A THESIS SUBMITTED TO:

THE UNIVERSITY OF LONDON

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE

FACULTY OF MEDICINE

MOLECULAR EPIDEMIOLOGY OF UNRELATED AND RELATED HIV INFECTIONS

Catherine Arnold

June 1996

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or institute of learning

Part of this work has already been published:

Arnold, C., Balfe, P., Clewley, J.P.: Sequence distances between genes of HIV-1 from individuals infected from the same source: implications for the investigation of possible transmission events. 1995. Virology **211**:198-203.

Arnold, C., Barlow, K.L., Parry, J.V., Clewley, J.P.: At least five HIV-1 sequence subtypes (A,B,C,D,A/E) occur in England. 1995. AIDS Res. Hum. Retroviruses **11**:427-429.

Arnold, C., Barlow, K.L., Kaye, S., Loveday, C., Balfe, P., Clewley, J.P.:HIV-1 sequence subtype G transmission from mother to infant: failure of variant sequence subspecies to amplify in the Roche Amplicor test. AIDS Res. Hum. Retroviruses **11**:999-1001.

ABSTRACT

Sequence data derived from human immunodeficiency virus infection can be used to investigate cases of possible virus transmission between individuals, and to determine phylogenetic relationships of the virus in the general population. The method of recovering the viral genome and the choice of region (gene) for sequencing both influence these analyses. Therefore, multiple single molecules were obtained from infected individuals and both the gp120 and p6/protease sequenced. The set of individuals chosen for sequencing included both related and unrelated HIV-1 infections. The transmission cases investigated were a surgeon and patient, two members of a sex circle, four pairs of heterosexual partners, a mother and child, a needlestick inoculation and an occupational exposure. Analysis of these sequences has allowed establishment of whether apparently epidemiologically connected infections show linkage at the molecular level. These sequences were also used for an assessment of the statistical validity of phylogenetic analyses of using various sub-fragments of gp120, containing differing proportions of conserved and variable regions. These subfragment analyses were compared with complete gp120 and inferences made about the minimum required data set for reliable phylogenies.

The data obtained for unrelated infections has been used to investigate the extent of the subtype diversity of HIV-1. It was shown that subtypes A, B, C, D, A/E and G occur in England. These non-B subtypes were obtained from patients in risk groups other than homosexual intercourse. The protein sequences derived from the nucleotide data were used to estimate base substitution rates for gp120. A transition/transversion rate of 1.96 was found for the 51 sequences analysed and the mean synonymous/nonsynonymous ratio for multiple sequences analysed from single individuals was found to be3.86. For comparative purposes, gp120 from chimaeric genomes used to express recombinant proteins were sequenced.

CONTENTS

LIST OF FIGURES

Figure **Title**

LIST OF TABLES

ABBREVIATIONS

AIDS	acquired immune deficiency syndrome
bp	base pairs
CD4$^+$	cluster of differentiation 4
DNA	deoxyribonucleic acid
env	envelope
ER	endoplasmic reticulum
gag	group-specific antigen
HIV	human immunodeficiency virus
HMA	heteroduplex mobility assay
kb	kilobase
kd	kilodalton
KS	Kaposi's sarcoma
LAS	lymphadenopathy syndrome
LTR	long terminal repeat
nef	negative regulatory factor
ORF	open reading frame
PBMC	peripheral blood mononuclear cells
PCP	*Pneumocystis carinii* pneumonia
PCR	polymerase chain reaction
pol	polymerase
RER	rough endoplasmic reticulum
rev	regulator of virion protein
RNA	ribonucleic acid
RRE	*rev*-responsive element
RT	reverse transcriptase
TAR	*trans*-activation element
tat	transactivator
vif	viral infectivity factor
vpr	viral protein R
vpu	viral protein U

ACKNOWLEDGMENTS

I would like to express my sincere thanks to my two supervisors, Jon Clewley and Peter Balfe, who were always there to give affable encouragement and first rate advice.

I'm also glad for an opportunity to be able to express my devotion to those closest to me for being themselves but especially to my Mum and my unborn child, without whom this thesis would never have been started, or finished.

CHAPTER 1

INTRODUCTION

1.1 General introduction

In 1981 reports were published describing a new syndrome, characterised by severe immune dysfunction, resulting in infection with unusual opportunistic pathogens and vulnerability to rare lymphomas (94, 169). Researchers investigating the causative agent of this so called Gay Related Immune Deficiency (GRID), subsequently defined as AIDS, initially focused their attention on several different viruses, including parvoviruses and herpesviruses, as well as retroviruses which are known to cause immune deficiency (218). In 1983 Barré-Sinoussi *et al.* at the Pasteur Institute in Paris (17) discovered a reverse-transcriptase-containing virus from a man with persistent lymphadenopathy syndrome (LAS), a syndrome that some physicians suspected was associated with AIDS. This was the first indication that AIDS could be caused by a retrovirus. At around the same time Gallo *etal* (85) also identified a retrovirus as the probable cause of AIDS. However, this virus subsequently turned out to be a contaminant by the French groups' isolate. The virus, initially named lymphadenopathy associated virus (LAV) by the French group, and human T-cell leukaemia virus type III (HTLV-III) by the American group, is now known as human immunodeficiency virus or HIV-1 (43).

HIV is a lentivirus, part of a separate genus of the *Retroviridae* family which infects human CD4$^+$ T lymphocytes and monocytes *in vivo* (65, 165). Other members of the lentiviridae family include equine infectious anaemia virus (EIAV), caprine arthritis-encephalitis virus (CAEV), feline immunodeficiency virus (FIV) and bovine immunodeficiency virus (BIV) and Visna virus. These viruses typically display long periods of latent or persistent infection in some infected cells. A second closely related virus (HIV-2) has been isolated which also causes AIDS (40).

Several days to weeks after infection with HIV-1 acute influenza-like symptoms develop (46), usually followed by a latent phase which may last for many years. After seroconversion there is a gradual decline in the number of CD4$^+$ T-cells. Initially the prolonged clinical latent phase characteristic of HIV-1 infection was thought to indicate a period of viral inactivity. However, recent reports suggest that this is not the case and that the latent phase is, in fact, a dynamic process in which cells are being infected and are dying in vast numbers (113, 274). The loss of CD4$^+$ cells, which play a central role in the immune response, accounts for most of the afflictions observed in AIDS. The onset of AIDS is indicated by various clinical signs including weight loss, general malaise, chronic fever, diarrhoea, hairy leukoplakia, oral thrush, *Pneumocystis carinii* pneumonia (PCP) and other opportunistic infections. These signs may appear separately, concurrently or successively. Patients with two or more of these signs will often die within 1-2 years unless treated with antiviral drugs.

As described above, AIDS was first observed in individuals infected with unusual opportunistic pathogens and who developed rare malignancies. In gay men, the two most commonly observed signs for the onset of AIDS are PCP and Kaposi's sarcoma (KS). In developing countries such AIDS-defining illnesses are used to diagnose HIV-1 infection in the absence of serological tests. PCP accounted for more than 50% of all the initial AIDS diagnoses, while KS occurs in 20-40% of AIDS cases. Another sign of AIDS which has been used in diagnosis is oral candidiasis, which is observed in over 80% of patients with AIDS. The development of oral candidiasis often heralds the impending succession of other more serious opportunistic infections, such as *Mycobacterium tuberculosis*, toxoplasmosis, cryptosporidiosis, and viral infections such asHerpes simplex, Cytomegalovirus, andVaricella zoster.

It has been calculated that 13 million people worldwide are infected with HIV-1 (278). The same report also estimated that a new infection takes place every 13 seconds and that a person dies from HIV-1 associated illness every 9 minutes. The WHO predict that by the year 2000, 30-100 million people worldwide will have been infected by HIV-1 or HIV-2.

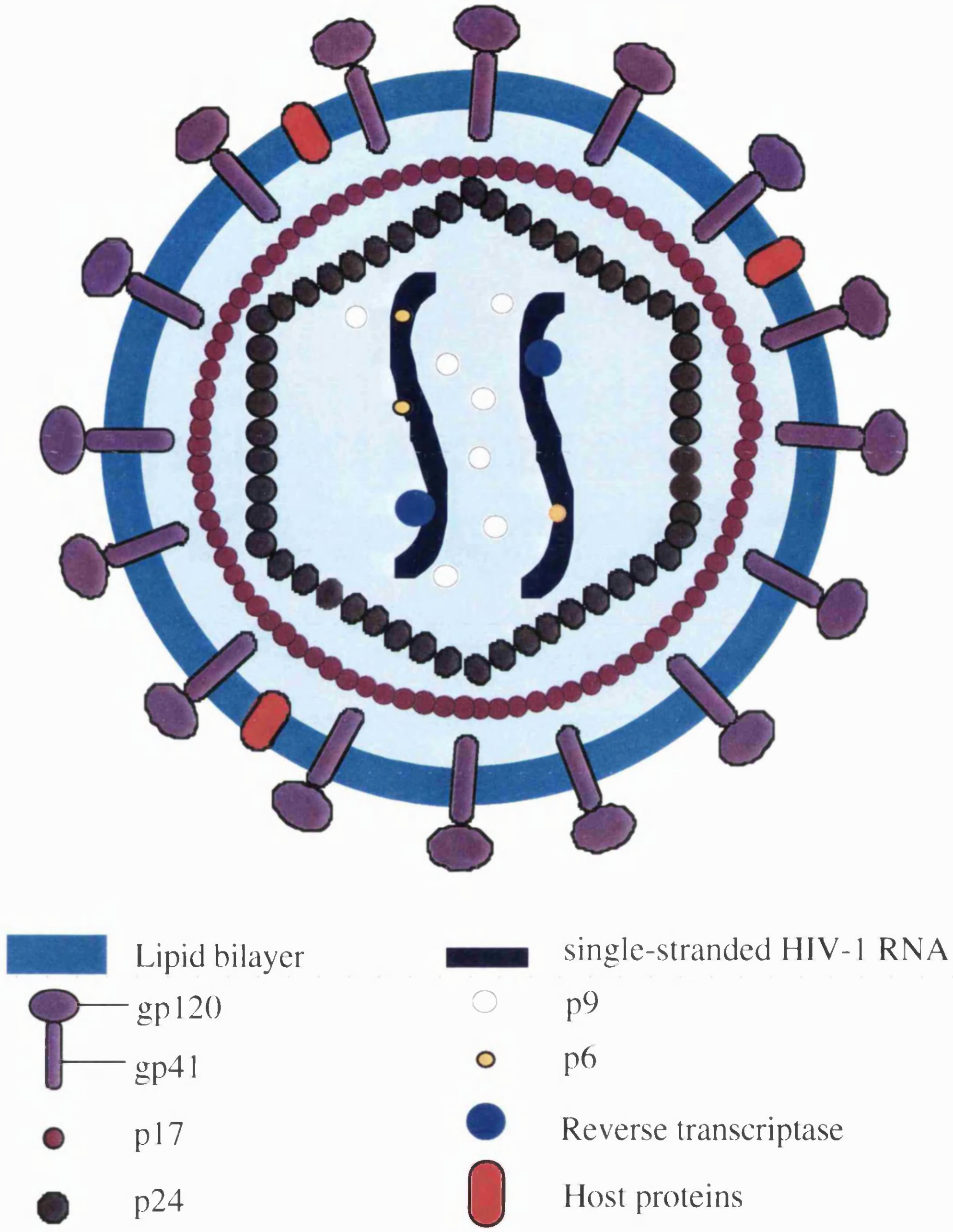
1.2 The HIV-1 virion

Structure

High-resolution electron microscopy has shown the HIV-1 virion to be an icosahedral structure (87) containing 72 external spikes made up of trimers or tetramers of the envelope glycoproteins gp120 and gp41 (see Figure 1.1) (87, 201, 209, 275). These spikes are derived from gp160 which is cleaved inside the cell into a gp120 external surface (SU) envelope protein and a gp41 transmembrane (TM) protein (170). The central region of the TM glycoprotein binds to the external viral gp120 in a noncovalent manner, probably at two hydrophobic regions in the amino and carboxy termini of gp120 (106). Some analyses suggest that a ball-and-socket type of structure is involved (229), and that several envelope regions, including V1/V2 and the C2 and V3 domains of gp120, help stabilise the association (254, 283). The gp120 contains the binding site for the cellular receptor(s) and the major neutralising domains. The HIV-1 lipid bilayer is also studded with various cell derived host proteins, including ß-2 microglobulin and Class I and Class II histocompatibility antigens acquired during virion budding (12). The core of HIV-1 is cone-shaped, characteristic of lentiviruses, and contains four nucleocapsid proteins, p24, p17, p9, and p6, each of which is proteolytically cleaved from a 53 kilodalton (kd) Gag precursor by the HIV-1 protease. Inside this capsid (CA) are two identical RNA strands with which the viral RNA-dependent DNA polymerase (Pol, also called reverse transcriptase or RT) and the nucleocapsid (NC) proteins (p6, p9) are closely associated. The core also contains integrase (IN) and protease (PR). The inner portion of the viral membrane is lined by a myristoylated p17 core (Gag) protein that provides the matrix (MA) for the viral structure and is crucial for the integrity of the virion (87, 88).

Figure 1.1

A schematic diagram of the HIV-1 virion

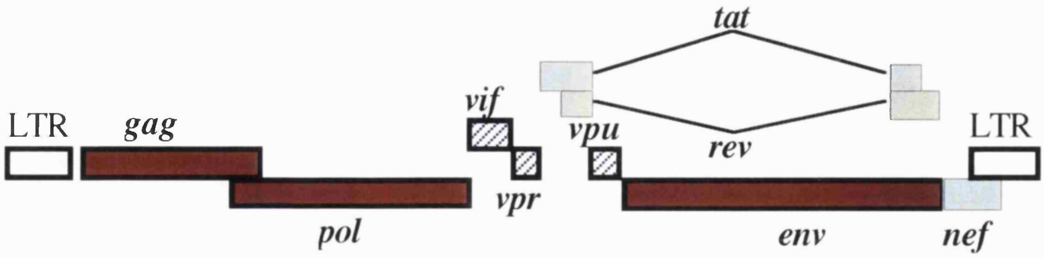


Genomic Organisation

The HIV-1 genome is a positive sense, single-stranded polyadenylated RNA of approximately 9.8 kilobases (kb) with open reading frames (ORFs) coding for several viral proteins (see Table 1.1 and Figure 1.1). The HIV-1 genome is flanked at both the 5' and 3' ends by long terminal repeat sequences (LTRs). The LTRs are non-coding regions and consist of regulatory elements for integration, transcription and polyadenylation of mRNAs (266). The primary transcript of HIV-1 is a full-length viral mRNA, which is translated into the Gag and Pol proteins. The Gag and the frame shift Gag-Pol products are synthesised in a ratio of about 20:1 (128, 198). The envelope proteins gp120 and gp41 are made from a precursor gp160 which is a singly spliced message from the full-length viral mRNA. Further splicing events subsequently produce smaller mRNAs important for the synthesis of other proteins. The relative amounts of the unspliced to singly and multiply spliced mRNAs appears to be determined by the *rev* gene, which is itself a product of a multiply spliced mRNA (66, 243, 247). Other spliced mRNAs are translated to a variety of viral regulatory and accessory proteins that can affect other aspects of HIV-1 replication (see below). *Tat, rev* and *nef* are involved in the regulation of viral replication and *vpu* and *vif* are involved in virion maturation and morphogenesis.

Figure 1.2

Genomic map of HIV-1



Function

The structural genes of HIV-1

The Gag precursor p53 gives rise by proteolytic cleavage to the smaller proteins p17, p24, p9, and p6. Some studies suggest that during maturation, the p53 protein is cleaved into six *gag* gene products: p17, p24, p2, p7, p1, p6 (107) though this is not universally accepted. The phosphorylated p24 polypeptide forms the chief component of the inner core of the nucleocapsid (see Figure 1.1) and the myristoylated p17 protein is associated with the inner surface of the lipid bilayer and probably stabilises the exterior and interior components of the virion. The p9 and p6 together form the nucleoid core. Various other functions have been proposed for HIV-1 p6 protein. It appears to be required for the incorporation of the HIV-1 accessory protein Vpr into virus particles (205), and it has been reported that p6 plays a role in virus particle production at a late stage in the budding process (95). Huang and colleagues confirmed a role for p6 late in the virus assembly process and defined a motif in the p6 required for virion production (Pro-Thr-Ala-Pro, located between residues 7 and 10) (122). Part of the work for this thesis (see chapter 3, Figure 3.27), involves subtyping and transmission studies of U.K. isolates using p6/protease sequences and confirms the presence of this motif in all specimens investigated. Huang *et al.* also demonstrated that mutational inactivation of the viral protease reverses the effect of p6 mutations on virus particle production, suggesting a link between p6 and protease functions during the budding of progeny virions.

The Pol precursor protein is cleaved into products consisting of the reverse transcriptase, the protease, and the integrase proteins. The reverse transcriptase enzyme has two functions: it can act as an RNA dependent DNA polymerase, and it has RNAase H activity required for the degradation of the RNA template during the synthesis of the double-stranded (proviral) DNA (266). The aspartyl protease processes the Gag and Pol polyproteins and is essential for the formation of infectious virions of HIV-1 (143). Recently the HIV-1 protease has become an important target for the design of antiretroviral drugs because of the essential role it plays in viral replication and the lack of an endogenous human counterpart. Winslow *et al.* sequenced protease genes from 24 isolates of HIV-1 and observed a maximum sequence variation of 10% at both the nucleic and amino acid levels (284). The degree of sequence variation in this region of the genome is reduced because only a limited repertoire of mutations can be tolerated by the enzyme whilst maintaining its activity in Gag-Pol processing (160). Data from other sequenced isolates confirm that major regions of the protease gene are highly conserved (77). The integrase is involved in catalysing the integration of proviral DNA into the host genome. This enzyme has three main activities: i) cleavage of the blunt-ended termini of the linear viral DNA; ii) cleavage of the host genomic DNA; iii) ligation of the ends of the proviral DNA with the ends of the host genomic DNA (204).

The envelope glycoproteins of HIV-1, gp120 and gp41, are derived from a 160 kd polypeptide which is modified, as with other glycoproteins destined for the plasma membrane, in the rough endoplasmic reticulum (RER) by the addition of 24 N-(asparagine) linked carbohydrate moieties, giving rise to a precursor polypeptide of 160 kd inserted into the lumen of the endoplasmic reticulum (ER) (80). Shortly after synthesis, gp160 monomers oligomerise (209), a process which is thought to be required for transport from the ER to the Golgi complex (281). Once in the Golgi, some of the high mannose, ER-acquired N-linked oligosaccharide side chains are modified to more complex forms, and gp160 is proteolytically cleaved to gp120 and gp41. The HIV-1 Env glycoprotein is extensively glycosylated; approximately half the molecular mass of gp120 being composed of oligosaccharides (6). Following gp160 cleavage, the oligomeric, noncovalently associated gp120-41 complexes are transported to the cell surface, where they are incorporated into budding virions.

The regulatory genes of HIV-1

HIV-1 is called a complex retrovirus because in addition to the Gag, Pol and Env genes common to all retroviruses it also has other ORFs which code for proteins that regulate the expression and replication of the virus. These are *tat* and *rev* (regulatory genes) and *vpr*, *vpu*, *vif* and *nef* (accessory genes). The *tat* gene codes for a strong transactivator protein which binds a *cis*-acting target sequence designated the *trans*-activator response element (TAR) located in the 3' portion of the LTR. This interaction is capable of increasing HIV-1 LTR-dependent gene expression by more than 100-fold (242). The 14 kd Tat polyprotein is translated from a doubly spliced viral mRNA species and is composed of 86 amino acids. The tat protein appears to contain three primary structural domains, including a proline-rich N-terminus, a cysteine-rich central portion, and a positively charged distal segment (220). The cysteine-rich domain probably mediates dimerisation of this transactivator (78), whereas the positively charged distal segment is responsible for both RNA binding and nuclear-nucleolar localisation (103).

Rev is a 19 kd protein which interacts with a *cis*-acting 200 nucleotide RNA loop structure called the Rev responsive element or RRE, located in the env mRNA (66, 243). This interaction involves cellular proteins and multimers of the Rev protein, and allows unspliced and partially spliced mRNA to enter the cytoplasm from the nucleus and give rise to the full-length viral proteins required for virus production. Normally these unspliced transcripts are excluded from the cytoplasm and thus their corresponding proteins are not synthesised. In the absence of Rev, only the fully spliced (2 kb) class of HIV-1 mRNAs is expressed (220). Rev deletion mutants of HIV have been shown to be incapable of inducing the synthesis of viral structural proteins (259). Rev functions as a negative regulator of its own production as well as controlling the expression of Tat mRNA. This results in the establishment of an equilibrium between viral structural and regulatory protein synthesis, which is probably required for efficient virus production (166).

Presumably, the balance of these viral gene products can determine whether HIV-1 infection will lead to a productive or latent state. High-level expression of Tat will activate substantial virus production (50). In contrast, expression of Nef could induce a latent state (164), depending on the effect of this viral protein on the intracellular environment. In 1992, Learmont *et al.* reported on long-term symptomless HIV-1 infection in six recipients of blood products from a single donor (151). A recent update regarding this investigation (150) analysed the HLA tissue types of the symptomless group and suggested that sharing of HLA alleles or haplotypes did not explain long-term nonprogression in this cohort as suggested by other workers. Recently, Huang and colleagues investigated the possibility that deletions in Nef resulted in long-term survival after HIV-1 infection (123). They found no gross deletion within Nef in the cases studied corresponding to long-term survival. In contrast, Deacon *et al.* reported that long-term survival after HIV-1 infection can be determined by deletions in Nef and the U3 region of the LTR, underlining the importance of Nef (or the U3 region of the LTR) in determining the pathogenicity of HIV-1 (52)

The accessory genes of HIV-1

Vif, Vpu, Vpr, and Nef have been shown to be dispensable or nonessential for *in vitro* replication and are usually referred to as accessory gene products (96). Nef appears to have several functions, including down-regulation of viral expression (1) and enhancement of the viral replication rate in culture (180). These observations have been confirmed *in* *vivo* by Kestler and colleagues (137) who showed that Nef was critical to the maintenance of high virus loads, and for the development of AIDS in rhesus monkeys. *

Vpr is thought to prevent proliferation of chronically infected cells (215) and is important for infection of macrophages by HIV-1 (45). The ability of HIV to infect terminally differentiated macrophages seems unusual because of the inability of other retroviruses such as avian and murine retroviruses to integrate into nonproliferating cells. It has been reported that viral strains of HIV-1 replicate in macrophages and stimulate peripheral blood lymphocytes to high levels with equal kinetics (105), the efficient growth of HIV-1 in macrophages cannot be explained by proliferation of a minor population of cells in the macrophage cultures. Work by Lewis and colleagues suggested that the ability of HIV-1 to infect nondividing cells was due to Vpr preventing mitosis and they suggested that HIV-1 has two distinct roles: early in the viral replication cycle Vpr acts by directing the preintegration complex to the nucleus and at the same time preventing infected cells from passing into mitosis (157, 158). Vpx, an accessory protein found in HIV-2 and some simian immunodeficiency viruses, is structurally very similar to Vpr. Tristem and colleagues suggested that *vpx* arose by a gene duplication of *vpr* (263). *In vivo, vpx* may be important for HIV-2/SIV persistence by facilitating infection and spread during the natural course of infection (263).

After the initial description of *vif* and the finding that its influence on the infectivity of progeny virions was found to be dispensable for virus infection of transformed T cell lines (71, 241, 250), little research was done to further the understanding of the function of this protein. Recently, however, several laboratories made the observation that *vif* is essential for HIV-1 infection in its primary target cells *in vivo*, CD4-bearing T lymphocytes (64, 272).

The Vpu protein appears to have roles in two unrelated functions:virus release (64, 272) and down regulation of CD4 after proteolysis in the endoplasmic reticulum (271, 282). The degradation of CD4 is sequence specific , and it is also compartment specific (ER). The other activity of Vpu, the release of virus particles, is independent of the ER degradation reaction (271, 282).

*Several groups have demonstrated that the *nef* gene is capable of downregulating cell surface CD4 molecules by the active sequestration of cell surface CD4 molecules in lysosomes, where they are subsequently degraded (293, 294, 295, 296, 297).

Table 1.1

HIV-1 proteins and their functions

Protein	Size (kd)	Function
Gag	p25(p24)	Capsid structural protein
	p17	Matrix protein - myristoylated
	p9	RNA binding protein (?)
	p6	RNA binding protein (?); incorporation of Vpr into virus particles; helps in virus budding.
Polymerase (Pol)	p66, p51	Reverse transcriptase; RNAse H - inside core
Protease	p10	Post-translation processing of viral proteins
Integrase	p32	Viral cDNA integration
Envelope	gp120	Envelope surface protein
	gp41	Envelope transmembrane protein
Tat[1]	p14	Transactivation
Rev[1]	p19	Regulation of viral mRNA expression
Nef[1]	p27	Pleiotropic, including virus suppression, myristoylated
Vif[1]	p23	Increases virus infectivity and cell-to-cell transmission; helps in proviral DNA synthesis and/or in virion assembly
Vpr	p15	transactivation
Vpu[1]	p16	Helps in virus release; disrupts gp160-CD4 complexes

See Figure 1.2 for location of the viral genes on the HIV-1 genome.

[1]Not found associated with the virion.

1.3 The HIV-1 replicative cycle

Virus attachment

The replicative cycle of HIV-1 is outlined in figure 1.3. The binding of viruses to specific viral receptors or attachment proteins on the cell surface is the first step in the majority of virus infections and one of the major determinants of virus host range and tissue tropism. The human CD4$^+$ T lymphocyte and monocyte are the major cellular targets for HIV-1 infection *in vivo* (65, 165). Maddon *et al.* provided conclusive evidence that CD4 is the receptor for HIV-1 by showing that human cells which normally lack CD4 are resistant to HIV-1 infection, but become susceptible when transfected with a vector encoding CD4. CD4$^+$ T cell killing induced by HIV-1 underlies the severe immunodeficiency characteristic of advanced AIDS (65). HIV-1 may also infect other types of cells, including glial cells, gut epithelium, and bone marrow progenitors (34), associated respectively with dementia, diarrhoea-wasting syndrome, and haematological abnormalities. The binding of virion-associated gp120 to CD4$^+$, the first step in HIV-1 infection, and membrane fusion are discussed in more detail in section 1.5.

Internalisation of the virion

CD4$^+$ receptor-bound HIV-1 virions are brought inside the cell one of two ways, either by classic receptor-mediated endocytosis (165), or possibly by virus-mediated membrane fusion (248). Membrane fusion involves gp120 (discussed in further detail below) and the gp41 envelope protein (270). The gp41 polypeptide, which is noncovalently associated with gp120, contains a region anchoring the envelope protein in the lipid bilayer and a fusogenic domain resembling the F proteins of the paramyxoviruses such as mumps and parainfluenza (214). Fusion of the viral and host-cell membranes occurs following activation of this domain promoting internalisation of the virion.

Reverse transcription and integration

The genetic material in an HIV-1 virion, similar to all lentiviruses, consists of a complex of two genomes of positive (mRNA) polarity (237). After internalisation of the virion, replication begins with the generation of a DNA copy containing two LTRs of the viral RNA mediated by the HIV-1 reverse transcriptase (Figure 1.3). Each LTR consists of a short repeat sequence (R) flanked by longer sequences that are derived from the 5' (U5) and 3' (U3) noncoding ends of the viral RNA. The cDNA or minus strand of the linear molecule is synthesised in a continuous manner from the tRNA primer while the plus strand is synthesised in a discontinuous manner (after partial degradation of the original RNA template by RNAse H) using multiple initiation sites (22, 120, 147). Subsequently, two forms of unintegrated covalently closed circular viral DNA molecules are detectable in the nucleus of the cell. One circular form contains tandem copies of the LTR and the second form contains a single LTR. It has been proposed that this form arises via a homologous recombination between two LTRs of a linear or circular molecule (267) or by premature repair and ligation of a circular replicative intermediate (132).

After translocation to the nucleus, the viral cDNA is inserted into the host genome by the viral integrase. The integration of a provirus into the DNA of the host cell is a recombination event whereby sequences near the ends of the LTRs in the viral DNA are joined to the host DNA in the reaction. Numerous sites in the host DNA can serve as targets for integration (237). Some viral DNA sequences, usually 2 base pairs from the end of the LTR, are lost during the integration reaction and there are sometimes small duplications of host cell DNA at either end of the proviruses (267).

Antiviral therapy against reverse transcriptase

Having no host counterpart, HIV-1 reverse transcriptase has been the target for most antiviral therapies to date including nucleoside analogues such as: zidovudine (AZT, 3'-azido-2',3'-dideoxythymidine); 3TC ((-)2'-deoxy-3'-thiacytidine); ddC (2',3'-dideoxycytidine); and ddI (2',3'-dideoxyinosine) (96). Each analogue has to be phosphorylated to an active triphosphate form following entry into the cell. They act either as chain terminators of the newly formed DNA strand produced by the reverse transcriptase or as competitors blocking the incorporation of the respective normal deoxynucleoside 5' triphosphate. Initially it was thought AZT substantially prolonged survival and reduced the frequency and severity of opportunistic infections in patients with AIDS, but this early indication of clinical benefit was only transient (70). The recent Concorde trial (44) failed to demonstrate any long term improvement in disease progression or survival when antiretroviral therapy was initiated in asymptomatic individuals rather than deferring it until the development of AIDS. The recent Delta trial (39), comparing combinations of zidovudine (AZT) with didanosine (ddI) or zalcitibane (ddC) against AZT alone in patients with HIV-1 infection was halted because routine monitoring of the data showed superiority of dual therapy in those naive to AZT. The on-going quattro trial, which includes a protease inhibitor as well as chain terminating nucleoside analogues, represents the first of a series of trials in which multiple (in quattro, 4) drugs, targeted to more than one protein, are employed.

Viral latency

Schnittman *et al.* reported that approximately 1 in 1000 peripheral-blood CD4$^+$ T cells from patients with AIDS express RNA (226). However, approximately 1 in 100 CD4$^+$ T cells contains detectable HIV-1 DNA. This apparent viral latency may be explained by the overall state of cellular activation. HIV-1 does not replicate in resting T cells, presumably because host factors are absent (47). Activation of these cells by antigens, mitogens, tumour necrosis factor (TNF), interleukin-1, or various gene products of different viruses, creates a permissive cellular environment that promotes a high level of HIV-1 replication (217). NF-kß proteins induced by these activating agents, which normally regulate the expression of various T-cell genes involved in growth, bind to and activate the duplicated kß enhancer element present in the U3 region of the LTR (190). This induced expression of NF-kß probably has a role in the initial stimulation of latent or persistent proviral forms of HIV-1. However, recent reports suggest that this prolonged clinical latent phase is, in fact, a dynamic process in which cells are being infected and dying in vast numbers (113, 274). Thus, the precise sequence of events during this asymptomatic phase of HIV-1 infection is yet to be determined.

Early expression of regulatory genes

During a single cycle of infection the HIV-1 genome is transcribed and processed into three classes of viral mRNA species (96, 138). The earliest mRNA species made in the infected cell are 2 kb doubly spliced mRNA transcripts that encode the major regulatory proteins Tat, Rev and Nef. The second and third classes of transcripts include the unspliced 9 kb and singly spliced viral mRNAs (4.5 kb) that encode the viral structural proteins Gag, Pol, and Env and the accessory proteins Vif, Vpr, and Vpu.

Late expression of structural and enzymatic genes

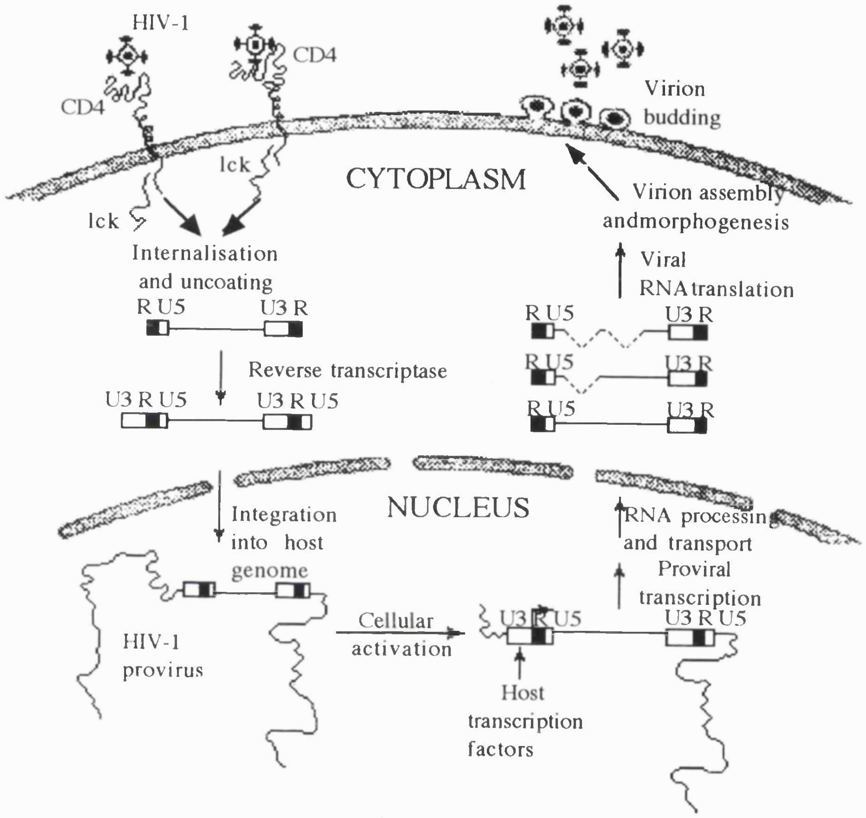
As noted above, after the first phase of expression, the second and third (or 'late') phases of expression of the HIV-1 genome is characterised by the cytoplasmic synthesis of the structural proteins that make up the HIV-1 virion and the accessory proteins (96). The transition between the synthesis of early and late gene products appears to be dependent on Rev (166). Rev achieves this by activating the cytoplasmic expression of the unspliced and singly spliced forms of mRNA that encode the products of *gag, pol,* and *env* genes, either by activating the nuclear export of the incompletely processed viral RNAs (167), or by disengaging the long mRNAs from the nuclear splicing apparatus (35).

Assembly of the HIV-1 virion

The Gag and Pol proteins form the core of the mature virion and the products of the *env* gene are the principal exterior-coat proteins. The Pol protein is translated from the same RNA transcript as the Gag precursor by a ribosomal frame-shifting mechanism (128), and is produced at approximately 20 times lower levels. The assembly of HIV virions is a process involving a series of events taking place close to the cell membrane. The core structure is formed by specific interactions of the two newly synthesised RNA genomes with the p55 Gag and p160 Gag-Pol polyprotein precursor molecules. Both the p55 and p160 molecules are post-translationally modified by covalent attachment of a myristoyl group. This modification is thought to aid in anchoring the polyproteins and their associated RNA genomes to the cytoplasmic side of the cell membrane. Consequently there is a marked increase in the local concentration of viral proteins and RNA leading to assembly and budding of the immature virion. During the budding process, gp120 and gp41 Env glycoproteins are captured along with the lipid bilayer from the cell membrane. The viral structural proteins are capable of spontaneous self-assembly. Formation of mature infectious HIV particles depends on separation of the p55 Gag and p160 Gag-Pol polyproteins. This is achieved by the protease domain of the Gag-Pol precursor. the protease acts first to cleave itself from the p160 Gag-Pol precursor, and then cleaves at seven further sites on the p55 and p160 precursors. The Vpu protein acts at around this time to promote efficient release of the virions from the surface of the cell (141). Vif appears to be necessary for full infectivity of the released virions (241, 250).

Figure 1.3

Replicative cycle of HIV-1 (taken from (96))



1.4 Variability of HIV-1

Genetic variability

The mutation rate of a virus depends on the number of replication cycles, the growth rate of the viral population and the fidelity of the viral polymerases. HIV-1 reverse transcriptase lacks proofreading activity and as such is error-prone, with estimates of a rate of up to 10 base changes per genome per round of replication (211). The rate of misincorporation is not uniform across the whole genome. Transitions between A and G occur frequently (especially G>A) when the template adenine residue is preceded by the doublets AT, TT, CT or AC and are suppressed when the preceding doublets are NG, GT, YC or GC (183, 213). The evolution rate, or the rate of fixation of these mutations, is dependent on various influences. These include: the mutation rate; positive selection for other environmental conditions such as host immune pressure and anti-viral drugs; and negative selection against variation imposed by the functional constraints of the virus. Hahn and Shaw and co-workers were the first to describe the genomic diversity of HIV-1 (100, 101, 231, 232) and found substantial heterogeneity, concentrated in *env*, among unlinked HIV-1 isolates. Subsequent studies revealed that isolates obtained from Zairian patients (ELI and MAL) demonstrated a much greater sequence divergence from the European/North American isolates (5). Moreover, the divergence between ELI/MAL was similar to the divergence between ELI/LAV and MAL/LAV, suggesting a longer evolution of the virus in Africa. All these studies found that the envelope gene exhibits a higher degree of variability than Gag and Pol, with amino acid sequences differing by up to 30% between different isolates (280). By comparing the sequence of different isolates of HIV-1, five domains of hypervariability (V1-V5) have been identified (see figure 2.3), interspersed with five conserved domains (C1-C5) (181, 246, 280). Saag and colleagues extended the work of Hahn *et al.* (102), and showed that intrapatient HIV-1 sequence variation was also extensive (219). The observation that HIV-1 isolates consisted of a complex mixture of genotypically distinguishable viruses was confirmed by Fisher and colleagues (72). In 1989 Goodenow used the term ‘quasispecies’ to describe the populations of closely related species (93). The quasispecies concept was introduced by Eigen (60).The Darwinian theory of natural selection was the basis for this concept. HIV-1 replicates with limited fidelity generating diversity and consequent phenotypic change. Goodenow showed that the complexity of HIV-1 quasispecies shows no correlation with the stage of disease (93).

Meyerhans and colleagues showed that this *in vivo* diversity is lost *in vitro* by the selective outgrowth of HIV-1 strains capable of rapid replication (177). This effect was subsequently observed by others (148, 168, 203, 268) and lead to the term 'to culture is to disturb' (93). In recognition of the observation that cultured virus is unrepresentative of that found in the patient the work described in this thesis is solely on uncultured virally infected cells.

Sequence variation of HIV-1 isolates between individuals infected from a common source is less heterogeneous than that between epidemiologically unlinked isolates. This observation has been the basis for various transmission investigations (4, 9, 14, 119, 129, 199). There are several reports indicating that within an individual, there are differences in the frequency and distribution of sequence variants in brain tissue versus blood cells (61, 203), indicating possible tissue-specific evolution of the viral quasispecies. Simmonds *et al.* demonstrated differences within the peripheral blood compartment between viral RNA from cell-free virus circulating in plasma and proviral DNA found in peripheral blood mononuclear cells (PBMCs) (236). They also found that new variants initially appeared in the plasma RNA populations and subsequently became detectable in the PBMCs (236).

Possible recombination adds further to the intricacy of genetic variation. Two models for retrovirus recombination have been suggested. Coffin proposed a switch of template during minus-strand generation (42). Junghans proposed a crossing over of the DNA plus-strand onto two DNA minus-strand copies (133), resulting in a heterozygous provirus. Moreover, Hu *et al.* proposed that such a heteroduplex is always repaired before integration (121). The detection of such recombinant viruses has been the subject of several recent studies, see below.

Phenotypic variability

Early studies of HIV-1 (7, 13, 40, 63, 69, 256) indicated that it is biologically and antigenically heterogeneous as well as genomically. These features of viral heterogeneity could influence disease and express themselves as characteristics such as: cellular tropism (13); replication kinetics/level of virus production , also defined as 'slow/low' (slow growing to low titres),or 'rapid/high' (rapid growth to high titres) (13); cytopathicity (7); syncytium-forming ability (256); sensitivity to neutralising or enhancing antibodies (36); genetic structure (101)

HIV-1 isolates recovered from individuals with AIDS showed higher replication rates than those recovered from asymptomatic individuals and were able to establish persistent infection in T-cell lines (13). The third biological property of isolates, found more frequently in advanced disease, was the ability to form syncytia (multinucleated cells resulting from the fusion of infected cells with uninfected CD4$^+$ cells) in primary PBMC cultures (69, 256). Several studies found that the appearance of syncytium inducing (SI) isolates generally correlates with disease progression (37, 257, 258). Tersmette and colleagues also found a strict correlation between the ability to induce syncytia in PBMC cultures and the ability to infect T-cell lines continuously and productively (257, 258). Conversely, non-syncytium inducing (NSI) isolates appeared to be much more monocytotropic than SI isolates and were not transmissible to T-cell lines (176, 228). Cells expressing low levels of CD4$^+$ receptors appear less susceptible to this form of cell death, a finding that probably explains the relative rarity of cytopathic effects produced by HIV-1 in monocytes and macrophages (96).

Antigenic or serologic variation of HIV-1 has also been demonstrated. Cheng-Meyer *et al.* reported a distinct sensitivity to serum neutralisation among HIV-1 isolates (36). Studies by McKeating and others showed that neutralising antibodies generated neutralisation resistant viruses *in vitro* (174, 212). Albert and colleagues showed viruses isolated early in infection were neutralised by both early and late sera. However, viruses isolated more than six months after primary infection were neutralised inadequately or not at all (2).

Moore *et al.* observed that inter- and intraclade neutralisation of primary isolates by HIV-1 positive sera was sporadic, with only some indication of clade-specific binding (299). The ability of sera to cross-neutralise across two or more genetic subtypes described by Kostrikis and colleagues (300) was confirmed by Weber *et al.* (301), leading to the conclusion that genetic subtypes of HIV-1 are not classical neutralisation serotypes.

1.5 The role of HIV-1 envelope glycoproteins in virus infection

CD4 binding

The first step in HIV-1 infection involves the binding of virion-associated gp120 to the cell surface molecule CD4, the major receptor for HIV-1 and 2 (49, 140, 165). CD4 binding to gp120, in particular, C3 and C4, induces conformational changes in both gp120 and gp41 that result in the exposure of Env domains thought to be involved in the membrane fusion reaction (135, 223, 224, 260). Once CD4 had been identified as the primary receptor for HIV-1 experiments involving soluble CD4 (sCD4) blocking were carried out to determine whether it could neutralise virus infectivity and be used for possible antiviral therapy (54, 73, 124, 239, 262). However, primary, non-laboratory-adapted isolates are neutralised poorly by soluble CD4 (48, 92), thereby diminishing the utility of sCD4 as a therapeutic agent. Env also associates with CD4 intracellularly soon after gp160 synthesis in the ER. This association of Env and CD4 early in the transport pathway leads to the down regulation of CD4 expression on the surface of Env-expressing cells (127). This decreases the level of cell surface CD4, reducing the ability of Env-expressing cells to become infected with additional virus (249).

Membrane fusion

As well as the ability to induce fusion between the lipid bilayer of the viral envelope and host cell membranes, Env expression in an infected cell can also lead to cell-to-cell fusion, or syncytium formation (described above) (256). A number of determinants in both gp120 and gp41 have been proposed to play a role in membrane fusion. When the Env glycoprotein undergoes conformational changes following CD4 binding a gp41 fusion peptide is exposed (224). Several fusion peptides may act together to destabilise the lipid bilayer of the target membrane by forming a 'fusion pore' between the two bilayers (277). In gp120 two regions appear to be involved in membrane fusion. Several studies have determined that antibodies directed to the V3 region were able to neutralise virus infectivity without affecting virus binding to CD4, for examples see (159, 238). Mutational analyses showed that single amino acid substitutions with the V3 loop blocked Env-induced syncytium formation (81-83) and virus infectivity (81). The V1/V2 region of gp120 has also been implicated in membrane fusion. Sullivan *et al.* showed that mutations within V1/V2 appeared to block syncytium formation without affecting the gp41-gp120 interaction or CD4 binding (254) and transfer of V2 sequences from syncytium-inducing Env glycoproteins conferred the ability to induce fusion on non-syncytium inducing Env glycoproteins (98). Work by McKeating *et al.* supported the idea of a role for V1/V2 in membrane fusion by demonstrating the neutralisation of viral infectivity using antibodies directed to this region. (174). Molecules other than CD4 have been shown to be necessary for membrane fusion induced by HIV-1 Env, although various suggestions, including a role for CD26 (32), have not received universal acceptance. Recent work by Feng and colleagues (68) may have identified the long-sought cofactor, designated 'fusin', using a novel functional cDNA cloning strategy. They showed that recombinant fusin enabled CD4-expressing nonhuman cell types to support HIV-1 Env-mediated cell fusion and HIV-1 infection. Moreover, antibodies to fusin blocked cell fusion and infection with normal CD4-positive human target cells.

Tissue tropism

A number of studies have concluded that sequences within gp120 are responsible for determining the tissue tropism of HIV-1. Macrophage tropism can be conferred upon T-cell line-tropic clones by the introduction of sequences from the V3 loop of macrophage tropic clones (38, 125, 196, 233). It is likely that a combination of sequences within and outside the V3 loop is required for optimal macrophage infection. Exchanging additional sequences adjacent to V3 enhances the ability of chimeric viruses to infect macrophages (276). Westervelt and colleagues also showed that the V3 loop conformation differs between macrophage tropic and T-cell line-tropic Env glycoproteins and that residues within and outside V3 influence conformation (276).

Env interactions

Although current attempts to obtain a crystal structure of HIV-1 Env have been unsuccessful, other data have provided information about Env interactions. Various groups have used antibody binding analysis to identify interacting regions within gp120. This approach has provided evidence for interactions between V1/V2 and C4; V3 and C1, C2 and C4; and C1, C2 and C5 (173, 174, 182, 208, 289). The work of McKeating *et al.* further supported the possibility of an interaction between V3 and C4 by the observation that treatment of gp120 with soluble CD4 enhances the binding of anti-V3 monoclonal antibodies (173).

1.6 HIV-1 Subtypes

The extensive biological heterogeneity of HIV-1 strains, introduced in section 1.4, is reflected in the genetic sequences of the virus. As discussed above, the diversity of HIV-1 arises because the viral reverse transcriptase is very error-prone. Also, on transcription to DNA, recombination between proviral DNA molecules may occur, providing another mechanism for variation.

HIV-1 and HIV-2

Three years after the initial characterisation of LAV (17) (now designated HIV-1), a second subtype was isolated from patients in Portugal originating from West Africa (40). Sequencing of the isolate revealed that it differed by more than 55% from previous HIV-1 isolates and it was thus designated HIV-2 (40). The major serologic differences between HIV-1 and HIV-2 reside in their envelope glycoproteins. Antibodies to HIV-2 cross-react with Gag and Pol proteins of HIV-1 but do not react with HIV-1 Env proteins and vice versa (89).

HIV-1 subtypes

Genetic variation may enable HIV-1 to escape immune surveillance. The human immune system is challenged by the irregular arrangement of conserved and variable domains in gp120 of HIV-1 (235, 280). The genetic variability of the virus is functionally constrained only by the necessity for successful replication and transmission. However, even the more conserved *gag* gene, encoding matrix, capsid and nucleocapsid proteins internal to the virion, exhibits considerable diversity (93). Most intervention strategies attempt to take the genetic variability of the virus into account. For effective intervention with vaccines and antiviral therapies it is essential to have a broad knowledge of the currently circulating HIV-1 strains in widely distributed geographic areas. This knowledge would provide some assurance that the current understanding of the genetic variability of the virus is both accurate and comprehensive. To this effect, the amplification of parts of the viral genome by PCR and DNA sequencing of the amplicons has allowed rapid comparisons of HIV-1 strains. By 1990 three main groups of HIV-1 had been recognised. The majority of isolates were from North America and Europe and constituted one group; the other isolates were one of two groups from Africa (ELI and MAL) (188). Louwagie and colleagues (163) were the first to carry out a comprehensive survey of currently circulating strains, expanding the HIV-1 *gag* sequence data set fivefold. They sequenced the *gag* gene from 55 individuals from 12 countries on four continents (Africa, South America, Asia and Europe) and compared the sequences with 15 sequences already in the HIV-1 database. Subtypes (referred to as genotypes in their paper) were identified by DNA distance and maximum parsimony methods, both with bootstrap resampling. Their *gag* data set confirmed the validity of the three original groups, added further isolates to a putative fourth subtype and identified three new subtypes. These subtypes were designated A-G.

When *env* sequences from these different strains were analysed phylogenetically similar subtype groupings to those seen in *gag* were observed (188). However they are not identical, as no subtype E exists for *gag*. Viruses with subtype E *env* sequences have *gag* sequences which cluster within subtype A (126). To date, 9 sequence subtypes (or clades) of HIV-1 *env* (A to I, which together form the M or Major group) have been identified (145, 163, 189), There are also several as yet unclassified subtypes (155, 189). The worldwide spread of HIV-1 has, in part, been understood by monitoring the distribution of these genotypic subtypes of the virus. For instance, the work of McCutchan and colleagues shows that B is the predominant subtype in the USA and Europe at present: in the Americas 95% of HIV-1 strains are subtype B (n=180) and in Europe 70% (n=162). In comparison in Asia only 29% (n=110) of characterised strains are of subtype B and in Africa 1% (n=257). The isolates studied by McCutchan *et al.* fell into the following subtypes: A= 123; B=336; C=60; D=67; E=78; F=24; G=15; H=4; O=2 (171). These subtypes are geographically distributed as shown in fig 1.4. It is likely that there are at least 10 times as many people infected with non-B subtype viruses worldwide than with B subtypes (278). WHO estimates indicate that there are more than 8 million HIV-1 infected adults in Africa, (278) most of whom are probably infected with subtypes A, C-H.

Two isolates have been obtained which are genetically distant from the M group. These originate from Cameroon and have been designated subtype O (Outlier) (99, 265). They are more divergent than other HIV-1 strains and appear to be equidistantly related to the HIV-1 M subtype group and the chimpanzee virus SIV_{cpz} (265).

HIV-1 Recombination

A recent study analysed sequence data from the Los Alamos HIV-1 database , searching for evidence of the existence of mosaic viral genomes of HIV-1 and HIV-2 (230). A mosaic genome is defined as a viral genome which consists of sequences derived from two or more distinct subtypes. If the reverse transcriptase of HIV-1 switches templates during proviral synthesis there is the possibility of recombination. The generation of recombinants in this manner requires two different genomes to infect the same cell (either by infection of two viruses at the same time or by superinfection). Then, on transcription to DNA, recombination can occur. Three key points regarding recombination were observed by Sharp and colleagues: 1) 10% of near full-length HIV-1 *gag* and *env* gene sequences are mosaic; 2) recombinants involve nearly all known group M sequences in the Los Alamos database; and 3) there is no evidence for recombination between the M and the O groups. The M and the O groups of HIV-1 are thought to be derived from independent cross-species transmission events, possibly from chimpanzees to humans. Another explanation for mosaic genomes could be that they are a result of *in vitro* or laboratory artifacts (178)). This is thought to be unlikely as the evidence supporting the theory that recombination occurs is internally consistent. For example, A and D are the most prevalent subtypes found in Africa, where A/D recombinants are the most common. A/E viruses (A *gag*, E *env*) were initially thought to be recombinants but, because there are no *gag* subtype E viruses, and A *gag* sequences are unlike most A sequences when analysed phylogenetically, it is now thought that A/E viruses have evolved more quickly in *env* than in *gag*. Thus the A/E viruses can be thought of as being in the process of becoming a distinct subtype.

Practical consequences of HIV-1 subtypes

i) Diagnostic tests

i) The diversity of HIV-1 subtypes may create difficulties in serological tests. This has been observed for HIV-1 and 2 (89). Also, a problem similar to that of HIV-1 and HIV-2 cross-recognition (89), is that antibodies against the O subtype may sometimes not react in serological tests using antigens derived from subtype B viruses, (161). Furthermore, although most HIV-1 diagnoses are accomplished by serology, DNA and RNA based tests such as PCR and NASBA (nucleic acid sequence based amplification) are used as supplemental and confirmatory assays in some laboratories. However, the primers used in these commercial kits were designed to amplify HIV-1 DNA based on the 1989 sequence data and may not amplify diverse genomes (10). DNA and RNA based tests are used to resolve ambiguous serological results, such as those in perinatal HIV-1 infections, where maternal antibody persists in the infant after birth, confounding HIV-1 diagnosis in the infant.

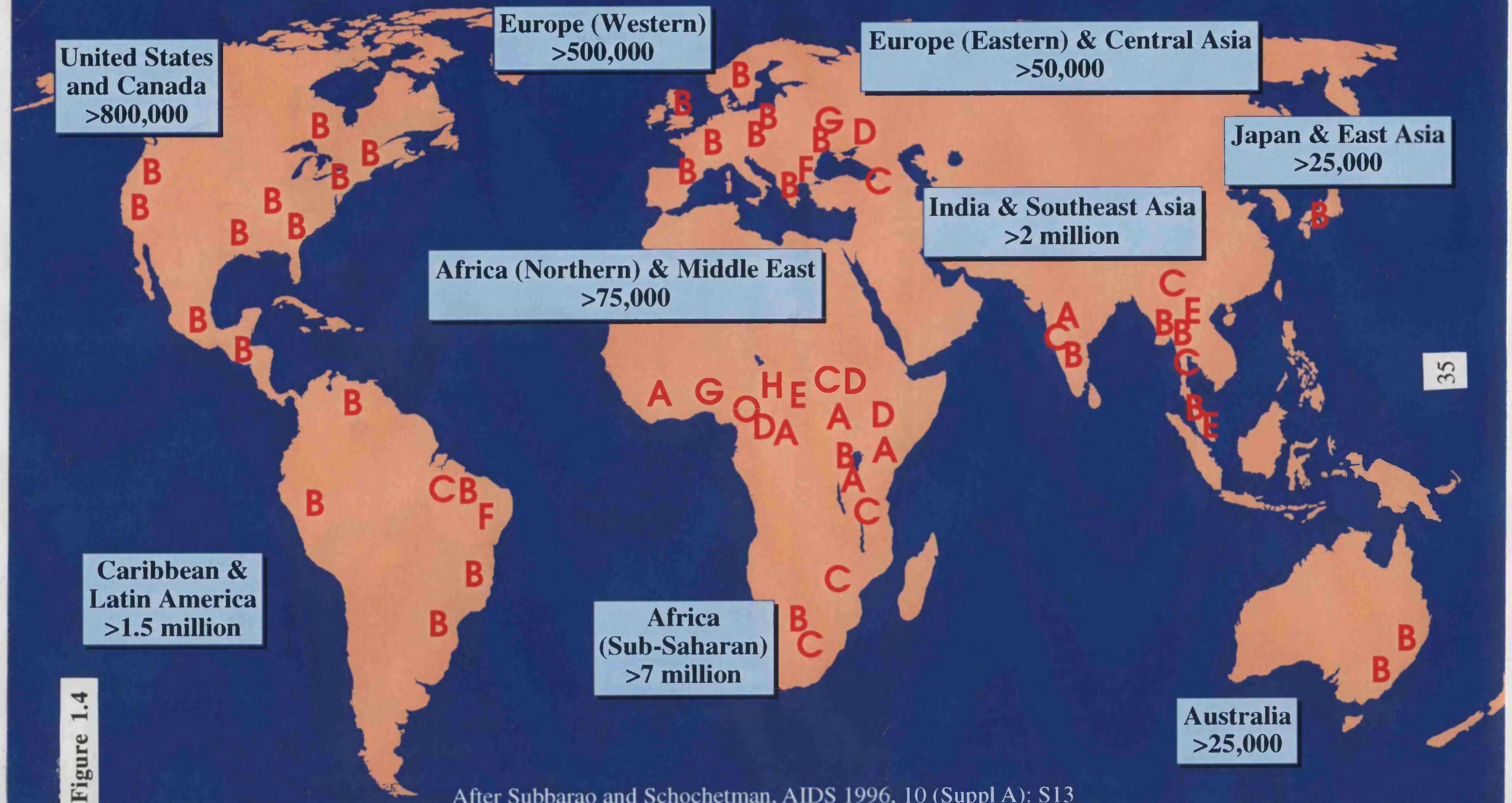
ii) vaccine development

The existence of several subtypes of HIV-1 may cause problems for vaccine design and preparation similar to those encountered for influenza virus (142). An influenza vaccine based on a single haemagglutinin (HA) does not protect from challenge by a virus with a different HA (104). Thus vaccine formulations based on only one HIV-1 strain or subtype may not elicit a broad enough immune response to protect against members of the other subtypes. In Thailand half of virus characterised from drug users is subtype B and half subtype E (28). Hence, plans to vaccinate with Genentech's subtype B-based eukaryotic-expressed gp120 subunit vaccine in Thailand have caused considerable debate, with speculation that the vaccine will not be protective for at least one, if not both, of these Thai subtypes (28).

iii) transmissibility

The African epidemic has affected both men and women equally and they are likely to be infected with a non-B virus, which can be argued to indicate a predominantly heterosexual mode of transmission. It has been claimed that subtypes other than B can be heterosexually transmitted more readily than non B subtypes, and that this is the reason for the lack of significant numbers of heterosexually acquired subtype B infections (245). Additionally, analysis of isolates from Thailand have suggested that subtype A/E may be transmitted more readily by sexual contact (134, 146), and subtype B by intravenous drug use (200). Biologically, the explanation may be that certain virus subtypes may have particular cell or tissue tropisms, for instance for Langerhans cells as described by Soto-Ramirez and colleagues (245). Additionally, differences in the rates of HIV-1 heterosexual transmission may also be attributable to sexual behaviour practices and possibly host genetic susceptibility. Also, certain cofactors may increase the efficiency of heterosexual transmission of HIV, including certain sexual behaviour and the presence of other sexually transmitted diseases (193, 210). It is however possible that this restriction of one particular subtype to one risk group may be the result of a founder effect. This occurs when spread of a subtype in a relatively isolated population (e.g. by culture or behaviour) is determined by an initial chance introduction of one subtype into that group.

WORLDWIDE DISTRIBUTION OF HIV-1



United States and Canada >800,000

Europe (Western) >500,000

Europe (Eastern) & Central Asia >50,000

Japan & East Asia >25,000

India & Southeast Asia >2 million

Africa (Northern) & Middle East >75,000

Caribbean & Latin America >1.5 million

Africa (Sub-Saharan) >7 million

Australia >25,000

After Subbarao and Schochetman, AIDS 1996, 10 (Suppl A): S13

Figure 1.4

For the three reasons described above, the detection of 'minority' subtypes in regions where they have not previously been observed is important and has already included the description of subtype C in India, (97) of subtypes F (minority) and B (majority) in Brazil, (162) and of subtype D in the United States (86). During the investigation of false negative results that arose in the course of an evaluation of the Roche Amplicor PCR kit (15, 261) and in molecular analysis of heterosexual and unexplained transmission events in England (10, 11, 41) HIV-1 infections of subtypes A, C, D, E, and G as well as the majority subtype B were found in this laboratory, and are described in this thesis.

Other methods for distinguishing subtypes

Subtyping by sequencing is an expensive and time consuming technique. Efforts have therefore been made to develop quicker and less costly alternatives. Non-sequencing gel based techniques used to detect differences between other viral genomes include restriction site length polymorphism (RFLP) analysis (207), SSCP (197) and DGGE (33). All of these rely on the existence of relatively few base differences between the DNAs under test. HIV-1 shows extensive sequence heterogeneity and length polymorphism which renders most of these techniques ineffective. However, the technique described by Pieniazek and colleagues, is based on RFLP of the protease gene, a relatively well conserved region of the HIV-1 genome (207). In their report they concede that, using their technique, highly divergent isolates could escape PCR amplification as a result of primer mismatches and, since a single nucleotide substitution could either generate or destroy a restriction site, sequence analysis remains the 'gold standard' in subtyping, though this assay could be applied to screen a large number of samples. The technique with the most potential appears to be the Heteroduplex Mobility Assay (HMA) (55).

Heteroduplex Mobility Assay (HMA)

This method offers a rapid procedure for studying genetic differences among HIV-1 strains and the evolution of genetic variants over time (55). A PCR product is amplified from the viral *env* gene in the sample of interest. The individual complementary single-strands are separated by melting and allowed to anneal with denatured PCR amplicons prepared from reference plasmid clones of representative individual subtypes. Heteroduplexes are formed between the single-stranded DNA from the two sources and these are resolved on polyacrylamide gels. If the two DNAs are closely related a relatively homogeneous migration pattern will be seen. Conversely, if they are less closely related, the heteroduplexes will display a more distinct migration pattern on the gel. However, Novitsky *et al.* (194) concluded that while non-B subtype viruses could be distinguished from B subtypes, it was more difficult to distinguish between the non-B subtypes and to assign a definitive subtype to them.

1.7 HIV-1 transmission studies

Recently sequence data has been used to investigate transmission of HIV-1 (4, 9, 14, 119, 129, 199). These studies have shown that HIV-1 strains isolated from individuals infected from a common source are more similar to one another than HIV-1 strains isolated from individuals with unrelated infections. This has been the basis for deciding whether apparently epidemiologically related cases represent actual transmissions. There are four important questions to consider before embarking on an investigation of HIV-1 transmission events. These are: i) which region of the genome to sequence?; ii) how should the phylogenetic analysis be carried out?; iii) what is the appropriate background population?; and iv) should the intra-sample HIV-1 variants be separated prior to amplification (i.e should one sequence cloned/single molecules rather than 'bulk' PCR products)?

In May 1992 Ou *et al.* described the transmission of HIV-1 from a Florida dentist to several of his patients. They directly sequenced and cloned PCR products from the C2-V5 region of *env* (680bp) from the dentist, patients and local controls (199). They chose this region as it had been used previously in transmission investigations (14, 31, 287) and because the Los Alamos HIV sequence database contained a relative abundance of C2-V3 sequences which could be used for comparative purposes. However, the choice of this region of the genome for transmission investigations has been criticised (53, 116, 202, 240) on the basis that it is too small and too variable for reliable conclusions to be drawn. Also, evolutionary convergence of unrelated sequences can occur in V3 (118, 251) which may confound the identification of linked infections. Strunnikova *et al.* observed temporal progression from early, phylogenetically unrelated sequences to late, convergent sequences in two infants (251). The V3 loop region of the *env* gene is a functional domain where convergence would be expected to occur. In Strunnikova's study the inferred phenotype of the convergent sequences was T-cell line tropic (previously associated with AIDS (179)) while that of the early sequences from the same infants was macrophage tropic. There has thus been continuing controversy about whether the dentist actually infected his patients (16, 187), though the most thorough analysis of the data (111) has upheld the original conclusion that the dentist transmitted HIV-1 to several of his patients. The most vocal criticism has come from journalists associated with the television programme *60 minutes*, which consists largely of assertions that contrary evidence might theoretically exist (16). Although unanswered questions about the case remain, the evidence continues to overwhelmingly support the CDC's theory (27).

In 1993 Holmes *etal.* described the use of p17 PCR products (316 bp), produced by a limiting dilution technique as an alternative to C2-V5 sequencing for determining genetic relatedness between genomes from an HIV-1 infected surgeon and a patient. Unlike the Florida dentist case, material from an alternative source of infection for the patient (an HIV-1 infected blood donor) was available for this investigation. The basis for choosing p17 for analysis was that, unlike V3, this region does not seem to show extensive evolutionary convergence. However, a possible disadvantage of using a small fragment of p17 may be that it is not variable enough to be a reliable marker for distinguishing between related and unrelated infections.

Another HIV-1 transmission, from a rapist to his victim, was investigated in 1993 by Albert and colleagues (4). They chose part of the *pol* gene (640 bp) and used the relatively simple method of direct DNA sequence analysis to show that virus populations harboured by the male and female were highly similar. Although the *pol* gene is relatively conserved they described an unusual amino acid signature pattern shared by the male and female and from this concluded that the two infections were related. This type of amino acid signature pattern analysis has also been used in the Florida dentist investigation (199).

The HIV-1 transmission investigations discussed above all had access to specimens from the likely source of infection, and the probability of transmission from the donor to the recipient was assessed, at least in the first three cases, by measuring sequence similarity using maximum likelihood methods. These HIV-1 transmission studies have usually included sequence data from the most likely source of infection, and the probability of transmission from the donor to the recipient has been assessed by measuring sequence similarity against control data using likelihood analysis. However, Jaffe and colleagues (129) have reported an absence of HIV-1 transmission in a second Florida dental practice. Unlike the three studies described previously, Jaffe *et al.* did not have specimens from the most probable source of infection for the patients. Therefore, their study involved an analysis of the genetic relatedness among HIV-1 strains together with identification of potential risks for acquisition of HIV-1 infection and an investigation of control measures in the dental practice. They compared V3 sequence distances, obtained by either cloning or direct sequencing of bulk PCR products, from the dentist, the HIV-1 positive patients and two of the patients' sexual partners. The importance of not obtaining a single direct consensus from a 'bulk' PCR was demonstrated in a study showing that in some cases a minor HIV-1 variant was transmitted between sexual partners (292). Therefore, if a direct consensus or 'bulk' sequence from PCR is obtained and not single molecules, it is possible that such minor variants could be missed. The amino acid alignments in chapter 6 from sequences obtained from single molecules also illustrate the point that the consensus sequence does not exist (for example CPHL1, position 1288-1290) or is represented by only one sequence (for example CPHL2, position 532-534). Also, the length polymorphism associated with HIV-1 is apparent in all multiple sequences examined for this thesis with the exception of CPHL2 and CPHL11, this has obvious implications for sequencing techniques which do not employ an initial dilution step such as cloning or limit dilution.

In the light of the above considerations and in order to avoid the controversies surrounding the Florida dentist investigation, the transmission investigations described in chapter 4 were done on full gp120 sequences amplified from single molecules. One aim of this work was to establish the optimum region of the HIV-1 genome for comparison of sequence data for the investigation of possible transmission events and to ascertain the most informative and reliable molecular analysis method.

1.8 Routes of HIV-1 transmission

The risk factors for the transmission of HIV-1 are well established. Almost all HIV-1 infections occur by one of four routes: sexual intercourse, sharing contaminated needles, treatment with infected blood or blood products, and perinatal transmission from mother to infant (8). Occasionally unusual modes of HIV-1 transmission appear to be the likely cause of subsequent infection. An example of an unusual mode is the transmission between two children, the first infected perinatally and the second probably infected through open skin lesions or mucous membranes by infected blood of the first child who had frequent episodes of bleeding (75). A further example of HIV-1 transmission between children from one home was described by Brownstein and colleagues (29). In this case an adolescent with haemophilia apparently acquired HIV-1 infection from his older brother who was HIV-1 positive and also had haemophilia. This probably occurred via a shared razor both brothers cut themselves with. Only two of the 8115 AIDS cases and 20,543 HIV-1 infections in the U.K. reported by the end of September 1993 may have occurred by other routes (8). Both were reported to be due to blood contact during fights: one through blood contact with open skin lesions and the other by a bite (8). Although these occurrences are rare, they emphasise the need to avoid exposure to HIV infected blood. The infamous case of the Florida dentist described above is, to date, the only documented case of a Health Care Worker (HCW) transmitting HIV-1 to his patients (202). The route of transmission in this case is still unknown and will probably remain unresolved. Still very rare, but more likely, are occupationally acquired infections.

Heterosexually acquired HIV-1 infection in the U.K.

Analysis of the envelope sequences of HIV-1 has allowed the subtypes A - I to be identified. In the U.S.A. and western Europe subtype B is prevalent for all transmission groups: blood-borne and homosexual or heterosexual transmissions. Heterosexually acquired infections in subsaharan Africa, in India or Thailand are mainly of subtype C or A or D or E (57, 62, 97, 126, 163, 172). There are occasional isolates of subtype B, but they are only frequent when there are drug users, for example the epidemics of subtype B among drug users in Bangkok (200)) and Manipur, India (18, 172). Ou and colleagues reported that most of the subtype B isolates they studied from Thailand had the GPGQ tetrapeptide at the tip of the V3 loop common in African HIV-1 strains but rare in North America and Europe, where the GPGR motif predominates. There are sporadic cases of subtype B in India or Africa, but this subtype is almost always confined to drug users, haemophiliacs, or blood transfusion cases. These cases account for approximately only 1% cases of infection in Africa (62). Soto-Ramirez and colleagues suggest that subtypes other than B can be heterosexually transmitted ten times more efficiently and give this as the reason for the lack of significant heterosexually acquired subtype B infection (245). According to CDSC statistics, of all reports of HIV-1 infected persons by exposure category in the U.K. to the end of December 1995 (206) 4644 out of 25635 (approximately 1 out of 5) HIV-1 infected persons acquired their infections heterosexually. Three thousand four hundred and forty three (approximately 3 out of 4) of these heterosexually acquired infections were thought to be through exposure to partners outside the U.K. Assuming the majority of infections acquired outside the U.K.were non B subtypes, as few as a quarter of heterosexually acquired HIV-1 infections in the U.K.may be of subtype B. The possibility of a heterosexually transmitted epidemic in the U.K may therefore be increased by these introductions of subtypes other than B. To gain insight into heterosexual transmission in England we investigated two transmission events (chapter 4). The work described in chapters 3 and 4 of this thesis suggests that some of the U.K. heterosexually acquired infections are second generation non-B infections (infections acquired outside the U.K. but transmitted to another individual in the U.K.).

HIV-1 infection via contaminated needles

Intravenous drug users (IVDUs) are the risk group with the largest number of HIV-1-infected patients in Scotland, with the majority of infections centred in Edinburgh (117). Approximately 49% of reported HIV-1 positive individuals in Scotland were thought to have become infected through intravenous drug use. This proportion is higher than that found in the rest of the U.K., where on average, less than 15% of HIV-1 infections within any community are related to drug use (117). Holmes *et al.* also suggested that the variants currently spreading through heterosexual contact came from the virus variant that first infected the the IVDU population in 1983-1984.

HIV-1 infection via contaminated blood/blood products

Haemophiliac patients infected by factor VIII, such as the cohort studied by the group at the Centre for HIV Research at the University of Edinburgh (14, 235, 236), constituted the largest group at risk via contaminated blood products prior to screening for HIV-1 antibody in blood and blood products. Holmes and colleagues' phylogenetic analysis suggested that the HIV-1 infections in the haemophiliac cohort were sufficiently different from the IVDU and the heterosexual populations to dispute claims that the haemophiliacs were infected from the IVDU community (117).

Perinatal HIV-1 infection

Studies on HIV-1 infections in pregnancy have demonstrated that virus is transmitted to the infant in 15-35% of cases (192). The majority of infections occur either late in pregnancy or at delivery with between 2-27% of infants being infected *in utero* (26, 58). The biological characteristics of maternal HIV-1 isolates seem to be an important factor in whether perinatal infection occurs. Mothers with virus isolates which grow rapidly to high titres in T cell lines (rapid/high) more frequently transmit virus to their children, although both rapid/high and slow/low (slow growing to low titres) viruses can be transmitted (3). Albert and colleagues also showed that HIV-1 with rapid/high replicative capacity does not appear to have a selective advantage during mother to child transmission. Using MT2 tropism as a marker they found, in both perinatal and sexual transmission, that rapid/high virus can be more often isolated from the transmitter than the slow/low type of virus. However, individual members of a virus population, when studied as single molecular clones, may show distinct biological properties from the virus population as a whole, thus biassing results.

Mother to child transmission is multi-factorial and maternal/fetal factors (clinically advanced disease, primary infection, twins, prematurity, chorioamnionitis and infant host immune response) should be considered as well as viral (load, phenotype and genotype) and immune factors (decreased CD4+ cells, cell-mediated and humoral responses). The presence of maternal neutralising antibody appears to be an important factor in transmission at delivery and is associated with increased protection (30). Strategies suggested for preventing/reducing mother to child transmission include: i) decreasing the viral load in pregnant women; ii) maintaining placental integrity; iii) providing passive anti-HIV-1 antibody to the pregnant woman, fetus and newborn. The number of RNA copies per mL of maternal plasma is thought to be important in determining whether the mother transmits to her infant (56). Dickover and colleagues noted that approximately fifty thousand copies per mL appears to be a threshold - 75% of mothers with this number or higher transmitted to their infants, whereas if the mothers' viral load was lower (< 20,000 copies per mL) there was no vertical transmission of HIV-1(56). As most vertical transmission occurs in late gestation or at birth, strategies to prevent a high viral load at this particular time are critical. Neutralising antibody production by the infants appears to be an important factor as to whether the child progresses to AIDS or not. Rapid progressors tend to have a high viral load. Infants infected via intrauterine transmission tend to progress more rapidly. By having a Caesarean section and not breast feeding, transmission of HIV-1 can be reduced by 50% and 14% respectively (252, 253). If primary maternal infection is postpartum and breast feeding occurs the estimated risk of transmission is 29% (59).

HIV-1 clearance in infants

As early as 1986, reports suggested the possibility that some children born to HIV-seropositive mothers may have been perinatally infected and have subsequently cleared the virus, leading to a stable seronegative status (19). A report by Bryson *et al.* described a case in which infection of the infant was strongly suggested by means of virus sequence similarity between the mother and the infant but the infant subsequently became seronegative (30). A recent report by Roques *et al.* describes clearance of HIV-1 in a group of 12 infants with perinatal exposure to HIV. Six infants tested HIV-positive in both coculture and PCR assays, the remaining six being repeatedly HIV-positive in PCR assays although culture negative (216). In all these children, clear positive HIV-1 detection was obtained in at least two successive samples. Roques *et al.* hypothesised that virus clearance could be due to humoral response in the mothers and/or children, leading to the transmission of neutralised virus particles or virus-harbouring cells. In this study they report data which show that HIV-1 clearance may not be due to high functional antibody titres in mothers or children. However, critics suggested that this phenomenon is most likely explained by contamination of the original blood samples or errors in labelling the samples (20). Although there is general agreement that HIV-1 clearance is biologically plausible, stricter diagnostic criteria are required before HIV-1 clearance is universally accepted.

Diagnosis of HIV-1 in infants

Before the advent of HIV-1 diagnostic PCR, the only way to determine if an infant born to an infected mother was not infected with HIV-1 was either by co-culture of possibly HIV-1 infected blood from the infant or through IgG testing of serial infant serum samples to demonstrate sero-reversion as maternal antibody declined. However, maternal antibodies can take up to 18 months to disappear. Commercial and in-house PCR for HIV-1 DNA or cDNA, NASBA for RNA, detection of p24 antigen following acid dissociation, antibody class-specific assays, (in particular IgA antibody) and virus isolation have all been applied to try and establish or exclude infant infection more rapidly, in order to maximise the benefits of specific anti-HIV-1 agents and meet the need for prophylaxis against opportunistic infections (184). Each assay has benefits and drawbacks with respect to speed, sensitivity, specificity and the volume of blood required. Diagnosis of infection should only be made after two independent tests on a single sample are positive and a second sample is also positive. If negative PCR results are obtained after screening monthly for six months, this suggests that the infant has escaped infection (184). However, it is difficult to equate the behaviour of adult and infant specimens in PCR tests. Smaller volumes of blood are collected from babies and the specimen is often less convenient than an adult specimen. Also, there may be a lower concentration of virus in the infant than the adult. The most frequent cause of PCR failures has been reported to be a low viral load (15, 290). Thus, the very specimens for which PCR should be diagnostically the most useful may be the most problematic (261).

A possible diagnostic PCR failure was noted in a specimen from an infant (CPHL11) born to an HIV-1 infected mother (CPHL10), described in chapter 4. The infant was thought to be HIV-1 infected from clinical observation (*Pneumocystis carinii* pneumonia, or PCP), but cell lysates from blood taken 3 months after birth gave negative reactions in the Roche Amplicor HIV-1 PCR assay. The same lysates were found to be reactive in a PCR using primer sets amplifying regions of the *gag*, *pol* and the LTR (136). The specimen was found to be positive for these primers using only 1/20th the sample input volume recommended for the Amplicor assay, suggesting that the PCR failure was due to mismatching rather than low input copy number (10). A specimen from the mother was positive in the Amplicor test.

Occupationally acquired HIV-1 infection in the U.K.

From 1985-1992 there were 176 occupational exposures to HIV-1 in the U.K. reported to the PHLS Communicable Disease Surveillance Centre (108). The outcome was reported for 134 of these incidents; two resulted in seroconversion, including one following zidovudine post exposure, giving an observed transmission rate of 2%. In this particular study the sample size was small and the rate appeared slightly higher than, though not inconsistent with, comparatively larger studies where rates ranged from 0.18% to 0.56% (108). By September 1993 there had been 182 reported occupationally acquired, or possibly occupationally acquired, HIV-1 infections worldwide (109). Part of the work of this thesis includes the investigation of two putative occupationally acquired HIV-1 infections in the U.K. using comparison of sequence data (chapter 4).

CHAPTER 2

MATERIALS AND METHODS

This chapter consists of two sections. The first section describes the buffers, solutions, and materials used during work carried out in this thesis and the second section describes the experimental methods used. The methods section is divided into two parts: the experimental methods used; and computer based analytical methods. Both parts include sufficient detail to enable accurate reproduction of all experiments carried out for this thesis.

MATERIALS

2.1 Buffers

2.1.1 **TE Buffer:**

10 mM Tris-HCl (pH 7.5)
1 mM disodium EDTA

2.1.2 **Phosphate Buffered Saline (PBS):**

8.5 g/l NaCl
1.07 g/l Na_2HP0_4
0.39 g/l NaH_2P0_4
(pH 7.1)

Electrophoresis Buffers

2.1.3 **10 x TBE Buffer:**

870 mM Tris-HCl
870 mM Boric acid
20 mM disodium-EDTA
(pH 8.5)

2.1.4 **Sequencing Gel Loading Buffer:**

10 mM disodium EDTA (pH 8.0)

Prepared in deionised formamide (2.2.4)

2.1.5 **Agarose Gel Loading Buffer:**

10% (w/v) Ficoll 400

0.25% (w/v) Orange G

Prepared in 1 x TE (2.1.1)

DNA Extraction Buffers

2.1.6 **Lymphocyte Lysis Buffer:**

10 mM Tris-HCl (pH 8.3)

1 mM EDTA

0.5% (v/v) Nonidet NP-40

0.5% (v/v) Tween 20

300 μg/mL proteinase K (Boehringer Mannheim)

2.1.7 **DNA Extraction Buffer (from post mortem tissue):**

500 μL 10 mM Tris-HCl (pH 7.4)

1 mM $MgCl_2$

0.5 % (w/v) SDS

200 μg/mL proteinase K

Magnetic Bead PCR Amplicon Purification Buffers

2.1.8 **Bead Wash Buffer:**

0.1% (w/v) BSA

Prepared in PBS (2.1.2)

2.1.9 **Binding and Washing Buffer (2 X B&W):**

10 mM Tris-HCl (pH 7.5)

1 mM EDTA

2 mM NaCl

RNA Extraction Buffers

2.1.10 **Lysis Buffer L6:**

120 g guanidinium thiocyanate

100 mL 0.1 M Tris-HCl (pH 6.4)

22 mL 0.2 M EDTA (pH 8.0)

2.6 g Triton X-100

Stirred overnight in the dark to dissolve. Stored in the dark and used within one month.

2.1.11 **Washing Buffer L2**

120 g guanidinium thiocyanate

100 mL 0.1 M Tris-HCl (pH 6.4)

Heat with shaking or stirring at 60-65°C

2.1.12 **Silica suspension**

60 g silicon dioxide

Prepared in 500 mL deionised water and left to stand at room temperature for 24 hours. Four hundred and thirty mL of the supernatant was removed by vacuum suction and the remainder was resuspended in 500 mL deionised water. The solution was then left to stand for 5 hours at room temperature. About 440 mL of the supernatant was then removed and 600 μL concentrated HCl to pH 2.0 added. The solution was then aliquoted, autoclaved and stored in the dark.

HLA Typing Buffers

2.1.13 **Citrate buffer**

18.4 g trisodium citrate dihydrate

The trisodium citrate dihydrate was dissolved in 800 mL deionised water and the pH was adjusted to 5.0 (±0.2) by the addition of citric acid monohydrate (approximately 6 g). The final volume was adjusted to 1 L using deionised water and the solution stored at room temperature.

2.1.14 **20 x SSPE buffer**

7.4. g $Na_2EDTA.2H_20$

210 g NaCl

27.6 g $NaH_2PO_4.H_20$

The $Na_2EDTA.2H_20$ was dissolved in 800 mL deionised water and the pH adjusted to 6.0 with 10 N NaOH to aid the dissolution of the EDTA. Two hundred and ten g of NaCl and 27.6 g $NaH_2PO_4.H_20$ were added and the pH adjusted to 7.4 (±0.2) with 10 N NaOH (about 10 mL) and the final volume was adjusted to 1 L with deionised water.

2.1.15 20% Sodium Dodecyl Sulphate (SDS)

200 g electrophoresis grade SDS

The SDS was dissolved in 800 mL of deionised water and heated to 35-50°C to aid dissolution. The final volume was adjusted to 1 L with deionised water.

2.1.16 Hybridisation Solution

250 mL SSPE (2.1.14)

25 mL 20% SDS (2.1.15)

The above solutions were combined and 725 mL deionised water added to a final volume of 1 L. The solids in the solution were dissolved before use by warming in a water bath at 60°C.

2.1.17 Wash Solution

250 mL SSPE (2.1.14)

10 mL 20% SDS (2.1.15)

The above solutions were combined and 1740 mL deionised water added to a final volume of 1 L. The solids in the solution were dissolved before use by warming in a water bath at 60°C.

2.2 Other Solutions

Gel Solutions

2.2.1 40% Polyacrylamide Stock Solution (19:1):

38 g/100 mL acrylamide

2 g/100 mL bis-acrylamide

These were dissolved in distilled water and stored at 4°C.

2.2.2 6% Denaturing Sequencing Gel:

15 mL 40% Polyacrylamide Stock Solution (19:1) (2.2.1)
50 g Urea
1-2 g Amberlite mixed bed ion exchange resin

These were dissolved in deionised water using a magnetic heater stirrer for 5 minutes, without allowing the temperature of the solution to go above room temperature. Ten mL 10 x TBE (2.1.3) was then added and the solution was then made up to 100 mL with deionised water and transferred to a 0.2 μM filtration unit and filtered and degassed under vacuum. To start polymerisation, 0.5 mL of freshly made 10% ammonium persulphate and 45 μL TEMED were added and the gel poured immediately. The gel was allowed to set for at least 2 hours before use.

2.2.3 1% Agarose Gel:

1 g SeaKem Agarose (FMC)
100 mL 1 x TBE buffer (2.1.3)

The above were combined in a 250 mL flask and gently swirled to aid hydration of the agarose particles and stoppered with a foam bung. The solution was then heated on a medium to high setting for 2.5 minutes in a microwave. After heating, and using a glove or some other protection, the hot liquid was gently swirled to ensure complete dissolution of the agarose and further heated if required. After cooling to 50-60°C, the solution was poured into a gel tray and allowed to set after the comb was put in place. Once set, 15μL of sample were combined with 5 μL of agarose gel loading buffer (2.1.5) and carefully loaded into a well. A suitable marker was also loaded, usually a 1 kb ladder (Gibco BRL). The gel was run at 100 mV for 45 minutes. Products were visualised by ethidium bromide staining (222).

2.2.4 Deionised Formamide:

Ten mL of formamide was stirred with 1 g of Amberlite mixed-bed resin for 30 minutes at room temperature followed by filtration through a Nalgene 0.2 μM filtration unit and stored in 100 μL aliquots at -20°C.

METHODS

2.3 Standard Procedures

These protocols are based on those in Sambrook *et al.* (222) except where otherwise stated.

2.3.1 **Organic Extractions**

Phenol Extraction

Phenol/water/chloroform (68:18:14; ABi) was used to remove protein contamination from DNA and also used to remove excess Dye Terminators from cycle sequencing reactions. The DNA solution was made up to a minimum volume of 100 μL and an equal volume of phenol/water/chloroform added; the mixture was vortexed and spun in a microfuge for 1-5 minutes. The upper, aqueous layer was removed to a clean microtube and the process repeated as required.

Chloroform Extraction

Chloroform extraction was used to remove traces of protein and phenol from DNA samples. An equal volume of chloroform (Analar, BDH) was added to the DNA, vortexed and spun in a microfuge to separate the layers. The upper aqueous layer was removed to a clean microtube and the process repeated

2.3.2 **Ethanol Precipitation**

DNA was precipitated by the addition of either 0.15 volumes 2 M sodium acetate (pH 4.5) or a half volume 7.5 M ammonium acetate and 2.5 volumes ice cold ethanol. The solution was mixed well and either stored at -20°C or centrifuged immediately in a microfuge (12000-14000 rpm) for 20 minutes.The pellet was washed in 70% ice cold ethanol by gentle vortexing to remove salt followed by centrifugation in a microfuge for 20 minutes. The majority of 70% ethanol was removed carefully using a disposable extended tip pastette, avoiding disturbance of the pellet. The pellet was then left to air dry or dried under vacuum.Pellets were either stored dry at -20°C or resuspended in either sterile distilled water, TE (2.1.1) or Sequencing gel loading buffer (2.1.4).

2.3.3 **Determination of DNA Concentration**

i) Determination of DNA Concentration by Spectrophotometer

The absorbance of the DNA sample at 260 nm and 280 nm was measured against a water blank in quartz cuvettes using a Beckman Du6 spectrophotometer. The DNA concentration was calculated according to the formula:

A_{260} = 50 μg/mL for ds DNA or 33 μg/mL for ss DNA.

DNA purity was estimated from the ratio of A_{260}/A_{280}, which is 1.8 for DNA free from protein contamination.

ii) Estimation of Concentration by Comparison with a known Standard

In the case of PCR amplicons, where primers and dNTPs confound absorbance readings, concentrations were estimated by running dilutions of the samples with a known standard on a 1% agarose gel (2.2.3) and visualised by ethidium bromide staining.

2.3.4 **Lymphocyte Purification from Whole Blood**

i) 'Micro Ficoll' Method

Using a screw cap microtube, 0.5 mL of PBS (2.1.2) was layered carefully over 0.5 mL of Ficoll Hypaque (Pharmacia). Half a mL of EDTA whole blood was added gently to the microtube without disturbing the Ficoll/aqueous interface. The sample was centrifuged in a microfuge for 1 minute (12000-14000 rpm). The 'buffy coat' (lymphocyte layer) was removed using a sterile wide bore pastette and transferred to a clean microtube containing 0.5 mL of PBS (2.1.2) and centrifuged for 3 minutes to pellet the cells. The excess PBS was removed and the cells resuspended in Amplicor wash buffer (Roche Diagnostic Systems) and vortexed to lyse any residual red blood cells. The sample was centrifuged for 3 minutes, excess wash buffer removed and the cell pellet stored at -20°C.

ii) Roche Method

Half a mL of EDTA whole blood was added to 1 mL of Amplicor wash buffer (Roche Diagnostic Systems) in a screw cap microtube and inverted 5-10 times. After a 5 minute incubation at room temperature the microtube was inverted 3-5 times and incubated for a further 5 minutes at room temperature. The sample was centrifuged in a microfuge for 3 minutes and the supernatant removed without disturbing the pellet. One mL of wash buffer was added to the microtube and it was vortexed. This process was repeated twice more. The supernatant was aspirated and the cell pellet was either stored at -20°C or extracted immediately (2.3.5).

2.3.5 **DNA Extraction**

a) **From lymphocytes**

i) Lymphocyte Lysis

The cell pellet (2.3.4) was washed in PBS (2.1.2), centrifuged and resuspended in 100 μL of Lymphocyte Lysis buffer (2.1.6) and incubated at 55°C for 1 hour. The proteinase K was inactivated by heating to 94°C for 15 minutes and the extract stored at -20°C.

ii) Roche Extraction Method

One hundred μL of extraction reagent was mixed, warmed to room temperature and added to the cell pellet. The sample was vortexed and incubated at 60°C (± 2°C) for 30 minutes in a dry heat temperature block followed by an incubation at 100°C (± 2°C) for 30 minutes. The extract was stored at -20°C.

b) **From post mortem tissue blocks**

One mL of xylene (BDH) was added to paraffin embedded tissue (approximately 1cm by 1 cm by 0.3 cm) in a screw cap tube, then vortexed sporadically for 20 minutes to remove excess paraffin. The xylene was then removed with a pastette and the tissue washed in 500 μL of ethanol (Analar, BDH). The tissue was then chopped up with a sterile scalpel and added to 500 μL 10 mM Tris pH 7.4, 1 mM $MgCl_2$, 150 mM NaCl, 10 μL 10 mg/mL proteinase K, 25 μL 10% SDS and incubated at 37°C for 24 to 72 hours. After incubation, the debris was pelleted and the supernatant was extracted with phenol/chloroform three times and once with chloroform/ IAA. The DNA was then precipitated overnight by adding a half volume of 7.5 M ammonium acetate and two volumes of ice cold ethanol. After centrifugation (20 minutes, 14000 rpm) the pellet was washed in 70% ethanol and resuspended in sterile distilled water overnight at $^+$4°C. Microconcentrators were used to concentrate and purify the samples (2.3.9).

2.3.6 Silica/GuSCN DNA/RNA Extraction (Boom method)

Sample preparation was carried out in a laboratory set aside solely for nucleic acid extraction for PCR. Tissue samples were homogenised in TE buffer to give a 10% suspension.

Fifty to two hundred microlitres of sample were added to 900 μL lysis buffer L6 (2.1.10) and 40 μL silica (2.1.12) in a screw capped microfuge tube. This silica suspension was vortexed before pipetting. Positive displacement pipettes or tips with filter plugs were used at all times. The silica/sample mix was vortexed for 5 seconds, and left to stand at room temperature for 5 minutes. The sample was then vortexed for a further 5 seconds, microfuged for 15 seconds and the supernatant removed with a plastic, fine tipped pastette. The supernatant was discarded into 10M NaOH in a fume hood to avoid production of HCN. One mL of buffer L2 (2.1.11) was added and vortexed, microfuged for 15 seconds and the supernatant removed as above. This step was repeated. The same step was further repeated: twice with 70% ethanol; once with acetone. The sample was dried for 10 minutes at 56°C on a heating block with the lid of the tube removed. Fifty microlitres of TE or RNase free water was added, the tube capped and vortexed and incubated at 56°C for 10 minutes. The sample was then vortexed and microfuged for 2 minutes. The supernatant was collected, ensuring no carry over of silica. If the sample was to be stored (at -70°C only) 1 μL RNase inhibitor was added.

2.3.7 PCR

i) DNA

Proviral DNA was prepared from EDTA blood samples by differential lysis or Micro Ficoll lymphocyte purification (2.3.4) followed by dilution to single molecules by limit dilution (2.3.8) (236). Nested PCR was carried out with one PCR primer tagged with biotin in the second round. PCR conditions were: 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 200 μM of each dNTP, 1.5 mM $MgCl_2$, 0.5 unit *Taq* polymerase and 5 pmol of each primer. First round primers 989L or 988L and 633L or 631L (Table 2.1) for gp120 and 469W and CA1 (Table 2.2) for p6/protease were used with the following cycling conditions: 94°C for 40 sec; 47°C for 35 sec; 72°C for 4 minutes for 30 cycles. Nested PCR was carried out using the same reaction conditions, using inner primers 944S and 609RE for gp120 (Table 2.1) and CA2 and BCA3 (Table 2.2) for p6/protease using the following cycling conditions: 95°C for 40 sec; 55°C for 35 sec; 72°C for 4 minutes for 35 cycles. Precise primer location maps for both regions are shown in figures 2.1 and 2.2 and a schematic of primer locations within gp120 are shown in figure 2.3. Samples which proved more difficult to amplify in gp120 were multiplexed with different combinations of four first round primers and two second round primers until sufficient amplification had been achieved.

ii) RNA (Reverse Transcription or cDNA step)

Fifty microlitres of specimen (serum or plasma) were extracted by the Boom method (2.3.6). After extraction 40 μL was recovered and used in the reverse transcription (cDNA) step: 22.2 μL of extracted RNA was added to 17.8 μL of cDNA mix (final concentration of 10 mM Tris pH 8.3 at 25°C, 50 mM KCl, 5 mM $MgCl_2$, 1 mM dNTP mix, 1 nmol random hexamer mix, 4 Units RNasin, 200 Units MLV RTase). The mixture was incubated at room temperature for 10 minutes for primer annealing followed by 42°C for 45 minutes for cDNA extension and 100°C for 5 minutes to inactivate the reverse transcriptase. It was then placed on ice. Twenty microlitres of this mixture was then combined with 80 μL of primary PCR mix to a final concentration of 10 mM Tris pH 8.3 at 25°C, 50 mM KCl, 1.25 mM $MgCl_2$, 5 pmoles of each primer, 2.5 Units of Taq. The tubes were overlayed with mineral oil and cycled: 94°C, 1'; 50°C, 1'; 72°C, 3'; x 35. After cycling, 2 μL of the primary PCR mix was added to a secondary PCR amplification mix.

Table 2.1

***env* PCR primers** (for location on gp120 see figure 2.1)

Sequence (5'->3')	Location w.r.t. **ATG** (fig 2.1) (-) indicates antisense	Primer name
TCATCAGAACAGTCAGACTCATCAAGC	-218>-202	989L
GTAGCAATAATAATAGCAATAG	-136>-103	988L
TCCCACTCCATCCAGGTC	1988>1971(-)	633L
CCAGACTGTGAGTTGCAACAGATGC	1805>1782(-)	631L
AGAAAGAGCAGAAGACAGTGGCAATG	-22>3	944S
CCCATAGTGCTTCCGGCCGCTCCCAAG	1685>1658(-)	609RE
GGGATATTGATGATCTGTAGTGC	75>95	627L
GTGGGTCACAGTCTATTATGGG	108>129	626L
CACCACGCGTCTCTTTGCCTTGGTGGG	1607>1582(-)	125Y
GAACCCAAGGAACA	1658>1646(-)	915N

Table 2.2

p6/protease PCR primers

Sequence (5'->3')	Location w.r.t. nuc. 1373 LAI (-) indicates antisense	Primer name
GCTACACTAGAAGAAATGATG	-18>4	CA1
TATTCCTAATTGAACTTCCC	1040>1021(-)	469W
GATGACAGCATGTCAGGGAG	1>20	CA2
GGCCATTGTTTAACTTTTGGG	854>835(-)	CA3

2.3.8 **Limit dilution PCR**

Limit dilution of HIV-1 required the initial determination of the dilution factor that allowed one amplifiable molecule of the genome to be amplified by the PCR (236). With this approach the selection bias associated with *in vitro* culture documented by Meyerhans et. al (178) was avoided. Also, copying errors due to *Taq* polymerase were minimised by amplification from a single proviral DNA molecule using nested primers and direct sequencing of the PCR product.

Using plugged tips, 5 μL of DNA extracted from lymphocytes (2.3.5) was added to 95 μL 1° PCR mix (2.3.7). The mixture was carefully mixed by gently pipetting up and down and a 20 μL aliquot taken and added to 80 μL of 1° PCR mix, making a 1 in 5 dilution. This step was repeated. One further dilution was made by adding 20 μL of the third dilution to 60 μL of 1° PCR mix. Each 1 in 5 dilution (total 80 μL) was divided into 4 PCR microtubes and cycled as described in section 2.3.7. This process created 4 duplicates of 4 serial dilutions of 1/5. The secondary PCR was carried out as described in section 2.3.7. The products were run on a 1% agarose gel (2.2.3) and visualised by ethidium bromide staining.

The Poisson distribution describes the sampling distribution of the number of occurrences, *r*, of an event during a period of time (or region of space or volume). It depends upon just one parameter, which is the mean number of occurrences, μ, in periods of the same length (or equal regions of space or volume):

The observed frequency (*fo*) of *r* occurrences = $e^{-\mu}$, where e is the mathematical constant 2.71828 (244). As the distribution of very dilute DNA between tubes is a random stochastic process, the Poisson formula was used by Simmonds *et al* (236) to calculate the likelihood of positive tubes having originally contained one, or more than one molecule of target DNA (Table 2.3). When the frequency of PCR-positive reactions is 0.2, approximately 95% of reactions can be predicted to have originated from single target sequences, whereas only half of reactions will be derived from single sequences when the frequency of positives is 0.7. As 5 μL of lymphocyte lysate was the initial volume used the first dilution gave the equivalent of 1 μL of lymphocyte lysate per tube, the second 1/5 μL of lymphocyte lysate, the third 1/25 μL of lymphocyte lysate and the final dilution 1/125 μL of lymphocyte lysate. If, for example, the third dilution was the one which gave 1 out of 4 bands positive in the gel, 48 1° PCR tubes were set up with the equivalent of 1/25 μL of lymphocyte lysate per tube (48 x 1/125 = 1.92 μL of lymphocyte lysate) to generate approximately 5-10 out of 48 tubes containing single amplifiable gp120 molecules after 2° amplification (2.3.7), electrophoresis (2.2.3) and visualised by ethidium bromide staining.

- Less than 5 positive tubes had an even greater chance of being generated from single rather than multiple molecules. More than 10 positives (i.e greater than 1/5th of the total number of tubes) indicated that some of the positives (> 1/30) would have been generated from more than one molecule.

Table 2.3

Quantitation and separation of sequences by limiting dilution (taken from 154)

Observed frequency positives	Calculated number DNA sequences[a]	Proportion single copies[b]
0.001	0.001	99.9%
0.01	0.010	99.5%
0.05	0.051	97.5%
0.10	0.105	94.8%
0.15	0.163	92.1%
0.20	0.223	89.3%
0.25	0.288	86.3%
0.30	0.357	83.2%
0.40	0.511	76.6%
0.50	0.693	69.3%
0.60	0.916	61.1%
0.70	1.204	51.2%
0.80	1.609	40.2%
0.90	2.302	25.6%
0.95	2.996	15.8%

[a] Actual frequency, f (in target molecules per replicate tube) calculated according to the formula $f = -\ln(fo)$, where (fo) is the observed frequency of negative reactions.
[b] Proportion of positive replicates, f^1 derived from a single copy of target DNA.

2.3.9 **PCR Amplicon Purification**

i) Magnetic Bead Purification

Twenty μL (200 μg) of streptavidin-coated magnetic beads (Dynal M280) were washed by adding 500 μL of Bead Wash buffer (2.1.8), placing the microtube into the magnetic separator for 60 seconds and removing the supernatant with a pipette. The microtube was removed from the separator and the beads resuspended in 20 μL of Binding and Washing buffer (2.1.9) by gentle mixing. The microtube was placed into the separator for 60 seconds and the supernatant removed with a pipette. The microtube was removed from the separator and the beads resuspended in 40 μL of Binding and Washing buffer (2.1.9). Forty μL of biotinylated PCR amplicon were added to the microtube and incubated for 15 minutes at room temperature. The beads were kept resuspended during this time by occasional inversion. The microtube was placed into the separator and the supernatant removed with a pipette. Forty μL of Binding and Washing buffer (2.1.9) were added to the microtube and the immobilised product was washed by turning the microtube in the magnetic separator. The supernatant was removed and the product resuspended in 30 μL of distilled water and stored at 4°C prior to cycle sequencing.

ii) Centricon-100 columns (Amicon Ltd)

Centricon 100 columns were used to concentrate and to remove PCR primers and dNTPs. Excess oil was removed by rolling PCR products on a parafilm strip or the sample was transferred by pipetting from underneath the oil, before reconstituting with 0.75 mL deionised water. The sample was centrifuged at 4000 rpm in an assembled Centricon concentration unit for 5 minutes. The sample was washed three times in deionised water by centrifugation, and recovered by inversion and centrifugation at 2000 rpm for 2 minutes. The concentrated PCR product was recovered in approximately 20 ul of water. The products were run on a 1% agarose gel (2.2.3) and visualised by ethidium bromide staining.

2.3.10 **Dye Terminator Cycle Sequencing**

After quantitation of amplicons (2.3.3), both strands were sequenced with a total of 11-15 sequencing primers (see Table 2) per PCR product on an Applied Biosystems 373A sequencer, using *Taq* DyeDeoxy Terminator sequencing kits, according to the manufacturer's instructions. Figure 2.1 shows a precise gp120 primer location map and figure 2.3 shows a schematic of sequencing primers located within gp120. Table 2.2 shows sequences of primers used for sequencing. Each base was sequenced at least 2.4 times. Good sequencing results were obtained with a 4:1 primer:template ratio, using 3.2 pmol primer in a terminator reaction. The amount of PCR product used was calculated from its size, for example: for 3.2 pmol of primer, the number of base-pairs in the PCR product was divided by 2 and the calculated amount of DNA in ng used. After cycling, unincorporated dye terminators were removed from the sequencing reactions by phenol extraction (2.3.1) and the purified product precipitated by ethanol precipitation (2.3.2). The pellets were washed in 70% ethanol and dried. The samples were either run immediately or stored dry at -20°C for up to 2 months. Prior to loading, the samples were resuspended in 3 μL of sequencing gel loading buffer (2.1.4), heated to 96°C in a heating block for 2 minutes and placed immediately on ice. The samples were then loaded onto a 6% sequencing gel (2.2.2) and run overnight.

2.3.11 **HLA Typing**

A kit was used according to the manufacturers instructions (Perkin Elmer Corporation). The AmpliType HLA DQα Forensic Kit amplified a region of the HLA DQα gene. The kit distinguished six alleles which defined twenty-one different groups. Essentially, DNA extracted from a sample was added to a tube containing amplification reagents. The DQα DNA sequences were amplified and the DNA produced was hybridised to DNA Probe strips. Capture probes were used to distinguish the six alleles. These probes were immobilised in a labelled pattern of dots on the Probe Strips. Captured DQα DNA was detected by a chemical reaction generating a blue colour in the probe dot, and the resulting pattern of blue dots on a Probe Strip revealed the DQα alleles present in the amplified sample.

2.3.12 **Heteroduplex Mobility Assay (HMA) Primers**

The Heteroduplex Mobility Assay (HMA) kit was provided by the NIH AIDS Research and Reference Reagents Program through the U.K. MRC Reagent Project. The first round primers were ED13 and ED14 which amplify an approximately 2000 bp fragment from the first exon of *rev* to the transmembrane protein gp41 coding region of *env*. Three sets of second round primers were used: i) ED5 and ED12 which amplify an internal fragment of about 1250 bp spanning the V1-V5 coding region of gp120 (generating approximately 1200 bp of sequence data); ii) ES7 and ES8 which amplify the 700 bp V3-V5 region of gp120 (generating approximately 600 bp of sequence data); iii) ED31 and ED33 which amplify the 500 bp C2-C3 region of gp120 (generating approximately 450 bp of sequence data). HMA was not performed for work carried out for this thesis, but the primers supplied in the kit were used to carry out PCR for some isolates where indicated.

Table 2.4

***env* sequencing primers**

Sequence (5'->3')	Location w.r.t. **ATG** (fig 2.1) (-) indicates antisense	Primer name
AAATAGACAGGTTAATTG	-51>-34	113V
GTGGGTCACAGTCTATTATGGG	108>129	626L
GAGGATATAATCAGTTTATGGG	322>343	625L
GGATCAAGCCTAAAGCCATGTG	342>364	624L
GCTCTTTCAATATCACCAC	494>512	016S
TACACAGGCCTGTCCAAAGG	666>685	623L
TGGCAGTCTAGCAGAAGAAG	849>868	619L
ACATTGTAACATTAGTAGAG	1056>1075	015S
ATCCTCAGGAGGGGACCCAG	1158>1177	618L
GGAAAAGCAATGTATGCCC	1366>1384	014S
TATGAGGGACAATTGGAG	1512>1529	013S
TCCCAAGAACCCAAGGAACA	1665>1646(-)	915
TTCACTTCTCCAATTGTCCCTC	1536>1515(-)	616L
GGGCATACATTGCTTTTCC	1384>1366(-)	325H
TTACAGTAGAAAAATTCCCCTC	1226>1205(-)	617L
CTCTACTAATGTTACAATGT	1075>1056(-)	CA22
GTACATTGTACTGTGCTGACATT	806>784(-)	621L
CAATAATGTATGGGAATTGGC	716>696(-)	622L
GTGGTGATATTGAAAGAGC	512>494(-)	CA23
CCCATAAACTGATTATATCCTC	343>322(-)	CA24
CATAATAGACCGTGACCCAC	127>108(-)	223K

Table 2.5

p6/protease sequencing primers

Sequence (5'->3')	Location w.r.t. nuc.1373 LAI (-) indicates antisense	Primer name
AAGCAAGAGTTTTGGCTGAA	35>54	CA4
ACAGGCTAATTTTTTAGGGA	271>290	CA5
ACAGGAGCAGATGATACAGT	558>578	CA6
GCCATCCATTCCTGGCTTTA	835>815(-)	CA7
CCTGTATCTAATAGAGCTTC	663>544(-)	CA8
TCCCTAAAAAATTAGCCTGT	290>271(-)	CA9

Figure 2.1 Primer location map of gp120

Primers located in the env 5' leader sequence (vpu).

```
6000
GCTCATCAGA ACAGTCAGAC TCATCAAGCT TCTCTATCAA AGCAGTAAGT AGTACATGTA
  |---      989L         --->
ATGCAACCTA TAATAGTAGC AATAGTAGCA TTAGTAGTAG CAATAATAAT AGCAATAGTT
                         |---        CG CC   988L          --->
                    |----  943S      --->
GTGTGGTCCA TAGTAATCAT AGAATATAGG AAAATATTAA GACAAAGAAA AATAGACAGG
                                                     |---   113V
6180                                        met
TTAATTGATA GACTAATAGA AAGAGCAGAA GACAGTGGCA ATGAGAGTGA AGGAGAAGTA
 113V->            |----      944S          --->
                          G CC  C
            Biotin |----      CA18          --->
```

Consensus alignment of 14 env sequences from HIV-1

Sequences included in alignment:

BH102; BRU; CDC42; ELI; JH32; MAL; MN; RF; SC; SF2; WMJ22; Z3; Z6.

```
1                                                   50
------ATGA GAGTGAAGG- GATAAAGAGG AATTATCAGC ACTTGGGGTG
51                                                 100
GAGATGGGGC ACCATGCTCC TTGGGATGTT GATGATCTGT AGTGCTGCAG
                          |---   627L        --->
                      3'end of signal peptide      |
101                                                150
AAAAATTGTG GGTCACAGTC TATTATGGGG TACCTGTGTG GAAAGAAGCA
       |---      C 626L  --->
       <----     G 223K  ---|(CA19)
           |BstEII|
151                                                200
ACCACCACTC TATTTTGTGC ATCAGATGCT AAAGCATATG ATACAGAGGT
201                                                250
ACATAATGTT TGGGCCACAC ATGCCTGTGT ACCCACAGAC CCCAACCCAC
251                                                300
AAGAAGTAGT ATTGGAAAAT GTGACAGAAA ATTTTAACAT GTGGAAAAAT
301                                                350
AACATGGTAG AACAGATGCA TGAGGATATA ATCAGTTTAT GGGATCAAAG
                                              |--624L
                       |---     625L       --->
                       <---     CA24       ---|
351                                                400
CCTAAAGCCA TGTGTAAAAT TAACCCCACT CTGTGTTACT TTAAATTGCA
624L        --->
401                                                450
CTGATTTGAA GAATGATACT A-TAC-AATA -TA-TACTAA TACCAATAGT
451                                                500
AGTA-CGGGG AAAAGATAAT GGAGAAAGGA GAAATGAAAA ACTGCTCTTT
                                               <---
501                                                550
CAATATCACC ACAAGCATAA GAGATAAG-T GCAGAAAGAA TATGCACTTT
016S   ---->
CA23   -----|
551                                                600
TTTATAAACT TGATGTAGTA CCAATAGAT- -G---AATA- TA----TA--
601                                                650
AATAATGATA --AATA-TAC TAATAGTACC AGCTATAGGT TGATAAATTG
651                                                700
TAA-ACCTCA GTCATTACAC AGGCCTGTCC AAAGGTATCC TTTGAGCCAA
                |---    623L      --->           <----
701                                                750
TTCCCATACA TTATTGTGCC CCGGCTGGTT TTGCGATTCT AAAGTGTAAT
 622L       ---|
751                                                800
GATAAGAAGT TCAATGGAAC AGGACCATGT ACAAATGTCA GCACAGTACA
                                    <---  621L     ---
801                                                850
ATGTACACAT GGAATTAGGC CAGTAGTGTC AACTCAACTG CTGTTAAATG
```

```
621L -|                                              |-
 851                                                900
 GCAGTCTAGC AGAAGAAGAG GTAGTAATTA GATCTGAAAA TTTCACAGAC
 ---    619L    --->
 901                                                950
 AATGCTAAAA CCATAATAGT ACAGCTGAAT GAATCTGTAG AAATTAATTG
 951                                               1000
 TACAAGACCC AACAACAATA CAAGAAAAAG TATAC-TATC CAGAGAGGAC
 1001                                              1050
 CAGGGAGAGC ATTTTATACA ACAGGAAAAG ATATAATAGG AAATATAAGA
 1051                                              1100
 CAAGCACATT GTAACATTAG TAGAGCAAAA TGGAATAACA CTTTAAAACA
      |---  015S       --->
      <---  CA22        ---|
 1101                                              1150
 GATAG-TAGA AAATTAAGAG AACAATTTGG -AA-AATAAA ACAATAATCT
 1151                                              1200
 TTAATCAATC CTCAGGAGGG GACCCAGAAA TTGTAACGCA CAGTTTTAAT
        |---     618L     --->
 1201                                              1250
 TGTGGAGGGG AATTTTTCTA CTGTAAT-CA ACACAACTGT TTAATAGTAC
     <---      617L      ---|
 1251                                              1300
 TTGGAATA-T AATAGT-CTA GTACTAAAGG TG--GCGTCA AATAACACTG
 1301                                              1350
 AAGGAAATGA CACAATCACA CTCCCATGCA GAATAAAACA AATTATAAAC
 1351                                              1400
 ATGTGGCAGG AAGTAGGAAA AGCAATGTAT GCCCCTCCCA TCGAAGGACA
                 |---      014S   --->
                 <---    325H     ---|(CA21)
 1401                                              1450
 AATTAGATGT TCATCAAATA TTACAGGGCT GCTATTAACA AGAGATGGTG
 1451                                              1500
 GTAATA---- ----GGTA-T GACAATAATG A-ACCGAGAT CTTCAGACCT
 1501                                              1550
 GGAGGAGGAG ATATGAGGGA CAATTGGAGA AGTGAATTAT ATAAATATAA
                <---   616L         ---|
             |--       013S --->
 1551                                              1600
 AGTAGTAAAA ATTGAACCAT TAGGAGTAGC ACCCACCAAG GCAAAGAGAA
                                   <---    125Y       C
                                                     |-
 1601                                              1650
 GAGTGGTGCA GAGAGAAAAA AGAGCAGTGG GAACAATAGG AGCTATGTTC
  C ---|                  |                       <--
MluI|
                          1st base of gp41 (GCA=Ala)
 1651                                              1700
 CTTGGGTTCT TGGGAGCAGC AGGAAGCACT ATGGGCGCAG CGTCAATGAC
 915N ---|(CA20)|---- 701 (SK68)  --->
         <---      G   C   609RE   ---|
 1701                                              1750
 GCTGACGGTA CAGGCCAGAC AATTATTGTC TGGTATAGTG CAACAGCA-A
 1751                                              1800
 ACAATTTGCT GAGGGCTATT GAGGCGCAAC AGCATCTGTT GCAACTCACA
                                   <---   631L
```

```
1801                                              1850
GTCTGGGGCA TCAAGCAGCT CCAGGCAAGA -TCCTGGCTG TGGAAAGATA
631L-|
1851                                              1900
CCTAAAGGAT CAACAGCTCC TAGGGATTTG GGGTTGCTCT GGAAAACTCA
1901                                              1950
TTTGCACCAC TACTGTGCCT TGGAATGCTA GTTGGAGTAA TAAATCTCTG
1951                                              2000
GATGAGATTT GGAATAACAT GACCTGGATG GAGTGGGAAA GAGAAATTGA
                      <---    633L   ---|
2001                                              2050
CAATTACACA AGCTTAATAT ACACCTTAAT TGAAGAATCG CAAAACCAGC
2051                                              2100
AAGAAAAGAA TGAACAAGAA TTATTGGAAT TGGATAA-TG GGCAAGTTTG
2101                                              2150
TGGAATTGGT TTAACATAAC AAATTGGCTG TGGTATATAA AAATATTCAT
2151                                              2200
AATGATAGTA GGAGGCTTGG TAGGTTTAAG AATAGTTTTT GCTGT-CTTT
2201                                              2250
CTATAGTGAA TAGAGTTAGG CAGGGATACT CACCATTATC GTTTCAGACC
2251                                              2300
C-CCTCCCAA CCCCGAGGGG ---ACCCGAC AGGCCCGAAG GAAT-GAAGA
2301                                              2350
AGAAGGTGGA GAGAGAGACA GAGACAGATC CGTTCGATTA GTGAA-GGAT
2351                                              2400
TCTTAGCACT TATCTGGGAC GATCTGCGGA GCCTGTGCCT CTTCAGCTAC
2401                                              2450
CACCGCTTGA GAGACTTACT CTTGATTGTA -CGAGGATTG TGGAACTTCT
2451                                              2500
GGGACGCAGG GGGTGGGAAG CCCTCAAATA TTGGTGGAAT CTCCTACAGT
2501                                              2550
ATTGGAGTCA GGAACTAAAG AATAGTGCTG TTAGCTTGCT TAATGCCACA
2551                                              2600
GC-ATAGCAG TAGCTGAGGG GACAGATAGG GTTATAGAAG TAGTACAAAG
2601                                              2650
AGCTTGTAGA GCTATTCTCC ACATACCTAG AAGAATAAGA CAGGGCTTGG
2651                                              2700
AAAGGGCTTT GCTATAA--- ---------- ---------- ----------
2701
```

Figure 2.2 Primer location map of p6/protease

Consensus alignment of 14 p6/protease sequences from HIV-1

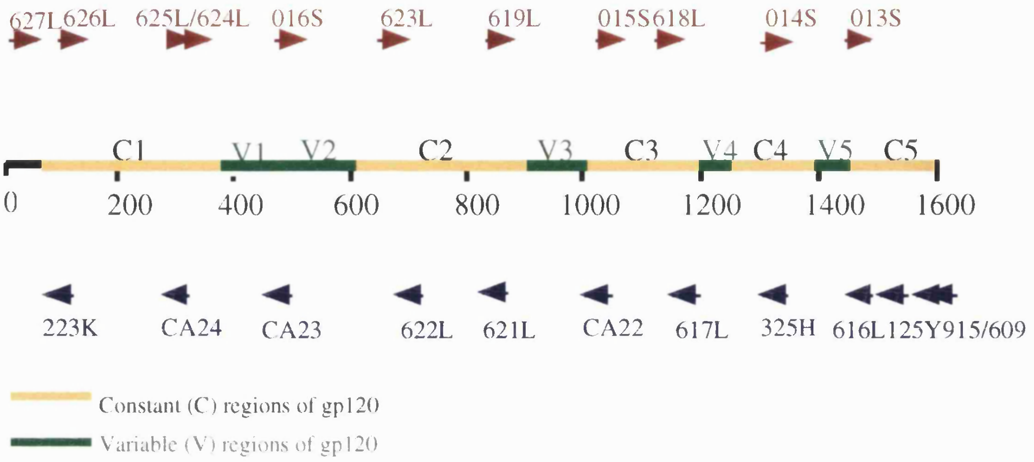
Sequences included in alignment:

BRU; HXB2; JRFL; HAN; NL43; SF2; OY1; RF; ELI; Z2Z6; NDK; MAL;U455; ANT70.

```
1                                                    50
GATGACAGCA TGTCAGGGAG TGGGGGGACC CGGCCATAAA GCAAGAGTTT
|---     CA2    ---->                |---    CA4
51                                                  100
TGGCTGAAGC AATG-----A GCCAAGTAA- ---CAAATTC AGC---TACC
--->
101                                                 150
A---TAATGA TGCAGAGAGG CAATTTTAGG AACCAAAGAA AGATTGTTAA
151                                                 200
GTGTTTCAAT TGTGGCAAAG AAGGGCAGAT AGCCAAAAAT TGCAGGGCCC
201                                                 250
CTAGGAAAAA GGGCTGTTGG AAATGTGGAA AGGAAGGACA CCAAATGAAA
251                                                 300
GATTGCACTG A-GAG----- ACAGGCTAAT TTTTTAGGGA AGATCTGGCC
                      |---    CA5      --->
                      <---    CA9      ---|
301                                                 350
TTCCCACAAG GGAAGGCCAG GGAATTTTCT TCAGAGCAGA CCAGAGCCA-
351                                                 400
---------- ---------- ---------- ---CAACAGC CCCACCAGAA
401                                                 450
GAGAGCTTCA GGTTTGGGGA AGAGACAACA ACTCCCTCTC AGAAGCAGGA
451                                                 500
GCCGATAGAC AAGGAACTGT ATCCTTTAAC TTCCCTCAAA TCACTCTTTG
501                                                 550
GCAACGACCC CTCGTCACAA TAAAGATAGG GGGGCAACTA AAGGAAGCTC
                                               <---
551                                                 600
TATTAGATAC AGGAGCAGAT GATACAGTAT TAGAAGAAAT GAATTTGCCA
 CA8     ----|
       |---     CA6       --->
601                                                 650
GGAAAATGGA AACCAAAAAT GATAGGGGGA ATTGGAGGTT TTATCAAAGT
651                                                 700
AAGACAGTAT GATCAGATAC TCATAGAAAT CTGTGGACAT AAAGCTATAG
701                                                 750
GTACAGTATT AGTAGGACCT ACACCTGTCA ACATAATTGG AAGAAATCTG
751                                                 800
TTGACTCAGA TTGGTTGCAC TTTAAATTTT CCCATTAGTC CTATTGAAAC
801                                                 850
TGTACCAGTA AAATTAAAGC CAGGAATGGA TGGCCCAAAA GTTAAACAAT
                                    <---        CA3
               <---      CA7      ---|
851
GGCC
---|
```

Figure 2.3

Schematic of sequencing primer locations within gp120


627L 626L 625L/624L 016S 623L 619L 015S 618L 014S 013S
C1 V1 V2 C2 V3 C3 V4 C4 V5 C5
0 200 400 600 800 1000 1200 1400 1600
223K CA24 CA23 622L 621L CA22 617L 325H 616L 125Y 915/609
Constant (C) regions of gp120
Variable (V) regions of gp120

2.4 Computer-based analytical methods

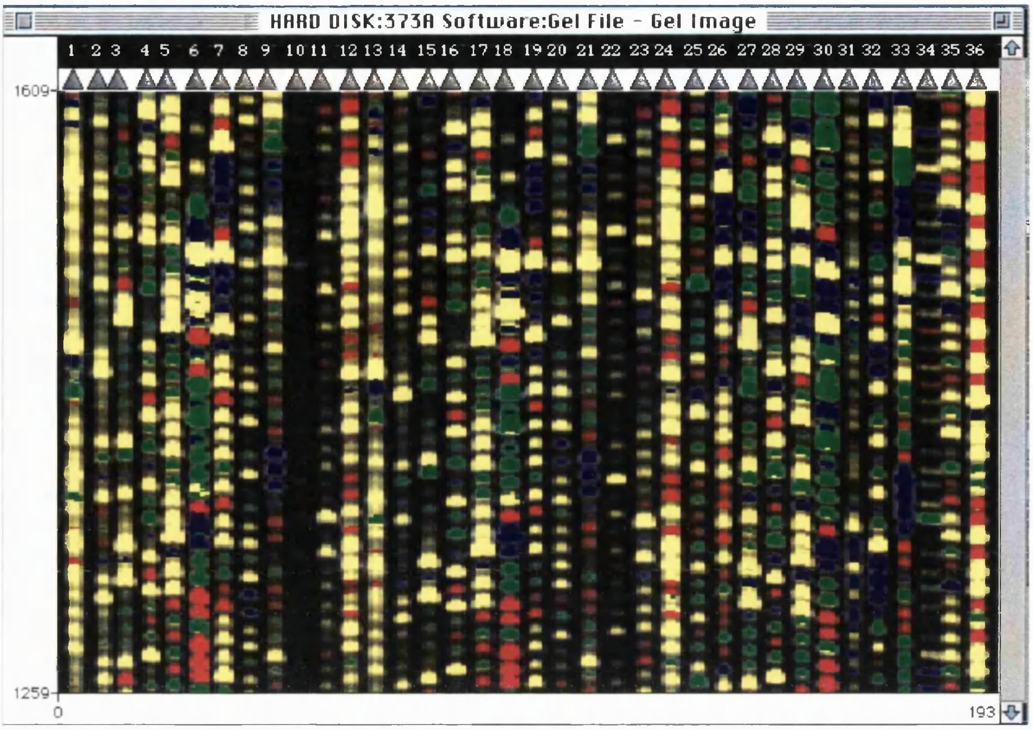
Both DNA strands were sequenced from each PCR product with a total of 11-15 sequencing primers per amplicon (see Table 2.1) on an Applied Biosystems 373A sequencer, using the DyeDeoxy Terminator sequencing kits. 373A-analysed sequence data was then further manipulated as follows:

2.4.1 **Revision of Dyedeoxy Terminator Sequence Data from the ABi 373A Analysis program**

This section describes the analysis of sequence data generated by the ABi automated sequencer after sequencing PCR products using *Taq* Dye Terminators. The choice of *Taq* Dye Terminators rather than Dye Primer sequencing was influenced by the type of template sequenced in this laboratory, i.e. PCR products of diverse viral genomes rather than cloned material. For our purposes Dye Terminator sequencing was more flexible and allowed the use of unmodified primers for sequencing. Although Dye Terminator sequencing was a more versatile method for this type of sequencing when compared with Dye Primers, the different chemistries used had different effects on the appearance of the data. When using Dye Terminators it was essential to sequence PCR products in both directions due to the low signal of some peaks in a chromatogram following certain sequences of nucleotides (listed below). If the PCR product was sequenced in only one direction across such a sequence, and the resulting base(s) was indeterminate, sequencing of the opposite strand was the only way to resolve any ambiguities.

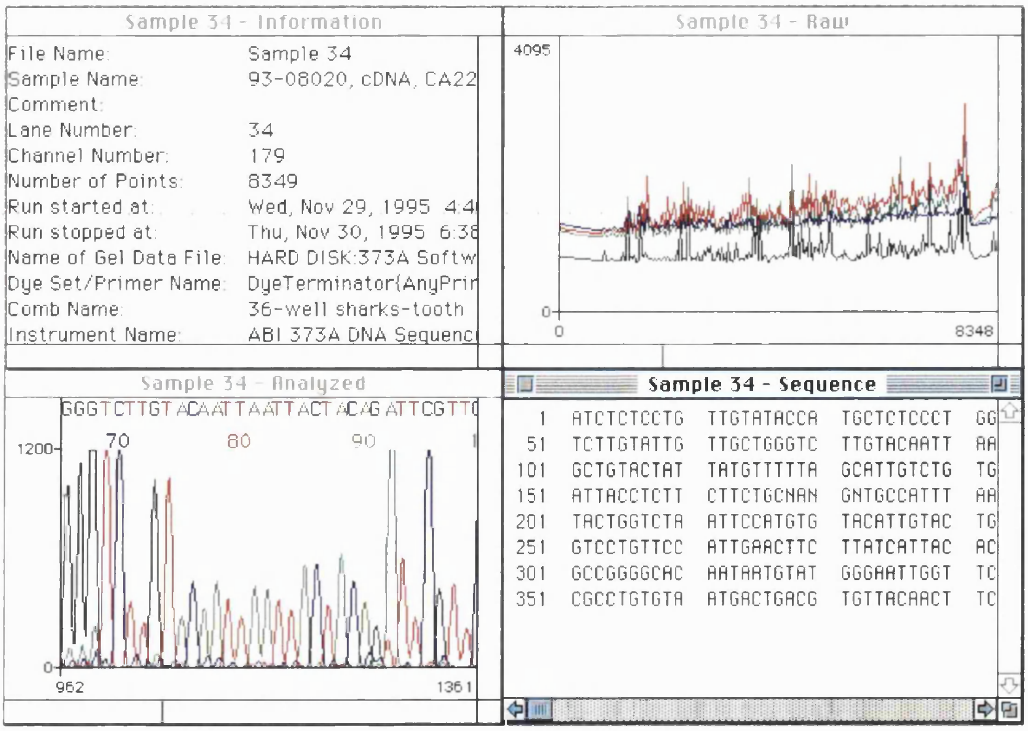
When the run was finished the 'Analysis' program was launched (see 'Analysed data' below). The 'gel file' or 'Gel Image' (Figure 2.4) displayed once 'Analysis' had finished was viewed in its entirety by using the scroll bars. 'Analysis' should have found the line of best fit through each lane of data, seen as the grey tracker lanes on the gel file. ABi Analysis was launched either by double-clicking on the 'Analysis icon' where individual sample files were opened by selecting 'Open' from the 'File' pull down menu or by double-clicking on the sample file being analysed.

Figure 2.4 typical 'Gel Image'



A = Green G = Blue C = Red T = Yellow

Figure 2.5 Sample file components



File Name: Sample 34
Sample Name: 93-08020, cDNA, CA22
Comment:
Lane Number: 34
Channel Number: 179
Number of Points: 8349
Run started at: Wed, Nov 29, 1995 4:4
Run stopped at: Thu, Nov 30, 1995 6:38
Name of Gel Data File: HARD DISK:373A Softw
Dye Set/Primer Name: DyeTerminator{AnyPrir
Comb Name: 36-well sharks-tooth
Instrument Name: ABI 373A DNA Sequenc

```
  1  ATCTCTCCTG  TTGTATACCA  TGCTCTCCCT  GG
 51  TCTTGTATTG  TTGCTGGGTC  TTGTACAATT  AA
101  GCTGTACTAT  TATGTTTTTA  GCATTGTCTG  TG
151  ATTACCTCTT  CTTCTGCNAN  GNTGCCATTT  AA
201  TACTGGTCTA  ATTCCATGTG  TACATTGTAC  TG
251  GTCCTGTTCC  ATTGAACTTC  TTATCATTAC  AC
301  GCCGGGGCAC  AATAATGTAT  GGGAATTGGT  TC
351  CGCCTGTGTA  ATGACTGACG  TGTTACAACT  TC
```

The sample files generated consisted of four components: (**i**) file information; (**ii**) raw data; (**iii**) analysed data; and (**i v**) sequence data (figure 2.5).

(**i**) File information

This file gave detailed information about the particular run. It was important to check that the base spacing was around 12 (i.e. if the gel was running correctly one base should have taken the same time to pass the read region as 12 scans of the laser); but good data was also obtained with spacing values between 9 and 15. A note was made of the base spacing in each case. The signal strength was checked to make sure it was within the specified ranges (see 'reanalysing a file' below).

(**ii**) Raw data

Raw data was data that had not been analysed by the 'Analysis' software. The raw data peaks were small, with a tall peak at the beginning of the sequencing run (this large peak was due to residual dye terminators remaining after phenol extraction).

(**iii**) Analysed data

After data collection and lane tracking, the data in the sample files were analysed. This was an automatic process involving preprocessing (signal strength analysis, finding base 1 location), first base calling, respacing of raw data and second base calling (based on respaced raw data).

(**i v**) Sequence data

This was the sequence data from the sample file in the standard IUPAC 5 letter code (A,G,C,T,N). It could be cut and pasted into different files if required.

Base Calling

Base calling of ABi sequence data was the selection of a region of raw data which was then further analysed by 'Analysis' software to produce an accurate recognisable chromatogram (also called a trace or an electropherogram). Chromatogram data was then assembled in the SeqEd v1.03 (ABi) (see 2.4.2) contig assembly program. The data were aligned, overlapped and cross-checked and the assembled sequence data was then exported and manipulated further using various DNA analysis software packages (see 2.4.3, 2.3.4). When sequencing PCR products, unlike cloned material, the sequence data generated were of a defined length. It was therefore important to accurately 'base call' raw data, defining the region of raw data which was then further analysed. Base calling was necessary if the improper assignment of base 1 had occurred and it would avoid unnecessary delays in contig assembly, as untidy data at either the 5' or 3' end of the sequence resulted in an assembly program not recognising clearly overlapping sequences, resulting in the creation of a contig with a consensus sequence containing added gaps and ambiguities.

Reanalysing a file

a). In 'Analysis', the 'File' menu was pulled down and 'Open' selected. A dialogue box appeared and enabled selection of the sample file required. Alternatively, double-clicking on a sample file would open it.

b). The 'Window' menu was pulled down and 'File Info' selected. The signal levels of the sample were checked as follows: ABi guidelines for ds DNA using *Taq* Terminators gave the following signal strength guidelines for good sequencing results: **A**=(200-300), **G**=(200-400), **T**=(50-150), **C**=(20-70). The base spacing was noted and checked to be within the required range. If the gel had run normally the spacing would be between 9-16 (see 'File Information' above).

c). The 'Window' menu was pulled down and 'Raw Data' and 'Controller' selected.

d). The Controller was used to find the start point of the data. The zoom or custom tools on the controller were used to position the cursor just after the large Dye Terminator peak towards the beginning of the raw data (this peak represented residual dye terminators after phenol extraction). A note was made of the X-axis number that appeared in the lower left-hand corner of the 'Scan' window (the start point for analysis).

e). The 'View' menu was pulled down and 'Full View' selected. Smaller PCR products that produced discrete raw data, i.e. with an unambiguous start and finish had an easily found end point using the zoom or custom tools as described above. For longer products the following was carried out: 'Analysed Data' from the 'Window' pull down menu was selected. Using 'custom view' the chromatogram was checked to gauge how much accurate sequence data could be obtained for that sample (for example approximately 400 bp). This number was multiplied by the base spacing, found under 'File Info' (e.g. base spacing = 11.3, multiplied by 400 = 4520). Add 4520 to your start point figure (e.g.

1050 + 4520 =5570) to obtain the end point figure. 'Analysis' was pulled down and 'Call Bases' selected. The 'Start' and 'End' figures were entered in the relevant boxes. The 'Use Start Point' and 'Print After Calling' boxes were checked if required. The other parameters did not need changing. When 'OK' was selected the 'Analysis Queue' appeared and the various stages of analyses and base calling were automatically carried out. When this was complete a new sample file was written over the old analysed data file. Raw data remained unchanged.

f). The tools on the Controller were used to edit the file, i.e., to zoom in on any area of the analysed data and to find or change bases. Each chromatogram was checked to ensure the computer had called the correct bases, if in doubt the base in question was changed to an N and was resolved when the sequence was aligned with the sequence of the opposite strand. The 'Delete to Last Base' command was used to delete unwanted bases from the end of a sequence by highlighting the last base required to remain in the sequence and selecting 'Delete to Last Base' under the 'Edit' pull down menu. All bases to the right of the highlighted base were deleted. Characteristic patterns were observed when using dye terminator chemistry because of a steric influence of the dyes when attached to ddNTPs. Some consistent patterns of ddNTP incorporation were seen in the chromatograms of analysed dye terminator runs (see Table 2.6).

Table 2.6

Characteristic patterns observed when using dye terminator chemistry

C's following **G**'s were weak

T's following **G**'s were weak.

C's following two or more **T**'s were enhanced.

In a string of four or more **G**'s, the third showed reduced signal.

G's following a string of **T**'s were enhanced.

The first **A** in a string was strong, the rest weak.

A's after **T**'s showed reduced signal.

Exporting and Importing Sequence data to and from different Files

The sequencer produced sequence data in two formats: sample files (previously described) and ASCII text files (files with the suffix .seq). The text files were exported for manipulation in MS.DOS based computers or any system requiring text files. The sequence sample files required specialised software in order to access chromatogram data, these included: ABi Analysis; SeqEd v.1.03; SeqMan (DNAStar). Once the sequence data was in a suitably refined condition it was imported into applications (usually SeqEd v. 1.03 (2.4.2) which created overlapping contigs using sequence data in both orientations.

2.4.2 **SeqEd v 1.03**

The SeqEd application was opened by double clicking the SeqEd icon. A new layout was opened by 'New Layout' from the 'File' pull down menu being selected. ABi sequences were imported into the file by 'Import Sequence' from the 'Sequences' pull down menu being selected and locating the sequences of interest in the dialogue box which was automatically presented. The sequences were selected either by being highlighted and double-clicking with the mouse or by being highlighted and the 'import' option selected. This process was repeated as necessary to import the required sequences into the SeqEd layout, in the order the sequences overlap each other. Any sequences which were antisense (i.e opposite strand) were reverse complemented by the sequence name in the ID panel on the left of the layout being highlighted and 'reverse complement sequence' selected using the 'Sequences' pull down menu.

Sequences were overlapped or aligned by the names of the sequences of interest being highlighted and 'overlap', 'comparative' or 'multiple' selected from the 'Align' pull down menu. After the dialogue box automatically presented was checked (the default settings given in the dialogue boxes are adequate for most applications) OK was selected and the alignment or overlap was carried out. The multiple alignment feature in the SeqEd application was not very powerful and larger alignments were carried out in MegAlign (2.4.4) or ClustalV (2.4.6) and PHYLIP (2.4.7) applications. After the alignment or overlap was complete, 'create shadows' under the 'sequences' pull down menu was used to select 'compare two sequences'. When OK had been selected, asterisked areas, i.e. areas indicating mismatch between two sequences, were highlighted and 'Display Chromatograms' selected from the 'Sequences' pull down menu to view areas of ambiguity. The associated chromatograms (or electropherograms) were viewed and the peaks which could not be assigned to one particular base were renamed N (unknown). This process was repeated until all the primers overlapped to create a contig (a continuous sequence consisting of overlapping sequences). The layout was then saved and all the named sequences highlighted and a 'unanimity sequence' was created by 'create shadow' being selected from the 'Sequences' pull down menu. The newly created unanimity sequence was 'frozen' by 'freeze shadow' being selected from the 'Sequences' pull down menu and then translated into the derived amino acid sequence by 'create shadow' followed by 'translate codons to amino acids' being selected under the 'Sequence' pull down menu. All three reading frames were checked and the 'universal' code selected in the dialogue box which was subsequently presented. The three letter amino acid or single letter amino acid code was selected as required and the unanimity sequence exported by 'export sequences' being selected from the 'Sequences' pull down menu. Sequences were exported in a various formats including; Chromatoref; text; Wisconsin (GCG); Intelligenetics; Staden. Different software packages require different formats and these different output formats allowed easy transfer of sequence data from SeqEd to different applications, although text files were used in the vast majority of cases.

2.4.3 **EDITSeq of *Lasergene***

EDITSeq was the starting point for most uses of *LASERGENE.* The other modules (except SEQMAN and ABI sequence traces) essentially required files in EDITSeq format for creating their own documents e.g. for a MegAlign alignment. DNA and protein files were imported from most common formats.

i) Opening EDITSeq

EDITSeq was opened by clicking on 'Sequence Editing' and 'Analysis' from the ***Lasergene** Navigator*, or by double-clicking the application icon. By default (N) an empty DNA sequence window was presented. A double helix icon in the top left of the window, in the status bar, indicated that it was a DNA window. To change to a protein window 'New' > 'New Protein' was selected from the 'File' pull down menu. In a DNA window the 'Triplet Indicator' next to the DNA icon in the status bar indicated whether a selected range of bases was an ORF or not The left and right pointing arrows showed the direction to move to return to an ORF. The sequence length and subsequence range selected was indicated in the middle of the status bar.

ii) Entering and editing data

Sequence data were entered either by cutting and pasting or exporting text files created in SeqEd (2.4.2) or by typing. The sequence could be proofread by choosing 'Macintosh Voice' from the 'Digitizer' pull down menu. Alternatively, the 'Open Mouth' icon on the bottom left of the window was selected; the 'Open Hand' icon or 'No Sound' from the 'Digitizer' menu halted proofreading. The speed of proofreading and tones instead of voice could be controlled from the 'Digitizer' menu. A sequence subrange from the 'Edit' pull down menu of EDITSeq could be chosen when'Go to Position' was selected and the range required typed in the box as the starting and finishing base numbers separated by a comma e.g. 13,99 for subrange bases 13 to 99. The sequence could be formatted from the 'Edit' pull down menu (case, font, size, spacing or blocks of sequence characters). 'Reverse Sequence' or 'Reverse Complement' of the active sequence, was generated when the required subrange (or all of it with 'Select All' from the 'Edit' menu) was chosen. The reversed/complemented sequence was displayed in a new window.

iii) Opening and importing sequence documents

Selecting 'Open' from the 'File' menu displayed a standard dialogue box for opening a document. The document had to be in DNAStar format. If it was not 'Import' could be selected from the 'File' pull down menu and the type of sequence file to import from the right window pane of the dialogue box specified, and whether it was a DNA or protein file using the radio buttons under the window pane. Only files of the specified type were displayed in the left window pane; these could be imported by double clicking on the file name or using 'Import'. EDITSeq interpreted ASCII files with a *.seq* extension as nucleotides, and those with a *.pro* as proteins.

iv) Set Ends

The standard method for opening or specifying a defined subrange of a sequence in ***Lasergene*** was using the 'Set Ends' button in the 'Open' dialogue box. If 'Set Ends' was selected a window was presented where the sequence subrange could be entered in text fields or, alternatively, thumbwheels could be manipulated with the mouse pointer to select the range. The 'Other Strand' could be selected, and an 'Other Segment' button allowed the unselected portion of a sequence to be specified, rather than the selected part. If the 'word length' was clicked on, the sequence was set to its full range.

v) Exporting sequence documents

'Export' from the 'File' pull down menu saved the sequence document as ASCII text. The program did not allow export in any file format except as a text file containing both the sequence and comments. However, the beginning of the sequence field was marked by a pair of colons (::), which was a standard symbol indicating the start of ASCII sequence data and was recognised by many other programs. Alternatively, standard 'cut' and 'paste' were used to transfer sequence information without the comments

vi) Searching for Open Reading Frames (ORFs)

ORFs were found in a sequence document by choosing 'Find ORF' from the 'Search' pull down menu and clicking 'Find Next'. The genetic code used was specified from the 'Genetic Codes' submenu from the 'Goodies' pull down menu. ORFs located in this manner were translated into a protein document by choosing 'Translate DNA' from the 'Goodies' pull down menu. The comments pane of the newly translated DNA displayed a statistical analysis of the protein.

vii) Protein analysis

When a region of a protein was selected, its MW, charge and isoelectric point were displayed in the status bar of the window. These properties and an amino acid analysis could be displayed in a separate window by choosing 'Protein Statistics' from the 'Goodies' pull down menu. The selected region could be 'Reverse Translated' from the 'Goodies' pull down menu; also the genetic code used could be specified from 'Genetic Codes' submenu. The genetic code was edited using 'Edit Selected Code'. This was useful for eliminating ambiguity in reverse translation.

2.4.4 **MegAlign of Lasergene**

The degree of similarity between different sequences, that is the extent of conserved nucleotide or amino acid residues, is used to make inferences about whether they share common ancestry, or have common structures and functions that may have arisen through convergent evolution. MegAlign was used to carry out alignments between nucleotide and their derived amino acid sequences as follows:

i) Opening MegAlign

MegAlign was opened by clicking on the 'Multiple Sequence Alignment' button on the Lasergene Navigator screen. Selecting 'New' from the 'File' pull down menu created a new worktable. Selecting 'Open' from the 'File' pull down menu opened a preexisting project. The Worktable was divided into three panes; the middle and right ones were scrollable windows that by default showed the beginning and end of the sequence. The left windowpane showed the sequence names and had palette tools. Sequences were added to an alignment by selecting 'Enter Sequences from the 'File' menu and using the '>> Add >>' button which added sequences from the left window pane to the right. A folder of sequences could be added. The sequences had to be DNAStar files, (EDITSeq documents, 2.4.3). After clicking 'Done' a Worktable appeared filled with the chosen sequences. Both DNA and protein sequences could be added to the same Worktable. If this was done the DNA sequences were automatically translated to amino acids, starting at the first base, whether or not it was in-frame. Double clicking a highlighted sequence name allowed a subrange of that sequence to be selected (see below). A long single click allowed the name to be changed from the keyboard; click and drag was used to move the sequence in the Worktable.

ii) Subranging sequences

The subrange of a sequence was set from the 'Options' pull down menu by selecting 'Set Sequence Limits' or was created in an EDITSeq document (2.4.3). A subsequence could also be selected choosing 'By Coordinates' from the 'Set Sequence Limits' selection under the 'Options' pull down menu. A standard LASERGENE thumbwheel dialogue box appeared (see section 2.4.3 (iv)).

iii) Aligning the sequences

'By Clustal Method' or 'By Jotun Hein Method' were chosen from the 'Align' pull down menu. The Jotun Hein method was chosen if the sequences were related by common descent, otherwise (and almost always) Clustal was used. The alignment parameters (k-tuple, gap penalty, gap penalty length, window and scoring diagonals) could be set from 'Method Parameters' under the 'Align' pull down menu.

After the alignment was carried out the project was saved and the *Alignment Report, Sequence Distances, Residue Substitutions* and *Phylogenetic Tree* were selected from the 'View' pull down menu and examined. 'Save As' under the 'File' menu was used to export the alignment as a PAUP/Nexus or GCG document from the 'Format' submenu of the 'Save As' dialogue box. To change the residues displayed in an alignment 'New Decoration' was selected from the 'Options' pull down menu. Either 'Hide' or 'Shade' the residues that match was a good way of emphasising sequence similarity in the alignment for figures and diagrams. The residues could be compared to a consensus or to an individual sequence from a pull-down menu in the 'Alignment Decoration' dialogue box.

'New Consensus' from the 'Options' pull down menu allowed the rules for defining a consensus sequence to be changed. The consensus could be set to be when all residues were identical or when a specified number matched. Also, the consensus could be set as a template group of amino acids.

iv) Realigning Residues

An alignment could be manually edited using the palette tools on the left side of the Worktable. A residue was selected by positioning the cursor over it; the cursor changed to a square tool which could be used to highlight individual residues. The sequence containing the selected residues was moved by the 'Straighten Columns', 'Shuffle Right' and 'Shuffle Left' palette tools. For example, by selecting one residue and the gap next to it with the square tool, the residue could be shifted into the gap by the appropriate 'Shuffle' palette button. 'Go to Position' from the 'Edit' pull down menu was used to move to a single residue or range of residues in one sequence, or in the consensus. Gaps and sequence disagreements were found using 'Find Disagreement' from the 'Edit' pull down menu.

v) Pairwise Alignments

Pairs of sequences were compared in MegAlign by selecting them by clicking once on their names in the sequence names field. The 'One Pair' option in the 'Align' pull down menu then became active. This lead to a submenu containing four pairwise methods. The Lipman-Pearson method was used for protein alignments; the Wilbur-Lipman and Martinez/Needleman-Wunsch are for DNA alignments; while the DotPlot method could be used for either DNA or protein. The Needleman-Wunsch algorithm is the basis for many alignment programs, both protein and DNA. The Wilbur-Lipman was used for global alignments, the Martinez/Needleman-Wunsch for local ones. For example, if searching a large sequence for similarity to a primer sequence, to which it is related but not identical, Martinez/Needleman-Wunsch was the better choice. In practice it was often best to use both methods. When an alignment method was chosen from the 'One Pair' submenu, a parameter dialogue box appeared. The default parameters were used for a first alignment. The effect of reducing the penalties was investigated afterwards. Higher gap penalty figures produced a more stringent alignment with a lower similarity index score.

After the alignment was completed, the Alignment view appeared. The alignment could be formatted with the Alignment Color button (a box with a cross in it) on the top left of the palette of the Alignment view. The Alignment view showed the subrange of the sequence that had formed the alignment. MegAlign did not display the entire sequence that went into the alignment, only that part of it that had significant similarity to the other sequence. The 'Show Context' box in the Alignment Color menu caused the display to change to show the complete alignment between both sequences.

The similarity index, and number and length of gaps were also shown in the Alignment view. The similarity index referred only to the aligned part of the two sequences, not to the entire sequences. It was calculated from the number of matching residues divided by the sum of the number of matching residues, plus the number of mismatching residues and plus the number of gaps. Since the number of gaps in the alignment was a function of the parameters chosen, the similarity index was only a relative value, not an absolute one. This could be seen by evaluating a subalignment of the aligned sequences. This was carried out by drag clicking on a region of the alignment, producing a highlighted sub-region. The similarity index for the selected region was displayed beneath that for the two aligned parent sequences. The 'Evaluate Subalignment' button on the palette now became active. This was selected by clicking and choosing the appropriate protein or DNA alignment method from the submenu. The parameter dialogue box appeared displaying the subranges for the two sequences. If the gap penalty or length parameters were changed and the subalignment reevaluated, then a new alignment window appeared. This usually resulted in a different pattern of gaps in the subalignment compared to the parent alignment, with a changed similarity score.

Amino acid groupings in MegAlign

There are four template groups of amino acids in MegAlign: functional, structural, chemical and charge. There are four functional groups of residues: a - acidic (DE), b - basic (HKR), f - hydrophobic (AFILMPVW) and p - polar (CGNQSTY); three structural groups: a - ambivalent (ACGPSTWY), e - external - (DEHKNQR) and i - internal (FILMV); eight chemical groups: a - acidic (DE), b - basic (HKR), f - aliphatic (AGILV), m - amide (HQ), o - aromatic (FWY), h - hydroxyl (ST), i - imino (P) and s - sulphur (CM); and three charge groups: a - acidic (DE), b - basic (HKR) and o - neutral (ACFGILMNPQSTVWY).

2.4.5 **Creation of sequence text files suitable for ClustalV analysis**

After assimilation using SeqEd v 1.03, sequences from SeqEd were either exported from SeqEd or cut and pasted into one text file in PIR format by starting each new sequence with '**>D1;**' (see below), typing the sequence name (ten characters or less) and beginning the sequence data two lines below the sequence name, as the line in between is reserved for comments, and finishing the individual sequence with an asterisk:

> = start reading
D = DNA sequence (P = protein sequence)
1 = linear (0 = circular)
; = label follows

2.4.6 **ClustalV Alignment**

This application ran with a PC interface and consequently each command had to be followed by the return or 'enter' key. ClustalV (110) was opened by double clicking on the icon. Option 1 (*Sequence input from disk*) was selected and the name of the text file to be aligned was typed in. The file to be analysed had to be in the same folder as the Clustal application or it would not be recognised. Option 2 was selected (*Multiple sequence alignments*) and option 4 toggled to produce a fast/approximate alignment. Option 9 was selected (*Output format options*) and option 4 toggled to turn on the PHYLIP output option. After pressing 'enter' to return the screen to the previous menu, option 1 was selected (*Do complete alignment now*). Alignment of HIV-1 gp120 sequences by ClustalV required the insertion of a large number of gaps due to length variation between individual samples. This complicated the interpretation of the phylogenetic relationships between individual species because there is, at present, no consensus on the correct interpretation of such gaps. Gaps in an alignment were effectively removed by giving any column including a gap zero weight and any column of data which included a character from every species in the alignment a weight of 1. In addition, alignments were edited by hand to align gaps to codon boundaries. Weights were invoked by placing a W on the first line of the file. The weights were then specified by a line or lines which started with W and then had enough characters or blanks to complete the full length of a species name.

2.4.7 Sequence analysis using the PHYLIP suite of programs

After alignment in ClustalV (110), the sequences were transferred to the Phylip suite of programs for the creation of distance matrix files and phylogenetic analysis (67) and subsequent generation of phylogenetic trees, described in more detail below:

Weighting

The weights option in PHYLIP allowed the specification of weights on individual characters. The weights caused a character to be counted as if it were n characters, where n is the weight. The values 0-9 gave weights 0 through 9, and the values A-Z give weights 10-35. By the use of weights overwhelming weight could be given to some characters and others dropped from analysis. In the molecular sequence programs only two values of the weights, 0 or 1, were allowed for nucleotides (see 2.4.6).

Bootstrapping

Bootstrapping was used as a resampling method to create new data sets by sampling N characters randomly with replacement, so that the resulting data set has the same size as the original, but some characters had been left out and others duplicated. The method assumed that the characters evolved independently, an assumption that may not be realistic for many kinds of data. However, this resampling method was the method of choice and is routinely used for statistical analysis of the sequence data sets used in phylogenetic analysis of HIV-1 by the majority of groups working in this field (4, 14, 119, 129, 199). For a review see (191)

ClustalV was used (see above) to align the sequences with the PHYLIP output format option selected, creating a '.phy' file which could then be used in subsequent manipulations in PHYLIP. All PHYLIP 3.5c applications expect to analyse a file called 'infile' and output data to a file called 'outfile' and, depending on the application, a treefile. Therefore, it was important to rename any files generated as the 'infile', 'outfile' or 'treefile' were repeatedly overwritten. The sequence alignment was then opened in **Seqboot** which read in the data set and produced multiple data sets from it by bootstrap resampling. The 'multiple data sets' option was usually set at 100. The 'multiple data sets' option had to be selected and the number of data sets specified in every PHYLIP application from Seqboot onwards. The output file from Seqboot was then renamed 'infile' and processed by **DNAdist** to generate distance matrices between species, based on Kimura's 2-parameter method (139) which is a method for estimating the number of nucleotide substitutions. Nucleotides A and G (purines) are structurally similar, as are T and C (pyrimidines). Base substitutions between purines and between pyrimidines are called transitions, while those between purines and pyrimidines are called transversions. It has been shown that for many genes the rate of transitional changes is considerably higher than that of transversions (90). There are two other distance methods in DNAdist which can be used: the Jukes-Cantor formula (131) which assumes that there is independent change at all sites with equal probability; and a maximum likelihood method using a similar model to that employed in DNAml , although this method is extremely slow with large data sets such as those described in this thesis. Distance matrices, composed of percentage pairwise differences (or distances) between individual molecules, generated by **DNAdist** from PHYLIP were used to calculate standard deviations from the mean for different groups of sequences. The distances were then used in the **Fitch** distance matrix program which estimated phylogenies from matrix data under the 'additive tree model', according to which the distances are expected to equal the sums of the branch lengths between the species. The Fitch program uses the Fitch-Margoliash criterion (74) and does not assume an evolutionary clock. Although it is recommended to carry out 1000 bootstraps for data sets, this proved impractical and very time consuming for the large data sets used in this work. One hundred data sets appeared to be an adequate substitution as bootstrapped data sets are randomly generated. A consensus tree was generated from the 100 trees generated in the treefile outputted by Fitch in **Consense**. The numbers at the nodes of the consensus tree indicated the number of times the group consisting of the species which were to the right of that node occurred among the 100 trees sampled, and thus acts as a crude statistic for the 'likelihood' of a particular grouping. Using these methods we were able to analyse the sequences generated in order to derive phylogenetic relationships. These consisted both of grouping of like sequences together (transmission events etc.) and identification of distinctions between sequences (identification of subtypes etc.).

RESULTS AND DISCUSSION

CHAPTER 3

MONITORING DIVERSITY AND SUBTYPING OF HIV-1 IN THE U.K. USING a) gp120 and b) p6/protease SEQUENCES: AT LEAST SIX HIV-1 SEQUENCE SUBTYPES (A, B, C, D, A/E, G) OCCUR IN ENGLAND

Background

To date, HIV-1 infection in the U.K. has been mainly of subtype B (see chapter 1), transmitted homosexually and by injecting drug use, whereas the African epidemic has equally affected both men and women who are likely to be infected with a non-B subtype virus. Soto-Ramirez and colleagues suggested that some subtypes have a particular tropism for cell types exposed during vaginal transmission (245). The possibility of a heterosexually transmitted epidemic in the U.K. may be increased by the introduction of strains of HIV-1 subtypes other than B, for example the African subtypes. There is a potential for these diverse subtypes of HIV to spread into groups who may not consider themselves to be at risk of HIV infection. This study aimed to determine the subtype of currently circulating strains, based on both a) *env* (gp120) and b) *gag/pol* (p6/protease) sequences amplified from lymphocytes, and hence to monitor in finer detail the spread of HIV-1 in the U.K. and to determine the optimum region of gp120 for subtyping.

a) gp120

Results

Sequences were determined for *env* (gp120 or partial gp120) regions of DNA amplified from lymphocytes prepared from specimens submitted to this laboratory for transmission studies, confirmation of HIV infection or subtyping of the virus (table 3.1), using the limit dilution method described in sections 2.3.4, 2.3.5, 2.3.7-10. A typical agarose gel, depicting an initial limit dilution of gp120 is shown in figure 3.1.

Figure 3.1

A typical agarose gel showing an initial limit dilution of gp120 (2.3.8). Lanes 1-4 are 4 secondary PCR amplicon replicates, initially amplified with 1 μL of lymphocyte lysate from CPHL14 (see table 3.1). PCR products in lanes 5-8 were initially amplified with the equivalent of 0.2 μL lysate, lanes 9-12 with the equivalent of 0.04 μL lysate and lanes 13-16 with the equivalent of 0.008 μL of lysate.

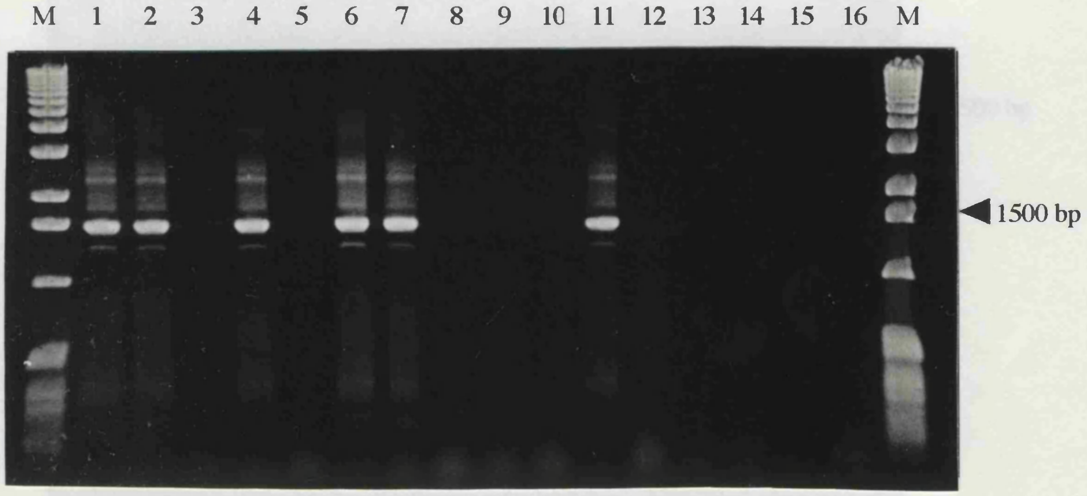


Figure 3.1 shows that three out of four tubes containing the first dilution were positive, 2 out of 4 from the second dilution were positive and 1 out of 4 from the third dilution was positive. Based on this result, 40 tubes of a 1 in 25 dilution were made and PCR carried out according to sections 2.3.5, 2.3.7-10. The PCR amplicons were run on an agarose gel, shown in figure 3.2.

Figure 3.2

A 1% agarose gel showing a 'bulk' gp120 dilution, as described in section 2.3.8.

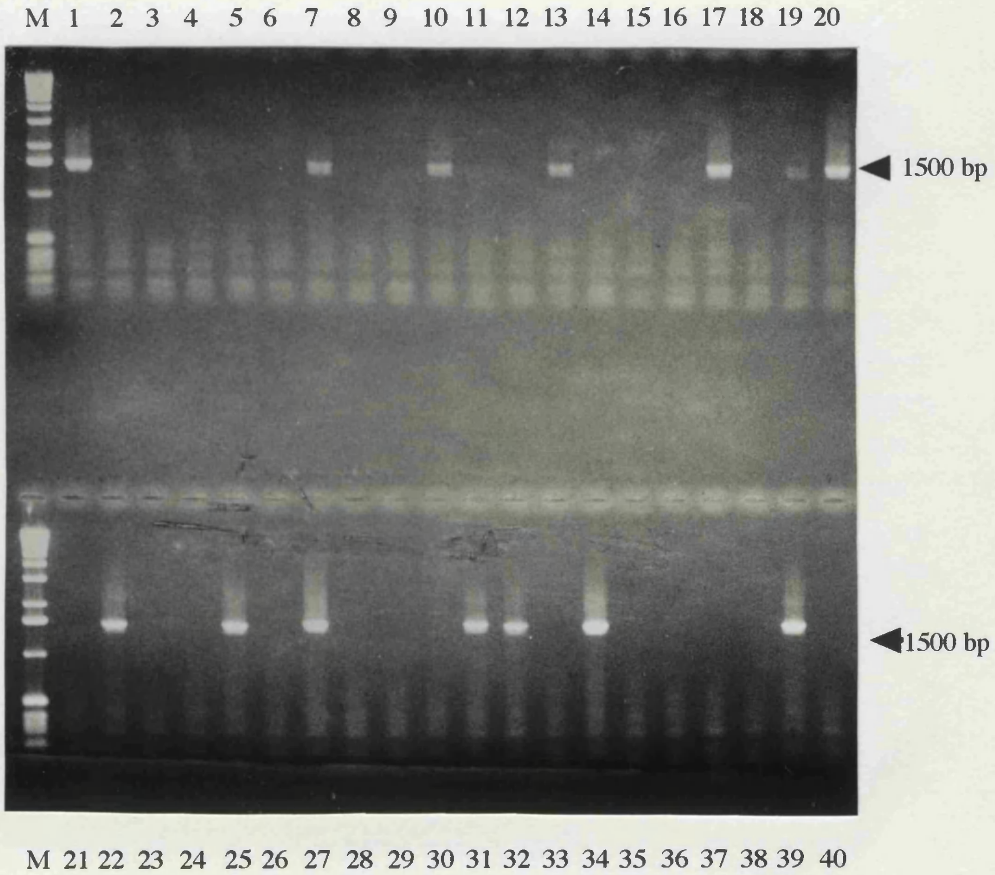

M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1500 bp
1500 bp
M 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

Figure 3.2 shows 13 out of 40 (0.325) PCR positive tubes with a 1 in 25 dilution of lysate. The faint band in lane 19 was ignored as it may have been carry over from lane 20. Table 2.3 (section 2.3.8) indicated that an observed frequency of 0.325 positives gives a proportion of approximately 80% of tubes containing single molecules.

Table 3.1

Clinical specimens

Patient no.
(CPHL no.
shown in bold)
	Reason for investigation
	T = transmission study
	R = Roche false negative
	S = subtyping
		Presumed
		place of infection
			Probable route
			of transmission
				gag/*env* subtype
				p6/gp120 seqs
				*p24 gag seqs

93-08020 (**1**)	T	U.K.	homosexual	B/B
93-17305 (**2**)	T	U.K.	heterosexual	B/B
94-4995 (**3**)	T	U.K.	heterosexual	B/B
93-43424 (**4**)	T	U.K.	heterosexual	D/D
93-43425 (**5**)	T	U.K.	heterosexual	D/D
94-27290 (**6**)	T	U.K.	sexual	B/B
94-24612 (**7**)	T	U.K.	heterosexual	B/B
94-26807 (**8**)	T	U.K.	heterosexual	B/B
94-27481 (**9**)	T	U.K.	heterosexual	B/B
94-47621 (**10**)	T	Rural Kenya/U.K.	heterosexual	G*/G
94-47622 (**11**)	T	Rural Kenya/U.K.	maternal	G*/G
95-24619 (**12**)	T	U.K.	homosexual	B/B
95-31191 (**13**)	T	U.K.	needlestick	B/B
96-10651 (**14**)	T	U.K.	heterosexual	B/B
93-00513 (**15**)	T	Africa	maternal	C/C
94-11643 (**16**)	T	Thailand	IVDU	A/E
93-33422 (**17**)	S	Cameroon	sexual	A/G
94-29517 (**18**)	S	n/k	maternal	D*/D
94-47971 (**19**)	R	Uganda/U.K.	heterosexual	D*/D
94-44501 (**20**)	R	Africa	maternal	A*/A
95-12310 (**21**)	R	U.K.	maternal	C*/C
94-12313 (**22**)	R	Somalia	heterosexual	A*/A
94-31296 (**23**)	R	n/k	IVDU	A*/n/a
93-8234 (**24**)	R	Africa/U.K.	maternal	A/D

n/k = not known
n/a = not amplifiable
* = partial p24 *gag* sequences (11)

i) subtypes deduced from all available gp120 sequences

A phylogenetic tree was constructed using the MegAlign neighbour-joining method described in section 2.4.4, using all the available gp120 sequence data generated for this thesis (figure 3.5). The full alignment of these sequences is shown in figure 3.3. The phylogenetic tree containing translated protein sequences is shown in figure 3.6. The alignment of the translated protein sequences is shown in figure 3.4. Though rerooted by an outgroup, the trees shown in this thesis are unrooted. The large number of sequences present in the tree makes it unfeasible to label an unrooted tree.

The alignment in figure 3.3 (opposite) was done on the DNA sequences, and not on the protein sequences with subsequent back translation from amino acids to nuclcotides. Thus the alignment is not in frame, and it does not map to the protein alignment in figure 3.4 which was done separately. However, it should be noted that the T_S/T_V and D_S/D_n ratios calculated later were done on inframe nucleotide alignments.

Figure 3.3

Alignment of all CPHL sequences

```
GCT--AGAA-AAT-TGTGGGTCACAGTCTATTATGGGGTACCTGTGTGGAAAGAAGCAACCACCACTCTA  Consensus
                       C
                 |BstEII|
.---------------------------------------------------------------------  CPHL1 cDNA
...AC....A...-........................................C...............  CPHL1 1
...AC....A...-........................................C...............  CPHL1 18
...AC....A...-........................................C...............  CPHL1 19
...AC....A..C-......................A.................C...............  CPHL1 4
...AC....A...-......................A.................C...............  CPHL1 43
...AC....A...-........................................C...............  CPHL1 7
...AA....C...-..G..T..................................................  CPHL2 11
...AC....C...-........................................................  CPHL2 18
...AC....C...-........................................................  CPHL2 25
...AC....C...-........................................................  CPHL2 3
...CA....A.----........................................A..............  CPHL7 13
...CA....A.----.....................A.......................T.........  CPHL7 17
...AA....A...-.................................................T......  CPHL8 2
...AA...CA.G.-.................................................T......  CPHL8 5
...AA....A...-.....................................................T..  CPHL6 3
...AA....A.G.-...............................A........................  CPHL6 41
...AA....A.G.-........................................................  CPHL6 6
...AC....A.G.-........................................................  CPHL9 1
...AC....A.G.-....................................................C...  CPHL9 17
...AC....A...-........................................................  CPHL9 6
...GA....C.GG-..........C.............................................  CPHL3 48
A.CGT----------------------...........................................  CPHL3 5
...GT....C.G.-........................................................  CPHL3 6
----------------------------------.......T........G........A.........   CPHL12
----------------------------......................G........A..........  CPHL13 2
-----------------------...........................G........A..........  CPHL13 8
----------------------------------------T.............................  CPHL14 15
...GC....A...-........................................................  CPHL14 37
...GC....A...-........................................................  CPHL14 44
...G-.AG.C...C.............T...........G.....A.....G..................  CPHL4 2
...G-.AG.C...C.............T...........G.....A.....G..................  CPHL4 35
...G-.AG.C...C.............T...........G.....A.....G..................  CPHL4 6
.-------------------------.............G.....A.....G..................  CPHL5 1
...G-.A---...C.......T.....T...........G.....A.....G..................  CPHL5 10
...G-.A---...C.............T...........G.....A.....G..................  CPHL5 12
.T-------------------..........................................T......  CPHL19
----------------------------------------------------------------------  CPHL18
.T.GC..G.C...-.............T..........................................  CPHL24
----------------------------------------------------------------------  CPHL20
-----------------------....---.......A.-------------------------------  CPHL16
...-C.A..C..CT....................................GG..T...GAT...C....C  CPHL17
----------------------------------------------------------------------  CPHL22
.-----------------...............................G....C...GAT.........  CPHL10 2
.-----------------...............................G....C...GAT.........  CPHL10 3
----------------------------------------------------------------------  CPHL10 7
T-----------------...............................G....T...GAT.........  CPHL11 1
----------------------------------------------------------------------  CPHL11 2
------------------...............................G....T...GAT.........  CPHL11 5
----------------------------------------------------------------------  CPHL15
----------------------------------------------------------------------  CPHL21
```

```
TTTTGTGCATCAGATGCTAAAGCATATGATACAGAGGTACATAATGTTTGGGCCACACATGCCTGTGTAC   Consensus

----------------------------------------------------------------------   CPHL1 cDNA
...........................TC................A........................   CPHL1 1
...........................TC................A........................   CPHL1 18
...........................TC................A........................   CPHL1 19
...........................TC................A........................   CPHL1 4
............................T................A........................   CPHL1 43
...........................TC................A........................   CPHL1 7
....................G......A.G......A.G...............................   CPHL2 11
....................G......A.G......A.G...............................   CPHL2 18
....................G......A.G......A.G...............................   CPHL2 25
..............C.....G......A.G......A.G...............................   CPHL2 3
......................................................................   CPHL7 13
....................G.................................................   CPHL7 17
......................................................................   CPHL8 2
..............................G...............................T.......   CPHL8 5
............................C.........................................   CPHL6 3
............................C.........................................   CPHL6 41
............................C.........................................   CPHL6 6
......................................................................   CPHL9 1
......................................................................   CPHL9 17
......................................................................   CPHL9 6
......................................................................   CPHL3 48
......................................................................   CPHL3 5
......................................................................   CPHL3 6
.................C.............................C......................   CPHL12
.................C....................................................   CPHL13 2
.................C....................................................   CPHL13 8
....................G........G......A.G...............................   CPHL14 15
....................G......A.G.....AA.G...............................   CPHL14 37
....................G......A.G.....AA.G...............................   CPHL14 44
.....................T.....A.A.......C.......A.C....................G.   CPHL4 2
.....................T.....A.A.......C.......A.C....................G.   CPHL4 35
.....................T.....A.A.......C.......A.C....................G.   CPHL4 6
.....................T....CA.A.......C.......A.C....................G.   CPHL5 1
.....................T.....A.A.......C.......A.C....................G.   CPHL5 10
.....................T.....A.A.......C.......A.C....................G.   CPHL5 12
.....................T.....C..G......C.........C.....T..............G.   CPHL19
----------------------------------------------------------------------   CPHL18
.....................T.....AGAG......C.........C.....T................   CPHL24
----------------------------------------------------------------------   CPHL20
----------------------------------------------------------------------   CPHL16
...........T.....C.........AG...T..AAA.........C.....T..C.............   CPHL17
----------------------------------------------------------------------   CPHL22
...........T....................T..AAGC..............T................   CPHL10 2
...........T....................T..AAGC..............T................   CPHL10 3
----------------------------------------------------------------------   CPHL10 7
...........T....................T..AAGC..............T................   CPHL11 1
----------------------------------------------------------------------   CPHL11 2
...........T....................T..AAGC..............T................   CPHL11 5
----------------------------------------------------------------------   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
CCACAGACCCCAACCCACAAGAAGTAGTATTGGAAAATGTGACAGAAAATTTTAACATGTGGAAAAATAA   Consensus

----------------------------------------------------------------------   CPHL1 cDNA
.............T.................................T..............G.......   CPHL1 1
.............T.................................T..............G.......   CPHL1 18
.............T.................................T..............G.......   CPHL1 19
.............T.................................T......................   CPHL1 4
.............T.................................T......................   CPHL1 43
.............T.................................T..............G.......   CPHL1 7
.......................A.......................G.A...G................   CPHL2 11
.......................A.......................G.A..................G.   CPHL2 18
.......................A.......................G.A....................   CPHL2 25
.......................A.......................G.A....................   CPHL2 3
....G............................G....................................   CPHL7 13
....G............................G....................................   CPHL7 17
..........................AG............................GC..........G.   CPHL8 2
.................................G........................A.........G.   CPHL8 5
.......T...............A.........G....................................   CPHL6 3
.......................A.........G....................................   CPHL6 41
.................................G....................................   CPHL6 6
....................A...........A.....................................   CPHL9 1
................................A.....................................   CPHL9 17
.......T........................A.....................................   CPHL9 6
.................................G.............G......................   CPHL3 48
...............................................G......................   CPHL3 5
...............................................G......................   CPHL3 6
...........G...........A...A..........................................   CPHL12
...........G...........A...A..........................................   CPHL13 2
...........G...........A...A..........................................   CPHL13 8
.......................A...A....A..............G.....G................   CPHL14 15
.......................T...A....A................A...G................   CPHL14 37
.......................T...A....A................A...G................   CPHL14 44
....G..................A...A.....................C....................   CPHL4 2
....G..................A...A..........................................   CPHL4 35
....G..................A...A..........................................   CPHL4 6
....G..................A...A.....................C....................   CPHL5 1
....G..................A...A.....................C....................   CPHL5 10
....G..................A...A.....................C....................   CPHL5 12
.......................A.GAA.C.A........C.......................G...G.   CPHL19
----------------------------------------------------------------------   CPHL18
..........................CAT...A................C.....T..............   CPHL24
----------------------------------------------------------------------   CPHL20
----------------------------------------------------------------------   CPHL16
......................GA..CCTC..........A............................G   CPHL17
----------------------------------------------------------------------   CPHL22
......................GT.G.ATC..A.......A.............................   CPHL10 2
......................GT.G.ATC..A.......A.............................   CPHL10 3
----------------------------------------------------------------------   CPHL10 7
......................GT.G.ATC..A.......A.............................   CPHL11 1
----------------------------------------------------------------------   CPHL11 2
.......................T.G.ATC..A.......A.............................   CPHL11 5
----------------------------------------------------------------------   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
CATGGTAGAACAGATGCATGAGGATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTAAAATTA  Consensus

-----------------------------------------.............................  CPHL1 cDNA
.................................C.................G..................  CPHL1 1
......................................................A..............G  CPHL1 18
......................................................................  CPHL1 19
.....................................................................G  CPHL1 4
.....................................................................G  CPHL1 43
......................................................................  CPHL1 7
......G.......................T.......................................  CPHL2 11
......G.......................T.......................................  CPHL2 18
......G.......................T.......................................  CPHL2 25
......G.......................T.......................................  CPHL2 3
..............................T.................T...........C.........  CPHL7 13
..............................T.............................C.........  CPHL7 17
......................................................................  CPHL8 2
......................................................................  CPHL8 5
......................................................................  CPHL6 3
......................................................................  CPHL6 41
................................................T...........C.........  CPHL6 6
..................................C.........................C.........  CPHL9 1
.........C............................................................  CPHL9 17
.........C..................................................C.........  CPHL9 6
................................................T.....................  CPHL3 48
................................................T.....................  CPHL3 5
................................................T.....................  CPHL3 6
................................A.....................................  CPHL12
................................N.....................................  CPHL13 2
................................A.....................................  CPHL13 8
......G...........G........C..T...C........G..........................  CPHL14 15
......G.......................T..................T....................  CPHL14 37
......G.......................T.......................................  CPHL14 44
......G.....A.........................................A............C..  CPHL4 2
......G.....A...................................T.....A............C..  CPHL4 35
......G.....A...................................T.....A............C..  CPHL4 6
......G.....A.........................................A..C........G...  CPHL5 1
......G.....A....................................TC...AG.C........G...  CPHL5 10
......G......................................C...TC...AG.C........G...  CPHL5 12
......G..G............................................A...............  CPHL19
----------------------------------------------------------------------  CPHL18
......G..G...........A................................A...............  CPHL24
----------------------------------------------------------------------  CPHL20
------------------------------------------C.....T.................G...  CPHL16
T...........A..............................G..........A..T.........C..  CPHL17
-------------------------------------------------..GC.................  CPHL22
.......C....A....................C.........G..G.A.....A...........GC..  CPHL10 2
.......C....A....................C.........G..G.A.....A...........GC..  CPHL10 3
----------------------------------------------------------------------  CPHL10 7
............A....................C.........G..G.A.....A...........GC..  CPHL11 1
----------------------------------------------------------------------  CPHL11 2
............C....................C.........G..G.A.....A...........GC..  CPHL11 5
---------------------------------------------------------.........G..G  CPHL15
----------------------------------------------------------------------  CPHL21
```

```
ACCCCACTCTGTGTTACTTTAAATTGCACTGA------T-------A-----------------CTAATG   Consensus

..............................A.TTTGA------ATG.TACTAATACCACTAGTG....GA   CPHL1 cDNA
..........................TG..A.TTTGAC.TTAAATTGCACTAAGAGGAATGGTA..C..A   CPHL1 1
..............................A.T------------------------------------A   CPHL1 18
..............................A.TTTGA------ATG.TACTAATACCACTAGTG....GA   CPHL1 19
..............................A.T------------------------------------A   CPHL1 4
..............................A.---------------------TGCCACTAGTG.....A   CPHL1 43
..............................A.GTTGA------ATG.TACTAATACCACTAGTG....GA   CPHL1 7
....................G.G.......A.------.GTAAA------------------------..   CPHL2 11
....................G.........A.------.GTAAA------------------------..   CPHL2 18
....................G.........A.------.GTAAA------------------------..   CPHL2 25
....................G.........A.------.GTAAA------------------------..   CPHL2 3
.......................C........------.TTGAAGA.TAATGGTACTA--------TG.-   CPHL7 13
.......................C........------.TTGAAGA.TAATGGTACTA--------.G.-   CPHL7 17
.......................C........------.TACAATA.TACTGATACTGCTA---...G..   CPHL8 2
.......................C........------.TACAATA.TACTAATTC--------------   CPHL8 5
.......................C........------.TATAATA.TACTAATCCCA--------.---   CPHL6 3
.......................C........------.TATAATA.TACTAATTCCA--------.---   CPHL6 41
.......................C........------.GTGAATG.TACTAATTCCA--------.---   CPHL6 6
.......................C........------.---------ACTAATTCCA--------TG--   CPHL9 1
.......................C........------.---------ACTAACTCCA--------TG--   CPHL9 17
.......................C........------.---------ACTAACTCCA--------TG--   CPHL9 6
..........................T.....------.T------.TGTGGGAAATACTA---.....A   CPHL3 48
..........................T.....------.T------.TGTGGGAAATACTA---.....A   CPHL3 5
..........................T.....------.T------.TGTGGGAAATACTA---.....A   CPHL3 6
..................C.............------.AAGTTAA.TATTATTAATACTA---....G.   CPHL12
................................------.AAGTTAA.TATTATTAATACTA---.....A   CPHL13 2
..................C.............------.AAGTTAA.TATTATTAATACTA---.....A   CPHL13 8
....................G.........A.------.GTAAA------------------------..   CPHL14 15
..................C.G..G......A.------.GTAAA------------------------..   CPHL14 37
..................C.G..A......A.------.GTAAA------------------------..   CPHL14 44
..............C...C...........A.--TGAA.G-G---A.CA------------GCA..GTG.   CPHL4 2
..............C...C...........A.--TGAA.G-G---A.CA------------GCA..GTG.   CPHL4 35
..............C...C...--..-...A---TGA-.G------.CA------------GCA..GTG.   CPHL4 6
..............C...C.............--TTAT.G-GGGGA.CT------------ACA..TCG.   CPHL5 1
..............C...C.............--TAAT.G-GGGGA.CT------------ACA..TCG.   CPHL5 10
..............C...C.............--TTAT.G-GGGGA.CT------------ACA..TCG.   CPHL5 12
..............C.....G..C........--CAGGAA-GAATG.-----------------------   CPHL19
----------------------------------------------------------------------   CPHL18
..............C........C........--TTGGAA-GAACA.TG------------CCA.....A   CPHL24
----------------------------------------------------------------------   CPHL20
..T..T.....C..............TG.---------------CA.GGCCAATTGGACCAATG.C..--   CPHL16
.....T.................C..T....TGTAAG-------------------GAATAACA..G...   CPHL17
.....T.....C..C..C........T.--..-------------------GATTCCAACAGTA.AG.CA   CPHL22
.....T.................C..T...A.TGTAA---------------------CTTATA......   CPHL10 2
.....T.................C..T...A.TGTAA---------------------CTTATA......   CPHL10 3
-----------------......C..T...A.TGTAA---------------------CTTGTG.....A   CPHL10 7
.....T.................C..T...A.TGTAA---------------------CTTATA......   CPHL11 1
--...T..........A......C..T...A.TGTAA---------------------CTTATA......   CPHL11 2
.....T.................C..T...A.TGTAA---------------------CTTATA......   CPHL11 5
..-...........C...........T..---------------AA.TATTGCCAAG---AACGGG..--   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
CCACTAATAATA-------------TACTAATACCAATAATAGTAGTGG-GAAACAATGGAGAA----G-   Consensus

......C..C..CCCC------TAG.GT........C..G.........G........ACG..GT-----   CPHL1 cDNA
...A.TGC.C..ATTTGGGGAATGA.G.........C..T.........A.....G..AC...GT-----   CPHL1 1
......T..C..------------G.G.........C..T.........A...C....AC....T-----   CPHL1 18
......C..C..CCCC------TAG.GT........C..G.........G........ACG..GT-----   CPHL1 19
......T..C..------------G.G.....C...C..T......G..A...C....AC...GT-----   CPHL1 4
......T..C.G------------G...........C..T.........A.....G..AC...GT-----   CPHL1 43
......C..C..CCCC------TAG.GT.......CC..G.........A........ACG..GT-----   CPHL1 7
.......G...G---------CTAC....T................C.AG....T.........A---.G   CPHL2 11
.......G...G---------CTAC....T................CAAGA...T.........A---.A   CPHL2 18
.......G...G---------CTAC....T................CAAGA...T.........A---.A   CPHL2 25
.......G...G---------CTAC....T................CAAG....T.........A---.A   CPHL2 3
...A........------------C.......G...CC.G......C..A.G.---------------.G   CPHL7 13
...A........------------C..A....G...CC.G.....AC..A.G.---------------.G   CPHL7 17
......G..G..AT---GGTACTGC.G..GG.G...C..G......C..A....TGT.A.....AAAA.G   CPHL8 2
-..A.G...C.GAT---GCCACTGG..A....G...C..G.........A....TGC.A.....AGAA.G   CPHL8 5
-TG...C.....------------C.......G...C..G.....TC.AG.G.GAG..A.....A---.G   CPHL6 3
-TG...C.G...------------C.......G...C..G.....TC..A.G.GAGG.A.....A---.G   CPHL6 41
-.-------...------------C.......G...C..G.....TC..A....AG..A....GA---.G   CPHL6 6
-----.......------------C..............G.....TC.AA.....G..A.....A---.G   CPHL9 1
-----.......------------C..............G.....TC..A.....G..A.....A---.G   CPHL9 17
-----.......------------C..............G.....TC..A.....G..A.....A---.G   CPHL9 6
...A.------------------AC........G..C..C....G.CAAT.G.-----------------   CPHL3 48
...A.------------------AC.......GG..C..C....G.CAAT.....T....G..---TA.C   CPHL3 5
...A.------------------AC....T...G..C..C....G.CAAT.....T....G.G---CA.G   CPHL3 6
TGGA.....G..GT---GGGG--------.----...........--..GTGG.TG.-.AG.GGTACA.G   CPHL12
...A.....T..GT---TTGG--------G.----...T......--AATTGG.TG----------CA.G   CPHL13 2
...A...A.G..GT---GGGG--------G.----..........--..GTGG.TG.-.AG.GGTACA.G   CPHL13 8
.....C.G...G---------CTAC....T..A...............AG.G..TG........A---.G   CPHL14 15
T....C.G...G---------CTAC....C................A.AG.G..T.......G.A---.G   CPHL14 37
T....C.G...G---------CTAC....C................C.AG.G..T.......G.A---.G   CPHL14 44
GG.AC.TC.---TTTTG---------------.GG..C.G...C.C--------..C.------------   CPHL4 2
GG.AC..C....CTTTAG---------TA....GGG.C.G...C.C--------..C.------------   CPHL4 35
G-.AC..C..G.CTT-AG---------TA....GG..C.G...C.C--------..C.------------   CPHL4 6
GG.GC..C.C--CCCTG------------A..--G..C.---AC.C--------..C.------------   CPHL5 1
GG.AC..C.C--CCCTG------------A..--G..C.---AC.C--------.CC.------------   CPHL5 10
GG.AC..C.C--CCTTG------------A..--G..C.---AC.C--------..C.------------   CPHL5 12
----------C.CTATCT----------....G---CC.C.GA.GC--------..TA------------   CPHL19
----------------------------------------------------------------------   CPHL18
...AC....CC.CTGCCA---------.....GT..CC.C..A.GA-----GG...TA------------   CPHL24
----------------------------------------------------------------------   CPHL20
----------------------CATA..CT..GT.CC...C.-..A.A.GA..TT--.A------ACA.A   CPHL16
TA.AG....GC.C------------------.GAGGTA....A.---------------..TGGCCGAAA   CPHL17
T..GC...GTC.CC---------------------------..C.C...---------------------   CPHL22
.A..C...TG..C------------------.GAG.....C.A---..----------......CAGA.A   CPHL10 2
.A..C...TG..C------------------.GAG.....C.A---..----------......CAGA.A   CPHL10 3
.A..C...TG..C------------------.GAG.......AC.C..---------..AGT.GCAGA.A   CPHL10 7
.A..C...TG..C------------------.GAG.....C.A---..----------......CAGA.A   CPHL11 1
.A..C...TG..C------------------.GAG.....C.A---..----------......CAGA.A   CPHL11 2
.A..C...TG..C------------------.GAG.....C.A---..----------......CAGA.A   CPHL11 5
----------------------TAT...CT.C.--------.-....A.CA------.-------GAG.G   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
AGAAATGAAAAACTGCTCTTTCAATATCACCACAGACATAAGAGATAAGGTGCAGAAAGAATATGCACTT   Consensus

-.................................AG.T.........GA...A.A...C...........   CPHL1 cDNA
-..C..............................AG.T.........GA...A.A...C...........   CPHL1 1
-..C..............................AG.T.........GA...A.A...C...........   CPHL1 18
-.......................A....G....AG.T...A.....GA...A.....C...........   CPHL1 19
-..C..............................AG.T.........GA...A.....C...........   CPHL1 4
-..C..............................AG.T.........GA...A.A...C...........   CPHL1 43
-.................................AG.T.........GA...A.A...C...........   CPHL1 7
......A..............................C................................   CPHL2 11
......A.................G............C................................   CPHL2 18
......A.................G............C................................   CPHL2 25
......A..............................C................................   CPHL2 3
..................................A........A.....T.T....G.............   CPHL7 13
.........G........................A........A.....T......G.............   CPHL7 17
......A.G........................A..................................C   CPHL8 2
......A.G........................A..............A..A........C........   CPHL8 5
...G..A..G.....T.........G........A............GAT....................   CPHL6 3
...G..A..G.....T.........G.....................G.T....................   CPHL6 41
...G..A..G...............G........A....G.......G.T....................   CPHL6 6
.........G...............G........A..............A....................   CPHL9 1
.........G...............G........A..............A....................   CPHL9 17
.........G...............G........A..............A....................   CPHL9 6
-...........................T.....A..T...........TAT.......T..........   CPHL3 48
............................T.....AG.T...........T.T.......T..........   CPHL3 5
............................T.....AG.T...........T.T.......T..........   CPHL3 6
......A.....T...........C......T..AGT...CA............T.GG.........T.G   CPHL12
......A.................C......T...G....CA.......AA...T..G.........T.G   CPHL13 2
......A.................C......T...G....CA.......AA...T.GG.........T.G   CPHL13 8
...................................GA..............A..................   CPHL14 15
....................................AG.............AA.................   CPHL14 37
....................................AG.............AA.................   CPHL14 44
-.G.....G...A............G.A.......CA............CAAA.AC...TGC........   CPHL4 2
-.G.....G...A............G.A.......CA............CAAA.AC...T.C........   CPHL4 35
-.G.....G...A............G.A.......CA............CAAA.AC...T.C........   CPHL4 6
-..C....T.....CA...G...G...T.......TAG..G........AAAA......TGAT.......   CPHL5 1
-..........................A.......TAG...........AAAA......TCAC.......   CPHL5 10
-..........................A......ATAG...........AAAA......TCCC.......   CPHL5 12
-..........................A........A..........GAAAAA..C...T.C........   CPHL19
----------------------------------------------------------------------   CPHL18
-.G.G....G.................A........AG.........G.AA.A..C...T..T.......   CPHL24
----------------------------------------------------------------------   CPHL20
T...G.A.G......T.....T.....G........A............AA......G.TCC.....A..   CPHL16
G..GC.A.CC.................A.A......A....A.....G.AG.A..........C..GA..   CPHL17
......A.............A......T.G......AC...........AGA...C...TC...T.GA.G   CPHL22
......A....................A........A.............A.G..G.......C..G...   CPHL10 2
......A....................A........A..........-.-A.-..........C..G...   CPHL10 3
...........................A........A............CA.A..........C..G...   CPHL10 7
...........................A........A............AA.A..........C..G...   CPHL11 1
...........................A........A............AA.A..........C..G...   CPHL11 2
...........................A........A............AA.A..........C..G...   CPHL11 5
....C.A.G...T............GCA........A............CAA.G......G.........   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
TTTTATAAACTTGATGTAGTACCAATAGATAATGA-----------ATA-T----ATAATA---------  Consensus

.......GC......A......A....A.......TAAAG------...A.AC------..---------  CPHL1 cDNA
.......GC......A......A.T..........TAAGG------...A.AC------..---------  CPHL1 1
.......GC......A......A.T..........TAATG------...A.AC------..---------  CPHL1 18
.......GC......A......A.T..........TAAGG------...A.AC------..---------  CPHL1 19
.......GC......A......A............TAAAG------...A.AC------..---------  CPHL1 4
.......GC......A......A.T..........TAATG------...A.AC------..---------  CPHL1 43
.......GC......A......A..........A.TAAAG------...A.AC------..---------  CPHL1 7
...............A.............A.....TAGGG------...G.G---......---------  CPHL2 11
.......C.......A.............A.....TAGGG------G..G.A---......---------  CPHL2 18
.......C.......A.............A.....TAGGG------G..G.A---......---------  CPHL2 25
...............A.............A.....TAGGA------...G.A---......---------  CPHL2 3
........G.....................G..----------------------......---------  CPHL7 13
..............................G..----------------------......---------  CPHL7 17
..C..............................ACTAATA------...C.ACTA..T...---------  CPHL8 2
.......GC......A....G..............TAAGG------.------AA......---------  CPHL8 5
...............A.............GG..----------------------......---------  CPHL6 3
..............................G.G----------------------......---------  CPHL6 41
..............................G..----------------------......---------  CPHL6 6
.....................AA.......---A---------------------....A.---------  CPHL9 1
.....................AA.......---A---------------------..G.A.---------  CPHL9 17
.....................AA.......---A---------------------..G.A.---------  CPHL9 6
.................................A.------------..C.ACCA...C..---------  CPHL3 48
.................................A.------------.G------...C.----------  CPHL3 5
.................................A.------------.G------...C.----------  CPHL3 6
......................A.....G.GGGA.TG------------------......---------  CPHL12
......................A.....G.GGAA.TG------------------......---------  CPHL13 2
......................A.....G.GGAA.TG------------------......---------  CPHL13 8
..G............A.............AG....TAAGG------...G.------....---------  CPHL14 15
......GC.......A...........A.A...A.---------------------....G---------  CPHL14 37
......GC.....................A...A.AAAGG---------------......---------  CPHL14 44
...............A...................TAATAGTGATA...G.ACCA...G..ACAAGAGTG  CPHL4 2
...............A...................TAATAGTGATA...G.ACCA...G..ACAAGAGTG  CPHL4 35
...............A...................TAATAGTGATA...G.ACCA...G..ACAAGAGTG  CPHL4 6
.......G.......A.............C.G------TAGTGATA...------------------GTT  CPHL5 1
.......G.......A.............C.G------TAGTGATA...------------------GGT  CPHL5 10
.......G.......A.............C.G------CAGTGATA...------------------GTT  CPHL5 12
...............A......A.......G..A.CAATAGT---AC-----------------------  CPHL19
----------------------------------------------------------------------  CPHL18
.................G....A...G...G....TAATAGT---AC..A.ACCA.CT...---------  CPHL24
----------------------------------------------------------------------  CPHL20
........G......A......A...---------TAAAGGTAATA...A.AATG...G..ATGAG----  CPHL16
..C..C....................---------------TGATA...G.---A..GCA.CTAG-----  CPHL17
........G.............A...------.AGTGAAAGTAGTA...A.ACCAGG.G..ATGATAGTG  CPHL22
..C......A..........G.....---------------TAATGG..G.------.G..GTGA-----  CPHL10 2
..C......A..........G.....---------------TAATGG..G.------.G..GTGA-----  CPHL10 3
.........G..........G...C.---------------TAATA...A.CTTA......GTGA-----  CPHL10 7
..C......A..........G.....---------------TAATGG..G.------.G..GTGA-----  CPHL11 1
..C......A..........G.....---------------TAATGG..G.------.G..GTGA-----  CPHL11 2
..C......A..........G.....---------------TAATGG..G.------.G..GTGA-----  CPHL11 5
...............A.....T..C.---------TAATGGCAAC--------TC----..GTGAG----  CPHL15
----------------------------------------------------------------------  CPHL21
```

```
CTAGTTATAGGTTGATAAGTTGTAA--CACCTCAGTCATTA-CACAGGCCTGTCCAAAGGTATCCTTTGA   Consensus

...AA........A.G.........--..............-....................A.......   CPHL1 cDNA
...CA........A.G.........--..............-....................A.......   CPHL1 1
...CA........A.G.........--..............-................A...A.......   CPHL1 18
...CA........A.G.........--..............-................A...A.......   CPHL1 19
...AA........A.G.........--..............-....................A.......   CPHL1 4
...CA........A.G.........--...A..........-....................A.......   CPHL1 43
...AA........A.G.........--..............-....................A.......   CPHL1 7
....C....................--..............-......................T.....   CPHL2 11
....C....................--.......AC.....-................A.....T.....   CPHL2 18
....C....................--.......AC.....-................A.....T.....   CPHL2 25
....C....................--.......A......-......................T.....   CPHL2 3
.........................--..............-.................A..........   CPHL7 13
.........................--..............-.................A..........   CPHL7 17
.........................--..............-............................   CPHL8 2
.........................--..............-..........C.................   CPHL8 5
.........................--..........C...-............................   CPHL6 3
.........................--..........C...-............................   CPHL6 41
.........................--..............-............................   CPHL6 6
.........................--..............-.................A..........   CPHL9 1
.........................--..............-.................A..........   CPHL9 17
.........................--..............-.................A..........   CPHL9 6
.C.AC....................--..............-....................A.......   CPHL3 48
-----....................--..............-....................A.......   CPHL3 5
-----....................--..............-....................A.......   CPHL3 6
.C..C....................--..............-.....A......................   CPHL12
.C..C....................--..............-.....A......................   CPHL13 2
.C..C....................--..............-.....A......................   CPHL13 8
....C....................AC..............A......................T.....   CPHL14 15
....C....................--.......A......-......................T.....   CPHL14 37
....C....................--.......A......-......................T.....   CPHL14 44
.C.A.........A....A......--T.......C.....-.......T....................   CPHL4 2
.C.A.........A....A......--T.......C.....-.......T....................   CPHL4 35
.C.A.........A...CA......--T.......C.....-.......T....................   CPHL4 6
.C.A.........A....A......--T.......C.....-.......T....................   CPHL5 1
.C.A.........A....A....C.--T.......C.....-.......T....................   CPHL5 10
.C.A....T....A....A......--T.......C.....-.......T.......C.........C..   CPHL5 12
-C.AA.....A..A....A......--T.......C.....-.......G............A.......   CPHL19
----------------------------------------------------------------------   CPHL18
.C.AC.....A..A...CA......--T.......C.....-.......G.........A..A.......   CPHL24
----------------------------------------------------------------------   CPHL20
-----........A....A......--T..T..........A-G.....T.........A..........   CPHL16
----.......C.A....A......--TGTTA..ACTG...A-...A..T............A.......   CPHL17
..CAG.....AC.A....A......--T.......C...C.-.......T..C.................   CPHL22
----....GTAC.A..G.A......--T..A...AC.....A-......T...........GAAT.....   CPHL10 2
----....GTAC.A....A......--T..A...AC.....A-......T...........GAAT.....   CPHL10 3
----.....TAC.A....A......--T..A...AC.....A-......T............AGT.....   CPHL10 7
----....GTAC.A....A......--T..A...AC.....A-......T......C....GAAT..G..   CPHL11 1
----....GTAC.A....A......--T..A...AC.....A-......T...........GAAT.....   CPHL11 2
----....GTAC.A....A......--T..A...AC.....A-......T...........GAAT.....   CPHL11 5
-----.....A.......A...C..--T......AC...A.C-...A..............C..T.....   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
GCCAATTCCCATACATTATTGTGCCCCGGCTGGTTTTGCGATTCTAAAGTGTAATGATAAGAAGTTCAAT   Consensus

A.....................................................................   CPHL1 cDNA
......................A...............................................   CPHL1 1
......................A...............................................   CPHL1 18
......................A...............................................   CPHL1 19
A.....................................................................   CPHL1 4
A.....................................................................   CPHL1 43
A.....................................................................   CPHL1 7
......................A..........G......C....................G......G.   CPHL2 11
........................................C....................G......G.   CPHL2 18
........................................C....................G......G.   CPHL2 25
........................................C....................G......G.   CPHL2 3
T...G..........A......A............................................C..   CPHL7 13
T...G.................A...............................................   CPHL7 17
T.....................................................A...........T...   CPHL8 2
T.....................................................CA..........T...   CPHL8 5
T.......T.......................................A...........C.........   CPHL6 3
T...G...........................................A...........C.........   CPHL6 41
T...G...........................................A...........C.........   CPHL6 6
T...G.................................................................   CPHL9 1
T...G.................................................................   CPHL9 17
T...G.................................................................   CPHL9 6
T......................................................A......C.......   CPHL3 48
T......................................................A......C.......   CPHL3 5
T......................................................A......C.......   CPHL3 6
...................................A...A..............A.....A.........   CPHL12
...................................A...A....................A.........   CPHL13 2
...................................A...A..............A........A......   CPHL13 8
......................A...............T.C....................G......G.   CPHL14 15
.............T..........................C...........................G.   CPHL14 37
........................................C...........................G.   CPHL14 44
A..........................A.........................GA...............   CPHL4 2
A..........................A.........................GA...............   CPHL4 35
A..........................A.........................GA...............   CPHL4 6
A..........................A........C.............TC.GA...............   CPHL5 1
A..........................A.......................C.GA...............   CPHL5 10
C..........................A.......................C.GA...............   CPHL5 12
.........T..............T..A.....G.....A........A......A..............   CPHL19
---------------............A.....A..............A.....C...............   CPHL18
...........................A.....A..............A..............A......   CPHL24
----------------------------------------------------------------------   CPHL20
T........T............A.T..A.......A.......T...................T......   CPHL16
........................T..A.......A.................GC...............   CPHL17
.........T............A....A............C.............G....C...A......   CPHL22
C.......................T..A...........A..............G...GCAG........   CPHL10 2
C.......................T..A...........A..............G...GCAG........   CPHL10 3
C.......................T..A...........A..............G...GC..........   CPHL10 7
C.......................T..A...........A..............G...GCAG........   CPHL11 1
C.......................T..A...........A..............G...GCAG........   CPHL11 2
C.......................T..A...........A..............G...GCAG......C.   CPHL11 5
C........T..............T..A.......A...................A....T.CA.C....   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
GGAACAGGACCATGTAAAAATGTCAGCACAGTACAATGTACACATGGAATTAAGCCAGTAGTATCAACTC    Consensus

..............C.....................................GA........G.......    CPHL1 cDNA
..............C.....................................GA........G.......    CPHL1 1
..............C.....................................GA........G.......    CPHL1 18
..............C.....................................GA........G.......    CPHL1 19
..............C.....................................GA........G.......    CPHL1 4
..............C.....................................GA........G.......    CPHL1 43
..............C.....................................GA........G.......    CPHL1 7
....A...........C...................................G.................    CPHL2 11
....A...........C...................................G.................    CPHL2 18
....A...........C...................................G.................    CPHL2 25
....A...........C...................................G.................    CPHL2 3
..............G.G..........G..........................................    CPHL7 13
................C.....................................................    CPHL7 17
......................................................................    CPHL8 2
......................................................................    CPHL8 5
................C.....................................................    CPHL6 3
................G......T............................G.................    CPHL6 41
................G.....................................................    CPHL6 6
................C.........T...........................................    CPHL9 1
................C.........T...........................................    CPHL9 17
................C.........T...........................................    CPHL9 6
....A...........C...................................G.................    CPHL3 48
....A...........C...................................GA................    CPHL3 5
....A...........C...................................G.................    CPHL3 6
......................................C...........C.GA........G.......    CPHL12
..................................................C.GA.....G..G.......    CPHL13 2
......................................C...........C.GA........G.......    CPHL13 8
....A....A..........................................G.................    CPHL14 15
....A..A.GT.........................................G.................    CPHL14 37
....A..A.TT.........................................G.................    CPHL14 44
........C.....C........T...................................G..........    CPHL4 2
........C.....C........T...................................G..........    CPHL4 35
........C.....C........T........................G..........G..........    CPHL4 6
........C.....C........T...................................G..........    CPHL5 1
........C.....C........T...................................G..........    CPHL5 10
........C.....C........T...................................G..........    CPHL5 12
..G..G..T.....C.....C..............G...........G..............G.......    CPHL19
.....G..T.....C........T...........G...........G.........AC...G.......    CPHL18
..G.....T.......CG..C..............G...........G....G.........G.......    CPHL24
----------------------------------------------------------------------    CPHL20
..G.....G..................T..........C....................G..........    CPHL16
........G.....C.......................C...........C...................    CPHL17
........G.....C........T...T...............................G.C........    CPHL22
..........................T.................................A.........    CPHL10 2
..........................T.................................A.........    CPHL10 3
.........A................T....C............................A.........    CPHL10 7
..........................T.................................A.........    CPHL11 1
..........................T.................................A.........    CPHL11 2
..........................T.................................A.........    CPHL11 5
..G...........C..T.........................................G..........    CPHL15
----------------------------------------------------------------------    CPHL21
```

```
AACTGCTGTTAAATG-GCAGTCTAGCAGAAGAAGAGG---TAATAATTAGATCTGAAAATTTCACAGACA   Consensus

...............-....C................---....G..........CC.............   CPHL1 cDNA
...............-.....................---....G..........CC.............   CPHL1 1
...............-..................G..---....G..........CC.............   CPHL1 18
...............-.T...................---..G.G..........CC.............   CPHL1 19
...............-.....................---....G..........CC.............   CPHL1 4
...............-.....................---....G..........CC.............   CPHL1 43
...............-.....................---....G..........CC.............   CPHL1 7
...............-..............A......---..G.............G.............   CPHL2 11
...............-..............A.G....---..G.............G.............   CPHL2 18
...............-..............A......---..G.............G.............   CPHL2 25
...............-..............A......---..G.............G.............   CPHL2 3
.......A.......-.T...................---..G.......G...C........T.GA...   CPHL7 13
.G.....A.......-.T...................---..G.......G...C........T.GA..G   CPHL7 17
.......A.......-.T...................---..G.......G...C........T.G....   CPHL8 2
.......A.......-.T...................---..G..C........C........T.G....   CPHL8 5
....A..........-.T............A....T.---..G.......G...C..........G....   CPHL6 3
....A..........-.T.................T.---..G.......G...C..........G....   CPHL6 41
.......A.......-.T.............G...A.---..G.......G...C..........G....   CPHL6 6
...............-.T.................T.---..G...........C....C....G....   CPHL9 1
...............-.T.................T.---..G...........C....C....G....   CPHL9 17
........G......-.T.................T.---..G...........C....C....G....   CPHL9 6
...............-.....................---..G.............C..CC....GA...   CPHL3 48
...............-....................A---..G...C.........C..CC....GA...   CPHL3 5
...............-.....................---..G.............C..CC....GA...   CPHL3 6
...............-....................A---..G.................A..T.GA...   CPHL12
...............-....................A---..G.................A..T.GA...   CPHL13 2
...............-....C...............A---..G.................A..T.GA...   CPHL13 8
...............-.....................---................G.........A...   CPHL14 15
...............-....C................---................G.........A...   CPHL14 37
...............-....C................---................G.........A...   CPHL14 44
..........G....-....................A---.C..................A.........   CPHL4 2
..........G....-....................A---.C..................A.........   CPHL4 35
..........G....-....................A---.C..................A.........   CPHL4 6
..........G....-....................A---.C..................A.........   CPHL5 1
..........G....-....................A---.C..................A.........   CPHL5 10
..........G....-....................A---.C........G.........A.........   CPHL5 12
..T..T....G....-....C...............A---....................C.....A.T.   CPHL19
....AT....G....-....C...............A---....................C.....A.T.   CPHL18
..........G....-....................A---.....G.............GC.T.....T.   CPHL24
----------------------------------------------------------------------   CPHL20
..T............-...............G....A---.......C............C.....A...   CPHL16
........C...C.TC...............-.....---..G.G.......T.C.G..C.....C....   CPHL17
........C.G....-.T.....T.......G.....AAA....G...............A.T.......   CPHL22
..T.A..T..G....-.....T........AG...AA---....G..............CA.........   CPHL10 2
..T.A..T..G....-.....T........AG...AA---....G..............CA.........   CPHL10 3
..T.A..T..G....-.....T.............CA---....GG.............CA.........   CPHL10 7
..T.A..T..G....-.....T............GCA---....G..............CA.........   CPHL11 1
..T.A..T..G....-.....T............GCA---....G..............CA.........   CPHL11 2
..T.A..T..G....-.....T............GCA---....G..............CA.........   CPHL11 5
....A..........-.T..C..........G....A---..........C.........C.G.......   CPHL15
----------------------...............---....................C.G..CA.T.   CPHL21
```

```
ATGCTAAAACCATAATAGTACAGCTGAATGAATCTGTAGAAATTAATTGTACAAGACCCAACAACAATAC   Consensus

.........A..................C..........T....................G.........   CPHL1 cDNA
.........A..................C..........C....................G........G   CPHL1 1
.........A..................C..........T....................G.........   CPHL1 18
.........A..................C..........T....................G.........   CPHL1 19
.........A..................C..........T....................G.........   CPHL1 4
.........A..................C..........T....................G.........   CPHL1 43
.........A..................C..........T....................G.........   CPHL1 7
...T......T............G...........................T..................   CPHL2 11
...T......T............G......................G....T..................   CPHL2 18
...T......T............G......................G....T..................   CPHL2 25
...T......T............G......................G....T..................   CPHL2 3
.........TT...........T.....C.........A.....G................T........   CPHL7 13
.........TT.................C................................T........   CPHL7 17
..A.........................C...........................G....T........   CPHL8 2
..........T.................C................................T........   CPHL8 5
..A.............................A............................T........   CPHL6 3
.............................................................T........   CPHL6 41
......................T......................................T........   CPHL6 6
............................A...C.............C..............T........   CPHL9 1
............................A...C.............C..............T........   CPHL9 17
............................A...C.............C..............T........   CPHL9 6
............................C.C.......................................   CPHL3 48
............................C.C.......................................   CPHL3 5
............................C.C.......................................   CPHL3 6
.........A............T............A...................G.T............   CPHL12
.........A............T............A...................G.T............   CPHL13 2
.........A............T............A...................G.T............   CPHL13 8
...T......T............T.......................................C......   CPHL14 15
...T..................................................................   CPHL14 37
...T..................................................................   CPHL14 44
...T.....A...............T............TT.....C....G....G...TC...T.....   CPHL4 2
...T.....A...............T............TT.....C....G....G...TC...T.....   CPHL4 35
...T.....A............A..T............TT.....C....G....G...TC...T.....   CPHL4 6
...T.....A...............T............CC.....C....GT...G...T....T.....   CPHL5 1
...T.....A...............T............CC.....C....GT...G...T....T.....   CPHL5 10
...T.....A...............T............CC.....C....GT.......T....T.....   CPHL5 12
.........T...G........T..T.....G......C..........C.....G...T..........   CPHL19
...TA....T.............G.T...A.G......AC....G....C.T...G..............   CPHL18
.........................T.....G......AC.........C.....G...T....T..C..   CPHL24
----------...............T...A.G.............T......C.....T...........   CPHL20
....C..............G..C..T...A..........G..C........C......TC........T   CPHL16
...................G..AT.T..CAGT......AG............C.....T...........   CPHL17
....C....A.............G.T.C...GC.....C.............C......G..........   CPHL22
....C....T.........G.....T..AACCC..A..A.T....CG...GTC.................   CPHL10 2
....C....T.........G.....T..AACCC..A..A.T....CG...GTC.................   CPHL10 3
....C....T.........G.....T..A..CC..A..A.T....CG...GTC.................   CPHL10 7
....C....T.........G.....T..AACCC..A..A.T....CG...GTC.................   CPHL11 1
....C....T.........G.....T..AACCC..A..A.T....CG...GTC.................   CPHL11 2
....C....T.........G.....T..AACCC..A..A.T....CG...GTC.................   CPHL11 5
....C....TA...........C..T..................GTG.................T.....   CPHL15
..ATC.....A...........T..C...................TG.....G........T..T.....   CPHL21
```

```
AAGAAAAAGTATACATATAGGACC------AGGGAGAGCATTTTATACAACAGGAGAAATAATAGGAGAC   Consensus

.....G.......TC.........------...........GG..............G............   CPHL1 cDNA
.....G...GG..AC...G.....------........T..GG..............G............   CPHL1 1
.....G.......TC.........------...........GG..............G............   CPHL1 18
.....G.......TC.........------...........GG...........................   CPHL1 19
.....G.......TC.........------...........GG..............G............   CPHL1 4
.....G.......TC.........------...........GG..............G.........A..   CPHL1 43
.....G.......TC.........------...........GG..............G............   CPHL1 7
........................------.........................C.............T   CPHL2 11
........................------.........................C.............T   CPHL2 18
..............C.........------.........................C.............T   CPHL2 25
........................------.........................C.............T   CPHL2 3
.......G................------......T..T................G....G........   CPHL7 13
.......G................------......T..T..G.............G....G........   CPHL7 17
.......G......G...T....ATATAGG.......T.................A..............   CPHL8 2
...C...G........C.......------....G....................AG..........A..   CPHL8 5
.....G.G................------.........................C..............   CPHL6 3
.....G.G................------......G..................C..............   CPHL6 41
.....G.G................------......G....................C............   CPHL6 6
.......G..........G....T------.........................C..............   CPHL9 1
.......G..........G....T------........................................   CPHL9 17
.......G..........G....T------........................................   CPHL9 6
.......G.G....G.........------................G........AG.........G...   CPHL3 48
.......G................------................G...................G..T   CPHL3 5
.......G................------................G...................G..T   CPHL3 6
..................G....T------.............C.............C...........T   CPHL12
..................G....T------.............C.............C...........T   CPHL13 2
..................G....T------.............C.............C...........T   CPHL13 8
.CT......G....C.........------.....C.T.................CG.......AA...T   CPHL14 15
..............C.........------.........................C.............T   CPHL14 37
..............C.........------.........................C.............T   CPHL14 44
....C......C...C........------....CA....C.C...........GC..............   CPHL4 2
....C......C............------....CA....C.C...........GC..............   CPHL4 35
....C......C............------....CA....C.C...........GC..............   CPHL4 6
....C......C............------....CA....C.C.......TGA.G---............   CPHL5 1
....C......C............------....CA....C.C.......TGAAG---............   CPHL5 10
....C......C..G.........------....CA....C.C.......TGA.G---............   CPHL5 12
....C..G................------...AC.....C.C.T........---.G....C......T   CPHL19
....CT......T...........------.....A....C.C...G......---.C..G........T   CPHL18
.......G...CT.....G.....------....CA.....GG.........A---G....G........   CPHL24
..T....G..G.C...........------...ACA.A....C...........TCG.G.......G..T   CPHL20
.....C....T.CAGG........------...ACA..T...CC...A.......AGC....C......T   CPHL16
.....G.G................------...ACA...T..C...G.......T..T........G..T   CPHL17
..A..G.GT.CG.AT...GC..ATTGGACC...ACA...C..C...G...---ATA.C........G..T   CPHL22
...............CT.C..T..------...ACA..........G.......T..C...........T   CPHL10 2
...............CT.C..T..------...ACA..........G.......T..C...........T   CPHL10 3
...............CT....T..------...ACA..........G.......T..C...........T   CPHL10 7
...............CT.C..T..------...ACA..........G.......T..C...........T   CPHL11 1
...............CT.C..T..------...ACA..........G.......T..C...........T   CPHL11 2
...............CT.C..T..------...ACA..........G.......T..C...........T   CPHL11 5
..........G.TAGG........------...ACA......C...G.....AAT..C............   CPHL15
.......G....GAGG........C------..ACA.A....C...G......................T   CPHL21
```

```
ATAAGACAAGCACATTGTAACATTAGTAGAGCAAAATGGAATAACACTTTACAACAGGTAGCTAAAAAAT   Consensus

......A..............C...........G.C........A......TGG...A...T........   CPHL1 cDNA
......A..............C...........G.C........A......TGG...A...T........   CPHL1 1
......A..............C...........G.C...............AGG...A...T........   CPHL1 18
......A..............C...........G.T...............AGG...A...T........   CPHL1 19
......A..............C...........G.C...............TGG...A...T...G....   CPHL1 4
......A............C.C........A..C.C...............TGG.......T........   CPHL1 43
......A..............C...........G.C........A......TGG...A...T........   CPHL1 7
......A....................GA..........G.G............................   CPHL2 11
...........................GA..A.........G............AG..............   CPHL2 18
...........................GA..A.........G............AG......C.......   CPHL2 25
...........................GA..A.........G.............G..............   CPHL2 3
..............C...............A....................G....AA............   CPHL7 13
..............C....................................G....AA............   CPHL7 17
....AGA.....T.C...G................................A.....A...T...G....   CPHL8 2
............T.C....................................A..A......T........   CPHL8 5
......A.....T.C....................................G.....A...T........   CPHL6 3
......A.....T.C....................................G.....A...T........   CPHL6 41
......A.....T.C....C...............................A.....A.......G....   CPHL6 6
..............C............G.......................AG....A............   CPHL9 1
..............C.............A.A....................AG....A.....G......   CPHL9 17
..............C.............A.A....................AG....A.....G......   CPHL9 6
.............................G.A.......C...........A.G...A...T.G......   CPHL3 48
.............................G.A...................A.G...A...T.G......   CPHL3 5
...........................G.G.A...................A.G...A...T.G......   CPHL3 6
.................C..........A..T.......G.A..T......GG.A..A............   CPHL12
.................C..........A..T.......G.A..T......GG.A..A.......G....   CPHL13 2
.................C.............T.......G.G..T......GG.A..A......G.....   CPHL13 8
......A.....................A..........G.A............................   CPHL14 15
......A.T...................A.A........G.A.....G..G...................   CPHL14 37
......A.T...................A.A........G.A............................   CPHL14 44
....................TG.....GA..A.........G..A..........GA.......T...G.   CPHL4 2
....................TG......A.AA.G.T....GC.............G........T...G.   CPHL4 35
....................TG......A.AA.G.T....GC.............GA.......T...G.   CPHL4 6
....................TG.....GC....C..........A...................T...G.   CPHL5 1
....................T......GC....C..........A...................T...G.   CPHL5 10
....................TG.....GC....C..........A...C...............T...G.   CPHL5 12
....................T......GA....TT.........A.........................   CPHL19
............T.......T......G...T.G......CC..AG..................GC....   CPHL18
...........................G......GT.....C..A................T........   CPHL24
....................TG.C....AGAA.G.T.......G.................A........   CPHL20
......A.....T.....G.G....A.G..A.............AG.....A...........G......   CPHL16
............T........G.C....A.AGCC............AG......G.........C.....   CPHL17
..................G.T..C..C......G.T........A.........A....G......C...   CPHL22
........G...........TG...A....A....G....CAG.G.TG...A.GG.....AAGG.C..GC   CPHL10 2
........G...........TG...A....A....G....CAG.G.TG...A.GG.....AAGG.C..GC   CPHL10 3
........G...T.......T....A....A...GT....CAGGG.TG...A.GA.....AAGG.C...C   CPHL10 7
........G...T.......T....A...GA...GT....CAGGG.TG...A.GG...GAGGG.....C   CPHL11 1
........G...T.......T....A...GA...GT....CAGGG.TG...A.GG...GAGGG.....C   CPHL11 2
........G...T.......T....A....A...GT....CAGGG.TG...A.GG...GAGGG.....C   CPHL11 5
.........................C.G.-----------------------------------------   CPHL15
...........................G..AAG..T....C...A.T.C.....GGA...AGCG......   CPHL21
```

```
TAAGAGAACAA---T--TTG----TGAA-AATAAAACAATAATCTTTAATCAA-C-TCCTCAGGAGGGGA   Consensus

...........AAACAA..ACAAT..G.G...........TG..........GT.G..............   CPHL1 cDNA
...........AAACAA..ACAAT..G.G............G..........GT.G..............   CPHL1 1
...........AAACAA..ACAAT..G.G............G..........GT.G..............   CPHL1 18
...........AAACAA..ACAAT..G.G............G..........GT.G..............   CPHL1 19
...........AAACAA..ACAAT..G.G............G..........GT.G..............   CPHL1 4
...........AAACAA..ACAAT..G.G...........TG..........GT.G..............   CPHL1 43
.G.........AAACAA..ACAAT..G.G............G..........GT.G..............   CPHL1 7
........A..---.--...----A..---...........GC.....C....---C.............   CPHL2 11
........A..---.--...----A..---...........GC.....C....---C.............   CPHL2 18
........A..---.--...----A..---...........GC.....C....---C.............   CPHL2 25
.....A..A..---.--...----A..---...........GC.....C....---C.............   CPHL2 3
...A.......---.--...----G...T............G........TC.---...........A..   CPHL7 13
...A.......---.--...----G...T............G.......ATC.---...........A..   CPHL7 17
...A..T....---.--AC.----A..---.............G...C.G...---C..........A..   CPHL8 2
...........---.--...----G..---...........G.....G.....---C..........A..   CPHL8 5
..G........---.--A..----G..GG....................G...---...........A..   CPHL6 3
..G........---.--A..----G..GG....................G...---...........A..   CPHL6 41
...A.......---.--...----G.G.G.............GA.........---...........A..   CPHL6 6
......C....---.--...----G..---.C.....................---...........A..   CPHL9 1
...........---.--...----A..---............C.........T---...........A..   CPHL9 17
...........---.--...----A..---............C.........T---...........A..   CPHL9 6
...........GAAC--CAT----.T..G........................---..............   CPHL3 48
...........GAAC--CAT----.T..G........................---..............   CPHL3 5
...........GAAC--CAT----.T..G........................---..............   CPHL3 6
..........C---.--...----..---............GAA.....AA..---..............   CPHL12
..........C---.--...----.A---....G.......GAA.....AA.T---..............   CPHL13 2
..........C---.--...----.A---............GAA.....AA..---..............   CPHL13 8
....G...A..---.--...----A..---...........GC..........---..............   CPHL14 15
........A..---.--...----A..---...........GC...C......---..............   CPHL14 37
........A..---.--...----A..---...........GC..........---..............   CPHL14 44
.....A.C.--------..----C.T..C..A.C...C....AA......T---.A..............   CPHL4 2
...A.A.C.--------..----C.T..C..A.C.........T......A---.A..............   CPHL4 35
...A.A.C.--------..----C.T..C..A.C.......G.T......A---.A..............   CPHL4 6
..G..A.C.--------..----C.T..C..A.C.GA.....AT......A---.A..............   CPHL5 1
..G..A.C.--------..----C.T..C..A.C..A......T......A---.A..............   CPHL5 10
..G..A.C.--------..----C.T..C..A.C.GA.....AT......A---.A..............   CPHL5 12
.......C.--------.C----C.T..C..G.C..A......T.....A.---.A.....G........   CPHL19
.....A.C.--------..----T.TG.C..C.GG........T.......---GA.....G........   CPHL18
..G...GC.--------..----C.T..C..G.C.GA.....AT.....A.---.ACA...G........   CPHL24
.......C.--------..----T.T..CC.G.C........AT.....A.---.A.....T........   CPHL20
...A...------------.CACT.T..T.....G.......G....---...C.AC..........A..   CPHL16
.....C..-----G.--C.----T.A.----C...............G..A---AACA............   CPHL17
....CA.C..-------C.TCCCT-----...................C.A..T.---............   CPHL22
..GA.ACC-----A.TAC.AAAAC----------..AT....C....G.CA---.AC.TG..........   CPHL10 2
..GA.ACC-----A.TAC.AAAAC----------..AT....C....G.CA---.AC.T...........   CPHL10 3
..GA.A.C-----A.TAC.AATAC----------..AT....C....G.CA---.AC.TG..........   CPHL10 7
...A.A.C-----A.TAC..ATAC...----.....AC....C....G.CA---.AC.TG..........   CPHL11 1
...A.A.C-----A.TAC..ATAC...----.....AC....C....G.CA---.AC.TG..........   CPHL11 2
...A.A.C-----A.TAC..ATAC...----.....AC....C....G.CA---.AC.TG..........   CPHL11 5
----------------------------------------------------------------------   CPHL15
..........-------C.TC--CCT..T..G....T..C.---.C.---...C.ACA............   CPHL21
```

```
CCCAGAAATTGTAATGCATAGTTTTAATTGTGGAGGGGAATTTTTCTACTGTAATACAACAAAACTGTTT   Consensus

..........A.......C.C...........................................T.....   CPHL1 cDNA
.............T..A.C.C............G..............................T.....   CPHL1 1
..................C.C...........................................T.....   CPHL1 18
..........A.......C.C...........................................T.....   CPHL1 19
..................C.C...........................................T.....   CPHL1 4
..........A.......C.C........................................G..T.....   CPHL1 43
..................C.C...........................................T.....   CPHL1 7
.A................C.................A........................C.G......   CPHL2 11
...........C......C.................A........................C.G......   CPHL2 18
...........C......C.................A........................C.G......   CPHL2 25
...........C.....CC.................AA.......................C.G......   CPHL2 3
...............................A......................................   CPHL7 13
...............................A......................................   CPHL7 17
G....T.GG...........T..........A................T............C........   CPHL8 2
....................T..........A.............................C........   CPHL8 5
.......G............T.....................................T..C........   CPHL6 3
.......G............T.....................................T..C........   CPHL6 41
...............................A.............................C........   CPHL6 6
.......G.......................A......................................   CPHL9 1
............G..................A......................................   CPHL9 17
....................A..........A......................................   CPHL9 6
..................C............A......................................   CPHL3 48
..................C............A......................................   CPHL3 5
..................C............A......................................   CPHL3 6
...........A......C..........................................C.CT.....   CPHL12
...........A......C.......................C..T...............C.CT.....   CPHL13 2
...........A......C..........................................C.CT.....   CPHL13 8
..................C.................A.....................T...........   CPHL14 15
....................................A.....................T...........   CPHL14 37
....................................A.....................T...........   CPHL14 44
..........AC..CA..C.......................C...............T...........   CPHL4 2
..........AC..CA..........................C...............T...G.......   CPHL4 35
..........AC..CA..........................C...............T...G.......   CPHL4 6
..........AC..CA..C.......................C...............T...GC......   CPHL5 1
..........AC..CA..C.......T...............C...............T...GC......   CPHL5 10
..........AC..CA..C.......................C...............T...GC......   CPHL5 12
..........AC..CA..C..C..............A..............C......T...........   CPHL19
.........A.C..CA....................A...........T...G.C...T..GCC......   CPHL18
..........AC..CA..C.......................C...............T..GGC......   CPHL24
..........AC..CA....................A...........T.....C...T..GGCT.....   CPHL20
T.T.......AC.......CA..........A................T..C..........G.......   CPHL16
TTT.......AC..C......C.........A....A...........T..C......T...........   CPHL17
.AT.......ACC.CA....................A...........T..C......T..GGC......   CPHL22
..T....G..AC..CAT..............A....A..G..C.....T............GG.......   CPHL10 2
.......G..AC..CA...............A....A..G..C.....T............GG......-   CPHL10 3
..T....G..AC..CA...............A....A..G..C.....T............GG.......   CPHL10 7
..T.......A.G.CA...............A....A..G........T.........T...........   CPHL11 1
..T.......A.G.CA...............A....A..G........T.........T...........   CPHL11 2
..T.......A.G.CA...............A....A..G........T.........T...........   CPHL11 5
----------------------------------------------------------------------   CPHL15
..T.......AC..CA...............A....A..G........T..C......T...........   CPHL21
```

```
AAT-AGTACTTGG---------AATGGTACTGAAAAGACGAATAATACTG--GAA-----TAACAAG--A   Consensus

G..-.........---------....T........GAC.T..C...GA.ACT.G.-----ATG..----.   CPHL1 cDNA
G..-.........---------....T........G.C.T......GT.ACT.G.-----ATG..----.   CPHL1 1
G..-.........---------....T........G.C.T......G..ACT.G.-----ATG..----.   CPHL1 18
G..-.........---------....T........G.C.T......GA.ACT.G.-----ATG..----.   CPHL1 19
G..-.........---------....T........G.C.T......G..ACT.G.-----ATG..----.   CPHL1 4
...-.........---------...AT.......GG.C.T......GT.ACT.G.-----GGG.C----.   CPHL1 43
G..-.........---------....T..........C.T......G..ACT.G.-----ATG..----.   CPHL1 7
...-.A.......---------..C.........TG...A.....-.TG.CCA..-----..G..CTG-.   CPHL2 11
...-.A.......---------..C......A..TG...A...G.-.TG.TCA..-----..G..CTG-.   CPHL2 18
...-.A.......---------..C......A..TG...A...G.-.TG.TCA..-----..G..CTG-.   CPHL2 25
...-.A.......---------..C........TTG...AT....-.TG.TCA..-----..G..CTG-.   CPHL2 3
...-.........G---GAGTG..........T.G.AG--------------------------...G-.   CPHL7 13
...-..C......G---GAGTG..........T...AG--------------------------...A-.   CPHL7 17
...-.........G---------------------GAGT....G......AA..G-----.T.A.TAA-.   CPHL8 2
...-.........A---------..A.....TGG.GAGT....G......AA..G-----.T.A.TAA-.   CPHL8 5
...-.........A---AAATG....A.....GGGGAGT....G......AA.GG-----.T.....G-.   CPHL6 3
...-.........A---AAATG....A.....GGGGAGT....G.G....AA.GG-----.T.....G-.   CPHL6 41
...-.........G---GAGTG...A......------------.A....-----------T.....G-.   CPHL6 6
...-.........---------...A......GGG.A.T.....G.....AA..G-----.C.A...A-.   CPHL9 1
...-.........---------...AA....TGGG.A.T.....GC....AA..G-----.C.A...A-.   CPHL9 17
...-.........---------...A.....TGGG.A.T.....GC....AA..G-----.C.A...A-.   CPHL9 6
...-.........G---A---G.......------------....C....TA--------A.TA.TAA-C   CPHL3 48
...-.........G---G---G.......------------....C....TA.G.-----A.TA.CAA-C   CPHL3 5
...-.........G---A---G.......------------....C....TA.G.-----A.TG.CAA-C   CPHL3 6
...------..AAA---GTTT-...A.....TGG..TG-...CT.C.GG.TTA..-----.G...C---.   CPHL12
...------..AAA---GTTT-...A.....TGG..TG-...CT.C.GG.TTA..-----.G...C---.   CPHL13 2
...------..AAA---GTTT-...A.....TGG..TG-...CT.C.GG.TTA..-----.G...C---.   CPHL13 8
...-.A.......---------..C.........TG...A...G.-.TG.GAA..-----.....CTG-.   CPHL14 15
..C-CA.......---------..C.........TG...A-------------.------..T..CTG-.   CPHL14 37
..C-CA.......---------..C.........TG...A-------------.------..T..CTG-.   CPHL14 44
...-.....A...GA-TGCT--...A....A.GG..TGTT.G..C..A.TCA...GAGGC......TA--   CPHL4 2
...-.A...A...CA-TGAT--...AA...ATGG..TG.TG...G...ATGGA.TGATAG...T..TAC.   CPHL4 35
...-.A...A...CA-TACT--...AA...ATGG..TG.TG...G...ATGGA.TTATAG...TCTTAA.   CPHL4 6
...-G....A...---------...A....ATGG..T.TT.GA....A.ACA...GGGGC...T..TAC.   CPHL5 1
...-G....A...---------...A....ATGG..T.TT.GA....A.ACA.GCGGGGC...T..TAC.   CPHL5 10
...-G....A...---------...A....ATGG..T.---------A.ACA...GGGGC...T..TAC.   CPHL5 12
...-.....A...GAG-AATTC..---C.AAT.GT.C.AAGG...--A.---A.TAGTGCAG...---C.   CPHL19
...-..C......GAT-AATGC---------C.GC.T..AGGC..--A.---AGCACGACGC------T.   CPHL18
...-..C......CAGCAGTCT.....C...TGGC.A.G..G.G.C.A.---.CCAGCACGC.GG..TC.   CPHL24
...-..C......-------------A..AATGGC..CATG.C.TC.A--------------.T.GCGCG   CPHL20
...-.A......-----------------.AT....----...G.A..CATG..GGGG--------TGT.   CPHL16
...-.-----GCACTTGGTTT-..........TG---CA....G.C..A.---------------G.TC.   CPHL17
...TG.C..A...AA-------.....C..ATTG-----.....-------T.GGGGAAA.GG..CT---   CPHL22
..------------------T-...AA..--AT.---CT..-G...G------------------..---   CPHL10 2
--------------------T-...AA..--AT.---CT..-G...G------------------..---   CPHL10 3
..------------------T-...AA..--AT.---CT..-G...G------------------..---   CPHL10 7
..A-.-----.A.AACTGTCT-...AA..--.T.---CT..-..G.G...---------------..---   CPHL11 1
..A-.-----.A.AACTGTCT-...AA..--.T.---CT..-..G.G...---------------..---   CPHL11 2
..A-.-----.A.AACTGTCT-...AA..--.T.---CT..-..G.G...---------------..---   CPHL11 5
----------------------------------------------------------------------   CPHL15
...-.....--A.GACTGTTT-...A..----.T.GT..A....G...------------------AAC.   CPHL21
```

```
AATGAAAATATCACACTCCCATGCAGAATAA-AACAAATTATAAACATGTGGCAGGAAGTAGGAAAAGCA    Consensus

.......................T.......-G.......G.....T.......................    CPHL1 cDNA
..A....................T.......-G....T..G.....C.......................    CPHL1 1
.......................T.......-G.......G.....T.......................    CPHL1 18
.......................T.......-..............T.......................    CPHL1 19
.......................T.......-G.............T.......................    CPHL1 4
---....................T.......-..............T.......................    CPHL1 43
.......................T.......-G.......G.....T.......................    CPHL1 7
.G-A...T-......................-......................................    CPHL2 11
.C-A...T-......................-......................................    CPHL2 18
.C-A...T-........T.............-......................................    CPHL2 25
..-A...T-......................-......................................    CPHL2 3
.....C..C......................-......................................    CPHL7 13
.....C..C......................-......................................    CPHL7 17
.C...C.CA......................-.....T..................G............T    CPHL8 2
.C...C.TA......................-.....T..................G.............    CPHL8 5
.......CA....T.................-......................................    CPHL6 3
.......CA....T.................-......................................    CPHL6 41
.....C.CA......................-......................................    CPHL6 6
.....C.CA..........A...........-......................................    CPHL9 1
.....T.CA......................-....G.................................    CPHL9 17
.....T.CA......................-....G.................................    CPHL9 6
.C.....T.......................-......................................    CPHL3 48
.C.---.TC......................-......................................    CPHL3 5
.C.....T.......................-......................................    CPHL3 6
..G............................-.....T................A.G......G......    CPHL12
.......................T.......-.....T................A.G......G......    CPHL13 2
..G............................-.....T..................G.............    CPHL13 8
.--------.........T............-........G.............................    CPHL14 15
.G-....T-........TA............-......................................    CPHL14 37
.G-....T-.........A............-......................................    CPHL14 44
----CT.....T.T.....A...........-.............T.................G......    CPHL4 2
.CA.CC.....T.......A...........-.............T........................    CPHL4 35
C.A.CC.T.T.T...A...A.......T...-..A.T........T....................T.T.    CPHL4 6
C.G.TT.---.........A...........-.....C.......TC.........G.......C.....    CPHL5 1
C.G.TT.---.........A...........-.............TC.........G.......C.....    CPHL5 10
C.G.TT.---.........A...........-C....T......CTC.........G.......C.....    CPHL5 12
....GG.TA....T.........T.......-...............G......................    CPHL19
T....CCC...A..T................-......................................    CPHL18
...AT..C...A..T................-.G....................................    CPHL24
G.G.GC.....A..T....A...........-......................................    CPHL20
.-..GC.C.........T.......AG....-.G......................G.......C.....    CPHL16
G....C.....A..T................-.G......G.G.GA.....T....G..........A..    CPHL17
------........T................-........G..............AG.......CG....    CPHL22
.....C.C...............T..G....-........G.G.GA.........AG.......C.....    CPHL10 2
.....C.C...............T..G....-........G.G.GA.........AG.......C.....    CPHL10 3
.....C.C...............T..G....-........G.G.GA.........AG.......C.....    CPHL10 7
....GC.C...............T.AG....-........G.G.GA.........AG.......C.....    CPHL11 1
....GC.C...............T.AG....-........G.G.GA.........AG.......C.....    CPHL11 2
....GC.C...............T.AG....-........G.G.GA.........AG.......C.....    CPHL11 5
----------------------------------------------------------------------    CPHL15
G..TC...C..........A...........T.............T..........GG.....GCG...G    CPHL21
```

```
ATGTATGCCCCTCCCATCAGAGGACTAATTAGATGTTCATCAAATATTACAGGGCTACTATTAACAAGAG  Consensus

....................C....A......G........................A............  CPHL1 cDNA
..................GA.....A......G.................C......A............  CPHL1 1
..................GA.....A......G........................A............  CPHL1 18
..............A.....C....A......G........................A............  CPHL1 19
....................T....A......G........................A............  CPHL1 4
..................GA.....A......G........................A............  CPHL1 43
....................T....A......G........................A............  CPHL1 7
....................T....A............................A.........T.....  CPHL2 11
....................T....A............................A........G......  CPHL2 18
....................T....A............................A........G......  CPHL2 25
....................T....A......G.....................A.G......GT.....  CPHL2 3
..............T.....T....C...C.T...C.................A.........GT.....  CPHL7 13
..............T.....T....C......T..C..............G..A.........GT.....  CPHL7 17
....................T....C.........C..............G...................  CPHL8 2
....................T...TC.........C..............G...................  CPHL8 5
....................T..............C..............G............GT.....  CPHL6 3
....................T..............C..............G............GT.....  CPHL6 41
....................T....C......T..C.T............G...A........GT.....  CPHL6 6
....................T....C.........C........G.....G............CT.....  CPHL9 1
....................T..GAC.........C..............G.............T.....  CPHL9 17
....................T..G.C.........C..............G.............T.....  CPHL9 6
..................GA.....C............................T.G.............  CPHL3 48
..................GA.....C............................T.G.............  CPHL3 5
..................GA.....C............................T.G.............  CPHL3 6
......................................................T.GA............  CPHL12
.................................GT...................T.GA............  CPHL13 2
.................................GT.....................GA............  CPHL13 8
....................T....A............................A........GT...G.  CPHL14 15
....................T.................................A........GT.....  CPHL14 37
....................T.................................A........GT.....  CPHL14 44
..................GC...........AC...A.......C.........A..A....G.......  CPHL4 2
..............T...GC...........A............C.........A..A....G.......  CPHL4 35
...........A......GC...........A............C.........A.......G...CCC.  CPHL4 6
..................GA.........C.CC...A.......C.........A..A....G.......  CPHL5 1
..................GA...........TC...A.......C.........A..A....G.......  CPHL5 10
....G.............C..A.GA......TC...A.......C.........A..A....G.......  CPHL5 12
.................TGA...........................A.....A...T.G..G.......  CPHL19
.................TGA....A....C.A............G........A...T.G..G.......  CPHL18
.................TGA.........C.AG....................A...T.G..G....A..  CPHL24
......----------------------------------------------------------------  CPHL20
........T...........T...AG.....AT...GT...............AA.......G.......  CPHL16
......A...........CC....GAG..A.AG...GA....C.C........A................  CPHL17
.................TGC....A....A.AG...A.......C........AA..A..A.G.......  CPHL22
..................GC....A.C.........GA.........C.....A...T............  CPHL10 2
..................GC....A.C.........GA.........C.....A...T............  CPHL10 3
..................GC....AAC.........GA.........C.....A...T............  CPHL10 7
..................GC....AAC.........GA.........C.....A..GT............  CPHL11 1
..................GC....AAC.........GA.........C.....A..GT............  CPHL11 2
..................GC....AAC.........GA.........C.....A..GT............  CPHL11 5
----------------------------------------------------------------------  CPHL15
..A..............-----------------------------------------------------  CPHL21
```

```
ATGGTGGT---AATAACAATAACACCAACAACAC-------GAGACCTTCAGACCTGGAGGAGGAGATAT   Consensus

.......C---..C.C....G.G......GC...C------..........................G..   CPHL1 cDNA
.......C---..C.C....G.G.....AGC...C------..........................G..   CPHL1 1
.......C---..C.C....G.G.....AGC...C------..........................G..   CPHL1 18
.......C---..C.C....G.G......GC...C------..........................G..   CPHL1 19
.......C---..C.C....G.G......GC...C------..........................G..   CPHL1 4
.......C---..C.C....G.G...........C------..........................G..   CPHL1 43
.......C---..C.C....G.G......GC...C------..........................G..   CPHL1 7
........---..C..T..C.G...T........------A....T...........C............   CPHL2 11
........---..C..T..C.G...T........------A....T...........C............   CPHL2 18
........---..C..T..C.G...T........------A....T...........C............   CPHL2 25
........---..C..T..C.G...T........------A....T...........C............   CPHL2 3
........---..........G...G....CTG.CACCGAG---.T...................A....   CPHL7 13
........---.G.....-----..G.----T..CACCGAG---.T...........A............   CPHL7 17
........---........AG.T..GG.......CACCGAG---.T........................   CPHL8 2
........---.........G.T..G......G.CACCGAG---.T........................   CPHL8 5
........---G.G..A........G...G.G..CACCGAG....T...................A....   CPHL6 3
........---G.G..A........G...G.G..CACCGAG....T...................A....   CPHL6 41
........---..G..A........G..TG.T..CACCGAG---.T...........T.......A....   CPHL6 6
........---.............------.G..C------....T........................   CPHL9 1
........---.........G....A...G.G..C------....T........................   CPHL9 17
........---.........G....A...G.G..C------....T........................   CPHL9 6
........---....G...C.GG......------------.............................   CPHL3 48
........---........CGGG......------------.............................   CPHL3 5
........---...GG...CGGG......------------.............................   CPHL3 6
........---....GT.................T------....T...T......AT.......A....   CPHL12
........---....GT.................T------....T...T......TT.......A....   CPHL13 2
........---....GT.................T------..-.T...T.-....TT.......A-...   CPHL13 8
........---..C..T....G...T........------A................C.......A....   CPHL14 15
........---..C..T..C.G...T....C...------A................C....--------   CPHL14 37
........---..C..T..C.....T....C...------A................C............   CPHL14 44
......AG---------..C..TT.---T.GTC.TAAC---.......C......G..............   CPHL4 2
.......G---------..C.GTT.---T..TT.TAGC---.............................   CPHL4 35
.......G---------..C.G.T.---T..TT.TAGC---.............................   CPHL4 6
........---------..C...------..TT.TAGA---.............................   CPHL5 1
........---------..C...------..TT.TAGC---.............................   CPHL5 10
........---------..C...------..TT.TAGA---....T........................   CPHL5 12
........GTA.G...T....GTCAG..T------------........T....................   CPHL19
........ACA.....T...GGT..A..T..T.ATAGTCAG.............................   CPHL18
........---.....T....GTC.G..T------------......A....G.................   CPHL24
----------------------------------------------------------------------   CPHL20
........TC------T.....TG.---------------------------------------------   CPHL16
.......GCCG...---.G.---.....T------------................C............   CPHL17
.......---G..C.GTG.G....TG.G...T---------.....T.C.--------------------   CPHL22
.......GAT---.G.....---..A..T------------..A..........................   CPHL10 2
.......GAT---.G.....---..A..T------------..A..........................   CPHL10 3
.......GAGG.........---..A..T------------..A..........................   CPHL10 7
.......GAGG.........---..A..T------------..A..........................   CPHL11 1
.......GAGG.........---..A..T------------..A...................-------   CPHL11 2
.......GAGG.........---..A..T------------..A..........................   CPHL11 5
----------------------------------------------------------------------   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
GAGGGACAATTGGAGAAGTGAATTATATAAATATAAAGTAGTAAAAATTGAACCATTAGGAGTAGCACCC   Consensus

...A........................................G.........................   CPHL1 cDNA
...A..................................................................   CPHL1 1
...A..................................................................   CPHL1 18
...A........................................G.........................   CPHL1 19
...A........................................G.........................   CPHL1 4
...A..................................................................   CPHL1 43
...A........................................G.........................   CPHL1 7
..................................................T...................   CPHL2 11
...A.....A........................................T...................   CPHL2 18
......................................................................   CPHL2 25
..................................................T...................   CPHL2 3
............................................G................A.C...T..   CPHL7 13
............................................G................A........   CPHL7 17
.............C..............................G.........................   CPHL8 2
............................................G......T..................   CPHL8 5
..A...T.....................................G.........................   CPHL6 3
..A.........................................G.........................   CPHL6 41
.........C..................................G.........................   CPHL6 66-
.......................................G....G.........................   CPHL9 1
.................................C..........G......C..................   CPHL9 17
.................................C..........G.........................   CPHL9 6
..............................................G.......................   CPHL3 48
..............................................G........G..............   CPHL3 5
..............................................G.......................   CPHL3 6
..A.........................................G.........................   CPHL12
..A..............T..........................G.........................   CPHL13 2
..A.-...-.....-.GT..........................G...G.............C.......   CPHL13 8
.................T...................T................................   CPHL14 15
----------------------------------------------------------------------   CPHL14 37
........................................A.........T...................   CPHL14 44
....................................G.........C........C..............   CPHL4 2
....................................G..G..T............C..............   CPHL4 35
..A.................................G..G...............C..............   CPHL4 6
....................................G..................C..............   CPHL5 1
....................................G..................C..............   CPHL5 10
........T....C......................G..............C...C.T...A........   CPHL5 12
...A.......................C...........................C....GC........   CPHL19
...AA......................C................G..........C....TA.......T   CPHL18
...A.................C......................G...C......C....TC........   CPHL24
----------------------------------------------------------------------   CPHL20
----------------------------------------------------------------------   CPHL16
.T............................G......A....GG.....C.....C..............   CPHL17
----------------------------------------------------------------------   CPHL22
......................C.......G............G....A---------------------   CPHL10 2
......................C.......G............G....A---------------------   CPHL10 3
.................-----------------------------------------------------   CPHL10 7
..............................G.........A.......A---------------------   CPHL11 1
----------------------------------------------------------------------   CPHL11 2
..............................G.........A........---------------------   CPHL11 5
----------------------------------------------------------------------   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
ACCAAGGCAAAGAGAAGAGTGGTGCAGAGAGAAAAAAGA     Consensus
               C C
             |MluI|
..............G...........A............     CPHL1 cDNA
....G.........G........................     CPHL1 1
..............G........................     CPHL1 18
....G.........G...........A............     CPHL1 19
....G.........G...........A.....G......     CPHL1 4
....G.........G...........A............     CPHL1 43
..............G...........A............     CPHL1 7
..........G.......................C....     CPHL2 11
..........G............................     CPHL2 18
..........G............................     CPHL2 25
..........G............................     CPHL2 3
....G......A...............C...........     CPHL7 13
....G......A...........................     CPHL7 17
...........C............A..............     CPHL8 2
...........A...........................     CPHL8 5
...........A...........................     CPHL6 3
...........A...........................     CPHL6 41
...........A...........................     CPHL6 6
....G......A...........................     CPHL9 1
....G......A....................G......     CPHL9 17
....G......A....................G......     CPHL9 6
.......................................     CPHL3 48
.......................................     CPHL3 5
.......................................     CPHL3 6
---------------------------------------     CPHL12
...............C.C......---------------     CPHL13 2
......A........C.C.C....G.-------------     CPHL13 8
---------------------------------------     CPHL14 15
---------------------------------------     CPHL14 37
....-----------------------------------     CPHL14 44
........................G.A............     CPHL4 2
........................G.A............     CPHL4 35
........................G.A............     CPHL4 6
....G...................GGA............     CPHL5 1
....G...................GGA............     CPHL5 10
C...G.....C.C...........GGA............     CPHL5 12
.........------------------------------     CPHL19
..T.......G.............G.-------------     CPHL18
....C......A............G.A............     CPHL24
---------------------------------------     CPHL20
---------------------------------------     CPHL16
..........GA............G..............     CPHL17
---------------------------------------     CPHL22
---------------------------------------     CPHL10 2
---------------------------------------     CPHL10 3
---------------------------------------     CPHL10 7
---------------------------------------     CPHL11 1
---------------------------------------     CPHL11 2
---------------------------------------     CPHL11 5
---------------------------------------     CPHL15
---------------------------------------     CPHL21
```

Figure 3.4

Alignment of all CPHL translated amino acid sequences

```
A-E-LWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTDPNPQEVVLENVTENFNMWKNN    Consensus

.T.K.............D..............S.....I.......................Y....E..    CPHL1 1
.T.K.............D..............S.....I.......................Y....E..    CPHL1 18
.T.K.............D..............S.....I.......................Y....E..    CPHL1 19
.T.K.............D..............S.....I.......................Y.......    CPHL1 4
.T.K.............D..............V.....I.......................Y.......    CPHL1 43
.T.K.............D..............S.....I.......................Y....E..    CPHL1 7
----------------------------------------------------------------------    CPHL1 cDNA
.K.Q.GF.........................K..M..................I.......E.D.....    CPHL2 11
.T.Q............................K..M..................I.......E......D    CPHL2 18
.T.Q............................K..M..................I.......E.......    CPHL2 25
.T.Q............................K..M..................I.......E.......    CPHL2 3
.Q.K-.............T......................................G............    CPHL7 13
.Q.K-....................................................G............    CPHL7 17
.K.K...................................................R.........A...D    CPHL8 2
.KDK.............................A.......................G.......I...D    CPHL8 5
.K.K............................A.....................I..G............    CPHL6 3
.K.K............................A.....................I..G............    CPHL6 41
.K.K............................A........................G............    CPHL6 6
.T.K.................................................K...K............    CPHL9 1
.T.K.....................................................K............    CPHL9 17
.T.K.....................................................K............    CPHL9 6
.E.QV....................................................G....D.......    CPHL3 48
-------.......................................................D.......    CPHL3 5
.V.Q..........................................................D.......    CPHL3 6
-----------..L..R..N..............................D...IE..............    CPHL12
---------.......R..N..............................D...IE..............    CPHL13 2
--------........R..N..............................D...IE..............    CPHL13 8
--------------..................E..M..................IE.K....D.D.....    CPHL14 15
.A.K............................K..M..................LE.K....K.D.....    CPHL14 37
.A.K............................K..M..................LE.K....K.D.....    CPHL14 44
.EDN..........................S.K..A..I...............IE..............    CPHL4 2
.EDN..........................S.K..A..I...............IE..............    CPHL4 35
.EDN..........................S.K..A..I...............IE..............    CPHL4 6
--------......................S.K..A..I...............IE..............    CPHL5 1
.EN-..........................S.K..A..I...............IE..............    CPHL5 10
.EN-..........................S.K..A..I...............IE..............    CPHL5 12
     ........................S.HA.A..................MK..............D    CPHL19
----------------------------------------------------------------------    CPHL18
VAGQ.........................S.RA.A....................H.K............    CPHL24
----------------------------------------------------------------------    CPHL20
----------------------------------------------------------------------    CPHL16
.QNN............RD.D.P..........S..K..................IP.............S    CPHL17
----------------------------------------------------------------------    CPHL22
------..........ED.D...............S..................LD.K............    CPHL10 2
------..........ED.D...............S..................LD.K............    CPHL10 3
----------------------------------------------------------------------    CPHL10 7
------..........ED.D...............S..................LD.K............    CPHL11 1
----------------------------------------------------------------------    CPHL11 2
------..........ED.D...............S..................LD.K............    CPHL11 5
----------------------------------------------------------------------    CPHL15
----------------------------------------------------------------------    CPHL21
```

```
MVEQMHEDIISLWDQSLKPCVKLTPLCVTLNCTN--------------T-N------T--S--E-I---G   Consensus

................................A.LTLNCTKRNGTHTNCT.LGNDAN.TI.SG.T.--QS   CPHL1 1
..................................----------------TTITSAN.TI.SG.P.--QN   CPHL1 18
..................................L----NDTNTTSAK.TTTTPSVN.TS.SG.T.--RS   CPHL1 19
..................................----------------TTITSANPTI.RG.P.--QS   CPHL1 4
..................................-----------ATSANTTITGTN.TI.SG.T.--QS   CPHL1 43
.................................KL----NDTNTTSAK.TTTTPSVN.PS.SG.T.--RS   CPHL1 7
-------------.....................L----NDTNTTSAK.TTTTPSVN.TS.SG.T.--RS   CPHL1 cDNA
..............................S...V----NATKNATTY.N.SS----------.E.MEK.   CPHL2 11
..................................V----NATKNATTY.N.SS----------KK.MEKE   CPHL2 18
..................................V----NATKNATTY.N.SS----------KK.MEKE   CPHL2 25
..................................V----NATKNATTY.N.SS----------KE.MEKE   CPHL2 3
.................................DL----KNNGTMSNNNTT------NAT.SSGG----.   CPHL7 13
.................................DL----KNNGTKSNNNT.------NAT.SNGG----.   CPHL7 17
.................................DY----NNTDTATSA.SSNGTAAGATS.SG.MLEKK.   CPHL8 2
.................................DY----NNTNSND----TDATGNNATS.SG.MLEKE.   CPHL8 5
.................................DY----NNTNPNATN.T.------ATS.IEGE.EK-.   CPHL6 3
.................................DY----NNTNSNATD.T.------ATS.IGGEVEK-.   CPHL6 41
.................................DV----NDTNSN--N.T.------ATS.IG.K.ER-.   CPHL6 6
.................................DT----NSMN----N.T.------.NS.IE.T.EK-.   CPHL9 1
..D..............................DT----NSMN----N.T.------.NS.IG.T.EK-.   CPHL9 17
..D..............................DT----NSMN----N.T.------.NS.IG.T.EK-.   CPHL9 6
.................................DY----VGNTTNTNT.NSTTSGN-------------.   CPHL3 48
.................................DY----VGNTTNTNT.NGTTSGNE.MGI--------A   CPHL3 5
.................................DY----VGNTTNTNT.YSTTSGNE.MGA--------.   CPHL3 6
..........N......................D-KLNIINTTKVDNSSGENNSRVDERGT--------.   CPHL12
..........N......................D-KLNIINTTNTNNISLGNYSKLD---A--------.   CPHL13 2
..........N......................D-KLNIINTTNTNKSSGGNNSRVDERGT--------.   CPHL13 8
.....Q........E...................V----NATQNATTYNN.SS----------.GMMEK.   CPHL14 15
..............................K...V----NVTQNATTH.N.SR----------.G.MEE.   CPHL14 37
..............................K...V----NVTQNATTH.N.SS----------.G.MEE.   CPHL14 44
................................-------------TNEWNSTVGN---IILRNSSTTG--   CPHL4 2
................................-------------TNEWNSTVGNNN.LVNRDSSTTG--   CPHL4 35
..............................MTM-----------MTALWE-----QQDLVNRNSSTTG--   CPHL4 6
................................-------------TDYWG.YTSGSN.PEMNN--TTD--   CPHL5 1
................S.A.............-------------TDNWG.YTSGNN.PEMNN--TPE--   CPHL5 10
..............H.S.A.............-------------TDYWG.YTSGNN.LEMNN--TTE--   CPHL5 12
................................--------------TDRK.DTISN---------ATDAI   CPHL19
----------------------------------------------------------------------   CPHL18
................................--------------TDWK.NATNTNNTTATNVTTNEDI   CPHL24
----------------------------------------------------------------------   CPHL20
--------------..................AKANW----------------TNANITYVPNIIGNLTD   CPHL16
..............E..................DVR-----------------NNTDVKN.TEVNNDGRK   CPHL17
----------------.Q..............RDSN-----------------------STDISNVTST.   CPHL22
..Q...........EG..................VT----------------YTNATNCTENNN-VENRE   CPHL10 2
..Q...........EG..................VT----------------YTNATNCTENNN-VENRE   CPHL10 3
-----------------------------.....VT----------------CANTTNCTENNNTVSSRE   CPHL10 7
..............EG..................VT----------------YTNATNCTENNN-VENRE   CPHL11 1
------------------------....N.....VT----------------YTNATNCTENNN-VENRE   CPHL11 2
...H..........EG..................VT----------------YTNATNCTENNN-VENRE   CPHL11 5
------------------------..........IA----------------KNGNI.Y---NSSMEG--   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
EMKNCSFNITTEIRDKVKKEYALFYKLDVVPIDND-----------TSYRLISCNTS-VITQACPKVSFE   Consensus

D..........SL..R...Q.....S..I.QL...K-----DNT--.T...R.....-.........T..   CPHL1 1
D..........SL..R...Q.....S..I.QL...N-----DNT--.T...R.....-.........T..   CPHL1 18
.......K.S.SLK.R...Q.....S..I.QL...K-----DNT--.T...R.....-.........T..   CPHL1 19
D..........SL..R...Q.....S..I.Q....K-----DNT--.K...R.....-.........T..   CPHL1 4
D..........SL..R...Q.....S..I.QL...N-----DNT--.T...R.....-.........T..   CPHL1 43
...........SL..R...Q.....S..I.Q...NK-----DNT--.K...R.....-.........T..   CPHL1 7
...........SL..R...Q.....S..I.Q.N..K-----DNT--.K...R.....-.........T..   CPHL1 cDNA
.I.........DL....Q..........I...E..R-----DSDN-...........-............   CPHL2 11
.I.....K...DL....Q.......T..I...E..R-----GSNN-...........-T...........   CPHL2 18
.I.....K...DL....Q.......T..I...E..R-----GSNN-...........-T...........   CPHL2 25
.I.........DL....Q..........I...E..R-----NSNN-...........-I...........   CPHL2 3
...........N..N.FQR..............D.N----------...........-........I..D   CPHL7 13
...........N..N.LQR..............D.N----------...........-........I..D   CPHL7 17
.I.........N.....Q................TN-----NTTNY...........-...........D   CPHL8 2
.I.........N....M...H....S..I.....-------DKENN...........-...........D   CPHL8 5
.I......V..N...RLQ..........I...ED.N----------...........-.L.........D   CPHL6 3
.I......V..D...RLQ...............E.N----------...........-.L.........D   CPHL6 41
.I......V..NM..RLQ...............D.N----------...........-...........D   CPHL6 6
........V..N....MQ............K...K-----------...........-........I..D   CPHL9 1
........V..N....MQ............K...E-----------...........-........I..D   CPHL9 17
........V..N....MQ............K...E-----------...........-........I..D   CPHL9 6
.........S.NL...YQ.V..............NT-----TNT--.N.........-.........T.D   CPHL3 48
.........S.SL...FQ.V..............ND-----T------.........-.........T.D   CPHL3 5
.........S.SL...FQ.V..............ND-----T------.........-.........T.D   CPHL3 6
.I........SS.Q...HR...........Q.GGND-----N----...........-....T.......   CPHL12
.I........SG.Q..KH............Q.GGND-----N----...........-....T.......   CPHL13 2
.I........SG.Q..KHR...........Q.GGND-----N----...........-....T.......   CPHL13 8
...........G.....Q.....L....I...ED.K-----DSN--........KPPQSL..........   CPHL14 15
............V............A..I...--KN-------NN-A..........-I...........   CPHL14 37
............V............A......E.KK-------DN-...........-I...........   CPHL14 44
-.RK....V..A....Q.QVH.......I......NSDNSTNSNKSAN....N....A-...........   CPHL4 2
-.RK....V..A....Q.QVH.......I......NSDNSTNSNKSAN....N....A-...........   CPHL4 35
-.RK....V..A....Q.QVH.......I......NSDNSTNSNKSAN....H....A-...........   CPHL4 6
-.I.S.VS...VVG..K..VI....R..I....S--------SDNSSN....N....A-...........   CPHL5 1
-..........VV...K..VT....R..I....S--------SDNRSN....N.H..A-...........   CPHL5 10
-..........IV...K..VP....R..I....S--------SDNSSN.W..N....A-......T...D   CPHL5 12
...............RK.QVH.......I.Q..DNNSTK---------....N....A-........T..   CPHL19
----------------------------------------------------------------------   CPHL18
GV..........V..RK.QVF.........QM.D.NSTNT----NY.N....H....A-.......IT..   CPHL24
----------------------------------------------------------------------   CPHL20
.VR.....M.......KQ.VH.I.....I.Q.KGNNNNDS------NE....N....-..K.....I..D   CPHL16
.LT......N...K.RR.....I...........SNA---------......N..VTT.-K......T..   CPHL17
.I....Y..S..L...RQQV.SM.......Q.SE--SSNNTRSNDSAQ....N....A-...........   CPHL22
.I..............EEE.......I.....NGS----------SSD.V.MN....T-.K......N.D   CPHL10 2
.I.............REK-.......I.....NGS----------SSD.V.IN....T-.K......N.D   CPHL10 3
................Q.........V....LN.N--------LNNSD.I..N....T-.K........D   CPHL10 7
................K.........I.....NGS----------SSD.V..N....T-.K....Q.NLD   CPHL11 1
................K.........I.....NGS----------SSD.V..N....T-.K......N.D   CPHL11 2
................K.........I.....NGS----------SSD.V..N....T-.K......N.D   CPHL11 5
.LR.....A.......QR..........I.SLNGN----------SSE....N....T-..........D   CPHL15
----------------------------------------------------------------------   CPHL21
```

```
PIPIHYCAPAGFAILKCNDKKFNGTGPCKNVSTVQCTHGIKPVVSTQLLLNGSLAE-EEVVIRSENFTDN   Consensus

.......T................................R...............-...M...A.....   CPHL1 1
.......T................................R...............-.G.M...A.....   CPHL1 18
.......T................................R...............-.......A.....   CPHL1 19
........................................R...............-...M...A.....   CPHL1 4
........................................R...............-...M...A.....   CPHL1 43
........................................R...............-...M...A.....   CPHL1 7
........................................R...............-...M...A.....   CPHL1 cDNA
.......T.....L......E.S.K...T...........R...............-K............   CPHL2 11
.............L......E.S.K...T...........R...............-K............   CPHL2 18
.............L......E.S.K...T...........R...............-K............   CPHL2 25
.............L......E.S.K...T...........R...............-K............   CPHL2 3
.V..Q..T..............H....WR...A.......................-.......Q..SN.   CPHL7 13
.V.....T....................T...........................-.......Q..SND   CPHL7 17
.................K......................................-.......Q..S..   CPHL8 2
..................N.....................................-....L..Q..S..   CPHL8 5
..L................N........T...........................-KD.....Q.....   CPHL6 3
.V.................N........R...........R...............-.D.....Q.....   CPHL6 41
.V.................N........R...........................-G......Q.....   CPHL6 6
.V..........................T...........................-.D.....Q.L...   CPHL9 1
.V..........................T...........................-.D.....Q.L...   CPHL9 17
.V..........................T....................V......-.D.....Q.L...   CPHL9 6
..................N.T...K...T...........R...............-.......D.L.N.   CPHL3 48
..................N.T...K...T...........R...............-..I.T..D.L.N.   CPHL3 5
..................N.T...K...T...........R...............-.......D.L.N.   CPHL3 6
...........Y.....K......................R...............-..I......ISN.   CPHL12
...........Y............................R...............-..I......ISN.   CPHL13 2
...........Y.....K......................R...............-..I......ISN.   CPHL13 8
.......T....VL......E.S.K.T.............R...............-...I.......N.   CPHL14 15
....Y........L........S.KEV.............R...............-...I.......N.   CPHL14 37
.............L........S.KEL.............R...............-...I.......N.   CPHL14 44
.................R......................................-..II.....I...   CPHL4 2
.................R......................................-..II.....I...   CPHL4 35
.................R.....................V................-..II.....I...   CPHL4 6
................FR......................................-..II.....I...   CPHL5 1
.................R......................................-..II.....I...   CPHL5 10
.................R......................................-..II.....I...   CPHL5 12
..................N.....................................-..II.....L.N.   CPHL19
-----.....................................T.............-..II.....L.N.   CPHL18
............................T...........R...............-..IIV...KL...   CPHL24
----------------------------------------------------------------------   CPHL20
.......T...Y........N...........S.......................-G.II.....L.N.   CPHL16
...........Y.....S................................TSQSSR-R.....FQ.....   CPHL17
.......T.....L...K.T............S..........A............G..IM.....I...   CPHL22
.................K.AE......................I............-R.IM.....I...   CPHL10 2
.................K.AE......................I............-R.IM.....I...   CPHL10 3
.................K.A......T......A.........I............-.DIMV....I...   CPHL10 7
.................K.AE......................I............-.GIM.....I...   CPHL11 1
.................K.AE......................I............-.GIM.....I...   CPHL11 2
.................K.AE.T....................I............-.GIM.....I...   CPHL11 5
...........Y......NNTS......N...........................-G.II.S...L...   CPHL15
------------------------------------------------------..-...I.....L.N.   CPHL21
```

```
AKTIIVQLNESVEINCTRPNNNTRKSIHIGP--GRAFYTTGEIIGDIRQAHCNISRAKWNNTLQQVAKKL   Consensus

..N.........A......S..R.RRVTM..--..VW...........K....L...D..K..W.IV...   CPHL1 1
..N.........V......S....R..S...--...W...........K....L...D.....R.IV...   CPHL1 18
..N.........V......S....R..S...--...W...........K....L...D.....R.IV...   CPHL1 19
..N.........V......S....R..S...--...W...........K....L...D.....W.IV...   CPHL1 4
..N.........V......S....R..S...--...W........N..K...TL..TH.....W..V...   CPHL1 43
..N.........V......S....R..S...--...W...........K....L...D..K..W.IV...   CPHL1 7
..N.........V......S....R..S...--...W...........K....L...D..K..W.IV...   CPHL1 cDNA
V......V........I..............--........Q......K......E...E..........   CPHL2 11
V......V......K.I..............--........Q.............EE..K....R.....   CPHL2 18
V......V......K.I..........P...--........Q.............EE..K....R.....   CPHL2 25
V......V......K.I..............--........Q.............EE..K....R.....   CPHL2 3
..I...H.....K.D..........G.....--.S......G.V............T......E.I....   CPHL7 13
..I......................G.....--.S.L....G.V...................E.I....   CPHL7 17
T.................A......G.R..HIG..S.....K.....KK.Y.D..........K.IV...   CPHL8 2
.......................S.G..L..GGA--.....R...N....Y............KK.V...   CPHL8 5
T.........T.............RG.....--........Q......K.Y............E.IV...   CPHL6 3
........................RG.....--........Q......K.Y............E.IV...   CPHL6 41
......H.................RG.....--........D......K.Y.T..........K.I....   CPHL6 6
........K.P..............G..M.L--........Q.............G.......R.I....   CPHL9 1
........K.P..............G..M.L--......................KT......R.I.E..   CPHL9 17
........K.P..............G..M.L--......................KT......R.I.E..   CPHL9 6
.........A...............G.R...--.....A..R..............E..H...K.IVE..   CPHL3 48
.........A...............G.....--.....A.................E......K.IVE..   CPHL3 5
.........A...............G.....--.....A................GE......K.IVE..   CPHL3 6
..N...H....I......L.........M.L--....H...D.............KV..E...GKI....   CPHL12
..N...H....I......L.........M.L--....H...D.............KV..E...GKI....   CPHL13 2
..N...H....I......L.........M.L--....H...D..............V..E...GKI.R..   CPHL13 8
V...................T..L.R.P...--.TS.....R..K...K......K...E..........   CPHL14 15
V..........................P...--........Q......N......KT..E..........   CPHL14 37
V..........................P...--........Q......N......KT..E..........   CPHL14 44
V.N.........L.T.A..S....Q.T....--.Q.L....Q...........V.EE..KK...R..I..   CPHL4 2
V.N.........L.T.A..S....Q.T....--.Q.L....Q...........V.KKD.S....R..I..   CPHL4 35
V.N.........L.T.A..S....Q.T....--.Q.L....Q...........V.KKD.S....R..I..   CPHL4 6
V.N.........P.T.V..Y....Q.T....--.Q.L..M-R...........V.A.Q..K......I..   CPHL5 1
V.N.........P.T.V..Y....Q.T....--.Q.L..M-K.............A.Q..K......I..   CPHL5 10
V.N.........P.T.V..Y....Q.TR...--.Q.L..M-R...........V.A.Q..K......I..   CPHL5 12
..IM..H.....Q......Y....QG.....--...LF..-..T...........E.L..K.........   CPHL19
V.I....V.K..T.D.I.......L......--.K.L.A.D-M.......Y....GVE.TKA.....S..   CPHL18
............T......Y.....GT.M..--.Q.W...-R.V...........G.S..K.....V...   CPHL24
---......K....I........I.GV....--.QT.....RV..........V.KKD..S.....D...   CPHL20
......H..K.........S..I.T.FR...--.QV.HK..S.T....K.Y.E.NGT...KA.K...E..   CPHL16
.......F.S..R...........RG.....--.Q...A..D........Y..V.KSQ...K..E..T..   CPHL17
..N....VT.P.Q......D...KRVRIMRIGP.Q...A-NN..........D....D..K...K...Q.   CPHL22
..I.....KTPIN.T.V...........F..--.Q...A..D...........VN.T..TEM.KE.KD..   CPHL10 2
V.I.....KTPIN.T.V...........F..--.Q...A..D...........VN.T..TEM.KE.KD..   CPHL10 3
..I.....KDPIN.T.V...........L..--.Q...A..D........Y...N.TS.TGM.KK.KD..   CPHL10 7
..I.....KTPIN.T.V...........F..--.Q...A..D........Y...N.TS.TGM.KE.RE..   CPHL11 1
..I.....KTPIN.T.V...........F..--.Q...A..D........Y...N.TS.TGM.KE.RE..   CPHL11 2
..I.....KTPIN.T.V...........F..--.Q...A..D........Y...N.TS.TGM.KE.RE..   CPHL11 5
..I...H.......V...........VR...--.Q...A.ND...........T                   CPHL15
I.....H.......M..........GMR...--.QT..A................GKN.TKI..G.SE..   CPHL21
```

```
REQ-----NNKTIIFNQSS-GGDPEIVMHSFNCGGEFFYCNTTKLFNSTWN-NST---NNTN--EENEN-   Consensus

...KQLQLE....V.....S.......LNT................D....VTER---P.NVTGMQK..-   CPHL1 1
...KQLQLE....V.....S.........T................D....VTER---P.NATGMQ...-   CPHL1 18
...KQLQLE....V.....S......I..T................D....VTER---P.NDTGMQ...-   CPHL1 19
...KQLQLE....V.....S.........T................D....VTER---P.NATGMQ...-   CPHL1 4
...KQLQLE....V.....S......I..T.............E.......ITEG---P.NVTG-GP..-   CPHL1 43
...KQLQLE....V.....S.........T................D....VTEK---P.NATGMQ...-   CPHL1 7
...KQLQLE....V.....S......I..T................D....VTER---P.NDTGMQ...-   CPHL1 cDNA
..K--FE--....A.T.P.-...T...................Q...N...-GTEWTNKWP.ST...---   CPHL2 11
..K--FE--....A.T.P.-......A................Q...N...-GTKWTNEWS.ST.Q.---   CPHL2 18
..K--FE--....A.T.P.-......A................Q...N...-GTKWTNEWS.ST.Q.---   CPHL2 25
.KK--FE--....A.T.P.-......A.P......K.......Q...N...-GTDWTYKWS.ST.K.---   CPHL2 3
K..--FG-.....V..S..-.............R................GV.G.VE-------.G.D.-   CPHL7 13
K..--FG-.....V.KS..-.............R................GV.G.VK-------...D.-   CPHL7 17
KV.--YE--....M.Q.P.-..E.VG...I...R.........Q......G------V.D.EELNKTDT-   CPHL8 2
...--FG--....V.D.P.-.........I...R.........Q.......--..WRV.D.EELNKTDI-   CPHL8 5
G..--YG-R......K...-.....V...I............SQ......KM.D.GGV.D.EGLQG..T-   CPHL6 3
G..--YG-R......K...-.....V...I............SQ......KM.D.GGV.E.EGLQG..T-   CPHL6 41
K..--FG-E....R.....-.............R.........Q......------GV.S.ETVQG.DT-   CPHL6 6
.A.--FG--..........-.....V.......R................--...GEM.S.EESK..DT-   CPHL9 1
...--FE--....T..H..-.............R................--.N.WEM.S.EESK..DT-   CPHL9 17
...--FE--....T..H..-.........N...R................--...WEM.S.EESK..DT-   CPHL9 6
...EPF--K..........-.............R................E-----NG...V-NNNT.I-   CPHL3 48
...EPF--K..........-.............R................G-----NG...VGNNNT-I-   CPHL3 5
...EPF--K..........-.............R................E-----NG...VGNDNT.I-   CPHL3 6
..H----FV....E.KK..-......E................H...L--KF...WNGTTGLN-DTK..-   CPHL12
..H----FV.R..E.KN..-......E................H...L--KF...WNGTTGLN-DT...-   CPHL13 2
..H----FV....E.KK..-......E................H...L--KF...WNGTTGLN-DTK..-   CPHL13 8
..K--FE--....A.....-......................S....N...-GTEWTNEWE.NT.-----   CPHL14 15
..K--FE--....A.....-......................S....H...-GTEWT----.IT.G.---   CPHL14 37
..K--FE--....A.....-......................S....H...-GTEWT----.IT.G.---   CPHL14 44
.NL----L.KT..K..-..S......TT..............S.......DA...--G.VSTNS..AN.T   CPHL4 2
KNL----L.KT.....-T.S......TT..............SR...N..HD.N.WNADS.WNDSN.TTA   CPHL4 35
KNL----L.KT..V..-T.S......TT..............SR...N..HT.N.WNADS.WNYSNLKQA   CPHL4 6
GNL----L.KTE.N..-T.S......TT..............SS...G..--...WNIR.NTEGAN.TQV   CPHL5 1
GNL----L.KTK....-T.S......TT...I..........SS...G..--...WNIR.NTGGAN.TQV   CPHL5 10
GNL----L.KTE.N..-T.S......TT..............SS...G..--...WN---NTEGAN.TQV   CPHL5 12
.DL----L.KTK...K-P.S......TT..............S.......ENSTNSTKV.NSADTNG---   CPHL19
.NL----FD.R.....-R.S......AT............D.SA......DNA.IQANST.LYDP-----   CPHL18
GGL----L.KTE.N.K-PHS......TT..............SG......QQSNGTWQRSD.ASTQES.I   CPHL24
.DL----F.QT..N.K-P.S......TT..............SG......--------I.GSMTSNSAEG   CPHL20
K.H----F.....S.-.PPS...L..T..H...R.........R...N.---------CIK.ETM.GC.G   CPHL16
.Q-----VF......DKH.-...L..TT.....R........S.......-------F.G.VQNDTGSDD   CPHL17
SNH----FP......TK..-...I..TT..............SG..---------WHME.GTLEYGGNGT   CPHL22
ETITK---T.--.T.DTPA-...L.VTTY....R.........G..-----N.NN----------T...D   CPHL10 2
ETITK---T.--.T.DTP.-.....VTT.....R.........G..-----N.NN----------T...D   CPHL10 3
ENITN---T.--.T.DTPA-...L.VTT.....R.........G..-----N.N-----------T...D   CPHL10 7
KNITD---TD.N.T.DTPA-...L..MT.....R........S...KIELSN..----------DSA..G   CPHL11 1
KNITD---TD.N.T.DTPA-...L..MT.....R........S...KIELSN..----------DSA..G   CPHL11 2
KNITD---TD.N.T.DTPA-...L..MT.....R........S...KIELSN..----------DSA..G   CPHL11 5
                                                                         CPHL15
..H----FP..K.TS-.PHS...L..TT....R........S.....S                         CPHL21
```

```
-ITLPCRIKQIINMWQEVGKAMYAPPISGLIRCSSNITGLLLTRDGG-NNN-TN-T--ETFRPGGGDMRD    Consensus

-.......R.FV.L.............E.Q..........I......-.T.E.KA.--........E...    CPHL1 1
-.......R..V.L.............E.Q..........I......-.T.E.KA.--........E...    CPHL1 18
-............L...............Q..........I......-.T.E..A.--........E...    CPHL1 19
-.......R....L...............Q..........I......-.T.E..A.--........E...    CPHL1 4
-............L.............E.Q..........I......-.T.E..N.--........E...    CPHL1 43
-.......R..V.L...............Q..........I......-.T.E..A.--........E...    CPHL1 7
-.......R..V.L...............Q..........I......-.T.E..A.--........E...    CPHL1 cDNA
-............................Q.........I..I....-...S..N.--.I...A......    CPHL2 11
-............................Q.........I..A....-...S..N.--.I...A......    CPHL2 18
-............................Q.........I..A....-...S..N.--.I...A......    CPHL2 25
-............................Q.........M..V....-...S..N.--.I...A......    CPHL2 3
-............................P.I..........V....-...S..TATE-I......N...    CPHL7 13
-............................P.S..........V....-S..D.---TE-I...E......    CPHL7 17
-.........F.....G............P.................-..KD.DN.TE-I..........    CPHL8 2
-.........F.....G............S.................-...D..NATE-I..........    CPHL8 5
-.I.......................................V....-EK.N..E.TE.I......N.K.    CPHL6 3
-.I.......................................V....-EK.N..E.TE.I......N.K.    CPHL6 41
-............................P.S.L.....I..V....-KK.N..D.TE-I...V..N...    CPHL6 6
-...Q........................P.....K......L....-...NK--.--.I..........    CPHL9 1
-............................T............I....-...D..E.--.I..........    CPHL9 17
-............................P............I....-...D..E.--.I..........    CPHL9 6
-..........................E.P.................-.S.R..E.----..........    CPHL3 48
-..........................E.P.................-...G..E.----..........    CPHL3 5
-..........................E.P.................-.G.G..E.----..........    CPHL3 6
-.........F.....G..........R............I......-.S.N..N.E--I...I..N.K.    CPHL12
-.........F.....G..........R....V.......I......-.S.N..N.E--I...L..N.K.    CPHL13 2
-.........F.....G..........R....V.......I......-.S.N..N.ESL...R---N.KT    CPHL13 8
-...S......V.................Q.........I..V....-...S..N.--.....A..N...    CPHL14 15
-...T..................................I..V....-...S..T.--.....A.         CPHL14 37
-...T..................................I..V....-...N..T.--.....A......    CPHL14 44
N.I.Q......................A...N.T.....II.....E-..S----SPN..S.........    CPHL4 2
N...Q......................A...K.......II......-.SS----NSS............    CPHL4 35
IF.IQ..L.N.........NV......A...K.......I...P...-.SS----NSS..........K.    CPHL4 6
-...Q.....L..L..G..Q.......E...T.T.....II......-..-----NSR............    CPHL5 1
-...Q........L..G..Q.......E...I.T.....II......-..-----NSS............    CPHL5 10
-...Q...T.F.TL..G..Q..C....RRI.I.T.....II......-..-----NSR.I..........    CPHL5 12
I.I..........R.............E...................VS..----SQN............    CPHL19
-..........................E.I.K...K...........T...G..NNSQ...........N    CPHL18
T..........................E...K...........K...-...----SPN..I.........    CPHL24
N...Q...................                                                  CPHL20
T.....K.........G..Q..........R.N.V.....I.......S..ASSE                   CPHL16
N..........VR.C.G...T..T...P.E.K.E.H...........P------NSTN.....A....W.    CPHL17
N..........V....R..R.......A.I.K.T.....IIM.....NT                         CPHL22
T..........VR...R..Q.......A.I...E.............ID.------TN............    CPHL10 2
T..........VR...R..Q.......A.I...E.............ID.------TN............    CPHL10 3
T..........VR...R..Q.......A.N...E.............R...-----TN............    CPHL10 7
T.....K....VR...R..Q.......A.N...E.............R...-----TN............    CPHL11 1
T.....K....VR...R..Q.......A.N...E.............R...-----TN.......A        CPHL11 2
T.....K....VR...R..Q.......A.N...E.............R...-----TN............    CPHL11 5
                                                                          CPHL15
                                                                          CPHL21
```

```
NWRSELYKYKVVKIEPLGVAPTKAKRRVVQREKR          Consensus

......................R...........          CPHL1 1
..................................          CPHL1 18
............R.........R...........          CPHL1 19
............R.........R...........          CPHL1 4
......................R...........          CPHL1 43
............R.....................          CPHL1 7
............R.....................          CPHL1 cDNA
..............V.........R.......T.          CPHL2 11
K.............V.........R.........          CPHL2 18
........................R.........          CPHL2 25
..............V.........R.........          CPHL2 3
............R.....I.S.R...........          CPHL7 13
............R.....I...R...........          CPHL7 17
............R...........N....K....          CPHL8 2
............R.D...................          CPHL8 5
............R.....................          CPHL6 3
............R.....................          CPHL6 41
............R.....................          CPHL6 6
............R.........R...........          CPHL9 1
............R.D.......R...........          CPHL9 17
............R.........R...........          CPHL9 6
.............V....................          CPHL3 48
.............V..V.................          CPHL3 5
.............V....................          CPHL3 6
............R.....YR                        CPHL12
...I........R................               CPHL13 2
-LEV........RM....A....T...A.               CPHL13 8
...I......L......                           CPHL14 15
                                            CPHL14 37
...........I..V.......SS                    CPHL14 44
.............L...............E....          CPHL4 2
.............................E....          CPHL4 35
.............................E....          CPHL4 6
......................R......G....          CPHL5 1
......................R......G....          CPHL5 10
I.............D...I..PR.T....G....          CPHL5 12
..................L...                      CPHL19
............R.....I.....R....D              CPHL18
....D.......R.....L...T......E....          CPHL24
                                            CPHL20
                                            CPHL16
..........I.E.Q.........R....E....          CPHL17
                                            CPHL22
............E.                              CPHL10 2
............E.                              CPHL10 3
...R.V                                      CPHL10 7
...........I.S                              CPHL11 1
                                            CPHL11 2
...........I.S                              CPHL11 5
                                            CPHL15
                                            CPHL21
```

Figure 3.5

Phylogenetic tree of all available CPHL sequences. The tree was constructed using the neighbour-joining method described in section 2.4.4. Numbered brackets at the right indicate the six main groups these sequences fall into. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.

Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

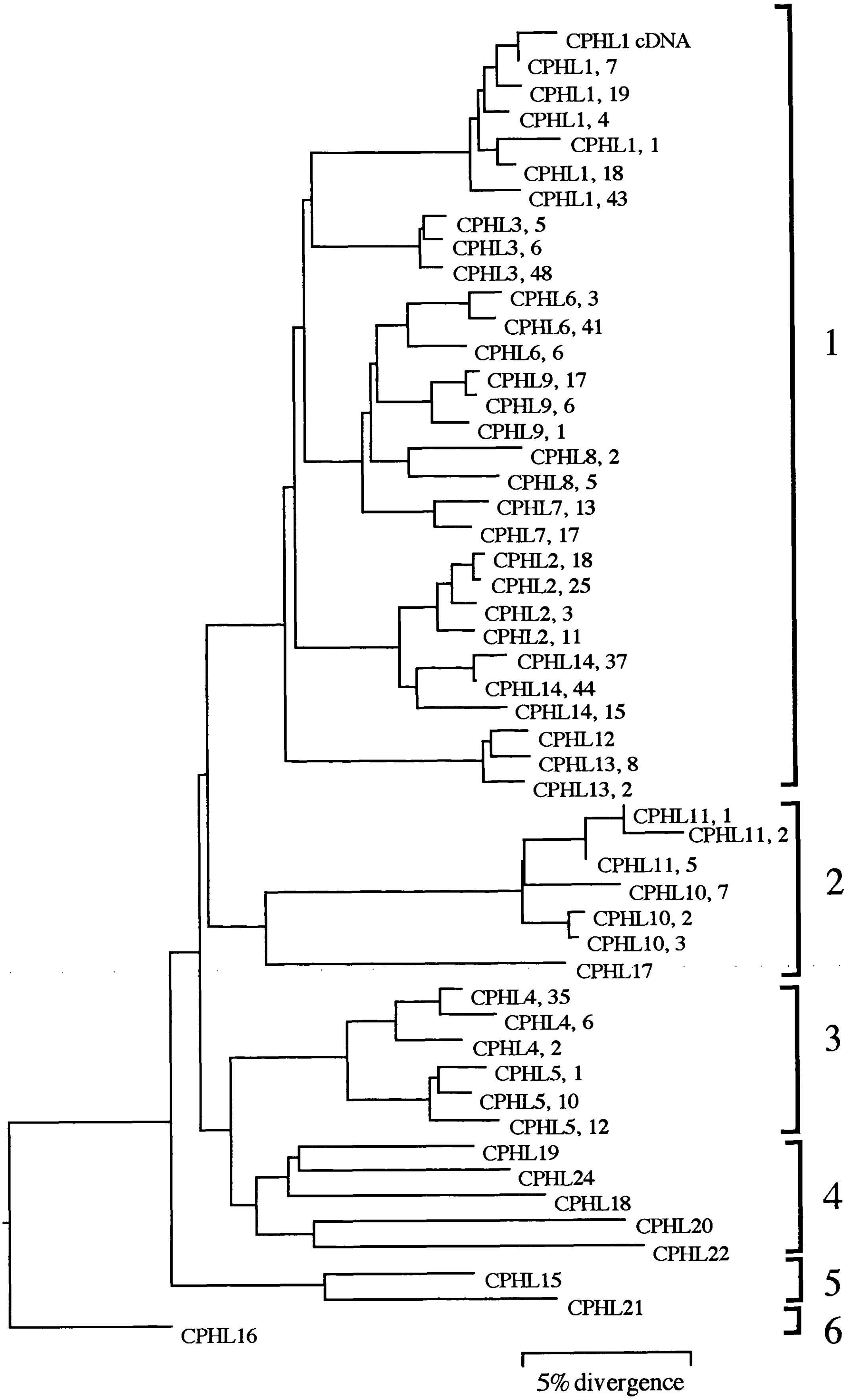
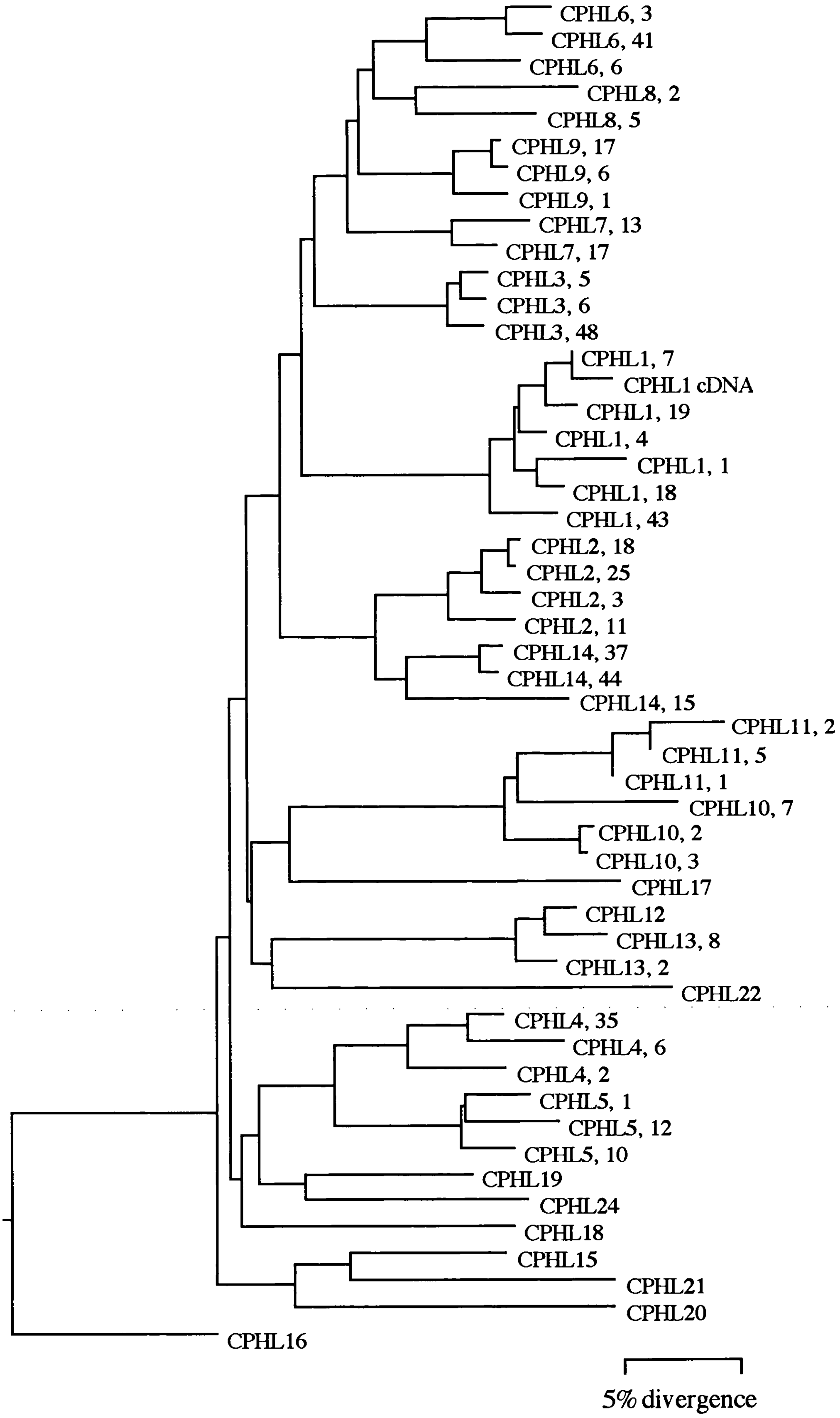

CPHL1 cDNA
CPHL1, 7
CPHL1, 19
CPHL1, 4
CPHL1, 1
CPHL1, 18
CPHL1, 43
CPHL3, 5
CPHL3, 6
CPHL3, 48
CPHL6, 3
CPHL6, 41
CPHL6, 6
CPHL9, 17
CPHL9, 6
CPHL9, 1
CPHL8, 2
CPHL8, 5
CPHL7, 13
CPHL7, 17
CPHL2, 18
CPHL2, 25
CPHL2, 3
CPHL2, 11
CPHL14, 37
CPHL14, 44
CPHL14, 15
CPHL12
CPHL13, 8
CPHL13, 2
CPHL11, 1
CPHL11, 2
CPHL11, 5
CPHL10, 7
CPHL10, 2
CPHL10, 3
CPHL17
CPHL4, 35
CPHL4, 6
CPHL4, 2
CPHL5, 1
CPHL5, 10
CPHL5, 12
CPHL19
CPHL24
CPHL18
CPHL20
CPHL22
CPHL15
CPHL21
CPHL16
1
2
3
4
5
6
5% divergence

Figure 3.6

Phylogenetic tree of all CPHL translated amino acid sequences. The tree was constructed using the neighbour-joining method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.

Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.


CPHL6, 3
CPHL6, 41
CPHL6, 6
CPHL8, 2
CPHL8, 5
CPHL9, 17
CPHL9, 6
CPHL9, 1
CPHL7, 13
CPHL7, 17
CPHL3, 5
CPHL3, 6
CPHL3, 48
CPHL1, 7
CPHL1 cDNA
CPHL1, 19
CPHL1, 4
CPHL1, 1
CPHL1, 18
CPHL1, 43
CPHL2, 18
CPHL2, 25
CPHL2, 3
CPHL2, 11
CPHL14, 37
CPHL14, 44
CPHL14, 15
CPHL11, 2
CPHL11, 5
CPHL11, 1
CPHL10, 7
CPHL10, 2
CPHL10, 3
CPHL17
CPHL12
CPHL13, 8
CPHL13, 2
CPHL22
CPHL4, 35
CPHL4, 6
CPHL4, 2
CPHL5, 1
CPHL5, 12
CPHL5, 10
CPHL19
CPHL24
CPHL18
CPHL15
CPHL21
CPHL20
CPHL16
5% divergence

Subtypes may be assigned on the basis of different groups of sequences clustering on discrete branches of the tree, figure 3.5 showed that the sequences appeared to fall into 6 distinct subtypes. McCutchan and colleagues described gp120 sequence variations of 20% between subtypes with up to 10% variation within a subtype in the M group (171). Based on these criteria, the sequences in groups 1 to 6 did not fall neatly within these definitions. This was most likely due to the small amount of sequence data available for some specimens. These findings were reflected in the amino acid tree (figure 3.6), which did not mirror the results shown in figure 3.5. In the amino acid tree (figure 3.6) several sequences or groups of sequences clustered onto different branches when compared with the nucleotide tree (figure 3.5). To determine the accepted subtypes of the CPHL sequences and to ascertain whether the inclusion in the nucleotide and amino acid trees of more diverse subtypes changed their topology, representative sequences of the currently known subtypes from the Los Alamos HIV database were included in the phylogenetic analysis. The results of the nucleotide analysis are shown in figure 3.7 and the amino acid analysis in figure 3.8.

Figure 3.7

Phylogenetic tree of all CPHL nucleotide sequences and subtype representatives from the Los Alamos HIV database (shown in bold). The tree was constructed using the neighbour-joining method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.

Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

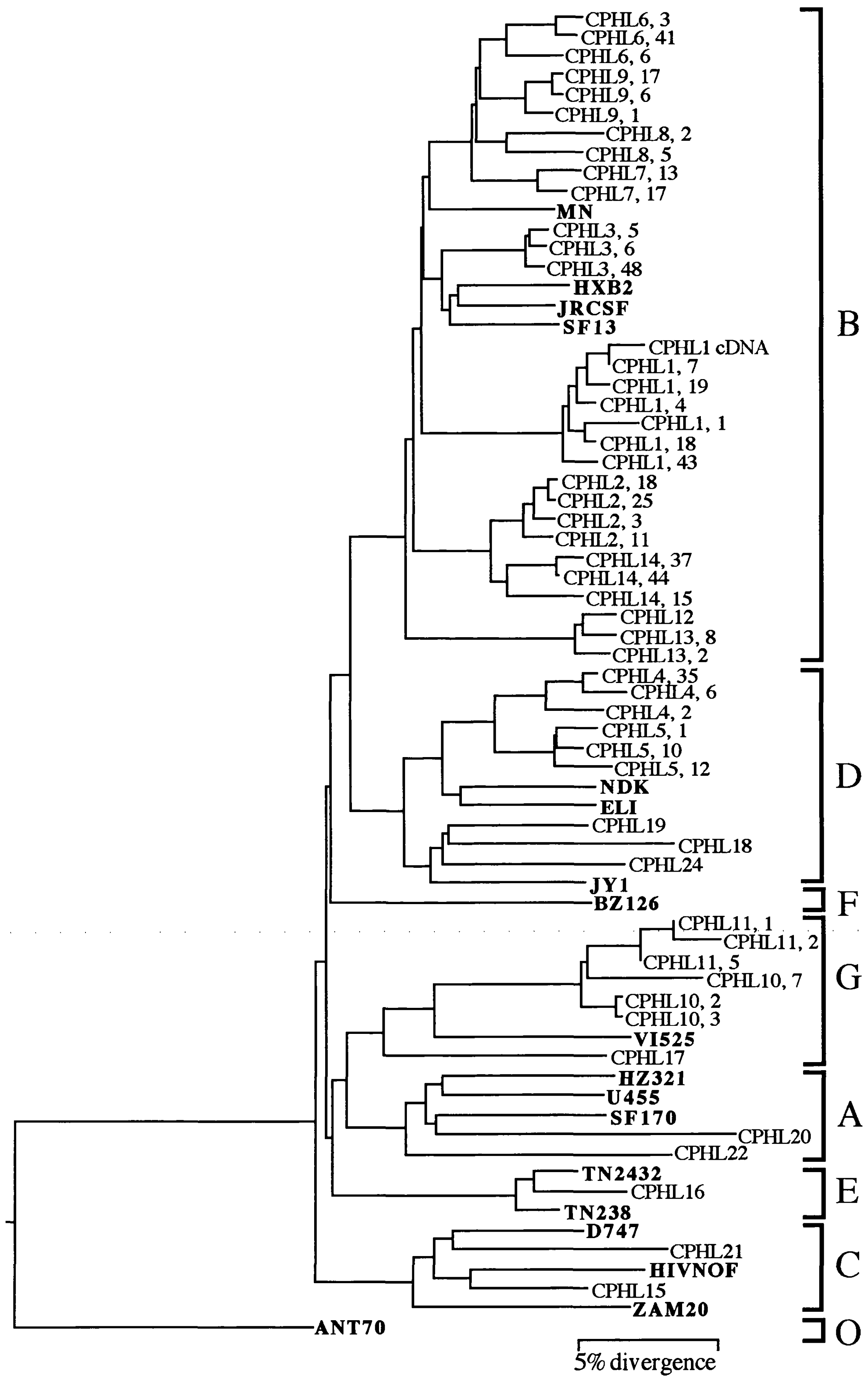
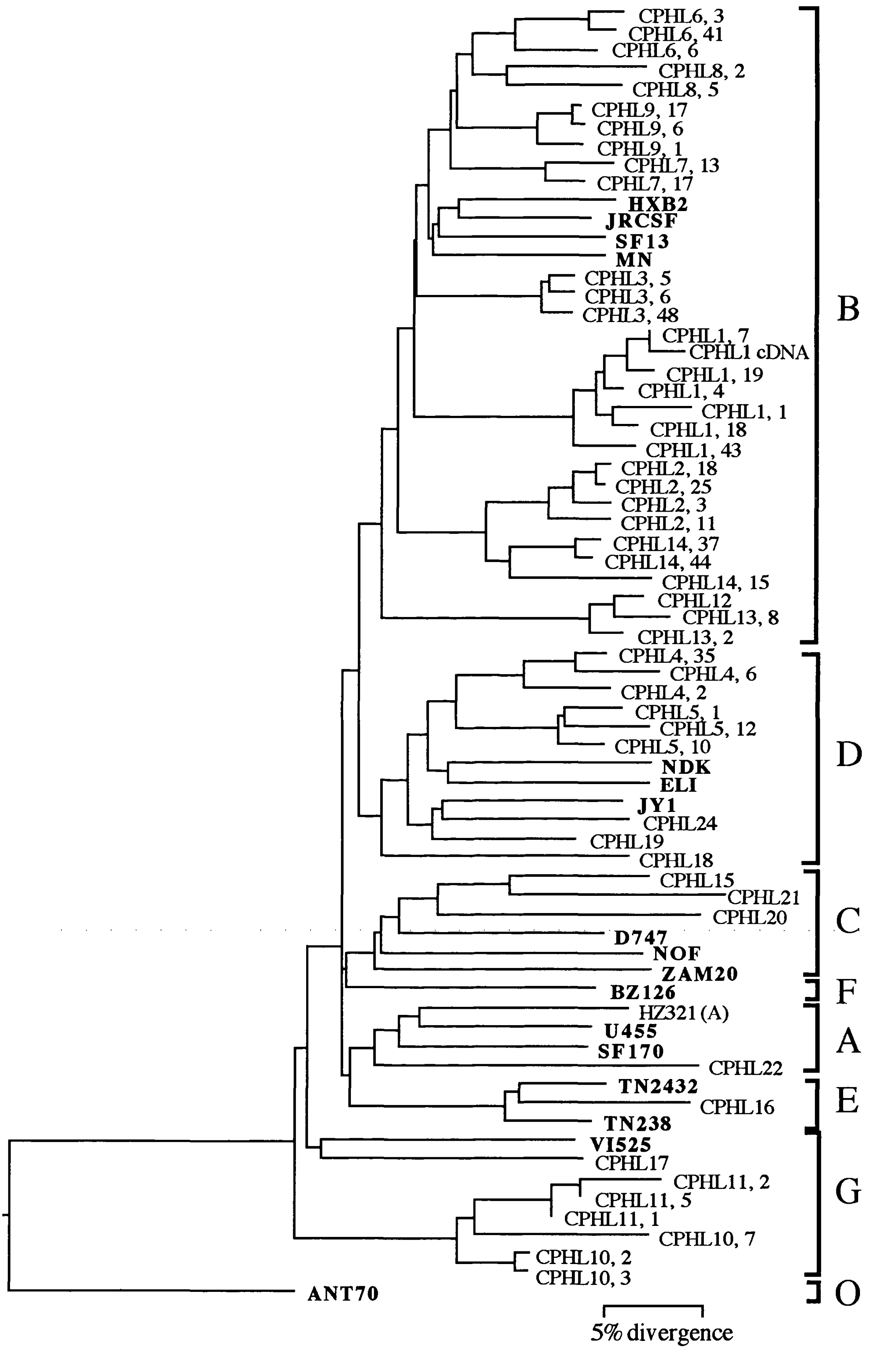

CPHL6, 3
CPHL6, 41
CPHL6, 6
CPHL9, 17
CPHL9, 6
CPHL9, 1
CPHL8, 2
CPHL8, 5
CPHL7, 13
CPHL7, 17
MN
CPHL3, 5
CPHL3, 6
CPHL3, 48
HXB2
JRCSF
SF13
CPHL1 cDNA
CPHL1, 7
CPHL1, 19
CPHL1, 4
CPHL1, 1
CPHL1, 18
CPHL1, 43
CPHL2, 18
CPHL2, 25
CPHL2, 3
CPHL2, 11
CPHL14, 37
CPHL14, 44
CPHL14, 15
CPHL12
CPHL13, 8
CPHL13, 2
CPHL4, 35
CPHL4, 6
CPHL4, 2
CPHL5, 1
CPHL5, 10
CPHL5, 12
NDK
ELI
CPHL19
CPHL18
CPHL24
JY1
BZ126
CPHL11, 1
CPHL11, 2
CPHL11, 5
CPHL10, 7
CPHL10, 2
CPHL10, 3
VI525
CPHL17
HZ321
U455
SF170
CPHL20
CPHL22
TN2432
CPHL16
TN238
D747
CPHL21
HIVNOF
CPHL15
ZAM20
ANT70
5% divergence
B
D
F
G
A
E
C
O

Figure 3.8

Phylogenetic tree of all CPHL amino acid sequences with subtype representatives from the Los Alamos HIV Database (shown in bold). Other details as for figure 3.7.

Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

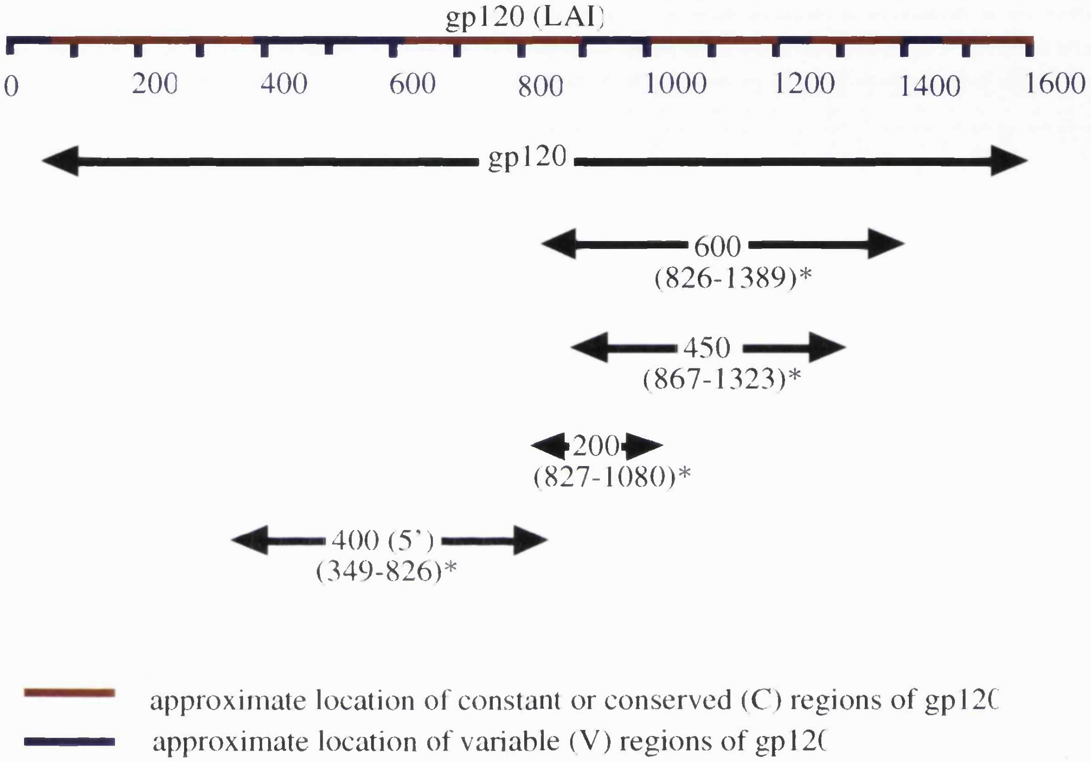

CPHL6, 3
CPHL6, 41
CPHL6, 6
CPHL8, 2
CPHL8, 5
CPHL9, 17
CPHL9, 6
CPHL9, 1
CPHL7, 13
CPHL7, 17
HXB2
JRCSF
SF13
MN
CPHL3, 5
CPHL3, 6
CPHL3, 48
CPHL1, 7
CPHL1 cDNA
CPHL1, 19
CPHL1, 4
CPHL1, 1
CPHL1, 18
CPHL1, 43
CPHL2, 18
CPHL2, 25
CPHL2, 3
CPHL2, 11
CPHL14, 37
CPHL14, 44
CPHL14, 15
CPHL12
CPHL13, 8
CPHL13, 2
CPHL4, 35
CPHL4, 6
CPHL4, 2
CPHL5, 1
CPHL5, 12
CPHL5, 10
NDK
ELI
JY1
CPHL24
CPHL19
CPHL18
CPHL15
CPHL21
CPHL20
D747
NOF
ZAM20
BZ126
HZ321 (A)
U455
SF170
CPHL22
TN2432
CPHL16
TN238
VI525
CPHL17
CPHL11, 2
CPHL11, 5
CPHL11, 1
CPHL10, 7
CPHL10, 2
CPHL10, 3
ANT70
B
D
C
F
A
E
G
O
5% divergence

ii) use of full gp120 and subsets of gp120 sequence for subtyping

The nucleotide tree (figure 3.7) showed that the sequences fell into subtypes A, B, C, D, E and G (see table 3.2 'HIV specimens in SE England'). To investigate how robust these subtype attributes were and how much gp120 sequence and from which region is needed to produce a reliable tree, subsets of gp120 sequences were analysed. Phylogenetic trees of 200, 450, 600 nucleotides (3'), 400 nucleotides (5') of gp120 as well as full length gp120 where available and their translated protein sequences were analysed and compared with reference sequences from the Los Alamos HIV database using the methods described in section 2.4 (figs 3.11-3.22 respectively). For these partial regions of gp120, bootstrapping (section 2.4.7) was used to produce multiple data sets (100 for each tree) which were analysed to produce a consensus nucleotide tree for each region. In the bootstrapped phylogenetic tree figures the numbers at the nodes indicate the number of times the group consisting of the species to the right of that node occurred out of 100 trees. Figure 3.9 shows a schematic diagram of approximate nucleotide positions of regions examined, with reference to gp120 of the LAI strain of HIV. The partial gp120 sequence PCR primers used were those described by Delwart *et al.* (55) (see section 2.3.12). These partial gp120 sequences were generated from PCR products concurrently amplified for another study comparing HMA and sequencing for subtyping (194). A region from *gag/pol* (p6/protease) from these specimens was also sequenced, these results are discussed in section b) of this chapter. Partial p24 *gag* sequences for these specimens were sequenced by K. Barlow (11) and it was found that by both nucleotide and amino acid sequence alignment, the viruses were grouped similarly for the three regions except for specimen CPHL16. This specimen had group A *gag* sequences (partial p24 and p6) and group E *env* (gp120) sequences. This is consistent with previous findings that the *env* E subtype is not found in *gag* coding sequences (126). This specimen came from a patient who is believed to have been infected in Thailand.

Figure 3.9

Schematic of gp120 regions analysed


gp120 (LAI)
0
200
400
600
800
1000
1200
1400
1600
gp120
600
(826-1389)*
450
(867-1323)*
200
(827-1080)*
400 (5')
(349-826)*
approximate location of constant or conserved (C) regions of gp12(
approximate location of variable (V) regions of gp12(

* denotes approximate nucleotide position of region analysed with reference to LAI gp12(

Table 3.2

HIV subtypes from patients in SE England: Bootstrap values

	number of gp120 nucleotides analysed									
	200		450		600		1500 (full)		400 (5')	
Patient	subtype	%*	subtype	%*	subtype	%*	subtype	%*	subtype	%*
95-12310	C	98	C	97	n/a	n/a	n/a	n/a	n/a	n/a
93-00513	C	98	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
94-11643	E	100	E	98	E	100	n/a	n/a	E	100
95-12313	A	79	A	47	A	98	n/a	n/a	A	40
94-33422	A	79	A	47	A	98	G	96	G	100
CPHL1	B	66	B	100	B	100	B	100	B	100
95-29517	D	46	A	47	D	76	n/a	n/a	n/a	n/a
94-47971	D	46	D	92	D	76	D	100	D	41
CPHL4	D	46	D	92	D	76	D	100	D	41
94-44501	n/a	n/a	A	45	n/a	n/a	n/a	n/a	n/a	n/a

*Percentages taken from figures 3.10-3.20. This figure represents the percentage of times the species to the right of the node (i.e. HIV-1 strains of a distinct subtype) occurred out of 100 randomly generated trees for each of the regions stated. Only percentages of 95 and above are considered statistically valid.

n/a = not amplifiable

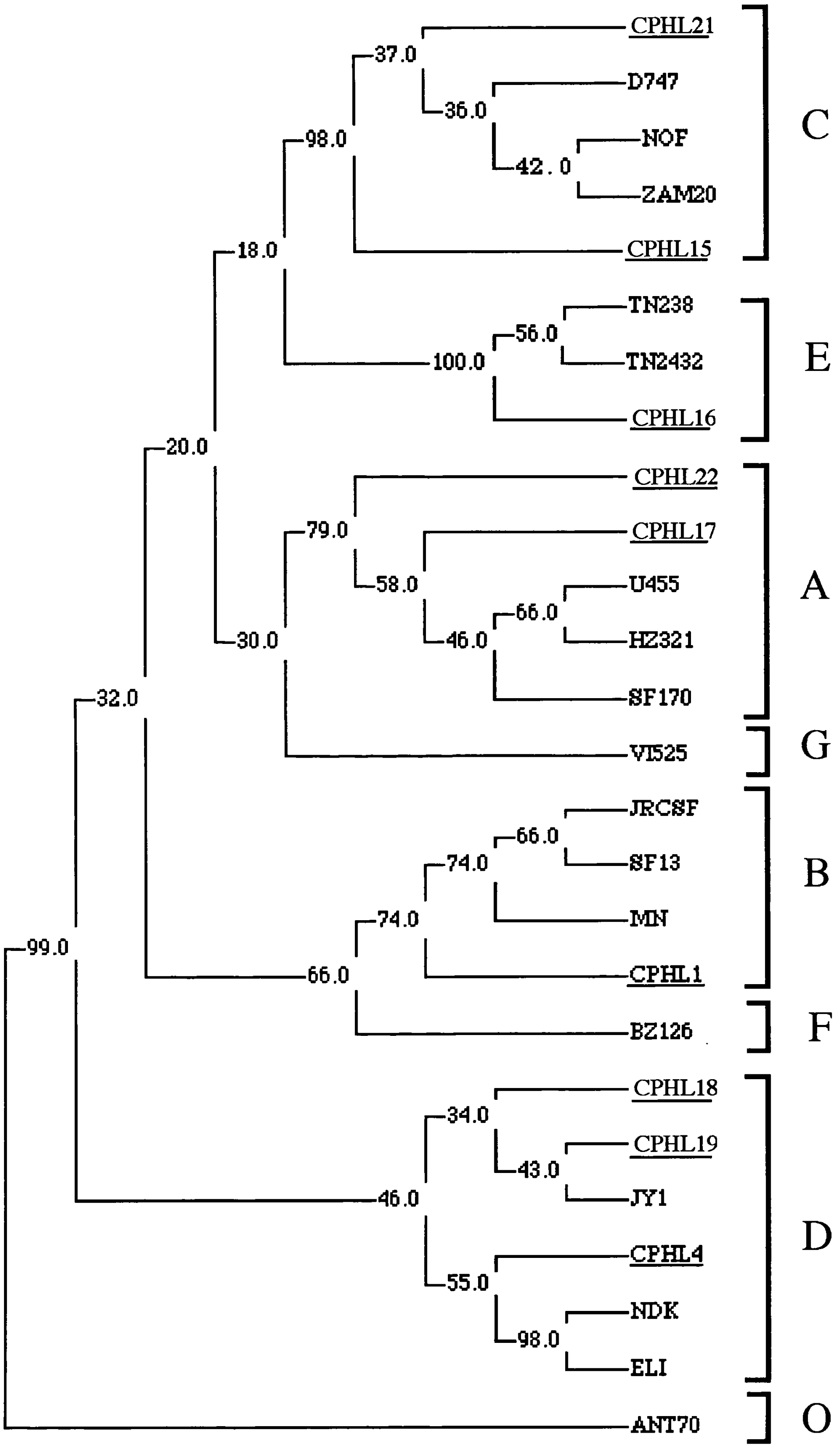
Use of subsets of gp120 for subtyping

Phylogenetic trees constructed with sequences of 200 nucleotides (see figure 3.9)

Figure 3.10

Phylogenetic tree of a 200 nucleotide region of gp120, translating to approximately 68 amino acids, comprising 25 amino acids of C2, V3, and 7 amino acids of C3. When translated to protein, this sequence comprises 47% amino acids from constant regions and 53% amino acids from variable regions. Sequences (underlined) were generated using the methods described in sections 2.3.4, 2.3.5, 2.3.7-10 and analysed and compared with reference sequences from the Los Alamos HIV database using the methods described in section 2.4. Bootstrapping (section 2.4.7) was used to produce multiple data sets (100 for each tree) which were analysed to produce a consensus nucleotide tree. The numbers at the nodes indicate the number of times the group consisting of the species which are to the right of that node occurred out of 100 trees.

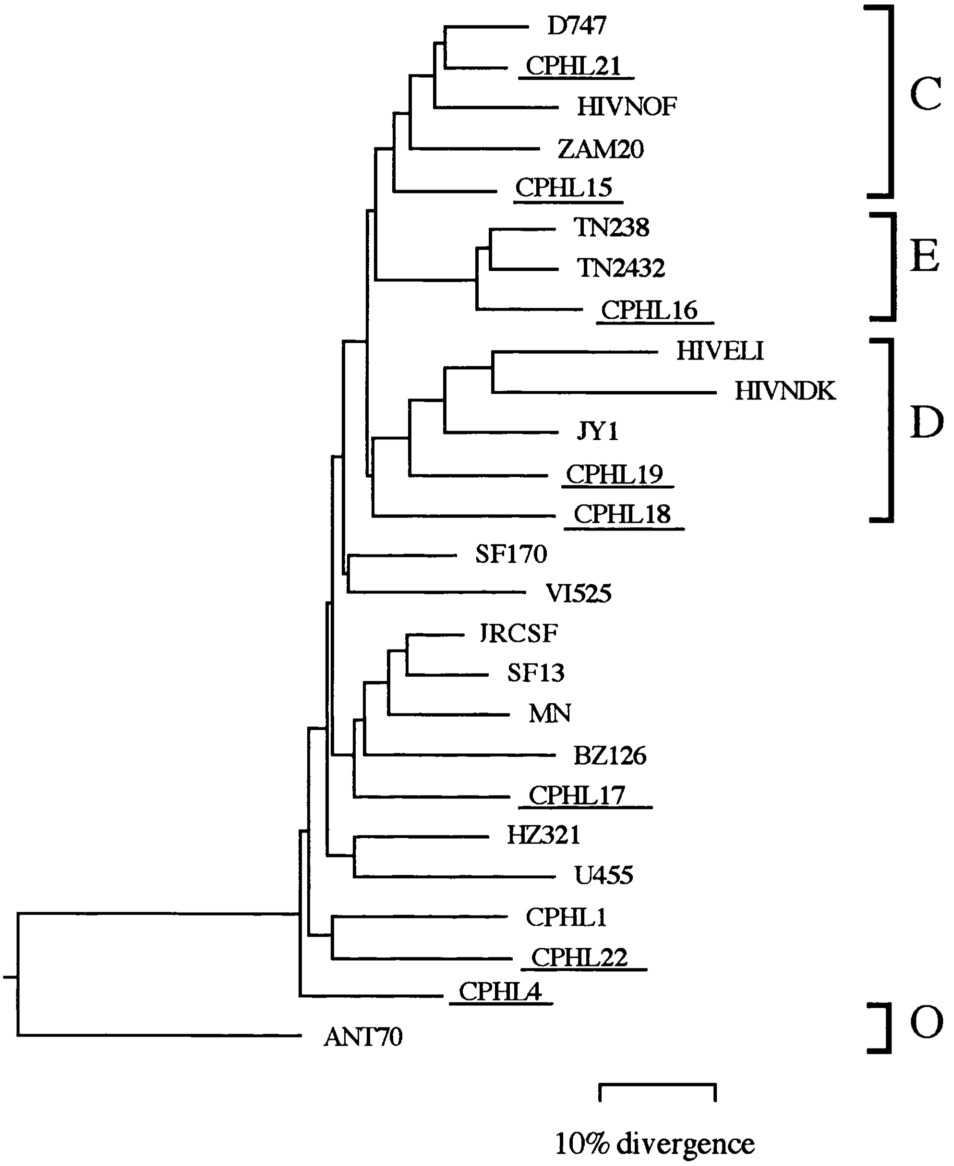
Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.



CPHL21
37.0
D747
36.0
98.0
NOF
42.0
ZAM20
CPHL15
C
18.0
TN238
56.0
TN2432
100.0
CPHL16
E
20.0
CPHL22
79.0
CPHL17
58.0
U455
66.0
HZ321
30.0
46.0
SF170
A
32.0
VI525
G
JRCSF
66.0
SF13
74.0
MN
74.0
99.0
CPHL1
B
66.0
BZ126
F
CPHL18
34.0
CPHL19
43.0
JY1
46.0
CPHL4
D
55.0
NDK
98.0
ELI
ANT70
O

Figure 3.11

Phylogenetic tree of translated amino acid sequences depicted in figure 3.10, constructed using the method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.

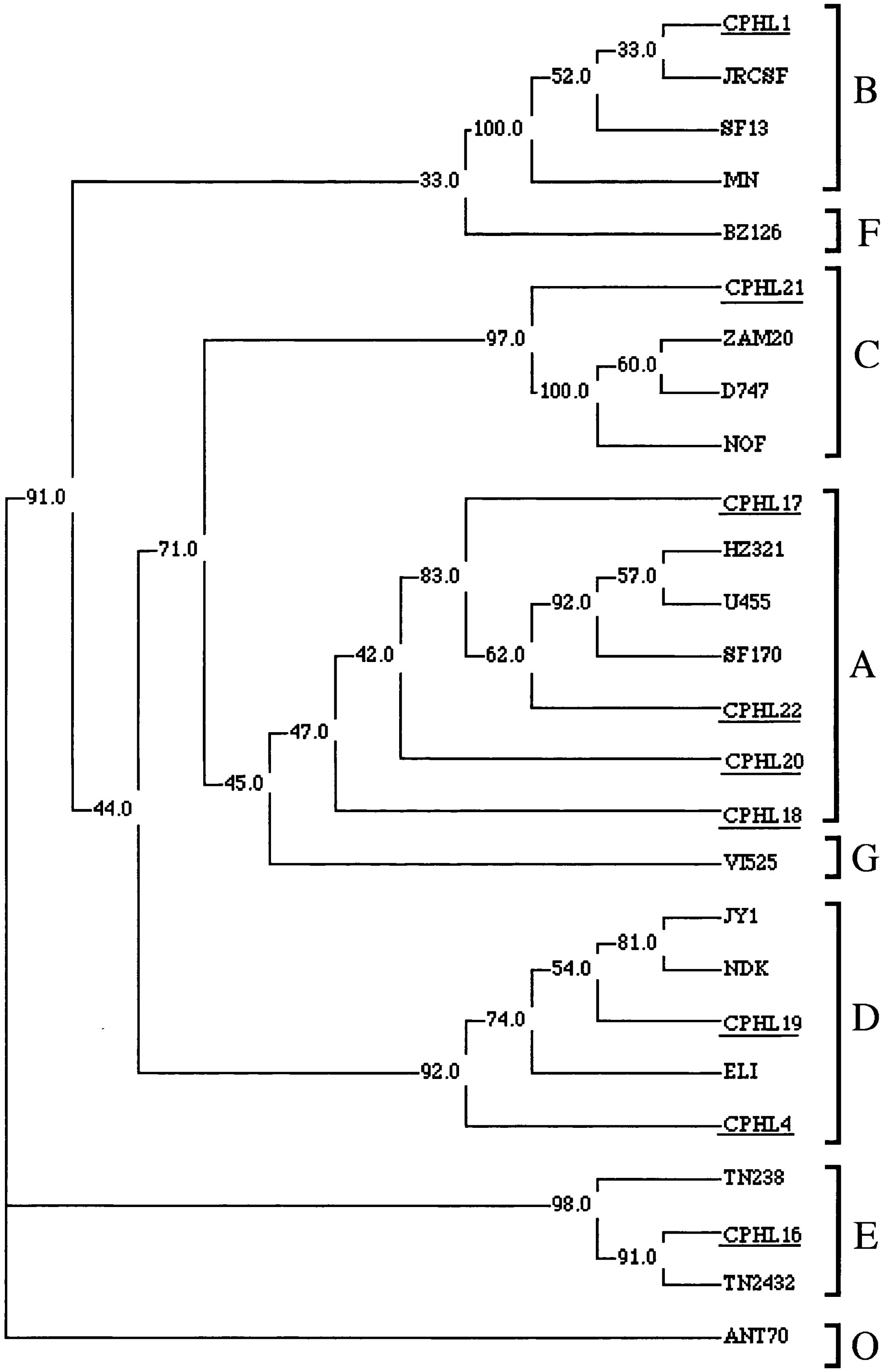


Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Phylogenetic trees constructed with sequences of 450 nucleotides (see figure 3.9)

Figure 3.12

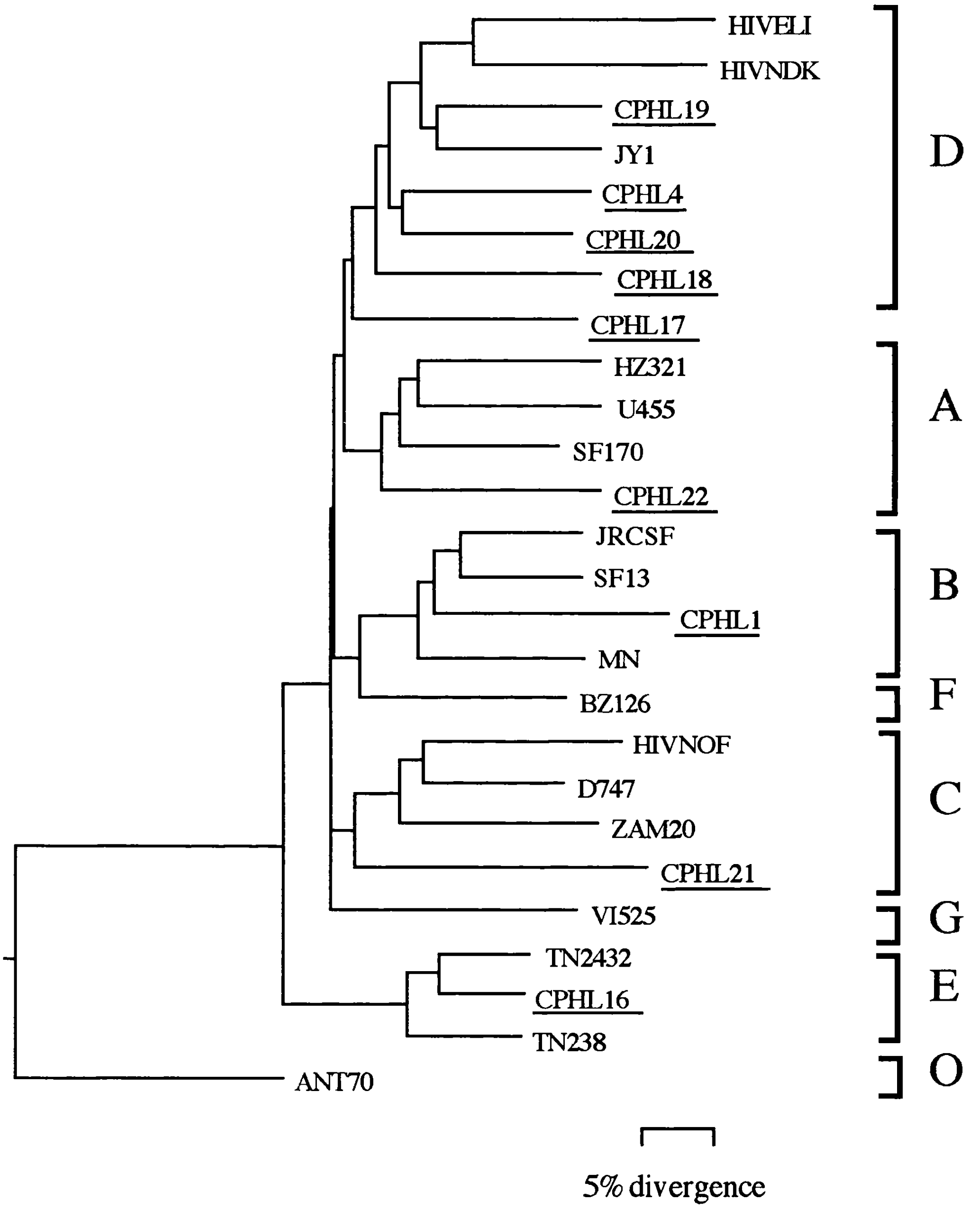
Phylogenetic tree of a 450 nucleotide region of gp120, translating to approximately 152 amino acids, comprising 12 amino acids of C2, V3, C3, V4, and 17 amino acids of C4. When translated to protein, this sequence comprises 55% amino acids from constant regions and 45% amino acids from variable regions. Other details as for figure 3.10. Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.



33.0
52.0
100.0
33.0
CPHL1
JRCSF
SF13
MN
B
BZ126
F
CPHL21
97.0
ZAM20
60.0
D747
100.0
NOF
C
91.0
CPHL17
71.0
HZ321
57.0
U455
83.0
92.0
42.0
62.0
SF170
A
CPHL22
47.0
CPHL20
45.0
44.0
CPHL18
VI525
G
JY1
81.0
NDK
54.0
CPHL19
74.0
D
92.0
ELI
CPHL4
TN238
98.0
E
CPHL16
91.0
TN2432
ANT70
O

Figure 3.13

Phylogenetic tree of translated amino acid sequences depicted in figure 3.12, constructed using the method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Phylogenetic trees constructed with sequences of 600 nucleotides (see figure 3.9)

Figure 3.14

Phylogenetic tree of a 600 nucleotide region of gp120, translating to approximately 189 amino acids, comprising 25 amino acids of C2, V3, C3, V4, C4 and 2 amino acids of V5. When translated to protein, this sequence comprises 62% amino acids from constant regions and 38% amino acids from variable regions. Other details as for figure 3.10.

Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

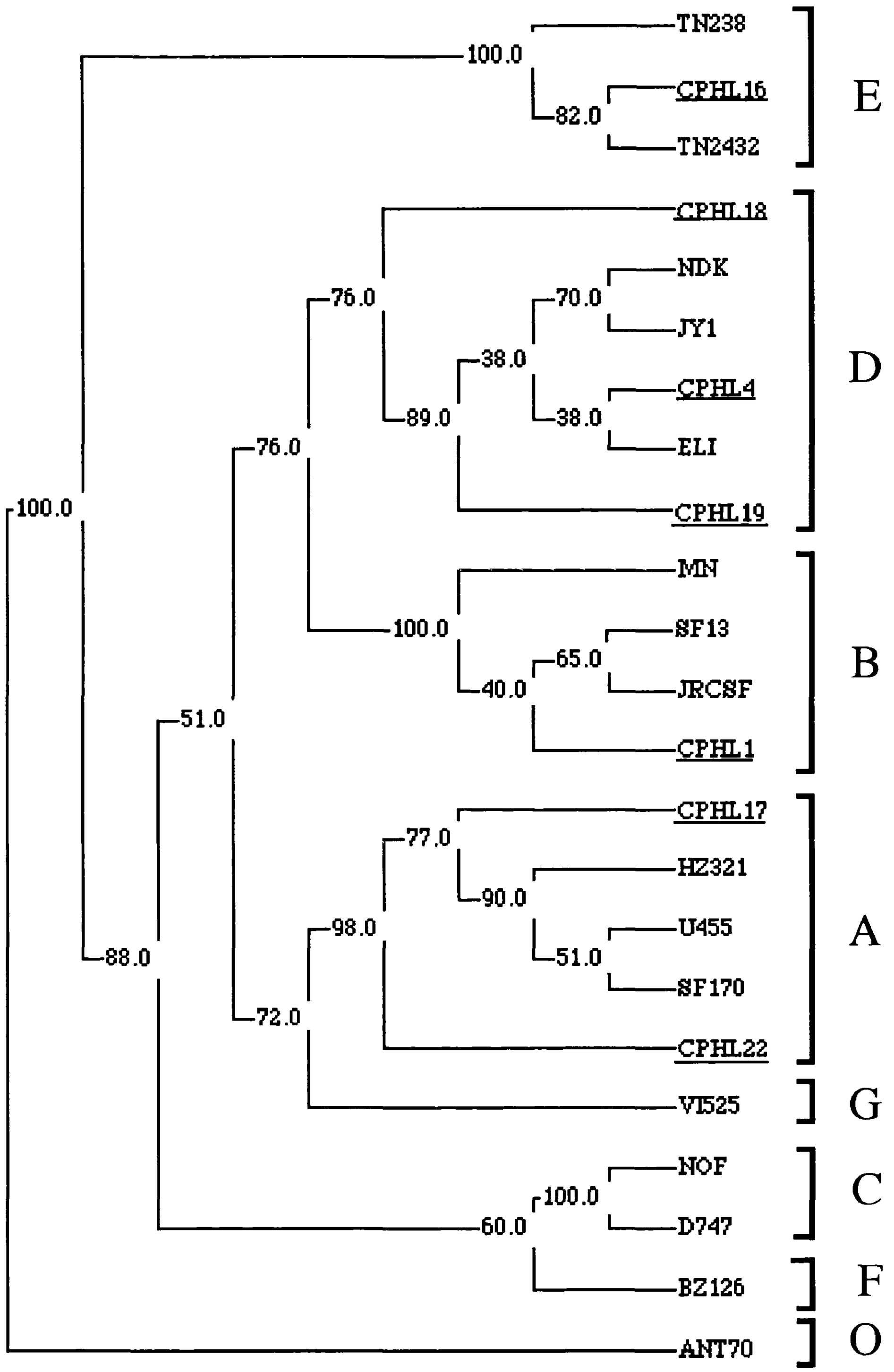
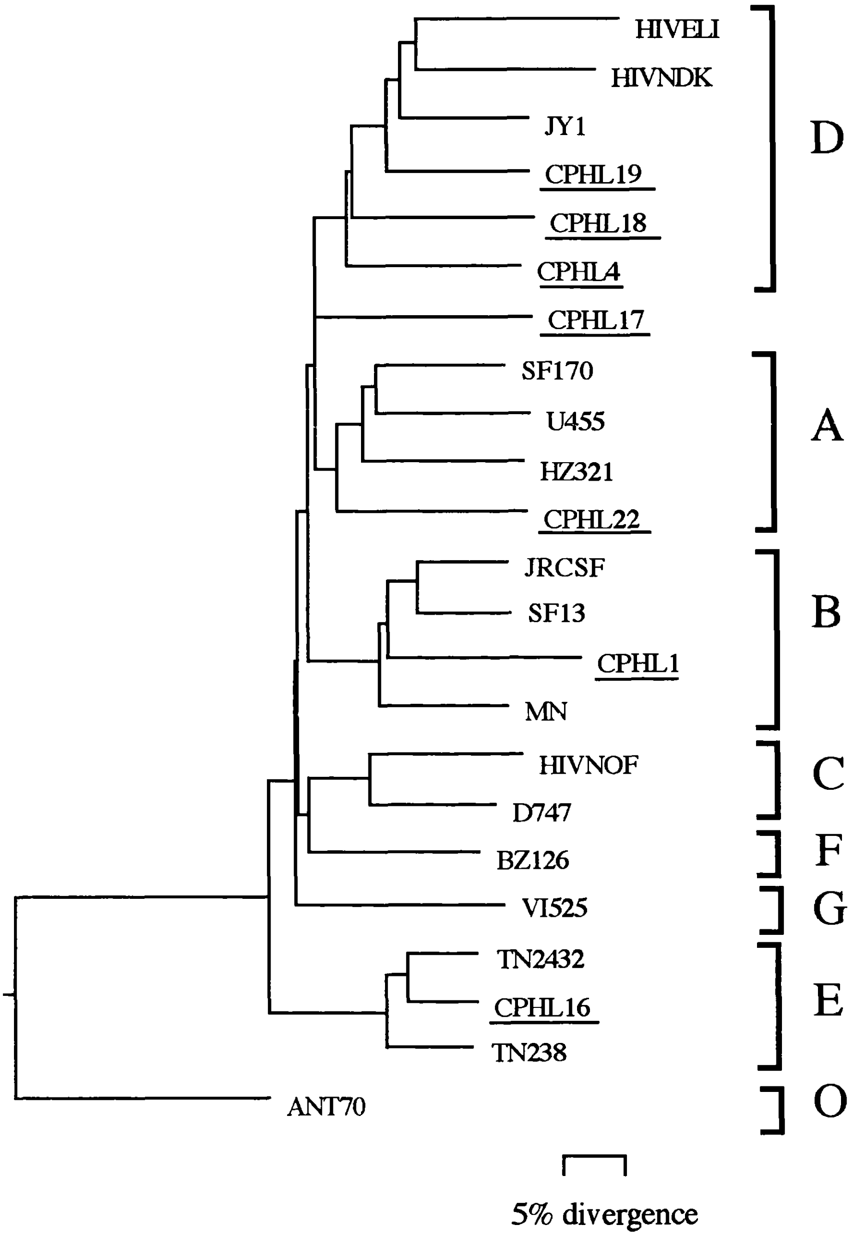

TN238
100.0
CPHL16
82.0
TN2432
E
CPHL18
NDK
76.0
70.0
JY1
38.0
D
CPHL4
89.0
38.0
76.0
ELI
100.0
CPHL19
MN
SF13
100.0
65.0
B
40.0
JRCSF
51.0
CPHL1
CPHL17
77.0
HZ321
90.0
98.0
U455
A
88.0
51.0
SF170
72.0
CPHL22
VI525
G
NOF
100.0
C
60.0
D747
BZ126
F
ANT70
O

Figure 3.15

Phylogenetic tree of translated amino acid sequences depicted in figure 3.14, constructed using the method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Phylogenetic trees constructed with sequences of 450 nucleotides (5' :see figure 3.9)

Figure 3.16

Phylogenetic tree of a 400 nucleotide region of the 5' end of gp120, translating to approximately 172 amino acids, comprising 16 amino acids of C1, V1, V2, and 73 amino acids of C2. When translated to protein, this sequence comprises 60% amino acids from constant regions and 40% amino acids from variable regions. Other details as for figure 3.10.

Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

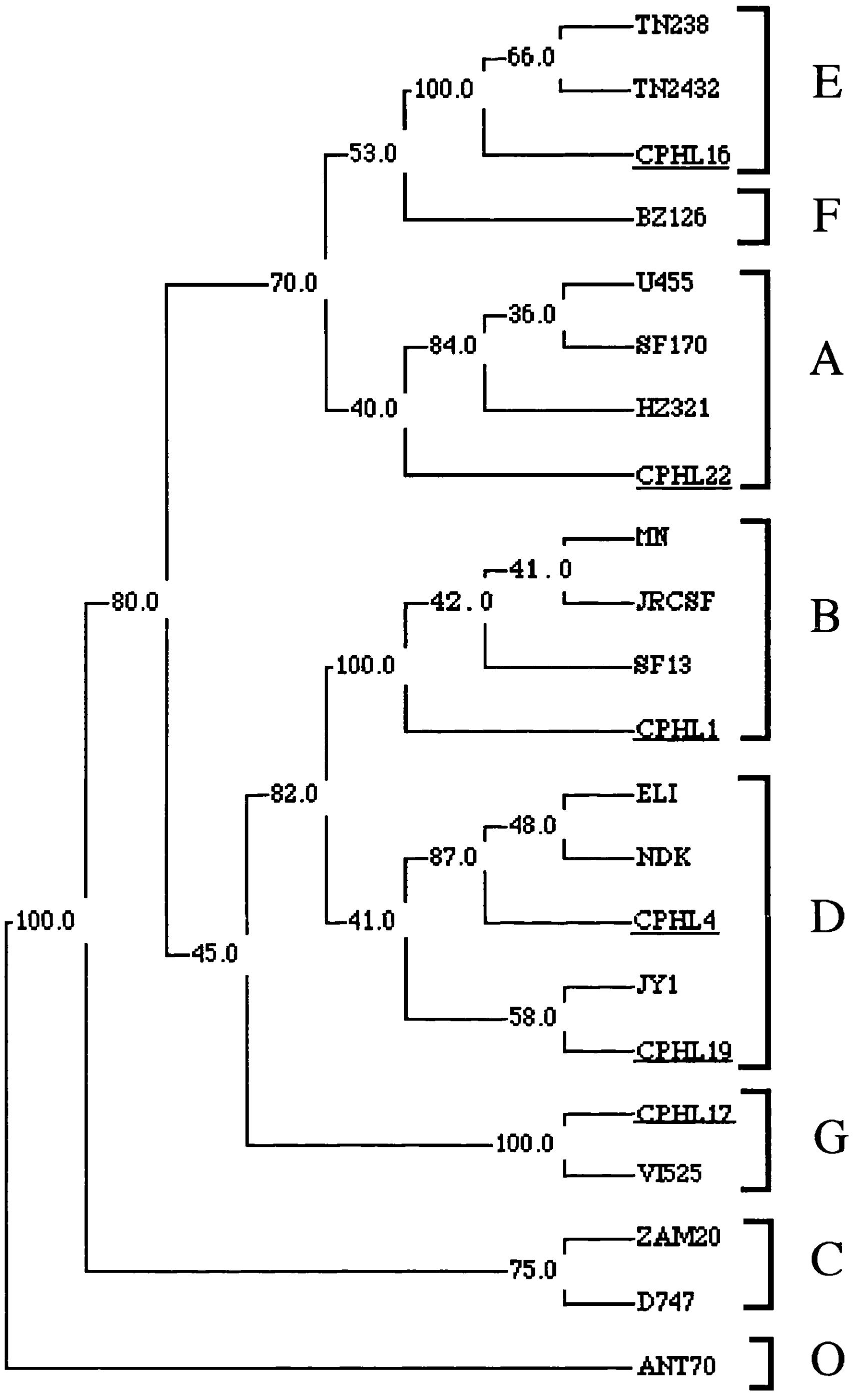
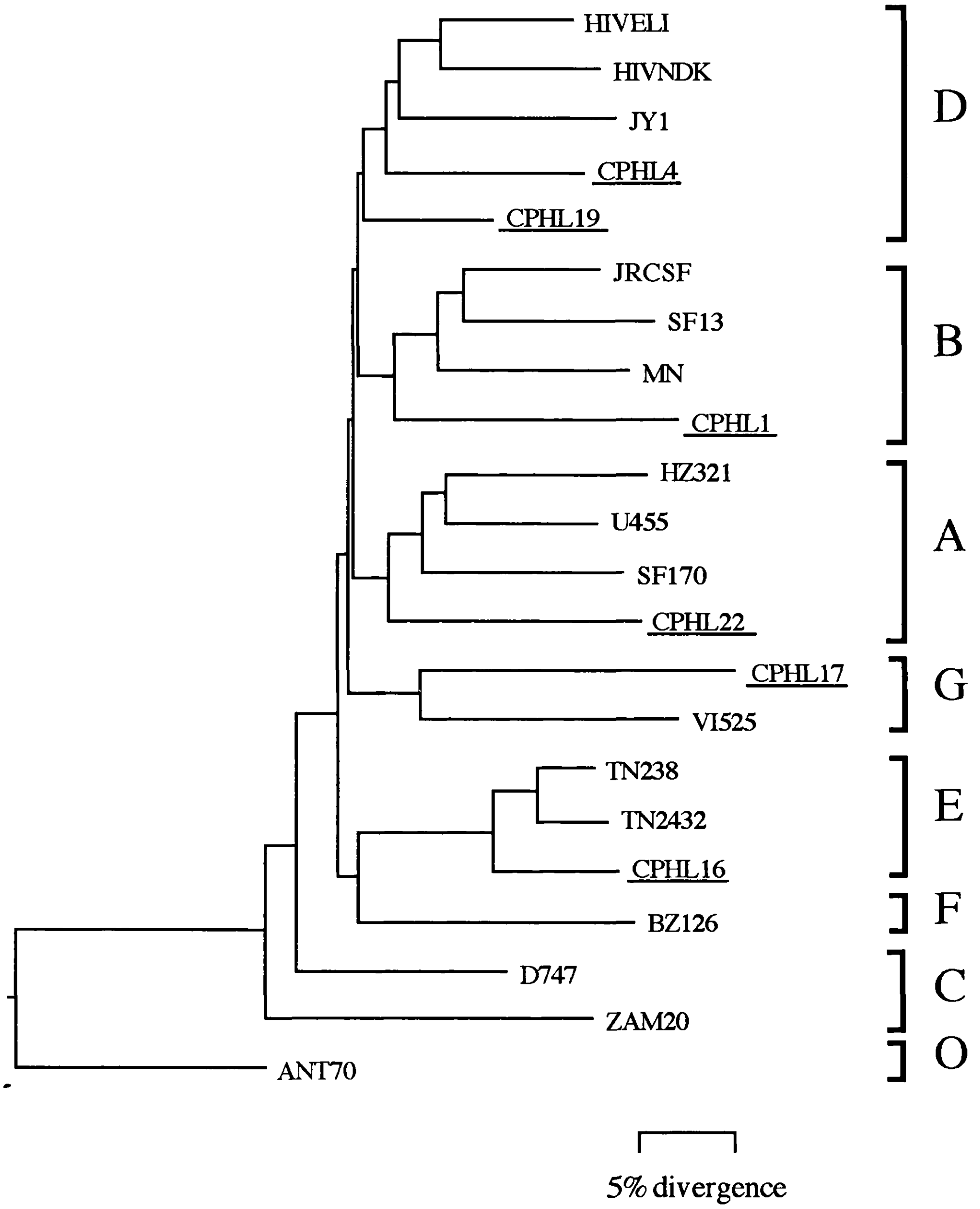

TN238
66.0
TN2432
100.0
CPHL16
53.0
BZ126
70.0
U455
36.0
SF170
84.0
HZ321
40.0
CPHL22
MN
41.0
JRCSF
42.0
SF13
100.0
CPHL1
80.0
ELI
82.0
48.0
NDK
87.0
CPHL4
100.0
41.0
45.0
JY1
58.0
CPHL19
CPHL17
100.0
VI525
ZAM20
75.0
D747
ANT70
E
F
A
B
D
G
C
O

Figure 3.17

Phylogenetic tree of translated amino acid sequences depicted in figure 3.16, constructed using the method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

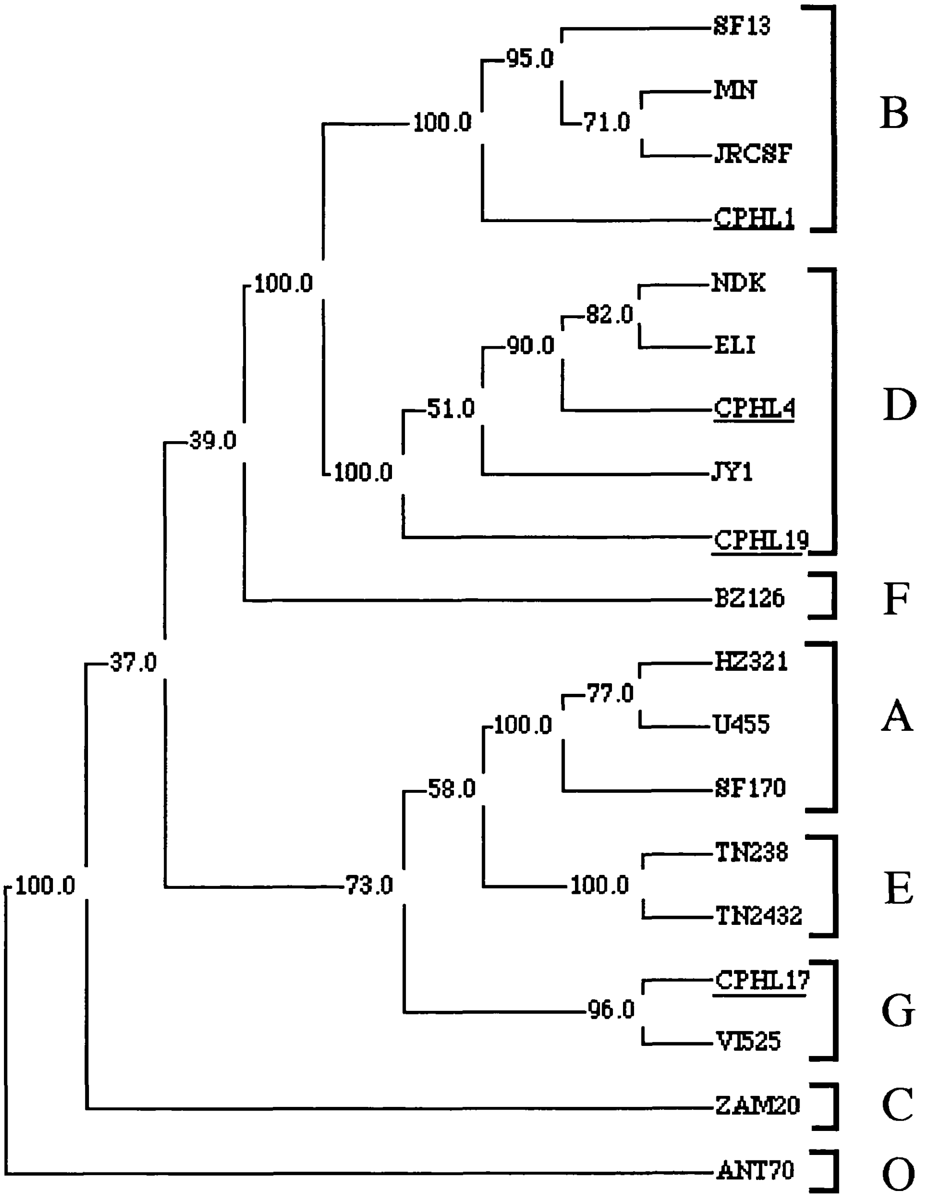
Use of full gp120 for subtyping

Phylogenetic trees constructed with sequences of 1400-1500 nucleotides (see figure 3.9)

Three complete gp120 nucleotide trees were analysed (figures 3.18, 3.20 and 3.22). Bootstrapping (see section 2.4.7) was used to produce multiple data sets (100 for each tree) which were analysed to produce a consensus nucleotide tree. Two trees were constructed using the bootstrapping method, one containing one CPHL subtype G sequence CPHL17 (figure 3.18) and one also containing the subtype G sequences from specimens CPHL10 and 11 (figure 3.20). The third nucleotide tree (figure 3.22) was constructed using the neighbour-joining method described in section 2.4.4, containing both CPHL17 and the CPHL10 and 11 sequences for a comparison of techniques. In the phylogenetic tree figures the numbers at the nodes indicate the number of times the group consisting of the species which are to the right of that node occurred out of 100 trees. Figure 3.19 shows the translated protein tree for nucleotide sequences depicted in figure 3.18 and figure 3.21 shows the translated protein tree for nucleotide sequences depicted in figure 3.20.

Figure 3.18

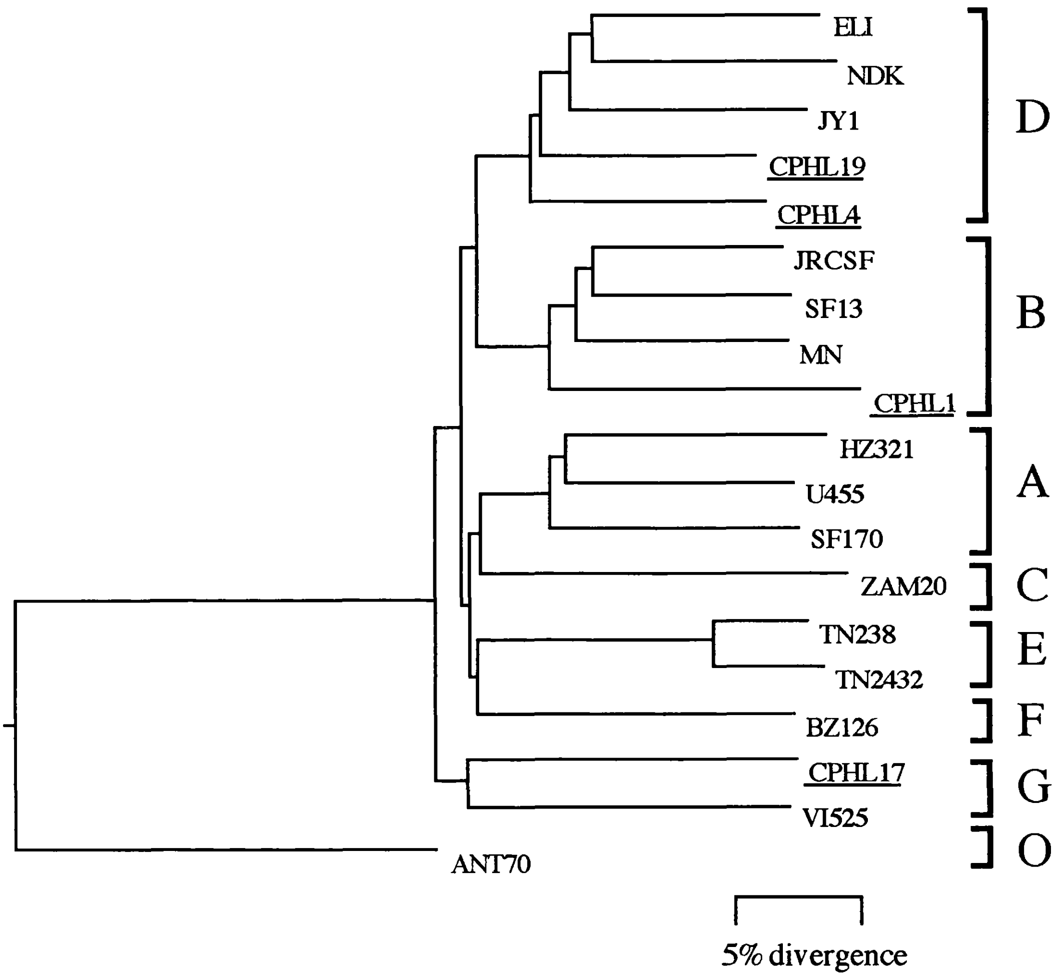
Phylogenetic tree of complete gp120 (approximately 1,400-1,500 nucleotides), translating to approximately 500 amino acids, comprising C1, V1, V2, C2, V3, C3, V4, C4, V5, and C5. When translated to protein, this sequence comprises 69% amino acids from constant regions and 31% amino acids from variable regions. Other details as for figure 3.10.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Figure 3.19

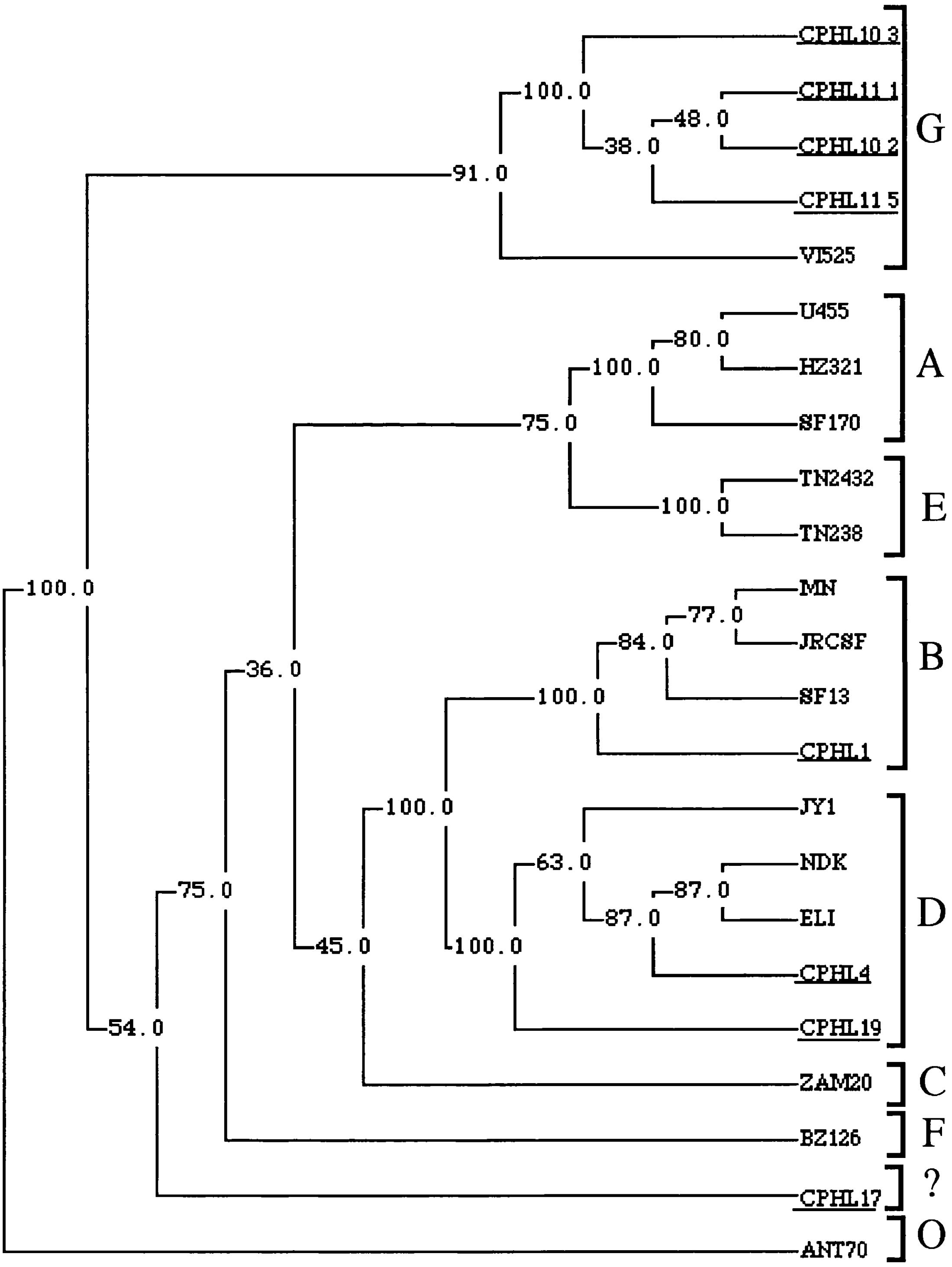
Phylogenetic tree of translated amino acid sequences depicted in figure 3.18 constructed using the method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Figure 3.20

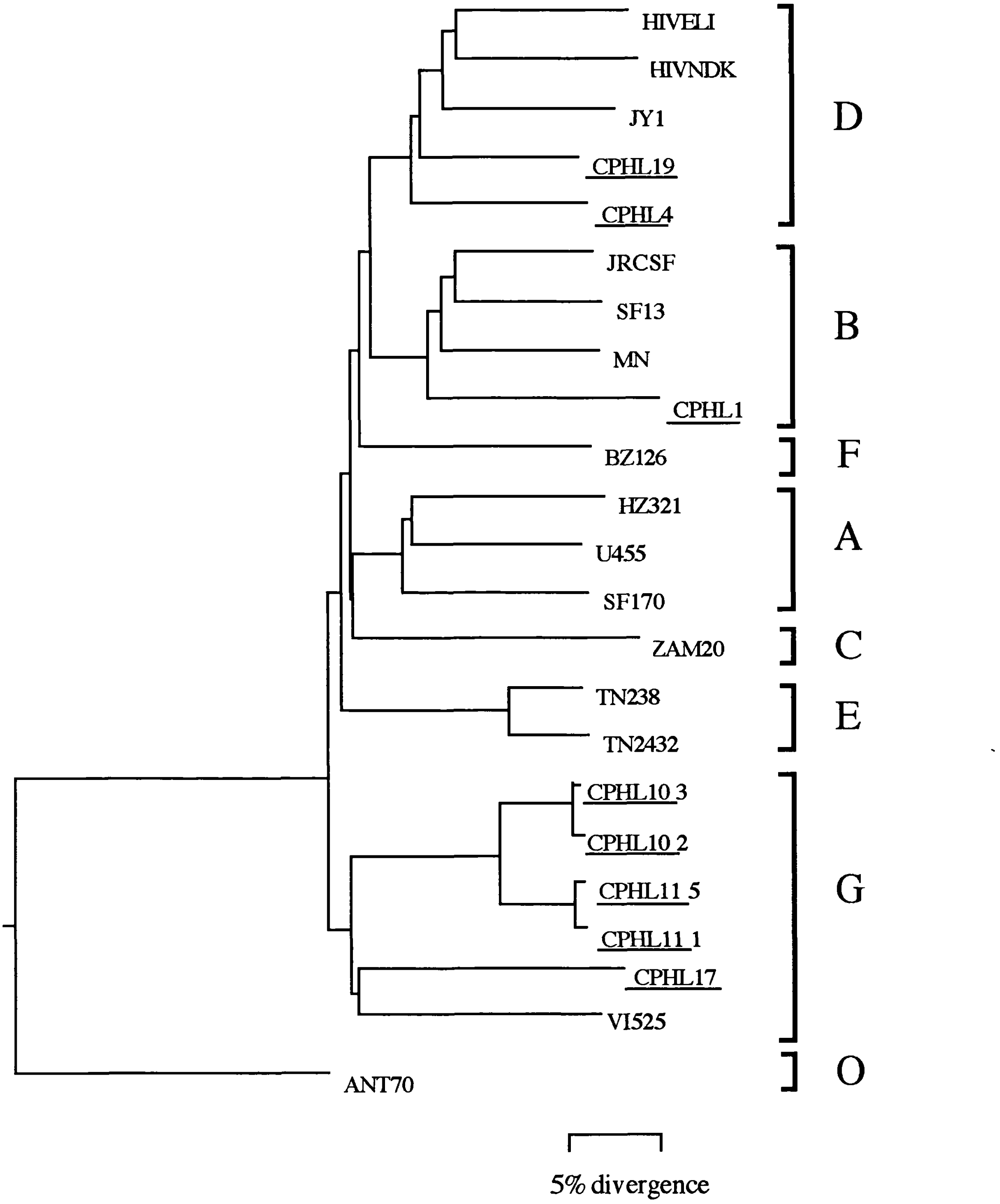
Phylogenetic tree of complete gp120. Other details as for figure 3.18. This tree includes subtype G mother/baby pair (CPHL10 and 11 respectively).



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Figure 3.21

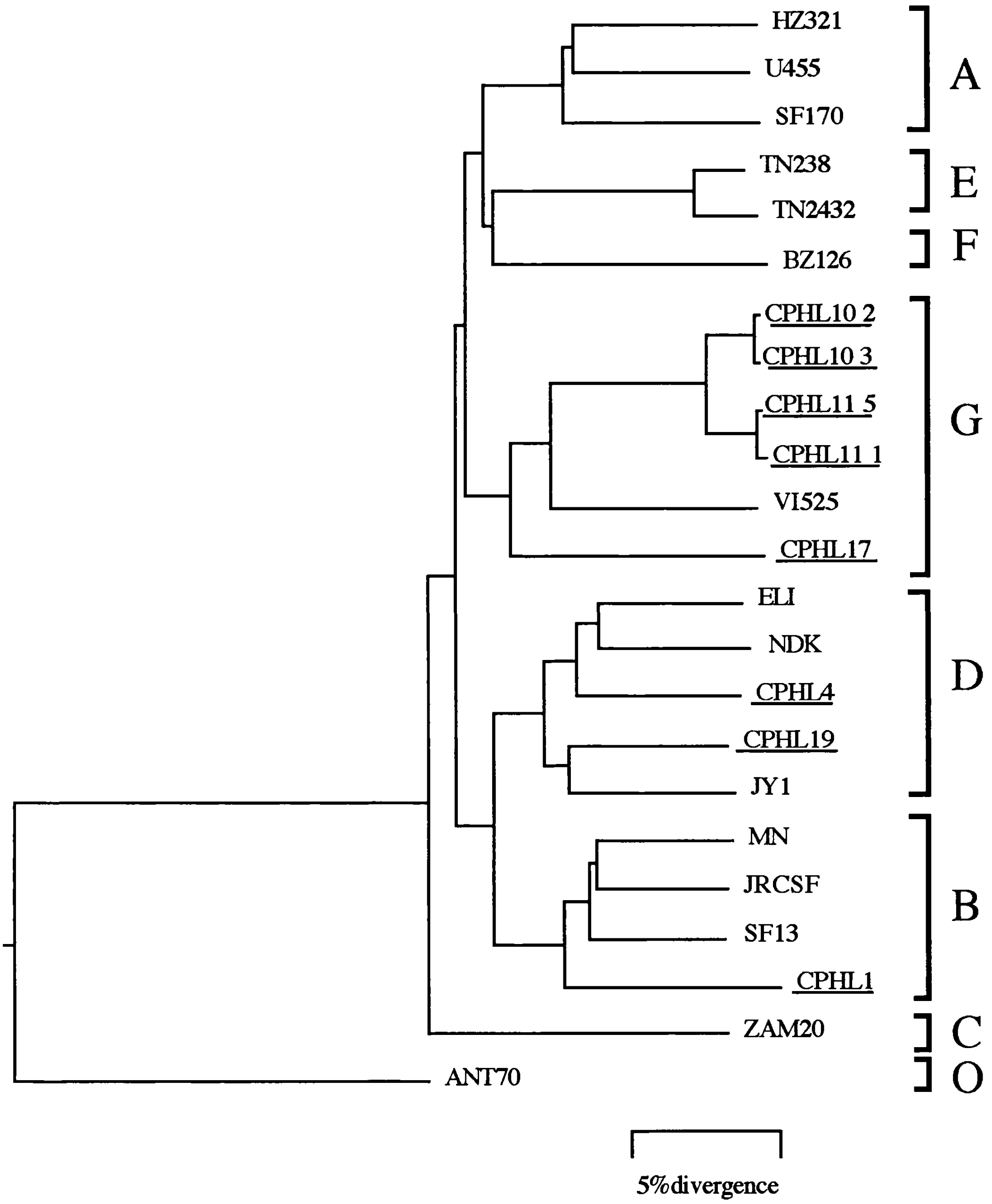
Phylogenetic tree of translated amino acid sequences depicted in figure 3.20 constructed using the method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Figure 3.22

Phylogenetic tree of complete gp120. This tree includes subtype G mother/baby pair (CPHL10 and 11 respectively) and was constructed using a neighbour joining method described in section 2.4.4. An indication of the degree of sequence dissimilarity is shown on the horizontal axis.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.3 for sequence data.

Discussion

i) subtypes deduced from all maximal available gp120 sequences

The number of European and North American heterosexually transmitted HIV infections is increasing (206). Injecting drug users and bisexual men may be a source of infection for heterosexual transmission and the HIV-1 epidemic may be spreading from them into a wider population. To date, HIV-1 infection in the U.K. has been mainly of subtype B, transmitted homosexually and by injecting drug use, whereas the African epidemic has equally affected both men and women who are likely to be infected with a non-B subtype virus. If the distribution of subtypes in particular risk groups is due to an initial chance founder effect this could also result in spread of virus into low risk groups in Europe and the United States. For example, if men infected with a non-B subtype by heterosexual intercourse or intravenous drug use outside Europe or the US returned to infect their female partners. Alternatively, it has been suggested that subtypes other than B are heterosexually transmitted ten times more efficiently than subtype B and this may account for the lack of large numbers of heterosexually acquired subtype B infections (245). Analyses of isolates from Thailand have suggested that subtype A/E may be transmitted more readily by sexual contact (134, 146), and subtype B by intravenous drug use (200). It might therefore be that the introduction of non-B subtype viruses into Europe and the United States will contribute to the dissemination of HIV into the heterosexual population. This could occur if there were any tropism of a subtype for vaginal transmission (245).

Soto-Ramirez and colleagues found striking differences in Langerhans cell tropism between different virus subtypes. Langerhans' cells express CD4 on their surfaces, are located in the oral and genital mucosa and are particularly abundant in the cervix, but are absent from the rectal mucosa (152). Subtype E viruses showed continuous increase in viral replication in Langerhans' cells which, they argued, could help to explain differences in the frequency of heterosexual transmission of different subtypes of HIV-1. Their study suggested that if subtype B viruses were transmitted heterosexually and replicated locally in genital Langerhans' cells, lower levels of replication might decrease the risk of virus establishing infection. However, differences in the rates of HIV-1 heterosexual transmission may also be attributable to sexual behaviour practices and host genetic susceptibility. Other cofactors may also increase the efficiency of heterosexual transmission of HIV, including the presence of other sexually transmitted diseases (193, 210). This conjecture is supported by the observation that in Thailand subtype E virus is associated with heterosexual transmission, and B with intravenous drug use (134). Soto-Ramirez and colleagues' summation also fits with the results obtained by Kunanusont *etal.* in Thailand, where HIV-1 heterosexual transmission in couples where one partner is HIV positive and the other is negative was less efficient when the index case was infected with a subtype B virus, compared to that in couples where subtype E viruses were involved (146). However, as described in this thesis, subtype B viruses in the U.K.may be recovered both from individuals who acquired the virus through heterosexual as well as homosexual intercourse. (see Chapter 4).

The heterosexually transmitted epidemic in the U.K includes strains of HIV-1 subtypes other than B, e.g. the African subtypes (245). The *envG/gag*A virus described here came from a patient infected in the Cameroon (CPHL17, see figure 3.1). One subtype A virus (CPHL22, see table 3.1) was possibly acquired heterosexually in Somalia. Two subtype C viruses came from infants whose mothers had resided in Africa. Specimens CPHL21 (subtype C, see table 3.1) and CPHL4 and 5 (subtype D, see table 3.1) viruses were most likely transmissions which occurred in England, the index case having acquired the infection originally in Africa, and are thus so-called second generation transmissions. For example, the CPHL4 sequences were from a female infected in England by her partner who probably acquired the virus in Africa. The adults infected in Africa most likely all acquired the virus heterosexually. The adult infected with the *gag* A, *env* E virus possibly acquired the virus in Thailand, but it is not known whether he was infected by injection or intercourse. As mentioned above, in Thailand, this type of virus has been associated with sexual transmission (134).

ii) use of subsets of gp120 for subtyping

Table 3.2 (generated from data taken from figures 3.10-20) shows that smaller regions of gp120, especially those with a low percentage (below 60%) of nucleotides in the constant regions do not give accurate subtyping results when analysed phylogenetically. However, the 5' 400 region, with approximately 60% of nucleotides in constant regions of gp120 gives the percentage occurrence of known subtype branches (bootstrap values) as low as 40%. This may indicate that some variable regions are more variable than others, i.e. in the 400 nucleotide sequences at the 5' end of gp120 or that the smaller regions analysed in some instances are too small for reliable conclusions to be made.

Table 3.2 shows specimen CPHL17 as subtype A for the 200, 450 and 600 bp fragments (figures 3.10, 3.12 and 3.14 respectively). For the 600 bp fragment, known subtype A strains occur on one branch 98% of the time, as shown in the consensus tree. However, with the 5' 400 (figure 3.16) and 1200 bp fragments (figure 3.18) it falls into subtype G, along with one known subtype G strain on one branch 96%-100% of the time. Moreover, when CPHL10 and 11 are included in the gp120 tree (figure 3.20), 54% of the time CPHL17 does not fall into any particular subtype. The neighbour-joining tree which includes CPHL17 and CPHL10 and 11 (figure 3.22), places CPHL17 in subtype G, but on a relatively short branch. This indicates that this genome is a possible A/G recombinant. Alternatively, it may reflect the extremely heterogeneous nature of subtype A (186). There are few subtype G gp120 sequences available in the database and the inclusion of several reference subtypes, rather than one, is likely to have an effect on the topology of the tree. It is has been suggested that subtype E is in the process of evolving into a separate subtype, with a slower rate of evolution in *gag*, which is still subtype A (230). It is possible that subtype G is more divergent with respect to evolution from subtype A. CPHL18 is also a possible A/D recombinant, although these results are not statistically significant.

b) p6/protease

Background

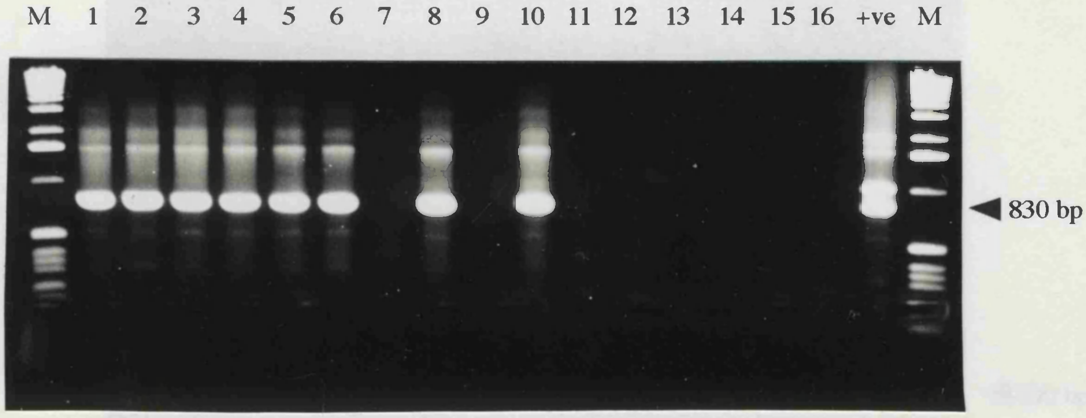
The *gag/pol* junction of HIV-1 , encoding the p6 and protease (p6/protease) with its lack of sequence variation in the protease gene (77, 284), and the observed length polymorphism in p6 depicted in figure 3.25 and 3.27, suggest that it may be a useful region from which to derive phylogenetic relationships for sequenced based subtyping and/or transmission studies. The arguments for not using partial regions of *env* (hypervariability, evolutionary convergence and the possibility of host immune response driven evolution), and for not using partial regions of *gag* (insufficiently variable), may, in part, be overcome by using the p6/protease region. To determine whether this region was useful for deriving phylogenetic relationships (subtyping and/or transmission studies), single molecules of the p6/protease region were amplified and sequenced from selected clinical specimens (see Table 3.1) using the methods described in 2.3.4, 5 and 2.3.7-10. Sequences were analysed using the methods described in section 2.4.

Results

Sequences were determined for the p6/protease region of DNA amplified from lymphocytes prepared from specimens submitted to this laboratory for transmission studies, confirmation of HIV infection or subtyping of the virus (see Table 3.1), using the limit dilution method described in sections 2.3.4, 2.3.5, 2.3.7-10. A typical agarose gel, depicting an initial limit dilution of p6/protease from sample CPHL15 (see table 3.1) is shown in figure 3.23.

Figure 3.23

A typical agarose gel showing an initial limit dilution of p6/protease (2.3.8)



For this particular lysate, 46 tubes of a 1 in 25 dilution were made and PCR carried out according to sections 2.3.5, 2.3.7-10. The PCR amplicons were run on an agarose gel, shown in figure 3.24.

Figure 3.24

A 1% agarose gel showing a 'bulk' p6/protease dilution, as described in section 2.3.8.

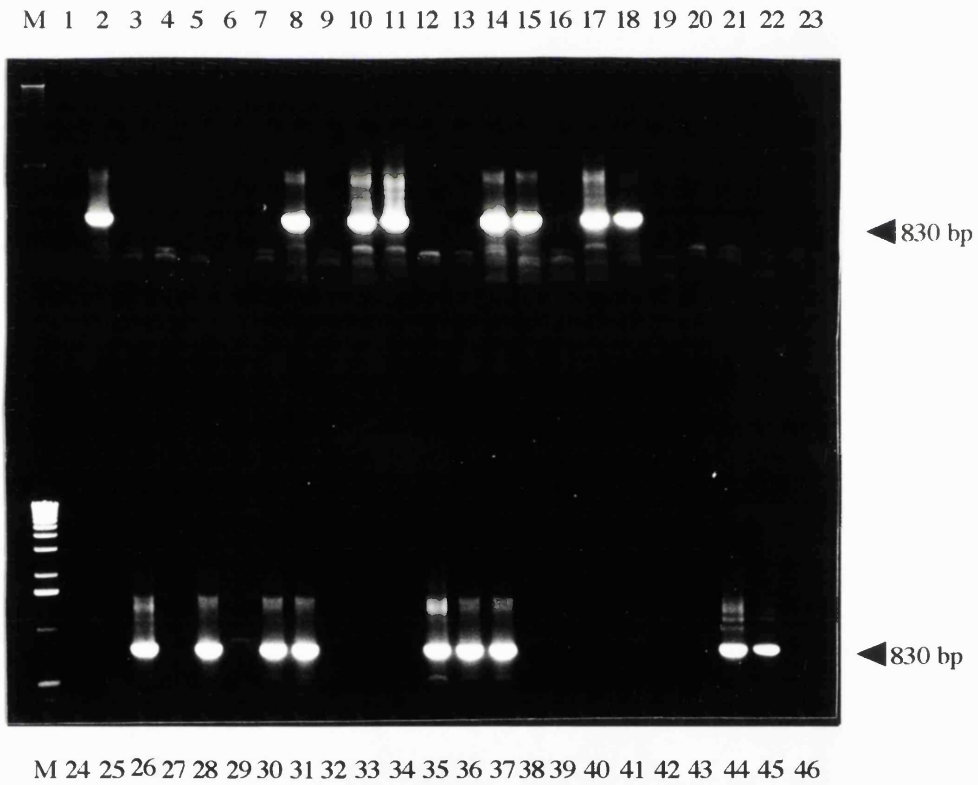

M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
830 bp
830 bp
M 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46

Figure 3.24 shows 17 out of 46 (0.37) PCR positive tubes with a 1 in 25 dilution of lysate. Table 2.3 (section 2.3.8) indicated that the observed frequency of 0.37 positives implies a proportion of approximately 80% of tubes containing single molecules.

i) Subtypes deduced from p6/protease sequences

A phylogenetic tree was constructed using the MegAlign neighbour-joining method described in section 2.4.4, using all the available p6/protease sequence data generated for this thesis together with representative sequences from the Los Alamos Database (Figure 3.26). Sequences from subtypes C, F and G are not available for this region and subtype E does not exist for *gag*. The full alignment of these sequences is shown in figure 3.25. The alignments of the translated protein sequences for p6 and protease are shown in figures 3.27 and 3.28 respectively.

Figure 3.25

Alignment of CPHL p6/protease sequences

```
TTCCTTCAGAGCAGACCAGAGCCAACAGCCC---------CACCAGAAGAGAGCTTCAGG Consensus

...............................---------.................... CPHL1, 10
...............................---------.................... CPHL1, 11
...............................---------.................... CPHL1, 14
...............................---------.................... CPHL1, 2
...............................---------.................... CPHL1, 8
...............................---------.................... CPHL2, 12
...............................---------.................... CPHL2, 2
..............G................CAACAGCCC.................... CPHL3, 20
..............G.....A..........---------.................... CPHL3, 23
..............G.....A..........CAACAGCCC.................... CPHL3, 28
..T............................---------......C..........G.. CPHL4, 1
..T............................---------......C..........G.. CPHL4, 42
..T............................---------......C..........G.. CPHL4, 45
..T............................---------.....AT..........G.. CPHL5, 1
..T............................---------.....AT..........G.. CPHL5, 34
..T............................---------.....AT..........G.. CPHL5, 35
..T............................---------................A.T. CPHL6
..T.......A...G..............A.---------......C............. CPHL15
..T.C..........................---------......C...A.A..GGG.. CPHL16, 1
..T..C........G................---------......C.......C..G.. CPHL17
...............................---------.G....C..........G.. CPHL24, 3

TTTGGGGAAGAGATAACAACTCCCTCTCAGAAGCAGGAGC------------CGA-AGAC Consensus

........................................------------...GG... CPHL1, 10
.............C..........................------------...GG... CPHL1, 11
.............C...C......................------------...GG... CPHL1, 14
.............C..........................------------...GG... CPHL1, 2
.............C...C......................------------...GG... CPHL1, 8
.............C..........................------------...T.... CPHL2, 12
.............C..........................------------A..T.... CPHL2, 2
.............C......C...........A.......------------...T.... CPHL3, 20
..........G..C..................A.......------------...T.... CPHL3, 23
..........G..C.........................A------------...T.... CPHL3, 28
.................---C.............C....AGAA------------.... CPHL4, 1
........G........---C.............C....AGAA------------.... CPHL4, 42
.................---C.............C....AGAA------------.... CPHL4, 45
.................---C...................CGAA------------.... CPHL5, 1
.................---C...................CGAA------------.... CPHL5, 34
.................---C...................CGAA------------.... CPHL5, 35
...............C........................------------...T.... CPHL6
..C.A...G------.....C...G...C...........CGAA------------.... CPHL15
A.G.............A---CT...TA.T...........AGAA-----AGA.A.GGAC. CPHL16, 1
..C.....G......G.---C.....C.C...A.......AGAA---------..AG... CPHL17
........G........---A...C.C.....A.......AGAA------------G... CPHL24, 3
```

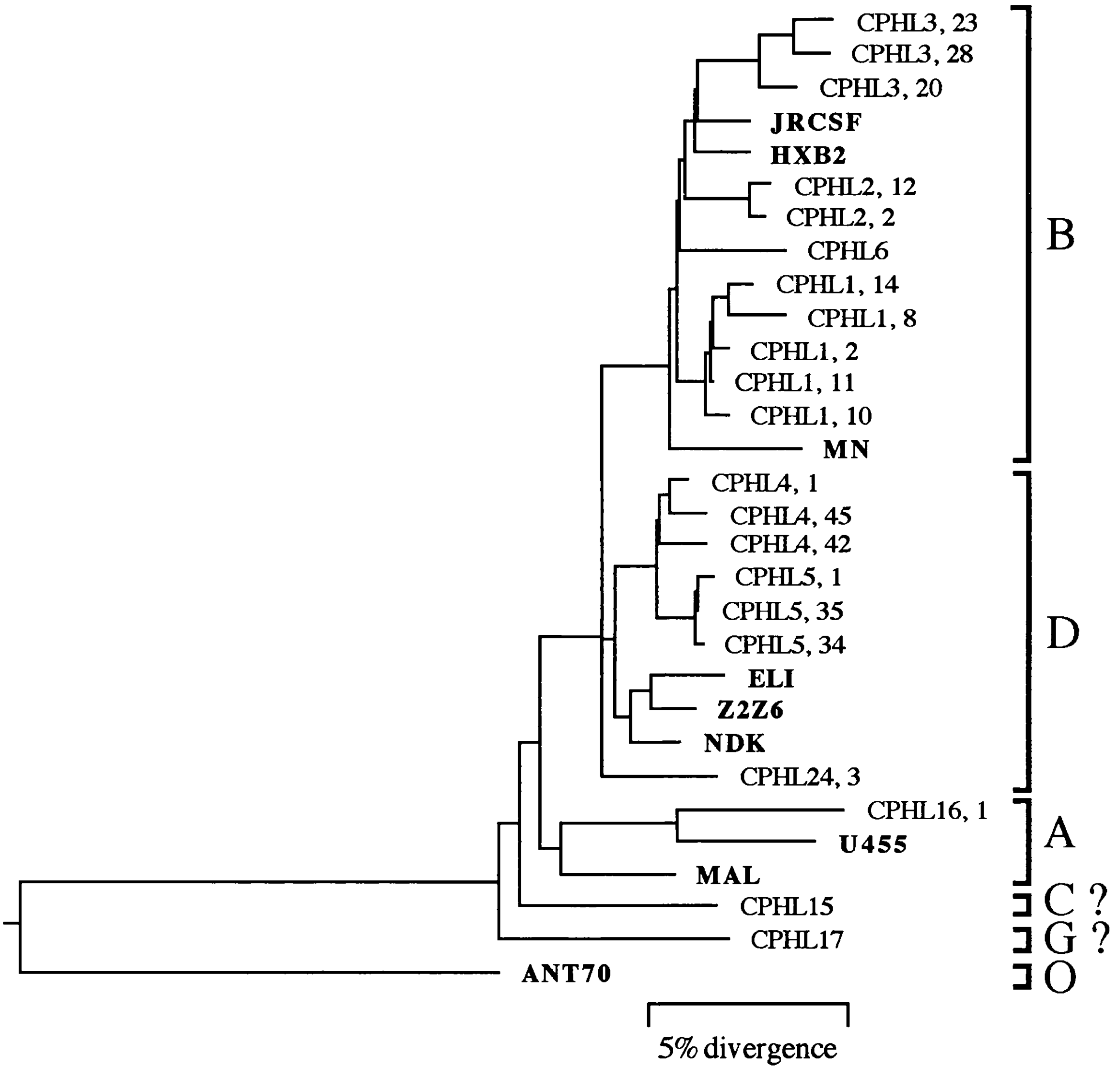
```
AAGGAACTGTATCCTTTAGCTTCCCTCAAATCACTCTTTGGCAACGACCCCTCGTCACAA Consensus

............................................................ CPHL1, 10
............................................................ CPHL1, 11
............................................................ CPHL1, 14
............................................................ CPHL1, 2
...A........................................................ CPHL1, 8
.................................................A.......... CPHL2, 12
............................G............................... CPHL2, 2
......ACA...................G..............................G CPHL3, 20
......AC....................G..............................G CPHL3, 23
......AC....................G..............................G CPHL3, 28
..............C.....................................T....... CPHL4, 1
..............C.....................................T......G CPHL4, 42
..............C.....................................T....... CPHL4, 45
..........C...C.....................................T......G CPHL5, 1
.....------...C.....................................T......G CPHL5, 34
.....------...C.....................................T......G CPHL5, 35
............................................................ CPHL6
.G....------..C...A........................G........T...T... CPHL15
.T---C..-CC........T................................T...C..G CPHL16, 1
---.....A......C..A.................................T..T..C. CPHL17
.....G........C.....................................T....... CPHL24, 3

TAAAAATAGGGGGGCAACTAAAGGAAGCTCTATTAGATACAGGAGCAGATGATACA-GTA Consensus

....G...................................................-... CPHL1, 10
....G...................................................-... CPHL1, 11
....G...................................................-... CPHL1, 14
....G........A..........................................-... CPHL1, 2
....G...................................................-... CPHL1, 8
....GG...A..............................................-... CPHL2, 12
....GG...A..............................................-... CPHL2, 2
..........A...............*.............................-... CPHL3, 20
..........A........G......T.............G............T..-... CPHL3, 23
..........A..........................................T..-... CPHL3, 28
.............A..G.......................................-... CPHL4, 1
.............A..G.......................................-... CPHL4, 42
.............A..G.......................................-... CPHL4, 45
.............A..G.......................................-... CPHL5, 1
.............A..G.......................................-... CPHL5, 34
.............A..G.......................................-... CPHL5, 35
....GG..................................................A... CPHL6
.....G...A...C..GA.............C.....C..................-... CPHL15
..........A..A..G..G..A.................................-... CPHL16, 1
.....G...CA.....G..G....................................-... CPHL17
....G........A..GA......................................-... CPHL24, 3
```

```
TTAGAAGAAATGAATTTGCCAGGAAGATGGAAACCAAAAATGATAGGGGG-AATTGGAGG Consensus

.............G....................................-......... CPHL1, 10
..................................................-......... CPHL1, 11
.......................G..........................-......... CPHL1, 14
..................................................-......... CPHL1, 2
...A...................G....A....................A-......... CPHL1, 8
........C...G.....................................-......... CPHL2, 12
........C...G.....................................-......... CPHL2, 2
..................................................-......... CPHL3, 20
.............G....................................-......... CPHL3, 23
.............C..............A.....................-....A.... CPHL3, 28
........C...G............A........................-......... CPHL4, 1
........T....G...........A......G.................-......... CPHL4, 42
........C..........A.....A........................-......... CPHL4, 45
........T................A........................-......... CPHL5, 1
........T................A........................-......... CPHL5, 34
........T................A........................-......... CPHL5, 35
........T...TG....................................G......... CPHL6
...........AC............A.....................A..-......... CPHL15
........T..A.............A........................-......... CPHL16, 1
...........AG............A.....G..................-......... CPHL17
.........................A........................-......... CPHL24, 3

TTTTATCAAAGTAAGACAGTATGATCAGATACTCATAGAAATCTGTGGACATAAAGCTAT Consensus

...................................................C........ CPHL1, 10
...................................................C........ CPHL1, 11
.................................AG.......T........C........ CPHL1, 14
..........................................T........C........ CPHL1, 2
.....................................A.............C........ CPHL1, 8
................T............................C.............. CPHL2, 12
.............................................C.............. CPHL2, 2
..................A......................................... CPHL3, 20
..................A......................................... CPHL3, 23
..................A......................................... CPHL3, 28
...........................A....C.....................GA.... CPHL4, 1
...........................A....C.........A............A.... CPHL4, 42
........................C..AG...C......................A.... CPHL4, 45
........................C..A....C................T.....A.... CPHL5, 1
........................C..A....C......................A.... CPHL5, 34
........................C..A....C......................A.... CPHL5, 35
........................G...G...AA.......................... CPHL6
..................A........A.....T........T......A.G..G..... CPHL15
.........G.....G..A..............T........T......A.A..G..... CPHL16, 1
........................GG.A...G..........TGAG..GA.A........ CPHL17
C..........................A.....AG..............T.......... CPHL24, 3
```

```
AGGTACAGTATTAGTAGGACCTACACCTGTCAACATAATTGGAAGAAATCTGTTGACTCA Consensus

.............................................G.............. CPHL1, 10
............................................................ CPHL1, 11
............................................................ CPHL1, 14
............................................................ CPHL1, 2
.................................................T.......... CPHL1, 8
.............A.............................................. CPHL2, 12
.............A.............................................. CPHL2, 2
............................................A............... CPHL3, 20
............................................................ CPHL3, 23
............................................................ CPHL3, 28
.................................................T.......... CPHL4, 1
.................................................T.......... CPHL4, 42
.................................T...............T.......... CPHL4, 45
.................................................T.......... CPHL5, 1
.............................................G...T.......... CPHL5, 34
.................................................T.......... CPHL5, 35
............................................................ CPHL6
..................G..............................A.......... CPHL15
.........G..............................A..C.....A.......... CPHL16, 1
.........G...A...................T..............CA.......... CPHL17
..............................A..................T.......... CPHL24, 3

GATTGGTTGCACTTTAAATTTT                                      Consensus

......................                                      CPHL1, 10
......................                                      CPHL1, 11
......................                                      CPHL1, 14
......................                                      CPHL1, 2
......................                                      CPHL1, 8
......C..............C                                      CPHL2, 12
......C..............C                                      CPHL2, 2
.............C........                                      CPHL3, 20
......................                                      CPHL3, 23
......................                                      CPHL3, 28
......C...............                                      CPHL4, 1
......C...............                                      CPHL4, 42
......C...............                                      CPHL4, 45
......C...............                                      CPHL5, 1
......C...............                                      CPHL5, 34
......C...............                                      CPHL5, 35
.............C........                                      CPHL6
.C....A.....AC........                                      CPHL15
...C.....T...........C                                      CPHL16, 1
......C..T............                                      CPHL17
......................                                      CPHL24, 3
```

Figure 3.26

Phylogenetic tree constructed using the neighbour-joining method described in section 2.4.4, using all the available p6/protease nucleotide sequence data generated for this thesis together with representative sequences from the Los Alamos Database (shown in bold). Sequences from subtypes C, F and G are not available for this region and subtype E does not exist for *gag*.



Gap stripping was not carried out for this analysis. For full details of the sequences used for this analysis see table 3.1 and figure 3.25 for sequence data.

Figure 3.27

Alignment of p6 proteins

```
FLQSRPEPTAP---PEESFRFGEETTTPSQKQE----P-DKELYPLASLKSLFGNDPSSQ   Consensus
...........---..........I........----.R.....................   CPHL1, 10
...........---...................----.R.....................   CPHL1, 11
...........---...................----.R.....................   CPHL1, 14
...........---...................----.R.....................   CPHL1, 2
...........---...................----.R..K..................   CPHL1, 8
...........---...................----.I.................H...   CPHL2, 12
...........---...................----QI..........R..........   CPHL2, 2
...........---...................----.I..........R......H...   CPHL2, 6
...........TAP...................----.I...T......R..........   CPHL3, 20
...........---.........G.........----.I...T......R..........   CPHL3, 23
...........TAP.........G.........----TI...T......R..........   CPHL3, 28
...N.......L----..LM....MP.......----.I.....................   CPHL6
...........---.A...G....I.-....P.QK----..................L..   CPHL4, 1
...........---.A...G....I.-....P.QK----..................L..   CPHL4, 42
...........---.A...G....I.-....P.QK----..................L..   CPHL4, 45
...........---.I...G....I.-......PK----....S.............L..   CPHL5, 1
...........---.I...G....I.-......PK----..D--.............L..   CPHL5, 34
...........---.I...G....I.-......PK----..D--.............L..   CPHL5, 35
...N.......---.A.....E.T.PA.---..----.K.R.--..T.......S..L..   CPHL15
.P.........---.A.NWGM...INSLL-...QK---DKDHPP..V..........L..   CPHL16
...........---.A..LG....IA-..P...QK----KD.....T..........LLP   CPHL17
...........---.A...G....I.-.P....QK----..................L..   CPHL24, 3
```

Figure 3.28

Alignment of protease proteins

```
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGKWKPKMIGGIGGFIKVRQYD   Consensus

....................................S...R...................   CPHL1, 10
........................................R...................   CPHL1, 11
........................................R...................   CPHL1, 14
........................................R...................   CPHL1, 2
..............VE..................D.D...R...................   CPHL2, 2
............V........S........S.....S...R...................   CPHL3, 23
..............V.....................C...R..................E   CPHL6
..................................D.D.......................   CPHL4, 1
............V.....................D.S.......................   CPHL4, 42
..................................D...Q.....................   CPHL4, 45
............V.....................D.........................   CPHL5, 1
............V.....................D.........................   CPHL5, 34
............V.....................D.........................   CPHL5, 35
...........S..VE..I................IH.......................   CPHL15
...........PV.....................DI........................   CPHL16 1
..............VA...................ID.....R................E   CPHL17
..................I.........................................   CPHL24, 3
```

```
QILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF     Consensus

.......................................     CPHL1, 10
.......................................     CPHL1, 11
...V...................................     CPHL1, 14
.......................................     CPHL1, 2
................I......................     CPHL2, 2
.......................................     CPHL3, 23
.VQ....................................     CPHL6
..P.......T............................     CPHL4, 1
..P.......T............................     CPHL4, 42
.VP.......T............................     CPHL4, 45
..P.....Y.T............................     CPHL5, 1
..P.......T............................     CPHL5, 34
..P.......T............................     CPHL5, 35
........K...................M...L.....S     CPHL15
........K................R..M..........     CPHL16 1
E.V...E.K.......I...........M.........S     CPHL17
...V....Y.............................S     CPHL24, 3
```

Discussion

Figure 3.26 indicates that it may be possible to subtype accurately with this region of HIV-1, as branch lengths separating the known different subtypes are relatively long, with bootstrap values at nodes separating known subtypes at 97-100%. However there are no database sequences available from subtypes C, F and G. The question marks on the figure indicate that specimens CPHL15 and 17 (see Table 3.1) are most likely to be subtypes C and G respectively for this region, as they fall into these subtypes after gp120 phylogenetic analysis. Specimen CPHL16 is subtype E in gp120, but A for this region, which is consistent with other strains also acquired in Thailand. A phylogenetic tree of amino acid sequences was not done for this region because, as it spans the *gag/pol* junction, a frame shift would have to be incorporated to maintain the ORF. The separate alignments of the derived p6 and protease amino acids, shown in figures 3.27 and 3.28 respectively, are too small to derive any meaningful phylogenetic information. Phylogenetic trees derived for both regions placed known linked samples from transmission investigations on separate branches. Open reading frames were obtained for p6 proteins from all samples, but four reading frames were found in which the protease gene was truncated. CPHL1, 8 and CPHL3, 28 protease sequences were both prematurely truncated by the same G>A mutation (at positions 247 and 256 respectively) relative to the start of p6, resulting in mutation from a tryptophan residue to a termination codon at amino acid 42 (TGG>TAG, in italics underlined on figure 3.26). CPHL2, 12 was prematurely truncated by a C>T mutation at position 294 relative to the start of p6, resulting in mutation from a glutamine residue to a termination codon at amino acid 58 (CAG>TAG, in italics underlined on figure 3.26).CPHL3, 20 sustained a G deletion at position 195 (asterisked on figure 3.26), causing a frameshift resulting in premature termination of the protein. The premature termination of these protease proteins would almost certainly have a severely detrimental effect on the activity of the functional enzyme, resulting in the loss of production of viral structural proteins and enzymes. It is interesting to note that the same G>A mutation, resulting in mutation from tryptophan to a termination codon, occurred in two separate specimens. As well as being a potential region for phylogenetic analysis, the protease protein itself is of course the target for antiretroviral therapy These sequences represent a cross section of protease genes from HIV-1 infections in the U.K. Monitoring of strain to strain variation within a locale is important for effective antiretroviral therapy and to detect naturally resistant strains in the antiviral drug-naive population (due to extensive variation within a subtype or the introduction of new subtypes), and to detect resistance as it arises within an individual on antiretroviral therapy.

In conclusion, this study has shown that diverse subtypes of HIV-1 are present in England in individuals not infected as a result of homosexual exposure (10, 11, 41). According to CDSC statistics of all reports of HIV-1 infected persons by exposure category in the U.K. to the end of December 1995 (206), 4644 out of 25635 (approximately 1 out of 5) HIV-1 infected persons acquired their infections heterosexually. Of these, 3443 (approximately 3 out of 4) were thought to be through exposure to partners outside the U.K. Assuming the majority of infections acquired outside the U.K.were non B subtypes, as few as a quarter of heterosexually acquired HIV infections in the U.K. may be subtype B. There is, therefore, potential for diverse subtypes of HIV to spread into groups who may not consider themselves to be at risk of HIV infection. Determining subtype, based on both *gag* and extensive *env* sequences (virtually whole gp120) amplified from lymphocytes, provides a way of monitoring in finer detail the spread of HIV-1 and better identifying any new sources of infection.

CHAPTER 4

THE MEASUREMENT OF gp120 SEQUENCE DIVERSITY FOR THE INVESTIGATION OF THE TRANSMISSION OF HIV-1

Background

The risk factors for the transmission of HIV are well established. Almost all HIV infections occur through four routes: i) sexual intercourse; ii) sharing contaminated needles; iii) treatment with infected blood or blood products; and iv) perinatal transmission from mother to infant (8). As described in chapter 1, most HIV transmission investigations have used comparison of sequence data generated from proviral DNA from apparently epidemiologically related cases to determine whether transmission had taken place (4, 9, 14, 119, 129, 199). In these investigations regions of *env*, *gag* and *pol* have been used to determine sequence relatedness between HIV-1 variants for establishing whether infections are linked (4, 119, 129, 199). The V3 loop of *env* has been extensively studied and has consequently been the region of choice for several investigations(185, 199, 279, 285, 287). However, as described in chapter 1, the V3 loop is under considerable immune pressure and evolves rapidly. This may lead to convergent evolution of otherwise unrelated sequences which may in turn confuse the determination of epidemiological relationships (115, 251). If, on the other hand, a region of *gag* or *pol* is used instead of V3 for determining sequence relatedness, it is possible that it may not show sufficient variation to be phylogenetically informative. Also, if a direct consensus or 'bulk' sequence from PCR is obtained and not single molecules, it is possible that minor variants could be missed.

In the light of the above considerations and in order to avoid the controversies surrounding the Florida dentist investigation, on average four full-length gp120 single molecules, including both variable and conserved domains, were sequenced for each individual involved in the transmission investigations subsequently described. The following studies include several modes of possible transmissions, including: i) heterosexual; ii) surgeon to patient; iii) occupationally acquired infection; and iv) perinatal.

As well as establishing whether or not transmission had taken place between the infected individuals, the aim was to establish the optimum region of the HIV-1 genome for comparison of sequence data for the investigation of transmission cases and to ascertain the most informative and reliable molecular analysis method.

i) gp120 sequence diversity following heterosexually acquired HIV-1 infection in the U.K.

Background

The number of European and North American heterosexually transmitted HIV infections is increasing. Some HIV-1 subtypes have been reported to display selective tissue tropisms resulting in an increased likelihood of infection during particular modes of transmission. For example, subtypes E and C appear to be more easily heterosexually transmitted (245). Thus the possibility of a heterosexually transmitted epidemic in the U.K. may be increased by the introduction of HIV-1 subtypes other than B, e.g. the subtypes prevalent in Africa. In order to gain insight into heterosexual transmission in England four transmission events were investigated by sequencing several single gp120 molecules from each of the six implicated individuals.

Patients studied

The initial investigation involved two individuals: a female, CPHL4, and her male partner, CPHL5 (Table 3.1). A second investigation involved four heterosexuals: CPHL6, A Moroccan male (see Table 3.1); who infected 3 females in the U.K., CPHL7; 8; and 9 (Table 3.1).

Results

Sequences were generated using the single molecule amplification methods described in sections 2.3.4, 2.3.5, 2.3.7-10 and were analysed and compared with reference sequences from the Los Alamos HIV database using the methods described in section 2.4. The epidemiological linkage of these cases was confirmed by their relative *env* sequence similarity, compared with unlinked sequences sequenced in this laboratory and with sequences from the Los Alamos HIV database. Figures 4.1 and 4.2 show the phylogenetic trees of nucleotide and derived amino acid sequences respectively, generated using the neighbour-joining method described in section 2.4.4. Comparison of their sequences with known subtypes revealed that one male had transmitted HIV-1 of subtype B (typically associated with Europe/ U.S.A.) and the other male had transmitted subtype D (typically associated with infection in Africa). Table 4.1 shows the pairwise gp120 distance between the subtype B donor (CPHL6) and recipients (CPHL7, 8 and 9) was 7.38% (±0.98, n=21). The pairwise gp120 distance between the subtype B recipients themselves (CPHL7, 8 and 9) was 7.94% (±0.79, n=16). The mean pairwise distance of gp120 between unlinked subtype B virus from material received and sequenced in this laboratory was 12.22% (±1.21, n=130). Mean pairwise distance for the gp120 molecules from a subtype D linked infection (CPHL4 and 5) was 8.8% (±1.25, n=9).

Table 4.1

Comparison of percentage pairwise nucleotide and amino acid distances between the gp120 data sets.

	Unlinked B	Linked B (CPHL6 to 7,8,9)	Linked D (CPHL4, 5)
MEAN	12.22 (19.85)[1]	7.38 (14.84)	8.8 (16.67)
S.D.	1.21 (2.34)	0.98 (1.43)	1.25 (1.42)
N	130	21	9

[1]: = amino acid sequence distances are in parentheses

S.D. = standard deviation

N = number of comparisons (number of molecules)

Figure 4.1

Phylogenetic tree of nucleotide sequences generated from heterosexually acquired HIV-1 infections in the U.K., constructed using the neighbour-joining method described in section 2.4.4. Further specimen details are shown in Table 3.1. The number following the CPHL number refers to the single molecule sequenced. Reference sequences shown in bold. The degree of sequence dissimilarity is shown on the horizontal axis

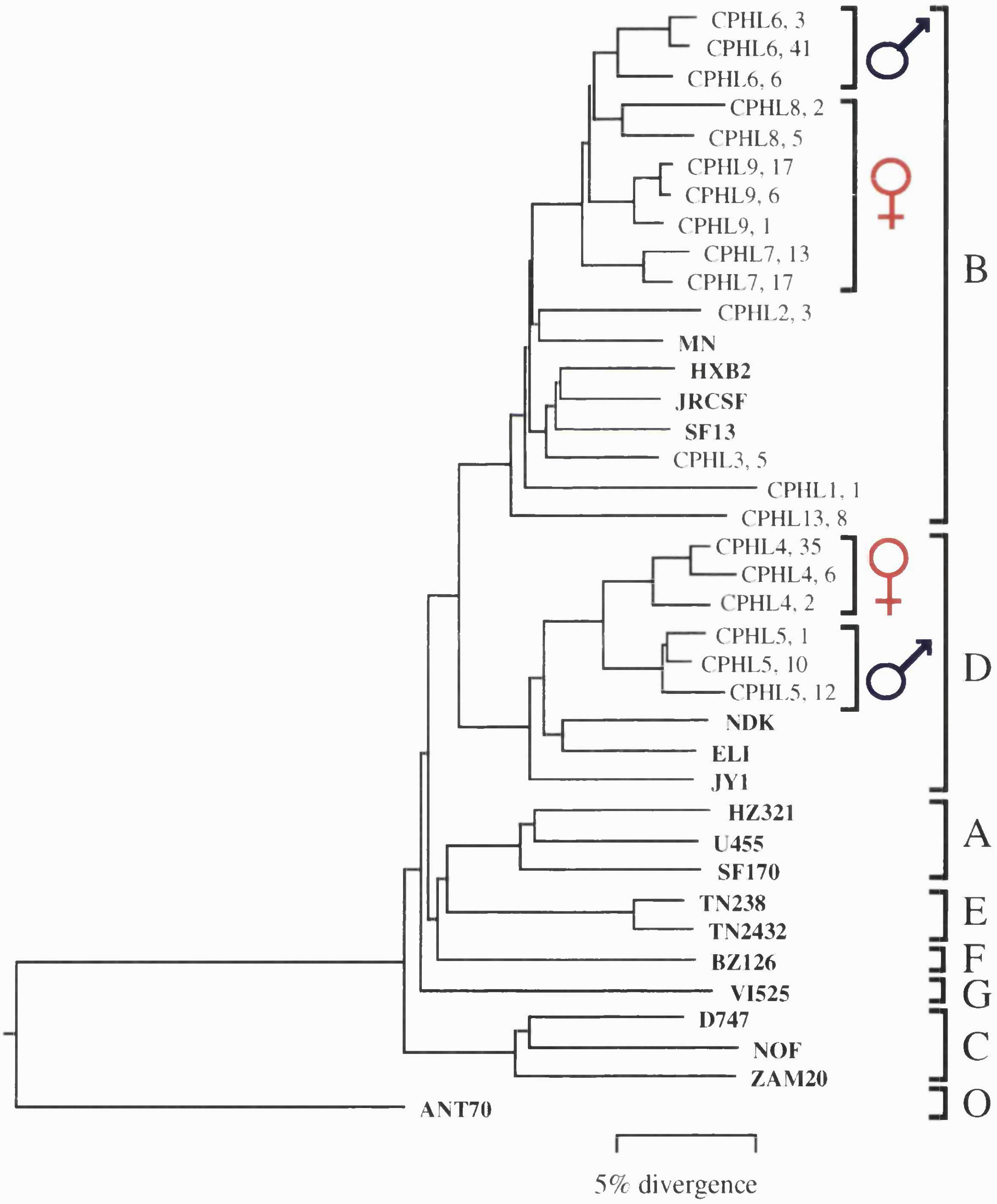
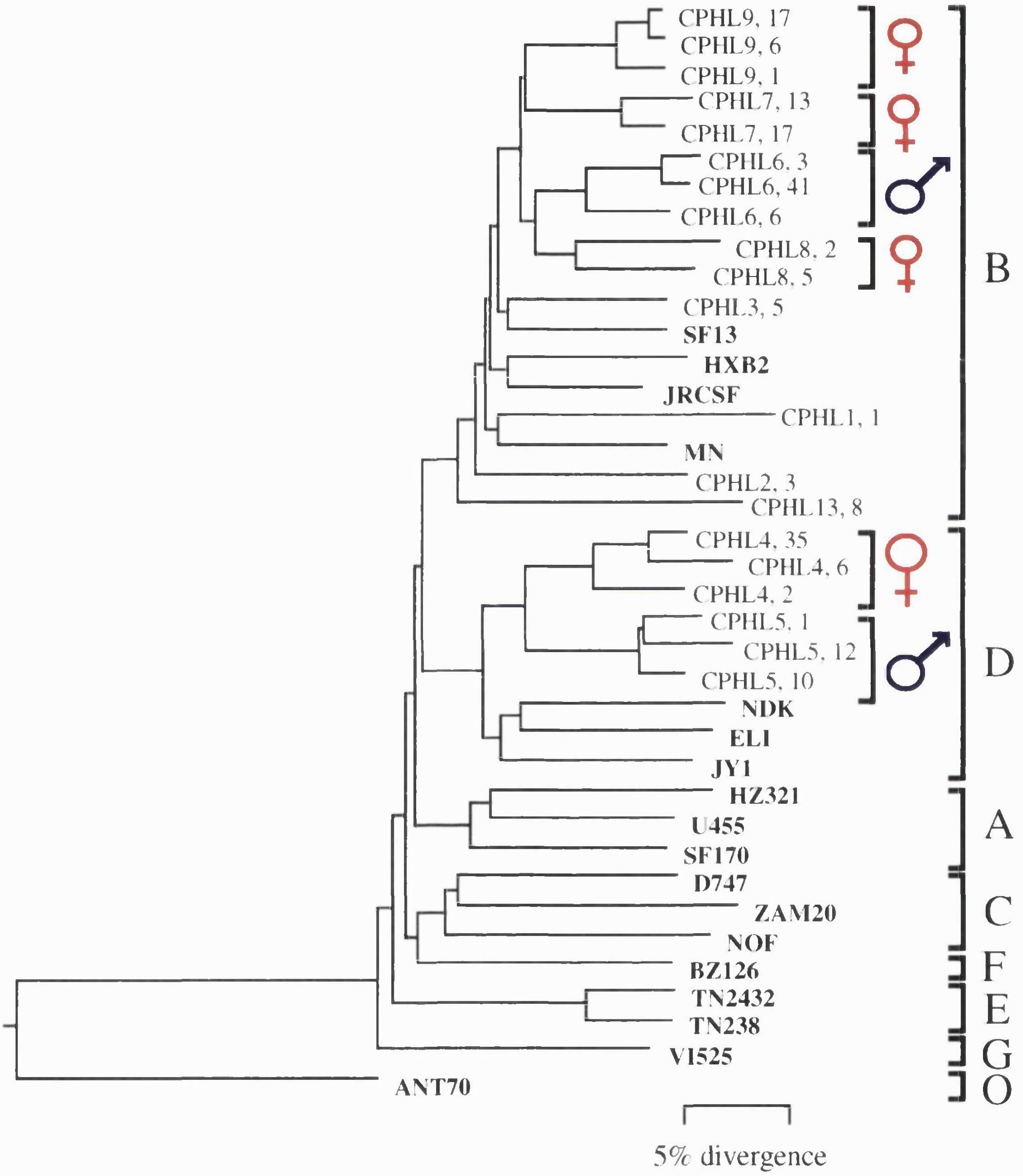


Figure 4.2

Phylogenetic tree of derived amino acid sequences from heterosexually acquired infections in the U.K. Other details as for figure 4.1.



Discussion

Results from the subtype B transmission case (CPHL6 to CPHL7, 8 and 9) show that mean pairwise sequence distances of individual gp120 molecules between recipients, at 7.94% (±0.79), is only marginally higher than that between donor and recipients (7.38% ±0.98) (Table 2). Therefore, if transmission to the recipients takes place at a similar time, and material for analysis is obtained relatively soon after infection (i.e. 2-3 years), it is possible that the HIV variants will be similar enough to establish a common source without obtaining material from the donor.

The results from the subtype D linked infection (CPHL5 to CPHL4) show that although material for analysis was obtained from the two individuals possibly 8 years after transmission, it was still possible to establish some similarity between their gp120 sequences. However, at 8.8% (±1.25), the mean pairwise sequence distance is approaching that between unrelated infections. There are fewer subtype D sequence data (i.e. other type D infections from the U.K.) available to provide a background population for comparative purposes, which may complicate analysis for non-B infections.

ii) Investigation of possible transmission from an HIV infected surgeon to a patient

Background

Patients studied

This investigation initially involved a possible HIV-1 transmission from an infected health care worker (HCW, CPHL1) and a female patient of CPHL1 (CPHL2). The epidemiological investigation in this case was described by Hochuli and colleagues (114). In 1986 CPHL2 had undergone three procedures performed or supervised by CPHL1 that carried a possible exposure risk. Thus transmission, if it occurred between CPHL1 and CPHL2, took place 6-7 years before specimens were collected from them both. Subsequent to the above investigation, material became available from two further females (CPHL3 and CPHL14) who were members of the same heterosexual sex circle as CPHL2 (a sex circle being here defined as a group of people who regularly swap sexual partners). Controls for this investigation include material from the male to female transmissions described above: CPHL4 and 5; and CPHL6, 7, 8 and 9.

Results

Using the methods described in sections 2.3.4, 2.3.5 and 2.3.7-10, several single gp120 molecules from these individuals were sequenced and the pairwise nucleotide distances compared between sequences. One cDNA sequence was generated from a serum sample from CPHL1, using the methods described in sections 2.3.6-7, 2.3.9, and 2.3.10. The mean pairwise distance between the cDNA molecule generated as control and the proviral DNA sequences for that individual (CPHL1) was 2.7% (±1.26, n = 6). The mean pairwise distance between CPHL1 and CPHL2 sequences was found to be 13.84% (±0.59, n=24). The mean pairwise distance of gp120 between unlinked subtype B virus from material received and sequenced in this laboratory was 12.22% (±1.21, n=130). As specific controls for the methods used here several gp120 molecules from heterosexual transmission cases described above were used. The mean pairwise gp120 distance between molecules from the donor and each of the recipients (CPHL6, 7, 8 and 9) was 7.38% (±0.98, n=21). For these cases, blood samples from the donor and recipients were taken approximately 2-3 years post-infection. The pairwise gp120 distance between the recipients themselves (CPHL7, 8 and 9) was 7.94% (±0.79, n=16). The mean pairwise distance for the gp120 molecules from a subtype D linked infection (CPHL4 and 5), where blood samples from both individuals were taken as much 8 years post-transmission, was 8.8% (±1.25, n=9). The sequence distances between CPHL1 and 2 were also compared with control sequences obtained from the Los Alamos database.

The mean pairwise distance between gp120 molecules from the three female members of the sex circle were compared (CPHL2, 3 and 14). The mean pairwise distance between CPHL2 and CPHL3 was 10.99% (± 0.36, n=12). The mean pairwise distance between CPHL2 and CPHL14 was 5.59% (± 0.48, n=12). The mean pairwise distance between CPHL3 and CPHL14 was 11.3% (±0.56, n=9).Table 4.2 shows a comparison of percentage pairwise nucleotide and amino acid distances between the gp120 data sets reported in this work. Alignment of the envelope sequences by ClustalV (110) required the insertion of a large number of gaps due to length variation between individual samples. This complicates the interpretation of the phylogenetic relationships between individual species because there is, at present, no consensus on the correct interpretation of such gaps. However, it was found that the results of our analyses were broadly similar whether or not gaps were accounted for. Simulation studies suggest that phylogenetic analysis using maximum likelihood inference is an effective way to discover evolutionary relationships (221) when material from the alternative possible source of infection is available, as it allows a statistical comparison of the relative likelihoods of the different transmission pathways. However, all methods of phylogenetic reconstruction assume that most substitutions are neutral (115) and it is therefore difficult to assess the performance of the different methods in the presence of natural selection, convergent evolution (118) and the extreme bias in base substitution seen in HIV-1, in particular the strong tendency for G to A changes (183). As no consensus exists about which analytical method is best both maximum likelihood analysis and neighbour joining methods were used. The methods described in section 2.4 6-7 were used to construct phylogenetic trees using both methods and the resultant trees were found to be very similar. The phylogenetic tree obtained was optimised by global branch swapping. In order to assess the robustness of the tree, bootstrap resampling was performed (section 2.4.7). The same analyses were repeated using just the C2-V3 region of the envelope gene, previously used in several studies to assess phylogenetic relationships. MegAlign (2.4.4.) was also used which employs the

Clustal method (110) to align the sequences, followed by the neighbour joining method to construct phylogenetic trees.

Table 4.2

Comparison of percentage pairwise nucleotide and amino acid distances between the gp120 data sets

	CPHL1/2	CPHL2/3	CPHL2/14	Unlinked B	Linked B (CPHL6 to 7,8,9)	Linked D (CPHL4, 5)
MEAN	13.84 (22.8)	10.99 (19.55)	5.59 (13.13)	12.22 (19.85)[1]	7.38 (14.84)	8.8 (16.67)
S.D.	0.59 (1.63)	0.36 (0.34)	0.48 (1.15)	1.21 (2.34)	0.98 (1.43)	1.25 (1.42)
N	24	12	12	130	21	9

[1]: = amino acid sequence distances are in parentheses

S.D. = standard deviation

N = number of comparisons (number of molecules)

Figure 4.3 shows the phylogenetic tree obtained from the neighbour joining analysis described in section 2.4.4 with nucleotide sequences from CPHL and subtype representatives from the Los Alamos database (shown in bold). CPHL1, 2, 3 and 14 sequences are shown in red. Analysis of the same data set using other phylogenetic methods gave essentially the same results. The separate branches containing the CPHL1, CPHL2 and CPHL3 gp120 sequences remained in all resamplings. The branch containing CPHL2 and CPHL14 remained in all resamplings. Phylogenetic analysis of amino acid sequences rather than nucleotides gave trees with the same branching order. Figure 4.4 shows the phylogenetic tree obtained from analysis of derived amino acid sequences from CPHL and subtype representatives from the Los Alamos database (shown in bold). CPHL1, 2, 3 and 14 sequences are shown in red. All gp120 nucleotide sequences obtained in this study gave open reading frames (ORFs).

In summary, all methods failed to group the HCW (CPHL1) and patient's (CPHL2) sequences together. All methods grouped the sequences from two of the three members of the sex circle together (CPHL2 and CPHL14). This was true for both nucleotide and amino acid sequence analysis. The CPHL3 sequences were found to be more closely related to one of the other sex circle members (CPHL2) at 10.99% (± 0.36, n=12) divergence than they are to any other of the sequences analysed. The analyses were also performed using just the C2-V3 region of the gp120 molecule, which has previously been used in other analyses of transmission events (199, 286). The resultant branching order for this reduced data was congruent, though the branch lengths were somewhat smaller overall. Bootstrap resampling of the subset did lead to several alternative topologies which were not seen when the whole data set was analysed (1000 trees, data not shown). For example, the CPHL3 set of three sequences which, when analysed as full length genes are separated from the rest of the sequence set (1000/1000 bootstrap resamplings), are only separated as a distinct group in two thirds of resamplings of the subset (634/1000), joining the CPHL8 data set in the remaining samplings. This result may be due either to the small number of nucleotide differences between the two data sets (10 bases), or to convergent evolution of these regions (118). This analysis led to the conclusion that the 312 bp C2-V3 subregion of gp120 was a less reliable data set than a full gp120 sequence. The extensive phylogenetic analysis carried out for these transmission data sets led to the conclusion that several single molecule full gp120 sequences generated from each individual would show linkage whichever phylogenetic method was used, whether gaps were included in the alignment or not. Using full gp120 to evaluate whether HIV transmission occurred, as this region is approximately 1500-1600 nucleotides in length, quickly generates computationally unwieldy data sets for both distance and parsimony methods once relevant controls are included. As the algorithms used to analyse nucleotide sequences behave differently when amino acid sequences are used, gp120 sequence data sets from transmission investigations were routinely analysed at both the nucleotide and amino acid level using the computationally less demanding, though still reliable, Clustal/neighbour joining method used by MegAlign (section 2.4.4). If, after following this method of analysis, the topologies of the trees were similar, a high level of confidence was achieved when large data sets such as those described in this thesis were involved.

Figure 4.3

Phylogenetic tree obtained from neighbour joining analysis described in section 2.4.4 of sequences from CPHL and subtype representatives from the Los Alamos database (shown in bold). Sequences generated from specimens from the HCW and 3 sex circle members (CPHL1, 2, 3 and 14 respectively) are shown in red.

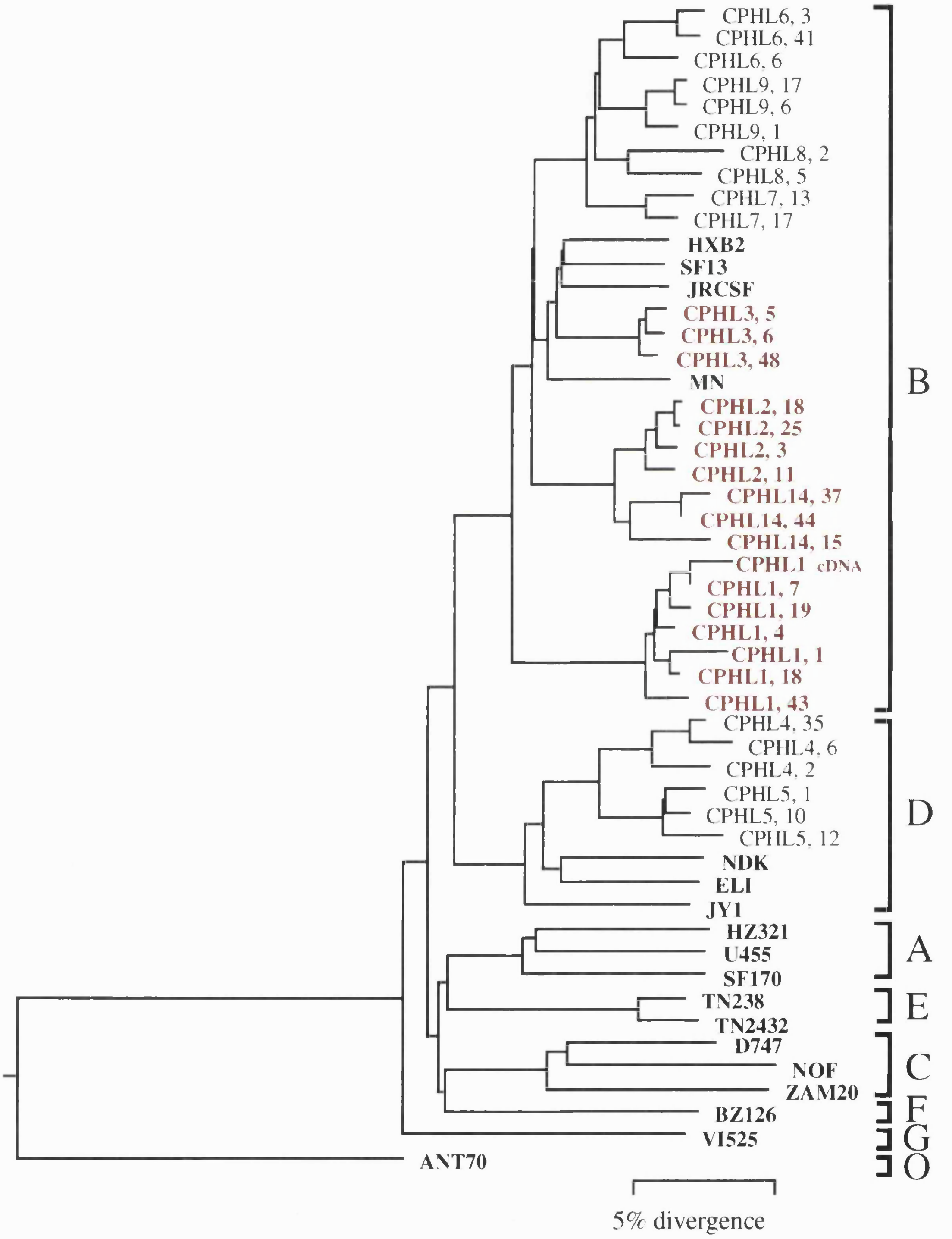
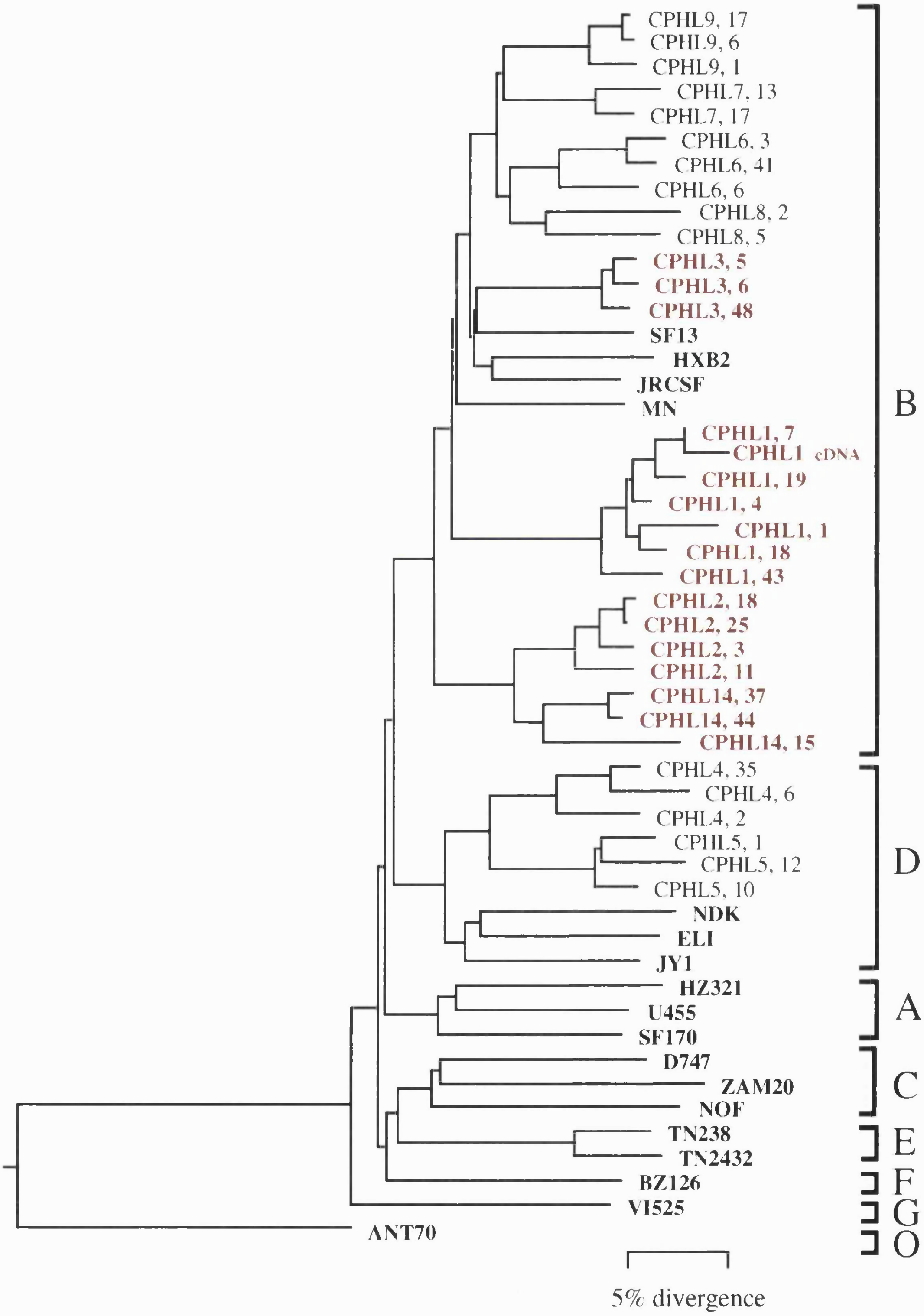


Figure 4.4

Phylogenetic tree obtained from analysis of derived amino acid sequences from CPHL and subtype representatives from the Los Alamos database (shown in bold). Other details as for figure 4.3.



Discussion

The gp120 region of the HIV-1 genome provides both variable and conserved regions. Theoretically, if transmission had occurred between CPHL1 and CPHL2, pairwise gp120 sequence variation from both might be expected to be around 8%. Mean pairwise differences between gp120 sequences from CPHL1 and CPHL2 were found to be 13.84% (±0.59), as great or greater than differences between unrelated subtype B sequences.

As CPHL2, 3 and 14 were members of the same sex circle, it is likely that their infections were linked. However, as all three members analysed thus far are female, direct transmission is unlikely. The mean pairwise sequence distance between single molecules of gp120 from CPHL2 and CPHL14 was 5.59% (±0.48). This was well within the range for linked subtype B infections seen in this laboratory. They were also within the range for intraperson sequence variation described in this thesis (1.45-6.52%). The mean pairwise sequence distance between single molecules of gp120 from CPHL2 and CPHL3 was 10.99% (±0.36). The routes and possible times of transmission of HIV-1 within this sex circle are unknown and so it cannot be definitively established whether the CPHL2 and 3 infections are linked. The higher mean percentage divergence between the sequences from these two individuals indicates that the sequences are unlikely to be related.

Results from some of the control data used in this study (CPHL6 transmission to CPHL7, 8 and 9) show that mean pairwise sequence distances of individual gp120 molecules between recipients, at 7.94% (±0.79), is only marginally higher than that between donor and recipients (7.38% ±0.98). Material obtained for analysis from this group was acquired approximately 2-3 years after infection in all 3 cases. All had developed advanced disease with CD4+counts below 200 within this time. These data, and the data from the two females CPHL2 and 14 between whom direct transmission was unlikely, show that it may be possible to investigate transmission events when material is not available from the index case if there is more than one presumed recipient. Therefore, if transmission to the recipients takes place at a similar time, and material for analysis is obtained relatively soon after infection (i.e. 2-3 years), it is possible that the HIV variants will be similar enough to establish a common source.

The results from the analysis of the C2-V3 region (studied by (199) in the Florida dentist investigation) compared with the complete gp120 of this data set showed that the branches of the two trees generated were identical when gaps in the alignment were omitted, but were discordant when gaps were included. Using the data presented here, the 312 bp amplicon analysed by Ou and colleagues has too few nucleotides for accurate phylogenetic analysis to be carried out as it was observed that one base change in a sequence reordered the tree. These results indicated that gp120 is a more informative region to study than C2-V3 for molecular transmission studies, also the importance of gap stripping prior to analysis in order to avoid length variation which, as indicated by the example shown above, may skew results.

iii) Occupationally acquired HIV-1 infection in the U.K.

Background

Occasionally unusual modes of HIV-1 transmission appear to be the likely cause of subsequent infection. It is estimated that only three of the 8115 AIDS cases and 20,543 HIV infections in the U.K. reported by the end of September 1993 may have occurred by these unusual routes (8). Although these occurrences are rare, they emphasise the need to avoid exposure to blood from HIV infected individuals. The much publicised case of the Florida dentist is, to date, the only case of a HCW transmitting HIV to his patients (202). The route of transmission in that case is still unknown and will probably remain unresolved. Still very rare, but more likely, are occupationally acquired infections.

Needlestick transmission investigation

Most HIV transmission investigations have used comparison of sequence data generated from proviral DNA from apparently epidemiologically related cases to determine whether actual transmission had taken place (4, 9, 14, 119, 129, 199). However, HIV sequence evolution can be discontinuous (236), in that the predominant virus strain isolated from PBMCs at a specific time may not be related to the predominant strains isolated before and since that time. Simmonds and colleagues found sequence changes over time in both the PBMC and the plasma RNA populations, but new variants appeared first in the plasma RNA population. Only subsequently did they become detectable in the PBMCs. This implies that the time interval since infection with HIV may be a determinant of the extent of sequence variation observed. This study compares sequence data generated from RNA from an AIDS patient with proviral DNA from a nurse exposed to his blood. This approach was necessary as there was only a stored serum sample available from the AIDS patient.

Clinical Case Report

A case of possible transmission of HIV by needlestick injury from a patient (CPHL12, see Table 3.1) dying of AIDS to a nurse (CPHL13, see Table 3.1) was investigated. Only a frozen serum sample was available from the presumed source of infection, who had recently died of AIDS. The serum sample from the presumed source had been taken and frozen 5 months before the needlestick incident took place and had not been freeze/thawed. The individual who sustained the needlestick injury was nursing CPHL12 at a London Hospice. On December 18th 1994 a needle attached to an empty 2 ml syringe lying in the bedclothes pricked his hand. The needlestick accident was not formally reported at the time. The incident was further complicated by the fact that CPHL13 is homosexual, though with no recent exposure. He sought hepatitis vaccination and HIV testing after the incident, and blood samples taken subsequently are consistent with HIV seroconversion. He returned for a second HIV antibody test on 7/2/95 which was positive. The EDTA blood specimen for sequencing was collected on 12/7/95.

Use of serum compared with lymphocytes

A control for serum (virion RNA) sample for which sequence data was available from the corresponding lymphocytes (proviral DNA) was available from a previous investigation (9). This serum was from CPHL1 (see Table 3.1), an individual who was at approximately the same disease stage as CPHL12 and for which proviral DNA sequences were available for pairwise comparison with the sequence from the viral RNA from the serum.

Results

A cDNA sequence was generated from the serum sample CPHL12, using the methods described in sections 2.3.6-7, 2.3.9, and 2.3.10. Two single molecules were sequenced using proviral DNA from CPHL13 as template generated using the methods described in sections 2.3.4-5 and 2.3.7-10. Sequences were then analysed using the methods described in section 2.4. The pairwise sequence distances between the cDNA from CPHL12 and the two proviral sequences from CPHL13 were 2.9% and 2.5%. The mean pairwise distance between the cDNA molecule generated as control and the proviral

DNA sequences for the same individual (CPHL1) was 2.7% (±1.26, n = 6). Figure 4.5 shows a neighbour joining phylogenetic tree of the CPHL12 and 13 gp120 sequences (shown in red), including database sequences (shown in bold), unlinked subtype B controls; CPHL1, 2 and 3; and transmission controls: subtype B; CPHL6, 7, 8 and 9; CPHL2 and 14 and subtype D; CPHL4 and 5. Figure 4.6 shows a neighbour joining phylogenetic tree of the derived amino acid sequences.

Figure 4.5

Neighbour joining phylogenetic tree of the CPHL12 and 13 gp120 nucleotide sequences (shown in red), including database sequences (shown in bold), unlinked subtype B controls; CPHL1, and 3; and transmission controls: subtype B; CPHL6, 7, 8 and 9; CPHL2 and 14 and subtype D; CPHL4 and 5.

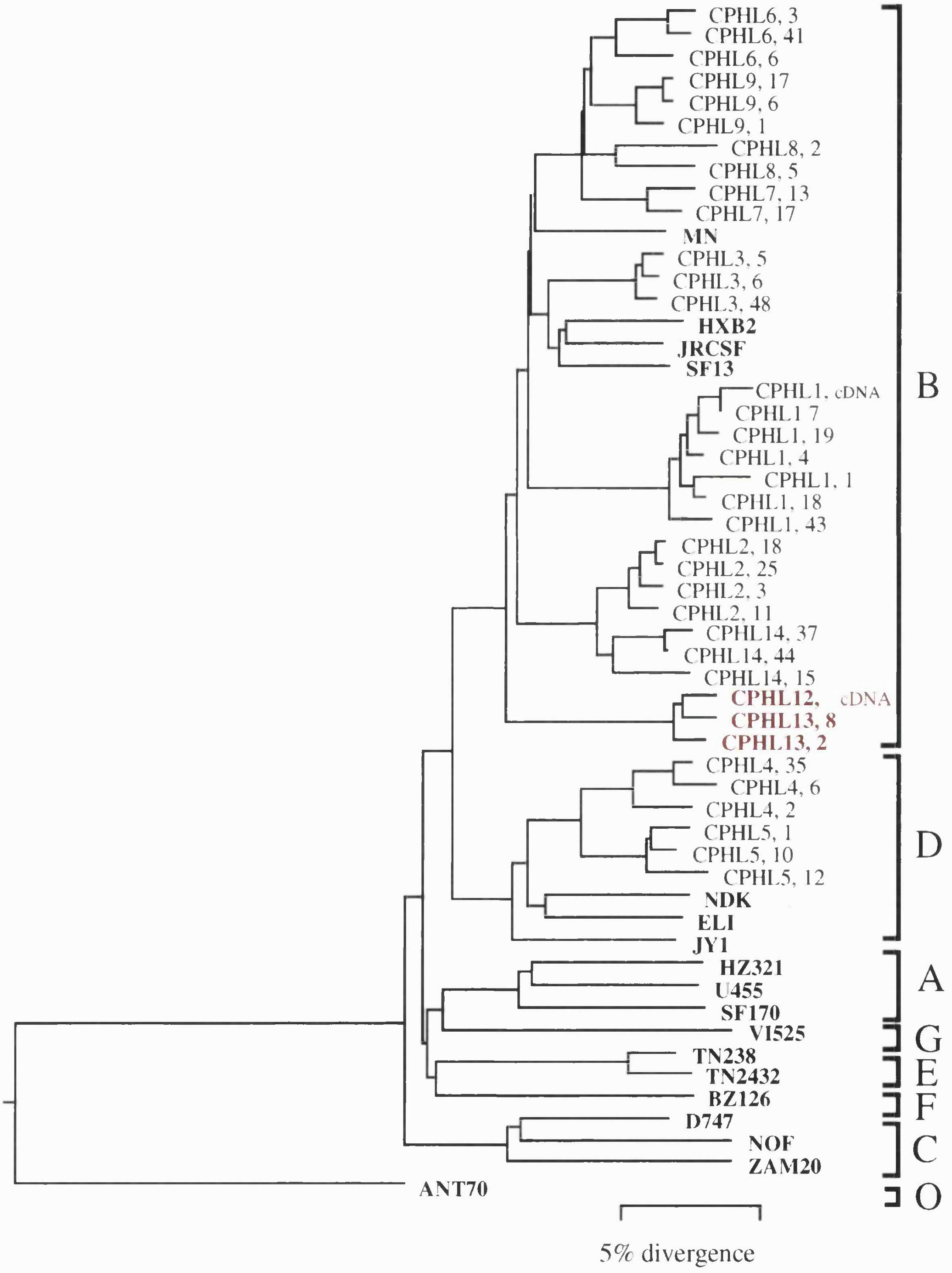
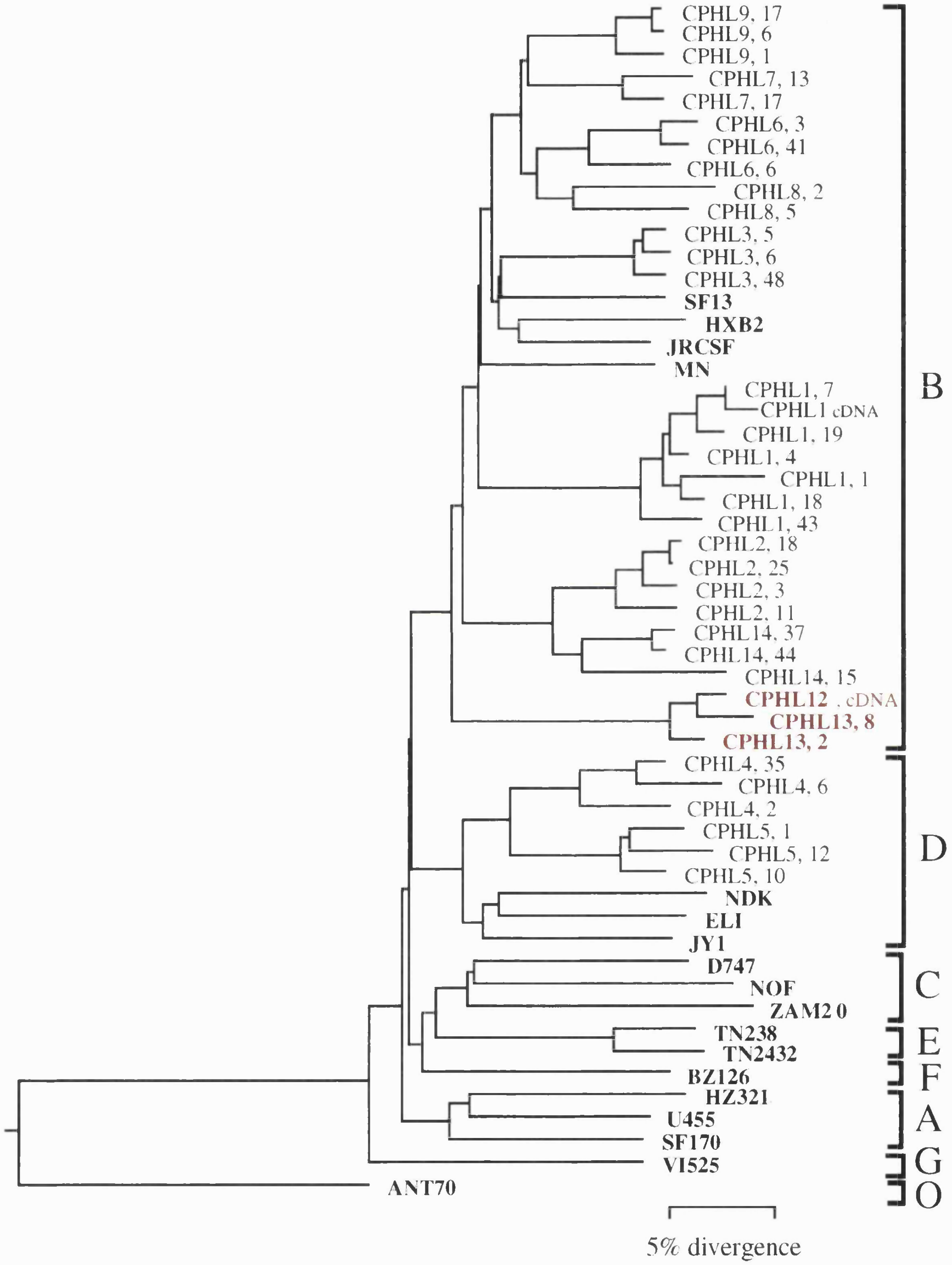


Figure 4.6

Neighbour joining phylogenetic tree of the CPHL12 and 13 gp120 derived amino acid sequences (shown in red). Other details as for Figure 4.5.



Discussion

As described in chapter 1, HIV-1 gp120 sequence data have been used previously to study transmission investigations (9). Work described in this thesis establishes intraperson gp120 sequence variation at between 1.45-6.52%. The mean pairwise distance of proviral gp120 sequences between unlinked subtype B infections from material received and sequenced for this thesis is 12.22% (± 1.21, n = 130). The mean pairwise distance between linked subtype B gp120 proviral DNA sequences is 7.38% (± 0.98, n = 21). The pairwise distances between sequences from CPHL12 and CPHL13 at 2.5% and 2.9% are well within the range for linked subtype B infections seen in this laboratory. They are also within the range for intraperson sequence variation.

Simmonds *et al* and others (185, 225, 236, 291) showed that there can be significant differences between the frequencies of sequence variants in DNA and RNA populations within the same sample. The cDNA/proviral DNA control pairwise distances from CPHL1 indicated that at later stages of disease (and possibly throughout the infection) cDNA and proviral DNA sequences, if generated from samples taken from the putative donor and recipient are within a few months of each other, are similar enough to be used to determine epidemiological linkage. Thus, serum as well as lymphocytes (or a combination of both) can be used as the starting material for this type of molecular transmission investigation. It is likely that this type of transmission investigation (i.e comparison of RNA and DNA) will become more common as serum is stored more routinely than PBMCs.

In conclusion, the pairwise distances and the phylogenetic trees (Figure 4.5 and 4.6) showing that sequences from the two individuals cluster together on the same branch are consistent with these two infections being closely linked. Until now transmission via a dry used needle is unprecedented in the U.K. especially after such an extended period since initial exposure to the virus. This case has implications for the safe handling of contaminated sharps even after extended periods of time since exposure

Cadaver handling investigation

Background

Heart, liver, and lung post mortem fixed tissue samples were received from two individuals who had committed suicide on a railway track. Another set of specimens was from a railway worker who had helped clean up after the suicides and was subsequently found to be HIV-positive. He had no risk factors for HIV infection other than being exposed to potentially infected blood/tissue from the two suicide cases. This study aimed to establish whether transmission of HIV could have occurred from either of the suicide cases to the railway worker.

DNA was extracted from the tissue blocks using the method described in section 2.3.5 b and in-house nested PCR (section 2.3.7)was carried out for the *gag/pol* junction (p6/protease), and partial p24 (*gag*) as described in (15). The Roche Amplicor HIV diagnostic kit was also used to determine the presence of HIV DNA. HLA tissue typing described in section 2.3.11 was carried out on the extracted DNA to establish both that the extraction method used yielded PCR-amplifiable DNA and to confirm that the tissue blocks were from different individuals. In the event of either of the tissue blocks giving a positive PCR result for HIV, EDTA blood from the railway worker would have been analysed. PCR products from the two individuals would have been sequenced and the data analysed phylogenetically to determine whether transmission was likely to have taken place between the infected individuals.

Human Leukocyte Antigen (HLA) Typing

HLA proteins and their polymorphisms are well characterised serologically and their pattern of Mendelian inheritance is known. They are frequently used in individual identification, particularly in paternity determination. The HLA proteins can be divided into two functionally and structurally distinct groups: Class I and Class II. The HLA D proteins, members of the Class II group, are found on the surface of some lymphocytes and macrophages, and differ from the classical tissue transplantation antigens, the "Class I" HLA A and B proteins found on the surface of most cells (264). Class II proteins and their genetics are well characterised due to their importance in bone marrow transplantation and their association with susceptibility to autoimmune and other diseases. The genes encoding the HLA Class II proteins are organised into three families:DP, DQ, and DR, and are positioned on human chromosome 6. Each HLA Class II protein is composed of two subunits, a and b, which are separately encoded in the DNA of each gene cluster.

The AmpliType HLA DQa Forensic Kit (Perkin Elmer) amplifies a region of the HLA DQa gene. The kit distinguishes six alleles which define twenty-one different groups. Essentially, DNA extracted from a sample is added to a tube containing amplification reagents. The DQa DNA sequences are amplified and the DNA produced is hybridised to DNA probe strips. Capture probes are used to distinguish the six alleles. These probes are immobilised on the probe strips. Base complementarity allows the probes to capture DNA sequences amplified from specific DQa alleles. Captured DQa DNA is detected by a non isotopic colour reaction and the resulting pattern of blue dots on a probe strip reveals the DQa alleles present in the amplified sample.

Results

DNA was extracted from the tissue blocks from both individuals for PCR and tissue typing. HIV PCR (*gag* and *gag/pol* in-house PCR and Roche Amplicor test) was negative for HIV-1 DNA in both cases. Tissue typing revealed the HLA types of the two individuals as 2,3 and 4,4 confirming that the DNA could be extracted and amplified from the fixed tissue blocks and could be shown to be from different individuals.

Discussion

Although amplifiable DNA was extracted from both post mortem samples, HIV DNA in the extracted tissue samples using both in-house and diagnostic PCR methods was not detectable. Therefore, further investigation of putative HIV-1 transmission could not be carried out. However, HIV-1 infection of either individual cannot be dismissed as HIV DNA is difficult to detect in cells other than lymphocytes.

iv) Perinatal transmission

Background

HIV infection of neonates is difficult to ascertain using serology due to the presence of maternal antibody. PCR-based diagnostic tests theoretically circumvent this problem by detecting of viral DNA or RNA (see chapter 1). However, a possible PCR failure was noted in a specimen from an infant (CPHL11, see Table 3.1) born to an HIV-1 infected mother (CPHL10, see Table 3.1) (10). The infant was thought to be HIV-1 infected from clinical observation (*Pneumocystis carinii* pneumonia, PCP), but cell lysates from blood taken 3 months after birth gave negative reactions in the Roche Amplicor HIV-1 PCR assay. The same lysates were found to be reactive in a PCR using primer sets amplifying regions of the *gag*, *pol* and the LTR (136). The specimen was found to be positive for these primers using only 1/20th the sample input volume recommended for the Amplicor assay, suggesting that the PCR failure was due to mismatching rather than low input copy number (10). A specimen from the mother was positive in the Amplicor test. Perinatal transmission was not in question in this investigation, but the lack of reactivity of the infants' specimen to amplify using the Roche Amplicor PCR assay suggested unusual sequences, highly divergent from the mother, or an unusual subtype. Three single molecules were sequenced from each individual using proviral DNA as template generated using the methods described in sections 2.3.4-5 and 2.3.7-10. Sequences were analysed using the methods described in section 2.4.

Results

Figure 4.7 shows a neighbour joining phylogenetic tree of the subtype G CPHL10 and 11 gp120 nucleotide sequences (shown in red), including database sequences (shown in bold), unlinked subtype B controls; CPHL1, and 3; and transmission controls: subtype B; CPHL6, 7, 8 and 9; CPHL2 and 14; CPHL12 and 13 and subtype D; CPHL4 and 5. The tree was generated using the methods described in section 2.4.4. Figure 4.8 shows the neighbour joining tree with the derived amino acids from the same sequences. At 4.7%, the mean pairwise distance between CPHL10 and 11 is within the range for linked B (2.5%-9.3%) and D (6.6%-9.9%) subtype infections described in this thesis.

Figure 4.7

Neighbour joining phylogenetic tree of the CPHL10 and 11 gp120 nucleotide sequences (shown in red), including database sequences (shown in bold), unlinked subtype B controls; CPHL1, and 3; and transmission controls: subtype B; CPHL6, 7, 8 and 9; CPHL2 and 14; CPHL12 and 13 and subtype D; CPHL4 and 5.

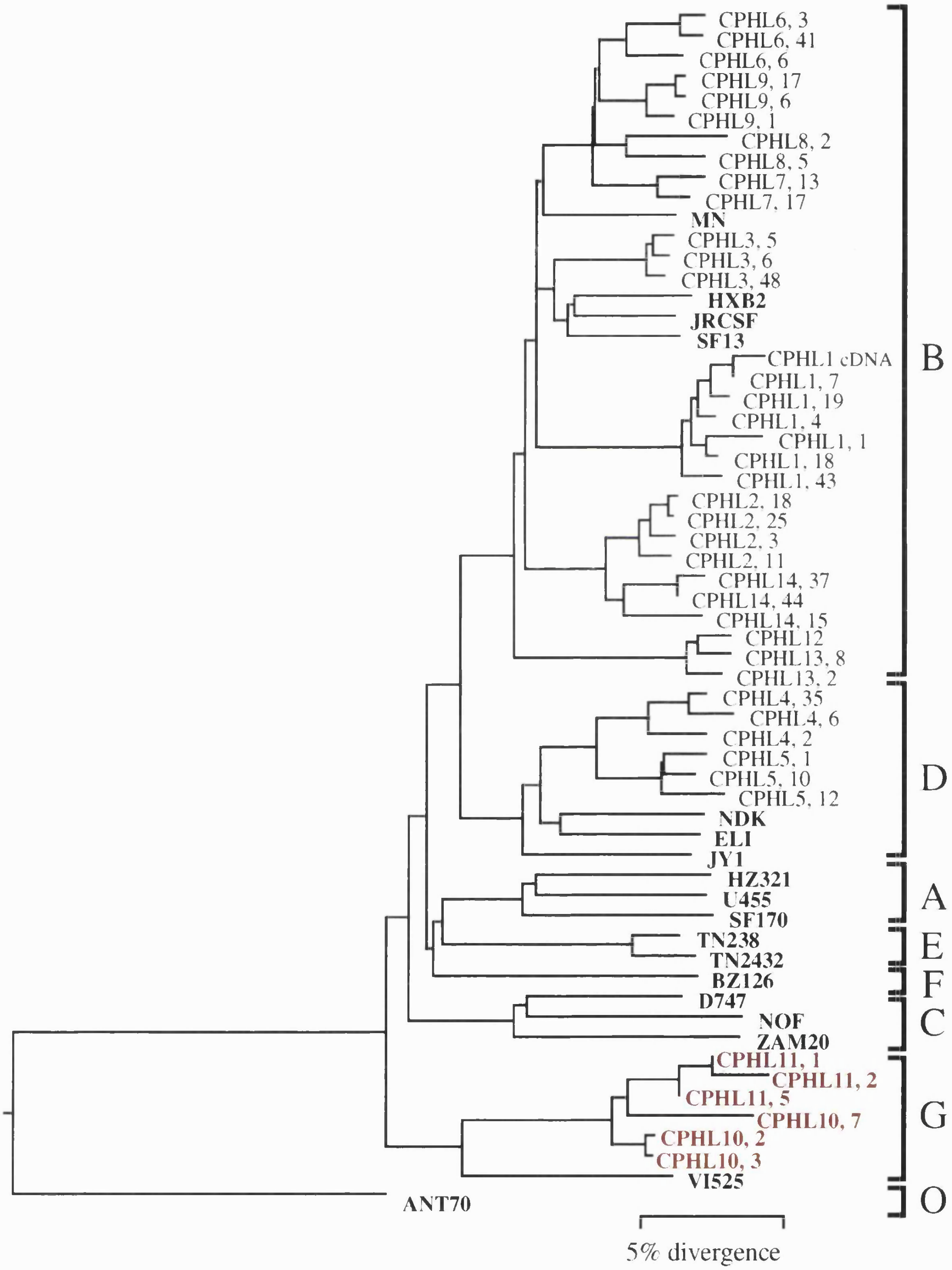
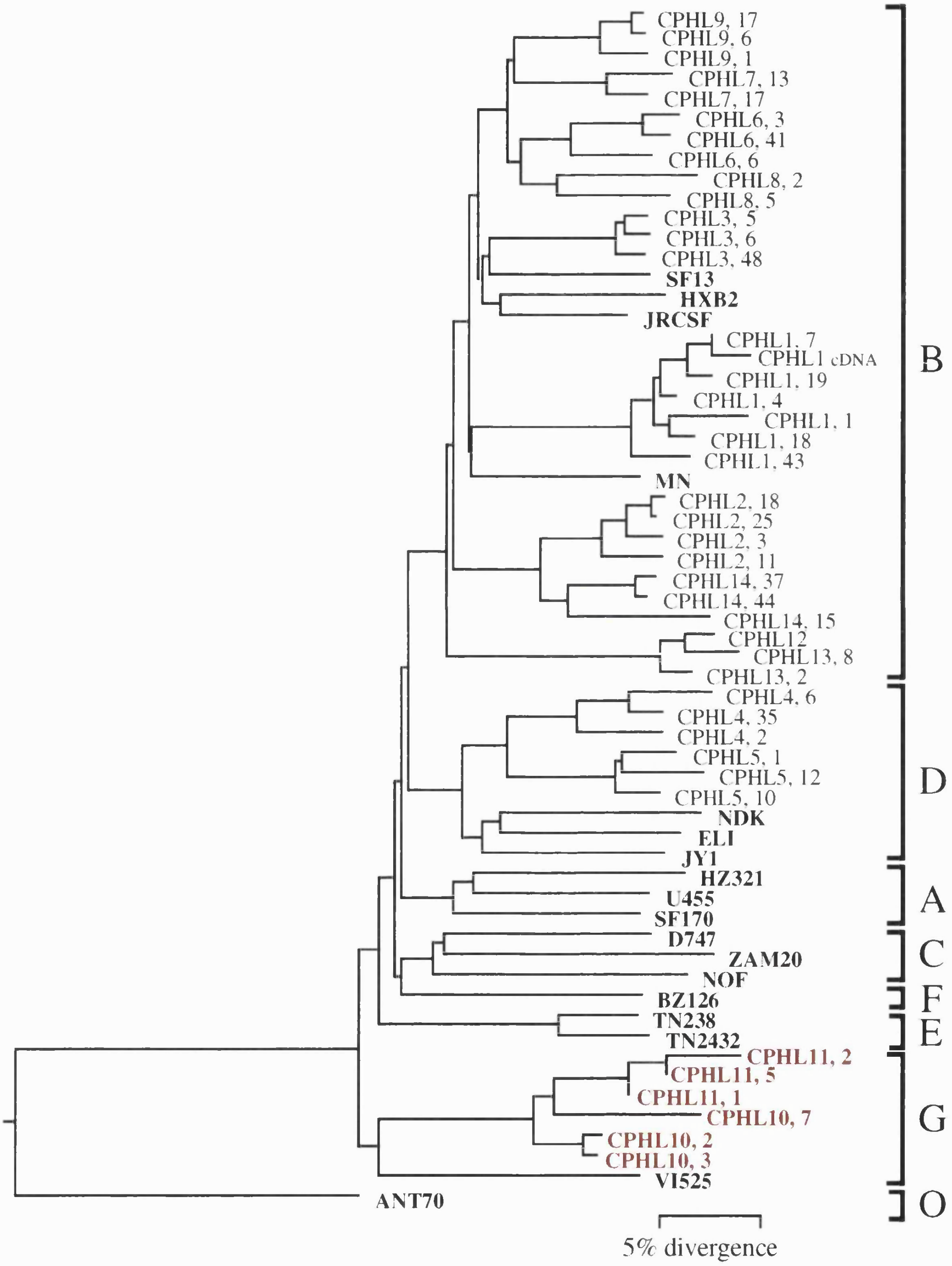


Figure 4.8

Neighbour joining phylogenetic tree of the CPHL10 and 11 gp120 amino acid sequences (shown in red). Other details as for Figure 4.7.



Discussion

The close similarity of sequences from the mother/infant (CPHL10 and 11) transmission results (Figures 4.7 and 4.8) agree with earlier observations that only one variant of the virus genome population, or quasispecies, may be transmitted from mother to infant (225, 287, 292). This variant may be either a minor or major component of the maternal quasispecies (225, 287, 292). In contrast, other reports examining envelope heterogeneity found that multiple maternal HIV-1 subtypes were transmitted to the infants (25, 149). Briant *et al.* reported that more than one maternal variant was responsible for infection in 3 out of four mother/infant pairs. Sequence artifacts generated in the laboratory may confuse some of these issues. For example, Korber *et al.* examined the sequence database and reported instances of identity or near identity in sequences from unlinked mother/infant pairs, suggesting contamination or mix-up of samples had occurred, stressing the need for stringent checks on sequence data, i.e. comparison with sequences of propagated viruses and molecular clones present in the laboratory where experiments are being conducted. They requested that journals encourage authors to describe the methods they have used and to provide the raw sequence data on request to facilitate the review process (144). One study in particular was mentioned by Korber *et al.* The authors of this work, Briant *et al*, answered the criticism by outlining the techniques used, stressing the unlikely scenario of contamination (24). However, as Briant *et al.* had cloned the material, contamination cannot be absolutely ruled out. Also, the choice of the V3 region of gp120 for transmission investigations has been criticised (53, 117, 202, 240) on the basis that it is not large enough and is too variable for reliable conclusions to be drawn. My analysis of the C2-V3 region, abstracted from complete gp120 data described earlier in this chapter, supports this criticism. Another important consideration is the phenomenon of evolutionary convergence of unrelated sequences which has been demonstrated in the V3 region of HIV-1 (118, 251) which may compromise the identification of linked infections. Strunnikova *et al.* observed temporal progression from early, phylogenetically unrelated sequences to late, convergent sequences in two infants (251).

CHAPTER 5

PROTEIN ANALYSIS

This chapter is divided into two sections: i) analysis of derived CPHL amino acid sequences, where 3 or more single molecules were sequenced; and ii) a summary of gp120 sequencing work carried out in collaboration with gp120 cloning, protein expression and monoclonal antibody characterisation work conducted by Jane McKeating at Reading University and Peter Balfe at U.C.L.

i) Sequence analysis of derived CPHL amino acid sequences

Background

As described in chapter 1 , HIV is characterised by great genetic flexibility. Several reports (reviewed in (153)) have shown that in the later stages of the disease the V3 region of *env* becomes more homogeneous in the infected individual. Also, the appearance of SI (syncytium inducing) variants shows that selection pressures drive the evolution of more virulent strains, eventually leading to the development of AIDS. It is known that the high error rate of viral reverse transcriptase (130) and the high turnover rate of RNA *in vivo* (113, 274) contribute to viral diversity within and between infected individuals. Viral diversity is also affected by selective forces, including the immune response of the host, cell tropism and irregular activation of infected cells. Bonhoeffer *et al.* (21) attempted to quantify selection pressures on HIV quasispecies by analysing the ratio of nucleotide substitutions which do not lead to amino acid alterations in a protein sequence (silent or synonymous changes) and those which do (nonsynonymous changes) in the envelope gene of an infected haemophiliac patient followed since seroconversion for 7 years (118). They derived the mean number of nucleotide substitutions per synonymous site, d_s, and per non-synonymous site, d_n, for all pairwise comparisons of sequences in samples taken at 3, 4, 5, 6 and 7 years after infection and showed that *ds* increases with time, due to accumulation of synonymous substitutions. They also noted that the pattern in increase in d_s / d_n corresponds with the pattern of decrease in the CD4 cell count, that is, where the CD4 count decreased, the d_s / d_n ratio increased.

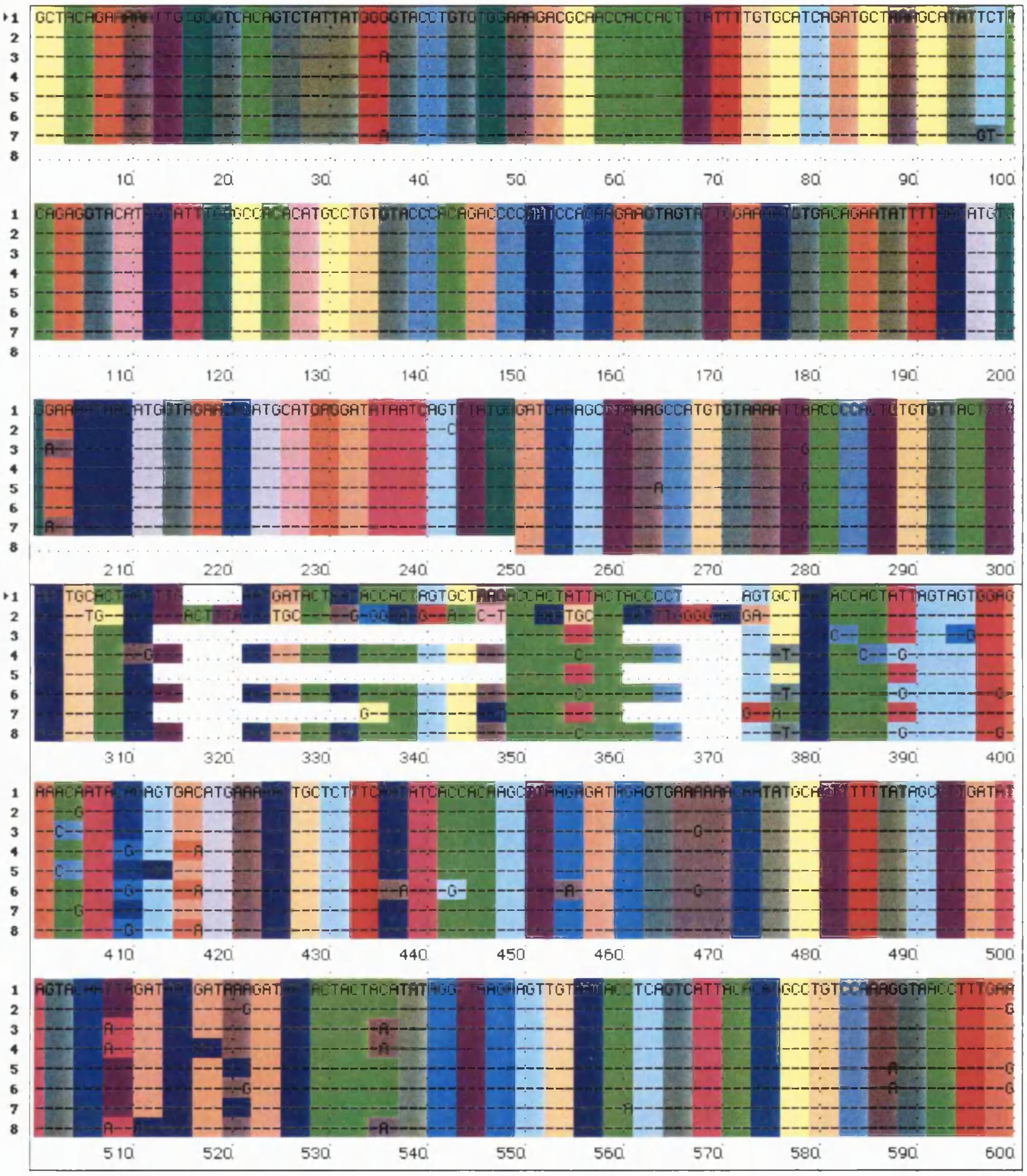
Although sequential samples were not available from the specimens studied in this thesis, further information about HIV-1 envelope d_s / d_n ratios can be gained by the analysis of multiple DNA sequences sampled concurrently. Multiple gp120 sequences from CPHL1, 2, 3, 4, 5, 6, 9, 10, 11, and 14 (for specimen details see figure 3.1) were each aligned (Figures 5.1-5.10 respectively) and the d_s / d_n ratios calculated for each individual by assuming deviation from the consensus to indicate a forward direction of mutation (Table 5.1). Sequences from CPHL11 (Figure 5.10) differed from each other by only 1 or 2 substitutions and therefore were not used for this analysis. Also included in the table are estimations of the expected Transition/Transversion ratios (T_s / T_v). An accurate estimate of T_s / T_v is important for phylogenetic analyses involving maximum likelihood and distance methods as it improves the accuracy of the phylogenetic tree produced from the data. The default value of T_s / T_v in the PHYLIP programs is 2.0, but it can vary from about 0.5 to 10.0 in real data sets (288). A new method of maximum likelihood, Quartet Puzzling, has been developed by Strimmer and von Haeseler (1996, in press). The method has not yet been evaluated in terms of accuracy of phylogenetic inference, but it quickly calculates the T_s / T_v ratio, even with large data sets. This method was also used to estimate the T_s / T_v ratio for the whole CPHL data set. To provide information regarding the pattern of nucleotide substitution for HIV-1 quasispecies within an individual, the data in figure 5.1 was used to calculate a nucleotide substitution table for the 7 CPHL1 sequences (Table 5.2).

Results

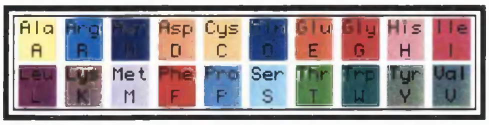
Figures 5.1-5.10 show aligned multiple gp120 sequences from CPHL1, 2, 3, 4, 5, 6, 9, 10, 11, and 14 respectively (for specimen details see figure 3.1). For each figure sequence 1 is the consensus sequence. Each coding triplet is colour coded according to the amino acid key at the bottom of the page. The nonsynonymous nucleotide changes are therefore indicated by change of amino acids (i.e. colour) relative to the consensus sequence. The second key indicates the location of each single molecule in the figure. The

d_s / d_n ratios were calculated for each individual by assuming deviation from the consensus to indicate a forward direction of mutation (Table 5.1). Also included in this table are estimations of the expected Transition/Transversion ratios (T_s / T_v).

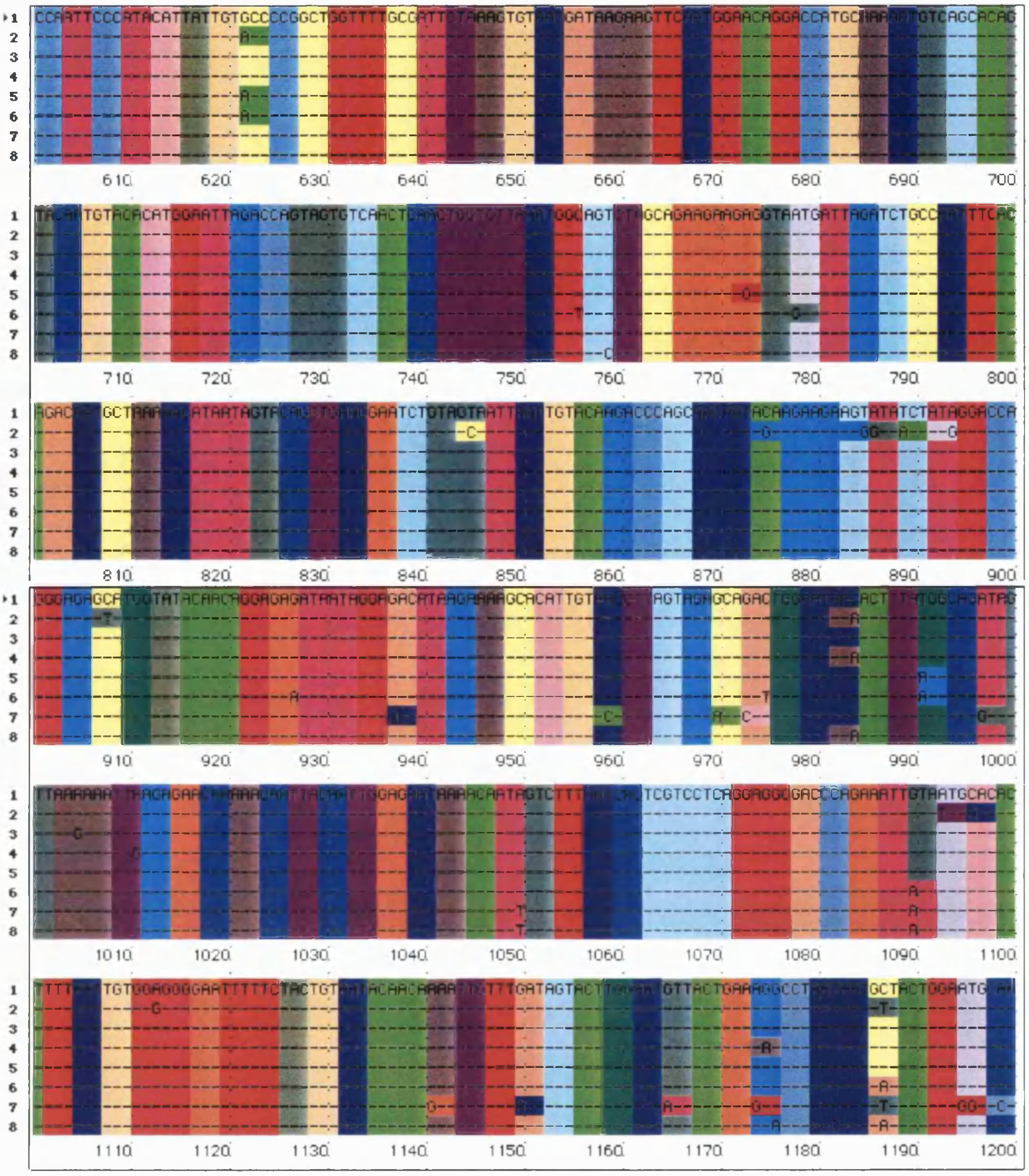
Figure 5.1 Amino acid alignment of CPHL1 nucleotide sequences showing synonymous and nonsynonymous mutations. Each triplet coding for an amino acid is colour coded according to the amino acid key at the bottom of the page. The second key indicates the location of each single molecule in the figure.



1 = CPHL1 consensus
2 = CPHL1. 1
3 = CPHL1. 4
4 = CPHL1. 7
5 = CPHL1. 18
6 = CPHL1. 19
7 = CPHL1. 43
8 = CPHL1 cDNA

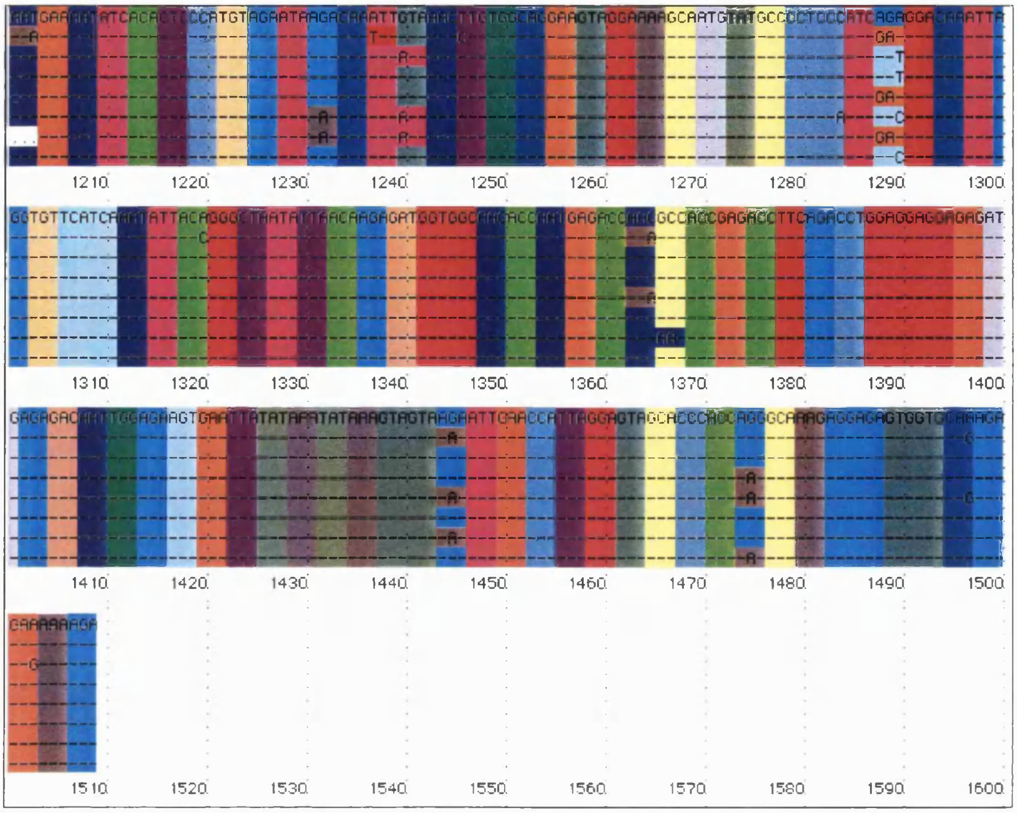


Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V



1 = CPHL1 consensus
2 = CPHL1, 1
3 = CPHL1, 4
4 = CPHL1, 7
5 = CPHL1, 18
6 = CPHL1, 19
7 = CPHL1, 43
8 = CPHL1 cDNA

Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

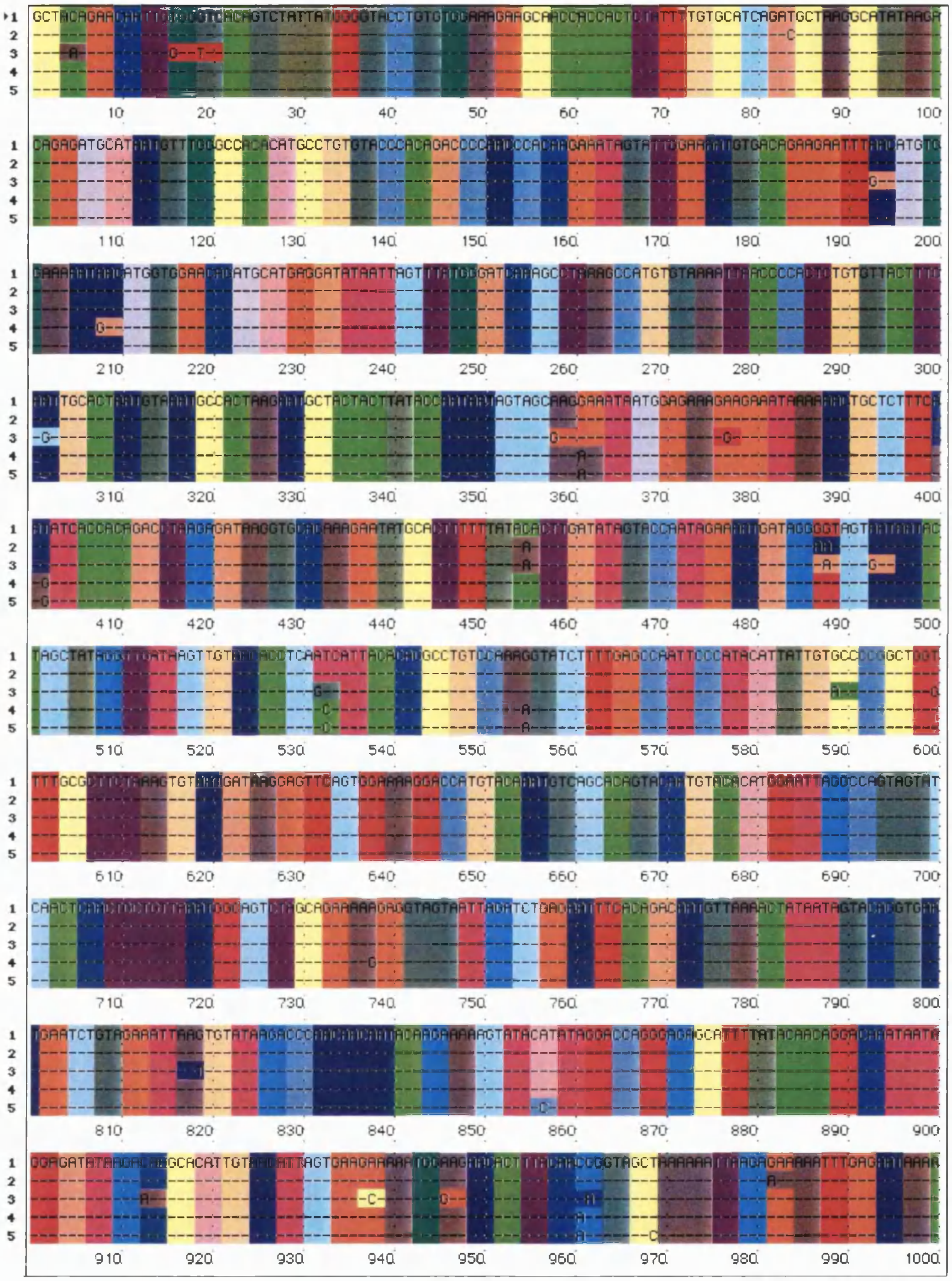

1
2
3
4
5
6
7
8
AATGAAAAATATCACAGTCCCATGTAGAATAAGACAAATTGTAAACTTGTGGCAGGAAGTAGGAAAAGCAATGTATTGCCCCCTCCCATCAGAGGACAAATTA
1210
1220
1230
1240
1250
1260
1270
1280
1290
1300
GGTGTTCATCAAATATTACAGGGCTAATATTAACAAGAGATGGTGGCAACACCAATGAGACCAACGCCAGCGAGACCTTCAGACCTGGAGGAGCAGAGAT
1310
1320
1330
1340
1350
1360
1370
1380
1390
1400
GAGAGACAATTGGAGAAGTGAATTATATAAATATAAAGTAGTAAGAATTGAACCATTAGGAGTAGCACCCAGCAGGGCAAAGAGGAGAGTGGTGCAAAGA
1410
1420
1430
1440
1450
1460
1470
1480
1490
1500
GAAAAAAGA
1510
1520
1530
1540
1550
1560
1570
1580
1590
1600

1 = CPHL1 consensus
2 = CPHL1, 1
3 = CPHL1, 4
4 = CPHL1, 7
5 = CPHL1, 18
6 = CPHL1, 19
7 = CPHL1, 43
8 = CPHL1 cDNA

Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
A	R	N	D	C	Q	E	G	H	I
Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
L	K	M	F	P	S	T	W	Y	V

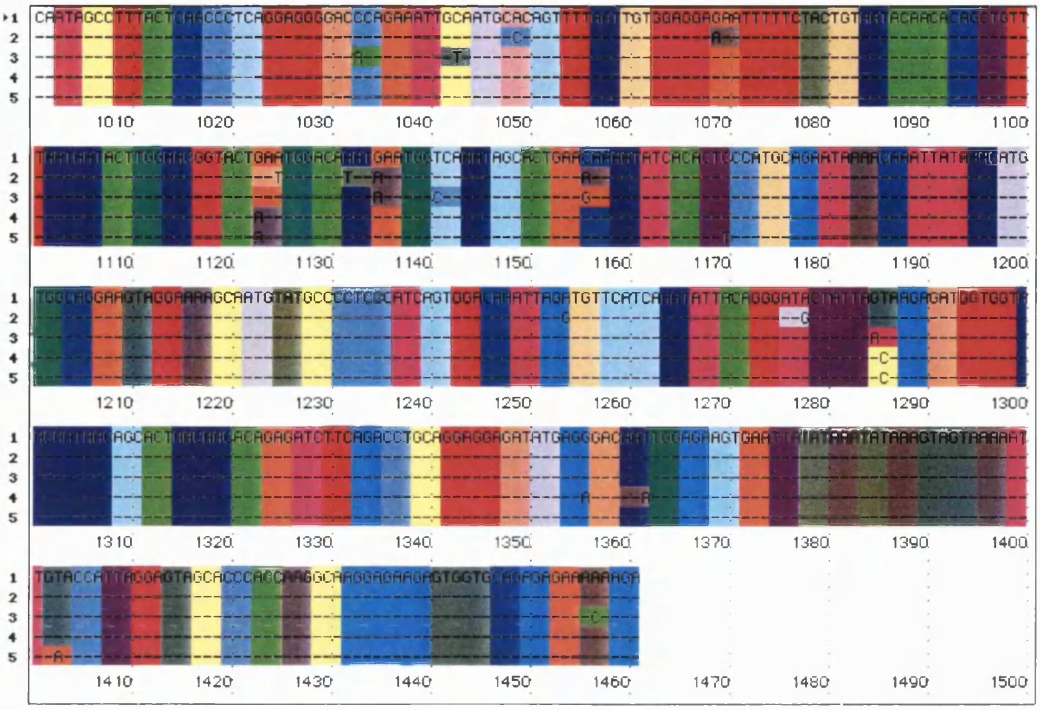
Figure 5.2

Amino acid alignment of CPHL2 sequences. Other details as for figure 5.1.



1 = CPHL2 consensus
2 = CPHL2, 3
3 = CPHL2, 11
4 = CPHL2, 18
5 = CPHL2, 25

Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V



1 = CPHL2 consensus
2 = CPHL2, 3
3 = CPHL2, 11
4 = CPHL2, 18
5 = CPHL2, 25

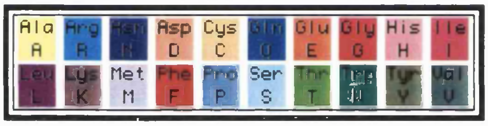
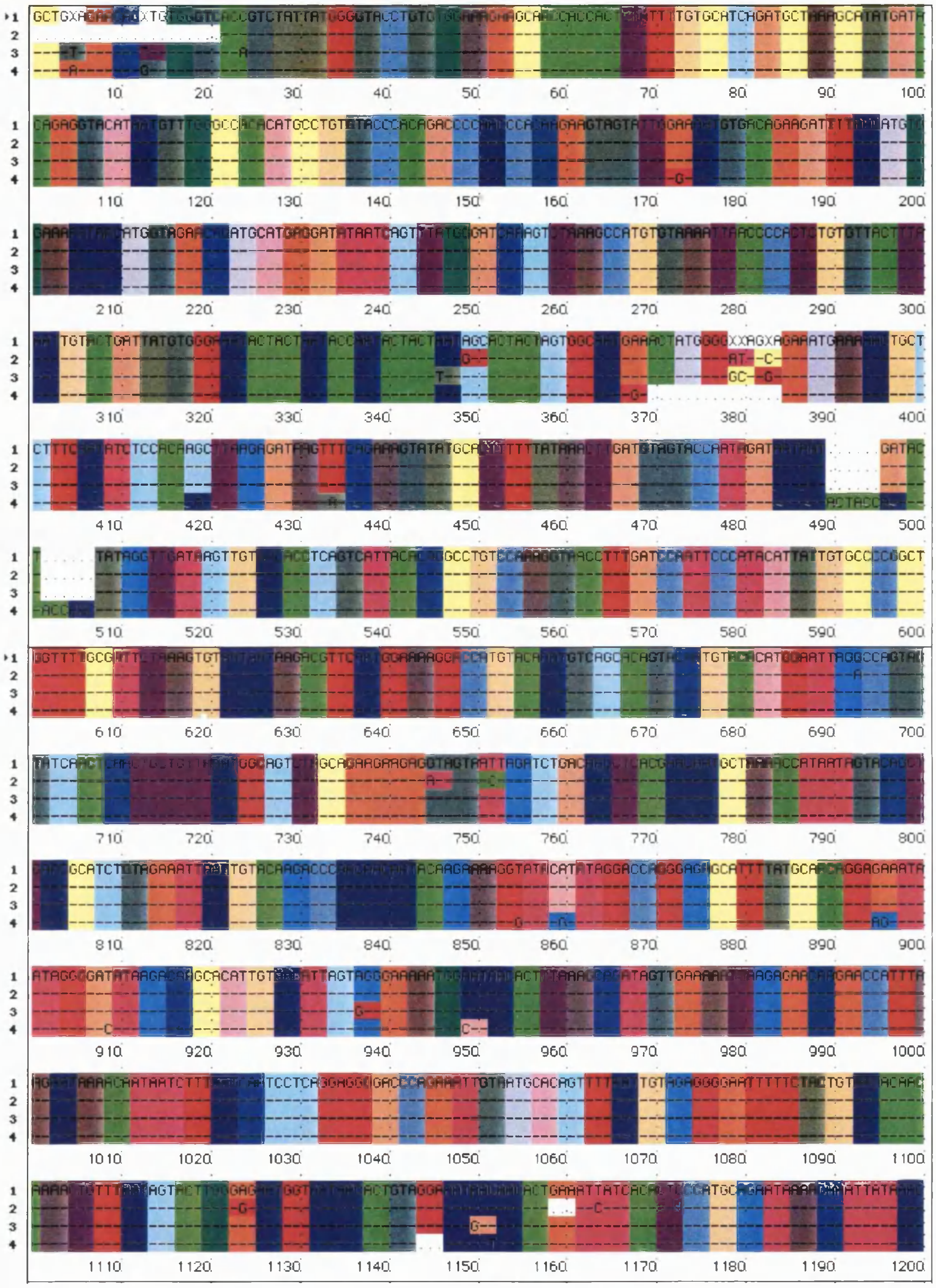
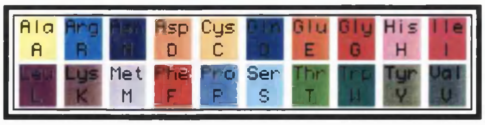
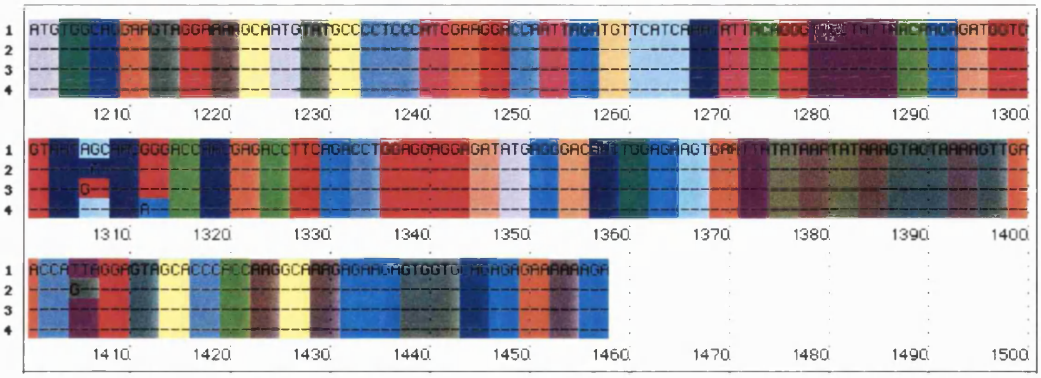


Figure 5.3 Amino acid alignment of CPHL3 sequences. Other details as for figure 5.1.



1 = CPHL3 consensus
2 = CPHL3. 5
3 = CPHL3. 6
4 = CPHL3. 48



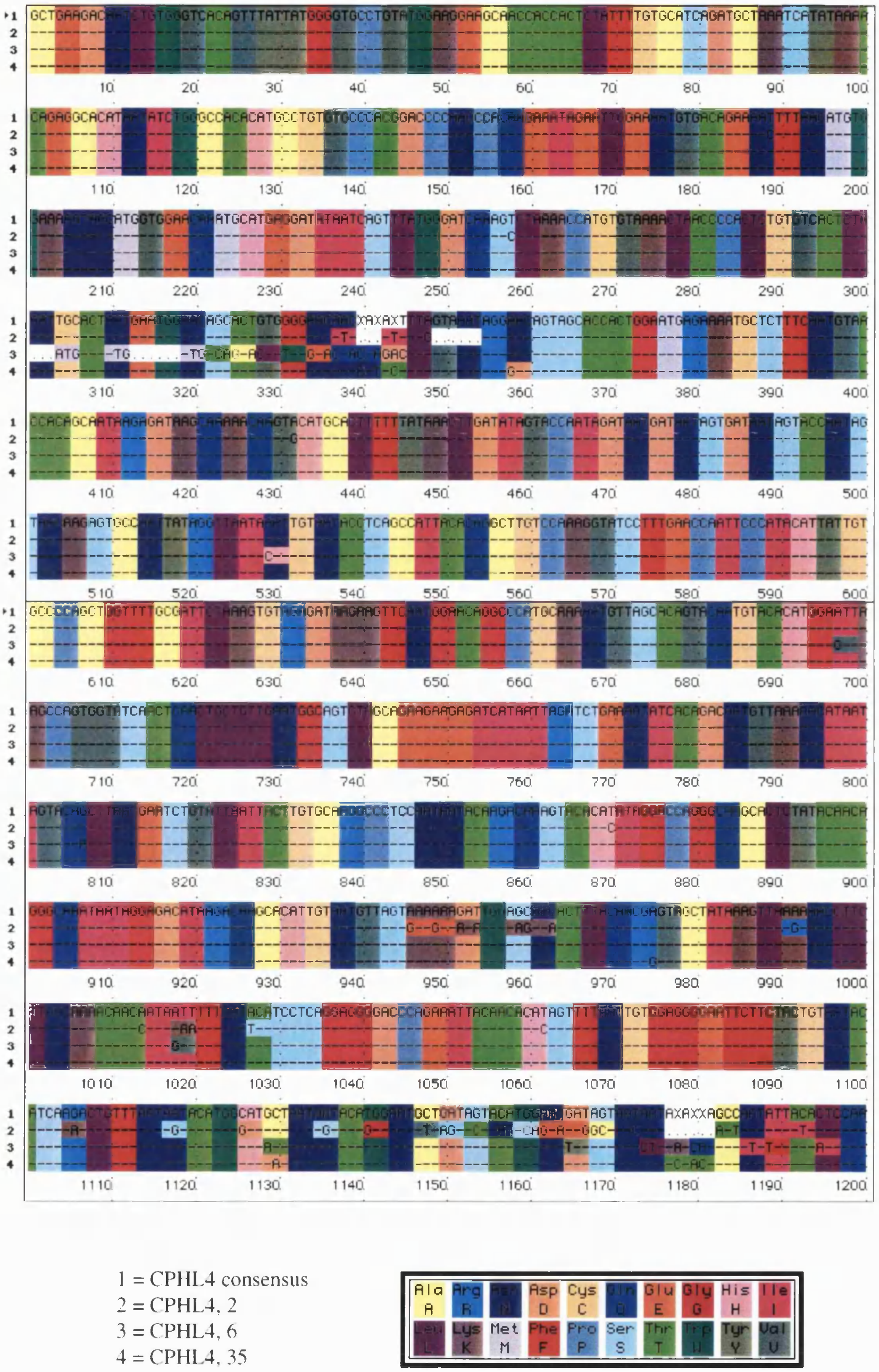


1 = CPHL3 consensus
2 = CPHL3, 5
3 = CPHL3, 6
4 = CPHL3, 48

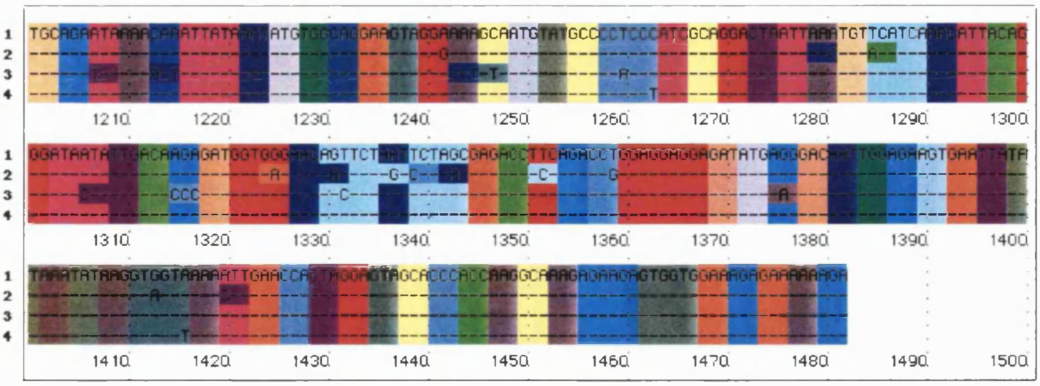
Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

Figure 5.4

Amino acid alignment of CPHL4 sequences. Other details as for figure 5.1.



1 = CPHL4 consensus
2 = CPHL4, 2
3 = CPHL4, 6
4 = CPHL4, 35



1 = CPHL4 consensus
2 = CPHL4, 2
3 = CPHL4, 6
4 = CPHL4, 35

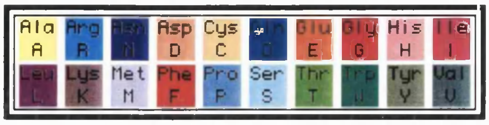
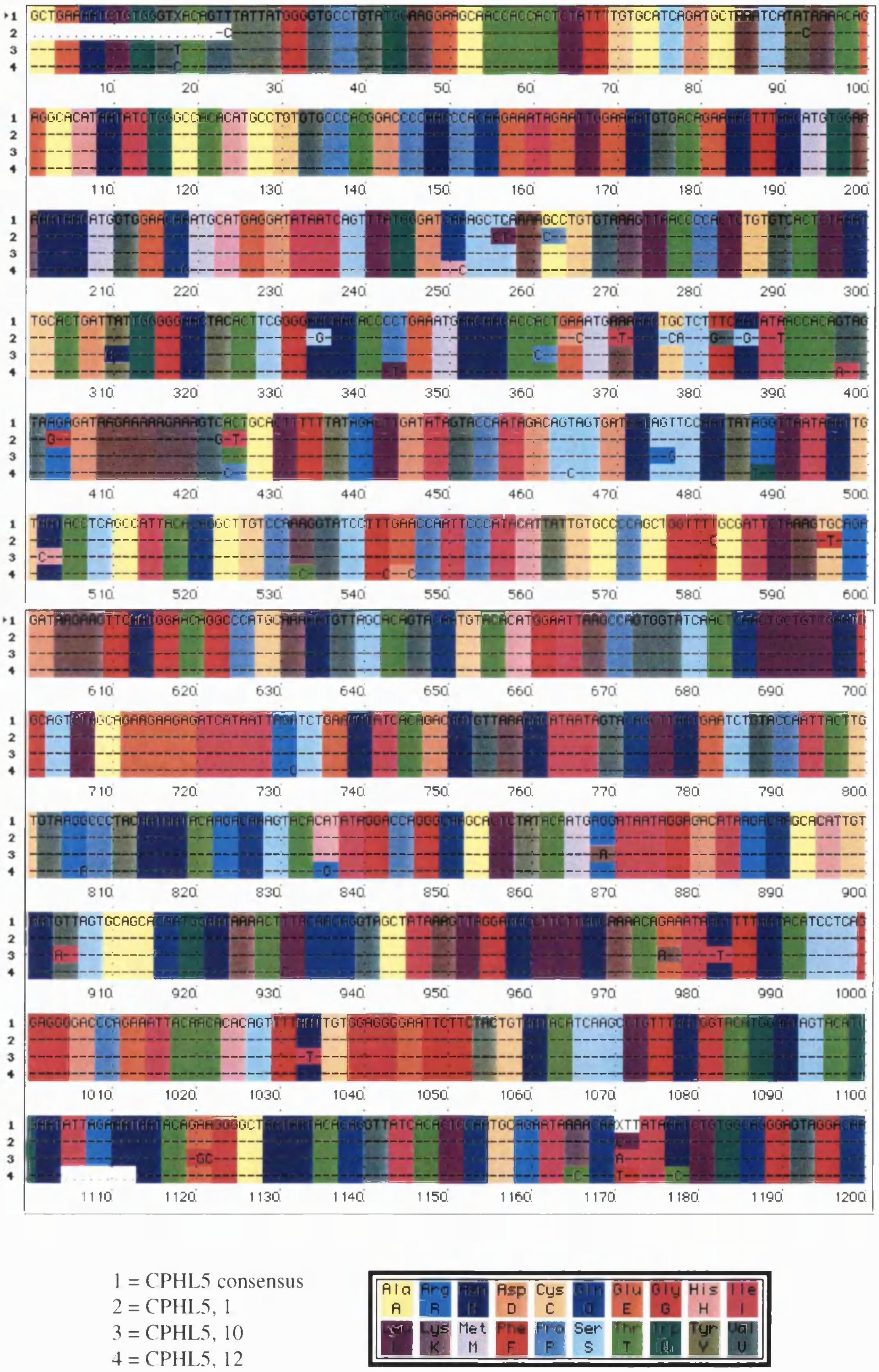
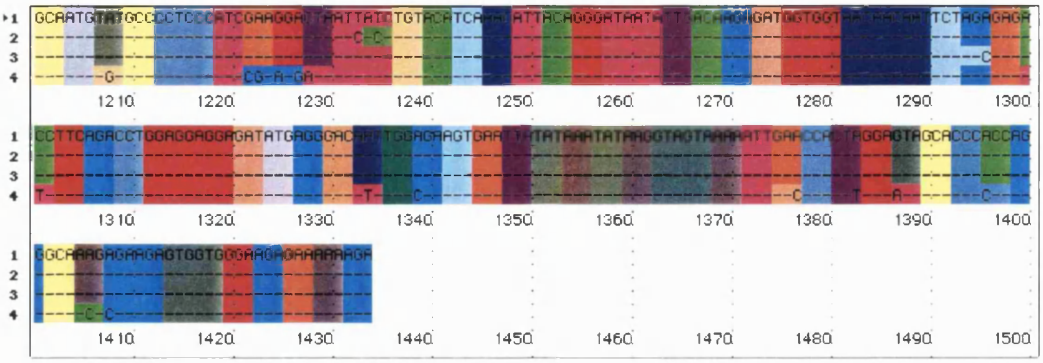


Figure 5.5

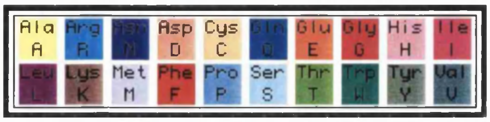
Amino acid alignment of CPHL5 sequences. Other details as for figure 5.1.



1 = CPHL5 consensus
2 = CPHL5, 1
3 = CPHL5, 10
4 = CPHL5, 12



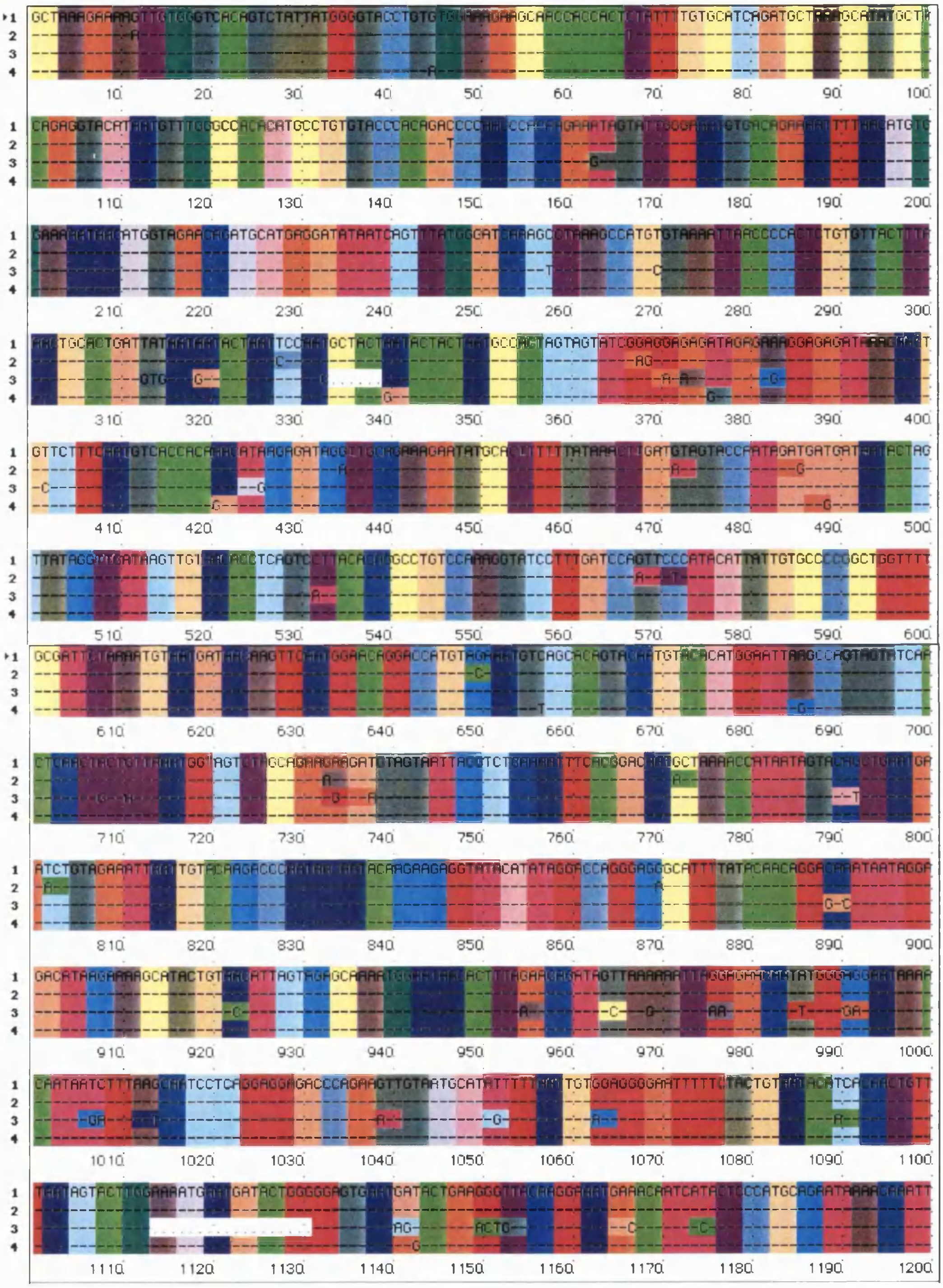
1 = CPHL5 consensus
2 = CPHL5, 1
3 = CPHL5, 10
4 = CPHL5, 12



Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

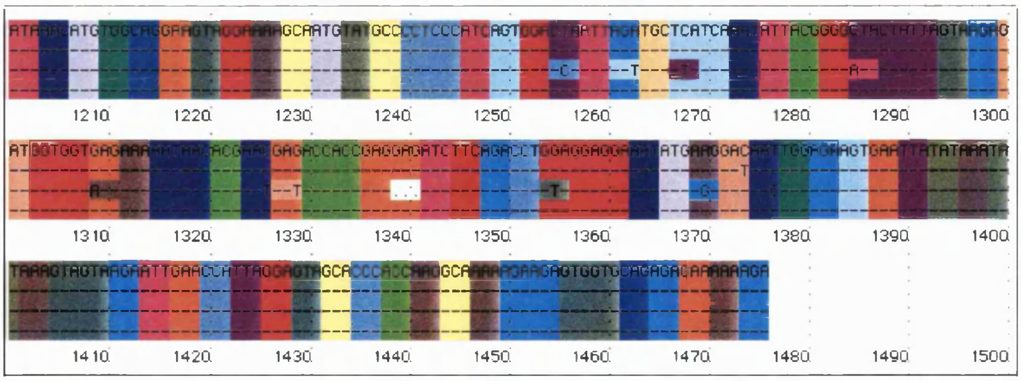
Figure 5.6

Amino acid alignment of CPHL6 sequences. Other details as for figure 5.1.



1 = CPHL6 consensus
2 = CPHL6, 3
3 = CPHL6, 6
4 = CPHL6, 41

Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

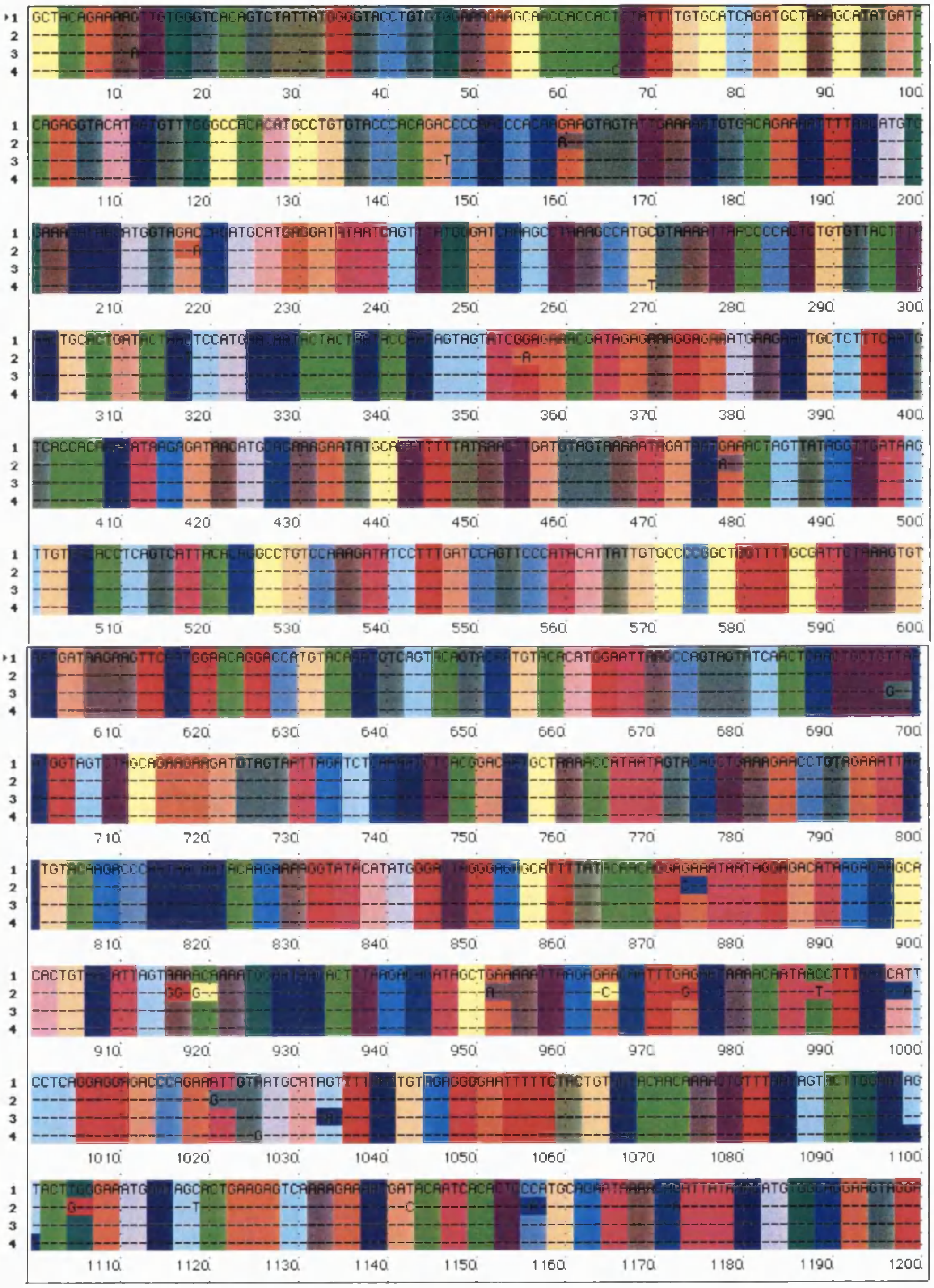


1 = CPHL6 consensus
2 = CPHL6, 3
3 = CPHL6, 6
4 = CPHL6, 41

Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

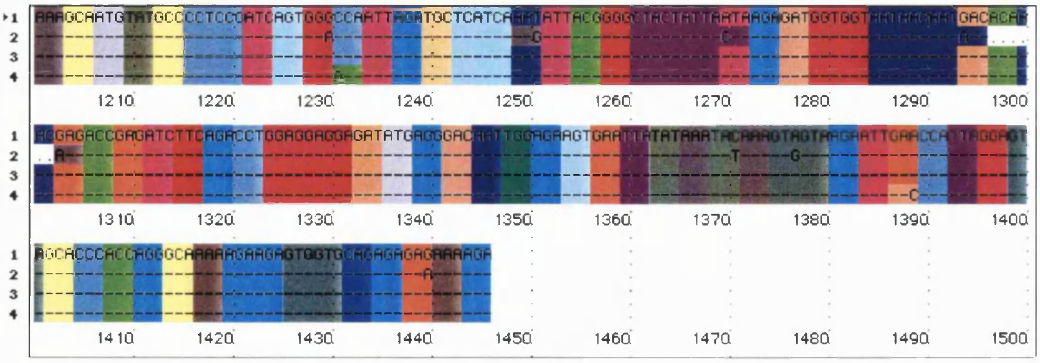
Figure 5.7

Amino acid alignment of CPHL9 sequences. Other details as for figure 5.1.



1 = CPHL9 consensus
2 = CPHL9, 1
3 = CPHL9, 6
4 = CPHL9, 17

Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

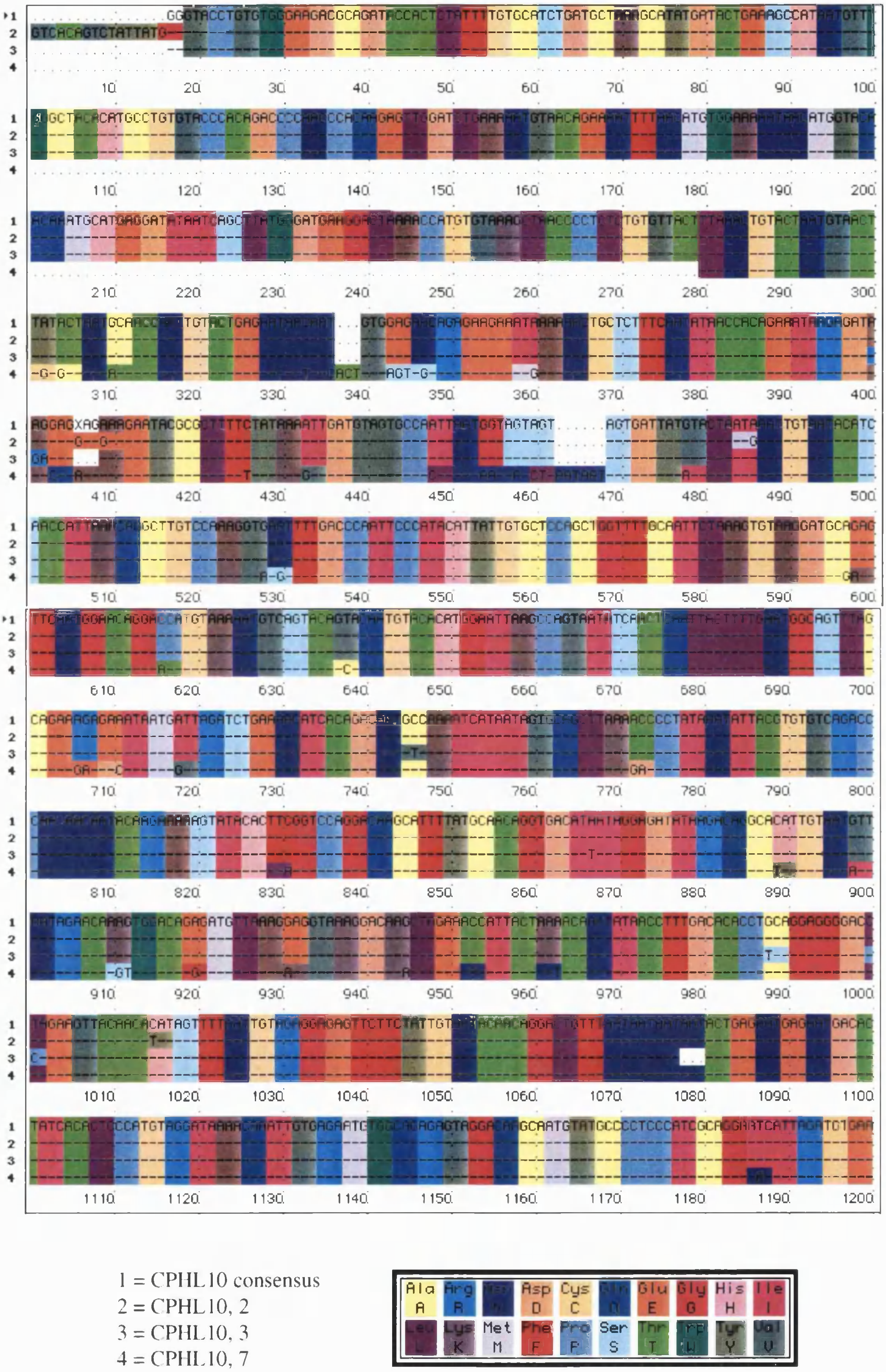

1 AAAGCAATGTATGCCCCTCCCATCAGTGGGCCAATTAGATGCTCATCAAATATTACGGGGCTACTATTAATAAGAAGATGGTGGTAATAACAATGACACAA
2 A G C A
1210 1220 1230 1240 1250 1260 1270 1280 1290 1300
1 AGGAGACCGAAGATCTTCAGACCTGGAGGAGGAGATATGAGGGACAATTGGAGAAGTGAATTATATAAATACAAAGTAGTAAGAATTGAACCATTAGGAGT
2 A T G
4 C
1310 1320 1330 1340 1350 1360 1370 1380 1390 1400
1 AGCACCCACCAGGGCAAAAAGAAGAGTGGTGCAGAGAGAGAAAAGA
2 A
1410 1420 1430 1440 1450 1460 1470 1480 1490 1500

1 = CPHL9 consensus
2 = CPHL9, 1
3 = CPHL9, 6
4 = CPHL9, 17

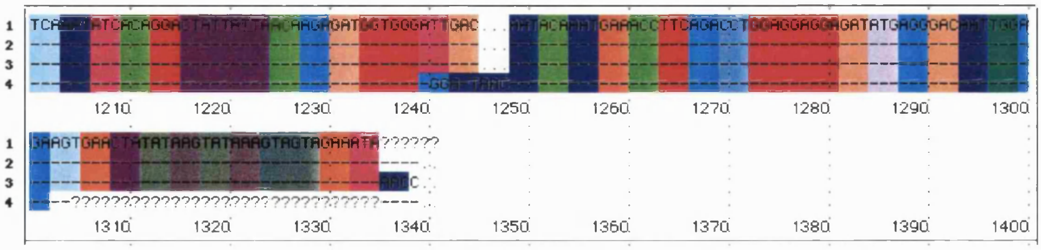
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
A	R	N	D	C	Q	E	G	H	I
Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
L	K	M	F	P	S	T	W	Y	V

Figure 5.8

Amino acid alignment of CPHL10 sequences. Other details as for figure 5.1.



1 = CPHL10 consensus
2 = CPHL10, 2
3 = CPHL10, 3
4 = CPHL10, 7

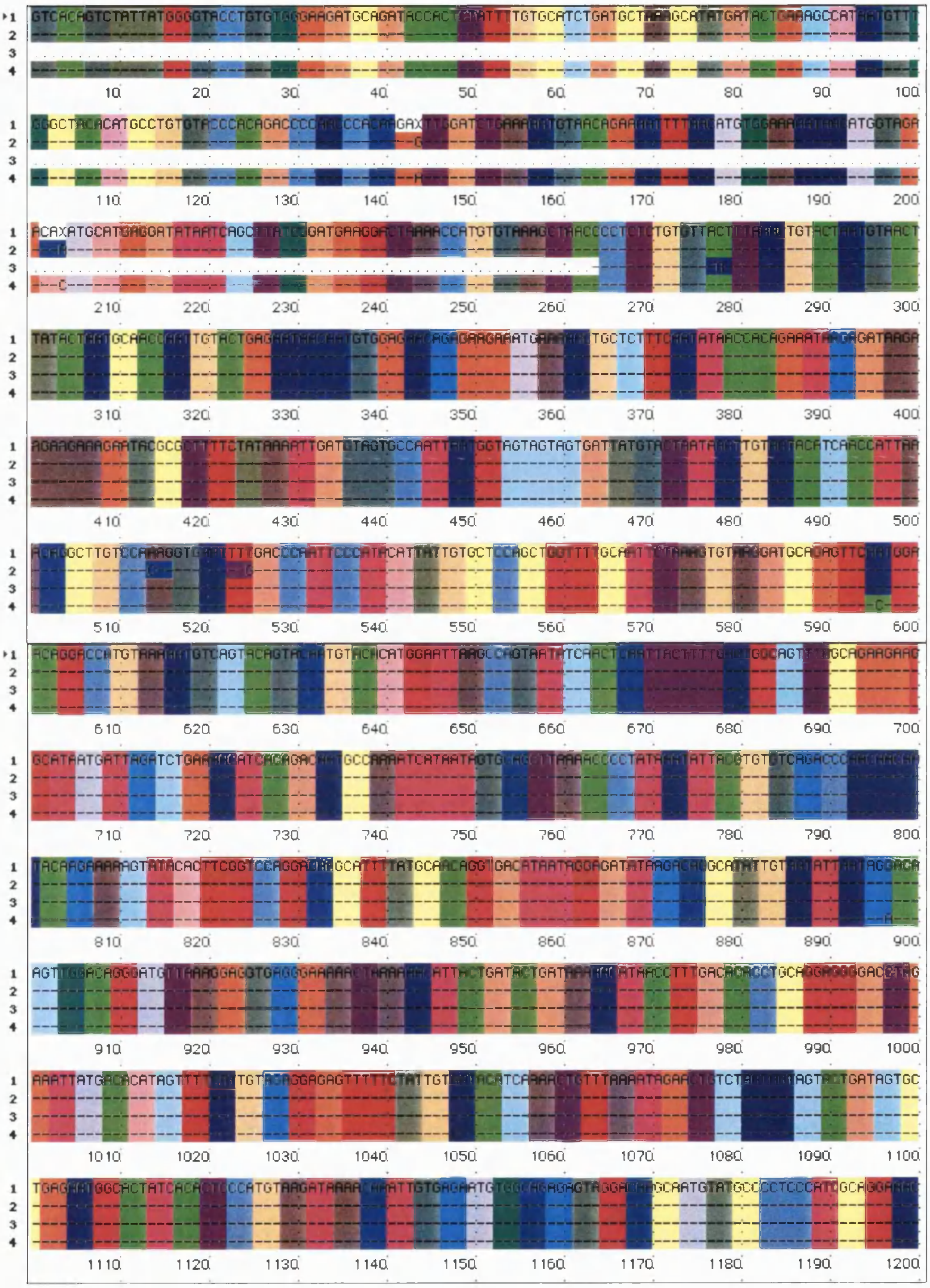

1 TCAAAAATCACAGGACTATTTATTAACAAGAGATGGTGGGATTGAC . . . AATACAAATGAAACCTTCAGACCTGGAGGAGGAGATATGAGGGACAATTGGA
2
3
4 GGA TAAC
1210
1220
1230
1240
1250
1260
1270
1280
1290
1300
1 GAAGTGAACTATATAAGTATAAAGTAGTAGAAATA??????
2
3 AAAC
4 ??????????????????????????????????
1310
1320
1330
1340
1350
1360
1370
1380
1390
1400

1 = CPHL10 consensus
2 = CPHL10, 2
3 = CPHL10, 3
4 = CPHL10, 7

Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

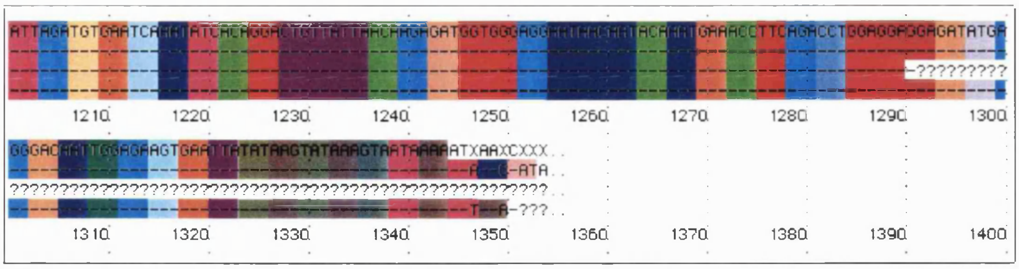
Figure 5.9

Amino acid alignment of CPHL11 sequences. Other details as for figure 5.1.



1 = CPHL11 consensus
2 = CPHL11, 1
3 = CPHL11, 2
4 = CPHL11, 5

Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
A	R	N	D	C	Q	E	G	H	I
Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
L	K	M	F	P	S	T	W	Y	V



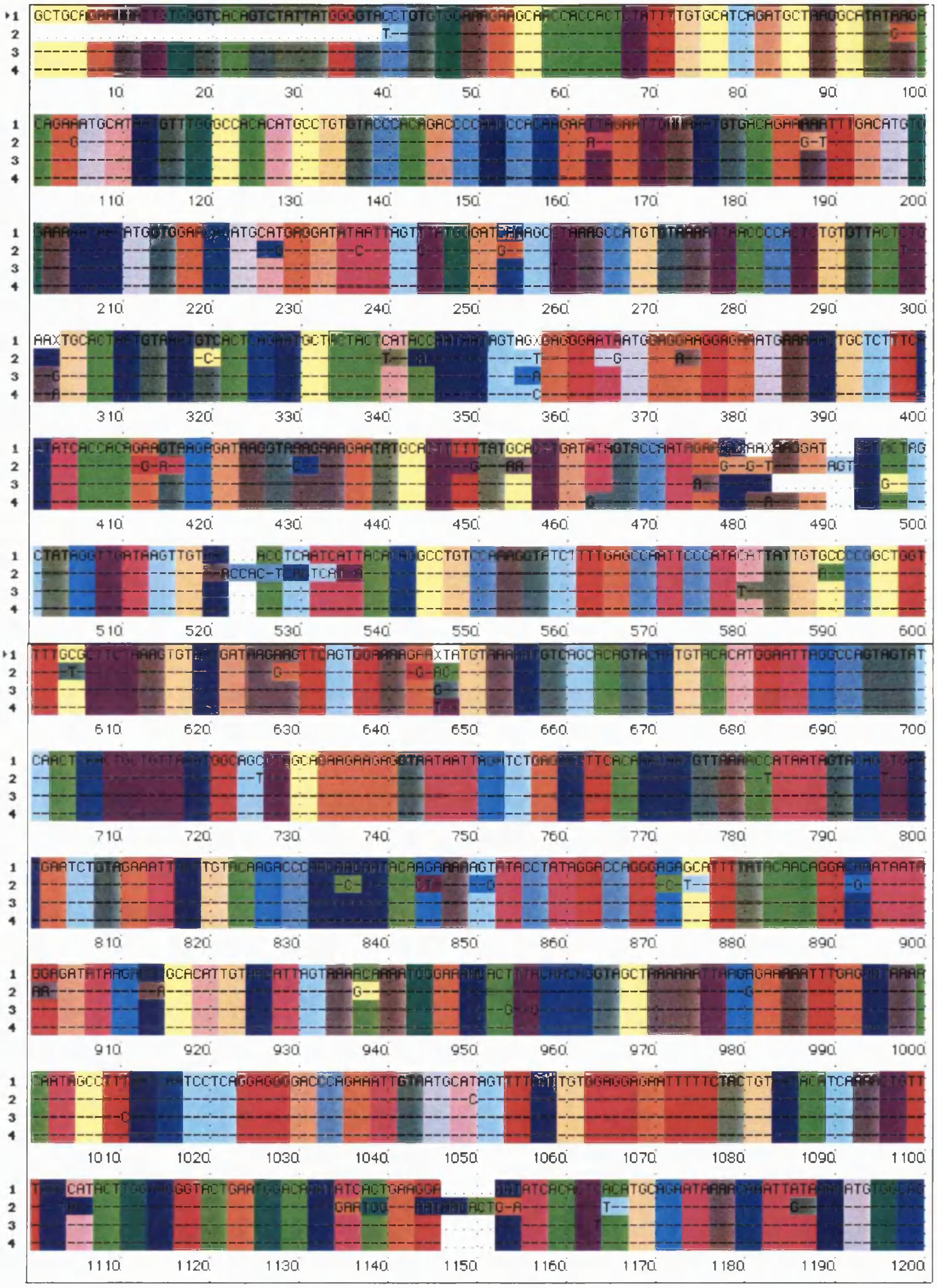
```
1 ATTAGATGTGAATCAAATATCACAGGACTGTTATTAACAAGAGATGGTGGGAGGAATAACAATACAAATGAAACCTTCAGACCTGGAGGAGGAGATATGA
2 ----------------------------------------------------------------------------------------------------
3 -------------------------------------------------------------------------------------------?????????
4 ----------------------------------------------------------------------------------------------------
          1210      1220      1230      1240      1250      1260      1270      1280      1290      1300

1 GGACAATTGGAGAAGTGAATTATATAAGTATAAAGTAATAAAAATXAAXCXXX..
2 ---------------------------------------------A--C-ATA..
3 ?????????????????????????????????????????????????????..
4 ---------------------------------------------T--A-???..
          1310      1320      1330      1340      1350      1360      1370      1380      1390      1400
```

1 = CPHL11 consensus
2 = CPHL11, 1
3 = CPHL11, 2
4 = CPHL11, 5

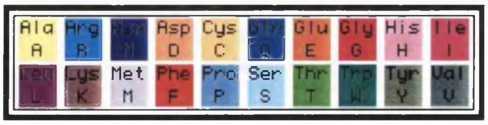
Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

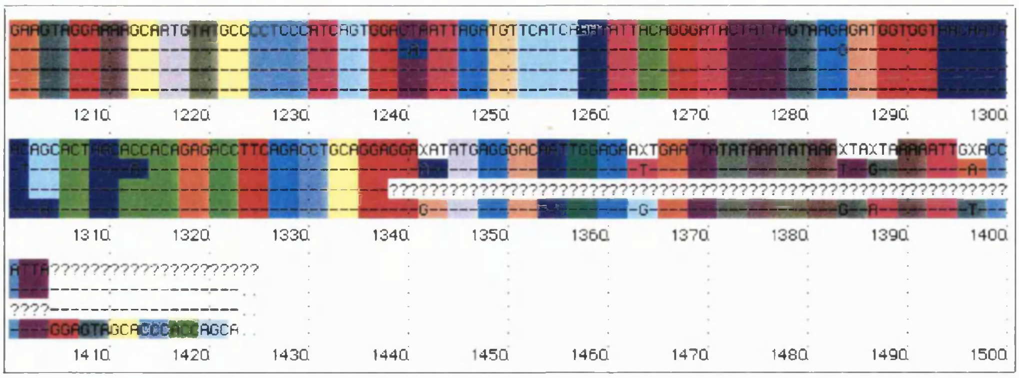
Figure 5.10

Amino acid alignment of CPHL14 sequences. Other details as for figure 5.1.



1 = CPHL14 consensus
2 = CPHL14, 15
3 = CPHL14, 37
4 = CPHL14, 44


Ala A
Arg R
Asn N
Asp D
Cys C
Gln Q
Glu E
Gly G
His H
Ile I
Leu L
Lys K
Met M
Phe F
Pro P
Ser S
Thr T
Trp W
Tyr Y
Val V



1 = CPHL14 consensus
2 = CPHL14, 15
3 = CPHL14, 37
4 = CPHL14, 44

Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I
Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

Table 5.1

Synonymous/nonsynonymous (d_s/d_n) and transition/transversion ratios (T_s/T_v) ratios calculated from multiple single molecules

CPHL No.	Disease stage	d_s / d_n	T_s / T_v
9	AIDS	3.90	2.36
1	AIDS (near death)	3.78	1.6
14	AIDS	3.99	2.12
4	n/k	3.90	1.68
6	AIDS	3.82	2.15
5	n/k	3.75	0.75
10	n/k	3.92	2.45
2	asymptomatic	3.90	1.52
3	asymptomatic	3.76	2.83

Mean d_s / d_n for all sequences above = 3.86

d_s / d_n ratio calculated using DNA Translator (298) with the unweighted pathway method of Nei and Gojobori (Gojobori *et al.* 1990. Meth. Enzym. 183:531-550).

Mean T_s / T_v for all sequences above = 1.94 (±0.59)0.41 (±0.14)

T_s / T_v ratio calculated using QPUZZLE (Strimmer and Haeseler, 1996, in press) = 1.96

n/k = not known

Table 5.2

Nucleotide substitution table for CPHL1 sequences relative to their consensus

	molecule no.							
Substitution	**1**	**4**	**7**	**18**	**19**	**43**	**cDNA**	**Total**
A>G	10	4	2	4	6	7	2	35
G>A	3	3	3	6	7	11	4	37
C>T	3	0	1	0	3	1	1	9
T>C	4	1	1	0	1	0	3	15
A>C	2	2	1	1	1	2	1	10
A>T	2	1	1	1	0	2	1	8
G>C	0	0	0	0	0	1	0	1
G>T	1	0	0	0	0	1	0	2
C>A	5	1	3	1	3	1	4	18
C>G	3	0	0	0	1	0	0	4
T>A	2	1	1	1	2	0	1	8
T>G	1	1	2	0	1	0	1	6

Discussion

The ratio of synonymous and nonsynonymous nucleotide substitutions gives information about the selection pressures on HIV-1 variants present in an individual at a given time. Varying selection pressure is consistent with the theory that the immune responses select for viral diversity (195, 286). In the first stages of infection the intact immune system will respond against dominant viral variants, favouring less common variants and thus providing strong selection pressure for diversification. As disease progresses, the immune system declines and these selection pressures become weaker. The work of Boyd *et al.* suggests that most viral diversity in the V3 region is caused by viral adaptation for various cell tropisms (23). Diversification of the initially homogeneous viral population occurs as HIV infects different cell types, e.g. initially macrophages and T cells when infection is under control and patients are asymptomatic followed by new variants that target only T cells when the virus begins to defeat the immune system, providing strong positive selection which declines once many variants with specific cell tropisms are generated. This lethal transformation that HIV-1 undergoes before it causes AIDS may be due to a change in the preference from one co-receptor to another. Fusin has recently been described as a second co-receptor by Feng *et al.* (68). T cells express fusin on their surface as well as CD4$^+$. Mutation of the virus to interact with fusin/CD4$^+$ in preference to a still-undiscovered co-receptor/ CD4$^+$ found on the surface of macrophages, may be part of the reason the immune system is defeated.

If the number of nucleotide differences between two sequences is small it is relatively simple to obtain the number of synonymous and nonsynonymous nucleotide differences by counting. However, when two or three nucleotide differences exist between corresponding codons of the sequences, the distinction between synonymous and nonsynonymous substitutions must be inferred using appropriate statistical methods. There are two different types of methods available for the purpose of estimating the substitution numbers: the unweighted and weighted pathway methods (91). The basic difference between these two methods is that in the unweighted pathway method an equal weight is given to two or more alternative evolutionary pathways, whereas in the weighted pathway method a greater weight is given to an evolutionary pathway involving synonymous substitutions than to one involving nonsynonymous substitutions. However, synonymous and nonsynonymous changes were relatively easy to calculate using sequence data obtained in this study as the majority of differences between two codons were effected by only one nucleotide substitution (e.g. see figure 5.1).Those few cases effected by more than one nucleotide substitution were discounted. Table 5.1 shows the calculated d_s /d_n and T_s / T_v ratios for some CPHL sequences. It is interesting to note that samples from individuals at a more advanced disease stage appear to have higher d_s / d_n ratios than asymptomatic individuals, in agreement with the observations of Bonhoeffer *et al.* (21). Also, the amino acid and nucleotide trees containing all CPHL sequences compared in chapter 3 would have different topologies if the true d_s / d_n ratios were very different from the expected or estimated values. The mean T_s / T_v ratio for 9 CPHL sequences and the computed T_s / T_v ratio for all CPHL sequences are in close agreement at 1.94 and 1.96 respectively, indicating that the default value of 2 used in PHYLIP programs for data analysed in this thesis should give accurate phylogenetic trees (Table 5.1).

The nucleotide substitution table shown in Figure 5.2 indicates that G>A and A>G (purine > purine) transitions were by far the most common substitution for this data set. Most of the nucleotide substitutions for molecule number 43 were nonsynonymous G>A transitions, suggesting possible G>A hypermutation. This phenomenon is thought to be an example of induced mutation, whereby the reverse transcriptase is forced into making errors by imbalances in the intracellular dCTP concentration (76, 268, 269).

Prediction of the specific effect of amino acid changes on the structure or function of a protein is always difficult, especially in the case of larger proteins where effect and counter effect operate throughout the molecule. Of course, changes such as the loss or gain of asparagine residues will affect the glycosylation arrangement, changing the shape of the 'cloud' of sugars which partially dictate the contours of the molecule's surface, though this cannot be estimated accurately. Also, the precise pattern of branching of these glycosylation sites cannot be predicted exactly (determined in the rough endoplasmic reticulum), further confounding modelling of these molecules. Loss or gain of cysteine residues is likely to have a profound, and perhaps a slightly more predictable effect on the shape of the protein. Figure 5.11 shows a schematic representation of gp120 HIV-1_{LAI}, showing disulphide bridges between cysteine residues and N-linked glycosylation sites.

Figure 5.11

Schematic representation of gp120 HIV-1_{LAI}, taken from (156). Glycosylation sites containing high-mannose and/or hybrid type oligosaccharides are indicated (green) as well as the glycosylation sites with complex-type oligosaccharide structures (blue). Variable regions are shown in red.

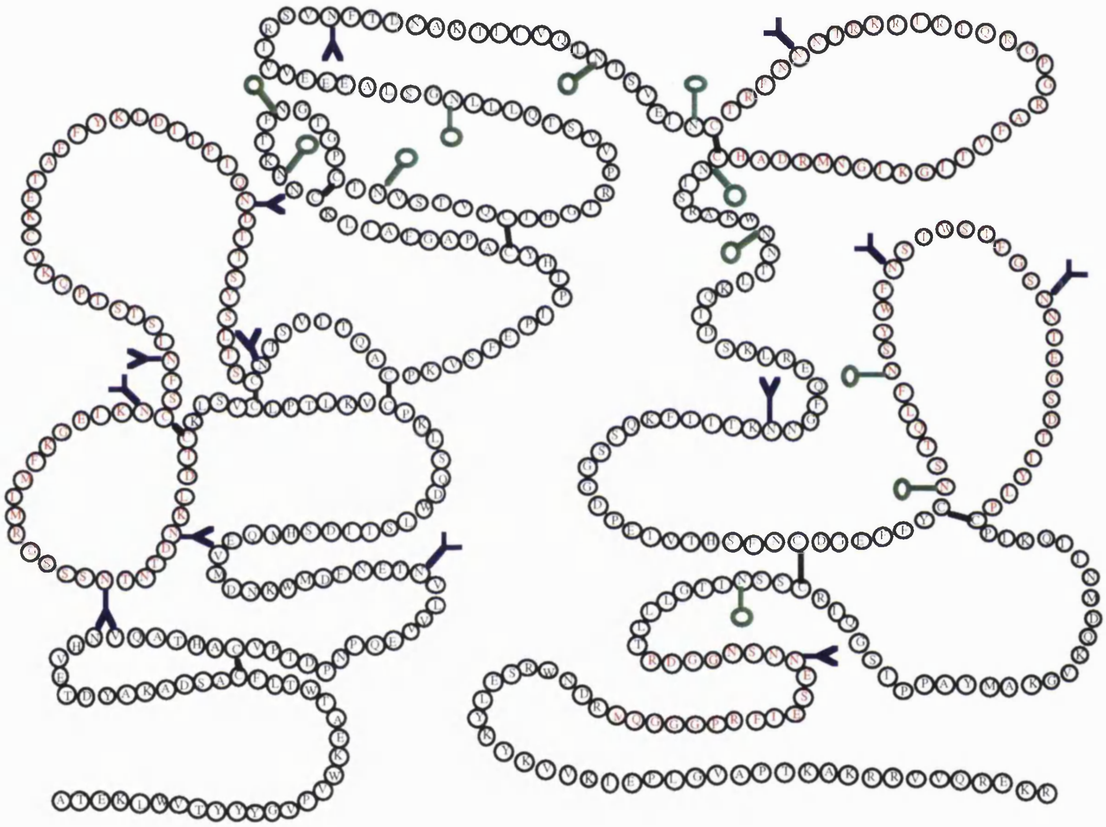


Figure 5.1 suggests that for CPHL1 molecule 1 two cysteine residues have arisen in V1, at nucleotide positions 325-7 and 355-7. These changes would most likely effect the conformation of the V1 loop severely. CPHL10, molecule 4 (Figure 5.9), according to the derived amino acid translation, also has an additional cysteine in C1 at nucleotide positions 301-3. CPHL5, molecule 1 (Figure 5.5) has lost 2 cysteine residues, one in V1 at nucleotide positions 376-8 and the other in C2, at nucleotide positions 596-8. CPHL 5, molecule 4 (Figure 5.5) has an additional cysteine, at nucleotide 1207-9, directly in the middle of the CD4 binding site in C4, which is likely to have a pronounced effect on the conformation of the protein and on the CD4 binding capability.

Although relatively similar patterns of substitution were seen for the 9 data sets analysed here, some groups of sequences were more homogeneous than others and yielded less information. For this type of analysis it is probably better to use more than 3 sequences from each sample to obtain more usable information. In this work most data was obtained from CPHL1 (7 sequences).

The exact pattern of nucleotide substitutions is still is still unclear for HIV-1. To date, none of the methods for phylogenetic analysis use models which take into account such factors as natural selection, the bias in base substitution in HIV and the tendency for G to A changes (183). Considering the great importance of the estimation of sequence divergence for studies of molecular evolution, further efforts are required to develop better statistical methods in this field.

ii) A summary of gp120 sequencing work carried out in collaboration with gp120 protein and monoclonal antibody characterisation

a) The work described in this section has been published by Shotton *et al.* (234).

Characterisation of polymorphism in T-cell line adapted virus

Background

Neutralising monoclonal antibodies (mAbs) mapping to the V2 region of HIV-1 gp120 have been reported (84, 112, 174, 273), implying that this variable loop may also have a functional role in the entry of viruses into target cells. Single-amino-acid changes in the V2 region have been reported to affect both the fusogenicity and the gp120-gp41 association (254). Also, the length of the carboxy side of the V2 loop has been correlated with a change from SI to NSI isolates (98). Shotten *et al.* identified six monoclonal antibodies mapping to both linear and conformation-dependent epitopes within the V2 region of HIV-1 clone HXB10 (234). Three of the mAbs raised neutralised HXB10 infectivity and epitope mapping of the mAbs identified three mAbs mapping to the crown of the V2 loop, all of which exhibited poor binding to cell surface-expressed gp120. However, the same mAbs bound well to recombinant gp120, which implied differential epitope exposure between recombinant monomeric gp120 and the oligomeric gp120 in the virion. As the V2 mAbs raised previously had failed to neutralise nonclonal preparations of virus, they cloned and expressed both gp120 and V1V2 from IIIB-, MN-, and RF- infected H9 cultures. The results of this work suggested that there were a number of polymorphic sites within V2 which affected mAb recognition, implying that after long-term passage *in vitro* the V2 domain is more variable than the V3 domain, as the V3 region from the clones of the same isolates showed only one, antigenically silent amino acid change. The polymorphism of the original isolate may account for the reduced antibody recognition observed.

The IIIB clones were generated by P.Balfe at UCL Dept of Virology. Briefly, purified DNA from the laboratory isolate IIIB was PCR amplified using primers containing the BstEII and MluI sites (see Figure 2.1). The PCR amplicons were then digested with BstEII and MluI restriction endonucleases, combined with cut vector pEE14, ligated, plated out and screened for insert by PCR. The clones were sequenced using the methods described in section 2.3.10.

Results

All sixteen gp120 clones were found to contain unique sequences, with an average pairwise distance of 2.3%. All reading frames were complete and no length polymorphism seen. A summary of the amino acid changes in the sequences when compared with the standard HXB2 sequence taken from the database is shown in table 5.3. As well as the amino acid changes shown in table 5.3 there were 11 silent substitutions, of which 6 were present in more than one clone. Sequence positions marked with an asterisk are associated with glycosylation site polymorphisms.

Table 5.3 Amino acid changes observed in 16 IIIB clones

IIIB clone number

Polymorphic sites (aa number)

	69	93	108	118	121	163	164	165	177	193	204	229	232	236	275	279	287	290	306	325	326
con	W	F	I	P	K	T	S	K	Y	L	A	N	T*	T*	V	D	Q	T	K	N	M
1	.	S	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	.	V	L	.	.	.	.	.	M	.	.	.	.	.	.	R	T	.	.	.
11	.	.	.	.	R	.	.	I	H	.	G	.	.	A	A	.	.	Q	R	S	.
16	.	.	.	.	.	.	.	I	.	.	.	.	.	.	A	.	.	Q	R	.	.
18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
34	.	.	.	.	.	.	.	.	.	.	.	.	.	I	.	.	.	.	.	.	.
36	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
38	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
42	.	.	.	.	.	A	.	I	.	.	.	.	.	.	A	.	.	Q	R	.	.
44	.	.	.	.	.	.	.	.	.	.	.	S	A	.	.	.	.	.	.	.	.
47	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T
51	.	.	.	.	I	.	.	.	.	.	.	.	.	.	A	.	.	.	R	.	.
52	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
56	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HXB2	.	.	.	.	I	.	.	.	.	.	.	.	.	.	.	.	.	.	R	.	.

IIIB clone number

Polymorphic sites (aa number)

	330	338	340	353	361	364	376	386	397	407	412	423	427	448	449	453	461	464	468	469	476
con	H	W	A	F	F	S	F	N*	N*	N	D	I	W	N*	I*	L	S	E	F	R	R
1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	R	.	.	.	.	.	S	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	D	.	.	F	.	.	.	.	N	.	.	.	K
16	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
18	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	N	G	.	.	.
20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	N	G	.	.	.
21	.	.	.	.	.	.	.	.	.	.	.	.	.	I	T	.	.	.	.	.	.
34	.	.	.	.	S	.	S	.	.	.	G	.	.	.	.	.	.	.	.	.	.
36	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
38	.	.	.	.	.	.	.	.	.	S	.	.	.	.	.	.	.	.	.	G	.
42	.	.	N	.	.	P	.	.	.	.	.	.	.	.	.	.	N	G	.	.	.
44	.	.	.	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.
47	.	.	.	L	.	.	.	.	.	.	.	.	.	.	.	P	.	.	.	.	.
51	R	.	.	.	.	.	.	S	.	.	.	.	.	.	.	.	.	.	S	.	.
52	.	.	.	.	.	.	.	.	.	.	.	.	R	.	.	.	N	G	.	.	.
56	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HXB2	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Discussion

The sequencing results show that the assumption that laboratory-adapted isolates are homogeneous is false. Each of the 16 clones sequenced was different. Also, the expression of the gp120 proteins of these clones and the subsequent monoclonal antibody characterisation carried out by Shotton *et al.* at Reading University (234) indicated that laboratory-adapted isolates of HIV are a mixture of antigenic variants, and suggest that multiple envelope clones must be sequenced before the diversity remaining even after long-term propagation can be appreciated. In a control experiment, 28 envelope genes were generated by PCR from the same starting copy number of pHXB2 plasmid, in this case no polymorphism was seen (A. Hammond, personal communication).

N.B. Pages 222-225 inclusive removed as requested by examiners.

General Summary

The comparison of HIV-1 sequence data has been important for elucidating phylogenetic relationships between HIV-1 strains taken from related and unrelated infections. How these relationships between different strains are arrived at has been the subject of much discussion. It is essential to use the most appropriate methods of generating and analysing sequence data to enable the correct assumptions concerning interrelationships to be made. The first major transmission investigation, the well-publicised Florida dentist case, compared the subfragment C2V3 of gp120 of the donor and potential recipients of the virus in order to determine any phylogenetic associations between the sequences. The choice of this fragment was initially criticised for being too small and too variable for this kind of analysis. Therefore, subsequent transmission studies used different regions of the HIV-1 genome, including both *gag* and *pol*, and involved differing methods of phylogenetic analysis. In order to determine which region of gp120 is the best to use, part of the work of this thesis examined the use of subfragments of gp120 compared with full gp120 for phylogenetic analysis. The most robust data set was generated using whole gp120, i.e. this region gave the highest percentage occurrence of known subtype branches (bootstrap values). Smaller regions of gp120, especially those with a low percentage (below 60%) of nucleotides in the constant regions do not give accurate subtyping results when analysed phylogenetically. However, the 5' 400 region, with approximately 60% of nucleotides in constant regions of gp120 gives bootstrap values as low as 40%, indicating that some variable regions are more variable than others, or that the smaller sequence regions analysed in some instances are too small for reliable conclusions to be made. The use of whole gp120 sequences in this study has also identified several possible *env* recombinants (A/G and A/D). The study of subsets of gp120 may not identify this type of variant.

HIV-1 transmission investigations have used various methods to generate sequence data from infected individuals. This thesis has shown that potential differences between viral RNA and proviral DNA are probably not significant with respect to phylogenetic analysis if clinical samples from the individuals involved are taken relatively soon after the transmission event. The use of sequence data generated from single molecules, although time consuming, is important for clean raw data when using the ABi automated sequencer. Also length polymorphism, especially in *env*, can produce unclear information. This thesis has shown that sequencing single molecules of gp120 can also be used to give estimations of transition/transversion ratios and synonymous/nonsynonymous ratios for HIV-1 strains within individuals. This information is necessary for accurate modelling of HIV-1 base substitution patterns. The protein sequences derived from the nucleotide data generated for work described in this thesis were used to estimate base substitution rates for gp120. A transition/transversion rate of 1.96 was found for the 51 sequences analysed and the mean synonymous/non synonymous ratio for multiple sequences analysed was found to be 3.86 The question of which base substitution model (s) to use when analysing sequence information phylogenetically is an active area of research. This thesis has shown that distance/neighbour joining methods are suitable for larger data sets (i.e. whole gp120 sequences) for which maximum likelihood may, arguably, be the most reliable but is too computationally demanding for routine use. The analysis by distance/neighbour joining can be given statistical credence by bootstrapping and relative branch lengths.

Other regions of the HIV-1 genome can also be used for determining phylogenetic relationships, in particular of indirectly linked infections as they are less variable than *env* and distant linkages are more apparent. The present work has analysed the p6/protease regions as a potential alternative to non-*env* regions for phylogenetic analysis. The variability of the p6 region due to length polymorphisms may enable more closely linked phylogenetic relationships to be determined.

Finally, sequences derived for these phylogenetic studies have been used to give an indication of the subtype diversity of HIV-1 in England. Recent studies have shown that some non-B subtypes appear to be more easily transmitted heterosexually. The present work has shown that these subtypes have already entered the heterosexual population in England and may represent the start of a non-B subtype epidemic in this country.

In conclusion, using larger subfragments of HIV sequence data, such as full gp120, is more reliable than smaller subfragments for subtyping and transmission investigations. However, when gap stripping of sequence data is carried out in order to avoid length variation, smaller subfragments are reliable enough to infer phylogenetic relationships, as confirmed by Leitner *et al.* (302); but, the smaller regions studied for subtyping are less likely to show whether the genome is recombinant or not. To do this either multiple smaller sequences or complete genomes have to be determined.

Bibliography

1. **Ahmad, N., and S. Venkatesan.** 1988. Nef protein of HIV-1 is a transcriptional repressor of HIV-1 LTR. Science. **241**:1481-1485.

2. **Albert, J., B. Abrahamsson, K. Nagy, E. Aurelius, H. Gaines, G. Nyström, and E.M. Fenyö.** 1990. Rapid development of isolate-specific neutralising antibodies after primary HIV-1 infection and consequent emergence of virus variants which resist neutralisation by autologous sera. AIDS. **4**:107-112.

3. **Albert, J., G. Scarlatti, T. Leitner, M. Uhlén, and E.M. Fenyö.** 1994. Low HIV-1 pathogenicity correlates with quality of immune response and structure of envelope protein, p. 146. Prevention and treatment of AIDS. Wiley-Liss, Hilton Head Island, South Carolina.

4. **Albert, J., J. Wahlberg, T. Leitner, D. Escanilla, and M. Uhlen.** 1994. Analysis of a rape case by direct sequencing of the Human Immunodeficiency Virus type 1 *pol* and *gag* genes. J. Virol. **68**:5918-5924.

5. **Alizon, M., S. Wain-Hobson, L. Montagnier, and P. Sonigo.** 1986. Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. Cell. **46**:63-74.

6. **Allan, J.S., J.E. Coligan, F. Barin, M.F. McLane, J.G. Sodroski, C.A. Rosen, W.A. Haseltine, T.H. Lee, and M. Essex.** 1985. Major glycoprotein antigens that induce antibodies in AIDS patients are encoded by HTLV-III. Science. **228**:1091-1094.

7. **Anand, R., F. Siegal, C. Reed, T. Cheung, S. Forlenza, and J. Moore.** 1987. Non-cytocidal natural variants of human immunodeficiency virus isolated from AIDS patients with neurological disorders. Lancet. **ii**:234-238.

8. **Anon.** 1993. HIV transmission between children at home. Communicable Disease Report. **3**:237.

9. **Arnold, C., P. Balfe, and J.P. Clewley.** 1995. Sequence distances between *env* genes of HIV-1 from individuals infected from the same source: implications for the investigation of possible transmission events. Virology. **211**:198-203.

10. **Arnold, C., K.L. Barlow, S. Kaye, C. Loveday, P. Balfe, and J.P. Clewley.** 1995. HIV type 1 sequence subtype G transmission from mother to infant:

failure of variant sequence species to amplify in the Roche Amplicor test. AIDS Res. Hum. Retroviruses. **11**:999-1001.

11. **Arnold, C., K.L. Barlow, J.V. Parry, and J.P. Clewley.** 1995. At least five HIV-1 sequence subtypes (A, B, C, D, A/E) occur in England. AIDS Res. Hum. Retroviruses. **11**:427-429.

12. **Arthur, L.O., J.W. Bess, R.C. Sowder, R.E. Benveniste, D.L. Mann, J.C. Chermann, and L.E. Henderson.** 1992. Cellular proteins bound to immunodeficiency viruses: implications for pathogenesis and vaccines. Science. **258**:1935-1938.

13. **Asjo, B., J. Albert, A. Karlsson, L. Morfeldt-Mamson, G. Biberfeld, K. Lidman, and E.M. Fenyo.** 1986. Replicative properties of human immunodeficiency virus from patients with varying severity of HIV infection. Lancet. **ii**:660-662.

14. **Balfe, P., P. Simmonds, C.A. Ludlam, J. O. Bishop, and A.J. Leigh Brown.** 1990. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. J. Virol. **64**:6221-6233.

15. **Barlow, K.L., J.H.C. Tosswill, and J.P. Clewley.** 1995. Analysis and genotyping of PCR products of the Roche Amplicor HIV-1 kit. J. Virol. Meth. **52**:65-74.

16. **Barr, S.** 1996. The 1990 Florida dental investigation: is the case really closed? Ann. Intern. Med. **124**:250-254.

17. **Barre-Sinoussi, F., J.-C. Chermann, F. Rey, M.T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier.** 1983. Isolation of a T-lymphotropic retrovirus fom a patient at risk for acquired immune deficiency syndrome (AIDS). Science. **220**:868-871.

18. **Baskar, P.V., S.C. Ray, R. Rao, T.C. Quinn, J.E.K. Hildreth, and R.C. Bollinger.** 1994. Presence in India of HIV type 1 similar to North American strains. AIDS Res. Hum. Retroviruses. **8**:1039-1041.

19. **Baur, A., N. Schwarz, S. Ellinger, K. Korn, T. Harrer, K. Mang, and G. Jahn.** 1986. Continuous clearance of HIV in a vertically infected child. Lancet. **2**:1045.

20. **Berger, A.** 1996. HIV babies shrug off infection. New Scientist. **149**:8.

21. **Bonhoeffer, S., E.C. Holmes, and M.A. Nowak.** 1995. Causes of HIV diversity. Nature. **376**:125.

22. **Boone, L.R., and A.M. Skalka.** 1981. Viral DNA synthesised in vitro by avian retrovirus particles permeabilised with melittin. Evidence for strand displacement mechanism in plus-strand synthesis. J. Virol. **37**:117-126.

23. **Boyd, M.T., G.R. Simpson, A.J. Cann, M.A. Johnson, and R.A. Weiss.** 1993. A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. J. Virol. **67**:3649-3652.

24. **Briant, L., J. Puel, C.M. Wade, A.J. Leigh Brown, and M. Guyader.** 1995. Protecting HIV databases - reply. Nature. **378**:243-244.

25. **Briant, L., C.M. Wade, J. Puel, A.J. Leigh Brown, and M. Guyader.** 1995. Analysis of envelope sequence variants suggests multiple mechanisms of mother-to-child transmission of human immunodeficiency virus type 1. J. Virol. **69**:3778-3788.

26. **Brossard, Y., J.-T. Aubin, L. Mandelbrot, C. Bignozzi, D. Brand, A. Chaput, J. Roume, N. Mulliez, F. Mallet, H. Agut, F. Barin, C. Brechot, A. Goudeau, J.-M. Huraux, J. Barrat, P. Blot, J. Chavinie, N. Ciraru-Vigneron, P. Engelman, F. Herve, E. Papiernik, and R. Henrion.** 1995. Frequency of early *in utero* HIV-1 infection: a blind DNA polymerase chain reaction study on 100 fetal thymuses. AIDS. **9**:359-366.

27. **Brown, D.** 1996. The 1990 Florida dental investigation: theory and fact. Ann. Intern. Med. **124**:255-256.

28. **Brown, P.** 1995. Will the strain show in Bangkok? New Scientist. **145**:11-12.

29. **Brownstein, A., and L. Augustyniak.** 1992. HIV infection in two brothers receiving intravenous therapy for haemophilia. MMWR. **41**:228-231.

30. **Bryson, Y.J., S. Pang, L. S.Wei, R. Dickover, A. Diagne, and I.S. Y. Chen.** 1995. Clearance of HIV infection in a perinatally infected infant. N. Engl. J. Med. **332**:833-838.

31. **Burger, H., B. Weiser, K. Flaherty, J. Gulla, P.N. Nguyen, and R.A. Gibbs.** 1991. Evolution of human immunodeficiency virus type 1 nucleotide sequence diversity among close contacts. Proc. Natl. Acad. Sci. USA. **88**:11236-40.

32. **Callebaut, C., B. Krust, E. Jacotot, and A.G. Hovanessian.** 1993. T cell activation antigen, CD26, as a cofactor for entry of HIV in CD4+ cells. Science. **262**:2045-2050.

33. **Cariello, N.F., J.K. Scott, A.G. Kat, W.G. Thilly, and P. Keohavong.** 1988. Resolution of a missense mutant in human genomic DNA by denaturing gradient gel electrophoresis and direct sequencing using in vitro DNA amplification. Am. J. Hum. Gen. **42**:726-734.

34. **Castro, B.A., C. Cheng-Mayer, L.A. Evans, and J.A. Levy.** 1988. HIV heterogeneity and viral pathogenesis. AIDS. **2**:S17-S27.

35. **Chang, D.D., and P.A. Sharp.** 1989. Regulation by HIV Rev depends upon recognition of splice sites. Cell. **59**:789-795.

36. **Cheng-Mayer, C., J.M. Homsy, L.A. Evans, and J.A. Levy.** 1988. Identification of HIV subtypes with distinct patterns of sensitivity to serum neutralization. Proc. Natl. Acad. Sci. USA. **85**. :2815-2819.

37. **Cheng-Meyer, C., D. Seto, M. Tateno, and J.A. Levy.** 1988. Biologic features of HIV-1 that correlate with virulence in the host. Science. **240**:80-82.

38. **Chesebro, B., J. Nishio, S. Perryman, A. Cann, W. O'Brien, I.S. Chen, and K. Wehrly.** 1991. Identification of human immunodeficiency virus envelope gene sequences influencing viral entry into CD4-positive HeLa cells, T-leukemia cells, and macrophages. J. Virol. **65**:5782-5789.

39. **Choo, V.** 1995. Combination superior to zidovudine in Delta trial. Lancet. **346**:895.

40. **Clavel, F., D. Guetard, F. Brun-Vezinet, S. Chamaret, M-A. Rey, M.O. Santos-Ferreira, A.G. Laurent, C. Dauget, C. Katlàma, C. Rouzioux, D. Klatzmann, J.L. Champalimaud, and L. Montagnier.** 1986. Isolation of a new human retrovirus from West African patients with AIDS. Science. **233**:343-346.

41. **Clewley, J.P., C. Arnold, K.L. Barlow, P.R. Grant, and J.V. Parry.** 1996. Diverse HIV-1 genetic subtypes in U.K. Lancet. **347**:1487.

42. **Coffin, J.M.** 1979. Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. J. Gen. Virol. **42**:1-26.

43. **Coffin, J.M., A. Haase, J.A. Levy, L. Montagnier, S. Oroszlan, N. Teich, H. Temin, K. Toyoshima, H. Varmus, P. Vogt, and R. Weiss.** 1986.

Human immunodeficiency viruses. Science. **232**:697.

44. **Seligmann, M., D. A. Warrell, J-P. Aboulker, C. Carbon, J.H. Darbyshire, J. Dormont, E. Eschwege, D. J. Girling, D. R. James, J-P. Levy, T E.A. Peto, D. Schwarz, A.B. Stone, I.V.D. Weller, R. Withnall, K. Gelmon, E. Lafon, A.M. Swart, V.R. Aber, A.G. Babiker, S. Lhoro, A.J. Nunn, and M. Vray.** 1994. Concorde: MRC/ANRS randomised double blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. Lancet. **343**:871-881.

45. **Connor, R.I., K.B. Chen, S. Choe, and N.R. Landau.** 1995. Vpr is required for efficient replication of human immunodeficiency virus type 1 in mononuclear phagocytes. Virology. **206**:935-944.

46. **Cooper, D.A., J. Gold, P. Maclean, B. Donovan, R. Finlayson, T.G. Barnes, H.M. Michelmore, P. Brooke, and R. Penny.** 1985. Acute AIDS retrovirus infection: definition of a clinical illness associated with seroconversion. Lancet. **i**:537-540.

47. **Cullen, B.R., and W.C. Greene.** 1989. Regulatory pathways governing HIV-1 replication. Cell. **58**:423-426.

48. **Daar, E.S., X.L. Li, T. Moudgil, and D.D. Ho.** 1990. High concentrations of recombinant soluble CD4 are required to neutralize primary human immunodeficiency virus type 1 isolates. Proc. Natl. Acad. Sci. USA. **87**:6574-6578.

49. **Dalgleish, A.G., P.C. Beverley, P.R. Clapham, D.H. Crawford, M.F. Greaves, and R.A. Weiss.** 1984. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature. **312**:763-767.

50. **Dayton, A.I., J.G. Sodroski, C.A. Rosen, W.C. Goh, and W.A. Haseltine.** 1986. The trans-activator gene of the human T cell lymphotropic virus type III is required for replication. Cell. **44**:941-947.

51. **De Jong, J.J., A. De Ronde, W. Keulen, M. Tersmette, and J. Goudsmit.** 1992. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. J. Virol. **66**:6777-6780.

52. **Deacon, N.J., A. Tsykin, A. Solomon, K. Smith, M. Ludford-Menting, D.J. Hooker, D.A. McPhee, A.L. Greenway, A. Ellet, and C.**

Chatfield. 1995. Genomic structure of an attenuated quasispecies of HIV-1 from a blood transfusion donor and recipients. Science. **270**:988-991.

53. **DeBry, R.W.** 1993. Dental HIV transmission? Nature. **361**:691.

54. **Deen, K.C., J.S. McDougal, R. Inacker, G. Folena-Wasserman, J. Arthos, J. Rosenberg, P.J. Maddon, R. Axel, and R.W. Sweet.** 1988. A soluble form of CD4 (T4) protein inhibits AIDS virus infection. Nature. **331**:82-84.

55. **Delwart, E.L., B. Herring, A.G. Rodrigo, and J.I. Mullins.** 1995. Genetic subtyping of Human Immunodeficiency Virus using a heteroduplex mobility assay. PCR Methods and Applications. **4**:S202-216.

56. **Dickover, R., S. Herman, E. Garrity, L. von Seidlen, P. Boyer, and Y. Bryson.** 1995. Maternal HIV levels are directly related to perinatal transmission risk and significantly reduced by ZDV treatment, p. 233. HIV pathogenesis. Wiley-Liss, Keystone, Colorado.

57. **Dietrich, U., M. Grez, H. von Briesen, B. Panhans, M. Geissendorfer, H. Kuhnel, J. Maniar, G. Mahambre, W.B. Becker, M.L. Becker.** 1993. HIV-1 strains from India are highly divergent from prototypic African and US/European strains, but are linked to a South African isolate. AIDS. **7**:23-7.

58. **Dunn, D.T., C.D. Brandt, and A. Kirvine.** 1995. The sensitivity of HIV-1 DNA polymerase chain reaction in the neonatal period and the relative contributions of intra-uterine and intra-partum transmission. AIDS. **1995**:F7-F11.

59. **Dunn, D.T., M.L. Newell, A.E. Ades, and C.S. Peckham.** 1992. Risk of transmission of human immunodeficiency virus through breast feeding. Lancet. **340**:558-585.

60. **Eigen, M., J. McCaskill, and P. Schuster.** 1988. Molecular quasi-species. J. Phys. Chemis. **92**:6881-6891.

61. **Epstein, L.G., C. Kuiken, B.M. Blumberg, S. Hartman, L.R. Sharer, M. Clement, and J. Goudsmit.** 1991. HIV-1 V3 domain variation in brain and spleen of children with AIDS: tissue-specific evolution within host-determined quasispecies. Virology. **180**:583-90.

62. **Essex, M.** 1995. North South epidemics. Different subtypes of HIV-1, p. 14-19, TB and HIV, vol. 8.

63. **Evans, L.A., T.M. McHugh, D.P. Stites, and J.A. Levy.** 1987.

Differential ability of human immunodeficiency virus isolates to productively infect cells. J. Immunol. **138**:3415-3418.

64. **Fan, L., and K. Peden.** 1992. Cell-free transmission of Vif mutants of HIV-1. Virology. **190**:19-29.

65. **Fauci, A. S.** 1988. The human immunodeficiency virus: infectivity and mechanisms of pathogenesis. Science. **239**:617-622.

66. **Feinberg, M.B., R.F. Jarrett, A. Aldovini, R.C. Gallo, and F. Wong-Staal.** 1986. HTLV-III expression and production involve complex regulation at the levels of splicing and translation of viral RNA. Cell. **46**:807-817.

67. **Felsenstein, J.** 1989. PHYLIP-Phylogeny Inference Package. Cladistics. **5**:164-6.

68. **Feng, Y., C.C. Broder, P.E. Kennedy, and E.A. Berger.** 1996. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G-protein-coupled receptor. Science. **272**:872-877.

69. **Fenyo, E.M., L. Morfeldt-Manson, F. Chiodi, B. Lind, A. von Gegerfelt, J. Albert, E. Olausson, and B. Asjo.** 1988. Distinct replicative and cytopathic characteristics of human immunodeficiency virus isolates. J. Virol. **62**:4414-4419.

70. **Fischl, M.A., D.D. Richman, M.H. Grieco, M.S. Gottlieb, P.A. Volberding, O.L. Laskin, J.M. Leedom, J.E. Groopman, D. Mildvan, R.T. Schooley, G.G. Jackson, D.T. Durack, and D. King.** 1987. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double blind placebo controlled trial. N. Engl. J. Med. **317**:185-191.

71. **Fisher, A.G., B. Ensoli, L. Ivanoff, M. Chamberlain, S. Petteway, L. Ratner, R.C. Gallo, and F. Wong-Staal.** 1987. The sor gene of HIV-1 is required for efficient virus transmission in vitro. Science. **237**:888-93.

72. **Fisher, A.G., B. Ensoli, D. Looney, A. Rose, R.C. Gallo, M.S. Saag, G.M. Shaw, B.H. Hahn, and F. Wong-Staal.** 1988. Biologically diverse molecular variants within a single HIV-1 isolate. Nature. **334**:444-7.

73. **Fisher, R.A., J.M. Bertonis, W. Meier, V. A. Johnson, D.S. Costopoulos, T. Liu, R. Tizard, B. D. Walker, M.S. Hirsch, R.T. Schooley, and R.A. Flavell.** 1988. HIV infection is blocked in vitro by recombinant

soluble CD4. Nature. **331**:76-78.

74. **Fitch, W.M., and E. Margoliash.** 1967. Construction of phylogenetic trees. Science. **155**:279-284.

75. **Fitzgibbon, J.E., S. Gaur, L. D. Frenkel, F. Laraque, B.R. Edlin, and D.T. Dubin.** 1993. Transmission from one child to another of human immunodeficiency virus type 1 with a zidovudine-resistance mutation. N. Engl. J. Med. **329**:1835-41.

76. **Fitzgibbon, J.E., S. Mazar, and D.T. Dubin.** 1993. A new type of G-->A hypermutation affecting human immunodeficiency virus. AIDS Res. Hum. Retroviruses. **9**:833-8.

77. **Fontenot, G., K. Johnston, J.C. Cohen, W.R. Gallaher, J. Robinson, and R.B. Luftig.** 1992. PCR amplification of HIV-1 proteinase sequences directly from lab isolates allows determination of five conserved domains. Virology. **190**:1-10.

78. **Frankel, A.D., D.S. Bredt, and C.O. Pabo.** 1988. Tat protein from human immunodeficiency virus forms a metal-linked dimer. Science. **240**:70-73.

79. **Fredriksson, R., P. Stålhanske, A. von Gegerfelt, B. Lind, P. Åman, E. Rassart, and Fenyö.** 1991. Biological characterization of infectious molecular clones derived from a human immunodeficiency virus type 1 isolate with rapid/high replicative capacity. Virology. **181**:55-61.

80. **Freed, E.O., and M.A. Martin.** 1995. The role of human immunodeficiency virus type 1 envelope glycoproteins in virus infection. J. Biol. Chem. **270**:23883-23886.

81. **Freed, E.O., and D.J. Myers.** 1992. Identification and characterization of fusion and processing domains of the human immunodeficiency virus type 2 envelope glycoprotein. J. Virol. **66**:5472-5478.

82. **Freed, E.O., D.J. Myers, and R. Risser.** 1991. Identification of the principal neutralizing determinant of human immunodeficiency virus type 1 as a fusion domain. J. Virol. **65**:190-194.

83. **Freed, E.O., and R. Risser.** 1991. Identification of conserved residues in the human immunodeficiency virus type 1 principal neutralizing determinant that are involved in fusion. AIDS Res. Hum. Retroviruses. **7**:807-811.

84. **Fung, M.S.C., C.R.Y. Sun, W.L. Gordon, R-S. Liou, T.W. Chang,**

W.N.C. Sun, E.S. Daar, and D.D. Ho. 1992. Identification and characterization of a neutralization site within the second variable region of human immunodeficiency virus type 1 gp120. J. Virol. **66**:848-856.

85. **Gallo, R.C., S.Z. Salahuddin, M. Popovic, G.M. Shearer, M. Kaplan, B.F. Haynes, T.J. Palker, R. Redfield, J. Oleske, and B. Safai.** 1984. Frequent detection and isolation of cytopathic retroviruses (HTLVIII) from patients with AIDS and at risk for AIDS. Science. **224**:500-503.

86. **Gao, F., L. Yue, S.C. Hill, D.L. Robertson, A.H. Graves, M.S. Saag, G.M. Shaw, P.M. Sharp, and B.H. Hahn.** 1994. HIV-1 sequence subtype D in the United States. AIDS Res. Hum. Retroviruses. **10**:625-627.

87. **Gelderblom, H.R., E.H.S. Hausmann, M. Ozel, G. Pauli, and M.A. Koch.** 1987. Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. Virology. **156**:171-176.

88. **Gelderblom, H.R., M. Ozel, and G. Pauli.** 1989. Morphogenesis and morphology of HIV. Structure-function relations. Arch. Virol. **106**:1-13.

89. **Gnann, J.W., J.B. McCormick, S. Mitchell, J.A. Nelson, and M.B. A. Oldstone.** 1987. Synthetic peptide immunoassay distinguishes HIV type 1 and HIV type 2 infections. Science. **237**:1346-1349.

90. **Gojobori, T., W-H. Li, and D. Graur.** 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360-369.

91. **Gojobori, T., E.N. Moriyama, and M. Kimura.** 1990. Statistical methods for estimating sequence divergence, p. 531-550. *In* R. F. Doolittle (ed.), Molecular evolution: computer analysis of protein and nucleic acid sequences, vol. 183. Academic Press, London.

92. **Gomatos, P.J., N.M. Stamatos, H.E. Gendelman, A. Fowler, D.L. Hoover, D.C. Kalter, D.S. Burke, E.C. Tramont, and M.S. Meltzer.** 1990. Relative inefficiency of soluble recombinant CD4 for inhibition of infection by monocyte-tropic HIV in monocytes and T cells. J.Immunol. **144**:4183-4188.

93. **Goodenow, M., T. Huet, W. Saurin, S. Kwok, J. Sninsky, and S. Wain-Hobson.** 1989. HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. J. AIDS. **2**:344-352.

94. **Gottlieb, M.D., R. Schroff, H.M. Schanker, J.D. Weisman, P.T.**

Fan, R.A. Wolf, and A. Saxon. 1981. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men. N. Engl. J. M. **305:**1425-1431.

95. **Göttlinger, H., T. Dorfman, J.G. Sodroski, and W. Haseltine.** 1991. Effect of mutations affecting the p6 *gag* protein on human immunodeficiency virus particle release. Proc. Natl. Acad. Sci. USA. **88:**3195-3199.

96. **Greene, W.C.** 1991. The molecular biology of human immunodeficiency virus type 1 infection. N. Engl.J. Med. **324:**308-317.

97. **Grez, M., U. Dietrich, P. Balfe, H. von Briesen, J.K. Maniar, G. Mahambre, E.L. Delwart, J.I. Mullins, and H. Rübsamen-Waigmann.** 1994. Genetic analysis of human immunodeficiency virus type 1 and 2 (HIV-1 and HIV-2) mixed infections in India reveals a recent spread of HIV-1 and HIV-2 from a single ancestor for each of these viruses. J. Virol. **68:**2161-2168.

98. **Groenink, M., R.A. Fouchier, S. Broersen, C.H. Baker, M. Koot, A.B. van't Wout, H.G. Huisman, F. Miedema, M. Tersmette, and H. Schuitemaker.** 1993. Relation of phenotype evolution of HIV-1 to envelope V2 configuration. Science. **260:**1513-1516.

99. **Gurtler, L.G., P.H. Hauser, J. Eberle, A. von Brunn, S. Knapp, L. Zekeng, J.M. Tsague, and L. Kaptue.** 1994. A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. J. Virol. **68:**1581-5.

100. **Hahn, B.H., M.A. Gonda, G.M. Shaw, M. Popovic, J.A. Hoxie, R.C. Gallo, and F. Wong-Staal.** 1985. Genomic diversity of the acquired immune deficiency syndrome virus HTLV- III: different viruses exhibit greatest divergence in their envelope genes. Proc. Natl. Acad. Sci. USA. **82:**4813-4817.

101. **Hahn, B.H., G.M. Shaw, S.K. Arya, M. Popovic, R.C. Gallo, and F. Wong-Staal.** 1984. Molecular cloning and characterization of HTLV-III virus associated with AIDS. Nature. **312:**166-169.

102. **Hahn, B.H., G.M. Shaw, M.E. Taylor, R.R. Redfield, P.D. Markham, S.Z. Salahuddin, F. Wong-Staal, R.C. Gallo, E.S. Parks, and W.P. Parks.** 1986. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science. **232:**1548-1553.

103. **Hauber, J., A. Perkins, E.P. Heimer, and B.R. Cullen.** 1988. Trans-activation of human immunodeficiency virus gene expression is mediated by nuclear

events. Proc. Natl. Acad. Sci. USA. **84**:6364-6368.

104. **Hay, A.J., A.R. Douglas, D.B. Sparrow, K.R. Cameron, and J.J. Skehel.** 1994. Antigenic and genetic characterisation of current influenza strains. European Journal of Epidemiology. **10**:465-466.

105. **Heinzinger, N., L. Baca Regen, M. Stevenson, and H.E. Gendelman.** 1995. Efficient synthesis of viral nucleic acids following monocyte infection by HIV-1. Virology. **206**:731-735.

106. **Helseth, E., U. Olshevsky, C. Furman, and J. Sodroski.** 1991. Human immunodeficiency virus type 1 gp120 envelope glycoprotein regions important for association with the gp41 transmembrane glycoprotein. J. Virol. **65**:2119-2123.

107. **Henderson, L.E., M.A. Bowers, R.C. Sowder, S.A. Serabyn, D.G. Johnson, J.W. Bess, L.O. Arthur, D.K. Bryant, and C. Fenselau.** 1992. Gag proteins of the highly replicative MN strain of human immunodeficiency virus type 1: posttranslational modifications, proteolytic processings, and complete amino acid sequences. J. Virol. **66**:1856-1865.

108. **Heptonstall, J., O.N. Gill, K. Porter, M.B. Black, and V.L. Gilbart.** 1993. Health care workers and HIV: surveillance of occupationally acquired infection in the United Kingdom. Communicable Disease Report Review. **3**:R147-R153.

109. **Heptonstall, J., K. Porter, and O.N. Gill.** 1993. Occupational transmission of HIV. Internal report of the PHLS, available from PHLS AIDS Centre at CDSC. Public Health Laboratory Service.

110. **Higgins, D.G., A.J. Bleasby, and R. Fuchs.** 1991. Clustal V: improved software for multiple sequence alignment. Computer Applications in the Biosciences. **8**:189-191.

111. **Hillis, D.M., and J.P. Huelsenback.** 1994. Support for dental HIV transmission. Nature. **369**:24-25.

112. **Ho, D.D., M.S.C. Fung, Y. Cao, C. Sun, T.W. Chang, and N.C. Sun.** 1991. Another discontinuous epitope on glycoprotein gp120 that is important in human immunodeficiency virus type 1 neutralization is identified by a monoclonal antibody. Proc. Natl. Acad. Sci. USA. **88**:8949-8952.

113. **Ho, D.D., A.U. Neumann, A.S. Perelson, W. Chen, J.M. Leonard, and M. Markowitz.** 1995. Rapid turnover of plasma virions and CD4 lymphocytes in

HIV-1 infection. Nature. **373**:123-126.

114. **Hochuli, V., O. Hyndman, and K. Porter.** 1995. Response to news that an obstetrician/gynaecologist has AIDS. Communicable Disease Report. **5**:7-11.

115. **Holmes, E.C.** 1993. The pattern and process of base substitution in immunodeficiency viruses. Binary. **5**:189-190.

116. **Holmes, E.C., and A.J. Leigh Brown.** 1993. Sequence data as evidence. Nature. **364**:766.

117. **Holmes, E.C., L.Q. Zhang, P. Robertson, A. Cleland, E. Harvey, P. Simmonds, and A.J. Leigh Brown.** 1995. The molecular epidemiology of Human Immunodeficiency Virus Type 1 in Edinburgh. The Journal of Infectious Diseases. **171**:45-53.

118. **Holmes, E.C., L.Q. Zhang, P. Simmonds, C.A. Ludlam, and A.J. Brown.** 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA. **89**:4835-4839.

119. **Holmes, E.C., L.Q. Zhang, P. Simmonds, A.S. Rogers, and A.J. Brown.** 1993. Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. J Infect Dis. **167**:1411-4.

120. **Hsu, T.W., and J.M. Taylor.** 1982. Single-stranded regions on unintegrated avian retrovirus DNA. J. Virol. **44**:47-53.

121. **Hu, W.S., and H.M. Temin.** 1990. Retroviral recombination and reverse transcription. Science. **250**:1227-1233.

122. **Huang, M., J.M. Orenstein, M.A. Martin, and E.O. Freed.** 1995. p6^{Gag} is required for particle production from full-length human immunodeficiency virus type 1 molecular clones expressing protease. J. Virol. **69**:6810-6818.

123. **Huang, Y., L. Zhang, and D.D. Ho.** 1995. Characterization of nef sequences in long-term survivors of human immunodeficiency virus type 1 infection. J. Virol. **69**:93-100.

124. **Hussey, R.E., N.E. Richardson, M. Kowalski, N.R. Brown, H.C. Chang, R.F. Siliciano, T. Dorfman, B. Walker, J. Sodroski, and E.L. Reinherz.** 1988. A soluble protein selectively inbibits HIV replication and syncytium formation. Nature. **331**:78-81.

125. **Hwang, S.S., T.J. Boyle, H.K. Lyerly, and B.R. Cullen.** 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science. **253**:71-74.

126. **Ichimura, H., S.C. Kliks, S. Visrutaratna, C-Y. Ou, M.L. Kalish, and J.A. Levy.** 1994. Biological, serological, and genetic characterization of HIV-1 subtype E isolates from Northern Thailand. AIDS Res. Hum. Retroviruses. **10**:263-269.

127. **Jabbar, M.A., and D.P. Nayak.** 1990. Intracellular interaction of human immunodeficiency virus type 1 (ARV- 2) envelope glycoprotein gp160 with CD4 blocks the movement and maturation of CD4 to the plasma membrane. J. Virol. **64**:6297-6304.

128. **Jacks, T., M.D. Power, F.R. Masiarz, P. Luciw, P.J. Barr, and H.E. Varmus.** 1988. Characterization of ribosomal frameshifting in HIV-1 *gag/pol* expression. Nature. **331**:280-283.

129. **Jaffe, H.W., J.M. McCurdy, M.L. Kalish, T. Liberti, G. Metellus, B.H. Bowman, S.B. Richards, A.R. Neasman, and J.J. Witte.** 1994. Lack of HIV Transmission in the Practice of a Dentist with AIDS. Annals of Internal Medicine. **121**:855-859.

130. **Ji, J., and L.A. Loeb.** 1994. Fidelity of HIV-1 reverse transcriptase copying a hypervariable region of the HIV-1 *env* gene. Virology. **199**:323-330.

131. **Jukes, T.H., and C.R. Cantor.** 1969. Evolution of protein molecules. Academic Press, New York.

132. **Junghans, R.P., L.R. Boone, and A.M. Skalka.** 1982. Products of reverse transcription in avian retroviruses analysed by electron microscopy. J. Virol. **43**:544-554.

133. **Junghans, R.P., L.R. Boone, and A.M. Skalka.** 1982. Retroviral DNA H structures: displacement-assimilation model of recombination. Cell. **30**:53-62.

134. **Kalish, M.L., C-C. Luo, B.G. Weniger, K. Limpakarnjanarat, N. Young, C-Y. Ou, and G. Schochetman.** 1994. Early HIV type 1 strains in Thailand were not responsible for the current epidemic. AIDS Res. Hum. Retroviruses. **10**:1573-1575.

135. **Kang, C-Y., K. Hariharan, M.R. Posner, and P. Nara.** 1993. Identification of a new neutralizing epitope conformationally affected by the attachment of CD4 to gp120. J. Immunol. **151**:449-457.

136. **Kaye, S., C. Loveday, and R.S. Tedder.** 1992. A microtitre format point mutation assay: application to the detection of drug resistance in human immunodeficiency virus type-1 infected patients treated with zidovudine. J Med Virol. **37**:241-6.

137. **Kestler, H.W., D.J. Ringler, K. Mori, D.L. Panicali, P.K. Sehgal, M.D. Daniel, and R.C. Desrosiers.** 1991. Importance of the nef gene for maintenance of high virus loads and for development of AIDS. Cell. **65**:651-662.

138. **Kim, S., R. Byrn, J. Groopman, and D. Baltimore.** 1989. Temporal aspects of DNA and RNA synthesis during human immunodeficiency virus infection: evidence for differential gene expression. J. Virol. **63**:3708-3713.

139. **Kimura, M.** 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. **16**:111-120.

140. **Klatzmann, D., E. Champagne, S. Chamaret, J. Gurest, D. Guetard, T. Hercend, J. C. Gluckman, and L. Montagnier.** 1984. T-lymphocyte T4 molecule behaves as a receptor for human retrovirus LAV. Nature. **312**:767-778.

141. **Klimkait, T., K. Strebel, M. D. Hoggan, M.A. Martin, and J.M. Orenstein.** 1990. The human immunodeficiency virus type 1-specific protein vpu is required for efficient virus maturation and release. J. Virol. **64**:621-629.

142. **Kodihalli, S., D.M. Justewicz, L.V. Gubareva, and R.G. Webster.** 1995. Selection of a single amino acid substitution in the haemagglutinin molecule by chicken eggs can render influenza A virus (H3) candidate vaccine ineffective. J. Virol. **69**:4888-4897.

143. **Kohl, N.E., E.A. Emini, W.A. Schlief, L.J. Davis, M.C. Hermbach, R.A.F. Dixon, E.M. Scolnik, and I.S. Sigel.** 1988. Active human immunodeficiency virus protease is required for viral infectivity. Proc. Natl. Acad. Sci. USA. **85**. :4686-4690.

144. **Korber, B.T.M., G. Learn, J.I. Mullins, B.H. Hahn, and S. Wolinsky.** 1995. Protecting HIV databases. Nature. **378**:242-244.

145. **Kostrikis, L.G., E. Bagdades, Y. Cao, L. Zhang, D. Dimitriou, and D.D. Ho.** 1995. Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: identification of a new subtype designated subtype I. J. Virol. **69**:6122-6130.

146. **Kunanusont, C., H.M. Foy, J.K. Kreiss, S. Rerks-Ngarm, P. Phanuphak, S. Raktham, C-P. Pau, and N.L. Young.** 1995. HIV-1 subtypes and male-to-female transmission in Thailand. Lancet. **345**:1078-1083.

147. **Kung, H-J., Y.K. Fung, J.E. Majors, J.M. Bishop, and H.E. Varmus.** 1981. Synthesis of plus strands of retroviral DNA in cells infected with avian sarcoma virus and mouse mammary tumour virus. J. Virol. **37**:127-138.

148. **Kusumi, K., B. Conway, S. Cunningham, A. Berson, C. Evans, A.K. Iversen, D. Colvin, M.V. Gallo, S. Coutre, E.G. Shpaer.** 1992. Human immunodeficiency virus type 1 envelope gene structure and diversity in vivo and after cocultivation in vitro. J. Virol. **66**:875-885.

149. **Lamers, S.L., J.W. Sleasman, J.X. She, K.A. Barrie, S.M. Pomeroy, D.J. Barrett, and M.M. Goodenow.** 1994. Persistence of multiple maternal genotypes of human immunodeficiency virus type I in infants infected by vertical transmission. J. Clin. Invest. **93**:380-90.

150. **Learmont, J., L. Cook, H. Dunckley, and J.S. Sullivan.** 1995. Update on long-term symptomless HIV type 1 infection in recipients of blood products from a single donor. AIDS Res. Hum. Retroviruses. **11**:1.

151. **Learmont, J., B. Tindall, L. Evans, A. Cunningham, P. Cunningham, J. Wells, R. Penny, J. Kaldor, and D.A. Cooper.** 1992. Long-term symptomless HIV-1 infection in recipients of blood products from a single donor. Lancet. **340**:863-867.

152. **Lehner, T., L. Hussain, J. Wilson, and M. Chapman.** 1991. Mucosal transmission of HIV. Nature. **353**:709.

153. **Leigh Brown, A.J.** 1991. Sequence variability in human immunodeficiency viruses: pattern and process in viral evolution. AIDS. **5 Suppl 2**:S35-42.

154. **Leigh Brown, A.J., and P. Simmonds.** 1995. Sequence analysis of virus variability based on the polymerase chain reaction (PCR), p. 161-188. *In* J. Karn (ed.), HIV A practical approach, vol. 1. Oxford University Press, Oxford.

155. **Leitner, T., A. Alaeus, S. Marquina, E. Lilja, K. Lidman, and J. Albert.** 1995. Yet another subtype of HIV type 1? AIDS Res. Hum. Retroviruses. **11**:995-997.

156. **Leonard, C.K., M.W. Spellman, L. Riddle, R.J. Harris, J.N.**

Thomas, and T.J. Gregory. 1990. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells. J. Biol. Chem. **265**:10373-10382.

157. **Lewis, P., M. Hensel, and M. Emerman.** 1992. Human immunodeficiency virus infection of cells arrested in the cell cycle. EMBO J. **11**:3053-3058.

158. **Lewis, P.F., and M. Emerman.** 1994. Passage through mitosis is required for oncoretroviruses but not for the human immunodeficiency virus. J. Virol. **68**:510-516.

159. **Linsley, P.S., J.A. Ledbetter, E. Kinney-Thomas, and S-L. Hu.** 1988. Effects of anti-gp120 monoclonal antibodies on CD4 receptor binding by the *env* protein of human immunodeficiency virus type 1. J. Virol. **62**:3695-3702.

160. **Loeb, D.D., C.A. Hutchison, M.H. Edgell, W.G. Farmerie, and R. Swanstrom.** 1989. Mutational analysis of human immunodeficiency virus type 1 protease suggests functional homology with aspartic proteinases. J. Virol. **63**:111-121.

161. **Loussert-Ajaka, I., T.D. Ly, M.L. Chiax, D. Ingerand, S. Saragosti, A.M. Couroucé, F. Brun-Vézinet, and F. Simon.** 1994. HIV-1/HIV-2 seronegativity in HIV-1 subtype O infected patients. Lancet. **343**:1393-1394.

162. **Louwagie, J., E.L. Delwart, J.I. Mullins, F.E. McCutchan, G. Eddy, and D.S. Burke.** 1994. Genetic analysis of HIV-1 isolates from Brazil reveals presence of two distinct genetic subtypes. AIDS Res. Hum. Retroviruses. **10**:561-567.

163. **Louwagie, J., F. McCutchan, J. Mascola, G. Eddy, K. Fransen, M. Peeters, G. van der Groen, and D. Burke.** 1993. Genetic subtypes of HIV-1. AIDS Res. Hum. Retroviruses. **9**:S147-S150.

164. **Luciw, P.A., C. Cheng-Mayer, and J.A. Levy.** 1987. Mutational analysis of the human immunodeficiency virus: the orf-B region down-regulates virus replication. Proc. Natl. Acad. Sci. USA. **84**:1434-1438.

165. **Maddon, P.J., A.G. Dalgleish, J.S. McDougal, P.R. Clapham, R.A. Weiss, and R. Axel.** 1986. The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. Cell. **47**:333-348.

166. **Malim, M.H., J. Hauber, R. Fenrick, and B.R. Cullen.** 1988. Immunodeficiency virus *rev trans*-activator modulates expression of the viral regulatory genes. Nature. **335**:181-183.

167. **Malim, M.H., J. Hauber, S-Y. Le, J.V. Maizel, and B.R. Cullen.** 1989. The HIV-1 *rev trans*-activator acts through a structured target sequence to activate the nuclear export of unspliced viral mRNA. Nature. **338**:254-257.

168. **Martins, L.P., N. Chenciner, B. Asjo, A. Meyerhans, and S. Wain-Hobson.** 1991. Independent fluctuation of human immunodeficiency virus type 1 rev and gp41 quasispecies in vivo. J. Virol. **65**:4502-7.

169. **Masur, H., M.A. Michelis, J.B. Greene, I. Onorato, R.A. vande Stouwe, R.S. Holzman, G. Wormser, L. Brettman, M. Lange, H.W. Murray, and S. Cunningham-Rundles.** 1981. An outbreak of community-acquired pneumocystis carinii pneumonia. N. Engl. J. Med. **305**:1431-1438.

170. **McCune, J.M., L.B. Rabin, M.B. Feinberg, M. Lieberman, J.C. Kosek, G.R. Reyes, and I.L. Weissman.** 1988. Endoproteolytic cleavage of gp160 is required for the activation of human immunodeficiency virus. Cell. **53**:55-67.

171. **McCutchan, F.E.** 1994. HIV Variation, p. 107-168. Prevention and treatment of AIDS. Wiley-Liss, Hilton Head Island, South Carolina.

172. **McCutchan, F.E., P.A. Hegerich, T.P. Brennan, P. Phanuphak, P. Singharaj, A. Jugsudee, P.W. Berman, A.M. Gray, A.K. Fowler, and D.S. Burke.** 1992. Genetic variants of HIV-1 in Thailand. AIDS Res. Hum. Retroviruses. **8**:1887-95.

173. **McKeating, J.A., J. Cordell, C.J. Dean, and P. Balfe.** 1992. Synergistic interaction between ligands binding to the CD4 binding site and V3 domain of human immunodeficiency virus type 1 gp120. Virology. **191**:732-742.

174. **McKeating, J.A., C. Shotton, J. Cordell, S. Graham, P. Balfe, N. Sullivan, M. Charles, M. Page, A. Bolmstedt, S. Olofsson, S.C. Kayman, Z. Wu, A. Pinter, C. Dean, J. Sodroski, and R.A. Weiss.** 1993. Characterization of neutralizing monoclonal antibodies to linear and conformation-dependent epitopes within the first and second variable domains of human immunodeficiency virus type 1 gp120. J. Virol. **67**:4932-4944.

175. **McKeating, J.A., Y-J. Zhang, C. Arnold, R. Fredriksson, E-M. Fenyo, and P. Balfe.** 1996. Chimeric viruses expressing primary envelope glycoproteins of human immunodeficiency virus type 1 show increased sensitivity to neutralisation by human sera. Virology. **In press.**

176. **Meltzer, M.S., D.R. Skillman, D.L. Hoover, B.D. Hanson, J.A. Turpin, D.C. Kalter, and H.E. Gendelman.** 1990. HIV and the immune system. Macrophages and the human immunodeficiency virus. Immunol. Today. **11**:217-223.

177. **Meyerhans, A., R. Cheynier, J. Albert, M. Seth, S. Kwok, J. Sninsky, L. Morfeldt-Manson, B. Asjo, and S. Wain-Hobson.** 1989. Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. Cell. **58**:901-10.

178. **Meyerhans, A., J.P. Vartanian, and S. Wain-Hobson.** 1990. DNA recombination during PCR. Nucleic Acids Res. **18**:1687-91.

179. **Milich, L., B. Margolin, and R. Swanstrom.** 1993. V3 loop of the human immunodeficiency virus type 1 env protein: interpreting sequence variability. J. Virol. **67**:5623-5634.

180. **Miller, M.D., M.T. Warmerdam, I. Gaston, W.C. Greene, and M.B. Feinberg.** 1994. The human immunodeficiency virus-1 nef gene product: a positive factor for viral infection and replication in primary lymphocytes and macrophages. J. Exp. Med. **179**:101-113.

181. **Modrow, S., B.H. Hahn, G.M. Shaw, R.C. Gallo, F. Wong-Staal, and H. Wolf.** 1987. Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. J. Virol. **61**:570-578.

182. **Moore, J.P., R.L. Willey, G.K. Lewis, J. Robinson, and J. Sodroski.** 1994. Immunological evidence for interactions between the first, second, and fifth conserved domains of the gp120 surface glycoprotein of Human Immunodeficiency Virus type 1. J. Virol. **68**:6836-6847.

183. **Moriyama, E.N., Y. Ina, K. Ikeo, N. Shimizu, and T. Gojobori.** 1991. Mutation pattern of human immunodeficiency virus genes. J. Mol. Evol. **32**:360-363.

184. **Mortimer, P.P., and A.G. Nicoll.** 1994. A demanding diagnosis: testing for HIV infection in infancy. PHLS Microbiology Digest. **11**:66-69.

185. **Mulder-Kampinga, G.A., C. Kuiken, J. Dekker, H.J. Scherpbier, K. Boer, and J. Goudsmit.** 1993. Genomic human immunodeficiency virus type 1 RNA variation in mother and child following intra-uterine virus transmission. J. Gen. Virol.

74:1747-56.

186. **Murphy, E., B. Korber, M.C. Georges-Courbot, B. You, A. Pinter, D. Cook, M.P. Kieny, A. Georges, C. Mathiot, F. Barré-Sinoussi.** 1993. Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic. AIDS Res. Hum. Retroviruses. **9**:997-1006.

187. **Myers, G.** 1994. Molecular investigation of HIV transmission. Ann. Int. Med. **121**:889-890.

188. **Myers, G., B. Korber, J.A. Berzofsky, T.F. Smith, and G.N.E. Pavlakis.** 1991. Human retroviruses and AIDS. Los Alamos National Laboratory, Los Alamos.

189. **Myers, G., B. Korber, S. Wain-Hobson, R.F. Smith, and G.N. Pavlakis.** 1994. Human Retroviruses and AIDS 1994. Los Alamos National Laboratory, Los Alamos, N.M.

190. **Nabel, G., and D. Baltimore.** 1987. An inducible transcription factor activates expression of human immunodeficiency virus in T cells. Nature. **326**:711-713.

191. **Nei, M.** 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

192. **Newell, M.L., and C. Peckham.** 1993. Risk factors for vertical transmission of HIV-1 and early markers of HIV-1 infection in children. AIDS. **7**:S591-S597.

193. **Nicolosi, A., M. Musicco, A. Saracco, and A. Lazzarin.** 1994. Risk factors for woman-to-man sexual transmission of the human immunodeficiency virus. Italian Study Group on HIV Heterosexual Transmission. J. AIDS. **7**:296-300.

194. **Novitsky, V., C. Arnold, and J.P. Clewley.** 1996. Heteroduplex mobility assay for subtyping HIV-1: improved methodology and comparison with phylogenetic analysis of sequence data. J. Virol. Meth. **in Press.**

195. **Nowak, M.A., R.M. May, and R.M. Anderson.** 1990. The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. AIDS. **4**:1095-1103.

196. **O'Brien, W.A., Y. Koyanagi, A. Namazie, J.Q. Zhao, A. Diagne, K. Idler, J.A. Zack, and I.S. Chen.** 1990. HIV-1 tropism for mononuclear phagocytes can be determined by regions of gp120 outside the CD4-binding domain. Nature. **348**:69-73.

197. **Orita, M., H. Iwahana, H. Kanazawa, K. Hayashi, and T. Sekiya.** 1989. Detections of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc. Natl. Acad. Sci. U.S.A. **86**:2766-2770.

198. **Oroszlan, S., and R.B. Luftig.** 1990. Retroviral proteinases. Current Topics in Microbiological Immunology. **157**:153-185.

199. **Ou, C.Y., C.A. Ciesielski, G. Myers, C.I. Bandea, C.C. Luo, B.T. Korber, J.I. Mullins, G. Schochetman, R.L. Berkelman, A.N. Economou, J.J. Witte, L.J. Furman, G.A. Satten, K.A. MacInnes, J.W. Curran and H.W. Jaffe.** 1992. Molecular epidemiology of HIV transmission in a dental practice. Science. **256**:1165-71.

200. **Ou, C.Y., Y. Takebe, B.G. Weniger, C.C. Luo, M.L. Kalish, W. Auwanit, S. Yamazaki, H.D. Gayle, N.L. Young, and G. Schochetman.** 1993. Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand [published erratum appears in Lancet 1993 Jul 24;342(8865):250]. Lancet. **341**:1171-1174.

201. **Ozel, M., G. Pauli, and H.R. Gelderblom.** 1988. The organization of the envelope projections on the surface of HIV. Archives of Virology. **100**:255-266.

202. **Palca, J.** 1993. CDC closes the case of the Florida dentist. Science. **255**:766.

203. **Pang, S., H.V. Vinters, T. Akashi, W.A. O'Brien, and I.S. Chen.** 1991. HIV-1 env sequence variation in brain tissue of patients with AIDS- related neurologic disease. J. AIDS **4**:1082-92.

204. **Pauza, C.D.** 1990. Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. Virology. **179**:886-9.

205. **Paxton, W.R., R.I. Connor, and N.R. Landau.** 1993. Incorporation of Vpr into human immunodeficiency virus type 1 virions: requirement for the p6 region of *gag* and mutational analysis. J. Virol. **67**:7229-7237.

206. **PHLS.** 1996. AIDS and HIV-1 infection in the United Kingdom: monthly report. Communicable Disease Report. **6**:25-28.

207. **Pienazek, D., L.M. Janini, A. Ramos, A. Tanuri, M. Schechter, J. M. Peraltal, A.C.P. Vicente, N.J. Pienazek, G. Schochetman, and M.A. Rayfield.** 1995. HIV-1 patients may harbor viruses of different phylogenetic subtypes: implications for the evolution of the HIV/AIDS pandemic. Emerging Infectious Diseases.

1:86-88.

208. **Pinter, A., W.J. Honnen, and S.A. Tilley.** 1993. Conformational changes affecting the V3 and CD4-binding domains of human immunodeficiency virus type 1 gp120 associated with env processing and with binding of ligands to these sites. J. Virol. **67**:5692-5697.

209. **Pinter, A., W.J. Honnen, S.A. Tilley, C. Bona, H. Zaghouani, M.K. Gorny, and S. Zolla-Pazner.** 1989. Oligomeric structure of gp41, the transmembrane protein of human immunodeficiency virus type 1. J. Virol. **63**:2674-2679.

210. **Plummer, F.A., J.N. Simonsen, D.W. Cameron, J.O. Ndinya-Achola, J.K. Kreiss, M.N. Gakinya, P. Waiyaki, M. Cheang, P. Piot, A.R. Ronald and E.N. Ngugi.** 1991. Cofactors in male-female sexual transmission of human immunodeficiency virus type 1 [see comments]. J. I. D. **163**:233-9.

211. **Preston, B.D., B.J. Poiesz, and L.A. Loeb.** 1988. Fidelity of HIV-1 reverse transcriptase. Science. **242**:1168-1171.

212. **Reitz, M.S., Jr., C. Wilson, C. Naugle, R.C. Gallo, and M. Robert-Guroff.** 1988. Generation of a neutralization-resistant variant of HIV-1 is due to selection for a point mutation in the envelope gene. Cell. **54**:57-63.

213. **Ricchetti, M., and H. Buc.** 1990. Reverse transcriptases and genomic variability: the accuracy of DNA replication is enzyme specific and sequence dependent. EMBO Journal. **9**:1583-1593.

214. **Richardson, C.D., and P.W. Choppin.** 1983. Oligopeptides that specifically inhibit membrane fusion by paramyxoviruses: studies on the site of action. Virology. **131**:518-532.

215. **Rogel, M., L. Wu, and M. Emerman.** 1995. The human immunodeficiency virus type 1 *vpr* gene prevents cell proliferation during chronic infection. J. Virol. **69**:882-888.

216. **Roques, P.A., G. Gras, F. Parnet-Mathieu, A. Mabondzo, C. Dollfus, R. Narwa, D. Marcé, J. Tranchot-Diallo, F. Hervé, G. Lasfargues, C. Courpotin, and D. Dormont.** 1995. Clearance of HIV infection in 12 perinatally infected children: clinical, virological and immunological data. AIDS. **9**:F19-F26.

217. **Rosenberg, Z.F., and A.S. Fauci.** 1990. Activation of latent HIV infection.

J. NIH Res. **2**:41-45.

218. **Rouse, B.T., and D.W. Horohov.** 1986. Immunosuppression in viral infections. Rev. Infect. Dis. **8**:850-873.

219. **Saag, M.S., B.H. Hahn, J. Gibbons, Y. Li, E.S. Parks, W.P. Parks, and G.M. Shaw.** 1988. Extensive variation of human immunodeficiency virus type-1 in vivo. Nature. **334**:440-4.

220. **Sadaie, M.R., T. Benter, and F. Wong-Staal.** 1988. Site-directed mutagenesis of two trans-regulatory genes (*tat*-III, *trs*) of HIV-1. Science. **239**:910-913.

221. **Saitou, N., and T. Imanishi.** 1989. Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum likelihood, minimum evolution and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. **6**:514-25.

222. **Sambrook, J., E.F. Fritsch, and T. Maniatis.** 1989. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York.

223. **Sattentau, Q., J. P. Moore, F. Vignaux, F. Traincard, and P. Poignard.** 1993. Conformational changes induced in the envelope glycoproteins of the human and simian immunodeficiency viruses by soluble receptor binding. J. Virol. **67**:7383-7393.

224. **Sattentau, Q. J., and J. P. Moore.** 1991. Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding. J. Exp. Med. **174**:407-415.

225. **Scarlatti, G., T. Leitner, E. Halapi, J. Wahlberg, P. Marchisio, M.A. Clerici-Schoeller, H. Wigzell, E.M. Fenyo, J. Albert, M. Uhlen and P. Rossi.** 1993. Comparison of variable region 3 sequences of human immunodeficiency virus type 1 from infected children with the RNA and DNA sequences of the virus populations of their mothers. Proc. Natl. Acad. Sci. USA. **90**:1721-5.

226. **Schnittman, S.M., M.C. Psallidopoulos, H.C. Lane, L. Thompson, M. Baseler, F. Massari, C.H. Fox, N.P. Salzman, and A.S. Fauci.** 1989. The reservoir for HIV-1 in human peripheral blood is a T cell that maintains expression of CD4. Science. **245**:305-308.

227. **Schuitemaker, H., M. Koot, N.A. Koostra, M.W. Dercksen, R.E. de Goede, R.P. van-Steenwijk, J.M. Lange, J.K. Schattenkerk, F. Miedema,**

and M. Tersmette. 1992. Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus populations. J. Virol. **66:**1354-1360.

228. **Schuitemaker, H., N.A. Kootstra, R.E. de Goede, F. de Wolf, F. Miedema, and M. Tersmette.** 1991. Monocytotropic human immunodeficiency virus type 1 (HIV-1) variants detectable in all stages of HIV-1 infection lack T-cell line tropism and syncytium-inducing ability in primary T-cell culture. J. Virol. **65:**356-363.

229. **Schulz, T.F., B.A. Jameson, L. Lopalco, A.G. Siccardi, R.A. Weiss, and J.P. Moore.** 1992. Conserved structural features in the interaction between retroviral surface and transmembrane glycoproteins. AIDS Res. Hum. Retroviruses. **8:**1571-1580.

230. **Sharp, P.M., D.L. Robertson, F. Gao, and B.H. Hahn.** 1994. Origins and diversity of human immunodeficiency viruses. AIDS. **8:**S27-S42.

231. **Shaw, G.M., B.H. Hahn, S.K. Arya, J.E. Groopman, R.C. Gallo, and F. Wong-Staal.** 1984. Molecular characterization of human T-cell leukemia (lymphotropic) virus type III in the acquired immune deficiency syndrome. Science. **226:**1165-1171.

232. **Shaw, G.M., M.E. Harper, B.H. Hahn, L.G. Epstein, D.C. Gajdusek, P.R.W., B.A. Navia, C.K. Petito, C.J. O'Hara, J.E. Groopman, E.S. Cho, J.M. Oleske, F. Wong-Staal, and R.C. Gallo.** 1985. HTLV-III infection in brains of children and adults with AIDS encephalopathy. Science. **227:**177-182.

233. **Shioda, T., J.A. Levy, and C. Cheng-Mayer.** 1991. Macrophage and T cell-line tropisms of HIV-1 are determined by specific regions of the envelope gp120 gene. Nature. **349:**167-169.

234. **Shotton, C., C. Arnold, Q. Sattentau, J. Sodroski, and J.A. McKeating.** 1995. Identification and characterization of monoclonal antibodies specific for polymorphic antigenic determinants within the V2 region of the human immunodeficiency virus type 1 envelope glycoprotein. J. Virol. **69:**222-230.

235. **Simmonds, P., P. Balfe, C.A. Ludlam, J.O. Bishop, and A.J. Brown.** 1990. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. J. Virol. **64:**5840-50.

236. **Simmonds, P., L.Q. Zhang, F. McOmish, P. Balfe, C.A. Ludlam, and A.J. Brown.** 1991. Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. J. Virol. **65**:6266-6276.

237. **Skalka, A.M.** 1988. Integrative recombination of retroviral DNA, p. 701-724. *In* R. Kucherlapati and G. R. Smith (ed.), Genetic recombination. American Society for Microbiology, Washington D.C.

238. **Skinner, M.A., A.J. Langlois, C.B. McDanal, J.S. McDougal, D.P. Bolognesi, and T.J. Matthews.** 1988. Neutralizing antibodies to an immunodominant envelope sequence do not prevent gp120 binding to CD4. J. Virol. **62**:4195-4200.

239. **Smith, D.H., R.A. Byrn, S.A. Marsters, T. Gregory, J.E. Groopman, and D.J. Capon.** 1987. Blocking of HIV-1 infectivity by a soluble, secreted form of the CD4 antigen. Science. **238**:1704-1707.

240. **Smith, T.F., and M.S. Waterman.** 1992. The continuing case of the Florida dentist. Science. **256**:1155-1156.

241. **Sodroski, J., W.C. Goh, C. Rosen, A. Tartar, D. Portelle, A. Burny, and W. Haseltine.** 1986. Replicative and cytopathic potential of HTLV-III/LAV with *sor* gene deletions. Science. **231**:1549-53.

242. **Sodroski, J., C. Rosen, F. Wong-Staal, S.Z. Salahuddin, M. Popovic, S. Arya, R.C. Gallo, and W.A. Haseltine.** 1985. Trans-acting transcriptional regulation of human T-cell leukemia virus type III long terminal repeat. Science. **227**:171-173.

243. **Sodroski, J.G., W.C. Goh, C. Rosen, A. Dayton, E. Terwilliger, and W.A. Haseltine.** 1986. A second posttranscriptional transactivator gene required for HTLV-III replication. Nature. **321**:412-417.

244. **Sokal, R.R., and F.J. Rohlf.** 1987. The Poisson Distribution, 2 ed. W.H Freeman and Company, San Francisco.

245. **Soto-Ramirez, L.E., B. Renjifo, M.F. McLane, R. Marlink, C. O'Hara, R. Sutthent, C. Wasi, P. Vithayasai, V. Vithayasai, C. Apichartpiyakul, P. Auewarakul, V. Peña Cruz, D-S. Chui, R. Osathanondh, K. Mayer, T-H. Lee, and M. Essex.** 1996. HIV-1 langerhans' cell

tropism associated with heterosexual transmission of HIV. Science. **271**:1291-1293.

246. **Starcich, B.R., B.H. Hahn, G.M. Shaw, P.D. McNeely, S. Modrow, H. Wolf, E.S. Parks, W.P. Parks, S.F. Josephs , and R.C. Gallo.** 1986. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell. **45**:637-648.

247. **Steffy, K.R., G. Kraus, D.J. Looney, and F. Wong-Staal.** 1992. Role of the fusogenic peptide sequence in syncytium induction and infectivity of human immunodeficiency virus type 2. J. Virol. **66**:4532-4535.

248. **Stein, B.S., S.D. Gowda, J.D. Lifson, R.C. Penhallow, K.G. Bensch, and E.G. Engelman.** 1987. pH-independent HIV entry into CD4-positive T cells via virus envelope fusion to the plasma membrane. Cell. **49**:659-668.

249. **Stevenson, M., C. Meier, A.M. Mann, N. Chapman, and A. Wasiak.** 1988. Envelope glycoprotein of HIV induces interference and cytolysis resistance in CD4$^+$ cells: mechanism for persistence in AIDS. Cell. **53**:483-496.

250. **Strebel, K., D. Daugherty, K. Clouse, D. Cohen, T. Folks, and M.A. Martin.** 1987. The HIV 'A' (sor) gene product is essential for virus infectivity. Nature. **328**:728-730.

251. **Strunnikova, N., S.C. Ray, R.A. Livingstone, E. Rubalcaba, and R.P. Viscidi.** 1995. Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. J. Virol. **69**:7548-7558.

252. **Anon. European Collaborative Study.** 1994. Caesarean section and the risk of vertical transmission on HIV-1 infection. Lancet. **343**:1447-1464.

253. **Anon. European Collaborative Study.** 1992. Risk factors for mother-to-child transmission of HIV-1. Lancet. **339**:1007-1012.

254. **Sullivan, N., M. Thali, C. Furman, D.D. Ho, and J. Sodroski.** 1993. Effect of amino acid changes in the V1/V2 region of the human immunodeficiency virus type 1 gp120 glycoprotein on subunit association, syncytium formation, and recognition by a neutralizing antibody. J. Virol. **67**:3674-3679.

255. **Tan, W., R. Fredriksson, A. Bjorndal, P. Balfe, and E.M. Fenyo.** 1993. Cotransfection of HIV-1 molecular clones with restricted cell tropism may yield progeny virus with altered phenotype. AIDS Res. Hum. Retroviruses. **9**:321-329.

256. **Tersmette, M., R.E.Y. de Goede, J.M. Bert, I.N. Al, R.A. Winkel, H.T.C. Gruters, H.G. Huisman, and F. Miedema.** 1988. Differential syncytium-inducing capacity of human immunodeficiency isolates: frequent detection of syncytium inducing isolates in patients with acquired immunodeficiency virus syndrome (AIDS) and AIDS-related complex. J. Virol. **62**:2026-2032.

257. **Tersmette, M., R.A. Gruters, F. de Wolf, R.E.Y. de Goede, B.J. M. Lange, P.T.H. Schellekens, J. Goudsmit, J. G. Huisman, and F. Miedema.** 1989. Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates. J. Virol. **63**:2118-2125.

258. **Tersmette, M., J.M.A. Lange, R.E.Y. de Goede, F. de Wolf, J. K. M. Eeftinck Shattenkerk, P.T.H. Schellekens, R.A. Coutinho, J.G. Huisman, J. Goudsmit, and F. Miedema.** 1989. Association between biological properties of human immunodeficiency virus variants and risk for AIDS and AIDS mortality. Lancet. **i.** :983-985.

259. **Terwilliger, E.F., E. Langhoff, D. Gabuzda, E. Zazopoulos, and W.A. Haseltine.** 1991. Allelic variation in the effects of the nef gene on replication of human immunodeficiency virus type 1. Proc. Natl. Acad. Sci. USA. **88**:10971-10975.

260. **Thali, M., J.P. Moore, C. Furman, M. Charles, D.D. Ho, J. Robinson, and J. Sodroski.** 1993. Characterization of conserved human immunodeficiency virus type 1 gp120 neutralization epitopes exposed upon gp120-CD4 binding. J. Virol. **67**:3978-3988.

261. **Tosswill, J.H.C., K.L. Barlow, J.V. Parry, and J.P. Clewley.** 1994. Polymerase chain reaction to diagnose HIV-1. Lancet. **343**:1431.

262. **Traunecker, A., W. Luke, and K. Karjalainen.** 1988. Soluble CD4 molecules neutralize human immunodeficiency virus type 1. Nature. **331**:84-86.

263. **Tristem, M., C. Marshall, A. Karpas, and F. Hill.** 1992. Evolution of the primate lentiviruses: evidence from vpx and vpr. EMBO J. **11**:3405-3412.

264. **Trowsdale, J., J.A. Young, A.P. Kelly, P.J. Austin, S. Carson, H. Meunier, A. So, H.A. Erlich, R.S. Speilman, J. Bodmer, and W.F. Bodmer.** 1985. Structure, sequence and polymorphism in the HLA-D region. Immunology Reviews. **85**:5-43.

265. **Vanden Haesevelde, M., J.L. Decourt, R.J. De Leys, B. Vanderborght, G. van der Groen, H. van Heuverswijn, and E. Saman.** 1994. Genomic cloning and complete sequence analysis of a highly divergent African human immunodeficiency virus isolate. J. Virol. **68:**1586-96.

266. **Varmus, H.** 1988. Retroviruses. Science. **240:**1427-1435.

267. **Varmus, H.E., and R. Swanstrom.** 1985. Replication of retroviruses, p. 75-134. *In* R. Weiss and N. Teich and H. Varmus and J. Coffin (ed.), RNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

268. **Vartanian, J.P., A. Meyerhans, B. Asjo, and S. Wain-Hobson.** 1991. Selection, recombination, and G-->A hypermutation of human immunodeficiency virus type 1 genomes. J. Virol. **65:**1779-88.

269. **Vartanian, J.P., A. Meyerhans, M. Sala, and S. Wain-Hobson.** 1994. G-->A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc. Natl. Acad. Sci. USA. **91:**3092-6.

270. **Veronese, F.D., A.L. DeVico, T. Copeland, S. Oroszlan, R.C. Gallo, and M. Sarngadharan.** 1985. Characterization of gp41 as the transmembrane protein coded by the HTLV-III/LAV envelope gene. Science. **229:**1402-1405.

271. **Vincent, M.J., N.U. Raja, and M.A. Jabbar.** 1993. The human immunodeficiency virus type 1 Vpu protein induces degradation of chimeric envelope glycoproteins bearing the cytoplasmic and anchor domains of CD4: role of cytoplasmic domain in Vpu-induced degradation in the endoplasmic reticulum. J. Virol. **67:**5538-5549.

272. **von Schwedler, U., J. Song, C. Aiken, and D. Trono.** 1993. Vif is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells. J. Virol. **67:**4945-55.

273. **Warrier, S.V., A. Pinter, W.J. Honnen, M. Girard, E. Muchmore, and S.A. Tilley.** 1994. A novel glycan-dependent epitope in the V2 domain of human immunodeficiency virus type 1 gp120 is recognized by a highly potent, neutralizing chimpanzee monoclonal antibody. J. Virol. **68:**4636-4642.

274. **Wei, X., S.K. Ghosh, M.E. Taylor, V.A. Johnson, E.A. Emini, P. Deutsch, J.D. Lifson, S. Bonhoeffer, M. Nowak, B.H. Hahn, and G. M. Shaw.** 1995. Viral dynamics in human immunodeficiency virus type 1 infection. Nature.

373:117-122.

275. **Weiss, C.D., J.A. Levy, and J.M. White.** 1990. Oligomeric organization of gp120 on infectious human immunodeficiency virus type 1 particles. J. Virol. **64:**5674-5677.

276. **Westervelt, P., D.B. Trowbridge, L.G. Epstein, B.M. Blumberg, Y. Li, B.H. Hahn, G.M. Shaw, R.W. Price, and L. Ratner.** 1992. Macrophage tropism determinants of human immunodeficiency virus type 1 in vivo. J. Virol. **66:**2577-2582.

277. **White, J.M.** 1990. Viral and cellular membrane fusion proteins. Annu. Rev. Physiol. **52:**675-697.

278. **WHO Global programme on AIDS.** 1994. The HIV/AIDS pandemic: 1994 overview. **WHO/GPA/SEF/94.4:**World Health Organisation, Geneva.

279. **Wike, C.M., B.T. Korber, M.R. Daniels, C. Hutto, J. Munoz, M. Furtado, W. Parks, A. Saah, M. Bulterys, and J.B. Kurawige.** 1992. HIV-1 sequence variation between isolates from mother-infant transmission pairs. AIDS Res. Hum. Retroviruses. **8:**1297-300.

280. **Willey, R., A. Rutledge, S. Dias, T. Folks, T. Theodore, and C.E. Buckler.** 1986. Identification of conserved and divergent domains within the envelope gene of the acquired immunodeficiency syndrome retrovirus of AIDS. Cell. **45:**637-648.

281. **Willey, R.L., T. Klimkait, D.M. Frucht, J.S. Bonifacino, and M.A. Martin.** 1991. Mutations within the human immunodeficiency virus type 1 gp160 envelope glycoprotein alter its intracellular transport and processing. Virology. **184:**319-329.

282. **Willey, R.L., F. Maldarelli, M.A. Martin, and K. Strebel.** 1992. Human immunodeficiency virus type 1 Vpu protein induces rapid degradation of CD4. J. Virol. **66:**7193-7200.

283. **Willey, R.L., and M.A. Martin.** 1993. Association of human immunodeficiency virus type 1 envelope glycoprotein with particles depends on interactions between the third variable and conserved regions of gp120. J. Virol. **67:**3639-3643.

284. **Winslow, D.L., S. Stack, R. King, H. Scarnati, A. Bincsik, and M.J. Otto.** 1995. Limited sequence diversity of the HIV type 1 protease gene from

References added in proof:

293. **Aiken, C., J. Konner, N.R. Landau, M.E. Lenburg, and D. Trono.** 1994. Nef induces CD4 endocytosis: requirement for a critical dileucine motif in the membrane-proximal CD4 cytoplasmic domain. Cell **7:**1015-1020.

294. **Anderson, S., D.C. Shugars, R. Swanstrom, and J.V Garcia.** 1993. Nef from primary isolates of human immunodeficiency virus type 1 suppresses surface CD4 expression in human and mouse T cells. J Virol. **67:**4923-4931.

295. **Benson, R.E., A. Sanfridson, J.S. Ottinger, C. Doyle, and B.R. Cullen.** 1993. Downregulation of cell-surface CD4 expression by simian immunodeficiency virus Nef prevents viral superinfection. J. Exp. Med. **177:**1561-1566.

296. **Brady, H.J., D.J. Pennington, C.G. Miles, and E. A. Dzierzak.** 1993. CD4 cell surface downregulation in HIV-1 transgenic mice is a consequence of intracellular sequestration. EMBO. **12**:4923-4932.

297. **Aiken, C., L. Krause, Y.L. Chen, and D. Trono.** 1996. Mutational analysis of HIV-1 Nef: identification of two mutants that are temperature-sensitive for CD4 downregulation. Virology. **217:**293-300.

298. **Eernisse, D.J.** 1992. DNA Translator and Aligner: Hypercard utilities to aid phylogenetic analysis. Computer Applications in the Biosciences. **8:**177-184.

299. **Moore, J.P., Y. Cao, J. Leu, L. Qin, B. Korber, and D.D. Ho.** 1996. Inter- and intraclade neutralisation of Human Immunodeficiency Virus type 1: Genetic clades do not correspond to neutralisation serotypes but partially correspond to gp120 antigenic serotypes. J Virol. **70:**427-444.

300. **Kostrikis, L.G., Y. Cao, H. Ngai, J.P. Moore, and D.D. Ho.** 1996. Quantitative analysis of serum neutralisation of Human Immunodeficiency Virus type 1 from subtypes A, B, C, D, E, F, and I: Lack of direct correlation between neutralisation serotypes and genetic subtypes and evidence for prevalent serum-dependent infectivity enhancement. J Virol. **70:**445-458.

301. **Weber, J., E-M. Fenyö, S. Beddows, P. Kaleebu, Å. Björndal, WHO Network for HIV isolation and characterisation.** 1996. Neutralisation serotypes of Human Immunodeficiency Virus type 1 field isolates are not predicted by genetic subtype. J Virol. **70:**7827-7832.

302. **Leitner, T., D. Escanilla, C. Franzén, M. Uhlén, and J. Albert.** 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. USA. **93:**10864-10869.