Global variation in the HIV-1 V3 region

Brian Gaschen, Bette T. Korber and Brian T. Foley

MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545

Introduction

Due to the immunogenicity and functional importance of the V3 loop, there has been a great deal of interest in the V3 region of the envelope protein, resulting in a large international effort to obtain V3 region sequences. This section, which includes sequences taken from more than 3,826 individuals, with complete references, provides an overview of the variation of sequences that span this region.

Sequences

To best summarize the spectrum of international HIV-1 variants, only one representative viral sequence was included per infected individual. A complete set of references accompanies the sequence alignments, and nomenclature was preserved from the original papers so individuals and isolates can be identified, when possible. HIV-1 was deleted from the sequence names in this section, as all sequences included here are HIV-1. Included with the references, when available, are brief descriptions of critical features of the sequences. This includes the health status of the individual from whom the virus was derived, whether or not the virus was cultured, and the year the blood sample was taken, if known.

All sequences are prefaced by a subtype association (see phylogenetic clustering below) and a two letter country code to identify the country in which the individual resided at the time that the blood sample was taken. If the person was a recent immigrant and this information was available, we included the country of origin in the references. The key to the country codes follows this introduction. Note that this key was updated for 1999, with several country codes for eastern european nations added since 1995.

Sometimes only one viral sequence was available from a person: a clone from an isolate, or a direct sequence of PCR amplified peripheral blood DNA. For other individuals, more than 100 viral sequences from PCR amplified DNA or RNA from blood samples were available. Consequently, over 13,000 sequences are represented by the 3,826 included in this section. When two or more sequences were available from a person, one of them was randomly selected. If a set of sequences from two or more individuals was epidemiologically linked, and genetically very similar, only one sequence from the set was included, preferably the most recently infected. One of the exceptions is found in sequences from the Kaliningrad and Ukranian IV drug user epidemic, which are included because these patients seem to be representative of an emerging new epidemic. Although these IV drug user sequences are very similar to each other, there is no evidence that the patients shared needles with each other; it appears that the virus was present in the drug preparation itself. In the sequence description and references section, the short hand “PCR-direct, peripheral blood DNA” is used to signify that viral DNA was amplified from PBMCs, without culturing, and a single “direct” sequence was obtained from the amplification reaction products. In a handful of cases, a particular gp160 clone from an isolate was shown to be expressed and functional using a vaccinia virus T7 expression system.

Phylogenetic clustering

Sequences have been organized according to the phylogenetic subtype association of their envelope V3 regions only. The original sequence subtype designations were defined based on the phylogenetic relationships determined by using both gag and env genes (when possible), are approximately genetically equidistant in envelope, except for subtypes B and D which are more closely related to each other, than to other subtypes, and have multiple members. The phylogenetic subtype designations and associations have generally been adopted by the HIV research community, and are now often presented with the publication of new sequences. We have either determined the subtype designations here, if

not specified in the original manuscript, or else confirmed the subtype designations of the original manuscripts, and then used the subtypes to organize this section. Generally, confirmations were done by aligning a set of HIV-1 V3 region sequences with longer env gene sequences (Part IIIC) that have clear subtype associations, and then using neighbor-joining trees to determine associations. Neighbor-joining trees were generated with F84 model distances (called maximum likelihood in dnadist) and a transition to transversion ratio of 1.3 using PHYLIP (Joseph Felsenstein, University of Washington). All available nucleotide sequence information was used for phylogenetic analysis; longer protein sequences were trimmed to be approximately the same length as the majority of the PCR fragments in this region, for the purposes of presentation. Some sequences were difficult to classify, and are included in the "U", or unclassified, section. The subtype A/G circulating recombinant form common in Nigeria, and exemplified by the IbNG isolate with accession number L39106, is subtype A in the env V3 region, but subtype G in parts of the pol gene. Many isolates from western Africa have only been sequenced in the env gene, and not in the pol regions needed to prove that the sequence is subtype G there. These sequences are sometimes classified as subtype A, but perhaps should be more correctly classified as CRF02(AG) recombinants.

The set of reference sequences used to help resolve subtype associations included at least two sequences from each subtype (A-D, F-H, J and K) or circulating recombinant form (CRF01-CRF04), plus two HIV-1 O group sequences as outgroup sequences. The reference sequences were selected based on being "typical" of the subtype they represent as determined by phylogenetic analysis. The set has changed over the last few years as more sequences have accumulated. Thus not all subtype designations were based on the same reference set.

Limitations of phylogenetic analyses

Most of the PCR derived sequences are rather short for accurate phylogenetic analyses, given the level of variability in this region – typically on the order of 250 to 300 nucleotides. Due to this limitation, some of the classifications in this section are uncertain and are our best estimate given the available information. Control studies were performed to compare the phylogenetic clustering of the V3 region using available longer sequences however, and these studies indicate that our subtype designations based on the V3 region are generally reliable. For many sequences, we had an approximately 700 base region of env available representing all of the subtypes. After removing positions in the alignment which included gaps, 519 bases were left. When a 298 base V3 region fragment was excised from this set, and neighbor joining trees were constructed using both the 519 base and 298 base long sequences, the phylogenetic subtype designations were consistent in each case. Further, when a subset of longer gp120 sequences was analyzed (92 of the 146), including 935 bases after removing positions in the alignment which included gaps, the subtype designations were again clear in neighbor-joining trees. This indicates that the limited V3 region PCR fragments, which include more than the V3-loop but not complete gp120, are generally able to serve as a reliable basis for subtype determination, given the limitation that the V3 loop sequence may be embedded in a recombinant genome.

Without detailed analysis, genetic recombination between subtypes may obscure phylogenetic relationships between sequences. A characteristic of recombination is an indeterminate place in phylogenetic analyses, and some of the "unclassified" category sequences may prove to be recombinant genomes upon further inspection. Also, while a subtype designation based on a gene or gene fragment may be correct, recombination events outside the region examined may have occurred. Therefore, care should be taken to not overinterpret the subtype designations. If one is to discuss the subtype designations of viral isolates based on the data presented here, they should refer to the designation as "B-like over the V3 loop region," rather than as "subtype B". This caveat is especially pertinant to sequences of subtype A and the CRF02(AG) circulating recombinant form which is subtype A-like in the V3 region. This year an attempt was made to re-classify some of the sequences formerly listed as subtype A, as CRF02(AG), based on the presence of signature sites in the V3 region which were thought to be reliable indicators of the subtype A vs. CRF02(AG) lineages.

Limitations of V3 amino acid consensus sequences

The V3 amino acid consensus sequences generated for each subtype have interesting features; however, one should be wary about assuming that any of the consensus sequences may broadly represent its subtype. Certainly many V3 loop variants in each of the subtypes, particulaly subtype D, are extremely divergent from the consensus sequences. These divergent forms may have very different biological and immunological characteristics from viruses which are similar to the consensus. Additionally, because of the relatively small sample size of some of the subtypes, notably H, J and K, consensus sequences can be dominated by a small group of highly similar sequences, which may in turn be a sampling artifact. Hence, these consensus sequences are "evolving" as new sequences from each subtype become available.