Identifying the important HIV-1 recombination breakpoints

John Archer[1], John W. Pinney[1], Jun Fan[1], Etienne Simon-Loriere[2], Eric J. Arts[3], Matteo Negroni[2] and David L. Robertson[1]

[1]Faculty of Life Sciences, University of Manchester, Manchester, UK; [2]Institut Pasteur, Paris, France; [3]Case Western Reserve University, Cleveland, Ohio, USA.

Contact:
Email: john.archer@postgrad.manchester.ac.uk.
Tel: +44 (0)161 275 1566
Post: Michael Smith Building, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

Abstract

Background: *Recombinant HIV-1 genomes contribute significantly to the diversity of variants within the HIV-1 pandemic. It is assumed that these mosaic genomes have novel properties that led to their prevalence. In regions of the HIV-1 genome where recombination conveys a selective advantage we predict that the distribution of breakpoints - the identifiable boundaries that delimit the mosaic structure - will deviate from the underlying distribution. To test this hypothesis we generate a probabilistic model of HIV-1 copy-choice recombination and compare the predicted breakpoint distribution to the real breakpoint distribution from the HIV/AIDS pandemic.*

Methods: *A "snapshot" of recombinant forms generated in the absence of selection has recently been obtained for the env gene by the analysis of 162 inter-subtype recombinants. We observed that significantly fewer breakpoints were located within five nucleotides or less of any mismatch between the aligned parental strains. We created a probabilistic model to describe the expected locations of breakpoints. Using this model, the pattern of HIV-1 inter-subtype breakpoints across the whole viral genome was predicted based on representative parental subtypes. The predicted distribution was compared to the distribution of known inter-subtype breakpoints.*

Results. *Across much of the HIV-1 genome we observe that the known inter-subtype breakpoint locations are predicted adequately. This observation strongly indicates that in these regions a mechanistic process is sufficient to explain breakpoint locations. In regions where there is a significant over- (either side of the env gene) or under- (short regions within gag, pol and most of env) representation of breakpoints natural selection must be influencing the pattern of recombination breakpoints. The paucity of recombination breakpoints within most of the envelope gene itself indicates the suppression of recombination.*

Conclusions. *We demonstrate the distribution of HIV-1 breakpoint locations across much of HIV's genome is adequately explained by a probabilistic model of the recombination process, based on local sequence identity. The exception is constituted by the envelope gene, identifying it as a region that has a tendency to be selectively transferred from one genetic background into another as an integral cassette. Our findings thus provide the first clear indication on how the majority of recombinant forms predominantly influence the viral population in the ongoing HIV/AIDS pandemics.*

Introduction

We have previously observed that recombination breakpoint locations amongst experimentally derived data are distributed across the HIV-1 envelope region of *gp120* non-randomly (Figure 1, Panel A) (1). Positioning of breakpoints is influenced by a number of factors including high sequence identity (1, 2), secondary RNA structure (3, 4) and the location of runs of identical nucleotides referred to as homopolymeric stretches (1, 5). The latter two are believed to increase the probability of a breakpoint occurring by stalling the reverse transcriptase complex during DNA synthesis, which in turn promotes the induction of strand-switching within regions of high sequence identity (Figure 1, Panel B).

Here we use a probabilistic model that takes into account local sequence identity in order to define an expectation for the distribution of breakpoints. Sequence identity is accounted for by not permitting breakpoints to occur directly on mismatches and by reducing the probability of a breakpoint occurring within a window of constant size anchored to the 5' of each mismatch (Figure 2).

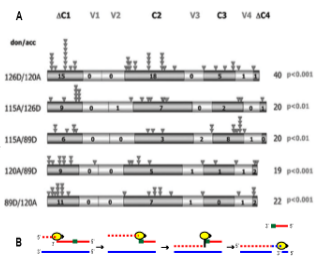

Figure 1. (A) Distribution of inter – subtype breakpoints along *gp120*. The constant regions are shaded dark grey. Parentals used are indicated on the left. Black numbers give the number of breakpoints identified within each region and grey triangles their approximate position. The total no. of recombinants analyzed for each pair is given on the right, together with the P-values for Chi-square tests. **(B)** Model of copy-choice recombination. The dotted red line represents the growing DNA strand. The full red line represents the donor RNA. Blue represents the acceptor RNA sequence. The green dot is some mechanistic feature on the donor strand that can stall the reverse transcriptase complex (yellow).

Methods

Laboratory Generated Recombinants. The experimental A/D inter- subtype recombinant sequences used are described in (1). The data from each parental pair was pooled and the frequency of breakpoints falling within breakpoint zones of particular sizes was calculated (Figure 2).

Breakpoints predictions based on full model. In the full model there are three different categories of site (Figure 2): (i) sites located on mismatches, (ii) sites located within windows and, (iii) sites not located within windows or on mismatches. At individual sites within these categories the probability of a breakpoint occurring is given by p_1, p_2 and p_3 respectively. Across the full alignment the sum of all p_1, p_2 and p_3 values is 1. Since breakpoints are not allowed on mismatches p1 is set to 0. The model can be summarized as follows:

$$(n-(m+w)).p_3 + w.\alpha.p_3 = 1 \quad \textit{where} \quad p_2 = \alpha.p_3$$

And w is the factor that the probability within windows is reduced by, n is the length of the alignment, m is the number of mismatches, w the number of nucleotides within windows and:

$$p_3 = \frac{1}{n-m-w(1-\alpha)} \quad \text{and} \quad p_2 = \frac{\alpha}{n-m-w(1-\alpha)}$$

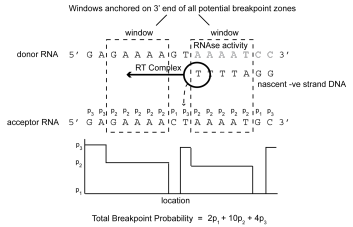

Figure 2. Diagrammatic model of HIV recombination. As the RT complex moves from the 3' end of the donor RNA to the 5' end it reverse transcribes the RNA sequence. The acceptor RNA strand is aligned alongside the donor strand and available for a potential crossover event – indicated by the dotted arrowed line. The dotted boxes indicate windows of decreased probability of crossover, that have been anchored on the mismatch at the 3' end of the potential breakpoint zone. The plot along the bottom is a representation of each of the probability values along the sequence.

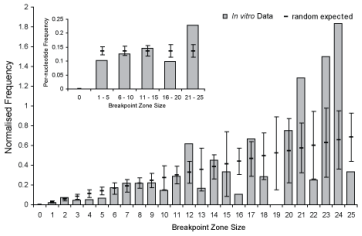

Figure 3. Normalized distribution of experimental breakpoints falling within zones ranging from size 0 to 25 (grey vertical bars). The horizontal bars indicate the expected random distribution of breakpoints within each individual zone. The associated error bars represent 1.645 standard deviations to include 90% of the distribution. The inset plot shows the per nucleotide frequency of both the experimental breakpoints and randomly generated breakpoints for zones up to size 25 (arranged in groups of 5).

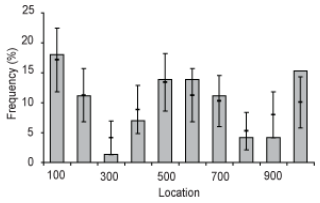

Figure 4. Recombination breakpoints across gp120. The experimental distributions are represented by the vertical grey bars. The horizontal bars indicate in the predicted distribution produced when using our model (Figure 2) and the associated error bars represent 1.645 standard deviations from the mean.

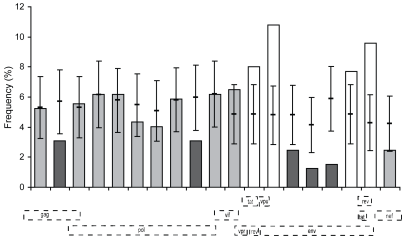

Figure 5. Predicted breakpoints locations using full length HIV-1 genomic sequences are displayed by the horizontal black bars. Error bars represent 1.645 standard deviations. The distribution of global breakpoints (6) is represented by the vertical grey bars. Dark grey indicates where the global data falls significantly above the predicted distribution, light grey indicates where the global data falls within the predicted distributions while white indicates where the global data falls below the predicted values. The frequency data has been divided up into window sizes of 400. Genomic regions are below the x-axis.

Results

The observed and expected distribution of breakpoints within different zone sizes is shown in figure 3. When the per nucleotide frequency of both the experimental breakpoints and the random expectations are organized into groups of size 5 (Figure 3 inset), a significant decrease is confirmed in zones of size 5 or less.

In figure 4, where the probability distributions produced by the model are displayed, 8 out of the 10 real breakpoint frequencies fall within the 1.654 standard deviations from the predicted values. There is no significant difference ($p>0.05$) between the expected frequencies and the experimental frequencies.

Discussion

Known breakpoint locations occurring within full length HIV-1 group M sequences that have been sampled from the group M pandemic (6) were compared to the distributions produced by the model. Across much of the HIV genome, it was observed that the model predicted distributions did not differ significantly from the global (*in vivo*) breakpoint distribution frequencies (Figure 5, light grey bars). That the frequency of recombination in some genomic regions does not depart significantly from the model predicted expectation indicates that a mechanistic process is often sufficient to explain breakpoint positioning.

The model and global distributions differ: (i) on either side of the envelope gene where a marked over-representation of breakpoints are present within the global data (Figure 5, white bars), and (ii) at regions in which breakpoint numbers are lower than expected, indicating a suppression of breakpoints within these regions (Figure 5, dark grey bars).

The over-representation of recombination breakpoints either side of env indicates a tendency for the shuttling of the entire envelope gene, or at least the coding region for extra-cellular *gp120*. Whether this is a result of coincidental or sequential recombination events, it indicates that selection is promoting *env*'s transfer from one genetic background into another effectively as an integral unit. This must be directly related to the envelope protein's functional significance in relation to viral fitness determinants, its propensity to be subject to high levels of positive selection, and the importance of the action of the immune response on HIV's envelope gene.

References

1. Baird HA, Galetto R, Gao Y, Simon-Loriere E, Abreha M, Archer J, Fan J, Robertson DL, Arts EJ, Negroni M (2006) *Nucleic Acids Res* 34(18):5203-5216.
2. Zhang J, Temin HM (1994) *J Virol* 68(4): 2409-2414.
3. Galetto R, Giacomoni V, Veron M, Negroni M (2006) *J Biol Chem* 281(5):2711-2720.
4. Moumen A, Polomack L, Roques B, Buc H, Negroni M (2001) *Nucleic Acids Res* 29(18):3814-3821.
5. Klarmann GJ, Schauber CA, Preston BD (1993) *J Biol Chem* 268(18):1376
6. Fan J, Negroni M, Robertson DL (2007) *Infect Genet Evol* 7(6): 717-23

Acknowledgements

JA is supported by BBSRC studentship, JP by a BBSRC project grant and JF by University of Manchester OSS award.