Divergent patterns of V3 loop evolution in M-group subtypes.

Introduction

Individual isolates or clones of HIV-1 from within any one subtype exhibit differences in cytopathic effect and cell tropism. Although subtype B is the most thoroughly studied, syncytium-inducing (SI) and non-syncytium inducing (NSI) isolates have been described for other subtypes as well. Despite this intraclade diversity of phenotypes, to date there have been few indications of intersubtype differences in pathogenicity or transmissibility [Kanki (1999)]. This may be due in part to a lack of controlled studies which could identify such differences. Subtype B virus infecting North Americans and Europeans cannot be directly compared to subtype A virus infecting central Africans, because of differences in factors such as diet, access to health care, and exposure to secondary pathogens. It is difficult to do longitudinal cohort studies in the developing world, where several subtypes are prevalent in the same population. The wide range of phenotypic variability within a single subtype, as well as ranges of human responses to infection, requires a large sample size for study. Finally, long-term studies including sero-incident cases are critical for detecting differences in pathogenicity, and this type of study requires many years to complete. This analysis and discussion is focussed on identifying differences in evolutionary trends of the envelope protein V3 region between the HIV-1 M group subtypes. We hope that this may help to identify subtypes of HIV-1 which are more likely to show differences in phenotype.

The evolution of HIV-1 is influenced not only by mutations in the viral genome and selective pressure by individual human hosts, but also by the overall temporal and geographical patterns of the epidemic. The rapid spread of subtype B into large serologically naive populations in the urban centers of North America and Europe resulted in a star phylogeny for subtype B, suggestive of large amounts of horizontal transfer of the virus in a relatively short peroid of time. Conversely, the evolution of subtype A has been influenced by less geographical spread, remaining largely in the sub-Saharan region of the African continent, within more rural populations, where dual infection with other subtypes of HIV-1 (and even HIV-2) has ocurred. Thus the subtype A phylogenetic tree is more "bushy" and intersubtype recombinants containing subtype A are more common than those containing subtype B.

Differences in modes of transmission might lead to different evolutionary trends for HIV-1. Transmission via contaminated needles may put different selection pressures on the virus than sexual transmission for example. Thus, the analysis of HIV-1 sequence variability should be accompanied by a consideration of the natural history of the various lineages of HIV-1, as well as an understanding of the biased sampling of HIV-1 isolates selected for sequence analysis.

A quick perusal of the tabulation of V3 loop amino acid sequences divided by subtype illustrates the variability of evolutionary patterns between subtypes. For example, while subtype C shows near perfect conservation of the GPGQ amino acid sequence at the tip of the V3 loop, subtype D has retained the proline in less than 65% of sequences sampled. An elevated rate of nonsynonymous substitutions in the third variable (V3) loop of subtype D viruses has previously been noted [Korber (1994b)]. Serological studies also indicated that the V3 loop of the D subtype was more diverse than other subtypes [Barin (1996)]. The V3 loop is a functionally and immunogenically important domain of the viral envelope. Positively charged amino acids in certain positions in the V3 loop correlate with a syncytium-inducing phenotype in culture [de Jong (1992), De Wolf (1994)], and usage of the CXCR4 second receptor [Cocchi (1996), Bieniasz (1997)].

We examined differences in patterns of envelope V3 region variability using the V3 data set from this 1999 HIV sequence compendium, with many more sequences available than when we previously completed this study in 1994 and 1997 [Korber (1994b)], [Dighe (1997)]. The data was explored using the following method: V3 region amino acid sequences were split into two regions: the V3 loop, typically including 35 amino acids between the cysteine residues which form a disulfide bond to produce the loop; and the 35 amino acids flanking the loop, including 20 amino acids from the region immediately adjacent to the NH2-terminal side of the loop (NFTDNARVIIVQLNESVEIN for example) and 15 amino acids from the region immediately adjacent to the COOH-side of the loop (NLSSTKWNNTLKQIT for example). Each of these two regions (loop or flanks) was aligned and

scored for pairwise similarity using the PIMA software package [Smith & Smith(1992)], with the protein similarity scoring scheme based on amino acid substitution matrices from protein blocks [Henikoff & Henikoff(1992), Henikoff & Henikoff(1993)]. The sequences in this set are compiled such that only a single sequence from each individual patient, or from each group of epidemiologically linked patients, is represented, thus eliminating the bias of multiple sequences from the same patient. A notable exception to the "one sequence per related group" rule are found in the IV drug user data sets (subtypes A, B and AB recombinant) from the former Soviet Union and nearby countries, where nearly identical HIV sequences are found in people with no direct epidemiological linkage. It has been speculated that the drug preparation itself was the source of HIV in at least some of these cases [Bobkov (1998a)].

The results from these analyses are shown in Figures 1 and 2. The median of the similarities found in the V3 loop of subtype D is much lower than in the other subtypes, an indication of much greater diversity, while the interquartile range (a measure of the variation around the median) is larger. The diversity in the flanking regions of subtype D, in contrast to the V3 loop, is very similar to that in subtypes B and C. This result suggests that the loop of subtype D is unique relative to the other subtypes, and is either under greater positive selective pressure for change, or alternatively, free from evolutionary constraints (negative selection presure) which restrict the variation of other subtypes. This effect is highly statistically significant, and is associated with the charge on the V3 loop (as discussed in the 1997 sequence compendium) [Dighe (1997)]. For example, using the non-parametric Wilcoxon or the Kolmogorov-Smirnov tests to compare the distributions of subtype A and D protein similarity scores for the V3 loop, gave p-values of $\ll 0.00001$. (To do the statistical test, the average value of each sequence compared to all others of that subtype was used to represent that sequence, so each sequence was only counted once for the statistical comparison, although all pairwise comparisons were made to construct the figures.) The A versus D comparison is of particular interest as the samples tend to come from the same studies, conducted in the same geographic regions.

The distribution of the similarities in CRF01(AE) V3 loops, originating generally either from Thailand or the Central African Republic, is different from the other subtypes, in that the distribution in these V3 loop similarity scores seems to be bimodal (Figure 1), reflecting the presence of two subpopulations of sequences. A highly conserved set of protein domains are apparent for the Thai sequences, but it is interesting to note that some of the phylogenetically similar Thai sequences are in the more divergent group of V3 loop protein sequences ([Yu (1995)], and the data underlying Figure 1), suggesting a potential for rapid divergence in this domain in the CRF01(AE) form of HIV-1.

It is possible that the observed differences in V3 variation between the subtypes could have resulted from systematic sampling artifacts from the different subtypes. However, this seems unlikely in light of the large number of individuals now sampled, and the efforts of the UNAIDS and DAIDS to collect international samples from recent seroconvertors. The data suggest that, in addition to the phylogenetic distinctions, there are distinctions between the subtypes resulting from different evolutionary pressures. Thus biological differences do exist, either in the history of the epidemic of the different subtypes, or in different selective pressures. Subtypes B and D appear to have shared a common ancestor, after some period of divergence from the other subtypes. Although phenotypic distinctions have not been clearly detected in the data accrued so far, potential differences are indicated by differences in genetic makeup and rate of evolution, at least within the functionally important V3 loop of the different subtypes. Despite these interesting distinctions, it will be important in the future not to focus the search for biological differences in viruses too narrowly on intersubtype differences. It has already been found that important phenotypic distinctions are found between different lineages within one subtype (such as SI and NSI isolates) and too much emphasis on the subtype distinctions could conceivably result in myopia with respect to potential intrasubtype differences.
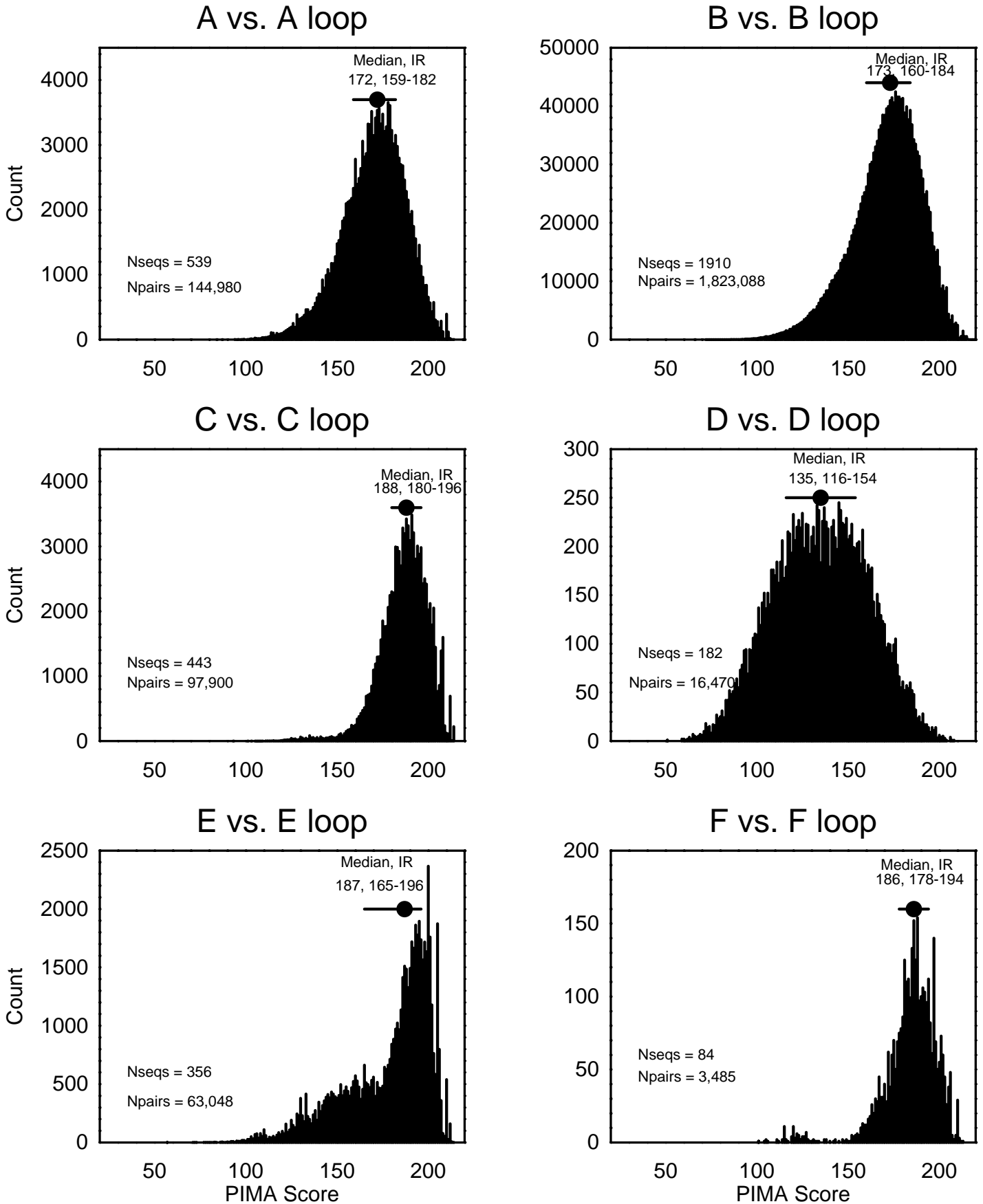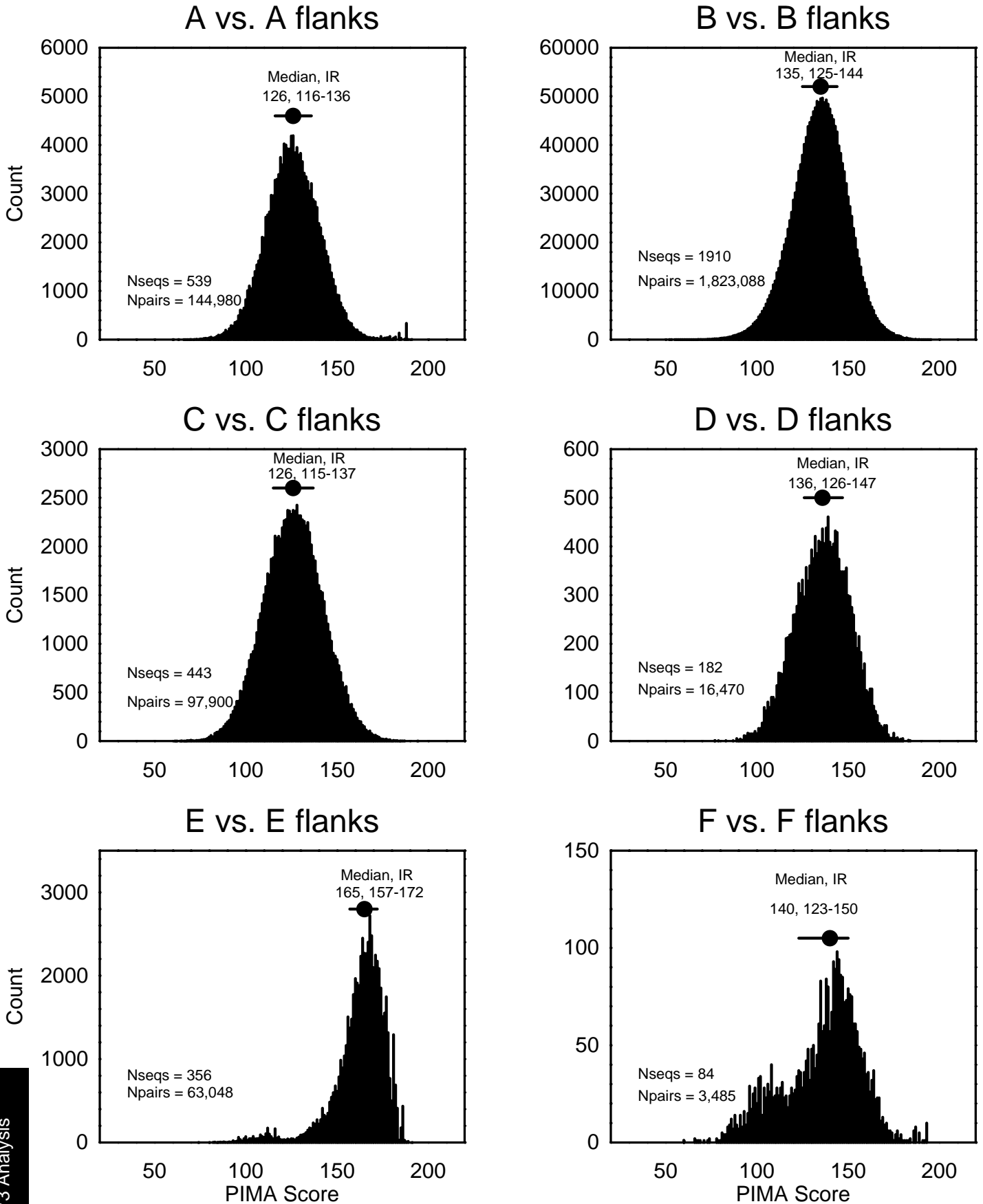
V3 Analysis

Figure 1

A vs. A flanks

Median, IR
126, 116-136

Nseqs = 539
Npairs = 144,980

Count

B vs. B flanks

Median, IR
135, 125-144

Nseqs = 1910
Npairs = 1,823,088

C vs. C flanks

Median, IR
126, 115-137

Nseqs = 443
Npairs = 97,900

Count

D vs. D flanks

Median, IR
136, 126-147

Nseqs = 182
Npairs = 16,470

E vs. E flanks

Median, IR
165, 157-172

Nseqs = 356
Npairs = 63,048

Count

PIMA Score

F vs. F flanks

Median, IR
140, 123-150

Nseqs = 84
Npairs = 3,485

PIMA Score

V3 Analysis

Figure 2