JVI Accepts, published online ahead of print on 30 May 2007
J. Virol. doi:10.1128/JVI.02683-06

Reconstruction and function of Ancestral Center-Of-Tree (COT) HIV-I Proteins

Running title: COT protein derivation and function

Morgane Rolland[1], Mark A. Jensen[1§], David C. Nickle[1], Jian Yan[2], Gerald H. Learn[1], Laura Heath[1], David Weiner[2] and James I. Mullins[1]*

[1]Department of Microbiology, University of Washington, Seattle, WA 98195-8070

[2]School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

§ Present address: Dept. of Epidemiology, College of Public Health, and Dept. of Genetics, Franklin College of Arts & Sciences, University of Georgia, Athens, GA 30602

* To whom correspondence should be addressed. Telephone: 206-732-6163; Fax: 206-732-6167; email: jmullins@u.washington.edu

Abstract: 248 words

Text (Introduction-Discussion): 5,201 words

Text (complete, with abstract): 5,449 words

Abstract

HIV-1's extensive diversity and its capacity to mutate and escape host immune responses are major challenges for AIDS vaccine development. Ancestral sequences, that minimize the genetic distance to circulating strains, provide an opportunity to design immunogens with the potential to elicit broad recognition of HIV epitopes. We developed a phylogenetics-informed algorithm to reconstruct ancestral HIV sequences, called Center-Of-the-Tree or COT. COT sequences have potentially significant benefits over isolate-based strategies as they minimize the evolutionary distances to circulating strains. COT sequences are designed to surmount the potential pitfalls stemming from sampling bias with the consensus method and outlier bias with the most recent common ancestor approach. We computationally derived COT sequences from circulating HIV-I subtype B sequences for the genes encoding the major viral structural proteins (Gag) and two regulatory proteins, Tat and Nef. COT genes were synthesized *de novo*, expressed in mammalian cells and the proteins characterized. COT Gag was shown to generate virus-like particles, while COT Tat transactivated gene expression from the HIV-1 LTR, and COT Nef mediated downregulation of cell surface MHC-I. Thus, retrodicted ancestral COT proteins can retain the biological function of extant HIV-I proteins. Additionally, COT proteins were immunogenic, as they elicited antigen specific CTL responses in mice. This data supports the utility of the COT approach to create novel and biologically active ancestral proteins as a starting point for studies of structure, function and biological fitness

of highly variable genes, as well as for the rational design of globally relevant vaccine candidates.

Introduction

Pauling and Zuckerkandl (32) suggested more than forty years ago that it should be possible to "synthesize…presumed components of extinct organisms…and study the physico-chemical properties of these molecules". However, it has only recently become relatively straightforward to perform ancestral state reconstructions and then evaluate the resurrected molecules experimentally. Reconstructed ancestral sequences have been used to analyze evolutionary pathways, correlate molecular changes with geographic or paleontological events (4, 6) identify functional divergences, for example proteins involved in vision and inflammation (7, 30), and investigate ancient features of life on earth (9). Recently, concomitant with the unprecedented increase of genome sequence data, ancestral state reconstructions have been expanding to encompass genome-wide characters, from the mega-base scale reconstruction of an ancestral mammalian genome (5) to the 10 kb genome of HIV (14).

The plasticity of the HIV genome allows generation of enormous numbers of viable mutants, resulting in circulating sequences that can differ by more than 30% in the maximally variable *env* gene. Since the genetic diversity of HIV-1 will continue to increase over the many years required for a vaccine to be developed, clinically tested, manufactured and deployed, it is crucial to focus on vaccine sequence designs that can mitigate the effect of this diversity, and to develop a substantially greater understanding of the structure-function relationships within the viral proteome to enhance development of antiretroviral therapeutics. Specifically, the efforts described here are directed toward designing HIV vaccine sequences that would embody as many common features of all circulating strains as possible while retaining functionality.

Several methods can be implemented to minimize the genetic distance to all known extant viruses. Consensus sequences (CON) correspond to the most frequent

amino acid or nucleotide at each site within a gene sequence alignment. However, a consensus sequence is subject to sampling bias and may not retain co-variable sites, since it does not consider evolutionary history (20). Our approaches seek to design antigens encompassing conserved structural features through phylogenetics-informed algorithms. HIV-1 phylogenies approximate a star-like shape in which most circulating strains have diverged approximately equally from a central point: the Most Recent Common Ancestor (ANC). A prototypical strain that embodies the ANC will be more genetically similar to all circulating strains than any one strain chosen at random (if star phylogenies are accurate for HIV evolutionary history). The ANC is also expected to conserve the amino acid co-variation required for proper protein folding (because the ancestor sequence is an estimate of a sequence that actually existed in viable viruses), thus improving the likelihood that the protein will function. However, the presence of outlier sequences in the phylogeny can bias the amino acid sequence against a large number of the intended circulating viral targets (10).

To ensure the robustness of ancestral reconstructions against such a long-branch-bias, we developed a novel algorithm, elaborated here, to identify the point on the phylogenetic tree that represents the minimum of a metric of evolutionary distance. This position, called the “Center Of Tree” or COT, minimizes the evolutionary distance to all sampled circulating strains, while still residing on an evolutionary path, to better capture the biological properties of circulating viruses (19). We have applied this novel phylogenetics-based design strategy to computationally derive genes encoding HIV-1 proteins that have both important biological functions and are frequent immunologic targets: Gag, Tat and Nef. COT antigens were synthesized *de novo* and then characterized experimentally to evaluate their functionality and their immunogenicity.

Materials and Methods

Algorithms. Below we describe a general method for identifying a position at a node or on a branch of a phylogenetic tree having completely specified branch lengths. This position, called the "center of tree" or COT, is a point at which a specified function *F* of the lengths from any point to all tips of the tree is minimized. Suppose a tree **T** has *n* tips or leaves, labeled a_1, a_2,..., a_n, and *p* is a point on a branch or is a node of the tree. Let l_i be the distance along the tree branches from *p* to a_i. Then a COT of **T** for the function *F*, is defined as a point $\hat{p}$ satisfying the following relationship:

$$F(\hat{p}:\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_n) \le F(p:l_1,l_2,\cdots,l_n), \text{for all points } p,$$

where the notation $F(p:l_1,l_2,\cdots,l_n)$ highlights the fact that the distances l_i depend on the point *p*.

In this description, the form of *F* is general; specific choices for *F* can be made based on the intended application. The general algorithm below is applicable for most useful continuous *F*s. For a given choice of *F*, a different algorithm may be more efficient. We describe such an algorithm to find COT when *F* is the mean of squares (MS) of the *l*s. Depending upon *F*, it is possible that more than one COT will exist for a given tree, but for many reasonable choices of *F* the COT will be unique.

General Algorithm. We first show that for a certain large class of functions $F: R^n \to R$ (the function does not have an infinite number of local minima, i.e. COT points), there is a finite number of points along the tree, each corresponding to a possible COT, and that we can enumerate these points and determine which are in fact a COT.

For an unrooted tree T of *n* tips, there are $u \le 2n-2$ nodes, counting tips and internal branches, and $w \le 2n-3$ branches, including internal and external branches (*u* and *w* are less than their maxima when polytomies (nodes with more than three

branches) exist in the tree). For each node $q_j, 1 \le j \le u$, we can calculate $c_j = F(q_j : l_1, l_2, \cdots, l_n)$. Each q_j is a candidate COT.

We then determine the candidate COT point for each branch b_k, enumerated $b_k, 1 \le k \le w$. Note that each branch, say branch *k*, is flanked by two nodes called R_k and L_k. Let the branch length of b_k be *l*. Now the tree is divided into right and left parts, so that if the tree had a root within branch *k*, the tips a_1, a_2,..., a_n would be divided into two groups: those descended from R_k and those descended from L_k. Suppose there are *s* right tips and *t* left tips. Let the distances from the right tips to R_k be written $\rho_1, \cdots \rho_s$, and the distances from the left tips to L_k be written $\lambda_1, \cdots \lambda_t$. Now, let a point *p* lying on branch b_k be a distance *x* from the right node R_k. Then the distance from *p* to L_k is $l - x$.

Then for branch b_k and *p* defined along it as described above, we have

$$F(p : l_1, \cdots, l_n) = F(p : \rho_1 + x, \cdots, \rho_s + x, \lambda_1 + l - x, \cdots, \lambda_t + l - x) = \tilde{F}_{b_k}(x), 0 < x < l .$$

In other words, on any branch *k* of T we can completely express the function *F* of *n* distances for every point *p* along that branch as a function of a single variable *x*. By our assumption that *F* has a finite number of extreme points, the functions $\tilde{F}_{b_k}(x)$ have a finite number of minima for *x* between 0 and *l*. Because *F* is continuous, those minima can be found by standard numerical methods, and each minimum $\hat{x}$ is associated with a point $p_{\hat{x}}$ as described above. Suppose there are *v* such points over all *w* branches, then we can write $d_i = F(p_i : l_1, \cdots, l_n) = \tilde{F}_{b(p_i)}(\hat{x}_i)$, for $1 \le i \le v$, where $b(p_i)$ is the branch associated with point p_i (not necessarily the *i*th branch). Then each p_i is a candidate COT, since if p_i is to minimize *F* among all points on the tree, it must at least minimize *F* on those points comprising the branch on which *p* resides. Since the nodes and

branches contain all points on the tree, we have enumerated all possible COTs in the q_j and the p_i.

Therefore, any and all points p $p \in \{q_1, \cdots, q_u, p_1, \cdots, p_v\}$ that satisfy $F(p : l_1, \cdots, l_n) = \min\{c_1, \cdots, c_u, d_1, \cdots, d_v\}$ are the only COTs for tree T given function *F*.

Phylogenetic trees can be expressed in computer programs as data structures that can be efficiently traversed by recursive routines that isolate each node and branch individually and systematically. The decomposition above formally describes the tasks to be performed upon consideration of each node and branch. While the algorithm is executed, the points and function values are stored, and the final determination of a COT sequence is accomplished by identifying the minima of the list of values and their associated points after the tree data structure has been completely traversed.

Algorithm to find points minimizing the mean squared distance from the points to tips. For the COT used in this study, we let $F(p : l_1, \cdots, l_n) = \frac{1}{n}\sum_{i=1}^{n} l_i^2$, the mean of the squared distances from the tips to point *p*. The COT obtained by minimizing this function essentially balances the average length of the branches on either side of point *p*, and thereby provides a point that yields a single reconstructed sequence with the maximum amount of sequence similarity to all the tips, given the evolutionary constraints of nucleotide change along the tree branches. As in the general algorithm, we decompose the tree into nodes and branches, enumerate all possible COT sequences, and calculate *F* for each possibility. The point with the minimum *F* is the COT. We can also express the function *F* in terms of quantities that can be efficiently calculated as the tree is traversed recursively; this allows the algorithm to accumulate the quantities c_i and d_i. Below, we describe the method of identifying possible COT points and calculating c_i and

d_i based on these quantities; then we describe the recursion equations for these quantities that can be used in the tree-traversal algorithm.

***Nodes*:** Consider each node q_i as a temporary root of the tree, and suppose q_i has k descendant branches, each of which defines a subtree with t_m tips, $1 \le m \le k$. Then F can be written

$$c_i = \frac{1}{n}\sum_{j=1}^{n} l_j^2$$
$$= \gamma_1\left(\frac{1}{t_1}\sum_{j=1}^{t_1}(l_j^{(1)})^2\right) + \cdots + \gamma_k\left(\frac{1}{t_k}\sum_{j=1}^{t_k}(l_j^{(k)})^2\right)$$
$$= \gamma_1 MS_1 + \cdots + \gamma_k MS_k,$$

where $\gamma_m = \frac{t_m}{n}$, the proportion of the n tips in the m^{th} subtree, and $(l_j^{(m)}), 1 \le j \le t_m$, is the distance from q_i to each of the t_m tips.

Each MS_m is therefore the mean of squared distances to the n tips of the entire tree associated with subtree m from node q_i, considering node q_i as the root, and each γ_m is the proportion of the n tips of the entire tree associated with subtree m.

***Branches*:** With this function F, there exists at most one possible COT on any branch. Consider a branch of length l with left and right nodes as described in the general algorithm above, and consider a point p within the branch. Let M_L be the simple average of distances from point p to the left tips, and M_R be the average of distances from p to the right tips. Suppose that there are t left tips and s right tips, and let $\gamma = \frac{t}{n}$. Now, define α as follows:

$$\alpha = \frac{(1-\gamma)M_R - \gamma M_L}{l} + 1 - \gamma .$$

There is a possible COT within the branch if $0 \le \alpha \le 1$, and it is the distance αl from the *left* node along the branch. If there is such a point, then the value of *F* at that point, d_i, can be written as

$$d_i = \gamma(1-\gamma)(M_L + M_R + l) - \gamma(M_L^2 - MS_L) - (1-\gamma)(M_R^2 - MS_R),$$

where *MS* is the mean of summed squared distances from the left or right nodes to their descendant tips, as indicated.

Finally, the COT is the point associated with the smallest value among the c_i and d_i.

Recursions to calculate M and MS. Suppose node *q* has *k* descendant nodes. Each of the *k* nodes is connected to *q* by a branch of length l_i, and each is the root of a subtree having s_i tips, $1 \le i \le k$. Suppose for each subtree, the mean distance M_i and the mean squared distances MS_i from node to tips have been calculated. Then the mean distance M_q and mean squared distance MS_q from *q* to all $s = s_1 + \cdots + s_k$ descendant tips are given by:

$$M_q = \gamma_1(M_1 + l_1) + \cdots + \gamma_k(M_k + l_k), \text{ and}$$
$$MS_q = \gamma_1(MS_1 + 2l_1M_1 + l_1^2) + \cdots + \gamma_k(MS_k + 2l_kM_k + l_k^2),$$
$$\text{where } \gamma_i = \frac{s_i}{s} \text{ for } 1 \le i \le k.$$

These quantities can thus be built up as the tree is recursively traversed, and can be used in the calculations described above.

The PERL scripts implementing the above algorithm have been combined into a web-based tool that is available at http://indra.mullins.microbiol.washington.edu/cgi-bin/COT/cot.cgi.

Application to data. HIV-1 subtype B nucleotide sequences and 3 subtype D sequences (used as an outgroup) were used to create data sets for Gag (39

sequences), the first exon of Tat (40 sequences) and Nef (37 sequences). Consensus sequences were derived from each data set without the outgroup {Maddison, 2001 #13641}. ANC were estimated from Maximum Likelihood (ML) trees generated with PAUP* (26) using the outgroup. The outgroup was then removed and the tree was imported back into PAUP to estimate the COT sequence. The three computationally derived nucleotide sequences, CON, ANC and COT, were added to the data sets and new ML trees were generated. For each, we estimated the average genetic distance between the derived sequence and the sequences used to generate the phylogenies.

Construction and *in vitro* expression of COT *gag, tat* and *nef* genes. COT gene nucleotide sequences were optimized for expression in human cells by changing codon usage to that of highly expressed human genes and by reducing the free energy to improve stability and translation efficiency of the transcripts (28). The optimized COT genes were synthesized by Blue Heron Corp. (Bothell, WA) and subcloned into pcDNA3.1(-) (Invitrogen) at Xba I and Not I sites.

Human embryonic kidney 293T cells were transfected with the COT *gag, tat* and *nef* gene constructs by the calcium phosphate coprecipitation method (3). Briefly, 3×10^5 293T cells were seeded per well in a 6-well plate and transfected the next day with 4 μg of plasmid DNA. 48 hours post-transfection, cells were lysed on ice with 0.5% NP-40 lysis buffer supplemented with protease inhibitors (leupeptin 2 μg/ml, aprotinin 2 μg/ml, 1 mM orthovanadate and 1 mM phenylmethylsulfonyl fluoride (PMSF)). Lysates were separated by electrophoresis on denaturing sodium dodecyl sulfate polyacrylamide (SDS-Page) gels containing 12 to 18% acrylamide, depending on the protein size. Proteins were transferred on a polyvinylidene difluoride PVDF membrane (Immobilon-P, Millipore) according to the Towbin transfer protocol (27). COT Gag and Nef proteins were detected with a 1:2000 dilution of HIV-1 p24-specific (#4121, NIAID,

http://www.aidsreagent.org/ecommerce) and of HIV-1 Nef- specific (#3689, NIAID) monoclonal antibodies, respectively, followed by a 1:4000 dilution of horseradish peroxidase (HRP)-labeled anti-mouse IgG. COT Tat was detected using a rabbit anti-serum (#1974, NIAID) and protein-bound antigens were detected with anti-rabbit HRP conjugate. Reactive protein bands were visualized by chemiluminescence using the ECLPlus Western blotting reagent (Amersham).

Virus-like Particle Production. Virus-like particles were obtained from 293T cell culture supernatants according to standard protocols (22). Briefly, 48 hours after transfection of 293T cells with COT Gag expression vectors, culture supernatants were clarified at 3,000 rpm at 4°C for 15 min and filtered through a 0.2 µm pore size filter. The filtrate was layered on top of a 20% sucrose cushion and spun at 27,000 rpm at 4°C for 1hr in a Beckman ultracentrifuge using an SW28 rotor. The pelleted virus-like particles were suspended in TNE buffer (10 mM Tris, 100 mM NaCl, 1 mM EDTA, pH 7.9).

Trypsin digestion assay. The virus-like particle fractions were incubated for 30 min at 37°C in the presence of trypsin (2 µg/ml) with or without triton X-100. Alternatively, a protease inhibitor cocktail (30 µM aprotinin, 435 µM leupeptin, 1 mM PMSF) was added to the virus-like particle fractions prior to the incubation with trypsin (200 µg/ml) with or without triton X-100 (1%). Reactions were stopped by adding Laemmli sample loading buffer (42 mM Tris-HCl, pH 6.8, 1.7%(w/v) SDS, 8.25% (v/v) glycerol, 0.6 M ß-mercaptoethanol) and heating at 100°C for 3 min. COT Gag was detected following the Western blot protocol described above.

HIV-1 LTR activity in the absence or presence of COT Tat. 1-2 x 10^5 293T cells were seeded in 6-well plates and transfected by the calcium phosphate coprecipitation technique (3) with various amounts of plasmids, but the total amount of DNA was kept constant (2 µg) by addition of the empty plasmid vector pcDNA3.1(-) DNA. Forty-eight hours post-transfection, cells were stained with X-Gal to detect β-galactosidase

expression and blue cells were scored. Basal transcription of pHIVlacZ (#151, NIAID) was determined with 200 and 400 ng of plasmid DNA in at least 3 different transfections. Tat activation of transcription was determined with 200 and 400 ng of pHIVlacZ cotransfected with 200, 400 or 800 ng of COT Tat exon 1 in at least 3 experiments.

MHC-I downregulation assay. 10^5 293T cells were seeded in 6-well plates and transfected by the calcium phosphate coprecipitation technique with 2 μg of pcDNA3.1-COTnef, pcDNA3.1nef-NL4-3 or pcDNA3.1nef-mock (defective reading frame with inactivating mutations at the 5' end of *nef*, (23)). Forty-eight hours post-transfection, the cell surface expression levels of the HLA-1 allele were analyzed with anti-HLA-ABC antigen/PE by flow cytometry as described (23).

Immunization of mice. The plasmids encoding COT proteins were digested with BamHI and NotI, the COT inserts were cloned in pVAX (invitrogen, Carlsbad, CA) to generate the constructs pVAX-COTGag, pVAX-COTTat and pVAX-COTNef. Immunizations were also done using a mock pVAX plasmid and pGag02CAM, which encodes a primary HIV-1 clade B Gag.

All immunizations were carried out with groups of three 5- to 6-week old female BALB/c mice. DNA vaccines were administered at a dose of 100 μg at days 0, 14 and 28. Immunizations were performed under anesthesia by injection into the anterior tibial muscle in the hind legs. Mice were sacrificed at day 35. The animals were cared for according to the regulations and guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

IFN-γ ELISpot. One week after the last immunization, splenocytes were isolated and red blood cells were lyzed by suspending in 2 ml Red Blood Cell Lysis Buffer/spleen for 5 min. The cells were then washed in PBS and resuspended in RPMI 1640 medium with 10% FBS. Cells were counted and prepared for analysis.

ELISpot assays were performed using High-Protein Binding IP 96 well Multiscreen™ plates coated with MAb to mouse IFN-γ. Responses were mapped using HIV overlapping peptide libraries corresponding to consensus sequences from subtypes A and B and group M. Briefly, 15-mer overlapping peptide library pools of either Gag, Tat or Nef were added to $2x10^5$ cells per well and incubated for 24h at 37°C in a 5% CO2 incubator. All tests were performed in triplicate. After addition of the detection antibody, color development was followed until spots were visible, and plates were air-dried. Wells were imaged and spots were counted by an automated ELISpot Reader (CTL Analyzers, Cleveland, OH) using the ImmunoSpot® software and analyzed as described. The average number of spot forming cells (SFC) was adjusted to $1x10^6$ splenocytes for data plotting. For each immunization experiment, pVAX was used as an internal control and the average number of SFC obtained with mock pVAX were subtracted from the numbers obtained with the COT plasmids.

Results

Identification of the "Center Of Tree" (COT). We have previously shown through simulations that computationally derived sequences do well at minimizing evolutionary distances to circulating HIV-1 strains (18), with the COT (see Methods for COT derivation procedures) and CON typically better than the ANC; A comparative study of the variation among HIV-1 *envelope* (*env*) gene sequences reconstructed with different methods was published elsewhere {Ross, 2006 #17261}. In our analysis, the CON and COT nucleotide sequences had similar average divergences from the sequence set (Table 1). The positions of ANC, COT, and CON in the inferred phylogeny of the *gag* sequence set is shown in Figure 1. We have found that COT proteins preserve more B-

cell and CTL epitopes than contemporary isolates {Frahm, Submitted #21185}, thus, we analyzed the AA differences between the COT proteins experimentally studied and central AA sequences corresponding to translated CON, ANC and COT nucleotide sequences reconstructions (Table 2). Before translation to protein sequences, the COT, ANC and CON nucleotide sequences were derived from alignments of circulating sequences (one sequence per individual) corresponding to 1) database sequences available in 2002, and 2) current database sequences, i.e., those available in 2002 and subsequently. In agreement with the genetic distances, experimental COT proteins showed the most AA differences when compared to ANC sequences (Table 2). Sixteen positions (out of 500) were found to vary in Gag, 5 (of 72) in Tat and 19 (of 206) in Nef. This highlights the key importance of the sequence dataset for accurate reconstructions; in retrospect, we consider that our COT sequences should have been generated based on larger sequence collections. To assess the biological significance of those changes, we analyzed whether mutations occurred at critical functional sites or within epitopes. Although we did not identify any changes known to abrogate the protein's functionality, it is remarkable that only 5 of the variable positions did not correspond to any known epitope; three in Gag, none in Tat, two in Nef. There was no specific indication whether one or the other variant found at one position was able to confer escape from the CTL immune response.

Humanization of COT sequences. HIV-1 subtype B COT *gag*, *tat* and *nef* genes were synthesized after optimization to encode proteins in mammalian cells (see Methods), and placed within DNA plasmid constructs. The COT *gag* gene sequence was optimized based on the codon usage of highly expressed human genes, resulting in an approximately 20% increase in the GC content (from 44.0 to 63.6%). The COT *tat* and *nef* gene sequences were optimized based on mRNA secondary structure stability (12, 28). The DNA sequence encoding an RNA with the lowest free energy was selected to

improve the stability and translation efficiency of COT transcripts, while the GC content of COT *tat* and *nef* remained unchanged at 40-45%.

***In Vitro* Expression of COT Gag, Tat and Nef proteins.** To evaluate the expression patterns of each of the three COT constructs, 293T cells were transiently transfected with plasmid DNA and cell lysates were analyzed by immunoblotting (Figure 2). An antibody specific to the viral core protein p24 recognized the Gag precursor p55 protein within cells and in cell supernatants (Figure 2A), as well as a cleavage product of approximately 41 kDa. COT Tat exon 1 migrated with the expected rate, corresponding to a molecular weight of approximately 14 KDa (Figure 2B). COT Nef was also well expressed and detected at 27 KDa and 25 KDa (Figure 2C), as expected due to the presence of alternative start codons. The codon-optimized Nef COT protein was also expressed at higher levels than Nef proteins corresponding to sequences from different HIV-1 patients or HIV-1 NL4-3 (data not shown).

Functionality of Gag, Tat and Nef proteins. The COT *gag* plasmid encodes Gag, but not the viral protease, and therefore expresses the p55Gag and not its mature cleavage products. Functional p55Gag is competent to bud from cells and form virus-like particles {Gheysen, 1989 #3; Wagner, 1992 #2}. Western blot of supernatants from transfected 293T cells showed that p55Gag was released from cells at high levels. And, as would be expected in the absence of protease, no processed p24Gag was observed in cell lysates. However, the protease mediated virus maturation process is not required for virus particle assembly and budding (reviewed in {Freed, 1998 #1}). Indeed, COT p55Gag generated extra-cellular virus-like particles, as shown by its resistance to proteolysis (1) (Figure 3). Furthermore, when a detergent that permeabilizes membrane vesicles was added, COT p55Gag became sensitive to proteolysis (Figure 3), implying that the proteolysis resistance is due to the inclusion of COT p55Gag in virus-like

particles. When a cocktail of protease inhibitors was added prior to the trypsin resistance assay, COT p55Gag was not degraded.

HIV-1 Tat expression results in high-level expression of viral proteins through transactivation of gene expression from the HIV-1 LTR (17). Since the second exon of Tat is dispensable for Tat transactivation (25), we assessed the transactivating potency of COT Tat exon 1 by comparing LTR expression in the absence or presence of COT Tat. 293T cells were transiently transfected with an LTR-LacZ construct in the absence or presence of a second plasmid encoding COT Tat exon 1. Basal and Tat-activated LTR activities were measured to determine the Tat-induced LTR activity. The LTRLacZ promoter demonstrated low basal and high COT Tat exon 1–induced transcription levels in the different experimental conditions that were tested. In the presence of COT Tat exon 1, LTRLacZ activities were increased 2 to 3 times over the basal level with LTRLacZ alone (Figure 4).

We also tested whether COT Nef was capable of mediating its reported function of downregulation of MHC-I from cell surfaces (Figure 5) (24). Expression plasmids encoding COT Nef, Nef from HIV-1 isolate NL4-3 or a defective reading frame of Nef NL4-3 were transfected in 293T cells and MHC-I expression was analyzed by flow cytometry. Average cell surface MHC-I expression was reduced 2-3 fold in cells transfected with either the COT Nef and Nef NL4-3 constructs compared with cells transfected with a pcDNA3-nef defective construct, consistent with previous reports (23).

Immunogenicity of Gag, Tat and Nef proteins. Mice were immunized with individual constructs, and cellular immune responses in splenocytes analyzed by IFNγ ELISpot using overlapping libraries of peptides corresponding to CON sequences from subtypes A and B, and group M. Each of the three COT DNA immunogens induced specific IFN-γ ELISpot responses when assayed with peptide pools corresponding to subtype B CON

sequences, and in some cases subtype A (Gag) and group M (Gag, Nef) CON peptides. Additionally, pCOTGag elicited stronger IFN-γ ELISpot responses than pGag02MAC, which encodes a circulating strain of HIV-1 subtype B (Figure 6).

Discussion

We used a novel algorithm to deduce phylogenetically informed, Center-of-Tree (COT) sequences, encoding the HIV-1 proteins Gag, Tat and Nef. We then synthesized the deduced sequences and experimentally characterized the encoded proteins to demonstrate that they retained important biological functions of the native proteins. Moreover, COT proteins elicited strong CTL immune responses in mice. We conclude that the COT algorithm infers functional ancestral proteins from sequences present in contemporary HIV-1 strains.

Phylogenetic-based ancestral sequence reconstructions like the COT may produce distinct products relative to CON sequence reconstructions, as recent results demonstrate that the behavior of an ancestral protein need not be an average of those of its descendants: Gaucher and colleagues (11), for example, have suggested that the ancestor of modern mesophiles lived at higher temperatures than its descendants, by showing that the ancestral proteins could function at 55°C. By comparing central protein sequences derived from different datasets, we found some variable positions among the sequences. It has previously been reported that differences in the ancestral sequences mostly derived from the method used to root the phylogenetic tree {Ross, 2006 #17261}. This emphasizes the need to carefully consider the sequence collection used to generate the tree and to assess the reliability of the tree prior to deriving a COT sequence.

We and others have suggested that reconstructions of central HIV sequences like COT, CON or ANC could be used to develop vaccine antigens (10, 19). Ancestral sequences are intended to be as similar as possible to all the strains of a given subtype, and therefore should induce immune system coverage broader than any individual native viral protein. We therefore recreated ancestral HIV-1 antigens for Gag, Tat and Nef, which are potentially critical immunologic targets given their immune reactivity (8) and their roles in the virus life cycle (29). Since we had recently derived an HIV-1 B ANC sequence for the Env gene and tested its immunogenicity in rabbits {Doria-Rose, 2005 #15726}, we chose not to reconstruct any COT Env. The structural HIV-1 protein Gag is highly conserved and is among the most common targets of the virus-specific cell mediated immune response, while Tat and Nef are more variable yet critical regulatory proteins important in viral gene expression and pathogenesis, and frequently induce T cell immune responses early in infection. Engineered COT proteins were well-expressed, a fundamental requirement for the immunogenicity of vaccine candidates, especially for the usually poorly expressed HIV-1 proteins (13, 33). Moreover, these proteins are capable of eliciting CTL immune responses in mice. It should be noted that COT Gag elicited strong cross-clade CTL responses when studied for reactivity against subtype B, A or group M peptide pools. Future detailed mapping studies need to be performed in order to formally demonstrate the ability of COT immunogens to induce broader CTL responses as compared to HIV-1 antigens. However, the increase in magnitude of the COT antigens observed for the CTL responses strongly suggests that CD4 help was improved by these designs suggesting a positive effect on Class II responses.

Our data confirm the utility of phylogenetic tools to select and construct novel functional ancestral gene sequences in the pursuit of understanding the core features of viral proteins required for function, and a broadly protective vaccine against HIV. Our

approach takes advantage of the rapid accumulation of sequence data to rationally design HIV antigens. Among the most vexing challenges of HIV therapeutics, as well as vaccine design, is the enormous capacity of HIV-1 to mutate and subsequently to become drug resistant or evade host immune responses. The critical problem posed by the extreme diversity of HIV is exacerbated by the long development and testing cycle of a new vaccine, which means that the variability of HIV will likely have changed considerably in the meantime. In this regard, it may be too optimistic to expect that the isolate-based candidate vaccines (currently in phase II and III human clinical trials) could be cross-reactive enough to protect against circulating viruses. Thus, using a central antigen such as COT might contribute to improvements over isolate-based vaccine approaches (21). Although more studies are necessary to determine whether central sequences will elicit cross-reactive responses sufficient to be protective, recent studies with a second generation CON group M *env* vaccine showed that it elicited improved levels neutralizing antibodies in guinea pigs, in some cases stronger and broader than contemporary isolates {Liao, 2006 #21271}. Further complicating the development of HIV antigens is the propensity of HIV to escape virus-specific cytotoxic T-lymphocytes (CTL), underscoring the importance of HIV-specific cellular immune responses in the control of the virus (2, 15, 16). Viral escape from host immunity thus represents a substantial hurdle for candidate CTL-vaccines. COT antigens could help avoid escape by potentially allowing more CTL responses to be generated by capitalizing on the larger representation of epitopes in COT designs compared to circulating isolates.

Apart from the critical importance of analyzing the genetic diversity of contemporaneous HIV on a global perspective, additional experimentation on ancestral state sequences is valuable for the study of HIV molecular evolution. Protein reconstruction provides an unusual opportunity to study the pathways and mechanisms of functional changes during molecular evolution, since the mechanistic basis and

dynamics of this process can be tracked in detail *in vitro*, allowing some fundamental questions to be rigorously examined (7, 30). Importantly, though analyzing ancestral proteins in the context of extant cells could lead to experimental artifacts, this risk is minimized in the case of HIV due to the enormous evolutionary rate of HIV compared to that of its host. That is, the virus' environment has remained essentially unchanged during viral evolution since its leap from non-human primates to humans in the last century (31). Ancestral reconstructions can shed additional light on the common functional elements of HIV strains, as well as the lineage-specific evolutionary changes that led to the multiple contemporaneous variations on these common elements. Hence, although biological properties are often studied using mutational analysis, we suggest that switching to the "ancestral-state" amino acid would be a pertinent means to investigate an amino acid's role. Our improving understanding and modeling of the molecular evolution of HIV along with a better knowledge of the correlates of immune protection will allow us to begin to predict how HIV sequences will change and to identify elements that increasingly successful therapies and an ultimately successful vaccine will need to incorporate.

Acknowledgments

The authors thank Frank Kirchhoff (Ulm, Germany) for the gift of the pNL4-3 plasmid. We thank Sherry McLaughlin, Angélique van 't Wout, Victor Swain and Fabienne Rayne for valuable discussions. The following reagents were obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH (http://www.aidsreagent.org/ecommerce): pHIVlacZ from Dr. Joseph Maio; Monoclonal Antibody to HIV-1 p24 (AG3.0) from Dr. Jonathan Allan; HIV-1 Nef Monoclonal Antibody (EH1) from Dr. James Hoxie; HIV-1BH10 Tat Monoclonal Antibody (15.1). This work was

supported by a gift from the Boeing Corporation and by the University of Washington Center for AIDS Research and STDs (NIH P30 AI27757).

References

1. **Agresta, B. E., and C. A. Carter.** 1997. Cyclophilin A-induced alterations of human immunodeficiency virus type 1 CA protein in vitro. J Virol **71:**6921-7.
2. **Allen, T. M., X. G. Yu, E. T. Kalife, L. L. Reyor, M. Lichterfeld, M. John, M. Cheng, R. L. Allgaier, S. Mui, N. Frahm, G. Alter, N. V. Brown, M. N. Johnston, E. S. Rosenberg, S. A. Mallal, C. Brander, B. D. Walker, and M. Altfeld.** 2005. De novo generation of escape variant-specific CD8+ T-cell responses following cytotoxic T-lymphocyte escape in chronic human immunodeficiency virus type 1 infection. J Virol **79:**12952-60.
3. **Bacchetti, S., and F. L. Graham.** 1977. Transfer of the gene for thymidine kinase to thymidine kinase-deficient human cells by purified herpes simplex viral DNA. Proc Natl Acad Sci U S A **74:**1590-4.
4. **Benner, S. A., M. D. Caraco, J. M. Thomson, and E. A. Gaucher.** 2002. Planetary biology--paleontological, geological, and molecular histories of life. Science **296:**864-8.
5. **Blanchette, M., E. D. Green, W. Miller, and D. Haussler.** 2004. Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res **14:**2412-23.
6. **Bleiweiss, R.** 1998. Slow rate of molecular evolution in high-elevation hummingbirds. Proc Natl Acad Sci U S A **95:**612-6.
7. **Chang, B. S., K. Jonsson, M. A. Kazmi, M. J. Donoghue, and T. P. Sakmar.** 2002. Recreating a functional ancestral archosaur visual pigment. Mol Biol Evol **19:**1483-9.
8. **Frahm, N., B. T. Korber, C. M. Adams, J. J. Szinger, R. Draenert, M. M. Addo, M. E. Feeney, K. Yusim, K. Sango, N. V. Brown, D. SenGupta, A. Piechocka-Trocha, T. Simonis, F. M. Marincola, A. G. Wurcel, D. R. Stone, C. J. Russell, P. Adolf, D. Cohen, T. Roach, A. StJohn, A. Khatri, K. Davis, J. Mullins, P. J. Goulder, B. D. Walker, and C. Brander.** 2004. Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities. J Virol **78:**2187-200.
9. **Galtier, N., N. Tourasse, and M. Gouy.** 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283:**220-1.
10. **Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber.** 2002. Diversity considerations in HIV-1 vaccine selection. Science **296:**2354-60.
11. **Gaucher, E. A., J. M. Thomson, M. F. Burgan, and S. A. Benner.** 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature **425:**285-8.
12. **Griswold, K. E., N. A. Mahmood, B. L. Iverson, and G. Georgiou.** 2003. Effects of codon usage versus putative 5'-mRNA structure on the expression of Fusarium solani cutinase in the Escherichia coli cytoplasm. Protein Expr Purif **27:**134-42.
13. **Hickman, H. D., A. D. Luis, W. Bardet, R. Buchli, C. L. Battson, M. H. Shearer, K. W. Jackson, R. C. Kennedy, and W. H. Hildebrand.** 2003. Cutting edge: class I presentation of host peptides following HIV infection. J Immunol **171:**22-6.

14. **Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham.** 1994. Application and accuracy of molecular phylogenies. Science **264:**671-77.
15. **Jones, N. A., X. Wei, D. R. Flower, M. Wong, F. Michor, M. S. Saag, B. H. Hahn, M. A. Nowak, G. M. Shaw, and P. Borrow.** 2004. Determinants of human immunodeficiency virus type 1 escape from the primary CD8+ cytotoxic T lymphocyte response. J Exp Med **200:**1243-56.
16. **Liu, Y., J. McNevin, J. Cao, H. Zhao, I. Genowati, K. Wong, S. McLaughlin, M. McSweyn, K. Diem, C. Stevens, J. Maenza, H. He, D. C. Nickle, D. Shriner, A. C. Collier, L. Corey, M. J. McElrath, and J. I. Mullins.** 2006. Selection on the human immunodeficiency virus type 1 proteome following primary infection. J Virol **80:**9519-29.
17. **Maio, J. J. B., F.L.** 1988. Regulation of Expression Driven by Human Immunodeficiency Virus Type 1 and Human T-Cell Leukemia Virus Type I Long Terminal Repeats in Pluripotential Human Embryonic Cells. J.Virol. **62:**1398-1407.
18. **Mullins, J. I., D. C. Nickle, L. Heath, A. G. Rodrigo, and G. H. Learn.** 2004. Immunogen sequence: The fourth tier of AIDS vaccine design. Expert Rev Vaccines **3 Suppl 1:**S151-9.
19. **Nickle, D. C., M. A. Jensen, G. S. Gottlieb, D. Shriner, G. H. Learn, A. G. Rodrigo, and J. I. Mullins.** 2003. Consensus and ancestral state HIV vaccines. Science **299:**1515-8; author reply 1515-8.
20. **Pagel, M.** 1999. Inferring the historical patterns of biological evolution. Nature **401:**877-84.
21. **Pitisuttithum, P., S. Nitayaphan, P. Thongcharoen, C. Khamboonruang, J. Kim, M. de Souza, T. Chuenchitra, R. P. Garner, D. Thapinta, V. Polonis, S. Ratto-Kim, P. Chanbancherd, J. Chiu, D. L. Birx, A. M. Duliege, J. G. McNeil, and A. E. Brown.** 2003. Safety and immunogenicity of combinations of recombinant subtype E and B human immunodeficiency virus type 1 envelope glycoprotein 120 vaccines in healthy Thai adults. J Infect Dis **188:**219-227.
22. **Rayne, F., A. V. Kajava, J. Lalanne, and R. Z. Mamoun.** 2004. In vivo homodimerisation of HTLV-1 Gag and MA gives clues to the retroviral capsid and TM envelope protein arrangement. J Mol Biol **343:**903-16.
23. **Schindler, M., S. Wurfl, P. Benaroch, T. C. Greenough, R. Daniels, P. Easterbrook, M. Brenner, J. Munch, and F. Kirchhoff.** 2003. Down-modulation of mature major histocompatibility complex class II and up-regulation of invariant chain cell surface expression are well-conserved functions of human and simian immunodeficiency virus nef alleles. J Virol **77:**10548-56.
24. **Schwartz, O., and e. al.** 1996. Endocytosis of major histocompatibility complex class I molecules is induced by HIV-1 Nef protein. Nature Medicine **2:**338-342.
25. **Seigel, L. J., L. Ratner, S. F. Josephs, D. Derse, M. B. Feinberg, G. R. Reyes, S. J. O'Brien, and F. Wong-Staal.** 1986. Transactivation induced by human T-lymphotropic virus type III (HTLV III) maps to a viral sequence encoding 58 amino acids and lacks tissue specificity. Virology **148:**226-31.
26. **Swofford, D. L.** 1999. PAUP* 4.0: Phylogenetic Analysis Using Parsimony (*and Other Methods), 4.0b2a ed. Sinauer Associates, Inc., Sunderland, MA.
27. **Towbin, H., T. Staehelin, and J. Gordon.** 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proc Natl Acad Sci U S A **76:**4350-4.
28. **Wu, X., H. Jornvall, K. D. Berndt, and U. Oppermann.** 2004. Codon optimization reveals critical factors for high level expression of two rare codon

genes in Escherichia coli: RNA stability and secondary structure but not tRNA abundance. Biochem Biophys Res Commun **313:**89-96.

29. **Yu, X. G., M. Lichterfeld, M. M. Addo, and M. Altfeld.** 2005. Regulatory and accessory HIV-1 proteins: potential targets for HIV-1 vaccines? Curr Med Chem **12:**741-7.
30. **Zhang, J., and H. F. Rosenberg.** 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proc Natl Acad Sci U S A **99:**5486-91.
31. **Zhu, T., B. T. Korber, A. J. Nahmias, E. Hooper, P. M. Sharp, and D. D. Ho.** 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature **391:**594-7.
32. **Zuckerkandl, E., and L. Pauling.** 1965. Molecules as documents of evolutionary history. J Theor Biol. **8:**357-66.
33. **zur Megede, J., M. C. Chen, B. Doe, M. Schaefer, C. E. Greer, M. Selby, G. R. Otten, and S. W. Barnett.** 2000. Increased expression and immunogenicity of sequence-modified human immunodeficiency virus type 1 gag gene. J Virol **74:**2628-35.

Figure Legends

Figure 1. Maximum likelihood phylogenetic tree for *gag* gene sequences. HIV-I subtype B *gag* sequences rooted with those from subtype D. The relative location of each computationally derived sequence is represented by a dot: gray for the ANC (Most Recent Common Ancestor), black for the COT (Center Of Tree) and unfilled for the CON (consensus).

Figure 2. Expression of GagCOT (A), TatCOT (B) and NefCOT (C) in transfected 293T cells. Cell and supernatant lysates were separated on SDS-PAGE gels, transferred on a PVDF membrane, and viral proteins detected using either anti-p24 monoclonal antibodies (A), anti-Tat serum (B) or anti-Nef monoclonal antibodies (C).

Figure 3. Trypsin resistance assay of COT Gag. Cell supernatants were collected from COT Gag transfected 293 T cells. The samples were then either untreated, or treated with trypsin in the absence or presence of a protease inhibitor cocktail. In parallel, samples were treated similarly in the presence of triton X100.

Figure 4. Transactivation of expression by COT Tat. A reporter gene encoding LacZ and under the control of the LTR was transfected in 293T cells with or without a plasmid encoding Tat. Following X-Gal staining, β-galactosidase expression was detected. Mean percentages of at least 3 independent experiments are shown.

Figure 5. Effect of Nef on MHC I cell surface expression in 293T cells. Cell surface expression of MHC I was recorded as mean fluorescence intensity by flow cytometry.

The mean and median values of 4 independent experiments are shown for COT Nef, and a vector containing a defective Nef reading frame used as a negative control, and the mean an median values of 3 independent experiments for NefNL4-3.

Figure 6. Immunogenicity of COT Gag, Tat and Nef. Antigen-specific T-cell responses induced by different plasmids were assessed by IFN-γ ELISPOT assays using different peptide pools. Mean (± standard deviation) numbers of SFC per 10^6 cells derived from groups of mice are shown on the *y*-axis. The antigens tested are shown on the *x*-axis.

Table 1.

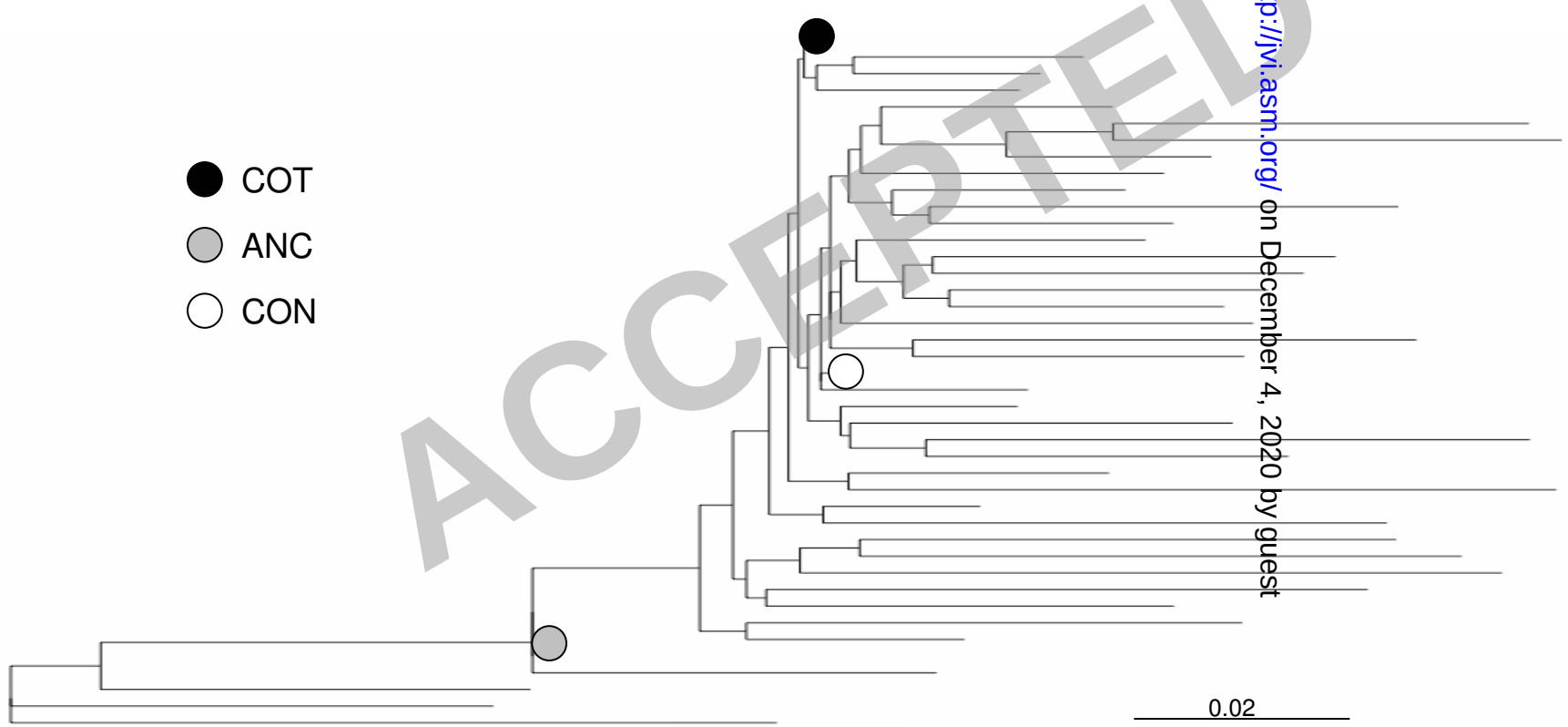
Average genetic distances between each of the central and ingroup sequences used to generate phylogenies. Shown are the average genetic distances as percentages with standard deviations shown in parenthesis.

	Gag	Tat	Nef
COT	0.0461 (0.00018)	0.0853 (0.00050)	0.0936 (0.00027)
CON	0.0464 (0.00018)	0.0860 (0.00050)	0.0993 (0.00027)
MRCA	0.0657 (0.00017)	0.0987 (0.00051)	0.1121 (0.00028)

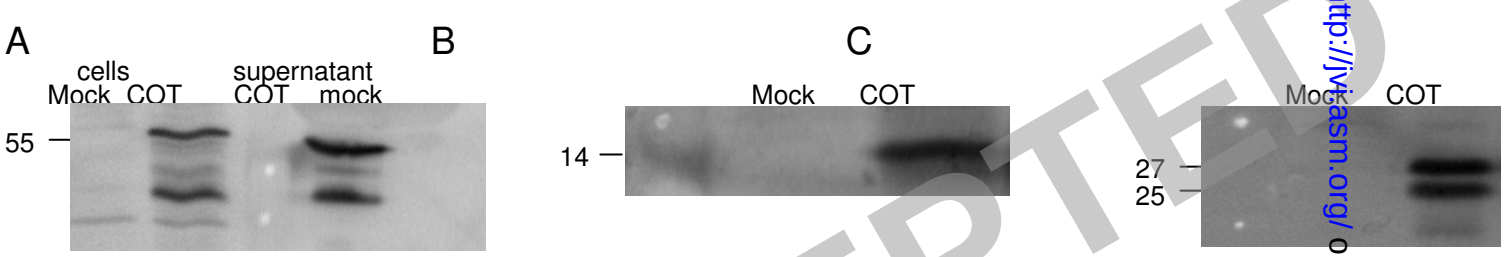
Table 2.

Amino Acid differences between the COT experimental proteins and each of the central sequences corresponding to sequences circulating currently and in 2002. The sequence alignments are based on sequences available at LANL. The AA positions of the mutations are indicated (based on HXB2 numbering), residues in bold correspond to mutations not previously reported in epitopes, mutations occurring within known epitopes are identified: Antibody (a), CTL (c), HTL (h).

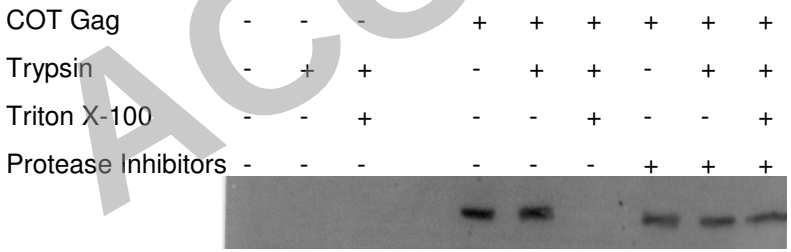
	Gag	Tat	Nef
COT 2002	15c-h, 138c-h, 403c-h	40c-h	11c, **152**, 178c-h
COT current	15c-h, 102c-h		85c-h, 163h, 170c-h, 205h
CON 2002	15c-h, 91c-h, 102c-h		15c-h, 51c-h, 85c-h, 178c-h
CON current	15c-h		10c, 85c-h
MRCA 2002	**67**, 102c-h, 279c-h, 312c-h, **385**, 389c, 403c-h, 490c-h	23c, 32c, 61h	8c, 10c, 11c, 21c-h, 61h, **157**, 169c-h, 170c-h
MRCA current	12c-h, **61,** 84c-h, 102c-h, 248c-h, 252c-h, 279c-h, 312c-h, 403c-h, 490c-h	23c, 32c, 40c-h, 59h-a, 61h	8c, 20c-h, 85c-h, 163h, 174c-h, 178c-h, 184c-h, 198c-h, 205h



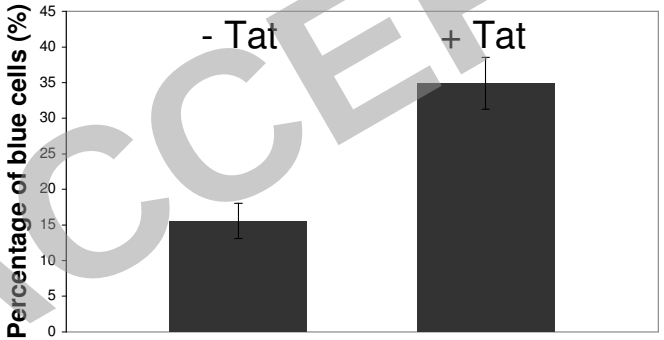
Rolland et al, Figure 2



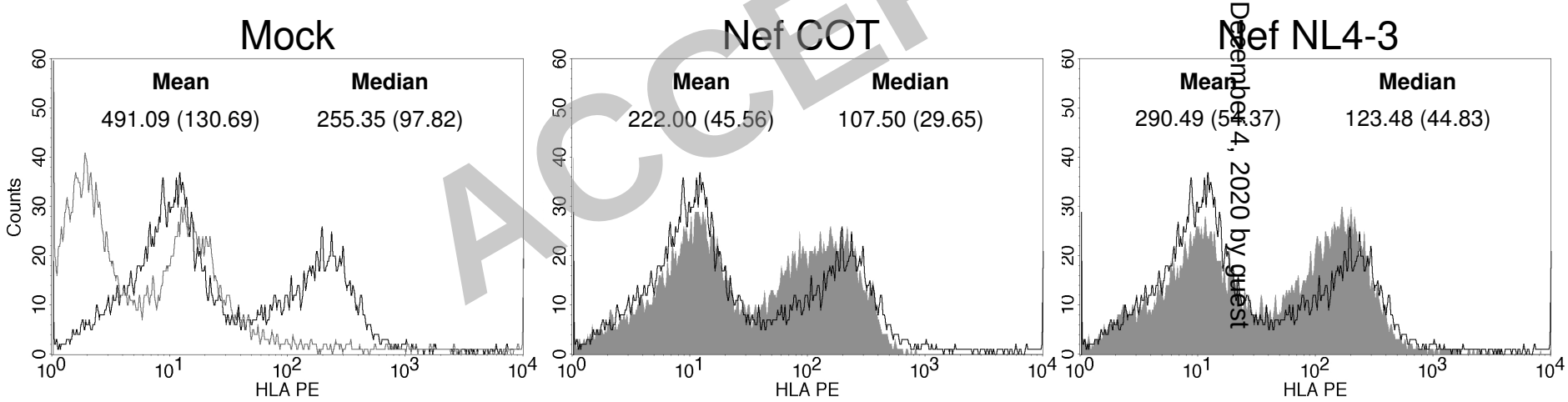
Rolland et al, Figure 3



Rolland et al, Figure 4



Rolland et al, Figure 5



Rolland et al, Figure 6