The Molecular Basis of pH-Modulated HIV gp120 Binding Revealed

Evolutionary Bioinformatics
Volume 15: 1–18
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934319831308
SAGE

Scott P Morton[1], Julie B Phillips[2] and Joshua L Phillips[1,3]

[1]Center for Computational Science, College of Basic and Applied Sciences, Middle Tennessee State University, Murfreesboro, TN, USA. [2]Department of Biology, Cumberland University, Lebanon, TN, USA. [3]Department of Computer Science, College of Basic and Applied Sciences, Middle Tennessee State University, Murfreesboro, TN, USA.

ABSTRACT: Decades of research has yet to provide a vaccine for HIV, the virus which causes AIDS. Recent theoretical research has turned attention to mucosa pH levels over systemic pH levels. Previous research in this field developed a computational approach for determining pH sensitivity that indicated higher potential for transmission at mucosa pH levels present during intercourse. The process was extended to incorporate a principal component analysis (PCA)-based machine learning technique for classification of gp120 proteins against a known transmitted variant called Biomolecular Electro-Static Indexing (BESI). The original process has since been extended to the residue level by a process we termed Electrostatic Variance Masking (EVM) and used in conjunction with BESI to determine structural differences present among various subspecies across Clades A1 and C. Results indicate that structures outside of the core selected by EVM may be responsible for binding affinity observed in many other studies and that pH modulation of select substructures indicated by EVM may influence specific regions of the viral envelope protein (Env) involved in protein-protein interactions.

KEYWORDS: HIV, Env, gp120, CD4, electrostatics, binding, pH, mucosa

RECEIVED: December 2, 2018. **ACCEPTED:** December 17, 2018.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Scott P Morton, Center for Computational Science, College of Basic and Applied Sciences, Middle Tennessee State University, Murfreesboro, TN 37130, USA. Email: spm3c@mtmail.mtsu.edu

Introduction

AIDS was discovered nearly 40 years ago, but a vaccine for the virus that causes the disease remains imponderable. The challenge that researchers face is the overwhelming mutation rate of the virus due to host immune system pressure after introduction into the body.

HIV is typically transmitted during sexual intercourse where an acidic mucosa pool exists. Because protein structures and their ability to interact with other proteins are affected by pH,[1] we focus our attention on this key component. HIV transmission occurs when the gp120 portion of the viral envelope protein (Env), attached to the periphery of the virus, makes contact with a CD4 protein receptor on host T-cell membranes. Interaction between the 2 structures initiates a binding process and subsequent introduction of the viral RNA.

A study completed by Boeras et al[2] in 2011 concluded that the highest populations of HIV variants are not the subspecies that transmit from one host to the next. Their determinations were backed by statistical analysis of population subspecies and transmission data through direct investigation of human volunteer donors. With the large pool of quasi-species extracted, and the capture of variants at the time of transmission, this data set presents a potential to determine differences in protein structure and the role of pH that may explain the transmission bottleneck.

We focus our efforts around the sequences provided by Boeras et al as the foundation of our latest theoretical methods in an effort to narrow the field of research to those Env quasi-species with a higher potential of producing an infection from host to host.

Background

The high rate of mutation obtained by HIV allows antigenic regions targeted by host immune responses to vary greatly across HIV variants. Most research has focused on inducing the so-called broadly neutralizing antibodies (bnAbs) that target protein antigenic regions conserved due to functional requirements of the binding process.[3] The gp120 extracellular subunit of Env is responsible for binding CD4 on the surface of host T cells to begin infection; this subunit is a common target for bnAbs.[4] Env fragments selected via computational optimization to potentially invoke the production of bnAbs are often employed in current work for vaccine production.[5] Studies using these methods have varied from successful[6] to unsuccessful.[7] One potential explanation is that environmental impacts on gp120-CD4 interactions are not considered during Env selection. In particular, isolating bnAbs from a blood/plasma environment (slightly basic pH) might obfuscate the impact of mucosal environments (often acidic pH) on transmission. Therefore, it is reasonable to assume that both Env structure and binding affinity with CD4 and/or bnAbs will be altered under physiological conditions which are more consistent with sexual transmission.

Recent experimental and computational studies have shown that pH does in fact impact both Env conformation and CD4 binding. In 2013, Stieh et al hypothesized that electrophoresis, which is commonly used to characterize and separate cells and micro-organisms,[8,9] could be applied at a protein level and performed direct experimentation to reveal a pattern of change in surface electrostatics across the pH range of the human body. Their findings produced a fingerprint of trimeric gp120 indicating a change in electrophoretic mobility from negative toward positive as pH increased.[1] The study was performed in a multidisciplinary, collaborative effort with computer scientists to develop a corresponding analytical protocol using off-the-shelf general public license (GPL) based software. The pipeline produced similar results to those of laboratory experiments developed by Stieh et al in that a determinable difference was seen from negative to positive with advancing pH. Stieh et al concluded that the experimental process and the computational data were in agreement.

In 2016/2017, Morton et al enhanced and refined the process introduced by Stieh et al to incorporate protein modeling via Modeler,[10] parallel processing, structure energy minimization by Gromacs,[11,12] and advanced floating point data compression through ZFP[13] that allowed for larger studies to be performed and a greater depth of analysis to take place.[14] A classification method called Biomolecular Electro-Static Indexing (BESI) was developed based on principal component analysis (PCA), cosine similarity analysis (CSA), and loosely based on latent semantic indexing (LSI). Nearly 1 million adaptive Poisson-Boltzmann solver (APBS)[15] calculations were executed by Morton et al with the entire computational process taking approximately 60 days to complete on a small compute cluster with 256 cores.

During 2016 to 2017, Howton and Phillips[16] introduced a prototype method that extended Stieh et al to the protein residue level. The approach used by these authors exercised the hypothesis that strains in chronic infection, the so-called chronic control (CC) strains, will likely have adapted to systemic pH and will be less efficient at binding CD4 under acidic conditions when compared with transmitted founder (TF) strains. Using computational modeling, some differences between subclasses (TF and CC) and clades (B and C) were discovered using a more extensive set of 28 Env proteins.[16] However, the specific molecular mechanism (eg, surface residues and mutations) responsible for the pH sensitivity of the gp120-CD4 interaction could not be determined using the resulting data. The main difficulty was assumed to stem from a small sample size and a broad range of sexual-transmission-type studies.

In 2018, Morton et al[17] developed a method of protein residue analysis that examines the surface charge fluctuations of amino acids called Electrostatic Variance Masking (EVM). This method aligns all sequence structures together and determines the charge variance of exposed surfaces across the set. This information is then used to image those amino acids via transparency against a representation of the structure in an alternate mode such as New Cartoon in VMD.[18] The imagery produces a unique view of charge active residues that are similar across all structures examined to date. The process reveals what were hypothesized to be the residues responsible for modulating the binding process by exposing the high variation of electrostatic charge across the pH range of the human body.

Table 1. List of donors taken from Boeras et al.[2]

SUBJECT	STATUS	CLADE	SCORES (HI/LO)
R56M	D	A1	0.914/0.069
Z153F[a]	D	C	0.781/0.400
Z185M[a]	D	C	0.758/0.499
Z201F	D	C	0.938/0.186
Z205M[a]	D	C	0.750/0.576
Z216F	D	C	0.777/0.443
Z221F	D	C	0.869/0.088
Z238F	D	C	0.892/0.352
Z242M	D	C	1.00/0.057
Z292F	D	A1	0.870/0.138

BESI, Biomolecular Electro-Static Indexing.
Subject indicates the country of origin, couple identifier, and sex, respectively.
D indicates the subject's status as the donor. Scores are the highest and lowest BESI scores for the sequence set.
[a]The subject pair is not mentioned in Boeras et al's study.

Target Data

From a pool of more than 900 HIV Env sequences, Boeras et al provided 252 gp120 protein assemblies drawn against 20 individuals from Rwanda and Zambia. The donors consisted of couples where one was known to be infected and the other was expected to acquire infection at some point. Samples were taken prior to communication of the disease and after infection of the recipient occurred. The naming conventions used for the sequences indicate the country of origin, the sex, a subject pair identifier, and a donor (D)/recipient (R) indicator. Our selection of sequences is based on the BESI scores of the donor sequence data for each couple and is represented in Table 1.

We use previously processed data from Morton et al[14] to reduce the overall processing time considerably.

Methods

Residue surface charges

We calculate the charges of individual amino acids that have solvent-accessible surfaces as described by Howton and Phillips,[16] enhanced and performed by Morton et al,[17] that include energy minimization steps performed by Gromacs[11,12] and compression levels approaching 2 orders of magnitude provided by ZFP.[13] With the latter enhancement, we are able to process larger studies across more solvation states that allow a more granular investigation of the substructures involving gp120.

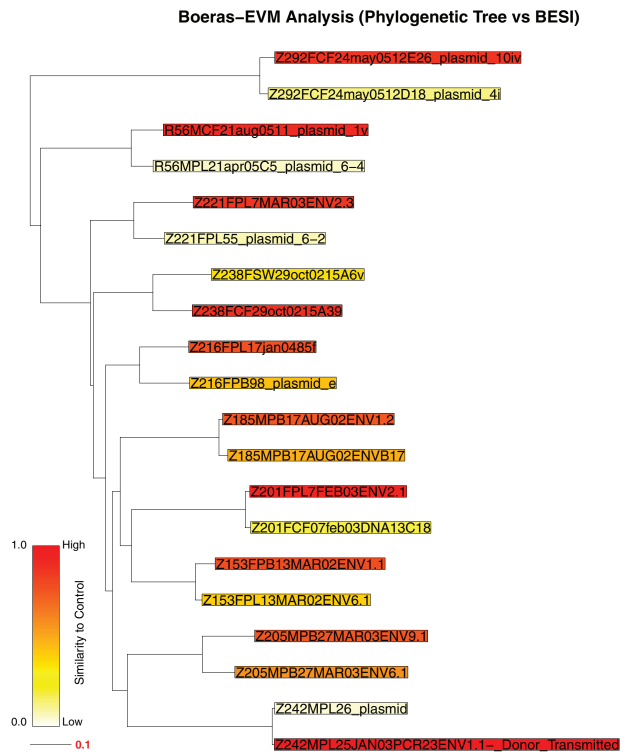


Figure 1. BESI vs phylogeny for the selected highest and lowest BESI scores applied as a gradient to the phylogenetic tree. As the BESI scores increase, the shading moves more toward the red. Each subclade of the tree is a specific donor in the study.
BESI, Biomolecular Electro-Static Indexing.

Phylogenetic tree

The phylogenetic tree inferred for the selected high and low BESI scores for each donor is constructed as follows. Sequences were aligned with MAFFT v7.273 using the L-INS-i strategy.[19] A maximum likelihood (ML) phylogenetic tree was inferred using the RAxML software, version 8.2.11,[20] with the HIVW amino acid model of substitution[21] and 100 bootstrap replicates. Trees were midpoint-rooted and rendered using APE version 5.0.[22] Expression of the phylogenetic tree involves minor differences from Morton et al[14] where recipient sequences are unused for this study.

BESI

With the focus of investigation being the transmission of the virus, our attention is directed to the donor group from Boeras et al. Using BESI as prescribed in Morton et al,[14] we select the maximum and minimum scores available from each donor into a correlation of BESI and phylogeny to produce Figure 1 which provides a graphical representation of BESI and evolution. One can see that, for each subclade of the tree, a higher and a lower score have been selected based on the gradient scale left of the inference. Note that the inferred tree also distinctly differentiates between donor categories where the sequence name represents the country Zambia (Z) or Rwanda (R) with a 3-digit code for a subject number. The fifth character is gender specific which is self-explanatory. All additional characters are attributes of the sequence that are explained in Boeras et al[2] if the reader chooses.

The reader should note that at this point no additional calculations have been made with the data; we have simply selected a subset of what was processed in Morton et al[14] and presented the results in a different manner.

Electrostatic variance masking

Selection of residues that show surface charge response to pH shifts involves calculating the electrostatic potential variance of each residue across all aligned sequences vertically. Where gaps are encountered in the alignment, a value of 0 is assigned. For each residue, the median value of individual residues for each model at a specific pH is taken to create a 1×61 vector for the pH range of 3.0 to 9.0 in 0.1 increments. The vectors are stacked row by row to create an array of dimensions $M \times 61$, where M is the number of sequences involved in the study. The mean value of each column is then calculated to produce a vector for which the variance is determined and stored. This is repeated for each alignment position. This method allows us to effectively filter out residues with small variations in mean surface charge across the pH shift.

For each sequence alignment, a reverse mapping is created to align selections with correct residue numbers on the individual proteins. Where a gap exists in the alignment, a hyphen (-) is assigned. This allows the determination of a cutoff value for variance where a selection of a gap in some determined sequence can easily be detected. To determine a starting value for selection, the ceiling of one-half the standard deviation is calculated for the variance data. Assuming a gap is selected, the value is incremented by 1 until a uniform selection across all sequences can be determined.

The selected residues of the gp120 protein are then applied to a VMD representation[18] to display the substructures involved. This method of imaging residue structures participating in the mechanistic functions of the binding process is EVM.

HXB2CG alignment

We align the assemblies to HXB2CG as described by Korber-Irrgang et al[23] in *Numbering positions in HIV relative to HXB2CG*. This provides a common numbering scheme for amino acids and allows us to describe those residues that EVM selects in a concise manner.

Comparing BESI and variable loop lengths

For each sequence used in this study, we extract residue information directly from amino-acid-based text files. The 5 variable loops associated with HIV are extracted by aligning to HXB2CG and clipping the loops inclusively at defined residue numbers provided by Los Alamos National Laboratory

(LANL)[24] in the *HXB2 annotated spreadsheet*. The information is correlated with BESI scores and presented via scatter plots grouped by variable loop number.

Tropism of loop V3

A method of prediction for a V3 tropism test of the major HIV-1 subtypes was developed by Cashin et al[25-29] to determine specificity for CCR5 and CXCR4 usage during the binding process. We extract this information using the provided web-based tool and present the data in table format as a comparison of co-receptor predicted binding mode, BESI score, and clade.

Results

Our results are focused around EVM, HXB2, variable loop lengths, and V3 tropism. We skip through BESI as the information used here is an extension of the analysis performed by Morton et al[14] and presented in *High-Throughput Structural Modeling of the HIV Transmission Bottleneck*. Sequence data are available in Appendix 1.

EVM

Performing the process prescribed by EVM produced a uniform selection of amino acids for each of the chosen sequences. Statistical information returned from this data set is as follows:

Standard deviation	123.7
1/2 Standard deviation	61.8
Number of selected residues	56.0
Variance cutoff selected	65.0
Percentage of variance selected	73.6
Percentage of residues selected	11.0

The variance data for the entire set of gp120 structures analyzed in this study are displayed in Figure 2. We separated Clades A1 and C into 2 graphs to display differences in Figures 3 and 4. We note similarities between the 2 representations and differences in amplitude for later analysis. A scree plot is generated to provide a sorted view of the data in Figure 5. The red horizontal line indicates the cutoff value chosen.

For the purposes of this discussion, we have selected a single gp120 structure as the subject of explanation for all the remaining graphics. The calculated sequence-based residue map for this Env is as follows:

- R56MCF21aug0511_plasmid_1v

14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 160 162 175 177 206 210 211 212 214 215 221 222 224 234 244 248 255 257 329 330 335 337 377 380 382 396 398 401 406 422 423 424 425 426 428 430 431 433 434 436 438

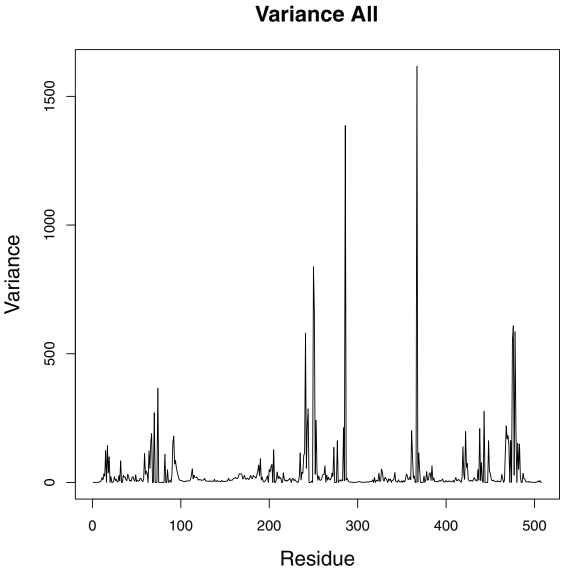


Figure 2. Raw EVM plot of the variance values for the aligned protein sequences. The small subset of amino acids (11.0%) experiencing surface charge modulation due to varying pH levels at or above the selection value (variance = 65) contain the largest amount of variation (73.6%).
EVM, Electrostatic Variance Masking.

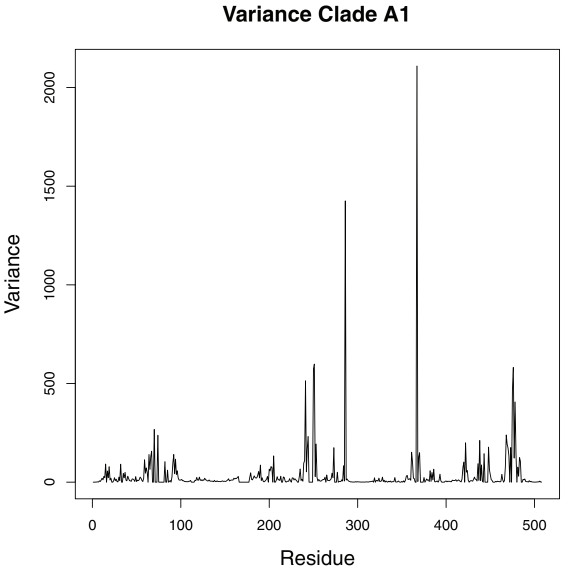


Figure 3. Raw EVM plot of the variance values for the aligned protein sequences of Clade A1. Compared with Figures 2 and 4, differences exist mainly in amplitude.
EVM, Electrostatic Variance Masking.

We apply the amino acid maps in VMD by first creating an additional representation in the interface. We use "New Cartoon" colored by secondary structure to represent the entire assembly. The second representation is limited to the selected residues provided by EVM as a single color (red) in transparency. Figure 6 is marked to present the $\alpha 2$ helix oriented left of the binding site and labeled accordingly. All the remaining images of Env structure and substructure are oriented identically for this article. All sequence pair imagery can be examined as shown in Appendix 1. The region of selection is highly

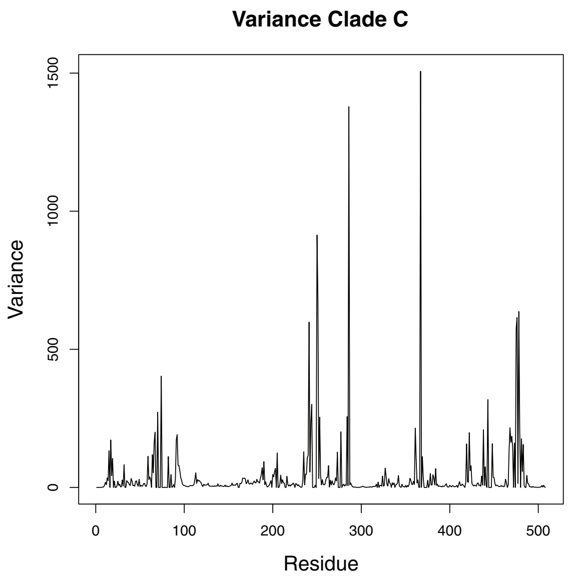


Figure 4. Raw EVM plot of the variance values for the aligned protein sequences of Clade C. Compared with Figures 2 and 3, differences exist mainly in amplitude.
EVM, Electrostatic Variance Masking.

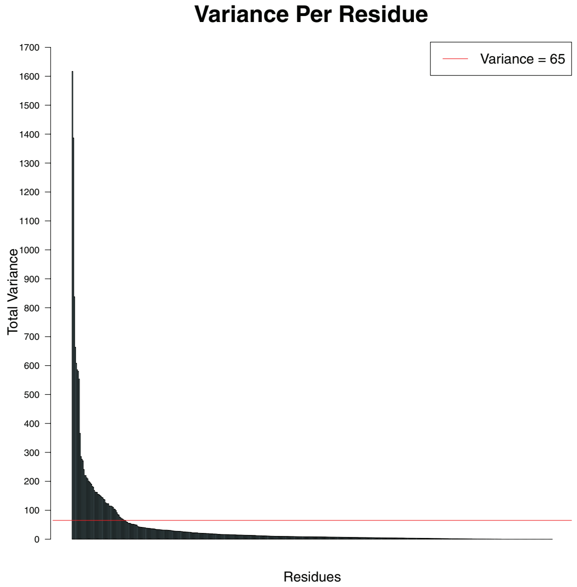


Figure 5. Scree plot of the variance values for the aligned protein sequences. The red line indicates the selected cutoff value displaying the large amount of variance (73.6%) across a small subset of amino acids (11.0%).

conserved and localized at the Env center. Assuming that the process continues to provide similar results for the other analyzed structures, the power of the tool to exhibit differences in assembly makeup will become apparent.

To further expound on the selection process, a WebLogo[30,31] representation is generated for the aligned sequences. Sequence logos present a unique method of graphical representation that displays the presence of like amino acids across the set of sequences by lettering height. Figure 7 displays the logos for all selected substructures in this study, and Figures 8 and 9 produce the logos for sequences in Clades A1 and C, respectively. We again note the minor discrepancies in content between the 2 clades for future analysis and disregard the differences in height due to the number of sequences present in each clade of this study.

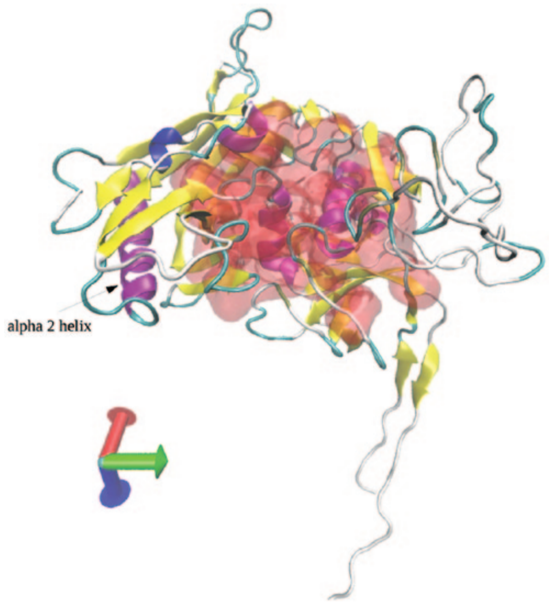


Figure 6. Electrostatic Variance Masking of R56MCF21aug0511_plasmid_1v. This figure indicates the orientation of the assembly with the $\alpha 2$ helix situated to the left and distinguished by the label and arrow to confirm the binding site position. All images of Env structure and substructure are oriented identically for this article.
Env, viral envelope protein.

HXB2CG characteristics

For this study, we aligned all assemblies to HXB2 using the procedure described by Korber-Irrgang et al[23] in *Numbering positions in HIV relative to HXB2CG*. Residue selections provided by EVM and mapped back to HXB2 position identification via the annotated spreadsheet[24] were identical containing the following list:

47 49 51 64 91 96 98 99 102 106 114 123 124 125 126 199 201 214 216 245 249 250 251 253 254 260 261 263 273 283 287 294 296 370 371 376 378 426 429 431 445 447 450 455 470 471 472 473 474 476 478 479 481 482 484 486

Per the annotated spreadsheet, we note the following pertinent EVM selections: Residues 64 and 91 are adjacent to 65 and 92, respectively, which are interface contacts with gp41; 123 is a co-receptor binding site outside of V3 and adjacent to 122 of the same function; 124 to 126 are CD4 contact residues; 199 is a co-receptor-specific (R5/X4) site; 201 is adjacent to 202 which is a co-receptor binding site outside of V3; 249 to 251 where 251 is a co-receptor-specific (R5/X4) site; 253 is adjacent to 252 which is an interface contact with gp41; 261

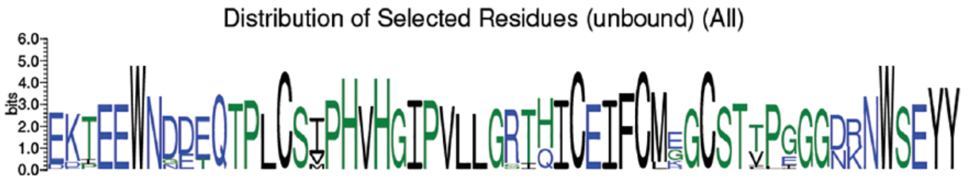


Figure 7. Sequence logos representation of the EVM selection process for all structures. The figure displays the conservation of residues in the EVM process across all sequences.
EVM, Electrostatic Variance Masking.

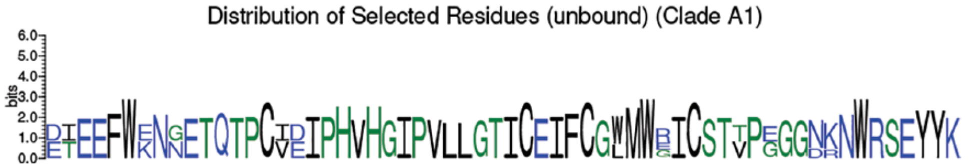


Figure 8. Sequence logos representation of the EVM selection process for structures in Clade A1. The figure displays the conservation of residues in Clade A1.
EVM, Electrostatic Variance Masking.

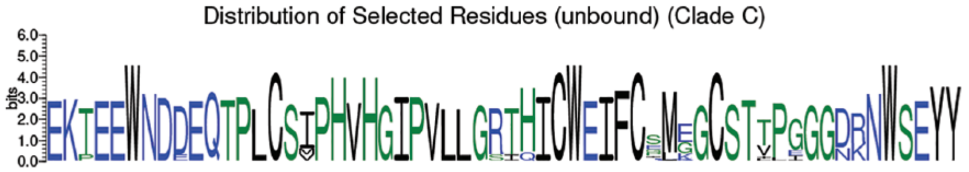


Figure 9. Sequence logos representation of the EVM selection process for structures in Clade C. The figure displays the conservation of residues in Clade C.
EVM, Electrostatic Variance Masking.

and 263 are adjacent to glycosite 262; 283 is a CD4 contact residue; 294 is adjacent to glycosite 295; 296 is the beginning of V3 loop; 370 is a CD4 contact residue and 371 is adjacent; 376 is adjacent to 377, a co-receptor binding site outside of V3; 378 is cystine linked to a counterpart at 445; 426, 429, and 431 are CD4 contact residues; 445 is cystine linked to a counterpart at 378; 447 is adjacent to glycosite 448; 455 is a CD4 contact residue; 470 is V5 loop end and adjacent to CD4 contact residue 469; 471 to 476 are CD4 contact residues.

Variable loop lengths

Derdeyn et al[32] observed that transmitted quasi-species appeared to have shorter variable loop lengths than the larger populations of the donor. Whereas our study focuses on the donor-specific envelope structures based on BESI score, we observe the research of Boeras et al[2] in that a small subset of quasi-species actually cross the transmission barrier. Our observation conjoins the 2 aforementioned examinations to reason that Derdeyn et al observed an attribute of the transmission bottleneck.

We applied BESI scores to variable loop lengths in scatter plots for the sequences used for this study. Figures 10 and 11 present a discernible correlation with BESI score and variable loop length. We note that our small sample size may influence other observations in this regard and distinguish loops V2 and V5 as exposing the potential need for further investigation. All variable loop graphs can be examined as shown in Appendix 1.

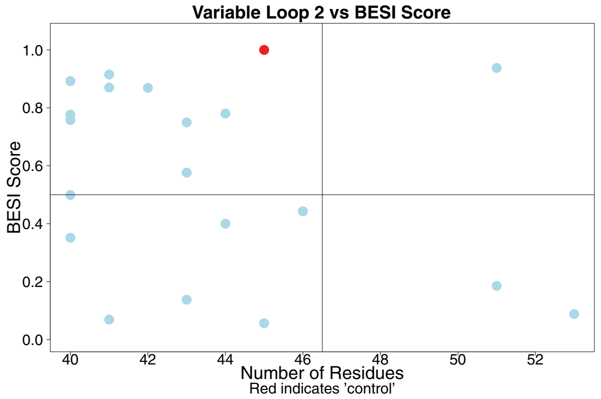


Figure 10. Scatter plot of variable loop V2 in comparison with BESI scores. Red indicates the control quasi-species. In total, 55% of the residues fall at or inside the quadrant containing our control variant (red dot) which is a BESI score of 0.5 or greater.
BESI, Biomolecular Electro-Static Indexing.

V3 co-receptor tropism

Our investigation into the usefulness of BESI, what the process is keying on and how to best describe using the method, requires the examination of peripheral attributes. Here we have applied a method of predicting the tropic mode of V3 co-receptor binding using a process developed by Cashin et al[25-29] (Table 2). Although the resulting data are inconclusive, we provide the same explanation of limited sample size as a potential detriment to the observations.

Discussion

HIV has been the focus of research around the world for nearly 4 decades. The virus has eluded scientists over this period due to the fast mutation rates and the ability to overwhelm the human immune response system. With some progress being made in the quality and length of life, a vaccine has still to be determined for the infectious disease.[33] Typical studies of the binding site overlook the significance of pH and the effects acidic fluids, common in genital mucous, have on antibody binding functions.[34]

The binding site of gp120 is correctly identified through residue-specific pH sensitivity without the need to determine the area based on the presence of CD4 or other counterpart protein structures. In addition, the selection process identifies highly reactive residues adjacent to common glycosite and specific co-receptor positions implying that pH modulation of these amino acids could influence activities common to those locations.

We note that our sample size is too small to evaluate against variable loop lengths although loops V2 and V5 do indicate the potential to produce interesting data. Furthermore, we observe no discernible characteristics in a comparison of V3 co-receptor tropism, BESI, and clade based on the small sample size used here.

These results suggest that the highly conserved and localized amino acid cluster is not responsible for variation in the ability of a particular mutation to infect another cell, but the variation of the

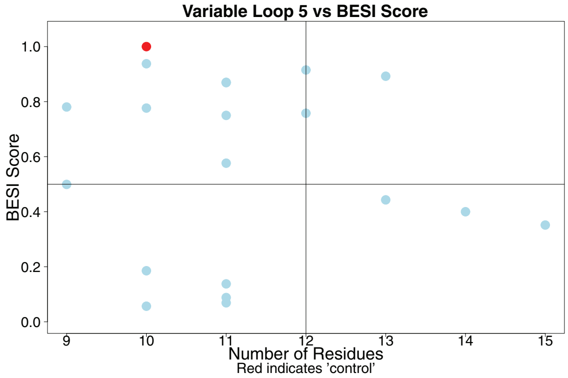


Figure 11. Scatter plot of variable loop V5 in comparison with BESI scores. Red indicates the control quasi-species. In total, 50% of the residues fall at or inside the quadrant containing our control variant (red dot) which is a BESI score of 0.5 or greater.
BESI, Biomolecular Electro-Static Indexing.

Table 2. Tropic mode of V3 co-receptor binding in comparison with BESI score and Clade revealing no pattern of distinction available with the current sample size.

SAMPLE	MODE	SCORE	CLADE
R56MCF21aug0511_plasmid_1v	R5	0.914998937284965	A1
R56MPL21apr05C5_plasmid_6-4	R5	0.0690380359279373	A1
Z153FPB13MAR02ENV1.1	R5	0.780562081434307	C
Z153FPL13MAR02ENV6.1	R5	0.400037903712386	C
Z185MPB17AUG02ENVB17	R5	0.499258854685399	C
Z185MPB17AUG02ENV1.2	R5	0.75789064171072	C
Z201FPL7FEB03ENV2.1	R5	0.937733719093409	C
Z201FCF07feb03DNA13C18	R5	0.185587639941212	C
Z205MPB27MAR03ENV9.1	R5	0.750062534639445	C
Z205MPB27MAR03ENV6.1	R5	0.576458537237502	C
Z216FPL17jan0485f	R5	0.776617270079055	C
Z216FPB98_plasmid_e	R5	0.443137527302722	C
Z221FPL55_plasmid_6-2	R5	0.088284202030072	C
Z221FPL7MAR03ENV2.3	R5	0.869016547898083	C
Z238FSW29oct0215A6v	R5	0.351697287784035	C
Z238FCF29oct0215A39	R5	0.892405202671353	C
Z242MPL25JAN03PCR23ENV1.1-_Donor_Transmitted	R5	1	C
Z242MPL26_plasmid	R5	0.0567616933542945	C
Z292FCF24may0512E26_plasmid_10iv	CXCR4-using	0.870187354349578	A1
Z292FCF24may0512D18_plasmid_4i	CXCR4-using	0.137851938524118	A1

BESI, Biomolecular Electro-Static Indexing.

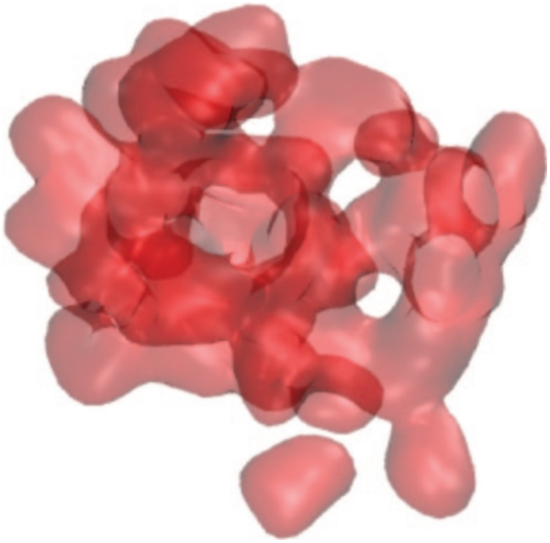


Figure 12. The core structure selected by EVM. The dark shading is the exposed surface at the CD4 binding site. The orientation of this image is identical to Figure 6 and is the same gp120.
EVM, Electrostatic Variance Masking.

remaining structure due to folding and loop lengths may. Figure 12 shows the core representation depicted in Figure 6. The darker shaded areas are exposed surfaces of the CD4 binding site.

We noted differences in variance data between Clades suggesting that some discernible variations may exist that provide additional insight into the binding process. Although these fluctuations are noted, the number of sequences selected for each clade precludes the useful comparison of the data and will need to be analyzed at a later date with a balanced set of sequences.

We conclude that BESI, in conjunction with EVM, provides a unique view of the gp120 Env and may provide additional focus on a subset of mutations for vaccine research. The process reveals differences in the outer structures of the protein and displays the power to distinguish features both visually and analytically.

Acknowledgements

We would like to thank Cynthia Derdeyn (Emory University) for providing the sequence data and Peter Hraber (Los Alamos National Laboratory) for helpful discussions about the sequence data and analyses used in the study by Boeras et al.

Author Contributions

SPM developed the methods of analysis, wrote the manuscript, performed all molecular simulations, simulation analysis and created all simulation data figures. JBP developed the alignment methods utilized for phylogeny. JLP developed the simulation protocols, reviewed manuscript, and simulation analysis.

ORCID iDs

Scott P Morton https://orcid.org/0000-0002-3777-7089

Joshua L Phillips https://orcid.org/0000-0002-4619-6083

REFERENCES

1. Stieh DJ, Phillips JL, Rogers PM, et al. Dynamic electrophoretic fingerprinting of the HIV-1 envelope glycoprotein. *Retrovirology*. 2013;10:33.
2. Boeras DI, Hraber PT, Hurlston M, et al. Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proc Natl Acad Sci U S A*. 2011;108:E1156–E1163.
3. Burton DR, Poignard P, Stanfield RL, Wilson IA. Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. *Science (New York, NY)*. 2012;337:183–186.
4. Wyatt R, Kwong PD, Desjardins E, et al. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature*. 1998;393:705–711.
5. Fischer W, Perkins S, Theiler J, et al. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat Med*. 2007;13:100–106.
6. Barouch DH, Stephenson KE, Borducchi EN, et al. Protective efficacy of a global HIV-1 mosaic vaccine against heterologous SHIV challenges in rhesus monkeys. *Cell*. 2013;155:531–539.
7. Liao HX, Bonsignori M, Alam SM, et al. Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. *Immunity*. 2013;38:176–186.
8. Mehrishi JN, Bauer J. Electrophoresis of cells and the biological relevance of surface charge. *Electrophoresis*. 2002;23:1984–1994.
9. Richmond DV, Fisher DJ. The electrophoretic mobility of micro-organisms. *Adv Microbial Physiol*. 1973;9:1–29.
10. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234:779–815.
11. Berendsen HJC, van der Spoel D, van Drunen R. Gromacs: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun*. 1995;91:43–56.
12. Lindahl E, Hess B, van der Spoel D. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model*. 2001;7:306–317.
13. Lindstrom P. Fixed-rate compressed floating-point arrays. *IEEE T Visual Comput Graph*. 2014;20:2674–2683.
14. Morton SP, Phillips JB, Phillips JL. High-throughput structural modeling of the HIV transmission bottleneck. Paper presented at: International Conference on Bioinformatics and Biomedicine; November 13-16, 2017; Kansas City, MO.
15. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A*. 2001;98:10037–10041.
16. Howton J, Phillips JL. Computational modeling of pH-dependent gp120-CD4 interactions in founder and chronic HIV strains. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York, NY: ACM Press; 2017:644–649.
17. Morton SP, Howton J, Phillips JL. Sub-class differences of PH-dependent HIV GP120-CD4 interactions. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York, NY: ACM Press; 2018:663–668.
18. Humphrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. *J Mol Graph*. 1996;14:33–38.
19. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30: 772–780.
20. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30:1312–1313.
21. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL. HIV-specific probabilistic models of protein evolution. *PLoS ONE*. 2007;2:e503.
22. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 2004;20:289–290.
23. Korber-Irrgang B, Foley BT, Kuiken C, et al. *Numbering Positions in HIV Relative to HXB2CG*. Berlin, Germany: ScienceOpen, Inc; 1998.
24. HXB2 annotated spreadsheet. 2017. HIV Sequence Database Website. https://www.hiv.lanl.gov/content/sequence/TUTORIALS/Tutorials.html.
25. Cashin K, Gray LR, Jakobsen MR, Sterjovski J, Churchill MJ, Gorry PR. CoRSeqV3-C: a novel HIV-1 subtype C specific V3 sequence based coreceptor usage prediction algorithm. *Retrovirology*. 2013;10:24.
26. Cashin K, Jakobsen MR, Sterjovski J, et al. Linkages between HIV-1 specificity for CCR5 or CXCR4 and in vitro usage of alternative coreceptors during progressive HIV-1 subtype C infection. *Retrovirology*. 2013;10:98.
27. Cashin K, Sterjovski J, Harvey KL, Ramsland PA, Churchill MJ, Gorry PR. Covariance of charged amino acids at positions 322 and 440 of HIV-1 Env contributes to coreceptor specificity of subtype B viruses, and can be used to improve the performance of V3 sequence-based coreceptor usage prediction algorithms. *PLoS ONE*. 2014;9:e109771.
28. Cashin K, Gray LR, Harvey KL, et al. Reliable genotypic tropism tests for the major HIV-1 subtypes. *Sci Rep*. 2015;5:8543.
29. Jakobsen MR, Cashin K, Roche M, et al. Longitudinal analysis of CCR5 and CXCR4 usage in a cohort of antiretroviral therapy-naïve subjects with progressive HIV-1 subtype C infection. *PLoS ONE*. 2013;8:e65950.

30. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14:1188–1190.
31. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*. 1990;18:6097–6100.
32. Derdeyn CA, Decker JM, Bibollet-Ruche F, et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science*. 2004;303: 2019–2022.
33. Korber B, Gnanakaran S. Converging on an HIV vaccine. *Science*. 2011;333:1589–1590.
34. Fahrbach KM, Malykhina O, Stieh DJ, Hope TJ. Differential binding of IgG and IgA to mucus of the female reproductive tract. *PLoS ONE*. 2013;8:e76176.

Appendix 1

Sequence data

Sequence names are taken from Boeras et al.[2]

- Z153FPB13MAR02ENV1.1
  - SLWVTVYYGVPVWKEAKATLFCASEAKAYEREVHNVWATHACVPTDPNPQEMVLENVTENFNMWKNDMVDQMHEDIISLWDQSLKPCVKLTPLCVTLNCTNAIFNNNITEEMKNCSFNITSELKDRKQKGSALFHSLDIVPLNSNSNSNYSEYRLISCNTSTITQACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGLGPCNNVSTVQCTHGIKPVVSTQLLLNGSLAEKDIVIRSENLTDNAKIIIVHLNESVEIVCIRPNNNTRKSMRIGPGQTFYATGAIIGDIRQAYCNISRKDWNTTLHKVKRKLGEHFPNTTKIKFEPSSGGDLEITTHSFNCRGEFFYCNTSELFNESFNGSDNGNITLPCRMKQIINMWQGVGRAMYAPPIAGKITCNSSITGLLLTRDGGEPGNETFRPGGGDMRDNWRSELYKYKVVEIKPLGIAPTKAKRRVVEREKR
- Z216FPL17jan0485f
  - SLWVTVYYGVPVWKEAKTTLFCASDAKAYEKEVHNVWATHACVPTDPNPQEIVLENVTESFNMWKNDMVDQMHEDIISLWDQSLQPCVKLTPLCVTLNCRDVTRNGTGNVTVDNSEGEIKNCSFNITTEIRDKKKNEYALFYKLDIVPLRNNSNEYRLINCNTSAIKQACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGTGPCNNVSTVQCTHGIKPVVSTQLLLNGSLAEEEIVIRSENLTDNTKTIIVHLNTSVEIVCTRPNNNTRKSVGIGPGQTFYATGDIIGDIRQAHCNINESNWNRTLQEVSRKLEEHFPNKAIQFQSPAGGDLEITTHSFNCRGEFFYCNTTKLFNGIYRANGTRNDTNKTLTLPCRIRQIINMWQEVGRAMYAPPIAGNIKCTSNITGILLTRDGGNTNNTEIFRPGGGNMKDNWRSELYKYKVVEIKPLGIAPTKAKRRVVEREKR
- Z216FPB98_plasmid_e
  - SLWVTVYYGVPVWKEAKTTLFCASDAKAYEKEVHNVWATHACVPTDPNPQEIVLENVTESFNMWKNDMVDQMHEDIISLWDQSLQPCVKLTPLCVTLNCSAVRNATDTNYNVTAKEEMKNCSFNITTEIRDKKKNEYALFYKLDIVPLNNNNSNAGNFSEYRLINCNTSAIKQACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGTGPCNNVSTVQCTHGIKPVVSTQLLLNGSLAEEEIVIRSENLTDNAKTIIVHLNESVRIECARPGNNTRKSVRIGPGQTFYATGDIVGDIRQAHCNISERDWNKTLQAVRKKLEKHFPNKTIQFKPPPPGGDLEITTHSFNCGGEFFYCNTSQLFNGTYNGTYMTNEAEGNANKTLTLPCRIRQIINMWQEVGRAMYAPPIAGNITCISNITGLLLTRDGGNTNDTNKTETFRPGGGNMKDNWRSELYKYKVVEIKPLGIAPTKAKRRVVEREKR
- Z242MPL25JAN03PCR23ENV1.1-_Donor_Transmitted
  - NLWVTVYYGVPVWKEAKATLFCASDAKAYDREVHNVWATHACVPTDPNPQELLLENVTENFNMWKNDMVDQMHEDVISLWDQSLKPCVKLTPLCVTLNCVNLIRNDTKNGTVMLDAKNCSFNATTEIKDRKRKEYALFYRLDIVPLESENSTNSSTKYRLINCNTSTVTQACPKVSFDPIPIHYCAPAGYAILKCNDETFNGTGPCNNVSTVQCTHGIKPVVSTQLLLNGSLTKEIIISSENITNNAKTIIVHLNESVAINCTRPSNNTRKSVRIGPGQAFYATNDIIGDIRQAHCNISRSQWNKTLERVKEKLEKQFHRNISFSSSSGGDLEITTHSFNCRGEFFYCNTTKLFLPNSNETENSTIILPCRIRQIINMWQEVGRAMYAPPIAGSIECKSNITGILLVRDGGINTTTEIFRPEGGNMKDNWRSELYKYKVVEIKPLGIAPTEAKRRVVEREKR
- Z238FSW29oct0215A6v
  - NLWVTVYYGVPVWKEAKTTLFCASDAKAYEKEVHNVWATHACVPTDPDPQEIVLGNVTENFNMWKNDMVDQMHEDVISLWDQSLKPCVKLTPLCVTLNCSNAKVNVTGNNTIDMQEEIKNCSFNATTEIQDKTKKVYALFYRADVVQLGSNKSEYRLINCNTSAITQACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGTGPCQKVSTVQCTHGIKPVVSTQLLLNGSPAEEEIIIRSKNLSDNTKTIIVHLNESVRIVCTRPGNNTRKSIRIGPGQTFYATGDIIGDIREAHCDVNATQWNKTLHQVQGKLREHFPNKTIEFKLPSGGDLEITMHSFNCRGEFFYCNTSGLFNRTYYPNGTEGANITRQNLPENITLPCRIKQIINMWQEVGRAMYAPPIAGNITCVSNITGLLLIRDGGGGTEASNETREIFRPGGGDMRDNWRSELYKYKVVEVQPLGVAPTKAKRRVVEREKR
- Z185MPB17AUG02ENVB17
  - ESWVTVYYGVPVWKEAKAPLFCASDAKAYEREVHNIWATHACVPTDPDPQEMVLKNV

    TENFNMWKNDMVDQMNEDIISLWDQS
    LKPCVKLTPLCVTLNCSNYNSTANSTGK
    STGSPNEEIKNCSFYTTTELRDKRKNESA
    LFNSLDIVKLDNNGSSYRLINCNTSTITQACP
    KVSFDPIPIHYCAPAGYAILKCNNKTFNG
    TGACNNVSTVQCTHGIKPVVSTQLLLNG
    SLAEEEIIIRSENLTNNAKTIIVQFTTPVG
    IVCVRPNNNTRKSVRIGPGQTFYATGDII
    GDIRQAHCNISEKTWNDTLQKVGKKLQE
    KFPNRTIEFARSSGGDPEITTHSFNCR
    GEFFYCNTSKLFNSTYMANSTNSTSN
    DTITLQCRIKQIINLWQKVGRAMYA
    PPIAGNITCKSNITGLLLTHDGGSNGTL
    IFRPGGGDMRDNWRSELYKYKVVE
    IRPLGVAPTKAKRRVVEREKR

- R56MCF21aug0511_plasmid_1v
    - NLWVNVYYGVPVWKDAETTLFCASD
      AKAYETEVHNVWATHACVPTDPNPQ
      EIHLENVTEEFNMWENNMVEQMHTD
      IISLWDQSLKPCVKLTPLCVTLKCSEAYN
      STVDSEVKGEIQNCSFNVTTEIRDKNQK
      VHALFYRPDIVPLSKGNGSEYRLIN
      CNTSAITQACPKVSFDPIPIHYCAPAGYA
      ILKCNNKTFNGTGPCNNVSTVQCTH
      GIKPVVSTQLLLNGSLAEKEIIIRSKNIT
      NNVNTIIVQLNSSVEINCTRPSNNT
      RKSIRIGPGQTFYATGDIIGDIRQAHCN
      LSRNLWNKTLSQIRNKLSKYFPNRTITF
      NTSSGGDLEITTHSFNCGGEFFYCNTSDL
      FNTNLVNDTDITNSTLTLPCKIKQIVRM
      WQGVGQAMYAPPIAGNITCRSKITG
      LLLVRDGGDTTDTDTETFRPGGG
      DMRDNWRSELYKYKVVKIEPIGVAP
      HRAKRRVVEREKR
- R56MPL21apr05C5_plasmid_6-4
    - NLWVNVYYGVPVWKDAKTTLFCASDA
      KAYDTEVHNVWATHACVPTDPNPQ
      EIHLENVTEEFNMWENNMVEQMHT
      DIISLWDQSLKPCVKLTPLCVTLNCSE
      FDNSTSPNTTVDSGMKGEIQNCSFNVT
      TEIRDKNQKVYALFYRPDIVPLSTGNG
      NEYRLINCNTSAITQACPKVSFDPIPIH
      YCAPAGYAILKCNNKTFNGTGPCNNVSTV
      QCTHGIKPVVSTQLLLNGSLAEKEIIIRSE
      NISDNVKTIIVQLNNSVEINCTRPG
      NNTRQSIRIGPGQTFYATGDIIGDIRQAH
      CNVSRNLWNKTLSQIRNKLSTYFLNKTI
      NFNTSSGGDLEITTHSFNCGGEFFYCNTSG
      LFNLNNTNITHITLPCRIKQIVRMWQEV
      GQAMYAPPIAGNITCRSNITGLLLVR
      DGGGTTNGSETFRPGGGNMKDNWRSE
      LYKYKVVKIEPIGIAPHRAKRRVVEREKR
- Z185MPB17AUG02ENV1.2
    - ESWVTVYYGVPVWKEAKAPLFCASDAKA
      YEREVHNIWATHACVPTDPDPQEMVLKN
      VTENFNMWKNDMVDQMNEDIISLWD
      QSLKPCVKLTPLCVTLNCSNYNSTANST
      GKNTGSPNEEIKNCSFYTTTELRDKRK
      NESALFNSLDIVSLDNNGSSYRLINCNT
      STITQACPKVSFDPIPIHYCAPAGYAI
      LKCNNKTFNGTGPCNNVSTVQCTH
      GIKPVVSTQLLLNGSLAEEEIIIRSENLT
      NNAKTIIVQFTTPVDIVCVRPNNNTRK
      SVRIGPGQTFYATGDIIGDIRQAHCNISE
      KTWNDTLQKVGEKLQEKFPNKT
      IVFARSSGGDLEITTHSFNCRGEFFYC
      NTSKLFNSTYMANSTNTNSTSNDTIT
      LQCRIKQIINLWQKVGRAMYAPPIAGN
      ITCKSNITGLLLTHDGTNPNNNNTLIF
      RPGGGDMRDNWRSELYKYKVVEIRPL
      GVAPTKAKRRVVEREKR
- Z201FPL7FEB03ENV2.1
    - NLWVTVYYGVPVWKEAKTTLFCASDAK
      AFENEVHNVWATHACVPTNPNPQELVL
      ENVTENFNMWENDMVEQMHEDIISLWD
      QSLKPCVKLTPLCVTLTCKNFTSKDANNVT
      VNNTQEIKNCSFNITTELRDKKKQESALF
      YRVDIVPLEESSGKNRSMNNSEYEEYRL
      INCNTSTITQACPKVTFDPIPIHYCVPAGY
      AILKCNNKTFNGSGPCNNVSTVQCTHGIKP
      VVSTQLLLNGSLAEEDIIIRSKNITDPSRTI
      IVHLKKAVEIACIRPGNNTRKSIRIGPGQT
      FYATGAIIGNIREAHCNISEKQWNETLYNV
      SKKLEGHFPNSIIKFESSSGGDLEIEMHS
      FNCRGEFFYCNTSQLFNSTYMPNSTR
      STGNASKIITLPCRIKQIVNMWQGVGQAM
      YAPPIAGNITCNSSITGLLLTRDGRKNNTEIF
      RPIGGDMKDNWRSELYKYKVVEIKPLGL
      APTKAKRRVVEREKR
- Z201FCF07feb03DNA13C18
    - NLWVTVYYGVPVWKEAKTTLFCASDAK
      AFDSEVHNVWATHACVPTDPNPQELVL
      ENVTENFNMWENDMVEQMHEDIISL
      WDQSLKPCVKLTPLCVTLTCKNFTSKDANN
      VTVNNTQEIKNCLFNITTELRDKKKQES
      ALFYRVDIVPLEESSGKNRSMNNSEYEE
      YRLINCNTSTITQACPKVTFDPIPIHYCVPA
      GYAILKCNNKTFNGSGPCNNVSTVQCTH
      GIKPVVSTQLLLNGSLAEEDIIIRSKNITDT
      FRTIIVHLKKAVEIACIRPGNNTR
      KSIRIGPGQTFYATGAIIGNIREAHCN
      ISEKLWNETLYNVSKKLEGHFPNSTIEF
      KPSSGGDLEIEMHSFNCRGEFFYCNTSQL
      FNSTYMPNSTRSTGNASKIITLPCRIKQIV

NMWQGVGQAMYAPPIAGNITCNSSITG
LLLTRDGRKNNTEIFRPIGGDMKDNWRS
ELYKYKVVEIKPLGLAPTKAKRRVVEREKR

- Z292FCF24may0512E26_plasmid_10iv
  - NLWVTVYYGVPVWREADTILFCATDAKTY
    DPEGHNVWATHACVPTDPNPQEIDLV
    NVTEDFNMWKNGMVEQMNTDITSLWD
    QSLKPCVSLTPLCVTLNCTSNITISNNTT
    TSNETVEDSIIKEMKNCSYNMTTELRDRRQ
    KVYSLFYKLDIVPIRENSSNEYRLINCNT
    SVVKQACPKTAFEPIPIHYCAPAGFAILKC
    KNKQFSGTGPCENVSSVQCTHGIKPVVS
    TQLLLNGSLAEEEIMIRSENFTDNAKT
    IIVQFVDPVEINCTRPGNNRRRSVHIGP
    GQAFYATGEVIGDIRKAHCNVSRTKWE
    NNLQKVAKKLRGKFKNGTTIIFANHSGGD
    LEITTHSFNCGGEFFYCNTSGLFNSTWNN
    DTESNSTQESNSTITLPCRIKQIVNMWQRV
    GQAIYAPPIEGVIRCESNITGLLLTRDGGG
    NNRTNETFRPEGGNMKDNWRSELYKYK
    VVKIEPLGVAPTPARRRVVMREKR
- Z205MPB27MAR03ENV6.1
  - NLWVTVYYGVPVWKEAKTTLFCASD
    AKAYEREVHNVWATHACVPTDPNP
    QEMELKNVTENFNMWKNDMVDQ
    MHEDIISLWDQSLKPCVKLTPLCVT
    LNCSNVTNYSNSSATNDSNYNATYV
    DEIKNCSFNATTEIRDKKRKEYALF
    YRPDIVPLNPNDGNSREYILINCNTS
    TIAQACPKVSFDPIPIHYCAPAGYAI
    LKCNDKNFNGTGPCDNVSTVQCT
    HGIKPVISTQLLLNGSLAEENIIIRS
    ENLANNVKTIIVHLNESVEINCTRP
    NNNTRKGIRIGPGQMFYAAGEIIGD
    IRRAHCNVNESKWNDTYQKIKKKLQ
    EHFPNKTIHFEPPAGGDLEITTHSFNCRG
    EFFYCNTSELFNSTRLTGQQNLSAIITLP
    CRIKQFINMWQGVGRAMYAPPIEGKITC
    NSSITGLLLTRDGGNVTSDNETFRLGG
    GDMRDNWRSELYKYKVVEIKPLGIAPT
    ESKRAVVEREKR
- Z238FCF29oct0215A39
  - NMWVTVYYGVPVWKEAKTTLFCASDAK
    AYEKEVHNVWATHACVPTDPNPQE
    IVLGNVTENFNMWKNDMVDQMHEDVIS
    LWDQSLKPCVKLTPLCVTLNCSNANV
    TEASNNILNMTEEIRNCSFNATTEI
    QDKTKKVYALFYKLDVVQLGSNTS
    EYRLINCNTSAITQACPKVSFDPIPI
    HYCAPAGYAILKGNNKTFKGTGPCQNV
    STVQCTHGIKPVVSTQLLLNGSLAEE
    GIIIRSENLTDNVKTIIVHLNESV
    PIVCTRPGNNTRKSIRIGPGQTFYAT
    GDIIGDIREAHCNINATQWNKTLQQVKG
    KLKEHFPDKTIKFESPSGGDLEITMHSFN
    CRGEFFYCNTSRLFNETYIEHHNATA
    NITLPCRIKQIINMWQEVGRAMYAPPV
    AGYITCNSSITGLLLLRDGGTSDNGTND
    ETFRPGGGDMRDNWRSELYKYKVV
    EVKPLGIAPTKAKRRVVEREKR
- Z292FCF24may0512D18_plasmid_4i
  - DLWVTVYYGVPVWREADTILFCASDAK
    TYNPEGHNVWATHACVPTDPNPQE
    IDLVNVTEDFNMWKNGMVEQMHTD
    IISLWDQSLKPCVSLTPLCVTLNCTSN
    ITISNNTTTSNETVEDSIIKEMKNCSY
    NMTTEVRDRRQKVYSLFYKLDMVPIRE
    DDNSSNEYRLINCNTSVVKQACPKIA
    FEPIPIHYCAPAGFAILKCKNKQFNG
    TGPCENVSSVQCTHGIKPVVSTQLLLNGS
    LAEEEVMIRSENFTNNAKTIIVQFVDP
    VKINCTRPGNNRRRSVHIGPGQAFYATGE
    VIGDIRKAHCNVSRTEWENTLQKVAK
    KLREKFKNGTTIIFANHSGGDLEITTH
    SFNCGGEFFYCNTSGLFNSTWNGTES
    NSTQELNSNITLPCRIKQIVNMWQRVGQ
    AIYAPPIEGVISCKSNITGLLLTRDGGGN
    NRTNETFRPEGGNMKDNWRSELYKY
    KVVKIEPLGIAPTPARRRVVMREKR
- Z205MPB27MAR03ENV9.1
  - NLWVTVYYGVPVWKEAKTTLFCASD
    AKAYEREVHNVWATHACVPTDPNPQEM
    FLKNVTEDFNMWKNDMVDQMHED
    IISLWDQSLKPCVKLTPLCVTLSCSNYSNCN
    DTMNSNHSTANCTSGGEIKNCSFNAT
    TEIRDKNRKEYALFYRPDIVPLKPNDSNSR
    EYILINCNTSTIAQACPKVSFDPIPIHYCAPA
    GYAILKCNDNKTFNGTGPCYNVSTVQCT
    HGIKPVISTQLLLNGSLAEEDIIIRSENLAN
    NVKTIIVHLNKSVEINCTRPNNNTSR
    GIRIGPGQTFFATGRIIGDIRQAYCSI
    NASKWNDTLQKIKRKLQEHFPNKTIQF
    APPAGGDLEITTHSFNCRGEFFYCNTSELF
    NISRLNSTSSIITLPCRIKQFINMWQKVGR
    AMYAPPIEGKITCNSSITGLLLTRDGGNN
    TNGTETFRPGGGDMRDNWRSELYKYKVVE
    IKPLGIAPTGSKRAVVEREKR
- Z221FPL55_plasmid_6-2
  - SLWVTVYYGVPVWKEAKTTLFCASDAKAY
    EKEMHNVWATHACVPTDPNPQELVLE
    NVTENFNMWKNDMVDQMHEDIISLWDQ
    SLKPCVKLTPLCVTLNCTNANITNNG
    TNHHNNGNGNTYNDTMAKEMKNCSFNV
    TTEIRDRQKNVYALFYKLDIVPIDNESKH
    NNSNESKHSNYSDYRLINCNTSAMTQAC
    PKVSFTPIPIHYCAPAGYAILKCNNKTFNG

TGPCHNVSTVQCTHGIKPVVSTQLLLNGSLAEPEIIIRSKNLTDNTKTIIVHLNQSVEIVCTRPGNNTRKSIRIGPGQTFYANDIIGDIRQAYCNISKRDWNNTLHWVSKKLREHFPNKPIKFENSSGGDIEITHHSFNCGGEFFYCNTSQLFNSTYMANSTYTENNSTKNITLPCRIKQIINMWQEVGRAMYAPPIAGNITCKSNITGLLLVRDGGGEINDTNGTETFRPGGGDMRDNWRSELYKYKVVEIKPLGIAPTKAKRRVVEREKR

- Z242MPL26_plasmid
  - NLWVTVYYGVPVWKEAKATLFCASDAKAYDREVHNVWATHACVPTDPNPQELLLENVTENFNMWKNDMVDQMHEDVISLWDQSLKPCVKLTPLCVTLNCVNLIRNDTKNGTVMLDAKNCSFNATTEIKDKKKKEYALFYRLDIVPLESENSTNSSTKYRLINCNTSTVTQACPKASFDPIPIHYCAPAGYAILKCNDETFNGTGPCSKVSTVQCTHGIKPVVSTQLLLNGSLTKEIIISSENITNNAKTIIVHLNESVAINCTRPSNNTRKSVRIGPGQAFYATNDIIGDIRQAHCNISRSQWNKTLERVKEKLEKQFHRNISFSSSSGGDLEITTHSFNCRGEFFYCNTTKLFLPNSNETENSTIILPCRIRQIINMWQEVGRAMYAPPIAGSIECKSNITGLLLVRDGGINTTTEIFRPEGGNMKDNWRSELYKYKVVEIKPLGIAPTEAKRRVVEREKR
- Z221FPL7MAR03ENV2.3
  - SLWVTVYYGVPVWKEAKTTLFCASDAKAYEKEMHNVWATHACVPTDPNPQEIVLGNVTENFNMWKNDMVDQMHEDIISLWDQSLKPCVKLTPLCVTLNCTNVNITSDGTTHNDISNGATYNDTTEMKNCSFNITTEVRDKKKNVYALFYELDIVPISNENTHIGYRLINCNTSAMTQACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGTGPCHNVSTVQCTHGIKPVVSTQLLLNGSLAEEEIIIRSKNLTDNTKTIIVHLNQSIEIVCTRPNNNTRKSIRIGPGQTFYATDGIIGNIRQAHCNVSTGNWSNTLQWVSEKLREHFPGKNIKFEPSSGGDLEITHHSFNCGGEFFYCDTSQLFNKTYPANSTDIRNGSNTPITLPCRIKQIINMWQEIGRAMYAPPIAGNITCKSNITGLLLVRDGGINGTNHTETFRPGGGDMRDNWRSELYKYKVVEIKPLGIAPTKAKRRMVEREKR
- Z153FPL13MAR02ENV6.1
  - SLWVTVYYGVPVWKEAKATLFCASDAKAYEREVHNVWATHACVPTDPNPQEMVLENVTENFNMWKNDMVDQMHEDIISLWDQSLKPCVKLTPLCVTLNCTNAIFNNNITEEMKNCSFNITSELKDRKQKESALFHSLDIVPLNNNSSNNYSEYRLISCNTSTITQACPKVSFDPIPIHYCAPAGYAILKCNNKTFNGSGPCNNVSTVQCTHGIKPVVSTQLLLNGSLAEKDIVIRSENLTDNAKIIIVHLNKSVEIKCIRPNNNTRKSVRIGPGQTFYATGAIIGDIRQAYCNISRKDWNTTLHEVKRKVREHFNATIKFEPSSGGDLEITTHSFNCRGEFFYCNTSKLFNESFNGSDNGNITLPCRIKQIINMWQGVGRAMYAPPIAGKITCNSSITGLLLTRDGGNRGNETNKTETFRPGGGDMRDNWRSELYKYKVVEIKPLGVAPTKAKRRVVEREKR

Variable loop comparison

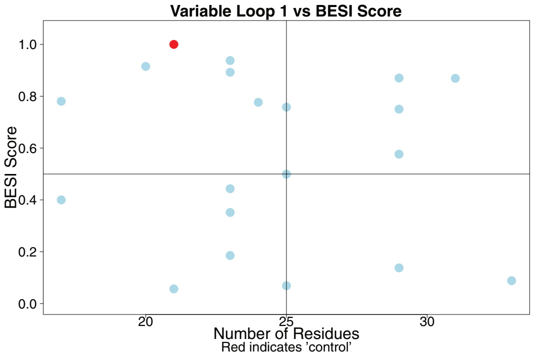


Figure A1. Scatter plot of variable loop V1 in comparison with BESI scores. Red indicates the control.
BESI, Biomolecular Electro-Static Indexing.

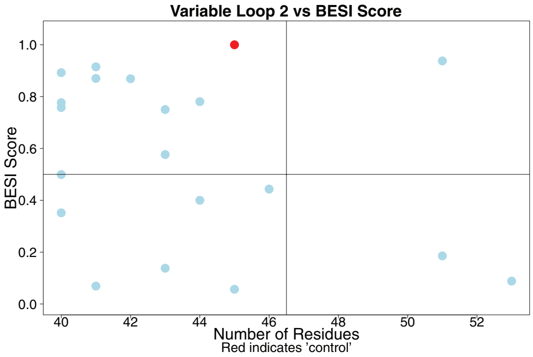


Figure A2. Scatter plot of variable loop V2 in comparison with BESI scores. Red indicates the control.
BESI, Biomolecular Electro-Static Indexing.

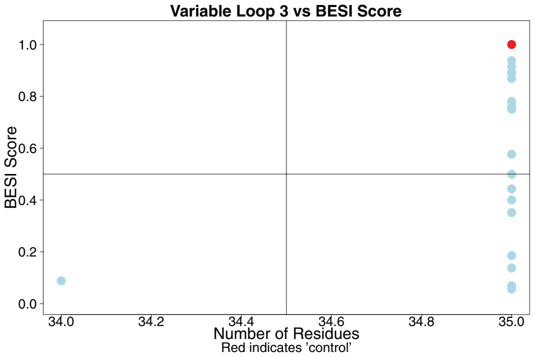


Figure A3. Scatter plot of variable loop V3 in comparison with BESI scores. Red indicates the control.
BESI, Biomolecular Electro-Static Indexing.

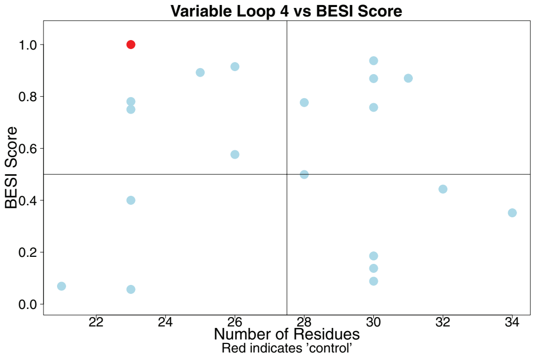


Figure A4. Scatter plot of variable loop V4 in comparison with BESI scores. Red indicates the control.
BESI, Biomolecular Electro-Static Indexing.

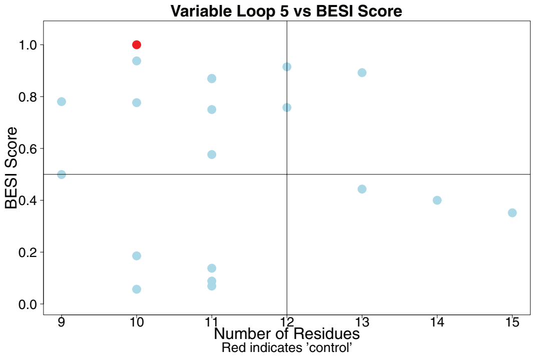
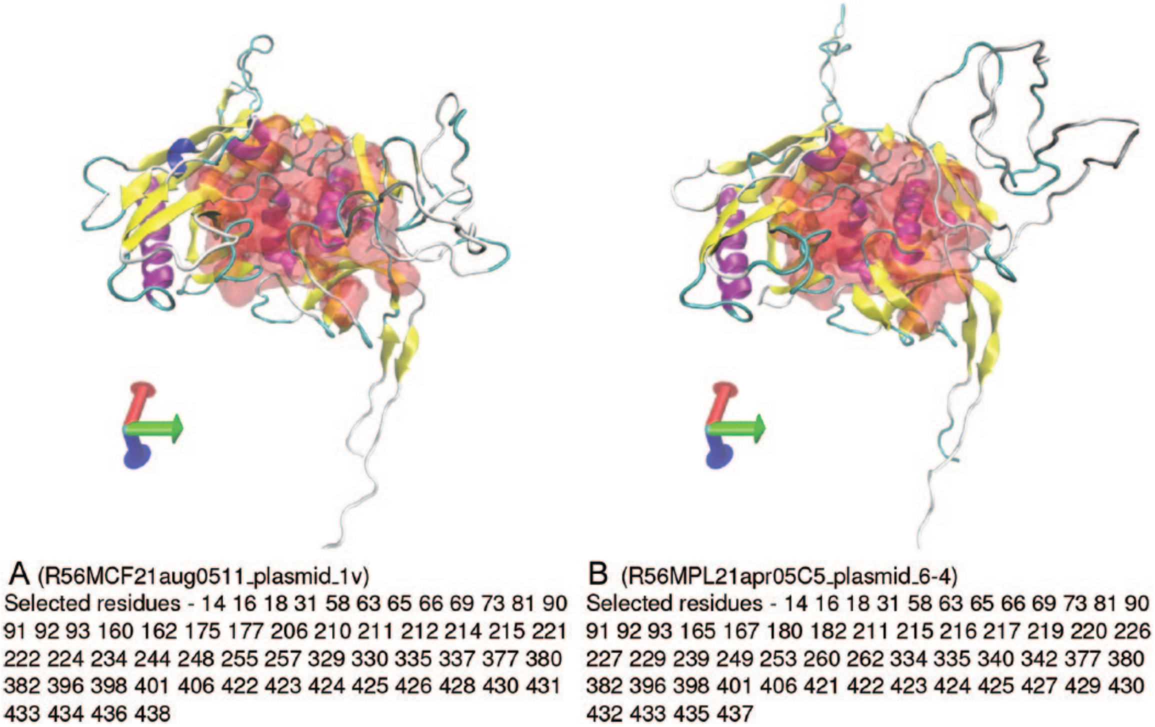


Figure A5. Scatter plot of variable loop V5 in comparison with BESI scores. Red indicates the control.
BESI, Biomolecular Electro-Static Indexing.

EVM imagery

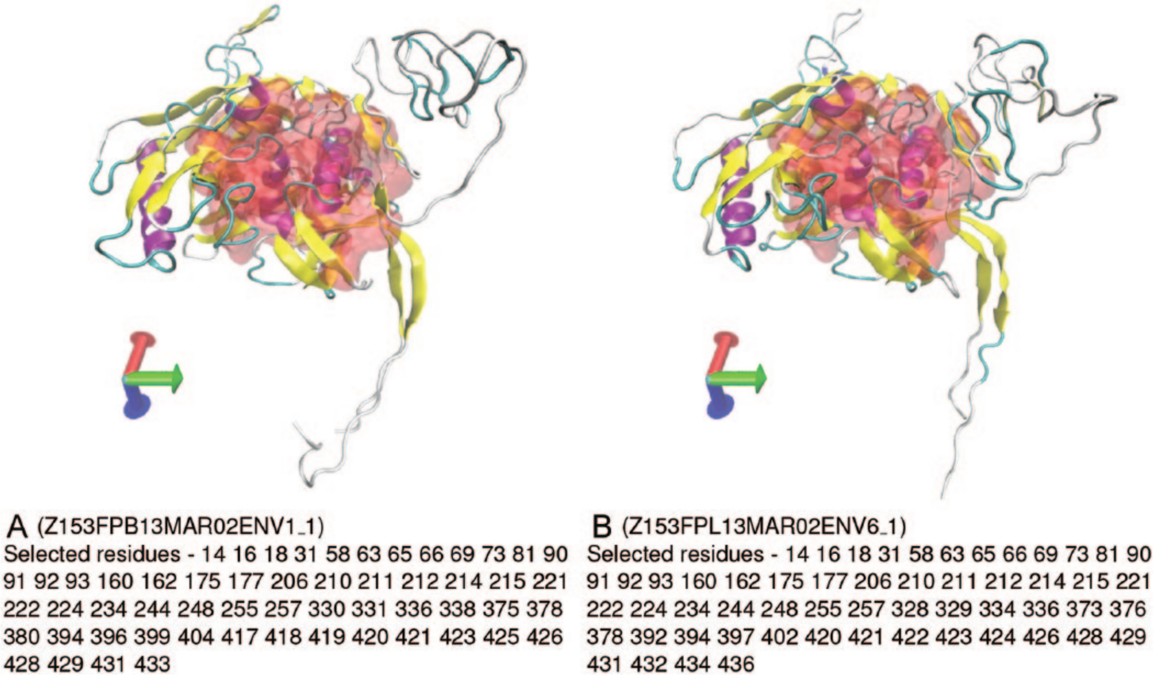
The figures display selected residues mapped to each sequence that are directly consumable in VMD and the EVM imagery associated with each assembly.



A (R56MCF21aug0511_plasmid_1v)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 160 162 175 177 206 210 211 212 214 215 221 222 224 234 244 248 255 257 329 330 335 337 377 380 382 396 398 401 406 422 423 424 425 426 428 430 431 433 434 436 438

B (R56MPL21apr05C5_plasmid_6-4)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 165 167 180 182 211 215 216 217 219 220 226 227 229 239 249 253 260 262 334 335 340 342 377 380 382 396 398 401 406 421 422 423 424 425 427 429 430 432 433 435 437

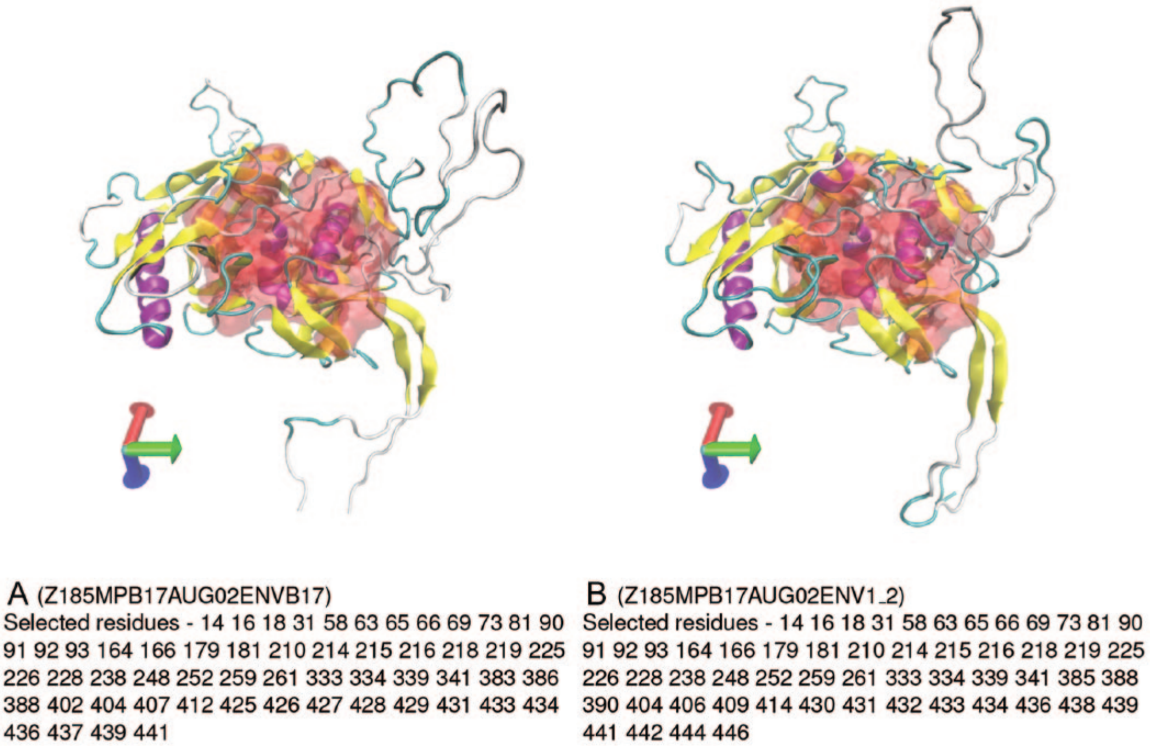
Figure A6. EVM imagery for donor R56M: (A) BESI score = 0.914 and (B) BESI score 0.069.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z153FPB13MAR02ENV1_1)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 160 162 175 177 206 210 211 212 214 215 221 222 224 234 244 248 255 257 330 331 336 338 375 378 380 394 396 399 404 417 418 419 420 421 423 425 426 428 429 431 433

B (Z153FPL13MAR02ENV6_1)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 160 162 175 177 206 210 211 212 214 215 221 222 224 234 244 248 255 257 328 329 334 336 373 376 378 392 394 397 402 420 421 422 423 424 426 428 429 431 432 434 436

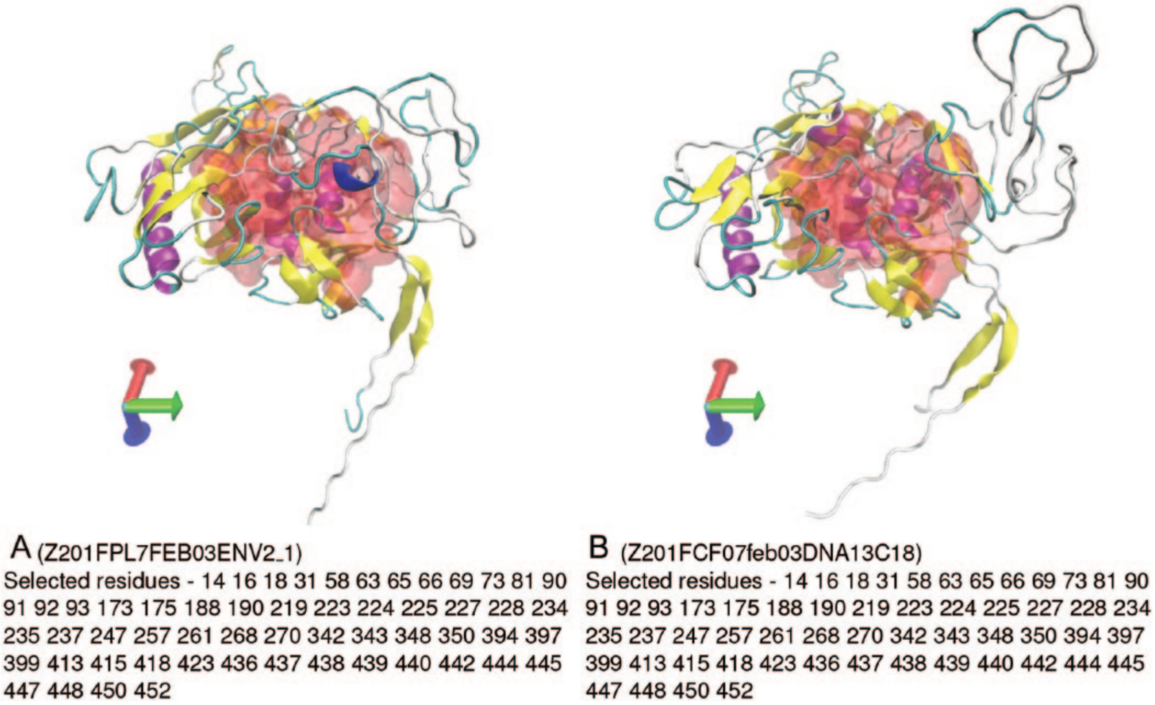
Figure A7. EVM imagery for donor Z153F: (A) BESI score = 0.781 and (B) BESI score = 0.400.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z185MPB17AUG02ENVB17)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 164 166 179 181 210 214 215 216 218 219 225 226 228 238 248 252 259 261 333 334 339 341 383 386 388 402 404 407 412 425 426 427 428 429 431 433 434 436 437 439 441

B (Z185MPB17AUG02ENV1_2)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 164 166 179 181 210 214 215 216 218 219 225 226 228 238 248 252 259 261 333 334 339 341 385 388 390 404 406 409 414 430 431 432 433 434 436 438 439 441 442 444 446

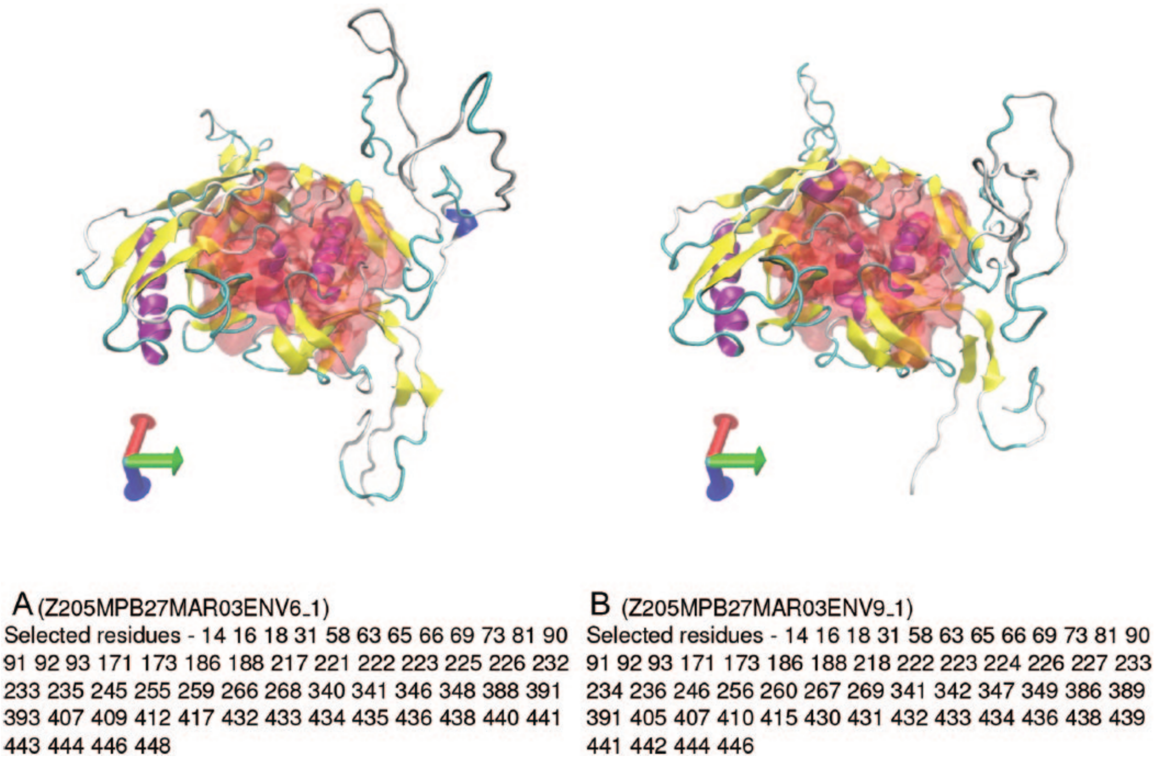
Figure A8. EVM imagery for donor Z185M: (A) BESI score = 0.758 and (B) BESI score = 0.499.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z201FPL7FEB03ENV2_1)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 173 175 188 190 219 223 224 225 227 228 234 235 237 247 257 261 268 270 342 343 348 350 394 397 399 413 415 418 423 436 437 438 439 440 442 444 445 447 448 450 452

B (Z201FCF07feb03DNA13C18)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 173 175 188 190 219 223 224 225 227 228 234 235 237 247 257 261 268 270 342 343 348 350 394 397 399 413 415 418 423 436 437 438 439 440 442 444 445 447 448 450 452

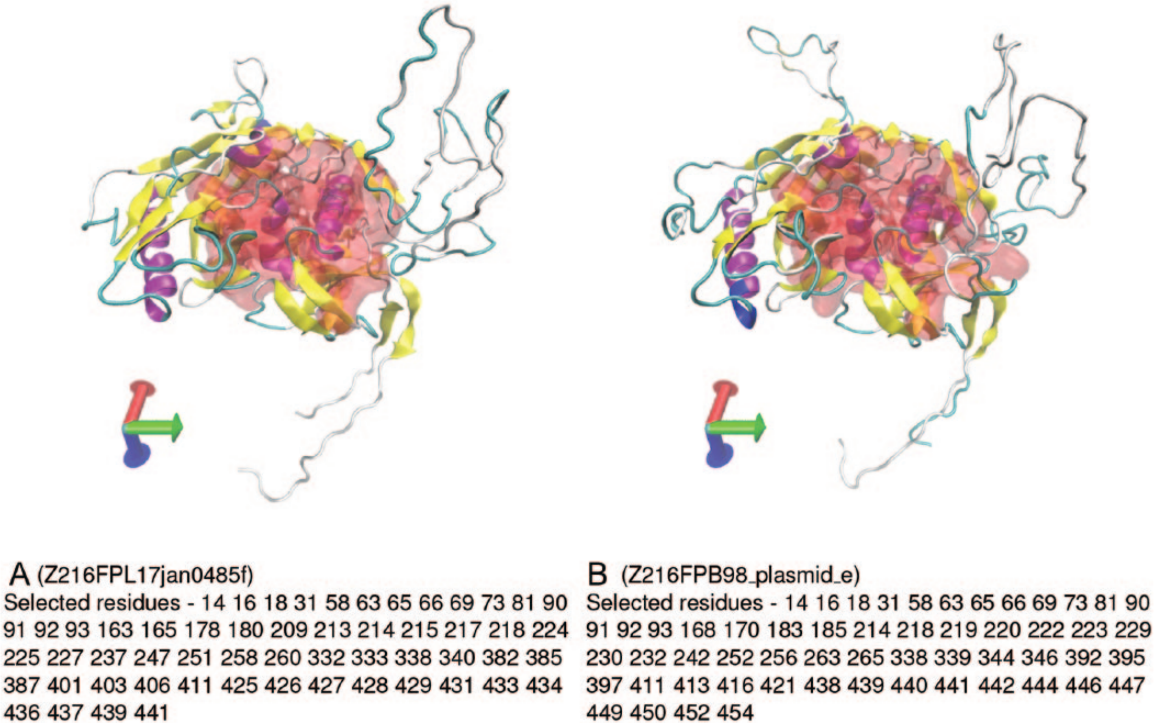
Figure A9. EVM imagery for donor Z201F: (A) BESI score = 0.938 and (B) BESI score = 0.186.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z205MPB27MAR03ENV6_1)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 171 173 186 188 217 221 222 223 225 226 232 233 235 245 255 259 266 268 340 341 346 348 388 391 393 407 409 412 417 432 433 434 435 436 438 440 441 443 444 446 448

B (Z205MPB27MAR03ENV9_1)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 171 173 186 188 218 222 223 224 226 227 233 234 236 246 256 260 267 269 341 342 347 349 386 389 391 405 407 410 415 430 431 432 433 434 436 438 439 441 442 444 446

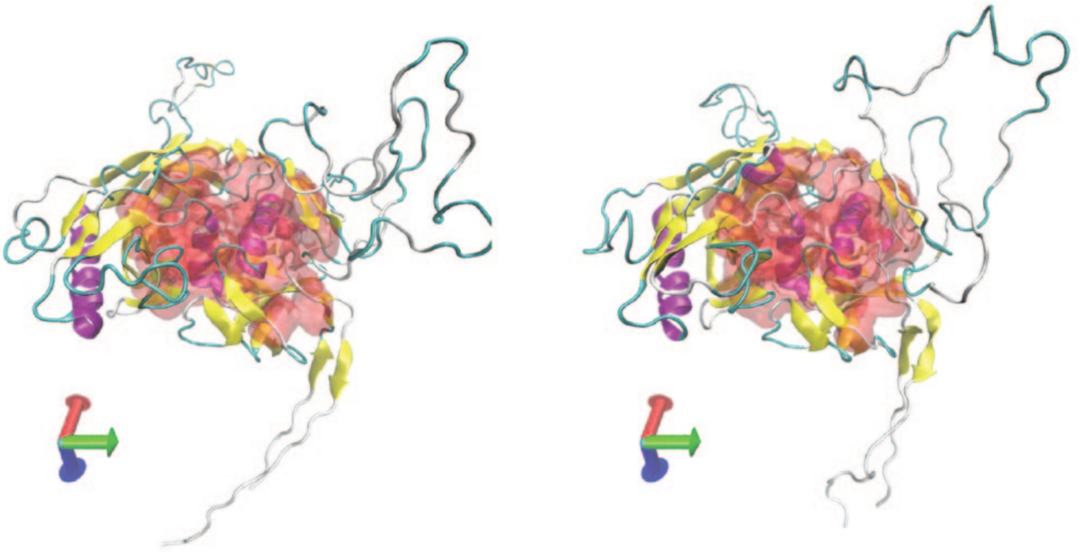
Figure A10. EVM imagery for donor Z205M: (A) BESI score = 0.750 and (B) BESI score = 0.576.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z216FPL17jan0485f)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 163 165 178 180 209 213 214 215 217 218 224 225 227 237 247 251 258 260 332 333 338 340 382 385 387 401 403 406 411 425 426 427 428 429 431 433 434 436 437 439 441

B (Z216FPB98_plasmid_e)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 168 170 183 185 214 218 219 220 222 223 229 230 232 242 252 256 263 265 338 339 344 346 392 395 397 411 413 416 421 438 439 440 441 442 444 446 447 449 450 452 454

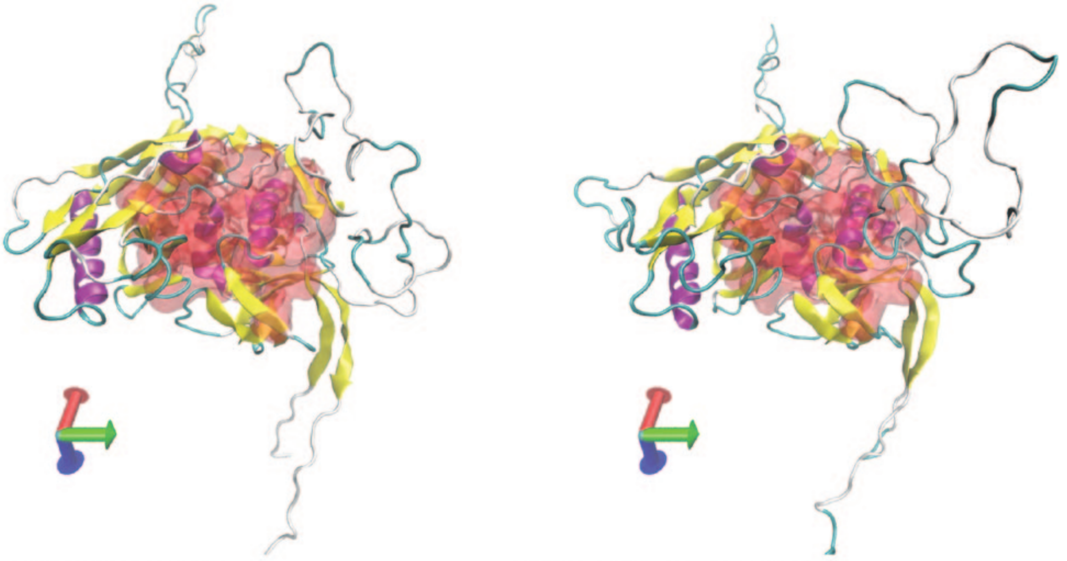
Figure A11. EVM imagery for donor Z216F: (A) BESI score = 0.777 and (B) BESI score = 0.443.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z221FPL55_plasmid_6-2)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 185 187 200 202 231 235 236 237 239 240 246 247 249 259 269 273 280 282 353 354 359 361 405 408 410 424 426 429 434 452 453 454 455 456 458 460 461 463 464 466 468

B (Z221FPL7MAR03ENV2_3)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 172 174 187 189 218 222 223 224 226 227 233 234 236 246 256 260 267 269 341 342 347 349 393 396 398 412 414 417 422 438 439 440 441 442 444 446 447 449 450 452 454

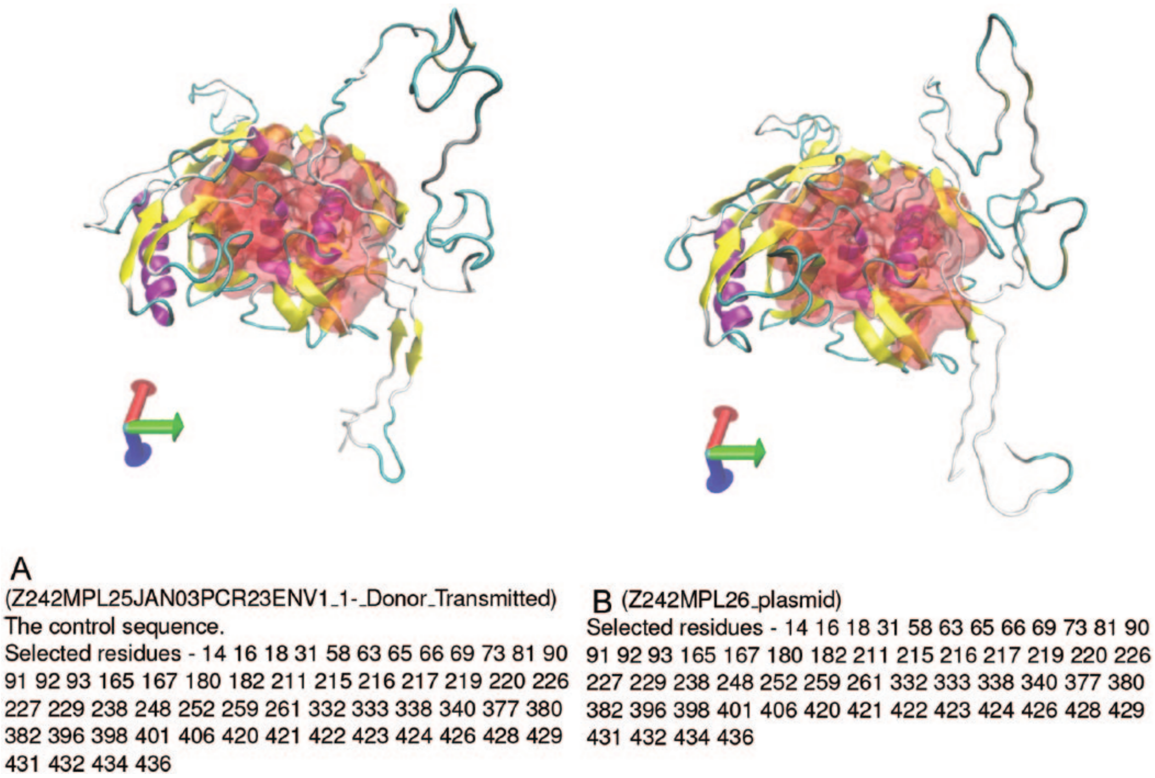
Figure A12. EVM imagery for donor Z221F: (A) BESI score = 0.869 and (B) BESI score = 0.088.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z238FCF29oct0215A39)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 162 164 177 179 208 212 213 214 216 217 223 224 226 236 246 250 257 259 331 332 337 339 378 381 383 397 399 402 407 424 425 426 427 428 430 432 433 435 436 438 440

B (Z238FSW29oct0215A6v)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 162 164 177 179 208 212 213 214 216 217 223 224 226 236 246 250 257 259 331 332 337 339 387 390 392 406 408 411 416 435 436 437 438 439 441 443 444 446 447 449 451

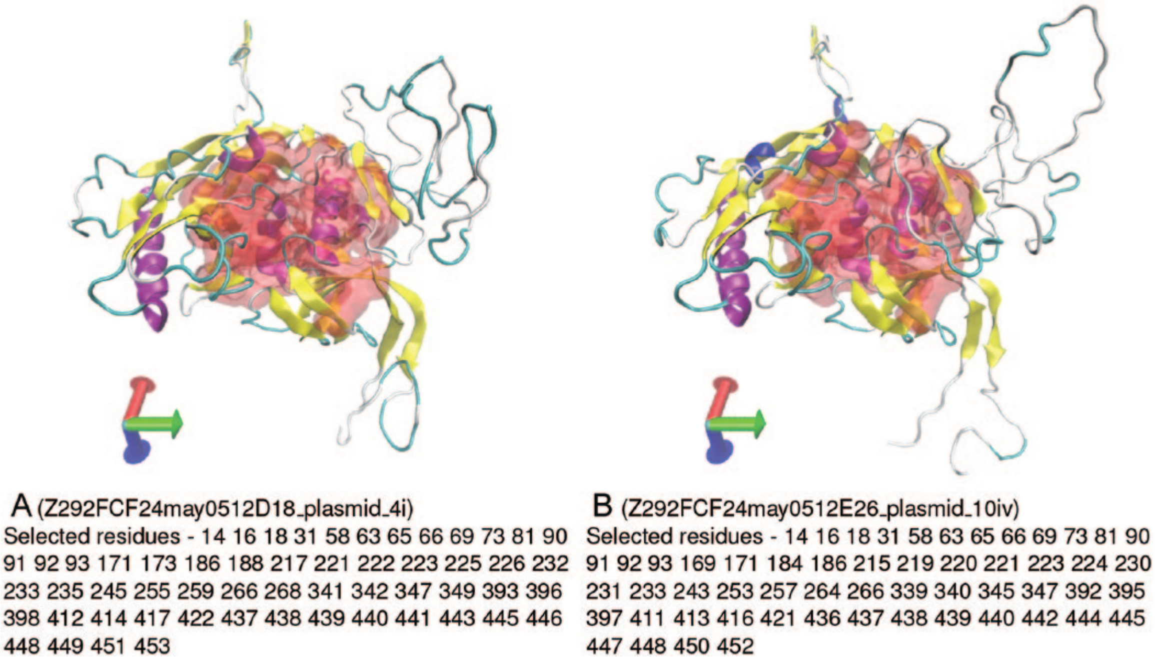
Figure A13. EVM imagery for donor Z238F: (A) BESI score = 0.892 and (B) BESI score = 0.352.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z242MPL25JAN03PCR23ENV1_1-_Donor_Transmitted)
The control sequence.
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 165 167 180 182 211 215 216 217 219 220 226 227 229 238 248 252 259 261 332 333 338 340 377 380 382 396 398 401 406 420 421 422 423 424 426 428 429 431 432 434 436

B (Z242MPL26_plasmid)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 165 167 180 182 211 215 216 217 219 220 226 227 229 238 248 252 259 261 332 333 338 340 377 380 382 396 398 401 406 420 421 422 423 424 426 428 429 431 432 434 436

Figure A14. EVM imagery for donor Z242M: (A) BESI score = 1.000 and (B) BESI score = 0.057.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.



A (Z292FCF24may0512D18_plasmid_4i)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 171 173 186 188 217 221 222 223 225 226 232 233 235 245 255 259 266 268 341 342 347 349 393 396 398 412 414 417 422 437 438 439 440 441 443 445 446 448 449 451 453

B (Z292FCF24may0512E26_plasmid_10iv)
Selected residues - 14 16 18 31 58 63 65 66 69 73 81 90 91 92 93 169 171 184 186 215 219 220 221 223 224 230 231 233 243 253 257 264 266 339 340 345 347 392 395 397 411 413 416 421 436 437 438 439 440 442 444 445 447 448 450 452

Figure A15. EVM imagery for donor Z292F: (A) BESI score = 0.870 and (B) BESI score = 0.138.
BESI, Biomolecular Electro-Static Indexing; EVM, Electrostatic Variance Masking.