Journal of General Virology (2008), **89**, 3119–3125 DOI 10.1099/vir.0.2008/003418-0

Short Communication

Recombination analysis and structure prediction show correlation between breakpoint clusters and RNA hairpins in the *pol* gene of human immunodeficiency virus type 1 unique recombinant forms

Andrea Galli,[1] Alessia Lai,[1] Stefano Corvasce,[1] Francesco Saladini,[2] Chiara Riva,[1] Lorenzo Dehò,[1] Ilaria Caramma,[1] Marco Franzetti,[1] Laura Romano,[2] Massimo Galli,[1] Maurizio Zazzi[2] and Claudia Balotta[1]

[1]Department of Clinical Sciences 'L. Sacco', University of Milan, Milan, Italy

[2]Department of Molecular Biology, University of Siena, Siena, Italy

Correspondence
Andrea Galli
Andrea.Galli@unimi.it

Received 24 April 2008
Accepted 25 July 2008

Recombination is recognized as a primary force in human immunodeficiency virus type 1 (HIV-1) evolution, increasing viral diversity through reshuffling of genomic portions. The strand-switching activity of reverse transcriptase is required to complete HIV-1 replication and can occur randomly throughout the genome, leading to viral recombination. Some recombination hotspots have been identified and found to correlate with RNA structure or sequence features. The aim of this study was to evaluate the presence of recombination hotspots in the *pol* gene of HIV-1 and to assess their correlation with the underlying RNA structure. Analysis of the recombination pattern and breakpoint distribution in a group of unique recombinant forms (URFs) detected two recombination hotspots in the *pol* region. Two stable and conserved hairpins were consistently predicted corresponding to the identified hotspots using six different RNA-folding algorithms on the URF parental strains. These findings suggest that such hairpins may play a role in the higher recombination rates detected at these positions.

Since its identification, human immunodeficiency virus type 1 (HIV-1) has been recognized as one of the fastest-evolving retroviruses. Its extensive genetic variability is due to several key factors. Firstly, HIV-1 reverse transcriptase (RT) has a very low accuracy, due to the lack of 3′-exonuclease proofreading activity, as well as to an innate low fidelity, and produces about one mutation per genome during each replication cycle (Preston *et al.*, 1988; Roberts *et al.*, 1988). Secondly, the virus has a high replication capacity, producing up to 10 billion new virions per day (Coffin, 1995), with a mean generation time of about 2.6 days (Perelson *et al.*, 1996). These two features greatly facilitate evasion of the immune response and antiretroviral therapy within the host and have led to the evolution of the nine subtypes (A–D, F–H, J, K) and six subsubtypes (A1–A4, F1, F2) recognized to date in the HIV-1 M group (Thomson *et al.*, 2002; Wainberg, 2004). In addition, genetic recombination was recognized early on as a strong driving force in HIV-1 evolution (Coffin, 1979; Hu & Temin, 1990; Robertson *et al.*, 1995; Temin, 1993) and is now considered one of the most important mechanisms of viral evolution (Rambaut *et al.*, 2004; Thomson & Nájera, 2005). During reverse transcription, the viral RT can switch with high frequency (Jetzt *et al.*, 2000; Levy *et al.*, 2004) between the two templates of the genomic RNA dimer (strand-switching activity), generating a recombinant DNA molecule when infection is sustained by a heterozygous virion. To date, as many as 43 circulating recombinant forms (CRFs) have been classified, and a growing number of unique recombinant forms (URFs) is being detected worldwide (Leitner *et al.*, 2005; Thomson & Nájera, 2005).

Recombination may occur randomly along the viral genome (Clavel *et al.*, 1989; Goodrich & Duesberg, 1990; Hu & Temin, 1990); nevertheless, some preferential sites have been identified and characterized accurately (Balakrishnan *et al.*, 2003; Dykes *et al.*, 2004; Galetto *et al.*, 2006; Kim *et al.*, 1997; Magiorkinis *et al.*, 2003; Moumen *et al.*, 2001; Zhuang *et al.*, 2002). Different models have been proposed to explain the molecular mechanism(s) underlying strand switching (Galetto & Negroni, 2005; Zhang *et al.*, 2000), some of which suggest a role of hairpin structures either in hampering the RT or in direct template interaction (Galetto *et al.*, 2006; Moumen *et al.*, 2003; Roda *et al.*, 2002), as well as in the mandatory minus-strand strong-stop DNA strand switch (Berkhout *et al.*, 2001).

The GenBank/EMBL/DDBJ accession numbers for the sequences determined in this study are EF488562–EF488607.

To assess the presence of recombination hotspots in the *pol* gene of HIV-1, we analysed the distribution of recombination points in the protease (PR) and RT regions of a group of 46 URFs. Sequences were obtained using in-house procedures (Peduzzi *et al.*, 2002) or commercial kits (Viroseq HIV-1 Genotyping System, Celera Diagnostics; Trugene HIV-1 Genotyping System, Bayer Healthcare) from clinical practice and specific molecular epidemiological studies (CASCADE Collaboration, 2000; Giuliano *et al.*, 2006; Wensing *et al.*, 2005). Treatment history was known for 29/46 subjects (63.0 %); 20 (43.5 %) of these were known to have been treated at the time of sampling, whilst nine were therapy naive.

Sequences were reduced to a length of 987 nt, starting with base 1 of the PR-coding region, and aligned using BioEdit version 7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). The standard 2005 HIV-1 reference alignment available at the Los Alamos National Laboratory (LANL) website (http://www.hiv.lanl.gov/content/hiv-db/mainpage.html) was used as the subtype reference. Sequences were recognized as URFs by phylogenetic analysis using PHYLIP version 3.63 (http://evolution.genetics.washington.edu/phylip.html) to build neighbour-joining trees with an F84 matrix distance and 1000 bootstrap replicates. Recombination patterns were then carefully determined by performing bootscan analysis with Simplot version 3.5.1 (Lole *et al.*, 1999), using a window sliding of 300 nt in 10 nt steps, with 1000 bootstrap replicates. The subtype pattern of each URF was confirmed by phylogenetic analysis of the recombination fragments using the same parameters as for full-length sequences. To compare breakpoint locations, the same analyses were performed on a control dataset of 50 URFs obtained from GenBank.

Phylogenetic and bootscanning analyses of URF sequences from both field and GenBank datasets showed wide heterogeneity regarding subtype composition. Overall, genomic fragments belonging to all known pure subtypes (A–D, F–H, J and K) and to four CRFs (01, 02, 06 and 09) were detected in the URFs. Recombination patterns were highly diverse and several low-frequency patterns could be observed in the two datasets (Table 1). Interestingly, some highly frequent subtype forms (B/F, A/G, A/D and B/C) were identified in both case files. Analysis of recombination patterns showed a broad distribution of breakpoints along the entire region in both our field isolates and the GenBank URFs. However, two clusters of breakpoints could be observed around nt 300 and 700, suggesting the presence of recombination hotspots in these areas (Fig. 1). In particular, 37/46 field isolates (80 %) and 39/50 GenBank sequences (78 %) had a recombination point in the region nt 200–420, whilst 18/46 field isolates (39 %) and 24/50 GenBank sequences (48 %) had a recombination point in the region nt 650–800. As bootscanning data could be questionable for about 150 nt at the 5′ and 3′ ends of the region, we repeated the analyses on all samples from GenBank and on nine field isolates for which a longer sequence was available. The same breakpoint distribution

Table 1. Subtype composition of URFs

Mosaic patterns are given as absolute numbers for the 46 URFs identified by us (dataset A) and for the 50 URFs obtained from GenBank (dataset B).

Patterns	A/B	A/C	A/D	A/F	A/G	A/G/K	A/H/J	A/K	B/C	B/C/F	B/D	B/F	B/G	C/D	C/J	D/F	F/G/K	F/K	G/H	H/J	J/U	01/B	02/A	02/B	02/G	02/06	02/09	09/G
Dataset A	1	–	2	–	5	4	–	3	2	1	2	15	2	–	–	–	1	–	–	–	–	–	2	2	–	1	2	1
Dataset B	1	4	11	2	3	–	1	1	–	3	–	13	1	1	1	1	–	1	1	1	1	1	–	–	1	–	–	–

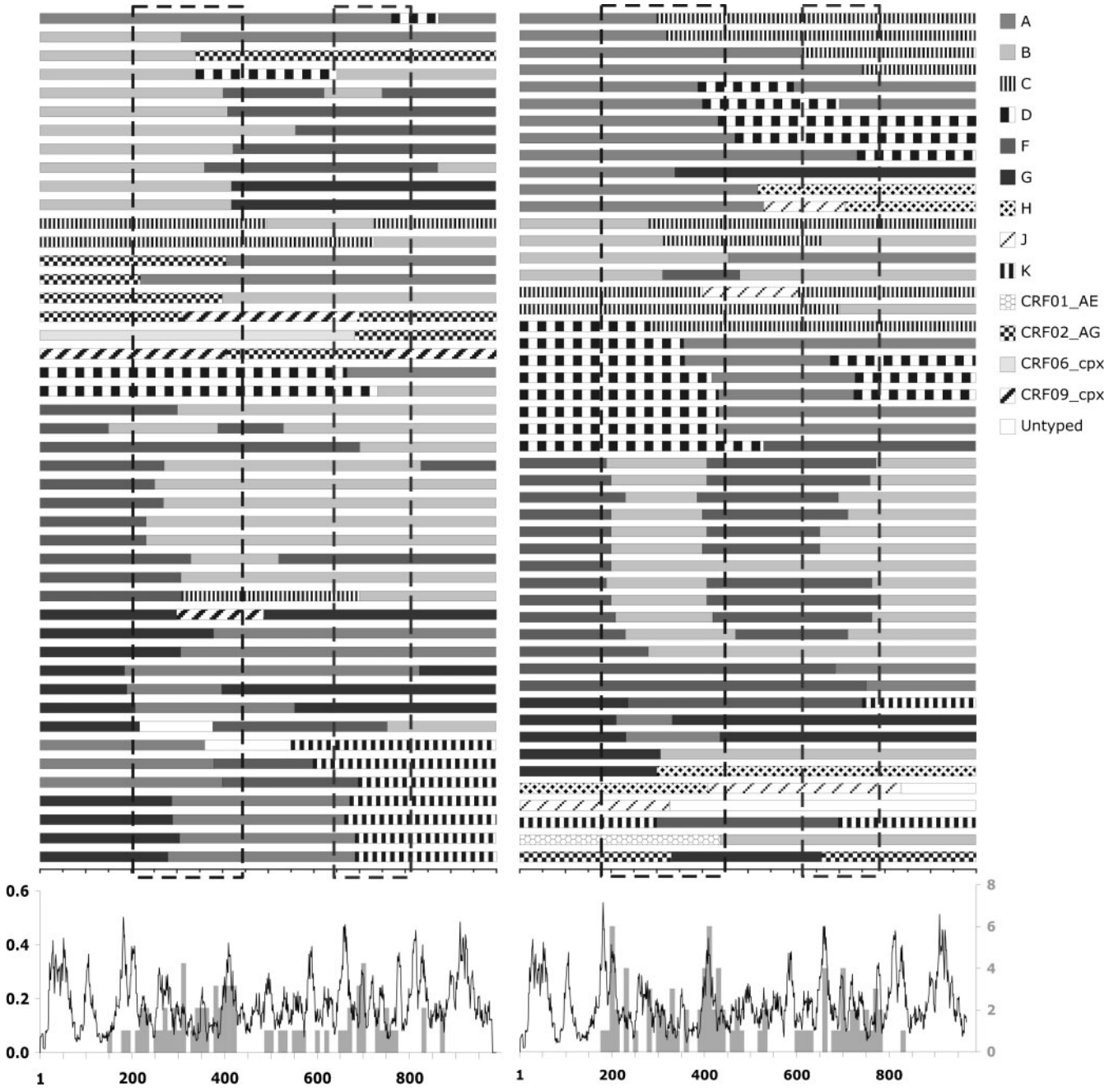

Fig. 1. Recombination patterns of URFs. Mosaic structures of field isolates (left) and GenBank sequences (right) are shown at the top. Hatched boxes indicate areas of high concentration of recombination points. At the bottom, entropy scores (solid line) and breakpoint frequency distributions (grey boxes) are shown superimposed for each dataset. The *x* axis shows the base numbering relative to nt 1 of PR, whilst the *y* axes indicate the Shannon entropy value (left, black font) and breakpoint frequency (right, grey font).

was confirmed when considering this subset of 59 longer sequences encompassing 1287 nt.

Although the vast majority of the URFs examined (94/96; 98 %) were found to be inter-subtype recombinants, thus containing enough genetic information to allow the detection of recombination, we were concerned that local stretches of highly homologous sequence could impair the detection of recombination sites. To evaluate this possibility, Shannon entropy was calculated for the two datasets as an estimate of genetic variability in the region under investigation, using the Entropy tool available on the LANL website

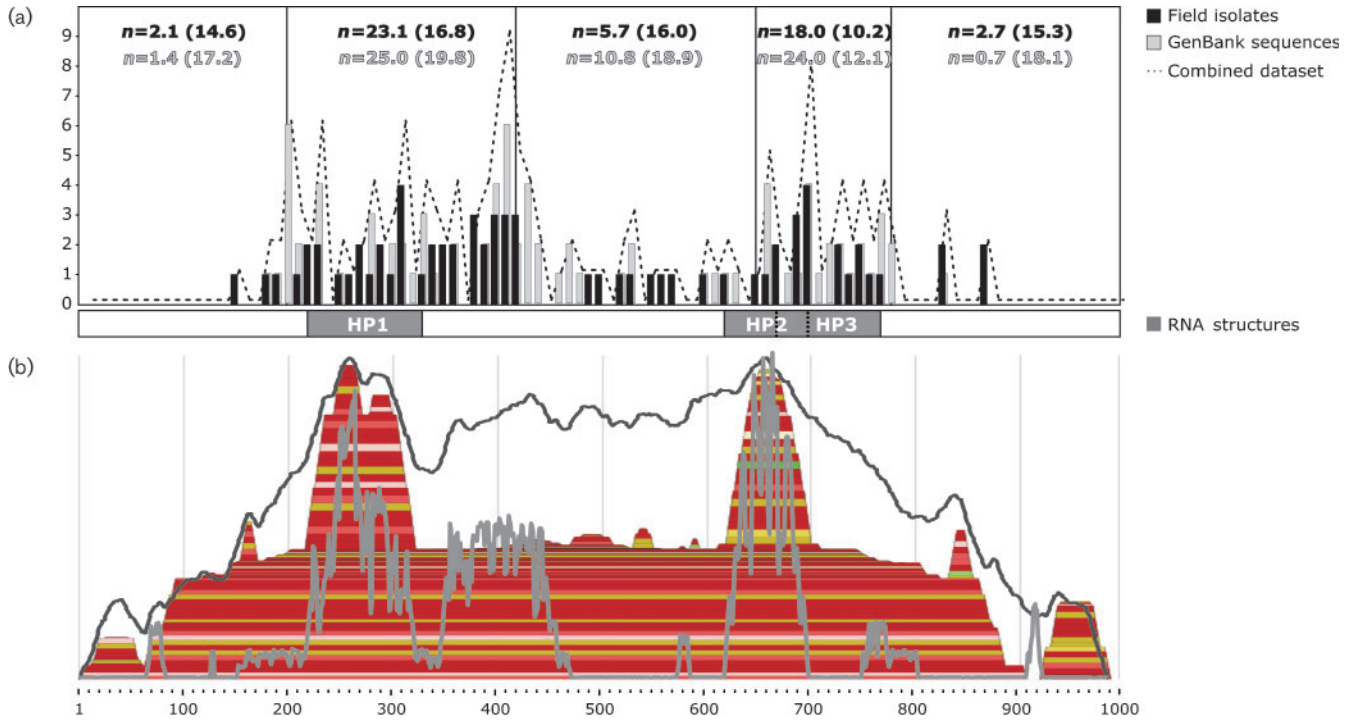

Fig. 2. Distribution of breakpoints and predicted hairpins. (a) The upper part of the diagram displays the base position (*x* axis) and absolute frequency (*y* axis) of the recombination points detected in our field isolates (shown in black), in the GenBank sequences (shown in grey) and in the combined dataset (dotted line). The normalized number of breakpoints detected in each region is indicated at the top, with the expected values for a random distribution shown in parentheses. At the bottom, the shaded boxes indicate the location of the identified hairpin-like structures (HP1, HP2 and HP3). (b) A mountain plot graph of the structures predicted by RNAalifold (full colour), Alidot (solid black line) and Pfold (solid grey line). Each base pair is represented as a horizontal box extending between the two bases; its height is relative to the thermodynamic likeliness of the base pair. The colour of the box indicates sequence variation at each position: red, ochre, green, cyan, blue and violet indicate one to six different base pairs consistent with the structure detected. Lighter shades of colour represent one or two non-compatible pairs. Alidot and Pfold structures are given as outlines. Positions are relative to nt 1 of PR in both panels.

(Korber *et al.*, 1994). Entropy data indicated that sequence variability fluctuated along the sequence but never dropped to zero, suggesting that the genetic information contained at every position was relevant for subtype assignment (Fig. 1). Moreover, the regions without recombination points corresponded to areas with a fairly high genetic variability, indicating that the detection of breakpoints, in these datasets, was not significantly biased by sequence variability or limitations of the bootscanning method used.

To investigate further the location of recombination hotspots, we plotted the position and frequency of breakpoints in both datasets and compared their distribution with the position of the breakpoint clusters identified by the visual analysis of recombination patterns (Fig. 2a). After normalizing the number of breakpoints to the length of each region, two recombination hotspots could be defined corresponding to the previously identified breakpoint clusters: one encompassing the PR/RT border (nt 200–420) and one in the RT (nt 650–780). A Poisson heterogeneity test was used to compare the observed layout with an expected distribution under the null hypothesis of homogeneous dispersion. This analysis indicated a significant and comparable non-random distribution of breakpoints in the field isolates, the GenBank URFs and the combined dataset, supporting the presence of recombination hotspots in the *pol* region of HIV-1 ($P<0.0001$ for each comparison).

Next, we assessed the possible association between breakpoint location and predicted RNA secondary structures. To evaluate the RNA structural features potentially favouring recombination at a specific position, ideally we needed to perform folding prediction on the parental strains of our field URFs. Therefore, through phylogenetic analysis, we identified the pure subtype reference strains most closely related to each recombination fragment of our field URFs. A reference dataset was then built, choosing the following reference strains (GenBank accession numbers): A (AF484509); B (K03455, AY423387); C (U52953, U46016); D (K03454, U88824); F (AJ249238, AY771447); G (AF061642); K (AJ249235, AJ249239). RNA structures

were predicted using two different approaches, conventional single-sequence algorithms and comparative folding predictions on the whole alignments, the latter strategy being considered more reliable. For single-sequence analyses, mFold (version 3.2) (Mathews *et al.*, 1999; Zuker, 2003), RNAfold (version 1.6.3) (Hofacker *et al.*, 1994; McCaskill, 1990) and sFold (version 2.0) (Ding *et al.*, 2004; Korber *et al.*, 1994) were used. Alignment foldings were predicted using RNAalifold (version 1.6.3) (Hofacker *et al.*, 2002), Alidot (version 2.0) (Hofacker & Stadler, 1999) and Pfold (Knudsen & Hein, 2003). The folding analysis of single sequences revealed two hairpin-like structures present in all parental subtypes, each corresponding to a recombination cluster (structures not shown). The position of such structures was invariable, even though their actual conformation could be very different among subtypes; the regions between them did not show any stable conserved structures. A hairpin-like structure was identified at position 218–324 (HP1), displaying a stable backbone with a varying number of lateral bulges. Two alternative hairpins could be detected at positions 620–702 (HP2) and 672–771 (HP3) (data not shown). HP2 was the most stable and defined structure among those observed, whilst HP3 was identified as a possible alternative to HP2, resulting from a slight downstream shift of the folded region, in subtypes B, F and K. HP2 showed an almost invariable structure with a high degree of similarity among different subtypes, whilst HP3 varied more broadly among the three strains in which it was identified.

Alignment folding analyses were then performed on the whole datasets. In Fig. 2(b), predicted foldings are shown using mountain plot graphics, which give additional information about base pair variation, as indicated by the colour associated with each base pair, and the thermodynamic likeliness of structure elements, as indicated by their height. The methods producing a consensus structure (Pfold and RNAalifold) and the algorithm seeking conserved structures (Alidot) gave similar results. In agreement with single-sequence analysis, the comparative methods predicted HP1 and HP2 as the most conserved and thermodynamically relevant hairpins, with structures very similar to those produced by the single-sequence algorithms.

Recent studies have demonstrated the uneven distribution of recombination breakpoints along the genome of HIV-1 and the presence of recombination hotspots in specific regions (Zhuang *et al.*, 2002). The dimer initiation sequence and the transactivating region hairpin in the long terminal repeat region, a hairpin located in the C2 portion of the gp120 region in *env* and a G-rich region in the *gag* gene have all been shown to promote recombination in single-cycle infection assays or *in vitro* reconstituted systems, i.e. in the absence of *in vivo* selective pressures (Balakrishnan *et al.*, 2003; Dykes *et al.*, 2004; Galetto *et al.*, 2006; Moumen *et al.*, 2001).

Our data showed that clusters of recombination breakpoints are detectable in specific PR and RT regions. Unlike the cited studies, our observations were based on sequences obtained from patients, some of whom were under therapy at the time of sampling or had been treated previously. These viruses had therefore evolved under *in vivo* selective pressures that could have biased the observed distribution of recombination points due to the advantage of some specific recombinants compared with others; further *in vitro* analyses are required to elucidate the possible role of *in vivo* selection on the distribution of breakpoints reported here.

Nevertheless, it is noteworthy that two very stable and conserved hairpin structures (HP1 and HP2) were predicted corresponding to the recombination clusters using two independent approaches, whilst no relevant structure could be detected in areas of low recombination frequency. These results suggest that the underlying RNA structure might play a role in the higher recombination rate detected at those positions. Recent studies have shown that RNA hairpin structures indeed correlate with recombination hotspots in other regions of the HIV-1 genome (Balakrishnan *et al.*, 2001; Galetto *et al.*, 2006; Moumen *et al.*, 2001). According to the proposed mechanisms, hairpins could promote recombination by direct interaction, through either their stems or loops, or by slowing down the RT during reverse transcription (Berkhout *et al.*, 2001; Galetto *et al.*, 2006; Lanciault & Champoux, 2006; Moumen *et al.*, 2001; Roda *et al.*, 2002).

In summary, we have predicted two conserved RNA hairpins in the *pol* gene of several HIV-1 subtypes that correlate with two recombination hotspots identified in a group of recombinant sequences isolated from patients. *In vitro* studies with different isolates and site-directed mutants are required to complement this evidence and clarify the role of the predicted hairpins, especially in the absence of selective pressure. A more detailed knowledge of HIV-1 recombination is important not only for a better understanding of HIV-1 biology and evolution, but also for its implications in the development of treatment strategies, as recombination of genome regions where different drug-resistance mutations may accumulate plays a role in the generation of multidrug-resistant viruses.

Acknowledgements

We wish to thank Drs M. Giuliano and A. D'Arminio Monforte for kindly providing samples from the SIMBA and TARGET studies, respectively, and J. Cibella for excellent technical assistance. This work was performed with the support of grants from Istituto Superiore di Sanità, Rome to C. B. (20G.18) and M. Z. (30G.58) and PAR/2005 of the University of Siena.

References

Balakrishnan, M., Fay, P. J. & Bambara, R. A. (2001). The kissing hairpin sequence promotes recombination within the HIV-I 5′ leader region. *J Biol Chem* **276**, 36482–36492.

Balakrishnan, M., Roques, B. P., Fay, P. J. & Bambara, R. A. (2003). Template dimerization promotes an acceptor invasion-induced transfer mechanism during human immunodeficiency virus type 1 minus-strand synthesis. *J Virol* **77**, 4710–4721.

Berkhout, B., Vastenhouw, N. L., Klasens, B. I. & Huthoff, H. (2001). Structural features in the HIV-1 repeat region facilitate strand transfer during reverse transcription. *RNA* **7**, 1097–1114.

CASCADE Collaboration (2000). Effect of ignoring the time of HIV seroconversion in estimating changes in survival over calendar time in observational studies: results from CASCADE. *AIDS* **14**, 1899–1906.

Clavel, F., Hoggan, M. D., Willey, R. L., Strebel, K., Martin, M. A. & Repaske, R. (1989). Genetic recombination of human immunodeficiency virus. *J Virol* **63**, 1455–1459.

Coffin, J. M. (1979). Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J Gen Virol* **42**, 1–26.

Coffin, J. M. (1995). HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **267**, 483–489.

Ding, Y., Chan, C. Y. & Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* **32**, W135–W141.

Dykes, C., Balakrishnan, M., Planelles, V., Zhu, Y., Bambara, R. A. & Demeter, L. M. (2004). Identification of a preferred region for recombination and mutation in HIV-1 *gag*. *Virology* **326**, 262–279.

Galetto, R. & Negroni, M. (2005). Mechanistic features of recombination in HIV. *AIDS Rev* **7**, 92–102.

Galetto, R., Giacomoni, V., Veron, M. & Negroni, M. (2006). Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. *J Biol Chem* **281**, 2711–2720.

Giuliano, M., Galluzzo, C. M., Germinario, E. A., Amici, R., Bassani, L., Deho, L., Vyankandondera, J., Mmiro, F., Okong, P. & Vella, S. (2006). Selection of resistance mutations in children receiving prophylaxis with lamivudine or nevirapine for the prevention of postnatal transmission of HIV. *J Acquir Immune Defic Syndr* **42**, 131–133.

Goodrich, D. W. & Duesberg, P. H. (1990). Retroviral recombination during reverse transcription. *Proc Natl Acad Sci U S A* **87**, 2052–2056.

Hofacker, I. L. & Stadler, P. F. (1999). Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput Chem* **23**, 401–414.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh Chem* **125**, 167–188.

Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J Mol Biol* **319**, 1059–1066.

Hu, W. S. & Temin, H. M. (1990). Retroviral recombination and reverse transcription. *Science* **250**, 1227–1233.

Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D. & Dougherty, J. P. (2000). High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J Virol* **74**, 1234–1240.

Kim, J. K., Palaniappan, C., Wu, W., Fay, P. J. & Bambara, R. A. (1997). Evidence for a unique mechanism of strand transfer from the transactivation response region of HIV-1. *J Biol Chem* **272**, 16769–16777.

Knudsen, B. & Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* **31**, 3423–3428.

Korber, B. T., Kunstman, K. J., Patterson, B. K., Furtado, M., McEvilly, M. M., Levy, R. & Wolinsky, S. M. (1994). Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J Virol* **68**, 7467–7481.

Lanciault, C. & Champoux, J. J. (2006). Pausing during reverse transcription increases the rate of retroviral recombination. *J Virol* **80**, 2483–2494.

Leitner, T., Korber, B., Daniels, M., Calef, C. & Foley, B. (2005). HIV-1 subtype and circulating recombinant form (CRF) reference sequences. In *HIV Sequence Compendium*. Los Alamos National Laboratory, NM, USA.

Levy, D. N., Aldrovandi, G. M., Kutsch, O. & Shaw, G. M. (2004). Dynamics of HIV-1 recombination in its natural target cells. *Proc Natl Acad Sci U S A* **101**, 4204–4209.

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W. & Ray, S. C. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* **73**, 152–160.

Magiorkinis, G., Paraskevis, D., Vandamme, A. M., Magiorkinis, E., Sypsa, V. & Hatzakis, A. (2003). In vivo characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity. *J Gen Virol* **84**, 2715–2722.

Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911–940.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105–1119.

Moumen, A., Polomack, L., Roques, B., Buc, H. & Negroni, M. (2001). The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination. *Nucleic Acids Res* **29**, 3814–3821.

Moumen, A., Polomack, L., Unge, T., Veron, M., Buc, H. & Negroni, M. (2003). Evidence for a mechanism of recombination during reverse transcription dependent on the structure of the acceptor RNA. *J Biol Chem* **278**, 15973–15982.

Peduzzi, C., Pierotti, P., Venturi, G., Romano, L., Mazzotta, F. & Zazzi, M. (2002). Performance of an in-house genotypic antiretroviral resistance assay in patients pretreated with multiple human immunodeficiency virus type 1 protease and reverse transcriptase inhibitors. *J Clin Virol* **25**, 57–62.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582–1586.

Preston, B. D., Poiesz, B. J. & Loeb, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science* **242**, 1168–1171.

Rambaut, A., Posada, D., Crandall, K. A. & Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nat Rev Genet* **5**, 52–61.

Roberts, J. D., Bebenek, K. & Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science* **242**, 1171–1173.

Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. (1995). Recombination in HIV-1. *Nature* **374**, 124–126.

Roda, R. H., Balakrishnan, M., Kim, J. K., Roques, B. P., Fay, P. J. & Bambara, R. A. (2002). Strand transfer occurs in retroviruses by a pause-initiated two-step mechanism. *J Biol Chem* **277**, 46900–46911.

Temin, H. M. (1993). Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. *Proc Natl Acad Sci U S A* **90**, 6900–6903.

Thomson, M. M. & Nájera, R. (2005). Molecular epidemiology of HIV-1 variants in the global AIDS pandemic: an update. *AIDS Rev* **7**, 210–224.

Thomson, M. M., Pérez-Alvarez, L. & Nájera, R. (2002). Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. *Lancet Infect Dis* **2**, 461–471.

Wainberg, M. A. (2004). HIV-1 subtype distribution and the problem of drug resistance. *AIDS* **18** (*Suppl. 3*), S63–S68.

Wensing, A. M., van de Vijver, D. A., Angarano, G., Asjo, B., Balotta, C., Boeri, E., Camacho, R., Chaix, M. L., Costagliola, D. & other authors (2005). Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. *J Infect Dis* **192**, 958–966.

Zhang, J., Tang, L. Y., Li, T., Ma, Y. & Sapp, C. M. (2000). Most retroviral recombinations occur during minus-strand DNA synthesis. *J Virol* **74**, 2313–2322.

Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D. & Dougherty, J. P. (2002). Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J Virol* **76**, 11273–11282.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406–3415.