

Methods

Model overview

For each sampled pair of cases i and j , **EpiLink** evaluates whether the observed testing-time difference and consensus-level genetic distance are compatible with a finite set of recent latent transmission histories. Let

$$t_{ij} = t_{\text{test},j} - t_{\text{test},i} \quad (1)$$

denote the observed difference in testing times, and let g_{ij} denote the observed consensus-level genetic distance. Candidate latent histories are drawn from

$$\mathcal{S}_M = \{H_{\text{AD}}(m) : m = 0, \dots, M\} \cup \{H_{\text{CA}}(m_i, m_j) : m_i, m_j \geq 0, m_i + m_j \leq M\}. \quad (2)$$

Here, $H_{\text{AD}}(m)$ is an ancestor–descendant history with m unsampled intermediates, and $H_{\text{CA}}(m_i, m_j)$ is a common-ancestor history in which the two sampled cases descend from a shared unsampled source with branch depths m_i and m_j . Direct transmission and co-primary infection are recovered as the special cases $H_{\text{AD}}(0)$ and $H_{\text{CA}}(0, 0)$, respectively (Fig. 1).

Latent transmission histories used by EpiLink

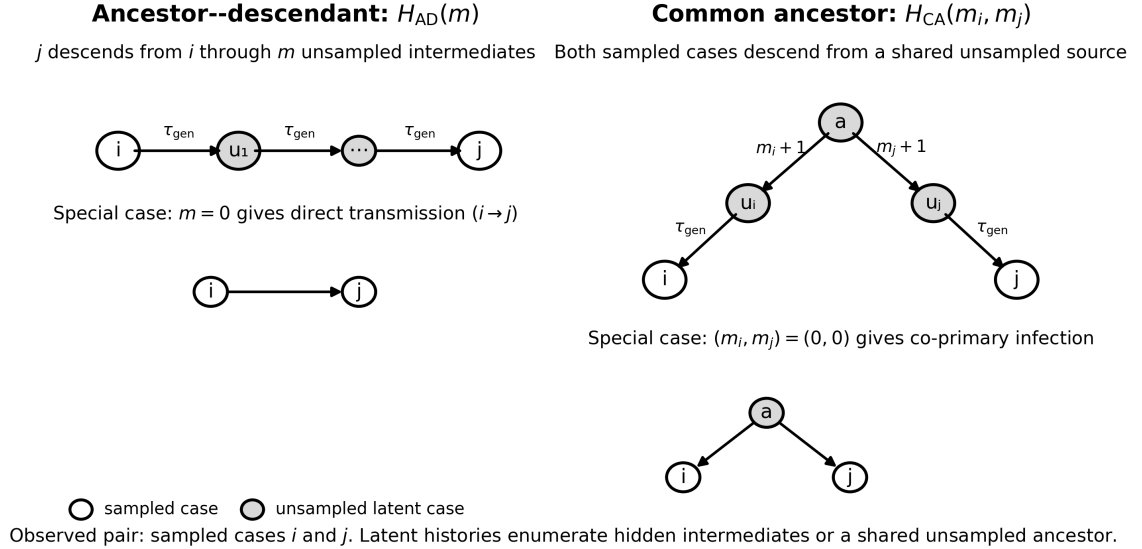


Figure 1: **Latent transmission histories considered by EpiLink.** Schematic representation of the two families of latent transmission histories linking a sampled pair of cases i and j . In the ancestor–descendant history $H_{\text{AD}}(m)$, case j descends from case i through m unsampled intermediates; the special case $m = 0$ corresponds to direct transmission. In the common-ancestor history $H_{\text{CA}}(m_i, m_j)$, both sampled cases descend from a shared unsampled source, with branch depths m_i and m_j ; the special case $(m_i, m_j) = (0, 0)$ corresponds to co-primary infection. Open circles denote sampled cases and shaded circles denote unsampled latent cases. These latent histories define the scenario set over which temporal and genetic compatibility scores are evaluated.

The temporal component is based on an $E/P/I$ infection-process model with Gamma-distributed latent, presymptomatic infectious, and symptomatic infectious stages [Hart et al., 2021], together with a Gamma-distributed delay from symptom onset to testing. Writing $\tau_{\text{inc}} = y_E + y_P$ for the incubation period and $\tau_{\text{gen}} = \tau_{\text{inc}} + y_{\text{toit}}$ for the generation interval, the model-implied testing-time difference under

scenario s is

$$T_s = A_j(s) - A_i(s) + \tau_{\text{inc},j} + x_{\text{test},j} - \tau_{\text{inc},i} - x_{\text{test},i}, \quad (3)$$

where $A_k(s)$ is the elapsed time from the latent reference point to infection of sampled case k . Under an ancestor–descendant history,

$$A_i(H_{\text{AD}}(m)) = 0, \quad (4)$$

$$A_j(H_{\text{AD}}(m)) = \sum_{r=0}^m \tau_{\text{gen},r}, \quad (5)$$

so that

$$T_{ij}(H_{\text{AD}}(m)) = \sum_{r=0}^m \tau_{\text{gen},r} + \tau_{\text{inc},j} + x_{\text{test},j} - \tau_{\text{inc},i} - x_{\text{test},i}. \quad (6)$$

Under a common-ancestor history,

$$A_i(H_{\text{CA}}(m_i, m_j)) = \sum_{r=1}^{m_i+1} \tau_{\text{gen},r}^{(i)}, \quad (7)$$

$$A_j(H_{\text{CA}}(m_i, m_j)) = \sum_{s=1}^{m_j+1} \tau_{\text{gen},s}^{(j)}, \quad (8)$$

giving

$$T_{ij}(H_{\text{CA}}(m_i, m_j)) = \sum_{s=1}^{m_j+1} \tau_{\text{gen},s}^{(j)} - \sum_{r=1}^{m_i+1} \tau_{\text{gen},r}^{(i)} + \tau_{\text{inc},j} + x_{\text{test},j} - \tau_{\text{inc},i} - x_{\text{test},i}. \quad (9)$$

For each scenario, Monte Carlo simulation yields draws from the induced temporal distribution.

The genetic component assumes that within-host divergence is negligible on the timescales of interest and that consensus divergence accumulates along transmission-linked lineages. Let B_s denote the effective transmission-related branch length under scenario s . For ancestor–descendant histories,

$$B_{ij}(H_{\text{AD}}(m)) = \sum_{r=0}^m \tau_{\text{gen},r} + \tau_{\text{inc},j} + x_{\text{test},j} - (\tau_{\text{inc},i} + x_{\text{test},i}), \quad (10)$$

whereas for common-ancestor histories,

$$B_{ij}(H_{\text{CA}}(m_i, m_j)) = \sum_{r=1}^{m_i+1} \tau_{\text{gen},r}^{(i)} + \sum_{s=1}^{m_j+1} \tau_{\text{gen},s}^{(j)} + \tau_{\text{inc},i} + x_{\text{test},i} + \tau_{\text{inc},j} + x_{\text{test},j}. \quad (11)$$

If r is the median substitution rate per site per year and L is genome length, then the corresponding per-genome daily rate is

$$\lambda = \frac{rL}{365}. \quad (12)$$

Given branch-length draw $B_s^{(n)}$, the expected genetic draw is

$$G_s^{(n)} = \lambda B_s^{(n)}, \quad (13)$$

or, under a stochastic mutational process,

$$G_s^{(n)} \mid B_s^{(n)}, \lambda \sim \text{Poisson}(\lambda B_s^{(n)}). \quad (14)$$

A relaxed uncorrelated log-normal clock can be used by drawing $r^{(n)}$ and replacing λ with $\lambda^{(n)} = r^{(n)}L/365$.

Observed temporal and genetic distances are scored against the simulated scenario-specific distributions using percentile-based compatibility measures. For any observed quantity x_{obs} and scenario-specific Monte Carlo draws $x_1^{(s)}, \dots, x_{n_s}^{(s)}$, define

$$p_s(x_{\text{obs}}) = \frac{\#\{i : x_i^{(s)} \leq x_{\text{obs}}\}}{n_s}, \quad C_s(x_{\text{obs}}) = 1 - 2|p_s(x_{\text{obs}}) - 0.5|. \quad (15)$$

Applying this separately to the temporal and genetic components yields $C_{T,s}(i, j)$ and $C_{G,s}(i, j)$, and the overall compatibility for scenario s is

$$C_s(i, j) = C_{T,s}(i, j) C_{G,s}(i, j). \quad (16)$$

To represent broader hypotheses, scores may then be summed across any user-defined target subset $\mathcal{S}_\star \subseteq \mathcal{S}_M$:

$$\text{score}_{\mathcal{S}_\star}(i, j) = \sum_{s \in \mathcal{S}_\star} C_s(i, j). \quad (17)$$

Because scenario-specific temporal and genetic draws are precomputed and cached, pairwise evaluation requires only percentile lookups against stored Monte Carlo samples, making the method scalable to large datasets.

References

William S Hart, Philip K Maini, and Robin N Thompson. High infectiousness immediately before covid-19 symptom onset highlights the importance of continued contact tracing. *eLife*, 10, April 2021. ISSN 2050-084X. doi: 10.7554/elife.65534. URL <http://dx.doi.org/10.7554/eLife.65534>.