Supplemental Digital Content

1. Description of the German seroconverter study

The German HIV-1 seroconverter study is a nationwide, multicenter, open, prospective, long-term observational cohort study which was initiated in 1997. Only HIV-1-infected individuals (≥ 18 years old) of whom the date of HIV-1 seroconversion is documented by laboratory diagnostic methods and who gave an informed consent are enrolled to the HIV-1 seroconverter cohort. Ethical approval was first received in 2005 and amended in 2013 by the ethics committee of the Charité University Medicine Berlin. Study patients were recruited by 22 clinics, 40 medical practices specialized in the care of HIV-1 patients, and seven public health offices. Epidemiological, clinical, laboratory diagnostic results and treatment data are collected at enrolment and at follow-up visits using standardized questionnaires. Inclusion criteria are seroconversion documented by a) a last negative and a first immunoblot-confirmed positive antibody test maximally 3 years apart ("documented seroconverter") or b) a first reactive test ("acute seroconverter"). The first reactive test of acute seroconverters is defined by the following laboratory diagnostic criteria: (1) detectable HIV-1 RNA or p24 antigen combined with a negative or indeterminate ELISA result or (2) a reactive HIV-1 ELISA combined with a negative or indeterminate immunoblot result and confirmation of seroconversion within six months. Note, that in the present study besides acute seroconverters only documented seroconverters were included of whom the maximal time period between the last negative and the first confirmed antibody tests was 12 months.

The arithmetic mean of the blood sampling dates for the last negative and first confirmed positive HIV-1 antibody test (documented HIV-1 seroconverter) or the blood sampling date for the first reactive test (acute HIV-1 seroconverter) is considered as the date of seroconversion and as the best proxy for the date of infection. The duration of infection is calculated as the difference between the date of blood sampling and the date of seroconversion.

2. Analysis of cluster size distribution

The analysis of the cluster size distribution was based on the power-law testing method proposed by Clauset *et al.*[1]. To this end, first a parameter $x_{\min} > 0$ is estimated, which defines the lower bound on the data exhibiting a discrete power-law behavior and then the scaling parameter α of the power-law distribution function given by $p(x) = x^{\alpha}$ is inferred. In order to test how plausible it is to fit a power-law model to the observed cluster size distribution, artificial sets of clusters are repeatedly generated from the power-law distribution, which are comparable in size with the observed set of clusters. Their distribution is parameterized by the values $x_{\min}$ and α estimated in the previous step. For each artificial distribution of clusters the distance to the true power-law distribution is computed using the Kolmogorov-Smirnov test (KS-test) and it is compared to the distance between the observed distribution of clusters and the true power-law. The relative amount of instances where the distance of the observed cluster size distribution to the true power-law is smaller than the distance of the artificial cluster data to the true power-law gives rise to a *P*-value. If this *P*-value is sufficiently large (here we used $P \geq 0.1$ in line with Clauset *et al.* [1]) then the power-law hypothesis cannot be rejected. If the *P*-value from the comparison of KS-statistics is smaller than the predefined threshold then the power-law hypothesis is considered not to be plausible. Note the different meaning of the *P*-value in this method, which, in contrast to usual null-hypothesis testing, is required to be large. For more details of this method see the paper of Clauset *et al.*[1]. For our analysis we used the Matlab code provided by Aaron Clauset (see http://tuvalu.santafe.edu/~aaronc/powerlaws/).

In addition to fitting a power-law model (described in the main manuscript), we tested if alternative distributions may yield a better fit to the cluster size data (Fig. S1). Using the method of Vuong[2], which quantifies if the log likelihood-ratio is significantly far from zero, we tested exponential, Poisson, Waring and Yule as alternative distributions. Starting from the smallest clustering threshold where the power-law distribution is plausible (θ =3%, see Fig. 1, right panel, main manuscript) we computed the log-likelihood ratios and *P*-values for the log-likelihood test. As shown in Fig. S1, the power-law distribution yielded a significantly better fit (greater log-likelihood ratio) than the exponential model for $\theta \geq 9\%$ ($P < 0.05$) and for $\theta \geq 7.5\%$ ($P < 0.1$). In addition, the power-law distribution exhibited a significantly better fit than

Poisson for $\theta \geq 5.5\%$ ($P < 0.05$). Below the indicated clustering thresholds, we did not identify a significantly higher likelihood for any model compared to the other, which suggests that a minimal amount of clusters is required for differentiating between alternative models. Similarly, power law provided a better fit as compared to the Waring distribution for a selected range of thresholds, while for a range of intermediate thresholds the two distributions could not be distinguished. In contrast, we could not determine a significant difference between power law and the Yule distribution independent of the choice of clustering threshold (see Fig. S1).

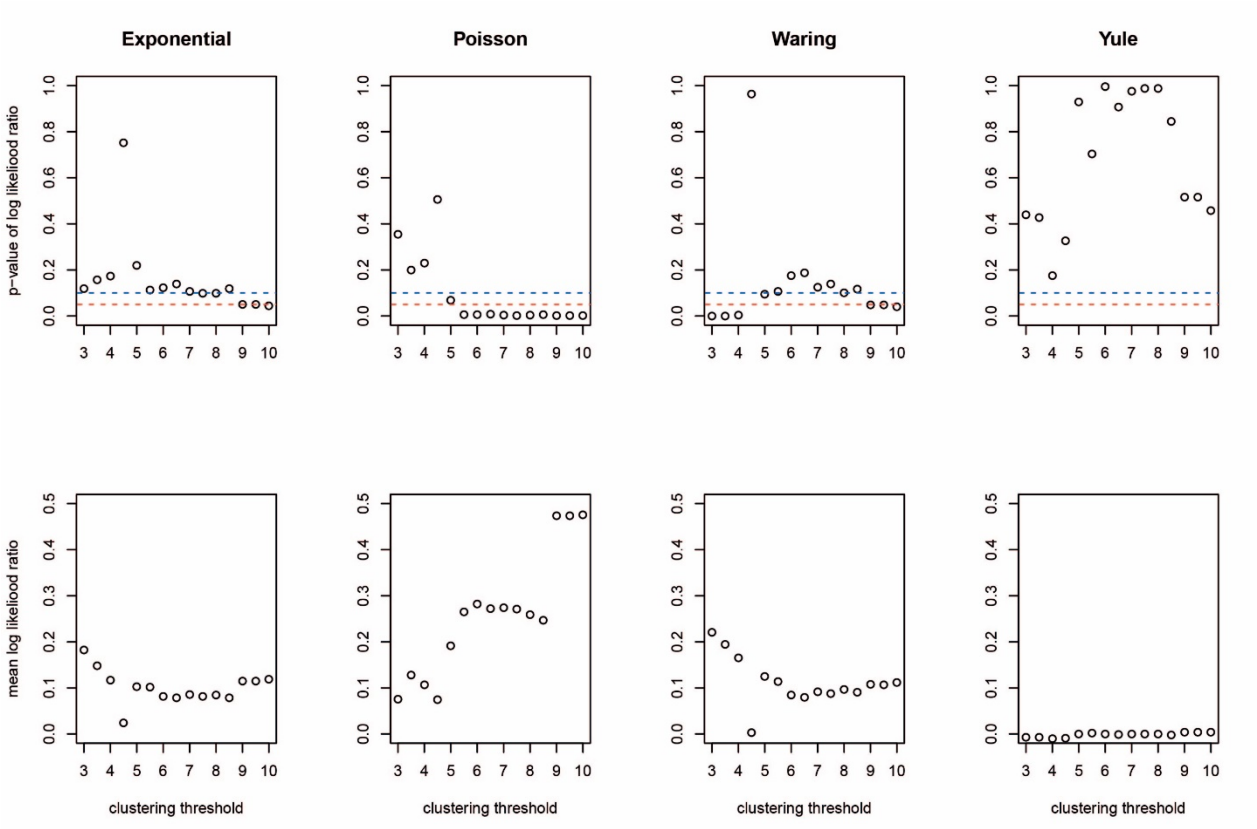

Figure S1: Analysis of the cluster size distribution. Top panel: *P*-values for the log-likelihood test of the power-law model vs. alternative distributions, as indicated. The blue line shows the significance level of 0.1 and the red line indicates the more conservative significance level of 0.05. **Bottom panel:** Values of the log-likelihood ratios between power-law and alternative distributions, ordered according to the top panel.

3. Hierarchical clustering and optimization of the clustering threshold

Phylogenetic clustering of viral sequences was conducted using a breadth-first search from the tree root to the leaves. Starting from the root, the algorithm recursively descends the phylogenetic tree along its inner nodes. For each node on this path it is first determined if it has at least 95% bootstrap support. If this is the case, it is then checked if the clade belonging to this node (i.e. all taxa for which this node is the putative most recent common ancestor) contains viral sequences from at least two different individuals and the mean of pairwise evolutionary distances μ between all individuals within this clade does not exceed a specified threshold θ_d. To this end, the mean pairwise divergence of a set of individuals belonging to a clade was computed according to

$$\mu = \binom{|\boldsymbol{N}|}{2}^{-1} \sum_{i,j \epsilon N} d_{ij},$$

where N denotes the set of (at least two) individuals whose sequences are included in a clade, $|N|$ denotes the size of this set. Furthermore, d_{ij} denotes the pairwise evolutionary distance between each two individuals i and j defined by the minimal patristic distance of any two sequences belonging to these two individuals. The search along the current descending path is stopped if all criteria are fulfilled and the individuals whose viral sequences descend from the current inner node are considered as a cluster. Otherwise the fulfilment of the criteria is recursively checked for each child node of the current node. As soon as the leaves of the tree are reached, the search is stopped indicating that no clusters are found along the current path. Similarly, the search is stopped without identifying a cluster if the clade belonging to the current node only consists of sequences from a single individual.

In order to obtain an optimal threshold θ_d, we assessed the clustering quality for a range of different thresholds. The challenge here is to find the balance between quantity and quality of putative transmission clusters. Within each cluster we measured the degree of uniqueness of the assignment of viral sequences of an individual to this cluster on the basis of the silhouette score for each individual i defined as:

$$\sigma_i = \frac{\min_k(b_{i,k}) - a_i}{\max\left(\min_k(b_{i,k}), a_i\right)},$$

with a_i denoting the average of minimal distances between viral sequences of individual i to viral sequences of all other individuals in the same cluster and $b_{i,k}$ denoting the average of minimal sequence distances of individual i to all individuals in a cluster $k \neq i$. Depending on the assignment quality of each individual, the silhouette score may range between -1 (completely wrong cluster assignment) and 1 (perfect cluster assignment). In order to obtain a representative measure for the distribution of silhouette scores of all individuals in clusters we used the minimal silhouette score $\sigma_{min} = \min(\sigma_i)$, i.e. a measure of the uniqueness of the most undetermined affiliation of an individual to a cluster.

Due to the hierarchical structure of the phylogenetic tree topology, smaller distance thresholds result in clusters tending to higher silhouette scores (see Figure 2 A, main manuscript) since these clusters consist of fewer sequences with a high similarity. In addition, small distance thresholds result in a small amount of clusters. In order to find a suitable trade-off between the silhouette score and the amount and size of clusters, we used a second optimality criterion based on the follow-up viral sequences of individuals. For our dataset (treatment-naïve *pol* sequences), we suspected weak diversifying selection pressure in contrast to several other studies [3-6]. Note that these studies [3-5] analyze within- and between-host evolution of HIV based on the *env* region of the HIV genome (coding for the HIV-1 surface proteins). The *env* region is under strong diversifying selection pressure in response to the adaptive immune system. In contrast, *pol* is expected to have a much lower observable evolutionary rate within the host in the absence of therapy targeting *pol*-encoded enzymes because the fitness of these enzymes could be near-optimal at the time of transmission. See also the reference[7] for an in depth comparison of evolutionary dynamics in different regions of the HIV-1 genome in untreated individuals. Note that although the recent paper[6] analyzes *pol* evolution, it considers sequences from treated individuals (implying a strong diversifying selective pressure). All of our samples originated from treatment-naïve individuals (see Material & Methods in main manuscript). To further confirm the absence of strong diversifying selection pressure, we analyzed the dN/dS ratio (ratio of non-

synonymous vs. synonymous substitutions), which was ≤ 1 in 73% of all follow-up sequence pairs, in line with[7]. In addition, the overall divergence was quite low in follow-up sequences of the same individual, as shown in Fig. S2. Acknowledging that longitudinal sequence data approximates the evolutionary dynamics of directly descended viruses, we assumed that we may use the corresponding divergence rate as a benchmark for clustering closely related viruses from different individuals (see also the reference in[8] which follows a similar idea).

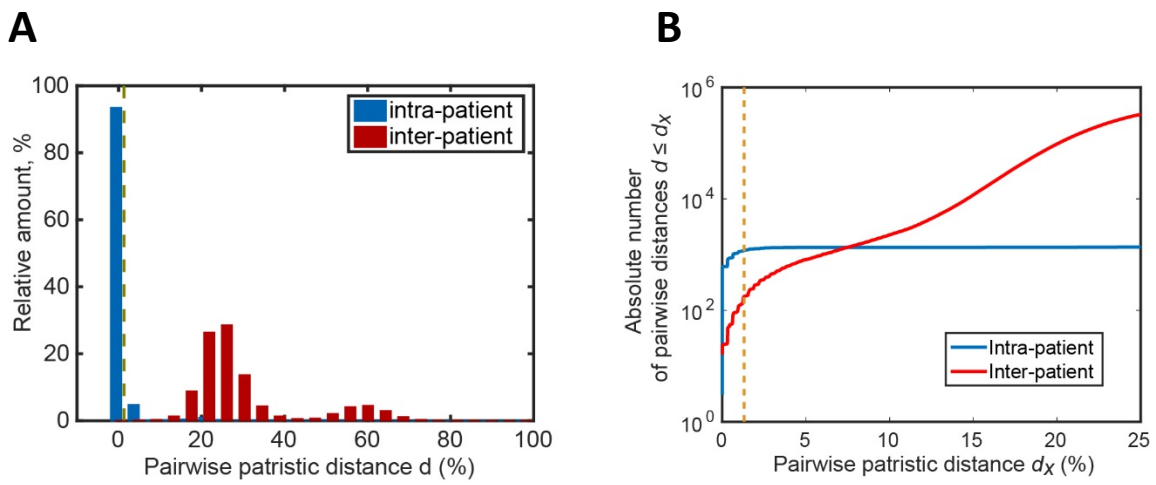

Figure S2: Comparison of intra- and interpatient sequence distances. A: Distribution of tree-based intra- and inter-patient sequence distances (blue and red, respectively). The dashed line indicates the threshold optimizing clustering modularity and the follow-up inclusion score (intra-patient distances). This optimal threshold was used within this study to infer the geographical and temporal spread of HIV. **B:** Cumulative number of pairwise sequence distances below the value indicated on the x-axis. Depicted is the (log10) absolute amount (y-axis) of distances not larger than a certain distance indicated on the x-axis.

Based on this observation, the second clustering criterion $\rho(\theta_d)$ measures for a given clustering threshold the amount of patients whose follow-up sequences are placed in the same cluster relative to the amount of all patients containing follow-up sequences (427 out of 1157 patients). An optimal clustering threshold should maximize the silhouette score *and* the criterion $\rho(\theta_d)$.

In order to combine both criteria (σ_{min} and ρ), we adjusted ρ (a ratio between 0 and 1) to the same range of values as σ_{min}, which can take values $-1 \leq \sigma_{min} \leq 1$. The overall score to be maximized in order to optimize the clustering threshold can be derived as the arithmetic mean of both values:

$$(\theta_d^*) = \underset{\theta_d}{\text{argmax}} \left(\frac{\rho(\theta_d) \cdot 2 - 1 + \sigma_{min}(\theta_d)}{2} \right).$$

We evaluated the results for the range of divergence thresholds $0.5 \leq \theta_d \leq 7$. The global maximum of the combined clustering score was reached at a mean pairwise divergence of 1.3% (Fig. 2, main manuscript). For comparison, throughout this work we used several locally optimal threshold values indicated in Fig. 2B

4. Statistical analysis of spatial transmission dynamics.

In order to assess the relation between regional proximity of patient's residences and viral transmission, the geographical affiliation of individuals (in terms of the federal states in Germany) within putative transmission clusters was analyzed. Individuals from 15 (out of 16) German federal states are represented in our dataset; for 5 patients the residence was not available (see Table 1, main manuscript).

Each putative transmission cluster may contain patients from different federal states A_i, e.g. a particular cluster may only contain patients that live in 'Berlin' and patients that live in 'Bavaria'. In the following we will refer to this property as the 'geographical cluster composition'. The number of clusters with an identical geographical composition $A_1 \wedge \ldots \wedge A_n$ (e.g. 'Berlin', 'Hamburg', 'Bavaria') is counted and normalized by the number of clusters containing patients from at least one of the respective federal states, i.e. by all clusters containing patients from 'Berlin', 'Hamburg' or 'Bavaria':

$$r_{obs}(A_1 \wedge \ldots \wedge A_n) = \frac{\#clusters\ with\ compostion\ A_1 \wedge \ldots \wedge A_n}{\#clusters\ containing\ A_1 \vee \ldots \vee A_n}.$$

where '#' stands for 'number of'. The amount of geographical compositions is not normalized by the number of *all* clusters, since this would lead to a bias due to the overrepresentation of patients from particular federal

states in our dataset, i.e. an overrepresentation of 'Berlin' in relation to other federal states (shown in Table 1).

We evaluated whether particular cluster compositions, e. g. 'Berlin' & 'Bavaria', are statistically over- or underrepresented. To this end we performed a bootstrap resampling: In order to construct a null model of cluster compositions, we randomly drew from the population of clustered patients with replacement and assigned the drawn patients to clusters, taking the size distribution of the putative transmission clusters into account. These randomly composed clusters were evaluated to obtain $r_{rand}(A_1 \wedge ... \wedge A_n)$ denoting the frequency of geographical cluster compositions after *random* assignment. We repeated the procedure $n = 10{,}000$ times, arriving at 10,000 estimates for r_{rand}.

We then tested whether the *observed* cluster composition ratio r_{obs} was significantly higher, or –lower than the one derived from *random* assignment of patients to putative transmission clusters. As an example, in the case of the Berlin/Bavaria cluster composition we assessed whether patients residing in federal states A_1 (e.g. 'Berlin') and A_2 (e.g. 'Bavaria') were more frequently clustered together in comparison to the random background:

$$H_0: r_{obs}(A_1 \wedge A_2) \leq r_{rand}(A_1 \wedge A_2)$$

$$H_1: r_{obs}(A_1 \wedge A_2) > r_{rand}(A_1 \wedge A_2).$$

The corresponding *P*-value (prob. of false positive prediction) is given by

$$P^{\uparrow} = \frac{\#\left(r_{rand}(A_1 \wedge A_2) \geq r_{obs}(A_1 \wedge A_2)\right)}{N},$$

H_0 can be rejected with significance level α if $P^{\uparrow} < \alpha$.

In order to evaluate whether patients residing in federal states A_1 and A_2 were less frequently clustered together than expected by chance, the test was devised analogously:

$$H_0: r_{obs}(A_1 \wedge A_2) \geq r_{rand}(A_1 \wedge A_2)$$

$$H_1 : r_{obs}(A_1 \wedge A_2) < r_{rand}(A_1 \wedge A_2)$$

with the *P*-value computed according to

$$P^{\downarrow} = \frac{\#\left(r_{rand}(A_1 \wedge A_2) \leq r_{obs}(A_1 \wedge A_2)\right)}{N}.$$

In order to compare the significance levels of the cluster composition frequencies, the test is devised with three different values α: 0.05 and 0.01. The results are shown in Fig. 3A-C (main manuscript).

5. Temporal properties of transmission dynamics

Fig. 4A in the main manuscript was generated in Matlab using the function *distributionPlot* and *P*-values were computed using a Wilcoxon Ranksum test (statistics toolbox). Fig. 4B shows a survival plot that was computed using the function *ecdf* in Matlab (statistics toolbox). *P*-values were computed using a log rank test. Hazard rates were estimated by fitting an exponential to the survival plots shown in Fig 4B with weighted least squares using the *lsqcurvefit* function in Matlab (optimization toolbox).

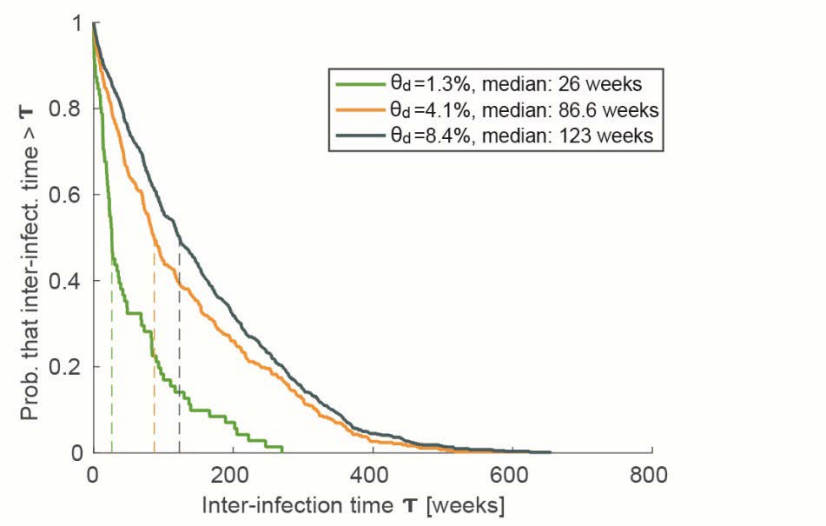

Figure S3: Probability that the inter-infection time is greater than the time indicated on the x-axis. The dashed lines represent the empirical medians of the inter-infection time distribution (corresponding to the values in the figure legend). The results are shown for three different

clustering thresholds θ_d which correspond to global or local minima of the combined score in Fig. 2 B (main manuscript).

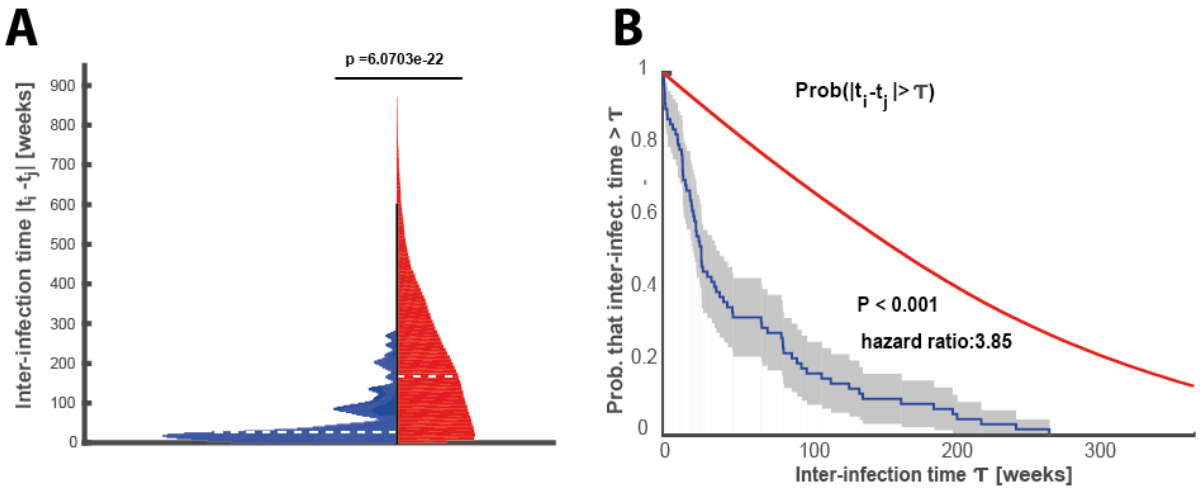

Figure S4: Temporal Transmission Dynamics when the background dynamics are defined in terms of all participants (clustered *and* unclustered). A: Probability distribution of all inter-infection times between patients belonging to the same transmission cluster (blue) and between all patients (red). The dashed white horizontal lines indicate the respective medians. **B:** Probability that the inter-infection time is greater than the time indicated on the x-axis (survival plot). Blue line: patients belonging to the same transmission cluster. Red: all patients. Grey areas represent the respective confidence areas which were computed using Greenwood's formula (vanishingly small for the red line).

6. Correspondence of inter-infection times in clusters with the time to onwards transmission.

In the following, we will outline several scenarios where inter-infection times in clusters and the time to onwards transmission either differ or coincide.

- If a cluster consists of two individuals, then their inter-infection time
    a. will coincide with the time to onwards transmission, if the two individuals are connected by a direct transmission event.

b. In the case where the two individuals are not connected by a direct transmission event, the time to onwards transmission statistics may either be over- or underestimated by the inter-infection times: E.g. in the case of a transmission chain, the time to onwards transmission will always be overestimated by the inter-infection times (e.g. assume individual #2 is not contained in the dataset in Figure S5C). It may be underestimated if individuals share a common source of infection, e.g. in Fig. S5A-B assume individual #1 is not contained in the dataset. Note that if the node-degree distribution in the transmission network is power-law distributed, then the latter scenario (common source of infection) is less likely.

The same considerations hold true, if more than two individuals were clustered together, in the case of missing transmission links or when all individuals are sampled. Hence, our statement "onwards transmission occurs shortly after infection", based on our analysis of inter-infection times, is reasonable.

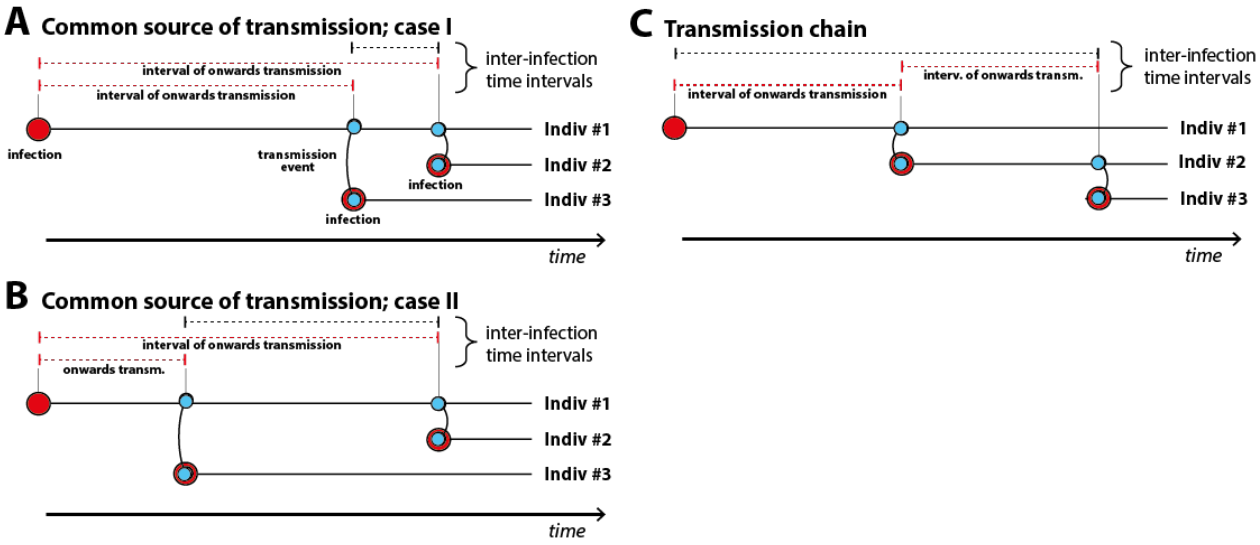

Figure S5: Correspondence of inter-infection times in clusters and time to onwards transmission. A & B: In the case of a common source of infection. **B:** In the case of a transmission chain. Blue dots connected by vertical bended lines indicate transmission events. Dashed red lines mark the actual time interval until onwards transmission. Dashed red and dashed black lines mark the estimated inter-infection times. Consequently, the dashed black lines mark inter-infection times with no correspondence to onwards transmission.

References

1. Clauset A, Shalizi CR, Newman MEJ. Power-Law Distributions in Empirical Data. *Siam Review.* 2009;51(4):661-703.
2. Vuong Q. Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses. *Econometrica.* 1989;57(2):307-333.
3. Lemey P, Kosakovsky Pond SL, Drummond AJ, et al. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput Biol.* 2007;3(2):e29.
4. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. *AIDS Rev.* 2006;8(3):125-140.
5. Shankarappa R, Margolick JB, Gange SJ, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol.* 1999;73(12):10489-10502.
6. Lorenzo-Redondo R, Fryer HR, Bedford T, et al. Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature.* 2016;530(7588):51-56.
7. Zanini F, Brodin J, Thebo L, et al. Population genomics of intrapatient HIV-1 evolution. *Elife.* 2015;4.
8. Poon AF, Joy JB, Woods CK, et al. The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada. *J Infect Dis.* 2015;211(6):926-935.