Analysis of HIV-1 and Foraminiferal Molecular Evolution

Christopher Mark Wade

PhD Thesis

Institute of Cell, Animal and Population Biology
University of Edinburgh

1997

TABLE OF CONTENTS

ABSTRACT

This thesis presents work examining aspects of both HIV-1 and foraminiferal evolution. The thesis is set out accordingly into two sections, each of which contains an introduction with individual chapters presented as a series of papers.

In section A, five papers are presented which examine the evolution of HIV-1 both within and between patients. The first paper presented examines the molecular epidemiology of HIV-1 within Scotland, Northern England, and Ireland (paper I), with attention focused on identifying risk group associated differences within the cohort. This work also provides important background information for the interpretation of molecular data from transmission clusters. The main focus of the work on HIV-1 evolution has been on the transmission of HIV-1, with particular emphasis placed on mother-child transmission. Four papers are presented which examine evolutionary aspects of HIV-1 transmission. The first of these (paper II) examines the viral variants transmitted from mother to child in four mother-child transmission pairs. The second (paper III) analyses similar data from five mother-child transmission pairs, focussing predominantly on viral evolution within the child over the first year of life. The final two papers investigating HIV-1 transmission examine viral variation within two transmission sets. Paper IV examines the vertical transmission of HIV-1 to two infected children born to the same mother at an approximately two year interval, while paper V examines the heterosexual transmission of HIV-1 from a male index to two female contacts and the subsequent vertical transmission of HIV-1 to their two children. The phylogenetic placement of these transmission sets within the Edinburgh cohort is also assessed.

In section B, four papers are presented which examine aspects of foraminiferal evolution. The first paper (paper I) focuses on the problems inherent in the amplification of foraminiferal DNA due to the association of large numbers of symbionts, commensals and food particles with each foraminifer. The amplification of foraminiferal sequences for the small subunit ribosomal RNA gene is then described, and the phylogenetic placement of the foraminifera within eukaryote evolution examined (papers II and III). Finally, the phylogenetic relationships within the foraminifera, in particular planktic foraminiferal evolution and the relationships between benthic and planktic foraminiferal species, are described (paper V).

DECLARATION:

Personal Contribution to Papers Presented in this Thesis

Section A: Genetic Analysis of HIV-1

Paper I

Leigh Brown, A. J., D. Lobidel, C. M. Wade, S. Rebus, A. N. Phillips, R. P. Brettle, A. J. France, C. S. Leen, J. McMenamin, A. McMillan, R. D. Maw, F. Mulcahy, J. R. Robertson, K. N. Sankar, G. Scott, R. Wyld, and J. F. Peutherer. 1997. The Molecular Epidemiology of Human Immunodeficiency Virus Type 1 In Six Cities in Britain and Ireland. Submitted for publication.

This paper represents a large collaboration over several years. In this project I performed some PCR amplification and sequencing, and all analyses on the sequence data (phylogenetic tree construction, calculation of intra and inter sample nucleotide distances, amino acid sequence comparisons and principal coordinates analysis). I also contributed the description of some methods to the final paper.

Paper II

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.

This paper presents collaborative work with L. Briant and M. Guyader of CHU Purpan, Toulouse, France. In this work, I performed all analyses of sequence data, including sequence alignment, phylogenetic tree construction, intra and inter sample and intra and inter patient distance comparisons and amino acid analyses. I also wrote a substantial part of the paper in conjunction with L. Briant and M. Guyader.

Paper III

Salvatori, F., S. Masiero, C. Giaquinto, C. M. Wade, A. J. Leigh Brown, L. Chieco-Bianchi, A. De Rossi. 1997. Evolution of human immunodeficiency virus type 1 in perinatally infected infants with rapid and slow progression to disease. J. Virol. 71:4694-4706.

This paper represents a collaboration with F. Salvatori and A. De Rossi of the Institute of Oncology, University of Padova, Italy. The majority of the work was carried out by the group in Italy. I performed the analyses of sequence data from mother-child pairs. This included

sequence alignment, phylogenetic tree construction and distance calculations.

Paper IV

Wade, C. M., D. Lobidel, and A. J. Leigh Brown. 1997. Analysis of human immunodeficiency virus type 1 *env* and *gag* sequence variants derived from a mother and two vertically infected children provides evidence for the transmission of multiple sequence variants. Submitted for publication.

and

Paper V

Wade, C. M., D. Lobidel, J. R. Robertson, J. Y. Q. Mok, D. Yirrel, and A. J. Leigh Brown. 1997. Evolution of human immunodeficiency virus type 1 during infection of a male index and transmission to two infected female partners and their children. Submitted for publication.

In these studies I performed almost all the PCR and nucleotide sequencing from viral nucleic acid extractions prepared by D. Lobidel, D. Yirrel or M. Aldhous. I performed all sequence assembly, alignment and analyses. This included all phylogenetic tree construction, intra and inter sample and intra and inter patient distance comparisons, synonymous and nonsynonymous distance estimation, amino acid sequence analyses. I wrote all drafts of the papers, receiving comments and corrections from A. J. Leigh Brown.

Section B. Molecular Evolution of the Planktic Foraminifera

Paper I

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996a. The isolation and amplification of the 18S ribosomal RNA gene from planktonic foraminifers using gametogenic specimens. *In* Whatley, R. C., and Moguilevsky, A. eds. Microfossils and Oceanic Environments. University of Wales, Aberystwyth Press. Chapter 3.1, pp. 249-259.

My involvement in this paper was in the choice of sequence region to examine for phylogenetic comparison of the foraminifera and in the development of the laboratory methods.

Paper II

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996b. Molecular evolution of planktic foraminifera. J. Foram. Res. 26:324-330.

As for paper I. I also carried out all the sequence assembly, alignment and sequence analyses.

The paper was written jointly by K. F. Darling and myself.

Paper III

Wade, C. M., K. F. Darling, D. Kroon, and A. J. Leigh Brown. 1996. Early evolutionary origin of the planktic foraminifera inferred from SSU rDNA sequence comparisons. J. Mol. Evol. 43:672-677.

For this work I performed the alignment and phylogenetic analysis of all sequence data. I wrote all drafts of the paper, with corrections from K. F. Darling and A. J. Leigh Brown.

Paper IV

Darling, K. F., C. M. Wade, D. Kroon, and A. J. Leigh Brown. in press. Planktic foraminiferal molecular evolution and their polyphyletic origins from benthic taxa. Marine Micropaleontology (in press).

As for paper II.

Paper V

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996c. Reading the history of the oceans in plankton DNA. NERC News Autumn 1996: 16-17.

NERC news article summarising the foraminiferal work to date was written by all authors jointly and rewritten by NERC news staff prior to publication.

Christopher Mark Wade

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Andrew Leigh Brown for his help and support throughout this project. I would also particularly like to thank Dr. Kate Darling for her guidance, advice and encouragement in the foraminiferal studies and Denis Lobidel for his support in the HIV studies. Thanks also to Dr. Dick Kroon, Dr. Eddie Holmes, Dr. Lin Qi Zhang, Sandy Cleland, Sarah Ashelford, Marian Aldhous, Sarah Lockett, Dr. Alicia Alonso, Dr. Marilyn Moore, Sharon Hutchinson, John Mokili, and everyone at the Centre for HIV Research.

Finally, special thanks to Karen Bowman and my mum and dad for all their encouragement.

LIST OF ABBREVIATIONS

ABI	Applied Biosystems
AIDS	acquired immunodeficiency syndrome
ARC	AIDS related complex
ARV	AIDS-associated retrovirus
bp	base pairs
BP	before present
BSA	bovine serum albumin
CDC	Center for Disease Control
cDNA	complementary DNA
CTAB	cetyltrimethylammonium bromide
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
DTT	dithiothreitol
env	envelope gene
EDTA	diaminoethanetetra-acetic acid
gag	group antigen
FCS	foetal calf serum
FIV	feline immunodeficiency virus
gp	glycoprotein
HIV	human immunodeficiency virus
kb	kilobase pairs
kDa	kilodalton
LAV	lymphadenopathy associated virus
LSU	large subunit
LTR	long terminal repeat
mRNA	messenger RNA
MYR	million years
nef	negative effector gene
nt	nucleotide
p	protein
PBMC	peripheral blood mononuclear cell

PCR	polymerase chain reaction
pol	polymerase gene
rev	regulator of expression of the virion gene
rDNA	ribosomal DNA
RNA	ribonucleic acid
rRNA	ribosomal RNA gene
RRE	*rev* responsive element
RT	reverse transcriptase
SDS	sodium dodecyly sulphate
SIV	simian immunodeficiency virus
SSU	small subunit
Taq	*Thermus aquaticus*
tat	trans activator gene
TBE	tris/borate EDTA
TEMED	N-N-N'-N'-tetra-methyl-1,2-diamino-ethane
vif	viral infectivity factor gene
vpr	viral protein r gene
vpu	viral protein u gene

Section A

Genetic Variation of HIV-1

GENERAL INTRODUCTION

The human immunodeficiency virus (HIV) causes a chronic debilitating infection which leads to severe immunosuppression, culminating as acquired immunodeficiency syndrome (AIDS), which ultimately results in death. The World Health Organisation has estimated that by December 1995 more than 16.8 million adults and 1 million children were infected with HIV (World Health Organisation, 1995) and it has been estimated that by the year 2000, 30-40 million people may become infected with HIV worldwide, of which over 90% will be found in developing countries (World Health Organisation, 1992).

1. Acquired Immunodeficiency Syndrome (AIDS): Identification of a New Clinical Syndrome

The first cases of a new clinical syndrome, subsequently termed acquired immunodeficiency syndrome (AIDS) (Centers for Disease Control, 1982a), were reported in 1981. Severe opportunistic infections (*Pneumocystis carinii* pneumonia and Kaposi's sarcoma) were reported amongst previously healthy homosexual men in Los Angeles and New York in the United States of America (Centers for Disease Control, 1981a and 1981b). Soon after these initial reports, a number of AIDS cases were reported among intravenous drug users (Masur *et al.*, 1981), haemophiliacs (Centers for Disease Control, 1982a), transfusion recipients (Centers for Disease Control, 1982b), female partners of infected individuals (Centers for Disease Control, 1982c) and children born to infected mothers (Centers for Disease Control, 1982d). AIDS cases have subsequently been reported worldwide and it soon became apparent that the major epidemic of the disease was in Central Africa. It is evident that AIDS clearly existed before the first reported cases in 1981 and the first retrospective evidence of infection has been provided by a serum specimen retrieved from archive frozen material taken from a patient in 1959 from Kinshasha in the Belgian Congo (now Zaire) (Nahmias *et al.*, 1986). Many other samples from the same archive were however negative and evidence for widespread infection in Africa before the mid 1970s is weak.

2. Human Immunodeficiency Virus (HIV): Identification of the Causative Agent of AIDS

The causative agent of AIDS, the human immunodeficiency virus (HIV), was first identified in 1983 by researchers in the Pasteur Institute, France, with the isolation of a new retrovirus, named lymphadenopathy virus (LAV), from the lymph node of a patient with

AIDS (Barré-Sinoussi *et al.*, 1983). The following year Gallo *et al.* (1984) reported the isolation of a retrovirus, named human T-cell leukaemia virus III (HTLV-III), from peripheral blood mononuclear cells (PBMCs) of patients with AIDS, and Levy *et al.* (1984) reported the isolation of a retrovirus, named AIDS-associated retroviruses (ARV). The three prototype viruses, LAV, HTLV-III and ARV, were soon recognised as variants of the same virus, which was distinct from HTLV, and in 1986 the International Committee on Taxonomy of Viruses gave the AIDS virus a separate name, human immunodeficiency virus (HIV) (Coffin *et al.* 1986). Subsequently, Clavel *et al.* (1986) isolated a new type of HIV from West African patients with AIDS, HIV-2. The first HIV type identified is now referred to as HIV-1. Immunodeficiency viruses have subsequently been identified in a number of other mammalian groups and include the simian (SIV), bovine (BIV) and feline (FIV) immunodeficiency viruses.

3. Classification of HIV

The human immunodeficiency viruses fall within a large group of immunodeficiency viruses, the lentiviruses, which belong to the retrovirus family (Retroviridae). Retroviruses have a characteristic life-cycle in which their single stranded RNA genome is reverse transcribed into a DNA copy by the viral-encoded enzyme, reverse transcriptase (Weiss *et al.*, 1985), upon entry of free virus into a host cell. This DNA copy is then integrated into the host genome as a provirus which is then replicated along with the host genome by the cellular enzyme DNA polymerase and may be transcribed to generate both the viral mRNA's and the full length genomic RNA molecule by the cellular enzyme RNA polymerase.

The retrovirus family is divided into three subfamilies, the oncoviruses, which include oncogenic viruses of animals (eg. human and bovine T cell leukaemia viruses (HTLV I and II), Rous sarcoma virus, murine and avian leukaemia viruses); the spumaviruses, which produce a 'foamy' cytopathic effect *in vitro* but have not been associated with disease in the original host (eg. human foamy virus); and the lentiviruses, which may cause a chronic infection in which symptoms of disease emerge after a long incubation period (eg visna-maedi virus of sheep, equine infectious anaemia virus (EIAV), caprine arthritis encephalitis virus (CAEV), and the immunodeficiency viruses). Lentiviruses are distributed widely in mammals causing chronic disease affecting the lungs, joints, nervous, haematopoietic and immune systems (Weiss *et al.*, 1985).

4. Evolutionary Origins of HIV

The closest evolutionary relatives of HIV are the non-human primate lentiviruses, the simian immunodeficiency viruses (SIVs). Sequence analyses have revealed the close relationship of HIV-2 to virus isolates from the sooty mangabey (SIV_{sm}) (Hirsch *et al.*, 1989; Gao *et al.*, 1992; Sharp *et al.*, 1994) and rhesus macaque (SIV_{mac}) (Chakrabarti *et al.*, 1987; Sharp *et al.*, 1994). The absence of immunodeficiency virus infection in wild Asian rhesus macaques (as opposed to captive macaques) makes the macaque an unlikely candidate as the source of HIV-2 infection. However, the close contact of the West African sooty mangabey species with humans, coupled with the occurrence of HIV-2 as the West African form of the virus, suggests the origin of HIV-2 by a transspecific infection from the sooty mangabey (Leigh Brown and Holmes, 1994). Further support is afforded to this hypothesis by the isolation of human virus isolates more closely related to SIV_{sm} than HIV-2, suggesting that the sooty mangabey is a natural reservoir and that human infection probably represents a zoonosis (Gao *et al.*, 1992). HIV-1 shows an apparent relationship with viral isolates from the chimpanzee (SIV_{cpz}) (Huet *et al.*, 1990; Sharp *et al.*, 1994) although the recent discovery of HIV-1 isolates more divergent from other HIV-1 sequences than SIV_{cpz} raises doubts to a SIV_{cpz} ancestor of HIV-1 (Leigh Brown and Holmes, 1994). This may suggest a more distant origin of HIV-1, perhaps as a branch of the African green monkey (SIV_{agm}) group (Fukasawa *et al.*, 1988). SIV isolates from the African green monkey show a high degree of diversity and the occurrence of viral isolates with mosaic genomes have indicated ancient recombination between the lineages found in African green monkeys and sooty mangabeys. This may indicate that the origins and diversity of the primate lentiviruses lie within the African green monkey group (Sharp *et al.*, 1994).

5. Structure of the HIV-1 Virion and Genomic Organisation

HIV is a small enveloped virus with the virion approximately 100nm in diameter. The genome consists of two positive stranded RNA molecules which are surrounded by a cone shaped protein inner core or capsid (CA) composed of the viral p24 *gag* protein. The viral RNA dependent DNA polymerase (*pol* p51 and p56 subunits; the reverse transcriptase) and the nucleocapsid proteins p9 and p6 are closely associated to the genomic RNA molecule contained within the p24 *gag* protein capsid. The capsid is surrounded by a membrane-associated matrix protein (MA), the p17 *gag* protein, which provides the matrix for the viral structure and is thought to stabilise both the capsid and the viral envelope and is vital for

integrity of the virion. The viral envelope is a lipid-bilayer derived from the host cell during the process of viral budding and contains the virus-encoded envelope proteins. The envelope is characteristically made up of 72 knobs containing trimers or tetramers of the envelope glycoproteins. They are derived from a 160 kDa precursor protein, gp160, which is cleaved to produce the external surface (SU) envelope glycoprotein, gp120, and a transmembrane (TM) protein, gp41. gp120 contains the binding sites for the cellular receptors and the major neutralising domains. The structure of the HIV-1 virion is detailed in Levy *et al.* (1993). A schematic representation of the virus is shown in Figure 1.

The HIV-1 genome is 9.7 kb in length and comprises a variety of viral genes which code for at least nine viral proteins. In addition to the genes encoding for the three structural proteins (*env*, *gag* and *pol*), which are common to all replicating retroviruses, HIV also contains an additional six (possibly eight) open reading frames which code for both regulatory proteins (*tat* and *rev*) and accessory proteins (*vif*, *vpr*, *vpu* and *nef*). In the proviral form, long terminal repeats (LTRs), which contain the promoter and enhancer elements involved in the regulation of viral gene expression, flank these nine genes. A schematic representation of the genome is shown in Figure 2.

5.1. HIV-1 Structural Genes

env

The *env* (envelope) gene encodes the surface (SU) and transmembrane (TM) envelope glycoproteins. Translation of the *env* gene yields a 160kDa precursor protein, gp160, which is cleaved by cellular proteases into two components, gp120, the external glycoprotein, and gp41, the transmembrane protein.

gag

The *gag* (group specific antigen) gene encodes the structural proteins of the viral core. Translation of the *gag* gene results in the generation of a precursor protein p55 which is cleaved by viral protease into four proteins, p17 (matrix protein), p24 (capsid protein), p7 and p6.

pol

The *pol* (polymerase) gene encodes the protease, reverse transcriptase and integrase viral enzymes. Translation of the *pol* gene results in the formation of the *pol* p66 polyprotein which is then cleaved into its individual products by viral protease to yield, the p22 protease, the p66 and p51 reverse transcriptase subunits, and the p32 integrase.

5.2. HIV-1 Regulatory Genes (for review see Cullen *et al.*, 1992)

tat

The *tat* (trans activator) gene is encoded by two exons. The translated protein is essential for viral replication and it is responsible for initiating and/or stabilising the elongation of primary HIV-1 mRNA transcripts by binding to a viral RNA target sequence (TAR) located in the LTR.

rev

The *rev* (regulator of expression of the virion) gene is also encoded by two exons and is translated into p19. The function of *rev* seems to be the promotion of nuclear export and cytoplasmic expression of the mRNA's which encode the viral structural proteins. Thus it plays a crucial role in the switching from "early regulatory" to "late structural" gene expression. *Rev* function is mediated by direct binding to an RNA structure, the rev responsive element (RRE), located within the transcribed *env* gene.

5.3. HIV-1 Accessory Genes (for review see Subbramanian and Cohen, 1994)

vif

The *vif* (viral infectivity factor) gene encodes a 23 KDa protein which is thought to play a role during virus assembly or maturation.

vpr

The *vpr* (viral protein r) gene encodes a 14 KDa protein which has two functions; it is necessary for replication in macrophages (early function) (Connor *et al.*, 1995) and also blocks lymphocytes in G2 of the cell cycle which results in apoptosis (He *et al.*, 1995)

vpu

The *vpu* (viral protein u) gene encodes a protein which is important for the proper maturation of virions and their efficient release or export from cells. It has also been demonstrated to degrade CD4 in the endoplasmic reticulum and has also been shown to decrease syncytium formation between infected cells.

nef

The *nef* (negative effector) gene encodes a 27 KDa protein which is important for viral replication. The protein is a downregulator of CD4 expression.

6. Life Cycle of HIV-1

Human immunodeficiency virus infection is initiated by successful binding of the

viral envelope glycoprotein, gp120 to the CD4 receptor on the surface of the $CD4^+$ cell (Dalgleish *et al.*, 1984; Klatzmann *et al.*, 1984). Following attachment, the envelope proteins mediate the fusion between virus and host cell membranes and the virus is internalised into the cell. Although human CD4 is sufficient for binding HIV-1 to cells it is not sufficient for HIV-1 penetration or for fusion of the viral envelope with the host cell membrane (Maddon *et al.*, 1986). Recently, secondary receptors of the 7tm type have been identified as coreceptors for HIV-1 infection (Weiss and Clapham, 1996). Non syncytium inducing (NSI) isolates of HIV-1 were shown to utilise a second receptor, CKR-5 (Deng *et al.*, 1996; Dragic *et al.*, 1996; Alkhatib *et al.*, 1996), a receptor for the attractant molecules (ß-chemokines) known as RANTES, MIP-1α and MIP-1ß, which are produced by $CD8^+$ T lymphocytes and have been shown to inhibit HIV-1 infection (Cocchi *et al.*, 1995). Another receptor, fusin, has been shown to act as the coreceptor for syncytium inducing (SI), T cell line adapted isolates of HIV-1 (Feng *et al.*, 1996). CD4 independent entry also appears to be possible due to the fact that a large number of cells lacking the CD4 receptor have been shown to be infected with HIV (Levy *et al.*, 1993).

Following fusion of the HIV-1 virion and host cell, the viral core is released into the cytoplasm of the host cell and the viral genome is reverse transcribed into a linear DNA copy by virus encoded *pol* reverse transcriptase. The reverse transcription process is highly error prone and leads to a high number of misincorporations, the creation of insertions, deletions and repetitions and recombination (see section 8.1). The viral double stranded DNA molecule is then translocated to the cell nucleus and the viral DNA is inserted into the cell genome by the viral *pol* encoded enzyme, integrase. Covalently closed circular viral DNA molecules have also been detected in addition to the linear form. Following integration, the virus may either remain latent or be expressed. In the active phase, viral DNA is transcribed into mRNA by the host cell enzyme RNA polymerase using signals in the LTR which are similar to those used by the cell for making it's own RNAs. Proviral transcription is regulated both by cellular transcription factors and the viral encoded *tat* and *rev* proteins (see section 3). *Rev*, in particular, plays a crucial role in switching transcription from "early" regulatory genes to "late" structural genes. Virion assembly occurs at the cell surface where the full length HIV-1 sense RNA strands assemble with the structural proteins. The virion then buds from the cell, acquiring its outer lipid membrane as it passes through the cytoplasm and *env* gene products, gp41 and gp120. The replication cycle of HIV-1 is illustrated in Figure 3.

7. Clinical Phases of HIV Infection

The course of HIV infection can be divided into three main phases with five clinical stages recognised.

7.1. Phase I - Primary Infection

Infection with HIV begins with primary infection. This period is characterised by rapid viral replication during which the virus establishes itself throughout the body. In the majority of cases (50-70% of infected individuals) primary infection is accompanied by an acute clinical syndrome ("seroconversion illness") which may include fever, rash, swelling of the lymph glands, ulcers in the mouth and sometimes genitalia, gastrointestinal tract disorders (anorexia, nausea, vomiting and diarrhoea), and not infrequently the nervous system may be affected (meningitis and encephalitis). There may be a profound suppression of the immune system which may be manifest by severe opportunistic infections, particularly thrush. This "seroconversion illness" coincides with the infected individual becoming positive for HIV antibodies, typically 2-4 weeks after infection, and is of approximately 2 weeks duration.

7.2. Phase II - Asymptomatic Infection

Following seroconversion, there follows a rapid decline in plasma viraemia and the patient then enters a period of clinical latency, the silent phase of HIV infection. During this period, viral infection appears to be confined to the lymphoid tissues and peripheral blood. The viral load does however increase during the asymptomatic period and the number of CD4 T cells slowly declines. Ultimately signs of immunological suppression start appearing and clinical symptoms develop and the patient then advances to the next stage. The asymptomatic phase of HIV infection lasts on average for a period of approximately 10 years.

7.3. Phase III - Symptomatic Infection

Following the asymptomatic phase, the patient begins to show clinical symptoms of HIV infection. The lymph nodes are usually the first tissue to show clinical signs of HIV infection, with the enlargement and swelling of lymph nodes, particularly in the head and neck region. This stage is referred to as progressive generalised lymphadenopathy (PGL). However, since lymphadenopathy is not specific for HIV infection, the clinical syndrome is not on its own a diagnostic criterion for AIDS. At a somewhat later stage, further clinical signs of HIV infection may become evident, including weight loss, diarrhoea, fever and

recurrent thrush infections of the oral and genital mucous membranes. Since these infections are not on their own a diagnostic criterion for AIDS, this clinical stage of HIV infection is often referred to as pre-AIDS or AIDS-related complex (ARC). The final clinical stage of HIV infection, AIDS, is diagnosed when the minimal criteria, depending on the classification system used, become manifest in the patient. The principal feature of AIDS is the dramatic reduction in the number of CD4 lymphocytes; the Centers for Disease Control has recently expanded its definition of AIDS to include HIV-infected patients with CD4 lymphocytes counts of less than 200 per µl. The profound immunodeficiency observed in AIDS results in the appearance of severe opportunistic infections and the development of neoplasms and death usually results within two years of developing AIDS. During the symptomatic phase of HIV infection, HIV replicates to high titre and high levels of virus can be detected in both lymphoid and non-lymphoid tissues.

8. HIV Sequence Variation and Host Interaction

8.1 Sources of Sequence Variation in HIV-1

The HIV genome is characterised by extensive genetic diversity both between and within patients (Alizon *et al.*, 1986; Hahn *et al.*, 1986; Starcich *et al.*, 1986; Saag *et al.*, 1988). The extreme variability of the genome is jointly attributable to both the high rate of virus replication and the error-prone nature of this process, particulary as a consequence of the viral enzyme reverse transcriptase. The viral population within an infected individual shows a remarkably high viral turnover, with estimates of the half-life of an evolving population *in vivo* (obtained by treating patients with highly potent inhibitors of the reverse transcriptase and protease) of approximately 2 days (Ho *et al.*, 1995; Wei, *et al.*, 1995; Wain-Hobson, 1995). The viral reverse transcriptase lacks 3' to 5' proof-reading exonuclease activity and has an exceptionally high misincorporation rate, with the average error rate per nucleotide incorporated estimated at 1/1700 (Roberts *et al.*, 1988) and 1/2000 to 1/4000 (Preston *et al.*, 1988). The creation of insertions, deletions and repetitions is also common. An additional mutational process, hypermutation, which involves monotonous base substitutions (typically G to A transitions) (Vartanian *et al.*, 1991) and appears to be a property of the reverse transcriptase has also been described. A further source of mutation is the cellular enzyme RNA polymerase II, another error-prone polymerase, which transcribes the provirus into RNA during the life cycle. Finally, an additional source of diversity within

the HIV genome is recombination (Howell *et al.*, 1991; Simmonds *et al.*, 1991; Coffin, 1992; Robertson *et al.*, 1995). This can occur when a cell is infected with two genetically distinct viruses with template switching by the reverse transcriptase enzyme during reverse transcription leading to the generation of a 'hybrid' provirus.

8.2. The HIV Quasispecies

Natural selection clearly plays an important role in the evolution of HIV and the interaction of high mutation rates and selection leads to what has been described as the HIV quasispecies; a mathematically-defined distribution of mutants generated by a mutation-selection process (Eigen and Winkler-Oswatitsch, 1990; Holland *et al.*, 1992; Nowak, 1992). In circumstances where there are no or very few errors during replication, Darwinian selection will lead to a homogeneous population consisting of the fastest replicating variant (the fittest variant) from amongst a pool of mutants with different replication rates. However, in HIV and other RNA viruses which display a high error rate, a distribution of mutants (the "quasispecies") is generated which will eventually reach an equilibrium at which a consensus sequence ("master variant") is produced. Thus, the HIV quasispecies represents a cloud of closely related but different variants around a central point, the variant with the highest fitness, the "master variant". These variants have different fitnesses and therefore different replication rates. The "master variant" may be present at a low frequency within the viral population but there may be many variants in the pool of variants that are close in terms of mutational distance to the "master variant". The fitnesses of variants will be increased by their proximity in mutational distance to the fittest form, even if their own intrinsic fitness is significantly lower. Due to the high mutation rate of HIV, the frequency of mutations required for these closely related variants to form the "master variant" will not be negligible and therefore the population of sequences around the "master variant" will replicate the most rapidly. Strong stabilising selection acts to constrain the spectrum of variation within the population countering the diversifying effect of the high mutation rate. Thus, HIV can be considered as consisting of a pool of variants within an infected individual. This provides the potential for rapid adaptation by selection for preexisting variants from the virus pool which forms the quasispecies. Human immunodeficiency virus can therefore quickly adapt to overcome new immune responses or rapidly develop resistance to antiretroviral therapies.

8.3. Genetic Variation of HIV-1

Human immunodeficiency virus type 1 shows considerable variation throughout the whole genome. Variation is not however uniform across the genome, with the *env* gene, encoding the envelope membrane proteins, showing considerably more variability, particularly in the hypervariable regions (14×10^{-3} nonsynonymous nucleotide substitutions per site per year) than the *gag* gene, which encodes the structural proteins of the virion, and the *pol* gene, which encodes the viral enzymes ($\approx 1.7 \times 10^{-3}$ nonsynonymous nucleotide substitutions per site per year) (Li *et al.*, 1988). This indicates that the *env* gene is evolving under different selection pressures from the rest of the genome. Whereas selection in the *gag* and *pol* genes appears to act against nonsynonymous variation, selection against amino acid change (negative selection), selection in the *env* gene appears to be adaptive, selection for amino acid change (positive selection) (Li *et al.*, 1988; Leigh Brown and Monaghan, 1988; Holmes *et al.*, 1992; Shpaer *et al.*, 1993). Comparison of the ratio of synonymous (silent) to nonsynonymous (amino acid changing) substitutions (d_s/d_n ratio) has been used to estimate a measure of the selection pressure on the assumption that synonymous substitutions are not affected by selection. A d_s/d_n ratio of 1 therefore indicates neutrality, a ratio of >1, as observed in the *gag* and *pol* genes, indicates purifying negative selection (amino acid change is selected against), and a ratio of <1, as observed in the *env* gene indicates positive selection, favouring new amino acid sequences. The relative contribution of genetic drift and natural selection to the evolution of HIV-1 is not however entirely clear. The predominance of synonymous over nonsynonymous substitutions in certain genes has been argued to provide support for the neutral theory (Gojobori *et al.*, 1990). Coffin (1992) has however argued that even small, nearly neutral selection pressures will have a large impact on the accumulation of mutations.

The pattern of evolution within the *env* gene can be explained by considering its structure and function. The gene encodes the 160kDa precursor protein, gp160, which is cleaved to produce a 41kDa transmembrane protein, gp41, and a 120kDa external protein, gp120. gp120 is responsible for mediating attachment to the cellular virus receptor (the CD4 molecule) (Lasky *et al.*, 1987; Jameson *et al.*, 1988) and encodes viral determinants for cell tropism and cytopathicity (Hwang *et al.*, 1991). It also forms the major target for neutralising antibodies (Rusche *et al.*, 1988). Variability within gp120 is not uniform and the region is divided into five major hypervariable regions, termed V1 to V5, which are interspersed with highly conserved regions and regions of intermediate variability, termed C1 to C4 (Modrow *et al.*, 1987). Of particular importance is the third hypervariable region, the V3 region, which

encodes the principal neutralisation determinant of HIV-1 (Rusche *et al.*, 1988; Javaherian *et al.*, 1989), an approximately 35 amino acid disulfide-bridged loop structure referred to as the V3 loop. The V3 region has also been identified as the principal determinant of cell tropism (Hwang *et al.*, 1991) and has been implicated as the principal region involved in the change from the slow-replicating, macrophage tropic, non-syncytium inducing (NSI) form to the fast-replicating, T-cell tropic, syncytium inducing (SI) form (Fouchier *et al.*, 1992; Milich *et al.*, 1993). As a consequence of its structure and function, evolution within the *env* gene, in particular V3, is thus a complex interplay between selection for variability and for conservation. Although there is positive selection for the replacement of amino acids which confer may affect immune recognition, there is also a selective constraint as to which amino acids are functionally viable (Holmes *et al.*, 1992).

9. HIV-1 Variation and the Evolution of HIV-1 Within a Single Patient During the Course of Infection

Three distinct phases of viral evolution have been identified within an infected patient during the course of natural infection with HIV-1 (Leigh Brown, 1991; Leigh Brown and Holmes, 1994). These stages are coincident with the three clinical phases of HIV infection (acute infection, asymptomatic infection and symptomatic infection) described in section 7.

9.1. Phase I - HIV-1 Transmission and Primary Infection

During the first phase of viral evolution, the period of time between infection and seroconversion, a huge expansion of the viral population occurs, with the viral titre in the plasma increasing from a few hundred particles at the time of infection to 10^8 or more (Zhang *et al.*, 1991) at seroconversion. The viral population of the infected patient has generally been shown to display a restricted level of sequence diversity during this period, particularly within the *env* gene, and such restricted levels of sequence diversity have been reported following infection via sexual, parenteral and vertical transmission routes (McNearney *et al.*, 1990; Wike *et al.*, 1992; Wolfs *et al.*, 1992; Wolinsky *et al.*, 1992; Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Zhang *et al.*, 1993; Zhu *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995). Such restricted levels of sequence diversity are in marked contrast to the heterogeneous sequence population typically observed within *env* in the long term infected donor and it has therefore been postulated that a sequence bottleneck occurs upon

infection, with infection initiated by a limited number of variants or even one particular variant. The viral homogeneity shortly after infection may be explained by a dilution model in which the inoculum is small and thus represents only a limited sample of the heterogeneity within the long term infected patient. The virus population observed within the recipient would then represent the progeny of only a single infectious unit. Alternatively, selective penetration may lead to a homogeneous virus population within the recipient due to certain variants preferentially infecting the cells lining the place of entry. The sequence heterogeneity in *env* is however initially lower than other regions of the genome and p17 *gag* gene sequences were not found to show such a restricted level of sequence diversity within the recipient upon seroconversion (Zhang *et al.*, 1993; Zhu *et al.*, 1993). This suggests that there is some heterogeneity within the inoculum but there is strong selection for specific *env* sequences in the interval between exposure and seroconversion (Zhang *et al.*, 1993). The rapid increase in viral population size coupled with the general sequence similarity in *env* would imply strong selection for the most rapidly replicating viruses (which appear to be non syncytium inducing macrophage tropic HIV-1 strains) during this period of primary infection (Leigh Brown, 1991; Leigh Brown and Holmes, 1994). Such strong selection would be expected to result in a reduction in sequence diversity across the *env* gene due to genetic hitch hiking (Leigh Brown and Holmes, 1994). The diversity observed in the p17 region of *gag* would then be explained by recombination between variants from a mixed pool of transmitted or newly arising variants (Leigh Brown and Holmes, 1994).

The transmission of more than one variant appears to be rare, although transmission of multiple variants has been reported in a small number of cases for both adults (one patient of the florida dentist case (Ou *et al.*, 1992, Korber and Myers, 1992); the victim in the Swedish rape case (Albert *et al.*, 1993); evidence for coinfection with multiple HIV-1 strains within an Australian homosexual male (Zhu *et al.*, 1995); and with heterogeneous virus populations also reported shortly after seroconversion within cervical secretions and/or peripheral blood in five women of a cohort of six Kenyan female sex workers (Poss *et al.*, 1995)) and some vertically infected children (Lamers *et al.*, 1994; Van't Wout *et al.*, 1994+; Briant *et al.*, 1995).

9.2. Phase II - Asymptomatic Phase

Following seroconversion, infectious virus is cleared from the peripheral circulation once the immune response, including both cytotoxic T lymphocytes (CTL) and humoral

responses, develops fully. The patient then enters the asymptomatic phase of infection. Despite the presence of a strong anti-viral immune response, there is continual viral replication in the peripheral blood and lymph nodes during this phase (Zhang *et al.*, 1991; Piatak *et al.*, 1993) and the virus population within the infected individual has been demonstrated to show a high degree of heterogeneity (Simmonds *et al.*, 1990; Wolfs *et al.*, 1990; Cichutek *et al.*, 1991; Simmonds *et al.*, 1991; Holmes *et al.*, 1992; McNearney *et al.*, 1992; Wolfs *et al.*, 1992; Lukashov *et al.*, 1995). Viral evolution during this period is characterised by a continual process of virus neutralisation and escape (Holmes *et al.*, 1992; Leigh Brown and Holmes, 1994), a selective process in which viral variants recognised by the immune system are replaced by antigenically distinct variants arising in the viral population of the patient, thus evading the host's immune response ("antigenic drift"). The higher the frequency a variant reaches in the viral population, the greater the probability of its immune recognition and neutralisation. Thus variants found at a low frequency will have a greater selective advantage and hence will increase in frequency until they also reach a level at which they are recognised and countered by the immune system.

The successive replacement of viral variants during HIV-1 phase II infection has clearly been demonstrated by Simmonds *et al.* (1991) and Cichutek *et al.* (1991). Simmonds *et al.* (1991) showed that V4 and V5 *env* sequence variants identified in the plasma at seroconversion were not detected in the plasma three years later but did persist within the proviral DNA population, thus indicating that the replacement of variants occurs more rapidly in the plasma with variants appearing to arise first in the plasma and then enter the PBMCs where they reside for longer periods of time. Cichutek *et al.* (1991) also clearly demonstrated the successive replacement of viral sequence variants through analyses of V1 and V2 *env* sequences in an infected haemophiliac. Variants found 11 months after seroconversion were found to replace those found 5 months after seroconversion. The most comprehensive study of viral evolution within an infected patient has however been performed by Holmes *et al.* (1992) who examined V3 sequence variants observed over the course of infection in an infected haemophiliac. This patient showed a single V3 loop sequence at seroconversion which was replaced by two distinct lineages at three years post seroconversion. These lineages persisted for the remainder of the infection but differed in frequency in successive years thus indicating that antigenic drift may involve a complex interplay between the different and competing lineages in the viral population of the patient instead of the sequential replacement of one antigenically distinct variant by another.

9.3. Phase III - Symptomatic Phase

The third phase of HIV evolution begins as the immune system collapses and the patient progresses to AIDS. During this period there is a progressive increase in viral load, presumably due to inefficient immune function. HIV infection spreads to many previously uninfected tissues and organs of the body including tissues of non-lymphoid origin such as the brain, lung and muscle.

Nowak *et al.* (1990) have proposed that the collapse of the immune system and consequent progression to AIDS is due to an asymmetric interaction between the HIV quasispecies and the $CD4^+$ T cell population. In this, $CD4^+$ T cells are directed against a specific HIV antigen whereas each virus strain can kill any $CD4^+$ cell. Nowak *et al.* (1990) has therefore predicted the existence of an "antigenic diversity" threshold below which the immune system can regulate the viral population but above which the immune system collapses and the patient progresses to AIDS. The existence of a threshold in viral diversity, coinciding with the rapid decline in $CD4^+$ cells and the onset of AIDS has nevertheless recently been questioned by viral sequence analyses (Wolinsky *et al.*, 1996; Miedema and Klein, 1996) and Wolinsky *et al.* (1996) have shown that in four patients progressing to AIDS, the two patients who progressed the fastest showed high levels of plasma virus with very limited viral diversity, whereas the two slowly progressing individuals showed high levels of viral diversity.

A reduced level of viral diversity would be expected to be observed during phase III with, similarly to phase I, the absence of an effective immune response leading to the predominance of the fastest replicating viruses within the patient's viral population.

10. Geographical Variation of HIV-1

10.1. Global Variation and Genetic Subtypes of HIV-1

Genetic analysis of global isolates of HIV-1 has identified two genetically distinct groups, the main group (M) and an outlier group (O) (Gurtler *et al.*, 1994; Myers *et al.*, 1995). Group M is the most prevalent and the subtypes of this group show a worldwide distribution. At least ten, approximately equidistant group M subtypes (A-J), and a number of as yet unclassified sequences have now been identified (Myers *et al.*, 1995; Louwagie *et al.*, 1993; Louwagie *et al.*, 1995). The European and North American virus isolates fall predominantly within subtype B. This subtype is also found in South America and Thailand,

and although subtype B is found in Africa it is rare on this continent. The HIV-1 population within Africa is considerably more diverse than that within Europe and North America with all HIV-1 subtypes presently identified found in Africa. In many African areas, often more than one HIV-1 subtype cocirculates. This may lead to recombination between subtypes in situations when coinfection of patients with more than one strain occurs and indeed a large number of apparently recombinant viruses have now been identified from analyses of published HIV-1 sequences (Robertson *et al.*, 1995; Salminen *et al.*, 1995).

10.2. Molecular Epidemiology of HIV-1

Sequence data have been used to reconstruct the epidemiological relationships of individuals within a community. Holmes *et al.* (1995) used part of the p17 coding region of the *gag* gene to describe the molecular epidemiology of HIV-1 within Edinburgh. The intravenous drug users (IDUs) were identified as forming a distinct infection cluster and patients infected following heterosexual transmission also clustered with the IDUs. Kuiken *et al.* (1993) examined the molecular epidemiology of HIV-1 in the IDU and homosexual risk groups in the Amsterdam cohort using sequences of the V3 region of the *env* gene. The Amsterdam IDU patients clustered within a single group distinct from the other risk groups (Kuiken *et al.*, 1993), in parallel to the situation observed in Edinburgh. The IDU group was consistently differentiated from the other risk groups at only two nonsynonymous nucleotide positions (Kuiken *et al.*, 1994) which suggests that the HIV-1 population in Amsterdam and Edinburgh were founded by a limited number of variants. Consistent risk-group associated differences have been identified within the Amsterdam cohort in analyses of *vpr* and *vpu* sequences supporting the V3 *env* data (Kuiken *et al.*, 1996). The initial molecular epidemiology study carried out by Holmes *et al.* (1995) has now been greatly extended by sequencing p17 *gag* gene sequences from representatives of all infected risk groups in Edinburgh and the inclusion of IDUs and homosexual men from six additional cities in Scotland, Northern England and Ireland (Leigh Brown *et al.*, submitted). HIV-1 sequences did not associate according to the city of sampling but did cluster according to risk group with the IDUs clustering predominantly within a single group in all analyses. This work is presented in paper I of this thesis.

11. Outline of Papers Presented in this Thesis

In this thesis, work is presented which examines aspects of the evolution of HIV-1

both within and between patients. The principal focus of this work has been to examine the transmission of HIV-1, with particular emphasis placed on mother-to-child transmission. The work is presented in the form of a series of papers.

In paper I (Leigh Brown *et al.*, submitted for publication), the molecular epidemiology of HIV-1 is examined through analysis of p17 *gag* gene sequences obtained from 211 seropositive patients from six cities in Scotland, Northern England and Ireland, infected either through injecting drug use or by sexual intercourse between men. Phylogenetic analyses revealed substantial heterogeneity in the sequences obtained from the homosexual men, however in contrast, the majority of sequences obtained from the injecting drug users clustered relatively tightly in the reconstructed phylogeny and were distinct both from sequences of published isolates and of the homosexual men. The analysis revealed no large-scale clustering of sequences by city in either risk group, although a number of close associations between pairs of individuals were observed.

Four papers are presented which examine the transmission of HIV-1 from a molecular evolutionary perspective. Papers II and III present evolutionary analyses of single mother-child transmission pairs.

Paper II (Briant *et al.*, 1995) examines the viral variants involved in the transmission of HIV-1 from mother-to-child in four mother-child transmission sets. The paper presents analyses of the V3 loop and flanking regions of the *env* gene derived from sequential proviral populations obtained from the mothers across pregnancy and from samples obtained from the children following delivery. We report that for one pair, sequences of the child were highly homogeneous and clustered in a single branch within reconstructed evolutionary trees, consistent with the selective transmission of a single maternal variant. However, a high level of sequence variability was observed in the viral populations of the other three children, with sequences of the child clustering in several separate branches in the phylogenetic tree. The data for these three pairs do not appear to be consistent with the selective transmission of a single maternal variant to the child suggesting that in these cases the transmission of several maternal variants may be responsible for initiating infection within the child.

Paper III (Salvatori *et al.*, 1997) addresses the relationship between the origin and evolution of HIV-1 variants and disease outcome in perinatally infected children. The paper examines sequence variations within the V3 loop and flanking regions of the *env* gene for five mother-child transmission pairs comprising 3 infants who progressed slowly to AIDS and two infants who progressed rapidly. For all transmission pairs, sequences of the child were

observed to cluster within a single monophyletic group, indicating that infection within the child was initiated by the transmission of a single maternal variant in both the rapid and slow progressing infants. Plasma HIV-1 RNA levels increased in all 5 infants during their first months of life, and then declined within the first semester of life only in the 3 slow progressors. V3 variability increased over time in all infants, but no differences in the pattern of V3 evolution in terms of potential viral phenotype were observed. The number of synonymous and nonsynonymous substitutions varied during the first semester of life regardless of viral load, CD4+ cell count and disease progression. However, during the second semester of life the rate of nonsynonymous substitutions was higher than that of synonymous substitutions in the slow, but not in the rapid progressors, thus suggesting a stronger host selective pressure in the slow progressing infants.

Papers IV and V examine viral evolution within two transmission sets, each set spanning a period of several years and comprising several patients and thus several distinct transmission events.

Paper IV (Wade *et al.*, submitted for publication) presents an analysis of HIV-1 evolution within a transmission set comprising a mother and two vertically infected children. Genetic variation was examined within the viral population of the mother and her two infected children for both the V3 loop and flanking regions of the *env* gene and the p17 region of the *gag* gene. Viral sequences of one child were highly homogeneous and clustered in a single branch within reconstructed phylogenetic trees, consistent with infection of the child being initiated by a single maternal variant. However, by contrast, substantial genetic heterogeneity was obseved even within the earliest samples obtained for the other child, with sequences of the child clustering in two distinct groups in phylogenetic trees and separated by sequences of the mother. These results are not consistent with the selective transmission of a single maternal variant to the child in this case and it is therefore proposed that the infection within this child is the result of the transmission of multiple sequence variants.

Paper V (Wade *et al.*, submitted for publication) presents an analysis of a transmission set comprising a male index and two female partners infected heterosexually at an approximately eight year interval. Each female partner subsequently gave birth to a vertically infected child. Viral evolution within the transmission set was examined by analysis of sequences of the V3 loop and flanking regions of the *env* gene and the p17 region of the *gag* gene for sequential samples obtained from the patients over a 10 year period. Within the male index, a progressive increase in the mean genetic distance to the seroconversion

sequence was observed over time, with an evolutionary diversification of the plasma viral sequences of the index into three lineages by year 7. Two distinct sequence populations were transmitted to the two female partners and the distance between the two mother-child pairs reflected the level of evolutionary change within the index in the period between the two transmission events. Both sexual and vertical transmission events led to a homogeneous sequence population within the newly infected individual immediately following infection. The phylogenetic placement of the sequences of the transmission set within a background population of Edinburgh patients revealed that the early sequences of the male index showed a closer relationship to a number of Edinburgh intravenous drug user sequences than to sequences obtained from the index later in the course of infection or indeed sequences obtained from the female transmission contacts.

13. References

Ahmad, N., B. M. Baroudy, R. C. Baker, and C. Chappey. 1995. Genetic analysis of human immunodeficiency virus type 1 V3 region isolates from mothers and infants after perinatal transmission. J. Virol. **69**:1001-1012.

Albert, J., J. Wahlberg, and M. Uhlén. 1993. Forensic evidence by DNA sequencing. Science **361**:595-596.

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

Alkhatib, G., C. Combadiere, C. C. Broder, Y. Feng, P. E. Kennedy, P. M. Murphy, E. A. Berger, 1996. CC CKR5: A RANTES, MIP-1α, MIP-1ß receptor as a fusin cofactor for macrophage-tropic HIV-1. Science **272**:1955-1958.

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

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.

Centers for Disease Control, 1981a. *Pneumocystis* pneumonia - Los Angeles. Morbid. Mortal. Weekly Rep. **30**:250-252.

Centers for Disease Control, 1981b. Kaposi's sarcoma and *Pneumocystis* pneumonia among homosexual men - New York City and California. Morbid. Mortal. Weekly Rep. **30**:305-308.

Centers for Disease Control, 1982a. Update on acquired immunodeficiency syndrome (AIDS).

Morbid. Mortal. Weekly Rep. 31:665-667.

Centers for Disease Control, 1982b. *Pneumocystis carinii* pneumonia among persons with haemophilia A. Morbid. Mortal. Weekly Rep. 31:3667.

Centers for Disease Control, 1982c. Possible transfusion associated Acquired Immune Deficiency Syndrome (AIDS) - California. Morbid. Mortal. Weekly Rep. 31:652-654.

Centers for Disease Control, 1982d. Immunodeficiency virus among female sexual partners of males with Acquired Immune Deficiency Syndrome (AIDS) - New York. Morbid. Mortal. Weekly Rep. 31:697-698.

Centers for Disease Control, 1982e. Unexplained immunodeficiency and opportunistic infections in infants - New York, New Jersey, California. Morbid. Mortal. Weekly Rep. 31:665-667.

Chakrabarti, L., M. Guyader, M. Alizon, M. D. Daniel, R. C. Desrosiers, P. Tiollais, and P. Sonigo, 1987. Sequence of simian immunodeficiency virus from macaque and its relationship to other human and simian retroviruses. Nature 328:543-547.

Cichutek, K., S. Norley, R. Linde, W. Kreuz, M. Gahr, J. Löwer, G. von Wagenheim, and R. Kurth, 1991. Lack of HIV-1 V3 region sequence diversity in two haemophiliac patients infected with a putative biologic clone of HIV-1 AIDS 5:1185-1187.

Clavel, F., D. Guétard, F. Brun-Vézinet, S. Charmaret, M. A. Ray, M. O. Santos-Ferreira, A. G. Laurent, C. Dauguet, C. Katlama, C. Rouzioux, D. Klatzman, J. L. Champalimaud, and L. Montagnier, 1986. Isolation of a new human retrovirus from West African patients with AIDS. Science 233:343-346.

Cocchi, F., A. L. DeVico, A. Garzino-Demo, S. K. Arya, R. C. Gallo, and P. Lusso, 1995. Identification of RANTES, MIP-1α, and MIP-1ß as the major HIV-suppresive factors produced by $CD8^+$ T cells. Science 270:1811-1815.

Coffin, J., 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.

Coffin, J. M. 1992. Genetic diversity and evolution of retroviruses. Curr. Top. Microbiol. Immunol. 176:143-164.

Connor, R.I, B. K. Chen, S. Choe, and N. R. Landau. 1995. Vpr is required for efficient replication of human immunodeficiency virus type-1 in mononuclear phagocytes. Virology 206:936-944.

Cullen, B. R., 1992. Mechanism of action of regulatory proteins encoded by complex retroviruses. Microbiol. Rev. 56:375-394.

Dalgleish, A. G., P. C. Beverley, P. R. Clapman, 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.

Deng, H., R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. D. Marzio, S. Marmon, R. E. Sutton, C. Mark Hill, C. B. Davis, S. C. Peiper, T. J. Schall, D. R. Littman, and N. R. Landau, 1996. Identification of a major coreceptor for primary isolates of HIV-1. Nature 381:661-666.

Dragic, T., V. Litwin, G. P. Allaway, S. R. Martin, Y. Huang, K. A. Nagashima, C. Cayanan, P. J. Maddon, R. A. Koup, J. P. Moore, and W. A. Paxton, 1996. HIV-1 entry into $CD4^+$ cells is mediated by the chemokine receptor CC-CKR-5. Nature 381:667-673.

Eigen, M. and R. Winkler-Oswatitsch, 1990. Statistical geometry in sequence space. Meth. Enzymol. 183:505-530.

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.

Fouchier, R. A. M., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Scuitemaker, 1992. Phenotype-associated sequence variation in the third hypervariable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. 66:3183-3187.

Fukasawaka, M., T. Miura, A. Hasegawa, S. Morikawa, H. Tsujimoto, K. Miki, T. Kitamura, and M. Hayami, 1988. Sequence of simian immunodeficiency virus from African green monkey, a new member of the HIV/SIV group. Nature 333:457-461.

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

Gao, F., L. Yue, A. T. White, P. G. Pappas, J. Barchue, A. P. Hanson, B. M. Greene, P. M. Sharp, G. M. Shaw, and B. H. Hahn, 1992. Human infection by genetically diverse SIV_{sm}-related HIV-2 in West Africa. Nature 358:495-499.

Gojobori, T., E. N. Moriyama, and M. Kimura, 1990. Molecular clock of viral evolution and the neutral theory. Proc. Natl. Acad. Sci. USA. 87:10015-10018.

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. 68:1581-1585.

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.

He, J., S. Choe, R. Walker, P. Di Marzio, D. O. Morgan, and N. A. Landau. 1995. Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell

cycle by inhibiting p34 ^{cdc2} activity. J. Virol. 69:6705-6711.

Hirsch, V. M., R. A. Olmsted, M. Murphey-Corb, R. H. Purcell, and P. R. Johnson, 1989. An African primate lentivirus (SIV_{sm}) closely related to HIV-2. Nature 339:389-392.

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.

Holland, J. J., J. C. de la Torre, D. A. Steinhauer, 1992. RNA virus populations as quasispecies. Curr. Top. Microbiol. Immunol. 176:1-20.

Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam, and A. J. Leigh Brown, 1992. Convergent and divergent 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.

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. J. Infect. Dis. 171:45-53.

Howell, R. M., J. E. Fitzgibbon, M. Noe, Z. J. Ren, D. J. Gocke, T. A. Schwartzer *et al.*, 1991. *In vivo* sequence variation of the human immunodeficiency virus type 1 *env* gene: evidence for recombination among variants found in a single individual. AIDS Res. Hum. Retroviruses 7:869-876.

Huett, T., R. Cheynier, A. Meyerhans, G. Roelants, and S. Wain-Hobson, 1990. Genetic organization of a chimpanzee lentivirus related to HIV-1. Nature 345:356-359.

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

Jameson, B. A., P. E. Rao, L. I. Kong, B. H. Hahn, G. M. Shaw, L. E. Hood, and S. B. Kent, 1988. Location and chemical synthesis of a binding site for HIV-1 on the CD4 protein. Science **240**:1335-1339.

Javaherian, K., A. J. Langlois, C. McDanal, K. L. Ross, L. I. Eckler, C. L. Jellis, A. T. Profy, J. R. Rusche, D. P. Bolognesi, S. D. Putney, and T. J. Matthews, 1989. Principal neutralization determinant of human immunodeficiency virus type 1 envelope protein. Proc. Natl. Acad. Sci. USA. **86**:6768-6772.

Klatzmann, D., E. Champagne, S. Chamaret, J. Gruest, D. Guetard, T. Hercend, J. C. Gluckman, and L. Montagnier, 1984. T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. Nature **312**:767-768.

Korber, B., and G. Myers. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retroviruses **8**:1549-1560.

Kuiken, C. L., G. Zwart, E. Baan, R. A. Couthino, J. A. R. van den Hoek, and J. Goudsmit, 1993. Increasing antigenic and genetic diversity of the V3 variable region of the V3 variable domain of the human immunodeficiency virus envelope protein in the course of the AIDS epidemic. Proc. Natl. Acad. Sci. USA. **90**:9061-9065.

Kuiken, C. L. and J. Goudsmit, 1994. Silent mutation pattern in V3 sequences distinguishes virus according to risk group in Europe. AIDS Res. Hum. Retroviruses **10**:319-320.

Kuiken, C. L., M. T. Cornelissen, F. Zorgdrager, S. Hartman, A. J. Gibbs, and J. Goudsmit. 1996. Consistent risk group-associated differences in human immunodeficiency virus type 1 vpr, vpu and V3 sequences despite independent evolution. J. Gen. Virol. 77:783-792.

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 1 in infants by vertical transmission. J. Clin. Invest. 93:380-390.

Lasky, L. A., G. Nakamura, D. H. Smith, C. Fennie, C. Shimasaki, E. Patzer, P. Berman, T.

Gregory, and D. J. Capon, 1987. Delineation of a region of the human immunodeficiency virus type 1 gp120 glycoprotein critical for interaction with the CD4 receptor. Cell 50:975-985.

Leigh Brown, A. J., and P. Monaghan, 1988. Evolution of the structural proteins of human immunodeficiency virus: selective constraints on nucleotide substitution. AIDS. Res. Hum. Retrovirus. 6:399-407.

Leigh Brown, A. J., 1991. Sequence variability in human immunodeficiency viruses: pattern and process in viral evolution. AIDS 5 (suppl. 2):S35-S42.

Leigh Brown, A. J. and E. C. Holmes, 1994. Evolutionary biology of human immunodeficiency viruses. Annu. Rev. Ecol. Syst. 25:127-165.

Leigh Brown, A. J., D. Lobidel, C. M. Wade, S. Rebus, A. N. Phillips, R. P. Brettle, A. J. France, C. S. Leen, J. McMenamim, A. McMillan, R. D. Maw, F. Mulcahy, J. R. Robertson, K. N. Sankar, G. Scott, and J. F. Peutherer. The molecular epidemiology of human immunodeficiency virus type 1 in six cities in Britain and Ireland. Submitted for publication.

Levy, J.A., A. D. Hoffman, S. M. Kramer, J. A. Landis, J. M. Shimabukuro, and L. S. Oshiro, 1984. Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. Science 225:840-842.

Levy, J. A., 1993. Pathogenesis of human immunodeficiency virus infection. Microbiol. Rev. 57:183-289.

Li, W-H., M. Tanimura, and P. M. Sharp, 1988. Rates and divergence between AIDS virus nucleotide sequences. Mol. Biol. Evol. 5:313-330.

Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, E. Sanders-Buell, G. A. Eddy, G. van der Groen, K. Fransen, G-M. Gershy-Damet, R. Deleys, and D. S. Burke, 1993. Phylogenetic analysis of *gag* genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS 7:769-780.

Louwagie, J., W. Jansens, J. Mascola, L. Heyndrickx, P. Hegerich, G. van der Groen, F. E. McCutchan, and D. S. Burke, 1995. Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. J. Virol. 69:263-271.

Lukashov, V. V., C. L. Kuiken, and J. Goudsmit. 1995. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J. Virol. 69:6911-6916.

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.

Masur, H., M. A. Michaelis, J. B. Greene, I. Onarato, R. A. van de Stouwe, R. S. Holzman, G. Wormser, L. Brettman, M. Lange, H. W. Murray, and S. Cunningham-Rudles, 1981. An outbreak of community acquired *Pneumocystis carinii* pneumonia: initial manifestation of cellular immune dysfunction. New Engl. J. Med 305:1431-1438.

McNearney, T., P. Westervelt, B. J. Thielan, D. B. Trowbridge, J. Garcia, R. Whittier, and L. Ratner L. 1990. Limited sequence heterogeneity among biologically distinct human immunodeficiency virus type 1 isolates from individuals involved in a clustered infectious outbreak. Proc. Natl. Acad. Sci. USA 87:1917-1921.

McNearney, T., Z. Hornickova, R. Markham, A. Birdwell, M. Arens, A. Saah, and L. Ratner. 1992. Relationship of human-immunodeficiency-virus type-1 sequence heterogeneity to stage of disease. Proc. Natl. Acad. Sci. USA. 89:10247-10251.

Miedema, F., and M. R. Klein, 1996. AIDS pathogeneis: a finite immune response to blame? Science 272:505-506.

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.

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. 61:570-580.

Mulder-Kampinga, G. A., C. Kuiken, 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. 74:1747-1756.

Mulder-Kampinga, G. A., A. Simonon, C. L. Kuiken, J. Dekker, H. J. Scherpbier, P. Van De Perre, K. Boer, and J. Goudsmit. 1995. Similarity in *env* and *gag* genes between genomic RNAs of human immunodeficiency virus type 1 (HIV-1) from mother and infant is unrelated to the time of HIV-1 RNA positivity in the child. J. Virol. 69:2285-2296.

Myers, G, B. Korber, B. H. Hahn, K. T. Jeang, J. W. Mellors, F. E. McCutchan, L. E. Henderson, and G. N. Pavlakis, 1995. Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, New Mexico.

Nahmias, A. J., J. Weiss, X. Yao, F. Lee, R. Kodsi, M. Schanfield, T. Matthews, D. Bolognesi, D. Durack, A. Motulsky, P. Kanki, and M. Essex, 1986. Evidence for human infection with and HTLV III/LAV-like virus in Central Africa, 1959. Lancet 327:1279-1280.

Nowak, M. A., 1992. What is a quasispecies ? Trends Ecol. Evol. 7:118-121.

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.

Ou, C.-Y., C. A. Ciesieleski, G. Myers, C. I. Bandea, C.-C. Luo, B. T. M. 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, H. W. Jaffe, the Laboratory Investigation Group, and the Epidemiological Investigation Group. 1992. Molecular epidemiology of HIV transmission in a dental practice. Science 256: 1165-1171.

Piatak, M. Jr., M. S. Saag, L. C. Yang, S. J. Clark, J. C. Kappes, K.-C. Luk, B. H. Hahn, G. M. Shaw, and J. D. Lifson, 1993. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. Science 259:1749-1754.

Poss, M., H. L. Martin, J. K. Kreiss, L. Granville, B. Chohan, P. Nyange, K. Mandaliya, and J. Overbaugh. 1995. Diversity in virus populations from genital secretions and peripheral blood from women recently infected with human immunodeficiency virus type 1. J. Virol. 69:8118-8122.

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

Roberts, J. D., K. Bebenek, and T. A. Kunkel, 1988. The accuracy of reverse transcriptase from HIV-1. Science 242:1171-1173.

Robertson, D. L., P. M. Sharp, F. E. McCutchan, B. H. Hahn, 1995. Recombination in HIV-1. Nature 374:124-126.

Rusche, J. R., K. Javaherian, C. McDanal, J. Petro, D. L. Lynn, R. Grimaila, A. Langlois, R. C. Gallo, L. O. Arthur, P. J. Fischinger, D. P. Bolognesi, S. D. Putney, and T. J. Matthews, 1988. Antibodies that inhibit fusion of human immunodeficiency virus-infected cells bind a 24-amino acid sequence of the viral envelope, gp120. Proc. Natl. Acad. Sci. USA. 85:3198-3202.

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 334:440-444.

Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan, 1995. Genotyping of HIV-1. *In* Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences (Myers, G., B. Korber, B. H. Hahn, K. T. Jeang, J. W. Mellors, F. E. McCutchan, L. E. Henderson, and G. N. Pavlakis, eds.) Los Alamos National Laboratory, Los Alamos, New Mexico.

Salvatori, F., S. Masiero, C. Giaquinto, C. M. Wade, A. J. Leigh Brown, L. Chieco-Bianchi, A. De Rossi. 1997. Evolution of human immunodeficiency virus type 1 in perinatally infected infants with rapid and slow progression to disease. J. Virol. 71:4694-4706.

Scarlatti, G., T. Leitner, E. Halapi, J. Wahlberg, P. Marchisio, M. A. Clerici-Schoeller, H. Wigzell, E. M. Fenyö, J. Albert, M. Uhlén, 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. 90:1721-1725.

Sharp, P. M., D. L. Robertson, F. Gao, and B. H. Hahn, 1994. Origins and diversity of human immunodeficiency viruses. AIDS 8 (suppl 1):

Shaw, G. M., F. Wong-Staal, R. C. Gallo. 1988. Etiology of AIDS: Virology, molecular biology, and evolution of human immunodeficiency viruses. *In* DeVita Jr, V. T., S. Hellman, and S. A. Rosenburg. eds. AIDS: Etiology, Diagnosis, Trestment and Prevention (second edition). J. B. Lippincott Company, Philadelphia, USA. pp.11-31.

Shpaer, E. G., and J. I. Mullins, 1993. Rates of amino acid change in the envelope protein correlate with athogenicity of primate lentiviruses. J. Mol. Evol. 37:57-65.

Simmonds, P., P. Balfe, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown, 1990. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. J. Virol. 64:5840-5850.

Simmonds, P., L. Q. Zhang, F. McOmish, P. Balfe, C. A. Ludlam, A. J. Leigh Brown, 1991. Disccontinuous 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.

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, R. C. Gallo, and F. Wong-Staal, 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.

Subbramanian, R. A., and E. A. Cohen, 1994. Molecular biology of the human immunodeficiency virus accessory proteins. J. Virol. 68:6831-6835.

Van't Wout, A. B., N. A. Kootstra, G. A. Mulder-Kampinga, N. A. Albrecht van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker. 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral and vertical transmission. J. Clin. Invest. 94:2060-2067.

Vartanian, J. P., A. Meyerhans, B. Asjo, S. Wain-Hobson, 1991. Selection, recombination and G-A hypermutation of human immunodeficiency virus type 1 genomes. J. Virol. 65:1779-1788.

Wade, C. M., D. Lobidel, and A. J. Leigh Brown. Analysis of human immunodeficiency virus type 1 *env* and *gag* sequence variants derived from a mother and two vertically infected children provides evidence for the transmission of multiple sequence variants. Submitted for publication.

Wade, C. M., D. Lobidel, J. R. Robertson, J. Y. Q. Mok, D. Yirrel, and A. J. Leigh Brown. Evolution of human immunodeficiency virus type 1 during infection of a male index and transmission to two infected female partners and their children. Submitted for publication.

Wain-Hobson, S., 1995. Virological mayhem. Nature News and Views 373:102.

Wei, X., S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw, 1995. Viral dynamics in human immunodeficiency virus type 1 infection. Nature 373:117-122.

Weiss, R. A., N. N. Teich, H. E. Varmus, and J. Coffin, 1985. RNA Tumor Viruses. Cold Spring Harbor Laboratory.

Weiss, R. A., and P. R. Clapham, 1996. Hot fusion of HIV. Nature News and Views

381:647-648.

Wike, C. M., B. T. M. Korber, M. R. Daniels, C. Hutto, J. Muñoz, M. Furtado, W. Parks, A. Saah, M. Bulterys, J.-B. Kurawige, and S. M. Wolinsky. 1992. HIV-1 sequence variation between isolates from mother-infant transmission pairs. AIDS Res. Hum. Retroviruses 8:1297-1300.

Wolfs, T. F. W., J. J. de Jong, H. van den Berg, J. M. Tijnagel, W. J. Krone, and J. Goudsmit, 1990. Evolution of sequences encoding the principal neutralization epitope of human immunodeficiency virus 1 is host independent, rapid, and continuous. Proc. Natl. Acad. Sci. USA 87:9938-9942.

Wolfs, T. F. W., G. Zwart, M. Bakker, and J. Goudsmit. 1992. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. Virology 189:103-110.

Wolinsky, S. M., C. M Wike, B. T. M. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Muñoz. 1992. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science 255:1134-1137.

Wolinsky, S. M., B. T. M. Korber, A. U. Neumann, M. Daniels, K. J. Kunstman, A. J. Whetsell, M. R. Furtado, Y. Cao, D. D. Ho, J. T. Safrit, and R. A. Koup, 1996. Adaptive evolution of human immunodeficiency virus type-1 during the natural course of infection. Science 272:537-542.

World Health Organization Global Programme on AIDS, 1992. Current and future dimensions of the HIV-1/AIDS pandemic: A capsule summary. WHO, Geneva.

World Health Organisation, 1995. AIDS - global data. Wkly. Epidem. Rec. 70:353-360.

Zhang, L. Q., P. Simmonds, C. A. Ludlam, and A. J. Leigh Brown. 1991. Detection, quantification and sequencing of HIV-1 from the plasma of seropositive individuals and from factor VIII concentrates. AIDS 5:675-681.

Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. 67:3345-3356.

Zhu, T. H. Mo., N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho., 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. Science 261:1179-1181.

Zhu, T., N. Wang, A. Carr, S. Wolinsky, and D. D. Ho., 1995. Evidence for coinfection by multiple strains of human immunodeficiency virus type 1 subtype B in an acute seroconvertor. J. Virol. 69:1324-1327.

Figure 1.

The HIV-1 Virion (reproduced from Shaw *et al.*, 1988).

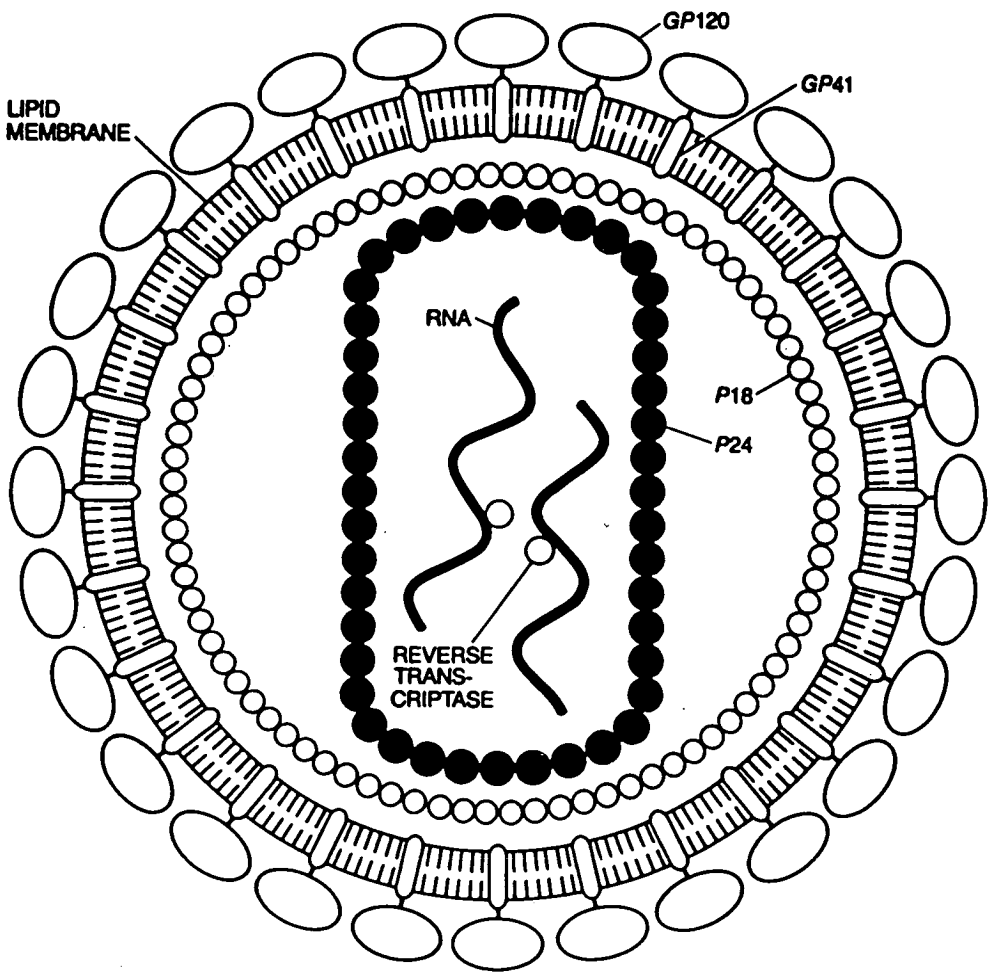

GP120
GP41
LIPID
MEMBRANE
RNA
P18
P24
REVERSE
TRANS-
CRIPTASE

Figure 2.

The HIV-1 Genome.

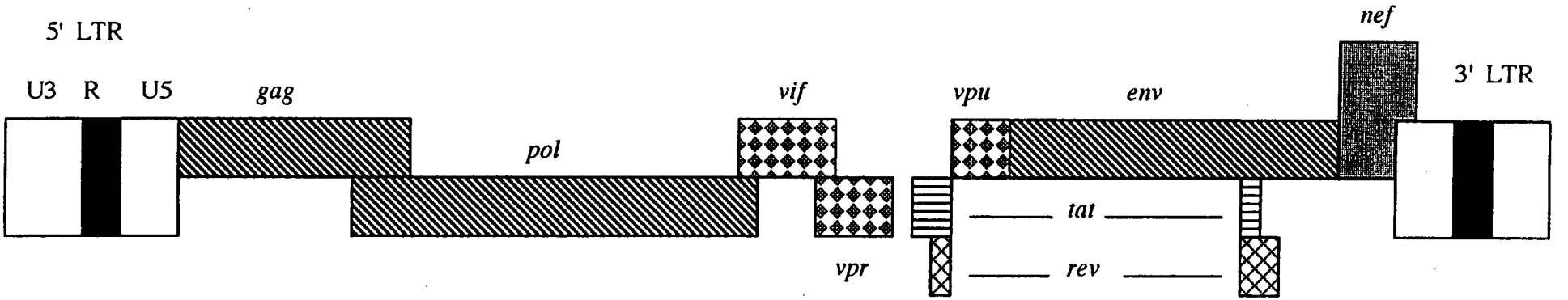
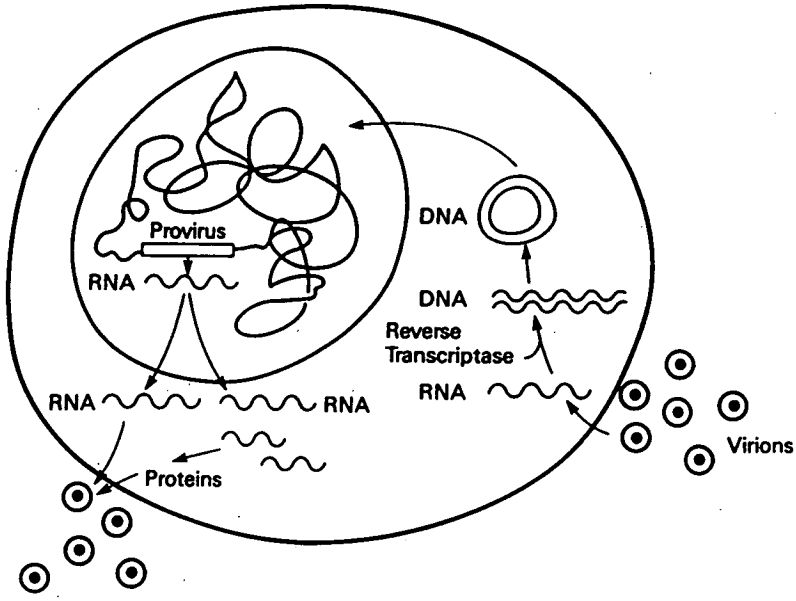

5' LTR
U3
R
U5
gag
pol
vif
vpr
vpu
env
nef
3' LTR
tat
rev

Figure 3.

Life Cycle of HIV-1 (reproduced from Shaw *et al.*, 1988).


Provirus
RNA
DNA
DNA
Reverse
Transcriptase
RNA
RNA
RNA
Proteins
Virions

Paper I

The Molecular Epidemiology of Human Immunodeficiency Virus Type 1 In Six Cities in Britain and Ireland

Andrew J. Leigh Brown[1*], Denis Lobidel[1], Christopher M. Wade[1], Selma Rebus[1], Andrew N. Phillips[2], R. P. Brettle[3], A. J. France[4], C. S. Leen[3], J. McMenamin[5], A. McMillan[6], R. D. Maw[7], F. Mulcahy[8], J. R. Robertson[9], K. N. Sankar[10], G. Scott[6], R. Wyld[3], and John F. Peutherer[11]

[1]Centre for HIV Research, Institute of Cell, Animal and Population Biology, University of Edinburgh; [2]Department of Public Health Medicine, Royal Free Hospital, London NW3; [3]Infectious Diseases Unit, City Hospital, Edinburgh; [4]King's Cross Hospital, Dundee; [5]Ruchill Hospital, Glasgow; [6]Department of Genito-Urinary Medicine, Royal Infirmary of Edinburgh, Edinburgh; [7]Department of Genito-Urinary Medicine, Royal Victoria Hospital, Belfast; [8]Department of Genito-Urinary Medicine, St James Hospital, Dublin; [9]Muirhouse Medical Group, Edinburgh; [10]Department of Genito-Urinary Medicine, Newcastle General Hospital, Newcastle; [11]Department of Medical Microbiology, University of Edinburgh

[*] Corresponding Author

Submitted for Publication

ABSTRACT

We have sequenced the p17 coding regions of the *gag* gene from 211 patients infected either through injecting drug use (IDU) or by sexual intercourse between men from 6 cities in Scotland, N. England, N. Ireland and the Republic of Ireland. All sequences were of subtype B. Phylogenetic analysis revealed substantial heterogeneity in the sequences from homosexual men. In contrast, sequence from over 80% of IDUs formed a relatively tight cluster, distinct both from those of published isolates and of the gay men. There was no large-scale clustering of sequences by city in either risk group, although a number of close associations between pairs of individuals were observed. From the known date of the HIV-1 epidemic among IDUs in Edinburgh, the rate of sequence divergence at synonymous sites is estimated to be about 0.8%. On this basis we estimate the date of divergence of the sequences among homosexual men to be about 1974, which may correspond to the origin of the B subtype epidemic.

Running head: Molecular epidemiology of HIV-1

The high genetic diversity which usually characterises HIV-1 has been exploited in many investigations of postulated linkage between infections. These include the identification of individuals belonging to an infection clusters of individuals attending a dental practice in Florida (Ou *et al.*, 1992; Hillis and Huelsenbeck, 1994), of haemophiliacs in Scotland and Germany (Balfe *et al.*, 1990; Holmes *et al.*, 1995; Kleim *et al.*, 1991; Chant *et al.*, 1993) and of a rape victim and their assailant in Sweden (Albert *et al.*, 1994). It has also allowed the exclusion of nosocomial infection as a source of HIV infection from an HIV positive surgeon (Rogers *et al.*, 1993; Holmes *et al.*, 1993), from a second U.S. dentist (Jaffe *et al.*, 1994) and from a British health care worker (Arnold *et al.*, 1995). Studies of emerging epidemics have revealed high levels of sequence similarity among patients infected in a short period of time (Ou *et al.*, 1993; McCutchan *et al.*, 1992; Dietrich *et al.*, 1993; Grez *et al.*, 1994) and in a recent investigation of a lengthy transmission chain, sequence data have been able to reconstruct the transmission routes reported (Leitner *et al.*, 1996).

These studies have been carried out with sequences from the V3 region of the *env* gene (Ou *et al.*, 1992; Hillis and Huelsenbeck, 1994; Balfe *et al.*, 1990; Kleim *et al.*, 1991; Chant *et al.*, 1993), from the p17 coding region of the *gag* gene (Holmes *et al.*, 1995; Holmes *et al.*, 1993; Albert *et al.*, 1994) and, more recently, with datasets consisting of the entire *env* gene (Arnold *et al.*, 1995) or of combined *env* and *gag* gene sequences (Leitner *et al.*, 1996). We have shown that the p17 coding sequence on its own can reconstruct known cases of epidemiological linkage, and have used it to describe the molecular epidemiology of HIV-1 within Edinburgh (Holmes *et al.*, 1995). In that investigation, injecting drug users (IDUs) and haemophiliacs were identified as forming distinct infection clusters, with patients infected following heterosexual transmission grouping with the IDUs. An investigation in Amsterdam based on V3 region sequences, also suggested a separation of IDUs from other risk groups, although only 2 synonymous nucleotide positions in the V3 loop were consistently associated with the distinction (Kuiken and Goudsmit, 1994).

We have analysed nucleotide sequence data in order to reconstruct and compare the HIV epidemics in Scotland, Ireland and the north of England. Samples were obtained from individuals infected by sexual contact between men and by injecting drug use from 3 cities in Scotland (Edinburgh, Glasgow and Dundee), from Newcastle in the north of England, Belfast (Northern Ireland) and Dublin (Republic of Ireland). Sequences of the p17 coding

region were obtained from over 200 individuals in this survey and we have analysed the phylogenetic relationships among these sequences using both parsimony and neighbour-joining (NJ) techniques. Our results indicate lasting distinctions between these risk groups which extend between cities and suggest that HIV-1 was sufficiently rapidly transmitted among IDUs in these cities for those patients to be defined as a single group from their viral genotypes.

Patients

Samples were obtained with informed consent from 211 HIV-1 seropositive homosexuals and injecting drug users (IDU) attending genito-urinary medicine and infectious diseases clinics in Edinburgh, Glasgow, Dundee, Newcastle, Belfast and Dublin between 1993 and 1995. All patients were resident in the city where they attended, although a few reported travelling overseas and were suspected to have acquired their infections outside the British Isles. Approximately 20 individuals from each risk group within each city were included in the study. When samples from more individuals were available, a subset was chosen at random for sequencing. Most were seroprevalent cases, although seroconversion dates were available for a minority. Where the number of patients available was less than 20, then all samples in that category (risk group and city) were sequenced.

Sample preparation and DNA Extraction

EDTA-treated whole blood samples were collected, shipped to Edinburgh and processed within 48 hours. The blood samples were separated into plasma and peripheral blood mononuclear cell (PBMC) fractions by ficoll Hypaque (Pharmacia) density gradient centrifugation at 1500g for 30 minutes and the PBMC fraction stored immediately in liquid nitrogen. DNA extraction was performed from 10^6 to 10^7 uncultured PBMCs as described by Simmonds *et al.* (1990).

PCR Amplification and sequence analysis

An approximately 390 base pair (bp) fragment of the p17 coding region of the *gag* gene (between positions 69 to 453 in the HIV-HXB2 genome (Myers *et al.*, 1995) was amplified for each patient by nested polymerase chain reaction (PCR) essentially as described by Leigh Brown and Simmonds (1995), using primers "*gag* 1-4" of Holmes *et al.* (1993). Single-stranded DNA was purified on streptavidin coated magnetic beads (Dynal-Dynabeads M280) and sequenced using an Applied Bioystems PRISM Sequenase Terminator Single Stranded DNA Sequencing Kit according to ABI operating instructions, as described elsewhere (Leigh Brown and Simmonds, 1995).

The raw nucleotide sequences were assembled with the TED and XBAP sequence editors (Staden, 1993) and aligned using the CLUSTAL V algorithm (Higgins and Sharp,

1988), as implemented in version 2.2 of the Genetic Data Environment (GDE) package (Smith *et al.*, 1994). The final alignment was improved manually. Phylogenetic analyses were performed using programs taken from version 3.52c of the Phylogeny Inference Package (PHYLIP; Felsenstein, 1989) using the neighbour-joining method (Saitou and Nei, 1987; program "NEIGHBOR"), maximum parsimony ("DNAPARS") and bootstrap resampling (Felsenstein, 1985) ("SEQBOOT" and "CONSENSE"). Nucleotide distances were estimated using the generalised two-parameter (maximum likelihood) model (Kishino and Hasegawa, 1989) ("DNADIST"). Principal coordinates analysis was performed using the program "PCOORD" (Higgins, 1992).

RESULTS

Two sequences of the p17 coding region of *gag* were obtained for all newly sequenced patients. Preliminary neighbour joining phylogenetic analysis of 411 sequences from 211 patients showed that for all patients for whom more than one sequence was obtained, the sequence which grouped most closely was the other from the same individual. This preliminary analysis also included sequences from 24 HIV-1 subtype B reference isolates, as a check for contamination. Subsequently, a single sequence was used from each patient for all further analyses so as to reduce computation time.

Of the approximately 390 nucleotide sites sequenced no fewer than 268 were variable in the dataset. At the amino acid level, 99 out of 130 residues showed some variation. The variable residues were distributed throughout the sequence, with some evidence for greater conservation towards the 3' end. The amino acid sequences obtained from 116 homosexual men and 84 IDUs are shown in Figure 1 aligned against HIV-1_{MN}; the sequences from additional patients analysed from Edinburgh will be presented elsewhere (Wade, C.M., *et al.* submitted).

Nucleotide distance comparisons

Nucleotide distances were calculated for all possible pairwise comparisons of the sequences from 211 new patients together with those of 24 reference isolates. The mean distances between individuals were calculated by city and risk group (Table 1); the mean distance among reference sequences was 6.03%. Among homosexual men, the mean distances for each city were very similar with an overall mean of 7.2% (range 6.7% (Belfast) - 7.5% (Edinburgh)). However, for 4 cities, there was even greater uniformity among IDUs (range 4.3% - 4.4%). The mean distance among Belfast IDUs was higher, at 7.3%, but only 4 IDUs were available from this city and only one from Newcastle. There was a marked difference between the means for the two risk groups: the means for homosexual men were significantly higher than those for IDUs, ($P<0.001$; based on binomial standard errors given in Table 1). In addition the mean distances among 16 haemophiliacs from Edinburgh was 3.3 ± 0.23%. This group, all of whom seroconverted following exposure to a common batch of factor VIII, had previously been found to show a low level of nucleotide diversity (Holmes *et al.*, 1995).

Comparisons between sequences from different cities reveal an unexpected feature in both risk groups. The mean divergence between patients from the same risk group for any

pair of cities was no greater than the within-city diversity for the same risk group. The range of inter-city mean distances among IDUs was 4.3%-4.6%, and for homosexual men was 6.8%-7.4%. Thus, overall, there was no effect whatever of geographical origin on nucleotide distance, but a highly significant and consistent effect of risk group.

Estimating the frequencies of synonymous (d_s) and nonsynonymous substitutions (d_n) separately revealed that the difference between risk groups came from both classes of substitution (Table 1b). However, the mean synonymous divergence among sequences from homosexuals was nearly twice that among IDUs, while the ratio for nonsynonymous substitutions was 1.4. A corresponding difference in the ratio d_s/d_n between the risk groups was observed. For samples from homosexual men this was 3.0, while for IDUs it was 2.3.

Phylogenetic analysis

The initial comparison of the 211 patients' sequences with HIV-1 reference sequences revealed that all belonged to subtype B. All further analyses were undertaken against a background only of subtype B reference sequences. An NJ tree was constructed for the sequences from all 211 IDUs and homosexual men together with 16 Edinburgh haemophiliacs and 24 subtype B reference sequences. The tree is characterised by long branches to the tips with few, and short, internal branches (Figure 2). Nevertheless, one subdivision is apparent, which appeared consistently with different numbers of sequences. This contains approximately 80% of the sequences from IDUs and a small number (5) from homosexual men, but none of the reference sequences, the closest being HIV-1_{PH136}. A total of 13 IDUs were located among sequences from homosexuals in the NJ tree (Figure 2). The location of the Edinburgh haemophiliacs amongst these sequences confirmed their separation from the IDUs proposed by Holmes *et al.* (1995) and specifically associated the source of this infection with the homosexual risk group (see below).

In order to test the division of the dataset into two risk group-associated clusters, a parsimony analysis was carried out and a majority-rule consensus of 100 of the most parsimonious trees is shown (Figure 3). This cladogram also split into two groups, again separating most of the IDUs from the remaining sequences, with none of the B subtype reference sequences among them. To investigate the stability of the IDU cluster, each of the 100 trees was examined, and the location of each IDU sequence recorded. Only the 9 sequences identified by asterisks in Figure 3, which moved in or out of the IDU group in different trees, showed any inconsistency in their clustering. We conclude that the parsimony

analysis strongly supports the existence of two phylogenetically distinct clusters, associated with the two risk groups.

A third method of analysing nucleotide distance data, based on the principal coordinates technique (Higgins, 1992), was employed to assess the division of the sequences by risk group. Following a transformation of the distances, principal coordinates analysis determines which combination of variables is most suited to reflect the variation in the dataset (the "principal coordinates"). Although the principal coordinates will be a poor reflection of the total variation in the dataset, they will reflect the major distinctions and consequently the method is particularly useful in resolving major trends within a dataset which may often be invisible or unreliable in a tree. Principal coordinates analysis of the 211 patients from the two risk groups (Figure 4) clearly distinguished the sequences into two groups containing sequences from IDUs and homosexual men, respectively, and with very little overlap between the two.

Detection of contact networks

As indicated above, although the main IDU cluster was repeatedly identified among the most parsimonious trees, bootstrap resampling of the data did not support it. However a total of 13 smaller clusters were supported at the 65% level or above, including 6 groups known previously to be linked. These included groups of reference sequences derived from the HIV-1_{LAI} strain, sequences from individuals who were part of an outbreak of HIV-1 infection in a Scottish prison in 1993 (Taylor *et al.*, 1995; Yirrell *et al.*, 1997) and sequences from the Edinburgh haemophiliac cohort (Table 2). The haemophiliacs grouped specifically with a sequence from a homosexual man attending the Edinburgh GUM clinic who was infected in 1984.

A total of 7 new groups were identified by bootstrapping of the entire tree with values exceeding 70%, 5 of which were supported in over 98% of bootstrap replicates (Table 2). Most of these groups contained a single pair of individuals, but one (99%) included 3 homosexual men from Edinburgh. In all clusters, the individuals identified came from the same city, consistent with the interpretation that they represent true contact networks.

DISCUSSION

In this study we have analysed nucleotide sequences of the p17 coding region of the *gag* gene obtained from over 200 patients representing the two major risk groups in six cities. Comparisons based on nucleotide distances revealed no evidence that subdividing the data by city led to significant heterogeneity. However, there was a consistent and highly significant effect of risk group, with the sequences of injecting drug users all being more similar to each other, regardless of their city of origin. Both the neighbour-joining and parsimony methods of tree construction separated the majority of IDUs into a single cluster; none of the 24 HIV-1 subtype B reference isolates grouped with IDUs, instead all were widely distributed in the rest of the tree.

There has been considerable discussion about the appropriateness of various methodologies for the reconstruction of HIV-1 phylogenies (Hillis *et al.*, 1994; Leitner *et al.*, 1996). An equally important issue is the choice of gene region for sequencing (Holmes *et al.*, 1995; Arnold *et al.*, 1995; Leitner *et al.*, 1996). The *gag* gene evolves more slowly than the V3 region of *env* (Leigh Brown and Monaghan, 1988) and it has been suggested that it does not contain sufficient information for the accurate reconstruction of phylogenies (Arnold *et al.*, 1995; Leitner *et al.*, 1996). However, when nucleotide distances are as great as has been observed in this study (Table 1), that is not likely to pose a significant problem. Indeed, we detected no less than 5 previously unknown linkages between patients, supported in >98% bootstrap replications and confirmed all 7 previously known clusters included in the dataset. We conclude that for community-based studies, the p17 coding sequence is sufficiently informative.

No previous study of the molecular epidemiology of HIV-1 has attempted a systematic comparison of HIV sequences from multiple cities and risk groups. Extensive studies in Thailand have concentrated on 2 centres of infection, in the north and south of the country, and have compared the viral sequences from IDUs and patients infected by heterosexual contact. These studies revealed a geographic division between two genetically very distinct strains (Ou *et al.*, 1993; McCutchan *et al.*, 1992), but also that subtype E sequences were associated with a predominantly heterosexual mode of transmission, while subtype B sequences were found in IDUs (Kunanusont *et al.*, 1995). These epidemics initiated from variants found in Thailand in 1989 and 1988, respectively (Weniger *et al.*, 1994), and the subtype E sequences are now increasing in frequency in the IDU risk group as well

(Kalish *et al.*, 1995). Overall, the effect of risk group appears to have been more significant than that of geographic location, as has clearly been the case in our study.

The second major study to compare HIV sequence variants in different risk groups has focused on two risk groups in Amsterdam infected with HIV-1 subtype B. These studies analysed V3 region sequences and detected a minor, but consistent, difference between sequences from IDUs and homosexual men in the nucleotide sequence of the V3 loop (Kuiken *et al.*, 1993; Kuiken and Goudsmit, 1994). Subsequently, differences were reported in other coding regions as well (Kuiken *et al.*, 1996). The statistically significant difference in nucleotide distances we have described, and the similar division in the phylogenetic analyses suggests that these studies appear to identify the same group of IDUs. We have extended our studies to include IDUs known to have been infected in Amsterdam and have found them to fall into the main IDU cluster we originally identified in Edinburgh (Holmes *et al.*, 1995). Intriguingly, this does not apply to IDUs from southern Europe, whose sequences scatter widely among those of homosexual men (Lobidel,D. and Soriano, V., unpublished data).

From our data, and there is a clear indication that a genetic bottleneck occurred when HIV-1 entered the IDU group we have studied, presumably reflecting infection of a single individual followed by very rapid spread. The major epidemic in Edinburgh occurred in 1983/1984 (Robertson *et al.*, 1986), which appears to be earlier than epidemics among IDUs in other cities, including Amsterdam (van Haastrecht *et al.*, 1991). The extreme rapidity of transmission (Robertson *et al.*, 1986), coupled with the lack of variability observed in the first few weeks of the infection (Zhang *et al.*, 1993; Zhu *et al.*, 1993) can account for the rapid spread of almost identical variants among highly susceptible populations, as has also been described in Thailand (Weniger *et al.*, 1992), Bombay (Dietrich *et al.*, 1993; Grez *et al.*, 1994; Pfutzner *et al.*, 1992) and recently among inmates of a Scottish prison (Yirrell *et al.*, 1997).

Although the overall nucleotide diversity observed in IDUs is restricted, the diversity observed among this sample of homosexual men is at least as great as that observed between subtype B sequences from across the United States. In particular, the branch lengths between some of these sequences are substantially greater than those between the reference isolates, and even exceed the divergence between the prototype B subtype strain HIV-1_{MN} and the prototype subtype D strain HIV-1_{ELI} (Myers *et al.*, 1995). There are three possible factors responsible: firstly that the virus has entered the homosexual risk group in northern Europe

on many occasions, each of which has established descendent lineages, i.e., there has been no recent bottleneck. Secondly, the reference sequences have all been obtained from tissue culture-adapted strains; it is possible that the known selection that is imposed on the virus in this process (Meyerhans *et al.*, 1989; Kusumi *et al.*, 1992) results in a restriction of diversity in p17. Restriction of amino acid divergence of p17 *in vivo* is indicated by the greater d_s/d_n ratio we have observed among homosexual men than IDUs. This indicates a slowing down of the divergence of amino acid sequences relative to synonymous sites (Table 1b). A final factor is that the isolates referred to were established in the early to mid 1980s, while patient sequences have continued to evolve during the intervening decade.

The availability of an independent date for the IDU epidemic in Edinburgh, approximately 10 years prior to sampling, and the point source origin of the virus in this population, allows the estimation of a mean divergence rate at synonymous sites of 0.83% (±0.08%) per year. The equivalent estimate amongst sequences from homosexual men was 15.5% (Table 1b). Assuming these have evolved at the same rate, the time between divergence and sampling in this group is 18.7 years (95%
confidence interval: 13.4 - 26.4 years), corresponding to a date of about 1975 (95% CI: 1968-1980). If our sample of homosexual men represent an unbiased sample of the B subtype, this is an estimate of the origin of the B subtype epidemic. It should however be noted that our estimates are based upon all pairwise synonymous distances within the IDU and homosexual groups. Consequently, many of these values are based on shared portions of the phylogenetic tree and are thus not independent of one another. This will therefore lead to an underestimation of the error associated with the estimates of time based on these data.

ACKNOWLEDGEMENTS

We are very grateful to the patients and the nursing staff of the clinical centres for provision of samples. We are also grateful to Anne P. Leigh Brown for development of the database and Dr David Goldberg for advice and comments on the manuscript. This work was supported by the Medical Research Council AIDS Directed Programme.

REFERENCES

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.

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.

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.

Chant, K., D. Lowe, G. Rubin, W. Manning, R. O'Donoughue, D. Lyle, M. Levy, S. Morey, J. Kaldor, R. Garsia, R. Penny, D. Marriott, A. Cunningham, and G. D. Tracy. 1993. Patient-to-patient transmission of HIV in private surgical consulting rooms. Lancet 342:1548-1549.

Dietrich, U., M. Grez, H. von Briesen, B. Panhans, M. Geissendorfer, H. Kuhnel, J. Maniar, G. Mahambre, W. B. Becker, M. L. B. Mecker, and H. Rubsamen Waigmann. 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 7:23-27.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein, J. 1989. PHYLIP-phylogeny inference package (version 3.2). Cladistics 5:164-166.

Grez, M., U. Dietrich, P. Balfe, H. von Briesen, J. K. Maniar, G. Mahambre, E. L. Delwart, J. I. Mullins, and H. Rubsamen 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.

Higgins, D. G. 1992. Sequence ordinations: a multivariate analysis approach to analysing large sequence datasets. CABIOS 8:15-22.

Higgins, D. G. and P. M. Sharp. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73:237-244.

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science 264:671-677.

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

Holmes, E. C., L. Q. Zhang, P. Simmonds, A. S. Rogers, and A. J. Leigh Brown. 1993. Molecular investigation of Human Immunodeficiency Virus (HIV) infection in a patient of an HIV-infected surgeon. J. Infect. Dis. 167:1411-1414.

Holmes, E. C., L. Q. Zhang, P. Robertson, A. Cleland, E. Harvey, P. Simmonds, and A. J. Leigh Brown. 1995. The molecular epidemiology of HIV-1 in Edinburgh, Scotland. J. Infect. Dis. 171:45-53.

Jaffe, H. W., J. M. McCurdy, M. L. Kalish, T. Liberti, G. Metellus, B. H. Bowman, A. R. Neasman, and J. J. Witte. 1994. Lack of transmission of human immunodeficiency virus in the practice of a dentist with AIDS. Ann. Intern. Med. 121:855-859.

Kalish, M. L., C. C. Luo, S. Raktham, C. Wasi, A. Baldwin, G. Schochetman, T. D. Mastro, N. Young, S. Vanichseni, H. Rubsamen Waigmann, H. von Briesen, J. I. Mullins, E. Delwart, B. Herring, J. Esparza, W. L. Heyward, and S. Osmanov. 1995. The evolving molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok, Thailand: implications for HIV vaccine trials. AIDS 9:851-857.

Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order in Hominoidea. J. Mol. Evol. 4:406-425.

Kleim, J.-P., A. Ackerman, H. H. Brackman, M. Gahr, and K. E. Schneweis. 1991. Epidemiologically closely related viruses from haemophilia B patients display high homology in two hypervariable regions of the HIV-1 *env* gene. AIDS Res. Hum. Retroviruses 7:417-421.

Kuiken, C. L., G. Zwart, E. Baan, R. A. Coutinho, J. A. R. van den Hoek, and J. Goudsmit, J. 1993. Increasing antigenic and genetic diversity of the V3 variable domain of the human immunodeficiency virus envelope protein in the course of the AIDS epidemic. Proc. Natl. Acad. Sci. U.S.A. 90:9061-9065.

Kuiken, C. L., M. T. Cornelissen, F. Zorgdrager, S. Hartman, A. J. Gibbs, and J. Goudsmit. 1996. Consistent risk group-associated differences in human immunodeficiency virus type 1 *vpr*, *vpu* and V3 sequences despite independent evolution. J. Gen. Virol. 77:783-792.

Kuiken, C. L., and J. Goudsmit. 1994. Silent mutation pattern in V3 sequences distinguishes virus according to risk group in Europe. AIDS Res. Hum. Retroviruses 10:319-320.

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.

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

Leigh Brown, A., and P. Monaghan. 1988. Evolution of the structural proteins of human immunodeficiency virus: selective constraints on nucleotide substitution. AIDS Res. Hum.

Retroviruses 4:399-407.

Leigh Brown, A. J., and P. Simmonds. 1995. HIV: a practical approach. *In* Karn, J. ed. Virology and Immunology. Oxford University Press, Oxford. Volime 1 pp.161-188.

Leitner, T., D. Escanilla, C. Franzen, M. Uhlen, 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.

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 8:1887-1895.

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 58:901-910.

Myers, G., B. Korber, B. H. Hahn, K. T. Jeang, J. W. Mellors, F. E. McCutchan, L. E. Henderson, and G. N. Pavlakis. 1995. Human retroviruses and AIDS 1995, Los Alamos National Laboratory.

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 256:1165-1171.

Ou, C. Y., Y. Takebe, B. G. Weniger, C. C. Luo, M. L. Kalish, W. Auwanit, S. Yamazaki, H. D. Gayle, N. I. Young, and G. Schochetman. 1993. Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. Lancet 341:1171-1174.

Pfutzner, A., U. Dietrich, U. von Eichel, H. von Briesen, H. D. Brede, J. K. Maniar, and H. Rubsamen Waigmann. 1992. HIV-1 and HIV-2 infections in a high risk population in Bombay, India: evidence for the spread of HIV-2 and presence of a divergent HIV-1 subtype.

J. Acquir. Immune Defic. Syndr. 5:972-977.

Robertson, J. R., A. B. V. Bucknall, P. D. Welsby, J. J. K. Roberts, J. M. Inglis, J. F. Peutherer, and R. P. Brettle. 1986. Epidemic of AIDS related virus (HTLV-III/LAV) among intravenous drug abusers. BMJ. 292:527-529.

Rogers, A. S., J. W. Froggatt III, T. Townsend, T. Gordon, A. J. Leigh Brown, E. C. Holmes, L. Q. Zhang, and H. Moses III. 1993. An investigation of potential HIV transmission to the patients of an HIV-infected surgeon. JAMA 269:1795-1801.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown. 1990. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. J. Virol. 64:864-872.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet. 1994. The genetic data environment: an expandable GUI for multiple sequence analysis. CABIOS 10:671-675.

Staden, R.(1993). Staden package update. Genome News 13:12-13.

Taylor, A., D. Goldberg, J. Emslie, J. Wrench, L. Gruer, S. Camerson, J. Black, B. Davis, J. McGregor, E. Follett, J. Harvey, J. Basson, and J. McGavigan. 1995. Outbreak of HIV infection in a Scottish prison. BMJ 310:289-292.

van Haastrecht, H. J. A., J. A. R. van den Hoek, G. H. Mientjes, and R. A. Coutinho. 1991. Did the introduction of HIV among homosexual men precede the introduction among injecting drug users in the Netherlands. AIDS 6:131-132.

Weniger, B. G., K. Limpakarnjanarat, K. Ungchusak, S. Thanprasertsuk, K. Choopanya, S. Vanichseni, T. Uneklabh, P. Thongcharoen, and C. Wasi. 1992. AIDS 1991 - A year in review. AIDS 5:S71-S85.

Weniger, B. G., Y. Takebe, C. Y. Ou, and S. Yamazaki. 1994. The molecular epidemiology of HIV in Asia. AIDS 8 (suppl 2):S13-S28.

Yirrell, D., P. Robertson, S. Cameron, D. Goldberg, and A. J. Leigh Brown. 1997. Molecular epidemiology of an HIV-1 outbreak in a Scottish prison. BMJ in press.

Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. 67:3345-3356.

Zhu, T., H. Mo, N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho. 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. Science 261:1179-1181.

Table 1.

Nucleotide distances among homosexual men and injecting drug users according to city. a) Overall nucleotide distances. b) Nucleotide distances for synonymous (d_s) and nonsynonymous (d_n) nucleotide substitutions considered separately for the two risk groups.

	City	IDUs		Homosexuals	
		n	Nucleotide distance % (± s.e.)	*n*	Nucleotide distance % (± s.e.)
a)	Edinburgh	30	4.3 ± 0.19	24	7.5 ± 0.28
	Glasgow	22	4.4 ± 0.23	25	7.2 ± 0.27
	Dundee	18	4.4 ± 0.25	8	7.1 ± 0.49
	Newcastle	1	-	19	6.9 ± 0.30
	Dublin	19	4.4 ± 0.24	19	6.9 ± 0.3
	Belfast	4	7.3 ± 0.76	22	6.7 ± 0.28
	Total	94	4.6 ± 0.11	117	7.2 ± 0.12
b)	Synonymous	94	8.3 ± 0.82	117	15.5 ± 1.1
	Nonsynonymous	94	3.6 ± 0.33	117	5.0 ± 0.35
	d_s/d_n	94	2.3	117	3.0

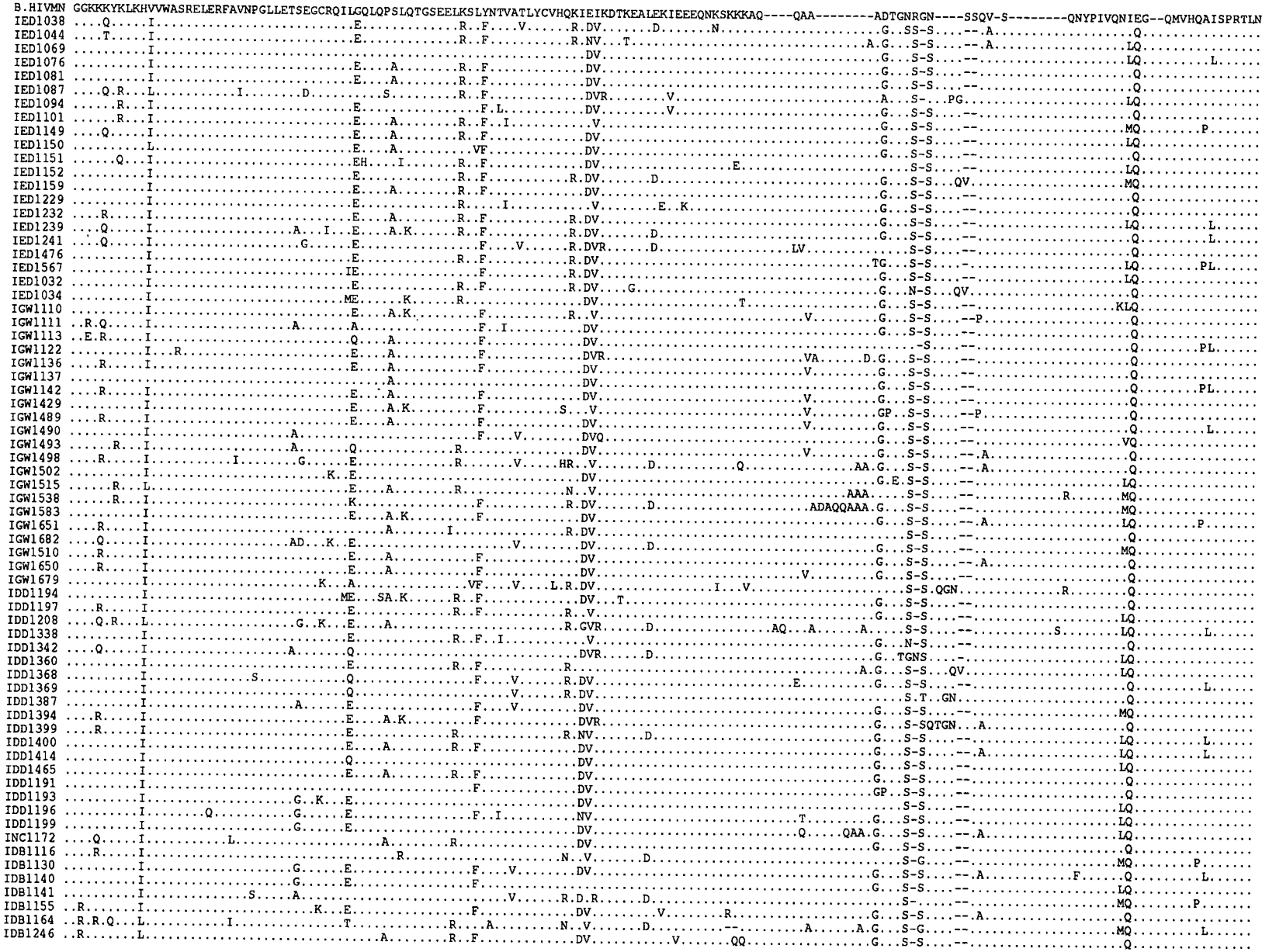
Table 2.

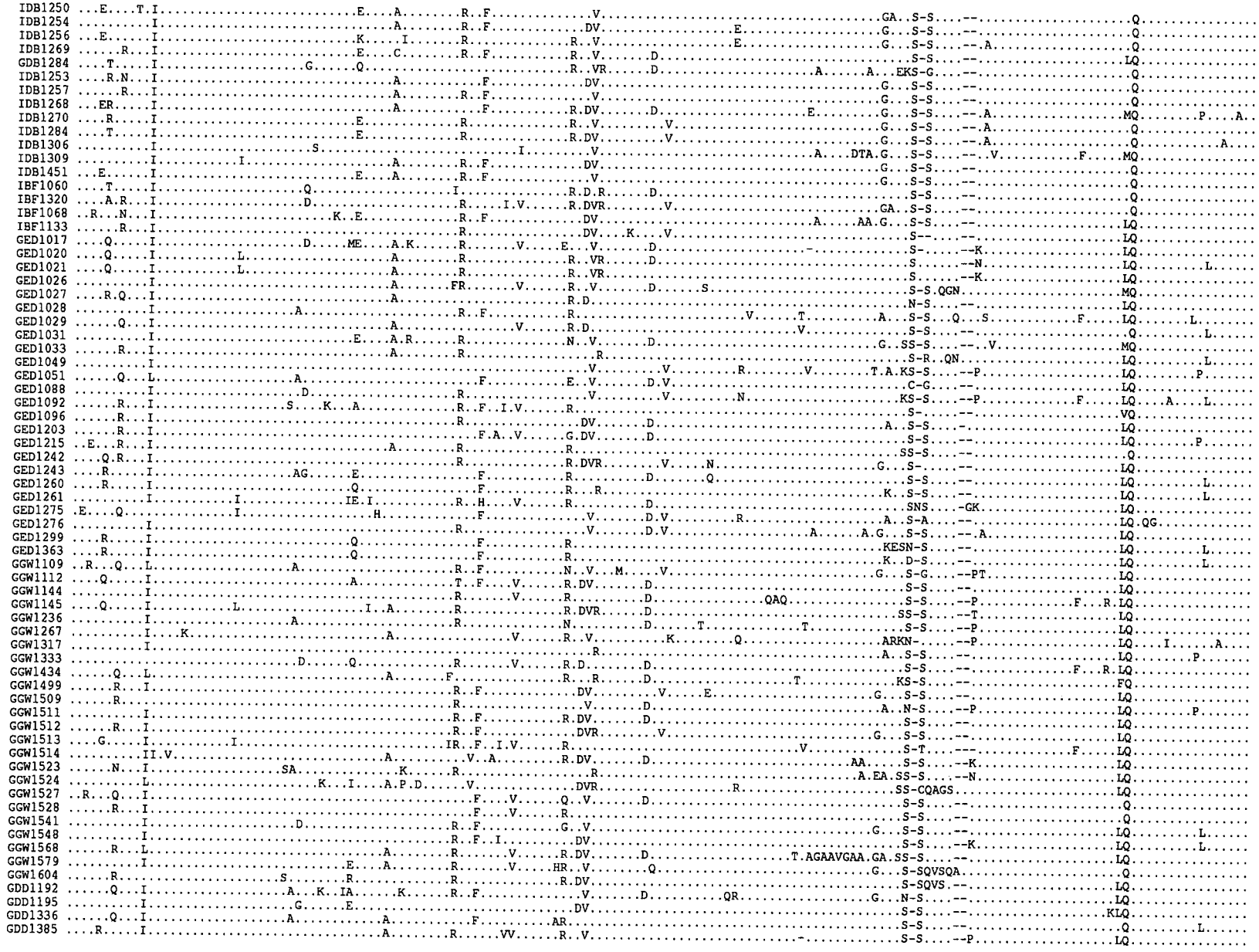
Significant associations between patients identified from phylogenetic analysis of *gag* p17 sequences. [1] % of bootstrap replications in which the cluster was observed. Those shown were all observed in >65% of bootstrap replicates. [2] Seropositive prisoner identified in Glenochil prison (Taylor *et al.*, 1995; Yirrell *et al.*, 1997). [3] Edinburgh Haemophilia Cohort (Holmes *et al.*, 1995).

Cluster	Patients	Bootstrap[1]	Risk Group	City
1	1397	100	Homosexual	Newcastle
	1294		Homosexual	Newcastle
2	1333	70	Homosexual	Glasgow
	1144		Homosexual	Glasgow
3	1066	100	Homosexual	Belfast
	1090		Homosexual	Belfast
4	Go8	100	Prisoner[2]	Glasgow
	1523		Homosexual	Glasgow
5	1122	71	IDU	Glasgow
	1142		IDU	Glasgow
6	1363	99	Homosexual	Edinburgh
	1299		Homosexual	Edinburgh
	1260		Homosexual	Edinburgh
7	1021	100	Homosexual	Edinburgh
	1020		Homosexual	Edinburgh
8	1028	66	Homosexual	Edinburgh
	EDI		Haemophilia cohort[3]	Edinburgh

Figure 1.

Amino acid sequences of the MA protein coding region (p17) of the *gag* gene from 200 patients, aligned against the sequence of HIV-1_{MN}. Patient risk group codes: I = Injecting drug user; G = Homosexual man. City codes: ED = Edinburgh; GW = Glasgow; DD = Dundee; NC = Newcastle; DB = Dublin; BF = Belfast.


B.HIVMN GGKKKYKLKHVVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQTGSEELKSLYNTVATLYCVHQKIEIKDTKEALEKIEEEQNKSKKKAQ----QAA--------ADTGNRGN----SSQV-S--------QNYPIVQNIEG--QMVHQAISPRTLN
IED1038
IED1044
IED1069
IED1076
IED1081
IED1087
IED1094
IED1101
IED1149
IED1150
IED1151
IED1152
IED1159
IED1229
IED1232
IED1239
IED1241
IED1476
IED1567
IED1032
IED1034
IGW1110
IGW1111
IGW1113
IGW1122
IGW1136
IGW1137
IGW1142
IGW1429
IGW1489
IGW1490
IGW1493
IGW1498
IGW1502
IGW1515
IGW1538
IGW1583
IGW1651
IGW1682
IGW1510
IGW1650
IGW1679
IDD1194
IDD1197
IDD1208
IDD1338
IDD1342
IDD1360
IDD1368
IDD1369
IDD1387
IDD1394
IDD1399
IDD1400
IDD1414
IDD1465
IDD1191
IDD1193
IDD1196
IDD1199
INC1172
IDB1116
IDB1130
IDB1140
IDB1141
IDB1155
IDB1164
IDB1246


IDB1250
IDB1254
IDB1256
IDB1269
GDB1284
IDB1253
IDB1257
IDB1268
IDB1270
IDB1284
IDB1306
IDB1309
IDB1451
IBF1060
IBF1320
IBF1068
IBF1133
GED1017
GED1020
GED1021
GED1026
GED1027
GED1028
GED1029
GED1031
GED1033
GED1049
GED1051
GED1088
GED1092
GED1096
GED1203
GED1215
GED1242
GED1243
GED1260
GED1261
GED1275
GED1276
GED1299
GED1363
GGW1109
GGW1112
GGW1144
GGW1145
GGW1236
GGW1267
GGW1317
GGW1333
GGW1434
GGW1499
GGW1509
GGW1511
GGW1512
GGW1513
GGW1514
GGW1523
GGW1524
GGW1527
GGW1528
GGW1541
GGW1548
GGW1568
GGW1579
GGW1604
GDD1192
GDD1195
GDD1336
GDD1385

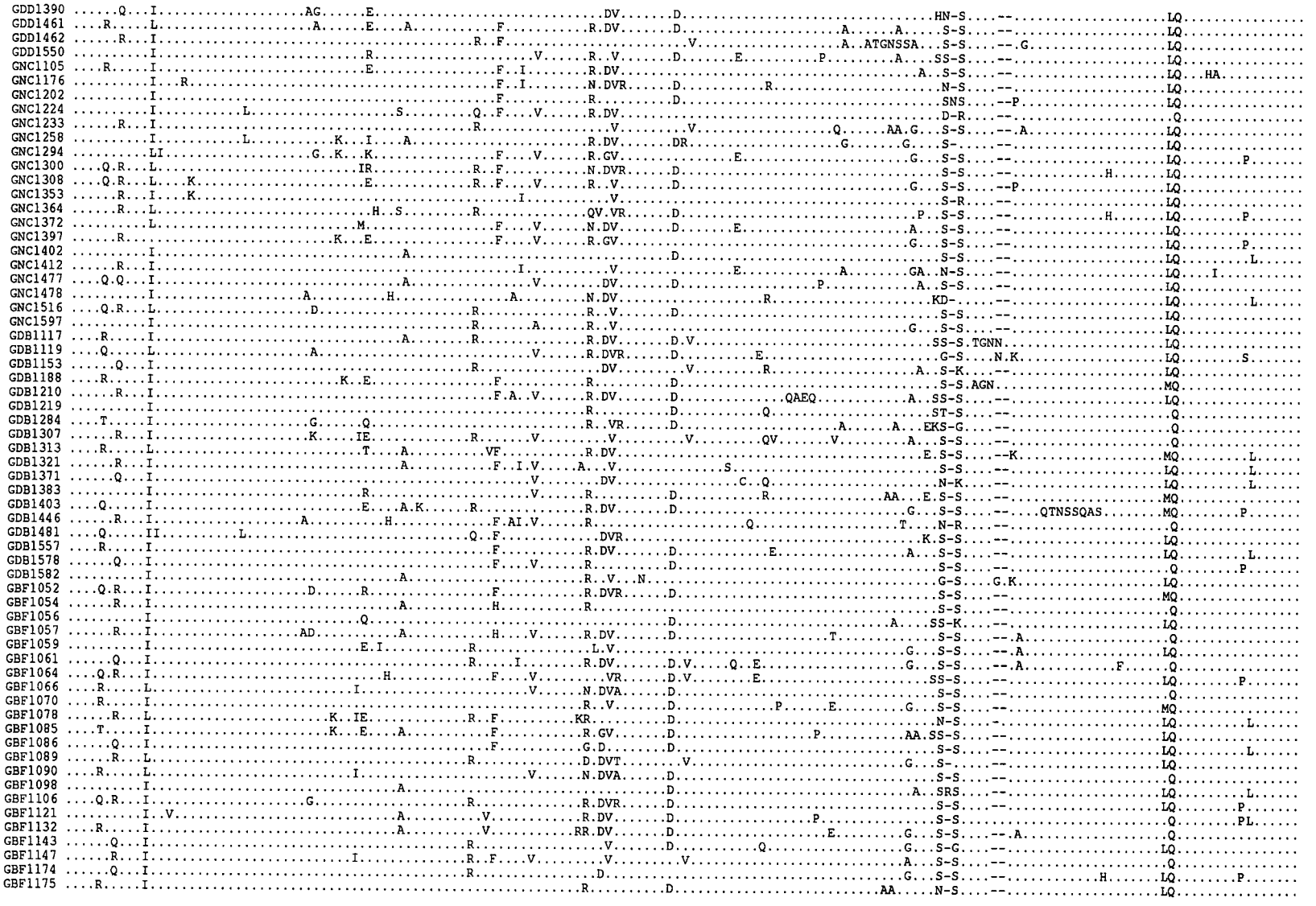
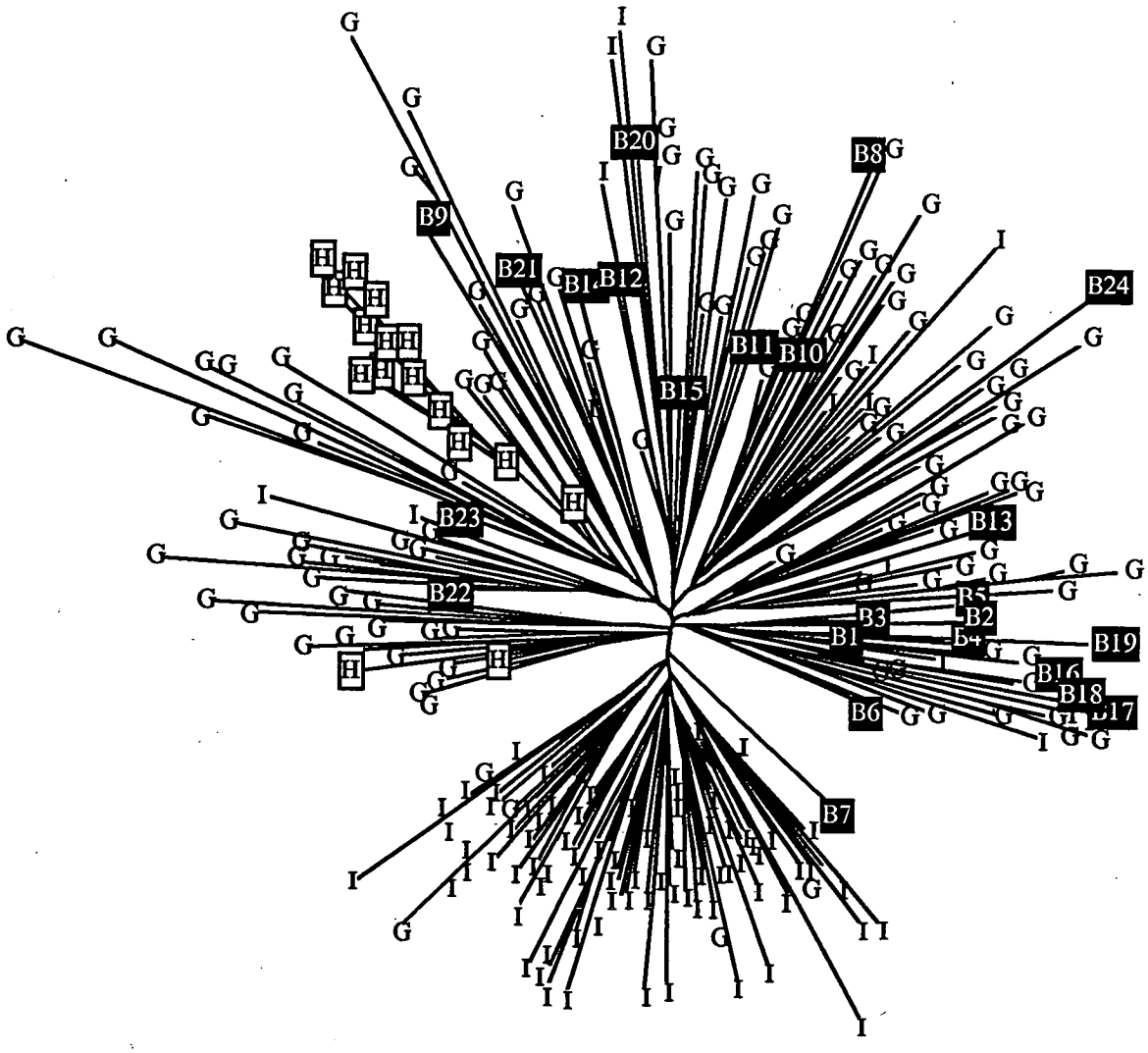


Figure 2.

Neighbour-joining phylogenetic tree of sequences of the MA protein coding region (p17) of the *gag* gene from 227 patients and 24 subtype B reference sequences. Risk group codes: H = haemophiliac; G = homosexual man; I = injecting drug user. B = Subtype B isolate sequences: 1, HIV-1_{LAI} sequence BRU; 2, HIV-1 BH102; 3, HIV-1_{PH22}; 4, HIV-1_{LAI}, clone HXB2; 5, HIV-1_{MN}; 6, HIV-1_{JH3}; 7, HIV-1_{PH136}; 8, HIV-1_{BZ190}; 9, HIV-1_{RF2}; 10, HIV-1_{NL4-3}; 11, HIV-1_{NY5}; 12, HIV-1_{CDC4}; 13, HIV-1_{D31}; 14, HIV-1_{HOY1}; 15, HIV-1_{YU2}; 16, HIV-1_{PH153}; 17, HIV-1_{JRCSF}; 18, HIV-1_{JRFL}; 19, HIV-1_{TB132}; 20, HIV-1_{BZ167}; 21, HIV-1_{HAN}; 22, HIV-1_{SF2}; 23, HIV-1_{CAM1}; 24, HIV-1_{BZ200}.



1%

Figure 3.

Parsimony tree of sequences of the p17 protein coding region (MA) of the *gag* gene from 227 patients and 24 subtype B reference sequences. The majority-rule consensus of 100 of the most parsimonious trees is shown. For risk group and subtype B isolate sequence codes, see Legend to Figure 2. Sequences marked * showed variable location with respect to the main IDU cluster among the 100 trees. All others showed a constant location.

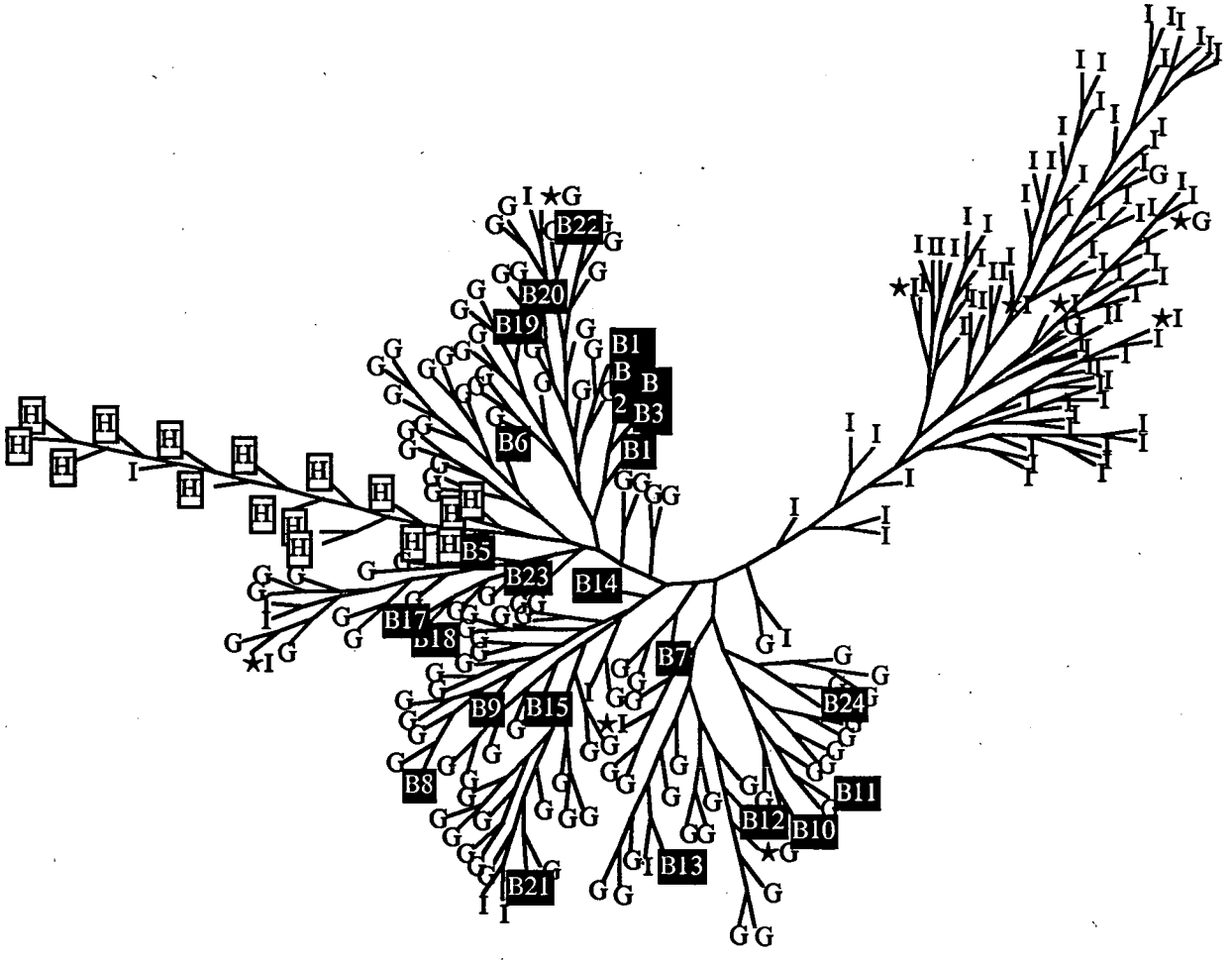
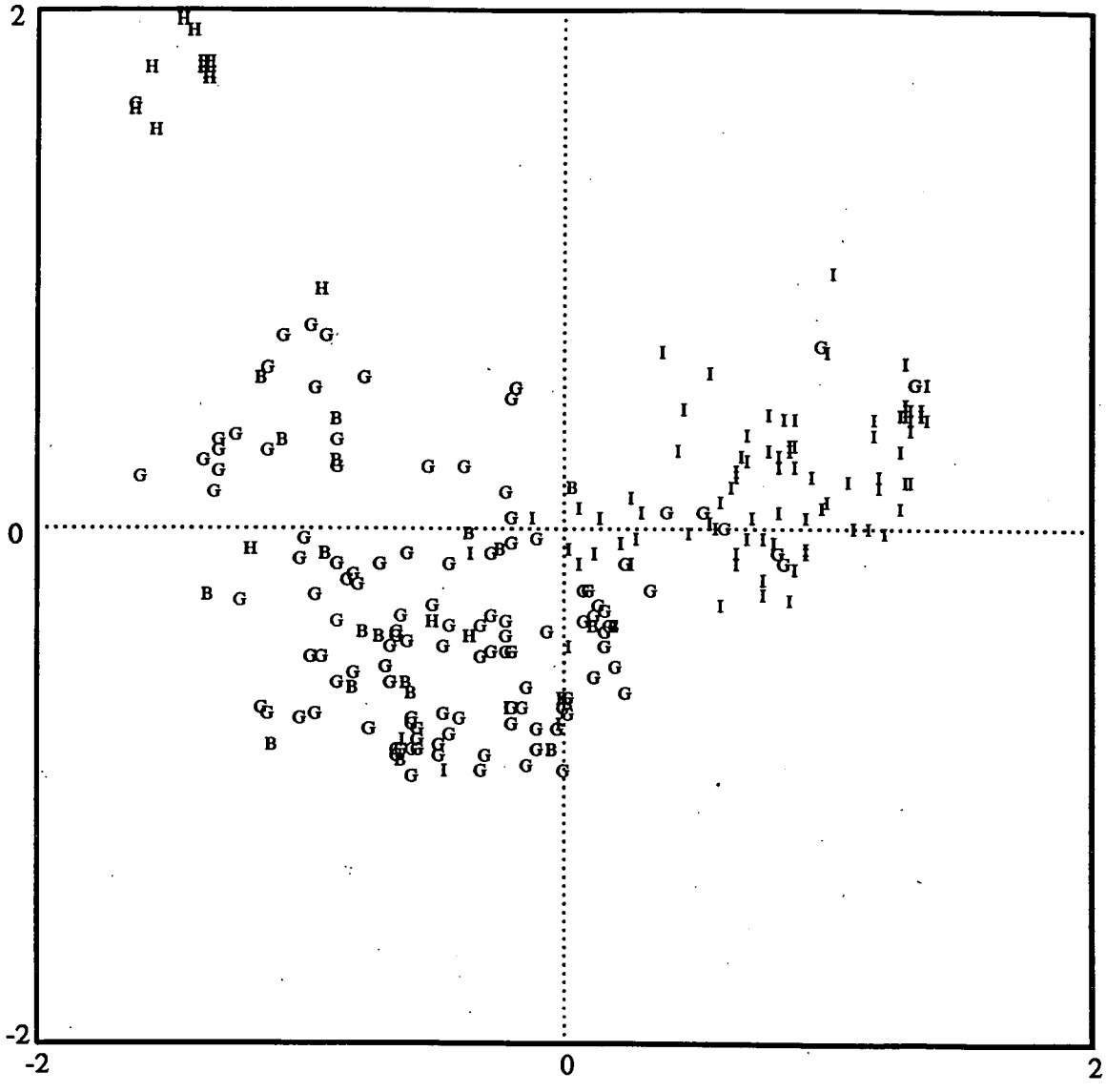

B22
B20
B19
B1
B3
B6
B5
B23
B14
B17
B18
B7
B9
B15
B24
B8
B11
B12
B10
B13
B21

Figure 4.

Principle coordinates analysis separates the main IDU cluster from other risk groups. The plot represents the 2 largest principle coordinates as ordinate and abscissa, with arbitrary units. The identifiers for risk groups are as for Figure 2.


2
0
-2
-2
0
2

Paper II

Analysis of Envelope Sequence Variants Suggest Multiple Mechanisms of Mother-to-Child Transmission of Human Immunodeficiency Virus Type 1

Laurence Briant[1a], Christopher M. Wade[2], Jacqueline Puel[1], Andrew J. Leigh Brown[2], and Mireille Guyader[1*]

[1] Laboratoire de Virologie, CHU Purpan, Place du Dr. Baylac, 31059 Toulouse, France.

[2] Centre for HIV Research, University of Edinburgh, Edinburgh EH9 3JN, United Kingdom.

[a] Present Address: Centre de Tri des Molécules anti-HIV, CRBM-CNRS, Boulevard Henri IV, 34060 Montpellier, France.

[*] Corresponding Author.

Journal of Virology (1995) 69:3778-3788

ABSTRACT

In order to elucidate molecular mechanisms involved in Human Immunodeficiency Virus (HIV-1) mother-to-child transmission, we have analysed genetic variations within the V3 hypervariable domain and flanking regions of the HIV-1 envelope gene in four mother-child transmission pairs. Phylogenetic analysis and amino acid sequence comparison were performed on sequences derived from maternal samples collected at different time points during pregnancy and after delivery, and from child samples collected from the time of birth until approximately one year of age. Heterogeneous sequence populations were observed in all maternal samples collected during pregnancy and post-delivery. In three newborns, viral sequence populations obtained within two weeks after birth revealed a high level of V3 sequence variability. In contrast, V3 sequences obtained from the fourth child (diagnosed at the age of one month) displayed a more restricted heterogeneity. The phylogenetic analysis performed for each mother-child sequence set suggested that several mechanisms may potentially be involved in HIV-1 vertical transmission. In one pair, child sequences were homogeneous and clustered in a single branch in the phylogenetic tree, suggesting that selective transmission of a single maternal variant may have occurred. In the other three pairs, the child sequences were more heterogeneous and clustered in several separate branches in the phylogenetic tree suggesting that in these cases the transmission of several maternal variants may be responsible for initiating infection within the child. In conclusion, distinct mechanisms may account for mother-to-child HIV-1 transmission including either the selective transmission of a single maternal variant or multiple transmission events.

Running title: Sequence variation in HIV mother-to-child transmission.

Extensive genetic diversity is a characteristic feature of the Human Immunodeficiency virus type-1 (HIV-1) (Alizon *et al.*, 1986; Myers *et al.*, 1991; Saag *et al.*, 1988). Considerable genomic diversity has been observed among independent isolates from epidemiologically unlinked infections and has also been reported to a lesser degree *in vivo*, among viral species from a single patient (Balfe *et al.*, 1990; Burger *et al.*, 1991; Holmes *et al.*, 1992; Kusumi *et al.*, 1992; Simmonds *et al.*, 1990; Simmonds *et al.*, 1991). Comparison of different isolates of HIV-1 has revealed a pronounced heterogeneity within the *env* gene (Hahn *et al.*, 1985). Sequence variability is unevenly distributed within *env* which is divided into five major hypervariable regions (V1 to V5), interspersed with conserved regions and regions of intermediate sequence variability (Modrow *et al.*, 1987; Willey *et al.*, 1986). Studies of the genetic diversity of the V3 hypervariable region have shown that the evolution of V3 is characterised by a high rate of nonsynonymous substitution, indicative of positive selection (Balfe *et al.*, 1990; Simmonds *et al.*, 1990; Wolfs *et al.*, 1990). The V3 region contains the major target for host neutralising antibodies (Javaherian *et al.*, 1989), encoding epitopes that elicit B-cell recognition (Van Tijn *et al.*, 1989), cytotoxic (Takahashi *et al.*, 1989; Takahashi *et al.*, 1992) and helper T-cell responses (Clerici *et al.*, 1989; Takahashi *et al.*, 1990). It has been suggested that the high rate of sequence change within the region may facilitate the immunological escape of the virus from the host's immune response (Gorny *et al.*, 1992; Nara *et al.*, 1990; Wolfs *et al.*, 1991; Zwart *et al.*, 1990). The significance of sequence variation within V3 is supported by the observation that the region encodes viral determinants for cell tropism, cytopathicity and the ability of the virus to grow in transformed T-cell lines (Cheng-Mayer *et al.*, 1991; Chesebro *et al.*, 1992; De Jong *et al.*, 1992a; De Jong *et al.*, 1992b; Fouchier *et al.*, 1992; Hwang *et al.*, 1991; Milich *et al.*, 1993; Shioda *et al.*, 1991).

Studies of sequence variation within the V3 region have described the heterogeneous viral population in long-term infected individuals and showed that sequences collected from an individual shortly after infection appear to be less variable than those found in the donor at the time of transmission (Simmonds *et al.*, 1991; Zhang *et al.*, 1993). This observation supports the hypothesis that sexual or parenteral HIV-1 infection may be initiated by a limited number of molecular variants. Few studies have analysed molecular characteristics of viruses vertically transmitted to children. Recent reports based on V3 sequence analyses of mother-child transmission pairs have suggested that a very limited number of variants or even one

particular variant could initiate infection within the child (Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Wolinsky *et al.*, 1992). This genotype may represent a minor maternal form, perhaps escaping a critical immune surveillance mechanism or with particular phenotypic properties. However, the molecular and biological properties of viruses transmitted perinatally to children have not yet been determined and the mechanism of mother to child HIV-1 transmission remains unclear. The detection of HIV nucleic acids in fetal tissues (Courgnaud *et al.*, 1991; Soeiro *et al.*, 1992) has suggested that transmission to the child may occur at an early stage of gestation. On the other hand, clinical studies have reported that a high proportion of perinatally infected children show no sign of infection at delivery (Ehrnst *et al.*, 1991; Krivine *et al.*, 1992). This would suggest transmission to the child either at late stage pregnancy or at delivery. In contrast, the early diagnosis of HIV infection in newborns (within the first few days following birth) has been interpreted to reflect infection early in pregnancy. Striking differences have also been reported regarding the evolution of disease in young perinatally infected children. AIDS has been shown to develop more rapidly in 20% of perinatally infected children than in adults (Auger *et al.*, 1988; Ryder *et al.*, 1989). Such rapid progression may be associated with infection of the child at an early stage of pregnancy.

In order to elucidate the molecular characteristics of HIV-1 variants involved in mother to child transmission, we have assessed the genetic diversity of V3 DNA sequences and flanking regions (313 bp) in four perinatally infected children and in their respective mothers. Since transmission may have occurred at any time during pregnancy or at delivery, we have analysed maternal variants detected at different time points throughout pregnancy and following delivery and compared these to variants detected in the child over a period of 16 months. The analysis of longitudinal samples collected from mothers during pregnancy was required to determine whether positive selection of maternal variants occurs in HIV-1 vertical transmission. This study confirmed that a single maternal variant may be selectively transmitted to the child. However, we also report the infection of the child by multiple maternal subtypes, with evidence provided for both early and late transmission events. Vertical transmission of HIV-1 is clearly a complex process and further characterisation of the transmitted viral species may provide further insight and a better understanding of this transmission route.

MATERIALS AND METHODS

Patients

Sequence variation within the V3 domain and flanking regions was assessed in four HIV-1 infected mother/newborn pairs. All four mothers gave birth at La Grave Hospital in Toulouse, France and were followed up over a one year period from the beginning of pregnancy. The clinical status of the patients studied, time of sampling, and immunological data are presented in Table 1.

Nucleic acid extraction

DNA was obtained from patients' uncultured peripheral blood mononuclear cells (PBMCs). To avoid any risk of contamination, DNA purification was carried out in laboratories free of PCR products, cultured HIV isolates or cloned HIV sequences. After separation on Ficoll-Hypaque (Pharmacia), cells were washed twice in RPMI-1640 (Whitakker) and lysed for two hours at 56°C in 10mM Tris pH 8.3, 50mM KCl, 2.5mM MgCl2, 0.45% NP40, 0.45% Tween 20 and 80mg/ml Proteinase K. After lysis, Proteinase K was inactivated by heating for 10 minutes at 95°C.

PCR amplification, cloning and sequencing of the V3 region

The region encoding the V3 loop and flanking sequences (positions 6615 to 6928 in the HIV-LAI genome, Myers *et al.*, 1991) was amplified in a nested PCR (Mullis and Faloona, 1987). Approximately 1µg of DNA was amplified in a 100 µl reaction mixture containing 10mM Tris-HCl pH8.3, 50mM KCl, 2mM MgCl2, 0.2mM each dNTP, 0.4µg each primer, 1.5 units Taq DNA polymerase (Perkin Elmer, Cetus) and overlaid with 25µl of mineral oil. The first amplification step was performed for 25 cycles with the outer primers E1: 5'-TACAATGTACACATGGAATT-3' and E2: 5'-TTACAGTAGAAAAATTCCCC-3' (positions 6551 to 6570 and 6955 to 6974 in the HIV-LAI genome (Myers *et al.*, 1991)). Ten microliters of the first amplification product were then used as the template in a second amplification step performed for 30 cycles with primers E3: 5'-GTATCGGAATTCCTGCTGTTGAATGGC-3' (position 6592 to 6618) and E4: 5'-TTAGCAAGCTTCTGGGTCCCCTCCGAGGA-3' (position 6907 to 6935) containing EcoR I and Hind III restriction sites respectively. Thermal cycling was performed using a Perkin Elmer Cetus 9600 thermalcycler with a 94°C 1 minute denaturation step, annealing at 55°C

for 1 minute and extension at 72°C for 1 minute. This was followed by a final 2 minutes 72°C extension step. Reagent controls and DNA from uninfected cells were included in each reaction. In addition, controls from the first round of amplification were included in the second amplification step. Ten percent of the secondary PCR product (313 bp) was analysed by electrophoresis on a 2% agarose gel with the expected band visualised following ethidium bromide staining. Amplified products were digested with EcoR I and Hind III restriction enzymes, purified using the Glassmax purification system (Life Technologies, BRL) and cloned in EcoR I / Hind III digested M13 vector. The ligated vector was used to transform DH5αF' competent cells. One to seventeen individual positive M13 clones were sequenced for each sample by the dideoxy-chain termination method (Sanger *et al.*, 1977) (T7 sequencing kit, Pharmacia).

Sequence analysis

The nucleotide sequences from each mother-child pair were aligned using the CLUSTAL V algorithm (Higgins *et al.*, 1992) as implemented in version 2.2 of the Genetic Data Environment (GDE) package (kindly provided by the Harvard Genome Laboratory). The final alignment was improved manually by preferring gaps to transition differences, transition differences to transversion differences and by the insertion of gaps to maintain the reading frame. Translation of nucleotide sequences to amino-acid was also undertaken within this package. Distance-based phylogenetic analyses were performed using programs taken from version 3.52c of the Phylogeny Inference Package (PHYLIP; Felsenstein, 1993). Nucleotide sequence distances were estimated using the generalised two-parameter (maximum likelihood) model which uses the transition probability formulas of Kishino and Hasegawa (1989) incorporating unequal rates of transition and transversion and allowing for different frequencies of the four nucleotides (program DNADIST). Phylogenies were reconstructed using both the neighbour-joining method (Saitou and Nei, 1987) (program NEIGHBOR) and Fitch-Margoliash (Fitch and Margoliash, 1967) (program FITCH) distance methods. Phylogenies were also reconstructed by maximum likelihood (Felsenstein, 1981) using the modified PHYLIP program FASTDNAML (kindly provided by Gary Olsen of the University of Illinois at Urbana-Champain and the Ribosomal Database Project) (data not shown). Settings for the transition/transversion ratio were estimated from each dataset. Bootstrap resampling (Felsenstein, 1985) was employed on the neighbour-joining trees (programs SEQBOOT and CONSENSE) to assign approximate confidence limits to the branches. Two

thousand bootstrap replications were performed. Alternative phylogenetic hypothesis were evaluated statistically using the Kishino-Hasegawa-Templeton likelihood ratio test (Kishino and Hasegawa, 1989) following the assignment of log likelihoods (program DNAML) to artificially generated topologies (program RETREE) (unpublished data).

Nucleotide sequence accession numbers.

Nucleotide sequences reported in this study have been assigned the GenBank accession numbers U24717 to U24999 and U25001 to U25025.

Clinical status of patients

Analysis of genetic diversity within the V3 region and flanking sequences was performed for four HIV-1 infected mother-child pairs. The clinical status of the mothers and children, time of sampling, immunological and clinical data are presented in Table 1.

Mothers: Mothers A, B and D were of European origin and former intravenous drug users. Mother C was of African origin (Angolan), had lived in Zaire, and is most likely to have been infected heterosexually. Nevertheless, since the HIV status of her partner was unknown, a risk factor could not reliably be identified for this woman. At the commencement of pregnancy, mothers A, B and D had been infected for at least 25 months, 5 years and 7 months respectively. Diagnosis of infection was performed at 5 months of pregnancy for mother C. All mothers were asymptomatic at the beginning of pregnancy (CDC stage II and III) (Centers for Disease Control, 1987) although mothers B and D became symptomatic (oral hairy leucoplakia and multidermal herpes zoster respectively) during late pregnancy at which point they were reclassified as CDC stage IVC2. The CD4+ cell counts within the mothers varied between 276/mm^3 and 858/mm^3 with no significant variation in the mothers during pregnancy (Table 1). However, a significant decrease in the CD4$^+$ cell count was observed in mother D one year after delivery (32/mm^3) although this was not accompanied by further modification in clinical status (data not shown). Mothers A, B and D were negative for plasma p24 antigen throughout pregnancy but mother C was positive at all time points (128, 8 and 6 pg/ml for time point MC1, MC2 and MC3 respectively).

Positive cell viraemia was observed at all time points during pregnancy in all four mothers studied (Puel, unpublished data). In contrast, plasma viraemia was negative in mothers B and D and only transiently observed in mothers A and C at low levels (Puel, unpublished data). In all cases, pregnancies were free of complications and all children were delivered vaginally. The mothers did not receive antiretroviral therapy before or during pregnancy.

Newborns: Evidence of infection in newborns was provided by co-culture and/or PCR using specific primers for the *gag* and *pol* genes (Gupta *et al.*, 1991; Puel, unpublished data). Children A, B and C were diagnosed as HIV positive (PCR and co-culture positive) during the first two weeks of life (6, 5 and 12 days of age respectively). Child D was negative by PCR and co-culture at birth but became HIV positive at one month of age. All children were

originally classified as CDC stage P2A (Centers for Disease Control, 1987). Children B, C and D remained asymptomatic for the duration of the study (24 months) (Table 1). In contrast, child A displayed neurological signs at 1 month of age and was reclassified as CDC stage P2B at 3 months (Table 1). This child developed AIDS within the first year of life, with evidence of cryptosporidiasis and multidermal herpes zoster and was reclassified as CDC stage P2D at 16 months of age. All children received AZT treatment (50 to 70 mg three times a day). Children A, B and C were treated with AZT from the age of 7 weeks whereas child D received antiretroviral therapy from the age of 14 weeks.

Sequence diversity

We have analysed genetic variations in 226 complete sequences spanning the V3 region and flanked 5' and 3' respectively by 102 bp and 114 bp. Four incomplete sequences lacking only a few bases at the 5' and/or 3' end which did not lack any variable region were also included.

The within-sample genetic diversity and between-sample genetic distance were calculated for each pairwise comparison between sequences from the four mother-child pairs (see Figure 1 and Table 2). To compare the diversity between the mothers' samples and those of the infants, we have plotted each value as a histogram in Figure 1. The overall means for all sequences obtained from each patient (Figure 1e) lay between 3.5% and 5.2% for the mothers' samples, with the between-mother distances (5.3%-8.2%; mean 7.2%) (Table 2e) being higher, as expected. The within-child diversity had a greater range (0.74%-8.8%) (Figure 1e) and the between-child distances were mostly higher (7.6%-10.8%; mean 8.8%) than for the mothers (Table 2f). Comparing the overall within-patient diversity for the mothers with their infants (Figure 1) revealed two cases (pair B and D) where the infants' samples were substantially less diverse than their mothers' (Figure 1b and 1d), one where they were similar (pair A, Figure 1a), and one (pair C, Figure 1c) where substantially greater diversity was found in the child. The within-sample genetic distances showed considerable variation within each patient. In three of the four pairs, the first child sample showed lower diversity than the mother's sample closest to delivery (A, B and D; Figure 1). However, exactly the opposite was true for pair C (Figure 1c). In the mothers' samples, a progressive increase of genetic distance with time was found in pair A (Table 2a; entries off-diagonal), with the distance from sample 1 to sample 6 (time interval: 7.5 months) almost twice that to sample 2 (interval: 1 month). Again, no such trend was observed in the other samples (Table

2b-d). Examination of individual values showed that divergence between samples from the patient can reach similar levels as found when comparing different patients. Clearly a more detailed form of analysis is required to identify any evolutionary pattern that may exist in these data.

Phylogenetic analyses

The phylogenetic trees obtained for the four mother-child pairs using the neighbour-joining method are presented in Figure 2.

The sequences of the four mother-child pairs were classified according to HIV-1 global subtype classification by phylogenetic comparison with reference sequences from the five designated *env* subtypes (Myers *et al.*, 1991). The sequences of all four mother-child pairs (including sequences from mother and child C who originated from Africa) were of subtype B (European / North American) in origin (data not shown).

Mother-child pair A

Analysis of V3 sequences from the first transmission pair (pair A) led to the generation of a "star" phylogeny in which the different lineages radiated from a single point. The neighbour-joining tree reconstructed from the dataset is presented in Figure 2a. Similar topologies were obtained with each of the three methods of tree construction employed (in addition to that reconstructed based upon the amino acid sequence data). Internal branches within the phylogeny were short and bootstrap values across the tree were very low. As a consequence, there was little support for any specific grouping within the tree, although three groups of child sequences were apparent. The first group (Figure 2a, child group 1) appeared to consist predominantly of early time point sequences (1 month and 2.5 months of age) although sequences from the child's 16 months sample were also associated with the group. There appeared to be some association of these sequences with a number of second and third trimester sequences from the mother (4.5, 6 and 7 months). The second child group (Figure 2a, child group 2) consisted of three 2.5 months sequences which clustered within the main group of maternal sequences detected during pregnancy (3.5 months and 4.5 months) but appeared to be most closely related to a small number of maternal sequences from the third trimester. The third group (Figure 2a, child group 3) consisted predominantly of late time point sequences (16 months) which appeared to be associated with sequences from the 2 months post-delivery sample of the mother.

Mother-child pair B

For the second transmission pair (pair B) a clear division of all sequences into two distinct groups was apparent from the neighbour-joining tree (Figure 2b). All three tree-building methods gave similar trees from the nucleotide sequences and the division into two distinct groups was also clearly apparent from the amino acid neighbour-joining tree (data not shown). The branch separating the two groups was resolved in 100% of bootstrap replicates and child sequences were associated with both groups. The first sequence group (Figure 2b, group 1) formed the predominant lineage present within the maternal population throughout pregnancy and included sequences from all time points from 3.5 months into pregnancy until delivery. Only three child sequences, isolated from the child at the age of 5 weeks, were associated with this group. The majority of the child sequences were associated with the second group (Figure 2b, group 2) with sequence isolates from all child time points (5 days, 5 weeks and 3.5 months). The mother sequences associated with this group were mainly from the 3 months post-delivery sample which formed a tight cluster but were present as a minor form throughout pregnancy. It was interesting to note that the group one maternal variant which was predominant at delivery had been replaced with the group two sequence type by 3 months post-delivery.

Mother-child pair C

Two child groups were also clearly apparent in the neighbour-joining tree for the third transmission pair (Figure 2c). The main child group (Figure 2c, child group 1), consisting solely of sequences found at 1.5 months and 2.5 months after birth, appeared to be descended from a lineage present within the mother during the second trimester (5 and 6.5 months maternal samples) (Figure 2c, mother group 1). The lineage was represented by four out of seven sequences from the 5 months maternal sample and in only one of eleven sequences from the 6.5 months maternal sample. It was not represented in the 7.5 months sample. A single child sequence from the 13 months sample was also weakly clustered with this group. Bootstrap support for the cluster incorporating the main group of child and ancestral mother sequences was reasonably high, with the cluster supported in 87.4% of replicates. Excluding one sequence (designated by an arrow in Figure 2c) the cluster was found in 96.1% of bootstrap replicates. The second minor child group (Figure 2c, child group 2) consisted of four (of five) late time point sequences (13 months of age). The group was located within the main cluster of the maternal sequences (Figure 2c, mother group 2), closest to those found late in pregnancy (7.5 months) although only low bootstrap support was indicated for this. The overall bootstrap support for the main mother group itself was high

(96.1% excluding intermediate sequences). The phylogenies reconstructed by alternative methods for mother-child pair C were again consistent (data not shown).

Mother-child pair D

For pair D, the three methods of phylogeny reconstruction employed all identified a single child group (Figure 2d). The sequences from both the 1.5 months and 11.5 months infant samples were clustered fairly tightly within the phylogeny and showed an association with a number of maternal sequences from the 8.5 months and 4.5 months post-delivery samples. These sequences represented a minor form in both maternal samples (two of nine sequences from the 8.5 months sample and two of ten sequences from the 4.5 months post-delivery sample). Four maternal sequences, three from the 4.5 months post-delivery sample and one from the 2 months sample fell intermediate between the mother-child group and the main maternal cluster. The presence of these intermediate sequences reduced the bootstrap support, but if they were excluded the branch separating the main maternal and the mother-child group was found in 99.2% of bootstrap replicates. Once again the three methods of phylogeny reconstruction employed were highly congruent in their inferred phylogenies and again the amino acid neighbour-joining tree was very similar to that reconstructed based upon nucleotide sequence data (data not shown).

Amino acid sequence heterogeneity between mothers and children

The amino acid sequence alignments for the four mother-child pairs are presented in Figure 3. For each pair, a consensus sequence was constructed for each time point by assigning the amino acid most frequently observed in the clones to each position. Given the intra-sample diversity observed in some of the pairs it was necessary to divide sequences from a single time point into clusters of highly related clones. In these cases, a consensus sequence was constructed from each individual cluster. As the first time point consensus of the child should theoretically be considered to be the most closely related to the transmitted viral species, all consensus sequences were aligned relative to the first time point consensus of the child. Two distinct groups of variants were observed in the first child C time point and as such a single consensus sequence could not be deduced in this case. The alignment for pair C was therefore based upon the overall consensus sequence from the child.

Mother-child pair A

In pair A, the three groups of child sequences apparent from the phylogenetic tree (Figure 2a) were also apparent from the amino acid sequence alignment (Figure 3a). Group

1 sequences were present in all child samples over a 16 months period and were detected at a high proportion in the early time points CA1 and CA2, but rare in the late child time point CA3. A high homology was observed between the child amino acid sequences CA2.c2 and CA3.c2 (Figure 3a; child groups 2 and 3). Acidic amino acid residues were present in both groups at positions 320 (E or D) and 339 (D) whereas uncharged residues were specifically found at these positions (A and N respectively) in the early child A sequences (Figure 3a; child group 1). Additionally, sequences from group 2 displayed lysine residues at positions 346 and 360 which were absent in other child groups (except clone CA3.1). Comparison of mother and child amino acid sequences from pair A showed a higher similarity between the maternal sequences and child sequences from groups 2 and 3 (Figure 3a). At position 320 and 339, the majority of maternal sequences showed the acidic residues (D or E) detected in child groups 2 and 3 and displayed basic amino acids (K or R) at positions 346 and 360 as observed in child group 2. Additionally, most of the mother A sequences had a proline residue at position 299 which was also observed in the child group 3 sequences. Amino acid sequences characteristic of the late child time point, CA3, were highly related to the post delivery maternal sequences (MA6). In particular, a high degree of similarity was observed between the 2 months post-delivery maternal clone, MA6.1, and the group 3 child sequence, CA3.c2. Six conserved potential N-linked glycosylation sites were observed at positions 276, 295, 301, 331, 338 and 354 in the mother and child sequences from the three groups. Two additional potential N-linked glycosylation sites were observed at positions 289 and 360 in a small number of sequences. Both sites were observed in the post-delivery maternal sequence, MA5.1, and in three of six late time point child sequences (CA3.c2). Other CA3 time point sequences contained the N-linked glycosylation site at position 360 only.

In summary, significant amino acid changes were observed between early (CA1) and late (CA2, CA3) time point child A sequences. This was mainly reflected by non conservative acidic and basic substitutions within the V3 and flanking regions and by the appearance of two additional glycosylation sites in the 5' and 3' V3 flanking regions.

Mother-child pair B

In agreement with the neighbour-joining tree (Figure 2b), two distinct groups of child sequences could also be distinguished for pair B on the basis of their amino acid sequences (Figure 3b). The first child group included three of five sequences from sample CB2 collected at 1.5 months of age. The second child group included all sequences from the 5 days and 3.5 months samples (CB1 and CB3 respectively) and the two remaining sequences from the CB2

time point. The two groups were characterised by the following amino acid sequence patterns: group 1: ^{283}T, ^{291}S, ^{300}S,^{305}R ^{306}S ^{308}T, ^{317}T, ^{342}K, ^{345}V, ^{360}T; group 2: ^{283}S, ^{291}T, ^{300}N, ^{305}K ^{306}G ^{308}H, ^{317}A, ^{342}R, ^{345}A, ^{360}N. The majority of amino acid substitutions between groups 1 and 2 were conservative with only a single non conservative (basic) substitution observed between groups 1 and 2 at position 308 (H to T). Furthermore, the substitution at position 360 (T to N) led to the appearance of a potential N-linked glycosylation site in the child group 2 sequences. An additional potential N-linked glycosylation site was also observed in this group at position 289. Six other potential glycosylation sites were perfectly conserved in both groups of sequences. The two distinct amino acid sequence patterns detected in the child were also observed in the maternal viral population. Amino acid sequences representative of group 1 were found in mother sequences MB1.c1, MB2, MB3.2, MB3.3 and MB4 where they differed only at residue 360 (K instead of T). The group 2 amino acid sequence pattern was observed in the MB5.c and MB1.c2 consensus sequences and in the single clone MB3.1 (except for residues ^{345}V and ^{360}K).

Mother-child pair C

For pair C (Figure 3c), the two groups of child sequences observed in the neighbour-joining tree (Figure 2c; child groups 1 and 2) were also apparent from the amino acid sequence alignment. Child group 1 included sequences detected at 1.5 and 2.5 months of age (CC1 and CC2 respectively). Within this group, a number of sequences (CC1.c1 and CC2.c2) were characterised by the presence of an additional glycosylation site at position 289 (subgroup 1a). This subgroup displayed the following amino acid sequence pattern: ^{275}D, ^{317}A, ^{320}E, ^{342}V, ^{353}N, ^{354}K and ^{360}K. A second group 1 subgroup (subgroup 1b), including the sequences CC1.c2 and CC2.c1, was characterised by residues ^{275}N, ^{317}K, ^{320}D, ^{342}E, ^{353}K, ^{354}N and ^{360}N. Interestingly, one distinct variant from child group 1 (CC1.1) contained a 5' sequence region (up to position 317) characteristic of subgroup 1a and a 3' region characteristic of subgroup 1b (see Figure 3). Child group 2 included all CC3 time point sequences. The main characteristics of this group of sequences was the presence of an aspartic acid residue at position 339 and of basic amino acid residues at positions 342 (R) and 313 (K) which modified the GPGR motif to GPGK. Mother C sequences were also clustered into two distinct groups, groups 1 and 2, including the consensus sequences MC1.c1, MC2.1 and MC1.c2, MC2.c1, MC2.2, MC3.c respectively. A high degree of similarity was observed between mother group 2 and child group 2. Both groups had uncharged residues at positions 281 (N) and 285 (S) and were characterised by an aspartic acid residue at position 339 and

by an additional glycosylation site at position 279. In contrast, mother group 1 was found to be more closely related to the single consensus sequence CC3.c1 and displayed an uncharged residue at position 308 (N) and a lysine at position 342. However, comparison of child group 1 sequences with mother group 1 sequences revealed no common amino acid pattern between the two groups (see Figure 3c). Five potential N-linked glycosylation sites, located at positions 276, 295, 301, 331 and 338 were perfectly conserved in all groups of sequences from pair C. In addition to the site located at position 289, two additional sites were identified at positions 354 and 360 in all sequences from pair C except in some child group 1 sequences (subgroup 1b) and in the maternal sequence MC2.2.

Thus highly divergent sequence patterns were identified from the mother-child pair C amino acid sequences. Sequence comparison revealed a stronger correlation between the CC3.c1 late child consensus and maternal sequences detected across pregnancy. Analysis of child sequence variations revealed minor changes within early time point sequences (CC1 and CC2) whereas more significant amino acid variations were observed between early and late child sequences (CC3). Finally, recombination events between divergent viral variants were observed in this mother.

Mother-child pair D

In pair D, all child sequences were clustered in a single group characterised by the following amino acid sequence pattern: ^{293}K,^{306}S, ^{313}K, ^{317}A, ^{320}D, ^{324}D, ^{341}K, ^{342}N, ^{346}N, ^{349}G, ^{360}K (Figure 3d). Thus all child sequences were characterised by the presence of a lysine residue at position 313, giving rise to the GPGK motif within V3 in contrast to the GPGR motif observed in the majority of sequences from the other mother-child pairs. The mother D sequences were more heterogeneous than those of the child and displayed two divergent amino acid sequence patterns. The first group of maternal sequences (including the consensus sequences MD3.c2 and MD4.c2) showed an amino acid pattern similar to that of the child (Figure 3d, group 1). By contrast, the second maternal group (including the consensus sequences MD1.c1, MD2.c1, MD3.c1, MD4.c1) (see Figure 3d, group 2) diverged from the child sequences by frequent amino acid substitutions and exhibited the following representative pattern: ^{293}E, ^{306}G, ^{313}R, ^{317}T, ^{320}E, ^{324}N, ^{349}R, ^{360}N. Moreover, a number of these sequences showed an insertion of two amino acids (Q-E) at position 350. This insertion was not observed in the child sequences or in mother group 1 sequences. Some individual clones (MD1.1, MD1.3 and MD4.1) showed an intermediate pattern between the two groups displaying amino acid variations specific for both mother group 1 and 2 simultaneously

(Figure 3d). Six potential N-linked glycosylation sites located at positions 276, 295, 301, 331, 338 and 354 were perfectly conserved between the sequences of the mother and child from pair D. An additional site located at position 360 was observed in the majority of maternal sequences from group 2.

Thus, in contrast to other pairs, child D showed a homogeneous amino acid sequence at distinct time points (1 month and 11.5 months of age). Marked differences were observed between the mother and child sequences, including the modification of the V3 region central motif and the specific loss of a potential N-linked glycosylation site in the 3' region of the child's sequences.

Common patterns of amino acid substitution

Analysis of V3 sequences showed a similar distribution of basic amino acids between the four pairs with residues appearing frequently both within V3 and in flanking regions (Figure 3). In all pairs, two positions within V3 (305 and 313) displayed basic conservative substitutions showing either an arginine residue or a lysine residue. In contrast, positions 298, 304 and 326 displayed only an arginine residue whereas position 329 displayed only a histidine residue. Non conservative basic substitutions were observed in the four pairs at position 308 (H to T or N or P). Such variations were commonly detected at position 317 in children A and C (T to K and A to K respectively) and at position 327 in both the mother and child of pairs A and B (Q to K or R respectively). Additional non conservative basic substitutions were observed in one child sequence from pair A (CA1.1) at positions 301 and 302 (N to K). Conserved basic residues were also observed between the four pairs in the V3 flanking sequences. In all pairs an arginine residue was observed at positions 273 and 334 and a lysine residue was found at positions 347 and 355 (except in pair D). A conserved lysine residue was also observed at position 282 in sequences from pairs B, C and D. Additional basic substitutions outside the common pattern of distribution observed for the four pairs were found in pair A where a higher proportion of basic residues were observed in the 3' flanking region. A different distribution of basic residues was also observed between the mother and child sequences from this pair with basic residues more frequent in the 3' region of the mother's sequences but mostly found within V3 in the child's sequences.

The distribution of acidic residues within V3 and flanking regions was also examined. Negatively charged residues were observed at position 320 (D or E) and an aspartic acid residue was also frequently observed at position 324 in sequences from pairs A, B and C.

Additional acidic residues were observed within V3 at positions 306 and 321 in a small number of sequences from pair A. Acidic residues were also observed in the four pairs at positions 339 and 350 within the 3' flanking region although a larger proportion of acidic residues were observed in the 5' flanking region (positions 267, 268, 269, 275, 290 and 293).

Correlation between amino acid sequence and potential phenotype

Since the presence of basic amino acids in V3 have been implicated in a change of virus tropism and appeared to be associated with the more virulent state of HIV-1 (Chesebro *et al.*, 1992; De Jong *et al.*, 1992a; De Jong *et al.*, 1992b; Fouchier *et al.*, 1992; Milich *et al.*, 1993; Tersmette *et al.*, 1988; Tersmette *et al.*, 1989), we were especially interested in the pattern of basic amino acid substitutions within V3 and flanking regions. Previous studies have suggested that a minimum of four amino acid substitutions at positions 306, 319, 320 and 327 within V3 (according to the HIV-LAI genome) can confer macrophage-tropism and alter T-cell-line tropism (Chesebro *et al.*, 1992; De Jong *et al.*, 1992a; 12). Furthermore, Fouchier *et al.* (1992) provided evidence that HIV isolates with a T-cell-line-tropic/syncytium inducing (SI), fast-replicating phenotype have a higher net charge in V3 because of the addition of basic amino acids at critical positions for phenotype. Among sequences from the four pairs (Figure 3), only three sequences from pair A and three from pair B had basic residues at critical positions for tropism (positions 306, 320 and 327 in pair A and positions 320 and 327 in pair B). In pair A, two individual clones with a lysine at position 320 (MA1.1) and 327 (MA5.2) were observed in the mother's sequences. The third sequence was a single child sequence (CA3.1) that had both R at position 320 and K at position 327. These variants may therefore have a potential T-cell tropic/SI phenotype. In pair B, two maternal sequences had either an arginine residue at position 320 or a lysine residue at position 327 and could represent a T cell-tropic/SI phenotype. Sequences with a basic amino acid residue at position 319 were not observed in any of the mother-child sequence sets.

In conclusion, no specific amino acid substitution appeared to be discriminant between the mother and child sequences of the four pairs. In two pairs (pairs C and D), a substitution at position 313 was observed in a number of the child's sequences which resulted in the modification of the GPGR motif located at the tip of the V3 loop. Despite the high variability observed in some putative N-linked glycosylation sites, no correlation was observed between the lack of any glycosylation site near or within the V3 loop and transmission to the child.

DISCUSSION

In order to examine the molecular mechanisms involved in HIV-1 mother-to-child transmission we have analysed the genetic relationship between virus populations detected during pregnancy in four HIV infected mothers and in their respective children. Analysis was performed on a 313 bp fragment containing the V3 region and flanking sequences. We compared sequences obtained at different time points during pregnancy and post-delivery from the mothers with child sequences obtained from birth.

Genetic diversity

Sequence data obtained from the V3 region in two out of the four children sampled in this study (pairs A and C) revealed substantial levels of genetic heterogeneity in their viral populations. Such heterogeneity has not been a prominent feature of earlier studies of mother-child transmission (Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Wolinsky *et al.*, 1992). Child B also showed significant viral heterogeneity when all samples from the child were included. Only child D showed substantially less diversity than found in its mother. The occurrence of such heterogeneity in the newly infected individual is quite unlike the situation in newly infected adults. Several studies have shown very restricted levels of variation in the *env* gene in the peripheral blood of haemophiliacs and patients infected by sexual contact. This has been interpreted as evidence for selection for particular viral variants from the heterogeneous pool present in most individuals later in the infection (Balfe *et al.*, 1990; Holmes *et al.*, 1992; Simmonds *et al.*, 1991; Zhang *et al.*, 1993). The striking difference between the situation described for these children and that observed in adults suggests mother-child transmission is a more complex process.

Phylogenetic analysis of the V3 sequence data has shed more light on the circumstances of transmission in each of the four mother-child pairs.

Evidence for multiple mother-to-child transmission.

In pair A, three groups of viral sequences were identified within the child, but as they were only weakly defined it was not possible to draw any specific conclusions regarding transmission. Nevertheless, the heterogeneity of the viral population of both the mother and child suggests that the infection within the child may have involved several transmission events. It is interesting to note that the mother showed high cell viraemia throughout

pregnancy (Puel, unpublished data) which may have facilitated the transmission of multiple variants to the child.

For pair B, both the reconstructed phylogeny and the amino-acid sequence alignments clearly indicated the occurrence of two divergent populations within the child, which were statistically significant. The simultaneous presence of both populations within the 1.5 months sample of the child was reflected by a particularly high within-sample diversity for this time point. In contrast, other samples from this child displayed a lower diversity due to the presence of samples from only one of these groups. The detection of both populations within the 1.5 months sample only is probably due to the limited number of clones sequenced for the other time points. The significant association of each of the two distinct populations in the child with one of the phylogenetically distinct populations in the mother clearly indicated that the infection within the child was the result of the transmission of two distinct maternal variants. Although we cannot infer when the two transmission events occurred, the association of the minor child group sequences with maternal sequences detected across pregnancy (from 4 months until delivery) (Figure 2b: group 1), and the lack of group 1 maternal variants in the maternal post-delivery sample implied that this transmission event may have occurred *in utero*. The close association of the major child group sequences with sequences from the maternal post-delivery sample (Figure 2b: group 2) might indicate transmission at delivery. Nevertheless, since maternal variants associated with the major child group were present at low number throughout pregnancy, it was not possible to evaluate the possibility of earlier intra-uterine transmission, nor was it possible to exclude the possibility of transmission of the minor child variant at delivery. The early diagnosis of infection within this child was also consistent with the hypothesis of an early child transmission event. This transmission case is therefore clear evidence of multiple child infection with transmission potentially involving both early contamination and late transmission events.

The analysis of pair C also revealed an unusually high level of genetic diversity within the child sequences obtained within the first year of life (8-10%). Two statistically significant groups of child sequences were apparent from the phylogenetic tree with specific maternal sequences clearly associated with each group. The major child lineage was associated with maternal sequences from 5 and 6.5 months, indicating intra-uterine contamination of the child in the second trimester. Other child sequences were more closely related to maternal sequences obtained later, perhaps suggesting a late pregnancy/delivery transmission event. Child C showed early evidence of infection (diagnosis was performed

within the first days of life) providing further evidence for early infection. Thus, as previously suggested for pair B, transmission may potentially reflect both early and late infection events, although as maternal sequences were not available prior to 5 months of pregnancy we can not reliably determine the time of the first transmission.

Although evidence for the transmission of more than one variant between adults is rare, we note that it was represented for one patient of the Florida dentist (Korber and Myers, 1992; Ou *et al.*, 1992), and also for the victim of transmission by rape (Albert *et al.*, 1993).

Evidence for selective mother-to-child transmission

In contrast to the other three cases we have studied, the sequences detected in child D were highly homogeneous at all time points. These sequences clustered within a single group in the neighbour-joining tree and were closely related to a small number of maternal sequences detected during late pregnancy and following delivery. This suggests that the infection within the child is the result of the transmission of a single maternal variant and the relationship observed between the child's sequences and late time point maternal sequences supports the hypothesis of transmission either late in pregnancy or at the time of delivery. The hypothesis of late transmission is further supported in this particular case by the fact that HIV infection within child D was not detected at birth but first diagnosed at 1 month of age. The transmitted subtype represented a minor variant present within the maternal sequence population during pregnancy and interestingly, sequences obtained from child D at the age of 11.5 months remained highly homogeneous and were closely associated with sequences detected at the time of diagnosis. Samples obtained at approximately one year of age from infected children in the other transmission pairs were more heterogeneous. These results therefore support the hypothesis of selective transmission of a minor maternal variant in pair D, as claimed in other mother-child transmission studies (Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Wolinsky *et al.*, 1992).

Implications of V3 amino acid sequence variation in transmission

In comparison of amino acid sequences from the four mothers with sequences from their respective children, we were not able to find any distinctive pattern between transmitted and untransmitted viral species. We did not observe the selective loss of any glycosylation site within the V3 region in variants preferentially transmitted to the children as observed in other studies (Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993). In particular, the potential

N-linked glycosylation site located upstream to the first cysteine of the V3 loop (position 302), the absence of which was implicated by Wolinsky *et al.* (1992) as a necessary requirement for perinatal HIV-1 transmission, was usually conserved between the mother and child sequences in our study.

Genetic variations within V3 have been found to influence antibody host immune responses (Nara *et al.*, 1990; Zwart *et al.*, 1990). Furthermore, antibody titers as well as affinities for epitopes within V3 have been implicated as factors determinant in HIV-1 perinatal infection (Devash *et al.*, 1990; Rossi *et al.*, 1989). Amino acid sequence analyses within this study showed systematic variations between mother and child sequences at defined amino acid positions within V3 and flanking regions. The GPGR motif showed a high degree of conservation within the maternal sequences, probably due to the functional importance of the region. In contrast, child sequences more frequently displayed the mutated GPGK motif. The transmitted subtypes detected in infected children from pairs A, B and C displayed the GPGR motif but a conversion to GPGK was observed in all children during the first year of life. In pair D, the GPGR motif was converted to GPGK in all child sequences and a glycosylation site in the V3 3' flanking region was selectively lost in all child sequences. The mutations observed in the V3 region between mother and child D may suggest that the transmitted variant had been subjected to immune selection within the mother with the variant perhaps escaping a critical immune surveillance mechanism.

Sequence variations observed in pair D do not however themselves localise the determinants accounting for the immune escape of the virus. Other biological tests, particularly assays based on analysis of maternal neutralising antibodies and maternal cellular immunity would be required to confirm the immune escape hypothesis.

The V3 region has also been shown to influence the ability of the virus to replicate in macrophages and to grow in transformed T-cell lines (Chesebro *et al.*, 1992; Hwang *et al.*, 1991; Shioda *et al.*, 1991). Several studies have reported that the emergence of basic amino acids at specific positions within V3 are associated with the phenotypic shift of the virus from the non syncytium-inducing (NSI), macrophage tropic form to the syncytium-inducing (SI), T-cell line tropic form (Chesebro *et al.*, 1992; De Jong *et al.*, 1992b; Milich *et al.*, 1993). More precisely, significant associations have been observed between the presence of basic amino acid residues at positions 306, 319, 320, and/or 327 within V3 and the SI phenotype. Moreover, macrophage-tropic viruses have been shown to be preferentially transmitted and to be responsible for establishing chronic infection (Zhang *et al.*, 1993). Our results appear

to be in agreement with these studies. Amino acid sequence analysis of the four pairs showed that in two of the four pairs studied (pairs A and B) a limited number of maternal sequences detected during pregnancy exhibited a potential T-cell tropic syncytium-inducing (SI) phenotype. However, in pair B only sequences with a potential macrophage-tropic non syncytium-inducing phenotype (NSI) were transmitted to the child remaining over a period of 3.5 months. Therefore, despite the presence of a limited number of variants with potential lymphotropic syncytium-inducing phenotype, transmission to this child involved sequences with amino acids characteristics of macrophage tropic non syncytium-inducing variants only. In pair A, a small number of sequences with a potential SI phenotype were detected in the child at 16 months of age. Additionally, a large proportion of basic substitutions, in comparison with the mother's sequences, were observed for this child. It could not be determined whether these SI variants were transmitted by the mother or whether they emerged from sequence variations of transmitted viral species. It is interesting to note that this child showed rapid evidence of neurological symptoms and developed AIDS within 24 months. Overall no association between HIV sequence diversity and disease evolution has been established, it is interesting to note that the rapid progression to AIDS observed in child A was associated with the presence of T-cell tropic syncytium-inducing variants soon after birth as has been reported for adult infections (Koot *et al.*, 1992; Tersmette *et al.*, 1988; Tersmette *et al.*, 1989). This observation is in good agreement with the rapid decline in $CD4^+$ T-cell count and progression to AIDS reported previously in patients showing a phenotypic shift of the virus from the NSI phenotype to SI.

In conclusion, our analysis has shown that more than one process may be involved in vertical HIV-1 transmission. The selective transmission of maternal variants was observed in one case from our study. However, in two other pairs, multiple infection of the child by at least two maternal variants was clearly demonstrated. These multiple transmissions may have occurred at different times during pregnancy. Since no amino acid sequence characteristics could be defined among transmitted viral species compared to non-transmitted variants, analysis of mother-to-child transmission of HIV infection in relation to maternal HIV antibodies as well as comparison of biological properties of variants transmitted and untransmitted to the child could lead to a better understanding of viral determinants involved in HIV vertical transmission.

ACKNOWLEDGEMENTS

Special thanks are given to Dr. Tricoire from the division of Paediatrics at C.H.U. Purpan and Dr. Berrebi from the division of Gynecology-Obstetrics at La Grave hospital who provided maternal and child blood samples. We also acknowledge the help provided by Denis Lobidel at the Centre for HIV Research, Edinburgh. Laurence Briant is a fellow of the "Agence Nationale de Recherches sur le SIDA" and this work was supported by the "Agence Nationale de Recherches sur le SIDA" and the "Medical Research Council AIDS Directed Programme".

REFERENCES

Albert, J., J. Wahlberg, and M. Uhlén. 1993. Forensic evidence by DNA sequencing. Science 361:595-596.

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

Auger, I., P. Thomas, V. de Gruttola, D. Morse, D. Moore, R. Williams, B. Truman, and C. E. Lawrence. 1988. Incubation periods for pediatric AIDS patients. Nature (London) 336:575-577.

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. 64:221-6233.

Burger, H., B. Weiser, K. Flaherty, J. Gulla, P. Nguyen, and R. Gibbs. 1991. Evolution of human immunodeficiency virus type 1 nucleotide sequence diversity among close contacts. Proc. Natl. Acad. Sci. USA. 88:11236-11240.

Cheng-Mayer, C., T. Shioda, and J. A. Levy. 1991. Host range, replication and cytopathic properties of human immunodeficiency virus type-1 are determined by very few amino-acid changes in tat and gp120. J. Virol. 65:6931-6941.

Centers for Disease Control. 1987. Morbid. Mortal. Wkly. Rep. 36:225-235.

Chesebro, B., K. Wehrly, J. Nishio, and S. Perryman. 1992. Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell tropic isolates: definition of critical amino acids involved in cell tropism. J. Virol. 66:6547-6554.

Clerici, M., N. I. Stocks, R. A. Zajac, R. N. Boswell, D. C. Bernstein, D. L. Mann, G. M. Shearer, and J. A. Berzofsky. 1989. Interleukin-2 production used to detect antigenic peptide recognition by T-helper lymphocytes from asymptomatic HIV-seropositive individuals. Nature (London) 339:383-385.

Courgnaud, V., F. Lauré, A. Brossard, C. Bignozzi, A. Goudeau, F. Barin, and C. Bréchot. 1991. Frequent early *in utero* HIV-1 infection. AIDS Res. Hum. Retroviruses. 7:337-341.

De Jong, J. J., J. Goudsmit, W. Keulen, B. Klaver, W. Krone, M. Tersmette, and A. de Ronde. 1992a. Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. J. Virol. 66:757-765.

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

Devash, Y., T. A. Calvelli, D. G. Wood, K. Reagan, and A. Rubinstein. 1990. Vertical transmission of human immunodeficiency virus is correlated with the absence of high affinity/avidity maternal antibodies to the gp120 principal neutralising domain. Proc. Natl. Acad. Sci. USA 87:3445-3449.

Ehrnst, E., S. Lindgren, M. Dictor, B. Johanson, A. Sonnerborg, J. Czajkowsky, G. Sundin, and A.B. Bohlin. 1991. HIV in pregnant women and their offspring: evidence for late transmission. Lancet 338: 203-206.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783-791.

Felsenstein, J. 1993. PHYLIP Manual Version 3.52c. Berkeley Univerity Herbarium,

University of California.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science 155:279-284.

Fouchier, R. A. M., M. Groenink, N. A. Koostra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. 1992. Phenotype-associated sequence variations in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. 66:3183-3187.

Gorny, M. K., A. J. Conley, S. Karswowska, A. Buchbinder, J. Y. Xu, E. A. Emini, S. Koenig, and S. Zolla-Pazner. 1992. Neutralisation of diverse Human Immunodeficiency Virus type-1 by an anti-V3 human monoclonal antibody. J. Virol. 66:7538-7542.

Gupta, P., M. Brady, M. Raabe, and A. Urbach. 1991. Detection of human immunodeficiency virus by virus culture and polymerase chain reaction in children born to seropositive mothers. J. AIDS. 4:1004-1009.

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

Higgins, D. G., A. J. Bleasby, and R. Fuschs. 1992. CLUSTAL V: improved software for multiple sequence alignment. CABIOS, 8: 189-191.

Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludham, 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.

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

Javaherian , K. A., J. Langlois, C. Mc Danal, K. L. Ross, L. I. Eckler, C. L. Jellis, A. T. Profy, J. R. Rusche, D. P. Bolognesi, S. D. Putney, and T. J. Matthews. 1989. Principal neutralising domain of the human immunodeficiency virus type 1 envelope protein. Proc. Natl. Acad. Sci. USA. **86**:6768-6772

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order in Hominoidea. J. Mol. Evol. **4**:406-425.

Koot, M., A. H. V. Vos, R. P. M. Keet, R. E. Y. de Goede, M. Wouter Dercksen, F. G. Terpstra, R. A. Coutinho, F. Miedema, and M. Tersmette. 1992. HIV-1 biological phenotype in long term infected individuals evaluated with an MT-2 cocultivation assay. AIDS **6**:49-54.

Korber, B., and G. Myers. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retroviruses. **8**:1549-1560.

Krivine, A., G. Firtion, L. Cao, C. Francon, R. Heurion, and P. Lebon. 1992. HIV replication during the first week of life. Lancet **339**:1187-1189.

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

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.

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

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. 74:1747-1756.

Mullis, K.B., and F.A. Faloona. 1987. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. Methods. Enzymol. 155: 335-350.

Myers, G., B. Korber, J. A. Berzofsky, R. F. Smith, and G. N. Pavlakis. 1991. Human Retroviruses and AIDS (Los Alamos Nat. Lab., Los Alamos, MN).

Nara, P. L., L. Smith, N. Dunlop, W. Hatch, M. Merges, D. Waters, J. Kelliher, R. C. Gallo, P. J. Fischinger, and J. Goudsmit. 1990. Emergence of viruses resistant to neutralisation by V3-specific antibodies in experimental human immunodeficiency virus type-1 IIIB infection of chimpanzees. J. Virol. 64:3779-3791.

Ou, C-Y., C. A. Ciesieleski, G. Myers, C. I. Bandea, C-C. Luo, B. T. M. 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, H. W. Jaffe, Laboratory Investigation Group, Epidemiologic Investigation Group. 1992. Molecular epidemiology of HIV Transmission in a dental practice. Science. 256:1165-1171.

Rossi, P., V. Moschese, P. A. Brolinden, C. Fundaro, I. Quinti, A. Plebani, C. Giaquinto, P. A. Tovo, K. Ljunggren, J. Rosen, H. Wigzell, M. Jondal, and B. Wahren. 1989. Presence of maternal antibodies to human immunodeficiency virus 1 envelope glycoprotein gp120 epitopes correlates with the uninfected status of children born to seropositive mothers. Proc. Natl. Acad. Sci. USA. 86:8055-8058.

Ryder, R. W, W. Nsa, S. E. Hassig, F. Behets, M. Rayfield, B. Ekungda, A. M. Nelson, U. Mulenda, H. Francis, K. Mwanda-Galirwa, F. Davachi, M. Rogers, N. Nzilambi, A. Greenberg, J. Mann, T. C. Quinn, P. Piot, and J. W. Curran. 1989. Perinatal transmission of the human immunodeficiency virus type 1 to infants of seropositive women in Zaire. N. Engl. J. Med. 320:1637-1642.

Saag, M. S., B. H. Hahn, J. Gibbons, Y. Li, E.S. Parks, W. P. Parks, and G. M. Shaw. 1988. Extensive variation of the human immunodeficiency virus type-1 *in vivo*. Nature (London). 334:440-444.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA Sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 74:5463-5467.

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. 90:1721-1725.

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 (London) 349:167-169.

Simmonds, J., P. Balfe, C. A. Ludlam, J. O. Bishop, and A. Leigh-Brown. 1990. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type-1. J. Virol. 64:5840-5850.

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

Soeiro, R., A. Rubinstein, W. K. Rashbaum, and W. D. Lyman. 1992. Materno-foetal transmission of AIDS: Frequency of human immunodeficiency virus type 1 nucleic acids sequences in human foetal DNA. J. Infect. Dis. 166:699-703.

Takahashi, H., S. Merli, S. D. Putney, R. Houghten, B. Moss, R. N. Germain, and J. A. Berzofsky. 1989. A single amino acid interchange yields reciprocal CTL specificities for HIV-1 gp160. Science **246**:118-121.

Takahashi, H., R. N. Germain, B. Moss, and J. A. Berzofsky. 1990. An immunodominant class I restricted T-lymphocyte determinant of human immunodeficiency virus type 1 induces CD4 class II-restricted help for itself. J. Exp. Med. **171**:571-576.

Takahashi, H., Y. Nakagawa, C. D. Pendleton, R. A. Houghten, K. Yokomuro, R. N. Germain, and J.A. Berzofsky. 1992. Induction of broadly cross-reactive cytotoxic T cells recognising an HIV-1 envelope determinant. Science **255**:333-336.

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

Tersmette, M., R. A. Gruters, F. De Wolf, R. E. Y. Goede, J. M. A. Lange, P. T. A. Schellekens, J. Goudsmit, H.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 of sequential HIV isolates. J. Virol. **63**:118-2125.

Van Tijn, D. A., C. A. Boucher, M. Bakker, and J. Goudsmit. 1989. Antigenicity of linear B cell epitopes in the C1, V1 and V3 region of HIV-1 gp120. Aquired Immune Defic. Syndr. **2**:303-306.

Willey, R. L., R. A. Rutledge, S. Dias, T. Folks, T. Theodore, C. E. Buckler, and M. A. Martin. 1986. Identification of conserved and divergent domains within the envelope gene of the AIDS retrovirus. Proc. Natl. Acad. Sci. USA. **83**:5038-5042.

Wolfs, T. F., J. J. De Jong, H. Van der Berg, J. M. G. H. Tijnagel, W. J. A. Krone, and J. Goudsmit. 1990. Evolution of sequences encoding the principal neutralisation epitope of

human immunodeficiency virus 1 is host dependent, rapid and continuous. Proc. Natl. Acad. Sci. USA. 87:9938-9942.

Wolfs, T. F. W., G. Zwart, M. Bakker, M. Valk, C. L. Kuiken, and J. Goudsmit. 1991. Naturally occurring mutations within HIV 1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. Virology 185:195-205.

Wolinsky, S. M., C. M. Wike, B. T. M. Korber, C. Hutto, W. P. Parks, L. L. Rosemblum, K. Kunstman, M. R. Furtado, and J. L. Munoz. 1992. Selective transmission of human immunodeficiency virus type-1 variant from mothers to infants. Science. 255:1134-1137.

Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh-Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J Virol 67:3345.3356.

Zwart, G., H. Langedijk, L. Van der Hoek, J. J. de Jong, T. F. Wolfs, C. Ramautarsing, M. Bakker, A. de Ronde, and J. Goudsmit. 1990. Immunodominance and antigenic variation of the principal neutralisation domain of HIV-1. Virology 181:481-489.

Table 1.

Time of sampling and clinical data from the mother-child pairs studied.

D = days ; M = months; del = delivery ; p-del = post-delivery.

* CD4 cell counts are expressed as cell/mm^3. The average normal CD4$^+$ cell count value for an adult is 1000 cell/mm^3. For children before the age of 11 months the normal CD4$^+$ cell count is between 1700-2880 cell/mm^3 (mean 2200) (Tricoire, personal communication). After 1 year of age and until 6 years the average CD4$^+$ cell count is between 1000 and 1800 cell/mm^3.

a

PAIR A

	MOTHER						CHILD		
Sample N°	MA1	MA2	MA3	MA4	MA5	MA6	CA1	CA2	CA3
Timepoint	3.5 M	4.5 M	6 M	7 M	8 M	2 M p-del	1M	2.5 M	16 M
CDC stage	III	III	III	III	III	III	P2A	P2A	P2D
CD4 count*	618	-	276	470	-	480	2204	-	1626
Symptoms	none	none	none	none	none	none	none	none	cryptosporidasis herpes

b

PAIR B

	MOTHER					CHILD		
Sample N°	MB1	MB2	MB3	MB4	MB5	CB1	CB2	CB3
Timepoint	3.5 M	4 M	6.5 M	Delivery	3 M p-del	5D	1.5 M	3.5 M
CDC stage	III	III	III	IVC2	IVC2	P2A	P2A	P2A
CD4 count*	576	612	-	672	858	-	-	1912
Symptoms	none	none	none	oral leucoplakia	oral leucoplakia	none	none	none

c

PAIR C

	MOTHER			CHILD		
Sample N°	MC1	MC2	MC3	CC1	CC2	CC3
Timepoint	5 M	6.5 M	7.5 M	1.5 M	2.5 M	13 M
CDC stage	II	II	II	P2A	P2A	P2A
CD4 count*	392	468	240	-	3159	1912
Symptoms	none	none	none	none	none	none

d

PAIR D

	MOTHER				CHILD	
Sample N°	MD1	MD2	MD3	MD4	CD1	CD2
Timepoint	2 M	3.5 M	8.5 M	4.5 M p-del	1.5M	11.5 M
CDC stage	II	II	IVC2	IVC2	P2A	P2A
CD4 count*	528	314	512	364	3500	1550
symptoms	none	none	zoster	zoster	none	none

Table 2.

Nucleotide distances in the V3 loop and flanking regions of HIV-1 among mother-child transmission pairs. Nucleotide distances were estimated for each pairwise sequence comparison using the generalised two parameter model. (a to d) Nucleotide distances between mother and child time point sequences. (e) Mean nucleotide distances between all maternal sequences. (f) Mean nucleotide distances between all child sequences.

a

PAIR *A*

	MA1	MA2	MA3	MA4	MA5	MA6	CA1	CA2
MA2	0.025							
MA3	0.026	0.031						
MA4	0.032	0.035	0.034					
MA5	0.033	0.037	0.032	0.040				
MA6	0.046	0.049	0.047	0.055	0.051			
CA1	0.044	0.045	0.046	0.042	0.051	0.064		
CA2	0.039	0.041	0.041	0.037	0.048	0.060	0.031	
CA3	0.061	0.065	0.065	0.069	0.073	0.070	0.067	0.068

b

PAIR *B*

	MB1	MB2	MB3	MB4	MB5	CB1	CB2
MB2	0.055						
MB3	0.057	0.041					
MB4	0.053	0.022	0.042				
MB5	0.055	0.088	0.069	0.085			
CB1	0.055	0.089	0.069	0.086	0.004		
CB2	0.058	0.054	0.058	0.051	0.057	0.057	
CB3	0.057	0.089	0.071	0.085	0.010	0.009	0.059

c

PAIR *C*

	MC1	MC2	MC3	CC1	CC2
MC2	0.063				
MC3	0.064	0.033			
CC1	0.115	0.110	0.111		
CC2	0.108	0.103	0.103	0.080	
CC3	0.072	0.048	0.043	0.106	0.099

d

PAIR D

	MD1	MD2	MD3	MD4	CD1
MD2	0.033				
MD3	0.037	0.036			
MD4	0.042	0.038	0.042		
CD1	0.054	0.057	0.046	0.045	
CD2	0.050	0.052	0.042	0.042	0.009

e

MOTHERS

	MA	MB	MC
MB	0.053		
MC	0.082	0.062	
MD	0.082	0.079	0.076

f

CHILDREN

	CA	CB	CC
CB	0.083		
CC	0.108	0.090	
CD	0.084	0.076	0.088

Figure 1.

Within sample and between sample nucleotide sequence diversity. (a to d) Within-sample diversity in mothers (dark grey columns) and children (light grey columns). (e) Mean within-patient diversity.

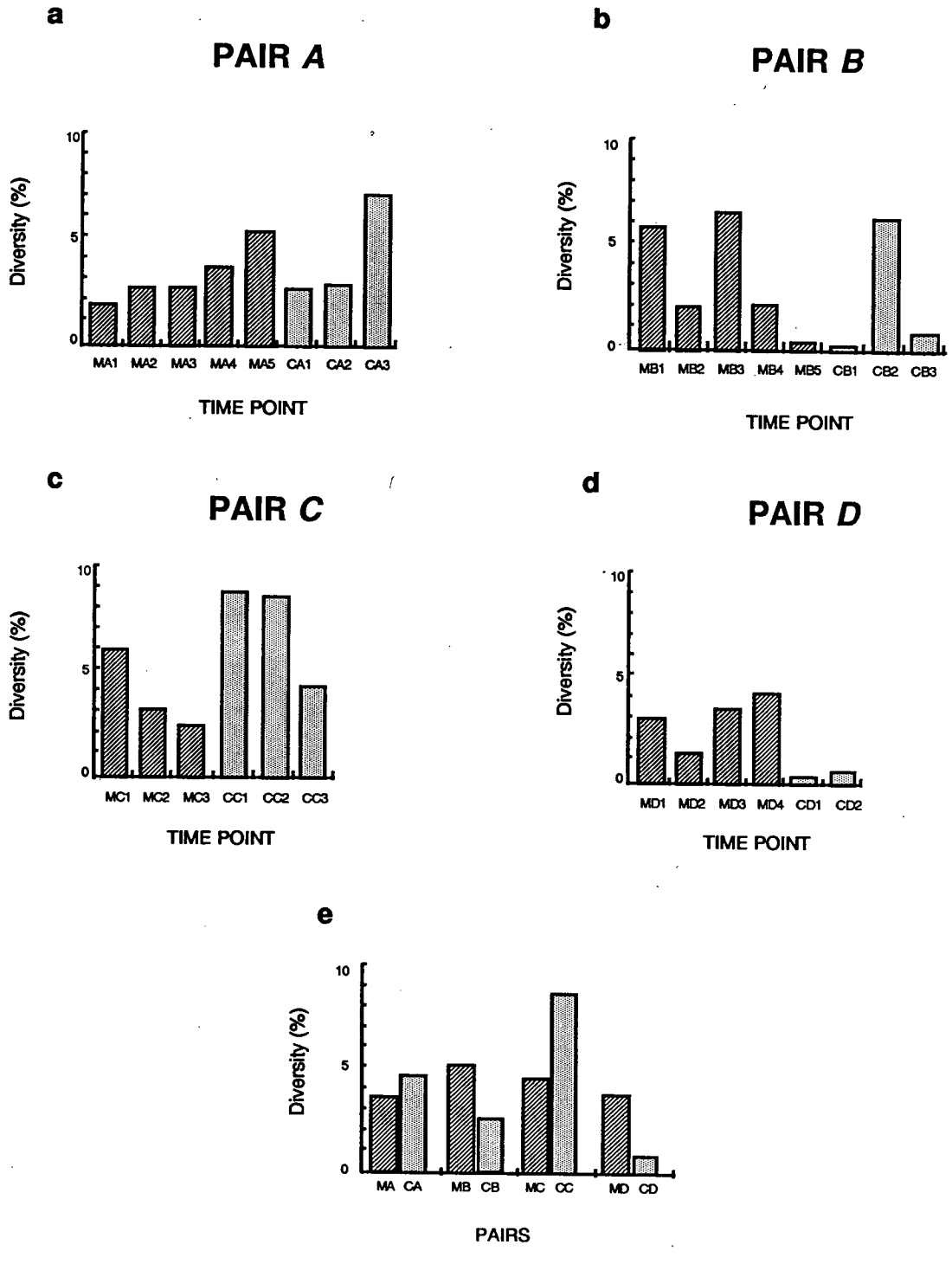

a
PAIR A
Diversity (%)
10
5
0
MA1 MA2 MA3 MA4 MA5 CA1 CA2 CA3
TIME POINT
b
PAIR B
Diversity (%)
10
5
0
MB1 MB2 MB3 MB4 MB5 CB1 CB2 CB3
TIME POINT
c
PAIR C
Diversity (%)
10
5
0
MC1 MC2 MC3 CC1 CC2 CC3
TIME POINT
d
PAIR D
Diversity (%)
10
5
0
MD1 MD2 MD3 MD4 CD1 CD2
TIME POINT
e
Diversity (%)
10
5
0
MA CA MB CB MC CC MD CD
PAIRS

Figure 2.

Unrooted neighbour-joining trees for the four mother-child transmission pairs. Symbols at the tips of each branch denote the time point to which the sequence belongs. Open symbols represent one individual child sequence, dashed symbols represent one individual maternal sequence. All branch lengths are drawn to scale.

Pair A: n = MA1; l = MA2; 9 = MA3; s = MA4; v = MA5; F = MA6; G = CA1; m = CA2; q = CA3. **Pair B**: n = MB1; l = MB2; 9 = MB3; s = MB4; F = MB5; G = CB1; m = CB2; q = CB3. **Pair C**: n = MC1; l = MC2; 9 = MC3; G = CC1; m = CC2; q = CC3. **Pair D**: n = MD1; l = MD2; s = MD3; F = MD4; G = CD1; m = CD2.

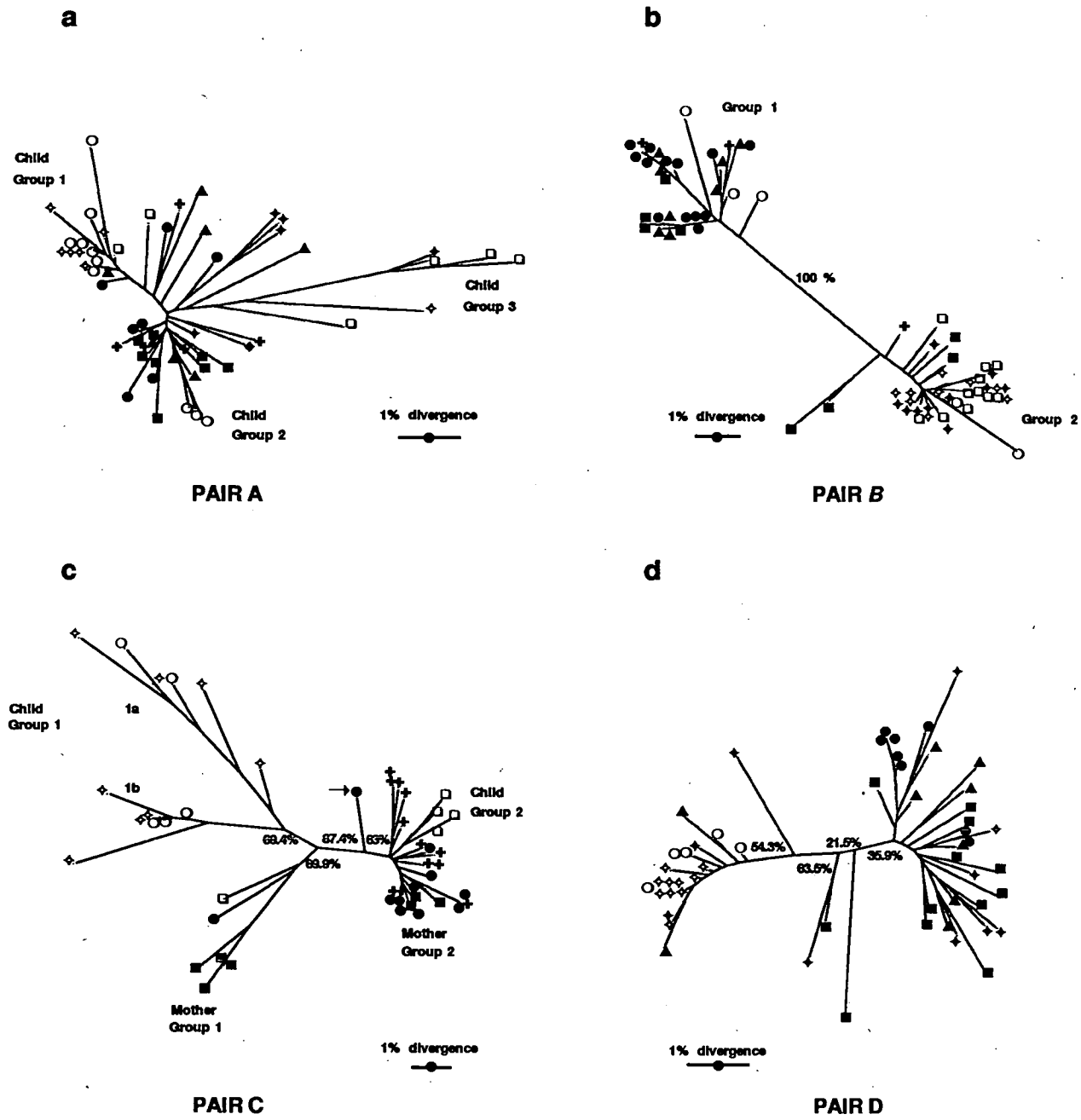
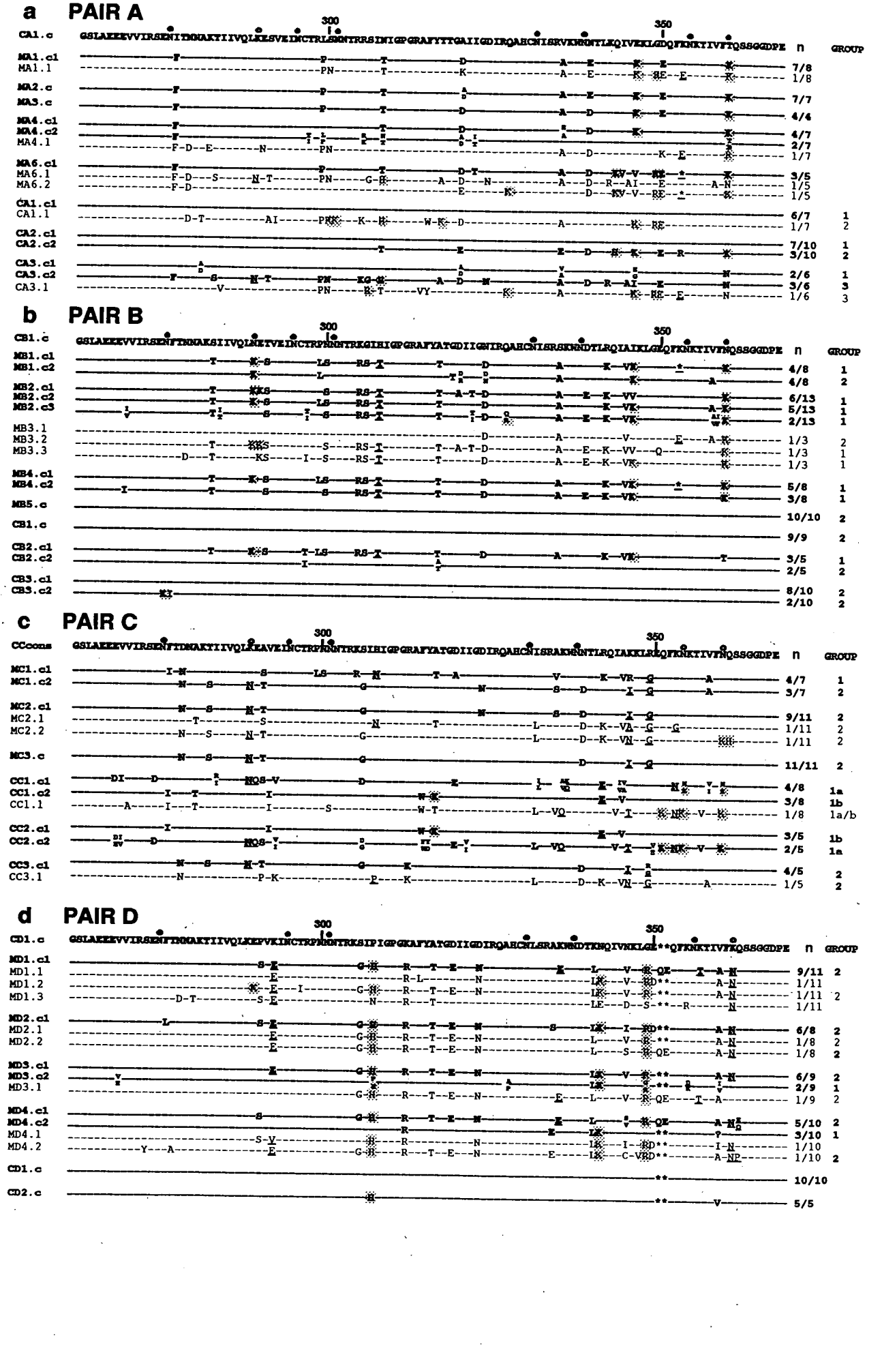

a
Child
Group 1
Child
Group 3
Child
Group 2
1% divergence
PAIR A
b
Group 1
100 %
1% divergence
Group 2
PAIR B
c
Child
Group 1
1a
1b
68.4%
87.4%
63%
89.9%
Child
Group 2
Mother
Group 2
Mother
Group 1
1% divergence
PAIR C
d
54.3%
21.5%
63.6%
35.9%
1% divergence
PAIR D

Figure 3. Amino acid sequence alignments. The amino-acid sequences of each sample are aligned with the deduced consensus sequence from the first child sample in the case of pairs A, B and D and with the consensus sequence deduced from all child time points (CCcons) in the case of pair C (see text). Consensus sequences are reported in bold text whereas individual clone sequences are reported in normal text. Amino acids are numbered according to their position in the HIV-LAI genome (Myers *et al.*, 1991). Conserved potential N-linked glycosylation sites are denoted by dark circles whereas open circles denote variable N-linked glycosylation sites. Shaded boxes indicate non conservative basic substitutions leading to the emergence of a positively charged residue as compared to the consensus. Underlined residues indicate non conservative substitutions leading to the loss of a positive charge as compared to the consensus. . indicates the deletion of a codon. n indicates the number of clones sequenced from each time point and represented by each consensus.


a PAIR A
b PAIR B
c PAIR C
d PAIR D
300
350
n
GROUP

Paper III

Evolution of Human Immunodeficiency Virus Type 1 in Perinatally Infected Infants with Rapid and Slow Progression to Disease

Francesca Salvatori[1], Sara Masiero[1], Carlo Giaquinto[2], Christopher M. Wade[3], Andrew J. Leigh Brown[3], Luigi Chieco-Bianchi[1], and Anita De Rossi[1*]

Department of Oncology and Surgical Sciences, Oncology Section, Interuniversity Center for Cancer Research, AIDS Reference Center[1], Pediatric Department[2], University of Padova, Italy, and Centre for HIV Research, University of Edinburgh, United Kingdom[3].

* Corresponding author

Journal of Virology (1997) **71**:4694-4706

ABSTRACT

We have addressed the relationship between the origin and evolution of HIV-1 variants and disease outcome in perinatally infected infants by studying the V3 region of viral variants in samples obtained from 5 transmitting mothers at delivery and sequentially over the first year of life from their infected infants, of whom 2 rapidly progressed to AIDS. Phylogenetic analyses disclosed that the V3 sequences from each mother-infant pair clustered together, and were clearly distinct from those of the other pairs. Within each pair, the child's sequences formed a monophyletic group, indicating that a single variant initiated the infection in both rapid and slow progressors. Plasma HIV-1 RNA levels increased in all 5 infants during their first months of life, and then declined within the first semester of life only in the 3 slow progressors. V3 variability increased over time in all infants, but no differences in the pattern of V3 evolution in terms of potential viral phenotype were observed. The number of synonymous and nonsynonymous substitutions varied during the first semester of life regardless of viral load, CD4+ cell count and disease progression. Conversely, during the second semester of life the rate of nonsynonymous substitutions was higher than that of synonymous substitutions in the slow, but not in the rapid progressors, thus suggesting a stronger host selective pressure in the former. In view of the proposal that V3 genetic evolution is driven mainly by host immune constraints, these findings suggest that while the immune response to V3 might contribute to regulating viral levels after the first semester of life, it is unlikely to play a determinant role in the initial viral decline soon after birth.

Running Title: HIV-1 Evolution in Perinatally infected infants.

INTRODUCTION

Mother-to-child transmission of human immunodeficiency virus type 1 (HIV-1) accounts for more than 95% of the cases of pediatric AIDS. Moreover, about one-third of these infected infants will develop severe symptoms of disease and/or severe immunodepression by one year of age. The timing of transmission, the pathogenic potential of the transmitted variants, and the host's capability to control the growth of the viral population have each been postulated to explain the observed differences in progression to disease (for review, see Pizzo *et al.*, 1995). Although the precise timing of vertical transmission cannot be pinpointed, it has been proposed that positive or negative detection of the virus in the child at birth might reflect viral transmission in utero or during the intrapartum period (Dickover *et al.*, 1994). By using highly sensitive polymerase chain reaction (PCR)-based methods to directly detect HIV-1, we and others (Borkoswky *et al.*, 1992; De Rossi *et al.*, 1992; Dunn *et al.*, 1994; Krivine *et al.*, 1992) demonstrated that a consistent proportion of infants who are subsequently recognized as infected do not have detectable virus levels in their peripheral blood cells at birth. Further studies on the dynamics of HIV-1 replication after birth (De Rossi *et al.*, 1996) and the appearance of the child's autochthonous antibodies (De Rossi *et al.*, 1993b; Rouzioux *et al.*, 1995) indicate that HIV-1 transmission mainly occurs late in pregnancy or at delivery. Most of the investigations that have evaluated the genetic diversity of the virus found a highly homogeneous viral population in newborns, thus indicating that a limited number of variants or even one initiated the infection in the infants (Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995; Scarlatti *et al.*, 1993; Strunnikova *et al.*, 1995; Wolinsky *et al.*, 1992). However, few studies have addressed the relationship between acquired HIV-1 variants and their evolution over time in relation to disease outcome (Hutto *et al.*, 1996; Strunnikova *et al.*, 1995).

An accumulating body of evidence indicates that disease progression in adults (Mellors *et al.*, 1996; Piatak *et al.*, 1993) as well as in infants (De Rossi *et al.*, 1996; Dickover *et al.*, 1994) is directly related to the level of virus replication; multiple viral factors, such as tropism and replicative capacity, and host factors, including cellular and humoral immune responses, are likely to determine virus levels.

Genetic analysis of the third variable region (V3) of the *env* gene has provided information about the potential phenotype of the virus and host-dependent constraints, since this region contains determinants involved in a number of biological properties of the virus,

including tropism and cytopathicity (Cann *et al.*, 1992; Chesebro *et al.*, 1992; de Jong *et al.*, 1992b; Hwang *et al.*, 1991; Shioda *et al.*, 1992), as well as recognition sites for both humoral and cellular T-cell immune responses (Palker *et al.*, 1988; Rusche *et al.*, 1988; Safrit *et al.*, 1994a; Safrit *et al.*, 1994b; Takahashi *et al.*, 1990; Takahashi *et al.*, 1992). Indeed, specific amino acid changes in V3 have been associated with the appearance of syncytium-inducing (SI) viral variants (de Jong *et al.*, 1992a; Fouchier *et al.*, 1992) and resistance to neutralizing antibodies (Wolfs *et al.*, 1991), and it has been proposed that a rapid evolution of antigenic sites such as the V3 region might exceed the capacity of the immune system to control virus growth (Nowak *et al.*, 1990; Nowak *et al.*, 1991). According to this theory, the V3 region would evolve more rapidly in individuals progressing to AIDS than in those who remain asymptomatic longer (Nowak *et al.*, 1990; Nowak *et al.*, 1991). A study of twins with different disease courses reported a higher increase in sequence diversity, paralleled by a lower immune response in the rapid progressor than in the slow progressor infant (Hutto *et al.*, 1996). Higher rates of sequence divergence and nonsynonymous nucleotide substitutions within the V3 region have also been reported in 2 children who progressed faster to AIDS compared to 4 slower progressing infants (Strunnikova *et al.*, 1995). However, it has recently been demonstrated that an increase in genetic diversity is correlated with a slower CD4+ cell decline and a prolonged asymptomatic period in adults who had been followed since primary infection (Wolinsky *et al.*, 1996). Furthermore, the observation that the rate of nonsynonymous nucleotide substitutions was higher in slow progressors than in subjects who progressed rapidly to AIDS and was correlated with the duration of the immunocompetent period (Lukashov *et al.*, 1995; Wolinsky *et al.*, 1996), leads to the proposal that the evolution of the viral population is driven mainly by host immune system selective forces.

In order to better clarify the transmitted variants and their evolution in relationship to disease outcome, we studied the V3 region of the viral variants present in samples collected at delivery from 5 mothers, and sequentially over the first year of life from their infected infants. The evolution of the V3 region was compared to the CD4+ cell number, plasma HIV-1 RNA levels, biological phenotype of the virus, and timing of seroconversion.

MATERIALS AND METHODS

Patients

Five HIV-1-infected mother-infant pairs were examined in this study. The women attended the Gynecology and Obstetrics Department of the University of Padova. Maternal samples were collected within 7 days of delivery. The clinical stage according to the Centers for Disease Control (CDC) (Centers for Disease Control, 1986) and the CD4+ cell count were provided with the samples. Peripheral blood samples from the children were supplied by the Pediatrics Department of the University of Padova. All were full-term infants born by spontaneous vaginal delivery; none were breast-fed. The infection status of the children was defined by virus isolation and PCR as previously described (De Rossi *et al.*, 1992; De Rossi *et al.*, 1993a). The children were followed clinically and immunologically every month during the first 3 months of life and then every 2-3 months; clinical and immunological status were defined according to the CDC criteria (Centers for Disease Control, 1994).

Sample preparation

Heparinized peripheral blood samples were centrifuged over Ficoll-Hypaque (Pharmacia, Uppsala, Sweden) density gradients; plasma was recovered from the upper phase, centrifuged at 1000 *g* for 15 min to ensure cell-free specimens, and stored at -80°C until further analysis. Peripheral blood mononuclear cells (PBMCs) were recovered from the top of the Ficoll gradient, washed twice with phosphate buffered saline, and counted. Two million PBMCs were lysed for 1 h at 56°C in 500 µl of TE buffer (10 mM Tris-HCl [pH 8] and 0.1 mM ethylenediamine tetraacetic acid [EDTA]) containing 0.001% Triton X-100, 0.0001% sodium dodecyl sulfate, and 600 µg of proteinase K/ml. After lysis, proteinase K was inactivated by heating the mixture for 15 min at 94°C. Detection of HIV-1 by PCR was performed directly on aliquots of the lysed cells as previously reported (De Rossi *et al.*, 1992). The remaining cells were resuspended in RPMI medium supplemented with 10% fetal calf serum (FCS) and 10% dimethyl sulfoxide, and then cryopreserved for biological assays.

Nested PCR and cloning

Nested PCR amplifications of proviral DNA were carried out on lysed PBMCs. Only one sample was processed at a time to avoid cross-contamination. The V3 region was amplified with the outer primer pair V3A (5' TACAATGTACACATGGAATT 3') and V3D

(5' ATTACAGTAGAAAAATTCCCC 3') and the inner primers V3B (5' TGGCAGTCTAGCAGAAGAAG 3') and V3C (5' CTGGGTCCCCTCCTGAGG 3'). The primer-binding sites were highly conserved between sequences of geographical variants of HIV-1 (Myers *et al.*, 1994); their positions in HIV-1_{IIIB} (clone HXB2) are as follows (5' to 3'): V3A, nt 6957 to nt 6976; V3B, nt 7009 to nt 7028; V3C, nt 7314 to nt 7331; and V3D, nt 7361 to nt 7381 (Starcich *et al.*, 1986). The first-round PCR reactions were performed using 25 µl of lysed PBMC, corresponding to 10^5 cells, in a 100 µl reaction mixture containing 20 nmol of each deoxynucleoside triphosphate (dNTP), 50 pmol of each of the outer primers, and 2 U of Taq DNA polymerase (Perkin-Elmer Cetus, Norwalk, CT). Forty cycles were carried out in a Gene Amp PCR System 9600 thermal cycler (Perkin-Elmer), each consisting of 30 sec at 94°C, 30 sec at 49°C and 30 sec at 72°C, followed by one cycle at 72°C for 5 min. Five microliters of the first-round PCR were amplified in a nested PCR using 80 pmol of the inner primers; 40 cycles were run, each of 30 sec at 94°C, 30 sec at 57°C, and 30 sec at 72°C, followed by a final extension at 72°C for 5 min. Appropriate negative controls (10^5 HIV-1 negative A301 cells, lysis buffer and distilled water) were included in each set of reactions. All negative controls from the first round of amplification were included in the second amplification step. The sensitivity of nested PCR was assessed by using 8E51 cells, which contain a single proviral DNA copy of HIV-1 per cell, serially diluted in 10^5 PBMC separated from heparinized blood of a healthy donor; 5 of 10 samples that were calculated to contain 1 8E51 cell yielded the expected product following nested PCR. This result is in agreement with the reported sensitivity of nested PCR to amplify a single HIV-1 copy (Mammano *et al.*, 1995).

Two separate PCR reactions for each sample were performed for subsequent cloning. The amplified products were separated on a 2.5% preparative low melting point agarose gel (NuSieve, FMC, Rockland, ME). The band of the appropriate size [322 base pairs (bp)] was excised and purified using the Wizard PCR Preps DNA Purification System (Promega, Madison, WI) according to the manufacturer's instructions, and then cloned using the TA cloning kit (Invitrogen, San Diego, CA). Four nanograms of the amplified products were ligated to 50 ng of pCR™ II vector with 4 U of T4 ligase for 16 h at 14°C, and transformed into competent cells of *Escherichia coli* strain INV α F'. Clones were color-selected on indicator plates containing 50 µg/ml of ampicillin and 40 µg/ml of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal). White colonies were selected and amplified in culture, and bacterial DNA was recovered using a DNA purification kit (BIO101, RPM Inc.,

Vista, CA), according to the manufacturer's instructions. The presence of the insert in the plasmid clones was screened by digestion with 10 U of EcoRI restriction enzyme (Boehringer Mannheim, Germany); the insert was screened for appropriate size and relative quantity by electrophoresis on an 0.8% agarose gel.

Sequencing

Nucleotide sequencing of the cloned PCR products (9 to 24 clones for each sample) was performed by the Sanger dideoxy nucleotide method, using the Sequenase version 2.0 kit (United States Biochemical Corp., Cleveland, OH). Each clone was sequenced in the forward and reverse directions using the V3B and V3C primers. One microgram of the double-stranded plasmid DNA was mixed with 12 pmol of primer and heat-denatured for 5 min; the annealing step was carried out for 1 min at room temperature. The labeling reaction was performed on ice for 3 min in a final volume of 20 µl containing 20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 25 mM NaCl, 5 mM dithiothreitol, 75 nM of each dNTP, 5 µCi [α-^{33}P]dATP (Amersham, UK) and 3 U of Sequenase. Four microliter aliquots of the mixtures were combined with 2.5 µl of each termination mixture (ddATP, ddCTP, ddTTP, ddGTP). After 5 min at 37°C, 4 µl of stop solution (95% formamide, 20 mM EDTA, 0.05% bromophenol blue and 0.05% xylene cyanol) were added. The sequencing products were denatured at 80°C for 3 min and 10 µl aliquots were resolved by electrophoresis on 8% polyacylamide denaturing gels in a sequencing apparatus (GIBCO-Bethesda Research Laboratories, Gaithersburg, MD); the gels were then exposed for 16 h to X-OMAT AR film (Eastman Kodak, Rochester, NY).

To calculate the rate of misincorporations introduced during PCR, cloning and sequencing, 10 clones from 2 different PCR amplifications of 8E51 cells were sequenced as above; only one point mutation was observed, corresponding to a misincorporation rate of 1/1890 (0.05%).

Sequence analysis

The sequence analysis program Microgenie (IntelliGenetics, Inc., Mountain View, CA) was used to record and translate V3 sequences. Nucleotide and translated amino acid sequences from each mother-child pair were aligned using the Clustal V program (Higgins *et al.*, 1992). The alignments were then adjusted by hand, and nucleotide gaps were introduced to maintain translation integrity. Consensus amino acid sequences were derived by assigning the deduced amino acid found in more than 50% of the clones to each position.

Phylogenetic analysis was conducted using programs from version 3.52c of the Phylogeny Inference Package (PHYLIP) (Felsenstein, 1993). Nucleotide sequence distances for all pairwise sequence comparisons were estimated by means of the generalized two-parameter (maximum likelihood) model, which uses the transition probability formulas of Kishino and Hasegawa (Kishino and Hasegawa, 1989) (program DNADIST). Intrasample and intersample sequence variations were expressed as the mean distance of all pairwise comparisons between sequences obtained within a sample, and from different samples, respectively. Phylogenies were reconstructed for each mother-child pair by both the neighbor-joining method (Saitou and Nei, 1987) (program NEIGHBOR) and the Fitch-Margoliash distance method (Fitch and Margoliash, 1967) (program FITCH); both tree construction methods were employed in order to increase confidence in the reconstructed phylogenies. Bootstrap resampling (Felsenstein, 1985) was applied to the neighbor-joining trees (programs SEQBOOT and CONSENSE) to assign approximate confidence limits to individual branches.

The genetic variability for each sample was calculated from the amino acid sequences by using the Simpson index, calculated as $D=\sum_i n_i^2/n^2$, where n_i denotes the number of type i sequences in the sample, and n denotes the total number of sequences (Nowak *et al.*, 1991). The number of synonymous substitutions per potential synonymous site (Ds) and the number of nonsynonymous substitutions per potential nonsynonymous site (Dn) for each set of sequences from the different time points was calculated by using the method of Nei and Gojobori (1986), incorporating the Jukes-Cantor correction for multiple substitutions, as implemented in the MEGA program (Kumar *et al.*, 1993).

Viral DNA and RNA quantitation

HIV-1 proviral DNA in cells and genomic HIV-1 RNA in plasma samples were quantified by using competitive DNA-PCR and competitive RT-PCR, respectively, as previously reported (De Rossi *et al.*, 1996). Briefly, to quantify proviral DNA, replicate portions of lysed cells (15 µl of cell sample, corresponding to 60,000 cells) were amplified along with 2 µl of increasing copy numbers of competitor plasmid pSPLI-II (De Rossi *et al.*, 1996). To quantify RNA, virion-associated RNA was obtained from plasma by an affinity capture method; replicate portions of RNA samples (2 µl, corresponding to 30 µl of plasma) were reverse-transcribed along with 2 µl of increasing copy numbers of competitor RNA transcribed from pSPLI-II. Amplification was carried out in a 100 µl final volume using 100 pg of the primer pair 1/2II, specific for highly conserved regions of HIV-1 and corresponding

to the following HIV-1 MN sequences (5'-3'): nt 696-723, upstream of the first 5' splice site (primer 1), and nt 914-889 within *gag* (primer 2II). To increase the sensitivity of the assay, the sense primer was radiolabeled with [γ-^{32}P]dATP. Forty-five amplification cycles were run, each consisting of 50 sec at 94°C, 45 sec at 62°C, and 50 sec at 72°C. A 30 µl aliquot of each 100 µl PCR reaction mixture was run on a polyacrylamide gel and exposed to an x-ray film for 2-4 h at -80°C. The sizes of the competitor and wild-type amplified products were 240 bp and 218 bp, respectively. The peak areas of the amplified bands were measured by densitometric scanning (ULTRASCAN XL, LKB, Pharmacia, Uppsala, Sweden). The logarithm of the ratio between the optical density values of competitor and wild-type amplified products (y-axis) was plotted against the logarithm of the competitor copy number (x-axis) and a linear regression curve was extrapolated.

Viral phenotype analysis

Viral isolation was performed as previously reported by culturing patient's PBMCs with PHA-stimulated donor PBMCs (De Rossi *et al.*, 1993a). Isolates were defined as rapid/high, slow/high, or slow/low according to the day of p24 antigen appearance and the levels of p24 in the supernatants of the primary coculture (De Rossi *et al.*, 1993a). HIV-1 isolates obtained from the primary cocultures were propagated by a single short-term passage (7 days) in PBMCs; supernatants were collected, centrifuged at 15000 *g* for 1 h, and then filtered through 0.22 µm filters (Millipore, Bedford, Massachussetts, USA). Virus content was quantified by measuring HIV-1 p24 protein. Individual isolates were aliquoted, stored at -80°C, and then used to evaluate infectivity in primary monocyte-derived macrophages (MDM) and PBL, and in the T-cell line MT-2, as previously described (Ometto *et al.*, 1995).

Autochthonous antibody production

De novo synthesis of antibodies to HIV-1 in infants was estimated by Western blot profiles, and/or an enzyme-linked immunosorbent assay (ELISA) using HIV-1 derived synthetic peptides as antigens, exactly as previously reported (De Rossi *et al.*, 1993b; De Rossi *et al.*, 1993c).

Nucleotide sequence accession numbers.

The sequences reported here have been deposited in the GenBank database under accession numbers U74767 to 74976 and U74987 to U75185.

RESULTS

Clinical status of patients

The characteristics of the 5 mother-child pairs studied are summarized in Table 1. Maternal blood samples were collected at delivery in 3 cases (M2, M3 and M4), and 7 days after delivery in 2 others (M1 and M5). All the mothers were of Italian origin, and were formerly intravenous drug users. All mothers but one were asymptomatic during their pregnancy, and none underwent antiretroviral therapy during pregnancy.

Four of the infants born to these mothers were tested for HIV-1 infection by PCR and virus culture at birth; infant C5 was tested on day 7, and was found to be HIV-1 positive at this first examination. The other infants were found to be HIV-1 negative at birth, but were HIV-1 positive at their second examination, which was performed within 1 month of birth (De Rossi *et al.*, 1993a; De Rossi *et al.*, 1996). According to the CDC criteria (Centers for Disease Control, 1994), the children were classified as asymptomatic (Category N), mildly symptomatic (Category A), moderately symptomatic (Category B), or severely symptomatic (Category C); the immunological status was defined as class 1 (no immunodepression), 2 (moderate immunodepression), or 3 (severe immunodepression) on the basis of the number of CD4+ cells/μl adjusted for age (Centers for Disease Control, 1994). Infants C4 and C5, who developed the symptoms listed in category C and/or severe immunodepression within the first year of life, were classified as rapid progressors, according to the definition of the Third Consensus Workshop on Pediatric AIDS. Infant C4 developed Pneumocystis carinii pneumonia by 3 months of age, while infant C5 showed severe immunodepression (< 750 CD4+ cells/μl) by 5 months of age, and developed encephalopathy by 10 months of age; infants C4 and C5 started therapy with AZT at 4 and 6 months, respectively. The remaining 3 infants (C1, C2, C3) were classified as slow progressors; within the first year of life they showed lymphadenopathy (C1 and C3) or hepatosplenomegaly (C2), and over the course of 4.5 years (mean follow-up) none has developed severe immunodepression or progressed to category C. Sequential samples collected over the first year of life at the time points listed in Table 1 were employed for genetic analysis.

Nucleotide sequence variability

To examine the origin and evolution of HIV-1 genotypes in infected infants, proviral sequences from the maternal PBMC sample at delivery and the infants' sequential samples

were amplified by nested PCR and cloned into the vector pCR™ II; a panel of 9 to 24 clones, each containing sequences spanning the V3 domain and flanked 5' and 3' by 60 bp and 21 bp respectively, was sequenced from each of the studied samples.

The infants' samples showed a more homogeneous viral population (0 to 1.53% intrasample mean nucleotide distances) than the maternal samples (1.34% to 4.12% intrasample mean nucleotide distances) (Figure 1, Panel A). The initial genetic variation of sequences from infants C1, C2, C3, C4, and C5 was 0.22%, 0.37%, 1.20%, 0.71%, and 0%, respectively. Heterogeneity increased with age in all the infants with the exception of C3; however, after an initial decrease (from 1.20% to 0.81%), an increase in sequence variability (from 0.81% to 1.53%) was detected in this child as well. The increase in HIV-1 variability over time was confirmed by an intersample analysis between sequences from the first sample and each subsequent sample, which showed an increase in the mean nucleotide distance values (Figure 1, Panel A).

The mean nucleotide distances between sequences from the mother and the child's first sample were 1.76%, 4.10%, 3.83%, 4.20%, and 0.95% for pairs 1, 2, 3, 4, and 5, respectively (Figure 1, panel A), while the lowest nucleotide distance values between sequences from mother's and child's first sample were 0.54%, 2.79%, 2.21%, 0%, and 0% for pairs 1, 2, 3, 4, and 5 (not shown); interestingly, these values correlated with the child's age at time of first sampling (Table 1). The divergence between the sequences of the child and the mother increased over time, with the exception of child 3, who showed the highest mean nucleotide variability (1.77%) over time (Figure 1, panel A).

Analysis of all sequences within and between pairs disclosed that mean nucleotide divergence values within each mother-child pair fell within the range of intrasample variability (from 0.65% to 2.58%); this was significantly lower ($p=0.0026$, Mann Whitney test) than inter-pair variability (8.85% to 14.05%) (Figure 1, panel B).

Phylogenetic analyses

A phylogenetic analysis of the sequences of the 5 mother-child pairs with reference sequences from the five *env* clades of HIV-1 disclosed that all of the sequences belonged to clade B (data not shown), in agreement with the Italian origin of the patients. A phylogenetic tree was reconstructed using the neighbor-joining method for all 409 sequences obtained for the mothers and infants in this study (Figure 2). Results showed that the sequences from each mother-infant pair clustered together; the five pairs were clearly discriminated and well

confined within subtrees. High bootstrap values were observed for the 5 mother-infant sequence sets, as would be expected for epidemiologically unrelated individuals, and excluded the possibility of contamination of the PCR reactions. Within the phylogenetic tree, the infants' sequences extended further from the putative ancestral node than those of the corresponding mothers, thus indicating a more divergent evolution. Within the subtrees, moreover, the maternal sequences were distributed into different branches, thus suggesting a pattern of multiple distinct lineages; conversely, in each infant' sequence set, most of the sequences from the first time point were clustered in a single branch, thus suggesting a monophyletic origin of the infection.

To further investigate the relationship between the maternal and infant sequences, phylogenetic analysis was performed separately for each mother-child nucleotide sequence set (Figure 3). In all pairs, the infants' sequences from the first sample were more homogeneous than those of the corresponding mother, and clustered tightly within a single branch in the phylogenetic tree, forming a monophyletic group. In pairs 2 and 3, viral variants of the child clustered distinctly from sequences of the mothers; this distinction was supported by relatively high bootstrap values (79 %, pair 2; 73 %, pair 3). In pairs 4 and 5, some maternal and infant sequences were intermingled: in pair 4, two minor maternal viral variants that were separated from the main group clustered with the child's sequence group, and in pair 5, the predominant maternal sequence was identical to the major sequence detected in the child at birth. In pair 1, maternal and infant sequences were distinct at the nucleotide level, but the maternal sequence displayed by 2 of 9 clones, and located in the phylogenetic tree nearest to the child's sequences, was identical at the amino acid level to the predominant viral variant detected in the first sample from the child (Figure 4). In infants C3 and C4, the sequences that showed the most divergence from those of the corresponding mothers were detected at the first time point (Figure 2 and Figure 3); in both cases, these sequences corresponded to variants with substitutions seen only once in the sequence set (clones C3.1.44 and C3.1.47 in child C3, and clone C4.1.12 in child C4; see Figure 4). These sequence changes could either represent in vitro introduced misincorporations or randomly occurring substitutions acquired during the first cycles of viral replication soon after birth.

The evolutionary pattern of the viral variants in C1, C2, C3, and C4 resembled a "star" phylogeny with branches radiating from the point at which the majority of the sequences detected at the first time point were clustered. In infant C3, some of the sequences derived from the latest sample were located within the phylogenetic tree nearest to the

maternal sequence (Figure 3). From this analysis, we could not discriminate between a phenomenon of convergent evolution or the emergence of additional transmitted maternal variants later in the child's life.

Amino acid sequence analysis

The deduced amino acid sequences of the V3 loop and flanking regions obtained for the 5 mother-child pairs are presented in Figure 4. The coding potential of the envelope open reading frame was maintained in most of the sequences, with 15 inactivating mutant stop codons (nonsynonymous substitutions TGG → TAG or TGA at position 62) detected in 75573 bases sequenced; this substitution was described at the same position in a previous report (Ahmad *et al.*, 1995). No frameshifts were observed among the 409 sequences, and the two cysteines involved in disulfidge bridge formation of the V3 loop (at positions 21 and 55, Figure 4) were conserved in all but one sequence (i.e., cysteine to arginine, in clone C1.1.1).

While the maternal virus population displayed a swarm of genetically distinct variants, in the first sample from each infant there was a predominant viral variant, representing from 42% (C3) to 100% (C5) of the viral population. A few minor viral variants were also detected; the differences between major and minor intrasample variants were due to randomly occurring substitutions, which were primarily observed in only one or a few sequences within each set of clones. A comparison of the child's amino acid sequences with those of the corresponding mother revealed that in all the infants the predominant variant was the one most closely related to the maternal sequence(s). In particular, in pairs 1 and 4, the child's predominant variant was identical to a minor maternal variant, while in pair 5 the child's predominant variant was identical to a major maternal variant. For mothers M4 and M5, these variants corresponded to those intermingled with the infants' sequences in the phylogenetic tree; for mother M1, this variant corresponded to the one (M1.5) showing the smallest nucleotide distance (0.54%) from the child's predominant sequence (C1.1.10). In infants C2 and C3, none of the sequences were found to be identical to those of the corresponding mother at either the nucleotide or the amino acid level; nevertheless, within each clone set, the major variant showed the smallest nucleotide and amino acid divergence from the maternal sequences. It is noteworthy that a length polymorphism was observed in pair 2, with all the child's sequences showing an amino acid insertion at position 45. Interestingly, specific amino acid residues, i.e., serine at position 38 and phenylalanine at position 40, were detected in most of the child's sequences, but only in a minor maternal genotype (M2.25).

In pair 3, a specific amino acid deletion at position 42 was observed in both the mother's and child's sequences; although no specific maternal variant more closely related to the child's sequences could be identified, the predominant variant in the child displayed the highest degree of homology with the maternal sequences. Therefore, it is likely that the predominant variant in the first sample constituted the form that initiated the infection in the child, and that minor variants were generated from the major one by point mutations.

Analysis of sequences derived from subsequent samples disclosed that the predominant variant persisted over time in each infant, even though it became less represented; indeed, its percentage within the total viral population from the first to the latest sample changed from 80% to 28%, 88% to 37%, 42% to 9%, 62% to 14%, and 100% to 52% in infants C1, C2, C3, C4 and C5, respectively. In addition to the expansion of minor variants, new variants were also detected in the subsequent samples. Based on a comparison of the phylogenetic and amino acid analyses, with the exception of child C3, none of these variants appeared more closely related to the maternal variant than the predominant variant detected in the first sample. Both conservation and divergence were observed in the central motif of the V3 loop within the mother-infant sequence sets. In pair 1, all sequences of both mother and child retained the GPGR motif; the GPGS motif was highly conserved in child C2, while all of the sequences of mother M2 but one (M2.25) showed the GPGR motif. Both the GPGK and GPGR motifs were detected in the maternal and child sequences of pair 4, while most of the sequences from pair 5 displayed the GPFR motif. In the sequence set derived from the samples of child C3, a shift in the central GPGK motif (detected in 100% of clones from the first sample) to GPGR (detected in 100% of maternal clones) was observed in 25% and 54% of clones obtained from the second and third sample, respectively.

The addition of carbohydrate chains to potential glycosylation sites may modulate the host immune response by obscuring linear epitopes, thus facilitating escape variants (Alexander and Elder, 1984; Jones and Jacob, 1991; Sodora *et al.*, 1989). Five potential N-linked glycosylation sites (NXT or NXS sequons), located at positions 1, 14, 20, 26, and 56, are present in the V3 and flanking regions (Figure 4). Four sites (located at positions 1, 20, 26, and 56) were fairly conserved among the clonal sequences of all five mother-child pairs. One site, located at position 14, was conserved in all the sequences of pairs 1, 2, and 5. In pair 4, this site was absent in all the child's sequences, and in 13 of 22 maternal sequences. In pair 3, this site was absent in the majority of the maternal sequences, and in all of the child's sequences from the first sample; in subsequent samples, however, this glycosylation

site was detected in an increasing number of sequences (2 of 24 and 9 of 22 in the second and third sample, respectively). Given that the GPGR and GPGK central motifs are common in the strains of clade B, and that the potential N-linked glycosylation sites are quite variable among primary HIV-1 isolates (Myers *et al.*, 1994), the features noted in the subsequent samples of child C3 may support a phenomenon of convergent evolution, rather than a transmission of multiple maternal genotypes to the infant.

V3 evolution in rapid and slow progressor infants

Figure 5 reports the intrasample variability and the number of synonymous and nonsynonymous nucleotide substitutions calculated for the V3 sequences from each child's sample, along with the sequential plasma HIV-1 RNA values and CD4+ cell counts (De Rossi *et al.*, 1996). As shown in Panel A of Figure 5, all infants exhibited a rapid increase in plasma viral RNA during the first 4-6 weeks of life. After this period, viral RNA levels decreased by at least 10-fold in slow progressor infants C1, C2, and C3, but remained at high levels (>1,000,000 RNA copies/ml plasma) in rapid progressor infants C4 and C5. Viral decline did not appear to correlate with the onset of HIV-1 humoral immune response, as autochthonous antibodies could be detected in 4 of the 5 infants, including the 2 rapid progressors. The time of seroconversion, estimated as the mean period between the age at the last negative and the first positive finding for autochthonous antibody production, ranged from 25 days in child C5 to 94 days in child C2 (De Rossi *et al.*, 1993b).

Intrasample variability increased progressively over time in both the rapid and slow progressor infants (Figure 5, Panel A). The pattern of synonymous and nonsynonymous substitutions was extremely varied in the first months of life, regardless of HIV-1 RNA levels, CD4+ cell count and disease progression (Figure 5, Panel B). Indeed, from the first to the second sample, the number of intrasample synonymous substitutions decreased in children C2, C3, and C4, and increased in infants C1 and C5; the number of nonsynonymous substitutions decreased in C1 and C3, and increased in children C2, C4, and C5. The mean Ds/Dn values from the first to the second sample varied from 0 to 7.1, 6.82 to 2.86, 1.57 to 1.54, and 1.83 to 0.23 in C1, C2, C3, and C4, respectively. During this same period of time, mean Ds/Dn values increased from 0 to 1.85 in C5, whose viral population was extremely homogeneous at days 7 and 48. Interestingly, the increase in the number of nonsynonymous substitutions was highest in C4, who was born to a mother seronegative for V3 epitopes, and in whom autochthonous antibodies to V3 epitopes could be documented at the estimated age

of 58 days (De Rossi *et al.*, 1993b). Therefore, although the changes in the V3 region during the first six months of life might reflect immune selective pressures, both the genetic data and antibody detection findings strongly argue against a role for the humoral response in curtailing the virus after the initial peak in slow progressor infants. In contrast, during the second six months of life, the increase in the number of nonsynonymous substitutions was comparable to that of synonymous substitutions in rapid progressor infants C4 and C5, but exceeded that of synonymous substitutions in all 3 slow progressors. Mean Ds/Dn values varied from 0.23 to 0.51 in C4 and from 1.85 to 1.09 in C5, but from 7.1 to 0, from 2.86 to 1.20 and from 1.54 to 0.61 in C1, C2, and C3 respectively, thus suggesting a stronger positive selection for changes in slow progressors than in rapid progressors.

No relationship emerged between the evolution of the V3 domain and the viral phenotype. Indeed, analysis of the viral phenotype at the time of first positive HIV-1 detection disclosed that 3 infants (C1, C4 and C5) had a rapid/high type virus according to the pattern of viral growth in primary coculture (De Rossi *et al.*, 1993a; Ometto *et al.*, 1995), but only 1 primary isolate (from C4) displayed SI capacity in MT-2 cells. Further studies performed at the last time point of genetic analysis disclosed that no shift in the viral phenotype had occurred in slow or rapid progressor infants (data not shown).

DISCUSSION

Infants with perinatally transmitted HIV-1 infection show two different patterns of disease outcome within their first months of life: some develop severe symptoms of disease and rapidly progress to AIDS, while others show a variable period of clinical latency. We studied the transmission and evolution of HIV-1 in 2 rapid and 3 slow progressor infants by analyzing maternal samples collected at delivery, and sequential samples from infants collected over their first year of life.

All but one of the infants studied were negative for viral markers at birth. This would suggest that they were most likely infected during the intrapartum period or close to the time of delivery. Only child C5 gave positive results by PCR and virus culture at first examination on day 7 after birth; in this case, transmission could have occurred either in utero or during the intrapartum period. Inter-pair phylogenetic analysis showed that sequences from each mother-infant pair were clearly distinct from one another, and clustered in a star-shaped phylogeny, indicating a pattern of evolution from a common ancestral sequence. Given the lack of direct epidemiological linkage between the pairs, this sequence may represent either a general ancestral subtype B virus, or a particular local ancestral strain. Intra-pair phylogenetic analysis supported the notion that a single V3 genetic variant initiated the infection in the infant, in agreement with observations made by others (Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995; Scarlatti *et al.*, 1993; Wolinsky *et al.*, 1992). This variant constituted the predominant form within the viral population in the first sample from the child, and with the exception of one case (C5), was found to be related to a minor maternal variant in the blood, thus suggesting a selective process during transmission or the first round of viral replication in the new host. Interestingly, the predominant variants at 1 month of age in both children C1 and C4 were identical at the amino acid level to minor variants within the maternal virus population, thus suggesting that selection occurred during transmission. None of the sequences in children C2 and C3 at 3 months of age showed identity with the maternal sequences. Although transmission early in utero could not be excluded in these two cases, it appears unlikely in light of the negative findings at birth; therefore, a selection following transmission might explain these results. Reports that viral variants show different tissue distribution, and that genital secretions may harbor distinct viral variants from blood (Overbaugh *et al.*, 1996; Zhu *et al.*, 1996), suggest the alternative explanation that transmitted variants originate from sites distinct from the peripheral blood compartment. In any case, the

comparison of the sequence data with the patterns of plasma HIV-1 RNA levels (Figure 5) suggests that viral replication in the first month of life leads to the expansion of the transmitted variant, while a selection within the host may occur only after this first phase.

The selective transmission of unglycosylated V3 variants has been demonstrated by some studies (Wolinsky *et al.*, 1992), although not confirmed in others (Ahmad *et al.*, 1995; Scarlatti *et al.*, 1993). Moreover, previous studies (Ometto *et al.*, 1995; van't Wout *et al.*, 1994) suggested that monocyte/macrophage-tropic variants are selectively transmitted from mother to child, and/or selectively replicated upon transmission. In the present study, we found that V3 glycosylation sites were highly conserved in both maternal and infant sequences, except for the site upstream of the first cysteine of the V3 domain (i.e., position 14, Figure 4), which was lacking in the majority of the sequences from mothers M3 and M4, and in all the sequences from the first sample of their infants. Furthermore, based on the presence of an uncharged or acidic amino acid at positions 11, 13, 25 and 29 within the V3 domain, most of the sequences in the mothers and infants were of potential NSI phenotype; in agreement with this, all the primary isolates from mothers and infants were able to grow in MDM cultures. A few maternal sequences predictive of the SI phenotype could be detected in M2, M3 and M4, whose isolates also displayed SI activity (Ometto *et al.*, 1995), but none of these sequences were found in their infants, although the isolate from C4 retained SI activity. Therefore, although no distinctive pattern could be seen between transmitted and non transmitted variants, nor between rapid and slow progressor infants, these findings might support the notion of a positive selection of unglycosylated monocyte/macrophage-tropic variants both during transmission and during the first cycles of viral replication in the newborn.

A major difference between rapid and slow progressor infants is that the latter are able to curtail and modulate viral replication after the first phase of replication soon after birth (De Rossi *et al.*, 1996), thus confirming the evidence that high and persistent levels of virus production are directly related to disease progression (Mellors *et al.*, 1996; Perelson *et al.*, 1996). All the infants in this study showed a rapid increase in plasma HIV-1 RNA content soon after birth, but this initial rise was followed by a decline in viral burden only in the 3 slow progressors. Both the immune response and viral phenotype may contribute in determining the different patterns of HIV-1 replication and production. Specific V3 sequences were not detected in rapid and slow progressor infants, suggesting that genetic polymorphism of the V3 region is not directly related to the early onset of AIDS in infants. The majority

of the sequences persisted over time as the NSI potential phenotype in both rapid and slow progressor infants, and accordingly, no changes in viral phenotype were documented. Only one (C4) of the two rapid progressors had an SI type virus, and viral isolates from this child retained SI characteristics despite the absence of SI type sequences within the V3 region. Although the presence of few variants with type SI sequences cannot be excluded, this finding, in agreement with previous observations (Mammano *et al.*, 1995), might indicate that epitopes other than V3 may contribute to the syncytium inducing capability of the virus.

As the V3 domain contains recognition sites for both humoral and cellular immune responses, variations within this region would be informative about the positive selective forces exerted by the host's immune system. We found that genetic variability increased over time, and did not correlate with disease progression or CD4+ cell decline. However, an analysis of the number of synonymous and nonsynonymous substitutions revealed differences between rapid and slow progressors. Indeed, during the second six months of life, the increase in the number of nonsynonymous substitutions was comparable to that of synonymous substitutions in the rapid progressor infants. Conversely, during the same period of age, the increase in the number of nonsynonymous substitutions was higher than that of synonymous substitutions in all 3 slow progressor infants, thus suggesting that an increase in genetic diversity in the slow, but not in the rapid progressors, correlated with positive selection for change. These findings differ from the observations of Strunnikova et al. (1995), and agree with recent studies in adults showing that a greater HIV-1 population complexity and a positive selection for nonsynonymous substitutions correlates with a slower CD4+ cell decline and a prolonged asymptomatic status (Delwart *et al.*, 1994; Lukashov *et al.*, 1995; Wolinsky *et al.*, 1996).

Recent data support that V3 genetic evolution is mainly driven by host immune constraints (Lukashov *et al.*, 1995; Wolinsky *et al.*, 1996). In this regard, it is noteworthy that of the 2 infants (C3 and C4) who lacked an N-linked glycosylation site at position 14 in their sequences soon after birth, only the slow progressor (C3) showed variants with a potential glycosylation site at this position in subsequent samples. As N-linked glycosylation sites could be important in the formation of conformational epitopes, and might mask linear epitopes, thus contributing to the immune escape of viral mutant forms (Alexander and Elder, 1984; Jones and Jacob, 1991; Sodora *et al.*, 1989), the different pattern observed in the two infants might reflect different immune pressure. However, it is noteworthly that during the first six months of life, the patterns of synonymous and nonsynonymous substitutions varied

from child to child regardless of the disease progression status; furthermore, mean Ds/Dn values were decreased in all 3 slow progressor infants at the last time-point, thus indicating that positive selection for changes occurred primarily in the second semester of life. This is of particular interest, because in the 3 slow progressor infants viral decline occurred earlier, during the third month of life. Taking into consideration the proposal that positive selection for amino acid changes in V3 is mainly driven by host immune system forces, these findings suggest that the immune response to V3 contributes to regulating viral levels after the first semester of life, but is unlikely to play a determinant role in curbing the virus after its rapid spread soon after birth. Interestingly, child C4 developed early AIDS despite the onset of antibody production against the V3 domain (De Rossi *et al.*, 1993b). It has been proposed that cellular immunity rather than humoral immune responses might play a role in restraining the virus after the primary infection (Borrow *et al.*, 1994; Ferbas *et al.*, 1996; Kroup *et al.*, 1994). The finding that HIV-1 specific CTL are only rarely detected in infected newborns (Luzuriaga *et al.*, 1991; Luzuriaga *et al.*, 1995), and our observation that positive selection for changes in V3 were mainly detected after the first semester of life suggest that factors other than the immune response to V3 play a crucial role in controlling virus replication soon after birth.

ACKNOWLEDGMENTS

We thank Patricia Segato and Donna D'Agostino for help in preparing the manuscript and Pierantonio Gallo for the artwork. This work was supported by grants from the Istituto Superiore di Sanita', Progetto AIDS No. 9302-04 and No. 9402-07.

REFERENCES

Ahmad, N., B. M. Baroudy, R. C. Baker, and C. Chappey. 1995. Genetic analysis of human immunodeficiency virus type 1 envelope V3 region isolates from mothers and infants after perinatal transmission. J. Virol. **69**:1001-1012.

Alexander, S., and J. H. Elder. 1984. Carbohydrate dramatically influences immune reactivity of antisera to viral glycoprotein antigens. Science **226**:1328-1332.

Borkoswky, W., K. Krasinski, H. Pollack, W. Hoover, A. Kaul, and T. Ilmet-Moore. 1992. Early diagnosis of human immunodeficiency virus infection in children <6 months of age: comparison of polymerase chain reaction, culture, and plasma antigen capture techniques. J.Infect.Dis. **166**: 616-619.

Borrow, P., H. Lewicki, B. H. Hahn, G. M. Shaw, and M. B. A. Oldstone. 1994. Virus-specific CD8+ CTL activity associated with control of viremia in primary HIV-1 infection. J. Virol. **68**:6103-6110.

Cann, A. J., M. J. Churcher, M. Boyd, W. O'Brien, J.-Q. Zhao, J. Zack, and L. S. Y. Chen. 1992. The region of the envelope gene of human immunodeficiency virus type 1 responsible for determination of cell tropism. J. Virol. **66**:305-309.

Centers for Disease Control. 1986. Classification system for human T lymphotropic virus type III/lymphadenopathy-associated virus infections. Morb. Mortal. Weekly Rep. **35**:334-339.

Centers for Disease Control. 1994. Revised classification system for HIV-1 infection in children less than 13 years of age. Morb. Mortal. Weekly Rep. **43**:1-10.

Chesebro, B., K. Wehrly, J. Nishio, and S. Perryman. 1992. Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: definition of critical amino acids involved in cell tropism. J. Virol. **66**:6547-6554.

de Jong, J. J., A. De Ronde, W. Keulen, M. Tersmette, and J.Goudsmit. 1992a. 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.

de Jong, J. J., J. Gudsmit, W. Keulen, B. Klaver, W. Krone, M. Tersmette, and A. de Ronde. 1992b. Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. J. Virol. 66:757-765.

De Rossi, A., L. Ometto, F. Mammano, C. Zanotto, C. Giaquinto, and L. Chieco-Bianchi. 1992. Vertical transmission of HIV-1: lack of detectable virus in peripheral blood cells of infected children at birth. AIDS 6:1117-1120.

De Rossi, A., C. Giaquinto, L. Ometto, F. Mammano, C. Zanotto, D. T. Dunn, and L. Chieco-Bianchi. 1993a. Replication and tropism of human immunodeficiency virus type 1 as predictors of disease outcome in infants with vertically acquired infection. J. Pediatr. 123: 929-936.

De Rossi, A., L. Ometto, F. Mammano, C. Zanotto, A. Del Mistro, C. Giaquinto, and L. Chieco-Bianchi. 1993b. Time course of antigenaemia and seroconversion in infants with vertically acquired HIV-1 infection. AIDS 7:1528-1529.

De Rossi, A., C. Zanotto, F. Mammano, L. Ometto, A. Del Mistro, and L. Chieco-Bianchi. 1993c. Pattern of antibody response against the V3 loop in children with vertically acquired immunodeficiency virus type 1 (HIV-1) infection. AIDS Res. Hum. Retroviruses 9:221-228.

De Rossi, A., S. Masiero, C. Giaquinto, E. Ruga, M. Comar, M. Giacca, and L. Chieco-Bianchi. 1996. Dynamics of viral replication in infants with vertically acquired human immunodeficiency virus type 1 infection. J.Clin. Invest. 97:323-330.

Delwart, E. L., H. W. Sheppard, B. D. Walker, J. Goudsmit, and J. I. Mullins. 1994. Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays. J. Virol. 68:6672-6683.

Dickover, R. E., M. Dillon, S. G. Gillette, A. Deveikis, M. Keller, S. Plaeger-Marshall, I. Chen, A. Diagne, E. R. Stiehm, and Y. Bryson. 1994. Rapid increases in load of human immunodeficiency virus correlate with early disease progression and loss of CD4 cells in vertically infected infants. J. Infect. Dis. **170**:1279-1284.

Dunn, D. T., C. D. Brandt, A. Krivine, S. A. Cassol, P. Roques, W. Borkowsky, A. De Rossi, E. Denamur, A. Ehrnst, C. Loveday, J.-A. Harris, K. McIntosh, A. M. Comeau, T. Rakusan, M.-L. Newell, and C. S. Peckham. 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 **9**:F7-F11.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783-791.

Felsenstein, J. 1993. PHYLIP manual version 3.52c. Berkeley University Herbarium, University of California, Berkeley.

Ferbas, J., E.S. Daar, K. Grovit-Ferbas, W.J. Lech, R. Detels, J.V. Giorgi, and A.H. Kaplan. 1996. Rapid evolution of human immunodeficiency virus strains with increased replicative capacity during the seronegative window of primary infection. J.Virol. **70**: 7285-7289.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science **155**:279-284.

Fouchier, R. A. M., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. **66**:3183-3187.

Higgins, D. G., A. J. Bleasby, and R. Fuschs. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Applic. Biosci. **8**:189-191.

Hutto, C., Y. Zhou, J. He, R. Geffin, M. Hill, W. Scott, and C. Wood. 1996. Longitudinal studies of viral sequence, viral phenotype, and immunologic parameters of human immunodeficiency virus type 1 infection in perinatally infected twins with discordant disease courses. J. Virol. 70:3589-3598.

Hwang, S. S., T. J. Boyle, K. H. Lyerly, and B. R. Cullen. 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science 253:1590-1593.

Jones, I. M., and G. S. Jacob. 1991. Anti-HIV drug mechanism. Nature 352:198 (letter).

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order in Hominoidea. J. Mol. Evol. 4:406-425.

Krivine, A., G. Firtion, L. Cao, C. Francoual, R. Henrion, and P. Lebon. 1992. HIV replication during the first weeks of life. Lancet 339:1187-1189.

Kroup, R. A., J. T. Safrit, Y. Cao, C. A. Andrews, G. McLeod, W. Borkowsky, C. Farthing, and D. D. Ho. 1994. Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. J. Virol. 68:4650-4655.

Kumar, S., K. Tamura, and M. Nei. 1993. MEGA: molecular evolutionary genetic analysis, version 1.02. Pennsylvania State University, University Park, Pa.

Lukashov, V. V., C. L. Kuiken, and J. Goudsmit. 1995. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J. Virol. 69:6911-6916.

Luzuriaga, K., R. Koup, C. Pikora, D. Brettler, and J. Sullivan. 1991. Deficient human immunodeficiency virus type-1 specific cytotoxic T cell responses in vertically infected children. J. Pediatr. 119:230-236.

Luzuriaga, K., D. Holmes, A. Hereema, J. Wong, D. L. Panicali, and J. L. Sullivan. 1995. HIV-1 specific cytotoxic T lymphocyte responses in the first year of life. J. Immunol. 154:433-443.

Mammano, F., F. Salvatori, L. Ometto, M. Panozzo, L. Chieco-Bianchi, and A. De Rossi. 1995. Relationship between the V3 loop and the phenotypes of human immunodeficiency virus type 1 (HIV-1) isolates from children perinatally infected with HIV-1. J. Virol. 69:82-92.

Mellors, J.W., C.R. Rinaldo, Jr., P.Gupta, R.M. White, J.A. Todd, and L.A: Kingsley. 1996. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science 272:1167-1170.

Mulder-Kampinga, G. A., A. Simonon, C. L. Kuiken, J. Dekker, H. J. Scherpbier, P. van de Perre, K. Boer, and J. Goudsmit. 1995. Similarity in *env* and *gag* genes between genomic RNAs of human immunodeficiency virus type 1 (HIV-1) from mother and infant is unrelated to time of HIV-1 RNA positivity in the child. J. Virol. 69:2285-2296.

Myers, G., S. Wain Hobson, L. E. Henderson, B. Korber, K.-T. Jeang, and G. N. Pavlakis (ed.). 1994. Human retroviruses and AIDS 1994. Los Alamos National Laboratory, Los Alamos, New Mexico.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418-426.

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.

Nowak, M. A., R. M. Anderson, A. R. McLean, T. F. W. Wolfs, J. Goudsmit, and R. M. May. 1991. Antigenic diversity thresholds and the development of AIDS. Science 254:963-969.

Ometto, L., C. Zanotto, A. Maccabruni, D. Caselli, D. Tuscia, C. Giaquinto, E. Ruga, L.

Chieco-Bianchi, and A. De Rossi. 1995. Viral phenotype and host-cell susceptibility to HIV-1 infection as risk factors for mother-to-child HIV-1 transmission. AIDS 9:427-434.

Overbaugh, J., R.J. Anderson, J.O. Ndinya-Achola, and J.K. Kreiss. 1996. Distinct but related Human Immunodeficiency Virus type 1 variant populations in enital secretions and blood. AIDS Res. Hum. Retroviruses 12: 107- 115.

Palker, T. J., M. E. Clark, A. J. Langlois, T. J. Matthews, K. J. Weinhold, R. R. Randall, D. P. Bolognesi, and B. F. Haynes. 1988. Type-specific neutralization of the human immunodeficiency virus with antibodies to env-encoded synthetic peptides. Proc. Natl. Acad. Sci. USA 85:1932-1936.

Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582-1586

Piatak, M., M. S. Saag, L. C. Yang, S. J. Clark, J. C. Kappes, K. C. Luk, B. H. Hann, G. M. Shaw, and J. D. Lifson. 1993. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. Science 259:1749-1754.

Pizzo, P. A., M. Wilfert, and the Pediatric AIDS Siena Workshop II. 1995. Markers and determinants of disease progression in children with HIV infection. J. Acquir. Immune Defic. Syndr. 8:30-44.

Rouzioux, C., D. Costagliola, M. Burgard, S. Blanche, M. J. Mayaux, C. Griscelli, A.-J. Valleron, and the HIV Infection in Newborns French Collaborative Study Group. 1995. Estimated timing of mother-to-child human immunodeficiency virus type 1 (HIV-1) transmission by use of a Markov model. Am. J. Epidemiol. 142:1330-1337.

Rusche, J. R., K. Javaherian, C. McDanal, J. Petro, D. L. Lynn, R. Grimaila, A. Langlois, R. C. Gallo, L. O. Arthur, P. J. Fischinger, D. P. Bolognesi, S. D. Putney, and T. J. Matthews. 1988. Antibodies that inhibit fusion of human immunodeficiency virus-infected cells bind a 24-amino acid sequence of the viral envelope gp120. Proc. Natl. Acad. Sci. USA 85:3198-

3202.

Safrit, J. T., A. Y. Lee, C. A. Andrews, and R. A. Koup. 1994a. A region of the third variable loop of HIV-1 gp120 is recognized by HLA-B7-restricted CTLs from two acute seroconversion patients. J. Immunol. 153:3822-3830.

Safrit, J. T., C. A. Andrews, T. Zhu, D. D. Ho, and R. A. Koup. 1994b. Characterization of HIV-1-specific CTL clones isolated during acute serum conversion: recognition of autologous virus sequences within a conserved immunodominant epitope. J. Exp. Med. 179:463-472.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Scarlatti, G., T. Leitner, E. Halapi, J. Wahlberg, P. Marchisio, M. A. Clerici-Schoeller, H. Wigzell, E. M. Fenyö, J. Albert, M. Uhlén, and P. Rossi. 1993. Comparison of variable region 3 sequences of human immunodeficiency virus type 1 from infected children with RNA and DNA sequences of the virus populations of their mothers. Proc. Natl. Acad. Sci USA 90:1721-1725.

Shioda, T., J. A. Levy, and C. Cheng-Mayer. 1992. Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell line and macrophage tropism of human immunodeficiency virus type 1. Proc. Natl. Acad. Sci. USA 89:9434-9438.

Sodora, D. L., G. H. Cohen, and R. J. Eisenberg. 1989. Influence of asparagine-linked oligosaccharides on antigenicity, processing, and cell surface expression of herpes simplex virus type 1 glycoprotein d. J. Virol. 63: 5184-5193.

Starcich, B. R., B. H. Hahn, G. M. Shaw, P. D. McNeely, S. Modrow, H. Wolf, E. S. Parks, W. P. Park, S. F. Josephs, R. C. Gallo, and F. Wong-Staal. 1986. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell 64:637-648.

Strunnikova, N., S. C. Ray, R. A. Livingston, 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.

Takahashi, H., R. N. Germain, B. Moss, and J. A. Berzofsky. 1990. An immunodominant class I-restricted cytotoxic T lymphocyte determinant of human immunodeficiency virus type 1 induces CD4 class II-restricted help for itself. J. Exp. Med. 171:571-576.

Takahashi, H., Y. Nakagawa, C. D. Pendleton, R. A. Houghten, K. Yokomuro, R. N. Germain, and J. A. Berzofsky. 1992. Induction of broadly cross-reactive cytotoxic T cells recognizing an HIV-1 envelope determinant. Science 255:333-336.

Third Consensus Workshop on Pediatric AIDS. Virological and Immunological features of HIV-1 infants defined as long-term non progressors. Spoleto, May 4-7, 1995.

van't Wout, A. B., N. A. Kootstra, G. A. Mulder-Kampinga, N. Albrecht-van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker. 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral, and vertical transmission. J. Clin. Invest. 94:2060-2067.

Wolfs, T. F. W., G. Zwart, M. Bakker, M. Valk, C. L. Kuiken, and J- Goudsmit. 1991. Naturally occurring mutations within HIV-1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. Virology 185:195-205.

Wolinsky, S. M., C. M. Wike, B. T. M. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Munoz. 1992. Selective transmission of human immunodeficiency virus type 1 variants from mothers to infants. Science 255:1134-1137.

Wolinsky, S. M., B. T. M. Korber, A. U. Neumann, M. Daniels, K. J. Kunstman, A. J. Whetsell, M. R. Furtado, Y. Cao, D. D. Ho, J. T. Safrit, R. A. Koup. 1996. Adaptative evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science 272:537-542.

Zhu, T., N. Wang, A.Carr, D.S. Nam, R.M. Jankowski, D.A.Cooper, and D. Ho. 1996.

Genetic characterization of human immunodeficiency virus type 1 in blood and genital secretions: evidence for viral compartmentalization and selection during sexual transmission. J.Virol. 70: 3098-3107.

Table 1.

Times of sampling and clinical data for mother-child transmission pairs.

		Patient code	Clinical stage[a]	HIV-1 detection at birth	Time of testing for genetic analyses (days after delivery/birth)
PAIR 1	M1	2222	IV		7
	C1	2243	N1	-	37
		2415	A1		184
		2564	A1		296
PAIR 2	M2	891	II		1
	C2	982	N1	-	77
		1143	A1		186
		1525	A1		400
PAIR 3	M3	892	II		1
	C3	996	A1	-	85
		1145	A1		190
		1461	A1		384
PAIR 4	M4	1101	II		1
	C4	1102	B1	-	29
		1223	C1		90
		1588	C2		315
PAIR 5	M5	2693	II		7
	C5	2663	A1	+	7
		2726	B1		48
		2848	B3		137
		3223	C3		390

Figure 1.

Nucleotide sequence variation in the V3 region of the *env* gene in five mother-infant pairs. Intrasample and intersample mean nucleotide distances within mother-infant pairs (A) and between mother-infants pairs (B). Shown on the diagonal are the mean nucleotide distances from all pairwise intrasample (A) and intra-pair (B) comparisons. Shown below the diagonal are the mean distances between intersample (A) and inter-pair (B) comparisons. n indicates the number of sequences analyzed for each sample (A) and pair (B).

A

Intrasample and intersample mean nucleotide distances within each pair

PAIR 1

	n	M1	C1.1	C1.2	C1.3
M1	9	1.34			
C1.1	20	1.76	0.22		
C1.2	14	1.77	0.22	0.23	
C1.3	14	2.41	0.85	0.86	0.72

PAIR 2

	n	M2	C2.1	C2.2	C2.3
M2	18	2.02			
C2.1	17	4.10	0.37		
C2.2	22	4.27	0.48	0.56	
C2.3	19	4.51	0.97	1.03	1.38

PAIR 3

	n	M3	C3.1	C3.2	C3.3
M3	22	1.62			
C3.1	19	3.83	1.20		
C3.2	24	3.27	1.10	0.81	
C3.3	22	3.11	1.77	1.37	1.53

PAIR 4

	n	M4	C4.1	C4.2	C4.3
M4	22	4.12			
C4.1	21	4.20	0.71		
C4.2	20	4.44	0.89	0.91	
C4.3	21	4.51	1.19	1.32	1.39

PAIR 5

	n	M5	C5.1	C5.2	C5.3	C5.4
M5	20	1.62				
C5.1	19	0.95	0.00			
C5.2	22	0.98	0.02	0.05		
C5.3	22	1.04	0.10	0.12	0.20	
C5.4	21	1.54	0.60	0.62	0.69	1.15

B

Mean nucleotide distances within and between pairs

	n	PAIR 1	PAIR 2	PAIR 3	PAIR 4	PAIR 5
PAIR 1	57	0.93				
PAIR 2	77	9.65	2.14			
PAIR 3	87	9.71	10.42	2.13		
PAIR 4	84	8.85	13.35	10.87	2.58	
PAIR 5	104	12.62	13.27	12.21	14.05	0.65

Figure 2.

Unrooted neighbor-joining tree of 409 V3 nucleotide sequences obtained from the five mother-child pairs. Branch lengths are drawn to scale. The scale bar corresponds to 1% nucleotide sequence divergence. Bootstrap values are expressed as percentages for each branch, and represent the percent occurrence of that branch per 100 bootstrap replicates. Open symbols denote maternal sequences; closed symbols indicate children's sequences. ● first sample; ▲ second sample; ■ third sample; ◆ fourth sample. Numbers refer to the number of identical sequences identified in a given sample. * denotes the position in the tree of the set of sequences included in the bracket.

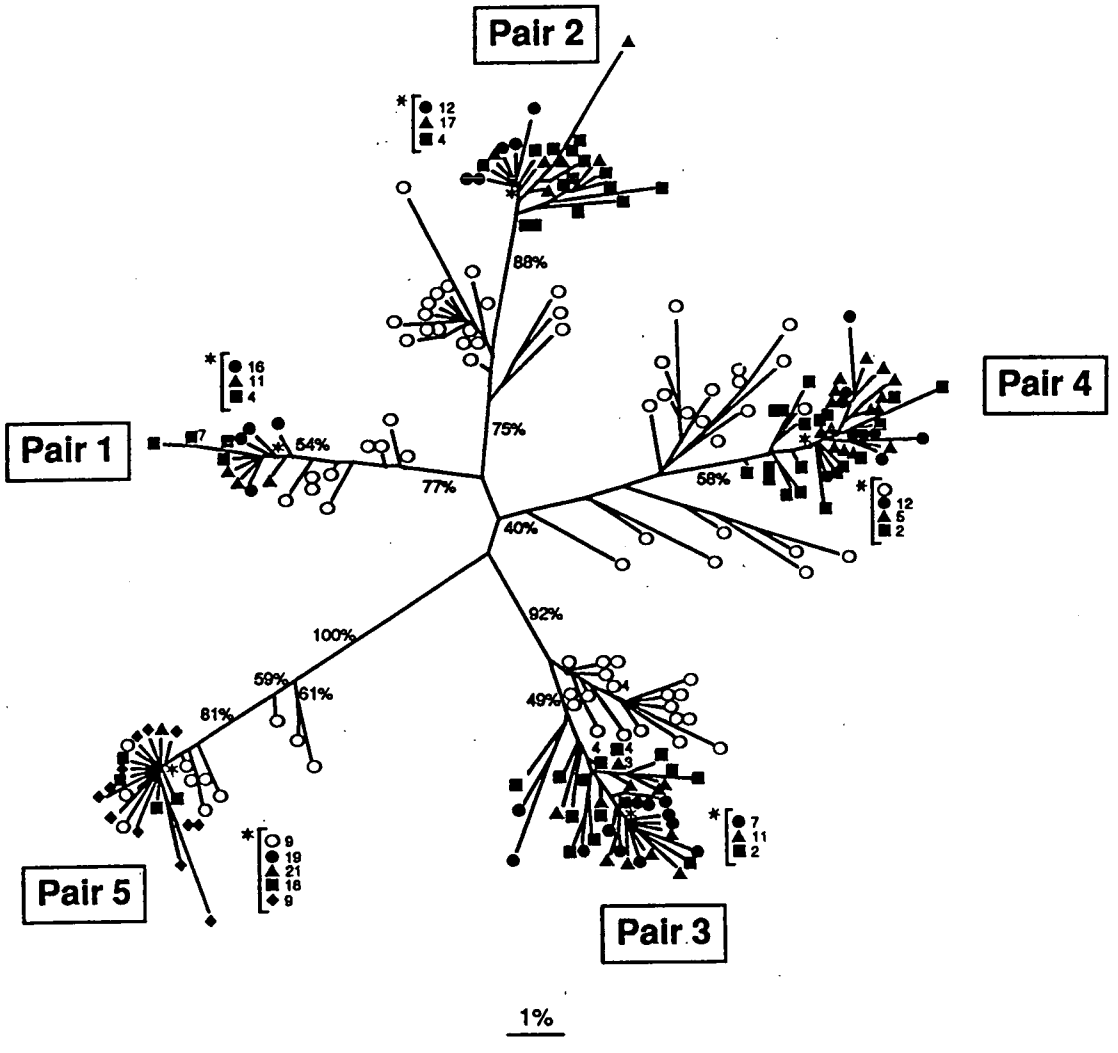

Pair 2
Pair 1
Pair 4
Pair 5
Pair 3
88%
75%
77%
54%
58%
40%
92%
100%
59%
61%
81%
49%
1%

Figure 3.

Neighbor-joining trees for each of the five mother-infant pairs. Branch lengths are drawn to scale. The scale bars correspond to 1% nucleotide sequence divergence. The number at the node indicates the proportion of support in 2000 bootstrap replicates. Open symbols indicate mother's sequences; closed symbols indicate child's sequences: ● first sample, ▲ second sample, ■ third sample, ◆ fourth sample. Numbers refer to the number of identical sequences identified in a given sample. * denotes the position in the tree of the set of sequences included in the bracket.

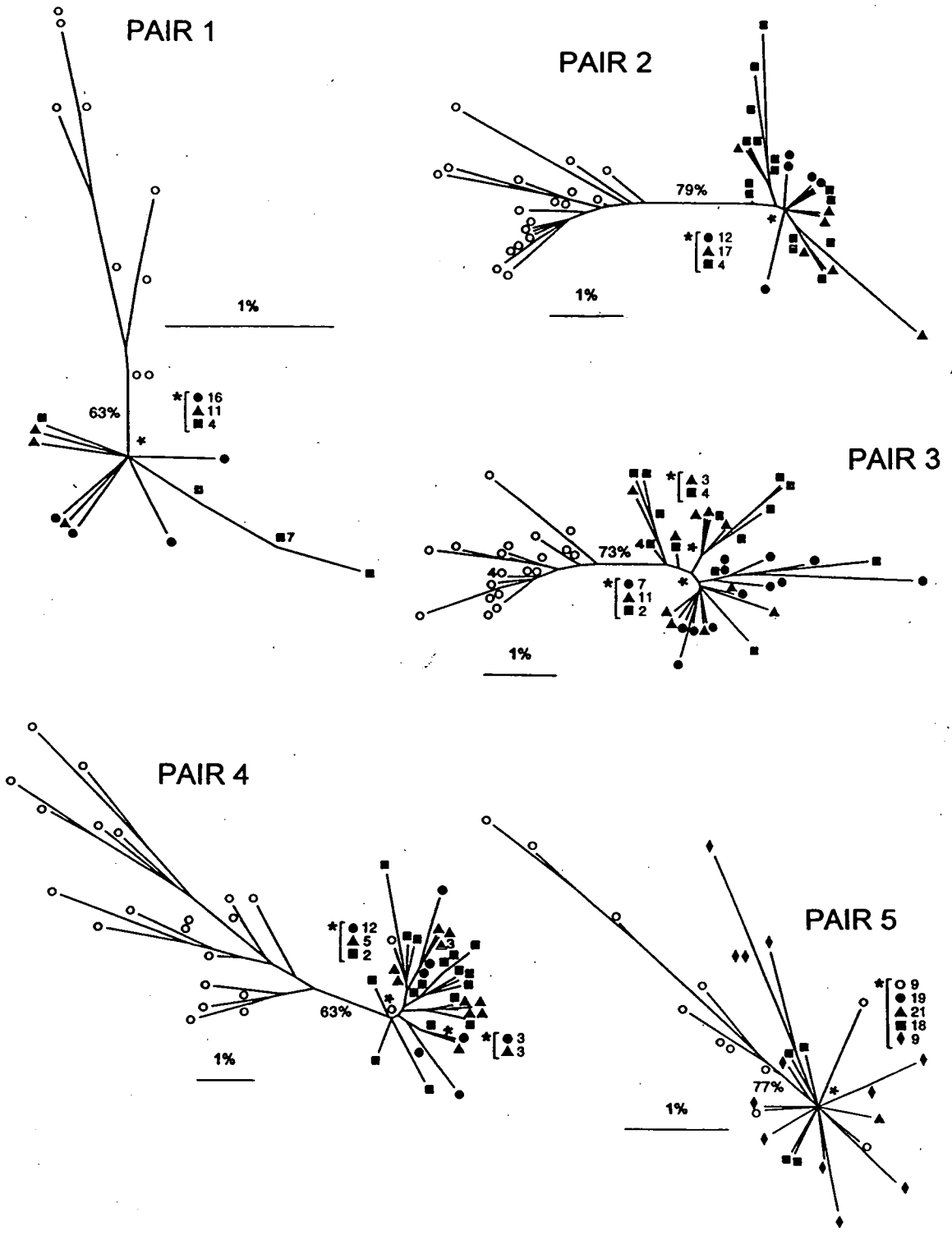
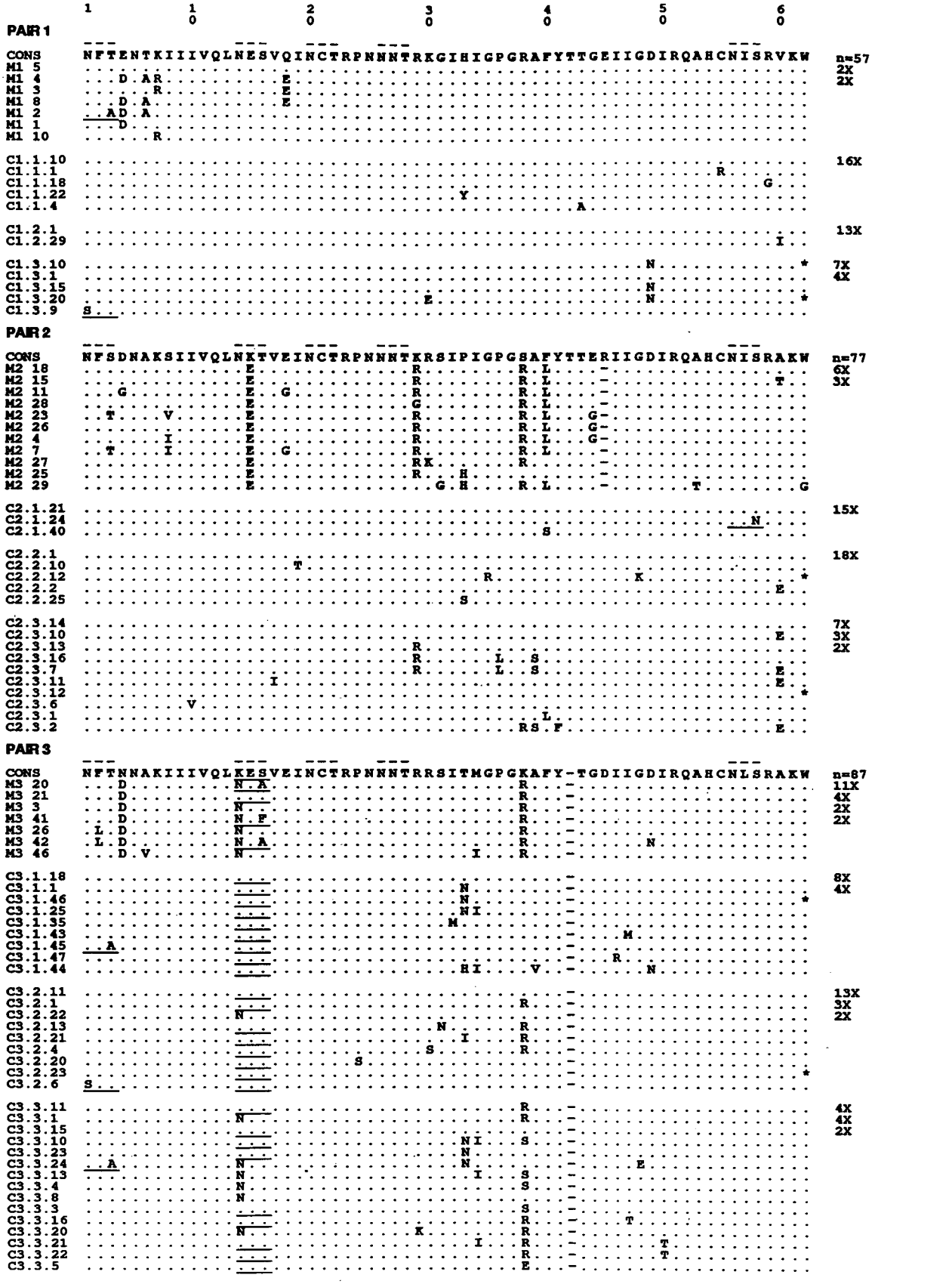

PAIR 1
1%
63%
16
11
4
7
PAIR 2
79%
12
17
4
1%
PAIR 3
73%
3
4
7
11
2
1%
PAIR 4
12
5
2
63%
3
3
1%
PAIR 5
9
19
21
18
9
77%
1%

Figure 4.

Multiple alignment of deduced amino acid sequences of the V3 region of 409 sequences derived from the 5 mother-infant pairs. Position 1 corresponds to amino acid 276 of the HXB2 envelope protein. Sequences are aligned against the consensus sequences derived for each mother-infant pair. The number of clones obtained from each pair is shown at the end of the consensus sequence. The frequency of clones with identical amino acid sequences for each sample is given at the end of the sequence. Dots indicate identity with the reference sequence, dashes represent gaps introduced to maximize alignment, and asterisks represent stop codons. Potential N-linked glycosylation sites are indicated by three dashes above the alignment. The loss of a potential N-linked glycosylation site is indicated by underlined letters.


PAIR 1
CONS NFTENTKIIIVQLNESVQINCTRPNNNTRKGIHIGPGRAFYTTGEIIGDIRQAHCNISRVKW n=57
M1 5 2X
M1 4 2X
M1 3
M1 8
M1 2
M1 1
M1 10
C1.1.10 16X
C1.1.1
C1.1.18
C1.1.22
C1.1.4
C1.2.1 13X
C1.2.29
C1.3.10 7X
C1.3.1 4X
C1.3.15
C1.3.20
C1.3.9
PAIR 2
CONS NFSDNAKSIIVQLNKTVEINCTRPNNNTKRSIPIGPGSAFYTTERIIGDIRQAHCNISRAKW n=77
M2 18 6X
M2 15 3X
M2 11
M2 28
M2 23
M2 26
M2 4
M2 7
M2 27
M2 25
M2 29
C2.1.21 15X
C2.1.24
C2.1.40
C2.2.1 18X
C2.2.10
C2.2.12
C2.2.2
C2.2.25
C2.3.14 7X
C2.3.10 3X
C2.3.13 2X
C2.3.16
C2.3.7
C2.3.11
C2.3.12
C2.3.6
C2.3.1
C2.3.2
PAIR 3
CONS NFTNNAKIIIVQLKESVEINCTRPNNNTRRSITMGPGKAFY-TGDIIGDIRQAHCNLSRAKW n=87
M3 20 11X
M3 21 4X
M3 3 2X
M3 41 2X
M3 26
M3 42
M3 46
C3.1.18 8X
C3.1.1 4X
C3.1.46
C3.1.25
C3.1.35
C3.1.43
C3.1.45
C3.1.47
C3.1.44
C3.2.11 13X
C3.2.1 3X
C3.2.22 2X
C3.2.13
C3.2.21
C3.2.4
C3.2.20
C3.2.23
C3.2.6
C3.3.11 4X
C3.3.1 4X
C3.3.15 2X
C3.3.10
C3.3.23
C3.3.24
C3.3.13
C3.3.4
C3.3.8
C3.3.3
C3.3.16
C3.3.20
C3.3.21
C3.3.22
C3.3.5

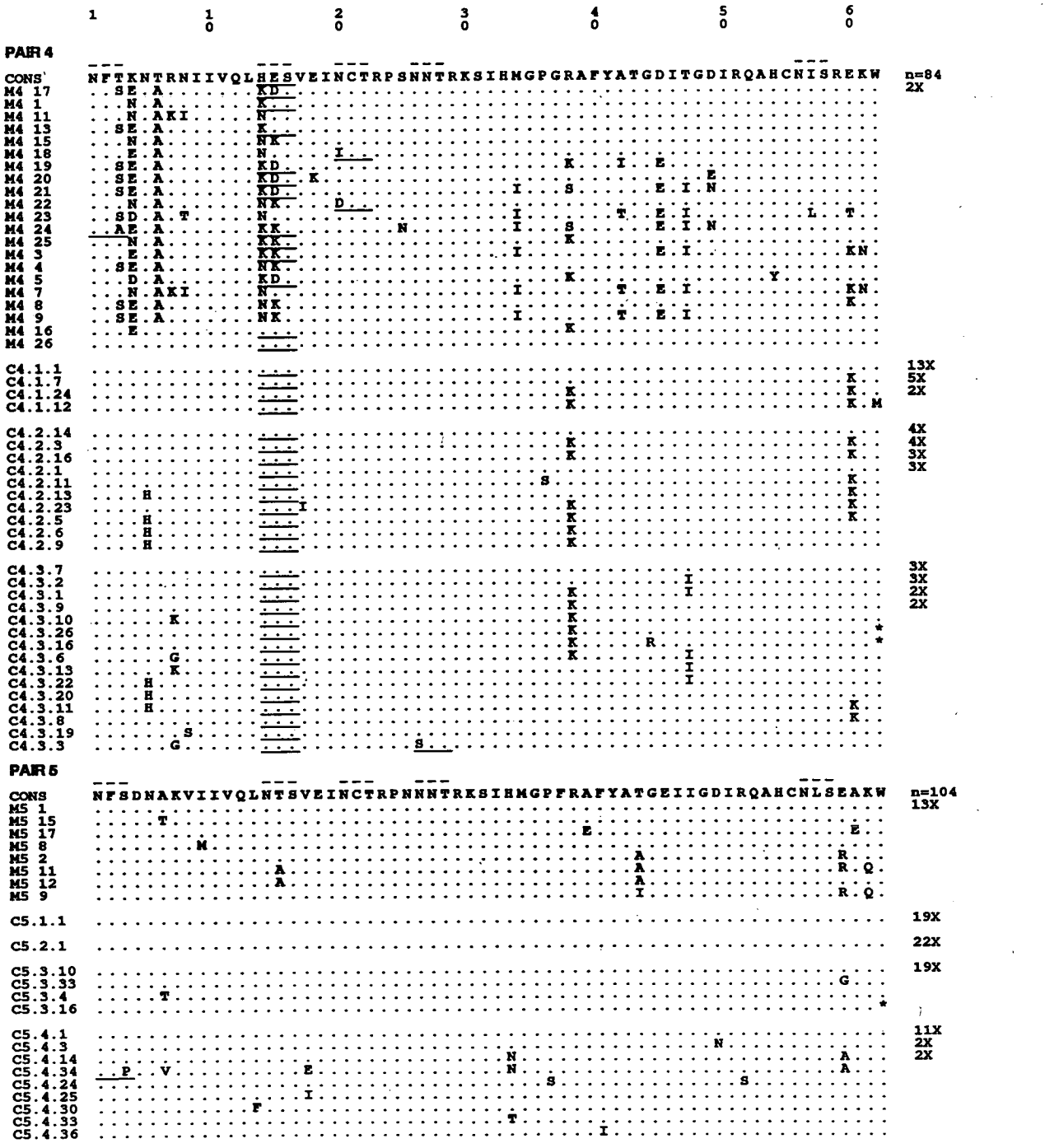
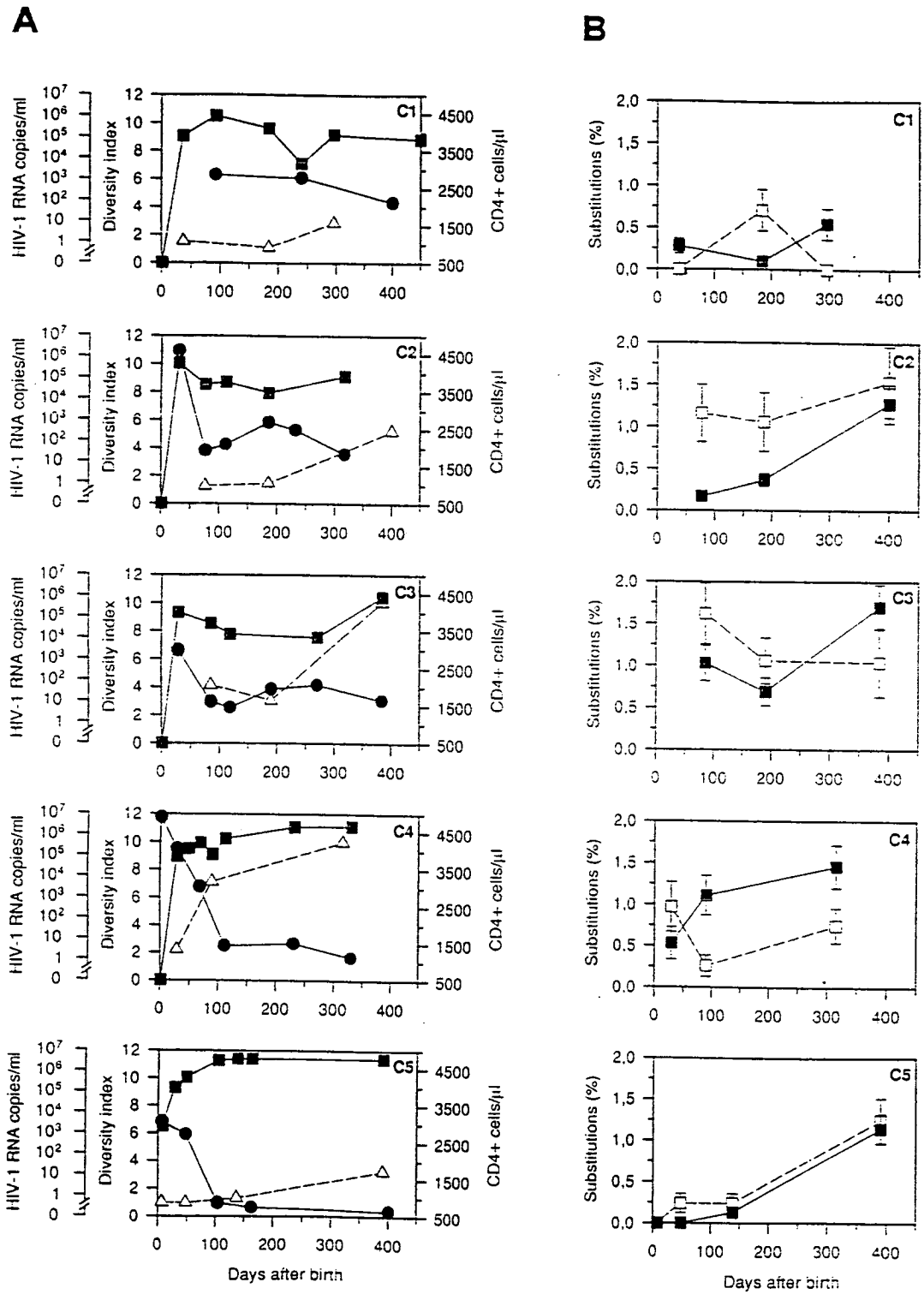

1
10
20
30
40
50
60
PAIR 4
CONS' NFTKNTRNIIVQLHESVEINCTRPSNNTRKSIHMGPGRAFYATGDITGDIRQAHCNISREKW n=84
M4 17 2X
M4 1
M4 11
M4 13
M4 15
M4 18
M4 19
M4 20
M4 21
M4 22
M4 23
M4 24
M4 25
M4 3
M4 4
M4 5
M4 7
M4 8
M4 9
M4 16
M4 26
C4.1.1 13X
C4.1.7 5X
C4.1.24 2X
C4.1.12
C4.2.14 4X
C4.2.3 4X
C4.2.16 3X
C4.2.1 3X
C4.2.11
C4.2.13
C4.2.23
C4.2.5
C4.2.6
C4.2.9
C4.3.7 3X
C4.3.2 3X
C4.3.1 2X
C4.3.9 2X
C4.3.10
C4.3.26
C4.3.16
C4.3.6
C4.3.13
C4.3.22
C4.3.20
C4.3.11
C4.3.8
C4.3.19
C4.3.3
PAIR 5
CONS NFSDNAKVIIVQLNTSVEINCTRPNNNTRKSIHMGPFRAFYATGEIIGDIRQAHCNLSEAKW n=104
M5 1 13X
M5 15
M5 17
M5 8
M5 2
M5 11
M5 12
M5 9
C5.1.1 19X
C5.2.1 22X
C5.3.10 19X
C5.3.33
C5.3.4
C5.3.16
C5.4.1 11X
C5.4.3 2X
C5.4.14 2X
C5.4.34
C5.4.24
C5.4.25
C5.4.30
C5.4.33
C5.4.36

Figure 5.

Molecular and genetic parameters in 3 slow progressor (C1, C2, and C3) and 2 rapid progressor (C4 and C5) HIV-1 infected infants. The patient code is shown within each plot. Panel A shows copy number of HIV-1 RNA/ml of plasma (■), CD4+ cell number (●) and intrasample viral diversity (Δ). Genetic diversity of the virus population was measured by the Simpson index (D), as detailed in Materials and Methods; the plot shows 1/D values for each sample at each time point. Panel B reports mean values (± standard errors) of synonymous (❑) and nonsynonymous (■) substitutions. The number of synonymous and nonsynonymous substitutions were calculated by using the method of Nei and Gojobori, incorporating the Jukes-Cantor correction for multiple substitutions, and expressed as percentages.


A
B
HIV-1 RNA copies/ml
Diversity index
CD4+ cells/µl
Substitutions (%)
C1
C2
C3
C4
C5
Days after birth

Paper IV

Analysis of Human Immunodeficiency Virus Type 1 *env* and *gag* Sequence Variants Derived from a Mother and Two Vertically Infected Children Provides Evidence for the Transmission of Multiple Sequence Variants

Christopher M. Wade[1*], Denis Lobidel[1], and Andrew J. Leigh Brown[1]

[1] Centre for HIV Research, Institute of Cell, Animal and Population Biology, Division of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3JN, United Kingdom.

[*] Corresponding Author

Submitted for Publication

ABSTRACT

In order to investigate the transmission of human immunodeficiency virus type 1 (HIV-1) from mother-to-child we have examined serial plasma RNA samples obtained from a mother over an eight year period spanning four pregnancies. Child 1 and 2 (born January 1987 and June 1990) were uninfected whilst child 3 and 4 (born July 1992 and February 1994) were HIV positive. Sequential plasma RNA and proviral DNA samples were obtained from the infected children. Genetic variation was examined within the viral population of the mother and her two infected children for both the V3 loop and flanking regions of the *env* gene and the p17 region of the *gag* gene. In one child (child 4) a highly homogeneous virus population was observed within both *env* and *gag* in contrast to the more heterogeneous virus population observed within the mother. Viral sequences of child 4 clustered within a single branch within the reconstructed phylogenetic tree. This is consistent with the transmission of a single maternal variant to the child in this case, which may indicate a selective process. By contrast, child 3 showed substantial genetic heterogeneity even within the first samples obtained shortly after birth. Sequences of child 3 clustered in two distinct groups within the phylogenetic tree and were separated by sequences of the mother. These results are not consistent with the selective transmission of a single maternal variant to the child in this case and we therefore propose that the infection within child 3 is the result of the transmission of multiple sequence variants to the child.

Running Head: Sequence variations in HIV-1 vertical transmission

Mother-to-child transmission of human immunodeficiency virus Type 1 (HIV-1) is estimated to occur at a rate of 13-32% of children born to HIV-1 infected mothers in industrialised countries and at a rate of 25-48% in developing countries (Dabis *et al.*, 1993). Mother-to-child transmission accounts for the majority of paediatric AIDS cases and more than 80% of perinatally infected children show symptoms of HIV infection by 18-24 months of age (Pizzo *et al.*, 1995). Transmission to the child may occur either *in utero* (prepartum) (Courgnaud *et al.*, 1991; Soeiro *et al.*, 1992), at delivery (intrapartum) (Ehrnst *et al.*, 1991; De Rossi *et al.*, 1992), or postnatally through breast feeding (postpartum) (Ziegler *et al.*, 1985; Lepage *et al.*, 1987). Evidence for *in utero* infection has been provided by the detection of viral nucleic acids in fetal tissues (Joviasis *et al.*, 1985; Sprecher *et al.*, 1986; Courgnaud *et al.*, 1991; Soeiro *et al.*, 1992) indicating that transmission to the child may occur at an early stage of gestation. Clinical studies have interpreted the early detection of HIV infection in newborns (within the first few days following birth) as indicative of transmission occurring early in pregnancy. Nevertheless, markers for HIV infection are absent at birth in a high proportion of perinatally infected children (Borkowsky *et al.*, 1992; Burgard *et al.*, 1992; Krivine *et al.*, 1992) and such clinical information has been interpreted as evidence for transmission either late in pregnancy or at the time of delivery in these cases (Ehrnst *et al.*, 1991; De Rossi *et al.*, 1992). Epidemiological data has indeed shown increased risks of transmission associated with vaginal delivery, with a significantly lower rate of perinatal transmission observed in children delivered by caesarean section (European Collaborative Study, 1994). Studies of vertical transmission in twins (Goedert *et al.*, 1991) also provide evidence for transmission at delivery and it has been proposed that the higher rate of transmission observed in first born twins is due to the exposure of the first twin to infectious material within the birth canal for a greater time than for the subsequent twin (Goedert *et al.*, 1991). Thus these data would suggest the occurrence of transmission as the child encounters the cervix and birth canal and indeed virus is present within the genital secretions of women with antibodies to HIV-1 (Wofsy *et al.*, 1986).

Vertical transmission of HIV-1 appears to be dependent upon the mothers immunological and virological status. Recent infection (Rossi, 1992), high viral load (Rossi, 1992), p24 antigenaemia (European Collaborative Study, 1992), low $CD4^+$ T cell counts (European Collaborative Study, 1992), low anti gp120 antibodies in the mothers serum

(Goedert *et al.*, 1989), absence of antibodies against specific domains of gp120 (Rossi, 1992) and gp41 (Ugen *et al.*, 1992), and advanced clinical stage of the mother (Rossi, 1992) appear to be associated with an increased risk of vertical transmission. Viral factors may also play a role in influencing vertical transmission.

Molecular analyses of HIV-1 sequence variants following sexual or parenteral transmission in adults have revealed that a highly homogeneous sequence population is observed within *env* in the recipient immediately following transmission (McNearney *et al.*, 1990; Wolfs *et al.*, 1992; Zhang *et al.*, 1993; Zhu *et al.*, 1993). This is in contrast to the heterogeneous sequence population typically observed within *env* in the long term infected patient. It has therefore been postulated that a sequence bottleneck occurs upon infection, with infection initiated by a limited number of variants or even one particular variant. p17 *gag* gene sequences were not found to show such a restricted level of sequence diversity within the recipient upon seroconversion (Zhang *et al.*, 1993; Zhu *et al.*, 1993). This suggests that the homogeneity observed in *env* is due to strong selection for specific *env* sequences either upon transmission or in the interval between exposure and seroconversion (Zhang *et al.*, 1993).

Molecular studies examining mother-to-child transmission of HIV-1 have suggested that infection within the child is initiated by a single maternal variant (Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995), a situation analogous to that observed for sexual or parenteral transmission in adults. A highly homogeneous sequence population is typically observed within the child which is considerably less diverse than that of the mother. The transmitted variant has been reported to represent a minor maternal form in the majority of mother-child transmission cases studied (Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995) and this has been interpreted as evidence that selection is playing a role in vertical transmission. A number of studies have however reported that the transmitted variant can represent either a major or a minor maternal form (Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Scarlatti *et al.*, 1993). The transmission of multiple HIV-1 genotypes from mother-to-child has nevertheless also been reported for a number of vertical transmission cases (Lamers *et al.*, 1994; Van't Wout *et al.*, 1994; Briant *et al.*, 1995). Multiple HIV-1 genotypes have also been observed to be transmitted to twins during a single pregnancy (Weiser *et al.*, 1993), with both molecular and clinical evidence provided for transmission occurring *in utero* for one

twin and at delivery for the other.

In order to examine the transmission of HIV-1 from mother-to-child we have examined serial plasma RNA samples obtained from a mother across an eight year period spanning four pregnancies (child 1 and 2 uninfected, child 3 and 4 infected) and compared the viral population of the mother with plasma RNA and proviral DNA sequences obtained from her two infected children. Evidence is provided that different transmission processes were in operation for the infection of the two infected children. One child was infected by a single maternal variant which may indicate a selective process. Sequence data for the other child, however, suggested transmission of multiple sequence variants to the child.

MATERIALS AND METHODS

Patients

Sequential plasma and peripheral blood mononuclear cell (PBMC) samples were obtained over a period of 9 years (1986-1994) from a transmission set comprising a mother and her two HIV-1 infected children. During the studied time period, the mother gave birth to four children of whom child 1 (born 1.87) and child 2 (born 1.6.90) were uninfected whilst child 3 (born 31.7.92) and child 4 (born 26.2.94) were infected. The mother first tested HIV positive on 17.10.85 with the seroconversion date estimated at May 1984. Her risk factor for HIV infection was intravenous drug use.

The times of sampling, clinical status and $CD4^+$ cell counts of the patients within the mother-child transmission set are presented in Table 1.

PCR amplification and sequencing.

EDTA treated and heparinized whole blood samples were separated on Ficoll-Hypaque (Pharmacia) and stored in liquid nitrogen (PBMCs) or at -20°C (plasma). DNA and RNA extraction, PCR amplification and automated DNA sequencing were performed essentially as described previously (Simmonds *et al.*, 1990; Zhang *et al.*, 1991; Leigh Brown and Simmonds, 1995).

An approximately 436 base pair (bp) fragment spanning the V3 loop and flanking regions of the *env* gene (positions 7029 to 7464 in the HIV-HXB2 genome (Genbank Accession Number: K03445) and an approximately 390 bp fragment of the p17 coding region of the *gag* gene (positions 857 to 1246 in the HIV-HXB2 genome) were amplified by limiting dilution nested PCR (Simmonds *et al.*, 1990), with 30 cycles in both the first and second rounds of amplification. The following primers were used:

env

1 (outer, +): 5'-TACAATGTACACATGGAATT-3' (nucleotides 6957-6976 HXB2 sequence)

2 (outer, -): 5'-GGAGGGGCATACATTGC-3' (7520-7537 HXB2 sequence)

3 (inner, +): 5'-TGGCAGTCTAGCAGAAGAAG-3' (7009-7028 HXB2 sequence)

4 (inner, -): 5'-ATTCTGCATGGGAGTGTG-3' (7465-7482 HIV-HXB2 sequence)

gag

1 (outer, +): 5'-GCGAGAGCGTCAGTATTAAGCGG-3' (795-817 HXB2 sequence)

2 (outer, -): 5'-TCTGATAATGCTGAAAACATGGG-3' (1296-1318 HXB2 sequence)

3 (inner, +): 5'-GGGAAAAAATTCGGTTAAGGCC-3' (833-856 HXB2 genome)

4 (inner, -): 5'-CTTCTACTACTTTTACCCATGC-3' (1247-1270 HXB2 sequence)

Both sense and antisense strands of the *env* and *gag* amplification products were sequenced using a direct solid phase automated sequencing approach using the T7 dye terminator sequencing chemistry (detailed in Leigh Brown and Simmonds, 1995). Sequencing products were run on an Applied Biosystems 373A automatic DNA sequencer.

Sequence analysis

Raw nucleotide sequences were assembled using the STADEN package (Staden, 1993). Sequences were then aligned using the CLUSTAL V algorithm (Higgins *et al.*, 1992), as implemented in version 2.2 of the Genetic Data Environment (GDE) package (Smith *et al.*, 1994). The final alignment was improved manually by preferring gaps to transition differences, transition differences to transversion differences and by the insertion of gaps to maintain the reading frame. Phylogenetic and distance analyses were performed using programs taken from version 3.52c of the Phylogeny Inference Package (PHYLIP; Felsenstein, 1993). Nucleotide distances were estimated using the generalised two-parameter (maximum likelihood) model (Kishino and Hasegawa, 1989) (program DNADIST) and phylogenetic trees were reconstructed using the neighbour-joining method (Saitou and Nei, 1987) (program NEIGHBOR). One hundred bootstrap replicates (Felsenstein 1985) (programs SEQBOOT and CONSENSE) were performed for each tree. Alternative phylogenetic hypotheses were evaluated statistically by a likelihood ratio test (Kishino and Hasegawa 1989). The number of synonymous substitutions per synonymous site (d_s) and the number of nonsynonymous substitutions per nonsynonymous site (d_n) were calculated using the Jukes Cantor one-parameter model (Jukes and Cantor, 1969) as implemented in the Molecular Evolutionary Genetics Analysis program version 1.01 (MEGA; Kumar *et al.*, 1993).

The sequence dataset was screened for the presence of potential contaminants by comparing the patient sequences with sequences of equivalent regions of all clones and other patients examined within the laboratory.

Nucleotide sequence accession numbers

Nucleotide sequences reported in this study have been assigned the GenBank accession numbers XXXXXX to YYYYYY.

We have obtained 124 sequences spanning the V3 loop and flanking regions of the *env* gene and 198 sequences of the p17 region of the *gag* gene from an infected mother and her two HIV-1 infected children. Nineteen plasma samples were obtained from the mother over a period of 8 years (1986-1993) although plasma RNA viral sequences were only obtained from a limited number of time points (Table 1). During this period the mother gave birth to four children. Two children, child 1 and 2, were uninfected whilst child 3 and 4 were infected vertically. Sequences were obtained from both plasma RNA and proviral DNA samples from the two infected children. Child 3 was followed for a period of two years following birth with child 4 followed for a five month period (Table 1).

Sequence variability in the mother and her children

The intra patient genetic diversities (Table 2, on diagonal), inter patient genetic distances (Table 2) and intra sample genetic diversities (Table 3) were calculated for each pairwise comparison between sequences of the mother-child transmission set. Relatively high intra patient sequence diversity was observed in child 3 within the first two years of life in both *env* (2.8%) and *gag* (1.2%) (Table 2). Diversity was relatively high even within the earliest samples obtained. The two sequences obtained in *env* from RNA sample 2R (approximately 2 months) differed at 1% of their nucleotide sites with a similar diversity observed in *gag* on a substantial sequence sample (n=11) (Table 3). Within the first proviral DNA sample 1D (approximately 1 month), 2.2% (n=9) diversity was observed in *env* with 1.1% (n=10) diversity observed in *gag* (Table 3). In contrast, child 4 showed much less variability with an overall diversity of 0.6% observed in *env* and 0.3% observed in *gag* (Table 2) from all samples. The earliest sample obtained for child 4, sample 2 (approximately 1 month), showed levels of sequence diversity of approximately 0.1% in both *env* and *gag* within both the plasma RNA and proviral DNA populations (Table 3). Sequences from the infected mother showed diversity levels of 1.7% (n=13) in *env* and 2.3% (n=36) in *gag* (Table 2). The lower levels of intrapatient diversity observed within the mother in *env* than in *gag* reflect the limited number of *env* sequences obtained from this patient.

The number of synonymous substitutions per synonymous site (d_s) and the number of nonsynonymous substitutions per nonsynonymous site (d_n) are presented for each sampling time point in Table 3. Synonymous and nonsynonymous distances were relatively high for

all time points in both *env* and *gag* in child 3 in contrast with low synonymous and nonsynonymous distances in both *env* and *gag* in child 4. Within child 3, the synonymous distances were typically greater than the nonsynonymous distances in both the p17 region of *gag* and within the V3 loop and flanking regions of the *env* gene. This resulted in d_s/d_n ratios typically greater than 1 for all time points in both *env* and *gag*, although d_s/d_n ratios were considerably greater in *gag* than *env*. Within child 4 d_s/d_n ratios were typically less than 1 in *env* but greater than 1 in *gag*.

Evolutionary relationships between the mother and her two HIV-1 infected children

Neighbour-joining phylogenetic trees inferred from analyses of all *env* and *gag* sequences derived from the mother and her infected children are presented in Figure 1. Sequences of the two children were observed to cluster distinctly from one another within both the *env* and *gag* trees (Figure 1a and 1b) with specific sequences of the mother showing an association with sequences derived from the two children. When sequences of the two children were considered independently of those of the mother, 84% (*env*) and 86% (*gag*) bootstrap support was provided for the branch separating the two children. Likelihood tests of alternative phylogenetic hypotheses in which sequences of one child were clustered with sequences of the other child confirmed the significance of the separation of sequences from the two children for both the *env* ($p<0.02$) and *gag* ($p<0.02$) analyses.

a) Child 3

The high levels of sequence diversity observed within child 3 immediately following birth were reflected by high degrees of sequence divergence within the inferred evolutionary trees. In *env*, two clearly distinct viral sequence lineages were identified within child 3 (Figure 1a, groups A and B). The distinction of these two groups was supported in 94% of bootstrap replicates when sequences of child 3 were considered on their own. Bootstrap support within the overall tree was reduced by the placement of maternal sequences on the branch to child 3 group A (Figure 1a, group A). Furthermore, likelihood tests of alternative phylogenetic hypotheses in which sequences of one child group were clustered with those of the other confirmed the significance of the two groups within the *env* tree ($p<0.03$).

Child 3 group A formed the minor group of sequence variants circulating within child 3, with a single sequence obtained from the viral RNA population of sample 3R (approximately 4 months following birth) and 12 sequences obtained from the proviral DNA population in samples 1D (5), 4D (6) and 5D (1) (approximately 1 month, 5 months and 11

months after birth respectively). The child 3 group A lineage did not persist within the child subsequent to these time points and was not observed in the most recent RNA (8R) or DNA (7D) samples in *env* which were taken approximately 2 years after birth. As noted above, child 3 group A variants appeared to persist within the proviral DNA population marginally longer than in the plasma RNA population. Child 3 group A sequences were observed to cluster, albeit loosely, with maternal sequences obtained from samples taken across pregnancy (samples 14R and 12R taken approximately 2 and 5 months respectively before the birth of child 3). Nonetheless, bootstrap support for this association was low (15%) and likelihood tests revealed that the maternal group could be associated alternatively with either child 3 lineage B or with child 4.

The major child group (Figure 1a, group B) showed a very strong association with a number of maternal sequences identified in RNA samples 18R and 19R which were obtained just over 1 year following the birth of child 3. Indeed, the maternal sequences were identical at the nucleotide level to a number of child 3 sequences identified in RNA samples 3R (approximately 4 months) and 7R (approximately 22 months) and DNA sample 5D (approximately 11 months) of the child. The mother-child 3 group B cluster was supported in 64% of bootstrap replicates in the overall tree although when all sequences of child 3 (groups A and B) were considered with the group B maternal sequences alone the cluster was supported in 94% of bootstrap replicates (also observed when child 3 was considered independently). Likelihood tests of alternative phylogenetic hypotheses in which the maternal sequences associated with child 3 lineage B were clustered alternatively with lineage A confirmed the significance of the association of these maternal sequences with child 3 group B ($p<0.01$). The poor storage conditions of plasma samples obtained from the mother made the generation of viral RNA sequences from maternal samples difficult. However, in the 4 *env* sequences that were obtained from samples taken across pregnancy (samples 14R and 12R, taken approximately 2 and 5 months respectively before the birth of child 3) the group B lineage was not identified. Child 3 lineage B persisted until the most recent *env* RNA (8R) and DNA (7D) samples obtained approximately 2 years following birth. During this period, viral sequences of the major child 3 lineage showed a clear increase in genetic distance from the transmitted maternal group B variant with time.

Two lineages were also observed within the reconstructed *gag* gene phylogeny for child 3 (Figure 1b, groups A and B). The branch to the main child 3 lineage A sequences (1 1R sequence, 4 1D sequences and 1 2R sequence) was supported in 88% of bootstrap

replicates when sequences of child 3 were considered on their own. The association of the individual 2R and 5D sequences with child 3 lineage A observed in the overall phylogeny was not however supported in bootstrap analyses. A greater number of sequences were obtained from plasma samples of the mother in *gag* than within *env* and thus consequently the relationship between maternal sequences and sequences of the child becomes somewhat clearer. As in *env*, child 3 group A (Figure 1b) consisted of sequences circulating within the plasma RNA and proviral DNA populations of the child earlier in the infection and did not persist until the most recent samples (8R and 8D, approximately 2 years following birth). The group consisted of 4 RNA sequences from time points 1R and 2R (approximately 1 month and 2 months after birth respectively) and 5 DNA sequences from samples 1D and 5D (approximately 1 month and 11 months after birth respectively). Child 3 group A sequences persisted within the proviral DNA population longer than within the plasma RNA population and clustered with a number of maternal sequences from all sequenced maternal time points (samples 5R, 12R, 13R, 15R taken approximately 2.5 years, 5 months, 3 months, 2 months, 2 weeks, respectively before the birth of child 3 and samples 18R and 19R taken just over 1 year after the birth of child 3). Group A formed the major variant present within the mother, particularly across the child 3 pregnancy.

The major group of child 3 sequences (Figure 1b, group B) contained representatives from all sequenced time points in *gag* (RNA and DNA samples 1 (approximately 1 month following birth) through 8 (approximately 2 years following birth)). A number of maternal sequences obtained predominantly from samples taken a year after pregnancy (18R and 19R) clustered within the child group although the group B sequence population was represented in the mother in sample 13R (approximately 3 months before birth). Child 3 group B sequences in *gag* did not show the clear increase in diversity observed in *env* over the 2 year study period in the child.

b) Child 4

Child 4 sequences, obtained from 5 samples taken between 1 month and 6 months of age, clustered tightly within both the inferred *env* and *gag* phylogenies (Figure 1a and 1b) as indicated by the intra patient and intra sample genetic diversities. In both the *env* and *gag* phylogenies, sequences of child 4 were associated with maternal sequences from time points 18R and 19R, approximately 6 and 2 months respectively before the birth of child 4 (by comparison approximately 1 year after the birth of child 3). In *env*, the maternal sequences associated with child 4 clustered outside the child group. In *gag*, the child sequences fell on

a maternal lineage consisting of sequences from time points 18R and 19R and 3 *gag* sequences of time point 18R were identical at the nucleotide level to the most common variant observed within the child.

c) Mother

Maternal sequences thus appear to fall within three groups within both the *env* and *gag* phylogenies (Figure 1a and 1b). Early sequences of the mother, including sequences obtained from samples taken during the child 3 pregnancy, clustered with a minor group of sequence variants within child 3 (child 3 lineage A, Figure 1). Later sequences of the mother, obtained a year after the birth of child 3 and during the child 4 pregnancy fell into two groups. One group was associated with the major group of sequence variants within child 3 (child 3 group B, Figure 1). The final group of maternal sequences was associated with viral sequences of child 4 (child 4, Figure 1).

Amino acid sequence heterogeneity between mothers and children

Deduced *env* and *gag* amino acid sequences for the mother-child transmission set are presented in Figure 2.

a) *Env* V3 loop and flanking regions

Three distinct amino acid sequence variants were identified within the deduced *env* amino acid sequences (Figure 2a) which corresponded to the three lineages (child 3 lineages A and B, and the child 4 lineage) observed within the phylogenetic tree (Figure 1a).

Child 3 lineage A could be differentiated from lineage B by the presence of specific amino acids at positions 66 and 67. An alanine (A) was observed at both these positions in child 3 lineage A with a glutamic acid (E) at position 66 and typically a glutamine (Q) at position 67 in child 3 lineage B (histidine (H) was observed at position 67 in a single child 3 lineage B sequence). Further, an alanine (A) and a lysine (K) were commonly observed within child 3 lineage A at positions 117 and 118 although these substitutions were not observed in all lineage A sequences. Arginine (R) was also fairly commonly observed within child 3 lineage B at position 84 but uncommon in child 3 lineage A (observed in 1 of 9 sequences). Within the later child 3 lineage B samples 7D, 7R and 8R (approximately 2 years post infection) amino acid substitutions were commonly observed at positions 39 (asparagine, N), 40 (methionine, M) and 91 (lysine, K).

Child 4 was characterised by amino acid substitutions at positions 39 (proline, P), 73 (lysine, K) and 117 (alanine, A) which were observed in all child 4 sequences. These amino

acids were fairly uncommon within the deduced amino acid sequences of child 3 and in no cases were all 3 amino acids observed within a single child 3 sequence. Threonine (T) at position 59 and aspartic acid (D) at position 121 were also commonly observed within the amino acid sequences of the earlier (1 month to 4 month) child 4 samples 2D, 2R and 3D.

Within the maternal *env* amino acid sequence population the earliest sequences obtained were closest to those observed within child 3 lineage A. Alanine (A) at position 66, characteristic of the child 3 lineage A sequence population, was observed within maternal sequences 12Ra and 14Ra, although the alanine (A) at position 67 observed in all child 3 lineage A sequences was not observed within these maternal sequences. The child 3 lineage B amino acid sequence was clearly identified in amino acid sequences 18Ra and 19Ra. These maternal amino acid sequence variants were identical to sequences 3Ra (2 sequences), 4Dc, 5Dd and 7Rg observed within child 3. Amino acid sequence patterns characteristic of child 4 were clearly identified in maternal sequences 14Rb, 18Rb and 19Rb. The maternal sequences did however contain an asparagine (N) at position 6 and a methionine (M) at position 9 which were not observed in the amino acid sequences of child 4. The GPGRAF motif at the crown of the V3 loop was conserved in all mother and child sequences.

Examination of putative N-linked glycosylation sites between the mother and her children revealed no pattern of selective loss of N-linked glycosylation sites between the mother and child 3 with the 10 glycosylation sites observed within the mother conserved between mother and child. The N-linked glycosylation site at position 91 was however generally absent from the later child 3 samples (7D, 7R and 8R) which were obtained approximately 2 years following birth. Child 4 showed the loss of the glycosylation site at position 115 in all sequences obtained from this child. This site was also absent in the maternal amino acid sequences of the variant associated with child 4. The N-linked glycosylation site at position 121 was also absent within the earlier (1 month to 4 month) child 4 samples 2D, 2R and 3D. Interestingly, the potential N-linked glycosylation site at position 7 was absent in the maternal sequences which were associated with child 4 but this site was present within all sequences obtained from the child. The potential N-linked glycosylation site proximal to the first cysteine of the V3 loop (position 26) which was absent from all infants' sequence sets described by Wolinsky *et al.* (1992) remained conserved between the mother and children.

The potential phenotype of the amino acid sequence variants was predicted on the basis of the global net charge of the V3 loop and the degree of sequence divergence from the

LaRosa subtype B consensus (Milich *et al.*, 1993; Donaldson *et al.*, 1994). The majority of sequences from the mother and her children were predicted to be of the macrophage-tropic, non-syncytium-inducing (NSI) phenotype. However, a small number of sequences obtained from child 3 between 15 months and 2 years (6D, 7D and 7R) were predicted to be of the T-cell-tropic, syncytium-inducing (SI) phenotype. These variants were observed to show a positively charged lysine (K) at position 51 as opposed to a negatively charged aspartic acid (D) or glutamic acid (E) which are more commonly observed at this position.

b) p17 *gag* region

The two lineages observed within child 3 were less clearly discernible within the deduced amino acid sequences of the p17 region of *gag* than within *env*. In general, a threonine (T) was observed at position 107 in child 3 lineage B with either an alanine (A) or a serine (S) observed at this position in lineage A. There were however exceptions to this with the child 3 lineage B 1Dd sequence showing an alanine at this position.

Child 3 amino acid sequences could however be clearly differentiated from those of child 4 at positions 58 (threonine (T) child 3, alanine (A) child 4), 100 (glycine (G) child 3, glutamic acid (E) child 4) and 102 (serine (S) child 3, asparagine (N) child 4). Position 107 was characterised typically by an alanine (A) or serine (S) in child 3 lineage A, an alanine (A) in child 4 and by contrast a threonine (T) at this position in child 3 lineage B.

Genetic Relationships with other Edinburgh HIV-1 sequences

A phylogenetic analysis of sequences from the mother-child transmission set with representative sequences of identified HIV-1 risk groups circulating in Edinburgh (Holmes *et al.*, 1995; Leigh Brown *et al.*, submitted), revealed that the patients fell within the intravenous drug user clade. This is consistent with the available clinical information for the mother. An analysis of p17 *gag* gene sequences from the mother and her two children with all available Edinburgh intravenous drug user sequences and B subtype reference isolates (Figure 3) revealed that all sequences from the mother and her children fell within a single group within the reconstructed phylogenetic tree and were not separated by sequences obtained from any other patient. Sequences were not available from these Edinburgh intravenous drug user patients for comparison within *env*.

We have analysed sequences spanning the V3 loop and flanking regions of the *env* gene and the p17 region of the *gag* gene from sequential plasma RNA populations obtained from a mother, and compared the viral characteristics of the maternal samples with both plasma RNA and proviral DNA populations obtained from her two vertically infected children.

Sequence variability and implications for transmission

a) Evidence for the transmission of a single maternal variant

Substantial differences were observed in the level of genetic heterogeneity within the viral populations of the two infected children. Child 4 showed a highly homogeneous sequence population within both the sequenced *env* and *gag* regions, with sequences of the child clustering tightly within a single group in the reconstructed phylogenetic trees. Such viral homogeneity upon transmission is consistent with the infection of child 4 being initiated by a single maternal variant. Similar levels of sequence diversity have been reported in the newly infected child immediately following birth in the majority of mother-child transmission cases studied (Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995), a situation analogous to the restricted levels of sequence diversity observed in recently infected adults upon seroconversion (McNearney *et al.*, 1990; Wolfs *et al.*, 1992; Zhang *et al.*, 1993; Zhu *et al.*, 1993). This has been interpreted as evidence for selection for particular viral variants from the heterogeneous pool present within the long term infected transmitter. Although on their own such observations do not distinguish selection from a founder effect, evidence that mother-child transmission can be a selective process is provided by the fact that the transmitted variant has been observed to represent a minor form within the maternal sequence population upon transmission in the majority of studied mother-child transmission cases (Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Mulder-Kampinga *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995) although the non-selective transmission of major maternal forms has also been reported in some cases (Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Scarlatti *et al.*, 1993). Based on the extremely limited number of sequences available, the maternal variant transmitted to child 4 appeared to be present in approximately 50% of the maternal sequence population during the child 4 pregnancy. The limited number of

maternal sequences obtained do not allow us to make a reliable estimate of the timing of the transmission event on the basis of sequence data alone. Virus was not however detected in child 4 until 1 month following birth (a sample taken at 6 days was negative), which is perhaps more consistent with infection occurring either late in pregnancy or at delivery (Ehrnst *et al.*, 1991; De Rossi *et al.*, 1992).

b) Evidence for the transmission of multiple sequence variants to the child

In contrast to child 4, substantial genetic heterogeneity was detected in child 3 within the first samples obtained following birth. The level of genetic diversity observed is considerably greater than diversity levels typically reported in patients, including vertically infected children, immediately following infection (McNearney *et al.*, 1990; Wike *et al.*, 1992; Wolfs *et al.*, 1992; Wolinsky *et al.*, 1992; Mulder-Kampinga *et al.*, 1993; Scarlatti *et al.*, 1993; Zhang *et al.*, 1993; Zhu *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995). Phylogenetic analysis revealed the existence of two groups of sequence variants within child 3 which were separated by sequences obtained from the mother. The main child group (group B) showed a clear association with maternal sequences in both *env* and *gag*, with a number of maternal sequence variants obtained in *env* identical to sequences obtained from the child. Again maternal sequences were associated with the minor child 3 group (group A) in both *env* and *gag*, although the association was not as clearly defined within *env*. The presence of both these groups within the first samples of the child, and their association with specific maternal sequence variants is indicative of the transmission of two distinct maternal variants to the child. Furthermore, the association of the two child lineages with predominantly early (across pregnancy, group A) or late (after pregnancy, group B) maternal variants is perhaps suggestive of transmission occurring at different times during pregnancy. Child 3 lineage A could possibly be the result of an early *in utero* transmission event with lineage B perhaps the result of transmission at delivery. The limited number of maternal sequences obtained for the transmission set do not however permit reliable inferences to be made regarding the time of transmission and it is possible that both lineages were present simultaneously, albeit with one lineage possibly present as a minor variant. Thus the results of the phylogenetic analysis for child 3 require either early infection with multiple variants or transmission of virus on more than one occasion, possibly during pregnancy and at delivery. The transmission of more than one variant appears to be rare, although transmission of multiple variants has been reported in a small number of cases for both adults (one patient

of the florida dentist case (Ou *et al.*, 1992, Korber and Myers, 1992); the victim in the Swedish rape case (Albert *et al.*, 1993); evidence for coinfection with multiple HIV-1 strains within an Australian homosexual male (Zhu *et al.*, 1995); and with heterogeneous virus populations also reported shortly after seroconversion within cervical secretions and/or peripheral blood in five women of a cohort of six Kenyan female sex workers (Poss *et al.*, 1995)) and some vertically infected children (Lamers *et al.*, 1994; Van't Wout *et al.*, 1994; Briant *et al.*, 1995).

The processes and opportunities for mother-to-child transmission of HIV-1 are undoubtedly very different from those operating in sexual or parenteral transmission. The association of mother and child for the nine month period of pregnancy would potentially provide a greater number of opportunities for transmission and thus would theoretically increase the possibility of the transmission of multiple variants to the child. The size of the inoculum has been shown to influence the likelihood of infection with multiple sequence variants, with a study of donor-recipient pairs indicating an increased risk of infection with multiple variants with a greater size inoculum (Van't Wout *et al.*, 1994). Mother-child transmission of HIV-1 may in effect increase the inoculum size either by allowing the easier transfer of infected fluids from the mother to the child or by providing more opportunities for the exchange of infected materials. Transmission to the child is known to occur both *in utero* and at delivery and there is no reason why viral transmission may not occur at both these times. In addition, if *in utero* transmission occurs as a result of placental damage then such damage may increase the likelihood of transmission on more than one occasion. It is possible that many unseen multiple transmissions occur between mother and child. The possibility that a second variant will become established may then be related to the relative fitnesses of the two variants. Only if two or more distinct viral variants present within the mother across pregnancy are transmitted to the child will evidence for multiple transmission be detected.

In conclusion, our analysis has provided evidence that mother-to-child transmission of HIV-1 is a complex process with different transmission processes occurring even within children born to a single mother. The patients studied have provided evidence for the transmission of a single maternal variant in the case of child 4, with the infection of child 3 apparently initiated by two distinct maternal variants.

ACKNOWLEDGEMENTS

We would like to thank Dr. J. Mok (Regional Infectious Diseases Unit, City Hospital, Edinburgh) and Dr. F. Johnstone (Department of Obstetrics and Gynaecology, Edinburgh Royal Infirmary, Edinburgh) for the provision of samples and clinical information. We would also like to thank Dr. David Yirrell and Marian Aldhous for identifying patients and samples. This work was supported by the Medical Research Council AIDS Directed Programme.

REFERENCES

Ahmad, N., B. M. Baroudy, R. C. Baker, and C. Chappey. 1995. Genetic analysis of human immunodeficiency virus type 1 V3 region isolates from mothers and infants after perinatal transmission. J. Virol. 69:1001-1012.

Albert, J., J. Wahlberg, and M. Uhlén. 1993. Forensic evidence by DNA sequencing. Science 361:595-596.

Borkowsky, W., K. Krasinski, H. Pollack, W. Hoover, A. Kaul and T. Ilmet-Moore. 1992. Early diagnosis of human immunodeficiency virus infection in children <6 months of age: comparison of polymerase chain reaction, culture, and plasma antigen capture techniques. J. Infect. Dis. 166:616-619.

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.

Burgard, M., M.-J. Mayaux, S. Blanche, A. Ferroni, M.-L. Guihard-Moscato, M.-C. Allemon, N. Ciraru-Vigneron, G. Firtion, C. Floch, F. Guilht, E. Lachassine, M. Vial, C. Griscelli, and C. Rouzioux. 1992. The use of viral culture and p24 antigen testing to diagnose human immunodeficiency virus infection in neonates. N. Engl. J. Med. 327: 1192-1197.

Courgnaud, V., F. Lauré, A. Brossard, C. Bignozzi, A. Goudeau, F. Barin and C. Bréchot. 1991. Frequent and early in utero HIV-1 infection. AIDS. Res. Hum. Retroviruses 7:83-88.

Dabis, F., P. Msellati, D. Dunn, P. Lepage, M-L. Newell, C. Peckham, P. Van de Perre, and the Working Group on Mother-to-Child Transmission of HIV. 1993. Estimating the rate of mother-to-child transmission of HIV. Report of a workshop on methodological issues Ghent (Belgium), 17-20 February 1992. AIDS 7:1139-1148.

De Rossi, A., L. Ometto, F. Mammano, C. Zanotto, C. Giaquininto, and L. Chiecho-Bianchi. 1992. Vertical transmission of HIV-1: lack of detectable virus in peripheral blood cells of

infected children at birth. AIDS 6:1117-1120.

Donaldson, Y. K., J. E. Bell, E. C. Holmes, E. S. Hughes, H. K. Brown and P. Simmonds. 1994. *In vivo* distribution and cytopathology of variants of human immunodeficiency virus type 1 showing restricted sequence variability in the V3 loop. J. Virol. 68:5991-6005.

Ehrnst, A., S. Lindgren, M. Dictor, B. Johansson, A. Sönnerborg, J. Czajkowski, G. Sundin, and A.-B. Bohlin. 1991. HIV in pregnant women and their offspring: evidence for late transmission. Lancet 338:203-207.

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

European Collaborative Study. 1994. Caesarean section and risk of vertical transmission of HIV-1 infection. Lancet 343:1464-1467.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein, J. 1993. PHYLIP manual version 3.52c. Berkeley University Herbarium, University of California, Berkeley.

Goedert, J. J., J. E. Drummond, H. L. Minkoff, R. Stevens, W. A. Blattner, H. Mendez, M. Robert-Guroff, S. Holman, A. Rubinstein, A. Willoughby, S. H. Landesman. 1989. Mother-to-infant transmission of human immunodeficiency virus type 1: association with prematurity or low anti-gp120. Lancet 2:1351-1354.

Goedert, J. J., A.-M. Duliège, C. I. Amos, S. Felton, R. J. Biggar, and The International Registry of HIV-Exposed Twins. 1991. High risk of HIV-1 infection for first-born twins. Lancet 338:1471-1475.

Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Applic. Biosci. 8:189-191.

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. J. Infect. Dis. 171:45-53.

Joviasis, E., M. A. Koch, A. Schäfer, M. Strauber, and D. Löwenthal. 1985. LAV/HTLV-III in 20-week fetus. Lancet 2:1129.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules Chapter 11, *In* Munro, H. N. ed. Mammalian protein metabolism, Academic Press, New York, pp. 21-132.

Kishino, H., and M. Hasegawa, 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order of the Hominoidea. J. Mol. Evol. 4:406-425.

Korber, B., and G. Myers. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retroviruses 8:1549-1560.

Krivine, A., G. Firtion, L. Cao, C. Francoual, R. Henrion, and P. Lebon. 1992. HIV replication during the first weeks of life. Lancet 339:1187-1189.

Kumar, S., K. Tamura, and M. Nei. 1993. MEGA: Molecular Evolutionary Genetics Analysis, version 1.0. The Pennsylvania State University, University Park, PA 16802, USA.

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 1 in infants by vertical transmission. J. Clin. Invest. 93:380-390.

Leigh Brown, A. J. and P. Simmonds. 1995. Analysis of HIV sequence variation. Chapter 11, *In* Karn, J. ed. HIV - A Practical Approach. Oxford University Press, pp. 161-188.

Leigh Brown, A. J., D. Lobidel, C. M. Wade, S. Rebus, A. N. Phillips, R. P. Brettle, A. J. France, C. S. Leen, J. McMenamim, A. McMillan, R. D. Maw, F. Mulcahy, J. R. Robertson, K. N. Sankar, G. Scott, and J. F. Peutherer. The molecular epidemiology of human

immunodeficiency virus type 1 in six cities in Britain and Ireland. Submitted for publication.

Lepage, P., P. Vande Perre, M. Carael, F. Nsengumuremyi, J. J. Nkurunziza, J.-P. Butler, and S. Sprecher. 1987. Postnatal transmission of HIV from mother to child. Lancet ii:400.

McNearney, T., P. Westervelt, B. J. Thielan, D. B. Trowbridge, J. Garcia, R. Whittier, and L. Ratner L. 1990. Limited sequence heterogeneity among biologically distinct human immunodeficiency virus type 1 isolates from individuals involved in a clustered infectious outbreak. Proc. Natl. Acad. Sci. USA 87:1917-1921.

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.

Mulder-Kampinga, G. A., C. Kuiken, 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. 74:1747-1756.

Mulder-Kampinga, G. A., A. Simonon, C. L. Kuiken, J. Dekker, H. J. Scherpbier, P. Van De Perre, K. Boer, and J. Goudsmit. 1995. Similarity in *env* and *gag* genes between genomic RNAs of human immunodeficiency virus type 1 (HIV-1) from mother and infant is unrelated to the time of HIV-1 RNA positivity in the child. J. Virol. 69:2285-2296.

Ou, C.-Y., C. A. Ciesieleski, G. Myers, C. I. Bandea, C.-C. Luo, B. T. M. 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, H. W. Jaffe, the Laboratory Investigation Group, and the Epidemiological Investigation Group. 1992. Molecular epidemiology of HIV transmission in a dental practice. Science 256: 1165-1171.

Pizzo, P. A., C. M. Wilfert, and the Pediatric AIDS Siena Workshop II. 1995. Markers and Determinants of Disease Progression in Children with HIV infection. Report of a consensus workshop, Siena, Italy, June 4-6, 1993. J. Acquired Immune Defic. Synd. 8:30-44.

Poss, M., H. L. Martin, J. K. Kreiss, L. Granville, B. Chohan, P. Nyange, K. Mandaliya, and

J. Overbaugh. 1995. Diversity in virus populations from genital secretions and peripheral blood from women recently infected with human immunodeficiency virus type 1. J. Virol. 69:8118-8122.

Rossi, P. 1992. Maternal factors involved in mother-to-child transmission of HIV-1. Report of a consensus workshop, Siena, Italy, 1992. J. Acquired Immune Defic. Synd. 5:1169-1178.

Saitou, N., and M. Nei, 1987. The neighbor-joining method: a new method for reconstructing evolutionary trees. Mol. Biol Evol. 4:406-425.

Scarlatti, G., T. Leitner, E. Halapi, J. Wahlberg, P. Marchisio, M. A. Clerici-Schoeller, H. Wigzell, E. M. Fenyö, J. Albert, M. Uhlén, 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. 90:1721-1725.

Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown. 1990. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. J. Virol. 64:864-872.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet, 1994. The genetic data environment and expandable GUI for multiple sequence analysis. Comput. Appl. Biosci. 10:671-675.

Soeiro, R., A. Rubinstein, W. K. Rashbaum, and W. D. Lyman. 1992. Materno-foetal transmission of AIDS: frequency of human immunodeficiency virus type 1 nucleic acids in sequences in human foetal DNA. J. Infect. Dis. 166:699-703.

Sprecher, S., G. Soumenkoff, F. Puissant, and M. Degueldre. 1986. Vertical transmission of HIV in 15-week fetus. Lancet 2:228-289.

Staden, R., 1993. Staden Package Update. Genome News 13:12-13.

Ugen, K. E., J. J. Goedert, J. Boyer, Y Refaeli, I. Frank, W. V. Williams, A. Willoughby, S. Landesman, H. Mendez, A. Rubinstein, T. Kieber-Emmons, and D. B. Weiner. 1992. Vertical transmission of human immunodeficiency virus (HIV) infection. Reactivity of maternal sera with glycoprotein 120 and 41 peptides from HIV type 1. J. Clin. Invest. **89**:1923-1930.

Van't Wout, A. B., N. A. Kootstra, G. A. Mulder-Kampinga, N. A. Albrecht van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker. 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral and vertical transmission. J. Clin. Invest. **94**:2060-2067.

Weiser, B., H. Burger, S. Nachman, Y. J. Hsu, and R. Gibbs. 1993. Use of serial HIV-1 sequences from a pregnant woman and her twins to study timing of vertical transmission. *In* Ginsberg, H. S., F. Brown, R. M. Chanock, and R. A. Lerner, eds. Vaccines 1993, modern approaches to new vaccines including prevention of AIDS. Cold Spring Habor Laboratory Press, Cold Spring Harbor, New York, pp. 161-188.

Wike, C. M., B. T. M. Korber, M. R. Daniels, C. Hutto, J. Muñoz, M. Furtado, W. Parks, A. Saah, M. Bulterys, J.-B. Kurawige, and S. M. Wolinsky. 1992. HIV-1 sequence variation between isolates from mother-infant transmission pairs. AIDS Res. Hum. Retroviruses **8**:1297-1300.

Wofsy, C. B., J. B. Cohen, L. B. Hauer, N. S. Padian, B. A. Michaelis, L. A. Evans, and J. A. Levy. 1986. Isolation of AIDS-associated retrovirus from genital secretions of women with antibodies to the virus. Lancet **1**:527-529.

Wolfs, T. F. W., G. Zwart, M. Bakker, and J. Goudsmit. 1992. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. Virology **189**:103-110.

Wolinsky, S. M., C. M Wike, B. T. M. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Muñoz. 1992. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science **255**:1134-1137.

Zhang, L. Q., P. Simmonds, C. A. Ludlam, and A. J. Leigh Brown. 1991. Detection,

quantification and sequencing of HIV-1 from the plasma of seropositive individuals and from factor VIII concentrates. AIDS 5:675-681.

Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. 67:3345-3356.

Zhu, T. H. Mo., N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho., 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. Science 261:1179-1181.

Zhu, T., N. Wang, A. Carr, S. Wolinsky, and D. D. Ho., 1995. Evidence for coinfection by multiple strains of human immunodeficiency virus type 1 subtype B in an acute seroconvertor. J. Virol. 69:1324-1327.

Ziegler, J. B., D. A. Cooper, R. Johnson, and G. Gold. 1985. Postnatal transmission of AIDS associated retrovirus from mother to infant Lancet i:896-897.

Table 1.

Times of sampling, clinical data, $CD4^+$ cell counts and the number of molecules sequenced per time point for the patients studied. ND, not determined.

[a] Time course provides details of the sampling time points relative to the dates of birth of child 3 (31.07.92) and child 4 (26.02.94). For the mother, the time is given relative to the delivery date of each child. - indicates sampling dates before delivery, + indicates sampling dates following delivery. Time course details are only provided for maternal samples for which either *env* or *gag* molecules were successfully sequenced. For the children, the age of the child at the time of sampling is given.

[b] The average normal $CD4^+$ cell count for an adult is 1000 cells per mm^3. For children under the age of 11 months, the normal $CD4^+$ cell count is between 1700 and 2880 cells per mm^3 (mean 2200). For children between 1 and 6 years of age, the average $CD4^+$ cell count is between 1000 and 1800 cells per mm^3.

Patient	Sample Name	Sample Date	Time Course[a]		Population	CD4+ Cell counts (cells/mm³)[b]	Clinical Information	CDC Classification	*Env* Sequences	*Gag* Sequences
			Child 3	Child 4						
Mother	1R	15.06.86			Plasma	ND	Asymptomatic	2	0	0
	2R	04.12.87			Plasma	563	"	2	0	0
	3R	15.02.89			Plasma	343	"	2	0	0
	4R	26.10.89			Plasma	250	"	2	0	0
	5R	30.01.90	-1.5 yrs	-4 yrs	Plasma	260	"	2	0	4
	6R	24.04.90			Plasma	330	"	2	0	0
	7R	15.06.90			Plasma	729	Persistent generalised lymphadenopathy	3	0	0
	8R	19.07.90			Plasma	300	"	3	0	0
	9R	15.10.90			Plasma	290	"	3	0	0
	10R	22.11.91			Plasma	ND	"	3	0	0
	11R	20.01.92			Plasma	212	"	3	0	0
	12R	28.02.92	-5 mths	-24 mths	Plasma	349	"	3	1	2
	13R	28.04.92	-3 mths	-22 mths	Plasma	210	"	3	0	3
	14R	08.06.92	-2 mths	-20 mths	Plasma	215	"	3	3	1
	15R	13.07.92	-18 days	-19 mths	Plasma	650	"	3	0	3
	16R	03.08.92			Plasma	342	"	3	0	0
	17R	25.06.93			Plasma	380	"	3	0	0
	18R	03.09.93	+14 mths	-5 mths	Plasma	330	"	3	5	12
	19R	21.12.93	+17 mths	-2 mths	Plasma	270	"	3	4	11
Child 3	1D	25.08.92	1 mth		PBMC	3580	Well	N1	9	10
	1R	"	"		Plasma	"	"	"	1	2
	2R	30.09.92	2 mths		Plasma	2340	Well	N1	2	11
	3R	10.11.92	4 mths		Plasma	2810	Lymphadenopathy, respiratory infection	A1	3	8
	4D	18.12.92	5 mths		PBMC	2510	Lymphadenopathy, respiratory infection	A1	9	11
	4R	"	"		Plasma	"	"	"	0	11
	5D	23.06.93	11 mths		PBMC	960	Respiratory infections, nappy rash, hepatomegaly	A2	11	11
	5R	"	"		Plasma	"	"	"	2	6
	6D	01.10.93	15 mths		PBMC	1750	Respiratory infection, acute parotitis	A2	7	9
	6R	"	"		Plasma	"	"	"	0	11
	7D	03.05.94	22 mths		PBMC	1340	Recurrent upper respiratory infections	A2	11	12
	7R	"	"		Plasma	"	"	"	11	5
	8D	17.08.94	25 mths		PBMC	1820	Lymphadenopathy	A2	0	4
	8R	"	"		Plasma	"	"	"	5	8
Child 4	1R	03.03.94		6 days	Plasma	ND	ND	ND	0	0
	2D	22.03.94		1 mth	PBMC	3550	Well	N1	9	12
	2R	"		"	Plasma	"	"	"	8	11
	3D	10.06.94		4 mths	PBMC	3210	Well	N1	11	11
	3R	"		"	Plasma	"	"	"	1	4
	4D	05.08.94		6 mths	PBMC	3590	Hepatomegaly	A1	10	10
	4R	"		"	Plasma	"	"	"	1	1

Table 2.

Inter patient distances and intra patient diversities for the infected mother and children within the V3 loop and flanking regions of *env* (a) and the p17 region of *gag* (b). Distances were calculated using the generalised two-parameter (maximum likelihood) distance estimate (program DNADIST; Felsenstein 1993). Intra patient distances are shown in bold on the diagonal. n indicates the number of sequences within each sample.

a.

	Mother	Child 3	Child 4
Mother	**0.0167**		
Child 3	0.0284	**0.0277**	
Child 4	0.0218	0.0385	**0.0056**

b.

	Mother	Child 3	Child 4
Mother	**0.0224**		
Child 3	0.0213	**0.0125**	
Child 4	0.0218	0.0270	**0.0027**

Table 3.

Nucleotide sequence diversity within the V3 loop and flanking regions of *env* and the p17 region of *gag* for the mother and her infected children. ML indicates a generalised two-parameter (maximum likelihood) distance estimate (program DNADIST; Felsenstein 1993), d_s, the proportion of synonymous nucleotide substitutions per synonymous site, and d_n, the proportion of nonsynonymous nucleotide substitutions per nonsynonymous site. d_s and d_n values were calculated using the Jukes-Cantor one-parameter model as implemented in MEGA version 1.01 (Kumar *et al.*, 1993). n indicates the number of sequences within each sample.

Patient	Sample Name	V3 loop and flanking regions of *env*					p17 Region of *gag*				
		n	ML	Ks	Ka	Ka/Ka	n	ML	Ks	Ka	Ka/Ka
Mother	5R	0	-	-	-	-	4	0.0160	0.0244	0.0138	1.7681
	12R	1	0.0000	0.0000	0.0000	-	1	0.0000	0.0000	0.0000	-
	13R	0	-	-	-	-	3	0.0199	0.0326	0.0165	1.9758
	14R	3	0.0102	0.0178	0.0082	2.1707	0	-	-	-	-
	15R	0	-	-	-	-	3	0.0203	0.0333	0.0169	1.9704
	18R	5	0.0182	0.0077	0.0209	0.3684	12	0.0166	0.0342	0.0119	2.8739
	19R	4	0.0202	0.0086	0.0232	0.3707	11	0.0181	0.0285	0.0154	1.8506
Child 3	1D	9	0.0221	0.0225	0.0219	1.0274	10	0.0113	0.0246	0.0079	3.1139
	1R	1	0.0000	0.0000	0.0000	-	2	0.0103	0.0243	0.0066	3.6818
	2R	2	0.0108	0.0128	0.0103	1.2428	11	0.0120	0.0208	0.0085	2.4470
	3R	3	0.0183	0.0213	0.0174	1.2241	8	0.0097	0.0289	0.0046	6.2826
	4D	9	0.0232	0.0216	0.0240	0.9000	11	0.0100	0.0306	0.0046	6.6522
	4R	0	-	-	-	-	11	0.0083	0.0260	0.0037	7.0270
	5D	11	0.0178	0.0234	0.0165	1.4182	11	0.0153	0.0373	0.0096	3.8854
	5R	2	0.0168	0.0265	0.0143	1.8531	6	0.0208	0.0448	0.0112	4.0000
	6D	7	0.0145	0.0260	0.0114	2.2807	9	0.0085	0.0334	0.0020	16.700
	6R	0	-	-	-	-	11	0.0077	0.0303	0.0015	20.200
	7D	11	0.0217	0.0236	0.0212	1.1132	12	0.0091	0.0337	0.0026	12.961
	7R	11	0.0213	0.0275	0.0197	1.3959	5	0.0093	0.0398	0.0013	30.615
	8R	5	0.0163	0.0257	0.0139	1.8489	8	0.0164	0.0428	0.0095	4.5053
Child 4	2D	9	0.0006	0.0000	0.0008	0.0000	12	0.0009	0.0000	0.0011	0.0000
	2R	8	0.0012	0.0056	0.0000	-	11	0.0014	0.0022	0.0012	1.8333
	3D	11	0.0042	0.0000	0.0053	0.0000	11	0.0022	0.0084	0.0006	14.000
	3R	1	0.0000	0.0000	0.0000	-	4	0.0000	0.0000	0.0000	-
	4D	10	0.0070	0.0049	0.0084	0.5833	10	0.0065	0.0164	0.0039	4.2051
	4R	1	0.0000	0.0000	0.0000	-	1	0.0000	0.0000	0.0000	-

Figure 1.

Unrooted neighbour-joining phylogenetic trees reconstructed from 124 *env* (a) and 198 p17 *gag* (b) gene sequences for the mother-child transmission set. Maternal sequences are represented by diagonals, sequences of child 3 by circles and sequences of child 4 by squares. Numbers within each symbol represent the sampling time point. Closed symbols denote plasma RNA sequences, open symbols, proviral DNA sequences. The scale bar corresponds to 1% nucleotide sequence divergence. Bootstrap values, based on 100 bootstrap replications, are expressed as a percentage.

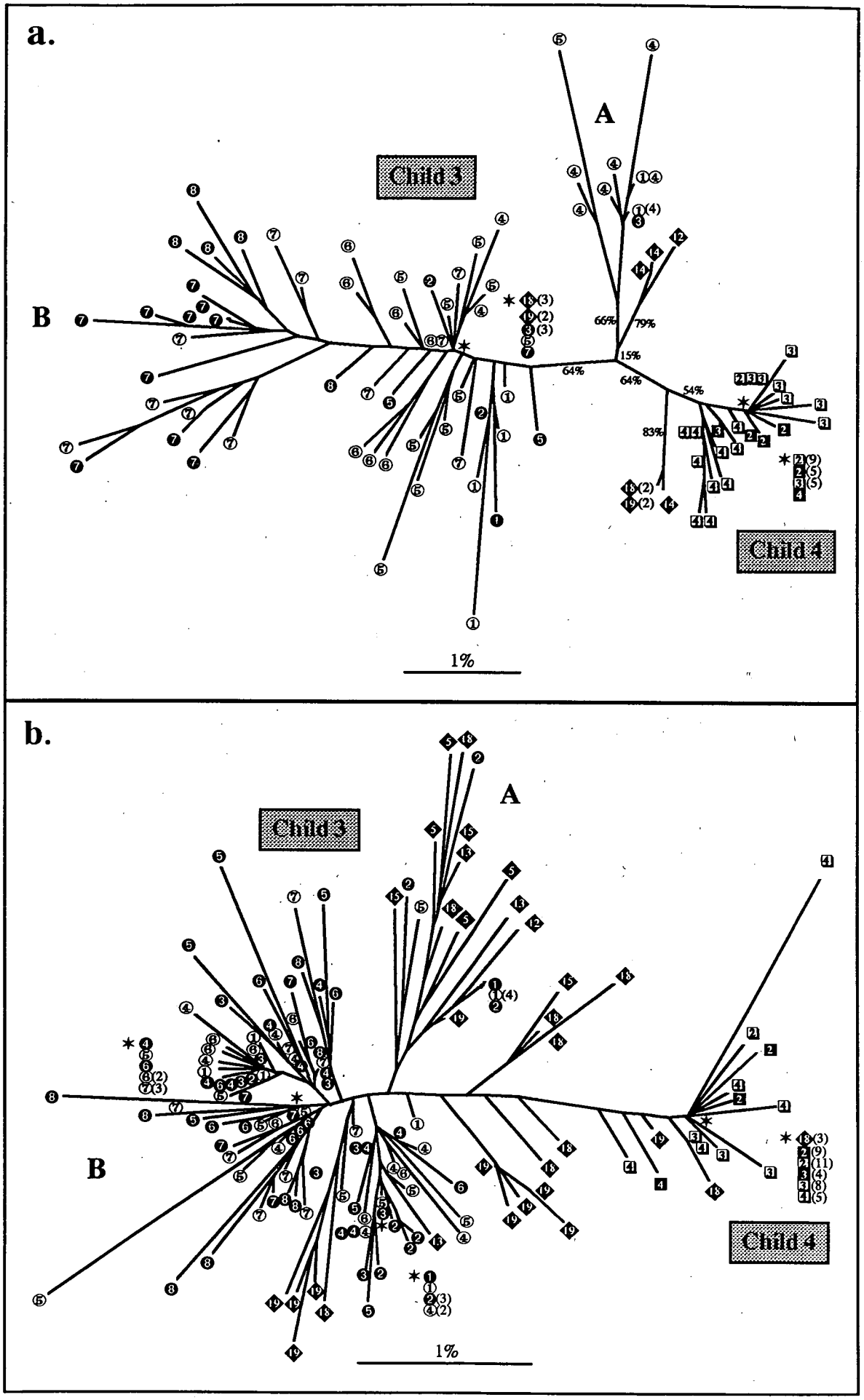
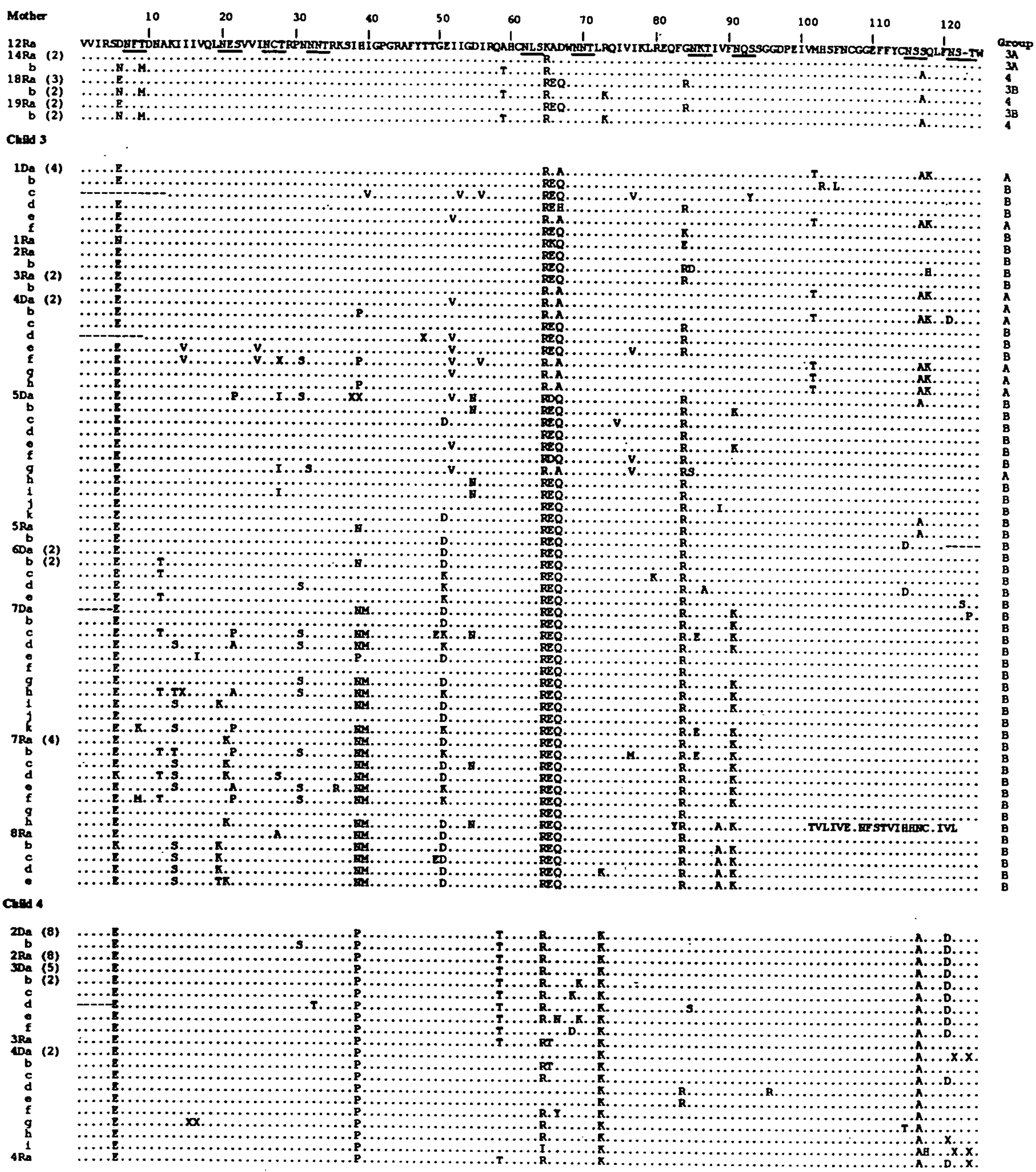

a.
A
B
Child 3
Child 4
66%
79%
15%
64%
64%
54%
83%
1%
b.
A
B
Child 3
Child 4
1%

Figure 2.

Multiple alignments of deduced amino acid sequences of the V3 loop and flanking regions of the *env* gene (a) and the p17 region of the *gag* gene (b) for the mother-child transmission set. Sequences are presented in order of the time of sampling and are aligned relative to the most common variant present within the first sample of the mother. Amino acids identical to the first sequence are replaced by a dot, dashes represent alignment gaps, and stop codons are represented by asterisks. X indicates unreadable bases in the nucleotide sequence at this position. Sequences are labelled according to sampling time point. R denotes plasma RNA sequences, D, proviral DNA sequences. The number of identical sequences for a time point are indicated in parentheses. Putative N-linked glycosylation sites are underlined. Sequences of child 3 were classified according to their phylogenetic grouping (lineages A and B, Figure 1a and 1b). Sequences of the mother were classified according to their phylogenetic association with a particular child and a particular child lineage (Figure 1a and 1b).

a.


Mother
10 20 30 40 50 60 70 80 90 100 110 120
12Ra VVIRSDNFTDNAKIIIVQLNESVVINCTRPNNNTRKSIHIGPGRAFYTTGEIIGDIRQAHCNLSKADWNNTLRQIVIKLREQFGNKTIVFNQSSGGDPEIVMHSFNCGGEFFYCNSSQLFNS-TW
Group
Child 3
Child 4

b.

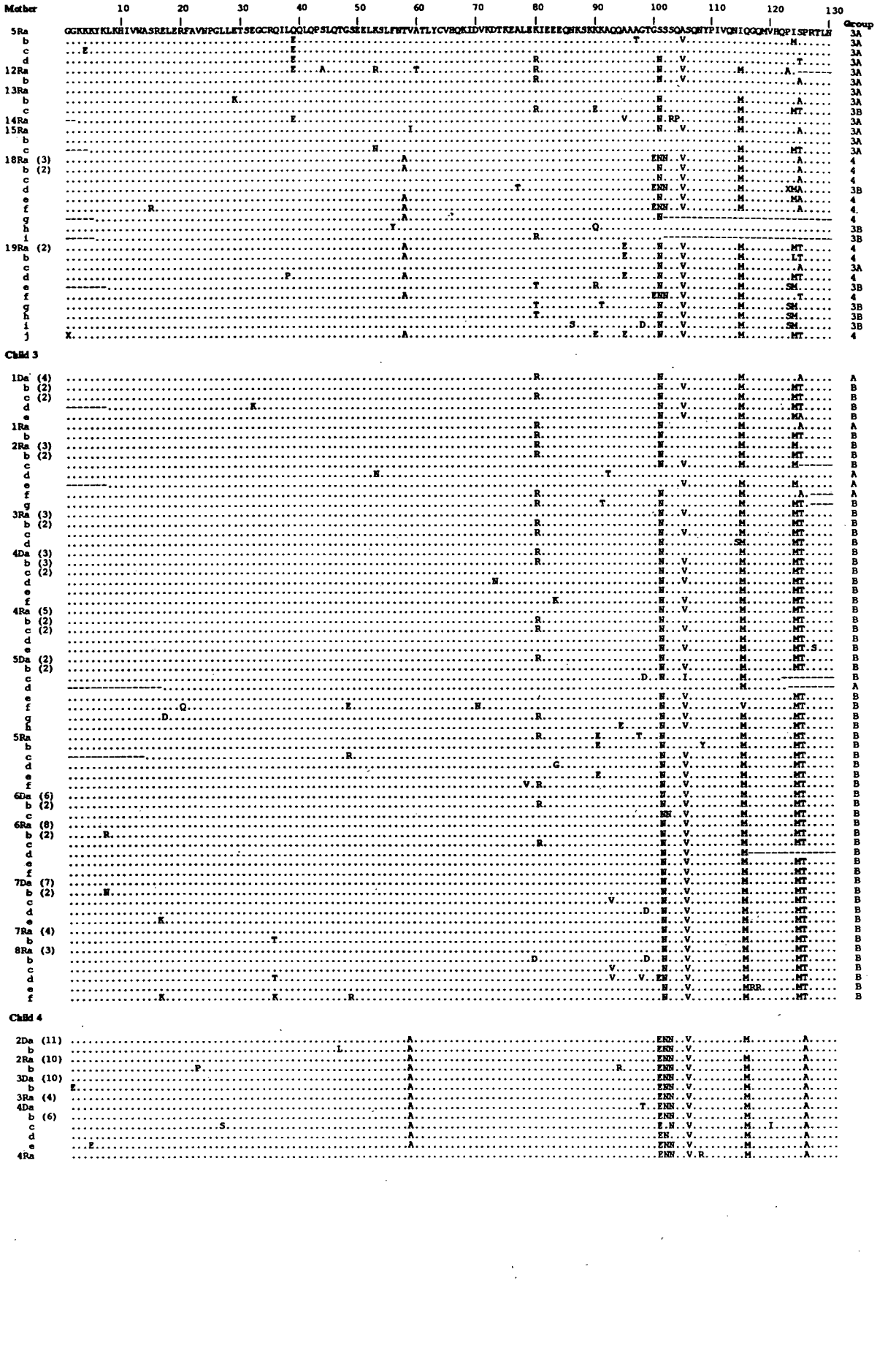
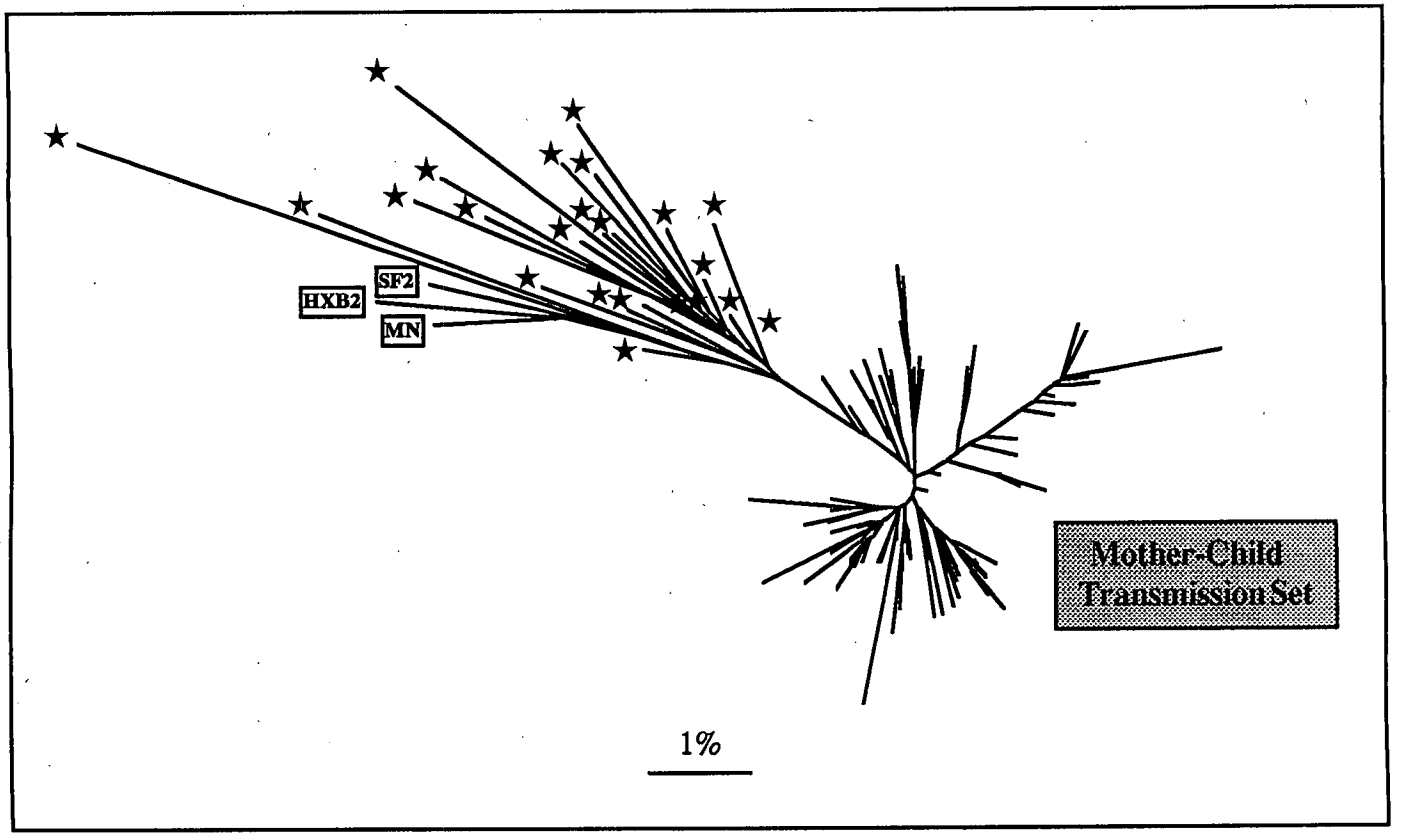

Mother
10
20
30
40
50
60
70
80
90
100
110
120
130
Group
5Ra GGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQTGSEELKSLFNTVATLYCVHQKIDVKDTKEALEKIEEEQNKSKKKAQQAAAGTGSSSQASQNYPIVQNIQGQMVHQPISPRTLN
Child 3
Child 4

Figure 3.

Unrooted neighbour-joining p17 *gag* gene phylogeny showing the relationships between the 198 sequences obtained from the mother-child transmission set and sequences of 23 epidemiologically unrelated HIV-1 infected intravenous drug users from the Edinburgh cohort and 3 HIV-1 subtype B reference isolates. Sequences of the subtype B reference isolates are labelled directly on the tree, stars denote individual Edinburgh intravenous drug user sequences. Sequences obtained from the mother-child transmission set are not labelled individually. The scale bar corresponds to 1% nucleotide sequence divergence.


HXB2
SF2
MN
Mother-Child
Transmission Set
1%

Paper V

Evolution of Human Immunodeficiency Virus Type 1 During Infection of a Male Index and Transmission to Two Infected Female Partners and Their Children

Christopher M. Wade[1*], Denis Lobidel[1], J. Roy Robertson[2], Jacqueline Y. Q. Mok[3], David Yirrel[1], and Andrew J. Leigh Brown[1]

[1] Centre for HIV Research, Institute of Cell, Animal and Population Biology, Division of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3JN, United Kingdom.

[2] Muirhouse Medical Group, 1 Muirhouse Avenue, Edinburgh.

[3] City Hospital, 51 Greenbank Drive, Edinburgh, EH10 5SB.

[*] Corresponding Author

Submitted for Publication

ABSTRACT

We have examined a transmission set comprising a male index, two infected female partners and their respective HIV-1 infected children. Nucleotide sequences for both the V3 loop and flanking regions of the *env* gene and the p17 region of the *gag* gene were obtained from sequential proviral DNA and plasma RNA samples donated by the patients over a 10 year period. Within the male index, an increase in genetic diversity was observed with time, with an evolutionary diversification of the plasma viral sequences of the index into three lineages by year 7. A progressive increase in the mean genetic distance to the seroconversion sequence was observed. Two distinct sequence populations were transmitted to the two female partners who were infected at least 8 years apart. The distance between the two mother-child pairs reflected the level of evolutionary change within the index in the period between the two transmission events. Both sexual and vertical transmission events led to a homogeneous sequence population within the newly infected individual immediately following infection. The phylogenetic placement of the sequences of the transmission set within a background population of Edinburgh patients revealed that the early sequences of the male index showed a closer relationship to a number of Edinburgh intravenous drug user sequences than to sequences obtained from the index later in the course of infection or indeed sequences obtained from the female transmission contacts.

Running Title: HIV-1 Evolution Within a Transmission Set

Molecular studies examining the transmission of human immunodeficiency virus type 1 (HIV-1) have reported that following sexual, parenteral, or vertical transmission, a highly homogeneous sequence population is generally observed within the recipient in *env* shortly after infection (McNearney *et al.*, 1990; Wolfs *et al.*, 1992; Zhang *et al.*, 1993; Zhu *et al.*, 1993; Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Scarlatti *et al.*, 1993; Mulder-Kampinga *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995). This is in contrast to the heterogeneous sequence population typically observed within a long term infected individual. It has therefore been proposed that a sequence bottleneck occurs upon infection, with infection initiated by a limited number or even one particular variant. The homogeneous *env* sequence population upon seroconversion may be due to either a dilution effect (in which only a single viable virus is transmitted), differential outgrowth (in which transmission leads to coinfection with several different viruses but one virus outgrows the others by chance) or selective penetration (in which certain variants preferentially infect cells lining the place of entry) (Zhang *et al.*, 1993; Zhu *et al.*, 1993). p17 *gag* gene sequences do not however show such a restricted level of sequence diversity within the recipient upon seroconversion (Zhang *et al.*, 1993; Zhu *et al.*, 1993) which suggests that the homogeneity observed in *env* is due to strong selection for specific *env* sequences either upon transmission or in the interval between exposure and seroconversion (Zhang *et al.*, 1993). Indeed, it has been suggested that V3 variants similar to the global consensus sequence for North American and European subtype B sequences, are selected for on the initiation of a new infection (Zhang *et al.*, 1993). The apparent transmission of multiple HIV-1 genotypes has nonetheless been reported recently for a small number of adult (Ou *et al.*, 1992, Korber and Myers, 1992; Albert *et al.*, 1993; Zhu *et al.*, 1995; Poss *et al.*, 1995) and vertical transmission cases (Lamers *et al.*, 1994; Van't Wout *et al.*, 1994; Briant *et al.*, 1995; Wade *et al.*, submitted).

Phylogenetic analyses examining the molecular epidemiology of HIV-1 have shown that epidemiologically related individuals consistently cluster together within a background of unrelated individuals in reconstructed evolutionary trees (Kuiken *et al.*, 1996; Leigh Brown *et al.*, submitted). Previously unknown contact networks, supported by high bootstrap values in the reconstructed trees, may even be identified from within a background of unrelated patients (Leigh Brown *et al.*, submitted). Phylogenetic analysis techniques have been shown to cluster sequences derived from related patients even at 10 years post transmission (Kuiken

et al., 1996) and have also been shown to reliably reconstruct the transmission events in a lengthy transmission chain consisting of 11 patients with 10 transmission events between 1981 and 1993 (Leitner *et al.*, 1996). HIV sequence data has also been used in a number of forensic applications to identify individuals belonging to transmission clusters (Ou *et al.*, 1992, Korber and Myers, 1992; Albert *et al.*, 1993; Holmes *et al.*, 1993); Jaffe *et al.*, 1994; Arnold *et al.*, 1995).

In this study we have examined a transmission set comprising a male index, two infected female partners and their respective HIV-1 infected children. Sequential proviral DNA and plasma RNA samples were donated by the patients over a 10 year period and nucleotide sequences of both the V3 loop and flanking regions of the *env* gene and the p17 region of the *gag* gene obtained. Viral sequence variations within the male index over time are described and the viral sequence population of the index is compared to those of the two infected female partners and their children. The phylogenetic relationships between viral sequences of the transmission set and sequences derived from a background population of HIV-1 infected patients from the Edinburgh intravenous drug user cohort is also described.

Patients

Sequential plasma and peripheral blood mononuclear cell (PBMC) samples were obtained over a period of 10 years from a transmission set comprising five HIV-1 infected patients. This set consisted of a male intravenous drug user (male index) who infected two female partners (mother 1 and 2) who subsequently gave birth to two HIV-1 infected children (child 1 and 2 respectively).

The male index first used drugs by injection in 1977. Although the patients first antibody test (which subsequently tested positive) was in June 1985 a blood sample obtained in December 1984 was PCR positive for HIV sequences (Table 1). Between late 1983 and 1984 a major epidemic of HIV occurred amongst drug users in Edinburgh (Robertson *et al.*, 1985). The patient has no AIDS defining illness at the present time. Mother 1 was diagnosed HIV antibody positive in May 1986 and her first AIDS defining illness was pneumocystis carinii pneumonia in October 1987. She had oesophageal candidiasis, a further episode of pneumonia, cerebral toxoplasmosis, and evidence of dementia in 1988 and died on 22. June 1989. Child 1 was born to mother 1 by instrumental vaginal delivery on 9. November 1984 and she was diagnosed HIV antibody positive in June 1986. The mother had a slight postpartum haemorrhage and lost about 500mls of blood. Since child 1 was not breast fed mother 1 must have been infected at least two years prior to her first HIV test. Child 1 was found to have persistent hepatosplenomegaly, iron deficiency anaemia, diarrhoea, recurrent thrombocytopaenia, behavioural abnormalities and generalised lymphadenopathy. Cardiomyopathy was diagnosed in August 1990 and the child deteriorated subsequent to this and died on 12. September 1991. Mother 2 was found to be HIV antibody positive in February 1993 when she attended for antenatal care. She remained symptomatically well with the exception of an abnormal cervical smear in June 1994. Child 2 was born to mother 2 by spontaneous vaginal delivery on 2. September 1993 with no apparent complications and he was confirmed HIV positive in December 1993 following positive virus culture on 2 occasions. The child had suspected meningitis in December 1993 and confirmed upper lobe pneumonia in February 1994 and was subsequently found to have obstructive emphysema of the upper left lobe and confirmed as having developed mental decay. The male index was the assumed father of both children.

The times of sampling, clinical status and $CD4^+$ cell counts of the patients studied are

presented in Table 1.

PCR amplification and sequencing.

EDTA treated and heparinized whole blood samples were separated on Ficoll-Hypaque (Pharmacia) and stored in liquid nitrogen (PBMCs) or at -20°C (plasma). DNA and RNA extraction, PCR amplification and automated DNA sequencing were performed essentially as described previously (Simmonds *et al.*, 1990; Zhang *et al.*, 1991; Leigh Brown and Simmonds, 1995).

An approximately 436 base pair (bp) fragment spanning the V3 loop and flanking regions of the *env* gene (positions 7029 to 7464 in the HIV-HXB2 genome (Genbank Accession Number: K03445) and an approximately 390 bp fragment of the p17 coding region of the *gag* gene (positions 857 to 1246 in the HIV-HXB2 genome) were amplified by limiting dilution nested PCR (Simmonds *et al.*, 1990), with 30 cycles in both the first and second rounds of amplification. The following primers were used:

env

1 (outer, +): 5'-TACAATGTACACATGGAATT-3' (nucleotides 6957-6976 HXB2 sequence)

2 (outer, -): 5'-GGAGGGGCATACATTGC-3' (7520-7537 HXB2 sequence)

3 (inner, +): 5'-TGGCAGTCTAGCAGAAGAAG-3' (7009-7028 HXB2 sequence)

4 (inner, -): 5'-ATTCTGCATGGGAGTGTG-3' (7465-7482 HIV-HXB2 sequence)

gag

1 (outer, +): 5'-GCGAGAGCGTCAGTATTAAGCGG-3' (795-817 HXB2 sequence)

2 (outer, -): 5'-TCTGATAATGCTGAAAACATGGG-3' (1296-1318 HXB2 sequence)

3 (inner, +): 5'-GGGAAAAAATTCGGTTAAGGCC-3' (833-856 HXB2 genome)

4 (inner, -): 5'-CTTCTACTACTTTTACCCATGC-3' (1247-1270 HXB2 sequence)

Both sense and antisense strands of the *env* and *gag* amplification products were sequenced using a direct solid phase automated sequencing approach using the T7 dye terminator sequencing chemistry (detailed in Leigh Brown and Simmonds, 1995). Sequencing products were run on an Applied Biosystems 373A automatic DNA sequencer.

Sequence analysis

Raw nucleotide sequences were assembled using the STADEN package (Staden, 1993). Sequences were then aligned using the CLUSTAL V algorithm (Higgins *et al.*, 1992), as implemented in version 2.2 of the Genetic Data Environment (GDE) package (Smith *et al.*,

1994). The final alignment was improved manually by preferring gaps to transition differences, transition differences to transversion differences and by the insertion of gaps to maintain the reading frame. Phylogenetic analyses were performed using programs taken from version 3.52c of the Phylogeny Inference Package (PHYLIP; Felsenstein, 1993). Distance-based phylogenetic analyses were carried out using the neighbour-joining (Saitou and Nei, 1987) (program NEIGHBOR) and Fitch-Margoliash (Fitch and Margoliash, 1967) (program FITCH) methods, with nucleotide sequence distances estimated for all pairwise sequence comparisons using the generalised two-parameter (maximum likelihood) model (Kishino and Hasegawa, 1989) (program DNADIST). Maximum likelihood (Felsenstein 1981) phylogenetic analyses were performed using the modified PHYLIP program FASTDNAML (kindly provided by Gary Olsen of the University of Illinois at Urbana-Champain and the Ribosomal Database Project), and maximum parsimony (Fitch, 1971) analyses were performed using the program DNAPARS. Bootstrap resampling (Felsenstein 1985) (programs SEQBOOT and CONSENSE) was employed to assign support to the neighbour-joining trees (100 bootstrap replicates). Alternative phylogenetic hypotheses were evaluated statistically by a likelihood ratio test (Kishino and Hasegawa 1989) following the assignment of log likelihoods (program DNAML) to artificially generated topologies (program RETREE). The number of synonymous substitutions per synonymous site (d_s) and the number of nonsynonymous substitutions per nonsynonymous site (d_n) were calculated using the Jukes Cantor one-parameter model (Jukes and Cantor, 1969) as implemented in the Molecular Evolutionary Genetics Analysis program version 1.01 (MEGA; Kumar *et al.*, 1993).

The sequence dataset was screened for the presence of potential contaminants by comparing the patient sequences with sequences of equivalent regions of all clones and other patients examined within the laboratory.

Nucleotide sequence accession numbers

Nucleotide sequences reported in this study have been assigned the GenBank accession numbers XXXXXX to YYYYYY.

RESULTS

We have obtained 185 sequences spanning the V3 loop and flanking regions of the *env* gene and 169 sequences of the p17 region of the *gag* gene from a transmission set comprising a male index, two infected female partners and their respective HIV-1 infected children. Sequences were obtained from 28 of the 33 sampling time points examined and spanned a period of 10 years, 1984 to 1994 (Table 1).

Viral evolution within the index

Viral plasma RNA sequences were obtained from six samples taken from the index between December 1984 (year 0) and February 1993 (year 9) (Table 1). Two samples obtained in year 8, 6 weeks apart (21.10.92 and 12.12.92) are treated as a single time point in all analyses. In addition to the RNA sequences, proviral DNA sequences were obtained in years 6 and 7. Throughout the period of study the patient remained asymptomatic; $CD4^+$ cell counts fluctuated between 320 and 6 (Table 1).

Four distinct amino acid sequence variants were observed within the viral population of the index within both the V3 loop and flanking regions of *env* and the p17 region of *gag* (groups A to D; Figure 1a and 1b). Within *env*, variant C was further subdivided into C and C' (Figure 1a). The potential phenotype of the amino acid sequence variants was predicted on the basis of the global net charge of the V3 loop and the degree of sequence divergence from the LaRosa subtype B consensus (Milich *et al.*, 1993; Donaldson *et al.*, 1994). The majority of sequences from the male index were predicted to be of the macrophage-tropic, non-syncytium-inducing (NSI) phenotype. However, a number of sequences obtained from the year 5 RNA sample and the year 6 DNA sample were predicted to be of the T-cell-tropic, syncytium-inducing (SI) phenotype.

The overall intrapatient sequence diversity for all samples obtained over the nine year period from the male index was 4.1% within the V3 loop and flanking regions of *env* and 2.1% within the p17 region of *gag*. In the samples large enough to calculate the intrasample sequence diversity (years 7-9) the index showed a slight increase in the level of intra sample sequence diversity over time (Table 2a). In *env*, sequence diversity within the plasma RNA population was 1.3% in year 5 (n=2), 3.6% (n=8) in year 7, 2.6% (n=10) in year 8, and 4.5% (n=12) in year 9. Although sample sizes in *gag* were smaller, an increase from 1.6% (n=3) to 2.6% (n=3) was seen from year 5 to year 9. Comparison of the mean distance between

later time points and the earliest time point obtained from the patient (Table 2b) revealed that the distance from the seroconversion 'A' type variant was approximately 4% in the year 7 plasma virus population in *env*, 4.2% in year 8 and 4.4% in year 9. In *gag*, the distance to the year 0 sequence, which was 1.6% in the plasma virus population in year 5, increased to 2.1% in the year 9 sample.

Maximum likelihood phylogenetic trees reconstructed for both the *env* and *gag* nucleotide sequence datasets of the index (Figure 2a and 2b) clearly depict the division of the index's sequences into four groups and the relationship between observed sequence variants and time. *Env* sequences from the earliest samples obtained from the index fell within a single group which comprised both year 5 plasma RNA and 6/12 year 6 proviral DNA sequences (*env* group A; Figure 2a). The three sequences of p17 *gag* obtained from year 0 plasma RNA also formed a distinct group (*gag* group A; Figure 2b). It was not possible to obtain RNA sequences within *env* from year 0.

There appeared to be an evolutionary diversification of the plasma viral sequences of the index into three lineages by year 7. Some of the year 6 proviral sequences were apparently intermediate between group A and the new groups, B, C and D (Figure 2A). A similar diversification occurred in *gag* (Figure 2b). The three *env* lineages persisted within the plasma virus population of the index until the most recent sample examined (year 9). In *gag*, lineages C and D were observed within the plasma virus population of the index from year 7, but lineage B was not observed in the plasma population in *gag* until year 9. All three lineages were nonetheless observed in *gag* proviral DNA sequences from both years 6 and 7.

All methods of phylogeny reconstruction employed (neighbour joining, Fitch-Margoliash and maximum likelihood) gave similar trees for the nucleotide sequences of the male index, and the division of the index's sequences into four groups was also clearly apparent from neighbour-joining trees reconstructed from amino acid sequence data for both *env* and *gag*.

Relationship between viral sequences of the index, and sequences of the two mother-child transmission pairs

Phylogenetic analysis of *env* and *gag* sequences of the infected female partners and their respective children (Figure 3a and 3b) revealed that these sequences cluster distinctly from one another supported in 100% of bootstrap replicates for both *env* and *gag* analyses

(data not shown).

Only a single sequence in *env* was obtained from mother 1, but from the child 9 *env* sequences and 10 *gag* p17 sequences were obtained late in infection (year 7). Nevertheless, for both genes these sequences were most closely associated with the earliest sequences found in the index ("Group A"; Figure 3a and b), although by a long branch. A nucleotide sequence distance of 4.7% was observed between mother 1 and male index group A, with a distance of 10.9% observed between child 1 and male index group A. A far more restricted level of sequence diversity was observed in *gag*, with a distance of 1.6% observed between child 1 and male index group A.

A large number of sequences of mother-child pair 2 were available from several time points. Nevertheless, all were observed to cluster with *env* and *gag* sequences of group D in the index (Figure 3a and 3b). Indeed, within *env*, 2 sequences from the year 8 sample of the index which are identical at the amino acid level (8Re, Figure 1a) cluster within the mother-child group (Figure 3a). One of these was identical in nucleotide sequence to the most common sequence variant observed within both mother and child 2, while the other sequence differed at only 1 nucleotide positions. These sequences would thus appear to represent the precise *env* sequence variants responsible for initiating infection within the mother. No *gag* sequence variants were identified in the male index which were identical to those sequences observed in mother-child pair 2.

Viral evolution within individual mother-child transmission pairs

Child 1 showed relatively high levels of sequence diversity in the single sample available (4.2% in V3 *env*; 1.4% in p17 *gag*, Table 3b). This sample was obtained from a single time point during the final stages of disease and at least 7 years after transmission. Viral sequences obtained from this sample were predicted to be of the T-cell-tropic, syncytium-inducing (SI) phenotype on the basis of the global net charge of the V3 loop and the degree of sequence divergence from the LaRosa subtype B consensus (Milich *et al.*, 1993; Donaldson *et al.*, 1994).

Intra patient sequence diversity was extremely low for both mother 2 (0.27% *env*; Table 3a and 0.30% *gag*; Table 3b) and child 2 (0.78% *env*; Table 3a and 0.53% *gag*; Table 3b) within both *env* and *gag* contrasting with overall sequence diversity levels of 4.1% in *env* and 2.1% in *gag* within the infected male index. Intra sample diversity varied between 0% and 0.9% in *env* and between 0.2% and 0.6% in *gag* within the mother in the samples

obtained across pregnancy. Child intra-sample diversity varied between 0.3% and 1.3% in *env* and between 0.2% and 0.7% in *gag* within the samples taken from the child spanning the first year of life. All viral sequences obtained from mother 2 and child 2 were predicted to be of the macrophage-tropic, non-syncytium-inducing (NSI) phenotype on the basis of the global net charge of the V3 loop and the degree of sequence divergence from the LaRosa subtype B consensus (Milich *et al.*, 1993; Donaldson *et al.*, 1994).

Viral evolution between transmissions

Comparison of evolutionary distances between the single sequence available from mother 1 and sequences from mother 2 revealed a distance of 7.4% within *env* (Table 3a). Samples from the two children were separated by an evolutionary distance of 11.9% in *env* (Table 3a) and 3.2% in *gag* p17 (Table 3b). The greater evolutionary distances between the children reflects both the degree of viral evolution within the male index in the 9 year period between the two transmission events and the long branch to child 1 sequences, probably related to the long period between transmission and sampling in this case.

Genetic Relationships with other Edinburgh HIV-1 sequences

Phylogenetic analyses of sequences from the transmission set with representative sequences of identified HIV-1 risk groups circulating in Edinburgh (Holmes *et al.*, 1995; Leigh Brown *et al.*, submitted), revealed that the patients fell within the intravenous drug user (IDU) clade. This is consistent with the available background data for the male index. This patient first used drugs by injection in 1977 and is believed to have formed part of the early Edinburgh cohort. While *env* sequences on other patients are unavailable, an analysis of *gag* gene sequences from the transmission set with all available Edinburgh IDU sequences (Figure 4) revealed that sequences obtained at different times during the infection did not consistently group together but mixed with sequences obtained from other unlinked Edinburgh IDU patients. In particular, sequences obtained from the male index in year 0 clustered tightly with sequences obtained from three unrelated Edinburgh IDU patients (group A; Figure 4) and were surprisingly more closely related to the sequences from these patients than to sequences obtained from later samples in the course of the male index's infection and consequently sequences obtained from the epidemiologically linked mothers and their children.

We have examined a transmission set comprising a male index, two infected female partners and their respective HIV-1 infected children. Analyses were performed on sequences generated for both the V3 loop and flanking regions of the *env* gene and the p17 region of the *gag* gene for sampling time points over a 10 year period from the 4 members of the transmission set.

Viral evolution within the male index

An increase in genetic diversity was observed in the male index with time, with a progressive increase in the genetic distance from the seroconversion sequence. Intrapatient sequence diversity within the index (4.1% within the V3 region and flanking sequences of *env* and 2.1% within the p17 region of *gag*) is similar to levels reported previously. Balfe *et al.* (1990) estimated a mean within patient diversity of 4.2% in the V3 region of *env* and 1.7% in the p24 region of *gag* for 6 haemophiliac patients. Intrapatient diversity levels of between 5 and 6% have been observed within *env* in later samples from this group (Leigh Brown and Cleland, 1996).

During the infection of the male index there has been a replacement of the seroconversion A type sequence by three new lineages, B, C, and D, by year 7. These three lineages persisted within the viral population of the index throughout the time period studied. Multiple lineages have frequently been described in long-term infected patients in the V3 region (Holmes *et al.*, 1992; Leigh Brown and Cleland, 1996; Wolinsky *et al.*, 1996). Holmes *et al.* (1992) followed a haemophiliac patient over a period of 7 years and showed that following a homogeneous V3 sequence population at seroconversion, an evolutionary diversification occurred at 3 years post infection with the lineages present at this time persisting throughout the remainder of the infection. Complex fluctuations in frequency of the different lineages present in the plasma occurred during infection, generally suggesting a negative frequency-dependent effect on fitness (Holmes *et al.* 1992). These fluctuations are compatible with the development of strain specific immune responses as hypothesised by Nowak *et al.* (1991), however Leigh Brown (1997) has found the stronger selective effect in this dataset to be for differences in cell tropism.

Relationship between viral sequences of the index, and sequences of the two mother-

child transmission pairs

Sequences obtained from the two mother-child transmission pairs clustered apart from each other within phylogenetic trees containing sequences from all 5 individuals. Each pair showed an association with a particular group of index sequences. For each mother-child pair, the variant transmitted to the child was a major form in the male index at or close to the inferred time of transmission. Sequences of mother-child pair 1 showed an association with the earliest sequences obtained from the index (group 'A' type variants in both *env* and *gag*). This variant was the only form identified within the male index at the inferred infection date for mother 1 who gave birth to the infected child (child 1) on 9. November 1984. As the child was not breast fed, mother 1 must have been infected prior to this date, although her infection was not detected until May 1986. In contrast, sequences of mother-child pair 2 associated with sequences from index group D, and particularly strongly with *env* sequences from the year 8 samples of the index. This correlates well with the date of the first positive HIV test (February 1993) for mother 2.

Viral evolution within individual mother-child transmission pairs

Restricted levels of sequence diversity were observed in mother and child 2 shortly after infection in contrast to the more heterogeneous sequence population observed within the long term infected male index. This is consistent with reported levels of sequence diversity immediately following infection for the majority of cases of transmission either via sexual contact (McNearney *et al.*, 1990; Wolfs *et al.*, 1992; Zhang *et al.*, 1993; Zhu *et al.*, 1993) or from mother to child (Wike *et al.*, 1992; Wolinsky *et al.*, 1992; Scarlatti *et al.*, 1993; Mulder-Kampinga *et al.*, 1993; Ahmad *et al.*, 1995; Mulder-Kampinga *et al.*, 1995). Mother 2 was first diagnosed HIV-1 infected when she attended for antenatal care (February 1993) and transmission to the mother is thought to have occurred immediately prior to or during the first stages of pregnancy. Evolutionary diversification of viral sequences occurred in the mother in the very short span of time between infection of the mother and transmission to the child, and therefore infection of both the mother and the child resulted from the transmission of the same sequence variant from the index.

No sequences were obtained from mother-child pair 1 until long after infection. Consequently it was not possible to make comparisons of levels of sequence diversity observed shortly after infection in mother and child 2 and reported following the vast majority of transmission events. The level of intrapatient sequence diversity and sequence

divergence from index group A sequences observed in child 1 in the sample collected one month before the death of the child was comparable to that observed within the long term infected male index over a similar time period.

Viral evolution between transmissions and implications for the identification of epidemiologically related individuals

This study has shown that evolution within the index can account for large initial differences between the virus populations of individuals infected from the same source. When sequences of child 1, obtained one month before the death of the child, were compared with sequences obtained from child 2, an evolutionary distance of 11.9% was observed in *env* (Table 3a), with a distance of 3.2% observed in *gag* (Table 3b), distances similar to those described from unlinked patients of the same subtype. As child 2 sequences were obtained shortly after infection and showed no evolutionary diversification from the index group D variant transmitted to mother 2 and consequently child 2 it is therefore likely that if sequences had been obtained late in infection from child 2, the distance between late samples of child 1 and late samples of child 2 would have been even greater. This has important implications for the identification of epidemiologically related patients. Arnold *et al.* (1995) report a mean level of diversity of 12.2% in *env* for 130 unrelated subtype B patients. It is thus quite conceivable that interpatient diversity within two related individuals separated by a limited number of transmission events would soon exceed the mean interpatient diversity for unrelated individuals.

There is no evidence that the restriction of variability at transmission constrains the extent of nucleotide sequence divergence, an inference also made by Kuiken *et al* (1993) from comparison of sequences obtained from acute seroconverters in Amsterdam 5 years apart. The single most important factor determining the extent of nucleotide sequence divergence in HIV-1 appears to be time.

Genetic Relationships with other Edinburgh HIV-1 sequences

Phylogenetic analyses in which *gag* sequences of the transmission set were placed within a background of all available Edinburgh intravenous drug users (IDUs), gay men and haemophiliacs, confirmed that the transmission set fell within the IDU clade as expected from the available clinical data. However, early sequences of the male index (seroconversion A type sequences, Figure 4), showed a closer relationship to a number of Edinburgh IDU

sequences, than to sequences obtained from the index later in the course of infection or indeed sequences obtained from the female transmission contacts. It is clear that the early stages of HIV-1 infection within the Edinburgh IDU population were characterised by very limited sequence variability and most probably the three Edinburgh IDU patients which were most closely related to the male index were separated from the male index by very few transmissions possibly forming members of the same shooting gallery. Nonetheless, the closer association of sequences of the male index with sequences of other IDU patients than with sequences obtained from later male index samples was unexpected. Previous molecular studies have shown that *env* sequences obtained early and late in infection are similar enough, even at 10 years post transmission, to cluster together within reconstructed evolutionary trees (Kuiken *et al.*, 1996). Phylogenetic analysis techniques have also shown that previously unknown contact networks can be reliably identified from within a background of unrelated patients by high bootstrap values within the reconstructed trees (Leigh Brown *et al.*, submitted).

ACKNOWLEDGEMENTS

We would like to thank Marian Aldhous for the identification of patients and samples. This work was supported by the Medical Research Council AIDS Directed Programme.

REFERENCES

Ahmad, N., B. M. Baroudy, R. C. Baker, and C. Chappey. 1995. Genetic analysis of human immunodeficiency virus type 1 V3 region isolates from mothers and infants after perinatal transmission. J. Virol. 69:1001-1012.

Albert, J., J. Wahlberg, and M. Uhlén. 1993. Forensic evidence by DNA sequencing. Science 361:595-596.

Arnold, C., P. Balfe, and J. P. Clewey. 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.

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.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein, J. 1993. PHYLIP manual version 3.52c. Berkeley University Herbarium, University of California, Berkeley.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science 155:279-284.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. 20:406-416.

Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Applic. Biosci. 8:189-191.

Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam, and A. J. Leigh 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. Aci. USA 89:4835-4839.

Holmes, E. C., L. Q. Zhang, P. Simmonds, A. S. Rogers, and A. J. Leigh Brown. 1993. Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. J. Infect. Dis. 167:1411-1414.

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.

Jaffe, H. W., J. M. McCurdy, M. L. Kalish, T. Liberti, G. Metellus, B. H. Bowman, A. R. Neasman, and J. J. White. 1994. Lack of transmission of human immunodeficiency virus in the practice of a dentist with AIDS. Ann. Intern. Med. 121:855-859.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules Chapter 11, *In* Munro, H. N. ed. Mammalian protein metabolism, Academic Press, New York, pp. 21-132.

Kishino, H., and M. Hasegawa, 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order of the Hominoidea. J. Mol. Evol. 4:406-425.

Korber, B., and G. Myers. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retroviruses 8:1549-1560.

Kuiken, C. L., G. Zwart, E. Baan, R. A. Couthino, J. A. R. van den Hoek, and J. Goudsmit, 1993. Increasing antigenic and genetic diversity of the V3 variable region of the V3 variable domain of the human immunodeficiency virus envelope protein in the course of the AIDS

epidemic. Proc. Natl. Acad. Sci. USA. **90**:9061-9065.

Kuiken, C. L., V. V. Lukashov, E. Baan, J. Dekker, J. A. M. Lenissen, and J. Goudsmit. 1996. Evidence for limited within-person evolution of the V3 domain of the HIV-1 envelope in the Amsterdam population. AIDS **10**: 31-37.

Kumar, S., K. Tamura, and M. Nei. 1993. MEGA: Molecular Evolutionary Genetics Analysis, version 1.0. The Pennsylvania State University, University Park, PA 16802, USA.

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 1 in infants by vertical transmission. J. Clin. Invest. **93**:380-390.

Leigh Brown, A. J. and P. Simmonds. 1995. Analysis of HIV sequence variation. Chapter 11, *In* Karn, J. ed. HIV - A Practical Approach. Oxford University Press, pp 161-188.

Leigh Brown, A. J. and A. Cleland. 1996. Independent evolution of the *env* and *pol* genes of HIV-1 during zidovudine therapy. AIDS **10**:1067-1073.

Leigh Brown, A. J., D. Lobidel, C. M. Wade, S. Rebus, A. N. Phillips, R. P. Brettle, A. J. France, C. S. Leen, J. McMenamim, A. McMillan, R. D. Maw, F. Mulcahy, J. R. Robertson, K. N. Sankar, G. Scott, R. Wyld, and J. F. Peutherer. The molecular epidemiology of human immunodeficiency virus type 1 in six cities in Britain and Ireland. Submitted for publication.

Leigh Brown, A. J. 1997. Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA **in press.**

Leitner, T. D. Escanilla, C. Franzen, M. Uhlen, and J. Albert. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. USA **in press.**

McNearney, T., P. Westervelt, B. J. Thielan, D. B. Trowbridge, J. Garcia, R. Whittier, and L. Ratner L. 1990. Limited sequence heterogeneity among biologically distinct human

immunodeficiency virus type 1 isolates from individuals involved in a clustered infectious outbreak. Proc. Natl. Acad. Sci. USA **87**:1917-1921.

Mulder-Kampinga, G. A., C. Kuiken, 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**:1747-1756.

Mulder-Kampinga, G. A., A. Simonon, C. L. Kuiken, J. Dekker, H. J. Scherpbier, P. Van De Perre, K. Boer, and J. Goudsmit. 1995. Similarity in *env* and *gag* genes between genomic RNAs of human immunodeficiency virus type 1 (HIV-1) from mother and infant is unrelated to the time of HIV-1 RNA positivity in the child. J. Virol. **69**:2285-2296.

Nowak, M. A., R. M. Anderson, A. R. McLean, T. F. W. Wolfs, J. Goudsmit, and R. M. May. 1991. Antigenic diversity thresholds and the development of AIDS. Science **254**:963-969.

Ou, C.-Y., C. A. Ciesieleski, G. Myers, C. I. Bandea, C.-C. Luo, B. T. M. 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, H. W. Jaffe, the Laboratory Investigation Group, and the Epidemiological Investigation Group. 1992. Molecular epidemiology of HIV transmission in a dental practice. Science **256**: 1165-1171.

Poss, M., H. L. Martin, J. K. Kreiss, L. Granville, B. Chohan, P. Nyange, K. Mandaliya, and J. Overbaugh. 1995. Diversity in virus populations from genital secretions and peripheral blood from women recently infected with human immunodeficiency virus type 1. J. Virol. **69**:8118-8122.

Robertson, J. R., A. B. V. Bucknall, P. D. Welsby, J. J. K. Roberts, J. M. Inglis, J. F. Peutherer, and R. P. Brettle. 1986. Epidemic of AIDS related virus (HTLV-III/LAV) among intravenous drug abusers. BMJ. **292**:527-529.

Saitou, N., and M. Nei, 1987. The neighbor-joining method: a new method for reconstructing evolutionary trees. Mol. Biol Evol. **4**:406-425.

Scarlatti, G., T. Leitner, E. Halapi, J. Wahlberg, P. Marchisio, M. A. Clerici-Schoeller, H. Wigzell, E. M. Fenyö, J. Albert, M. Uhlén, 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**:1721-1725.

Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown. 1990. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. J. Virol. **64**:864-872.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet, 1994. The genetic data environment and expandable GUI for multiple sequence analysis. Comput. Appl. Biosci. **10**:671-675.

Staden, R., 1993. Staden Package Update. Genome News 13:12-13.

Van't Wout, A. B., N. A. Kootstra, G. A. Mulder-Kampinga, N. A. Albrecht van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker. 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral and vertical transmission. J. Clin. Invest. **94**:2060-2067.

Wade, C. M., D. Lobidel, and A. J. Leigh Brown. Analysis of Human Immunodeficiency Virus Type 1 *env* and *gag* Sequence Variants Derived from a Mother and Two Vertically Infected Children Provides Evidence for the Transmission of Multiple Sequence Variants. Submitted for publication.

Wike, C. M., B. T. M. Korber, M. R. Daniels, C. Hutto, J. Muñoz, M. Furtado, W. Parks, A. Saah, M. Bulterys, J.-B. Kurawige, and S. M. Wolinsky. 1992. HIV-1 sequence variation between isolates from mother-infant transmission pairs. AIDS Res. Hum. Retroviruses **8**:1297-1300.

Wolfs, T. F. W., G. Zwart, M. Bakker, and J. Goudsmit. 1992. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. Virology **189**:103-110.

Wolinsky, S. M., C. M Wike, B. T. M. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Muñoz. 1992. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science **255**:1134-1137.

Wolinsky, S. M., B. T. M. Korber, A. U. Neumann, M. Daniels, K. J. Kuntsman, A. J. Whetsell, M. R. Furtado, Y. Cao, D. D. Ho, J. T. Safrit, and R. A. Koup. 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science **272**:537-542.

Zhang, L. Q., P. Simmonds, C. A. Ludlam, and A. J. Leigh Brown. 1991. Detection, quantification and sequencing of HIV-1 from the plasma of seropositive individuals and from factor VIII concentrates. AIDS **5**:675-681.

Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. **67**:3345-3356.

Zhu, T., H. Mo, N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho. 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. Science **261**:1179-1181.

Zhu, T., N. Wang, A. Carr, S. Wolinsky, and D. D. Ho., 1995. Evidence for coinfection by multiple strains of human immunodeficiency virus type 1 subtype B in an acute seroconvertor. J. Virol. **69**:1324-1327.

Table 1.

Times of sampling, clinical data, $CD4^+$ cell counts and the number of molecules sequenced per time point from the patients studied. ND, not determined.

[a] Sample names of the male index reflect the year of sampling, year 0 (seroconversion) to year 9. For mother-child pair 1, sample names reflect years following the estimated date of seroconversion for mother 1 and years following birth for child 1. For mother-child pair 2, sample names reflect months following seroconversion for mother 2 and months following birth for child 2.

[b] Time course provides details of the sampling time points relative to the dates of seroconversion and delivery of the other patients within the transmission set. For the male index, the time is given relative to the seroconversion dates of the mothers. For the mothers, the time is given relative to the delivery date of each child. - indicates sampling dates before delivery, + indicates sampling dates following delivery. For the children, the age of the child at the time of sampling is given.

[c] The average normal $CD4^+$ cell count for an adult is 1000 cells per mm^3. For children under the age of 11 months, the normal $CD4^+$ cell count is between 1700 and 2880 cells per mm^3 (mean 2200). For children between 1 and 6 years of age, the average $CD4^+$ cell count is between 1000 and 1800 cells per mm^3.

Patient	Sample Name[a]	Sample Date	Time Course[b]		Population	CD4[+] Cell Counts (cells/mm^3)[c]	Clinical Information	CDC Classification	*Env* Sequences	*Gag* Sequences
			Mother 1	Mother 2						
Male Index	0R	.12.84	SC M1	-9 yrs	Plasma	ND	Asymptomatic		0	3
	5R	30.08.89	+5 yrs	-3 yrs	Plasma	557	"		2	3
	6D	28.09.90	+6 yrs	-2 yrs	PBMC	580	"		12	11
	7D	28.11.91	+7 yrs	-1 yrs	PBMC	360	"		9	10
	7R	"	"	"	Plasma	"	"		8	4
	8Ra	21.10.92	+8 yrs	-4 mths	Plasma	380	"		7	6
	8Rb	12.12.92	+8 yrs	-2 mths	Plasma	624	"		3	3
	9R	26.02.93	+9 yrs	+1 mth	Plasma	244	"		12	3
Mother 1	2R	.05.86	+2 yrs		Plasma	ND	Asymptomatic		1	0
Child 1	3R	02.10.87	3 yrs		Plasma	1690	Lymphadenopathy, recurrent respiratory infections, diarrhoea, hepatosplenomegaly	A1	0	0
	7D	02.08.91	7 yrs		PBMC	200	Cardiac failure, cardiomyopathy, severe wasting	B3	10	10
Mother 2	2D	10.02.93		-7 mths	PBMC	ND	Asymptomatic		2	2
	2R	"		"	Plasma	"	"		20	7
	3Ra	08.03.93		-6 mths	Plasma	565	"		9	9
	3Rb	17.03.93		-6 mths	Plasma	823	"		0	0
	5D	12.05.93		-4 mths	PBMC	ND	"		6	6
	5R	"		"	Plasma	"	"		0	0
	7D	07.07.93		-2 mths	PBMC	ND	"		10	3
	7R	"		"	Plasma	"	"		2	2
	8D	06.08.93		-1 mth	PBMC	480	"		1	2
	8R	"		"	Plasma	"	"		3	0
	9D	03.09.93		DEL C2	PBMC	1230	"		12	10
	9R	"		"	Plasma	"	"		0	0
	14D	18.02.94		+5 mths	PBMC	700	"		0	0
	14R	"			Plasma	"	"		0	2
Child 2	2R	26.11.93		2mths	Plasma	3170	Lymphadenopathy, candidal nappy rash	A1	12	12
	4R	14.01.94		4 mths	Plasma	2210	Viral illness with irritability	A1	6	14
	5D	25.02.94		5 mths	PBMC	2810	Recent RUL pneumonia, nappy rash	B1	12	11
	8D	20.05.94		8 mths	PBMC	2590	Lymphadenopathy, hepatosplenomegaly	B1	12	12
	8R	"		"	Plasma	"	"	"	2	4
	11D	19.08.94		11 mths	PBMC	1890	Developmental decay	B1	0	8
	11R	"		"	Plasma	"	"	"	5	0
	12R	28.09.94		12 mths	Plasma	1030	LUL emphysema	B1	7	12

Table 2.

Within-sample nucleotide sequence diversity (a) and cumulative sequence divergence (b) within the V3 loop and flanking regions of *env* and the p17 region of *gag* of the index patient. Cumulative distances are calculated as the mean pairwise distance between a given sample and the earliest group A plasma RNA sequence of the index (*env*, year 5; *gag*, year 0). ML indicates a generalised two-parameter (maximum likelihood) distance estimate (program DNADIST; Felsenstein 1993), d_s, the proportion of synonymous nucleotide substitutions per synonymous site, and d_n, the proportion of nonsynonymous nucleotide substitutions per nonsynonymous site. d_s and d_n values were calculated using the Jukes-Cantor one-parameter model as implemented in MEGA version 1.01 (Kumar *et al.*, 1993). n indicates the number of sequences within each sample. Diversity and distance estimates for year 8 reflect the mean of the diversities and distances for year 8 samples a and b.

a.

Patient	Population	Year of Sampling	V3 Loop and Flanking regions of *env*					p17 Region of *gag*				
			n	ML	ds	dn	ds/dn	n	ML	ds	dn	ds/dn
Male Index	Plasma	0	0	-	-	-	-	3	0.0017	0.0078	0.0000	-
		5	2	0.0133	0.0249	0.0104	2.3942	3	0.0156	0.0485	0.0066	7.3485
		7	8	0.0363	0.0218	0.0411	0.5304	4	0.0184	0.0550	0.0086	6.3953
		8	10	0.0260	0.0286	0.0322	0.8882	9	0.0198	0.0500	0.0117	4.2735
		9	12	0.0452	0.0360	0.0191	1.8848	3	0.0260	0.0752	0.0132	5.6970
	PBMC	6	12	0.0298	0.0182	0.0329	0.5532	11	0.0155	0.0479	0.0068	7.0441
		7	9	0.0360	0.0380	0.0363	1.0468	10	0.0218	0.0611	0.0113	5.4071

b.

Patient	Population	Year of Sampling	V3 loop and flanking regions of *env*					p17 Region of *gag*				
			n	ML	ds	dn	ds/dn	n	ML	ds	dn	ds/dn
Male Index	Plasma	0	0	-	-	-	-	3	0.0009	0.0039	0.0000	-
		5	2	0.0067	0.0125	0.0052	2.4038	3	0.0164	0.0485	0.0077	6.2987
		7	8	0.0391	0.0253	0.0436	0.5803	4	0.0156	0.0333	0.0108	3.0833
		8	10	0.0422	0.0402	0.0439	0.9157	9	0.0181	0.0359	0.0132	2.7197
		9	12	0.0435	0.0395	0.0442	0.8937	3	0.0208	0.0449	0.0144	3.1180
	PBMC	6	12	0.0280	0.0274	0.0284	0.9648	11	0.0132	0.0296	0.0085	3.4823
		7	9	0.0411	0.0355	0.0434	0.8180	10	0.0171	0.0400	0.0109	3.6697

Table 3.

Inter patient distances and intra patient diversities for the infected mothers and children within the V3 loop and flanking regions of *env* (a) and the p17 region of *gag* (b). Sequences were unavailable for mother 1 in *gag*. Distances were calculated using the generalised two-parameter (maximum likelihood) distance estimate (program DNADIST; Felsenstein 1993). Intra patient distances are shown in bold on the diagonal.

a.

	Mother 1	Child 1	Mother 2	Child 2
Mother 1	**0.0000**			
Child 1	0.0978	**0.0421**		
Mother 2	0.0743	0.1172	**0.0027**	
Child 2	0.0760	0.1187	0.0057	**0.0078**

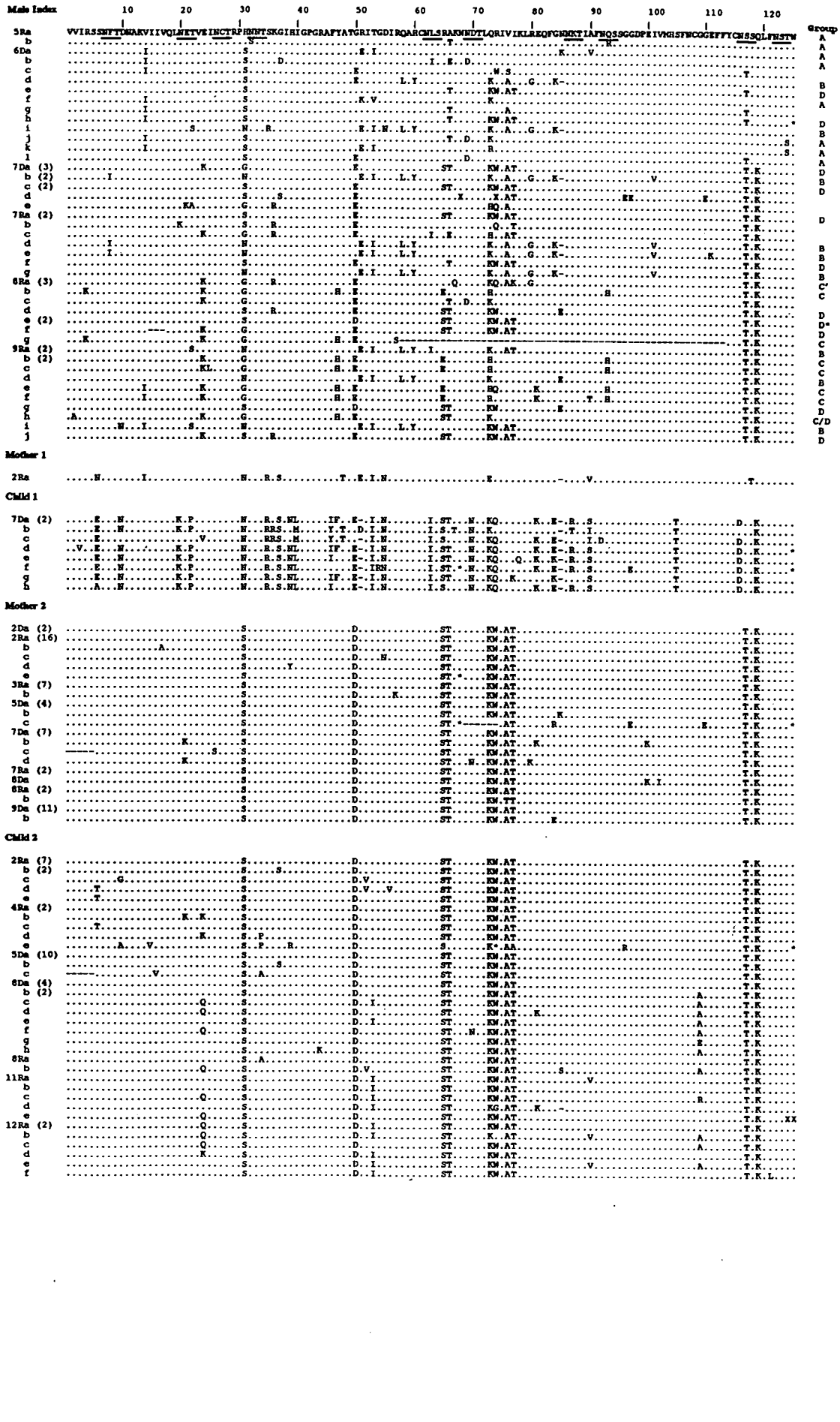
b.

	Child 1	Mother 2	Child 2
Child 1	**0.0144**		
Mother 2	0.0306	**0.0030**	
Child 2	0.0320	0.0047	**0.0053**

Figure 1.

Multiple alignments of deduced amino acid sequences of the V3 loop and flanking regions of the *env* gene (a) and the p17 region of the *gag* gene (b) for the five patients of the transmission series. Sequences are presented in order of the time of sampling and are aligned relative to the most common variant present within the first sample of the index. Amino acids identical to the first sequence are replaced by a dot, dashes represent alignment gaps, and stop codons are represented by asterisks. X indicates unreadable bases in the nucleotide sequence at this position. Sequences are labelled according to sampling time point. R denotes plasma RNA sequences, D, proviral DNA sequences. The number of identical sequences is indicated in parentheses. Putative N-linked glycosylation sites are underlined. Sequences of the index were assigned to 4 groups (A-D) on the basis of their inferred amino acid sequences and placement within the reconstructed phylogenetic tree.

a.


Male Index
10
20
30
40
50
60
70
80
90
100
110
120
Group
5Ra
6Da
7Da (3)
7Ra (2)
9Ra (2)
Mother 1
2Ra
Child 1
7Da (2)
Mother 2
2Da (2)
2Ra (16)
3Ra (7)
5Da (4)
7Da (7)
7Ra (2)
8Da
8Ra (2)
9Da (11)
Child 2
2Ra (7)
4Ra (2)
5Da (10)
8Da (4)
8Ra
11Ra
12Ra (2)

b.

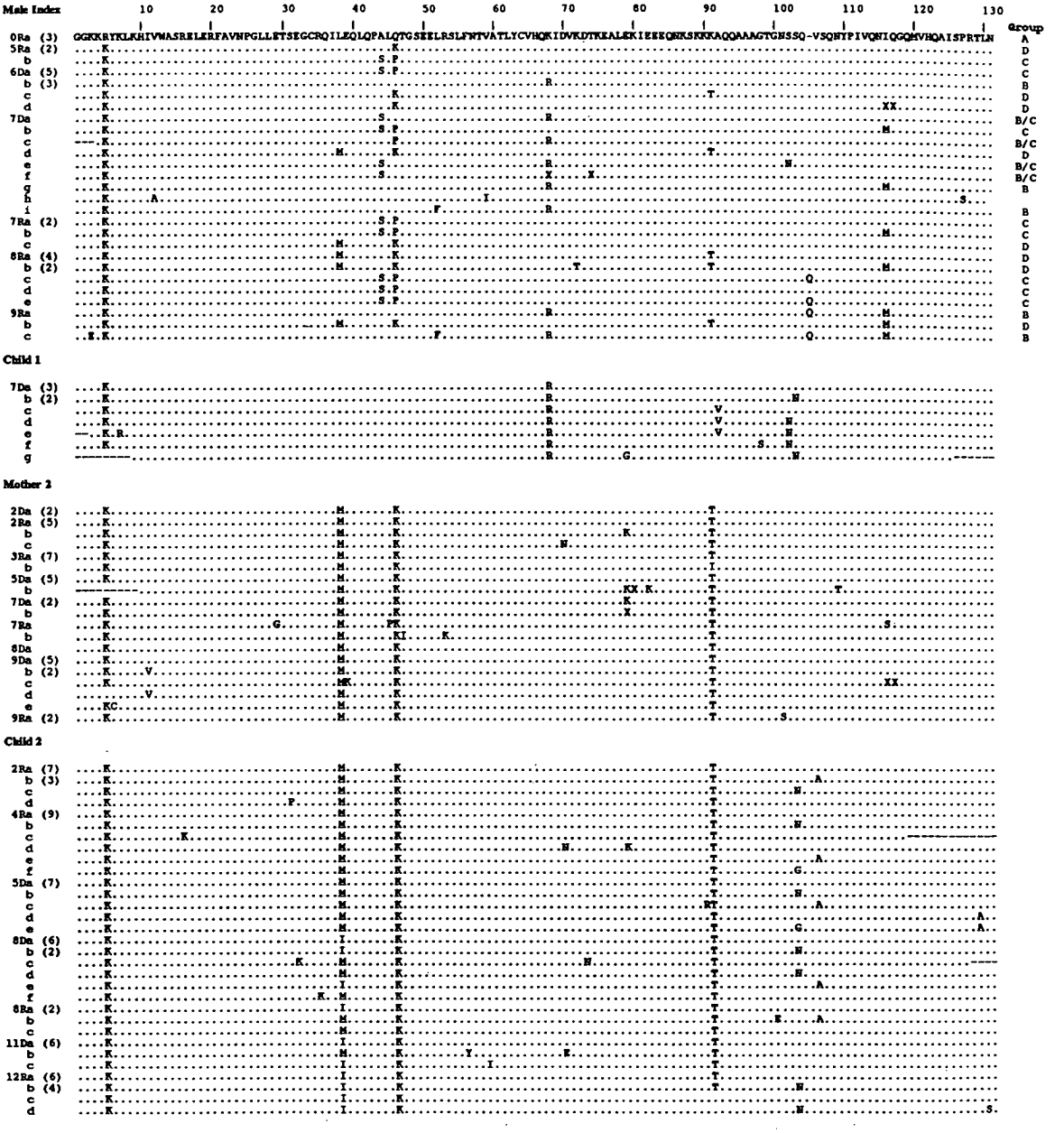

Male Index
10
20
30
40
50
60
70
80
90
100
110
120
130
Group
0Ra (3) GGKKRYKLKHIVWASRELERFAVNPGLLETSEGCRQILEQLQPALQTGSEELRSLFNTVATLYCVHQKIDVKDTKEALEKIEEEQNKSKKKAQQAAAGTGNSSQ-VSQNYPIVQNIQGQMVHQAISPRTLN A
5Ra (2) B
b C
6Da (5) C
b (3) B
c D
d D
7Da B/C
b C
c B/C
d D
e B/C
f B/C
g B
h
i B
7Ra (2) C
b C
c D
8Ra (4) D
b (2) D
c C
d C
e C
9Ra B
b D
c B
Child 1
7Da (3)
b (2)
c
d
e
f
g
Mother 2
2Da (2)
2Ra (5)
b
c
3Ra (7)
b
5Da (5)
b
7Da (2)
b
7Ra
b
8Da
9Da (5)
b (2)
c
d
e
9Ra (2)
Child 2
2Ra (7)
b (3)
c
d
4Ra (9)
b
c
d
e
f
5Da (7)
b
c
d
e
8Da (6)
b (2)
c
d
e
f
8Ra (2)
b
c
11Da (6)
b
c
12Ra (6)
b (4)
c
d

Figure 2.

Unrooted maximum likelihood phylogenetic trees reconstructed from 53 *env* (a) and 43 p17 *gag* (b) gene sequences of the index. Each sequence is labelled individually and numbered according to the year of sampling. Closed symbols denote plasma RNA sequences, open symbols, proviral DNA sequences. The scale bar corresponds to 1% nucleotide sequence divergence. Sequences of the index were assigned to 4 groups (A-D) on the basis of their inferred amino acid sequences and placement within the phylogenetic tree.

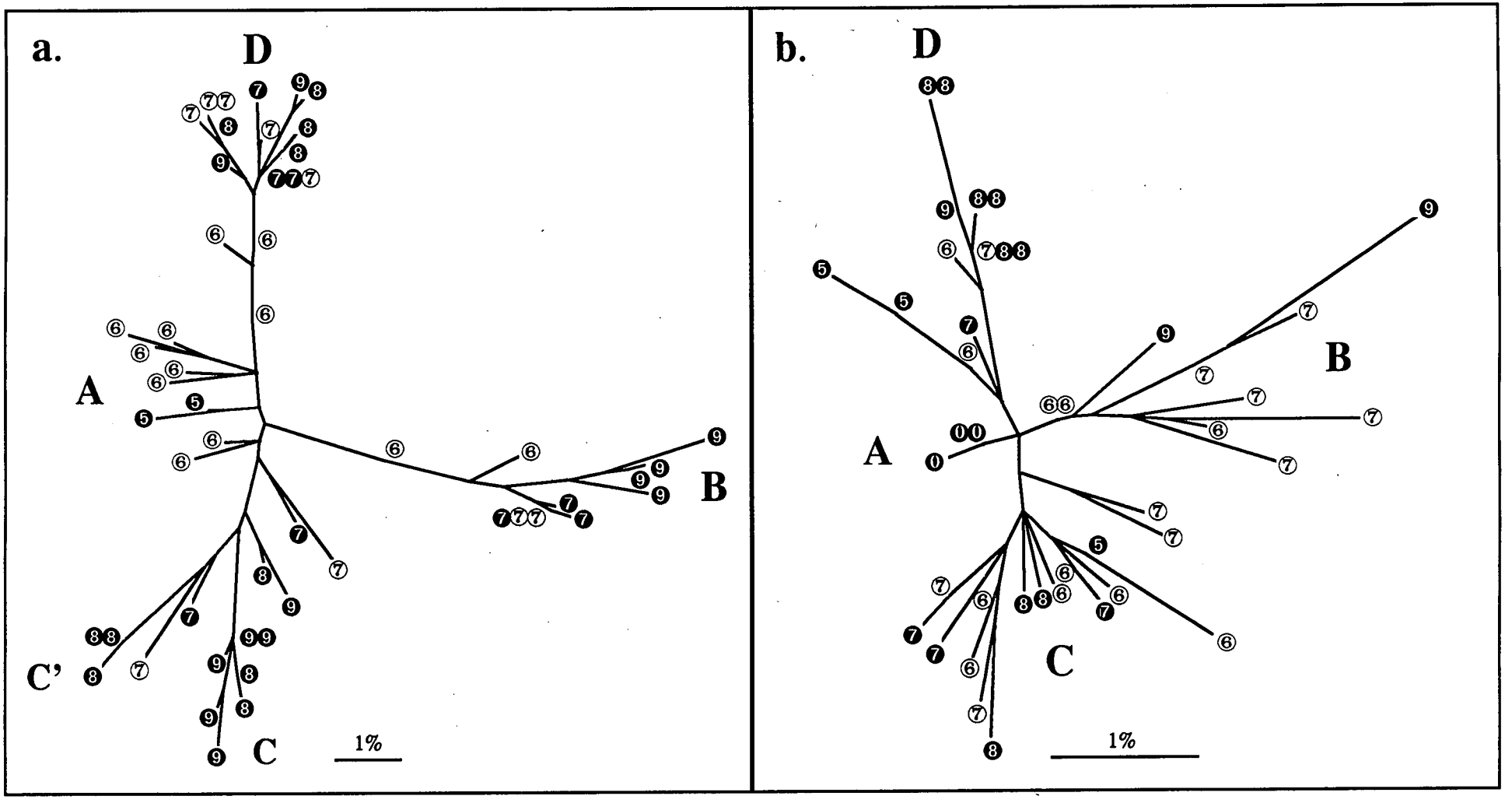

a.
D
A
B
C'
C
1%
b.
D
A
B
C
1%

Figure 3.

Unrooted neighbour-joining phylogenetic trees reconstructed from 185 *env* (a) and 169 p17 *gag* (b) gene sequences for all five patients of the transmission set. Sequences of the index are labelled individually and numbered according to year of sampling. Closed symbols denote plasma RNA sequences, open symbols, proviral DNA sequences. Sequences of the mothers and children are not labelled individually. The scale bar corresponds to 1% nucleotide sequence divergence. Bootstrap values, based on 100 bootstrap replications, are expressed as a percentage. Groups A through D denote the four groups identified within the viral population of the index.

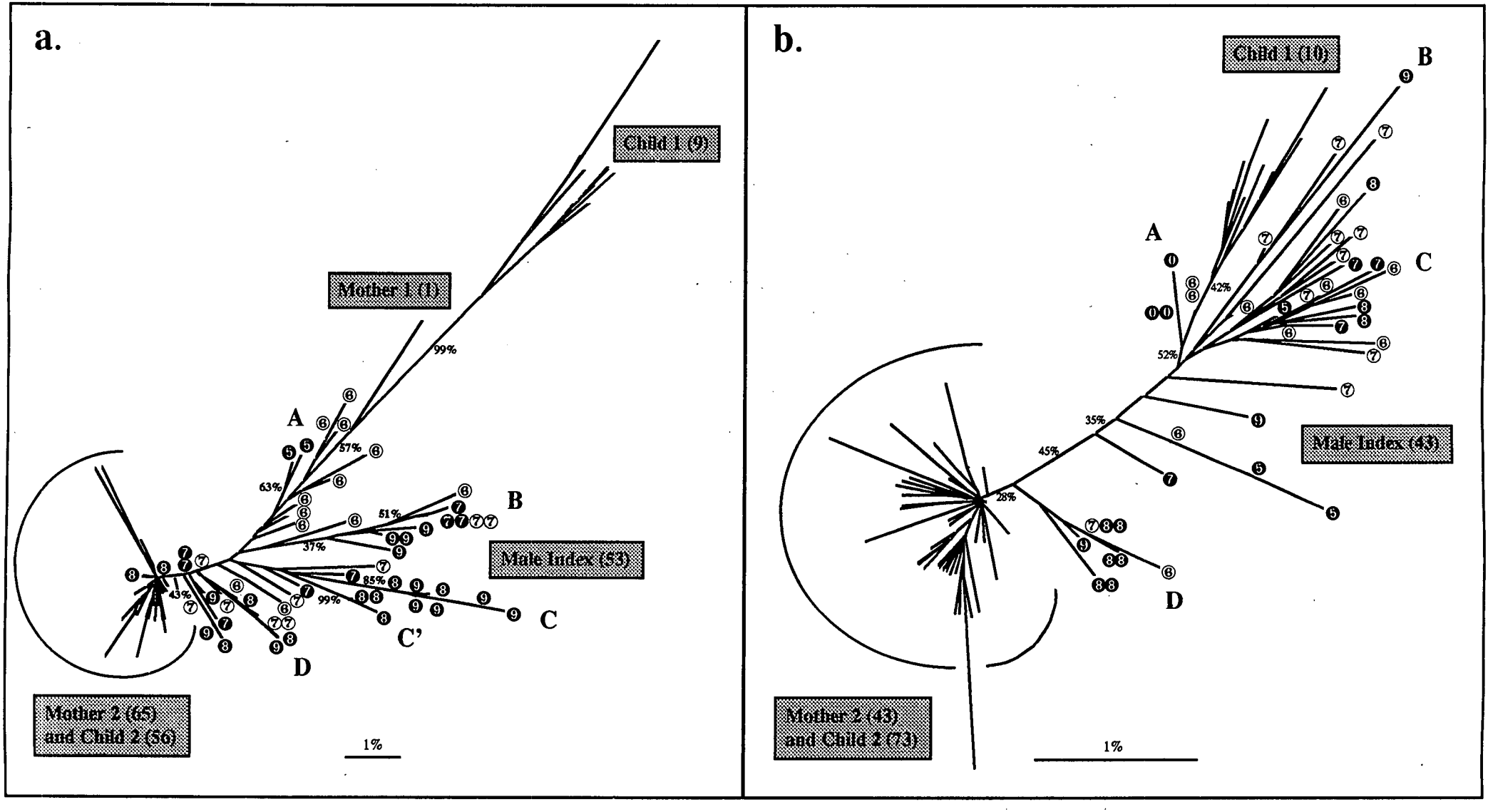
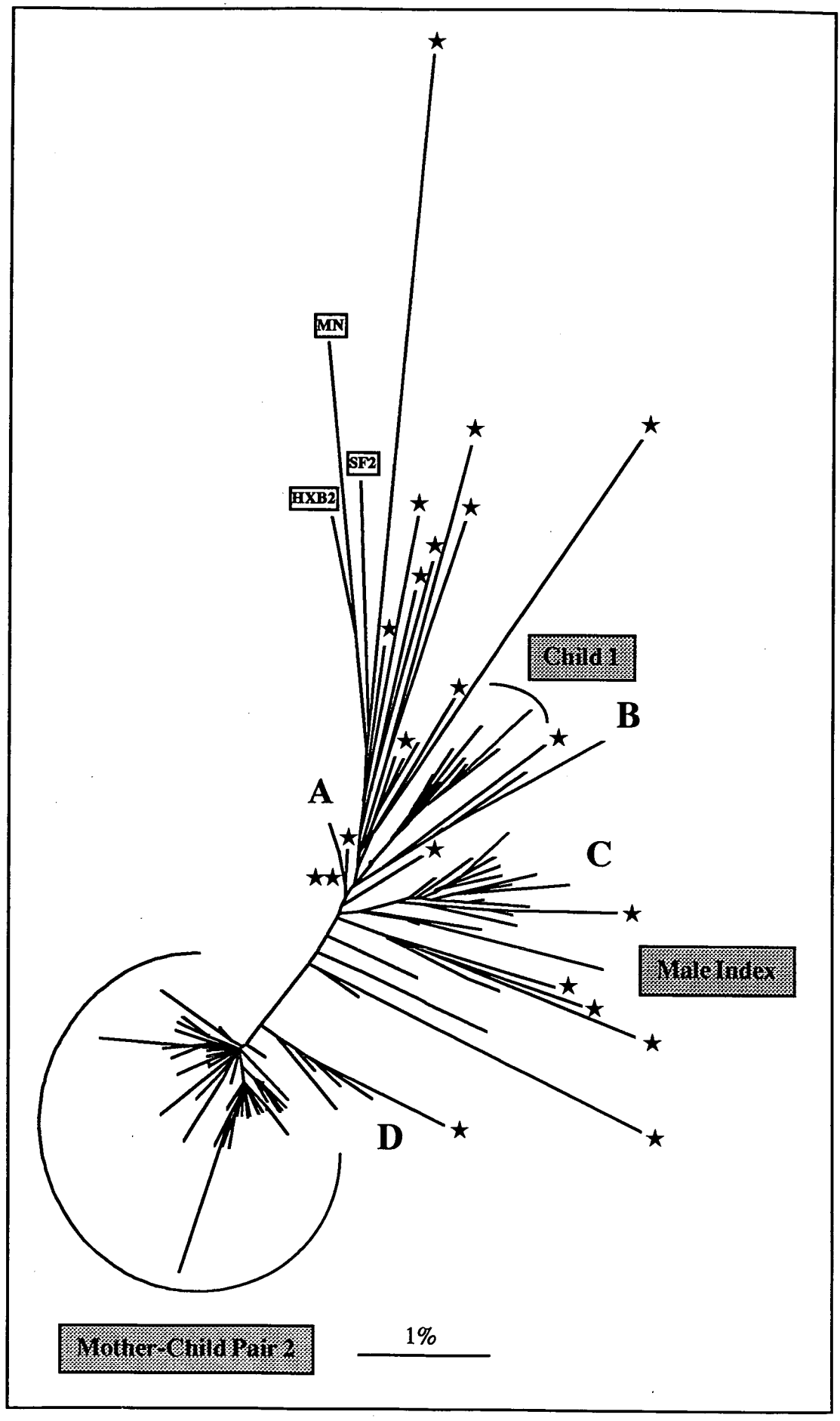

a.
Child 1 (9)
Mother 1 (1)
99%
A
57%
63%
B
51%
37%
Male Index (53)
85%
43%
99%
C
C'
D
Mother 2 (65)
and Child 2 (56)
1%
b.
Child 1 (10)
B
A
42%
C
52%
35%
45%
Male Index (43)
28%
D
Mother 2 (43)
and Child 2 (73)
1%

Figure 4.

Unrooted neighbour-joining p17 *gag* gene phylogeny showing the relationship between the 169 sequences obtained from the transmission set and sequences of 21 epidemiologically unrelated HIV-1 infected intravenous drug users from the Edinburgh cohort and 3 HIV-1 subtype B reference isolates. Sequences of the subtype B reference isolates are labelled directly on the tree, stars denote individual Edinburgh intravenous drug user sequences. Sequences obtained from the mother-child transmission set are not labelled individually. The scale bar corresponds to 1% nucleotide sequence divergence.


MN
SF2
HXB2
Child 1
B
A
C
Male Index
D
Mother-Child Pair 2
1%

Section B

Molecular Evolution of the Planktic Foraminifera

GENERAL INTRODUCTION

1. The Foraminifera

The foraminifera are a group of single-celled organisms that include both benthic species, which live on or within the first few centimetres of the ocean sediments, and planktic species, which form a large part of the marine plankton throughout the world's oceans (Figure 1). The earliest benthic species appeared in the fossil record in the Early Cambrian (540 million years (MYR) before present (BP)) (Culver, 1991), with the planktic species entering the plankton during the Mid Jurassic (170 MYR BP) (Loeblich and Tappan, 1988; Caron and Homewood, 1983).

The calcitic skeletons of planktic foraminiferal species are discarded upon reproduction, and thus constitute a substantial part of the continuous flux of biogenic particulate matter. The skeletons eventually settle down on the deep sea floor and their vast quantities accumulate to produce a pelagic sediment cover that occupies roughly one third of the entire earth's surface. The skeletons of fossil foraminifers are often well preserved and micropalaeontologists have built up precise planktic foraminiferal biostratigraphies (succession of planktic foraminiferal species through time), by studying cores of ocean sediment taken from throughout the world's oceans. The evolutionary progression of planktic foraminifera has now become established as the backbone of an important part of the geological time scale with the geochronology of the planktic foraminifera used widely in determining the age of marine sediments (Kennett and Srinivasan, 1983; Berggren *et al.*, 1985a and b; Bolli and Saunders, 1985).

(CHECK NEW REF Berggren, 1995)

2. The Importance of the Planktic Foraminifera to Palaeoceanography

The planktic foraminifera have more recently been used as an important tracer of changes in the oceanic environment and thus indirectly as indicators of past climatological and geological processes and events (Bé, 1977; Brummer and Kroon, 1988; Hemleben *et al.*, 1989). Knowledge of the distribution and ecology of foraminiferal species in present day environments (the current foraminiferal assemblages) in association with modern water mass distribution can be used to infer past marine environments based on the species found in the fossil record at a given geological time period. Planktic foraminiferal species are adapted to environments with specific water conditions and this limits their distribution to regions of the

ocean where their preferred temperature and salinity conditions are found. The planktic foraminifers living in such "faunal provinces" accumulate in the ocean sediment as fossil assemblages and thus the oceanographic conditions prevalent at that time can be inferred.

The evolution of the planktic foraminifera is thought to be environmentally controlled (Kennett, 1976; Kroon *et al.*, 1988; Williams *et al.*, 1988), being closely tied to changes in oceanic circulation patterns in concert with the ever changing global climate. Patterns of planktic foraminiferal evolution can be linked to specific environmental events in the past. For example, a high species diversity is observed in tropical oceans, where the formation of a deep stratified layer at the surface, several hundred meters thick, floats above the dense cooler water providing extra niche space for evolutionary radiation. Periods in the past which are characterised by a high species diversity can therefore be associated with similar warm stable climates. High numbers of planktic foraminiferal specimens with low species diversity may reflect regions with sporadically high nutrient levels, such as in polar waters where large numbers of a single foraminiferal species feed on the seasonal algal blooms.

3. Limited Biological Information for the Foraminifera and the Requirement for Molecular Data in Resolving a Foraminiferal Phylogeny

The interpretation of the fossil foraminiferal data therefore requires a detailed knowledge of living planktic foraminifers. Laboratory and field experiments have however as yet only partly resolved fundamental biological issues, such as the life-span, mode of reproduction and turn-over-rates of planktic foraminifers (Hemleben *et al.*, 1989) and the effects of the chemico-physical environment on the morphology and chemical composition of the test (Kroon and Ganssen, 1989; Kroon and Darling, 1995). Current taxonomic criteria are based on the morphology of the test (Bé, 1977) but it is not known whether different morphotypes, assumed to be the same species using the current taxonomic criteria, reflect environmental conditions or are perhaps in reality different genetic species with different ecological requirements. The relationships between the planktic foraminiferal species and indeed the relationships between benthic and planktic foraminiferal species remain unclear. A common palaeontological interpretation of foraminiferal macroevolution implies the transformation of the test from a primitive membraneous wall to the agglutinated and finally to the secreted calcareous wall (Haynes, 1981). However, the macroevolutionary relationships are very difficult to trace from the fossil record with the emergence of a new test type leading to extreme changes rendering the identification of ancestral or intermediate forms difficult.

Furthermore, the phylogenetic placement of the foraminifera within eukaryote evolution remains unresolved. The application of molecular methods to the foraminifera will therefore prove greatly informative in resolving a phylogeny for this group. Work presented in this thesis examines the phylogenetic placement of the foraminifera within eukaryote evolution and also the phylogenetic relationships within the foraminifera.

4. Three Domains of Life and the Origin of Eukaryotes

The division of life into three domains, the eubacteria (which includes gram positive and gram negative bacteria), the archaebacteria (which includes methanobacteria, halobacteria and sulfobacteria) and the eukaryotes, has been proposed (Woese *et al.*, 1990), with each domain postulated to be completely distinct from the others.

Controversy surrounded the division of life into three domains, with the traditional subdivision of the prokaryotes into the bacteria and blue green algae being replaced by the Archaebacteria (which have cell membranes made of isoprenoidal ether lipids) and the Eubacteria (with membranes of fatty acid ester lipids, similar to those of the eukaryotes). Lake (1988) questioned the validity of the archaebacteria based upon an analysis of small subunit (SSU) ribosomal (r) RNA gene sequences using a new method of tree construction, "evolutionary parsimony", which suggested that the eocytes (sulfobacteria) were more closely related to eukaryotes than to other archaebacteria, and that halobacteria were closer to eubacteria than to other archaebacteria and therefore that the archaebacteria should be split into two kingdoms. Gouy and Li (1989) have however discounted Lake's hypothesis showing that the archaebacteria fall within a single branch in reconstructed trees based on analysis of both small subunit (SSU) rRNA and large subunit (LSU) rRNA sequence datasets. Distance and maximum parsimony methods of tree construction gave the same results for both the SSU and LSU rRNA sequences ie all archaebacteria fall on a single branch. However, while evolutionary parsimony showed a closer association for the eocytes and eukaryotes in analyses of the SSU, evolutionary parsimony was consistent with the other tree construction methods in placing the archaebacteria on a single branch in analyses of the LSU rRNA gene. Evolutionary parsimony uses only transversions and Guoy and Li (1989) therefore concluded that Lake's SSU rRNA evolutionary parsimony tree was most likely incorrect due to using a small number of the total available sites.

Attempts to place a root on the universal tree of life have proved problematic due to the absence of an outgroup. However, through the analysis of duplicated protein-coding genes

(elongation factors Tu and G and the α and ß subunits of ATPase) which are conserved in all organisms in the three primary lines of descent and are thus the result of a gene duplication event prior to the separation of the three kingdoms, the root of the tree is placed within the eubacterial line of descent (Iwabe, 1989). The eukaryotes are thus proposed to have evolved from an archaebacterial ancestor (Iwabe, 1989; Woese, 1990). Sequence data for a number of conserved proteins have nevertheless been shown to contradict this phylogeny (Forterre *et al.*, 1993) and Golding and Gupta (1995) have provided evidence to support a chimeric origin of the eukaryote genome based upon analysis of sequences for all available proteins and rRNA sequence data. Sequence data for the rRNA genes and genes for proteins involved in the transcription/translation process suggest that the transcription/translation unit originated as a single unit from an archaebacterial ancestor. Several other protein coding genes independent of the transcription/translation apparatus also support a gram positive-gram negative eubacterial clade/archaebacterial-eukaryotic clade. However, a number of other proteins support a significant gram positive eubacterial-archaebacterial/gram negative eubacterial-eukaryote clade.

The complete sequence of the genome of an archaebacteria, *Methanococcus jannaschii*, is now available (Bult *et al.*, 1996) and this has provided tremendous insights into the relationships between the three kingdoms (Morell, 1996). The genes that control the translation and transcription of DNA and the replication of the genome were found to show a striking similarity to the eukaryotes. The histone genes also show a strong similarity to eukaryotes which would suggest that the archaebacteria organise their DNA in a way similar to the eukaryotes. Genes involved in metabolic processes were however found to be more similar to those of the eubacteria implying that both the archaebacteria and the eubacteria derived central biochemical pathways from a common ancestor.

Molecular phylogenies based on rRNA gene sequences and sequences of a number of protein-coding genes imply that the eukaryote lineage emerged early in the history of life (Sogin, 1991; Knoll, 1992). The earliest eukaryote fossils identified are however between 1700 and 1900 million years old (Knoll, 1992). Nevertheless, the absence of eukaryote fossils in rocks prior to this date possibly reflects the lack of a good fossil record in early rocks and the low preservational potential of most lower eukaryotes (Knoll, 1992).

5. Eukaryote Evolution

Early eukaryote evolution is characterised by a series of independent protistan

branches. The amitochondriate microsporidian, tritrichomonad and diplomonad lineages represent the earliest, second and third offshoots respectively (Leipe *et al.*, 1993) and it has been proposed that these deep diverging eukaryotes should be classified as a separate kingdom, the Archezoa (Cavalier-Smith, 1989). The archezoa consists predominantly of aerotolerant anaerobes, the majority of which live parasitically within animal hosts. A number of species are however free-living and also lack mitochondria which suggests that mitochondria are primitively absent from the archezoa rather than lost by a secondary reduction due to a parasitic life style (Cavalier-Smith, 1987). The archezoa are very primitive eukaryotes and are highly divergent from other eukaryote species. Similarly to the prokaryotes, the archezoan ribosome is small (70S) and contains 16S and 23S rRNA (also comprising the 5.8S rRNA) genes as opposed to the 18S, 5.8S and 28S rRNA genes characteristic of higher eukaryotes (Vossbrink and Woese, 1986). This indicates that the 80S ribosome evolved after the origin of the eukaryotes and also after the evolution of sex and meiosis (Cavalier-Smith 1987). The branching order of the archezoan taxa is however problematic due to marked differences in GC content (discussed in paper III, Wade *et al.*, 1996). Analyses of a free living diplomonad species, *Hexamita inflata*, which has a GC content of 55%, has however provided evidence that the microsporidians represent the earliest branch of the eukaryote lineage followed by the tritrichomonads and then the diplomonads (Leipe *et al.*, 1993). Ultrastructural data is not convincing regarding the identification of the earliest archezoan branch. Although microsporidia are the least complex, lacking dictyosomes, kinetosomes or flagella, the mitotic spindle in tritrichomonads is entirely extranuclear, and thus if the tritrichomonads represent the earliest branch, microsporidia must have secondarily lost their kinetosomes and flagella (Leipe *et al.*, 1993).

Good ultrastructural and biochemical evidence exists to support the endosymbiotic theory for the origin of both the mitochondria and chloroplasts. Phylogenetic comparison of mitochondrial 16S-like rRNA sequences with 16S rRNA sequences of eubacterial phyla has localised the origin of mitochondria to the α subdivision of the purple bacteria (Yang *et al.*, 1985). Comparisons were made using plant mitochondrial rRNA sequences since these retain many eubacterial rRNA characteristics unlike the mitochondrial rRNA sequences of other eukaryote phyla which have diverged considerably over time. Phylogenetic comparison of 16S-like rRNA chloroplast sequences with 16S rRNA sequences of eubacterial phyla has similarly located the origin of chloroplasts to the cyanobacteria (Giovannoni *et al.*, 1988).

The first representatives of mitochondriate protist lineages within eukaryote evolution

are the kinetoplastids and euglenoids (Leipe *et al.*, 1993) and evidence is provided within this thesis for the early evolutionary origin of the foraminifera (paper III, Wade *et al.*, 1996). The foraminifera contain mitochondria and thus represent one of the earliest mitochondriate lineages. Eukaryotes with mitochondria are believed to have evolved from a swimming biciliated eukaryote monad with a rigid cytoskeleton, of which the closest living relatives are the archezoan metamonad protozoa of the class Anaxostylea, represented by the diplomonads (Cavalier-Smith, 1987). Diplomonads contain a number of free living species and are thus good candidates for the host that engulfed purple bacteria by phagocytosis to form the first mitochondria.

Aerobic metabolism could not have occurred in the oxygen-poor environments hypothesised for the early earth and would not have been possible until the partial pressure of oxygen (pO_2) in the atmosphere increased to 1 to 2% of the present atmospheric level (PAL) following the onset of cyanobacterial photosynthesis 2800 to 2400 MYR BP (Knoll, 1992). Knoll (1992) suggests that the onset of an aerobic atmosphere is the critical factor for the evolution of mitochondrial symbioses; oxygen is toxic to cells without mitochondria and thus the acquisition of mitochondria would have provided a great evolutionary advantage. Knoll (1992) suggests that the anaerobic nature of the early diverging amitochondriate eukaryote protist lineages is consistent with the oxygen-poor environments proposed for the early earth. However, Cavalier-Smith (1987) suggests that there is no good reason to connect the apparently later origin of mitochondria with an aerobic atmosphere but instead suggests that the origin of mitochondrial symbioses occurred perhaps as much as 2000 million years after oxygen became widespread and that the origin of phagocytosis was the critical factor. He also suggests that the anaerobic character of many archezoa is more likely to be a secondary adaptation to competition by aerobically more efficient eukaryotes than a relic from early eukaryote evolution (Cavalier-Smith, 1987).

The timing of chloroplast acquisition within eukaryote evolution is unclear. Two theories exist. The serial endosymbiosis theory suggests that mitochondria were acquired first, with the acquisition of chloroplasts considerably later. This might be explained by the atmospheric oxygen and fixed nitrogen concentrations; although most bacterial autotrophs can fix nitrogen, nitrogen fixation does not occur in chloroplasts within eukaryote cytoplasm (see Knoll, 1992). The acquisition of protochloroplast symbionts may thus not have been favoured until atmospheric oxygen reached levels where nitrate production is higher, about 10% PAL or more (Knoll, 1992). Cavalier-Smith (1987) however, suggests mitochondria and

chloroplasts originated at about the same time, distinctly after the origin of the nucleus. In support of this, one of the earliest lineages in eukaryote evolution with chloroplasts is the euglenozoa, which also represents one of the earliest mitochondriate protist lineages. It is however possible that euglenid chloroplasts may be derived from symbiotic green algae which would imply a relatively late acquisition of photosynthesis within this early diverging eukaryote group. Both mitochondria and chloroplasts may have arisen on more than one occasion. In particular, good molecular and ultrastructural evidence exists to support multiple origins of chloroplasts within eukaryote evolution, with chloroplasts possibly arising as many as six or more times, with separate symbioses giving rise to the photosynthetic euglenids, rodophytes, chlorophytes, chromophytes, cryptophytes and photosynthetic dinoflagellates (Knoll, 1992).

The kinetoplastids and euglenoids are followed by a number of other early diverging protist lineages, including the slime moulds and amoebae, which diverge prior to the separation of the major eukaryotic groups. The major eukaryotic assemblages, the animals, fungi, plants, alveolates and stramenophiles, appear to have evolved nearly simultaneously in a rapid burst of evolution forming the "crown" of the eukaryotic tree (Knoll, 1992; Wainwright *et al.*, 1993). The "crown" diversification is dated from the fossil record at approximately 1100 to 1000 MYR BP (Knoll, 1992). The reason for such a relatively late evolutionary diversification of the major eukaryote groups remains unclear. The acquisition of chloroplasts, fundamental to many "crown" group species, may be the trigger for this evolutionary diversification. However, chloroplasts may well have been acquired well before the "crown" diversification, which would suggest that other critical factors could have been inhibiting diversification following chloroplast acquisition. One possibility is that the explosive diversification observed within the "crown" occurred long after the acquisition of chloroplasts and the evolution of multicellularity and was perhaps due to the establishment of sexual population structures (see Knoll, 1992). Alternatively, the "crown" group diversification observed from the fossil record may reflect rapid diversification within a few easily fossilised groups rather than among all branches in the crown (Knoll, 1992). Environmental conditions again may have played an important part in the "crown" diversification with an increase in pO_2 to 15% PAL at approximately 1900 MYR BP and with major biogeochemical change, possibly including major increases in pO_2, shortly prior to the emergence of the macroscopic animals (Knoll, 1992). Branching orders within the "crown" do however remain uncertain due to the rapid diversification with all groups evolving

essentially simultaneously (Sogin, 1991).

6. Outline of Papers Presented in this Thesis

In this thesis, a series of papers is presented which examines the molecular evolutionary relationships of the planktic foraminifera.

Paper I (Darling *et al.*, 1996a) presents preliminary work which evaluates two extraction procedures for the isolation of foraminiferal DNA and procedures for reducing contamination from symbionts, commensals and prey organisms. Foraminifers were cultured to gametogenesis in order to minimise contamination with non foraminiferal genomes; symbionts, commensals and food particles are expelled at gametogenesis. However, despite efforts to minimise contamination at the extraction stage, preliminary amplification of extracted DNA using universal primers for the small subunit (SSU) ribosomal (r) RNA gene resulted in the generation of more than one amplification product, indicating the presence of contaminants within the sample. Sequence investigations of the amplification products are detailed in subsequent papers.

Papers II (Darling *et al.*, 1996b) and III (Wade *et al.*, 1996) present partial SSU rRNA sequences for five species of planktic foraminifera and examine the phylogenetic placement of the foraminifera within eukaryote evolution. Paper II discusses difficulties in the isolation and amplification of foraminiferal DNA due to the presence of large numbers of contaminant genomes (symbionts, commensals and food particles) and full details of the collection and culture of planktic foraminifers to gametogenesis, critical to overcoming problems of contaminants, and the molecular methods employed are presented. Amplification of SSU rDNA using universal primers resulted in the generation of two distinct amplification products indicating the presence of contaminants within the sample. Only sequences of the larger amplification product clustered within a single monophyletic group within reconstructed evolutionary trees, distinct from sequences of known symbionts, commensals and food particles. Evidence that these SSU rDNA sequences originated from foraminiferal nuclear genomes is presented and discussed in detail in paper III. Foraminiferal SSU rDNA sequences could be clearly distinguished from those of potential contaminants by unique foraminiferal specific insertions as well as considerable nucleotide distance in aligned regions. Phylogenetic analysis of the five planktic foraminiferal SSU rDNA sequences with representatives of a diverse range of eukaryote, archaebacterial and eubacterial taxa revealed that the evolutionary origin of the foraminiferal lineage precedes the rapid eukaryote diversification represented by

the "crown" of the eukaryotic tree, and probably represents one of the earliest splits among extant free living aerobic eukaryotes. The phylogenetic placement of the planktic foraminifera within the "tree of life" is examined in detail in paper III.

Paper IV (Darling *et al.*, 1997) presents SSU rDNA sequences for an additional seven planktic foraminiferal species and examines the evolutionary relationships between the planktic foraminiferal genera, comprising representatives of both non spinose and spinose groups, and the relationships of the planktic lineages to SSU rDNA sequences of the benthic suborders available to date. The phylogenetic analysis showed that the planktic foraminifers were polyphyletic in origin, not evolving solely from a single "globigerinid like" lineage in the Mid-Jurassic, but derived from at least two ancestral benthic lines, with the benthic ancestor of one species, *Neogloboquadrina dutertrei*, possibly entering the plankton later than the Mid-Jurassic. The divergences of the planktic spinose species generally support recent phylogenies based on the fossil record, which infer a radiation from a globigerinid common ancestor in the Mid to Late Oligocene. The paper also examines the relationship between closely related morphotypes of two species, *Globigerinella siphonifera* Type I and II and *Globigerinoides ruber* "pink" and "white". Morphotypes of *Ge. siphonifera* can not be distinguished using traditional palaeontological techniques yet exhibit extreme genetic distances, indicative of species level variations within the SSU rDNA. This has profound implications for tracing fossil lineages and for the estimation of molecular evolutionary rates based on the fossil record. In addition, Caribbean and Western Pacific specimens of three species, *Globigerinoides sacculifer*, *Orbulina universa* and *Globigerinella siphonifera* Type II, were compared. Caribbean and Western Pacific specimens of both *Globigerinoides sacculifer* and *Orbulina universa* showed complete sequence identity within both the conserved core regions and expansion segments whilst *Globigerinella siphonifera* Type II Caribbean and Western Pacific specimens showed one polymorphic position in the conserved regions and substantial variability in the variable regions.

Finally, paper V (Darling *et al.*, 1996c) provides a summary of the work carried out thus far examining foraminiferal evolution.

7. General Discussion

The isolation and amplification of foraminiferal DNA has proved extremely complex due to the association of many non-foraminiferal nuclei, including symbionts, commensals, prey particles and adherent organisms, with the foraminiferan. In papers presented in this

thesis, partial SSU rDNA sequences are presented for a number of planktic foraminiferal species. It is contended that these sequences originated from foraminiferal nuclear genomes and evidence is provided in support of this conclusion. Briefly, all foraminiferal sequences could be distinguished from other eukaryote sequences by the inclusion of large insertions at common regions within the SSU rRNA and clustered within a monophyletic group distinct from sequences of possible symbionts, commensals, prey particles and adherent organisms. The phylogenetic placement of the foraminiferal SSU rDNA sequences prior to the "crown" diversification was also in agreement with analyses of partial LSU sequences of both benthic and planktic foraminifera (Pawlowski *et al.*, 1994; Merle *et al.*, 1994) and an identical phylogenetic placement for the benthic foraminifera has recently been inferred from analyses of the complete SSU gene (Pawlowski *et al.*, 1996).

An alternative, although unlikely explanation for the monophyly of the SSU sequences obtained from the foraminifera is that the sequences derived from a previously unknown eukaryotic organism symbiotic with both benthic and planktic foraminifers. Nevertheless, the absence of an alternative candidate sequence meeting these criteria renders this hypothesis unlikely leaving the foraminiferal origin of the SSU sequences as the most likely explanation of the data.

In situ hybridization would address the issue of the origin of the amplified DNA directly and would be a useful approach to finally resolve the origin of the sequences. However, it is worth noting that a previous attempt to use in situ hybridisation to locate the origin of possible foraminiferal SSU rDNA sequences (Wray *et al.*, 1995) proved ambiguous since the localisation of the foraminiferal nuclei was not cytogenetically demonstrated and the possibility therefore remained that what was stained was a symbiont, parasite, or food organism (Pawlowski *et al.*, 1996).

The early evolutionary origin of the foraminiferal lineage inferred from analyses of the SSU rRNA gene is of importance in the study of early eukaryote evolution and in particular in the evolution of the mitochondria. In the analyses presented here, a phylogenetic placement of the foraminiferal lineage prior to the mitochondriate kinetoplastids and euglenoids is inferred, with the foraminifera thus representing the earliest mitochondriate protist lineage. The placement of lineages in the middle of the tree is however unclear although there is significant support for the placement of the foraminiferal lineage prior to the "crown" diversification, approximately 1000 to 1100 Myr ago. Recent analyses of full length benthic foraminiferal SSU sequences have resolved an identical placement for the

benthic foraminifera with stronger support provided for the placement of the foraminifera prior to the kinetoplastids and euglenoids based upon analysis of the full length gene (Pawlowski *et al.*, 1996).

The earliest foraminifera to appear in the fossil record were agglutinated-walled benthic species which appeared in the Lower Cambrian, approximately 560 Myr ago (Culver, 1991). However, the placement of the foraminifera prior to the "crown" diversification presented here suggests that the origin of the foraminifera predates the Lower Cambrian. Agglutinated benthic foraminifers are believed to have evolved from ancestral membraneous-walled forms similar to *Allogromia* (Tappan and Loeblich, 1988) and the molecular data may therefore suggest that some unfossilised membraneous-walled forms may have existed long before the first fossilised forms in the Lower Cambrian. However, an alternative explanation for the early branching of the foraminifera may lie in the high rates of rRNA evolution within the group. Artefactual early positions due to high evolutionary rates have been suggested recently for the Euglenoids and some "rhizopods" (Philippe and Adoutte, 1995). Analyses of conserved protein-coding sequences such as the elongation factor genes (Hasegawa *et al.*, 1993; Hashimoto *et al.*, 1994), will provide vital additional information in resolving the phylogenetic placement of the foraminifera.

The first planktic foraminifers appeared in low numbers in the Mid Jurassic, approximately 170 Myr ago (Caron and Homewood, 1983; Tappan and Loeblich, 1988), and are thought to have evolved from benthic foraminiferal ancestors of the suborder *Robertinina* (Loeblich and Tappan, 1974). Current interpretations of the fossil record suggest that all extant planktic foraminiferal species evolved from an ancestral "globigerine" stock. Following major faunal disruptions, a similar succession of morphotypes consistently reappear (Cifelli, 1969; Norris, 1991) and the assumption has been made that since the Mid-Jurassic, planktic foraminifers have evolved from lineages already inhabiting the plankton. The molecular data presented in this thesis contradicts this assumption and provides strong evidence indicating that the non spinose planktic foraminifera *Neogloboquadrina dutertrei* is not descended from an ancestral planktic "globigerine" lineage but is instead more closely related to benthic groups. This implies an entry into the plankton independently of the other spinose planktic species studied. Such re-seeding from the benthos may possibly account for the appearance of other planktic species into the fossil record for which ancestral forms cannot be identified. New, as yet unpublished analyses have also indicated that an additional non spinose species *Globigerinita glutinata* has arisen independently from the main spinose planktic group. This

therefore suggests multiple independent episodes of benthic-planktic transitions in the history of the modern planktic foraminifera contrary to previous interpretations of the fossil record.

Despite the discrepancies observed between the inferred phylogenetic placement of the non spinose planktic species based upon molecular data and the foraminiferal taxonomy based upon the fossil record, the divergences of the planktic spinose species inferred from the SSU rRNA molecular phylogeny generally support recent phylogenies based on paleontological data. The rapid branching of the planktic spinose lineages observed from the SSU phylogeny and the consequent poorly resolved structure in this area of the tree is consistent with phylogenies based on the fossil record which infer a rapid radiation from a globigerinid ancestor in the Mid to Late Oligocene. The molecular data supports the paleontological data documenting the relationship between *Orbulina universa* and *Globigerinoides sacculifer* (Bolli and Saunders, 1985) and the close relationship between *Globigerinoides ruber* and *Globigerinoides conglobatus* (Cordey, 1967).

To date, no morphometric analyses of the foraminifera have been carried out for comparison with the molecular data. Work is currently underway at Edinburgh to examine the morphological differences between the different genotypes of foraminiferal species. Several morphotypes of planktic foraminiferal species have been identified which may now prove to be genetically distinct. Genotypic variants have also been identified which do not have detectable morphological differences in their shell using current paleontological approaches. Closer examination of the shell using electron microscopy, chemical analysis and stable isotope analysis may show detectable differences which could be used to differentiate them in the sediments. Only if this proved successful for all identified genotypes could a fossil phylogeny based upon the morphology and chemical composition of the test be related to the divergences within a SSU rDNA molecular phylogeny.

8. References

Bé, A. W. H. 1977. An ecological, zoogeographic and taxonomic review of recent planktonic foraminifera. *In* Ramsay, A. T. S. ed. Oceanic Micropaleontology. Academic Press, London. Volume 1, pp. 1-100.

Berggren, W. A., D. V. Kent, and J. J. Flynn. 1985a. Paleogene geochronology and chronostratigraphy. *In* Snelling, N. J. ed. Geochronology and the Geologic Time Scale. Geol. Soc. London, Mem. **10**:141-195.

Berggren, W. A., D. V. Kent, and J. A. Van Couvering. 1985b. Neogene geochronology and chronostratigraphy. *In* Snelling, N. J. ed. Geochronology and the Geologic Time Scale. Geol. Soc. London Mem. **10**:211-260.

Bolli, H. M. and J. B. Saunders. 1985. Oligocene to Holocene low latitude planktonic foraminifera. *In* Bolli, H. M., J. B. Saunders, and K. Perch-Nielsen. eds. Plankton Stratigraphy (1). Cambridge University Press. pp 155-262.

Brummer, G. J. A., and D. Kroon. 1988. Planktonic foraminifers as tracers of ocean-climate history. Free University Press, Amsterdam. pp. 6-346.

Bult, C. J. O. White, G. J. Olson, L. Zhu, R. D. Fleischmann *et al.* 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273**:1058-1073.

Caron, M., and P. Homewood. 1983. Evolution of early planktic foraminifers. Marine Micropaleontology, 7:453-462.

Cavalier-Smith, T. 1987. Eukaryotes with no mitochondria. Nature News and Views **326**:332-333.

Cavalier-Smith, T. 1989. Archaebacteria and archezoa. Nature News and Views **339**:100-101.

Cifelli, R. J. 1969. Radiation of Cenozoic planktonic foraminifera. Syst. Zool. 18:154-168.

Cordey, W. G. 1967. The development of *Globigerinoides ruber* (d'Orbigny, 1839) from the Miocene to recent. Paleontology 10:647-659.

Culver, S. J. 1991. Early cambrian foraminifera from West Africa. Science 254:689-691.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996a. The isolation and amplification of the 18S ribosomal RNA gene from planktonic foraminifers using gametogenic specimens. *In* Whatley, R. C., and Moguilevsky, A. eds. Microfossils and Oceanic Environments. University of Wales, Aberystwyth Press. Chapter 3.1, pp. 249-259.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996b. Molecular evolution of planktic foraminifera. J. Foram. Res. 26:324-330.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996c. Reading the history of the oceans in plankton DNA. NERC News Autumn 1996: 16-17.

Darling, K. F., C. M. Wade, D. Kroon, and A. J. Leigh Brown. 1997. Planktic foraminiferal molecular evolution and their polyphyletic origins from benthic taxa. Marine Micropaleontology in press.

Forterre, P., N. Benachenhou-Lahfa, F. Confalonieri, M. Duguet, C. Elie, and B. Labedan. 1993. The nature of the last universal ancestor and the root of the tree of life, still open questions. BioSystems 28:15-32.

Giovannoni, S. J., S. Turner, G. J. Olsen, S. Barns, D. J. Lane, and N. R. Pace. 1988. Evolutionary relationships among cyanobacteria and green chloroplasts. J. Bacteriol. 170:3584-3592.

Golding, G. B., and R. S. Gupta. 1995. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. 12:1-6.

Gouy, M. and W.-H. Li. 1989. Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. Nature **339**:145-147.

Hasegawa, M., T. Hashimoto, J. Adachi, N. Iwabe, and T. Miyata. 1993. Early branchings in the evolution of eukaryotes: ancient divergence of Entamoeba that lacks mitochondria revealed by protein sequence data. J. Mol. Evol. **36**:380-388.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Ojkamoto, and M. Hasegawa. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. Mol. Biol. Evol. **11**:65-71.

Haynes, J. R. 1981. Foraminifera. Macmillan, London.

Hemleben, C., M. Spindler, O. R. and Anderson. 1989. Modern planktonic foraminifera. Springer-Verlag, New York. pp. 1-335.

Iwabe, N., K. I. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. 1989. Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA. **86**:9355-9359.

Kennett, J. P. 1976. Phenotypic Variation in some Recent and Late Cenozoic Planktonic Foraminifera. *In* Hedley, R. H. and C. G. Adams eds. Foraminifera, Volume 2. Academic Press, London. pp. 1-60.

Kennett, J. P. and M. S. Srinivasan. 1983. Neogene planktonic foraminifera. Hutchinson Ross, Stroudsburg.

Kroon, D. and K. F. Darling. 1995. Size and upwelling control of the stable isotope composition of *Neogloboquadrina dutertrei* (D'Orbigny), *Globigerinoides ruber* (D'Orbigny) and *Globigerina bulloides* D'Orbigny: Examples from the Panama Basin and the Arabian Sea. J. Foram Res. **25**:39-52.

Kroon, D and G. Ganssen. 1989. Northern Indian Upwelling cells and the stable isotope composition of living planktonic foraminifers. Deep Sea Research 36:1219-1236.

Kroon, D., P. F. Wouters, L. Moodley, G. Ganssen, and S. R. Troelstra. 1988. Phenotypic variation of *Turborotalita quinqueloba* (Natland) tests in living populations and in the Pleistocene of an eastern Mediterranean piston core. *In* Brummer, G. J. A., and D. Kroon. eds. Planktonic foraminifers as tracers of ocean climate history. Free University Press, Amsterdam pp. 131-147.

Knoll, A. H. 1992. The early evolution of eukaryotes: a geological perspective. Science 256:622-627.

Lake, J. A. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature 331:184-186.

Leipe, D. D., J. H. Gunderson, T. A. Nerad, and M. L. Sogin. 1993. Small subunit ribosomal RNA^{+} of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. Mol. Biochem. Parasit. 59:41-48.

Loeblich, A. R. Jr., and H. Tappan. 1974. Recent advances in the classification of the Foraminiferida. *In* Hedley, R. H., and C. G. Adams, eds. Foraminifera, 1. Academic Press, London. pp. 1-53.

Merle, C., M. Moullade, O. Lima, and R. Perasso. 1994. An attempt to phylogenetically characterise some planktonic foraminifers on the basis of 28S rDNA partial sequences. Comptes rendus de l'Academiedes Sciences Serie II Sciences del la Terre et des Planetes, 319:149-153.

Morell, V. 1996. Life's last domain. Science 273:1043-1045.

Norris, R. D. 1991. Biased extinction and evolutionary trends. Paleobiology, 17(4):388-399.

Pawlowski, J., I. Bolivar, J. Guiard-Maffia, and M. Gouy. 1994. Phylogenetic position of foraminifera inferred from LSU rDNA sequences. Mol. Biol. Evol. 11:929-938.

Pawlowski, J., I. Bolivar, J. Fahrni, T. Cavalier-Smith, and M. Gouy. 1996. Early origin of foraminifera suggested by SSU rRNA gene sequences. Mol. Biol. Evol. 13:445-450.

Philippe, H., and A. Adoutte. 1995. How reliable is our current view of eukaryotic phylogeny ? *In* Brugerolle, G., and J. P. Mignot, eds. Protistological Actualities, II ECP, Clemont-Ferrand, France. pp. 17-33.

Sogin, M. L. 1991. Early evolution and the origin of eukaryotes. Curr. Opin. Genet. Dev. 1:457-463.

Tappan, H., and A. R. Loeblich. 1988. Foraminiferal evolution, diversification, and extinction. J. Paleontol. 62(5):695-714.

Vossbrink, C. R., and C. R. Woese. 1986. Eukaryote ribosomes that lack a 5.8S RNA. Nature 320:287-288.

Wade, C. M., K. F. Darling, D. Kroon, and A. J. Leigh Brown. 1996. Early evolutionary origin of the planktic foraminifera inferred from SSU rDNA sequence comparisons. J. Mol. Evol. 43:672-677.

Wainwright, P. O., G. Hinkle, M. L. Sogin, and S. K. Stickel. 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. Science 260:340-342.

Williams, D. F., R. Ehrlich, H. J. Spero, N. Healy-Williams, and A. C. Gary. 1988. Shape and isotopic differences between conspecific foraminiferal morphotypes and resolution of palaeoceanographic events. Palaeogeography, Palaeoclimatology, Palaeoecology 64:153-162.

Woese, C. R. O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains archaea, bacteria and eucarya. Proc. Natl. Acad. Sci.

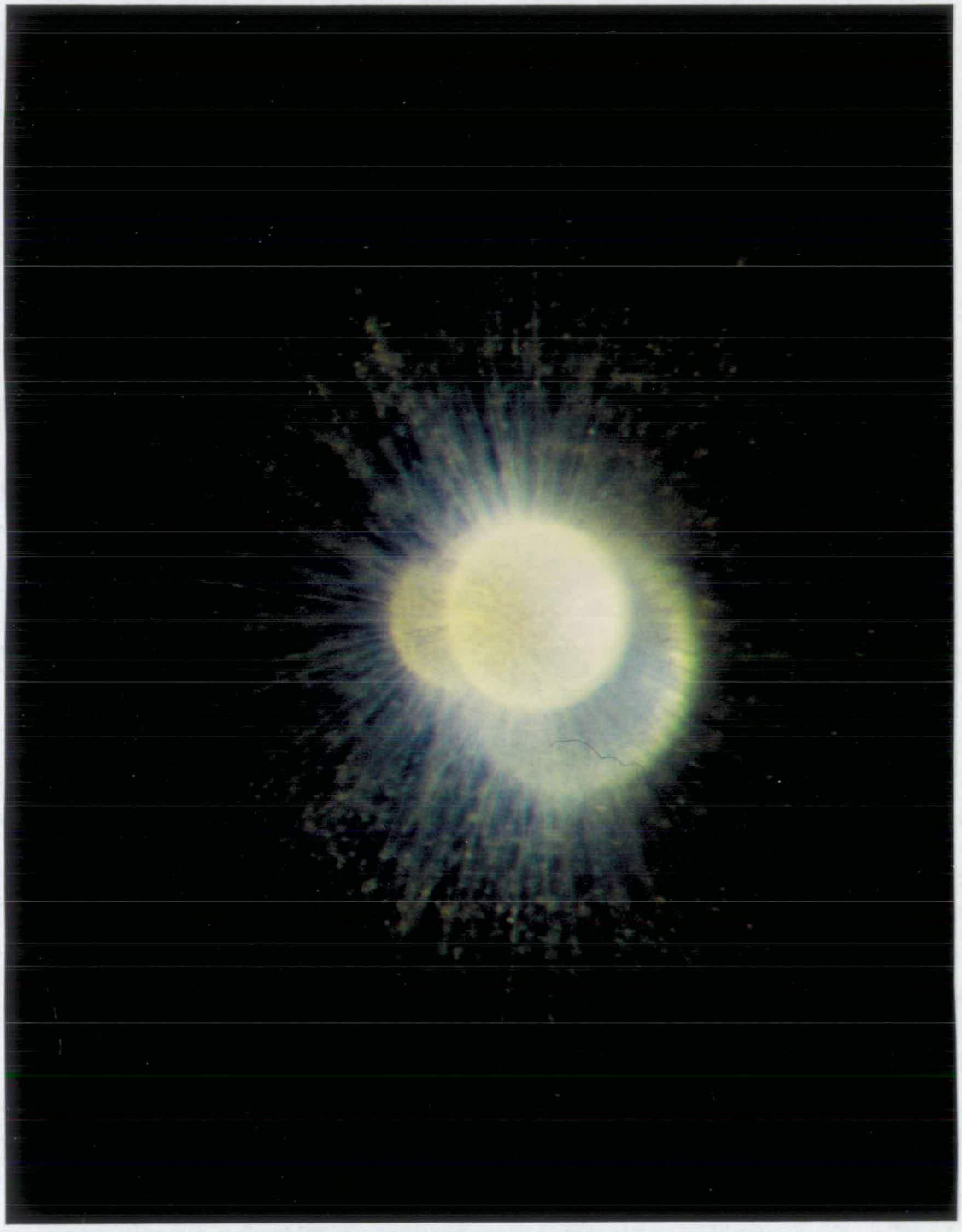
USA. 87:4576-4579.

Wray, C. G., M. R. Langer, R. DeSalle, J. J. Lee, and J. H. Lipps. 1995. Origin of the foraminifera. Proc. Natl. Acad. Sci. 92:141-145.

Yang, D., Y. Oyaizu, H. Oyaizu, G. J. Olson, and C. R. Woese. 1985. Mitochondrial origins. Proc. Natl. Acad. Sci. USA. 82:4443-4447.

Figure 1.

The planktic foraminiferal species, *Globigerinoides saculifer* (Brady).



Paper I

The Isolation and Amplification of the 18S Ribosomal RNA Gene from Planktic Foraminifers Using Gametogenic Specimens

Kate F. Darling[1*], Dick Kroon[1], Christopher M. Wade[2] and Andrew. J. Leigh Brown[2]

[1]Department of Geology and Geophysics, Grant Institute,University of Edinburgh, West Mains Road, Edinburgh, EH9 3JW.

[2]Centre for HIV Research, Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JN.

[*] Corresponding Author

In Whatley, R. C., and Moguilevsky, A. eds. (1996) Microfossils and Oceanic Environments. University of Wales, Aberystwyth Press. Chapter 3.1, pp.249-259.

ABSTRACT

Nine tropical species of planktic foraminifers were collected and cultured to gametogenesis. The gametogenic foraminifers were used to provide multiple planktic foraminiferal genomes for ribosomal (r)DNA analysis with minimal contamination from endosymbionts and food particles in order to facilitate the preferential amplification of foraminiferal rDNA. Genomic DNA was extracted from the foraminifers using the modified techniques of Langer *et al.* (1993) and Wray *et al.* (1993). Both techniques were successful in providing template DNA for the amplification of fragments of the small subunit (SSU) rRNA gene. The Langer *et al.* (1993) technique consistently produced single bands on electrophoresis gels some of which were subsequently found to contain different sequences of the SSU within the same band. The more rigorous extraction procedure of Wray *et al.*, (1993) produced multiple bands of SSU fragments, indicating greater contamination following amplification. Preliminary direct sequencing results from a variable region of the SSU shows considerable variation between sequences obtained from the different planktic species. Alignment with the SSU sequences of benthic foraminifers and other protists should indicate which sequences are most likely to be foraminiferal in origin.

INTRODUCTION

Planktic foraminifers are single-celled protists. Their calcitic tests constitute a substantial part of the continuous flux of biogenic particulate matter settling to the sea floor and have been used extensively by micropalaeontologists to determine the age of marine sediments (Kennett and Srinivasan, 1983; Berggren *et al.*, 1985a and b; Bolli and Saunders, 1985). More recently, attention has focused on Quaternary and modern sediment assemblages in order to exploit the palaeoceanographical and palaeoclimatological information contained in the fossil record (CLIMAP, 1976; Vincent and Berger, 1981).

Test morphology represents the foundation of both foraminiferal taxonomic study and phylogenetic investigation. Taxonomists can identify morphotypes of a 'species', but cannot determine whether morphological variants reflect changes in environmental conditions or are genetically different 'species' with distinct ecological requirements (Kroon *et al.*, 1988; Brummer and Kroon, 1988). Micropalaeontologists are therefore limited in their interpretation of the immense amount of palaeoceanographical information contained within the sediments.

With the advent of nucleotide sequencing and the ability to amplify large quantities of specific sequences of DNA using the polymerase chain reaction (PCR; Mullis and Faloona, 1987; Saiki *et al.*, 1988), it has now become practical to investigate foraminiferal DNA sequences. The highly conserved, SSU rRNA gene (Figure 1) has been used extensively for phylogenetic studies of protists and other eukaryotes. It has also been used to ascertain patterns of divergence within the eukaryotic radiation (Cedergren *et al.*, 1988; Sogin, 1989; Hillis and Dixon, 1991). An extensive and growing rDNA sequence database (GenBank, Benson *et al.*, 1994) can be drawn on to compare new sequences, and as such the rDNA SSU is the obvious candidate gene for foraminiferal sequence comparisons. Wray *et al.* (1994) and Langer and Lipps (1994) have amplified and sequenced the 18S rDNA of the benthic foraminifer *Ammonia beccarii* (Linné). The sequence was verified as being foraminiferal in origin by Wray *et al.* (1994) using an *in situ* hybridization technique. Phylogenetic reconstructions place *A. beccarii* within the alveolate clade which includes ciliates, dinoflagellates and apicomplexans. The placement of a benthic foraminifer within the alveolate clade is not, however, supported by the results of Pawlowski *et al.* (1994) who designed specific foraminiferal primers from extracted RNA and used them to amplify 1600-1800 base pairs of the large subunit (LSU) rRNA gene. Phylogenetic analysis using sequences from the LSU rDNA database places the benthic foraminifera close to the plasmodial and

cellular slime moulds. The problem of foraminiferal classification therefore remains unresolved.

A range of 'universal' eukaryotic primers are available for the polymerase chain reaction (PCR) amplification of the SSU 18S rDNA (Medlin *et al.*, 1988; White *et al.*, 1990). Such primers, however, are capable of amplifying both the target 18S of the foraminifers and also any contaminant rDNA present in the extracted sample. Foraminiferal contaminants include symbionts, facultative symbionts, commensals and prey particles and often have associated parasitic protists (Hemleben *et al.*, 1989). In order to avoid the amplification of non-foraminiferal rDNA, it would be desirable to obtain a monoculture of planktic foraminifers. Unfortunately, although planktic foraminifers can be maintained in culture, their gametes have never been observed to fuse to produce a further generation.

The extraction of genomic DNA from individual planktic foraminiferal specimens of *Globigerina bulloides* d'Orbigny, *Orbulina universa* d'Orbigny and *Globigerinella siphonifera* (d'Orbigny), collected directly from the ocean, has been reported by Langer *et al.* (1993). Originally, phylogenetic analysis of the derived 18S sequences showed that they were significantly different from the dinoflagellate sequences within the database. Further investigation however, suggested that these sequences clustered with *Gymnodinium* dinoflagellates (Langer and Lipps, 1994). It is surprising that sequences of *chrysophytes*, which are the known endosymbionts of *Gl. siphonifera* (Faber *et al.*, 1988), were not also found. That single foraminiferal genomes were not amplified is due, probably, to the overwhelming competition from the dinoflagellate templates. In order to obtain a large sample of foraminiferal genomic DNA with as little contamination as possible we have relied on the foraminiferal gametogenic process to produce a package of foraminiferal genomic DNA which may contain up to 250,000 genomes per test depending on the species (Spindler *et al.*, 1978; Bé *et al.*, 1983).

Our principal interest in the investigation of planktic foraminiferal DNA is to develop a technique which will enable us to carry out molecular systematic studies of closely related species and morphotypes. Limited quantities of genomic DNA rules out the use of standard 'fingerprinting' techniques, which require large quantities of genomic DNA for restriction digests. Our initial investigations involved the method of Random Amplified Polymorphic DNA (RAPD, Williams *et al.*, 1990; Welsh and McClelland, 1990) a PCR based DNA fingerprinting technique. However, little progress was made using RAPD priming with foraminifera, and the attempt was abandoned.

Although the SSU rDNA is highly conserved, it also contains areas of greater variability which have been used to study relationships at the species level (Medlin *et al.*, 1991). The identification of sequences within the 18S specific to foraminifera should allow the construction of species specific primers and provide a means of obtaining foraminiferal sequences adjacent to the SSU in the internal transcribed spacer (ITS), which shows much greater variability (Figure 1; Hillis and Dixon, 1991). As an alternative to a fingerprinting approach, the study of the variable parts of the foraminiferal 18S and also the ITS, may prove equally useful in investigating the close relationships of interest.

We have evaluated the extraction techniques of Langer *et al.* (1993) and Wray *et al.* (1993) using 9 species of planktic foraminifera cultured to gametogenesis. Fragments of the SSU 18S rDNA have been amplified using the primers of White *et al.* (1990) and direct sequencing of the PCR products has begun using a fluorescent terminator automated sequencing system (Applied Biosystems). Here, we outline our evaluation of the extraction techniques as used with our sampling method. In addition, the PCR techniques are discussed where they differ from those of Langer *et al.* (1993) and Wray *et al.* (1993). The sequencing methods are also reported and preliminary results discussed.

Collection and culture of gametogenic foraminifers

In order to obtain planktic foraminifera at the stage of gametogenesis, it is necessary to culture them in the laboratory. The procedures are as described in Hemleben *et al.* (1989). The tropical planktic foraminifera were collected two miles off the west coast of the island of Curaçao, Dutch Antilles and cultured onshore at the Caribbean Marine Biological Institute (CARMABI).

The spinose species were caught individually by Scuba divers at a depth of 3-5 metres. The spinose species, *O. universa*, *Gl. siphonifera* Type I and Type II (Faber *et al.*, 1988), *Globigerinoides sacculifer* (Brady), *Globigerinoides ruber* (d'Orbigny, pink forma) and *Globigerinoides conglobatus* (Brady) were identified in the water column and collected individually in glass jars. A quantity of open ocean water was obtained at the same time for use in the culturing procedures. Individual spinose foraminifers were transferred to flat bottomed, closed top, glass culturing vials containing 32ml of sea water, filtered (0.45µm pore size Millipore filter) to reduce the risk of contamination from other protists. The identity of the specimens was confirmed using an inverted compound microscope and the foraminifers were maintained in culture at a constant temperature (25°C). The temperature, light intensity and day length were controlled to simulate conditions in the water column. Individual foraminifers were fed a single, one-day-old, *Artemia salina* nauplius (brine shrimp) per day, to supplement the photosynthetic products obtained from their symbionts and optimise growth rate. The spinose foraminifers were maintained in culture until they reached gametogenesis.

The non-spinose species, *Pulleniatina obliquiloculata* (Parker and Jones), *Neogloboquadrina dutertrei* (d'Orbigny) and *Globorotalia menardii* (d'Orbigny), were collected in plankton drift nets (mesh size 75µm) at a depth of 5 metres. Nets were deployed for periods of only 2 minutes to reduce crushing and entanglement of the specimens. The sample was immediately decanted from the cod end of the net into a 1000ml vessel and diluted with sea water to aid dispersal and prevent flocculation. Ashore, the samples (100ml aliquots) were poured into 9cm diameter glass culture dishes and the non-spinose foraminifera were separated from the rest of the plankton sample using a binocular dissecting microscope (×50). The non-spinose foraminifera do not float in small quantities of sea water and sink to the bottom of the sorting vessel. They are often found surrounded by a mass of rhizopodia entangled with debris. The adherent material may be a combination of prey particles on which

the foraminifers were feeding before capture and the detritus acquired during containment. The debris was gently separated from the tests using dissecting needles and the foraminifera were transferred to clean culture dishes containing 300ml of filtered sea water. Pasteur pipettes were used for the transfer and the foraminifera were left to shed any remaining debris by cytoplasmic streaming. Healthy, non-spinose foraminifera adhering to the bottom of the dish and extending a halo of rhizopodia were transferred into new culture dishes, approximately 10 to a culture. Non-toxic plastic covers were placed over each dish and maintained in the same light and temperature conditions as those used for the spinose species.

Non-spinose species of foraminifera are omnivorous and are known to consume more algal than animal prey (Anderson *et al.*, 1979; Hemleben *et al.*, 1985). They do survive and grow however, when fed only animal prey. A drop of finely chopped *Artemia* nauplii was discharged as a slurry above their rhizopodial network and the 3 species cultured grew normally and produced extra chambers. The feeding conditions were sufficient to successfully culture them to their gametogenic stage.

Identification of the gametogenic stage

The recognition of gametogenesis in spinose species is well documented and is detailed in a diagrammatic timetable by Bé *et al.* (1983) for the species *G. sacculifer* (Figure 2). The first sign of gametogenesis is indicated by the test sinking to the bottom of the vessel followed by spine shortening, rhizopodial withdrawal and the expulsion of symbiont debris. At this stage, the specimens were transferred to a culture vial containing fresh filtered sea water to reduce the contamination from detrital feeding protists during the later stages of nuclear proliferation. The exchange of water also removes any associated parasitic dinoflagellates from the surrounding medium.

Nuclear proliferation occurs over a period of hours and changes can be recognised which indicate the imminent release of the gametes. A mass of granular cytoplasm emerges from the aperture of the test; or in the case of *O. universa* small bulges from the large pores of the final chamber wall. At this stage, the tests were thoroughly cleaned with fine sable brushes in filtered sea water to remove any adherent material, rinsed again, and placed into the bottom of a 0.5ml microtube using a pasteur pipette. The remaining sea water was drained off and the specimens were immediately frozen at -25^{0}C. Speed at this stage is essential to prevent the rupture of the gametogenic bulge and loss of gametes.

Signs of gametogenesis in non-spinose species are less well documented. In the case of *Gl. menardii* the test appears to turn from a pink to a milky cream colour, which may be attributed to the consumption of prey particles which colour the cytoplasm and also to the secondary thickening of the test wall prior to gametogenesis (Bé, 1980). The flagellated gametes (Bé and Anderson, 1976a; Bé *et al.*, 1977; Spindler *et al.*, 1978) can be clearly seen as a shimmering mass through the test wall, making it possible to pick and freeze at exactly the correct stage. *N. dutertrei* also shows signs of secondary thickening; changing from a tawny light brown to a textured white. It is not possible, however, to see through the test wall and imminent gametogenesis is hard to detect. At gametogenesis a flocculant white mass appears at the aperture and then quickly disperses. Vigilant observation is necessary to interrupt the gametogenic process at the correct stage for freezing before the gamete mass emerges. *P. obliquiloculata* changes from a bright red to a textured light pink before gametogenesis due to test wall thickening. Few specimens were present in the plankton samples and only 2 individuals were observed to reach the correct stage of gametogenesis for DNA analysis. All non-spinose specimens were processed for freezing using the same procedure as in the spinose species.

The samples were transported in dry ice to the UK and preserved at -70°C to prevent nuclease activity.

Foraminiferal genomes

The reproductive cycle of planktic foraminifera is not understood. Their gametes (or asexual swarmers) have never been observed to fuse and it is unknown whether they have the biphasic life cycle found in some benthic foraminifera (Grell, 1973). If the cycle is sexual, the flagellated cells released from planktic foraminifera could be products of the mitotic division of a mononucleate haploid cell. Diploid zygotes would be formed following fusion of the gametes which in turn would divide meiotically at some stage to produce single cell haploid individuals. If the meiotic division following gamete fusion occurs immediately prior to proloculus formation (the first chamber), it would indicate that no diploid multinucleate stage exists within the planktic foraminiferal life cycle. On the other hand, it is possible that a multinucleate diploid phase of the life cycle remains unrecognised or even that planktic foraminifers do not have alternation of generations.

Haploid gametes released from a haploid parent cell should possess identical copies of genomic DNA. The number of gametes present in each foraminiferal test is specific to the

species. In spinose *G. sacculifer* there are at least 2.8 x 10^5 gametes per test (Bé *et al.*, 1977). It is possible that *Gl. menardii* also contains gamete numbers within this range as the test is of equivalent size and was observed to be completely filled by the gamete mass. *N. dutertrei*, however, appears to have a slightly smaller gamete mass. It is most likely that all species contain numbers of gametes in the order of at least 10^5 genomes which is far in excess of any contaminant genomes which may be present.

Extraction of genomic DNA from foraminifers and the amplification of the SSU rRNA gene

The microtubes containing the single specimens of each foraminifer were taken from the -70°C freezer and the tip of the tubes immediately cut off to give instant access to the frozen tests. The tests were lifted with a sable brush, leaving behind any residual sea water, and placed in an microtube with 20µl of TE buffer (10mM Tris-Cl., pH 8.0 and 1mM EDTA [disodium ethylenediaminetetra-acetate]) and immediately crushed.

Extraction method of Wray *et al.* (1993)

After crushing, the protocol of Wray *et al.* (1993) was used, which involved a complete digest using Proteinase K and incubation with 10% CTAB (cetyltrimethylammonium bromide). The incubation was followed by phenol/chloroform/isoamyl alcohol extraction and precipitation in ethanol. The DNA pellet was resuspended in 50µl of DEPC (diethyl pyrocarbonate) treated sterile distilled water.

Extraction method of Langer *et al.* (1993).

Single specimens of all species collected were placed in a clean microtube, as described, and lightly crushed in 10µl of a mixture containing low EDTA TE buffer + 1% β- Mercaptoethanol. The sample volume was adjusted to 50µl and the samples incubated at 70°C for 15 min. The β -Mercaptoethanol was then allowed to evaporate at 50°C for 3 hrs. The remaining liquid was centrifuged for 5 seconds and transferred to a fresh tube.

PCR amplification

The isolates were serially diluted to remove the effects of inhibitors of the PCR found to be present within the extract. Although there were a large number of foraminiferal 18S templates present in our samples, we were unable to visualise the whole SSU gene on agarose gels following the primary PCR even though the PCR conditions were extensively modified

in order to maximise yield. As the whole of the gene could not be isolated, a dilution and internal priming ('nested') approach was adopted to amplify the selected region in two rounds (Figure 1). The 'universal' amplification primers of White *et al.* (1990) were used and the amplification performed in a 25µl volume using the Langer *et al.* (1993) PCR reaction mixture. The PCR conditions were modified in order to optimise amplification and were found to produce consistent results (Table 1).

PCR products were identified by gel electrophoresis (100v, 1.5hrs) using a 1% agarose (Flowgen) gel in a 1 X TBE (Tris-borate-EDTA) buffer. Products were visualized on a UV trans-illuminator following ethidium bromide staining and the size of the amplified products was estimated relative to a pGEM DNA molecular weight marker (Promega; Figure 3). Successful PCR amplifications were replicated using 100µl volumes and run on a 0.8% low melting point agarose gel (Sea Plaque) at 40v for 6 hrs. The bands were excised and purified using a Wizard 'magic' PCR clean up kit (Promega) and the concentration of the products determined by gel electrophoresis against known standards in preparation for sequencing.

PCR product comparison

The initial investigation involved extracts obtained using the Wray *et al.* (1993) protocol. The 'nested' primers often produced more than one band in the correct size range for the SSU fragment being amplified (White et al., 1990), indicating the presence of contaminants within the sample. It is most likely that such a complete digest releases even the smallest amount of contamination for amplification, which, although insignificant in the primary amplification, becomes amplified in the secondary PCR. Bright single bands were obtained by 'nesting' internal primers using a 1µl aliquot of primary product of the White *et al.* (1990) extraction and were of a similar size, equivalent to those described for the appropriate primer pair (Figure 3; White *et al.*, 1990).

Sequencing Methodology

The SSU rDNA gene region chosen for the initial investigation is located between primers N5 and N6 (Figure 1). This fragment contains areas which are highly conserved and also a short variable region which may be of particular use in comparing the closer relationships within the planktic foraminifers. A direct, enzymatic, 'cycle sequencing' method (Leigh Brown and Simmonds, 1994), which incorporates a dye label into the DNA along

with the terminating bases ('Taq' Dye Deoxy™ Terminator Cycle Sequencing Kit, ABI), was used to sequence the isolated fragments. Thermal cycling of the sequencing reaction increases signal intensity and decreases sensitivity to reaction conditions. Following 25 cycles, the products were purified and run on an acrylamide gel within an Applied Biosystems 373A automated sequencing system which reads the fluorescence labelled fragments by laser. The whole of the N5/N6 fragment can be sequenced on a single gel using this method. The method has also proved particularly robust in preventing secondary structure formation as bonds are consistently broken at the denaturing temperatures of every cycle. Complementary strands were sequenced for cross checking each base pair and each strand duplicated.

Sample quality

The use of "universal" eukaryotic rDNA primers is potentially hazardous when investigating the phylogenetic placement of foraminifers within the protist lineages. Foraminifers are thought to be closely related to their symbionts and prey and amplification of symbiont and prey rDNA complicates the identification of foraminiferal sequences and the interpretation of phylogenetic analysis. Most spinose foraminifers sequester dinoflagellate or chrysophyte symbionts. Many non-spinose species possess facultative chrysophytes (Hemleben et al., 1989). Contamination could be minimised by selecting species of foraminifers that do not contain symbionts when taking foraminifera direct from the ocean. The spinose foraminifers which do not harbour endosymbionts, however, are often associated with free-swimming dinoflagellates between their spines, as in *G. bulloides* (Spero and Angel; 1991), or with large numbers of commensals, as in *Hastigerina pelagica* (d'Orbigny) (Spindler and Hemleben, 1980). The cytoplasm of foraminifera, collected directly from the ocean, will also contain food particles of unknown eukaryotes and intact genomes may be present which cannot be eliminated before extraction. One of the benefits of selecting gametogenic foraminifers is that they are thought to consume or expel their symbionts prior to gametogenesis (Hemleben *et al*, 1989). In addition, prey particles would most likely be consumed to maximise nutrient resources in order to generate the extra energy needed for the intensive cellular activity and division during gamete production. The contamination in gametogenic foraminifera should therefore be minimal and foraminiferal genomes should far outnumber those of the contaminants.

Target SSU 18S templates

The copy number of the rRNA gene within the foraminiferal genome is not known. The rDNA array of a eukaryote nuclear genome typically consists of several hundred tandemly repeated copies (Long and Dawid, 1980), although they vary between one to several thousand. Such a number may also be present in contaminant rDNA. The use of a single foraminiferal nuclear genome as a template to amplify foraminiferal rDNA is potentially hazardous when the contaminant rDNA templates may be present in far greater numbers and hence be preferentially amplified by PCR. The amplification of SSU rRNA genes from planktic foraminifers has been achieved by primarily diluting single individuals by 10^{-1} and

10^{-2} (Langer *et al.*, 1993). At these dilutions the samples contain very few copies of the target gene and the sequences have since been shown to be solely dinoflagellate in origin (Langer *et al.*, 1994). In contrast, amplification of the 18S of the benthic foraminifera *A. beccarii*, a foraminifer which does not contain endosymbionts, has been achieved from single specimens (Langer *et al.*, 1994). The sequence consensus was supported by Wray *et al.* (1993 and 1994) using 100-500 individual specimens of *A. beccarii* and confirmed as being foraminiferal in origin by *in situ* hybridisation (Wray *et al.*, 1994). In gametogenic planktic foraminifers, the proportion of foraminiferal templates compared to any contaminant should be overwhelming and the foraminiferal templates preferentially amplified by PCR.

PCR product analysis

Preliminary sequencing data indicate that, although the bands of the White *et al.* (1990) extraction appear to be single when run on an agarose gel, they sometimes contain strands with more than one sequence. Many protists, including potential contaminants, possess SSU rDNA of a similar size. It is therefore, quite clear that whichever extraction method is chosen initially in an attempt to overcome the contamination problem, the potential of the PCR to amplify even a single contaminant molecule present in the primary amplification may introduce substantial quantities of contaminant 18S into the secondary product. This is a serious problem when attempting to sequence the full 18S gene using a nested primer approach as separate contigs may not be derived from the same genome. A problem which would be compounded should the 'universal' primers preferentially amplify contaminant rDNA.

It is now clear that the inherent problem of cross-species contamination will not be overcome at the level of extraction and 2 approaches are currently being adopted to ensure the amplification of single foraminiferal genomes. Firstly, the fragment targeted using primers N5/N6 contains a highly variable region which may be variable enough to design specific primers which should allow the amplification of the rest of the 18S relating specifically to the sequenced fragment. The second strategy will be to dilute the samples to single molecules using a statistical, titration approach (Leigh Brown and Simmonds, 1990). The single molecules will then be amplified using the nested PCR.

The fragment chosen for the initial investigation contains approximately 300 bases, which will not be sufficient to obtain a reliable estimate of the phylogeny. A phylogenetic tree, constructed from known protist sequences, including the ones obtained from the planktic

foraminifers, should however, provide an indication of whether the generated sequences show different phylogenetic origins within the eukaryotes.

It is not possible to choose between the merits of the extraction techniques of Wray *et al.* (1993) and Langer *et al.* (1993) until we have fully sequenced the 18S gene fragments and carried out extensive phylogenetic analyses to determine whether planktic foraminiferal 18S was amplified following extraction. The less rigorous Langer *et al.* (1993) extraction consistently produces single bands of similar size although these have subsequently been shown to contain multiple 18S sequences. It is most likely that this variation is due to amplification of contaminants within the primary PCR. The single molecule and foraminiferal-specific primer approach will hopefully resolve these problems. The multiple bands observed on gels following the Wray *et al.* (1993) extraction will be sequenced as there is considerable variation in 18S size within protists and it is possible that the foraminiferal N5/N6 fragment is larger than the benthic foraminiferal fragment (Langer *et al.*, 1994; Wray *et al.*, 1994).

The benthic foraminiferal sequence will be important for phylogenetic comparison as planktic foraminifers most likely evolved from benthic foraminifers and it should therefore provide a useful reference. On the other hand, it is possible that these groups are phylogenetically distinct. It is also quite possible that spinose and non-spinose species of planktic foraminifera will prove to be substantially different from one another. GenBank now contains sequences from an extensive array of protists and it should be possible, using complete 18S sequences, to determine the source of the sequence under study. Wray *et al.* (1994) confirmed their benthic foraminiferal sequence by using an *in situ* hybridisation technique. It may be possible to use a similar approach to confirm candidate planktic foraminiferal sequences.

The lack of success in amplifying 18S rDNA from single non-gametogenic planktic foraminifers (Langer *et al.*, 1994) is disappointing. Collecting foraminifers directly from the ocean would simplify the investigation into morphotype discrimination, as only a few specimens of each would be required. It may be possible to overcome this problem by using foraminiferal specific primers when the sequences of the relevant species become available. Intra-species variable sequences would then be targeted for amplification which may eventually lead to determine whether foraminiferal morphotypes are genotypic expressions or modulated by the environment.

ACKNOWLEDGEMENTS

We thank Jelle Bijma and Brian Huber for their advice and help during the collection of the planktic foraminifers and also Heather Austin for her help in culturing them at the Caribbean Marine Biological Institute (CARMABI, Curaçao, Dutch Antilles). Prof. Hemleben is acknowledged for making essential equipment available for this project. We would like to thank Petra zur Lage and Sarah Ashelford for their help during the pilot study. We would also like to thank Sandy Cleland, Elizabeth Harvey, Pamela Robertson and Sarah Ross for their invaluable technical assistance. Financial support was provided by the Carnegie Trust for the field work in Curaçao

REFERENCES

Anderson, O. R., M. Spindler, A. W. H. Bé, and Ch. Hemleben. 1979. Trophic activity of planktic foraminifera. J. Mar. Biol. Assoc. U.K. 59:791-799.

Bé, A. W. H. 1980. Gametogenic calcification in a spinose planktic foraminifer, *Globigerinoides sacculifer* (Brady). Marine Micropaleontology 5:283-310.

Bé, A. W. H. and O. R. Anderson. 1976a. Gametogenesis in planktic foraminifera. Science 192:890-892.

Bé, A. W. H., O. R. Anderson, and W.W. Faber, jr. 1983. Sequence of morphological and cytoplasmic changes during gametogenesis in the planktic foraminifer *Globigerinoides sacculifer* (Brady). Micropaleontology 29:310-325.

Bé, A. W. H., C. Hemleben, O. R. Anderson, M. Spindler, J. Hacunda, and S. Tuntivate-Choy. 1977. Laboratory and field observations of living planktic foraminifera. Micropaleontology 23:155-179.

Benson, D. A., M. Boguski, D. J. Lipman, and J. Ostell. 1994. GenBank. Nucl. Acids Res. 22:3441-3444.

Berggren, W. A., D. V. Kent, and J. J. Flynn. 1985a. Paleogene geochronology and chronostratigraphy. *In* Snelling, N. J. ed. Geochronology and the Geologic Time Scale. Geol. Soc. London Mem. 10:141-195.

Berggren, W. A., D. V. Kent, and J. A. Van Couvering. 1985b. Neogene geochronology and chronostratigraphy. *In* Snelling, N. J. ed. Geochronology and the Geologic Time Scale. Geol. Soc. London Mem. 10:211-260.

Bolli, H. M., and J. B. Saunders. 1985. Oligocene to Holocene low latitude planktic foraminifera. *In* Bolli, H. M., J. B. Saunders, and K. Perch-Nielsen. eds. Plankton Stratigraphy, (1). Cambridge University Press. pp.155-262.

Brummer, G. J. A and D. Kroon. 1988. Genetically controlled planktic foraminiferal coiling ratios as tracers of past ocean dynamics. *In* Brummer, G. J. A., and D. Kroon. eds. Planktic foraminifers as tracers of ocean-climate history. Free University Press, Amsterdam. pp.293-298.

Cedergren, R., M. W. Gray, A. Abel, and D. Sankoff. 1988. The evolutionary relationships among known life forms. J. Mol. Evol. **28**:98-112.

CLIMAP-Project Members. 1976. The surface of the ice-age earth. Science **191**:131-137.

Faber, W. W., Jr., O. R. Anderson, and D. A. Caron. 1988. Algal foraminiferal symbiosis in the planktic foraminifera *Globigerinella aequilateralis.* 1. Occurance and stability of two mutually exclusive chrysophyte endosymbionts and their ultrastructure. J. Foram. Res. **18**:334 -343.

Grell, K. G. 1973. Protozoology. Berlin, Heidelberg, New York. pp.96-100.

Hemleben, C., M. Spindler, and O. R. Anderson. 1989. Modern planktic foraminifera. Springer-Verlag, New York.

Hemleben, C., M. Spindler, I. Breitinger, and W. G. Deuser. 1985. Field and laboratory studies on the ontogeny and ecology of some *globorotaliid* species from the Sargasso sea off Bermuda. J. Foram. Res. **15**:254-272.

Hillis, D. M., and M. T. Dixon. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Quarterly Review Biology **66**:411-446.

Kennett, J. P. and M. S. Srinivasan. 1983. Neogene planktic foraminifera, Hutchinson Ross, Stroudsburg.

Kroon, D., P. F. Wouters, L. Moodley, G. Ganssen, and S. R. Troelstra. 1988. Phenotypic

variation of *Turborotalita quinqueloba* (Natland) tests in living populations and in the Pleistocene of an eastern Mediterranean piston core. *In* Brummer, G. J. A., and D. Kroon. eds. Planktic foraminifers as tracers of ocean climate history. Free University Press, Amsterdam. pp.131-147.

Langer, M. R., and J. H. Lipps. 1994. Molecular evolution and phylogenetic status of foraminifera and dinoflagellate symbionts as inferred from nuclear 18S-rDNA sequences. PaleoBios **16**:42.

Langer, M. R., J. H. Lipps, and W. E. Piller. 1993. Molecular paleobiology of protists: amplification and direct sequencing of foraminiferal DNA. Micropaleontology **39**:63-68.

Leigh Brown, A. J., and P. Simmonds. 1994. Analysis of HIV sequence variation. *In* Karn, J. ed. HIV - a Practical Approach. Oxford University Press.

Long, E. O., and I. B. Dawid. 1980. Repeated genes in eukaryotes. Annu. Rev. Biochem. **49**:727-764.

Medlin, L. K., H. J. Elwood, S. Stickel, and M. L. Sogin. 1988. The characterization of enzymatically amplified eukaryotic 16-S like rRNA-coding regions. Gene **71**:491-499.

Medlin, L. K., H. J. Elwood, S. Stickel, and M. L. Sogin. 1991. Morphological and genetic variation within the diatom *Skeletonema costatum* (Bacillariophyta): evidence for a new species, *Skeletonema pseudocostatum*. J. Phycol. **27**:514-524.

Mullis, K. B., and F. A. Faloona. 1987. Specific synthesis of DNA *in vitro* via a polymerase-catalysed chain reaction. Methods in Enzymology **155**:335-350.

Palowski, J., I. Bolivar, and J. Guiard-Maffia. 1994. Molecular phylogeny of foraminifera inferred from partial LSU rDNA sequences. PaleoBios **16**:52.

Saiki, R. K., D. H. Gelfland, S. Stoffel, S. J. Scharf, R. Higuchi, K. B. Mullis, and H. A. Ehrlich. 1988. Primer-directed enzymatic amplification of DNA with thermostatic DNA

polymerase. Science 239:487-490.

Sogin, M. L. 1989. Evolution of eukaryotic microorganisms and their small subunit ribosomal RNA's. Amer. Zool. 29:487-499.

Spero, H. J., and D. L. Angel. 1991. Planktic sarcodines - microhabitat for oceanic dinoflagellates. J. Phycology 27:187-195.

Spindler, M., O. R. Anderson, C. Hemleben, and A. W. H. Bé. 1978. Light and electron microscopic observations of gametogenesis in *Hastigerina pelagica* (Foraminifera). J. Protozool. 25:427-433.

Spindler, M., and Ch. Hemleben. 1980. Symbionts in planktic foraminifera (Protozoa). *In* Schwemmler, W., and H. E. A. Schenk. eds. Endocytobiology, Endosymbiosis and Cell Biology. Berlin, New York. 1, pp.133-140.

Vincent, E. and W. H. Berger. 1981. Planktic foraminifera and their use in Paleoceanography. *In* Emiliani, C. ed. Series The Sea, The Oceanic Lithosphere. Wiley and Sons, New York. 12, pp.1025-1119.

Welsh, J. and M. McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucl. Acids Res. 18:7213-7218.

White, T. J., T. Bruns, S. Lee, and J. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* Innis, M. A., D. H. Gelfland, J. J. Sninsky, and T. J. White. eds. PCR Protocols: A guide to methods and applications. Harcourt Brace Jovanovich, San Diego. pp.315-322.

Williams, J. G. K., A. R. Kubelik, K. J. Lival, J. A. Rafalski, and S. V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucl. Acids Res. 18:6531-6535.

Wray, C. G., R. DeSalle, and J. L. Lee. 1994. A phylogenetic placement of the

foraminifera based upon nuclear small-subunit rDNA sequence and DNA in-situ hybridization. PaleoBios 16:67.

Wray, C. G., J. L. Lee, and R. DeSalle. 1993. Extraction and enzymatic characterization of foraminiferal DNA. Micropaleontology 39:69-73.

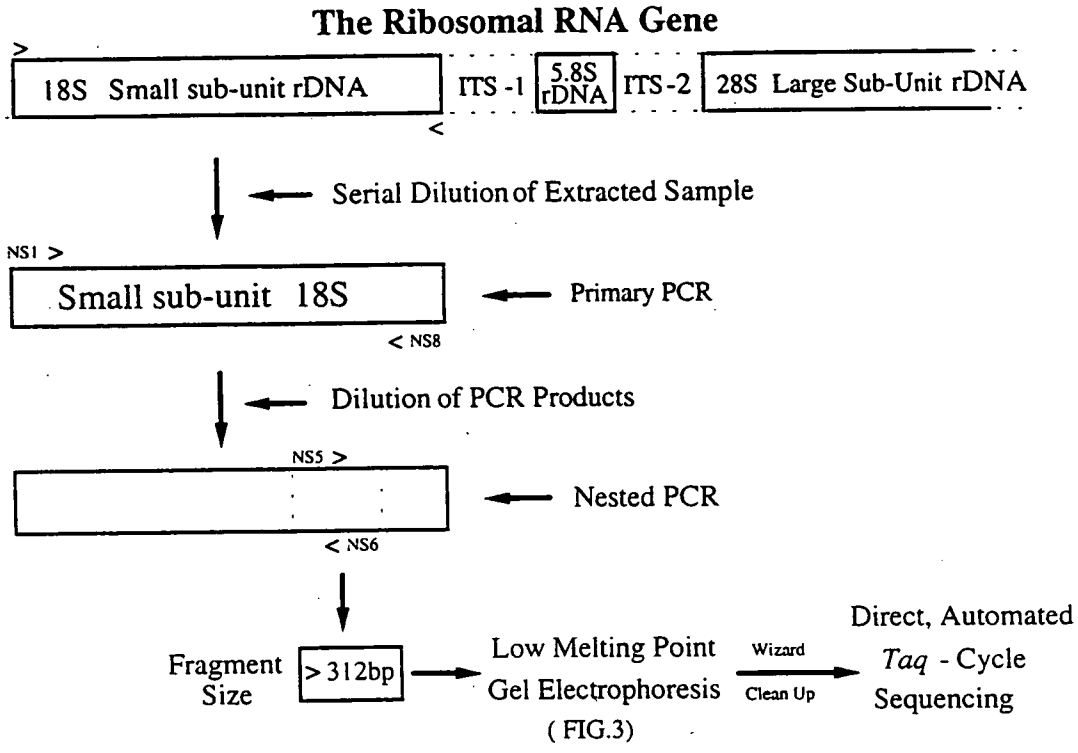
Table 1.

Conditions used for the PCR.

Denature	Anneal	Extend	
94°C, 2min	48°C, 5 min	72°C, 4 min	1 Cycle
94°C, 25 sec	48°C, 35 sec	72°C, 4 min	30 Cycles
		72°C, 7 min	Final Extension

Figure 1.

Flow diagram of the amplification and sequencing of fragment N5/N6 of the rRNA gene using the 'universal' primers designed by White *et al.* (1990).


The Ribosomal RNA Gene
18S Small sub-unit rDNA
ITS -1
5.8S rDNA
ITS -2
28S Large Sub-Unit rDNA
Serial Dilution of Extracted Sample
NS1 >
Small sub-unit 18S
< NS8
Primary PCR
Dilution of PCR Products
NS5 >
< NS6
Nested PCR
Fragment Size
> 312bp
Low Melting Point Gel Electrophoresis
Wizard Clean Up
Direct, Automated Taq - Cycle Sequencing

(FIG.3)

Figure 2.

The stages of planktic foraminiferal gametogenesis as observed in *Globigerinoides sacculifer* (Brady; Bé *et al.*, 1983, with permission). The foraminifers are picked at stage 4, which is towards the end of the nuclear proliferation phase, before the expansion of the gametogenic bulge. At this stage genome content is maximal and tests must be picked and frozen immediately to prevent gamete release.

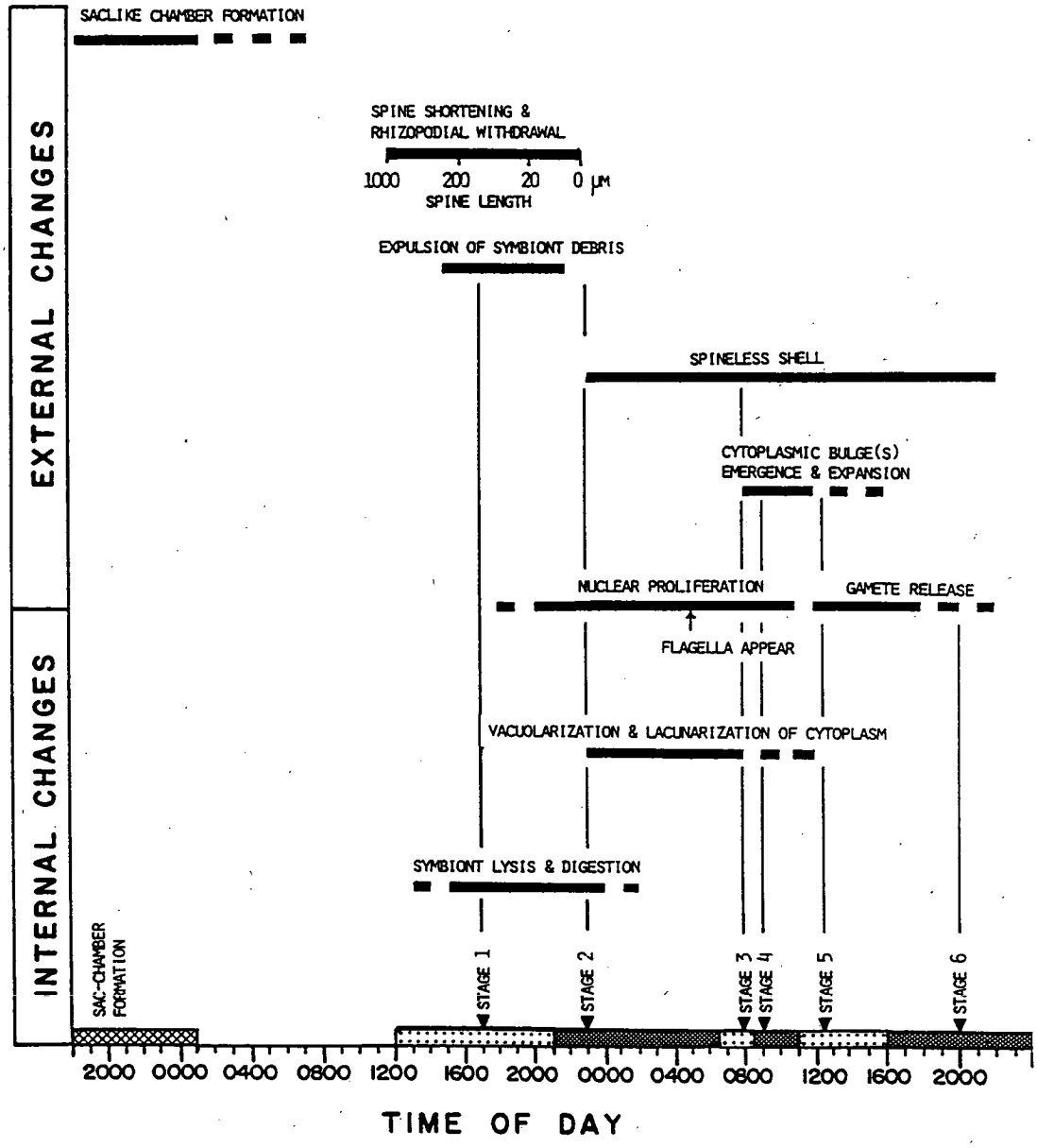
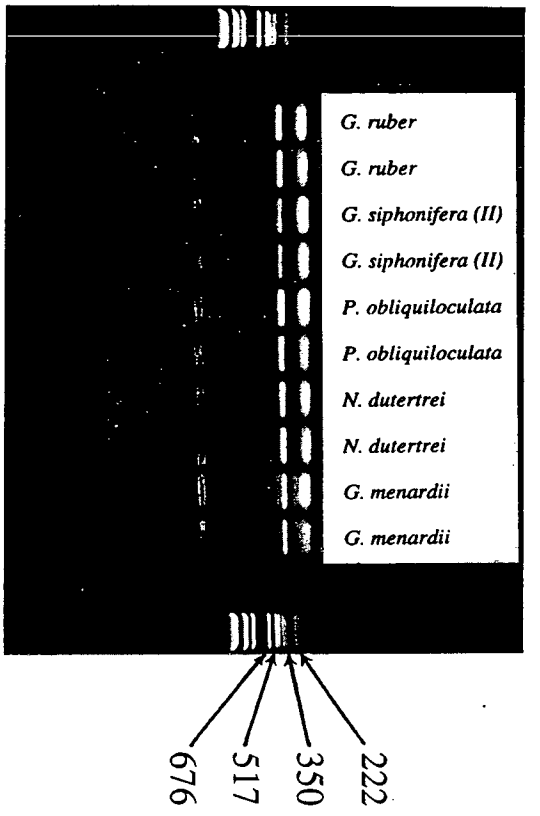

EXTERNAL CHANGES
INTERNAL CHANGES
SACLIKE CHAMBER FORMATION
SPINE SHORTENING &
RHIZOPODIAL WITHDRAWAL
1000 200 20 0 µM
SPINE LENGTH
EXPULSION OF SYMBIONT DEBRIS
SPINELESS SHELL
CYTOPLASMIC BULGE(S)
EMERGENCE & EXPANSION
NUCLEAR PROLIFERATION
GAMETE RELEASE
FLAGELLA APPEAR
VACUOLARIZATION & LACUNARIZATION OF CYTOPLASM
SYMBIONT LYSIS & DIGESTION
SAC-CHAMBER FORMATION
STAGE 1
STAGE 2
STAGE 3
STAGE 4
STAGE 5
STAGE 6
2000 0000 0400 0800 1200 1600 2000 0000 0400 0800 1200 1600 2000
TIME OF DAY

Figure 3.

A low melting point agarose gel, calibrated with a pGEM (Promega) molecular weight marker (base pairs), showing the amplified fragments (> 312 base pairs) of 5 planktic foraminifera using the 'universal' primers N5/N6 (White *et al*, 1990).


G. ruber
G. ruber
G. siphonifera (II)
G. siphonifera (II)
P. obliquiloculata
P. obliquiloculata
N. dutertrei
N. dutertrei
G. menardii
G. menardii
676
517
350
222

Paper II

Molecular Phylogeny of the Planktic Foraminifera

Kate F. Darling[1*], Dick Kroon[1], Christopher M. Wade[2], and Andrew J. Leigh Brown[2]

[1]Department of Geology and Geophysics, Grant Institute, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JW, UK.

[2]Centre for HIV Research, Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JN, UK.

* Corresponding Author

Journal of Foraminiferal Research (1996) **26**:324-330

ABSTRACT

Planktic foraminifers are an important group of marine zooplankton whose fossil record has been used extensively to reconstruct past climatic and oceanographic changes. Knowledge of the molecular phylogenetics of the planktic foraminifera will allow us to determine whether a link exists between individual water masses and specific foraminiferal genotypes and help unravel the relationship between planktic foraminiferal evolution and the history of ocean circulation. However, the isolation of foraminiferal genomic DNA has proved technically complex due to major difficulties in obtaining uncontaminated DNA from living specimens. We have minimized this problem by extracting genomic DNA from gametogenic foraminifers, and successfully amplified an approximately 1000 base pair fragment of the foraminiferal small subunit (SSU) ribosomal (r)RNA gene for five planktic species. A molecular phylogeny was constructed which shows that the planktic foraminifera form a distinct monophyletic group within the eukaryotic SSU rDNA tree, probably representing one of the earliest splits among free-living aerobic eukaryotes. This separation preceded the rapid eukaryote diversification, represented by the "crown" group, dated at 1100-1000 million years before present (Knoll, 1992). The monophyly of the planktic foraminiferal group conclusively confirms the foraminiferal origin of the amplified DNA. The planktic foraminiferal SSU rRNA gene also contains a number of highly variable regions of which some are unique to the foraminifera. These areas should permit genetic comparisons at the species and morphotype level, allowing the interpretation of the relationships between genotype, phenotype and environmental change.

INTRODUCTION

The planktic foraminifera are an important group of marine zooplankton and are among the most widely used microfossils for reconstructing oceanic environments of the past (Bé, 1977; Brummer and Kroon, 1988; Hemleben *et al.*, 1989). The assemblages of planktic foraminifera are excellent indicators of different ocean water masses, yet their evolutionary origin, their relationship to the benthic foraminifera and the evolutionary relationships within the planktic foraminifera remain poorly understood. Knowledge of the molecular phylogenetics of planktic foraminifera will allow us to determine whether a link exists between individual water masses and specific foraminiferal genotypes and help unravel the relationship between planktic foraminiferal evolution and the history of ocean circulation.

The ability to amplify large quantities of specific sequences of DNA using the polymerase chain reaction (PCR) (Mullis and Faloona, 1987; Saiki *et al.*, 1988) permits analyses of the foraminifera at the genotype level through foraminiferal DNA sequence analysis. The highly conserved, SSU rRNA gene (Figure 1) has been used extensively for phylogenetic studies of protists and other eukaryotes, showing the patterns of divergence within the eukaryotic radiation (Cedergren *et al.*, 1988; Sogin, 1989; Hillis and Dixon, 1991). An extensive rDNA sequence database (Maidak *et al.*, 1994) is available for the comparison of new sequences, making the SSU rRNA gene the obvious candidate for foraminiferal sequence comparisons.

The extraction of genomic DNA from individual planktic foraminiferal specimens collected directly from the ocean, has been reported by Langer *et al.* (1993). Analysis of their amplified SSU rDNA sequences showed that they differed from known dinoflagellate sequences by approximately 10% (Langer *et al.*, 1993) . This level of diversity is not significantly greater than the diversity levels observed between the dinoflagellates shown in the present study (9% diversity between *Symbiodinium pilosum* and *Crypthecodinium cohnii*, Figure 2). Attempts to determine the relative position of benthic foraminifera within SSU rDNA eukaryotic phylogenetic trees has produced contradictory results. Analysis of SSU rDNA sequences from two species of the genus *Ammonia*, sequenced independently in separate laboratories (Wray *et al.*, 1995), suggested a position for the benthic foraminifera within the alveolate clade, part of the eukaryotic "crown" group of the SSU rDNA tree. The identity of the Wray *et al.* (1995) sequences was also supported by an investigation into their location within the foraminiferal cell, using an *in-situ* hybridization approach, which indicated

that they were positioned within the foraminiferal nuclei. The "crown" group placement of the benthic foraminifera was not supported in phylogenetic trees inferred from partial sequences of the large subunit (LSU) rRNA gene (Pawlowski *et al.*, 1994a), which placed the *Ammonia* genus close to the plasmodial and cellular slime molds, prior to the separation of the major eukaryote lineages. Furthermore, this placement is supported by analyses of the LSU rRNA gene for three additional benthic genera (Pawlowski *et al.*, 1994a) and two planktic genera (Merle *et al.*, 1994). The problem of foraminiferal phylogeny has therefore remained in dispute.

A range of 'universal' eukaryotic primers are available for the PCR amplification of eukaryotic SSU rDNA (Medlin *et al.*, 1988; White *et al.*, 1990). The primers are complementary to the most highly conserved regions of the rRNA gene and are capable of amplifying both the target SSU of the foraminifera and also any contaminant rDNA, which may include endosymbionts, parasites, commensals or prey organisms. In order to avoid the amplification of non-foraminiferal rDNA by the "universal" primers, it would be desirable to obtain a monoculture of planktic foraminifera. Unfortunately, the reproductive cycle of planktic foraminifera is not understood. Their gametes (or asexual swarmers) have never been observed to fuse in culture (Anderson *et al.*, 1991), and it is unknown whether they have the biphasic life cycle found in some benthic foraminifera (Grell, 1973; Lee *et al.*, 1991). It is, therefore, not possible to obtain quantities of planktic foraminiferal genomic DNA by monoculturing.

We have developed an alternative procedure which has enabled us to extract large quantities of planktic foraminiferal genomic DNA by utilizing the foraminiferal gametogenic process. By interrupting the process immediately prior to gamete release, a package of foraminiferal genomic DNA can be obtained which contains typically 300,000-400,000 genomes per test, depending on the species (Spindler *et al.*, 1978; Bé *et al.*, 1983). In addition, gametogenic foraminifera have been observed to consume or expel their symbionts and prey particles prior to gametogenesis (Hemleben *et al.*, 1989) which significantly reduces the number of contaminating genomes. Maximization of the ratio of foraminiferal genomes to those of the contaminants makes the amplification of foraminiferal DNA possible, despite the observed preferential amplification of "crown" group organisms by the "universal" primers.

The ability to obtain multiple copies of planktic foraminiferal genomic DNA has enabled us to amplify an approximately 1000 base pair region of the SSU rRNA gene of 5

tropical planktic foraminifera. Comparisons with a diverse range of eukaryote taxa, including known symbionts and prey organisms, have revealed that the foraminiferal sequences do not cluster with any other known eukaryote sequence. This, in addition to the monophyly of the planktic foraminiferal group, provides strong evidence for the foraminiferal origin of the amplified sequences.

MATERIALS AND METHODS

Collection and Culture of Planktic Foraminifera

In order to obtain planktic foraminifera at the stage of gametogenesis, we utilized the procedures as described in Hemleben *et al.* (1989) and Darling *et al.* (1996). Five species of tropical planktic foraminifera; the spinose species, *Orbulina universa* d'Orbigny, *Globigerinella siphonifera* (d'Orbigny) Type I (Faber *et al.*, 1988), *Globigerinoides sacculifer* (Brady), *Globigerinoides ruber* (d'Orbigny), and the non-spinose species, *Neogloboquadrina dutertrei* (d'Orbigny), were collected off the west coast of the Caribbean island of Curaçao and cultured onshore at the Caribbean Marine Biological Institute (CARMABI). The identity of the spinose specimens was confirmed using an inverted compound microscope and selected specimens were maintained in culture. The temperature, light intensity and day length were controlled to simulate conditions in the water column. Individual foraminifers were fed a single, one-day-old *Artemia salina* nauplius (brine shrimp) per day to supplement the photosynthetic products obtained from their symbionts and optimise growth rate. Non-spinose species of foraminifera are omnivorous and are known to consume more algal than animal prey (Anderson *et al.*, 1979). Nevertheless, they do survive and grow even when fed only animal prey, and *N. dutertrei* was successfully cultured to the gametogenic stage with finely chopped *Artemia* nauplii discharged as a slurry above the rhizopodial network.

Identification of the Gametogenic Stage

The recognition of gametogenesis in spinose species is well documented (Spindler *et al.*, 1978; Bé *et al.*, 1983). The first indication of gametogenesis is the sinking of the test to the bottom of the vessel followed by spine shortening, rhizopodial withdrawal, and the expulsion of symbiont debris. At this stage, the specimens were transferred to a culture vial containing fresh, filtered sea water to reduce contamination from detrital feeding protists during the later stages of nuclear proliferation. The exchange of water also aids in the removal of any contaminant organisms from the surrounding medium. Nuclear proliferation occurs over a period of hours and changes can be recognized which indicate the imminent release of the gametes. A mass of granular cytoplasm emerges from the aperture of the test; or in the case of *O. universa,* small bulges from the large pores of the final chamber wall. At this stage, the tests were thoroughly cleaned with fine sable brushes in filtered sea water to remove any adherent material, rinsed again, and placed into the bottom of a 0.5ml

microtube using a pasteur pipette. The remaining sea water was drained off and the specimens were immediately frozen at -25°C. Speed at this stage is essential to prevent the rupture of the gametogenic bulge and loss of gametes. The samples were transported in dry ice to the UK and preserved at -70°C to prevent nuclease activity.

Signs of gametogenesis in the non-spinose species *N. dutertrei* are less defined than for the spinose species. The test starts to show secondary thickening, changing from a tawny light brown to a textured white. At this stage, it is not possible to see through the test wall and imminent gametogenesis is hard to detect. At gametogenesis, a flocculant white mass appears at the aperture and then quickly disperses. Vigilant observation is necessary to interrupt the gametogenic process at the correct stage for freezing before the gamete mass emerges. All *N. dutertrei* specimens were processed for freezing using the same procedure as in the spinose species.

Extraction of Genomic DNA

Pilot studies were carried out to evaluate alternative extraction procedures (Darling *et al.*, 1996), and the protocol of Wray *et al.* (1993) was found to be successful in lysing the gamete cells.

Three microtubes containing individual specimens of each foraminiferal species were taken from the -70°C freezer and the tips of the tubes immediately cut off to give instant access to the frozen tests. The tests were lifted with a sable brush, leaving behind any residual sea water, and placed in a microtube containing 20µl TE buffer (10mM Tris-Cl. pH 8.0 and 1mM EDTA [disodium ethylenediaminetetra-acetate]). They were crushed and the volume adjusted to 250µl with 0.1M EDTA and 0.25% SDS (sodium dodecyl sulphate). A digest was carried out using Proteinase K (0.5µg/ml final concentration) at 65°C for 2 hours followed by a one hour incubation with 10% CTAB (cetyltrimethylammonium bromide; Clark, 1992). DNA was extracted using standard phenol/chloroform/isoamyl alcohol procedures and precipitated in ethanol (3hrs at -20°C). After centrifugation, the DNA pellet was dried and resuspended in 50µl of DEPC (diethyl pyrocarbonate) treated sterile distilled water and stored at -20°C.

PCR Amplification

Foraminiferal rDNA was amplified using the PCR (Mullis and Faloona, 1987; Saiki *et al.*, 1988). "Universal" eukaryotic SSU rDNA primers (Medlin *et al.*, 1988; White *et al.*,

1990) were unsuccessful in amplifying the whole gene, even when the amplification conditions were extensively modified. Nevertheless, the terminal 3' primer NS8 (White *et al.*, 1990) does amplify a fragment of the gene when paired with the internal primer NS5 (White *et al.*, 1990). This enabled the amplification of an approximately 1000 base pair (bp) fragment of the foraminiferal gene (equivalent positions 1151-1767 in the *Saccharomyces cerevisiae* SSU rDNA sequence; Figure 1), corresponding to the 30-48 region of the eukaryotic SSU rRNA secondary structure model (Neefs *et al.*, 1990). Amplification of this region often produces additional, smaller amplification products, indicating the presence of contaminants within the sample. Subsequent sequencing and phylogenetic analyses of these fragments identified them as primarily symbiont in origin (unpublished data).

PCR conditions were varied to optimise amplification. The conditions consisted of an initial cycle with denaturation at 94°C for 2 minutes, a 5 minute 48°C annealing step followed by extension at 72°C for 4 minutes. This was followed by 30 cycles with denaturation at 94°C for 25 seconds, a 35 second 48°C annealing step and extension at 72°C for 4 minutes; with a final extension step at 72°C for 7 minutes.

PCR products were identified by gel electrophoresis (100v, 1hr) using a 1% agarose (Flowgen) gel in a 1 X TBE (Tris-borate-EDTA) buffer. Products were visualized on a UV trans-illuminator following ethidium bromide staining and the size of the amplified products was estimated relative to a pGEM DNA molecular weight marker (Promega).

Successful PCR amplifications were replicated using 100µl volumes and run on a 0.8% low melting point agarose gel (Sea Plaque) at 40v for 6 hrs. The bands were excised and purified using a Wizard 'magic' DNA clean up kit (Promega) and the concentration of the products determined by gel electrophoresis against known standards in preparation for direct sequencing.

Sequencing Methodology

A direct, enzymatic, 'cycle sequencing' method (Leigh Brown and Simmonds, 1994), which incorporates a dye label into the DNA along with the terminating bases ('Taq' Dye Deoxy™ Terminator Cycle Sequencing Kit, ABI), was used to sequence the amplified fragments. Thermal cycling of the sequencing reaction increases signal intensity and decreases sensitivity to reaction conditions. Following 25 cycles, the products were purified and run on an acrylamide gel within an Applied Biosystems 373A automated sequencing system (Leigh Brown and Simmonds, 1994). The method has also proved particularly robust in preventing

secondary structure formation as bonds are consistently broken at the denaturing temperatures of every cycle. Complementary strands were sequenced for cross checking each base pair and each strand duplicated.

Sequence Analysis

Sequences were aligned relative to representatives of the diverse range of eukaryote taxa (obtained from the ribosomal database project SSU rDNA database; Maidak *et al.*, 1994) within the Genetic Data Environment (GDE) package (Smith *et al.*, 1994). Alignment involves identifying the homologous sites between sequences of the selected taxa. Only those positions which were unambiguously aligned were used for subsequent phylogenetic analyses.

The phylogenetic tree was reconstructed by neighbor-joining (Saitou and Nei, 1987) analysis (program NEIGHBOR, Phylogeny Inference Package (Felsenstein, 1993)) of 575 unambiguously aligned sites (corresponding to the conserved regions in Figure 1). Nucleotide sequence distances were estimated for all pairwise sequence comparisons using the generalized two parameter (maximum likelihood) model (program DNADIST (Felsenstein, 1993)). Fitch-Margoliash (1967) (program FITCH (Felsenstein, 1993)) analysis produced a topology consistent with the neighbor-joining tree. Bootstrap values (Felsenstein, 1985) (programs SEQBOOT and CONSENSE (Felsenstein, 1993)), expressed as a percentage, were based on 2000 replications.

The planktic foraminiferal SSU rDNA sequences presented in this study are deposited in GenBank accession numbers U65631-U65635.

RESULTS

We have sequenced an approximately 1000bp region (983bp for *O. universa* to 1052bp for *G. siphonifera* Type I) of the SSU rRNA gene for the five planktic foraminiferal species. An alignment of the foraminiferal sequences with 438 SSU rDNA sequences, representing a diverse range of taxa within the SSU rDNA data base (Maidak *et al.*, 1994) revealed that all the foraminiferal sequences contained four variable length expansion segments and three insertions unique to the planktic foraminifera (Figure 1). A high degree of sequence variability and length variation was observed between foraminiferal species in both the expansion segments and foraminiferal specific insertions. Preliminary investigations into the level of within species sequence variation in *G. siphonifera* Type I. has nevertheless shown complete sequence identity in five individual gametogenic specimens.

A phylogenetic analysis of the foraminifera with 26 representatives of the diverse range of eukaryote taxa, including symbionts and prey organisms, revealed that the planktic foraminifera form a clearly distinct monophyletic group (Figure 2). This group is supported in 100% of bootstrap replications. The planktic foraminifera consistently branch early in the tree, falling between the euglenoid/trypanosome branch and the slime-molds and amoebae. The precise location of the foraminiferal branch is not however supported statistically, although bootstrap analysis, excluding sequences intermediate between the foraminiferal branch and the eukaryote "crown" group, provide strong support (90% of bootstrap replications) for a location outside of the "crown" group. In our analysis, the published benthic foraminiferal sequence for *A. beccarii* (Wray *et al.*, 1995), was clearly distinct from the planktic foraminiferal sequences, showing a minimum evolutionary distance of 39% from the planktic group in comparison with a maximum within-group distance of 27%.

Analysis of the SSU rRNA gene did not provide a reliable estimate of the evolutionary relationships between the planktic foraminiferal genera with no support for any specific topology in preference to any other.

DISCUSSION

The isolation of planktic foraminiferal DNA has proved to be complex. Our initial approach, to develop a "fingerprinting" technique using genomic DNA (Darling *et al.*, 1996), proved impossible due to considerable symbiont and food particle contamination. In order to avoid these problems, we targeted the ribosomal RNA genes, since non-specific "universal" primers were available for the amplification of eukaryotic ribosomal RNA genes using the PCR. Unfortunately, the "universal" primers preferentially amplified the foraminiferal symbionts, which made it necessary to obtain samples with minimal symbiont contamination by using gametogenic specimens. The gametogenic specimens have a very high ratio of planktic foraminiferal genomes compared with the contaminants, and amplification of a large foraminiferal SSU fragment was achieved, though often accompanied by a smaller "crown" group fragment. The presence of two bands made it necessary to sequence all observed bands and compare them with known eukaryotic sequences in order to determine their origin. Only when we identified sequences which clustered together forming a monophyletic group, and which did not cluster with any other previously sequenced organisms, did we have any evidence that foraminiferal DNA may have been sequenced. The candidate bands, consistently larger than others identified on the gel, were found in all foraminiferal species studied. Sequencing of the shorter bands revealed that these fragments were mainly algal in origin, and therefore we targeted and sequenced the larger bands which we believed to be foraminiferal in origin. The larger band sequences proved highly unusual, containing large insertions not found in any other known eukaryote group. The foraminifers were subsequently found to hold a unique position in eukaryotic evolution; forming a monophyletic group completely separate from any other protistan lineage.

To locate the position of the foraminiferal group within a eukaryotic phylogenetic tree, we chose representatives of major eukaryotic organisms, including known symbionts and prey. It appears that the monophyletic planktic foraminiferal group represents one of the earliest known evolutionary divergences amongst free-living aerobic eukaryotes (Figure 2). Our results contradict the placement of the benthic genus *Ammonia* within the "crown" group (see also Figure 2), inferred from analysis of the SSU rDNA (Wray *et al.*, 1995). Nevertheless, our placement of the planktic foraminifers outside the "crown" group is strongly supported by analyses of partial sequences of the LSU rRNA gene of two planktic genera (Merle *et al.*, 1994) and four benthic genera, including *Ammonia* (Pawlowski *et al.*, 1994a).

This indicates that planktic and benthic foraminifera have a close evolutionary relationship and suggests that the Wray *et al.* (1995) *Ammonia* sequences are not foraminiferal in origin. Pawlowski *et al.* (1994a) LSU rDNA placement of *Ammonia beccarii* outside the "crown" group has subsequently been confirmed by their sequencing of *A. beccarii* SSU rDNA (Pawlowski, - unpublished data, GenBank accession no. X86094). Recent analyses have shown that this sequence clusters with the planktic sequences presented in our analyses (unpublished data), confirming the close evolutionary relationship of benthic and planktic foraminifera. A study of the structure and composition of tests from the fossil record suggests that the ancestors of the planktic foraminifers were benthic foraminifers (Tappan and Loeblich, 1988), which first appeared in the fossil record in the Lower Cambrian (Culver, 1991; Lipps, 1992a and b). Our placement of the planktic foraminifera outside the "crown" group suggests that foraminiferal origin pre-dates the Lower Cambrian (Knoll, 1992).

Partial sequences of the SSU rRNA gene did not provide a reliable estimate of the evolutionary relationships between the four planktic foraminiferal genera used in this study. A full length SSU rDNA sequence or analyses of the LSU rRNA gene may prove more informative at the genus level. At species and intraspecies level, the unusually large expansion segments and foraminiferal specific insertions found in the planktic foraminiferal SSU rRNA gene may provide valuable information for studies of phylogenetic relationships within the group. These possibilities need further study to evaluate their full potential.

The amplification and sequencing of planktic foraminiferal genomic DNA will generate a new avenue for oceanographic research dependent on planktic foraminiferal evolution. To date, all such studies have been based on untested assumptions concerning the genotypic relationships of different morphotypes. The highly variable regions within the planktic foraminiferal SSU rDNA may prove useful for the study of the distribution of extant foraminiferal genotypes. This will only be possible following an extensive study of intra-species, morphotypic sequence variability of the SSU rDNA within and between specific biogeographical regions. Using this approach, it should be possible to determine whether genotypes are contained within water masses or whether gene flow occurs across ocean fronts. If distinct genotypes (and their respective morphotypes) are indeed associated with specific oceanic water masses, then this would permit the evolution of palaeo-water masses to be traced from the fossil record.

In summary, the gametogenic approach has been shown to be successful in providing sufficient foraminiferal genomes to enable amplification of the SSU rRNA gene. Now that

the sequences have been identified, specific primers can be developed which would target only the foraminiferal SSU gene, facilitating amplification from non-gametogenic specimens. Amplification from single specimens may be possible using this approach and this would permit the study of gene variation within and between morphotypes of a specific species. Such information could help resolve whether foraminiferal morphotypes are genotypic expressions or are generated by environmental change.

ACKNOWLEDGEMENTS

Research was supported by NERC (GR3/09736) with additional financial support provided by the Carnegie Trust. We thank H. Austin, J. Bijma and B. Huber for their assistance during the collection and culture of foraminifera at the Caribbean Marine Biological Institute (CARMABI), Curaçao. We greatly appreciate the technical assistance of P. Robertson, D. Lobidel and A. Cleland in the laboratory.

REFERENCES

Anderson, O. R., M. Spindler, A. W. H. Bé, and C. Hemleben. 1979. Trophic activity of planktonic foraminifera. Journal of the Marine Biological Association, U.K. 59:791-799.

Anderson, O. R., J. J. Lee, and W. W. Faber, Jr. 1991. Collection, maintenance and culture methods for the study of living foraminifera. *In* Lee, J. J., and O. R. Anderson. eds. Biology of Foraminifera. Academic Press, New York and London. pp. 335-357.

Bé, A. W. H. 1977. An ecological, zoogeographic and taxonomic review of recent planktonic foraminifera. *In* Ramsay, A. T. S. ed. Oceanic Micropaleontology. Academic Press, London. Volume 1, pp. 1-100.

Bé, A. W. H., O. R. Anderson, and W. W. Faber, Jr. 1983. Sequence of morphological and cytoplasmic changes during gametogenesis in the planktonic foraminifer *Globigerinoides sacculifer* (Brady). Micropaleontology **29**:310-325.

Brummer, G. J. A., and D. Kroon. 1988. Planktonic foraminifers as tracers of ocean-climate history. Free University Press, Amsterdam. pp. 6-346.

Cedergren, R., M. W. Gray, A. Abel, and D. Sankoff. 1988. The evolutionary relationships among known life forms. J. Mol. Evol. **28**:98-112.

Clark, C. G. 1992. DNA purification from polysaccharide-rich cells. *In* Lee, J. J. and A. T. Soldo. eds. Protocols in Protozoology. Society of Protozoologists, Lawrence. Volume 1, D-3.1-2.

Culver, S. 1991. Cambrian foraminifera. Science **254**:689-691.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996a. The isolation and amplification of the 18S ribosomal RNA gene from planktonic foraminifers using gametogenic specimens. *In* Whatley, R. C., and Moguilevsky, A. eds. Microfossils and Oceanic Environments. University of Wales, Aberystwyth Press. Chapter 3.1, pp. 249-259.

Faber, W. W. Jr., O. R. Anderson, and D. A. Caron. 1988. Algal foraminiferal symbiosis in the planktonic foraminifera *Globigerinella aequilateralis*. 1. Occurrence and stability of two mutually exclusive chrysophyte endosymbionts and their ultrastructure. J. Foram. Res. **18**:334-343.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783-791.

Felsenstein, J. 1993, PHYLIP, Manual version 3.52c. Berkeley University Herbarium, University of California, Berkeley.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees: A method based on mutation distances as estimated from cytochrome c sequences is of general applicability Science **155**:279-284.

Grell, K. G. 1973. Protozoology. Berlin, Heidelberg, New York. pp. 96-100.

Hemleben, C., M. Spindler, and O. R. Anderson. 1989. Modern planktonic foraminifera. Springer-Verlag, New York. pp. 1-335.

Hillis, D. M., and M. T. Dixon. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Quarterly Review Biology **66**:411-446.

Knoll, A. H. 1992. The early evolution of eukaryotes: A geological perspective. Science **256**:622-627.

Langer, M. R., J. H. Lipps, and W. E. Piller. 1993. Molecular paleobiology of protists: amplification and direct sequencing of foraminiferal DNA. Micropaleontology **39**:63-68.

Lee, J. J., W. W. Faber Jr., O. R. Anderson, and J. Pawlowski. 1991. Life-cycles in foraminifera. *In* Lee, J.J., and O. R. Anderson. eds. Biology of Foraminifera. Academic Press, New York and London. pp.285-334.

Leigh Brown, A. J., and P. Simmonds. 1994. Analysis of HIV sequence variation. *In* Karn, J. ed. HIV - a practical approach. Oxford University Press. pp.161-188.

Lipps, J. H. 1992a. Origin and early evolution of foraminifera. *In* Takayanagi, Y., and T. Saito. eds. Studies in Benthic Foraminifera. Tokai University Press, Japan. pp.3-9.

Lipps, J. H. 1992b. Proterozoic and Cambrian skeletonized protists. *In* Schopf, J. W., and C. Klein. eds. The Proterozoic Biosphere. Cambridge University Press. pp.237-242.

Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olson, K. Fogel, J. Blandy, and C. R. Woese. 1994. The ribosomal database project. Nucl. Acids Res. **22**:3484-3487.

Medlin, L. K., H. J. Elwood, S. Stickel, and M. L. Sogin. 1988. The characterization of enzymatically amplified eukaryotic 16-S like rRNA-coding regions. Gene **71**:491-499.

Merle, C., M. Moullade, O. Lima, and R. Perasso. 1994. Essai de caractérisation phylogénétique de Foraminifères planctoniques à partir de séquences partielles d'ADNr28S: C. R. Acad. Sci. Paris, **319 serie II**:149-153.

Mullis, K. B., and F. A. Faloona. 1987. Specific synthesis of DNA *in vitro* via a polymerase-catalysed chain reaction. Methods in Enzymology **155**:335-350.

Neefs, J., Y. Van de Peer, L. Hendriks, and R. De Wachter. 1990. Compilation of small ribosomal subunit RNA sequences. Nucl. Acids Res. **18**:2237-2242.

Pawlowski, J., I. Bolivar, J. Guiard-Maffia, and M. Gouy. 1994a. Phylogenetic position of Foraminifera inferred from LSU rRNA gene sequences. Mol. Biol. Evol. **11**:929-938.

Pawlowski, J., I. Bolivar, J. Fahrni, and L. Zaninetti. 1994b. Taxonomic identification of foraminifera using ribosomal DNA sequences. Micropaleontology **40**:373-377.

Saiki, R. K., D. H. Gelfland, S. Stoffel, S. J. Scharf, R. Higuchi, K. B. Mullis, and H. A. Ehrlich. 1988. Primer-directed enzymatic amplification of DNA with thermostatic DNA polymerase. Science 239:487-490.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet. 1994. The genetic data environment: an expandable GUI for multiple sequence analysis.: Comput. Appl. Biosci. 10:671-675.

Sogin, M. L. 1989. Evolution of eukaryotic microorganisms and their small subunit ribosomal RNA's. American Zoology 29:487-499.

Spindler, M., O. R. Anderson, C. Hemleben, and A. W. H. Bé. 1978. Light and electron microscopic observations of gametogenesis in *Hastigerina pelagica* (Foraminifera). Journal of Protozoology 25:427-433.

Tappan, H., and A. R. Loeblich. 1988. Foraminiferal evolution, diversification, and extinction. Journal of Paleontology 62:695-714.

Wade, C. M., K. F. Darling, D. Kroon, and A. J. Leigh Brown. 1996. Early evolutionary origin of the planktonic foraminifera inferred from SSU rDNA sequence comparisons. J. Mol. Evol. 43:672-677.

White, T. J., T. Bruns, S. Lee, and J. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* Innis, M. A., D. H. Gelfland, J. J. Sninsky, and T. J. White. eds. PCR Protocols: A guide to methods and applications. Harcourt Brace Jovanovich, San Diego. pp.315-322.

Wray, C. G., M. R. Langer, R. DeSalle, J. J. Lee, and J. H. Lipps. 1995. Origin of the foraminifera. Proc. Natl. Acad. Sci. 92:141-145.

Wray, C. G., J. L. Lee, and R. DeSalle. 1993. Extraction and enzymatic characterization of foraminiferal DNA. Micropaleontology 39:69-73.

Figure 1.

Flow diagram of the amplification and sequencing of an approximately 1000 base pair fragment of the foraminiferal SSU rRNA gene and a schematic representation of the amplified region. Amplification was performed using the "universal" primers designed by White *et al.* (1990):-

NS5: 5'-AACTTAAAGGAATTGACGGAAG-3'

NS6: 5'-GCATCACAGACCTGTTATTGCCTC-3'

NS7: 5'-GAGGCAATAACAGGTCTGTGATGC-3'

NS8: 5'-TCCGCAGGTTCACCTACGGA-3'

The relative positions of the conserved regions (black), expansion segments (grey) and foraminiferal specific insertions (white) in *G. siphonifera* (Type I) are mapped against the *Saccharomyces cerevisiae* reference SSU rDNA sequence.

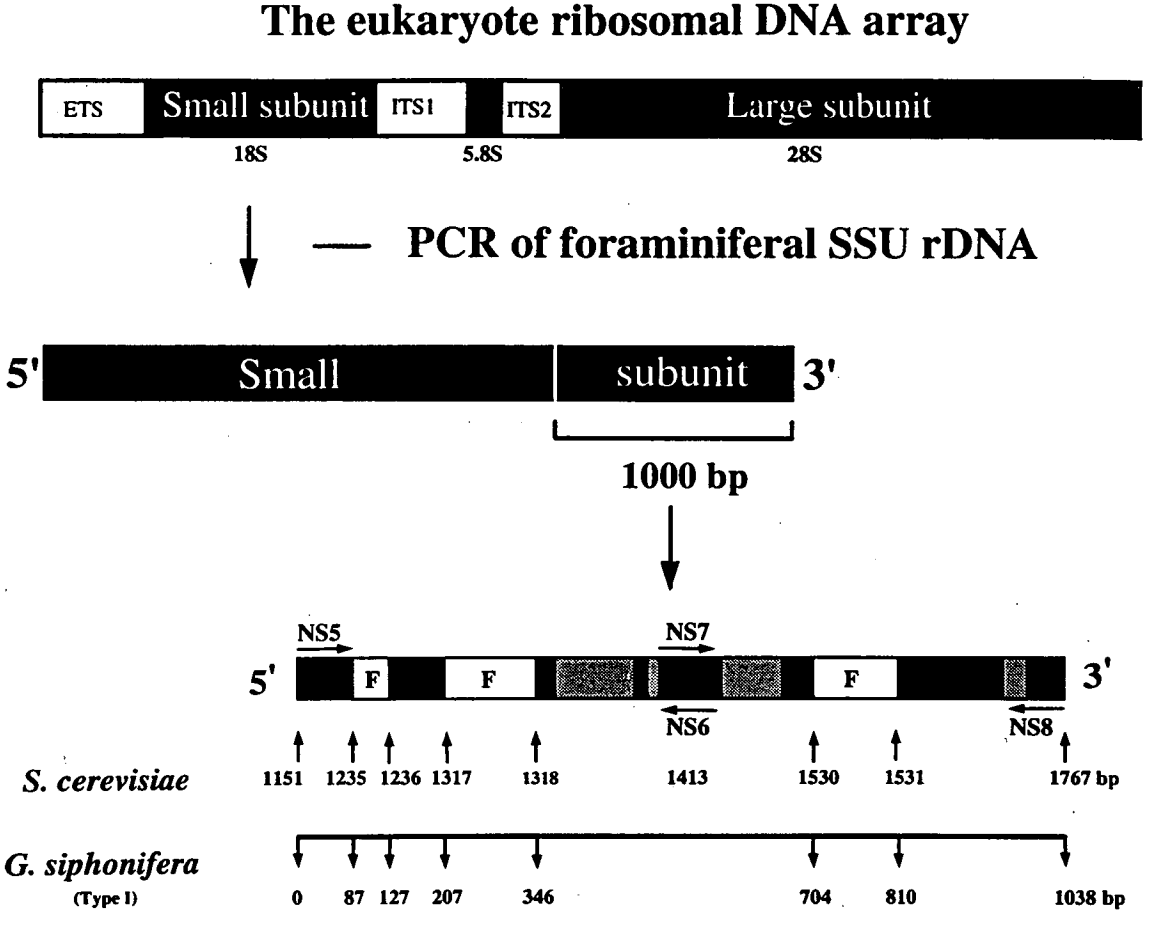
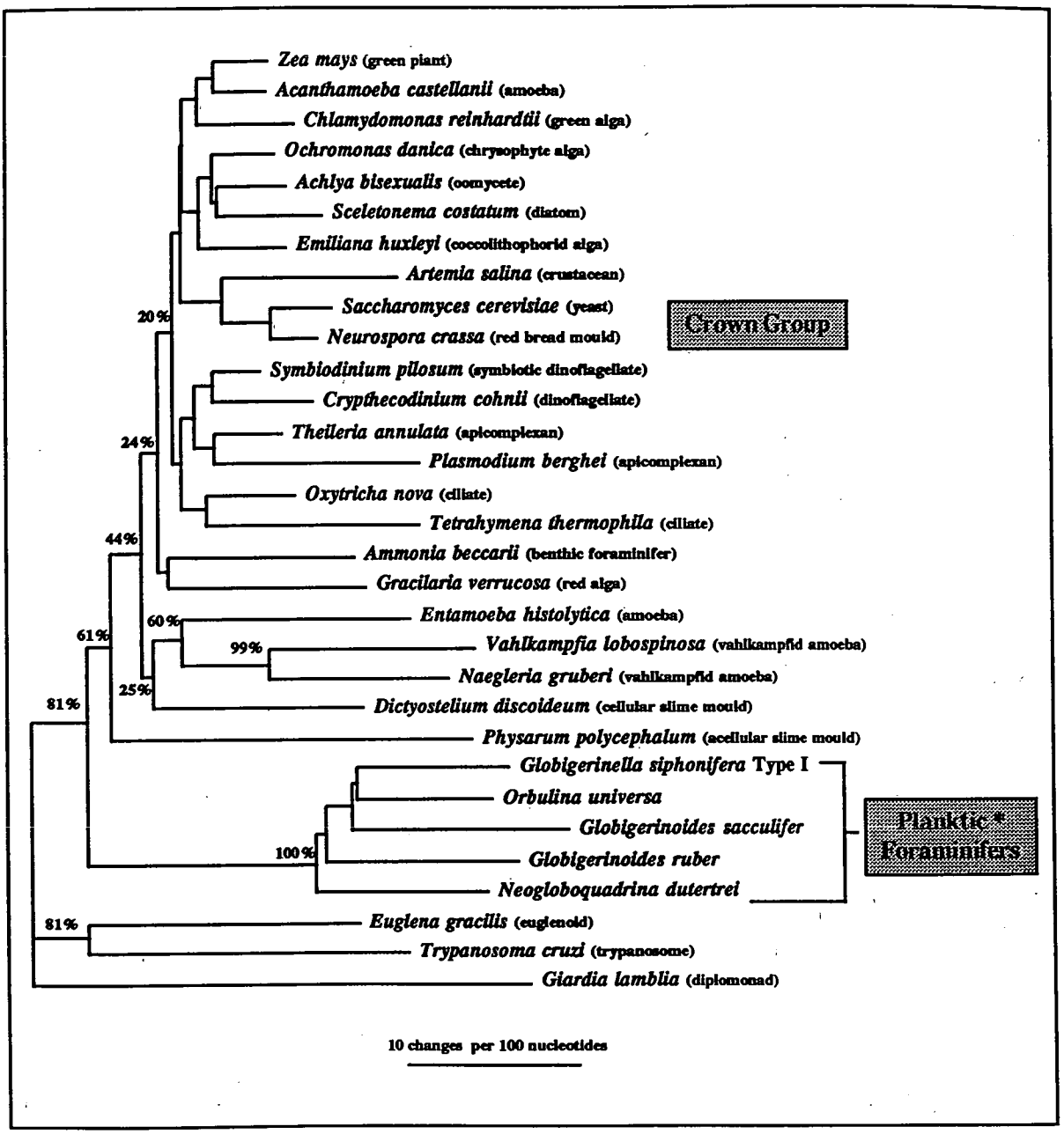

The eukaryote ribosomal DNA array
ETS
Small subunit
ITS1
ITS2
Large subunit
18S
5.8S
28S
PCR of foraminiferal SSU rDNA
5'
Small
subunit
3'
1000 bp
NS5
NS7
5'
F
F
F
3'
NS6
NS8
S. cerevisiae
1151
1235
1236
1317
1318
1413
1530
1531
1767 bp
G. siphonifera
(Type I)
0
87
127
207
346
704
810
1038 bp

Figure 2.

Phylogenetic placement of the planktic foraminifera in a eukaryotic phylogeny inferred from partial sequences (575 unambiguously aligned nucleotide sites) of the SSU rRNA gene. Branch lengths represent the evolutionary distance between organisms calculated from substitutional mutations/site. The scale bar corresponds to 10 changes per 100 nucleotide positions. Bootstrap values based on 2000 replications are expressed as a percentage. Partial sequences of the SSU rRNA gene do not provide a reliable estimate of the evolutionary relationships within the "crown" group, although the phylogenetic placement of organisms diverging prior to the "crown" conform with analyses based on full length SSU rDNA sequences. The lack of resolution within the "crown" leads to the erroneous placement of both *G. verrucosa* and the *A. beccarii* sequence of Wray *et al.* (1995). The placement of the planktic foraminiferal sequences within SSU rDNA phylogenies under a range of different methods of phylogenetic analysis is discussed in Wade *et al.* (1996).

* No support is provided for any specific planktic foraminiferal topology in preference to any other.


Zea mays (green plant)
Acanthamoeba castellanii (amoeba)
Chlamydomonas reinhardtii (green alga)
Ochromonas danica (chrysophyte alga)
Achlya bisexualis (oomycete)
Sceletonema costatum (diatom)
Emiliana huxleyi (coccolithophorid alga)
Artemia salina (crustacean)
Saccharomyces cerevisiae (yeast)
Neurospora crassa (red bread mould)
Crown Group
Symbiodinium pilosum (symbiotic dinoflagellate)
Crypthecodinium cohnii (dinoflagellate)
Theileria annulata (apicomplexan)
Plasmodium berghei (apicomplexan)
Oxytricha nova (ciliate)
Tetrahymena thermophila (ciliate)
Ammonia beccarii (benthic foraminifer)
Gracilaria verrucosa (red alga)
Entamoeba histolytica (amoeba)
Vahlkampfia lobospinosa (vahlkampfid amoeba)
Naegleria gruberi (vahlkampfid amoeba)
Dictyostelium discoideum (cellular slime mould)
Physarum polycephalum (acellular slime mould)
Globigerinella siphonifera Type I
Orbulina universa
Globigerinoides sacculifer
Globigerinoides ruber
Neogloboquadrina dutertrei
Planktic * Foraminifers
Euglena gracilis (euglenoid)
Trypanosoma cruzi (trypanosome)
Giardia lamblia (diplomonad)
20%
24%
44%
61%
60%
99%
25%
81%
100%
81%
10 changes per 100 nucleotides

Paper III

Early Evolutionary Origin of the Planktic Foraminifera Inferred from SSU rDNA Sequence Comparisons

Christopher M. Wade[1*], Kate F. Darling[2], Dick Kroon[2] and Andrew J. Leigh Brown[1]

[1] Centre for HIV Research, Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JN, UK.

[2] Department of Geology and Geophysics, Grant Institute, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JW, UK.

[*] Corresponding Author

Journal of Molecular Evolution (1996) **43**:672-677

SUMMARY

Phylogenetic analysis of five partial planktic foraminiferal small subunit (SSU) ribosomal (r) DNA sequences with representatives of a diverse range of eukaryote, archaebacterial and eubacterial taxa has revealed that the evolutionary origin of the foraminiferal lineage precedes the rapid eukaryote diversification represented by the "crown" of the eukaryotic tree, and probably represents one of the earliest splits among extant free living aerobic eukaryotes. The foraminiferal rDNA sequences could be clearly separated from known symbionts, commensals and food organisms. All five species formed a single monophyletic group distinguished from the "crown" group by unique foraminiferal specific insertions as well as considerable nucleotide distance in aligned regions.

Keywords: Planktic Foraminifera, Eukaryote Evolution, SSU rDNA Phylogeny

INTRODUCTION

The planktic foraminifera are an important group of marine zooplankton and are among the most widely used microfossils for reconstructing oceanic environments of the past (Bé 1977; Hemleben *et al.*, 1989). Attempts to determine the evolutionary origin of the foraminifera have produced contradictory results. Analysis of a small subunit (SSU) ribosomal (r) DNA sequence of the benthic foraminiferal species *Ammonia beccarii* (Linné) (Wray *et al.*, 1995) suggested a position for the foraminifera within the alveolate clade, part of the eukaryotic "crown" group of the SSU phylogenetic tree. This placement was not supported in phylogenetic trees inferred from sequences of the large subunit (LSU) rRNA gene, which placed both benthic (Pawlowski *et al.*, 1994) and planktic (Merle *et al.*, 1994) foraminiferal species early in the eukaryotic LSU tree, prior to the separation of the major eukaryote lineages which form the "crown". Our analysis of SSU rRNA gene sequences of the planktic foraminifera (Darling *et al.*, 1996b) is in strong agreement with a phylogenetic placement of the foraminifera early in eukaryote evolution, prior to the "crown" diversification.

Amplification of foraminiferal DNA is made extremely complex due to the presence of large numbers of symbionts (dinoflagellates, chrysophytes and diatoms), commensals (dinoflagellates), prey particles and adherent organisms (Langer and Lipps 1993; Darling *et al.*, 1996a; Darling *et al.*, 1996b). The potentially high number of "contaminant" genomes and the phylogenetic placement of many of these organisms within the "crown" group (SSU rDNA Tree, Ribosomal Database Project; Maidak *et al.*, 1994) may lead to the preferential amplification of contaminant rDNA using SSU rDNA "universal" primers. This situation makes the identification of foraminiferal genomic DNA from amongst the pool of additional amplification products extremely complex. Consequently, the phylogenetic placement of the foraminifera has been a subject of considerable controversy.

The rRNA genes have been used extensively in the reconstruction of eukaryote phylogenetic relationships; with the SSU rRNA gene in particular being utilised for resolving branching patterns for the early diverging eukaryote lineages (Sogin 1989, 1991; Schlegel 1991; Knoll 1992; Leipe *et al.*, 1993). Eukaryotic SSU rDNA trees indicate that early eukaryote evolution is characterised by a series of independent protistan branches, with the amitochondriate microsporidian, tritrichomonad and diplomonad lineages representing the earliest, second and third offshoots respectively (Leipe *et al.*, 1993). Eukaryotes with mitochondria are believed to have evolved from a swimming biciliated eukaryote monad with

a rigid cytoskeleton, of which the closest living relatives are the archezoan metamonad protozoa of the class Anaxostylea, represented by the diplomonad *Giardia* (Cavalier-Smith 1987). The first representatives of mitochondriate protist lineages within the eukaryotic SSU rDNA tree are the kinetoplastids and euglenoids. These are followed by the slime moulds and amoebae which precede the nearly simultaneous separation of the major eukaryotic assemblages, the animals, fungi, plants, alveolates and stramenophiles, which form the "crown" of the eukaryotic tree (Knoll 1992; Wainwright *et al.*, 1993). The placement of the early protist lineages within SSU rDNA phylogenies is nonetheless extremely complex, with rate variations and marked differences in GC content making the reliable placement of lineages extremely difficult (Loomis and Smith, 1990; Schlegel 1991; Leipe *et al.*, 1993; Yang and Roberts 1995). Protein phylogenetic analyses may prove more reliable in some cases. Such analyses have recently questioned the placement of the mitochondrion-lacking amoeba, *Entamoeba histolytica*, with analyses of elongation factor-1α suggesting an earlier placement for this species, prior to the separation of the mitochondriate protist lineages (Hasegawa *et al.*, 1993; Hashimoto *et al.*, 1994).

We have attempted to resolve the problems of foraminiferal phylogeny by utilising the SSU rRNA gene to infer the evolutionary origins of four spinose and one non-spinose planktic foraminiferal species, belonging to four separate foraminiferal genera. In order to overcome inherent problems associated with the amplification of contaminant rDNA, we have developed a procedure that enabled us to extract large quantities of foraminiferal genomic DNA, with minimal contamination, by utilising the foraminiferal gametogenic stage (Darling *et al.*, 1996a). Using this approach, coupled with the sequencing of all amplification products and their phylogenetic comparison with representatives of all known foraminiferal contaminants, we believe that we have identified foraminiferal nuclear SSU rDNA (Darling *et al.*, 1996b). Here we present a detailed phylogenetic analysis to ascertain the placement of the planktic foraminifera within the early branching eukaryote lineages.

MATERIALS AND METHODS

Foraminiferal Species

Five species of tropical planktic foraminifera were collected off the west coast of the Caribbean island of Curaçao, Dutch Antilles. These consisted of four spinose species, *Orbulina universa* d'Orbigny, *Globigerinella siphonifera* (d'Orbigny) Type I, *Globigerinoides sacculifer* (Brady) and *Globigerinoides ruber* (d'Orbigny), and the non-spinose species, *Neogloboquadrina dutertrei* (d'Orbigny). Foraminifers were cultured in the laboratory (Bé *et al.*, 1977; Hemleben *et al.*, 1987) to induce gametogenesis (Darling *et al.*, 1996a).

DNA Extraction, PCR Amplification and Sequencing

Foraminiferal genomic DNA was extracted, amplified using the polymerase chain reaction (PCR) and sequenced automatically using an Applied Biosystems 373A automatic DNA sequencer. Full details of these procedures are provided in Darling *et al.* (1996b). Polymerase chain reaction amplification was performed using "universal" eukaryotic SSU rDNA primers (White *et al.*, 1990). "Universal" eukaryotic primers were unsuccessful in amplifying the whole gene, however, amplification of a 3' terminal region, corresponding to the 30-48 region of the eukaryotic SSU rRNA secondary structure model (Neefs *et al.*, 1990), was possible using the primers NS5 and NS8 (White *et al.*, 1990).

Sequence Analysis

The partial planktic foraminiferal SSU rDNA sequences were aligned relative to 438 eukaryote taxa, 3 archaebacterial taxa and 3 eubacterial taxa (obtained from the Ribosomal Database Project; Maidak *et al.*, 1994). Alignment was performed using the Genetic Data Environment (GDE) package (Smith *et al.*, 1994). Foraminiferal phylogenetic trees were reconstructed using 35 representatives of a diverse range of eukaryotes (including known foraminiferal symbionts and prey organisms) and 6 prokaryotes. Five hundred and forty six unambiguously aligned nucleotide sites were used. Phylogenetic analyses were performed with four different methods using programs taken from version 3.52c of the Phylogeny Inference Package (PHYLIP) (Felsenstein 1993). Distance based phylogenetic analyses were performed using both the neighbour-joining (Saitou and Nei 1987) (program NEIGHBOR) and Fitch-Margoliash (Fitch and Margoliash 1967) (program FITCH) methods, with nucleotide sequence distances calculated for all pairwise sequence comparisons using the

generalised two-parameter (maximum likelihood) model (program DNADIST). Maximum likelihood (Felsenstein 1981) phylogenetic analyses were performed using the modified PHYLIP program FASTDNAML (kindly provided by Gary Olsen of the University of Illinois at Urbana-Champain and the Ribosomal Database Project), and maximum parsimony (Fitch, 1971) analyses were performed using the program DNAPARS. Bootstrap resampling (Felsenstein 1985) (programs SEQBOOT and CONSENSE) was employed to assign support to the neighbour-joining (2000 replicates) and Fitch-Margoliash (100 replicates) trees. Alternative phylogenetic hypotheses were evaluated statistically by a likelihood ratio test (Kishino and Hasegawa 1989) following the assignment of log likelihoods (program DNAML) to artificially generated topologies (program RETREE).

Nucleotide Sequence Accession Numbers

Foraminiferal nucleotide sequences reported in this study have been assigned GenBank accession numbers U65631 to U65635.

RESULTS

Polymerase chain reaction amplification of the foraminiferal DNA extract using primers NS5/NS8 resulted in the generation of two distinct amplification products, indicating the presence of contaminants within the sample. Sequencing was originally focused on the smaller amplification products, which corresponded in size with the expected eukaryotic NS5/NS8 sequence length. However, phylogenetic analyses of these fragments, amplified from a number of different foraminiferal species, revealed an association with a known foraminiferal symbiont, commensal or prey organism in the majority of cases (data not shown). Furthermore, no evidence of monophyly was observed for any of the foraminiferal species for the smaller amplification products, with the sequences being distributed throughout the eukaryotic "crown" group. In addition, we did not identify any sequence which clustered with the published sequence for the benthic foraminifera, *A. beccarii* (Wray *et al.*, 1995). Preliminary sequencing of the larger amplification product revealed a high degree of similarity between the five planktic foraminiferal species. Subsequent phylogenetic analysis has shown that the sequences derived from this amplification product form a monophyletic group and do not cluster with any known foraminiferal symbiont, commensal or prey particle (Darling *et al.*, 1996b). This strongly suggests that the larger amplification product is foraminiferal in origin.

We have sequenced an approximately 1000 base pair (bp) region from the 3' terminus of the SSU rRNA gene for five species of tropical planktic foraminifera. Comparisons of the foraminiferal sequences with 438 full length published SSU rDNA sequences (Ribosomal Database Project database; Maidak *et al.*, 1994) revealed that the foraminiferal amplification products were considerably larger than the corresponding regions for any other eukaryote taxa. The length of the amplified foraminiferal sequences varied between 980 bp for *O. universa* and 1048 bp for *G. siphonifera* Type I in comparison with a typical eukaryotic sequence length of 618 bp for the homologous region of the chrysophyte, *Ochromonas danica*. All foraminiferal sequences were found to contain four variable length expansion segments and three insertions which were unique to the foraminifera. These regions showed considerable length and sequence variability, even within the foraminifera, with several regions displaying a level of diversity great enough to prevent their alignment.

Phylogenetic analysis of the foraminiferal sequences with 35 representatives of the diverse range of eukaryote taxa, including symbionts and prey organisms, revealed that the

planktic foraminiferal sequences formed a clearly distinct monophyletic group (Fig. 1). This group was supported in 100% of 2000 neighbour-joining and 100% of 100 Fitch-Margoliash bootstrap replications and resolved consistently with all four phylogeny reconstruction methods employed. Likelihood tests of alternative "polyphyletic" phylogenetic hypotheses were carried out in which the branch to the outlying foraminifera, *N. dutertrei*, was located at two positions outside the foraminiferal group. These also showed a high level of support for the monophyly of the foraminiferal group ($p<0.01$) (Test 1; Table 1). While analysis of the SSU rDNA sequences presented here did not provide a reliable estimate of the evolutionary relationships between the planktic foraminiferal genera we are currently undertaking additional analyses to resolve the molecular taxonomic relationships of a number of planktic foraminiferal species (Darling *et al.*, in preparation). The planktic foraminiferal sequences showed considerable evolutionary distances, reflected in the lengths of the branches to the extant foraminiferal genera from their common ancestors. Such high levels of sequence diversity are a characteristic feature of the early diverging eukaryote lineages.

The planktic foraminiferal sequences consistently branched early in the tree, falling between the diplomonads and the euglenoids and kinetoplastids. The precise location of the foraminiferal branch was not however supported statistically (Fig. 1), with relatively short branch lengths between internodes and low bootstrap support for the branches separating the highly divergent early protist lineages. Nevertheless, the location outside the "crown" group was supported in over 99% of bootstrap replicates for both the neighbour-joining and Fitch-Margoliash trees (excluding sequences intermediate between the foraminiferal branch and the eukaryote "crown" group) as well as in likelihood ratio tests ($p<0.05$) (Test 2; Table 1). With the exception of parsimony, all tree construction methods employed resolved an identical placement for the planktic foraminifera and were in close agreement with the full length SSU rDNA tree of Leipe *et al.* (1993) in their placement of the protist lineages diverging before the "crown". Furthermore, phylogenetic trees reconstructed with the amitochondriate eukaryote lineages removed consistently placed the foraminiferal branch before the euglenoid/kinetoplastid branch. Inconsistencies in the parsimony analysis in the placement of the protist lineages might be explained by an insufficient number of available sites given the sensitivity of this method to unequal rates (Felsenstein 1978). Nevertheless, parsimony analysis was still congruent in its placement of the foraminiferal sequences outside of the "crown" diversification. The GC content of the foraminiferal sequences would not be expected to bias the inferred placement of the foraminifera. Foraminiferal GC content was relatively

homogeneous, varying between 47% for *N. dutertrei* and 50% for *G. ruber*. Overall, the mean GC content of all sequences in the phylogenetic analysis was 51%, but unusual GC contents are observed in SSU rDNA sequences of the diplomonad *Giardia lamblia* (75% GC) and the microsporidian *Vairimorpha necatrix* (35% GC). The inclusion of rDNA sequences of the diplomonad species, *Hexamita inflata* (54% GC), a specific relative of *G. lamblia*, has been shown to provide a more reliable estimate of the phylogenetic position of the diplomonads (Leipe *et al.*, 1993). Thus rDNA sequences of *H. inflata* and the microsporidian *Encephalitozoon cuniculi* (52% GC) were included in these analyses.

The published SSU rDNA sequence for the benthic foraminifera *A. beccarii* (Wray *et al.*, 1995), was clearly distinct from the planktic foraminiferal sequences presented here. In our analysis of partial SSU rDNA sequences, this sequence clustered with that of the red alga, *Gracilaria verrucosa*, just outside the "crown" group. We nevertheless concur with the Wray *et al.* (1995) placement of the *A. beccarii* sequence within the "crown" based upon analysis of the full length gene. With the exclusion of sequences intermediate between the foraminiferal branch and the *G. verrucosa/A. beccarii* cluster, the branch separating the planktic foraminifera from *A. beccarii* was supported in 99.8% of 2000 neighbour-joining and 100% of 100 Fitch-Margoliash bootstrap replicates. Furthermore, attempts to cluster the *A. beccarii* sequence as an outlier to the planktic foraminiferal group resulted in a significantly worse topology ($p<0.05$) (Test 3; Table 1).

DISCUSSION

Sequence analysis of foraminiferal DNA is rendered substantially more complex by the association of many non-foraminiferal nuclei with the organism. Our strategy of amplifying gametogenic stage individuals substantially reduces the level of contamination and markedly increases the ratio of foraminiferal genomic DNA to that of contaminants. Nevertheless, PCR amplification produced amplification products of different lengths. Phylogenetic analysis of the smaller amplification products revealed close associations with specific symbiont, commensal or food organisms. Only the larger amplification product sequences clustered within a monophyletic group when analysed in a background of eukaryotic SSU rDNA sequences.

There are three lines of evidence which favour the sequences analysed here as originating from foraminiferal nuclear genomes. Firstly, they are all structurally distinguished from other eukaryotic sequences by the inclusion of large foraminiferal specific insertions at common regions within the foraminiferal SSU rRNA molecule. Secondly, phylogenetic analysis of the conserved regions of the larger foraminiferal amplification product sequences with those from other eukaryotes has shown that these and only these sequences give a monophyletic grouping for the planktic foraminifera, as would be expected for this taxonomically distinct group. Thirdly, the placement of the planktic foraminiferal SSU sequences prior to the "crown" diversification is congruent with the placement of both benthic and planktic foraminiferans inferred by Pawlowski *et al.* (1994) and Merle *et al.* (1994) from analyses of partial sequences of the LSU gene. Again this would be expected from current taxonomic understanding of these groups. While a formal possibility remains that a previously unknown early eukaryotic organism, symbiotic with both benthic and planktic foraminiferans, has given rise to all these sequences, we contend that the absence of an alternative candidate sequence meeting these criteria leaves the early origin of the foraminifera as the most reasonable interpretation of the molecular data. An SSU rDNA sequence obtained by Wray *et al.* (1995) from the benthic foraminiferan *A. beccarii* does not group with the planktic foraminifers in our tree. In contrast, an LSU sequence obtained by Pawlowski *et al.* (1994) from the same species was placed by them together with other benthics in the same position relative to the "crown" group as we have placed the planktics. In our view the likely explanation is that the SSU sequence obtained by Wray *et al.* (1995) is not foraminiferal in origin.

The actual divergence inferred from the phylogenetic analysis is one of the earliest known evolutionary divergences amongst free-living aerobic eukaryotes. We have inferred a phylogenetic placement for the foraminiferal branch between the amitochondriate microsporidian lineage and the mitochondriate kinetoplastids and euglenoids. Nevertheless, the phylogenetic placement of lineages which branch in the middle of the tree is unclear. The foraminiferal lineage does however fall significantly outside of the "crown" group indicating conclusively that this lineage originated before the separation of the major eukaryotic lineages, approximately 1000 to 1100 Myr ago (Sogin 1991; Knoll 1992).

The precise location of the foraminiferal branch within the global eukaryotic SSU rDNA phylogeny may possibly be resolved more reliably by the analysis of a complete length foraminiferal SSU rDNA sequence. Additionally, analyses of conserved protein-coding sequences such as the elongation factor genes (Hasegawa *et al.*, 1993; Hashimoto *et al.*, 1994), may prove informative in the placement of the foraminifera. Analysis of mitochondrial sequences, when it is feasible, will also provide further valuable information about the relationships among the early eukaryotes and the evolution of the mitochondria in general.

ACKNOWLEDGEMENTS

Research was supported by NERC (GR3/09736) with additional financial support provided by the Carnegie Trust. We thank Heather Austin, Jelle Bijma and Brian Huber for their assistance during the collection and culture of foraminifers at the Caribbean Marine Biological Institute (CARMABI), Curaçao. We also greatly appreciate the assistance of Denis Lobidel.

REFERENCES

Bé, A. W. H. 1977. An ecological, zoogeographic and taxonomic review of recent planktonic foraminifera. *In* Ramsay A. T. S. ed. Oceanic micropaleontology. Academic Press, New York and London. pp.1-100.

Bé, A. W. H., C. Hemleben, O. R. Anderson, M. Spindler, J. Hacunda, and S. Tuntivate-Choy. 1977. Laboratory and field observations of living planktonic foraminifera. Micropaleontology **23**:155-179.

Caron, M., and P. Homewood. 1983. Evolution of early planktic foraminifers. Marine Micropaleontology 7:453-462.

Cavalier-Smith, T. 1987. Eukaryotes with no mitochondria. Nature **326**:332-333.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996a. The isolation and amplification of the 18S ribosomal RNA gene from planktonic foraminifers using gametogenic specimens. *In* Whatley, R. C., and Moguilevsky, A. eds. Microfossils and Oceanic Environments. University of Wales, Aberystwyth Press. Chapter 3.1, pp. 249-259.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996b. Molecular evolution of planktic foraminifera. J. Foram. Res. **26**:324-330.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401-410.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein, J. 1993. PHYLIP manual version 3.52c. Berkeley University Herbarium,

University of California, Berkeley.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science 155:279-284.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. 20:406-416.

Hasegawa, M., T. Hashimoto, J. Adachi, N. Iwabe, and T. Miyata. 1993. Early branchings in the evolution of eukaryotes: ancient divergence of Entamoeba that lacks mitochondria revealed by protein sequence data. J. Mol. Evol. 36:380-388.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Ojkamoto, and M. Hasegawa. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. Mol. Biol. Evol. 11:65-71.

Hemleben, C., M. Spindler, I. Breitinger, and R. Ott. 1987. Morphological and physiological responses of *Globigerinoides sacculifer* (Brady) under varying laboratory conditions. Marine Microaleontol. 12: 305-324.

Hemleben, C., Spindler, M., and O. R. Anderson. 1989. Modern planktonic foraminifera. Springer-Verlag, New York.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order of the Hominoidea. J. Mol. Evol. 4:406-425.

Knoll, A. H. 1992. The early evolution of eukaryotes: a geological perspective. Science 256:622-627.

Leipe, D. D., J. H. Gunderson, T. A. Nerad, and M. L. Sogin. 1993. Small subunit ribosomal

RNA^{+} of *Hexamita inflata* and the quest for the first branch in the eukaryottic tree. Mol. Biochem. Parasit. **59**:41-48.

Loomis, W. F, and D. W. Smith. 1990. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc. Natl. Acad. Sci. USA. **87**:9093-9097.

Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olsen, K. Fogel, J. Blandy, and C. R. Woese. 1994. The ribosomal database project. Nucleic Acids Research **22**:3484-3487.

Merle, C., M. Moullade, O. Lima, and R. Perasso. 1994. An attempt to phylogenetically characterise some planktonic foraminifers on the basis of 28S rDNA partial sequences. Comptes rendus de l'Academiedes Sciences Serie II Sciences del la Terre et des Planetes, **319**:149-153.

Neefs, J., Y. Van de Peer, L. Hendriks, and R. De Wachter. 1990. Compilation of small ribosomal subunit RNA sequences. Nucl. Acids. Res. **18**:2237-2242.

Pawlowski, J., I. Bolivar, J. Guiard-Maffia, and M. Gouy. 1994. Phylogenetic position of foraminifera inferred from LSU rDNA sequences. Mol. Biol. Evol. **11**:929-938.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing evolutionary trees. Mol. Biol Evol. **4**:406-425.

Schlegel, M. 1991. Protist evolution and phylogeny as discerned from small subunit ribosomal RNA sequence comparisons. Europ. J. Protistol. **27**:207-219.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, P. M. Gillevet. 1994. The genetic data environment and expandable GUI for multiple sequence analysis. Comput. Appl. Biosci. **10**:671-675.

Sogin, M. L. 1989. Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. Amer. Zool. **29**:487-499.

Sogin, M. L. 1991. Early evolution and the origin of eukaryotes. Curr. Opin. Genet. Dev. 1:457-463.

Wainwright, P. O., G. Hinkle, M. L. Sogin, and S. K. Stickel. 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. Science 260:340-342.

White, T. J., T. Bruns, S. Lee, and J. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* Innis, M. A., D. H. Gelfand, J. J. Sninsky, and T. J. White. eds. PCR protocols: a guide to methods and applications. Harcourt Brace Jovanovich, San Diego, pp.315-322.

Wray, C. G., M. R. Langer, R. DeSalle, J. J. Lee, and J. H. Lipps. 1995. Origin of the foraminifera. Proc. Natl. Acad. Sci. 92:141-145.

Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451-458.

Table 1.

Bootstrap and likelihood support for alternative phylogenetic hypotheses for the placement of the planktic foraminifera. Bootstrap values, expressed as a percentage, represent the support for a particular hypothesis per 2000 neighbour-joining or 100 Fitch-Margoliash (in brackets) bootstrap replications. Likelihood evaluation of alternative phylogenetic hypotheses was performed using a likelihood ratio test (Kishino and Hasegawa, 1989). All hypotheses tested were based on the reconstructed neighbour-joining tree.

** Significantly worse than alternative hypotheses ($p<0.01$)

* Significantly worse than alternative hypotheses ($p<0.05$)

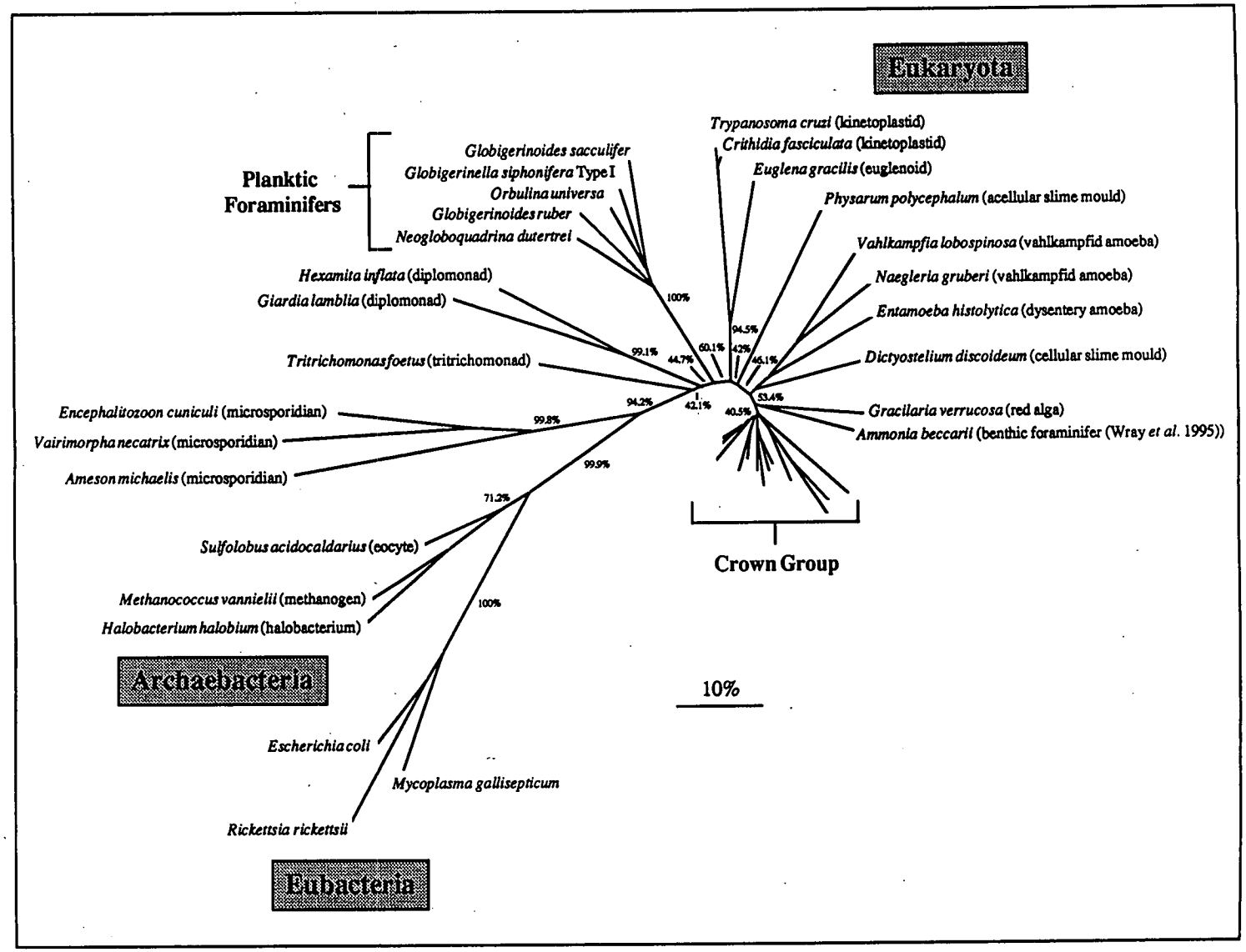
+ Bootstrap value excludes sequences intermediate between the foraminiferal branch and the eukaryote "crown" group.

++ Bootstrap value excludes sequences intermediate between the foraminiferal branch and the *G. verrucosa*/*A. beccarii* branch.

Competing Hypotheses	Bootstrap Support	Log Likelihood
Test 1		
Planktic foraminifera form a monophyletic group	100% (100%)	-12694.082
VS		
Planktic foraminifera are polyphyletic		
a) *N. dutertrei* placed after *T. cruzi/C. fasciculata/ E. gracilis* branch	-	-12819.638 (SD: 19.955)**
b) *N.dutertrei* placed before *G. lamblia/H. inflata* branch	-	-12804.421 (SD: 19.372)**
Test 2		
Planktic foraminifera fall outside the crown group	99.9% (100%)+	-12694.082
VS		
Planktic foraminifera fall within the crown group	-	-12722.030 (SD: 13.394)*
Test 3		
Benthic foraminifer, *A. beccarii* (Wray *et al.* 1995), branches distinctly from the planktic foraminifera	99.8% (100%)++	-12694.082
VS		
Benthic foraminifer, *A. beccarii* (Wray *et al.* 1995), clusters with the planktic foraminifera to form a monophyletic foraminiferal group (*A. beccarii* placed on the planktic foraminiferal branch)	-	-12724.604 (SD: 14.371)*

Figure 1.

Small Subunit rDNA phylogeny for the planktic foraminifera. The phylogenetic tree was reconstructed by neighbour-joining analysis of 546 unambiguously aligned nucleotide sites. The scale bar corresponds to 10 changes per 100 nucleotide positions. Bootstrap values, based on 2000 replications, are expressed as a percentage. The "crown" group taxa correspond to *Symbiodinium pilosum* (symbiotic dinoflagellate), *Theileria annulata* (apicomplexan), *Tetrahymena thermophila* (ciliate), *Ochromonas danica* (chrysophyte), *Achlya bisexualis* (oomycete), *Sceletonema costatum* (diatom), *Emiliania huxleyi* (prymnesiophyte), *Chlamydomonas reinhardtii* (green alga), *Zea Mays* (maize), *Acanthamoeba castellanii* (acanthamoeba), *Saccharomyces cerevisae* (yeast), *Artemia salina* (brine shrimp) and *Homo sapiens* (human).


Eukaryota
Trypanosoma cruzi (kinetoplastid)
Crithidia fasciculata (kinetoplastid)
Euglena gracilis (euglenoid)
Physarum polycephalum (acellular slime mould)
Vahlkampfia lobospinosa (vahlkampfid amoeba)
Naegleria gruberi (vahlkampfid amoeba)
Entamoeba histolytica (dysentery amoeba)
Dictyostelium discoideum (cellular slime mould)
Gracilaria verrucosa (red alga)
Ammonia beccarii (benthic foraminifer (Wray et al. 1995))
Planktic Foraminifers
Globigerinoides sacculifer
Globigerinella siphonifera Type I
Orbulina universa
Globigerinoides ruber
Neogloboquadrina dutertrei
Hexamita inflata (diplomonad)
Giardia lamblia (diplomonad)
Tritrichomonas foetus (tritrichomonad)
Encephalitozoon cuniculi (microsporidian)
Vairimorpha necatrix (microsporidian)
Ameson michaelis (microsporidian)
Sulfolobus acidocaldarius (eocyte)
Methanococcus vannielii (methanogen)
Halobacterium halobium (halobacterium)
Archaebacteria
Escherichia coli
Mycoplasma gallisepticum
Rickettsia rickettsii
Eubacteria
Crown Group
10%
100%
99.1%
44.7%
60.1%
42.1%
94.5%
42%
46.1%
53.4%
40.5%
94.2%
99.8%
99.9%
71.2%
100%

Paper IV

Planktic Foraminiferal Molecular Evolution and their Polyphyletic Origins from Benthic Taxa

Kate F. Darling[1*], Christopher M. Wade[2], Dick Kroon[1], and Andrew J. Leigh Brown[2]

[1] Department of Geology and Geophysics, Grant Institute, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JW, UK.

[2] Centre for HIV Research, Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JN, UK.

[*] Corresponding Author

Marine Micropaleontology (in press)

ABSTRACT

Phylogenetic analyses based on partial sequences of the small subunit (SSU) ribosomal (r) RNA gene have shown that the planktic and benthic foraminifera form a distinct monophyletic group within the eukaryotes. In order to determine the evolutionary relationships between benthic and planktic foraminifers, representatives of spinose and non-spinose planktic genera have been placed within a molecular SSU rDNA phylogeny containing sequences of the benthic suborders available to date. Our phylogenetic analysis shows that the planktic foraminifers are polyphyletic in origin, not evolving solely from a single "globigerinid like" lineage in the Mid-Jurassic, but derived from at least two ancestral benthic lines. The benthic ancestor of *Neogloboquadrina dutertrei* possibly entered the plankton later than the Mid-Jurassic, and further investigation of related extant species may provide an indication of the timing of this event. The evolutionary origin of the non-spinose species *Globorotalia menardii* remains unclear. The divergences of the planktic spinose species generally support recent phylogenies based on the fossil record, which infer a radiation from a globigerinid common ancestor in the Mid to Late Oligocene. The branching pattern indicates that there are possibly four distinct groups within the main spinose clade, with large evolutionary distances being observed between them. *Globigerinoides conglobatus* clusters strongly with *Globigerinoides ruber* and are divergent from *Globigerinella siphonifera, Orbulina universa* and *Globigerinoides sacculifer.*

Conserved regions of the SSU rRNA gene show sufficient variation to discriminate foraminifers at the species level. Large genetic differences have been observed between the pink and white forms of *Gs. ruber* and between *Ge. siphonifera* Type I and II. The two types of *Ge. siphonifera* cannot be discriminated by traditional paleontological methods, which has considerable implications for tracing fossil lineages and for the estimation of molecular evolutionary rates based upon the fossil record. The conserved regions show a high degree of sequence identity within a species, providing signature sequences for species identification. The variable regions of the gene may prove informative for population level studies in some species although complete sequence identity was observed in *G. sacculifer* and *O. universa* between specimens collected from the Caribbean and Western Pacific.

INTRODUCTION

Molecular phylogenetic analysis of partial small subunit (SSU) ribosomal (r) DNA sequences amplified from four planktic foraminiferal genera, has shown that they form a distinct monophyletic group within the eukaryotic phylogeny (Darling *et al.*, 1996b). The evolutionary origin of the planktic foraminifers precedes the rapid eukaryote diversification ("crown" group) in the Cambrian (Sogin, 1989; Wainwright *et al.*, 1993; Knoll, 1992), and probably represents one of the earliest divergences amongst extant free living eukaryotes (Darling *et al.*, 1996b; Wade *et al.*, 1996b; Figure 1). Molecular phylogenetic analysis of benthic foraminifers from both SSU and large subunit (LSU) rDNAs has independently indicated an early origin for the foraminifera (Pawlowski *et al.*, 1994 and 1996).

It has been suggested that the earliest planktic foraminifers evolved from a single benthic lineage in the Mid-Jurassic or earlier (Loeblich and Tappan, 1974; Caron and Homewood, 1983). Subsequent planktic radiations following major extinctions are thought to have evolved from surviving planktic forms (Tappan and Loeblich, 1988; Norris, 1991; Olsson, 1992) rather than from new adaptation to the planktic habitat from the benthos. In order to determine the precise evolutionary relationships between the planktic and benthic foraminiferal genera, the planktic species have been placed phylogenetically within a background of benthic species, representing a diverse range of benthic sub-orders. This analysis incorporates thirteen representatives of six different planktic foraminiferal genera, including both spinose and non-spinose species, into a SSU rDNA phylogeny with representatives of five benthic foraminiferal sub-orders.

This extended phylogenetic analysis will enable evaluation of the evolutionary relationships between extant species and morphotypes of planktic foraminifers and permit the level of variation associated with different taxonomic levels to be assessed. The extensive knowledge of the morphological relationships between the foraminifera, based upon the fossil record, will allow a direct comparison between interpretations of the fossil record and relationships inferred from the molecular tree. Current classifications of the planktic foraminifera are based largely upon the placement of species within evolutionary lineages exhibiting gradational morphological transitions (Kennett and Srinivasan, 1983; Pearson, 1993). The present study will enable us to determine whether the current morphological grouping of species into genera is reflected in the inferred relationships from the molecular tree.

The determination of the evolutionary distances between foraminiferans may permit more precise differentiation of foraminiferal species and their morphotypes. This is of great importance for palaeoceanographic studies. Close examination of morphologically similar individuals, defined as being within a single species, indicates considerable variation in form. Such morphotypic variability within a species is considered by many workers to be environmentally controlled (Kennett, 1976; Kroon *et al.*, 1988; Williams, *et al.*, 1988). This assumption has prompted the use of morphotypic variability in palaeoceanographic studies as an indicator of environmental change, without any direct evidence that morphotypic variability is solely a function of the environment. Molecular phylogenetic analysis can establish the genetic relationships between morphologically related forms and help clarify the true relationship between phenotypic variability and the environment. However, morphotypes of the same species are unlikely to show sufficient variation within the conserved regions of the SSU rRNA gene to permit their differentiation. The region of the foraminiferal SSU rRNA gene chosen for this study consists of eight conserved regions, used for phylogenetic analysis, interspersed by seven variable regions (Darling *et al.*, 1996b). The variable regions may provide additional information for an investigation of the closer relationships between foraminiferal populations and morphotypic variants.

MATERIALS AND METHODS

Collection and Culture of Planktic Foraminifers

Nine species of planktic foraminifers were collected from two tropical localities. The non-spinose species *Globorotalia menardii* (d'Orbigny) and *Neogloboquadrina dutertrei* (d'Orbigny) and the spinose species *Globigerinoides sacculifer* (Brady), *Globigerinoides ruber* (d'Orbigny) "pink" form, *Orbulina universa* d'Orbigny and *Globigerinella siphonifera* (d'Orbigny) Type I and Type II were collected off the west coast of the Caribbean island of Curaçao and cultured onshore at the Caribbean Marine Biological Institute (CARMABI) as described in Darling *et al.*, (1996a). The spinose species *Globigerina bulloides* d'Orbigny, *Globigerinoides conglobatus* (Brady), *Gs. sacculifer*, *Gs. ruber* "white" form, *O. universa* and *Ge. siphonifera* Type II were collected off the Great Barrier Reef (GBR), Australia, 0.8 nautical miles due east of Ribbon Reef 10. They were cultured onshore at Lizard Island Research Station located in the Cairns section of the GBR.

Planktic foraminiferal specimens were collected either by drift net at an approximate depth of 5m or individually by SCUBA divers at a depth of 3-8m. Taxonomic identification was confirmed using a stereo microscope or an inverted compound microscope as described in Darling *et al.*, (1996a). *Globigerinella siphonifera* Type I and Type II are quite distinct, and the differences between them are outlined in Faber *et al.*, (1988) and Huber *et al.*, (submitted). They have a characteristic spine, rhizopodial and symbiont distribution which can be used to differentiate them. *Globigerinella siphonifera* Type I has a light brown test with a highly anastomosing network of rhizopodia extending outwards to approximately the diameter of the foraminiferal test with the symbionts distributed throughout the network. In contrast, the test of Type II has a dark brownish appearance and the spines generally radiate in two tufts on opposite sides of the test. The rhizopods of Type II are aligned along the spines and the symbionts dispersed far out from the test. Identification of *G. bulloides* was confirmed during culture. *Globigerina bulloides* did not possess algal symbionts, clearly separating it from *Globigerina falconensis* Blow, which is known to be symbiont bearing (Hemleben *et al.*, 1989). Further confirmation was made by stereomicroscopic examination following spine loss prior to gametogenesis. The exposed tests possessed the wide aperture of *G. bulloides* with no diminutive final chamber or apertural lip characteristic of the species *G. falconensis* (Malmgren and Kennett, 1977). As *G. bulloides* is typically a transitional/ polar species and may be distinct from the semi-tropical species, it will be referred to

throughout the present text as *Globigerina sp.*

All species were cultured in a controlled environment culture chamber until they reached an advanced stage of gametogenesis as described in Darling *et al.* (1996a), except for the GBR specimens of *O. universa* and *Gs. conglobatus.* Amplification of DNA was achieved in these species from single non-gametogenic specimens. All gametogenic samples were frozen at -25°C prior to gamete release. The samples from the Caribbean were transported in dry ice to the UK and preserved at -70°C before DNA extraction. The samples from the GBR were transported to the mainland in dry ice where DNA extractions were carried out at James Cook University, Townsville.

DNA Extraction

Nucleic acid extractions were carried out as outlined in Darling *et al.*, (1996a; 1996b). Strict procedures were observed in the laboratory to prevent contamination during extraction.

The SSU rRNA gene and its choice for use in molecular systematic studies

The principal aim of this investigation is to carry out molecular systematic studies of closely related species and morphotypes. The limited quantities of genomic DNA and problems with contaminants rule out the use of a standard "fingerprinting" approach (Darling *et al.*, 1996a). The PCR technique however, allows the rapid amplification of large quantities of specific sequences of DNA, and this approach has been used to target the ribosomal RNA genes, present in multiple copies within the genome.

The ribosome is a cytoplasmic organelle present in all living cells, consisting of three RNA subunits and protein. Proteins are synthesised within the ribosome by translating the messenger RNA code derived from the DNA within the nucleus. The genes encoding the subunits of the ribosome are highly conserved, as they are required for the maintenance of the integrity of the ribosomal structure. As a consequence of their high level of conservation, the rRNA genes, in particular, have been used to create a global phylogeny ("tree of life") from representatives of a cross-section of all living organisms. Studies of the SSU rRNA gene has shown that the foraminifera form one of the earliest of all eukaryote groups sequenced to date (Pawlowski *et al.*, 1996; Wade *et al.*, 1996; Figure 1). The SSU rRNA gene sequences contain sufficient evolutionary information to allow the determination of both close and distant evolutionary relationships, as the conserved regions are interspersed by more variable regions. These have been used to provide additional evolutionary information for phylogenetic

analyses within the foraminiferal group.

PCR Amplification and Sequencing

The PCR amplification of an approximately 1000bp region of the terminal 3' end of the foraminiferal SSU rRNA gene (equivalent position 1151-1767 in the *Saccharomyces cerevisiae* complete SSU rDNA sequence) was carried out using the "universal" primers of White *et al.*, (1990), as described in Darling *et al.*, (1996a; 1996b). A sense primer (FS3), specific to our planktic foraminiferal sequences, was designed for sequencing across the "universal" primer positions (FS3; 5'- GTGATCTGTCTGCTTAATTGC - 3', equivalent to base positions 206-227 in *Ge. siphonifera* Type I). A direct, enzymatic, "cycle sequencing" method (see Leigh Brown and Simmonds, 1994) was used to sequence the amplified fragments. The products of the sequencing reaction were run on an acrylamide gel within an Applied Biosystems 373A automated sequencing system. Complementary strands were sequenced for cross checking each base position and each strand was duplicated. To confirm sequence uniformity within a single species, a within species sequence investigation was carried out on *Ge. siphonifera* Type I.

In all species examined in this investigation, no multiple base calling was observed from any amplification product, indicating that cross species contamination had not occurred in any sample. If morphotypically similar but genetically distinct individuals had been extracted together in the same sample, multiple base calling would immediately indicate the presence of mixed templates. To confirm that the *N. dutertrei* sequences were not benthic foraminiferal contaminants, *N. dutertrei* specimens were also collected off the GBR, Australia. The partial sequences amplified showed complete identity with those obtained in the Caribbean (unpublished data).

Sequence Analysis

Partial SSU rDNA sequences of 7 Caribbean planktic foraminifera, 6 Australian planktic foraminifera and 15 published benthic foraminifera (Pawlowski *et al.*, benthic species GenBank accession nos: *Haynesina germanica* Z69615; *Textularia* sp. Z69610; *Trochammina hadai* Z69612; *Peneroplis pertusus* Z69604; *Archaias angulatus* Z69603; *Quinqueloculina* sp. Z69605; *Glabratella opercularis* Z69614; *Elphidium aculeatum* Z69618; *Bolivina* sp. Z69613; *Bigerina* sp. Z69611; *Astrorhiza triangularis* Z69609; *Astrammina rara* Z69608; *Massilina secans* Z69606; *Allogromia* sp. X86093; *Ammonia beccarii* X86094) were aligned

manually within the Genetic Data Environment (GDE) package (Smith *et al.*, 1994). Phylogenetic trees were reconstructed using 604 unambiguously aligned nucleotide sites (including 538bp from the eight conserved regions and 66bp from alignable areas within the variable regions and foraminiferal specific insertions). Phylogenetic analyses were performed using programs taken from version 3.52c of the Phylogeny Inference Package (PHYLIP) (Felsenstein 1993). Distance based phylogenetic analyses were performed using both the neighbour-joining (Saitou and Nei 1987) (program NEIGHBOR) and Fitch-Margoliash (Fitch and Margoliash 1967) (program FITCH) methods, with nucleotide sequence distances estimated for all pairwise sequence comparisons using the generalised two-parameter (maximum likelihood) model (program DNADIST). Maximum likelihood (Felsenstein 1981) phylogenetic analyses were performed using DNAML, and maximum parsimony (Fitch, 1971) analyses were performed using the program DNAPARS. Bootstrap resampling (Felsenstein 1985) (programs SEQBOOT and CONSENSE) was employed to assign support to the neighbour-joining (2000 replicates) and Fitch-Margoliash (100 replicates) trees. Alternative phylogenetic hypotheses, based upon the neighbour-joining and maximum likelihood trees, were evaluated statistically by a likelihood ratio test (Kishino and Hasegawa 1989) following the assignment of likelihoods (program DNAML) to artificially generated topologies (program RETREE).

The planktic foraminiferal SSU rDNA sequences presented in this study are deposited in GenBank, accession numbers X-Y.

Sequence variability in the 3' terminal fragment of the foraminiferal SSU rRNA gene

Alignment of the 1000bp 3' terminal fragment of the planktic foraminiferal SSU rRNA gene (Darling *et al.*, 1996b) with representatives of the diverse range of eukaryote taxa within the SSU rDNA database (Ribosomal data base project (RDP); Maidak *et al.*, 1994), has shown that the amplified region consists of eight conserved nucleotide regions which are interspersed by seven variable regions. Comparison against all eukaryote SSU rDNA sequences (438) presently within the RDP database show that the variable regions fall into two categories. The amplified fragment contains four variable length expansion segments (V7a and b, V8 and V9) which are present in most eukaryotes and also three insertions unique to the foraminifera (F1-F3), (Figure 2 and 3). The foraminiferal specific insertion F1 may represent the V6 variable region observed only in prokaryote sequences (Neefs *et al*, 1990). Extensive sequence variations were observed between the planktic foraminiferal species within both the foraminiferal specific insertions and expansion segments. This was manifest in extensive length variations (Figure 2) and substitutional changes (Figure 3) which rendered alignment between the more distantly related planktic foraminiferal species impossible within the variable regions.

The most distinctive feature of the alignment is a two base deletion in all planktic spinose species corresponding to position 2093/4 in *Allogromia sp.* In all the non-spinose and benthic species, two adenine bases are present in these positions. Three additional substitutional changes, at positions 2436 (A-T), 2546 (T-C) and 2985 (G/A-T), were also characteristic of the benthic/non-spinose to spinose split. A cytosine at position 2576 (T/A-C) was specific to all the planktic foraminiferal species, excluding *N. dutertrei.*

Phylogenetic relationships between benthic and planktic foraminiferal genera

A phylogenetic tree of 13 planktic foraminiferal sequences is presented in Figure 4. The planktic foraminiferal species are placed within a background of benthic species representing the five sub-orders, *Allogromiina, Textulariina, Astrorhizida, Milioliina* and *Rotaliina* (Pawlowski *et al.*, 1996; Pawlowski *et al.*, GenBank accession nos. Z69603, Z69604, Z69605, Z69606, Z69608, Z69609, Z69610, Z69611, Z69612, Z69613, Z69614, Z69615 and Z69618). The planktic species consist of 8 representatives of the spinose genera *Globigerinella, Globigerinoides, Orbulina* and *Globigerina,* and 2 representatives of the non-

spinose genera *Globorotalia* and *Neogloboquadrina*. The direction of evolution within the tree can be inferred to be from benthic to planktic from knowledge of the fossil record. The *Textularia* and *Allogromia* suborders are thought to represent the earliest benthic foraminifers of the Cambrian (Culver, 1993), with the planktic foraminifers evolving much later in the Mid-Jurassic.

The planktic and benthic foraminiferal species formed separate clusters within the reconstructed phylogeny (Figure 4), with the exception of the non-spinose species *N. dutertrei*, which falls within the benthic cluster. These phylogenetic relationships were consistently resolved with all tree construction methods employed (neighbour joining, Fitch-Margoliash, maximum parsimony and maximum likelihood). The separation of the main benthic group from the planktic species is supported in 94% of neighbour-joining and 88% of Fitch-Margoliash bootstrap replicates. The *N. dutertrei* / benthic cluster, contains benthic sequences of two sub-orders, *Textulariina* (Delage and Hérouard, 1896) and *Rotaliina* (Delage and Hérouard, 1896). *Neogloboquadrina dutertrei* shows an evolutionary distance of 3.3% to *Trochammina hadai*, 4.0% to *Textularia sp.* and 4.0% to *Bigerina sp.* which are of the sub-order *Textulariina* (Delage and Hérouard, 1896) and a distance of 5.2% to *Glabratella opercularis*, 5.6% to *Bolivina sp.* and 7.1% to *Haynesina germanica* of the sub-order *Rotaliina* (Delage and Hérouard, 1896). These distances contrast markedly with those observed between the planktic genera and approximate to the level of diversity observed between planktic sister species.

Phylogenetic relationships within the planktic genera

The planktic genera show a high level of sequence diversity, which is reflected in the long branch lengths observed between these species. The *Globorotalia* and *Globigerina* species are particularly divergent and cluster away from the main spinose planktic foraminiferal group. *Globorotalia menardii* and *Globigerina sp.* are separated by an evolutionary distance of 35%, which is similar to the 33% distance between *Gl. menardii* and the main spinose group and the 35% distance between *Globigerina sp.* and the main spinose group. The *Globorotalia* and *Globigerina* species cluster together in the neighbour-joining, maximum parsimony and maximum likelihood trees, with the branch supported in 86% of bootstrap replicates in the neighbour-joining tree. This branch is not however resolved in the Fitch-Margoliash tree, with *Gl. menardii* falling prior to *G. bulloides* on the lineage to the main planktic group. The remaining spinose species, *Globigerinella*, *Globigerinoides* and

Orbulina, cluster separately from *Globorotalia* and *Globigerina*, with the clade supported in 99% of both neighbour-joining and Fitch-Margoliash bootstrap replicates. The main spinose cluster is resolved with all tree construction methods.

Since the association between the main spinose group and the spinose *Globigerina sp.* remains unclear, we have attempted to clarify the relationship by examining a number of alternative phylogenetic hypotheses (Table 1). The placement of the main spinose group on the *Globigerina* lineage does not result in a significantly worse topology (P>0.40) nor does the placement of *Gl. menardii* on the main ancestral branch (P>0.40) (equivalent to the inferred placement for *Gl. menardii* in the Fitch-Margoliash tree). This indicates that the *Globigerina* lineage may form the ancestral lineage to the spinose species as indicated by the fossil record (Kennett and Srinivasan, 1983; Pearson, 1993).

The genera within the main spinose group are characterised by long branch lengths separated by relatively short internodes. These were supported in 87% and 82% of neighbour-joining bootstrap replicates and in 79% and 74% of Fitch-Margoliash bootstrap replicates. The order of branching of the main spinose group was not however consistent with all phylogeny reconstruction methods employed, with the placement of the *Gs. conglobatus*/*Gs. ruber* and *Ge. siphonifera* groups inconsistent in parsimony and maximum likelihood analyses. Rigorous likelihood testing of alternative branching orders could not resolve whether one topology was supported in preference to any other (unpublished analyses). Strong support is, however, provided for the fine structures within the main spinose group. *Globigerinoides conglobatus* shows a strong association with *Gs. ruber* (99% of neighbour-joining and 100% of Fitch-Margoliash bootstrap replicates) with the two species separated by an evolutionary distance of 14%. *Globigerinoides sacculifer* clusters with *O. universa* with the association supported in 82% of neighbour-joining and 74% of Fitch-Margoliash bootstrap replicates. *Globigerinoides sacculifer* shows an evolutionary distance of 22% to *O. universa*, a distance of 27% to *Gs. ruber* and a distance of 24% to *Gs. conglobatus*. *Orbulina universa* shows a distance of 23% to *Gs. ruber* and 21% to *Gs. conglobatus*.

Species level sequence variations

Within *Globigerinoides*, the pink and white forms of *Gs. ruber* show an evolutionary distance of 5.6%. This level of diversity is similar to that observed in *Globigerinella*, with a distance of 6.2% observed between the *Ge. siphonifera* Type I and Type II morphotypes. This degree of evolutionary change may be characteristic of a single planktic speciation

event and contrasts with complete sequence identity at the intra-species level among 5 individuals of *Ge. siphonifera* Type I (unpublished data).

Within the conserved regions (C1-C8, Figure 3), thirty substitutional changes were observed between *Ge. siphonifera* Type I and Type II. This compares to 26 substitutional changes between the pink and white forms of *Gs. ruber*. Changes were more commonly observed at the 3' end of the amplified fragment, being primarily concentrated in C6, C7 and C8. A greater degree of variability was observed between *Ge. siphonifera* Type I and Type II in the variable regions and foraminiferal specific insertions than observed between the pink and white forms of *Gs. ruber* (Figure 3). Most of the variable regions and foraminiferal specific insertions could not be aligned between *Ge. siphonifera* Type I and Type II, whereas variation within these regions between the pink and white forms of *Gs. ruber* was not sufficient to prevent alignment. The F3 region was alignable between *Ge. siphonifera* Type I and Type II, although not between *Gs. ruber* pink and white. The expansion segments V7a, V8 and V9 have a high level of base substitution and insertion/deletion events between Types I and II. The short region V7b has only one substitution and one deletion. The F3 alignment showed only four substitutional changes accompanied by two deletions in Type II, while F1 and F2 could not be aligned.

Intra-species sequence variations between Caribbean and Australian foraminiferal populations

Comparison of specimens of the Caribbean and Australian populations of *Ge. siphonifera* Type II, *Gs. sacculifer* and *O. universa* revealed complete sequence identity within the 8 conserved regions, with the exception of a single polymorphic site within the C2 region of *Ge. siphonifera* Type II (Figure 3). In addition, Caribbean and Australian populations of *Gs. sacculifer* and *O. universa* were also identical within both the expansion segments and foraminiferal specific insertions. Interestingly, the populations of *Ge. siphonifera* Type II showed a greater degree of variability within both the expansion segments and foraminiferal specific insertions. Nevertheless, this level of variability is substantially less than that observed at the species level between *Ge. siphonifera* Type I and Type II and between the pink and white forms of *Gs ruber*.

DISCUSSION

We have obtained partial SSU rDNA sequences of both spinose and non-spinose planktic foraminiferal genera. The phylogenetic placement of these sequences within a background of benthic species (Figure 4), revealed that the planktic species had a polyphyletic origin, with *N. dutertrei* clustering with benthic forms and away from the other planktic species. In addition, the phylogenetic tree indicates that there may be 4 different groups within the main spinose clade with large, approximately equidistant, evolutionary distances being observed between them. The molecular data also permits the genetic differentiation of the pink and white morphotypes of *Gs. ruber* and the two types of *Ge. siphonifera*. The Caribbean and Australian populations of *Gs. sacculifer* and *O. universa* show complete sequence identity, although Caribbean and Australian populations of *Ge. siphonifera* Type II show some differences in the variable regions.

Origins of the planktic foraminifera

The first planktic foraminifers appeared in low numbers in the Mid-Jurassic (Caron and Homewood, 1983; Tappan and Loeblich, 1988), and are thought to have possibly evolved from a group of trochospiral, benthic calcareous, hyaline foraminifera, the *Oberhauserellidae* of the suborder *Robertinina* (Loeblich and Tappan, 1974). The SSU rRNA gene of species of this suborder have yet to be sequenced. It is proposed that the ancestral stock of "globigerine" morphotypes gave rise to the diverse planktic assemblages that include the "globorotalid-like" morphotypes (Tappan and Loeblich, 1988). A similar succession of morphotypes consistently reappear following major faunal disruptions (Cifelli, 1969; Norris, 1991). The assumption has been made that since the Mid-Jurassic, planktic foraminifers have evolved from the lineages already inhabiting the plankton. This is brought into question by the sequence data presented here, which indicate that *N. dutertrei* is not descended from a "globigerine" lineage. *Neogloboquadrina dutertrei* is most closely related to benthic species of the suborders *Textulariina* and *Rotaliina* with an evolutionary distance of between 3.3% and 5.6%. This is considerably less than the 6.2% evolutionary distance observed between the planktic sibling species *Ge. siphonifera* Types I and II and the 5.6% distance observed between the pink and white forms of *Gs. ruber*. This implies either a more recent benthic habit or a much slower rate of evolution for the planktic species *N. dutertrei*, similar to that of the benthic sub-orders. In either case, *N. dutertrei* entered the plankton independently from

the other planktic species studied. Evolution of a planktic habit may, therefore, not always represent diversification from a previously existing planktic species, but instead, a separate derivation from the benthos. Such re-seeding from the benthos could account for the appearance of some planktic species into the fossil record where ancestral forms cannot be identified.

The precise benthic origin of *N. dutertrei* remains unclear. It seems highly unlikely that *N. dutertrei* originated from an agglutinated foraminiferal group like the *Textulariina*. Some calcareous foraminifers of the *Rotaliina* sub-order do however also cluster fairly closely with *N. dutertrei* and many other benthic groups have yet to be sequenced which may show a closer relationship. Examination of the Neogene *Neogloboquadrina* lineage (Parker, 1967; Blow, 1969 and Kennett and Srinivasan, 1983) shows that *N. dutertrei* has ancestors with remarkably similar morphology. The earliest member of the *Neogloboquadrina* is most likely *Neogloboquadrina acostaensis*. However, Kennett and Srinivasan (1983) suggested that *Neogloboquadrina continuosa* was possibly the first species of this lineage, but this link has never been proven. The most likely time for a reseeding from the benthos would therefore appear to be the origin of *N. acostaensis*. Alternatively, the *Neogloboquadrina* lineage may have derived from a larger clade including various cancellate non-spinose groups which arose in the Paleocene (Pearson, 1993). A thorough investigation of the extant non-spinose species will be necessary to determine the extent of the *N. dutertrei* clade. In addition, such an investigation would determine whether any other extant non-spinose lineages have arisen from benthic foraminifers rather than from planktic species.

The fossil record indicates that globigeriniid lineages have given rise to the globorotaliids on more than one occasion (Cifelli, 1969 and Norris, 1991). The molecular data of the present study does not link *Gl. menardii* with any of the benthic foraminiferal suborders from which sequence data is available, and *Gl. menardii* has been shown to be highly divergent from the lineage of *Globigerina sp.* There are however, a number of positions, including three substitutional changes and an insertion/deletion event, which are specific to the spinose planktic species and tie all the benthic sub-orders and the non-spinose neogloboquadriniid and globorotaliid planktic lineages together. All benthic species sequenced so far, and the two sequenced planktic taxa, *Gl. menardii* and *N. dutertrei*, possess two extra bases within the conserved region C1. It is highly unlikely that the globorotaliid lineage would have re-acquired such an insertion in an identical position. This indicates that the globorotaliid ancestors of *Gl. menardii* may have diverged prior to the divergence of the

globigerinid spinose lineage of *Globigerina sp.* and could equally well be located on the main branch of the tree. The separate origin of the globorotaliids is also supported by the statistical analysis shown in Table 1 (test 2). Although there is support for the common ancestry of *Gl. menardii* and *Globigerina sp.* with all phylogenetic methods excluding Fitch-Margoliash, the point of their divergence from the common ancestor is ambiguous due to their high evolutionary distance of 35%. Molecular phylogenetic analysis may be able to resolve the true order of divergence only when more extant species have been sequenced to provide further structure within the central section of the tree.

Comparison of spinose species phylogenetic relationships against the fossil record

The ancestry of the modern spinose species is unclear. Kennett and Srinivasan, (1983) suggest that the modern spinose species may have evolved from a *Globigerina* ancestor in the Late Oligocene. They subdivide the lineage into the sub-genera *Globigerina* (*Globigerina*) and *Globigerina* (*Zeaglobigerina*), from which the different lineages of spinose species are thought to have evolved. However, Pearson (1993) suggested that this split may have occurred much earlier in the Paleocene or Eocene, when there was a great diversification of spinose species within the genus *Subbotina*. The two lineages can be separated by differences in the test wall structure; *Globigerina* (*Globigerina*; Gg.) has a hispid surface and *Globigerina* (*Zeaglobigerina*, Zg.) a cancellate surface. The modern representatives of the two lineages are thought to be *Globigerinella* (Gg.) and *Globigerinoides* (Zg.). The molecular data, however, does not support this sub-division within the spinose group. The tree indicates that there are four relatively equidistant groups which share a common ancestor. Each group could be considered to be of genus level status, as the evolutionary distance of 27% between *Gs. sacculifer* and *Gs. ruber* is greater than the genus level differences observed between *Gs. sacculifer*/*Ge. siphonifera* (22%) and *Gs. sacculifer*/*O. universa* (22%). *Globigerinoides conglobatus* clusters with *Gs. ruber*, with an evolutionary distance of 14%. There is strong statistical support for this association and may indicate that they diverged from a common ancestor at a later time than the other species within the proposed Zg. lineage.

Differences in branch lengths between the planktic and benthic species

One of the striking features of the foraminiferal molecular phylogeny, is the apparent difference in evolution rates within and between planktic and benthic foraminiferal species. Within the planktic species there are large variations in branch lengths which suggest

extensive variation in evolutionary rates. Since the divergences of the Neogene planktic foraminifers are relatively accurately dated, it should be possible to estimate an evolutionary rate for these lineages. We have made an attempt to estimate evolutionary rates for two lineages, *Globigerinella* and *Orbulina*, for which we have good fossil evidence for the date of their morphological divergence. The divergence of *Globigerinella* at 31 Ma. from *G. praebulloides* gives an estimated evolutionary rate of 1%/4.6Ma. In contrast, the divergence of *O. universa* from *Gs. triloba* at 17 Ma. gives an evolutionary rate of 1%/1.8Ma. It is possible that apparent rate differences may reflect a lack of correlation between the point at which genetic, as opposed to morphological, divergence occurred. The first appearance of a lineage within the fossil record may not always reflect the point at which molecular divergence occurred, as distinct lineages may remain cryptic for a period of time prior to morphotypic change. This is likely to underestimate the divergence rate. The apparent evolutionary rate differences may also be a consequence of the number of speciation events occurring within the lineage from the point of divergence to the present time.

Branch lengths for all benthic foraminifers in the rRNA tree are shorter than those among the planktic lineages, despite the fact that benthic foraminifers have a longer fossil record. This suggests that benthic evolution rates are slower than those observed in planktic species. The anomalous planktic foraminifera *N. dutertrei*, shows branch lengths that are more characteristic of the benthic species. It is possible that the evolutionary processes within the benthos are entirely different from those of the plankton.

Molecular data provides unique markers for morphotypic investigation

Morphotypic variation at the species level and within species level is of particular interest for paleoceanography and paleoclimatology. Many changes in foraminiferal morphology have been attributed to the effects of environmental change on their phenotype (Kennett, 1976; Robbins, 1988; Kroon *et al.*, 1988; Williams, *et al.*, 1988). Such "ecophenotypes" have been used as indicators of past climatic change. It is most likely, however, that phenotypic variability is due to a combination of environmental, metabolic and genetic factors (Brummer and Kroon, 1988). Molecular data now provides us with species-specific molecular markers that can be used to resolve the question of whether such morphological differences are paralleled by genetic divergence.

The two biologically distinct forms of *Ge. siphonifera* (Faber *et al.*, 1988) have very subtle differences in test morphology that do not justify species level distinction using

traditional paleontological species concepts. Biological and geochemical evidence nevertheless suggests overwhelmingly that *Ge. siphonifera* Type I and Type II are sibling species (Knowlton, 1993), despite the morphological similarity of their empty shells (Huber *et al.*, submitted). The molecular data supports this, with a high level of sequence diversity (5.9%) observed within the conserved regions between *Ge. siphonifera* Type I and II. In contrast, no differences were observed in either the conserved or variable regions between individuals of *Ge. siphonifera* Type I in the Caribbean. Complete sequence identity, with the exception of a single polymorphic base, was also observed within the conserved regions between *Ge. siphonifera* Type II from the Caribbean and Australia, despite their geographic separation by the Isthmus of Panama. The combined evidence clearly suggests that *Ge. siphonifera* Type I and II should be reclassified as distinct species. The speciation mechanisms responsible for their divergence also remain unclear (discussed in Huber *et al.*, submitted). This speciation event has been termed cryptic speciation, and possibly heralds the first of many such events previously hidden from taxonomic scrutiny, but which can now be identified by molecular phylogenetic analysis. If such events are common, they will prove challenging in the phylogenetic interpretation of the fossil record based on morphology.

One possible speciation event that is not "cryptic" is that observed between the pink and white forms of *Gs. ruber*. The two forms live in close proximity in the water column in the Caribbean, though the pink form became extinct 125,000 years ago in the Pacific (Thompson *et al.*, 1979). The level of nucleotide variation between these two morphotypes is 5.6%, very similar to the level of diversity observed between *Ge. siphonifera* Type I and Type II. The two morphotypes of *Gs. ruber* were collected in the Caribbean (pink) and the GBR (white) and are therefore geographically isolated by the Isthmus of Panama. Although the pink form of *Gs. ruber* has not been observed in sediments older than the Pleistocene (1.66 Ma), its presence on both sides of the Panama Isthmus indicates that it probably evolved before the closure, and that its absence in earlier sediments is most likely due to lack of preservation of the pigment. In order to test whether such isolation could account for the 5.6% difference observed between them, it would have been desirable to compare the pink and white forms from the Caribbean. Unfortunately, *Gs. ruber* is difficult to keep in culture and only the pink form was successfully cultured to gametogenesis. *Globigerinella siphonifera* Type II, *Gs. sacculifer* and *O. universa* collected in the Caribbean and off the GBR, and therefore subjected to the same isolation scenario, can however be compared. Analysis showed that *Ge. siphonifera* Type II had only a single polymorphic change in 604

base positions and *Gs. sacculifer* and *O. universa* showed complete sequence identity (Figure 3). This would suggest that Caribbean/Australian isolation does not account for the differences observed between *Gs. ruber* morphotypes.

The fact that *Ge. siphonifera* Type II, *Gs. sacculifer* and *O. universa* show a remarkable degree of conservation in the conserved regions between the Caribbean and the western Pacific Ocean, indicates that the sequences can be used as unique species specific molecular markers for these species in future studies. This will be of particular importance in discriminating between similar "cryptic" speciation events as seen within the *siphonifera*. Although the variability between the expansion segments and foraminiferal specific insertions makes their use in phylogenetic analyses impractical at the species level, these regions may prove informative for analyses at the population level. Differences observed in the variable regions between *Ge. siphonifera* Type II from the Caribbean and Australia (Figure 3), indicate their potential for such investigations. Nevertheless, complete sequence identity was observed within the variable regions between the Caribbean and Australian populations of *Gs. sacculifer* and *O. universa*, rendering them uninformative for population level investigation in these species. The LSU rRNA gene or the ribosomal internal transcribed spacers (regions lying between the SSU and LSU genes) may prove more informative in population level studies.

The plankton ecosystem and the passage from benthos to plankton

The fossil record indicates that the movement of foraminifers from the benthos to the plankton significantly post-dates the major diversification of benthic foraminifers (Tappan and Loeblich, 1988). It is highly probable that the pelagic environment, which existed prior to the appearance of the first planktic foraminifers in the Mid-Jurassic, was an ecosystem not suited to foraminiferal nutrient requirements or life cycle (Signor and Vermeij, 1994). However, by the late Triassic and Jurassic, coccolithophoriids and diatoms were present in the plankton and the mid Cretaceous saw a large increase in the diversity of coccolithophoriids, diatoms and dinoflagellates (Tappan and Loeblich, 1973). High competition and predation pressures in the benthos probably offered a major evolutionary advantage in diversifying into a planktic environment. The movement from the benthos to the plankton is a major step and possibly necessitates considerable adaptation. Foraminifers are most likely to have been restricted from adopting a planktic life style by the negative buoyancy of their adult forms. The fact that benthic foraminifers are negatively buoyant is particularly indicated by the strategy of the

gamont forms of some species of *Discorbidae* and *Cymbaloporidae* to adopt a planktic mode of existence for part of their life cycle. The evolution of gas filled float chambers has enabled these largely benthic species to release their gametes in the surface layers, an adaptation which is believed to have evolved independently on at least four occasions (Banner *et al.*, 1985). Planktic foraminifers may however, have relied upon the acquisition of an organelle to control their buoyancy. It has been postulated that fibrillar bodies (Lee *et al.*, 1965), present in large numbers in the cytoplasm of spinose and non-spinose planktic foraminifers, including *N. dutertrei* (Hemleben, 1989), may have a buoyancy function (Anderson and Bé, 1976). Fibrillar bodies are unknown in benthic foraminifers, which adds further support to their function as a flotation device (Anderson and Lee, 1991), although Spero (1988) suggests an alternative function, as a source of stored calcium ions for calcification. Evidence from the present study indicates that the planktic foraminifera have achieved buoyancy at different times throughout their history. If the acquisition of fibrillar bodies is required by planktic foraminifers before buoyancy can be achieved, their appearance within the planktic foraminifera may not represent a single evolutionary event. Alternatively, all lineages of benthic foraminifers which gave rise to the planktic foraminifers may possess fibrillar body precursors in some as yet unrecognised form.

CONCLUSIONS

Benthic and planktic foraminifers form a monophyletic group, early in eukaryote evolution. Planktic foraminifers derive from at least two benthic ancestral lines. The majority of planktic foraminifers cluster distinctly from the benthic foraminifers sequenced to date. *Neogloboquadrina*, however, shows a close relationship with early evolving benthic species. Within our reconstructed phylogeny, the non-spinose species *Gl. menardii* clusters with the spinose species *Globigerina*. This clustering is, however, equivocal and statistical tests show that *Gl. menardii* could be placed on the main branch. The spinose species sequenced within this study may have a common ancestor, with *Globigerina* falling ancestral to the main group of spinose species. There are possibly four distinct groups within the main spinose clade with large evolutionary distances being observed between them. *Globigerinoides conglobatus* clusters with *Gs. ruber* with strong statistical support for this association. Large evolutionary distances, which are indicative of species level differences, were observed between the pink and white forms of *Gs. ruber* and *Ge. siphonifera* Types I and II. The Caribbean and Australian populations of *Gs. sacculifer* and *O. universa* have complete sequence identity, although *Ge. siphonifera* Type II populations have some differences in the variable region, the reasons for which remain unclear.

Additional foraminiferal sequences are required before the precise phylogenetic relationships between the foraminifera can be resolved. In particular, benthic species which are thought likely ancestors of planktic groups and planktic species which have an unclear ancestry from the fossil record, should be targeted. The sequencing of key planktic foraminiferal species should provide further structure within the tree, allowing a more extensive comparison against planktic fossil phylogenies.

ACKNOWLEDGEMENTS

We would like to thank Hugh Sweatman (Australian Institute of Marine Science) for his advice and introductions and also David Blair and his staff, Department of Zoology, James Cook University, for the use of their molecular laboratory facilities. In particular, we would like to thank Anne Hoggett and Lyle Vail, Co-Directors of Lizard Island Research Station, Cairns, for their invaluable help and support. Permission for planktic foraminiferal collection in Australia was granted by the Great Barrier Reef Marine Park Authority, Townsville. We thank Heather Austin, Jelle Bijma and Brian Huber for their assistance during the collection and culture of foraminifera at the Caribbean Marine Biological Institute (CARMABI), Curaçao and also Rebecca Darling for her assistance on Lizard Island. We are very grateful to Paul Pearson and two anonymous reviewers for their comprehensive and constructive reviews. Research was supported by NERC (GR3/09736) with additional financial support provided by the Carnegie Trust.

REFERENCES

Anderson, O. R., and A. W. H. Bé. 1976. The ultrastructure of a planktonic foraminifer *Globigerinoides sacculifer* (Brady), and its symbiotic dinoflagellates. J. Foram. Res. **6**:1-21.

Anderson, O. R., and J. J. Lee. 1991. Cytology and fine structure. *In* J. J. Lee, and R. O. Anderson. eds. Biology of Foraminifera. Academic Press, New York and London. pp. 7-40.

Banner, F. T., C. P. G. Pereira, and D. Desai. 1985. "Tretomphaloid" float chambers in the Discorbidae and Cymbaloporidae. J. Foram. Res. **15(3)**:159-174.

Blow, W. H. 1969. Late Middle Eocene to Recent Planktonic Foraminiferal Biostratigraphy. *In* Bronnimann, P., and H. H. Renz. eds. Proceedings of the First International Conference on Planktonic Microfossils, Volume 1, Leiden, E. J. Brill. pp 199-422.

Brummer G. A., and D. Kroon. 1988. Genetically controlled planktonic foraminiferal coiling ratios as tracers of past ocean dynamics. *In* Brummer, G. A., and D. Kroon. eds. Planktonic foraminifers as tracers of ocean climate history, Free University Press, Amsterdam. pp. 131-147.

Caron, M., and P. Homewood. 1983. Evolution of early planktic foraminifers. Marine Micropaleontology **7**:453-462.

Cifelli, R. J. 1969. Radiation of Cenozoic planktonic foraminifera. Syst. Zool. **18**:154-168.

Culver, S. J. 1993. Foraminifera. *In* Lipps, J. H. eds. Fossil Prokaryotes and Protists. Blackwell, Boston. pp. 203-247.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996a. The isolation and amplification of the 18S ribosomal RNA gene from planktonic foraminifers using gametogenic specimens. *In* Whatley, R. C., and Moguilevsky, A. eds. Microfossils and Oceanic Environments. University of Wales, Aberystwyth Press. Chapter 3.1, pp. 249-259.

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown. 1996b. Molecular evolution of planktic foraminifera. J. Foram. Res. **26**:324-330.

Faber, W. W. Jr., O. R. Anderson, J. L. Lindsay, and D. A. Caron. 1988. Algal foraminiferal symbiosis in the planktonic foraminifera *Globigerinella aequilateralis*. 1. Occurrence and stability of two mutually exclusive chrysophyte endosymbionts and their ultrastructure. J. Foram. Res. **18**:334 -343.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368-376.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783-791.

Felsenstein, J. 1993. PHYLIP, Manual version 3.52c. Berkeley University Herbarium, University of California, Berkeley.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees: A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science **155**:279-284.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. **20**:406-416.

Hemleben, C., M. Spindler, and O. R. Anderson. 1989. Modern Planktonic Foraminifera. Springer-Verlag, New York, pp 335.

Huber, B. T., J. Bijma, and K. F. Darling. Cryptic speciation in the living planktic foraminifer *Globigerinella siphonifera* (d'Orbigny). Submitted to Paleobiology.

Kennett, J. P. 1976. Phenotypic Variation in some Recent and Late Cenozoic Planktonic Foraminifera. *In* Hedley, R. H., and C. G. Adams. eds. Foraminifera, Volume 2. Academic Press, London. pp. 1-60.

Kennett, J. P., and M. S. Srinivasan. 1983. Neogene planktonic foraminifera. Hutchinson Ross, Stroudsburg.

Kishino H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order of the Hominoidea. J. Mol. Evol. 4:406-425.

Knoll, A. H. 1992. The early evolution of eukaryotes: A geological perspective. Science 256:622-627.

Knowlton, N. 1993. Sibling species in the sea. Annu. Rev. Ecol. Syst. 24:189-216.

Kroon, D., P. F. Wouters, L. Moodley, G. Ganssen, and S. R. Troelstra. 1988. Phenotypic variation of *Turborotalita quinqueloba* (Natland) tests in living populations and in the Pleistocene of an eastern Mediterranean piston core. *In* Brummer, G. J. A., and D. Kroon. eds. Planktonic foraminifers as tracers of ocean climate history. Free University Press, Amsterdam pp. 131-147.

Lee, J. J., H. D. Freudenthal, V. Kossoy, and A. W. H. Bé. 1965. Cytological observations on two planktonic foraminifera, *Globigerina bulloides* d'Orbigny, 1826 and *Globigerinoides ruber* (d'Orbigny, 1839) Cushman, 1927. J. Protozool. 12:531-542.

Leigh Brown, A. J., and P. Simmonds. 1994. Analysis of HIV sequence variation. *In* Karn, J. ed. HIV: a practical approach. Oxford University Press. pp. 161-188.

Loeblich, A. R. Jr., and H. Tappan. 1974. Recent advances in the classification of the Foraminiferida. *In* Hedley, R. H., and C. G. Adams. eds. Foraminifera, 1. Academic Press, London. pp. 1-53.

Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olson, K. Fogel, J. Blandy, and C. R. Woese. 1994. The ribosomal database project. Nucl. Acids Res. 22:3484-3487.

Malmgren, B. A., and J. P. Kennett. 1977. Biometric differentiation between recent

Globigerina bulloides and *Globigerina falconensis* in the southern Indian Ocean. J. Foram. Res. 7:130-148.

Neefs, J., Y. Van de Peer, L. Hendriks, and R. De Wachter. 1990. Compilation of small ribosomal subunit RNA sequences. Nucl. Acids Res. 18:2237-2242.

Norris, R. D. 1991. Biased extinction and evolutionary trends. Paleobiology, 17(4):388-399.

Olsson, R. K., C. Hemleben, W. A. Berggren, and C. Liu. 1992. Wall texture classification of planktonic foraminifera genera in the lower Danian. J. Foram. Res. 22(3):195-213.

Parker, F. L. 1967. Late Tertiary biostratigraphy (planktonic foraminifer) of tropical Indo-Pacific deep-sea cores. Bulletins of American Paleontology 52:115-208.

Pawlowski, J., I. Bolivar, J. Guiard-Maffia, and M. Gouy. 1994. Phylogenetic position of foraminifera inferred from LSU rRNA gene sequences. Mol. Biol. Evol. 11:929-938.

Pawlowski, J., I. Bolivar, J. Fahrni, T. Cavalier-Smith, and M. Gouy. 1996. Early origin of foraminifera suggested by SSU rRNA gene sequences. Mol. Biol. Evol. 13:445-450.

Pearson, P. N. 1993. A lineage phylogeny for the Paleogene planktonic foraminifera. Micropaleontology 39:193-222.

Robbins, L. L. 1988. Environmental significance of morphotypic variability in open-ocean versus ocean-margin assemblages of *Orbulina universa*. J. Foram. Res. 18(4):326-333.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Signor, P. W., and G. J. Vermeij. 1994. The plankton and the benthos: origins and early history of an evolving relationship. Paleobiology 20(3):297-319.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet. 1994. The genetic

data environment: an expandable GUI for multiple sequence analysis. Comput. Appl. Biosci. **10**:671-675.

Sogin, M. L. 1989. Evolution of eukaryotic microorganisms and their small subunit ribosomal RNA's. American Zoology **29**:487-499.

Spero, H. 1988. Ultrastructural examination of chamber morphogenesis and biomineralization in the planktonic foraminifer *Orbulina universa*. Marine Biology **99**:9-20.

Tappan, H., and A. R. Loeblich. 1973. Evolution of the oceanic plankton. Earth Science Reviews **9**:207-240.

Tappan, H., and A. R. Loeblich. 1988. Foraminiferal evolution, diversification, and extinction. J. Paleontol. **62(5)**:695-714.

Thompson, P. R., A. W. H. Duplessy, and N. J. Shackleton. 1979. Disappearance of pink-pigmented *Globigerinoides ruber* at 120,000 years BP in the Indian and Pacific Oceans. Nature **280**:554-558.

Wade, C. M., K. F. Darling, D. Kroon, and A. J. Leigh Brown. 1996. Early evolutionary origin of the planktic foraminifera inferred from SSU rDNA sequence comparisons. J. Mol. Evol. **43**:672-677.

Wainwright, P. O., G. Hinkle, M. L. Sogin, and S. K. Stickel. 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. Science **260**:340-342.

White, T. J., T. Bruns, S. Lee, and J. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* Innis, M. A., D. H. Gelfland, J. J. Sninsky, and T. J. White. eds. PCR Protocols: A Guide to Methods and Applications. Harcourt Brace Jovanovich, San Diego. pp.315-322.

Williams, D. F., R. Ehrlich, H. J. Spero, N. Healy-Williams, and A. C. Gary. 1988. Shape

and isotopic differences between conspecific foraminiferal morphotypes and resolution of paleoceanographic events. Paleogeography, Palaeoclimatology, Palaeoecology **64**:153-162.

Table 1.

Likelihood support for alternative phylogenetic hypotheses for the relationships between the planktic genera within the tree. Likelihood evaluation of alternative phylogenetic hypotheses was performed using a likelihood ratio test (Kishino and Hasegawa, 1989) with hypotheses tested based on either a neighbour-joining or maximum likelihood tree.

	Neighbour-joining		Maximum likelihood	
Competing Hypotheses	**log likelihood**	**Probability**	**Log likelihood**	**Probability**
Test 1. **Inferred phylogeny** **VS** **Main spinose group placed on Globigerina lineage**	-5181.884 -5186.696 (SD: 8.045)	 P>0.40	-5139.360 -5145.382 (SD: 8.580)	 P>0.40
Test 2. **Inferred Phylogeny** **VS** **Gl. menardii placed on main ancestral branch**	-5181.884 -5186.697 (SD: 8.052)	 P>0.40	-5139.360 -5145.399 (SD: 8.580)	 P>0.40

Figure 1.

Small subunit rDNA phylogeny for the planktic foraminifera modified from Wade *et al.* (1996). The phylogenetic tree was reconstructed by neighbour-joining analysis of 546 unambiguously aligned nucleotide sites. The scale bar corresponds to 10 changes per 100 nucleotide positions. The "crown" group taxa consist of *Symbiodinium pilosum* (symbiotic dinoflagellate), *Theileria annulata* (apicomplexan), *Tetrahymena thermophila* (ciliate), *Ochromonas danica* (chrysophyte), *Achlya bisexualis* (oomycete), *Sceletonema costatum* (diatom), *Emiliania huxleyi* (prymnesiophyte), *Chlamydomonas reinhardtii* (green alga), *Zea Mays* (maize), *Acanthamoeba castellanii* (acanthamoeba), *Saccharomyces cerevisae* (yeast), *Artemia salina* (brine shrimp) and *Homo sapiens* (human).

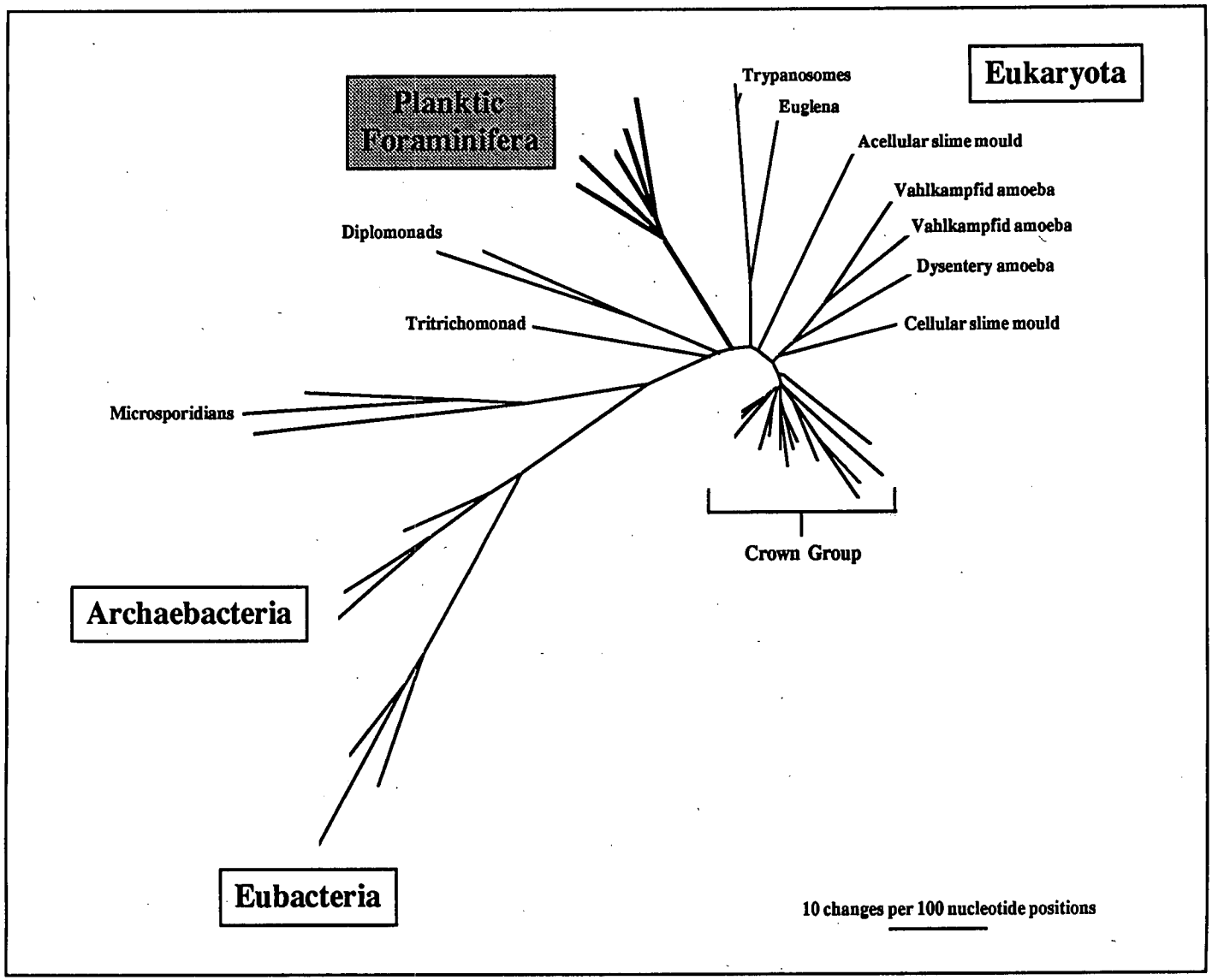

Eukaryota
Planktic Foraminifera
Trypanosomes
Euglena
Acellular slime mould
Vahlkampfid amoeba
Vahlkampfid amoeba
Dysentery amoeba
Cellular slime mould
Diplomonads
Tritrichomonad
Microsporidians
Crown Group
Archaebacteria
Eubacteria
10 changes per 100 nucleotide positions

Figure 2.

A schematic representation of the 3' terminal region of the SSU rRNA gene (approximately 1000 base pairs) showing length variations within the foraminiferal specific insertions and expansion segments. V7-V9 represent variable length expansion segments present in most eukaryotes and F1-F3 represent three insertions which are unique to the foraminifera. F1 may represent the V6 variable region observed only in prokaryote sequences (Neefs *et al.*, 1990). The length of each region is given in base pairs.

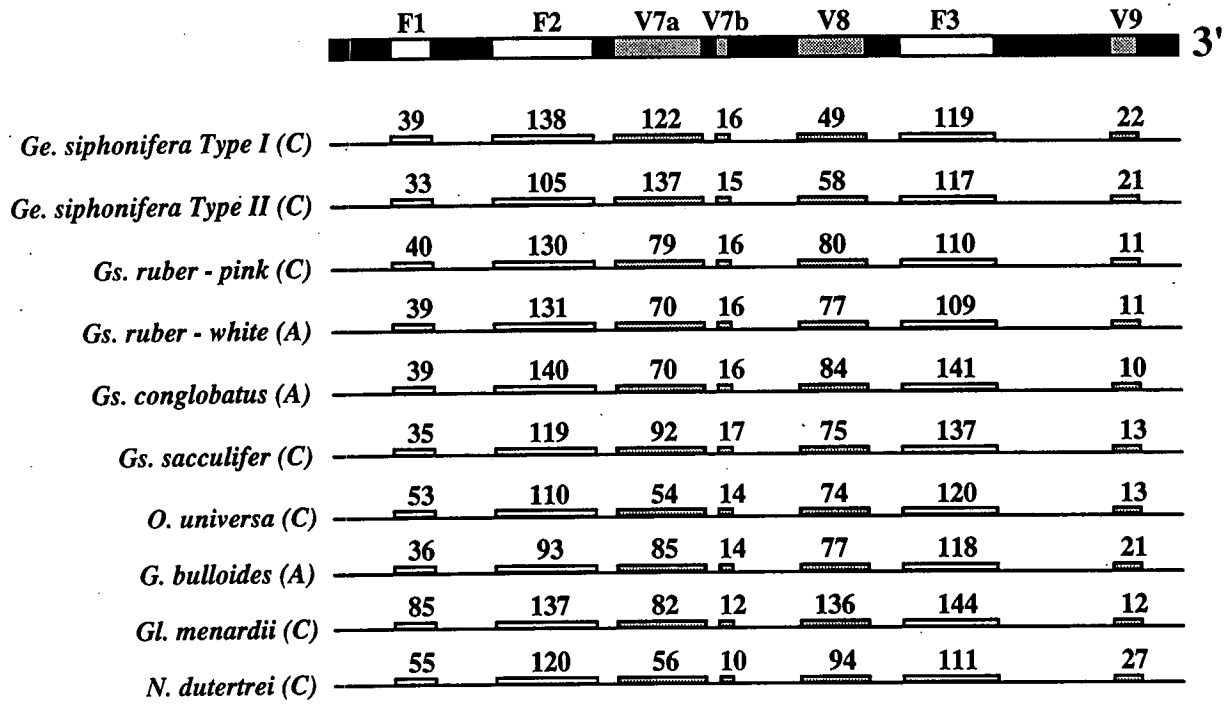

F1
F2
V7a
V7b
V8
F3
V9
3'
Ge. siphonifera Type I (C)
39 138 122 16 49 119 22
Ge. siphonifera Type II (C)
33 105 137 15 58 117 21
Gs. ruber - pink (C)
40 130 79 16 80 110 11
Gs. ruber - white (A)
39 131 70 16 77 109 11
Gs. conglobatus (A)
39 140 70 16 84 141 10
Gs. sacculifer (C)
35 119 92 17 75 137 13
O. universa (C)
53 110 54 14 74 120 13
G. bulloides (A)
36 93 85 14 77 118 21
Gl. menardii (C)
85 137 82 12 136 144 12
N. dutertrei (C)
55 120 56 10 94 111 27

Figure 3.

A schematic representation of the 3' terminal region of the SSU rRNA gene (approximately 1000 base pairs) showing substitutional changes within the foraminiferal conserved regions, specific insertions and expansion segments. C1-C8 represent the highly conserved regions which were aligned relative to comparable regions present in all eukaryotes. V7-V9 represent variable length expansion segments present in most eukaryotes and F1-F3 represent three insertions which are unique to the foraminifera. P represents a polymorphic base position. U denotes that a region is unalignable.

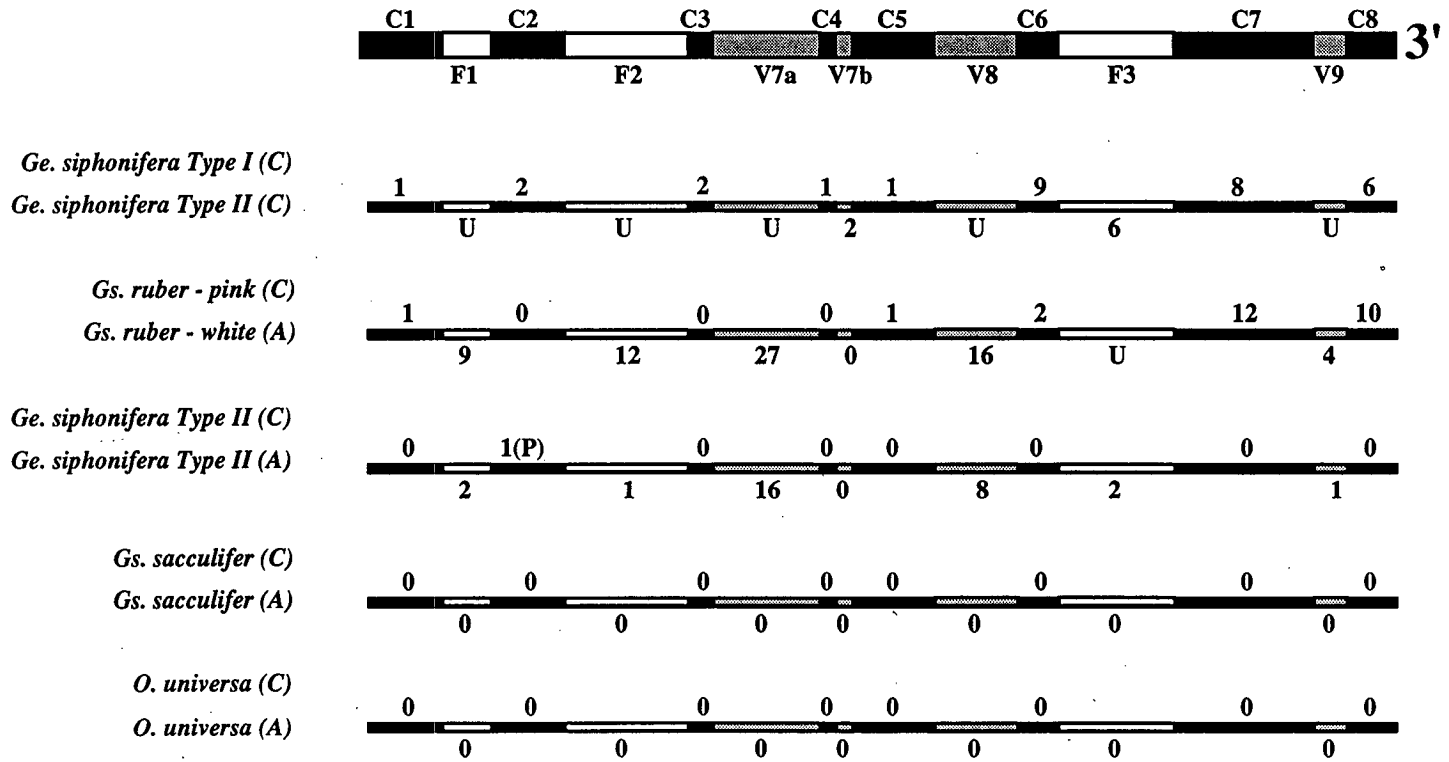
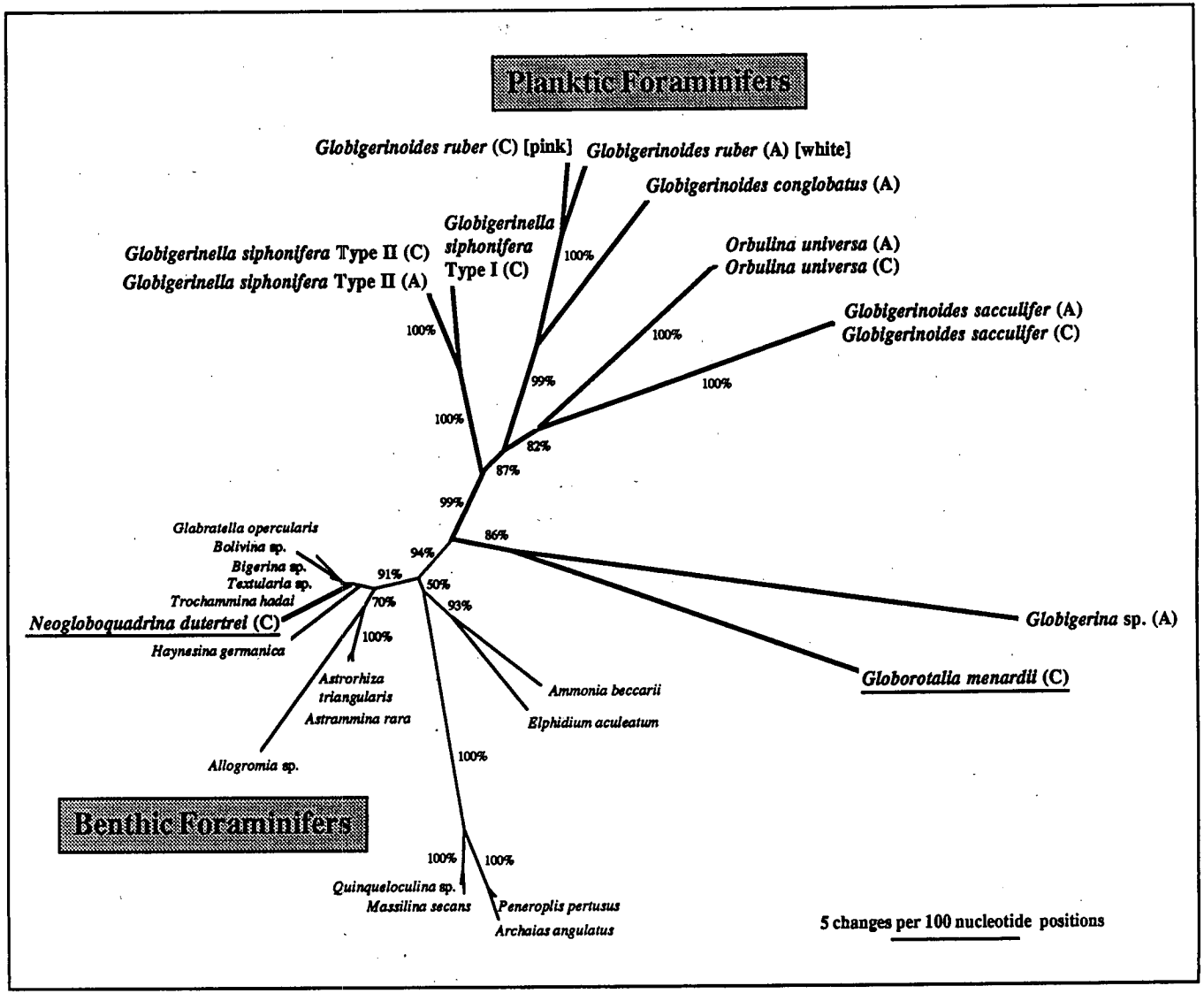

C1
C2
C3
C4
C5
C6
C7
C8
3'
F1
F2
V7a
V7b
V8
F3
V9
Ge. siphonifera Type I (C)
Ge. siphonifera Type II (C)
1 2 2 1 1 9 8 6
U U U 2 U 6 U
Gs. ruber - pink (C)
Gs. ruber - white (A)
1 0 0 0 1 2 12 10
9 12 27 0 16 U 4
Ge. siphonifera Type II (C)
Ge. siphonifera Type II (A)
0 1(P) 0 0 0 0 0 0
2 1 16 0 8 2 1
Gs. sacculifer (C)
Gs. sacculifer (A)
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0
O. universa (C)
O. universa (A)
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0

Figure 4.

Neighbour joining, unrooted phylogenetic tree showing the relationships between planktic spinose (*Globigerinella*, *Globigerinoides*, *Orbulina* and *Globigerina*) and non-spinose (*Globorotalia* and *Neogloboquadrina*) genera within a background of benthic species (representing 5 sub-orders, *Allogromiina*, *Textulariina*, *Astrorhizida*, *Milioliina* and *Rotaliina*). The phylogenetic tree was reconstructed based upon 604 unambiguously aligned sites of the terminal 3' region of the SSU rRNA gene (including 538bp from the eight conserved regions and 66bp from alignable areas of the variable regions/foraminiferal specific insertions). The planktic sequences obtained in this study are highlighted in bold, with the non-spinose planktic species underlined. Branch lengths represent the evolutionary distance between species calculated from substitutional mutations/site. The scale bar corresponds to 5 changes per 100 nucleotide positions. Bootstrap values, which reflect the support for a particular branch within the tree, are expressed as a percentage. Neighbour-joining bootstrap values are based on 2000 bootstrap replications. C denotes a Caribbean specimen and A, Australian.


Planktic Foraminifers
Globigerinoides ruber (C) [pink]
Globigerinoides ruber (A) [white]
Globigerinoides conglobatus (A)
Globigerinella siphonifera Type I (C)
Globigerinella siphonifera Type II (C)
Globigerinella siphonifera Type II (A)
Orbulina universa (A)
Orbulina universa (C)
Globigerinoides sacculifer (A)
Globigerinoides sacculifer (C)
100%
100%
100%
100%
99%
100%
82%
87%
99%
86%
94%
91%
50%
93%
70%
100%
Glabratella opercularis
Bolivina sp.
Bigerina sp.
Textularia sp.
Trochammina hadai
Neogloboquadrina dutertrei (C)
Haynesina germanica
Astrorhiza triangularis
Astrammina rara
Allogromia sp.
Globigerina sp. (A)
Globorotalia menardii (C)
Ammonia beccarii
Elphidium aculeatum
100%
Benthic Foraminifers
100%
100%
Quinqueloculina sp.
Massilina secans
Peneroplis pertusus
Archaias angulatus
5 changes per 100 nucleotide positions

Paper V

Reading the h story o the oceans in plankton DNA

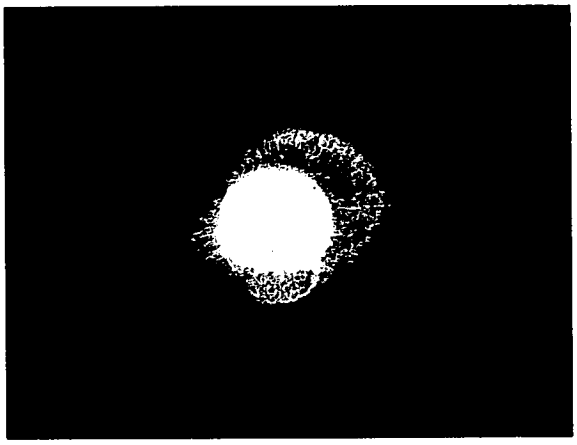
Kate Darling, Dick Kroon, Chris Wade and Andrew Leigh Brown

At the University of Edinburgh, oceanographers have teamed up with molecular biologists to apply DNA sequencing techniques to special types of plankton collected from oceans around the world. A comparison of the geographical and genetic relationships between different species of plankton is providing new information about the history of ocean circulation.

Marine plankton - tiny animals and plants floating in the world's oceans - include large numbers of single-celled animals called the Foraminifera. These are distinguished from other plankton by their hard shell of calcium carbonate that allows them to take on complex symmetrical shapes. Planktic (floating) foraminifers are thought to have evolved from benthic species (which live on the ocean floor) during the mid Jurassic period, 170 million years ago.

Fossil records

The skeletons of dead plankton sink to the ocean floor and accumulate in sediments, which have been laid down to immense depths over millions of years. These accumulations of dead marine organisms can now be observed on land as massive deposits of chalk and limestone, where there were once seas. The skeletons of fossil foraminifers are often well preserved in sediments, and can be studied in cores of ocean sediments from around the world. By noting the succession of foraminifer fossil shapes through the layers of sediment, and therefore through time, scientists have constructed a precise 'biostratigraphy' based on these species. This evolutionary progression of planktic foraminifers now allows us to identify the age of particular rock strata, providing the backbone of part of our geological timescale.



The spiny protrusions on this hard-shelled plankton acts as a net for snaring large mobile prey. The different shapes of the 'skeletons' of this large family of plankton can be used to study the history of the oceans.

Species adaptation and diversity reflect ancient conditions

Information bound up in the fossil record of planktic foraminifers can be used to reconstruct the circulation and climate of past oceans. Each species of foraminifer is adapted, by its shape and behaviour, to an environment with specific conditions of temperature, food and salt concentration. Knowing the distribution and ecology of foraminiferal species in different environments today, we can estimate when ancient conditions were the same, because similar-looking species would be found in the fossil record at that time.

When the environment changes, natural selection favours the organisms best adapted to live in the new conditions. For the planktic foraminifers, changing patterns of ocean circulation and global climate are thought to have been the driving forces behind rapid bursts of evolution.

A high species diversity is today observed in tropical oceans. Periods in the past with a high species diversity in the fossil record can therefore be associated with similar warm, stable climates. On the other hand, many individual foraminifers with low species diversity may suggest a region with sporadically high nutrient levels. This could happen in cold polar waters, where large numbers of a single foraminifer species feed on massive seasonal blooms of algae.

Genetic relationships

The above studies are based on looking at the shapes (morphology) of modern and ancient foraminifers, to assess whether species are related. However, this is not the whole story because shape cannot reveal the genetic relationships between species. Two distantly related species might look the same by chance, while two different-looking species may be closely related.

We are using a new technique to study these genetic relationships. This involves examining the DNA sequence of a single gene in each species. The gene codes for ribosomal RNA (rRNA), which is an essential component of the protein synthesis machinery in all

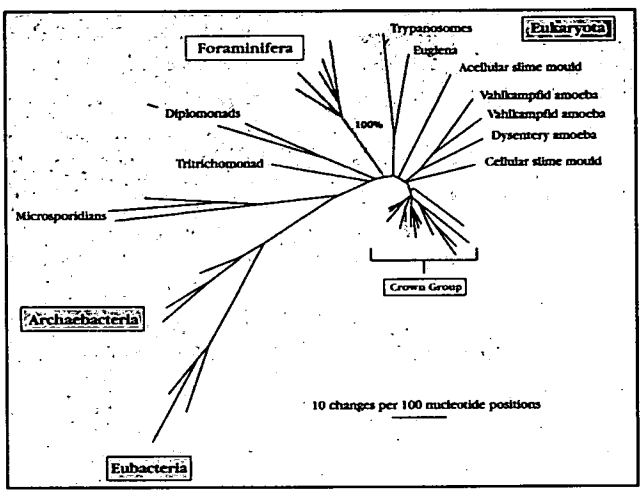
types of cell. With the passage of time many sequence changes are incorporated into the gene, as random mutations occur. The more sequence differences there are between this gene in two different organisms, the greater the time that has elapsed since they had a common ancestor.

The technique

Obtaining pure plankton DNA for accurate sequencing was difficult. Each foraminifer, being a single cell, has only one nucleus containing its genes. However, it may also contain several other plants or algae, which live in symbiosis within the cell, or have been taken in as food. These plants contain their own DNA. In order to separate out the foraminiferal DNA, we collected and cultured individual specimens until they reached the reproductive stage. In the reproductive process, the foraminifer digests any remaining food particles and the majority of its symbionts. At this point, the foraminifer contains hundreds of thousands of copies of its own genes but negligible quantities of contaminating DNA. We then extracted and copied the foraminifer DNA, and sequenced its rRNA gene.

Tree of life

The ribosomal RNA gene has previously been used to construct an evolutionary tree showing the relationships between all living organisms. We added our information on the rRNA gene sequences for several sub-tropical planktic foraminifer species to this evolutionary tree for comparison. We found that the Foraminifera form a single group distinct from all other eukaryotes (organisms which have a nucleus). In addition, the Foraminifera diverged very early in eukaryote evolution, before the main explosion of life forms (including all multicellular eukaryotes) that occurred 1100 million years ago, in the late Precambrian era. This is interesting because even the ancestral benthic foraminifers do not appear in the fossil record until 550 million years ago, and therefore they must have existed during the intervening time without the protection of a solid shell, leaving us no trace.



Evolutionary tree showing the early origins of the Foraminifera as a genetically distinct group (based on the sequence of the ribosomal RNA gene). The Foraminifera diverged from the 'trunk' of the tree well before the 'crown group', which consists of all higher animals and plants, as well as many other types of plankton.

Further investigation of the genetic relationships between foraminifers indicates that, contrary to the generally accepted view, planktic foraminifers evolved from more than one benthic group. Our findings suggest that non-spine-bearing foraminifers may have may have evolved earlier than the main spine-bearing species. The latter probably arose in the early Miocene, 22 million years ago, during a period of warm, stable climate. The genetic analysis has already turned up some interesting differences between species, confirming that morphology is not always a good guide to relatedness. When evolutionary changes are hidden within identical morphologies, organisms that prefer different environmental conditions will be treated as identical for the purposes of the fossil record, leading to inaccuracies in reconstructing past climates.

Future studies

We have started to map the variability of the gene in a particular species of tropical foraminifer from the Caribbean and Coral Seas. As the closure of the seaway between North and South America, 3 million years ago, prevented mixing of populations, the degree of difference in the rRNA gene can be calibrated against time. We also plan to collect planktic foraminifers from the polar oceans. The Arctic and Antarctic species look morphologically similar. But analysis of the rRNA gene will show whether the populations have remained isolated over a long period of time. If they have diverged genetically, that will suggest they have been isolated. If not, the polar populations of foraminifers must have crossed the equator during the glacial periods when the entire globe was much colder and the intensity of ocean circulation was stronger. Some cold water species that look very similar to their equivalents in the polar oceans can, surprisingly, be found as isolated populations in cool upwelling water masses in the tropical seas. The gene sequences will show whether or not these species are genetically distinct, and suggest how evolving species moved within the oceans in relation to past current systems. Information from such studies will contribute substantially to our understanding of how ocean circulation has changed through the millennia.

Kate Darling and Dick Kroon are at the Department of Geology and Geophysics, Grant Institute, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JW. Chris Wade and Andrew Leigh Brown are at the Centre for HIV Research, Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JN.

Appendix I

Materials and Methods

1. Nucleic Acid Extraction

1.1. Viral RNA and DNA Extraction

EDTA treated and heparinized whole blood samples were separated on Ficoll-Hypaque (Pharmacia). The PBMC fraction was stored in liquid nitrogen with the plasma fraction stored at -70°C. Early plasma samples analysed in the mother-child transmission studies were however stored at -20°C. The long term storage of plasma samples at -20°C made the recovery of viral RNA from these samples inefficient.

1.1.1. Viral DNA Extraction from Peripheral Blood Mononuclear Cells

DNA extraction was performed from 10^5 to 10^7 uncultured PBMCs (modified from Simmonds *et al.*, 1990).

1. Buffy coat containing approximately 10^5 to 10^7 cells was washed in 20mls RPMI and centrifuged at 1500rpm for 10 minutes.
2. The resulting cell pellet was incubated in a 400µl solution containing 110mM NaCl, 55mM Tris pH 8.0, 1.1mM EDTA, 0.5% SDS, 10mg/ml proteinase K, 20µl/ml poly A at 37°C for 2 hours.
3. Following incubation the DNA was extracted once with an equal volume of phenol, once with phenol-chloroform (1:1) and once with chloroform-isoamylalchohol (50:1).
4. The DNA was then precipitated with 1ml ethanol/40µl 3M sodium acetate pH 5.2 overnight at -20°C.
5. Viral DNA was then pelleted by centrifugation at 14000g for 15 minutes, the supernatant discarded, and the DNA pellet dried under vacuum.
6. Finally, extracted viral DNA was resuspended in 20µl diethylpyrocarbonate (DEPC) treated H_2O and stored at -70°C.

1.1.2. Viral RNA Extraction from Plasma

Viral RNA extraction was performed essentially as described previously (Zhang *et al.*, 1991).

1. 200µl plasma was subjected to ultracentrifugation at 45,000g, at 4°C for 2 hours, to concentrate viral particles.
2. The resulting pellet was resuspended in 1ml denaturing solution (2M guanidinium thiocyanate; 12.5mM sodium citrate pH7.0; 0.25% sarcosyl; 0.05M 2-

mercaptoethanol, 50% water-saturated distilled phenol and 1ug carrier RNA), 200ul of chloroform, mixed vigorously, and then incubated on ice for 15 minutes.

3. The solution was centrifuged at 14,000g for 15 minutes and the aqueous phase was then added to an equal volume of isopropanol and the RNA allowed to precipitate at -20°C for 45 minutes.
4. The RNA was then pelleted by centrifugation at 14,000g at 4°C for 15 minutes.
5. The RNA pellet was then washed in 1ml 75% ethanol and dried at 42°C for 10 minutes.
6. Finally, extracted viral RNA was resuspended in 20µl DEPC treated sterile distilled H_2O. cDNA syntheses were performed from RNA samples immediately with remaining RNA stored at -70°C.

1.2. Foraminiferal DNA Extraction

Foraminiferal DNA extraction was performed using a CTAB (cetyltrimethylammonium bromide) extraction procedure (Clark, 1992) essentially as described by Wray *et al.* (1993). Pilot studies (Darling *et al.*, 1996a) showed that this method proved to be successful in lysing foraminiferal gamete cells.

1. Three microtubes containing individual specimens of each foraminiferal species were taken from the -70°C freezer and the tips of the tubes immediately cut off to give instant access to the frozen tests.
2. The tests were lifted with a sable brush, leaving behind any residual sea water, and placed in a 0.5ml microtube containing 20µl TE buffer (10mM-Tris-Cl. pH 8.0 and 1mM EDTA [disodium ethylenediaminetetra-acetate]).
2. The tests were then crushed and the volume adjusted to 250µl with 0.1M EDTA and 0.25% SDS (sodium dodecyl sulphate).
3. A digest was carried out using Proteinase K (0.5µg/ml final concentration) at 65°C for 2 hours followed by a 1 hour incubation with 10% CTAB.
4. The incubation was followed by phenol/chloroform/isoamyl alcohol extraction and the DNA was then precipitated by incubation in ethanol at -20°C for 3 hours.
4. The DNA was pelleted by centrifugation at 13,000g for x minutes, the supernatant discarded and the pellet dried.
5. Extracted foraminiferal DNA was resuspended in 50µl of DEPC treated sterile distilled H_2O and stored at -20°C

2. Viral cDNA Synthesis

Viral RNA was reverse transcribed into cDNA by specific viral primer initiated cDNA synthesis, essentially as described by Zhang *et al.* (1991). Due to the poor storage conditions of the plasma samples obtained from many of the patients examined in these studies modifications were made to the reverse-transcription procedure of Zhang *et al.* (1991) in order to enhance cDNA synthesis from the low concentration RNA samples examined.

7.5µL of the viral RNA extract was incubated at 42°C for 1 hour in a 20µl reaction mixture containing 50mM Tris-Cl pH 8.0, 5mM $MgCl_2$, 5mM dithiothreitol (DTT), 50mM KCl, 0.05mg/ml bovine serum albumin (BSA), 10% dimethyl sulphoxide (DMSO), 600µM each dNTP, 1.5µM outer antisense primer (*env* 2 and *gag* 2; see section I.3.1), 15 units RNAsin (Promega), 10 units AMV reverse-transcriptase (Promega) and overlaid with 25µl of paraffin. Generated cDNA was then used in PCR amplification as described in section I.3.1.

3. PCR Amplification

3.1. Viral PCR Amplification

Two approaches were taken to examine HIV sequence variation. A single molecule sequencing strategy was employed to examine the viral sequence variation both within individual patients and between patients in HIV-1 transmission studies. Single target molecules were obtained through a limiting dilution approach (Simmonds *et al.*, 1990; Leigh Brown and Simmonds, 1995) in which both the DNA and cDNA were titrated prior to PCR amplification to the limit dilution point. This is the DNA/cDNA dilution at which less than 20% of subsequent PCR reactions are positive. By application of the Poisson distribution this means that only single target molecules were present at the outset of the PCR reaction in tubes which were positive following PCR amplification. By contrast, a consensus sequencing approach was employed to examine sequence variation between patients in the molecular epidemiology studies. In this, PCR amplification was performed directly, without limiting dilution of the template DNA, such that a single consensus sequence was obtained for each patient following PCR amplification.

Sequences were generated for regions of both the *env* and *gag* genes in studies presented here. An approximately 436 base pair (bp) fragment spanning the V3 loop and flanking regions of the *env* gene (positions 7029 to 7464 in the HIV-HXB2 genome (Genbank

Accession Number: K03445)) and an approximately 390 bp fragment of the p17 coding region of the *gag* gene (positions 857 to 1246 in the HIV-HXB2 genome) were amplified by limiting dilution nested polymerase chain reaction (PCR). PCR amplifications were performed in a 20µl reaction mixture containing 10mM Tris-HCl pH8.8, 50mM KCl, 1.5mM $MgCl_2$, 0.2mM each dNTP, 0.4µg each primer*, 0.5 units Taq DNA polymerase (Promega) and overlaid with 25µl of paraffin. Two rounds of PCR amplification were performed, with the oligonucleotide primers used in a nested fashion. This was necessary as it increases the sensitivity and specificity of the reaction and thus is the only method sensitive enough to amplify from a single copy of target sequence. The first amplification step was performed for 30 cycles with the outer primers, *env* 1: 5'-TACAATGTACACATGGAATT-3' (sense; position 6957-6976 HIV-HXB2 genome) and *env* 2: 5'-GGAGGGGCATACATTGC-3' (antisense; position 7520-7537 HIV-HXB2 genome) or *gag* 1: 5'-GCGAGAGCGTCAGTATTAAGCGG-3' (sense; position 795-817 HIV-HXB2 genome) and *gag* 2: 5'-TCTGATAATGCTGAAAACATGGG-3' (antisense; position 1296-1318 HIV-HXB2 genome). One microlitre of the first amplification product was then used as the template in a second amplification step performed for 30 cycles with primers, *env* 3: 5'-TGGCAGTCTAGCAGAAGAAG-3' (sense; position 7009-7028 HIV-HXB2 genome) and *env* 4: 5'-ATTCTGCATGGGAGTGTG-3' (antisense; position 7465-7482 HIV-HXB2 genome) or *gag* 3: 5'-GGGAAAAAATTCGGTTAAGGCC-3' (sense; position 833-856 HIV-HXB2 genome) and *gag* 4: 5'-CTTCTACTACTTTTACCCATGC-3' (antisense; position 1247-1270 HIV-HXB2 genome). Thermal cycling was performed with a 94°C 25 second denaturation step, annealing at 55°C for 35 seconds and extension at 68°C for 2.5 minutes. This was followed by a final 6.5 minutes 68°C extension step.

Following two rounds of amplification, the secondary PCR products were analysed by visualisation on an ethidium bromide stained agarose gel (see section I.4).

* All primers were synthesised by the Oswel DNA Service, Department of Chemistry, University of Edinburgh, on an Applied Biosystems 394 Synthesiser. All primers were HPLC purified.

3.1. Foraminiferal PCR Amplification

"Universal" eukaryotic SSU rDNA primers (Medlin *et al.*, 1988; White *et al.*, 1990) were used in the amplification of foraminiferal DNA. They were however unsuccessful in

amplifying the whole gene, even when the amplification conditions were extensively modified. Nevertheless, the terminal 3' antisense primer NS8: 5'-TCCGCAGGTTCACCTACGGA-3'* (White *et al.*, 1990) was successful in amplifying a fragment of the gene when paired with the internal sense primer NS5: 5'-AACTTAAAGGAATTGACGGAAG-3' (White *et al.*, 1990). This enabled the amplification of an approximately 1000 base pair (bp) fragment of the foraminiferal gene (equivalent positions 1151-1767 in the *Saccharomyces cerevisiae* SSU rDNA sequence) (Figure I.1), corresponding to the 30-48 region of the eukaryotic SSU rRNA secondary structure model (Neefs *et al.*, 1990). Amplification of this region often produced additional, smaller amplification products, indicating the presence of contaminants within the sample which were subsequently shown to be primarily symbiont in origin.

PCR conditions were varied to optimise amplification. Thermal cycling consisted of an initial cycle with denaturation at 94°C for 2 minutes, a 5 minute 48°C annealing step followed by extension at 72°C for 4 minutes. This was followed by 30 cycles with denaturation at 94°C for 25 seconds, a 35 second 48°C annealing step and extension at 72°C for 4 minutes; with a final extension step at 72°C for 7 minutes.

* All primers were synthesised by the Oswel DNA Service, Department of Chemistry, University of Edinburgh, on an Applied Biosystems 394 Synthesiser. All primers were HPLC purified.

4. Agarose Gel Electrophoresis

PCR products were visualised on ethidium bromide stained agarose gels. Gels were cast and run in 1x TBE buffer and 0.5µg/ml ethidium bromide was added to each gel. Generally, PCR products were screened for positives by running on a 1.5% agarose gel for 20 minutes at 150 volts. When necessary, PCR amplification products were run more slowly on 2% agarose gels at 50 volts. This permitted the separation of bands similar in size and was necessary for the gel extraction of single bands from the foraminiferal amplification products which typically consisted of multiple bands. Seakem agarose (FMC) was typically used for gel electrophoresis, although low melting point Nusieve agarose (FMC) was used when it was necessary to extract bands from the gel.

DNA fragment lengths were estimated by comparison with molecular weight markers of known size on an agarose gel.

300ml 1.5% Agarose Gel:

1. 4.5 g agarose was dissolved in 30mls 10x TBE buffer (0.89M Tris-borate, 0.89M boric acid, 0.1M EDTA pH 8.5) and 270mls distilled H_20 .
2. The gel was heated to boiling point in a microwave, cooled to approximately 50°C and 15µl ethidium bromide (0.5µg/ml) added.
3. The gel was then poured into a prelevelled electrophoresis plate and allowed to set for about 30 minutes, at room temperature.
4. Electrophoresis was carried out in 1 X TBE buffer following which DNA bands were visualised on a UV transilluminator.

5. Sequencing

All sequences were generated using an Applied Biosystems 373A automated DNA sequencer. Both Taq and T7 dye terminator sequencing chemistries were employed (detailed in Leigh Brown and Simmonds 1995).

5.1. Sequencing of HIV-1 Amplification Products

Both the sense and antisense strands of the *env* and *gag* amplification products were sequenced using a direct solid phase automated sequencing approach (detailed in Leigh Brown and Simmonds 1995). Single-stranded DNA was purified for sequencing on streptavidin coated magnetic beads (Dynal-Dynabeads M280) and sequencing was performed using an Applied Biosystems PRISM Sequenase Terminator Single Stranded DNA Sequencing Kit. The internal *env* and *gag* primers 3 and 4, used in PCR amplification, were used for sequencing (see section I.3.1).

Purification of single stranded amplified DNA on magnetic beads

1. 3µl of the primary amplification products (*env* and *gag* primers 1 and 2) were amplified in a 120µl secondary amplification using a combination of biotinylated and non-biotinylated internal primers 3 and 4.
2. 200µg (20µl) dynabeads were used to immobilise each biotinylated amplification product. The dynabeads were washed once in an equal volume of binding and washing buffer (B&W; 10mM Tris HCl pH 7.5, 1mM EDTA, 3.5M NaCl) and then resuspended in twice their initial volume of B&W buffer.
3. 100µl of the biotinylated secondary amplification product was immobilised by

incubation with 40µl of the beads solution at 48°C for 30 minutes.

4. The dynabeads, coated with the biotinylated secondary amplification product, were then placed in a magnetic separator (Dynal-MPC) and washed with 40µl B&W buffer.
5. The DNA strands were then separated by incubation of the dynabeads in 50µl 0.1M NaOH at room temperature for 10 minutes.
6. The NaOH supernatant, containing the non-biotinylated strand, was discarded, and the dynabeads, coated with the biotinylated strands, were washed once with 50µl B&W buffer and once with 50µl TE buffer (10mM Tris HCl pH 8.0, 1mM EDTA).
7. The dynabeads were finally resuspended in 14µl DEPC treated H_20.

T7 Dye Terminator Sequencing

8. Annealing of the sequencing primer was carried out by incubation of the 14µl template DNA with 5µl 5X SS MOPS Buffer (equal volume of MOPS/Mn^{2+} isocitrate solution) and 1µl of the complementary non-biotinylated primer at 65°C for 2 minutes followed by subsequent slow cooling to 30°C.
9. 4µl T7 dye terminator mix was then added to the annealed template reactions and the mix incubated at 37°C for 2 minutes. The extension reaction was then performed by the addition of 1µl T7 DNA polymerase (1.5 units) and incubation at 37°C for 10 minutes.
10. Sequencing products were then washed twice with 40µl 0.01M Tris pH 8.0, 0.1% tween 20 pH 8.0 and once with 40µl TE buffer.
11. Finally, the beads were resuspended in 3-4µl formamide/EDTA, the sequencing products denatured by heating at 90°C for 2 minutes, and the samples were then placed on ice and loaded on a 6% acrylamide sequencing gel. See section I.5.3 for details of sequencing gel preparation and sample loading.

5.2. Sequencing of Foraminiferal rDNA Amplification Products

Foraminiferal rDNA amplification products were sequenced using an automated Taq DyeDeoxy cycle sequencing approach. Amplified PCR products were first purified using a Wizard™ PCR Preps DNA Purification System (Promega) and the purified products were then sequenced using an Applied Biosystems Taq DyeDeoxy sequencing kit. Sequencing was performed using the "universal" primers NS5 (sense) and NS8 (antisense) which were used in PCR amplification of the approximately 1000 bp amplification product. In addition, internal "universal" primers NS6: 5'-GCATCACAGACCTGTTATTGCCTC-3'(sense) and

NS7: 5'-GAGGCAATAACAGGTCTGTGATGC-3' (antisense) were used to sequence from the centre of the amplified fragment to the 5' and 3' end of the fragment. Sequencing across the NS6/NS7 primer site was carried out using a foraminiferal specific primer FS3: 5'-GTGATCTGTCTGCTTAATTGC-3'(sense) which was designed 5' of the NS6/NS7 site. The location of the primer sites on the amplified foraminiferal SSU rDNA fragment is shown in Figure I.1.

Wizard™ Purification of PCR Products

1. The aqueous (lower) phase for each completed PCR reaction was transferred to a fresh 1.5ml microcentrifuge tube, avoiding the excessive carry-over of mineral oil, and 100µl of direct purification buffer was added to the PCR reaction (30-300µl) and vortexed briefly.
2. 1ml of Magic PCR Preps Resin was then added and this was vortexed briefly 3 times over a 1 minute period.
3. The resin/DNA mix was then pipetted into the syringe barrel and a vacuum applied to draw the resin/DNA mix into the minicolumn.
4. The column was then washed by adding 2ml of 80% isopropanol to the syringe barrel, and a vacuum applied to draw the solution through the minicolumn. The vacuum was drawn for an additional 1-2 minutes in order to dry the resin.
5. The minicolumn was then transferred to a 1.5ml microcentrifuge tube and the column centrifuged for 20 seconds at 12,000 g to remove any residual isopropanol.
6. The minicolumn was then transferred to a new microcentrifuge tube, 30µl DEPC treated distilled H_2O added to the column and allowed to stand for 1 minute.
7. The minicolumn was centrifuged for 20 seconds at 12,000 g to elute the bound DNA fragment.

Taq DyeDeoxy Thermal Cycling

8. For each sequencing reaction a mix was prepared containing 4µl 5X terminator ammonium cycle sequencing (TACS) buffer, 4µl dNTP mix, 4µl DyeDeoxy™ A Terminator, 4µl DyeDeoxy™ C Terminator, 4µl DyeDeoxy™ G Terminator, 4µl DyeDeoxy™ T Terminator, 4 units *Taq* DNA polymerase, 4µl sequencing primer (3.2 pmol), 1µg PCR template and the volume was adjusted to 20µl with sterile distilled H_2O. The reagents were mixed thoroughly by pipetting back and forth to ensure that the enzyme was evenly distributed and the reaction mix then overlaid with a drop of

mineral oil. The DyeDeoxy Terminators are light sensitive and care was therefore taken to ensure that they were stored in the dark whenever possible.

9. Thermal cycling was then carried out for 25 cycles in total with denaturation at 96°C for 30 seconds, annealing at 50°C for 15 seconds, and 60°C for 4 minutes.

Phenol/Chloroform Purification of Extension Products

10. 90µl sterile distilled H_2O was added to each reaction tube, vortex briefly and then centrifuge for a few seconds. The reaction was removed from below the oil with a pipet and transferred to a new tube.
11. The excess terminators were then extracted with 100µl of phenol:$H_2$0:chloroform (68:18:14). The sample was vortexed and then centrifuged for 2 minutes.
12. The upper aqueous phase was then transferred to a fresh tube and reextracted with a second 100µl aliquot of phenol:$H_2$0:chloroform. The sample was vortexed and then centrifuged for a further 2 minutes and once again the aqueous upper layer was transferred to a fresh tube.
13. The extension products were precipitated by adding 15µl of 2 M sodium acetate, pH 4.5, and 300µl of ice cold 100% ethanol and incubating on ice for 30 minutes.
14. The mixture was then centrifuged for 20 minutes at room temperature.
15. The ethanol was removed with a pipette and the pellets dried for 2 minutes on a hotblock at 90°C.
16. Finally, the dried pellets were resuspended in 3-4µl formamide/EDTA, the sequencing products denatured by heating at 90°C for 2 minutes, and the samples were then placed on ice and loaded on a 6% acrylamide sequencing gel. See section I.5.3 for details of sequencing gel preparation and sample loading. If samples were not loaded immediately, dried pellets were stored in the dark at -20°C until sequencing.

5.3. Acrylamide Sequencing Gel Electrophoresis

Sequencing products generated using both the T7 dye terminator and Taq DyeDeoxy methods were run on an Applied Biosystems 373A DNA sequencer.

Preparing Gel Casting Equipment

1. The glass plates, spacers and comb were washed with Alconox™ (Aldridge) and warm water. The equipment was then rinsed thoroughly with warm water followed by distilled water and the plates were then allowed to air dry.

2. The plates were assembled with the side and bottom edges sealed with electrical tape.

Preparation of a 6% Acrylamide Solution

3. For a 60ml gel solution (enough to pour one gel) 30g Urea, 9ml 40% acrylamide stock solution (BioRad), 20ml distilled H_20 and 0.5g mixed bed resin (Sigma) were mixed and stirred for approximately one hour until all the urea was dissolved.
4. The acrylamide solution was transferred to a 100ml graduated cylinder containing 6ml of filtered 10X TBE buffer and the volume adjusted to 60ml with distilled H_20.
5. The acrylamide solution was then filtered through a 0.2µm cellulose acetate filter unit under vacuum, and degassed for 5 minutes.

Casting the Gel

6. The acrylamide solution was poured into a 150ml beaker on ice and 300µl 10% ammonium persulfate (freshly made) and 33µl TEMED were added and mixed gently to avoid the generation of air bubbles.
7. The solution was then immediately poured between the plates using a 50ml syringe and filled to about 3-5cm from the top edge of the notched plate.
8. Once all air bubbles had risen to the surface two clamps were placed on either side of the plates. The gel casting comb was then inserted and then plates were laid in a horizontal position and the comb secured with three clamps.
9. The gel was then allowed to polymerise for a minimum of two hours.
10. Following polymerization the clamps, tape and casting comb were removed from the gel and the plates were then washed thoroughly with tap water followed by distilled H_2O to remove any excess acrylamide on the outside surface and the plates allowed to air dry.

Preparing Gels for Loading on the 373A Automatic Sequencer

11. The lower buffer chamber was placed in the automatic sequencer and the prepared sequencing gel was then locked in place.
12. The filter set was then selected from the "Configure" menu on the ABI 373A sequencer. For T7 dye terminator sequencing filter set B was selected. For Taq dyedeoxy cycle sequencing filter set A was selected.
13. The plates were then scanned to verify that they were clean. This was done using the "Plate Check" option from the "Start Pre Run" menu on the ABI 373A. If the base line was not flat in the scan window on the Macintosh, the plates were further cleaned with distilled H_2O, dried and re-scanned until the base line appeared flat.

14. The PMT setting was checked from the scan window (Y-axis of the lowest (typically the blue) line). This should be within the range of 800-1000. If outside this range then the PMT was adjusted by adjusting the PMT voltage from the "Configure" options on the sequencing machine.

15. Once complete, plate checking was concluded using the "Main menu" option, "Abort Run".

16. The sharks-tooth comb was then placed on the gel surface and the electrophoresis equipment assembled fully.

17. The upper and lower buffer chambers were filled with 1X TBE, and the wells rinsed with a syringe filled with 1X TBE buffer.

Preparing and Loading the Samples

18. A mixture containing 5µL deionized formamide to 1µL 50mM EDTA, pH 8.0 was prepared.

19. 4µL of this mixture was added to each tube for a 24 lane gel with 3µl added for a 36 lane gel. The mixture was agitated vigorously to dissolve the dry residue and the liquid was then collected at the bottom of the tube by centrifugation.

20. Prior to loading, the samples were denatured by heating at 90°C for 2 minutes, and then transferred immediately onto ice and then loaded on the sequencing gel. The samples were loaded in two batches, first the odd sample numbers and then the even. Sample wells were flushed prior to each load with 1X TBE buffer. The first batch of samples were run for 5 minutes ("Start Run" on the keypad). The run was then interrupted ("Intrpt Run") and the second batch of samples loaded and ran ("Resume Run").

21. The sample sheet for the sequence run was then completed on the Macintosh and data collection commenced.

6. References

Darling, K. F., D. Kroon, C. M. Wade, and A. J. Leigh Brown, 1996a. The isolation and amplification of the 18S ribosomal RNA gene from planktonic foraminifers using gametogenic specimens. *In* Whatley, R. C. and A. Moguilevsky. eds. Microfossils and Oceanic Environments. University of Wales, Aberystwyth Press.

Leigh Brown, A. J. and P. Simmonds. 1995. Analysis of HIV sequence variation. Chapter 11, *In* Karn, J. ed. HIV - A Practical Approach. Oxford University Press, pp 161-188.

Medlin, L. K., H. J. Elwood, S. Stickel, and M. L. Sogin, 1988., The characterization of enzymatically amplified eukaryotic 16-S like rRNA-coding regions. Gene **71**:491-499.

Neefs, J., Y. Van de Peer, L. Hendriks, and R. De Wachter, 1990. Compilation of small ribosomal subunit RNA sequences. Nucleic Acids Research **18**:2237-2242.

Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown. 1990. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. J. Virol. **64**:864-872.

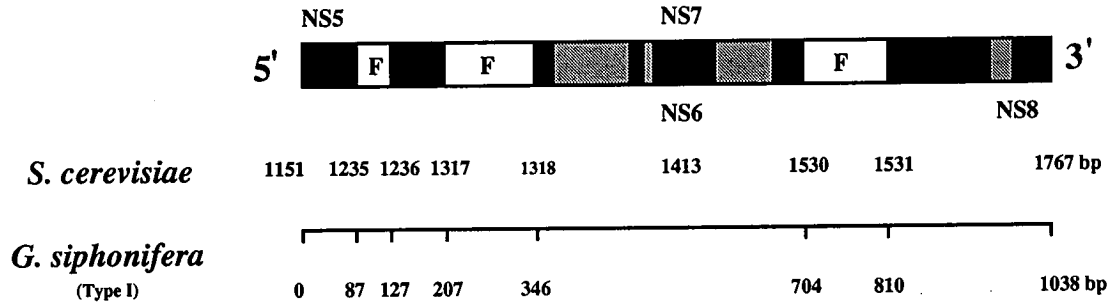
White, T. J., T. Bruns, S. Lee, and J. Taylor, 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *In* Innis, M. A., D. H. Gelfland, J. J. Sninsky, and T. J. White, eds. PCR Protocols: A guide to methods and applications. Harcourt Brace Jovanovich, San Diego. pp.315-322.

Wray, C. G., J. L. Lee, and R. DeSalle, 1993. Extraction and enzymatic characterization of foraminiferal DNA. Micropaleontology **39**:69-73.

Zhang, L. Q., P. Simmonds, C. A. Ludlam, and A. J. Leigh Brown. 1991. Detection, quantification and sequencing of HIV-1 from the plasma of seropositive individuals and from factor VIII concentrates. AIDS 5:675-681.

Figure 1.

Schematic representation of the approximately 1000 bp foraminiferal SSU rDNA amplification product indicating the positions of the primer sites used for PCR amplification and sequencing. The relative positions of the conserved regions (black), expansion segments (grey) and foraminiferal specific insertions (white) in *G. siphonifera* (Type I) are mapped against the *Saccharomyces cerevisiae* reference SSU rDNA sequence.


NS5
NS7
5'
F
F
F
3'
NS6
NS8
S. cerevisiae
1151 1235 1236 1317 1318 1413 1530 1531 1767 bp
G. siphonifera
(Type I)
0 87 127 207 346 704 810 1038 bp

Appendix II

Sequence Analysis

1. Sequence Assembly, Alignment and Translation to Amino Acid

Raw nucleotide sequences, generated using an Applied Biosystems 373A DNA sequencer, were assembled and processed using the programs TED and XBAP of the STADEN package (Staden, 1993). The sequences of the sense and antisense strands of each molecule were edited with reference to the trace data and a consensus sequence was output. Processed sequences were then aligned within version 2.2 of the Genetic Data Environment (GDE) package (Smith *et al.*, 1994), with preliminary alignment of sequences performed using the CLUSTAL V algorithm (Higgins *et al.*, 1992), and with the alignment then improved manually. During alignment, gaps were preferred to transition differences, and transition differences preferred to transversion differences. Translation of nucleotide sequences to amino acid was also undertaken within the GDE package using the universal codon table.

Alignment of HIV-1 *env* and *gag* sequences was not problematic. Sequences were aligned with reference to the translated amino acid sequences, with alignment gaps inserted to maintain the reading frame. Comparison was made with published alignments of HIV-1 subtype reference sequences (Myers *et al.* 1995).

Foraminiferal SSU rDNA sequences were aligned relative to a published reference SSU rDNA sequence alignment, including 438 eukaryote, 3 archaebacterial and 3 eubacterial taxa (obtained from the Ribosomal Database Project; Maidak *et al.* 1994), and constructed with reference to the SSU rRNA secondary structure model. Sites homologous to SSU rDNA core regions, conserved across all eukaryote taxa, were aligned relative to the published SSU rDNA alignment. Where possible, variable regions were aligned between the foraminifera. However, for many sites alignment was not possible even between closely related foraminifers. The variable regions were located in seven variable length, highly diverse expansion segments. Four expansion segments were present in all eukaryote taxa but three were unique to the formainifera. Foraminiferal SSU rDNA sequences are of unusual length, approximately twice the size of most other eukaryotes due to length variations in the expansion segments, which renders alignment complex. Sites of questionable positional homology were exluded from subsequent analyses.

2. Phylogenetic Analysis

A range of phylogenetic inference methods are available for reconstructing

evolutionary trees. These may be broadly divided into two main categories; the first based on the principal of parsimony and the second on more statistical, model-based approaches including both distance based methods and maximum likelihood.

Phylogenetic methods seek to obtain the best phylogenetic estimate for the data available based upon a defined tree selection criterion. In the case of algorithmic methods, which include all pair-group cluster analysis distance methods (for example UPGMA; Sneath and Sokal, 1973) and the neighbour-joining distance method (Saitou and Nei, 1987), the phylogenetic criterion is defined by the tree-building algorithm and the method proceeds to the inferred tree directly, with tree inference and the definition of the preferred tree combined into a single statement. Such methods generate a single best estimate phylogeny for the data. By contrast, optimality criterion methods, which include the Fitch-Margoliash distance method (Fitch and Margoliash, 1967), parsimony (Fitch, 1971) and maximum likelihood (Felsenstein, 1981), separate the phylogenetic criterion from the algorithm. Such methods define a criterion, or an objective function, by which alternative trees may be compared, with an algorithm used merely to compute the value of the objective function for a given tree and for searching for the trees that optimise this value. The Fitch-Margoliash method, defines the objective function as the net disagreement (lack of fit) between the data and the inferred phylogeny and seeks to minimise this value. Maximum likelihood attempts to maximise the objective function, the probability of the data given a tree, whilst parsimony attempts to minimise the objective function, the number of character state changes (steps) required to account for the variation in the dataset. The separation of the phylogenetic criterion for evaluating a particular tree from the search algorithm for finding the optimal tree permits all examined trees to be assigned a score and ranked according to the criterion. In this way, sub-optimal trees, which may possibly explain the data almost equally as well as the optimal tree, may also be evaluated (used in the evaluation of alternative phylogenetic hypotheses using likelihood; section 2.5.1). Thus in principle, optimality criterion methods are superior to algorithmic methods which result only in the generation of a single best estimate tree.

Optimality criterion methods may use a number of different search strategies in order to find the optimal tree under the defined phylogenetic criterion. Exact search algorithms guarantee the discovery of all optimal trees and include both exhaustive search procedures, which evaluate every possible tree under the selected criterion, and branch-and-bound methods, which reduce the number of trees evaluated by traversing the search tree and then eliminating parts of the search tree that only contain suboptimal solutions. Exact search

algorithms are however extremely computer intensive and can therefore only be applied to relatively small datasets. Consequently, this necessitates the use of approximate heuristic search algorithms, which generally operate by hill climbing methods, for most datasets. The optimality criterion programs of the PHYLIP package used in the analyses presented in this thesis use a stepwise addition, heuristic search algorithm. This works by the addition of each taxon in a stepwise manner, with the placement of each taxon optimised according to the phylogenetic criterion of the tree-building method. The optimal tree is then saved and the next taxon added. Heuristic search strategies are however prone to entrapment in local optima. Thus it is necessary to employ branch swapping algorithms and rearrangements of the sequence input order in order to maximise the probability of obtaining the optimal tree.

Simulation studies and experimental phylogenies have demonstrated that many tree construction methods are powerful enough to accurately reconstruct evolutionary histories, provided that the rates of change for the observed characters are appropriate for analysis. Such studies have also demonstrated that many methods are fairly robust to violations of their underlying assumptions (for review see Hillis *et al.*, 1994). Although all methods of phylogeny reconstruction seek to infer the evolutionary relationships between sequences in the form of a phylogenetic tree, different approaches differ in the assumptions which they make and the mechanism by which the phylogeny is reconstructed. As such, although there is only one true phylogeny for a given dataset, different phylogeny reconstruction methods may differ in the evolutionary trees which they infer. Phylogenetic analysis methods perform best under a neutral (poisson based) model of molecular evolution and it is unclear how they perform when evolution is characterised by natural selection which may lead to the occurrence of homoplasy (parallelism or convergence) in the dataset. Human immunodeficiency virus *env* gene sequences in particular are clearly not evolving in a neutral manner and as such it is unclear how different methods of phylogeny reconstruction will perform. For this reason it is therefore particularly important to employ a range of phylogeny reconstruction techniques and not rely on one particular method. As such, nucleotide sequence phylogenies have been reconstructed using distance-based methods (neighbour joining (Saitou and Nei, 1987) and Fitch-Margoliash (Fitch and Margoliash, 1967)), maximum likelihood (Felsenstein, 1981), and maximum parsimony (Fitch, 1971). Neighbour joining was also used to reconstruct phylogenies based on amino acid sequences.

All phylogenetic analyses were performed using programs taken from version 3.52c of the Phylogeny Inference Package (PHYLIP) (Felsenstein, 1993).

2.1. Estimation of the Transition/Transversion Ratio

Based on a preliminary phylogeny reconstructed for each nucleotide sequence dataset using a default transition/transversion ratio of 2.0, 12-15 "representative" sequences from points across each tree were selected. The ratio of transitions to transversions was estimated for this sample by likelihood. In this, maximum likelihood phylogenies (see section 2.3) were reconstructed using a range of transition/transversion ratios for the sample of sequences using the program DNAML, and the alternative phylogenies assigned a likelihood value. The optimal phylogeny and thus the best estimate of the transition/transversion ratio is the tree with the highest likelihood. The process was then repeated with a range of transition/transversion ratios around the optimal until the ratio was estimated to an accuracy of 0.01. Subsequent analyses were performed using the derived ratio of transitions to transversions for each dataset.

2.2. Distance Based Methods of Phylogeny Reconstruction

Distance based methods do not use discrete character state data to infer phylogenetic relationships but instead transform the original dataset into a pairwise matrix of evolutionary distances to which the phylogenetic tree is fitted. In analysing distance data, all distance matrix programs implicitly make the assumption that each distance is measured independently from the others. On the basis of their assumptions, distance methods may be subdivided into cluster analysis techniques and additive tree methods. Cluster analysis techniques, which include the unweighted pair group method with arithmetic mean (UPGMA) (Sneath and Sokal, 1973) are somewhat more stringent in their assumptions than additive tree methods. Although they make few assumptions they make the very strong assumption that the data be approximately ultrametric. Ultrametric distances are the most constrained and are defined mathematically by satisfaction of the three-point condition. In practical terms, ultrametric distances can be precisely fitted to a tree such that all pairwise distances are equal to the sum of the branch lengths that connect the respective taxa (ie. the distances are additive), but in addition, the tree can be rooted such that all taxa are equidistant from the root. Thus when applied to molecular data, the underlying assumption is that the expected rate of gene substitution is constant (ie. a universal molecular clock is assumed). It is therefore essential that the rates of evolution are approximately constant among the different lineages. For this reason cluster analyses were not undertaken in these analyses. Taken to the extreme, cluster analysis techniques form the basis for the phenetic approach to phylogeny reconstruction in

which it is asserted that the extent of similarity is all important biologically with the consideration of the historical branching order of secondary interest. Additive tree methods, of which neighbour-joining (Saitou and Nei, 1987) and the Fitch-Margoliash method (Fitch and Margoliash, 1967) have been applied in analyses presented here, cannot however clearly be classified as phenetic as although they reconstruct phylogenies based on distance measures they do not assume a direct connection between similarity and evolutionary relationship. Additive distances are less constrained than ultrametric distances and mathematically satisfy the four-point condition. Additive tree construction methods make the assumption that the data is additive (ie. distances can be fitted to an unrooted tree such that all pairwise distances are equal to the sum of the branch lengths that connect the respective taxa) but do not make the somewhat stronger assumption, characteristic of cluster analyses, that the tree can be rooted such that all taxa are equidistant from the root. Thus additive tree methods do not assume a molecular clock and as such lineage to lineage variation in the average clock rate will be tolerated. Nevertheless, it is assumed that the data come close to fitting an additive tree, and therefore correction for multiple substitutions is particularly important for data that might include lineage specific differences in average rate.

2.2.1. Transformation of Sequence Data to Distances

Nucleotide Sequences

It is essential that for the application of additive tree construction methods to distance data that correction be made for unseen multiple nucleotide substitutions at the same site (parallel or back mutations). This becomes particularly important as the distance between sequences increases as the net effect of parallelisms and reversions is that sequence similarity does not decline uniformly with the number of events. The probability of superimposition increases as the sequences become more diverse and thus if correction is not made then distances will move further from additivity as divergence increases. Ultimately, the noise to signal ratio will become so high that methods of correction can not compensate for the accumulation of noise and the relationships between the sequences can no longer be represented by a tree.

The generalised two-parameter model (maximum likelihood model) was used to correct for multiple substitutions in the sequence analyses presented here (program DNADIST). This model uses the transition probability formula of Kishino and Hasegawa (1989) which incorporate unequal rates of transition and transversion similarly to the two-parameter model of Kimura (1980), but in addition also allow for different frequencies of the

four nucleotides.

The model makes the following assumptions (modified from Swofford and Olsen, 1990):

1. All substitutions at a given sequence position are independent.
2. Every sequence position is equally subject to change.
3. The base composition is not shifting over time.
4. The rate of change to each residue type is proportional to the residue's equilibrium abundance but independent of the identity of the starting residue.
5. Insertions or deletions have not occurred. Although this assumption is obviously violated it is unclear how to treat insertion/deletion events and as such pairwise alignment gaps were excluded from each pairwise distance estimation.

Amino Acid Sequences

Distance matrices were constructed from amino acid sequence alignments under the Dayhoff PAM model of amino acid replacement using the program PROTDIST.

2.2.2. Reconstruction of Phylogenies Using Neighbour Joining

The additive tree construction method, neighbour joining (Saitou and Nei, 1987) is an algorithmic method of phylogeny reconstruction. It does not use an implicit mathematical model but works on the principal of minimum evolution. The method is fairly robust and is extremely fast. As such, neighbour joining is suitable for the analysis of large sequence datasets and has proved particularly useful in analysing the large HIV-1 transmission and epidemiology datasets presented here. Neighbour joining is a highly reliable method and produces nearly optimal trees in most circumstances, and performs well with datasets that contain high levels of noise.

Working from a matrix of estimated distances between all pairs of sequences, the net divergence (r_i) of each sequence from all the other sequences is calculated using the formula:

$$r_i = \sum_{k=1}^{N} d_{ik}$$

where N is the number of terminal nodes in the distance-matrix and d_{ik} is the distance between sequences i and k. k represents each of the sequences in the matrix in turn. The separation between each pair of sequences (nodes) is then adjusted on the basis of their divergence from all the other nodes using the formula:

$$M_{ij}=d_{ij}-(r_i-r_j)/(N-2)$$

for all i and with $j>i$. N is the number of terminal nodes, d_{ij} is the distance between nodes i and j and r_i and r_j are the net divergence of i and j respectively. Thus a modified rate-corrected distance matrix is constructed. This has the effect of normalising the divergence of each sequence for its average clock rate and thus the assumption of ultrametricity avoided. The tree is then constructed by linking the least distant pair of nodes from the modified matrix thus defining a new node u which joins nodes i and j and the rest of the tree. The lengths (s) of the branches from u to i and u to j are defined as:

$$s_{iu}=d_{ij}/2+(r_i-r_j)/[2(N-2)]$$

$$s_{ju}=d_{ij}-s_{iu}$$

where N is the number of terminal nodes. The distance from u to each other terminal node (k) is then calculated from the following equation:

$$d_{ku}=(d_{ik}+d_{jk}+d_{ij})/2$$

Thus when two nodes are linked, their common ancestral node is added to the tree and the terminal nodes with their respective branches removed. This pruning process converts the newly added common ancestor into a terminal node on a tree of reduced size. The process is then repeated (decreasing the value of N by one each time) until only two nodes (i and j) remain. This remaining branch length is given by the equation:

$$s_{ij}=d_{ij}$$

Neighbour joining phylogenies were constructed using the neighbour-joining option of the program NEIGHBOR. Negative branch lengths, which are sometimes assigned by the algorithm as it seeks to represent the data by an additive tree, were converted to zero length before plotting.

2.2.3. Reconstruction of Phylogenies Using the Fitch-Margoliash Method

Phylogenies were reconstructed using the Fitch-Margoliash least-squares method (Fitch and Margoliash, 1967) using the program FITCH. The global rearrangement and jumble options, which maximises the probability of finding the optimal phylogeny, were used in all reconstructions. Global rearrangements attempt to optimize the phylogeny by removing all possible subtrees from the tree and reinserting them in all possible positions until no further rearrangement can improve the topology. The jumble option results in the rearrangement of the input order which may affect the phylogeny inferred due to the stepwise manner in which sequences are added to the heuristic search algorithm employed by the

program.

Although the Fitch-Margoliash least-squares method is an additive phylogeny reconstruction technique it differs quite dramatically from neighbour joining in its operation. Unlike neighbour joining, the Fitch-Margoliash method is an optimality criterion method and thus does not proceed directly to the final solution by following a specific algorithm but instead defines an objective function, the net disagreement (lack of fit) between the data and the inferred tree which it attempts to minimise. The measure of lack of fit is provided by the sum of squares of the distances and thus the objective of least-squares methods is to find the tree which minimises the following sum of squares:

$$\text{Sum of Squares} = \sum_i \sum_j n_{ij} (D_{ij}-d_{ij})^2/D_{ij}^p$$

where D is the observed distance between species *i* and *j* and d is the expected distance, computed as the sum of the lengths (amounts of evolution) of the branches of the tree from species *i* to species *j*. The quantity n is the number of times each distance has been replicated (1 in this study) and p, which distinguishes the various least-squares methods is set to 2 for the Fitch-Margoliash method.

The principal behind the least-squares approach to phylogeny reconstruction is theoretically better than that of neighbour joining and other algorithmic methods. Although all additive tree methods will produce the correct phylogeny if the data is additive, without the definition of an objective function (amount to be minimised) it is difficult to determine how a method will perform when the data do not perfectly fit an additive tree. Least squares methods seek to find the optimal value for the topology and branch lengths under the additive tree model for nonideal data.

2.3. Reconstruction of Phylogenies Using Maximum Likelihood

Maximum likelihood is an optimality criterion method which uses a statistical approach to choose a phylogeny that maximises the probability of the data given a tree. Although it was first applied to the estimation of phylogenies from molecular sequences by Neyman (1971), where it was used to infer simple three-species trees, it was later developed by Felsenstein (1981) for the inference of phylogenies with an arbitrary number of sequences. The method is a character-based method of evolutionary analysis and as such does not convert the data to an overall distance but examines each individual character position. Maximum likelihood requires three elements, a particular model of molecular evolution (M),

in this case the Felsenstein maximum likelihood model, the data (D), the aligned nucleotide sequences, and the competing hypotheses (H), which are simply the alternative tree topologies with associated branch lengths. In likelihood theory there is always assumed to be a competing hypothesis. The likelihood of a hypothesis (H), the tree, is the probability of the data given the hypothesis and the model, P(D;H.M), considered as a function of the tree. The probability of all possible sets of data must add up to one, but when the data is held constant and the hypothesis (tree) varied, the different values of P(D;H.M) need not add up to one and are called likelihoods rather than probabilities. Thus the likelihood of a hypothesis (H) given the experimentally determined data (D) on the given model (M), is given by:

$$L_D(H)=P(D;H.M)$$

The maximum likelihood method simply chooses that hypothesis (H), the tree, which maximises the likelihood thus maximising the probability that the observed data would have occurred under the given model.

The model employed by the maximum likelihood method is the generalised two-parameter model (maximum likelihood model), which uses the transition probability formulas of Kishino and Hasegawa (1989). The model is similar to the two-parameter model of Kimura, but in addition to incorporating different rates of transition and transversion it also allows for different frequencies of the four nucleotides, so long as the substitution rates are balanced so as to maintain the equilibrium abundance. The same model was also used in the estimation of distances for use in the neighbour joining and Fitch-Margoliash methods. A number of assumptions are inferred by the model (taken from Felsenstein, 1993):

1. Each site in the sequence evolves independently.
2. Different lineages evolve independently.
3. Each site undergoes substitution at an expected rate which is chosen from a series of rates (each with a probability of occurrence) which we specify.
4. All relevant sites are included in the sequence, not just those that have changed or those that are "phylogenetically informative".
5. A substitution consists of the replacement of the existing base with either:
    a. A base drawn from a pool of purines or a pool of pyrimidines (depending on whether the base being replaced was a purine or a pyrimidine). This can lead to either no change or to a transition.
    b. A base drawn at random from a pool of bases at known frequencies, independently of the identity of the base which is being replaced. This could

lead to either no change, a transition or a transversion.

Maximum likelihood phylogenies were generated using the program DNAML. The global rearrangement and jumble options, discussed earlier in the section on the Fitch-Margoliash method (section 2.2.3) were used in all reconstructions.

2.4. Reconstruction of Phylogenies Using Parsimony

Parsimony is a character-based optimality criterion method of phylogeny reconstruction which attempts to find the tree topology that minimises the number of character state changes (steps) required to account for the observed variation in the dataset, thus minimising the total length of the tree. Thus for sequence data, parsimony attempts to find the unrooted phylogeny which requires the minimum number of base substitutions. The data are first reduced by the elimination of all invariant sites and then secondly by the elimination of those sites where a base change is found only in a single sequence. The phylogenetic tree is then constructed using the remaining 'phylogenetically informative' sites (sites where residues appear in more than one sequence) with the tree topology which minimises the total number of substitutions at these sites selected.

If within a group only a single change had occurred at each site then it would clearly be possible to construct a tree in which the newly arisen base at each site would be shared by all species descended from the lineage in which the change occurred. Within the tree, sets of species having the new bases would be either perfectly nested or disjoint. However, when character conflicts occur assumptions of homoplasy (convergence, parallelism, or reversal) must be invoked in order to explain the data. The principal of parsimony essentially maintains that simpler hypotheses are preferable to more complicated hypotheses. Thus a tree which minimises the number of steps also minimises the number of additional hypotheses (homoplasies) required to explain the data. Parsimony methods typically perform well when evolution is generally divergent in pattern but perform poorly when there are a large number of homoplasies in the dataset. This has important implications for the analysis of HIV-1 sequence data, in particular analyses of the V3 region of the *env* gene. *Env* V3 has been implicated as being involved in escape from immune selection and if there are also constraints on the direction of sequence change then convergent substitutions will occur (Holmes *et al.*, 1992).

There are a range of parsimony approaches to phylogeny reconstruction and although they all attempt to minimise some significant evolutionary quantity they vary in the

evolutionary assumptions which they make (Swofford and Olsen, 1990). In this study parsimony phylogenies were reconstructed using the maximum parsimony program DNAPARS. The global rearrangement and jumble options, discussed earlier in the section on the Fitch-Margoliash method (section 2.2.3) were used in all reconstructions. Where more than one equally most parsimonious tree was found they were represented as a single majority-rule consensus tree using the program CONSENSE. DNAPARS carries out unrooted parsimony, analogous to Wagner parsimony (Kluge and Farris, 1969, Farris, 1970) but generalised (Fitch, 1971) to allow unordered multistate characters (eg. nucleotide and amino acid sequences) to be analysed. The assumptions of the method are as follows (taken from Swofford and Olsen, 1990 and Felsenstein, 1993):

1. Any transformation from one character state to another also implies a transformation through any intervening states, as defined by the ordering relationship.
2. Character states are freely reversible (ie. character state change in either direction is assumed to be equally probable and character states may change from one state to another and back again).
3. Each site evolves independently.
4. Different lineages evolve independently.
5. The probability of a base substitution at a given site is small over the lengths of time involved in a branch of the phylogeny.
6. The expected amounts of change in different branches of the phylogeny do not vary by so much that two changes in a high-rate branch are more probable than one change in a low-rate branch.
7. The expected amounts of change do not vary enough among sites that two changes in one site are more probable than one change in another.

These assumptions are however controversial and are not accepted by all authors. One argument for the justification of parsimony methods is the hypothetico-deductive approach, favoured by the systematic approach to phylogeny reconstruction. This makes the assertion that each additional character state change is considered to require an additional evolutionary hypothesis and thus if each extra hypothesis is considered to be refuting the correctness of the phylogenetic tree, then parsimony can be viewed as accepting the least rejected hypothesis. Along with this approach to the justification of parsimony methods goes the assertion that the use of parsimony requires no substantive assumptions about evolutionary processes. However, the validity of the hypothetico-deductive approach to the justification of

parsimony is itself in question with the relation of parsimony to Popper's hypothetico-deductive model of falsification of scientific hypotheses in question, as falsification, in the case of parsimony, is not absolute (for review see Felsenstein, 1988). Thus statistical concepts for the justification of parsimony, favoured by Felsenstein (see Felsenstein, 1988), must be considered. Such justifications appear to be valid only when expected amounts of character change are low and thus parsimony methods would be expected to provide a reasonable estimation of the phylogeny when divergence is low.

2.5. Reliability of Inferred Phylogenies

Two approaches were employed in the assessment of the reliability of the inferred phylogenies. The first, the bootstrap, was used to provide a general indication of support across the tree, whilst the second, likelihood, was used to evaluate specific alternative hypotheses.

2.5.1. Placing Confidence Limits on Phylogenies Using the Bootstrap

The bootstrap is a general purpose resampling technique which can be used to place confidence limits on statistics which are estimated without knowledge of the underlying distribution. It attempts to infer the variability in the distribution from which the data was drawn (in the case of sequence data the underlying sequence distribution from which the sampled sequence region was taken) by resampling points from the data. For its application to the estimation of confidence limits of internal branches in phylogenetic analyses (Felsenstein, 1985) it requires the assumption that the data points can be treated as having evolved independently on the same phylogeny according to a stochastic process. The bootstrap works by resampling points from the dataset (sites in the sequence alignment) at random with replacement until a new dataset of the same size as the original is generated. Thus, within a given bootstrap replication some sites will be represented several times whereas others will not be included. This resampling procedure is repeated a given number of times and for each of these resampled datasets, an estimate of the statistic (phylogeny) generated. Hedges (1992) has shown that 2000 bootstrap replications are required to assign significance at the 95% level (1825 bootstrap replications are required to attain an accuracy of 1% at a bootstrap proportion of 0.95). However, work by Hillis and Bull (1993) has shown that although the number of bootstrap iterations are of importance if a precise measure of the bootstrap proportion is required, the statistical meaning of such a measure is obscure and the

bootstrap does not provide a direct measure of the biologically more important concepts of repeatability (the probability that a given result will be repeated again given a new sample of characters) and accuracy (the probability that a given result represents the true phylogeny). Their work has indicated that under conditions thought to be typical of most phylogenies a clade supported in more than 70% of bootstrap replicates is likely to be correct (probability that the clade is correct is greater than 95%) whereas below 30% the branch is likely to be incorrect.

The inferred monophyletic groups that occur in the majority of bootstrap replicates may be visualised by the generation of a majority-rule consensus tree from the bootstrap phylogenies. Confidence limits may then be placed on the occurrence of particular monophyletic groups by the percentile method, in which a group represented in 95% of bootstrap replicates can be considered to be supported (monophyletic) at the 95% confidence level. Nevertheless, this application is only valid if the group to be tested for monophyly is chosen *a priori* as otherwise there is the possibility of running into a "multiple tests" problem (Felsenstein, 1985, 1988). At the 95% level one branch of twenty would be expected to show significance purely by chance. Hillis and Huelsenbeck (1994) have also examined problems of summarising the bootstrap results as the percentage occurrence of clades within a majority-rule consensus tree. Their work, which was applied to the Florida dentist HIV transmission case, showed that individual sequences may be consistent with a given hypothesis (in this case consistent with the dental transmission hypothesis) in every bootstrap replicate and yet not be present in the majority rule consensus tree. They concluded in this case that the majority-rule consensus tree was not adequate for identifying the patients who may have been infected by the dentist and that whether or not a tree is consistent with the dental transmission hypothesis for any given patient should be dependent only on the placement of the patient sequences with respect to sequences of the dentist and local controls and not those of other patients.

The application of the bootstrap to the phylogenies presented here was performed using the programs SEQBOOT and CONSENSE. Bootstrap values were assigned to neighbour joining (2000 replications), Fitch-Margoliash (100 replications), maximum parsimony (100 replications) and maximum likelihood (10 replications) trees when possible. Neighbour joining is a relatively fast method of tree construction and as such 2000 bootstrap replicates were applied routinely to all neighbour-joining trees irrespective of the size of the dataset. It was not possible to apply the bootstrap to other methods in analyses of large

datasets.

2.5.1. Evaluation of Alternative Phylogenetic Hypotheses Using Likelihood

Alternative phylogenetic hypotheses (phylogenetic trees) were evaluated statistically using the Kishino-Hasegawa-Templeton likelihood ratio test (Kishino and Hasegawa, 1989). This is a nonparametric test which uses the mean and variance of log-likelihood differences between trees, taken across sites, to assess the significance of observed differences in likelihood between the trees. If the mean is more than 1.96 standard deviations different then the two trees are significantly different at the 0.05% level. The test was performed using the maximum likelihood phylogeny inference program DNAML into which the phylogenies to be compared were input as user defined trees and assigned a likelihood value (log-likelihood). Branch lengths were optimised under the constraint of the given tree topology within DNAML. The test was then performed by the comparison of each alternative phylogeny with that phylogeny assigned the highest likelihood (the optimal phylogeny) and on the basis of this test topologies declared to be either significantly or not significantly worse. Alternative user-defined hypotheses for evaluation were generated using the program RETREE which permits the movement of sequences within the tree to generate alternative sequence topologies.

One criticism of the likelihood ratio tests employed is that they evaluate a specific, alternative user-generated tree topology, but do not optimise the topology in order to obtain the best tree under the evolutionary constraint under test. Although branch lengths were reestimated for the user-defined tree topology, the topology was held constant. Thus, whilst the method provides a test of the 2 specific trees under study, it does not permit the optimisation of the topology under the topological constraint imposed. Consequently this may lead to the erroneous rejection of the alternative hypothesis based upon a sub-optimal tree. The latest version of PAUP provides an option for the generation of optimal trees under a given constraint. However, this option was not available for the likelihood evaluation of alternative tree topologies presented in this thesis with the latest version of PAUP still currently unavailable on general release.

3 Principal Coordinates Analysis

Principal coordinates analysis (PCOORD) is an ordination method which can be used to take a set of objects, initially arranged in a high-dimensional space (defined by the

measurements made on the objects) and represent these in a small number (two or three) principal dimensions. Principal coordinates analysis has been applied to the analysis of sequence data by Higgins *et al.* (1992). The raw distances are initially transformed to ensure that they approach Euclidean distances and principal coordinates analysis is then used to determine which combinations of variables (principal coordinates) is best suited to reflect the variation in the data. Unless the objects are equidistant, as would happen if all objects are randomly related to each other, some dimensions account for far more of the information in the datset than others. By plotting the objects along the most significant two or three dimensions, the major trends and groupings in the data can be observed.

Principal coordinates analysis makes very few assumptions about the structure of the data, unlike cluster analysis methods which all assume an underlying hierarchial structure in the data as a result of the evolutionary process. It therefore forms a highly useful method for the identification of major trends in a dataset where there is little structure. Principal coordinates analysis has been employed here to examine the relationships between HIV-1 sequences obtained from intravenous drug users, homosexual men and haemophiliacs in six cities in Britain and Ireland (Section A, Paper I). Although traditional phylogenetic analysis techniques clustered intravenous drug user and homosexual/haemophiliac sequences distinctly in the phylogenetic tree, the inferred trees showed very little internal structure and consequently low support for the intravenous drug user clade. Principal coordinates analysis provided additional support for the the distinction of risk group sequences.

4. References

Farris, J. S., 1970. Methods for computing Wagner trees. Syst. Zool., 19:83-92.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet., 22:521-565.

Felsenstein, J. 1993. PHYLIP manual version 3.52c. Berkeley University Herbarium, University of California, Berkeley.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. Science 155:279-284.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. 20:406-416.

Hedges, S. B., 1992. The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies. Mol. Biol. Evol. 9:366-369.

Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam, and A. J. Leigh Brown, 1992. Convergent and divergent 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.

Higgins, D. G., 1992. Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. Comput. Applic. Biosci. 8:15-22.

Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Applic. Biosci. 8:189-191.

Hillis, D. M., and J. J. Bull, 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182-192.

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

Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham. 1994. Application and accuracy of molecular phylogenies. Science 264:671-677.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol., 16:111-120.

Kishino, H., and M. Hasegawa, 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order of the Hominoidea. J. Mol. Evol. 4:406-425.

Kluge A. G and J. S. Farris, 1969. Quantitative phyletics and the evolution of Anurans. Syst. Zool., 18:1-32.

Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olson, K. Fogel, J. Blandy, and C. R. Woese. 1994. The ribosomal database project. Nucl. Acids Res. 22:3484-3487.

Myers, G., B. Korber, B. H. Hahn, K. T. Jeang, J. W. Mellors, F. E. McCutchan, L. E. Henderson, and G. N. Pavlakis. 1995. Human retroviruses and AIDS 1995, Los Alamos National Laboratory.

Neyman, J., 1971. Molecular studies of evolution: a source of novel statistical problems. In "Statistical Decision Theory and Related Topics" (Gupta, S.S. and Yackel, J., eds.). Academic Press, New York. pp. 1-27.

Saitou, N., and M. Nei, 1987. The neighbor-joining method: a new method for reconstructing evolutionary trees. Mol. Biol Evol. 4:406-425.

Smith, S. W., R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet, 1994. The genetic data environment and expandable GUI for multiple sequence analysis. Comput. Appl. Biosci. 10:671-675.

Sneath, P. H. A., and R. R. Sokal, 1973. Numerical Taxonomy. Freeman, San Francisco.

Staden, R., 1993. Staden Package Update. Genome News 13:12-13.

Swofford, D. L. and G. J. Olsen, 1990. Phylogeny reconstruction. *In* "Molecular Systematics" (Hillis, D. M. and Moritz, C., eds.). Sinauer Associates Inc. Publishers, Sunderland, Massachusetts, USA. pp 411-501.