Nondisclosed MSM in UK HIV transmission networks: phylogenetic analysis of surveillance data

Manon Ragonnet-Cronin, Stéphane Hué, Emma B. Hodcroft, Anna Tostevin, David Dunn, Tracy Fawcett, Anton Pozniak, Alison E. Brown, Valerie Delpech and Andrew J. Leigh Brown on behalf of the UK HIV Drug Resistance Database

Addresses

Manon Ragonnet-Cronin, PhD
University of California, San Diego, Department of Medicine
220 Dickinson street, San Diego, CA 92103, USA

Stéphane Hué, PhD
London School of Hygiene & Tropical medicine, Department of Infectious Disease Epidemiology
Keppel Street, London WC1E 7HT, UK

Emma B. Hodcroft, PhD
University of Edinburgh, Institute of Evolutionary Biology,
Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK

Anna Tostevin, MSc
University College London, Institute for Global Health
Mortimer Market Centre, Capper Street, London WC1E 6JB, UK

Professor David Dunn, PhD
University College London, Institute for Global Health
Mortimer Market Centre, Capper Street, London WC1E 6JB, UK

Tracy Fawcett, MSc
Virology, Old Medical School, Leeds General Infirmary, Leeds LS1 3EX, UK

Anton Pozniak, MD
Chelsea and Westminster Hospital, 369 Fulham Road, London, SW10 9NH, UK

Alison E. Brown, PhD
Public Health England, Wellington House, 133-155 Waterloo Road, London SE1 8UG, UK

Valerie Delpech, PhD
Public Health England, Wellington House, 133-155 Waterloo Road, London SE1 8UG, UK

Professor Andrew J. Leigh Brown, PhD
University of Edinburgh, Institute of Evolutionary Biology,
Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK
Research in context

Evidence before this study
A search on PubMed with the terms “genetic”, “cluster”, “HIV”, “epidemiology” and “network” in January 2017 returned 25 articles. We focused our attention on analyses of densely sampled epidemics in Europe and North America. The present UK epidemic among men who have sex with men (MSM) arose through the introduction of subtype B viruses in the early 1980s. In contrast, in heterosexuals, the majority of viruses are non-B subtypes. Within MSM transmission networks, there is a subgroup of men who self-report as heterosexual but whose HIV genetic sequences link only to MSM. This group is very poorly understood: they are inaccessible to classical epidemiological approaches and were only revealed in the UK recently, in a single molecular epidemiology study.

Added value of this study
We focused on this group of self-reported heterosexual men who link only to MSM adopting a novel network based method which allowed us investigate transmission across multiple subtypes. Analysing viral sequences from over 50,000 individuals, we identified men who self-report as heterosexual in MSM HIV transmission networks (potential nondisclosed MSM), examined their position and compared them to the MSM with whom they clustered. We found that these men tended to link to each other and were not involved in rapid chains of MSM transmission. They were more likely to be non-White than the MSM they clustered with. We found more examples of heterosexual men linking heterosexuals and MSM than heterosexual women in that position, although the total number of such chains was small.

Implications of all the available evidence
The tendency of nondisclosed MSM in our network to link to each other indicates that they may preferentially seek out other nondisclosed MSM partners for sex. This result aligns with conclusions from small scale interview-based studies, which found that they attend different venues to MSM. The current study is the first to characterise the behaviour of nondisclosed MSM at a national scale. Our results suggest that nondisclosed MSM may be underestimating their risk and should be encouraged to test more frequently with better targeted prevention messages. However, it should be emphasised that the extent of bridging between the heterosexual and MSM epidemics appears to be limited.

Summary
Background
Patients who do not disclose their sexuality, for example men who do not disclose same-sex behaviour, are extremely difficult to characterise through traditional epidemiological approaches. Using a recently developed method to detect large networks of viral sequences from time-resolved trees, we have localised these men in UK transmission networks, gaining critical insights into the behaviour of this group.

Methods
We obtained HIV pol sequences from the UK HIV Drug Resistance Database (UKRDB), a central repository for resistance tests performed as part of routine clinical care throughout the UK. Sequence data are linked to demographic and clinical data held by the UK Collaborative HIV Cohort study and the national HIV/AIDS Reporting System database. Initially we reconstructed maximum likelihood phylogenies from sequences from over 50,000 individuals in the UK. Sequences were selected for time-resolved analysis in BEAST if they were clustered with at least one other sequence at a genetic

2
distance ≤4.5% with support ≥ 90%. We used time-resolved phylogenies to create networks by linking together nodes if sequences shared a common ancestor within the previous 5 years. We identified potential nondisclosed MSM (pnMSM) as self-reported heterosexual men who clustered only with men. We measured the network position of pnMSM, including betweenness (a measure of connectedness and importance) and assortativity (the propensity for nodes sharing attributes to link).

Findings
14,405 individuals were represented in the network, including 8,452 MSM, 1743 heterosexual women and 1341 male heterosexuals. 249 pnMSM were identified (18.6% of all clustered heterosexual men). pnMSM were more likely to be Black-African (p<0.0001) and were slightly older (39.00 vs. 36.38 p=0.002) than the MSM they clustered with. Betweenness centrality was lower for pnMSM than for MSM (1.31 vs 2.24, p=0.002), indicating that they were in peripheral positions in MSM clusters. Assortativity by risk group was higher than expected (0.037 vs - 0.037, p=0.01) signifying that pnMSM linked to each other. We found that self-reported male heterosexuals were more likely than heterosexual women to link MSM and heterosexuals (Fisher’s exact test; p=0.0004; OR 2.24) but the number of such transmission chains was small (only 54 in total compared to 32 in women).

Interpretation
We have shown that pnMSM are a subgroup distinct from both MSM and from male heterosexuals. Male heterosexuals are the group most likely to be diagnosed in late stage disease, and nondisclosed MSM may put female partners at higher risk. Thus, pnMSM require specific consideration to ensure they are included in public health interventions.

Funding
NIH GM110749 (MIDAS).

Introduction
Men who have sex with men (MSM) are the group at highest risk for HIV infection, with higher HIV prevalence than the general population in every world region (1). In the UK, MSM account for half of those living with HIV despite increased testing and uptake of treatment (2).

Among MSM, those who do not disclose their same-sex behaviour for fear of stigmatisation, rejection or prejudice (3) are more likely to exhibit risky sexual behaviour, for example not using condoms (4), and are more likely to have sex with women (5). Insights into the behaviour of nondisclosed MSM are difficult to obtain but central to our understanding of HIV transmission. Phylogenetic analysis of viral sequences has provided a valuable route to elucidating transmission dynamics (6). Phylogenies can be combined with epidemiological, demographic and behavioural data to elucidate spread of HIV through time, geographical regions and different risk groups (7). Phylogenies have been used to investigate the role of age, disease stage, viral load and treatment on transmission (8).

In the UK, analysis of subtype B viruses demonstrated the existence of six independent transmission chains among MSM following separate introductions in the 1980s (9). MSM were much more likely than heterosexuals to fall within groups of closely linked infections, or clusters (6, 10). Within clusters, 25% of transmissions happen within six months of infection among MSM (6), compared to 2% in heterosexuals (10). In a recent analysis we used phylogenetics to quantify the extent of nondisclosure of sex with men among heterosexual men (11). Eleven percent of heterosexual men
clustered only with MSM compared to only 5% of heterosexual women. This discrepancy suggests around 6% (1-11%) of men who report only heterosexual behaviour became infected through sex with men. The difference was most significant among Black African men, where up to 21% may have been infected by men.

The structure of contact networks provides information on the process that led to their formation (12). The transmission network is a subset of the contact network whose structure is shaped by the underlying contact network and by the transmission characteristics of the pathogen (13). Thus, an isolated node in an HIV transmission network may represent an individual who had no sexual contacts after infection or one who had many sexual contacts but never transmitted the virus. Nonetheless, aggregate differences in network characteristics between groups can reveal differences in behaviour: MSM clusters are larger than heterosexual clusters (6, 10), indicating on average more sexual partners and more onward transmission events.

Here we have further exploited our network reconstruction methodology (14) in order to describe the network positions of nondisclosed MSM for the first time. We hypothesised that their position within network clusters would differ from that of the MSM with whom they cluster, revealing differences in sexual or transmission behaviour.

Methods

1. Study population
We obtained 63,065 aligned HIV pol sequences, each representing a distinct HIV positive individual, from the UK HIV Drug Resistance Database (UKRDB; http://www.hivrdb.org; sequences up to mid-2013). The UKRDB was established in 2001 as a central repository for resistance tests performed as part of routine clinical care throughout the UK. Clinical centres that contribute to the UKRDB are listed in the Appendix, p5. Sequences covering protease and 900 bases of reverse transcriptase are received from laboratories. Sequences are aligned by the database managers at UKRDB through the Stanford University Genotypic Resistance Interpretation Algorithm (hivdb.stanford.edu). Sequence data are linked to demographic and clinical patient data held by the UK Collaborative HIV Cohort study (15) and the national HIV/AIDS Reporting System database held at Public Health England (16). After linking sequences to age, sex, risk group, ethnicity, avidity score and stage of infection, the data were anonymised. Infections are classified as recent (<155 days) if their avidity is <80%. UKRDB data collection and processing is explained on the UKRDB website (http://www.hivrdb.org.uk/data-collection-and-processing). We assigned subtypes using SCUEAL (17). The three most common subtypes, B (49.5%), C (24.9%), and A1 (4.5%), were analysed, comprising 50,025 individuals diagnosed between 1997 and 2013, for each of whom we used the earliest available sequence. The London Multicentre Research Ethics Committee gave ethical approval for non-consented studies of anonymised data (MREC/01/2/10).

In parallel, we retrieved over 130,000 pol (HXB2 positions 2253-3870, minimum length 900bp) sequences of HIV subtypes A1, B, and C, from the Los Alamos National Laboratory (LANL) database (January 2014). To limit the size of alignments, we selected the ten closest non-UK sequences to each UK sequence. All sequences were stripped of 44 sites associated with drug resistance (18).

2. Phylogenetic analysis
For each subtype, we reconstructed phylogenies in RaxMLv8.2 with 100 bootstrap replicates (19). We considered 100 bootstraps to be sufficient because the maximum likelihood phylogeny was the first step towards deciding whether to include sequences in a time-resolved analysis. We identified clusters in phylogenies using Cluster Picker v1.2.3 if they were supported by a bootstrap ≥90% and
had a maximum genetic distance ≤4.5% \((6)\), with gaps ignored in pairwise comparisons and ambiguities counted as mismatches. The phylogenetic relationships between clustered sequences were then time-resolved in BEASTv1.8 \((20)\).

We sorted previously identified clusters into pools of up to 300 sequences so that all sequences from the same cluster were analysed together \((14)\), but limiting the size of datasets. Seventy-three pools were created altogether. For each pool, we generated summary statistics, including the number of sequences, the number of (RaxML-identified) clusters and the breadth of years covered (the difference between the oldest and the youngest sequence). We tested settings in two pools each of subtype B and subtype C sequences with maximum breadth of sequence dates \((12\) years). We compared clock (strict and relaxed) and demographic models (constant, exponential, logistic, Bayesian skyline and Bayesian skyride) in the test pools. We used an SRD06 substitution model as codon models give the best fit to HIV \(\text{pol}\) sequence data \((21)\). A normally distributed molecular clock rate prior with a mean of \(2.55 \times 10^{-3}\) was used \((9)\). The best model was selected as the skyride with an uncorrelated lognormal clock by means of its Bayes Factor (the ratio of the marginal likelihoods of two models), as estimated through path sampling. Chains were run for 100,000,000 generations in duplicate, checked for convergence in Tracer v1.6 and combined. We generated maximum clade credibility trees in Tree Annotator v1.8.4.

We reconstructed networks from phylogenies as previously described \((14)\) using R scripts (available at github.com/hxnx-sam/Rcode_utils/). In brief, each sequence is represented by a node in the network and nodes are linked together (clustered) if their time to most recent common ancestor is ≤5 years in the time-resolved maximum clade credibility tree. In contrast to clusters defined phylogenetically, nodes in network clusters are not all connected to each other, and clusters can contain sequences which are not monophyletic in the tree. Different depths of network reconstruction were evaluated in a previous analysis \((14)\), and here we selected five years as a cutoff for consistency with the National Survey of Sexual Attitudes and Lifestyles (NATSAL) \((22)\).

### 3. Procedures

We constructed a multivariate logistic regression model with the outcome being either clustered in the network \((1)\) or not \((0)\). Subtype, sex, risk group and ethnicity were included as predictor variables, and we included terms for the interactions between variables.

We conducted network analysis using the \(\text{igraph}\) v1.1 \((23)\) and \(\text{network}\) v1.13 \((24)\) packages in R v3.4.1. Networks for all subtypes were amalgamated and nodes were linked to demographic data. While the transmission dynamics of B and non-B subtypes are undoubtedly different in the UK \((6,10)\), subtype was maintained as a variable associated with each node so that results could be interpreted by subtype.

We classified self-reported heterosexual men who clustered only with MSM and with no women as potential nondisclosed MSM (pnMSM; see results) and we compared their positions within networks to that of the MSM with whom they clustered based on their centrality and linkage (Figure 1).

Centrality measures \((\text{degree}, \text{betweenness centrality and eigenvector centrality})\) quantify how important a node is in keeping a network connected. Degree is the number of edges connected to a node. Betweenness centrality is the number of shortest paths linking every pair of nodes that go through the node examined. Eigenvector centrality weighs the centrality of neighbouring nodes so that a node with high eigenvector centrality may not have high degree itself but is central to the structure of the network. Betweenness and eigenvector centrality measure how connected a node is, and how important that node is to the network. Eigenvector centrality is more conservative than
degree and betweenness and is less affected by network and cluster size. Low centrality indicates that a node is peripheral to a network (Figure 1A).

We evaluated the propensity for nodes to link to other nodes sharing their characteristics in order to test whether pnMSM were more likely to link to each other than to MSM (Figure 1B). This was done by calculating network assortativity as available in the network package (24). Assortativity varies between -1 (completely disassortative) and 1 (completely assortative) and is influenced by the ratio of node labels (here, MSM and HET/pnMSM). To ensure both assortative and disassortative links were possible in each cluster, we analysed only clusters comprising at least two MSM and two male self-reported heterosexuals. To generate a null distribution of expected assortativity given the relative representation of MSM and pnMSM, we randomised risk group 10,000 times across the sub-network and recalculated assortativity each time.

Finally, we tested whether pnMSM might link the MSM and heterosexual epidemics (Figure 1C) by counting the number of relationships they brokered between MSM and heterosexual women and comparing it to the number of relationships brokered by heterosexual women (Figure 1D). We performed a Gould-Fernandez brokerage analysis available in R (25) to count occurrences in the network of heterosexual men linking MSM and female heterosexuals, and compared this to the incidence of females linking MSM and heterosexual men (Figure 1D). This analysis was performed on the entire network as it required clusters in which females were also present.

4. Data sharing
As submission of the entire national cohort to public databases would permit transmission network identification and risk breaching patient confidentiality, we have followed earlier practice (6, 10) and submitted a random sample of 10% of each subtype to GenBank (accession numbers available in the online Appendix p4). The UKRDB Steering Committee encourages analysis proposals. Instructions for applying are available at http://www.hivrdb.org.uk/.

5. Role of the funding source
The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results
The earliest sequence analysed originated from an individual sampled in January 1997, the most recent sequence analysed here originated from an individual diagnosed and sampled in July 2013 (Supplementary Figure 3, Appendix p2). Of 50,025 sequences analysed, 21,053 (42.1%) sequences linked to at least one other at a genetic distance of 4.5% in the RaxML phylogeny. At a time-depth of five years, 14,405 patients shared an ancestor within 5 years and were therefore represented in the resulting network (28.8%, Table 1, Figure 2). Of 4,836 clusters, 3,344 (69.1%) comprised only two sequences, and maximum cluster size was 87 (Supplementary Figure 1, Appendix p1).

Based on the logistic regression, MSM were more likely to cluster than heterosexuals (OR: 2.6, p<0.0001, Table 1), subtype B sequences were more likely to cluster than subtypes A1 and C (OR: 2.8, p<0.0001) and white individuals were more clustered than Black Africans (OR: 1.9, p=0.0002). Some interactions between predictors were significant but effects were small (Supplementary Table 1, Appendix p3). Subtype B, MSM risk group and white ethnicity were independently predictive of clustering even after their interactions were taken into account.
We analysed all clusters that contained one or more self-reported heterosexual men and one or more MSM but no women. We describe these men as potential nondisclosed MSM (pnMSM). There were 223 clusters comprising 955 MSM and 249 self-identified heterosexual men (Figure 3). These pnMSM represented 18.6% of a total of 1341 clustered self-reported heterosexual men. As a comparison, 131/1743 (7.5%) of clustered heterosexual women clustered only with MSM. In the present dataset, 22,296/25,492 (87.5%) MSM were infected with subtype B, while for other risk groups, only 9,353/37,671 (24.8%) were infected with subtype B (Fisher’s exact t-test, p<0.0001).

We compared pnMSM to the MSM with whom they clustered: pnMSM were more likely to be Black African, less likely to be infected with subtype B and were slightly older than the MSM they clustered with (Table 2). There was no difference in the proportion of recent infections between the two groups.

We wanted to test whether the position of pnMSM and MSM in the network differed because a difference in transmission network characteristics is indicative of a difference in underlying transmission dynamics. In clusters comprising MSM and pnMSM, betweenness (1.31 vs 2.24) and eigenvector centrality (0.50 vs 0.52) were significantly lower for pnMSM than for the MSM they clustered with (paired Wilcoxon signed rank test, p=0.002 and p=0.039, respectively; Table 3), indicating that pnMSM were in peripheral positions in clusters. However, mean degree was not different between MSM and pnMSM (paired Wilcoxon signed rank test, 3.02 vs 2.83, p=0.07) and so pnMSM were not less connected overall than MSM.

We then investigated whether pnMSM might be more likely to link to each other in clusters, indicating that pnMSM might be more likely to partner with and be infected by other pnMSM. There were 26 clusters containing at least two MSM and at least two self-reported heterosexual men, comprising 310 MSM and 87 heterosexual men. Among the 26 clusters in the true network, the assortativity coefficient was 0.037, and an assortativity coefficient higher than this was seen in only 110/10,000 of the permutations (p=0.011; simulation mean: -0.037; simulation 95% CI -0.004 - -0.034; Supplementary Figure 2, Appendix p2), indicating that pnMSM linked to each other.

Finally, we tested whether pnMSM might link the MSM and heterosexual epidemics. We found that self-reported heterosexual men were more likely to broker relationships between MSM and heterosexuals (Fisher’s exact test, p=0.0004; OR 2.24 95%CI 1.41-3.61) than heterosexual women but numbers were small for both categories. Of 1341 self-reported heterosexual men represented by nodes in the network, 54 (4.0%) were located in between an MSM and a heterosexual of the opposite gender. This was true for only 32/1711 (1.9%) of heterosexual women.

**Discussion**

We identified potential nondisclosed MSM in the UK epidemic and describe their position in transmission networks. In contrast to previous phylogenetic analyses, which were limited to a single subtype at a time (6, 10, 14), the network-based approach allowed for analysis of multiple subtypes concurrently, leading to a substantial benefit in power. We found that pnMSM were more likely to be in peripheral positions in clusters and to link to each other. pnMSM were more likely to link MSM and heterosexuals of the opposite gender than were heterosexual women. Taken together, these differences in the reconstructed transmission network suggest differences in the transmission dynamics of pnMSM as compared to both MSM and heterosexual men.

A limitation of this study was that the reconstruction method creates a network with more links than the true transmission network because the order of transmissions is difficult to resolve when successive infections occur close together in time. Sexual networks in which individuals became
infected within a short time will appear as fully connected clusters. For this reason, degree here is
not solely related to the number of onward transmissions, but is also a measure of the speed of
transmission within a group. In this context, the equivalent degree but peripheral positions of
pnMSM suggest that they are involved in closely linked MSM sexual networks but transmit less
between these sexual networks, and probably transmit less overall. These differences might reflect
differences in the contact networks of pnMSM or differences in transmission. One explanation might
be that they have a lower partner exchange rate, but another possibility is that they use condoms
more consistently or have sex with lower risk partners. Linking to each other indicates that they may
not find sexual partners in the same venues as MSM and/or seek out other undisclosed MSM for
sex. This finding is supported by interviews, with undisclosed MSM in the USA (26) and men who
have sex with men and women (MSMW) in the UK (22) reporting avoiding venues considered too
public or openly gay, and is important because prevention messages displayed in these venues will
not reach them. MSMW in NATSAL were less likely than MSM to visit sexual health clinics and to
receive prevention there, despite similar rates of sexually transmitted infections, and nearly 50% had
never been to a gay club (22). The Gay Men’s Sex Survey 2014 similarly found that bisexual and non-
gay identified men were 2.5 times less likely to have ever been tested for HIV (27).

If undisclosed MSM have sex with women, they may put female partners at risk. We found that
self-reported heterosexual men more frequently linked MSM to heterosexual women than
heterosexual women linked MSM to heterosexual men. This finding is surprising as men who report
both male and female partners are classified as MSM in the database, so we expect linkage between
MSM and women. However, the number of such potential transmission chains was small, only 54 in
total (4% of all heterosexual men clustered in the network), in agreement with a model quantifying
the extent of bisexual bridging in large cities in the USA (28). In NATSAL data, MSMW reported
unprotected sex with men and women, but lower rates of unprotected sex with men than those
observed among MSM (22), which could explain why this sexual behaviour does not lead to more
transmissions.

The absence of women in clusters formed the basis of our procedure for identifying pnMSM. Our
inclusion of clusters containing multiple self-reported heterosexuals enabled us to test whether
pnMSM linked to each other. We did not analyse clusters involving only heterosexual men (n>200)
because we cannot determine whether women are missing. Even among the clusters examined, we
cannot exclude the possibility that female intermediaries are missing. The higher frequency of non-B
subtypes, most common among heterosexuals, among pnMSM, could indicate that a proportion of
them were in fact infected heterosexually. However, given that MSM and women are both
diagnosed at much higher rates than heterosexual men (2), it is likely that our estimates of
parameters relating to pnMSM remain conservative. If the additional >200 clusters containing only
heterosexual men and >2500 clusters comprising only MSM had been included in the assortativity
analysis, assortativity based on self-reported risk group would have been stronger but even with our
conservative approach it was significant.

Based on the proportion of individuals diagnosed with recent infections, pnMSM were not
diagnosed later than MSM. In contrast, heterosexual men in the UK tend to be diagnosed later than
MSM (2) thus pnMSM may be testing more regularly than heterosexual men. However, the number
of individuals with recency test results was small, which may limit our ability to detect any
difference. We previously found that pnMSM were more likely to be Black African than other
heterosexuals (11) and here found that they were more likely to be Black African than the MSM they
cluster with. These results are in agreement with findings from interviews with Black gay men in the
UK, who are less likely to disclose their sexuality than White gay men (29).
Our results reinforce the need for prevention campaigns appropriate for nondisclosed MSM. They will not be reached by messages displayed in traditional gay venues if they do not frequent them and as a consequence may underestimate their own risk. Interventions that target risk behaviours without focusing on disclosure will be more effective; however, encouraging patients to build trusting relationships with their healthcare providers, in which they feel able to disclose their sexual behaviour, could be a valuable focus for Public Health. In parallel, our findings support the deployment of real-time HIV phylogenetic analyses for targeted public health responses (30). In this case, identification of pnMSM based on their position in networks could help tailor partner services interviews towards them, for example. Further investigation into the behaviour of nondisclosed MSM in the UK, in collaboration with social scientists, should be carried out to better tailor health promotion towards this group.
**Table 1: Risk group, sex, ethnicity and subtype of sequences clustered in the network as compared to the rest of the UK HIV database (UKdb).**

<table>
<thead>
<tr>
<th></th>
<th>UKDB</th>
<th>NODES IN NETWORK</th>
<th>OR (95% CI)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td>50025</td>
<td>14405 (28.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Risk group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HET</td>
<td>17072</td>
<td>3084 (18.1%)</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>MSM</td>
<td>23128</td>
<td>8452 (36.5%)</td>
<td>2.6 (1.6-4.3)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Other/ NA</td>
<td>9825</td>
<td>2869 (28.8%)</td>
<td>0.9 (0.4-1.8)</td>
<td></td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>11461</td>
<td>1881 (16.4%)</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>31360</td>
<td>10305 (32.9%)</td>
<td>1.1 (0.8-1.4)</td>
<td></td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black African</td>
<td>12136</td>
<td>1701 (14.0%)</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>25092</td>
<td>8745 (34.9%)</td>
<td>1.9 (1.3-2.6)</td>
<td>0.0002</td>
</tr>
<tr>
<td>Other/ NA</td>
<td>12797</td>
<td>3959 (31.0%)</td>
<td>2.2 (1.5-3.2)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Subtype</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>2512</td>
<td>405 (16.1%)</td>
<td>Ref</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>31649</td>
<td>11326 (35.8%)</td>
<td>2.8 (2.1-3.8)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>C</td>
<td>15864</td>
<td>2674 (16.9%)</td>
<td>1.0 (0.8-1.2)</td>
<td></td>
</tr>
</tbody>
</table>

UKdb UK HIV Database, HET heterosexual, MSM men who have sex with men, NA not available, F female, M male, CI confidence Interval.

**Table 2: Ethnicity, HIV subtype, recency of infection and age of men who have sex with men (MSM) and potential nondisclosed MSM (pnMSM).**

<table>
<thead>
<tr>
<th></th>
<th>MSM</th>
<th>pnMSM</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td>955</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black African</td>
<td>20 (2.1%)</td>
<td>42 (16.9%)</td>
<td>&lt;0.0001*</td>
</tr>
<tr>
<td>White</td>
<td>813 (85.1%)</td>
<td>147 (59.0%)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>114 (11.9%)</td>
<td>58 (23.3%)</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>8(0.8%)</td>
<td>2(0.8%)</td>
<td></td>
</tr>
<tr>
<td><strong>Subtype</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>9 (1.0%)</td>
<td>4 (1.6%)</td>
<td>0.006*</td>
</tr>
<tr>
<td>B</td>
<td>922 (96.5%)</td>
<td>229 (92.0%)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>24 (2.5%)</td>
<td>16 (6.4%)</td>
<td></td>
</tr>
<tr>
<td><strong>Recency</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recent</td>
<td>236 (83.7%)</td>
<td>53 (84.1%)</td>
<td>1#</td>
</tr>
<tr>
<td>Established</td>
<td>46 (16.3%)</td>
<td>10 (15.9%)</td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mean +/- sd)</td>
<td>36.38 +/- 10.46</td>
<td>39.00 +/- 11.96</td>
<td>0.002*</td>
</tr>
</tbody>
</table>

* fisher’s exact test, * two sample t-test, NA not available

**Table 3: Centrality indicators for MSM and potential nondisclosed MSM (pnMSM). P values were calculated using a paired Wilcoxon signed rank test.**

<table>
<thead>
<tr>
<th></th>
<th>MSM (MEAN, 95%CI)</th>
<th>pnMSM (MEAN, 95%CI)</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betweenness Centrality</td>
<td>2.24 (0.98 – 3.51)</td>
<td>1.31 (0.48-2.15)</td>
<td>0.002*</td>
</tr>
<tr>
<td>Eigenvector Centrality</td>
<td>0.52 (0.50 - 0.55)</td>
<td>0.50 (0.47 - 0.53)</td>
<td>0.039*</td>
</tr>
<tr>
<td>Degree</td>
<td>3.02 (2.65 – 3.40)</td>
<td>2.83 (2.46 – 3.20)</td>
<td>0.073</td>
</tr>
</tbody>
</table>

Note: statistical significance was calculated using a non-parametric paired Wilcoxon signed rank test.
Figures

Figure 1: Cartoon depicting the characteristic positions within men who have sex with men (MSM) clusters of potential nondisclosed MSM (pnMSM). Nodes are coloured by self-reported risk group. A. Based on betweenness and eigenvector, pnMSM tended to be on the periphery of clusters. B. Based on assortativity, pnMSM linked to each other in MSM clusters. C. Based on a brokerage analysis, heterosexual men were more likely to link MSM and heterosexual women than heterosexual females were to link MSM and self-reported heterosexual men. D. The brokerage analysis counts the number of occurrences of each of the two configurations in the network.

Figure 2: Data processing flow chart showing the number of sequences at each stage of analysis. Only clusters comprising at least one UK sequence were counted in the RaxML tree. Percentages were calculated with the total number of sequences from that subtype as the denominator. GD: genetic distance, tMRCA: time to most recent common ancestor.
Figure 3: Clusters containing men who have sex with men (MSM) and heterosexual men. Clusters were selected if they contained no women and at least one heterosexual man and one MSM. Nodes are coloured by risk group with heterosexuals in orange and MSM in blue. Nodes in grey are those for which risk is unknown. All nodes represent men.
Acknowledgments
The authors would like to thank Samantha J. Lycett and Joel O. Wertheim for advice relating to data analysis. The list of UK HIV Drug Resistance Database Steering committee members is available in the Appendix p5.

Author contributions
MRC, SH and ALB devised the study. TF, AP collected patient data and contributed to UK RDB and provided insights into patient behaviour. AEB and VD provided demographic data for all UK patients and provided insights into epidemiological trends. AT and DD collated and processed the viral sequences. EH wrote code and gave advice throughout the project. MRC conducted all analyses. MRC and ALB wrote the manuscript. All authors reviewed and accepted the final manuscript.

Conflicts of interest
MRC is currently receiving funding from Gilead for work on Hepatitis C transmission through a UCSD grant, but was not at the time this research was done. AP has received honoraria for advisory boards and symposia from ViV, Gilead, Merck and Janssen and has acted a consultant for Janssen within the past 4 years. All other authors report no competing interests.

Role of funding source
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics committee approval
The London Multicentre Research Ethics Committee gave ethical approval (MREC/01/2/10).

References
Appendix
Supplementary Figures

Supplementary Figure 1: Cluster size distribution of the UK database HIV transmission network.
Supplementary Figure 2: Assortativity by self-reported risk group was calculated among the clusters (in turquoise). In parallel, risk group labels were randomised 10,000 times within clusters and assortativity was re-estimated in each network to generate a null distribution given risk group frequencies (in pink).
**Supplementary Figure 3**

Sample dates for the 50,025 individuals included in the study, broken down by month.

**Supplementary Tables**

**Supplementary Table 1: Correlates of clustering and interaction between predictors in a multivariate logistic regression model**

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>95%CI (lower)</th>
<th>95%CI (upper)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subtype</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2.8384***</td>
<td>2.0962</td>
<td>3.8429</td>
<td>1.47E-11</td>
</tr>
<tr>
<td>C</td>
<td>0.9928</td>
<td>0.8152</td>
<td>1.2173</td>
<td>0.9438</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.0846</td>
<td>0.8124</td>
<td>1.4405</td>
<td>0.5779</td>
</tr>
<tr>
<td><strong>Risk group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PWID</td>
<td>0.9273</td>
<td>0.2297</td>
<td>3.0538</td>
<td>0.9073</td>
</tr>
<tr>
<td>MSM</td>
<td>2.6300***</td>
<td>1.6119</td>
<td>4.2709</td>
<td>0.0001</td>
</tr>
<tr>
<td>Other/NA</td>
<td>0.875</td>
<td>0.3895</td>
<td>1.7543</td>
<td>0.7253</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other/NA</td>
<td>2.1835***</td>
<td>1.4882</td>
<td>3.1634</td>
<td>4.78E-05</td>
</tr>
<tr>
<td>White</td>
<td>1.8811***</td>
<td>1.344</td>
<td>2.6122</td>
<td>0.0002</td>
</tr>
<tr>
<td><strong>Subtype/ Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B/Male</td>
<td>1.4795*</td>
<td>1.0724</td>
<td>2.0476</td>
<td>0.0175</td>
</tr>
<tr>
<td>C/Male</td>
<td>1.2552</td>
<td>0.9364</td>
<td>1.6898</td>
<td>0.1309</td>
</tr>
<tr>
<td><strong>Subtype/ Risk group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B/PWID</td>
<td>0.4349**</td>
<td>0.2399</td>
<td>0.8135</td>
<td>0.0073</td>
</tr>
<tr>
<td>C/PWID</td>
<td>1.123</td>
<td>0.556</td>
<td>2.311</td>
<td>0.7489</td>
</tr>
<tr>
<td>B/MSM</td>
<td>0.4136***</td>
<td>0.2651</td>
<td>0.6455</td>
<td>0.0001</td>
</tr>
<tr>
<td>C/MSM</td>
<td>0.7259</td>
<td>0.4541</td>
<td>1.1611</td>
<td>0.1805</td>
</tr>
<tr>
<td><strong>Subtype/ Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B/Other</td>
<td>1.4702</td>
<td>0.6787</td>
<td>3.5008</td>
<td>0.3528</td>
</tr>
<tr>
<td>C/Other</td>
<td>0.9131</td>
<td>0.4335</td>
<td>2.1142</td>
<td>0.8201</td>
</tr>
<tr>
<td>B/White</td>
<td>0.6342*</td>
<td>0.4331</td>
<td>0.9332</td>
<td>0.0200</td>
</tr>
<tr>
<td></td>
<td>C/White</td>
<td>1.3528</td>
<td>0.9599</td>
<td>1.9185</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Sex/ Risk group</strong></td>
<td>M/PWID</td>
<td>1.0662</td>
<td>0.756</td>
<td>1.5142</td>
</tr>
<tr>
<td></td>
<td>M/MSM</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>M/Other</td>
<td>0.8306</td>
<td>0.5627</td>
<td>1.2304</td>
</tr>
<tr>
<td><strong>Sex/ Ethnicity</strong></td>
<td>M/Other</td>
<td>0.7657</td>
<td>0.6016</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>M/White</td>
<td>0.678***</td>
<td>0.5438</td>
<td>0.8448</td>
</tr>
<tr>
<td><strong>Risk group/ Ethnicity</strong></td>
<td>PWID/Other</td>
<td>1.5572</td>
<td>0.5203</td>
<td>5.7936</td>
</tr>
<tr>
<td></td>
<td>MSM/Other</td>
<td>1.0848</td>
<td>0.7804</td>
<td>1.5116</td>
</tr>
<tr>
<td></td>
<td>Other/Other</td>
<td>0.9384</td>
<td>0.5385</td>
<td>1.6213</td>
</tr>
<tr>
<td></td>
<td>PWID/White</td>
<td>1.2976</td>
<td>0.4697</td>
<td>4.5893</td>
</tr>
<tr>
<td></td>
<td>MSM/White</td>
<td>1.0646</td>
<td>0.784</td>
<td>1.4494</td>
</tr>
<tr>
<td></td>
<td>Other/White</td>
<td>0.9772</td>
<td>0.5745</td>
<td>1.6595</td>
</tr>
</tbody>
</table>

Reference groups are not shown: subtype A1, female sex, heterosexuals, Black-African ethnicity.

PWID people who inject drugs, MSM men who have sex with men, NA not available, CI confidence interval. ***p<0.001, **p<0.01, *p<0.05
Genbank accession codes

1  KU498303 - KU499301, EU236439 - EU236538, JN100661 - JN101948, KX661407 - KX663330,
2  Q462040, Q462050, Q462058, Q462067, Q462088, Q462094, Q462095, Q462096, Q462098,
3  Q462100, Q462107, Q462109, Q462129, Q462141, Q462144, Q462156, Q462157, Q462159,
4  Q462164, Q462166, Q462172, Q462175, Q462183, Q462184, Q462216, Q462226, Q462231,
5  Q462234, Q462260, Q462264, Q462267, Q462272, Q462300, Q462333, Q462335, Q462353,
6  Q462403, Q462413, Q462416, Q462417, Q462446, Q462463, Q462464, Q462501, Q462526,
7  Q462532, Q462027, Q462029, Q462030, Q462031, Q462032, Q462033, Q462034, Q462036,
8  Q462042, Q462046, Q462047, Q462049, Q462054, Q462057, Q462059, Q462062, Q462064,
9  Q462070, Q462071, Q462072, Q462074, Q462075, Q462077, Q462081, Q462082, Q462083,
10 Q462085, Q462086, Q462087, Q462089, Q462090, Q462091, Q462093, Q462097, Q462101,
11 Q462102, Q462103, Q462111, Q462112, Q462113, Q462116, Q462118, Q462120, Q462121,
12 Q462123, Q462124, Q462126, Q462128, Q462134, Q462137, Q462139, Q462143, Q462147,
13 Q462148, Q462149, Q462150, Q462152, Q462153, Q462154, Q462155, Q462158, Q462161,
14 Q462163, Q462167, Q462171, Q462173, Q462176, Q462178, Q462179, Q462182, Q462185,
15 Q462186, Q462187, Q462189, Q462190, Q462191, Q462195, Q462196, Q462197, Q462198, Q462199,
16 Q462200, Q462208, Q462209, Q462210, Q462211, Q462212, Q462213, Q462219, Q462222, Q462227,
17 Q462228, Q462230, Q462235, Q462239, Q462240, Q462241, Q462242, Q462244, Q462245, Q462247,
18 Q462248, Q462254, Q462262, Q462265, Q462266, Q462269, Q462273, Q462274, Q462278,
19 Q462280, Q462282, Q462283, Q462285, Q462289, Q462290, Q462291, Q462294, Q462295, Q462299,
20 Q462301, Q462302, Q462305, Q462306, Q462308, Q462309, Q462311, Q462312, Q462314,
21 Q462315, Q462316, Q462319, Q462320, Q462321, Q462322, Q462325, Q462326, Q462327,
22 Q462328, Q462329, Q462330, Q462331, Q462332, Q462334, Q462337, Q462338, Q462341,
23 Q462342, Q462343, Q462344, Q462346, Q462347, Q462348, Q462349, Q462350, Q462351,
24 Q462352, Q462356, Q462357, Q462358, Q462359, Q462363, Q462364, Q462366, Q462369,
25 Q462371, Q462372, Q462374, Q462375, Q462376, Q462377, Q462382, Q462383, Q462387,
26 Q462392, Q462393, Q462394, Q462395, Q462396, Q462397, Q462398, Q462399, Q462400,
27 Q462401, Q462404, Q462406, Q462407, Q462411, Q462412, Q462414, Q462415, Q462418,
28 Q462419, Q462423, Q462424, Q462425, Q462426, Q462428, Q462441, Q462443, Q462444,
29 Q462447, Q462448, Q462450, Q462454, Q462455, Q462457, Q462458, Q462460, Q462461,
30 Q462466, Q462467, Q462468, Q462473, Q462474, Q462476, Q462477, Q462481, Q462482,
31 Q462483, Q462484, Q462485, Q462486, Q462487, Q462488, Q462489, Q462491, Q462496,
32 Q462497, Q462498, Q462500, Q462502, Q462503, Q462504, Q462505, Q462506, Q462507,
33 Q462508, Q462509, Q462510, Q462511, Q462514, Q462515, Q462516, Q462518, Q462521,
34 Q462522, Q462523, Q462524, Q462527, Q462529, Q462530, Q462531
UK HIV Drug Resistance Database Steering committee members

Steering Committee: David Asboe, Anton Pozniak (Chelsea & Westminster Hospital, London); Patricia Cane (Public Health England, Porton Down); David Chadwick (South Tees Hospitals NHS Trust, Middlesbrough); Duncan Churchill (Brighton and Sussex University Hospitals NHS Trust); Duncan Clark (Barts Health NHS Trust, London); Simon Collins (HIV i-Base, London); Valerie Delpech (National Infection Service, Public Health England); Samuel Douthwaite (Guy’s and St. Thomas’ NHS Foundation Trust, London); David Dunn, Esther Fearnhill, Kholoud Porter, Anna Tostevin, Oliver Stirrup (Institute for Global Health, UCL); Christophe Fraser (University of Oxford); Anna Maria Geretti (Institute of Infection and Global Health, University of Liverpool); Antony Hale (Leeds Teaching Hospitals NHS Trust); Stéphane Hué (London School of Hygiene and Tropical Medicine); Steve Kaye (Imperial College, London); Linda Lazarus (Expert Advisory Group on AIDS Secretariat, Public Health England); Andrew Leigh-Brown (University of Edinburgh); Tamyo Mbisa (National Infection Service, Public Health England); Nicola Mackie (Imperial NHS Trust, London); Samuel Moses (King’s College Hospital, London); Andrew Leigh-Brown (University of Edinburgh); Eleni Nastouli, Deenan Pillay, Anthony Phillips, Caroline Sabin (University College London, London); Erasmus Smit (Public Health England, Birmingham Heartlands Hospital); Kate Templeton (Royal Infirmary of Edinburgh); Peter Tilston (Manchester Royal Infirmary); Ian Williams (Mortimer Market Centre, London); Hongyi Zhang (Addenbrooke’s Hospital, Cambridge).

Coordinating Centre: Institute for Global Health, UCL (David Dunn, Keith Fairbrother, Esther Fearnhill, Kholoud Porter, Anna Tostevin, Oliver Stirrup)

Centres contributing data: Clinical Microbiology and Public Health Laboratory, Addenbrooke’s Hospital, Cambridge (Jane Greatorex); Guy’s and St Thomas’ NHS Foundation Trust, London (Siobhan O’Shea, Jane Mullen); PHE – Public Health Laboratory, Birmingham Heartlands Hospital, Birmingham (Erasmus Smit); Antiviral Unit, National Infection Service, 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 (Tracy Fawcett); Specialist Virology Centre, Liverpool (Mark Hopkins); Department of Clinical Virology, Manchester Royal Infirmary, Manchester (Peter Tilston); Department of Virology, Royal Free Hospital, London (Clare Booth, Ana Garcia-Diaz); Edinburgh Specialist Virology Centre, Royal Infirmary of Edinburgh (Lynne Renwick); Department of Infection & Tropical Medicine, Royal Victoria Infirmary, Newcastle (Matthias L Schmid, Brendan Payne); South Tees Hospitals NHS Trust, Middlesbrough (David Chadwick); Department of Virology, Barts Health NHS Trust, London (Jonathan Hubb); Molecular Diagnostic Unit, Imperial College, London (Steve Kaye); University College London Hospitals (Stuart Kirk); West of Scotland Specialist Virology Laboratory, Gartnavel, Glasgow (Rory Gunson, Amanda Bradley-Stewart).

Funding: This work is currently supported by the UK Medical Research Council (Award Number 164587).