**Supplementary Methods**

*Cell Sorting of HLA-DR+/- Memory CD4+ T-Cells*

Peripheral blood mononuclear cells (PBMCs) were obtained from leukapheresis samples by Lymphoprep (STEMCELL) density gradient centrifugation followed by isolation of total CD4+ T-cells by magnetic negative selection, according to the manufacturer’s protocol (Stem Cell Technologies). Cells were sorted using the following antibodies: CD3-Brilliant Violet 711 (clone OKT3, BioLegend), CD4-APC-eFluor 780 (clone OKT4, ThermoFisher), CD14-V500 (clone M5E2, BD# 561391), LIVE/DEAD Aqua marker (Invitrogen# L34957), CD45RA-PE-CF594 (clone HI100, BD Biosciences), and HLA-DR-Brilliant Violet 421 (clone L243, BioLegend). CD3+CD4+ T-cells were gated on memory cells defined as CD45RA negative. HLA-DR+/- cells were then sorted for on a BD FACS ARIA-II (Supplementary Figure 1). The HLD-DR+/-populations were sorted to >99% purity. Following sorting, the cell subsets were processed at 4oC and stored as a dry cell pellet at -80oC until analysis.

*Statistical Methods*

Infection frequencies across memory cell subsets and participants were compared with a multivariable logistic regression on the number of cells positive for virus (or intact virus where relevant), and the number of cells used in the FLIPs assay for that participant and cell subset. All intact sequences were considered for this calculation, including those that were 100% identical to another intact sequence. The covariates in this regression were participant, and memory subset, and an interaction term to test effect modification between these. Average infection frequencies for each cell subset were calculated by calculating the infection frequency for each subset for each participant, and then taking the mean across participants. Note that while the overall conclusions of our analysis of this data remain the same as those found previously [8]*,* estimates of infection frequency (total and intact) for individual participants and subsets have changed due to improving the way this data is analysed. Previously, infection frequency was calculated for each sequence run and this estimate was analysed with a linear model that was used to predict average infection frequency within each participant / cell subset. This article instead uses a logistic model on the number of cells positive and cells used, effectively weighting the estimate of infection frequency by the number of cells used in each sequencing run to derive that estimate. A consequence of this improved method is improved precision and improved power to detect effect modification between participants and memory subsets. This has resulted in infection frequencies being calculated individually for each participant as opposed to the average trend over all participants.

The proportion of intact viral sequences, sequences resulting from an expansion event, and tropism was compared across participants and cell subsets with a logistic regression. The covariates were participants, cell subsets, and interaction between these two terms to test for effect modification. To test the relationship between intactness and tropism/expansion event, a Fisher exact test was performed. Statistical analysis was carried out in *R: A language and environment for statistical computing*, version 3.4.3.