Journal of Virology

# A Reliable Phenotype Predictor for Human Immunodeficiency Virus Type 1 Subtype C Based on Envelope V3 Sequences

**Mark A. Jensen, Mia Coetzer, Angélique B. van 't Wout, Lynn Morris and James I. Mullins**
***J. Virol.* 2006, 80(10):4698. DOI: 10.1128/JVI.80.10.4698-4704.2006.**

JOURNAL OF VIROLOGY, May 2006, p. 4698–4704 Vol. 80, No. 10
0022-538X/06/$08.00+0 doi:10.1128/JVI.80.10.4698–4704.2006


# A Reliable Phenotype Predictor for Human Immunodeficiency Virus Type 1 Subtype C Based on Envelope V3 Sequences†

Mark A. Jensen,[1]‡* Mia Coetzer,[2]‡ Angélique B. van 't Wout,[1]§ Lynn Morris,[2] and James I. Mullins[1]

*Department of Microbiology, University of Washington, Seattle, Washington,[1] and AIDS Virus Research Unit, National Institute for Communicable Diseases, Johannesburg, South Africa[2]*

Received 2 December 2005/Accepted 22 February 2006

**In human immunodeficiency virus type 1 (HIV-1) subtype B infections, the emergence of viruses able to use CXCR4 as a coreceptor is well documented and associated with accelerated CD4 decline and disease progression. However, in HIV-1 subtype C infections, responsible for more than 50% of global infections, CXCR4 usage is less common, even in individuals with advanced disease. A reliable phenotype prediction method based on genetic sequence analysis could provide a rapid and less expensive approach to identify possible CXCR4 variants and thus increase our understanding of subtype C coreceptor usage. For subtype B V3 loop sequences, genotypic predictors have been developed based on position-specific scoring matrices (PSSM). In this study, we apply this methodology to a training set of 279 subtype C sequences of known phenotypes (228 non-syncytium-inducing [NSI] CCR5$^+$ and 51 SI CXCR4$^+$ sequences) to derive a C-PSSM predictor. Specificity and sensitivity distributions were estimated by combining data set bootstrapping with leave-one-out cross-validation, with random sampling of single sequences from individuals on each bootstrap iteration. The C-PSSM had an estimated specificity of 94% (confidence interval [CI], 92% to 96%) and a sensitivity of 75% (CI, 68% to 82%), which is significantly more sensitive than predictions based on other methods, including a commonly used method based on the presence of positively charged residues (sensitivity, 47.8%). A specificity of 83% and a sensitivity of 83% were achieved with a validation set of 24 SI and 47 NSI unique subtype C sequences. The C-PSSM performs as well on subtype C V3 loops as existing subtype B-specific methods do on subtype B V3 loops. We present bioinformatic evidence that particular sites may influence coreceptor usage differently, depending on the subtype.**

Viruses establishing human immunodeficiency virus type 1 (HIV-1) infection generally use CCR5 as a coreceptor for entry into host cells (11, 12). In some individuals, viruses are transmitted or viral genetic changes arise over time that permit the virus to use other coreceptors, particularly CXCR4. Viruses using CCR5 exclusively (also known as R5 viruses) are typically non-syncytium-inducing (NSI) in MT-2 cell assays, while those that use CXCR4, either exclusively (X4 viruses) or together with CCR5 (R5X4), are syncytium-inducing (SI) viruses (2). The ability of single virus clones to use both CCR5 and CXCR4 as coreceptors in HIV-1 is also well known for most subtypes (32). The emergence of viruses able to use CXCR4 is associated with accelerated CD4 T-cell decline and more rapid disease progression (18, 27). Coreceptor switching from NSI/CCR5 to SI/CXCR4 is seen in at least 50% of HIV-1 subtype B-infected individuals (27, 28). However, in HIV-1 subtype C infections, responsible for 56% of global infections (UNAIDS [http://www.unaids.org]), CXCR4-using viruses appear to be far less common, even in individuals with more advanced disease (3, 8, 9, 24). The ability to screen large subtype C virus-infected cohorts for R5X4 and X4 viruses is vital to a better understanding of these differences.

In many cases, however, it is not economically or practically possible to obtain biologically active virus samples to test in vitro, or the only available source of information is a sampling of viral population sequences. Although *env* (encoding the envelope) is not currently routinely sequenced (unlike *pol* [encoding the protease and reverse transcriptase], which is sequenced for drug resistance testing), it is likely that this will become more common with the imminent use of CCR5 antagonists in clinical trials, and the clinical interpretation of such sequences will become more important. Sequence-based methods for predicting coreceptor usage, if reliable, can provide a useful surrogate for biological phenotyping in these situations. Some genotype-based methods (such as the one presented here), in contrast to common phenotype-based assays, also score V3 sequences on a continuous scale. Changes in a continuous index have been correlated with shifts in phenotype and could be useful for predicting pathogenic changes within individuals that have not yet emerged biologically (17). This could become important for guiding therapeutic decisions and is a rationale for a larger role of *env* sequencing in the molecular surveillance of HIV-1.

Studies have shown that certain amino acid sites in the *env* gene, specifically in the third variable (V3) loop, are involved in coreceptor binding. This region plays an integral role in virus infectivity, and variations in the region have been correlated with changes in cell tropism, syncytium formation, and the progression of disease (6, 19, 21, 31). The V3 loop consists of approximately 35 amino acids with a conserved disulfide bridge at the base. Distinct genetic differences between CCR5-

* Corresponding author. Present address: Department of Global Health, Rollins School of Public Health, Emory University, 1518 Clifton Road, Atlanta, GA 30322. Phone: (404) 727-9776. Fax: (404) 727-4590. E-mail: mark.jensen@emory.edu.

† Supplemental material for this article may be found at http://jvi.asm.org/.

§ Present address: Department of Clinical Viro-Immunology, Sanquin Research, Amsterdam, The Netherlands.

‡ M.A.J. and M.C. contributed equally to this work.

and CXCR4-using viruses have been described that influence coreceptor usage (10, 14, 30). These differences have been exploited in bioinformatic approaches to predict tropism, with various degrees of success (17). These approaches include noting the presence or absence of positively charged amino acids at V3 sites 11 and/or 25 (the "11/25 rule") to distinguish between SI and NSI viruses (14), a multiple regression method based on positive, negative, and net V3 charges (4), a neural network strategy (26), a machine-learning method (23), and a subtype B position-specific scoring matrix (B-PSSM) (16). The PSSM showed improved predictive power over the other methods (17) and was also useful in the analysis of the transition from R5 to X4 in subtype B viruses (16).

For this work, we tested four of these existing predictors of viral tropism on a subtype C data set of V3 sequences with known phenotypes to determine their applicability to subtype C sequences. We found an initially poor performance of these methods, which were developed based on knowledge obtained from subtype B viruses/sequences, in predicting SI virus CXCR4 usage, suggesting that a predictor based on subtype C sequences was necessary. Since the B-PSSM was shown to have an improved positive predictive value (17), we analyzed PSSM predictors constructed from V3 sequences of subtype C isolates of known phenotypes (C-PSSM). Predictions based on the C-PSSM exhibited increased reliability and sensitivity over subtype B-based predictors. We also found that the previously described B-PSSM (16) performed comparably to the C-PSSM when the predictor cutoff score was optimized for subtype C sequences. Further bioinformatic analysis indicated that V3 sites influencing coreceptor usage may differ between the two subtypes.

## MATERIALS AND METHODS

**Data sets.** A training set of 279 HIV-1 subtype C V3 sequences was compiled from published literature, the Entrez nucleotide database (http://www.ncbi.nlm.nih.gov/entrez/), the Los Alamos HIV sequence database (http://hiv-web.lanl.gov), and our own laboratory (7–9; unpublished data) (see the supplemental material for sequences, accession numbers, and references). Only sequences for which the corresponding biological phenotypes were determined on coreceptor-transfected cell lines (for CCR5 and/or CXCR4 usage) or MT-2 cells (for NSI/SI phenotypes) were used in this study. In some cases, only MT-2 data were available, and therefore the training set was divided into NSI (R5) and SI (R5X4 and X4) viruses. There were 228 NSI V3 sequences (from 200 subjects) and 51 SI V3 sequences (from 20 subjects). We used sequence logos (29) to visualize the site-wise differences in amino acid distributions between SI and NSI sequences.

Our training set constituted our best effort to obtain all available phenotyped subtype C sequences as of June 2004. To secure a validation set on which to test the C-PSSM predictor, we searched the Los Alamos database in January 2006 to obtain new phenotyped V3 sequences not contained in our training set. This search yielded 24 additional X4/SI sequences and 47 additional R5/NSI sequences. These sequences were all distinct. Several V3 sequences occurred multiple times in the database, with each instance associated with a different isolate. In these cases, if phenotypic information was available for more than one accession number, we required that the phenotype be consistent across reports for inclusion in the validation set. Alignments, accession numbers, and other details for all sequences in this study are available in the supplemental material and at http://mullinslab.microbiol.washington.edu/.

Most (>97%) phenotype determinations for the sequences in the training set were obtained by bulk phenotyping methods (see the supplemental material). Specifically, primary patient peripheral blood mononuclear cells or heterogeneous primary isolates were typically cocultured with HIV-negative peripheral blood mononuclear cells, and the resultant viruses were assayed for phenotype. This is not ideal, as sequences obtained from a heterogeneous isolate may not be representative of the phenotype manifested in the assay. For example, NSI viruses may predominate in a mixed isolate, but rarer SI viruses in the isolate may generate syncytia in an MT-2 cell culture. The isolate would thus be assigned an SI phenotype, but a sequence sampled from the isolate would likely belong to an NSI virus. Obtaining sequence and phenotype information from biological or molecular clones would obviate this problem. Unfortunately, a large data set of clonal isolates with associated sequences and phenotypes does not currently exist for subtype C.

More precisely, then, the test that we developed is a predictor of the phenotype of a bulk isolate based on the genotype of a single sequence sampled from that isolate. In one sense, this makes the test more practical, as most phenotyping studies involve bulk isolates. We have also observed for subtype B viruses (M. Curlin, J. I. Mullins, et al., unpublished data) that there is little difference in predictive power for PSSM predictors generated using assay data for either bulk or cloned isolates, provided that the size of the data set is large enough. Moreover, the success in prediction we obtained in this study also suggests that mismatches between sequence and bulk phenotypes are relatively uncommon, possibly due to selection at the level of coculture.

**Genotypic algorithms.** A subset of the data representing only the unique sequences within 220 subjects (constituting 200 NSI sequences and 23 SI sequences) was subjected to the following four prediction methods: the 11/25 rule (14), a multiple regression method referred to as the Briggs method (4), a machine-learning method referred to as the Pillai method (23), and the B-PSSM, as publicly implemented (16; http://mullinslab.microbiol.washington.edu/computing/pssm). The percentage of sequences with correctly predicted phenotypes was calculated for each algorithm.

**Development of C-based scoring matrices.** We derived a predictor from PSSM, calculated as described previously (16), based on the subtype C training set of 279 V3 sequences. Distributions of specificities (fractions of correctly predicted NSI sequences) and sensitivities (fractions of correctly predicted SI sequences) were estimated by combining data set bootstrapping with leave-one-out cross-validation. In this procedure, the target sequence and all sequences of that phenotype from the same infected individual were removed from the data set, and single sequences of each phenotype from the remaining individuals were randomly sampled with replacement. The random sample was used to calculate a PSSM predictor, and the PSSM was used in turn to predict the phenotype of the target. The prediction was made by comparing the score of the target sequence to a cutoff score, as follows: an SI prediction was called if the target score was greater than the cutoff, and an NSI prediction was called if the score was less than the cutoff. Cutoffs were calculated as scores which maximized the product of the sensitivity and specificity of predictions made by applying the PSSM to the sequences used to produce the PSSM (16). Resampling was repeated 100 times to obtain an empirical prediction probability for the target sequence. Each sequence in the data set was treated as a target in turn. The Bernoulli random variables represented by the prediction probabilities for each sequence were then sampled to obtain a set of simulated predictions, and sensitivity and specificity were calculated using the simulated predictions. Bernoulli sampling was repeated 100 times to obtain the reported empirical distributions for sensitivity and specificity. All analyses were performed using scripts written in PERL and R (25; http://cran.r-project.org). Scripts are available upon request.

For between-subtype comparisons, we performed this analysis on an HIV-1 subtype B data set consisting of 187 NSI and 70 SI sequences from 107 infected subjects (26).

**ROC analysis.** We used receiver-operator characteristic curves (ROCs) (1) to compare the C-PSSM and B-PSSM on the basis of the overall ability to discriminate between SI and NSI sequences. We used a C-PSSM based on the entire data set and a B-PSSM based on SI/NSI sequences as described previously (16). For a particular set of target sequences, each target was scored using one of the matrices. For any cutoff score, false-positive results were those NSI sequences whose scores lay above the cutoff, and positive results were SI sequences with scores above the cutoff. The ROC was generated by plotting the pair (false-positive fraction of NSI sequences and positive fraction of SI sequences = 1 − specificity and sensitivity) for a set of 100 cutoff scores evenly spanning the range of PSSM scores. To get a sense of the variation in ROC over the set of infected individuals represented in our data, we calculated an ROC for each of 100 sampled data sets comprised of one randomly selected sequence per patient per phenotype.

We analyzed the performances of both matrices on both subtype C and B V3 sequences and plotted the distributions of the areas under the ROC for each of the four cases. The area under the curve is a measure of the test's ability to correctly predict the phenotype of any sequence: an area of 0.5 indicates that the test is not better than a random guess, while a perfect test (for those sequences analyzed) has an area of 1. We calculated the area using Euler's method of integration.

**Overlap coefficient analysis.** To investigate the potential differences between subtypes B and C, we examined the overlap coefficients (OCs) (13) between SI and NSI amino acid profiles. The training sets described the amino acid fre-

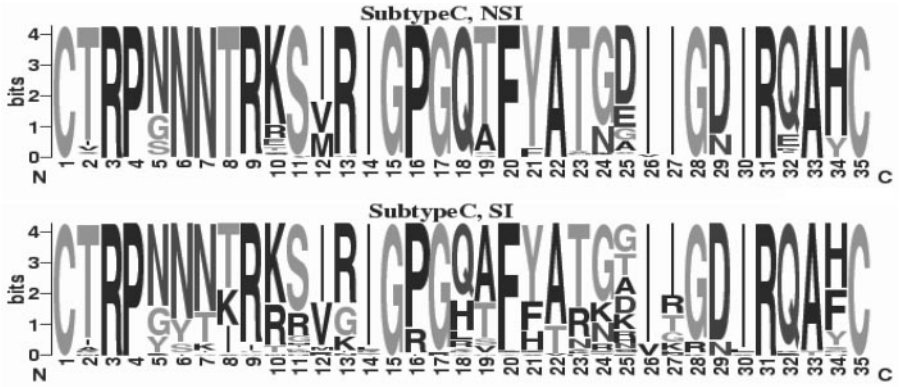


FIG. 1. Sequence logos of HIV-1 subtype C V3 sequences used in this study. The character and size of each logo represent the proportion of an amino acid at the specific site. The subtype C NSI data set is represented by 228 NSI V3 sequences, and the SI data set corresponds to 51 SI V3 sequences.

quency distribution for each site. The OC, in this context, is a site-wise measure of the difference between the SI and NSI amino acid distributions for V3 sequences (equations for the OC are given in reference 13). If the OC is 0, the distributions are identical, and if the OC is 1, there is no amino acid overlap between the distributions (i.e., the SI amino acids at a site are completely distinct from NSI amino acids at that site). Thus, the OC is a measure of the ability of a V3 site to discriminate between the two phenotypes based on the training set.

To determine whether an OC was significantly high, we compared it to a distribution of OCs generated by randomly assigning training set sequences to the SI or NSI category. OCs were calculated for 250 random permutations, and the *P* value of the training set OC was reported as 1 minus its percentile within the random OC distribution.

**HTA.** The V3-based heteroduplex tracking assay (V3-HTA) has been used as a rapid genotype-based method to identify genetic variation associated with NSI- and SI-like viruses in subtypes B and C (22, 24). The mobility of the heteroduplex reflects the differences between the probe and the sample sequence. This is measured by the mobility ratio *k*, which is the distance traveled by the heteroduplex divided by the distance traveled by the homoduplex. The greater the genetic difference between the probe and the sample, the slower the migration of the heteroduplex and the smaller the mobility ratio. V3-HTA mobility ratios were available for 13 NSI and 8 SI subtype C viral isolates from South Africa using an NSI probe (9). The C-PSSM score of each of these isolates was calculated and correlated to the V3-HTA mobility ratio to determine the relationship between the genotypic algorithm and a genotype-based molecular assay.

**C-PSSM for public use.** We have made a C-PSSM predictor available online at http://mullinslab.microbiol.washington.edu/computing/pssm/, based on the computational techniques presented in this paper. The details of this implementation, including improved handling of data set sampling issues, are beyond the scope of this report and will be described separately (M. A. Jensen, unpublished data). Briefly, all sequences in the data set were used to generate the matrix, but the contribution of each infected individual's sequence was weighted by the reciprocal of the number of sequences sampled from that individual. We refer to this as the "sample-averaged" C-PSSM.

## RESULTS

**HIV-1 subtype C sequences.** The data set compiled for this study consisted of 279 HIV-1 subtype C V3 sequences with known biological phenotypes consisting of 228 NSI V3 sequences (from 200 subjects) and 51 SI V3 sequences (from 20 subjects). A graphical representation of these sequences using sequence logos (29) is shown in Fig. 1. A higher degree of amino acid variation within the V3 loops of subtype C SI sequences distinguished them from the NSI sequences. Overall, 29 of the 35 amino acid loci among SI viruses showed variation, compared to only 12 among NSI viruses. Given that more than four times the number of NSI sequences were available, the limited V3 variation of NSI viruses compared to SI viruses is unlikely to result from sampling error. The GPGQ crown motif, which is typical of subtype C viruses, was highly conserved among the NSI sequences but showed variation at positions 16 and 18 among SI sequences; in particular, the Q at position 18 was heavily substituted. There was also increased variation at positions 11, 13, 25, and 27 in the SI sequences. In subtype B viruses, positions 11 and 25 are often positively charged amino acids in SI viruses (15), but in this subtype C data set, this was less evident.

**Predicting phenotypes from genetic sequence data.** The performances of four commonly used phenotype predictors were evaluated with a subset of 223 sequences representing unique sequences (one randomly selected from each individual) from the data set (Table 1). All algorithms predicted the NSI phenotype with a high degree of accuracy (99.5%), except for the Briggs method, where only 52% of the NSI sequences were correctly identified. For SI viruses, all four algorithms performed poorly in predicting biological phenotypes. This may be due to the fact that the 11/25, Briggs, and B-PSSM methods were developed with subtype B sequences and the Pillai method used a majority of B sequences. Among these algorithms, the PSSM had an increased predictive power (17), and we therefore chose this method for the development of a subtype C-specific predictor.

**Performance of C-PSSM prediction.** A subtype C PSSM was derived using the data set of 279 V3 sequences, with each

TABLE 1. Comparison of performances of public implementations of available prediction methods to determine viral phenotypes of the subtype C unique data set[a]

| Method | % of sequences with correctly predicted phenotype (no. correct/total no.) | |
|---|---|---|
| | NSI ($n$ = 200) | SI ($n$ = 23) |
| 11/25 | 99.5 (199/200) | 47.8 (11/23) |
| Briggs | 53.0 (106/200) | 39.1 (9/23) |
| Pillai (SVM)[b] | 99.5 (199/200) | 52.2 (12/23) |
| B-PSSM (sinsi)[c] | 99.5 (199/200) | 56.5 (13/23) |

[a] The Pillai method and B-PSSM are web-based tools, and the pretrained classifiers selected for analysis are shown in parentheses.
[b] http://genomiac2.ucsd.edu:8080/wetcat/index.html.
[c] http://mullinslab.microbiol.washington.edu/computing/pssm/.

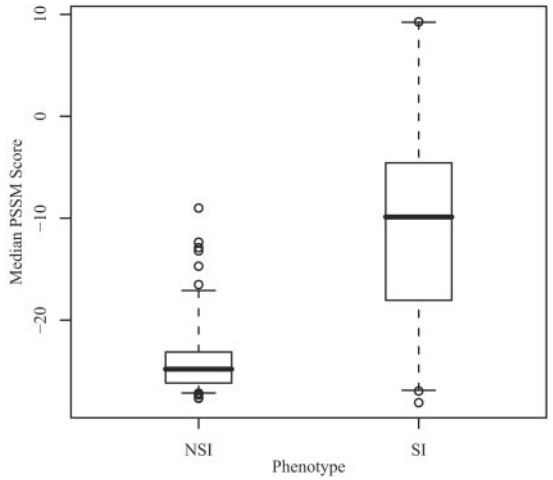


FIG. 2. Comparison of C-PSSM score distributions of 229 NSI and 51 SI subtype C sequences. Scores are the median PSSM scores over 100 bootstrapped data sets, as described in Materials and Methods. Box boundaries are interquartile ranges, and central lines are medians over sequences. Error bars extend from the 2.5th and the 97.5th percentiles; beyond these, sequences are represented as outlier points. In a Kruskal-Wallis test, $\chi^2 = 78.9$ and $P < 10^{-15}$.

containing 35 amino acids. The C-PSSM scores of the NSI viruses ranged from −27.7 to −9.0 (median, −24.8), while those for SI viruses ranged from −28.1 to 9.3 (median, −9.8) (Fig. 2). The median C-PSSM score distributions for NSI and SI sequences differed significantly ($P < 10^{-15}$) by the Kruskal-Wallis test.

We compared the performances of subtype C-based and subtype B-based matrices in several ways. The distributions in Fig. 3 suggest that a C-PSSM test is likely to be more specific and less sensitive than a B-PSSM test on homotypic sequences (i.e., sequences of the same subtype as those used to create the PSSM). But the sensitivity and specificity are not independent as the cutoff score varies, as indicated in the ROCs in Fig. 4A. The ROC area for the B-PSSM test on subtype B sequences is significantly higher than that for the C-PSSM test on subtype C sequences (Fig. 4B)

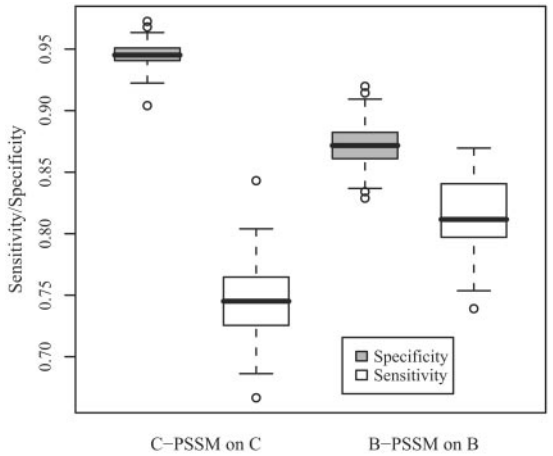


FIG. 3. Comparison of C-PSSM on subtype C data set with B-PSSM on subtype B data set, using leave-one-out/bootstrap predictions. The boxes show the values described in the legend to Fig. 2.

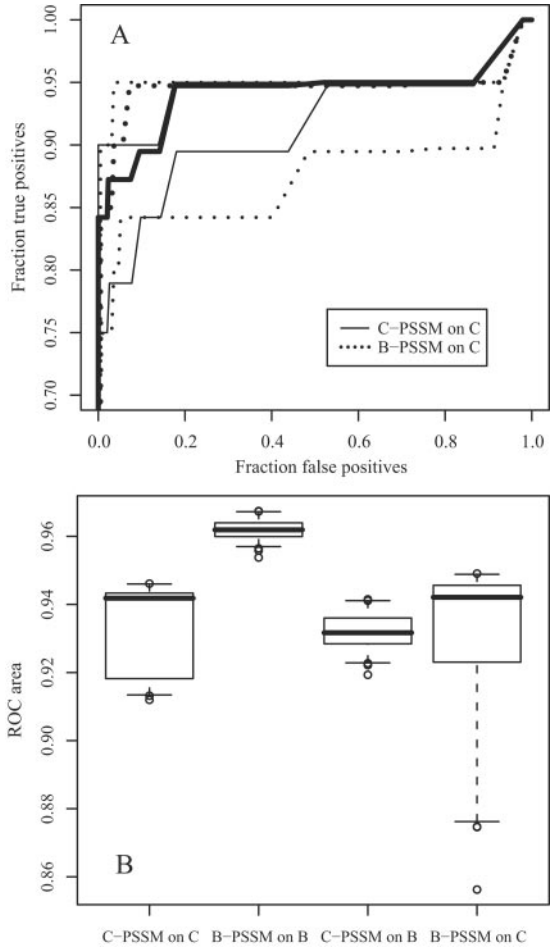


FIG. 4. ROC analysis. (A) ROCs for C-PSSM and B-PSSM tests on subtype C sequences. *x* axis, false-positive fractions of true NSI sequences (equivalent to 1 − specificity); *y* axis, positive fractions of true SI sequences (equivalent to sensitivity). Thicker lines indicate the median positive fraction over 100 bootstrapped data sets (see text), and thinner lines above and below are 97.5th and 2.5th percentiles, respectively. Solid lines, C-PSSM; dotted lines, B-PSSM. (B) Areas under the ROCs over 100 bootstrapped data sets for all combinations of matrices and data.

(Kruskal-Wallis test; $P < 10^{-15}$), suggesting that the B-PSSM is more likely to make a correct prediction, regardless of phenotype, than the C-PSSM for homotypic sequences. This is likely due to the larger number of unique SI sequences available for subtype B.

To test the hypothesis that using subtype C sequences to build a PSSM yields an improved predictor for subtype C viruses, we also considered how well the existing B-PSSM predicts subtype C sequences. We found that the B-PSSM in its public implementation, which uses a cutoff score optimized for subtype B sequences, was highly specific (99.5%) but not sensitive (52%) (Table 1). However, our ROC analysis indicated that neither C nor B matrices had an advantage in correctly predicting subtype C phenotypes; the means of ROC areas in these two cases were not significantly different (Fig. 4B) (Kruskal-Wallis test; $P = 0.11$). This implies that a B-PSSM with a cutoff optimized for subtype C sequences could improve the sensitivity of that test, with a limited cost in specificity. In

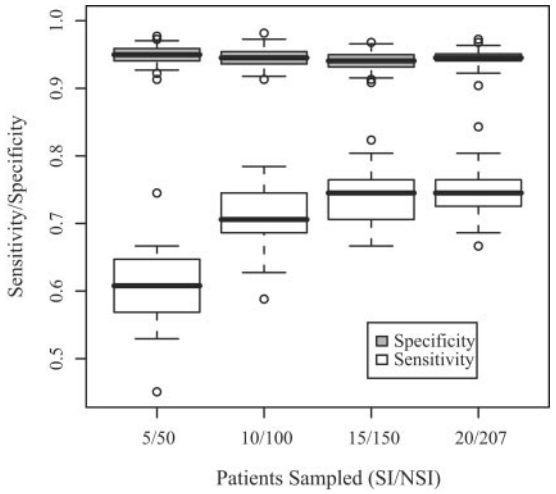


FIG. 5. Effect of sample size on sensitivity and specificity of C-PSSM. Boxes show the values described in the legend to Fig. 2.

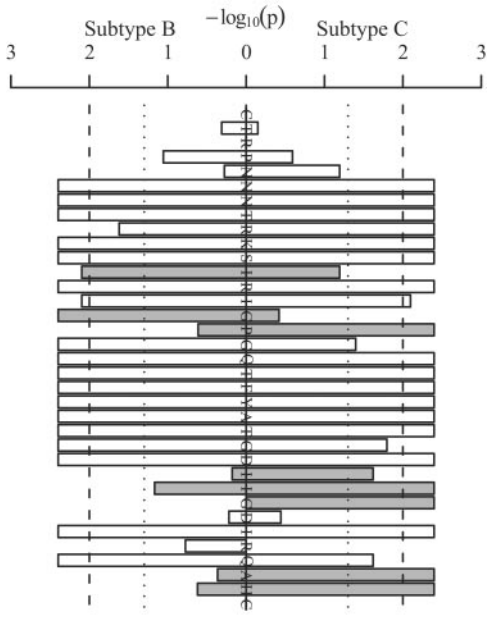


FIG. 6. V3 overlap coefficient *P* values for subtypes C and B. Dotted lines, $P = 0.05$; dashed lines, $P = 0.01$. On the main axis, the V3 site is labeled with subtype C NSI consensus residues. Gray pairs show sites in which the measured OC was significant in one subtype and nonsignificant in the other.

fact, the B-PSSM had a sensitivity of 84% and a specificity of 93% with the subtype C data set, with an optimal cutoff calculated for that data set.

The ROCs in Fig. 4A show that the distributions of sensitivity for a given specificity overlap for the C-PSSM and B-PSSM tests on subtype C sequences. Thus, neither test has a clear advantage, although the B-PSSM has a better median sensitivity for low levels of false-positive results. The variances of the ROC area distributions (Fig. 4B) in these two cases differed significantly (Levene-Carroll-Schneider test [1]; $F = 6.74$; $P = 0.01$). Since the B-PSSM test response was more variable, this suggests that tests on subtype C sequences using a B-PSSM will be more dependent on the sequences being analyzed. That is, some subtype C virus-infected patients may be very well predicted, while others will be rather poorly predicted, using a B-PSSM, while this effect may be ameliorated by using a C-PSSM.

Finally, the ROCs can be used to compare the C-PSSM to the publicly implemented methods shown in Table 1. Each of the methods examined (excluding the Briggs method) gave an almost perfect specificity; the sensitivities at that level ranged between 47.8% and 52.0%. Inspection of the C-PSSM ROC at the 2.5th percentile shows that a sensitivity of approximately 75% can be attained for a specificity of 100%. The C-PSSM thus represents a significant performance improvement over these methods in this respect.

To determine whether the sensitivity and specificity might be significantly improved by increasing the number of training sequences, we performed leave-one-out/bootstrap analysis on subsets of the data set. The total size of the subsets was increased incrementally, and unique SI and NSI sequences were randomly selected in a ratio of 1:10, comparable to the ratio in the total data set (Fig. 5). The specificity of the C-PSSM for predicting NSI phenotypes was high at even the smallest total sample size, and it declined slightly when the sample number was increased ($P = 0.001$ for a Kruskal-Wallis test between the smallest and largest sample sizes). The sensitivity of the C-PSSM for predicting SI phenotypes with small sample numbers was poor but appeared to approach a limit as the sample size increased to approximately 100 (Kruskal-Wallis test between sample size pairs; 5/50 versus 10/100, $P < 10^{-15}$; 10/100 versus 15/150, $P = 6 \times 10^{-6}$; 15/150 versus 20/207, $P = 0.025$).

**Validation.** As further validation, the sample-averaged C-PSSM (described in Materials and Methods) was applied to a separate set of unique subtype C sequences that were recently reported. To make predictions, we used an optimal cutoff score calculated based on the leave-one-out methodology described here, which will be detailed separately (Jensen, unpublished data). The C-PSSM achieved 83.3% (20/24) sensitivity and 83.0% (39/47) specificity, with an ROC area of 88.1%. The B-PSSM with the cutoff optimized for the subtype C training set gave 54.1% (13/24) sensitivity and 70.7% (37/47) specificity with an ROC area of 70.9%. The validation sequences and details of the analysis are included in the supplemental material.

The fact that the ROC areas for both subtypes fall below the estimated confidence intervals (Fig. 4) indicates that a prediction bias persists in the cross-validation analysis, despite the use of the leave-one-out technique. This bias should not affect the relative comparisons between subtypes that we have made above, but the estimated values for optimal specificity and sensitivity should be considered upper bounds of the actual values.

**Site-wise differences in phenotype between subtypes B and C.** Because SI viruses are reported to be much less prevalent in subtype C than in subtype B populations, it is possible that different, less evolutionarily labile sites influence the manifestation of phenotype for subtype C viruses. The availability of both subtype B and subtype C training sets afforded us a chance to investigate potential differences using OCs. Figure 6 displays the *P* value for the OC for each V3 site, comparing

subtype B (using the SI/NSI data set of Resch et al. [26]) and subtype C sequences (using our data set). Sites with OCs that exceed the 95th percentile of the random permutation distribution (depicted in Fig. 6 as bars that extend beyond the dotted lines) have amino acid distributions that are significantly different between SI and NSI viruses. Sites with nonsignificant OCs are less informative for purposes of discriminating between SI and NSI viruses by genotype. Under this interpretation, the OC analysis highlights sites that are potentially different in their influence on phenotype between the two subtypes. In particular, V3 sites 12, 15, 16, 26, 27, 28, 33, and 34 (shown in gray in Fig. 6) have significant OC values in one but not the other subtype.

**Comparison of C-PSSM scores by HTA.** The V3-HTA has been used to identify NSI- and SI-like viruses when hybridized with an R5 probe (9, 22, 24). Samples with a similar sequence to the probe have high mobility ratios ($k > 0.90$) and are usually NSI, whereas samples with low mobility ratios ($k < 0.90$) are different from the probe and are frequently SI (9). The mobility ratios of 8 SI (5 R5X4 and 3 X4) and 13 NSI isolates (9) were compared to their C-PSSM scores. A highly significant correlation between the mobility ratios and PSSM scores of the samples ($P < 0.0001$; $r^2 = 0.53$) was found (data not shown). NSI viruses had high mobility ratios and decreased PSSM scores, and most of these clustered tightly together. Two NSI samples had unusually low mobility ratios (0.76 and 0.80). However, they also had amino acid deletions that probably accounted for this, a property that did not affect the C-PSSM scores. Overall, the SI samples had a broader distribution of mobility ratios and C-PSSM scores.

## DISCUSSION

We tested four available phenotype prediction methods developed for HIV-1 subtype B on HIV-1 subtype C sequences with known phenotypes and found them to be highly accurate in predicting NSI usage, but less so in predicting CXCR4 usage. We therefore derived a subtype C-specific phenotype predictor that increased the accuracy of predicting CXCR4 usage and that performs nearly as well on subtype C V3 loops as existing subtype B-specific methods do on subtype B V3 loops. This correlated well with a genotype-based method for detecting NSI and SI viruses (V3-HTA), suggesting that the C-PSSM can be applied to subtype C V3 sequences of unknown coreceptor usage.

Sequence logos highlighted appreciable differences between V3 sequences of NSI and SI subtype C viruses. The NSI data set was very homogeneous, with little or no variation at many of the amino acid sites, while in the SI data set there was greater variation at most sites. These data suggested that sufficient genetic variation between NSI and SI subtype C sequences exists to allow the sequences to be used to differentiate these phenotypes. However, none of the available prediction methods were able to adequately exploit these differences in differentiating NSI from SI viruses. The 11/25 rule is based on the presence of positively charged amino acids at positions 11 and/or 25 (14, 20). However, >50% of the subtype C SI sequences in this study did not have a positively charged amino acid at either of these positions (Table 1, percent correctly predicted for the 11/25 method). Furthermore, while the 11/25 method is considered a reliable sequence-based phenotype predictor, other studies have shown that more than two amino acid positions need to be considered when assigning phenotype (26). The Briggs method performed the least well of all the algorithms evaluated. This method is based on genotype variables in the V3 region derived from subtype B sequences, with NSI viruses having a net charge of <4 and SI viruses having a net charge of >4. The net charges of the subtype C NSI data set ranged from 0 to 6, and the SI data set had net charges from 3 to 9, which probably explains the poor performance of this algorithm. The Pillai method is a two-way classification method that differentiates between viruses able or unable to use CXCR4. The limitations of this method include the fact that it misclassifies R5X4 viruses (23). These dual-tropic viruses may represent an intermediate stage of coreceptor evolution in subtype B viruses (16, 33, 34) and are usually grouped into the SI data set, as was done in this study. Our SI data set contained nine dual-tropic viruses, six of which were incorrectly predicted as NSI using the Pillai method.

While most methods could accurately identify NSI viruses, it was clear that a new method was needed to improve the sensitivity of prediction of SI viruses from subtype C sequences. The subtype C-specific PSSM addressed some of the limitations of other prediction methods, including the B-PSSM as publicly implemented (16). In particular, the C-PSSM identified SI viruses more reliably than the other methods, resulting in a significant increase in sensitivity when low levels of false-positive results are required. We also found that the B-PSSM performed comparably to the C-PSSM when the cutoff score was optimized for subtype C sequences. In this case, however, the variability of performance, as measured by ROC areas, was greater for the B-PSSM.

It was previously suggested (16) that the PSSM score represents the "X4 potential" of a sequence, in that intermediate scores track the temporal evolution of viruses within an individual. This method has also contributed to a better understanding of the role of intermediates (R5X4) in the transition from R5 to X4 in subtype B viruses (16) and can now be applied to subtype C viruses, where this transition has not often been reported. For this application, and to improve the prediction quality of the C-PSSM, vigorous sampling of more patients will be required, in at least the present ratio of SI to NSI sequences. Therefore, future sampling should focus on the acquisition of more X4 sequences from new individuals.

Potential site differences between subtypes B and C within the V3 loop were investigated by overlap coefficient analysis using current training sets. This suggested that changes in the crown at site 15 will influence coreceptor usage in subtype B viruses but that changes at site 16 will have more influence in subtype C viruses. Other sites that had significant OC values in one but not the other subtype were sites 12, 26, 27, 28, 33, and 34. These are unlikely to be simple artifacts of sampling, since the majority of sites are congruent (either both significant or both nonsignificant) between the subtypes and since both possibilities (significant for subtype B and nonsignificant for subtype C and vice versa) are represented at the incongruent sites. However, we are not claiming that incongruent sites constitute a rejection of any explicit null model. Rather, this simple analysis suggests the possibility of differential phenotypic effects of mutations at certain sites that could be evaluated in future studies.

The C-PSSM represents an improvement over currently available methods for predicting SI viruses in subtype C populations. This could lead to improved detection of SI viruses and to insights into the pathogenic role of coreceptor usage in this subtype. With the increased availability of small-molecule fusion inhibitors and the use of antiretroviral therapy in patients infected with subtype C viruses, there is concern that certain therapies may increase the risk of developing X4 viruses during subtype C infections. PSSM scores may be a useful tool for assessing baseline risk for this possibility (17) and may have prognostic value for treatment outcomes in general (5). Although the number of HIV-1 subtype C SI viruses available is a limiting factor, this study has shown that currently available data provide a good initial basis for a subtype C coreceptor usage predictor.

## ACKNOWLEDGMENTS

We thank G. Pollakis for data clarification and discussion and our anonymous reviewers for helpful comments.

This work was funded by grants from the South African AIDS Vaccine Initiative (SAAVI), The Wellcome Trust, and the Poliomyelitis Research Foundation and by grants to J.I.M. and A.B.W. from the U.S. Public Health Service, including grants to the University of Washington Center for AIDS Research. L.M. is a Wellcome Trust International Senior Research Fellow in Biomedical Science in South Africa. M.C. received travel support from the Fogarty Training Fellowship (TWO-0231). M.A.J. was supported in part by NIH award GM33782 to Bruce R. Levin.

## REFERENCES

1. **Armitage, P., G. Berry, and J. N. S. Matthews.** 2002. Statistical methods in medical research, 4th ed. Blackwell Scientific Publications, Oxford, United Kingdom.
2. **Berger, E. A., R. W. Doms, E. M. Fenyo, B. T. Korber, D. R. Littman, J. P. Moore, Q. J. Sattentau, H. Schuitemaker, J. Sodroski, and R. A. Weiss.** 1998. A new classification for HIV-1. Nature **391:**240.
3. **Bjorndal, A., A. Sonnerborg, C. Tscherning, J. Albert, and E. M. Fenyo.** 1999. Phenotypic characteristics of human immunodeficiency virus type 1 subtype C isolates of Ethiopian AIDS patients. AIDS Res. Hum. Retrovir. **15:**647–653.
4. **Briggs, D. R., D. L. Tuttle, J. W. Sleasman, and M. M. Goodenow.** 2000. Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages). AIDS **14:**2937–2939.
5. **Brumme, Z. L., W. W. Dong, B. Yip, B. Wynhoven, N. G. Hoffman, R. Swanstrom, M. A. Jensen, J. I. Mullins, R. S. Hogg, J. S. Montaner, and P. R. Harrigan.** 2004. Clinical and immunological impact of HIV envelope V3 sequence variation after starting initial triple antiretroviral therapy. AIDS **18:**F1–F9.
6. **Cann, A. J., M. J. Churcher, M. Boyd, W. O'Brien, J. Q. Zhao, J. Zack, and I. S. Chen.** 1992. The region of the envelope gene of human immunodeficiency virus type 1 responsible for determination of cell tropism. J. Virol. **66:**305–309.
7. **Choge, I., T. Cilliers, P. Walker, N. Taylor, M. Phoswa, T. Meyers, J. Viljoen, A. Violari, G. Gray, P. Moore, M. Papathanasopoulos, and L. Morris.** Genotypic and phenotypic characterization of viral isolates from HIV-1 subtype C infected children with slow and rapid disease progression. AIDS Res. Hum. Retrovir., in press.
8. **Cilliers, T., J. Nhlapo, M. Coetzer, D. Orlovic, T. Ketas, W. C. Olson, J. P. Moore, A. Trkola, and L. Morris.** 2003. The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. J. Virol. **77:**4449–4456.
9. **Coetzer, M., T. Cilliers, R. Swanstrom, and L. Morris.** What genetic changes in V3 are associated with CXCR4 usage in HIV-1 subtype C isolates? Submitted for publication.
10. **De Jong, J. J., A. De Ronde, W. Keulen, M. Tersmette, and J. Goudsmit.** 1992. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. J. Virol. **66:**6777–6780.
11. **Deng, H., R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. Di Marzio, S. Marmon, R. E. Sutton, C. M. Hill, C. B. Davis, S. C. Peiper, T. J. Schall, D. R. Littman, and N. R. Landau.** 1996. Identification of a major co-receptor for primary isolates of HIV-1. Nature **381:**661–666.
12. **Dragic, T., V. Litwin, G. P. Allaway, S. R. Martin, Y. Huang, K. A. Nagashima, C. Cayanan, P. J. Maddon, R. A. Koup, J. P. Moore, and W. A. Paxton.** 1996. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. Nature **381:**667–673.
13. **Dybul, M., M. Daucher, M. A. Jensen, C. W. Hallahan, T. W. Chun, M. Belson, B. Hidalgo, D. C. Nickle, C. Yoder, J. A. Metcalf, R. T. Davey, L. Ehler, D. Kress-Rock, E. Nies-Kraske, S. Liu, J. I. Mullins, and A. S. Fauci.** 2003. Genetic characterization of rebounding human immunodeficiency virus type 1 in plasma during multiple interruptions of highly active antiretroviral therapy. J. Virol. **77:**3229–3237.
14. **Fouchier, R. A., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker.** 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. **66:**3183–3187.
15. **Hoffman, N. G., F. Seillier-Moiseiwitsch, J. Ahn, J. M. Walker, and R. Swanstrom.** 2002. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. J. Virol. **76:**3852–3864.
16. **Jensen, M. A., F. S. Li, A. B. van 't Wout, D. C. Nickle, D. Shriner, H. X. He, S. McLaughlin, R. Shankarappa, J. B. Margolick, and J. I. Mullins.** 2003. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 Env V3 loop sequences. J. Virol. **77:**13376–13388.
17. **Jensen, M. A., and A. B. van 't Wout.** 2003. Predicting HIV-1 coreceptor usage with sequence analysis. AIDS Rev. **5:**104–112.
18. **Koot, M., I. P. Keet, A. H. Vos, R. E. de Goede, M. T. Roos, R. A. Coutinho, F. Miedema, P. T. Schellekens, and M. Tersmette.** 1993. Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. Ann. Intern. Med. **118:**681–688.
19. **Maas, J. J., S. J. Gange, H. Schuitemaker, R. A. Coutinho, R. van Leeuwen, and J. B. Margolick.** 2000. Strong association between failure of T cell homeostasis and the syncytium-inducing phenotype among HIV-1-infected men in the Amsterdam Cohort Study. AIDS **14:**1155–1161.
20. **Milich, L., B. Margolin, and R. Swanstrom.** 1993. V3 loop of the human immunodeficiency virus type 1 Env protein: interpreting sequence variability. J. Virol. **67:**5623–5634.
21. **Milich, L., B. H. Margolin, and R. Swanstrom.** 1997. Patterns of amino acid variability in NSI-like and SI-like V3 sequences and a linked change in the CD4-binding domain of the HIV-1 Env protein. Virology **239:**108–118.
22. **Nelson, J. A., S. A. Fiscus, and R. Swanstrom.** 1997. Evolutionary variants of the human immunodeficiency virus type 1 V3 region characterized by using a heteroduplex tracking assay. J. Virol. **71:**8750–8758.
23. **Pillai, S., B. Good, D. Richman, and J. Corbeil.** 2003. A new perspective on V3 phenotype prediction. AIDS Res. Hum. Retrovir. **19:**145–149.
24. **Ping, L. H., J. A. Nelson, I. F. Hoffman, J. Schock, S. L. Lamers, M. Goodman, P. Vernazza, P. Kazembe, M. Maida, D. Zimba, M. M. Goodenow, J. J. Eron, S. A. Fiscus, M. S. Cohen, and R. Swanstrom.** 1999. Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. J. Virol. **73:**6271–6281.
25. **R Development Core Team.** 2004. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
26. **Resch, W., N. Hoffman, and R. Swanstrom.** 2001. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. Virology **288:**51–62.
27. **Richman, D. D., and S. A. Bozzette.** 1994. The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression. J. Infect. Dis. **169:**968–974.
28. **Scarlatti, G., E. Tresoldi, A. Bjorndal, R. Fredriksson, C. Colognesi, H. K. Deng, M. S. Malnati, A. Plebani, A. G. Siccardi, D. R. Littman, E. M. Fenyo, and P. Lusso.** 1997. In vivo evolution of HIV-1 co-receptor usage and sensitivity to chemokine-mediated suppression. Nat. Med. **3:**1259–1265.
29. **Schneider, T. D., and R. M. Stephens.** 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18:**6097–6100.
30. **Shioda, T., J. A. Levy, and C. Cheng-Mayer.** 1991. Macrophage and T cell-line tropisms of HIV-1 are determined by specific regions of the envelope gp120 gene. Nature **349:**167–169.
31. **Stamatatos, L., and C. Cheng-Mayer.** 1993. Evidence that the structural conformation of envelope gp120 affects human immunodeficiency virus type 1 infectivity, host range, and syncytium-forming ability. J. Virol. **67:**5635–5639.
32. **Tscherning, C., A. Alaeus, R. Fredriksson, A. Bjorndal, H. Deng, D. R. Littman, E. M. Fenyo, and J. Albert.** 1998. Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. Virology **241:**181–188.
33. **van Rij, R. P., H. Blaak, J. A. Visser, M. Brouwer, R. Rientsma, S. Broersen, A. M. de Roda Husman, and H. Schuitemaker.** 2000. Differential coreceptor expression allows for independent evolution of non-syncytium-inducing and syncytium-inducing HIV-1. J. Clin. Investig. **106:**1039–1052.
34. **Yi, Y., F. Shaheen, and R. G. Collman.** 2005. Preferential use of CXCR4 by R5X4 human immunodeficiency virus type 1 isolates for infection of primary lymphocytes. J. Virol. **79:**1480–1486.