Detecting Selection in the HIV-1 Genome during Sexual Transmission Events

Journal Article

Author(s):
Seifert, David; Joos, Beda; Braun, Dominique L.; Oberle, Corinna S.; Schenkel, Corinne D.; Kuster, Herbert; Grube, Christina; Böni, Jürg; Yerly, Sabine; Aubert, Vincent; Klimkait, Thomas; Günthard, Huldrych F.; Beerenwinkel, Niko iD; Metzner, Karin J.; Swiss HIV Cohort Study

viruses

MDPI

Article

Detecting Selection in the HIV-1 Genome during Sexual Transmission Events

David Seifert [1,2], Beda Joos [3], Dominique L. Braun [3,4], Corinna S. Oberle [3,4], Corinne D. Schenkel [3,4], Herbert Kuster [3,4], Christina Grube [3], Jürg Böni [4], Sabine Yerly [5], Vincent Aubert [6], Thomas Klimkait [7], Huldrych F. Günthard [3,4,*], Niko Beerenwinkel [1,2,*], Karin J. Metzner [3,4,*] and on behalf of the Swiss HIV Cohort Study [†]

1   Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland; david.seifert@bsse.ethz.ch
2   SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland
3   Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, 8091 Zurich, Switzerland; beda.joos@bluewin.ch (B.J.); dominique.braun@usz.ch (D.L.B.); corinna_oberle@hotmail.com (C.S.O.); corinneschenkel@hotmail.com (C.D.S.); herbert.kuster@usz.ch (H.K.); christina.grube@usz.ch (C.G.)
4   Institute of Medical Virology, University of Zurich, 8091 Zurich, Switzerland; boeni.juerg@virology.uzh.ch
5   Laboratory of Virology, University Hospital Geneva, 1205 Geneva, Switzerland; sabine.yerly@hcuge.ch
6   University Hospital Lausanne, Service of Immunology and Allergy, University Hospital Center, 1011 Lausanne, Switzerland; vincent.aubert@chuv.ch
7   Molecular Virology, Department of Biomedicine-Petersplatz, University of Basel, 4009 Basel, Switzerland; thomas.klimkait@unibas.ch
*   Correspondence: huldrych.guenthard@usz.ch (H.F.G.); niko.beerenwinkel@bsse.ethz.ch (N.B.); karin.metzner@usz.ch (K.J.M.)
†   The list of the Swiss HIV Cohort Study study members can be found at the end of the manuscript (Acknowledgements).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: Little is known about whether and how variation in the HIV-1 genome affects its transmissibility. Assessing which genomic features of HIV-1 are under positive or negative selection during transmission is challenging, because very few virus particles are typically transmitted, and random genetic drift can dilute genetic signals in the recipient virus population. We analyzed 30 transmitter–recipient pairs from the Zurich Primary HIV Infection Study and the Swiss HIV Cohort Study using near full-length HIV-1 genomes. We developed a new statistical test to detect selection during transmission, called Selection Test in Transmission (SeTesT), based on comparing the transmitter and recipient virus population and accounting for the transmission bottleneck. We performed extensive simulations and found that sensitivity of detecting selection during transmission is limited by the strong population bottleneck of few transmitted virions. When pooling individual test results across patients, we found two candidate HIV-1 genomic features for affecting transmission, namely amino acid positions 3 and 18 of Vpu, which were significant before but not after correction for multiple testing. In summary, SeTesT provides a general framework for detecting selection based on genomic sequencing data of transmitted viruses. Our study shows that a higher number of transmitter–recipient pairs is required to improve sensitivity of detecting selection.

Keywords: HIV-1; transmission; transmitter–recipient pairs; Selection Test in Transmission (SeTesT); Vpu; ZPHI; SHCS

1. Introduction

While major achievements have been made in containing the HIV/AIDS pandemic, the worldwide prevalence of people living with human immunodeficiency virus type 1 (HIV-1) is continuing to increase [1]. HIV-1 is transmitted via body fluids containing blood, semen, vaginal secretions, or breast milk. The main route of transmission occurs via sexual

intercourse. On a global scale, 80% of transmissions are due to heterosexual transmission, with over 90% occurring in the resource-limited world [2].

The human body presents multiple barriers during and after transmission that HI virions have to overcome to establish a persistent infection [3]. Crossing the genital and rectal mucosa presents itself as the first physical barrier. Once in the body, HIV-1 has to evade innate and adaptive immune responses and enter a target cell in order to replicate [3,4].

All of the aforementioned hurdles have the potential to exert selective pressure on the HIV-1 genome during the establishment of the infection. Whether selection acts on the HIV-1 genome during transmission is poorly understood [5]. It depends on host factors and on the intrinsic fitness landscape of the virus itself [6]. The viral fitness landscape is an association of a positive value, the replicative fitness of the virus, to each viral genotype. Furthermore, the fitness landscape may also depend on the host environment, and can therefore also change over time [7]. It has been shown that intra-host fitness landscapes can be inferred from either cross-sectional [8] or time-series [9] data. In contrast to the HIV-1 fitness landscape during replication in a host, much less is known about the fitness landscape at transmission to a new host, i.e., the transmissibility landscape of HIV-1. This high-dimensional evolutionary parameter can only be studied in transmission events, which are much more difficult to sample compared to longitudinal samples of one patient. A better understanding of the transmissibility landscape is crucial, for instance, for designing broadly neutralizing antibodies in the quest for an HIV vaccine [10]. Broadly neutralizing antibodies work by attacking epitopes under negative selection, that is, antigenic determinants that are conserved and that need to traverse a deep fitness valley in order to escape the immune system.

Historically, finding loci of the HIV-1 genome under selection during transmission has been already approached in the mid 1990s by analyzing the HIV-1 genetic makeup in sexual transmission pairs and in mother-to-child transmissions [11,12]. Later, with the discovery of HIV-1 co-receptors, it has been observed that viral strains preferring the CCR5 co-receptor are selected for during transmission, which appears to highlight a significant difference between the transmission and the replication fitness landscape [4]. Studies on intra-host diversification indicate an incomplete switch of co-receptor usage, suggesting higher intra-host fitness for HIV-1 populations employing a mix of both CCR5 and CXCR4 as co-receptors. Other studies have shown that shorter segments of the variable regions of the envelope protein gp120 and less potential N-linked glycosylation sites are preferred during transmission and may hence confer a selective advantage particularly in transmissions of HIV-1 subtypes A, C, and D and possibly also subtype B [13–15]. Studying heterosexual transmission of HIV-1 has shown that on a cohort level, amino acids that are abundant in the population of HIV-infected individuals are selected for during transmission [16]. This observation supports the notion that there are unique mechanisms of selection during transmission, which differ from the selective forces acting during the arms race between the virus and the host immune response as the untreated disease progresses to AIDS.

A great deal of effort has been invested into determining the size of the transmission bottleneck, that is, the number of transmitted viruses causing a clinically persistent infection. This number cannot be observed directly and is difficult to estimate. Different studies have supported a very small number of viruses—in most cases just one—that establish infection [17–20]. This is in line with statistical expectations under a neutral transmission model, given that the likelihood of transmitting HIV-1 is very low in general [2]. Studying selection during sexual transmission events is complicated by several factors. The correct pairing of transmitter–recipient partners in sexual transmission events has been hampered greatly due to incomplete sampling, especially of the transmitting partners, as these are often unknown to the recipient. The issue is exacerbated by legal and social complications that arise from admitting transmission of an incurable disease [21].

With the advent of next-generation sequencing (NGS), viral populations can be studied in great detail, where previous single-genome amplification assays were extremely

laborious and could not sample the viral population to the required depth [22]. With sufficient depth of coverage even minor variants can be detected reliably [23].

Probabilistic models of the transmission bottleneck have been formulated that account for selection in the presence of strong selection. Bergstrom et al. [24] formulated the transmission problem using the quasispecies model. Sobel Leonard et al. [25] devised a model that estimates the bottleneck size without estimating selective pressures during the early stage of establishing the disease. Lumby et al. [26] proposed a fully probabilistic model that infers selection and the transmission bottleneck. More recent models have combined intra- and inter-host selection into an overall model for global influenza evolution [27].

Here, we develop a new statistical test that accounts for the strong bottleneck the virus population experiences during HIV-1 transmission. The test takes viral genotype counts derived from patient transmitter and recipients as input to detect selection. We found that in practice detecting selection requires either a dominant variant in the transmitter disappearing in the recipient or a minor variant in the transmitter reaching fixation in the recipient after transmission. In order to test for selection, we analyzed in 30 resolved transmitter–recipient pairs all amino acid loci of the HIV-1 genome for which more than one amino acid was found in at least six transmitters. We further assessed the amino acid sequence of the V1/V2, V3, V4 and V5 loops of the gp120 locus for signatures of selection.

2. Materials and Methods

2.1. Study Design

Transmission pairs were obtained from the Swiss HIV Cohort Study (SHCS) and in the Zurich Primary HIV Infection Study (ZPHI). The SHCS is a nationwide, clinic-based cohort enrolling at least 53% of all HIV-1-infected adults ever diagnosed in Switzerland [28,29]. Approximately 80% of those patients were included in genotypic HIV-1 drug resistance testing since 1996 [28,29]. The ZPHI is an observational, non-randomized, single-center cohort enrolling patients diagnosed with an acute or recent primary HIV-1 infection (www.clinicaltrials.gov accessed on 9 December 2021; ID NCT00537966) [20,30,31]. Blood collections and sampling of plasma are scheduled every 3 and 6 months in the ZPHI and SHCS, respectively. The SHCS and the ZPHI are approved by the ethics committee of the participating institutions (Kantonale Ethikkommission Bern (#21/88), Kanton St. Gallen Ethikkommission (#12/003), Comite departemental d'éthique des specialites medicales et de medecine communautaire et de premier recours, Geneva (#01-142), Kantonale Ethik-Kommission Zürich (#EK-793), Comitato etico cantonale, Bellinzona (#Rif. CE 813), Commission cantonale d'éthique de al recherche sur l'étre humain, Lausanne (#131/01), Ethikkommission beider Basel (#688)) and written informed consent was obtained from all participants. Patients harboring HIV-1 subtypes other than subtype B were excluded in all further analyses.

2.2. Determining Transmitter–Recipient Relationships

Genotypic HIV-1 drug resistance testing, i.e., protease and reverse transcriptase sequences, were used to identify transmission pairs by assessing phylogenetic relationships, as previously described [31]. To confirm potential transmission pairs, near full-length HIV-1 genome sequences, clinical and laboratory data, for instance, viral loads and transmission risk group, were used. For the potential recipients, we used the first available plasma sample for sequencing of near full-length HIV-1 genomes. For potential transmitters, we chose the closest plasma sample to the estimated date of transmission (EDT), which could also be prior to transmission (Figure 1). This procedure resulted in 30 high-confidence transmitter–recipient pairs (Supplementary Materials, Section S1).
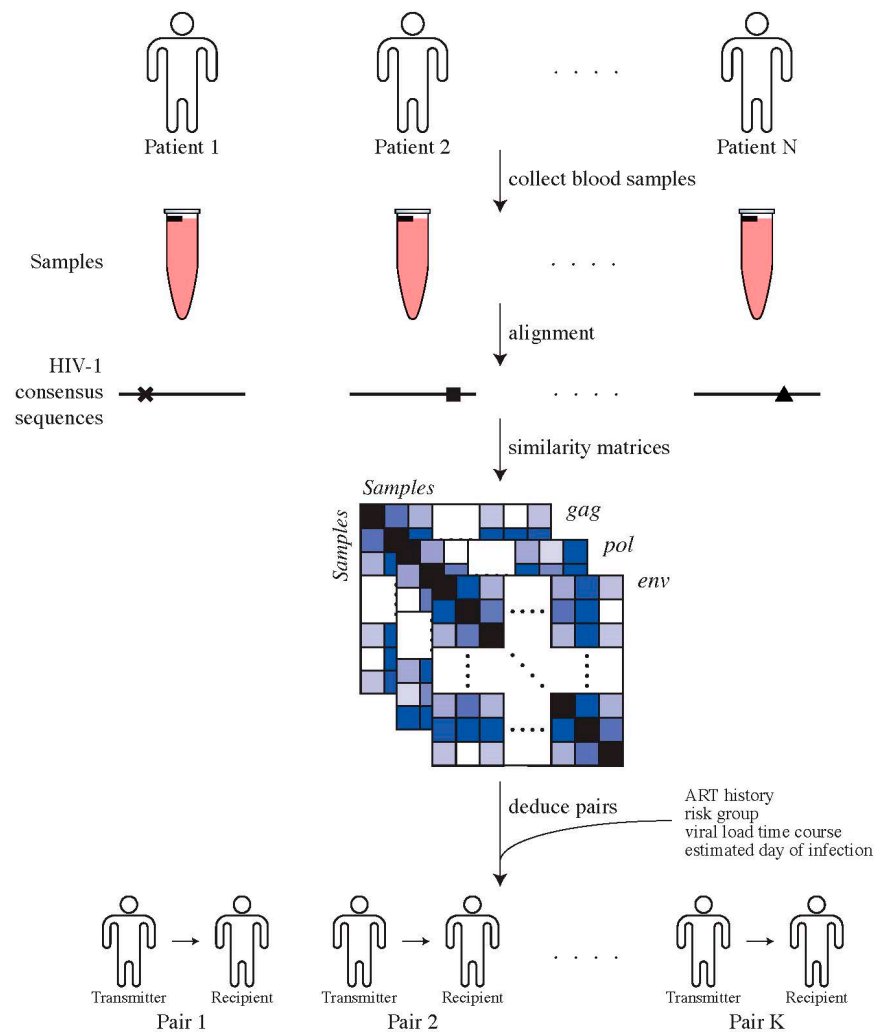
Figure 1. Determining transmitter–recipient relationships in the ZPHI and SHCS cohorts. We determine direct transmitter–recipient relationships by calling HIV-1 consensus sequences for all patients' samples, generated using ngshmmalign. Pairwise similarity distances between patients' samples were computed and combined with additional clinical and epidemiological data in order to identify transmitter–recipient pairs.

The estimated date of transmission was determined as previously described [30]. Briefly, clinical and laboratory data such as known risk situations, appearance of first symptoms, earlier negative HIV-1 test results, avidity assays and Western blot results were considered in the estimation. Infections within 90 and 180 days after the EDT were defined as acute and recent, respectively [30].

2.3. Next-Generation Sequencing (NGS) Data Generation and Analysis

Viral genomic sequences were generated via the Illumina MiSeq Desktop Sequencer v2 2 × 250 kit for primary virus isolate and plasma full-length sequencing as described in [32]. Sequencing was performed using five amplicons covering the HIV-1 genome from the 5′ LTR to the 3′ LTR. In the first step of our analysis, we merged all Illumina MiSeq runs per patient. We performed clipping and quality filtering using PRINSEQ [33], where we clipped bases with a Phred score below 30 in a sliding window of 10 bases from the 5′ and 3′ ends of every read. After clipping, we required reads to be at least 200 nt in length, allowed for a maximum of 4 ambiguous 'N' bases and required an average Phred score of at least 30 over the whole read. Reads missing a mate, either due to missing it right from the start or due to previously performed quality checks, were removed from the preprocessed data.

We aligned the quality-trimmed paired-end reads using the custom-made aligner ngshmmalign (https://github.com/cbg-ethz/ngshmmalign accessed on 9 December 2021) [23]. Briefly, this alignment tool employs a profile Hidden Markov Model (HMM) to account for insertions and deletions that occur as part of the natural evolution of HIV-1. Certain loci, such as p6, gp120 and nef are prone to structural variants. The hypervariable V1 to V5 loops in gp120 are especially likely to include a large number of insertions and deletions relative to the reference strain HIV-1 HXB2 (GenBank accession number K03455). ngshmmalign accounts for biological insertions and removes conserved deletions with respect to the cohort consensus sequence to build a new consensus sequence. This patient-specific consensus sequence contains all insertions with respect to each patient, such that deletions make up all real structural variants per patient.

In order to standardize all patients to the same reference genome, we performed a multiple sequencing alignment of all patient consensus sequences and HIV-1 HXB2 using MAFFT [34]. The resulting multiple sequence alignment was used to convert each patient's alignment to a standardized alignment against HIV-1 HXB2, resulting in comparable loci between patients with respect to a standardized HIV-1 reference. From these alignments, we extracted the amino acid composition at every residue site of all open reading frames. We retained only amino acids with frequency above 0.75% at every locus, in order to avoid calling spurious variants arising due to sequencing or RT-PCR errors. The cutoff 0.75% represents a tradeoff between losing minor variants and calling variants erroneously (false positives). In all but the most extreme datasets analyzed by Schirmer et al. [35], 0.75% is higher than the mean error rate across both mate pairs.

After running our analysis pipeline, we have genomic consensus sequences of majority nucleotide bases, genomic consensus sequences with ambiguous nucleotide bases of more than 5% minor allele frequency, amino acid allele frequencies at every position of every open reading frame above 0.75% minor allele frequency and amino acid haplotypes in the hypervariable regions of gp120. Details of the next-generation sequencing characteristics are given in Supplementary Materials, Section S2.

2.4. Testing for Selection during Transmission

To determine whether the composition of the transmitted population in the recipient deviates from the one of the presumed transmitters, we developed a statistical test, which we coined Selection Test in Transmission (SeTesT, https://github.com/cbg-ethz/SeTesT accessed on 9 December 2021); a detailed description can be found in Supplementary Materials, Section S3). The genotypes we test for can be single-locus amino acids, nucleotides, or haplotype blocks, i.e., sequences of nucleotides or amino acids spanning multiple loci. The input data is a $2 \times K$ contingency table of absolute counts, where K denotes the number of genotypes. Entries in the two rows denote the number of times a genotype occurs among all sequencing reads obtained from the transmitter and recipient, respectively (Figure 2).

As the transmitted population undergoes a strong population bottleneck due to physical and immunological barriers, general methods for $2 \times K$ contingency tables cannot be employed, as they only capture the sampling variance and fail to account for other sources of variation, most notably downsampling during the transmission bottleneck. Therefore, in practice, such tests suffer from massively inflated false positive rates (Supplementary Materials, Section S4).
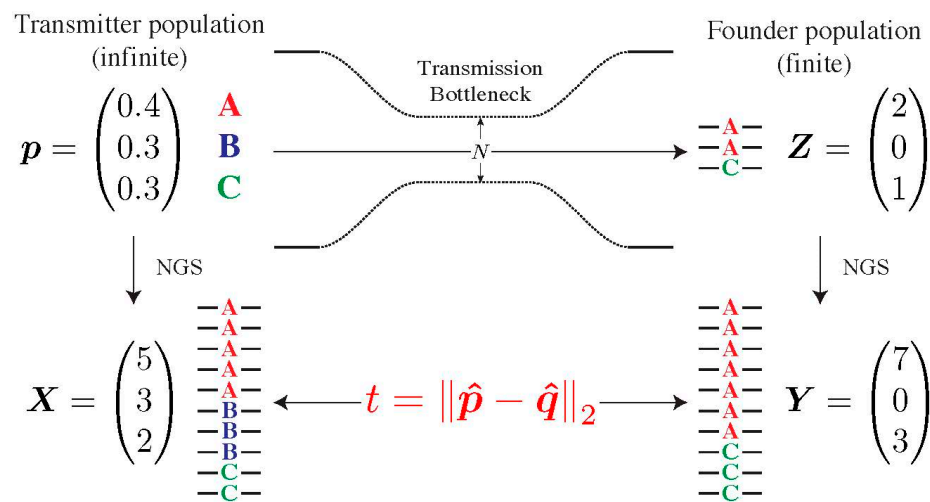
Figure 2. Modelling the transmission bottleneck. The transmission bottleneck of the HIV-1 population shapes its composition in the recipient drastically. Starting with an infinite pool of three genotypes A, B and C in the transmitter p, our model estimates deviations from neutrality by first passing this population through a strong bottleneck, yielding a small founder population in the recipient Z. Both the transmitter and founder population cannot be estimated directly, and can only be sampled by NGS, yielding the vector of counts X and Y. The Euclidean distance t between vectors p and q of relative abundances defines our test statistic.

To address this limitation, we developed SeTesT, a bespoke statistical test that accounts for variation in the unknown size of the transmission bottleneck (Supplementary Materials, Section S3). Briefly, we developed a probabilistic graphical model to derive the genotype counts of both the transmitter and recipient virus population from the viral NGS data (Supplementary Materials, Section S3.1). We tested the null hypothesis that the fitness landscape during transmission is flat, that is, all genotypes have the same fitness (Supplementary Materials, Section S3.2). The alternative hypothesis is that there exists at least one genotype that possesses a selective advantage during transmission. Additionally, we modelled the bottleneck explicitly using a single-generation Wright-Fisher process, where we assumed a mean bottleneck size of around 1.78 virions per transmission event [17] (Supplementary Materials, Section S3.3). SeTesT test is based on comparing the genotype distributions directly using the Euclidean distance, rather than estimating any model parameters. Significant differences between the two distributions indicate deviation from neutrality during transmission. For multi-locus genotypes, we adapted the test statistic to account for divergent evolution between transmitter and recipient genotypes during the time from transmission to sample collection (Supplementary Materials, Section S3.4).

2.5. Assessing the Specificity of Our Model

A crucial part for every statistical test is ensuring that it is statistically correct, that is, the probability of yielding a p-value smaller than α is at most α. Such tests are also referred to as unbiased [36]. Tests that yield on average a higher fraction of p-values below α for some given α are considered liberal or anti-conservative and do not allow for controlling the false positive rate. We simulated data from our model for different population settings and for different read coverages in the transmitter and analyzed whether our test yields inflated p-values or not (Supplementary Materials, Section S4).

2.6. Assessing the Sensitivity of SeTesT

In order to detect selection, we tested the statistical power, or sensitivity, of SeTesT by simulating a multinomial sampling process, followed by a read simulation step, and then running our test on these simulated data (Supplementary Materials, Section S5). We tested the sensitivity over a range of frequencies and fitness values of a single genotype in

the population having a positive or negative selective advantage. In addition, we assessed sensitivity for different read coverages in the recipient. Finally, we assessed sensitivity given 2, 3, 5, and 10 different genotypes in the transmitter.

2.7. Combining Data Sets to Increase Power

As different pairs of transmitters and recipients can be regarded as independent, multiple pairs can be combined into one summary statistic in order to gain power. We employed Fisher's method to combine multiple *p*-values of different transmitter–recipient pairs for the same locus. Fisher's sum-of-logs test has been shown to be one of the most powerful tests for combining independent tests in the setting of genome-wide association studies [37]. In order to assess the gain of power through pooling of pairs of patients, we simulated from our model and assessed the sensitivity of our model when using differing numbers of pooled *p*-values.

2.8. Assessing the Effect of Early Sampling

In order to estimate the effect of an evolving viral population on the results of our test, we have selected transmitter–recipient pairs 1, 21, and 29 (Supplementary Materials, Section S2), which all have less than 30 days to the estimated date of infection. By focusing on just these pairs, we can exclude selection affecting the population and changes in genotype distribution can more reliably be attributed to transmission alone.

3. Results

We developed a statistical test called Selection Test in Transmission (SeTesT) on the basis of a probabilistic model for detecting selection during viral transmission using viral NGS data of transmitter–recipient pairs (Figure 2). The test accounts for the strong population bottleneck during transmission, the finite sampling of the underlying virus populations that NGS data provides, and divergence of multi-locus genotypes due to the time-span between transmission and sampling. Simulating from the probabilistic model, we confirmed that SeTesT is capable of controlling the false positive rate. Below we first assess the statistical sensitivity of SeTesT under various parameters and then analyze NGS data from 30 curated transmitter–recipient pairs to detect selection during transmission in the HIV-1 genome.

3.1. Detecting Selection Is Compromised by the Transmission Bottleneck

We performed simulations with different transmission fitness landscapes in order to assess the power of our test to detect selection during transmission. We found that a significant obstacle in detecting transmission selection is the strong population bottleneck (Supplementary Materials, Section S4), while read coverage of the transmitted viral population does not improve statistical power beyond 100–1000 reads. SeTesT is very conservative for a regime in which only 1–2 viruses are likely transmitted (Supplementary Materials, Section S4.1). We found *p*-values to be inflated by a factor of 17.4 (Supplementary Materials, Section S4.1). In general, with an increasing number of genotypes, the test becomes less conservative (Supplementary Materials, Section S4.3).

The statistical power of SeTesT strongly depends on the composition of the population at the time of transmission (Supplementary Materials, Section S5), but it is practically independent of the read coverage of the recipient viral population (Supplementary Materials, Section S5.1). Detecting selection at a single locus is only possible when this specific genotype (i.e., nucleotide or amino acid) with high selective advantage exists at a low frequency in the transmitter and then reaches a very high frequency, which typically means fixation, in the recipient after transmission. On the other hand, detecting selection against a single genotype requires the other genotypes that are selected for to be minor variants in the transmitter. Detecting selection against a multi-locus genotype becomes more and more feasible for an increasing number of loci (Supplementary Materials, Section S5.2), as the increasing number of possible genotypes makes the test less conservative. At the

same time, detecting selection for a single locus is not aided by an increasing number of observed genotypes in the population, because only a situation in which the selected genotype reaches fixation can lead to a significant result.

Aggregating *p*-values for the same locus across multiple patients allows for near-perfect sensitivity with more than 30 transmitter–recipient pairs when either a genotype selected for in the recipient had a low frequency in the transmitter, or a highly frequent transmitter genotype was selected against and in the recipient. On the other hand, when a selected genotype is already abundant in the transmitter population or a genotype being selected against is absent from the transmitter population, little sensitivity can be achieved, even with a large aggregation of *p*-values across patients (Supplementary Materials, Section S5.3). Thus, we can detect selection during transmission if we observe an aggregate of minor variants being the source of the founder population. Detecting weak selection is significantly more challenging than detecting strong selection, because the required proportion of the selectively advantageous variant must pre-exist in the transmitter at a certain fraction that is neither too low nor too high (Figure 3). At extremely low abundances, the selectively advantageous variant cannot rich fixation with only weak selection, whereas under strong selection the selected variant can reach fixation over a larger range of relative abundances.
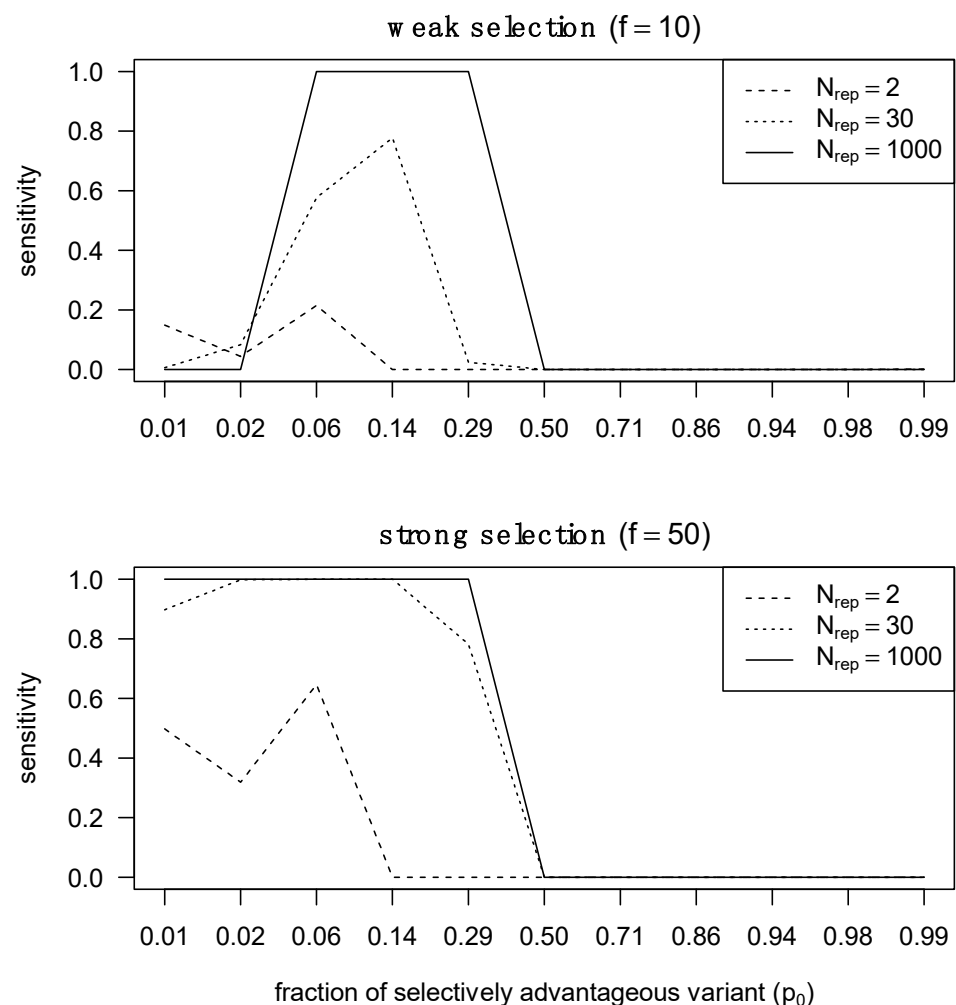

Figure 3. Sensitivity of detecting selection in different selection regimes with different numbers of pooled transmitter–recipient pairs. We assessed statistical sensitivity of detecting selection across 2, 30 and 1000 pairs. On the right, we performed the same analysis but in the presence of strong selection.

3.2. Assessing Selection in the HIV-1 Genomes of ZPHI and SHCS Transmitter–Recipient Pairs

We proceeded to determine likely transmitter–recipient pairs from previous exploratory studies on the ZPHI/SHCS cohorts based on the whole HIV-1 genome (Figure 2). We matched recipients to their prospective transmitters by determining empirical pairwise distances and acceptable cut-off thresholds on the overall distribution of these distances. Combined with ART history, risk group, viral load time course, and estimated date of infection we determined the final transmitter–recipient relationships (Supplementary Materials, Section S1). In total, we determined 30 high-confidence transmitter–recipient pairs. As part of this analysis, we could recover seven out of eight confirmed transmitter–recipient pairs that have been determined and comprehensively validated in another study on the basis of multiple assays on the env locus [31]. The most common route of transmission was between MSM (men who have sex with men), 28/30 (93%). The majority of transmitters were in the chronic phase of their HIV-1 infection at the estimated date of infection, 21/39 (70%). The median sampling date was 53 (IQR: −4; 96) and 49 (IQR: 31; 77) days after the estimated day of infection for transmitters and recipients, respectively (Supplementary Materials, Section S1).

We applied SeTesT in a whole-genome fashion to each amino acid at each single locus of the HIV-1 genome, excluding the heterogeneous V1/V2, V3, V4 and V5 loops of the gp120 open reading frame, because these hypervariable regions showed a very high degree of heterogeneity in multiple sequence alignments [38,39] leading to spurious calls (Supplementary Materials, Section S2). Instead, we considered multi-locus genotypes in these regions and analyzed the complete protein sequences of the V1/V2, V3, V4 and V5 loops of gp120 open reading frame. We analyzed a total of 2773 individual amino acid loci of the HIV-1 genome (Supplementary Materials, Section S6). Approximately half of the loci (1296) could not be tested, due to lack of multiple amino acids in the transmitter or failed amplicons. A further 579 loci allowed only for one test across the 30 pairs. We could perform a maximum of 12 tests for four loci (Figure 4). Assessing the impact of early HIV-1 infection (sampled within 30 days) compared to established HIV-1 infection (over 30 days) yielded no significant locus under selection after pooling across pairs.
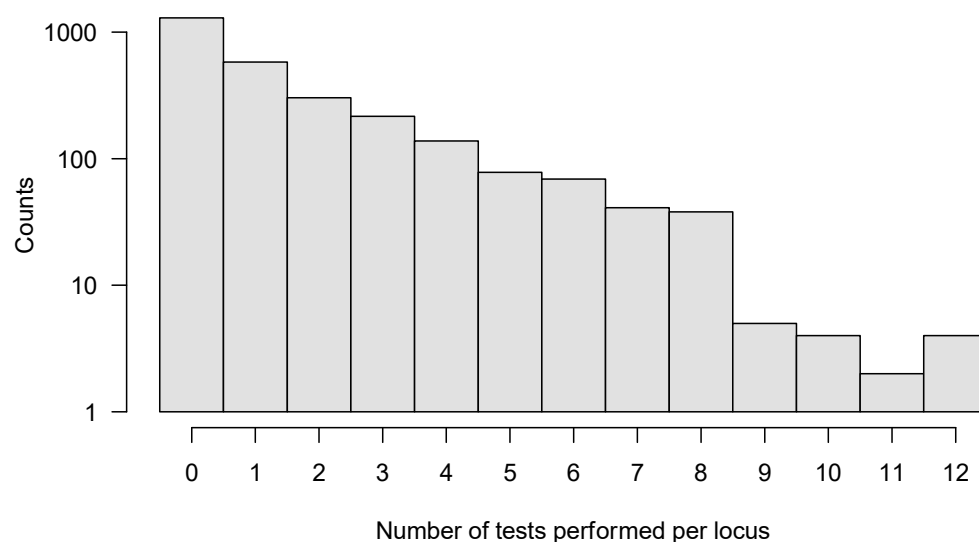

Figure 4. Histogram of number of tests per locus that could be performed for the whole HIV-1 genome. In total, we analyzed 2773 individual amino acid loci, with a maximum of 12 tests performed for any locus. In order to improve sensitivity, we only analyzed those loci further where six or more tests could be performed out of all 30 pairs.

After aggregating *p*-values by pooling across patients for each locus, we required at least six pairs out of 30 to have at least two amino acids in the transmitter and one in the recipient in order to be able to perform the test, resulting in 163 potential sites across

protein-coding regions of the HIV-1 genome. After pooling the data in this fashion, we identified two amino acid loci, namely positions 3 and 18, of Vpu that are statistically significant before multiple testing correction ($p = 0.042$ and 0.04, respectively) but not after Benjamini–Hochberg FDR correction (Figure 5, Supplementary Materials, Section S6). At amino acid position 3, we found serine-to-alanine and proline-to-serine transitions being potentially selected for during transmission. At residue 18, we found the leucine-to-isoleucine transition to be selected for. None of the immunogenic loops in gp120 showed any signatures of selection when pooled across all pairs of patients (Figure 6).
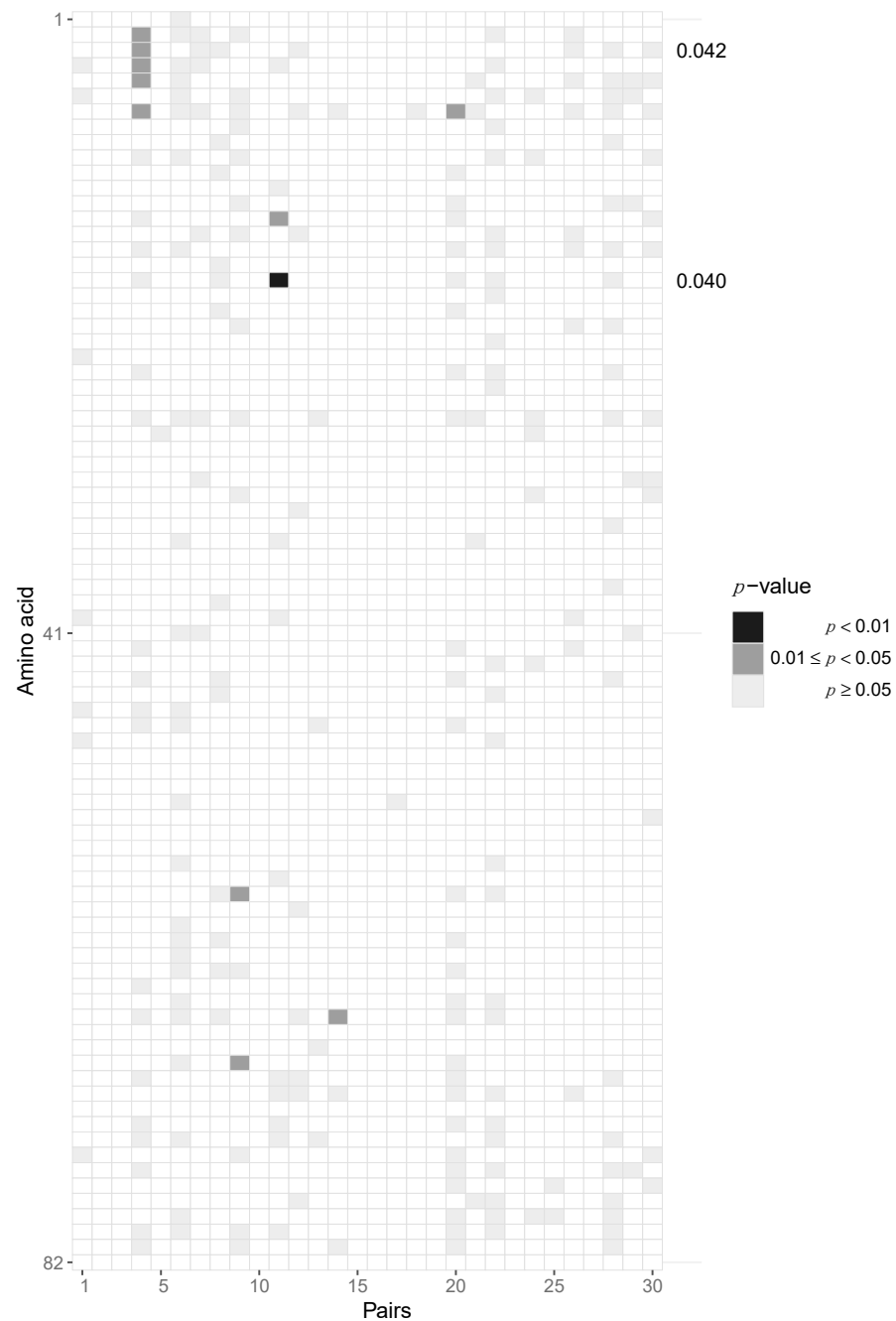
Figure 5. Potential selection of Vpu variants during HIV-1 transmission. Heatmap of *p*-values of the Vpu protein. Each column represents the test outcomes for a transmitter–recipient pair and every row represents one amino acid locus out of 82 across all 30 recipient and transmitters. Significant *p*-values after pooling across pairs without Benjamini–Hochberg multiple testing correction are shown on the right axis. Amino acid loci 3 and 18 were found to be significant.
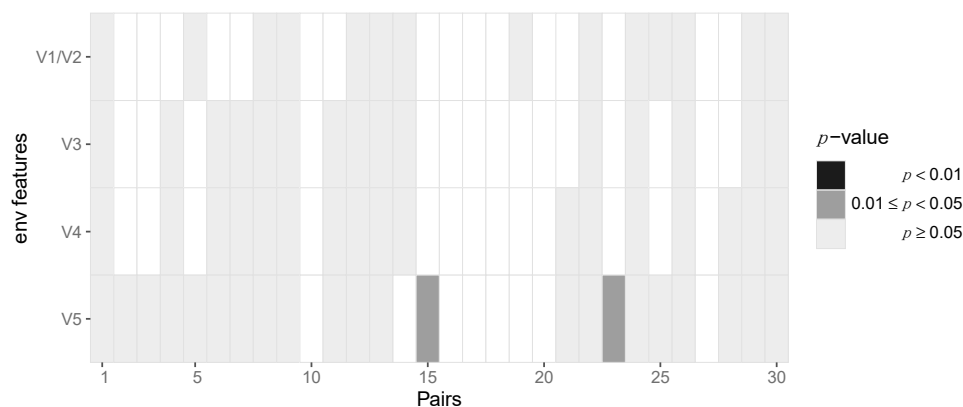
Figure 6. No evidence for selection of Env gp120 V region variants during HIV-1 transmission. Heatmap of *p*-values of the gp120 immunogenic loci. Each column represents the test outcomes for a transmitter–recipient pair. The *p*-values across the 30 transmitter–recipient pairs for the four immunogenic loci (V1/V2, V3, V4 and V5) are shown on the x and y axis, respectively.

4. Discussion

We have developed SeTesT, a customized deep sequencing-based statistical test for deviations from a neutral transmission model taking into account the population bottleneck of viral transmission and potential divergence of transmitted multi-locus genotypes. We considered the entire HIV-1 genome in 30 curated transmitter–recipient pairs and tested those loci for which at least six tests could be performed across pairs. We found two loci under selection before multiple testing correction.

With our whole-genome analysis, we found parts of Vpu to be possible candidates for being under selection. Given that at amino acid position 3 no single amino acid was selected for, the locus may be under positive selection. This is also corroborated by the high inherent heterogeneity of Vpu, which is more in line with positive selection than purifying negative selection. Vpu has been shown to affect virion release in transmitter–recipient pairs [40,41]. Both loci are in the transmembrane part of Vpu, which are determinants of CD4 downregulation [42].

SeTesT models the transmission event by sampling from the transmitter's viral population a very small number of virions that are potentially passed on to the recipient. We model this stochastic process to derive a test statistic over the observed data. We assume that the estimates of the viral variants derived from the read data are unbiased and drawn from a multinomial distribution. In case the NGS population frequency estimators are known or believed to be biased with respect to the true frequencies, the primerID protocol allows for certain error corrections [43,44].

We have shown that SeTesT produces meaningful *p*-values and in particular controls the type I error rate. It is, however, very conservative when assuming that on average 1.78 viruses are transmitted per event for successful transmission events. The power of the test is effectively determined by this bottleneck, with all other parameters having much smaller effects on sensitivity. Ignoring this bottleneck would lead to a massively inflated false positive rate, as can be seen, for instance, when just using Fisher's exact test. This effect seems intuitive when regarding the number of transmitted viruses as independent replicates of an experiment. Having just one observation, i.e., one transmitted virus, will not be enough evidence to falsify the null hypothesis of neutral transmission and stochastic sampling alone. A proper experimental way to increase power would be to repeat the transmission with the same transmitter population, conditioned on knowing the true transmission network. Such a repeated transmission under controlled conditions is, however, for ethical, experimental, biological and practical reasons impossible in humans. An in vitro model, where the different genetic bottlenecks are separated, could be a viable validation strategy of candidate loci under selection. With a two compartment in vitro model, where the two compartments are separated by an artificial mucosa [45], characteristics of the virus

may be tested that make it more permeable to the mucosal layer. In a second in vitro system one might utilize a well-characterized culture of cells of the human innate immune system, in which the transmissibility with respect to the immune system could be tested.

The number of transmitted virions in our model corresponds to the known infection risk per intercourse for men who have sex with men [46]. With approximately 1.78 viruses transmitted during such a transmission event, under a Poisson model, the estimated probability of transmission is approximately 40%. This probability is much larger than any known figure from epidemiological studies [47], suggesting that the transmission bottleneck may be even smaller than 1.78 viruses in practice when not conditioning on successful transmission. In this sense, our bottleneck parameter can be considered to be on the liberal side. We further make the implicit assumption that the population frequencies at the sampling time are representative of the frequencies during transmission. To what extent this assumption is violated is hard to judge, given that we have no longitudinal data. To our knowledge, there are no comprehensive studies that have analyzed the stability of variant frequencies from time of infection to when blood is sampled.

We have introduced an amino acid substitution model for cases where whole open reading frames are to be tested for selection and recipient sequences are not a subset of transmitter sequences. In general, substitution models assume that all sites in a sequence are not under any selective pressure. While we cannot ascertain neutrality for all sites, even in cases of non-neutral transmission we use the substitution model only to match recipient sequences to transmitter sequences and not for any phylogenetic inference or the test itself. Furthermore, the sampling times are two orders of magnitude lower than the reciprocal of the substitution rate, such that mismatching transmitters to recipients is very unlikely.

It should be noted here that we have analyzed our test in a fashion similar to genome-wide association studies (GWAS). Importantly, pooling across patients is statistically valid, as different pairs of transmitter and recipients are independent of each other given the transmission network. Such pooling methods are common practice in the field of GWAS [37]. Because most likely only one virus is transmitted in the majority of transmissions, linkage disequilibrium is maximal, as initially the newly established infection cannot reduce linkage equilibrium by recombination for a lack of other haplotypes. With this strong linkage disequilibrium, *p*-values along the genome are not statistically independent, making our test on a whole-genome scale substantially more conservative than if sites were independent. Our analysis is in contrast to other analyses on the level of cohort features, such as those performed in [16]. Our goal is to detect selection on individual amino acids and the different immunogenic loops of gp120, whereas the authors in [16] have focused on identifying summary statistics that are supported on a population level, without determining individual loci under selection. Furthermore, their study has excluded the env locus, whereas we include all of gp120 and all of gp41 in our analysis.

While the final result involves 163 sites on which multiple testing correction was applied, the effective number of tests is much lower, that is, the hypothetical number of independent tests if all correlation due to linkage disequilibrium between tests is taken into account. Furthermore, due to the conservative nature of our test, most *p*-values are larger than statistically necessary due to strong discrete effects in the test statistic [48]. For these reasons, the two sites in Vpu may still be regarded as interesting candidates for selection despite not reaching statistical significance after correction for multiple testing. Further investigations are required to falsify or confirm their involvement in non-neutral transmission.

Another study has also concluded that minor variants are preferentially transmitted, corroborating that selection acts during transmission, without being able to pinpoint any one locus that appears to be under selection [49]. The authors found that the transmitted variant is not the majority variant of the transmitter's genital tract. A limitation of our method is that we cannot detect complex higher-order patterns of selection, where the selective advantage derives not from the effect of a single allele, but the interaction of multiple alleles that are not observable within the same sequencing reads. In this scenario, the space of the test statistic space grows exponentially and we would lose all statistical power.

Another limitation of our method is that we assume that we have uncovered the true direct transmitter–recipient pairs. Phylogenetic methods on the other hand do not require explicit pairs of transmitter–recipient relationships because they estimate the transmission network implicitly. While being able to estimate evolutionary divergence and the potential indirect transmission network, separating the evolutionary pressures from transmission and intra-host evolution becomes more involved if only the former is to be investigated. Furthermore, early immune escape and reversion are potential confounding factors that are impossible to control for in our setting, because this would require longitudinal data to tease apart different evolutionary forces [50,51].

The number of patients in our study is limited but still considerable given the difficulty of reliably matching transmitters with recipients, not least due to involved legal and ethical intricacies. We have validated the pairs using multiple sources of anamnestic records, and have found that a subset of pairs in our study are in close agreement with pairs from our previous study [31]. We believe that increasing sensitivity in detecting transmission selection can only be achieved through a concerted effort of the HIV community to pool cohorts. Only when the number of properly paired transmitter–recipient samples is drastically increased can we hope to improve sensitivity of detecting selection. Our statistical analysis shows, for example, that if the transmitting populations are assumed to have the same population composition, then for 1000 aggregated patients, selection can be detected with sensitivity 1.0.

5. Conclusions

We have developed SeTesT, a novel statistical test to determine whether selection acts on certain parts of the HIV-1 genome during sexual transmission events. Applying the test to 30 transmitter–recipient pairs, we have found new potential candidates for selection in Vpu. Our probabilistic model also shows that statistical inference in general is very challenging and we quantify its limitations to help design future studies. Our results suggest that a higher number of transmitter–recipient pairs is required to improve sensitivity of detecting selection.

SeTesT can be used in other settings than sexual transmission events and for different viruses. For instance, it could also be used for estimating selection on the genome of the Hepatitis C virus, given that the genetic bottleneck for an intravenous transmission mode could be estimated. Thus, SeTesT provides a tool based on a statistically sound approach for exploring the role of selection during viral transmission events.

Author Contributions: D.S., H.F.G., N.B. and K.J.M. conceived and designed this study. B.J., D.L.B., C.S.O., C.D.S., H.K., C.G., J.B., S.Y., V.A. and T.K. were responsible for data acquisition and sequencing. D.S., B.J. and K.J.M. analyzed the data. D.S. developed SeTesT. D.S. visualized the data. All authors interpreted the data. D.S. and K.J.M. wrote this paper. All authors reviewed and edited this paper. All authors have read and agreed to the published version of this manuscript.

References

1. UNAIDS. Global HIV & AIDS Statistics—2020 Fact Sheets. Available online: https://www.unaids.org/en/resources/fact-sheet (accessed on 9 December 2021).
2. Shattock, R.J.; Moore, J.P. Inhibiting sexual transmission of HIV-1 infection. *Nat. Rev. Microbiol.* **2003**, *1*, 25–34. [CrossRef]
3. Joseph, S.B.; Swanstrom, R.; Kashuba, A.D.; Cohen, M.S. Bottlenecks in HIV-1 transmission: Insights from the study of founder viruses. *Nat. Rev. Microbiol.* **2015**, *13*, 414–425. [CrossRef]
4. Shaw, G.M.; Hunter, E. HIV Transmission. *Cold Spring Harb. Perspect. Med.* **2012**, *2*, a006965. [CrossRef]
5. Sagar, M. HIV-1 transmission biology: Selection and characteristics of infecting viruses. *J. Infect. Dis.* **2010**, *202* (Suppl. S2), S289–S296. [CrossRef]
6. Quinones-Mateu, M.E.; Arts, E.J. Fitness of drug resistant HIV-1: Methodology and clinical implications. *Drug Resist. Updat.* **2002**, *5*, 224–233. [CrossRef]
7. Elena, S.F.; Agudelo-Romero, P.; Lalic, J. The evolution of viruses in multi-host fitness landscapes. *Open Virol. J.* **2009**, *3*, 1–6. [CrossRef]
8. Seifert, D.; Di Giallonardo, F.; Metzner, K.J.; Gunthard, H.F.; Beerenwinkel, N. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. *Genetics* **2015**, *199*, 191–203. [CrossRef]
9. Bonhoeffer, S.; Barbour, A.D.; De Boer, R.J. Procedures for reliable estimation of viral fitness from time-series data. *Proc. Biol. Sci.* **2002**, *269*, 1887–1893. [CrossRef]
10. Mouquet, H.; Nussenzweig, M.C. HIV: Roadmaps to a vaccine. *Nature* **2013**, *496*, 441–442. [CrossRef]
11. Ahmad, N.; Baroudy, B.M.; Baker, R.C.; Chappey, C. Genetic analysis of human immunodeficiency virus type 1 envelope V3 region isolates from mothers and infants after perinatal transmission. *J. Virol.* **1995**, *69*, 1001–1012. [CrossRef]

12. Zhu, T.; Mo, H.; Wang, N.; Nam, D.S.; Cao, Y.; Koup, R.A.; Ho, D.D. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* **1993**, *261*, 1179–1181. [CrossRef]

13. Chohan, B.; Lang, D.; Sagar, M.; Korber, B.; Lavreys, L.; Richardson, B.; Overbaugh, J. Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1–V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels. *J. Virol.* **2005**, *79*, 6528–6531. [CrossRef]

14. Liu, Y.; Curlin, M.E.; Diem, K.; Zhao, H.; Ghosh, A.K.; Zhu, H.; Woodward, A.S.; Maenza, J.; Stevens, C.E.; Stekler, J.; et al. Env length and N-linked glycosylation following transmission of human immunodeficiency virus Type 1 subtype B viruses. *Virology* **2008**, *374*, 229–233. [CrossRef]

15. Choi, J.Y.; Pond, S.L.K.; Anderson, C.M.; Richman, D.D.; Smith, D.M. Molecular Features of the V1-V4 Coding Region of Sexually Transmitted Human Immunodeficiency Virus Type 1. *J. Infect. Dis.* **2017**, *215*, 1506–1513. [CrossRef]

16. Carlson, J.M.; Schaefer, M.; Monaco, D.C.; Batorsky, R.; Claiborne, D.T.; Prince, J.; Deymier, M.J.; Ende, Z.S.; Klatt, N.R.; DeZiel, C.E.; et al. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* **2014**, *345*, 1254031. [CrossRef]

17. Keele, B.F.; Giorgi, E.E.; Salazar-Gonzalez, J.F.; Decker, J.M.; Pham, K.T.; Salazar, M.G.; Sun, C.; Grayson, T.; Wang, S.; Li, H.; et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 7552–7557. [CrossRef]

18. Derdeyn, C.A.; Decker, J.M.; Bibollet-Ruche, F.; Mokili, J.L.; Muldoon, M.; Denham, S.A.; Heil, M.L.; Kasolo, F.; Musonda, R.; Hahn, B.H.; et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science* **2004**, *303*, 2019–2022. [CrossRef]

19. Haaland, R.E.; Hawkins, P.A.; Salazar-Gonzalez, J.; Johnson, A.; Tichacek, A.; Karita, E.; Manigart, O.; Mulenga, J.; Keele, B.F.; Shaw, G.M.; et al. Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS Pathog.* **2009**, *5*, e1000274. [CrossRef]

20. Rieder, P.; Joos, B.; Scherrer, A.U.; Kuster, H.; Braun, D.; Grube, C.; Niederost, B.; Leemann, C.; Gianella, S.; Metzner, K.J.; et al. Characterization of Human Immunodeficiency Virus Type 1 (HIV-1) Diversity and Tropism in 145 Patients With Primary HIV-1 Infection. *Clin. Infect. Dis.* **2011**, *53*, 1271–1279. [CrossRef]

21. Magnani, R.; Sabin, K.; Saidel, T.; Heckathorn, D. Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS* **2005**, *19* (Suppl. S2), S67–S72.

22. Posada-Cespedes, S.; Seifert, D.; Beerenwinkel, N. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus. Res.* **2017**, *239*, 17–32. [CrossRef] [PubMed]

23. Posada-Cespedes, S.; Seifert, D.; Topolsky, I.; Jablonski, K.P.; Metzner, K.J.; Beerenwinkel, N. V-pipe: A computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics* **2021**, *37*, 1673–1680. [CrossRef] [PubMed]

24. Bergstrom, C.T.; McElhany, P.; Real, L.A. Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 5095–5100. [CrossRef] [PubMed]

25. Sobel Leonard, A.; Weissman, D.B.; Greenbaum, B.; Ghedin, E.; Koelle, K. Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *J. Virol.* **2017**, *91*, e00171-17. [CrossRef]

26. Lumby, C.K.; Nene, N.R.; Illingworth, C.J.R. A novel framework for inferring parameters of transmission from viral sequence data. *PLoS Genet.* **2018**, *14*, e1007718. [CrossRef]

27. Morris, D.H.; Petrova, V.N.; Rossine, F.W.; Parker, E.; Grenfell, B.T.; Neher, R.A.; Levin, S.A.; Russell, C.A. Asynchrony between virus diversity and antibody selection limits influenza virus evolution. *bioRxiv* **2020**, *9*, e62105. [CrossRef]

28. Schoeni-Affolter, F.; Ledergerber, B.; Rickenbach, M.; Rudin, C.; Gunthard, H.F.; Telenti, A.; Furrer, H.; Yerly, S.; Francioli, P. Cohort profile: The Swiss HIV Cohort study. *Int. J. Epidemiol.* **2010**, *39*, 1179–1189. [CrossRef]

29. Yang, W.L.; Kouyos, R.; Scherrer, A.U.; Boni, J.; Shah, C.; Yerly, S.; Klimkait, T.; Aubert, V.; Furrer, H.; Battegay, M.; et al. Assessing the Paradox Between Transmitted and Acquired HIV Type 1 Drug Resistance Mutations in the Swiss HIV Cohort Study From 1998 to 2012. *J. Infect. Dis.* **2015**, *212*, 28–38. [CrossRef]

30. Rieder, P.; Joos, B.; von Wyl, V.; Kuster, H.; Grube, C.; Leemann, C.; Boni, J.; Yerly, S.; Klimkait, T.; Burgisser, P.; et al. HIV-1 transmission after cessation of early antiretroviral therapy among men having sex with men. *AIDS* **2010**, *24*, 1177–1183. [CrossRef]

31. Oberle, C.S.; Joos, B.; Rusert, P.; Campbell, N.K.; Beauparlant, D.; Kuster, H.; Weber, J.; Schenkel, C.D.; Scherrer, A.U.; Magnus, C.; et al. Tracing HIV-1 transmission: Envelope traits of HIV-1 transmitter and recipient pairs. *Retrovirology* **2016**, *13*. [CrossRef]

32. Giallonardo, F.D.; Töpfer, A.; Rey, M.; Prabhakaran, S.; Duport, Y.; Leemann, C.; Schmutz, S.; Campbell, N.K.; Joos, B.; Lecca, M.R.; et al. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Res.* **2014**, *42*, e115. [CrossRef] [PubMed]

33. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **2011**, *27*, 863–864. [CrossRef] [PubMed]

34. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]

35. Schirmer, M.; Ijaz, U.Z.; D'Amore, R.; Hall, N.; Sloan, W.T.; Quince, C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **2015**, *43*, e37. [CrossRef] [PubMed]

36. Neyman, J.; Pearson, E.S. Contributions to the theory of testing statistical hypotheses. *Stat. Res. Mem.* **1936**, *1*, 1–37.

37. Derkach, A.; Lawless, J.F.; Sun, L. Robust and Powerful Tests for Rare Variants Using Fisher's Method to Combine Evidence of Association From Two or More Complementary Tests. *Genet. Epidemiol.* **2013**, *37*, 110–121. [CrossRef]

38. Guglietta, S.; Pantaleo, G.; Graziosi, C. Long sequence duplications, repeats, and palindromes in HIV-1 gp120: Length variation in V4 as the product of misalignment mechanism. *Virology* **2010**, *399*, 167–175. [CrossRef]

39. Prakash, S.S.; Kalra, R.; Lodha, R.; Kabra, S.K.; Luthra, K. Diversity of HIV type 1 envelope (V3-V5) sequence in HIV type 1-infected Indian children. *AIDS Res. Hum. Retrovir.* **2012**, *28*, 505–509. [CrossRef]

40. Jafari, M.; Guatelli, J.; Lewinski, M.K. Activities of Transmitted/Founder and Chronic Clade B HIV-1 Vpu and a C-Terminal Polymorphism Specifically Affecting Virion Release. *J. Virol.* **2014**, *88*, 5062–5078. [CrossRef]

41. Gonzalez, M.E. Vpu Protein: The Viroporin Encoded by HIV-1. *Viruses* **2015**, *7*, 4352–4368. [CrossRef]

42. Magadan, J.G.; Bonifacino, J.S. Transmembrane domain determinants of CD4 Downregulation by HIV-1 Vpu. *J. Virol.* **2012**, *86*, 757–772. [CrossRef] [PubMed]

43. Zhou, S.; Jones, C.; Mieczkowski, P.; Swanstrom, R. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next Generation Sequencing of HIV-1 Genomic RNA Populations. *J. Virol.* **2015**, *89*, 8540–8555. [CrossRef] [PubMed]

44. Seifert, D.; Di Giallonardo, F.; Topfer, A.; Singer, J.; Schmutz, S.; Gunthard, H.F.; Beerenwinkel, N.; Metzner, K.J. A Comprehensive Analysis of Primer IDs to Study Heterogeneous HIV-1 Populations. *J. Mol. Biol.* **2016**, *428*, 238–250. [CrossRef]

45. Cavarelli, M.; Foglieni, C.; Rescigno, M.; Scarlatti, G. R5 HIV-1 envelope attracts dendritic cells to cross the human intestinal epithelium and sample luminal virions via engagement of the CCR5. *EMBO Mol. Med.* **2013**, *5*, 776–794. [CrossRef] [PubMed]

46. Patel, P.; Borkowf, C.B.; Brooks, J.T.; Lasry, A.; Lansky, A.; Mermin, J. Estimating per-act HIV transmission risk: A systematic review. *AIDS* **2014**, *28*, 1509–1519. [CrossRef] [PubMed]

47. Boily, M.C.; Baggaley, R.F.; Wang, L.; Masse, B.; White, R.G.; Hayes, R.J.; Alary, M. Heterosexual risk of HIV-1 infection per sexual act: Systematic review and meta-analysis of observational studies. *Lancet Infect. Dis.* **2009**, *9*, 118–129. [CrossRef]

48. Agresti, A.; Gottard, A. Reducing conservatism of exact small-sample methods of inference for discrete data. In *Compstat 2006, Proceedings in Computational Statistics, Rome, Italy, 23 March 2006*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 245–260.

49. Boeras, D.I.; Hraber, P.T.; Hurlston, M.; Evans-Strickfaden, T.; Bhattacharya, T.; Giorgi, E.E.; Mulenga, J.; Karita, E.; Korber, B.T.; Allen, S.; et al. Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, E1156–E1163. [CrossRef]

50. Borrow, P.; Lewicki, H.; Wei, X.; Horwitz, M.S.; Peffer, N.; Meyers, H.; Nelson, J.A.; Gairin, J.E.; Hahn, B.H.; Oldstone, M.B.; et al. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat. Med.* **1997**, *3*, 205–211. [CrossRef]

51. Price, D.A.; Goulder, P.J.; Klenerman, P.; Sewell, A.K.; Easterbrook, P.J.; Troop, M.; Bangham, C.R.; Phillips, R.E. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 1890–1895. [CrossRef]