Optimization of PCR Amplicons to Predict Clonality of Full-Length HIV-1 Sequences

339

Sarah B. Laskey[1], Christopher W. Pohlmeyer[1], Oliver Elliott[2], Katherine M. Bruner[1] & Robert F. Siliciano[1,3]

[1]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
[2]Department of Biomedical Informatics, Columbia University Medical Center, New York, New York, USA
[3]Howard Hughes Medical Institute, Baltimore, Maryland, USA

Introduction

The proliferation of latently infected cells may lead to the persistence of HIV-1 in infected individuals despite suppressive antiretroviral therapy. Because the HIV-1 replication process is highly error-prone, the detection of identical viral genomes in distinct host cells provides evidence for clonal expansion. Although it is common to amplify and sequence short, subgenomic regions of the viral genome for phylogenetic analysis, sequence identity of these amplicons does not guarantee clonality across the full viral genome. On the contrary, the sequence diversity in a subgenomic region is best understood as a minimum estimate of the total sequence diversity in the sample analyzed. We show that although longer amplicons capture more diversity, no subgenomic region can recapitulate the diversity of full viral genomes (middle panel). Consequently, some identical subgenomic amplicons should be expected even from the analysis of completely unique viral genomes, and the presence of identical amplicons alone is not proof of clonally expanded HIV-1. Subgenomic sequences are contextualized by the primer set used to generate them, and we present here a strategy for the evaluation of sequences in the context of specific primer sets (right panel).

Methods: Clonal Prediction Score

Clonal prediction score (CPS) quantifies how well a primer set is able to distinguish identical HIV-1 genomes. The CPS is equal to the fraction of unique genomes that are correctly identified as unique using only the region amplified by that particular primer set.

Table 1. Sequence alignments analyzed to calculate clonal prediction scores. Each alignment is composed of unique, full-length sequences isolated from a single HIV-infected individual.

Sample	Source	Reference	PMID	Number of subjects	Sequences per subject (range)
Acute treated – DNA	DNA from resting CD4$^+$ T cells of subjects treated during acute infection	Bruner *et al.*, manuscript under review, CROI Abstract 83	-	6	7-14
Chronic treated – DNA	DNA from resting CD4$^+$ T cells of subjects treated during chronic infection	Bruner *et al.*, manuscript under review, CROI Abstract 83	-	5	8-13
VOA – DNA	DNA from p24 antigen-negative quantitative viral outgrowth assay culture wells	Ho *et al.*, 2013, Cell.	24243014	5	6-12
Longitudinal – RNA	Plasma RNA from 2-10 longitudinal samples per subject	Herbeck *et al.*, 2011, J. Virol.	21593162	6	28-121
Acute – RNA	Plasma RNA collected during acute infection	Salazar-Gonzalez *et al.*, 2009, J. Exp. Med.	19487424	9	5-14

How to calculate the CPS for a primer set (Fig. 1):

- Start with an alignment of unique, full-length HIV-1 sequences from a single subject (Table 1).
- Determine the amplicons that would be produced in a PCR using a given primer set, using that alignment of sequences as templates. (Sequences with gaps overlapping the primer sites will not be amplified.)
- Count the number of unique amplicons (in the Fig. 1 example, 2) and the total number of amplicons produced (in the Fig. 1 example, 7). The ratio of these numbers multiplied by 100 is equal to the CPS. In the example in Fig. 1, the CPS = 28.6.
- The maximum possible CPS is 100, when all amplicons produced are unique. The minimum possible CPS is 100/N, where N is the number of amplicons produced, in the case when all amplicons produced are identical.
- The CPS depends both on the sequence alignment and the primer set analyzed.

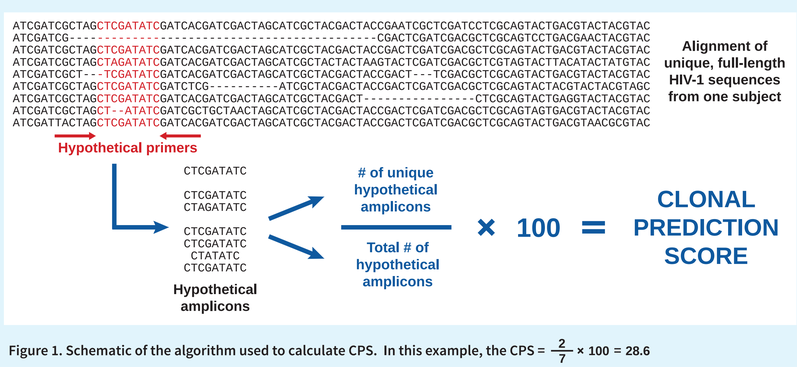


Figure 1. Schematic of the algorithm used to calculate CPS. In this example, the CPS = $\frac{2}{7} \times 100 = 28.6$

ACKNOWLEDGMENTS AND FUNDING

The authors would like to thank Dr. Christine Durand, Dr. Andrew Redd, Dr. Stuart Ray, and Dr. Daniel Rosenbloom for helpful discussions and advice. This work was supported by the Martin Delaney CARE and DARE Collaboratories (NIH grants AI096113 and 1U19AI096109), by an ARCHE Collaborative Research Grant from the Foundation for AIDS Research (amfAR 108165-50-RGRL), by the Johns Hopkins Center for AIDS Research (P30AI094189), by NIH grant 43222, and by the Howard Hughes Medical Institute and the Bill and Melinda Gates Foundation.

REFERENCES

(1) Bruner K. et al. Rapid accumulation of defective proviruses complicates HIV-1 reservoir measurements. *Submitted to Nat. Med.* in December 2015, under review in January 2016. See CROI Abstract 83. **(2)** Ho, Y-C. et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–51 (2013). **(3)** Herbeck, J. T. et al. Demographic Processes Affect HIV-1 Evolution in Primary Infection before the Onset of Selective Processes. *J. Virol.* **85**, 7523–7534 (2011). **(4)** Salazar-Gonzalez, J. F. et al. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J. Exp. Med.* **206**, 1273–89 (2009). **(5)** Evering, T. H. et al. Absence of HIV-1 evolution in the gut-associated lymphoid tissue from patients on combination antiviral therapy initiated during primary infection. *PLoS Pathog.* **8**, e1002506 (2012). **(6)** Palmer, S. et al. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J. Clin. Microbiol.* **43**, 406–13 (2005). **(7)** Bailey, J. R. et al. Residual human immunodeficiency virus type 1 viremia in some patients on antiretroviral therapy is dominated by a small number of invariant clones rarely found in circulating CD4+ T cells. *J. Virol.* **80**, 6441–57 (2006). **(8)** Delwart, E. L. et al. Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes. *Science* **262**, 1257–61 (1993). **(9)** Wagner, T. A. et al. An increasing proportion of monotypic HIV-1 DNA sequences during antiretroviral treatment suggests proliferation of HIV-infected cells. *J. Virol.* **87**, 1770–8 (2013). **(10)** Palmer, S. et al. New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. *J. Clin. Microbiol.* **41**, 4531–6 (2003). **(11)** Maldarelli, F. et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014). **(12)** Kearney, M. F. et al. Lack of Detectable HIV-1 Molecular Evolution during Suppressive Antiretroviral Therapy. *PLoS Pathog.* **10**, e1004010 (2014). **(13)** Wagner, T. A. et al. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–3 (2014).

Results I: Which primer sets are best?

Are certain regions of the HIV-1 genome better able to differentiate unique, full-length sequences?

To investigate how CPS varies across the HIV-1 genome and across infected individuals, we calculated CPS values for hypothetical 1 kb amplicons spanning the HIV-1 genome at 10 bp intervals. These amplicons were defined by hypothetical 10 bp forward and reverse primers. While the amplicons defined by these hypothetical primers have a length of 1 kb in the HXB2 reference genome, they may have different lengths in sequences containing insertions or deletions between the primer sites. Along with sequence differences, length polymorphisms can be used to differentiate unique sequences. The results of this analysis for six Acute treated – DNA samples (Table 1) are shown in Fig. 2a. There was **substantial variation among subjects**, suggesting that a primer set optimized to differentiate sequences in one sample may not be optimal in another sample from a different subject.

CPS values spanning the HIV-1 genome, averaged over all of the subjects within each sample group, are shown in Fig. 2b. The purpose of these averaged scores was to determine whether any region of the viral genome yields consistently higher or lower CPS than other regions across different subjects or sample types. CPS values averaged over several subjects are **consistent across the viral genome** for all five sample types evaluated.

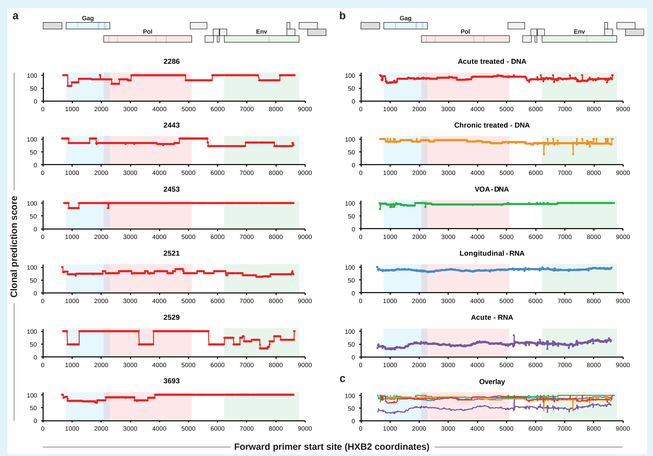


Figure 2. Clonal prediction scores of hypothetical 1 kb amplicons spanning the HIV-1 genome. Schematics of the HIV-1 genome are aligned to charts to highlight viral gene locations. **(a)** CPS of 1 kb-wide amplicons spanning the HIV-1 genome for six Acute treated - DNA subjects. **(b)** CPS of 1 kb-wide amplicons spanning the HIV-1 genome, averaged over all subjects in five sample categories (see Table 1) **(c)** Overlay of the five plots in part b.

How is the length of an amplicon related to its capacity to differentiate unique, full-length sequences?

To determine the effect of amplicon length on CPS, we calculated the CPS for hypothetical amplicons of different lengths spanning the viral genome. We summarized the analysis of all amplicons of a given length by taking the average CPS over all amplicons of that length across the genome (Fig. 3). For example, all of the points plotted in the top chart from Fig. 2a are averaged together to yield the subject 2286, 1000 bp data point in the Acute treated – DNA Average CPS plot in Fig. 3. Individual results for the 31 subjects are grouped by sample type. For every subject, regardless of sample type, longer amplicons yield higher CPS. This is unsurprising, as a longer amplicon should be more likely than a short amplicon to contain sequence diversity. However, the precise relationship between amplicon length and CPS varies dramatically by sample type.

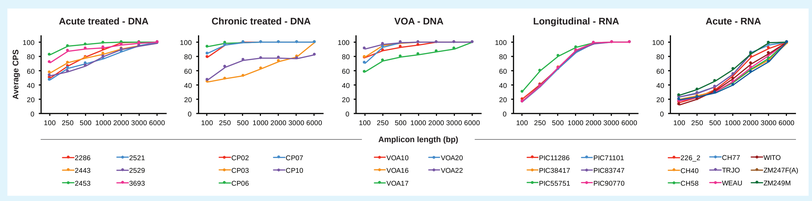


Figure 3. Relationship between amplicon length and CPS. Average CPS value over all amplicons of a given length, spanning the HXB2 reference genome at 10 bp intervals. Amplicons with undefined CPS were not included in the average.

Conclusions I:

- There is substantial variation among subjects; no single subgenomic amplicon can distinguish HIV-1 genomes across a variety of sample types or subjects.
- There is no optimal region of the genome best suited for differentiating unique sequences.
- Longer amplicons distinguish unique HIV-1 genomes better than shorter amplicons.

Results II: How to interpret sequence data?

We calculated the CPS for eight primer sets used in published studies of HIV-1 diversity, evolution, and clonal expansion (Table 2) using the sequence alignments from Table 1. Figure 4 shows the results, averaged over all subjects for each sample type.

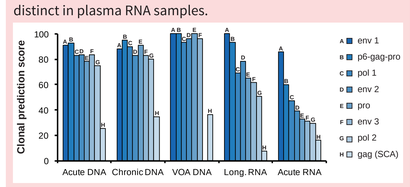


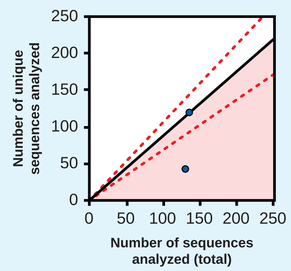
Figure 4. Clonal prediction scores of previously published primer sets. CPS calculated for each of the primer sets listed in Table 2, averaged over all subjects within each sample type (Table 1). Consistent with the results in Fig. 3, larger amplicons typically have higher CPS values, and this relationship is most distinct in plasma RNA samples.

Table 2. Primer sets evaluated using CPS analysis

Primer set	Reference	PMID	Amplicon length (bp)	Amplicon HXB2 coordinates
env 1	Evering *et al.*, 2012, PLoS Pathog.	22319447	2948	5956 – 8903
p6-gag-pro	Palmer *et al.*, 2005, J. Clin. Microbiol.	15635002	1565	1870 – 3434
pol 1	Bailey *et al.*, 2006, J. Virol.	16775332	664	2598 – 3261
env 2	Delwart *et al.*, 1993, Science	8235655	667	7001 – 7667
pro	Bailey *et al.*, 2006, J. Virol.	16775332	537	2056 – 2592
env 3	Bailey *et al.*, 2006, J. Virol.	16775332	471	1059 – 7529
pol 2	Wagner *et al.*, 2013, J. Virol.	23175380	416	2200 – 2615
gag (SCA)	Palmer *et al.*, 2003, J. Clin. Microbiol.	14532178	79	1298 – 1376

CPS Plots: Using primer CPS to interpret amplicon clonality

Subgenomic amplicons cannot provide definitive evidence of full-length clonal HIV-1 genomes, even when identical amplicon sequences are detected. The CPS for a primer set indicates how much **false-positive clonality** is found by that primer set in alignments of unique genomes (Table 1). Thus the CPS for alignments of unique genomes can be understood as an **estimate of the background signal expected** when using a given primer set to evaluate clonality. We used the CPS of published primer sets to estimate the presence or absence of true clonality in published phylogenetic trees (Fig. 5).



1. Define axes:
- x-axis: Total number of sequences analyzed
- y-axis: Number of unique sequences analyzed

2. Plot primer set CPS (an estimate of false-positive clonality):
- On these axes, the CPS is plotted as a line with a slope equal to CPS/100. The maximal CPS = 100 is plotted as the diagonal line x = y.
- In the plots in Fig. 5, the solid black line shows the CPS of a primer set, calculated using the RNA or DNA alignments in Table 1 (also shown in Fig. 4). Dashed red lines show the CPS ± one standard deviation.

3. Plot sequence data alongside the CPS of the primer set used to generate the sequences:
- Count the total number of sequences in an alignment or phylogenetic tree. Count the number of unique sequences in the alignment/tree.
- Each phylogenetic tree (or sequence alignment) can be plotted as a single point on these axes. (Two points are shown in this example plot, representing two different phylogenetic trees.)
- Points plotted in the (shaded) region below the black line represent sequences with more clonality than the expected background level for that primer set.

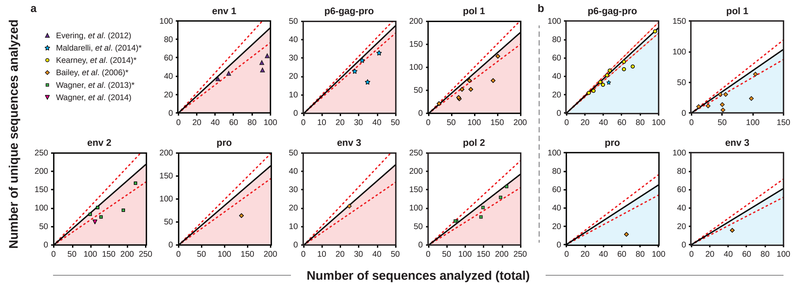


Figure 5. Using CPS plots to interpret amplicon clonality. Evaluation of previously published phylogenetic trees in the context of the CPS of the primer sets used to generate those trees. (See blue box above explaining the design and interpretation of these plots.) **(a)** Proviral DNA samples. **(b)** Plasma RNA samples. Each plotted point indicates the actual number of total sequences and unique sequences present in a published phylogenetic tree. Phylogenetic trees containing both proviral DNA and plasma RNA sequences were counted separately for the two sequence types and plotted separately in parts **(a)** and **(b)**. **References in which hypermutated sequences were not explicitly included in phylogenetic trees, artificially increasing the apparent clonality in those trees.*

Conclusions II:

- Primer set CPS dramatically impacts the apparent clonality of sequences generated using that primer set.
- The context provided by calculating CPS for a primer set is essential for evaluating evidence of clonal expansion based on subgenomic amplicons generated using that primer set.
- Identical HIV-1 sequence fragments must be validated to demonstrate full-genome clonality conclusively.