

## Accepted Manuscript

Title: Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations

Authors: Michael R. Jordan, Mary Kearney, Sarah Palmer, Wei Shao, Frank Maldarelli, Eoin P. Coakley, Colombe Chappey, Christine Wanke, John M. Coffin



PII: S0166-0934(10)00166-7  
DOI: doi:10.1016/j.jviromet.2010.04.030  
Reference: VIRMET 11227

To appear in: *Journal of Virological Methods*

Received date: 18-11-2009  
Revised date: 27-4-2010  
Accepted date: 29-4-2010

Please cite this article as: Jordan, M.R., Kearney, M., Palmer, S., Shao, W., Maldarelli, F., Coakley, E.P., Chappey, C., Wanke, C., Coffin, J.M., Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations, *Journal of Virological Methods* (2008), doi:10.1016/j.jviromet.2010.04.030

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Table 1. Patient demographic data**

<b>Patient Number</b>	<b>Gender<sup>1</sup></b>	<b>Risk Factor<sup>2</sup></b>	<b>HIV-1 RNA copies/ml</b>	<b>CD4 cells/<math>\mu</math>l</b>	<b>Estimated Year of Infection</b>	<b>Estimated Time Seroconversion to Specimen Collection (years)</b>
1	F	Unknown	99,000	NA	2000	1
2	F	HS	19,467	393	1997	3
3	M	HS	45,708	351	1997	3
4	M	MSM or HS	61,022	1,196	1999	2
5	M	MSM or HS	200,000	89	1988	12
6	M	HS	17,796	NA	1998	3
7	M	IDU	154,000	421	1988	12
8	M	IDU or HS	34,000	NA	1996	4
9	M	HS	5,323	735	1996	4
10	M	IDU	3,446	530	1996	4
11	M	IDU	54,018	135	2001	1
12	F	IDU	3,401	384	1994	6
13	F	IDU	679	293	1990	11
14	M	HS	490	677	1999	2
15	M	MSM	300,000	194	2003	0.5
16	M	MSM	34,000	NA	2003	1
17	M	IDU	750	751	1998	2

Median HIV RNA 34,000 copies/ml and median CD4 count 393 cells/ $\mu$ l. Estimated year of infection: Range (1998-2003). All specimens were obtained from July 2000 to July 2001. NA=Data not available; <sup>1</sup>F=Female; M=Male; <sup>2</sup>HS=heterosexual contact; MSM=men who have sex with men; IDU=intravenous drug user

**Table 2. Genetic distance of HIV-1 sequences derived by SGS and PCR/cloning**

<b>Patient</b>	<b>HIV RNA copies/ml</b>	<b>Average Pairwise Difference SGS, %</b>	<b>Average Pairwise Difference PCR/Cloning, %</b>	<b>Average Pairwise Difference Between Assays, %</b>	<b>p(K*s)</b>
1	99,000	1.27%	1.19%	0.08%	0.76
2	19,467	1.30%	0.87%	0.43%	0.06
3	45,708	0.52%	0.23%	0.29%	0.27
4	61,022	0.20%	0.46%	0.26%	0.024
5	200,000	2.04%	1.95%	0.09%	0.95
6	17,796	0.20%	0.40%	0.20%	0.01
7	154,000	1.95%	2.08%	0.13%	0.15
8	34,000	0.95%	0.82%	0.13%	0.67
9	5,323	0.81%	0.77%	0.04%	0.01
10	3,446	1.33%	1.20%	0.13%	0.0001 <sup>+</sup>
11	54,018	0.71%	1.20%	0.49%	0.0008 <sup>+</sup>
12	3,401	0.66%	0.63%	0.03%	0.0001 <sup>+</sup>
13	679	1.30%	1.23%	0.07%	0.034
14	490	0.74%	1.16%	0.42%	0.0031
15	300,000	0.18%	0.36%	0.18%	0.0268
16	34,000	0.32%	0.54%	0.22%	0.18
17	750	1.14%	0.99%	0.15%	0.62

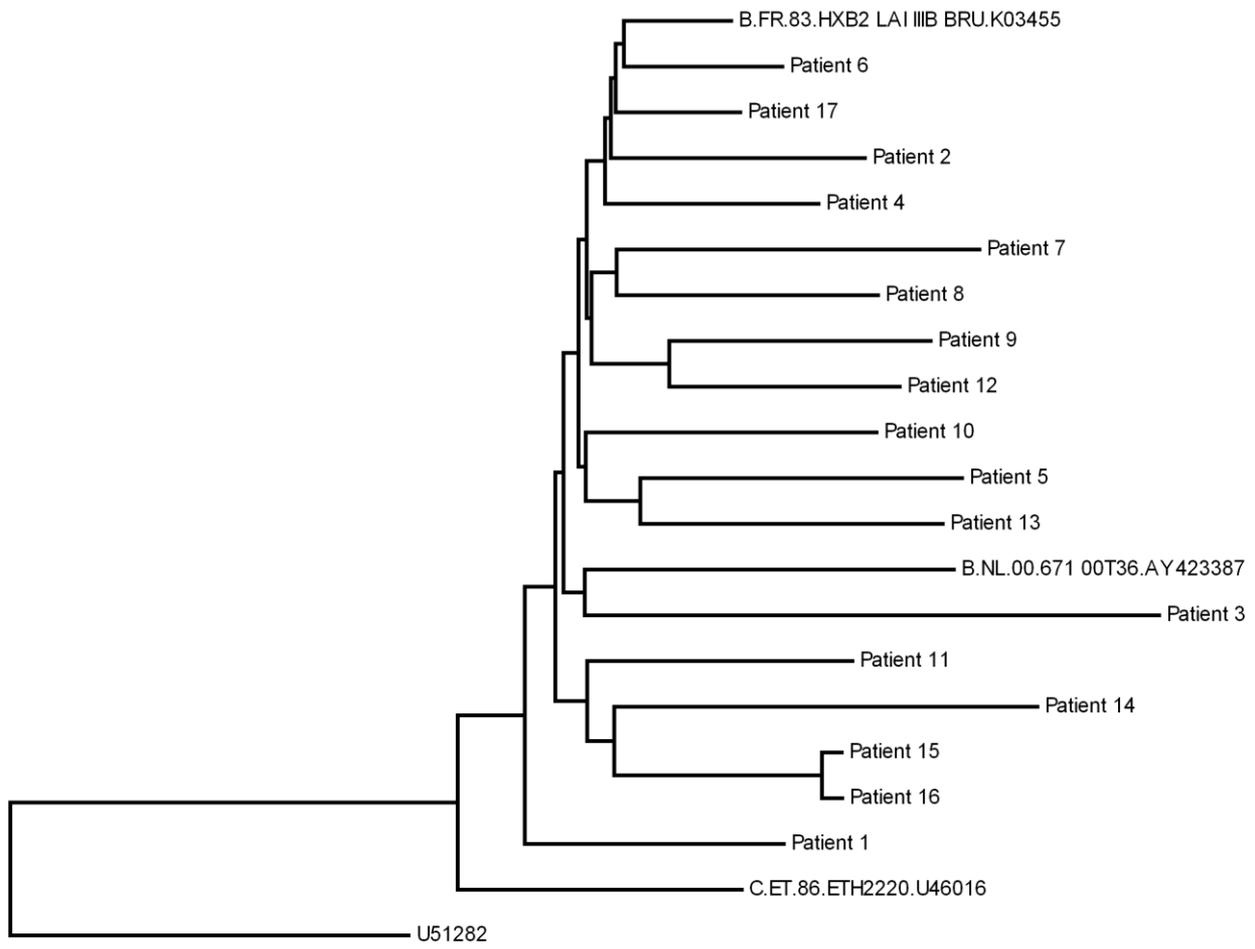
<sup>+</sup>Statistically significant difference  $p(\mathbf{K*s}) < 0.003$

**Table 3. Calculation of sequence entropy**

Patient	Nucleic Acid			Amino Acid		
	Position	Query consensus	P-value at this position	Position	Query consensus	P-value at this position
9	6 (RT)	C	0.002			
11	489 (RT)	A	0.001			
	389 (RT)	A	0.004			
12				64 (RT)	K	0.004
	543 (RT)	T	0.004			
	69 (PR)	A	0.005			
13						
	336 (RT)	A	0.005			

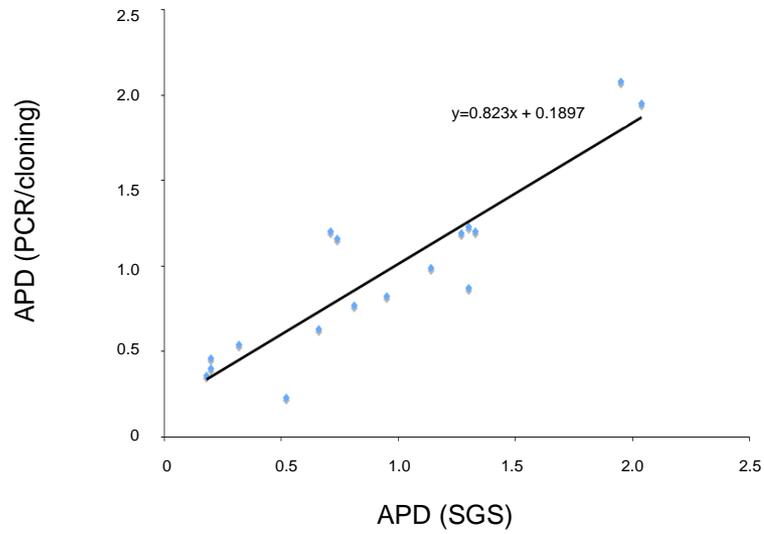
Sequences derived by cloning and sequencing were classified as background sequences and sequences derived by SGS as query sequences. One thousand randomizations were performed comparing each set of sequences with statistical significance defined as  $p < 0.005$ .

Figure(s)



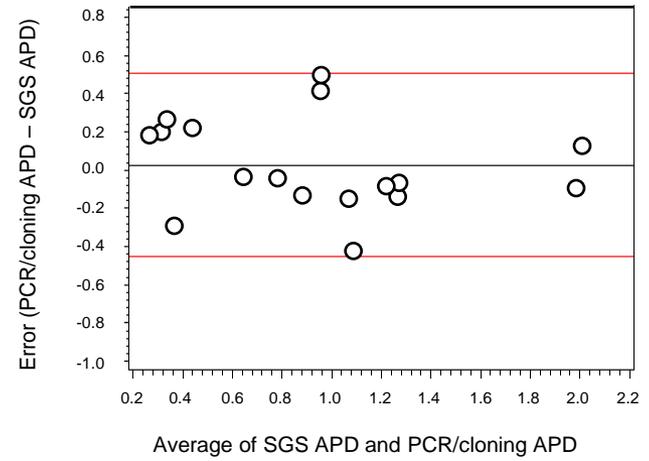
10

a Correlation of APD by SGS and PCR/cloning

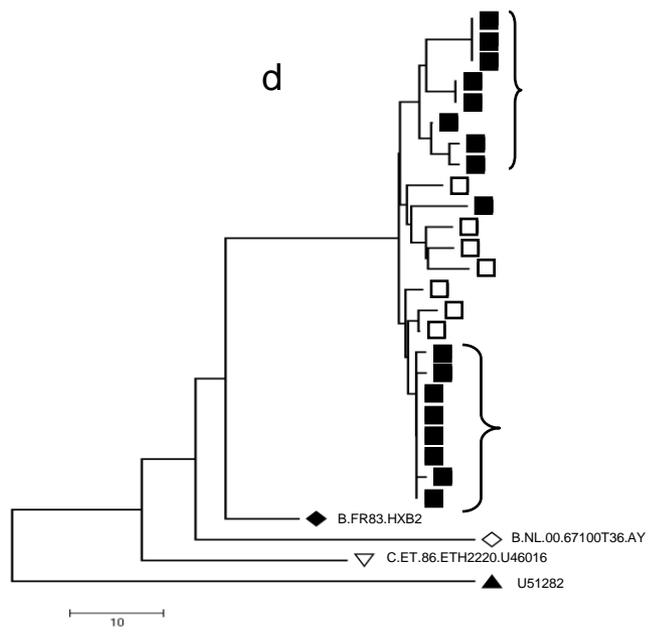
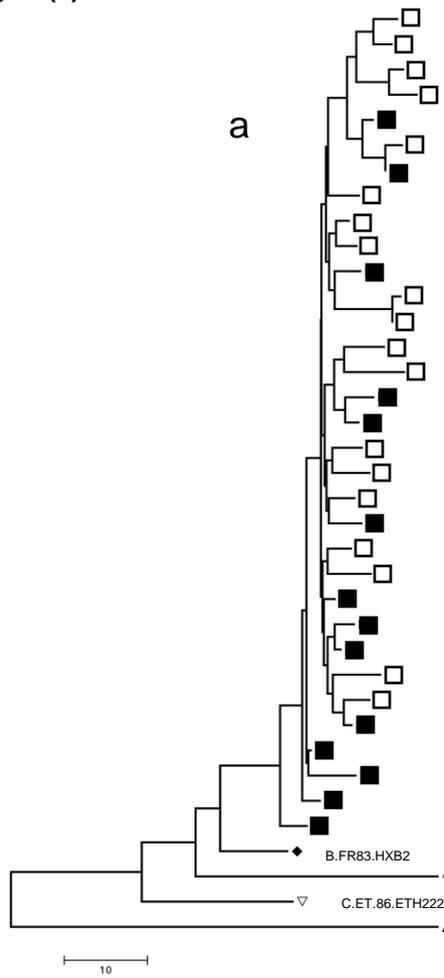


b

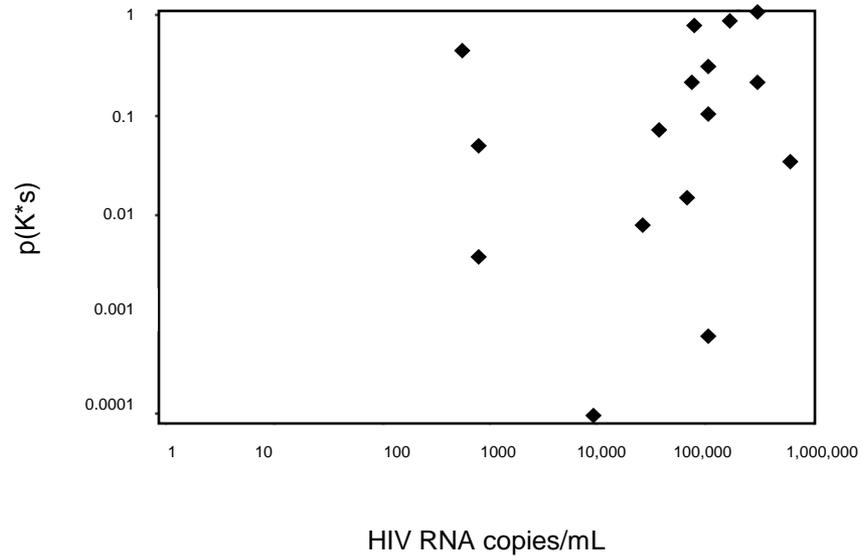
Bland Altman: Comparison line is the mean of the difference between SGS APD and PCR/cloning APD



Figure(s)



P(K\*s) vs. HIV RNA



1

2 **Title:** Comparison of standard PCR/cloning to single genome sequencing for analysis of  
3 HIV-1 populations

4

5 **Authors:** Michael R. Jordan <sup>a\*</sup>, Mary Kearney <sup>b</sup>, Sarah Palmer <sup>b,c</sup>, Wei Shao <sup>b</sup>, Frank  
6 Maldarelli <sup>b</sup>, Eoin P. Coakley <sup>d</sup>, Colombe Chappey <sup>e</sup>, Christine Wanke <sup>a</sup>, John M. Coffin  
7 <sup>a</sup>

8

9 **Affiliations:**

10 <sup>a</sup>Tufts University School of Medicine, Tufts Medical Center, Boston, MA, USA; <sup>b</sup>HIV  
11 Drug Resistance Program, National Cancer Institute, National Institutes of Health,  
12 Frederick, MD, USA; <sup>c</sup>Swedish Institute for Infectious Disease Control, Karolinska  
13 Institute, Stockholm, Sweden; <sup>d</sup>Monogram Biosciences, South San Francisco, CA, USA;  
14 <sup>e</sup>Genentech, South San Francisco, CA, USA.

15

16 **\*Corresponding Author:** Michael R. Jordan MD MPH, Division of Geographic Medicine  
17 and Infectious Disease, Tufts Medical Center, Tufts University School of Medicine, 800  
18 Washington Street, Box 41, Boston, MA, 02111, USA; [mjordan@tuftsmedicalcenter.org](mailto:mjordan@tuftsmedicalcenter.org);  
19 fax: 1-617-636-3216

20

21 **Full address of co-authors:**

22 Mary Kearney, HIV Drug Resistance Program, National Cancer Institute-Frederick, P.O.  
23 Box B, Building 535, Frederick, MD 21702-1201, USA; [kearney@ncifcrf.gov](mailto:kearney@ncifcrf.gov)

24

25 Sarah Palmer, HIV Drug Resistance Program, National Cancer Institute-Frederick, P.O.  
26 Box B, Building 535, Frederick, MD 21702-1201, USA; [sarah.palmer@smi.ki.se](mailto:sarah.palmer@smi.ki.se)

27

28 Wei Shao, HIV Drug Resistance Program, National Cancer Institute-Frederick, P.O. Box  
29 B, Building 535, Frederick, MD 21702-1201, USA; [shaow@ncifcrf.gov](mailto:shaow@ncifcrf.gov)

30

31 Frank Maldarelli, HIV Drug Resistance Program, National Cancer Institute-Frederick, P.O.  
32 Box B, Building 535, Frederick, MD 21702-1201, USA; fmalli@mail.nih.gov

33

34 Eoin P Coakley, Monogram Biosciences, Inc., 345 Oyster Point Blvd., South San  
35 Francisco, CA 94080-1913, USA; ecoakley@monogrambio.com

36

37 Colombe Chappey, Genentech, INC., 1 DNA Way, South San Francisco, CA 94080-4990,  
38 USA; colombe.chappey@gmail.com

39

40 Christine Wanke, Tufts University School of Medicine, 136 Harrison Ave., Boston MA  
41 02111, USA; Christine.wanke@tufts.edu

42

43 John M. Coffin, Tufts University School of Medicine, 136 Harrison Ave., Boston MA  
44 02111, USA; john.coffin@tufts.edu

45

46

47 **Running Title:** HIV-1 viral diversity PCR/cloning vs. single genome sequencing

48

49 **Word Count:** Abstract- 150 body- 2,574

50

51

52

53 **Abstract**

54

55 To compare standard PCR/cloning and single genome sequencing (SGS) in their ability to  
56 reflect actual intra-patient polymorphism of HIV-1 populations, a total of 530 HIV-1 *pro-*  
57 *pol* sequences obtained by both sequencing techniques from a set of 17 ART naïve patient  
58 specimens was analyzed. For each specimen, 12 and 15 sequences, on average, were  
59 characterized by the two techniques. Using phylogenetic analysis, tests for panmixia and  
60 entropy, and Bland-Altman plots, no difference in population structure or genetic diversity  
61 was shown in 14 of the 17 subjects. Evidence of sampling bias by the presence of subsets  
62 of identical sequences was found by either method. Overall, the study shows that neither  
63 method was more biased than the other, and providing that an adequate number of PCR  
64 templates is analyzed, and that the bulk sequencing captures the diversity of the viral  
65 population, either method is likely to provide a similar measure of population diversity.

66

67

68 **Keywords:** HIV, Single genome sequencing (SGS), pro-pol diversity, cloning and  
69 sequencing, treatment naïve

70

## 71 1. Introduction

72

73 Human immunodeficiency virus type 1 (HIV-1) exists as an evolving population in  
74 infected individuals (Coffin 1995). The genetic diversity of HIV-1 results from rapid, high-  
75 level virus turnover (approximately  $10^{11}$  virions per day and  $10^8$  infected cells per day) and  
76 from nucleotide misincorporation during replication of the HIV-1 genome by error prone  
77 reverse transcriptase (RT), (Mansky and Temin 1995; Menendez-Arias 2002; Preston et al.,  
78 1988; Roberts et al., 1988) as well as mutagenic host factors (Smith 2005). Importantly,  
79 many mutations do not have a deleterious impact on viral fitness and thus accumulate  
80 during successive rounds of viral replication. To characterize variants making up a viral  
81 population, it has been a common practice to obtain multiple sequences by performing RT-  
82 PCR on a region of the viral genome, cloning the amplified products, and selecting at  
83 random a number of clones for sequencing. Because primer DNA sequences used in PCR  
84 are pre-defined, PCR imposes a selection which may underestimate actual intra-patient  
85 diversity (Liu et al., 1996). If the number of RT-PCR templates in the original specimen is  
86 low, (or poorly reactive with the primers), it is unlikely that all sequences subsequently  
87 obtained by cloning will be derived from different input templates resulting in the  
88 resampling of individual genomes in the population. PCR-based recombination has also  
89 been observed, generating sequences that are not present in the original virus population  
90 (Liu et al., 1996; Shao et al., 2009). Single genome sequencing (SGS, also called SGA)  
91 permits individual cDNA molecules derived from defined portions of the genome to be  
92 PCR amplified and sequenced in bulk thus eliminating the effects of PCR-based  
93 recombination and the re-sampling of multiple clones from the same initial template  
94 molecule; and greatly reducing the error rate due to PCR (Palmer et al., 2005). The SGS  
95 assay error rate has been estimated to be 0.003% and the assay recombination rate was  
96 estimated to be less than one crossover between two closely related templates in 66,000 bp  
97 analyzed (Palmer et al., 2005). Previously a comparison of genetic diversity obtained from  
98 sequences derived by SGS and PCR was published (Salazar-Gonzalez et al., 2008);  
99 however, no comparative analysis of the PCR/cloning and SGS in their ability to reflect  
100 intra-patient HIV-1 diversity has been published. The present study compares the genetic

101 diversity among HIV-1 *pro-pol* sequences derived from a set of patient specimens using  
102 these two methods.

103

## 104 **2. Materials and Methods**

105

### 106 2.1 Patients and virological endpoints

107 Single plasma specimens from seventeen ART naïve individuals over the age of 18  
108 were obtained from patients attending the Tufts Medical Center infectious disease clinic or  
109 from an established cohort of ART naïve HIV-1 infected prisoners in the Commonwealth  
110 of Massachusetts (Table 1) (Stone et al., 2002). The study was approved by the Institutional  
111 Review Board at Tufts Medical Center, the Human Research Review Committee for the  
112 Massachusetts Department of Public Health, Lemuel Shattuck Hospital and the  
113 Massachusetts Department of Corrections Health Service Unit, and the Office of Human  
114 Subjects Protection at the National Institutes of Health. All subjects provided written  
115 informed consent for participation and testing of specimens. All patients were  
116 antiretroviral naïve by self-report, chart review, and/or primary physician report. The  
117 median HIV-1 RNA level was 34,000 copies/ml (490- 300,000 copies/ml); and the median  
118 CD4 count cells was 393 cells/ $\mu$ l. Subjects' estimated year of HIV infection, by self-report,  
119 ranged from 1988-2003. All plasma specimens were obtained from July 2000 to July 2001  
120 except for the specimens from patient 15 and patient 16 which were obtained in 2004.  
121 Estimated times from seroconversion to specimen collection ranged from 6 months to 12  
122 years.

123

### 124 2.2 PCR/Cloning and sequencing

125 HIV RNA was harvested using a standard guanidinium isothiocyanate extraction  
126 method (Zhang et al., 1991). Population based sequencing was performed using a  
127 previously described protocol using MULV reverse transcriptase and platinum Taq  
128 (Invitrogen, Carlsbad, CA, USA). A 1.4 kb fragment of *gag-pol* was amplified by a 35-  
129 cycle RT-PCR and subsequent 25-cycle nested PCR (NPCR) using a previously described  
130 protocol and primer sets initially designed to amplify HIV-1 subtype B at low levels of  
131 viraemia (Coakley et al., 2002).

132 PRL- f (nt. 1800 HXB2; 5'GGGACCAGCGGCTACACTAGAAGAAATGATGACAG  
133 CATGTCAGG3'),  
134 pRev (nt. 2514 HXB2; 5' AATCTGAGTCAACAGATTTCTTCC3) and  
135 Pro1.8-f (nt. 1897 HXB2; 5'GAAGCAATGAGCCAAGTAACAAAT3'),  
136 pRev (nt. 2514 HXB2; 5' AATCTGAGTCAACAGATTTCTTCC3) (Coakley et al., 2002).

137

138 NPCR products generated as described above were cloned using a TOPO TA cloning  
139 vector (Invitrogen, Carlsbad, CA, USA) following manufacturer's instructions.

140 Sequencing of plasmid DNA isolated from randomly chosen individual bacterial colonies  
141 (7-20 per specimen) was performed by standard dideoxy methods using conserved primers  
142 (Macrogen, Rockville, MD, USA).

143

### 144 2.3 Single Genome Sequencing

145 HIV RNA was extracted using standard guanidinium extraction methods [7]; cDNA  
146 was synthesized using random hexamers and diluted to an average of one amplifiable  
147 molecule per 3 wells of a microtiter plate and PCR amplified using a previously described  
148 methodology and primer sets (Palmer et al., 2005). A 1.4 kb fragment of *gag-pol* was (p6-  
149 RT region; HXB2 bases 2253-3257) was amplified and analyzed. Sequencing of DNA  
150 produced by SGS was performed by standard dideoxy methods using conserved primers  
151 (Macrogen, Rockville, MD, USA).

152

### 153 2.4 Sequence alignment and distance measurements

154 A total of 530 sequences, 1.4 kb in length, was analyzed from the seventeen  
155 patients. For each specimen, a mean of 12 and 15 sequences was characterized by  
156 PCR/cloning and by SGS respectively. Nucleotide sequences were aligned using Clustal X  
157 (Chenna et al., 2003). All alignments were visually inspected and frameshifts were  
158 removed using BioEdit sequence editor (<http://www.mbio.ncsu.edu/BioEdit/BioEdit.html>).  
159 A consensus sequence for each patient sequence set was generated by the BioEdit sequence  
160 editor (<http://www.mbio.ncsu.edu/BioEdit/BioEdit.html>). Genetic diversity was measured  
161 by average pairwise differences (APD) within and between sequence sets derived from  
162 each specimen using MEGA 4.0 (<http://www.megasoftware.net>). Neighbor-joining (NJ)

163 tree construction with 1,000 bootstrap replicates was performed using MEGA 4.0  
164 (<http://www.megasoftware.net>).

165

## 166 2.5 Testing for Divergence

167 A series of tests for population subdivisions described by Hudson et al. (Hudson et  
168 al., 1992) and adapted to biological sequences by Achaz et al. (Achaz et al., 2004) was  
169 performed. This test determines the probability that HIV-1 sequences derived by SGS and  
170 by cloning are derived from the same or different populations within the viral quasispecies.  
171 The method is a non-parametric test that computes pairwise differences between all  
172 sequences from both samples and calculates the probability of panmixia ( $p(K^*_s)$ ) (i.e. the  
173 probability that two different populations are not statistically different from each other). A  
174 p-value greater than the nominal level of significance ( $p(K^*_s) > 0.003$ ) suggests that each  
175 set of sequences is unlikely to have been derived from different populations. In this  
176 analysis, 10,000 permutations were used to obtain p-values. Testing was performed using a  
177 web-based program (<http://www.wabi.snv.jussieu.fr/achaz/hudsonstest.html>). Bland-Altman  
178 analysis was used to determine the limits of agreement between SGS and cloning  
179 measurements. Limits of agreement between methods were defined as the mean difference  
180  $\pm 2$  SD (Bland and Altman 1999; Dewitte et al., 2002).

181

## 182 2.6 Testing for Entropy

183 To further characterize and understand the differences observed between sequence  
184 diversity obtained by both methods a test of Shannon entropy, which applies a measure of  
185 variation in sequence alignments and compares two sets of aligned sequences to determine  
186 if there is variability in one set relative to the other, was performed; statistical confidence is  
187 achieved using a Monte Carlo randomization strategy (Efron and Tibshirani 1991; Leitner  
188 et al., 1993). One thousand randomizations were performed, comparing each set of  
189 sequences with statistical significance defined as  $p < 0.005$ . Analyses were performed on  
190 both nucleic and amino acid sequences.

191

## 192 3. Results

### 193 3.1 Overall sequence relationships and drug resistance

194 The NJ tree showed no evidence of relatedness among the virus consensus  
195 sequences from different patients with the exception of patients 15 and 16, a known  
196 transmission pair (Fig. 1). Sequences had no evidence of major HIV drug resistance  
197 mutations based on the Stanford HIV drug resistance mutation algorithm  
198 (<http://hivdb.stanford.edu>) except for patient 15 and 16, both of whom had K103N  
199 mutations, encoding resistance to non-nucleoside RT inhibitors. Sequences with K103N  
200 mutations were reverted to wild type when analyses were performed.

201

### 202 3.2 Average pairwise distance observed within and between assays

203 Intrapopulation APD observed by SGS ranged from 0.20% to 2.04% with a median  
204 of 0.81%. APD values obtained by PCR/cloning had a similar range, 0.23% to 2.08% with  
205 a median of 0.87% (Table 2). The mean pairwise difference between the two assays ranged  
206 from 0.03% to 1.27% with a median difference of 0.15%. The diversity values obtained by  
207 SGS and by PCR/cloning were highly correlated,  $r^2=0.82$ ;  $p<0.000001$  (Student t-test)  
208 (Fig 2a); the correlation was robust irrespective of plasma RNA level, and remained  
209 statistically significant with removal of the values with highest diversity ( $p<0.0003$ )  
210 suggesting that outliers were not driving the correlation.

211

### 212 3.3 Assessment of sampling bias using an automated test of panmixia

213 Sampling bias between SGS and PCR/cloning was further assessed using a web-  
214 based test of panmixia (<http://www.abi.snv.jussieu.fr/achaz/hudsonstest.html>). The  
215 algorithm assumes that if a bias had been introduced by the method of analysis, the groups  
216 of sequences obtained by SGS and cloning would be significantly different with  
217 probabilities of panmixia ( $K^*s$ ) less than 0.003. Fourteen of the seventeen sets of sequences  
218 demonstrated no such sampling bias. For three patients, namely 10, 11, and 12, sequences  
219 had probabilities of panmixia less than this value suggesting the possibility of bias in one  
220 method relative to the other (Table 2). Bland-Altman analysis showed no evidence of bias  
221 (i.e. the difference in APD between the two assays) as a function of degree of diversity  
222 (Fig. 2b). The mean bias was 0.0271 and 95% limits of agreement ranging from -0.45 to  
223 0.51. All values were within the 95% limits of agreement.

224

### 225 3.4 Assessment of genetic distances using Neighbor-joining trees

226 Neighbor-joining trees with all the sequences from each patient were generated to  
227 describe the genetic distance among the sequences obtained by the two techniques.  
228 Sequences from four patients, 2, 10, 11, and 12, are used to illustrate the different tree  
229 configurations observed. The NJ tree derived from the virus in patient 2 was typical of the  
230 NJ trees observed for 14 of 17 patient specimens and showed intermingling sequences with  
231 no overall difference in diversity (Fig. 3a). By contrast, for patient 10, PCR/cloning  
232 identified a relatively distinct clade of six genetically closely related sequences, none of  
233 which was similar to the genomes derived by SGS (Fig. 3b). The tree structure shows that a  
234 subpopulation of the sequences in this patient was preferentially amplified using cloning,  
235 although the bootstrap support was low and no difference in APD was observed. The APDs  
236 were similar at 1.33% and 1.20% for SGS and by PCR/cloning respectively (Student t-test,  
237  $p=0.07$ , Table 2) with a  $p(K*S)$  of 0.001. Several assay artifacts could explain the  
238 preferential amplification including PCR amplification error, primer selectivity of both  
239 assays and/or PCR-based recombination.

240 In patient 11 (Fig 3c), the APD for SGS and PCR/cloning were 0.71 % and 1.20%  
241 respectively, with  $p(K*S)$  of 0.008. With a similar tree configuration, the higher APD may  
242 be a result of the small number of sequences (11) derived by PCR/cloning. This finding  
243 highlights the importance of characterizing a large number of genomes when analyzing  
244 differences in population diversity.

245 In patient 12, the APD was 0.66% for sequences obtained by SGS and 0.63% for  
246 PCR/ cloning, with  $p(K*S)$  of 0.001. As in patient 10, the overall diversity measured by  
247 SGS and cloning was comparable; however, the NJ tree (Fig. 3d) demonstrated two sub-  
248 populations of virus present in cloned sequences. The first sub-population detected by  
249 PCR/cloning is a cluster of identical or highly similar sequences which likely reflects re-  
250 sampling of a single template during PCR. A second cluster was detected by PCR/cloning  
251 and not by SGS. The two PCR/cloning sequence clusters and the SGS sequence cluster  
252 may reflect preferential amplification by either method.

253

### 254 3.5 Assessment of position-specific differences between methods using Shannon Entropy

255 To investigate whether there were any position-specific differences in SGS or  
256 cloning-derived measures of diversity, each nucleotide position was analyzed using an  
257 automated test for Shannon entropy  
258 (<http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html>) (Table 3). At the  
259 nucleic acid level, differences in entropy using a Monte Carlo randomization strategy were  
260 found to be statistically significant ( $p < 0.005$ ) in patients 9, 11, 12, 13, indicating that SGS  
261 and cloning differed in detecting genetic diversity at individual positions in *pro-pol*.  
262 Importantly, there was no evidence of systematic position specific bias by either method in  
263 determining genetic diversity. At the amino acid level, no differences in entropy were  
264 detected, with the exception of patient 12 at amino acid position 64 of reverse transcriptase,  
265 where all 17 PCR/clonal sequences were lysine, while arginine was present in three of the  
266 seven SGS derived sequences and lysine in the remaining four.

267 Of interest, no evidence of correlation was observed between the APD and the  
268 plasma RNA level over the range of viral loads studied 490-300,000 copies/ml (Fig. 4).

269

#### 270 **4. Discussion**

271 Describing HIV population diversity is increasingly important for the assessment of  
272 intrahost virus evolution and its relationship to disease progression, including the existence  
273 and development of low frequency drug resistance mutations and their impact on treatment  
274 outcomes. Additionally, the accurate assessment of population diversity is essential in  
275 understanding the effects of micro-environmental pressure on population genetic variation  
276 over time and in estimating dates of HIV seroconversion based on estimates of viral  
277 diversity. The PCR/cloning technique is widely used to describe HIV-1 population  
278 diversity and detect low frequency mutations. SGS is a newer technique which is gaining  
279 popularity and this study assesses the genetic diversity obtained from techniques on plasma  
280 specimens from 17 patients.

281 An adequate sample size is required to estimate genetic diversity. Adequate  
282 sampling to detect minor species has been estimated using probability considerations  
283 (Salazar-Gonzalez et al., 2008); with 14 sequences there is a 10 % probability of not  
284 sampling sequences present at a frequency of  $< 15$  % (Salazar-Gonzalez et al., 2008).  
285 Carrying this model forward, binomial probability suggests that when a population is

286 composed of two variants A and B present in equal amounts, the probability of detecting  
287 each in exactly a 50:50 mixture is 0.2. Likewise, the probability of detecting 3 of variant A  
288 and 11 of B is 0.001. Overall, both PCR/cloning and SGS detected similar levels of genetic  
289 diversity in the patients sampled, even in circumstances where sensitive statistical analysis  
290 revealed sampling differences.

291 Of importance is the occurrence of viral subpopulations detected by one technique  
292 and not the other. A subpopulation of homogeneous viruses may be the actual result of the  
293 selection of a fit virus variant, or on contrary it may be an artifact of the amplification step  
294 of either technique. Each technique was found to miss a viral sub-population reported by  
295 the other. Each assay employed different sets of HIV-1 subtype B specific primers and the  
296 number of subjects in the study was too small to determine if either sequencing technique  
297 preferentially selected specific subpopulations because of the primer sequences. Due to low  
298 genetic diversity among viral populations within any one individual patient and the  
299 likelihood of recombination during virus replication in vivo, it was not possible to assess  
300 PCR-based recombination. If significant recombination events had occurred during PCR in  
301 PCR/cloning derived sequences, the overall measure of diversity would not be affected, but  
302 the tree topology would be severely compromised. Similarly, PCR/cloning does not permit  
303 the assessment of mutation linkage, which is feasible with SGS. Finally, while only clearly  
304 observed in one patient 10 (Fig. 3d), PCR resampling could lead to an underestimation of  
305 genetic diversity. Overall, the study demonstrates that neither method was more biased than  
306 the other, and that providing an adequate number of genomes are analyzed, either method  
307 is likely to provide similar measures of population diversity.

308

### 309 **Acknowledgements**

310 MRJ was supported by the National Institute for Allergy and Infectious Disease:  
311 T32 AI07389; CFAAR 1P30A142853-10; K24 A1055293-06A1; and K23 AI074423-03;  
312 and the Center for Drug Abuse and AIDS Research: P30 DA013868. JMC was a Research  
313 Professor of the American Cancer Society, with support from the George Kirby  
314 Foundation.

315

### 316 **References**

- 317 Achaz, G., Palmer, S., Kearney, M., Maldarelli, F., Mellors, J.W., Coffin, J.M., Wakeley,  
318 J., 2004. A robust measure of HIV-1 population turnover within chronically  
319 infected individuals. *Mol. Biol. Evol.* 21(10): p. 1902-12.
- 320 Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. *Stat*  
321 *Methods Med Res.* 8(2): p. 135-60.
- 322 Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson,  
323 J.D., 2003. Multiple sequence alignment with the Clustal series of programs  
324 *Nucleic. Acids. Res* 2003. 31 (13): p. 3497-5000.
- 325 Coakley, E.P., Doweiko, J.P., Bellossilo, N.A., D'Agata, E.M., Albrecht, M.A., 2002. HIV  
326 Drug Resistance Profiles and Clinical and Virologic Outcomes among HIV-Infected  
327 Subjects with Stable Detectable Plasma Viral Loads < 1000 Copies/mL for at least  
328 12 Months 9th Conference on Retroviruses and Opportunistic Infections. 556-T.
- 329 Coffin, J.M., 1995. HIV population dynamics in vivo: implications for genetic variation,  
330 pathogenesis, and therapy. *Science.* 267(5197): p. 483-9.
- 331 Dewitte, K., Fierens, C., Stöckl, D., Thienpont, L.M., 2002. Application of the Bland-  
332 Altman plot for interpretation of method-comparison studies: a critical investigation  
333 of its practice. *Clin. Chem.* 48(5): p. 799-801; author reply 801-2.
- 334 Efron, B., Tibshirani, R., 1991. *Statistical Data Analysis in the Computer Age.* *Science.*  
335 253(5018): p. 390-395.
- 336 Hudson, R.R., Boos, D.D., Kaplan, N.L., 1992. A statistical test for detecting geographic  
337 subdivision. *Mol. Biol. Evol.* 9(1): p. 138-51.
- 338 <http://www.abi.snv.jussieu.fr/achaz/hudsonstest.html>; [Last Accessed November 3, 2009].  
339 <http://hivdb.stanford.edu>. [Last Accessed November 3, 2009].  
340 <http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html> [Last Accessed  
341 November 3, 2009].  
342 <http://www.megasoftware.net>; [Last Accessed April 22, 2010].  
343 <http://www.mbio.ncsu.edu/BioEdit/BioEdit.html>; [Last Accessed April 22, 2010].
- 344 Leitner, T., Halapi, E., Scarlatti, G., Rossi, P., Albert, J., Fenyö, E.M., Uhlén, M., 1993.  
345 Analysis of heterogeneous viral populations by direct DNA sequencing.  
346 *Biotechniques.* 15(1): p. 120-7.
- 347 Liu, S.L., Rodrigo, A.G., Shankarappa, R., Learn, G.H., Hsu, L., Davidov, O., Zhao, L.P.,  
348 Mullins, J.I., 1996. HIV quasispecies and resampling. *Science.* 273(5274): p. 415-6.
- 349 Mansky, L.M., Temin, H.M., 1995. Lower in vivo mutation rate of human  
350 immunodeficiency virus type 1 than that predicted from the fidelity of purified  
351 reverse transcriptase. *J. Virol.* 69(8): p. 5087-94.
- 352 Menendez-Arias, L., 2002. Molecular basis of fidelity of DNA synthesis and nucleotide  
353 specificity of retroviral reverse transcriptases. *Prog Nucleic. Acid. Res. Mol. Biol.*  
354 71: p. 91-147.
- 355 Palmer, S., Kearney, M., Maldarelli, F., Halvas, E.K., Bixby, C.J., Bazmi, H., Rock, D.,  
356 Falloon, J., Davey, R.T., Jr, Dewar, R.L., Metcalf, J.A., Hammer, S., Mellors, J.W.,  
357 Coffin J.M., 2005. Multiple, linked human immunodeficiency virus type 1 drug  
358 Resistance mutations in treatment-experienced patients are missed by standard  
359 genotype analysis. *J. Clin. Microbiol.* 43(1): p. 406-13.
- 360 Preston, B.D., Poiesz, B.D., Loeb, L.A. 1988. Fidelity of HIV-1 reverse transcriptase.  
361 *Science.* 242(4882): p. 1168-71.
- 362 Roberts, J.D., Bebenek, K., Kunkel, T.A., 1988. The accuracy of reverse transcriptase from

- 363 HIV-1. *Science*. 242(4882): p. 1171-3.
- 364 Salazar-Gonzalez, J.F., Bailes, E., Pham, K.T., Salazar, M.G., Guffey, M.B., Keele, B.F.,  
365 Derdeyn, C.A., Farmer, P., Hunter, E., Allen, S., Manigart, O., Mulenga, J.,  
366 Anderson, J.A., Swanstrom, R., Haynes, B.F., Athreya, G.S., Korber, B.T., 2008.  
367 Sharp, P.M., Shaw, G.M., Hahn, B.H. Deciphering human immunodeficiency virus  
368 type 1 transmission and early envelope diversification by single-genome  
369 amplification and sequencing. *J. Virol.* 82(8): p. 3952-70.
- 370 Shao, W., Boltz, V.F., Kearney, M., Maldarelli, F., Mellors, J.W., Stewart, C., Levitsky,  
371 A., Volfovsky, N., Stephens, R.M., Coffin, J.M., 2009. Characterization of HIV-1  
372 sequence artifacts introduced by bulk PCR and detected by 454 sequencing. XVIII  
373 international HIV drug resistance workshop, Fort Myers, FL, USA. Abstract 104
- 374 Smith, R.A., Loeb, L.A., Preston, B.D., 2005. Lethal HIV mutagenesis. *Virus. Res.*  
375 107(2): p. 215-228.
- 376 Stone, D.R., Corcoran, C., Wurcel, A., McGovern, B., Quirk, J., Brewer, A., Sutton, L.,  
377 D'Aquila, R.T., 2002. Antiretroviral drug resistance mutations in antiretroviral-  
378 naive prisoners. *Clin. Infect. Dis.* 35(7): p. 883-6.
- 379 Zhang, L.Q., Simmonds, P., Ludlham, C.A., Brown, A.J., 1991. Detection, quantification  
380 and sequencing of HIV-1 from the plasma of seropositive individuals and from  
381 factor VIII concentrates. *AIDS*. 5: p. 675-681.
- 382

383

384 **Figure Legends**

385

386 **Fig. 1 Relationship among virus populations in the studied patients.** All sequences  
387 obtained in this study were compiled and a single NJ tree was constructed to check for  
388 sequence overlap. The NJ tree and bootstrap resampling of 1,000 trees demonstrated  
389 separate clustering of sequences from each patient; thus excluding contamination (data not  
390 shown). An NJ tree was prepared for the consensus sequences of the virus in all patients.  
391 Patients 15 and 16 are a known transmission pair. All sequences are HIV-1 subtype B.  
392 Clustering of SGS derived sequences and sequences derived by PCR/cloning within each  
393 patient cluster was evident with no evidence of contamination (data not shown). Additional  
394 subtype B and C reference sequences obtained from the Los Alamos National HIV  
395 Database were included in the analysis for comparison.

396

397

398 **Fig. 2 Relationship of sequence diversity to virus load for the two methods a)**

399 Correlation of APD values obtained by SGS and PCR/cloning. b) Bland-Altman plot of the  
400 difference in average pairwise difference between the two assays as a function of diversity.

401

402 **Fig. 3. Neighbor-joining trees of HIV-1 populations from selected patients.** Solid  
403 squares represent sequences derived by PCR/cloning and open squares represent sequences  
404 derived by single genome sequencing. a) NJ tree of patient 2, representative of 14/17 trees  
405 obtained in the analyses showing intermingling of sequences obtained by PCR/cloning and  
406 sequences obtained by single genome sequencing with no overall difference in diversity by  
407 topology. Trees of sequences from the patients showing distinct populations by the two  
408 methods are shown in b-d. b) Patient 10; A cluster of 6 sequences amplified selectively by  
409 cloning and sequencing is denoted on the tree by a bracket. c) Patient 11. d) Patient 12; two  
410 distinct sub-populations of virus found by in PCR/cloning and not observed in SGS derived  
411 sequences are denoted by brackets. Additional reference sequences obtained from the Los  
412 Alamos National HIV database were included in the analysis..

413

414 **Fig. 4: Probability of  $p(K^*S)$** , between SGS and clonal sequences are plotted as a function  
415 of viral load. No correlation between viremia and  $pK$  is observed.

Accepted Manuscript