Available online at www.sciencedirect.com
ScienceDirect
Infection, Genetics and Evolution 7 (2007) 717–723
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

The distribution of HIV-1 recombination breakpoints

Jun Fan [a], Matteo Negroni [b], David L. Robertson [a,*]

[a] Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK
[b] Unité de Régulation Enzymatique des Activités Cellulaires, Institut Pasteur, Paris 75724, Cedex 15, France

Received 30 March 2007; received in revised form 18 July 2007; accepted 24 July 2007
Available online 28 July 2007

Abstract

We find that recombination breakpoints are non-randomly distributed across the genomes of HIV-1 intersubtype recombinants. In particular we find two recombination prone regions, ''hot spots'', located approximately either side of the envelope gene. To investigate this, we test whether there is a correlation between the distribution of the recombinant breakpoints with (1) genetic similarity, (2) predicted locations of secondary RNA structure, (3) regions identified as recombinant hot spots from experimental studies and (4) the predicted locations of positively selected sites. No detectable relationship with RNA secondary structure was found. A weak relationship with genetic similarity exists but it does not account for the recombination hot spots. The comparison with the published experimental studies indicated that the identified recombination hot spots differ in their locations, indicating that selection is having an impact on HIV-1 recombinant structures in infected individuals. We observe an association between recombination prone regions and strong positive selection across the envelope gene in support of this hypothesis.

Keywords: HIV-1; AIDS; Recombination; Breakpoints; Strand-switching

1. Introduction

HIV has a highly diverse viral population both within an infected individual and in the human population (Rambaut et al., 2004). Interestingly, it is not the DNA-to-RNA transcriptional stage of replication that is the key to the persistence of the AIDS virus (HIV), rather it is the viral RNA-to-DNA (reverse transcription) stage, prior to integration, at which the majority of new viral variants are generated. These mutant HIVs arise both as a result of reverse transcriptase's (RT) error-prone nature while copying and RT's propensity to ''jump'' on the same RNA strand or switch between the two viral RNA strands, generating a recombinant form (Temin, 1993; Negroni and Buc, 2001). It has been directly measured that this error rate is as high as 5.4×10^{-5} mutations per site per replication (Gao et al., 2004) and that two to three recombination events occur per genome per replication cycle (Jetzt et al., 2000). This on-going generation of divergent viruses and diversifying selection prior to progression to AIDS produces the variation that permits HIV to escape the host immune response (Wolinsky et al., 1996; Yang et al., 2003; Choisy et al., 2004).

Early in the identification of divergent HIV-1 strains it was clear that some viruses formed distinct clusters or clades in phylogenetic trees (Myers et al., 1992). For example, the HIV strains first discovered, and responsible for the majority of Western infections, clustered together and came to be known as subtype B. Currently nine HIV-1 group M subtypes are defined: A–D, F–H, J and K, while former subtypes E and I have been redesignated as circulating recombinant forms CRF01 and CRF04, respectively, of which 34 are now designated (Robertson et al., 2000). The CRFs are defined as recombinants descended from the same recombination event(s). They are considered as more significant than the unique recombinant forms (URFs) as some, particularly CRF01 and CRF02, are responsible for disproportionate numbers of infections globally.

The generation of HIV recombinants is thought to occur by a copy choice mechanism during reverse transcription (Vogt, 1971; Negroni and Buc, 2001) and high sequence identity in the region of a potential breakpoint has been shown to promote strand-switching by reverse transcriptase (Zhang and Temin, 1994; Baird et al., 2006). In addition features of the RNA such as homopolymeric runs (Klarmann et al., 1993) and secondary RNA structure (Moumen et al., 2003; Galetto et al., 2004) have been found to promote strand-switching and hence recombination.

* Corresponding author. Tel.: +44 161 275 5089; fax: +44 161 275 5082.
E-mail address: david.robertson@manchester.ac.uk (D.L. Robertson).

1567-1348/$ – see front matter
doi:10.1016/j.meegid.2007.07.012

Here we investigate potential factors influencing the location of breakpoints in HIV-1 intersubtype recombinants (both URFs and CRFs) obtained from the LANL HIV Sequence Database. We find, as others have (Magiorkinis et al., 2003), that there is non-random distribution of recombination breakpoints across HIV's genome and a weak relationship with genetic similarity. Furthermore, we find that the peaks in the breakpoint distribution are not explained by predicted locations of secondary RNA structure, or regions identified as recombinant hot spots in experimental studies. Interestingly a strong association was observed between the strongest recombination prone regions and the genomic region (the envelope) exhibiting the highest levels of positive selection.

2. Methods

2.1. Sequence alignments

The HIV-1 group M near complete genome alignment was retrieved from the Los Alamos National Laboratory (LANL) HIV Sequence Database (http://hiv-web.lanl.gov/). The genome sequence of the recombinant strain Z321 was concatenated from available sequences (accession numbers U76035 and M15896, respectively) and included in the alignment using the profile alignment function in CLUSTAL W (Thompson et al., 1994). The genome alignment was split into nine alignments corresponding to the open reading frames of *gag*, *pol*, *vif*, *tat*, *rev*, *vpr*, *vpu*, *env* and *nef*. Ambiguous regions of the alignment, regions with high numbers of insertions and deletions, and stop codons were excluded. Each gene sequence alignment was altered so that codon boundaries corresponded to the amino acid sequence alignment and concatenated to form the new genome alignment. This alignment is available on request.

Alignments were analysed with appropriate reference sequences for each sequence that has been identified as an intersubtype recombinant in the LANL alignment. These include the recombinant, a representative of each of the subtypes implicated in the recombination structure (indicated in the recombinant's name in the LANL alignment) and an outgroup sequence. The ''parental'' representative sequences used were either a consensus sequence, or if less than three sequences are available for a particular subtype, one strain was chosen. For each CRF only one representative strain was included in our analysis and, if available in the LANL alignment, the CRF consensus sequence used.

2.2. Breakpoint detection

For each alignment, modified versions of informative sites analysis and diversity plotting (Robertson et al., 1995; van Cuyck et al., 2005; Fan et al., in preparation) were used to initially detect intersubtype recombination breakpoints. The main modifications were the use of a *T*-test to aid in the detection of crossovers in the diversity plots and the scanning of polymorphic sites in optimised windows, in addition to informative sites, to increase the number of sites used in the maximum Chi-squared test (Fan et al., in preparation). Breakpoints were further verified by performing 1000 neighbor-joining bootstrap replicates, implemented with PAUP* 4.0b10 (Swofford, 2002), on alignments from either side of putative breakpoints. Breakpoints were confirmed if the bootstrap replicates on both sides of the breakpoint were 90% or higher. In ambiguous cases, or where multiple breakpoints are in close proximity, the program SimPlot (Lole et al., 1999) was also used to determine whether or not a breakpoint was present.

2.3. Distribution of breakpoints

The numbers of recombination breakpoints present in each gene region and in incremented windows of sizes 200 and 500 nucleotides were counted. A Chi-squared goodness of fit test was used to see whether the observed numbers of recombination breakpoints differ significantly from a random distribution. The expected numbers of breakpoints for each window were assumed to follow a Poisson discrete distribution. The appropriateness of this distribution was tested by simulating the locations of breakpoints randomly (10,000 times) and for each performing the Chi-squared goodness of fit test. The number of times the Chi-squared value for the simulated data exceeded the value for the Poisson distributed was counted.

2.4. Similarity analysis

Genetic similarity across the alignment was calculated as the mean pairwise-intersubtype similarity in 200 and 500 nucleotide windows between nine subtype representative sequences for HIV-1 group M.

2.5. RNA secondary structure

RNAfold from the Vienna RNA package version 1.4 (Hofacker et al., 1994) was used to predict the RNA secondary structure from the HIV genome sequence. The predicted structure is represented in a string of dots and brackets, where ''.'' indicates the single strand and ''('' and '')'' are the 5′ and 3′ complementary base of double strands, which form the RNA secondary structure. The percentages of dots and brackets at each site are calculated along the genome, and used to obtain the mean and standard deviation of the percentages of ''('' and '')''. If the stem in the RNA structure exists, there will be lots of pairs of the ''('' and '')'' within a small region. If the RNA structure is conserved, the percentage of the ''('' or '')'' will be significantly greater than the mean value. Here, the threshold value is equal to the mean plus 1.96 times the standard deviation ($t_{0.95 \text{ with d.f. } >120} = 1.96$). A sliding window is used to calculate the distribution of the number of the significant brackets. The window size and the step size are set to 200 and 50, respectively.

Another 'local' approach for predicting RNA secondary structure (Forsdyke, 1995) was also implemented. A sliding window was incremented along the sequence with a window size of 200 nucleotides and step size of 50 nucleotides. For each window RNAfold was then used to estimate the entropy for the window. This entropy is determined by the local base