

# A RELIABLE PHENOTYPE PREDICTOR FOR HIV-1 SUBTYPE C BASED ON ENV V3 SEQUENCES

M. Coetzer<sup>\*1</sup>, M. A. Jensen<sup>2</sup>, A. B. van 't Wout<sup>3</sup>, L. Morris<sup>1</sup>, J. I. Mullins<sup>3</sup>

National Institute for Communicable Diseases, Johannesburg, South Africa<sup>1</sup>; Emory University, Atlanta, GA<sup>2</sup>; University of Washington, Seattle, WA<sup>3</sup>

## INTRODUCTION

Transmitted viruses generally use CCR5 as a coreceptor for entry into host cells. In many HIV-1 infected individuals, genetic changes arise that allow the virus to use other coreceptors, CXCR4 in particular. The use of both CCR5 and CXCR4 as coreceptors during host cell entry in HIV-1 subtypes A, B and D is well known. In these subtypes, there is a strong association between CXCR4 usage and accelerated disease progression. HIV-1 can be classified according to coreceptor usage, which correlates strongly with the ability of the virus to induce syncytia in the MT-2 cell line. Viruses using CCR5 exclusively (also known as R5 viruses) are typically non-syncytium inducing (NSI), while viruses that use CXCR4, either exclusively (X4 viruses), or along with CCR5 (R5X4), are syncytium inducing (SI). In subtype B, more than 50% of patients develop X4 viruses, but in subtype C, responsible for 42% of global infections, CXCR4 usage is rarely seen. The reason for X4 development is uncertain. It is not clear, for instance, whether X4 viruses are the cause or consequence of disease progression, how common X4 viruses are *in vivo*, or whether there are virological and/or immunological constraints selecting against these viruses early in infection. The infrequency of X4 viruses in subtype C suggests virological differences compared to other subtypes, but there have been no cohort studies on subtype differences, that relate to disease progression or coreceptor usage. Clearly, X4 viruses are not required for disease progression in subtype C infections. This gives rise to important questions. Is pathogenesis in subtype C infections fundamentally different from that of subtype B infections? If not, does the rarity of X4 virus in progressive subtype C infection support the idea that X4 development in B infections is a by-product and not a cause of end-stage disease? The ability to screen large subtype C-infected cohorts for X4 viruses, and relate their presence to disease status, is vital to a systematic attack on such questions. However, coreceptor phenotype assays are expensive, particularly in the developing countries where subtype C predominates. A reliable phenotype prediction method, based on sequence, could provide for rapid and less expensive screening.

Studies have shown that certain mutations in *env*, specifically the third variable region (V3) are associated with the MT-2 phenotype and coreceptor usage of HIV-1. This region plays an integral part in virus infectivity, and variations of the region have been correlated with changes in cell tropism, syncytia induction, as well as the progression of disease. The V3 region consists of approximately 35 amino acids, with a conserved disulfide bridge at the base of the loop. As in other subtypes where X4 viruses have been described, there is a high degree of amino acid variation within the V3 region of subtype C viruses between the different phenotypes, that could be used to predict viral tropism (Figure 1).

Bioinformatic approaches predicting viral tropism have increased the understanding of R5 to X4 transition and the evolution of X4 viruses. Most of these prediction methods have been developed using subtype B sequences, and although they have been applied to some subtype C isolates, they do not perform as well. We hypothesized that, using V3 sequences of C isolates of known phenotype, we could develop a reliable C-specific phenotype predictor.

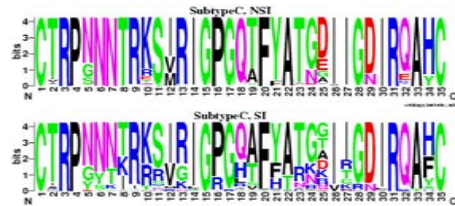


Figure 1: Graphical distribution of amino acids at V3 sites. (Top) Subtype C NSI V3 sequences and (bottom) subtype C SI V3 sequences. Note that, for example, the NSI phenotype mostly has a GPGR crown, whereas the SI phenotype has more variability within this region, particularly at positions 11 and 25.

## MATERIALS AND METHODS

**Sequences.** A training set was compiled of HIV-1 subtype C sequences with known phenotype, containing 229 R5 V3 sequences (from 200 subjects) and 51 X4 V3 sequences (20 subjects), from GENBANK and unpublished work.

**Previously described predictors.** A subset of this dataset representing only unique sequences (200 R5 and 23 X4 using) was applied to other available prediction methods: the 11/25 rule to distinguish between NSI and SI-like viruses (Fouchier *et al.* 1992); a multiple regression method based on positive, negative and net V3 charge (Briggs *et al.* 2000); a machine-learning method (Pillai *et al.* 2003), and the B-PSSM (Jensen *et al.* 2004). Due to limited numbers this dataset was divided into NSI (R5) viruses and SI (R5X4 and X4) viruses.

**C-PSSM.** We derived predictors from position specific scoring matrices (PSSM) based on the subtype C training set of 280 V3 sequences. Distributions of specificity (correct R5 prediction) and sensitivity (correct X4 prediction) were estimated by combining dataset bootstrapping with leave-one-out cross-validation, as follows. In this procedure, the target sequence was removed from the data set, and the remaining sequences randomly sampled with replacement from each phenotype category. The random sample was used to calculate a PSSM predictor, and used to predict the phenotype of the target. Resampling was repeated 100 times to obtain an empirical prediction distribution. Each sequence in the data set was treated as a target in turn. The prediction distribution for each sequence was then sampled to obtain a single instance of sensitivity and specificity. This sampling was repeated 100 times to obtain the final empirical distributions. Non-independence of sequences was attenuated by randomly sampling single sequences from individuals on each bootstrap iteration.

**V3-HTA.** V3-based heteroduplex tracking assays (V3-HTA) were performed on 12NSI and 8 SI subtype C viral isolates, using a known subtype C R5 as probe. We compared mobilities to PSSM scores using linear regression.

## RESULTS

The performance of available phenotype predictors was tested on a subset of subtype C sequences (Table 1). All the predictors had a high specificity (99.5%) in predicting the NSI phenotype correctly, except the method used by Briggs, predicting only 52% of the NSI sequences. Generally the sensitivity of these methods when applied to SI sequences was very low, with the Briggs method lowest.

Table 1: Comparison of some phenotype predictors on subtype C sequences with known phenotype.

Method	% Phenotype correctly predicted	
	NSI (n=200)	SI (n=23)
11/25	99.5	47.8
Briggs	52	34.8
Pillai (SVM)	99.5	52
B-PSSM	99.5	52

In order to increase the sensitivity and specificity of predicting subtype C phenotype, we developed a subtype C specific predictor using position specific scoring matrix (PSSM). The C-specific predictor (C-PSSM) had a specificity of 91% [C.I. 89%-93%] and sensitivity of 84% [C.I. 79%-89%], based on leave-one-out bootstrap using all sequences. The performance of the subtype C predictor was compared to the subtype B predictor using the subtype B dataset as describe in Jensen *et al.* 2003, as shown in Figure 2.

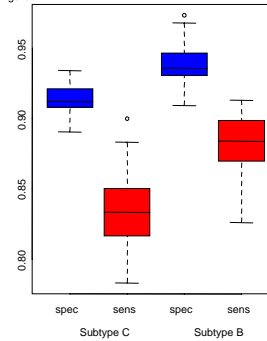


Figure 2: Leave-one-out/bootstrap predictions of subtype C compared to subtype B, indicating that subtype C sequences perform comparably well in the C-PSSM as the subtype B sequences does in the B-PSSM.

Analysis based on single sequences per individual gave comparable specificity (94% [92%-96%]) and somewhat lower sensitivity (75% [68%-82%]). These are similar to B-PSSM results on subtype B targets (median specificity 93%; sensitivity 88%). The C-PSSM is significantly more sensitive than the B-PSSM applied to the C sequences (B-PSSM on C, sensitivity: 37%, data not shown). Note that [in the single-sequence analysis] the median sensitivity and specificity appear to approach a limit as the number of unique patients sampled increases (Figure 3). This suggests that sampling more patients may not improve the prediction quality by a large amount.

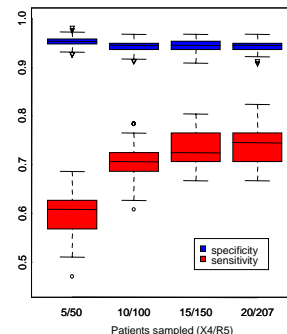


Figure 3: Random sampling in proportion to R5 vs X4 in dataset of single subtype C sequences per patient.

## C-PSSM and V3-HTA

Heteroduplex tracking assay (HTA) has been used to identify SI and NSI-like viruses when hybridized with a R5 probe (Nelson *et al.* 1997). Samples with a similar sequence to the probe have a low mobility ratio and are usually NSI, whereas samples with a high mobility ratio are different from the probe and are frequently SI. The mobility ratio of 8 SI and 12 NSI samples were compared to their PSSM score. There was a strong correlation between the mobility ratio and PSSM score of samples ( $p < 0.0001$ ;  $r_2: 0.56$ ).

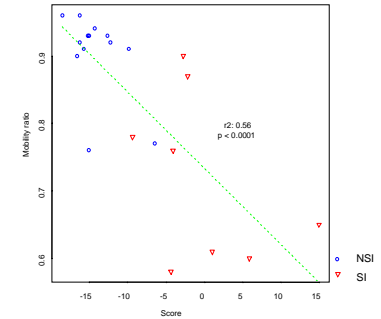


Figure 4: Comparison between the HTA mobility ratio vs. PSSM score of subtype C NSI and SI samples. Mobility ratios were deduced from the distance the heteroduplex vs homoduplex separated in a gel.

## CONCLUSION

The PSSM has proven to be an effective predictor of biological coreceptor phenotype, and has contributed to a better understanding of the role of intermediates (R5X4) in the transition of R5 to X4 in subtype B. We derived a C-specific phenotype predictor that performs nearly as well on C V3 loops as do existing B-specific methods on B V3 loops, and can thus be applied to characterize subtype C V3 sequences of unknown coreceptor usage. Other methods are very specific in predicting NSI tropism, but not very sensitive in detecting SI phenotype. Although the number of SI viruses available is a limiting factor in improving the performance of prediction methods for subtype C viruses, this study has shown that currently available data provide a good initial basis for subtype C coreceptor usage prediction. We also found that V3-HTA mobilities correlated well with PSSM score and phenotype. This technique could allow inexpensive scoring and prediction for large numbers of subtype C V3 samples.

## REFERENCES AND ACKNOWLEDGEMENTS

- \*Briggs D, Tuttle D, Slessman J, Goodenow M. Envelope V3 amino acid sequence predicts HIV-1 phenotype (coreceptor usage and tropism for macrophages). *AIDS* 2000, 14:2837-9.
- \*Fouchier R, Groenik M, Kootstra N, *et al.* Phenotype associated sequence variation in the third variable domain (V3) of HIV type 1 gp120 molecule. *J Virol* 1992, 66:3183-7.
- \*Jensen MA, Li F, Van't Wout AB, Nickle DC *et al.* Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of HIV-1 *env* V3 loop sequences. *J Virol* 2003, 77:13376-13388.
- \*Jensen MA and Van't Wout AB. Predicting HIV-1 coreceptor usage with sequence analysis. *AIDS Rev* 2003, 5:104-112.
- \*Nelson J A, Fiscus S A and Swanstrom R. Evolutionary variants of the human immunodeficiency virus type 1 V3 region characterized by using a heteroduplex tracking assay. *J Virol* 1997, 71:8750-8.
- \*Pillai S, Good B, Richman D, Corbell J. A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses* 2003, 19:145-9.

This work was conducted during a visit to Dr Jim Mullin's laboratory funded by South African Fogarty Training and Research Program.