

# Derivation of EpiLink compatibility score

For a sampled pair  $(i, j)$ , let  $t_{ij}$  denote the testing-time difference and  $g_{ij}$  the observed consensus-level genetic distance. EpiLink compares a finite set of recent latent transmission scenarios linking the pair, computes a compatibility score for each scenario, and reports the compatibility of a user-specified target scenario.

## Observed quantities and timing model

Let  $t_{\text{test},i}$  and  $t_{\text{test},j}$  denote the testing dates, measured in days on a common time scale, and define

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

Let  $g_{ij}$  denote the observed consensus-level genetic distance, measured in nucleotide differences.

We implement the variable infectiousness  $E/P/I$  model of Hart et al. [2021]. Infection is partitioned into latent ( $E$ ), presymptomatic infectious ( $P$ ), and symptomatic infectious ( $I$ ) stages with Gamma-distributed durations  $y_E$ ,  $y_P$ , and  $y_I$ , and we extend this temporal model with a Gamma-distributed testing delay  $x_{\text{test}}$  from symptom onset to testing. Writing  $y_{\text{toit}}$  ( $y^*$  in Hart et al. [2021]) for the time from onset of infectiousness to transmission, the incubation period and generation interval are

$$\tau_{\text{inc}} = y_E + y_P, \quad (2)$$

$$\tau_{\text{gen}} = \tau_{\text{inc}} + y_{\text{toit}}, \quad (3)$$

so that the total time from infection to testing is  $\tau_{\text{inc}} + x_{\text{test}}$ . The temporal model therefore provides latent random variables such as incubation periods, generation intervals, and testing delays, which are integrated over by Monte Carlo.

## Latent transmission scenarios

Let  $M$  denote the maximum hidden depth allowed in the latent relationship between cases  $i$  and  $j$ . The candidate set is

$$\mathcal{S}_M = \{H_{\text{AD}}(m) : m = 0, \dots, M\} \cup \{H_{\text{CA}}(m_i, m_j) : m_i, m_j \geq 0, m_i + m_j \leq M\}. \quad (4)$$

Here,  $H_{\text{AD}}(m)$  denotes an ancestor–descendant scenario in which case  $j$  descends from case  $i$  through  $m$  unsampled intermediates, whereas  $H_{\text{CA}}(m_i, m_j)$  denotes a common-ancestor scenario in which both sampled cases descend from an unsampled source with branch depths  $m_i$  and  $m_j$ . The cases  $H_{\text{AD}}(0)$  and  $H_{\text{CA}}(0, 0)$  correspond to direct transmission and co-primary infection, respectively.

## Temporal draws

For a given scenario  $s \in \mathcal{S}_M$ , let  $t_x$  denote the infection time of the latent reference point: the infection time of case  $i$  under an ancestor–descendant scenario, or the infection time of the shared unsampled

source under a common-ancestor scenario. Let  $A_k(s)$  denote the elapsed time from this reference point to infection of sampled case  $k \in \{i, j\}$ . Then

$$t_{\text{test},k} = t_x + A_k(s) + \tau_{\text{inc},k} + x_{\text{test},k}, \quad (5)$$

so the model-implied testing-time difference is

$$T_s = A_j(s) - A_i(s) + \tau_{\text{inc},j} + x_{\text{test},j} - \tau_{\text{inc},i} - x_{\text{test},i}. \quad (6)$$

Under  $H_{\text{AD}}(m)$ ,

$$A_i(H_{\text{AD}}(m)) = 0, \quad (7)$$

$$A_j(H_{\text{AD}}(m)) = \sum_{r=0}^m \tau_{\text{gen},r}, \quad (8)$$

and therefore

$$T_{ij}(H_{\text{AD}}(m)) = \sum_{r=0}^m \tau_{\text{gen},r} + \tau_{\text{inc},j} + x_{\text{test},j} - \tau_{\text{inc},i} - x_{\text{test},i}. \quad (9)$$

Under  $H_{\text{CA}}(m_i, m_j)$ ,

$$A_i(H_{\text{CA}}(m_i, m_j)) = \sum_{r=1}^{m_i+1} \tau_{\text{gen},r}^{(i)}, \quad (10)$$

$$A_j(H_{\text{CA}}(m_i, m_j)) = \sum_{s=1}^{m_j+1} \tau_{\text{gen},s}^{(j)}, \quad (11)$$

so that

$$T_{ij}(H_{\text{CA}}(m_i, m_j)) = \sum_{s=1}^{m_j+1} \tau_{\text{gen},s}^{(j)} - \sum_{r=1}^{m_i+1} \tau_{\text{gen},r}^{(i)} + \tau_{\text{inc},j} + x_{\text{test},j} - \tau_{\text{inc},i} - x_{\text{test},i}. \quad (12)$$

For each scenario  $s$ , Monte Carlo simulation yields draws  $T_s^{(1)}, \dots, T_s^{(N)}$  from the induced temporal distribution.

## Genetic draws

The genetic component assumes that within-host divergence is negligible on the timescales of interest and that consensus divergence accumulates along transmission-linked lineages. Under scenario  $s$ , let  $B_s$  denote the effective transmission-related branch length, in days, separating the sampled genomes.

For ancestor–descendant scenarios,

$$B_{ij}(H_{\text{AD}}(m)) = \sum_{r=0}^m \tau_{\text{gen},r} + \tau_{\text{inc},j} + x_{\text{test},j} - (\tau_{\text{inc},i} + x_{\text{test},i}). \quad (13)$$

For common-ancestor scenarios,

$$B_{ij}(H_{\text{CA}}(m_i, m_j)) = \sum_{r=1}^{m_i+1} \tau_{\text{gen},r}^{(i)} + \sum_{s=1}^{m_j+1} \tau_{\text{gen},s}^{(j)} + \tau_{\text{inc},i} + x_{\text{test},i} + \tau_{\text{inc},j} + x_{\text{test},j}. \quad (14)$$

For each scenario  $s$ , Monte Carlo simulation yields draws  $B_s^{(1)}, \dots, B_s^{(N)}$ .

Let  $r$  denote the median substitution rate per site per year and  $L$  the genome length in sites. The corresponding per-genome daily substitution rate is

$$\lambda = \frac{rL}{365}. \quad (15)$$

Under a relaxed uncorrelated log-normal clock (UCLN), one instead samples

$$r^{(n)} \sim \text{LogNormal}(\mu_r, \sigma_r^2), \quad (16)$$

with  $\mu_r$  chosen so that the median equals the specified substitution rate  $r$ , and sets

$$\lambda^{(n)} = \frac{r^{(n)}L}{365}. \quad (17)$$

Given branch-length draw  $B_s^{(n)}$  and clock-rate draw  $\lambda^{(n)}$ , the genetic draw used for scoring is

$$G_s^{(n)} = \lambda^{(n)} B_s^{(n)} \quad (18)$$

under a deterministic mutational process, and

$$G_s^{(n)} \mid B_s^{(n)}, \lambda^{(n)} \sim \text{Poisson}(\lambda^{(n)} B_s^{(n)}) \quad (19)$$

under a stochastic mutational process.

## Compatibility scores

For any observed quantity  $x_{\text{obs}}$  and scenario  $s$  with Monte Carlo draws  $x_1^{(s)}, \dots, x_{n_s}^{(s)}$ , define the percentile score

$$p_s(x_{\text{obs}}) = \frac{\#\{i : x_i^{(s)} \leq x_{\text{obs}}\}}{n_s}, \quad (20)$$

and the corresponding compatibility score

$$C_s(x_{\text{obs}}) = 1 - 2|p_s(x_{\text{obs}}) - 0.5|. \quad (21)$$

Applying this definition to the temporal draws gives

$$p_{T,s}(i, j) = \frac{\#\{n : T_s^{(n)} \leq t_{ij}\}}{N}, \quad C_{T,s}(i, j) = 1 - 2|p_{T,s}(i, j) - 0.5|, \quad (22)$$

and applying it to the genetic draws gives

$$p_{G,s}(i, j) = \frac{\#\{n : G_s^{(n)} \leq g_{ij}\}}{N}, \quad C_{G,s}(i, j) = 1 - 2|p_{G,s}(i, j) - 0.5|. \quad (23)$$

The overall compatibility assigned to scenario  $s$  is the product of the temporal and genetic compatibilities,

$$C_s(i, j) = C_{T,s}(i, j) C_{G,s}(i, j). \quad (24)$$

## Target subset and computation

Let  $\mathcal{S}_\star \subseteq \mathcal{S}_M$  denote a user-specified target subset chosen when the **EpiLink** object is constructed. The reported target score is

$$\text{score}_{\mathcal{S}_\star}(i, j) = \sum_{s \in \mathcal{S}_\star} C_s(i, j). \quad (25)$$

When  $\mathcal{S}_\star$  contains a single element, this reduces to the original single-scenario score. The implementation also returns  $C_s(i, j)$  for every  $s \in \mathcal{S}_M$ . For common-ancestor scenarios, the labels are ordered:  $H_{\text{CA}}(m_i, m_j)$  denotes a scenario in which case  $i$  is  $m_i$  generations and case  $j$  is  $m_j$  generations from their most recent common ancestor. Because temporal compatibility depends on the signed testing-time difference  $t_{ij} = t_{\text{test},j} - t_{\text{test},i}$ , the scenarios  $H_{\text{CA}}(m_i, m_j)$  and  $H_{\text{CA}}(m_j, m_i)$  are generally not equivalent and may receive different compatibilities. If an unordered interpretation is desired for a mirrored common-ancestor relationship, the user can include both orderings in  $\mathcal{S}_\star$ .

For each scenario  $s \in \mathcal{S}_M$ , **EpiLink** precomputes and stores  $N$  draws of  $T_s$  and  $B_s$  at construction time. It then converts the cached branch-length draws into cached mutation draws  $G_s^{(n)}$  using the chosen mutational process. Each pairwise evaluation therefore only requires percentile calculations against the cached draws, which is well suited 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>.