Identifying HIV Sequences that Escape Antibody Neutralization using Random Forests and Collaborative Targeted Learning

Yutong Jin*, David Benkeser

March 30, 2022

EMORY
ROLLINS SCHOOL OF PUBLIC HEALTH
Department of Biostatistics and Bioinformatics

Background: HIV Vaccine

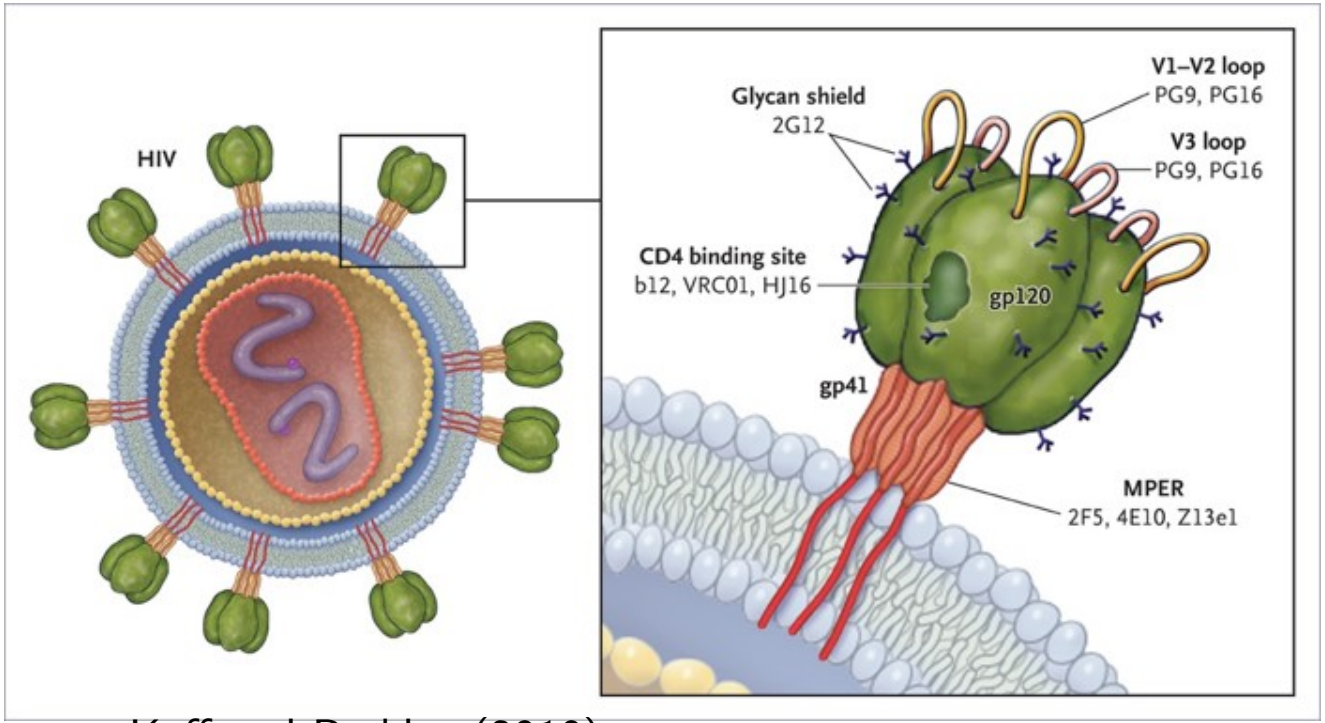
ABOUT

- In 2019, an estimated 1.1 million individuals were living with HIV in the United States and 36,801 new HIV diagnoses were reported (Centers for Disease Control and Prevention 2021).

"Despite nearly four decades of effort by the global research community, an effective vaccine to prevent HIV remains an elusive goal."

– Anthony S. Fauci, M.D.

Background: HIV Vaccine



source: Koff and Berkley (2010)

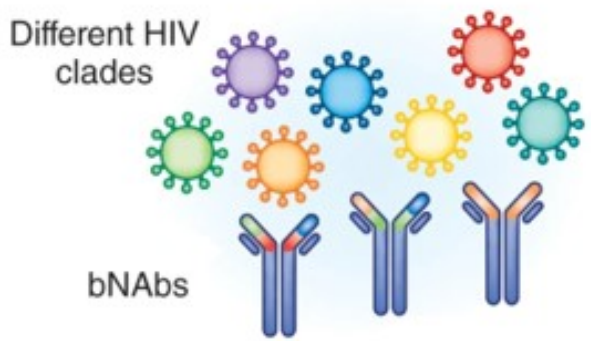
WHY is it so hard to make an HIV vaccine?

- HIV is highly genetically variable
  - **rapid** replication
  - prone to **constant mutations**
  - difficult to make a vaccine that can neutralize **broad variety of viral strains**

Background: Broadly Neutralizing Antibodies

Recent advances in HIV vaccine science and prevention focus on **broadly neutralizing antibodies (bnAbs)**

- neutralize a wide variety of HIV strains
- confer protective immune response
- can be optimized outside the human body



source: Tomaras and Haynes (2014)

This has shifted focus in the research pipeline towards:

- Vaccines that induce bnAbs, and
- Monoclonal antibody "cocktails" that provide protective breadth.

Motivation: Main Challenges

Key question

Which mutations on the HIV envelope (Env) protein will lead to a resistance to a certain antibody?

If we can identify such mutations...

We can combine multiple antibodies that target different antigens:

- If one antibody fails, choose others to fill holes
- Will antibodies that are effective in South Africa be effective in Thailand?

Motivation: Available Data

A glance of the data set

sensitivity	origin	subtype	hxb2.6	hxb2.8	...	hxb2.856
0	Asia	A1	N	Q	...	A
1	Asia	B	other	P	...	other
0	Africa	D	other	P	...	A
1	America	C	N	Q	...	A
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	Europe	B	N	Q	...	A

In the present settings, amino acid (AA) residues in the Env protein:

- are high dimensional,
- exhibit structural constraints (resulting in strong correlation)

Counterfactuals and Notations

Notation:

- Y: a dichotomous resistance outcome of whether the virus is sensitive to the particular antibody
- $\boldsymbol{W}$: a collection of J Env AA residues
  - W_j: a particular AA residue of interest
  - $\boldsymbol{W}_{-j}$: all residues except j
- $Y(W_j = w)$: a counterfactual resistance outcome that fixes the AA at residue j to $w \in \mathcal{W}_j$

Counterfactuals and Notations

Statistical problem of interest

- Given a particular AA at a given residue, how likely is it that the virus can be neutralized by a particular antibody?
  - **Estimation problem**: the probability of the sensitivity given certain AA
- Are there any AA residues that are important to the antibody resistance?
  - **Hypothesis testing**, e.g., $H_0 : \mu_j(w)$ is constant in w

Parameter of interest

We suggest answering these questions using estimation and inference about the parameter

$$\mu_j(w) = E[E(Y \mid W_j = w, \boldsymbol{W}_{-j})]$$

If certain key causal assumptions hold, then $\mu_j(w)$ equals the counterfactual probability of interest:

- Interpretation: the proportion of viruses that would be sensitive to neutralization if they had amino acid w at residue j

If causal assumptions do not hold, then $\mu_j(w)$ does not have a causal interpretation.

- Interpretation: "importance" of AA substitution (adjusting for other sequence features)

Causal Assumptions

The **main assumptions** needed for causal interpretation are:

Consistency Assumption

- the potential outcome under AA w at residue j is the outcome that will actually be observed when residue j is AA w.

Ignorability Assumption

- $Y(w) \perp W_j \mid \boldsymbol{W}_{-j}$

Positivity Assumption

- At each residue of interest, it is possible for all HIV Env sequences in the population to have various amino acids present.

General Templates for TMLE

Our test builds on the targeted minimum loss-based estimators (TMLE) (van Der Laan and Rubin 2006).

A TMLE procedure is used for estimating $\mu_j(w)$ for a particular j and w:

This procedure can be repeated for each $w \in \mathcal{W}_j$.

Efficient Influence Function (EIF)

Inference for $\hat{\mu}_j(w)$ can be drawn with the variance derived from its **influence function**.

The covariance of the vector of estimates $\hat{\boldsymbol{\mu}}_j = \{\hat{\mu}_j(w) : w \in \mathcal{W}_j\}$ can be consistently estimated by:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}} = n^{-1} \boldsymbol{D}_j^\top \boldsymbol{D}_j \ ,$$

where $\boldsymbol{D}_j$ is an $n \times |\mathcal{W}_j|$ matrix formed by stacking the row vectors of estimated IF of $\hat{\mu}_j(w)$ evaluated on each observation.

Hypothesis Testing: Wald-type Test

The null hypothesis $H_0 : \mu_j(w)$ is constant in w can be written:

$$H_0 : \boldsymbol{A}\boldsymbol{\mu_j} = 0$$

where $\boldsymbol{A}$ defines a contrast matrix that conduct pairwise comparison between potential AAs at residue j, e.g.,

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

The Wald-type test statistics with a $\binom{|\mathcal{W}_j|}{2} - 1$ degree-of-freedom:

$$T_j = (\boldsymbol{A}\hat{\boldsymbol{\mu}}_{\boldsymbol{j}})'(\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}}\boldsymbol{A})^{-1}(\boldsymbol{A}\hat{\boldsymbol{\mu}}_{\boldsymbol{j}})$$

Multiplicity corrections (e.g., Bonferroni) can be used to avoid spurious positives.

Additional Challenge in TMLE

Recalling that AA residues in the Env protein are:

- high dimensional
- highly correlated

Potential challenges

1. The estimated GPS may be extremely small for some pseudo-virus sequences.
2. The resulting $\hat{\mu}_j$ could be highly biased, with correspondingly inflated type I errors.

A Tentative Solution: Variable Pre-screening for GPS

Data-driven PS model-building

We propose using **variable importance measures** from an OR model to select variables to include in the GPS model.

Key question:

How many features from OR should be advanced into GPS?

One possible solution: collaborative TMLE (CTMLE)!

Provides an objective criteria for selecting the number of features to advance.

Details of CTMLE

Outcome-adaptive CTMLE implementation:

1. Fit OR model using random forests to get an initial estimator $\bar{Q}_n^{(1)}$.
   - The covariates are ranked by their feature importance.
2. Propose K potential values, $r_1, \ldots, r_K$, of the number of covariates to be included in the GPS model.
3. A sequence of GPS estimators can be constructed as $g_{n,j}^{(1)}, \ldots, g_{n,j}^{(K)}$.
4. Obtain $\bar{Q}_n^{*,(k)}$ by performing similar TMLE steps using $\bar{Q}_n^{(k)}$ and $g_{n,j}^{(k)}$, with $\mathrm{L}_j(\bar{Q}_n^{*,(k)}) \leq \mathrm{L}_j(\bar{Q}_n^{*,(k-1)})$.

Cross-validated CTMLE

Once K triplets have been derived, the best triplet k_n is selected through cross-validation.

A similar Wald-type hypothesis testing can be performed for the cross-validated CTMLE estimate.

Simulation Studies

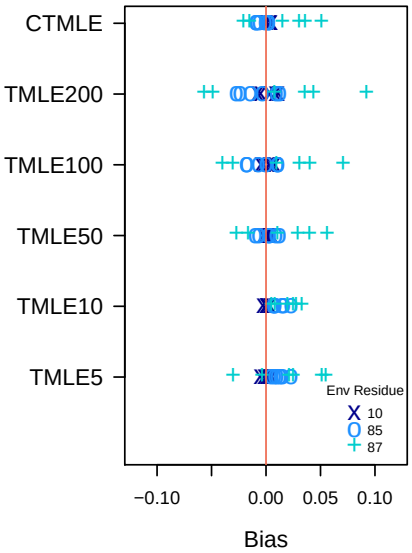
Simulation setup:

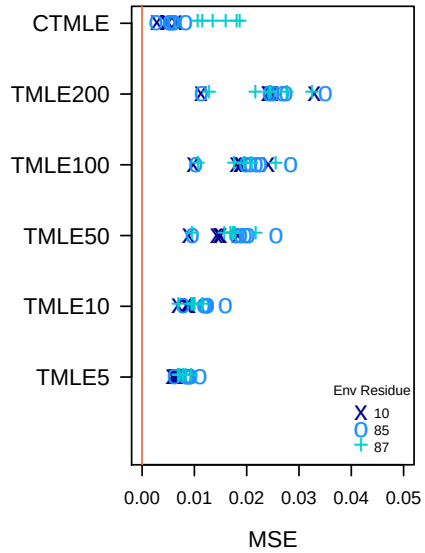
- Sample size: **500 / 1000**
- Number of features (Residues): **200**
- Number levels for each feature: **4**
- **AR-1 correlation** with $\rho = 0.75$ (moderate correlation)
- True signals: AA_{37}, AA_{87}, AA_{94}, AA_{135}, AA_{151}
- Number of possible PS features: $\{5, 10, 50, 100, 200\}$

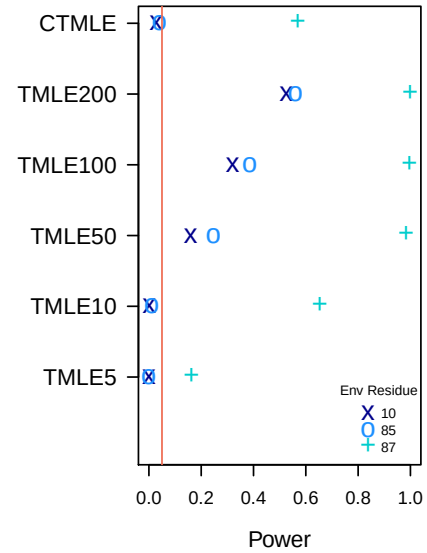
	β_{j1}	β_{j2}	β_{j3}	β_{j4}
W_{37}	0.160	-0.321	-0.492	0.214
W_{87}	0.181	0.521	-0.612	0.321
W_{94}	0.104	0.414	-0.789	-0.117
W_{135}	0.178	0.350	-0.453	-0.433
W_{151}	0.072	0.311	0.638	-0.320

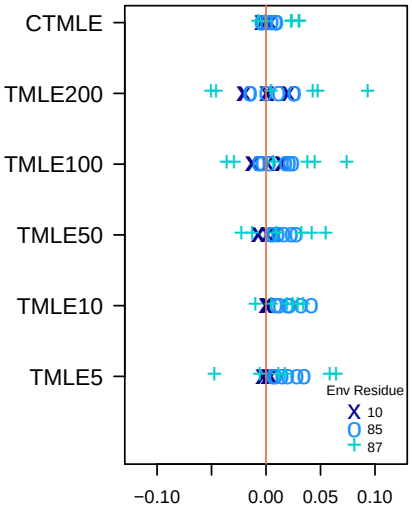
Table: True coefficients (β) used in simulation study

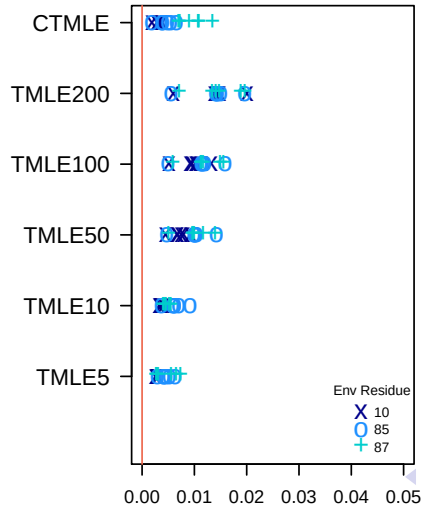
Simulation Results

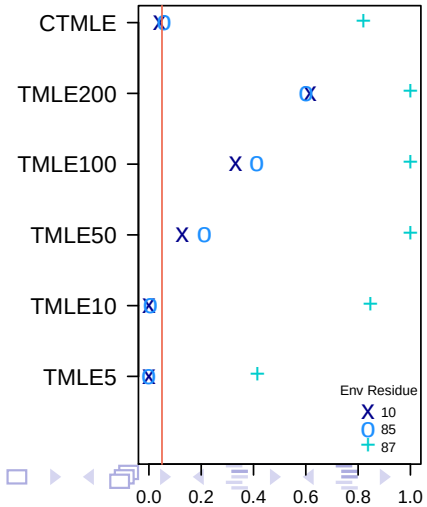










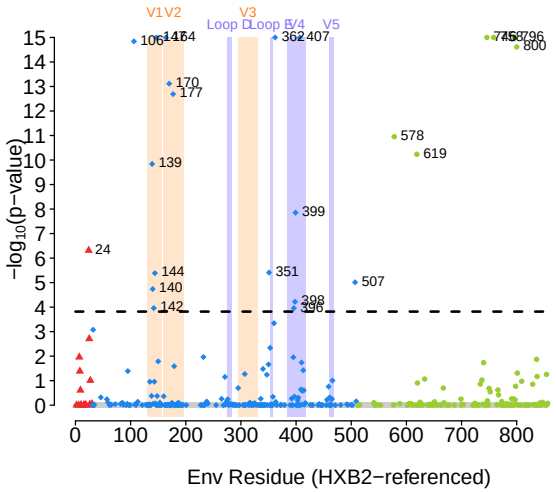


Real Data Analysis Results

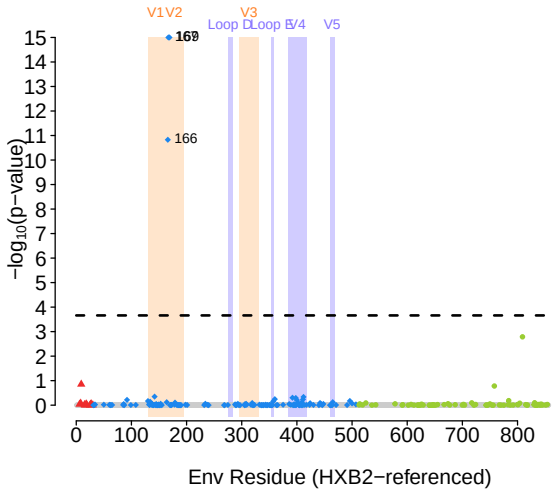
The Compile, Analyze and Tally NAb Panels (CATNAP) database (Yoon et al. 2015) consists of:

- binary antibody sensitivity (sensitive = 1)
- site-specific Env AA sequences
- other demographic records of virus
  - geographic origin
  - subtype
  - viral size
  - ...

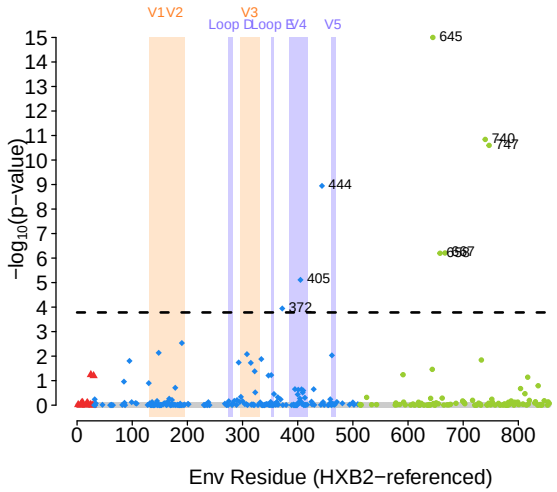
HIV Residue Results



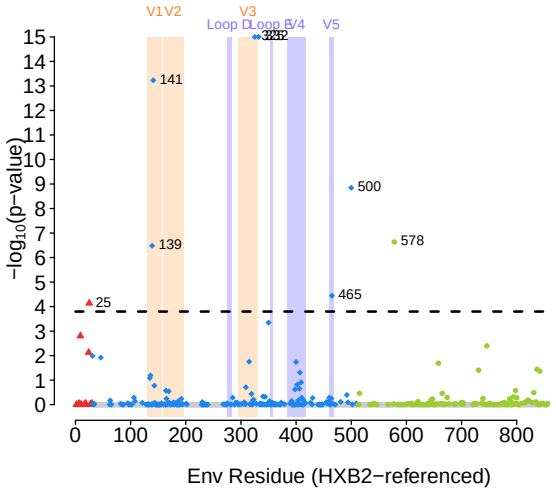
(A) VRC01



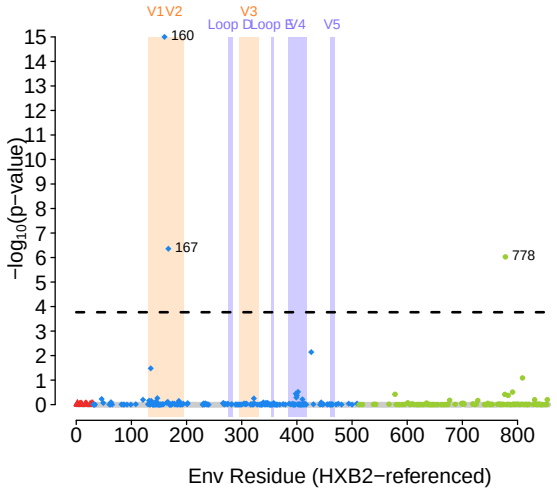
(B) VRC26.08



(C) 10−1074



(D) PGT121



(E) PGT145

signal peptide
gp120
gp41

Discussion

Future directions:

- Feature importance: Random Forest ⇒ other algorithms
- Pairwise comparisons ⇒ family-wise error rate control methods

References I

Centers for Disease Control and Prevention. Diagnoses of hiv infection in the united states and dependent areas, 2019. `http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html`, 2021. Accessed [June 21, 2021].

W. C. Koff and S. F. Berkley. The renaissance in hiv vaccine development—future directions. *New England Journal of Medicine*, 363(5):e7, 2010.

G. D. Tomaras and B. F. Haynes. Lessons from babies: inducing hiv-1 broadly neutralizing antibodies. *Nature medicine*, 20(6):583–585, 2014.

M. J. van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

H. Yoon, J. Macke, A. P. West Jr, B. Foley, P. J. Bjorkman, B. Korber, and K. Yusim. Catnap: a tool to compile, analyze and tally neutralizing antibody panels. *Nucleic acids research*, 43(W1):W213–W219, 2015.

Thank You!