nature immunology

ARTICLES

https://doi.org/10.1038/s41590-019-0581-0

Aberrant B cell repertoire selection associated with HIV neutralizing antibody breadth

Krishna M. Roskin[1,2,3], Katherine J. L. Jackson[4], Ji-Yeun Lee[5], Ramona A. Hoh[5], Shilpa A. Joshi[5], Kwan-Ki Hwang[6], Mattia Bonsignori[6], Isabela Pedroza-Pacheco[7], Hua-Xin Liao[6], M. Anthony Moody[6], Andrew Z. Fire[5,8], Persephone Borrow[7], Barton F. Haynes[6,9,10,11]* and Scott D. Boyd[5]*

A goal of HIV vaccine development is to elicit antibodies with neutralizing breadth. Broadly neutralizing antibodies (bNAbs) to HIV often have unusual sequences with long heavy-chain complementarity-determining region loops, high somatic mutation rates and polyreactivity. A subset of HIV-infected individuals develops such antibodies, but it is unclear whether this reflects systematic differences in their antibody repertoires or is a consequence of rare stochastic events involving individual clones. We sequenced antibody heavy-chain repertoires in a large cohort of HIV-infected individuals with bNAb responses or no neutralization breadth and uninfected controls, identifying consistent features of bNAb repertoires, encompassing thousands of B cell clones per individual, with correlated T cell phenotypes. These repertoire features were not observed during chronic cytomegalovirus infection in an independent cohort. Our data indicate that the development of numerous B cell lineages with antibody features associated with autoreactivity may be a key aspect in the development of HIV neutralizing antibody breadth.

bNAbs against HIV are able to recognize the envelope glycoproteins of diverse viral variants, often by targeting conserved epitopes that play important functional roles in the viral life cycle. After years of chronic infection, between 10% and 50% of HIV-infected individuals develop antibodies with neutralizing breadth in their serum, depending on the criteria used to define breadth[1]. Many HIV bNAb lineages have stereotyped genetic and structural features, such as high frequencies of somatic hypermutation (SHM) and long heavy-chain complementarity-determining region 3 (CDR-H3) sequences. These sequence features are rare in healthy human antibody repertoires, but contribute to targeting vulnerable epitope sites on the HIV envelope glycoprotein[2]. Understanding the mechanisms contributing to bNAb formation is a priority for HIV vaccine development: vaccines eliciting bNAb-like antibodies could potentially protect against a wide range of HIV strains.

Why do some HIV-infected individuals generate neutralizing antibody breadth, while others do not? It is possible that all individuals who are infected have a similar potential for producing bNAbs, but that formation of these antibodies is rate-limited by very rare stochastic events such as formation of naive B cells with particular CDR-H3 sequences, or the acquisition of particular somatic mutations during affinity maturation. In this model, bNAb lineages in each patient would be unusual outliers from the rest of their antibody repertoire. Alternatively, some infected individuals could be predisposed to generating bNAb-like antibodies due to systematic differences in their B cell repertoire formation or selection processes. In such patients, bNAbs would arise from a pool of many antibody lineages with similar features. Evaluation of germline allelic variants in antibody heavy-chain variable gene segments found no correlation with bNAb production[3]. However, other features of repertoire formation such as V(D)J segment recombination frequencies, CDR-H3 junction production, SHM frequency or differences in B cell selection could also affect the likelihood of producing bNAb-like antibodies. Factors such as the T cell populations providing help to B cells; regulatory T (T_{reg}) cell populations that may control autoreactive B cells; or the mixtures and frequencies of HIV viral variants in the patient, could also lead to systematic differences in neutralizing antibody breadth between patients.

Here, we analyze antibody heavy-chain (*IGH*) gene repertoires in 96 patients with chronic HIV infection (46 patients with extensive neutralizing breath (bNAb) and 50 patients without neutralizing breadth (noNAb)) and 43 HIV-uninfected controls from the same geographical regions, to assess whether the two HIV-infected groups show systematic differences in antibody repertoire formation or selection. Our results based on extensive sequencing of each patient's *IGH* repertoire indicate that HIV-infected individuals show significant perturbations of their antibody repertoires, including suppressed average SHM frequencies and increased frequencies of B cells expressing antibodies with features associated with autoreactivity. BNAb individuals each have hundreds to thousands of antibody lineages with long CDR-H3s and very high SHM frequencies. In contrast, noNAb subjects and HIV-uninfected individuals appear to select against antibody lineages with long CDR-H3 and high SHM frequencies. Further, we note that patient B cell and T cell phenotypes are linked, with a negative correlation seen between CTLA-4$^+$ T_{reg} cells and bNAb patient CDR-H3 lengths.

[1]Department of Pediatrics, University of Cincinnati, College of Medicine, Cincinnati, OH, USA. [2]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [3]Division of Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [4]Department of Immunology, Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia. [5]Department of Pathology, Stanford University, Stanford, CA, USA. [6]Duke Human Vaccine Institute, Duke University School of Medicine, Durham, NC, USA. [7]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK. [8]Department of Genetics, Stanford University, Stanford, CA, USA. [9]Department of Medicine, Duke University School of Medicine, Durham, NC, USA. [10]Department of Immunology, Duke University School of Medicine, Durham, NC, USA. [11]Duke Global Health Institute, Duke University School of Medicine, Durham, NC, USA. *e-mail: barton.haynes@duke.edu; sboyd1@stanford.edu

An unrelated chronic viral infection, human cytomegalovirus, shows no evidence of eliciting the *IGH* repertoire deviations seen in bNAb HIV-infected individuals. Our results indicate that production of HIV neutralizing antibody breadth is associated with specific systematic differences in individual B cell repertoire formation and selection.

Results

IgG SHM in HIV-infected bNAb and noNAb individuals. We first evaluated SHM frequencies in *IGH* repertoires encoding each antibody isotype and subclass, to assess the effects of HIV infection, as well as to test for differences between bNAb and noNAb individuals. bNAb individuals were those whose serum showed the broadest inhibition of a panel of 12 HIV isolates in a pseudovirus infection assay, with 95% inhibition of at least seven isolates (see Methods). HIV-infected individuals showed overall decreases in mean SHM frequencies in *IGH* expressed as IgG but not in other isotypes, compared to uninfected individuals (Fig. 1a and Supplementary Fig. 1a). bNAb individuals, however, had mean IgG SHM frequencies closer to those of uninfected people. Notably, the most highly mutated antibodies (represented here by the SHM value of the 99th percentile in the distribution of SHM frequencies in each participant) were more mutated in IgG subtypes of bNAb individuals compared to both uninfected controls and noNAb individuals (Fig. 1b and Supplementary Fig. 1b). The 99th percentile SHM values of IgG in bNAb individuals approach the SHM frequencies of well-described bNAb monoclonal antibodies isolated in previous studies (Fig. 1c), and are higher than the SHM of non-neutralizing antibodies or those elicited by vaccination in previous studies (Supplementary Fig. 1c). We noted that the standard deviation of SHM values for the clones within each individual was also higher in IgG subtypes in bNAb individuals compared to noNAb or HIV-uninfected individuals (Supplementary Fig. 1d). The antibody clonal lineages at the 99th percentile of mutation or higher, represent an average of 790 distinct clones (range 287–1,390) in each bNAb individual, rather than small numbers of outlier clones.

Frequencies of insertions and deletions (indels) in the antibody variable regions have been previously shown to be correlated with SHM frequency[4]. Here, insertion frequency in Ig genes correlated with SHM frequency and was decreased in all HIV-infected individuals compared to uninfected individuals, but deletion rates in IgG1 in bNAb individuals were higher, even than in uninfected individuals (Supplementary Fig. 1e). IgA antibodies behaved differently from the other isotypes, having increased indel rates in HIV-infected individuals compared to uninfected individuals, even though their SHM frequencies were similar to those of uninfected controls. These results highlight further differences in IgG SHM or selection processes compared to the other isotypes in antibody repertoires of patients with HIV infection.

CDR-H3 lengths in HIV-infected individuals are increased. Antibodies with long CDR-H3 regions have been previously reported to be enriched for autoreactive or polyreactive sequences, although this is not the only determinant of antibody autoreactivity[5]. We found that IgG subtype antibodies in both bNAb and noNAb individuals had longer CDR-H3s compared to uninfected individuals, as previously reported (Fig. 2a and Supplementary Fig. 2a)[6]. The fraction of very long CDR-H3 (>19 amino acids) was also higher in IgG subtypes HIV-infected individuals compared to uninfected controls but did not differ between bNAb and noNAb individuals (Fig. 2b and Supplementary Fig. 2b). Clones with >19 amino acids in their CDR-H3s represented thousands of distinct lineages in each participant (mean of 8,548 for bNAb individuals, 7,540 for noNAb individuals and 7,891 for uninfected individuals). To evaluate whether these CDR-H3 length differences could be attributed to the early steps of B cell antigen receptor (BCR) rearrangement in the bone marrow, we sequenced *IGH* rearrangements from a genomic DNA template and analyzed the sequences of unproductive rearrangements (the remnants of unsuccessful first attempts by the B cell precursor to rearrange the *IGH* locus). Nonfunctional rearrangement CDR-H3s were longer than productive ones but surprisingly, all HIV-infected individuals had shorter nonproductive CDR-H3s than uninfected individuals, with bNAb individuals having the shortest nonproductive CDR-H3s of all (Fig. 2c). Therefore, the increased proportion of IgG antibodies with CDR-H3 lengths exceeding 19 amino acids in HIV-infected individuals does not arise during BCR repertoire formation but is a consequence of subsequent B cell selection.

Long CDR-H3s and high SHM in bNAb IgG repertoires. Given the observed differences in SHM and CDR-H3 features between antibody repertoires in HIV-infected individuals and those of uninfected individuals, we sought to determine whether any interactions between antibody characteristics were specific to bNAb individuals, as many known bNAb antibody lineage types require both long CDR-H3 and high SHM frequencies in their development. We plotted the average CDR-H3 lengths for antibodies at the 70th, 80th, 90th, 95th and 99th or higher percentiles of SHM frequency in each patient's repertoires of each antibody isotype (Fig. 3 and Supplementary Fig. 2c). bNAb-producing individuals showed a significant increase in CDR-H3 lengths in the most-mutated fractions of their IgG repertoires, in contrast to noNAb or uninfected individuals. Therefore, bNAb producers are distinct from noNAb individuals in the selection or persistence of B cells expressing antibodies with long CDR-H3 regions, particularly when these are highly somatically mutated. Previous studies have measured frequencies of gp120$^+$ B cells in bNAb patient blood to be in the order of 30 per 30,000 (0.1%) IgG$^+$ memory B cells and bNAbs to be 1–4 per 30,000 (0.003–0.01%) IgG$^+$ memory B cells, suggesting that the SHM and CDR-H3 differences that we observe between bNAb and noNAb individuals are not due solely to HIV-specific gp120-binding B cell lineages[7–10].

Decreased CTLA-4$^+$ T_{regs} in bNAb individuals with long CDR-H3. We evaluated whether the SHM frequencies or CDR-H3 lengths in bNAb, noNAb or uninfected individuals were related to HIV viral load or frequencies of CD4$^+$ T cells or other T cell subsets. Average SHM frequencies in IgG antibodies within the noNAb group were correlated with their CD4$^+$ T cell frequencies, but this was not the case for the bNAb or uninfected individuals, suggesting that noNAb individuals alone show a dependence on CD4$^+$ T cell frequencies for maintaining SHM (Supplementary Fig 3a). As expected, both bNAb and noNAb individuals had lower CD4$^+$ T cell frequencies than uninfected individuals (Supplementary Fig. 3b) and as previously reported, bNAb individuals had higher viral loads than noNAb individuals, although these average differences between groups were not sufficient to predict bNAb status (Supplementary Fig. 3c)[11]. The 99th percentile IgG SHM frequencies were positively correlated with CD4$^+$ T cell frequencies and were negatively correlated with higher viral loads in noNAb individuals, but not in bNAb individuals (Supplementary Fig. 3d,e). The data suggest that bNAb producers respond differently to their immunological environment (viral load and frequencies of CD4$^+$ T cells; Supplementary Fig. 3f) and achieve very high SHM frequencies in numerous clones despite having lower CD4$^+$ T cell frequencies than healthy individuals.

We next examined T cell subsets of known helper or regulatory functions. Resting memory follicular helper (T_{FH}) cell (CXCR3$^-$PD-1$^+$ within CD4$^+$CXCR5$^+$) frequencies in the blood were not correlated with 99th percentile IgG SHM frequencies (Supplementary Fig. 4a). Given that the SHM frequencies in the most highly mutated IgG fractions were higher in bNAb individuals than in noNAb or healthy individuals, generation of very high SHM

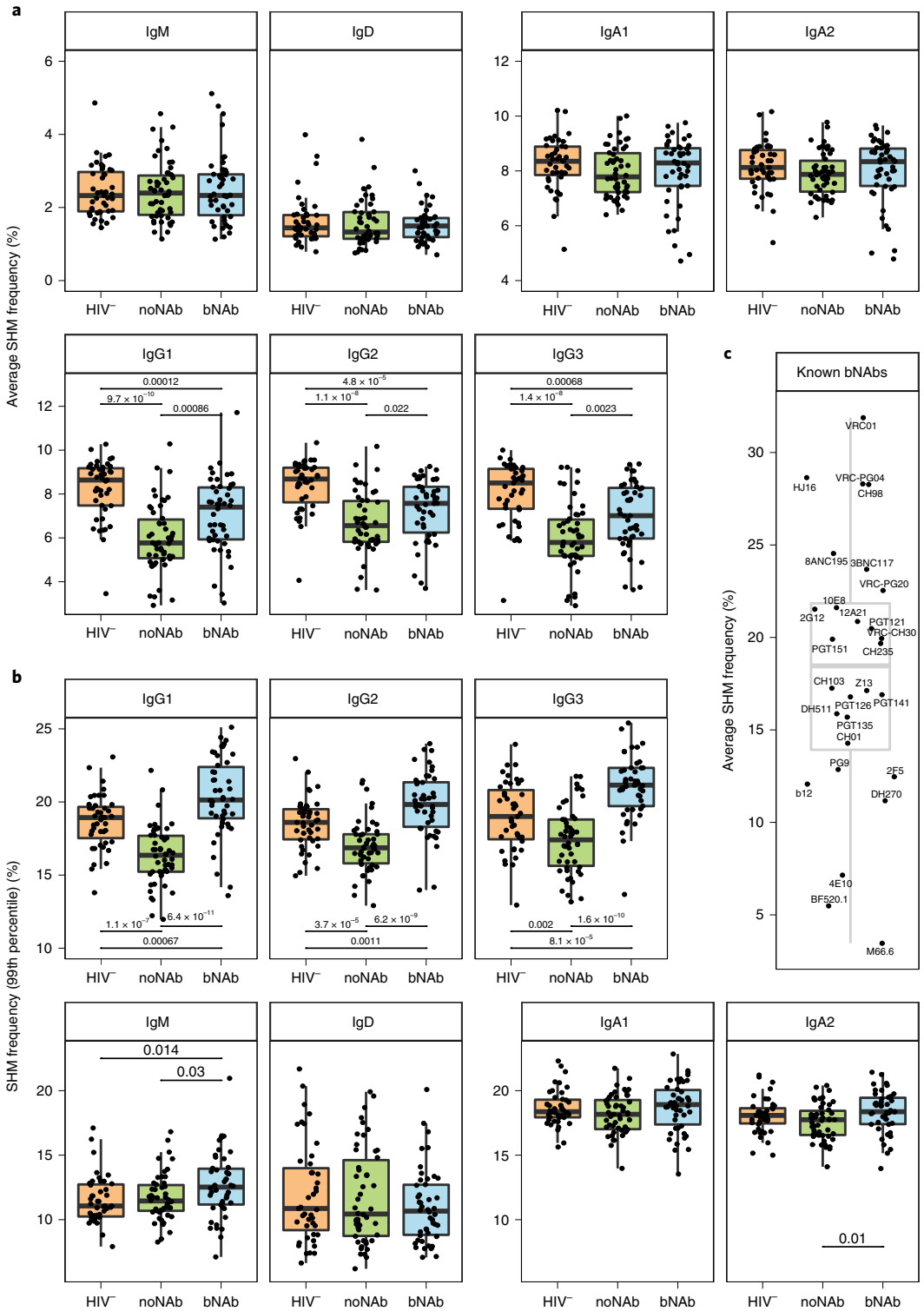

Fig. 1 | Differences in SHM frequency in antibody heavy-chain transcripts in HIV-uninfected controls and HIV-infected bNAb and noNAb individuals. The clonal structure of the expressed antibody heavy-chain repertoire is described in the Methods. Each dot represents the mean SHM frequency (the number of nucleotides mutated from the germline sequence of IGHV, divided by the total number of nucleotides in the IGHV sequence) for clones with the indicated isotype within a participant. **a**, The overall SHM frequency for IgM, IgD, IgG and IgA isotypes. The SHM frequencies in the low abundancy isotypes, IgG4 and IgE, are shown in Supplementary Fig. 1a. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected (HIV⁻), $n=43$; noNAb, $n=50$; bNAb, $n=46$. **b**, The SHM frequency of the 99th percentile of most-mutated IgM, IgD, IgG and IgA IGHV sequences for each individual. Data for IgG4 and IgE are shown in Supplementary Fig. 1b. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$. **c**, SHM frequencies of a collection of known HIV bNAbs. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers).

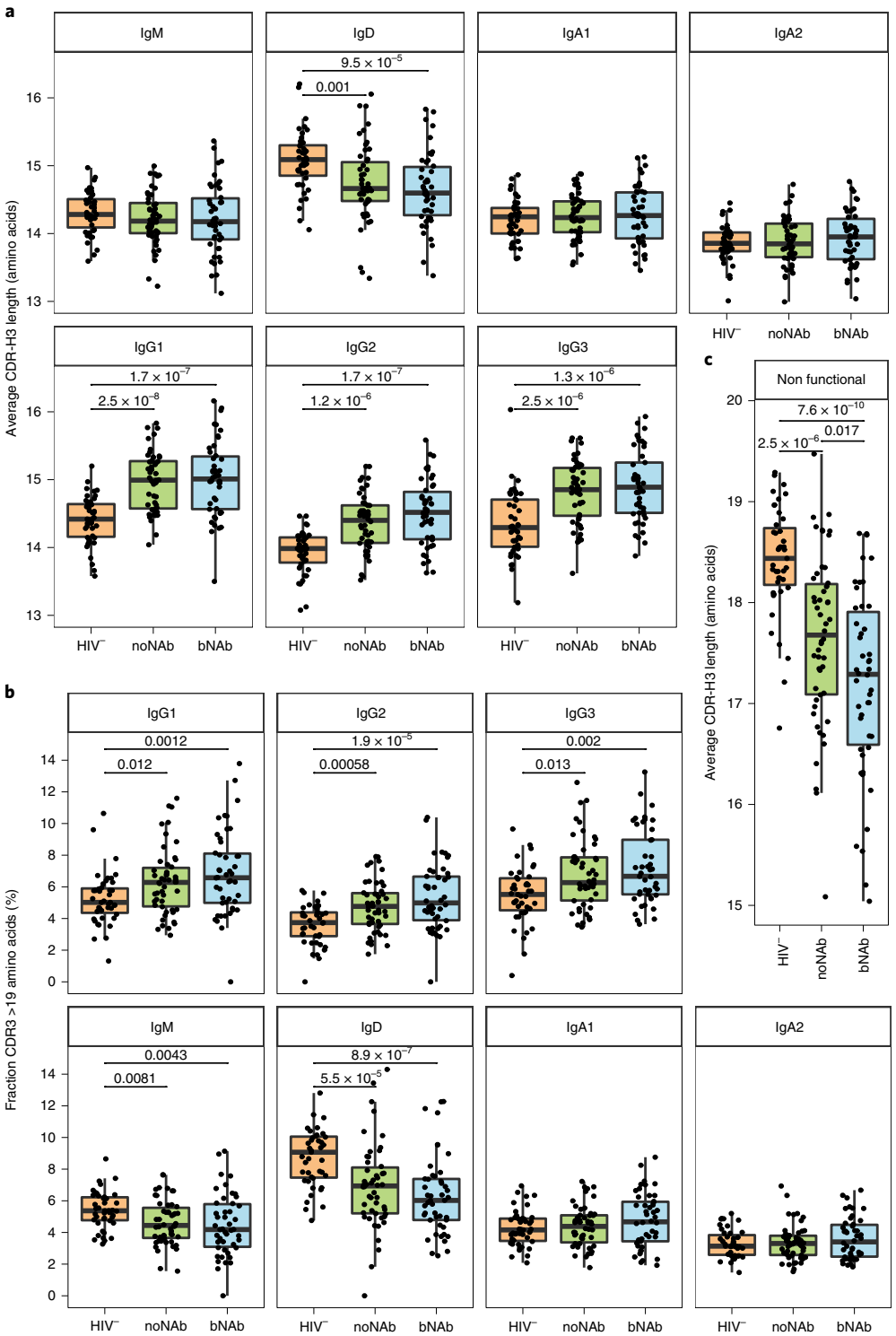

Fig. 2 | Differences in the length of the CDR-H3 of antibody heavy chains in healthy controls compared to HIV-infected bNAb and noNAb individuals. **a**, CDR-H3 amino acid lengths in HIV-infected individuals and uninfected controls in several isotypes. Each dot represents the mean CDR-H3 length of clones expressing the indicated isotype within an individual. Data for IgG4 and IgE are shown in Supplementary Fig. 2a. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$. **b**, Fraction of clones with long CDR-H3s, that is, CDR-H3s longer than 19 amino acids. Data for IgG4 and IgE are shown in Supplementary Fig. 2b. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$. **c**, CDR-H3 lengths of unproductive *IGH* rearrangements from a genomic DNA template. Because these rearrangements are not expressed as a functional protein, they provide V(D)J recombination data free of the effects of protein-based selection. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$.

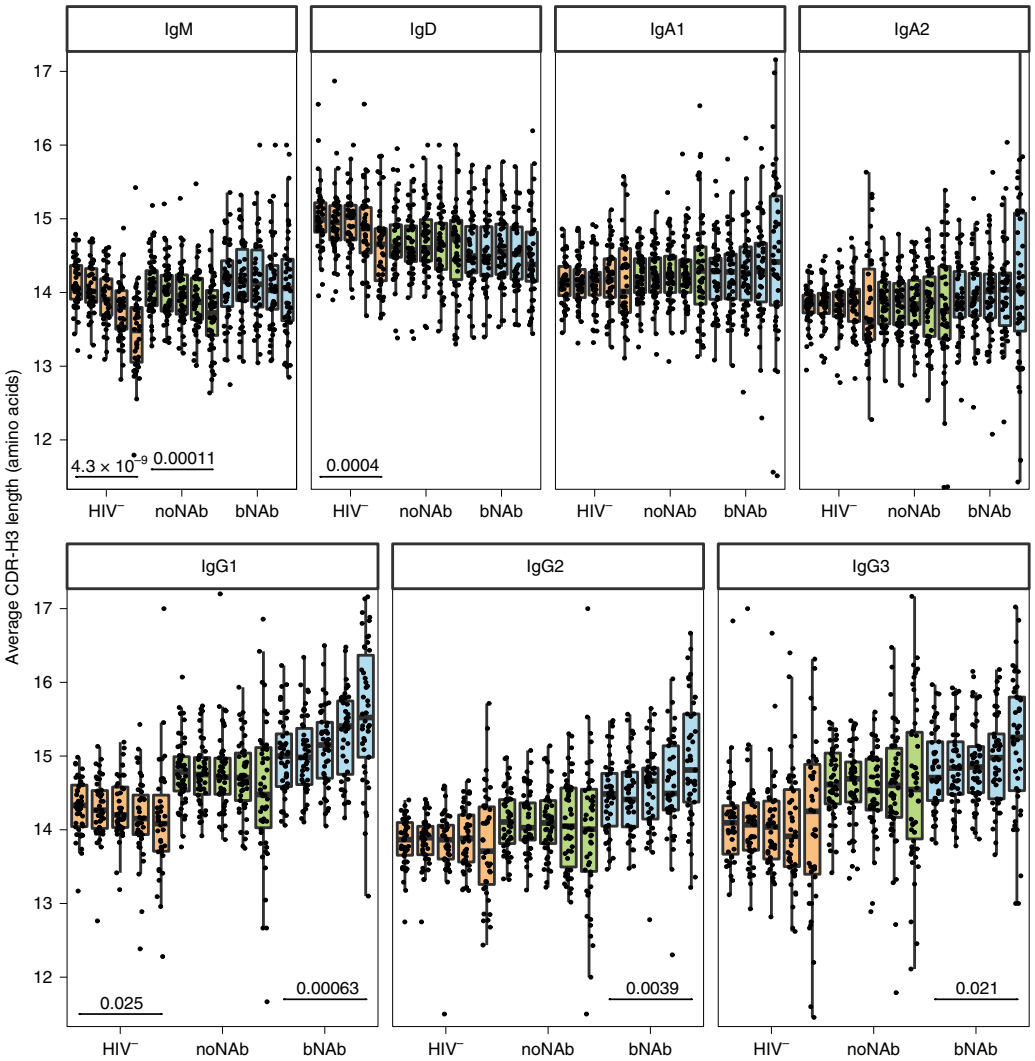

Fig. 3 | Average heavy-chain CDR-H3 lengths in *IGH* repertoire fractions with increasing frequencies of SHM. Within each group, the columns from left to right indicate the mean CDR-H3 lengths in the 30%, 20%, 10%, 5% or 1% most-mutated sequences in each individual, respectively. Data for the rare isotypes IgG4 and IgE are shown in Supplementary Fig. 2c. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$.

frequencies was not limited by CD4+ T cell or memory CD4+ T_{FH} cell frequencies in the bNAb individuals.

There were no significant relationships between the frequency of regulatory CD4+ T_{reg} cells (CD25+Foxp3+CD4+ cells) and extreme (99th percentile) IgG SHM frequencies (Fig. 4a) or CDR-H3 lengths in any subject group (Fig. 4b). Circulating T follicular regulatory (T_{FR}) cells (CD25+Foxp3+ within CXCR5+CD4+ T cells) showed no correlation with CDR-H3 lengths (Supplementary Fig. 4b), and only a weak negative correlation with the 99th percentile IgG SHM frequencies in both bNAb and noNAb individuals (Supplementary Fig. 4c). The 99th percentile IgG SHM frequencies had no relationship to the proportion of T_{reg} cells expressing CTLA-4 (Fig. 4c), but there was a significant negative correlation between the frequency of CTLA-4+ T_{reg} cells and mean CDR-H3 lengths in bNAb individuals ($P=0.0016$; Fig. 4d). CTLA-4 is an activation marker in CD4+ T cells, and plays an important role in the control of humoral responses by CD4+ T_{reg} and T_{FR} cells[12]. HIV viral loads were usually higher in bNAb individuals than noNAb individuals (Supplementary Fig. 3d). While this analysis cannot prove causation, and it is possible that these T cell populations are coordinately regulated with the B cell repertoires in study participants rather than directly affecting the B cells, the bNAb patients with the lowest frequencies of CTLA-4+ T_{reg} cells had the longest CDR-H3s. The length of the longest CDR-H3s (99th percentile) in bNAb individuals showed a similar negative correlation with CTLA-4+ T_{reg} cell frequencies. We observed a weak but nonsignificant correlation between CTLA-4+ T_{reg} cell frequencies and patient viral loads in bNAb individuals (Supplementary Fig. 4d).

bNAb *IGHV4-34* usage and selection against autoreactivity. Most known bNAbs use a restricted set of IGHV gene segments. We examined IGHV segment usage frequency in the 1% most-mutated clones of each individual, focusing on segments such as *IGHV4-34*, which possesses natural autoreactivity in the unmutated state, to determine whether there were differences that might predispose them to forming bNAbs (Fig. 5a–c and Supplementary Fig. 5a). After Bonferroni multiple hypothesis correction, we identified a statistically significant difference in the usage of *IGHV4-34* between groups, in IgG, IgM and IgD antibodies, but not in unproductive *IGH* rearrangements or in the IgA compartments. HIV-infected

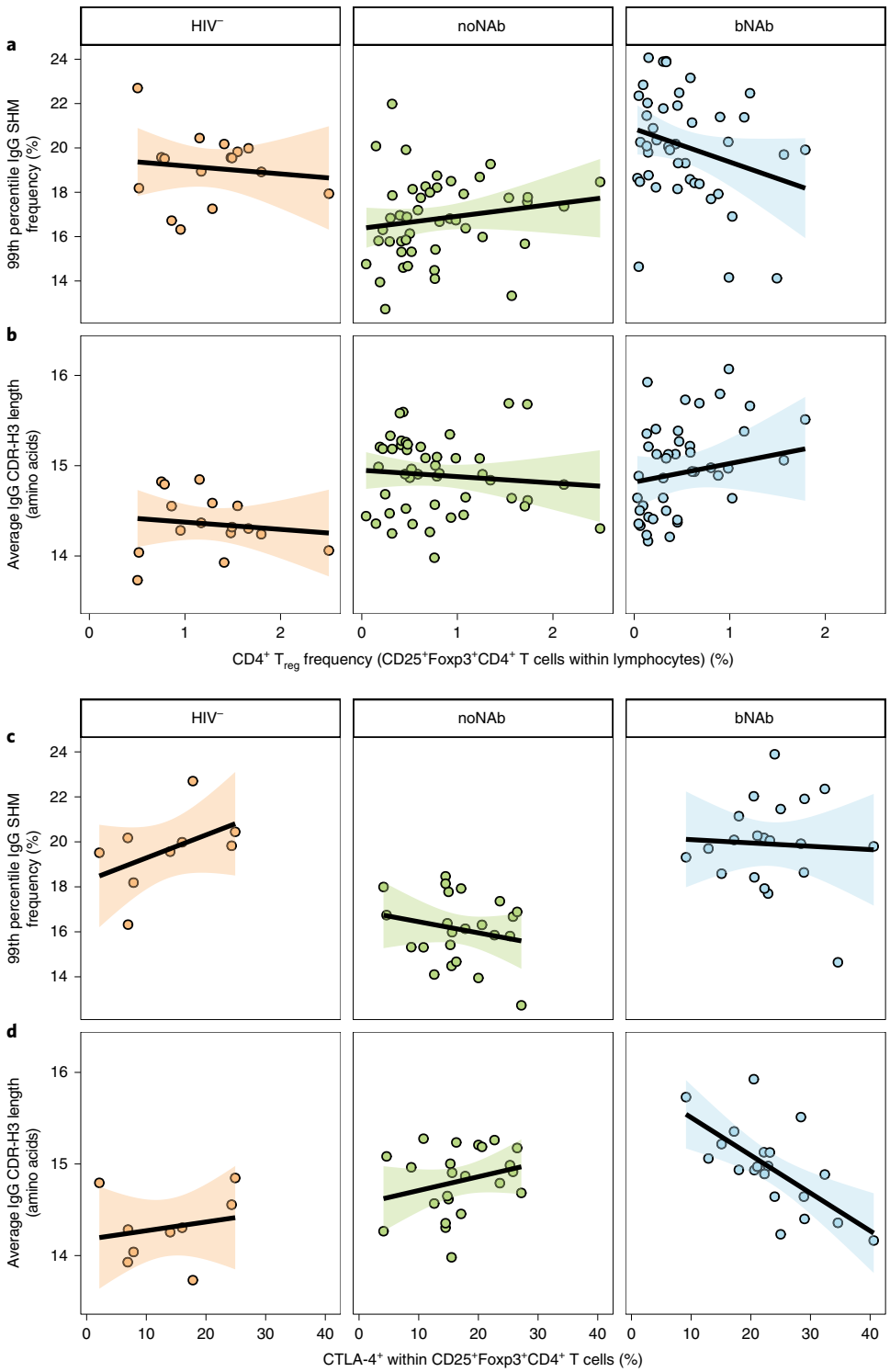

Fig. 4 | Correlation of Ig heavy-chain features and T cell subsets. a, 99th percentile IgG SHM frequencies and percentage of CD4$^+$ T_{reg} cells (CD25$^+$Foxp3$^+$CD4$^+$ cells) within lymphocytes in individuals with no HIV infection, or HIV-infected noNAb or bNAb individuals. Shaded regions in all panels are the 95% confidence interval from the linear regression model. **b**, Average IgG CDR-H3 length and the percentage of T_{reg} cells within lymphocytes. Shaded regions in all panels are the 95% confidence intervals from the linear regression model. **c**, Percentage of CD25$^+$Foxp3$^+$CD4$^+$ T cells expressing immune-checkpoint protein CTLA-4 and the 99th percentile SHM frequency in IgG. Shaded regions in all panels are the 95% confidence intervals from the linear regression model. **d**, Percentage of CD25$^+$Foxp3$^+$CD4$^+$ T cells expressing CTLA-4 showed a negative correlation with heavy-chain CDR-H3 length in bNAb individuals (slope = −0.04125, s.e.m. = 0.01108, P = 0.0016 for linear regression model). Shaded regions in all panels are the 95% confidence intervals from the linear regression model.

individuals used *IGHV4-34* more than uninfected controls and bNAb individuals used *IGHV4-34* more frequently than other groups. This difference was most evident in the most-mutated clones in each individual (Fig. 5c and Supplementary Fig. 5b).

B cells expressing antibodies using *IGHV4-34* convey information about the immune system's control of autoreactive clones. Unmutated *IGHV4-34* encodes a tyrosine residue at codon 26 (in the IMGT numbering system) at the junction between framework

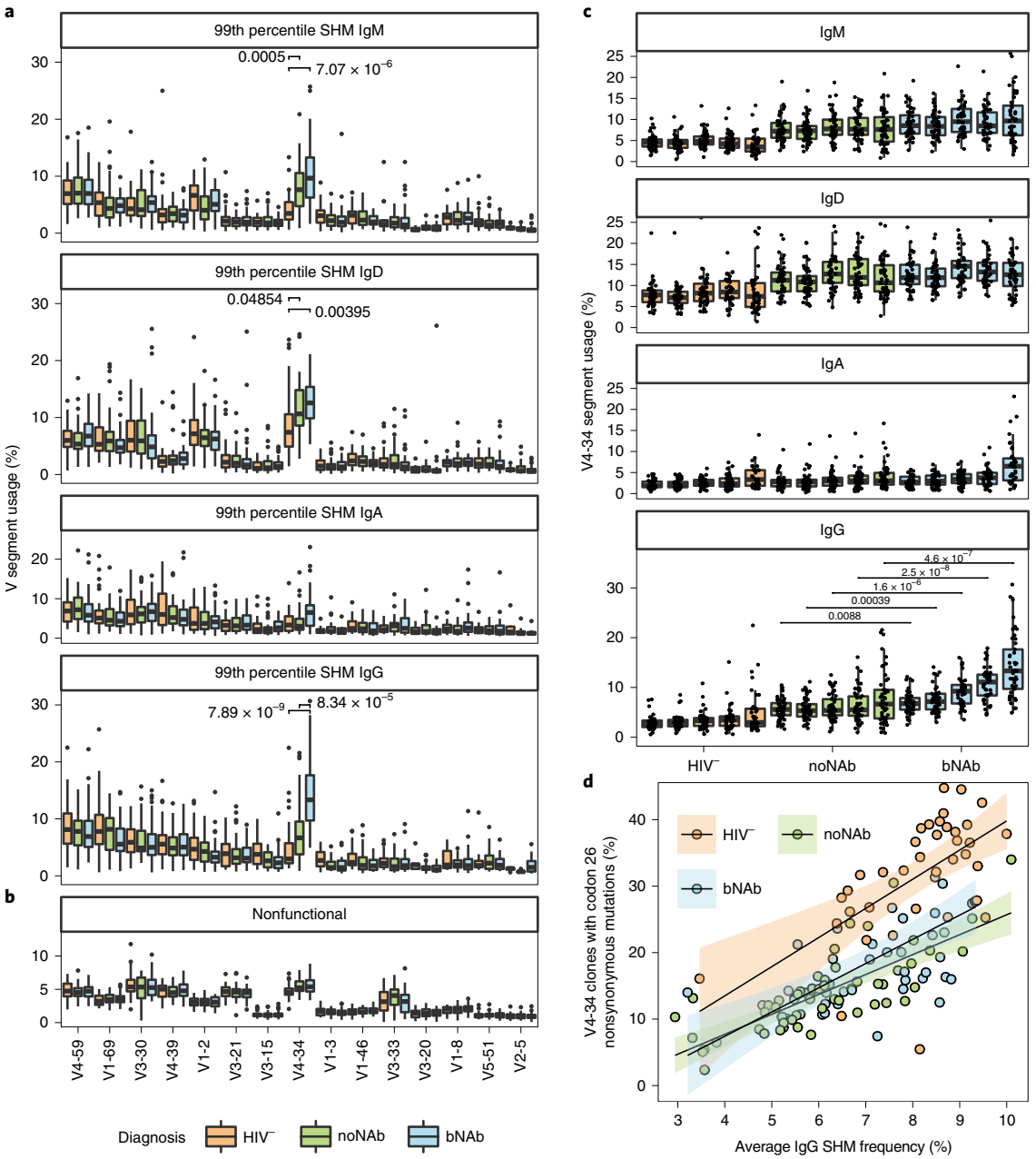

Fig. 5 | IGHV segment usage and SHM analysis of clones using *IGHV4-34*. **a**, Differences in IGHV segment usage in the 1% most-mutated sequences. The IGHV segments used by known HIV bNAbs are shown, as well as the known autoreactivity-associated *IGHV4-34*. After Bonferroni multiple hypothesis correction, statistically significant differences were those in *IGHV4-34* segment usage in IgG, IgM and IgD. Germline *IGHV4-34* is known to have a hydrophobic patch in the vicinity of the amino acid encoded by codon 26 that contributes to B cell receptor autoreactivity. Data for the low abundance isotype IgE are shown in Supplementary Fig. 5a. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$. **b**, Unproductive, and thus not subject to protein-based selection, *IGH* rearrangements showed no bias in *IGHV4-34* usage. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$. **c**, *IGHV4-34* gene segment usage in the 30%, 20%, 10%, 5% or 1% most-mutated clones in each individual is shown for IgM, IgD, IgA and IgG. IgE isotype data are shown in Supplementary Fig. 5b. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$. **d**, The fraction of clones using *IGHV4-34* with hydrophobic patch SHM at IMGT codon 26 plotted against average per person IgG SHM frequency. Shaded regions in all panels are the 95% confidence intervals from the linear regression model.

1 and CDR-H1 that contributes to a hydrophobic patch that binds poly-*N*-acetyllactosamine carbohydrate self-antigens (known as 'big I' and 'little i' antigens) on red blood cells[13,14]. SHM that alters codon 26 can remove self-reactivity and is very common in B cells that use *IGHV4-34* (ref. [13]). HIV-infected individuals have decreased codon 26 mutation frequencies compared to uninfected individuals, when

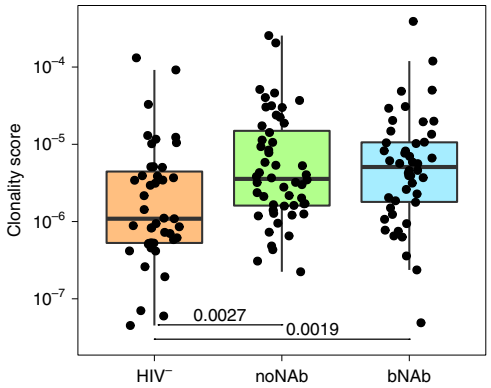

Fig. 6 | Clonality analysis. Clonality scores in *IGH* replicate sequencing datasets generated from a genomic DNA template (see Methods) from individuals with and without HIV infection. *P* values for comparisons between groups are for the two-sided Wilcoxon–Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$.

normalized for total antibody mutation frequencies, suggesting that HIV infection is associated with a permissive state in which potentially autoreactive clones that would otherwise be eliminated can persist (Fig. 5d).

Increased B cell clonal expansion in HIV infection. Previous studies with relatively small cohorts have reported lower BCR sequence diversity in individuals with HIV as compared to uninfected controls; however, it can be challenging to distinguish between true decreases in species richness versus the presence of some larger clones[15]. By sequencing replicate *IGH* libraries from six independent genomic DNA (gDNA) template aliquots for each individual, we calculated a normalized 'clonality score' equivalent to the probability that randomly selected reads from two replicate *IGH* libraries from a sample will be from the same B cell clone. Unrearranged IGHV gene segments do not appear in these gDNA libraries as they are too far upstream of the IGHJ gene segments to give PCR amplicons. Individuals who were HIV-infected had elevated clonality compared to individuals who were HIV-uninfected (Fig. 6). In all groups, SHM frequency was increased in expanded clones compared to those showing no evidence of clonal expansion (Supplementary Fig. 6). There was no difference between bNAb and noNAb individuals in this measure, even when differences in $CD4^+$ T cell counts and viral loads were included in the analysis (Supplementary Fig. 3 and ref. [11]).

***IGH* convergence and repertoire features in HIV infection.** We hypothesized that *IGH* repertoire differences as well as antigen-driven selection would give rise to distinctive common or 'convergent' rearrangements distinguishing individuals who were HIV-infected from those who were HIV-uninfected and bNAb from noNAb individuals. Such convergent sequences could be useful for diagnostic purposes, identifying B cell repertoires that have been shaped by exposure to HIV and those that give rise to bNAb antibody types. Some convergent sequences would be expected to be HIV-specific, while others could be polyreactive, autoreactive or specific for other antigens. We searched for shared *IGH* sequences that were informative for distinguishing individuals who were HIV-infected from those who were HIV-uninfected. Individuals were scored based on the number of such shared sequences that they demonstrated. Receiver operating characteristic (ROC) plots showed that this classification scheme was effective for distinguishing individuals with HIV infection from those who were uninfected,

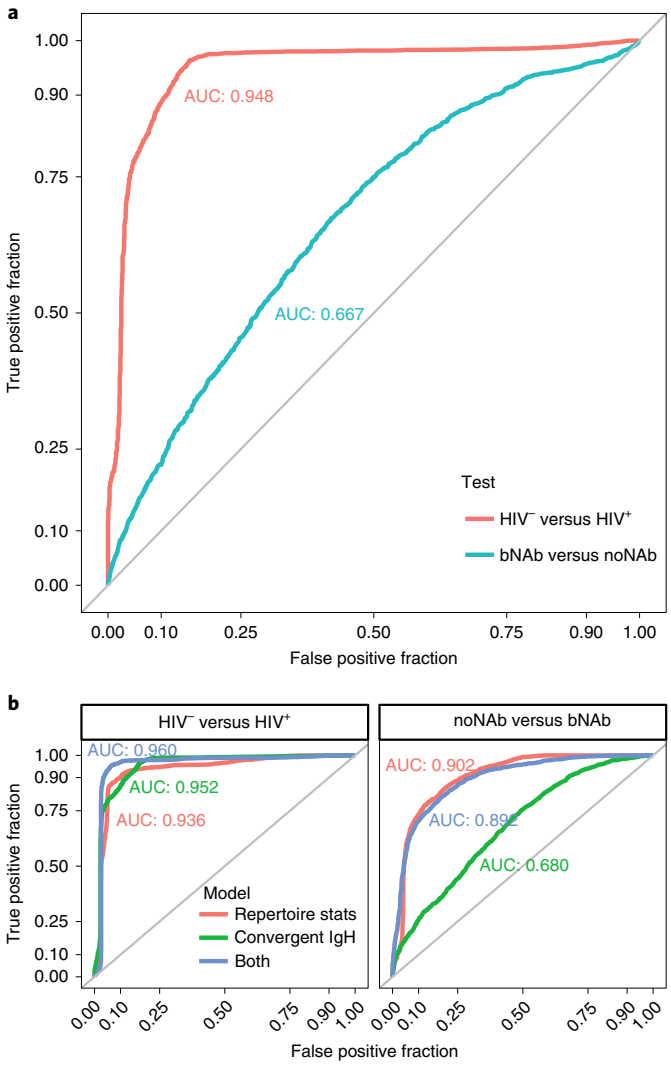

Fig. 7 | Prediction of HIV and bNAb status. a, ROC curve for prediction of HIV status (HIV-infected versus HIV-uninfected) (red curve) and bNAb versus noNAb status (blue) using the number of predictive *IGH* sequence clusters found in each individual. Fourfold cross validation was used to prevent overfitting: predictive clusters were selected from three folds and were tested on the held-out fold. Random partitioning of the data into folds was performed 40 times. Curves shown are the average over all 160 partitions. AUC, a summary of overall predictive power, is shown for each classification problem. **b**, Predictive accuracy for random forest models that use overall repertoire statistics (red), predictive *IGH* clusters (green) or both (blue) as features for the prediction of HIV infection status (left) and bNAb versus noNAb status (right). The numbers of participants analyzed were: HIV-uninfected, $n=43$; noNAb, $n=50$; bNAb, $n=46$.

achieving a >90% true positive rate with a <10% false positive rate, with an area under the curve (AUC) of 0.95 (Fig. 7a). The same approach, attempting to distinguish bNAb individuals from noNAb individuals performed less well, with an AUC of 0.72, but showed some discrimination between these groups. Selection of convergent clonotypes for these classifications was subjected to cross validation to minimize overfitting of the classifier, but we did not empirically test the binding specificity of the convergent antibodies identified. When we searched for the sequences of known monoclonal bNAbs that have been previously isolated from a few of the individuals in the bNAb patient cohort in this study, using a threshold of 80% nucleotide identity in CDR-H3, equal CDR-H3 length and the same

Table 1 | Summary of bNAb and noNAb heavy-chain repertoire features

	noNAb	bNAb
Average SHM frequency		
IgG subtypes	Lower than HIV-uninfected	Lower than HIV-uninfected, but higher than noNAb
Extreme SHM frequency (99th percentile)		
IgG subtypes	Lower than HIV-uninfected	Higher than HIV-uninfected
CDR-H3 length		
IgD/unselected	Shorter than HIV-uninfected	Shorter than HIV-uninfected
IgG subtypes	Longer than HIV-uninfected	Longer than HIV-uninfected
CDR-H3 length at increasing SHM frequency		
IgM	Nondecreasing, unlike HIV-uninfected	Nondecreasing, unlike HIV-uninfected
IgG subtypes	Comparable to HIV-uninfected	Increasing
IGHV segment usage		
IGHV4-34 in IgG	Higher than HIV-uninfected	Higher than HIV-uninfected and noNAb, increasing with SHM
'Redemption' of *IGHV4-34* clones	Less than HIV-uninfected	Less than HIV-uninfected
Clonality	Higher than HIV-uninfected	Higher than HIV-uninfected
$CD4^+$ T cell frequencies	Less than HIV-uninfected	Less than noNAb and HIV-uninfected
Effect on SHM frequency	Increasing with $CD4^+$ T cell count	Unaffected
Viral load	–	Higher than noNAb
Effect on extreme SHM frequencies	Decreases extreme SHM	Unaffected

IGHV and IGHJ gene segment usage, we identified bNAbs CH01, CH02, CH03 and CH04 in the *IGH* repertoire data from individual 0219; bNAbs DH511.1/2/3/6/7P/8P and DH511.4/5/10P in individual CH0210; and CH105 in individual CH0505 (refs. [9,16,17]). At a CDR-H3 identity threshold of 75%, we also identified bNAbs CH103, CH104 and CH106 in individual CH505 (ref. [17]). We did not find convergent antibodies to the known bNAb lineages in other broad-neutralizing individuals at these CDR-H3 identity thresholds. We infer that the antibody lineages responsible for HIV neutralizing breadth in these other individuals do not share high sequence identity with known bNAbs.

To further evaluate the ability of *IGH* repertoire data to classify HIV infection status and neutralizing breadth, we combined the presence or absence of convergent *IGH* sequences with general repertoire statistics listed in Table 1: mean IgG mutation level; 99th percentile IgG mutation level; mean IgG, IgM and IgD CDR-H3 lengths; fraction of 'long' (>19 amino acids) IgG, IgM and IgD CDR-H3s; mean nonfunctional CDR-H3 length from gDNA libraries; and usage of *IGHV4-34* in the 1% most-mutated IgG sequences, in a random forest model (Fig. 7b) with the same cross-validation scheme as above[18]. While this more complex model yielded only a small improvement on the already accurate classification of HIV infection status (AUC of 0.96), the addition of repertoire statistics markedly improved the categorization of bNAb and noNAb individuals (AUC of 0.89).

Cytomegalovirus minimally alters *IGH* repertoires. Finally, we assessed whether the *IGH* repertoire features differing with HIV infection state and neutralizing breadth were specific to HIV, by considering an unrelated chronic viral infection, human cytomegalovirus (CMV), a herpesvirus causing a lifelong persistent infection that is highly prevalent in many populations and almost ubiquitous in HIV-infected cohorts. We sequenced *IGH* repertoires in HIV-seronegative blood donors, of whom $n=52$ were CMV-seropositive and $n=61$ were CMV-seronegative. Individuals who were CMV-seropositive compared to those who were CMV-seronegative showed none of the *IGH* repertoire features seen in individuals who were infected with HIV and bNAb individuals, such as overall SHM differences in IgG isotypes (Supplementary Fig. 7a), 99th percentile SHM values in IgG isotypes (Supplementary Fig. 7b), CDR-H3 lengths (Supplementary Fig. 7c) or CDR-H3 fraction longer than 19 amino acids (Supplementary Fig. 7d), CDR-H3 lengths in progressively higher-SHM fractions of the IgG repertoires (Supplementary Fig. 7e) or *IGHV4-34* usage (Supplementary Fig. 7f). These results indicate that the HIV-associated or bNAb-associated *IGH* repertoire features we observed in the HIV-infected indviduals are not a generic consequence of all chronic viral infections.

Discussion

Many human immune responses show heterogeneity, such as when memory of previous exposures shapes responses to new influenza viral strains[19] or host genetic background influences hepatitis B vaccination responses[20,21]. It has been unclear whether an individual's HIV neutralizing antibody breadth depends on rare stochastic events giving rise to very few bNAb clones or whether bNAb producers have systematic BCR repertoire differences leading to many potential bNAb lineages. Our data favor the latter possibility, showing numerous lineages with high SHM and long CDR-H3s in bNAb individuals. Notably, the most highly mutated fraction of IgGs in bNAb individuals have higher SHM frequencies than those of noNAb or uninfected individuals and number in the hundreds to thousands per individual, rather than representing rare outlier clones.

Human pre- or pro-B cells with long CDR-H3s are removed from the repertoire before the naive B cell stage, and antigen-experienced B cell populations are further selected to have shorter CDR-H3s compared to the naive repertoire[5,22]. We found that CDR-H3s in unselected, unproductive *IGH* in HIV-infected individuals were shorter than in uninfected controls, similarly to findings in common variable immune deficiency[22] (CVID) and potentially reflecting altered V(D)J recombination in the bone marrow. In contrast, CDR-H3s in productive IgG antibodies were longer in HIV-infected individuals compared to uninfected individuals, showing a strong selection for longer CDR-H3s. However, only the bNAb individuals showed selection favoring long CDR-H3 in the most highly mutated fractions of their antibody repertoires.

These findings do not exclude a major role for the course of the viral infection and the number and variety of HIV antigenic variants that arise in each patient in driving antibody repertoire phenotypes. bNAb producers had higher viral loads, suggesting that greater antigen drive could contribute to bNAb lineage development[23]. Previous estimates of HIV broadly neutralizing or gp120$^+$ B cells in bNAb individuals ranged from 0.003–0.1% of memory IgG$^+$ B cells, indicating that these cells are unlikely to account for the differences in SHM and CDR-H3 values that we observed between bNAb and noNAb individuals[7–10]. However, gp120$^+$ and bNAb B cell measurements may underestimate frequencies of cells specific for other unknown HIV antigenic variants in participants.

From T cell phenotyping data previously published[11], we identified T cell correlates of BCR repertoire features in individuals who were infected with HIV. SHM frequencies in bNAb individuals were not correlated with total CD4$^+$ T cell or memory T_{FH} frequencies in the blood. CTLA-4$^+$ T_{reg} frequencies correlated with CDR-H3 lengths in bNAb producers, opening the possibility that these cells may be involved, either directly or indirectly, in controlling CDR-H3 lengths in the maturing BCR repertoire of bNAb individuals. Alternatively, increased CTLA-4 expression in T_{reg} cells in this context may primarily indicate that higher levels of T_{reg} activation are present under conditions that also lead to shorter CDR-H3s. Transient modulation of immune checkpoints such as the CTLA-4 pathway during HIV vaccination has been proposed as a new approach to enable HIV-infected individuals to generate bNAbs. Any attempt to test this idea would need to include strategies to minimize autoimmune side effects, such as those seen in cancer immunotherapy trials[24,25]. Our analysis raises the hypothesis that CTLA-4 blockade could lead to HIV-specific antibodies with longer CDR-H3s.

Antibodies using unmutated *IGHV4-34* are implicated in cold agglutinin disease, an autoimmune hemolytic anemia associated with HIV infection and other immune disorders[26,27]. IgG in HIV-infected individuals used *IGHV4-34* at higher frequencies than in uninfected controls, particularly in bNAb individuals. SHM of codon 26 in *IGHV4-34* decreases this autoreactivity, a process termed 'clonal redemption' and is seen at high frequencies in HIV-infected individuals[13]. Serum reactivity with an anti-idiotype antibody that binds unmutated *IGHV4-34* has been loosely correlated with broad HIV neutralization[28]. We found that bNAb producers show significantly lower SHM frequencies of codon 26 in *IGHV4-34* gene segments compared to uninfected controls, suggesting that the lack of selection against autoreactive antibodies is associated with the development of HIV neutralizing breadth, consistent with previous reports of bNAb producers showing various autoantibodies in their sera[29,30].

In summary, our evidence points to the loss of normal selection against B cells expressing IgG with long CDR-H3 regions and high SHM frequencies, and a general loss of selection against autoreactive antibodies, in the development of HIV neutralizing breadth. CTLA-4$^+$ T_{reg} cells show a strong negative correlation with CDR-H3 lengths in the BCR repertoires of bNAb producers, suggesting that this cell population, or CTLA-4 pathway manipulation more generally, could be targets for future strategies to improve neutralizing breadth generated by HIV vaccination. A major topic of current study is the way that differences in the HIV viral populations, such as the diversity, concentrations or epitope features of viral subpopulations, may affect the B cell phenotypes we observe. HIV-infected individuals produce characteristic 'convergent' clones with highly similar sequences detected in two or more individuals that are rarely detected in healthy people. These convergent clones indicate whether an individual is infected by HIV, although we have not determined whether all such clones are specific to HIV. We also find other clones that more weakly classify an individual's HIV neutralizing breadth. The accuracy of bNAb classification is enhanced in models that incorporate additional repertoire characteristics, but convergent clones alone are sufficient to predict HIV infection status accurately. As sequence data accumulate for human antibodies specific for various pathogens, vaccines and other antigens, it may become possible to use antibody repertoire data to predict an individual's responses to future immune stimuli and to improve vaccination strategies for challenging pathogens such as HIV.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41590-019-0581-0.

Received: 28 November 2018; Accepted: 12 December 2019;
Published online: 20 January 2020

References

1. Hraber, P. et al. Prevalence of broadly neutralizing antibody responses during chronic HIV-1 infection. *AIDS* **28**, 163–169 (2014).
2. Mascola, J. R. & Haynes, B. F. HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunol. Rev.* **254**, 225–244 (2013).
3. Scheepers, C. et al. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J. Immunol.* **194**, 4371–4378 (2015).
4. Kepler, T. B. et al. Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies. *Cell Host Microbe* **16**, 304–313 (2014).
5. Wardemann, H. et al. Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
6. Hicar, M. D. et al. Low frequency of broadly neutralizing HIV antibodies during chronic infection even in quaternary epitope targeting antibodies containing large numbers of somatic mutations. *Mol. Immunol.* **70**, 94–103 (2016).
7. Bonsignori, M. et al. Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. *Cell* **165**, 449–463 (2016).
8. Bonsignori, M. et al. Staged induction of HIV-1 glycan-dependent broadly neutralizing antibodies. *Sci. Transl. Med.* **9**, eaai7514 (2017).
9. Bonsignori, M. et al. Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. *J. Virol.* **85**, 9998–10009 (2011).
10. Bonsignori, M. et al. Two distinct broadly neutralizing antibody specificities of different clonal lineages in a single HIV-1-infected donor: implications for vaccine design. *J. Virol.* **86**, 4688–4692 (2012).
11. Moody, M. A. et al. Immune perturbations in HIV-1-infected individuals who make broadly neutralizing antibodies. *Sci. Immunol.* **1**, aag0851 (2016).
12. Sage, P. T., Paterson, A. M., Lovitch, S. B. & Sharpe, A. H. The coinhibitory receptor CTLA-4 controls B cell responses by modulating T follicular helper, T follicular regulatory, and T regulatory cells. *Immunity* **41**, 1026–1039 (2014).
13. Reed, J. H., Jackson, J., Christ, D. & Goodnow, C. C. Clonal redemption of autoantibodies by somatic hypermutation away from self-reactivity during human immunization. *J. Exp. Med.* **213**, 1255–1265 (2016).
14. Li, Y., Spellerberg, M. B., Stevenson, F. K., Capra, D. J. & Potter, K. N. The I binding specificity of human VH4-34 (VH4-21) encoded antibodies is determined by both VH Framework Region 1 and complementarity determining region 3. *J. Mol. Biol.* **256**, 577–589 (1996).
15. Hoehn, K. B. et al. Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. *Phil. Trans. R. Soc. Lond. B* **370**, 20140241 (2015).
16. Williams, L. D. et al. Potent and broad HIV-neutralizing antibodies in memory B cells and plasma. *Sci. Immunol.* **2**, eaal2200 (2017).
17. Liao, H.-X. et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476 (2013).
18. Ho, T. K. Random decision forests. In *Proc. International Conference on Document Analysis and Recognition* ICDAR (eds Kavanaugh, M. & Storms, P.) 278–282 (IEEE Computer Society Press, 1995).
19. Fonville, J. M. et al. Antibody landscapes after influenza virus infection or vaccination. *Science* **346**, 996–1000 (2014).
20. Hennig, B. J. et al. Host genetic factors and vaccine-induced immunity to hepatitis B virus infection. *PLoS One* **3**, e1898 (2008).
21. Höhler, T. et al. Differential genetic determination of immune responsiveness to hepatitis B surface antigen and to hepatitis A virus: a vaccination study in twins. *Lancet* **360**, 991–995 (2002).

22. Roskin, K. M. et al. IgH sequences in common variable immune deficiency reveal altered B cell development and selection. *Sci. Transl. Med.* **7**, 302ra135 (2015).
23. Gray, E. S. et al. The neutralization breadth of HIV-1 develops incrementally over four years and is associated with $CD4^+$ T cell decline and high viral load during acute infection. *J. Virol.* **85**, 4828–4840 (2011).
24. Haynes, B. F. et al. HIV–host interactions: implications for vaccine design. *Cell Host Microbe* **19**, 292–303 (2016).
25. Bertrand, A., Kostine, M., Barnetche, T., Truchetet, M.-E. & Schaeverbeke, T. Immune related adverse events associated with anti-CTLA-4 antibodies: systematic review and meta-analysis. *BMC Med.* **13**, 211 (2015).
26. Ciaffoni, S. et al. Presence and significance of cold agglutinins in patients with HIV infection. *Haematologica* **77**, 233–236 (1992).
27. Pruzanski, W., Roelcke, D., Donnelly, E. & Lui, L. C. Persistent cold agglutinins in AIDS and related disorders. *Acta Haematol.* **75**, 171–173 (1986).
28. Kobie, J. J. et al. 9G4 autoreactivity is increased in HIV-infected patients and correlates with HIV broadly neutralizing serum activity. *PLoS One* **7**, e35356 (2012).
29. Haynes, B. F. et al. Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. *Science* **308**, 1906–1908 (2005).
30. Liu, M. et al. Polyreactivity and autoreactivity among HIV-1 antibodies. *J. Virol.* **89**, 784–798 (2015).

Methods

Patient cohorts. Participants were recruited in a multisite study of acute and chronic HIV-1 infection that included an uninfected control arm, as previously reported[11]. The study was approved by the Duke Medicine Institutional Review Board as well as the ethics boards of the local sites. All participants provided informed consent. On average, participants in the chronic arm of the study were followed for >400 d. Patients defined clinically as virus controllers were excluded from the study. Participants who were infected with HIV and uninfected controls were predominantly African and regions of origin of these individuals are described[11]. Matching of nonneutralizing HIV-infected controls to broadly neutralizing HIV-infected individuals was carried out using propensity score-matching with age, sex and country of origin as input variables, and resulted in a nonneutralizing set of control individuals that did not significantly differ in age, sex or country of origin compared to the broadly neutralizing individuals.

bNAb patient status was determined by serological measurement of the reciprocal inhibitory dilution for 50% reduction of infection (ID_{50}) for a panel of 12 HIV isolates in the TZM-bl pseudovirus neutralization assay, a measure of HIV neutralization based on decreases in expression of a Tat-regulated firefly luciferase (Luc) reporter gene after one round of infection with Env-pseudotyped viruses[31]. To be categorized as a bNAb individual, the participant's principal component score (PC1) from the neutralization panel data was required to exceed a value of 3 (representing the top 14% of neutralizers) and their ID_{50} score was required to be at least 95 for 7 of the 12 viral isolates in the panel[11]. All nonneutralizing HIV-infected controls had PC1 scores <0.1 and did not exceed an ID_{50} threshold of 95 for any HIV isolate in the serological panel.

For comparison of CMV-seropositive and CMV-seronegative participant *IGH* repertoires in Supplementary Fig. 7, a separate cohort of 113 healthy adult blood donors from the Stanford Blood Center was used. All participants provided informed consent. These participants were negative for pathogen-screening testing for HIV, HCV, HBV, HTLV, West Nile virus, TPPA (syphilis) and *Trypanosoma cruzi*, and had no symptoms of acute illness or diseases, including malignancies. Participant ages ranged from 17–87 years (median: 52 years and mean: 49 years). Sequence data from these participants are reported by Nielsen et al.[32].

T cell phenotyping by flow cytometry. Flow cytometric analysis of T cell populations in peripheral blood mononuclear cells (PBMCs) was carried out as previously reported[11]. For analysis of circulating resting memory T_{FH} frequencies, PBMCs were surface stained with CXCR3 allophycocyanin (APC) (1C6) and CXCR5 Alexa Fluor (AF)488 (RF8B2) (both from BD Biosciences) at 37 °C for 10 min, then washed and surface stained with CD3 phycoerythrin (PE)-Texas Red (S4.1; Life Technologies); CD4 APC-H7 (SK3), CD45RO PE-Cy7 (UCHL1), CCR6 PE (11A9) and PD-1 Brilliant Violet (BV)421 (EH12.1) (all from BD Biosciences), plus a biotinylated antibody to ICOS (ISA-3; eBioscience) at 20 °C for 15 min; then washed and stained with fluorescent-conjugated streptavidin Peridinin chlorophyll protein (PerCP) (BD Biosciences) and Live/Dead Fixable Aqua Dead Cell Stain (Life Technologies) at 20 °C for 10 min, before washing and fixation with 4% paraformaldehyde for 15 min.

For analysis of circulating $CD4^+$ T_{reg} and T_{FR} cells, monoclonal antibodies used for surface staining included CD3 PE-Texas Red (S4.1; Life Technologies), CD4 APC-H7 (SK3), CD25 PE-Cy7 (2A3), CXCR5 AF488 (RF8B2) and PD-1 BV421(EH12.1) (all from BD Biosciences). PBMCs were stained for 15 min at 37 °C before washing and staining for 10 min at 20 °C with Live/Dead Fixable Aqua Dead Cell Stain (Life Technologies). The Foxp3 Fix/Perm kit (eBioscience) was then used to fix and permeabilize cells at 20 °C for 15 min before staining with Foxp3 APC (PCH101) and CTLA-4 PerCP-eFluor710 (14D3) (both from eBioscience) at 20 °C for 45 min, after which they were washed twice with permeabilization buffer.

Flow cytometry data were acquired with a Cyan ADP Analyzer (Beckman Coulter) or LSR Fortessa X-20 Analyzer (BD Biosciences) with standard filter sets and data were analyzed using FlowJo software (v.8.8.6; Treestar). The T cell flow cytometry data were previously published[11] and details of the gating strategies used for identification of cell subsets are described there.

Isolation of nucleic acids from peripheral blood mononuclear cells. Heparinized whole-blood samples (5 to 8 ml) were collected and PBMCs were isolated by centrifugation of diluted blood over Histopaque-1077 (Sigma-Aldrich) or Ficoll-Paque (GE Healthcare). gDNA was isolated via column purification (Qiagen), magnetic bead-based isolation (MagNA Pure, Roche Life Science) or centrifugation-based purification (ArchivePure, 5 Prime). RNA was isolated using the AllPrep DNA/RNA kit (Qiagen).

PCR amplification of cDNA for *IGH* library sequencing. For RNA-derived antibody isotype libraries, cDNA was synthesized with SuperScript II reverse transcriptase (Thermo Fisher Scientific) with random hexamer primers. Templates were amplified by PCR using isotype-specific primers located in the first exon of the constant (C1) regions (Supplementary Table 2) and BIOMED-2 IGHV primers in the framework 1 (FR1) region (Supplementary Table 4)[22,33]. These primers also encoded approximately half of the Illumina linker sequences needed for cluster generation and sequencing on the MiSeq instrument. Sample identity was encoded by eight-nucleotide multiplex identifier barcodes in each primer. For Illumina cluster recognition, four randomized nucleotides were encoded in the primers immediately after the Illumina linker sequence in the constant region primers. Each antibody isotype for each sample was amplified in a separate PCR reaction to prevent formation of cross-isotype chimeric PCR products. PCR was carried out with AmpliTaq Gold (Thermo Fisher Scientific) following the manufacturer's instructions using a program of: 94 °C for 7 min, 35 cycles of (94 °C for 30 s, 58 °C for 45 s and 72 °C for 120 s) and final extension at 72 °C for 10 min. A second PCR step with primers listed in Supplementary Table 5 was used to add the remaining portion of the Illumina linkers to the amplicons and was carried out with the Qiagen Multiplex PCR kit (Qiagen) according to the manufacturer's instructions, using 0.4 µl of the first PCR product as template in a 30-µl reaction. The PCR program for the second PCR step was 94 °C for 15 min, 12 cycles of (94 °C for 30 s, 60 °C for 45 s and 72 °C for 90 s) and final extension at 72 °C for 10 min. The products of each PCR reaction were quantitated, pooled in equimolar amounts, electrophoresed on agarose gels and gel extracted with QIAquick kits (Qiagen). High-throughput sequencing was performed with an Illumina MiSeq instrument using 600-cycle sequencing kits.

PCR amplification of gDNA IGH libraries and sequencing. For gDNA-templated libraries, PCR reactions were prepared from 100 ng gDNA aliquots to generate three independent libraries per sample. Multiplexed primers to the IGHJ (Supplementary Table 3) and FR1 or FR2 framework regions, as per the BIOMED-2 design (Supplementary Table 4), were used, with additional sequence representing the first half of the Illumina linkers[22,33]. Eight-nucleotide barcode sequences were included in the primers to indicate sample and replicate identity and four randomized bases were included upstream of the barcodes on the IGHJ primer for Illumina clustering. PCR was performed with AmpliTaq Gold polymerase as per the manufacturer's instructions. The PCR program was as follows: 94 °C for 5 min, 35 cycles of (94 °C for 30 s, 60 °C for 45 s and 72 °C for 90 s) and final extension at 72 °C for 10 min. A second PCR reaction to complete the Illumina linker sequence was carried out using the protocol in the previous section with the primers from Supplementary Table 5. Libraries were pooled in equimolar amounts, gel extracted and sequenced in two runs of the Illumina MiSeq instrument.

IGH sequence processing and analysis. Sequences with exact matches to IGHV and IGHJ (gDNA) or isotype constant region (for cDNA-derived data) barcodes were assigned to the corresponding samples and replicate libraries. Barcodes and primer sequences were trimmed. Matches to gene-specific primer sequences allowed for a one-nucleotide mismatch. Trimmed sequences had their V, D and J gene segments and junctional bases annotated with IgBLAST 1.3.0 (ref. [34]) and a sequence database of germline gene segments from the international ImMunoGeneTics information system (IMGT)[35]. Annotation of antibody isotype and subtype of RNA-derived sequences was accomplished by looking for perfect matches between constant regions from the IMGT database and sequences upstream of the constant region primer. CDR-H3 sequences were identified on the basis of the conserved cysteine-104 and the motif downstream of the conserved tryptophan-118 residue[36]. Non-*IGH* artifactual sequences and those with poor IGHV matches (IGHV segment match bit score less than 140) were removed from the data. The Python scripts and R code used for secondary analyses of the data can be provided on request.

In total, 25,507,784 sequences from HIV-infected individuals (mean: 265,706 per individual) and 8,393,313 sequences from healthy controls (mean: 195,193 per individual) were obtained and analyzed. For the CMV cohort analysis, at total of 36,720,077 sequences from individuals who were CMV-seropositive and 45,429,593 sequences from those who were CMV-seronegative was obtained and analyzed.

Inference of clonal lineages. Sequences were clustered into putative clonal lineages using single-linkage clustering. Briefly, the process starts with all sequences in their own lineages and iteratively, two lineages are merged if any two reads, one from each lineage, satisfy the following four criteria: (1) come from the same individual; (2) share the same V and J segment annotations (not including allele designation); (3) have equal CDR-H3 length; and (4) CDR-H3 nucleotide sequences match with 90% or more identity. The process ends when there are no lineages satisfying the criteria. This clonal inference procedure produced 6,821,567 clones from HIV-infected individuals (mean: 710,578 per person) and 2,400,986 clones from healthy controls (mean: 66,349 per person). The mean number of reads per lineage was 3.68.

Clonality analysis and clonality scores. To calculate a summary measure of the contribution of clonally expanded B cells to the repertoire of each individual, normalizing for sequencing depth in each individual, we calculated a clonality score as:

$$\frac{\sum_{ijk,i\neq k} N_{ij} \times N_{ik}}{\sum_{jk,j\neq k} T_j \times T_k}$$

where N_{ij} and N_{ik} are the copy numbers of clone i observed in independent replicate PCR libraries, j and k, generated from independent aliquots of template DNA and

T_j and T_k are total read numbers in the corresponding replicate libraries. The log base 10 of this score is shown in the figures.

Selection of disease-specific *IGH* clusters. To determine antibody amino acid sequences that distinguish individuals with and without HIV infection or bNAb from noNAb-producers, we clustered *IGH* amino acid sequences to identify convergent antibodies, that is antibodies elicited by different subjects with similar *IGH* sequences. Heavy-chain sequences annotated with the same V and J segment (not considering alleles) and the same CDR-H3 length were clustered based on 90% CDR-H3 amino acid sequence similarity. Partitioning of sequences by V, J and CDR-H3 length was performed using custom software, while the CDR-H3 amino acid sequence clustering was performed using cd-hit[37] with options -c 0.90 -l 4 -S 0 -g 1 -b 1. For discriminating individuals with and without HIV infection, clusters were selected as informative if (1) they contained sequences from at least three individuals; (2) 80% of subjects with the sequences were HIV-infected; and (3) the cluster had an expected mutual information of 0.06 bits or greater[38]. For discriminating bNAb from noNAb-producers, clusters were selected if (1) they contained sequences from at least three individuals; (2) 80% of individuals with the sequences were bNAb producers; and (3) the cluster had an expected mutual information of 0.05 bits or greater. A lower threshold was used to separate bNAb from noNAb-producers due to fewer individuals in this comparison.

Classifying individuals with predictive clusters. For each of the predictive clusters, a consensus CDR-H3 amino acid sequence was calculated by taking the most common amino acid at each position, where each cluster member was weighted by the number of participants with that sequence. A new individual was scored by co-clustering their annotated *IGH* sequences with the consensus sequences of the predictive clusters. This clustering was performed using the same method as detailed above. The new individual was assigned a score equal to the number of predictive clusters that have sequences from the individual co-clustering with the sequences of the cluster.

To test the classification accuracy of this prediction method, we performed fourfold cross validation. Subject groups (individuals who were infected with HIV versus those who were not or bNAb producers versus noNAb-producers) were separated into four sets using stratified random sampling. The first three sets were used to select predictive *IGH* sequences as described above. The selected sequences were then used to score the fourth set as described above. To give an accurate estimate of the generalization error of this method, this process was repeated for 160 partitions.

Repertoire biomarkers of HIV infection and bNAb status. Features in the random forest classifier[39] included: the mean and 99th percentile mutation level of the IgG compartment; the mean CDR-H3 length and fraction of long CDR-H3s (>19 amino acids) of the IgG, IgM, IgD and out-of-frame sequences; clonality score; fraction of IgG sequences using *IGHV4-34* and the presence or absence of the convergent *IGH* sequences from the previous section. The same cross-validation scheme as above was used here. Class probabilities were used to calculate the ROC curve and AUC.

Bioinformatical and statistical analysis. Analysis of heavy-chain data was performed with custom Python[40] code using Pandas[41] and NumPy[42]. Statistical tests and generation of plots was performed in R[43] using Rstudio[44], ggplot[45] and plyr[46]. *P* values were calculated by using the two-sided Wilcoxon Mann–Whitney *U*-test. Box-whisker plots show the median (horizontal line), interquartile range (box) and 1.5-times the interquartile range (whiskers).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The B cell heavy-chain sequences analyzed here are available through the Short Read Archive. Data from the HIV cohort can be found in the BioProject PRJNA486667. Data from CMV-seropositive and seronegative healthy controls are deposited as BioProject PRJNA491287.

References

31. Seaman, M. S. et al. Tiered categorization of a diverse panel of HIV-1 Env pseudoviruses for assessment of neutralizing antibodies. *J. Virol.* **84**, 1439–1452 (2010).
32. Nielsen, S. C. A. et al. Shaping of infant B cell receptor repertoires by environmental factors and infectious disease. *Sci. Transl. Med.* **11**, eaat2004 (2019).
33. van Dongen, J. J. M. et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17**, 2257–2317 (2003).
34. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
35. Lefranc, M.-P. et al. IMGT, the international ImMunoGeneTics information system 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2015).
36. North, B., Lehmann, A. & Dunbrack, R. L. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* **406**, 228–256 (2011).
37. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
38. Manning, C. D., Raghavan, P. & Schütze, H. *Introduction to Information Retrieval* (Cambridge Univ. Press, 2008).
39. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R. News* **2–3**, 18–22 (2002).
40. Python Language Reference v.2.7 (Python Software Foundation, 2013).
41. McKinney, W. Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 51–56 (Python in Science Conference, 2010).
42. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).
43. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2019).
44. RStudio Team. *RStudio: Integrated Development Environment for R* (RStudio, 2016).
45. Wickham, H. ggplot2: Elegant Graphics For Data Analysis (Springer, 2009).
46. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29 (2011).

Acknowledgements

We gratefully acknowledge the participants who volunteered for this study. Support for this work was provided by grants from the National Institutes of Health, National Institute of Allergy and Infectious Diseases, Division of AIDS; UM-1 grant for the Duke Center for HIV/AIDS Vaccine Immunology-Immunogen Discovery AI100645; National Institutes of Health grant Nos. R21-AI100696, CHAVI-AI0678501, R01AI127877 and R01AI130398; and MRC Programme grant No. MR/ K012037.

Author contributions

K.M.R., A.Z.F., P.B., B.F.H. and S.D.B. conceptualized the study. K.M.R., K.J.L.J., I.P.-P., B.F.H. and S.D.B. were responsible for the methodology. K.M.R. and K.J.L.J. were responsible for the software. K.M.R., J.-Y.L., R.A.H., I.P.-P., K.-K.H., H.-X.L. and S.A.J. managed the investigation. K.M.R., K.J.L.J., S.D.B., I.P.-P., P.B., M.A.M., M.B. and K.-K.H. handled the formal analysis. K.-K.H., M.B., H.-X.L., M.A.M., P.B., B.F.H. and S.D.B. collected the resources. K.M.R. curated the data. K.M.R. and S.D.B. wrote the original draft of the manuscript. K.M.R., K.J.L.J., B.F.H., M.B., M.A.M., P.B., S.A.J. and S.D.B. reviewed and edited the manuscript. K.M.R. was responsible for the visualization. P.B., B.F.H. and S.D.B. supervised the project. P.B., B.F.H. and S.D.B. were responsible for funding acquisition.

Competing interests

A patent application related to computational methods used in this paper is in preparation by K.M.R. and S.D.B.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41590-019-0581-0.

Correspondence and requests for materials should be addressed to B.F.H. or S.D.B.

Peer review information L. A. Dempsey was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Boyd, Scott; Haynes, Barton

Last updated by author(s): Dec 3, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Flow Cytometry: Cyan ADP Analyzer (Beckman Coulter), data analysis FlowJo software(v8.8.6; Treestar); Illumina MiSeq Software v2.5, 2.6.
Data analysis	Paired-end read merging with FLASH v1.2; clustering with cd-hit v4.6; Immunoglobulin sequence alignments with IgBLAST v1.3 (NCBI). Flow cytometry analysis with FlowJo software(v8.8.6; Treestar). Python scripts and R code used for secondary analyses of the data will be provided on request to reviewers or readers.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The immunoglobulin cell heavy-chain sequences analyzed here are available through the Short Read Archive (SRA) in the BioProjects PRJNA486667 and PRJNA491287.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size predetermination for this exploratory study was performed. The sample size (46 broad neutralizing subjects, 50 non-neutralizing subjects, 43 uninfected controls from the same population groups) was the largest number of subjects available with these serological and infection criteria from the prospective cohort that was the source of the specimens. The statistically significant differences in IGH repertoire features detected in the different infection status and serological status groups in the study support the adequacy of the number of participants analyzed. A second cohort of comparable size of HIV-uninfected subjects from a different population group (California blood donors) was used to evaluate effects of CMV seropositivity status on IGH repertoire features: 52 were CMV-seropositive and 61 were CMV-seronegative.
Data exclusions	No data were excluded from the analysis.
Replication	The number of broadly-neutralizing HIV-infected patients and appropriate control subjects in this study is larger than in other antibody repertoire analyses on this topic in the literature. We analyzed all available subjects in the study, but did not analyze a separate replication cohort.
Randomization	There was no intervention for randomization in this study.
Blinding	There was no group allocation in the study; patients were classified based on their serological and virological data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	CXCR3 allophycocyanin (APC) (1C6) and CXCR5 Alexa Fluor® (AF)488 (RF8B2) (BD Biosciences); CD3 phycoerythrin (PE)-Texas Red (S4.1; Life Technologies); CD4 APC-H7 (SK3), CD45RO PE-Cy7 (UCHL1), CCR6 PE (11A9) and PD-1 Brilliant Violet™ (BV)421 (EH12.1) (BD Biosciences); biotinylated anti-ICOS (ISA-3; eBioscience); streptavidin Peridinin chlorophyll protein (PerCP) (BD Biosciences); CD3 PE-Texas Red (S4.1; Life Technologies); CD4 APC-H7 (SK3), CD25 PE-Cy7 (2A3), CXCR5 AF488 (RF8B2) and PD-1 BV421(EH12.1) (BD Biosciences); Foxp3 APC (PCH101) and CTLA-4 PerCP-eFluor710 (14D3) (eBioscience);
Validation	Commercial antibodies from BD Biosciences, Life Technologies and eBioscience were validated by the manufacturer.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patients defined clinically as virus controllers were excluded from the study. HIV-infected participants and uninfected controls were predominantly African, and regions of origin of these individuals are described in Moody et al., Sci Immunol. 2016. Matching of non-neutralizing HIV-infected controls to broadly-neutralizing HIV-infected individuals was carried out using propensity score matching with age, sex, and country of origin as input variables, and resulted in a non-neutralizing set of control individuals that did not significantly differ in age, sex or country of origin compared to the broadly-neutralizing individuals.

Recruitment	Donor enrollment was based on standard clinical diagnosis of HIV-1 infection, at which time subjects were all antiretroviral naïve. The HIV study donors were enrolled at clinical sites in Tanzania, South Africa, Malawi, and the United Kingdom. The CMV study donors were enrolled at Stanford Blood Center in Stanford, California, and represent a self-selected population in the sense that they volunteered to donate blood.
Ethics oversight	This study was approved by the institutional review boards at Duke University, King's College London, and Stanford University.

Note that full information on the approval of the study protocol must also be provided in the manuscript.