Phylogenetic Analysis of HIV Type 1 CRF02_AG in Multiple Genes in Italian and African Patients Living in Italy

Stefania Paolucci,^{1,*} Antonio Piralla,^{1,*} Loretta Fiorina,¹ Roberto Gulminetti,² Stefano Novati,²
Alessia Lai,³ and Fausto Baldanti¹

Abstract

Human immunodeficiency virus type 1 (HIV-1) circulating recombinant form (CRF) 02_AG is a major recombinant variant in different geographic areas and is predominant in West and Central Africa. Of particular interest is the increased frequency of CRF02_AG in patients living in Italy. In the present study, phylogenetic analyses were performed on *gag*, *pol* (*integrase*), and *env* (*gp120* and *gp41*) gene sequences from 34 CRF02_AG-infected patients living in Italy. Thirty out of 34 (89.4%) patients were from western Africa, 3/34 (8.8%) were born in Italy, and 1/34 (2.9%) was from Cuba. Phylogenetic analysis revealed the presence of a well-supported clade (aLRT score > 0.75) of sequences only in *gp120* and *gp41* trees. Evolutionary rate estimation showed a faster evolution for the *gp120* gene with respect to the *gag*, *integrase*, and *gp41* genes. This finding was confirmed by the analysis of interpatient variability. Inpatient variability was greater in *gp120* gene sequences; 10/19 (52.6%; $p < 0.001$) patients had a ratio of $dN/dS > 1$ as compared with *gag*, *integrase*, and *gp41* gene sequences with dN/dS ratios < 1. In summary, phylogenetic analyses of CRF02_AG strains offer a perspective on intrapatient and interpatient variability among CRF02_AG-infected patients living in Italy. In addition, divergent phylogenetic relationships were observed among different genomic regions.

HUMAN IMMUNODEFICIENCY VIRUS type 1 (HIV-1) is characterized by wide genetic variability in different geographic regions as a result of different factors: its high replication rate, reverse transcriptase (RT) error rate, recombination events, and multiple selective factors.

HIV-1 strains are differentiated into four major groups (M, N, O, and P). Within the M group, different subtypes (A–D, F–H, J, and K), 55 circulating recombinant forms (CRFs), and a growing number of unique recombinant forms (URFs) have been reported.^{1,2} The prevalence and geographic distribution of groups, subtypes, and CRFs are rapidly evolving due to increased migration flow, international travel, and global trade.^{3,4} Subtype C is responsible for the majority of HIV-1 infections worldwide (48%), followed by subtype A (12%), subtype B (11%), CRF02_AG (8%) and CRF01_AE (5%), subtype G (5%), and subtype D (2%), while subtypes F, H, J, and K cause fewer than 1% of infections.² In particular, the CRF02_AG variant is present in different geographic areas, and is predominant in West and Central Africa.³

The molecular epidemiology of HIV-1 subtypes in Italy was recently updated, and revealed a modified picture since the initial spread of the epidemic.^{5,6} Of particular interest was the increased frequency of CRFs in patients living in Italy,^{7,8} mainly consequent to an increased immigration flow from African countries. Among HIV-1 CRFs, CRF02_AG is the most frequent in Italy.⁵ In the present study, a series of CRF02_AG genotypes was examined by phylogenetic analyses of four genomic regions. Evolutionary rates and interpatient variability were estimated for each sequence dataset. Moreover, inpatient variability was assessed in a subset of patients with multiple time point sampling.

In the period 2001–2011, 1,806 baseline genotypic resistance tests were performed from as many patients referred to the Institute of Infectious Diseases, University of Pavia, Fondazione IRCCS Policlinico San Matteo Pavia, Italy. Demographic data, including country of origin, were collected from all CRF02_AG patients. All patients were naive to integrase inhibitors, entry inhibitors (T20), and CCR5 ligand

¹Molecular Virology Unit, Microbiology and Virology Department, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy.

²Institute of Infectious Diseases, University of Pavia, Pavia, Italy.

³Department of Biomedical and Clinical Sciences, L. Sacco Hospital, Section of Infectious Diseases and Immunopathology, University of Milan, Milan, Italy.

*These authors contributed equally to this work.

inhibitors (maraviroc) at all sampling time points. Administered treatments included RT and protease (PR) inhibitors only.

The study was approved by the Institutional Review Board of the Fondazione IRCCS Policlinico San Matteo according to guidelines on the use of biological specimens for scientific purposes and in keeping with Italian law (art.13 D.Lgs 196/2003), and after having obtained informed written consent from the patients.

Virus strain subtyping was initially based on full-length protease sequences and partial RT sequences (codons 41 to 237) performed for the evaluation of RT and PR inhibitor resistance. Sequences were analyzed using the HIV-1 subtyping tool at the Monitoring Service of the Department of Molecular Biology of the University of Siena, Siena, Italy (www.hivarc.net/hiv_resistenza.asp). Virus subtyping was also confirmed by comparison of *integrase*, *gag*, *gp120*, and *gp41* sequences with the most recent reference sequences available at the Los Alamos database (www.hiv.lanl.gov/), as previously described.⁹

Stored peripheral blood mononuclear cell (PBMC) samples from patients infected with CRF02_AG were used for this study. Proviral sequence analysis of the CRF02_AG strains was performed on the genes *gag*, *pol* (*integrase*), and *env* (*gp120* and *gp41*) not submitted to antiviral drug pressure in these patients. After viral DNA extraction using the automatic Easy Mag extractor (Biomerieux, Lyon, France), parts of the *gag* gene¹⁰ (36–281 aa; p17 and p24 partial region), *pol* gene¹¹ (716–982 aa; *integrase*), and *env* gene V3 domain (254–378 aa; *gp120*)¹² and (511–683 aa; *gp41*) were amplified by nested polymerase chain reaction (PCR) as previously reported. For the amplification of the *gp41* gene the following primers were used: gp41-1 5'-TAGGAGCTTT GTTCCTTGGGTTCTTG-3' and gp41-2 5'-GGTGAATAT CCCTGCCTAACTCTATT-3' in the first-round PCR and gp41-3 5'-GGTTCTTGGGAGCAGCAGGAA-3' and gp41-4 5'-CCTACCAAGCCTCCTACTATCA-3' in the nested PCR.

All amplifications were carried out in a 100 μ l reaction volume containing 10 μ l of extracted DNA, 2 μ l AmpliTaq Gold DNA Polymerase enzyme (Applied Biosystems, Foster City, CA), and 200 μ M (each) dNTPs and 10 μ l 10 $\times$ PCR reaction buffer. The thermal profile was 10 min at 94°C for DNA polymerase activation, followed by 50 cycles consisting of 1 min at 95°C, 1 min at 52°C, 1 min and 10 s at 72°C, and a final 7 min extension at 72°C. Subsequently, the nested PCR was performed using 3 μ l from the first PCR and the following thermal profile: 10 min at 94°C for DNA polymerase activation, followed by 30 cycles consisting of 1 min at 95°C, 1 min at 50°C, 1 min and 10 s at 72°C, and a final 7 min extension at 72°C. β_2 -Microglobulin was amplified in parallel as a housekeeping gene standard.¹³ Direct sequencing of PCR products was performed using nested PCR primers and the BigDye Terminator v1.1 Cycle Sequencing kit (Applied Biosystems, Foster City, CA) with automatic

sequencer (ABI PRISM 3100 genetic analyzer DNA Sequencer, Applied Biosystems, Foster City, CA). The sequence data were assembled and edited using Sequencer version 5.0 (Gene Codes Corporation, Ann Arbor, MI) and compared with available HIV complete genome sequences using the Blast program (<http://blast.ncbi.nlm.nih.gov>).

Sequences were aligned using the Clustal W program integrated in the MEGA package, version 5.0.¹⁴ The best-fitting nucleotide substitution model was estimated using the Modeltest program and a hierarchical likelihood ratio for each of the analyzed genome regions. Maximum likelihood phylogenetic trees were constructed by using the online PhyML program.¹⁵ The reliability of specific clusters in the inferred tree was evaluated by using the SH-like approximate likelihood ratio test (aLRT).¹⁵ Cluster was defined as ≥ 3 sequences showing the node of tree with an aLRT score > 0.75 . Phylogenetic trees were drawn with FigTree v.1.4.0 (tree.bio.ed.ac.uk/software/figtree/).

HIV-1 reference sequences (including CRF02_AG sequences) were obtained from the Los Alamos HIV sequence database. Data sets of 88 complete genome CRF02_AG reference sequences from different geographic regions were compiled. Additional HIV-1 CRF02_AG sequences from other geographic regions (33 *gag* sequences, 171 *integrase* sequences, and 143 *gp120* and *gp41* sequences) were also included in each alignment. Breakpoints of recombination could be misleading in the interpretation of evolutionary dynamics among CRF02_AG strains. Thus, to avoid discordant results, recombination breakpoints were excluded from the analysis of the genomic regions as suggested by Abecasis *et al.*¹⁶ The evolutionary rates for HIV-1 CRF02_AG strains were estimated for each sequence dataset by using Bayesian Evolutionary Analysis Sampling Trees (BEAST) software version 1.7.4.¹⁷ The substitution rates were generated using a general-time-reversible substitution model with gamma distribution and invariable rates among sites, with Markov chain Monte Carlo (MCMC) runs of 100,000,000 steps sampled every 10,000 steps and analyzed with Tracer v.1.5 with a discarded burn-in of 20%. Different coalescent priors (constant population size, exponential growth, logistic growth, Bayesian skyline plot, and skyline) were tested using both strict and relaxed molecular clock models. Different clock models were compared by calculating the Bayes factor (BF).¹⁸ The difference (in $\log_e$ space) of marginal likelihood between any two models is $\log_e$ BF. Evidence against the null model (i.e., the one with a lower marginal likelihood) was indicated by $6 > (2 \log_e \text{BF}) > 2$ (positive), $10 > (2 \log_e \text{BF}) > 6$ (strong), and $(2 \log_e \text{BF}) > 10$ (very strong). BF calculations were performed with Tracer v.1.5 software (<http://beast.bio.ed.ac.uk/Tracer>). In particular, BF analysis showed that the relaxed lognormal molecular clock fitted the data better than the strict clock model. The effective sample size (ESS value) for each parameter was > 300 , indicating sufficient mixing of the Markov chain.

FIG. 1. Maximum likelihood analysis of CRF02_AG *gag* and CRF02_AG *integrase* sequences. Each tree shows the relationship between Italian sequences and reference sequences from other countries. Sequence labels include the number of patients, year of sample, and country of origin. The three-letter abbreviation country codes are as follows: Burkina Faso (BUR), Ivory Coast (CIV), Cameroon (CMR), Ghana (GHA), Italy (ITA), and Togo (TOG). Branch lengths were scaled in nucleotide substitutions per site as indicated by the bar at the bottom. Approximate likelihood ratio test (aLRT) SH-like *p*-values for supported clades ($p > 0.75$) are also indicated.

Interpatient variability was calculated as the mean proportion of nucleotide identity between all sequence pairs. Only the baseline virus sequence was taken into account in the divergence estimation. To estimate the inpatient variability, the nonsynonymous-to-synonymous substitution (dN/dS) ratio and the mean proportion of nucleotide differences between the baseline virus sequence and the more recent virus sequence for each patient were used. The ratio of dN to dS substitutions is an indication of selection pressure; values < 1 indicate negative selection, values $= 1$ indicate neutral evolution, and values > 1 indicate positive selection. The rates of synonymous (dS) and nonsynonymous (dN) substitutions were calculated using the SNAP program implemented in the Los Alamos HIV Databases.

During the 10-year study period, 1,806 patients underwent HIV-1 subtyping. Among these, 214 (13.4%) were infected with HIV non-B subtype. Of these, 34/214 (15.8%) were infected with the CRF02_AG subtype and were included in the study. Thirty out of 34 (89.4%) were from western Africa (23 from Cote D'Ivoire, three from Cameroon, two from Togo, one from Ghana, and one from Burkina Faso), 3/34 (8.8%) were born in Italy, and 1/34 (2.9%) was from Cuba. We assumed that HIV-1 infection was acquired in the country of origin. The three patients born in Italy reported sexual intercourse with Ivorian sex workers.

Four phylogenetic trees were constructed with sequences from different HIV-1 genes. In the *gag* tree, sequences from patients living in Italy were dispersed among sequences from other countries. Only one cluster including sequences from native Italian and Ivorian patients showed an aLRT score > 0.75 (Fig. 1). Likewise, in the *integrase* tree a single cluster with an aLRT score > 0.75 was composed of sequences from two native Italian and three Ivorian patients that were different from those observed in the *gag* tree. The remaining sequences observed were intermixed with other sequences. In the *gp120* tree 21/34 (61.8%) sequences were included in four clusters supported by an aLRT value > 0.75 (Fig. 2). All of these clusters included sequences from patients from Western Africa and one native Italian patient. Of note, the III-*gp120* cluster was composed only of sequences from Ivorian patients. Similarly, in the *gp41* tree, 30/34 (88.2%) sequences belonged to four different clusters (Fig. 2). None of the sequences obtained in this study clustered with sequences from other countries.

Notably, a close phylogenetic relationship was observed in four pairs of patients (#17 and #26, #29 and #39, #5 and #38, and #2 and #27). The sequences from each pair (all Ivorian) were situated on the same branch for all trees analyzed. However, a potential epidemiologic correlation in these patients could not be established.

Substitution rates were estimated as the median of substitutions per site per year of infection for each sequence dataset. Bayesian MCMC analyses under a skyline for the prior tree were used to estimate the rate of evolution. Ana-

lysis of the BF showed that the model using a lognormal relaxed clock fit the dataset better than models assuming an exponential relaxed clock or a strict clock. With the lognormal relaxed clock model, the estimated mean rate of evolution was 1.05×10^{-3} substitutions per site per year (s/s/y) (range 8.32×10^{-5} – 2.13×10^{-3}) for the *gag* gene, 4.05×10^{-4} s/s/y (range 3.59×10^{-5} – 9.68×10^{-4}) for *integrase*, 1.67×10^{-3} s/s/y (range 1.41×10^{-4} – 3.66×10^{-3}) for *gp120*, and finally 1.34×10^{-3} s/s/y (range 2.30×10^{-4} – 2.77×10^{-3}) for *gp41*.

The interpatient variability was estimated in each gene alignment by comparison with sequences from the baseline sample. Nucleotide identities ranged from 84.7% to 97.9% (mean 90.6%) in the *gag* sequences, 89.3% to 98.1% (mean 94.7%) in the *integrase* sequences, 50.4% to 100% (mean 73.6%) in the *gp120* sequences, and 85.0% to 95.4% (mean 90.1%) in the *gp41* sequences.

The inpatient variability over time (range 1–10 years) was assessed in 19/34 (55.9%) subjects (Table 1). The median inpatient nucleotide distance measured on the basis of pairwise comparison was higher (5.59%; $p < 0.05$) in *gp120* sequences with respect to *gag* (1.77%), *integrase* (1.28%), and *gp41* (1.82%) sequences. The majority of nucleotide changes within *gag*, *integrase*, and *gp41* genes involved synonymous substitutions, resulting in a ratio of $dN/dS < 1$, while in the *gp120* sequence 10/19 (52.6%; $p < 0.001$) patients had a ratio of $dN/dS > 1$.

Over the past decade, the epidemiology of HIV in Europe has been investigated at length because of its implications in public health, research, and clinical aspects.^{2,9} In the present study, we investigated the phylogenetic relationship between HIV-1 CRF02_AG strains in patients presently living in Italy, the majority of whom were immigrants from the Ivory Coast. Four genomic regions not subjected to drug selective pressure in the study group were analyzed.

Phylogenetic analyses showed that CRF02_AG strains from Ivorian patients, from other African patients, and from native Italian patients were dispersed among the sequences from Africa, Europe, and the United States, with only single well-supported clusters in the *gag* and *integrase* trees. The same diversity observed within the *gag* gene was previously reported in CRF02_AG strain sequences from Cameroonian patients,¹⁹ while no information about the phylogenetic evolution of the *integrase* gene in CRF02_AG strains is available. However, a recent study analyzing *pol* sequences of CRF02_AG strains from Cameroonian patients living in Italy⁷ showed the same intermixing of Italian sequences observed in this study for *integrase* gene sequences. However, in the *gp120* and *gp41* trees four well-supported clusters were observed. However, in the *gp120* tree the sequences outside the clusters were interspersed with other sequences from different countries, while in the *gp41* tree, almost all sequences were included in the four clusters observed.

This finding suggests the closest evolutionary origin for *gp41* gene sequences with respect to the other genes

FIG. 2. Maximum likelihood analysis of CRF02_AG V3(*gp120*) and CRF02_AG *gp41* sequences. Each tree shows the relationship between Italian sequences and reference sequences from other countries. Sequence labels include the number of patients, year of sample, and country of origin. The three-letter abbreviation country codes are as follows: Burkina Faso (BUR), Ivory Coast (CIV), Cameroon (CMR), Cuba (CUB), Ghana (GHA), Italy (ITA), and Togo (TOG). Branch lengths were scaled in nucleotide substitutions per site as indicated by the bar at the bottom. Approximate likelihood ratio test (aLRT) SH-like p -values for supported clades ($p > 0.75$) are also indicated.

V3 (gp120)

gp41

TABLE 1. INPATIENT VARIABILITY

Patient ID	Sampling time	<i>gag</i>		<i>integrase</i>		<i>gp120</i>		<i>gp41</i>	
		Sequence diversity (%)	dN/dS	Sequence diversity (%)	dN/dS	Sequence diversity (%)	dN/dS	Sequence diversity (%)	dN/dS
1	2005–2010	2.60	0.20	2.08	0.15	1.29	0.08	1.82	0.32
2	2004–2011	3.67	0.46	1.96	0.09	4.80	0.36	2.22	0.09
3	2009–2010	0.31	0.26	0.38	0.13	1.64	0.50	0.79	NA
5	2009–2010	1.92	0.09	1.28	0.11	4.49	1.51	2.63	0.30
9	2008–2011	1.92	0.26	1.57	0.42	18.84	0.71	2.88	0.45
10	2007–2009	1.94	0.37	0.13	NA	18.12	0.53	0.78	0.26
14	2003–2010	1.42	0.03	2.63	0.08	8.50	1.50	3.31	0.50
17	2005–2011	2.61	0.96	1.56	0.09	7.79	5.65	3.75	4.54
18	2005–2008	1.25	0.09	1.41	0.06	8.08	1.37	0.59	0.52
22	2006–2011	0.31	NA	3.01	0.45	5.30	0.75	1.61	0.43
25	2006–2010	0.31	0.27	1.02	0.09	10.43	2.29	1.81	0.33
26	2002–2009	2.26	0.14	1.02	0.16	11.59	1.11	4.22	0.45
27	2001–2010	1.77	0.71	1.15	0.32	5.59	1.26	1.39	0.64
28	2008–2009	0.47	0.53	0.38	NA	1.97	1.37	0.19	NA
36	2005–2011	1.26	0.08	1.81	0.10	3.39	NA	2.22	0.14
37	2005–2010	5.45	0.13	3.20	0.18	6.82	1.04	2.22	0.17
38	2008–2009	0	NA	0.38	0.52	3	0.99	0	NA
40	2007–2009	3.59	0.47	0.89	0.19	6.97	2.52	2.22	0.21
41	2008–2009	0.31	NA	0.64	0.39	1.30	0.84	0.80	0.78

Significant dN/dS values are reported in bold.
NA, not available.

analyzed. It may be possible that these clusters were determined by their role in viral replication and pathogenesis of the *env* gene. In fact, the *gp41* is a complex transmembrane protein that contains various well-defined conserved domains that play a role in fusion pore formation.²⁰ Mutations in this region have been observed to alter the fusion activity.²¹ In addition, the results of this study could be explained by the occurrence of convergent evolution in sequences of the *env* gene as also suggested by others.²² Nevertheless, CRF02_AG of African origin is broadly distributed across west/central Africa and has apparently been circulated stably there for many years. Thus, the CRF02_AG subtype likely underwent multiple recombination events during its evolution.²³ The entire *env* gene is sequenced and studied in comparison with sequences from other genomic regions, such as the *pol* and *gag* regions.^{3,24} In this study, *gp120* and *gp41* gene sequences have been analyzed separately, and a different pattern of evolution was observed. In accordance with previous observations^{3,25} *gp120* gene sequences showed higher variability and evolutionary rates, while the *gp41* gene sequences were observed to evolve with a rate similar to the *gag* gene sequences. The lower variation of *gp41* gene with respect to *gp120* gene has been also described by Steckbeck and colleagues.²⁶

In a subset of patients, an analysis of sequential samples was performed and intrapatient variability was calculated. As expected, our estimates of intrahost evolution based on nucleotides were observed to be similar to those of previous studies on other HIV-1 subtypes.^{27,28} Indeed, some clustering might be expected because the people who chose to move to Italy may have common roots in Africa, either a shared social or family link. Of note, the *gag* and *gp41* sequences exhibited similar nucleotide diversity and dN/dS ratios over time. The lower variation in the *gp41* is likely due to more strict re-

quirements for structure and function due to large conformational rearrangements that occur during the fusion process. All genomic regions selected for the analyses were not under drug pressure selection; thus, it could be hypothesized that the *gag* and *gp41* sequences were under negative selection. Most likely, the accumulation of nucleotide changes occurred before the introduction of the strains in Italy, including the three native Italian patients who acquired their CRF02_AG infection from Ivorian sexual partners. Nevertheless, closer surveillance will be necessary to better evaluate the evolution of HIV subtypes in Italy as a common route of migration in European regions.

In summary, this study offers a perspective on intrapatient and interpatient variability among CRF02_AG-infected patients living in Italy. In addition, divergent phylogenetic relationships were highlighted between different genomic regions. Thus, it is particularly important to explore evolutionary reconstruction among CRF HIV-1 strains.

Sequence Data

The sequences reported in this study have been submitted to the GenBank database under accession numbers KF922511–KF922646 and KJ612138–KJ612379.

Acknowledgments

The authors thank Daniela Sartori for manuscript editing and Laurene Kelly for revision of the English. The work was supported by the Ministero della Salute, Ricerca Corrente Grant 80207.

Author Disclosure Statement

No competing financial interests exist.

References

- Geretti AM: HIV-1 subtypes: Epidemiology and significance for HIV management. *Curr Opin Infect Dis* 2006; 19:1–7.
- Hemelaar J, Gouws E, Ghys PD, *et al.*: Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS* 2011;25:679–689.
- Faria NR, Suchard MA, Abecasis A, *et al.*: Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. *Infect Genet Evol* 2012;12:453–460.
- Taylor BS, Sobieszczyk ME, McCutchan FE, *et al.*: The challenge of HIV-1 subtype diversity. *N Engl J Med* 2008; 358:1590–1602.
- Lai A, Riva C, Marconi A, *et al.*: Changing patterns in HIV-1 non-B clade prevalence and diversity in Italy over three decades. *HIV Med* 2010;11:593–602.
- Ciccozzi M, Santoro MM, Giovanetti M, *et al.*: HIV-1 non-B subtypes in Italy: A growing trend. *New Microbiol* 2012; 35:377–386.
- Véras NM, Santoro MM, Gray RR, *et al.*: Molecular epidemiology of HIV type 1 CRF02_AG in Cameroon and African patients living in Italy. *AIDS Res Hum Retroviruses* 2011;27:1173–1182.
- Monno L, Brindicci G, Lai A, *et al.*: An outbreak of HIV-1 BC recombinants in Southern Italy. *J Clin Virol* 2012;55: 370–373.
- Abecasis AB, Wensing AM, Paraskevis D, *et al.*: HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology* 2013; 10:7.
- Paolucci S, Folli A, Gulminetti R, *et al.*: HIV-1 plasma variants encoding truncated reverse transcriptase (RT) in a patient with high RT-specific CD8⁺ memory T-cell response. *Current HIV Res* 2009;7:302–310.
- Baldanti F, Paolucci S, Gulminetti R, *et al.*: Early emergence of raltegravir resistance mutations in patients receiving HAART salvage regimens. *J Med Virol* 2010;82: 116–122.
- Svicher V, Alteri C, Montano M, *et al.*: Performance of genotypic tropism testing on proviral DNA in clinical practice: Results from the DIVA study group. *New Microbiol* 2012;35:17–25.
- Watzinger F, Suda M, Preuner S, *et al.*: Real-time quantitative PCR assays for detection and monitoring of pathogenic human viruses in immunosuppressed pediatric patients. *J Clin Microbiol* 2004;42:5189–5198.
- Tamura K, Peterson D, Peterson N, *et al.*: MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;28:2731–2739.
- Guindon S, Dufayard JF, Lefort V, *et al.*: New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 2010; 59:307–321.
- Abecasis AB, Lemey P, Vidal N, *et al.*: Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: Subtype G is a circulating recombinant form. *J Virol* 2007;81:8543–8551.
- Drummond AJ, Suchard MA, Xie D, *et al.*: Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012;29:1969–1973.
- Drummond AJ and Rambaut A: BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007; 7:214.
- Tongo M, Martin DP, Zembe L, *et al.*: Characterization of HIV-1 gag and nef in Cameroon: further evidence of extreme diversity at the origin of the HIV-1 group M epidemic. *Virol J* 2013;10:29.
- Freed EO, Delwart EL, Buchschacher GL Jr, and Paganin AT: A mutation in the human immunodeficiency virus type 1 transmembrane glycoprotein gp41 dominantly interferes with fusion and infectivity. *Proc Natl Acad Sci USA* 1992;89:70–74.
- Cao J, Bergeron L, Helseth E, Thali M, Repke H, and Sodroski J: Effects of amino acid changes in the extracellular domain of the human immunodeficiency virus type 1 gp41 envelope glycoprotein. *J Virol* 1993;67:2747–2755.
- Strunnikova N, Ray SC, Livingston RA, Rubalcaba E, and Viscidi RP: Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. *J Virol* 1995;69:7548–7558.
- Zhang M, Foley B, Schultz AK, *et al.*: The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology* 2010;7:25.
- Ajoge HO, Gordon ML, de Oliveira T, *et al.*: Genetic characteristics, coreceptor usage potential and evolution of Nigerian HIV-1 subtype G and CRF02_AG isolates. *PLoS One* 2011;6:e17865.
- Abecasis AB, Vandamme AM, and Lemey P: Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J Virol* 2009;83:12917–12924.
- Steckbeck JD, Craigo JK, Barnes CO, and Montelaro RC: Highly conserved structural properties of the C-terminal tail of HIV-1 gp41 protein despite substantial sequence variation among diverse clades: Implications for functions in viral replication. *J Biol Chem* 2011;286:27156–27166.
- Novitsky V, Wang R, Rossen Khan R, *et al.*: Intra-host evolutionary rates in HIV-1C env and gag during primary infection. *Infect Genet Evol* 2013;19:361–368.
- Walker PR, Pybus OG, Rambaut A, *et al.*: Comparative population dynamics of HIV-1 subtypes B and C: Subtype-specific differences in patterns of epidemic growth. *Infect Genet Evol* 2005;5:199–208.

Address correspondence to:

Fausto Baldanti

Molecular Virology Unit

Microbiology and Virology Department

Fondazione IRCCS Policlinico San Matteo

27100 Pavia

Italy

E-mail: f.baldanti@smatteo.pv.it