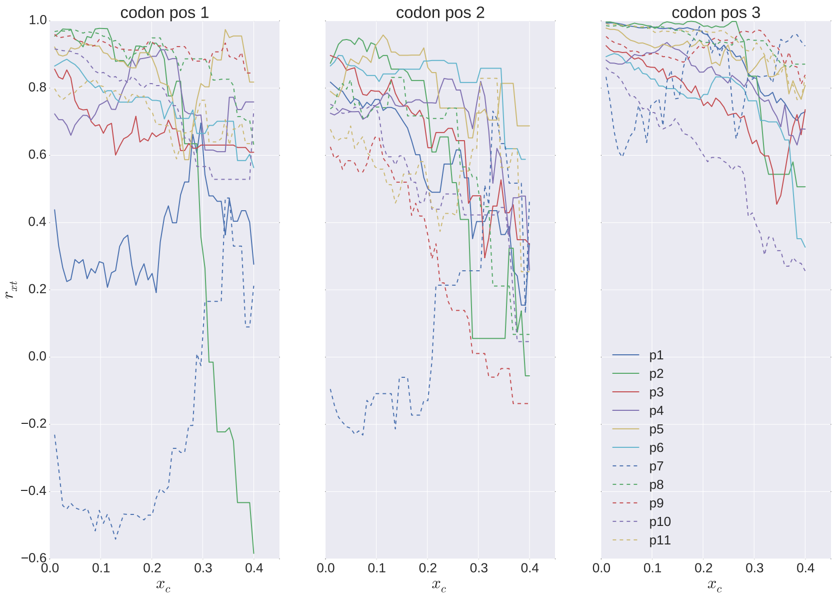
S2 Fig. Diversity in *env* as a function of the time since infection (TI). (Diversity measure: Hamming distance, $x_c = 0$.)



S3 Fig. Pearson correlation coefficients for different diversity measures, including all sites. (Genetic region: *pol*.)

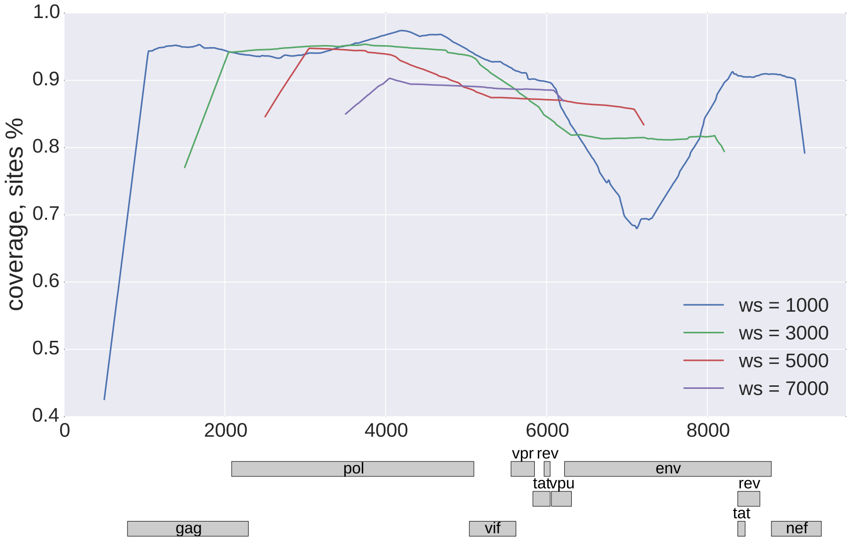


S4 Fig. Pearson correlation coefficients for Hamming distance (diversity), by codon position. (Genetic region: *pol*.)



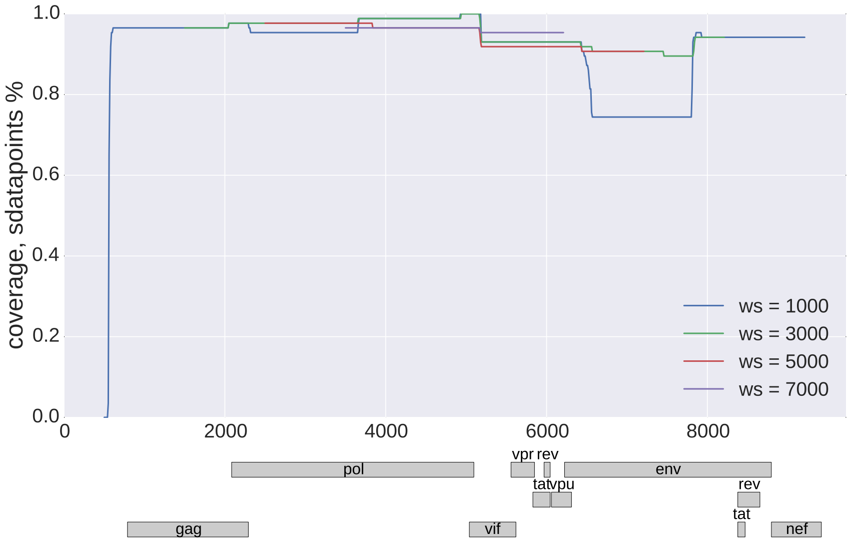
371

S5 Fig. Average percentage of usable sites in a genome window, along the genome. (Genetic region: *pol*, diversity measure: Hamming distance, $x_c = 0$.) 372 373

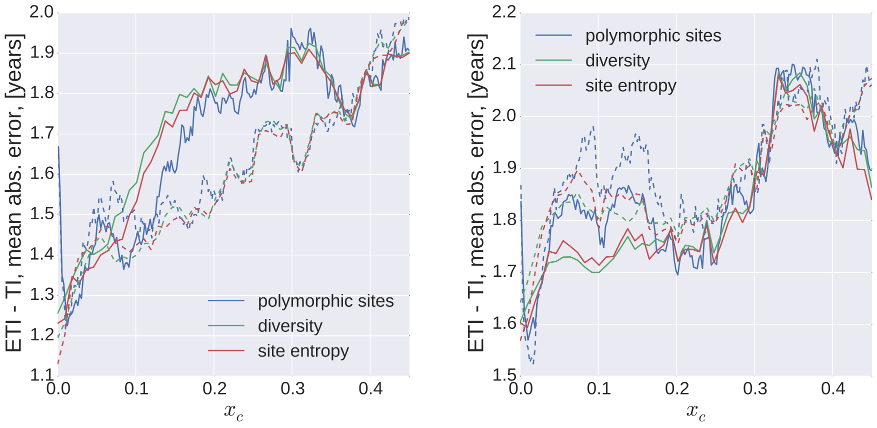


374

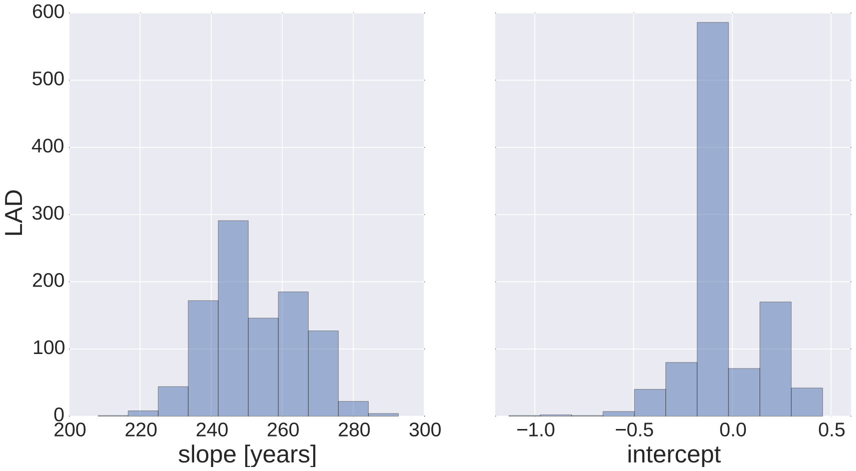
S6 Fig. Percentage of usable data-points along the genome. (Genetic region: *pol*, diversity measure: Hamming distance, $x_c = 0$.) 375 376



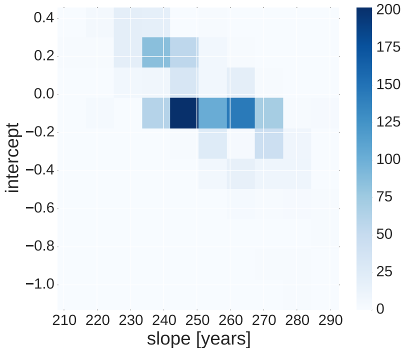
S7 Fig. Mean absolute error as a function of the low-frequency cutoff (x_c). Different diversity measures perform very similarly when the cut-off x_c is large. Hamming distance and entropy outperform fraction of polymorphic sites for low x_c. This graph is based on diversity in *gag* (left) and *env* (Right). Solid lines correspond to using all sites, dashed - only 3rd codon positions.



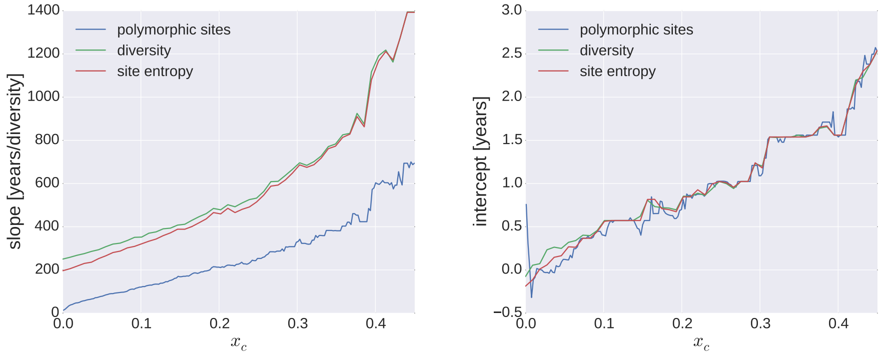
S8 Fig. Bootstrap of slopes and intercept values. (Gene: 3rd codon positions in *pol*, diversity measure: Hamming distance, $x_c = 0$.)



S9 Fig. Two-dimensional histogram of bootstrapped slopes and intercepts. (Genetic region: 3rd codon positions in *pol*, diversity measure: Hamming distance, $x_c = 0$.)



S10 Fig. Dependence of the slope and intercept in the cutoff. (Genetic region: 3rd codon positions in *pol*, diversity measure: Hamming distance.)



S1 Appendix Linear fitting procedures Given a set of K time points (t_k) and corresponding values of mean/average diversity (D_k) we infer slope and intercept pair (s, t_0) by performing linear regression. Here we used least absolute deviation (LAD) method, which means minimizing the mean absolute deviation (MAE)

$$S_{LAD} = \frac{1}{K}\sum_{k=1}^{K} |sD_k + t_0 - t_k| \qquad (5)$$

in respect to parameters s and t_0;.

LAD has an advantage over the traditional optimal least squares approach (OLS) in that it is less sensitive to the possible outliers. Minimization can be done either using one of the efficient linear programming techniques, or by considering all pairs s, t_0 corresponding to straight lines passing through all possible pairs of data points, and choosing the pair corresponding to the least error. (It is known that the regression line corresponding to the minimum of S_{LAD} passes through at least two of the data points.) We used the latter approach, as it is quite efficient for rather small number of data points that we had to deal with, but also could be modified to perform in parallel calculations for several genetic regions, which was useful for the sliding window analysis of the error along the genome.

S2 Appendix Moving average The moving average for the error dependence on the time of infection was calculated using the mean absolute error for a set of n points:

$$S_q = \frac{1}{n}\sum_{k=q+1}^{q+n} |sD_k + t_0 - t_k| \qquad (6)$$

while the corresponding times of infection where evaluated as

$$\tau_q = \frac{1}{n}\sum_{k=q+1}^{q+n} t_k. \qquad (7)$$

The advantage of this technique over other time averaging approaches is that one avoids significant variation of the averaging error, since the number of data points in the average is fixed to n.

S1 Table Recommended slope and intercept values depending on the cutoff.

	slope and intercept			**slope only**	
cutoff (x_c)	slope (s)^a	intercept (t_0)^b	MAE^b	slope (s)^a	MAE^b
0.05	78.16	0.12	1.02	80.50	1.01
0.10	120.79	0.40	1.01	129.92	1.05
0.15	168.42	0.52	1.05	183.90	1.11
0.20	211.29	0.70	1.18	233.76	1.23
0.25	253.30	1.03	1.32	312.03	1.43
0.30	328.22	1.09	1.45	404.20	1.60
0.35	381.41	1.56	1.59	565.60	1.72
0.40	603.82	1.54	1.68	834.63	1.85
0.45	694.28	2.53	1.87	1157.02	2.62

Genetic region: 3rd codon positions in *pol*, diversity measure: average number of polymorphic sites. ^{a}in years/diversity; ^{b}in years.

S2 Table Recommended slope and intercept values depending on the cutoff.

	slope and intercept			slope only	
cutoff (x_c)	slope (s)^a	intercept (t_0)^b	MAE^b	slope (s)^a	MAE^b
0.00	196.34	-0.19	0.85	189.31	0.85
0.05	256.99	0.26	0.95	271.60	0.95
0.10	321.13	0.41	1.01	350.22	1.04
0.15	387.34	0.70	1.11	447.40	1.15
0.20	465.61	0.74	1.21	526.40	1.27
0.25	518.50	1.02	1.34	678.49	1.44
0.30	674.63	1.18	1.46	831.50	1.61
0.35	773.04	1.54	1.59	1170.50	1.72
0.40	1211.36	1.54	1.70	1676.82	1.85
0.45	1392.68	2.54	1.87	2310.37	2.61

Genetic region: 3rd codon positions in *pol*, diversity measure: average number of polymorphic sites ^{a}in years/diversity; ^{b}in years.

S3 Table Recommended slope and intercept values depending on the cutoff.

	slope and intercept			slope only	
cutoff (x_c)	slope (s)^a	intercept (t_0)^b	MAE^b	slope (s)^a	MAE^b
0.00	574.51	-0.52	0.85	517.61	0.88
0.05	691.67	-0.02	0.94	673.42	0.92
0.10	795.07	0.11	0.91	830.76	0.92
0.15	972.00	0.21	1.02	1029.23	1.03
0.20	1111.40	0.32	1.10	1217.34	1.10
0.25	1292.90	0.57	1.20	1434.69	1.25
0.30	1547.65	0.79	1.30	1788.47	1.37
0.35	1987.89	0.92	1.43	2531.70	1.46
0.40	2415.57	1.31	1.68	3345.15	1.75
0.45	3131.26	2.23	1.85	5204.78	2.27

Genetic region: all sites in *pol*, diversity measure: Hamming distance. ^{a}in years/diversity; ^{b}in years.

S4 Table Recommended slope and intercept values depending on the cutoff.

	slope and intercept			slope only	
cutoff (x_c)	slope (s)^a	intercept (t_0)^b	MAE^b	slope (s)^a	MAE^b
0.05	181.68	-0.17	1.08	171.09	1.10
0.10	283.41	0.10	0.90	287.39	0.88
0.15	389.90	0.17	0.96	401.41	0.96
0.20	490.18	0.28	1.07	528.22	1.09
0.25	587.25	0.57	1.21	652.39	1.23
0.30	734.72	0.78	1.30	862.30	1.36
0.35	968.38	0.88	1.44	1229.70	1.45
0.40	1208.07	1.26	1.70	1663.00	1.77
0.45	1595.22	2.23	1.82	2599.17	2.27

Genetic region: 3rd codon positions in *pol*, diversity measure: average number of polymorphic sites. ^{a}in years/diversity; ^{b}in years.

S5 Table Recommended slope and intercept values depending on the cutoff.

	slope and intercept			slope only	
cutoff (x_c)	slope (s)^a	intercept (t_0)^b	MAE^b	slope (s)^a	MAE^b
0.00	445.61	-0.75	0.91	390.96	0.93
0.05	595.29	-0.12	0.97	582.31	0.95
0.10	745.57	0.07	0.89	755.98	0.88
0.15	909.35	0.19	0.99	944.97	1.00
0.20	1089.94	0.29	1.09	1160.02	1.10
0.25	1236.10	0.57	1.20	1370.86	1.24
0.30	1511.44	0.79	1.29	1772.83	1.35
0.35	1941.61	0.89	1.42	2456.16	1.44
0.40	2405.31	1.35	1.66	3229.49	1.74
0.45	3062.32	2.28	1.84	5200.83	2.26

Genetic region: 3rd codon positions in *pol*, diversity measure: average number of polymorphic sites. ^{a}in years/diversity; ^{b}in years.

Acknowledgments

We would like to express our gratitude to the study participants and to Johanna Brodin, Fabio Zanini, Lina Thebo, Christa Lanz and Göran Bratt for important contributions in the generation of the published data that was analyzed in this study.

Additional information

Competing interests: The authors declare that no competing interests exist.

Funding

European Research Council, Stg. 260686, principal investigator Richard A Neher; Swedish Research Council , K2014-57X-09935, principal investigator Jan Albert. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Conception and design: RAN, JA; Acquisition of data, Analysis and interpretation of data, Drafting or revising the article: VP, RAN, JA

References

1. Le Vu S, Pillonel J, Semaille C, Bernillon P, Le Strat Y, Meyer L, et al. Principles and uses of HIV incidence estimation from recent infection testing–a review. Euro surveillance: bulletin europeen sur les maladies transmissibles= European communicable disease bulletin. 2008;13(36):537–545.

2. Group IACPW, et al. More and better information to tackle HIV epidemics: towards improved HIV incidence assays. PLoS Med. 2011;8(6):e1001045.

3. World Health Organization. When and how to use assays for recent infection to estimate HIV incidence at a population level. 2011.

4. Busch MP, Pilcher CD, Mastro TD, Kaldor J, Vercauteren G, Rodriguez W, et al. Beyond detuning: 10 years of progress and new challenges in the development and application of assays for HIV incidence estimation. Aids. 2010;24(18):2763–2771.
5. Murphy G, Parry J. Assays for the detection of recent infections with human immunodeficiency virus type 1. Euro Surveill. 2008;13(36):314–320.
6. 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–F19.
7. Moyo S, Vandormael A, Wilkinson E, Engelbrecht S, Gaseitsiwe S, Kotokwe KP, et al. Analysis of Viral Diversity in Relation to the Recency of HIV-1C Infection in Botswana. PloS one. 2016;11(8):e0160649.
8. Guy R, Gold J, Calleja JMG, Kim AA, Parekh B, Busch M, et al. Accuracy of serological assays for detection of recent infection with HIV and estimation of population incidence: a systematic review. The Lancet infectious diseases. 2009;9(12):747–759.
9. Bärnighausen T, McWalter TA, Rosner Z, Newell ML, Welte A. Review Article: HIV Incidence Estimation Using the BED Capture Enzyme Immunoassay: Systematic Review and Sensitivity Analysis. Epidemiology. 2010; p. 685–697.
10. Lazarus J, Jürgens R, Weait M, Phillips A, Hows J, Gatell J, et al. Overcoming obstacles to late presentation for HIV infection in Europe. HIV medicine. 2011;12(4):246–249.
11. Sabin CA, Schwenk A, Johnson MA, Gazzard B, Fisher M, Walsh J, et al. Late diagnosis in the HAART era: proposed common definitions and associations with mortality. AIDS (London, England). 2010;24(5):723–727.
12. Goujard C, Bonarek M, Meyer L, Bonnet F, Chaix ML, Deveau C, et al. CD4 cell count and HIV DNA level are independent predictors of disease progression after primary HIV type 1 infection in untreated patients. Clinical Infectious Diseases. 2006;42(5):709–715.
13. Lodi S, Phillips A, Touloumi G, Geskus R, Meyer L, Thiébaut R, et al. Time from human immunodeficiency virus seroconversion to reaching CD4+ cell count thresholds¡ 200,¡ 350, and¡ 500 cells/mm3: assessment of need following changes in treatment guidelines. Clinical infectious diseases. 2011;53(8):817–825.
14. Lodi S, Guiguet M, Costagliola D, Fisher M, de Luca A, Porter K. Kaposi sarcoma incidence and survival among Hiv-infected homosexual men after Hiv seroconversion. Journal of the National Cancer Institute. 2010;102(11):784–792.
15. Minga A, Lewden C, Gabillard D, Bomisso G, Toni Td, Emième A, et al. CD4 eligibility thresholds: an analysis of the time to antiretroviral treatment in West African HIV-1 seroconverters. AIDS (London, England). 2011;25(6):819–823.
16. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. Journal of virology. 1999;73(12):10489–10502.

17. Kouyos R, von Wyl V, Yerly S, Böni J, Rieder P, Joos B, et al. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2011;52(4):532–539.
18. Ragonnet-Cronin M, Aris-Brosou S, Joanisse I, Merks H, Vallée D, Caminiti K, et al. Genetic Diversity as a Marker for Timing Infection in HIV-Infected Patients: Evaluation of a 6-Month Window and Comparison With BED. Journal of Infectious Diseases. 2012;206(5):756–764.
19. Andersson E, Shao W, Bontell I, Cham F, Wondwossen A, Morris L, 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. Infection, Genetics and Evolution. 2013;18:125–131.
20. Allam O, Samarani S, Ahmad A. Hammering out HIV-1 incidence with Hamming distance. AIDS. 2011;25(16):2047–2048.
21. Cousins MM, Konikoff J, Laeyendecker O, Celum C, Buchbinder SP, Seage GR, et al. HIV diversity as a biomarker for HIV incidence estimation: including a high-resolution melting diversity assay in a multiassay algorithm. Journal of clinical microbiology. 2014;52(1):115–121.
22. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. eLife Sciences. 2016;4:e11282. doi:10.7554/eLife.11282.
23. Schuurman R, Brambilla D, de Groot T, Huang D, Land S, Bremer J, et al. Underestimation of HIV type 1 drug resistance mutations: results from the ENVA-2 genotyping proficiency program. AIDS research and human retroviruses. 2002;18(4):243–248.
24. Parkin N, Bremer J, Bertagnolio S. Genotyping external quality assurance in the World Health Organization HIV drug resistance laboratory network during 2007–2010. Clinical infectious diseases. 2012;54(suppl 4):S266–S272.
25. Land S, Cunningham P, Zhou J, Frost K, Katzenstein D, Kantor R, et al. TREAT Asia Quality Assessment Scheme (TAQAS) to standardize the outcome of HIV genotypic resistance testing in a group of Asian laboratories. Journal of virological methods. 2009;159(2):185–193.
26. Zanini F, Brodin J, Albert J, Neher RA. Error rates, PCR recombination, and sampling depth in HIV-1 Whole Genome Deep Sequencing. Virus research. 2016;doi:10.1016/j.virusres.2016.12.009.
27. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and. Bioinformatics. 2009;25(11):1422–1423. doi:10.1093/bioinformatics/btp163.
28. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science Engineering. 2011;13(2):22–30. doi:10.1109/MCSE.2011.37.
29. Hunter JD. Matplotlib: A 2D graphics environment. Computing In Science & Engineering. 2007;9(3):90–95.
30. Zanini F, Puller V, Brodin J, Albert J, Neher RA. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus. 2017;3:1.

31. Moyo S, Wilkinson E, Novitsky V, Vandormael A, Gaseitsiwe S, Essex M, et al. Identifying recent HIV infections: From serological assays to genomics. Viruses. 2015;7(10):5508–5524.

32. Laeyendecker O, Kulich M, Donnell D, Komárek A, Omelka M, Mullis CE, et al. Development of methods for cross-sectional HIV incidence estimation in a large, community randomized trial. PloS one. 2013;8(11):e78818.