Transmitted HIV-1 is more pathogenic in heterosexual individuals than homosexual men

Ananthu James[a] and Narendra M. Dixit[a,b,*]

[a]Department of Chemical Engineering, Indian Institute of Science, Bengaluru, India
[b]Centre for Biosystems Science and Engineering, Indian Institute of Science, Bengaluru, India

[*]Corresponding author; E-mail: narendra@iisc.ac.in

Manuscript details:
Abstract: 150 words
Text: ∼1800 words
Figures: 3
Table: 1
References: 27
Methods: ∼1000 words

Supplementary Materials:
Text notes: 2
Figures: 1
Tables: 7
References: 41

Abstract

Transmission bottlenecks introduce selection pressures on HIV-1 that vary substantially with the mode of transmission. Recent studies on small cohorts have suggested that stronger selection pressures lead to fitter transmitted/founder (T/F) strains. Manifestations of this selection bias at the population level have remained elusive. Here, we analysed early CD4 cell count measurements reported from $\sim$340,000 infected heterosexual individuals (HSX) and men-who-have-sex-with-men (MSM), across geographies, ethnicities and calendar years and found them to be consistently lower in HSX than MSM ($P<0.05$). The corresponding average reduction in CD4 counts relative to healthy adults was 86.5% in HSX and 67.8% in MSM ($P<10^{-4}$). This difference could not be attributed to differences in age, HIV-1 subtype, viral load, gender, ethnicity, time of transmission, or diagnosis delay across the groups. We concluded that the different selection pressures arising from the different predominant transmission modes have resulted in more pathogenic T/F strains in HSX than MSM.

Introduction

The bottlenecks in HIV-1 transmission result in a 'selection bias' favoring fitter transmitted/founder (T/F) viruses over less fit ones[1,2]. Several recent studies have presented evidence of genetic, phenotypic, and clinical manifestations of the selection bias in small cohorts[1,3–6]. From 137 heterosexual (HSX) donor-recipient pairs, T/F viruses were found to carry higher than average frequencies of amino acids associated with high *in vivo* fitness[1]. Similarly, from 127 discordant couples, lower viral replication capacity (vRC), indicative of lower viral fitness, early in infection was associated with slower decline of CD4 T cell counts[4,6]. The selection bias varies with the mode of transmission[3]. The stronger the bottlenecks, the fitter the corresponding T/F viruses are likely to be[1,2]. Anal intercourse is over 10-fold more permissive on average than penile-vaginal intercourse[7]. Analysis of T/F genomes from 131 subjects revealed that the T/F genomes were under greater positive selection in heterosexual individuals (HSX), in whom the penile-vaginal mode predominates[8], than homosexual men, or men-who-have-sex-with-men (MSM), who transmit predominantly through anal intercourse[3]. Among HSX, men had T/F viruses with higher predicted fitness *in vivo* than women[1], consistent with the asymmetry of the bottlenecks between insertive and receptive penile-vaginal intercourse[7].

An important question that follows is whether the differential selection bias across modes of transmission is manifested at the wider population level. Such differential bias could contribute to variations in disease progression and treatment outcomes and underlie the diverse trajectories of the HIV-1 pandemic across infected groups in which different modes of transmission predominate.

Results and Discussion

To answer this question, we decided to compare early CD4 T cell count measurements between HSX and MSM. Immediately following infection, CD4 T cell counts fall steeply, recover partially, and then settle within a few weeks/months to a value smaller than in the pre-infection state[9] (Fig. 1(a)). Subsequent changes in the CD4 counts occur slowly, over many months to years. Thus, CD4 count measurements made early in infection tend to be close to the value to which the counts settle after the initial dynamics. These early CD4 counts are expected to be minimally affected by host-specific adaptive mutations[1] and, therefore, representative of the fitness of the T/F strain in the recipient. The fitter the strain, the lower would be the CD4 count. The CD4 count is also a more robust marker of disease state than other commonly used markers such as set-point viral load (SPVL). High vRC of the

T/F viruses was associated with low CD4 counts at 3 months post-infection (which roughly coincides with seroconversion) and rapid CD4 count decline for ∼5 years, independently of SPVL[4,6].

HSX and MSM are the two major groups driving the global HIV-1 epidemic[9]. They use predominant modes of transmission with a substantial difference in the selection bias[7]. Importantly, they display little inter-mixing in most geographical regions. We inferred the latter from the distinct prevalence of HIV-1 subtypes in the two groups, which we found across geographical regions and calendar years (Fig. 2; Text S1; Tables S1 and S2). Together, these characteristics allow for the difference in the selection bias to be sustained long-term, potentially amplified, and manifested in sample sizes large enough for detection with statistical significance. We thus hypothesized that the stronger selection bias associated with penile-vaginal transmission than anal transmission would result in lower early CD4 counts in HSX than in MSM.

To test this hypothesis, we collated available data of CD4 count measurements either at seroconversion or at diagnosis from all large studies, which amounted to a total of ∼340,000 patients across four geographical regions followed over a total period of nearly four decades, and examined the differences between HSX and MSM (Methods; Table 1). We found that HSX consistently had lower CD4 counts than MSM (Fig. 1(b); Tables 1 and Tables S3-S5). For instance, measurements from ∼120,000 patients across 21 countries in the European Union and European Economic Area (EU/EEA) indicated, following population-weighted averaging of yearly data during 2010-2018, that the mean CD4 count in MSM at diagnosis was ∼440 cells/μL, whereas it was substantially lower, ∼300 cells/μL, in HSX ($P<10^{-4}$)[10]. The numbers were similar in the preceding 5 year period (2002-2007) reported by a smaller study involving a few thousand patients[11]. In the UK, measurements from close to 9000 patients during 1990-1998 showed that the counts at diagnosis were ∼330 cells/μL in MSM and ∼230 cells/μL in HSX ($P<10^{-3}$)[12]. In China, during 2006-2012, the mean CD4 counts at diagnosis from ∼180,000 patients were ∼370 cells/μL in MSM and ∼270 cells/μL in HSX ($P<10^{-4}$)[13]. Similarly, in the US, from over 25,000 patients during 2006-2015, the counts at diagnosis were ∼400 cells/μL in MSM and ∼300 cells/μL in HSX ($P<10^{-4}$)[14]. We also examined/estimated the counts at seroconversion where available. In the CASCADE study, involving ∼4000 patients during 1979-2000 in Europe and Australia, the mean cell counts at seroconversion were ∼620 cells/μL in MSM and ∼590 cells/μL in HSX ($P=0.027$)[15]. Further, using the reported diagnosis delays and the slopes of CD4 count decline

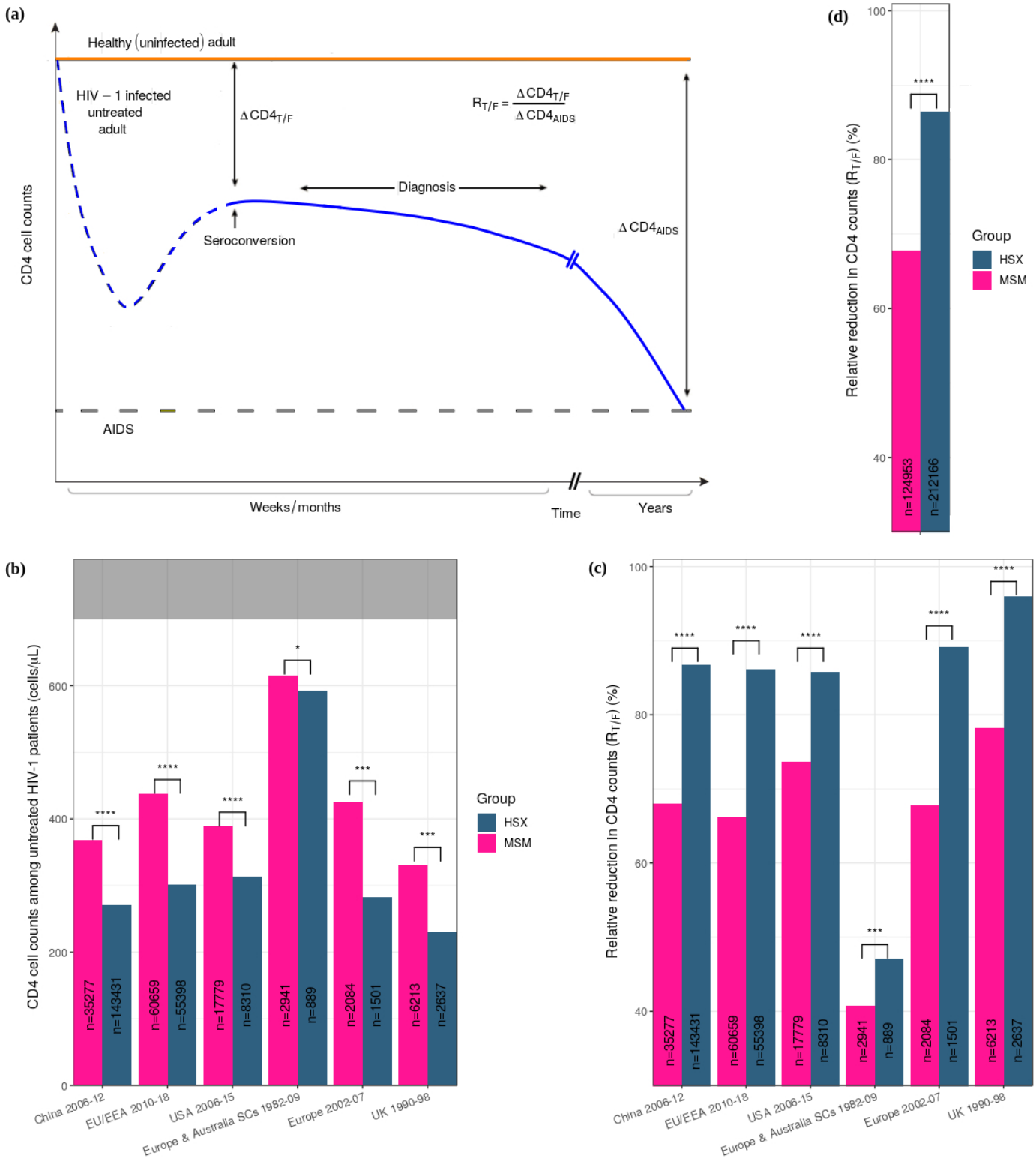

Figure 1: Early CD4 T cell counts and the associated relative reduction ($R_{T/F}$) in MSM and HSX. (a) Schematic of typical CD4 count changes post HIV-1 infection (blue), before (dashed) and after (solid) diagnosis/seroconversion. The reduction at diagnosis/seroconversion relative to uninfected individuals (orange) and that associated with AIDS (grey dashed line) yields $R_{T/F}$, the reduction attributable to the T/F virus. **(b)** Early mean CD4 cell counts and **(c)** the corresponding $R_{T/F}$ in untreated infected adult HSX and MSM from different geographical regions and calendar years (see Methods, Tables 1 and S3-S5 for details). The grey region indicates counts in uninfected, healthy individuals. **(d)** Population-weighted average of $R_{T/F}$ across all the datasets in (c). The sample sizes (n) are indicated. SCs indicate seroconverters. ****, ***, ** and * indicate $P < 10^{-4}$, $P < 10^{-3}$, $P < 10^{-2}$ and $P < 0.05$, respectively.

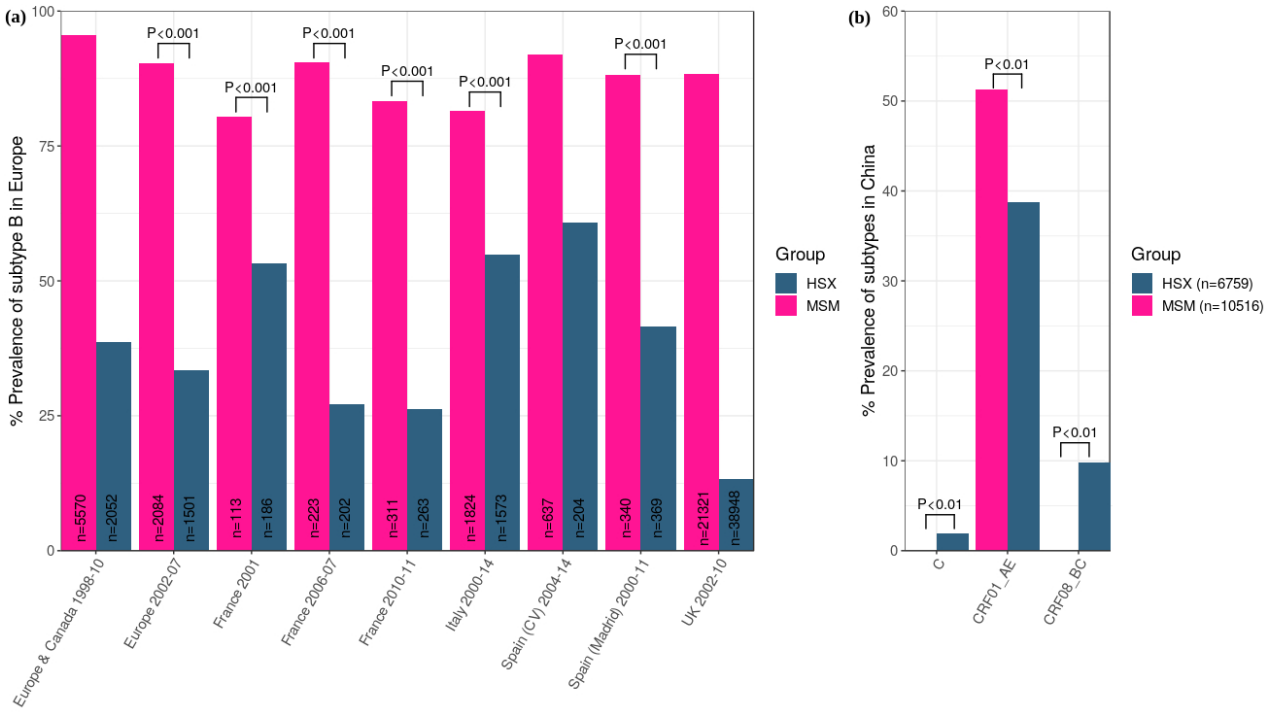

Figure 2: Subtype prevalence of HIV-1 in MSM and HSX populations. Prevalence of **(a)** subtype B in different regions in Europe and Canada and **(b)** all the subtypes in China. The sample sizes (n) along with the time periods of the surveys are indicated. P values are listed where available in the original sources. Sources of the data and additional details are in Tables S1 and S2. The different prevalence indicates little mixing between MSM and HSX in the populations studied.

in the US population above[14], we estimated that the cell counts at seroconversion, for the age group 13-29 years, were $\sim$550 cells/μL in MSM and $\sim$480 cells/μL in HSX ($P<10^{-4}$) (Methods). Remarkably, we did not find any large study (sample size $\gtrsim 1000$) that reported higher early CD4 cell counts in HSX than MSM.

While the evidence from absolute CD4 count comparisons was thus overwhelming, differences in CD4 counts in healthy (uninfected) individuals across gender, ethnicity and geographical regions could render absolute CD4 counts only an approximate measure of the fitness of the T/F strains. Two individuals may have similar early CD4 counts but may still have been infected by T/F strains of different fitness if their pre-infection CD4 counts were different, with the individual with the higher pre-infection count infected by the fitter T/F strain. To overcome this limitation, we constructed a metric to quantify the relative reduction in the CD4 cell count, R, corresponding to the absolute CD4 count T as $R = \dfrac{T_{healthy} - T}{T_{healthy} - T_{AIDS}} \times 100 = \dfrac{\Delta CD4}{\Delta CD4_{AIDS}} \times 100$, where $T_{healthy}$ was the count pre-infection, and $T_{AIDS} = 200$ cells/μL the count defining AIDS. Thus, R was 0% when $T = T_{healthy}$ and 100% when $T = T_{AIDS}$ and decreased linearly with T between these extremes. Choosing $T_{healthy}$ specific to the

Table 1. Early CD4 cell counts in infected adults at diagnosis or seroconversion, estimated from data collated from several studies (see text and Tables S3-S5). The corresponding relative reduction in CD4 count, $R_{T/F}$, calculated as described in the text, are also listed. P values for comparisons of the CD4 counts (and $R_{T/F}$) between MSM and HSX indicate a significantly higher CD4 counts and lower $R_{T/F}$ in MSM throughout. The last row represents the population-weighted average of all the datasets.

Sample & duration	Risk group	CD4 counts ($cells/\mu L$)		Sample size (n)	$R_{T/F}$ (%)	
		Mean (SD)	P value		Mean (SD)	P value
EU/EEA 2010 – 18	MSM*	437 (397)	$0.00^{\dagger\dagger}$	60,659	66.2 (58.7)	$0.00^{\dagger\dagger}$
	HSX	302 (464)[d]		55,398	86.2** (63.0)	
	HSX men*	270 (464)[d]	2.0×10^{-81}	27,822	90.0 (66.3)	1.7×10^{-51}
	HSX women	335 (325)		27,576	82.9 (42.0)	
USA 2006 – 15	MSM (6 m.o. diagnosis)	390 (243)	3.0×10^{-97}	17,779	73.7 (34.4)	5.2×10^{-143}
	HSX (6 m.o. diagnosis)	314 (284)		8,310	85.8 (35.9)	
	MSM (SCs, age: 13-29)	553 (315)	1.2×10^{-18}	6,328	51.0 (45.5)	1.8×10^{-37}
	HSX (SCs, age: 13-29)	483 (377)		2,958	64.7 (49.0)	
China 2006 – 12	MSM	368 (222)	$0.00^{\dagger\dagger}$	35,277	68.0 (45.1)	$0.00^{\dagger\dagger}$
	HSX	270 (260)		143,431	86.7 (50.0)	
Europe 2002 – 07	MSM	426 (242)	[c] $< 10^{-3}$	2,084	67.8 (37.5)	2.5×10^{-63}
	HSX	283 (258)		1,501	88.8 (35.3)	
UK 1990 – 98	MSM	331 (242)[d]	[c] $< 10^{-3}$	6,213	78.2 (42.0)	3.8×10^{-92}
	HSX	230 (258)[d]		2,637	96.0 (34.7)	
Europe & Australia (CASCADE) SCs 1979 – 00	MSM† (age<40)	621 (323)	4.3×10^{-3}	2,570	40.0 (53.7)	0.010
	HSX men[a] (age<40)	576 (327)		428	46.4 (52.7)	
	HSX women†,[a] (age<40)	623 (229)		349	46.4 (38.7)	
	MSM‡ (age>40)	578 (221)	0.017	371	46.2 (40.1)	0.073
	HSX men[b] (age>40)	534 (134)		62	52.4 (29.1)	
	HSX women‡,[b] (age>40)	580 (113)		50	51.8 (27.1)	
	MSM (total)	616 (323)[d]	0.027	2,941	40.7 (53.7)	6.9×10^{-4}
	HSX (total)	592 (327)[d]		889	47.1 (51.7)	
Overall	MSM	–		124,953	67.8 (58.7)[d]	$0.00^{\dagger\dagger}$
	HSX	–		212,166	86.5 (63.0)[d]	

*$P = 0.00^{\dagger\dagger}$ for the comparisons of both the cell counts and $R_{T/F}$ between MSM and HSX men.

**$R_{T/F}$ in the EU/EEA region changed from 86.2% (SD = 63.0%) to 84.9% (SD = 69.5%) in HSX when we used CD4 counts in healthy individuals from Tanzania (representative of sub-Saharan Africa, 746 cells/μL (Table S6)), instead of 941 cells/μL from Italy (representative of EU/EEA), corresponding to the 35% of HSX, still substantially higher than $R_{T/F}$ of 66.2% in MSM ($P = 0.00^{\dagger\dagger}$).

††These P values were below the lower representation limit ($= 2.23 \times 10^{-308}$) of R (and Excel).

$^{\dagger}P = 0.44$ for CD4 count and $P = 3.1 \times 10^{-3}$ for $R_{T/F}$ comparisons between MSM and HSX women aged < 40 years.

$^{\ddagger}P = 0.46$ for CD4 count and $P = 0.10$ for $R_{T/F}$ comparisons between MSM and HSX women aged > 40 years.

[a] $P = 0.010$ for CD4 count and $P = 0.50$ for $R_{T/F}$ comparisons between HSX men and HSX women aged < 40 years.

[b] $P = 0.026$ for CD4 count and $P = 0.46$ for $R_{T/F}$ comparisons between HSX men and HSX women aged > 40 years.

[c] Reported in the original sources.

[d] See Methods and the footnotes in Tables S3 and S5.

respective geographies, ethnicities, and genders (Table S6), we estimated R corresponding to the early cell count measurements above, which we denoted as $R_{T/F}$, indicative of the relative reduction in CD4 count due to the T/F virus (Fig. 1(c)). The higher the $R_{T/F}$, the fitter would be the T/F strain, regardless of the pre-infection CD4 count, rendering $R_{T/F}$ a more robust marker of T/F viral fitness than the associated early absolute CD4 counts. (Note that $R_{T/F}$ is a static measure and is not indicative of the 'speed' of disease progression; cell count decline can be faster despite higher early CD4 counts in MSM than HSX[15,16].)

We found that in EU/EAA, during 2010-18, $R_{T/F}$ was 86.2% in HSX and 66.2% in MSM ($P<10^{-4}$). During 2002-07, these numbers were 88.8% and 67.8% ($P<10^{-4}$), respectively. The corresponding numbers were 96.0% and 78.2% in the UK ($P<10^{-4}$), and 86.7% and 68.0% in China ($P<10^{-4}$). In the US, the difference was smaller but still substantial, with $R_{T/F}$ of 85.8% in HSX and 73.7% in MSM ($P<10^{-4}$). At seroconversion, these numbers were 64.7% and 51.0%, respectively ($P<10^{-4}$). For the seroconverters from the CASCADE study, the trend was consistent, with $R_{T/F}$ of 47.1% in HSX and 40.7% in MSM ($P<10^{-3}$). Overall, thus, $R_{T/F}$ comparisons showed more significant differences between MSM and HSX than absolute CD4 count comparisons (Fig. 1(b) and (c)). Further, $R_{T/F}$ allowed comparison across the different datasets. Thus, while the HSX all had $R_{T/F}$ >85% at diagnosis, the MSM displayed a range from ∼65% to a little under 80%. We could also combine the datasets, including those at diagnosis and seroconversion, and estimate an overall $R_{T/F}$. Using a population-weighted average across the datasets, we estimated the overall $R_{T/F}$ to be 86.5% in HSX and 67.8% in MSM ($P<10^{-4}$) (Fig. 1(d)). This overall comparison provides strong evidence of greater cell count reduction due to, and hence greater pathogenicity of, the T/F viruses in HSX than in MSM.

To attribute the differences in $R_{T/F}$ between HSX and MSM to the differential selection bias at transmission in the two groups, we considered and ruled out all the major potential confounding factors. First, MSM are typically diagnosed at a younger age than HSX. In the two European studies, MSM were 5 (Table S5)[10] and 1.6 years[11] younger on average than HSX at diagnosis. Given the cell count decrease of ∼7 cells/μL per year of age at diagnosis[12], the CD4 counts should have been higher in MSM by only ∼35 and ∼11 cells/μL, whereas they were higher by 135 and 143 cells/μL (Fig. 1(b)), respectively, a difference that could not be explained by the age at diagnosis. Second, MSM are often predominantly infected by subtype B[17], whereas HSX are by subtypes B and C (Fig. 2; Text S1). This

subtype difference should have resulted in lower CD4 counts in MSM than HSX because of the higher virulence of subtype B[18,19], a trend opposite of what is observed. Moreover, in the US where subtype B dominates both HSX and MSM (Text S1), $R_{T/F}$ was lower among MSM (Fig. 1(c)). In agreement, an independent study found that subtype B T/F viruses had higher fitness among HSX than MSM[3]. Third, the CD4 counts could not be explained as an indirect manifestation of variations in SPVL; in the European study, CD4 counts were higher in MSM despite higher SPVL in MSM than HSX (Table S3). Fourth, healthy men had lower CD4 counts than HSX and healthy women everywhere except China (Table S6), and infected HSX men displayed higher $R_{T/F}$ than MSM (Table 1 and Fig. S1), two reasons to rule out gender as the cause of lower $R_{T/F}$ in MSM. Fifth, in Europe (EU/EEA), while MSM are predominantly Caucasian, 30-35% of infected HSX are of sub-Saharan African origin[10,11]. In China, however, where no differences in ethnicity exist between MSM and HSX, a substantial difference in $R_{T/F}$ is seen between them (Fig. 1(c)), ruling out ethnicity as a confounding factor. Further, accounting for baseline CD4 count differences across ethnicities in EU/EEA did not alter our findings (Table 1). Sixth, early onward transmission may limit donor-specific adaptations in the T/F strain and allow it to cause more severe cell count reduction in the recipient. Early transmissions, however, are more common to MSM than HSX[18,20], in keeping with the greater association of MSM with transmission clusters[17] (Fig. 3; Table S7), and should have led to higher $R_{T/F}$ in MSM than HSX, in contrast to our findings. Seventh, although MSM tend to be diagnosed earlier than HSX[14] and may thus suffer a lower loss of CD4 counts at diagnosis, the differences are seen also in CD4 counts at seroconversion[14,15], which would occur at similar times post infection in the two groups. Besides, MSM had lower cell counts in China too, where, owing to social stigma, MSM may not get diagnosed earlier than HSX[13]. The difference in $R_{T/F}$ between MSM and HSX was thus not attributable to any of the above factors. We concluded therefore that the difference originated from the variations in the fitness of the T/F strains in the two groups arising from the different selection biases at transmission.

Our findings establish the selection bias at transmission as an important underlying factor shaping HIV-1 adaptation at the population level. The differential adaptation of HIV-1 to MSM and HSX, which in most geographical regions show little inter-mixing, may have led over the years to the selection and, possibly, fixation of different adaptive mutations in the T/F viruses in the two groups. Genetic differences have been observed between T/F strains in MSM and HSX in small cohorts[3].

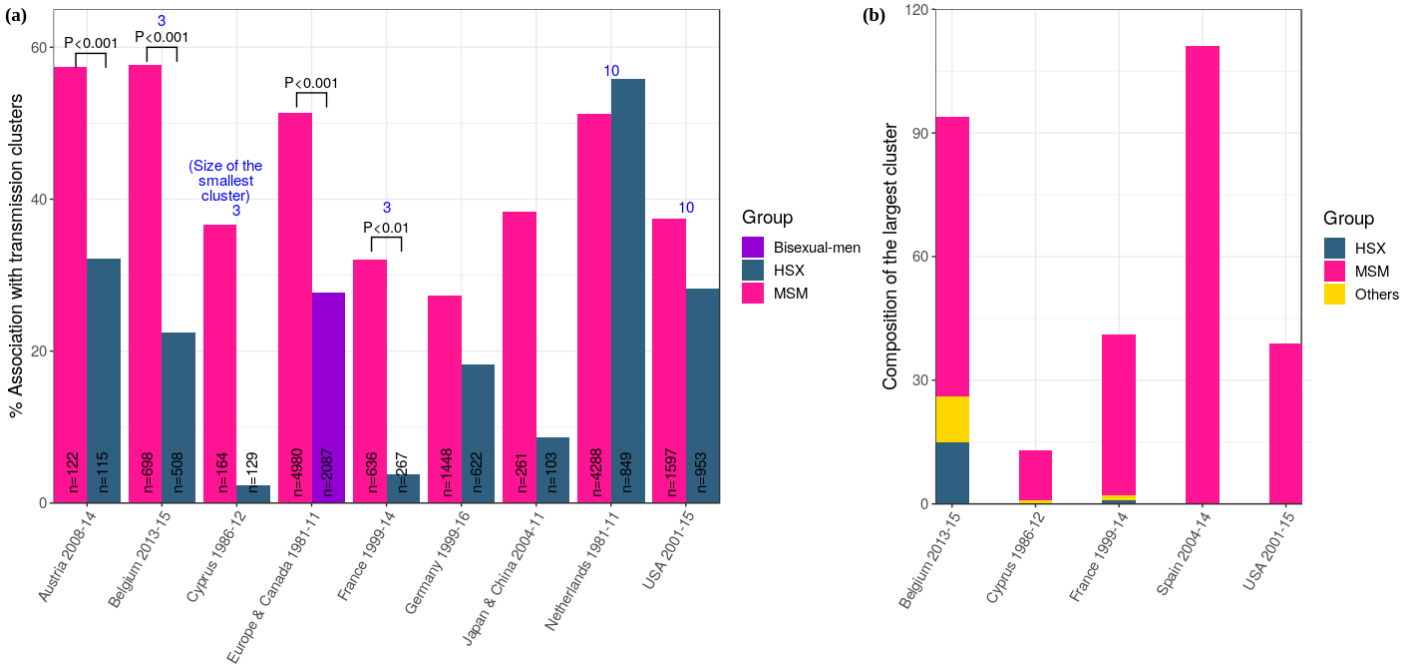

Figure 3: Association with transmission clusters. **(a)** The fraction of HSX and MSM (or bisexuals in one case) associated with transmission clusters and **(b)** the composition of the largest clusters in different geographical regions. The sample sizes (n) along with the time periods of the surveys are indicated. P values and the minimum sizes of the clusters (blue text) where available from the original sources are listed. Sources of the data and additional details are in Table S7. MSM are thus far more likely to be associated with clusters and also tend to form large clusters.

Future studies may establish them at the population level, as sequencing technologies that allow facile identification of T/F viruses emerge. The technologies may also serve to elucidate such differences between other infected groups, which are likely to be present to lower degrees than between MSM and HSX, depending on the differences in the selection bias between the groups, the exclusivity of the associated modes of transmission, and the extent of mixing between the groups. Our findings also suggest that heritable viral traits such as SPVL[21] may have evolved differently in MSM and HSX, potentially driving differential spread of the HIV-1 epidemic in the two groups. The extent of these differences may determine whether intervention strategies, including the development and use of preventive vaccines, may have to be tailored to individual infected groups.

Methods

Data of CD4 counts. To test our hypothesis that early CD4 counts in HSX would be higher than in MSM at the population level, we collated data from all large studies ($n \gtrsim 1{,}000$) that reported CD4 counts either at diagnosis or seroconversion in both these groups. The data are summarized along with our analysis in Table 1 and details are in Tables S3-S5. From reports on coun-

tries in the EU/EEA and China[10,13], we digitized the median CD4 counts using WebPlotDigitizer ($https://automeris.io/WebPlotDigitizer$). For our analysis, we averaged the data over the study duration. To obtain sample sizes, we multiplied the diagnosed cases with the reported fraction of diagnoses contributing to the annual CD4 counts in the entire EU/EEA (Table S5). The fraction was assumed to be the same across the risk groups and the set of 21 countries studied. We also assumed the proportions of men and women in HSX to remain the same during 2010-18. In the CASCADE study[15], which segregated data into age groups, we averaged over age groups. To obtain the population-weighted average CD4 counts, we assumed that the proportions of the populations in the different transmission categories were the same across age groups and that the fractions of men and women remained conserved (except in MSM and hemophiliacs) (Table S3). To calculate $R_{T/F}$, we also collated data of CD4 counts from healthy, uninfected adults in the USA, UK, Italy (which was used for the three studies involving European populations), Tanzania, and China, which are listed in Table S6. For $R_{T/F}$ calculations pertaining to the UK, CD4 counts from healthy MSM and HSX were available, which we used. We found the counts in MSM comparable to those from healthy HSX men. As a result, for other populations, we used the cell counts for healthy HSX men where counts from healthy MSM were unavailable.

Estimation of mean CD4 counts and their standard deviations. When the median, m, and interquartile range (IQR), $(q_l,\ q_u)$, of CD4 counts were available, we estimated the corresponding mean, μ, and standard deviation (SD), σ, using $\mu = \frac{m+q_u+q_l}{3}$ and $\sigma = \frac{q_u-q_l}{1.35}$, following the widely used method[22] applicable to large sample sizes, as considered here. When 95% confidence intervals (CIs), $(c_l,\ c_u)$, were available instead of IQR, we evaluated SD using another method[23] which yielded $\sigma = \frac{\sqrt{n}\ (c_u-c_l)}{3.92}$ when the sample size $n \gtrsim 100$. When IQR was unavailable, we approximated the medians as the means, assuming the distributions to be normal. For data from China and EU/EEA, where σ was available for the total population, consisting of all the transmission categories, we estimated σ for MSM and HSX using the ratios between σ corresponding to MSM or HSX and the total population reported from other studies (see footnote in Table S4). Similarly, for obtaining σ for HSX men and women, we employed the corresponding ratios of maximum σ from the CASCADE study. When information necessary to estimate σ was unavailable, we used the highest σ available from the most relatable dataset, as with the UK and the CASCADE study. To estimate the SD of $R_{T/F}$, we employed the

error propagation equation[24] and derived $\sigma(R_{T/F}) = \frac{\sqrt{\sigma^2_{infected}(\mu_{healthy}-T_{AIDS})^2+\sigma^2_{healthy}(\mu_{infected}-T_{AIDS})^2}}{(\mu_{healthy}-T_{AIDS})^2}$, where μ, σ are given in Tables 1 and S6. For $\sigma(R_{T/F})$ of all the data combined, we chose σ from EU/EEA, involving data from 21 countries.

Estimation of CD4 counts at seroconversion. In the US study[14], a model of CD4 count decline following seroconversion has been proposed, which allowed us to estimate CD4 counts at seroconversion from measurements at diagnosis. According to the model, the CD4 count T in an untreated individual at time t from seroconversion follows $\sqrt{T} = a_0 + b_1 \times t + e_{1t}$, where a_0 and b_1 are constants and e_{1t} is an error term. At seroconversion, the CD4 count, T_0, was obtained by setting $t = 0$, so that $\sqrt{T_0} = a_0 + e_{10}$. Assuming that $e_{1t} = e_{10}$, it followed that $T_0 = \left(\sqrt{T} - b_1 \times t\right)^2$. The values of b_1 for different age groups and transmission categories were available[25]. Also, the median delays (and IQR) in diagnosis following seroconversion, t_d, have been estimated[14], using which we calculated the corresponding mean and SD. For MSM and HSX, we took the mid-value of the means of t_d in 2006 and 2015 and chose the largest SD, and obtained $t_d = 4.05 \pm 6.67$ and $t_d = 5.40 \pm 9.04$ years, respectively, for the duration 2006-15. If T_d is the CD4 count at diagnosis, then $T_0 = \left(\sqrt{T_d} - b_1 \times t_d\right)^2$. We applied the analysis to data from the most populated age group (13-29 years) and used the mid-value, 21 years, for which b_1 was -0.93, -0.77, and -0.80 year^{-1} for MSM, HSX men, and HSX women, respectively. Furthermore, we assumed the fractions of females to be the same among all non-MSM groups, in order to obtain the population sizes of MSM and HSX in this age group (footnote in Table S3). Correspondingly, we obtained b_1 = -0.79 year^{-1} for HSX. To obtain uncertainties in the estimates of T_0, we repeated the above analysis with T_d and t_d set at values $\pm\sigma$ away from their respective means, but ensuring that their lowerbounds ≥ 0 and omitting terms that are second order in σ. Half the difference between the resulting maximum and minimum values of T_0 yielded the σ corresponding to seroconversion.

Statistical analysis. To examine whether the mean CD4 counts (or mean $R_{T/F}$) were significantly higher (or lower) in MSM than HSX, we employed the one-tailed t-test with unequal variance with the test statistic $t = \frac{\mu_{HSX}-\mu_{MSM}}{\sqrt{\sigma^2_{HSX}/n_{HSX}-\sigma^2_{MSM}/n_{MSM}}}$ and degrees of freedom $d = \frac{\left[\frac{1}{n_{HSX}} + \frac{(\sigma_{MSM}/\sigma_{HSX})^2}{n_{MSM}}\right]^2}{\left[\frac{1}{n^2_{HSX}(n_{HSX}-1)} + \frac{(\sigma_{MSM}/\sigma_{HSX})^4}{n^2_{MSM}(n_{MSM}-1)}\right]}$, where n_{HSX} and n_{MSM} were the two sample sizes, respectively[26]. The tests were performed using the R package[27], which yielded corresponding P values.

Data of HIV-1 subtype prevalence. To assess the extent of mixing between MSM and HSX, we

collated data of the prevalence of HIV-1 subtypes in the two groups across relevant geographical regions and calendar years. The data are summarized in Fig. 2 and Tables S1-S2 and discussed in Text S1.

Data of association with transmission clusters. Finally, we considered the extent of association of MSM and HSX with transmission clusters as an indicator of the time of onward transmission post-infection. The corresponding data we collated along with data of the compositions of the largest transmission clusters in different settings are in Fig. 3 and Table S7 and are discussed in Text S2.

Acknowledgments

We thank Pranesh Padmanabhan, Rajat Desikan, and Pradeep Nagaraja for comments. This work was supported by the DBT/Wellcome Trust India Alliance Senior Fellowship IA/S/14/1/501307 (NMD).

References

1 Carlson, J. M. *et al.* Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* **345** (2014). URL https://science.sciencemag.org/content/345/6193/1254031.

2 Joseph, S. B., Swanstrom, B., Kashuba, A. D. M. & Cohen, M. S. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. *Nat. Rev. Microbiol.* **13**, 414–425 (2015). URL https://doi.org/10.1038/nrmicro3471.

3 Tully, D. C. *et al.* Differences in the selection bottleneck between modes of sexual transmission influence the genetic composition of the HIV-1 founder virus. *PLOS Pathogens* **12**, 1–29 (2016). URL https://doi.org/10.1371/journal.ppat.1005619.

4 Claiborne, D. T. *et al.* Replicative fitness of transmitted HIV-1 drives acute immune activation, proviral load in memory CD4+ T cells, and disease progression. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1480–1489 (2015). URL https://www.pnas.org/content/112/12/E1480.

5 Carlson, J. M. et al. Impact of pre-adapted HIV transmission. *Nat. Med.* **22(6)**, 606–13 (2016). URL https://doi.org/10.1038/nm.4100.

6 Selhorst, P. *et al.* Replication capacity of viruses from acute infection drives HIV-1 disease progression. *Journal of Virology* **91** (2017). URL https://jvi.asm.org/content/91/8/e01806-16.

7 Patel, P. *et al.* Estimating per-act HIV transmission risk: a systematic review. *AIDS* **28(10)**, 1509–19 (2014).

8 Owen, B. N. *et al.* Prevalence and frequency of heterosexual anal intercourse among young people: a systematic review and meta-analysis. *AIDS Behav.* **19(7)**, 1338–60 (2015). URL https://doi.org/10.1007/s10461-015-0997-y.

9 Deeks, S. G., Overbaugh, J., Phillips, A. & Buchbinder, S. HIV infection. *Nat. Rev. Dis. Primers* **1**, 15035 (2015).

10 European Centre for Disease Prevention and Control. *HIV/AIDS surveillance in Europe 2019* (HIV/AIDS surveillance in Europe 2019 - 2018 data). URL https://www.ecdc.europa.eu/en/publications-data/hivaids-surveillance-europe-2019-2018-data. [Online; accessed 01-August-2020].

11 Frentz, D. et al. Patterns of transmitted HIV drug resistance in Europe vary by risk group. *PLoS One* **9(4)**, e94495 (2014). URL https://doi.org/10.1371/journal.pone.0094495.

12 Gupta, S. B., Gilbert, R. L., Brady, A. R., Livingstone, S. J. & Evans, B. G. CD4 cell counts in adults with newly diagnosed HIV infection: results of surveillance in England and Wales, 1990-1998. *AIDS* **14(7)**, 853–861 (2000).

13 Tang, H. *et al.* Baseline CD4 cell counts of newly diagnosed HIV cases in China: 2006–2012. *PLoS ONE* **9**, e96098 (2014). URL https://doi.org/10.1371/journal.pone.0096098.

14 Robertson, M. M., Braunstein, S. L., Hoover, D. R., Li, S. & Nash, D. Estimates of the time from seroconversion to ART initiation among people newly diagnosed with HIV from 2006 to 2015, New York City. *Clin. Infect. Dis.* **ciz1178** (2019).

15 CASCADE Collaboration. Differences in CD4 cell counts at seroconversion and decline among 5739 HIV-1–infected individuals with well-estimated dates of seroconversion. *J. Acquir. Immune Defic. Syndr.* **34**, 76–83 (2003). URL https://journals.lww.com/jaids/Fulltext/2003/09010/Differences_in_CD4_Cell_Counts_at_Seroconversion.12.aspx.

16 Lodi, S. *et al.* Time from human immunodeficiency virus seroconversion to reaching CD4+ cell count thresholds $< 200, < 350,$ and < 500 cells/mm^3 : assessment of need following changes in treatment guidelines. *Clin. Infect. Dis.* **53(8)**, 817–825 (2011). URL https://doi.org/10.1093/cid/cir494.

17 Beyrer, C. *et al.* Global epidemiology of HIV infection in men who have sex with men. *Lancet (London, England)* **380(9839)**, 367–377 (2012). URL https://doi.org/10.1016/

S0140-6736(12)60821-6.

18 Ariën, K. K., Vanham, G. & Arts, E. J. Is HIV-1 evolving to a less virulent form in humans? *Nat. Rev. Microbiol.* **5**, 141–151 (2007). URL https://doi.org/10.1038/nrmicro1594.

19 Shet, A., Nagaraja, P. & Dixit, N. M. Viral decay dynamics and mathematical modeling of treatment response: evidence of lower in vivo fitness of HIV-1 subtype C. *J. Acquir. Immune Defic. Syndr.* **73**, 245–251 (2016). URL https://journals.lww.com/jaids/Fulltext/2016/11010/Viral_Decay_Dynamics_and_Mathematical_Modeling_of.1.aspx.

20 Villabona-Arenas, C. J. *et al.* Number of HIV-1 founder variants is determined by the recency of the source partner infection. *Science* **369**, 103–108 (2020). URL https://science.sciencemag.org/content/369/6499/103.

21 Fraser, C. *et al.* Virulence and pathogenesis of HIV-1 infection: An evolutionary perspective. *Science* **343** (2014). URL https://science.sciencemag.org/content/343/6177/1243727.

22 Wan, X., Wang, W., Liu, J. & Tong, T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med. Res. Methodol.* **14** (2014). URL https://doi.org/10.1186/1471-2288-14-135.

23 Higgins J. P. T., Thomas J., Chandler J., Cumpston M., Li T., Page M. J. & Welch V. A. (ed.) *Cochrane Handbook for Systematic Reviews of Interventions* (Cochrane, 2019). URL www.training.cochrane.org/handbook. [version 6.0 (updated July 2019)].

24 Bevington, P. R. and Robinson, D. K. *Data Reduction and Error Analysis for the Physical Sciences* (McGraw Hill, 2003), 3 edn.

25 Song, R., Hall, H. I., Green, T. A., Szwarcwald, C. L. & Pantazis, N. Using CD4 data to estimate HIV incidence, prevalence, and percent of undiagnosed infections in the United States. *J. Acquir. Immune Defic. Syndr.* **74**, 3–9 (2017).

26 Ruxton, Graeme D. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behav. Ecol.* **17**, 688–690 (2006). URL https://doi.org/10.1093/beheco/ark016.

27 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2013). URL http://www.R-project.org/.

SUPPLEMENTARY INFORMATION

Transmitted HIV-1 is more pathogenic in heterosexual individuals than homosexual men

Ananthu James[a] and Narendra M. Dixit[a,b,*]

[a]Department of Chemical Engineering, Indian Institute of Science, Bengaluru, India
[b]Centre for Biosystems Science and Engineering, Indian Institute of Science, Bengaluru, India

*Corresponding author; E-mail: narendra@iisc.ac.in

Contents:

Text S1: Distinct subtype prevalences indicate minimal mixing between MSM and HSX

In many geographical locations, mixing between MSM and HSX appears minimal. This is evident from the different prevalences of HIV-1 subtypes in the two groups. MSM in western nations are dominated by HIV-1 subtype B, whereas HSX comprise a mixture of subtypes[1], with subtypes B and C being the predominant ones[2]. For instance, in the United Kingdom, from 2002-2010, MSM had nearly 90% subtype B infections, whereas HSX had a little over 10% subtype B. Mixing between the two groups would have led to a more similar distribution of subtypes in the two. The two groups thus appear to have sustained their respective infections over the years in near complete isolation. The difference in subtype prevalences holds also in Canada, Spain, France, and other nations (Fig. 2(a); Table S1). In China, the dominant subtype is CRF01-AE, which is present in MSM with a frequency of $>$50% but in HSX at $<$40% (Fig. 2(b); Table S2)[3], perhaps indicative of more mixing than in Europe. In Korea, the extent of mixing could not be assessed using subtypes because over 80% of all infections were subtype B[4]. In USA, though subtype B dominates both MSM and HSX[5,6], mixing between the groups has been argued not to be common[7]. In the Nordic states, some mixing between MSM and HSX is evident[8]. Overall, little mixing between MSM and HSX is evident in most geographical settings, suggesting that the different selection biases between the groups may have been sustained over the course of the epidemic.

Text S2: Clustering and transmission patterns

MSM are known to engage in different sexual contact patterns compared to HSX. They tend to have more partners than HSX[9,10]. They are also far more likely to belong to transmission clusters compared to HSX[1]. A transmission cluster comprises individuals carrying viral genomes that cluster together in a phylogenetic tree[11], suggesting that the viral sequences isolated from the individuals are closely related. In Japan and China, an infected MSM had a nearly 40% chance of being part of a cluster, whereas an infected HSX had $<$10% chance[12]. In France, the corresponding numbers were $\sim$35% and $\sim$4%, respectively[13]. This trend was true for all the countries with data available except the Netherlands (Fig. 3(a); Table S7). MSM also formed larger clusters than HSX. The largest clusters reported in Belgium and Spain comprised nearly 100 individuals each, with the Belgian cluster containing $\sim$70 MSM and the Spanish cluster exclusively MSM (Fig. 3(b); Table S7)[14,15]. Together, these

data suggest greater similarity in the viral strains in MSM than HSX. One way in which this greater similarity could arise is by onward transmission occurring sooner after infection in MSM than HSX, allowing lesser individual host-specific adaptation before transmission.

Supplementary Figures

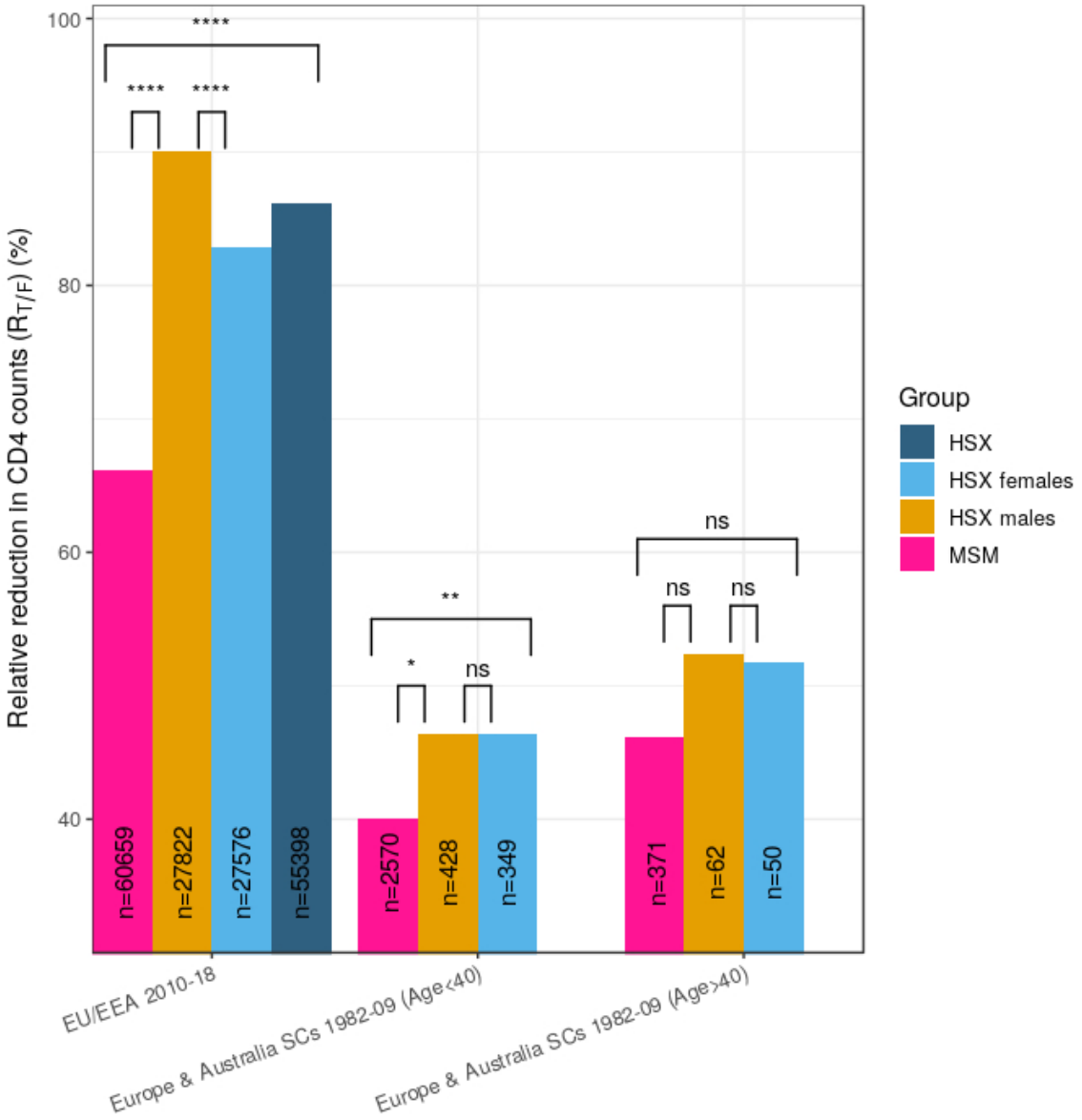

Figure S1. Effect of gender on the relative reduction in early CD4 counts. $R_{T/F}$ from EU/EEA during 2010-18[16] and the CASCADE study[17] indicate that HSX men have higher $R_{T/F}$ than MSM, with this difference achieving significance with large sample sizes, ruling out gender as a cause of the lower CD4 count reduction in MSM than HSX. The sample sizes (n) are indicated along with the P values. (****, ***, ** and * indicate P$< 10^{-4}$, P$< 10^{-3}$, P$< 10^{-2}$ and P< 0.05, respectively, while ns (not significant) implies P> 0.05.) Additional details are in Table 1 (main text).

Supplementary Tables

Table S1. Prevalence of subtype B in Europe. Data from different regions in Europe show substantially higher subtype B percentage prevalence in MSM than HSX ($P < 0.001$ in each study, unless specified). The sample sizes (n) are in parantheses.

Region [Reference]	Survey year(s)	Prevalence of subtype B (%)	
		MSM (n)	HSX (n)
France[18]	2001	80.5 (113)	53.2 (186)
	2006 – 07	90.6 (223)	27.2 (202)
	2010 – 11	83.3 (311)	26.2 (263)
Italy[19]	2000 – 14	81.5 (1,824)	54.9 (1,573)
UK[20]	2002 – 10	88.4‡ (21,321)	13.3‡ (38,948)
Madrid, Spain[21]	2000 – 11	88.2 (340)	41.5 (369)
Comunidad Valenciana (CV), Spain[15]	2004 – 14	92.0‡ (637)	60.8‡ (204)
Europe[22]	2002 – 07	90.4 (2,084)	33.5 (1,501)
Canada* & Europe[23]	1998 – 10	95.5‡ (5,570)	38.7‡ (2,052)
Japan[12]	2004 – 11	96.9‡ (261)	39.8‡ (103)

‡P value not specified.
*In this study, 15.9% patients were from Canada.

Table S2. Prevalence of subtypes in China. A recent review[3] of 130 published articles, together involving of 10,516 MSM and 6,759 HSX individuals, has examined the prevalence of different subtypes in China, which is reproduced below. P values indicate significant differences in the prevalences of 3 subtypes.

Subtype	Prevalence (%) (95% CI)		
	MSM	HSX	P value
CRF01_AE	51.28 (46.15 – 56.40)	38.78 (33.08 – 44.63)	< 0.01
CRF07_BC	19.98 (16.17 – 24.07)	14.88 (10.96 – 19.23)	0.083
CRF08_BC	0.00 (0.00 – 0.00)	9.81 (6.44 – 13.70)	< 0.01
B\B′	17.74 (12.78 – 23.26)	15.41 (11.15 – 20.16)	0.508
C	0.00 (0.00 – 0.00)	1.89 (0.81 – 3.28)	< 0.01
Others	1.56 (0.98 – 2.24)	2.39 (1.33 – 3.66)	0.2109

Table S3. Early median CD4 cell counts in infected adults from several large population studies*. The sources of the studies, the periods of study, measurement times, and other details are mentioned. The USA and European studies report IQRs, whereas the CASCADE study provides 95% CIs. NA - not available.

Region	Time period	Measurement time	Risk group	CD4 counts $(cells/\mu L)$	Sample size (n)	Other details
USA‡ [24]	$2006-15$	Within 6 m.o. diagnosis	MSM	$387\ (227-555)$	$17,779$	
			HSX	$308\ (125-508)$	$8,310$	
Europe† [22]	$2002-07$	Within 6 m.o. diagnosis	MSM	$435\ (259-585)$	$2,084$	$P<10^{-3}$
			HSX	$280\ (110-458)$	$1,501$	
England & Wales [25]	$1990-98$	Within 6 m.o. diagnosis	MSM	331 (NA)**	$6,213$	$P<10^{-3}$
			HSX	230 (NA)**	$2,637$	
Europe & Australia [17] (CASCADE) SCs††	$1979-00$	Seroconversion	MSM (age<40)	$621\ (609-634)$	$2,941$	n=5739. 54.7% (MSM & hemophiliac men) only men. 20.4% women. 12.6% aged>40.
			MSM (age>40)	$578\ (556-601)$		
			HSX men (age<40)	$576\ (545-607)$	889	
			HSX men (age>40)	$534\ (501-569)$		
			HSX women (age<40)	$623\ (599-647)$		
			HSX women (age>40)	$580\ (548-612)$		

*We did not include studies comparing CD4 counts in MSM and the general population, the latter consisting of not only MSM and HSX, but also considerable fractions of injection drug users and hemophiliacs [26]. The latter two groups may have significantly different early CD4 counts than MSM and HSX [17], precluding a direct comparison between MSM and HSX. When many articles were available for overlapping durations, corresponding to the same geographic location (as in the case of New York City), we chose data from the study having a higher sample size [24]. Interestingly, we found no study with n $\gtrsim 1000$ where mean/median early cell counts among HSX were significantly higher than those among MSM [4].

‡Corresponding to this population, the proportion of females was 22.2% and the age group 13-29 had 35.6% of $n = 28,162$ individuals.

†Median viral load (with IQR) in log copies/mL is $4.88\ (4.3-5.4)$ for MSM and $4.79\ (4.2-5.3)$ for HSX $(P<0.001)$ [22].

**We chose the SDs from the Europe (2002-07) study (see Methods).

††For the whole sets of MSM and HSX, including all the age groups, we used the SDs corresponding to age <40 years. For HSX, we used the largest SD among HSX men and women.

Table S4. Early median CD4 cell counts in infected MSM and HSX from China[27]. The sample sizes (n) were directly available, while the cell counts were estimated using WebPlotDigitizer[28]. The last row represents estimates (see Methods) for the entire period 2006-12.

Year	MSM		HSX	
	Cell counts $(cells/\mu L)$	Sample size (n)	Cell counts $(cells/\mu L)$	Sample size (n)
2006	315	244	230	3,305
2007	345	672	249	7,251
2008	362	2,025	266	12,378
2009	364	3,624	265	18,208
2010	368	5,648	267	24,425
2011	370	9,316	271	35,211
2012	371	13,748	281	42,653
Average	368 ± 222 (SD†)	35,277	270 ± 260 (SD†)	143,431

†Corresponding to the total population that contributed to the CD4 count data in China, consisting of MSM, HSX and other transmission categories, during $2006-12$, the IQR was $130-454$ $cells/\mu L$. Using this, we obtained SD $= 240$ $cells/\mu L$ (see Methods). To calculate the SDs corresponding to MSM and HSX, we recall that in the USA[24] the SDs corresponding to the CD4 counts from MSM, HSX and the total population were 243, 284 and 262 $cells/\mu L$, respectively. Assuming that the ratios between SDs remain the same in China as well, we evaluated the SD for MSM to be 222 and HSX to be 260 $cells/\mu L$ in China.

Table S5. Early median CD4 cell counts in infected adults from EU/EEA[16]. The cell counts were estimated using WebPlotDigitizer[28]. The median ages, whenever available, are also provided. The last row provides the mean cell counts (with SDs) and total numbers of MSM, HSX men and women, respectively, estimated as in Methods.

Year	No. of diagnoses*		Fraction with	Median age		Median CD4 count $(cells/\mu L)$		
	MSM	HSX	CD4 counts††	MSM	HSX	MSM	HSX men	HSX women
2010	10,348	11,693	0.591	-	-	420	261	312
2011	10,411	11,011	0.561	-	-	422	260	317
2012	11,238	10,782	0.553	-	-	436	270	319
2013	11,553	10,187	0.607	-	-	443	271	339
2014	11,821	10,049	0.614	-	-	450	281	346
2015	11,484	9,267	0.747	-	-	454	284	345
2016	10,508	9,003	0.673	34	39	440	274	351
2017	9,723	8,518	0.717	34	39	434	268	349
2018	8,049	7,267	0.695	36	41	423	262	340
Total	$n^{**}_{MSM} = 60,659$, $n^{**}_{HSXM} = 27,822$, $n^{**}_{HSXW} = 27,576$					$(437 \pm 397)^{\ddagger}$	$(270 \pm 464)^{\ddagger}$	$(335 \pm 325)^{\ddagger}$

*The number of diagnoses specific to each country during 2009-18 was provided for MSM and HSX in the latest annual report[16]. Using these, we estimated the numbers above.

††These are taken from the annual surveillance reports[29]. Since this number was unavailable for 2009, we omitted the data for the calculation of the mean CD4 counts for the combined data.

**The numbers of HSX men and women diagnosed from the 21 EU/EEA countries in 2018 were available in the 2019 annual report[16]. Using this, we calculated the fractions of HSX men and women and assumed that they remained the same during 2010-18. Also, $n_{HSX} = n_{HSXmen} + n_{HSXwomen} = 55,398$. (There were a small number of transgenders among HSX in column 3, but the CD4 counts were reported only for HSX men and women.)

‡In 2018, for 16,694 adults across risk groups, the median cell counts with 95% CIs were 365 $(359 - 372)$ $cells/\mu L$. Using this, we computed their SD to be ~ 429 $cells/\mu L$ and evaluated the SDs for MSM and HSX. To obtain the SDs for HSX men and women, we used the SD ratio for the age group < 40 years corresponding to the CASCADE study (see Methods).

Table S6. CD4 T cell counts in healthy adults. Mean CD4 counts in healthy adults from different population groups which define baseline counts for estimating the relative reduction in early cell count following HIV-1 infection. Sample sizes are in brackets. SD is standard deviation.

Region	Category (sample size, n)	Cell counts ($cells/\mu L$)	Other details
China[30,31] (Hong Kong[30] & Shanghai[31])	Men (78[30] & 377[31])	725	Combined mean; SD $= 258$[30] & 256[31].
	Women (130[30] & 237[31])	724	Combined mean; SD $= 254$[30] & 255[31].
	HSX (822)	725	SD $= 255$ for Shanghai[31]
Sub-Saharan Africa (Tanzania)[32]	Men* (42)	666	SD $= 247$
	Women* (60)	802	SD $= 250$
	HSX (102)	746	SD $= 257$‡‡
USA[33]	Men** (33)	921	SD $= 188$
	Women (67)	1041	SD $= 340$
	HSX** (100)	1001	SD $= 305$
UK[34]	HSX men† (50)	840	SD $= 285$
	HSX women (50)	1050	SD $= 377$
	HSX‡ (100)	945	We used the SD for women.
	MSM†,‡ (100)	800	SD $= 324$
EU/EEA (& Europe) & Australia (Italy)[35]	General population or HSX†† (965)	941	SDs were not reported. We used the SDs from the UK study (see Methods).
	Men†† (532)	902	
	Women (436)	989	

*$P = 8 \times 10^{-3}$ for the comparison between men and women (reported in the original study[32]).
**$P = 0.04$ for the comparison between healthy men and HSX from USA.
†$P = 0.22$ for the comparison between MSM and HSX men from UK.
‡$P = 6.6 \times 10^{-4}$ for the comparison between healthy MSM and HSX from UK, using SD ($= 305$) for HSX from USA. Using SD (=377) corresponding to HSX women from UK instead yielded $P = 2 \times 10^{-3}$.
††$P = 0.018$ using the SDs from UK.
‡‡For our calculation involving the EU/EEA 2010-18 population (footnote in Table 1), we used the SD (=377) for the HSX from Europe/UK.

Table S7. Association of MSM and HSX with transmission clusters. The percentages of individuals found to be associated with transmission clusters in MSM and HSX in several studies are collated. In the second column, the numbers in parantheses indicate sample sizes examined. Where available, P values and the largest cluster sizes are indicated in other details.

Region (Study period)	Group (n)	Associated with clusters (%)	Other details
Japan[12] (2004 – 11)	MSM (261)	38.3	Clusters are also linked to China.
	HSX (103)	8.7	
Austria[36] (2008 – 14)	MSM (122)	57.4	$P < 0.001$
	HSX (115)	32.2	
Belgium[14] (2013 – 15)	MSM (698)	57.7	$P < 0.001$. Cluster size ≥ 3. The largest cluster is sized 94, consisting of 68 MSM and 15 HSX.
	HSX (508)	22.4	
Cyprus[37] (1986 – 2012)	MSM (164)	36.6	Cluster size ≥ 3. The two largest clusters are sized 13, having $11-12$ MSM.
	HSX (129)	2.3	
France[13] (1999 – 2014)	MSM (636)	32.1	$P < 0.01$. Cluster size ≥ 3. The largest cluster, sized 41, has 39 MSM and a HSX.
	HSX (267)	3.8	
Germany[38] (1999 – 2016)	MSM (1,448)	27.3	
	HSX (622)	18.2	
Netherlands[39] (1981 – 2011)	MSM (4,288)	51.2	Cluster size ≥ 10.
	HSX (849)	55.8	
Spain[15] (2004 – 14)	MSM (637)	–	8 of 12 clusters sized ≥ 10 consist only of MSM. The largest cluster has a size of 111, and is composed only of MSM.
	HSX (204)	–	
USA[40] (2001 – 15)	MSM (1,597)	37.4	The largest cluster has a size of 39, and has only MSM.
	HSX (953)	28.3	
Canada and 9 European countries[41] (1981 – 2011)	MSM (4,980)	51.4	The 9 countries are Italy, Greece, Netherlands, Germany, France, UK, Norway, Austria, and Spain. $P < 0.001$.
	Bisexual men (2,087)	27.7	

References

1 Beyrer, C., Baral, S. D., van Griensven, F., Goodreau, S. M., Chariyalertsak, S., Wirtz, A. L., and Brookmeyer, R. Global epidemiology of HIV infection in men who have sex with men. *Lancet (London, England)* **380(9839)**, 367–377 (2012). URL https://doi.org/10.1016/S0140-6736(12)60821-6.

2 Abecasis, A. B. et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology* **10** (2013). URL https://doi.org/10.1186/1742-4690-10-7.

3 Yuan, R., Cheng, H., Chen, L. S., Zhang, X., and Wang, B. Prevalence of different HIV-1 subtypes in sexual transmission in China: a systematic review and meta-analysis. *Epidemiol. Infect.* **144(10)**, 2144–2153 (2016).

4 Kim, G. J. *et al.* Estimating the origin and evolution characteristics for Korean HIV type 1 subtype B using Bayesian phylogenetic analysis. *AIDS Res. Hum. Retroviruses* **28(8)**, 880–884 (2012). URL https://doi.org/10.1089/aid.2011.0267.

5 Junqueira, D. M. and Almeida, S. E. d. M. HIV-1 subtype B: traces of a pandemic. *Virology* **495**, 173–184 (2016).

6 Dennis, A. M. *et al.* Rising prevalence of non-B HIV-1 subtypes in North Carolina and evidence for local onward transmission. *Virus Evol.* **3(1)**, vex013 (2017). URL https://doi.org/10.1093/ve/vex013.

7 Satcher, A. J., Durant, T., Hu, X., and Dean, H. D. AIDS cases among women who reported sex with a bisexual man, 2000-2004–United States. *Women & Health* **46**, 23–40 (2007).

8 Esbjörnsson, J. *et al.* HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus Evolution* **2** (2016). URL https://doi.org/10.1093/ve/vew010.

9 Glick, S. N. *et al.* A comparison of sexual behavior patterns among men who have sex with men and heterosexual men and women. *J. Acquir. Immune Defic. Syndr.* **60(1)**, 83–90 (2012). URL https://doi.org/10.1097/QAI.0b013e318247925e.

10 Kenyon, C. R., Wolfs, K., Osbak, K., van Lankveld, J., and Van Hal, G. Implicit attitudes to sexual partner concurrency vary by sexual orientation but not by gender—A cross sectional study of

Belgian students. *PLoS One* **13(5)**, e0196821 (2018).

11 Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., and Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS* **31(9)**, 1211–1222 (2017).

12 Kondo, M. *et al.* Emergence in Japan of an HIV-1 variant associated with transmission among men who have sex with men (MSM) in China: first indication of the international dissemination of the Chinese MSM lineage. *Journal of Virology* **87**, 5351–5361 (2013). URL https://jvi.asm.org/content/87/10/5351.

13 Chaillon, A. et al. Spatiotemporal dynamics of HIV-1 transmission in France (1999-2014) and impact of targeted prevention strategies. *Retrovirology* **14(1)**, 15 (2017). URL https://doi.org/10.1186/s12977-017-0339-4.

14 Verhofstede, C. et al. Phylogenetic analysis of the Belgian HIV-1 epidemic reveals that local transmission is almost exclusively driven by men having sex with men despite presence of large African migrant communities. *Infect. Genet. Evol.* **61**, 36–44 (2018). URL https://doi.org/10.1016/j.meegid.2018.03.002.

15 Patino-Galindo, J. A. et al. The molecular epidemiology of HIV-1 in the Comunidad Valenciana (Spain): analysis of transmission clusters. *Sci. Rep.* **7**, 11584 (2017). URL https://doi.org/10.1038/s41598-017-10286-1.

16 European Centre for Disease Prevention and Control. *HIV/AIDS surveillance in Europe 2019* (HIV/AIDS surveillance in Europe 2019 - 2018 data). URL https://www.ecdc.europa.eu/en/publications-data/hivaids-surveillance-europe-2019-2018-data. [Online; accessed 01-August-2020].

17 CASCADE Collaboration. Differences in CD4 cell counts at seroconversion and decline among 5739 HIV-1–infected individuals with well-estimated dates of seroconversion. *J. Acquir. Immune Defic. Syndr.* **34**, 76–83 (2003). URL https://journals.lww.com/jaids/Fulltext/2003/09010/Differences_in_CD4_Cell_Counts_at_Seroconversion.12.aspx.

18 Descamps, D. et al. National sentinel surveillance of transmitted drug resistance in antiretroviral-naive chronically HIV-infected patients in France over a decade: 2001–2011. *J. Antimicrob. Chemother.* **68(11)**, 2626–2631 (2013). URL https://doi.org/10.1093/jac/dkt238.

19 Fabeni, L. et al. Dynamics and phylogenetic relationships of HIV-1 transmitted drug resistance ac-

cording to subtype in Italy over the years 2000–14. *J. Antimicrob. Chemother.* **72(10)**, 2837–2845 (2017). URL https://doi.org/10.1093/jac/dkx231.

20 The UK Collaborative Group on HIV Drug Resistance. The increasing genetic diversity of HIV-1 in the UK, 2002–2010. *AIDS* **28(5)**, 773–780 (2014). URL https://journals.lww.com/aidsonline/Fulltext/2014/03130/The_increasing_genetic_diversity_of_HIV_1_in_the.15.aspx.

21 Yebra, G. et al. Different trends of transmitted HIV-1 drug resistance in Madrid, Spain, among risk groups in the last decade. *Arch. Virol.* **159(5)**, 1079–87 (2014). URL https://doi.org/10.1007/s00705-013-1933-y.

22 Frentz, D. et al. Patterns of transmitted HIV drug resistance in Europe vary by risk group. *PLoS One* **9(4)**, e94495 (2014). URL https://doi.org/10.1371/journal.pone.0094495.

23 Klein, M. B. et al. The effects of HIV-1 subtype and ethnicity on the rate of CD4 cell count decline in patients naive to antiretroviral therapy: a Canadian-European collaborative retrospective cohort study. *CMAJ OPEN* **2(4)**, E318–E329 (2014). URL https://doi.org/10.9778/cmajo.20140017.

24 Robertson, M. M., Braunstein, S. L., Hoover, D. R., Li, S., and Nash, D. Estimates of the time from seroconversion to ART initiation among people newly diagnosed with HIV from 2006 to 2015, New York City. *Clin. Infect. Dis.* **ciz1178** (2019).

25 Gupta, S. B., Gilbert, R. L., Brady, A. R., Livingstone, S. J., and Evans, B. G. CD4 cell counts in adults with newly diagnosed HIV infection: results of surveillance in England and Wales, 1990-1998. *AIDS* **14(7)**, 853–861 (2000).

26 Pantazis, N. et al. Temporal trends in prognostic markers of HIV-1 virulence and transmissibility: an observational cohort study. *Lancet HIV* **1(3)**, e119–26 (2014). URL https://doi.org/10.1016/S2352-3018(14)00002-2.

27 Tang, H. *et al.* Baseline CD4 cell counts of newly diagnosed HIV cases in China: 2006–2012. *PLoS ONE* **9**, e96098 (2014). URL https://doi.org/10.1371/journal.pone.0096098.

28 Rohatgi, A. *WebPlotDigitizer*. Pacifica, California, USA (July, 2020). URL https://automeris.io/WebPlotDigitizer.

29 European Centre for Disease Prevention and Control. *Annual HIV/AIDS surveillance re-*

ports (HIV/AIDS surveillance in Europe). URL https://www.ecdc.europa.eu/en/all-topics-zhiv-infection-and-aidssurveillance-and-disease-data/annual-hivaids-surveillance-reports. [Online; accessed 01-August-2020].

30 Kam, K. M. *et al.* Lymphocyte subpopulation reference ranges for monitoring human immunodeficiency virus-infected Chinese adults. *Clinical and Vaccine Immunology* **3**, 326–330 (1996). URL https://cvi.asm.org/content/3/3/326.

31 Jiang, W. *et al.* Normal values for CD4 and CD8 lymphocyte subsets in healthy Chinese adults from Shanghai. *Clinical and Vaccine Immunology* **11**, 811–813 (2004). URL https://cvi.asm.org/content/11/4/811.

32 Ngowi, B. J., Mfinanga, S. G., Bruun, J. N., and Morkve, O. Immunohaematological reference values in human immunodeficiency virus-negative adolescent and adults in rural northern Tanzania. *BMC Infect Dis.* **9** (2009).

33 Valiathan, R. et al. Reference ranges of lymphocyte subsets in healthy adults and adolescents with special mention of T cell maturation subsets in adults of South Florida. *Immunobiology* **219(7)**, 487–496 (2014). URL https://doi.org/10.1016/j.imbio.2014.02.010.

34 Bofill, M. e. a. Laboratory control values for CD4 and CD8 T lymphocytes. Implications for HIV-1 diagnosis. *Clinical & Experimental Immunology* **88(2)**, 243–252 (1992). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2249.1992.tb03068.x.

35 Santagostino, A. et al. An Italian national multicenter study for the definition of a reference ranges for normal values of peripheral blood lymphocyte subsets in healthy adults. *Haematologica* **84(6)**, 499–504 (1999). URL https://pubmed.ncbi.nlm.nih.gov/10366792/.

36 Hoenigl, M. et al. Characterization of HIV transmission in South-east Austria. *PLoS ONE* **11(3)**, e0151478 (2016). URL https://doi.org/10.1371/journal.pone.0151478.

37 Pineda-Peña, A. et al. HIV-1 infection in Cyprus, the Eastern Mediterranean European frontier: a densely sampled transmission dynamics analysis from 1986 to 2012. *Sci. Rep.* **8** (2018). URL https://doi.org/10.1038/s41598-017-19080-5.

38 Stecher, M. et al. Molecular epidemiology of the HIV epidemic in three German metropolitan regions – Cologne/Bonn, Munich and Hannover, 1999–2016. *Sci. Rep.* **8** (2018). URL https://doi.org/10.1038/s41598-018-25004-8.

151 39 Bezemer, D. *et al.* Dispersion of the HIV-1 epidemic in men who have sex with men in the Nether-
152 lands: a combined mathematical model and phylogenetic analysis. *PLoS Med.* **12(11)**, e1001898
153 (2015). URL https://doi.org/10.1371/journal.pmed.1001898.

154 40 Dennis, A. M. et al. HIV-1 transmission clustering and phylodynamics highlight the important role
155 of young men who have sex with men. *AIDS Res. Hum. Retroviruses* **34(10)**, 879–88 (2018). URL
156 https://doi.org/10.1089/aid.2018.0039.

157 41 Paraskevis, D. et al. HIV-1 molecular transmission clusters in nine European countries and Canada:
158 association with demographic and clinical factors. *BMC Medicine* **17** (2019). URL https://doi.
159 org/10.1186/s12916-018-1241-1.