Supporting Online Material

IMPACT OF CLADE, GEOGRAPHY AND AGE OF THE EPIDEMIC ON HIV-1 NEUTRALIZATION BY ANTIBODIES

Peter Hraber[a], Bette T. Korber[a,b], Alan S. Lapedes[a], Robert T. Bailer[c], Michael S. Seaman[d], Hongmei Gao[e], Kelli M. Greene[e], Francine McCutchan[f,m], Carolyn Williamson[g], Jerome H. Kim[f], Sodsai Tovanabutra[f], Beatrice H. Hahn[h], Ronald Swanstrom[i], Michael M. Thomson[j], Feng Gao[e], Linda Harris[k], Elena Giorgi[a], Nicholas Hengartner[a], Tanmoy Bhattacharya[a,l], John R. Mascola[c], David C. Montefiori[e#]

[a]Los Alamos National Laboratory and [b]New Mexico Consortium, Los Alamos, New Mexico, United States of America
[c]Vaccine Research Center, National Institutes of Health, Bethesda, Maryland, United States of America
[d]Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America
[e]Duke University Medical Center, Durham, North Carolina, United States of America
[f]US Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, Maryland, United States of America.
[g]Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa
[h]Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America
[i]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America
[j]Instituto de Salud Carlos III, Madrid, Spain
[k]Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America
[l]Santa Fe Institute, Santa Fe, New Mexico, United States of America
[m]Recently retired from the Bill & Melinda Gates Foundation, Seattle, Washington, United States of America

Text S1: Summarizing hypervariable region characteristics

Here we characterized features within a hypervariable regions excised from NSDP protein alignment, summarizing and reporting the following peptide characteristics: lengths, number of N-linked glycosolation sites (the sequence pattern NX[ST], where X can be any amino acid except P), and net charge. We have designed a web-based tool called the "Variable Region Characteristics" tool and made it available through the Los Alamos database to automate this process (http://www.hiv.lanl.gov/content/sequence/VAR_REG_CHAR/#)). The web tool is general in that can be used to characterize full-length proteins or excised regions from any protein alignment, not just HIV, by the user specifying alignment positions. As it was an approach initially developed for HIV Env, however, if an Env alignment is the input, we have automated the excision and characterization of either the full V1, V2, V3, V4, V5 loops, or the hypervariable regions within them. How the boundaries of the hypervariable regions are treated in the alignment is critical for a meaningful result, and we explain this fully in the documentation at the web site. Here we present a brief discussion of the strategy we used to define the hypervariable regions for this study. The hypervariable region definitions are the same boundaries as those specified in the database. All are noted below and include the V1, V2, V4, and V5 loops; the V3 loop does not have a hypervariable region.

We do not recommend use of the V region hypervariable boundaries in all situations. We have found these boundaries useful for other population-level alignments. However, some alignments, particularly from within-subject longitudinal samples, may benefit by a modified selection of hypervariable region boundaries. In particular, regions where insertions and deletions dictate the evolutionary trajectory may be more narrowly defined in a within-subject setting than the population based boundaries used here.

V1:
The V1 loop includes positions 131-157 in HXB2, and is bounded by a disulfide bond in the Cysteines at the base (C). The V1 loop is highlighted in blue in the HXB2 Env protein fragment shown below. The hypervariable region in V1 as defined in this study (the region where the alignment begins to breaks down), as it is found in HXB2, is marked in red, from T132 through G152. There is extreme length variation in such regions, and the program will extract everything thing between, but not including, the more readily aligned C131 and E153 in HXB2, which bound the hypervariable region.

```
         1                         1
         3                         5
         1                         7
LTPLCVSLKCTDLKNDTNTNSSSGRMIMEKGEIKNCSFNISTSIR  V1 loop, HXB2
          TDLKNDTNTNSSSGRMIMEKG                V1 hypervariable region
```

V2:
The V2 region begins where V1 ends, starting at S158 and continuing through C196 in HXB2. Like V1, V2 is bounded by Cys bonds, however the C196 is linked with the C at 126, giving a "rabbit ear" structure to the region. The V2 hypervariable region is marked in red; in HXB2 it spans D185 and S190.

```
         1                                     1
         5                                     9
         8                                     6
MEKGEIKNCSFNISTSIRGKVQKEYAFFYKLDIIPIDNDTTSYKLTSCNTSVITQA V2 loop, HXB2
                                    DNDTTS            V2 hypervariable region
```

V1V2:

As described in the main text, the V1 and V2 regions are summarized independently before adding the results, rather than treating the entire region or concatenating fragments. This avoids introducing spurious N-linked glycosylation sites from concatenating V1 and V2 hypervariable fragments.

V4:

V4 is bounded by the disulfide bond between C385 and C418, using HXB2 numbering. The V4 hypervariable region is marked in red, in HXB2 it spans F396 and G410:

```
       3                                4
       8                                1
       5                                8
CGGEFFYCNSTQLFNSTWFNSTWSTEGSNNTEGSDTITLPCRIKQIINMW V4 loop, HXB2
                  FNSTWSTEGSNNTEG                  V4 hypervariable region
```

V5:

The V5 loop is defined based on a loop-like projection in the gp120 structure, and is not bound by Cysteine disulfide bridges at its base. It is located in positions N460 to R469 in HXB2. The V5 hypervariable region is marked in red; in HXB2 it spans N460 and S465:

```
         4        4
         6        6
         0        9
GLLLTRDGGNSNNESEIFRPGGG      V5 loop, HXB2
         NSNNES              V5 hypervariable region
```

Considerations for defining hypervariable loops

If hypervariable regions based on the positions noted in HXB2 are simply excised from an alignment, the extent of the region in other proteins with longer hypervariable sections than HXB2 will not be captured, and depending on the input alignment, even regions with the same length or shorter hypervariable regions than HXB2 may not be fully represented. We include this illustration to clarify our approach, providing an example, as this is the first time we have published statistical comparisons of loop characteristics. The V2 hypervariable region is shown in bold below, and the HXB2 hypervariable region is highlighted in red:

```
B.FR.83.HXB2_K03455      QKEYAFFYKLDIIPI--------DNDTTSYKLTSCNTSVITQACPKVSFEPIPIHYCAPA
B.US.98.1058_AY173951    QKQYALFYKLDVVQMN-------NNNN-SYRLISCNTSVITQACPKVSFEPIPIYYCAPA
B.NL.00.671_00T_AY331295 QREFALLSKLDIVPIDNDSY--------SYMLINCNTSVITQACPKVSFQPIPIHYCTPA
C.BR.92.BR025_AY423387   EKVHALFYRLDIVPLKNESS---NTSGD-YRLINCNTSAITQACPKVSFDPIPIHYCAPA
C.IN.95.95IN210_U52953   QTVYALFYKLDIVPLDNEEQENDSNSSGYYRLINCNTSALTQACPKVTFDPIPIHYCAPA
C.ZA.04.04ZASK1_AF067155 QKVNALFYRSDIVPL--EK------NSSEYILINCNTSTITQACPKVSFDPIPIHYCAPA
```

If the region from the alignment that spans D185 and S190 in HXB2 is simply extracted, the following peptides would be pulled from the alignment, and most of the hypervariable regions in most sequences would be missed.

INCORRECT SUMMARY:

Fragment	Length	Charge	Glycosylation
DNDTTS	6	-2	1
NNNN-S	5	0	1
-----S	1	0	0
NTSGD-	5	-1	1
SNSSGY	6	0	1
--NSSE	4	-1	1

If instead the entire region between the two more conserved "alignable" positions, located just outside the bounds between I184 and Y191 in the hypervariable stretch in HXB2, including the gaps inserted into HXB2 to maintain the alignment, is excised, then the full region is captured, and we get very different, and appropriate, results.

CORRECT SUMMARY:

Fragment	Gaps Removed	Length	Charge	Glycosylation
--------DNDTTS	DNDTTS	6	-2	1
N-------NNNN-S	NNNNNS	6	0	1
DNDSY--------S	DNDSYS	6	-2	1
KNESS---NTSGD-	KNESSNTSGD	10	-1	2
DNEEQENDSNSSGY	DNEEQENDSNSSGY	14	5	2
—-EK------NSSE	EKNSSE	6	-1	1

Prior to this analysis, we took care to confirm that the alignment was sensible in the boundary regions. Because insertions often in part carry direct repeats, and regions vary in length extensively (Wood et al., PLoS Pathog. 2009 May;5(5):e1000414.), automated multiple alignment programs sometimes can give grossly inappropriate results in the hypervariable regions of HIV Env, particularly when the multiple alignment program is challenged with a very large and diverse data set as input. Thus, manual review of these regions is critical if they are important to a particular study, as in this case.

Text S2: Mixed-effects logistic regression of ID50 response

Of the 44,758 ID50 values, 19,169 (42.8%) were below the level of detection. To resolve the issue of estimating population parameters with censored observations present, we model the binary variable Y as an indicator function I: $Y = I$ (ID50 > 28), which is True when the measured ID50 is above the median, and relate Y to observed covariates X with mixed-effects logistic regression. This definition of Y ensures that positive coefficients in the model correspond to an increase in the expected proportion of ID50s above the median.

We utilized several statistical models, one to investigate the effects of infection stage and overall clade-matching, one to check that the main effects remain significant when plasma screening effects are resolved by post-screening, and third to quantify magnitude of effects between particular clades.

Our first analysis relates the logistic transformation $\Theta_i = \log P[Y_i = 1|X_i] - \log P[Y_i = 0|X_i]$ to the following vector of covariates X:

1. Stage: Two-level factor that indicates the infection stage from which the virus was isolated:
   Level E: early
   Level L: late

2. Match: Four-level factor measuring closeness between virus and plasma clades:
   Level 1: clade of virus and clade of plasma are both from CRF07
   Level 2: clade of virus and clade of plasma are both from CRF01
   Level 3: other matching clades of virus and plasma, with C and CRF07 pooled
   Level 4: clade of virus does not match clade of plasma

3. Screened: Two-level factor for plasma screening. The baseline is “False”.

4. pS12 & vS12: Length of V1/V2 hypervariable regions; pS12 refers to plasma, vS12 to virus.

The first model considers a linear combination of the above explanatory variables. Analyses of geometric mean ID50s suggested an interaction between virus stage (Stage variable) and relatedness of plasma and virus clades (Match variable). Analysis of deviance showed no statistical support for the interaction term:

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
Linear	10	20790	20870	-10385			
Interaction	13	20793	20897	-10384	2.7263	3	0.4358

As a result, we select the simpler model without the interaction.
That model is specified in R (**Text S2**) as follows:

```
M.1 = formula(Y ~ Stage + Match + Screened + pS12 + vS12
                 + (1|plasma) + (1|virus))
fit0.mer = glmer(M.1, data=nsdp, family=binomial)
```

Estimated standard deviations for the random effects of plasma and virus are 1.8137 and 0.9935, respectively, so the random effect for plasma is about twice the random effect for viruses. Furthermore, the magnitude of the estimated fixed effects is typically smaller in magnitude (compared to the standard error) than the random effect (**Table S2**).

The stage of the virus had a statistically significant effect, with late-stage viruses increasing the number of ID50 values over the median value of 28 (**Table S2**). The Match variable also had a significant effect, with plasma and virus matches in clades CRF07 and CRF01 increasing the expected number of ID50 values above the median, while matching of the other virus/plasma clades had a statistically significant decrease in the expected number of ID50 values above 28.

While we give no statistical p-value for the mismatch between virus and plasma Env clade, that effect had a large negative impact on the expected number of ID50s above the median.

In addition, the effects of the sum of the lengths of the V1/V2 regions of both virus and plasma Envs were statistically supported at the 5% significance level. As discussed in the text, this is consistent with our expectation that long V1/V2 insertions may inhibit access to some neutralizing epitopes in Env, resulting in the observed association of length with resistance in the virus. In contrast, contemporaneous Envs isolated from plasma samples are likely to reflect resistance to antibodies in that plasma, and the more potent plasma have greater pressure to accrue resistance mutations in the Env population within the host. Finally, also as expected, screening increased the expected proportion of ID50 assay outcomes above the median.

To assess the robustness of these findings, we applied the post-screening criterion by excluding 58 plasma samples with geometric mean ID50s below 20, simplified representations of the Match variable to is.matched as either True (Levels 1-3, including C plasma against CRF07 virus, and C virus against CRF07 plasma) or False (Level 4) and the virus infection stage to is.chronic, either True (Level L, i.e. late Fiebig VI–Chronic) or False (Level E and also Intermediate-stage viruses, i.e. Fiebig I–early VI), increased the threshold for the response variable to the median ID50 value with low-titer plasma samples excluded, and repeated the analysis (**Text S2**).

The estimated standard deviations for the random effects of plasma and virus were 1.23 and 1.03, respectively, so the random effect for plasma was still greater but much closer to the random effect for viruses when low-titer plasmas were excluded. Magnitudes of the estimated fixed effects were still smaller in than random effects. **Table S3** summarizes significance levels of fixed effects in a simplified model:

```
 Y ~ is.late * is.matched + pS12 + vS12 + (1|plasma) + (1|virus).
```

Overall, patterns of significance were consistent with results from all plasmas, regardless of the screening effect (**Table S2**): strong support for clade-mismatched assay results, earlier-stage viruses, having fewer above-median ID50s. The less obvious associations of viruses with longer V1/V2 loop lengths having greater neutralization resistance and, conversely, plasma Envs with longer V1/V2 loop lengths having greater neutralization potency, were also supported.

Estimates for the Match variable in Model 1 indicated a significant interaction between the virus clade and Plasma clade for predicting the expected probability for ID50 above the median. We therefore consider the following model for the ID50 response above the median:

```
M.3=formula(Y ~ virus.clade * plasma.clade + plasma.screened
           + pS12 + vS12 + (1|plasma) + (1|virus))
fit2.mer=glmer(M.3, data=nsdp, family=binomial)
```

This analysis showed that the variable “screening” and total lengths of the V1 and V2 hypervariable regions for the virus are not statistically significant at the 5% level. Removing these variables and refitting the fixed-effect model yields interactions between virus and plasma clades as presented **Table 4** in the main text.

The standard deviation the plasma and virus random effects were 1.673 and 1.052, respectively. Again the variation attributed to plasma Env clade was larger than that attributed to virus clades. Also, the magnitude of estimated fixed effect was again commensurate with the standard deviation of the random effects, suggesting that this is a very noisy situation, in part due to screening. Despite the challenges presented by covariates with virus and plasma in the checkerboard experimental design, the biological signals were clearly and consistently significant under mixed-effects logistic regression.

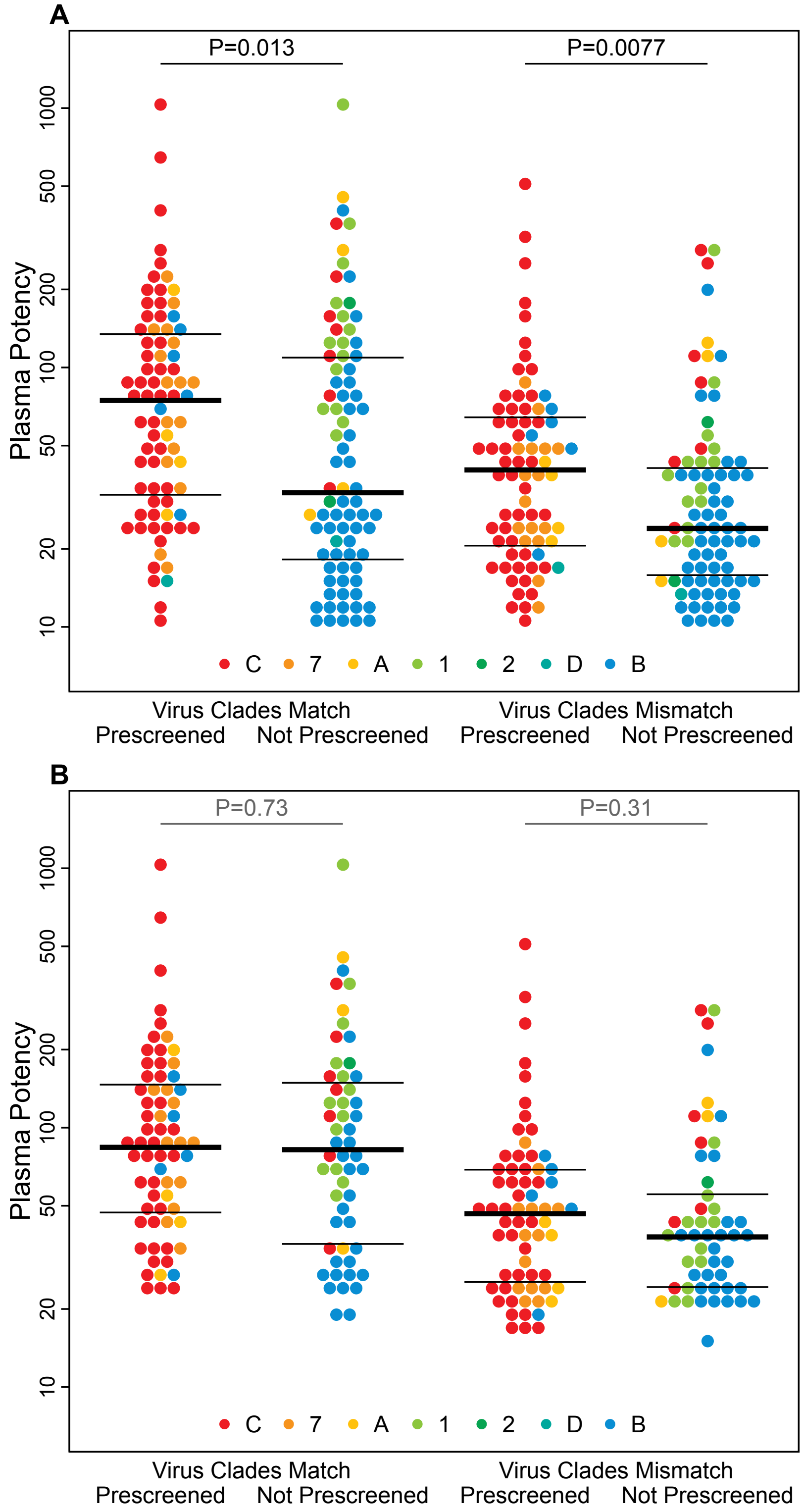

Figure S1. Computational post-screening mitigates screening bias. (A) Screened plasmas are more potent than non-screened plasmas against clade-matched (left, Wilcoxon p=0.013) and mismatched (right, p=0.0077) viruses. (B) Excluding 40 plasmas with overall geometric mean titers below the assay sensitivity limit minimizes the screening bias against clade-matched (left, p=0.73) and clade-mismatched (right, p=0.31) viruses. Thick black lines indicate median ID50s. Thin black lines indicate 25th and 75th percentiles.

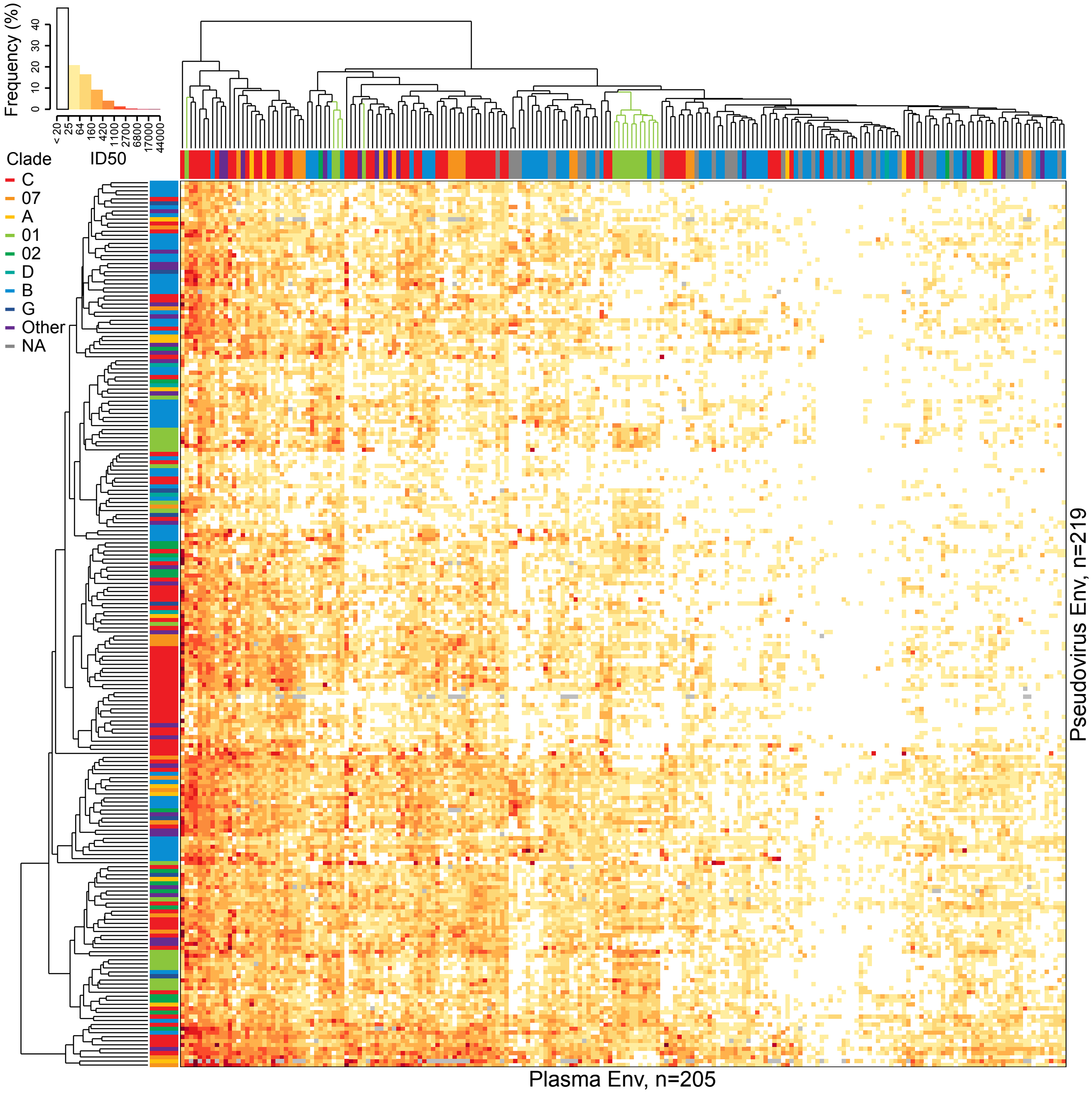

Figure S2. Hierarchically clustered heatmap of ID50 values. Green lines on the dendrogram above the heatmap indicate plasmas from Thailand. One B-clade plasma from Thailand (T500108_503963) clusters with ten CRF01 plasmas from Thailand (T500206_614109, T500207_502102, T500207_509989, T500107_535902, T500208_504258, T500104_276248, T500204_502281, T500105_293735, T500207_503006, and T500105_500617), and the other B-clade plasma from Thailand (T500105_286588) clusters with two CRF01 plasmas from Thailand (T500106_501602 and T500104_256254). Bootstrap support was 80% of 1000 replicates for the first cluster, on the node that includes all 11 plasmas, and 91% for the smaller cluster. Two other CRF01 plasmas, both from Thailand (R163b and T500107_357545), did not cluster with any of these.

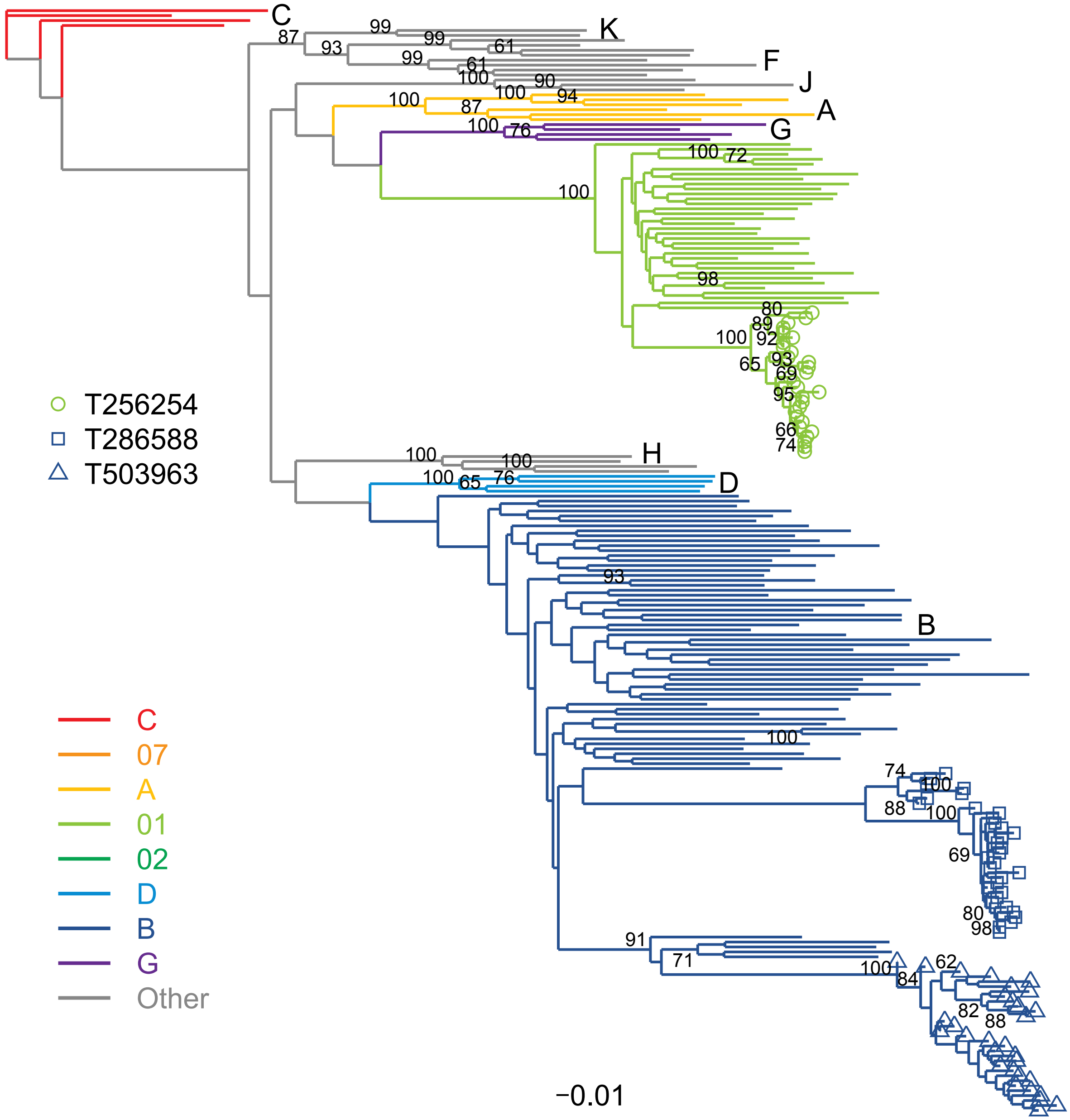

Figure S3. Single-clade infections in three Thai plasmas. Maximum-likelihood phylogeny of *env* nucleotide sequences from 59 B-clade plasmas, 34 CRF01 plasmas and pseudoviruses, and 39 non-recombinant M-group subtype reference sequences. Sequences from the 3 plasmas with shared profiles (T256254, T286588, and T503963) are indicated by geometric symbols. Where shown, node labels indicate over 60% bootstrap support from 100 resampled replicates.

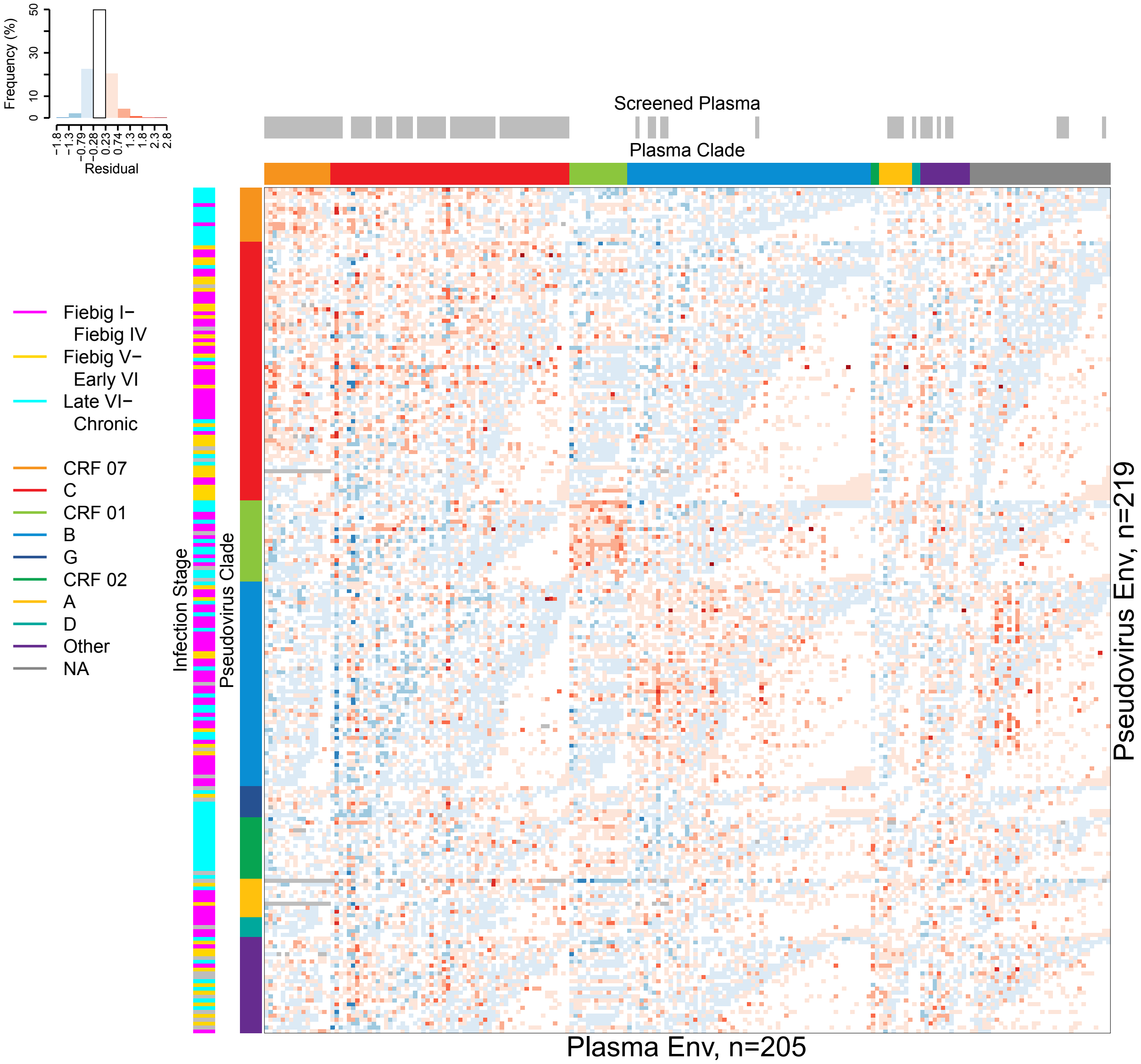

Figure S4. Residual ID50s obtained by subtracting the non-specific row/column effect. Heatmap of residual ID50s shows the difference between observed and approximated values after removing non-specific effects of the approximation.

Table S1. Correlation of geometric mean ID80 with hypervariable loop properties per virus and plasma Env.

	All Data[1]			Positive Values[2]			No Low Plasma[3]			Pos/No Low Plasma[4]		
	τ[5]	p[6]	q[7]	τ	p	q	τ	P	q	τ	p	q
Virus ID80[8]		n=219			n=219			n=219			n=219	
V1 length[9]	-0.21	4e-06	1e-04***	-0.24	2e-07	9e-06***	-0.22	3e-06	8e-05***	-0.23	4e-07	2e-05***
V1 netchg[10]	0.18	3e-04	0.005***	0.12	0.016	0.061*	0.18	3e-04	0.005***	0.12	0.015	0.060*
V1 glycos[11]	-0.11	0.038	0.111	-0.18	3e-04	0.005***	-0.11	0.038	0.111	-0.18	4e-04	0.006***
V2 length	-0.07	0.160	0.291	-0.12	0.009	0.042**	-0.07	0.155	0.285	-0.12	0.009	0.043**
V2 netchg	0.10	0.057	0.143	0.13	0.012	0.048**	0.10	0.053	0.136	0.13	0.010	0.046**
V2 glycos	-0.10	0.049	0.129	-0.16	0.003	0.026**	-0.10	0.046	0.128	-0.16	0.003	0.028**
V1/V2 length	-0.26	3e-08	2e-06***	-0.29	4e-10	1e-07***	-0.26	2e-08	2e-06***	-0.29	7e-10	1e-07***
V1/V2 netchg	0.17	5e-04	0.007***	0.14	0.004	0.031**	0.17	4e-04	0.006***	0.14	0.004	0.030**
V1/V2 glycos	-0.16	0.001	0.013**	-0.27	7e-08	4e-06***	-0.16	0.001	0.013**	-0.27	1e-07	5e-06***
V4 length	-0.03	0.500	0.578	-0.03	0.499	0.578	-0.04	0.461	0.555	-0.03	0.590	0.633
V4 netchg	0.05	0.307	0.438	0.06	0.215	0.351	0.06	0.248	0.384	0.06	0.265	0.395
V4 glycos	-0.02	0.684	0.691	-0.04	0.438	0.545	-0.02	0.672	0.689	-0.04	0.423	0.538
V5 length	-0.08	0.096	0.205	-0.14	0.006	0.034**	-0.08	0.097	0.206	-0.14	0.005	0.034**
V5 netchg	0.00	0.982	0.773	-0.03	0.628	0.663	0.00	0.977	0.773	-0.02	0.683	0.691
V5 glycos	-0.01	0.815	0.739	-0.15	0.005	0.031**	-0.01	0.834	0.742	-0.15	0.005	0.031**
Plasma ID80		n=170			n=164			n=129			n=129	
V1 length	0.07	0.203	0.334	0.05	0.323	0.444	0.13	0.027	0.088*	0.06	0.312	0.439
V1 netchg	0.00	0.953	0.770	0.06	0.316	0.441	0.03	0.680	0.691	0.04	0.524	0.597
V1 glycos	0.09	0.137	0.263	0.12	0.042	0.119	0.17	0.012	0.048**	0.12	0.073	0.169
V2 length	0.12	0.023	0.078*	0.07	0.178	0.308	0.07	0.261	0.394	0.04	0.512	0.590
V2 netchg	0.12	0.034	0.102	0.05	0.408	0.525	0.12	0.062	0.152	0.05	0.423	0.538
V2 glycos	0.01	0.835	0.742	0.00	0.955	0.770	-0.02	0.756	0.729	-0.04	0.527	0.597
V1/V2 length	0.14	0.010	0.046**	0.09	0.108	0.223	0.18	0.003	0.028**	0.09	0.162	0.292
V1/V2 netchg	0.06	0.259	0.394	0.07	0.200	0.334	0.10	0.134	0.261	0.08	0.219	0.355
V1/V2 glycos	0.09	0.123	0.245	0.10	0.080	0.180	0.14	0.031	0.096*	0.08	0.237	0.370
V4 length	-0.03	0.595	0.633	-0.06	0.266	0.395	-0.03	0.591	0.633	-0.06	0.311	0.439
V4 netchg	0.00	0.964	0.770	0.06	0.320	0.444	0.02	0.771	0.733	0.09	0.175	0.305
V4 glycos	-0.06	0.302	0.437	-0.06	0.355	0.470	-0.01	0.918	0.763	-0.04	0.552	0.618
V5 length	0.02	0.772	0.733	0.00	0.957	0.770	0.09	0.154	0.285	0.01	0.934	0.763
V5 netchg	-0.03	0.574	0.627	-0.04	0.453	0.552	0.00	0.971	0.773	-0.04	0.593	0.633
V5 glycos	-0.12	0.048	0.129	-0.07	0.235	0.370	-0.07	0.349	0.468	-0.09	0.203	0.334

[1] Censored values were taken as given, i.e. placeholder constants of 10 for ID50s below 20.

[2] Censored values treated as missing, i.e. only positive assay results were used.

[3] Plasma with geometric mean ID50 below 20 were excluded.

[4] Both plasma with low geometric mean ID50s and censored values were excluded.

[5] Kendall's τ as computed by the eponymous R package.

[6] Two-sided p value for the null hypothesis of no correlation.

[7] False-discovery rates computed from 360 p-values by the qvalue package. Significance levels: * q<0.10; ** q<0.05; *** q<0.01.

[8] Sample size (n) is listed for each data subset.

[9] Hypervariable loop boundaries are defined in Materials and Methods and not simply C-C.

[10] Net charge, i.e. number of K, H, and R sites minus the number of D and E sites.

[11] Number of potential N-linked glycosylation sites following the Nx[ST] motif, with x not P.

Table S2. Fixed-effects estimates of log-odds ratios for ID50s above the median. Where values were determined from other estimates, given the constraint that the total of fixed effects per factor equals zero, no standard error is reported.

Factor	**Estimate**	**Std Error**	**Z Value**	**Pr(>|Z|)**[1]
Intercept	0.69622	0.79275	0.878	0.3798
Stage: Early	-0.20943	0.08604	-2.434	0.0149 *
Stage: Late	0.20943			
Match: CRF07 (1)	0.46637	0.18851	2.474	0.0134 *
Match: CRF01 (2)	1.04076	0.19650	5.297	1.18×10^{-7} ***
Match: Other (3)	-0.22463	0.09178	-2.448	0.0144 *
Match: NOT (4)	-1.28250			
Screened: No	-0.45068	0.14652	-3.076	0.0021 **
Screened: Yes	0.45068			
Plasma V1+V2	0.04457	0.02089	2.134	0.0328 *
Virus V1+V2	-0.02944	0.01257	-2.343	0.0191 *

[1] Significance levels: * p<0. 05; ** p<0.01; *** p<0.001.

Table S3. Fixed-effects estimates of log-odds ratios for above-median post-screened ID50s.

Effect	Estimate	Std Error	Z Value	Pr(>|Z|)[1]
(Intercept)	-0.05733	0.64848	-0.088	0.92956
Is.LateF	-0.26115	0.08446	-3.092	0.00199 **
Is.MatchF	-0.57418	0.02423	-23.700	$< 2\times10^{-16}$ ***
Plasma V1+V2	0.03912	0.01629	2.402	0.01632 *
Virus V1+V2	-0.02965	0.01168	-2.539	0.01113 *
Is.LateF:Is.MatchF	0.03371	0.02338	1.441	0.14947

[1] Significance levels: * p<0. 05; ** p<0.01; *** p<0.001.

Table S4. Fixed-effects log-odds ratio estimates per clade. As in Table S2, standard errors are not listed for entries inferred from the constraint that the sum of estimated effects per factor must be zero.

Factor	**Estimate**	**Std Error**	**z Value**	**Pr(>|z|)**[1]
Intercept	-0.66471	0.77895	-0.853	0.393468
Virus Clade: 7	0.66518	0.28625	2.324	0.020140 *
Virus Clade: 1	-0.18279	0.24519	-0.745	0.455971
Virus Clade: 2	0.78036	0.27341	2.854	0.004314 **
Virus Clade: A	0.15275	0.33566	0.455	0.649069
Virus Clade: B	-0.20809	0.17744	-1.173	0.240902
Virus Clade: C	0.02807	0.16632	0.169	0.865990
Virus Clade: D	-0.88457	0.45183	-1.958	0.050263 .
Virus Clade: G	0.35096			
Plasma Clade: 7	0.39927	0.52111	0.766	0.443558
Plasma Clade: 1	1.08505	0.52026	2.086	0.037016 *
Plasma Clade: 2	0.12513	1.06683	0.117	0.906631
Plasma Clade: A	0.38690	0.57906	0.668	0.504041
Plasma Clade: B	-0.62793	0.36428	-1.724	0.084753 .
Plasma Clade: C	0.66427	0.39240	1.693	0.090486 .
Plasma Clade: D	-2.03269			
Screened: No	-0.11932	0.22549	-0.529	0.596711
Screened:Yes	0.11932			
Plasma V1+V2 Length	0.03924	0.01932	2.031	0.042273 *
Virus V1+V2 Length	-0.01985	0.01172	-1.693	0.090445 .

[1] Significance levels: . p<0.10; * p<0. 05; ** p<0.01.

Table S5. Comparison of empirical and permuted correlations for row/column effect.

In each case, the observed r^2 was significantly greater than any obtained from 10,000 randomizations.

Dataset	**Observed r**[1]	**Observed r^2**	**Min Permuted r^2**	**Max Permuted r^2**
Montefiori	0.7	0.49	0.006	0.014
Doria-Rose et al.	0.82	0.6724	0.03	0.096
Seaman et al.	0.68	0.4624	0.042	0.12
Kulkarni et al.	0.76	0.5776	0.056	0.2

[1] Correlation coefficients between neutralization data and the corresponding row/column effect.