# Primer ID Deep Sequencing of HIV-1 *env* Region to Reveal Genetic Structure of Viral Population

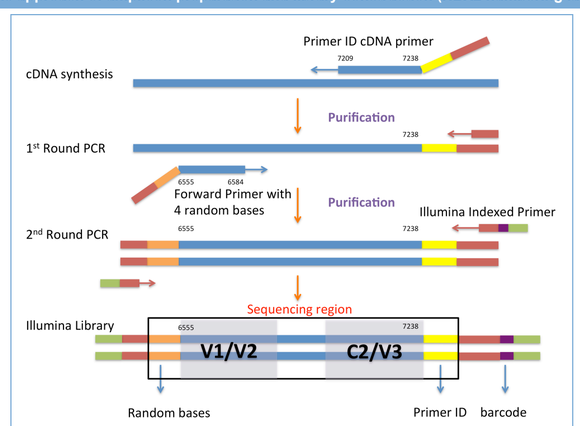
**Shuntai Zhou**[1], Maria Bednar[1], Christa Sturdevant[1,2], Serena Spudich[3], Richard Price[4], Ronald Swanstrom[1,5,6]

1. Lineberger Cancer Center, 2. Department of Microbiology and Immunology, 5. Department of Biochemistry and Biophysics, 6. UNC Center For AIDS Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 3. Department of Neurology, Yale University, School of Medicine, New Haven, Connecticut, 4. Department of Neurology, University of California at San Francisco, School of Medicine, San Francisco, California, USA.

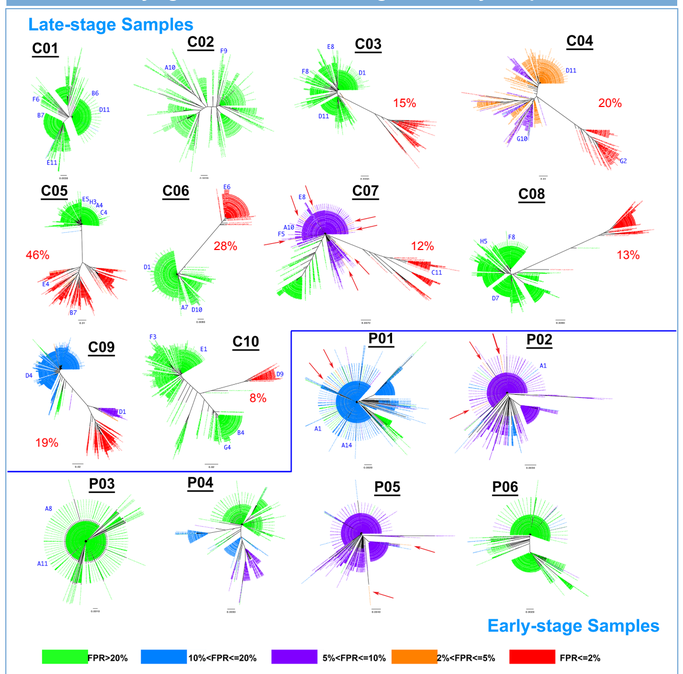
## Abstract

**Background**: HIV env is the major determinant of entry phenotype. We use a random sequence tag (Primer ID) to tag each cDNA molecule prior to PCR and next generation sequencing (NGS) to reduce PCR and sequencing errors and provide accurate frequencies of minority variants. We sequenced the V1-V3 region of env to detect X4 variants and define the genetic structure of the viral population. **Methodology:** Viral RNA was extracted from plasma obtained from 10 late-stage and 6 early-stage. We used MiSeq paired-end sequencing from the ends of a V1-V3 amplicon. Entry tropism was predicted by V3 sequences using Gene2Pheno. Entry assays were done using SGA env amplicon clones to confirm phenotypes. A control plasmid was amplified, sequenced, and analyzed along with clinical samples. **Results**: Sequencing data from the control plasmid showed that the residual error rate of this approach was 1 error in 4,000 bases. The number of consensus sequences was linearly correlated with template numbers (R2 = 0.999). X4 viruses were detected in 8 of 10 late-stage samples but none in early-stage. Phylogenetic analysis of C2/V3 sequences revealed that the most stringently called X4 viruses (Geno2Pheno, FPR <= 2%) formed distinct lineages in samples containing X4 viruses. Entry phenotype analysis suggested that FPR <= 2% were true X4 lineages. Analysis of both V1/V2 and C2/V3 regions showed very little recombination between the X4 and R5 lineages, while within some R5 lineages or X4 lineages there could be extensive recombination. **Conclusions:** Primer ID combined with NGS provides an improved approach to deep sequencing highly variable regions of HIV-1. Clustering of sequences with low FPR of X4 is an additional type of information to predict X4 calls based on sequence analysis. Restricted recombination of X4 and R5 viruses suggest these variants are genetically isolated either producing nonviable recombinants or replicating in distinct cell types.

## The approach of amplicon preparation and library construction (HBX2 numbering labeled)




## Phylogenetic trees of C2/V3 region of study samples



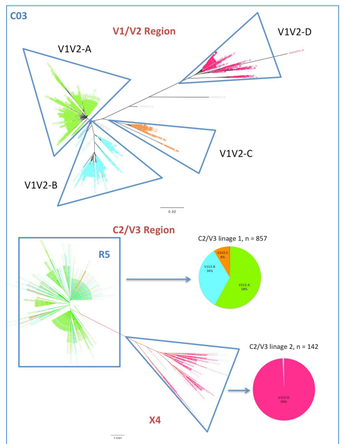

Showing Different level of Geno2pheno$_{[CORECEPTOR]}$ false positive rate (FPR) among viral population. Color of taxa names indicates level of FPR: red, FPR <= 2%; orange, FPR 2-5%; purple, FPR 5-10%; blue, FPR 10-20%; green, FPR > 20%. Percentage of X4 lineage (FPR <= 2%) in the viral population are also listed

Red arrows indicate sequences with relatively low FPR scores among lineages with intermediate FPR scores.

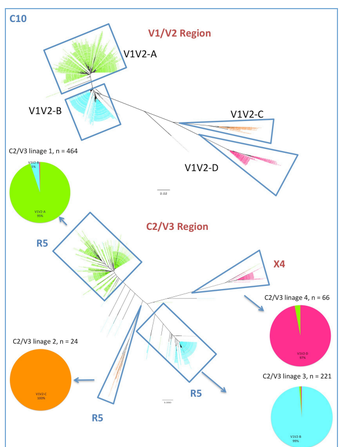
Names in blue indicate the identical or closest sequences in the tree of the clones that we obtained from single genome amplification and were examined for entry phenotypes *in vitro*.

- **The most stringently called X4 viruses (Geno2Pheno, FPR <= 2%) formed distinct lineages in samples containing X4 viruses.**
- **Entry phenotype analysis agreed with the most genotype prediction (Geno2Pheno, FPR <= 2%)**

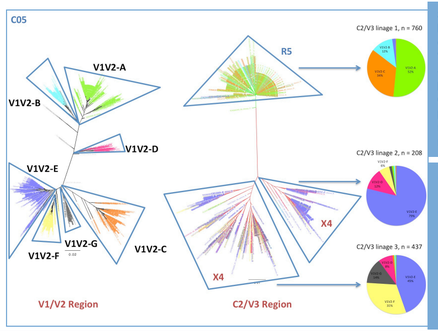
## Linages of V1/V2 region and C2/V3 region and their recombination




In sample C03, there is a distinct lineage in the V1/V2 tree (V1V2-D) linked to the X4 C2/V3 lineage, and the three other V1/V2 lineages form the major R5 lineage in the C2/V3 tree. Thus recombinants between the X4 and R5 viruses between V1 and V3 were extremely rare.



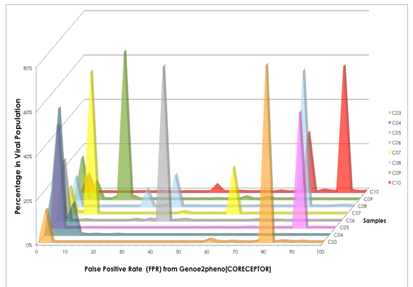

C10, there were 4 major V1/V2 lineages and each one was strongly associated with a separate C2/V3 lineage, three of which were R5 and one X4. In this case all four lineages appeared to be replicating independently




Sample C05 had seven distinct V1/V2 lineages (A-G), of which 3 were associated with the major R5 virus lineage (lineage V1V2-A, B, and C with C2/V3 lineage 1), and 4 were associated with the two X4 virus lineages (lineage V1V2-D, E, F, and G with C2/V3 lineages 2 and 3); there appeared to be extensive mixing/recombination between the two X4 lineages but very little between the X4 lineages and the R5 lineage.

- **Low recombination of X4 and R5 lineages at V1V2 region**
- **Recombination of X4 or R5 lineages by themselves observed**
- **Suggesting independent replication of X4 and R5 lineages, i.e. compartmentalization of X4 and R5 viruses in the lymphoid system**
- **Or, producing nonviable recombinants**

## FPR distribution among viral populations




**Sequence populations were clustered in their FPR values, with there being a wide range of populations with intermediate values and what appeared to be a tight clustering of low FPR values for those predicted to be X4 variants, suggesting that these sequence variants represent distinct lineages within the host rather than a continuum of variability.**

## Conclusions

- Primer ID combined with NGS provides a novel approach to deep sequencing highly variable region of HIV-1 with low error rate and accurate allelic frequency.
- Clustering of sequences with low FPR of X4 is an additional type of information to predict X4 calls based on sequence analysis.
- Restricted recombination of X4 and R5 viruses suggests these variants are genetically isolated either producing nonviable recombinants or replicating in distinct cell types.

## Acknowledgements

We thank our study subjects and clinical investigators. We thank UNC CFAR for the support of the study.

This study is funded by NIH grant 5R37AI044667-13 and grant 1R21AI108539-01.

shuntai.zhou@unc.edu