AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 23, Number 10, 2007, pp. 1283–1292

DOI: 10.1089/aid.2007.0098

HIV Type 1 Vaccines for Worldwide Use: Predicting In-Clade and Cross-Clade Breadth of Immune Responses

ADAM C. FINNEFROCK, XIAOMEI LIU, DAVID W. OPALKA, JOHN W. SHIVER, DANILO R. CASIMIRO, and JON H. CONDRA

ABSTRACT

One of the greatest challenges in HIV vaccine development is accommodating the worldwide sequence diversity of the HIV-1 virus. To understand how viral sequence diversity may affect the potential breadth of HIV-1 vaccines designed to elicit antiviral T cell immunity, we have developed novel approaches to assess sequence conservation at the amino acid level, where vaccine effects are exerted. Taking each sequence from the LANL 2004 amino acid alignments as a potential vaccine or as a challenge virus, all pairwise combinations of sequences were evaluated by two methods: first, a traditional comparison of aligned sequences, and second, by a new walking 9-mer algorithm chosen to emphasize the typical length of an MHC-I epitope. The rules for comparing mismatched 9-mer pairs between vaccine and challenge sequences were empirically deduced from an experiment on Nef-specific CD8 epitopes and the viral sequences from naturally HIV-1-infected patients. Results were weighted such that each clade contributed in proportion to its global prevalence. Cross-clade breadth of response is best maintained for vaccines encoding Pol and Gag, while commonly proposed Env- and Tat-based vaccines would be more clade sensitive. We evaluated the additional breadth that could be expected from multiclade vaccines including consensus and ancestral sequences. For more diverse proteins, adding a second strain can add a significant increase in breadth, although for three or more strains the intrinsic diversity of the protein leads to diminishing improvement.

INTRODUCTION

With the discovery 25 years ago of the first cases of AIDS[1,2] and the identification of the HIV virus[3–5] there was initial optimism that an effective vaccine would be forthcoming. The search has been frustrated because the virus hides its epitopes, employs several mechanisms to downregulate host immune responses, and attacks the very immune system that would be responsible for its neutralization. There has been increasing evidence that antiviral T cell immunity can potentially impact viral replication.[6,7] In recent years, several vaccine candidates designed to elicit this type of immunity have entered clinical trials.[8]

The challenge of developing an effective HIV vaccine for worldwide use is complicated by viral sequence diversity. This results from several factors, including high viral replication and error rates, prolonged courses of infection, viral adaptation to immune and drug pressures, and the deposition of infecting virus and its descendants into long-lived proviral reservoirs from which they may ultimately reemerge. Besides evading the humoral and cell-mediated immune response in a single host, this leads to an astonishing diversity in the HIV virus within a local population[9] and globally.[10] In the face of geographic and social isolation of infected individuals, HIV-1 replication has given rise to multiple independently evolving viral lineages. To date, 15 major HIV-1 clades and numerous interclade circulating recombinant forms have been recognized worldwide.[11]

Given this great diversity, the selection of vaccine immunogens that are most likely to elicit broad immunity, both within and across clades, becomes very challenging. Several investigators have addressed the issue of viral diversity with vaccines that incorporate one copy of the most conserved genes (gag, pol) and/or multiple versions[12] of the less conserved genes (env) taken from the major clades. It is also unclear what types of antigen sequences should be used. For example, consensus or putative ancestral sequences[13,14] and center-of-tree modifications thereof[15,16] have been proposed as potential immunogens in order to minimize overall genetic distances between vaccine

Vaccine and Biologics Research, Merck Research Laboratories, West Point, Pennsylvania 19486.

and target viruses. However, these sequences are composites of multiple natural viral sequences that do not necessarily represent existing viruses and more specifically, could present artificial T cell epitopes. To identify the best potential sequences for inclusion in a vaccine, it is necessary to assess the distribution of amino acid sequence conservation among the naturally circulating sequences that will be the ultimate targets of an HIV-1 vaccine. For a vaccine designed to elicit cellular immunity, the broadest responses would be expected from immunogens that encode the largest number of conserved cytotoxic T-lymphocyte (CTL) epitopes, both to maximize the strength of response and to minimize the probability of viral escape.

In this article, we describe a systematic approach to model the impact of both intraclade and interclade sequence variation of selected HIV-1 antigens on vaccine responses. We also systematically evaluate the vaccines that encode proteins from multiple clades to broaden coverage. Such approaches will be essential to guide the selection of antigens for inclusion into a worldwide HIV-1 vaccine.

MATERIALS AND METHODS

Sequence alignments treatment

Amino acid sequences were downloaded from the 2004 alignments provided by the Los Alamos National Laboratory HIV Sequence database.[11] Data were cleaned by removing terminal stop codons ($), replacing internal stop codons with gaps, and unifying unknown characters (e.g., #, x were translated into X). Consensus sequences were calculated directly from the amino acid sequences, ambiguous positions were resolved by examining neighboring positions, and choosing the amino acid that was most frequent in contiguous segments.

Sequence comparisons

Two methodologies were used to estimate HIV sequence conservation within and between clades: traditional similarities of aligned sequences and N-mer set similarities. A standard sequence similarity was performed by considering each vaccine and target sequence pair and comparing corresponding amino acids. Pairs of sequences were aligned by a globally optimized Needleman–Wunsch algorithm[17] implemented with the EMBOSS suite[18] to yield a unique best score. This is preferred to scoring multiple alignments that do not in general produce the best pairwise matches. When performing alignments, typical affine gap penalties of 1 to open a gap and 0.1 to extend it were used consistently. No penalty was exacted where either sequence contained an "unknown" amino acid; the unknown was treated as a wild card during alignment but ignored when computing the score. For comparison by "identity," identical amino acids were scored as 1, while mismatched were either scored as 0 (identity matrix) or values between 0 and 1 (chemical similarity matrix). For comparison by "similarity," the EMPAR matrix of protein characteristics[19] was chosen to approximate immune recognition through MHC/peptide binding and protein processing. For instance, comparing amino acids G and M would produce a score of 0.25, while G and S would produce a score of 0.69. We anticipate that data from systematic epitope discovery and epitope mapping will inform and improve the matrix. Scores were normalized by the length of the shorter sequence.

Sequence weighting and normalization

The publicly available sequence data on HIV-1 infections is extensive, but not necessarily representative of the global pandemic. First, some patients have been repeatedly sampled (e.g., through longitudinal or tissue-specific studies) and have contributed many sequences to the databases. Second, due to scientific resource constraints, most sequences derive from the developing world and clade B in particular, although the geographic and clade representation breadth is improving with recent studies in the developing world. Because of these factors, when discussing vaccines for global distribution, we must weight the data to avoid bias toward well-studied individuals and populations. Only sequences that could be linked with a uniquely identified patient were included in the analyses. For each patient that contributed multiple sequences (M), each sequence was weighted $1/M$, so that all patients contribute equally to the overall sum. For comparisons where a mean global coverage was calculated (Figs. 2 and 4), sequences were grouped by clade, and weights for each sequence within a clade were multiplied by a constant so that each clade contributed to the total according to its relative frequency in new HIV-1 infections across the world as estimated by Osmanov *et al.*[20] This produces an overall score that is proportional to its presumed globally averaged representation.

ELISpot assay

The interferon (IFN-γ) ELISpot assay has been described previously in detail.[21] For this communication, each 9-mer peptide that could be derived from JRFL nef was synthesized (Synpep, CA), that is, for the 216 amino acid sequence, peptides with amino acids spanning 1–9, 2–10, 3–11, . . . , 207–215, and 208–216. Each 9-mer peptide was individually tested against peripheral blood mononuclear cells (PBMCs) from each of the five subjects described in the Results. For an ELISpot response to be considered as positive, the number of spot-forming cells must be $\geq$55 spots/10^6 PBMCs and $\geq$4-fold the media-only negative control wells.[22]

RESULTS

N-mer analyses of sequences

One of the difficulties of designing vaccines is that the discovery of HIV-directed T cell epitopes is still incomplete, both in terms of the diversity of HIV-1 viral sequences and the HLA backgrounds of infected patients. This is particularly true for viral isolates and HLA types prevalent in less-developed countries. Because the majority of infections occur in the developing world,[23] the need for prospective screening for identification of HIV epitopes, HLA types, and MHC binding motifs is particularly acute, and thus we used an approach that does not rely upon specific epitopes known to date but nonetheless incorporates elements relevant to T cell immunology.

Because the fundamental antigenic units of the cellular immune system are N-mer stretches internally processed and then

apropriately presented on the cell surface, we developed an "N-mer homology" algorithm designed around this principle. This is unlike traditional homology methods that consider single amino acids as the smallest meaningful unit. While our N-mer homology algorithm is applicable to N-mers of any length *N*, encompassing either MHC-II or MHC-I presentation, for brevity we focus in this communication on class I. The T cell receptor recognizes MHC-I with bound peptides of 8–10 amino acids, with most peptides 9 amino acids in length. With this choice, we are explicitly evaluating vaccines and target isolates for their homology according to what may elicit CTL recognition.

As illustrated in Fig. 1, for every pair of vaccine/target sequences, we compare the set of all successive 9-mers (aa 1–9, aa 2–10 . . .) from the vaccine sequence with the set of all successive 9-mers (aa 1–9, aa 2–10 . . .) from the target sequence. For instance, JRFL Nef protein [24,25] (GenBank accession number U63632) would generate the following 208 9-mers: 1–9, MGGKWSKRS; 2–10, GGKWSKRSV; 3–11, GKWSKRSVP; . . . ; 206–214, RELHPEYYK; 207–215, ELHPEYYKD; and 208–216, LHPEYYKDC. These 9-mers comprise a set that will be considered as potential 9-mer CTL epitopes in either a proposed vaccine or potential target (infecting virus).

All possible 9-mers from the vaccine sequence are compared with those from the target sequence. Each 9-mer in the first set is compared against every 9-mer in the second set, and the closest match is selected. The number of matches (signifying potential CTL epitope responses) between the vaccine and target sets is summed and normalized by the number of 9-mers in the target set. By considering contiguous 9-mers excised from the vaccine and target sequences, we mimic antigen processing and epitope presentation but make no sequence-specific assumptions about the likelihood of peptide cleavage or binding into the MHC-I groove, which depend sensitively on HLA types. Moreover, alignments are unnecessary and scoring is unaffected by misalignment or insertions/deletions and is independent of position. It is also readily extensible to multicomponent (multiclade) vaccines, as shown below.

In an effort to determine what degree of permissiveness within the 9-aa window is most biologically relevant, PBMCs were collected from five HIV-infected patients from Thailand and tested by IFN-γ ELISpot[21] for responses to all 208 possible 9-mer peptides derived from JRFL Nef. From these same patients, at least five independent molecular proviral DNA clones were sequenced and their Nef amino acid sequences were deduced. The sequences of all positive 9-mer ELISpot peptides (see Materials and Methods) were then compared with the most closely matched patient proviral sequence. In Table 1, ELISpot data (spots/10^6 PBMCs) for test 9-mer peptides are shown in bold and the fraction of identical amino acids between the test peptides and the patient sequences is given.

When mixed sequences were present, the best matching viral sequence was assumed to be responsible for the full ELISpot response; this generates a conservative estimate of the ability of a mismatched peptide immunogen to elicit a functional response. (*In vitro* assays with exogenous peptide cannot capture all aspects of epitope recognition; in particular, natural processing is omitted.) Of the positive responses, 21 were attributed to 9-mers fully matched between the ELISpot peptide and viral sequence, 13 were attributed to 9-mers with 1 mismatch, 6 were attributed to 9-mers with 2 mismatches, and 1 to a 9-mer with 3 mismatches, as shown in Table 1. These results indicate that immunogens that are mismatched at up to two positions within a nine amino acid span may be tolerated, in some cases, between a vaccine and challenge virus and still yield a potentially significant functional response. The specific tolerance for mismatch between two 9-mers will be sequence specific; the identity of the amino acids and their positions can be relevant, and the allowable number of mismatches can in general be more or less than 8 amino acids. To derive a generalized guideline for the N-mer homology analyses, we elected to require that 8/9 or 9/9 amino acids must match between two 9-mers to score a positive epitope response.

Protein conservation

We also performed traditional pairwise amino acid comparisons for several HIV-1 antigens using sequences from the LANL (Los Alamos National Laboratories) 2004 database as described in Materials and Methods. In this analysis, each pair of sequences was aligned, identity at each amino acid position was scored either 1 or 0, and the score was averaged

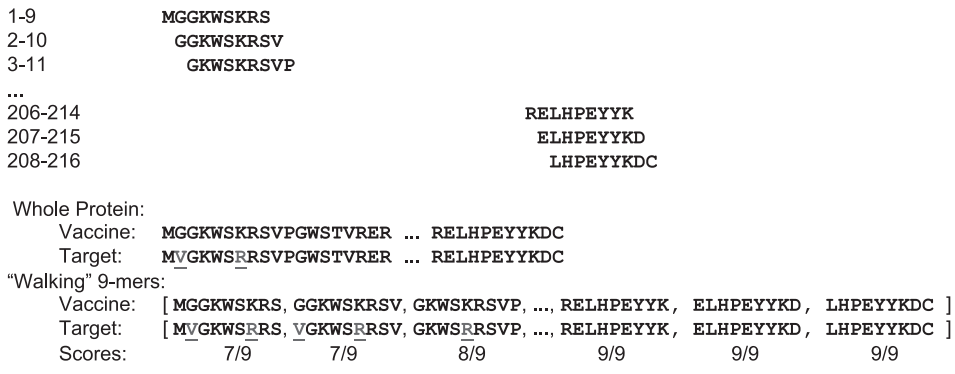

FIG. 1. Illustration of the N-mer scoring method. Top: generation of N-mers from a linear amino acid sequence. Bottom: comparison of two amino acid sequences and the number of matching amino acids in each 9-mer. The total score between two full amino acid sequences is the number of matching 9-mers exceeding a threshold score (e.g., 8/9 or 9/9).

Table 1. ELISpot Data for 9-mer Peptides and Fraction of Identical Amino Acids Between Test Peptides and Patient Sequences[a]

Subject 1	*Subject 2*		*Subject 3*		*Subject 4*	*Subject 5*	
RPQVPLRPM **350**	FPDWQNYTP **70**	PVRPQVPLR **270**	LKEKGGLEG **1130**	KRQDILDLW **310**	GAVDLSHFL **80**	GYFPDWQNY **270**	KRQDILDLW **90**
....	....	....	.R.......	R.......M	..L......	....	....
9/9	9/9	9/9	8/9	7/9	8/9	9/9	9/9
....	....	....	.R.......	R.......M	..L......	....	....
9/9	9/9	9/9	8/9	7/9	8/9	9/9	9/9
....	....	..K......	.R.......	R.......M	..L......	......H..	....
9/9	9/9	8/9	8/9	7/9	8/9	8/9	9/9
....	....	..KK.....	.R.......	R.......M	..L......	......H..	....
9/9	9/9	8/9	8/9	7/9	8/9	8/9	9/9
....	....	....	.R.E.....	R.......M	..L......	......H..	....
9/9	9/9	9/9	7/9	7/9	8/9	8/9	9/9
DEEVGFPVR **140**	DEEVGFPVR **90**	VRPQVPLRP **280**	KGAVDLSHF **400**	RQDILDLWV **150**	KEKGGLEGL **870**	YFPDWQNYT **290**	VPLRPMTYK **870**
..G......	E.G......	....	...L.....	......M.I	....	....	..V......
8/9	7/9	9/9	8/9	7/9	9/9	9/9	8/9
..G......	E.G......	....	...I.....	......M.I	....	....	..V......
8/9	7/9	9/9	8/9	7/9	9/9	9/9	8/9
..G......	E.G.....K	.K.......	...I.....	......M.I	....	......H..	....
8/9	6/9	8/9	8/9	7/9	9/9	8/9	9/9
..G......	E.G.....K	.K.......	...L.....	......M.I	....	......H..	....
8/9	6/9	8/9	8/9	7/9	9/9	8/9	9/9
..G......	E.G......	....	...I.....	......M.I	....	......H..	..V......
8/9	7/9	9/9	8/9	7/9	9/9	8/9	8/9
QDILDLWVY **180**	EEVGFPVRP **70**	QDILDLWVY **780**	DWQNYTPGP **930**	QDILDLWVY **1000**		QVPLRPMTY **180**	GAVDLSHFL **570**
.E......H	.G.......	.E.....I.	....	....		...V.....	..F......
7/9	8/9	7/9	9/9	7/9		8/9	8/9
.E......H	.G.......	.E.....I.	....	......M.I		...V.....	..W......
7/9	8/9	7/9	9/9	7/9		8/9	8/9
.E......H	.G......K	.E.....I.	....	......M.I		....	..F......
7/9	7/9	7/9	9/9	7/9		9/9	8/9
.E......H	.G......K	.E.....I.	....	......M.I		....	..L......
7/9	7/9	7/9	9/9	7/9		9/9	8/9
.E......H	.G.......	.E.....I.	....	......M.I		...V.....	..L......
7/9	8/9	7/9	9/9	7/9		8/9	8/9
EDEEVGFPV **100**	EVGFPVRPQ **180**		YHTQGYFPD **830**	TYKGAVDLS **240**		VGFPVRPQV **200**	NADCAWLEA **100**
...G.....	G.......		....	.....L...		....	...V.....
8/9	8/9		9/9	8/9		9/9	8/9
...G.....	G.......		....	.....I...		....	..A......
8/9	8/9		9/9	8/9		9/9	8/9

...G..... 8/9 ...G..... 8/9 ...G..... 8/9	G.....K.. 7/9 G.....K.. 7/9 G..... 8/9 VGFPVRPQV **240** ..... 9/9 ..... 9/9 .....K... 8/9 .....K... 8/9 ..... 9/9 GFPVRPQVP **380** ..... 9/9 ..... 9/9 ....K.... 8/9 ....K.... 8/9 ..... 9/9 FPVRPQVPL **370** ..... 9/9 ..... 9/9 ...K..... 8/9 ...K..... 8/9 ..... 9/9		..... 9/9 ..... 9/9 ..... 9/9 GLIHSQKRQ **950** .....R.. 8/9 .....R.. 8/9 .....R.. 8/9 .....R.. 8/9 GAITSSNTA **510** ..... 9/9 ..... 9/9 ..... 9/9 ..... 9/9 HGAITSSNT **120** ..... 9/9 ..... 9/9 ..... 9/9 ..... 9/9	.....I... 8/9 .....L... 8/9 .....I... 8/9 YKGAVDLSH **130** .....L... 8/9 .....I... 8/9 ....I.... 8/9 ....L.... 8/9 ....I.... 8/9 DRVRRTEPA **160** ..MK.A... 6/9 E.MK.A... 5/9 E.MK.A... 5/9 E.MK.A... 5/9 E.MK.A... 5/9 GAVDLSHFL **1220** ..L..... 8/9 ..I..... 8/9 ..I..... 8/9 ..L..... 8/9 ..I..... 8/9		..... 9/9 ..... 9/9 ..... 9/9 FPVRPQVPL **360** .....V 8/9 .....V 8/9 ..... 9/9 ..... 9/9 .....V 8/9 PVRPQVPLR **70** .....V. 8/9 .....V. 8/9 ..... 8/9 ..... 8/9 .....V. 8/9 VRPQVPLRP **70** .....V.. 8/9 .....V.. 8/9 ..... 9/9 ..... 9/9 .....V.. 8/9	...V..... 8/9 ...V..... 8/9 ...V..... 8/9

[a]For each of the patients whose epitopes were mapped by progressive ELISpot assays to single 9-mer peptides, at least five independent molecular proviral DNA clones were sequenced and their Nef amino acid sequences were deduced. The sequences of all ELISpot-positive 9-mer peptides were then compared with the most closely matched patient proviral sequence. ELISpot data (spot-forming cells/10^6 PBMCs) for test 9-mer pepties are shown in bold and the fraction of identical amino acids between test peptides and the patient sequences is given. Immunogens mismatched at up to two positions within a nine amino acid span may be tolerated between a vaccine and challenge virus and still yield a potentially significant functional response.

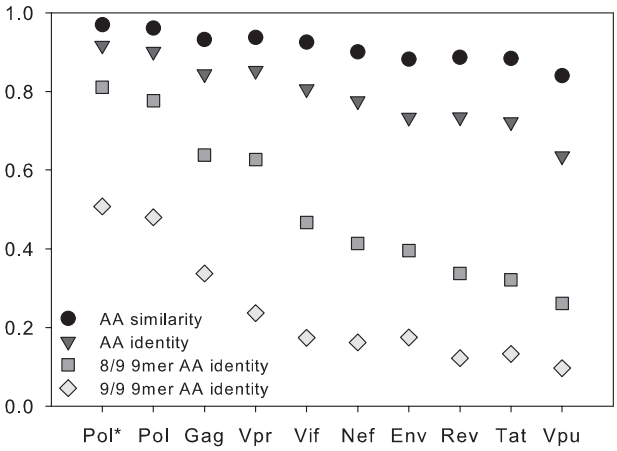

FIG. 2. Mean amino acid conservation between HIV-1 isolates coding a given protein. Pol* indicates Pol(RT-IN) with protease removed for safety considerations; this also enhances interisolate conservation as compared with Pol. Conservation scores were calculated with either a traditional homology (strict amino acid identity or amino acid chemical similarity) or the N-mer method described in the text where either eight or nine amino acids in a 9-mer were required to count a match. Although the relative score diminishes by requiring identity vs. chemical similarity and by requiring 9-mers to match (instead of only single amino acids), the hierarchy of conservation among the proteins is broadly similar. Target isolates (patient sequences) were weighted according to each clade's relative prevalence in the global population, as described in the text.

over the length of the aligned sequence. Figure 2 shows the overall amino acid conservation for each HIV-1 protein. Pol* indicates the portion of Pol spanning only the reverse transcriptase (RT) and integrase proteins; in our current vaccine candidate, the highly variable protease was deleted for potential clinical safety considerations. Each calculation is normalized to protein length to permit comparisons between proteins. Four different scoring thresholds are shown, considering whole protein comparisons by either amino acid identity or chemical similarity; for 9-mer comparisons, peptide pairs were considered matched if identical either at 9/9 or at $\geq$8/9 residues. In general, whole-protein comparisons are more forgiving than N-mer calculations because a single amino acid substitution affects only a single position in a protein while *N* N-mers are affected by the same substitution in an N-mer calculation. The harsh penalty is mitigated somewhat by 9-mer rules that permit eight out of nine to match.

Figure 3 illustrates intraclade and interclade 9-mer similarities for the commonly considered protein antigens Env, Gag, Pol*, and Nef with respect to clades A, B, and C. Sequences designated as clade A, A1, or A2 were all considered to be clade A due to the relative paucity of sequences in these clades. For brevity only the intermediate case that requires eight or nine amino acids to match is shown. These data show that the hierarchy of either intraclade or cross-clade similarity is Pol* > Gag > Nef, Env. However, the intraclade similarities for Nef or Env approximate the cross-clade similarities found for Gag. Therefore, achieving cross-clade coverage for Nef or Env with a single vaccine immunogen is predicted to prove to be more challenging than for Gag.

Multiclade vaccines

The N-mer method can be generalized to multistrain vaccines, something that is difficult to do with traditional similarity comparisons. For instance, consider a multistrain vaccine encoding two versions of the same protein, perhaps archetypes from different clades. It is possible to compute the homology between vaccine strain 1 and a natural isolate and between vaccine strain 2 and the same isolate; however, we are not aware of any unambiguous and unique way to produce a single figure of merit that combines the two independent homology scores. In contrast, for any given choice of function for N-mer cross-recognition, the N-mer method readily lends itself to a unique score, as follows. Each vaccine N-mer is compared against each natural isolate N-mer, as shown in Fig. 1. For the two vaccine strains, each shares the same natural isolate. Then, the best of the two matching scores is kept for each natural isolate N-mer. By summing these best scores over all isolate N-mers (just as is done for a single vaccine strain), a unique score that incorporates the closest matching N-mers from any vaccine strain is calculated. This can be generalized immediately to comparing three or more vaccine strains against a target isolate.

Figure 4 shows the globally weighted scores for vaccines encoding one, two, or three versions of the proteins Env, Gag, Pol*, and Nef. The scores represent the coverage that a potential vaccine would be expected to have, on average, against new HIV-1 infections globally. The scores implicitly take each subunit and each version to be equally immunogenic and noninterfering. For vaccines containing only a single sequence per gene, a weighted consensus produced the highest score, followed by a clade C consensus (clade C is more common worldwide than A or B). Putative ancestral sequences for group M scored less well but nonetheless exceeded single clade A or clade B consensus sequences. Mean coverage scores were enhanced in all cases when two sequences per gene were considered, if one sequence was a consensus of clade C. Triple sequence per antigen type generated only marginally higher scores. Because the coverage is calculated in terms of potential N-mer epitopes, if an N-mer is not included within one strain of a multistrain vaccine, it may be included within a second (or third) strain, increasing the mean global coverage score. Interestingly, the cumulative ability of three Nef or Env copies to cover the global sequence diversity of these antigens is comparable to that provided by a single copy of Gag.

It is also possible to compare clade-specific ancestral with clade-specific consensus sequences. In most cases, consensus and ancestral sequences, either single or multisequence vaccines, yielded scores that were different by less than 3% weighted according to the global distribution. One exception is the single-strain Nef vaccine where the ancestor derived[26] from A1 scored higher than did the A consensus. This may be a consequence of generating a clade A consensus for the distinct A1/A2 subtypes. In general, we have grouped A1 and A2 together in our analyses due to the relatively small number of

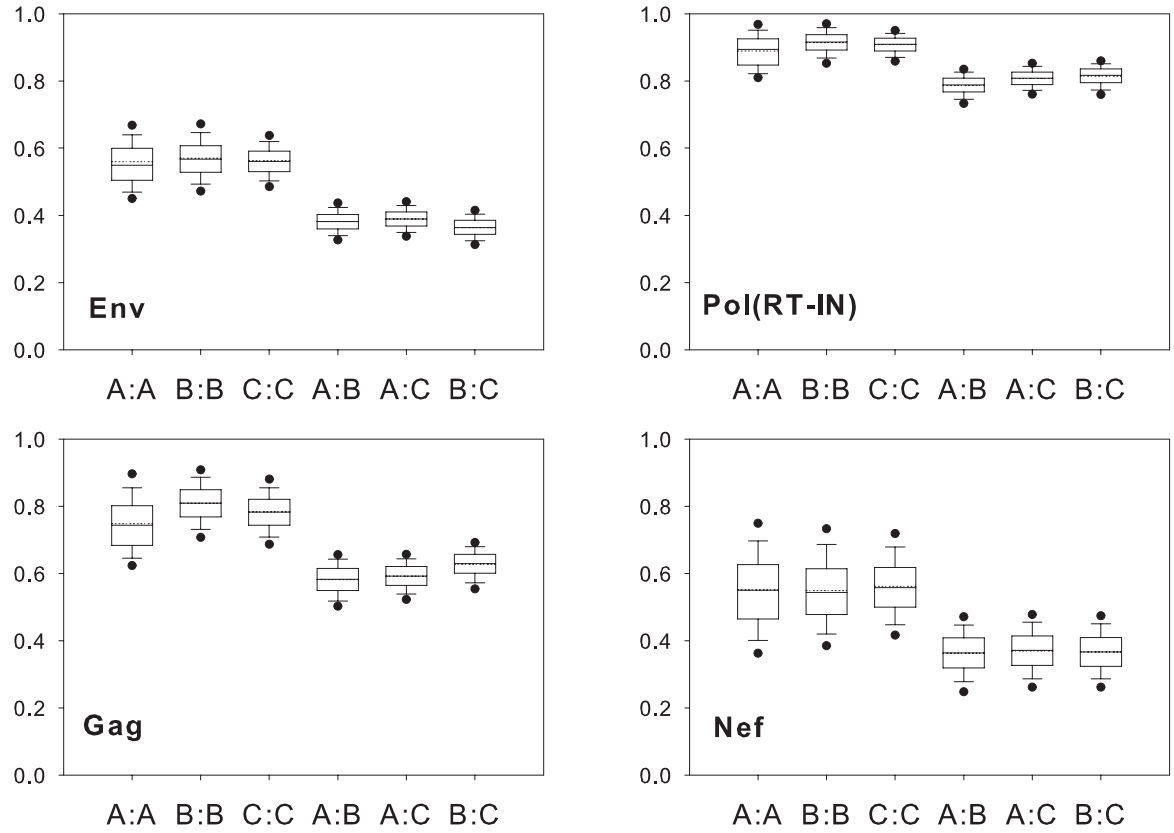

FIG. 3. Amino-acid conservation between HIV-1 proteins. Conservation scores were calculated with the N-mer method where eight or nine identical positions between two 9-mers is considered a match. Box center lines indicate the median (solid) and mean (dashed); these overlap in most cases so only one line is visible. Box boundaries indicate 25th and 75th percentiles, whiskers are at 10% and 90%, and dots are at 5% and 95%. Intraclade distributions are distinguishably higher from interclade. Due to heterogeneity between A1 and A2, intraclade (A–A) scores tend to be lower than other intraclade comparisons (B–B and C–C). Notably, the intraclade means from Nef and Env are comparable to the interclade means from Gag. This emphasizes the high conservation present in Pol and Gag relative to all other HIV-1 proteins.

A1 and especially A2 isolates[27] that have been sequenced. Notably, the Pol* global scores are dramatically higher (5–7%) for single or multistrain vaccines when consensuses are used instead of ancestral sequences. The maximum score of 96% for a combination of A, B, and C consensuses holds the possibility that properly constructed consensus vaccines against well-conserved targets can have almost complete global coverage.

DISCUSSION

The major issues in selecting the immunogens for a worldwide HIV-1 vaccine concern (1) the choice of antigens, (2) the number of copies of the same antigen type, and (3) the clades from which the antigen would be selected. These questions have been difficult to address due to the lack of a quantitative framework. Here, we have proposed a method to rationally evaluate hypothetical vaccines in a manner specifically relevant to vaccine design. The phylogenetic approach used before is ideally suited to viral evolution, the origins of diversity, the timing of significant events, and the effects of recombination as subtypes merge. The N-mer approach does not answer these questions but instead addresses how to evaluate vaccines against present-day real HIV-1 isolates. Beyond the fact that N-mers are the fundamental unit for cellular immunity, the approach requires a model for N-mer cross-recognition that can be derived from empirical data as performed in this communication. We have made no attempt to incorporate known epitopes or HLA-specific motifs. This is a potential limitation, but a necessary one given the currently limited number of known epitopes, particularly for non-clade B strains and non-Caucasian HLA types. For a worldwide vaccine, this would be especially inappropriate because it may bias the analysis toward a minor subset of viral infections and of host antiviral recognition motifs. Significant efforts are being made to overcome these limitations including epitope mapping by ELISpot[28,29] and flow cytometry.[29,30] As the list of epitopes becomes more complete and prospective screens are performed, then those epitopes can be readily incorporated into this type of analysis. For example, the N-mer scores of known epitopes could be increased relative to nonepitopes to refine the analysis. Additionally, as continued cross-recognition epitope mapping studies are performed, either in vaccine trials comparing the vaccine immunogen to non-

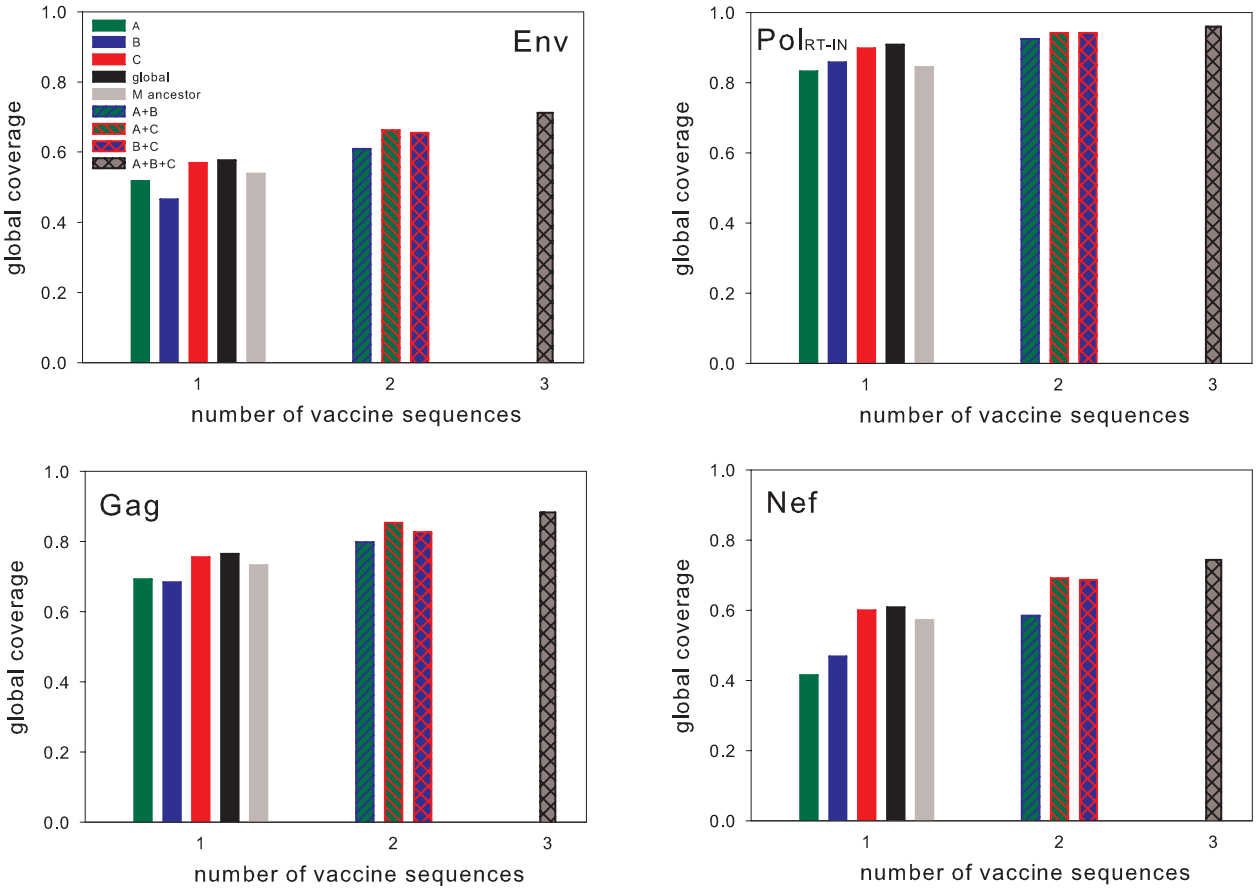

FIG. 4. Mean amino acid conservation for the vaccines indicated in the legend. Conservation scores were weighted according to the global prevalence of clades A, B, and C as described in the text. For vaccines containing only a single sequence per gene, a weighted consensus produced the highest score, followed by a clade C consensus (clade C is more common than A and B). Putative ancestral sequences for group M scored less well but nonetheless exceeded clade A and clade B consensus sequences. Mean conservation scores were enhanced in all cases when two sequences per gene were considered, if one sequence was a consensus of clade C. Triple sequence per gene constructs generated only marginally higher scores.

homologous responses, or epidemiological studies comparing HIV infection sequences to a defined set of peptides with known HLA types, these data can inform a more complex model function for N-mer cross-recognition.

Sequences such as consensuses derived from sequence alignments or putative ancestral sequences pose the concern that vaccine responses may be elicited against either artificial junctions or historical epitopes that are no longer prevalent in the present-day epidemic. Another method has recently been proposed[31] to construct cocktails of mosaic sequences stochastically derived from populations of present-day sequences. The analyses of the cocktails includes an N-mer analysis similar to that described here, and indicate that an artificial sequence can improve overall N-mer coverage within clade C or group M sequences. Regardless of the choice of antigen sequence—consensus, ancestor, mosaic cocktail, or natural isolate—the candidates should be assessed relative to sequence data properly weighted to account for the number of persons likely to be infected by HIV-1 virus of each clade to address potential efficacy in a worldwide setting.

The hierarchy of in-clade and cross-clade amino acid sequence conservation among potential vaccine immunogens is Pol > Gag,Vpr > Vif > Nef,Env > Rev,Tat > Vpu. The conservation of Pol is further enhanced by omitting protease. That this ranking is consistent across multiple homology methods and assumptions of stringency for cross-recognition enables us to answer comparative questions between genes, clades, and potential vaccines. Incorporating highly conserved proteins in a vaccine dramatically improves the likelihood of cross-clade efficacy, and Pol and Gag meet this criterion. When selecting less well-conserved genes for inclusion, including two or more versions from different clades can enhance overall coverage, although each additional version adds diminishing benefit.

In summary, we present a novel general approach to assess the antigenicity of a candidate vaccine immunogen(s) in the context of known viral sequence diversity. This method, for the first time, provides the ability to score sequences without multiple sequence alignments, without additional parameters to deal with the resultant artificial gaps, and can evaluate multiple vaccine sequences against the same target protein to yield a

uniquely determined score unlike traditional homologies that do not facilitate the combination of scores into a single figure of merit. These features are particularly useful for evaluating and comparing vaccines that contain multiple versions of the same antigen to address cross-clade variability and against a large number of sequences in a systematic manner.

HIV antigens such as Pol and Gag can provide significant coverage both within a clade and across different clades. Moving to other antigens such as Env or Nef reduces both intraclade and interclade coverage, although this can be improved by including additional versions of these antigens, particularly if the additional antigen is selected from clade C, the globally most prevalent clade. Adding a third version (e.g., encompassing clades A, B, and C) yields further improvement. However, this approach has limits, as adding further versions of the same antigen yields diminishing results in each case and increases the challenge imposed by the packaging limits of any delivery vector. Breadth of coverage is a crucial question for vaccines against the global pandemic of HIV-1. Not only is vaccine breadth essential to ensure that a diverse set of viral sequences can be addressed, but increasing the number of epitopes is associated with lower viral setpoints and better control during chronic infection. The various strategies for increasing vaccine breadth, including multistrain and consensus/ancestral vaccines, can be evaluated and guided through the analyses described here in a manner relevant to T cell recognition.

ACKNOWLEDGMENTS

The authors acknowledge DNA sequence analysis and review from Holly Hammond, Dana Wood, and Zhiyu Fang, helpful discussions with Dr. Emilio Emini, and advice on the ELISpot assay from Sheri Dubey and Kelly Collins.

REFERENCES

1. Gottlieb MS, Schroff R, Schanker HM, *et al.*: Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: Evidence of a new acquired cellular immunodeficiency. N Engl J Med 1981;305:1425–1431.
2. Masur H, Michelis MA, Greene JB, *et al.*: An outbreak of community-acquired Pneumocystis carinii pneumonia: Initial manifestation of cellular immune dysfunction. N Engl J Med 1981; 305:1431–1438.
3. Barre-Sinoussi F, Chermann JC, Rey F, *et al.*: Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 1983;220:868–871.
4. Popovic M, Sarngadharan MG, Read E, and Gallo RC: Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science 1984;224:497–500.
5. Levy JA, Hoffman AD, Kramer SM, Landis JA, Shimabukuro JM, and Oshiro LS: Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. Science 1984;225:840–842.
6. Koup RA, Safrit JT, Cao Y, *et al.*: Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. J Virol 1994;68:4650–4655.
7. Shiver JW, Fu TM, Chen L, *et al.*: Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. Nature 2002;415:331–335.
8. Girard MP, Osmanov SK, and Kieny MP: A review of vaccine research and development: The human immunodeficiency virus (HIV). Vaccine 2006;24:4062–4081.
9. McCutchan FE, Viputtigul K, de Souza MS, *et al.*: Diversity of envelope glycoprotein from human immunodeficiency virus type 1 of recent seroconverters in Thailand. AIDS Res Hum Retroviruses 2000;16:801–805.
10. McCutchan FE: Global epidemiology of HIV. J Med Virol 2006;78(Suppl. 1):S7–S12.
11. Leitner T, Foley B, Hahn B, *et al.*: *HIV Sequence Compendium 2005*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 2005.
12. Zhan X, Martin LN, Slobod KS, *et al.*: Multi-envelope HIV-1 vaccine devoid of SIV components controls disease in macaques challenged with heterologous pathogenic SHIV. Vaccine 2005;23: 5306–5320.
13. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, and Detours V: Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 2001;58:19–42.
14. Gaschen B, Taylor J, Yusim K, *et al.*: Diversity considerations in HIV-1 vaccine selection. Science 2002;296:2354–2360.
15. Nickle DC, Jensen MA, Gottlieb GS, *et al.*: Consensus and ancestral state HIV vaccines. Science 2003;299:1515–1518; author reply 1515–1518.
16. Mullins JI, Nickle DC, Heath L, Rodrigo AG, and Learn GH: Immunogen sequence: The fourth tier of AIDS vaccine design. Expert Rev Vaccines 2004;3:S151–159.
17. Needleman SB and Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
18. Rice P, Longden I, and Bleasby A: EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 2000;16: 276–277.
19. Rao JM: New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int J Pept Protein Res 1987;29:276–281.
20. Osmanov S, Pattou C, Walker N, Schwardlander B, and Esparza J: Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr 2002;29:184–190.
21. Coplan PM, Gupta SB, Dubey SA, *et al.*: Cross-reactivity of anti-HIV-1 T cell immune responses among the major HIV-1 clades in HIV-1-positive individuals from 4 continents. J Infect Dis 2005;191:1427–1434.
22. Fu T-M, Dubey SA, Mehrotra DV, *et al.*: Evaluation of cellular immune responses in subjects chronically infected with HIV-1. AIDS Res Hum Retroviruses 2007;23:67–76.
23. UNAIDS: 2006 Report on the global AIDS epidemic. At http://www.unaids.org/en/HIV_data/2006GlobalReport/default.asp, 2006.
24. Clark GL and Vinters HV: Dementia and ataxia in a patient with AIDS. West J Med 1987;146:68–72.
25. Koyanagi Y, Miles S, Mitsuyasu RT, Merrill JE, Vinters HV, and Chen IS: Dual infection of the central nervous system by AIDS viruses with distinct cellular tropisms. Science 1987;236: 819–822.
26. Korber B, Muldoon M, Theiler J, *et al.*: Timing the ancestor of the HIV-1 pandemic strains. Science 2000;288:1789–1796.
27. Visawapoka U, Tovanabutra S, Currier JR, *et al.*: Circulating and unique recombinant forms of HIV type 1 containing subsubtype A2. AIDS Res Hum Retroviruses 2006;22:695–702.
28. Currier JR, Visawapoka U, Tovanabutra S, *et al.*: CTL epitope distribution patterns in the Gag and Nef proteins of HIV-1 from subtype A infected subjects in Kenya: Use of multiple peptide sets increases the detectable breadth of the CTL response. BMC Immunol 2006;7:8.

29. Karlsson AC, Martin JN, Younger SR, *et al.*: Comparison of the ELISPOT and cytokine flow cytometry assays for the enumeration of antigen-specific T cells. J Immunol Methods 2003;283:141–153.
30. Suni MA, Dunn HS, Orr PL, *et al.*: Performance of plate-based cytokine flow cytometry with automated data analysis. BMC Immunol 2003;4:9.
31. 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.

Address reprint requests to:
Adam C. Finnefrock
Vaccine Basic Research
Merck Research Laboratories
770 Sumneytown Pike
West Point, Pennsylvania 19486

E-mail: adam_finnefrock@merck.com