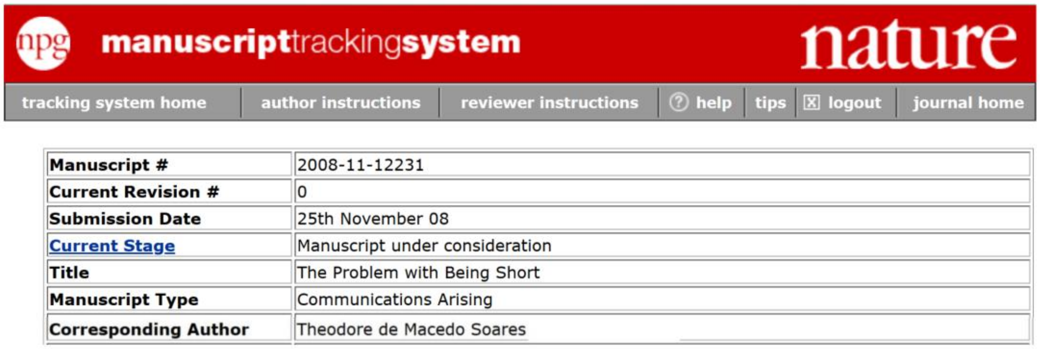

Manuscript #	2008-11-12231
Current Revision #	0
Submission Date	25th November 08
Current Stage	Manuscript under consideration
Title	The Problem with Being Short
Manuscript Type	Communications Arising
Corresponding Author	Theodore de Macedo Soares

The Problem with Being Short

Theodore de Macedo Soares

After phylogenetic analyses of the two earliest HIV-1 sequences from 1959 (ZR59) and 1960 (DRC60) and other HIV, Worobey et al.[1] estimated that the HIV-1 M-group originated from a common ancestor "near the beginning of the twentieth century." Their key analyses, however, are based on just 163 matching envelope gene sites, less than 2% of HIV's genome, which raises questions about the validity of their results.[2, 3, 4] Analyses of all DRC60 segments do not support their conclusions.

Worobey et al. states that the 11.7% nucleotide difference between DRC60 and ZR59 "indicates that the HIV-M group founder virus began to diversify [in humans]…decades before 1960." Applying a 95% confidence interval (CI) of the population proportion (applied here, to my knowledge for the first time, to gauge the reliability of sequence sample prior to phylogenetic analysis), results in a difference of 11.7% with wide CI ranging 7.6% - 17.5%.

$$\tilde{p}_s \pm z_{a/2}\sqrt{\frac{\tilde{p}_s(1-\tilde{p}_s)}{n+\left(z_{a/2}\right)^2}}\sqrt{\frac{N-n}{N-1}} \qquad \tilde{p}_s = \frac{X+\frac{\left(z_{a/2}\right)^2}{2}}{n+\left(z_{a/2}\right)^2} \cong \frac{X+2}{n+4}$$

Figure1| Adjusted Wald Interval[5, 6] with Finite Population Correction[7]
DRC60/ZR59 Confidence Interval: $z_{a/2}$ =1.96 (for 95% CI), X=19 (# differences), n=163, N=9719 (Genome size of the reference HIV1-HXB2 sequence). Finite Population Correction [(N-n)/(N-1)]$^{1/2}$ negligible when sampled population is less than 5% of population.[7]

A 7.6% difference between the two samples is much less the average HIV-1 M-group between-subtypes p-distance of 12.9% and would be consistent with a recent common ancestor. A 17.5% difference would either place the common ancestor much further back in time (inconsistent with an epidemic beginning in the late 1970's) or would be consistent with recent separate transmissions of SIVcpz from different chimpanzees as the three most closely related SIVcpz (LB7, MB66, MB897), differ by 19.8%.

The 95%CI applied to differences ranging 0.6%-19% between each of all the sequences analyzed by Worobey et al., result in the true difference varying 517%-28%, 205%-17%, 119%-12%, for their alignments consisting of 163, 492, 994 nucleotide sites respectively, than measured (the smaller measured difference resulting in the larger variance). A measured difference of 8% would necessitate a sample size of 3,000 nucleotides to result in the true difference varying by about ±10%. If care is taken to assure that the sampling methodology results in a phylogenetically representative sample of HIV's genome then the Adjusted Wald Interval proposed here serves to estimate the margin of error and confidence interval associated with sample size. If as here, there is no such assurance (the author's alignment determining DRC60 and ZR59 chance convenience samples were recovered from randomly degraded DNA); the width of this confidence interval functions as a minimum estimate.

As the true differences in the author's alignments range widely and as these CI nucleotides are unknown, a large variation in possible branch lengths and time to most recent common ancestor result from a character-based phylogenetic method such as used by them. As this confidence interval affects variability and not length, phylogenetic method sensitivity to short sequence lengths[2, 3, 4] would additionally apply.

Neither do analyses of all DRC60 (507bp) segments added to 235 full-length HIV-1 M sequences including the theoretical M group ancestor[8] at the approximate center of HIV's radial phylogeny, supports their conclusions. P-distance between DRC60 (501 sites after gap-stripping) and the ancestor of 7.2% (CI 5.3-9.8) was found longer than 53 (8 subtype A [as is DRC60]) other HIV sequences by an average of 8.9% (2.9-28.6). Corrected distances, using NJ Maximum Composite Likelihood methods[9, 10] determined 58 other sequences closer to the ancestor.

Illustrating the problem with being short: p-distance calculations for the 163bp sites used by the authors determined 206 of 234 sequences closer to M-ancestor than DRC60 by a median of 74.9%. Just as 163 sites are found not to be representative of 501 sites, 501 or 994 sites are not necessarily representative of ~10,000 HIV sites. The wide 95% confidence interval associated with the short 163bp alignment supports conclusions ranging from a recent to ancient common ancestor. DRC60 is found amidst modern sequences. DRC60's upper 95%CI—exacerbating this finding—further undermines its probative value and indicates that even all 507 sites of DRC60 is too short for dependable phylogenetic analyses. These findings caution against the use of such short sequences in phylogenetic analysis as they may, as here, lead to overconfident results.

Methods

The 2004 HIV group M ancestor was downloaded from Los Alamos National Laboratory (LANL), aligned using ClustalX(1.83), manually adjusted, and fitted to 234 HIV-1 M group sequences (named recombinants excluded) downloaded pre-aligned from LANL. The average between-subtype distances were determined from the 234 HIV sequences containing 7668 sites after gap-stripping. SIVcpz distances were determined from an alignment containing 8816 sites after gap-stripping. All analyses were conducted using Mega4.[10] The NJ distances used Maximum Composite Likelihood with heterogeneous setting between lineages and gamma (0.85)[11] distributed along sites.

[1] Worobey M et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature.* 2008 Oct 2;**455**(7213):661-4.

[2] Cummings MP, Otto SP, Wakeley J. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol Biol Evol*. 1995 Sep;**12**(5):814-22.

[3] Rosenberg MS, Kumar S. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc Natl Acad Sci U S A.* 2001 Sep 11;**98**(19):10751-6. Epub 2001 Aug 28.

[4] Kolaczkowski B, Thornton JW. Effects of branch length uncertainty on Bayesian posterior probabilities for phylogenetic hypotheses. *Mol Biol Evol*. 2007 Sep;24(9):2108-18. Epub 2007 Jul 17.

[5] Agresti, A., & Coull, B. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician,* **52**, 119-126.

6 Brown, L. D., Cai, T. and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statist. Sci.* **16**: 101–133.

7 Cochran, William G. Sampling techniques. Third edition. Wiley Series in Probability and Mathematical Statistics. *John Wiley & Sons, New York-London-Sydney,* 1977.

[8] 2004 M Group ancestor as constructed by the Los Alamos National Laboratory: http://www.hiv.lanl.gov/ .

[9] Tamura K, Nei M & Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences (USA)* **101**:11030-11035.

[10] Tamura K, Dudley J, Nei M & Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24:**1596-1599.

[11] Posada D, Crandall KA.Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1).*Mol Biol Evol*. 2001 Jun;**18**(6):897-906.

Author:
Theodore de Macedo Soares
[redacted]

NOTES:

1. For the reviewers/referees: All analyses and sequence alignments used will be made immediately available upon request.
2. Although the modified version submitted to *Nature* on December 14, 2008 is accurate and, if necessary, suffices, this version reflects the following improvements:
    a. Directly references all of the authors' alignments.
    b. Better explains the use of the Modified Wald Interval.
    c. Uses the 2004 HIV-1 M group ancestor which is better annotated by LANL in their website.
    d. Deletes reference to ZR59 alignment distances to the ancestor, as additional research indicates this alignment biases ancestor distances towards the B/D node.
    e. Incorporates analyses using 234 full-length HIV sequences that provide robust supporting data. (The authors had supplied 153 concatenated sequences 994 nucleotides in length).

Note: This submission has been reviewed by two university professors who are teaching or have taught statistics—one with a PhD in statistics the other with a PhD in physics—and two professors teaching and working in the biosystematics and bioinformatics fields. Their helpful comments assisted the fine-tuning of this submittal.