Coping with Viral Diversity in HIV Vaccine Design

David C. Nickle[1], Morgane Rolland[1], Mark A. Jensen[1¤a], Sergei L. Kosakovsky Pond[2], Wenjie Deng[1], Mark Seligman[1¤b], David Heckerman[3], James I. Mullins[1*], Nebojsa Jojic[3]

1 Department of Microbiology, University of Washington School of Medicine, Seattle, Washington, United States of America, 2 Department of Pathology, University of California San Diego, La Jolla, California, United States of America, 3 Microsoft Research, Redmond, Washington, United States of America

The ability of human immunodeficiency virus type 1 (HIV-1) to develop high levels of genetic diversity, and thereby acquire mutations to escape immune pressures, contributes to the difficulties in producing a vaccine. Possibly no single HIV-1 sequence can induce sufficiently broad immunity to protect against a wide variety of infectious strains, or block mutational escape pathways available to the virus after infection. The authors describe the generation of HIV-1 immunogens that minimizes the phylogenetic distance of viral strains throughout the known viral population (the center of tree [COT]) and then extend the COT immunogen by addition of a composite sequence that includes high-frequency variable sites preserved in their native contexts. The resulting COT$^+$ antigens compress the variation found in many independent HIV-1 isolates into lengths suitable for vaccine immunogens. It is possible to capture 62% of the variation found in the Nef protein and 82% of the variation in the Gag protein into immunogens of three gene lengths. The authors put forward immunogen designs that maximize representation of the diverse antigenic features present in a spectrum of HIV-1 strains. These immunogens should elicit immune responses against high-frequency viral strains as well as against most mutant forms of the virus.

Citation: Nickle DC, Rolland M, Jensen MA, Pond SLK, Deng W, et al. (2007) Coping with viral diversity in HIV vaccine design. PLoS Comput Biol 3(4): e75. doi:10.1371/journal.pcbi.0030075

Introduction

The failure of AIDS vaccine efforts in the past 20-plus years is thought to be due, in part, to the enormous viral antigenic diversity found within and among patients with human immunodeficiency virus type 1 (HIV-1) infection. However, until recently, relatively little effort had been devoted to choosing particular viral variant sequences or designing sequences to include within vaccines [1,2]. There were early attempts to design vaccines by concatenating commonly recognized T cell and antibody epitopes [3], but these did not produce a viable vaccine candidate. New methods of combining epitopes are being explored in vaccine design, including production of pseudoprotein strings of T cell epitopes [4], and the synthetic scrambled antigen vaccine (SAVINE) [5], which employs consensus overlapping peptide sets from HIV-1 proteins scrambled together. Focusing on the use of whole viral protein sequences, natural strains (NSs) as well as consensus (CON) sequences are being used as a means to minimize the abrogating effect of antigenic diversity in vaccine antigens [2,6,7], as are the inferred most recent common ancestors (MRCA, or ANC) [6,8–10] of targeted virus populations defined as sequences that reside at the basal node of the set of in-group sequences in a phylogenetic tree reconstruction [11]. HIV-1 *env* sequences representing both the CON and ANC have been prepared and studied, but neither has generated exceptionally broad humoral immune reactivity in initial small animal studies [7,12].

In an effort to develop antigens that capture both the summary of circulating variation found in CON estimates, and the coupling of mutations generated with inferred ANC sequences, we have developed an alternative computational method that reconstructs the ancestral state sequence at the center of tree (COT) ([13] and Rolland M, Jensen MA, Nickle DC, Learn GH, Heath L, et al., unpublished data). The COT sequence explicitly minimizes genetic distance, as does the CON, and because it is derived from a phylogenetic tree, it embodies the most likely mutational coupling relationships found in the ANC. Despite these efforts, it may be that no single unit-length antigen, including any NS, CON, ANC, or COT, will encompass sufficient antigenicity to elicit protective immune responses against a broad array of viruses [7,12], as will be required of an AIDS vaccine. This led us to hypothesize that we would need more than one antigenic sequence, or greater than one gene length of the antigen, to elicit protection against the broad antigenic diversity encountered in natural infections. However, cocktails of large numbers of native, full-length NS antigens would quickly become unmanageably complex for practical use as vaccines.

Here, we propose a means to cope with HIV-1 diversity by engineering vaccine antigen constructs to include short protein sequences present at high frequencies in natural

Editor: Sebastian Bonhoeffer, ETH Zürich, Switzerland

Received July 11, 2006; **Accepted** March 6, 2007; **Published** April 27, 2007

Abbreviations: ANC, ancestors; CON, consensus; COT, center of tree; CTL, cytotoxic T lymphocyte; HIV-1, human immunodeficiency virus type 1; NS, natural strain

* To whom correspondence should be addressed. E-mail: jmullins@u.washington.edu

¤a Current address: Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, United States of America

¤b Current address; Department of Statistics, University of Washington, Seattle, Washington, United States of America

Author Summary

The ability of human immunodeficiency virus type 1 (HIV-1) to acquire mutations that preserve virus viability yet evade immune responses contributes to the current failure in producing a vaccine. We describe the generation of candidate HIV-1 immunogens that include multiple forms of variable elements of the virus including some that retain colinearity with the virus and thus are expected to retain protein function. These antigens compress the variation found in many viral strains into lengths suitable for vaccine immunogens. For example, we can capture 62% of the variation found in the Nef protein and 82% of the variation in the Gag protein into immunogens of three gene lengths. We put forward immunogen designs that maximize representation of the diverse antigenic features present in a spectrum of HIV-1 strains. These immunogens should elicit immune responses against high frequency viral strains as well as against most mutant forms of the virus.

viral populations. Currently, this method is explicitly directed toward developing $CD8^+$ cytotoxic T lymphocyte (CTL) responses, which are critical to controlling viremia during infection [14–17]. Because the cumulative strength of the CTL-mediated immune response depends on the presence of recognizable epitopes (often approximately nine amino acids in length) in the target proteins, it is logical to seek to maximize epitope coverage within a vaccine antigen. However, although substantial, our current catalog of known CTL epitopes appears to be woefully incomplete [18], hence our strategy relies on the universe of HIV sequences and not solely on known epitope content. Thus, here we will define *coverage* as the sum of the frequencies of all nine amino acid segments (9mers) where the frequency is derived from random independent HIV-1 subtype B isolates found in the vaccine construct. As our epitope catalog increases and our knowledge of protein degradation, CTL epitope binding, and HLA presentation is expanded, this epitope-specific data can be integrated into the measure of coverage (e.g., by weighting epitope frequencies in accordance to their relative "importance" when computing coverage). In this study, we applied our method to Nef because it is highly variable and is potentially very difficult to design an immunogen against, and to Gag because it is immunologically important yet more conserved. We considered subtype B sequences because more immunological information is available about this subtype than any other. This clearly makes the vaccine construct described here as region-specific because of the biogeographic nature of the distribution of viral subtypes across the globe [19]. However, our purpose is to illustrate and demonstrate that this method has promise at producing a vaccine against highly variable infectious agents such as HIV.

Methods

Vaccination with all known viral sequences would capture all known viral sequence variation, but realistic vaccine constructs might at best include several sequence lengths, each length containing major variants for immune presenta-

Table 1. GenBank IDs of Sequences Used

Gag Sequences

AB078005	AF538307	AY206660	AY751406	AY835762	DQ127542
AB078703	AJ271445	AY206661	AY751407	AY835764	DQ127548
AB078704	AJ437030	AY206662	AY779550	AY835765	DQ295192
AB078709	AJ437033	AY206663	AY779553	AY835766	DQ295193
AB078711	AJ437038	AY206664	AY779556	AY835769	DQ295195
AB097870	AJ437039	AY247251	AY779557	AY835770	DQ487188
AB221005	AJ437044	AY275555	AY779563	AY835772	DQ487189
AF003887	AJ437047	AY275556	AY779564	AY835774	DQ487190
AF004394	AJ437050	AY275557	AY786790	AY835776	DQ487191
AF042100	AJ437051	AY308760	AY786830	AY835777	K02007
AF042101	AJ437058	AY308762	AY786870	AY835778	L02317
AF042102	AY173951	AY314044	AY786910	AY835779	M13136
AF042103	AY173952	AY314063	AY786919	AY835780	M17451
AF042104	AY173954	AY331283	AY786920	AY839827	M19921
AF042105	AY173955	AY331285	AY786949	AY857022	M26727
AF049494	AY173956	AY331287	AY786952	AY857144	M38429
AF049495	AY180905	AY331290	AY786962	AY857165	M38431
AF069140	AY206647	AY331292	AY818644	AY970946	M93258
AF075719	AY206648	AY331297	AY819715	AY970950	U21135
AF086817	AY206649	AY332236	AY835748	CQ958304	U23487
AF146728	AY206651	AY352275	AY835751	D10112	U26546
AF224507	AY206652	AY423387	AY835753	DQ085869	U34603
AF256204	AY206653	AY560107	AY835754	DQ097739	U34604
AF286365	AY206654	AY560108	AY835755	DQ097744	U39362
AF538302	AY206656	AY560109	AY835757	DQ097745	U43096
AF538303	AY206657	AY560110	AY835758	DQ097747	U43141
AF538304	AY206658	AY679786	AY835759	DQ127536	U69584
AF538305	AY206659	AY682547	AY835761	DQ127539	U71182
AF538306					

Nef Sequences

AB012824	AF129350	AF203153	AF538302	AY786750	L15518
AB034257	AF129351	AF203161	AF538304	AY835748	M17451
AB034272	AF129352	AF203165	AF538305	AY835751	M21098
AB078005	AF129354	AF203172	AF538306	AY835753	M26727
AB221005	AF129355	AF203180	AJ271445	AY835762	M58173
AF004394	AF129362	AF203188	AJ430664	AY835765	M93259
AF011471	AF129364	AF203192	AY037269	AY835770	U03295
AF011474	AF129369	AF203194	AY037282	AY835772	U03338
AF011481	AF129370	AF203198	AY116676	AY835776	U03343
AF011487	AF129372	AF219672	AY116713	AY835779	U12055
AF011493	AF129373	AF219685	AY116714	AY835780	U16863
AF042101	AF129375	AF219691	AY116727	AY857022	U16875
AF047082	AF129376	AF219729	AY116781	AY857144	U16934
AF063926	AF129377	AF219755	AY116805	AY899356	U23487
AF069139	AF129378	AF219760	AY116830	AY899382	U24455
AF120745	AF129379	AF219765	AY121441	DQ007902	U26087
AF120772	AF129382	AF219771	AY173951	DQ085869	U26110
AF120840	AF129388	AF219782	AY308762	DQ121815	U26119
AF120851	AF129389	AF219792	AY314063	DQ121883	U26138
AF120867	AF129390	AF219800	AY331285	DQ127537	U34603
AF120887	AF129392	AF219812	AY331290	DQ127548	U43106
AF120898	AF129394	AF219819	AY331293	DQ487191	U44444
AF120909	AF203108	AF219845	AY352275	DQ659737	U44450
AF129334	AF203111	AF238268	AY444311	L07422	U44462
AF129335	AF203116	AF252897	AY713408	L15482	U44468
AF129342	AF203126	AF252910	AY739040	L15489	U66543
AF129343	AF203137	AF462708	AY779550	L15500	U69584
AF129346	AF203141	AF462753	AY786630	L15515	U71182
AF129347					

doi:10.1371/journal.pcbi.0030075.t001

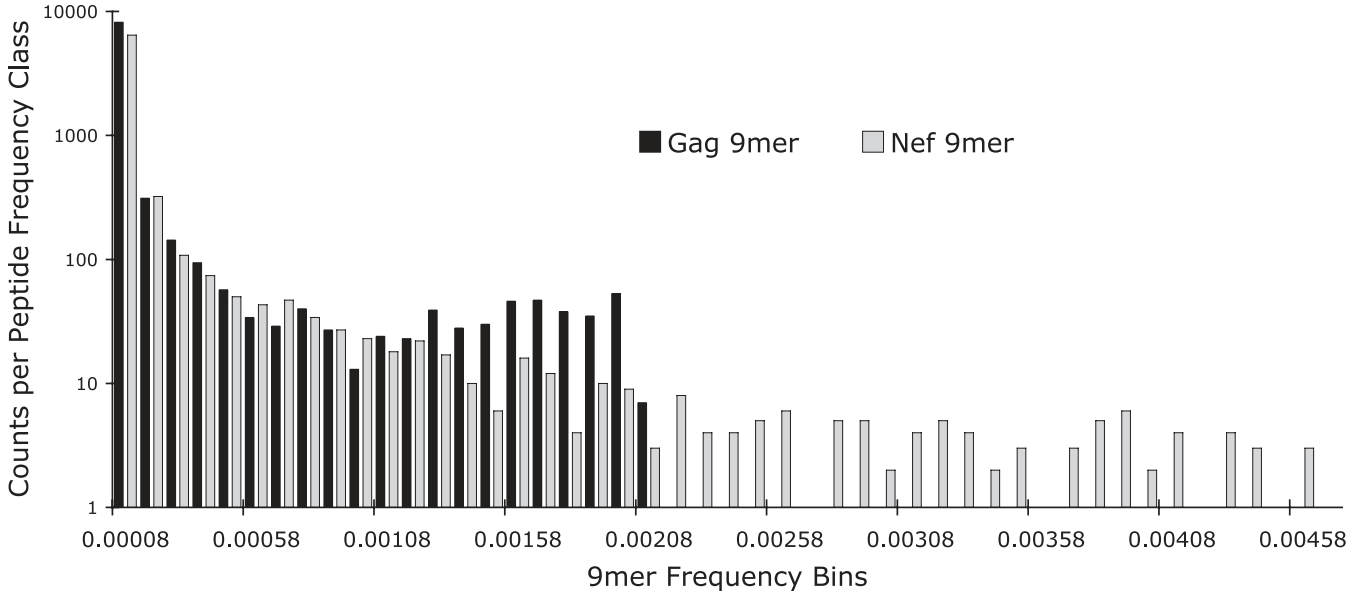

Figure 1. 9mer Peptide Distribution Derived from 169 HIV-1 Subtype B Gag and Nef Protein Sequences

Each bin in the histogram represents the number of 9mers from a particular frequency class plotted on a log scale. There are only a few peptides found at high frequencies, whereas most of the 9mers occur only once or twice. The score of a given frame is the sum of the frequencies of each unique 9mer contained by the frame. The possible extreme value frequencies for each peptide from all rare to all common is $1.198 \times 10^{-5} - 0.0020$ for Gag (black bars) and $2.988 \times 10^{-5} - 0.0051$ for Nef (gray bars). The differences in the two distributions can be explained by the differences in gene length and levels of conservation.

doi:10.1371/journal.pcbi.0030075.g001

tion. To quantify variant representation and rationally choose the included variation on this basis, Jojic and colleagues have proposed a method based on machine-learning for the compression of sequence variation into a sequence of minimal length (the "epitome"; [20,21]). Below, we describe an alternative, more transparent algorithm also designed to attain optimized sequence coverage over a fixed-length antigen. We refer to the constructs generated by our method as COT^{+} because they consist of COT antigens augmented by the addition of high-frequency 9mers. We demonstrate the performance of our approach on the highly variable and epitope-rich viral Nef protein as well the epitope-rich major structural protein, Gag. The algorithm consists of five steps applied to a sample of viral nucleotide sequences, each isolated from a separate patient. We started with all publicly available *nef* and *gag* gene sequences from HIV-1 subtype B [22]. By excluding sequences with more than two stop codons and with large indels, and including only independent single sequences from a given individual to avoid sampling bias, we obtained a 169-sequence dataset for Gag; the Nef data set was also constrained to 169 sequences for comparative purposes (Table 1 includes the GenBank IDs of all sequences used). The algorithm, however, can rapidly process datasets with thousands of sequences when such datasets become available.

The Algorithm

(1) A COT sequence is calculated as described ([13 and Rolland M, Jensen MA, Nickle DC, Learn GH, Heath L, et al., unpublished data) from a phylogenetic tree that captures the relationships among genes in the sample using maximum likelihood methods [23]. Briefly, from aligned sequences we estimate a maximum likelihood tree under a HKY + Γ + I model of evolution in PAUP*v4beta10 [24]. The resulting tree is re-rooted at the point that describes the least-squares distance to all the tips on the phylogeny (the COT node). We then infer the maximum likelihood state using the same model of evolution as above.

(2) A table of unique 9mer peptides [20,21] with their corresponding frequencies (the 9mer distribution) is constructed from translated protein sequences. To illustrate this, note that if our sample contained N identical sequences of length L each, but every 9mer in the COT peptide library was unique, then each peptide would be at equal frequency $\frac{1}{L-8}$. On the other hand, if every sequence were different from all others, to the extent that no 9mer was represented twice, the frequency for each peptide would equal $\frac{1}{N(L-8)}$. Actual samples will yield an intermediate distribution that can be exploited for vaccine design (see Figure 1). We used this distribution to compute "coverage"; that is, as we select candidate fragments to be included in the potential vaccine, we will select only those fragments that are highly represented under the 9mer curve.

(3) Unique or rare 9mers, which by definition are unlikely to be common in circulating viral strains, are likely to derive from low-fitness variants [25,26] and, because of their low frequency, have low probability of being incorporated in our vaccine constructs. Specifically, we calculate the frequency of all observed mutations at each site, and revert any mutation with a frequency below a fixed "smoothing" threshold, M, to the corresponding character in the COT sequence. All 9mers present in the COT sequence are then removed from the 9mer distributions before proceeding to the next step.

(4) Given a fixed window size F (ranging from 9 to L, where L = the length of the protein sequence [we start with 9 because that is the size of the peptide that is most often found to encode epitope sequences] and a stride parameter S

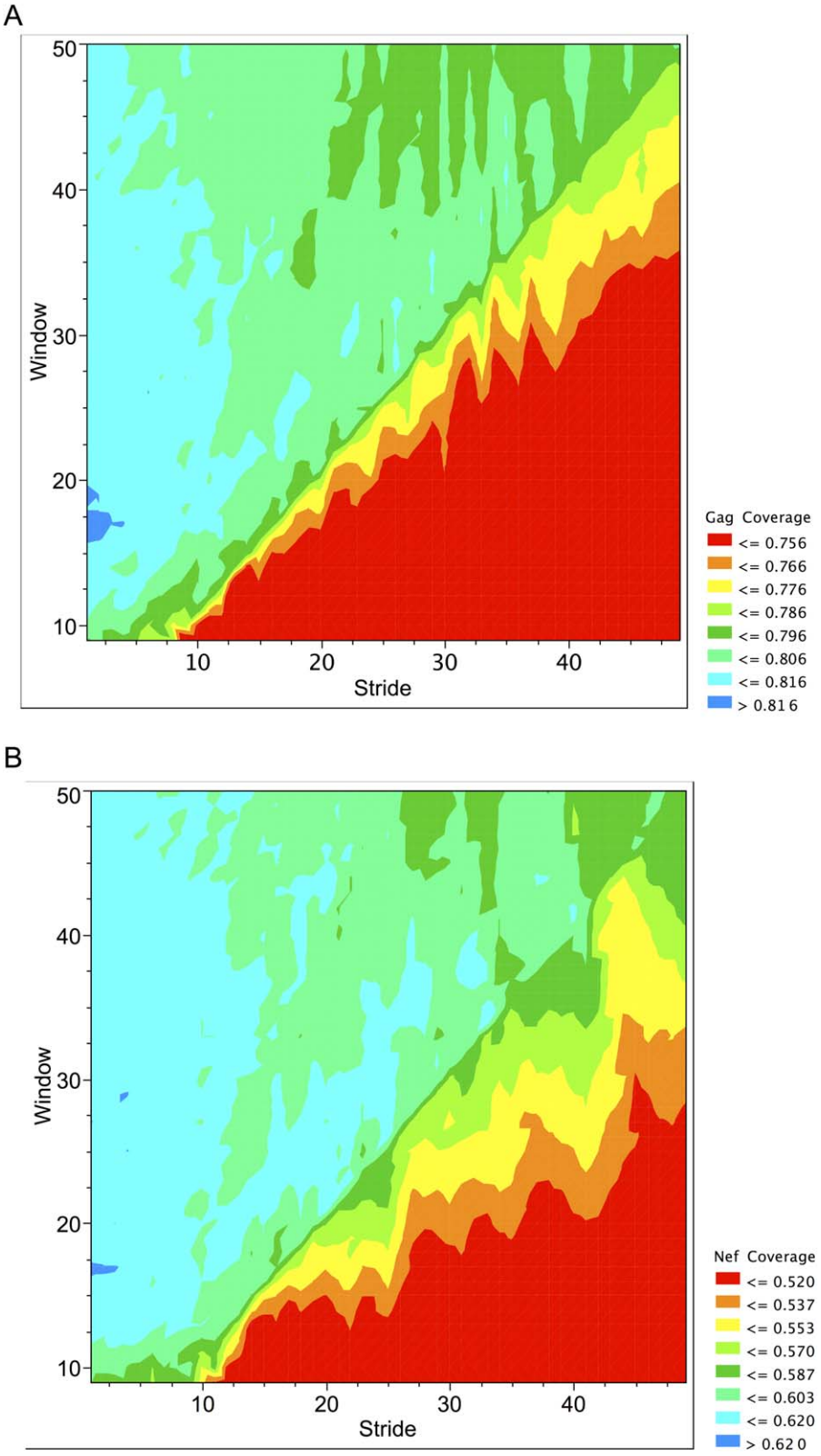

Figure 2. The Effects of Stride versus Window Length on the Measure of Coverage

In each graph a three-gene-length COT^{+} construct is evaluated for coverage. Cold (blue) colors indicate high levels of coverage, and hot (red) colors indicate low levels of coverage. The diagonal in the topography represents the transition from strides shorter than window length to strides longer than window length. The maximal coverage at three gene lengths occurs with a window size of 17 with a stride of 1 with no smoothing for both genes—where 82% of the 9mer area is captured for Gag (A) and 62% of the 9mer area is captured for Nef (B). It should be noted that in the area of window of 17 and a stride of 1 the surface is quite flat, and there are several pairs of parameters that give similar results.

doi:10.1371/journal.pcbi.0030075.g002

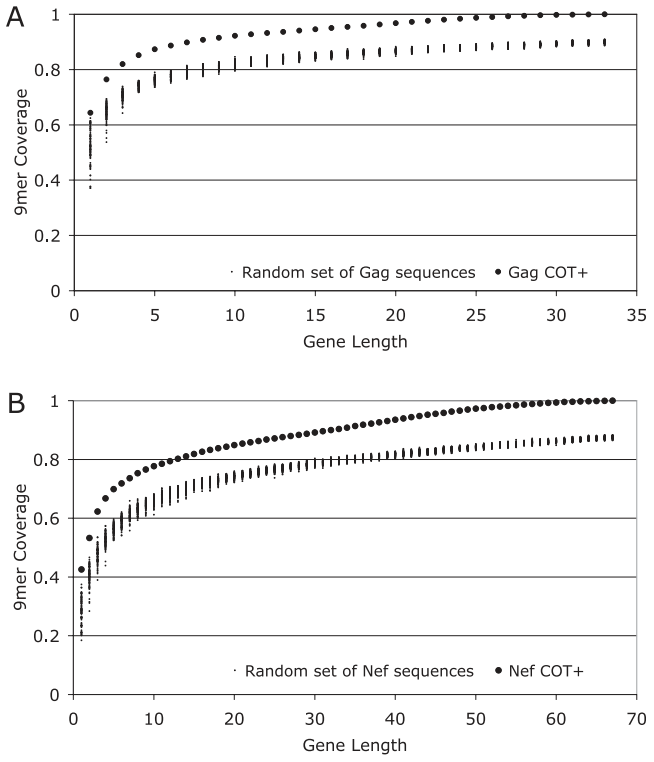

Figure 3. Coverage Comparison between COT^+ and 100 Randomly Sampled (without Replacement) Sets of Sequences of the Same Length

The comparison at the single gene length is for HIV-1 subtype B Gag and Nef, and measures the COT sequence against randomly sampled database sequences. The COT^+ captures all known variation in the training dataset at 33 gene lengths for Gag (A) and 67 gene lengths for Nef (B). Neither Gag nor Nef randomly sampled datasets will reach 100% coverage until 100% of the data is sampled.

doi:10.1371/journal.pcbi.0030075.g003

[ranging from 1 to *L*, the length of the protein]), we generate all sequence fragments from the sampled sequence by iteratively shifting the frame *S* residues at a time. We then compute the coverage for each sequence fragment not already present in the COT sequence, and append the sequence fragment to the COT string, compressing with possible overlap to yield a COT^+ molecule with the highest ratio of coverage per length. Specifically, fragments are chosen by their level of coverage and whether or not they have differences with respect to the COT sequences. The highest coverage fragments are chosen first, with subsequent fragments with lower coverage being chosen subsequently. This process is repeated until the sequence of desired length is derived. The length of the COT^+ sequence is arbitrarily chosen by taking into account plasmid size limitations for producing and delivering an antigen construct and the amount of variability that can be efficiently incorporated as the length is extended, which in turn depends on the variability found in circulating strains that have been sampled for a particular gene. We note that it is possible to arrange the order in which sequence fragments are added to COT^+ to maximize the overlap of consecutive fragments, thereby further compressing the antigen.

(5) The values of window size *F*, stride step *S*, and smoothing threshold *M* are varied to achieve maximum coverage (Figure 2A and 2B).

Comparison with Random Sequences

We compared our constructs of various lengths to randomly drawn sequences from the curated dataset of 169 sequences using the optimal values for *F* and *S*. We generated COT^+ for both Gag and Nef at ever-increasing unit protein lengths until we reached 100% coverage. For comparison, we concatenated randomly sampled protein sequences 100 times at ever-increasing unit protein lengths from both Gag and Nef and measured 9mer coverage across the same gene lengths (Figure 3A and 3B). We chose protein unit length for our comparison, but COT^+ can be derived for any partial unit protein length desired.

Cross-Validation

To ensure that we were not overestimating the coverage of our constructs due to the finite size of our dataset, we repeated our approach using 10-fold cross-validation. We partitioned the data into ten sets, and for each we estimated COT^+ from the remaining 90% of the data and then measured its coverage of the sequences in the chosen set. Thus, given that our assessment of coverage is on a set of sequences not seen in training, we yield an estimated lower bound on the coverage we would obtain for a larger population. We report this lower bound as a percentage of similarity to the estimated upper-bound COT^+, derived from training and testing on all 169 sequences for both Gag and Nef. This study is geared to understand the effect of sample size on the on the COT^+ estimation and to show that we are not overfitting the estimations.

Known Epitope Coverage

Although the list of known HIV-specific CD8 T-cell epitopes is far from complete [18], we sought to determine how well our 9mer coverage-based constructs identified known epitopes. To this end, we obtained all available HIV CTL epitopes from the Los Alamos National Laboratory (LANL) HIV immunology database [27] and counted the perfect matches to our constructs. Because many true epitopes are listed multiple times and larger peptides are reported frequently where the true epitope is embedded, we curated the database to remove any larger epitope that had a smaller embedded known epitope with the same supertype HLA response pattern, and removed any duplicates.

Results

We inferred COT sequences from databases of Gag and Nef protein sequences from HIV-1 subtype B from 169 independently infected individuals, and then added frequently observed variant 9mer peptides to create COT^+ sequences. The frequencies of unique 9-mer peptides are shown in Figure 1. We find that maximal coverage occurs when the window size, *F*, is 17, the stride length, *S*, is 1, and when smoothing *M* is 0 (Figure 2A and 2B). One possible reason for why an *S* value equal to 1 leads to the highest coverage is that it gives every amino acid in the sequences a chance to be in every possible position in a high-scoring peptide. Counter-intuitive to this is the observation that *S* values greater than 1 do not get penalized with big drops in 9mer coverage. We think the explanation for this observation has to do with the fact that even with *S* larger than 1, every amino acid in the

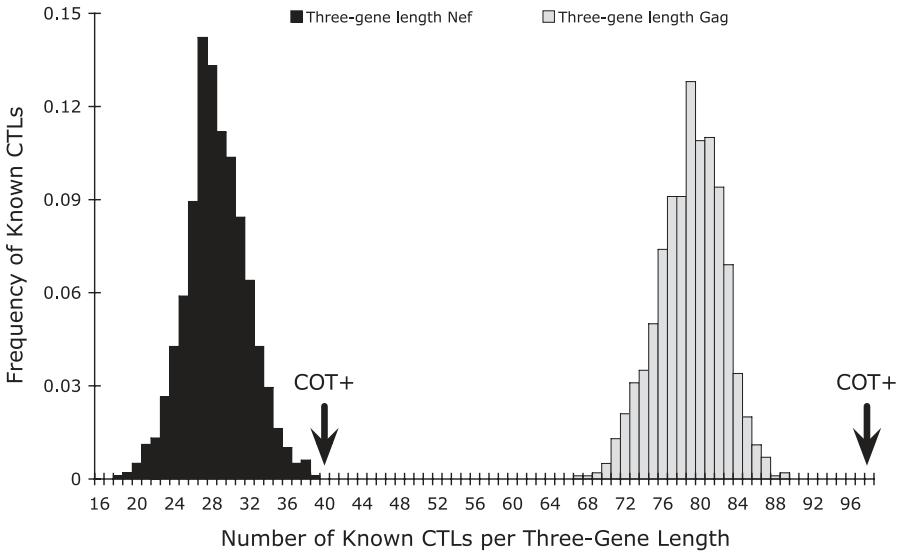

Figure 4. The Distribution of the Number of Known Epitopes in Three Randomly Chosen *Gag* (Right-Side Distribution) and *Nef* (Left-Side Distribution) Genes from the Los Alamos National Laboratory Database

The COT^+ sequence at three gene lengths for Gag has 98 out of 102 known CTL epitopes, and Nef has 40 out of 49 known CTL epitopes.
doi:10.1371/journal.pcbi.0030075.g004

sequences is still considered when building a construct. This is exemplified by the fact that the biggest drops in 9mer coverage come when S is larger than F, because it is in this parameter space that some amino acids have the probability of not being considered at all in the resulting construct.

Adding peptides to generate a three-gene-length COT^+ construct achieved 82% 9mer coverage for Gag and 62% for Nef, whereas an antigen constructed from several random concatenated database sequences [22] needed to achieve the same level of coverage required ten gene lengths for Gag and approximately 11 for Nef (Figure 3A and 3B). When COT^+ is compared with 100 constructs of the same length obtained by concatenating randomly selected sequences from the Los Alamos National Laboratory database [22], the COT^+ estimate had a higher level of coverage in every case (randomization test, $p < .01$) for both Gag and Nef. The flattening of the curves in Figure 3A and 3B suggests that after the COT^+ construct has grown past a few gene lengths, the benefit of adding more length is dramatically reduced. For example, the extension of the COT^+ construct from one to three gene lengths results in a 16% increase in coverage for Gag and a 13% increase in coverage for Nef. However, extending COT^+ from three to five gene lengths yields only 5% additional coverage for both Gag and Nef. The COT^+ sequence reaches 100% coverage at 33 gene lengths for Gag and 67 gene lengths for Nef, while the randomly sampled sets reach 100% coverage only after all 169 sequences are included. The latter observation is due to the fact that many of the mutations found in HIV are private (i.e., found only within the lineage infecting a particular person).

When applied to small datasets, our algorithm generates COT^+ constructs with high coverage. An extreme example is making a three-gene construct from just three genes in the training set. In this scenario, we can trivially achieve 100% coverage. The larger the training set, the lower the coverage in a three-gene-length vaccine construct. A 10-fold cross-validation study was therefore designed to determine the effects of sample size on our COT^+ constructs. Specifically, at three protein lengths, the cross-validated coverage of Gag and Nef are 96% and 93%, respectively. This suggests that for both proteins these inferences are generalizable across HIV-1 subtype B and that adding more sequence data into the training dataset would add very little to these estimations. That is to say, 10% of the original 169 sequences produce estimations of the COT^+ that are highly consistent with the estimations from the entire dataset, supporting the notion that there is a saturation effect and that adding sequences beyond the 169 will not give rise to better estimations.

Assessing the inclusion of functional CTL epitopes in our constructs is problematic. The majority of the known CTL epitopes were mapped using peptides derived from a limited number of HIV strains (e.g., laboratory-adapted strains and consensus sequences). The CTL database is also incomplete (e.g., a recent study that used a subset of autologous peptides from a single patient enabled recognition of 28% more epitopes in the virus than were previously reported [18]), and it is unclear whether characterized epitopes form an unbiased sample of naturally occurring antigenic peptides. It is also necessary that the epitope be presented in the proper context of adjacent amino acids for efficient immunoproteasome cleavage. We therefore assessed the overall size of the peptides needed to obtain maximal coverage of included 9mers. As shown in Figure 2A and 2B, maximal coverage of both the Gag and Nef datasets was obtained with a window size of 17 amino acids and a stride of one amino acid and no smoothing required (see Methods). Hence, we are able to construct immunogens that preserve much of the extended local amino acid environment of the epitope without sacrificing coverage. This enhances the likelihood that the desired peptide epitope will be properly cleaved by cellular proteases and presented efficiently on HLA molecules.

Next, we assessed the inclusion of known CTL epitopes in our constructs by comparing the number of known HIV-1 Nef and Gag epitopes [27] contained in the three-gene-length

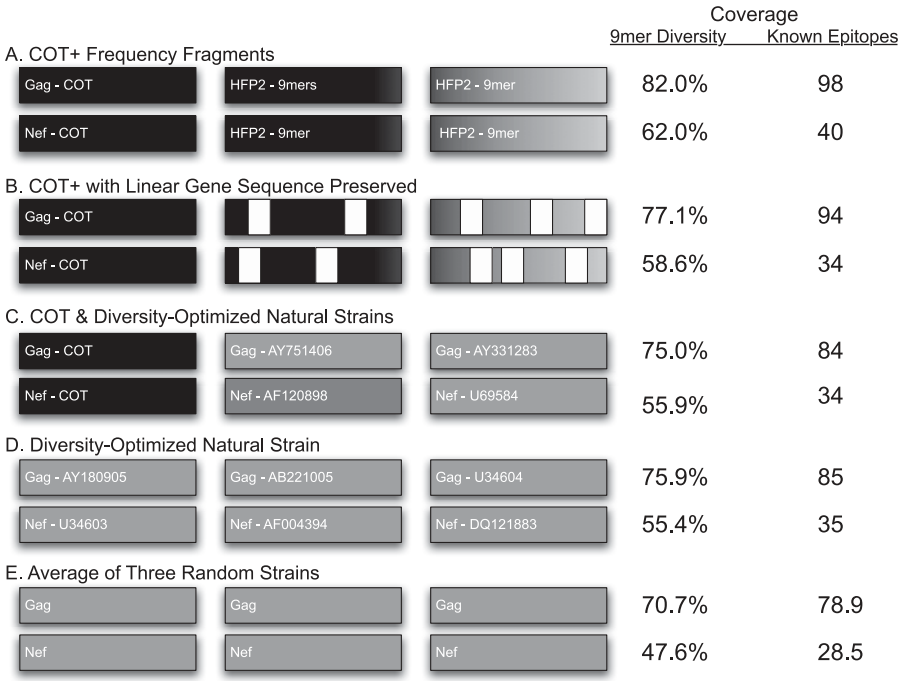

Figure 5. Possible Configurations for Vaccine Constructs

Each bar represents one unit-length gene. The fill intensity of each bar represents the density of unique peptides and known CTL epitopes. The coverage that each construct captures of the amino acid diversity of the dataset is shown on the right for both 9mers and epitopes.
(A) COT^+, composed of the estimated COT plus the appended high-frequency peptides (HFPs) composing the second and third gene lengths.
(B) COT plus HFPs placed into a gene collinear fashion on the second and third gene lengths.
(C) COT plus two NSs chosen to maximize 9mer coverage.
(D) All NSs of Gag and Nef sequences chosen such that 9mer coverage is maximized and for comparative reasons (E) is average coverage across all NSs. The GenBank IDs of the NSs are written inside each bar.
doi:10.1371/journal.pcbi.0030075.g005

COT^+ constructs to that of 1,000 combinations of three randomly selected database sequences (Figure 4). Sequences from the viral strains used to map these epitopes were excluded from the randomization study. Although our algorithm does not attempt to explicitly enrich for known CTL epitopes, the number of known epitopes in COT^+ is significantly higher than in a random three-gene construct ($p < 0.001$) for both Gag and Nef. This suggests that COT^+ provides a substantial boost in the number of epitopes shared between the immunogen and a random circulating database variant, and thus may have enhanced potential as an immunogen.

Discussion

COT^+ constructs provide a means to extensively compress epitope variation into an immunogen of minimal size. Much of the known variation of both the relatively conserved HIV-1 Gag gene and the quite variable Nef gene can be successfully compressed into COT^+ constructs of a few gene lengths. Little increases in variation coverage are noted, however, beyond three to four gene lengths. Coverage grows with length approximately in a $y = m\log(x) + b$ form where y is coverage and x is length of the construct. The difference between COT^+ construct of Gag and Nef can be broken down into these terms. The coverage intercept parameter b is higher for Gag constructs than for Nef simply because Gag is a more conserved protein than Nef. However, the parameter m is larger for Nef than it is for Gag because the benefits of 9mer compression on coverage are higher with constructs made from variable proteins.

Our COT^+ generation algorithm is a rapid, computationally efficient heuristic approximation, though it is not guaranteed to find the antigen that achieves maximal epitope coverage for a fixed length. More computationally intensive approaches, such as genetic algorithm searches or approximate solutions to the classic Traveling Salesman problem (see http://mathworld.wolfram.com/TravelingSalesmanProblem.html), could also be brought to bear on the problem of antigen design. Surprisingly, selecting the high-frequency 9mers alone and appending them to the COT sequence does poorly in terms of total coverage (unpublished data). This observation is due to the fact that many of the 9mers do not overlap, and therefore the fragments cannot be efficiently joined. By going back and selecting high-coverage peptide windows from the original data, we obtain better compression in the vaccine construct leading to higher coverage constructs for the same length.

It is a reasonable assumption that the retention of native protein structures might be advantageous in generating CTL epitopes, since epitopic peptides are generated in vivo by protein degradation within infected cells. Nef and Gag COT clearly adopt a native structure, as they retain biological activity (Rolland M, Jensen MA, Nickle DC, Learn GH, Heath L, et al., unpublished data). However, the extended COT^+ component of antigens generated in the manner proposed

here does not preserve a sequence that is necessarily collinear with the native gene over the second and third gene lengths (Figure 5A). Hence, we have also considered additional means of optimizing immunogen structures that also preserve native structure. First, we can assemble high-frequency variable elements in a pattern collinear with the native gene, with some segments redundant with COT to retain collinearity (Figure 5B). We can also use NS sequences in combination with the COT sequence to optimize coverage (Figure 5C). We can also do very well in generating coverage by exclusive use of NS sequences that maximize 9mer coverage (Figure 5D). Although it is not guaranteed, these additional constructs (Figure 5B–5D) should have biologically acceptable tertiary structures. The COT^{+} approach captures more of the 9mer distribution and more of the known CTL epitopes than any of the potential constructs presented here. Applying high-frequency peptides onto COT to create a collinear pattern provides the second highest level of diversity and epitope enrichment, but the use of COT plus two NSs is not beneficial relative to judicious choice of three NSs. Last, it should be noted that all of these methods substantially exceed the coverage afforded by the use of a single strain as a vaccine.

Immunodominance gives rise to a rank order of immune responses to specific epitopes [28], and the underlying biological mechanisms giving rise to these rank orders are poorly understood. The antigen designs we report here do not take immunodominance into account. One can argue that the combination of epitopes we have derived could elicit an immunodominant response that does not reflect what is found in circulating HIV strains and hence could be a poor choice for vaccine design. However, since the strings of peptides in our immunogen design are captured by their frequency in the circulating viral population, we surmise that these antigens have epitopes that are shared across many potential challenge strains and could thus lead to potentially broad immune response. However, immunodominance rank order patterns can be partially illuminated by expressing epitopes from different vaccine vectors [29–31]. By vaccinating with different combinations of vectors encoding a single or more antigens, they found that using separate vectors elicited broader $CD8^{+}$ T cell responses. Because COT^{+} is directed towards capturing high-frequency fragments from a variable protein, it is well-suited to being expressed as segments on separate vectors. The COT^{+} algorithm can be generalized to produce *sets* of immunogens that can take advantage of this phenomenon.

COT^{+} constructs are able to capture significantly more known epitopes and potential antigen variability than much longer constructs composed by combining circulating strains. Considering the substantial expense and difficulty involved in production and testing of candidate vaccines, careful crafting of potential antigens using computational methods, including that shown here, may be beneficial. Furthermore, this approach is applicable not only to HIV vaccine design, but to the design of vaccines targeting any pathogen capable of rapid escape from immune recognition.

Acknowledgments

We thank Vladimir Jojic and Carl Kadie for assisting NJ and DH in developing the algorithms for epitome optimization. Based on Vladimir Jojic's initial approach, Carl Kadie has implemented and tested an improved epitome optimization algorithm, which we used for comparisons during the development of the COT^{+} algorithm. The COT^{+} sequences described here are available upon request.

Author contributions. DCN, MR, DH, JIM, and NJ conceived and designed the experiments. DCN, WD, and MS performed the experiments. DCN, MR, WD, and JIM analyzed the data. DCN, MAJ, SLKP, WD, MS, DH, and NJ contributed reagents/materials/analysis tools. DCN, MR, and JIM wrote the paper.

Funding. This work was supported by a gift from the Boeing Corporation to JIM.

Competing interests. The authors have declared that no competing interests exist.

References

1. Mullins JI, Nickle DC, Heath L, Rodrigo AG, Learn GH (2004) Immunogen sequence: The fourth tier of AIDS vaccine design. Expert Rev Vaccines 3 (Supplement 1): S151–S159.
2. Gao F, Korber BT, Weaver E, Liao HX, Hahn BH, et al. (2004) Centralized immunogens as a vaccine strategy to overcome HIV-1 diversity. Expert Rev Vaccines 3: S161–S168.
3. Palker TJ, Matthews TJ, Langlois AJ, Tanner ME, Martin ME, et al. (1989) Polyvalent human immunodeficiency virus synthetic immunogen comprised of envelope gp120 T-helper cell sites and B-cell neutralization epitopes. J Immunol 142: 3612–3619.
4. De Groot AS, Marcon L, Bishop EA, Rivera D, Kutzler M, et al. (2005) HIV vaccine development by computer assisted design: The GAIA vaccine. Vaccine 23: 2136–2148.
5. Thomson SA, Jaramillo AB, Shoobridge M, Dunstan KJ, Everett B, et al. (2005) Development of a synthetic consensus sequence scrambled antigen HIV-1 vaccine designed for global use. Vaccine 23: 4647–4657.
6. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, et al. (2002) Diversity considerations in HIV-1 vaccine selection. Science 296: 2354–2360.
7. Gao F, Weaver EA, Lu Z, Li Y, Liao HX, et al. (2005) Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein. J Virol 79: 1154–1163.
8. Learn G, Mullins JI (2000) Inferring an ancestral Asian HIV-1 subtype E *env* sequence for use as a vaccine immunogen. In Proceedings of the Meeting of the AIDS Panels for the U.S.-Japan Cooperative Medical Science Program. 12th Joint Scientific Meeting of the U.S.-Japan Cooperative Medical Science Program; 22–24 March 2000; Santa Fe, New Mexico, United States.
9. Learn G, Mullins JI (2000) The use of an inferred epidemic ancestral sequence as a vaccine immunogen [abstract]. 7th Annual International Discussion Meeting on HIV Dynamics and Evolution; 28–30 April 2000; Seattle, Washington, United States. Available: http://ubik.microbiol.washington.edu/Seattle2000/abstracts/abstract30.html. Accessed 28 March 2007.
10. Mullins JI, Rodrigo AG, Learn GH, Doria-Rose N, Haigwood N (2002) How should HIV strains be chosen for inclusion in vaccines [abstract]? Keystone Symposium, HIV-1 Protection and Control by Vaccination, 5–11 April 2002; Keystone, Colorado, United States.
11. Stewart CB (1995) Molecular evolution. Active ancestral molecules. Nature 374: 12–13.
12. Doria-Rose N, Learn GH, Rodrigo AG, Nickle DC, Li F, et al. (2005) Human immunodeficiency virus type 1 subtype B ancestral envelope protein is functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating subtype B envelope. J Virol 79: 11214–11224.
13. Nickle DC, Jensen MA, Gottlieb GS, Shriner D, Learn GH, et al. (2003) Consensus and ancestral state HIV vaccines. Science 299: 1515–1518.
14. Schmitz JE, Kuroda MJ, Santra S, Sasseville VG, Simon MA, et al. (1999) Control of viremia in simian immunodeficiency virus infection by $CD8^{+}$ lymphocytes. Science 283: 857–860.
15. Jin X, Bauer DE, Tuttleton SE, Lewin S, Gettie A, et al. (1999) Dramatic rise in plasma viremia after CD8(+) T cell depletion in simian immunodeficiency virus-infected macaques. J Exp Med 189: 991–998.
16. Koup RA, Safrit JT, Cao Y, Andrews CA, McLeod G, et al. (1994) Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. J Virol 68: 4650–4655.
17. Borrow P, Lewick H, Hahn BH, Shaw GM, Oldstone MBA (1994) Virus-specific $CD8^{+}$ cytotoxic T-lymphocyte activity associated with control of viremia in primary HIV-1 infection. J Virol 68: 6103–6110.
18. Liu Y, McNevin J, Cao J, Zhao H, Genowati I, et al. (2006) Selection on the human immunodeficiency virus type 1 proteome following primary infection. J Virol 80: 9519–9529.
19. McCutchan FE (2006) Global epidemiology of HIV. J Med Virol 78 (Supplement 1): S7–S12.
20. Jojic N, Jojic V, Frey B, Meek C, Heckerman D (2005) Using epitomes to model genetic diversity: Rational design of HIV vaccine cocktails. In Proceedings of Neural Information Processing Systems (NIPS). 19th Annual

Conference on Neural Information Processing Systems; 5–10 December 2005; Whistler, British Columbia, Canada. Available: http://books.nips.cc/papers/files/nips18/NIPS2005_0759.pdf. Accessed 28 March 2007.
21. Jojic N, Frey B, Kannan A (2003) Epitomic analysis of appearance and shape. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Volume 2. Washington (D. C.): IEEE Computer Society. pp. 34–43.
22. Leitner T, Foley B, Hahn B, Marx P, McCutchan F, et al., editors (2006) HIV Sequence Compendium 2005. Los Alamos (New Mexico): Theoretical Biology and Biophysics Group, Los Alamos National Laboratory. 648 + viii p.
23. Hillis DM, Moritz C, Mable BK, editors (1996) Molecular systematics. 2nd edition Sunderland (Massachusetts): Sinauer Associates. 655 p.
24. Swofford DL (1999) PAUP* 4.0: Phylogenetic Analysis Using Parsimony (*and Other Methods). 4.0b2a ed. Sunderland (Massachusetts): Sinauer Associates.
25. Jones NA, Wei X, Flower DR, Wong M, Michor F, et al. (2004) Determinants of human immunodeficiency virus type 1 escape from the primary $CD8^+$ cytotoxic T lymphocyte response. J Exp Med 200: 1243–1256.
26. Allen TM, Altfeld M, Geer SC, Kalife ET, Moore C, et al. (2005) Selective escape from $CD8^+$ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. J Virol 79: 13239–13249.
27. Korber BTM, Brander C, Haynes BF, Koup R, Moore JP, et al. (2005) HIV Molecular Immunology 2005. Los Alamos (New Mexico): Los Alamos National Laboratory, Theoretical Biology and Biophysics. 1,158 p.
28. Yewdell JW, Bennink JR (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. Annu Rev Immunol 17: 51–88.
29. Bartholdy C, Stryhn A, Christensen JP, Thomsen AR (2004) Single-epitope DNA vaccination prevents exhaustion and facilitates a broad antiviral $CD8^+$ T cell response during chronic viral infection. J Immunol 173: 6284–6293.
30. Chen W, Anton LC, Bennink JR, Yewdell JW (2000) Dissecting the multifactorial causes of immunodominance in class I-restricted T cell responses to viruses. Immunity 12: 83–93.
31. Egan MA, Megati S, Roopchand V, Garcia-Hand D, Luckay A, et al. (2005) Rational design of a plasmid DNA vaccine capable of eliciting cell-mediated immune responses to multiple HIV antigens in mice. Vaccine 24: 4510–4523.