

Machine learning approaches identify immunologic signatures of total and intact HIV DNA during long-term antiretroviral therapy

Reviewed Preprint

v2 • August 6, 2024

Revised by authors

Reviewed Preprint

v1 • March 22, 2024

Lesia Semenova, Yingfan Wang, Shane Falcinelli, Nancie Archin, Alicia D Cooper-Volkheimer, David M Margolis, Nilu Goonetilleke, David M Murdoch, Cynthia D Rudin , Edward P Browne 

Department of Computer Science, Duke University • UNC HIV Cure Center UNC Chapel Hill • Department of Microbiology and Immunology, UNC Chapel Hill • Department of Medicine, UNC Chapel Hill • Department of Medicine, Duke University

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

Understanding the interplay between the HIV reservoir and the host immune system may yield insights into HIV persistence during antiretroviral therapy (ART) and inform strategies for a cure. Here, we applied machine learning approaches to cross-sectional high-parameter HIV reservoir and immunology data in order to characterize host-reservoir associations and generate new hypotheses about HIV reservoir biology. High-dimensional immunophenotyping, quantification of HIV-specific T cell responses, and measurement of genetically intact and total HIV proviral DNA frequencies were performed on peripheral blood samples from 115 people with HIV (PWH) on long-term ART. Analysis demonstrated that both intact and total proviral DNA frequencies were positively correlated with T cell activation and exhaustion. Years of ART and select bifunctional HIV-specific CD4 T cell responses were negatively correlated with the percentage of intact proviruses. A Leave-One-Covariate-Out (LOCO) inference approach identified specific HIV reservoir and clinical-demographic parameters, such as age and biological sex, that were particularly important in predicting immunophenotypes. Overall, immune parameters were more strongly associated with total HIV proviral frequencies than intact proviral frequencies. Uniquely, however, expression of the IL-7 receptor alpha chain (CD127) on CD4 T cells was more strongly correlated with the intact reservoir. Unsupervised dimension reduction analysis identified two main clusters of PWH with distinct immune and reservoir characteristics. Using reservoir correlates identified in these initial analyses, decision tree methods were employed to visualize relationships among multiple immune and clinical-demographic parameters and the HIV reservoir. Finally, using random splits of our data as training-test sets, machine learning algorithms predicted with approximately 70% accuracy whether a given participant had qualitatively high or low levels of total or intact HIV DNA. The techniques described here may be useful for assessing global patterns within the increasingly high-dimensional data used in HIV reservoir and other studies of complex biology.

eLife assessment

Semenova et al. have studied a large cross-sectional cohort of people living with HIV on suppressive antiretroviral therapy and performed high dimensional flow-cytometry for analysis with data science/machine learning approaches to investigate associations of immunological and clinical parameters and intact/total HIV DNA levels (and categorizations). The study is **useful** in introducing these new methods and large data set and appears mostly **solid**, though some of the claims were **incompletely** supported by the modeling results. The authors have revised the text to fairly reflect their results, yet open questions remain about utility, particularly as to the value of categorical classification (vs continuous measurement) of reservoir size.

<https://doi.org/10.7554/eLife.94899.2.sa2>

Introduction

The advent of ART has significantly decreased morbidity and mortality associated with HIV infection. However, a long-lived proviral reservoir precludes a cure of infection (1–4). While a number of cell types have been proposed as contributing to the HIV reservoir, the largest and most well-characterized reservoir is within CD4 T cells. Molecular analyses of residual HIV proviruses in CD4 T cells during ART have shown that the reservoir is comprised of a heterogeneous mixture of full-length replication-competent proviruses and a larger set of defective proviruses containing internal deletions or APOBEC-mediated hypermutation(s) (5–7). Infected cells can undergo clonal expansion and contraction over time in response to homeostatic cues or antigen-driven activation (8–10). In addition to the presence of a latent HIV reservoir, PWH on therapy exhibit a persistent level of T cell immune activation and other altered phenotypes compared to uninfected individuals, involving cells of both the innate and adaptive immune systems (11). Specifically, HIV infection is associated with increased levels of inflammatory biomarkers and increased immune cell activation (12–15). PWH also exhibit a depleted naïve T cell compartment (16) and increased expression of activation/exhaustion and senescence markers such as CD38, PD-1 and KLRG1 (17–21). The molecular mechanisms underlying persistent immune activation and dysfunction for PWH on ART are unclear, but it is possible that ongoing viral gene expression and reverse transcription generates ligands for pattern recognition receptors in the innate immune system (22). However, another important potential cause of ongoing immune activation during ART is persistent sequelae of the massive disruptions to T cell homeostasis, intestinal permeability and lymphoid structures that occur during untreated HIV infection (23–25).

It is possible that the HIV reservoir and the host immune system interact in multiple complex ways during therapy. The dynamics of how the HIV reservoir affects the immune system and vice versa are important to characterize and may lead to novel ideas to promote reservoir depletion or to inhibit immune dysfunction that may be related to ongoing HIV expression in persistently infected cells. A number of previous studies have examined the relationship between reservoir size and the immune system (21, 26–44). Consistent major findings from these studies include enrichment of the HIV reservoir in CD4 T cells expressing activation and exhaustion markers, positive correlation of the reservoir with central memory T cells (T_{cm}) and transitional memory (T_{tm}) T cell subsets, and inverse correlation of the HIV reservoir with the CD4 nadir and the CD4/CD8 T cell ratio. Recent studies indicate enrichment of the intact HIV reservoir in cells expressing VLA-4, but no correlation with classic T cell activation markers (26, 36). One important limitation of most of these studies is that they were not able to distinguish intact

proviruses from defective proviruses, making the relationship between the immune system and the intact reservoir unclear. Furthermore, most of these studies examined a limited number of surface markers and may have missed associations that would be revealed by a higher-resolution examination of immunophenotypes. Additional assessment of immune correlates of the intact HIV reservoir in additional cohorts is thus needed. Finally, there is a recent, growing appreciation for the profound influence of ART initiation timing and length of treatment on the size and composition of the HIV reservoir that must be accounted for in assessing persistent immune activation in PWH on therapy (11 [↗](#), 45 [↗](#)).

In this study, we use a mixture of traditional statistical approaches and novel data science methods to assess the relationship between the immune system and the HIV reservoir across a cross-sectional study of 115 PWH. We identify numerous significant correlations between immune cell populations and the size of the total HIV reservoir, as well as smaller number of associations with the intact reservoir, including CD127 expression in CD4 T cells. This study highlights the potential of machine learning approaches to visualize global patterns in high-parameter studies of the HIV reservoir. Additionally, these approaches corroborate recent findings in the HIV persistence field regarding preferential intact proviral decay and immune dynamics and highlight the complex immunologic signatures of HIV latency.

Results

Cohort description, reservoir quantification, and immunophenotyping

This cross-sectional study evaluated 115 participants with HIV infection for at least one year and who had been receiving ART for at least 0.9 years (median 9 years, IQR 5.2-16.6) (**Table 1** [↗](#)). The study cohort was 77 percent male, 60 percent black and median participant age was 45 years.

The Intact Proviral DNA assay (IPDA) was performed on isolated total CD4 T cells (46 [↗](#)). The IPDA estimates of the frequency of total proviruses, intact proviruses, and the percent intact of total proviruses for each participant. Immunophenotyping was performed using 25-color spectral flow cytometry on peripheral blood mononuclear cells (PBMCs) from each study participant. A representative flow gating is shown in **Figure S1** [↗](#). Briefly, cells within the live lymphocytes gate were first defined as T cells (CD3+/CD56-), NK cells (CD3-/CD56+), or NKT cells (CD3+/CD56+). T cells were then further subdivided as CD4+ or CD8+ T cells, and then as naïve T cells (Tn, CD45RA+/CCR7+), central memory T cells (Tcm, CD45RA-/CCR7+), effector memory cells (Tem, CD45RA-/CCR7-) and terminally differentiated effector cells (Teff, CD45RA+/CCR7-). Expression of various surface markers within each of these subsets was then examined (CD38, HLA-DR, PD-1, KLRG1, CD127, CD27, NKG2A). In addition to measuring the baseline abundance of surface proteins, we also stimulated PBMCs from each participant with or without a pool of HIV-derived peptides and measured intracellular cytokines TNF α , IL-2, and IFN γ as well as the surface expression of the degranulation marker CD107a. HIV-specific T cell responses were defined as the difference in frequency of cytokine/CD107a+ cells between HIV stimulated and nonstimulated cells. Frequencies of CD4 or CD8 T cells that were single or double-positive for TNF α , IL-2, IFN γ , or CD107a were used for analysis.

These analyses resulted in the determination of a total of 144 parameters for the cohort of n = 115 PWH on ART. The 144 parameters include 133 immunophenotypic cell frequencies, 3 HIV reservoir parameters (frequency of intact HIV DNA, frequency of total HIV DNA, and percentage intact of total HIV DNA), and 8 clinical-demographic parameters. Clinical-demographic variables (**Table 1** [↗](#)) included age, biological sex, years of ART, estimated years of HIV infection prior to ART (here, we use categorical variables: years before ART=NA, years before ART< 1, years before ART $\geq$ 1), CD4

	Percentage (count)	Median	[Q1, Q3]	[Min, Max]
Age		45	[37, 53]	[23,65]
Sex (% male)	76.52% (88)			
Race				
Black	60% (69)			
White	37.39% (43)			
Other	2.61% (3)			
Years of HIV		11	[7, 19.85]	[1, 33.6]
Years before ART < 1	55.65% (64)			
Years before ART ≥ 1	38.26% (44)			
Years before ART=NA	6.09% (7)			
Years of ART		9	[5.23, 16.63]	[0.9,33.5]
Recent CD4 count		799	[624.5, 962]	[319, 1970]
CD4 Nadir		313.5	[163.25, 463.25]	[2, 1080]

Table 1

Participant demographic and clinical characteristics.

For demographics and clinical information, we report percentage for categorical variables, medians and [Q1, Q3] for real-value variables. ART is antiretroviral therapy. CD4 counts reported in cells/mm³. Years of HIV has 1 missing value, Years of ART has 7, and CD4 Nadir has 3; consequently, these missing values are not included in median and quantiles computations. Years before ART means years of HIV infection before ART initiation.

nadir, most recent CD4 T cell count, and race (Caucasian, African-American, other). No significant differences in the reservoir based on biological sex or race were observed (**Table S1**). The median and interquartile range (IQR) for all immune cells population variables are presented in **Table S2**.

Correlation of total, intact, and percent intact HIV DNA with immunophenotypes

For the initial analysis, Spearman correlation of the clinical-demographic and immunophenotypic data with HIV reservoir metrics was assessed. 69 host variables correlated (unadjusted $p < 0.05$) with one or more characteristics of the HIV reservoir (**Table S3**). Following Benjamini-Hochberg p -value adjustment for multiple comparisons using a false discovery rate correction at 5%, 31 variables were found to be correlated with total HIV DNA, 3 with intact HIV DNA, and 4 with the percentage of intact HIV DNA (**Table 2**). Of note, the correlations observed were generally weak to moderate in magnitude (the mean of the absolute values of Spearman correlation coefficients was 0.31 for total HIV DNA, 0.33 for intact HIV DNA, and 0.35 for percent intact HIV DNA).

With regard to clinical-demographic variables, CD4 nadir prior to ART initiation was negatively correlated with total HIV DNA (Spearman $r = -0.32$), whereas age ($r = 0.31$) and years of ART ($r = 0.31$) were positively correlated with total HIV DNA (**Table 2**). Years of ART ($r = -0.45$) was negatively correlated with the percentage of intact proviruses (**Table 2**), consistent with the hypothesis that intact proviruses are progressively eliminated from the reservoir over time. We then examined the association between the reservoir and time on ART further (**Figure 1**). Correlation plots of total reservoir frequency, intact reservoir frequency, and percentage intact HIV DNA versus years of ART were visualized (**Figure 1A-C**). The plots showed no significant correlation for intact HIV DNA frequency versus years of ART (**Figure 1B**), while total HIV DNA reservoir frequency was positively correlated with the length of ART treatment. (**Figure 1A**, Spearman $r = 0.31$). When we examined a plot of percent intact proviruses versus time on therapy (**Figure 1C**), we observed a biphasic decay pattern, consistent with previous reports (47–49). Indeed, when we fitted piece-wise linear functions with allowances for up to 2 breaks, the best-fit model had decay slope changes at 1.6 and 9 years of ART. The R^2 score of a piece-wise linear model with two breaks was 44.18%, while the score for a linear model without any breaks was 16.3%. In PWH within approximately 0-6 years of ART initiation, a significant proportion of the participants exhibited high fractions of intact proviruses (50-90%). After approximately 6 years of ART, however, the majority of participants had reservoirs for which intact proviruses represented a minor fraction of the overall reservoir (0-30%). Of note, several immune features also demonstrated statistically significant correlations with years of ART, including CD8 T cell expression of CD38 (negative correlation), CD8 Tcm frequency (positive correlation), and CD107a+IL-2+IFN γ -TNF α -CD8 T cells (positive correlation)(**Figures S2**, **S3**, **S4**, **Table S4**).

Examination of immunophenotypic parameter correlations with HIV DNA metrics demonstrated several notable correlations (**Table 2**, **Figures S2**, **S3**, **S4**). As previously reported, CD4 T cell frequencies were negatively correlated with total HIV DNA frequencies ($r = -0.36$), whereas CD8 T cells were positively correlated ($r = 0.41$). This relationship held true for intact HIV DNA frequencies but not for the percentage of intact HIV DNA (**Table 2**). Frequencies of CD4 and CD8 T naïve cells, as well as their expression of CD38, were significantly negatively correlated with total HIV DNA frequency, consistent with the hypothesis that larger reservoirs are associated with depletion of the naïve T cell subsets and increased immune activation. By contrast, CD4 and CD8 Tcm and Tem cells were weakly positively correlated with total HIV DNA, possibly due to increased differentiation of the naïve compartment into these memory subsets in people with larger reservoirs. Markers of T cell activation and exhaustion were weakly but significantly positively correlated with total HIV DNA, including NKG2A+ CD4 T, PD-1+ CD4 T, HLA-DR+ CD4 T, PD-1+ CD8 T, and PD-1+/CCR7+ CD8 T frequencies, among others (**Table 2**, **Figures S2**, **S3**, **S4**). For both intact and total HIV DNA, an inverse correlation with the CD4/CD8 T cell ratio was observed (**Figure S5**). Interestingly, for the percentage of intact HIV DNA, negative correlations

Variable	Total	Intact	Percent intact
%CD8 T	0.4052	0.3562	0.0068
%CD38+/HLA-DR- CD4 T	-0.3891	-0.1098	0.2664
%KLRG1-/PD-1- CD4 T	-0.3808	-0.1289	0.2334
%Tn CD4 T	-0.3802	-0.2031	0.1714
%NKG2A+ CD4 T	0.3618	0.2904	0.0179
%PD-1-/CCR7+ CD4 T	-0.3590	-0.1082	0.2283
%CD4 T	-0.3564	-0.3195	-0.0079
%Tcm CD8 T	0.3466	0.1752	-0.1814
%CD38+ CD4 T	-0.3366	-0.0611	0.2829
%PD-1-/CCR7- CD4 T	0.3300	0.1824	-0.0938
%CD38+/HLA-DR- CD8 T	-0.3267	-0.0837	0.2636
%PD-1+ CD4 T	0.3222	0.0664	-0.2470
Age	0.3172	0.1669	-0.1471
CD4 Nadir	-0.3164	-0.1512	0.1927
%PD-1+ Tn CD4 T	0.3119	0.1246	-0.0962
%Tn CD8 T	-0.3028	-0.2697	0.0154
Years of ART	0.3062	-0.0688	-0.4523
%PD-1+/CCR7+ CD8 T	0.2926	0.1361	-0.1254
%CD38-/HLA-DR+ CD4 T	0.2849	0.0920	-0.1500
%KLRG1-/PD-1- CD8 T	-0.2757	-0.2182	0.0162
%PD-1+ Tn CD8 T	0.2738	0.1411	-0.0705
%CD38+/HLA-DR- Tn CD8 T	-0.2676	-0.0610	0.2119
%PD-1+/CCR7+ CD4 T	0.2665	0.0859	-0.1673
%KLRG1+/CD27+ CD8 T	0.2606	0.2212	0.0319
%CD27+ CD4 T	-0.2602	-0.0531	0.1767
%KLRG1-/CD27+ CD4 T	-0.2575	-0.0464	0.2171
%HLA-DR+ CD4 T	0.2565	0.1110	-0.0970
%PD-1+ CD8 T	0.2541	0.1969	0.0035
%CD38+ CD8 T	-0.2418	0.0109	0.3114
%CD38-/HLA-DR+ Tn CD8 T	0.2404	0.0738	-0.1346
%Tem CD4 T	0.2402	0.0888	-0.1669
%CD127+ CD4 T	-0.1298	-0.3160	-0.2539
%CD107a-IFN γ -IL-2+TNF α + CD4 T	0.0568	-0.1775	-0.3223
%CD107a-IFN γ -IL-2+TNF α - CD4 T	0.0285	-0.2455	-0.3265

Table 2

PWH features correlate with HIV reservoir characteristics.

The abundance of 144 immune cell populations were determined by flow cytometry and the HIV reservoir was quantified by intact proviral DNA assay for a cohort of 115 people with HIV (PWH). Each abundance and clinical and demographic variable was correlated with total HIV reservoir frequency, intact reservoir frequency and the percentage of intact proviruses. Spearman correlation coefficients (**bold**) are shown for 36 variables that had significant p-values (< 0.05) after Benjamini-Hochberg correction for multiple comparisons. Each feature/subset is ranked by the absolute value of the correlation coefficient for the total reservoir frequency. For years of ART, we compute correlation based on 108 participants, excluding participants with missing years of ART values.

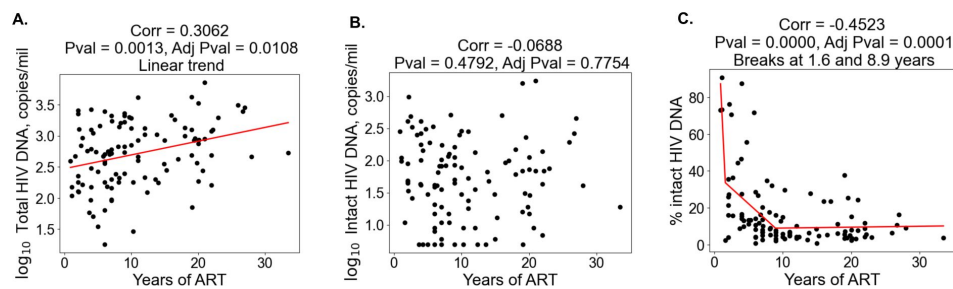


Figure 1

Duration of treatment and the HIV reservoir.

Scatterplots for years of ART vs total HIV reservoir frequency (**A**), intact reservoir frequency (**B**) and percent intact (**C**) are shown. Each dot represents an individual study participant. Correlation coefficients and corresponding p-values are shown for each plot. Participants that have missing values of years of ART were not included in the plot. For percent intact, piecewise linear function with two breaks is fitted. For total HIV reservoir frequency a linear function is fitted.

were observed with the frequencies of 2 subsets of bifunctional HIV peptide-stimulated CD4 T cells (IL-2+TNF α + and IFN γ +IL-2+), as well as with years on ART. Notably, the magnitude of correlation with immune parameters was typically stronger for total HIV DNA compared to intact HIV DNA (Table 2). This pattern may reflect a larger influence of total HIV DNA on immunophenotype of the host, due to the overall higher abundance of total HIV DNA relative to the intact reservoir. A notable exception to this pattern was CD127-CD4 T cell frequencies, which were significantly positively correlated with intact but not total HIV DNA.

Overall these findings support the notion that larger HIV reservoirs are associated with increased levels of immune activation and depletion of the naive T cell compartment. Furthermore, the data indicate that there may be some immune features, such as CD127 expression on CD4 T cells that are uniquely associated with the intact reservoir rather than the total reservoir.

Leave-One-Covariate-Out (LOCO) inference identifies specific HIV reservoir and clinical-demographic parameters important for the prediction of immunophenotypes

Given the evidence that reservoir and immune recovery dynamics likely occur in concert, we next used a variable importance approach - Leave-One-Covariate-Out (LOCO) inference analysis, to account for potential confounding variables and more carefully assess the relative importance of a given variable to an immunophenotype. The LOCO inference approach is described in Figure 2A. First, a least-squares regression linear model that predicts the dependent variable is fitted on a set of independent variables, and an R^2 value is generated (Figure 2A Step 1). Next, one independent variable is excluded from the model, the model is refitted, and an adjusted R^2 value is determined (Figure 2A Step 2). Finally, the change in R^2 scores, ΔR^2 , is calculated (Figure 2A Step 3). This process is then repeated for all independent variables in the model, and the ΔR^2 values are plotted in a heatmap for the variables of interest (Figure 2B-D).

We applied the LOCO inference analysis to all 133 immunophenotype parameters (dependent variables). Three separate analyses were conducted for a set of independent variables consisting of age, sex, years of ART, CD4 nadir, recent CD4 count, years of HIV infection prior to ART (=NA, < 1, $\geq$ 1), and one of the three HIV reservoir metrics (total HIV DNA, intact HIV DNA, or percentage intact HIV DNA). Multicollinearity analysis confirmed that independent variables from this analysis were not correlated with each other to a degree that would interfere with model generation (Table S5). Each analysis produced a heatmap of 133 dependent immunophenotype variables, with visualization of the variables for each immunophenotype (Figure S6, Tables S6-S8). For each of these analyses, we displayed the top 10 variables with the largest ΔR^2 value for LOCO inference incorporating either total (Figure 2B), intact (Figure 2C), or percentage intact (Figure 2D) HIV DNA.

The least-squares linear regression models were generally weakly explanatory (R^2 values in range [-11.11%, 35.36%]) of the global variability in immunophenotypes (Tables S6-S8, Figure 2B-D, Figure S7). Across all 3 analyses, age and race emerged as an important covariate of several immunophenotypes related to T cell subsets, particularly the naïve and central memory compartments, as well as KLRG-1, a marker of immune senescence (Figure 2 and Figure S6). Interestingly, biological sex was a key variable for model prediction of NK cell NKG2A expression, and race was important for Tem CD8, KLRG1+/CD27+ CD8 frequency for all three models. Recent CD4 count was important for predicting the frequencies of HLA-DR-CD8 Tn cells and CD4, CD8 T cell frequencies, but more so for the models incorporating total HIV DNA and percent intact than the intact HIV DNA frequency (Figure 2A-C). Total HIV DNA frequency was an important model input for prediction of PD-1+ CD4 T naïve cells, HIV-specific IL-2+TNF α -IFN γ -CD107-CD4 T cells, and CD4/CD8 T cell frequencies, among others (Figure 2B, Figure S6). Total HIV DNA frequency and intact HIV DNA frequency were also important predictors of CD4 T effector memory HLA-DR+CD38+ expression (Figure 2B and 2C).

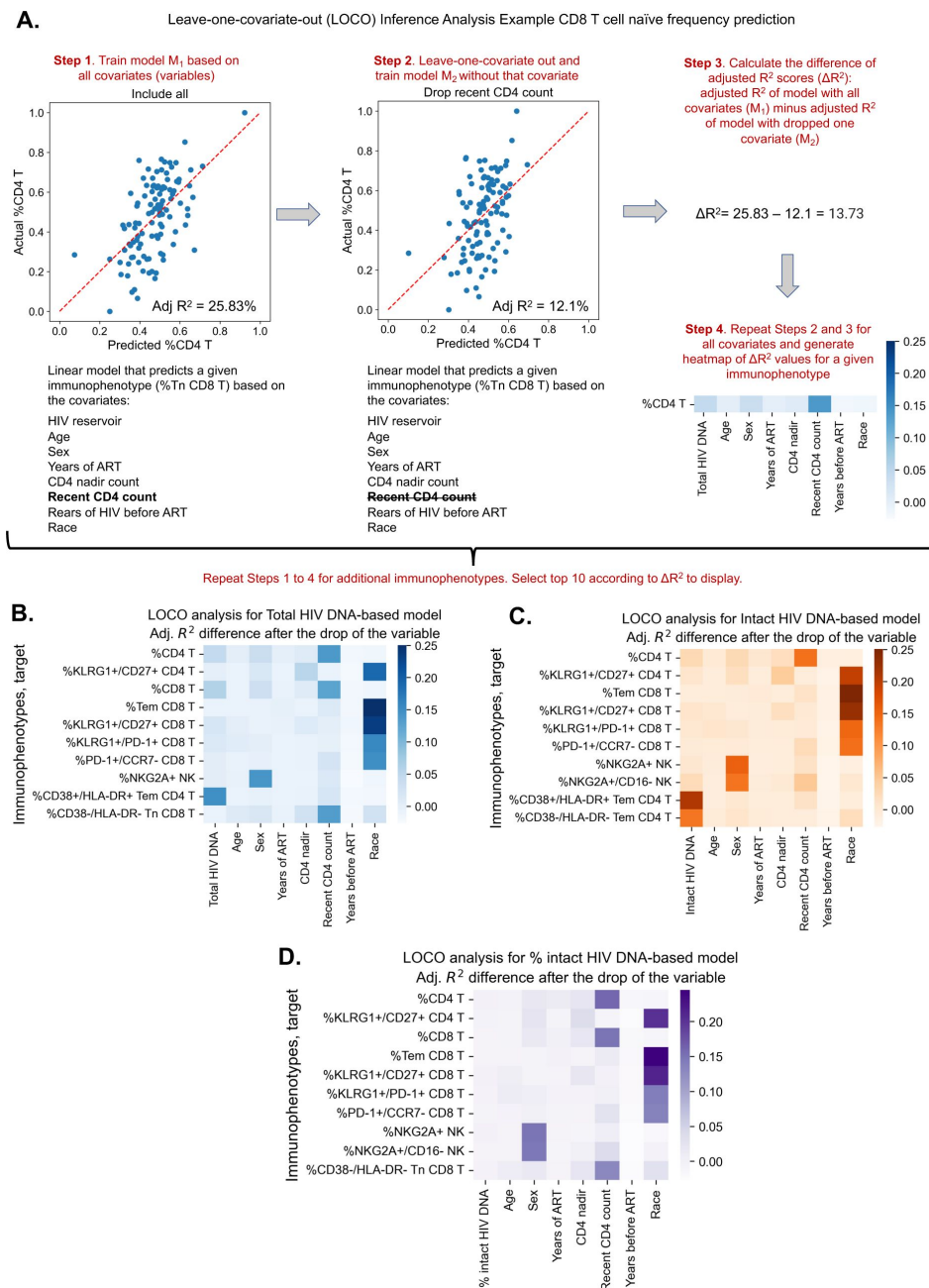


Figure 2

Leave-one-covariate-out (LOCO) analysis for clinical-demographic features and reservoir characteristics while predicting immunophenotypes.

A. Explanation of leave-one-covariate-out (LOCO) analysis based on example of %CD4 T for clinical-demographic features and reservoir characteristics while predicting immunophenotypes. Analysis was performed for all 133 immunophenotypes considered in the study. The top 10 biggest drops in adjusted R^2 scores are reported for models that use total reservoir frequency (**B.**), intact reservoir frequency (**C.**), or percent intact (**D.**) as features in addition to clinical and demographic information. Participants with missing years of ART values are excluded from this analysis. The missing value of the CD4 nadir for one participant is imputed. In **Figure S6**, the LOCO analysis results are visualized for all 133 immunophe-

In general, total and intact HIV DNA were predictive of similar immune features (**Figure S6** [↗](#), **Tables S6** [↗](#)–**S8** [↗](#)); however, total HIV DNA generally typically had greater magnitude ΔR^2 than intact HIV DNA. The percentage of intact HIV DNA did not appear to contribute greatly to the predictive power of the model for the variance in the immunophenotypes examined (**Figure S6C** [↗](#)), most likely due to its relatively high correlation with years of ART. Overall these findings further demonstrate the existence of a number of clinical parameters that are associated with immune signatures in PWH, including age, gender, and years of ART. Additionally, these results highlight key potential confounding variables that should be considered when interpreting the association of virological parameters with the immune system of PWH.

Receiver operator characteristic curve analysis demonstrates distinct immune parameters associated with intact and total reservoir frequencies

The complex relationship between the HIV reservoir and the immune system led us to evaluate if machine learning (ML) approaches that combine several parameters into models of the HIV reservoir could be a useful approach to understand this interaction. Before this could be attempted, we first identified a defined set of the most valuable variables from which ML models could be built. To achieve this, we first binarized the three HIV reservoir metrics (total HIV DNA frequency, intact HIV DNA frequency, %intact HIV DNA) into high (above median) or low (below median) reservoir groups and generated receiver operator (ROC) curves for the 144 clinical-demographic and immunophenotype parameters (**Figure 3** [↗](#)). The area under the curve (AUC) of the ROC curve indicates the ability of the feature (immune parameter or clinical/demographic information) to correctly identifying a participant as having qualitatively low or high total HIV DNA frequency (**Figure 3A** [↗](#)), intact HIV DNA frequency (**Figure 3B** [↗](#)), or percentage intact HIV DNA (**Figure 3C** [↗](#)). A model that randomly guesses high or low reservoir has an AUC of 0.5. In this analysis, 57 host variables had an AUC higher than 0.6 for one or more of total HIV DNA, intact HIV DNA, or percent intact DNA (**Table S9**). The most effective individual immune markers for classification of high versus low total HIV DNA frequency included %NKG2A⁺ CD4 T (AUC = 0.70), %PD-1⁺ Tn CD4 T (AUC = 0.68), and %CD38⁺/HLA-DR-CD8 T (AUC = 0.68). In contrast, the most effective markers for classifying based on the frequency of intact HIV DNA frequency included %CD127⁺ CD4 T (AUC = 0.71), %CD8 T (AUC = 0.66), and %Tn CD8 T (AUC = 0.65). Finally, for percentage intact HIV DNA, years of ART (AUC = 0.72), %CD107-IFN γ -IL-2+TNF α ⁺ CD4 T (AUC = 0.65) and %KLRG1⁺/PD1⁺ CD4 T (AUC = 0.65) were the most effective (**Figure 3** [↗](#)). Similar to the findings in the Spearman correlations (**Table 2** [↗](#)) and LOCO inference analysis (**Figure 2** [↗](#)), there were more variables that had AUC values > 0.6 for total reservoir frequency (**Table S9**): 44 variables above this threshold for the total reservoir, but only 19 variables for the intact reservoir and 23 variables for percent intact. Overall, this approach allowed us to derive a ranked list of the most predictive immune parameters for each aspect of the HIV reservoir, and these highly ranked features were thus used for subsequent dimension reduction and ML modeling.

Dimension reduction reveals two clusters of PWH with distinct HIV reservoirs

To examine the overall structure of dataset we employed an unsupervised dimension reduction (DR) machine learning approach (PaCMAP ([50](#) [↗](#))) using the 10 immune cell features with the highest AUC values for classifying participants based on total HIV DNA frequency, intact HIV DNA frequency, or percentage intact HIV DNA (**Figure 3** [↗](#), **Table S9** [↗](#)). Interestingly, while no clear clustering was observed when we use intact reservoir frequency or percentage intact HIV DNA associated features, we observed two distinct clusters of PWH (cluster 1 and cluster 2, **Figure 4A** [↗](#)) when using total HIV DNA associated features. These clusters were of roughly equal size with 50 participants in cluster 1 and 65 in cluster 2. Projecting total reservoir frequency (above or below the median) onto the clustering plot, we observe a strong distinction between the two

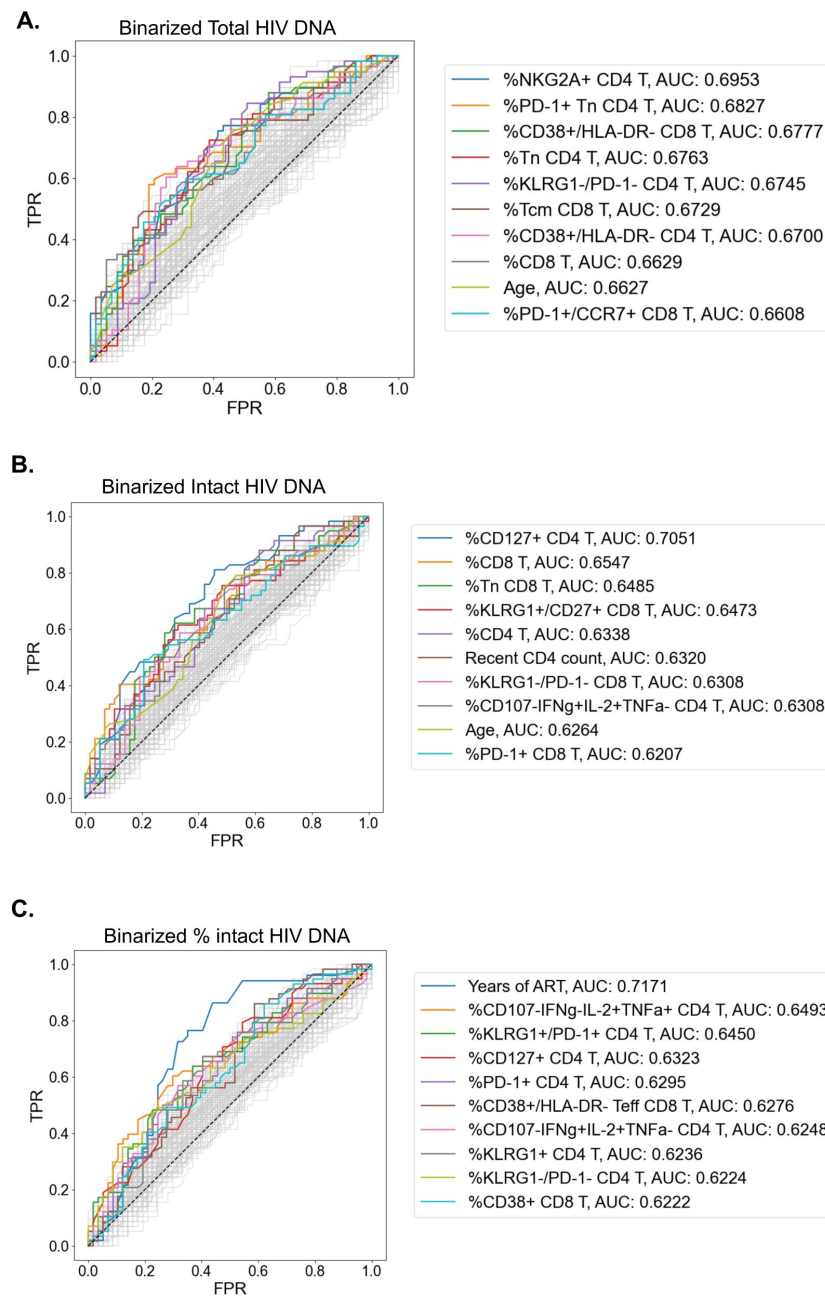


Figure 3

Receiver operating characteristic (ROC) curves identify PWH parameters that can classify reservoir characteristics.

For total reservoir frequency (**A**), intact reservoir frequency (**B**), and percent intact (**C**), ROC curves are plotted for all 144 immune markers, demographics, and clinical variables (shown in gray). Axes represent the true positive rate (TPR) and the false positive rate (FPR) for each variable for classifying study participants into low (below median) versus high (above median) reservoir frequency. ROC curves for ten variables with the highest AUC values are shown in color for each HIV reservoir characteristic. Striped black lines represent the ROC curves of a random model. For years of ART ROC curves, we exclude participants with missing years of ART val₂u₇es.

clusters in terms of total HIV DNA reservoir characteristics, (**Figure 4B,C**). Analysis of quantitative reservoir frequencies across clusters (**Figure 4C-E**) demonstrated that cluster 1 is characterized by a smaller total HIV reservoir frequency, but greater percentage of intact proviruses (**Figure 4C-E**). In contrast, cluster 2 is defined by a larger total reservoir frequency, but lower percentage of intact proviruses (**Figure 4C-E**). The frequency of intact proviruses also tended to be lower for cluster 1 but this difference was not significant. We also visualized the dataset using principal component analysis (PCA) (**Figure S9**), and observed that participants from each of the clusters identified by PaCMAP occupied different areas of the plot, although the separation was not as clear as with PaCMAP.

To gain insight into the immune cell features that distinguish these two clusters, we generated ROC curves for the 133 immune parameters based on their ability to identify membership in cluster 1 versus cluster 2 for each participant (**Figure 4F**). Many features that distinguished the clusters overlapped with the features selected for the clustering analysis, including expression patterns of CD38, HLA-DR, KLRG-1 and PD-1 on CD4 and CD8 T cells and CD4 T naïve cell frequencies. However, some novel features emerged, including CD4 KLRG1-CD27+ frequency, CD8 KLRG1-PD1-frequency, and CD8 Tn frequency (**Figure 4F** and **Figure S8**). Cluster 2, which had a higher total and intact absolute reservoir frequency but lower percentage of intact proviruses, was associated with a lower fraction of naïve CD4 and CD8 T cells. Additionally, cluster 2 had generally higher expression of immune exhaustion markers, including KLRG1 and PD-1 (**Figure S8**). These data indicate that cluster membership is highly associated with the overall level of immune activation and exhaustion for the individual. Notably, with regard to other clinical variables, cluster 2 had an older median age, longer time on ART, lower CD4 nadir, and lower current CD4 T cell count (**Figure S8**). We have previously shown that cannabis use is associated with lower levels of immune activation markers in PWH, despite having minimal impact on the size of the HIV reservoir (51). Interestingly, a subset of participants in this study had clinical data regarding cannabis (CB) use. CB users were enriched in cluster 1, which had lower levels of immune exhaustion markers (**Figure S8**). Overall, these data demonstrate that, within the cohort of PWH on long-term ART, there were two major clusters of participants with distinct reservoir characteristics and immunophenotypes.

Decision tree visualization of PWH with respect to reservoir characteristics

Since the interaction of the immune system and the HIV reservoir is multifactorial, we hypothesized that models that consider multiple parameters simultaneously could more accurately describe the overall dataset, and provide insights regarding the biology of the HIV reservoir and the host immune system. To accomplish this, we employed a decision tree approach to visualize combinations of variables that classify participants as having high (above median) or low (below median) reservoir frequency. Compared to the dimension reduction technique, decision tree visualization is an interpretable supervised approach and does not require post-hoc analysis.

We first selected 35 variables with the highest ROC AUC values for either total reservoir frequency, intact reservoir frequency, or percentage intact HIV DNA to be considered for model generation (**Figure 4**). Using these parameters, we fitted Generalized and Scalable Optimal Sparse Decision Trees (GOSDT) (52) to the data in order to classify high versus low total HIV DNA, intact HIV DNA, or percentage intact HIV DNA. We required the trees to achieve at least 80% accuracy for classifying PWH in the cohort, as well as have at least five PWH in each leaf. Since these trees are based on the entire dataset, these models are thus descriptive rather than predictive.

For the total HIV DNA frequency decision tree, only four immune variables were required to accurately describe high versus low total HIV DNA status (**Figure 5A**, **Figure S10**): CD8 T cell frequency, CD4 nadir, %CD38+HLA-DR-CD8 T cells, and %NKG2A+ CD4 T cells. The tree divided the

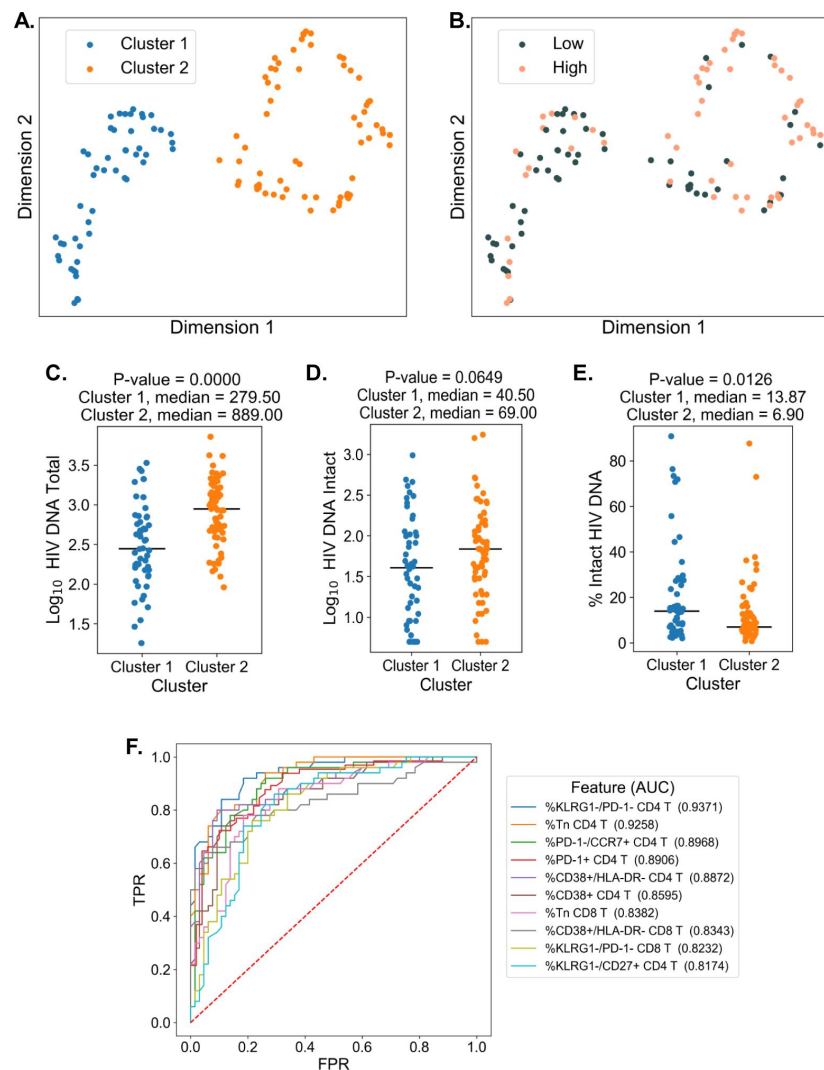


Figure 4

Dimension reduction reveals two major clusters of PWH with distinct immune systems and reservoirs.

A. PaCMAP was applied to the data using the ten immune cell features with the highest AUC values for classifying participants based on total reservoir frequency, and two clusters (cluster 1 and cluster 2) are identified. **B.** Same as A but data points are color-coded by total reservoir frequency (high=pink, low=gray). Total reservoir frequency (**C.**), intact reservoir frequency (**D.**), and percent intact (**E.**) is shown for participants within each cluster. **F.** Key immune cell features that distinguish cluster 1 from cluster 2 are identified by visualizing the features with the highest AUC values with respect to classifying cohort participants based on cluster membership. Axes represent the true positive rate (TPR) and the false positive rate (FPR) for each variable. Immune markers and clinical-demographic features are shown for each cluster in [Figure S8](#).

cohort into five subgroups (leaves), among which three have high total reservoir frequency and two have low total reservoir frequency. Comparing the labels provided by GOSDT model with the actual data, the tree achieved 83.5% accuracy (i.e. misclassifying 19 PWH among the overall cohort of 115 PWH). Notably, when we combined all samples from “high total reservoir” leaves and all samples from “low total reservoir” leaves, we observed a significant difference in the actual median total reservoir frequencies for these two groups (266/M for low total and 1288.5/M for high total, Mann Whitney U test p-value is 3.56×10^{-13} , **Figure 5B** [↗](#)).

We then repeated this approach for classifying the cohort participants with respect to intact reservoir frequency (**Figure 5C** [↗](#), **Figure S11** [↗](#)). Despite the overall lower correlation between the immune cell phenotypes and the intact reservoir size, this tree nevertheless achieved 82.6% accuracy. The tree had six leaves (three with low intact reservoir, three with high intact reservoir) and relied on %Tn CD8 T cells, %CD107a-IFN γ +IL-2+TNF α -CD4 T cells, %CD127+ CD4 T cells, %CD38-HLA-DR+ Tcm CD8 T cells, and CD4 Nadir. When we combined the ‘high intact’ leaves together and the ‘low intact’ leaves together, we observed a significant difference in median intact reservoir frequency between the groups (29.5/M for low intact and 101/M for high intact, Mann Whitney U test p-value is 9.59×10^{-8} , **Figure 5D** [↗](#)).

Finally, a third iteration of GOSDT generation was performed, classifying the cohort participants with respect to percentage intact of total proviruses (**Figure 5E** [↗](#), **Figure S12** [↗](#)). This tree achieved 82.4% accuracy and relied on years of ART treatment, %KLRG1+CD27-CD4 T cells, %CD127+ CD4 T cells, and recent clinical CD4 count. This tree has two leaves with low percent intact reservoir and three leaves with high percent intact reservoir. We observed a significant difference in the actual percent intact values between the model generated high percent intact and low percent intact leaves (median 5.7% for low percent intact leaves and median 15.3% for high percent intact leaves: Mann Whitney U test, p-value is 5.5×10^{-10} , **Figure 5F** [↗](#)).

Collectively, these decision trees highlight combinations of immune parameters that can accurately describe qualitatively high versus low total, intact, and percentage intact HIV DNA in a cohort of $n = 115$ PWH. This visualization serves as a basis for mechanistic hypotheses about the interactions of the immune system and HIV reservoir during long-term ART.

Machine learning algorithms identify immune feature combinations that predict high versus low reservoir metrics with ~70% test accuracy

Although clustering and decision tree analysis permit visualization and understanding of global structures within a dataset, we were curious if combinations of immune and clinical-demographic parameters could actually accurately predict, rather than only visualize, the size of the HIV reservoir. We considered five machine learning algorithms including Logistic Regression with L2 regularization (LR), CART, Support Vector Machines with RBF kernel (SVM), Random Forest (RF), and Gradient Boosted Trees (GBT). Initially, we attempted to predict reservoir frequency as a continuous variable, but found that models trained for this purpose performed poorly and tended to overfit to training sets (**Figure S13** [↗](#)). Thus, we focused our efforts on predicting high (above median) versus low (below median) reservoir frequency. For reservoir characteristics, we measured accuracy of the models over 10 random splits of our data into training and test sets (**Figure 6A** [↗](#) and **Table S10** [↗](#)). Overall, we found that for total reservoir frequency, Logistic Regression achieved highest mean classification accuracy (69.31%) in test data. The accuracy of these models is likely limited by the sample size ($n=115$) and noise in the data, as evidenced by lower R^2 score of machine learning models for non-binarized HIV characteristics (**Figure S13** [↗](#)).

To examine the contribution of individual immune features to model performance, we examined log-istic regression coefficients for each immune cell variable in the model. Since Logistic Regression coefficients are associated with expected change in log odds (based on $\log_e$), we can

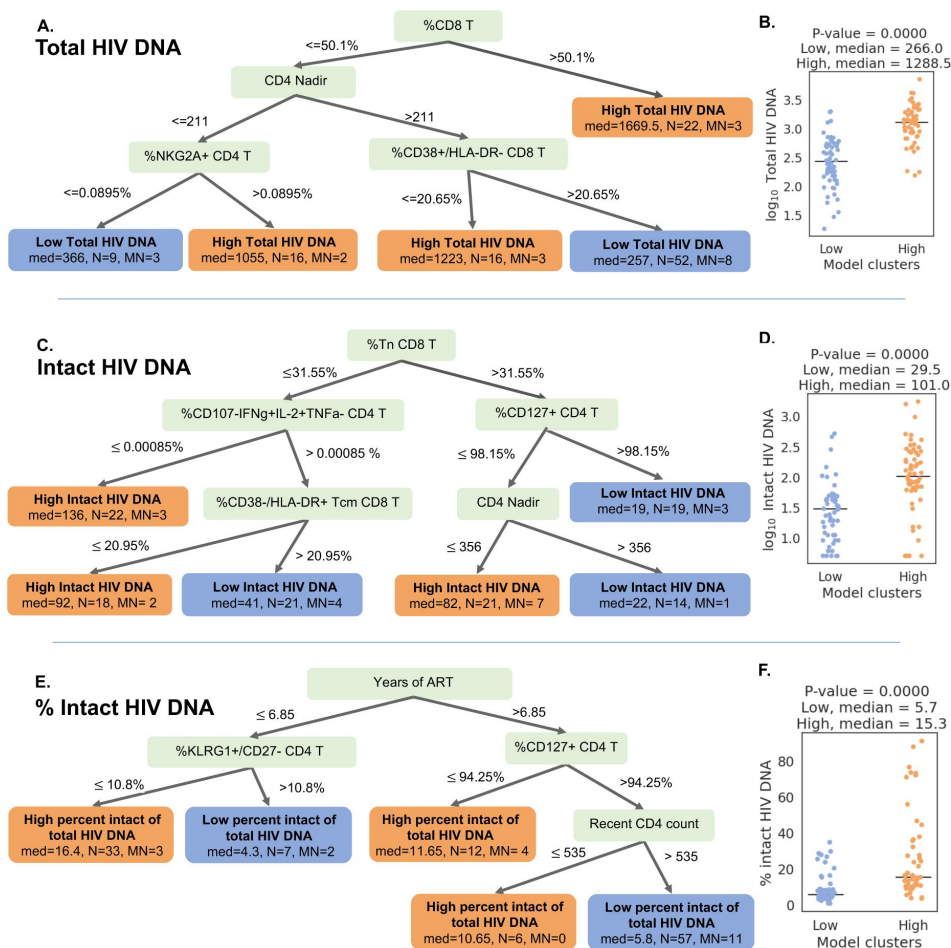


Figure 5

Decision tree visualization of the association of immune cell subsets with reservoir characteristics.

(A, C, E) Host variables (immune cell frequencies, demographic, and clinical information) were used to visualize the PWH dataset using the optimal sparse decision trees algorithm GOSDT. The overall set of PWH was classified as likely having high (above median, orange “leaves”) or low (below median, blue “leaves”) total reservoir frequency (A), intact reservoir frequency (C), and percent intact (E). In each leaf, “med” denotes the median HIV characteristic of PWH, N is the number of PWH in the leaf, and MN is the number of mislabeled PWH. (B, D, F) PWH in model leaves associated with high (orange) or low (blue) reservoir frequency characteristics were aggregated and a Mann-Whitney U test was performed to determine statistical significance between the actual total reservoir frequency of the “high” and “low” groups for total reservoir frequency (B), intact reservoir frequency (D), and percent intact (F). For the percent intact tree we exclude participants with missing values of years of ART. For total and intact reservoir frequency, missing values of years of ART were imputed, however, since the trees do not use this variable, imputations do not influence results. Visualization trees are explained with sets of rules in Supplemental Materials: [Figure S10](#) for total reservoir frequency, [Figure S11](#) for intact reservoir frequency, and [Figure S12](#) for percent intact.

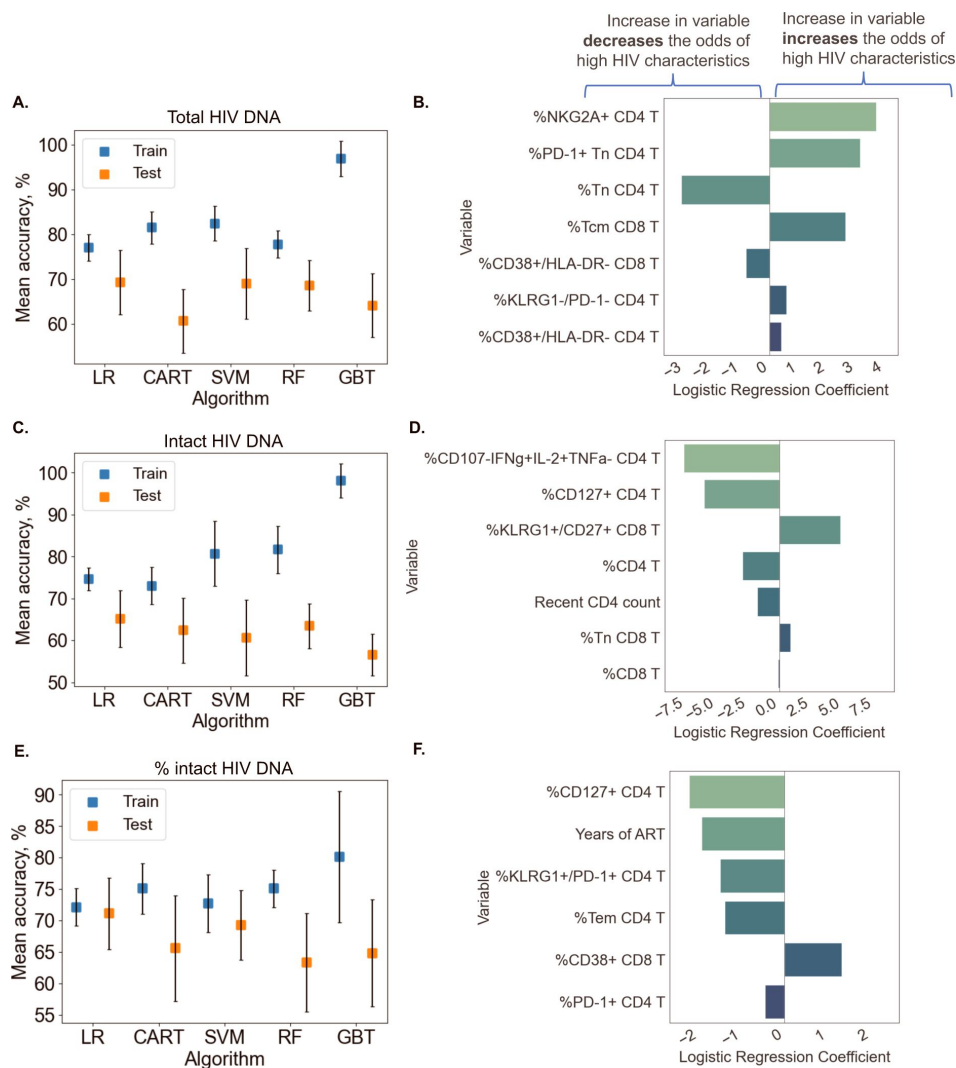


Figure 6

Predicting HIV reservoir characteristics with machine learning.

Average training and test accuracies over ten training and test data splits for Random Forest (RF), Gradient Boosted Trees (GBT), Support Vector Machines with RBF kernel (SVM), Logistic Regression (LR), and CART models for total reservoir frequency (A), intact reservoir frequency (C), and percent intact (E) are shown. For one split of training and test sets, logistic regression models are visualized for total reservoir (B), intact reservoir (D), and percent intact (F). On the y-axis we show variables used by the model, while the x-axis displays coefficient values for individual variables used by models. For percent intact models we exclude participants with missing values of years of ART. For total and intact reservoir frequency, missing values of years of ART were imputed. The missing value of the CD4 nadir for one participant was imputed using the MICE algorithm.

think about the coefficient β for variable X in the following way: increasing variable X by one unit multiplies the odds of high reservoir frequency (probability that the reservoir size is high divided by the probability that the reservoir frequency is low) by e^β . In **Figure 6B** [we visualize the logistic regression model for one data split among ten we considered for total reservoir frequency](#). For this split we observe that higher values of %NKG2A+ CD4 T, %PD-1+ Tn CD4 T, and %Tcm CD8 T are associated with a increased probability of total reservoir frequency being high. On the other hand, an increase in %Tn CD4 T decreases the odds of high reservoir frequency. The model visualized in **Figure 6B** [achieved 75.86% training and 75% test accuracy](#).

Similarly, logistic regression models performed best for predicting intact reservoir frequency compared to other methods (average 65.17% test accuracy, see **Figure 6C** [we visualize the Logistic Regression model for one fixed data split](#)). In **Figure 6D** [we observed that higher values of %CD107a-IFN \$\gamma\$ +IL-2+TNF \$\alpha\$ -CD4 T and %CD127+ CD4 T, and % CD4 T are associated with lower probability of intact reservoir frequency being high, while higher values of %KLRG1+CD27+ CD8 T are associated with increased probability of intact reservoir frequency being high](#).

For analysis of percentage intact HIV DNA, we display training and test accuracy values in **Figure 6E** [For the visualized model for one data split \(Figure 6F\)](#), an increase in %CD127+ CD4 T, years of ART, %KLRG1+PD-1+ CD4 T, and %Tem CD4 T leads to a lower probability of percent intact HIV DNA being high.

Overall, these analyses demonstrate that we can use machine learning tools to construct models that can predict with approximately 70 percent accuracy whether a given PWH has qualitatively low or high total HIV DNA frequency, intact HIV DNA frequency or percentage of HIV DNA. Further studies with larger cohorts will likely improve the accuracy of these models.

Discussion

In this cross-sectional study of peripheral blood samples from a well-characterized clinical cohort of 115 PWH on long-term ART, we used machine learning analysis of IPDA and high parameter flow cytometry to better understand the associations between the HIV reservoir and a broad range of immunophenotypes. Using these approaches, specific HIV reservoir features (total HIV DNA frequency, intact HIV DNA frequency, percentage intact HIV DNA) and clinical-demographic variables (age, sex, CD4 nadir, recent CD4 count, years on ART, years of HIV infection prior to initiating ART) were identified that predict specific immunophenotypes in PWH on long-term ART. Distinct host immune correlates of intact versus total HIV DNA frequency were also identified. Further, combinations of immune parameters that classified and predicted whether a given PWH has high or low total, intact, or percentage intact HIV DNA were defined. Collectively, this study confirms existing knowledge about HIV reservoir biology and identifies novel associations for further investigation. Additionally, from a methodological perspective, this work also demonstrates the utility of specific data science and machine learning approaches for HIV reservoir studies. Development and application of these analytic approaches are important to enable biological insights from high-parameter studies of the HIV reservoir and other complex biology, particularly with the advent of high-dimensional techniques such as CYTOF and scRNAseq.

Previous studies have sought to identify immune and other correlates of the total HIV DNA reservoir ([21](#), [26](#)–[44](#)). These studies have identified several activation/exhaustion markers that correlate with total HIV DNA, including the CD4/CD8 ratio, T cell expression of HLA-DR, CD38, PD-1, CTLA-4, and LAG-3, and NK cell KLRG1 expression. However, not all studies have found associations of total HIV DNA and immune subsets, particularly for CD4 T cell HLA-DR and CD38 expression ([39](#), [53](#)–[55](#)) during longterm ART. There is also growing appreciation for

additional factors that correlate with the HIV reservoir. Specifically, the size of the reservoir has been linked to the abundance of specific bacteria species in the intestinal microbiome (56). Also, genetic studies have indicated the existence of single nucleotide polymorphisms (SNPs) that are associated with reservoir size, including SNPs in MX1 and IRF7 (57–59). The correlative results from this present study corroborate many of these studies, and provide additional insights. A number of correlations between total HIV DNA and immune activation and exhaustion were identified. There were weak to moderate positive correlations observed between total HIV DNA frequency and CD4 T cell expression of NKG2A, PD-1, KLRG1 and HLA-DR as well as CD8 T cell expression of PD-1 and CCR7. Total HIV DNA frequency was inversely correlated with CD8 T cell expression of CD38, CD4 T cell expression of CD27, the frequency of naïve CD4 and CD8 T cells, and the CD4/CD8 ratio. Overall our results support the hypothesis that larger HIV reservoir size is associated with increased frequencies of cells with an activated or exhausted phenotype, with depletion of naïve cells and with an increased proportion of central memory CD8 T cells. The correlation of NKG2A expression in CD4 T cells with the total reservoir size represents a novel observation that merits further investigation. Only a very small percentage of CD4 T cells express NKG2A (less at 1 percent), and the biological significance for the correlation of these cells with the HIV reservoir is unknown. NKG2A is known to function as an inhibitory receptor for NK cells and for CD8 T cells, but its role in CD4 T cells is less clear (60).

When we examined immune correlates of intact HIV DNA measured by the IPDA, many of the same relationships as described above were noted, but with decreased magnitude and often lacking statistical significance. This is consistent with a recent study using the IPDA that did not find associations of intact HIV DNA with CD4 T cell HLA-DR and CD38 co-expression (61), as well as a study using near-full-length viral sequencing in single cells that found only slight enrichment of HLA-DR or PD-1 in cells with genetically intact, inducible proviruses (36). The stronger relationship between the immune cell phenotypes and the total reservoir, rather than the intact reservoir, should be investigated further. We speculate that this association could be driven by the more numerous defective proviruses promoting inflammation or immune activation through innate immune sensing pathways, or a greater level of transcriptional activity for defective proviruses. This is certainly plausible giving the growing appreciation for the immunogenicity and high frequency of defective proviruses within the HIV reservoir (62–65). It is important to note, however, that it is also possible that the stronger association of total proviral burden with immunophenotype reflects persistent damage caused to the immune system before therapy, and that this association between the total reservoir and immune cells may not result from ongoing interactions between HIV and the immune system during therapy.

Uniquely, CD127 expression on CD4 T cells was significantly associated with the intact reservoir frequency but not the total reservoir. Specifically, the frequency of CD127-CD4 T cells was positively correlated with the frequency of intact proviruses. CD127 is the alpha-chain IL-7 receptor and is expressed on long-lived memory T cells that are depleted during untreated HIV infection. Post-ART, the frequency of CD127+ memory CD4 subsets in PWH increases over time, albeit slowly (11). Interestingly CD127+ cells are enriched in latent HIV infection in tissues (66), and IL-7 is associated with slower natural reservoir decay (27). Furthermore, IL-7 signaling is known to drive expression of VLA-4, which was recently reported to highly enrich for genetically intact, inducible proviruses (36). This observational finding that intact but not total HIV DNA inversely correlates with CD127 expression on CD4 T cells requires further investigation.

Dimension reduction machine learning approaches identified two robust clusters of PWH when using total HIV DNA reservoir-associated immune cell frequencies. These clusters exhibited distinct reservoir characteristics, with one cluster being enriched with participants with a larger total reservoir size, a lower percentage of intact viruses, a lower frequency of naïve T cells and elevated expression of activation and exhaustion markers, while the other contained PWH with smaller total reservoirs, higher percentage of intact viruses, higher frequency of naïve T cells and lower expression of activation and exhaustion markers. The cluster membership was notably

associated with age and time on therapy. The existence of two distinct clusters of PWH with different immune features and reservoir characteristics could have implications for HIV cure strategies. These two clusters are also interesting in the context of a previous study examining reservoir characteristics which proposed two major types of viral reservoir within PWH: one type with smaller reservoirs that were enriched in Tcm cells and associated with lower levels of Ki67+ cells and immune activation, and another type with larger reservoirs that were enriched in Ttm cells and associated with higher levels of Ki67+ cells and immune activation (27²⁸). It is possible that these two proposed types of reservoirs represent the two clusters we observe using immune markers and dimension reduction, although we were unable to examine the frequency of HIV proviruses specifically within Tcm and Ttm cells in this study.

In addition to unsupervised dimension reduction approaches, we were able to construct simple, in-terpretable decision trees that describe our cohort of PWH with more than 80% accuracy with respect to the size of the reservoir. Furthermore, when we constructed decision trees to classify PWH separately based on total reservoir frequency, intact reservoir frequency and percentage intact, distinct sets of features were important (though some were the same) for each of these models. The decision tree analyses highlight complex immune parameter combinations that serve as a basis for unique hypotheses about HIV reservoir biology. One striking example of machine learning approaches identifying potential reservoir biology is the percentage intact HIV DNA decision tree. As expected, years of ART was critical to predict the percentage of intact proviruses, but CD4 T cell CD127 expression, CD4 T cell KLRG1 expression and recent CD4 count also played a significant role.

Logistic regression models that we trained to classify high versus low total reservoir, intact reservoir, and percentage intact were able to predict out-of-sample with ~70% accuracy on average. Each of these models relied on five features only and were simple enough for us to visualize and assess how important each variable is for every characteristic of reservoir size. These models corroborated the potential biological relationships identified in the decision tree analysis, and confirmed that these variables could actually be used to predict out-of-sample whether total, intact or percentage intact HIV DNA was qualitatively high or low. The success of our predictive models was likely limited by the relative small size of the datasets used to train and test them (n=115). As such, larger datasets will be required to generate more accurate models that can be used for reliable prediction of reservoir characteristics. Additionally, it is likely that unmeasured factors and biological noise also contribute to reservoir size.

Our findings should be considered in the light of some inherent limitations and caveats. This study is cross-sectional in nature and is primarily observational, so the findings should be interpreted with caution. As with any observational study, confounding bias may influence our results, particularly for correlative analyses. The associations between the HIV reservoir and the immune system we observed were fairly weak, despite being statistically significant. Similarly, the predictive ability of our ML models was modest (70 percent) for classifying PWH as having an above or below median reservoir. As mentioned above, our cohort is likely too small to generate accurate predictive ML models, and consists of mostly male participants, which could bias the results. We also cannot rule out the influence of additional unmeasured confounding factors. In particular, the length of untreated infection prior to ART could be a major contributor to the immune signatures we observe. In this study, the participants had all been treating during chronic infection, and different associations might be observed in PWH who are treated during acute infection. Another limitation to this study is that our analyses only considered cells and proviruses in the peripheral blood, and this may not represent HIV reservoirs and immune cells located in tissues known to harbor latent HIV such as gut, spleen and brain. Our analysis of surface protein phenotypes also does not detect potentially important associations between the reservoir and transcriptional pathways within these cell types. Future work using single cell RNAseq from a similar cohort may help to reveal deeper layers of associations that are invisible to the current analysis. As is always the case with correlation studies, it is important to be aware that these

associations are only correlations and do not necessarily represent functional or mechanistic relationships. Even so, these correlations are still useful as a source of novel hypotheses that could be tested in subsequent studies.

Nevertheless, our study does have some strengths compared to previous work in this area, specifically a highly detailed flow panel, allowing us to quantify numerous immune subsets with a high degree of resolution. Also, the use of the intact proviral DNA assay allows us to separately examine the association of the immune system with intact and total viral DNA. It is however, important to note that IPDA does not measure all proviral sequences in PWH and is likely failing to identify some sequences due to proviral polymorphisms or dual deletions in the two IPDA amplicons. Also, a minor fraction of IPDA+ proviruses are actually defective due to the limited specificity of using two short amplicons to select for full-length intact proviruses (46 [↗](#), 67 [↗](#)).

Overall, these findings suggest a complex concert of immune recovery, HIV reservoir dynamics, and intrinsic host factors (age, biological sex) that shapes host immunophenotype, even after years of ART. Mechanistic work will be needed to fully dissect the dynamic relationship between the immune system and the reservoir. Nevertheless, these findings support a model in which ongoing interaction between the HIV reservoir and the host immune cells continue to drive an association, albeit a relatively minor one, with persistent CD4 T cell activation during long-term ART. Notably, both intact and defective proviruses appear to contribute to this immune signature of HIV latency during long-term ART. This study identifies several areas for future investigation. The question of whether the HIV reservoir directly drives persistent immune activation could be addressed directly by specifically suppressing viral transcription during therapy and examining the impact on the immune system. The stronger association of the intact reservoir with CD127 expression in CD4 T cells is an intriguing observation that warrants further validation and explanation. This observation could, for example, be explained by a model in which the intact reservoir is preferentially located in cells with a distinct pattern of CD127 expression. These ideas, along with other future studies exploring the interplay of the host immune system and HIV reservoir might leverage new high-dimensional technologies such as scRNAseq and mass cytometry. The machine learning approaches described and applied in this study may be particularly useful for gaining biological insights into these high parameter datasets.

Methods

Cohort and sample collection

115 people with HIV (PWH) were recruited from two clinical sites. 66 PWH were recruited at Duke University medical center and 49 PWH at UNC Chapel Hill. Peripheral blood was collected and peripheral blood mononuclear cells (PBMCs) were obtained by Ficoll separation, then frozen in 90% fetal bovine serum 10% dimethylsulfoxide (DMSO) (51 [↗](#), 68 [↗](#)–70 [↗](#)).

Flow cytometry

Cell preparation and staining followed previously described methods (71 [↗](#)). All antibodies were titrated to optimize signal-to-noise ratio on PBMCs prior to use, assuming a 50µL staining volume. Cell viability was examined using Zombie-NIR Fixable Viability Dye (0.4µL per 50µL staining volume; Biolegend). Antibodies used for surface staining were as follows: KLRG1-BV421 (SA231A2 clone, Biolegend), CD45RA-PacBlue (H100 clone, BL), CD8-BV570 (RPA-T8 clone, Biolegend), CD127-BV605 (A019D5 clone, Biolegend), CD56-BV650 (5.1H11 clone, Biolegend), CCR7-BV711 (G043H7 clone, Biolegend), CD27-BV750 (O323 clone, Biolegend), PD1-VioBright515 (REA165 clone, Miltenyi), NKG2A-PE-Vio615 (REA110 clone, Miltenyi), CD16-PerCP-Cy5.5 (33G8 clone, Biolegend), CD38-PCPeF710HB7 clone, TF), CD14-SparkNIR685 (63D3 clone, Biolegend), CD19-SparkNIR685 (HIB19 clone, Biolegend), HLA-DR-APC-F750 (L243 clone, Biolegend). Antibodies used for intracellular staining were as follows: CD3-BV480 (UCHT1 clone, BD), CD4-PerCP (L200 clone, BD), IFN-g-PE-Cy7

(4S.B3 clone, Biolegend), IL-2-APC (MQ1-17H12 clone, Biolegend), TNF α -AF700 (Mab11 clone, Biolegend). For HIV peptide stimulation, the cells were incubated with peptide mixes from Gag, Pol and Env proteins at 0.2 μ g/mL in the presence of brefeldin A and monensin for 6 hrs. Samples were analyzed using a Cytex Aurora spectral flow cytometer. Flow cytometry analysis was performed in FlowJo v10.8 software.

Intact Proviral DNA Assay (IPDA)

Cryopreserved samples of PBMCs from each study participant were viably thawed. A portion was used for immunophenotyping as described above, and the remainder were subjected to total CD4 T cell negative selection with the StemCell Technologies EasySepTM Human CD4⁺ T Cell Enrichment Kit (Cat# 19052). CD4 T cell DNA was extracted using the QIAamp DNA Mini Kit and quantified on a NanoDrop 1000 (Thermo Fisher Scientific). IPDA was performed as originally described (46), with a validated PCR annealing temperature modification to increase signal to noise ratio (70). Gating for positive droplets was set using negative (DNA elution buffer and HIV-seronegative CD4 T-cell DNA), and positive (Integrated DNA Technologies gblock amplicon) positive control wells processed in parallel. DNA shearing index values were similar to those reported previously (median, 0.33; interquartile range [IQR], 0.31–0.34) (68–70). A median of 1.04×10^6 ($Q1 - 8.71 \times 10^5$, $Q3 - 1.16 \times 10^6$) cell equivalents were assessed for each donor.

Statistics

Variable data analysis

To create binarized labels that represent reservoir characteristics, we split the data at the median, which is 553/M for total reservoir frequency, 53/M for intact reservoir frequency, and 8.64 percent for percent intact. For every variable (immune cell frequency, demographics, and clinical information), we computed Spearman correlation and an Area Under the Curve (AUC) value. To compute the AUC value, we first created a Receiver Operating Characteristic (ROC) curve by plotting the true positive rate versus the true negative rate for every cell subset frequency. AUC is then computed as the area under the ROC curve using the trapezoidal rule.

To assess the importance of clinical, demographic information and HIV reservoir characteristics in describing the immune markers, we performed a leave-one-covariate-out (LOCO) analysis. We trained a linear regression model (M_1) with intercept based on all variables. We removed each variable one at a time and trained a new model without this variable. Then we computed the difference in R^2 score before and after dropping the variable. Variance Inflation Factors (VIFs) for age, sex, years of ART, CD4 nadir, recent CD4 account, and years of HIV infection prior to HIV infection were less than 2, indicating an acceptable level of correlation between these independent variables (72). However, higher VIF values were observed for intact and total HIV DNA, therefore separate LOCO inference analyses were conducted in order to avoid artificial fluctuations in model fit due to multicollinearity.

Data visualization

We used PaCMAP (50) to reduce the dimensionality of the dataset to a 2-dimensional space. Due to the relatively small sample size of our dataset ($n=115$), we set the number of neighbors to 5. Additionally, we used GOSDT (52) which computes sparse optimal trees to identify patterns in the dataset. For all the visualization trees we set the depth budget parameter to 4 and the regularization parameter to 0.02.

Training and generalization

Our training procedure for classification is described in [Algorithm 1](#) in Supplemental Materials, which is a standard machine learning pipeline with an added search for the number of variables. For the full hyperparameters list see [Table S8](#). We normalized continuous variables to have values between 0 and 1. For logistics regression models, we set the regularization parameter of ℓ_2 regularization based on cross-validation. The values are 10 for total reservoir frequency and percent intact and 100 for intact reservoir frequency.

Study approval

Written informed consent was obtained for all study participants. The study design was reviewed and approved by IRB for both Duke University and UNC Chapel Hill.

Data availability

Underlying data available at: <https://doi.org/10.15139/S3/3CFVIU>.

Code is available at: https://github.com/lesiasemenova/ML_HIV_reservoir.

Author contributions

Designed the study – EPB, DMM, CDR. Conducted experiments – SDF, ADVC. Analyzed the data – LS, YW, SDF. Wrote the manuscript: All authors.

Conflict of interest statement

The authors have declared that no conflict of interest exists.

Acknowledgements

This work was supported by the following grants from the National Institutes of Health: NIAID R01 AI143381 (EPB), NIAID UM1 AI164567 (DM Margolis), NIDA R61 DA047023 (EPB), NIDA R01 DA054994 (CDR), NIAID F30 AI145588 (SDF).

Supplementary Methods

CD4/CD8 ratio regression analysis

To examine the connection between CD4/CD8 and (CD127+ CD4)/CD8 ratio and HIV characteristics ([Figure S5](#)), we first computed the natural logarithm ($\log_e$) of the intact, total and percent intact as the outcome and normalized outcome of each immune feature (meaning that our values of ratio and outcome are between 0 and 1). We removed outliers using the DBSCAN algorithm ([1](#)), which is an unsupervised clustering algorithm that groups data points based on density into a single

cluster. For DBSCAN, we set the parameter that determines radius of a circle around each point (that is used to compute density) as 0.15 and minimum number of samples in each cluster as 10. Then we fitted the remaining datapoints using Linear Regression.

Regression machine learning models to predict reservoir frequency from immune cell frequencies

To predict reservoir frequency based on immune cell frequencies, clinical and demographic information (**Figure S13** [↗](#)), similar to classification, we sought to find the set of model parameters and types of models that would be able to fit the dataset without overfitting. We used several approaches: Linear Regression (LR), Ridge Regression (RR), Kernel Regression with RBF kernel (KR), Decision Tree Regressor (DT), Random Forest (RF), and Gradient Boosted Trees (GBT). We followed the procedure described in **Algorithm 1** [↗](#), optimized the R^2 score, returned the mean and standard deviation of the R^2 score of training and test sets, and chose features based on the absolute value of the correlation coefficient. We normalized variables to have values between 0 and 1. We then fitted the models to the normalized natural logarithm ($\log_e$) of total, intact reservoir frequency and percent intact and then further normalization makes outcomes to be between 0 and 1. For total and intact reservoir frequency we visualized the Ridge Regression model, which performed the best based on mean test R^2 score (**Figure S13 D, E, G, H** [↗](#)). We set the regularization parameter to 1, as this value was best on average according to cross-validation for total and intact reservoir frequency. For percent intact, we visualized Linear Regression based on 4 variables (**Figure S13 F, I** [↗](#)) as it achieved the second highest R^2 score after Kernel Ridge based on 25 variables.

Algorithm 1 Training procedure for classification (regression)

Algorithm 1 Training procedure for classification (regression)

Input: Outcome, algorithm, set of hyperparameters from **Table S10** [↗](#)

Output: Mean and standard deviation of train and test accuracies (R^2 scores)

```

1: Compute product of hyperparameters for grid search, let  $H$  be one hyperparameter setting
2: for  $k$  in the set of top AUC (correlation) variables to consider do
3:   Take the first  $k$  variables from the set of variables
4:   for test seed in the array of random seeds for train/test splits do
5:     Sort data according to test seed
6:     Split data into test and train in proportion 25% to 75%
7:     for every hyperparameters settings  $H$  do
8:       for validation seed in the array of random seeds for train/valid splits do
9:         Sort data according to validation seed
10:        Split data into validation and train for validation in proportion 25% to 75%
11:        Train algorithm with hyperparameter setting  $H$  on train for the validation set
12:        Compute accuracy ( $R^2$  score) on validation set
13:      end for
14:    Compute mean validation accuracy ( $R^2$  score) on all validation sets for hyperparameter
    setting  $H$ 
15:  end for
16:  Find hyperparameter setting ( $H_m$ ) with the highest mean validation accuracy
17:  Train algorithm with hyperparameter setting  $H_m$  on train data
18:  Compute train and test accuracy ( $R^2$  score)
19: end for
20: Compute accuracy ( $R^2$  score) mean and std over different train and test sorting
21: end for
22: Among mean validation accuracies ( $R^2$  scores) for different  $k$  find the maximum
23: Return train and test accuracies ( $R^2$  scores) mean and standard deviation that corresponds to the
    number of variables  $k$  with the highest mean validation accuracy ( $R^2$  score)
  
```


Representative flow cytometry gating is shown for one sample from the 115-person cohort.



We report medians and [Q1, Q3] for total reservoir frequency, intact reservoir frequency (per million CD4 T cells), and percent intact. Mann-Whitney U test was used to compute p-values for gender and Kruskal-Wallis H-test for race.

Group	Size	Intact HIV DNA	Total HIV DNA	% intact HIV DNA
Male	88	63.5 [18.75 146.25]	681 [198 1433.75]	8.73 [4.9 20.58]
Female	27	41 [19, 92.5]	489 [317.5, 637]	8.64 [4.86, 15.56]
p-value		0.4521	0.1937	0.9451
Black	69	43 [19, 113]	502 [181, 1246]	9.77 [4.64, 16.4]
White	43	69 [26 191.5]	605 [289.5 1415.5]	8 [5.31 20.81]
Other	3	67 [41, 520]	1055 [720, 1165,]	6.39 [5.20, 41.37]
p-value		0.5079	0.4916	0.9914

Immune marker	Med	[Q1, Q3]	Immune marker	Med	[Q1, Q3]
%CD4 T	52.90	[43.85, 59.60]	%CD27+ CD4 T	86.20	[79.15, 90.10]
%CD38+ CD4 T	46.60	[39.65, 54.30]	%CD127+ CD4 T	97.10	[95.20, 98.30]
%HLA-DR+ CD4 T	12.30	[9.76, 16.40]	%KLRG1+ CD4 T	17.20	[10.85, 26.25]
%NKG2A+ CD4 T	0.10	[0.06, 0.20]	%PD-1+ CD4 T	45.40	[39.70, 54.25]
%Tcm CD4 T	49.90	[42.30, 56.65]	%Tn CD4 T	30.40	[22.10, 39.90]
%Teff CD4 T	1.73	[0.89, 3.50]	%Tem CD4 T	15.70	[10.95, 19.70]
%CD38-/HLA-DR+ CD4 T	8.16	[6.22, 11.00]	%CD38+/HLA-DR+ CD4 T	4.37	[3.16, 5.94]
%CD38+/HLA-DR- CD4 T	40.20	[32.90, 47.75]	%KLRG1-/CD27+ CD4 T	74.40	[67.30, 83.05]
%KLRG1+/CD27+ CD4 T	6.70	[3.67, 11.65]	%KLRG1+/CD27- CD4 T	7.68	[4.54, 13.40]
%KLRG1-/CD27- CD4 T	6.01	[4.43, 8.04]	%KLRG1-/PD-1+ CD4 T	32.80	[27.00, 37.95]
%KLRG1+/PD-1+ CD4 T	12.60	[7.46, 17.30]	%KLRG1+/PD-1- CD4 T	4.79	[3.12, 8.63]
%KLRG1-/PD-1- CD4 T	49.30	[38.35, 55.90]	%PD-1-/CCR7+ CD4 T	50.90	[40.50, 56.35]
%PD-1+/CCR7+ CD4 T	29.70	[25.00, 35.05]	%PD-1+/CCR7- CD4 T	14.60	[10.40, 18.90]
%PD-1-/CCR7- CD4 T	3.77	[2.56, 5.25]	%CD8 T	40.30	[33.35, 47.45]
%CD27+ CD8 T	65.40	[54.05, 75.30]	%CD38+ CD8 T	35.70	[26.05, 45.45]
%CD127+ CD8 T	99.20	[98.40, 99.60]	%HLA-DR+ CD8 T	25.70	[18.75, 36.05]
%KLRG1+ CD8 T	62.10	[47.40, 73.40]	%PD-1+ CD8 T	37.50	[26.10, 47.50]
%Tn CD8 T	30.40	[21.60, 41.50]	%Teff CD8 T	26.90	[20.35, 36.70]
%Tem CD8 T	19.00	[13.00, 27.15]	%CD38-/HLA-DR+ CD8 T	16.10	[10.55, 22.75]
%CD38+/HLA-DR+ CD8 T	8.43	[5.86, 13.45]	%CD38+/HLA-DR- CD8 T	19.60	[14.45, 31.15]
%KLRG1-/CD27+ CD8 T	35.00	[23.45, 48.80]	%KLRG1+/CD27+ CD8 T	25.90	[19.35, 33.45]
%KLRG1+/CD27- CD8 T	30.30	[21.45, 43.40]	%KLRG1-/CD27- CD8 T	2.42	[1.58, 4.57]
%KLRG1-/PD-1+ CD8 T	7.40	[5.17, 10.95]	%KLRG1+/PD-1+ CD8 T	27.80	[18.35, 36.75]
%KLRG1+/PD-1- CD8 T	30.20	[23.35, 38.25]	%Tcm CD8 T	14.60	[10.30, 22.00]
%KLRG1-/PD-1- CD8 T	27.70	[17.75, 43.70]	%PD-1-/CCR7+ CD8 T	30.70	[22.45, 42.95]
%PD-1+/CCR7+ CD8 T	13.30	[10.10, 16.70]	%PD-1+/CCR7- CD8 T	22.00	[14.85, 27.50]
%PD-1-/CCR7- CD8 T	28.30	[18.20, 36.40]	%NK	6.37	[3.98, 10.50]
%NKG2A+ NK	31.60	[21.40, 49.85]	%NKG2A-/CD16+ NK	52.20	[34.25, 64.90]
%NKG2A+/CD16+ NK	16.50	[10.21, 24.35]	%NKG2A+/CD16- NK	13.10	[8.26, 23.75]
%NKG2A-/CD16- NK	13.80	[9.26, 19.15]	%NKT	4.33	[2.33, 6.56]
%NKG2A+ NKT	9.31	[4.79, 19.70]	%NKG2A-/CD16+ NKT	6.31	[1.66, 12.95]
%NKG2A+/CD16+ NKT	0.43	[0.14, 1.23]	%NKG2A+/CD16- NKT	8.48	[4.57, 17.65]
%NKG2A-/CD16- NKT	77.40	[65.85, 89.35]	%CD38-/HLA-DR+ Tcm CD4 T	9.93	[7.52, 12.25]
%CD38+/HLA-DR+ Tcm CD4 T	4.71	[3.13, 6.30]	%CD38+/HLA-DR- Tcm CD4 T	22.90	[17.00, 28.45]
%CD38-/HLA-DR- Tcm CD4 T	61.80	[55.20, 66.75]	%PD-1+ Tcm CD4 T	58.90	[52.55, 65.80]
%CD38-/HLA-DR+ Tn CD4 T	0.66	[0.30, 1.12]	%CD38+/HLA-DR+ Tn CD4 T	7.29	[5.21, 9.38]
%CD38+/HLA-DR- Tn CD4 T	81.30	[76.20, 87.55]	%CD38-/HLA-DR- Tn CD4 T	8.52	[5.80, 14.20]
%PD-1+ Tn CD4 T	7.26	[4.40, 12.65]	%CD38-/HLA-DR+ Teff CD4 T	17.50	[11.20, 25.75]
%CD38+/HLA-DR+ Teff CD4 T	19.00	[10.75, 28.15]	%CD38+/HLA-DR- Teff CD4 T	16.00	[11.25, 27.15]
%CD38-/HLA-DR- Teff CD4 T	38.50	[27.40, 50.50]	%PD-1+ Teff CD4 T	56.40	[39.30, 72.25]
%CD38-/HLA-DR+ Tem CD4 T	27.40	[22.45, 33.80]	%CD38+/HLA-DR+ Tem CD4 T	7.31	[5.31, 10.10]
%CD38+/HLA-DR- Tem CD4 T	4.77	[3.49, 6.96]	%CD38-/HLA-DR- Tem CD4 T	57.00	[50.75, 65.80]
%PD-1+ Tem CD4 T	82.20	[76.00, 86.55]	%CD38-/HLA-DR+ Tcm CD8 T	22.10	[16.70, 28.20]
%CD38+/HLA-DR+ Tcm CD8 T	7.11	[4.23, 12.10]	%CD38+/HLA-DR- Tcm CD8 T	5.28	[3.71, 7.18]
%CD38-/HLA-DR- Tcm CD8 T	62.70	[51.65, 72.10]	%PD-1+ Tcm CD8 T	63.60	[53.80, 71.10]
%CD38-/HLA-DR+ Tn CD8 T	6.02	[3.29, 9.82]	%CD38+/HLA-DR+ Tn CD8 T	3.68	[2.42, 5.36]
%CD38+/HLA-DR- Tn CD8 T	45.40	[31.70, 57.25]	%CD38-/HLA-DR- Tn CD8 T	43.00	[31.75, 54.45]
%PD-1+ Tn CD8 T	15.00	[9.96, 24.35]	%CD38-/HLA-DR+ Teff CD8 T	27.80	[21.55, 37.25]
%CD38+/HLA-DR+ Teff CD8 T	16.60	[11.35, 25.95]	%CD38+/HLA-DR- Teff CD8 T	12.40	[7.18, 18.45]
%CD38-/HLA-DR- Teff CD8 T	39.20	[26.30, 47.20]	%PD-1+ Teff CD8 T	33.90	[24.10, 41.85]
%CD38-/HLA-DR+ Tem CD8 T	29.70	[21.55, 36.05]	%CD38+/HLA-DR+ Tem CD8 T	13.80	[9.31, 21.10]
%CD38+/HLA-DR- Tem CD8 T	5.31	[3.27, 9.88]	%CD38-/HLA-DR- Tem CD8 T	46.30	[35.30, 56.80]
%PD-1+ Tem CD8 T	62.40	[49.70, 71.80]	%IL-2+ CD8 T	0.00	[-0.03, 0.03]
%IL-2+ CD4 T	0.02	[0.01, 0.04]	%TNF α + CD4 T	0.04	[0.01, 0.06]
%TNF α + CD8 T	0.04	[0.01, 0.12]	%IFN γ + CD4 T	0.02	[0.00, 0.04]
%IFN γ + CD8 T	0.05	[0.01, 0.21]	%CD107+ CD4 T	0.02	[0.01, 0.04]
%CD107+ CD8 T	0.23	[0.12, 0.55]	%CD107+IFN γ +IL-2-TNF α - CD4 T	0.00	[0.00, 0.00]
%CD107+IFN γ -IL-2-TNF- CD4 T	0.00	[0.00, 0.00]	%CD107+IFN γ -IL-2-TNF α + CD4 T	0.00	[0.00, 0.00]
%CD107+IFN γ -IL-2-TNF α - CD4 T	0.01	[0.01, 0.02]	%CD107-IFN γ +IL-2-TNF α - CD4 T	0.00	[0.00, 0.00]
%CD107-IFN γ +IL-2-TNF α + CD4 T	0.00	[-0.01, 0.01]	%CD107-IFN γ +IL-2-TNF α - CD4 T	0.00	[-0.01, 0.00]
%CD107-IFN γ -IL-2-TNF α + CD4 T	0.01	[0.01, 0.02]	%CD107-IFN γ -IL-2-TNF α - CD4 T	0.00	[-0.01, 0.00]
%CD107-IFN γ -IL-2-TNF α - CD4 T	0.01	[0.00, 0.02]	%CD107+IFN γ +IL-2-TNF α - CD8T	0.03	[0.01, 0.07]
%CD107+IFN γ -IL-2-TNF- CD8 T	0.00	[0.00, 0.00]	%CD107+IFN γ -IL-2-TNF α + CD8 T	0.01	[0.00, 0.03]
%CD107+IFN γ -IL-2-TNF α - CD8 T	0.14	[0.08, 0.30]	%CD107-IFN γ +IL-2-TNF α - CD8 T	0.00	[0.00, 0.00]
%CD107-IFN γ +IL-2-TNF α + CD8 T	0.00	[0.00, 0.01]	%CD107-IFN γ +IL-2-TNF α - CD8 T	0.00	[-0.01, 0.01]
%CD107-IFN γ -IL-2-TNF α + CD8 T	0.00	[0.00, 0.00]	%CD107-IFN γ -IL-2-TNF α - CD8 T	-0.01	[-0.04, 0.00]
%CD107-IFN γ -IL-2-TNF α - CD8 T	0.00	[0.00, 0.00]			

Table S2

Immune subsets characteristics

. We report medians, 25% and 75% percentiles for every immune variable used in the study.

Variable	Total HIV DNA	Intact HIV DNA	% intact HIV DNA	Variable	Total HIV DNA	Intact HIV DNA	% intact HIV DNA
%CD8 T	0.4052	0.3562	0.0068	%CD38+/HLA-DR- CD8 T Tcm	-0.2169	-0.0182	0.1970
%CD38+/HLA-DR- CD4 T	-0.3891	-0.1098	0.2664	%CD107-IFN γ +IL-2-TNF α - CD4 T	-0.2154	-0.1633	-0.0599
%KLRG1-/PD-1- CD4 T	-0.3808	-0.1289	0.2334	%KLRG1-/CD27+ CD8 T	-0.2112	-0.1894	-0.0021
%Tn CD4 T	-0.3802	-0.2031	0.1714	%KLRG1+/PD-1+ CD8 T	0.2089	0.1922	0.0473
%NKG2A+ CD4 T	0.3618	0.2904	0.0179	%CD38+/HLA-DR- CD4 T Tn	-0.2066	0.0140	0.1909
%PD-1-/CCR7+ CD4 T	-0.3590	-0.1082	0.2283	%PD-1+/CCR7- CD4 T	0.1971	0.0410	-0.1338
%CD4 T	-0.3564	-0.3195	-0.0079	%CD38-/HLA-DR+ CD8 T	0.1947	0.0572	-0.0860
%Tcm CD8 T	0.3466	0.1752	-0.1814	%KLRG1+ CD8 T	0.1887	0.1709	0.0048
%CD38+ CD4 T	-0.3366	-0.0611	0.2829	%KLRG1-/PD-1+ CD4 T	0.1877	0.0818	-0.0963
%PD-1-/CCR7- CD4 T	0.3300	0.1824	-0.0938	%NKG2A-/CD16+ NK	0.1855	0.0433	-0.1222
%CD38+/HLA-DR- CD8 T	-0.3267	-0.0837	0.2636	%KLRG1+/PD-1- CD4 T	0.1849	0.0616	-0.1388
%PD-1+ CD4 T	0.3222	0.0664	-0.2470	%CD38+/HLA-DR+ CD4 T Tem	0.1798	0.2668	0.1961
Age	0.3172	0.1669	-0.1471	%KLRG1+ CD4 T	0.1771	0.0002	-0.2034
CD4 Nadir	-0.3164	-0.1512	0.1927	%KLRG1+/PD-1+ CD4 T	0.1650	-0.0392	-0.2355
%PD-1+ Tn CD4 T	0.3119	0.1246	-0.0962	Years before ART < 1	-0.1645	0.0527	0.2513
%Tn CD8 T	-0.3028	-0.2697	0.0154	% NKT	0.1616	-0.0223	-0.1942
Years of ART	0.3062	-0.0688	-0.4523	%CD38-/HLA-DR+ CD4 T Tn	0.1612	-0.0675	-0.2160
%PD-1+/CCR7+ CD8 T	0.2926	0.1361	-0.1254	%PD-1+/CCR7- CD8 T	0.1352	0.1899	0.1057
%CD38-/HLA-DR+ CD4 T	0.2849	0.0920	-0.1500	%CD38-/HLA-DR- CD8 T	0.1349	-0.0452	-0.2308
%KLRG1-/PD-1- CD8 T	-0.2757	-0.2182	0.0162	%CD38-/HLA-DR- CD8 T Tcm	0.1302	-0.0048	-0.2000
%PD-1+ Tn CD8 T	0.2738	0.1411	-0.0705	Recent CD4 count	-0.1302	-0.1944	-0.0970
%CD38+/HLA-DR- CD8 T Tn	-0.2676	-0.0610	0.2119	%CD127+ CD4 T	-0.1298	-0.3160	-0.2539
%PD-1+/CCR7+ CD4 T	0.2665	0.0859	-0.1673	Years before ART = NA	0.1183	-0.0312	-0.1950
%KLRG1+/CD27+ CD8 T	0.2606	0.2212	0.0319	%CD127+ CD8 T	0.1136	-0.1217	-0.2458
%CD27+ CD4 T	-0.2602	-0.0531	0.1767	%CD38+/HLA-DR+ CD8 T Tcm	-0.0920	0.0701	0.2340
%KLRG1-/CD27+ CD4 T	-0.2575	-0.0464	0.2171	%CD107-IFN γ -IL-2+TNF α + CD8 T	0.0909	-0.0499	-0.2270
%HLA-DR+ CD4 T	0.2565	0.1110	-0.0970	%CD38+/HLA-DR- CD8 T Teff	-0.0833	0.1328	0.2462
%PD-1+ CD8 T	0.2541	0.1969	0.0035	%CD38+/HLA-DR- CD8 T Tem	-0.0660	0.1177	0.1958
%CD38+ CD8 T	-0.2418	0.0109	0.3114	%CD107-IFN γ -IL-2+TNF α + CD4 T	0.0568	-0.1775	-0.3223
%CD38-/HLA-DR+ CD8 T Tn	0.2404	0.0738	-0.1346	%IL-2+ CD4 T	-0.0515	-0.1936	-0.1868
%Tem CD4 T	0.2402	0.0888	-0.1669	%CD38+/HLA-DR+ CD8 T Tem	0.0510	0.1694	0.2107
%PD-1-/CCR7+ CD8 T	-0.2321	-0.2173	-0.0237	%CD38+/HLA-DR+ CD8 T	0.0461	0.1660	0.1865
%CD38-/HLA-DR- CD4 T Tem	-0.2302	-0.2365	-0.1349	%CD107-IFN γ +IL-2+TNF α - CD4 T	0.0285	-0.2455	-0.3265
%KLRG1+/CD27- CD4 T	0.2271	0.0207	-0.1935	%CD38+/HLA-DR- CD4 T Tem	-0.0134	0.1551	0.2013
%Tcm CD4 T	0.2262	0.1745	-0.0492				

Table S3

Host features correlate with HIV reservoir characteristics.

The abundance of 133 immune cell populations was determined by flow cytometry and the HIV reservoir was quantified by intact proviral DNA assay for a cohort of 115 people with HIV (PWH). Each abundance and clinical and demographic variable was correlated with total HIV reservoir frequency, intact reservoir frequency, and the percentage of intact proviruses. Spearman correlation coefficients are shown for 69 variables that were correlated ($p < 0.05$) with one or more reservoir characteristics. Each feature/subset is ranked by the absolute value of the correlation coefficient for the total reservoir frequency. Variables that had significant p-values (< 0.05) after Benjamini–Hochberg correction for multiple comparisons (144 comparisons in our case) with one of the characteristics are shown in bold. For years of ART, correlations are computed based on 108 participants, excluding participants with missing years of ART values. For CD4 Nadir, correlations are computed based on 114 patients, excluding participants with the missing CD4 Nadir value.

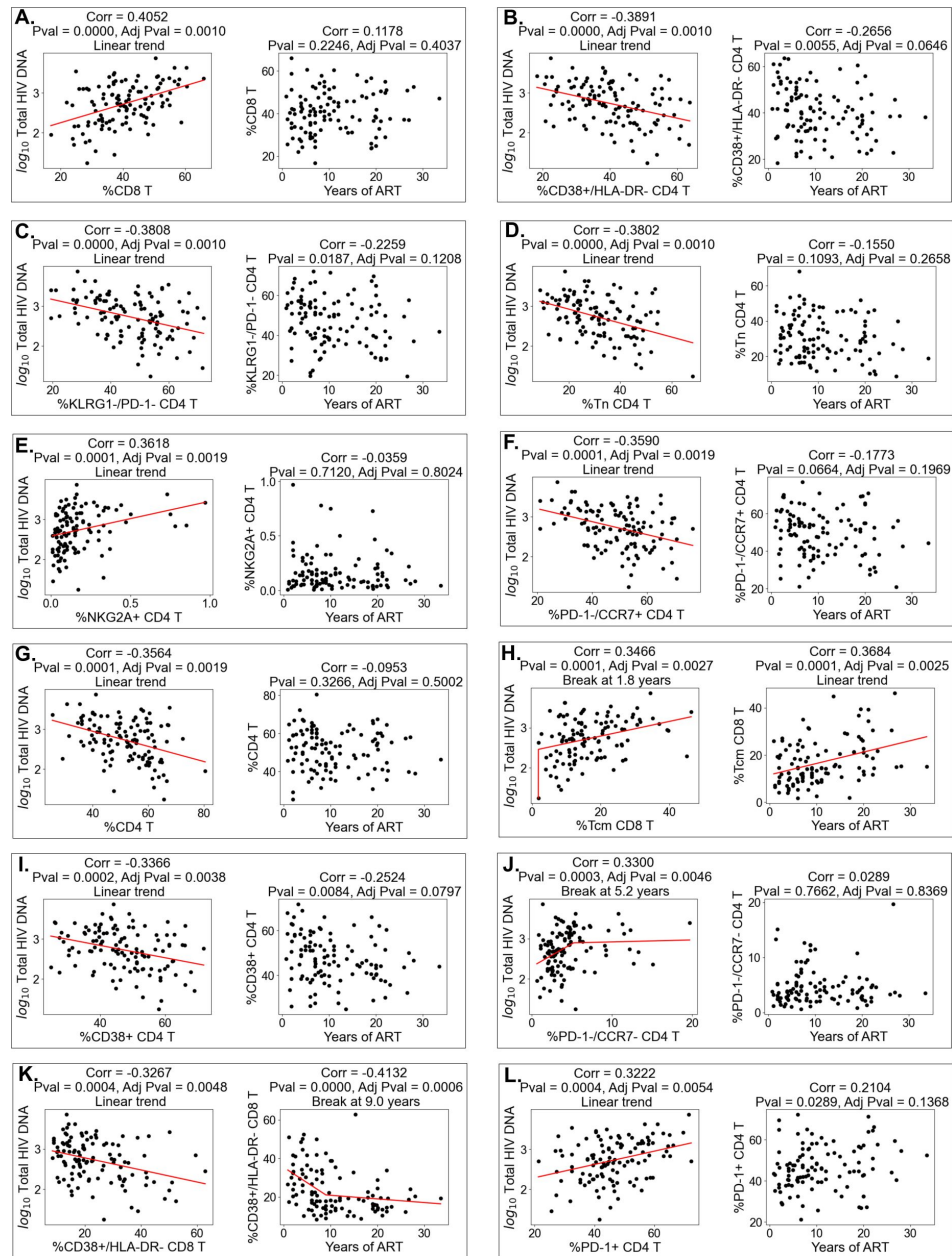


Figure S2

Abundance of immune cell subsets correlates with HIV reservoir. Continuation in Figures S3, S4.

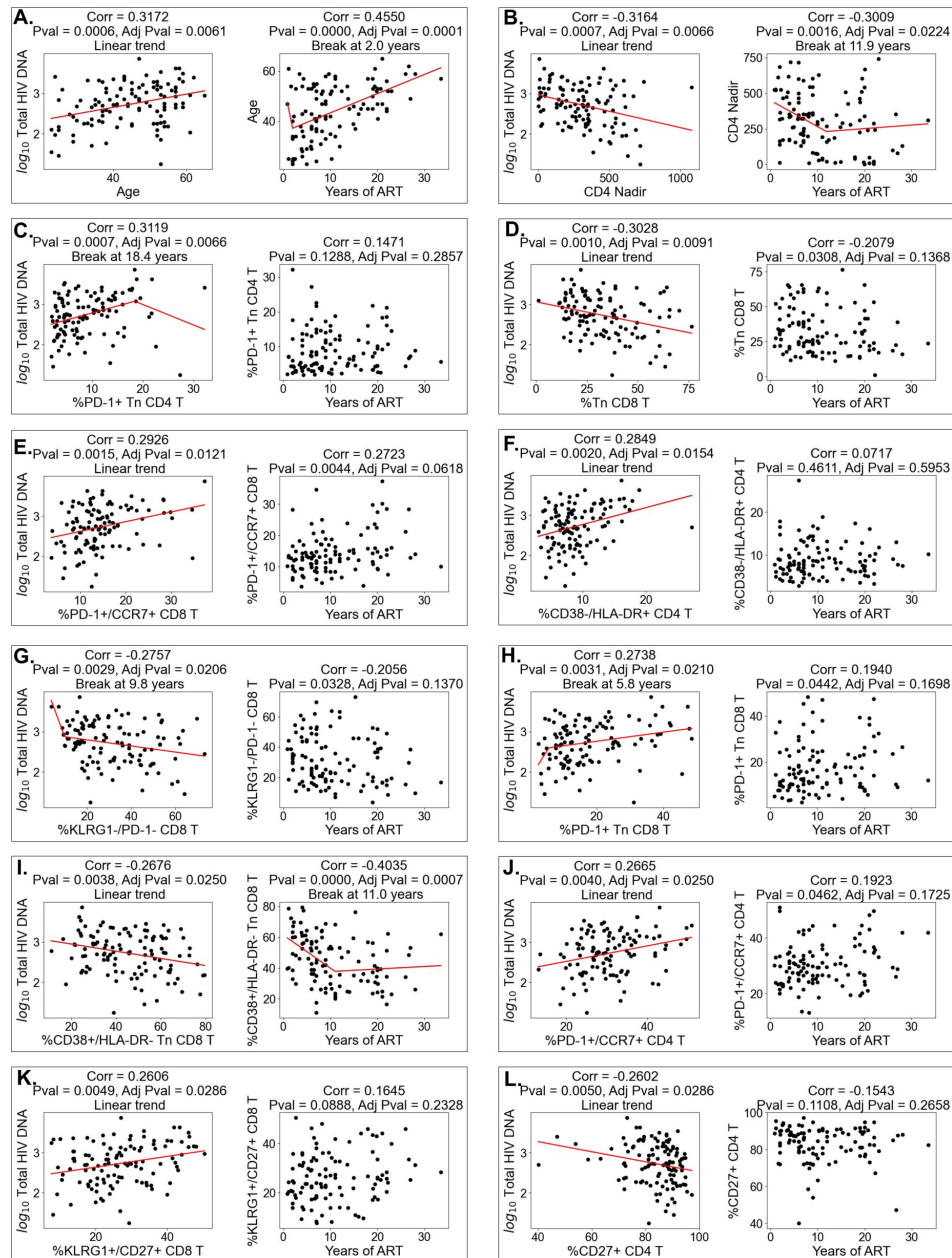


Figure S3

Abundance of immune cell subsets correlates with HIV reservoir. Continuation in Figures S2, S4.

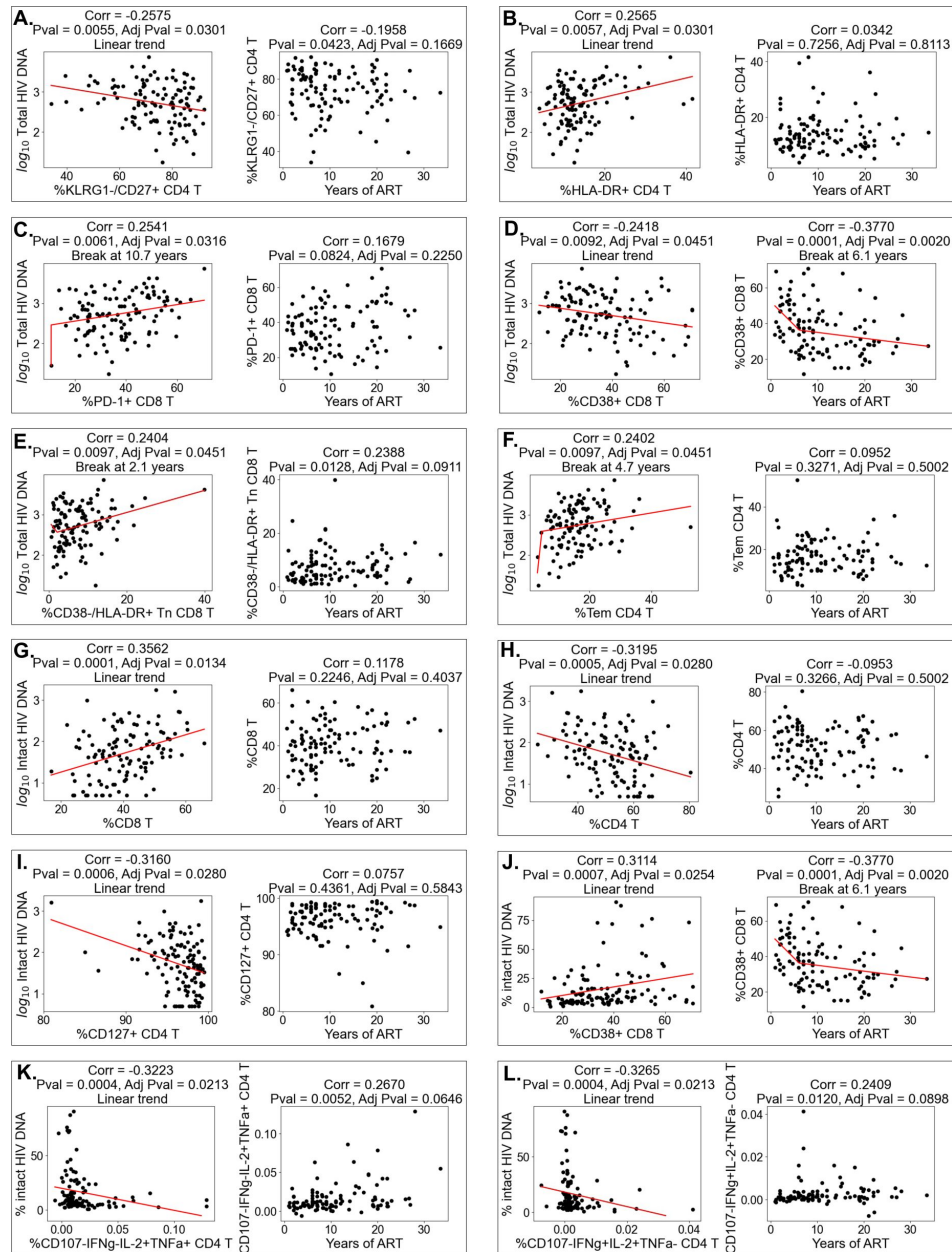


Figure S4

Abundance of immune cell subsets correlates with HIV reservoir. Continuation in Figures S2, S3.

Figure S5

CD4/CD8 and (%CD127+ CD4T)/CD8 ratios correlate with total and intact reservoir frequency.

Normalized (values transformed to be between 0 and 1) CD4/CD8 (**A**) and (%CD127+ CD4T)/CD8 (**B**) ratios are shown on the x-axis and the normalized natural logarithm ($\log_e$) of the total reservoir, intact reservoir and percent intact on the y-axis. Spearman correlation was computed between ratios and HIV reservoir characteristics. Outliers (red data points) were removed with the DBSCAN clustering algorithm and a linear regression model was fitted (black line) to the remaining data points. Spearman correlations, R^2 scores and mean squared error after outlier removal are displayed.

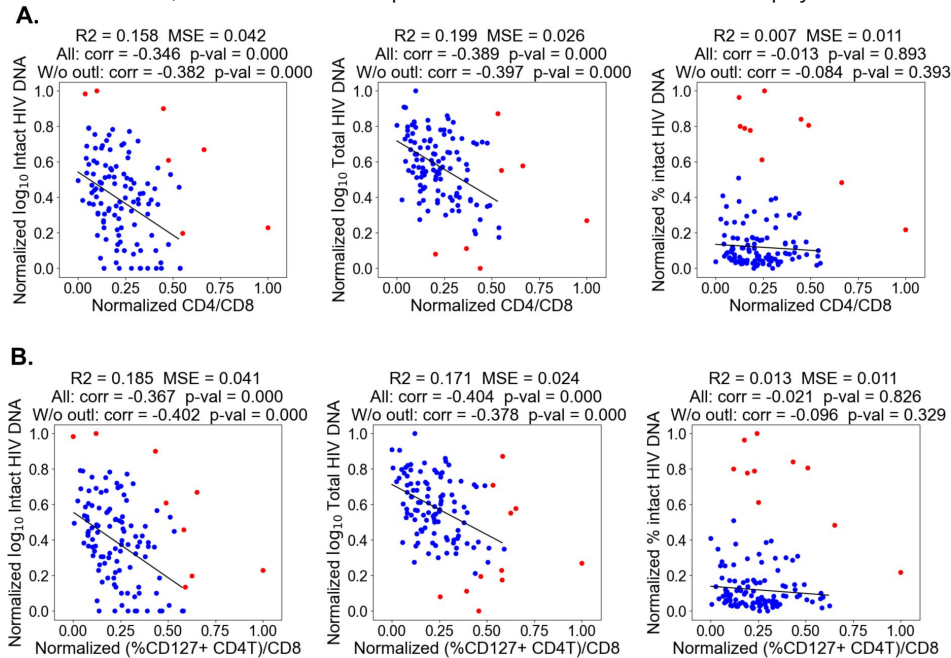


Table S4

PWH immune features correlate with years of ART.

The abundance of 133 immune cell populations was determined by flow cytometry and the HIV reservoir was quantified by intact proviral DNA assay for a cohort of 115 people with HIV (PWH). Each abundance and clinical and demographic variable was correlated with years of therapy for all participants (middle column) and for participants who had been on therapy for less than 10 years (right column). Spearman correlation coefficients are shown for all variables that were correlated ($p < 0.05$) after Benjamini-Hochberg correction with years of treatment for all participants or for participants that were on ART for less than 10 years. Variables that had significant p-value (< 0.05) after correction for multiple comparisons are shown in bold. Correlation is computed based on 108 PWH for all participants and on 60 PWH for PWH who were on therapy for less than 10 years). Computations exclude PWH with missing years of ART values. For CD4 Nadir, correlations are computed excluding the participant with missing CD4 Nadir value.

Variable	Spearman correlation coefficient, All participants	Spearman correlation coefficient, Participants with years of ART < 10 years
Age	0.4550	0.2629
%CD38+/HLA-DR- CD8 T	-0.4132	-0.3788
%CD38+/HLA-DR- Tn CD8 T	-0.4035	-0.3075
%CD38+ CD8 T	-0.3770	-0.3025
%Tcm CD8 T	0.3684	0.1595
%CD38-/HLA-DR- Tn CD8 T	0.3580	0.2222
%CD38+/HLA-DR- Tcm CD8 T	-0.3138	-0.2666
CD4 Nadir	-0.3078	0.1265
%CD107a+IFN γ -IL-2+TNF- CD8 T	0.2903	0.1728

Total HIV DNA model		Intact HIV DNA model		% intact HIV DNA model	
Variable	VIF value	Variable	VIF value	Variable	VIF value
Total HIV DNA	1.2545	Intact HIV DNA	1.0517	% intact HIV DNA	1.3469
Age	1.7426	Age	1.7243	Age	1.7220
Sex	1.1948	Sex	1.1616	Sex	1.1795
Race	1.5080	Race	1.4577	Race	1.6348
Years of ART	1.3048	Years of ART	1.3004	Years of ART	1.3181
CD4 nadir	1.3188	CD4 nadir	1.2875	CD4 nadir	1.3773
Recent CD4 count	1.3314	Recent CD4 count	1.3374	Recent CD4 count	1.3134
Years before ART	1.1367	Years before ART	1.1138	Years before ART	1.1611

Table S5

Multicollinearity analysis for variables used in models to predict immunophenotypes

. Variance Inflation Factor (VIF) for every variable is computed for models that fit immunophenotypes as a target and are based on clinical, demographic information and HIV characteristics (total, intact reservoir frequency or percent intact) as features. Note that here “Years before ART” is a continuous variable that we used for VIF computations only. In our other analysis we categorize this variable and use “Years before ART=NA”, “Years before ART< 1”, and “Years before ART≥ 1” instead. We excluded 3 participants with race indicated as “Other” to avoid near-binary categorical variables in VIF computation.

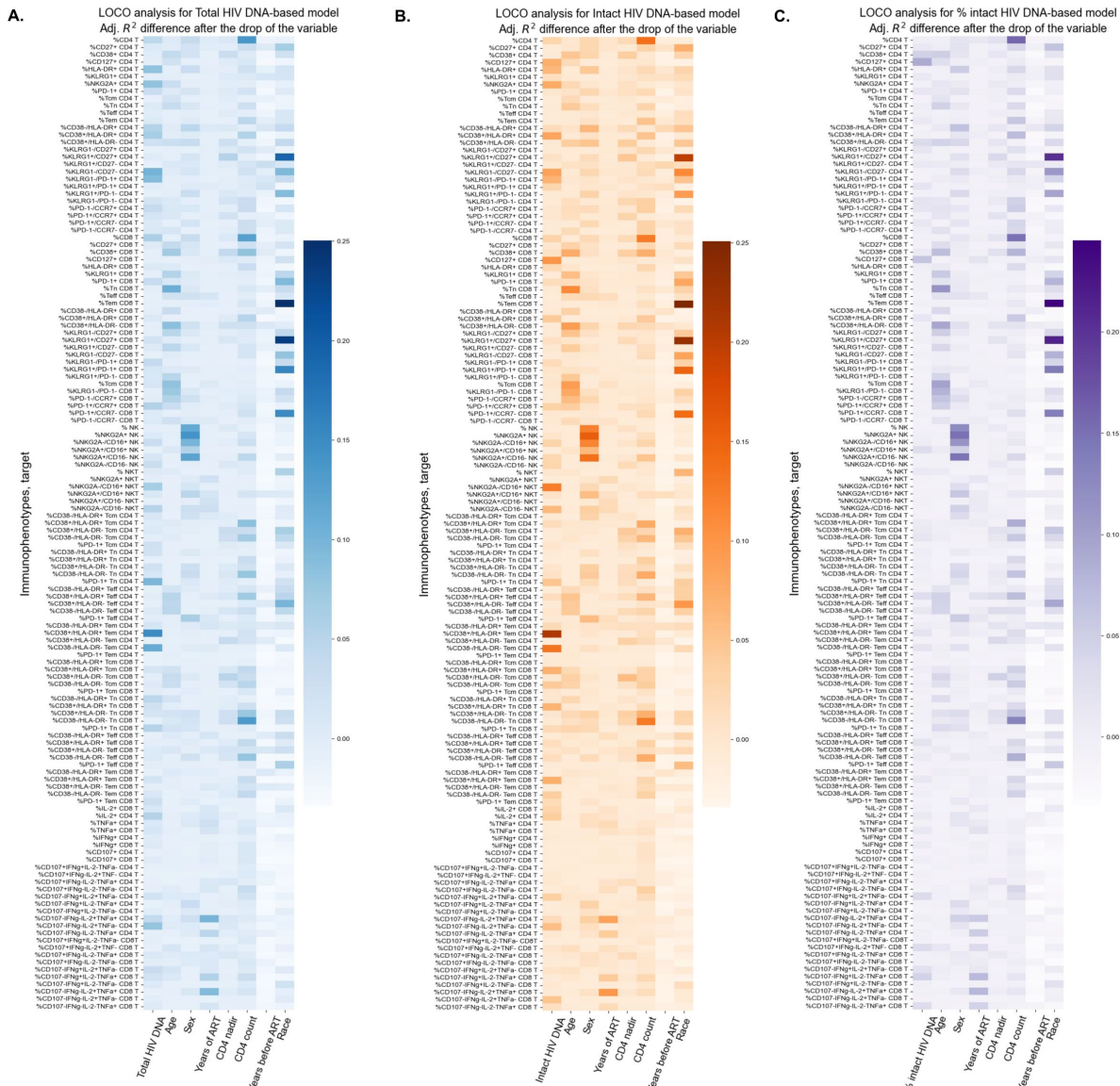


Figure S6

Leave-one-covariate-out (LOCO) analysis for clinical-demographic features and reservoir characteristics while predicting immunophenotypes.

Analysis that is described in [Figure 2](#) was performed for all 133 immunophenotypes considered in the study. Drops in adjusted R^2 scores are reported for models that use total reservoir frequency (A.), intact reservoir frequency (B.), or percent intact (C.) as features in addition to clinical and demographic information such as age, biological sex, race, years of treatment, CD4 nadir, recent CD4 count, and years of HIV before treatment (=NA, <1, ≥1). On the x-axis, we show features that were dropped from the model. On the y-axis, we display immunophenotypes, which are targets (outcomes) for the linear regression models. In [Tables S6](#)–[S8](#) we show the actual values of drops in adjusted R^2 score.

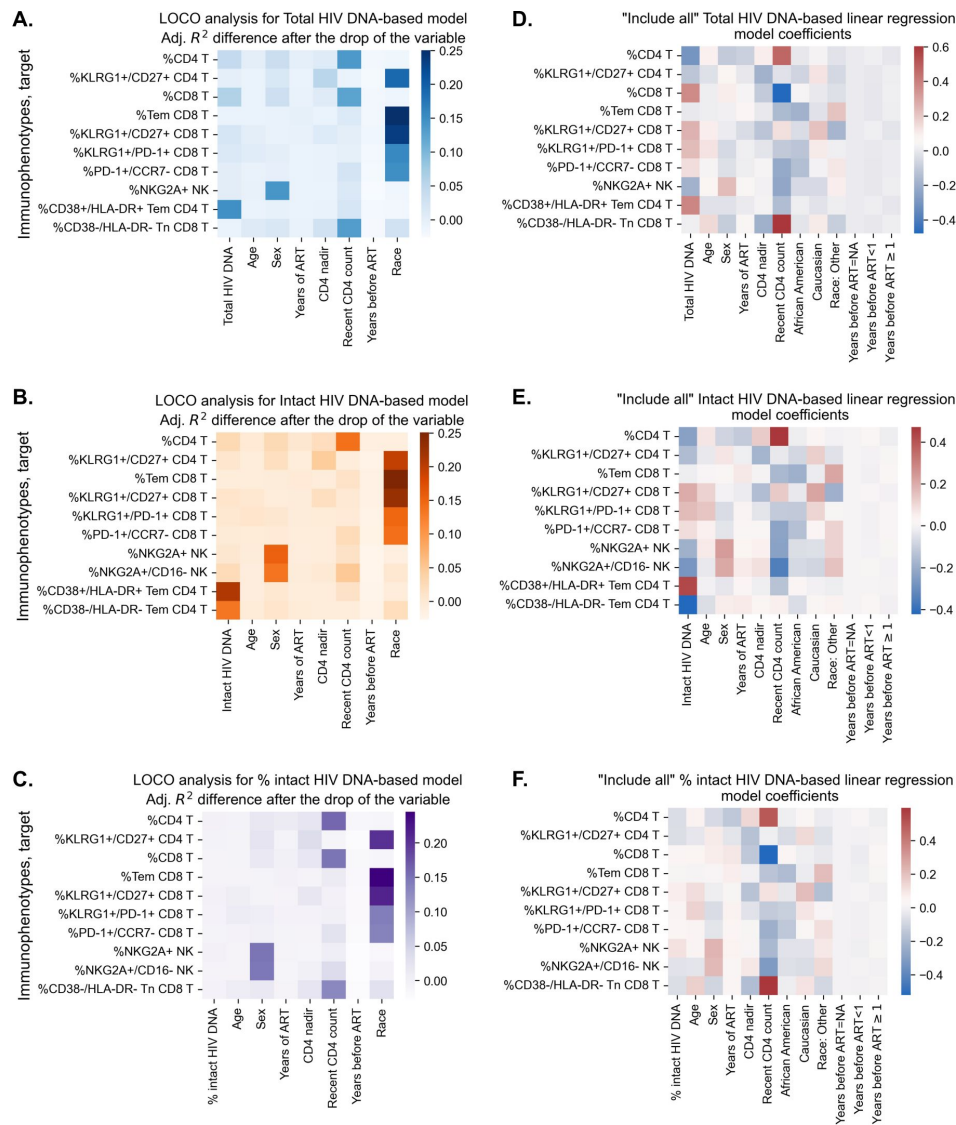


Figure S7

Coefficient visualization for linear regression models that predict immunophenotypes based on clinical, demographic information and HIV characteristics.

(A., B., C.) LOCO analysis from [Figure 3](#) for total-reservoir based (A.), intact reservoir-based (B.) and percent intact-based (C.) models. The drops in adjusted R^2 scores are shown after removing a feature and training a new model without it. (D., E., F.) Coefficient visualization for models that include clinical and demographic information such as age, biological sex, years of treatment, CD4 nadir, recent CD4 count, and years of HIV before treatment = NA, years of HIV before treatment < 1, years of HIV before treatment ≥ 1 and total reservoirs frequency (D.) or intact reservoir frequency (E.), or percent intact (F.). No features are dropped from these models, they are "Include all" models from [Tables S6](#)–[S8](#). On the x-axis, the feature is shown, and on the y-axis the target (immunophenotypes from (A., B., C.)). The heatmap displays the coefficient in front of that variable in the model (if the model is $\%CD4\ T = \beta_1 Total + \beta_2 Age + \beta_3 Sex + \dots$, then $\beta_1, \beta_2, \beta_3, \dots$ are visualized), where positive coefficients are shown in red and negative in blue.

Immunophenotype \ Model	Adj. R^2	ΔR^2 , difference in adjustable R^2 after dropping a covariate							
	Include all	Total HIV DNA	Age	Sex	Years of ART	CD4 nadir	Recent CD4 count	Years before ART	Race
%CD4 T	0.2583	0.0443	0.0024	0.0365	-0.0006	0.0079	0.1374	-0.0184	-0.0153
%CD27+ CD4 T	0.1031	0.0050	-0.0015	0.0191	-0.0017	-0.0028	0.0282	0.0026	0.0651
%CD38+ CD4 T	0.0970	-0.0011	0.0390	0.0201	-0.0092	0.0315	0.0017	-0.0241	0.0233
%CD127+ CD4 T	-0.0229	-0.0057	0.0235	-0.0076	-0.0105	0.0004	0.0085	-0.0299	-0.0203
%HLA-DR+ CD4 T	0.1862	0.0816	-0.0083	0.0294	-0.0061	-0.0043	0.0236	-0.0117	0.0161
%KLRG1+ CD4 T	0.0638	-0.0044	-0.0047	-0.0089	-0.0076	0.0029	0.0049	0.0135	0.0154
%NKG2A+ CD4 T	0.1082	0.0820	0.0303	-0.0043	-0.0077	-0.0020	-0.0034	-0.0214	-0.0097
%PD-1+ CD4 T	0.1175	0.0304	0.0017	0.0152	-0.0041	0.0025	0.0189	-0.0270	0.0019
%Tcm CD4 T	0.0192	-0.0101	0.0191	-0.0083	-0.0086	-0.0010	-0.0101	-0.0110	-0.0186
%Tn CD4 T	0.0960	0.0044	0.0343	0.0102	-0.0072	-0.0071	0.0249	-0.0276	-0.0179
%Teff CD4 T	-0.0298	-0.0013	-0.0073	-0.0033	-0.0090	0.0010	-0.0041	-0.0193	0.0111
%Tem CD4 T	0.0585	0.0046	-0.0072	0.0026	-0.0084	-0.0098	0.0450	-0.0060	0.0105
%CD38-/HLA-DR+ CD4 T	0.1711	0.0546	0.0004	0.0451	0.0052	0.0070	-0.0041	0.0099	0.0372
%CD38+/HLA-DR+ CD4 T	0.1159	0.0655	-0.0030	0.0012	-0.0076	-0.0089	0.0544	-0.0270	-0.0175
%CD38+/HLA-DR- CD4 T	0.1603	0.0246	0.0290	0.0351	-0.0077	0.0269	-0.0085	-0.0217	0.0383
%KLRG1-/CD27+ CD4 T	0.0438	-0.0090	-0.0078	-0.0094	-0.0056	0.0045	0.0198	0.0002	-0.0180
%KLRG1+/CD27+ CD4 T	0.2228	-0.0001	-0.0071	0.0153	-0.0080	0.0532	-0.0077	-0.0217	0.1907
%KLRG1+/CD27- CD4 T	0.0351	-0.0100	0.0064	0.0064	-0.0044	-0.0004	0.0117	0.0287	0.0109
%KLRG1-/CD27- CD4 T	0.2694	0.1007	-0.0005	0.0139	-0.0042	-0.0068	0.0213	-0.0046	0.0948
%KLRG1-/PD-1+ CD4 T	0.1404	0.0904	-0.0042	-0.0021	-0.0089	-0.0053	0.0005	0.0089	0.0346
%KLRG1+/PD-1+ CD4 T	0.0112	-0.0013	-0.0068	0.0024	-0.0003	-0.0048	0.0022	0.0200	-0.0254
%KLRG1+/PD-1- CD4 T	0.1251	-0.0087	-0.0053	0.0304	-0.0078	0.0066	-0.0006	-0.0134	0.0893
%KLRG1-/PD-1- CD4 T	0.1403	0.0178	0.0064	-0.0079	-0.0070	0.0170	0.0299	-0.0216	-0.0252
%PD-1-/CCR7+ CD4 T	0.1334	0.0369	0.0092	0.0089	-0.0066	-0.0038	0.0321	-0.0251	0.0135
%PD-1+/CCR7+ CD4 T	0.0690	0.0157	0.0073	-0.0067	-0.0061	0.0188	-0.0083	-0.0079	-0.0197
%PD-1+/CCR7- CD4 T	0.0433	0.0040	-0.0098	0.0164	-0.0085	-0.0096	0.0284	-0.0052	0.0045
%PD-1-/CCR7- CD4 T	0.0038	0.0072	0.0134	-0.0100	-0.0076	-0.0029	0.0276	-0.0181	0.0040
%CD8 T	0.2546	0.0603	-0.0077	0.0334	-0.0066	0.0068	0.1239	-0.0188	-0.0138
%CD27+ CD8 T	0.0584	0.0020	0.0364	0.0119	-0.0091	-0.0097	0.0243	-0.0058	0.0162
%CD38+ CD8 T	0.2543	-0.0068	0.0692	0.0085	-0.0062	0.0382	0.0806	-0.0005	-0.0191
%CD127+ CD8 T	-0.0073	-0.0073	-0.0099	-0.0075	-0.0037	0.0053	0.0000	-0.0168	0.0226
%HLA-DR+ CD8 T	-0.0012	0.0039	0.0033	-0.0104	-0.0103	-0.0067	0.0299	-0.0201	-0.0255
%KLRG1+ CD8 T	0.2143	-0.0082	0.0547	0.0041	-0.0067	-0.0039	-0.0017	-0.0176	0.0456
%PD-1+ CD8 T	0.2642	0.0358	0.0090	0.0010	-0.0071	-0.0073	-0.0048	-0.0222	0.0932
%Tn CD8 T	0.2975	-0.0052	0.1074	0.0081	-0.0070	-0.0073	0.0088	-0.0061	0.0343
%Teff CD8 T	-0.0151	-0.0089	0.0207	0.0237	0.0126	-0.0103	-0.0037	-0.0251	0.0106
%Tem CD8 T	0.2716	-0.0076	-0.0075	-0.0061	-0.0020	-0.0076	0.0118	-0.0186	0.2503
%CD38-/HLA-DR+ CD8 T	0.0074	-0.0061	0.0323	-0.0099	-0.0103	-0.0102	-0.0035	0.0015	-0.0256
%CD38+/HLA-DR+ CD8 T	0.0112	0.0069	-0.0099	-0.0091	-0.0100	0.0023	0.0509	-0.0301	-0.0184
%CD38+/HLA-DR- CD8 T	0.3161	-0.0044	0.0852	0.0236	-0.0038	0.0223	0.0228	0.0063	-0.0009
%KLRG1-/CD27+ CD8 T	0.2091	-0.0082	0.0484	0.0075	-0.0063	-0.0038	0.0037	-0.0153	0.0421
%KLRG1+/CD27+ CD8 T	0.3370	0.0190	0.0007	-0.0069	-0.0055	0.0097	0.0013	-0.0166	0.2342
%KLRG1+/CD27- CD8 T	0.0396	0.0025	0.0452	0.0089	-0.0097	-0.0098	0.0148	-0.0104	-0.0074
%KLRG1-/CD27- CD8 T	0.0770	-0.0091	0.0004	-0.0058	-0.0083	-0.0096	0.0147	-0.0243	0.0798
%KLRG1-/PD-1+ CD8 T	0.0189	0.0384	-0.0052	-0.0098	-0.0091	-0.0102	-0.0097	-0.0299	0.0340
%KLRG1+/PD-1+ CD8 T	0.3163	0.0130	0.0057	0.0017	-0.0070	-0.0069	-0.0025	-0.0206	0.1568
%KLRG1+/PD-1- CD8 T	-0.0235	0.0086	0.0497	-0.0063	-0.0064	-0.0046	-0.0083	-0.0219	-0.0302
%Tcm CD8 T	0.2293	0.0124	0.0817	-0.0076	0.0043	-0.0076	-0.0031	-0.0230	-0.0154
%KLRG1-/PD-1- CD8 T	0.2662	-0.0041	0.0740	0.0077	-0.0067	-0.0024	-0.0014	-0.0147	0.0375
%PD-1-/CCR7+ CD8 T	0.1968	-0.0032	0.0539	0.0226	-0.0035	-0.0075	0.0040	-0.0130	0.0190
%PD-1+/CCR7+ CD8 T	0.1735	0.0550	0.0300	-0.0011	-0.0076	0.0003	0.0133	-0.0252	-0.0145

Table S6

Adjusted R^2 scores and differences in adjusted R^2 for LOCO analysis for the model that contains total reservoir frequency.

The first column contains all immunophenotypes and is a target for the regression model. The second column contains adjusted R^2 of the linear regression model with features such as total reservoir frequency, age, biological sex, race, years of treatment, CD4 nadir, recent CD4 count, and years of HIV before treatment (=NA, < 1, ≥ 1). The next seven columns show differences in adjusted R^2 score (ΔR^2) after removing a specific feature. Participants with missing years of ART values are excluded from this analysis. The missing value of the CD4 nadir for one participant is imputed.

%PD-1+/CCR7- CD8 T	0.2468	0.0035	-0.0063	-0.0034	-0.0078	-0.0061	0.0226	-0.0219	0.1493
%PD-1-/CCR7- CD8 T	-0.0373	0.0053	0.0244	0.0060	0.0023	-0.0104	-0.0033	-0.0173	-0.0119
% NK	0.1125	-0.0065	-0.0081	0.1126	-0.0063	0.0048	-0.0084	-0.0247	-0.0051
%NKG2A+ NK	0.0933	0.0035	-0.0053	0.1413	-0.0080	-0.0088	0.0130	-0.0264	-0.0165
%NKG2A-/CD16+ NK	0.0840	0.0297	-0.0071	0.0991	-0.0080	-0.0027	0.0271	-0.0262	-0.0148
%NKG2A+/CD16+ NK	-0.0016	-0.0095	0.0071	0.0554	-0.0103	-0.0096	-0.0104	-0.0259	-0.0296
%NKG2A-/CD16- NK	0.1141	0.0121	-0.0089	0.1223	-0.0046	-0.0050	0.0449	-0.0271	-0.0038
%NKG2A-/CD16- NK	-0.0205	0.0306	-0.0099	-0.0025	-0.0106	0.0065	0.0015	-0.0309	-0.0301
% NKT	-0.0062	0.0020	0.0002	-0.0105	-0.0099	-0.0088	0.0003	-0.0297	0.0564
%NKG2A+ NKT	-0.0306	-0.0060	-0.0104	0.0000	0.0248	-0.0099	-0.0098	-0.0314	-0.0192
%NKG2A-/CD16+ NKT	0.0151	0.0655	-0.0070	0.0127	-0.0074	-0.0077	-0.0086	-0.0264	-0.0178
%NKG2A+/CD16+ NKT	0.0394	0.0089	-0.0100	0.0468	0.0118	-0.0096	0.0140	0.0052	0.0017
%NKG2A+/CD16- NKT	-0.0391	-0.0082	-0.0105	-0.0057	0.0204	-0.0096	-0.0075	-0.0315	-0.0205
%NKG2A-/CD16- NKT	0.0722	0.0513	-0.0068	0.0266	0.0279	-0.0096	-0.0097	-0.0274	-0.0004
%CD38-/HLA-DR+ Tcm CD4 T	-0.0030	0.0104	-0.0102	0.0238	0.0046	-0.0005	-0.0103	-0.0254	-0.0014
%CD38+/HLA-DR+ Tcm CD4 T	0.0657	-0.0001	0.0020	-0.0031	-0.0044	-0.0024	0.0839	-0.0286	-0.0213
%CD38-/HLA-DR- Tcm CD4 T	0.0952	0.0071	-0.0030	-0.0034	-0.0090	0.0325	0.0299	-0.0276	0.0636
%CD38-/HLA-DR- Tcm CD4 T	0.1248	-0.0062	0.0012	-0.0090	-0.0091	0.0276	0.0680	-0.0264	0.0442
%PD-1+ Tcm CD4 T	0.0118	0.0225	-0.0098	-0.0102	-0.0089	0.0017	-0.0019	-0.0302	0.0069
%CD38-/HLA-DR+ Tn CD4 T	-0.0410	0.0206	-0.0080	0.0040	-0.0045	-0.0059	-0.0078	-0.0260	-0.0213
%CD38+/HLA-DR+ Tn CD4 T	-0.0505	-0.0065	-0.0105	-0.0107	-0.0097	-0.0106	0.0317	-0.0317	-0.0137
%CD38+/HLA-DR- Tn CD4 T	0.0408	0.0226	-0.0095	0.0273	-0.0094	0.0053	0.0224	-0.0285	0.0064
%CD38-/HLA-DR- Tn CD4 T	0.0616	0.0115	-0.0097	0.0341	-0.0093	0.0094	0.0772	-0.0276	-0.0104
%PD-1+ Tn CD4 T	0.0946	0.0970	-0.0085	0.0065	-0.0074	0.0011	-0.0072	-0.0254	0.0230
%CD38-/HLA-DR+ Teff CD4 T	0.0287	-0.0086	0.0171	0.0008	-0.0100	-0.0017	0.0122	-0.0125	0.0352
%CD38+/HLA-DR+ Teff CD4 T	0.1811	-0.0032	0.0474	-0.0080	0.0024	0.0106	0.0711	-0.0241	0.0464
%CD38+/HLA-DR- Teff CD4 T	0.1149	-0.0078	0.0454	-0.0064	-0.0090	-0.0089	0.0035	0.0016	0.0993
%CD38-/HLA-DR- Teff CD4 T	0.1315	0.0023	0.0462	-0.0089	-0.0055	0.0132	0.0289	-0.0265	0.0438
%PD-1+ Teff CD4 T	0.0153	-0.0022	-0.0046	0.0616	0.0091	-0.0077	0.0027	-0.0246	-0.0116
%CD38-/HLA-DR+ Tem CD4 T	0.0211	0.0126	-0.0041	0.0020	0.0123	0.0045	0.0083	0.0021	0.0096
%CD38+/HLA-DR+ Tem CD4 T	0.1387	0.1481	-0.0077	-0.0040	-0.0090	-0.0064	0.0186	-0.0200	-0.0172
%CD38+/HLA-DR- Tem CD4 T	-0.0344	-0.0091	-0.0088	-0.0083	-0.0106	0.0357	0.0169	0.0010	-0.0308
%CD38-/HLA-DR- Tem CD4 T	0.0779	0.1106	-0.0071	0.0007	0.0022	-0.0093	-0.0064	-0.0279	0.0054
%PD-1+ Tem CD4 T	-0.1111	-0.0115	-0.0050	-0.0108	-0.0093	-0.0116	-0.0116	-0.0340	-0.0273
%CD38-/HLA-DR+ Tcm CD8 T	-0.0878	-0.0112	-0.0087	-0.0079	-0.0101	-0.0107	-0.0063	-0.0325	-0.0232
%CD38+/HLA-DR+ Tcm CD8 T	0.0513	0.0153	0.0206	-0.0067	-0.0088	0.0086	0.0399	-0.0201	-0.0203
%CD38+/HLA-DR- Tcm CD8 T	0.0880	-0.0057	0.0342	0.0043	-0.0094	0.0495	0.0284	-0.0255	-0.0266
%CD38-/HLA-DR- Tcm CD8 T	0.0719	0.0021	0.0298	-0.0068	-0.0096	0.0164	0.0489	-0.0249	-0.0271
%PD-1+ Tcm CD8 T	-0.0177	-0.0038	-0.0065	-0.0009	-0.0090	-0.0104	0.0083	-0.0310	-0.0102
%CD38-/HLA-DR+ Tn CD8 T	0.0511	0.0455	0.0163	-0.0066	-0.0075	-0.0021	-0.0089	-0.0285	-0.0246
%CD38+/HLA-DR+ Tn CD8 T	-0.0188	0.0216	-0.0024	-0.0098	-0.0044	-0.0009	0.0257	-0.0305	-0.0295
%CD38+/HLA-DR- Tn CD8 T	0.2816	0.0145	0.0168	0.0247	-0.0051	0.0255	0.0751	-0.0209	0.0336
%CD38-/HLA-DR- Tn CD8 T	0.2317	-0.0079	0.0090	0.0168	-0.0059	0.0274	0.1329	-0.0223	0.0274
%PD-1+ Tn CD8 T	0.0747	0.0523	0.0190	0.0119	-0.0077	-0.0075	-0.0011	-0.0283	-0.0223
%CD38-/HLA-DR+ Teff CD8 T	0.0617	-0.0096	0.0031	-0.0047	0.0126	-0.0085	-0.0091	0.0011	0.0306
%CD38+/HLA-DR+ Teff CD8 T	-0.0083	-0.0044	-0.0046	0.0012	-0.0102	0.0004	0.0389	-0.0280	0.0003
%CD38+/HLA-DR- Teff CD8 T	0.1117	0.0032	0.0100	-0.0056	0.0071	0.0161	0.0178	0.0111	0.0278
%CD38-/HLA-DR- Teff CD8 T	0.0079	0.0048	-0.0070	0.0218	-0.0080	0.0140	0.0789	-0.0294	-0.0279
%PD-1+ Teff CD8 T	0.0093	-0.0035	-0.0057	-0.0074	-0.0099	-0.0093	0.0000	-0.0249	0.0604
%CD38-/HLA-DR+ Tem CD8 T	-0.0024	-0.0095	0.0033	-0.0021	-0.0020	-0.0099	-0.0104	0.0058	-0.0132
%CD38+/HLA-DR+ Tem CD8 T	-0.0234	0.0178	-0.0045	-0.0106	-0.0083	-0.0042	0.0334	-0.0310	-0.0304
%CD38+/HLA-DR- Tem CD8 T	0.0303	-0.0046	-0.0054	-0.0096	0.0046	-0.0015	0.0181	0.0005	-0.0015
%CD38-/HLA-DR- Tem CD8 T	-0.0335	0.0082	-0.0108	-0.0053	-0.0074	0.0029	0.0410	-0.0282	-0.0278
%PD-1+ Tem CD8 T	-0.0275	0.0265	-0.0085	0.0005	-0.0106	-0.0076	-0.0102	-0.0309	-0.0142
%IL-2+ CD8 T	-0.0375	0.0109	-0.0099	-0.0092	0.0091	-0.0108	-0.0092	-0.0303	0.0064
%IL-2+ CD4 T	0.0355	0.0529	-0.0051	0.0031	0.0119	-0.0015	-0.0090	-0.0166	-0.0128

Table S6 (continued)

%TNFa+ CD4 T	-0.0180	0.0065	-0.0106	0.0065	0.0256	-0.0099	0.0059	-0.0144	-0.0149
%TNFa+ CD8 T	-0.0562	0.0118	-0.0046	-0.0086	0.0152	-0.0041	0.0119	-0.0294	-0.0309
%IFNg+ CD4 T	-0.0646	-0.0003	-0.0099	-0.0067	-0.0085	-0.0110	-0.0055	-0.0254	-0.0133
%IFNg+ CD8 T	-0.0831	0.0019	-0.0096	-0.0110	-0.0039	-0.0100	0.0092	-0.0285	-0.0230
%CD107+ CD4 T	-0.0805	-0.0101	-0.0080	-0.0112	-0.0037	-0.0109	0.0090	-0.0263	-0.0263
%CD107+ CD8 T	-0.0856	0.0007	-0.0078	-0.0077	-0.0007	-0.0088	0.0098	-0.0332	-0.0311
%CD107+IFNg+IL-2-TNFa- CD4 T	-0.0835	-0.0105	-0.0076	-0.0097	-0.0098	-0.0111	0.0007	-0.0272	-0.0260
%CD107+IFNg-IL-2-TNF- CD4 T	-0.0745	-0.0108	-0.0107	-0.0109	-0.0107	-0.0075	0.0036	-0.0157	-0.0255
%CD107+IFNg-IL-2-TNFa+ CD4 T	-0.0750	-0.0097	-0.0112	-0.0100	0.0067	0.0115	-0.0030	-0.0170	-0.0284
%CD107+IFNg-IL-2-TNFa- CD4 T	-0.0386	-0.0064	-0.0001	-0.0098	0.0090	-0.0094	0.0339	-0.0318	-0.0249
%CD107-IFNg+IL-2-TNFa- CD4 T	-0.0170	0.0203	0.0030	0.0028	-0.0102	-0.0050	0.0214	-0.0262	-0.0130
%CD107-IFNg+IL-2-TNFa+ CD4 T	-0.0478	-0.0108	-0.0051	0.0295	-0.0068	-0.0029	0.0007	-0.0319	-0.0244
%CD107-IFNg+IL-2-TNFa- CD4 T	-0.0048	-0.0009	-0.0100	0.0054	-0.0052	-0.0104	-0.0023	-0.0304	-0.0055
%CD107-IFNg-IL-2-TNFa+ CD4 T	0.1865	0.0390	-0.0016	0.0255	0.0991	0.0104	0.0024	-0.0133	-0.0177
%CD107-IFNg-IL-2-TNFa- CD4 T	0.0568	0.0783	0.0015	0.0032	-0.0087	0.0079	-0.0073	-0.0222	-0.0152
%CD107-IFNg-IL-2-TNFa+ CD4 T	-0.0205	-0.0085	-0.0105	0.0024	0.0467	-0.0025	0.0011	-0.0100	-0.0184
%CD107+IFNg+IL-2-TNFa- CD8T	-0.0906	-0.0061	-0.0113	-0.0113	-0.0109	-0.0112	0.0061	-0.0301	-0.0221
%CD107+IFNg-IL-2-TNF- CD8 T	-0.0248	-0.0103	0.0132	-0.0107	0.0117	-0.0019	-0.0086	-0.0266	-0.0106
%CD107+IFNg-IL-2-TNFa+ CD8 T	-0.0177	0.0067	0.0014	0.0051	0.0033	0.0101	0.0005	-0.0285	-0.0218
%CD107+IFNg-IL-2-TNFa- CD8 T	-0.0361	-0.0035	-0.0048	0.0036	0.0063	-0.0091	0.0084	-0.0136	-0.0292
%CD107-IFNg+IL-2-TNFa- CD8 T	-0.0046	0.0323	0.0150	-0.0090	0.0069	-0.0104	0.0018	-0.0238	0.0026
%CD107-IFNg+IL-2-TNFa+ CD8 T	0.0194	0.0193	0.0068	-0.0102	0.0553	-0.0058	0.0312	-0.0293	-0.0262
%CD107-IFNg+IL-2-TNFa- CD8 T	-0.0374	0.0055	-0.0105	-0.0108	-0.0065	-0.0092	0.0036	-0.0301	0.0213
%CD107-IFNg-IL-2-TNFa+ CD8 T	0.0283	-0.0101	-0.0063	0.0087	0.0924	-0.0004	0.0149	-0.0269	-0.0066
%CD107-IFNg-IL-2-TNFa- CD8 T	-0.0558	-0.0022	-0.0102	-0.0100	-0.0109	-0.0084	-0.0081	-0.0278	-0.0006
%CD107-IFNg-IL-2-TNFa+ CD8 T	-0.0037	0.0017	-0.0094	-0.0092	0.0149	-0.0103	-0.0104	-0.0248	-0.0188

Table S6 (continued)

Immunophenotype	Model	Adj. R^2	ΔR^2 , difference in adjustable R^2 after dropping a covariate							
			Intact HIV DNA	Age	Sex	Years of ART	CD4 nadir	Recent CD4 count	Years before ART	Race
%CD4 T		0.2423	0.0283	-0.0013	0.0266	0.0047	0.0204	0.1378	-0.0138	-0.0131
%CD27+ CD4 T		0.0896	-0.0086	0.0016	0.0264	0.0021	-0.0056	0.0325	0.0069	0.0708
%CD38+ CD4 T		0.0906	-0.0075	0.0482	0.0276	-0.0083	0.0373	-0.0021	-0.0235	0.0275
%CD127+ CD4 T		0.0550	0.0722	0.0202	-0.0074	-0.0095	-0.0028	0.0008	-0.0254	-0.0242
%HLA-DR+ CD4 T		0.1658	0.0612	-0.0072	0.0450	-0.0085	0.0059	0.0233	-0.0010	0.0286
%KLRG1+ CD4 T		0.0818	0.0136	-0.0043	-0.0085	-0.0076	-0.0003	0.0084	0.0084	0.0192
%NKG2A+ CD4 T		0.1013	0.0751	0.0221	0.0017	-0.0093	0.0112	-0.0041	-0.0130	0.0017
%PD-1+ CD4 T		0.0859	-0.0012	0.0072	0.0252	0.0007	0.0109	0.0229	-0.0271	0.0068
%Tcm CD4 T		0.0191	-0.0102	0.0202	-0.0080	-0.0084	-0.0006	-0.0101	-0.0111	-0.0187
%Tn CD4 T		0.0829	-0.0088	0.0421	0.0161	-0.0051	-0.0050	0.0290	-0.0277	-0.0173
%Teff CD4 T		-0.0378	-0.0093	-0.0058	-0.0005	-0.0100	-0.0021	-0.0030	-0.0169	0.0145
%Tem CD4 T		0.0448	-0.0092	-0.0053	0.0073	-0.0067	-0.0098	0.0504	-0.0024	0.0126
%CD38-/HLA-DR+ CD4 T		0.1345	0.0180	0.0058	0.0626	-0.0016	0.0201	-0.0033	0.0220	0.0482
%CD38+/HLA-DR+ CD4 T		0.1327	0.0823	-0.0052	0.0081	-0.0044	-0.0078	0.0497	-0.0243	-0.0094
%CD38+/HLA-DR- CD4 T		0.1300	-0.0057	0.0394	0.0484	-0.0054	0.0395	-0.0091	-0.0197	0.0452
%KLRG1-/CD27+ CD4 T		0.0473	-0.0055	-0.0066	-0.0088	-0.0040	0.0054	0.0249	0.0002	-0.0179
%KLRG1+/CD27+ CD4 T		0.2304	0.0074	-0.0068	0.0183	-0.0078	0.0457	-0.0074	-0.0226	0.1975
%KLRG1+/CD27- CD4 T		0.0426	-0.0026	0.0088	0.0084	-0.0032	0.0006	0.0158	0.0271	0.0133
%KLRG1-/CD27- CD4 T		0.2528	0.0842	-0.0044	0.0268	0.0015	0.0005	0.0204	-0.0152	0.1125
%KLRG1-/PD-1+ CD4 T		0.1170	0.0671	0.0001	0.0053	-0.0082	0.0051	0.0003	-0.0061	0.0500
%KLRG1+/PD-1+ CD4 T		0.0391	0.0266	-0.0063	0.0020	-0.0005	-0.0072	0.0064	0.0130	-0.0212
%KLRG1+/PD-1- CD4 T		0.1276	-0.0062	-0.0051	0.0315	-0.0078	0.0057	0.0004	-0.0143	0.0908
%KLRG1-/PD-1- CD4 T		0.1160	-0.0065	0.0123	-0.0059	-0.0043	0.0266	0.0353	-0.0195	-0.0264
%PD-1-/CCR7+ CD4 T		0.0957	-0.0008	0.0170	0.0182	-0.0028	0.0025	0.0378	-0.0238	0.0195
%PD-1+/CCR7+ CD4 T		0.0520	-0.0013	0.0123	-0.0040	-0.0031	0.0294	-0.0078	-0.0135	-0.0172
%PD-1+/CCR7- CD4 T		0.0305	-0.0088	-0.0092	0.0231	-0.0068	-0.0101	0.0323	-0.0013	0.0073
%PD-1-/CCR7- CD4 T		-0.0127	-0.0092	0.0196	-0.0105	-0.0095	-0.0062	0.0322	-0.0151	0.0073
%CD8 T		0.2259	0.0315	-0.0068	0.0220	-0.0036	0.0206	0.1265	-0.0141	-0.0085
%CD27+ CD8 T		0.0558	-0.0007	0.0334	0.0088	-0.0083	-0.0089	0.0241	-0.0101	0.0191
%CD38+ CD8 T		0.2703	0.0092	0.0742	0.0098	-0.0057	0.0352	0.0711	-0.0026	-0.0173
%CD127+ CD8 T		0.0994	0.0993	-0.0094	-0.0074	-0.0043	0.0017	-0.0059	-0.0076	0.0073
%HLA-DR+ CD8 T		0.0171	0.0221	0.0046	-0.0100	-0.0099	-0.0089	0.0254	-0.0148	-0.0276
%KLRG1+ CD8 T		0.2143	-0.0082	0.0556	0.0043	-0.0066	-0.0038	-0.0018	-0.0176	0.0461
%PD-1+ CD8 T		0.2452	0.0167	0.0147	0.0066	-0.0055	-0.0046	-0.0046	-0.0210	0.0797
%Tn CD8 T		0.2965	-0.0061	0.1125	0.0100	-0.0072	-0.0072	0.0090	-0.0045	0.0315
%Teff CD8 T		-0.0162	-0.0100	0.0196	0.0225	0.0150	-0.0101	-0.0041	-0.0259	0.0091
%Tem CD8 T		0.2719	-0.0073	-0.0075	-0.0061	-0.0018	-0.0076	0.0123	-0.0188	0.2507
%CD38-/HLA-DR+ CD8 T		0.0079	-0.0057	0.0350	-0.0101	-0.0103	-0.0098	-0.0038	0.0053	-0.0255
%CD38+/HLA-DR+ CD8 T		0.0464	0.0422	-0.0096	-0.0080	-0.0093	-0.0028	0.0427	-0.0291	-0.0225
%CD38+/HLA-DR- CD8 T		0.3133	-0.0071	0.0917	0.0274	-0.0026	0.0258	0.0206	0.0081	-0.0019
%KLRG1-/CD27+ CD8 T		0.2091	-0.0082	0.0498	0.0081	-0.0063	-0.0036	0.0039	-0.0152	0.0423
%KLRG1+/CD27+ CD8 T		0.3275	0.0095	0.0035	-0.0069	-0.0067	0.0184	0.0008	-0.0190	0.2280
%KLRG1+/CD27- CD8 T		0.0342	-0.0029	0.0412	0.0056	-0.0091	-0.0090	0.0140	-0.0143	-0.0055
%KLRG1-/CD27- CD8 T		0.0784	-0.0077	-0.0008	-0.0047	-0.0087	-0.0096	0.0178	-0.0242	0.0830
%KLRG1-/PD-1+ CD8 T		-0.0004	0.0190	-0.0020	-0.0081	-0.0066	-0.0081	-0.0101	-0.0291	0.0330
%KLRG1+/PD-1+ CD8 T		0.3072	0.0040	0.0091	0.0055	-0.0064	-0.0055	-0.0022	-0.0199	0.1476
%KLRG1+/PD-1- CD8 T		-0.0333	-0.0012	0.0441	-0.0084	-0.0037	-0.0079	-0.0088	-0.0253	-0.0310
%Tcm CD8 T		0.2214	0.0045	0.0914	-0.0081	0.0087	-0.0059	-0.0035	-0.0220	-0.0179
%KLRG1-/PD-1- CD8 T		0.2653	-0.0051	0.0785	0.0099	-0.0070	-0.0006	-0.0014	-0.0131	0.0347
%PD-1-/CCR7+ CD8 T		0.1951	-0.0049	0.0583	0.0265	-0.0045	-0.0065	0.0041	-0.0107	0.0155
%PD-1+/CCR7+ CD8 T		0.1522	0.0336	0.0400	0.0051	-0.0049	0.0113	0.0122	-0.0255	-0.0152

Table S7

Adjusted R^2 scores and differences in adjusted R^2 for LOCO analysis for the model that contains intact reservoir frequency.

The first column contains all immunophenotypes and is a target for the regression model. The second column contains adjusted R^2 of the linear regression model with features such as intact reservoir frequency, age, biological sex, race, years of treatment, CD4 nadir, recent CD4 count, and years of HIV before treatment (=NA, < 1, ≥ 1). The next seven columns show differences in adjusted R^2 score (ΔR^2) after removing a specific feature. Participants with missing years of ART values are excluded from this analysis. The missing value of the CD4 nadir for one participant is imputed.

%PD-1+/CCR7- CD8 T	0.2403	-0.0030	-0.0054	-0.0013	-0.0074	-0.0074	0.0237	-0.0210	0.1407
%PD-1-/CCR7- CD8 T	-0.0454	-0.0028	0.0204	0.0024	0.0066	-0.0109	-0.0040	-0.0212	-0.0134
% NK	0.1107	-0.0083	-0.0085	0.1207	-0.0054	0.0073	-0.0086	-0.0253	-0.0060
%NKG2A+ NK	0.0942	0.0044	-0.0062	0.1548	-0.0087	-0.0072	0.0137	-0.0273	-0.0137
%NKG2A-/CD16+ NK	0.0842	0.0300	-0.0083	0.1167	-0.0092	0.0051	0.0283	-0.0278	-0.0092
%NKG2A+/CD16+ NK	-0.0026	-0.0104	0.0060	0.0601	-0.0102	-0.0098	-0.0103	-0.0262	-0.0294
%NKG2A-/CD16- NK	0.1266	0.0246	-0.0084	0.1358	-0.0060	-0.0001	0.0488	-0.0259	0.0025
%NKG2A-/CD16- NK	-0.0232	0.0279	-0.0089	-0.0059	-0.0104	0.0185	0.0020	-0.0312	-0.0279
% NKT	-0.0088	-0.0007	0.0024	-0.0104	-0.0104	-0.0101	0.0002	-0.0285	0.0641
%NKG2A+ NKT	-0.0352	-0.0106	-0.0107	0.0026	0.0216	-0.0104	-0.0093	-0.0313	-0.0180
%NKG2A-/CD16+ NKT	0.0736	0.1239	-0.0076	0.0219	-0.0087	0.0019	-0.0061	-0.0281	-0.0046
%NKG2A+/CD16+ NKT	0.0377	0.0072	-0.0100	0.0558	0.0081	-0.0078	0.0143	0.0130	0.0013
%NKG2A+/CD16- NKT	-0.0417	-0.0108	-0.0108	-0.0042	0.0180	-0.0101	-0.0065	-0.0317	-0.0200
%NKG2A-/CD16- NKT	0.0807	0.0598	-0.0082	0.0382	0.0197	-0.0062	-0.0095	-0.0281	0.0111
%CD38-/HLA-DR+ Tcm CD4 T	-0.0175	-0.0041	-0.0095	0.0322	0.0003	0.0052	-0.0102	-0.0230	0.0030
%CD38+/HLA-DR+ Tcm CD4 T	0.0821	0.0163	0.0020	-0.0016	-0.0032	-0.0053	0.0767	-0.0275	-0.0209
%CD38-/HLA-DR- Tcm CD4 T	0.0802	-0.0079	0.0000	0.0000	-0.0096	0.0423	0.0252	-0.0277	0.0702
%CD38-/HLA-DR- Tcm CD4 T	0.1225	-0.0085	0.0037	-0.0091	-0.0088	0.0312	0.0615	-0.0266	0.0469
%PD-1+ Tcm CD4 T	-0.0079	0.0028	-0.0089	-0.0095	-0.0066	0.0097	-0.0009	-0.0302	0.0134
%CD38-/HLA-DR+ Tn CD4 T	-0.0677	-0.0061	-0.0101	0.0111	-0.0082	-0.0011	-0.0093	-0.0233	-0.0203
%CD38+/HLA-DR+ Tn CD4 T	-0.0540	-0.0099	-0.0109	-0.0110	-0.0086	-0.0109	0.0375	-0.0319	-0.0141
%CD38+/HLA-DR- Tn CD4 T	0.0214	0.0031	-0.0102	0.0379	-0.0078	0.0143	0.0196	-0.0272	0.0123
%CD38-/HLA-DR- Tn CD4 T	0.0586	0.0084	-0.0098	0.0424	-0.0082	0.0182	0.0772	-0.0258	-0.0053
%PD-1+ Tn CD4 T	0.0467	0.0491	-0.0099	0.0184	-0.0099	0.0165	-0.0083	-0.0292	0.0387
%CD38-/HLA-DR+ Teff CD4 T	0.0272	-0.0101	0.0198	0.0026	-0.0101	-0.0005	0.0103	-0.0116	0.0372
%CD38+/HLA-DR+ Teff CD4 T	0.1986	0.0143	0.0492	-0.0079	0.0035	0.0162	0.0636	-0.0224	0.0374
%CD38+/HLA-DR- Teff CD4 T	0.1137	-0.0090	0.0429	-0.0071	-0.0092	-0.0090	0.0052	0.0009	0.0979
%CD38-/HLA-DR- Teff CD4 T	0.1298	0.0006	0.0432	-0.0085	-0.0038	0.0201	0.0284	-0.0266	0.0369
%PD-1+ Teff CD4 T	0.0117	-0.0058	-0.0057	0.0574	0.0128	-0.0058	0.0030	-0.0228	-0.0141
%CD38-/HLA-DR+ Tem CD4 T	0.0289	0.0205	-0.0024	0.0059	0.0088	0.0131	0.0100	0.0113	0.0171
%CD38+/HLA-DR+ Tem CD4 T	0.1995	0.2090	-0.0082	0.0038	-0.0067	-0.0076	0.0135	-0.0242	-0.0069
%CD38+/HLA-DR- Tem CD4 T	-0.0349	-0.0096	-0.0079	-0.0091	-0.0108	0.0346	0.0205	0.0004	-0.0311
%CD38-/HLA-DR- Tem CD4 T	0.1010	0.1337	-0.0040	0.0098	-0.0038	-0.0069	-0.0073	-0.0225	0.0213
%PD-1+ Tem CD4 T	-0.1078	-0.0082	-0.0037	-0.0110	-0.0098	-0.0115	-0.0115	-0.0339	-0.0261
%CD38-/HLA-DR+ Tcm CD8 T	-0.0815	-0.0049	-0.0080	-0.0081	-0.0098	-0.0108	-0.0077	-0.0319	-0.0244
%CD38+/HLA-DR+ Tcm CD8 T	0.1044	0.0684	0.0211	-0.0048	-0.0076	0.0011	0.0311	-0.0231	-0.0173
%CD38+/HLA-DR- Tcm CD8 T	0.0941	0.0004	0.0356	0.0058	-0.0094	0.0575	0.0310	-0.0242	-0.0261
%CD38-/HLA-DR- Tcm CD8 T	0.1129	0.0431	0.0318	-0.0059	-0.0092	0.0097	0.0395	-0.0257	-0.0267
%PD-1+ Tcm CD8 T	-0.0122	0.0017	-0.0069	0.0008	-0.0094	-0.0105	0.0067	-0.0310	-0.0108
%CD38-/HLA-DR+ Tn CD8 T	0.0043	-0.0012	0.0267	-0.0018	-0.0102	0.0061	-0.0080	-0.0279	-0.0247
%CD38+/HLA-DR+ Tn CD8 T	0.0255	0.0659	-0.0027	-0.0084	-0.0016	-0.0064	0.0194	-0.0298	-0.0256
%CD38+/HLA-DR- Tn CD8 T	0.2663	-0.0007	0.0226	0.0331	-0.0027	0.0361	0.0704	-0.0198	0.0304
%CD38-/HLA-DR- Tn CD8 T	0.2331	-0.0066	0.0105	0.0186	-0.0055	0.0280	0.1271	-0.0224	0.0265
%PD-1+ Tn CD8 T	0.0427	0.0203	0.0282	0.0225	-0.0098	-0.0017	-0.0028	-0.0286	-0.0213
%CD38-/HLA-DR+ Teff CD8 T	0.0641	-0.0072	0.0038	-0.0047	0.0135	-0.0087	-0.0087	0.0001	0.0294
%CD38+/HLA-DR+ Teff CD8 T	0.0110	0.0149	-0.0041	0.0009	-0.0099	-0.0027	0.0329	-0.0288	-0.0059
%CD38+/HLA-DR- Teff CD8 T	0.1132	0.0047	0.0082	-0.0068	0.0047	0.0101	0.0167	0.0052	0.0250
%CD38-/HLA-DR- Teff CD8 T	0.0174	0.0143	-0.0076	0.0188	-0.0067	0.0074	0.0742	-0.0300	-0.0289
%PD-1+ Teff CD8 T	0.0040	-0.0088	-0.0069	-0.0085	-0.0093	-0.0100	0.0010	-0.0264	0.0573
%CD38-/HLA-DR+ Tem CD8 T	-0.0032	-0.0103	0.0026	-0.0010	-0.0027	-0.0097	-0.0104	0.0047	-0.0138
%CD38+/HLA-DR+ Tem CD8 T	0.0310	0.0722	-0.0042	-0.0098	-0.0066	-0.0081	0.0253	-0.0295	-0.0280
%CD38+/HLA-DR- Tem CD8 T	0.0251	-0.0098	-0.0067	-0.0101	0.0021	-0.0034	0.0204	-0.0021	-0.0033
%CD38-/HLA-DR- Tem CD8 T	-0.0031	0.0386	-0.0104	-0.0062	-0.0058	-0.0026	0.0343	-0.0237	-0.0234
%PD-1+ Tem CD8 T	-0.0347	0.0193	-0.0098	0.0059	-0.0106	-0.0104	-0.0103	-0.0317	-0.0171
%IL-2+ CD8 T	-0.0207	0.0277	-0.0100	-0.0077	0.0064	-0.0098	-0.0081	-0.0274	-0.0024
%IL-2+ CD4 T	0.0228	0.0403	-0.0076	0.0109	0.0046	-0.0080	-0.0091	-0.0078	-0.0203

Table S7 (continued)

%TNFa+ CD4 T	-0.0307	-0.0063	-0.0103	0.0120	0.0197	-0.0083	0.0039	-0.0104	-0.0183
%TNFa+ CD8 T	-0.0743	-0.0064	-0.0072	-0.0061	0.0091	0.0008	0.0087	-0.0278	-0.0323
%IFNg+ CD4 T	-0.0703	-0.0060	-0.0092	-0.0047	-0.0097	-0.0104	-0.0061	-0.0230	-0.0165
%IFNg+ CD8 T	-0.0946	-0.0096	-0.0107	-0.0102	-0.0065	-0.0084	0.0065	-0.0270	-0.0252
%CD107+ CD4 T	-0.0808	-0.0104	-0.0076	-0.0113	-0.0041	-0.0106	0.0089	-0.0255	-0.0268
%CD107+ CD8 T	-0.0976	-0.0112	-0.0095	-0.0053	-0.0041	-0.0069	0.0060	-0.0335	-0.0322
%CD107+IFNg+IL-2-TNFa- CD4 T	-0.0843	-0.0113	-0.0071	-0.0100	-0.0101	-0.0112	0.0000	-0.0267	-0.0263
%CD107+IFNg-IL-2-TNF- CD4 T	-0.0663	-0.0026	-0.0104	-0.0109	-0.0104	-0.0082	0.0010	-0.0136	-0.0260
%CD107+IFNg-IL-2-TNFa+ CD4 T	-0.0674	-0.0020	-0.0111	-0.0098	0.0072	0.0152	-0.0012	-0.0144	-0.0295
%CD107+IFNg-IL-2-TNFa- CD4 T	-0.0425	-0.0103	-0.0017	-0.0104	0.0125	-0.0101	0.0360	-0.0318	-0.0245
%CD107-IFNg+IL-2-TNFa- CD4 T	-0.0007	0.0366	0.0012	-0.0003	-0.0091	-0.0088	0.0181	-0.0210	-0.0172
%CD107-IFNg+IL-2-TNFa+ CD4 T	-0.0467	-0.0098	-0.0043	0.0321	-0.0071	-0.0027	-0.0007	-0.0319	-0.0240
%CD107-IFNg+IL-2-TNFa- CD4 T	-0.0123	-0.0084	-0.0094	0.0095	-0.0029	-0.0105	-0.0013	-0.0309	-0.0085
%CD107-IFNg-IL-2-TNFa+ CD4 T	0.1650	0.0176	-0.0047	0.0371	0.0844	0.0012	0.0011	-0.0069	-0.0218
%CD107-IFNg-IL-2-TNFa- CD4 T	0.0343	0.0558	-0.0033	0.0125	-0.0101	-0.0037	-0.0074	-0.0144	-0.0231
%CD107-IFNg-IL-2-TNFa+ CD4 T	-0.0145	-0.0025	-0.0100	0.0061	0.0416	-0.0017	-0.0028	-0.0099	-0.0158
%CD107+IFNg+IL-2-TNFa- CD8T	-0.0959	-0.0114	-0.0114	-0.0111	-0.0113	-0.0109	0.0037	-0.0294	-0.0233
%CD107+IFNg-IL-2-TNF- CD8 T	-0.0124	0.0022	0.0118	-0.0105	0.0111	-0.0030	-0.0096	-0.0272	-0.0069
%CD107+IFNg-IL-2-TNFa+ CD8 T	-0.0303	-0.0059	-0.0016	0.0103	-0.0005	0.0175	-0.0011	-0.0303	-0.0198
%CD107+IFNg-IL-2-TNFa- CD8 T	-0.0415	-0.0090	-0.0072	0.0084	0.0022	-0.0081	0.0035	-0.0151	-0.0298
%CD107-IFNg+IL-2-TNFa- CD8 T	-0.0309	0.0060	0.0086	-0.0063	0.0005	-0.0087	-0.0003	-0.0196	-0.0055
%CD107-IFNg+IL-2-TNFa+ CD8 T	-0.0086	-0.0086	0.0011	-0.0094	0.0433	-0.0016	0.0242	-0.0289	-0.0285
%CD107-IFNg+IL-2-TNFa- CD8 T	-0.0450	-0.0021	-0.0109	-0.0104	-0.0083	-0.0106	0.0027	-0.0287	0.0147
%CD107-IFNg-IL-2-TNFa+ CD8 T	0.0301	-0.0083	-0.0065	0.0088	0.0963	0.0004	0.0166	-0.0265	-0.0053
%CD107-IFNg-IL-2-TNFa- CD8 T	-0.0049	0.0488	-0.0099	-0.0094	-0.0104	-0.0097	-0.0096	-0.0287	-0.0089
%CD107-IFNg-IL-2-TNFa+ CD8 T	0.0054	0.0108	-0.0089	-0.0097	0.0188	-0.0093	-0.0102	-0.0214	-0.0187

Table S7 (continued)

Immunophenotype \ Model	Adj. R^2	ΔR^2 , difference in adjustable R^2 after dropping a covariate							
	Include all	% intact HIV DNA	Age	Sex	Years of ART	CD4 nadir	Recent CD4 count	Years before ART	Race
%CD4 T	0.2110	-0.0030	-0.0037	0.0187	0.0119	0.0220	0.1593	-0.0141	-0.0116
%CD27+ CD4 T	0.1079	0.0098	0.0032	0.0209	-0.0054	0.0002	0.0401	-0.0046	0.0630
%CD38+ CD4 T	0.1428	0.0447	0.0499	0.0169	-0.0072	0.0137	-0.0031	-0.0260	0.0174
%CD127+ CD4 T	0.0716	0.0888	0.0268	0.0014	-0.0004	-0.0096	0.0059	-0.0159	-0.0171
%HLA-DR+ CD4 T	0.0959	-0.0087	-0.0057	0.0532	-0.0093	0.0034	0.0388	-0.0066	0.0241
%KLRG1+ CD4 T	0.0783	0.0101	-0.0056	-0.0061	-0.0095	-0.0057	0.0054	-0.0013	0.0253
%NKG2A+ CD4 T	0.0201	-0.0061	0.0138	0.0068	-0.0092	0.0105	0.0027	-0.0164	-0.0055
%PD-1+ CD4 T	0.1005	0.0134	0.0105	0.0206	-0.0062	-0.0003	0.0337	-0.0275	0.0010
%Tcm CD4 T	0.0193	-0.0100	0.0202	-0.0079	-0.0082	-0.0004	-0.0101	-0.0129	-0.0192
%Tn CD4 T	0.0930	0.0014	0.0447	0.0126	-0.0083	-0.0081	0.0349	-0.0277	-0.0169
%Teff CD4 T	-0.0314	-0.0029	-0.0049	-0.0021	-0.0081	0.0025	-0.0004	-0.0220	0.0088
%Tem CD4 T	0.0634	0.0095	-0.0043	0.0037	-0.0096	-0.0092	0.0593	-0.0121	0.0118
%CD38-/HLA-DR+ CD4 T	0.1082	-0.0083	0.0094	0.0650	-0.0027	0.0133	0.0016	0.0126	0.0426
%CD38+/HLA-DR+ CD4 T	0.0474	-0.0030	-0.0085	0.0154	0.0022	-0.0081	0.0714	-0.0268	-0.0129
%CD38-/HLA-DR- CD4 T	0.1674	0.0317	0.0444	0.0388	-0.0087	0.0164	-0.0081	-0.0243	0.0350
%KLRG1-/CD27+ CD4 T	0.0690	0.0162	-0.0066	-0.0096	-0.0093	-0.0033	0.0265	-0.0100	-0.0133
%KLRG1+/CD27+ CD4 T	0.2192	-0.0037	-0.0062	0.0233	-0.0062	0.0366	-0.0080	-0.0234	0.2033
%KLRG1+/CD27- CD4 T	0.0552	0.0101	0.0080	0.0028	-0.0090	0.0077	0.0156	0.0134	0.0068
%KLRG1-/CD27- CD4 T	0.1606	-0.0081	-0.0077	0.0316	0.0021	-0.0040	0.0390	-0.0143	0.1081
%KLRG1+/PD-1+ CD4 T	0.0408	-0.0092	0.0046	0.0078	-0.0084	-0.0006	0.0106	-0.0032	0.0417
%KLRG1+/PD-1- CD4 T	0.0291	0.0166	-0.0078	-0.0041	-0.0092	-0.0099	0.0025	0.0010	-0.0208
%KLRG1-/PD-1- CD4 T	0.1291	-0.0047	-0.0054	0.0357	-0.0057	0.0014	0.0000	-0.0173	0.0951
%KLRG1-/PD-1- CD4 T	0.1413	0.0188	0.0152	-0.0073	-0.0086	0.0100	0.0456	-0.0242	-0.0250
%PD-1-/CCR7+ CD4 T	0.1133	0.0168	0.0212	0.0138	-0.0080	-0.0057	0.0513	-0.0267	0.0129
%PD-1+/CCR7+ CD4 T	0.0476	-0.0057	0.0153	-0.0044	-0.0057	0.0193	-0.0057	-0.0094	-0.0206
%PD-1+/CCR7- CD4 T	0.0502	0.0109	-0.0085	0.0177	-0.0097	-0.0084	0.0405	-0.0118	0.0038
%PD-1-/CCR7- CD4 T	-0.0034	0.0000	0.0218	-0.0103	-0.0068	-0.0020	0.0389	-0.0209	0.0045
%CD8 T	0.1876	-0.0067	-0.0058	0.0155	-0.0005	0.0192	0.1499	-0.0159	-0.0113
%CD27+ CD8 T	0.0466	-0.0099	0.0303	0.0067	-0.0079	-0.0092	0.0195	-0.0103	0.0142
%CD38+ CD8 T	0.2689	0.0078	0.0702	0.0041	-0.0076	0.0222	0.0787	-0.0088	-0.0163
%CD127+ CD8 T	0.0614	0.0614	-0.0091	-0.0004	0.0194	-0.0077	-0.0016	0.0034	0.0164
%HLA-DR+ CD8 T	-0.0054	-0.0004	0.0078	-0.0089	-0.0069	-0.0102	0.0340	-0.0130	-0.0257
%KLRG1+ CD8 T	0.2160	-0.0064	0.0557	0.0054	-0.0076	-0.0026	-0.0021	-0.0163	0.0425
%PD-1+ CD8 T	0.2212	-0.0072	0.0185	0.0096	-0.0037	-0.0049	-0.0017	-0.0220	0.0786
%Tn CD8 T	0.2964	-0.0063	0.1148	0.0114	-0.0073	-0.0071	0.0100	-0.0039	0.0309
%Teff CD8 T	-0.0163	-0.0101	0.0189	0.0226	0.0111	-0.0103	-0.0048	-0.0252	0.0102
%Tem CD8 T	0.2718	-0.0074	-0.0075	-0.0062	-0.0018	-0.0075	0.0116	-0.0184	0.2452
%CD38-/HLA-DR+ CD8 T	0.0034	-0.0102	0.0376	-0.0103	-0.0102	-0.0098	-0.0021	0.0032	-0.0249
%CD38+/HLA-DR+ CD8 T	0.0175	0.0133	-0.0102	-0.0044	-0.0029	-0.0076	0.0550	-0.0290	-0.0223
%CD38-/HLA-DR- CD8 T	0.3145	-0.0059	0.0932	0.0257	-0.0042	0.0205	0.0198	0.0042	-0.0007
%KLRG1-/CD27+ CD8 T	0.2101	-0.0073	0.0497	0.0089	-0.0071	-0.0028	0.0034	-0.0143	0.0393
%KLRG1+/CD27+ CD8 T	0.3146	-0.0034	0.0055	-0.0064	-0.0071	0.0200	-0.0017	-0.0200	0.2143
%KLRG1+/CD27- CD8 T	0.0270	-0.0101	0.0382	0.0042	-0.0090	-0.0094	0.0103	-0.0139	-0.0088
%KLRG1-/CD27- CD8 T	0.0794	-0.0066	-0.0003	-0.0060	-0.0073	-0.0095	0.0174	-0.0255	0.0810
%KLRG1-/PD-1+ CD8 T	-0.0295	-0.0101	0.0008	-0.0076	-0.0069	-0.0096	-0.0106	-0.0310	0.0404
%KLRG1+/PD-1+ CD8 T	0.2978	-0.0055	0.0113	0.0081	-0.0050	-0.0051	-0.0001	-0.0200	0.1414
%KLRG1+/PD-1- CD8 T	-0.0427	-0.0106	0.0403	-0.0089	-0.0042	-0.0072	-0.0100	-0.0244	-0.0313
%Tcm CD8 T	0.2086	-0.0082	0.0980	-0.0082	0.0093	-0.0065	-0.0056	-0.0229	-0.0179
%KLRG1-/PD-1- CD8 T	0.2640	-0.0063	0.0811	0.0117	-0.0076	0.0002	-0.0004	-0.0125	0.0326

Table S8

Adjusted R^2 scores and differences in adjusted R^2 for LOCO analysis for the model that contains percent intact.

The first column contains all immunophenotypes and is a target for the regression model. The second column contains adjusted R^2 of the linear regression model with features such as percent intact, age, biological sex, race, years of treatment, CD4 nadir, recent CD4 count, and years of HIV before treatment ($=NA$, < 1 , ≥ 1). The next seven columns show differences in adjusted R^2 score (ΔR^2) after removing a specific feature. Participants with missing years of ART values are excluded from this analysis. The missing value of the CD4 nadir for one participant is imputed.

%PD-1-/CCR7+ CD8 T	0.1925	-0.0075	0.0609	0.0288	-0.0063	-0.0062	0.0059	-0.0107	0.0154
%PD-1+/CCR7+ CD8 T	0.1092	-0.0093	0.0479	0.0077	-0.0039	0.0074	0.0038	-0.0271	-0.0202
%PD-1-/CCR7- CD8 T	0.2370	-0.0063	-0.0049	0.0002	-0.0065	-0.0077	0.0271	-0.0207	0.1373
%PD-1+/CCR7- CD8 T	-0.0536	-0.0109	0.0178	0.0011	0.0060	-0.0110	-0.0062	-0.0207	-0.0151
% NK	0.1105	-0.0085	-0.0086	0.1232	-0.0043	0.0078	-0.0087	-0.0258	-0.0073
%NKG2A+ NK	0.0882	-0.0015	-0.0075	0.1503	-0.0075	-0.0092	0.0070	-0.0259	-0.0181
%NKG2A-/CD16+ NK	0.0570	0.0028	-0.0096	0.1143	-0.0083	-0.0034	0.0147	-0.0265	-0.0179
%NKG2A+/CD16+ NK	0.0380	0.0302	0.0055	0.0474	-0.0080	-0.0043	-0.0094	-0.0176	-0.0293
%NKG2A-/CD16- NK	0.0943	-0.0077	-0.0077	0.1467	-0.0086	-0.0004	0.0378	-0.0269	-0.0054
%NKG2A-/CD16- NK	-0.0583	-0.0072	-0.0073	-0.0070	-0.0109	0.0089	-0.0049	-0.0322	-0.0316
% NKT	-0.0004	0.0077	0.0053	-0.0104	-0.0083	-0.0069	0.0066	-0.0305	0.0496
%NKG2A+ NKT	-0.0329	-0.0083	-0.0107	0.0041	0.0134	-0.0107	-0.0094	-0.0308	-0.0168
%NKG2A-/CD16+ NKT	-0.0613	-0.0109	-0.0109	0.0282	-0.0110	-0.0035	-0.0110	-0.0310	-0.0169
%NKG2A+/CD16+ NKT	0.0217	-0.0088	-0.0098	0.0569	0.0070	-0.0093	0.0078	0.0048	0.0043
%NKG2A-/CD16- NKT	-0.0382	-0.0073	-0.0107	-0.0031	0.0101	-0.0107	-0.0071	-0.0318	-0.0183
%NKG2A-/CD16- NKT	0.0118	-0.0092	-0.0100	0.0464	0.0081	-0.0073	-0.0094	-0.0302	0.0046
%CD38-/HLA-DR+ Tcm CD4 T	-0.0226	-0.0093	-0.0090	0.0320	0.0005	0.0012	-0.0094	-0.0255	-0.0004
%CD38+/HLA-DR+ Tcm CD4 T	0.0719	0.0061	-0.0002	0.0033	0.0060	-0.0083	0.0876	-0.0266	-0.0242
%CD38+/HLA-DR- Tcm CD4 T	0.0986	0.0105	0.0016	-0.0025	-0.0077	0.0236	0.0199	-0.0275	0.0589
%CD38-/HLA-DR- Tcm CD4 T	0.1477	0.0167	0.0044	-0.0085	-0.0079	0.0146	0.0587	-0.0236	0.0370
%PD-1+ Tcm CD4 T	-0.0158	-0.0051	-0.0079	-0.0097	-0.0086	0.0023	0.0045	-0.0311	0.0068
%CD38-/HLA-DR+ Tn CD4 T	-0.0586	0.0029	-0.0105	0.0082	-0.0041	-0.0070	-0.0105	-0.0282	-0.0237
%CD38+/HLA-DR+ Tn CD4 T	-0.0486	-0.0045	-0.0109	-0.0107	-0.0105	-0.0101	0.0389	-0.0310	-0.0156
%CD38+/HLA-DR- Tn CD4 T	0.0197	0.0014	-0.0101	0.0349	-0.0097	0.0039	0.0117	-0.0295	0.0029
%CD38-/HLA-DR- Tn CD4 T	0.0445	-0.0056	-0.0097	0.0420	-0.0093	0.0094	0.0639	-0.0282	-0.0120
%PD-1+ Tn CD4 T	-0.0061	-0.0037	-0.0100	0.0194	-0.0103	0.0054	-0.0105	-0.0296	0.0223
%CD38-/HLA-DR+ Teff CD4 T	0.0283	-0.0090	0.0202	0.0015	-0.0099	-0.0027	0.0102	-0.0142	0.0342
%CD38+/HLA-DR+ Teff CD4 T	0.2131	0.0288	0.0458	-0.0081	0.0215	0.0301	0.0699	-0.0178	0.0344
%CD38+/HLA-DR- Teff CD4 T	0.1154	-0.0072	0.0444	-0.0075	-0.0092	-0.0092	0.0040	-0.0037	0.0926
%CD38-/HLA-DR- Teff CD4 T	0.1441	0.0149	0.0415	-0.0065	0.0062	0.0316	0.0307	-0.0252	0.0311
%PD-1+ Teff CD4 T	0.0080	-0.0094	-0.0064	0.0527	0.0149	-0.0054	0.0052	-0.0227	-0.0145
%CD38-/HLA-DR+ Tem CD4 T	0.0130	0.0045	-0.0003	0.0127	-0.0035	0.0193	0.0054	0.0162	0.0199
%CD38+/HLA-DR+ Tem CD4 T	0.0081	0.0175	-0.0098	0.0149	0.0040	-0.0077	0.0333	-0.0303	-0.0089
%CD38+/HLA-DR- Tem CD4 T	-0.0345	-0.0093	-0.0081	-0.0085	-0.0105	0.0353	0.0202	0.0014	-0.0308
%CD38-/HLA-DR- Tem CD4 T	-0.0151	0.0177	-0.0002	0.0223	-0.0106	-0.0047	-0.0022	-0.0236	0.0215
%PD-1+ Tem CD4 T	-0.1108	-0.0111	-0.0045	-0.0107	-0.0090	-0.0115	-0.0116	-0.0340	-0.0273
%CD38-/HLA-DR+ Tcm CD8 T	-0.0865	-0.0099	-0.0087	-0.0070	-0.0109	-0.0111	-0.0064	-0.0319	-0.0226
%CD38+/HLA-DR+ Tcm CD8 T	0.0358	-0.0002	0.0136	-0.0008	-0.0022	-0.0030	0.0470	-0.0263	-0.0244
%CD38+/HLA-DR- Tcm CD8 T	0.0853	-0.0084	0.0398	0.0060	-0.0094	0.0467	0.0247	-0.0259	-0.0265
%CD38-/HLA-DR- Tcm CD8 T	0.0702	0.0004	0.0250	-0.0026	-0.0069	0.0038	0.0533	-0.0278	-0.0280
%PD-1+ Tcm CD8 T	-0.0203	-0.0065	-0.0078	0.0040	-0.0106	-0.0104	0.0103	-0.0311	-0.0141
%CD38-/HLA-DR+ Tn CD8 T	-0.0041	-0.0096	0.0303	-0.0015	-0.0102	0.0023	-0.0061	-0.0292	-0.0269
%CD38+/HLA-DR+ Tn CD8 T	-0.0474	-0.0069	-0.0067	-0.0065	0.0049	-0.0080	0.0339	-0.0320	-0.0311
%CD38+/HLA-DR- Tn CD8 T	0.2610	-0.0061	0.0256	0.0327	-0.0041	0.0278	0.0630	-0.0212	0.0289
%CD38-/HLA-DR- Tn CD8 T	0.2343	-0.0054	0.0101	0.0157	-0.0071	0.0214	0.1311	-0.0229	0.0292
%PD-1+ Tn CD8 T	0.0124	-0.0100	0.0344	0.0251	-0.0102	-0.0044	-0.0070	-0.0301	-0.0259
%CD38-/HLA-DR+ Teff CD8 T	0.0652	-0.0060	0.0033	-0.0031	0.0056	-0.0094	-0.0089	-0.0046	0.0260
%CD38+/HLA-DR+ Teff CD8 T	0.0166	0.0205	-0.0054	-0.0047	-0.0035	-0.0082	0.0387	-0.0301	-0.0058
%CD38+/HLA-DR- Teff CD8 T	0.1050	-0.0035	0.0062	-0.0083	-0.0027	0.0049	0.0213	-0.0001	0.0194
%CD38-/HLA-DR- Teff CD8 T	0.0102	0.0071	-0.0086	0.0104	0.0012	-0.0003	0.0843	-0.0302	-0.0296
%PD-1+ Teff CD8 T	0.0031	-0.0097	-0.0073	-0.0085	-0.0097	-0.0097	0.0030	-0.0256	0.0573
%CD38-/HLA-DR+ Tem CD8 T	-0.0027	-0.0098	0.0026	-0.0004	-0.0047	-0.0093	-0.0104	0.0014	-0.0147
%CD38+/HLA-DR+ Tem CD8 T	-0.0297	0.0115	-0.0076	-0.0080	0.0039	-0.0105	0.0391	-0.0307	-0.0283
%CD38+/HLA-DR- Tem CD8 T	0.0254	-0.0095	-0.0068	-0.0101	-0.0008	-0.0048	0.0212	-0.0048	-0.0044
%CD38-/HLA-DR- Tem CD8 T	-0.0409	0.0007	-0.0105	-0.0094	0.0017	-0.0064	0.0466	-0.0231	-0.0249
%PD-1+ Tem CD8 T	-0.0635	-0.0095	-0.0108	0.0097	-0.0097	-0.0108	-0.0089	-0.0326	-0.0203

Table S8 (continued)

%IL-2+ CD8 T	-0.0499	-0.0014	-0.0109	-0.0080	0.0090	-0.0109	-0.0108	-0.0313	0.0079
%IL-2+ CD4 T	-0.0078	0.0097	-0.0093	0.0203	-0.0071	-0.0102	-0.0104	-0.0037	-0.0201
%TNFa+ CD4 T	-0.0314	-0.0069	-0.0101	0.0153	0.0098	-0.0070	0.0028	-0.0084	-0.0184
%TNFa+ CD8 T	-0.0553	0.0127	-0.0081	-0.0077	0.0208	-0.0070	0.0034	-0.0312	-0.0301
%IFNg+ CD4 T	-0.0723	-0.0080	-0.0087	-0.0029	-0.0110	-0.0097	-0.0070	-0.0218	-0.0170
%IFNg+ CD8 T	-0.0793	0.0057	-0.0108	-0.0108	-0.0005	-0.0109	0.0028	-0.0307	-0.0211
%CD107+ CD4 T	-0.0816	-0.0112	-0.0073	-0.0113	-0.0053	-0.0107	0.0080	-0.0260	-0.0269
%CD107+ CD8 T	-0.0859	0.0005	-0.0097	-0.0070	0.0017	-0.0100	0.0037	-0.0330	-0.0306
%CD107+IFNg+IL-2-TNFa- CD4 T	-0.0839	-0.0109	-0.0069	-0.0098	-0.0098	-0.0111	-0.0003	-0.0277	-0.0259
%CD107+IFNg+IL-2-TNFa- CD4 T	-0.0692	-0.0055	-0.0107	-0.0103	-0.0111	-0.0098	0.0031	-0.0113	-0.0241
%CD107+IFNg+IL-2-TNFa+ CD4 T	-0.0745	-0.0091	-0.0112	-0.0089	0.0007	0.0160	-0.0032	-0.0141	-0.0293
%CD107+IFNg+IL-2-TNFa- CD4 T	-0.0426	-0.0103	-0.0021	-0.0103	0.0088	-0.0098	0.0393	-0.0319	-0.0243
%CD107+IFNg+IL-2-TNFa+ CD4 T	-0.0050	0.0323	-0.0015	-0.0067	0.0021	-0.0104	0.0251	-0.0139	-0.0216
%CD107-IFNg+IL-2-TNFa+ CD4 T	-0.0451	-0.0081	-0.0049	0.0334	-0.0089	-0.0004	0.0015	-0.0320	-0.0256
%CD107-IFNg+IL-2-TNFa- CD4 T	-0.0144	-0.0104	-0.0091	0.0097	-0.0036	-0.0106	0.0005	-0.0309	-0.0070
%CD107-IFNg+IL-2-TNFa+ CD4 T	0.1469	-0.0006	-0.0062	0.0458	0.0550	-0.0027	-0.0022	-0.0051	-0.0221
%CD107-IFNg+IL-2-TNFa- CD4 T	-0.0319	-0.0104	-0.0070	0.0173	-0.0103	-0.0040	-0.0026	-0.0193	-0.0197
%CD107-IFNg+IL-2-TNFa+ CD4 T	-0.0144	-0.0024	-0.0102	0.0020	0.0526	-0.0055	-0.0014	-0.0149	-0.0166
%CD107+IFNg+IL-2-TNFa- CD8T	-0.0858	-0.0013	-0.0113	-0.0113	-0.0098	-0.0113	0.0020	-0.0316	-0.0203
%CD107+IFNg+IL-2-TNFa- CD8 T	0.0178	0.0324	0.0137	-0.0094	0.0343	-0.0090	-0.0091	-0.0296	-0.0017
%CD107+IFNg+IL-2-TNFa+ CD8 T	-0.0309	-0.0065	-0.0029	0.0091	0.0018	0.0096	-0.0038	-0.0288	-0.0219
%CD107+IFNg+IL-2-TNFa- CD8 T	-0.0397	-0.0071	-0.0068	0.0056	0.0063	-0.0095	0.0043	-0.0129	-0.0298
%CD107-IFNg+IL-2-TNFa- CD8 T	-0.0143	0.0226	0.0049	-0.0078	0.0114	-0.0105	-0.0058	-0.0274	0.0065
%CD107-IFNg+IL-2-TNFa+ CD8 T	0.0437	0.0436	-0.0001	-0.0099	0.0737	-0.0091	0.0168	-0.0292	-0.0236
%CD107-IFNg+IL-2-TNFa- CD8 T	-0.0468	-0.0039	-0.0109	-0.0106	-0.0063	-0.0091	-0.0015	-0.0310	0.0221
%CD107-IFNg+IL-2-TNFa+ CD8 T	0.0308	-0.0076	-0.0062	0.0110	0.0759	0.0021	0.0162	-0.0255	-0.0052
%CD107-IFNg+IL-2-TNFa- CD8 T	-0.0628	-0.0092	-0.0092	-0.0082	-0.0106	-0.0104	-0.0070	-0.0302	-0.0033
%CD107-IFNg+IL-2-TNFa+ CD8 T	0.0244	0.0298	-0.0081	-0.0101	0.0444	-0.0040	-0.0101	-0.0132	-0.0230

Table S8 (continued)

Variable	Bin Total HIV DNA	Bin Intact HIV DNA	Bin % intact HIV DNA	Variable	Bin Total HIV DNA	Bin Intact HIV DNA	Bin % intact HIV DNA
%NKG2A+ CD4 T	0.6953	0.6084	0.5416	%PD-1-/CCR7+ CD8 T	0.6207	0.6134	0.5115
%PD-1+ Tn CD4 T	0.6827	0.5749	0.5313	%CD38+ CD8 T	0.6142	0.5425	0.6222
%CD38+/HLA-DR- CD8 T	0.6777	0.5797	0.6084	%CD38-/HLA-DR- Teff CD4 T	0.6122	0.5587	0.5017
%Tn CD4 T	0.6763	0.5933	0.5665	%CD38-/HLA-DR+ Tn CD8 T	0.6122	0.5420	0.5933
%KLRG1-/PD-1- CD4 T	0.6745	0.5789	0.6224	%KLRG1-/CD27+ CD8 T	0.6096	0.6101	0.5265
%Tcm CD8 T	0.6729	0.6205	0.5541	%CD38+/HLA-DR- Tn CD4 T	0.6078	0.5050	0.5531
%CD38+/HLA-DR- CD4 T	0.6700	0.5507	0.5889	%CD107-IFN γ +IL-2-TNF α - CD8 T	0.6065	0.5712	0.5133
%CD8 T	0.6629	0.6547	0.5484	%CD38+/HLA-DR+ Tem CD4 T	0.6057	0.5829	0.5941
Age	0.6627	0.6264	0.5476	%CD38+/HLA-DR- Tcm CD8 T	0.6039	0.5150	0.5750
%PD-1+/CCR7+ CD8 T	0.6608	0.5907	0.5328	%CD38-/HLA-DR+ Tcm CD8 T	0.6030	0.6053	0.5015
%PD-1+ CD8 T	0.6562	0.6207	0.5070	%KLRG1-/PD-1+ CD4 T	0.6019	0.5346	0.5290
%PD-1+ Tn CD8 T	0.6553	0.5880	0.5597	%NKG2A-/CD16+ NK	0.6012	0.5257	0.5848
%Tn CD8 T	0.6550	0.6485	0.5138	%KLRG1+ CD8 T	0.5954	0.6021	0.5251
%PD-1-/CCR7+ CD4 T	0.6537	0.5473	0.6178	%KLRG1-/CD27+ CD4 T	0.5945	0.5322	0.6205
%PD-1+/CCR7+ CD4 T	0.6475	0.5656	0.5740	%CD27+ CD4 T	0.5776	0.5018	0.6021
%Tcm CD4 T	0.6452	0.5998	0.5104	%KLRG1+/CD27- CD4 T	0.5723	0.5017	0.6181
%KLRG1-/PD-1- CD8 T	0.6440	0.6308	0.5234	%PD-1+ Tcm CD4 T	0.5705	0.5005	0.6158
%PD-1+ CD4 T	0.6434	0.5286	0.6295	%PD-1+ Tem CD4 T	0.5688	0.6124	0.5489
%CD4 T	0.6429	0.6338	0.5322	%KLRG1+ CD4 T	0.5649	0.5243	0.6236
%CD107-IFN γ +IL-2-TNF α - CD4 T	0.6420	0.5670	0.5185	Recent CD4 count	0.5620	0.6320	0.6104
%CD38+ CD4 T	0.6414	0.5299	0.5932	%KLRG1+/PD-1+ CD4 T	0.5602	0.5100	0.6450
%CD127+ CD4 T	0.6367	0.7051	0.6323	%Tem CD4 T	0.5576	0.5230	0.6145
%CD38+/HLA-DR- Tn CD8 T	0.6364	0.5558	0.5724	%IL-2+ CD4 T	0.5416	0.6022	0.5687
%PD-1-/CCR7- CD4 T	0.6340	0.6010	0.5473	%CD107-IFN γ +IL-2+TNF α - CD4 T	0.5342	0.6308	0.6248
%KLRG1+/PD-1+ CD8 T	0.6335	0.6098	0.5020	%CD38+/HLA-DR- Teff CD8 T	0.5306	0.5613	0.6276
CD4 Nadir	0.6303	0.6110	0.5563	Years before ART < 1	0.5299	0.5222	0.6092
Years of ART	0.6296	0.5069	0.7171	%CD127+ CD8 T	0.5195	0.6051	0.6048
%CD38-/HLA-DR- Tem CD4 T	0.6249	0.5513	0.5867	%CD107-IFN γ -IL-2+TNF α + CD4 T	0.5003	0.6000	0.6493
%KLRG1+/CD27+ CD8 T	0.6224	0.6473	0.5153				

Table S9

Host features classify PWH with respect to HIV reservoir characteristics.

The frequency of each of the 144 variables was analyzed for performance in classifying each PWH within the cohort as having reservoir frequency above or below the median using a receiver operating characteristics (ROC) curve. The area under the curve (AUC) is reported. The table includes variables that had an AUC value larger than 0.6 for one of the reservoir frequency characteristics. We highlight variables that had an AUC value higher than 0.65 in **bold**. Variables are ranked by AUC for the total reservoir frequency. For years of ART, AUC values are computed based on 108 PWH, excluding participants with missing years of ART values. For CD4 Nadir, AUC values are computed based on 114 PWH, excluding one participant] with the missing CD4 Nadir value.

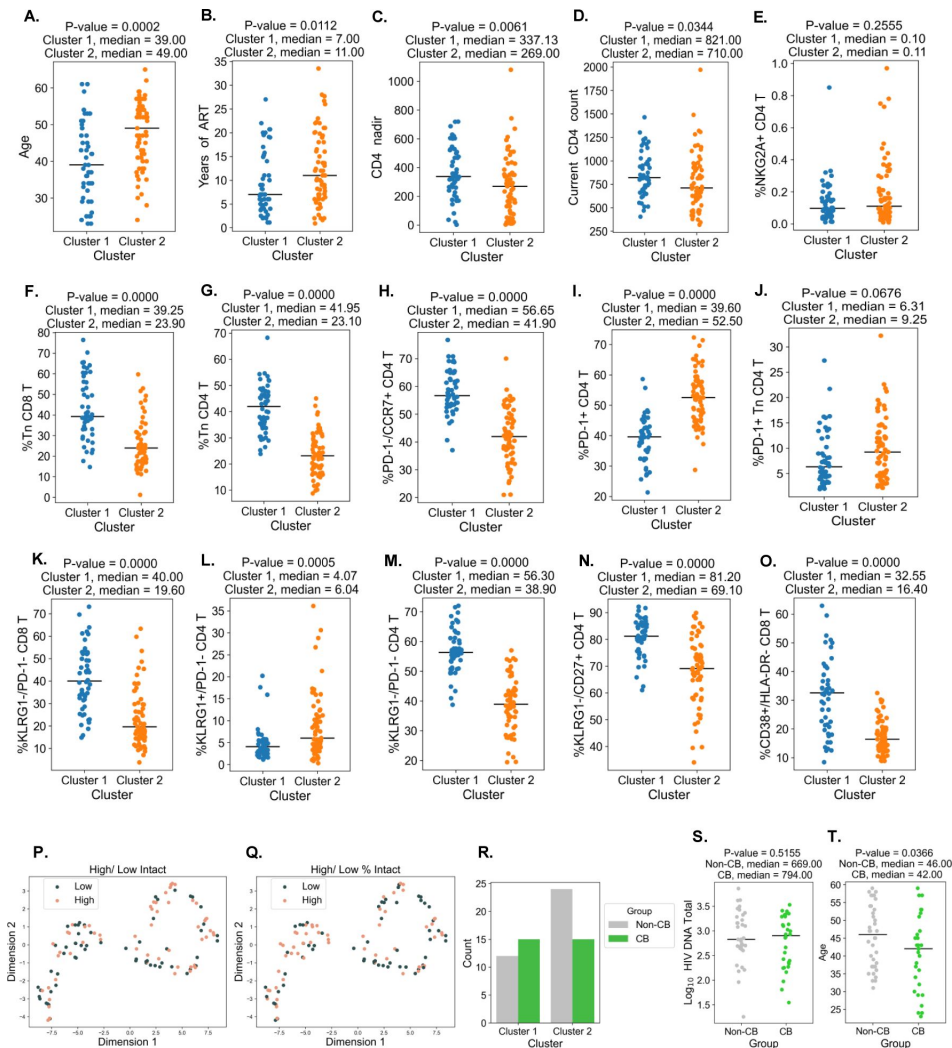


Figure S8

Dimension reduction supplemental figures.

A. Participant age is shown within each cluster. **B.** Participant years of ART are shown within each cluster. **C.** Participant CD4 nadir is shown within each cluster. **D.** Participant CD4 count is shown within each cluster. **E-O.** Participant immune features of interest are shown, where plots of immune features with similar names are placed nearby. **P.** Clusters with data points color-coded by intact reservoir frequency (high=pink, low=gray). **Q.** Clusters with data points color-coded by percent intact reservoir frequency (high=pink, low=gray). **R.** Relative proportions of cannabis (CB) users and non-users (non-CB) are shown for each cluster. **S.** Total reservoir frequencies (per million CD4 T cells) for non-users and CB users are shown. **T.** The ages of study participants for non-users and CB users are shown.

Figure S9

Left: PCA plot with data points color-coded by the membership of clusters identified in [Figure 4](#). **Right:** PCA plot with data points color-coded by total reservoir frequency (high=pink, low=gray).

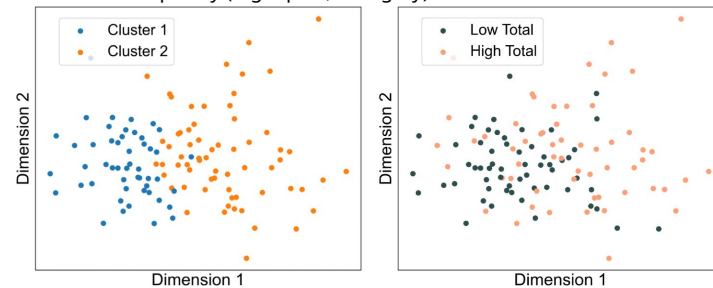
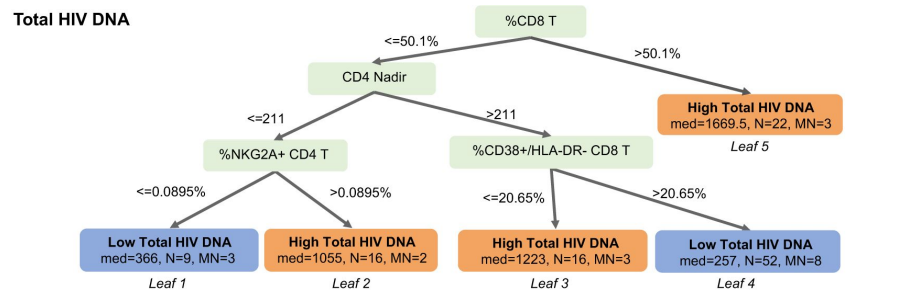


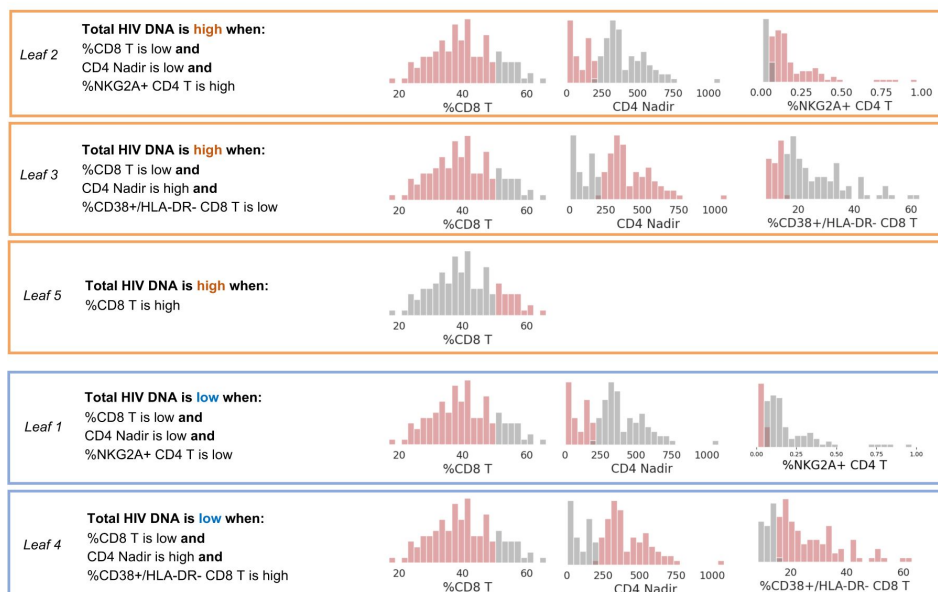
Figure S10

The total reservoir frequency visualization tree from [Figure 6](#) is explained with a set of rules.

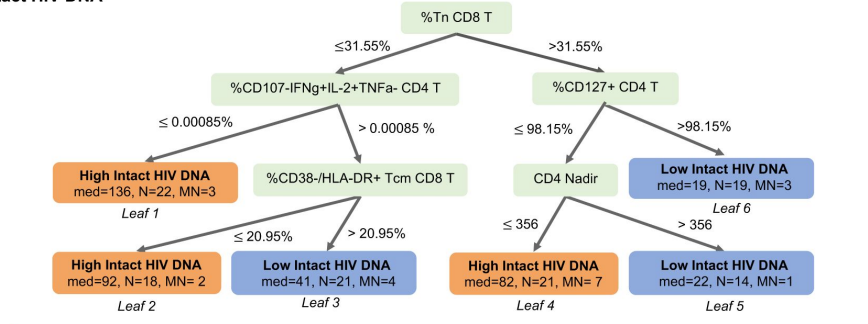
For every leaf, the path that leads to this leaf is described. The histogram of the variable and split value for every node used in the tree is shown.



To read the tree, consider each leaf and follow the path from the root to this leaf.
All features that are encountered on the path will construct a rule.



Intact HIV DNA



To read the tree, consider each leaf and follow the path from the root to this leaf.
All features that are encountered on the path will construct a rule.

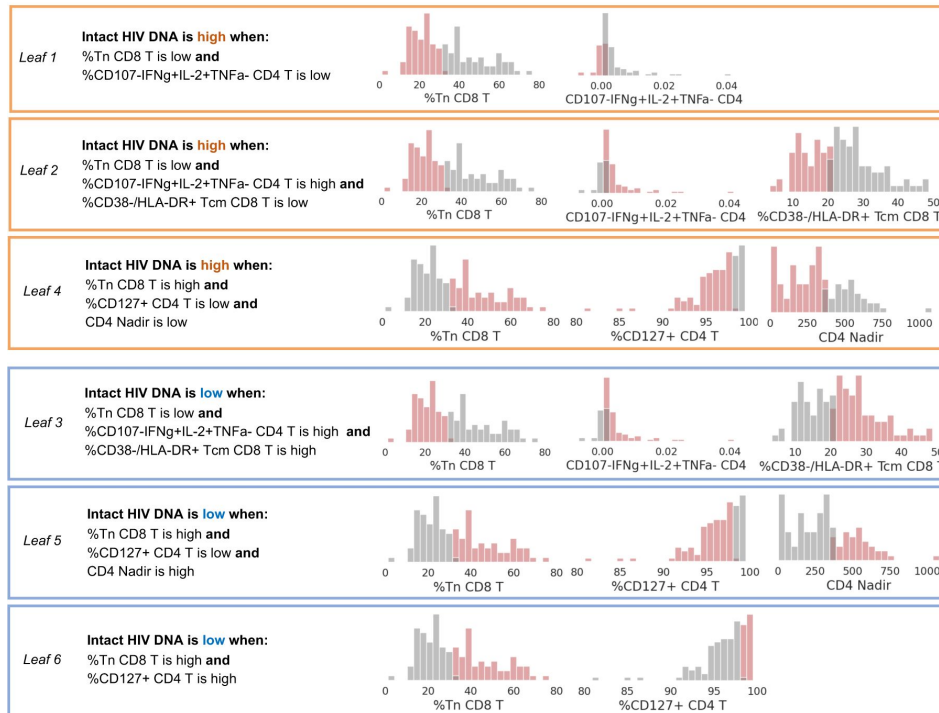
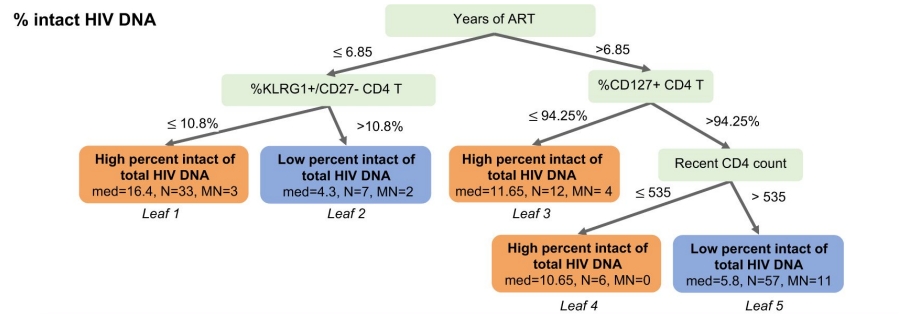


Figure S11

The intact reservoir frequency visualization tree from Figure 6 is explained with a set of rules.

For every leaf, the path that leads to this leaf is described. The histogram of the variable and split value for every node used in the tree is shown.



To read the tree, consider each leaf and follow the path from the root to this leaf.
All features that are encountered on the path will construct a rule.

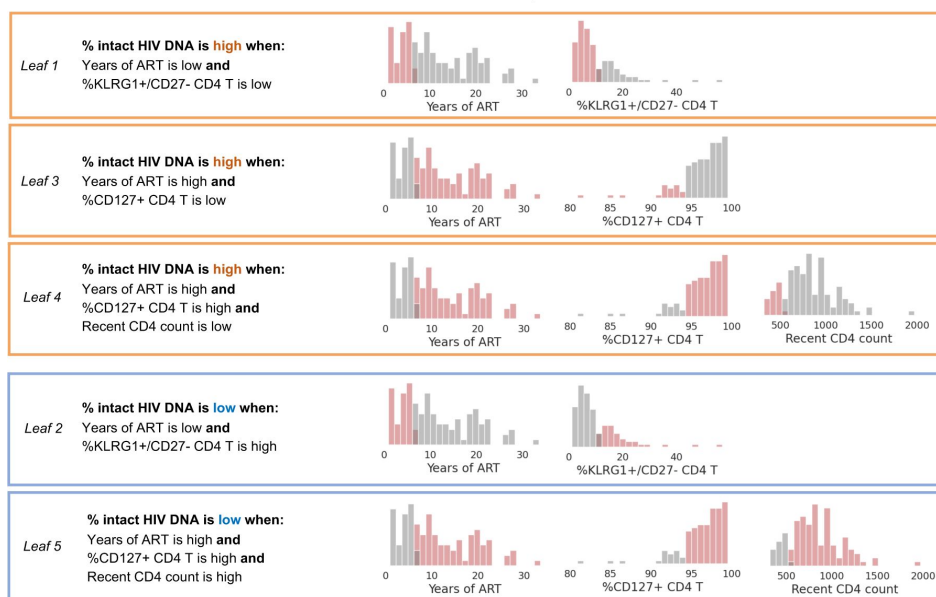


Figure S12

The percent intact visualization tree from Figure 6 is explained with a set of rules.

For every leaf, the path that leads to this leaf is described. The histogram of the variable and split value for every node used in the tree is shown.

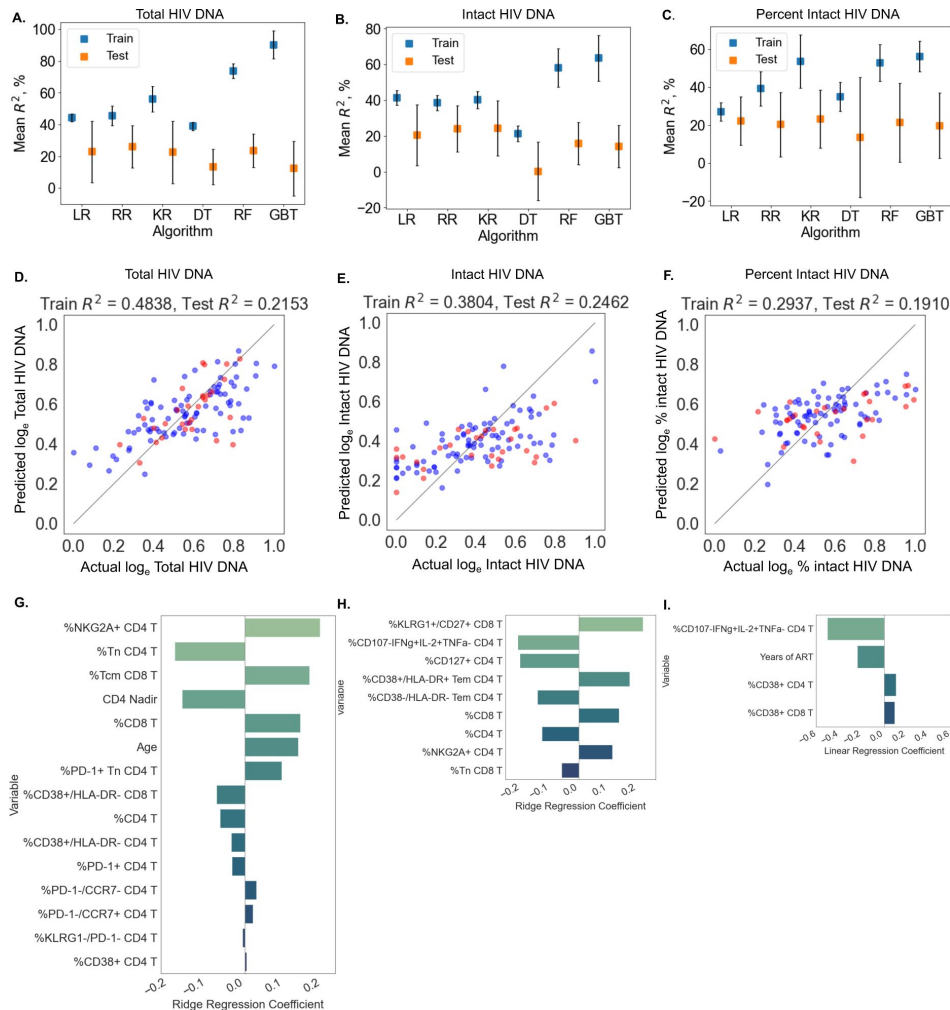


Figure S13

Using machine learning to predict reservoir frequency.

Average training and test R^2 scores over different training and test data splits for Linear Regression (LR), Ridge Regression (RR), Kernel Regression with RBF kernel (KR), Decision Tree Regressor (DT), Random Forest (RF), and Gradient Boosted Trees (GBT) models are shown for predicting total reservoir frequency (**A**), intact reservoir frequency (**B**), and percentage intact (**C**). For one split of training and test sets, ridge regression performance is shown for total reservoir frequency (**D**), intact reservoir frequency (**E**), and linear regression for percentage intact (**F**). On the x-axis, the actual values of HIV reservoir characteristics are shown, while on the y-axis the outputs of models are shown for training data (blue dots) and test data (red dots). For the same training and test split, ridge regression model feature coefficients are visualized for total reservoir (**G**), intact reservoir (**H**), and linear regression model feature coefficients are visualized for percent intact (**I**). On the y-axis we show variables used by the model, while on the x-axis coefficient values for linear models based on these variables. For percent intact models we exclude participants with missing values of years of ART. For total and intact reservoir frequency, missing values of years of ART were imputed. The missing value of the CD4 nadir for one participant was imputed as well using the MICE algorithm.

General	
Random seeds for train/test splits	0..10
Random seeds for train/valid splits	0..20
Set of the top AUC variables to consider	2..9,10,15,20,25,30,40,50
Set of the top absolute value of correlation coefficient to consider	2..9,10,15,20,25,30,40, all
Logistic Regression (classifier)	
Penalty	L2
Regularization parameter	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000
SVM with RBF kernel (classifier)	
Regularization parameter	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000
Ridge Regression (regressor)	
Regularization parameter	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000
Kernel Ridge Regression (regressor)	
Regularization parameter	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000
Kernel parameter	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000
CART (classifier) and Decision Tree (regressor)	
Maximum depth	2..7, 10
Minimum number of samples in the leaf	4, 6, 8, 10
Random Forest and Gradient Boosted Trees (classifier and regressor)	
Maximum depth	2..7, 10
Minimum number of samples in the leaf	4, 6, 8, 10
Maximum number of estimators	2, 5, 10, 20, 50, 100, 200

Table S10

Ranges of hyperparameters values that we used to perform grid search for classification and regression.

References

1. Finzi D, et al. (1997) **Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy** *Science* **278**:1295–1300
2. Chun TW, et al. (1997) **Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy** *Proc Natl Acad Sci U S A* **94**:13193–13197
3. Chun TW, et al. (1998) **Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection** *Proc Natl Acad Sci U S A* **95**:8869–8873
4. Wong JK, et al. (1997) **Recovery of replication-competent HIV despite prolonged suppression of plasma viremia** *Science* **278**:1291–1295
5. Ho YC, et al. (2013) **Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure** *Cell* **155**:540–551
6. Imamichi H, et al. (2016) **Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy** *Proc Natl Acad Sci U S A* **113**:8783–8788
7. Bruner KM, et al. (2016) **Defective proviruses rapidly accumulate during acute HIV-1 infection** *Nat Med* **22**:1043–1049
8. Maldarelli F, et al. (2014) **HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells** *Science* **345**:179–183
9. Wang Z, et al. (2018) **Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane** *Proc Natl Acad Sci U S A* **115**:E2575–E2584
10. Collora JA, et al. (2022) **Single-cell multiomics reveals persistence of HIV-1 in expanded cytotoxic T cell clones** *Immunity* **55**:1013–1031
11. Sponaugle A, et al. (2023) **Dominant CD4+ T cell receptors remain stable throughout antiretroviral therapy-mediated immune restoration in people with HIV** *Cell Rep. Med* **10**:1268
12. Hunt PW, et al. (2003) **T cell activation is associated with lower CD4+ T cell gains in human immunodeficiency virus-infected patients with sustained viral suppression during antiretroviral therapy** *J Infect Dis* **187**:1534–1543
13. Bastard JP, et al. (2012) **Circulating interleukin-6 levels correlate with residual HIV viraemia and markers of immune dysfunction in treatment-controlled HIV-infected patients** *Antivir Ther* **17**:915–919
14. Kottlilil S, et al. (2006) **Innate immune dysfunction in HIV infection: effect of HIV envelope-NK cell interactions** *J Immunol* **176**:1107–1114
15. Lichtfuss GF, et al. (2011) **Biomarkers of immune dysfunction following combination antiretroviral therapy for HIV infection** *Biomark Med* **5**:171–186

16. Prescott LM. (1995) **Loss of naive T cells may play major role in weakened immune response** *J Int Assoc Physicians AIDS Care* **1**
17. Day CL, et al. (2006) **PD-1 expression on HIV-specific T cells is associated with T-cell exhaustion and disease progression** *Nature* **443**:350–354
18. Ibegbu CC, et al. (2005) **Expression of killer cell lectin-like receptor G1 on antigen-specific human CD8+ T lymphocytes during active, latent, and resolved infection and its relation with CD57** *J Immunol* **174**:6088–6094
19. Tavenier J, et al. (2015) **Immunosenescence of the CD8(+) T cell compartment is associated with HIV-infection, but only weakly reflects age-related processes of adipose tissue, metabolism, and muscle in antiretroviral therapy-treated HIV-infected patients and controls** *BMC Immunol* **16**
20. Wang S, et al. (2020) **An atlas of immune cell exhaustion in HIV-infected individuals revealed by single-cell transcriptomics** *Emerg Microbes Infect* **9**:2333–2347
21. Hatano H, et al. (2013) **Cell-based measures of viral persistence are associated with immune activation and programmed cell death protein 1 (PD-1)-expressing CD4+ T cells** *J Infect Dis* **208**:50–56
22. Alter G, et al. (2007) **Single-stranded RNA derived from HIV-1 serves as a potent activator of NK cells** *J Immunol* **178**:7658–7666
23. Lederman MM, et al. (2013) **Residual immune dysregulation syndrome in treated HIV infection** *Adv. Immunol* **119**:51–83
24. Heather JM, et al. (2015) **Dynamic perturbations of the T-cell receptor repertoire in chronic HIV infection and following antiretroviral therapy** *Front. Immunol* **6**
25. Hunt PW, et al. (2014) **Gut epithelial barrier dysfunction and innate immune activation predict mortality in treated HIV infection** *J. Infect. Dis* **210**:1228–1238
26. Horsburgh BA, et al. (2020) **High levels of genetically intact HIV in HLA-DR+ memory T cells indicates their value for reservoir studies** *AIDS* **34**:659–668
27. Chomont N, et al. (2009) **HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation** *Nat Med* **15**:893–900
28. Cockerham LR, et al. (2014) **CD4+ and CD8+ T cell activation are associated with HIV DNA in resting CD4+ T cells** *PLoS One* **9**
29. Ruggiero A, et al. (2015) **During Stably Suppressive Antiretroviral Therapy Integrated HIV-1 DNA Load in Peripheral Blood is Associated with the Frequency of CD8 Cells Expressing HLA-DR/DP/DQ** *EBioMedicine* **2**:1153–1159
30. Lee E, et al. (2019) **Memory CD4 + T-cells expressing HLA-DR contribute to HIV persistence during prolonged antiretroviral therapy** *Front. Microbiol* **10**
31. Huang Y, et al. (2023) **Immune activation and exhaustion marker expression on T-cell subsets in ART-treated adolescents and young adults with perinatal HIV-1 infection as correlates of viral persistence** *Front. Immunol* **14**

32. Banga R, et al. (2016) **PD-1(+) and follicular helper T cells are responsible for persistent HIV-1 transcription in treated aviremic individuals** *Nat. Med* **22**:754–761
33. Fromentin R, et al. (2019) **PD-1 blockade potentiates HIV latency reversal ex vivo in CD4+ T cells from ART-suppressed individuals** *Nat. Commun* **10**
34. Fromentin R, et al. (2016) **CD4+ T cells expressing PD-1, TIGIT and LAG-3 contribute to HIV persistence during ART** *PLoS Pathog* **12**
35. Pardons M, et al. (2019) **Single-cell characterization and quantification of translation-competent viral reservoirs in treated and untreated HIV infection** *PLoS Pathog* **15**
36. Dufour C, et al. (2023) **Phenotypic characterization of single CD4+ T cells harboring genetically intact and inducible HIV genomes** *Nat. Commun* **14**
37. Zhu T, Cao W, Li T. (2023) **HIV DNA positively correlates with HLA-DR+CD8+ T lymphocytes over 8-year suppressive antiretroviral therapy** *AIDS* **37**:1335–1337
38. Wu VH, et al. (2023) **Profound phenotypic and epigenetic heterogeneity of the HIV-1-infected CD4+ T cell reservoir** *Nat. Immunol* **24**:359–370
39. Gálvez C, et al. (2021) **Atlas of the HIV-1 reservoir in peripheral CD4 T cells of individuals on successful antiretroviral therapy** *MBio* **12**
40. Bernal S, et al. (2023) **Impact of obefazimod on viral persistence, inflammation, and immune activation in people with HIV on suppressive antiretroviral therapy** *J. Infect. Dis*
41. Astorga-Gamaza A, et al. (2023) **KLRG1 expression on natural killer cells is associated with HIV persistence, and its targeting promotes the reduction of the viral reservoir** *Cell Rep. Med* **4**
42. Fisher K, et al. (2023) **Unequal distribution of genetically-intact HIV-1 proviruses in cells expressing the immune checkpoint markers PD-1 and/or CTLA-4** *Front. Immunol* **14**
43. Dubé M, et al. (2023) **Spontaneous HIV expression during suppressive ART is associated with the magnitude and function of HIV-specific CD4+ and CD8+ T cells** *Cell Host Microbe* **31**:1507–1522
44. Takata H, et al. (2023) **An active HIV reservoir during ART is associated with maintenance of HIV-specific CD8+ T cell magnitude and short-lived differentiation status** *Cell Host Microbe* **31**:1494–1506
45. Siliciano JD, Siliciano RF. (2022) **In vivo dynamics of the latent reservoir for HIV-1: New insights and implications for cure** *Annu. Rev. Pathol* **17**:271–294
46. Bruner KM, et al. (2019) **A quantitative approach for measuring the reservoir of latent HIV-1 proviruses** *Nature* **566**:120–125
47. Peluso MJ, et al. (2020) **Differential decay of intact and defective proviral DNA in HIV-1-infected individuals on suppressive antiretroviral therapy** *JCI Insight* **5**
48. Gandhi RT, et al. (2023) **Varied patterns of decay of intact human immunodeficiency virus type 1 proviruses over 2 decades of antiretroviral therapy** *J. Infect. Dis* **227**:1376–1380

49. McMyn NF, et al. (2023) **The latent reservoir of inducible, infectious HIV-1 does not decrease despite decades of antiretroviral therapy** *J. Clin. Invest* **133**
50. Wang Y, et al. (2021) **Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization** *Journal of Machine Learning Research* **22**:1–73
51. Falcinelli SD, et al. (2023) **Impact of cannabis use on immune cell populations and the viral reservoir in people with HIV on suppressive antiretroviral therapy** *The Journal of Infectious Diseases*
52. Lin J, et al. (2020) **Generalized and scalable optimal sparse decision trees** *International Conference on Machine Learning* :6150–6160
53. Poizot-Martin I, et al. (2013) **Lack of correlation between the size of HIV proviral DNA reservoir and the level of immune activation in HIV-infected patients with a sustained undetectable HIV viral load for 10 years** *J Clin Virol* **57**:351–355
54. Gandhi RT, et al. (2017) **Levels of HIV-1 persistence on antiretroviral therapy are not associated with markers of inflammation or activation** *PLoS Pathog* **13**
55. Besson GJ, et al. (2014) **HIV-1 DNA decay dynamics in blood during more than a decade of suppressive antiretroviral therapy** *Clin. Infect. Dis* **59**:1312–1321
56. Borgognone A, et al. (2022) **Gut microbiome signatures linked to HIV-1 reservoir size and viremia control** *Microbiome* **10**
57. Siegel DA, et al. (2023) **Host variation in type I interferon signaling genes (MX1), C-C chemokine receptor type 5 gene, and major histocompatibility complex class I alleles in treated HIV+ noncontrollers predict viral reservoir size** *AIDS* **37**:477–488
58. Thorball CW, et al. (2020) **Host Genomics of the HIV-1 Reservoir Size and Its Decay Rate During Suppressive Antiretroviral Treatment** *J Acquir Immune Defic Syndr* **85**:517–524
59. Zhang Z, et al. (2021) **IRF7 and RNH1 are modifying factors of HIV-1 reservoirs: a genome-wide association analysis** *BMC Med* **19**
60. Creelan BC, Antonia SJ. (2019) **The NKG2A immune checkpoint – a new direction in cancer immunotherapy** *Nat. Rev. Clin. Oncol* **16**:277–278
61. Gandhi RT, et al. (2021) **Selective Decay of Intact HIV-1 Proviral DNA on Antiretroviral Therapy** *J Infect Dis* **223**:225–233
62. Tumpach C, et al. (2023) **Adaptation of droplet digital PCR-based HIV transcription profiling to digital PCR and association of HIV transcription and total or intact HIV DNA** *Viruses* **15**
63. Singh K, et al. (2023) **Long-term persistence of transcriptionally active ‘defective’ HIV-1 proviruses: implications for persistent immune activation during antiretroviral therapy** *AIDS* **37**:2119–2130
64. Ishizaka A, et al. (2016) **Short intracellular HIV-1 transcripts as biomarkers of residual immune activation in patients on antiretroviral therapy** *J. Virol* **90**:5665–5676

65. Martin HA, et al. (2022) **New assay reveals vast excess of defective over intact HIV-1 transcripts in antiretroviral therapy-suppressed individuals** *J. Virol* **96**
66. Hsiao F, et al. (2020) **Tissue memory CD4+ T cells expressing IL-7 receptor-alpha (CD127) preferentially support latent HIV-1 infection** *PLoS Pathog* **16**
67. Gaebler C, et al. (2021) **Sequence evaluation and comparative analysis of novel assays for intact proviral HIV-1 DNA** *J. Virol* **95**
68. Falcinelli SD, et al. (2020) **Impact of biological sex on immune activation and frequency of the latent HIV reservoir during suppressive antiretroviral therapy** *J. Infect. Dis* **222**:1843–1852
69. Gay CL, et al. (2022) **Stable latent HIV infection and low-level viremia despite treatment with the broadly neutralizing antibody VRC07-523LS and the latency reversal agent vorinostat** *J. Infect. Dis* **225**:856–861
70. Falcinelli SD, et al. (2021) **Longitudinal Dynamics of Intact HIV Proviral DNA and Outgrowth Virus Frequencies in a Cohort of Individuals Receiving Antiretroviral Therapy** *J Infect Dis* **224**:92–100
71. Healy ZR, Murdoch DM. (2016) **OMIP-036: Co-inhibitory receptor (immune checkpoint) expression analysis in human T cell subsets** *Cytometry A* **89**:889–892
72. James G, et al. (2013) **An introduction to statistical learning**
1. Ester M, et al. (1996) **A density-based algorithm for discovering clusters in large spatial databases with noise** *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)* **96**:226–231

Editors

Reviewing Editor

Tadatsugu Taniguchi

University of Tokyo, Tokyo, Japan

Senior Editor

Tadatsugu Taniguchi

University of Tokyo, Tokyo, Japan

On responding to the first round of reviews, the authors have nicely adjusted their wording and fairly describe the results of their study. Certain markers were identified for further investigation. Yet, an overall non-obvious relationship between immune markers and HIV reservoirs has been shown previously, and despite the attempt to leverage powerful ML algorithms, they are not magical and cannot reveal strong relationships that fundamentally do not exist. In addition, categorical classification is for now hard to interpret and the more powerful ML algorithms do not seem to outperform more classic regression methods. Therefore, it remains relatively hard to evaluate the utility of this kind of study.

Initial summary:

Semenova et al. have studied a large cross-sectional cohort of people living with HIV on suppressive ART, N=115, and performed high dimensional flow-cytometry to then search for associations between immunological and clinical parameters and intact/total HIV DNA levels.

A number of interesting data science/ML approaches were explored on the data and the project seems a serious undertaking. However, like many other studies that have looked for these kinds of associations, there was not a very strong signal. Of course the goal of unsupervised learning is to find new hypotheses that aren't obvious to human eyes, but I felt in that context, there were (1) results slightly oversold, (2) some questions about methodology in terms mostly of reservoir levels, and (3) results were not sufficiently translated back into meaning in terms of clinical outcomes.

Strengths:

The study is evidently a large and impressive undertaking and combines many cutting edge statistical techniques with a comprehensive experimental cohort of people living with HIV, notably inclusive of populations underrepresented in HIV science. A number of intriguing hypotheses are put forward that could be explored further. Data will be shared and could be a useful repository for more specific analyses.

Weaknesses:

Despite the detailed experiments and methods, there was not a very strong signal for variable(s) predicting HIV reservoir size. The spearman coefficients are ~0.3, (somewhat weak, and acknowledged as such) and predictive models reach 70-80% prediction levels, though of sometimes categorical variables that are challenging to interpret.

There are some questions about methodology, as well as some conclusions that are not completely supported by results, or at minimum not sufficiently contextualized in terms of clinical significance. Edit, authors have substantially revised the text.

On associations: the false discovery rate correction was set at 5%, but data appear underdetermined with fewer observations than variables (144vars > 115pts), and it isn't always clear if/when variables are related (e.g inverses of one another, for instance %CD4 and %CD8).

The modeling of reservoir size was unusual, typically intact and defective HIV DNA are analyzed on a log10 scale (both for decays and predicting rebound). Also sometimes in this analysis levels are normalized (presumably to max/min?, e.g. S5), and given the large within-host variation of level we see in other works, it is not trivial to predict any downstream impact of normalization across population vs within person. Edit, fixed.

Also, the qualitative characterization of low/high reservoir is not standard, and naturally will split by early/late ART if done as above/below median. Given the continuous nature of these data it seems throughout that predicting above/below median is a little hard to translate into clinical meaning.

Lastly, work is comprehensive and appears solid, but the code was not shared to see how calculations were performed. Edit, fixed.

<https://doi.org/10.7554/eLife.94899.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:**Reviewer #1 (Public Review):***Summary:*

Semenova et al. have studied a large cross-sectional cohort of people living with HIV on suppressive ART, N=115, and performed high dimensional flow cytometry to then search for associations between immunological and clinical parameters and intact/total HIV DNA levels.

A number of interesting data science/ML approaches were explored on the data and the project seems a serious undertaking. However, like many other studies that have looked for these kinds of associations, there was not a very strong signal. Of course, the goal of unsupervised learning is to find new hypotheses that aren't obvious to human eyes, but I felt in that context, there were (1) results slightly oversold, (2) some questions about methodology in terms mostly of reservoir levels, and (3) results were not sufficiently translated back into meaning in terms of clinical outcomes.

We appreciate the reviewer's perspective. In our revised version of the manuscript, we have attempted to address these concerns by more adequately explaining the limitations of the study and by more thoroughly discussing the context of the findings. We are not able to associate the findings with specific clinical outcomes for individual study participants but we speculate about the overall biological meaning of these associations across the cohort. We cannot disagree with the reviewer, but we find the associations statistically significant, potentially reflecting real biological associations, and forming the basis for future hypothesis testing research.

Strengths:

The study is evidently a large and impressive undertaking and combines many cutting-edge statistical techniques with a comprehensive experimental cohort of people living with HIV, notably inclusive of populations underrepresented in HIV science. A number of intriguing hypotheses are put forward that could be explored further. Sharing the data could create a useful repository for more specific analyses.

We thank the reviewer for this assessment.

Weaknesses:

Despite the detailed experiments and methods, there was not a very strong signal for the variable(s) predicting HIV reservoir size. The Spearman coefficients are ~0.3, (somewhat weak, and acknowledged as such) and predictive models reach 70-80% prediction levels, though sometimes categorical variables are challenging to interpret.

We agree with the reviewer that individual parameters are only weakly correlated with the HIV reservoir, likely reflecting the complex and multi-factorial nature of reservoir/immune cell interactions. Nevertheless, these associations are statistically significant and form the basis for functional testing in viral persistence.

There are some questions about methodology, as well as some conclusions that are not completely supported by results, or at minimum not sufficiently contextualized in terms of clinical significance. On associations: the false discovery rate correction was set at 5%, but data appear underdetermined with fewer observations than variables (144vars >

115ppts), and it isn't always clear if/when variables are related (e.g inverses of one another, for instance, %CD4 and %CD8).

When deriving a list of cell populations whose frequency would be correlated with the reservoir, we focused on well-defined cell types for which functional validation exists in the literature to consider them as distinct cell types. For many of the populations, gating based on combinations of multiple markers leads to recovery of very few cells, and so we excluded some potential combinations from the analysis. We are also making our raw data available for others to examine and find associations not considered by our manuscript.

The modeling of reservoir size was unusual, typically intact and defective HIV DNA are analyzed on a log10 scale (both for decays and predicting rebound). Also, sometimes in this analysis levels are normalized (presumably to max/min?, e.g. S5), and given the large within-host variation of level we see in other works, it is not trivial to predict any downstream impact of normalization across population vs within-person.

We have repeated the analysis using log10 transformed data and the new figures are shown in Figure 1 and S2-S5.

Also, the qualitative characterization of low/high reservoir is not standard and naturally will split by early/late ART if done as above/below median. Given the continuous nature of these data, it seems throughout that predicting above/below median is a little hard to translate into clinical meaning.

Our ML models included time before ART as a variable in the analysis, and this was not found to be a significant driver of the reservoir size associations, except for the percentage of intact proviruses (see Figure 2C). Furthermore, we analyzed whether any of the reservoir correlated immune variables were associated with time on ART and found that, although some immune variables are associated with time on therapy, this was not the case for most of them (Table S4). We agree that it is challenging to translate above or below median into clinical meaning for this cohort, but we emphasize that this study is primarily a hypothesis generating approach requiring additional validation for the associations observed. We attempted to predict reservoir size as a continuous variable using the data and this approach was not successful (Figure S13). We believe that a significantly larger cohort will likely be required to generate a ML model that can accurately predict the reservoir as a continuous variable. We have added additional discussion of this to the manuscript.

Lastly, the work is comprehensive and appears solid, but the code was not shared to see how calculations were performed.

We now provide a link to the code used to perform the analyses in the manuscript, https://github.com/lesiasemenova/ML_HIV_reservoir.

Reviewer #2 (Public Review):

Summary:

Semenova et. al., performed a cross-sectional analysis of host immunophenotypes (using flow cytometry) and the peripheral CD4+ T cell HIV reservoir size (using the Intact Proviral DNA Assay, IPDA) from 115 people with HIV (PWH) on ART. The study mostly highlights the machine learning methods applied to these host and viral reservoir datasets but fails to interpret these complex analyses into (clinically, biologically) interpretable findings. For these reasons, the direct translational take-home message from this work is lost amidst a large list of findings (shown as clusters of associated markers) and sentences

such as "this study highlights the utility of machine learning approaches to identify otherwise imperceptible global patterns" - lead to overinterpretation of their data.

We have addressed the reviewer's concern by modifications to the manuscript that enhance the interpretation of the findings in a clinical and biological context.

Strengths:

Measurement of host immunophenotyping measures (multiparameter flow cytometry) and peripheral HIV reservoir size (IPDA) from 115 PWH on ART.

Major Weaknesses:

(1) Overall, there is little to no interpretability of their machine learning analyses; findings appear as a "laundry list" of parameters with no interpretation of the estimated effect size and directionality of the observed associations. For example, Figure 2 might actually give an interpretation of each X increase in immunophenotyping parameter, we saw a Y increase/decrease in HIV reservoir measure.

We have added additional text to the manuscript in which we attempt to provide more immunological and clinical interpretation of the associations. We also have emphasized that these associations are still speculative and will require additional validation. Nevertheless, our data should provide a rich source of new hypotheses regarding immune system/reservoir interaction that could be tested in future work.

(2) The correlations all appear to be relatively weak, with most Spearman R in the 0.30 range or so.

We agree with the review that the associations are mostly weak, consistent with previous studies in this area. This likely is an inherent feature of the underlying biology – the reservoir is likely associated with the immune system in complex ways and involves stochastic processes that will limit the predictability of reservoir size using any single immune parameter. We have added additional text to the manuscript to make this point clearer.

(3) The Discussion needs further work to help guide the reader. The sentence: "The correlative results from this present study corroborate many of these studies, and provide additional insights" is broad. The authors should spend some time here to clearly describe the prior literature (e.g., describe the strength and direction of the association observed in prior work linking PD-1 and HIV reservoir size, as well as specify which type of HIV reservoir measures were analyzed in these earlier studies, etc.) and how the current findings add to or are in contrast to those prior findings.

We have added additional text to the manuscript to help guide the readers through the possible biological significance of the findings and the context with respect to prior literature.

(4) The most interesting finding is buried on page 12 in the Discussion: "Uniquely, however, CD127 expression on CD4 T cells was significantly inversely associated with intact reservoir frequency." The authors should highlight this in the abstract, and title, and move this up in the Discussion. The paper describes a very high dimensional analysis and the key takeaways are not clear; the more the author can point the reader to the take-home points, the better their findings can have translatability to future follow-up mechanistic and/or validation studies.

We appreciate the reviewer's comment. We have increased the emphasis on this finding in the revised version of the manuscript.

(5) The authors should avoid overinterpretation of these results. For example in the Discussion on page 13 "The existence of two distinct clusters of PWH with different immune features and reservoir characteristics could have important implications for HIV cure strategies - these two groups may respond differently to a given approach, and cluster membership may need to be considered to optimize a given strategy." It is highly unlikely that future studies will be performing the breadth of parameters resulting here and then use these directly for optimizing therapy.

Our analyses indicate that membership of study participants in cluster1 or cluster 2 can be fairly accurately determined by a small number of individual parameters (KLRG1 etc, Figure 4F), and measuring the cells of PWH with the degree of breadth used in this paper would not be necessary to classify PWH into these clusters. As such, we feel that it is not unrealistic to speculate that this finding could turn out to be clinically useful, if it becomes clear that the clusters are biologically meaningful.

(6) There are only TWO limitations listed here: cross-sectional study design and the use of peripheral blood samples. (The subsequent paragraph notes an additional weakness which is misclassification of intact sequences by IPDA). This is a very limited discussion and highlights the need to more critically evaluate their study for potential weaknesses.

We have expanded on the list of limitations discussed in the manuscript. In particular, we now address the size of the cohort, the composition with respect to different genders and demographics, lack of information for the timing of ART and the lack of information regarding intracellular transcriptional pathways.

(7) A major clinical predictor of HIV reservoir size and decay is the timing of ART initiation. The authors should include these (as well as other clinical covariate data - see #12 below) in their analyses and/or describe as limitations of their study.

All of the participants that make up our cohort were treated during chronic infection, and the precise timing of ART initiation is unclear in most of these cases. We have added additional information to explain this in the manuscript and include this in the list of limitations.

Reviewer #3 (Public Review):

Summary:

This valuable study by Semenova and colleagues describes a large cross-sectional cohort of 115 individuals on ART. Participants contributed a single blood sample which underwent IPDA, and 25-color flow with various markers (pre and post-stimulation). The authors then used clustering, decision tree analyses, and machine learning to look for correlations between these immunophenotypic markers and several measures of HIV reservoir volume. They identified two distinct clusters that can be somewhat differentiated based on total HIV DNA level, intact HIV DNA level, and multiple T cell cellular markers of activation and exhaustion.

The conclusions of the paper are supported by the data but the relationships between independent and dependent variables in the models are correlative with no mechanistic work to determine causality. It is unclear in most cases whether confounding variables could explain these correlations. If there is causality, then the data is not sufficient to infer directionality (ie does the immune environment impact the HIV reservoir or vice versa or both?). In addition, even with sophisticated and appropriate machine learning approaches, the models are not terribly predictive or highly correlated. For these

reasons, the study is very much hypothesis-generating and will not impact cure strategies or HIV reservoir measurement strategies in the short term.

We appreciate the reviewer's comments regarding the value of our study. We fully acknowledge that the causal nature and directionality of these associations are not yet clear and agree that the study is primarily hypothesis generating in nature. Nevertheless, we feel that the hypotheses generated will be valuable to the field. We have added additional text to the manuscript to emphasize the hypothesis generating nature of this paper.

Strengths:

The study cohort is large and diverse in terms of key input variables such as age, gender, and duration of ART. Selection of immune assays is appropriate. The authors used a wide array of bioinformatic approaches to examine correlations in the data. The paper was generally well-written and appropriately referenced.

Weaknesses:

(1) The major limitation of this work is that it is highly exploratory and not hypothesis-driven. While some interesting correlations are identified, these are clearly hypothesis-generating based on the observational study design.

We agree that the major goal of this study was hypothesis generating and that our work is exploratory in nature. Performing experiments with mechanism testing goals in human participants with HIV is challenging. Additionally, before such mechanistic studies can be undertaken, one must have hypotheses to test. As such we feel our study will be useful for the field in helping to identify hypotheses that could potentially be tested.

(2) The study's cross-sectional nature limits the ability to make mechanistic inferences about reservoir persistence. For instance, it would be very interesting to know whether the reservoir cluster is a feature of an individual throughout ART, or whether this outcome is dynamic over time.

We agree with the reviewer's comment. Longitudinal studies are challenging to carry out with a study cohort of this size, and addressing questions such as the one raised by the reviewer would be of great interest. We believe our study nevertheless has value in identifying hypotheses that could be tested in a longitudinal study.

(3) A fundamental issue is that I am concerned that binarizing the 3 reservoir metrics in a 50/50 fashion is for statistical convenience. First, by converting a continuous outcome into a simple binary outcome, the authors lose significant amounts of quantitative information. Second, the low and high reservoir outcomes are not actually demonstrated to be clinically meaningful: I presume that both contain many (?all) data points above levels where rebound would be expected soon after interruption of ART. Reservoir levels would also have no apparent outcome on the selection of cure approaches. Overall, dividing at the median seems biologically arbitrary to me.

The reviewer raises a valid point that the clinical significance of above or below median reservoir metrics is unclear, and that the size of the reservoir has potentially little relation to rebound and cure approaches. In the manuscript, we attempted to generate models that can predict reservoir size as a continuous variable in Figure S13 and find that this approach performs poorly, while a binarized approach was more successful. As such we have included both approaches in the manuscript. It is possible that future studies with larger sample sizes and more detailed measurements will perform better for continuous variable prediction. While this is a fairly large study (n=115) by the standards of HIV reservoir analyses, it is a

small study by the standards of the machine learning field, and accurate predictive ML models for reservoir size as a continuous variable will likely require a much larger set of samples/participants. Nevertheless, we feel our work has value as a template for ML approaches that may be informative for understanding HIV/immune interactions and generates novel hypotheses that could be validated by subsequent studies.

(4) The two reservoir clusters are of potential interest as high total and intact with low % intact are discriminated somewhat by immune activation and exhaustion. This was the most interesting finding to me, but it is difficult to know whether this clustering is due to age, time on ART, other co-morbidity, ART adherence, or other possible unmeasured confounding variables.

We agree that this finding is one of the more interesting outcomes of the study. We examined a number of these variables for association with cluster membership, and these data are reported in Figure S8A-D. Age, years of ART and CD4 Nadir were all clearly different between the clusters. The striking feature of this clustering, however, is the clear separation between the two groups of participants, as opposed to a continuous gradient of phenotypes. This could reflect a bifurcation of outcomes for people with HIV, dynamic changes in the reservoir immune interactions over time, or different levels of untreated infection. It is certainly possible that some other unmeasured confounding variables contribute to this outcome and we have attempted to make this limitation clearer.

(5) At the individual level, there is substantial overlap between clusters according to total, intact, and % intact between the clusters. Therefore, the claim in the discussion that these 2 cluster phenotypes may require different therapeutic approaches seems rather speculative. That said, the discussion is very thoughtful about how these 2 clusters may develop with consideration of the initial insult of untreated infection and / or differences in immune recovery.

We agree with the reviewer that this claim is speculative, and we have attempted to moderate the language of the text in the revised version.

(6) The authors state that the machine learning algorithms allow for reasonable prediction of reservoir volume. It is subjective, but to me, 70% accuracy is very low. This is not a disappointing finding per se. The authors did their best with the available data. It is informative that the machine learning algorithms cannot reliably discriminate reservoir volume despite substantial amounts of input data. This implies that either key explanatory variables were not included in the models (such as viral genotype, host immune phenotype, and comorbidities) or that the outcome for testing the models is not meaningful (which may be possible with an arbitrary 50/50 split in the data relative to median HIV DNA volumes: see above).

We acknowledge that the predictive power of the models generated from these data is modest and we have clarified this point in the revised manuscript. As the reviewer indicates, this may result from the influence of unmeasured variables and possible stochastic processes. The data may thus demonstrate a limit to the predictability of reservoir size which may be inherent to the underlying biology. As we mention above, this study size (n=115) is fairly small for the application of ML methods, and an increased sample size will likely improve the accuracy of the models. At this stage, the models we describe are not yet useful as predictive clinical tools, but are still nonetheless useful as tools to describe the structure of the data and identify reservoir associated immune cell types.

(7) The decision tree is innovative and a useful addition, but does not provide enough discriminatory information to imply causality, mechanism, or directionality in terms of

whether the immune phenotype is impacting the reservoir or vice versa or both. Tree accuracy of 80% is marginal for a decision tool.

The reviewer is correct about these points. In the revised manuscript, we have attempted to make it clear that we are not yet advocating using this approach as a decision tool, but simply a way to visualize the data and understand the structure of the dataset. As we discuss above, the models will likely need to be trained on a larger dataset and achieve higher accuracy before use as a decision tool.

(8) Figure 2: this is not a weakness of the analysis but I have a question about interpretation. If total HIV DNA is more predictive of immune phenotype than intact HIV DNA, does this potentially implicate a prior high burden of viral replication (high viral load &/or more prolonged time off ART) rather than ongoing reservoir stimulation as a contributor to immune phenotype? A similar thought could be applied to the fact that clustering could only be detected when applied to total HIV DNA-associated features. Many investigators do not consider defective HIV DNA to be "part of the reservoir" so it is interesting to speculate why these defective viruses appear to have more correlation with immunophenotype than intact viruses.

We agree with the reviewer that this observation could reflect prior viral burden and we have added additional text to make this clearer. Even so, we cannot rule out a model in which defective viral DNA is engaged in ongoing stimulation of the immune system during ART, leading to the stronger association between total DNA and the immune cell phenotypes. We hypothesize that the defective proviruses could potentially be triggering innate immune pattern recognition receptors via viral RNA or DNA, and a higher burden of the total reservoir leads to a stronger apparent association with the immune phenotype. We have included text in the discussion about this hypothesis.

(9) Overall, the authors need to do an even more careful job of emphasizing that these are all just correlations. For instance, HIV DNA cannot be proven to have a causal effect on the immunophenotype of the host with this study design. Similarly, immunophenotype may be affecting HIV DNA or the correlations between the two variables could be entirely due to a separate confounding variable

We have revised the text of the manuscript to emphasize this point, and we acknowledge that any causal relationships are, at this point, simply speculation.

(10) In general, in the intro, when the authors refer to the immune system, they do not consistently differentiate whether they are referring to the anti-HIV immune response, the reservoir itself, or both. More specifically, the sentence in the introduction listing various causes of immune activation should have citations. (To my knowledge, there is no study to date that definitively links proviral expression from reservoir cells in vivo to immune activation as it is next to impossible to remove the confounding possible imprint of previous HIV replication.) Similarly, it is worth mentioning that the depletion of intact proviruses is quite slow such that proviral expression can only be stimulating the immune system at a low level. Similarly, the statement "Viral protein expression during therapy likely maintains antigen-specific cells of the adaptive immune system" seems hard to dissociate from the persistence of immune cells that were reactive to viremia.

We updated the text of the manuscript to address these points and have added additional citations as per the reviewer's suggestion.

(11) Given the many limitations of the study design and the inability of the models to discriminate reservoir volume and phenotype, the limitations section of the discussion

seems rather brief.

We have now expanded the limitations section of the discussion and added additional considerations. We now include a discussion of the study cohort size, composition and the detail provided by the assays.

Recommendations for the authors:

Reviewer #1 (Recommendations For The Authors):

A few specific comments:

"This pattern is likely indicative of a more profound association of total HIV DNA with host immunophenotype relative to intact HIV DNA."

Most studies I have seen (e.g. single cell from Litherfeld/Yu group) show intact proviruses are generally more activated/detectable/susceptible to immune selection, so I have a hard time thinking defective proviruses are actually more affected by immunotype.

We hypothesize that this association is actually occurring in the opposite direction – that the defective provirus are having a greater impact on the immune phenotype, due to their greater number and potential ability to engage innate or adaptive immune receptors. We have clarified this point in the manuscript

"The existence of two distinct clusters of PWH with different immune features and reservoir characteristics could have important implications for HIV cure strategies - these two groups may respond differently to a given approach, and cluster membership may need to be considered to optimize a given strategy."

I find this a bit of a reach, given that the definition of 2 categories depended on the total size.

We have modified the language of this section to reduce the level of speculation.

"This study is cross-sectional in nature and is primarily observational, so caution should be used interpreting findings associated with time on therapy".

I found this an interesting statement because ultimately time on ART shows up throughout the analysis as a significant predictor, do you mean something about how time on ART could indicate other confounding variables like ART regimen or something?

We have rephrased this comment to avoid confusion. We were simply trying to make the point that we should avoid speculating about longitudinal dynamics from cross sectional data.

"As expected, the plots showed no significant correlation for intact HIV DNA versus years of ART (Figure 1B), while total reservoir size was positively correlated with the time of ART (Figure 1A, Spearman $r = 0.31$)."

Is this expected? Studies with longitudinal data almost uniformly show intact decay, at least for the first 10 or so years of ART, and defective/total stability (or slight decay). Also probably "time on ART" to not confuse with the duration of infection before ART.

We have updated the language of this section to address this comment. We have avoided comparing our data with respect to time on ART to longitudinal studies for reasons given above.

On dimensionality reduction, as this PaCMAP seems a relatively new technique (vs tSNE and UMAP which are more standard, but absolutely have their weaknesses), it does seem important to contextualize. I think it would still be useful to show PCA and assess the % variance of each additional dimension to assess the effective dimensionality, it would be helpful to show a plot of % variance by # components to see if there is a cutoff somewhere, and if PaCMAP is really picking this up to determine the 2 dimensions/2 clusters is ideal. Figure 4B ultimately shows a lot of low/high across those clusters, and since low/high is defined categorically it's hard to know which of those dots are very close to the other categories.

We have added this analysis to the manuscript – found in Figure S9. The PCA plot indicates that members of the two clusters also separate on PCA although this separation is not as clear as for the PaCMAP plot.

Minor comments on writing etc:

Intro

-Needs some references on immune activation sequelae paragraph.

We have added some additional references to this section.

-"promote the entry of recently infected cells into the reservoir" -- that is only one possible mechanistic explanation, it's not unreasonable but it seems important to keep options open until we have more precise data that can illuminate the mechanism of the overabundance.

We have modified the text to discuss additional hypotheses.

-You might also reference Pankau et al Ppath for viral seeding near the time of ART.

We have added this reference.

-"Viral protein expression during therapy likely maintains antigen-specific cells of the adaptive immune system" - this was unclear to me, do you mean HIV-specific cells that act against HIV during ART? I think most studies show immunity against HIV (CD8 and CD4) wanes over time during ART.

The Goonetilleke lab has recently generated data indicating that antiviral T cell responses are remarkably stable over time on ART, but we agree with the reviewer that the idea that ongoing antigen expression in the reservoir maintains these cells is speculative. We have modified the text to make this point clearer.

-Overall I think the introduction lacked a little bit of definitional precision: i.e. is the reservoir intact vs replication competent vs all HIV DNA and whether we are talking about PWH on long-term ART and how long we should be imagining? The first years of ART are certainly different than later, in terms of dynamics. The ultimate implications are likely specific for some of these categorizations.

-"persistent sequelae of the massive disruptions to T cell homeostasis and lymphoid structures that occur during untreated HIV infection" needs a lot more context/referencing. For instance, Peter Hunt showed a decrease in activation after ART a long time ago.

-Heather Best et al show T cell clonality stays perturbed after ART.

We have updated the text of the introduction and added references to address the reviewer's comments.

Results

-It would be important to mention the race of participants and any information about expected clades of acquired viruses, this gets mentioned eventually with reference to the Table but the breakdown would be helpful right away.

We have added this information to the results section.

-"performed Spearman correlations", may be calculated or tested?

We have corrected the language for this sentence.

Comments on figures:

-Figure 1 data on linear scale (re discussion above) -- hard to even tell if there is a decay (to match with all we know from various long-term ART studies).

-Figure 4 data is shown on $\ln(\log_e)$ scale, which is hard to interpret for most people.

-Figures 4 C,D, and E should have box plots to visually assess the significance.

-Figure 4B legend says purple/pink but I think the colors are different in the plot, could be about transparency

-Figure 5 it is now not clear if $\log_e$ (?).

-Figure 6 "HIV reservoir characteristics" might be better to make this more explicit. Do you mean for instance in the 6B title Total HIV DNA per million CD4+ T cells I think?

We have made these modifications.

Reviewer #2 (Recommendations For The Authors):

Minor Weaknesses:

(1) The Introduction is too long and much of the text is not directly related to the study's research question and design.

We have streamlined the introduction in the revised manuscript.

(2) While no differences were seen by age or race, according to the authors, this is unlikely to be useful since the numbers are so small in some of these subcategories. Results from sensitivity analyses (e.g., excluding these individuals) may be more informative/useful.

We agree that the lower numbers of participants for some subgroupings makes it challenging to know for sure if there are any differences based on these variables. Have added text to clarify this. We have added age, race and gender to the LOCO analysis and to the variable inflation importance analysis (Table S5).

(3) For Figure 4, based on what was described in the Results section of the manuscript, the authors should clarify that the figures show results for TOTAL HIV DNA only (not intact DNA): "Dimension reduction machine learning approaches identified two robust clusters of PWH when using total HIV DNA reservoir-associated immune cell frequencies (Figure 4A), but not for intact or percentage intact HIV DNA (Figure 4B and 4C)".

We have added this information.

(4) The statement on page 5, first paragraph, "Interestingly, when we examined a plot of percent intact proviruses versus time on therapy (Figure 1C), we observed a biphasic decay pattern," is not new (Peluso JCI Insight 2020, Gandhi JID 2023, McMyn JCI 2023). Prior studies have clearly demonstrated this biphasic pattern and should be cited here, and the sentence should be reworded with something like "consistent with prior work", etc.

We have added citations to these studies and rephrased this comment.

(5) The Cohort and sample collection sections are somewhat thin. Further details on the cohort details should include at the very minimum some description of the timing of ART initiation (is this mostly a chronic-treated cohort?) and important covariate data such as nadir CD4+ T cell count, pre-ART viral load, duration of ART suppression, etc.

The cohort was treated during chronic infection, and we have clarified this in the manuscript. Information regarding CD4 nadir and years on ART are included in Table 1. Unfortunately, pre-ART viral load was not available for most members of this cohort, so we did not use it for analyses. The partial pre-ART viral load data is included with the dataset we are making publicly available.

Reviewer #3 (Recommendations For The Authors):

Minor points:

(1) What is meant by CD4 nadir? Is this during primary infection or the time before ART initiation?

We have clarified this description in the manuscript. This term refers to the lowest CD4 count recorded during untreated infection.

(2) The authors claim that determinants of reservoir size are starting to emerge but other than the timing of ART, I am not sure what studies they are referring to.

We have updated the language of this section. We intended to refer to studies looking at correlates of reservoir size, and feel that this is a more appropriate term than 'determinants'

(3) The discussion does not tie in the model-generated hypotheses with the known mechanisms that sustain the reservoir: clonal proliferation balanced by death and subset differentiation. It would be interesting to tie in the proposed reservoir clusters with these known mechanisms.

We have added additional text to the manuscript to address these mechanisms.

(4) Figure 1: Total should be listed as total HIV DNA.

We have updated this in the manuscript.

(5) *Figure 1C: Worth mentioning the paper by Reeves et al which raises the possibility that the flattening of intact HIV DNA at 9 years may be spurious due to small levels of misclassification of defective as intact.*

We have added this reference.

(6) *"Total reservoir frequency" should be "total HIV DNA concentration"*

We respectfully feel that “frequency” is a more accurate term than “concentration”, since we are expressing the reservoir as a fraction of the CD4 T cells, while “concentration” suggests a denominator of volume.

(7) *Figure S2-5: label y-axis total HIV DNA.*

We have updated this figure.

<https://doi.org/10.7554/eLife.94899.2.sa0>