A novel methodology for large-scale phylogeny partition

Mattia C.F. Prosperi1,2,3, Massimo Ciccozzi4, Iuri Fanti1, Francesco Saladini5, Monica Pecorari6, Vanni Borghi7, Simona Di Giambenedetto1, Bianca Bruzzone8, Amedeo Capetti9, Angela Vivarelli10, Stefano Rusconi11, Maria Carla Re12, Maria Rita Gismondo13, Laura Sighinolfi14, Rebecca R. Gray2,3, Marco Salemi2,3, Maurizio Zazzi5 & Andrea De Luca15 on behalf of the ARCA collaborative group

Understanding the determinants of virus transmission is a fundamental step for effective design of screening and intervention strategies to control viral epidemics. Phylogenetic analysis can be a valid approach for the identification of transmission chains, and very-large data sets can be analysed through parallel computation. Here we propose and validate a new methodology for the partition of large-scale phylogenies and the inference of transmission clusters. This approach, on the basis of a depth-first search algorithm, conjugates the evaluation of node reliability, tree topology and patristic distance analysis. The method has been applied to identify transmission clusters of a phylogeny of 11,541 human immunodeficiency virus-1 subtype B pol gene sequences from a large Italian cohort. Molecular transmission chains were characterized by means of different clinical/demographic factors, such as the interaction between male homosexuals and male heterosexuals. Our method takes an advantage of a flexible notion of transmission cluster and can become a general framework to analyse other epidemics.
The spread of type-1 human immunodeficiency virus (HIV-1) infection is influenced by a long duration of the asymptomatic stage, a high viral evolutionary and replication rate, as well as population movements, social behaviour and geopolitical factors such as infrastructure development and accessibility. Phylogenetic analysis can be used to investigate HIV-1 transmission events, although the long period of infectivity, coupled with often non-uniform spatial and temporal sampling, can limit its efficiency in tracing the infection chains.

Several epidemiological factors, for example the mode of HIV-1 transmission or the disease stage, might be associated to different transmission lineages. Antiretroviral therapy (ART) usually suppresses HIV-1 replication, reducing the probability of virus transmission. Drug-resistant strains can be selected by ART exposure, and transmitted from patients failing ART or from drug-naive patients already carrying a drug-resistant virus. Given the decreasing prevalence of virological failures among HIV-1-infected patients and of vireamie patients carrying drug resistance over time, a greater proportion of transmitted resistance may be derived from ART-naive subjects.

Transmission clusters among HIV-1-infected patients have been inferred and analysed across various study cohorts, including the United Kingdom, Switzerland, Canada and the Netherlands. Nowadays, parallel computation techniques permit the estimation of highly resolved phylogenies for large populations. However, the partition of phylogenetic trees is a challenging task and there is not yet an established automated consensus methodology to identify transmission clusters, particularly when considering large-scale phylogenies. Furthermore, the definition of a transmission cluster itself is not standardized. Transmission clusters have been associated to sub-trees of phylogenies containing sequences from at least x distinct subjects (from a minimum of two patients to more than ten patients) with a sub-tree reliability of ≥ y% (usually bootstrap support or posterior probability, ranging from 80 to 99%). Geographical constraints, that is, ≥ z% of sequences from the same country in a cluster (usually ≥ 80%), have been also introduced. Often, unrelated viral strains from different countries are added as controls, and clusters are identified by visually inspecting an initial tree or by re-estimating smaller trees after reducing the original data set. Cluster selection based on absolute genetic and patristic distance values has also been used.

This work proposes a new approach for the automated partition of large phylogenies, by introducing a flexible (yet statistically robust) definition of a transmission cluster, which fits also the previously introduced definitions. The method is then applied for identifying transmission clusters within the HIV-1 subtype B epidemic in Italy. Data were drawn from the Antiretroviral Resistance Cohort Analysis (ARCA), a national observational cohort of HIV-1-infected patients followed up at 105 clinical centres. At the time of this study, data from > 20,000 patients and > 23,000 HIV-1 pol gene sequences were available. A phylogeny on the Italian HIV-1 subtype B isolates was estimated by parallel computation and, after applying the new partition method, clustering was linked with patients' demographic and clinical information. As a result, several factors associated with transmission chains were identified, suggesting strategies to monitor the epidemic. The methodology proposed here provides a general and robust framework for analysing large-scale phylogenies that can be applied in investigating the molecular epidemiology and the intra-host evolutionary patterns of a broad range of pathogens.

Results

Automated partition of phylogenetic trees. The definition of transmission cluster introduced here requires a rooted phylogeny and assumes that two or more patients belong to a potential transmission cluster if their viral sequences are monophyletic and more closely related than those from two randomly selected individuals. The meaning of 'more closely related' refers to the comparison with a distance threshold formally defined as follows.

Given a rooted phylogeny, the distribution of all patristic distances between pairs of taxa (whole-tree distribution) and the distribution of all pairwise patristic distances within any sub-tree (sub-tree distribution) are defined. Then, a specific sub-tree is classified as a transmission cluster if the median value of its distance distribution is below a t-percentile threshold (which needs to be properly tuned, see below) of the whole-tree distribution.

As complimentary constraints, a sub-tree is considered to be eligible as transmission cluster only if it contains sequences from ≥ 2 subjects and has a reliability ≥ 90%. It is important to notice that both the threshold (either t-percentile of the whole-tree distribution or a specific cutoff value) and the other constraints can easily be modified so that they conform to previously introduced definitions.

To extract clusters from a phylogeny, a depth-first algorithm was used, which is a general strategy for traversing a graph or a tree. In brief, after calculating the whole-tree distance distribution, the depth-first (Fig. 1) starts from a root node and moves into a sub-tree, checking its reliability, number of leaves and calculating the sub-tree distribution. If the clustering conditions are not met, the algorithm goes deeper into another child sub-tree, otherwise it stops, and the leaves of the current sub-tree are placed in a cluster. Then the search restarts from the most recent sub-tree that has not been visited. If a leaf is reached, no clusters are identified for that path.

![Figure 1 | Automated partition of a phylogenetic tree.](image)

**Figure 1 | Automated partition of a phylogenetic tree.** Graphical example of a depth-first tree search for automated phylogenetic tree partition. The method considers nodes/sub-trees with a reliability ≥ 90% and ≥ 2 distinct patients, recognizing a sub-tree as a cluster when the median sub-tree pairwise patristic distance is below a percentile threshold of the whole-tree patristic distance distribution (let it be the 10th percentile). (a) An example of a phylogenetic tree, where each patient/sequence is identified by a letter (A–Z) and each tree node has an associated value of reliability (which might be bootstrap support). (b) Histogram of the whole-tree patristic distance distribution. The vertical black line corresponds to the 10th percentile distance threshold. The partition method identifies three clusters (yellow, red and green) and discards the grey sub-tree.
To analyse the effect of the threshold \( t \), two independent data sets of different sample sizes (\( n = 127 \) and \( n = 1,204 \) sequences) were used, comprising HIV-1 subtype B \( pol \) sequences obtained either from patients belonging to epidemiologically confirmed transmission clusters (plus a set of control sequences) or by visually selecting the transmission chains from the phylogenetic tree. Our method yielded rates of agreement with the epidemiologically confirmed/visually determined transmission events up to 93%. The optimal threshold interval ranged from the 10th to the 35th percentile, with a corresponding absolute distance range of 0.04–0.08 nucleotide substitutions per site (Supplementary Fig. S1). Additional sensitivity analyses were conducted by comparing cluster sizes, cluster numbers, distance and branch length distribution of the study and the validation sets, and by comparing against randomly generated trees (Methods).

**Application of the partition method to HIV-1 phylogeny.** A total of 11,541 HIV-1 subtype B \( pol \) sequences from 7,350 patients were included in the analysis, plus two outgroups (subtype C and J). Multiple sequence alignment and maximum-likelihood phylogeny, excluding drug-resistance mutation sites, were carried out using parallel computation. Figure 2 shows the phylogenetic tree, whereas Figure 3 depicts the median and interquartile (IQR) values of branch lengths from a parent node to the child and of number of nodes across the tree levels, starting from the root. Branch lengths and node levels showed a minimal inverse correlation (\( r = -0.023, P < 0.0001 \)). There was also a low but statistically significant correlation between the calendar dates of the leaves and the root-to-tip distances (\( r = 0.25, P < 0.0001 \), Supplementary Fig. S2), confirmed on a multivariable analysis (Supplementary Table S1).

By applying our partition method, the number of clusters retrieved using three distinct absolute distance thresholds of 0.04, 0.07 and 0.08 nucleotide substitutions per site (corresponding to the 1st, 15th and 30th percentile thresholds) were 842, 587 and 292, respectively. For each threshold, each cluster had a median (IQR) number of 3 (2–6), 4 (2–9), and 4 (2–10) distinct patients, respectively (Supplementary Fig. S3). The proportion of clustered isolates was 41.71, 73.44 and 92.23%. Of note, at the 30th percentile threshold, about a half of the sequences were placed in two distinct clusters (\( n = 5,785 \) and \( n = 1,538 \)), both exceeding in size the number of un-clustered isolates (\( n = 897 \)). At thresholds below the 30th percentile, the maximum cluster size was always inferior to the number of un-clustered isolates (Supplementary Fig. S4). Thus, the 30th percentile threshold behaved as a break-point: all subsequent analyses yielded similar results across lower thresholds, whereas the same did not always hold when considering both thresholds above and below the break-point. The partitions retrieved at different thresholds exhibited a mild rate of agreement, with a peak at the 15th percentile threshold (Supplementary Fig. S5). Notably, two previously published cases\(^{5,26} \) of drug-resistance transmission from ART-naive to ART-naive patients in ARCA were correctly identified at any of the thresholds considered.

**Transmission clusters and factors associated with clustering.** The phylogeny partition was linked to the patients’ epidemiological and clinical information. Descriptive statistics of the study population is shown in Table 1.

Table 2 describes the composition of clusters when considering the threshold distance of 0.07 (15th percentile, highest rate of agreement). Sequences without epidemiological information were ignored, and only the earliest available sequence per patient was considered. Each cluster was evaluated checking if it was composed exclusively either by a single or by a mixture of different demographic factors. The estimated proportions were then compared against those that would be expected by chance after randomly shuffling the cluster indices. We found an excess in the proportion of clusters composed solely by patients living in the Northern Italy (31 versus 17%, \( P < 0.0001 \)), by those living in the Central Italy (25 versus 16%, \( P = 0.0008 \)), by male homosexuals (27 versus 16%, \( P < 0.0001 \)), by male homosexuals and heterosexuals together (20 versus 10%, \( P < 0.0001 \)), by patients infected recently (within three years from the first HIV-1 positive test, 35 versus 19%, \( P < 0.0001 \)), by ART-naive patients without any drug-resistance mutation (15 versus 6%, \( P < 0.0001 \)), by patients with high HIV-1 RNA load (above 10,000 copies per ml, 43 versus 34%, \( P = 0.02 \)), by patients carrying at least one resistance mutation to nucleoside/nucleotide reverse transcriptase inhibitors (NRTI) and protease inhibitors (PI)
### Table 1 | Study population.

| Variables | n | % |
|-----------|---|---|
| Sequences | 11,541 | 100 |
| Patients | 7,350 | 63.69 |
| Patients with known seroconversion date | 258 | 3.51 |
| Patients with known date of first HIV-1 positive test | 2,720 | 37.01 |
| **Country area** | | |
| Central Italy | 3,274 | 44.54 |
| Northern Italy | 3,528 | 48.00 |
| Southern Italy | 443 | 6.03 |
| Unknown | 105 | 1.43 |
| **Mode of HIV-1 transmission** | | |
| Injecting drug user | 1,418 | 19.29 |
| Heterosexual | 1,439 | 19.58 |
| Male homosexual | 1,313 | 17.86 |
| Other/unknown | 3,180 | 43.57 |
| **Gender** | | |
| Female | 1,677 | 22.82 |
| Male | 5,111 | 69.54 |
| Unknown | 562 | 7.65 |
| **Country of birth** | | |
| Italy | 4,210 | 57.28 |
| Other countries | 229 | 3.12 |
| Unknown | 2,911 | 39.60 |
| **ART, antiretroviral therapy; IQR, interquartile range; NNRTI, non-nucleoside reverse transcriptase inhibitors; nRTI, nucleoside/nucleotide reverse transcriptase inhibitors; PI, protease inhibitors.**

#### Characteristics of subtype B HIV-1 infected patients enrolled in the Italian ARCA cohort.

| Variables | Median | IQR |
|-----------|--------|-----|
| Sequence year | 2004 | 2002–2007 |
| Number of sequences per patient | 1 | 1–2 |
| **HIV-1 RNA Log_{10} copies per ml** | | |
| ART-naive | 4.64 | 4.04–5.23 |
| ART-experienced | 3.98 | 3.32–4.65 |
| **CD4 T cells per mm³** | | |
| ART-naive | 361 | 196–556 |
| ART-experienced | 334 | 191–504 |
| **Age (years)** | | |
| ART-naive | 37 | 31–44 |
| ART-experienced | 41 | 37–46 |
| **Time passed from the first HIV+ test (years)** | | |
| ART-naive | 0 | 0–1 |
| ART-experienced | 10 | 6–15 |
Table 2 | Demographic factors in transmission clusters.

| Factor strata | Clusters | Random | Adjusted P value |
|---------------|----------|--------|-----------------|
|                | n        | %      | n               | %      | < 0.0001 |
| Country area of residence |          |        |                 |        |          |
| Northern Italy | 183      | 31.28  | 96              | 17.30  | 0.0001   |
| Central Italy | 146      | 24.96  | 90              | 16.22  | 0.0341   |
| Southern Italy | 8        | 1.37   | 7               | 1.26   | 0.3898   |
| Northern and Central Italy | 158      | 27.01  | 230             | 41.44  | < 0.0001 |
| Northern and Southern Italy | 19       | 3.25   | 13              | 2.34   | 0.4504   |
| Central and Southern Italy | 9        | 1.54   | 12              | 2.16   | 0.5302   |
| Northern, Central and Southern Italy | 62       | 10.60  | 107             | 19.28  | 0.0011   |
| Mode of HIV transmission |          |        |                 |        |          |
| Male homosexual | 135      | 27.05  | 77              | 15.88  | < 0.0001 |
| Heterosexual | 93       | 18.64  | 72              | 14.85  | 0.1792   |
| IDU | 56       | 11.22  | 79              | 16.29  | 0.0415   |
| Male homosexual and heterosexual | 99       | 19.84  | 48              | 9.90   | < 0.0001 |
| Male homosexual and IDU | 13       | 2.61   | 46              | 9.48   | < 0.0001 |
| Heterosexual and IDU | 50       | 10.02  | 42              | 8.66   | 0.5564   |
| Male homosexual, heterosexual and IDU | 53       | 10.62  | 121             | 24.95  | < 0.0001 |
| ART status |          |        |                 |        |          |
| Naive | 99       | 19.08  | 48              | 9.78   | 0.0001   |
| Experienced | 227      | 43.74  | 231             | 47.05  | 0.0883   |
| Naive and experienced | 193      | 37.19  | 212             | 43.18  |          |
| Presence of at least one resistance mutation to NRTI |          |        |                 |        |          |
| Yes | 153      | 26.15  | 102             | 18.65  | 0.0069   |
| No | 156      | 26.67  | 87              | 15.90  | < 0.0001 |
| Yes and no | 276      | 47.18  | 358             | 65.45  | < 0.0001 |
| Presence of at least one resistance mutation to NNRTI |          |        |                 |        |          |
| Yes | 67       | 11.45  | 58              | 10.34  | 0.5998   |
| No | 243      | 41.54  | 183             | 32.62  | 0.0011   |
| Yes and no | 275      | 47.01  | 320             | 57.04  | 0.0012   |
| Presence of at least one resistance mutation to PI |          |        |                 |        |          |
| Yes | 61       | 10.43  | 34              | 6.09   | 0.0181   |
| No | 300      | 51.28  | 225             | 40.32  | 0.0007   |
| Yes and no | 224      | 38.29  | 299             | 53.58  | < 0.0001 |
| Presence of at least one drug-resistance mutation to any ARV class |          |        |                 |        |          |
| Yes | 188      | 32.14  | 161             | 28.50  | 0.2574   |
| No | 119      | 20.34  | 58              | 10.27  | < 0.0001 |
| Yes and no | 278      | 47.52  | 346             | 61.24  | < 0.0001 |
| Years passed from the first HIV-1 positive test |          |        |                 |        |          |
| Below 3 | 143      | 34.71  | 76              | 18.86  | < 0.0001 |
| Between 3 and 9 | 53       | 12.86  | 48              | 11.91  | 0.7235   |
| Between 9 and 14 | 33       | 8.01   | 49              | 12.16  | 0.0851   |
| Above 14 | 26       | 6.31   | 37              | 9.18   | 0.1964   |
| Mixtures with ≥2 factors | 157      | 38.11  | 193             | 47.08  | 0.0121   |
| HIV-1 RNA load (copies per ml) |          |        |                 |        |          |
| Below 1,000 | 19       | 3.85   | 28              | 6.38   | 0.1300   |
| Between 1,000 and 10,000 | 55       | 11.16  | 61              | 13.90  | 0.2834   |
| Above 10,000 | 212      | 43.00  | 151             | 34.40  | 0.0169   |
| Mixtures with ≥2 factors | 207      | 41.99  | 199             | 45.33  | 0.4017   |
| Mode of HIV transmission and gender |          |        |                 |        |          |
| Male homosexual | 130      | 26.69  | 76              | 15.90  | 0.0001   |
| Male heterosexual | 66       | 13.55  | 19              | 3.97   | < 0.0001 |
| Male homosexual | 45       | 9.24   | 40              | 8.37   | 0.6846   |
| Male IDU | 36       | 7.39   | 48              | 10.04  | 0.2173   |
| Male homosexual, heterosexual | 22       | 4.52   | 10              | 2.09   | 0.0649   |
| Male and female heterosexual, male homosexual | 21       | 4.31   | 8               | 1.67   | 0.0338   |
| Other mixtures | 167      | 34.29  | 277             | 57.95  | < 0.0001 |
| ART status and presence of at least one drug-resistance mutation to any ARV class |          |        |                 |        |          |
| Naive and no-resistance | 78       | 15.00  | 30              | 5.93   | < 0.0001 |
| Naive and resistance | 12       | 2.31   | 11              | 2.17   | 0.3898   |
| Naive, treated and no-resistance | 28       | 5.38   | 11              | 2.17   | 0.0169   |
| Naive, treated and resistance | 22       | 4.23   | 24              | 4.74   | 0.7247   |
| ART status and presence of at least one drug-resistance mutation to any ARV class |          |        |                 |        |          |

ART, antiretroviral therapy; ARV, antiretroviral; IDU, injecting drug user; NNRTI, non-nucleoside reverse transcriptase inhibitors; NRTI, nucleoside/nucleotide reverse transcriptase inhibitors; PI, protease inhibitors. Clusters compositions by different demographic factors, using a clustering threshold of 0.07 nucleotide substitutions per site. Observed proportions have been compared with a data randomization.
In various configurations with respect to ART-naive patients, cluster more than IDU, whereas the contrary held at higher thresholds. This might be explained by the fact that IDU have a larger pool of contacts as compared with homosexuals/heterosexuals and thus are more likely to be included in larger transmission chains characterized by greater genetic diversity.

**Discussion**

A new methodology for the partition of large-scale phylogenies and the inference of transmission clusters was introduced. This approach conjugates the evaluation of node reliability, tree topology and patristic distance analysis. Previous approaches have lacked, in general, not only a standardized definition of a transmission cluster, but also a formal algorithmic procedure for the detection of reliable transmission clusters within a phylogeny. In the studies by Lewis et al. and Hughes et al., clusters were identified on large data sets using a fixed threshold on genetic distances, and then confirmed by phylogenetic trees constructed afterwards. Another approach was the CTree algorithm, originally applied to viral sub-typing, which does not account for node reliability and is not applicable to large data sets in its current implementation. Other available methods have used different cluster selection schemes by performing nested phylogenetic analyses, and/or adding criteria for geographical consistency, but in most cases the assessment of transmission clusters is still subject to a visual tree inspection. The definition of a transmission cluster proposed here is general and can be tuned to accommodate any of the previous definitions. The method can be of interest for epidemiologists because obtained partitions could be used in multi-level modelling.

The depth-first approach was applied to a large study population, composed of viral sequences of HIV-1-infected patients followed up at the Italian clinics, with the aim to gain knowledge about viral transmission clusters and their associated factors. The transmission clusters detected in our study population were preferentially composed either by male homosexuals only or by male homosexuals and heterosexuals together. As male homosexuals (and presumably male bisexuals) are a primary source of the cluster composition, this could be considered for designing targeted population screenings and interventions. On the other hand, there was a negative association between male homosexuals/heterosexuals and IDU. We found that ART-naive patients cluster both with other ART-naive and with ART-experienced patients, in various configurations with respect to the carriage of drug-resistance mutation. Although the direction of the transmission cannot easily be inferred, the proportion of clusters composed exclusively by ART-naive patients carrying at least one drug-resistance mutation in our population study (so presumable transmitting the resistance to each other) was ~2%, a low proportion, comparable to a previous estimation in the United Kingdom. A substantial responsibility of drug-resistance transmission is still carried by the ART-experienced population, suggesting that any decrease in the prevalence of drug-resistance in the ART-experienced population could translate into a reduction of the proportion of new infections with drug-resistant virus.

Multivariable analysis revealed several factors independently associated with the transmission clustering, including a higher HIV-1 RNA load and CD4+ count, but not the presence of drug-resistance. These results partially agree with those reported from the Swiss cohort study, in which the strongest predictor of clustering was the detection of drug-resistance, along with HIV-1 RNA load and CD4+ count. However, that study population was composed by newly diagnosed individuals.

The association of a recent HIV-1 positive test with transmission clustering could be simply due to the fact that in these individuals the intra-host viral evolution is at an early stage and thus it is easier to trace the possible transmission source. Another possibility, however, is that some newly infected patients might not be aware of being infected, and therefore continue to engage in risky sexual behaviours. Recently, infected and untreated patients are also more likely to have higher viral loads, resulting in higher probability of transmission. Thus, efforts to expand HIV-1 testing programmes should be a successful strategy to reduce the rates of HIV-1 transmission. Concomitantly, the association of higher HIV-1 RNA loads and CD4+ counts with transmission clustering emphasizes the need of early diagnoses and suggests how early treatment may be useful in reducing HIV-1 transmission events.

This work has some limitations. First, epidemiological conclusions may change slightly or dramatically depending on the choice of the threshold. By evaluating the algorithm on different data sets, and using different definitions of a transmission cluster, the optimal threshold spanned a large range, from 0.04 to 0.08 nucleotide substitutions per site, whereas previous studies used more stringent criteria (either 0.015 or 0.031). Ideally, a transmission chain, at least in the context of a single subtype phylogeny of HIV-1 isolates, should be defined by an absolute threshold. At the moment, the lack of epidemiologically confirmed data precludes such a precise estimation. Nonetheless, the present method offers a rational validation strategy to identify an appropriate cutoff, depending on the specific data set under investigation and on the (strict or loose) definition of a transmission cluster. Second, complete demographic and clinical information were missing for a relevant proportion of patients enrolled in ARCA. A non-uniform sequence sampling has most likely occurred overtime. Most of the sequences included in our analysis were derived from ART-experienced patients tested at virological failure. Although repeated measures and convergent evolution due to treatments were accounted for, the analysis lacked a sample contribution from recently infected patients. Sampling could potentially distort results when analysing factors associated with transmission chains. Finally, the partition method is fully dependent on the phylogenetic tree: if the tree is unreliable, then so is the cluster identification. The estimation of highly resolved and reliable phylogenies from large data sets is still a challenge and requires massively parallelized routines, especially for maximum likelihood or Bayesian tree inference. For this reason, the approach based on nested phylogenies might be still a preferable choice.

Some limitations of phylogenetic efficiency in detecting transmission chains could also be overcome by coupling phylogeny inference with models derived from the complex network theory. For instance, a phylogeny could be used as a previous knowledge to infer sexual links when designing a social network model for HIV-1 infection.
The table below lists factors associated with transmission clustering.

| Factor/threshold | 1st Percentile (0.04 ns s\(^{-1}\)) | 15th Percentile (0.07 ns s\(^{-1}\)) | 30th Percentile (0.08 ns s\(^{-1}\)) |
|------------------|-------------------------------------|-------------------------------------|-------------------------------------|
|                   | OR       | 95% CI       | P value       | OR       | 95% CI       | P value       | OR       | 95% CI       | P value       |
| Calendar year of genotyping  (per more recent) | 1.04 (1.01–1.06) | 0.0011 | 0.98 (0.96–1.01) | 0.1163 | 0.98 (0.95–1.01) | 0.2626 |
| Italian country area of residence  (ref. Central) | Northern  | 0.52 (0.46–0.6) | < 0.0001 | 0.73 (0.63–0.85) | < 0.0001 | 0.109 (0.85–1.39) | 0.4936 |
| Southern  | 0.58 (0.45–0.75) | < 0.0001 | 0.78 (0.6–1.01) | 0.0578 | 1.07 (0.68–1.68) | 0.7860 |
| Unknown  | 0.86 (0.56–1.34) | 0.5151 | 1.05 (0.63–1.76) | 0.8583 | 0.65 (0.32–1.32) | 0.2341 |
| Mode of HIV-1 transmission  (ref. IDU) | Heterosexual | 1.64 (1.29–2.09) | < 0.0001 | 1.08 (0.86–1.16) | 0.4993 | 0.27 (0.17–0.44) | < 0.0001 |
| Male homosexual | 1.58 (1.24–2.01) | 0.0002 | 1.05 (0.82–1.33) | 0.7112 | 0.16 (0.01–0.25) | < 0.0001 |
| Other/unknown | 1.52 (1.19–1.94) | 0.0007 | 1.05 (0.84–1.32) | 0.6658 | 0.33 (0.21–0.52) | < 0.0001 |
| Gender  (ref. female) | Male  | 1.88 (1.56–2.26) | < 0.0001 | 1.00 (0.83–1.19) | 0.9564 | 0.54 (0.4–0.73) | < 0.0001 |
| Unknown  | 2.27 (1.73–2.97) | < 0.0001 | 1.08 (0.82–1.42) | 0.5801 | 0.52 (0.32–0.83) | 0.0058 |
| Country of birth  (ref. Italy) | Other than Italy | 0.64 (0.45–0.91) | 0.0142 | 0.53 (0.37–0.75) | 0.0004 | 0.28 (0.18–0.43) | < 0.0001 |
| Unknown  | 1.03 (0.86–1.23) | 0.7390 | 1.13 (0.93–1.36) | 0.2096 | 1.12 (0.82–1.53) | 0.4867 |
| Age  (ref. > 46 years old) | ≤36  | 1.49 (1.23–1.81) | < 0.0001 | 1.20 (0.98–1.47) | 0.0790 | 1.49 (1.1–2.04) | 0.0111 |
| > 36 and ≤41 | 1.06 (0.88–1.26) | 0.5617 | 1.00 (0.83–1.21) | 0.9565 | 1.70 (1.26–2.29) | 0.0006 |
| > 41 and ≤46 | 0.99 (0.84–1.18) | 0.9366 | 1.00 (0.84–1.19) | 0.9965 | 1.44 (1.1–1.9) | 0.0082 |
| Unkown  | 1.02 (0.82–1.26) | 0.8910 | 1.10 (0.87–1.38) | 0.4303 | 1.50 (1.01–2.21) | 0.0424 |
| ART status (ref. ART-naive) | ART-experienced | 0.68 (0.56–0.84) | 0.0003 | 1.08 (0.85–1.37) | 0.5361 | 1.18 (0.82–1.7) | 0.3708 |
| Unknown  | 0.72 (0.59–0.89) | 0.0020 | 1.04 (0.82–1.31) | 0.7447 | 0.93 (0.65–1.32) | 0.6677 |
| Time from first HIV + test  (ref. > 14 years) | ≤3 years  | 2.56 (1.87–3.51) | < 0.0001 | 1.78 (1.3–2.45) | 0.0004 | 1.33 (0.75–2.36) | 0.3261 |
| > 3 and ≤9 years | 1.70 (1.27–2.26) | 0.0003 | 1.14 (0.87–1.5) | 0.3524 | 0.87 (0.52–1.45) | 0.6022 |
| > 9 and ≤14 years | 1.30 (1–1.7) | 0.0537 | 1.18 (0.93–1.51) | 0.1804 | 0.99 (0.62–1.57) | 0.9519 |
| Unknown  | 1.38 (1.04–1.81) | 0.0223 | 0.90 (0.7–1.17) | 0.4394 | 0.77 (0.47–1.24) | 0.2809 |
| Presence of at least one drug-resistance mutation* | NRTI  | 0.68 (0.6–0.78) | < 0.0001 | 0.72 (0.62–0.82) | < 0.0001 | 0.99 (0.8–1.24) | 0.9378 |
| NNRTI  | 0.93 (0.83–1.04) | 0.2271 | 0.99 (0.88–1.12) | 0.8875 | 0.95 (0.78–1.16) | 0.6196 |
| PI  | 0.93 (0.82–1.07) | 0.3092 | 0.80 (0.7–0.92) | 0.0012 | 0.77 (0.61–0.97) | 0.0274 |
| HIV-1 RNA per one Log\(_{10}\) copies per ml higher | 1.08 (1.01–1.15) | 0.0344 | 1.01 (0.94–1.09) | 0.7574 | 1.04 (0.92–1.17) | 0.5722 |
| CD4+ count per 50 cells/mm\(^3\) higher | 1.03 (1.01–1.04) | 0.0002 | 1.02 (1.01–1.04) | 0.0057 | 1.03 (1–1.06) | 0.0616 |

ART, antiretroviral therapy; CI, confidence interval; IDU, injecting drug user; NNRTI, non-nucleoside reverse transcriptase inhibitors; NRTI, nucleoside/tide reverse transcriptase inhibitors; ns s\(^{-1}\), nucleotide substitutions per site; OR, odds ratio; PI, protease inhibitors.

*Mixtures of wild-type and resistant mutants are classified as resistant

Adjusted OR of transmission clustering evidence (clustered versus un-clustered isolates) from fitting a multivariable logistic generalized-estimating-equations model, by considering different percentile thresholds.

The analysis of transmission clusters using the method discussed herein can easily be extended to non-B subtype HIV-1, whose circulation is increasing in the Italy and Europe\(^{12}\), as well as other viral epidemics, such as hepatitis B or C viruses (HBV/ HCV), influenza and coronavirus infection. In particular, HCV is one of the fastest growing pandemics and its evolutionary rate, genetic heterogeneity and risk transmission factors are similar to HIV-1. Given the upcoming availability of HCV antivirals, surveillance of drug-resistance and analysis of transmission clusters using standardized algorithms are expected to be of outstanding interest in the near future. Finally, the algorithm described here can potentially be used in the study of intra-host phylogenies—especially for the characterization of infected cellular reservoirs and compartments in viruses such as HIV\(^{33,36}\) or simian immunodeficiency virus\(^{37}\)—in terms of viral fitness, emergence of drug resistance and pathogenesis.

**Methods**

**Study population.** HIV-1 pol gene sequences (encompassing at least amino acidic positions 1–99 of protease and 1–250 of reverse transcriptase) of ART-naive and ART-experienced patients were extracted from ARCA, with sampling dates, plus corresponding information (where available) on patients’ age, gender, country of origin, country area of residence, mode of HIV-1 transmission, date of first HIV-1-positive test, date of first ART, HIV-1 RNA load and CD4+ T cell count contemporary to the sampling date (between -30/+7 days for HIV-1 RNA, between -30/+30 days for CD4 +).

**Sequence analysis.** Viral subtype was assigned by using the Rega subtyping tool (http://www.bioafrika.net/rega-genotype/html/subtypinghiV.html). All sequences classified as subtype B, excluding any putative recombinant form, were aligned...
using the parallel implementation of ClustalW\(^\text{\textregistered}\), and manually edited. Resistance to an antiretroviral class (nucleoside-tide/non-nucleoside/protease inhibitors) was defined as the presence of at least one major mutation conferring resistance to at least one drug belonging to that class according to the 2009 update of the Internationa\(\text{\textregistered}\) AIDS Society reference list (http://www.lanl.gov/content/mutations/\text{\textregistered}\_mutations\_figures.pdf). Deduced amino-acid mixtures including a major resistan\(\text{\textregistered}\) mutation were considered as indicative of resistance, unless otherwise specified. Columns of the multiple alignment corresponding to codon positions associated to drug-resistance were removed.

**Phylogenetic analysis.** Maximum-likelihood phylogenetic analysis was performed on the alignment, adding HIV-1 subtypes J and C as outgroups. The parallel implementation of FastTree software\(^\text{\textregistered}\) was used, setting up a general-time-reversible model, with a 20-parameter gamma optimization, and a mix of nearest-neighbour interchanges and sub-tree/prune/graft moves for tree topology search. Similarity of each tree split was calculated by a Shimodaira–Hasegawa test. To visualize the large tree as a whole, three-dimensional hyperbolic geometry software was used (http://www.caida.org/tools/visualization/walrus/).

Parallel ClustalW multiple alignment required ~10 h on a two quad-core 64-bit Intel Xeon X5550@2.66GHz, with Hyper-Threading technology, using a total of 16 processing units and 24GB DDR3 RAM. Parallel FastTree software ran for ~40 min.

**Automated partition of phylogenies.** After obtaining the maximum-likelihood phylogenetic tree, rooted on HIV-1 subtype J, tree topology was analysed with a depth-first visit\(^\text{\textregistered}\) by considering the number of sub-trees with a node reliability ≥ 90%, and an associated number of leaves with at least two distinct patients. At each step of the depth-first visit, a sub-tree was identified as a cluster if the median value of the sub-trees divergence was below a t-percentile of the whole-tree distance distribution. If this condition was met in a node, the search at that node was stopped, ignoring the children nodes, passing to analyse other node siblings. The threshold \(t\) was evaluated and optimized over the range [5th, 50th] percentile of the whole-tree distance distribution, with a step of 0.05. Cluster partitions were compared by using the adjusted Rand index\(^\text{\textregistered}\).

**Software implementation and computational complexity.** Software has been implemented using the java (http://java.sun.com/) programming language, released as a free platform-independent standalone application, executable both from the command-line and with a graphical user interface (downloadable as a Supplementary Software 1, and Supplementary Note 1, Supplementary Data 1 and 2, and Supplementary Fig. S6). The software requires a phylogenetic tree input file in newick format. Three different options are available: (i) threshold-based partition algorithm, specifying a threshold value and a maximum number of distance comparisons; (ii) topological analysis, with median (IQR) values for branch length, reliability and number of nodes at each tree level; (iii) patristic distance calculator.

In the partition algorithm (i), the calculation of the branch length difference distribution for a tree with \(k\) leaves, requires \(k(k-1)/2\) pairwise comparisons, corresponding to 66.6 million for the ARCA phylogenetic tree. In the worst case, the depth-first search may require a number of accesses to the distance matrix, which is cubic in the number of entries. However, the implemented software has an efficiency routing index that permit the exact calculation of a whole-tree distance matrix even for huge trees, at a price of a high random access memory usage. If the java virtual machine is initialized with at least 8GB of RAM, it is possible to compute exactly a full matrix for a tree with ~10,000 leaves. Our routine for patristic distance calculation was validated by comparing it against the `cophenetic` function of the `ape` library in the R software (http://www.R-project.org).

**Validation of partition method and threshold optimization.** The depth-first visit procedure for the automated partition of a phylogenetic tree was evaluated by using two different independent data sets.

First, a maximum-likelihood phylogenetic tree was estimated (with the same alignment and tree fit procedures as previously described) by considering a set of group M subtype B HIV-1 pol sequences obtained from patients with a confirmed and known transmission history, plus a set of random control sequences from distinct patients of the same geographical area, and two outgroups (HIV-1 subtypes J and C). This data set was produced by querying the Los Alamos HIV repository (http://www.hiv.lanl.gov/content/index), in which 62 sequences from the pol dataset were reported to be involved in five distinct transmission events. The number of control sequences was the same as that of sequences linked to the transmission events. Thus, a total of 127 sequences were considered in the subsequent phylogenetic analysis. The phylogenetic tree (Supplementary Fig. S7) was also visually inspected to confirm the epidemiological evidence of transmission clustering, as follows.

Identification of clusters was performed by examining leaves and nodes exhaustively. Only at a most recent common ancestor (MRCA) node with a reliability ≥ 90%, all the children leaves were iteratively grouped together, maximizing the cluster size, ensuring that any group included isolates from at least two distinct patients. If in a theoretical cluster with a reliability of ≥90% were scored as transmission chains, then any highly supported epidemiological cluster with several nested highly supported transmission chains would also be considered as a transmission chain. However, if two potential transmission clusters happened to share a MRCA with a reliability value of ≥ 90%, but the two lineages departing from that MRCA exhibited long branch lengths, then they were considered as two independent transmission chains. Leaves that did not meet these criteria were left un-clustered.

All transmission events were identified as distinct clusters, although some of the control sequences were placed in one or more of the five transmission clusters and additional clusters were identified (n = 12). The adjusted Rand index that compared the visual inspection of the phylogenetic tree against the epidemiological evidence was equal to 0.55. The depth-first procedure was executed optimizing the threshold value over a percentile interval range (5th, 50th) of the whole-tree distance distribution, with a step of 0.05, obtaining a maximum adjusted Rand index of 0.69 (11 clusters) at \(t = 100\) (0.04 nucleotide substitutions per site) percentile with respect to the epidemiological evidence and of 0.93 (10 clusters) at \(t = 20\) (0.062 nucleotide substitutions per site) percentile with respect to the visual clustering (Supplementary Fig. S1).

Successively, another maximum-likelihood phylogenetic tree was estimated on the whole set of group M subtype B HIV-1 pol sequences from the Los Alamos HIV repository, adding subtypes J and C as outgroups (n = 1,204), and transmission clusters were inferred by means of a visual inspection. The highest adjusted Rand index between the depth-first method and the visual inspection was 0.88, corresponding to a threshold of 0.08 nucleotide substitutions per site, that is, the 35th percentile (Supplementary Fig. S1). As previously found\(^\text{\textregistered}\), the proportion of sequences that clustered together from the same country was significantly higher than expected by chance.

We also compared the distributions of patristic distances and branch lengths in the validation sets and in the ARCA data set. Although significantly different in the sample size, the same threshold percentile (95th percentile) showed a remarkably low absolute difference in the average patristic distance values, which was below 0.004 nucleotide substitutions per site. The same held when considering the average branch length, with an absolute difference of 0.018 (Supplementary Fig. S8).

As a final evaluation, we compared the distributions of cluster number and cluster sizes across different percentile thresholds by considering several subtypes B HIV-1 trees estimated on data sub-samples, against randomly generated trees (at different birth rates). At all thresholds below the 45th percentile, both the average number of clusters and the cluster size of real trees were significantly different from the values obtained by a random tree set (Supplementary Figs S9 and S10).

**Epidemiological analysis.** Multivariate regression analysis was performed to identify prognostic factors of belonging to a transmission cluster (using a binomial link), and of root-to-tip distance (using linear regression). Generalized-estimating-equations models were used, using the `geepack` library in the R software, with patient codes as identifiers. Covariate of interests were: sequence year (numeric), presence of for nucleoside-tide/non-nucleoside/protease inhibitor resistance mutations (binary), patient’s area of residence (Northern/Southern/Central Italy or unknown), mode of HIV-1 transmission (heterosexual, male homosexual, IDU, other, unknown), gender (male/female/unknown), country of origin (Italian/non-Italian/unknown), age (categorized on the IQR range or unknown), ART status (ART-experienced/ART-naive/unknown), time passed from the first HIV-1 positive determination (categorized on the IQR range or unknown), HIV-1 RNA load and CD4+ cell count (numeric) at the time of genotyping.

**References**

1. Perrin, L., Kaiser, L. & Yerby, S. Travel and the spread of HIV-1 genetic variants. *Lancet Infect. Dis.* 3, 22–27 (2003).
2. Gray, R. R. et al. Spatial phydynamics of HIV-1 epidemic emergence in east Africa. *AIDS* 23, F9–F17 (2009).
3. Resik, S. et al. Limitations to contact tracing and phylogenetic analysis in establishing HIV type 1 transmission networks in Cuba. *AIDS Res. Hum. Retroviruses* 23, 347–356 (2007).
4. Brown, A. E. et al. Phylogenetic reconstruction of transmission events from individuals with acute HIV infection: toward more-rigorous epidemiological definitions. *J. Infect. Dis.* 199, 427–431 (2009).
5. Hué, S., Clewley, J. P., Cane, P. A. & Pillay, D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 18, 719–728 (2004).
6. Wilson, D. P., Law, M. G., Grulich, A. E., Cooper, D. A. & Kaldor, J. M. Relation between HIV viral load and infectiousness: a model-based analysis. *Lancet* 373, 492–499 (2004).
7. Wilson, D. P., Law, M. G., Grulich, A. E., Cooper, D. A. & Kaldor, J. M. Relation between HIV viral load and infectiousness: a model-based analysis. *Lancet* 373, 492–499 (2004).
8. Marcellin, A. G. et al. Detection of HIV-1 RNA in seminal plasma from untreated patients with undetectable HIV-1 RNA in blood plasma. *AIDS* 22, 1677–1679 (2008).
9. Attia, S., Egger, M., Mueller, M., Zwahlen, M. & Low, N. Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. *AIDS* 23, 1397–1404 (2009).
10. Wensing, A. M. et al. Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. *J. Infect. Dis.* 192, 958–966 (2005).
10. Pillay, D. et al. The impact of transmitted drug resistance on the natural history of HIV infection and response to first-line therapy. AIDS 20, 21–28 (2006).
11. Di Giambenedetto, S. et al. Decline prevalence of HIV-1 drug resistance in treatment-failing patients: a clinical cohort study. Antivir. Ther. 12, 835–839 (2007).
12. Marks, A. J., Pillay, D. & Mclean, A. R. The effect of intrinsic stochastics on transmitted HIV drug resistance patterns. J. Theor. Biol. 262, 1–13 (2010).
13. Brenner, B. G. et al. Transmission networks of drug resistance acquired in early stage HIV infection. AIDS 22, 2905–2911 (2008).
14. Recordon-Pinson, P. et al. HIV type 1 transmission dynamics in recent seroconverters: relationship with transmission of drug resistance and viral diversity. Antivir. Ther. 14, 551–556 (2009).
15. Lewis, F., Hughes, G. J., Rambaut, A., Romaniuk, A. & Leigh Brown, A. J. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 7, e1000389 (2010).
16. Hué, S., Gifford, R. J., Dunn, D., Fernhill, E. & Pillay, D. Demonstration of sustained drug-resistant human immunodeficiency virus type 1 lineages circulating among treatment-naïve individuals. J. Virol. 83, 2645–2654 (2009).
17. Hughes, G. J. et al. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog. 5, e1000909 (2010).
18. Fisher, M. et al. Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. AIDS 24, 1739–1747 (2010).
19. Yerly, S. et al. The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection. AIDS 23, 1415–1423 (2009).
20. Koopman, J. S. et al. Macroparasite epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. J. Infect. Dis. 201, 1488–1497 (2010).
21. Ragonnet-Cronin, M. et al. Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics. J. Acquire. Immune. Defic. Syndr. 55, 102–108 (2010).
22. Bezemédo, D. et al. Transmission networks of HIV-1 among men having sex with men in the Netherlands. AIDS 24, 271–282 (2010).
23. Wicker, N., Perrin, G. R., Thierry, J. C. & Poch, O. Seactor: a program for inferring protein subfamilies from phylogenetic trees. Mol. Biol. Evol. 18, 1435–1441 (2001).
24. Archer, J. & Robertson, D. L. Ctree: comparison of clusters between phylogenetic trees made using Bioinfomart. Bioinf. 23, e29–e38 (2007).
25. Razzolini, F. et al. Three-class-resistant human immunodeficiency virus type 1 variant in a drug-naive heterosexual couple. J. Clin. Microbiol. 46, 3856–3859 (2008).
26. Romano, L., Venturi, G., Vivarelli, A., Galli, L. & Zazzi, M. Detection of a drug-resistant human immunodeficiency virus variant in a newly infected heterosexual couple. Clin. Infect. Dis. 54, 116–117 (2007).
27. Salemi, M. et al. Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected users. Proc. Natl Acad. Sci. USA 96, 13233–13238 (1999).
28. Goodreau, S. M. Assessing the effects of human mixing patterns on HIV infection and response to first-line therapy. AIDS 20, 21–28 (2006).
29. Price, M. N., Delah, D. S. & Arkin, A. P. FastTree 2—Approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e10810 (2010).
30. Knuth, D. E. The Art of Computer Programming 3rd edn, Vol 1 (Addison-Wesley, 1997).
31. Hubert, L. & Arabie, P. Comparing partitions. J. Classif. 2, 193–218 (1985).
Bonora (Torino—Malattie Infettive Amedeo di Savoia), Palma Delle Foglie (Trento—Malattie Infettive), Cristina Rossi (Treviso—Malattie Infettive e Tropicale), Elena Seminari (Varese—Virologia), Federica Poletti (Verbania—Malattie Infettive Verbania), Vincenzo Mondino (Varese—Virologia), Marina Malena (Verona—Centro di Medicina Preventiva-ULSS 20), Emanuela Lattuada (Verona—Malattie Infettive).

Author contributions
Manuscript writing, epidemiological analysis, ideation of partition algorithm were done by M.C.F.P.; M.C., M.S., R.R.G. performed the phylogenetic analysis, theoretical support, manuscript review and editing; I.F. carried out installation and execution of parallel computation software for sequence alignment and phylogenetic analysis, implementation of partition algorithm and graphical user interface; F.S., M.P., V.B., S.Di G., B.B., A.C., A.V., S.R., M.C.R., M.R.G., L.S. contributed to data collection, clinical data collection, molecular sequencing, data monitoring, coordination with local centres participating the ARCA cohort, manuscript evaluation; M.Z. carried out molecular sequencing, manuscript editing; A.De L. designed the study, manuscript editing and project leading.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: A.De L. received speakers' honoraria, served as consultant or participated in advisory boards for GlaxoSmithKline, Gilead, Bristol-Myers Squibb, Abbott Virology, Janssen-Cilag Tibotec, Siemens Diagnostics and Monogram Biosciences. S.R. has received research grants and has been involved in advisory boards or educational courses supported by the following companies: Abbott, Boehringer-Ingelheim, Bristol-Myers Squibb, Gilead, GlaxoSmithKline, ViV Healthcare, Merck, and Janssen-Cilag. M.Z. has received research funding from Pfizer; served as a consultant for Abbott Molecular, Boehringer Ingelheim, Gilead Sciences and Janssen-Cilag; and served on speakers' bureaus for Abbott, Bristol-Myers Squibb, Merck, and Pfizer. The remaining authors declare no competing financial interests. B.B. has received funds for speaking, consultancy and travel from ViV Healthcare, Gilead Sciences, Abbott Molecular, Janssen-Cilag and Siemens Health Care.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Prosperi, M.C.F., et al. A novel methodology for large-scale phylogeny partition. Nat. Commun. 2:321 doi: 10.1038/ncomms1325 (2011).