Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions

Stéphane Hué[a], Alison E. Brown[b], Manon Ragonnet-Cronin[c], Samantha J. Lycett[c,d], David T. Dunn[e], Esther Fearnhill[e], David I. Dolling[e], Anton Pozniak[f], Deenan Pillay[a], Valerie C. Delpech[b], Andrew J. Leigh Brown[c], on behalf of the UK Collaboration on HIV Drug Resistance and the Collaborative HIV, Anti-HIV Drug Resistance Network (CHAIN)

Objective: HIV-1 subtype B infections are associated with MSM in the UK. Yet, around 13% of subtype B infections are found in those reporting heterosexual contact as transmission route. Using phylogenetics, we explored possible misclassification of sexual exposure among men diagnosed with HIV in the UK.

Design: Viral gene sequences linked to patient-derived information were used to identify phylogenetic transmission chains.

Methods: A total of 22 481 HIV-1 subtype B *pol* gene sequences sampled between 1996 and 2008 were analysed. Dated phylogenies were reconstructed and transmission clusters identified as clades of at least two sequences with a maximum genetic distance of 4.5%, a branch support of at least 95% and spanning 5 years. The characteristics of clusters containing at least one heterosexually acquired infection were analysed.

Results: Twenty-nine percent of the linked heterosexuals clustered exclusively with MSM. These were more likely to be men than women. Estimated misclassification of homosexually acquired infections ranged between 1 and 11% of the reported male heterosexuals diagnosed with HIV. Black African heterosexual men were more often phylogenetically linked to MSM than other ethnic group, with an estimated misclassification range between 1 and 21%.

Conclusion: Overall, a small proportion of self-reported heterosexual men diagnosed with HIV could have been infected homosexually. However, up to one in five black African heterosexual men chose not to disclose sex with men at HIV diagnosis and preferred to be identified as heterosexual. Phylogenetic analyses can enhance surveillance-based risk information and inform national programmes for monitoring and preventing HIV infections.

AIDS 2014, **28**:1967–1975

Keywords: heterosexual transmission, HIV, molecular epidemiology, MSM, phylogenetics

[a]Department of Infection, University College London, [b]HIV and STI Department, Public Health England, London, [c]School of Biological Sciences, University of Edinburgh, Edinburgh, [d]Institute of Biodiversity, University of Glasgow, Glasgow, [e]Medical Research Council Clinical Trial Unit, and [f]HIV and Sexual Health Clinic, Chelsea and Westminster Hospital, London, UK.

Correspondence to Dr Stéphane Hué, Department of Infection, University College London, Cruciform Building, Gower Street, London WC1E 6BT, UK.

Tel: +44 203 3108 2131; e-mail: stephane.hue@ucl.ac.uk

Received: 10 March 2014; revised: 11 June 2014; accepted: 11 June 2014.

DOI:10.1097/QAD.0000000000000383

ISSN 0269-9370 © 2014 Wolters Kluwer Health | Lippincott Williams & Wilkins

Introduction

By the end of 2012, an estimated 98 400 people were living with HIV in the UK [1]. Like many resource-rich countries, the highest prevalence rate is amongst MSM [2]. Although the overall number of new HIV diagnoses has been on the decline since 2005, new diagnoses in this group continue to rise, surpassing the number of diagnoses among heterosexuals in 2011. In 2012, 3250 MSM were newly diagnosed, which is the highest level ever reported.

HIV-1 subtype B virus causes an estimated 40% of HIV infections diagnosed in the UK. Whilst this strain is mainly found in MSM [3], 10–13% of persons diagnosed annually with HIV-1 subtype B infections between 2002 and 2010 reported heterosexual contact as their most probable route of infection. This trend was most notable among men (men-to-women ratio: 55 : 45). It is unclear whether the profile of the subtype B epidemic over the past decade is the result of increased mixing between MSM and heterosexual communities, or signifies potential nondisclosure of sex with other men among reported heterosexuals.

The diversity of RNA viral genomes is a valuable source of information when studying epidemiological trends and makes it possible to reconstruct the trajectory of specific viral strains within an infected population by phylogenetic methods. This approach is now well established and has been extensively applied to the study of historical epidemics (e.g. [4,5]), sub-epidemics within risk or demographic groups (e.g. [6,7]) or even discrete transmission chains (e.g. [8–10]).

The characterization of transmission chains within a viral phylogeny involves the identification of clusters, or subtrees, fulfilling criteria empirically determined so as to represent linked transmissions. Although a variety of criteria are used, they usually include a minimum number of clustered sequences (e.g. two or more), minimal intra-cluster genetic differences defined in various ways (e.g. $\leq$0.045 substitutions per sites in [11]) and/or strong support for the branch leading to the most recent common ancestor of the clade of viruses (e.g. Bayesian posterior probability $\geq$1.00 in [12]). Dated phylogenies constructed under a Bayesian statistical framework with molecular clock inference can also be used to determine a fixed time frame within which transmission events occur [6,13].

We hypothesized that a proportion of HIV-1 subtype B infections amongst male heterosexuals may be misclassified MSM infections. We therefore identified and analysed transmission chains involving reported heterosexuals, and quantified the proportion of HIV subtype B infections that may be incorrectly ascribed to heterosexual risk rather than sex between men. This study is the first attempt to quantify nondisclosure of homosexually acquired infections reported through national surveillance and illustrates how phylogenetic inference can complement traditional epidemiological analyses.

Methods

Sequence data

A total of 22 481 HIV-1 subtype B partial *pol* gene sequences sampled between 1996 and 2008 by the UK HIV Drug Resistance database (HDRdb) were analysed. The HDRdb is a central repository for resistance tests performed as part of routine clinical care throughout the UK (http://www.hivrdb.org.uk/). Sequences span the entire protease (297 nucleotides) and first 1248 nucleotides of the reverse transcriptase of the virus. Sequences were sub-typed using the algorithm SCUEAL [14]. For patients with multiple sequences, the first sequence available was selected for analysis. At the time of sampling, 61 and 28% of the studied patients were recorded as being antiretroviral treatment-naive and experienced, respectively. The remaining 11% had unknown treatment status.

Patient information

Each sequence was linked to clinical and demographic data gathered by the UK Collaborative HIV Cohort Study (UK CHIC) [15] or the HIV and AIDS Reporting System at Public Health England (HARS) (http://www.hpa.org.uk/). UK CHIC has been collecting demographic, clinical and laboratory data on HIV-positive patients from 13 of the largest clinics within the UK since 2002. HARS is a comprehensive national cohort of patients newly diagnosed and retained in HIV care over time. The available information included sex, risk group, ethnicity, age group, treatment status, as well as the codified geographical location of the patients. Exposure groups were classified as: MSM ($n = 14\,651$), heterosexual contact ($n = 2153$), other (including injecting drug users, contact with blood product and mother to child transmission; $n = 773$) and not known ($n = 4904$). Ethnic groups were divided into black African ($n = 398$), black Caribbean ($n = 773$), white ($n = 14\,997$), other (including other black, Indian/Pakistani/Bangladeshi, other Asian/Oriental and other/mixed; $n = 1474$) and not known ($n = 4839$). A total of 4531 sequences (20%) could not be linked to the demographic data. All information was pseudo-anonymized prior to the analysis. The sequences curated by the HDRdb encompass around 46% of cumulative HIV infections in the UK since 1996, and are representative of the known UK HIV epidemic in terms of risk groups, sex, ethnicity and age at diagnosis (Supplementary Fig. 1, http://links.lww.com/QAD/A547). The individuals reporting heterosexually acquired HIV infections in the study cohort are detailed in Table 1.

Table 1. Characteristics of clustered heterosexuals (≤5 years).

		All heterosexuals		Female		Male				
		Total	Clustered (%)	Total	Clustered (%)	Total	Clustered (%)	Odds ratio	95% CI	*P*
Total *N* (%)		2153	566 (26)	1116	295 (26)	1037	271 (26)	1.02	0.84–1.23	0.874
Clustered with	Heterosexuals only	–	293 (14)	–	182 (16)	–	111 (11)	1.63	1.26–2.09	**0.0002**
	MSM only	–	163 (8)	–	52 (5)	–	111 (11)	0.41	0.29–0.57	**<0.0001**
	Other heterosexuals and MSM	–	10 (<1)	–	2 (<1)	–	8 (<1)	0.23	0.05–1.10	0.064
	Other (including IDU)	–	26 (1)	–	16 (1)	–	10 (1)	1.49	0.67–3.31	0.322
	Not known	–	84 (4)	–	43 (4)	–	31 (3)	1.30	0.81–2.08	0.273
Ethnicity	Black African	230	71 (31)	120	34 (28)	110	37 (34)	0.78	0.45–1.36	0.385
	Black Caribbean	416	120 (29)	267	82 (31)	149	38 (26)	1.29	0.82–2.03	0.261
	White	1200	295 (25)	581	142 (24)	619	153 (25)	0.55	0.20–1.53	0.860
	Other	263	70 (27)	129	34 (26)	134	36 (27)	0.97	0.56–1.68	0.093
	Not known	44	10 (23)	19	3 (16)	25	7 (28)	0.98	0.75–1.27	0.344
Age at diagnosis (years)	<15	0	0 (0)	0	0 (0)	0	0 (0)	n/a	n/a	n/a
	15–24	314	92 (29)	233	68 (29)	81	24 (30)	0.98	0.56–1.70	0.940
	25–34	678	161 (24)	367	83 (23)	311	78 (25)	0.87	0.61–1.24	0.452
	>35	814	199 (25)	322	82 (26)	492	139 (28)	0.87	0.63–1.19	0.382
	Not known	347	114 (20)	194	62 (32)	153	52 (34)	0.91	0.58–1.43	0.690
Treatment status	Naive	1309	372 (28)	653	181 (28)	656	191 (29)	0.93	0.73–1.19	0.580
	Experienced	792	178 (22)	433	104 (24)	359	74 (21)	1.22	0.87–1.71	0.253
	Not known	52	16 (31)	30	10 (33)	22	6 (27)	1.33	0.40–4.46	0.640

The statistical significance of gender imbalance in each category was assessed by odds ratio calculation ($P < 0.05$). P values < 0.05 are indicated in bold. Categories with less than 100 individuals were collated into the 'Other' category. CI, confidence interval; IDU, injecting drug user.

Phylogenetic reconstruction

Sequences were aligned and manually edited using the programs ClustalX [16] and Se-Al v.2.0a11 (http://tree.bio.ed.ac.uk/software/seal/), respectively. Pairwise genetic distances were calculated for the entire dataset, using an in-house R script for uncorrected distance calculations, and sequences sharing at least 95.5% nucleotide similarity with at least one other sequence were selected for the study. An initial approximate maximum likelihood phylogeny of the selected sequences was built, under the General Time Reversible model of nucleotide substitutions and varying substitution rates across sites (GTR + CAT), with the software FastTree v2.1.5 [17]. Branch support was calculated by Shimodaira–Hasegawa-like local branch support (SH-like test), as implemented in FastTree.

Identification of transmission clusters

Putative transmission clusters were identified as follows. First, all phylogenetic clades of at least two sequences, with a maximum genetic distance of 4.5%, and a SH-like local branch support of at least 95% were extracted from the maximum likelihood phylogeny, using the Cluster Picker [18]. In order to control for selection-derived false-positives, the phylogeny of the putative clusters was reconstructed after removing from the alignment 38 codon positions associated with antiretroviral drug resistance [19] and considering third codon positions only. Subsequent analyses were restricted to those clusters that remained monophyletic in all cases.

The putative clusters were then pooled in alignments of about 150 sequences and confirmed by Bayesian Markov chain Monte-Carlo (MCMC) phylogenetic inference, as implemented in the package BEAST v1.7.4 [20]. Two independent runs of 50 000 000 generations, sampling every 1000th tree, were performed on each cluster pool. Divergence times, used as a surrogate of infection time, were estimated using an uncorrelated log-normal (UCLN) model of molecular evolutionary rate heterogeneity, a Bayesian skyline tree coalescent prior [21] and the SRD06 model of nucleotide substitution [22]. This combination of models was selected after testing several alternative models for each prior category on a random subset of the data. For each pairwise model comparison, a Bayes factor greater than 3 was deemed as a strong support for the favoured model [23]. A log-normal prior was set on the rate, with a mean value of 2.5×10^{-3} substitutions per sites per year, and a log SD of 0.1. Marginal posterior probabilities were plotted with Tracer v1.5 (http://tree.bio.ed.ac.uk/software/tracer/) in order to assess convergence of the model parameter values. An effective sample size (ESS) of at least 200 was considered a satisfactory convergence estimator. Maximum clade credibility trees (MCCTs) of the dated phylogenies were reconstructed using TreeAnnotator v1.7.0, available within the BEAST package (http://beast.bio.ed.ac.uk). Trees were edited with FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). Only clusters spanning a transmission period of up to 5 years in the MCCTs were selected for the statistical analyses. Binomial confidence intervals (CIs) were calculated with VassarStats (http://vassarstats.net/prop1.html).

Results

Heterosexual transmission of HIV-1 subtype B

A total of 13 699/22 481 (61%) HIV-1 subtype B sequences were linked to at least one other sequence in the database, forming 2860 putative transmission clusters (maximum intra-cluster genetic distance $\leq$4.5% and branch support $\geq$95%), a similar level to that estimated for an earlier database release [11]. The 39% of sequences with no match in the database suggest infections acquired abroad or from individuals whose UK partners were not diagnosed, were not included in the database or had too distant a connection to be identified.

Patients reporting heterosexual contact as their most likely route of infection represented 9% of all HIV-infected patients in clusters (1207/13 699) and 56% of all heterosexuals (1207/2153; Fig. 1). These patients were distributed across 671 clusters, 59% of which (399/671) formed transmission chains spanning less than 5 years (Bayesian posterior probability $\geq$0.95). For the remainder of the analysis, only these 399 heterosexual clusters spanning 5 years were taken into account, which involved 26% (566/2153) of the studied heterosexuals.

The characteristics of linked heterosexuals are shown in Table 1. Linked heterosexuals did not differ from unlinked heterosexuals with respect to ethnicity, age at diagnosis and treatment status. They were predominantly treatment-naïve at the time of sampling, white or black Caribbean of both sexes. The sex distribution was evenly split, with a male-to-female ratio of 0.92 (271/566) (Fig. 1). Over half of these heterosexuals (293/566; 52%)

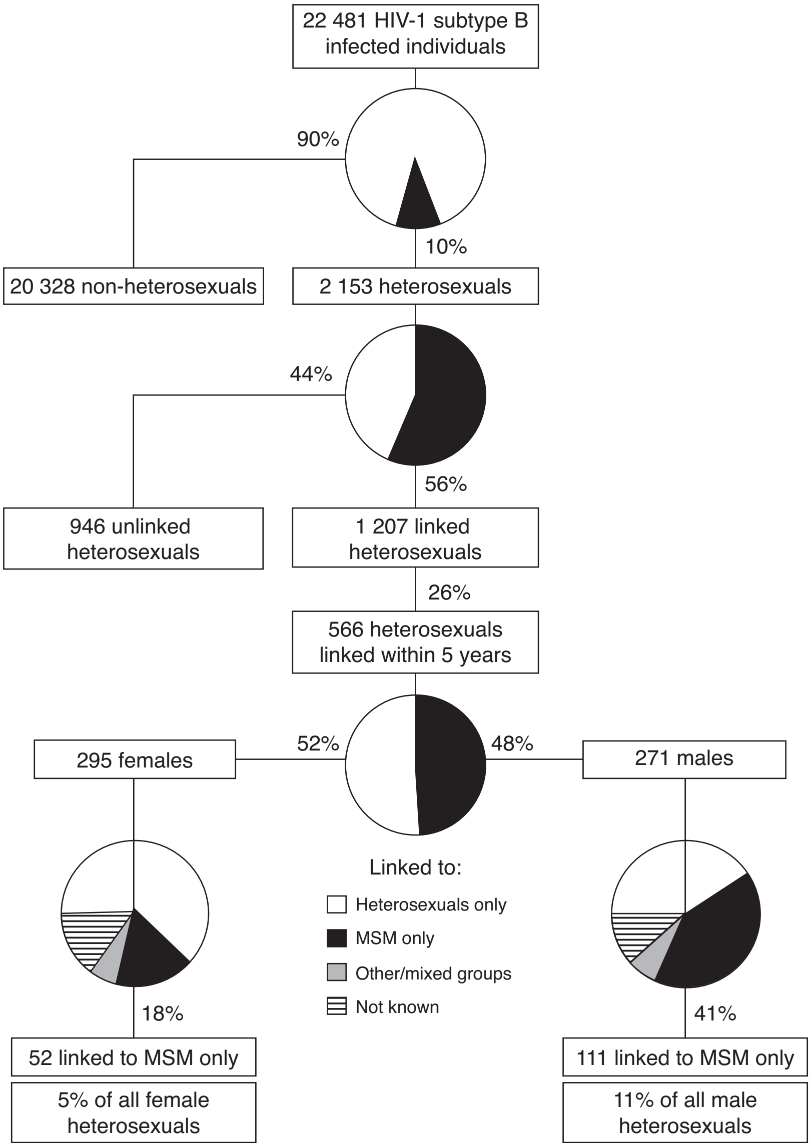

Fig. 1. Linkage among reported heterosexuals in the UK HIV Drug Resistance Database.

were found in clusters involving only other heterosexuals (see, e.g. Fig. 2a), 62% (182/293) being female.

Potential non-disclosure of homosexual contact among reported heterosexuals

A third of the linked heterosexuals (173/566; 31%) belonged to a cluster that included both heterosexuals and MSM (e.g. Fig. 2b), and 29% (163/566) were solitary heterosexuals in a cluster otherwise exclusively comprised of MSM (e.g. Fig. 2c). The 173 heterosexuals linked to at least one MSM formed 167 independent transmission clusters, ranging in size from 2 (91/167 for the clusters) to 13 (one cluster of 1 male heterosexual linked to 12 MSM).

The characteristics of the heterosexuals exclusively linked to MSM within 5 years are given in Table 2. Heterosexuals linked only to MSM were more likely to be men (111/163) than women (52/163) [odds ratio (OR) 0.41, 95% CI 0.29–0.57, $P < 0.001$], and represented 11 and 5% of the studied male and female heterosexuals, respectively (Figs 1 and 2d). Assuming that the proportion of female heterosexuals solely linked to MSM represents the level of 'disassortative' mixing expected in the UK cohort, the difference between the male and female heterosexuals linked only to MSM (i.e. 6%) is likely to reflect the proportion of misreported MSM infections occurring at diagnosis.

Demographics of non-disclosed MSM

A substantial majority of the individuals found in heterosexual/MSM clusters linked to persons from the same geographical area, with infections diagnosed in the London area accounting for over 50% of the infections in these clusters (data not shown). This is in agreement with national reports on HIV diagnoses in the country (see for instance the 2013 Public Health England report on HIV in the United Kingdom: http://www.hpa.org.uk/).

Half of the men identified as heterosexuals linked solely to other men were white (83/163, 51%; 95% CI 43–58%), reflecting the over-representation of that ethnicity in the cohort. However, within each ethnic group, the highest level of linkage with MSM was found amongst male heterosexuals of black African origin: 21% of the patients in that category (23/110; 95% CI 14–29%) were exclusively linked to MSM. By contrast, only 10% (59/619; 95% CI 7–12%) of the white men reporting heterosexual contact as the route of infection were linked to only MSM (Table 2). In addition, the proportion of individuals reporting an unknown route of infection was significantly greater for black African men (8/259, 3%) compared to white men (182/20 997, 0.9%; $P = 0.003$, Fisher's exact test). No significant difference was seen between black Caribbean and white men ($P = 0.111$).

By comparison, 24 of 52 female heterosexuals linked to only MSM were white (46%; 95% CI 33–59%) and

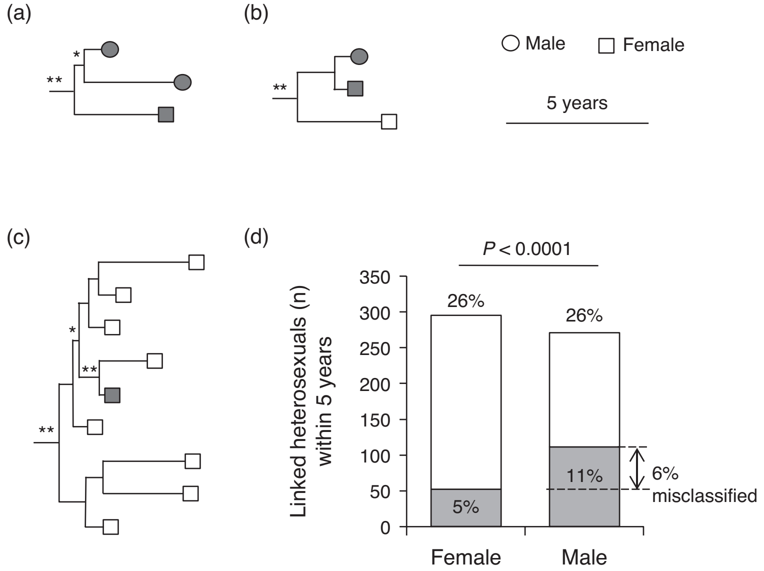

Fig. 2. Phylogenetic clusters representative of heterosexual transmission patterns. (a) Transmission chain involving heterosexuals only. (b) Transmission chain involving both heterosexuals and MSM. (c) Transmission chain involving a single heterosexual linked to MSM only. Male and female individuals are represented by squares and circles, respectively. MSM and heterosexuals were indicated in white and grey, respectively. Branch lengths express unit of time in calendar-years, as indicated by the scale. Branches with a local support of at least 90% and equal to 100% are labelled with one or two asterisks, respectively. (d) Heterosexuals linked to at least one other individual in the cohort within 5 years, per sex. The proportion of heterosexuals linked to MSM only is coloured in grey. The excess of male-to-female heterosexuals in the latter category, representing misclassified MSM infections, is indicated by the dashed lines. Statistical significance of the sex imbalance was assessed by Fisher's exact test.

Table 2. Characteristics of heterosexuals clustered with only MSM within 5years or less.

		All heterosexuals		Female		Male				
		Total	Clustered (%)	Total	Clustered (%)	Total	Clustered (%)	Odds ratio	95% CI	*P*
Total *N* (%)		2153	163 (8)	1116	52 (5)	1037	111 (11)	0.41	0.29–0.57	**<0.001**
Ethnicity	Black African	230	42 (18)	120	19 (15)	110	23 (21)	0.71	0.36–1.39	0.321
	Black Caribbean	416	11 (3)	267	5 (2)	149	6 (4)	0.45	0.13–1.51	0.200
	White	1200	83 (7)	581	24 (4)	619	59 (10)	0.41	0.25–0.67	**<0.001**
	Other	263	23 (9)	129	4 (3)	134	19 (14)	0.19	0.06–0.59	**0.003**
	Not known	44	4 (9)	19	0 (0)	25	4 (16)	1.12	0.01–2.42	0.168
Age at diagnosis (years)	<15	0	0 (0)	0	0 (0)	0	0 (0)	n/a	n/a	n/a
	15–24	314	22 (7)	233	8 (3)	81	14 (17)	0.17	0.07–0.42	**<0.001**
	25–34	678	55 (8)	367	15 (4)	311	40 (12)	0.28	0.15–0.53	**<0.001**
	>35	814	61 (7)	322	19 (6)	492	42 (9)	0.67	0.38–1.18	0.164
	Not known	347	25 (7)	194	10 (5)	153	15 (10)	0.50	0.22–1.15	0.102
Treatment status	Naive	1309	122 (9)	653	36 (6)	656	86 (13)	0.39	0.26–0.58	**<0.001**
	Experienced	792	39 (5)	433	15 (3)	359	24 (7)	0.59	0.26–0.97	**0.040**
	Not known	52	2 (4)	30	1 (3)	22	1 (5)	0.72	0.04–12.2	0.823

The statistical significance of gender imbalance in each category was assessed by odds ratio calculation ($P < 0.05$). *P* values < 0.05 are indicated in bold. Categories with less than 100 individuals were collated into the 'Other' category. CI, confidence interval.

19/52 were of black African (37%; 95% CI 25–50%) ethnicity. The marital status of these female heterosexuals linked to MSM was not known. These HIV-positive women most likely represent sporadic infections through sex with either unsampled heterosexual MSM or self-reported bi-sexual men (classified as MSM).

Discussion

We demonstrate the epidemiological utility of phylogenetic inference when supplementing surveillance methods to monitor the HIV epidemic. We provide evidence of misclassified MSM HIV infections together with an estimate of the number of HIV-positive men who may not have disclosed homosexual contact at the time of diagnosis in the UK. In our cohort, incorrectly identified infection routes amount to 6% of the male heterosexuals infected with HIV subtype B in the UK (i.e. 11% minus the 5% representing the expected disassortative mixing seen in female heterosexuals). If we assume that the same level of misreporting occurs in HIV-1 subtype B and non-B infections, we extrapolate that at least 6% of all infections in male heterosexuals are misclassified MSM infections. On the contrary, if we assume that misreporting in non-B infections is close to null, and since HIV-1 subtype B represents 19.8% of all infections found in heterosexual men in the UK [3], about 1% of all infections amongst men reported as heterosexually acquired were acquired homosexually. This would amount to 142–852 of the 14 200 heterosexual men living with diagnosed HIV in the UK by 2012 [1]. This low figure largely validates the reliability of sexual exposure information collected by routine surveillance.

In addition to heterosexuals exclusively linked to MSM, 54 male heterosexuals were linked to only other male heterosexuals. Whereas these could reflect transmission chains with unsampled female partners, they could, as an upper estimate, be comprised entirely of MSM misclassified as heterosexuals. Adding all these 54 male heterosexuals to the 111 ones linked solely to MSM, the proportion of male heterosexuals linked to only other men increases to 16% (165/1037) and the proportion of incorrectly identified infection rises to 11% (i.e. 16% minus 5% of expected disassortative mixing seen in female heterosexuals).

Our estimates are likely to be conservative. Firstly, our survey focused on subtype B, the most prevalent subtype circulating in the UK, and did not investigate infections with other subtypes. Subtype C infections, for instance, are also highly prevalent in the UK (i.e. 34.3% of all HIV diagnoses made between 2002 and 2010) and mainly associated with heterosexual contact. Only 12.2% of these infections are reported in MSM [3], although our data for subtype B infections suggest that this proportion may be an underestimate. Secondly, opting for a conservative approach, we have only considered those transmission clusters linking a single reported heterosexual to one or more MSM. By doing so, we have excluded clusters with more than one heterosexual infection linked to MSM, where the directionality of transmission could not be unambiguously established. The example shown in Fig. 1b shows a heterosexual male and female linked to a MSM. In this case, we cannot determine whether the male heterosexual was infected by a MSM prior to infecting a female partner (which would count as a misclassified MSM transmission), or if the female heterosexual was infected by a MSM first, then in turn infected a male partner. In the latter scenario, the reported heterosexual infection risk would be genuine. Thirdly, it has been estimated that 67% of heterosexual adults born abroad and diagnosed with HIV in the UK acquired their infection outside the UK [24]. Viral sequences obtained from these patients would have no match in the database, and the most likely route of acquisition for these infections could not be estimated with the methodology used in this study.

In our cohort, both exclusive linkage with MSM and unreported route of infection were significantly greater for black African heterosexual men compared with men of other ethnicities. When extrapolated to the total number of black African heterosexual men living with diagnosed HIV in the UK, our estimates suggest that between 1% (no misreporting in non-B viruses, with subtype B infections accounting for 2.8% of the HIV diagnoses made amongst black African men [3]) and 21% (same level of misreporting in B and non-B viruses) of black African men reported as heterosexuals most likely acquired HIV through sex with other men. Moreover, if the seven male heterosexuals of Black African ethnicity linked to only other male heterosexuals are added to these figures, the number of potentially misclassified infections reached 26% of all Black African HIV-positive male heterosexuals in the study group. These estimates amount to 75–1575/7500 men [1]. Such observations support the notion that black African men are less likely to disclose sex between men as a route of potential exposure compared to other ethnic groups. Factors hampering disclosure of same-sex sexuality commonly include social–cultural barriers and experiences of discrimination. This seems to be particularly true for MSM of black or minority ethnicity. Africans testing for HIV at a London hospital, for instance, were twice as likely as white patients to be concerned about future discrimination if they tested positive, and four times more likely to be worried about meeting someone they knew at the clinic [25]. Behavioural studies have also shown that MSM from a black African background were more likely to have sexual intercourse with a woman than white MSM [26,27]. These trends are in line with our estimates. Taken together, our findings indicate that MSM unwilling to disclose their route of infection are more likely to be found amongst black African male heterosexuals than any other group. This is of importance since MSM of black or minority ethnicity are of highest risk of acquiring HIV in the UK, and MSM perceived as heterosexuals may be missed by targeted prevention programmes. Apart from MSM being under social–cultural constraints, male heterosexuals who have sex with men, but are not identified as 'gay', are also likely to misreport their potential exposure. Discordance between reported sexual behaviour and sexual identity has been reported before, accounting for 12% of the men interviewed in a recent US study [28].

The sequences used for this study represent at least 46% of cumulative HIV infections in the UK since 1996. Despite such dense sampling of the UK HIV-positive population, a proportion of the identified clusters will be incomplete. Clearly, in common with other convenience cohorts, incomplete data pose the potential for sampling bias that might in principle explain the excess of men seen amongst heterosexuals linked only to MSM. However, we have observed a large proportion of heterosexual clusters comprising only women (42% of the heterosexual clusters, data not shown). Female-to-female transmission clusters are likely to reflect a sampling bias against HIV-transmitting male heterosexuals. Late HIV diagnoses (i.e. a $CD4^{+}$ cell count $<350/\mu l$ within 3 months of diagnosis) may contribute to this sampling bias, as it is highest amongst heterosexual men in the UK [1]. Heterosexual HIV transmissions involving undiagnosed men and diagnosed women could result in the phylogenetic pattern seen in the female-to-female clusters. Male-to-male heterosexual clusters, on the contrary, were less frequent (20% of the heterosexual clusters, data not shown), suggesting that under-sampling of female individuals cannot explain the trends observed. Quantifying the proportion of incomplete clusters is difficult with the adopted methodology. Although others have attempted to quantify and assess the impact of missing transmitters in phylogenetic clusters, these methods were not applicable to our cohort at the time of this study [29]. Further work will concentrate on developing an appropriate framework for this task.

In order to control as much as possible for confounding factors linked to unsampled transmitting individuals, our estimates were based on the analysis of transmission clusters occurring within a 5-year time span. An excess of male-to-female heterosexuals exclusively linked to MSM was also observed in the larger set of identified transmission clusters, that is, those identified on the basis of intra-cluster mean genetic distance at of least 4.5% only (see Methods section). This proportion amounted to 8% of the men reported heterosexuals infected with subtype B, and was remarkably consistent with that of the 'time-spanned' approach.

We have shown that phylogenetic analyses coupled with epidemiological data can identify and quantify nondisclosure of homosexual contact among heterosexual men. Although phylogenetic inference is increasingly present in work focusing on prevention strategies (e.g. [30]), it remains to establish itself as a tool for routine HIV surveillance. To date, the monitoring of the HIV epidemic at national levels, as well as the development of prevention strategies, solely relies on information gathered from patients and/or clinicians reports. In the present study, we have shown how the integration of molecular data to epidemiological, clinical and demographic information can contribute to the identification of groups particularly vulnerable to HIV. Accurate risk factor information is necessary to inform and evaluate the public health response to the epidemic. Phylogenetic approaches can provide an adjustment factor of this information (as illustrated here) and help infer missing social–demographic profiles when coupled with probabilistic models in the convenience cohorts. This framework can be applied to local HIV epidemics where comprehensive molecular surveillance is conducted through routine resistance testing at diagnosis or prior to treatment.

Acknowledgements

The UK Collaborative Group on HIV Drug Resistance is a collaboration between the UK HIV Drug Resistance Database; UK CHIC; Public Health England HARS and participating academic centres, clinics and laboratories.

Steering Committee: Celia Aitken, Gartnavel General Hospital, Glasgow; David Asboe, Anton Pozniak, Chelsea & Westminster Hospital, London; Daniel Webster, Royal Free NHS Trust, London; Patricia Cane, Public Health England, Porton Down; Hannah Castro, David Dunn, David Dolling, Esther Fearnhill, Kholoud Porter, MRC Clinical Trials Unit, London; David Chadwick, South Tees Hospitals NHS Trust, Middlesbrough; Duncan Churchill, Brighton and Sussex University Hospitals NHS Trust; Duncan Clark, St Bartholomew's and The London NHS Trust; Simon Collins, HIV i-Base, London; Valerie Delpech, Public Health England, London; Anna Maria Geretti, University of Liverpool; David Goldberg, Health Protection Scotland, Glasgow; Antony Hale, Leeds Teaching Hospitals NHS Trust; Stéphane Hué, University College London; Steve Kaye, Imperial College London; Paul Kellam, Wellcome Trust Sanger Institute & UCL Medical School; Linda Lazarus, Expert Advisory Group on AIDS Secretariat, Public Health England, London; Andrew Leigh-Brown, University of Edinburgh; Nicola Mackie, Imperial NHS Trust; Chloe Orkin, St. Bartholomew's Hospital, London; Philip Rice, St George's Healthcare Trust, London; Deenan Pillay, Andrew Phillips, Caroline Sabin, University College London Medical School; Erasmus Smit, Public Health England, Birmingham Heartlands Hospital; Kate Templeton, Royal Infirmary of Edinburgh; Peter Tilston, Manchester Royal Infirmary; William Tong, Guy's and St. Thomas' NHS Foundation Trust, London; Ian Williams, Mortimer Market Centre, London; Hongyi Zhang, Addenbrooke's Hospital, Cambridge; Mark Zuckerman, King's College Hospital, London.

Centres contributing data: Clinical Microbiology and Public Health Laboratory, Addenbrooke's Hospital, Cambridge (Jane Greatorex); HIV/GUM Research Laboratory, Chelsea and Westminster Hospital, London (Adrian Wildfire); Guy's and St. Thomas' NHS Foundation Trust, London (Siobhan O'Shea, Jane Mullen); Public Health England – Public Health Laboratory, Birmingham Heartlands Hospital, Birmingham (Erasmus Smit); Public Health England London (Tamyo Mbisa); Imperial College Health NHS Trust, London (Alison Cox); King's College Hospital, London (Richard Tandy); Medical Microbiology Laboratory, Leeds Teaching Hospitals NHS Trust (Tony Hale, Tracy Fawcett); Specialist Virology Centre, Liverpool (Mark Hopkins, Lynn Ashton); Department of Clinical Virology, Manchester Royal Infirmary, Manchester (Peter Tilston); Department of Virology, Royal Free Hospital, London (Daniel Webster, Ana Garcia-Diaz); Edinburgh Specialist Virology Centre, Royal Infirmary of Edinburgh (Jill Shepherd); Department of Infection & Tropical Medicine, Royal Victoria Infirmary, Newcastle (Matthias L. Schmid, Brendan Payne); South Tees Hospitals NHS Trust, Middlesbrough (David Chadwick); St George's Hospital, London (Phillip Hay, Phillip Rice, Mary Paynter); Department of Virology, St Bartholomew's and The London NHS Trust (Duncan Clark, David Bibby); Molecular Diagnostic Unit, Imperial College, London (Steve Kaye); University College London Hospitals (Stuart Kirk); West of Scotland Specialist Virology Lab Gartnavel, Glasgow (Alasdair MacLean, Celia Aitken, Rory Gunson).

Coordinating Centre: Medical Research Council Clinical Trials Unit (MRC CTU), London (Kate Coughlin, David Dolling, David Dunn, Esther Fearnhill, Kholoud Porter).

Funding: The UK HIV Drug Resistance Database is supported by the Medical Research Council (grant number G0900274), and is partly funded by the Department of Health. This project was also funded by the University College London Hospitals Biomedical Research Centre and the Collaborative HIV and Anti-HIV Drug Resistance Network (CHAIN), by the Wellcome Trust (grant number 092807) and a studentship from the Biotechnology and Biological Science Research Council (MR-C). Additional support for the HIVRDB is provided by Boehringer Ingelheim, Bristol-Myers Squibb, Gilead, Tibotec (a division of Janssen-Cilag) and Roche.

Conflicts of interest

The views expressed in the publication are those of the authors and not necessarily those of the Department of Health.

References

1. Public Health England. HIV in the United Kingdom: 2012 Report. 2013. http://www.hpa.org.uk/Publications/InfectiousDiseases/HIVAndSTIs/1311HIVintheUk2013report/.
2. Phillips AN, Cambiano V, Nakagawa F, Brown AE, Lampe F, Rodger A, *et al*. **Increased HIV incidence in men who have sex with men despite high levels of ART-induced viral suppression: analysis of an extensively documented epidemic.** *PLoS One* 2013; **8**:e55312.

3. The UKCGoHIVDR. **The increasing genetic diversity of HIV-1 in the UK, 2002–2010.** *AIDS* 2014; **28**:773–780.
4. Robbins KE, Lemey P, Pybus OG, Jaffe HW, Youngpairoj AS, Brown TM, *et al.* **U. S. Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains.** *J Virol* 2003; **77**:6359–6366.
5. Hue S, Pillay D, Clewley JP, Pybus OG. **Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups.** *Proc Natl Acad Sci U S A* 2005; **102**:4425–4429.
6. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ. **Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom.** *PLoS Pathog* 2009; **5**: e1000590.
7. Dennis AM, Hue S, Hurt CB, Napravnik S, Sebastian J, Pillay D, *et al.* **Phylogenetic insights into regional HIV transmission.** *AIDS* 2012; **26**:1813–1822.
8. de Oliveira T, Pybus OG, Rambaut A, Salemi M, Cassol S, Ciccozzi M, *et al.* **Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak.** *Nature* 2006; **444**:836–837.
9. Leitner T, Kumar S, Albert J. **Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history.** *J Virol* 1997; **71**:4761–4770.
10. Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT, *et al.* **Molecular epidemiology of HIV transmission in a dental practice.** *Science* 1992; **256**:1165–1171.
11. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. **Episodic sexual transmission of HIV revealed by molecular phylodynamics.** *PLoS Med* 2008; **5**:e50.
12. de Silva TI, van Tienen C, Onyango C, Jabang A, Vincent T, Loeff MF, *et al.* **Population dynamics of HIV-2 in rural West Africa: comparison with HIV-1 and ongoing transmission at the heart of the epidemic.** *AIDS* 2013; **27**:125–134.
13. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT. **Transmission network parameters estimated from HIV sequences for a nationwide epidemic.** *J Infect Dis* 2011; **204**:1463–1469.
14. Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AF, Hughes G, *et al.* **An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1.** *PLoS Comput Biol* 2009; **5**:e1000581.
15. UK Collaborative HIV Cohort Steering Committee. **The creation of a large UK-based multicentre cohort of HIV-infected individuals: The UK Collaborative HIV Cohort (UK CHIC) Study.** *HIV Med* 2004; **5**:115–124.
16. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997; **25**:4876–4882.
17. Price MN, Dehal PS, Arkin AP. **FastTree 2: approximately maximum-likelihood trees for large alignments.** *PLoS One* 2010; **5**:e9490.
18. Ragonnet-Cronin M, Hodcroft E, Hue S, Fearnhill E, Delpech V, Brown AJ, *et al.* **Automated analysis of phylogenetic clusters.** *BMC Bioinformatics* 2013; **14**:317.
19. Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer RW, *et al.* **Update of the drug resistance mutations in HIV-1: March 2013.** *Top Antivir Med* 2013; **21**:6–14.
20. Drummond AJ, Suchard MA, Xie D, Rambaut A. **Bayesian phylogenetics with BEAUti and the BEAST 1.7.** *Mol Biol Evol* 2012; **29**:1969–1973.
21. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. **Bayesian coalescent inference of past population dynamics from molecular sequences.** *Mol Biol Evol* 2005; **22**:1185–1192.
22. Shapiro B, Rambaut A, Drummond AJ. **Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences.** *Mol Biol Evol* 2006; **23**:7–9.
23. Kass RE, Raftery AE. **Bayes factors.** *J Am Stat Assoc* 1995; **90**:773–795.
24. Rice BD, Elford J, Yin Z, Delpech VC. **A new method to assign country of HIV infection among heterosexuals born abroad and diagnosed with HIV.** *AIDS* 2012; **26**:1961–1966.
25. Erwin J, Peters B. **Treatment issues for HIV+ Africans in London.** *Soc Sci Med* 1999; **49**:1519–1528.
26. Prost A, Elford J, Imrie J, Petticrew M, Hart GJ. **Social, behavioural, and intervention research among people of Sub-Saharan African origin living with HIV in the UK and Europe: literature review and recommendations for intervention.** *AIDS Behav* 2008; **12**:170–194.
27. Soni S, Bond K, Fox E, Grieve AP, Sethi G. **Black and minority ethnic men who have sex with men: a London genitourinary medicine clinic experience.** *Int J STD AIDS* 2008; **19**:617–619.
28. Pathela P, Hajat A, Schillinger J, Blank S, Sell R, Mostashari F. **Discordance between sexual behavior and self-reported sexual identity: a population-based survey of New York City men.** *Ann Intern Med* 2006; **145**:416–425.
29. Volz EM, Frost SD. **Inferring the source of transmission with phylogenetic data.** *PLoS Comput Biol* 2013; **9**:e1003397.
30. Eshleman SH, Hudelson SE, Redd AD, Wang L, Debes R, Chen YQ, *et al.* **Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial.** *J Infect Dis* 2011; **204**:1918–1926.