Evolutionary analyses support that ASP (Anti Sense Protein) overlapping ORF could be the 10th gene of HIV-1 M pandemic group

JOBIM 2015

Elodie Cassan¹,²,³  Anne-Muriel Arigon Chifolleau¹,²
Antoine Gross³  Olivier Gascuel¹,²

¹LIRMM, UMR5506 CNRS, Université de Montpellier, France
²Computational Biology Institute, Montpellier, France
³CPBS, FRE3689 CNRS, Montpellier, France
HIV-1

- 2 types of Human Immunodeficiency Virus (HIV): **HIV-1** and HIV-2
- More than 35 million people infected
- Fast evolution + attack of immune cells→ no vaccine today

- Origin: inter-species transmissions of simian viruses (SIV)
- HIV-1: Group M, N, O et P
- Group M: pandemic group
**HIV-1 genome**

- **9 genes** (+ ASP gene ??)
- **Env gene region**: presence of a long **overlapping** and **anti-sens** Open Reading Frame (ORF)
Overlapping genes and evolutionary constraints

- Strategy to reduce the size of the viral genome
- Strong evolutionary constraints especially on frame -2
Overlapping genes and evolutionary constraints

- Strategy to reduce the size of the viral genome
- Strong evolutionary constraints especially on frame -2
Overlapping gene, and constraints

- Strategy to reduce the size of the viral genome
- Strong evolutionary constraints especially on frame -2

<table>
<thead>
<tr>
<th>Frame</th>
<th>+2/ +3</th>
<th>−1</th>
<th>−2</th>
<th>−3</th>
</tr>
</thead>
<tbody>
<tr>
<td># of amino acids</td>
<td>4</td>
<td>2.5</td>
<td>&lt; 2</td>
<td>10</td>
</tr>
</tbody>
</table>

**Table**: Average number of amino acids choice due to local constraints (considering +1 frame as fixed)
Presence of ASP ORF?

Group M

- Presence of ASP ORF
- Subtype A: ASP ORF is shorter
Presence of ASP ORF?

**Group M**

- Presence of ASP ORF
- Subtype A: ASP ORF is shorter
Presence of ASP ORF?

Group M

- Presence of ASP ORF
- Subtype A: ASP ORF is shorter
Presence of ASP ORF?

Group out-of-M:

- No ASP ORF
Emergence of ASP ORF?

- Phylogenetic tree
  - Group M: Los Alamos reference sequences
  - Groups out-of-M: selection of 10 sequences in group O + all others sequences
- Repartition of Start (red triangle) and Stop codons (black cross) in the ASP region.
- For each group or subtype: the frequency of Start codon and the average length of ASP ORF (obtained from our whole data by weighting the sequences so that each patient has total weight 1).
The presence of ASP ORF: emergence and selective pressure

Elodie CASSAN

ASP, the 10th gene of HIV-1 M pandemic group?
## Presence of ASP ORF

- **Emergence of ASP ORF**
- **Measuring the selective pressure**

### Summary

<table>
<thead>
<tr>
<th>Group / Subtype</th>
<th>Frequency Start</th>
<th>Mean Length</th>
<th>Prevalence (WHO 2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIVcpz Pts</td>
<td>0.77</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0.03</td>
<td>45</td>
<td>&lt;0.1 %</td>
</tr>
<tr>
<td>SIVgor</td>
<td>1.00</td>
<td>80</td>
<td>0 %</td>
</tr>
<tr>
<td>P</td>
<td>1.00</td>
<td>83</td>
<td>0 %</td>
</tr>
<tr>
<td>SIVcpz Ptt</td>
<td>0.16</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1.00</td>
<td>127</td>
<td>0 %</td>
</tr>
<tr>
<td>SIVcpz Ptt</td>
<td>1.00</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1.00</td>
<td>172</td>
<td>4.6 %</td>
</tr>
<tr>
<td>J</td>
<td>1.00</td>
<td>174</td>
<td>0.1 %</td>
</tr>
<tr>
<td>A</td>
<td>0.99</td>
<td>149</td>
<td>12 %</td>
</tr>
<tr>
<td>H</td>
<td>0.25</td>
<td>77</td>
<td>0.1 %</td>
</tr>
<tr>
<td>F</td>
<td>0.88</td>
<td>138</td>
<td>0.4 %</td>
</tr>
<tr>
<td>K</td>
<td>1.00</td>
<td>144</td>
<td>0.02 %</td>
</tr>
<tr>
<td>C</td>
<td>0.96</td>
<td>162</td>
<td>48.2 %</td>
</tr>
<tr>
<td>D</td>
<td>0.93</td>
<td>145</td>
<td>2.49 %</td>
</tr>
<tr>
<td>B</td>
<td>0.97</td>
<td>173</td>
<td>11.3 %</td>
</tr>
</tbody>
</table>

---

**ASP, the 10th gene of HIV-1 M pandemic group?**

Elodie CASSAN
Introduction

Presence of ASP ORF

Emergence of ASP ORF

Measuring the selective pressure

Summary

Elodie CASSAN

ASP, the 10th gene of HIV-1 M pandemic group?

<table>
<thead>
<tr>
<th>Group / Subtype</th>
<th>Frequency Start</th>
<th>Mean Length</th>
<th>Prevalence (WHO 2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIVcpz Pts</td>
<td>0.77</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0.03</td>
<td>45</td>
<td>&lt;0.1 %</td>
</tr>
<tr>
<td>SIVgor</td>
<td>1.00</td>
<td>80</td>
<td>0 %</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>83</td>
<td>0 %</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>136</td>
<td></td>
</tr>
<tr>
<td>SIVcpz Ptt</td>
<td>1.00</td>
<td>128</td>
<td>0 %</td>
</tr>
</tbody>
</table>

One sequence with ASP ORF

<table>
<thead>
<tr>
<th>Group</th>
<th>Frequency Start</th>
<th>Mean Length</th>
<th>Prevalence (WHO 2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1.00</td>
<td>172</td>
<td>4.6 %</td>
</tr>
<tr>
<td>J</td>
<td>1.00</td>
<td>174</td>
<td>0.1 %</td>
</tr>
<tr>
<td>A</td>
<td>0.99</td>
<td>149</td>
<td>12 %</td>
</tr>
<tr>
<td>H</td>
<td>0.25</td>
<td>77</td>
<td>0.1 %</td>
</tr>
<tr>
<td>F</td>
<td>0.88</td>
<td>138</td>
<td>0.4 %</td>
</tr>
<tr>
<td>K</td>
<td>1.00</td>
<td>144</td>
<td>0.02 %</td>
</tr>
<tr>
<td>C</td>
<td>0.96</td>
<td>162</td>
<td>48.2 %</td>
</tr>
<tr>
<td>D</td>
<td>0.93</td>
<td>145</td>
<td>2.49 %</td>
</tr>
<tr>
<td>B</td>
<td>0.97</td>
<td>173</td>
<td>11.3 %</td>
</tr>
</tbody>
</table>
Elodie CASSAN
ASP, the 10th gene of HIV-1 M pandemic group?
Measuring the selective pressure

Selection pressure?

- Most constrained frame
- Unique solution for 55% of ASP sites (HXB2)
- Mir & Shober 2014, selection pressure in frame +1 and -2 are \( \sim \) undistinguishable.
- Detection of selection pressure induced by ASP protein is difficult with conventional methods.

\[
\begin{align*}
&Ala \\
+1 &\quad -\quad -\quad GCx &\quad -\quad \rightarrow \\
&\leftarrow -xCG &\quad -\quad -\quad -\quad -2
\end{align*}
\]

Ala
Start codon may appear/disappear

- Ile or Met on env $\implies$ Ile or Met on ASP
- xyC $\Leftrightarrow$ xyT synonymous for every xy

\[
\text{Start} \quad \text{Ile} \\
\text{(GA)TA} \quad -2 \text{ ASP}
\]
Start codon may appear/disappear

- Ile or Met on $env \Rightarrow$ Ile or Met on ASP
- $xyC \Leftrightarrow xyT$ synonymous for every $xy$
Selection pressure to conserve the Start codon in group M

**Group M**

- **Strong conservation**

**Groups out-of-M**

- **As expected ~50/50**
What about Stop codons?

Some Stop codons are potential and can synonymously appear/disappear.
Potential Stop Codons

- Sites with appearance of Stop codons **without modification of env**. Blue = potential Stops, Red = existing Stops
- Presence of Stop codons in out-of-M groups
- Statistical test

**Figure**: Potential Stops (blue) and existing Stops (red) on ASP
Potential Stop Codons

- Sites with appearance of Stop codons **without modification of env**. Blue = potential Stops, Red = existing Stops
- Presence of Stop codons in out-of-M groups
- Statistical test

**Figure**: Potential Stops (blue) and existing Stops (red) on ASP
Potential Stop Codons

- Sites with appearance of Stop codons **without modification of env**. Blue = potential Stops, Red = existing Stops
- Presence of Stop codons in out-of-M groups
- Statistical test

**Figure**: Potential Stops (blue) and existing Stops (red) on ASP
**Principle**

- Pairwise sequences analysis (Inspired by the method of Firth, Nucleic Acids Research, 2014)

![Diagram](image)

- Count synonymous mutations $\rightarrow$ Stop codon
- Comparison with theoretical count based on an evolutionary model
- Z test

$$Z = \frac{Theo - Obs}{2 \times \text{variance}}$$
**Group M results**

- **ASP region**:
  - $Z = 7.12$, p-value $= 10^{-12}$ (with RRE) ✓
  - $Z = 3.6$, p-value $= 2.10^{-4}$ (without RRE) ✓

- **env-ASP region**:
  - $Z = -1.85$, p-value $= 0.063$ ✗
### Group M results

- **ASP region:**
  - $Z = 7.12$, p-value $= 10^{-12}$ (with RRE) ✓
  - $Z = 3.6$, p-value $= 2.10^{-4}$ (without RRE) ✓

- **env-ASP region:**
  - $Z = -1.85$, p-value $= 0.063$ ×

### Group out-of-M results

- **ASP region:**
  - $Z = 1.46$, p-value $= 0.14$ (with RRE) ×
  - $Z = -0.1$, p-value $= 0.92$ (without RRE) ×

- **env-ASP region:**
  - $Z = -0.43$, p-value $= 0.66$ ×
Summary

- Existence of the ASP ORF in pandemic subtypes (C, B ...) of Human Immunodeficiency Virus 1 M.
- Absence of the ASP ORF in other non-pandemic human and simian groups.
- Phylogenetic analyses indicate that ASP ORF appeared recently.
- ASP protein is subjected to significant selection pressure.
- **ASP ORF could be the 10th gene of HIV-1 M and seems to be correlated with human pandemy.**
Merci de votre attention