Research

Molecular characterization of the HIV-1 *gag* nucleocapsid gene associated with vertical transmission

Brian P Wellensiek, Vasudha Sundaravaradan, Rajesh Ramakrishnan and Nafees Ahmad*

Address: Department of Microbiology and Immunology, College of Medicine, The University of Arizona Health Sciences Center, Tucson, Arizona, USA

Email: Brian P Wellensiek - bwellen1@u.arizona.edu; Vasudha Sundaravaradan - vasudha@u.arizona.edu; Rajesh Ramakrishnan - ramakris@bcm.tmc.edu; Nafees Ahmad* - nafees@u.arizona.edu

* Corresponding author

Abstract

Background: The human immunodeficiency virus type 1 (HIV-1) nucleocapsid (NC) plays a pivotal role in the viral lifecycle: including encapsulating the viral genome, aiding in strand transfer during reverse transcription, and packaging two copies of the viral genome into progeny virions. Another *gag* gene product, p6, plays an integral role in successful viral budding from the plasma membrane and inclusion of the accessory protein Vpr within newly budding virions. In this study, we have characterized the *gag* NC and p6 genes from six mother-infant pairs following vertical transmission by performing phylogenetic analysis and by analyzing the degree of genetic diversity, evolutionary dynamics, and conservation of functional domains.

Results: Phylogenetic analysis of 168 *gag* NC and p6 genes sequences revealed six separate subtrees that corresponded to each mother-infant pair, suggesting that epidemiologically linked individuals were closer to each other than epidemiologically unlinked individuals. A high frequency (92.8%) of intact open reading frames of NC and p6 with patient and pair specific sequence motifs were conserved in mother-infant pairs' sequences. Nucleotide and amino acid distances showed a lower degree of viral heterogeneity, and a low degree of estimates of genetic diversity was also found in NC and p6 sequences. The NC and p6 sequences from both mothers and infants were found to be under positive selection pressure. The two important functional motifs within NC, the zinc-finger motifs, were highly conserved in most of the sequences, as were the *gag* p6 Vpr binding, AIP1 and late binding domains. Several CTL recognition epitopes identified within the NC and p6 genes were found to be mostly conserved in 6 mother-infant pairs' sequences.

Conclusion: These data suggest that the *gag* NC and p6 open reading frames and functional domains were conserved in mother-infant pairs' sequences following vertical transmission, which confirms the critical role of these gene products in the viral lifecycle.

Background

Mother-to-infant (vertical) transmission of HIV-1 occurs at a rate of 30%, and accounts for 90% of infections in children worldwide. Transmission of the virus can occur

at three stages: prepartum (in utero), intrapartum (during birth), and postpartum (breast feeding). Several factors have been linked to vertical transmission including: low CD4 count and high viral load of the mother, advanced maternal disease status, invasive procedures, infections during pregnancy and prolonged exposure of the infant to blood and ruptured membranes during birth [1-8]. The exact molecular mechanisms of vertical transmission are not well understood, however we and others have shown that the minor HIV-1 genotypes are transmitted from mother to infant [9,10]. It has also been shown that the macrophage-tropic (R5) phenotype is involved in transmission [11]. Analysis of several HIV-1 accessory and regulatory genes, including *vif, vpr, vpu, nef, tat and rev* has revealed conservation of functional domains of these genes during vertical transmission [12-17]. In addition, transmitting mothers' *vif and vpr* sequences were more heterogeneous and the functional domain more conserved than non-transmitting mothers' sequences [12-17]. However, other HIV-1 genes may also play a crucial role in virus transmission and pathogenesis.

One such gene product, the *gag* nucleocapsid (NC) plays a pivotal role in the viral lifecycle, including encapsulating the viral genome, aiding in the reverse transcription process, protecting the viral genome from nuclease digestion and packaging two copies of the viral genome into progeny virions [18-23]. The NC gene product, also termed p7, is translated as a Pr55 Gag precursor and when cleaved is 55 amino acids long. It contains one major functional domain, consisting of two zinc finger like motifs. These motifs allow the NC to bind the packaging signal, or Ψ site, on viral RNA, as well as coat the viral genome [18,24,25]. They contain the sequence $C\text{-}X_2\text{-}C\text{-}X_4\text{-}H\text{-}X_4\text{-}C$ with the critical residues consisting of three cystines and one histidine [20]. When these critical zinc binding amino acids are mutated to non-zinc binding residues, it results in virions that are defective in RNA packaging and replication [18,21,26]. Several basic amino acid residues throughout the NC gene product are also associated with RNA binding, and aid in NC function [18,21]. These basic residues are responsible for interaction with the side chains of the viral nucleic acids. NC plays several roles during the reverse transcription step of the HIV-1 lifecycle. It is responsible for ensuring proper annealing of the tRNA[Lys] primer to the primer binding site to initiate reverse transcription, and also aids in strand transfer so that reverse transcription can continue [20,21,23,27,28]. During and after reverse transcription, it has been shown that NC binds to the newly generated viral DNA and protects it from cellular nucleases until it can integrate into the host cell genome [22,29]. Due to the importance of this gene any alterations to the NC may affect transmission and pathogenesis of the virus.

Another example of a crucial gene product is p6, which plays an integral role in successful viral budding from the plasma membrane and inclusion of the accessory protein Vpr within newly budding virions [30-35]. The p6 gene product is also initially translated as a Pr55 Gag precursor and is 52aa long when cleaved by the viral protease. The p6 protein contains a viral late (L) domain with the sequence PTAPP, which is necessary for viral budding [36,37]. It has been shown that the late domain interacts with the host cell factor Tsg101 which is involved in regulating intracellular trafficking [32,35,38,39]. The late domain has also been shown to be crucial for detachment of virions from the host cell surface. Defects and mutations in the late domain can result in chains of immature virions that cannot release from the host cell surface [36,40]. The p6 gene product also contains a region with the sequence DKELYPLASLRSLFG that is responsible for interacting with the host cell factor AIP1 [31,41,42]. AIP1 has been shown to interact with Tsg101 and host factor ESCRT-III to function in a late-acting endosomal sorting complex that is essential for viral budding [31,41-43]. There are two domains that could possibly be required for inclusion of Vpr, either the FRFG domain [30] or the $(LXX)_4$ domain [33,34,44,45]. Defects within the Vpr binding domains could result in virions that lack Vpr. This would affect the ability of the virus, upon infection, to replicate in nondividing cells such as macrophages, and would affect the ability of the viral DNA to localize to the host cell nucleus for integration. The p6 gene product is also critical in the viral lifecycle, and therefore any changes within it may effect the transmission and pathogenesis of the virus.

In this study, we have characterized and analyzed the genetic diversity and population dynamics of the *gag* NC and p6 genes from six mother-infant pairs following vertical transmission. Our findings suggest that these gene products are mostly conserved during mother-infant transmission. Furthermore, the critical functional domains were conserved in most sequences analyzed. These results help to further our understanding of the molecular mechanisms that are involved in vertical transmission of HIV-1.

Results
Phylogenetic analysis of NC and p6 sequences from mother-infant pairs
Multiple independent polymerase chain reactions (PCRs) were performed on peripheral blood mononuclear cell (PBMC) DNA from six mother-infant pairs, a total of 13 patients including one mother who gave birth to HIV-1 positive twins. Eight to eighteen clones from each patient were obtained and sequenced. The phylogenetic analysis was performed using a neighbor-joining tree of the 168 NC and p6 sequences from the mother-infant pairs (Fig.
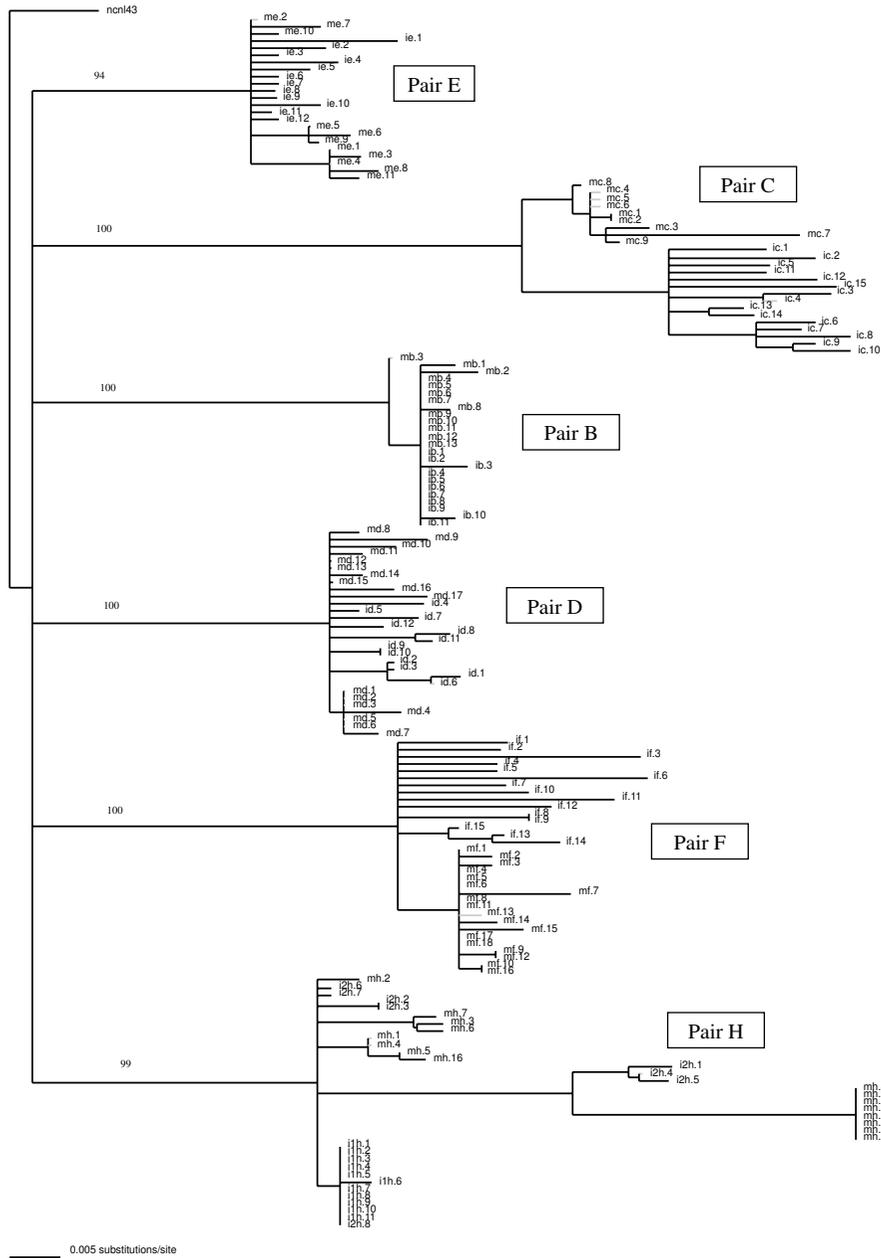
Figure 1
Phylogenetic analysis of 168 HIV-1 NC and p6 sequences from six mother-infant pairs; pairs B, C, D, E, F, and H. The neighbor-joining tree is based on the distance calculated between the nucleotide sequences from the six mother-infant pairs. Each terminal node represents one sequence. The values on the branches represent the occurrence of that branch over 1,000 bootstrap resamplings. Each pair formed a distinct subtree, and within each subtree the mother and infant sequences were generally separated into clusters, although some intermingling was observed. The formation of subtrees indicated that epidemiologically linked mother-infant pairs were closer to each other evolutionarily than to epidemiologically unlinked pairs, and that there was no PCR cross-contamination. The placement of the HIV-1 lab control strain NL4-3 indicates that no PCR contamination occurred.

Figure 2
Multiple sequence alignment of deduced amino acids of NC and p6 from mother-infant pairs B and C. Within the alignment, the top sequence is the NC consensus B (ConBNC) sequence to which the mother-infant pair sequences are compared. Each line of the alignment represents one clone sequence, and is identified by a clone number with M referring to mother and I referring to infants. The dots represent agreement with the consensus sequence, while substitutions are represented by a single letter amino acid code. Stop codons are shown as asterisks (*). The functional domains within the sequence are indicated above the alignment.

1). This neighbor-joining tree was generated by incorporating a best-fit model of evolution into PAUP [46], and the resulting tree was then bootstrapped 1000 times to ensure fidelity. Analysis of the tree demonstrated that the sequences from the six mother-infant pairs form distinct, well separated subtrees, and all pairs were separate from the lab control strain HIV-1 isolate NL4-3. Within each subtree the sequences for the mother and infant are generally well separated into subtrees, however some intermingling was observed in pairs B, D, E, and H. The intermingling of mother-infant sequences suggests that the isolates from these patients are very closely related, and had not as of yet evolved to form separate, distinct subtrees. Taken together the data indicates that epidemiologically linked (mother-infant) patient sequences are closer to each other evolutionarily than epidemiologically unlinked sequences. The separation of the mother-infant

sequences from each pair and NL4-3 indicates that no PCR contamination occurred.

Coding potential of NC and p6 gene sequences
The multiple sequence alignment of the deduced amino acid sequences of the HIV-1 NC and p6 genes is shown in Figs. 2, 3, 4. Of the 168 sequences analyzed, 156 contained an intact open reading frame (ORF), yielding a frequency of 92.8%. This high frequency indicates that the coding potential of the NC and p6 genes was maintained in most of the sequences analyzed. Looking more closely, the frequency of an intact ORF for the mothers' sequences was 89.4%, while the infants' sequences yielded a frequency of 96.3%. Several clones within mother-infant pair H were found to be defective due to a single nucleotide substitution, insertion or deletion, which resulted in the formation of a stop codon. There were several patient

```
                  NUCLEOCAPSID GENE PRODUCT (p7)                                              p6 GENE PRODUCT
        ◄──────────────────────────────────────────►                    ◄──────────────────────────────────────────────────────►
              Zinc finger #1        Zinc finger #2                  Late Domain       Vpr binding domains
            ◄──────────►          ◄──────────►                       ↔          ↔                  ◄──────────►
        1                                                                                                                      133
ConBNC  MMQRGNFRNQ RKTVKCFNCG KEGHIAKNCR APRKKGCWKC GKEGHQMKDC TERQANFLGK IWPSHKGRPG NFLQSRPE.. .PTAPPE... ...ESFRFGE ETTTPSQKQE PIDKELYPLA SLRSLFGNDP SSQ
 MD-1   .......K.. .......... ......R... .......... .......... .......... .......... .......... .......... .......... ...R.M...T .......... ..T
 MD-2   .......K.. .......... ......R... .......... .......... .......... .......... .......... .......... .......... ...R.M...T .......... ..T
 MD-3   .......K.. .......... ......R... .......... .......... .......... .......... .......... .......... .......... ...R.M...T .......... ..T
 MD-4   .......K.. .......... ....VR... .......... .......... .......... .......... .......... ..E....D .......... ...R.M...T .......... ..T
 MD-5   .......K.. .......... ......R... .......... .......... .......... .......... .......... .......... .......... ...R.M...T .......... ..T
 MD-6   .......K.. .......... ......R... .......... .......... .......... .......... .......... .......... .......... ...R.M...T .......... ..T
 MD-7   .......K.. .......... ......R... .......... .......... .......... .......... .......... .......... .......... ...W.M...T .......... ..T
 MD-8   .......K.. .......... ......R... .......... .......... .......... ...N...... PTA PPE.....E. .......... ...R.M...T .......... ..T
 MD-9   .......K.. .......... ....VR... ..G....... .......... .......... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 MD-10  .......K.. .......... ......R... .......... .......... .....S.... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 MD-11  .......K.. .......... ......R... .......... .......... .......... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 MD-12  .......K.. .......... ......R... .......... .......... .......... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 MD-13  .......K.. .......... ......R... .......... .......... .......... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 MD-14  .......K.. .......... ......R... .......... .......... .......... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 MD-15  .......K.. .......... ......R... .......... .......... .......... ...N...... PTV PPE...... .......... ...R.M...T .......... ..T
 MD-16  .......K.. .......... ......R... .......... .......... .......... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 MD-17  .......K.. .......... ....VR... ....R..... ......R... .......... ...N...... PTA PPE...... .......... ...R.M...T .......... ..T
 ID-1   .......K.. .......... ......R... .......... .......... .......... ...N...... .....V..... .......... ...R.M...T ...K...... ..T
 ID-2   .......K.. .......... ......R... .......... .......... .......... ...N...... .....V..... .......... ...R.M...T ..K...... ..T
 ID-3   .......K.. .......... ......R... .......... .......... .......... ...N...... .....V..... .......... ...R.M...T ...K...... ..T
 ID-4   ....D..K.. .......... ......R... .......... .......... .......... ...N...... .....L..... .......... ...R.M...T .......... ..T
 ID-5   .......K.. .......... ......R... .......... .......... .......... ...N...... .....V..... .......... ...R.M...T .......... ..T
 ID-6   .......K.. .......... ......R... .......... .......... .......... ...N...... .....V..... .......... ...R.M...T .......... ..T
 ID-7   .......K.. .......... ......R... ....R..... .......... .......... ...N...... .....V..... .......... .K.R.M...T .......... ..T
 ID-8   .......K.. .......... ......R... .......... .......... .......... ...N...... ..T...V..... .......... ...R.M...T A ...... ..T
 ID-9   .......K.. .......... ......R... .......... .......... .......... ...N...... .....L..... .......... ...R.M...T A ...... ..T
 ID-10  .......K.. .......... ......R... .......... .......... .......... ...N...... .....L..... .......... ...R.M...T A ...... ..T
 ID-11  .......K.. .......... ......R... .......... .......... .......... ...N...... .....V..... .......... ...R.M...T .......... ..T
 ID-12  .......K.. .......... ......R... .......... .......... .......... ...N...... .....L..... .......... .K.R.M...T .......... ..T

 ME-1   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .S........ .......... .....P.... .T...V.... ...
 ME-2   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .......... .......... .......... .....V.... ...
 ME-3   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .S........ .......... .....P.... .T...V.... ...
 ME-4   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .S........ .......... .....P.... .T...V.... ...
 ME-5   ...K...... KRN....... ......R... .......... ........E. .......... ...N...... .......... .......... .......... .....V.... ...
 ME-6   ...K...... KRN....... ......R... .......... ........E. .......... ...N...... .S........ .......... .......... .....V.... ...
 ME-7   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .......... .......... .....LT...V.... ...
 ME-8   ...K...... ..N....... ......R... ....R..... ........E. I......... ...N...... .S........ .......... .....P.... .T...V.... ...
 ME-9   ...K...... KRN....... ......R... .......... ........E. .......... ...N...... .......... .......... S....V.... ...
 ME-10  ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .......... .M........ .......... .....V.... ...
 ME-11  ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .S........ .G... .....P.... .T...V.... ...
 IE-1   ...K...... ..N....... .K...R... .......... ........E. .......... ..F.N..... .......... .......... .K..V.... ...
 IE-2   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .....L..... .......... .K..V.... ...
 IE-3   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .....L..... .......... .....V.... ...
 IE-4   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .....L..... .......... .K........ .P.
 IE-5   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... ......I.... .......... .....V.... ...
 IE-6   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .....L..... .......... .....V.... ...
 IE-7   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .....L..... .......... .....V.... ...
 IE-8   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .......... .......... .....V.... ...
 IE-9   ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .......... .......... .....V.... ...
 IE-10  ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .......... .......... .....P.... ..N..V.... ...
 IE-11  ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .......... .......... .K..V.... ...
 IE-12  ...K...... ..N....... ......R... .......... ........E. .......... ...N...... .....L..... .......... .....V.... ...
```
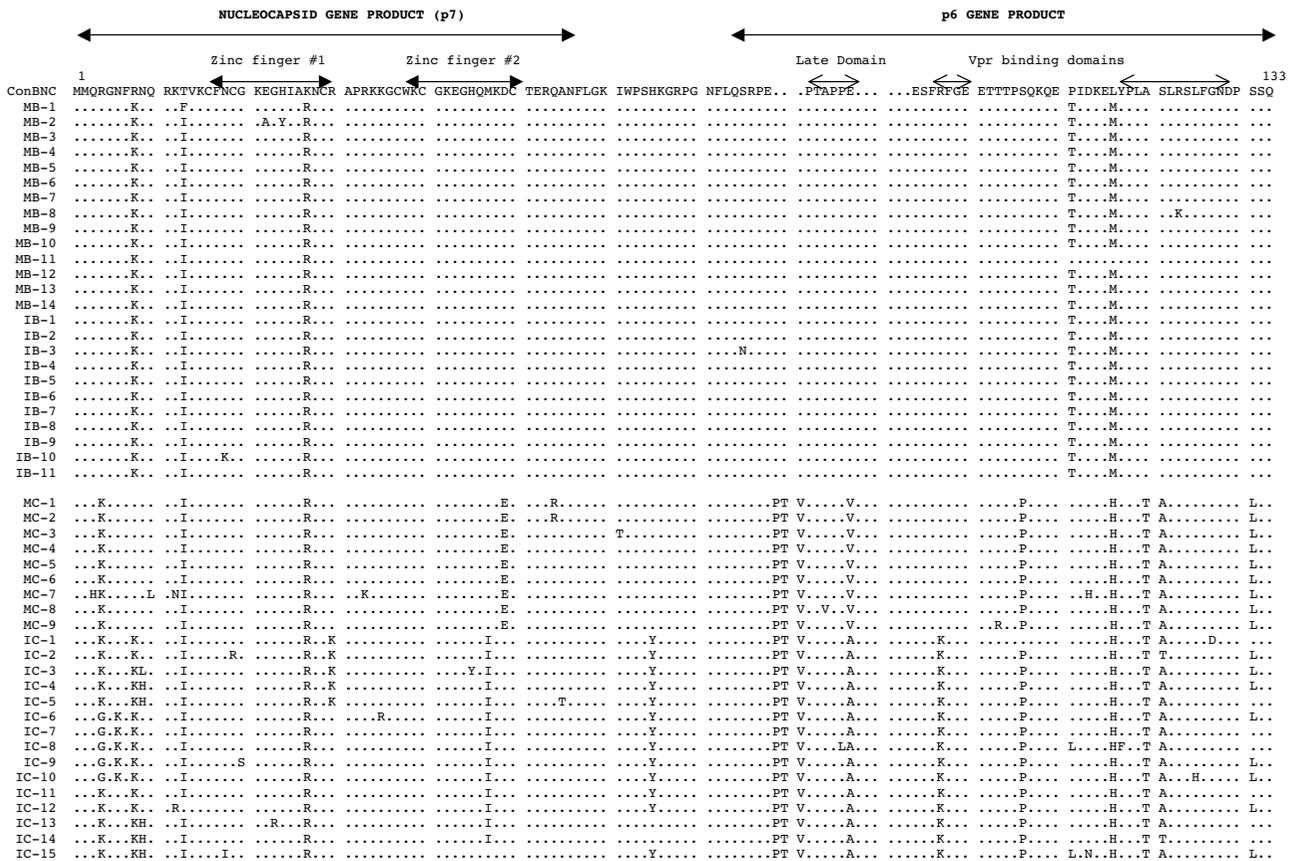
Figure 3
Multiple sequence alignment of deduced amino acids of NC and p6 from mother-infant pairs D and E. Within the alignment, the top sequence is the NC consensus B (ConBNC) sequence to which the mother-infant pair sequences are compared. Each line of the alignment represents one clone sequence, and is identified by a clone number with M referring to mother and I referring to infants. The dots represent agreement with the consensus sequence, while substitutions are represented by a single letter amino acid code. Stop codons are shown as asterisks (*). The functional domains within the sequence are indicated above the alignment.

and pair specific sequence patterns within the NC sequences analyzed. An insertion of proline-threonine-valine (PTV) was seen in the sequences of mother-infant pair C at position 78, and an insertion of proline-threonine-alanine-proline-proline-glutamate (PTAPPE) was observed within several sequences of mother D at position 84. This resulted in a duplication of the PTAP motif within this patient. An amino acid substitution was also present in most of the sequences when compared as a whole, a leucine (L) was replaced with a methionine (M), valine (V), histidine (H), arginine (R) or glutamine (Q) at position 116.

Variability of NC and p6 gene sequences in mother-infant pairs
The nucleotide and amino acid distances, which measure the degree of genetic variability based on pairwise comparison, were calculated for the six mother-infant pairs' sequences (Table 2). The nucleotide sequences within mothers B, C, D, E, F, and H varied by 0.26, 0.53, 0.84, 1.13, 0.27, and 5.04% (median values) respectively, ranging from 0 to 6.30%. The infant (B, C, D, E, F, I1H, and I2H) sequences differed by 0, 2.59, 0.88, 1.11, 1.78, 0, 3.22% (median values) respectively, ranging from 0 to 5.03%. Moreover, the nucleotide sequence variability

```
                    NUCLEOCAPSID GENE PRODUCT (p7)                                    p6 GENE PRODUCT
       <───────────────────────────────────────────>              <──────────────────────────────────────────────>

              Zinc finger #1        Zinc finger #2                    Late Domain        Vpr binding domains
       1                <───────>        <───────>                      <─>         <─>          <──────>         133
ConBNC MMQRGNFRNQ RKTVKCFNCG KEGHIAKNCR APRKKGCWKC GKEGHQMKDC TERQANFLGK IWPSHKGRPG NFLQSRPE.. .PTAPPE... ...ESFRFGE ETTTPSQKQE PIDKELYPLA SLRSLFGNDP SSQ
  MF-1  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
  MF-2  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ..I....... .......N..C... .......... A....M.... .......... ...
  MF-3  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... ...P...... ...
  MF-4  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
  MF-5  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
  MF-6  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
  MF-7  ...K...G.K G.I....... RV.T..... T......... .......... .......... ....Y..... .......... ...N..C... .......... A....V.... .......... ...
  MF-8  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
  MF-9  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .........T ...
 MF-10  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
 MF-11  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
 MF-12  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .........T ...
 MF-13  ...K...G.K G.I....... .V........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
 MF-14  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... ....I.N... ...
 MF-15  ...K...G.K G.I....... RV........ .......... ..R....... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
 MF-16  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
 MF-17  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
 MF-18  ...K...G.K G.I....... RV........ .......... .......... .......... ....Y..... .......... ...N..C... .......... A....M.... .......... ...
  IF-1  ......G.K G.I....... RV........ .......... .......... .......... .......... .......... ...N..C... .......... T....M...T .......... ...
  IF-2  ......G.K G.I....... RV....I.. .......... .......... .......... .......... .......... ...N..C... .......... T....M.... .......... ...
  IF-3  .S.....G.K G.I....... RA........ .......... .......... .......... ......Q... .......... ...N..C... .......... T....M.... .......... ...
  IF-4  ......G.K G.I....... RA........ .......... .......... .......... .......... .......... ...N..C.G .......... T....M.... .......... ...
  IF-5  ......G.K G.I....... RA........ .......... .......... .......... .......... .......... ...N..C... .......... T....M.... .......... ...
  IF-6  ......G.K G.IF..... RA........ .......... .......... .......... .......... .......... ...N..C... .......... T....M.... ....DN... .A.
  IF-7  ....K.G.K G.I....... RA........ .......... .......... .......... .......... .......... ...N..... .......... T....M.... .......... ...
  IF-8  ......G.K G.I....... RV........ .......... .......... .......... .......... .......... ...N..C... .......... T....M.... .......... ...
  IF-9  ......G.K G.I....... RV........ .......... .......... .......... .......... .......... ...N..C... .......... T....M.... .......... ...
 IF-10  ......G.K G.I....... RA........ .......... .......... .......... .......... .......... ...N..C... .......... T....M.... .......... ...
 IF-11  ...K...G.K G.I....... RA........ .......... .......... .......R.. .......... .......... ...N..CR.. .......... T....M.... ..I.....H. ...
 IF-12  ......G.K G.I....... RV........ .......... .K........ .......... .......... .......... ...N..C... .......... T....M.... .......... ...
 IF-13  ......G.K G.I....... RV........ .......... .......... .......... .......... .......... ...N..C... .......... A....V.... .......... ...
 IF-14  ......G.K G.I...D RA........ .......... .......... .......... .......... .......... ...N..C... .......... A....V.... .......... ...
 IF-15  ......G.K G.I....... RV........ .......... .......... .......... .......... .......... ...N..C... .......... A....V.... .......... ...

  MH-1  .......... .......... .......... .K........ .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
  MH-2  .......... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
  MH-3  .......... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... QT...R.... .P........ L..
  MH-4  .......... .......... .......... .K........ .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
  MH-5  .......... .......... .......... .K........ .......... I........R .......... .......... .......... ..A....*. S....R.... .......... L..
  MH-6  .......... .......... .Y... .......... .......... I........R .......... .......... .......... ..A....... QT...R.... .......... L..
  MH-7  .......... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... QT...R...P .......... L..
  MH-8  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
  MH-9  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
 MH-10  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
 MH-11  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
 MH-12  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
 MH-13  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
 MH-14  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
 MH-15  ......K.. .......S ..R....... .......... *.. K..R.... IK......R. ....R.... .......... ....K. ..A....... S....Q.... .......... L..
 MH-16  .N........ .......... .K........ .......... I........R .......... .......... .......... ..A....... S....R.... .........*. L..
 I1H-1  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-2  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-3  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-4  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-5  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-6  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-7  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-8  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I1H-9  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
I1H-10  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
I1H-11  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I2H-1  ...K...... .A....... ..R....... .......... *.. R..... I........R. .......R .......... ....K. ..A....... S....R.... .......... L..
 I2H-2  .......... .......... .......... .......... .......... I........R .R.... .......... .......... .......... S....R.... .......... L..
 I2H-3  .......... .......... .......... .......... .......... I........R .R.... .......... .......... ..A....... S....R.... .......... L..
 I2H-4  .......... .R....... .......... .......... *.. R..... I........R. .R.... ....R .......... ....K. ..A....... S....R.... .......... L..
 I2H-5  .......... .......... ..R....... .......... *.. R..... I...T...R. ....R .......... ....R .......... ..A....... S....R.... .......... L..
 I2H-6  .......... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I2H-7  .......... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
 I2H-8  ...K...... .......... .......... .......... .......... I........R .......... .......... .......... ..A....... S....R.... .......... L..
```

Figure 4
Multiple sequence alignment of deduced amino acids of NC and p6 from mother-infant pairs F and H, including both infant H twins (I1H and I2H). Within the alignment, the top sequence is the NC consensus B (ConBNC) sequence to which the mother-infant pair sequences are compared. Each line of the alignment represents one clone sequence, and is identified by a clone number with M referring to mother and I referring to infants. The dots represent agreement with the consensus sequence, while substitutions are represented by a single letter amino acid code. Stop codons are shown as asterisks (*). The functional domains within the sequence are indicated above the alignment.

between epidemiologically linked mother-infant pairs (pairs B, C, D, E, F, and H) varied by 0, 3.16, 1.13, 1.12, 1.99, and 1.87% (median values) respectively, and ranged from 0 to 6.66%. In addition, the deduced amino acid sequence variability of NC and p6 within mothers (B, C, D, E, F, and H) differed by 0, 0.80, 0.81, 2.47, 0.81, and 4.12% (median values) respectively, ranging from 0 to

13.05%. Furthermore, the infants' (B, C, D, E, F, I1H, and I2H) amino acid sequences varied by 0, 4.05, 1.63, 1.63, 2.45, 0, and 2.04% (median values) respectively, and ranged from 0 to 9.31%. The amino acid sequence variability between epidemiologically linked mother-infant pairs (pairs B, C, D, E, F, and H) varied by 0, 5.74, 1.63, 2.47, 3.28, and 3.28% (median values) respectively, and

Table 1: Patient demographics, clinical, and laboratory parameters of six HIV-1 infected mother-infant pairs involved in vertical transmission.

Patient	Age	Sex	CD4+ cells/mm3	Length of infection	Antiviral Drug	Clinical Evaluation
MB	28 yr	F	509	11 mo	None	Asymptomatic
IB	4.75 mo	M	1942	4.75 mo	None	Asymptomatic
MC	23 yr	F	818	1 yr 6 mo	None	Asymptomatic
IC	14 mo	F	772	14 mo	ZDV	Symptomatic AIDS
MD	31 yr	F	480	2 yr 6 mo	None	Asymptomatic
ID	28 mo	M	46	28 mo	ddC	Symptomatic AIDS; failed ZDV therapy
ME	26 yr	F	395	2 yr	ZDV	Symptomatic AIDS
IE	34 mo	M	588	34 mo	ZDV	Symptomatic AIDS
MF	23 yr	F	692	2 yr 10 mo	None	Asymptomatic
IF	1 wk	M	2953	1 wk	ZDV	Asymptomatic
MH	33 yr	F	538	5 mo	None	Asymptomatic
IH1	7 mo	F	3157	7 mo	ACTG152	Hepatosplenomegaly, lymphadenopathy
IH2	7 mo	F	2176	7 mo	ACTG152	Hepatosplenomegaly, lymphadenopathy

M: Mother; I: Infant
Length of infection: The closest time of infection that could be documented was the first positive HIV-1 serology date or the first visit of the patient to the AIDS treatment center, where all the HIV-1 positive patients were referred to as soon as an HIV-1 test was positive. As a result, these dates may not reflect the exact dates of infection.
Clinical evaluation for the infants is based on CDC criteria [70]
Mother and infant samples for each pair were collected at the same time

ranged from 0 to 14.55%. The nucleotide and amino acid sequence variability was also calculated between epidemiologically unlinked individuals. It was determined that the nucleotide distances gave a median value of 7.68, while the amino acid distances produced a median of 14.68. A comparison revealed that the variability between epidemiologically linked mother-infant pairs was lower than the variability between epidemiologically unlinked individuals. This suggests that epidemiologically linked sequences were closer to each other evolutionarily than to unlinked sequences.

We also evaluated if the low variability of NC sequences seen in our mother-infant pair isolates was due to errors

made by Platinum *Pfx* Taq polymerase used in our study. We did not find any errors made by the Taq polymerase when we used a known sequence of HIV-1 NL4-3 for PCR amplification and DNA sequencing of the NC gene.

Dynamics of HIV-1 NC and p6 gene evolution in mother-infant pairs

Different models of evolution were suggested by Modeltest 3.06 [47] based on maximum likelihood estimates and chi square tests that were performed by the program. The estimates of genetic diversity of the NC and p6 sequences obtained were determined using the Watterson model, which assumes segregated sites, and the Coalesce model, which assumes a constant population size. These

Table 3: Estimates of genetic diversity of the NC and p6 sequences from six HIV-1 infected mother-infant pairs involved in vertical transmission.

Patient	Within Mothers		Within Infants	
	θ_W	θ_C	θ_W	θ_C
B	0.004	0.009	0.003	0.002
C	0.011	0.017	0.031	0.075
D	0.022	0.017	0.011	0.020
E	0.016	0.015	0.015	0.024
F	0.009	0.011	0.025	0.061
H	0.023	0.016	-	-
I1H	-	-	0.001	0.001
I2H	-	-	0.017	0.017
Total	0.014	0.014	0.015	0.029

θ_W: viral diversity as estimated by the Watterson method.
θ_C: viral diversity as estimated by the Coalesce method.
Totals were calculated as the average of all values.

Table 2: Nucleotide and amino acid distances of the NC and p6 sequences from mother sets, infant sets, and between mother-infant pairs.

Nucleotide Distances								
Pair	Within Mother			Within Infant			Between Mother and Infant		
	Min	Med	Max	Min	Med	Max	Min	Med	Max
B	0	0.26	0.80	0	0	0.80	0	0	1.10
C	0	0.53	2.12	0	2.26	3.77	0	2.94	5.44
D	0	0.84	2.03	0	0.88	2.57	0	1.13	3.18
E	0	1.13	2.37	0	1.11	2.39	0	1.12	2.68
F	0	0.27	1.62	0	1.78	3.70	0	1.99	4.82
H	0	1.89	6.30	-	-	-	0	1.62	6.28
I1H	-	-	-	0	0	0.27	-	-	-
I2H	-	-	-	0	3.22	5.03	-	-	-
Total	0	0.53	6.30	0	1.16	5.03	0	1.38	6.28

Amino Acid Distances								
Pair	Within Mother			Within Infant			Between Mother and Infant		
	Min	Med	Max	Min	Med	Max	Min	Med	Max
B	0	0	4.12	0	0	1.63	0	0	4.12
C	0	0.8	4.89	0	4.05	7.49	0	5.74	14.55
D	0	0.81	4.12	0	1.63	4.12	0	1.63	5.87
E	0	2.47	6.74	0	1.63	4.97	0	2.47	6.74
F	0	0.81	3.28	0	2.45	5.82	0	3.28	8.43
H	0	4.12	13.05	-	-	-	0	3.28	13.05
I1H	-	-	-	0	0	0	-	-	-
I2H	-	-	-	0	2.04	9.31	-	-	-
Total	0	0.81	13.05	0	1.63	9.31	0	2.45	14.55

M: Mother; I: Infant. Min: Minimum; Med: Median; Max: Maximum.
Totals were calculated for all pairs together.

estimates of genetic diversity are displayed as theta values, and represent the rate of mutation per site per generation (Table 3). The Watterson model estimated the level of genetic diversity within infected mothers to be 0.014, and within infected infants to be 0.015. Slightly greater estimates were obtained using the Coalesce method, with the genetic diversity between mothers being 0.014, and between infants 0.029. Together these data suggest that both the mother and infant populations evolved slowly and at similar rates. The difference between the estimates of genetic diversity between the mother and infant sequences, using either method, is not statistically significant.

Rates of accumulation of non-synonymous and synonymous substitutions
The ratio of the accumulation of non-synonymous (dn) to synonymous substitutions (ds) was used to estimate the

selection pressure on the NC and p6 gene by using a model modified by Nielson and Yang [48], which was then implemented by codeML [49]. The advantage of the codeML method lies in the fact that this model views the codon as the unit of evolution, as opposed to the nucleotide which is used in other models [50]. Moreover, the Nielson and Yang model does not assume that all sites within a sequence are under the same selection pressure. This gives a more realistic view of evolution because mutations, in some cases leading to only a single amino acid change, can be more advantageous or deleterious in some regions of a protein compared to others, and thus undergoes positive or purifying selection. In addition the dn/ds ratio that is calculated determines the selection pressure acting upon the changes within the codon, with a dn/ds ratio of greater than 1 indicating that positive selection pressure is present. Not only does this model determine positive selection pressure, it also calculates the percent-

age of mutations that are selected. The percent of mutations that are conserved fall in the p1 category, the neutral mutations are in the p2 category, and the positively selected mutations are in the p3 category. The estimations of the dn/ds ratio as well as the percentages in each category (p1, p2, and p3) for each patient sample are given in Table 4. All of the sequence populations analyzed displayed a dn/ds ratio greater than or equal to 1.

In general, the mother sequences displayed a higher percentage of positively selected p3 sites compared to the infants. Within mothers, almost 100% of the mutations in mothers B, C, and F were positively selected. Although mother D and mother H have the highest dn/ds values, less than 1% of the mutations are positively selected. Most of the mutations in mother D and mother H are neutral. When compared to the mothers, infants have less than 3% of mutations that are positively selected, with the exceptions of infant D and the second infant H twin (I2H). In contrast to the mothers, the infants have a more even distribution of conserved and neutral mutations. It is interesting to note that in four of the seven infants, over 50% of the mutations observed were neutral mutations. This higher proportion of p2 sites in infants was also seen in analysis of the *nef* and reverse transcriptase (RT) genes [12,51]. The positive selection pressure acting on these patient sequences was estimated in codeML using both neutral models and positive selection models. In patients where a substantial proportion of mutations were in the p3 category, the positive selection model was significant over the neutral model (data not shown). These data indicate that a higher percentage of mutations are positively selected in mothers as compared to infants, however positive selection pressure was observed when analyzing the NC gene sequences from both the mother and infant patient samples.

Analysis of functional domains of NC and p6 within mother-infant pairs

The function of the HIV-1 NC protein is to bind to viral RNA and DNA. This protein contains two zinc fingers and many basic amino acids that allow it to interact with the viral nucleic acids. The critical residues of the zinc fingers consist of three cysteines and one histidine, and have the sequence $C-X_2-C-X_4-H-X_4-C$, with X representing any amino acid, and are located at positions 16 to 29 and 37 to 50 within the NC protein [20]. The critical residues within these zinc fingers are located at positions 16, 19, 24, and 29 in the first zinc finger and positions 37, 40, 45, and 50 in the second zinc finger. A mutation at any of these critical residues abolishes the ability of these functional domains to bind the zinc cofactor, which will lead to improper folding of the protein [24,29]. Analysis of the first zinc finger sequence from the six mother-infant pairs shows that of the 168 sequences acquired, only two con-

tained mutations at the critical residues (Figs. 2, 3, 4). Infant C clone 2 (IC-2) contained the substitution C19R, and mother B clone 2 (MB-2) (Fig. 2) contained the substitution H24Y. Furthermore, the second zinc finger contained substitutions at the critical residues in only one clone; infant C clone 3 (IC-3) contained an H45Y substitution (Fig. 2). However some sequences within mother H and the second infant H twin (I2H) contain substitutions that resulted in the formation of a stop codon at position 38 within the second zinc finger (Fig. 4). These stop codons would result in a truncation in the second zinc finger, and would result in only one functional zinc finger (the first zinc finger) within the NC protein of these clones. When two zinc fingers are present, the first generally tends to play a more critical role [18,20], however removal of the second zinc finger function has been shown to greatly decrease the annealing capacity of the NC protein [20,29]. Despite these exceptions, the critical residues of both zinc fingers within the mother-infant NC sequences were highly conserved.

There are several basic residues, arginine (R), lysine (K), or histidine (H), within the NC protein that also allow it to function. Of the 56 amino acids that make up the NC protein, 17 are basic [21]. These basic residues spread throughout the protein and are responsible for interacting with the side chains on viral nucleic acids [18,52]. Mutations in these basic residues has been shown to reduce RNA binding and encapsidation [21]. Analysis of the sequences from the mother-infant pairs shows that there are substitutions at many of the basic residues. However looking more in depth, a majority of the substitutions are from one basic amino acid to another. Furthermore, there are several substitutions from non-basic to basic residues throughout the protein sequences obtained, and some of these substitutions are compensatory mutations for changes from a basic amino acid elsewhere within the sequence (Figs. 2, 3, 4). While there are several substitutions involving basic amino acids within the NC protein sequences from the six mother-infant pairs, the presence of several basic residues throughout the protein sequences is highly conserved.

The p6 gene was also sequenced as a result of sequencing the NC gene. The p6 protein contains two major functional domains, the viral late domain located at positions 79 to 83, and the Vpr binding domains located at positions 87 to 90 and 107 to 118 [30,33,45]. The late domain contains the sequence proline-threonine-alanine-proline-proline (PTAPP) and is responsible for ensuring proper budding of a newly formed virion from the host cell membrane [32,53]. The prolines at positions 82 and 83 have especially been shown to be critical for Tsg101 binding [32]. Analysis of the p6 protein sequences from the six mother-infant pairs revealed that the late domains, espe-

Table 4: Ratio of nonsynonymous (dn) to synonymous (ds) substitutions in NC and p6 sequences from six HIV-1 infected mother-infant pairs involved in vertical transmission.

Pair	Mothers					Infants				
	N	p1	p2	p3	dn/ds	N	p1	p2	p3	dn/ds
B	13	0	0	100	1.37	11	0	100	0	1.00
C	9	0	0	100	1.51	15	11.72	86.09	2.19	22.84
D	17	0	99.02	0.98	68.79	12	87.26	0	12.74	6.75
E	11	83.68	0	16.32	33.75	12	48.04	51.02	0.94	37.57
F	18	0.03	0	99.97	1.20	15	21.48	77.20	1.32	9.00
H	16	0	100	0	89.00	-	-	-	-	-
I1H	-	-	-	-	-	11	100	0	0	21.24
I2H	-	-	-	-	-	8	68.65	0	31.34	5.47
Total	84	13.95	33.17	52.88	32.60	84	48.16	44.90	6.94	14.82

N: Number of clones sequenced, Totals were calculated as the average of all values.
p1: proportion of conserved codons as a percent
p2: proportion of neutral codons as a percent
p3: proportion of positively selected codons as a percent; dn/ds = dn/ds ratio at p3

cially the critical prolines, are conserved in most of the sequences obtained (Figs. 2, 3, 4). Interestingly, in several sequences from mother D there is a duplication of the late domain (Fig. 3). It has been shown that duplication of this domain could be linked to antiretroviral drug resistance [54,55]. However since mother D has not been exposed to antiretroviral drugs (Table 1), this duplication must have arisen naturally or was present in the virus that was initially transmitted to mother D. In general, the late domain of the p6 protein from the mother-infant pairs was highly conserved.

The Vpr binding domain could be located in two possible positions within the p6 protein sequences of the six mother-infant pairs, either positions 87 to 90 or 107 to 118 [30,33,45] (Fig. 2). The domain located at positions 87 to 107 has the sequence phenylalanine-arginine-phenylalanine-glycine (FRFG) [30], while the domain at positions 107 to 118 has the sequence leucine-XX-leucine-XX-leucine-XX-leucine-XX ((LXX)$_4$)[45], with X representing any amino acid. These Vpr binding domains are responsible for inclusion of the viral accessory protein Vpr into newly forming virions. Analysis of the protein sequences from the mother-infant pairs revealed that while the FRFG Vpr binding domain was mostly conserved, there were some notable exceptions. There were single amino acid substitutions within the domain in every clone of mother and infant F (pair F), infant C (IC), and infant D (ID) (Figs. 2, 3, 4). It has been shown that mutations at either of the two phenylalanines within the FRFG domain, which is seen in pair F and infant D, causes a loss of Vpr packaging within virions; while a substitution at the

arginine site, which is seen in infant C, seems to have little to no effect [30]. In spite of these exceptions however, the FRFG Vpr binding domain within the six mother-infant pairs analyzed was mostly conserved. Analyzing the protein sequences also showed that the (LXX)$_4$ domain was also mostly conserved within the sequences obtained, except for the first leucine in every clone. This first leucine was substituted with either a methionine (M), a valine (V), a histidine (H), an arginine (R), or a glutamine (Q) (Figs. 2, 3, 4). A change in this first leucine has been shown to decrease Vpr binding [45]. The third and fourth leucine have been shown to be critical for Vpr inclusion [33,34], and these residues are highly conserved within the mother-infant sequences obtained. As with the FRFG domain, the (LXX)$_4$ Vpr binding was mostly conserved within the sequences of the mother-infant pairs analyzed.

The p6 gene product also contains a region, from amino acid positions 31–46 with the sequence DKELYPLASLRSLFG that is responsible for interacting with the host cell factor AIP1 [31]. This motif within the mother-infant pair sequences was mostly conserved, however every clone analyzed contained a substitution at the first leucine, as also seen in the (LXX)$_4$ domain (Figs 2, 3, 4). Mother and infant C (pair C) (Fig. 2) and mother and infant D (pair D) (Fig. 3) also contained additional substitutions within the AIP1 binding domain. It is not known at this time what effect these substitutions would have on the interaction of p6 with AIP1. Despite these exceptions, the AIP1 binding domain was mostly conserved within the six mother-infant pairs' sequences obtained.

In addition to the substitutions mentioned, there were several other substitutions that occurred outside of the functional domains. The effect that these changes would have is not known at this time.

Analysis of immunologically relevant mutations within the CTL epitopes of NC and p6

The cytotoxic T-lymphocyte (CTL) response is known to contribute a significant portion of the body's immune response to an HIV-1 infection. In patients with a strong CTL response, there has been shown to be a decrease in the viral load within the patient and a long-term non-progressor disease status [56]. It has also been shown that mothers who transmit the virus to their infants have an increased number of CTL escape variants when compared to non-transmitting mothers [57]. This could demonstrate a correlation between the amount of CTL escape variants circulating in the mothers' bloodstream and the likelihood of vertical transmission of the virus. There have been several CTL epitopes identified within the NC and p6 proteins. The first epitope within the NC protein has the sequence CRAPRKKGC and is located between amino acid positions 28 and 36 [58]. This epitope is recognized by HLA-B14 and contains the last cystine of the first zinc finger, and the first cystine of the second zinc finger. Analysis of the NC amino acid sequences from the six mother-infant pairs revealed that this epitope was highly conserved in most of the clones that were obtained (Figs. 2, 3, 4). Another CTL epitope, KEGHQMKDCTERQANF, is located at amino acid positions 42–57 and is recognized by several HLA types [58]. This epitope spans the last 14 amino acids of the NC protein and contains the histidine and final cystine of the second zinc finger. Again this epitope was mostly conserved when the sequences from the mother-infant pairs was analyzed. The next motif, CTERQANFL, is located from positions 50 to 56 and is recognized by HLA-B61 [58]. This epitope contains the last cystine of the second zinc finger and was highly conserved within the mother-infant sequences obtained.

The first motif within the p6 gene sequences, GNFLQS-RPEPTAPPF, is located at amino acid positions 70–84 and is recognized by several HLA types [58]. Analysis of the mother-infant sequences revealed that this epitope was mostly conserved, with the exception of mother and infant C (pair C). Pair C contains a PTV insertion beginning at position 78 (Fig. 2). It is not known at this time what effect on CTL recognition this insertion would have. The next epitope is located at amino acid positions 105 to 114 and has the sequence KELYPLTSL [58]. This epitope is recognized by HLA-B60 and is positioned within the $(LXX)_4$ Vpr binding domain. Within the mother-infant sequences obtained this CTL epitope was mostly conserved, however the first lysine within the epitope was substituted within every clone analyzed (Figs. 2, 3, 4). It is

not known at this time what effect this substitution would have on recognition of this epitope. Another epitope, YPLTSLRSLF, is located at positions 108 to 117 and is recognized by HLA-B7 [58]. Analysis of the six mother-infant pairs' sequences revealed that this epitope was mostly conserved (Figs. 2, 3, 4). Overall, analysis of the CTL recognition epitopes within the sequences of the mother-infant pairs' displayed that these epitopes which are involved in immune recognition of the virus were mostly conserved.

Discussion

In this study, we have shown that the *gag* p17, NC and p6 genes, of HIV-1 were mostly conserved during vertical transmission. Six mother-infant pairs were analyzed and the NC and p6 open reading frames were found to be conserved with a frequency of 92.8%. When distance analysis and population dynamics were performed, it was found that there was a low degree of viral heterogeneity as well as genetic diversity. In spite of this, a positive selection pressure was found be to acting on the NC and p6 gene sequences. Comparison of the mother-infant pairs to each other revealed that the NC and p6 sequences from epidemiologically linked individuals were more closely related to each other than to epidemiologically unlinked individuals. While the individual pairs could clearly be distinguished from one another based on sequence analysis, the functional domains within all sequences analyzed remained mostly conserved. These findings suggest that maintenance of conserved NC and p6 genes during vertical transmission is important for pathogenesis of HIV-1 in mothers and infants.

The open reading frames (ORF) of the NC and p6 genes sequenced from six mother-infant pairs involved in vertical transmission were highly conserved. Of the 168 sequences analyzed, 156 contained an intact ORF (Fig 2). The remaining sequences contained a substitution that resulted in the formation of a premature stop codon. This data is comparable to previously analyzed genes, including *gag* p17, *vif, vpr, tat, vpu, nef*, and reverse transcriptase [12-16,51,59,60]. Maintenance of the NC and p6 ORFs confirms the importance of this gene in the viral lifecycle, and suggests its importance in vertical transmission of the virus.

Several pair and patient specific sequence motifs were observed when a multiple sequence alignment was performed. There were also three substitutions that were present in every sequence analyzed (Figs 2, 3, 4). These universal mutations could provide a possible target for future diagnosis or for antiretroviral therapy development. The NC has already become a promising target for antiviral therapy [24,61-63], perhaps the information that this study provides can help further the knowledge of

HIV-1 vertical transmission and help to design preventative methods. Analysis of the nucleotide and amino acid distances revealed a low degree of viral heterogeneity within the sequences analyzed (Fig 3). A low degree of genetic diversity was also found by the Watterson and Coalesce methods (Fig 4). The mutation rate per site per generation (the θ value) was slightly, although not significantly, higher in infants than in mothers. This slight increase in mutation rate could account for the slightly higher heterogeneity within the infant sequences. These values could be higher within the infants due to the fact that the virus within the infants has had less time to adapt to its new host than in the mothers. This is also supported by the observation that mothers have a higher percentage of mutations that are positively selected as compared to infants, where most of the mutations were neutral or conserved. A higher percentage of positively selected mutations could indicate that the virus has evolved for a longer duration in the mothers. Since it is likely that the infants have been infected for a shorter duration than the mothers, the viral variants in the infants have been exposed to the selection pressure of the new host for a shorter time. Although the distance values indicate that the infants were slightly more diverse, phylogenetic analysis reveals that each of the six mother-infant pairs clustered together in distinct subtrees (Fig 1). This reveals that although the virus is evolving in separate hosts, it is still more similar to the virus in its epidemiologically linked host, than to the virus in an epidemiologically unlinked host. Information of this kind can be helpful in establishing an epidemiological relationship between transmitter-recipient pairs.

The functional domains in the NC and p6 genes from the six mother-infant pairs were mostly conserved during vertical transmission. The critical residues of the NC zinc fingers were highly conserved, while the basic residues throughout the NC protein displayed more variability. However, these changes within the basic residues did not result in an overall loss of these basic amino acids. In fact, most of the basic amino acids that were substituted were replaced by another basic residue. While some changes did result in a loss of a basic residue, other compensatory mutations elsewhere in the amino acid sequence replaced the residue lost. This conservation of critical zinc finger residues and presence of many basic residues implies that it is very important to the virus evolutionarily to maintain the function these motifs provide. The viral late domain, Vpr binding sites, and AIP1 binding site were all mostly conserved as well, when the mother-infant p6 sequences were analyzed. Critical residues within the late domain, including the two prolines vital for Tsg101 binding, were highly conserved during vertical transmission. This domain was not only conserved, but duplicated within some mother D sequences. It has been shown that duplication of this domain could be linked to antiretroviral

drug resistance [54,55,64]. However since mother D has not been exposed to antiretroviral drugs (Table 1), this duplication must have arisen naturally or was present in the virus that was initially transmitted to mother D. It is not known at this point, however, what effect this duplication would have on the budding ability of the virus. Both of the Vpr binding domains were mostly conserved in the mother-infant sequences obtained, as was the AIP1 binding motif. Again, the conservation of these functional domains during vertical transmission suggests that it is important to the virus evolutionarily to maintain the functions that these regions provide.

Several substitutions were seen outside of the functional domains mentioned. While the relevance of these changes is not known at this time, the effects could be studied by performing biological studies using the NC clones obtained within this study. Another study of interest would be to characterize the same NC region in mothers who naturally failed to transmit the virus to their offspring and compare the results to that of this study.

There were several CTL epitopes, which were recognized by several different HLA types, within the mother-infant NC sequences obtained. It has been shown that transmitting mothers have larger numbers of CTL escape variants as compared to non-transmitting mothers, but the transmitted viruses carrying epitopes are not escape variants [57]. It is entirely possible however, that the CTL responses studied are tissue specific. A representation of peripheral blood, and the virus and CTL variants in the placenta, birth canal, and breast milk are different [65]. There were substitutions within the CTL epitopes of the NC sequences from the six mother-infant pairs; however it is unknown what effect these changes may have. It is possible that the substitutions observed in this study may influence vertical transmission by resulting in differential responses in a tissue specific manner.

Although great strides have been made in the prevention of HIV-1 vertical transmission in countries such as the United States, the infection of children vertically in developing countries remains a large problem. In order deal with this problem, a better understanding of the mechanisms involved needs to be established. A characterization of many of the viral genes during vertical transmission has already been completed [1,2,11,12,14-16,51,59,60], and has shed new light on the molecular mechanisms of an HIV-1 vertically acquired infection. Using this previous information, new targets and strategies may be able to be developed in order to prevent the spread of this disease. The data presented in this study may be able to provide new insights into the molecular characteristics of HIV-1 vertical transmission and further our understanding so

that novel treatment and prevention strategies may be developed.

Conclusion

In this study we have shown that the HIV-1 NC and p6 genes were mostly conserved in six mother-infant pairs following vertical transmission. Phylogenetic analysis revealed that virus from epidemiologically linked mother-infant pairs was closer to each other than to epidemiologically unlinked pairs. The NC and p6 open reading frames were highly conserved and estimates of viral heterogeneity and genetic diversity were low. Several patient and pair specific substitutions were seen within the NC amino acid sequence. The functional domains within the NC and p6 sequences, which include two zinc fingers, several basic residues, a viral late domain, two possible Vpr binding sites, and a binding site for AIP1, were widely conserved following vertical transmission. The data presented in this study provides evidence that supports the critical role of the NC gene product in the viral lifecycle and in pathogenesis of HIV-1 during vertical transmission.

Materials and methods

Patient population

Blood samples were collected from six mother-infant pairs following vertical transmission, including a set of twins (I1H and I2H) in the case of mother H. The demographics, clinical and laboratory findings on these mother-infant pairs is summarized in Table 1. The Human Subjects Committee of the University of Arizona (Tucson, AZ) and the Institutional Review Board of the Children's Hospital Medical Center (Cincinnati, Ohio) approved this study. Written informed consent to participate in this study was obtained from the mothers of the mother-infant pairs.

PCR amplification, cloning, and sequencing

Peripheral blood mononuclear cells (PBMCs) were isolated using a single-step Ficoll-Hypaque method (Pharmacia-LKB) from whole blood samples of HIV-1 infected mother-and-infant pairs that were involved in vertical transmission. The PBMCs DNA, which contains the integrated HIV-1 genome, was isolated as previously described [9]. The HIV-1 NC gene was amplified using a two-step polymerase chain reaction (PCR) method according to the modified protocol described by Ahmad et al [9]. Outer primers NC-1 (5'GAAGAAATGATGACAG-CATGTCAGGGAGTGGG, 1819 to 1851, sense) and NC-2 (5'CCATCTTCCTGGCAAATTCATTTCTTCTAATACT, 2344 to 2378, antisense) were first used, followed by nested primers NC-3 (5'CACCGGCCATAAAGCAA-GAGTTTTGGCTGAAGC, 1854 to 1887, sense) and NC-4 (5' CATCTGCTCCTGTATCTAATAGAGCTTCCTT, 2310 to 2341, antisense). Equal amounts of PBMC DNA (approximately 25 to 50 copies, minimum) as determined by end-point dilution was subjected to multiple (four to six) PCRs to obtain clones that were sequenced and analyzed. PCR was carried out in a 25μl reaction mixture containing 2.5 μl 10× *Pfx* Amplification buffer, 2.5 mM MgSO$_4$, 400 μM of each dATP, dCTP, dGTP, and dTTP, 0.2 μM of each outer primer, and 2.5 units (U) of Platinum *Pfx* DNA polymerase (Invitrogen Inc.). The reaction was initiated at 94°C for 2 minutes (min), and then cycled at 94°C for 30 seconds (sec), 50°C for 30 sec, and 72°C for 2 min, for 35 cycles, with an addition extension period of 10 min at 72°C to end the reaction. Following the first round of PCR, 5–8 μl of the first-PCR product was used for nested PCR using the same reagents and the inner primers. This nested PCR was cycled at 94°C for 30 sec, 55°C for 30 sec, and 72°C for 2 min, for 35 cycles. A negative control was included with each PCR which used sterile water in place of DNA. PCR was also performed on HIV-1 NL4-3, of which the sequence is known (GenBank accession number M19921), to assess any errors made by the Platinum *Pfx* DNA polymerase. To avoid contamination, all samples, reagents and PCR products were kept separately and dispensed in a separate room free of all laboratory-used DNA.

The PCR products were visualized on a 1% agarose gel and cloned into the TOPO TA cloning system (pCR 2.1-TOPO vector) as per manufacturer's instructions (Invitrogen Inc.). Positive bacterial colonies were determined by blue-white screening, and the presence of correct-sized inserts was confirmed by restriction digest. Eight to eighteen clones from each mother and infant, from multiple independent PCRs, were sequenced using the Thermosequenase Cycle Sequencing protocol (USB) and the University of Arizona Biotechnology Center automated system (ABI Prism® 3700 DNA automated sequencing system).

Sequence analysis

The nucleotide sequences of the HIV-1 NC gene (approximately 375 bp) from six mother-infant pairs were analyzed using the Wisconsin package version 10.1 of the Genetics Computer Group (GCG) and were translated to corresponding deduced amino acid sequences (125 amino acids). Both the nucleotide and amino acid sequences were aligned, using the HIV-1 NL4-3 NC sequence as a reference, by Clustal X. A model of evolution was optimized for the entire nucleotide sequence data set using the Huelsenbeck and Crandall approach [66]. Likelihood models of evolution were calculated using PAUP [46] and a chi square (χ^2) test was performed using Modeltest 3.06 [47]. The model with the highest likelihood was incorporated into PAUP to generate a neighbor-joining tree, which was bootstrapped 1000 times to ensure fidelity. The tree was based on the nucleotide sequences from the six mother-infant pairs, as well as HIV-1 NL4-3 which was used as a reference sequence

and the outgroup for the tree. Using Modeltest and the Akaike Information Criterion (AIC) [67], all the null hypotheses were rejected except the best-fit model (GTR+G). The base frequencies for the NC gene were calculated as: freq A = 0.37, freq C = 0.23, freq G = 0.21 and freq T = 0.19. Five rate categories were also figured, and were as follows: R(A- C)= 1.00, R (A-G)= 5.34, R(A-T) = 0.79, R(C-G) = 0.79, R(C-T) = 5.34, R(G-T) = 1.00. The rate heterogeneity was taken into account using a gamma distribution with a shape parameter (α) of the distribution estimated from the data via maximum likelihood. This shape parameter had a value of 0.7085. A model of evolution for each patient was also generated and optimized to estimate corrected pairwise nucleotide distances using PAUP [46]. Amino acid distances were also estimated using the Jukes-Cantor model within the Wisconsin package version 10.1 of GCG. The minimum, median, and maximum distance values were calculated for both nucleotides and amino acids for each patient as well as for the linked and unlinked patient pairs. The dynamics of HIV-1 evolution was assessed using techniques of population genetics. In population genetics, genetic diversity is defined as $\theta = 2N_e\mu$, where N_e is the effective population size and μ is the per nucleotide mutation rate per generation. The Watterson model, which assumes segregated sites, and the Metropolis-Hastings model, which assumes a constant population size, was used to estimate the differences in genetic diversity using the Coalesce 1.5 program [68,69]. To analyze the evolutionary processes acting upon the NC gene, we estimated the ratio of non-synonymous (dn) to synonymous (ds) substitutions by a maximum likelihood model using codeML, which is part of the PAML package [49]. The Nielsen and Yang [48] model considers the codon instead of the nucleotide as the unit of evolution and thus incorporates three distinct categories of sites. The first category represents the sites that are invariable or conserved (p1, dn/ds = 0); the second category represents sites that are neutral (p2, dn/ds = l), at which dn and ds are fixed at the same rate; and the third category represents sites that are under positive selection, where dn has a higher fixation rate than ds (p3, dn/ds >1). The dn/ds was estimated for each patient using both neutral and positive selection models in codeML.

Nucleotide sequence accession numbers
The nucleotide accession numbers of the sequences submitted to GenBank are <u>DQ026833</u> through <u>DQ027000</u>.

Competing interests
The author(s) declare that they have no competing interests.

Authors' contributions
BW carried out the PCR, cloning, and sequencing. BW, VS, and RR performed the sequence analysis by computer programs. BW and NA participated in the experimental design, data interpretation, and writing of the manuscript. All authors read and approved the final manuscript.

References
1. Ahmad N: **The vertical transmission of human immunodeficiency virus type 1: molecular and biological properties of the virus.** *Crit Rev Clin Lab Sci* 2005, **42**:1-34.
2. Ahmad N: **Molecular mechanisms of human immunodeficiency virus type 1 mother-infant transmission.** *Adv Pharmacol* 2000, **49**:387-416.
3. Ahmad N: **Maternal-Fetal Transmission of Human Immunodeficiency Virus.** *J Biomed Sci* 1996, **3**:238-250.
4. Blanche S, Rouzioux C, Moscato ML, Veber F, Mayaux MJ, Jacomet C, Tricoire J, Deville A, Vial M, Firtion G: **A prospective study of infants born to women seropositive for human immunodeficiency virus type 1. HIV Infection in Newborns French Collaborative Study Group.** *N Engl J Med* 1989, **320**:1643-1648.
5. Mok JQ, Giaquinto C, De Rossi A, Grosch-Worner I, Ades AE, Peckham CS: **Infants born to mothers seropositive for human immunodeficiency virus. Preliminary findings from a multicentre European study.** *Lancet* 1987, **1**:1164-1168.
6. Ryder RW, Nsa W, Hassig SE, Behets F, Rayfield M, Ekungola B, Nelson AM, Mulenda U, Francis H, Mwandagalirwa K: **Perinatal transmission of the human immunodeficiency virus type 1 to infants of seropositive women in Zaire.** *N Engl J Med* 1989, **320**:1637-1642.
7. Lepage P, Van de Perre P, Carael M, Nsengumuremyi F, Nkurunziza J, Butzler JP, Sprecher S: **Postnatal transmission of HIV from mother to child.** *Lancet* 1987, **2**:400.
8. **Mother-to-child transmission of HIV infection. The European Collaborative Study.** *Lancet* 1988, **2**:1039-1043.
9. Ahmad N, Baroudy BM, Baker RC, Chappey C: **Genetic analysis of human immunodeficiency virus type 1 envelope V3 region isolates from mothers and infants after perinatal transmission.** *J Virol* 1995, **69**:1001-1012.
10. Wolinsky SM, Wike CM, Korber BT, Hutto C, Parks WP, Rosenblum LL, Kunstman KJ, Furtado MR, Munoz JL: **Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants.** *Science* 1992, **255**:1134-1137.
11. Matala E, Hahn T, Yedavalli VR, Ahmad N: **Biological characterization of HIV type 1 envelope V3 regions from mothers and infants associated with perinatal transmission.** *AIDS Res Hum Retroviruses* 2001, **17**:1725-1735.
12. Hahn T, Ramakrishnan R, Ahmad N: **Evaluation of genetic diversity of human immunodeficiency virus type 1 NEF gene associated with vertical transmission.** *J Biomed Sci* 2003, **10**:436-450.
13. Husain M, Hahn T, Yedavalli VR, Ahmad N: **Characterization of HIV type 1 tat sequences associated with perinatal transmission.** *AIDS Res Hum Retroviruses* 2001, **17**:765-773.
14. Yedavalli VR, Chappey C, Matala E, Ahmad N: **Conservation of an intact vif gene of human immunodeficiency virus type 1 during maternal-fetal transmission.** *J Virol* 1998, **72**:1092-1102.
15. Yedavalli VR, Chappey C, Ahmad N: **Maintenance of an intact human immunodeficiency virus type 1 vpr gene following mother-to-infant transmission.** *J Virol* 1998, **72**:6937-6943.
16. Yedavalli VR, Husain M, Horodner A, Ahmad N: **Molecular characterization of HIV type 1 vpu genes from mothers and infants after perinatal transmission.** *AIDS Res Hum Retroviruses* 2001, **17**:1089-1098.
17. Yedavalli VR, Ahmad N: **Low conservation of functional domains of HIV type 1 vif and vpr genes in infected mothers correlates with lack of vertical transmission.** *AIDS Res Hum Retroviruses* 2001, **17**:911-923.

18. Poon DT, Wu J, Aldovini A: **Charged amino acid residues of human immunodeficiency virus type 1 nucleocapsid p7 protein involved in RNA packaging and infectivity.** *J Virol* 1996, **70:**6607-6616.
19. Zuber G, McDermott J, Karanjia S, Zhao W, Schmid MF, Barklis E: **Assembly of retrovirus capsid- nucleocapsid proteins in the presence of membranes or RNA.** *J Virol* 2000, **74:**7431-7441.
20. Heath MJ, Derebail SS, Gorelick RJ, DeStefano JJ: **Differing roles of the N- and C-terminal zinc fingers in human immunodeficiency virus nucleocapsid protein-enhanced nucleic acid annealing.** *J Biol Chem* 2003, **278:**30755-30763.
21. Freed EO: **HIV-1 gag proteins: diverse functions in the virus life cycle.** *Virology* 1998, **251:**1-15.
22. Krishnamoorthy G, Roques B, Darlix JL, Mely Y: **DNA condensation by the nucleocapsid protein of HIV-1: a mechanism ensuring DNA protection.** *Nucleic Acids Res* 2003, **31:**5425-5432.
23. Feng YX, Campbell S, Harvin D, Ehresmann B, Ehresmann C, Rein A: **The human immunodeficiency virus type 1 Gag polyprotein has nucleic acid chaperone activity: possible role in dimerization of genomic RNA and placement of tRNA on the primer binding site.** *J Virol* 1999, **73:**4251-4256.
24. Rice WG, Baker DC, Schaeffer CA, Graham L, Bu M, Terpening S, Clanton D, Schultz R, Bader JP, Buckheit RW Jr, Field L, Singh PK, Turpin JA.: **Inhibition of multiple phases of human immunodeficiency virus type 1 replication by a dithiane compound that attacks the conserved zinc fingers of retroviral nucleocapsid proteins.** *Antimicrob Agents Chemother* 1997, **41:**419-426.
25. Azoulay J, Clamme JP, Darlix JL, Roques BP, Mely Y: **Destabilization of the HIV-1 complementary sequence of TAR by the nucleocapsid protein through activation of conformational fluctuations.** *J Mol Biol* 2003, **326:**691-700.
26. McGrath CF, Buckman JS, Gagliardi TD, Bosche WJ, Coren LV, Gorelick RJ: **Human cellular nucleic acid-binding protein Zn2+ fingers support replication of human immunodeficiency virus type 1 when they are substituted in the nucleocapsid protein.** *J Virol* 2003, **77:**8524-8531.
27. Heath MJ, Destefano JJ: **A complementary single-stranded docking site is required for enhancement of strand exchange by human immunodeficiency virus nucleocapsid protein on substrates that model viral recombination.** *Biochemistry* 2005, **44:**3915-3925.
28. Levin JG, Guo J, Rouzina I, Musier-Forsyth K: **Nucleic acid chaperone activity of HIV-1 nucleocapsid protein: critical role in reverse transcription and molecular mechanism.** *Prog Nucleic Acid Res Mol Biol* 2005, **80:**217-286.
29. Lyonnais S, Gorelick RJ, Mergny JL, Le Cam E, Mirambeau G: **G-quartets direct assembly of HIV- 1 nucleocapsid protein along single-stranded DNA.** *Nucleic Acids Res* 2003, **31:**5754-5763.
30. Zhu H, Jian H, Zhao LJ: **Identification of the 15FRFG domain in HIV-1 Gag p6 essential for Vpr packaging into the virion.** *Retrovirology* 2004, **1:**26.
31. Strack B, Calistri A, Craig S, Popova E, Gottlinger HG: **AIP1/ALIX is a binding partner for HIV-1 p6 and EIAV p9 functioning in virus budding.** *Cell* 2003, **114:**689-699.
32. VerPlank L, Bouamr F, LaGrassa TJ, Agresta B, Kikonyogo A, Leis J, Carter CA: **Tsg101, a homologue of ubiquitin-conjugating (E2) enzymes, binds the L domain in HIV type 1 Pr55(Gag).** *Proc Natl Acad Sci USA* 2001, **98:**7724-7729.
33. Kondo E, Gottlinger HG: **A conserved LXXLF sequence is the major determinant in p6gag required for the incorporation of human immunodeficiency virus type 1 Vpr.** *J Virol* 1996, **70:**159-164.
34. Bachand F, Yao XJ, Hrimech M, Rougeau N, Cohen EA: **Incorporation of Vpr into human immunodeficiency virus type 1 requires a direct interaction with the p6 domain of the p55 gag precursor.** *J Biol Chem* 1999, **274:**9083-9091.
35. Goff A, Ehrlich LS, Cohen SN, Carter CA: **Tsg101 control of human immunodeficiency virus type 1 Gag trafficking and release.** *J Virol* 2003, **77:**9173-9182.
36. Gottlinger HG, Dorfman T, Sodroski JG, Haseltine WA: **Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release.** *Proc Natl Acad Sci U S A* 1991, **88:**3195-3199.
37. Huang M, Orenstein JM, Martin MA, Freed EO: **p6Gag is required for particle production from full-length human immunodefi-**

38. ciency virus type 1 molecular clones expressing protease. *J Virol* 1995, **69:**6810-6818.
39. Demirov DG, Ono A, Orenstein JM, Freed EO: **Overexpression of the N-terminal domain of TSG101 inhibits HIV-1 budding by blocking late domain function.** *Proc Natl Acad Sci U S A* 2002, **99:**955-960.
39. Garrus JE, von Schwedler UK, Pornillos OW, Morham SG, Zavitz KH, Wang HE, Wettstein DA, Stray KM, Cote M, Rich RL, Myszka DG, Sundquist WI: **Tsg101 and the vacuolar protein sorting pathway are essential for HIV-1 budding.** *Cell* 2001, **107:**55-65.
40. Demirov DG, Orenstein JM, Freed EO: **The late domain of human immunodeficiency virus type 1 p6 promotes virus release in a cell type-dependent manner.** *J Virol* 2002, **76:**105-117.
41. Martin-Serrano J, Yarovoy A, Perez-Caballero D, Bieniasz PD: **Divergent retroviral late-budding domains recruit vacuolar protein sorting factors by using alternative adaptor proteins.** *Proc Natl Acad Sci U S A* 2003, **100:**12414-12419.
42. von Schwedler UK, Stuchell M, Muller B, Ward DM, Chung HY, Morita E, Wang HE, Davis T, He GP, Cimbora DM, Scott A, Krausslich HG, Kaplan J, Morham SG, Sundquist WI: **The protein network of HIV budding.** *Cell* 2003, **114:**701-713.
43. Strack B, Calistri A, Gottlinger HG: **Late assembly domain function can exhibit context dependence and involves ubiquitin residues implicated in endocytosis.** *J Virol* 2002, **76:**5472-5479.
44. Jenkins Y, Pornillos O, Rich RL, Myszka DG, Sundquist WI, Malim MH: **Biochemical analyses of the interactions between human immunodeficiency virus type 1 Vpr and p6(Gag).** *J Virol* 2001, **75:**10537-10542.
45. Lu YL, Bennett RP, Wills JW, Gorelick R, Ratner L: **A leucine triplet repeat sequence (LXX)4 in p6gag is important for Vpr incorporation into human immunodeficiency virus type 1 particles.** *J Virol* 1995, **69:**6873-6879.
46. Swofford DI: *PAUP* Phylogenetic analysis using parsimony and other methods, vol 4.0.0b2.* Sunderland, MA: Sinauer associated; 1999.
47. Posada D, CK : **MODELTEST: testing the model of DNA substitution.** *Bioinformatics* 1998, **14:**817-818.
48. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148:**929-936.
49. Yang Z: *Phylogenetic Analysis of Maximum Likelihood (PAML) 3.0th edition.* London UCo.; 2000.
50. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3:**418-426.
51. Sundaravaradan V, Hahn T, Ahmad N: **Conservation of functional domains and limited heterogeneity of HIV-1 reverse transcriptase gene following vertical transmission.** *Retrovirology* 2005, **2:**36.
52. Dorfman T, Luban J, Goff SP, Haseltine WA, Gottlinger HG: **Mapping of functionally important residues of a cysteine-histidine box in the human immunodeficiency virus type 1 nucleocapsid protein.** *J Virol* 1993, **67:**6159-6169.
53. Freed EO: **Viral late domains.** *J Virol* 2002, **76:**4679-4687.
54. Peters S, Munoz M, Yerly S, Sanchez-Merino V, Lopez-Galindez C, Perrin L, Larder B, Cmarko D, Fakan S, Meylan P, Telenti A: **Resistance to nucleoside analog reverse transcriptase inhibitors mediated by human immunodeficiency virus type 1 p6 protein.** *J Virol* 2001, **75:**9644-9653.
55. Ibe S, Shibata N, Utsumi M, Kaneda T: **Selection of human immunodeficiency virus type 1 variants with an insertion mutation in the p6(gag) and p6(pol) genes under highly active antiretroviral therapy.** *Microbiol Immunol* 2003, **47:**71-79.
56. Yu XG, Shang H, Addo MM, Eldridge RL, Phillips MN, Feeney ME, Strick D, Brander C, Goulder PJ, Rosenberg ES, Walker BD, Altfeld M: **Important contribution of p15 Gag-specific responses to the total Gag-specific CTL responses.** *Aids* 2002, **16:**321-328.
57. Wilson CC, Brown RC, Korber BT, Wilkes BM, Ruhl DJ, Sakamoto D, Kunstman K, Luzuriaga K, Hanson IC, Widmayer SM, Wiznia A, Clapp S, Ammann AJ, Koup RA, Wolinsky SM, Walker BD: **Frequent detection of escape from cytotoxic T-lymphocyte recognition in perinatal human immunodeficiency virus (HIV) type 1 transmission: the ariel project for the prevention of transmission of HIV from mother to infant.** *J Virol* 1999, **73:**3975-3985.

58. *HIV-1 Immunology CTL, CD8+ T cells on World Wide Web URL* [http://www.hiv.lanl.gov/content/immunology/ctl_search].
59. Hahn T, Ahmad N: **Genetic characterization of HIV type 1 gag p17 matrix genes in isolates from infected mothers lacking perinatal transmission.** *AIDS Res Hum Retroviruses* 2001, **17:**1673-1680.
60. Hahn T, Matala E, Chappey C, Ahmad N: **Characterization of mother-infant HIV type 1 gag p17 sequences associated with perinatal transmission.** *AIDS Res Hum Retroviruses* 1999, **15:**875-888.
61. Segal DJ, Goncalves J, Eberhardy S, Swan CH, Torbett BE, Li X, Barbas CF 3rd: **Attenuation of HIV-1 replication in primary human cells with a designed zinc finger transcription factor.** *J Biol Chem* 2004, **279:**14509-14519.
62. Qiu JT, Liu B, Tian C, Pavlakis GN, Yu XF: **Enhancement of primary and secondary cellular immune responses against human immunodeficiency virus type 1 gag by using DNA expression vectors that target Gag antigen to the secretory pathway.** *J Virol* 2000, **74:**5997-6005.
63. Sei S, Yang QE, O'Neill D, Yoshimura K, Nagashima K, Mitsuya H: **Identification of a key target sequence to block human immunodeficiency virus type 1 replication within the gag-pol transframe domain.** *J Virol* 2000, **74:**4621-4633.
64. Ibe S, Hotta N, Takeo U, Tawada Y, Mamiya N, Yamanaka K, Utsumi M, Kaneda T: **Prevalence of drug-resistant human immunodeficiency virus type 1 in therapy-naive patients and usefulness of genotype testing.** *Microbiol Immunol* 2003, **47:**499-505.
65. Sabbaj S, Edwards BH, Ghosh MK, Semrau K, Cheelo S, Thea DM, Kuhn L, Ritter GD, Mulligan MJ, Goepfert PA, Aldrovandi GM: **Human immunodeficiency virus-specific CD8(+) T cells in human breast milk.** *J Virol* 2002, **76:**7365-7373.
66. Huelsenbeck JPCK: **Phylogeny estimation and hypothesis testing using maximum likelihood.** *Annu Rev Ecol Sys* 1997:437-466.
67. Akaike H: **A new look at the statistical model identification.** *IEEE Trans Autom Contr* 1974, **19:**716-723.
68. Kuhner MK, Yamato J, Felsenstein J: **Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling.** *Genetics* 1995, **140:**1421-1430.
69. *Fetching the COALESCE program on World Wide Web URL* [http://evolution.genetics.washington.edu/lamarc/coalesce.html].
70. **Centers for Disease Control and Prevention Classification system for human immunodeficiency virus (HIV) in children under 13 years of age.** *MMWR* 1987, **36:**225-236.