# USING PHYLOGENETICS TO ACCURATELY INFER HIV-1 TRANSMISSION DIRECTION

*Christian Julian Villabona-Arenas[1,2], Stéphane Hué[1,2], James Baxter[3], Matthew Hall[4], Katrina A. Lythgoe[4], John Bradley[1], Katherine E. Atkins[1,2,3] **

[1]Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

[2]Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, London, UK

[3]Centre for Global Health, Usher Institute of Population Health Sciences and Informatics, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK

[4]Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK

* Corresponding author: Katherine.Atkins@ed.ac.uk

## Abstract

Inferring the direction of transmission between linked individuals living with HIV provides unparalleled power to understand the epidemiology that determines transmission. State-of-the-art approaches to infer directionality use phylogenetic ancestral state reconstruction to identify the individual in whom the most recent common ancestor of the virus populations originated. However, these methods vary in their accuracy when applied to different datasets and it is currently unclear under what circumstances inferring directionality is inaccurate and when bias is more likely. To evaluate the performance of phylogenetic ancestral state reconstruction, we inferred directionality for 112 HIV transmission pairs where the direction of transmission was known, and detailed additional information was available. Next, we fit a statistical model to evaluate the extent to which epidemiological, sampling, genetic and phylogenetic factors influenced the outcome of the inference. Third, we repeated the analysis under real-life conditions when only routinely collected data are available. We found that the inference of directionality depends principally on the topology class and branch length characteristics of the phylogeny. Specifically, directionality is most correctly inferred when the phylogenetic diversity and the minimum root-to-tip distance in the transmitter is greater than that of the recipient partner and when the minimum inter-host patristic distance is large. Similarly, under real-life conditions, the probability of identifying the correct

transmitter increases from 52%—when a monophyletic-monophyletic or paraphyletic-polyphyletic tree topology is observed, when the sample size in both partners is small and when the tip closest to the root does not agree with the state at the root—to 93% when a paraphyletic-monophyletic topology is observed, when the sample size is large and when the tip closest to the root agrees with the state at the root. Our results support two conclusions. First, that discordance between previous studies in inferring transmission direction can be explained by differences in key phylogenetic properties that arise due to different evolutionary, epidemiological and sampling processes; and second that easily calculated metrics from the phylogenetic tree of the transmission pair can be used to evaluate the accuracy of inferring directionality under real-life conditions for use in population-wide studies. However, given that these methods entail considerable uncertainty, we strongly advise against using these methods for individual pair-level analysis.

## Background

Identifying transmission chains via contact tracing is a cornerstone of infectious disease control. It provides an opportunity to test potential cases, treat infections early and break ongoing transmission. Moreover, identifying the direction of transmission (DoT), provides paramount knowledge for understanding risk factors of transmission and susceptibility [1–3], household transmission [4], spread [5–7] and early pathogenesis events [8]. Yet inferring directionality is challenging. In rare instances, the DoT can be inferred from the comparison of symptoms onset time or testing histories of the partners. However, this method is restricted to cases with known contact histories, and for whom other sources of infection can be ruled out, such as for sexually transmitted infections occurring between self-reported sexual partners.

Comparing the ancestral relationship between pathogen genomes sampled from a putative transmission pair has been proposed as a method to identify the DoT [9]. Specifically, current approaches propose that the DoT may be inferred from ancestral state reconstructions along phylogenies of pathogen sequences sampled from a pair of linked individuals using parsimony-based algorithms [9,10]. Under this framework, the transmitting individual corresponds to the state at the root—i.e., individual A or B—after minimizing

the number of state changes along the phylogeny necessary to explain the observed state distribution at the tips. For example, when paraphyly is observed—i.e., when all sequences from one partner form a monophyletic cluster embedded within the pathogen population of the other partner— it would be concluded that the monophyletic clade represents the recipient's viral population [9]. While simulations suggest that using the topology of a phylogeny reconstructed from multiple viral sequences sampled from a known transmission pair can correctly identify the DoT, empirical tests of this hypothesis varied in accuracy. For example, one study reported a correct inference of the direction of HIV transmission in 31/32 transmission pairs, with no direction incorrectly inferred [11] but other studies incorrectly identified the direction of HIV transmission in 4/31 couples [12] or 4/36 couples [13]. Moreover, the relative contribution of different factors to the inference of DoT remains elusive and thus the likely success of identifying transmission directionality is not always obvious. While the scientific, ethical and legal implications of identifying the transmitting partner likely limit analysis to the population level rather than the individual level, these methods require consistent and verifiable accuracy if they are to be widely adopted [14].

In this study, we analyzed HIV transmission pairs—for which both the DoT and detailed epidemiological information are known, and multiple virus sequences are available—to test the predictive value of current phylogenetic approaches of direction inference. Next, we fit a statistical model to evaluate the extent to which epidemiological, sampling, genetic and phylogenetic factors influence the outcome of the inference. Third, we developed a statistical model that predicts the likely success of identifying DoT under real-life conditions when only routinely collected data are available. Using this model, we provide a framework to suggest how the accuracy of determining the DoT can be incorporated in population-level transmission studies.

## Methods

### Ancestral state reconstruction

We used publicly available HIV-1 sequence data from 112 transmission pairs for which the DoT is known from epidemiological records and where at least five sequences were available per individual [8]. For each pair in our base case analysis, we inferred a Maximum Likelihood (ML) phylogenetic tree under a general

time-reversible nucleotide substitution site model with the addition of heterogeneity of substitution rates among sites using a parametric Gamma model (GTR+G) with IQ-Tree [15]. For each ML tree, we calculated the probability, $p_i$, that ancestral state reconstruction correctly identifies the transmitting partner in each pair, *i*, while using the partner's role in the transmission event as states. That is, after labelling the tips of each tree as sampled from either the transmitter or the recipient partner, we estimated the state probabilities at the root that maximized the likelihood of observing the state distribution at the tips using a joint estimation procedure (i.e., calculating the most likely state for each internal node in the tree while integrating over all the possible states along the other nodes in proportion to their probability) and assuming equal rates of transition between the two states. These analyses were conducted with the R package Ape [1617].

We conducted sensitivity analyses to assess the role of phylogenetic reconstruction in assigning the DoT. First we assessed the role of the branch lengths estimated with a parametric Gamma model by calculating $p_i$ from ML trees built under a four category non-parametric rate heterogeneity model (FreeRates) [18,19] using IQ-Tree [15]. We then evaluated whether Bayesian approaches improve the accuracy of the inferences. For this, we used hierarchical Bayesian inference (BI) with MrBayes 3.2.7 [20,21] and simultaneously calculated a distribution of trees and the corresponding ancestral state posterior probabilities at the root, $X_i$; we then defined the probability density $p_i = Pr(X_i > 0.5)$. Finally, we inferred the ancestral states of each ML tree using the most parsimonious reconstruction which, instead of providing state probabilities, selects the state at the root that incurs the smallest number of state changes that are needed to observe the state distribution at the tips. For this, we used the Sankoff algorithm with the R package phangorn 2.5.5 [22–24].

**Phylogenetic inference of DoT**

In our base case analysis, we classified the inferred direction of transmission (I-DoT) as "consistent" with the known transmission direction if $p_i \geq 0.5$, or "inconsistent" otherwise. In a sensitivity analysis, we accounted for a third "equivocal" outcome by classifying the I-DoT for each transmission pair, *i*, as either "consistent" if $p_i \geq t$, "inconsistent" if $p_i \leq 1\text{-}t$, or "equivocal" otherwise. We used both a relaxed threshold of

115 *t*=0.6 and a conservative threshold of *t*=0.95 for this ordinal three-class outcome. For the parsimony-
116 based approach, we classified the I-DoT as either "consistent" if the state at the root was the transmitting
117 partner, "inconsistent" if the state at the root was the recipient partner, or "both" if both partners were
118 equally parsimonious at the root.
119
120 **Explaining the accuracy of phylogenetic inference of transmission direction**
121 We evaluated in what circumstances ancestral state reconstruction succeeds in identifying the
122 transmitting partner. For this, we built a suite of logistic regression models to predict the I-DoT as a
123 function of information available from all transmission pairs. That is, for the base case binary outcome,
124 the probability that the inferred DoT is consistent with the known DoT, while for the three-class outcome,
125 the probability that the inferred DoT is consistent or inconsistent with the known DoT. We used 13
126 covariates organized into four classes (Table 1).
127
128 **Table 1**. Covariates used in the two models

| **Class** | **Covariate** | **Values** (units where applicable) * | |
|---|---|---|---|
| | | **Model with all data** | **Model with routinely available data** |
| *Epidemiological information (E)* | Sexual risk exposure group | Men-who-have-sex-with-men (*MSM*) or Heterosexual (*HET*) | |
| | Recency of the transmitter's infection | *Acute* (transmission occurred up to 90 days after the transmitter's infection), or *Chronic* (otherwise) | Excluded |
| *Sampling information (S)* | Sample size | *Low* (number of sampled haplotypes in at least one the partners is less than 10) or *High* (otherwise) | |
| | Sample size difference | Difference* in the number of sampled haplotypes between the partners | Absolute difference in the number of sampled haplotypes between the partners |

| | | | |
|---|---|---|---|
| | Time from transmission | Sum of the absolute time-to-sampling of both partners relative to transmission time (days) | Excluded |
| *Genetic information (G)* | Sequence alignment length | Number of base pairs | |
| | Intra-host nucleotide diversity difference | Difference* between the within-partner mean pairwise sequence diversity (substitutions per site) | Absolute difference between the within-partner mean pairwise sequence diversity (substitutions per site) |
| | Multiplicity of infection | *Single* (probability of one founder haplotype in the recipient is greater than or equal to 0.75) or *Multiple* (otherwise)† | Excluded |
| *Phylogenetic information (P)*** | Topology class | Paraphyletic-polyphyletic (PP), Paraphyletic-monophyletic (PM) or Monophyletic-monophyletic (MM)‡ | |
| | Phylogenetic diversity difference | Difference* between the sum of the branch lengths of each partner subtree§ (substitutions per site) | Absolute difference between the sum of the branch lengths of each partner subtree§ (substitutions per site) |
| | Root-to-tip difference | Difference* between the minimum root-to-tip distances of the partners' sequences (substitutions per site) | Absolute difference between the minimum root-to-tip distances of the partners' sequences (substitutions per site) |
| | Most basal tip identity | *Transmitter*, *Recipient* or *Both*. The identity of the tip(s) that minimizes the number of internal nodes along the paths between itself and the root | *Agree*, *Disagree* or *Ambiguous*. Whether the identity of the tip(s) that minimizes the number of internal nodes along the paths between itself and the root matches the identity of the individual with the higher ancestral state probability at the root |

| | | | |
|---|---|---|---|
| | Inter-host patristic distance | The shortest patristic distance between tips from the partners (substitutions per site) | The shortest patristic distance between tips from the partners (substitutions per site) |

† As in [8]
‡ As in Romero-Severson et al, 2016
§ We used the sum of the edge lengths that give rise to only one tip in the subtree as in [25]
* Subtraction of the recipient's value from the transmitter's value.
** Illustrated in **Figure S1**; to build these covariates when using a posterior distribution of trees, we selected either the most frequent observation (in the case of qualitative covariates) or the mean shift mode (in the case of the quantitative covariates).

We fitted 16 separate models built from all possible combinations of these four classes to identify the best set of I-DoT predictors. That is, one model with all four classes of predictors (ESPG), four models with three classes (ESG, ESP, SGP and EGP), six models with two classes (EG, ES, SG, SP, GP and EP) and four single-class models (E, S, G and P).

**Increasing the accuracy of transmission direction inference**

The previous suite of statistical models assumes knowledge of the transmitter and recipient's identity in addition to epidemiological information not typically known. To evaluate how to interpret the I-DoT under 'real-life' conditions when this information is unknown, we developed a second suite of models with a reformulated set of eight covariates which are described in Table 1.

To evaluate how transmission pair characteristics influence the accuracy of inferring the DoT, we created simulated datasets that represent all possible combinations of the eight covariates. We used the respective categories of the discrete covariates, while for the continuous covariates we used a range of values evenly distributed between the minimum and the maximum of the original data. Using the best model from the binary suite of models and the best model from each of the ordinal suites of models ($t$=0.6 or $t$=0.95), we calculated the probability of I-DoT across the range of transmission pair characteristics.

**Model fitting, comparison and selection**

We fitted all statistical models using least absolute shrinkage and selection operator (Lasso) regression with the R package glmnet [26] for the binary models and with the R package ordinalNet [27] for the ordinal models. Using this approach, the resulting coefficients can be interpreted as evidence against the

inclusion of a covariate if the coefficient shrinks to zero [28]. The shrinkage coefficient was estimated using leave-one-out cross validation. To compare the binary models, we calculated the area under the curve (AUC) statistic with the R package pROC [29]. For ordinal models, we calculated a macro-AUC by averaging all results (one versus the rest) with linear interpolation between points using the R package multiROC [30]. We considered models with AUC > 0.9 to have high discriminatory power and selected the best ranking models as those with the highest AUC within three decimal places.

## Results

### Data

The 112 transmission pairs exhibited wide variation across all the epidemiologic, sampling, genetic, and phylogenetic characteristics evaluated (**Figure S2**). Specifically, the transmitter was in the acute stage at the time of transmission in 11/112 pairs, while 36/112 pairs were reported as MSM (as previously described, [8]). The sample size was low (i.e., fewer than 10 sequences in the least sampled individual) in 60/112 pairs, while the median sample size difference was 5.5 haplotypes (interquartile range—IQR—1.00-13.25), and the median sampling time of both partners relative to recipient infection was 173 days (IQR 84-410 days).

The median sequence alignment length was 1,534 base pairs (IQR 747-2,591); a total of 103/112 of the pairs had sequences that spanned the *env* region while 9/112 spanned the *gag* region; the median difference of intra-host nucleotide diversity was 0.013 substitutions per site (IQR 0.005-0.030%), while 84/112 recipient's infections were more probably seeded by a single variant.

The most frequent topology class was PM (62/112), followed by PP (30/112) and MM (20/112), while the median difference in phylogenetic diversity was 0.051 substitutions per site (IQR 0.010-0.122), the median difference in minimum root-to-tip distances was 0.007 substitutions per site (IQR 0.002-0.020) and the median of the minimum inter-host patristic distance was 0.009 substitutions per site (IQR 0.003-0.020). In terms of the most basal tip identity, the tip closest to the root (i.e., the one separated by the

least number of internal nodes) belonged to the transmitter partner in 86/112 pairs, to the recipient partner in 12/112 pairs, and tips from both partners were equally closer to the root in 14/112 pairs.

**Phylogenetically inferred DoT (l-DoT)**

We found that probabilistic ancestral state reconstruction tends to correctly infer the DoT, with 83.9% (94/112) of the pairs being consistent and 16.1% (18/112) of the pairs inconsistent with the known DoT (**Figure 1, Table S1**). There were significant differences in the topology class by outcome (Pearson's Chi-squared $P < 0.001$) with a PM topology class being more frequently observed (59/94) when the DoT was correctly inferred, while MM (8/18) and PP (7/18) were more frequently observed when the DoT was incorrectly inferred.

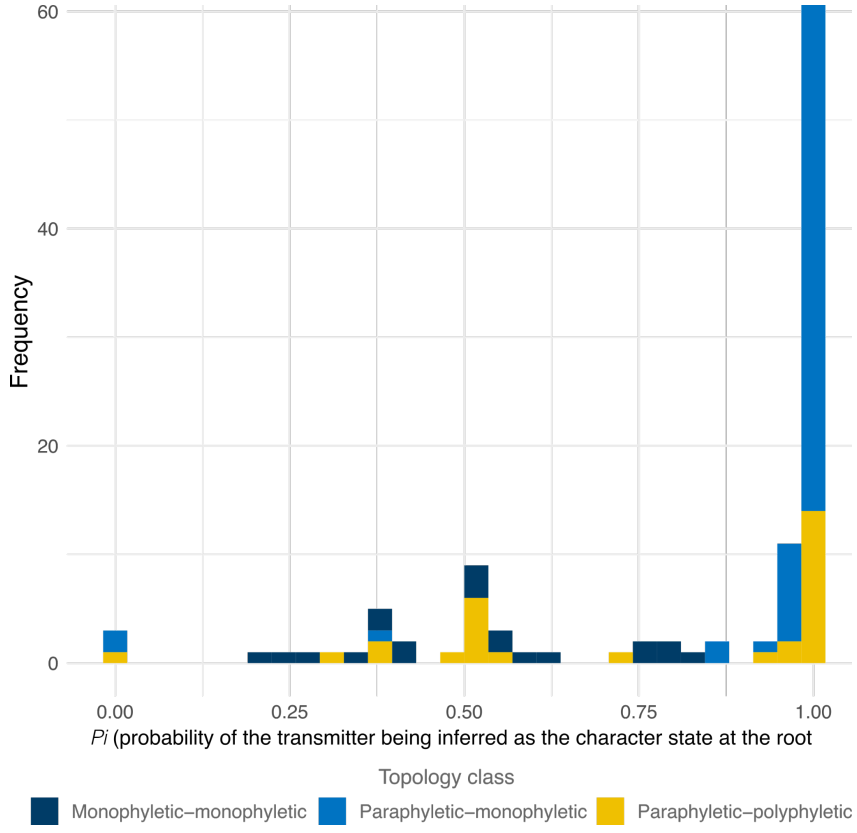


**Figure 1.** The probability for each transmission pair, *i*, that the transmitting partner is correctly identified using ML ancestral state reconstruction. Observations are colored by the observed topology class.

**Explaining the accuracy of phylogenetic inference of transmission direction**

We found that the 16 logistic models varied greatly in their discriminatory power to detect when the phylogenetically inferred DoT was correct, with the AUC values ranging between 0.723 and 0.976 (**Figure 2A**). There were seven models with an AUC greater than 0.9, with a median AUC of 0.974 and with little separating their discriminatory power (the maximum ΔAUC was 0.006); these seven models all included at least four covariates from the phylogenetic class (P) after variable selection and regularization (**Figure 2B**).

The model P had the highest discriminatory power (0.976, **Table S2**). In this model, the probability of correctly inferring the DoT increases, (i) when we observe a PM or a PP topology (compared to MM), (ii) when the most basal tip in the tree corresponds to a sample from the transmitter (compared to both partners being equally basal), (iii) when the difference in the phylogenetic diversity and (iv) the minimum inter-host patristic distance get larger, and (v) when the difference between the minimum root-to-tip distances of the pair's sequences gets smaller. In contrast, the probability of correctly inferring the DoT decreases (i) when the most basal tip corresponds to a sample from the recipient (compared to both partners being equally basal).

## Increasing the accuracy of transmission direction inference

### *Base case analysis*

When we re-analyzed the data after masking the identity of each partner and using only routinely available information, in our base case analysis, the fitted models were either single-class models (S, G, P) or the dual-class models SG and SP (**Figure 2C**). Model SP was the best-fitting model (AUC = 0.827), with sample size, sample size difference, topology class, and the identity of the most basal tip, being the predictor covariates (**Figure 2D**). Specifically, model SP suggests that the probability that the transmitting partner is correctly identified is higher (i) when the sample size is large (compared to small), (ii) when the difference in sample size gets larger, (iii) when we observe a PM topology (compared to either MM or PP), (iv) when the identity of the most basal tip agrees (compared to either disagreeing or being ambiguous) with the identity of the individual with the highest probability at the root.

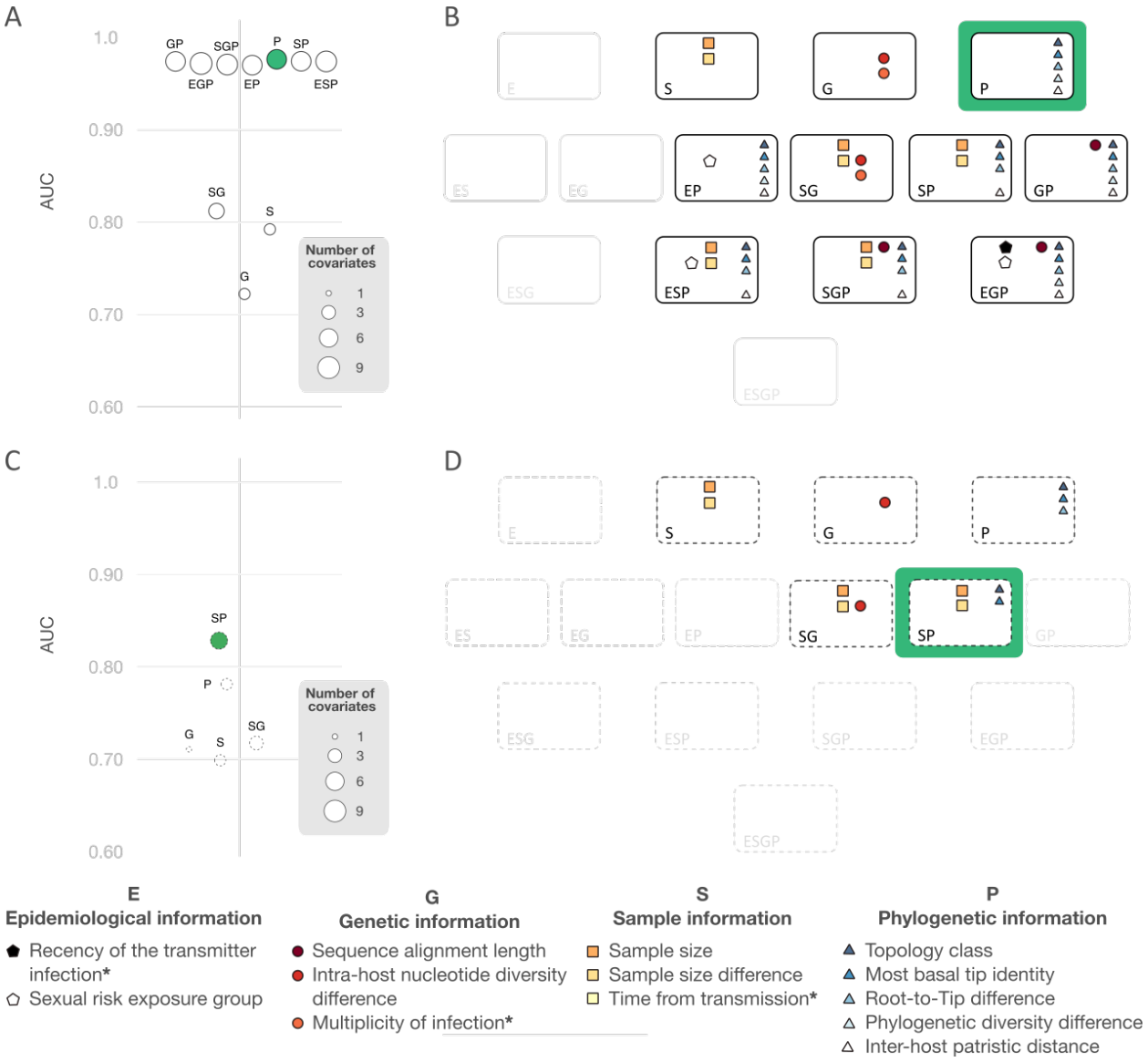


**Figure 2. Model results.** (A) AUC of the masked models, with green fill for models with the highest AUC. (B) The final subset of predictor covariates after Lasso regression fits. (C) as in (B) but using only predictor covariates that are routinely available and where the direction of transmission is unknown are used (i.e., excluding predictors marked with '*') models.

***Sensitivity analyses***

*Equivocal outcomes*: When we classify the inferred DoT to be either consistent, inconsistent or equivocal with a relaxed probability threshold, the I-DoT is consistent with the known transmission direction in 74.1% (n=83) of the pairs, equivocal in 13.4% (n=15) and inconsistent in 12.5% (n=14) (**Table S1**). Similarly, using a conservative threshold increased the proportion of pairs that are classified as equivocal to 33.0% (n=37) and reduced the proportions of consistent and inconsistent pairs to 64.3% (n=72) and

2.7% (n=3), respectively. Regardless of the threshold, the ordinal model with the highest macro-AUC was model P (the phylogenetic class only model), but the discriminatory power of this model was lower for the conservative threshold than that of the relaxed one (AUC = 0.765 vs. 0.843 respectively, **Figure S3**, **Table S3**). Similar to the binary model, the ordinal models show that the probability of the I-DoT being correct is higher when we observe a PM topology (compared to either MM or PP), and when the identity of the basal tip agrees (compared to either disagree or ambiguous) with the identity of the individual with the highest probability at the root. In addition, this probability increases as the difference in phylogenetic diversity between the individuals gets larger and, in the case of the conservative threshold, also increases as the difference between the minimum root-to-tip distances gets larger and the minimum inter-host patristic distance gets smaller.

*Most parsimonious ancestral state reconstruction*: When we used the most parsimonious reconstruction to calculate the inferred DoT from the ML trees, the model SP had the highest macro-AUC and an equivalent discriminatory power than the best-case scenario of the probabilistic ordinal approach (AUC = 0.844; ΔAUC of 0.004) (**Figure S3**, **Table S3)**. This model suggests that the probability of the I-DoT being correct is higher when the observed difference in sample size gets larger, when we observe a PM topology (compared to either MM or PP), when the identity of the most basal tip agrees (compared to either disagreeing or being ambiguous) with the most parsimonious state at the root.

*Tree reconstruction methods*: When we used either ML tree reconstruction with a non-parametric rate heterogeneity model (R4) or BI under a GTR+G4 model, we found that the binary model with the highest macro-AUC was the model GP (AUC of 0.853 and 0.867 for ML+R4 and BI, respectively). On the other hand, with ML under GTR+R4 model the best ranking ordinal model was model P regardless of the threshold (an AUC of 0.835 and 0.821 for $t$=0.6 and $t$=0.95, respectively). for BI, the top ranked models were model P with t=0.6 and model SP with t=0.95 with an AUC of 0.837 and 0.747, respectively (**Figure S4**, **Table S3**). The GP and SP models included the covariate difference in intra-host nucleotide diversity and difference in the number of sampled haplotypes, respectively. All GP and P models included the two topological covariates, that's it the topology class and whether the identity of the most basal tip agrees or

disagrees with the identity of the individual with the higher probability at the root. When the equivocal outcome was considered, the covariates that rely on branch lengths (i.e., phylogenetic diversity, root-to-tip and patristic distances) became additional predictors for the correct identification of the DoT with the same effects as for the base case ordinal models, with the exception of the model built under ML with GTR+R4 and t=0.6 in which only the two topological covariates remain as important.

**Implications for bias within population studies**

We next evaluated whether routinely undisclosed epidemiological characteristics are associated with the probability of correctly identifying the direction of transmission. Specifically, we found that the stage of the transmitter's infection at the time of transmission is associated with the topology class of the phylogenetic tree. That is, PP topologies—that are associated with less chance of accurately predicting the transmitting partner—are more frequently observed (72.7%) when transmission occurred during the transmitter's acute stage and PM topologies—that are associated with more chance of accurately predicting the transmitting partner—are more frequently observed (58.4%) when transmission occurred during the transmitter's chronic stage (Pearson's chi-squared test $P < 0.001$). Therefore, because the stage of the transmitting partner's infection is likely to influence the topology class of the phylogenetic tree, which, in turn influences the probability of correctly identifying the transmitting partner, there is a risk of overrepresentation chronic stage infections in the set of correctly identified transmission pairs.

**Implications for inference of transmission direction**

Our analysis suggests that transmission pair characteristics influence the likelihood of correctly identifying the DoT using ancestral state reconstruction. To estimate the practical importance of this result, we used the best binary and ordinal models (SP and P) to predict the chance of inferring the correct DoT using a simulated dataset.

Our results suggest that, in our binary and ordinal model with a equivocal threshold of 0.6, observing a PM topology is sufficient to provide at least a 75% chance of correctly identifying the transmitting partner (**Figure 3**). If the identity of the most basal tip agrees with the identity of the individual with the higher

state probability at the root, this chance increases to at least 90%. If the classification is between consistent or inconsistent, this probability further increases by observing a minimum difference in the number of haplotypes in the samples. (**Figure 3A**). Conversely when the classification is between consistent, inconsistent and equivocal, this probability further increases by observing a minimum difference in phylogenetic diversity.

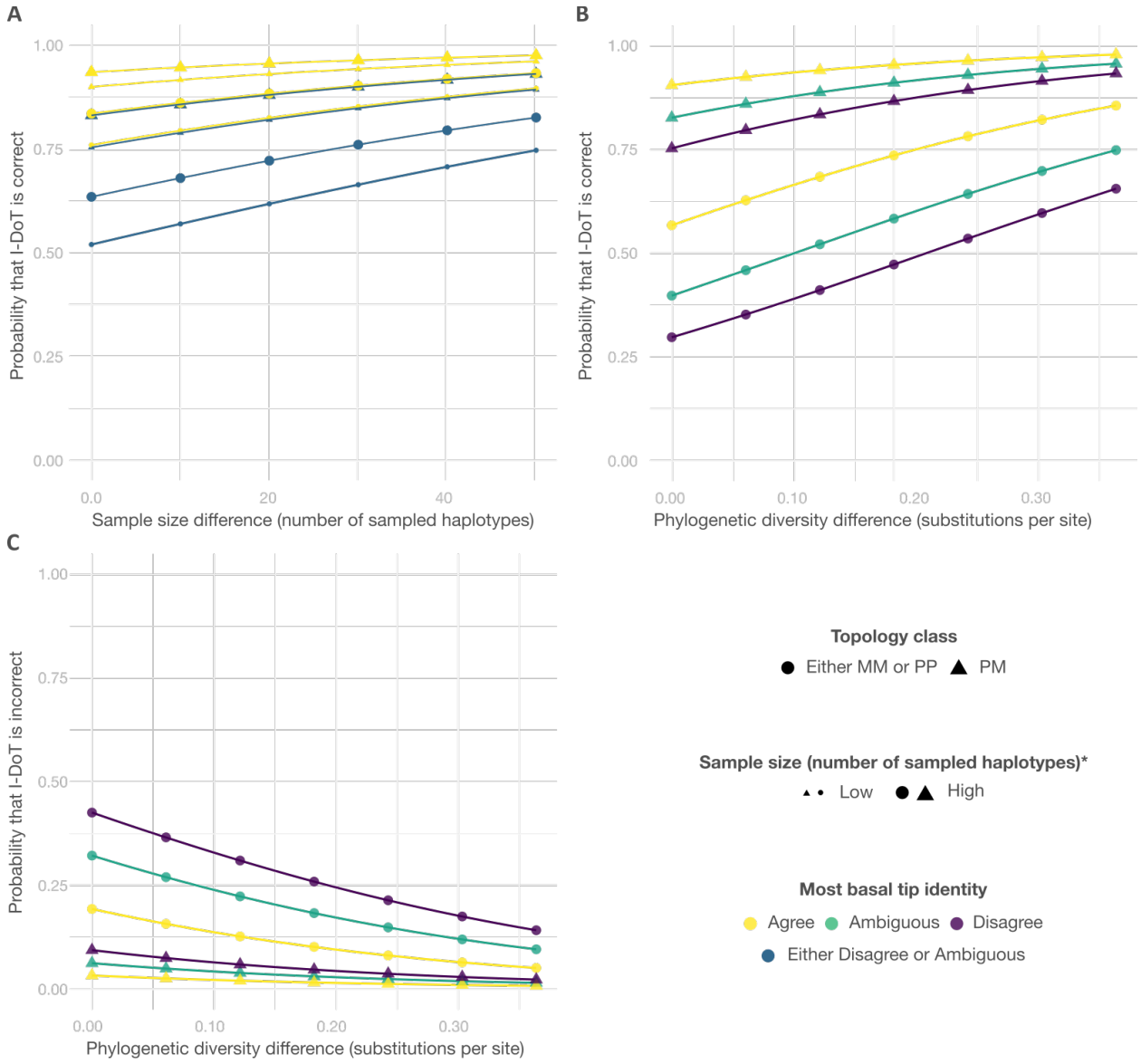


**Figure 3. Predicting the success of inferring the direction of transmission.** (A) The binary 'SP' model with four predictor covariates. (B) and (C) ordinal model 'P' model (with relaxed threshold for direction of transmission classification $t$=0.6) with three predictor covariates. (*) Sample size only applies to binary 'SP model' in (A).

## Discussion

We have combined empirical data on well-characterized HIV transmission pairs with statistical modelling to determine the conditions under which probabilistic phylogenetic analysis correctly infer the direction of HIV transmission. Our results suggest that, while ancestral state reconstruction correctly identifies the transmission direction in the majority of known transmission pairs, this success is determined by the epidemiological, sampling, genetic and phylogenetic characteristics of the individuals and their viral populations. We show that topological and branch-length metrics—such as root-to-tip distances—from the phylogenetic tree of the transmission pair, affect the chances of successfully inferring the transmission direction.

To guide future work on identifying the transmitting partner within a linked HIV pair, we quantified the probability of correctly inferring the transmission direction as a function of readily obtainable information. Under these circumstances, a PM topology and a match between the identity of the tip closer to the root (i.e., the one separated by the least number of internal nodes) and the identity of the state assigned to the root were highly predictive of inferring the correct transmission direction. This result agrees with the theoretical prediction that when multiple viral sequences per individual are available, the relative ordering of sequence clusters from the two individuals should inform DoT inference [9]. Moreover, our results suggest that using a relative metric of the difference in intra-host diversity between the partners improved discriminatory power (with larger differences indicative of a greater chance of correctly identifying the transmitting partner), which is consistent with previous work [11,12].

There is a noticeable drop in discriminatory power when our models only include readily obtainable data, which is likely due to the loss of discriminatory information. Indeed, two variables that are not typically known and are not included under our real-life conditions model—the recency of the transmitter's infection at the time of transmission and the time from transmission—have been shown to influence the topology class [8,9], which in turn influences the chance of correctly identifying the transmitting partner. In

the absence of such data, our results confirm that inferences about directionality can entail considerable uncertainty [31]. Nonetheless, our results shed light into the possible reasons for variability between studies. For instance, studies that most successfully infer the DoT used next generation sequencing data from heterosexual serodiscordant couples, where transmission occurs during the chronic stage of the transmitter [11,13] and there is a higher likelihood of a PM topology that our model suggests is indicative of correctly identifying the transmitting partner.

Here we show that even when we are conservative about attributing the DoT using ancestral state reconstruction, requiring 95% of the trees in a distribution to support a direction, a small percentage of cases still have an incorrect prediction. A recent study tested whether the prediction of the DoT could be improved by using NGS and in the best-case scenario the prediction was incorrect for 4/33 (12%) pairs [12]. Nonetheless, easily calculated measures of relative intra-host diversity can increase the probability that the transmitting partner is correctly identified (the difference in haplotype numbers, the difference in mean pairwise nucleotide diversity, the difference in phylogenetic diversity or the minimum root-to-tip distance). Our results suggest that these metrics can be evaluated for each pair concurrently with ancestral state reconstruction, which would provide a probability of incorrectly inferring DoT. In turn, these probabilities could be used to either select a subset of pairs for which a smaller probability of incorrect assignment is likely, or as weights to adjust further analysis which could be useful to move beyond identifying potential transmission pairs or clusters and allow to interpret transmission inferences by considering the degree of uncertainty in the DoT.

Our results suggest that there was little difference in the ordinal classification performance of ancestral state reconstruction methods (either probabilistic or parsimony-based algorithms) when a Maximum Likelihood tree is used. However, we did find differences in the nature of the data that were able to predict whether the inference was correct. That is, while differences in the sampled haplotypes remain high using a parsimony-based approach, only phylogenetic information is required for probability-based approaches. These differences likely occur because probabilistic methods, unlike parsimony-based algorithms,

incorporate information about branch lengths during the inference, which are indicators of nucleotide diversity.

This study has some limitations: first well characterized transmission pairs are scarce, and we were not able to test our models out of our sample but instead followed approaches to minimize overfitting and avoid biased estimates of model performance. Second, we used relational metrics that summarized the magnitude of differences in the intra-host diversity (i.e., differences in the number of sampled haplotypes, in intra-host nucleotide diversity, and in minimum root-to-tip distance), in the inter-host-diversity (i.e., differences in minimum patristic distance), or in a composite measure of diversity (e.g. differences in phylogenetic diversity). However, there are alternative ways to conceptualize diversity and there may be other factors that affect the DoT while not represented in the available data. Finally, we did not consider the effects of processes such as superinfection and recombination, which impact on diversity and phylogenetic interpretation.

While the use of phylogenetic analysis to infer the transmission, direction has recently shown promise, there has been considerable uncertainty about the consistency in accuracy across studies. Here we provide a statistical framework to help explain these differences and to improve the reliability in future work We stress that while phylogenies provide rich and important information about transmission, conclusions on directionality must be considered cautiously and with full adherence to the strictest ethical standards of data use.

**Acknowledgements**

CJVA and KEA were funded by an ERC Starting Grant (award number 757688) awarded to KEA. MH was funded by The HIV Prevention Trials Network (grant number H5R00701.CR00.01) and The Bill and Melinda Gates Foundation (grant number OPP1175094). JACB was supported by the MRC Precision Medicine Doctoral Training Programme (ref: 2259239). KAL was supported by The Wellcome Trust and The Royal Society grant no. 107652/Z/15/Z. JB received support from the UK MRC and the UK DFID

(#MR/R010161/1) under the MRC/DFID Concordat agreement and as part of the EDCTP2 Programme supported by the European Union.

## Competing Interest Statement

The authors declare no competing interests.

## References

1. Volz, E. M. & Frost, S. D. W. Inferring the source of transmission with phylogenetic data. *PLoS Comput. Biol.* **9**, e1003397 (2013).
2. Robert, A. *et al.* Determinants of Transmission Risk During the Late Stage of the West African Ebola Epidemic. *Am. J. Epidemiol.* **188**, 1319–1327 (2019).
3. Faye, O. *et al.* Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *Lancet Infect. Dis.* **15**, 320–326 (2015).
4. Lalor, M. K. *et al.* Recent household transmission of tuberculosis in England, 2010–2012: retrospective national cohort study combining epidemiological and molecular strain typing data. *BMC Medicine* vol. 15 (2017).
5. Rockett, R. J. *et al.* Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat. Med.* **26**, 1398–1404 (2020).
6. Prem, K. *et al.* Inferring who-infected-whom-where in the 2016 Zika outbreak in Singapore—a spatio-temporal model. *Journal of The Royal Society Interface* vol. 16 20180604 (2019).
7. Ratmann, O. *et al.* Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. *Lancet HIV* **7**, e173–e183 (2020).
8. Villabona-Arenas, C. J. *et al.* Number of HIV-1 founder variants is determined by the recency of the source partner infection. *Science* **369**, 103–108 (2020).
9. Romero-Severson, E. O., Bulla, I. & Leitner, T. Phylogenetically resolving epidemiologic linkage. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2690–2695 (2016).
10. Wymant, C. *et al.* PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Mol. Biol. Evol.* **35**, 719–733 (2018).
11. Zhang, Y. *et al.* Evaluation of Phylogenetic Methods for Inferring the Direction of Human Immunodeficiency Virus (HIV) Transmission: HIV Prevention Trials Network (HPTN) 052. *Clinical Infectious Diseases* (2020) doi:10.1093/cid/ciz1247.
12. Rose, R. *et al.* Phylogenetic Methods Inconsistently Predict the Direction of HIV Transmission Among Heterosexual Pairs in the HPTN 052 Cohort. *J. Infect. Dis.* **220**, 1406–1413 (2019).

13. Ratmann, O. *et al.* Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nature Communications* vol. 10 (2019).
14. Mutenherwa, F., Wassenaar, D. R. & de Oliveira, T. Ethical issues associated with HIV phylogenetics in HIV transmission dynamics research: A review of the literature using the Emanuel Framework. *Dev. World Bioeth.* **19**, 25–35 (2019).
15. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
16. Pupko, T., Pe'er, I., Shamir, R. & Graur, D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.* **17**, 890–896 (2000).
17. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* vol. 35 526–528 (2019).
18. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005 (1995).
19. Soubrier, J. *et al.* The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol.* **29**, 3345–3358 (2012).
20. Huelsenbeck, J. P. & Bollback, J. P. Empirical and Hierarchical Bayesian Estimation of Ancestral States. *Systematic Biology* vol. 50 351–366 (2001).
21. Ronquist, F. Bayesian inference of character evolution. *Trends Ecol. Evol.* **19**, 475–481 (2004).
22. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
23. Hanazawa, M., Narushima, H. & Minaka, N. Generating most parsimonious reconstructions on a tree: A generalization of the Farris-Swofford-Maddison method. *Discrete Applied Mathematics* vol. 56 245–265 (1995).
24. Narushima, H. & Hanazawa, M. A more efficient algorithm for MPR problems in phylogeny. *Discrete Applied Mathematics* vol. 80 231–238 (1997).
25. Orme, D. *et al. Caper: Comparative Analyses of Phylogenetics and Evolution in R*. (2018).
26. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* vol. 33 (2010).
27. Wurm, M. J., Rathouz, P. J. & Hanlon, B. M. Regularized Ordinal Regression and the ordinalNet R Package. (2017).
28. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 58 267–288 (1996).
29. Robin, X. *et al.* pROC: an open-source package for R and S to analyze and compare ROC curves. *BMC Bioinformatics* vol. 12 (2011).
30. Wei, R., Wang, J. & Jia, W. *multiROC: Calculating and Visualizing ROC and PR Curves Across Multi-Class Classifications*. (2018).
31. Wu, J. *et al.* The inference of HIV-1 transmission direction between HIV-1 positive couples based on the sequences of HIV-1 quasi-species. *BMC Infect. Dis.* **19**, 566 (2019).