Male-male and female-female pairs
in HIV phylogenetic source attribution analysis

Lucie Abeler-Dörner and Matthew Hall

9 June 2023

Introduction

Phylogenetic source attribution analysis aims to inform HIV prevention strategies by studying the characteristics of the source population, to learn more about the groups of people that contribute most to onward transmission. These analyses are conducted at the population level on large datasets and never aim to focus on single transmission pairs. While methods for source attribution are getting more precise, there is always some uncertainty and thus they are unsuitable for providing conclusive evidence that any one person infected any other. At the population level, however, individual uncertainty is less important and a strong correlation between being identified as a source by the method and being one in reality is sufficient to obtain fairly precise estimates for the extent that larger groups in a population, for example men and women of certain ages, are involved in transmission.

Here, we will explore why there are almost always male-male and female-female phylogenetic pairs in a genetic analysis and when they are real and when a technical by-product of a given analysis.

Phylogenetic distance thresholds of likely direct HIV transmission pairs

Phylogenetic techniques analyse the similarity between a set of genetic sequences and, using models on how mutations occur, generate phylogenetic trees based on how closely the sequenced organisms are likely to be related to each other. In the context of pathogens, a close relationship also indicates proximity of hosts in the chain of transmission. The distance in the phylogenetic tree between sequences from pathogens infecting two hosts is smallest for direct transmissions and increases for sequences that have one or more intermediates in the chain (Figure 1).

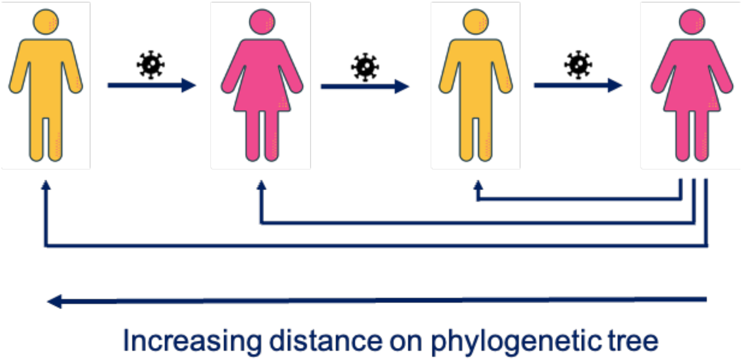

Figure 1: Phylogenetic relationships of viruses, and thus their hosts, in a transmission chain. The more transmissions have occurred, the larger the phylogenetic distance between them. Icons in this and all other figures from freepick.com and flaticon.com

The HPTN 052 study looked at transmission in cohabitating heterosexual couples. In this group, it is very easy to determine pairs that were phylogenetically linked by transmission and pairs that were not based on phylogenetic distance.

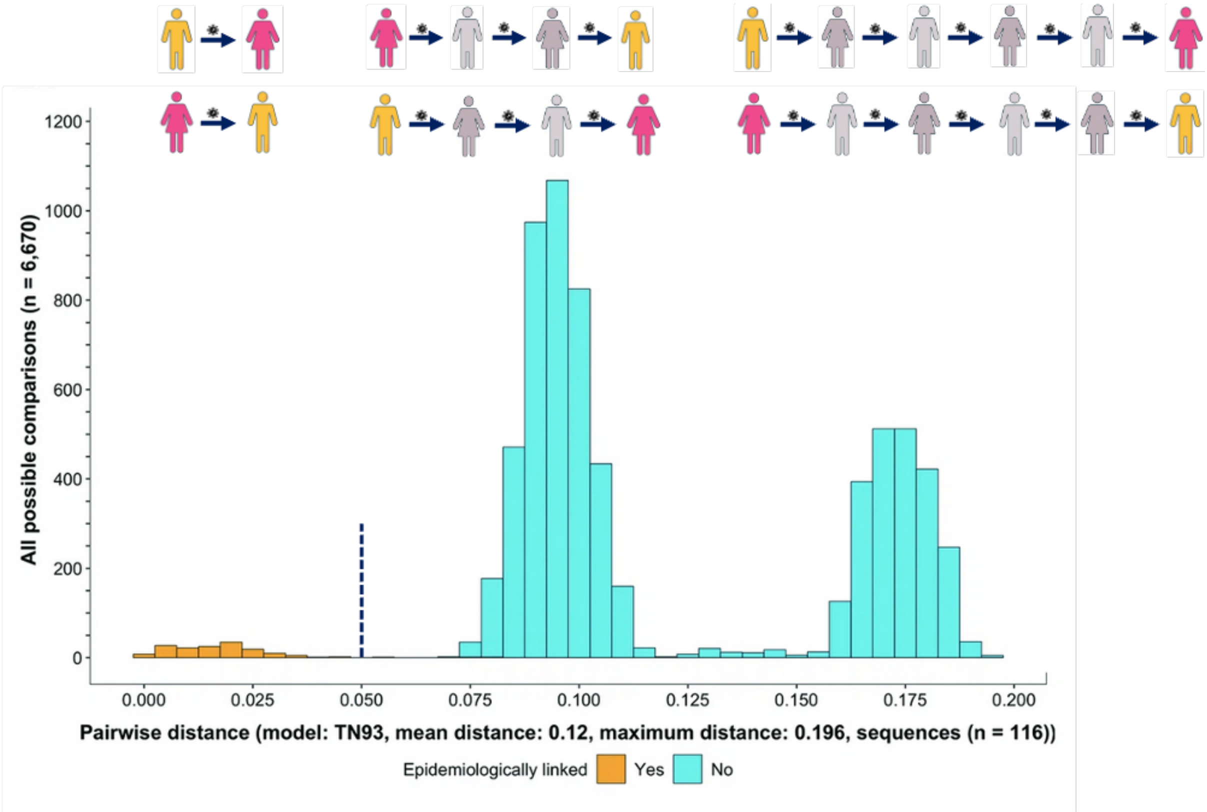

Figure 2: Distribution of genetic distances between viral whole-genome consensus sequences couples in the HPTN 052 study, graph modified from Magosi et al 2022, *eLife* **11**:e72657. Sampled individuals are depicted in colour.

In this case, the yellow peak of individuals involved in direct transmission is clearly distinguishable from the turquoise peaks of those for indirect transmissions and it is easy to pick a distance threshold (the dark blue line) below which the transmission is considered direct.

In most other studies, however, which often recruit a representative, random or opportunistically sampled fraction of the population, a graph like this would be less clear cut. In this case, choosing the best threshold depends on the data, but also on additional information we have on the population. In particular, they depend on

- whether we expect to find a large fraction of non-heterosexual transmission,
- whether we expect to find a large fraction of transmission through injected drug use,
- and whether we expect most transmissions to be recent or not.

Same-sex transmission pairs in datasets with predominantly heterosexual HIV transmission

Let's take the easiest case, case A: a dataset from an epidemic in rural or peri-urban sub-Saharan Africa in which we expect the vast majority of transmissions to be heterosexual, with few MSM who might choose not to be part of the study, and no or a negligible amount of injected drug use.

A plot of pairwise distances might look like Figure 3.

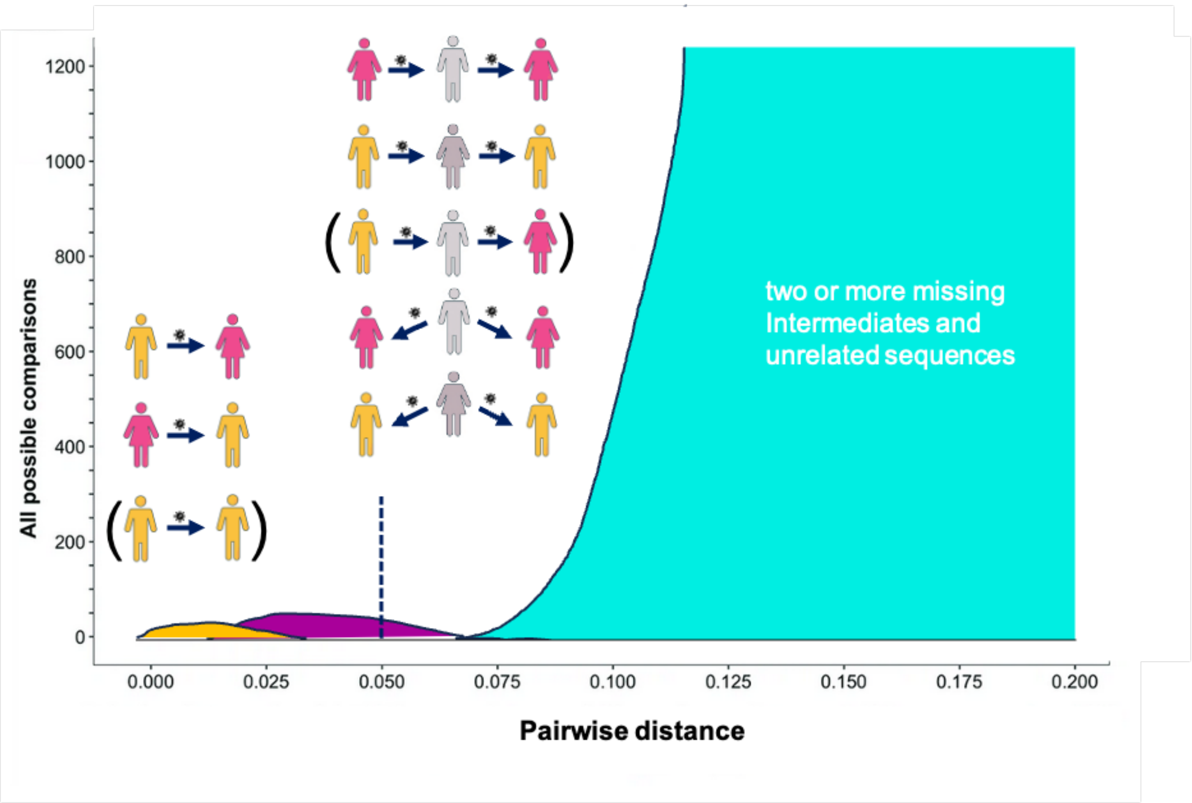

Figure 3: Schematic depiction of a pairwise distance distribution found in a dataset characterised by overwhelmingly heterosexual HIV transmission. Sampled individuals are depicted in colour. Unlikely scenarios under the assumptions are depicted in brackets.

The true transmission pairs (yellow) are still separated from chains with two or more missing intermediates and unrelated sequences (turquoise), but the picture is less clear because of chains with one intermediate or cases where one person infected two people who were then sampled. The threshold from the HPTN 052 study is still the best threshold to separate the yellow and the turquoise peak, but keeping the threshold means that a large fraction of the purple peak ends up below our threshold for direct transmission. However, if only one missing intermediary is possible and all transmission is heterosexual, then it is easy to identify direct transmissions because the people involved are of opposite sexes. The individuals on either side of an unsampled individual are of the same sex as each other. These are the male-male and female-female pairs which are reported in many phylogenetic studies. We would still like to only keep working with the pairs in the yellow peak. Under the assumption that the dataset contains overwhelmingly heterosexual transmissions, the best possible separation we can get between the yellow peak and the peak is to keep all opposite-sex pairs and exclude all same-sex pairs.

This does not mean that there are no same-sex relationships in the dataset or that there is no MSM transmission. It just means that we focus on the heterosexual pairs to reduce the overall error. Depending on the dataset and the research question, it might be possible to estimate the contribution to transmission of true homosexual pairs in the dataset by estimating the expected number of male-male pairs based on the sample size, the number of observed opposite-sex pairs and the number of female-female pairs which we know are not cases of biological transmission. If the number of observed male-male pairs is much higher than the expected number, the assumptions and analysis choices might have to be revisited. Whether male-male pairs should be included in the analysis plan also depends on the legal situation and social and cultural norms of the study country.

Another special case are mother-to-child transmissions. Fortunately, these are by now rare almost everywhere in the world and can usually be identified by the age-gap in the pair which is much larger than usual age gaps in transmission pairs with a female source.

In summary, in data sets from largely heterosexual epidemics, same-sex pairs occur as a result of to how the phylogenetic distance threshold is chose which is used to differentiate direct transmissions and individuals separated by two missing intermediates. The vast majority of these pairs will have one missing intermediary and they can be excluded from further analysis under the assumptions. If desired, mathematical models can be to test the assumption of a largely heterosexual epidemic by estimating the expected number of male-male pairs, depending on the dataset and the research question.

Same-sex transmission pairs in datasets of homosexual HIV transmission

Let us now consider case B, a dataset of self-identified homosexual male participants, for example in a big city in Europe. All participants in the study are male, so all pairs are male-male pairs. A plot of pairwise distances might look like Figure 4.

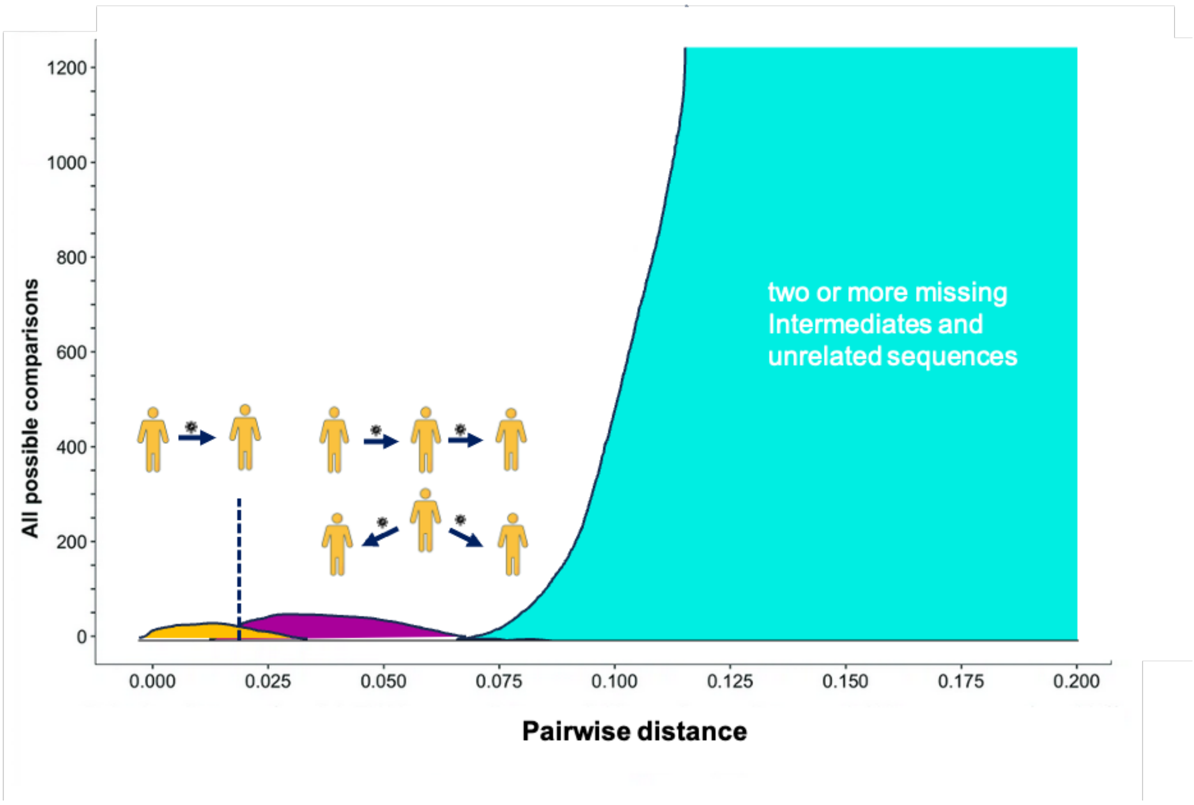

Figure 4: Schematic depiction of a pairwise distance distribution found in a dataset characterised by MSM transmission.

The best distance threshold for direct transmission pairs has now changed (assuming similar preferences of sensitivity versus specificity, as sex can no longer be used as a criterion to eliminate indirect transmission pairs. The group of predicted direct transmissions has now more false positives and more false negative pairs are excluded from analysis.

Same-sex transmission pairs in datasets of HIV transmissions in groups with injected drug use

HIV transmission in groups with injected drug use can occur through needle sharing but also through sexual transmission. Assuming that we look at a dataset in which transmission has predominantly occurred through use of shared needles, the same threshold would be used than for a dataset with homosexual transmission (Figure 5). The difference between the two is that in this particular case, female-female pairs can also be genuine transmission pairs, although they have been reported to be rate, and many more combinations are possible for the scenario of one missing intermediate.

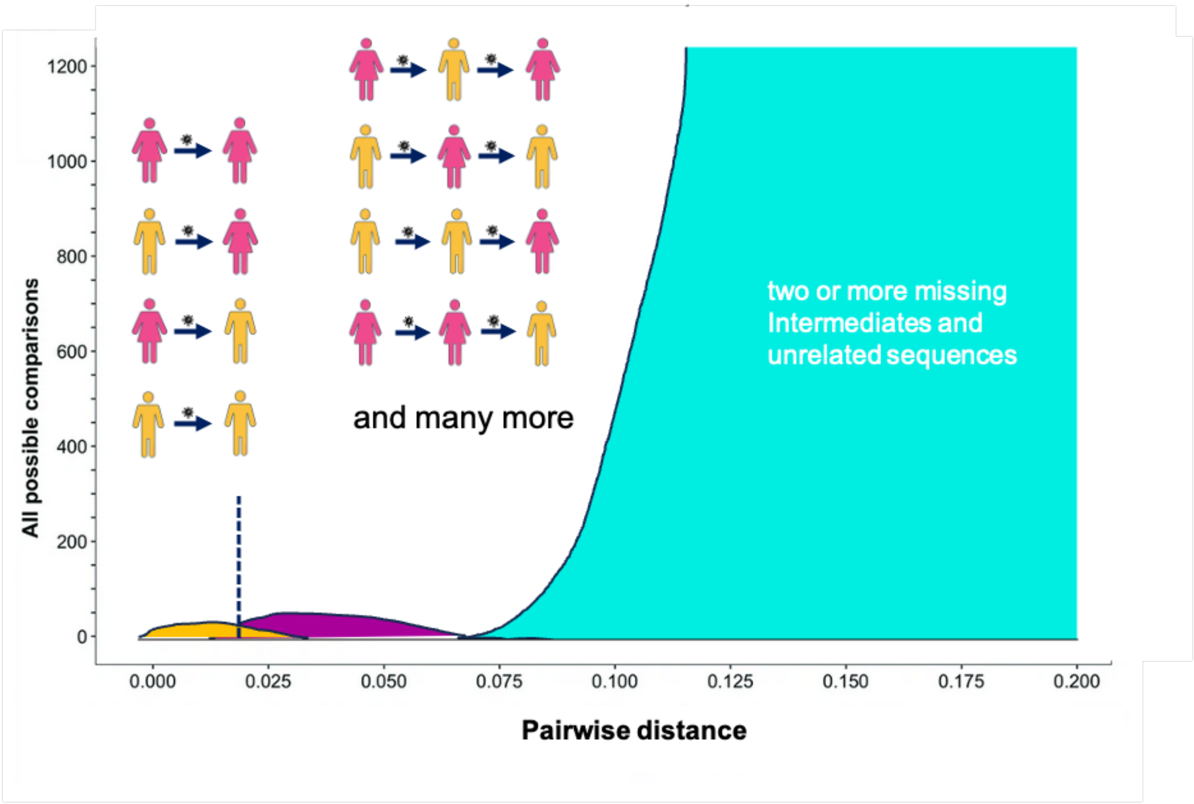

Figure 5: Schematic depiction of a pairwise distance distribution found in a dataset characterised by HIV transmission through needle sharing in communities who inject drugs

Besides the occurrence of plausible female-female transmission pairs, datasets with transmission originating from injected drug use can be (but do not have to be) characterised by the appearance of several very similar sequences, a higher frequency of dual transmission and a higher frequency of viruses with CXCR4 receptor usage.

Further comments

Datasets with mixed heterosexual and homosexual transmission and possibly even transmission through shared needles are difficult to analyse. In the absence of any additional information, the best threshold would be somewhere between the one used in Figures 2/3 and 4/5. It would however be preferable to obtain additional metadata to inform analysis choices.

Other additional information can also be helpful to inform analysis choices. The phylogenetic distance between sequences from a transmission pair increases with time. Knowledge about whether most inferred transmissions are likely to have happened recently or some time ago can therefore also help to determine the best threshold.

Of course, the best threshold also depends on the research question and criteria for sensitivity and specificity. For example, if the focus is not on direct transmission pairs but on clusters, smaller networks and sub-epidemics, a larger threshold should be chosen.

Phylogenetic transmission analyses create highly sensitive data and should only be used where fully informed consent has been obtained from participants, and in contexts where researchers and/or public health officials can guarantee that participants will not be subject to criminalisation or stigmatisation, or at risk of reducing the interaction with their health provider because they fear criminalisation or stigmatisation. For further information on the ethical implications of phylogenetic transmission analyses, please see Coltard et al, Lancet HIV. 2018, 5(11):e656-e666 and Jamrozik et al 2023 (forthcoming in BMJ Global Health).

Conclusion

The significance of same-sex pairs observed in a phylogenetic analyses of HIV transmission is highly dependent on the context of the analyses. They may represent genuine transmission in datasets of HIV transmission between MSM and can represent genuine transmission pairs in analyses of outbreaks caused by use of shared needles in groups who inject drugs. In both cases it is hard to separate direct transmissions from cases of one missing intermediary. In analyses of predominantly heterosexual epidemics, they are however most likely to be artefactual and a by-product of the chosen distance threshold between phylogenetic pairs. Their existence or even a relatively high percentage of same-sex pairs do not represent a flaw in the analysis. Depending on the analysis and the research question, the fraction of male-male pairs in comparison to the fraction of female-female pairs and opposite-sex pairs can be used, in combination with the fraction of male and female participants in the study, to calculate the likelihood that male-male pairs represent genuine MSM transmissions.