Med Microbiol Immunol
DOI 10.1007/s00430-011-0228-8

ORIGINAL INVESTIGATION

Endogenous or exogenous spreading of HIV-1 in Nordrhein-Westfalen, Germany, investigated by phylodynamic analysis of the RESINA Study cohort

Glenn Lawyer · Eugen Schülter · Rolf Kaiser · Stefan Reuter · Mark Oette · Thomas Lengauer · The RESINA Study Group

Received: 9 September 2011 / Accepted: 27 December 2011
© Springer-Verlag 2012

Abstract HIV's genetic instability means that sequence similarity can illuminate the underlying transmission network. Previous application of such methods to samples from the United Kingdom has suggested that as many as 86% of UK infections arose outside of the country, a conclusion contrary to usual patterns of disease spread. We investigated transmission networks in the Resina cohort, a 2,747 member sample from Nordrhein-Westfalen, Germany, sequenced at therapy start. Transmission networks were determined by thresholding the pairwise genetic distance in the *pol* gene at 96.8% identity. At first blush the results concurred with the UK studies. Closer examination revealed four large and growing transmission networks that encompassed all major transmission groups. One of these formed a supercluster containing 71% of the sex with men (MSM) subjects when the network was thresholded at levels roughly equivalent to those used in the UK studies, though methodological differences suggest that this threshold may be too generous in the current data. Examination of the endo- versus exogenesis hypothesis by testing whether infections that were exogenous to Cologne or to Dusseldorf were endogenous to the greater region supported endogenous spread in MSM subjects and exogenous spread in the endemic transmission group. In intravenous drug using group subjects, it depended on viral strain, with subtype B sequences appearing to have origin exogenous to the Resina data, while non-B sequences (primarily subtype A) were almost completely endogenous to their local community. These results suggest that, at least in Germany, the question of endogenous versus exogenous linkages depends on subject group.

Keywords Transmission network · Network epidemiology · Disease spread · Endogenous

This study is conducted on behalf of the RESINA Study Group.

G. Lawyer (✉) · T. Lengauer
Department of Computational Biology,
Max Planck Institute for Informatics, Saarbrücken, Germany
e-mail: lawyer@mpi-inf.mpg.de

E. Schülter · R. Kaiser
Institute of Virology, University of Cologne, Cologne, Germany

S. Reuter
Clinic for Gastroenterology, Hepatology and Infectious Diseases,
University Hospital, Düsseldorf, Germany

M. Oette
Clinic for General Medicine, Gastroenterology and Infectious
Diseases, Augustinerinnen Hospital, Cologne, Germany

Introduction

The Human Immunodeficiency Virus (HIV) is only known to transmit via intimate social contact, mimicking and following social networks. The rapid mutation rate of RNA viruses such as HIV means that their evolution and epidemiological spread occur on the same timescale [1, 2]. Thus, given good sampling coverage of a population, measures of genetic relatedness can capture much of the structure of the underlying transmission networks. An understanding of these networks has profound implications for disease prevention strategies. One of the earliest applications of molecular markers to tracking disease networks provided the key for stopping the 1991–1992 San Francisco tuberculosis outbreak [3]. More recently, mathematical models have suggested that targeting individuals which bridge communities is a very effective control strategy [4].

Phylodynamics has provided valuable insights into the spread of HIV. An investigation of primary infections in

Published online: 20 January 2012

Quebec concluded that early infection accounts for approximately half of onward transmission [5]. Further data from Quebec showed that disease transmission during the early infection phase substantially contributed to transmitted drug resistance [6]. Similar results have been found in Switzerland [7]. In the United Kingdom (UK), the relatively high genetic distance between the majority of samples has been interpreted as evidence that most infections have exogenous origin [8–10]. Distribution of subtypes showed travel-related exogenous transmission in the Middle East and North Africa [11]. Phylodynamics have also helped elucidate the different network structure seen in different patient groups. In the UK, the spread of the epidemic was estimated to be approximately twice as fast in the men having sex with men (MSM) compared to the heterosexual group [10]. A study of high-risk individuals in Alabama suggested that people in the intravenous drug using group (IVDA) were significant drivers of the epidemic [12].

The evidence from the UK against a sustained endogenous and spreading disease population [8–10], produced by two independent research groups, is counter-intuitive. Strong arguments suggest that disease spread (in a generic sense) is inherently spatial (i.e., [13]). HIV has the additional feature that it only transmits via intimate contact. Analysis of social networks with geographic location suggests that interactions increase with geographical closeness (i.e., [14]). Thus, the dynamics of both disease spread in general and interactions in social networks with geographical structure suggest that an ongoing epidemic, such as is present in the UK, would have a large endogenous base.

We here investigate whether data suggest that the HIV epidemic in Germany has an endogenous or exogenous structure. Data come from the RESINA Study, an ongoing prospective multi-center investigation monitoring transmitted drug resistance in Nordrhein-Westfalen, Germany [15]. Nordrhein-Westfalen, at 17.8 million inhabitants, is the most populous federal state in Germany and accounts for 21% of documented HIV cases in Germany. Since the structure of HIV transmission networks can be expected to vary considerably in different patient groups, the analysis considers sample-wide and patient-group specific patterns. Along with investigating endo- versus exogenesis, the analysis investigates the structure of transmission networks in the data.

Materials and methods

Subjects

The RESINA Study is an ongoing prospective multi-center investigation carried out in Nordrhein-Westfalen. Thirty-five out of 40 initially contacted centers providing specialized care for HIV-positive patients contribute to the study. The current data include 2,747 patients. The mean age of the patients at the time of sampling was 39 years (standard deviation 10 years). The majority (2,191) are male, 551 are female, and four patients do not have their gender recorded. Inclusion criteria are documented HIV infection, eligibility for treatment, and agreement of the patient and the treating physician on the start of highly active antiretroviral therapy (HAART). Exclusion criteria are prior exposure to antiretroviral drugs and unwillingness to participate. Enrolled subjects gave written informed consent. All relevant institutional review boards have issued positive approval.

Table 1 Groups and subtypes

	Count	Percent (%)
MSM	1,396	51
Endemic	382	14
Hetero	494	18
IVDA	184	7
Bisexual	40	1
Blood	16	–
Unknown	235	9
TOTAL	2,747	100
A	180	7
B	1,962	71
C	94	3
CRF 01_AE	117	4
CRF 02_AG	191	7
CRF other	42	2
D	48	2
G	51	2
Other	62	2
Total	2,747	100

The transmission group and HIV-1 subtype of the 2,747 patients

Consensus sequences of the HIV *pol* gene, which codes for protease and reverse transcriptase, were gathered as part of routine diagnostics. In addition to viral genotype, the data also records gender, age, the first two digits of their postal code, reported transmission route, and HIV-1 subtype. Distribution of the different transmission groups and subtypes are given in Table 1.

Sequences and genetic distance

Sequence alignment was performed as previously described [16]. Genetic distance was computed by counting the number of nucleotide mismatches between subjects. To account for possible convergent evolution due to immune system pressure, suspected HIV escape mutations were removed from the data. These were taken from Brumme et al.'s report [17] and consisted of nine protease codons: (10, 12,

14, 15, 35, 17, 63, 64, and 93) and 22 reverse transcriptase codons: (11, 35, 102, 123, 135, 162, 165, 173, 174, 177, 200, 203, 207, 211, 245, 250, 275, 277, 309, 321, 329, and 335). Brumme et al. list six additional escape mutations associated with the reverse transcriptase protein. These locations are not sequenced in the RESINA Study data. In addition, convergent evolution due to (transmitted) drug resistance was accounted for by removing codons with known drug resistance associations. These were taken from the Stanford HIV Drug Resistance Database [18] and include 17 protease codons: (23, 24, 30, 32, 33, 46, 47, 48, 50, 53, 54, 73, 76, 82, 84, 88, and 90) and 30 reverse transcriptase codons: (41, 67, 70, 210, 215, 219, 65, 74, 75, 115, 69, 151, 62, 77, 116, 98, 100, 101, 103, 106, 108, 179, 181, 188, 190, 225, 227, 230, 236, and 238). As each codon represents three nucleotides, the removal of immune escape and drug resistance codons results in distance measures being computed based on 1,083 nucleotides.

Similarity thresholds

Transmission networks are determined by thresholding the matrix of pairwise genetic distances. Sequences with genetic distance equal to or less than the threshold are linked. A cluster, which is a set of sequences each of which can be reached from each other via links, is considered to represent a transmission network.

No one threshold can fully capture the true underlying transmission networks. Rather, we determine a range of thresholds that capture essential structures of these networks and base our results on tests over this range. Several measures, described below, suggest that thresholds in the range of 35 to 40 base-pair differences (bpd) well capture the network structures. A threshold of 35 bpd represents 96.8% identity.

The first measure considers the relationship between threshold and the number of clusters, where singletons are not considered as a cluster. At a small threshold, very few sequences will be linked. As the threshold is raised, more and more sequences will join to form networks until instead of individual sequences being joined, networks of sequences are joined. Raising the threshold above this point will cause the overall number of networks to fall. An appropriate threshold should be at or just past the peak of this curve. In the current data, this approach suggested that a threshold between 35 and 40 bpd would best reveal the network structure in the subtype B sequences, with similar values for the other subtypes. Figure 1 plots the relationship between threshold and number of clusters for the common HIV subtypes in the data.

A second measure is based on the observation that even at higher thresholds, one observes a few large networks and many small networks and singletons. Under the assumption that these large networks represent real transmission structures, the smallest threshold at which these large networks remain clearly larger than the other networks would capture their kernels. The criteria "clearly larger" is soft; we here consider a twofold increase in size as clearly larger. At a threshold of 50 or less base-pair difference, the data show 6 large clusters with 80, 80, 82, 84, 91, and 1,288 members. The next largest network contains 44 subjects and the next 19. Two of these large networks appear to be artifacts of the high threshold. They drop from 80 and 84 members at a threshold of 50 bpd to 6 and 9 members at a threshold of 40 bpd. By contrast, the remaining large networks have 66, 75, 75, and 429 members at a threshold of 40 bpd. These are clearly larger than the remaining networks in the data, as the next largest has only 24 members. These four largest networks remain clearly larger than any other until a threshold of 36 bpd.

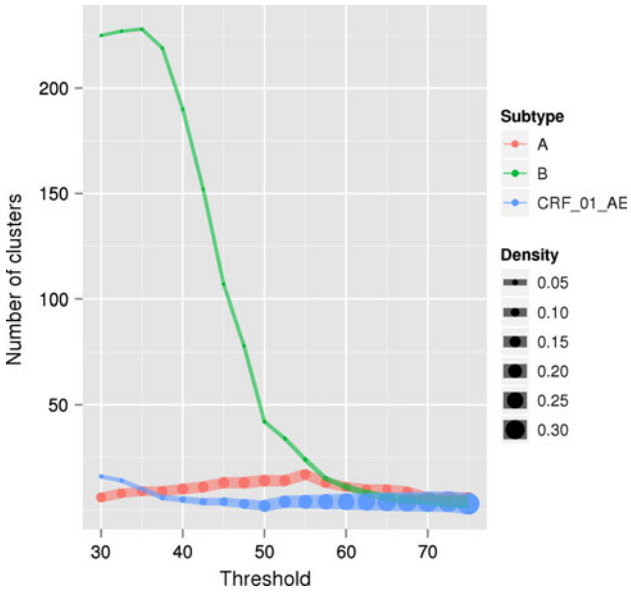

Fig. 1 Number of networks at different thresholds. Plotting the number of connected components at different thresholds identifies the point at which small clusters begin combining into transmission networks. For subtype B sequences, the peak is at a threshold of 35 bpd (96.8% identity), while for circulating recombinant form 01_AE the peak is at or below 30 bpd. Subtype A does not reach its peak until 55 bpd, suggesting that the thresholds used in the current study may not capture the full extent of the network. Mean cluster density remains approximately constant for all reported subtypes over the range 35–45 bpd

Lewis and Hughes et al. [9, 10] compared ambiguous and exact pairwise differences between samples, assuming that ambiguous differences would reflect sequences with shared alleles whereas exact differences would reflect fixation of different alleles in the respective viral populations. A density plot of these values showed a peak at just over 25% bpd, interpreted as inter-subtype differences, and a second peak at approximately 12% bpd, interpreted as intra-subtype

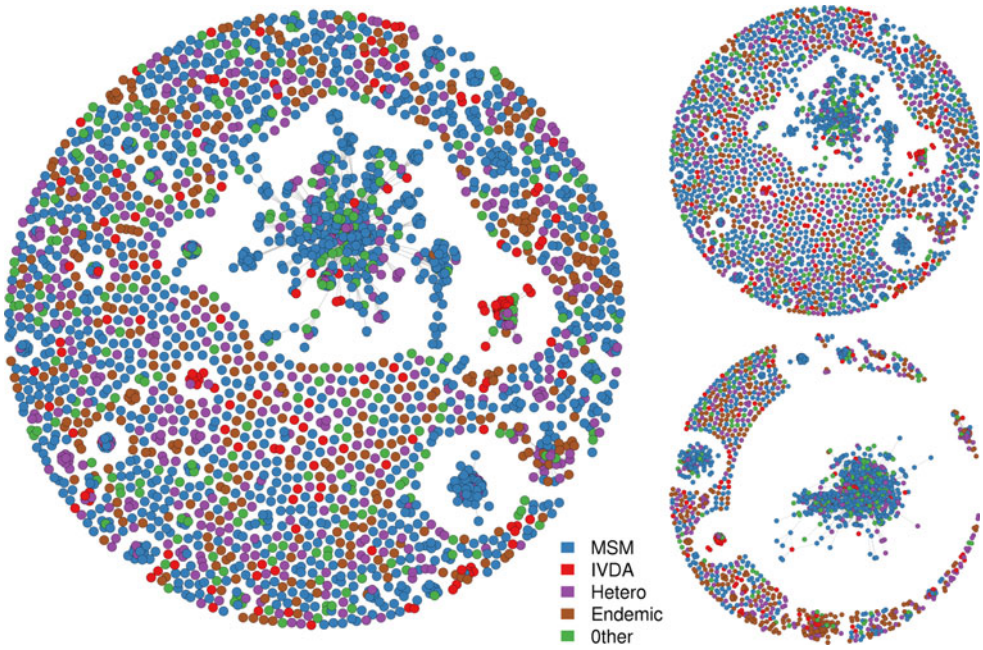

Fig. 2 Petridish plot. The RESINA Study sequences, colored by patient group, shown at thresholds of 40 bpd (*large figure*), 35 bpd (*top right*), and 54 bpd (*bottom right*). At the lower thresholds, most nodes are in small networks or not linked to other nodes. The large central network expands to 1,463 members (53% of the data) at a threshold of 54 bpd. Table 1 gives the proportion of subjects in each transmission group

differences. A third, smaller peak at approximately 5% bpd, was interpreted as groups of closely related sequences, that is, transmission chains. To avoid convergent evolution effects due to drug pressure, the authors apply the 5% rule to third codon positions. The 5% rule, which translates to 54 bpd, appears to be overly generous in the current data. Nonetheless, we occasionally report results at this level to better illuminate the data.

Figure 2 plots the entire data at thresholds of 35, 40, and 54 bpd, illustrating the clustering patterns at the different thresholds.

Network structure

Following the example of the studies of HIV transmission in the UK [8–10], we first investigate the distribution of cluster sizes, including the proportion of unclustered sequences, in the full data and in the different patient groups.

We next investigate the structure of the four large networks. The patient-group composition of each network is determined. The density of each network, a graph-theoretic measure of its interconnectedness, is obtained. In addition, we measure each network's temporal span, measured as the difference between the earliest and latest sequencing dates for included patients. Linear models are used to determine the growth rate of each cluster over time as well as the rate of genetic drift within the cluster. Fit of the growth rate model is computed using the r-squared statistic. The significance of the observed genetic drift is measured using an ANOVA-based p-value.

Endo- versus exo-genous structure

To further investigate the question of endogenous spread versus exogenous/migration-based influx, the following approach is used. Transmission networks are created using only subjects in Cologne, and again using subjects only in Dusseldorf. HIV is primarily an urban disease. Cologne and Dusseldorf represent the majority of the patients in the study, and, while not geographically distant, the two cities are separated by a substantial cultural divide.

For each sequences, which is not part of a local transmission network (i.e., which is exogenous to the city), we compute the distance to the next closest sequence in the data. Distances that are below the 40 bpd threshold evidence that the infection is part of a region-wide transmission network. The proportion of such sequences estimates the proportion of endogenous (to Nordrhein-Westfalen) HIV transmission. The proportion of such distances which indicates genetic similarity is indicative of the amount of endogenous spread at the regional level. The results are segmented both by city and by patient group. Similarity or divergence between groups and cities in these distance distributions is established using a two-sided Kolmorgorov-Smirnov test.

Results

Transmission networks

The initial investigation found results that are roughly concordant with the UK studies. Sequences that are not part of

Table 2 Cluster size distribution

Group	*n*	% Clustered	% Small clusters
MSM	1,396	53–63	35–27
Endemic	382	34–42	30–34
Hetero	493	47–56	33–31
IVDA	184	54–62	25–17
Bisexual	40	50–62	32–30
Unknown	252	53–60	38–31
TOTAL	2,747	49–58	34–29

Cluster sizes are unevenly distributed according to transmission group. The table shows, for each transmission group, the number of subjects, percent clustered subjects, and percent small (10 or fewer member) clusters in that group. The figure before the dash is at a threshold of 35 bpd, after is at 40 bpd. Transmission groups with less than 20 subjects (blood products, $n = 16$, pre/perinatal, $n = 1$, prostitute, $n = 1$) are placed in the "Unknown" transmission group when computing this table

a transmission network are common, as are small clusters that imply non-perpetuating transmission chains. Of the 2,747 sequences, 1,391 (51%) are more than 35 bpd from all other sequences in the data, that is, not part of a transmission cluster. This value falls to 1,149 (42%) at a threshold of 40 bpd. Most of the observed clusters are small, with 924 subjects in clusters of size 10 or less (34% of total sequences and 68% of clustered sequences). At a threshold of 40 bpd, the number of subjects in clusters of size 10 or less is 787 (29% of total, 49% of clustered sequences). Combined, then, between 71 and 85% of all sequences in the data do not appear to be part of a large transmission network, at least at this level of analysis and within the defined threshold range.

The clustering proportion is not consistent between the different transmission groups. IVDA and MSM subjects are more likely to have a similar sequence in the data (IVDA: 54–62%, MSM: 53-63%), while heterosexual and endemic subjects are less likely (heterosexual: 47–56%, endemic: 34–42%). The proportions of unclustered and small clusters by transmission group are given in Table 2. Kolmorgorov-Smirnov testing suggested that the cluster size distributions are indeed different between these different groups ($p < 0.001$, all comparisons).

Large networks

At the upper threshold of 40 bpd, four networks have more than 50 members. The next largest network contains 24 individuals. Two of these large networks are principally composed of patients from the MSM transmission group, and carry subtype B infections. One network drew three quarters of its members from the IVDA transmission group and is composed of subtype A (A1) strains. The remaining large network contained equal proportions of patients from the heterosexual and the endemic transmission groups and is composed of CRF 01_AE strains. The large MSM and the heterosexual/endemic networks have densities below 0.10, indicating loosely connected networks. The smaller MSM network's density grows from 0.08 at the low threshold to 0.17 at the high threshold. Raising the threshold increases the number of connections between member much faster than it increases the number of members. The IVDA network is the densest of the four, with densities in the range 0.20–0.26. In biological terms, this means that all of the HIV strains in this cluster are tightly related. Figure 3 shows the four networks at varying thresholds. Table 3 summarizes the transmission group and density estimates of the networks, along with the temporal extent and growth rates of the networks.

Network growth rates over time are strongly linear for all networks at all thresholds. The r-squared values, a measure of the proportion of variability explained by the model, range from 94 to 99%, depending on the network and threshold. The temporal extent of all four large networks was consistent over all thresholds tested. We do not draw conclusions regarding the relationship between growth rate and threshold. Given that temporal extent does not depend on threshold, straightforward algebra shows that growth rate at different thresholds is dependent on cluster size at different thresholds, which is reported in Table 3. A plot of the network growth rates at the 40 bpd threshold is contained in Fig. 4.

Genetic drift was only measured at the 40 bpd threshold. At this level, it is small but highly significant ($p < 0.001$) for three of the networks. The large MSM network has an estimated genetic drift of 0.002 bpd/day. The heterosexual/endemic network shows an estimated drift of 0.003 bpd/day. The IVDA network, in contrast, coalesces with time. The estimated drift is a decrease of 0.004 bp/day. Plots of the relationships are given in Fig. 5. The smaller of the MSM networks spans 3 years, and does not show a significant relationship between distance and time, possibly due to its short duration. It is excluded from the figure.

The two cities

The first measure taken from the two cities was the percentage of subjects with a similar sequence in the city, that is, those that appear to belong to an endogenous (to the city) transmission network. For MSM subjects, these percentages are similar to the values seen in the full data. For subjects in the heterosexual group, the percentages in the cities are slightly smaller than observed in the full data. For the endemic transmission group, they are noticeably larger.

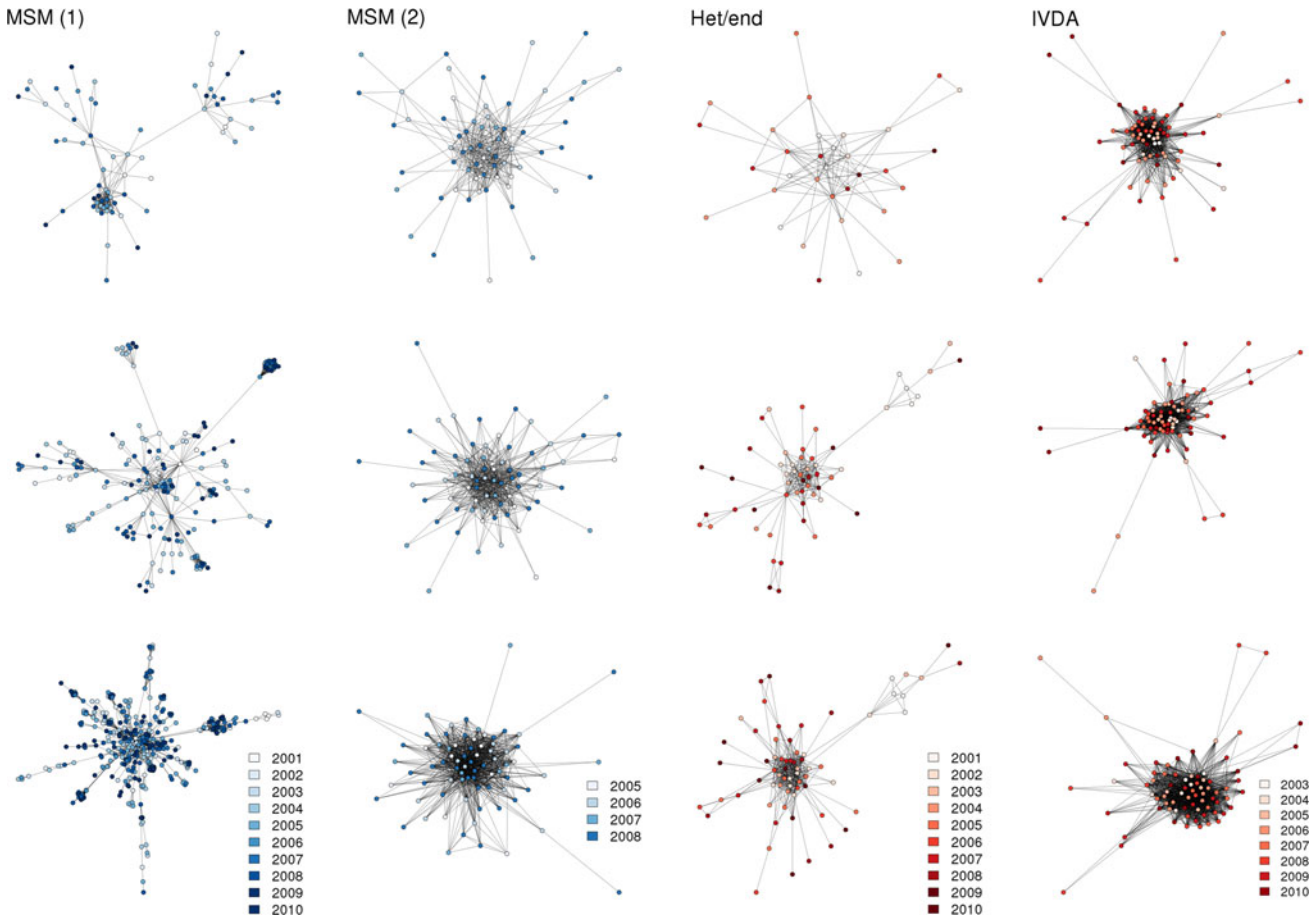

Fig. 3 The four large networks. The study identified four large transmission networks. Two are composed primarily of subjects in the MSM transmission group (MSM (1) and MSM (2)), the third contains primarily subjects from the heterosexual and endemic transmission groups, and the fourth is composed primarily of IVDA subjects. The *upper row* shows the networks at a threshold of 35 bpd, the *middle row* at 37 bpd, and the *lower row* at 40 bpd, illustrating how the core of the network expands as the threshold is increased. Shading the nodes according to the year in which they were sequenced shows that earlier nodes tend to be more central to the network. The four networks are further characterized in Figs. 4 and 5, and in Table 3

Table 3 Large network characteristics

Transmission group	*N*	Subtype	Density	Growth	Duration
MSM (69–72%)	70–429	B	0.05–0.01	8–51	9
MSM (75–77%)	64–75	B	0.08–0.17	17–19	3
Het/end[a] (52/33–42/32%)	33–66	CRF 01_AE	0.10–0.09	4–7	8–9
IVDA (65–64%)	71–75	A	0.20–0.26	11	7

The four large networks can be characterized by the predominant transmission group, their size, the subtype, and the network density. Numbers before the dash are at a threshold of 35 bpd, after at 40 bpd. Growth rate is number of new cases per year. Duration is number of years between first and last observed member

[a] Heterosexual/endemic

The second measure considers the proportion of subjects who do not belong to a city-specific transmission network, but who do have a similar sequence elsewhere in the RESINA data. For all transmission groups except IVDA, this second measure has no significant difference between the two cities. The difference between transmission groups, however, was strongly significant. MSM patients show a several fold increase in the proportion of sequences with a similar sequence elsewhere in the data compared to endemic patients. The rate for heterosexual patients lies in

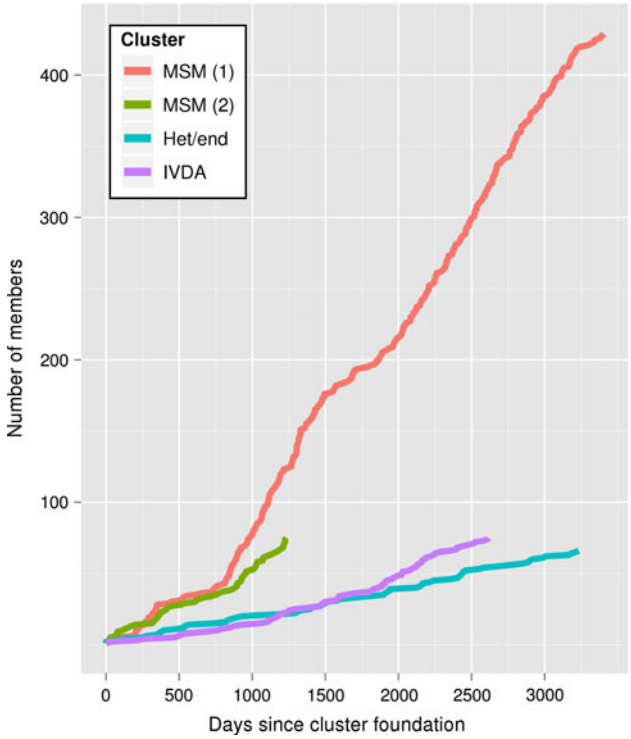

Fig. 4 Network growth. Growth rates in the four large networks are strongly linear. The plot shows the cumulative membership over time for each network, normalized to the date of the first sequencing date in the network. Network names are as in Fig. 3 and are based on the predominant transmission group of its members

between that of MSM and endemic patients. Kolmorgorov-Smirnov testing confirms that the distribution of next closest sequences differs markedly between the MSM and endemic group ($p < 0.001$, both cities), and between the MSM and heterosexual group (Dusseldorf $p = 0.001$, Cologne $p = 0.007$). Table 4 summarizes the number of clustered and unclustered sequences for each city, as well as the fraction of sequences that are close to another sequence in the data. Figure 6 shows the clustering and the distributions of distances to the next closest sequence for the MSM and endemic subjects.

The IVDA group has different clustering patterns in each city. At a threshold of 35 bpd, 76% of IVDA subjects in Dusseldorf are linked to another subject in the city. Most of these belong to a single network of 34 individuals. This large cluster is part of the large subtype A cluster discussed in the preceding section. The 47% of clustered IVDA subjects in Cologne, by contrast, form small clusters (five pairs, one triplet). The difference is also reflected in the relative proportion of subtype B to subtype A patients in the two cities. In the full data, half of the IVDA patients (91 out of 184) carry subtype B strains, with subtype A being the next most common ($n = 57$). Dusseldorf, however, has more subtype A ($n = 36$) than subtype B ($n = 25$) IVDA patients. In Cologne, subtype B predominates ($n = 23$ vs. $n = 3$ subtype A) in the IVDA community.

Discussion

The fundamental question of this paper is whether HIV spread in Nordrhein-Westfalen is best described as endogenous spread or repeated re-introduction of the virus to the region. At the level of the full dataset, our results support the repeated re-introduction (exogenous) explanation and concur with the UK studies. We estimated that 74–83% of sequences were not part of large transmission networks. The two UK investigations of subtype B sequences found that 86% of 1,645, respectively, and 82% of 2,126 subtype B sequences were not part of endogenous transmission networks [8, 9]. A third UK study investigating non-B subtypes also found clustering to be rare, with 75% of the samples reported as not having a genetically similar sequence recorded [10]. The UK studies interpreted the lack of observed clustering as evidence for widespread and worldwide geographical mixing and migration of HIV [8, 9]. None of these studies mentioned IVDA patients.

A closer look, however, suggests that at least in the MSM and IVDA communities the spread is largely endogenous. From the viewpoint of epidemiology, a consistent clustering would indicate distinctly bounded epidemiological entities [9]. Under this definition, the large MSM cluster is not bounded, as its size increases dramatically with increased thresholds. In fact, when the threshold is relaxed to 54 bpd, the threshold suggested by the 5% rule used in the UK studies, this cluster contains 1,463 individuals, or 53% of the full data and 71% of the MSM subjects in the data ($n = 991$). The data from the two cities also support a trend toward endogenous spread of HIV in the MSM patient group. At a threshold of 40 bpd, one-quarter of the patients with no similar sequence in their city are in fact linked to another patient elsewhere in Nordrhein-Westfalen.

The divergence from the conclusions of the UK studies could be explained by gaps in their subject samples. Hue et al.'s [8] UK wide study contained less than 3% of the number of infections in the country. Such low coverage makes local networks difficult to detect. Lewis et al.'s [9] samples were all taken from one large hospital in London. While the size of the hospital and study population would suggest that local networks should have been uncovered, the study does not allow one to rule out transmission from other parts of London or England.

A general weakness of both the Lewis et al. study and the current RESINA Study investigation is that not all infected individuals are aware of their status. One study found that 45% of HIV-positive men active in Miami,

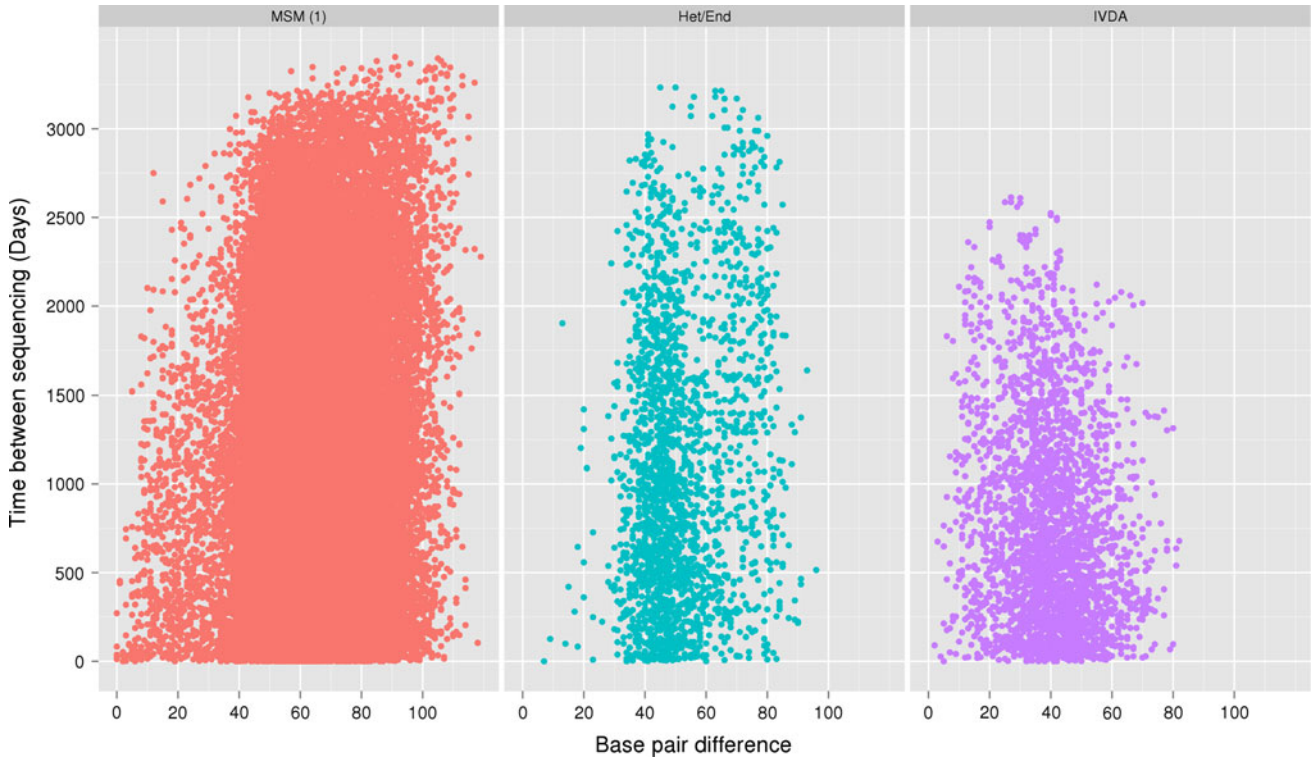

Fig. 5 Genetic drift within networks. Comparing genetic distance to the number of days between sequencing dates, pairwise for all pairs in each transmission network, finds that three of the four networks have a small but highly significant relationship between sample date and genetic distance. The IVDA network shows decreasing genetic distance with temporal separation, while the remaining networks show increasing distance with time. Sequencing dates in the MSM (2) network span only 3 years, and no significant genetic drift is observed. It is not shown in this plot. Rates of drift are given in the main text

Florida's MSM culture were HIV positive without awareness of their infection status [19]. Until they seek medical attention, these people are not present in the data. A study of people in the primary, acute phase of HIV infection found that such cases tended to form transmission clusters, and the findings are consistent with other studies showing that such patients play a key role in the onward spread of the disease [6]. This suggests that the number and/or size of actual transmission networks is underestimated in the data.

Our data evidenced a large transmission network among IVDA patients. This network contains the majority of subtype A carrying IVDA patients in the data (48 out of 57). Its growth rate of 11 cases/year is especially alarming given the relatively small percentage of German residents who regularly inject illicit substances. The coalescence over time of the HIV genotypes in this IVDA cluster is also noteworthy. One interpretation of this result is that patients with long-standing infection re-transmit HIV. This interpretation is further supported by the density of the network. Both the speculative interpretation and the density of the network are strong arguments in favor of needle exchange and other harm reduction efforts to stop the spread of HIV. The enhanced connectivity of a dense graph facilitates transmission of the HIV virus [4]. Increasing safe injecting practices would break links and reduce the density of the network.

The results from studying transmission networks in the two cities suggest that HIV transmission within the endemic group in Germany is strongly biased toward urban settings. Nearly 60% of endemic patients in Cologne and Dusseldorf belonged to a locally constrained transmission network, compared to 42% in the full data. Yet few endemic subjects belonged to networks that extended outside of the city. These results are concurrent with the third UK study [10]. Despite high coverage of the UK patient base, transmission networks were reported to be largely absent in patients carrying non-B strains. In the UK, non-B subtypes are strongly associated with subjects in the endemic group [10]. These findings are also concurrent with what one would expect from this subject group, which is by definition biased toward people who are likely to have acquired the infection during sustained time spent in a country outside of Europe.

This study found evidence of four large transmission networks in the RESINA Study patients, encompassing all major transmission groups. All networks have strongly linear growth rates. This argues that HIV spread is not under control in the region and that the disease has several strong,

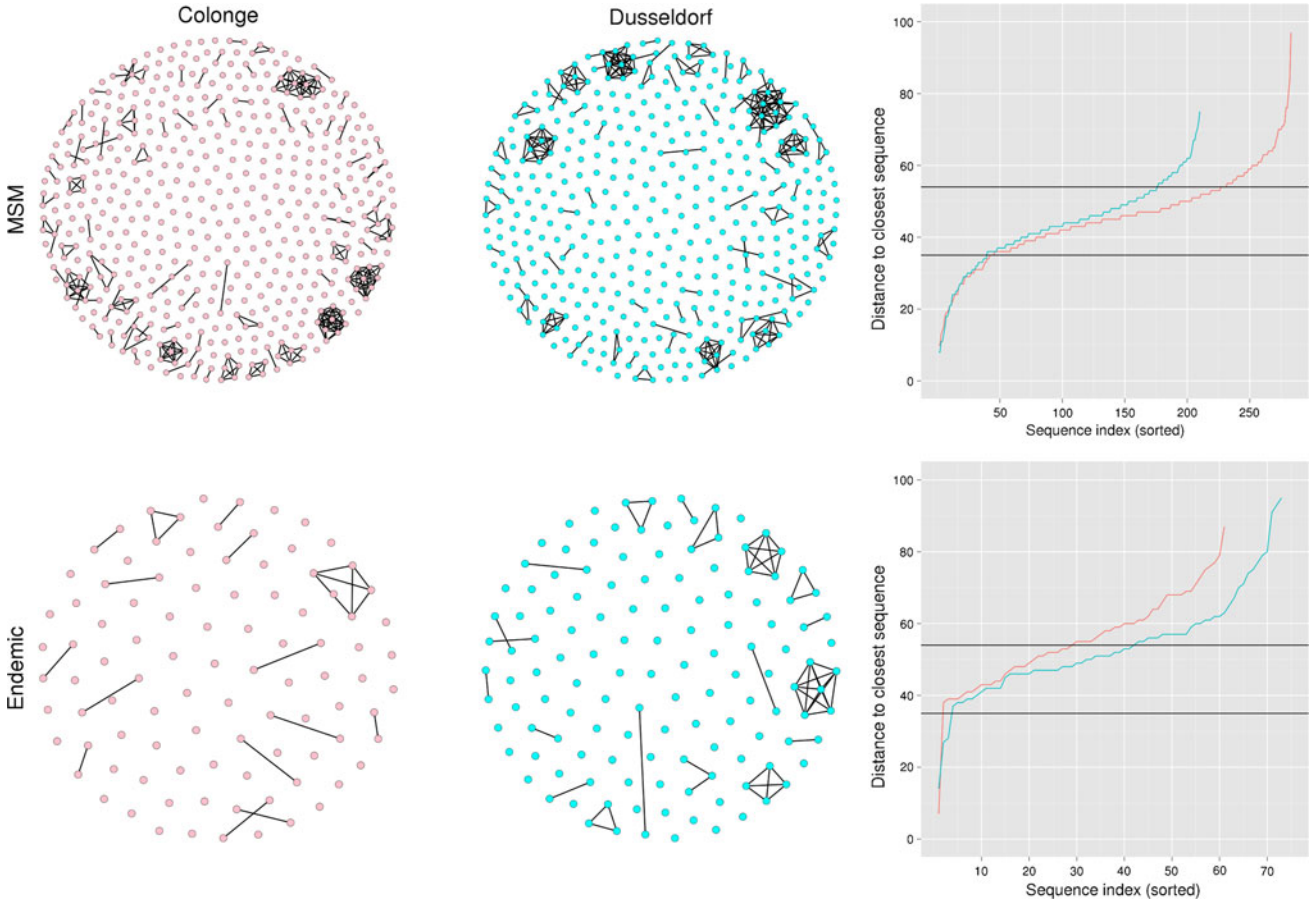

Fig. 6 Endogenous versus exogenous spread. MSM patients (*top row*) and endemic patients (*bottom row*) show similar network structures in both cities, though the smaller population of endemic patients means that this group has a smaller absolute number of networks. The charts on the right of the figure show, for sequences without a close match in the city, the cumulative distribution of the distance to the next closest sequence in the entire RESINA Study cohort. MSM subjects without a close link in their city (*top*) are still likely to be in a transmission network located within Nordrhein-Westfalen. Endemic subjects (*bottom*), by contrast, are much less likely to be in such a network. See also Table 4

Table 4 Endo versus exogenous characteristics

Group	City	*N*	*N*-end	% Endo	% 35	% 40	% 54
MSM	Cologne	510	252	0.49	0.14	0.22	0.79
MSM	Dusseldorf	418	184	0.44	0.17	0.25	0.82
Hetero	Cologne	144	63	0.44	0.03	0.06	0.65
Hetero	Dusseldorf	150	70	0.47	0.09	0.11	0.70
Endemic	Cologne	96	57	0.59	0.02	0.04	0.44
Endemic	Dusseldorf	131	72	0.55	0.03	0.08	0.54
IVDA	Cologne	34	16	0.47	0.12	0.12	0.69
IVDA	Dusseldorf	68	15	0.22	0.20	0.27	0.67
Unknown	Cologne	66	31	0.47	1.29	2.06	9.23
Unknown	Dusseldorf	125	54	0.43	0.80	1.20	4.74
Total	Cologne	850	419	0.49	0.10	0.16	0.71
Total	Dusseldorf	892	395	0.44	0.13	0.19	0.73

For each patient group and city, this table gives the count, the count of clustered sequences (threshold 40 bpd), and the percentages of these unclustered sequences with a similar sequence elsewhere in the RESINA Study cohort at thresholds of 35, 40, and 54 bpd

endemic, and spreading transmission networks in Nordrhein-Westfalen. The current data concur with the Hughes et al. [10] findings that the growth rate in patients in the endemic/heterosexual transmission groups is half that observed in patients in the MSM transmission group.

The current investigation has several limitations. Patients were enrolled in the study at the initiation of HAART, suggesting that they have knowingly carried the HIV infection for a number of years before the HIV sequence was recorded. Assuming a transmission link between two subjects, this delay implies that the virus transmitted is not identical to the virus sequenced. We mitigate this by only considering sequence similarity of the *pol* gene. This gene codes proteins crucial to viral replication, placing it under much more consistent adaptive pressure than the *env* or *nef* genes, which mutate rapidly in response to the host immune system [20, 17]. We also remove known escape and drug resistance mutations before computing similarity scores. Nonetheless, the genetic drift

makes determination of direction of transmission impossible in the current data.

Further aspects of the demographic data are also less precise than desirable. The geographical location of the patient is limited to the first two digits of their postcode, and the recorded postcode is for the patient's current residence, not their residence at time of infection. Group labels were also imprecise. A recent investigation, also involving the RESINA Study cohort, has found that 59% of the Turkish men who report the mode of transmission as heterosexual have HIV sequences that are closely linked with samples from the MSM group [21]. We note, however, that prevalence of HIV in people of Turkish origin is lower than in the remaining population.

The strengths of the study include a large, well characterized data set that provides thorough coverage of the most populous state of Germany. The data include all major patient groups and HIV subtypes, though the discussion in the current work focuses on those which were most strongly represented in the implied transmission networks. Geospatial information allows closer investigation into the source and spread of the disease. Temporal dynamics are explored.

Conclusions

Phylodynamic investigation of the HIV epidemic in Germany suggests the presence of large endogenous HIV transmission networks in the MSM and IVDA communities. The IVDA network is composed of HIV-1 group M subtype A, while IVDA subjects carrying subtype B tend not to be involved in transmission networks. Infections in endemic subjects do not seem to form large transmission networks within the region.

Acknowledgments The authors wish to thank Claudia Mueller, Saleta Sierra, Nadine, Luebke, Melanie Balduin, Doerte Hammerschmidt, Jens Verheyen from the Institute of Virology, and the cooperating HIV centers. This work was supported by the RESINA Study Project (BMG IIA5-2010-2510AUK361), CHAIN Project (EU-223131), EURESIST project (IST-4-027173), MedSys project (BMBF-0315489C), and CORUS Project (BMBF 01ES0712). All experiments and use of data comply with the current laws of Germany and were approved by the relevant ethical boards.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303(5656):327–332. doi:10.1126/science.1090727
2. Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 10(8):540–550. doi:10.1038/nrg2583
3. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N Engl J Med 330(24):1703–1709. doi:10.1056/NEJM199406163302402
4. Salathé M, Jones JH (2010) Dynamics and control of diseases in networks with community structure. PLoS Comput Biol 6(4):e1000,736. doi:10.1371/journal.pcbi.1000736
5. Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C, Baril JG, Thomas R, Rouleau D, Bruneau J, Leblanc R, Legault M, Tremblay C, Charest H, Wainberg MA, Group QPHIS (2007) High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis 195(7):951–959
6. Brenner BG, Roger M, Moisi DD, Oliveira M, Hardy I, Turgel R, Charest H, Routy JP, Wainberg MA, Cohort MP, Groups HIVPS (2008) Transmission networks of drug resistance acquired in primary/early stage HIV infection. AIDS 22(18):2509–2515. doi:10.1097/QAD.0b013e3283121c90
7. Castro E, Khonkarly M, Ciuffreda D, Bürgisser P, Cavassini M, Yerly S, Pantaleo G, Bart PA (2010) HIV-1 drug resistance transmission networks in southwest Switzerland. AIDS Res Hum Retroviruses 26(11):1233–1238. doi:10.1089/aid.2010.0083
8. Hué S, Pillay D, Clewley JP, Pybus OG (2005) Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc Natl Acad Sci USA 102(12):4425–4429. doi:10.1073/pnas.0407534102
9. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Brown AJL (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 5(3):e50. doi:10.1371/journal.pmed.0050050
10. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Brown AJL, U.K.HIV Drug Resistance Collaboration (2009) Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog 5(9):e1000,590
11. Mumtaz G, Hilmi N, Akala FA, Semini I, Riedner G, Wilson D, Abu-Raddad LJ (2011) HIV-1 molecular epidemiology evidence and transmission patterns in the Middle East and North Africa. Sex Transm Infect 87(2):101–106. doi:10.1136/sti.2010.043711
12. Volz E, Frost SDW, Rothenberg R, Meyers LA (2010) Epidemiological bridging by injection drug use drives an early HIV epidemic. Epidemics 2(3):155–164. doi:10.1016/j.epidem.2010.06.003
13. Biek R, Real LA (2010) The landscape genetics of infectious disease emergence and spread. Mol Ecol 19(17):3515–3531. doi:10.1111/j.1365-294X.2010.04679.x
14. Helsley RW, Zenou Y (2011) Social networks and interactions in cities. SSRN eLibrary
15. R.E.S.I.N.A. Study Team: (2007) Trends of prevalence of primary HIV drug resistance in Germany. J Antimicrob Chemother 60(4):843–848
16. Oette M, Kaiser R, Däumer M, Akbari D, Fätkenheuer G, Rockstroh JK, Stechel J, Rieke A, Mauss S, Schmalöer D, Göbels K, Vogt C, Wettstein M, Häussinger D (2004) Primary drug-resistance in HIV-positive patients on initiation of first-line antiretroviral therapy in Germany. Eur J Med Res 9(5):273–278
17. Brumme ZL, Brumme CJ, Heckerman D, Korber BT, Daniels M, Carlson J, Kadie C, Bhattacharya T, Chui C, Szinger J, Mo T, Hogg RS, Montaner JSG, Frahm N, Brander C, Walker BD, Harrigan PR (2007) Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. PLoS Pathog 3(7):e94. doi:10.1371/journal.ppat.0030094
18. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31(1):298–303

19. Center for Disease Control and Prevention (2005) HIV prevelance, unrecognized infection, and HIV testing among men who have sex with men–five U.S. cities, June 2004–April 2005. Morb Mortal Wkly Rep 54. U.S. Department of Health and Human Services, Center for Disease Control and Prevention
20. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73(12):10,489–10,502
21. Schülter E, Oette M, Balduin M, Reuter S, Rockstroh J, Fätkenheuer G, Esser S, Lengauer T, Agacfidan A, Pfister H, Kaiser R, Akgül B (2011) HIV prevalence and route of transmission in Turkish immigrants living in North-Rhine Westphalia, Germany. Med Microbiol Immunol. doi:10.1007/s00430-011-0193-2