HIV Type 1 Mutational Patterns in HIV Type 1 Subtype C-Infected Long-Term Survivors in Karonga District Malawi: Further Analysis and Correction

Ishla Seager,¹ Michael D. Leeson,¹ Amelia C. Crampin,^{2,3} Dominic Mulawa,² Neil French,^{2,3} Judith R. Glynn,³ Simon A.A. Travers,^{1,4} and Grace P. McCormack¹

Abstract

Here we present new sequence data from HIV-1 subtype C-infected long-term survivors (LTS) from Karonga District, Malawi. *Gag* and *env* sequence data were produced from nine individuals each of whom has been HIV-1 positive for more than 20 years. We show that the three amino acid deletion in *gag* p17 previously described from these LTS is not real and was a result of an alignment error. We find that the use of dried blood spots for DNA-based studies is limited after storage for 20 years. We also show some unlikely amino acid changes in *env* C2-V3 in LTS over time and different patterns of genetic divergence among LTS. Although no clear association between mutations and survival could be shown, amino acid changes that are present in more than one LTS may, in the future, be shown to be important.

THE NATURAL HISTORY OF HIV-1 infection and disease progression has been well established in adults, with 10 years being the median time from initial infection to the development of AIDS in the absence of therapeutic intervention.^{1,2} However, this can vary widely within individuals, with rapid progressors developing AIDS symptoms in as little as 6 months, and other individuals, such as long-term survivors (LTS), who progress to AIDS over a much longer period of time.³ LTS are those individuals characterized as having survived for >10 years, without antiretroviral therapy (ART), but who also show a steady decline in the number of circulating CD4 cells.⁴

Some studies have shown viral factors play an important role in the survival of LTS, for example, viruses that contain defects in particular HIV-1 genes, such as *env*, *gag*, *nef*, *vpu*, *vif*, *rev*, and *tat*.⁵ Currently there is very little information for LTS found in sub-Saharan Africa.⁶ McCormack *et al.*⁷ described a three amino acid deletion at the end of *gag* p17 found in 15 LTS from Karonga District, Malawi. In each LTS the deletion was observed in sequences dating from the late 1990s but was not present in any sequences dating from the 1980s. It was also described in two-thirds of the other HIV-1-positive individuals from Karonga District included in that study from the late 1990s and it was suggested that the deletion could be

associated with longer survival and onward transmission.⁷ In this work we sought to further characterize the viral factors involved in long-term survival in Karonga District, Malawi. We find that the previous observation of a three amino acid deletion⁷ was, in fact, erroneous, and we describe our follow-up study of the 38 long-term survivors.

Thirty-eight HIV-positive individuals were seen in Karonga District Malawi in both the late 1980s and late 1990s.² Seventeen of them were still alive in 2004.⁷ Fourteen of these were sought again in 2010 (three were not sought as they had refused to participate the previous two times) when it was found that three had died, one had left the region, and one refused to provide a sample. Of the nine individuals seen, eight were infected with subtype C and one (LTS30) was infected with the unclassifiable strain that was described for this region.⁸ No amplification was possible from dried blood spots (DBS) from this latter individual prior to 2004. Five individuals had begun ART (one in 2005, two in 2006, one in 2008, and one in 2009). Four individuals had not begun ART and are thus HIV positive without treatment for a minimum of 21 years (although one of these has also now been referred) (Table 1).

DNA was reextracted from DBS collected between 1986 and 1989 from LTS and the wider population with a view to

¹Molecular Evolution and Systematics Laboratory, Zoology, Ryan Institute, School of Natural Sciences, National University of Ireland, Galway, Ireland.

²Karonga Prevention Study, Chilumba, Malawi.

³Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom.

⁴South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa.

TABLE 1. SUMMARY OF SEQUENCE DATA, CD4 COUNTS, AND ART INFORMATION AVAILABLE FOR THOSE LTS FOUND TO BE STILL PRESENT IN KARONGA DISTRICT IN 2004 AND IN 2010

Survivor	Year	Sequence data available	CD4 count (cells/mm ³)	ART start date
LTS1	1999	<i>gag</i> and <i>env</i>		
	2004		586	
	2010	<i>gag</i> and <i>env</i>	449	
LTS2	1999	<i>gag</i> and <i>env</i>		
	2010	<i>gag</i> and <i>env</i>	244	ART 2009
LTS5	1989	<i>gag</i>		
	1999	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i>		
	2010	<i>gag</i>	789	ART 2008
LTS8	1989	<i>gag</i> and <i>env</i>		
	1999	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i> and <i>env</i>	265	
	2010	<i>gag</i> and <i>env</i>	734	ART 2005
LTS9 ^a	1988	<i>gag</i> and <i>env</i>		
	1999	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i> and <i>env</i>	56	
LTS10	1989	<i>gag</i> and <i>env</i>		
	1999	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i> and <i>env</i>	328	
	2010	<i>gag</i> and <i>env</i>	138	Referred for ART 2010
LTS12	1988	<i>env</i>		
	1999	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i> and <i>env</i>	439	
	2010	<i>gag</i> and <i>env</i>	390	ART 2006
LTS17 ^b	1989	<i>gag</i> and <i>env</i>		
	1998	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i> and <i>env</i>	452	
LTS20	1988	<i>gag</i> and <i>env</i>		
	2000	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i> and <i>env</i>		
	2010	<i>gag</i> and <i>env</i>	139	ART 2006
LTS21	1989	<i>gag</i> and <i>env</i>		
	1999	<i>gag</i> and <i>env</i>		
	2004		47	
	2010	<i>gag</i> and <i>env</i>	32	Refused ART
LTS22 ^a	1989	<i>gag</i> and <i>env</i>		
	1998	<i>gag</i> and <i>env</i>		
	2004	<i>gag</i> and <i>env</i>	475	
LTS30 ^c	2004	<i>gag</i>	656	
	2010	<i>gag</i> and <i>env</i>	362	

^aHad died by 2010.

^bHad left the Karonga District by 2010.

^cUnclassifiable subtype.

ART, antiretroviral therapy; LTS, long-term survivors.

cloning polymerase chain reaction (PCR) products to explore evidence of the three amino acid deletion in that time period. DBS were also available from 10 of the LTS seen in 2004 and cell pellets from 9 in 2010. Plasma or cell pellets were utilized from 100 individuals randomly chosen from samples collected between 2008 and 2010, from existing studies in the District to explore the frequency of the three amino acid deletion in the 2008–2010 time period. Proviral DNA was extracted from 200 μ l of cell pellet or plasma using the QIAamp DNA Blood Mini Kit (Qiagen) blood and from the DBS using the QIAamp DNA Micro Kit (Qiagen). Nested PCR and sequencing of a 750-bp region of *gag* p17p24 and a 500-bp

region of *env* C2V3 were carried out as previously described.⁹ Amplification from the DBS from the 1980s was largely unsuccessful. DBS had been frozen at -20°C for over 20 years and exposed to a number of freeze-thaw events and it is highly probable that this has led to fragmentation of the DNA present. Previous studies found -20°C suitable for long-term storage of DBS (6 years)^{10–13} but our work suggests an upper limit to the length of time such samples can be stored successfully in this way for DNA-based studies. Cloning was carried out using the Topo TA cloning kit (Biosciences). Automatic sequencing in both directions was carried out by Eurofins Genetic Services Ltd or by LGC Genomics. Sequence chromatographs were examined and manually edited in Seqman (DNASTar Inc.).

The three amino acid deletion was not found in any *gag* sequences produced from LTS in 2004 and 2010 or in over 100 sequences produced from blood samples collected in 2008–2010. Furthermore, none of 50 sequences from 50 clones produced from the 1998 sample from LTS2 (which had previously showed the deletion) contained the deletion. We then reexamined all of the raw data from the sequences used in McCormack *et al.*⁷ and the deletion was not found in any of the sequences. We suggest that an alignment error was made at an early stage of the multiple alignment assembly of the relevant sequences in the original study. This serves as a stark reminder of the dangers of such errors when handling large numbers of sequences. All affected sequences from McCormack *et al.*⁷ that were submitted to GenBank have been reexamined and the correct sequences redeposited.

Multiple alignments of all sequential *env* (74) and *gag* (65) sequences from subtype C-infected LTS along with 40 control sequences were assembled and optimized in MacClade 4 (Sinauer Associates). Phylogenetic trees were reconstructed under the GTR+gamma model of DNA substitution implemented by RAXML 7.0.3¹⁴ with all parameters optimized by RAXML. Confidence levels in the groupings in the phylogeny were assessed using 1000 bootstrap replicates as part of the RAXML phylogeny reconstruction. The subtype C ancestral sequence derived in previous work¹⁵ was employed as the out-group for both *gag* and *env* trees. Both gene trees showed that for most individuals the sequences from the different time points grouped together (8/10 for *env* and 10/11 for *gag*) but only half grouped with significant bootstrap support (Fig. 1a and b). Sequences from LTS21 formed multiple clusters on both gene trees, which is consistent with the pattern seen in McCormack *et al.*⁷

To further explore this, additional *gag* and *env* consensus sequences were produced from DNA reextracted from the DBS collected in 1989 and 1999 for this individual (LTS21) with additional *env* sequences produced from the 2010 DNA sample. In *gag* the 1990s sequences were ancestral to the 2010 sequences and the two 1989 sequences grouped distantly from them (Fig. 1b). In *env* the 2010 sequences showed further variation with two sequences grouping away from the 1999 sequences (Fig. 1a). The average genetic distance between the eight *env* sequences (across all time points) was 12%, which was higher than the 8.8% genetic distance between all other LTS sequences from all individuals at all the different time points. The genetic distance between two of the sequences collected from 2010 was even higher (17.5%) than between the sequences collected in the 1980s and 1990s (7.8%) from the same individual.

FIG. 1. Maximum Likelihood trees generated from (a) *env* and (b) *gag* gene sequences from long-term survivors (LTS) dating from 2004 and 2010 along with local control sequences. Only the LTS sequences are labeled and are named by the LTS number (see Table 1) and the year the sample was collected. Sequences labeled a, b, c, and d refer to multiple consensus sequences generated from the same time point. Bootstrap values of over 70 are marked on the relevant branches.

(Figure Continued →)

FIG. 1. (Continued).

Although sample mislabeling is very unlikely, as the individual's name was written on the filter paper as well as a unique identifier, we cannot exclude the possibility. Superinfection is also possible. The sequences came from a female who was 21 when she was first identified as being HIV-1 positive. She has maintained a very low CD4 count for the past 6 years, 47 cells/mm³ in 2004 and 32 cells/mm³ in 2010. At both visits she was described as being healthy and showing no signs of AIDS and she refused ART on both occasions. Divergent *env* sequences were also found among sequences from the 2010 sample of LTS1, a female who was 37 when she

was first identified as being HIV-1 positive. Her CD4 count was 586 cells/mm³ in 2004 and had fallen slightly to 449 cells/mm³ in 2010 and at that time she had not been referred for ART.

The BLOSUM62 matrix was used to assess the likelihood of amino acid substitutions between sequential sequences of *env* for each LTS, with indels and shared deletions relative to the other sequences also noted. A graphic representation of these observed amino acid substitutions within each LTS sample from one time period to another shows the large amount of change that was apparent across most of *env* C2-V3 in all individuals (Fig. 2). There are a number of positions that

FIG. 2. A graphic representation of observed amino acid substitutions that have occurred in *env* within survivors over time. Substitutions are color coded by likelihood according to the BLOSUM62 matrix with green being the most likely and red the least (green > yellow > orange > red); blue indicates an insertion in one of the sequences relative to the other, pale gray indicates a shared deletion, and dark gray indicates an ambiguous site, e.g., a stop codon. The sequence collected at the earliest available time point was compared to all available *env* sequences from subsequent time points. The comparisons are labeled by the LTS number and the years being compared. Those labeled with an a or b refer to multiple sequences generated from the same time point. **Mark positions 268 and 269 in *env* using HXB2 numbering.

changed within nearly all of the LTS, e.g., HXB2 positions 268–269 showed substitutions in eight LTS (asterisk on Fig. 2). Some of these changes are less likely mutations, e.g., in six individuals (LTS8, 9, 10, 12, 20, and 22) there was a change from glycine to glutamic acid or vice versa. In one individual (LTS17) there a deletion at position 269 within the 1980s and one of the 1990s *env* sequences, which became a lysine in a second 1990s and 2004 sequence. Mutations at positions 11, 24, and 25 within the V3 loop have been associated with a change in coreceptor usage with a shift from negatively charged amino acids to positively charged amino acids being suggested to result in a switch from the use of CCR5 to CXCR4 usage. Only LTS17 showed evidence of a change to a positive charge in this region by 2004 but had left the district by

2010 and so we do not have any additional information on this person. A large amount of change was seen just after the V3 loop in the C3 region. Indeed the high degree of genetic divergence seen between multiple *env* sequences from the same time point in some LTS rendered comparisons of pairwise genetic distances between time points meaningless (even when we excluded the individuals with possible superinfection mentioned above). For example, two consensus sequences from LTS17 from 1999 showed a genetic distance of 4%. The population of viruses within individuals is currently being explored further by sequencing multiple clones.

Comparing amino acid substitutions in *gag* sequences showed higher numbers of substitutions within the *gag* p17

FIG. 3. The genetic divergence seen in *gag* over time in LTS5, 8, 9, 10, and 20. All sequences from one individual from different time points were compared to the sequence generated from the earliest time point available. LTS9 and 22 had died before 2010. LTS5 had begun antiretroviral therapy (ART) in 2008, LTS8 in 2005, LTS20 in 2006, and LTS10 had been referred for ART in 2010 as indicated by the vertical lines.

domain when compared to p24 as might be expected and, using the BLOSUM62 matrix as a reference, most of these substitutions were changes that were more likely to occur (data not shown). Different patterns of genetic divergence in *gag*, calculated using pairwise distances, were apparent among the LTS over time. Two LTS (LTS5 and LTS8) showed an overall trend of an increase in genetic divergence in *gag* over time as can be seen in Fig. 3. Both LTS10 and LTS20 showed a general increase in genetic divergence from 1989 to 2004; however, in 2010 the amount of genetic divergence from 1989 decreased for LTS20 and plateaued in LTS10. Three of these individuals had begun ART between 2004 and 2010 and one has since been referred. Two individuals (LTS9 and LTS22) who were alive in 2004 but had died by 2010 showed very different patterns of divergence. LTS9 showed an increase in divergence from 1989 to 1999, which decreased in 2004, while in LTS22 there was a linear increase in divergence from 1989 to 2004 (Fig. 3).

In summary, we present new sequence data from HIV-1 subtype C-infected long-term survivors from Malawi, Africa. This work highlights some of the pitfalls associated with sequence analysis. We show that the three amino acid deletion in *gag* p17 previously described from these LTS⁷ was the result of an alignment error. Extreme caution must be used while making inferences from sequence data as some errors may be very well hidden. However, data on long-term survivors infected with subtype C virus are important to accumulate. We show many amino acid changes in *env* C2-V3 in LTS over time. Although no clear association between mutations and survival could be shown, amino acid changes that are present in a number of LTS may in the future be shown to be important for survival, but future work in this regard will require data from virus and host.

Sequence Data

All sequences have been deposited into GenBank, accession numbers JN393505–393554, and raw data and alignments employed in this work are available from the authors on request. JN393505–393554

Acknowledgments

This material is based upon works supported by Science Foundation Ireland under Grant 07/RFP/EEEEBF424 and by an IRCSET EMBARK Scholarship to Ishla Seager (RS/2007/203). The Karonga Prevention Study is funded primarily by the Wellcome Trust, with contributions from LEPR. Permission for the study was received from the National Health Sciences Research Committee, Malawi, and the Ethics Committee of the London School of Hygiene and Tropical Medicine, United Kingdom.

Author Disclosure Statement

No competing financial interests exist.

References

- Langford SE, Ananworanich J, and Cooper DA: Predictors of disease progression in HIV infection: A review. *AIDS Res Ther* 2007;4:11.
- Crampin AC, Floyd S, Glynn JR, *et al.*: Long-term follow-up of HIV-positive and HIV-negative individuals in rural Malawi. *AIDS* 2002;16(11):1545–1550.
- Pantaleo G and Fauci AS: Immunopathogenesis of HIV infection. *Annu Rev Microbiol* 1996;50:825–854.
- Learmont J, Tindall B, Evans L, *et al.*: Long-term symptomless HIV-1 infection in recipients of blood products from a single donor. *Lancet* 1992;340(8824):863–867.
- Poropatich K and Sullivan DJ Jr: Human immunodeficiency virus type 1 long-term non-progressors: The viral, genetic and immunological basis for disease non-progression. *J Gen Virol* 2010;92(Pt 2):247–268.
- Laeyendecker O, Redd AD, Lutalo T, *et al.*: Frequency of long-term nonprogressors in HIV-1 seroconverters from Rakai Uganda. *J Acquir Immune Defic Syndr* 2009;52(3):316–319.
- McCormack GP, Glynn JR, Clewley JP, *et al.*: Emergence of a three codon deletion in *gag* p17 in HIV type 1 subtype C long-term survivors, and general population spread. *AIDS Res Hum Retroviruses* 2006;22(2):195–201.
- McCormack GP, Glynn JR, Sibande F, *et al.*: Highly divergent HIV-1 group M sequences in Karonga District, Malawi in the early 1980s. *AIDS Res Human Retroviruses* 2003;19(5):441–445.
- McCormack GP, Glynn JR, Crampin AC, *et al.*: Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. *J Virol* 2002;76(24):12890–12899.
- Cassol S, Salas T, Arella M, Neumann P, Schechter MT, and O'Shaughnessy M: Use of dried blood spot specimens in the detection of human immunodeficiency virus type 1 by the polymerase chain reaction. *J Clin Microbiol* 1991;29(4):667–671.
- Cassol S, Salas T, Gill MJ, *et al.*: Stability of dried blood spot specimens for detection of human immunodeficiency virus DNA by polymerase chain reaction. *J Clin Microbiol* 1992;30(12):3039–3042.
- McNulty A, Jennings C, Bennett D, *et al.*: Evaluation of dried blood spots for human immunodeficiency virus type 1 drug resistance testing. *J Clin Microbiol* 2007;45(2):517–521.
- Steege K, Luchters S, Demecheleer E, *et al.*: Feasibility of detecting human immunodeficiency virus type 1 drug resistance in DNA extracted from whole blood or dried blood spots. *J Clin Microbiol* 2007;45(10):3342–3351.
- Stamatakis A: RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006;22(21):2688–2690.
- Travers SA, Clewley JP, Glynn JR, *et al.*: Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. *J Virol* 2004;78(19):10501–10506.

Address correspondence to:

Grace P. McCormack
Molecular Evolution and Systematics Laboratory
Department of Zoology
Ryan Institute, School of Natural Sciences
National University of Ireland
Galway
Ireland

E-mail: grace.mccormack@nuigalway.ie