Supporting Information File 3: Validation of the mathematical model using a North American cohort

To observe how well our model performs when applied to other cohorts, we used data from a previously described cohort where daily EBV shedding in the oral mucosa was measured in 26 adult men from Seattle, Washington [?]. The distribution of viral loads found in positive swabs for each participant is shown in Fig A.

Fig A. Distribution of virus loads detected in oral swabs from participants of the Seattle cohort. Each box and whisker represents the distribution of the viral loads in daily oral swabs testing positive for EBV for an individual participant. Since many swabs did not test positive for EBV via qPCR, the percentage of a participant’s oral samples that did test positive is indicated by the colour of the box. The threshold of detection (150 copies/ml) is indicated by the horizontal red line.

We applied the ABC fitting algorithm, as described in the Methods of the main paper, to fit individual participant data from the Seattle Study to our model. We were able to determine parameter fits for 23 of the 26 participants. Those individuals whose data could not be fit either had no EBV viral shedding detected (1 participant), only 1 positive swab for EBV (1 participant) or an insufficient number of data points (only three swabs for 1 participant). The distribution of $\rho$ for each participant is shown in Fig B. We found that the $\rho$ values are similar for participants in both studies. While the mean $\rho$ value was significantly lower ($p<0.001$) for participants in our Uganda study (mean of 0.19) than for participants in the Seattle study (mean of 0.33), this was mainly due to the model fitting better to individuals with higher viral loads. When continuously correcting for unit increases in the log$_{10}$ of the participants’ median EBV load, there was no significant difference between $\rho$ values from Ugandan participants and the participants in the Seattle cohort (mean increase in $\rho$ of 0.05 for Seattle study compared to our study, 95% CI = [-0.01,0.11], p-value = 0.096).
Fig B. Distribution of the goodness of fit (\(\rho\)) statistics for participants of the Seattle study.
The data from 23 of the 26 Seattle study participants were able to fit to our model. For each participant, the 1000 parameter sets that produced the best agreement between the summary statistics of the data and model simulations were chosen. The \(\rho\) values, a measure of the fit, is shown for each parameter set. \(\rho\) values for participants of the Seattle study are similar to those for participants of our study, indicating the model works equally well for both sets of data.

As described in the Methods of the main paper, we used importance sampling to determine distributions for the parameters governing infection traits in different participant groups. Specifically, we looked at the median number of tonsillar crypts actively infected in an individual (Fig Ca and Cb), the median amount of virus within an active crypt (Fig Cc and Cd), the rate of B cell reactivation leading to epithelial infection (\(b\)) (Fig Da and Db), and the rate at which EBV-specific cytotoxic T cells proliferate in response to infection (\(\theta\)) (Fig Dc and Dd).

Like in the Uganda cohort, HIV-1 coinfected Seattle cohort participants are expected to have more actively infected crypts and more virus per actively infected crypt than HIV-1 uninfected Seattle cohort participants (Fig Ca and Cc). Similarly, higher median viral loads are related to more infection and virus replication in crypts within the tonsils (Fig Cb and Cd). However, since Seattle cohort participants had significantly lower viral loads than participants of the Uganda cohort study, on average, fewer crypts within the tonsils are predicted to be actively infected, and each one produces lower viral loads.

When evaluating differences between the densities of parameters \(b\) and \(\theta\) in individuals of different HIV-1 statuses and with different median EBV loads, the Seattle study again gives similar predictions to those achieved with the Uganda data. Differences between groups of individuals within the Seattle cohort are shown in Fig D.
Fig C. Numbers of actively infected crypts and EBV viral loads from the Seattle study. We present distributions of the median number of crypts actively producing EBV within an individual, stratified by (a) HIV-1 status and (b) the log$_{10}$ median viral load of the individual. We also show distributions of the median amount of virus within a crypt actively producing virus stratified by (c) HIV-1 status and (d) log$_{10}$ median viral load. Directional arrows and numbers by figure legends indicate the probability that a randomly selected individual of one group has a higher parameter value (be it the number of active crypts or viral load per active crypt) than a randomly selected individual in a second group. Arrows show the direction of comparison.
Fig D. Distribution of parameters $b$ and $\theta$, stratified by HIV-1 status and median EBV viral load in the Seattle study. Fitting our mathematical model to participant data revealed that parameter $b$ is greater in HIV-1 coinfected participants (a), increasing with median viral load (b), and parameter $\theta$ is lower in HIV-1 coinfected participants (c), decreasing with median viral load (d). Directional arrows and numbers by figure legends indicate the probability that a randomly selected individual of one group has a higher parameter value (be it $b$ or $\theta$) than a randomly selected individual in a second group. Arrows show the direction of comparison.

The Seattle study also collected data on CD4+ T cell count and HIV-1 RNA load in the plasma of HIV-1 coinfected participants. We sought to understand how this data correlated with our predicted values of parameters $b$ and $\theta$. We performed importance sampling on the parameter sets chosen for each participant based on the corresponding goodness of fit ($\rho$) values to find the posterior distribution of $b$ and $\theta$ for each participant. We then took the median value for each participant and, using generalized linear models, determined how each 100-cell increase in CD4+ T cell count and each log$_{10}$ increase in HIV-1 RNA load affected these values. Results are shown in Table A. While these traits did not cause significant changes to $b$ or $\theta$, they did cause changes in the same direction as those found when performing the same analysis on the
Uganda cohort data. The lack of significance may be due to the smaller cohort size of the Seattle study.

**Table A. Effect of CD4+ T cell count and HIV-1 load on the median value of parameter \( b \) and \( \theta \) for individuals in the Seattle study.** In HIV-1 co-infected participants, median values of parameters \( b \) and \( \theta \) are influenced by the CD4+ T cell count and HIV-1 load. Fold-change is shown.

<table>
<thead>
<tr>
<th>Trait</th>
<th>FC</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>HIV-1 RNA</td>
<td>0.33</td>
<td>0.12-0.95</td>
</tr>
<tr>
<td></td>
<td>CD4+ T cell</td>
<td>0.96</td>
<td>0.52-1.75</td>
</tr>
<tr>
<td>( b )</td>
<td>HIV-1 RNA</td>
<td>1.00</td>
<td>0.66-1.50</td>
</tr>
<tr>
<td></td>
<td>CD4+ T cell</td>
<td>0.91</td>
<td>0.74-1.11</td>
</tr>
</tbody>
</table>

**References**