

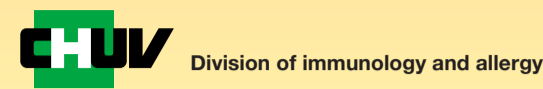
NON-RANDOM DISTRIBUTION OF CRYPTIC REPEATING TRIPLETS OF PURINES AND PYRIMIDINES (RNY)_N AND RECOMBINATION IN GP120 OF HIV-1

E. De Crignis¹, S. Guglietta¹, B. T. Foley², M. Negroni³, G. Pantaleo¹, C. Graziosi¹

¹Laboratory of AIDS Immunopathogenesis, Division of Immunology and Allergy, Department of Medicine, CHUV, Lausanne

²Theoretical Biology and Biophysics Group, Los Alamos National Laboratories, Los Alamos, NM

³Architecture et Reactivite de l'ARN, Université de Strasbourg, CNRS, IBMC, Strasbourg



Introduction

Infection by HIV is an extremely dynamic process with high viral turnover rates. During HIV-1 replication cycle a large number of mutations, including transversions and transitions as well as insertions and deletions, can occur. An important source of these mutations is the reverse transcription process. Two features of retrotranscription facilitate mutation: the low proofreading ability of the enzyme and the multiple strand transfer steps necessary to create the dsDNA from the viral genome. Strand transfer promotes the occurrence of recombination and this process is able to boost sequence variability by creating new combinations of mutated and non mutated sequences.

We previously demonstrated that events of length polymorphism involving indels spanning multiples of three bases occur in the V4 region of gp120. In this work we have investigated intra-host length polymorphism in gp120 in patients with early HIV-1 infection. Indels were found all across the molecule in regions enriched in trinucleotides repetitions, suggesting that they may arise as a consequence of misalignment processes. Finally we found several evidences of intra patient recombination.

Materials and Methods

Seven patients naive to therapy in early infection, harboring subtypes B, C and CRF02AG were included in the study. An Env RNA fragment spanning C1-C5 was cloned and sequenced. Sequence analysis was performed using Sequencher and Bioedit together with the tools available at the HIV Sequence database. Recombination analysis was performed using Splitstree package and by the program RDP3Beta30. We evaluated the distribution of codons inside indels and along constant and variable regions. Differences in RNY distribution between constant and variable regions were evaluated by comparing the observed results with 1000 randomly generated sequences. The codon usage was calculated using Bioedit, significancy in codon usage between different regions were evaluated with a T test.

Results

Intra-host sequences of gp120 are characterized by the occurrence of major indels in V1, V2, V4 and V5 and to a lesser extent in C3: major indels affecting at various extent the distribution of potential N-glycosylation sites of the protein can be observed in all gp120 variable regions (Fig. 1).

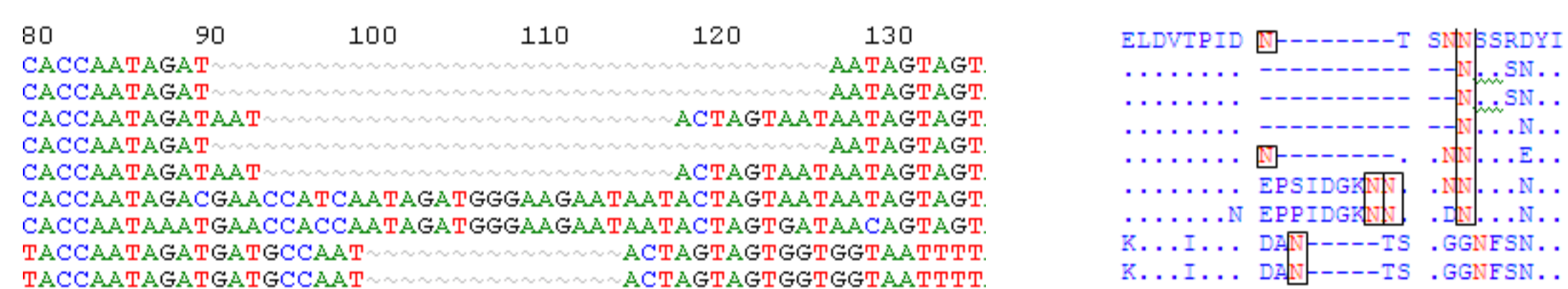


Fig.1: Indels (left panel) and glycosylation site variation (right panel) in patient 1

Variable regions of gp120 are enriched in RNY trinucleotides: Patterns of alternating purines (R) and pyrimidines (Y) are often linked to mutation potential of the sequences. The frequency of stretches of RNY (spanning from four to seventeen repeats) is significantly higher in variable regions than in constant regions (Fig. 2).

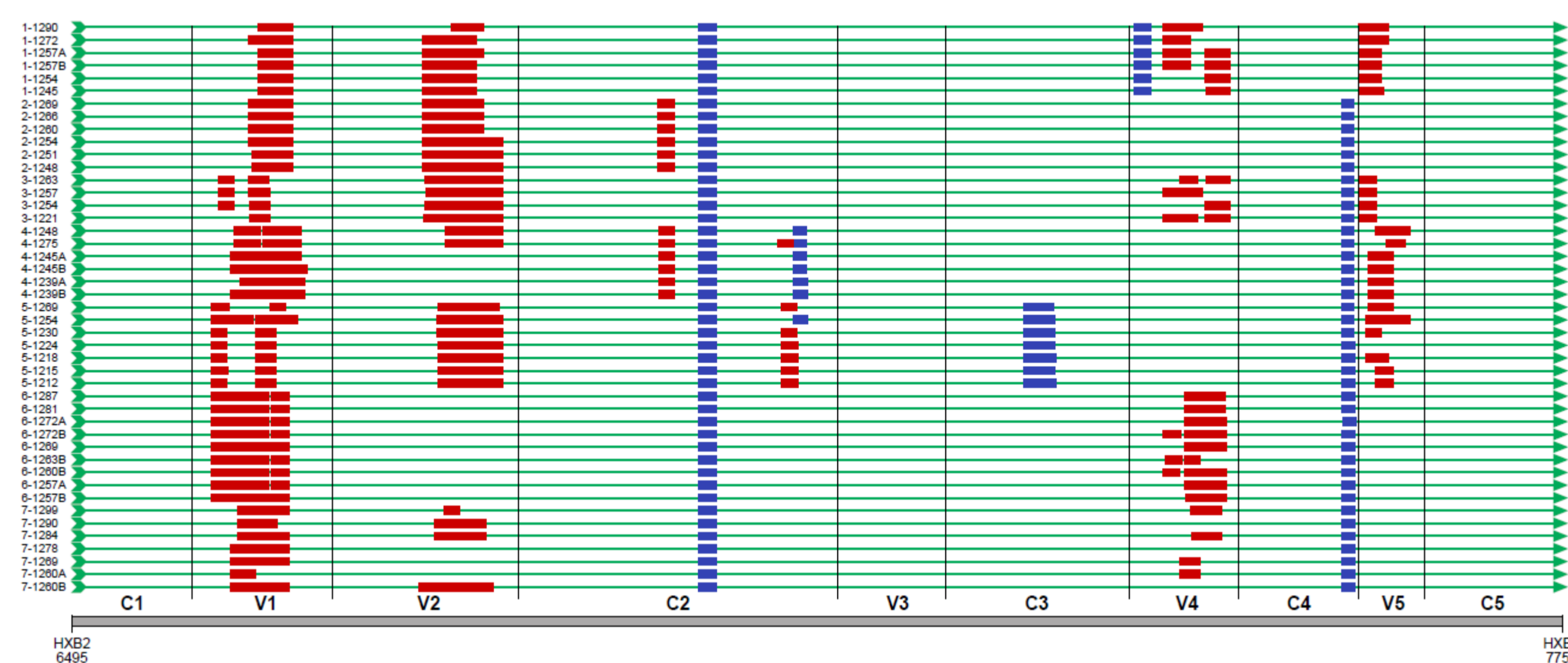
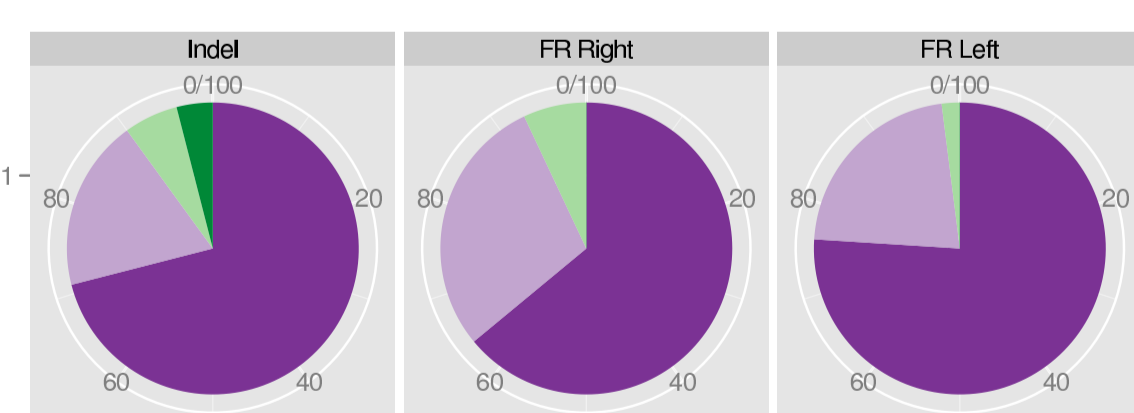


Fig.2: Stretches of contiguous in-frame RNY codons (red bars) along gp120



Sequences comprised within indels consist mainly of RNY codons and the indel events are usually flanked by RNY codons as well (Fig. 3).

Fig.3: Codon composition of indels and indels flanking regions (FR)

A linear relationship between length of the indels and number of RNY codons can be identified ($R^2=0.90$). The association is weaker when only asparagine is considered ($R^2=0.55$) and absent when considering other codons (Fig. 4).

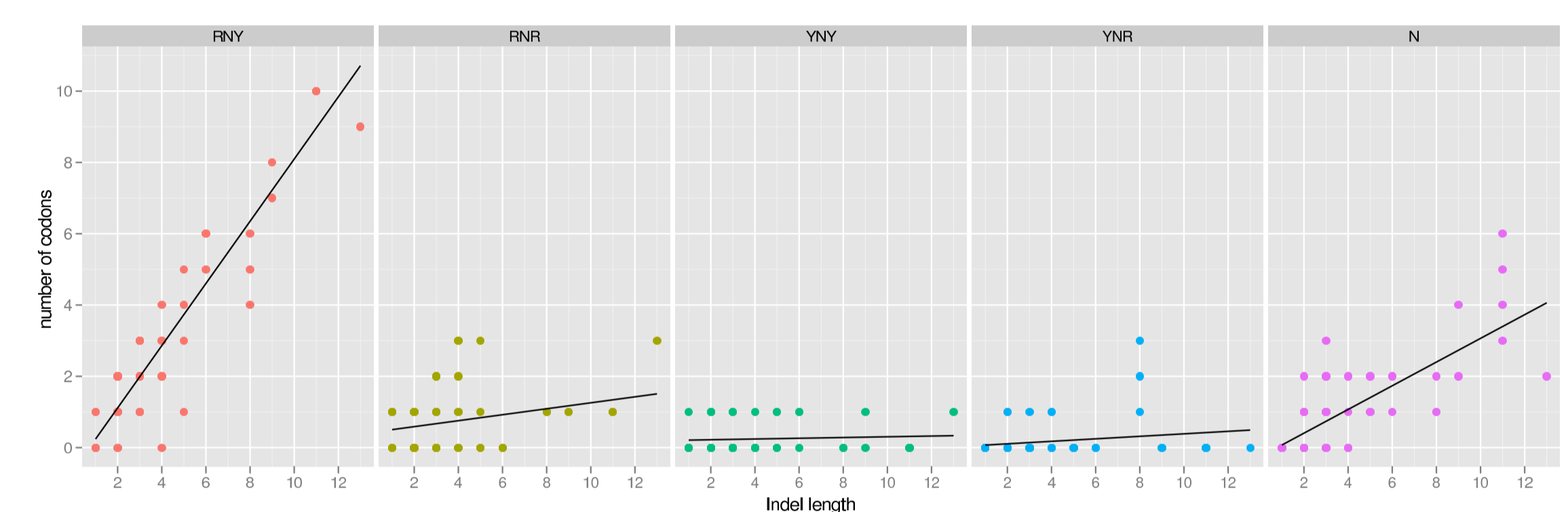


Fig.4: Linear relationship between length of the indels and number of codons. The last panel (N) shows the relationship between length of the indels and the number of asparagine residues.

Codon usage is markedly different among gp120 regions: Aminoacids that can be encoded either by RNY or not-RNY codons are mainly encoded by RNY codons inside indels (Fig. 5).

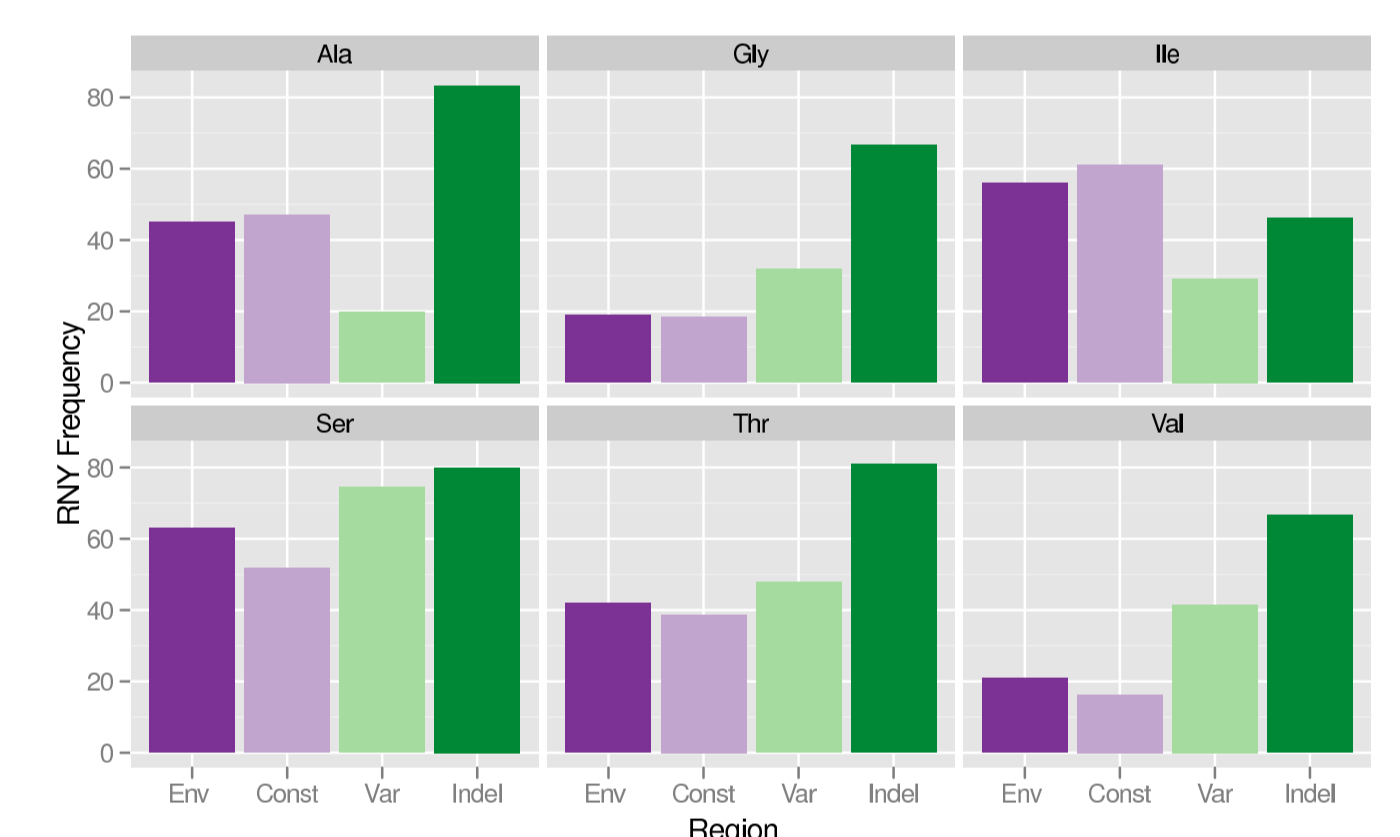


Fig.5: Frequency of residues encoded by RNY codons in Env, Constant Regions, Variable Regions and inside indels.

RNY distribution differs between HIV-1 genes Analysis of an alignment comprising 127 sequences from the HIV database reveals that stretches of 4 or more RNY are mainly located in Env. Moreover, even considering variable and constant regions together, the frequency of aminoacids encoded by RNY is significantly ($p < 0.01$, Kruskal-Wallis test) different among HIV regions with a higher prevalence of RNY codons in Env (Fig.6).

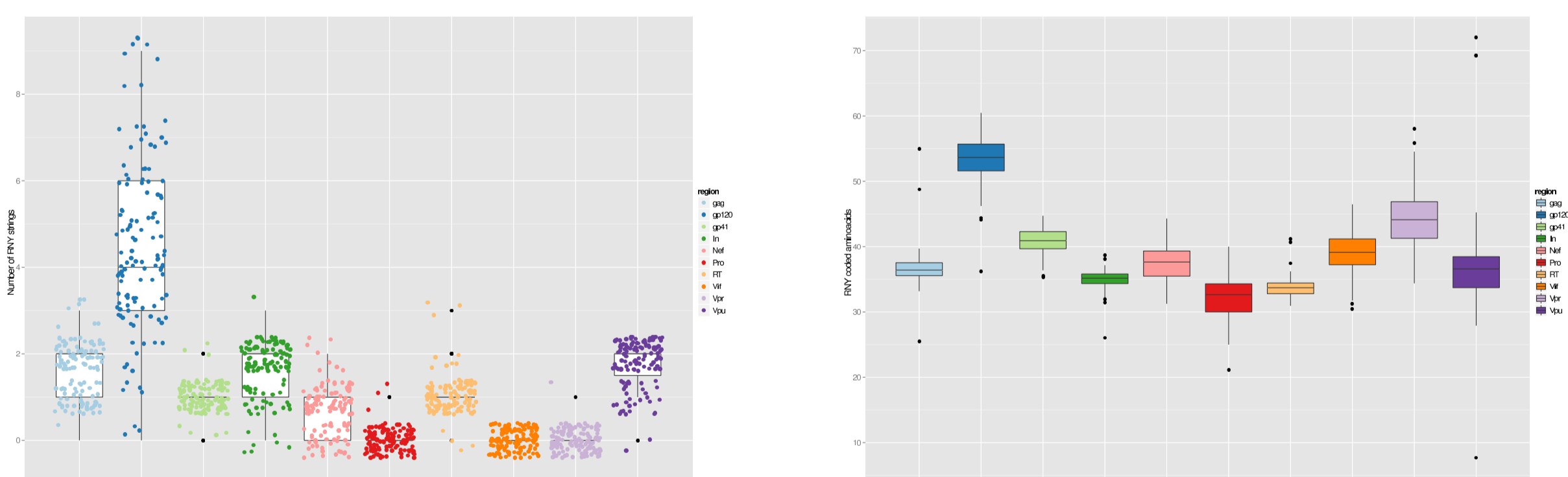


Fig.6: Number of RNY stretches (4 or more codons) along HIV genome (left panel). Percentage of RNY coded aminoacids in different HIV genes (right panel).

Recombination occurs among individual clones in the same patient PHI-test and RDP analysis were significant for all the patients analyzed. Analysis of clone alignments by RDP identified with certainty twelve breakpoints.

Acknowledgements: The authors wish to acknowledge the statistical support and advice of Dr. Antonio Di Narzo, Swiss Institute for Bioinformatics, University of Lausanne.

Contacts: Elisa De Crignis, Department of Medicine, Division of Allergy and Immunology, CHUV BT02-36,1011 Lausanne-Switzerland. E-mail: elisa.decrignis@gmail.com

Conclusions

Variable regions are responsible for length variation polymorphisms in gp120 and are characterized by the presence of an unusual distribution of RNY with respect to the rest of the molecule. Interestingly, all the codons coding for the components of glycosylation sites (N, T, S) can be encoded by RNY codons. If RNY codons are able to promote indels by means of duplication this could lead to a change of glycosylation patterns that can be advantageous for generating new variants able to escape antibody pressures. Even if less evident, intra-host recombination is highly likely and can be observed in the sequences we analyzed. All these mechanisms contribute to the endless generation of an array of ever changing gp120 that are simultaneously produced and selected for or against depending on the virus-host interaction at any given time during infection.