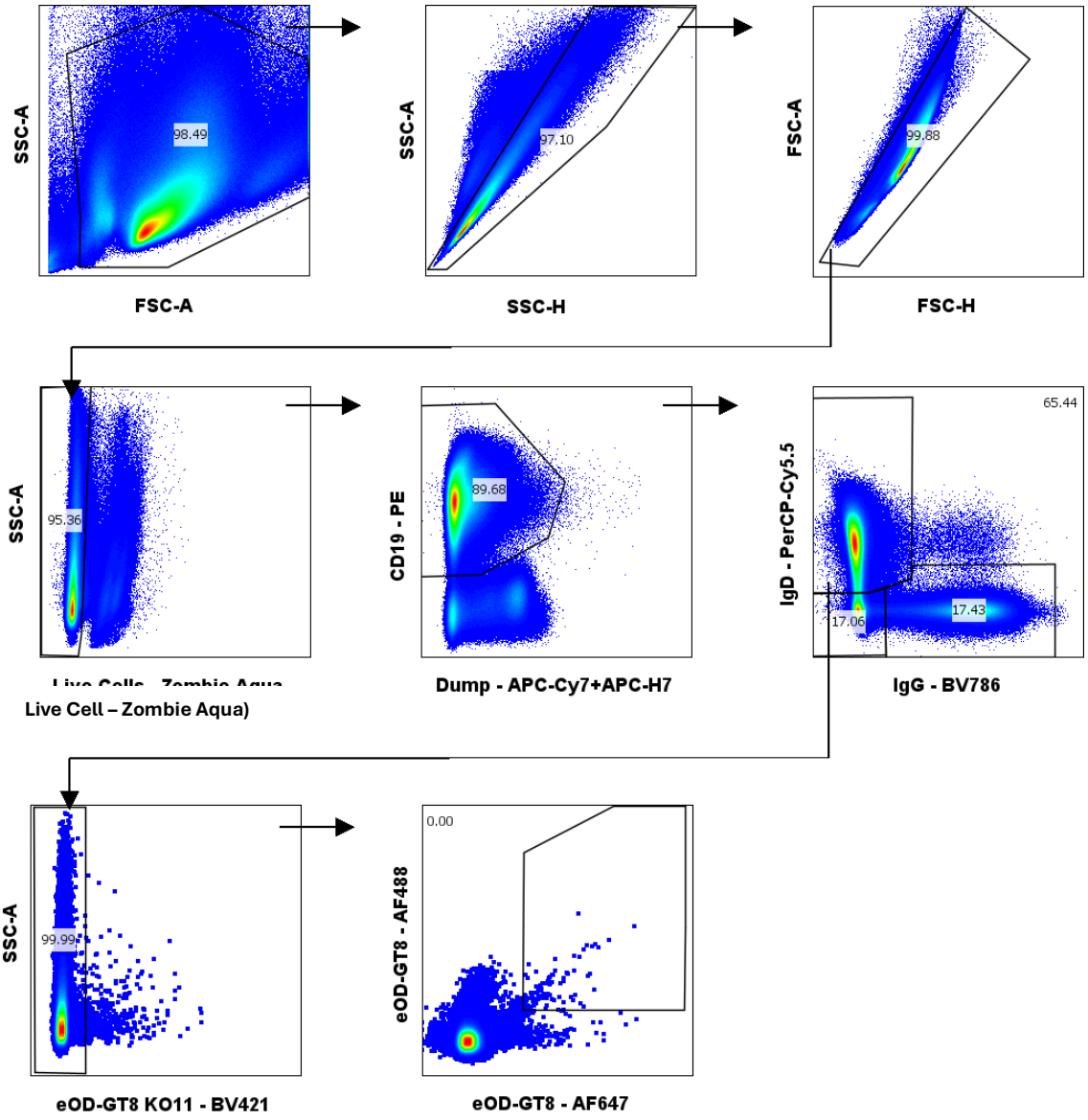
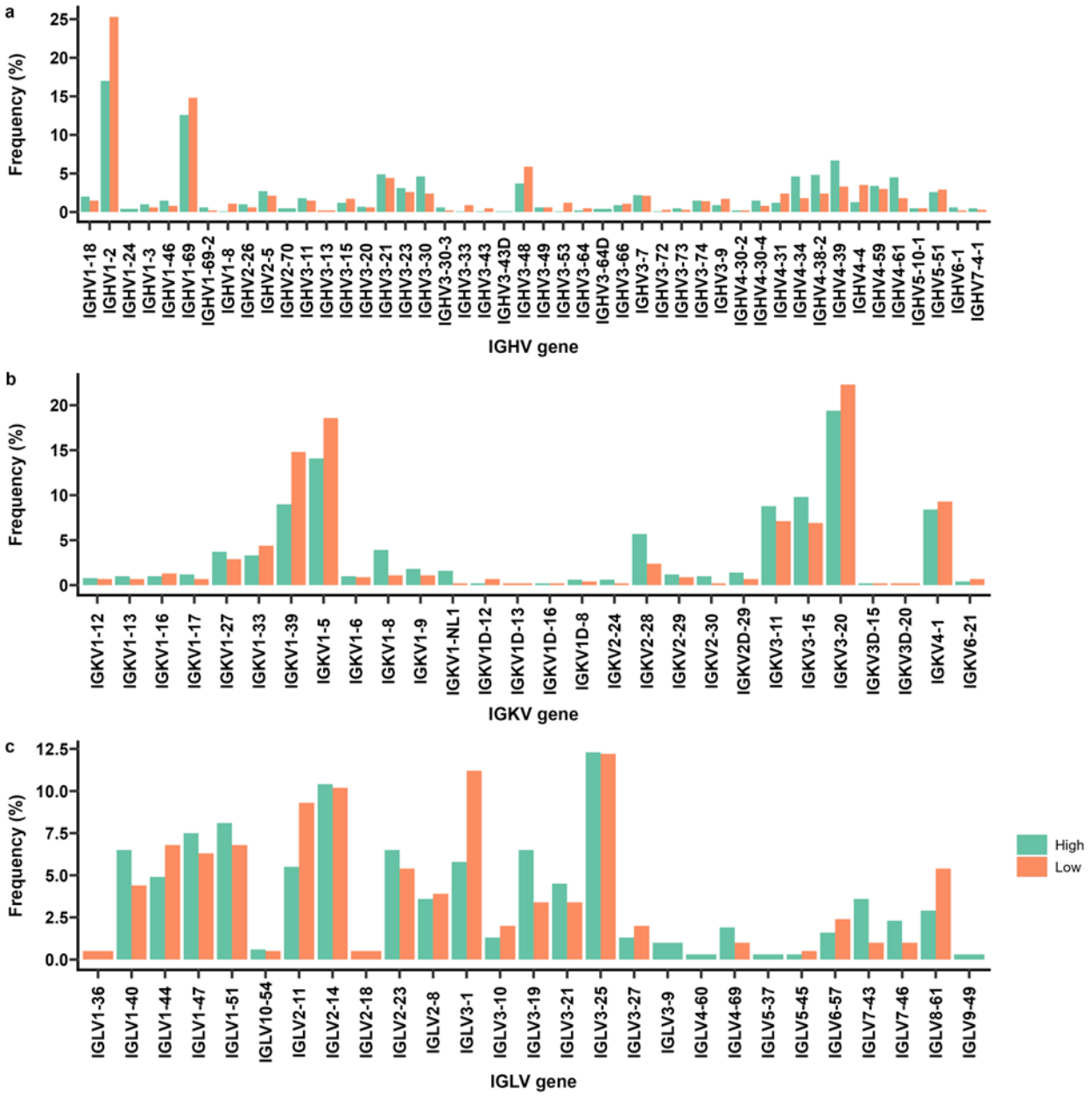
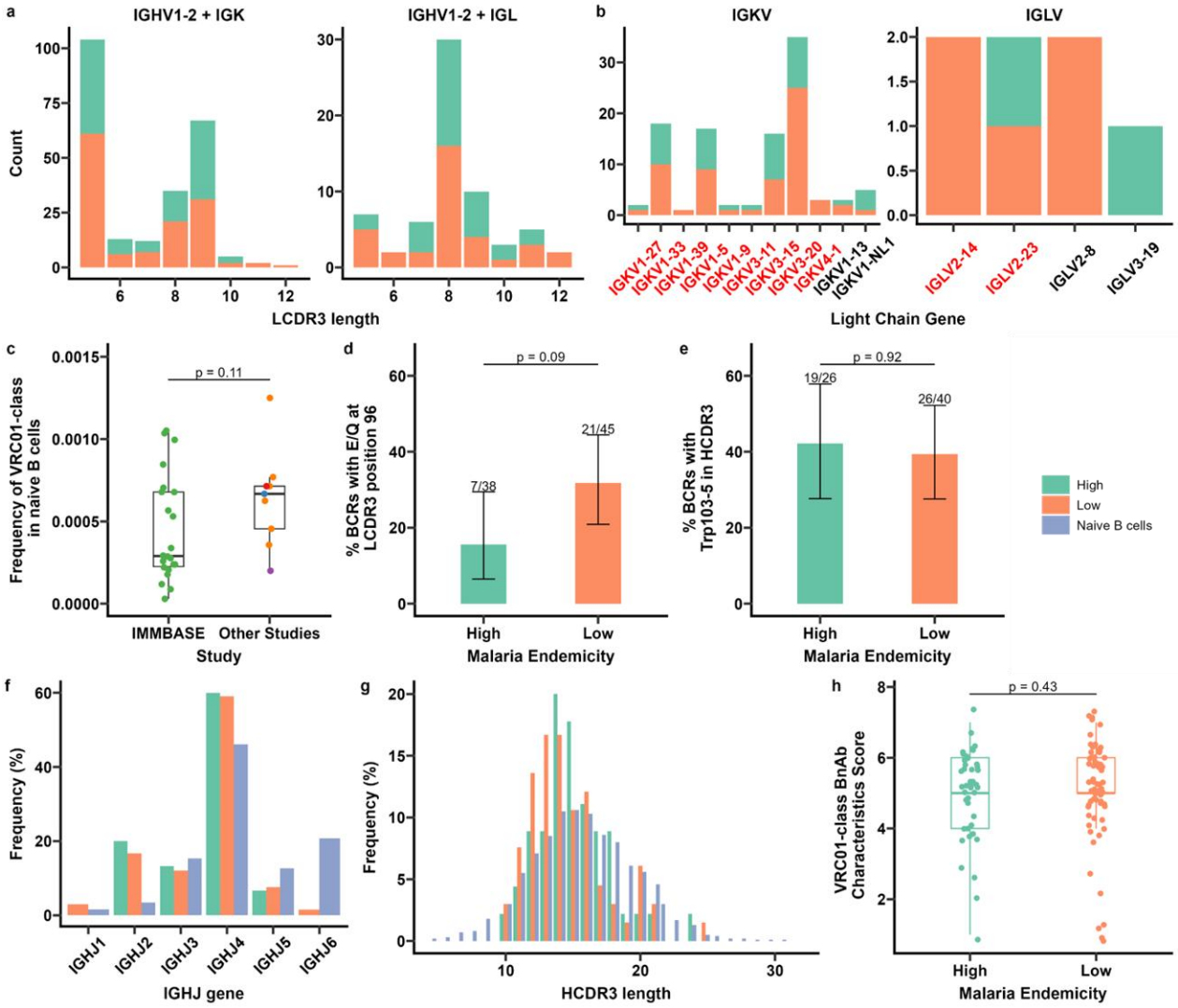


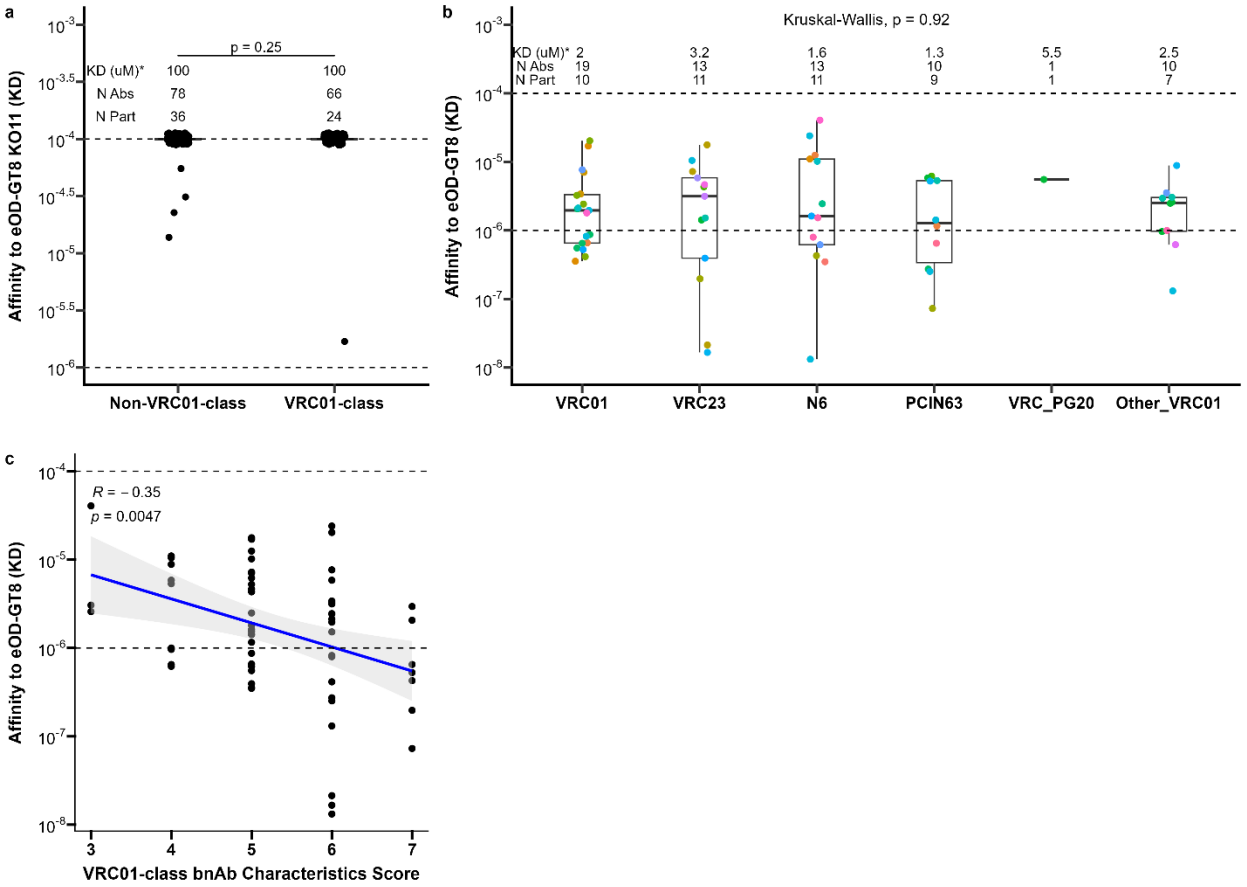
Supplementary Figure 1: Sort gating strategy to isolate eOD-GT8 KO^{-}/eOD-GT8^{++} naïve B cells (CD4bs-specific naïve B cells).



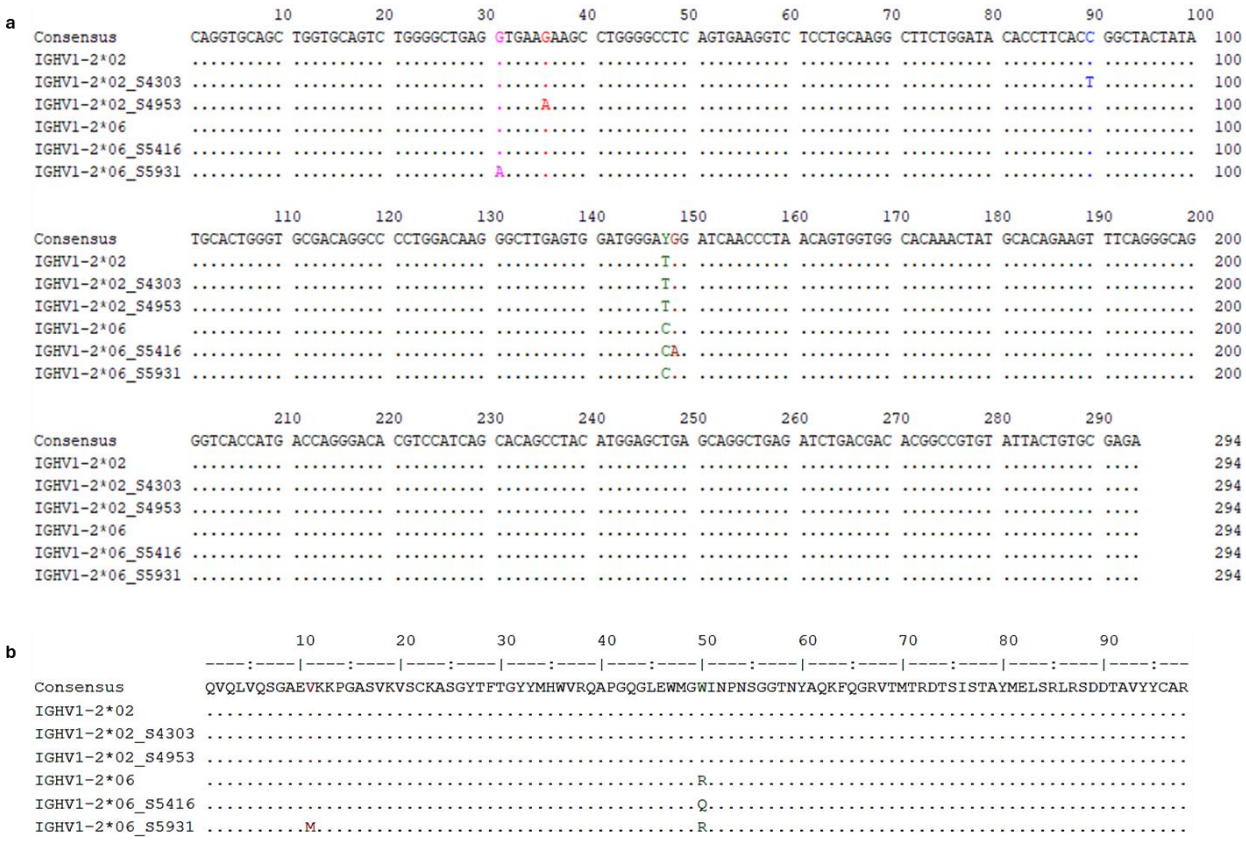
Supplementary Figure 2: V Gene usage distribution in the epitope-specific (eOD-GT8 KO^{-}/eOD-GT8^{++}) cells. Frequency of the **a** IGHV, **b** IGKV, and **c** IGLV gene usage in the eOD-GT8-specific B cells stratified by malaria endemicity.



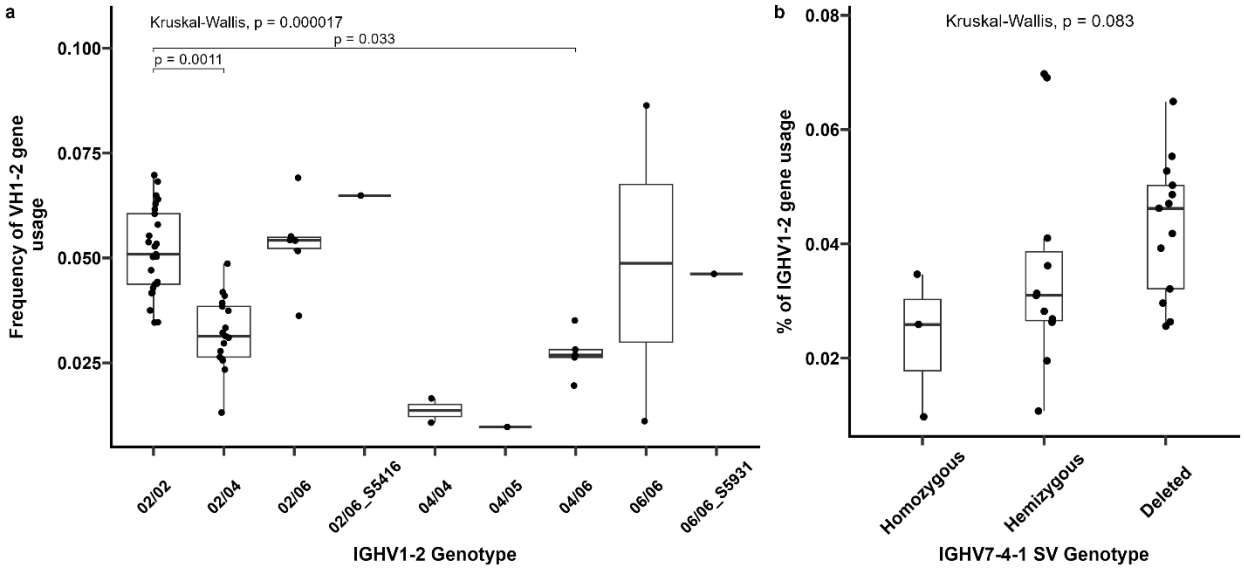
Supplementary Figure 3: Characteristics of sequences from epitope-specific and VRC01-class naïve B cells. **a** LCDR3 length distribution of IGK and IGL light chains paired with an IGHV1-2 gene. **b** Light chain gene usage of the VRC01-class precursors. Genes highlighted in red have previously been identified in naïve VRC01-class precursors, while those in black have not previously been identified in precursors. **c** Comparison of precursor frequency estimates between this study and other studies with publicly available data. Points are coloured by the respective study: green – this study, purple pooled data from Jardine *et al.,* 2015[1] corrected for losses during sorting and library preparation, blue – pooled data from Havenar-Daughton *et al.,* 2018 (BCR RNAseq method), red – Havenar-Daughton *et al.,* 2018 (BCR PCR method)[2] and orange – individual participants from Lee *et al.,* 2021[3] . **d** Percentage of BCRs using a glutamate (E) or glutamine (Q) at LC position 96 (Kabat numbering) in IGK 5aa-LCDR3s from VRC01-class precursors. **e** Percentage of BCRs using a tryptophan residue at position -5 of the HCDR3 (Trp_{103-5}) in VRC01-class precursors. **f** IGHJ gene usage of VRC01-class precursors compared to random naïve B cells. **g** HCDR3 length distribution in VRC01-class precursors compared to random naïve B cells. **h** Variation in VRC01-class bnAb characteristics score by malaria endemicity. For the boxplots, each dot represents an individual, with the thick lines showing the median value, the boxes indicating the 25% and 75% quartiles and the 'whiskers' highlighting the minimum and maximum values. Statistical comparisons were carried out using the Wilcoxon rank-sum test. In the bar graphs, the whiskers represent the 95% Clopper-Pearson confidence intervals with statistical significance determined using a Chi-Square test.



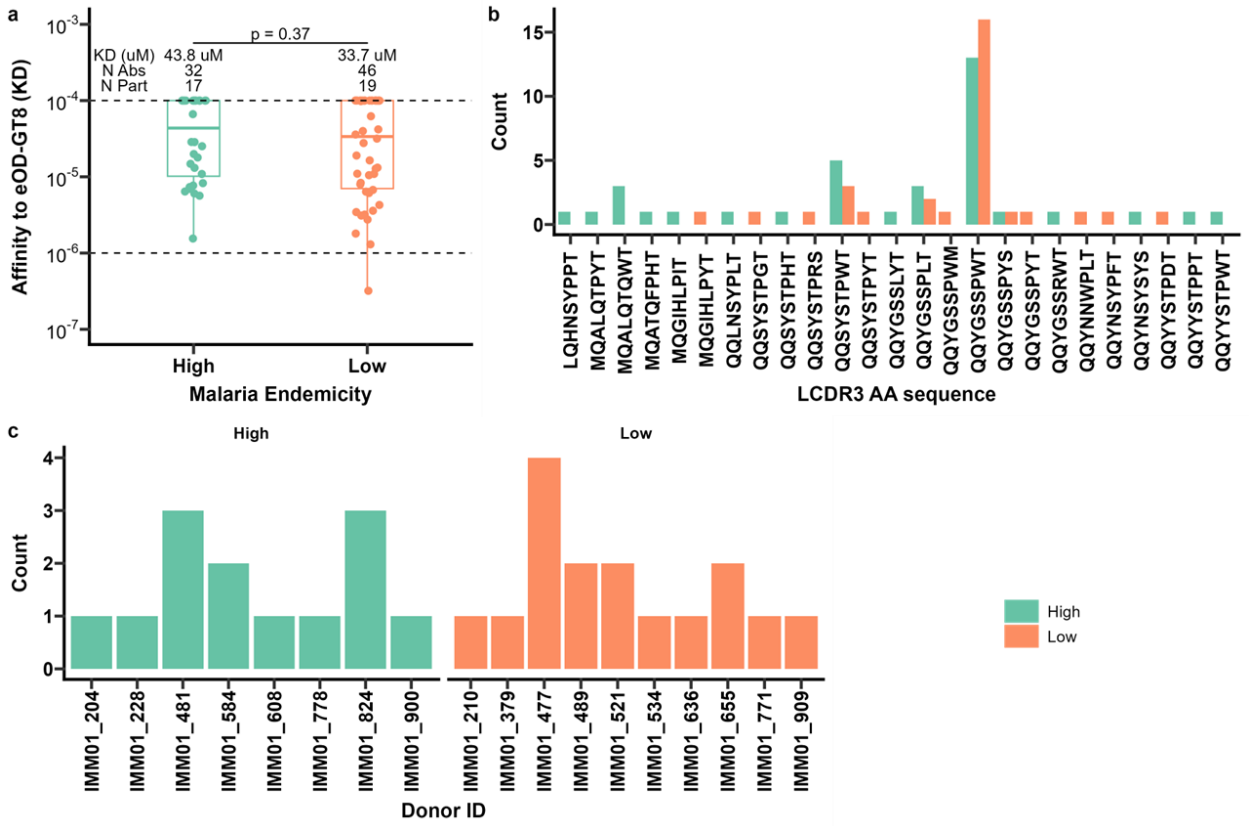
Supplementary Figure 4: Monovalent affinity of select mAbs produced from VRC01-class and non-VRC01-class sequences. a Affinity to eOD-GT8 KO11 by VRC01-class and non-VRC01-class mAbs. Statistical significance determined using a two-sided Wilcoxon rank sum test. **b** Affinity to eOD-GT8 by VRC01-class mAbs from the different VRC01-subclasses (VRC01 (Vκ3-20+), VRC23 (Vκ3-15+), N6 (Vκ1-33+), PCIN63 (Vκ1-5+) and VRC-PG20 (Vλ2-14)), with points coloured by participant. Each dot represents an individual mAb, with the thick lines showing the median value, the boxes indicating the 25% and 75% quartiles and the 'whiskers' highlighting the minimum and maximum values. Statistical significance determined using a Kruskal-Wallis test. **c** Correlation between bnAb characteristics score and monovalent affinity to eOD-GT8 (n = 64, excluded two Lambda light chain mAbs as these cannot be scored on matching to the 'QQYEF' motif). Correlation coefficient and p value determined by Spearman test. 'K_D (μM)' – median K_D in μm, 'N Part' – number of participants, 'N Abs' – number of mAbs tested.



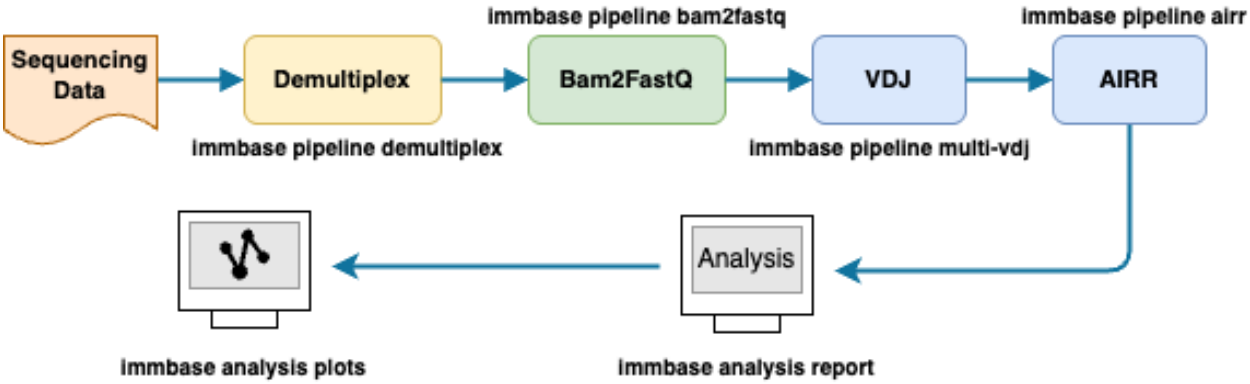
Supplementary Figure 5: Multiple alignment of reference IGHV1-2*02 and IGHV1-2*06 alleles with the novel *02 and *06 alleles. a Nucleotide alignment. **b** Amino acid alignment. Coloured residues indicate mutations.



Supplementary Figure 6: Influence of IGHV genotypes on IGHV1-2 gene usage and VRC01-class precursor frequency. a Variation in IGHV1-2 gene usage frequency by IGVH1-2 genotype (N = 60) and **b** by IGHV7-4-1 structural variation genotype (n = 27). Statistical comparisons were carried out using the Kruskal-Wallis test with pot hoc analysis done using Dunn's test with Bonferroni correction applied for multiple comparison adjustment. Each dot represents a participant, with the thick lines showing the median value, the boxes indicating the 25% and 75% quartiles and the 'whiskers' highlighting the minimum and maximum values.



Supplementary Figure 7: Characteristics of non-VRC01-class epitope-specific naïve B cells. a Monovalent affinity of non-VRC01-class mAbs stratified by malaria endemicity. **b** Frequency of the different amino acid motifs identified in 9-aa LCDR3s of non-VRC01-class eOD-GT8++/KO- cells. **c** Count of B cells encoding the public LC lineage (Vκ3-20 + QQYGSSPWT LCDR3) observed in 9-aa LCDR3 sequences in multiple donors (n = 18). Each dot represents a mAb, with the thick lines showing the median value, the boxes indicating the 25% and 75% quartiles and the 'whiskers' highlighting the minimum and maximum values.



Supplementary Figure 8: Summary of the 'immbase' data analysis pipeline used to pre-process and analyse the data. The 'immbase' pipeline was used to process the sequencing data from demultiplexing to VDJ assignment and finally to sequence characterisation in order to classify VRC01-class sequences.

Supplementary Table 1: Demographic characteristics of the study participants

	Ahero[1]	Nairobi[1]	Ngerenya[1]	Junju[1]	High Endemicity[1]	Low Endemicity[1]	Overall
Participant summary							
No. of participants	15	15	15	15	30	30	60
Age (Years)	24 (23, 34)	21 (19, 23)	27 (22, 31)	32 (30, 40)	30 (24, 37)	23 (21, 27)	26 (22, 31)
Sex							
Female	7 (47%)	9 (60%)	0 (0%)	3 (20%)	10 (33%)	9 (30%)	19 (32%)
Male	8 (53%)	6 (40%)	15 (100%)	12 (80%)	20 (67%)	21 (70%)	41 (68%)

[1] – Median (IQR) or Frequency (%)

Supplementary Table 2: Inclusion and exclusion criteria for enrolment into the study

Inclusion criteria:	
	HIV negative result
	Haemoglobin level above 12.5 g/dL
	Age between 18 – 60 years
	A normal blood pressure (systolic 90 –120 mmHg; diastolic 60–80 mmHg)
	Pulse rate of 60 – 100 beats per minute
	Resident within the recruitment area for at least 10 years
Exclusion Criteria:	
	HIV positive
	Weight below 50 kg
	History of severe illness in the last 6 months (defined as any clinically significant acute or chronic medical condition that has occurred over the past 6 months that is considered progressive or in the opinion of the involved medical doctor makes the volunteer unsuitable for participation in the study).
	Blood donation within the last 3 months for men and last 4 months for women
	Pregnant or breastfeeding
	Delivered or had a termination of pregnancy within the last 6 months.

Supplementary Table 3: Summary of sorting and sequencing metrics by malaria endemicity

	High Endemicity[1]	Low Endemicity[1]	Overall
Study population summary			
No. of participants	30	30	60
No. of samples analysed	27	30	57[a]
No. of enriched B cells in millions	44.77 (22.2 – 57.9)	30.47 (17.6 – 38.7)	30.87 (18 – 50.1)
Sorting summary			
No. of IgD+ B cells screened in millions	8.76 (4.2 – 15.3)	8.97 (2.9 – 13.8)	8.76 (3.6 – 13.9)
No. of epitope-specific cells sorted	226 (96 – 358)	173 (80 – 274)	178 (92 – 321)
Sequencing summary			
No. of participants with sequencing data	26	28	54[b]
No. of sequences per participant	24 (12 – 46)	21 (9 - 51)	23 (10 – 51)
No. of sequences per participant after filtering to retain naïve sequences	9 (6 – 32)	13 (5 – 29)	11 (6 – 32)
No. of participants with a VRC01 Class BCR[c]	16	14	30
No. of participants with at least 10 sequences[d]	11	17	28

[1] – Median (IQR)

[a] – 3 samples not enough PBMCs for analysis

[b] – 3 samples had no sequence data available owing to failed GEM generation

[c] – analysis of VRC01-class BCRs done after filtering to retain sequences from naïve B cells

[d] – analysis of precursor frequencies restricted to participants with at least 10 sequences

Supplementary Table 4: Summary of sorting and sequencing metrics by location

	Ahero[1]	Nairobi[1]	Ngerenya[1]	Junju[1]	Overall[1]
Participant summary					
No. of participants	15	15	15	15	60
No. of samples analysed	12	15	15	15	57[a]
No. of enriched B cells in millions	26.77 (23.2 – 48.9)	39.45 (28.9 – 50.0)	17.5 (8.6 – 30.57)	52.85 (22 – 67.9)	30.87 (18.0 – 50.1)
Sorting summary					
No. of IgD^+ B cells screened in millions	5.99 (3.9 – 9.0)	12.33 (8.4 – 16.0)	4.39 (1.2 – 12.2)	11.29 (4.5 – 18.8)	8.35 (3.2 – 13.8)
No. of epitope-specific cells sorted	136 (87-272)	235 (180-334)	116 (54-173)	317 (124-558)	176 (89-320)
Sequencing summary					
No. of participants with sequencing data	11	13	15	15	54[b]
No. of sequences per participant	15 (13 – 32)	34 (17 – 51)	17 (8 – 33)	29 (16 – 80)	23 (10 – 51)
No. of sequences per participant after filtering to retain naïve sequences	9 (7 – 23)	18 (12 – 36)	9 (5 – 20)	18 (6 – 59)	11 (6 – 32)
No. of participants with at least 10 sequences[c]	5	9	5	11	30
No. of participants with VRC01 Class BCR	3	10	7	8	28

[1] – Median (IQR)

[a] – 3 samples not enough PBMCs for analysis

[b] – 3 samples had no sequence data available owing to failed GEM generation

[c] – analysis of VRC01-class BCRs done after filtering to retain sequences from naïve B cells

[d] – analysis of precursor frequencies restricted to participants with at least 10 sequences

Supplementary Table 5: Sequence features used to classify other non-VRC01-class CD4bs-targeted bnAb precursor lineages

Type	bnAb class	HC V gene	HCDR3 Length	n	Reference
HCDR3-dominated	CH103-like	IGHV4-59	13	5	[4]
	VRC13-like	IGHV1-69	21	8	[4]
	VRC16-like	IGHV3-23	20	1	[4]
	VRC33-like	IGHV4-34	14	6	[4]
	HJ16-like	IGHV3-30	18	5	[4]
IGHV gene-restricted	VH1-46-derived	IGHV1-46	-	17	[4]

LCDR3 length was not considered an inclusion criterion

Supplementary Table 6: Flow cytometry PBMC phenotyping antibody panel

Marker	Fluorophore	Clone	Vendor	Catalogue No.
Live/Dead stain	BV510		Biolegend	423102
CD3	APC-Cy7	OKT3	BD	317342
CD14	APC-H7	M5E2	BD	561384
CD16	APC-H7	3G8	BD	560195
CD56	APC-Cy7	5.1H11	Biolegend	362512
CD19	PE	HIB19 (RUO)	BD	555413
IgD	PerCP-Cy5.5	IA6-2	BD	561315
IgG	BV786	G18-145	BD	564230

Supplementary Data 1: Sequence characteristics the VRC01-class naive B cells isolated using eOD-GT8

Supplementary Data 2: Summary of IGHV1-2 genotypes, IGHV1-2 gene usage frequency and IGHV7-4-1 haplotypes of the study participants

Supplementary Data 3: Characteristics of epitope-specific (KO^{-}/$eOD\text{-}GT8^{++}$) $IGHV1\text{-}2^{+}$ naïve B cells using 9 aa and 8 aa LCDR3s.

References

1. Jardine, J. G. *et al.* HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science* **351**, 1458–1463 (2016).

2. Havenar-Daughton, C. *et al.* The human naive B cell repertoire contains distinct subclasses for a germline-targeting HIV-1 vaccine immunogen. *Science Translational Medicine* **10**, (2018).

3. Lee, J. H. *et al.* Vaccine genetics of IGHV1-2 VRC01-class broadly neutralizing antibody precursor naïve human B cells. *npj Vaccines* **6**, 1–12 (2021).

4. Zhou, T. *et al.* Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. *Cell* **161**, 1280–1292 (2015).