Identifying the Important HIV-1 Recombination Breakpoints

John Archer

Faculty of Life Sciences
University of Manchester
Overview

- An introduction to HIV-1.
- The diversity across the pandemic.
- How recombination contributes to this diversity.
- *In vitro* recombination in the absence of selection.
- Recombination within the pandemic.
A brief introduction to HIV-1

- HIV is an RNA virus with two identical copies of the genome.
- Viruses classified as HIV-1 fall into three groups.
- Group M is responsible for the HIV pandemic.
The virus is characterized by its:
- mutation rate \((10^{-5}/\text{site})\)
- recombination rate
- turnover \((10^9/\text{day})\)

- The result is nine divergent subtypes.
- Within a subtype divergence can be 20%.
The Democratic Republic of Congo

- All subtypes observed
- A large amount of inter subtype diversity
- Many unclassified forms.
- Geographic distribution of chimpanzees
The role of recombination

- Recombinant forms add to the complexity.
- These consist of mosaic genomes derived from different subtypes.
- Intra subtype recombinant forms are generated when a host is dually infected by strains form different subtypes.
- The generation process is largely dependent on sequence identity.
- However breakpoint persistence is dependent on selection.
- Unique Recombinant Forms (URF’s) are continually emerging.
- Circulating Recombinant Form’s (CRF’s) are recombinant forms with at least three complete genomes sampled from different individuals.
The persistence of recombinant forms

- Currently there are 43 CRF’s in the HIV LANL database.

- CRF_01 and CRF_02 alone make up 7% of the pandemic.

- The persistence of so many CRF’s suggests that recombination is playing an important role in distributing beneficial characteristics.
Recombination in the absence of selection

- Our collaborators took strains from subtypes A and D and, using a single cycle system, generated recombinant breakpoints across the envelope region.

- In total 162 breakpoints were generated across 7 pairs of parentals.

- There was no host selection influencing breakpoint persistence.

- Because the exact parentals were known sequence identity at the breakpoint location could be determined.
Breakpoints occurred at higher frequency within conserved regions.

Additionally potential breakpoint zones of size 5 or less had significantly fewer breakpoints than random.
To start with we wanted to replicate this distribution given the parentals.

Windows anchored on 3’ end of all potential breakpoint zones

Importantly windows are allowed to overlap so that regions with many mismatches will have a uniformly reduced rate of recombination.

Total Breakpoint Probability = 2p₁ + 10p₂ + 4p₃
The three different categories of sites are those:
- located on a mismatch ($p_1$)
- located within windows ($p_2$) and
- not located on a mismatch or within a window ($p_3$)

Across the full alignment the probability of a breakpoint is 1. Thus:

$$mp_1 + wp_2 + (n - (m + w))p_3 = 1,$$

but from our data $p_1$ is 0, additionally we define the ratio of $p_2$ to $p_3$ as:

$$\alpha = \frac{p_2}{p_3}$$

By rearrangement we get:

$$p_2 = \frac{\alpha}{n-m-w(1-\alpha)}$$ and $$p_3 = \frac{1}{n-m-w(1-\alpha)}.$$
Applying this model to the parental pairs used *in vitro*, and normalizing for the number of breakpoints observed for each pair, we could replicate the observed distribution.

But this was in the absence of selection.

We wanted to look at this distribution in relation to the pandemic.
Recombination in the presence of selection

- 324 global breakpoints had been previously classified using 80 inter-subtype recombinants.

- The number of breakpoints observed between each subtype pair had been recorded e.g.: AB, 2; AC, 50; AD, 83; etc.

- We applied our model to subtype representative parentals and weighted the probabilities according to numbers observed for a given pair.

- The location of breakpoints that are persistent within the pandemic could then be compared to our expected distribution.

- Areas where the two distributions deviate are those where selection is influencing breakpoint persistence.
Many of the observed breakpoint locations within the pandemic fall into the expected distribution.
However within the envelope gene as well as within small areas of \textit{gag} and \textit{pol} there was an under representation of breakpoints.

This is possible due to structural constraints being involved in folding proteins.
Additionally at either side of the envelope gene there was a significant over representation of breakpoints.

This indicates that genomes where the envelope region is being exchanged in one piece are being favorably selected for.
Conclusion

- Initial breakpoint generation is dependent on sequence identity.
- However breakpoint persistence is strongly influenced by selection.
- The selection pressure on breakpoint persistence across the genome is not uniform and in large portions of the genome the occurrence of breakpoints can be explained by simple sequence identity.
Acknowledgements

David Robertson, University of Manchester (Supervisor)

John Pinney, Imperial College London (Lab post doc)

Etienne Simon-Loriere and Matteo Negroni (Université Louis Pasteur de Strasbourg)

Kathryn Else, University of Manchester (Advisor)

Members of the Robertson Group

Organizers of Symposium

BBSRC for funding