Molecular Epidemiology of HIV-1 in the United Kingdom

by

Stéphane Hué, M.Sc.

Centre for Virology

Royal Free and University College Medical School

University College London

Thesis submitted for the degree of Doctor in Philosophy

February 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

HIV infection is now the fastest-growing serious health hazard in the United Kingdom (UK), with an estimated 53,000 infected adults at the end of 2003. Despite a recent increase in heterosexually acquired infections, the most prevalent clade of virus within the country remains subtype B, from the main group of HIV-1, which is mainly transmitted through sex between men. To date, very little is known about how subtype B successfully invaded the British population, and how the virus has subsequently spread and evolved. Given that molecular data on HIV-1 is becoming increasingly available since the introduction of routine gene sequencing for drug-resistance monitoring, the present thesis proposes to assess the reliability of the HIV-1 *pol* gene for molecular analyses of epidemiological relevance. Identification of transmission networks by phylogenetic means were primarily conducted, with the further goal to investigate the dynamics of HIV-1 transmission at both individual and population level in the UK. Evolutionary and epidemiological approaches were then combined in order to assess the correlates of transmission within a population of primary HIV-1 infected individuals within a localised risk group, exploiting both molecular and clinical data. Finally, the epidemic history of HIV-1 subtype B in the UK was reconstructed from sampled HIV-1 *pol* gene sequences, providing new insights into the complexity of HIV-1 epidemics that must be considered when developing monitoring and prevention initiatives. The analyses presented in these pages emphasizes the advantage of combining state-of-the-art epidemiological studies to phylogenetic frameworks when investigating the dynamics of a viral epidemic as complex as HIV-1.

Acknowledgments

The reasons to thank the persons I wish to thank are legion. In their own ways, they all provided me with unconditional support, trust and anxiety. Not only did this thesis blossom through contact with them, but I did too. I am first and foremost indebt to my supervisor Dr Deenan Pillay for his trust, aid and irony. I owe Dr Jonathan Clewley for his expertise, Dr Patricia Cane for her assistance. I owe Dr Paul Kellam and the Viral Genomic and Bioinformatics Group for their warmth and proficiency. I owe Dr Oliver Pybus and Dr Andy Rambaut for their insight. I owe many people at the University of Birmingham, Birmingham Heartlands Hospital, Central Health Protection Agency, Windeyer Institute of Medical Research and the sordid King & Queen for their generosity. I owe Catherine Gale for her kind-hearted altruism. I owe my friends for their praiseworthy patience and my family for the unreasonable faith they have in me. I owe Sarah Abraham for making me do it.

Contributions to this Thesis

Chapter III-IV-V

The HIV-1 *pol* gene used in this thesis were extracted, amplified and sequenced by Judith Workman and Daina Ratcliffe at the Health Protection Agency Antiviral Susceptibility Reference Unit, Birmingham Heartlands Hospital, Birmingham, UK.

Chapter IV

Dr David Pao, Dr Martin Fischer and Dr Gillian Dean from the Department of GU Medicine, Brighton and Sussex University Hospitals, Brighton, UK, recruited the study cohort and collected clinical data for each patient. Professor Caroline Sabin from the Department of Primary Care and Population Sciences, Royal free & University College Medical School, London, UK performed all statistical analyses.

Chapter V

The sequences used for the estimation of the HIV-1 subtype B *pol* gene evolutionary rate were generated by the Department of Human Retrovirology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. Dr Oliver Pybus and Dr Andrew Rambaut provided assistance in the implementation of the methods described in this chapter and contributed to the interpretation of the results.

Contents

Figures

Chapter I 1

Chapter II 31

Chapter III **63**

Chapter IV **82**

Chapter V **103**

Tables

To study history one must know in advance that one is attempting something fundamentally impossible, yet necessary and highly important. To study history means submitting to chaos and nevertheless retaining faith in order and meaning. It is a very serious task, young man, and possibly a tragic one.

Hermann Hesse, The Glass Bead Game

CHAPTER I

A Biased Introduction to HIV-1

1. Human Immunodeficiency Virus & AIDS

The first clinical evidence of a new immunodeficiency disease was reported in 1981 in the United States, as an outbreak of opportunistic infections among immunocompromised gay men (Gottlieb et al. 1981). The syndrome, initially referred to as Gay Related Immune Deficiency (GRID), was renamed AIDS, for Acquired Immune Deficiency Syndrome. An association between the disease and a retroviral agent was proposed soon after (Barre-Sinoussi et al. 1983; Gallo et al. 1983). The pathogen, initially named human T-cell leukaemia virus type III (HTVL-III) (Gallo et al. 1984), is nowadays known as human immunodeficiency virus (HIV). From that (yet early) point, it became obvious that the epidemic had already grown out of proportion worldwide (Jaffe et al. 1983; Pape et al. 1983; Selik et al. 1984; Curran et al. 1985; Hardy et al. 1985; Mccormick et al. 1987). Two divergent AIDS-related viruses have been distinguished to date, namely HIV type 1 (HIV-1) and type 2 (HIV-2), whose closest relatives, the simian immunodeficiency viruses (SIVs), infect other primates (Hahn et al. 2000; Sharp et al. 2005). While HIV-2 remains most prevalent in West Africa and former Portuguese colonies (Wilkins et al. 1993; Schim Van Der Loeff and Aaby. 1999), HIV-1 has spread uncontrollably in human populations and accounts nowadays for an overwhelming majority of the global epidemic.

1.1. Organisation of HIV-1

Human immunodeficiency viruses belong to the primate lentivirus serogroup, genus *Lentivirus*, family *Retroviridae*. Mature virions (reviewed in Turner and Summers. 1999) have an average diameter of 80 to 100 nm and are enveloped by a lipid bilayer derived from the host-cell membrane. Surface projections are exposed evenly over the envelope, comprising viral surface glycoproteins (gp120 and gp41) as well as cellular membrane proteins derived from the host cell. Between the membrane and a cone-shaped viral core lies a shell of matrix proteins (p17). The capsid core, which consists of capsid proteins (p24), lies at the centre of the virus and encloses two copies of the viral genome stabilized as a ribonucleoprotein complex with nucleocapsid proteins (p7). Essential, virally encoded enzymes (i.e. protease, reverse transcriptase and integrase) as well as accessory regulatory proteins (i.e. Nef, Vif and Vpr) are also encapsidated. A schematic structure of HIV-1 is presented in Fig.1.1.

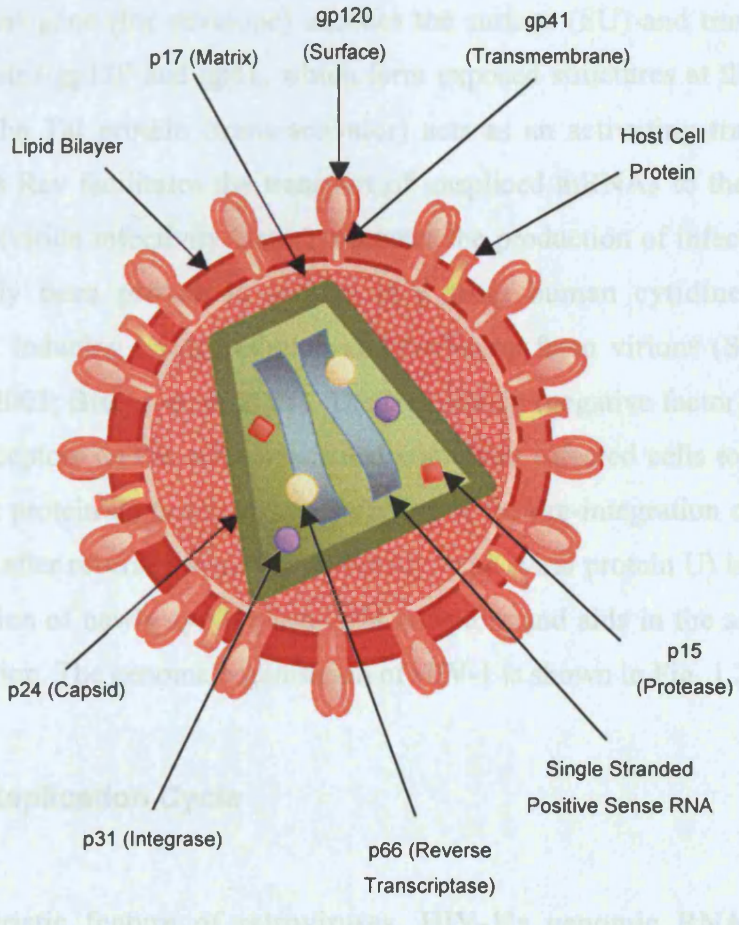


Fig.1.1. Schematic representation of the HIV-1 virion

HIV-1 genome consists of two copies of positive sense, single stranded RNA, of approximately 9200 nucleotides (nt). Each RNA strand bears nine genes, encoding for 14 viral proteins (reviewed in Frankel and Young. 1998) and nested between terminal repeated sequences of about 600 nt, known as long terminal repeats (LTR). A brief description of the genes and their products is given in Table 1.1.

The three principal genes of HIV-1 genome, namely *gag, pol* and *env*, are responsible for the synthesis of the structural and enzymatic proteins of the virus, whereas six accessory genes encode for regulatory (*tat* and *nef*) and auxiliary (*vif, nef, vpr* and *vpu*) factors (Ratner et al. 1985). The *gag* gene (for group specific antigen) encodes for the precursor of the internal structural proteins of HIV-1, processed to form mature proteins of the matrix (MA), the capsid (CA) and the nucleocapsid (NC). The *pol* gene (for polymerase) encodes the enzymatic proteins of the virus, such as the reverse transcriptase (RT, DNA polymerase coupled with RNase H activity), the protease (PR, mediator of the Gag-Pol polyproteins cleavage and maturation) and the integrase (IN, responsible for the integration of the DNA provirus into the host genome). The *env* gene (for envelope) encodes the surface (SU) and transmembrane (TM) glycoproteins gp120 and gp41, which form exposed structures at the surface of the host cell. The Tat protein (trans-activator) acts as an activating transcriptional protein, whereas Rev facilitates the transport of unspliced mRNAs to the cytoplasm. The Vif protein (virion infectivity factor) promotes the production of infectious virions and has recently been proven to protect HIV from human cytidine deaminase APOBEC3G by inducing its degradation and exclusion from virions (Sheehy et al. 2002; Yu et al. 2003; Bishop et al. 2004). The Nef protein (negative factor) reduces the level of CD4 receptors on the cell surface and stimulates infected cells to divide. The Vpr factor (viral protein R) promotes the transport of the pre-integration complex into the host nucleus after reverse transcription. Finally, Vpu (viral protein U) is responsible for the degradation of newly synthesised CD4 receptors and aids in the assembly and release of the virion. The genome organisation of HIV-1 is shown in Fig. 1.2.

1.2. HIV-1 Replication Cycle

Characteristic feature of retroviruses, HIV-1's genomic RNA is reverse transcribed by the RT protein into viral DNA prior to integration into the host genome.

Table 1.1. HIV-1 genes and their products

Gene		Protein			
Name	Position	Name	Size	Fonction	localisation
gag	**790-1186**	**Membrane anchoring (MA)**	**p17**	**env interaction; nuclear transport of viral core**	**virion**
	1186-1879	**Capsid (CA)**	**p24**	**core capsid**	**virion**
	1921-2086	**Nucleocapsid (NC)**	**p7**	**RNA binding**	**virion**
	2134-2292		**p6**	**Vpr binding**	**virion**
pol	**2253-2550**	**Protease (PR)**	**p15**	**gag/pol cleavage and maturation**	**virion**
	2550-4230	**Reverse transcriptase (RT)**	**p66**	**reverse transcription**	**virion**
	4612-5096	**Integrase (IN)**	**p31**	**proviral DNA integration**	**virion**
vif	5041-5619	Vif	p23	virion maturation and infectivity; protection from human APOBEC3G	cytoplasm
vpr	5559-5850	Vpr	p10-15	nuclear localisation of preintegration complex; inhibition of cell division	cytoplasm; virion virion; nucleus
tat	5831-6045, 8379-8469	Tat	p16/p14	viral transcriptional factor	nucleolus; nucleus
rev	5970-6045, 8379-8653	Rev	p19	RNA transport; stability and utilisation factor	nucleolus; nucleus
vpu	6045-6310	Vpu	p23	extracellular release of viral particules; CD4 degradation in the ER*	integral membrane protein
env	**6225-7758**	**Surface (SU)**	**gp120**	**external glycoprotein; CD4 and co-receptors binding**	**plasma membrane; virion envelope**
	7758-8795	**Transmembrane (TM)**	**gp41**	**transmembrane glycoprotein**	**plasma membrane; virion envelope**
nef	8797-9417	Nef	p27-p25	CD4 and HLA class I protein down regulation	plasma membrane; cytoplasm

Principal genes are indicated in bold

* Endoplasmic reticulum

HIV-1 infects CD4 bearing macrophages and T-helper lymphocytes, through a replication cycle involving several steps, divided into two phases. The early phase encompasses a succession of processes leading to the integration of the proviral DNA into the host cell genome. The late phase includes regulation of the viral gene expression, synthesis, maturation and release of viral particles.

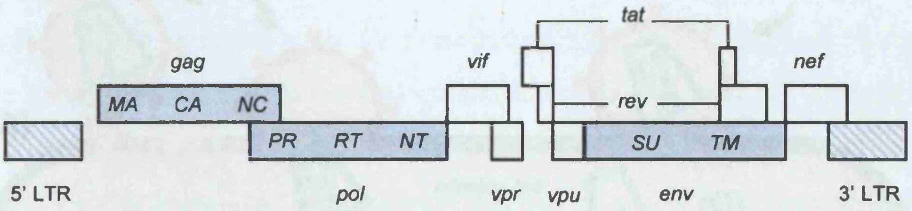


Fig. 1.2. Organisation of HIV-1 genome. Each RNA strand harbors three principal genes (*gag, pol* and *env;* in blue), and 6 accessory genes (*vif, vpr, vpu, tat, rev, nef*; in white), nested between long terminal repeats (LTR, hashed boxes). The *gag* gene encodes for proteins of the matrix (MA), the capsid (CA) and the nucleocapsid (NC); *pol* encodes for the reverse transcriptase (RT), the protease (PR) and the integrase (IN); *env* encodes the surface (SU) and transmembrane (TM) glycoproteins.

1.2.1. Early Phase of Replication

HIV infection begins with a specific interaction between the viral glycoprotein gp120 and the host-cell surface molecule CD4, a protein ordinarily involved in antigen recognition (Dalgleish et al. 1984). A supplementary interaction with the host chemokine co-receptors CCR5 or CXCR4, depending on viral strains (Choe et al. 1996), is further required to trigger membrane fusion, allowing the internalization of the viral core (Fig. 1.3A) (Turner and Summers. 1999). Partial uncoating of the viral core follows, releasing the viral RNA into the host cell (Fig. 1.3B). The ribonucleoprotein complex is then released into the cell cytoplasm (Fig. 1.3C), initiating the conversion of the viral RNA into a double-stranded DNA copy, known as provirus, by the reverse transcriptase (Fig. 1.3D). The consecutive steps of the reverse transcription, during which several template switching occur between the two RNA genomes, are summarised in Fig. 1.4.

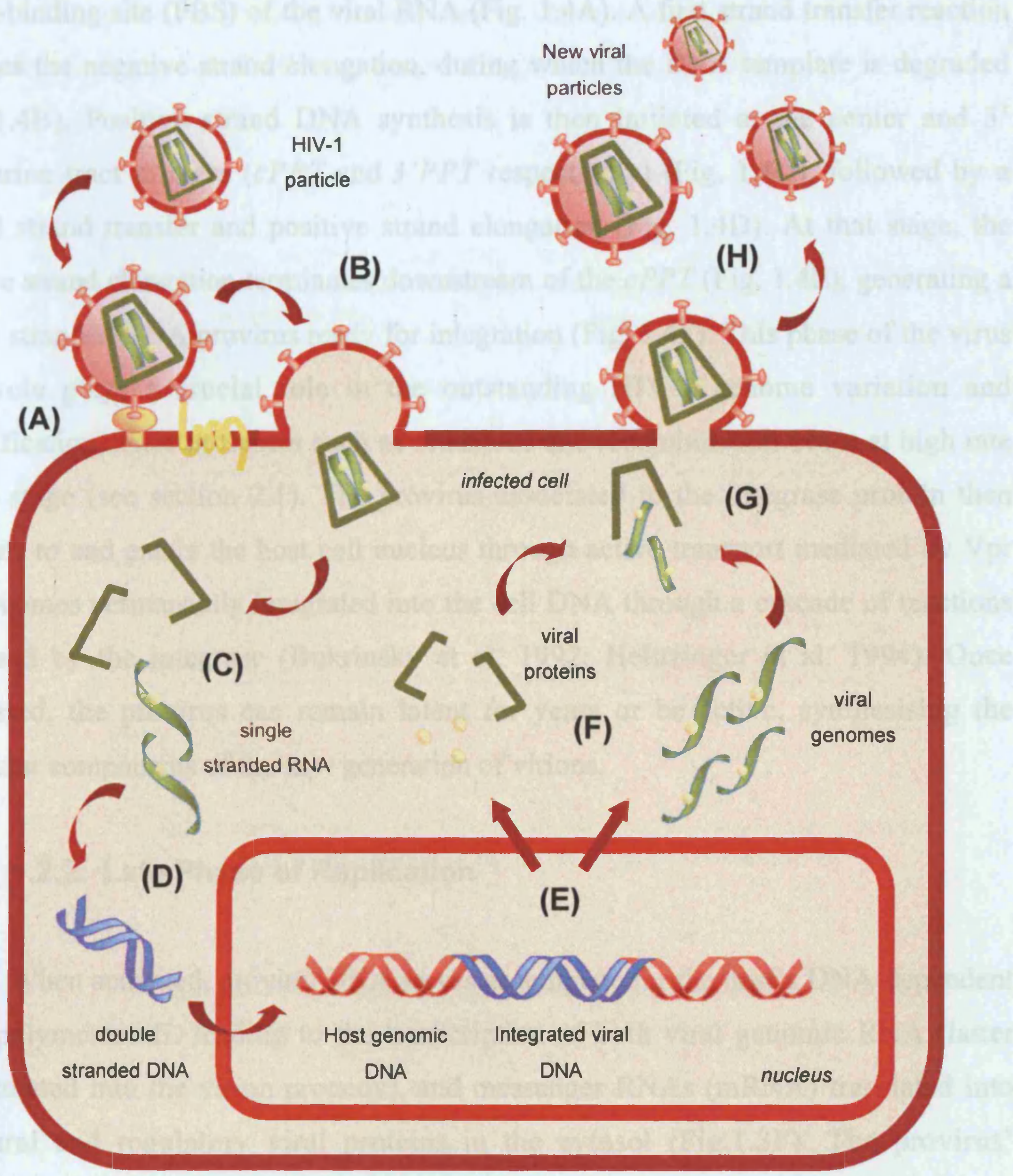


Fig. 1.3. The HIV-1 replication cycle. HIV-1 infection begins with an interaction between HIV envelope proteins and both the CD4 and chemokine receptors of the cell (A), triggering membrane fusion (B). After entry of the virion, partial uncoating (C) and reverse transcription of the viral RNA (D) occur in the cytoplasm of infected cells. The subsequent double-stranded DNA product is transported to the nucleus, where it integrates into chromosomal DNA (E). The integrated viral DNA serves as a template for DNA-dependent RNA polymerase and leads to the production of mRNAs that are translated into viral proteins in the cytosol (F). Viral protein precursors and genomic RNA are transported to the inner region of plasma membrane, where progeny virus particles "bud" from cells (G) and are released as immature particles (H). Subsequent proteolysis will generate mature particles.

First, negative strand DNA synthesis occurs, initiated by the binding of a tRNA^{lys} to the primer-binding site (PBS) of the viral RNA (Fig. 1.4A). A first strand transfer reaction precedes the negative strand elongation, during which the RNA template is degraded (Fig. 1.4B). Positive strand DNA synthesis is then initiated at the center and 3' polypurine tract primers (*cPPT* and *3'PPT* respectively) (Fig. 1.4C), followed by a second strand transfer and positive strand elongation (Fig. 1.4D). At that stage, the positive strand elongation terminates downstream of the *cPPT* (Fig. 1.4E), generating a double stranded DNA provirus ready for integration (Fig. 1.4F). This phase of the virus life cycle plays a crucial role in the outstanding HIV-1 genome variation and diversification, since processes such as mutations and recombination occur at high rate at this stage (see section 2.1). The provirus associated to the integrase protein then migrates to and enters the host cell nucleus through active transport mediated by Vpr and becomes permanently integrated into the cell DNA through a cascade of reactions catalysed by the integrase (Bukrinsky et al. 1992; Heinzinger et al. 1994). Once integrated, the provirus can remain latent for years or be active, synthesising the molecular components of the new generation of virions.

1.2.2. Late Phase of Replication

When activated, proviral DNA serves as template for the host's DNA-dependent RNA polymerase II, leading to the transcription of both viral genomic RNA (latter encapsulated into the virion progeny), and messenger RNAs (mRNA) translated into structural and regulatory viral proteins in the cytosol (Fig.1.3F). The provirus' transcription is regulated by the interaction of host factors with the viral promoter located in the 5′ LTR. Transcription is also enhanced by the viral protein Tat, which binds to the transactivating responsive sequence (TAR), an RNA element responsible for viral transcription initiation and elongation from LTR promoter (Feng and Holland. 1988; Roy et al. 1990; Feinberg et al. 1991). Unspliced or partially spliced transcripts are exported from the nucleus to the cytoplasm by active transport mediated by the viral Rev protein. Translation of the gp160 Env precursor is undergone within the endoplasmic reticulum and Golgi apparatus, whereas the Gag and Gag-Pol polyproteins are synthesized by cytoplasmic ribosomes. Translation starts by the *gag* domain and Gag-Pol transcripts are generated via a frameshift process, which allows the termination codon between the two genes to be bypassed (Ratner et al. 1985; Jacks et al. 1988).

HIV-1 polyproteins Gag and Gag-Pol are produced at estimated 9:1 to 19:1 stoichiometry ratios, depending on whether in vitro or in vivo systems are used to analyze the phenomenon (Park and Morrow. 1991; Parkin et al. 1992).

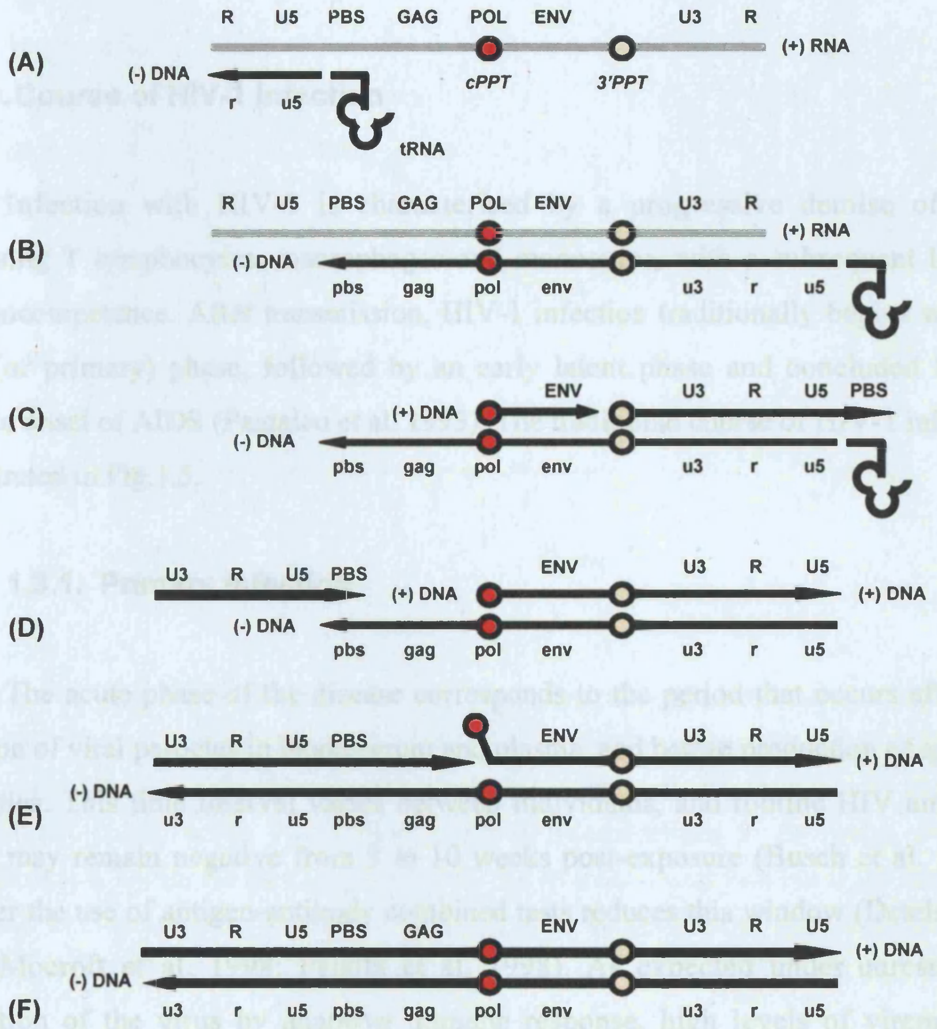


Fig. 1.4. Mechanism of HIV-1 reverse transcription. (A) Negative strand DNA synthesis, initiated by the binding of a tRNA^{lys} to the primer binding site (PBS) of the viral RNA (B) First strand transfer reaction, followed by negative strand elongation (simultaneously with the degradation of the RNA template); (C) Positive strand DNA synthesis, initiated at the center and 3' polypurine tract primers (*cPPT* and *3'PPT* respectively); (D) Second strand transfer and positive strand elongation; (E) Termination of the positive strand elongation downstream of the *cPPT*; (F) Double stranded DNA, termed provirus. *After Rausch et Le Grice, 2004.*

Viral proteins are next transported to the inner surface of the plasma membrane, where they accumulate and condense to form an immature virion (Fig. 1.3G). As the particle extrudes from the cell, it acquires a lipid coat expressing mature TM and SU envelope glycoproteins, causing cellular death (Fig. 1.3H). Proteolytic processing of the Gag and Pol proteins by encapsidated proteases concludes the maturation of the virion soon after it is released.

1.3. Course of HIV-1 Infection

Infection with HIV-1 is characterised by a progressive demise of CD4 expressing T lymphocytes, macrophages and monocytes, with a subsequent loss of immunocompetence. After transmission, HIV-1 infection traditionally begins with an acute (or primary) phase, followed by an early latent phase and concluded by the ultimate onset of AIDS (Pantaleo et al. 1993). The traditional course of HIV-1 infection is illustrated in Fig.1.5.

1.3.1. Primary Infection

The acute phase of the disease corresponds to the period that occurs after the detection of viral particles in blood serum and plasma, and before production of specific antibodies. This time interval varies between individuals, and routine HIV antibody testing may remain negative from 3 to 10 weeks post-exposure (Busch et al. 1995), however the use of antigen-antibody combined tests reduces this window (Detels et al. 1998; Mocroft et al. 1998; Palella et al. 1998). As expected under unrestrained replication of the virus by adaptive immune response, high levels of viremia are observed during this phase, reaching levels of up to 100 million copies of HIV-1 RNA/ml (Daar et al. 1991; Piatak et al. 1993). Concurrently, the pool of CD4+ T lymphocytes starts to decline (Pedersen et al. 1990; Gupta. 1993), until the CD4+ T cell counts drops below the level under which opportunistic infections can develop (Vento et al. 1993). CD4 count traditionally rebounds after primary infection (Fig.1.5A), yet the loss of HIV-specific CD4 T-cell response experienced during that phase of the disease is never fully recovered, even when treatment is administered (Autran et al. 1997; Pitcher et al. 1999).

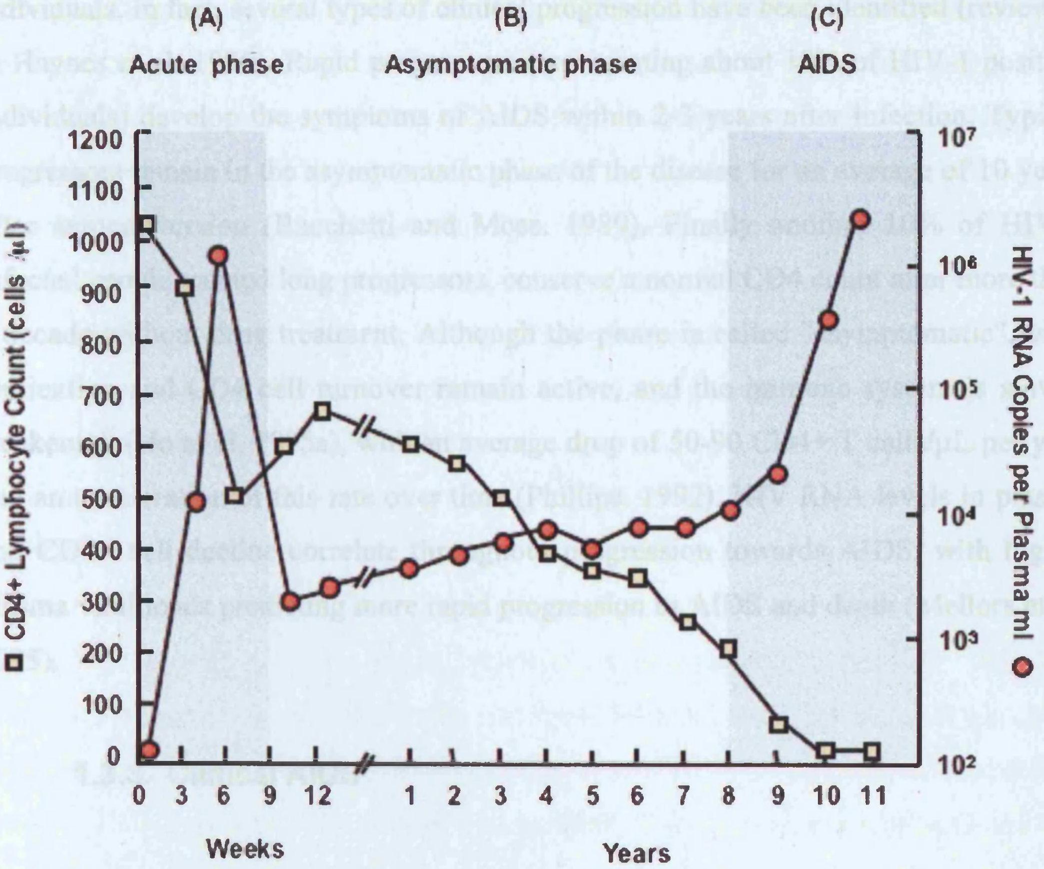


Fig. 1.5. Progression of HIV-1 infection. (A) During primary infection, HIV-1 actively replicates and disseminates in the host's body, causing an abrupt decrease in CD4+ T cells. (B) As specific immune response to HIV is initiated, an asymptomatic phase follows, during which the number of CD4+ cells continue to decrease, while detectable viremia remains constant. (C) By the time CD4 count falls below the critical level of 200 copies/ml, the host's immune system collapses and AIDS is declared. *Adapted from Pantaleo et al., 1993.*

1.3.2. Chronic Asymptomatic Infection

Following seroconversion, HIV-1 infection progresses into an asymptomatic phase during which a relative equilibrium between viral replication and the host immune response is reached. During that stage, specific anti-HIV-1 antibodies are produced, while detectable HIV-1 antigens dramatically decrease in blood serum and plasma (Fig.1.5B). There is usually no manifestation of clinical symptoms, and the 'silent' progression of the disease may cover months to years according to the individuals. In fact, several types of clinical progression have been identified (reviewed in Haynes et al. 1996). Rapid progressors (representing about 10% of HIV-1 positive individuals) develop the symptoms of AIDS within 2-3 years after infection. Typical progressors remain in the asymptomatic phase of the disease for an average of 10 years after seroconversion (Bacchetti and Moss. 1989). Finally another 10% of HIV-1 infected people, named long progressors, conserve a normal CD4 count after more than a decade without drug treatment. Although the phase is called "asymptomatic", viral replication and CD4 cell turnover remain active, and the immune system is slowly weakening (Ho et al. 1995a), with an average drop of 50-90 CD4+ T cells/µL per year and an acceleration of this rate over time (Phillips. 1992). HIV RNA levels in plasma and CD4+ cell decline correlate throughout progression towards AIDS, with higher plasma viral loads predicting more rapid progression to AIDS and death (Mellors et al. 1995).

1.3.3. Clinical AIDS

The onset of AIDS is conventionally defined by measurement of CD4 levels below 200 cells/µL (Fig.1.5C). At that point, the collapse of the host's immune system allows for opportunistic infection to declare, eventually leading to death. In the absence of antiretroviral therapy, survival times after diagnosis of AIDS is on average 10-12 months (Gail et al. 1997).

2. Mechanisms of HIV-1 Variation

The first evidence of retroviral variation was published as early as in 1913, when Rous and Murphy observed dissimilar tumours as a result of infection by chicken sarcoma virus no. 1 (Rous and Murphy. 1913). Since then, viral genetic plasticity has been extensively investigated, mainly through molecular approaches, and it is generally accepted that the diversity exhibited by a viral population is a reflection of the virus' natural history.

In the context of AIDS, the evidence that genomic heterogeneity exists among different isolates rapidly followed the characterisation of the disease's causative pathogen (Benn et al. 1985; Hahn et al. 1985; Wong-Staal et al. 1985). With an average evolution rate of nearly 2.4 x 10^{-3} substitutions per base pair per year (Korber et al. 2000), HIV-1 genome diversifies a million times faster than mammalian genes (Kumar and Subramanian. 2002). This feature makes HIV one of the fastest evolving organisms known to date. Such an outstanding genetic plasticity results from a complex combination of conflicting evolutionary forces expressed via molecular adaptation and random genetic drift, intimately linked to both the host and virus biology.

2.1. Causes of HIV-1 Evolution

2.1.1. Mutations

RNA viruses share high mutation rates, ranging from 10^{-3} to 10^{-5} misincorporations per nucleotide site per round of replication (Holland et al. 1982; Drake et al. 1998). Hence, the introduction of insertion, deletion or base mismatches into the HIV-1 genome is a crucial determinant of the virus' variability (Katz and Skalka. 1990). The HIV-1 reverse transcriptase has two distinct enzymatic activities: a RNA- or DNA-dependent DNA polymerase and a ribonuclease (RNase) H. The DNA polymerase activity of the molecule is responsible for the transcription of viral RNA while the RNase H activity involves the degradation of the RNA strand from RNA-DNA hybrids formed during reverse transcription. The enzyme functions as both an endonuclease and exonuclease in hydrolyzing its target (Schatz et al. 1990).

The HIV-1 reverse transcriptase displays a relatively poor processivity in *in vitro* studies (Bebenek et al. 1989; Bebenek et al. 1998), and, in the absence of 3' to 5' exonuclease proofreading activity, it exhibits a remarkably high error rate (Roberts et al. 1988). Because of its inability to excise mispaired nucleotides, the molecule has a 100-fold lower fidelity than the cellular DNA polymerases, which possess the proofreading 3'-exonuclease activity. Although other RTs also lack this proofreading function, HIV-1 RT is even 10- to 100-fold more error-prone (Drake et al. 1998; Drake. 1999).

When estimated, HIV-1 mutation rates vary according to the experimental system used. On average, reported mutation rates are high with *in vitro* purified HIV-1 RT, ranging from 3 x 10^{-4} to 6 x 10^{-4} substitutions per site per round of replication (Preston et al. 1988; Roberts et al. 1988; Boyer et al. 1992; Hubner et al. 1992), while they are 10- to 20- fold lower if estimated in single-round infection systems (Pathak and

Temin. 1990; Mansky and Temin. 1995; Kim et al. 1996). More recently, Gao et al. (2004) found an overall HIV-1 mutation rate of 5.4 x 10^{-5} substitutions per base per replication cycle, using single-round infection systems on near-full-length HIV-1 genomes.

Interestingly, the incorporation of mutations is far from being a random process. Early sequence analysis showed that diversity is not evenly distributed throughout the genome, with the greatest genetic heterogeneity found in the envelope gene (Hahn et al. 1985; Starcich et al. 1986). Furthermore, pairwise comparisons of intra-subtype HIV-1 protein sequences showed significant variation across the genome, with a median percentage of amino acid differences of 17% (range 4–30%) in Env, 15% (3–30%) in Tat, and 8% (2-15%) in Gag (Korber et al. 2001). Substitution bias can also be qualitative, as illustrated by the extreme tendency toward G to A mutations observed in retroviral genomes, a phenomenon known as hypermutation (Vartanian et al. 1991). It is only recently that the deamination of cytosine residues in nascent retroviral cDNA by the host cell apolipoprotein B editing complex protein (APOBEC) 3G has been identified as at the origin of the phenomenon (reviewed in Vartanian et al. 2003). In some hypermutated segments of the genome, up to 60% of guanine residues can be substituted (Vartanian et al. 2002), heavily contributing to the virus genetic diversity. Also, the effect of HIV-1's RT prohibitive error rate is aggravated by its fast replicative dynamic: with a generation time approximating 2.6 days *in vivo* (Perelson et al. 1996a), the production of viral particles in an untreated HIV-1 positive individual exceeds 10^9 copies per day (Ho et al. 1995b; Wei et al. 1995; Perelson et al. 1996b). Nonetheless, the HIV-1 RT alone is far from being a unique source of mutation. Retroviral replication is also mediated by cellular DNA polymerases and RNA polymerases II, each of which may contribute to the incorporation of mutations at later stages. Although the contribution of the high-fidelity DNA polymerases is unlikely to have a substantial impact in the mutation rate of HIV-1 genome, RNA polymerase II mediated replication may be a significant error-prone step (reviewed in Preston and Dougherty. 1996).

2.1.2. Recombination

Recombination is a process of genetic exchange, through which a hybrid, or mosaic, nucleic acid sequence is generated from two or more genetically distinct parental genomes. Recombination is frequent in retroviruses and accounts for at least

10% of the circulating strains of HIV-1 (Sharp et al. 1996; Mccutchan. 2000; Peeters and Sharp. 2000). Recombination requires the co- or super-infection of a single host cell by more than one viral particle, during which one copy of each parental genome is encapsidated into a heterozygous virion (Hu and Temin. 1990). When infection by the chimera virus occurs, a recombinant genome is generated by RT switches from a parental genome to another (Goodrich and Duesberg. 1990). The mechanistic origin of recombination lies in frequent interruptions of the reverse transcriptase during polymerisation. This phenomenon, called "pausing," leads occasionally to the dissociation of the enzyme from the primer-template complex (Bebenek et al. 1993; Wu et al. 1995) and is sequence specific (Harrison et al. 1998). The ability of RT/primer/template complexes to bind an additional single-stranded RNA molecule also promotes recombination-prone template switching (Peliska and Benkovic. 1992).

To date, several models have been proposed to explain the mechanism of recombination during reverse transcription (reviewed in Negroni and Buc. 2001). Amongst these, the 'forced copy-choice' model proposes that template switching occurs during the synthesis of the negative DNA strand, and is driven by breaks in the viral RNA, forcing the reverse transcriptase to jump from one RNA copy to another (Vogt. 1971; Coffin. 1979). Alternatively, the 'strand displacement assimilation' model suggests that recombination occurs during the synthesis of the positive DNA strand, when an internal initiated fragment is displaced by upstream growing fragments and hybridises to parallel DNA synthesis complexes (Boone and Skalka. 1981; Junghans et al. 1982).

Recombination is a ubiquitous phenomenon in retroviral biology, and holds a major responsibility in HIV-1 molecular variability (Robertson et al. 1995a; Robertson et al. 1995b). Thus, the reverse transcriptase is known to be highly recombination prone, with an estimated 3 recombination events occurring per genome per round of replication (Jetzt et al. 2000; Zhuang et al. 2002). This rate, higher than the actual mutation rate of the virus (Jung et al. 2002), is thought to be the highest of all organisms. In practice, however, recombination can be difficult to detect, especially amongst closely related strains or genomes. The recent re-analysis of empirical datasets suggests that recombination is more common in HIV-1 genomes than presently thought (Posada. 2002). If true, underestimating the frequency of recombination in HIV-1 is likely to have repercussions on the reliability of previous inferences about the virus' evolutionary history and dynamics. This finding might also have serious implications on vaccine development (Gaschen et al. 2002).

2.1.3. Selection

Nucleotide misincorporation and genetic recombination alone are not sufficient to explain HIV-1 molecular evolution (see for instance Gao et al. 2004). Indeed, the evolution of the virus is a composite phenomenon also encompassing the rate at which genetic changes get fixed within the population. By fixation, one shall understand the process through which the frequency of a genetic polymorphism increases up to 100% in a given population. According to the theory of natural selection (Darwin. 1959), if a particular genetic change increases the chances of an organism to survive in a given environment, it will be subject to positive selective pressure. If, in contrary, a genetic change decreases the changes of survival of the organism, it will be subject to negative selective pressure and be eliminated. Under this balancing selection, advantageous mutations get eventually fixed in a population while deleterious ones are eliminated. A distinction has thus to be made between rates of mutation and rates of substitution. As seen earlier, the mutation rate of an organism is traditionally expressed as the number of substitutions per nucleotide site per round of replication. By contrast, the rate of substitution of an organism corresponds to the rate at which newly acquired nucleotide substitutions become fixed and spread within a population. Rates of substitution are expressed as number of nucleotide substitution per site per unit of time (i.e. day, year or generation). The substitution rate of a virus (or a gene) is driven by diverse evolutionary forces such as selective adaptation or random genetic drift, and reflects the relative proportion of advantageous, neutral or prejudicial mutational forces exerted on the genome. When looking at HIV-1 evolutionary dynamics, one needs to distinguish between intra- and inter-host environments.

Intra-host evolution of HIV-1

Soon after infection, the transmitted HIV-1 particles and subsequent viral populations are subject to strong, non-random pressures within the host, exerted by dynamics such as induced immune response or antiretroviral therapy. Under such adverse environmental conditions, adaptation is a key determinant of HIV-1 evolution within the host, promoting the selection and fixation of mutations. The relative frequency, strength and exact location of positively selected substitutions have been extensively investigated in specific regions of HIV-1 genome (Seibert et al. 1995; Wolinsky et al. 1996; Yamaguchi-Kabata and Gojobori. 2000; Yang et al. 2000; De Oliveira et al. 2004; Leal et al. 2004), and a large body of evidence suggests that Darwinian evolution is both fast and widespread in HIV populations within infected individuals. This is particularly obvious in the context of immune escape (Zanotto et al. 1999; Ross and Rodrigo. 2002) and antiretroviral drug resistance (Richman et al. 1994; Frost et al. 2001). Thus, fixation of adaptive changes within the envelope gene, where most of the amino-acid substitutions confer a strong selective advantage in evading immune recognition, occur every 2.5 months on average, constituting what is though to be the fastest adaptation rate ever recorded for a single protein-coding gene (Williamson. 2003). In the case of the *pol* gene, the emergence of resistance-conferring polymorphisms has been reported soon after administration of all available antiretroviral drugs (Pillay et al. 2000). Despite the decreased fitness these mutations are associated with in the absence of therapy, most of them become fixed in the viral population for the adaptive benefit they confer, the occurrence of compensatory changes counterbalancing the loss of replicative fitness of the resistant mutants (reviewed in Quinones-Mateu and Arts. 2002).

Despite the strong influence of positive selection on HIV-1 evolution, stochastic fluctuations also drive allele fixation in HIV populations, creating a genetic drift (Leigh-Brown and Richman. 1997; Holmes and Zanotto. 1998; Frost et al. 2000; Shriner et al. 2004). Mutation frequency can vary simply by chance, as the result of a random sampling process and regardless of the relative advantage (or disadvantage) mutations confer. For instance, in a population bottleneck, where the population suddenly contracts to a small size and then grows again to a large population, genetic drift can result in sudden and dramatic changes in allele frequencies (Kitrinos et al. 2005). Similarly, migration across the host's compartments may produce founder effects, where rare mutations in the originating population get fixed in the next generation of virus (Poss et al. 1998). Hence changes in cellular conditions, compartmentalization, or migration dynamics are amongst the factors exerting a constant purifying force on viral population, resulting in survival of lineages on the basis of pure chance rather than fitness (Leigh Brown. 1997; Rouzine and Coffin. 1999; Ribeiro and Bonhoeffer. 2000).

The idea of a genetic drift implies that all mutations are treated neutrally in respect to fitness, and is directly influenced by the effective size of the viral population.

By effective population size, one should understand the proportion of the total population size that successfully contributes to the next generation of virus (see Chapter II, section 4.2). Indeed, deterministic evolution has a greater impact on large populations, where the abundance of viral variants is subject to fitness competition, while in a small population, the fate of mutants is more sensitive to the influence of random events, independent of their fitness.

Whether intra-host effective population size of HIV-1 is large or small is still open to debate. The relative influence of deterministic or stochastic models of molecular evolution has yet to be unambiguously determined. On one hand, authors argue in favor of a large effective population size within a host, reaching an order of magnitude of 10^5 (Coffin. 1995). On the other hand, observations seem more consistent with a smaller effective population size, involving an average of 1500 reproductive particles (Leigh Brown. 1997; Leigh-Brown and Richman. 1997). However, the two models are not necessary mutually exclusive and may reflect population patterns under different conditions of HIV-1 natural history. If HIV-1 populations may behave in a mostly deterministic manner under intra-host constraints, the loss in population size resulting from transmission may increase the impact of stochastic forces. To date, whether HIV-1's intra-host population size is large or small is still contentious.

Inter-host evolution of HIV-1

If HIV-1 intra-host evolution is clearly shaped by the successive gain and loss of advantageous and disadvantageous mutations, the virus' evolution between hosts shows little evidence of being driven by positive selection. Indeed, host-to-host transmission of HIV-1 is traditionally accompanied by a loss of genetic diversity, through what has been termed a bottleneck effect (Cichutek et al. 1992; Mcnearney et al. 1992; Wolfs et al. 1992). Effectively, only a minor subset of the donor's viral population will successfully be passed on to the recipient, and the new host population will rise from the limited genetic pool transmitted in that way (Wolinsky et al. 1992). The strong purifying selection exerted by transmission on HIV-1 intra-host populations is particularly obvious in the highly variable V3 loop of the envelope gene, where a significant loss of allelic diversity is observed in primarily infected individuals compared to chronically infected ones (Zhu et al. 1993). Thus, the characterization of homogenous populations within HIV-1 primarily infected individuals is a accurate indicator of recent transmission bottlenecks (Delwart et al. 2002).

An additional feature of HIV-1 evolution during transmission is the decrease of population fitness following bottleneck events. After infection, the reduced diversity of the transmitted sub-population allows stochastic forces to fix deleterious mutations via genetic drift (Muller. 1934). This phenomenon, known as Muller's Ratchet, is well characterised in RNA viruses (Duarte et al. 1992: Chao. 1997) and predicts that, when mutation rates are high and a significant proportion of mutations are deleterious, an irreversible 'ratchet' mechanism will gradually decrease the fitness of a small asexual population. The loss of the fittest genotypes will be irreparable unless some other process recreates individuals of comparable fitness, one such process being recombination (Muller. 1964; Felsenstein. 1974). Experimental evidence shows that genetic reassortments such as recombination can reduce the mutational load in a population and promote the clearance of accumulated deleterious effects (Chao et al. 1997).

Although transmission bottlenecks are profoundly liable to HIV-1 evolution at the inter-host level, alternative stochastic determinants are at work. Amongst these, host behavioral factors such as difference in transmission dynamics may generate a strong genetic drift influencing allele fixation. For instance, an advantageous mutation arising within an individual with low risk behavior or partner exchange rate may fail to be successfully transmitted and selected for, accentuating the purifying effect induced by transmission on HIV populations.

2.2. Consequences of HIV-1 Evolution

2.2.1. HIV-1 Subtypes

A striking outcome of the HIV-1 fast rate of evolution is the extensive genetic diversification the virus went through in a few decades, enforcing the need for a nomenclature system, and leading to the classification of lineages on the basis of genetic distances and phylogenetic clustering (Robertson et al. 2000).

So far, HIV-1 encompasses three distinctive genetic groups, termed M (Main), O (Outgroup) and N (new or non-M, non-O), suspected to result from three independent cross-species introductions (Sharp et al. 2001). While groups O and N represent a small fraction of the HIV-1 strains identified worldwide and remain restricted to West-Central Africa (De Leys et al. 1990; Gurtler et al. 1996; Simon et al. 1998). The ubiquitous group M is responsible for a vast majority of the HIV-1 epidemic worldwide. Since considerable diversity was accumulated within the group M itself (Louwagie et al. 1993), the latter was further divided into nine clades or subtypes, designated A, B, C, D, F, G, H, J and K. These clades are approximately equidistant in phylogenetic terms, and can differ from each other from up to 30% in the *env* region (Korber et al. 2000). In the light of the overall high number of HIV-1 subtypes cocirculating in the Democratic Republic of Congo (formerly Zaire), together with the high intrasubtype found in this area, it has been suggested that this region is the epicentre of HIV-1 group M (Vidal et al. 2000). Alternatively, although the vast majority of group O infections are restricted to Cameroon, recent studies have identified diversity and genetic substructure within the group (Roques et al. 2002; Yamaguchi et al. 2002). By contrast, only six infections by group N viruses have been reported to date (Simon et al. 1998; Ayouba et al. 2000; Roques et al. 2004). Besides, inter-subtype mosaic genomes were identified in regions where distinct subtypes co-circulate. These are now designated as 'circulating recombinant forms', or CRFs (Carr et al. 1998). To date, up to 15 CRFs have been recognised (Los Alamos HIV sequence database, *http://www.hiv.lanl.gov/*).

HIV-1 subtypes are unequally distributed across risk groups and geographic areas (see Table 1.2). For example, while most subtypes, including CRFs, circulate in Africa, subtype B is mainly predominant in North America and Western Europe, where the HIV-1 epidemic is dominated by homosexual and injecting drug use (IDU) transmission (Tatt et al. 2001). Inversely, subtype C is highly prevalent in populations where heterosexual contact is the main source of infection and represents about 50% of the circulating strains. Consequently, in increasing spread of non-B strains has recently been reported within the heterosexual population in Western Europe and North America, where the epidemic is dominated by subtype B (UNAIDS, *www.unaids.org*). On a global scale, the most prevalent HIV-1 clades are subtypes C (47%), A (27.2%), B (12.3%), D (5.3%) and CRF01_AE (3,2%) (Osmanov et al. 2002).

Whether subtypes have intrinsic biological differences is subject to debate, and several studies have emphasized the difficulty in discriminating between viral, host and societal factors (reviewed in Hu et al. 1999). Yet there is a large body of evidence suggesting that discrepancies exist across HIV-1 groups and subtypes with respect to transmissibility and pathogenesis (Blackard et al. 2001), disease progression (Wolinsky et al. 1992; Soto-Ramirez et al. 1996; Van Harmelen et al. 1997), co-receptor usage and cell tropism (Zhang et al. 1996; Tscherning et al. 1998; Peeters et al. 1999) or antiretroviral drugs susceptibility (Apetrei et al. 1998; Descamps et al. 1998; Wainberg. 2004). Moreover, different substitution patterns of positively selected sites were reported across subtypes, with a correlation within clades despite differences in the strength of selection (Gaschen et al. 2002; Choisy et al. 2004). For all that, a vast majority of the efforts made in developing diagnostic tests, antiretroviral drugs, and HIV-1 vaccines were on the basis of subtype B viruses.

2.2.2. Drug Resistance

The first effective drug against HIV was azidovudine (AZT), a nucleoside analogue inhibiting the virus' reverse transcriptase (Yarchoan et al. 1986). Soon after, came the introduction of highly active antiretroviral therapy (HAART), based on 'cocktails' of nucleoside and non-nucleoside reverse transcriptase inhibitors (NRTIs and nNRTIs), protease inhibitors (PIs), and more recently inhibitors of viral entry into susceptible cells. As chain terminators, NRTIs compete with natural desoxynucleoside triphosphates (dNTPs) for incorporation into the synthesised DNA chain, while NNRTIs and PIs inhibit replication by binding to the active site of reverse transcriptase and protease respectively. Fusion inhibitors represent the most recent generation of anti-retroviral drugs and block the fusion of the viral envelope to the cell membrane. HAART has significantly reduced HIV transmission, morbidity and mortality (Palella et al. 1998), but has also led to the problem of drug resistance (Larder and Kemp. 1989; Najera et al. 1994; Ribeiro and Bonhoeffer. 2000).

Resistance to PIs and nNRTIs originates from conformational changes reducing the affinity between the inhibitor and the binding site of the mutated molecule. In the case of PIs, mutations in protease cleavage sites have also been reported to be responsible for resistance (Cote et al. 2001). Alternatively, resistance to NRTIs either come from mutations preventing the addition of nucleotide analogues to the synthesised DNA chain (Huang et al. 1998; Sarafianos et al. 1999), or mutations in RT increasing the cleanse of drug from the DNA (Arion et al. 1998). To date, 27 nucleotide positions in the protease gene are known to harbour mutations associated with PI resistance, and almost that many are involved in drug resistance in the RT gene (Shafer et al. 2000; Johnson et al. 2003). It is understood that both selective pressures exerted by the drug regimens and underlying genetic drift play an active role in the emergence of drug resistance mutations (Frost et al. 2000; Ibanez et al. 2000; Frost et al. 2001; Leal et al. 2004). Nevertheless, the respective weight of deterministic and stochastic forces in the process has still to be clarified. While authors are in favour of a largely deterministic acquisition of advantageous alleles under a drug-rich environment (Rouzine and Coffin. 1999), others support a neutral model (Brown and Richman. 1997). The two hypotheses are far from being mutually exclusive, and epidemiological evidence suggests that drug resistance acquired during adaptation to sub-optimal HAART is randomly transmitted between individuals (Gomez-Cano et al. 1998; Pillay et al. 2000; Blower et al. 2001).

It has been shown that many mutations conferring drug resistance to reverse transcriptase and protease inhibitors have deleterious effects on the replicative capacity of the virus, so that resistant mutants have a decreased fitness compared to wild-type viruses in a drug-free environment (Harrigan et al. 1998; Zennou et al.1998; Martinez-Picado et al. 1999). However, the benefit drug resistance-associated mutations confer during therapy is such that these mutants are eventually selected for and fixed within the viral population, the presence of compensatory mutations frequently sustaining the fitness loss of the mutants (Nijhuis et al. 1998; Menendez-Arias et al. 2003). As a result, deleterious allelic changes come fixed by the action of genetic drift while they are expected to recede by purifying selection. This is illustrated, for instance, by the *in vivo* fitness of HIV-1 subpopulations harboring mutations at codons 41 and 215 of reverse transcriptase (related to zidovudine resistance). While these clones are highly fit in the presence of drug, reduced fitness involving their RT function was reported for the mutants in an environment requiring competition with zidovudine-sensitive strains (Goudsmit et al. 1997; Yerly et al. 1998).

3. Epidemiology of HIV-1

HIV-1 has been proven successful in exploiting various means of transmission adapted to key aspect of modern life, and is transmitted through three principal routes: unprotected sexual contact (Gottlieb et al. 1981; Royce et al. 1997), contact with infected blood or blood products (Des Jarlais et al. 1992; Lackritz et al. 1995; Schreiber et al. 1996) and prenatal transmission from mother to child (Rogers et al. 1987; Gabiano et al. 1992). Twenty years after its identification, almost 40 million people (range 35.9 –

44.3 million) live with HIV worldwide. In 2004, 4.9 million (4.3 million–6.4 million) individuals were newly infected, annunciating the biggest increase in new infections ever recorded since the beginning of the epidemic. The global AIDS epidemic killed a total of 20 million people, 3.1 million (2.8 million–3.5 million) of whom died in the past year (UNAIDS, *http://www.unaids.org/*). A key feature of the epidemic is its remarkable heterogeneity within regions, countries, or niches. For instance, an estimated 25 million people are living with HIV in sub-Saharan Africa, embodying almost two-thirds of all people living with HIV (UNAIDS, *http://www.unaids.org/*). There is nonetheless a vast epidemiological diversity of HIV across the African countries, with prevalences ranging from 2 to 20% of the adult population. By contrast, 580 000 people are living with HIV in Western Europe, including almost 50,000 infections diagnosed in the United Kingdom (Health Protection Agency's monthly report 2004, *http://www.hpa.org.uk/*).

3.1. HIV-1 in the United Kingdom

Epidemiological studies coupled with phylogenetics has shed light on the introduction of HIV-1 into the Western world (Kuiken et al. 2000). In essence, it was hypothesised that a HIV subtype B strain was carried out of Africa and introduced into the North American homosexual community by a gay airline steward, 'patient 0', infected in the late seventies (Auerbach et al. 1984). It is however doubtful that one individual alone is at the origin of the initial spread of HIV in the United States (US) of America, and it has been suggested that 'patient 0' belonged to a group of homosexual men involved in frequent sex tourism who died of AIDS between 1980 and 1982 (Hooper. 2000). If the identification of the US epidemic founder effect(s) remains subject to conjecture, it is most certain that the virus arrived through gay sex tourism from Haiti, where HIV prevalence was advanced at the time the first case of AIDS was identified in the US (Greco. 1983; Johnson and Pape. 1989). From the United States, the virus presumably spread from one European country to one other, including the United Kingdom (UK). AIDS was first characterised in the Western world amongst men having sex with men, suggesting that the epidemic expanded from this risk group (Cheingsong-Popov et al. 1984; Van Haastrecht et al. 1992). The disease, however, affected other population groups, and injecting drug users are thought to have played a role in the spread of the virus in the early 80's (Masur et al. 1981).

Table.1.2. Global distribution and predominance of HIV-1 subtypes and circulating recombiant forms (CRFs)

Group	Subtype	Distribution	References
M	A	Eastern & Central Africa	*Nkengasong et al. 1995; Carr et al. 1999; Hu et al. 2000*
	B	Northern & South Amercia, Western Europe, Oceania	*Brodine et al. 1995; Ramos et al. 1999; Kuiken et al. 2000; Oelrichs et al., 2000a*
	C	Southern & Estern Africa, India	*Lole et al, 1999; van Harmelen et al. 1999; Hussein et al. 2000*
	D	Central Africa	*Hu et al.2000; Hierholzer et al. 2002; Vidal et al.2003*
	F	Central Africa, Southern America	*Triques et al. 1999; Laukkanen et al. 2000*
	G	Central Africa	*Potts et al. 1993; Janssens et al. 1994; Delaporte et al. 1996*
	H	Central Africa	*Murphy et al. 1993; Janssens et al. 1994; Nkengasong et al.1994*
	J	Central Africa	*Bikandou et al. 2000; Cham et al. 2000; Vidal et al. 2000*
	K	Western Africa	*Triques et al. 2000; Vidal et al. 2000*
	CRF01_AE	Western & Central Africa, South-East Asia	*Montavon et al. 2000; McCutchan et al. 1999; Kato et al. 2001*
	CRF02_AG	Western & Central Africa	*Carr et al. 1998; Montavon et al. 2000; Carr et al. 2001a*
	CRF03_AB	Eastern Europe	*Bobkov et al. 1998; Liitsola, 1998*
	CRF04_cpx	Cyprus & Greece	*Gao, 1998; Nasioulas, 1999*
	CRF05_DF	Democratic Republic of Congo	*Laukkanen, 2000*
	CRF06_cpx	Western Africa	*Montavon, 1999; Baldrich-Rubio et al. 2001*
	CRF07_BC	China	*McCutchan at al. 2002; Yang et al. 2002*
	CRF08_BC	China	*McCutchan at al. 2002; Yang et al. 2002*
	CRF09_cpx	Senegal	*Burda, 2004; McCutchan et al. 2004*
	CRF10_CD	Tanzania	*Kulinska et al. 2001*
	CRF11_cpx	Central Africa	*Montavonet al. 2002*
	CRF12_BF	Southern America	*Carr et al., 2001b; Thomson et al. 2002;*
	CRF13_cpx	Cameroon	*Wilbe et al. 2002*
	CRF14_BG	Spain, Portugal	*Thomson et al. 2001; Delgado et al. 2002*
	CRF15_01B	Thailand	*Tovanabutra et al. 2003*
O		Cameroon, Gabon, France	*Nkengasong et al. 1993; Zekeng et al. 1994; Gurtler et al. 1994*
N		Cameroon	*Simon, 1998*

The first case of AIDS in the UK was documented in December 1981 (Dubois et al. 1981). Twenty years later, 49,500 adults lived with HIV in the United Kingdom, including 5,542 new infections acquired that year (almost double the number identified in 1997) (Health Protection Agency's monthly report 2004, *http://www.hpa.org.uk/*). Although men having sex with men (MSM) remain the acquisition group at greatest risk in Britain with 29,362 prevalent infections (53%) in 2003, an overwhelming increase in heterosexually acquired infections has been reported since 1999 (49% of new HIV diagnoses in 2003 were in people infected during heterosexual intercourse), probably acquired outside the UK (see Fig. 1.6) (UNAIDS, *http://www.unaids.org/*). Indeed, three quarters of the heterosexual infections diagnosed in recent years have been in people originating from high prevalence countries, such as sub-Saharan Africa (Hamers and Downs. 2004). In contrast, the prevalence of HIV infection amongst injecting drug users (IDUs) remains low (Aceijas et al. 2004). If such a rise in new HIV diagnoses seems partly attributable to a significant increase in HIV testing in Britain, a large share of HIV infections still remain undiagnosed. Indeed, despite considerable efforts in surveillance and monitoring of the disease, an estimated 11000 infections (one third of people living with HIV in the UK) are believed to remain undiagnosed, and are likely to discover their condition only once afflicted by AIDS-related illnesses (Department of Health United Kingdom, *http://www.publications.doh.gov.uk/*). HIV-1 predominantly infects young adults in the UK, with 79% aged 15 to 39 years at diagnosis. Through the last 15 years, the main focus of infection in Britain has been in the London region, with nearly two thirds of the infected individuals residing in this area. With the decline in deaths observed since the administration of highly active antiretroviral therapy (HAART) and the ongoing rise of newly diagnosed infections, the prevalence of HIV-1 in Britain is increasing. HIV infection is now the fastest-growing health hazard in England (Department of Health United Kingdom, http:*//www.publications.doh.gov.uk/).*

3.2. Molecular Epidemiology of HIV-1

Although epidemiology has been focussing on the occurrence, origin and spread of epidemics long before the causal agents of diseases were identified, the recent emergence of molecular techniques has given the discipline second wind. Since the first applications of molecular epidemiology to infectious diseases in general (Kilbourne. 1973), and HIV in particular (Smith et al. 1988), the on-going availability of sequence data has allowed in-depth analyses of epidemics' features previously out of reach for traditional techniques. In regard to HIV-1, the comparative analysis of gene sequence variation has become a standard practice since the early days of the epidemic, exploiting bioinformatics methodologies of increasing sophistication. Thus, molecular phylogenetics (see Chapter II, section 3) permitted the investigation of key determinants of the epidemic as diverse as the origin of emerging populations (Korber et al. 2000a; Salemi et al. 2001; Sharp et al. 2001), the sources and transmission patterns of localized outbreaks (Ou et al. 1992; Albert et al. 1994; Hayman et al. 2001), or the inference of the demographic histories of HIV-1 lineages (Grassly et al. 1999; Pybus et al. 1999; Kurbanov et al. 2003; Lemey et al. 2003; Robbins et al. 2003).

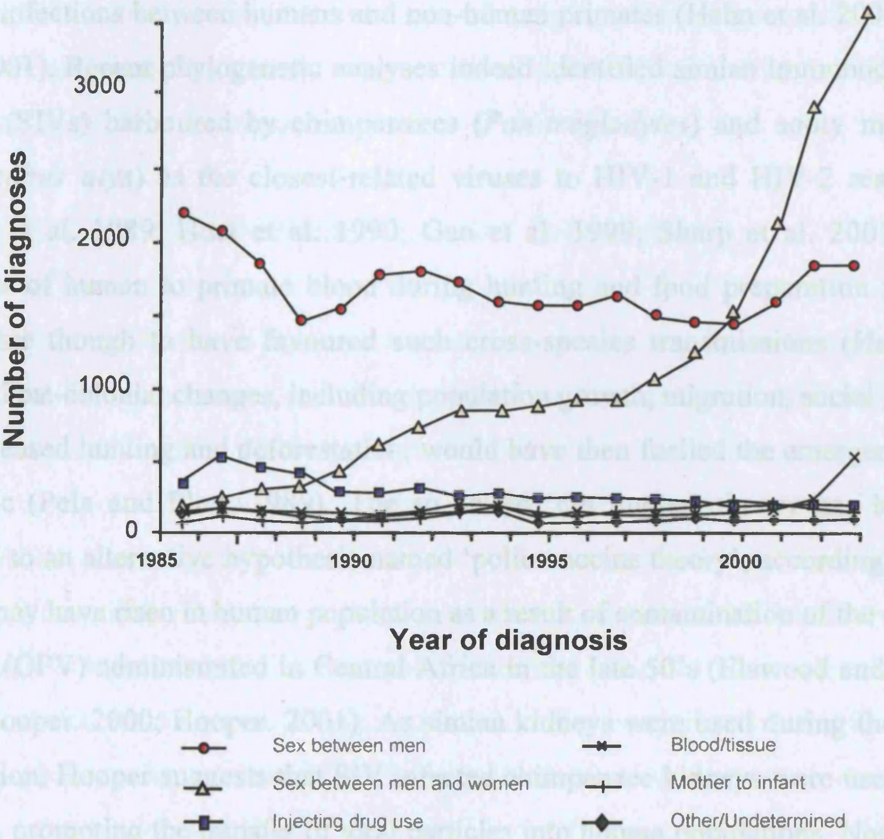


Fig. 1.6. Number of new HIV-1 diagnoses by year of diagnosis in the UK, and they probable route of infection. From the *HPA Annual Report, January 2004.*

Overall, molecular epidemiology has proven a powerful tool for the design and evaluation of preventative programs for public health monitoring, and is increasingly used by clinicians, molecular biologists, phylogeneticists and epidemiologists with an interest in HIV-1 research. Two examples of direct relevance for the work presented in this thesis will be developed here, namely the investigation of the origins of human immunodeficiency viruses and the first use of HIV-1 sequences data for forensic purposes.

3.2.1. Controversial Origin of HIV-1

Molecular studies permitted to unveil the mystery around the origin of both human immunodeficiency viruses and it is now generally accepted that the colonisation of the human population by HIV-1 and -2 resulted from several, independent, cross-species infections between humans and non-human primates (Hahn et al. 2000b; Sharp et al. 2001). Recent phylogenetic analyses indeed identified simian immunodeficiency viruses (SIVs) harboured by chimpanzees (*Pan troglodytes*) and sooty mangabays (*Cercocebus atys*) as the closest-related viruses to HIV-1 and HIV-2 respectively (Hirsch et al. 1989; Huet et al. 1990; Gao et al. 1999; Sharp et al. 2001). Direct exposure of human to primate blood during hunting and food preparation in central Africa are though to have favoured such cross-species transmissions (Hahn et al. 2000a). Post-colonial changes, including population growth, migration, social upheaval, and increased hunting and deforestation, would have then fuelled the emergence of the epidemic (Pela and Platt. 1989). The so called 'cut hunter' theory has long been opposed to an alternative hypothesis named 'polio vaccine theory', according to which HIV-1 may have risen in human population as a result of contamination of the oral polio vaccine (OPV) administrated in Central Africa in the late 50's (Elswood and Stricker. 1994; Hooper. 2000; Hooper. 2001). As simian kidneys were used during the vaccine preparation, Hooper suggests that SIV infected chimpanzee kidneys were used for this purpose, promoting the transfer of viral particles into human populations. Nonetheless, testimony of eyewitnesses, virological data and epidemiological analyses concur to reject this hypothesis as false (Plotkin. 2001). More recently, molecular evidence has finally proven the OPV theory to be erroneous (Worobey et al. 2004). While vaccines were prepared from chimpanzee tissues endemic from the region of Kisangani (Democratic Republic of Congo), Worobey *et al.* showed that the virus circulating within Kisangani Chimpanzees (SIVcpzDRC) is phylogenetically distinct from all strains of HIV-1. Phylogenetic studies further suggest that HIV-1 group M emerged in the human population around 1930, probably in west equatorial Africa, where the epicentre of the pandemic appears to be (Korber et al. 2000; Sharp et al. 2000). The latter findings correlate with the isolation of the earliest HIV-1 sequences known to date from plasma sampled in Leopoldville, Belgian Congo (today's Kinshasa, Democratic Republic of the Congo) in 1959 (Nahmias et al. 1986; Zhu et al. 1998).

Nevertheless, the cut hunter theory fails to find unanimous support amongst the scientific community. Marx *et al.* (2001), for instance, suggested that the probability of five or more zoonotic transitions (i.e. HIV-1 groups M, O, N and HIV-2 subtypes A B) occurring in a brief period is exceedingly small, and that increased unsterile injections in Africa during the period 1950-1970 is likely to have promoted serial human passage of SIV, generating HIV by a series of multiple genetic transitions. The timing of the origin of HIV-1 subtype B was also criticized for not incorporating measures of unequal rates of evolution amongst viral lineages (Salemi et al. 2001; Lukashov and Goudsmit. 2002), fuelling further the controversy on the origin of HIV.

3.2.2. The Florida Dentist Case and Followers

The identification of the source of an outbreak for forensic purposes was amongst the first applications of HIV-1 molecular epidemiology. In the early 90's, Ou *et al.* (1992) identified Dr David Acer, an HIV-infected dentist from Florida, USA, as the source of infection of five of his patients. This conclusion was reached by epidemiologic investigation and phylogenetic comparison of HIV-1 envelope gene sequences amplified from the dentist, his patients and local HIV-infected control individuals. The molecular relatedness of the samples led to the conclusion that five of the patients harboured viruses closely related to the dentist's one, while unrelated viruses infected other patients. David Acer died before any criminal charges were brought, but the so-called Florida dentist case started a vivid controversy (Abele and Debry. 1992; Smith and Waterman. 1992; Debry et al. 1993; Crandall. 1995).

Phylogenetic evidence were effectively produced in a criminal court for the first time in 1998, as a doctor of Louisiana, USA, was accused of infecting his former lover by injecting her with HIV infected blood in an act of vengeance (Metzker et al.1998). DNA samples of the virus in the victim's blood and that of the HIV-positive patient in question were found to be phylogenetically similar. On that ground, the Louisiana doctor was found guilty and sentenced to 50 years.

Since then, molecular epidemiology has been used in several criminal HIV transmission trials, including a rape case (Albert et al. 1994), reckless multiple transmission (Birch et al. 2000) and a sexual assault on children (Machuca et al. 2001). In the UK, phylogenetic evidence was produced in what was the first ever HIV transmission conviction in England and Wales (*http://news.bbc.co.uk/*). Mohammed Dica was found guilty of reckless, rather than deliberate, biological Grievous Bodily Harm against two women and was sentenced to 8 years in prison in November 2003.

Furthermore, phylogenetic analyses have been used to establish relatedness between HIV strains in various non-legal contexts. Reports include an outbreak of infection in a Scottish prison (Yirell et al. 1997; Yirell et al. 1999) as well as in a small heterosexual community (Hayman et al. 2001), the characterisation of transmitted drug resistance (Taylor et al. 2003), father-to-child horizontal transmissions (Ceballos et al. 2004) and investigations by sanitary authorities following nosocomial transmissions (Holmes et al. 1993; Arnold et al. 1995; Blanchard et al. 1998; Goujon et al. 2000; Yerly et al. 2001a).

The Florida dentist case rapidly initiated a debate about the practicalities and difficulties in establishing transmission networks from the analysis of viral gene sequence data, particularly about the choice for the most informative genetic region on one hand and the choice of the best phylogenetic methodology on the other (Holmes et al. 1993). Ideally, full-length sequences should be used for the investigation of potential linkages by phylogenetic means, however practicalities preclude such an approach. The use of *env* gene sequences is often recommended (see for example Leitner *et al.,* 1996), the extensive variation of which has made it attractive for such analyses. However, the exploitation of *env* is far from unproblematic. First, convergent evolution (i.e. identical mutational patterns in unlinked sequences) has repeatedly been observed in the V3 loop of the *env* gene (Holmes et al. 1992; Zhang et al. 1993). More importantly, the rapid genetic diversification of this region is likely to compromise identification of linked sequences in distantly sampled individuals. Indeed, both divergence and diversity of the HIV-1 *env* gene have been shown to increase linearly in early stages of infection (Shankarappa et al. 1999). Alternatively, the *pol* gene is traditionally considered as sub-optimal in terms of phylogenetic signal for its genetic conservation. The *gag* gene would offer a good intermediate, yet its exploitation for molecular epidemiology remains anecdotal and is almost exclusively done in conjunction with *env* (Albert et al. 1994; Birch et al. 2000).

The second issue to consider concerns the method of analysis. If the choice of a phylogenetic method for reconstructing transmissions is less sensitive than the choice of a genetic region (Leitner et al. 1996), the misuse of a model of evolution (see Chapter II section 3.4) can have major consequences on the accuracy of the reconstruction. Indeed, since rates of evolution differ across HIV-1 lineages, populations, or genetic regions, the selection of an optimal model must be a prerequisite when estimating HIV-1 phylogenies. The systematic (and often unjustified) use of over-simplistic models of evolution is unfortunately frequent in HIV molecular analyses. When using such models, features of importance in the context of HIV-1 transmission, such as branch length within the reconstructed genealogy, may be underestimated (Yang et al. 1994). Close attention is therefore required in dealing with HIV-1 sequences for epidemiological, clinical or forensic purposes.

4. The Present Thesis

The work presented in this thesis aims to shed light on the dynamics of the subtype B HIV-1 epidemic in the United Kingdom through the exploitation of routinely available molecular data. The general methods and principles used for that purpose are developed in Chapter II. Chapter III proposes to assess and validate the reliability of the HIV-1 *pol* gene for the characterisation of transmission networks by phylogenetic means. For that purpose, HIV-1 *pol* gene sequences were used to characterise possible transmission chains between patients represented within a nationwide resistance-testing database. The subsequent phylogenies were compared to genealogies obtained with more variable genetic regions of the same HIV-1 samples to confirm relatedness. Chapter IV applies the previous findings to the investigation HIV-1 transmission within a cohort of newly infected men having sex with men from a discrete geographical area, with a focus on the impact of primary infection in HIV-1 transmissibility. Finally, Chapter V focuses on the exploration of the HIV-1 epidemic history amongst gay men in Britain using a coalescent-based approach, estimating of the date of introduction of several lineages of epidemiological significance, the fluctuation of the viral population over time and the growth rate of the epidemic amongst the risk group. A general discussion of the results presented within is proposed in Chapter VI.

CHAPTER II

Methods & Principles

1. Study Population: the ASRU *pol* Database

1.1. Specimen Collection

The *pol* sequences used for this study were generated from plasma samples collected from HIV-1 infected people in the United Kingdom by the Antiviral Susceptibility Reference Unit (ASRU), Health Protection Agency (HPA), Heartlands Hospital, Birmingham, UK. The laboratory provides a service to clinics serving approximately 4000 treated patients (~20% of UK treated population), of which 10-20% are tested for resistance per year. The samples were submitted for routine genotypic resistance testing between 1999 and 2001, and include samples from acute infections, chronic but drug naïve infections and from patients at the time of therapy failure. Clinical information on the patients was available for most samples, including the date of collection, geographic area, reason for analysis and viral load, as well as molecular information such as subtype of the virus or genotypic patterns of drug resistance.

The number of entries in the database expanded from around 2000 entries at the time the present body of work was initiated to reach up to 4000 sequences to date. On average, 10% of these sequences were classified as incident infections, an incident infection corresponding to a seroconversion within six months of an HIV-1 negative test. Follow up samples, i.e. samples collected from a same patient at different time points of monitoring, were available for approximately 10% of the patients represented in the database.

1.2. Ethics Committee Approval and Patient Consent

This research was approved by the Health Protection Agency Ethics Committee, allowing the submission of results generated by ASRU to non-commercial databases in anonymous form. Thus, clinical information such as the age group, sex or risk group of the patients was stored in a secure independent database in order to preserve patient anonymity prior to analysis. However, specific consent was requested from patients appearing within transmission clusters in order to document potential sexual contacts, whilst blinding clinicians and patients to the laboratory data. Such epidemiological information was only obtained from a minority of patients.

2. Bench Work

Amplification and sequencing of the *pol* region of plasma-derived HIV-1 was carried out for routine genotypic resistance testing by the Antiviral Reference Unit of Public Health Laboratory Service, Heartlands Hospital, Birmingham.

2.1. Extraction of HIV-1 RNA from Blood Samples

Viral RNA was extracted from 1 ml of plasma stored in EDTA at -20°C, using QIAamp Viral RNA Extraction kit (Qiagen) according to the manufacturer's instructions. Explicitly, 200 ml of blood plasma aliquot was centrifuged for one hour at 17,000 rpm and at + 4°C. The supernatant was carefully discarded and the pellet re-suspended in 140 μl of RPMI Cell Culture Medium. The mixture was then vortexed for approximately 15 seconds. 560 μl of Qiagen AVL Lysis Buffer, containing carrier RNA, was added to the suspension, vortexed and incubated at room temperature for 10 minutes. 560 μl of 100% ethanol was further added to the suspension and vortexed, then 630 μl of the mixture was added to a Qiagen Spin Column and centrifuged for 1 min at 13,000 rpm. Filtrate was discarded and a further 630 μl of the suspension was added to the spin column for an additional centrifugation, as described above. After discard of the filtrate, 500 µl of AW1 Wash Buffer was added to the column and the suspension was centrifuged for 1min at 13,000 rpm. The previous step was then repeated with addition of 500 µl of AW2 Wash Buffer. The RNA was finally eluted by adding 40 µl of Elution Buffer to the column, which was incubated at room temperature for 1 min, then centrifuged for 1 min at 13,000 rpm.

2.2. Reverse Transcription (RT) of HIV-1 RNA

The reverse transcriptase reaction was carried out immediately after extraction. cDNA was generated from the extracted HIV genomic RNA. The reaction was performed by reverse transcriptase polymerase chain (RT-PCR), using commercial Qiagen RT-PCR amplification kit. For this purpose, 15 µl of RNA was added to 5 µl of 5x RT buffer, 2.5 µl of each dNTP at 5 mM, 2 µl of random primers at 100 ng/µl, 0.1 µl RNase inhibitor, 0.5 µl of MMLV reverse transcriptase and 4.9 µl of dH_2O. The cycling conditions for the reverse transcription were 37°C for 60 min, followed by 94°C for 10 min.

2.3. Polymerase Chain Reaction (PCR) Amplification

pol gene

The entire protease gene and the first 220 codons of RT gene of the samples were amplified by nested polymerase chain reaction (PCR), using the Qiagen Taq PCR mastermix kit. The sequence and orientation of the primers used are given in Table 2.1 and Fig. 2.1.

To perform the primary round of the nested PCR, 5 µl of cDNA were added to 10 µl of Qiagen mastermix, 2.5 µl of each primer and 5 µl of dH_2O. For the secondary reaction, 5 µl of first round product were added to 50 µl of Qiagen mastermix, 20 µl of each primer and 10 µl of dH_2O. Conditions for both first and second round reactions were: 15 sec at 94°C for 1 cycle, followed by 30 sec at 95°C, 30 sec at 54°C, and 45 sec at 72°C for 40 cycles, finally followed by 10 min at 72°C.

Table 2.1. Primer sequences for the amplification of the protease and reverse transcriptase genes.

Primer	Position*	Sequence	Orientation
Pout3	2019-2038	AAG GGC TGT TGG AAA TGT GG	first round sense primer
Pout4	3298-3275	GTC TTT TTC TGG CAG CAG TAT AGG	first round anti-sense primer
Pin1	2503-2525	AAT TGG AAG AAA TCT GTT GAG TC	RT second round sense
Pin2	3276-3254	GGC TGT ACT GTC CAT TTA TCA GG	RT second round anti-sense
Pin3	2604-2585	GGG CCA TCC ATT CCT GGC TT	PR second round anti-sense
Pin4	2147-2167	CAG AGC CAA CAG CCC CAC CAG	PR second round sense primer
Pin8	3021-3002	GCT GGT GAT CCT TTC CAT CC	RT sequencing
Pin9	2864-2883	GTA ACA GTA CTG GAT GTG GG	RT sequencing

* Position with respect to HXB2 reference strain.

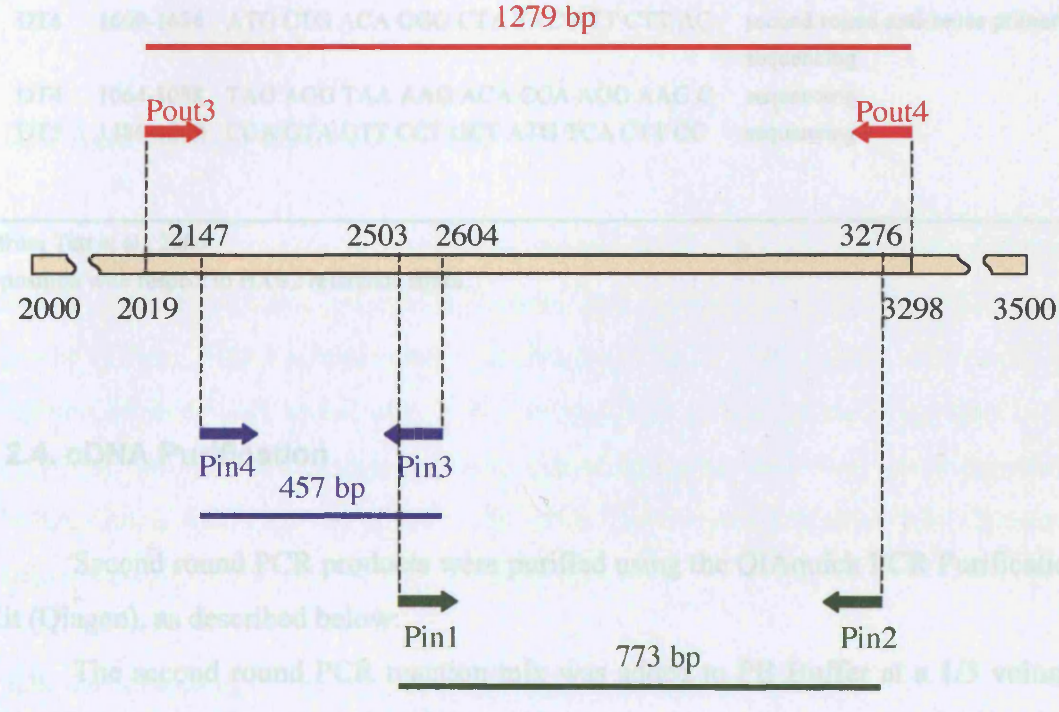


Fig. 2.1. Position of the primers used for the amplification of the PR and RT genes. First round primers are indicated in red. Second round primers for the amplification of the protease and reverse transcriptase genes are shown in blue and green respectively. Positions are given with respect to HXB2 strain.

gag & env gene

Two fragments of 690 and 550 base pairs, partially covering the p17/p24 region of the *gag* gene and the V3 loop region of the *env* gene respectively, were amplified by multiplex nested PCR from cDNA already used for *pol* gene amplification using Qiagen Taq PCR mastermix. Sequences of the primers used for the amplification of the *gag* and *env* genes are detailed in Table 2.2 and Table 2.3 respectively. Cycling conditions are given in Table 2.4.

Table 2.2: Primers for *gag* genes amplification and sequencing

Primer[a]	Position[b]	Sequence	Orientation
DT1	790-812	ATG GGT GCG AGA GCG TCA GTA TT	first round sense primer
DT7	1818-1844	CCC TGA CAT GCT GTC ATC ATT TCT TCT	first round anti-sense primer
DT3	886-908	CAT CTA GTA TGG GCA AGC AGG GA	second round sense primer; sequencing
DT6	1609-1634	ATG CTG ACA GGG CTA TAC ATT CTT AC	second round anti-sense primer; sequencing
DT4	1064-1088	TAG AGG TAA AAG ACA CCA AGG AAG C	sequencing
DT5	1486-1509	CGA GTA GTT CCT GCT ATG TCA CTT CC	sequencing

[a] from Tatt et al., 2000

[b] position with respect to HXB2 reference strain.

2.4. cDNA Purification

Second round PCR products were purified using the QIAquick PCR Purification Kit (Qiagen), as described below:

The second round PCR reaction mix was added to PB Buffer at a 1/5 volume ratio. The suspension was placed in a QIAquick spin column and centrifuged at 13,000 rpm for 1 minute. This step was repeated until process of the entire reaction volume was processed. 750 µl of PE wash buffer was then added to the solution and the QIAquick column centrifuged at 13,000 rpm for 1 minute. The filtrate was discarded and the column centrifuged again 10,000 rpm for 1 minute to ensure complete elimination of the wash buffer. Finally, DNA was eluted in 30 µl of distilled water by centrifugation at 10,000 rpm for 1 minute, after a brief incubation at room temperature.

Table 2.3: Primers for *env* genes amplification and sequencing

Primer	Position[c]	Sequence	Orientation
ED5[a]	6557-6582	ATG GGA TCA AAG CCT AAA GCC ATG TG	first round sense primer
ED12[a]	7782-7811	AGT GCT TCC TGC TGC TCC CAA GAA CCC	first round anti-sense primer
ED31[a]	6817-6845	ACC TCA GCC ATA ACA CAA GCC TGT CCA	second round sense primer; sequencing
ED33[a]	7360-7381	TTG CAA TAG AAA AAT TCC CCT C	second round anti-sense primer; sequencing
621[b]	6945-6967	GTA CAT TGT ACT GTG CTG ACA TT	sequencing
623[b]	6827-6846	TAC ACA AGC CTG TCC AAA GG	sequencing
ES7[b]	7001-7020	CTG TTA AAT GGT AGC CTA GC	sequencing
ES8[b]	7647-7667	CAC TTC TCC AAT TGT CCC TCA	sequencing

[a] modified from *Delwart et al., 1993.*
[b] modified from *Arnold et al., 1995.*
[c] position with respect to HXB2 reference strain.

2.5. Agarose Gel Electrophoresis

In order to estimate the amplicons' concentration, 2 µl of the purified product was visualised by agarose gel electrophoresis. In a microwave oven, 1% "Ultra Pure" agarose (Gibco, Life Technologies) was dissolved in 1x TBE buffer. After cooling, ethidium bromide was added at a final concentration of 0.5 µg/ml. Migration of the cDNA samples was carried on at 100 volts in loading buffer, until clear band separation. The amplicons were then visualised under Ultra Violet transilluminator (LKB Bromma Macrovue).

2.6. Sequencing

Sequencing of the amplicons was undertaken using Beckman Coulter CEQ2000 protocols. Approximately 100 fmoles of DNA template were added to sequencing master mix provided by the manufacturer and primers at a concentration of 2 pmol/µl.

Table 2.4. Thermocycler conditions for *gag* and *env* genes amplification by PCR

		1st round			2d round		
		Number of cycles	Temperature (°C)	Time (min)	Number of cycles	Temperature (°C)	Time (min)
gag	*Step 1*	1	94	0.15	1	94	0.15
	Step 2	40	95	0.30	40	95	0.30
			54	0.30		54	0.30
			72	0.45		72	0.45
	Step 3	1	72	10.00	1	72	10.00
	Hold	1	4		1	4	
env	*Step 1*	1	94	2.00	1	94	2.00
	Step 2	35	94	0.15	35	94	0.15
			50	0.30		55	0.30
			72	1.00		72	1.00
	Step 3	1	72	10.00	1	72	10.00
	Hold	1	4		1	4	

Cycling conditions for the sequencing reaction were:

96°C for 20 sec

50°C for 20 sec

60°C for 4 min

for 30 cycles followed by holding at 4°C.

Chromatograms of the resulting forward and reverse sequences were edited and aligned using Sequencher software (Gene Codes, Ann Arbor, Michigan), from which a consensus sequence was obtained for each amplified region.

2.7. Subtyping & Drug Resistance Testing

The subtype of the samples was determined on the basis of the *pol* genetic variability by submission of the generated sequence to the Stanford HIV RT and Protease Sequence Database *(http://hivdb.stanford.edu/)*. When submitted to the database subtyping tool, a query sequence is compared to a list of reference sequences representing each of the 10 known subtypes of HIV-1 Main group, as well as the most common circulating recombination forms known to date. The subtype of the closest reference sequence is then assigned to the query, as determined by calculation of pairwise uncorrected-nucleotide distance.

A drug resistance interpretation was performed for each generated *pol* sequence, using the HIVseq program (beta version) available on the Stanford HIV RT and Protease Sequence Database website. HIVseq accepts user-submitted RT and protease sequences, compares them to the consensus subtype B reference sequence, and uses the differences as query parameters for interrogating the Stanford HIV Drug Resistance database (Shafer et al. 2000b). This test allows the prediction of genotypic drug resistance to 16 available drugs.

3. Molecular Evolution and Phylogenetics

According to the classical view of taxonomy, the relationship between species (i.e. their phylogeny) is inferred from the comparison of morphological characters (Linnaeus. 1758). The 'molecular revolution' that took place in the past decades, however, dramatically changed the general perception of evolution and hierarchical classification of organisms. The origin of molecular evolution as a science appeared long before the support for heredity was characterised (Watson and Crick; 1953). The field was in fact pioneered in the early 20th century by the work of Georges Nuttall (1902; 1904). Nuttall attempted to characterize a 'blood relationship' between organisms by mixing sera and anti-sera from different species. The more closely related the species, the stronger the cross-reaction between sera and anti-sera. Long after Nuttall's work, the understanding of genetic changes driving evolution rose with the formulation of the synthetic theory of Neo-Darwinism (Huxley. 1942), which connected the discovery of the molecular units of evolution (i.e. genes) with the mechanisms of genetic variation (i.e. natural selection).

The idea of macromolecules carrying information only took off decades later, however, with the first successful sequencing of a complete protein, insulin, by Frederick Sanger and his colleagues (Sanger. 1959). The onset of sequencing, for which Sanger was awarded the 1958 Nobel Prize in chemistry, gave substantial way to molecular systematics and many biologists began to argue that the best way to answer questions about evolution was to study them at the molecular level. Hence Sanger and colleagues published the first ever comparison of amino acid sequences from different species (Brown et al. 1955), while Walter Fitch and Emanuel Margoliash (1967) showed in a seminal article how to use molecular information to reconstruct phylogenies that were remarkably similar to the ones based on more traditional taxonomic characters. Herein the era of molecular phylogenetics was initiated.

It is now accepted that nucleic acid sequences hold valuable phylogenetic signal. Closely related organisms share a high degree of agreement in their molecular structure, while the molecules of distantly related organisms usually show patterns of dissimilarity. Molecular phylogenetics exploits these evolutionary 'footprints' accumulated through time by DNA or protein sequences in order to reconstruct their phylogeny. Phylogenies are traditionally represented in the shape of a schematic tree, i.e. a phylogenetic tree, built on the basis of a sequence alignment. Historically, the primary interest in constructing phylogenetic trees was the pattern of evolutionary relationships itself. More recently, however, trees have been generated to derive information regarding the processes responsible for the observed pattern of evolutionary relationships, and the tree topology becomes the framework upon which further inference can be drawn. As such, phylogenetics facilitates analyses of rates of evolution (Drummond et al. 2003a), recombination (Posada et al. 2002), divergence of lineages and population demographics (Grassly and Holmes. 1999).

Although sequence comparisons can be done using either nucleotide or amino-acid sequences, only evolutionary changes between nucleotide sequences will be covered in the present chapter.

3.1. Sequence Alignments

Comparisons between two stretches of nucleic acids are only valid if the considered sequences are homologous in the evolutionary sense of the term. That is, if the two regions of interest are directly derived from a common ancestor. Hence, constructing a sequence alignment is a prior requirement to any molecular evolution analysis and must be considered carefully. Aligning two or more sequences allows the identification and localisation of specific evolutionary alterations, such as single nucleotide polymorphisms (SNP), insertions or deletions accumulated by the different lineages since their divergence from a shared ancestor.

A sequence alignment allows three types of base differences to be recognised: matches, mismatches and gaps (Fig. 2.2). A match occurs when the same base is encountered at a given position; a mismatch is found when at least one substitution has occurred since the two sequences diverged from each other; a gap indicates that a deletion or an insertion has occurred in one of the compared sequences.

When comparing sequences with low genetic divergence, alignments can easily be performed manually, with the help of sequence editing softwares such as Bioedit or MacClade (Maddison and Maddison. 1989; Hall. 2000). Alternatively, a plethora of programs implementing sequence alignments algorithms are publicly available, the most widely used being ClustalX (Thompson et al. 1997). ClustalX implements a scoring system where base matches (or mismatches) are assigned a positive score, whereas gaps are assigned negative scores. The severity of the gap penalty varies when either a gap is introduced or extended. A heuristic search is then performed to select for the best alignment, i.e. the alignment with the highest score.

In the present thesis gene sequences were initially aligned in-frame using ClustalX version 1.81. Alignments were subsequently subjected to visual inspection and manually improved, using the sequence editor Bioedit.

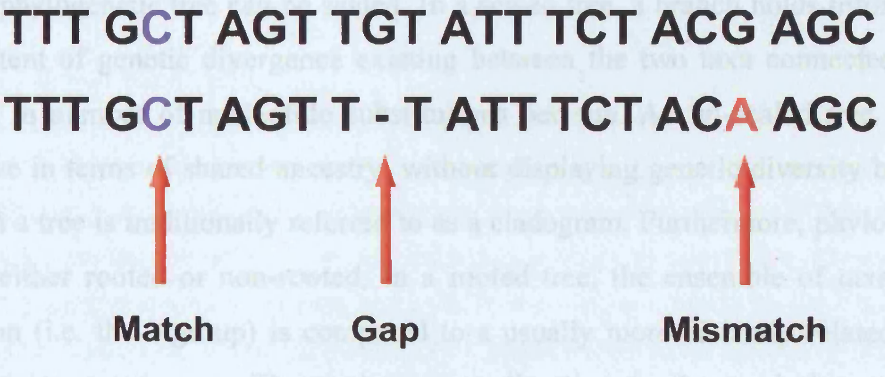

TTT GCT AGT TGT ATT TCT ACG AGC
TTT GCT AGT T - T ATT TCT ACA AGC
Match
Gap
Mismatch

Fig. 2.2. Different types of bases impairments. A match occurs when the same base is encountered at a given position; a mismatch is found when at least one substitution occurred since the two sequences diverged from each other; a gap indicates that a deletion or an insertion occurred in one of the compared sequences.

3.2. Phylogenetic Trees

A phylogenetic tree is a schematic representation of plausible evolutionary relationships existing amongst a group of sampled organisms. Within the frame of molecular analysis, a phylogenetic tree is directly inferred from the specific evolutionary patterns held by a sequence alignment.

An example of phylogenetic tree is given in Fig. 2.3, illustrating the relationship between four species of Hominidae. A phylogenetic tree essentially consists of nodes linked together by branches. Nodes represent taxonomic units and may be either internal or external nodes. Terminal nodes, or tips, represent known sequences of extant or extinct organisms, also known as Operational Taxonomic Units (OTUs). Internal nodes represent theoretical ancestors. They rely on no empirical records and are referred to as Hypothetical Taxonomic Units (HTUs). Nodes are interconnected by branches, which symbolise relationships between the taxa in terms of descent and ancestry. Methods are also available enabling the reconstruction of networks of relatedness, where ancestry is not restricted to a single taxon (Bandelt and Dress. 1992). The branching patterns seen within a phylogenetic tree can be divided in groups of two or more taxa including both their common ancestors and their descendents. These branches patterns are referred to as clades or clusters.

A phylogenetic tree can be scaled. In a scaled tree, a branch holds information on the extent of genetic divergence existing between the two taxa connected by it, expressed in number of nucleotide substitutions per site. An un-scaled tree is only informative in terms of shared ancestry, without displaying genetic diversity between taxa. Such a tree is traditionally referred to as a cladogram. Furthermore, phylogenetic trees are either rooted or non-rooted. In a rooted tree, the ensemble of taxa under comparison (i.e. the ingroup) is compared to a usually more distantly related taxon called a root or outgroup. The root gives a direction to the evolution pathway reconstructed within the tree, since it represents the common ancestor from which a unique path leads to any other node. An unrooted tree only specifies the relationship among species, without identifying a common ancestor, or directional evolution.

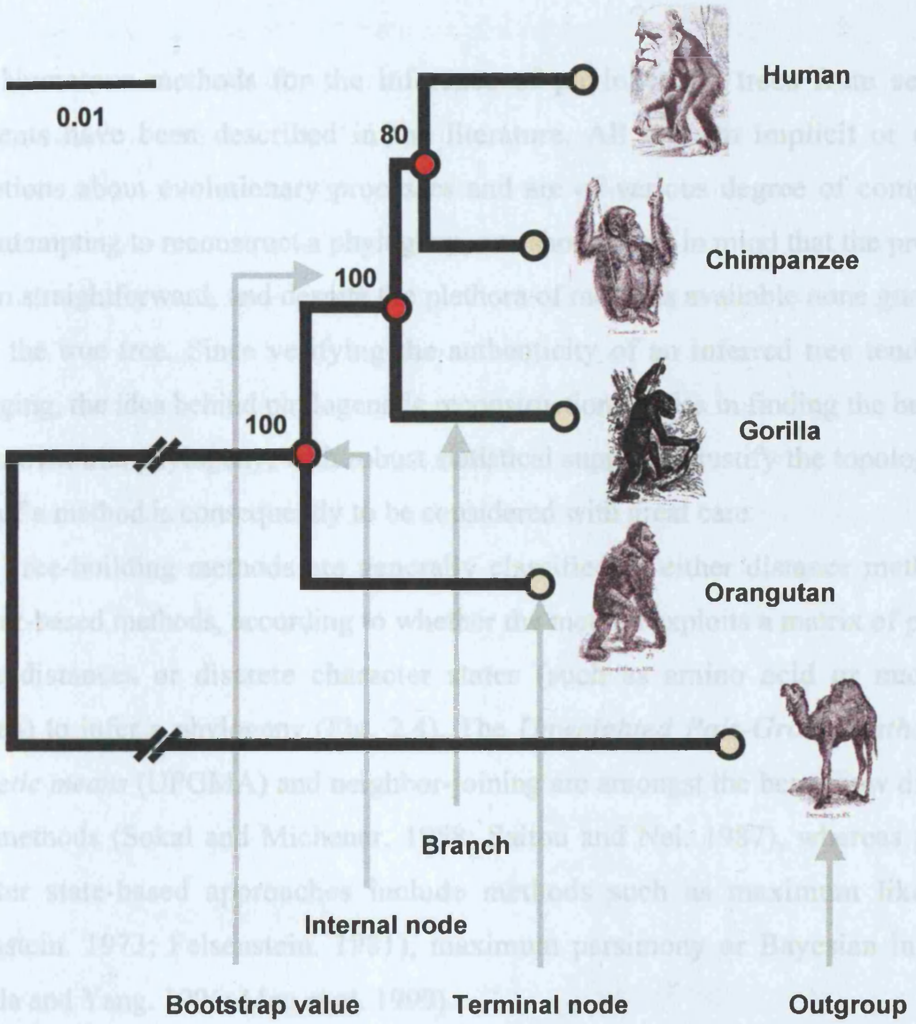


Fig. 2.3. Example of phylogenetic tree, illustrating the relationship between four species of Hominidae. Terminal nodes represent true taxonomic units (i.e. existing gene or protein sequences) and are labelled by orange dots. Internal nodes represent hypothetical ancestors, as labelled by red dots. Nodes are interconnected by branches symbolising the extend of genetic divergence between two units. In a phylogram, branches are scaled and expressed in number of nucleotide substitutions per site (see scale bar). Only horizontal branches have a biological significance, vertical branches being for clarity. Figures at internal nodes are bootstrap values supporting the corresponding branch. By convention, only values above 50 are indicated. The tree is rooted against a distantly related specie, or outgroup, here an Artiodactyl sequence. *After Glazko & Nei, 2003.*

3.3. Phylogenetic Methods

Numerous methods for the inference of phylogenetic trees from sequence alignments have been described in the literature. All rely on implicit or explicit assumptions about evolutionary processes and are of various degree of complexity. When attempting to reconstruct a phylogeny, one should bear in mind that the process is far from straightforward, and despite the plethora of methods available none guarantees to find the true tree. Since verifying the authenticity of an inferred tree tends to be challenging, the idea behind phylogenetic reconstruction resides in finding the best fit to the unknown true phylogeny, with robust statistical support to justify the topology. The choice of a method is consequently to be considered with great care.

Tree-building methods are generally classified as either distance methods or character-based methods, according to whether the method exploits a matrix of pairwise genetic distances or discrete character states (such as amino acid or nucleotide positions) to infer a phylogeny (Fig. 2.4). The *Unweighted Pair-Group method with Arithmetic means* (UPGMA) and neighbor-joining are amongst the best know distance-based methods (Sokal and Michener. 1958; Saitou and Nei. 1987), whereas popular character state-based approaches include methods such as maximum likelihood (Felsenstein. 1973; Felsenstein. 1981), maximum parsimony or Bayesian inference (Rannala and Yang. 1996; Mau et al. 1999).

Distance Methods

Distance methods, also known as algorithmic methods, first convert aligned sequences into a matrix of pairwise genetic distances. For every pair of sequences in the alignment, a parametric distance is calculated as the fraction of positions in which the two sequences differ, providing a measure of dissimilarity. Distance matrix methods involve two steps. First, the evolutionary distance is calculated for every possible sequence pair in the given alignment. Secondly, the tree is inferred on the basis of the relationship between the distance values.

The simplest, and oldest, distance method is Unweighted Pair-Group method with Arithmetic means (UPGMA) (Sokal and Michener. 1958). Programs implementing this method first find the pair of taxa with the smallest genetic distance between them and define the branching between then as half of that distance. These two taxa are combined in a cluster, reducing the number of entries by one. A new distance matrix is then computed with the distance from the cluster to each of the remaining sequences. That process is repeated on the new matrix and reiterated until there is only one entry left in the matrix. An additive tree is finally built by adding up clusters and respective branch lengths from the root to the node of the last cluster generated. UPGMA trees are said to be ultrametric, i.e. all taxa are equally distant from the root. This is a consequence of the main assumption the method is based on, that is the existence of a constant molecular clock amongst lineages. According to this assumption, substitutions accumulate at the same rate in all lineages diverging from a common ancestor. Since a large body of evidence tend to prove that a global molecular clock does not exist, or remains circumscribed to groups of species (Page and Holmes. 1998; Li and Graur. 2000), such an assumption limits the reliability of the method. For that reason, amongst others, UPGMA tends not to be used anymore.

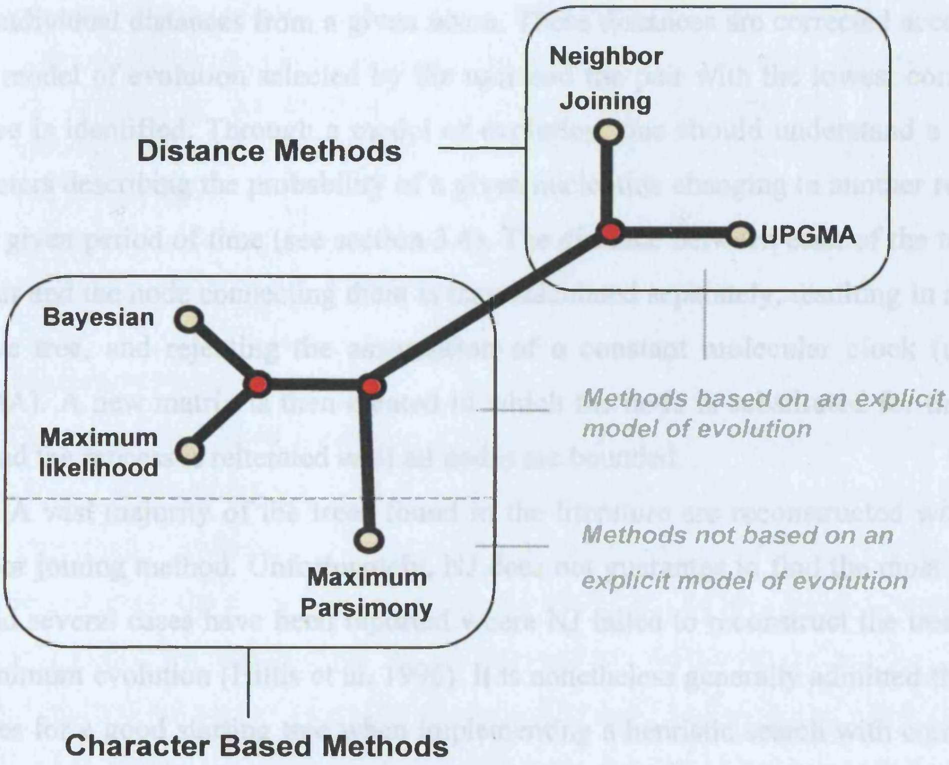
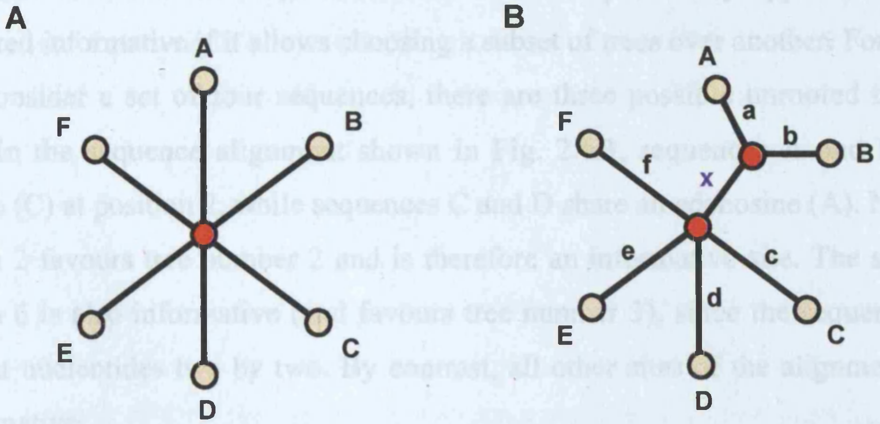


Fig. 2.4. Relationship between some common phylogenetic methods.

Developed by Saitou and Nei in 1987 (Saitou and Nei. 1987), and latter modified by Studier and Keppler (Studier and Keppler. 1988), the neighbor joining method (NJ) remains to date the most popular distance method. Its popularity is mainly due to the accessibility and rapidity of its implementation. Like UPGMA, NJ employs a matrix of genetic distances, on the basis of which a tree is built by gradually finding neighbours exhibiting the minimum genetic distance. The implementation starts with a star-like tree with no internal branches or hierarchical structure. Internal branches are gradually added to that tree by considering every possible pairs of taxa and selecting the one pair that gives the smallest sum of branch lengths (see Fig. 2.5). The selected pair is then regarded as a single composite taxon and a new matrix of pairwise genetic distance is computed, as with UPGMA. The next pair with the smallest branch length is selected again and the process is repeated until all internal branches are found. The main difference between UPGMA and NJ methods lies on that NJ does not construct intermediate clusters with nodes at the mid point but directly calculates for each entry a 'net' distance from all other entries in the matrix. This distance is expressed as the sum of all individual distances from a given taxon. These distances are corrected according to the model of evolution selected by the user and the pair with the lowest corrected distance is identified. Through a model of evolution, one should understand a set of parameters describing the probability of a given nucleotide changing to another residue over a given period of time (see section 3.4). The distance between each of the taxa in that pair and the node connecting them is then calculated separately, resulting in a non-additive tree, and rejecting the assumption of a constant molecular clock (unlike UPGMA). A new matrix is then created in which the node is substituted for the two taxa, and the process is reiterated until all nodes are bounded.

A vast majority of the trees found in the literature are reconstructed with the neighbor joining method. Unfortunately, NJ does not guarantee to find the most likely tree and several cases have been reported where NJ failed to reconstruct the tree with the minimum evolution (Hillis et al. 1996). It is nonetheless generally admitted that NJ provides for a good starting tree when implementing a heuristic search with computer intensive character-based methods such as parsimony or maximum likelihood. This procedure was followed for the reconstruction of the trees presented in this thesis.



$d_{AB} = a + b$ $d_{AC} = a + x + c$

$d_{AD} = a + x + d$ ...

$d_{EF} = e + f$

Fig. 2.5. Implementation of the NJ algorithm. (A) A star-like tree is used at a starting topology. Letters on the branches represent the genetic distance between the two nodes connected by that branch. (B) Internal branches are progressively inserted (like branch x in our example) and the total length of the tree is calculated by summing up the distances between each external nodes (B). Branches minimizing the total tree length are retained and the shortest tree selected. *After Li and Grauer, 1999.*

Maximum Parsimony

Originally developed for amino acid sequence data by Eck and Dayhoff (Eck and Dayhoff. 1966) and latter extended for the analysis of nucleotide sequences (Fitch. 1977), parsimony analyses have been the predominant approach in molecular phylogenetics from the early 70's to relatively recently. The idea behind parsimony inference is to find the tree topology for a given sequence alignment that can be explain with the smallest number of character changes. Starting from an initial topology, the maximum parsimony algorithm infers the minimum number of mutations required to justify all nodes of the tree at every sequence position. The process is then repeated for all theoretically possible tree topologies and the tree requiring the minimum number of evolutionary changes, called the minimum tree, is selected as the best tree. Two or more trees with the same number of minimum changes can potentially be found this way.

The notion of informative sites underlies the parsimony approach. A site is considered informative if it allows choosing a subset of trees over another. For instance, if we consider a set of four sequences, there are three possible unrooted trees (Fig. 2.6A). In the sequence alignment shown in Fig. 2.6B, sequences A and B share a cytosine (C) at position 2, while sequences C and D share an adenosine (A). Nucleotide position 2 favours tree number 2 and is therefore an informative site. The same way, position 6 is also informative (and favours tree number 3), since the sequences share different nucleotides two by two. By contrast, all other sites of the alignment remain uninformative.

The identification of all informative sites in an alignment is the fist step in parsimony inference, followed by the calculation of the minimum number of substitutions at each informative site. The sum of the number of changes across all informative sites for each possible tree will then allow designating the most parsimonious tree, i.e. the tree with the smallest number of nucleotide substitutions.

The main limitation of parsimony lies in its inconsistency. Simulations have shown that maximum parsimony leads to incorrect results with an infinite amount of data (Felsenstein. 1978). Particularly, species at the ends of long branches in a parsimony tree have a tendency to be made artificially close to each other, due to the high frequency of parallel changes that arrive at a same position. This phenomenon is called 'long-branch attraction' and the situation in which this inconsistency occurs is often referred to as the 'Felsenstein zone'.

Another limitation of parsimony is the assumption that a minimal number of changes reflects a minimal evolution between two taxa. Given that all evolutionary changes between two sequences may not be visible (as in the case of multiple substitutions at a same nucleotide position), this assumption is only valid for closely related sequences, and the dissimilarity between two distantly related taxa might as well be underestimated in a parsimony tree. Also, since a small proportion of the sites in an alignment are truly informative, most of the sequence information may not be used and the inference process remains therefore sub-optimal. For these reasons parsimony was not used in the present study.

Maximum Likelihood

The first applications of a maximum likelihood (ML) approach to tree reconstruction were accomplished in the 70's when Joe Felsenstein developed new algorithms for molecular data (Felsenstein. 1973; Felsenstein. 1981). The method relies on sophisticated statistical theory and exploits the concept of likelihood (Swofford et al. 1996; Li. 1997). In statistics, likelihood is the probability P of observing the data D given the hypothesis H, and is noted

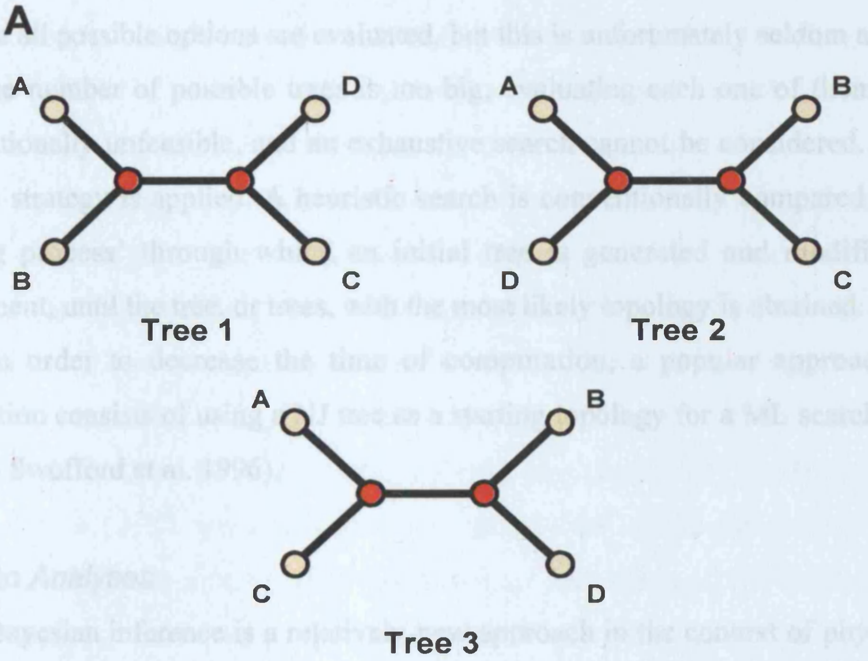
$$L = P\,(D \mid H)$$

In terms of phylogenetic reconstruction, D corresponds to the sequence alignment of interest, while H is a given phylogenetic tree. Thus the likelihood of a tree corresponds to the probability of that tree describing the patterns of the sequence alignment, given a specific model of nucleotide substitution. A maximum likelihood approach will result in the calculation of the likelihood of all possible unrooted trees for the specified alignment and selecting the one(s) associated with the highest, or maximum, likelihood. Since likelihood values are usually of very small order of magnitude, they are traditionally expressed as natural logarithms and referred to as log-likelihood.

Maximum likelihood approaches require three elements, namely a sequence alignment, a user-defined explicit model of nucleotide substitution and an initial tree topology. In order to compute the likelihood value of that tree to describing the alignment under the selected model of evolution, the likelihoods of observing the substitution patterns are calculated at each nucleotide position of the alignment, and then summed. This is to say, given a model of evolution, the probability P_{ij}(D | H) that two sequences would have nucleotide i and j, respectively, at the given position is calculated for each possible nucleotide substitution, and for each possible pair of sequences in the alignment. The log likelihood of observing the sequences is then obtained by calculating the sum of the log likelihoods at each individual site.

The main advantage of ML is that it allows the user to specify the model of molecular evolution for the computation of the data. It is an advantage in that it allows the user to control on the assumptions made during computation. However, such a dependence on a model can easily turn to a disadvantage, since the application of a particular model of evolution can seriously bias the outcome of the search, and ML trees may not be reliable if the model used is not selected with care.

Moreover, maximum likelihood remains a computationally intensive method and implementing a ML tree rapidly becomes a time consuming process as the number of taxa increases. Since character-based methods in general, and ML in particular, search for the best fit to the data amongst all possible trees, the time of computation of these methods is highly dependent on the number of sequences under consideration.



B

Position	1	2	3	4	5	6	7	8	9
Sequence A	C	C	T	C	A	A	A	T	C
Sequence B	C	C	A	C	A	T	A	T	C
Sequence C	C	A	G	C	A	T	A	T	C
Sequence D	C	A	C	C	A	A	A	T	G

Fig. 2.6. (A) When considering a set of four sequences, three possible unrooted trees are possible. In the sequence alignment shown in (B), sequences A and B share a cytosine at position 2, while sequences C and D share an adenosine. Nucleotide position 2 favours tree number 1 and is therefore an informative site. The same way, position 6 is also informative (and favours tree number 2), since the sequences share different nucleotides two by two. By contrast, all other sites of the alignment remain uninformative.

Indeed, for three sequences, four possible unrooted trees are possible, whereas for ten sequences, the number of possible unrooted trees reaches 2.03 x 10^6. Hence, when the number of taxa to be compared remains small, an exhaustive search of all individual trees is technically feasible. Exhaustive searches provide a guarantee of finding the best tree since all possible options are evaluated, but this is unfortunately seldom achievable. When the number of possible trees is too big, evaluating each one of them becomes computationally unfeasible, and an exhaustive search cannot be considered. Instead, a heuristic strategy is applied. A heuristic search is conventionally compared to a 'hill-climbing process' through which an initial tree is generated and modified by re-arrangement, until the tree, or trees, with the most likely topology is obtained.

In order to decrease the time of computation, a popular approach of tree construction consists of using a NJ tree as a starting topology for a ML search (Hillis et al. 1996; Swofford et al. 1996).

Bayesian Analyses

Bayesian inference is a relatively new approach in the context of phylogenetics but its strength and accessibility makes it an increasingly popular method for tree searching. Like maximum likelihood, Bayesian inference incorporates an explicit model of sequence evolution and looks for the trees that correlate best to the sequence alignment under the given model. However, while maximum likelihood searches for the tree that, under a hypothesis (i.e. a tree topology), maximises the chance of observing the data (i.e. the sequence alignment), Bayesian reasoning works the problem in a reverse fashion and searches for the tree that maximises the chances of seeing that particular tree given the data and the model of evolution. In other words, maximum likelihood searches for the tree that maximise the probability *P (Data | Tree),* while Bayesian inference searches for the tree that maximise the probability *P (Tree | Data).*

Bayesian inference relies on the use of a simple mathematical formula used for calculating conditional probabilities and is named after the British cleric Thomas Bayes who developed it during the 18th century (Bayes. 1764). According to the Bayes theorem, the (posterior) probability P of an hypothesis H , given the conditional event E is:

$$P(H_i \mid E) = P(E \mid H_i)\, P(H_i) \,/\, \Sigma_j P(E \mid H_j)\, P(H_j)$$

where $P(E \mid H_i)$ is the probability of observing the event E under the hypothesis H_i, $P(H_i)$ the prior probability of the hypothesis H_i before observing event E, and $\Sigma_j P(E \mid H_j)\ P(H_j)$ the average probability of event E across all probable hypotheses.

Traditionally Bayesian inference is implemented by the use of the Metropolis-Hastings Markov Chain Monte Carlo (MCMC) methods (Metropolis et al. 1953; Hastings. 1970). The image of a hill-climbing process is often used to describe the heuristic search of MCMC algorithms, as the process is comparable to a random walk over the space of all possible tree combinations. Indeed, looking for the tree that, amongst all other trees, exhibits the highest likelihood is comparable to searching for the highest peak when blindly walking through a hilly environment. This 'hill-climbing' process follows several steps: A random tree T1 is selected as the current tree and compared to a second tree T2; if the likelihood $L1$ of T1 is inferior to the likelihood $L2$ of T2, then T2 replaces T1 as the current solution (the climber goes one step up-hill); in contrary if $L1 > L2$, T1 is maintained as the current solution; the current tree is sampled and the whole process is reiterated a significant amount of times. The number of time a particular tree is 'visited' is proportional to its likelihood given the data, biasing the solutions by their likelihood score, and the program yields a set of trees that the algorithm has repeatedly visited (i.e. the top of the highest hill).

Like maximum likelihood, Bayesian inference allows control of the specified model of evolution. Moreover, the beauty of the method relies is that it allows for simultaneous independent searches that occasionally exchange information, increasing the efficiency of the computation. Also, MCMC gives a measure of statistical support for any sampled solution, given as the number of time a feature is represented in the sampled solutions. If, for instance, a clade is present in 75% of the sampled trees, there is a 75% chance that this clade is correct according to the assumption of our evolutionary model. Finally, Bayesian MCMC inference not only yields a set of best fitting tree topologies, but also estimates the parameters of the selected model, all at once. Several programs implement Bayesian phylogenetics, the most popular of these being MrBayes (Huelsenbeck. 2000).

3.4. Models of Molecular Evolution

We saw that many methods for phylogenetic reconstruction are dependent on an explicit model of molecular evolution. Modelling for evolutionary changes arose from the need to 'correct' observed genetic distances between two sequences into better measures of actual evolutionary distances. Indeed, the extent of observed genetic dissimilarity between two sequences is not linear with time, but curvilinear due to multiple substitutions, or multiple hits, occurring at the same site. As the number of substitutions increases, the chances that the same site may go through more than one change become higher, and failing to correct for these multiple hits may result in underestimating the true evolutionary distance between the two sequences.

To date, several models of evolution have been developed in order to describe the dynamics of nucleotide substitution between DNA sequences. These models rely on mathematical matrices and statistical inference techniques, reflecting with uneven accuracy the biological phenomena responsible for the mutational disparity found in molecular datasets. The relative complexity of these models is a function of the extent of the biological, biochemical and evolutionary knowledge they incorporate. Hence, it is generally admitted that more complex models give a better statistical fit to observed evolutionary patterns of gene sequences, and therefore lead to a better phylogenetic reconstruction (Goldman. 1993; Yang et al. 1994).

Three classes of parameters are implemented in the models of sequence evolution developed to date, namely base frequency, base exchangeability and rate heterogeneity among sites. Base frequency accounts for the respective frequency of the four bases (A, G, C and T) over all sequence sites. Allowing for certain bases to emerge more likely than other, when substitutions occur, is thought to reflect the compositional constraints nucleic acids are under, such as G-C content or secondary structures. For instance, the HIV-1 genome is deeply biased toward G to A transitions (Vartanian et al. 1991; Vartanian et al. 2002). Base exchangeability describes the tendency of bases to be substituted for one another. For instance, transitions (substitutions between purines or between pyrimidines) have been proven to occur at a higher rate than transversions (substitutions between purines to pyrimidines or visa versa) – see Fig.2.7. Base exchangeability reflects the biochemical similarity bases share and its effect on mutational bias.

Rate heterogeneity accounts for the difference in substitution rates across different taxa or DNA regions. Typically, a gamma distribution is used to describe such a heterogeneity in nucleotide substitution rate across sequences (Yang. 1994b). The range of the rate variation amongst sites is dictated by the shape parameter α of the distribution (Fig. 2.8). Small values of α will result in L-shape distributions, indicating extreme rate variation across the sequences, whereas higher values of α will reflect bell-shape distribution, as seen when most of the sites remain invariable and few have high rates of substitutions. Models featuring a gamma distribution of rate heterogeneity are conventionally given the suffix '+Γ'.

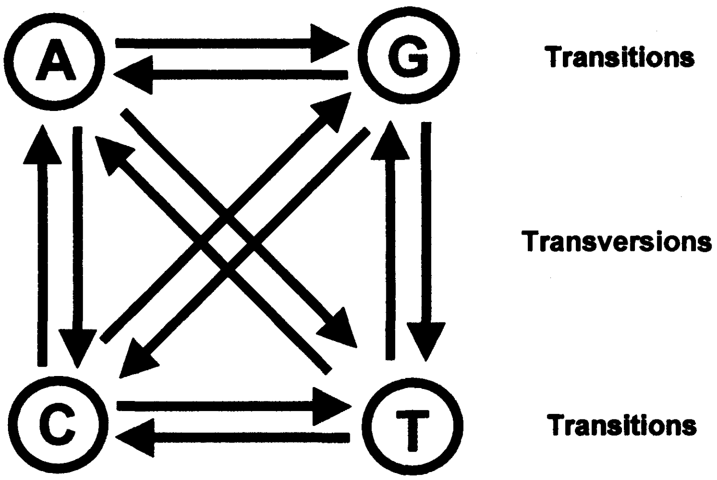


Fig.2.7. Different types of base exchangeability. Substitutions between purines or between pyrimidines, i.e. between adenosine and guanine or cytosine and thymidine, are termed transitions. Substitutions between purines to pyrimidines, or visa versa, are termed transversions.

Current models of evolution range from very simplistic to more intricate ones, varying on the type and number of parameters taken into account. Since it assumes an equal frequency of substitution for the four bases, the Jukes and Cantor one parameter model (JC) is the simplest model of sequence evolution (Jukes and Cantor. 1969). Kimura's two parameter model (K2P) assumes that the base frequency is equal along sites but that transitions occur at a higher rate than transversions (Kimura. 1980), while Felsenstein's F81 model assumes an equal transition/transversion rate but allows unequal frequency of base substitutions (Felsenstein. 1981). The HKY85 model allows both base frequency and transition/transversion rate to differ (Hasegawa et al. 1985).

Finally, the general time reversible (GTR) model assumes that all six pairs of substitutions occur at different rate (Yang. 1994a). The relationship between several models of nucleotide substitutions is illustrated in Fig.2.9.

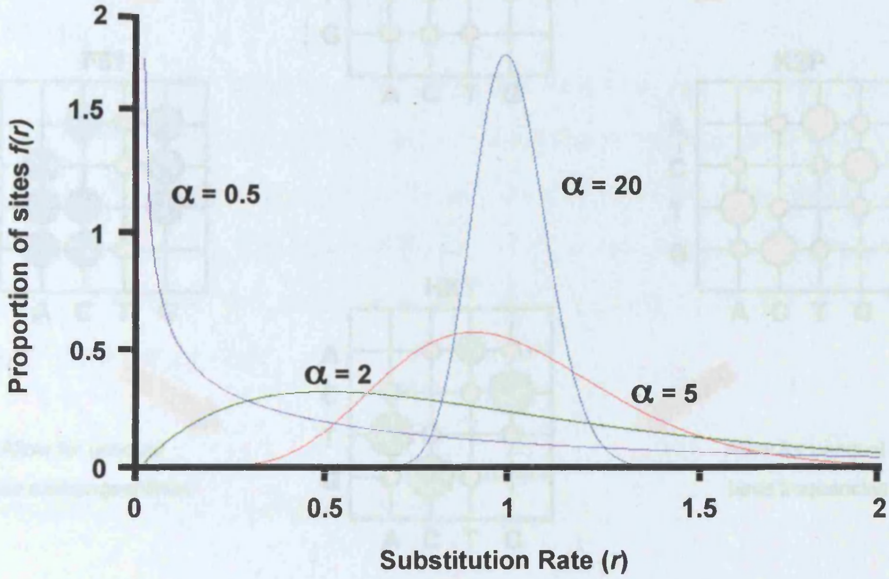


Fig. 2.8. Substitution rate heterogeneity amongst sites according to the shape parameter α of a gamma distribution. For values of α < 1, distributions are L-shaped, indicating extreme rate variation across the sequences. For values of α > 1, distributions are bell-shaped, distribution, as seen when most of the sites remain invariable. *After Yang (1996).*

Given the dependence of some phylogenetic methods on a model of sequence evolution, it is critical to accurately select the model with the best fit for a given dataset. The choice of an optimal model is usually derived from the patterns of the sequence dataset itself, and can be achieved by statistical hypothesis testing, such as likelihood ratio testing. The likelihood ratio test (LRT) is a statistical assessment of the goodness-of-fit between two models (i.e. models of evolution) and yields a likelihood ratio statistic Δ, which corresponds to the ratio of the likelihood of the alternative hypothesis (i.e. model 1) to the null hypothesis (i.e. model 0). Since likelihoods are of very small order of magnitude, likelihood scores are usually expressed as log-likelihoods:

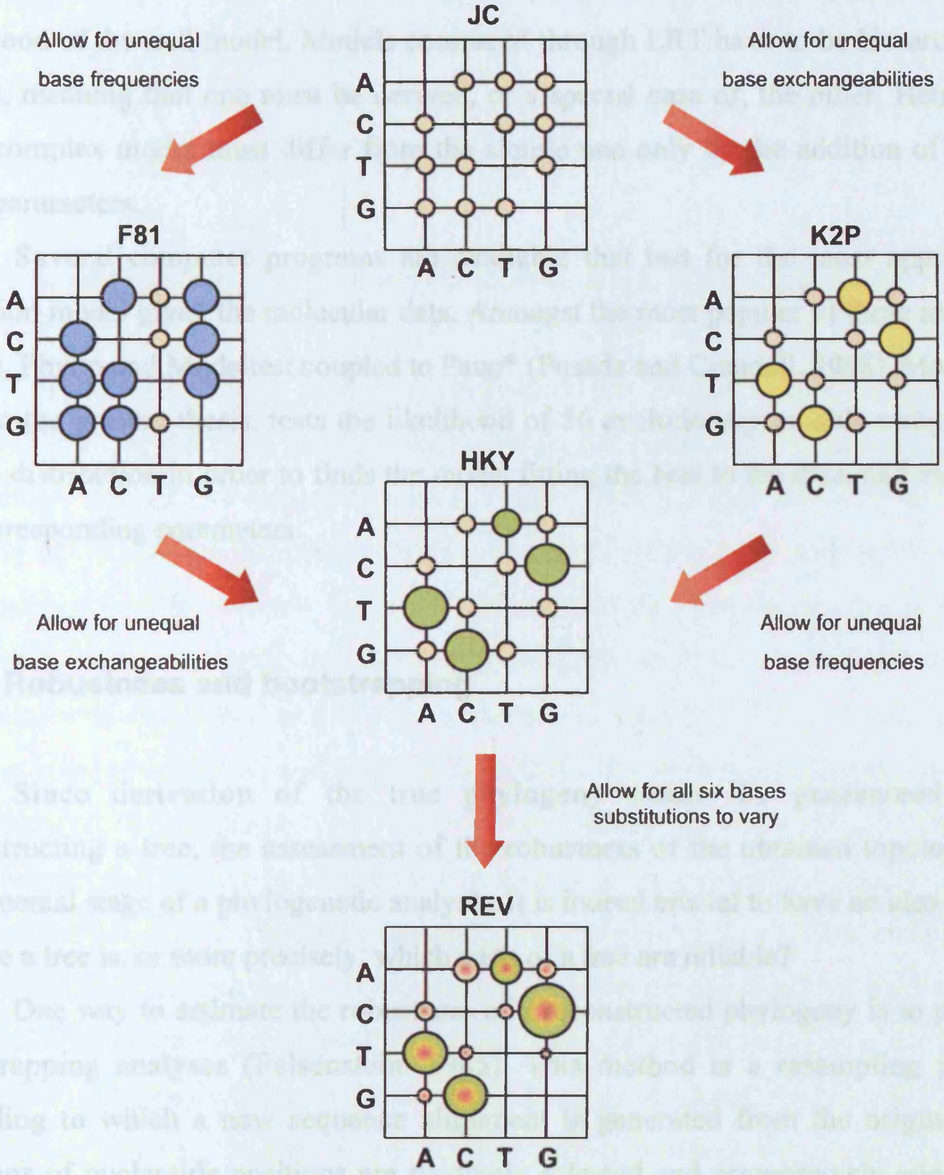

JC
Allow for unequal base frequencies
Allow for unequal base exchangeabilities
F81
K2P
HKY
Allow for unequal base exchangeabilities
Allow for unequal base frequencies
Allow for all six bases substitutions to vary
REV
A
C
T
G
A C T G

Fig. 2.9. Relationships amongst five models of nucleotide substitution, namely the Jukes-Cantor (JC), Felsenstein (F81), Kimura 2 parameters (K2P), Hasegawa-Kishino-Yano (HKY) and the General Reversible (REV) models. In each matrix, the bubble area is proportional to the rate of substitution. Red Arrows indicate the way models are nested within each other, allowing statistical model comparison by Likelihood Ratio Testing. *After Whelan et al., 2001.*

$$\Delta = 2 x (lnL_1\text{-}lnL_2)$$

where L_1 is the maximum likelihood of the alternative model, and L_2 the maximum likelihood of the null model. Models compared through LRT have to be hierarchically nested, meaning that one must be derived, or a special case of, the other. Hence, the most complex model must differ from the simple one only by the addition of one or more parameters.

Several computer programs are available that test for the most appropriate evolution model given the molecular data. Amongst the most popular of these are Tree-Puzzle, Phylip and Modeltest coupled to Paup* (Posada and Crandall. 1998). Modeltest, used in the present thesis, tests the likelihood of 56 evolutionary models using a Chi-square distribution in order to finds the model fitting the best to the data, and estimates the corresponding parameters.

3.5. Robustness and bootstrapping

Since derivation of the true phylogeny cannot be guaranteed when reconstructing a tree, the assessment of the robustness of the obtained topology is a fundamental stage of a phylogenetic analysis. It is indeed crucial to have an idea of how reliable a tree is, or more precisely, which parts of a tree are reliable?

One way to estimate the robustness of a reconstructed phylogeny is to perform bootstrapping analyses (Felsenstein. 1985). This method is a resampling process according to which a new sequence alignment is generated from the original one. Columns of nucleotide positions are randomly selected and progressively added to a new alignment, with repeats allowed, until the pseudoreplicate reaches the size of the template alignment. This process of generating new alignments is repeated a substantial amount of times, usually 1000 times, and trees are generated from these multiple new datasets. The number of time a particular branch of the original tree is found in the pseudo-trees will give an estimation of the reliability of it. The robustness of the branch is usually expressed as a bootstrap value indicated on the branch itself, representing the percentage of times that particular branch was present in the total number of pseudo-trees. For instance, a bootstrap value of 100 indicates that the branch associated to it was present in all trees, and is therefore extremely robust. By convention, only bootstrap scores greater than 50 are shown. Bootstrapping is a built-in feature implemented by most of the wildly used phylogenetic packages software, such as Paup* or Phylip.

4. Population Dynamics

4.1. Neutral Theory of Molecular Evolution & the Molecular Clock

The first definitive evidence supporting the neutral theory of molecular evolution was the discovery that synonymous substitutions occur at a much higher rate than non-synonymous changes (Kimura. 1968). With it arose the idea that a vast majority of DNA mutations may not have a functional relevance and may get fixed in a population simply by chance. In sharp contrast with the neo-Darwinian belief that natural selection is the main driving force of molecular evolution, the neutral theory of molecular evolution was first proposed by Motoo Kimura in the 60's (Kimura. 1968), and also addressed by King and Jukes (King and Jukes. 1969). According to this theory, a majority of the mutations fixed in a genome confer no selective advantage (or disadvantage) and are lost or fixed purely by a random sampling effect, or genetic drift. If we accept the idea of a randomly driven molecular evolution, a critical correlate has to be considered, according to which sequence evolution, embodied by nucleotide or amino-acid substitutions, follows a constant rate, or molecular clock. Under this assumption, the genetic diversity between two lineages of a population is a function of the mutation rate and the size of the population. Furthermore, the degree of genetic dissimilarity between two sequences can be used to date to the time at which these sequences diverged from their shared ancestor.

At this stage, a distinction has to be made between rates of mutation and rates of substitution. The rate of mutation corresponds to the rate at which mutational errors are incorporated into a genome during replication. This rate is traditionally expressed as the number of substitutions per nucleotide site per round of replication. The rate of mutation of an organism is known to be dictated by the specific efficiency (or lack of efficiency) of its polymerases. For instance, the lack of poof-reading activity of the reverse transcriptase of RNA viruses largely determines the rate of mutation of these viruses. By contrast, the rate of substitution of an organism corresponds to the rate at which newly acquired nucleotide substitutions become fixed and spread within a population. Hence, rates of substitution are expressed as number of nucleotide substitutions per site per unit of time (i.e. day, year or generation). The substitution rate of a virus or a gene is shaped by diverse evolutionary forces such as natural selection or random genetic drift and can be considered as the relative proportion of advantageous, neutral or advantageous mutational forces.

In the context of viral evolution, it is of great importance to accurately measure the mutation rate of a population, given that it is a major determinant of an infectious agent's adaptive capacity, and it is a key parameter for population genetics analyses and phylogenetics applied to population studies over time. However, there is in reality a large body of evidence suggesting that rates of molecular substitutions differ among species, genes, and even among different regions of the same gene, contradicting the idea of a molecular clock (see for example Li and Graur. 2000). Also, despite the accepted idea that negative selection is the main driving force in the evolution of RNA viruses (Leigh-Brown and Richman. 1997), a recent study showed that only a small minority of these organisms obey a strict molecular clock (Jenkins et al. 2002). Nonetheless, the authors showed that substitution rates estimated from large datasets should be reliable indicators of average rates of evolution, and models relaxing the molecular clock have been recently developed in order to allow molecular rates to vary over time and across lineages. These models make possible the estimation of an average rate of evolution, as well as allowing analyses requiring a clock-like environment (Thorne et al. 1998; Huelsenbeck et al. 2000).

4.2. Population Dynamics & The Coalescent

It is a well-established fact that evolution is a progressive process responsible for the patterns and characteristics of a population over time. By observing these evolutionary changes in sampled individuals from a given population, biological issues can retrospectively be addressed concerning the history of that population and the evolutionary mechanisms responsible for the observed patterns, i.e. the dynamics of that population.

It was the need for approaches that allowed the inference of these historical features from contemporary character states that lead to the development of the coalescent theory. First described by Kingman (Kingman. 1982a; Kingman. 1982b), the coalescent was concurrently discovered by Hudson (Hudson. 1983) and Tajima (Tajima. 1989), then generalised by Griffiths and Tavare (Griffiths and Tavare. 1994). It describes a stochastic process allowing the inference of historical states of a population from the genealogy of individuals randomly sampled from it. Since nucleic and amino acid sequences are known to hold inherent information about ancestral relatedness of individuals, or lineages, the coalescent is particularly well suited for the analyses of molecular data. Under these considerations, one can estimate the changes undergone by a population through time and, in the context of molecular epidemiology, reconstruct the history of viral epidemics (Kurbanov et al. 2003; Lemey et al. 2003; Pybus et al. 2003; Robbins et al. 2003; Twiddy et al. 2003). Concretely, the framework of the coalescent theory allows us to estimate specific demographic parameters such as the rate of nucleotide substitution μ, the current effective population size *Ne* or even date the origin of the most recent common ancestor (MRCA) of a given population, on the basis of the topology of a phylogenetic tree based on sample sequences.

The idea behind the coalescent theory is that, in the absence of selection, sampled lineages are assumed to randomly 'choose' their parent as we go back in time. Whenever two descendants 'pick' the same parent, their lineages is said to coalesce (see Fig. 2.10). The rate at which lineages coalesce depends on how many lineages are coalescing (the more lineages, the faster the rate), and on the size of the population (the more parents to choose from, the slower the rate). If we consider a sample of *n* gene sequences from a population of *N*, one can reconstruct the genealogy, or phylogeny tree, of these sequences, the root of which corresponds to the MRCA shared by the *n* taxa.

Looking backward in time, the number of ancestral sequences decreases as the lineages coalesce, until all lineages coalesce into the MRCA of the sample. Through this process, the probability of coalescence at the previous generation (i.e. the probability that two sequences in the current generation share a single ancestor in the previous generation) is $1/(2N)$, where N is the effective population size. The probability that coalescence occurred $t + 1$ generations ago is given by the distribution $1/2N\ (1 - 1/2N)^t$. If we assume that the number of mutations that occurred on a sequence in a given period of time is a Poisson variable, the mean time of $2N$ generations separating the two sequences implies that the mean number of mutations in the two sequences is $\theta = 4N\mu$, where μ is the mutation rate per sequence per generation.

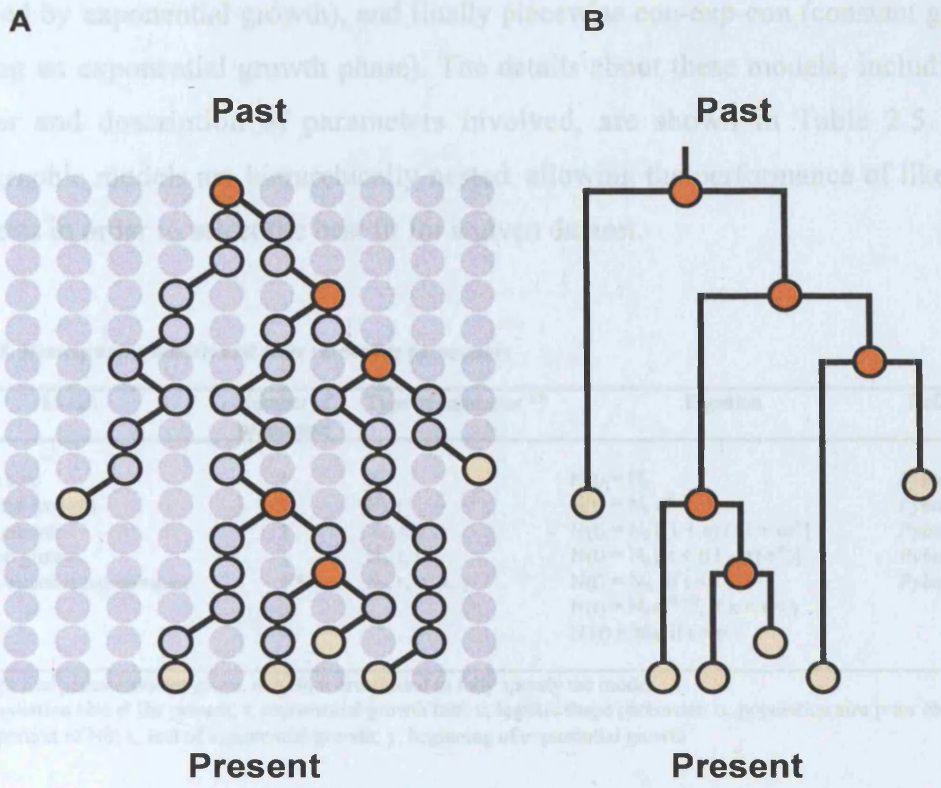


Fig. 2.10. Relationship between the demographic history (A) and the genealogy (B) of individuals sampled from a constant-sized population. Our example represents the genealogy of 6 individuals (yellow dots), over 15 generations (rows). Red dots correspond to hypothetical common ancestors. Moving back from present to past, the number of lineages in each generation decreases when two individuals shared a common ancestor (coalescent event), and increases when sampled individuals are encountered (sampling event). *After Drummond et al., 2003b.*

In the same way that phylogenetic inference is tightly bound to the modelling of molecular evolution, the coalescent is highly dependent on the assumption of a demographic model, i.e. a mathematical function describing the evolution of the size of a population over time. The choice of a model will determine the set of discrete parameters needed to estimate in order to accurately reconstruct of the population history of the sampled lineages, and is a crucial prerequisite. Five possible models have been described in the literature (see Fig. 2.11). They are, in order of increasing complexity: constant population size, exponential growth, logistic (exponential growth followed by constant population size) or expansion growth (constant population size followed by exponential growth), and finally piecewise con-exp-con (constant growths flanking an exponential growth phase). The details about these models, including the number and description of parameters involved, are shown in Table 2.5. These demographic models are hierarchically nested, allowing the performance of likelihood ratio tests in order to select the best fit for a given dataset.

Table 2.5. Demographic models and their respective parameters

Model	Number of parameters	Type of parameter **	Equation	References
Constant size	1	N_0	$N(t) = N_0$	*Pybus, 2000*
Exponential growth	2	N_0, r	$N(t) = N_0\, e^{-rt}$	*Pybus, 2000*
Logistic growth	3	N_0, r, c	$N(t) = N_0\, [(1 + c) / (1 + ce^{rt}]$	*Pybus, 2000*
Expansion growth	3	N_0, r, α	$N(t) = N_0\, [\alpha + ((1 - \alpha)\, e^{-rt})]$	*Pybus, 2000*
Constant-exponential-constant	5 *	N_0, r, α, x, y	$N(t) = N_0$, if $t < x$ $N(t) = N_0 e^{-r(t-x)}$, if $x < t < y$ $N(t) = Na$, if $t > y$	*Pybus, 2003*

* although five parameters are given, only four are needed to fully specify the model.

** N_0, population size at the present; r, exponential growth rate; c, logistic shape parameter; α, population size prior change, as a proportion of N0; x, end of exponential growth; y, beginning of exponential growth

Measurably evolving populations, i.e. populations from which molecular sequences can be collected at different time points in time with significant genetic differences (Drummond et al. 2003b), provide a particularly good framework for the application of the coalescent. Given the remarkably fast rate of evolution characterising HIV-1 and other RNA viruses, sequences from these organisms are particularly suitable for use with coalescent models. Moreover, since the importance of random genetic drift and neutral selection has been recently highlighted in the evolution of the virus (Leigh-Brown and Richman. 1997), the coalescent appears to be particularly suitable for the study of both between- and within-host dynamics of HIV-1.

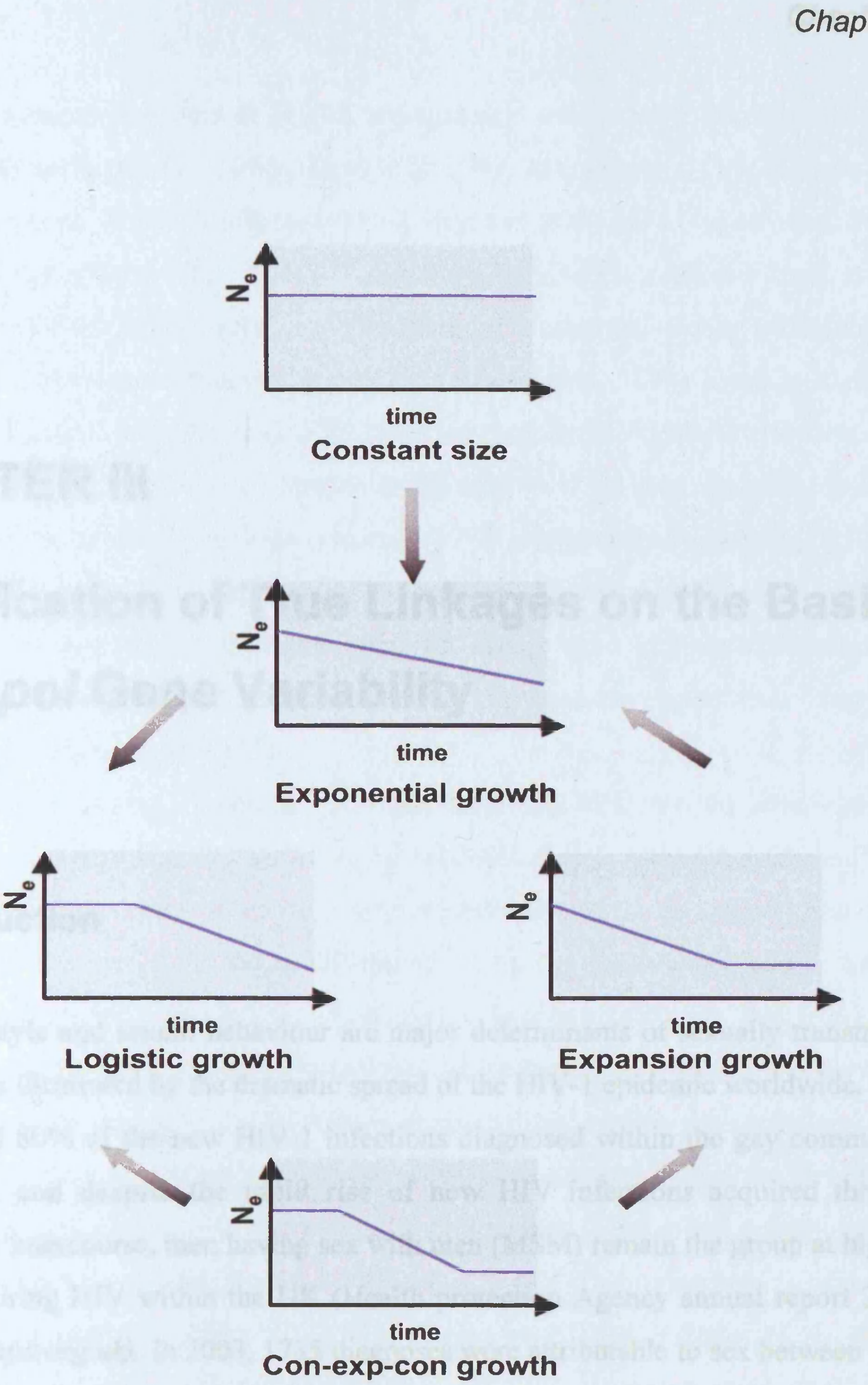


Fig. 2.11. Demographic models describing different patterns of population size evolution through time. Five demographic models are currently available, that is, in order of increasing complexity: constant growth, exponential growth, logistic and expansion growth, constant-exponential-constant growth. Each model is presented as the variation of the effective population size Ne over time. These models are nested, i.e. a model is a special case of another model in the direction indicated by

CHAPTER III

Identification of True Linkages on the Basis of the *pol* Gene Variability

1. Introduction

Lifestyle and sexual behaviour are major determinants of sexually transmitted infections, as illustrated by the dramatic spread of the HIV-1 epidemic worldwide. With an estimated 80% of the new HIV-1 infections diagnosed within the gay community since 1999, and despite the rapid rise of new HIV infections acquired through heterosexual intercourse, men having sex with men (MSM) remain the group at highest risk of acquiring HIV within the UK (Health protection Agency annual report 2003, *http://www.hpa.org.uk)*. In 2003, 1735 diagnoses were attributable to sex between men, increasing the incidence of HIV infections in the later risk group to 3.7% per year (Health Protection Agency annual report 2004, *http://www.hpa.org.uk/*). Despite this escalation in HIV incidence, transmission between risk-groups remains negligible, resulting in certain HIV-1 subtypes being associated with specific modes of transmissions (Tatt et al. 2001). Subtype B, for instance, remains the most prevalent clade within men having sex with men (MSM) in the western world. At the light of these considerations, the identification and characterisation of HIV-1 transmission networks amongst this risk group is of profound relevance for the public health in the UK, and the information held by molecular data is particularly suitable for that purpose.

Phylogenetic analyses of HIV-1 transmission events have been the focus of abundant studies (Balfe et al. 1990; Kleim et al. 1991; Arnold et al. 1995; Holmes et al. 1995; Leitner et al. 1996; Yirrell et al. 1997; Hayman et al. 2001; Taylor et al. 2003). With increasing availability of HIV-1 sequence data, such analyses have proved themselves notably useful in the reconstruction of transmissions, including the resolution of legal issues (Rogers et al. 1993; Albert et al. 1994; Birch et al. 2000; Machuca et al. 2001; Metzker et al. 2002). To that respect, HIV-1 sequences were used for the first time in 1991 to corroborate the infection of patients attending a dental surgery in Florida by their practician (Ou et al. 1992). Beyond the controversy it fuelled (Smith and Waterman. 1992; Debry et al. 1993; Hillis and Huelsenbeck. 1994; Crandall. 1995), the 'Florida dentist case' raised important issues to be considered before using HIV-1 gene sequences for the establishment of transmission chains (see Chapter I, section 3.2.2) (Holmes et al. 1993).

One of these considerations concerned the choice of a suitable genetic region. There is indeed a significant heterogeneity in nucleotide substitutions across the HIV-1 genome, resulting in unequal phylogenetic signals from gene to gene (Leitner and Albert. 1999; Korber et al. 2000). Of course, complete genome analysis is ideally applied to transmission studies. However, since there are relatively few full-length sequences available and phylogenetic analyses are restricted by the cost of sequencing appropriate background material as well as computational power, the sequence length and genetic region of choice need to be carefully considered together in order to guarantee the best estimate of phylogenetic relatedness. Most phylogenetic studies undertaken to date have relied on the V3 loop region of the *env* gene, taking advantage of its hypervariability (Balfe et al. 1990; Kleim et al. 1991; Chant et al. 1993). Alternatively, the entire *env* gene (Arnold et al. 1995; Hayman et al. 2001) and fragments of the *gag* gene (such as the p17 region) have been exploited, sometimes together (Arnold et al. 1995; Leitner et al. 1996; Leigh Brown et al. 1997; Hayman et al. 2001). Nonetheless it has been argued that fragments covering the V3 loop are too short or too variable to allow robust inferences on the genetic relatedness of specimens (Debry et al. 1993). Also, convergent evolution, i.e. genetic similarity in unrelated patients, has been reported in *env*, with the potential to bias phylogenetic inferences into false positives (Zhang et al. 1993). As for the *gag* gene, the limited number of sequences available in public databases makes it use problematic. By contrast, the region spanning the protease and RT genes is routinely sequenced in the clinical context of genotypic drug resistance testing and a large body of data is now being generated. Successful attempts to determine HIV-1 subtypes on the basis of the protease and the RT genes have been reported, so long as the fragment used is long and variable enough to counterbalance the lack of genetic constraint (Kessler et al. 2001; Pasquier et al. 2001; Yahi et al. 2001; Gale et al. 2004). However, the *pol* gene remains unpopular for phylogenetic analyses due to its extreme genetic conservation and the *pol* gene is commonly considered sub-optimal for the study of HIV-1 transmission histories (Albert et al. 1993; Palmer et al. 2002). Furthermore, an additional difficulty is encountered in the body of drug resistance related mutations when considering the *pol* region for the conduction of phylogenetic inference. It is indeed not infrequent that unrelated viruses harbour similar mutations associated with drug resistance after exposure to highly active antiretroviral treatment (HAART). Such convergent evolution could potentially bias the clustering of the viral sequences compared in the tree and lead to false relatedness between unrelated viruses.

The present study aimed to determine whether the *pol* gene holds sufficient genetic variability to allow the useful study of potential patterns of transmissions. For these purposes, potential linkages were identified within the ASRU *pol* sequence database, compared with clusters obtained from more variable genetic regions of HIV-1 (i.e. the *gag* and *env* genes), and the influence of drug resistance related mutations in the process of phylogenetic reconstruction was assessed.

The work presented in this chapter was published in Hué *et al.* 2004.

2. Material and Methods

2.1. Study Cohort

The *pol* sequences used for this study were extracted from the Health Protection Agency Antiviral Susceptibility Reference Unit database. There were generated from plasma samples collected from HIV-1 infected people in the United Kingdom between 1999 and 2003. For the purposes of the study, data were anonymised prior to analysis, according to the Health Protection Agency Ethics Committee's policy. Specific consent was nonetheless requested from patients appearing within transmission clusters in order to document potential sexual contacts, whilst blinding clinicians and patients to the laboratory data. Despite large-scale screening, epidemiological information was only obtained from a minority of patients.

2.2. Gene Amplification and Sequencing

pol variability

The region spanning the protease gene and the 235 first codons of the reverse transcriptase were amplified from plasma-derived viruses by random primed reverse transcription and nested PCR at the Antiviral Reference Unit of the Health Protection Agency, Heartlands Hospital, Birmingham. Details of the procedure are given in Chapter II, section 1.3.

gag and *env variability*

Where cDNA was available, regions spanning the *gag* and *env* genes were amplified and sequenced, as described in Chapter II, section 1.3. Thus, *gag* and *env* sequencing was undertaken for samples involved in clusters of *pol* sequences (n=23), sequential samples from the same individuals used as controls (n=6), and randomly selected samples where the pol gene was unrelated to other sequences (n = 23).

2.3. Genetic Distances

A minimum genetic diversity was expected between samples generated from patients involved in transmission networks. Hence, in order to characterise and identify sexual linkages amongst the nearly 2500 entries of the database, a pre-selection of closely related *pol* sequences was undertaken by computing pairwise genetic distances between all sequences, using the program Paup*. The genetic distances were calculated according to the general reversible time model with invariable sites and gamma distribution (GTR+I+G). The GTR model was selected over 57 alternative models of nucleotide substitution by likelihood ratio testing, using the software Modeltest and Paup* (see Chapter II, section 3.4). This model allows each possible substitution to have a different rate, with the constraint of being symmetrical, so that a substitution from a nucleotide *i* to *j* has to be the same as a substitution from *j* to *i*.

2.4. Phylogenetic Reconstruction

The general procedure followed for the phylogenetic reconstruction of the selected sequences is represented in Fig. 3.1. The methodology was consecutively applied to the *pol*, *gag* and *env* datasets. First, in-frame multiple alignments of the nucleotide sequences were constructed with the program ClustalX, with gap-opening and -extension penalties of 10 and 0.30 respectively, then manually adjusted using the editing software BioEdit. Sequences that could not be unambiguously aligned or were of insufficient length were excluded from the study. Phylogenetic relationships between the sequences were reconstructed using successively the neighbor joining (NJ) and maximum likelihood (ML) methods. The alignment matrices were imported into the tree building software Paup*, and an initial neighbor-joining tree was built under the Hasegawa-Kishino-Yang (HKY85) model of evolution with a ratio of transversion to transitions of 2:1. The best fitting model of nucleotide substitution was then estimated on the basis of the NJ tree topology, using a maximum likelihood ratio test to compare the different models implemented by Modeltest version 3.06. The parameters of the selected model of DNA substitution, together with the initial neighbor-joining tree, were finally used to perform a heuristic search for a ML tree. The trees were rooted against the corresponding region of an HIV-1 subtype K sequence (GenBank accession number AJ249239) extracted from the Los Alamos HIV-1 Database (*http://www.hiv-web.lanl.gov/*), and the robustness of the topologies was evaluated by bootstrap analysis, with 1000 rounds of replication. The models of nucleotide substitution used for the reconstruction of each ML trees are detailed in Table 3.1.

In order to assess the potential bias induced by drug resistance associated substitutions on the reconstruction of the samples relatedness, 46 codon positions known to be related to antiretroviral resistance (Shafer et al. 2000a) were then excluded from the previous *pol* sequences alignment and a maximum likelihood tree was implemented. Resistance mutation positions known as primary (or major) and secondary (or minor) were excluded. Primary mutations are known to lead to an alteration in drug binding by themselves, whereas secondary mutations do not have a significant effect on phenotype by themselves (D'aquila et al. 2003). The positions excluded from the *pol* alignment, together with the related drug resistance, are listed in Table 3.2. The phylogeny estimation, model testing and bootstrap procedures were performed with Paup*, as described above.

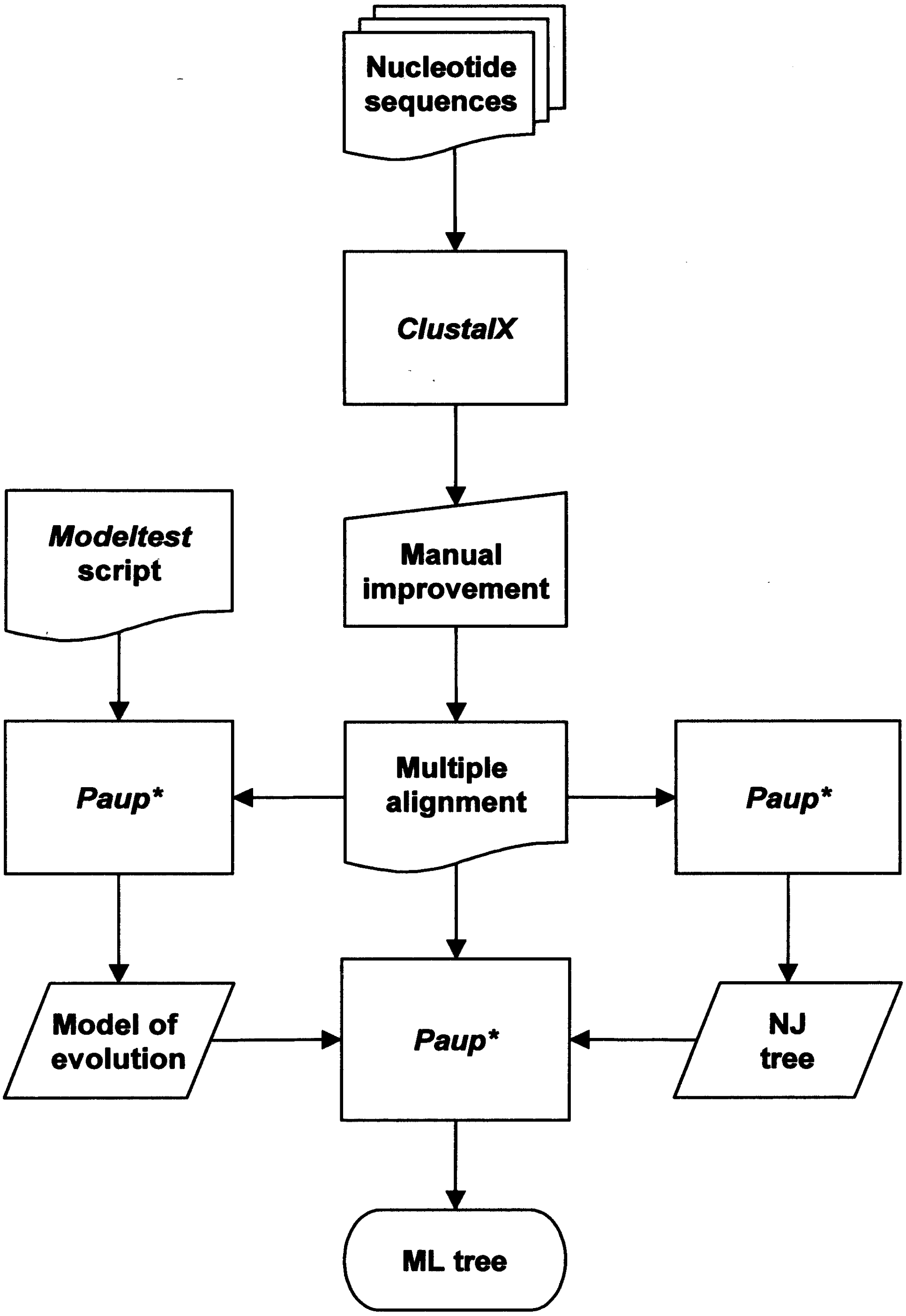


Fig. 3.1. Flow diagram summarising the methodology used for the reconstruction of phylogenetic trees. A multiple alignment of the HIV-1 *pol* sequences was first generated using the software ClustalX. After manual improvement on the sequence editor Bioedit, the alignment was then used to simultaneously select the model of evolution with the best fit to the data (with the software Modeltest) and construct an initial Neighbor joining tree. Both tasks were performed with the program Paup*. The multiple alignment, model and initial tree were finally used in order to generate a definite maximum likelihood tree in Paup*.

2.5. Sequence Data

The HIV-1 nucleotide sequences used in the present study were deposited into GenBank (*http://www.ncbi.nlm.nih.gov/)* under the accession numbers AY362043-AY362180, AY360862-AY360910 and AY360911-AY360959. The first, second and third series of accession numbers correspond to the *pol, gag* and *env* sequences alignments respectively.

Table 3.1 Models of evolution selected for the *pol, env* and *gag* datasets

	Sequence Alignment			
	pol	*pol-dr**	*gag*	*env*
*Model selected:***	GTM+I+G	GTM+I+G	GTM+I+G	GTM+I+G
- lnL =	13116.73	10355.26	6208.02	7903.65
Substitution model:				
[A-C] =	2.3700	1.9958	0.5737	1.5545
[A-G] =	8.9795	8.9787	1.6363	3.1150
[A-T] =	0.7548	0.7298	0.4937	0.7500
[C-G] =	1.5364	1.3601	0.2858	0.9318
[C-T] =	10.612	11.771	2.7136	2.3225
[G-T] =	1.0000	1.0000	1.0000	1.0000
Base frequencies:				
freqA =	0.4226	0.4227	0.3887	0.3742
freqC =	0.1422	0.1525	0.2277	0.1861
freqG =	0.1950	0.2005	0.2332	0.2016
freqT =	0.2402	0.2242	0.1505	0.2380
Proportion of invariable sites:	0.4766	0.5160	0.2005	0.2060
Gamma distribution shape parameter:	1.0441	1.0828	0.6520	0.9436

* *pol* alignment after exclusion of 46 codon positions associated with drug resistance

** *according to the Akaike Information criterion (AIC), as implemented in Modeltest 3.06*

Abbreviations: GTM, general time reversible model; +I, with invariable sites; +G, with gamma distribution; -lnL, log likelihood score; [i-j]: rate of substitution between base i and base j; freq i, frequency of base i in the dataset.

Table 3.2 Drug Resistance Mutations in HIV-1 (from d'Aquila *et al.* 2003)

Amino acid substitutions associated with resistance to					
Pi		NRTi		nNRTi	
Mutation	Prevalence in the data	Mutation	Prevalence in the data	Mutation	Prevalence in the data
L10F/V/I/R	15	**M41L**	**19**	**L100I**	**1**
K20M/R	11	E44D	2	**K103N**	**13**
L24I	0	**A62V**	**2**	**V106A/M**	**0**
D30N	**0**	**K65R**	**0**	**V108I**	**0**
V32I	0	**D67N**	**3**	**Y181C/I**	**2**
L33F	1	**T69D**	**4**	**Y188C/L/H**	**2**
M36I	28	**K70R**	**3**	**G190A/S**	**3**
M46I/L	**0**	**L74V**	**1**		
I47V	0	**V75I**	**0**		
G48V	**0**	**F77L**	**0**		
I50V/L	**0**	Y115F	0		
F53L	0	**F116Y**	**0**		
I54V/M/L	1	V118I	0		
L63P	68	**Q151M**	**0**		
A71V/T	21	**M184V/I**	**18**		
G73S/A	0	L210W	6		
V77I	22	**T215Y/F**	**15**		
V82A/F/T/S	**0**	K219Q/E	2		
I84V	**0**				
N88D/S	0				
L90M	**4**				

Primary mutations are indicated in bold

3. Results

Out of the 2500 *pol* sequences generated on samples dated from 1999-2003, 140 were selected on the basis of the closest pairwise genetic distances. Thus sequences sharing more than 95% similarity with one or more other entries from the database were selected for the study. Overall, the average inter-patient genetic variation amongst the sequences was 5.1% (range 0-12.4%). Although several subtypes were represented within the subset of sequences, including subtype A, B, C, D, G and CRF01-AE, the vast majority were of subtype B (88%), reflecting the subtype distribution of prevalent infections in the UK at the time of the study.

The maximum likelihood tree derived from the selected *pol* sequences is presented in Fig. 3.2. Twelve pairs or triplets of sequential sequences from a same patient were used as controls. Bootstrap values higher than 50% are indicated on the branches, reflecting the frequency with which a given branch occurred in 1000 bootstrap resampling. A total of 23 possible transmission clusters were identified from the tree topology. The criteria used to select these linkages were determined by plotting the supporting bootstrap score of each terminal cluster against the within-average branch length calculated from the ML tree topology (Fig. 3.3). Threshold criteria for the validation of the putative transmission chains were decided in light of the tree topology. Considering the extreme conservation of the *pol* gene, the bootstrap values supporting the branches were expected to be high, and the genetic distance between sequences involved in clusters to be low. Therefore, clusters were considered true linkage when fulfilling the following two conditions: 1) a bootstrap value equal or greater to 99%, and 2) an average genetic distance (i.e. branch length) lower than 0.015 nucleotide substitutions per sites within the cluster. There was no significant distinction between intra-patient (i.e. control) and inter-patient (i.e. linked) sequences in terms of genetic distance. All controls conformed to these criteria, with the exception of the multiple sequences belonging to patient 7 and 8, whose clusters were supported by lower bootstrap values (i.e. 95% and 92% respectively). The reason why these two clusters failed to fit in the criteria remains unclear. The relative low bootstrap score attributed to samples from patient 7 could be explained by the presence of an archive sequence, subsequently becoming the majority plasma population within the follow-up samples. For instance, a virus originating many years previously may emerge following a treatment interruption, resulting in a genetic distance between the two serial samples greater than expected. Unfortunately, matched *gag* and *env* sequences could not be generated for these samples due to PCR difficulties.

All putative transmission events involved subtype B viruses. Most of the 'non-B clades' of the ML tree were supported by high bootstrap scores, but since the bootstrap resampling process is known to be influenced by the number of taxonomic units considered in a tree (Felsenstein. 1985), the scores associated with these branches are likely to be artificially high due to under-representation within the dataset, and these clusters were excluded from the categorisation of potential clusters.

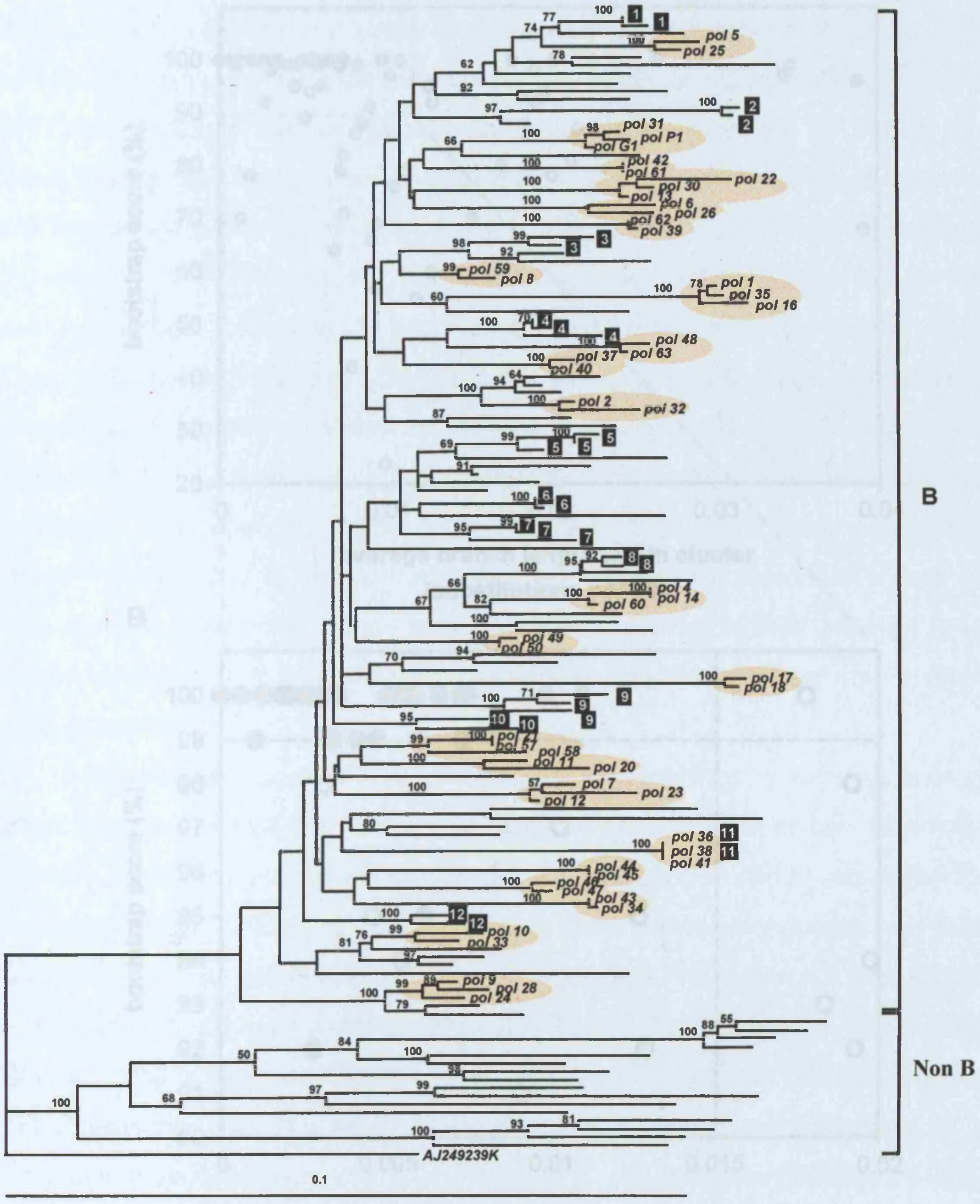


Fig. 3.2. Maximum likelihood tree representing the phylogenetic relationships between HIV *pol* sequences extracted from the HPA resistance database. The tree was constructed according to the GTR+I+G model of evolution and rooted against a HIV-1 subtype K sequence (AJ249239K) from the Los Alamos HIV database. Bootstrap values higher than 50% are indicated on the branches. Clusters involving potential transmission events are circled. Twelve pairs or triplets of multiple sequences from a same patient were used as control. These sequences are tagged by figures in black boxes (e.g. 1 indicates multiples sequences from patient 1). For clarity, only branches involved in possible linkages are labelled.

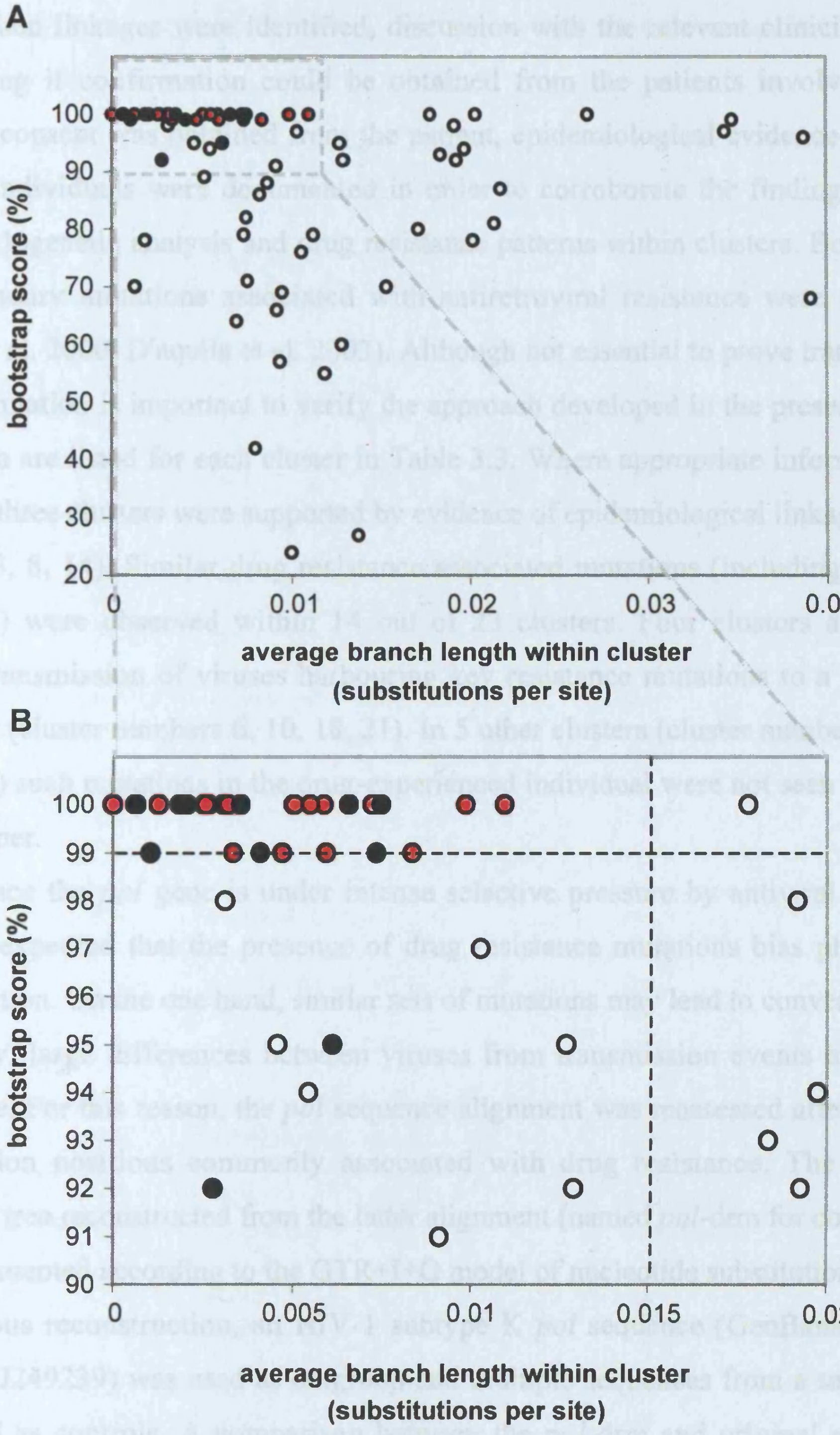


Fig. 3.3. (A) Average branch length within the terminal clusters of the maximum likelihood *pol* tree plotted against the bootstrap scores supporting the clusters. Possible transmission clusters and controls (i.e. clusters comprising intra-patient follow up sequences) are indicated by red and black dots respectively. (B) The cut-off values for the characterisation of linkages were a supporting bootstrap score higher than 99% and a mean genetic distance of 0.015 nucleotide substitutions per site.

When linkages were identified, discussion with the relevant clinicians helped determining if confirmation could be obtained from the patients involved. Where informed consent was obtained from the patient, epidemiological evidence of linkage between individuals were documented in order to corroborate the findings from the initial phylogenetic analysis and drug resistance patterns within clusters. Both primary and secondary mutations associated with antiretroviral resistance were considered (Hirsch et al. 2000; D'aquila et al. 2003). Although not essential to prove transmissions, such information is important to verify the approach developed in the present analysis. These data are listed for each cluster in Table 3.3. Where appropriate information was obtained, three clusters were supported by evidence of epidemiological linkage (clusters numbers 3, 8, 14). Similar drug resistance associated mutations (including secondary mutations) were observed within 14 out of 23 clusters. Four clusters appeared to identify transmission of viruses harbouring key resistance mutations to a drug naïve individual (cluster numbers 6, 10, 18, 21). In 5 other clusters (cluster numbers 1, 4, 11, 12 and 16) such mutations in the drug-experienced individual were not seen in the drug naïve partner.

Since the *pol* gene is under intense selective pressure by antiviral therapy, it might be expected that the presence of drug resistance mutations bias phylogenetic reconstruction. On the one hand, similar sets of mutations may lead to convergence, and conversely, large differences between viruses from transmission events may lead to divergence. For this reason, the *pol* sequence alignment was reassessed after exclusion of 46 codon positions commonly associated with drug resistance. The maximum likelihood tree reconstructed from the latter alignment (named *pol*-drm for convenience) was implemented according to the GTR+I+G model of nucleotide substitutions. As with the previous reconstruction, an HIV-1 subtype K *pol* sequence (GenBank accession number AJ249239) was used as outgroup and multiple sequences from a same patient were used as controls. A comparison between the *pol*-drm and original *pol* trees is shown in Fig. 3.4. Despite the deletion of 46 highly variable sites, the two topologies were congruent and the 23 putative transmission clusters identified within the *pol* tree were conserved in the *pol*-drm tree. Moreover, no additional clusters to those based on *pol* sequences were strongly supported by bootstrap scores. This suggests that mutations induced by antiretroviral therapy are unlikely to bias the reconstruction of transmission networks, and that unrelated virus harbouring identical drug resistance patterns are unlikely to cluster together within a phylogenetic *pol* tree, leading to false positives.

Table 3.3. Epidemiological and drug resistance mutation information for the 23 clusters of *pol* sequences.

Cluster	Sequence	Year of sampling	Drug history	Resistance associated mutations to		*gag/env*
				PIs	RTi	linkage
1	pol 5	2000	experienced	L10V, L63T	G190A	yes
	pol 25	2001	naïve	L10V, L63S	none	yes
2	pol 29	2001	naïve	None	none	n/a
	pol 31	2001	naïve	None	none	n/a
	pol P1	2001	naïve	None	none	n/a
3	pol 42	2001	naïve	L63P	None	n/a
	pol 61	2001	naïve	L63P	None	n/a
4	pol 13	2000	naïve	L10V†, M36I	none	yes
	pol 22	2001	experienced	L10V†, M36I	**M184V, Y188L**	yes
	pol 30	2001	naïve	L10V†, M36I	T69I†	yes
5	pol 6	2000	experienced	none	none	n/a
	pol 26	2001	naïve	none	none	n/a
6	pol 39	2002	experienced	M36L, L63P	**T69N**	n/a
	pol 62	2000	naïve	M36L, L63P	**T69N**	n/a
7	pol 8	2000	experienced	L63P	none	n/a
	pol 59	1999	experienced	L63P	none	n/a
8	pol 1	2000	experienced	L63T	none	yes
	pol 16	2001	naïve	L63T	none	yes
	pol 35	2002	naïve	L63T	none	n/a
9	pol 48	2001	naïve	L63H, A71V, V77I, I93L	none	n/a
	pol 63	2001	experienced	L63H, A71V, V77I, I93L	none	n/a
10	pol 37	2002	naïve	L63P	**M41L, T215Y**	yes
	pol 40	1998	experienced	L63P	**M41L, T215C**	yes
11	pol 2	2000	naïve	I93L	none	n/a
	pol 32	2001	experienced	I93L	**A62V, K65R, L74V, G190S**	n/a
12	pol 4	2000	naïve	L10V, I93L	none	yes
	pol 14	2000	naïve	L10V, I93L	none	yes
	pol 60	2001	experienced	L10V, L63P, A71V, I93L	K103N	n/a
13	pol 49	2000	naïve	M36I, L63P, I93L	none	n/a
	pol 50	2001	experienced	M36I, L63P, I93L	none	n/a
14	pol 17	2001	naïve	M36I, L63P, V77I, I93L	none	yes
	pol 18	2001	experienced	L63P, V77I, I93L	none	yes

Abreviations: PIs, protease inhibitors; RTIs, reverse transcriptase inhibitors;
Primary mutations are indicated in bold
† atypical mutation at the given codon

Table 3.3. (continued)

Cluster	Sequence	Year of sampling	Drug history	Resistance associated mutations to		*gag/env* linkage
				PIs	RTIs	
15	pol 21	2001	experienced	L10I, K20R, M36I, L63S, I93L	none	n/a
	pol 57	2001	experienced	L10I, L63C, I93L	none	n/a
	pol 58	2000	naïve	L10I, K20R, L63S, A71T, I93L	none	n/a
16	pol 11	2000	naïve	L10I, L63C, I93L	none	n/a
	pol 20	2001	experienced	L10I, L63C, I93L	**M41L, V118I, L210W, T215Y**	n/a
17	pol 7	2000	experienced	L10I, L63P, V73I, I93L	none	n/a
	pol 12	1998	naïve	L10I, L63P, V73I, I93L	none	n/a
	pol 23	2001	experienced	L10I, L63P, V73I, I93L	L210F	n/a
18	pol36	2001	naïve	K20R, M36I, L63A	**M41L, T215E †**	yes
	pol41	2001	naïve	K20R, M36I, L63A	**M41L, T215E †**	yes
19	pol 44	2002	experienced	M36I	**T215D**	yes
	pol 45	2002	experienced	M36I	**T215D**	yes
20	pol 46	2002	experienced	L63P	**T69A**	yes
	pol 47	2002	experienced	L63P	**T69A**	yes
21	pol 34	2002	experienced	L10V, L63P	**T215D**	yes
	pol 43	2000	naïve	L10V, L63P	**T215D**	yes
22	pol 10	2000	naïve	L63P, I93L	none	n/a
	pol 33	2001	naïve	L63P, I93L	none	n/a
23	pol 9	2001	naïve	L10I, L33I, L63T, A71T, I93L	A98S	yes
	pol 24	2000	naïve	L10I, L33I, L63T, A71T, I93L	A98S	yes
	pol 28	2000	naïve	L10I, L33I, L63T, A71T, I93L	A98S	n/a

Abbreviations: PIs, protease inhibitors; RTIs, reverse transcriptase inhibitors;
Primary mutations are indicated in bold
† atypical mutation at the given codon

Finally the relatedness of the sequences within an identified transmission cluster was further confirmed by constructing maximum likelihood trees based on the *env* and *gag* genes of the samples. A total of 49 sequences were used for the reconstruction of both *gag* and *env* trees, comprising 23 out of the 53 sequences involved in possible linkages (where stored samples or cDNA were available), coupled to 3 pairs of controls and 23 background unrelated sequences. The resulting *gag* and *env* alignment lengths were 747 base pairs and 557 base pairs respectively. The GTR+I+G model of molecular evolution was found to be the most appropriate for both datasets. Maximum likelihood

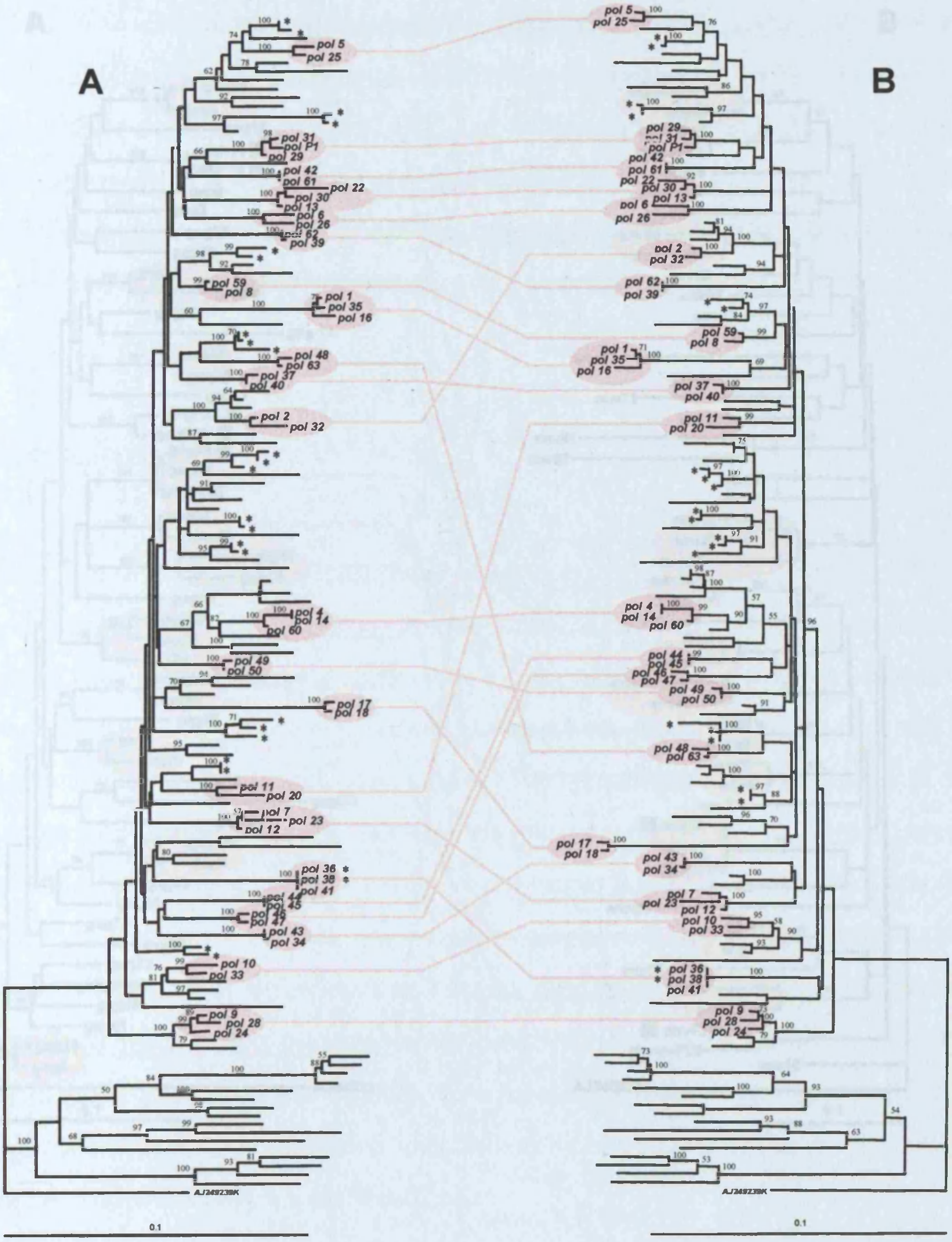


Fig. 3.4. Maximum likelihood trees constructed on the basis of *pol* sequences before (**A**) and after (**B**) exclusion of codon positions associated with drug resistance. The trees were reconstructed under the GTM+I+G model of evolution, and rooted against a HIV-1 subtype K sequence (AJ249239K). Transmission clusters are circled and controls (i.e. multiple sequences from a patient) are indicated by a star. Bootstrap values above 50% are indicated on the branches. Only branches involved in possible linkages are labelled.

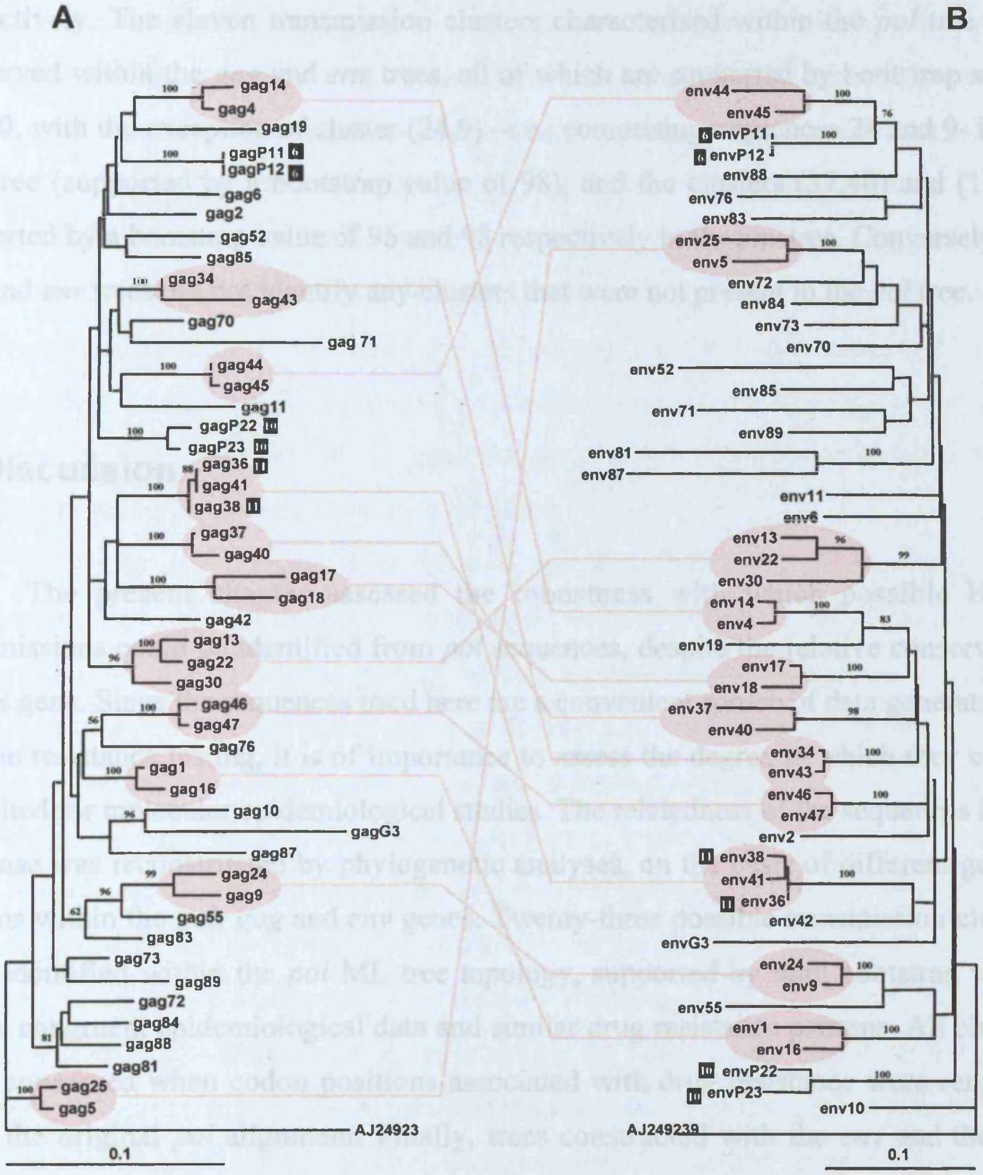


Fig. 3.5. Maximum likelihood trees derived from the *gag* (A) and *env* (B) regions of the samples, under the GTR+I+G model of nucleotide substitution. The trees are rooted against an HIV-1 subtype K sequence (AJ249239K) from the Los Alamos HIV database. Possible transmission clusters previously identified within the *pol* tree presented in Fig. 1 are circled in red. The pairs of multiple sequences from a same patient used as controls are indicated by figures in black boxes. Bootstrap values of 50% or greater are indicated on the trees.

trees constructed from the *gag* and *env* sequences are shown in Fig. 3.5A and 3.5B respectively. The eleven transmission clusters characterised within the *pol* tree were conserved within the *gag* and *env* trees, all of which are supported by bootstrap scores of 100, with the exception of cluster (24,9) –i.e.: comprising sequences 24 and 9- in the *gag* tree (supported by a bootstrap value of 98), and the clusters (37,40) and (13,22) supported by a bootstrap value of 96 and 98 respectively in the *env* tree. Conversely, the *gag* and *env* trees did not identify any clusters that were not present in the *pol* tree.

4. Discussion

The present chapter assessed the robustness with which possible HIV-1 transmissions could be identified from *pol* sequences, despite the relative conservation of this gene. Since the sequences used here are a convenient source of data generated for routine resistance testing, it is of importance to assess the degree to which they can be exploited for molecular epidemiological studies. The relatedness of the sequences in our database was reconstructed by phylogenetic analyses, on the basis of different genetic regions within the *pol, gag* and *env* genes. Twenty-three possible transmission clusters were identified within the *pol* ML tree topology, supported by high bootstrap values (>99), congruent epidemiological data and similar drug resistance patterns. All clusters were conserved when codon positions associated with drug resistance were removed from the original *pol* alignment. Finally, trees constructed with the *env* and the *gag* regions of the samples were consistent with the results obtained with the *pol* region and the same transmission clusters were identified.

It has been suggested that the *pol* gene is suboptimal for reconstructing transmission events (Palmer et al. 2002), since the genetic distance between protease and RT sequences from unrelated individuals may not always be significantly different from the distance between related individuals. The present study compared the topologies of tree obtained with three HIV-1 genes known to undergo distinctive evolutionary dynamics (i.e. *pol, gag* and *env*), *pol* having the lowest and *env* the highest rate of substitution (Li et al. 1988; Korber et al. 2000). Clustering patterns were identical within the three phylogenetic trees, with a similar range of statistical significance. Consequently, these results suggest that HIV-1 *pol* gene holds sufficient intrinsic genetic variability to permit the reconstruction of transmission histories by phylogenetic means. Whether or not phylogenetic relationships characterised from protease and RT sequences should be confirmed by more variable genetic regions of HIV-1 is open to debate. The present work clearly indicates that congruent results are obtained whichever of the three principal genes of HIV-1 are considered, the trees obtained only differing by the length of their branches and the clustering patterns of distant unrelated sequences. These findings could have an immediate consequence in the monitoring of HIV-1 epidemiology. In view of the preponderance of HIV *pol* sequence data consequent on routine HIV resistance genotypic testing, these sequences could also be utilised effectively to track the presence of transmission clusters within the communities from which there were obtained.

It is also worth noting that most of the sequences used for the study were generated from plasma samples obtained within a period of 3 years. The characterisation of transmission patterns within a group of HIV-1 infected individuals might be more problematic when using sequences collected over a longer time span, because of within-individual evolution. Indeed, we noted a greater than average genetic distance in *pol* from sequential samples taken from control patients number 7 and 8. Also, when based on a single genetic region, the interpretation of inferred linkage might be undermined by the presence of recombination in the genomes considered. A further concern relates to the bottleneck represented by transmission of a single, or narrow spectrum of virions, especially when appreciating that within-host compartmentalisation may lead to sexual transmission of genital rather than blood virus species (Taylor et al. 2003). Given that the maximum likelihood inference could not be performed on the whole data set, only sequences sharing more than 95% identity with at least one other sequence from the database were used. Such a pre-processing of the data could potentially have an impact on the results and favoured the presence of strongly supported clusters within the tree.

Although comparison with epidemiological data is important for the validation of the linkages characterised at the molecular level, this information remains hard to obtain and only 3 of the transmission clusters could be confirmed. This can mainly be attributed to the difficulty encountered when consent from the patients is requested. Furthermore, the presence of multiple sexual partners often compromises the characterisation of linkages between HIV-1 infected individuals and networks can be problematical to establish. It is important to distinguish between epidemiological and individual purposes for undertaking these analyses. It is essential that informed consent is obtained from individual patients prior to the potential identification of their source of infection, and that appropriate security is afforded to HIV-1 sequence databases.

Finally, a number of instances of transmitted drug resistance through this analysis were identified, as described elsewhere (Pillay et al. 2000b; Ammaranond et al. 2003; Taylor et al. 2003). It is self evident that the presence of key mutations themselves is insufficient to virologically prove transmission. It could be suggested that the *pol* gene sequence, itself generated for purposes of resistance testing, is adequate for such phylogenetic studies.

CHAPTER IV

Correlates of Sexual Risk and HIV-1 Transmission during Primary Infection

1. Introduction

The natural progression of an HIV-1 infection traditionally begins with an acute, (or primary) phase, followed by an early clinical latent phase (spanning 3 to 10 years), ultimately followed by the onset of AIDS. The acute phase of the infection is conventionally described as the interval during which HIV epitopes can be detected in blood serum and plasma before the production of specific antibodies, which occurs approximately 30 days after infection. Nonetheless, when looking at the incidence of HIV-1 infections, it is difficult to distinguish between new diagnoses of chronic infections and recently acquired infections, and the infectiousness of acutely infected individuals, while of major relevance for public health, remains difficult to assess.

High plasma and genital tract viral load (VL), viral tropism, host susceptibility and opportunistic sexually transmitted infections (STIs) are amongst the clinical factors believed to increase HIV-1 infectiousness (Blaak et al. 1998; Kaufmann et al. 1998; Pesenti et al. 1999; Vernazza et al. 1999; Wahl et al. 1999; Pilcher et al. 2004b). Susceptibility to HIV-1 seems also to be influenced by genetic factors such as HLA type, co-receptor type, and/or gender (Long et al. 2000; Ray and Quinn. 2000; Glynn et al. 2001; Al Jabri. 2002; Tang et al. 2002; Trachtenberg et al. 2003; Koning et al. 2004;

Quayle et al. 2004). Moreover, it has been hypothesised that HIV infected persons may be more infectious at the early stages of the disease (Yerly et al. 2001b). Both mathematical modelling and empirical epidemiological data seem to support this hypothesis. Hence, individuals with primary HIV-1 infections (PHIs) are suspected to be up to 1000 times more infectious than during any other stage of the disease (Koopman et al. 1997). Also, Jacquez *et al* estimated that between 25 and 47% of new homosexually acquired infections may be transmitted during the 2 first months of infection (Jacquez et al. 1994). Furthermore, it has been suggested that sexual risk behaviour during and after HIV seroconversion has a significant impact on the spread of the epidemic. Since high infectivity may precede symptoms in primary infection (Kahn and Walker. 1998; Pilcher et al. 2001), HIV-1 infected individuals may be unaware of the risk they expose partners. High-risk sexual intercourse amongst acutely infected men having sex with men have indeed been reported, involving alarming rates of partner changes, sexual concurrency and unprotected anal intercourses (Colfax et al. 2002), suggesting that primary infections play a more important role in transmission from casual partners than in transmission from steady partners (Xiridou et al. 2004). As a consequence, rates of transmission of resistant HIV strains, which compromise treatment success, are up to 20% in many countries (Little et al. 2002), including the UK (Pillay et al. 2000). It is therefore of clinical and epidemiological relevance to efficiently identify newly acquired infections and to measure the rate of transmission amongst primary infected individuals.

Besides, it has been suggested that risky sexual behaviour such as high rates of partner change and concurrent partnership, coupled with high infectivity, promote the emergence of superinfection (i.e. re-exposure with HIV-1 after an initial infection) (Blackard and Mayer. 2004). As a consequence, recurrent exposure to HIV amongst seropositive individuals through high-risk behaviour increases the likelihood of recombination (Fang et al. 2004), with implications for public health (such as the emergence of multi-drug resistant recombination forms). Numerous molecular tools enable a reliable categorization of mosaic genomes (Posada. 2002) and the identification of recombinant forms is relatively straightforward when the parental viruses belong to distinct clades of the same HIV-1 group, or are even more distantly related. However, detection of recombination is more challenging amongst viruses of the same clade, and reports of intrasubtype HIV-1 recombination are rare in the literature (De Baar et al. 2003; Pollakis et al. 2003). If partial *env* and *gag* gene sequences are traditionally used for the characterisation of recombinant form when full-length sequence is not available, the *pol* gene (partially or in its integrity) offers a good alternative for subtype assignment since it has become increasingly available through routine resistance testing. Its reliability for phylogenetic reconstructions has been recently shown (Hué et al. 2004), and, with all but two circulating recombination forms exhibiting break points in the protease and RT genes (Kuiken et al. 2002), the *pol* region is now considered as adequate for subtype classification (Barlow et al. 2001; Yahi et al. 2001; Pandrea et al. 2002; Njouom et al. 2003). Nonetheless, the *pol* gene alone has rarely been utilized for recombination analyses to date.

In order to understand further the impact of primary infections on the spread of the HIV-1 epidemic, molecular and epidemiological analyses of PHI was undertaken within a geographically discrete area of the UK, with a focus on newly infected MSM. Potential transmission clusters were identified by phylogenetic means and related to clinical and epidemiological data, in order to identify significant determinants of the HIV transmission at early stages of the disease. When large networks of transmission were characterised, viral genomes isolated from the individuals involved were tested for intra-subtype recombination, under the assumption that super-infection might have occurred.

The work presented in this chapter was published in Pao *et al.* 2005.

2. Material and Methods

2.1. Study Cohort

Subjects with primary HIV-1 infection (PHI) were recruited from a cohort of 1235 HIV-positive individuals attending a Genitourinary Medicine Unit in Brighton, UK, for follow up between 1999 and 2003 (the department is the unique local provider for HIV and STI care). Of these, 86% were Caucasian and were 89% are male. The predominant route of infection within the cohort was sex between men (79%). National surveillance data confirms that over 90% of the diagnosed HIV infected patients resident in the area attend this clinic. Primary HIV infections were identified by at least one of the following tests:

- Previous negative HIV antibody test within 18 months
- Evolving Western Blot or HIV antibody response.
- Positive HIV-1 antibody test in association with a negative "detuned" HIV antibody assay (suggestive of infection within the previous 4-6 months)

The western blot test detects antibodies to specific denaturated HIV-1 proteins, including the core (p17, p24, and p55), polymerase (p31, p51, p66), and envelope (gp41, gp120, gp160) proteins (Carlson et al. 1985; Schwartz et al. 1988). The test is considered negative in the absence of all bands, and positive if reactivity is detected to gp41 and gp120/160 env bands or to either of these env bands plus the p24 gag band. The presence of any bands that do not meet the criteria for a positive result is considered an indeterminate result.

A detuned assay consists in an enzyme-linked immunosorbent assay (ELISA) test of low sensitivity following a standard one, in order to distinguish patients who have seroconverted within the past 129 days from patients who seroconverted sometimes beyond this point (Janssen et al. 1998; Mcfarland et al. 1999). The assay takes advantage of the progressive increase in HIV antibody titre during the initial phase of infection: a subject recently infected will have lower antibody level and will test negative on the less sensitive ELISA. Detuned assays were performed using the bioMérieux Vironostika HIV-1 assay (bioMérieux UK Ltd., Basingstoke, UK) as previously described (Kothe et al. 2003).

In total, 103 subjects with PHI diagnosed between 1999 and 2003 consented to the study and were included in the following epidemiological and phylogenetic analyses.

2.2. Epidemiological and Clinical Data

Markers reflecting the subjects' clinical status were recorded at the clinics for each patient taking part in the study. These included CD4 cell count, CD4 percentage, HIV-1 viral load (VL), as well as the presence or absence of PHI symptoms. Measurements of HIV-1 viral load, or level of plasma viral RNA, reflects the cumulative production of virions from the various cellular reservoirs and turnover of virus-producing cells in those reservoirs. Additional epidemiological information was obtained directly from the patients, such as the HIV acquisition risk group and details of the sexual behaviour (i.e. estimated frequency and nature of sexual contacts within 3 months prior to PHI diagnosis). In addition, the presence and nature of sexually transmitted infections (STIs) within 3 months prior to HIV-1 diagnosis was recorded. Prevalence of gonorrhoea, chlamydia, non-specific urethritis, early syphilis and genital ulcer diseases was noted. Confidentiality and anonymity of the patients were protected by irreversibly unlinking clinical and laboratory identification numbers using a firewall system managed by the local Public Health Laboratory. The Brighton and Hove Local Research Ethics Committee and the Health Protection Agency Ethics Committee approved the present study. Written, informed consent was obtained from all participants.

2.3. Statistical Analyses

As appropriate, chi square tests or Fisher's exact test were used to determine the significance of epidemiological and clinical differences across individuals involved or not in putative transmission networks. The chi square test is a non-parametric test of statistical significance, allowing the estimation of the degree of confidence one can have in accepting or rejecting a null hypothesis (Siegel and Castellan Jr. 1988). Typically, the hypothesis tested is whether or not two different samples are different enough in some characteristic to allow a generalization on the populations from which the samples are drawn (here, subjects involved or not in transmission clusters). The chi square test returns a value that has to be compared to critical values of chi square distributions for the appropriate degrees of freedom and the chosen probability of error threshold (e.g., $p < 0.05$). If the returned chi square value is larger than the critical value, the data present a statistically significant relationship between the tested variables. Since the null hypothesis H_0 conventionally states that the relation across the data does not exist, the relationship does exist if H_0 is rejected. In our case, the null hypothesis was that clinical factors found in both linked and unlinked individuals do significantly differ. The Fisher exact test of significance is used in place of the chi-square test in small datasets (Siegel and Castellan Jr. 1988). It tests the probability of getting the observed data simply by chance. By convention, the Fisher exact test is computed when 5 or less values are to be tested by category, or when the total sample size is inferior to 20. Multivariate logistic regressions were used to identify factors independently associated with belonging to a transmission cluster. Logistic regressions are traditionally used to estimate the probability of a certain event occurring, allowing the assessment of interaction between variables, and yield odds ratios. These are calculated by dividing the odds in the group of interest by the odds in the control group. An odds ratio of 1 implies that the event is equally likely in both groups. An odds ratio greater than one implies that the event is more likely in the first group. An odds ratio less than one implies that the event is less likely in the first group.

All statistical analyses were performed by Professor Caroline Sabin, Department of Primary Care and Population Sciences, Royal free & University College Medical School, London, UK, using the software SAS version 8.

2.4. Recombination Analyses

Two sets of analyses were conducted in parallel on transmission clusters involving four sequences of more. On one hand, phylogenetic networks were generated using the split decomposition method implemented in the software Splitstree (Huson. 1998). Phylogenetic networks originated from the idea that a tree may be an inappropriate evolutionary model when conflicting signals of relatedness are encountered. By enabling multiple ancestries for a single taxon, networks are particularly appropriate for the visualisation of complex patterns of evolution such as recombination, where different genome partitions may support different phylogeny. Networks are constructed on the basis of genetic distance matrices, from which the branching order is determined, using the split decomposition method developed by Bandelt (Bandelt and Dress. 1992). A simplified illustration of how a split decomposition network is implemented is given in Fig 4.1. Following this procedure, the network has a tree-shape, with a unique parent (i.e. internal node) for each group of descendant, if no conflictual phylogenetic signal is encountered (Fig 4.1b). With less ideal data, the algorithm yields a network that can be interpreted as possible evidence for different and conflicting phylogenies (Fig. 4.1c).

On the other hand, bootscanning analyses (Salminen et al. 1995) were conducted using the software Simplot version 2.5 (*http://sray.med.som.jhmi.edu/RaySoft/Simplot/).* When performing such analyses, a query sequence is compared to an alignment of reference sequences (here, *pol* gene sequences) in a sliding-window fashion, i.e. for a successive set of overlapping sub-regions of the alignment. Within each window, the phylogenetic relationship between all sequences (including the query) is determined using bootstrap resampling. That is a NJ tree is calculated for the stretch of genome spanning each window, and the bootstrap value of the phylogenetic cluster including the query sequence in each tree is plotted along the genome as a XY plot, where the X axis represents the bootstrap values and the Y axis the genome position at the midpoint of each window. If the query sequence happen to be a recombinant form of two or more of the references sequences, a progressive switch of the highest bootstrap value from one reference to another will be observed at the recombination break points.

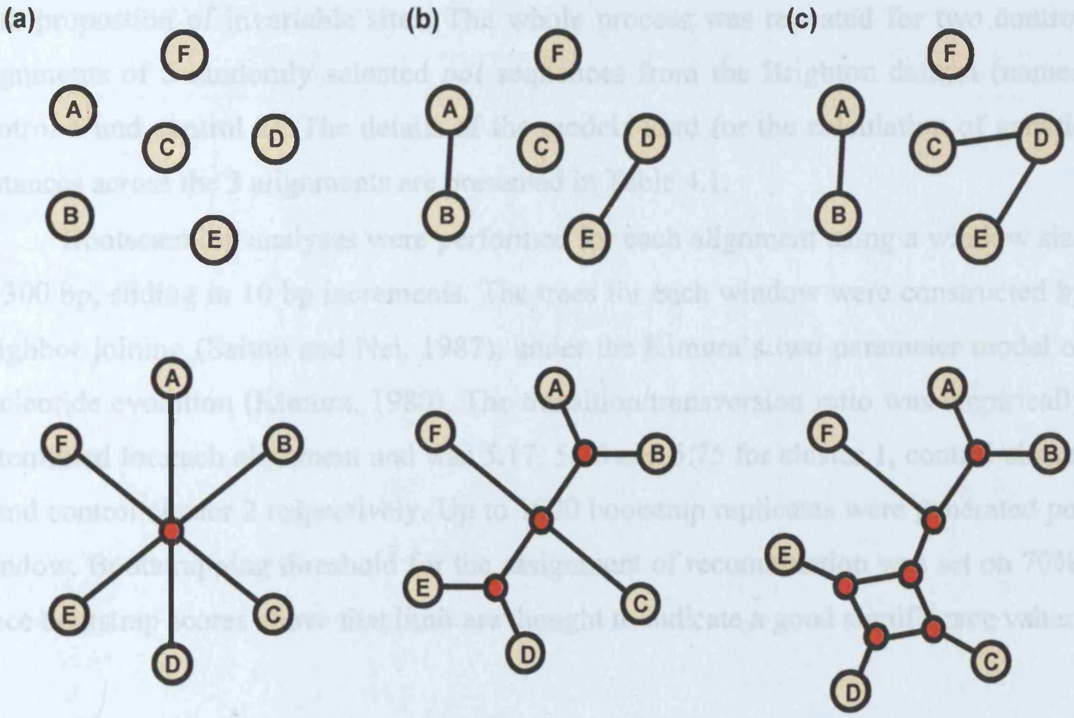


Fig. 4.1 Construction of a phylogenetic network. (a) Lets imagine five unlinked taxa. A star-like tree is used at a starting topology. (b) After calculation of the genetic distances between all taxa, A and B are found to be closely related and are linked together. The process is then reiterated for D and E. (C) As D appears to be closely related to C as well as E, the three taxa are linked by a network. *After Bryant and Moulton, 2004.*

Since a minimum of four sequences are needed to perform a bootscan test, the search for intrasubtype recombination within the phylogeny concerned cluster 1 and 6 only, which include 5 and 4 sequences respectively (see Fig. 4.2). Strictly speaking, cluster 6 involves 3 distinct patients only, two sequences being follow-up samples from the same individual. Moreover, 3 out of the total 4 sequences were fully identical, and differed from the forth one by 3 synonymous changes only. Under these conditions, the phylogenetic signal was too discrete to allow the analysis, and cluster 1 alone was investigated for intrasubtype recombination. The 5 sequences were manually aligned using the sequence editor Bioedit and the alignment translated into file formats accepted by Splitstree (i.e. modified nexus file) and Simplot (i.e. Phylip file). Genetic distances across the sequences were calculated in Paup* within a maximum likelihood framework, according to the general time reversible model of nucleotide substitution, with proportion of invariable sites. The whole process was repeated for two control alignments of 5 randomly selected *pol* sequences from the Brighton dataset (named control 1 and control 2). The details of the models used for the calculation of genetic distances across the 3 alignments are presented in Table 4.1.

Bootscanning analyses were performed for each alignment using a window size of 300 bp, sliding in 10 bp increments. The trees for each window were constructed by neighbor-joining (Saitou and Nei. 1987), under the Kimura's two parameter model of nucleotide evolution (Kimura. 1980). The transition/transversion ratio was empirically determined for each alignment and was 5.17, 5.76 and 5.75 for cluster 1, control cluster 1 and control cluster 2 respectively. Up to 1000 bootstrap replicates were generated per window. Bootstrapping threshold for the assignment of recombination was set on 70%, since bootstrap scores above that limit are thought to indicate a good significance value.

3. Results

3.1. Epidemiological Data

Amongst the 103 subjects who consented to the study, 73 (71%) had a detuned HIV antibody assay suggestive of infection 4 to 6 months prior testing. A total of 99 (96.1%) were male, and the age of the cohort ranged from 21 to 67 years, with a median age of 36 years. In terms of risk groups, 90 of the 99 males were MSM (90.9%), while 6 of the subjects (5.8%) reported a history of intravenous drug use (2 MSM, 2 heterosexual males and 2 heterosexual females). Where information was available, STIs were diagnosed concurrently with PHI in 34 of 89 (34.3%) individuals. When reporting sexual practices, 61 out of the 90 MSM (68%) mentioned unprotected anal intercourse in the 3 months prior to diagnosis. No information was available regarding sexual practices in the period preceding that time point. The CD4 count was available for 101 out of 103 patients and had an average value of 526 (range 195-1477) cells per ml. Median HIV viral plasma load was log 4.95 (range 2.03-6.00) copies per ml.

Table 4.1 Models of nucleotide substitutions for the Brighton datasets

	Datasets			
	Global ML tree	*Cluster 1*	*Control 1*	*Control 2*
Model selected: *	GTR+I+G	GTR+I	GTR+I	GTR+G
- lnL =	12876.04	1416.01	1943.52	1878.7827
Substitution model:				
[A-C] =	2.5800	1.0000	1.0000	1.0000
[A-G] =	9.9856	5.9039	8.4111	4.7650
[A-T] =	0.6978	1.0000	1.0000	0.2431
[C-G] =	1.3455	1.0000	1.0000	0.2431
[C-T] =	12.783	34.405	7.4351	4.7650
[G-T] =	1.0000	1.0000	1.0000	1.0000
Base frequencies:				
freqA =	0.4158	0.3947	0.3970	0.3939
freqC =	0.1489	0.1551	0.1621	0.1583
freqG =	0.1939	0.2078	0.2042	0.2077
freqT =	0.2414	0.2424	0.2367	0.2402
Proportion of invariable sites:	0.4807	0.8942	0.6552	-
Gamma distribution shape parameter:	1.0604	-	-	0.1255

* estimated according to the Akaike Information criterion (AIC), as implemented in Modeltest 3.06

Abbreviations: ML, maximum likelihood; GTM, general time reversible model; F81, Felsenstein 81 model; +I, with invariable sites; +G, with gamma distribution; -lnL, log likelihood; [i-j]: rate of substitution between base i and base j; freq i, frequency of base i in the dataset.

A large majority of the strains were of subtype B (n=96), non-B isolates being distributed as follow: 2 subtype C, 1 subtype A, 2 subtype G, and 3 recombinant circulating forms (i.e. 1 CRF01_AE and 2 CRF02_AG). Finally, 13 of the 103 *pol* sequences used in the study (12.5%) harboured primary antiretroviral resistance-associated mutations. The epidemiological data across the study cohort are detailed in Table 4.2.

3.2. Phylogenetic analyses

From the topology of the ML phylogenetic tree presented in Fig 4.2, viruses from 35 of 103 individuals (34%) grouped into 15 distinct transmission chains. As detailed in Chapter III, section 3, the criteria used to select the putative transmission chains were arbitrary determined from the tree topology. Thresholds for the validation of true linkages were decided by plotting the bootstrap scores supporting each terminal cluster the within-average branch. Clusters were indicative of true linkages when fulfilling the following two conditions: 1) a bootstrap value equal or greater to 99%, and 2) an average genetic distance (i.e. branch length) lower than 0.015 nucleotide substitutions per sites within the cluster (see Chapter IV).

3.3. Statistical Analyses

The 15 possible linkages comprised 1 cluster of 5 individuals, 2 of 3 and 12 of 2. All transmission chains involved male patients, 32 of which (97%) were MSM. For individuals within 11 out of 15 of the clusters, PHI was diagnosed within 12 months of each other, giving supporting evidence that transmission occurred during the primary phase of the infection. When comparing clinical data between linked (n=35) and unlinked (n=68) individuals, patients involved in transmission clusters were younger, with a median age of 34 years compared to 37 years amongst unlinked subjects (p=0.05). As for the informed sexual practices, linked individuals reported a greater number of different sexual contacts within 3 months prior census, with an average of 3 sexual partners against 2 partners amongst unlinked subjects (p=0.006), and were more likely to have engaged unprotected anal intercourses in the previous 3 months (87.5 vs. 65.3%; p=0.05). In addition, high rates of STIs at the time of PHI diagnosis were observed in both groups with a trend of higher prevalence in linked individuals (42.9%

Table 4.2 Epidemiological and clinical data on the study cohort (n = 104)

Category		n	%
Sex of the patients:	Male	100	96.2
	Female	4	3.7
Age within the cohort:	Median	36	n/a
	Range	(21-67)	n/a
Risk group:	Homosexual	91	87.5
	Heterosexual	11	10.6
	IDU	6	5.8
	Other	4	3.8
	Not known	2	1.9
Number of sexual activity within 3 months prior to diagnosis of PHI:	1	29	27.9
	2	20	19.2
	3 to 5	20	19.2
	6 to 10	10	9.6
	>10	12	11.5
	Not known	13	12.9
Reported sexual intercourses within 3 months prior to diagnosis of PHI:	Unprotected oral	62	59.6
	Unprotected vaginal	8	7.7
	Unprotected anal	61	58.7
	Protected anal	7	6.7
	Other	1	1.0
	Not known	22	21.2
STD within 3 months prior to diagnosis of PHI:	None	56	53.8
	1	23	22.1
	2	7	6.7
	3	4	3.8
	Not known	14	13.5
Symptoms at time of infection:	Yes	38	36.5
	No	37	35.6
	Not known	29	27.9
Clinical markers at diagnosis:	CD4 count	526.00	n/a
	CD4 %	28.50	n/a
	viral load	4.92	n/a

Abbreviations :IDU,injecting drug user; PHI, primary HIV-1 infection; STD, sexualy transmitted disease.

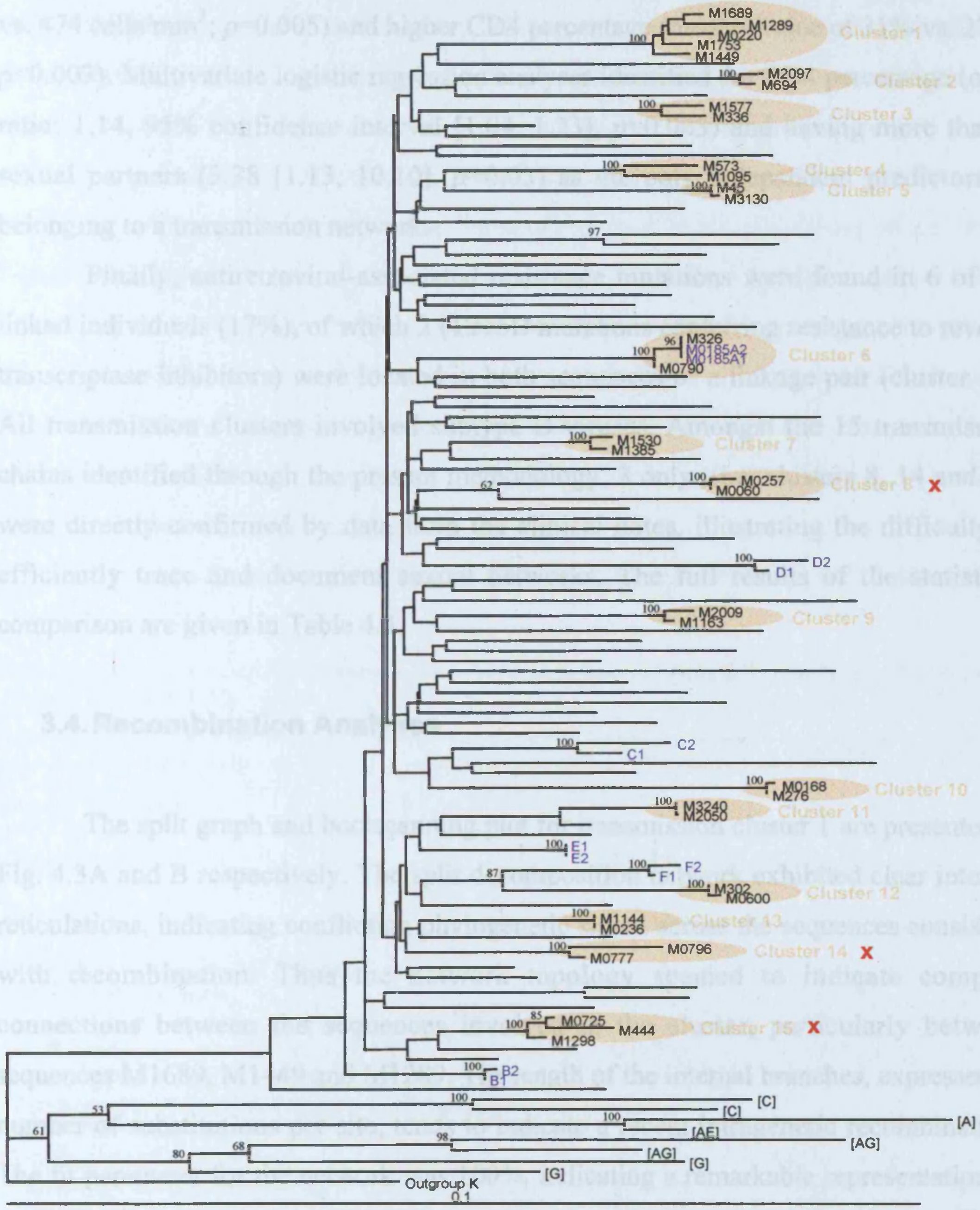


Fig. 4.2 Maximum likelihood tree representing the phylogenetic relationships between HIV-1 *pol* sequences from the Brighton dataset. The tree was constructed according to the GTR+I+G model of evolution and rooted against a HIV-1 subtype K sequence (AJ249239K), extracted from the Los Alamos HIV database. Bootstrap values higher than 50% are indicated on the branches. Six pairs of follow-up sequences from the same individuals were used as controls of relatedness. These are indicated in blue, following a letter code, i.e. sample C1 and C2 correspond to the first and second time point respectively sampled from patient C.

vs. 27.9%; p=0.13). Linked patients also had higher CD4 counts (median value of 612 vs. 474 cells/mm^3; p=0.005) and higher CD4 percentage (median value of 31% vs. 27%; p=0.003). Multivariate logistic regression analyses identified the CD4 percentage (odds ratio: 1.14, 95% confidence interval [1.04, 1.23], p=0.003) and having more than 5 sexual partners (3.38 [1.13, 10.10], p=0.03) as the only independent predictors of belonging to a transmission network.

Finally, antiretroviral-associated resistance mutations were found in 6 of the linked individuals (17%), of which 2 (T215D mutations conferring resistance to reverse transcriptase inhibitors) were located in both sequences of a linkage pair (cluster 11). All transmission clusters involved subtype B viruses. Amongst the 15 transmission chains identified through the present methodology, 3 only (i.e. clusters 8, 14 and 15) were directly confirmed by data from the clinical notes, illustrating the difficulty to efficiently trace and document sexual networks. The full results of the statistical comparison are given in Table 4.3.

3.4. Recombination Analyses

The split graph and bootscanning plot for transmission cluster 1 are presented in Fig. 4.3A and B respectively. The split decomposition network exhibited clear internal reticulations, indicating conflicting phylogenetic signal across the sequences consistent with recombination. Thus the network topology seemed to indicate complex connections between the sequences involved in the cluster, particularly between sequences M1689, M1449 and M1289. The length of the internal branches, expressed as number of substitutions per site, tends to indicate a recent intragenetic recombination. The fit parameter for the network was 100%, indicating a remarkable representation of the data's phylogenetic signal by the split graph. Evidence of potential recombination were also found in the bootscanning plot when comparing sequence M1689 to the other taxa of the cluster. The sequence clustered with sequence M1449 in the 5' region of the gene, and with sequence M1289 in the 3' end, with a clear-cut switch around nucleotide position 640. The presence of such a break point is traditionally regarded an evidence for recombination. Traditionally, the bootstrap threshold for the identification of recombination is set to 70% when performing bootscanning analyses (Salminen et al. 1995). Despite a bootstrap score significantly higher for cluster M1689-M1289 compared to the other clusters of the phylogeny, the bootstrap values for this cluster barely reached the cut-off value of 70%. By contrast, values for cluster M1689-M1289 ranged from 75% to 85%. The mosaic profile of sequence M1689 was corroborated by the corrected genetic distance *d* calculated between the three sequences up- and downstream the potential breakpoint (i.e. nucleotide position 640). Thus, sequence M1689 shared higher similarity with M1449 ($d = 0.00697$ substitutions/sites) than with M1289 ($d = 0.01079$ subs/sites) in the 5'end of the gene. Inversely, the sequence shared higher similarity with M1289 ($d = 0.00652$ subs/sites) than with M1449 ($d = 0.01100$ subs/sites) in the 3'end of the gene. Distances were corrected according to the general time reversible model of nucleotide substitution, with proportion of invariable sites (see Table 4.1). No discriminatory patterns were found when looking at drug resistance-related mutations held within cluster 1's sequences, all of which harbored the same polymorphisms in both protease and RT genes.

In comparison, the patterns exhibited by control clusters 1 and 2 were more

Table 4.3 Comparision of epidemiological data between linked and unlinked patients (significant associations are indicated in bold)

	Linked	Unlinked	*p* value*
Number of patients	35	68	-
Male gender	35 (100%)	64 (94.1%)	0.29
Age: median (range)	**34 (23,54)**	**37 (21,67)**	**0.05**
Number of contacts within 3 months prior to PHI diagnosis:	**3 (1, 100)**	**2 (1, 36)**	**0.006**
MSM risk:	32 (97%)	58 (85%)	0.09
Higher reported risk in 3 months prior to PHI diagnosis:			
Unprotected oral intercourse	25 (78%)	36 (73.5%)	0.83
Unprotected anal intercourse	**28 (87.5%)**	**32 (65.3%)**	**0.05**
Protected anal intercourse	2 (6.3%)	5 (10.2%)	0.70
STIs within 3 months prior to PHI diagnosis:			
Yes	15 (42.9%)	19 (27.9%)	0.33
No	18 (51.4%)	37 (54.4%)	0.33
Not known	2 (5.7%)	12 (17.7%)	0.33
Clinical markers at diagnosis:			
CD4 count: median (range)	**612 (195, 1477)**	**474 (196, 1259)**	**0.005**
CD4%: median (range)	**31 (12, 40)**	**27 (7, 42)**	**0.003**
Viral load: median (range)	4.97 (2.03, 6.00)	4.94 (2.30, 6.00)	0.70

Abbreviations: PHI, primary HIV-1 infection; MSM,men having sex withmen; STI, sexually transmitted disease ambiguous, as shown in Fig. 4.4 and 4.5 respectively. Both split graphs presented minor reticulations at the centre of a star-like topology, with terminal branches significantly longer than the internal splits. These patterns are more likely to be an artefact induced by the conserved nature of the data than an evidence for true recombination. The reticulations seen in the split-trees could be, for instance, the result of insufficient correction for multiple hits, or insufficient mutation patterns. Neither precise breakpoints, nor bootstrap value above the traditional cut-off limit of 70% was apparent within the bootscan plots, reinforcing the idea that the reticulations at the centre of the networks result from intrinsic phylogenetic 'noise' rather than from true mosaic patterns.

4. Discussion

The present analysis aimed to describe via molecular and epidemiological means HIV-1 primary infections amongst men having sex with men in Brighton, UK, as well as to characterise significant determinants of transmission at early stages of the disease. The relatedness of the viruses infecting 103 patients attending the Brighton clinics was reconstructed by phylogenetic means on the basis of *pol* gene sequences and interpreted under the light of epidemiological data.

With 35 primary infected individuals involved in transmission chains (34% of the cohort), the present study supports the assertion that primary HIV-1 infections may be associated with increased risk of onward transmission. There was a significant positive association between early transmission and young age, high rate of unprotected anal intercourses and high sexual partner change. The large representation of reported unprotected (mainly anal) intercourses and high partner changes in the linked cohort corroborates epidemiological reports about the recent trends in sexual behaviours in the UK, where an increase in high-risk sexual practice has been registered (Johnson et al. 2001). Nonetheless, only 31 (64.6%) of the unlinked patients reported unprotected anal intercourse, possibly reflecting the lack of information prior to 3 months prior diagnosis. There was a trend towards higher rates of STIs amongst linked individuals on a background of extremely high STI rates in the study population. The rising incidence of sexually transmitted diseases currently observed is almost certainly a consequence of the changing sexual attitudes in modern Britain, strongly supporting the argument for a improved STI surveillance, particularly of high-risk groups. In terms of disease progression markers, CD4 count was positively correlated with transmission, while plasma viral load failed to be an efficient indicator for transmissibility. This latter point could be explained by the discordance between blood and seminal VL, which appears to be a more consistent correlate of infectiousness in men (Pilcher et al. 2004b). The initial CD4+ cell count recorded in the cohort (i.e. 526 copies/µl) was lower than expected in acutely infected individuals, in whom CD4+ cells have not been strongly depleted yet (Pantaleo et al. 1993). This value is, however, consistent with previous measurements in similar studies (Weiss et al. 1992; Fidler et al. 2001; Deschamps et al. 2005). When looking at antiretroviral resistance motifs, one cluster out of the 15 characterised exhibited transmitted drug resistance-related mutations. Neither of the individuals involved in the transmission pair were drug experienced but still harboured the same resistance mutations, illustrating the potential for secondary spread of resistance strains, as previously reported (Yerly et al. 2001b; Taylor et al. 2003).

Despite the high rate of potential transmission exhibited within the tree (i.e. 34% of the sequences involved in potential transmission), the present results require to be seen in the light of limitations induced by data sampling. In fact, the involvement of two or more subjects in a transmission chain does not exclude the possibility of a common source of infection, rather than transmission within clusters. Individuals currently involved in transmissions may have been infected through a third party who has not been sampled, and yet harbour viruses with remarkably high genetic similarity. In that case, whether or not transmission occurred within the primary phase of the transmitter's infection is difficult to assess. Conversely, a fraction of the primary infections represented in the tree may have come from epidemiologically unlinked transmitters. That would result in no obvious clustering patterns within the tree, irrespective of the age of the transmitting infection, and the extent of linkage would be under-estimated. Moreover, identifying sexual partnership in a homogeneous population sub-group such as MSM appeared to be problematic, limiting the recognition of potential non-sampled transmitters. The surprisingly small number of linkages confirmed by clinical notes (3 out of the 15 clusters identified) emphasizes the difficulty in obtaining a reliable sexual history from the patients, aggravated by high rates of anonymous sexual partners.

The above findings support the view that as a disease stage, primary HIV-1 infection represents a major public health threat, and suggest that a substantial number of newly acquired infections may result from a limited pool of highly infectious sexually active individuals, unaware of their infectious condition. This is aspect of the epidemic prevention programs would benefit taking into account. The results presented in this chapter highlight the need to provide efforts in identification, counseling and possibly early treatment of individuals with primary HIV-1 infection. Indeed, a significant proportion of PHI remains undiagnosed in the community and an estimated 31% of the HIV-1 infected adults in the UK in 2001 were unaware of their infection (Brown et al. 2004). So far, HIV prevention programs have been heavily focused on protecting susceptible individuals (Pilcher et al. 2004a). However, reducing infectiousness of HIV-positive subjects may be an effective strategy. An efficient diagnosis of individuals during PHI, timely contact tracing, management of STIs and possibly treatment with antiretrovirals may all be useful methods not only to improve individual patient care but also to interrupt chains of transmission during this unique and possibly crucial stage of HIV infection.

The present analysis identified unprotected anal intercourse as a behaviour risk positively correlated with HIV transmission at the early stages of the disease. Preliminary studies suggest that high-risk sexual practices increase the incidence of HIV superinfection, and therefore the probability for recombination to occur. Despite its obvious significance for public health, intrasubtype recombination has been poorly addressed in the literature to date, probably on account of the practical difficulty of identifying such events. Evidence of potential recombination between *pol* sequences was found in our dataset using two distinct methodologies, despite a weak phylogenetic signal. By comparison, for split decomposition graphs constructed on control clusters (i.e sets of 5 randomly selected sequences from the dataset), no substantial evidence for recombination was found, despite a negligible reticulation at the root of the tree. This probable 'background noise' induced by the high degree of conservation of the *pol* gene illustrates the obstacles encountered when conducting such analyses. The lack of polymorphism across the sequences, and the *de facto* weakness of the phylogenetic signal it induces, may represent the main limitation of the present analysis. Not only is the HIV-1 *pol* gene extremely conserved, but variations within a subtype might also be insufficient to capture evidence of recombination without ambiguity. Furthermore, an accumulation of point mutations, such as drug resistance-related mutations, may have occurred after the recombination, masking the genetic specificity of the two parental elements of the mosaic sequence. Finally, the number of individuals involved in the potential transmission chains may be insufficient to really characterise recombination. In conclusion, the overall degree of recombination occurring amongst transmission chains is likely to be underestimated, and inversely the support for recombination found in the sequence M1689 must be considered with caution, as it may be artificially induced by the quality of the molecular data, and the restricted sensitivity of existing tools.

Intrasubtype recombination raise serious concerns regarding the monitoring of disease progression, future therapeutic options or even vaccine design. Superinfection occurring in antiretroviral-experienced individuals could have serious consequences for subsequent treatment. Thus, recombination between 2 or more HIV virions with differing drug resistance profiles could for instance result in a multi-drug resistant recombinant form. The practical difficulties encountered when investigating intrasubtype recombination, which leads to a critical lack of data on the topic, are likely to have a perverse effect a global underestimation of the true rate of recombination in HIV-1 and its real impact on worldwide public health.

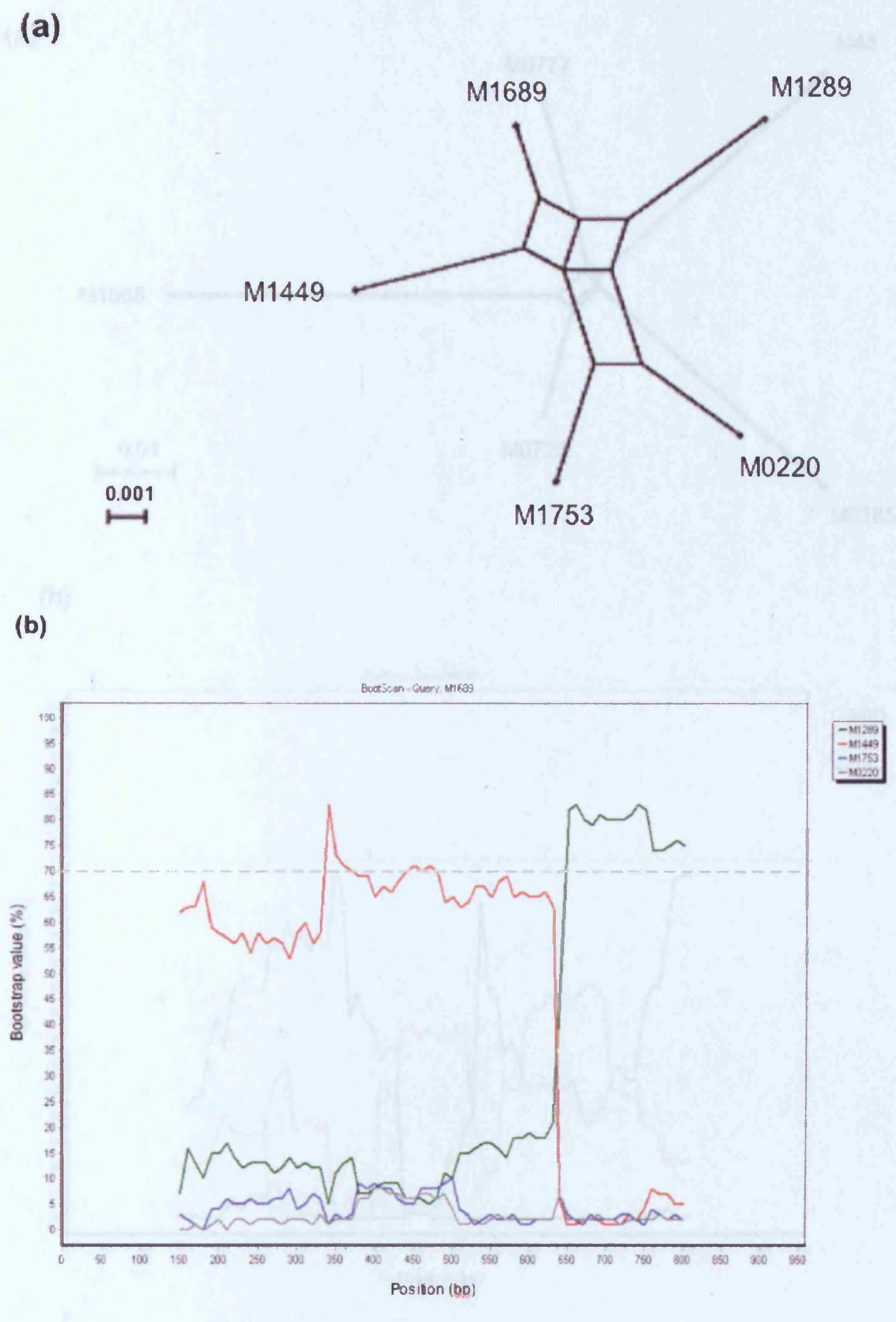


Fig 4.3. (a) Splits graph for the *pol* gene sequences involved in transmission cluster 1. Branch lengths are expressed in nucleotide substitutions per site (b) Bootscanning analysis of sequences involved in cluster 1. Sequence M1689 was used a query sequence. Bootscan was preformed with a sliding window of 300 nucleotides (incremented by 10 nucleotides per step) and 1000 bootstrap replicates. Bootscan threshold for potential recombination was set to 70%, as indicated by the dashed horizontal line.

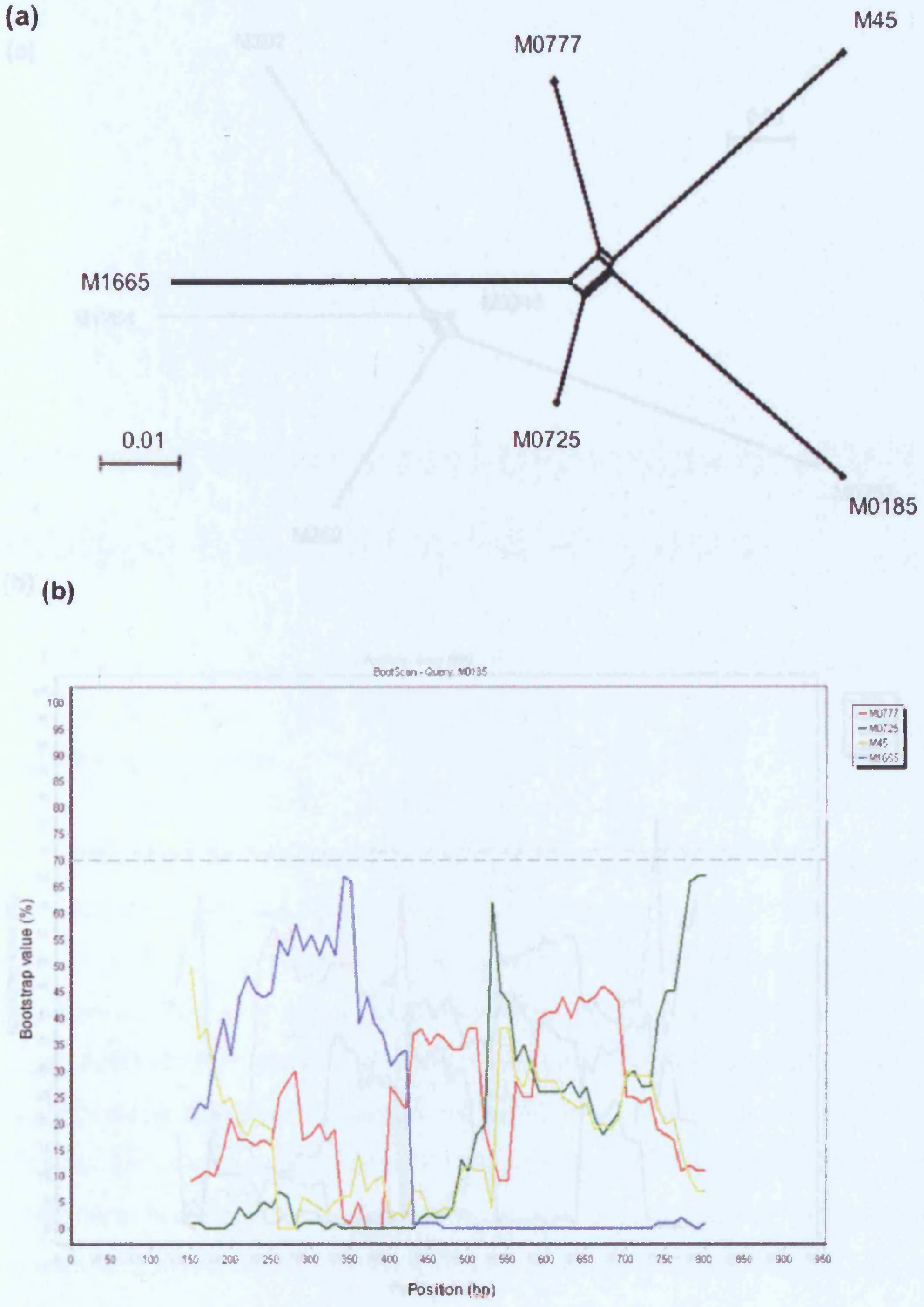


Fig 4.4. (a) Splits graph for the *pol* gene sequences randomly selected as control cluster 1. Branch lengths are expressed in nucleotide substitutions per site (b) Bootscanning analysis of sequences used for control cluster 1. Sequence M0185 was used a query sequence. Bootscan was preformed with a sliding window of 300 nucleotides (incremented by 10 nucleotides per step) and 1000 bootstrap replicates. Bootscan threshold for potential recombination was set to 70%, as indicated by the dashed horizontal line.

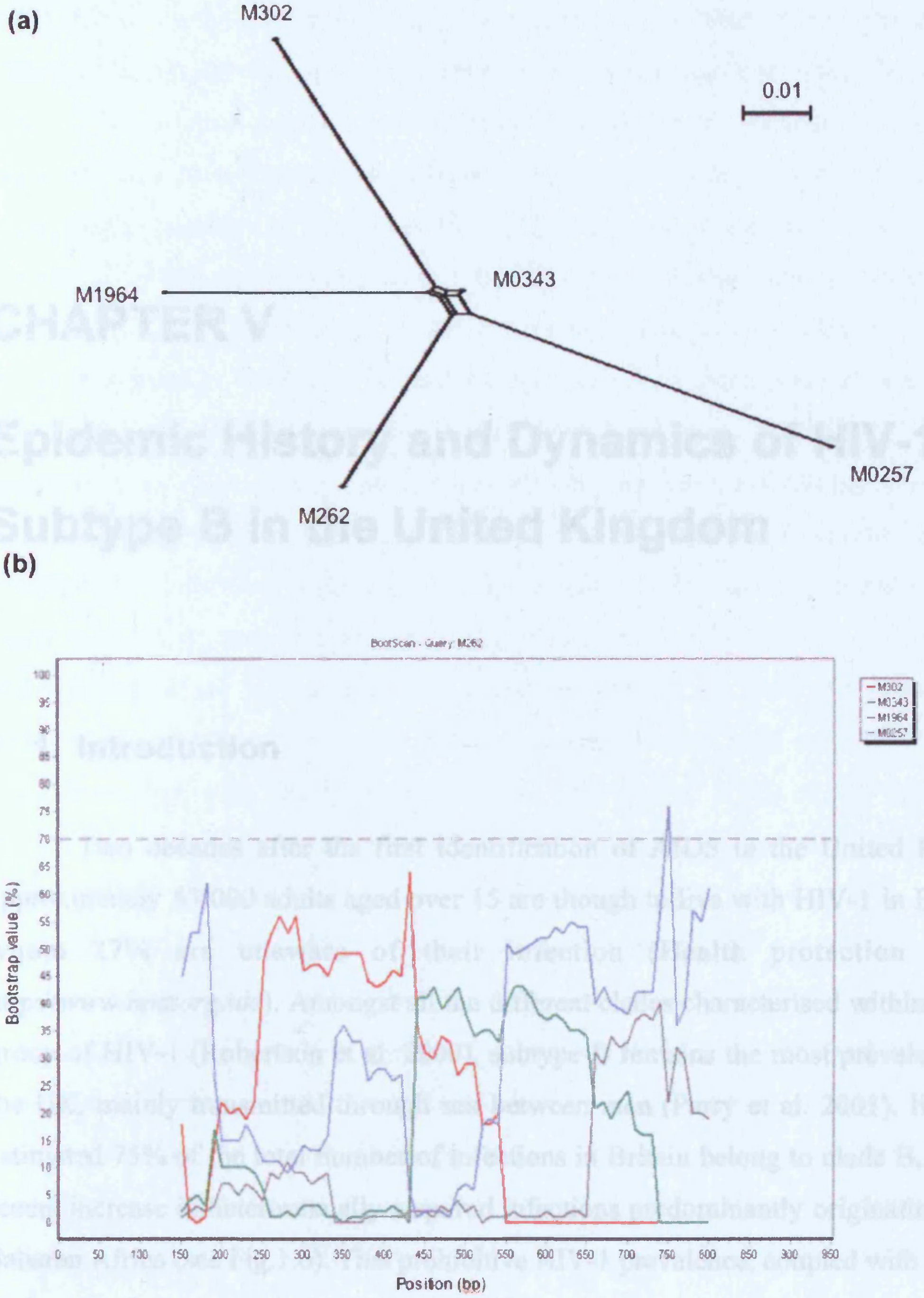


Fig 4.5. (a) Splits graph illustrating the genetic relationship between the *pol* gene sequences randomly selected as control cluster 2. Branch lengths are expressed in nucleotide substitutions per site (b) Bootscanning analysis of sequences used for control cluster 2. Sequence M262 was used a query sequence. Bootscan was preformed with a sliding window of 300 nucleotides (incremented by 10 nucleotides per step) and 1000 bootstrap replicates. Bootscan threshold for potential recombination was set to 70%, as indicated by the dashed horizontal line.

CHAPTER V

Epidemic History and Dynamics of HIV-1 Subtype B in the United Kingdom

1. Introduction

Two decades after the first identification of AIDS in the United Kingdom, approximately 53 000 adults aged over 15 are though to live with HIV-1 in Britain, of whom 27% are unaware of their infection (Health protection Agency, http://*www.hpa.org.uk/*). Amongst all the different clades characterised within the main group of HIV-1 (Robertson et al. 2000), subtype B remains the most prevalent within the UK, mainly transmitted through sex between men (Parry et al. 2001). Indeed, an estimated 75% of the total number of infections in Britain belong to clade B, despite a recent increase in heterosexually acquired infections predominantly originating in sub-Saharan Africa (see Fig.1.6). This prohibitive HIV-1 prevalence, coupled with continual high rate of new infections recorded year after year for the past decade, makes men having sex with men (MSM) the acquisition group at highest risk in the UK (Murphy et al. 2004). However, very little is known about how subtype B successfully invaded the British population, and more importantly, how the virus has subsequently spread and evolved.

Phylogenies reconstructed from sampled viral gene sequences are known to hold valuable information about the past structure of a population and can therefore be used to understand the course of a viral epidemic over time (Holmes et al. 1995; Nee et al. 1995). Hence the history of a pathogen population can be inferred from the genealogy of randomly sampled strains (as represented by a phylogenetic tree) using the coalescent theory of population genetics (see Chapter II, section 4.2). By this means, one can reconstruct the changing number of infected individuals through time and estimate the demographic parameters that shape the epidemic, such as the rate of growth in the number of infections and the date of introduction of a lineage into a host population (Kuhner et al. 1995). While the coalescent framework assumes neutral evolution, the HIV-1 *pol* gene is known to be under strong selection, both positive and negative (Richman et al. 1994; Rouzine and Coffin. 1999; Frost et al. 2000; Leal et al. 2004). However, sites under strong selective pressure only represent a small proportion of the sequence compared to neutral sites, and previous demographic analyses yielded similar demographic estimates when considering different HIV-1 genes, regardless of the variable selective pressures they are subject to (Lemey et al. 2003). The coalescent recently established itself as a state-of-the-art framework for molecular epidemiology and has previously been applied to the investigation of pathogens such as *Plasmodium falciparum* (Joy et al. 2003), hepatitis C virus (Pybus et al. 2003) or HIV type 1 and 2 (Holmes et al. 1995; Grassly et al. 1999; Yusim et al. 2001; Lemey et al. 2003; Robbins et al. 2003) providing new insights into those epidemics.

With the introduction of the routine generation of HIV-1 gene sequences for drug-resistance monitoring, molecular data on HIV-1 within the UK has become increasingly available, and amenable to modelling techniques for the study of virus evolution. Hitherto, the genetic variability of the HIV-1 *env* or *gag* genes has made these regions attractive for evolutionary studies, compared with genes under stronger evolutionary constraints such as *pol*. However, it has been recently demonstrated that the *pol* gene encodes sufficient variation to conduct phylogenetic analyses and reconstruct transmission events, despite the potential bias conferred by emergence of drug resistance-associated mutations (Hué et al. 2004).

In the present chapter, the history of the HIV-1 subtype B epidemic in the UK was reconstructed from the demographic information contained within a large dataset of contemporary *pol* gene sequences. The complexity of the epidemic within a defined risk group (i.e. men having sex with men) was explored, dating the introduction of epidemiologically significant lineages and estimating their rates of spread. The results were interpreted in the light of epidemiological data so as to understand the impact of the variation of the viral populations over time on public heath.

The work presented in this chapter was published in Hué *et al.* 2005.

2. Methods

2.1. Study Population

A total of 1645 HIV-1 subtype B *pol* gene sequences from the United Kingdom (UK) were used for the study. These were generated from plasma samples collected from over the UK by the Health Protection Agency's Antiviral Susceptibility Reference Unit, Birmingham, UK, as described in Chapter II. The samples were submitted for routine genotypic drug resistance testing between 1999 and 2003 and included samples from acute infections, chronic, drug naïve infections and from patients at the time of therapy failure. The sequences were 952 bp long, including the full-length protease gene as well as the first 218 codons of the reverse transcriptase gene. Around 85% of these sequences were from men who had sex with men (MSM).

2.2. Phylogenetic Reconstruction

According to surveillance data, only 1 out of 3 newly diagnosed HIV infections in the UK has been acquired within the country (Brown et al. 2004). In the light of the continuous mixing of HIV-1 strains worldwide, clusters of sequences deriving from single independent introductions of HIV within the British population had to be identified by phylogenetic means in order to further study within-UK transmission. In other words, in order to investigate the modality of spread of the epidemic in Britain, the extant of sporadically introduced sequences from abroad had to be assessed and these excluded from the analysis. An initial neighbor-joining (NJ) tree was constructed with 1645 UK and 1784 worldwide subtype B *pol* sequences, according to the Hasegawa-Kishino-Yano model of nucleotide substitution, with gamma distribution of rate heterogeneity. Due to the size of the alignment (n = 3429), the model was selected on the basis of a subset of 100 randomly selected sequences from it, using the software

Modeltest. The size of the dataset also precluded the application of a more complex model for computational reasons. The evolutionary parameters used for the computation are detailed in Table 5.1. The non-UK sequences used in the study were extracted from public resources, such as GenBank (*http://www.ncbi.nlm.nih.gov/*) and the Los Alamos HIV Sequence Database (*http://www.hiv.lanl.gov/*). Only subtype B *pol* sequences for which the date and country of sampling are documented were used.

After identification of UK transmission clusters, sequences of non-UK origin were stripped out and the phylogenies of the clusters were re-estimated with the program Paup*, using a maximum likelihood approach (Felsenstein. 1973). The trees were reconstructed according to the General Time Reversible model of nucleotide substitution (Yang. 1994), with proportion of invariable sites and substitution rate heterogeneity, as estimated with the software Modeltest (Posada and Crandall. 1998). The detail of the selected models of evolution is given in Table 5.1. In order to give an evolutionary direction to the lineages, each tree was rooted against a subtype D *pol* sequence from our dataset. The robustness of the ML topologies was statistically assessed for each ML trees by bootstrapping, with 1000 rounds of replication. The sequences involved in transmission clusters have been deposited into GenBank under the accession numbers AY669865 to AY670087.

2.3. Estimation of HIV-1 Subtype B *pol* Gene Rate of Evolution

Distances between two nodes of a phylogenetic tree are traditionally measured in units of substitutions per sites. Besides, assuming a constant molecular clock, the expected difference in the number of substitutions accumulated along to homologous lineages is expressed as $\delta = \mu t$, where μ and t stand for the specific rate of nucleotide substitution (i.e. the number of substitutions per site per unit of time) and the sampling interval respectively. Hence, in order to scale the transmission trees into calendar years, the evolutionary rate of the HIV-1 subtype B *pol* gene had to be estimated and applied to the branch length. When inferring rates of evolution, the accuracy of the estimation is highly dependant on the time window spanned by the sampled sequences. Preliminary analyses determined that our UK sequence database did not hold enough temporal signal (i.e. 5 years of sampling) for the estimation of μ. The evolution rate was therefore inferred from an independent dataset of 106 subtype B *pol* gene sequences and fixed

Table 5.1 Models of nucleotide substitutions applied to the global tree and transmission clusters

	Global tree	Transmission clusters					
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Model selected: *	HKY+G	GTR+G	GTR+I+G	GTR+I+G	GTR+I+G	GTR+I+G	GTR+I+G
- lnL =	243.23	1416.01	7045.27	4567.2646	4041.1064	3911.9917	4223.3477
Substitution model:							
[A-C] =	1.0000	1.0000	3.1795	2.2572	1.0000	2.8755	2.333
[A-G] =	0.8460	5.9039	9.8343	8.0336	4.6063	8.3961	11.6698
[A-T] =	1.0000	1.0000	0.9817	0.6271	0.3676	1.1073	0.8805
[C-G] =	1.0000	1.0000	1.4745	0.5368	0.3676	0.9935	1.3884
[C-T] =	0.163	34.405	9.8343	8.0336	4.6063	8.3961	11.6698
[G-T] =	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Base frequencies:							
freqA =	0.3808	0.3947	0.3821	0.395	0.4002	0.3884	0.3974
freqC =	0.1784	0.1551	0.1503	0.1564	0.158	0.1615	0.1559
freqG =	0.1539	0.2078	0.2178	0.2055	0.2107	0.2099	0.2079
freqT =	0.2869	0.2424	0.2498	0.2432	0.2312	0.2402	0.2388
Proportion of invariable sites:	-	0.8942	0.4201	0.4853	0.4752	0.3924	0.5134
Gamma distribution shape parameter:	1.6368	-	0.6779	0.8816	0.7161	0.5843	0.7041

** estimated according to the Akaike Information criterion (AIC), as implemented in Modeltest 3.06*

Abbreviations: HKY, Hasegawa-Kishino-Yano model; GTM, general time reversible model; +I, with invariable sites; +G, with gamma distribution; -lnL, log likelihood; [i-j]: rate of substitution between base i and base j; freq i, frequency of base i in the dataset.

as a prior probability density for Bayesian demographic analyses. The sequences used for this purpose were generated between 1983 and 2000 from men having sex with men (n = 44) and injecting drug users (IDUs; n = 62) participating in cohort studies at the Academic Medical Centre of Amsterdam, as shown in Fig. 5.1a (Lukashov and Goudsmit. 2002). The sequences were 804 bp long, spanning the entire protease gene (294 bp) and the first 510 bp of the RT gene. A total of 89% of the sequences harboured drug resistance mutations (see Fig. 5.1b). The evolution rate was estimated by Bayesian Markov Chain Monte Carlo inference using the program Beast (Drummond et al. 2002; Drummond and Rambaut. 2003), for a MCMC chain length of 10,000,000 states with sampling every 100th generation. GenBank accession numbers of these sequences are available in the original publication (Lukashov and Goudsmit. 2002).

2.4. Demographic History and Population Dynamics

The investigation of the epidemic history of the six UK clusters involved two steps. Firstly, five demographic models, each of which illustrate effective numbers of infections through time (see Fig. 2.11), were evaluated to select the model that best describes the epidemiological history of the UK transmission clusters. These models were compared by likelihood ratio test from likelihoods calculated by the program Genie (Pybus and Rambaut. 2002). Constant growth, exponential growth, logistic growth (exponential growth followed by constant population size), expansion growth (constant population size followed by exponential growth), and 'con-exp-con' growth (constant growth periods flanking an exponential growth phase) models were tested. The detail of these models is available in Pybus & Rambaut, 2002. Since Genie requires an input tree calculated under the assumption of a constant molecular clock, the program TipDate (Rambaut. 2000) was used to rescale each transmission tree under the Single Rate Dated Tip (SRDT) model. This model of nucleotide substitution assumes a constant rate of evolution across branches but relaxes the assumption of contemporaneous tips (Rambaut. 2000).

Secondly, the demographic and evolutionary parameters of the epidemic, with their confidence intervals, were estimated by Bayesian MCMC inference for a chain length of 10,000,000 states with sampling every 100th generation, using the program Beast. The estimated parameters include the date of the most recent common ancestor (MCRA) of the cluster, the effective number of infections at the most recent time of sampling *Ne* (i.e. the effective number of prevalent infections), and the growth rate during the exponential phase *r*. The Bayesian MCMC results were used to calculate a marginal posterior distribution of the demographic model for each cluster, i.e. a graphical representation of the effective number of infections through time, generated using the program Tracer (*http://evolve.zoo.ox.ac.uk/tracer/)*. An overview of the methodology used in this chapter is given in Fig. 5.2.

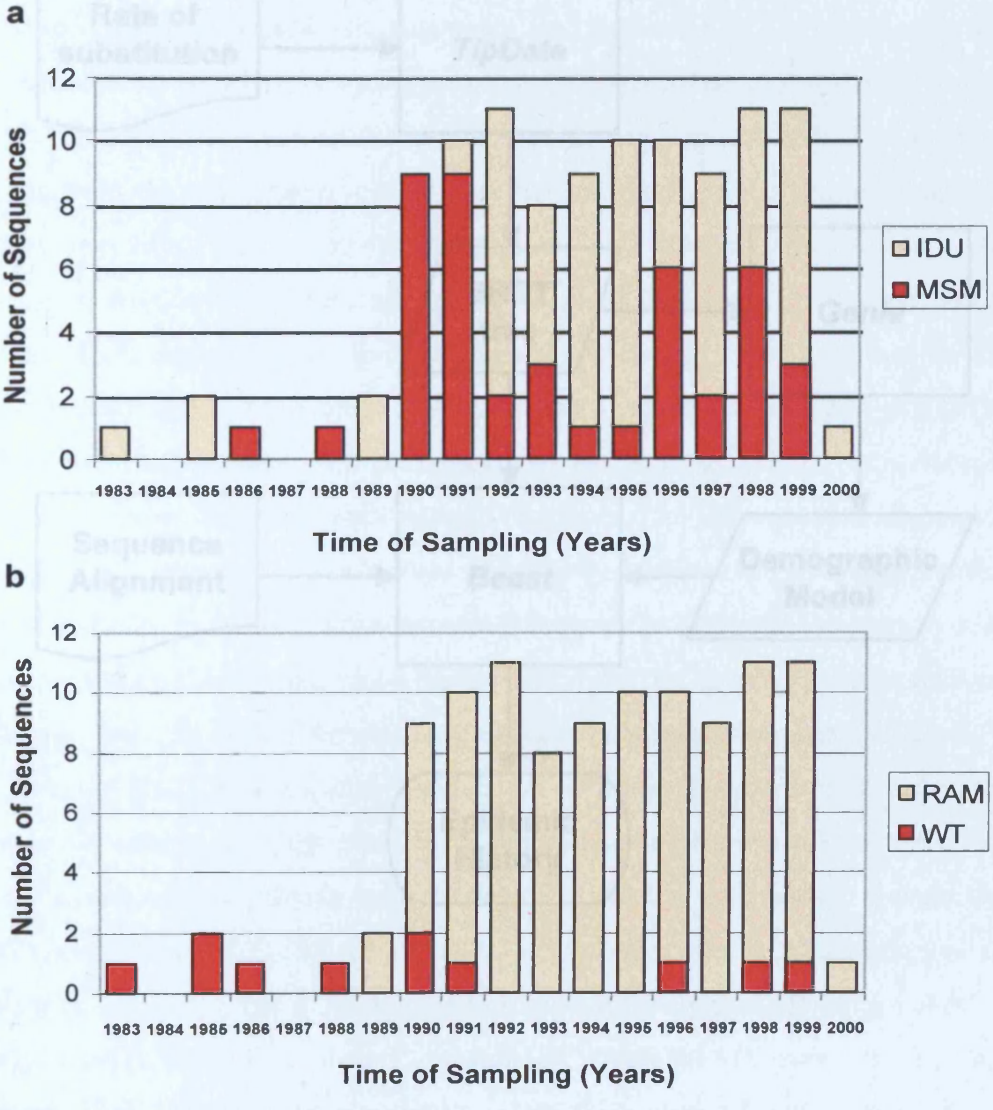


Fig. 5.1 Description of the Amsterdam cohort (n = 44). (a) Relative proportion of sequences collected from men having sex with men (MSM) and from intravenous drug users (IDV) in the cohort. (b) Prevalence of resistance associated mutations (RAM) compared to wild type (WT) in the cohort.

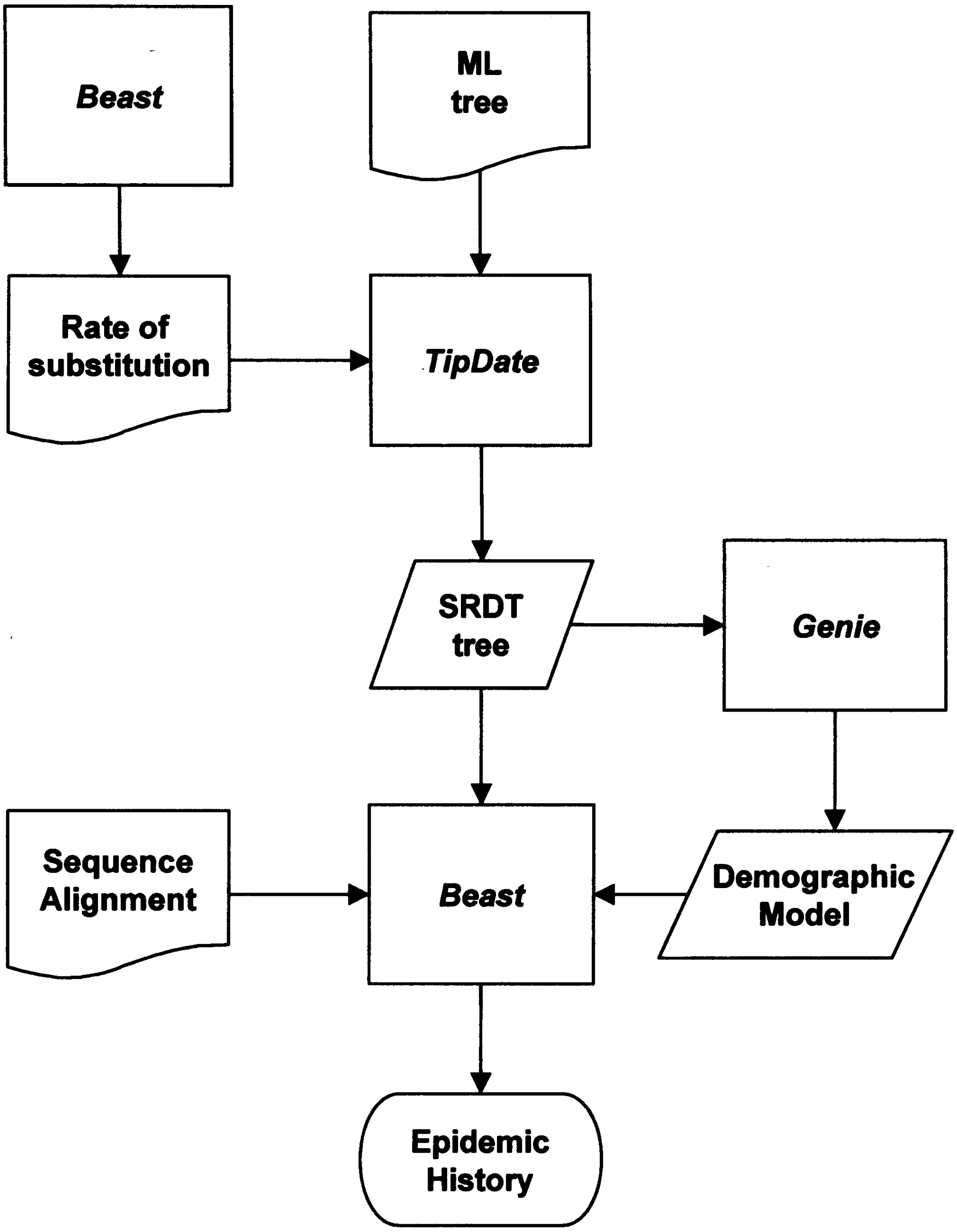


Fig. 5.2. Summary of the methodology used for the investigation of the spread of 6 HIV-1 lineages amongst MSM in the UK. Maximum likelihood tree topologies were rescaled in calendar year units with TipDate, according to a specific evolutionary rate estimated with Beast. The optimal demographic model was selected for each cluster on the basis on the rescaled topologies with the software Genie, and used for the estimation of the evolutionary and demographic parameters shaping the epidemic history of the strains, using the software Beast.

3. Results

3.1. Introduction of HIV-1 Subtype B in the UK

The initial NJ phylogenetic tree constructed from 3429 UK and worldwide subtype B pol sequences is too large to display here, but a schematic representation of the phylogeny is presented in Fig. 5.3. Three clustering patterns were distinguished:sporadic UK sequences, non-UK transmission clusters, and UK transmission clusters. Sporadic UK sequences (i.e. those that do not group with other UK lineages in the tree) probably represent single, independent introductions of the virus without subsequent spread. Transmission clusters were identified as clades of sequences from the same geographical area that descend from a common ancestor, indicating spread of the virus in that region. UK transmission clusters were differentiated from non-UK clusters on the basis of the size of the clade and the proportion of UK sequences within it: UK transmission clusters were defined as those clades with more than 25 sequences, 90% or more of which were of UK origin (Fig. 5.4). The relative arbitrariness of these criteria was based on the authors' experiences in coalescent inference. Empirical data showed that more than 20 sequences are required to reliably infer demographic trends (data not shown). Moreover, a bootstrap threshold of 90% was chosen to provide a manageable number of transmission clusters to study, together with undeniable confidence on the UK origin of these. We note that such methodology probably underestimates the number of transmission chains identified.

Most of the UK sequences represented sporadic lineages, scattered among sequences from other geographical areas, suggesting much geographical mixing of subtype B strains on a worldwide scale. Nonetheless, six UK transmission clusters were identified, involving 45, 61, 28, 28, 26 and 33 sequences. These transmission chains were distinct, indicating that at least six independent introductions of subtype B HIV-1 have succeeded in sustaining onward transmission within the UK over time, and until the present. Each UK transmission chain contained an array of sequences of diverse origin within Britain and no significant regional pattern was observed within a given UK cluster. The robustness of the clusters within the overall tree could not be statistically evaluated due to the huge size of the dataset. Nonetheless, the six UK lineages showed statistical robustness when compared to subsets of worldwide sequences for bootstrap analyses (neighbor-joining search with 1000 replicates implemented in the software Paup*; data not shown). To further explore the history of the six successful viral lineages, sequences of non-UK origin were removed from the six clusters and the phylogenetic histories of the UK sequences were re-estimated using a maximum likelihood approach. The ML trees are displayed in Fig. 5.5.

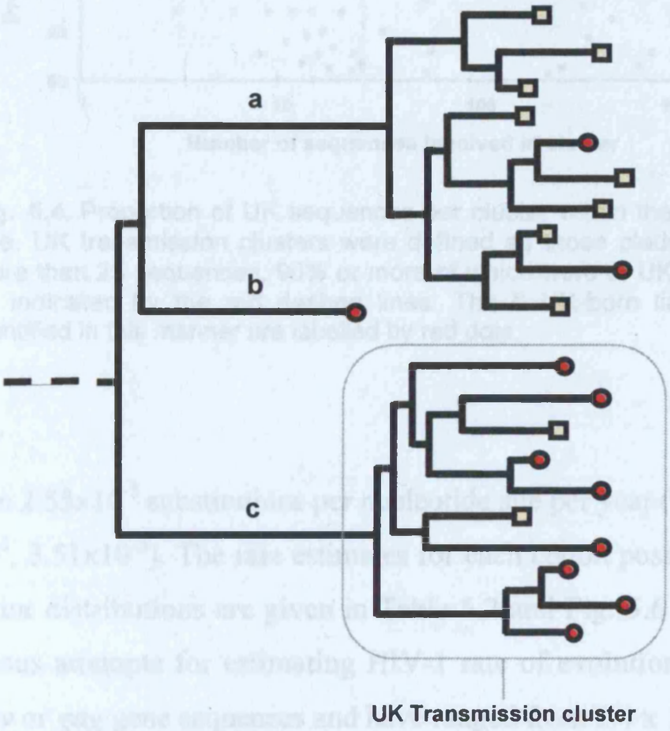


Fig. 5.3. Schematic representation of the phylogeny generated from 3429 UK and worldwide HIV-1 subtype B *pol* sequences. Red circles and yellow squares represent UK and non-UK sequences respectively. Three branching patterns were distinguished: (a) non-UK transmission clusters, (b) sporadic UK infections, and (c) UK transmission clusters. Transmission clusters are sequences from a particular location that descend from a common ancestor, indicating a successful spread of the virus. UK transmission clusters are defined as clades that include at least 25 sequences, 90% or more of which are of UK origin.

3.2. Estimation of HIV-1 Subtype B *pol* Gene Rate of Evolution

The evolution rate of the subtype B HIV-1 *pol* gene was calculated using an independent dataset of 106 sequences, sampled between 1983 and 2000 in Amsterdam (Lukashov and Goudsmit. 2002). Using a Bayesian MCMC framework, the average rate

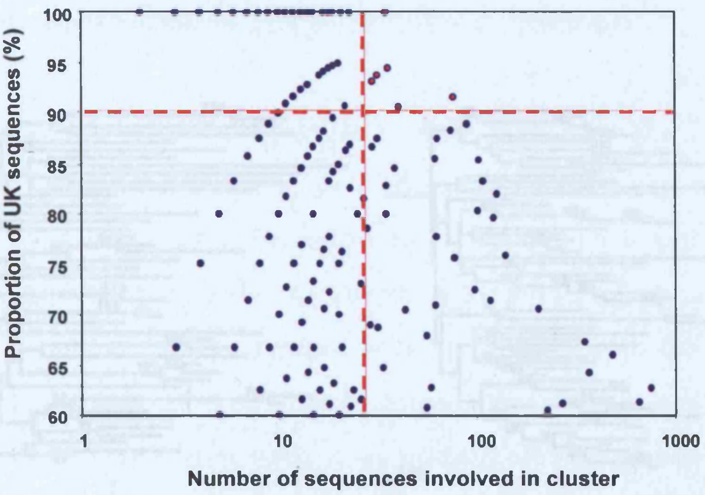


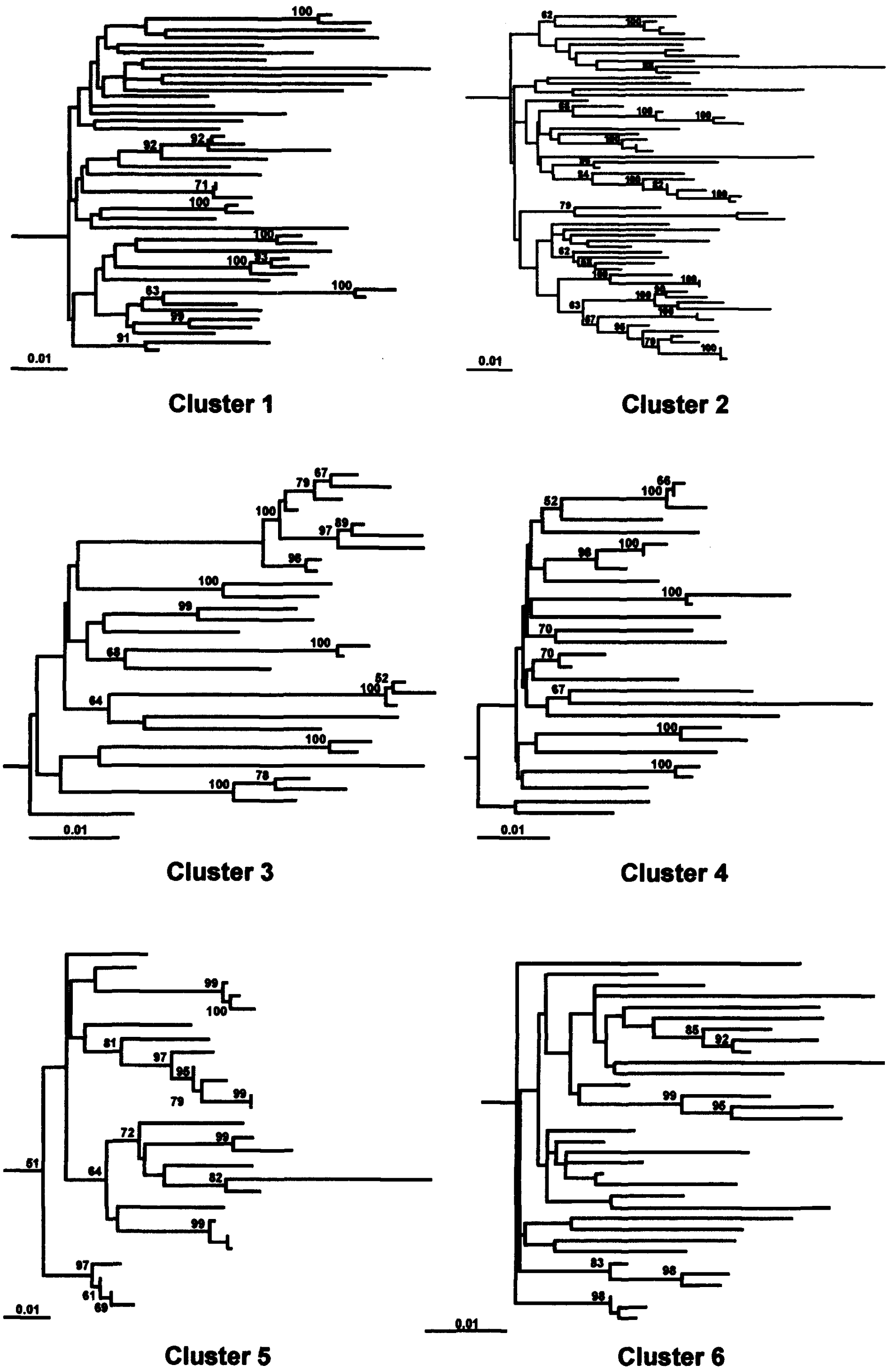
Fig. 5.4. Proportion of UK sequences per cluster within the global tree. UK transmission clusters were defined as those clades with more than 25 sequences, 90% or more of which were of UK origin, as indicated by the red dashed lines. The 6 UK-born lineages identified in this manner are labelled by red dots.

was estimated to be 2.55×10^{-3} substitutions per nucleotide site per year (95% confidence intervals: 1.74×10^{-3}, 3.51×10^{-3}). The rate estimates for each codon position of the gene and values' posterior distributions are given in Table 5.2 and Fig. 5.6 respectively. In comparison, previous attempts for estimating HIV-1 rate of evolution have typically relied on partial *env* or *gag* gene sequences and have ranged from 2.4×10^{-3} to 6.7×10^{-3} subst./site/year (see Table 5.3). Our estimate is consistent with the order of magnitude of 10^{-3} expected for an HIV-1 gene. The ML trees of the 6 UK clusters were thus rescaled on the assumption of a molecular clock with a rate of 0.0025 subst./site/year, and their topologies on a timescale of years are shown in Fig. 5.7.

Table 5.2. Estimated rates of evolution of the HIV-1 subtype B *pol* gene at the first, second and third codon positions, in substitution per site per year.

	Mean	Standard deviation	95% HPD * lower	95% HPD upper
1st codon position	2.28×10^{-3}	1.46×10^{-3}	1.42×10^{-3}	3.29×10^{-3}
2d codon position	1.43×10^{-3}	8.25×10^{-6}	8.91×10^{-4}	2.00×10^{-3}
3d codon position	3.93×10^{-3}	2.64×10^{-5}	2.90×10^{-3}	5.24×10^{-3}
Overall rate	2.55×10^{-3}	4.97×10^{-4}	1.74×10^{-3}	3.51×10^{-3}

* Higher posterior density



Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

Cluster 6

Fig. 5.5. Maximum likelihood trees for the 6 UK transmission clusters. The trees were constructed according to the GTR+I+G model of evolution and rooted against a subtype D HIV-1 pol sequences from the ASRU database, using the software Paup*. Bootstrap values above 50 are indicated on the branches.

3.3. Demographic History and Parameter Estimation

The likelihoods of the demographic models compared for the study are presented in Table 5.4. For each of the six clusters, a model of logistic population growth best fitted the demographic information contained in the tree topologies. Under the logistic model, the effective number of infections *Ne* grows exponentially at rate *r* from time t_a (time of the most recent common ancestor of the cluster) then decreases in growth rate towards the present. A schematic representation of the logistic model is given in Fig.5.8. By convention, time scale is represented with the present at the origin, going back into the past along the X-axis from left to right. One should bear in mind that *Ne* reflects the number of infections contributing to new infections, rather than the total number of prevalent infections within the transmission cluster.

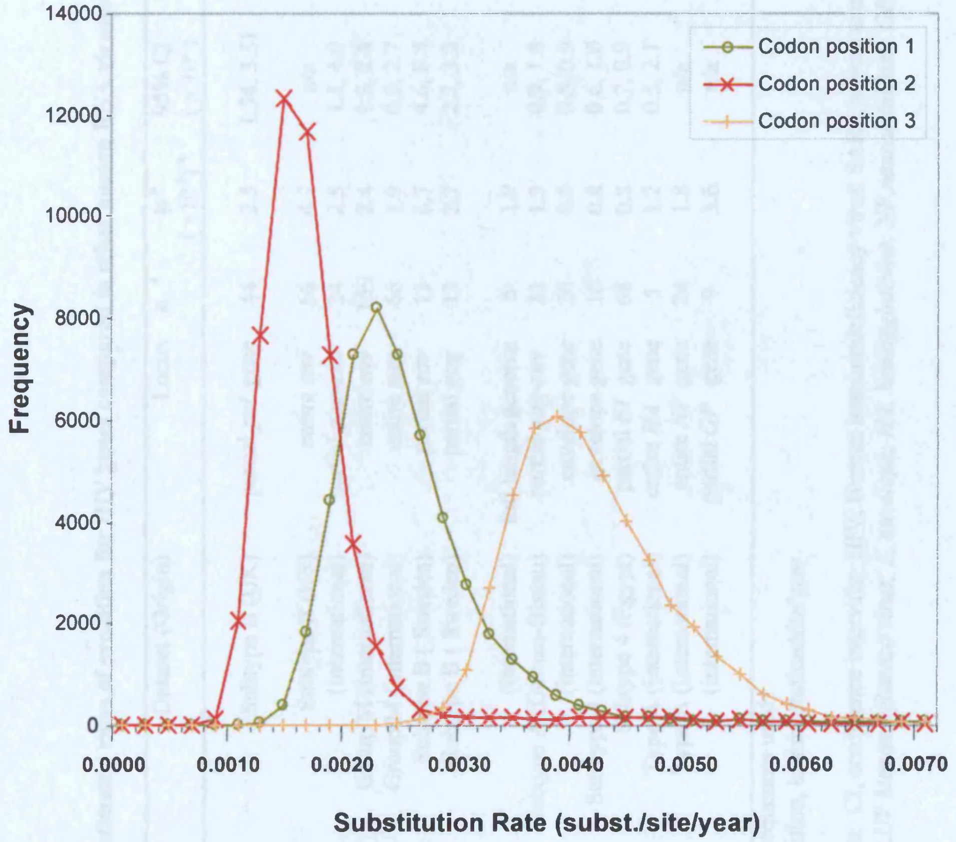


Fig. 5.6. Bayesian posterior distributions of values for the subtype B HIV-1 *pol* gene's substitution rate at each codon positions. The rates were estimated by Bayesian MCMC inference for a MCMC chain length of 10,000,000 states and are expressed as the number of nucleotide substitutions per site per calendar years.

Table 5.3. Estimated rates of evolution for HIV genes compared to other human RNA viruses

Organism	Dataset (Origin)	Locus	n [a]	μ [b] ($\times 10^{-3}$) [b]	95% CI ($\times 10^{-3}$)	Method used	Reference
HIV-1	Subtype B (UK)	partial *pol* gene	44	2.5	1.74, 3.51	Bayesian MCMC inference	*Hué et al. (2005)*
HIV-1	Subtype B (US)	entire *env*	66	4.7	n/a	Maximum likelihood estimation	*Robbins et al. (2003)*
HIV-1	(international)	partial *gag-env*	24	2.5	1.1, 4.0	Maximum likelihood estimation	*Jenkins et al. (2002)*
HIV-1	Group M (international)	entire *env*	159	2.4	1.8, 2.8	Root-to-tip linear regression	*Korber et al. (2000)*
HIV-1	Group M (international)	entire *gag*	66	1.9	0.9, 2.7	Root-to-tip linear regression	*Korber et al. (2000)*
HIV-1	Subtype B (Sweden)	partial *env*	13	6.7	4.6, 8.8	Pairwise distance linear regression	*Leitner & Albert (1999)*
HIV-1	Subtype B (Sweden)	partial *gag*	13	2.7	2.2, 3.2	Pairwise distance linear regression	*Leitner & Albert (1999)*
SARS virus	(international)	full length genome	6	1.9	n/a	Least square method	*Lu et al. (2004)*
HIV-2	Subtype A (Guinea-Bissau)	partial *gag-env*	33	1.3	0.9, 1.8	Maximum likelihood estimation	*Lemey et al. (2003)*
DENV-4	(international)	envelope gene	20	0.6	0.5, 0.9	Maximum likelihood estimation	*Twiddy et al. (2003)*
DENV-4	Serotype 4 (international)	envelope gene	16	0.8	0.6, 1.0	Bayesian MCMC inference	*Drummond et al.(2003)*
HCV	Subtype 4 (Egypt)	partial *E1* gene	68	0.8	0.7, 0.9	Bayesian MCMC inference	*Pybus et al. (2003)*
HFLUV-A	Type A (international)	entire *HA* gene	5	1.2	0.5, 2.1	Root-to-tip linear regression	*Suzuki & Nei (2002)*
HFLUV-A	Type A (international)	entire *NP* gene	24	1.8	n/a	Maximum likelihood estimation	*Jenkins et al. (2002)*
Ebola virus	(international)	partial *GP* gene	9	3.6	n/a	Root-to-tip linear regression	*Suzuki Y. et al. (1997)*

[a] number of sequences used

[b] rate of evolution, in substitution/site/year

Abbreviations: CI, confidence intervals; HIV, Human immunodeficiency virus; SARS, severe acute respiratory syndrome; DENV, Dengue virus; HCV, Hepatitis *C virus; HFLUV human influenza virus; E, envelope; HA, hemagglutinine; NP,neuraminidase; GP, glycoprotein; MCMC, Markov Chain Monte Carlo.*

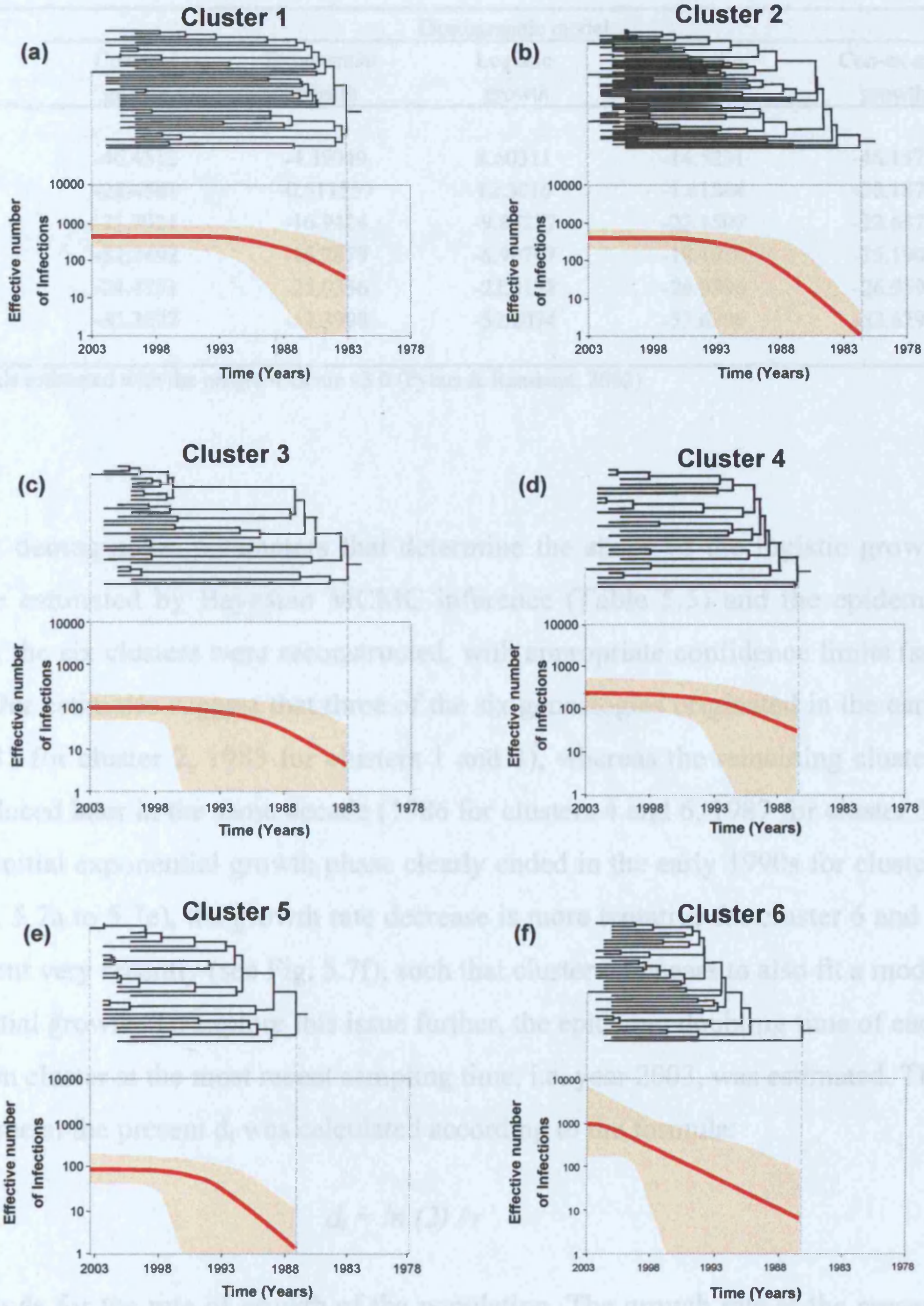


Fig. 5.7. Phylogenetic trees of the six UK transmission clusters and their corresponding estimated epidemic histories (all shown on the same timescale). The trees represent the ancestral relationships of sequences belonging to each cluster. (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, (e) cluster 5, (f) cluster 6. The demographic histories were estimated by Bayesian MCMC inference using a model of logistic growth and show change in the effective number of infections through time (in calendar years). The red line represents the median estimate of the effective number of infections, whereas the yellow shaded area delimitates the 95% confidence limits of the estimate.

Table 5.4. Likelihoods of 5 demographic models for the 6 UK transmission clusters

Cluster	Demographic model				
	Constant growth	Exponential growth	Logistic growth	Expansion growth	Con-ex-con growth
1	-46.4512	-4.39049	8.60311	-14.5231	-46.1573
2	-28.4581	-0.511559	12.5016	-1.61864	-30.1872
3	-21.7921	-16.9424	-9.85277	-22.1507	-22.6473
4	-32.7492	-16.2879	-6.99797	-18.1014	-35.1909
5	-24.4751	-25.0356	-22.0138	-26.9596	-26.9596
6	-82.3527	-52.3998	-52.2074	-53.6396	-53.6299

Log likelihoods estimated with the program Genie v3.0 (Pybus & Rambaut, 2002)

The demographic parameters that determine the shape of the logistic growth curve were estimated by Bayesian MCMC inference (Table 5.5) and the epidemic histories of the six clusters were reconstructed, with appropriate confidence limits (see Fig. 5.7). Our estimates suggest that three of the six genealogies originated in the early 1980s (1981 for cluster 2, 1983 for clusters 1 and 3), whereas the remaining clusters were introduced later in the same decade (1986 for clusters 4 and 6, 1987 for cluster 5). While the initial exponential growth phase clearly ended in the early 1990s for clusters 1 to 5 (Fig. 5.7a to 5.7e), the growth rate decrease is more tentative for cluster 6 and is only apparent very recently (see Fig. 5.7f), such that cluster 6 appears to also fit a model of exponential growth. To explore this issue further, the epidemic doubling time of each transmission cluster at the most recent sampling time, i.e. year 2003, was estimated. The doubling time at the present d_t was calculated according to the formula:

$$d_t = ln\,(2)\,/\,r$$

where r stands for the rate of growth of the population. The growth rate at the present $r(t_0)$ was calculated as a function of d_t and t_{50}, the time in the past when the population was half of its size) :

$$r(t_0) = (ln\,(2)\,/\,d_t)\;x\;(\,1 - (\,1\,/\,(1{+}C)))$$

where:

$$C = 1\,/\,(exp(\,(ln(2)\;x\;t_{50})\,/\,d_t\,) - 2)$$

These formulas were adapted from Pybus et al., 2001. The current doubling time was more likely to be <20 years (equal to an exponential growth rate >0.035 years^{-1}) for cluster 6 in comparison to the other clusters, as shown in Table 5.6. In marked contrast, the exponential growth rates at the time of initiation of each cluster (r) were very similar, with an average of 0.80 years^{-1}. Finally, the current effective number of infections *Ne* varied from cluster to cluster, ranging from 94 (cluster 5) to 1350 (cluster 6) effective infections.

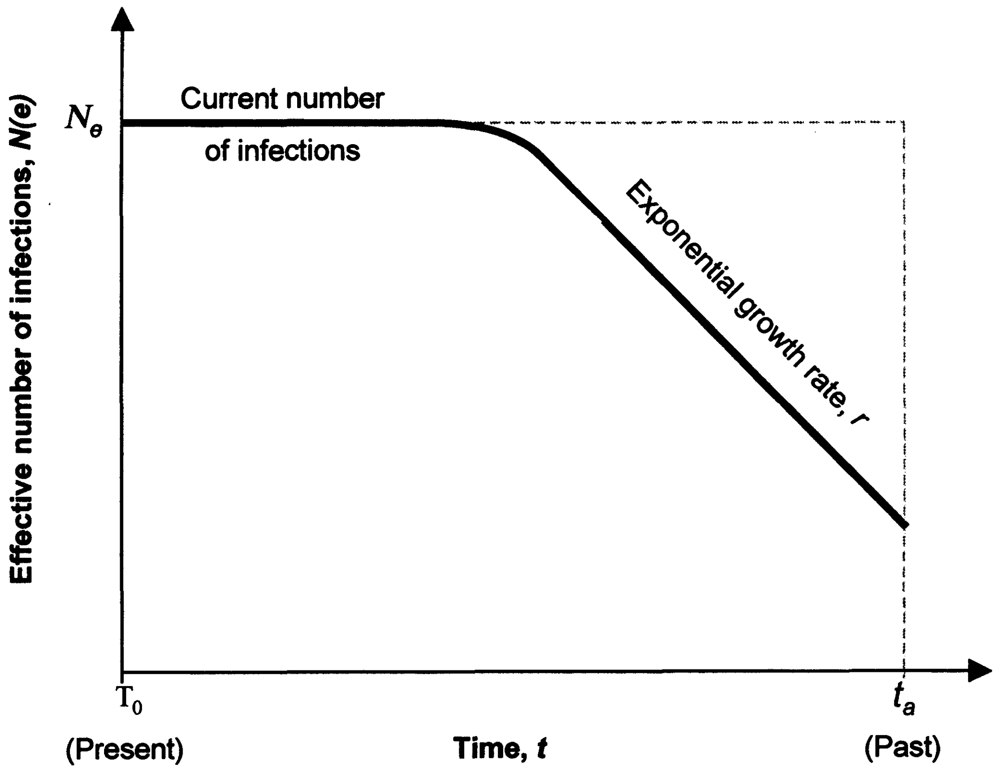

Effective number of infections, N(e)
N_e
Current number of infections
Exponential growth rate, r
T_0
t_a
(Present)
Time, t
(Past)

Fig. 5.8. Schematic representation of the logistic model of population growth. According to this model, the number of infections grows exponentially at rate *r* from time t_a (time of the most recent common ancestor of the sampled sequences). The growth rate slows as time moves towards the present, such that *Ne* represents the effective number of infections at the present. *Ne* can be thought of as the number of infections contributing to new infections, rather than the total number of prevalent infections within the cluster.

Table 5.5. Parametric estimates (with 95% confidence intervals) under the logistic growth demographic model

Cluster	μ [a]	Ne [b]	r [c]	Origin of the tree (*yrs)*
Cluster 1	2.55 x 10^{-3} (0.0017, 0.0035)	493 (201, 833)	1.08 (0.66, 2.56)	1983 (1978, 1988)
Cluster 2	2.55 x 10^{-3} (0.0017, 0.0035)	386 (190, 655)	0.47 (0.30, 0.95)	1981 (1976, 1987)
Cluster 3	2.55 x 10^{-3} (0.0017, 0.0035)	98 (42 , 171)	0.50 (0.19, 4.62)	1983 (1977, 1988)
Cluster 4	2.55 x 10^{-3} (0.0017, 0.0035)	250 (88, 483)	1.38 (0.63, 2.50)	1986 (1982, 1991)
Cluster 5	2.55 x 10^{-3} (0.0017, 0.0035)	94 (36, 85)	0.68 (0.35, 2.10)	1987 (1983, 1991)
Cluster 6	2.55 x 10^{-3} (0.0017, 0.0035)	1350 (109, 5489)	0.67 (0.37, 3.85)	1986 (1981, 1991)
US cluster [d]	6.7 x 10^{-3} (n/a, n/a)	4830 (1995, 26 750)	0.834 (0.72, 0.945)	1968 (1966, 1970)

[a] Rate of evolution, in substitutions per site per year, independently estimated and fixed as a prior density probability
[b] Effective number of infections
[c] Rate of exponential growth, in *years* $^{-1}$
[d] from Robbins *et al.* , 2003

4. Discussion

The history of the HIV-1 subtype B epidemic within the UK was explored through the demographic information contained within contemporary molecular data. A timescale for the spread of strains presently circulating in the UK was estimated, and compared to epidemiological data so as to understand causes of their variation over time.

The present estimates suggest that subtype B viruses currently circulating within the UK are comprised of at least six established chains of transmission, introduced in the early and mid 1980's. This demonstrates the existence of distinct, possibly non-overlapping sexual networks within the predominant MSM risk group and argues against the hypothesis that one initial entry of HIV-1 was responsible for the spread of the subtype B epidemic. It also emphasises the preponderant role of migration in the HIV-1 epidemic in Britain, as illustrated by the overwhelming prevalence of sporadic lineages (86% of the total UK samples) in the genealogy, representing the proportion of imported viruses imported viruses not leading to a discernable cluster. The transmission clusters we characterised had a similar epidemic curve and geographic distribution within the UK, indicating a simultaneous spread under the same demographic pressures. The introduction of the early founder viruses in the early 1980's (i.e. clusters 1-3) seems to coincide with the explosion of new infections reported by epidemiological data at the time (Health Protection Agency, ***http://www.hpa.org.uk/)***. This coupling of HIV introduction with epidemiological changes is likely to have favoured the emergence and persistence of the transmission chains presently circulating amongst MSM. However, since the first cases of AIDS were reported in this country in 1982 (World Health Organisation, *http://www.who.int/emc-hiv/fact_sheets/*), with a likely original infection of these individuals at least 10 years previously, currently circulating strains may not represent the original lineages established within the country. If earlier strains existed they may have been unsuccessful in sustaining transmission chains until the present, and may no longer be of epidemiological significance. The absence of older strains could also reflect a sampling bias.

Table 5.6. The probabilities for the clusters' doubling time at the present to be <20 years (equal to an exponential growth rate >0.035 years^{-1})

	Cluster 6	Cluster 5	Cluster 4	Cluster 3	Cluster 2	Cluster 1
$p(r < 0.0001)$	0.2624	0.3451	0.8801	0.5094	0.0869	0.9478
$p(0.0001 < r < 0.01)$	0.2521	0.5336	0.1003	0.3443	0.7361	0.0491
$p(r > 0.01)$	0.4855	0.1213	0.0197	0.1462	0.1771	0.0031
$p(d_{50} < 20\text{years})$	0.3630	0.0443	0.0087	0.0777	0.0393	0.0030

Abbreviations: p, probability; r, growth rate; d_{50}, population doubling time.

For all six lineages, the exponential growth phase coincides with a significant augmentation of newly acquired HIV-1 infections reported within MSM and IDU in the UK (Health protection Agency, *http://www.hpa.org.uk/)*. The average growth rate during the exponential phase was estimated to be 0.80 years^{-1} (ranging from 0.47 to 1.38), approximating to a doubling time of around 1 year. This value is similar to the estimate of the exponential growth rate of the US epidemic (0.83 years^{-1}, 0.72 to 0.94) (Robbins et al. 2003), suggesting that the two epidemics obey similar demographic pressure. This idea is also supported by the similarities found in terms of effective number of infections of the two epidemics. Despite a wide heterogeneity across the six UK-born transmission clusters in *Ne* (spanning from 94 to 1350), the average effective population size amongst the six viral populations reaches 445 and approximates to 2.5% of the infected population in Britain. This is remarkably similar to that observed in the US, where an estimated effective population size of 5000, with ~200,000 infections in 1995, was estimated (Robbins et al. 2003). These values represent the number of infections contributing to onward transmission, rather than the larger number of actual infections. Despite the accuracy and sensitivity of the methodology used, the time scales estimated here should be regarded with caution, as they are extremely sensitive to sampling error. Sampling time of sequences is generally only recorded to the precision of a year, inducing inaccuracy in time frames. Also, despite the known time of collection of the samples, the cumulative time a given sequence had to evolve within a host is hardly even known. HIV infected cells may harbour proviral DNA for a certain period of time, during which the replication, and therefore the evolution, of the organism are suspended (Wolinsky et al. 1996; Rodrigo et al. 1999; Shankarappa et al. 1999). As a result, sequenced genes may be older than the expected according to their sampling date, and the time calibration consequently biased.

Since 1990, there have been important changes in Britain's demographic structure, social attitude and awareness of HIV-1/AIDS (Johnson et al. 2001). Despite an increase in high-risk behaviour among men having sex with men (such as the number of sexual partners or concurrent partnerships), a significant increase in consistent condom use has been reported since 1990. Such a change in sexual health, coupled to large scaled campaigns against AIDS over the past decade, could explain the equilibrium reached by the effective number of prevalent infections. The effect of antiretroviral therapy on the parameters of the epidemic dynamic needs also to be considered. Although therapy is instituted primarily to reduce progression of disease, it may also impact on transmission through reduction of infectivity. If so, we would expect evidence of a plateau in the late (rather then early) 1990's at the time that highly active antiretroviral therapy became widely used, as well as a down turn in the estimated number of new infections within the UK. In fact Health Protection Agency epidemiological data reports no significant changes in the incidence of HIV-1 within gay men since the late 1980's, and an actual increase over the past 3 years. That suggests that antiviral therapy has not had a significant impact on the growth of the epidemic; indeed, it has been proposed that the epidemic is driven by transmissions in primary infection, before therapy is usually initiated (see Chapter IV). The recent increase in new infections is not reflected in the growth dynamics of any of the six populations identified by this analysis. On-going analyses of the type undertaken here will clarify whether the recent increase in new subtype B infections derive from longstanding vial lineages, or newly introduced viruses.

In conclusion, state-of-the-art statistical methods were applied to HIV-1 molecular data to identify some key parameters of the dynamics and growth of the subtype B epidemic in the UK. This demonstrated that currently circulating viruses within MSM entered the UK in the mid 1980's and that a slow down of epidemic growth for these lineages occurred in the early 1990's. It is often assumed that the HIV-1 epidemic within the UK represents smaller, independent epidemics defined by risk group. The existence of multiple epidemics (i.e. at least six) within MSM was demonstrated here, with comparable evolution over time and obeying related demographic constrains. The identification of these multiple lineages within the predominant risk group of the HIV-1 epidemic in the UK suggests that such heterogeneity must be considered when developing HIV monitoring and prevention initiatives.

CHAPTER VI

General Discussion

The present thesis aimed to investigate the reliability of the HIV-1 *pol* gene for the identification of transmissions networks by phylogenetic means, on the basis of which molecular analysis of epidemiological relevance were further conducted. Evolutionary and epidemiological approaches were combined in order to assess the correlates of transmission within a population of primary HIV-1 infected individuals within a localised risk group (i.e. men having sex with men in Brighton, UK), exploiting both HIV-1 *pol* genes sequences and clinical data. When large networks of transmission were characterised, the viral genomes involved were tested for intra-subtype recombination, under the assumption that super-infection might have occurred. Finally, the epidemic history of HIV-1 subtype B in the UK was reconstructed from sampled HIV-1 *pol* gene sequences, providing new insights into the complexity of HIV-1 epidemics that must be considered when developing HIV monitoring and prevention initiatives.

The collection of analyses presented in these pages emphasizes the advantage of combining state-of-the-art epidemiological studies to phylogenetic frameworks when investigating the dynamics of a viral epidemic as complex as HIV-1. In fact, phylogenetics and population genetics have already played a central role. Rapid accumulation of DNA sequence data since the 1980s has transformed the focus of HIV-1 research, and phylogenetics not only shed light on the origins of the virus (Sharp et al. 2001), but also on the evolutionary processes that shape viral genetic diversity within and among patients (Crandall et al. 1999), and have provided evidence in forensic cases of HIV transmission (Ou et al. 1992). Evolution is a forward process, causing organisms as well as populations to changes their characteristics over time, with significant effect on the biology of a pathogen. Thus, unifying the epidemiological and evolutionary dynamics of HIV-1 helps understanding the past and predicting the future of the AIDS epidemic. Understanding the host helps understanding the virus, both at the infra- or inter-host level. Nonetheless the use study of molecular sequences in an epidemiological context is not exempt from pitfalls, and the accuracy of evolutionary or historical estimates is sensitive to various levels of biases.

1. Violated Assumptions

Amongst these biases is the burden of assumptions. Several simplifying assumptions had to be made that may alter the accuracy of the estimates presented here. First, the number of infected hosts within a local population was assumed to be homogeneous and epidemiologically close. However, reports on the evolution of the HIV-1 epidemic clearly show a discrepancy in HIV-1 prevalence and incidence worldwide (UNAIDS, *http://www.unaids.org/*), as well as significant variation in rates of exposure between HIV-1 infected individuals (Plummer et al. 1991; Service and Blower. 1995). Moreover, viral lineages were exclusively sampled from individuals in whom viral load was detectable. Since administration of anti-retroviral therapy suppresses the replicative capacity of the virus (Frenkel et al. 2003), it is sensible to assume that the overall rate of evolution of the virus is reduced during therapy, as a result of both genetic bottlenecks (Martinez-Picado et al. 1999) and pharmacological barriers (Perno et al. 1998; Durant et al. 2000). If, as it was suggested, the administration of anti-retroviral therapy reduces the overall HIV-1 rate of evolution (Fraser et al. 2001), estimates obtained from a population of treated individuals (such as our *pol* gene sequences database) are likely to be flawed. The estimating of evolution rates across cohorts of drug-naïve and -experienced individuals could help to assess the extent of this potential bias. Finally, while coalescent-based inference assumes neutral evolution, there is plethora of evidence according to which both positive and negative selective pressure is heavily exerted on HIV-1 genes (Yamaguchi-Kabata and Gojobori. 2000; Yang et al. 2000). However, selection on HIV genes within infected individuals does not appear to impact on the shape of genealogies at the epidemiological (i.e. inter-host) level and therefore the bias in coalescent estimates would remain trivial. Importantly, previous coalescent analyses have yielded similar demographic estimates from different HIV-1 genes, which are under considerably different selection pressure (Lemey et al. 2003).

2. Selection Bias

Sample selection represents another obstacle of importance when conducting molecular epidemiology analyses. The selection of a suitable genetic region is the first step at which bias can be encountered, and the reliability of conclusions drawn from the given gene can potentially be altered when generalized to the organism scale. The study reported in Chapter III stresses the point that, when approaching the use of HIV-1 sequences to characterise linkage, the key issue is for a molecular dataset to contain sufficient variability as defined by phylogenetic criteria, regardless of whether the sequence is *gag, pol* or *env*. And that genetic variability should not alone be a criterion to consider. Hypervariable regions, for instance, may be subject to convergent evolution (i.e. identical mutational patterns in unlinked sequences), such as in the V3 loop of the *env* gene (Holmes et al. 1992; Zhang et al. 1993). The rapid genetic diversification of this region is also likely to compromise identification of linked sequences in distantly sampled individuals. Indeed, both divergence and diversity of the HIV-1 *env* gene have been shown to increase linearly in early stages of infection (Shankarappa et al. 1999). Hence, the choice for an appropriate genetic target for such studies must not only be considered in light of the intrinsic variability of a given dataset, but also of the possible time span separating the samples under comparison. A sensible way to determine such criteria would be to incorporate positive controls within the sequence alignment, such as sequences from known transmission pairs or intra-patient follow-up samples. Close attention is therefore required in dealing with HIV-1 gene sequences for epidemiological, clinical or forensic purposes. It is nonetheless sensible to conclude from the results shown in Chapter III that the *pol* gene, widely available since the onset of drug resistance testing, holds sufficient genetic variation to allow phylogenetic analyses and offers an attractive alternative to more variable regions. It could also be hypothesized that hypervariable gene sequences (such as *env*, or to a lesser degree *gag*) are more suitable for the analysis of intra-host viral evolution, while more conserved regions are more informative for evolution analysis at a larger scale, i.e. amongst groups, populations or even species.

Equal attention is required when selecting a study cohort, both qualitatively and quantitatively. When investigating HIV-1 transmission clusters in Chapter VI, a minimum clade size of 25 was used because smaller sample sizes are unlikely to give reliable coalescent estimates under complex demographic models, and a minimum fraction of 90% UK sequences was chosen to ensure that the clusters that were identified represent chains of transmission that have overwhelmingly occurred in the UK. It is however noteworthy that this methodology probably underestimates the number of transmission chains identified. The use of parallel datasets may also avoid the potential peculiarity of a given sample and relax the sensitivity of conclusions drawn from a single dataset. The study of six independent HIV-1 transmission clusters in Chapter VI, yielding similar patterns of demographic history, added robustness to the conclusion drawn from their dynamics. Finally, the *pol* sequences exploited in this thesis are unlikely to have been sampled randomly with respect to geographical, social or ethnic origin and certain samples may have a disproportionate influence on the estimations. It is consequently of importance to implement rational sampling protocols, in order to reduce selection biases without undermining the advantage of the present abundance of molecular information. Developing resampling schemes into Bayesian analyses, or performing randomization tests of genetic diversity, could secure this issue.

3. Recombination

Another obvious limitation of the methodology used in the present thesis would be the sensitivity of evolutionary inference to recombination. With a rate of recombination per base exceeding that of mutation (Jetzt et al. 2000), it becomes crucial to consider the effect of recombination events when studying HIV-1 molecular evolution. Phylogenetic trees can be seriously affected by recombination events, thus rendering less reliable the estimation of population histories or event timings. In the presence of recombination, sequences have dissimilar phylogenetic histories in the different parts of their locus, infringing the assumption of a strictly bifurcating genealogy (i.e. where one descendant has two ancestors only), and do evolve along a set of correlated trees rather than a single tree (Hudson. 1983; Anisimova et al. 2003). In regard to the high rates of superinfection and recombination characterising HIV genes, bifurcating phylogenetic trees may be suboptimal when representing transmission dynamics. Also, since recombination leads to apparent substitution rate heterogeneity among sites (Worobey. 2001), mosaic sequences can compromise the reliability of phylogenetic reconstructions, Bayesian inference or molecular clock tests. Nonetheless, recombination in early internal branches is more disruptive in the genealogy of the tree than at the tips of the tree. In a star-like tree such as seen in the context of inter-host HIV-1 transmission, recombination is thus likely to have a less pronounced impact on estimates. On a population dynamic standpoint, recombination can also have a potential adverse effect and bias analyses toward the underestimation of the time of most recent common ancestors, the underestimation of the amount of recent divergence, or apparent sign of exponential growth (Shierup and Hein. 2000). A recent study, however, suggests that coalescent-based analyses of HIV population histories do not significantly differ when assuming a single or multiple genealogy for all loci (Lemey et al. 2004).

There is a need for methods that allow the incorporation of recombination in phylogenetic or coalescent-based analyses. Work along that line has recently started, Work has recently started to that respect, but the challenge is such that further developments are needed before a successful application in standard evolutionary analyses (Griffiths and Marjoram. 1996; Fearnhead and Donnelly. 2001). The incorporation of models of recombination into molecular evolution analyses would represent a robust alternative to the difficult detection of recombination in poorly variable genes (Taylor and Korber. 2005). Although good estimates can be obtained by applying population-genetic methods to DNA sequences (reviewed in Stump and McVean. 2003), a high-resolution measure of HIV-1 recombination rates remains an experimental challenge, especially in well-conserved genes such as *pol*. And since phylogenetic approaches are becoming increasingly significant in HIV-1 research, understanding the recombination process would help understanding the dynamics of the epidemic.

4. Who's Next?

Extending the methodology used in Chapter V to alternative risk groups would provide a useful comparison of the population dynamics and evolution of HIV-1 across risk-groups in the UK. The mechanisms of HIV-1 transmission are so versatile that the establishment of new infections across risk groups is likely to exhibit different dynamics. For instance, transmission networks through needle sharing are likely to be initiated more rapidly than through sexual contact. While transmission dynamics amongst IDUs are susceptible to rate of needle sharing, rates of partner exchange or high-risk sexual behavior are more specific determinant of transmission in the latter risk group (Anderson and May. 1988; Kaplan. 1989; Blower et al. 1991). Thus, studies showed that injection frequency was positively and highly significantly associated with HIV-1 *env* genetic diversity (Carneiro et al. 1999) and mutation rate in patients who had injected at least once a day during the previous 6 months was estimated to be 62% greater than the rate in those who had not injected at all. Furthermore, heroin and cocaine have been reported to enhance HIV replication in vitro (Peterson et al. 1990). It is therefore sensible to expect discrepancies in the dynamics of the two sub-epidemics. Understanding these would provide useful material for prevention and monitoring programs in diverse risk groups, as illustrated by the recent molecular analyses done of the explosive IDU epidemics in Eastern Europe (Roudinskii et al. 2004).

Alternatively, numerous studies showed that *in vitro* fitness and transmissibility varies across HIV-1 subtypes (Kunanusont et al. 1995; Hu et al. 1999; Ball et al. 2003), although it is poorly understood whether, if extrapolated to *in vivo* behaviour, it results from intrinsic features of the variants enhancing transmissibility or from epidemiological factors. HIV-1 subtype C, for instance, accounts for more than 47% of the worldwide HIV-1 new infections in 2000 (Osmanov et al. 2002), and is suspected to represent 55% of the global number of HIV-1 infections (Esparza and Bhamarapravati. 2000). It remains unclear whether this predominance is a reflection of a founder effect or higher fitness conferred by genetic specificities (Oelrichs et al. 2000; Ndung'u et al. 2001). In England and Wales, subtype C accounts for the majority (32%) of new HIV-1 diagnosis, mainly acquired in Sub-Saharan Africa through heterosexual transmission (Tatt et al. 2004). The number of heterosexually acquired HIV infections diagnosed in the UK has risen hugely over the last 15 years, and took over the rate of infection in men who have sex with men for the first time in 1999 (see Fig. 1.6). Considering the difference in epidemiological history and mode of transmission, viruses of subtype B and C, both highly prevalent in the UK, are likely to show different patterns of genetic diversity and dynamics. Since molecular data routinely generated from subtype C strains is increasingly available in Britain, molecular analyses of the latter would help understanding the relative influence of biological and behavioral factors on the spread of the subtype C epidemic in the UK. For instance, the estimation of the rate of spread and effective population size of subtype C sequences from the UK would help determining the relative role of genetic drift and selection in the shape of the sub-epidemic The comparison with population dynamics of subtype C sequences from Asian or Sub-Saharan African countries, where epidemiological determinants differ form those of the UK, would be equally informative to that respect.

Bibliography

Abele L.G. and R.W. Debry (1992). "Florida dentist case: research affiliation and ethics." Science 255(5047): 903.

Aceijas C., G.V. Stimson, M. Hickman and T. Rhodes (2004). "Global overview of injecting drug use and HIV infection among injecting drug users." Aids 18(17): 2295-2303.

Al Jabri A.A. (2002). "HLA and in vitro susceptibility to HIV infection." Mol Immunol 38(12-13): 959-967.

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(9): 5918-5924.

Albert J., J. Wahlberg and M. Uhlen (1993). "Forensic evidence by DNA sequencing." Nature 361(6413): 595-596.

Ammaranond P., P. Cunningham, R. Oelrichs, K. Suzuki, C. Harris, L. Leas, A. Grulich, D.A. Cooper and A.D. Kelleher (2003). "Rates of transmission of antiretroviral drug resistant strains of HIV-1." J Clin Virol 26(2): 153-161.

Anderson R.M. and R.M. May (1988). "Epidemiological parameters of HIV transmission." Nature 333(6173): 514-519.

Anisimova M., R. Nielsen and Z. Yang (2003). "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites." Genetics 164(3): 1229-1236.

Apetrei C., D. Descamps, G. Collin, I. Loussert-Ajaka, F. Damond, M. Duca, F. Simon and F. Brun-Vezinet (1998). "Human immunodeficiency virus type 1 subtype F reverse transcriptase sequence and drug susceptibility." J Virol 72(5): 3534-3538.

Arion D., N. Kaushik, S. Mccormick, G. Borkow and M.A. Parniak (1998). "Phenotypic mechanism of HIV-1 resistance to 3'-azido-3'-deoxythymidine (AZT): increased polymerization processivity and enhanced sensitivity to pyrophosphate of the mutant viral reverse transcriptase." Biochemistry 37(45): 15908-15917.

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(1): 198-203.

Auerbach D.M., W.W. Darrow, H.W. Jaffe and J.W. Curran (1984). "Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact." Am J Med 76(3): 487-492.

Autran A., G. Carcelain, T. S. Li, C. Blanc, D. Mathez, R. Tubiana, C. Katlama, P. Debré and J. Leibowitch (1997). “Positive effects of combined antiretroviral therapy on CD4+ T cell homeostasis and function in advanced HIV disease.” Science 277(5322): 112-6.

Ayouba A., S. Souquieres, B. Njinku, P.M. Martin, M.C. Muller-Trutwin, P. Roques, F. Barre-Sinoussi, P. Mauclere, F. Simon and E. Nerrienet (2000). "HIV-1 group N among HIV-1-seropositive individuals in Cameroon." Aids 14(16): 2623-2625.

Bacchetti P. and A.R. Moss (1989). "Incubation period of AIDS in San Francisco." Nature 338(6212): 251-253.

Baldrich-Rubio E., S. Anagonou, K. Stirrups, E. Lafia, D. Candotti, H. Lee and J.P. Allain (2001). "A complex human immunodeficiency virus type 1 A/G/J recombinant virus isolated from a seronegative patient with AIDS from Benin, West Africa." J Gen Virol 82(Pt 5): 1095-1106.

Balfe P., P. Simmonds, C.A. Ludlam, J.O. Bishop and A.J. 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(12): 6221-6233.

Ball S.C., A. Abraha, K.R. Collins, A.J. Marozsan, H. Baird, M.E. Quinones-Mateu, A. Penn-Nicholson, M. Murray, N. Richard, M. Lobritz, P.A. Zimmerman, T. Kawamura, A. Blauvelt and E.J. Arts (2003). "Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C." J Virol 77(2): 1021-1038.

Bandelt H.J. and A.W. Dress (1992). "Split decomposition: a new and useful approach to phylogenetic analysis of distance data." Mol Phylogenet Evol 1(3): 242-252.

Barlow K.L., I.D. Tatt, P.A. Cane, D. Pillay and J.P. Clewley (2001). "Recombinant strains of HIV type 1 in the United Kingdom." AIDS Res Hum Retroviruses 17(5): 467-474.

Barre-Sinoussi F., J.C. Chermann, F. Rey, M.T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum and L. Montagnier (1983). "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." Science 220(4599): 868-871.

Bayes T. (1764). "An essay toward solving a problem in the doctrine of chances." Philosophical Transactions of the Royal Society of London 53: 370-418.

Bebenek K., J. Abbotts, S.H. Wilson, T.A. Kunkel (1993). “Error-prone polymerization by HIV-1 reverse transcriptase. Contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots.” J Biol Chem 268(14) :10324–10334.

Bebenek K., J. Abbotts, J.D. Roberts, S.H. Wilson, T.A. Kunkel (1989). “Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase.” J Biol Chem 264(28) :16948–16956.

Benn S., R. Rutledge, T. Folks, J. Gold, L. Baker, J. McCormick, P. Feorino, P. Piot, T. Quinn and M. Martin (1985). “Genomic heterogeneity of AIDS retroviral isolates from North America and Zaire.” Science 230 (4728), 949–951.

Bikandou B., J. Takehisa, I. Mboudjeka, E. Ido, T. Kuwata, Y. Miyazaki, H. Moriyama, Y. Harada, Y. Taniguchi, H. Ichimura, M. Ikeda, P.J. Ndolo, M.Y. Nzoukoudi, R. M'vouenze, M. M'pandi, H.J. Parra, P. M'pele and M. Hayami (2000). "Genetic subtypes of HIV type 1 in Republic of Congo." AIDS Res Hum Retroviruses 16(7): 613-619.

Birch C.J., R.F. Mccaw, D.M. Bulach, P.A. Revill, J.T. Carter, J. Tomnay, B. Hatch, T.V. Middleton, D. Chibo, M.G. Catton, J.L. Pankhurst, A.M. Breschkin, S.A. Locarnini and D.S. Bowden (2000). "Molecular analysis of human immunodeficiency virus strains associated with a case of criminal transmission of the virus." J Infect Dis 182(3): 941-944.

Bishop K.N., R.K. Holmes, A.M. Sheehy and M.H. Malim (2004). "APOBEC-mediated editing of viral RNA." Science 305(5684): 645.

Blaak H., A.B. Van't Wout, M. Brouwer, M. Cornelissen, N.A. Kootstra, N. Albrecht-Van Lent, R.P. Keet, J. Goudsmit, R.A. Coutinho and H. Schuitemaker (1998). "Infectious cellular load in human immunodeficiency virus type 1 (HIV-1)-infected individuals and susceptibility of peripheral blood mononuclear cells from their exposed partners to non-syncytium-inducing HIV-1 as major determinants for HIV-1 transmission in homosexual couples." J Virol 72(1): 218-224.

Blackard J.T. and K.H. Mayer (2004). "HIV superinfection in the era of increased sexual risk-taking." Sex Transm Dis 31(4): 201-204.

Blackard J.T., B. Renjifo, W. Fawzi, E. Hertzmark, G. Msamanga, D. Mwakagile, D. Hunter, D. Spiegelman, N. Sharghi, C. Kagoma, and M. Essex (2001). “HIV-1 LTR subtype and perinatal transmission.” Virology 287(2): 261-5.

Blower S.M., A.N. Aschenbach, H.B. Gershengorn and J.O. Kahn (2001). "Predicting the unpredictable: transmission of drug-resistant HIV." Nat Med 7(9): 1016-1020.

Blower S.M., D. Hartel, H. Dowlatabadi, R.M. Anderson and R.M. May (1991). "Drugs, sex and HIV: a mathematical model for New York City." Philos Trans R Soc Lond B Biol Sci 331(1260): 171-187.

Bobkov A., E. Kazennova, L. Selimova, M. Bobkova, T. Khanina, N. Ladnaya, A. Kravchenko, V. Pokrovsky, R. Cheingsong-Popov and J. Weber (1998). "A sudden epidemic of HIV type 1 among injecting drug users in the former Soviet Union: identification of subtype A, subtype B, and novel gagA/envB recombinants." AIDS Res Hum Retroviruses 14(8): 669-676.

Boone L.R. and A.M. Skalka (1981). "Viral DNA synthesized in vitro by avian retrovirus particles permeabilized with melittin. II. Evidence for a strand displacement mechanism in plus-strand synthesis." J Virol 37(1): 117-126.

Boyer J.C., K. Bebenek and T. A. Kunkel (1992). “Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA templates.” Proc Natl Acad Sci USA 89(13): 6919–6923.

Brodine S.K., J.R. Mascola, P.J. Weiss, S.I. Ito, K.R. Porter, A.W. Artenstein, F.C. Garland, F.E. Mccutchan and D.S. Burke (1995). "Detection of diverse HIV-1 genetic subtypes in the USA." Lancet 346(8984): 1198-1199.

Brown A.E., K.E. Sadler, S.E. Tomkins, C.A. Mcgarrigle, D.S. Lamontagne, D. Goldberg, P.A. Tookey, B. Smyth, D. Thomas, G. Murphy, J.V. Parry, B.G. Evans, O.N. Gill, F. Ncube and K.A. Fenton (2004). "Recent trends in HIV and other STIs in the United Kingdom: data to the end of 2002." Sex Transm Infect 80(3): 159-166.

Brown H., Sanger F and Kitai R. (1955). "The structure of pig and sheep insulins." Biochem J 60(4): 556-565.

Bryant D. and V. Moulton (2004). "Neighbor-net: an agglomerative method for the construction of phylogenetic networks." Mol Biol Evol 21(2): 255-265.

Bukrinsky M.I., N. Sharova, M.P. Dempsey, T.L. Stanwick, A.G. Bukrinskaya, S. Haggerty and M. Stevenson (1992). "Active nuclear import of human immunodeficiency virus type 1 preintegration complexes." Proc Natl Acad Sci U S A 89(14): 6580-6584.

Burda S.T., F.A. Konings, C.A. Williams, C. Anyangwe and P.N. Nyambi (2004). "HIV-1 CRF09_cpx Circulates in the North West Province of Cameroon Where CRF02_AG Infections Predominate and Recombinant Strains Are Common." AIDS Res Hum Retroviruses 20(12): 1358-1363.

Busch M.P., L.L. Lee, G.A. Satten, D.R. Henrard, H. Farzadegan, K.E. Nelson, S. Read, R.Y. Dodd and L.R. Petersen (1995). "Time course of detection of viral and serologic markers preceding human immunodeficiency virus type 1 seroconversion: implications for screening of blood and tissue donors." Transfusion 35(2): 91-97.

Carlson J.R., M.L. Bryant, S.H. Hinrichs, J.K. Yamamoto, N.B. Levy, J. Yee, J. Higgins, A.M. Levine, P. Holland, M.B. Gardner and Et Al. (1985). "AIDS serology testing in low- and high-risk groups." Jama 253(23): 3405-3408.

Carneiro M., X.F. Yu, C. Lyles, A. Templeton, A.E. Weisstein, M. Safaeian, H. Farzadegan, D. Vlahov and R.B. Markham (1999). "The effect of drug-injection behavior on genetic evolution of HIV-1." J Infect Dis 180(4): 1025-1032.

Carr J.K., J.N. Torimiro, N.D. Wolfe, M.N. Eitel, B. Kim, E. Sanders-Buell, L.L. Jagodzinski, D. Gotte, D.S. Burke, D.L. Birx and F.E. Mccutchan (2001). "The AG recombinant IbNG and novel strains of group M HIV-1 are common in Cameroon." Virology 286(1): 168-181.

Carr J.K., T. Laukkanen, M.O. Salminen, J. Albert, A. Alaeus, B. Kim, E. Sanders-Buell, D.L. Birx and F.E. Mccutchan (1999). "Characterization of subtype A HIV-1 from Africa by full genome sequencing." Aids 13(14): 1819-1826.

Carr J.K., B. Foley, T. Leitner, M. Salminen, B.T. Korber and F.E. Mccutchan (1998). Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic. Human Retrovirus and AIDS. L.A.N. Los Alamos National Laboratory.

Ceballos A., G. Andreani, S.E. Ayala, Y. Romer, I. Rimoldi, M.R. Agosti and L.M. Peralta (2004). "Epidemiological and molecular evidence of two events of father-to-child HIV type 1 horizontal transmission." AIDS Res Hum Retroviruses 20(8): 789-793.

Cham F., L. Heyndrickx, W. Janssens, G. Van Der Auwera, K. Vereecken, K. De Houwer, S. Coppens, H. Whittle and G. Van Der Groen (2000). "Study of HIV type 1 gag/env variability in The Gambia, using a multiplex DNA polymerase chain reaction." AIDS Res Hum Retroviruses 16(17): 1915-1919.

Chant K., D. Lowe, G. Rubin, W. Manning, R. O'donoughue, D. Lyle, M. Levy, S. Morey, J. Kaldor, R. Garsia and Et Al. (1993). "Patient-to-patient transmission of HIV in private surgical consulting rooms." Lancet 342(8886-8887): 1548-1549.

Chao L. (1997). “Evolution of sex and the molecular clock in RNA viruses.” Gene 205(1): 301-308.

Cheingsong-Popov R., R.A. Weiss, A. Dalgleish, R.S. Tedder, D.C. Shanson, D.J. Jeffries, R.B. Ferns, E.M. Briggs, I.V. Weller, S. Mitton and Et Al. (1984). "Prevalence of antibody to human T-lymphotropic virus type III in AIDS and AIDS-risk patients in Britain." Lancet 2(8401): 477-480.

Choe H., M. Farzan, Y. Sun, N. Sullivan, B. Rollins, P.D. Ponath, L. Wu, C.R. Mackay, G. Larosa, W. Newman, N. Gerard, C. Gerard and J. Sodroski (1996). "The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates." Cell 85(7): 1135-1148.

Choisy M., C.H. Woelk, J.F. Guegan and D.L. Robertson (2004). "Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes." J Virol 78(4): 1962-1970.

Cichutek K., H. Merget, S. Norley, R. Linde, W. Kreuz, M. Gahr and R. Kurth (1992). "Development of a quasispecies of human immunodeficiency virus type 1 in vivo." Proc Natl Acad Sci U S A 89(16): 7365-7369.

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

Coffin J.M. (1995). "HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy." Science 267(5197): 483-489.

Colfax G.N., S.P. Buchbinder, P.G. Cornelisse, E. Vittinghoff, K. Mayer and C. Celum (2002). "Sexual risk behaviors and implications for secondary HIV transmission during and after HIV seroconversion." Aids 16(11): 1529-1535.

Cote H.C., Z.L. Brumme and P.R. Harrigan (2001). "Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir." J Virol 75(2): 589-594.

Crandall K.A. (1995). "Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus." J Virol 69(4): 2351-2356.

Crandall K.A., D.A. Vasco, D. Posada and H. Imamichi (1999). "Advances in understanding the evolution of HIV." Aids 13(Suppl A): S39-47.

Curran J.W., W.M. Morgan, A.M. Hardy, H.W. Jaffe, W.W. Darrow and W.R. Dowdle (1985). "The epidemiology of AIDS: current status and future prospects." Science 229(4720): 1352-1357.

Daar E.S., T. Moudgil, R.D. Meyer and D.D. Ho (1991). "Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection." N Engl J Med 324(14): 961-964.

D'aquila R.T., J.M. Schapiro, F. Brun-Vezinet, B. Clotet, B. Conway, L.M. Demeter, R.M. Grant, V.A. Johnson, D.R. Kuritzkes, C. Loveday, R.W. Shafer and D.D. Richman (2003). "Drug resistance mutations in HIV-1." Top HIV Med 11(3): 92-96.

Dalgleish A.G., P.C. Beverley, P.R. Clapham, D.H. Crawford, M.F. Greaves and R.A. Weiss (1984). "The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus." Nature 312(5996): 763-767.

Darwin C. (1859). On the origin of species. John Murray, London.

De Baar M.P., A. Abebe, A. Kliphuis, G. Tesfaye, J. Goudsmit and G. Pollakis (2003). "HIV type 1 C and C' subclusters based on long terminal repeat sequences in the Ethiopian type 1 subtype C epidemic." AIDS Res Hum Retroviruses 19(10): 917-922.

Debry R.W., L.G. Abele, S.H. Weiss, M.D. Hill, M. Bouzas, E. Lorenzo, F. Graebnitz and L. Resnick (1993). "Dental HIV transmission?" Nature 361(6414): 691.

Delaporte E., W. Janssens, M. Peeters, A. Buve, G. Dibanga, J.L. Perret, V. Ditsambou, J.R. Mba, M.C. Courbot, A. Georges, A. Bourgeois, B. Samb, D. Henzel, L. Heyndrickx, K. Fransen, G. Van Der Groen and B. Larouze (1996). "Epidemiological and molecular characteristics of HIV infection in Gabon, 1986-1994." Aids 10(8): 903-910.

De Leys R., B. Vanderborght, M. Vanden Haesevelde, L. Heyndrickx, A. Van Geel, C. Wauters, R. Bernaerts, E. Saman, P. Nijs, B. Willems and Et Al. (1990). "Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin." J Virol 64(3): 1207-1216.

Delgado E., M.M. Thomson, M.L. Villahermosa, M. Sierra, A. Ocampo, C. Miralles, R. Rodriguez-Perez, J. Diz-Aren, R. Ojea-De Castro, E. Losada, M.T. Cuevas, E. Vazquez-De Parga, R. Carmona, L. Perez-Alvarez, L. Medrano, L. Cuevas, J.A. Taboada, R. Najera, N. Manjon, A. Marino and I. Herrero (2002). "Identification of a newly characterized HIV-1 BG intersubtype circulating recombinant form in Galicia, Spain, which exhibits a pseudotype-like virion structure." J Acquir Immune Defic Syndr 29(5): 536-543.

Delwart E., M. Magierowska, M. Royz, B. Foley, L. Peddada, R. Smith, C. Heldebrant, A. Conrad and M. Busch (2002). "Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection." Aids 16(2): 189-195.

Delwart E.L., E.G. Shpaer, J. Louwagie, F.E. Mccutchan, M. Grez, H. Rubsamen-Waigmann and J.I. Mullins (1993). "Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes." Science 262(5137): 1257-1261.

De Oliveira T., M. Salemi, M. Gordon, A.M. Vandamme, E.J. Van Rensburg, S. Engelbrecht, H.M. Coovadia and S. Cassol (2004). "Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design?" Genetics 167(3): 1047-1058.

Descamps D., M.L. Chaix, P. Andre, V.Brodard, J. Cottalorda, C. Deveau, M Harzic, D. Ingrand, J. Izopet, E. Kohli, B. Marquelier, S. Mouajjah, P. Palmer, I. Pellegrin, J.C. Plantier, C. Poggi, S. Rogez, A. Ruffault, V. Schneider, A. Signori-Schmuck, C. Tamalet, M. Virden, C. Rouzioux, F. Brun-Vezinet, L. Meyer, D. Costagliola (2005). "French national sentinel survey of antiretroviral drug resistance in patients with HIV-1 primary infection and in antiretroviral-naive chronically infected patients in 2001-2002." Acquir Immune Defic Syndr 38(5): 545-552.

Descamps D., C. Apetrei, G. Collin, F. Damond, F. Simon and F. Brun-Vezinet (1998). "Naturally occurring decreased susceptibility of HIV-1 subtype G to protease inhibitors." Aids 12(9): 1109-1111.

Des Jarlais D.C., S.R. Friedman, K. Choopanya, S. Vanichseni, and T.P. Ward (1992). "International epidemiology of HIV and AIDS among injecting drug users." AIDS 6(10): 1053-1068.

Detels R., A. Munoz, G. Mcfarlane, L.A. Kingsley, J.B. Margolick, J. Giorgi, L.K. Schrager and J.P. Phair (1998). "Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. Multicenter AIDS Cohort Study Investigators." Jama 280(17): 1497-1503.

Domingo E. and J.J. Holland (1997). "RNA virus mutations and fitness for survival." Annu Rev Microbiol 51, 151– 178.

Drake J.W. (1999). “The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes.” Ann N Y Acad Sci 870: 100-107.

Drake J.W., B. Charlesworth, D. Charlesworth and J.F. Crow (1998). “Rates of spontaneous mutation.” Genetics 148(4): 1667-86.

Drummond A.J., O.G. Pybus A. Rambaut (2003a). “Inference of viral evolutionary rates from molecular sequences.” Adv Parasitol 54:331-358.

Drummond A.J., O.G. Pybus, A. Rambaut, R. Forsberg and A.G. Rodrigo (2003b). "Measurably evolving populations." Trends Ecol. Evol. 18(9): 481-488.

Drummond A. and A. Rambaut (2003). BEAST v1.0, available from *http://evolve.zoo.ox.ac.uk/beast/.*

Drummond A.J., G.K. Nicholls, A.G. Rodrigo and W. Solomon (2002). "Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data." Genetics 161(3): 1307-1320.

Duarte E., D. Clarke, A. Moya, E. Domingo, and J. Holland (1992). “Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet.” Proc Natl Acad Sci USA 89(13): 6015–6019.

Dubois R.M., M.A. Braitwaite, J.R. Mikhail and J.C. Batten(1981). “Primary Pneumocystis Carinii and Cytomegalovirus Infections.” Lancet 2(8259): 1339.

Durant J., P. Clevenbergh, R. Garraffo, P. Halfon, S. Icard, P. Del Giudice, N. Montagne, J.M. Schapiro and P. Dellamonica (2000). "Importance of protease inhibitor plasma levels in HIV-infected patients treated with genotypic-guided therapy: pharmacological data from the Viradapt Study." Aids 14(10): 1333-1339.

Eck R.V. and M.O. Dayhoff (1966). Atlas of protein Sequence and Structure. Silver Spring, MD.

Elswood B.F. and R.B. Stricker (1994). “Polio vaccines and the origin of AIDS.” Med Hypotheses 42(6): 347-354.

Esparza J. and N. Bhamarapravati (2000). "Accelerating the development and future availability of HIV-1 vaccines: why, when, where, and how?" Lancet 355(9220): 2061-2066.

Fang G., B. Weiser, C. Kuiken, S.M. Philpott, S. Rowland-Jones, F. Plummer, J. Kimani, B. Shi, R. Kaul, J. Bwayo, O. Anzala and H. Burger (2004). "Recombination following superinfection by HIV-1." Aids 18(2): 153-159.

Fearnhead P. and P. Donnelly (2001). “Estimating recombination rates from population genetic data.” Genetics 159(3): 1299-1318.

Feinberg M.B., D. Baltimore and A.D. Frankel (1991). “The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation.” Proc Natl Acad Sci USA 88: 4045-4049.

Felsenstein J. (1985). "Confidence limits on phylogenies: An approach using the bootstrap." Evolution 39: 783-791.

Felsenstein J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." J Mol Evol 17(6): 368-376.

Felsenstein J. (1978). “Cases in which Parsimony or Compatibility Methods Will be Positively Misleading.” Syst Zool 27(4): 401-410.

Felsenstein J. (1974). “The evolutionary advantage of recombination.” Genetics 78:737-756.

Felsenstein J. (1973). "Maximum-likelihood estimation of evolutionary trees from continuous characters." Am J Hum Genet 25(5): 471-492.

Feng S. and E.C. Holland (1988). "HIV-1 tat trans-activation requires the loop sequence within tar." Nature 334: 165-167.

Fidler S., A. Oxenius, M. Brady, J. Clarke, I. Cropley, A. Babiker, H.-T. Zhang, D. Price, R. Phillips and J. Weber (2002). "Virological and immunological effects of short-course antiretroviral therapy in primary HIV infection." AIDS 16(15): 2049-2054.

Fitch W.M. (1977). "On the problem of discovering the most parsimonious tree." Am. Nat. 111: 223-257.

Fitch W.M. and E. Margoliash (1967). "Construction of phylogenetic trees." Science 155 (760) 279–284.

Fraser C., N.M. Ferguson and R.M. Anderson (2001). "Quantification of intrinsic residual viral replication in treated HIV-infected patients." Proc Natl Acad Sci U S A 98(26): 15167-15172. Epub 12001 Dec 15111.

Frankel A.D. and J.A. Young (1998). "HIV-1: fifteen proteins and an RNA." Annu Rev Biochem 67: 1-25.

Frenkel L.M., Y. Wang, G.H. Learn, J.L. Mckernan, G.M. Ellis, K.M. Mohan, S.E. Holte, S.M. De Vange, D.M. Pawluk, A.J. Melvin, P.F. Lewis, L.M. Heath, I.A. Beck, M. Mahalanabis, W.E. Naugler, N.H. Tobin and J.I. Mullins (2003). "Multiple viral genetic analyses detect low-level human immunodeficiency virus type 1 replication during effective highly active antiretroviral therapy." J Virol 77(10): 5721-5730.

Frost S.D., H.F. Gunthard, J.K. Wong, D. Havlir, D.D. Richman and A.J. Leigh Brown (2001). "Evidence for positive selection driving the evolution of HIV-1 env under potent antiviral therapy." Virology 284(2): 250-258.

Frost S.D., M. Nijhuis, R. Schuurman, C.A. Boucher and A.J. Brown (2000). "Evolution of lamivudine resistance in human immunodeficiency virus type 1-

infected individuals: the relative roles of drift and selection." J Virol 74(14): 6262-6268.

Gail M.H., W.Y. Tan, D. Pee and J.J. Goedert (1997). "Survival after AIDS diagnosis in a cohort of hemophilia patients. Multicenter Hemophilia Cohort Study." J Acquir Immune Defic Syndr Hum Retrovirol 15(5): 363-369.

Gale C.V., R. Myers, R.S. Tedder, I.G. Williams and P. Kellam (2004). "Development of a novel human immunodeficiency virus type 1 subtyping tool, Subtype Analyzer (STAR): analysis of subtype distribution in London." AIDS Res Hum Retroviruses
20(5): 457-464.

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 and Et Al. (1984). "Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS." Science 224(4648): 500-503.

Gallo R.C., P.S. Sarin, E.P. Gelmann, M. Robert-Guroff, E. Richardson, V.S. Kalyanaraman, D. Mann, G.D. Sidhu, R.E. Stahl, S. Zolla-Pazner, J. Leibowitch and M. Popovic (1983). "Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)." Science 220(4599): 865-867.

Gao F., Y. Chen, D.N. Levy, J.A. Conway, T.B. Kepler and H. Hui (2004). "Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious." J Virol 78(5): 2426-33.

Gao F., E. Bailes, D.L. Robertson, Y. Chen, C.M. Rodenburg, S.F. Michael, L.B. Cummins, L.O. Arthur, M. Peeters, G.M. Shaw, P.M. Sharp and B.H. Hahn (1999). "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes." Nature 397(6718): 436-441.

Gao F., D.L. Robertson, C.D. Carruthers, Y. Li, E. Bailes, L.G. Kostrikis, M.O. Salminen, F. Bibollet-Ruche, M. Peeters, D.D. Ho, G.M. Shaw, P.M. Sharp and B.H. Hahn (1998). "An isolate of human immunodeficiency virus type 1 originally classified as subtype I represents a complex mosaic comprising three different group M subtypes (A, G, and I)." J Virol 72(12): 10234-10241.

Gaschen B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B.H. Hahn, T. Bhattacharya and B. Korber (2002). "Diversity considerations in HIV-1 vaccine selection." Science 296(5577): 2354-2360.

Gabiano C., P.A. Tovo, M. de Martino, L. Galli, C. Giaquinto, A. Loy, M.C. Schoeller, M. Giovannini, G. Ferranti, L. Rancilio , et al (1992). "Mother-to-child transmission of human immunodeficiency virus type 1: risk of infection and correlates of transmission." Pediatrics 90(3): 369-374.

Glazko G.V. and M. Nei (2003). "Estimation of divergence times for major lineages of primate species." Mol Biol Evol 20(3): 424-434.

Glynn J.R., M. Carael, B. Auvert, M. Kahindo, J. Chege, R. Musonda, F. Kaona and A. Buve (2001). "Why do young women have a much higher prevalence of HIV than young men? A study in Kisumu, Kenya and Ndola, Zambia." Aids 15 Suppl 4: S51-60.

Goldman N. (1993). "Statistical tests of models of DNA substitution." J Mol Evol 36(2): 182-198.

Gomez-Cano M., A. Rubio, T. Puig, M. Perez-Olmeda, L. Ruiz, V. Soriano, J.A. Pineda, L. Zamora, N. Xaus, B. Clotet and M. Leal (1998). "Prevalence of genotypic resistance to nucleoside analogues in antiretroviral-naive and antiretroviral-experienced HIV-infected patients in Spain." Aids 12(9): 1015-1020.

Goodrich D.W. and P.H. Duesberg (1990). "Retroviral recombination during reverse transcription." Proc Natl Acad Sci U S A 87(6): 2052-2056.

Gottlieb M.S., R. Schroff, H.M. Schanker, J.D. Weisman, P.T. Fan, R.A. Wolf and A. Saxon (1981). "Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency." N Engl J Med 305(24): 1425-1431.

Goudsmit J., A. De Ronde, E. De Rooij and R. De Boer (1997). "Broad spectrum of in vivo fitness of human immunodeficiency virus type 1 subpopulations differing at reverse transcriptase codons 41 and 215." J Virol 71(6): 4479-4484.

Grassly N.C., P.H. Harvey and E.C. Holmes (1999). "Population dynamics of HIV-1 inferred from gene sequences." Genetics 151(2): 427-438.

Greco R.S. (1983). "Haiti and the stigma of AIDS." Lancet 2(8348): 515-516.

Griffiths R.C. and P. Marjoram (1996). “Ancestral inference from samples of DNA sequences with recombination.” J Comput Biol 3(4): 479-502.

Griffiths R.C. and S. Tavare (1994). "Sampling theory for neutral alleles in a varying environment." Philos Trans R Soc Lond B Biol Sci 344(1310): 403-410.

Gupta K.K. (1993). "Acute immunosuppression with HIV seroconversion." N Engl J Med 328(4): 288-289.

Gurtler L.G., L. Zekeng, J.M. Tsague, A. Van Brunn, E. Afane Ze, J. Eberle and L. Kaptue (1996). "HIV-1 subtype O: epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV." Arch Virol Suppl 11: 195-202.

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(3): 1581-1585.

Hahn B.H., G.M. Shaw, K.M. De Cock and P.M. Sharp (2000). "AIDS as a zoonosis: scientific and public health implications." Science 287(5453): 607-614.

Hahn B.H., M.A. Gonda, G.M. Shaw, M. Popovic, J.A. Hoxie, R.C. Gallo and F. Wong-Staal (1985). “Genomic diversity of the acquired immune deficiency syndrome virus HTLV-III: different viruses exhibit greatest divergence in their envelope genes.” Proc Natl Acad Sci USA 82 (14), 4813–4817.

Hall T.A. (2000). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symp Series 41: 95-98.

Hamers F.F. and A.M. Downs (2004). "The changing face of the HIV epidemic in western Europe: what are the implications for public health policies?" Lancet 364(9428): 83-94.

Hardy A.M., J.R. Allen, W.M. Morgan and J.W. Curran (1985). "The incidence rate of acquired immunodeficiency syndrome in selected populations." Jama 253(2): 215-220.

Harrigan P. R., S. Bloor and B. A. Larder (1998). “Relative replicative fitness of zidovudine-resistant human immunodeficiency virus type 1 isolates in vitro.” J Virol 72(5): 3773-3778.

Harrison G.P., M.S. Mayo, E. Hunter, A.M. Lever (1998). “Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5' and 3' of the catalytic site.” Nucleic Acids Res 26(14): 3433-3442.

Hasegawa M., H. Kishino and T. Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." J Mol Evol 22(2): 160-174.

Hastings W.K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." Biometrika 57: 97-109.

Hayman A., T. Moss, G. Simmons, C. Arnold, E.C. Holmes, L. Naylor-Adamson, J. Hawkswell, K. Allen, J. Radford, J. Nguyen-Van-Tam and P. Balfe (2001).

"Phylogenetic analysis of multiple heterosexual transmission events involving subtype B of HIV type 1." AIDS Res Hum Retroviruses 17(8): 689-695.

Haynes B.F., G. Pantaleo and A.S. Fauci (1996). "Toward an understanding of the correlates of protective immunity to HIV infection." Science 271(5247): 324-328.

Heinzinger N.K., M.I. Bukinsky, S.A. Haggerty, A.M. Ragland, V. Kewalramani, M.A. Lee, H.E. Gendelman, L. Ratner, M. Stevenson and M. Emerman (1994). "The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells." Proc Natl Acad Sci U S A 91(15): 7311-7315.

Hierholzer M., R.R. Graham, I. El Khidir, S. Tasker, M. Darwish, G.D. Chapman, A.H. Fagbami, A. Soliman, D.L. Birx, F. Mccutchan and J.K. Carr (2002). "HIV type 1 strains from East and West Africa are intermixed in Sudan." AIDS Res Hum Retroviruses 18(15): 1163-1166.

Hillis D.M.C., C. Moritz and B.K. Mable (1996). Applications of molecular systematics. Molecular systematics. B.K. Mable, Sinauer Associates, Sunderland, MA.

Hillis D.M. and J.P. Huelsenbeck (1994). "Support for dental HIV transmission." Nature 369(6475): 24-25.

Hirsch M.S., F. Brun-Vezinet, R.T. D'aquila, S.M. Hammer, V.A. Johnson, D.R. Kuritzkes, C. Loveday, J.W. Mellors, B. Clotet, B. Conway, L.M. Demeter, S. Vella, D.M. Jacobsen and D.D. Richman (2000). "Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel." Jama 283(18): 2417-2426.

Hirsch V.M., R.A. Olmsted, M. Murphey-Corb, R.H. Purcell and P.R. Johnson (1989). "An African primate lentivirus (SIVsm) closely related to HIV-2." Nature 339(6223): 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(6510): 123-126.

Holland, J. J., K. Spindler, F. Horodyski, E. Grabau, S. Nichol, S. VandePol (1982). "Rapid evolution of RNA genomes." *Science* 215, 1577–1585.

Holmes E.C., S. Nee, A. Rambaut, G.P. Garnett and P.H. Harvey (1995). "Revealing the history of infectious disease epidemics through phylogenetic trees." Philos Trans R Soc Lond B Biol Sci 349(1327): 33-40.

Holmes E.C., A.J. Brown and P. Simmonds (1993). "Sequence data as evidence." Nature 364(6440): 766.

Holmes E.C., L.Q. Zhang, P. Simmonds, C.A. Ludlam and A.J. Brown (1992). "Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient." Proc Natl Acad Sci U S A 89(11): 4835-4839.

Holmes E.C. and Zanotto P.M. (1998). "Genetic drift of human immunodeficiency virus type 1?" J Virol 72(1): 886-887.

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(6): 1411-1414.

Hopper E. (2001). "Experimental oral polio vaccines and acquired immune deficiency syndrome." Philos Trans R Soc Lond B Biol Sci 356(1410): 803-814.

Hooper E. (2000). The river: a journey back to the source of HIV and AIDS. Penguin Books, London.

Hu D.J., J. Baggs, R.G. Downing, D. Pieniazek, J. Dorn, C. Fridlund, B. Biryahwaho, S.D. Sempala, M.A. Rayfield, T.J. Dondero and R. Lal (2000). "Predominance of HIV-1 subtype A and D infections in Uganda." Emerg Infect Dis 6(6): 609-615.

Hu D.J., A. Buve, J. Baggs, G. Van Der Groen and T.J. Dondero (1999). "What role does HIV-1 subtype play in transmission and pathogenesis? An epidemiological perspective." Aids 13(8): 873-881.

Hu W.S. and H.S. Temin (1990). “Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination.” Proc Natl Acad Sci USA 87(4): 1556-1560.

Huang H., R. Chopra, G.L. Verdine and S.C. Harrison (1998). "Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance." Science 282(5394): 1669-1675.

Hubner A., M. Kruhoffer F. Grosse and G. Krauss (1992). “Fidelity of human immunodeficiency virus type I reverse transcriptase in copying natural RNA.” J Mol Biol 223(3): 595–600.

Hudson R.R. (1983). "Properties of a neutral allele model with intragenic recombination." Theor Popul Biol 23(2): 183-201.

Hué S., D. Pillay, J.P. Clewley and O.G. Pybus (2005). “Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups.” Proc Natl Acad Sci USA 102(12): 4425-4429.

Hué S., J.P. Clewley, C. P.A. and D. Pillay (2004). "HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy." AIDS 18: 719-728.

Huelsenbeck J.P. (2000). Mr Bayes: Bayesian inferences of phylogeny. University of Rochester, N.Y.

Huelsenbeck J.P., B. Larget and D. Swofford (2000). "A compound poisson process for relaxing the molecular clock." Genetics 154(4): 1879-1892.

Huet 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.

Hussein M., A. Abebe, G. Pollakis, M. Brouwer, B. Petros, A.L. Fontanet and T.F. Rinke De Wit (2000). "HIV-1 subtype C in commerical sex workers in Addis Ababa, Ethiopia." J Acquir Immune Defic Syndr 23(2): 120-127.

Huxley J. S. (1942). Evolution: The Modern Synthesis. Allen and Unwin Ltd, London.

Ibanez A., B. Clotet and M.A. Martinez (2000). "Human immunodeficiency virus type 1 population bottleneck during indinavir therapy causes a genetic drift in the env quasispecies." J Gen Virol 81(Pt 1): 85-95.

Jacks T., M.D. Power, F.R. Masiarz, P.A. Luciw, P.J. Barr and H.E. Varmus (1988). "Characterization of ribosomal frameshifting in HIV-1 gag-pol expression." Nature 331(6153): 280-283.

Jacquez J.A., J.S. Koopman, C.P. Simon and I.M. Longini, Jr. (1994). "Role of the primary infection in epidemics of HIV infection in gay cohorts." J Acquir Immune Defic Syndr 7(11): 1169-1184.

Jaffe H.W., D.J. Bregman and R.M. Selik (1983). "Acquired immune deficiency syndrome in the United States: the first 1,000 cases." J Infect Dis 148(2): 339-345.

Janssen R.S., G.A. Satten, S.L. Stramer, B.D. Rawal, T.R. O'brien, B.J. Weiblen, F.M. Hecht, N. Jack, F.R. Cleghorn, J.O. Kahn, M.A. Chesney and M.P. Busch (1998). "New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes." Jama 280(1): 42-48.

Janssens W., L. Heyndrickx, K. Fransen, J. Motte, M. Peeters, J.N. Nkengasong, P.M. Ndumbe, E. Delaporte, J.L. Perret, C. Atende and Et Al. (1994). "Genetic and phylogenetic analysis of env subtypes G and H in central Africa." AIDS Res Hum Retroviruses 10(7): 877-879.

Jenkins G.M., A. Rambaut, O.G. Pybus and E.C. Holmes (2002). "Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis." J Mol Evol 54(2): 156-165.

Jetzt A.E., H. Yu, G.J. Klarmann, Y. Ron, B.D. Preston and J.P. Dougherty (2000). "High rate of recombination throughout the human immunodeficiency virus type 1 genome." J Virol 74(3): 1234-1240.

Johnson V.A., F. Brun-Vezinet, B. Clotet, B. Conway, R.T. D'aquila, L.M. Demeter, D.R. Kuritzkes, D. Pillay, J.M. Schapiro, A. Telenti and D.D. Richman (2003). "Drug resistance mutations in HIV-1." Top HIV Med 11(6): 215-221.

Johnson A.M., C.H. Mercer, B. Erens, A.J. Copas, S. Mcmanus, K. Wellings, K.A. Fenton, C. Korovessis, W. Macdowall, K. Nanchahal, S. Purdon and J. Field (2001). "Sexual behaviour in Britain: partnerships, practices, and HIV risk behaviours." Lancet 358(9296): 1835-1842.

Johnson W.D., Jr. and J.W. Pape (1989). "AIDS in Haiti." Immunol Ser 44: 65-78.

Joy D.A., X. Feng, J. Mu, T. Furuya, K. Chotivanich, A.U. Krettli, M. Ho, A. Wang, N.J. White, E. Suh, P. Beerli and X.Z. Su (2003). "Early origin and recent expansion of Plasmodium falciparum." Science 300(5617): 318-321.

Jukes T.H. and C.R. Cantor (1969). Evolution of protein molecules. Mammalian Protein Metabolism III. H.N. Munro, Academic Press: New York: 21-132.

Jung A., R. Maier, J.P. Vartanian, G. Bocharov, V. Jung, U. Fischer, E. Meese, S. Wain-Hobson and A. Meyerhans (2002). "Multiply infected spleen cells in HIV patients." Nature 418(6894): 144.

Junghans R.P., L.R. Boone and A.M. Skalka (1982). "Retroviral DNA H structures: displacement-assimilation model of recombination." Cell 30(1): 53-62.

Kahn J.O. and B.D. Walker (1998). "Acute human immunodeficiency virus type 1 infection." N Engl J Med 339(1): 33-39.

Kaplan E.H. (1989). "Needles that kill: modeling human immunodeficiency virus transmission via shared drug injection equipment in shooting galleries." Rev Infect Dis 11(2): 289-298.

Kato K., S. Kusagawa, K. Motomura, R. Yang, T. Shiino, K. Nohtomi, H. Sato, K. Shibamura, T.H. Nguyen, K.C. Pham, H.T. Pham, C.T. Duong, D.T. Bui, T.L. Hoang, Y. Nagai and Y. Takebe (2001). "Closely related HIV-1 CRF01_AE variant among injecting drug users in northern Vietnam: evidence of HIV spread across the Vietnam-China border." AIDS Res Hum Retroviruses 17(2): 113-123.

Katz R. A. and A. M. Skalka (1990). Generation of diversity in retroviruses. Annu. Rev. Genet. 24:409–445.

Kaufmann G.R., P. Cunningham, A.D. Kelleher, J. Zaunders, A. Carr, J. Vizzard, M. Law and D.A. Cooper (1998). "Patterns of viral dynamics during primary human immunodeficiency virus type 1 infection. The Sydney Primary HIV Infection Study Group." J Infect Dis 178(6): 1812-1815.

Kessler H.H., D. Deuretzbacher, E. Stelzl, E. Daghofer, B.I. Santner and E. Marth (2001). "Determination of human immunodeficiency virus type 1 subtypes by a rapid method useful for the routine diagnostic laboratory." Clin Diagn Lab Immunol 8(5): 1018-1020.

Kilbourne E.D. (1973). "The molecular epidemiology of influenza." J Infect Dis 127(4): 478-487.

Kim T., R.A. Mudry Jr., C.A. Rexrode II, and V. K. Pathak (1996). “Retroviral mutation rates and A-to-G hypermutations during different stages of retroviral replication.” J Virol 70(11): 7594–7602.

Kimura M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." J Mol Evol 16(2): 111-120.

Kimura M. (1968). "Evolutionary rate at the molecular level." Nature 217(129): 624-626.

King J.L. and T.H. Jukes (1969). "Non-Darwinian evolution." Science 164(881): 788-798.

Kingman J.F. (1982a). "The coalescent." Stoch. Process. Their Appl. 13: 235-248.

Kingman J.F. (1982b). "On the genealogy of large populations." J. Appl. Probab. 19: 27-43.

Kitrinos K.M., J.A. Nelson, W. Resch and R. Swanstrom (2005). “Effect of a protease inhibitor-induced genetic bottleneck on human immunodeficiency virus type 1 env gene populations.” J Virol 79(16): 10627-10637.

Kleim J.P., A. Ackermann, H.H. Brackmann, M. Gahr and K.E. Schneweis (1991). "Epidemiologically closely related viruses from hemophilia B patients display high homology in two hypervariable regions of the HIV-1 env gene." AIDS Res Hum Retroviruses 7(4): 417-421.

Koning F.A., C.A. Jansen, J. Dekker, R.A. Kaslow, N. Dukers, D. Van Baarle, M. Prins and H. Schuitemaker (2004). "Correlates of resistance to HIV-1 infection in homosexual men with high-risk sexual behaviour." Aids 18(8): 1117-1126.

Koopman J.S., J.A. Jacquez, G.W. Welch, C.P. Simon, B. Foxman, S.M. Pollock, D. Barth-Jones, A.L. Adams and K. Lange (1997). "The role of early HIV infection in the spread of HIV through populations." J Acquir Immune Defic Syndr Hum Retrovirol 14(3): 249-258.

Korber B., B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir and V. Detours (2001). “Evolutionary and immunological implications of contemporary HIV-1 variation.” Br Med Bull 58:19-42.

Korber B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B.H. Hahn, S. Wolinsky and T. Bhattacharya (2000). "Timing the ancestor of the HIV-1 pandemic strains." Science 288(5472): 1789-1796.

Kothe D., R.H. Byers, S.P. Caudill, G.A. Satten, R.S. Janssen, W.H. Hannon and J.V. Mei (2003). "Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence." J Acquir Immune Defic Syndr 33(5): 625-634.

Koulinska I.N., T. Ndung'u, D. Mwakagile, G. Msamanga, C. Kagoma, W. Fawzi, M. Essex and B. Renjifo (2001). "A new human immunodeficiency virus type 1 circulating recombinant form from Tanzania." AIDS Res Hum Retroviruses 17(5): 423-431.

Kuhner M.K., J. Yamato and J. Felsenstein (1995). "Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling." Genetics 140(4): 1421-1430.

Kuiken C., B. Foley, B.H. Hahn, F. Mccutchan, J. Mellors, W., J.I. Mullins, S. Wolinsky and B. Korber (2002). The 2002 HIV Sequence Compendium. Los Alamos, New Mexico.

Kuiken C., R. Thakallapalli, A. Esklid and A. De Ronde (2000). "Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus." Am J Epidemiol 152(9): 814-822.

Kumar S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." Proc Natl Acad Sci U S A 99(2): 803-808. Epub 2002 Jan 2015.

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(8957): 1078-1083.

Kurbanov F., M. Kondo, Y. Tanaka, M. Zalalieva, G. Giasova, T. Shima, N. Jounai, N. Yuldasheva, R. Ruzibakiev, M. Mizokami and M. Imai (2003). "Human immunodeficiency virus in Uzbekistan: epidemiological and genetic analyses." AIDS Res Hum Retroviruses 19(9): 731-738.

Lackritz E.M., G.A. Satten, J. Aberle-Grasse, R,Y, Dodd, V.P. Raimondi, R.S. Janssen, W.F. Lewis and E.P. Notari 4th, L.R. Peterson (1995). "Estimated risk of transmission of the human immunodeficiency virus by screened blood in the United States. N Engl J Med 333(26): 1721-1725.

Larder B.A. and S.D. Kemp (1989). "Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT)." Science 246(4934): 1155-1158.

Laukkanen T., J.K. Carr, W. Janssens, K. Liitsola, D. Gotte, F.E. Mccutchan, E. Op De Coul, M. Cornelissen, L. Heyndrickx, G. Van Der Groen and M.O. Salminen (2000). "Virtually full-length subtype F and F/D recombinant HIV-1 from Africa and South America." Virology 269(1): 95-104.

Leal E.D.S., E.C. Holmes and P.M. Zanotto (2004). "Distinct patterns of natural selection in the reverse transcriptase gene of HIV-1 in the presence and absence of antiretroviral therapy." Virology 325(2): 181-191.

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 U S A 94(5): 1862-1865.

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." Virology 235(1): 166-177.

Leigh Brown A.J. and D.D. Richman (1997). "HIV-1: gambling on the evolution of drug resistance?" Nat Med 3(3): 268-271.

Leitner T. and J. Albert (1999). "The molecular clock of HIV-1 unveiled through analysis of a known transmission history." Proc Natl Acad Sci U S A 96(19): 10752-10757.

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 U S A 93(20): 10864-10869.

Lemey P., O.G. Pybus, A. Rambaut, A.J. Drummond, D.L. Robertson, P. Roques, M. Worobey and A.M. Vandamme (2004). “The molecular population genetics of HIV-1 group O. Genetics.” 167(3): 1059-68.

Lemey P., O.G. Pybus, B. Wang, N.K. Saksena, M. Salemi and A.M. Vandamme (2003). "Tracing the origin and history of the HIV-2 epidemic." Proc Natl Acad Sci U S A 100(11): 6588-6592.

Li W.H. (1997). Molecular evolution, Sinaur Associates, Sunderland, MA.

Li W.H. and D. Graur (2000). Fundamentals of molecular evolution, second edition. Sunderland, Massachusetts.

Li W.H., M. Tanimura and P.M. Sharp (1988). "Rates and dates of divergence between AIDS virus nucleotide sequences." Mol Biol Evol 5(4): 313-330.

Liitsola K., I. Tashkinova, T. Laukkanen, G. Korovina, T. Smolskaja, O. Momot, N. Mashkilleyson, S. Chaplinskas, H. Brummer-Korvenkontio, J. Vanhatalo, P. Leinikki and M.O. Salminen (1998). "HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad." Aids 12(14): 1907-1919.

Linnaeus C. (1758): "Systemata Naturae." 10th ed. Stockholm.

Lole K.S., R.C. Bollinger, R.S. Paranjape, D. Gadkari, S.S. Kulkarni, N.G. Novak, R. Ingersoll, H.W. Sheppard and S.C. Ray (1999). "Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination." J Virol 73(1): 152-160.

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 Et Al. (1993). "Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes." Aids 7(6): 769-780.

Lukashov V.V. and J. Goudsmit (2002). "Recent evolutionary history of human immunodeficiency virus type 1 subtype B: reconstruction of epidemic onset based on sequence distances to the common ancestor." J Mol Evol 54(5): 680-691.

Machuca R., L.B. Jorgensen, P. Theilade and C. Nielsen (2001). "Molecular investigation of transmission of human immunodeficiency virus type 1 in a criminal case." Clin Diagn Lab Immunol 8(5): 884-890.

Maddison W.P. and D.R. Maddison (1989). "Interactive analysis of phylogeny and character evolution using the computer program MacClade." Folia Primatol (Basel) 53(1-4): 190-202.

Mansky L. M., and H.M. Temin (1995). "Lower in-vivo mutation-rate of human immunodeficiency-virus type-1 than that predicted from the fidelity of purified reversetranscriptase." J Virol. 69: 5087-5094.

Marx P.A., P.G. Alcabes and E. Drucker (2001). "Serial human passage of simian immunodeficiency virus by unsterile injections and the emergence of epidemic human immunodeficiency virus in Africa." Philos Trans R Soc Lond B Biol Sci (1410): 911-920.

Martinez-Picado J., A.V. Savara, L. Sutton and R.T. D'aquila (1999). "Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1." J Virol 73(5): 3744-3752.

Masur H., M.A. Michelis, J.B. Greene, I. Onorato, R.A. Stouwe, R.S. Holzman, G. Wormser, L. Brettman, M. Lange, H.W. Murray and S. Cunnigham-Rundles (1981). "An Outbreak of community acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction." N Engl J Med 305(24): 1431-1438.

Mau B., M.A. Newton and B. Larget (1999). "Bayesian phylogenetic inference via Markov chain Monte Carlo methods." Biometrics 55(1): 1-12.

Mccormick A., H. Tillett, B. Bannister and J. Emslie (1987). "Surveillance of AIDS in the United Kingdom." Br Med J (Clin Res Ed) 295(6611): 1466-1469.

Mccutchan F.E. (2000). "Understanding the genetic diversity of HIV-1." Aids 14(Suppl 3): S31-44.

Mccutchan F.E., J.L. Sankale, S. M'boup, B. Kim, S. Tovanabutra, D.J. Hamel, S.K. Brodine, P.J. Kanki and D.L. Birx (2004). "HIV type 1 circulating recombinant form CRF09_cpx from west Africa combines subtypes A, F, G, and may share ancestors with CRF02_AG and Z321." AIDS Res Hum Retroviruses 20(8): 819-826.

Mccutchan F.E., J.K. Carr, D. Murphy, S. Piyasirisilp, F. Gao, B. Hahn, X.F. Yu, C. Beyrer and D.L. Birx (2002). "Precise mapping of recombination breakpoints suggests a common parent of two BC recombinant HIV type 1 strains circulating in China." AIDS Res Hum Retroviruses 18(15): 1135-1140.

Mccutchan F.E., J.K. Carr, M. Bajani, E. Sanders-Buell, T.O. Harry, T.C. Stoeckli, K.E. Robbins, W. Gashau, A. Nasidi, W. Janssens and M.L. Kalish (1999). "Subtype G and multiple forms of A/G intersubtype recombinant human immunodeficiency virus type 1 in Nigeria." Virology 254(2): 226-234.

Mcfarland W., M.P. Busch, T.A. Kellogg, B.D. Rawal, G.A. Satten, M.H. Katz, J. Dilley and R.S. Janssen (1999). "Detection of early HIV infection and estimation of incidence using a sensitive/less-sensitive enzyme immunoassay testing strategy at anonymous counseling and testing sites in San Francisco." J Acquir Immune Defic Syndr 22(5): 484-489.

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 U S A 89(21): 10247-10251.

Mellors J.W., L.A. Kingsley, C.R. Rinaldo, Jr., J.A. Todd, B.S. Hoo, R.P. Kokka and P. Gupta (1995). "Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion." Ann Intern Med 122(8): 573-579.

Menendez-Arias L., M.A. Martinez, M.E. Quinones-Mateu and J. Martinez-Picado (2003). "Fitness variations and their impact on the evolution of antiretroviral drug resistance." Curr Drug Targets Infect Disord 3(4): 355-371.

Metropolis N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller (1953). "Equations of states calculations by fast computing machines." Journal of Chemical Physics 21: 1087-1091.

Metzker M.L., D.P. Mindell, X.M. Liu, R.G. Ptak, R.A. Gibbs and D.M. Hillis (2002). "Molecular evidence of HIV-1 transmission in a criminal case." Proc Natl Acad Sci U S A 99(22): 14292-14297.

Mocroft A., S. Vella, T.L. Benfield, A. Chiesi, V. Miller, P. Gargalianos, A. D'arminio Monforte, I. Yust, J.N. Bruun, A.N. Phillips and J.D. Lundgren (1998).

"Changing patterns of mortality across Europe in patients infected with HIV-1. EuroSIDA Study Group." Lancet 352(9142): 1725-1730.

Montavon C., L. Vergne, A. Bourgeois, E. Mpoudi-Ngole, G. Malonga-Mouellet, C. Butel, C. Toure-Kane, E. Delaporte and M. Peeters (2002). "Identification of a new circulating recombinant form of HIV type 1, CRF11-cpx, involving subtypes A, G, J, and CRF01-AE, in Central Africa." AIDS Res Hum Retroviruses 18(3): 231-236.

Montavon C., C. Toure-Kane, F. Liegeois, E. Mpoudi, A. Bourgeois, L. Vergne, J.L. Perret, A. Boumah, E. Saman, S. Mboup, E. Delaporte and M. Peeters (2000). "Most env and gag subtype A HIV-1 viruses circulating in West and West Central Africa are similar to the prototype AG recombinant virus IBNG." J Acquir Immune Defic Syndr 23(5): 363-374.

Montavon C., F. Bibollet-Ruche, D. Robertson, B. Koumare, C. Mulanga, E. Esu-Williams, C. Toure, S. Mboup, E. Saman, E. Delaporte and M. Peeters (1999). "The identification of a complex A/G/I/J recombinant HIV type 1 virus in various West African countries." AIDS Res Hum Retroviruses 15(18): 1707-1712.

Muller H. J. (1964). "The relation of recombination to mutational advance." Mutat Res 106: 2-9.

Muller H.J. (1934). "Some genetic aspects of sex." American Naturalist 66: 118.

Murphy E., B. Korber, M.C. Georges-Courbot, B. You, A. Pinter, D. Cook, M.P. Kieny, A. Georges, C. Mathiot, F. Barre-Sinoussi and Et Al. (1993). "Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic." AIDS Res Hum Retroviruses 9(10): 997-1006.

Murphy G., A. Charlett, L.F. Jordan, N. Osner, O.N. Gill and J.V. Parry (2004). "HIV incidence appears constant in men who have sex with men despite widespread use of effective antiretroviral therapy." Aids 18(2): 265-272.

Nahmias A.J., J. Weiss, X. Yao, F. Lee, R. Kodsi, M. Schanfield, T. Matthews, D. Bolognesi, D. Durack, A. Motulsky and Et Al. (1986). "Evidence for human infection with an HTLV III/LAV-like virus in Central Africa, 1959." Lancet 1(8492): 1279-1280.

Najera I., D.D. Richman, I. Olivares, J.M. Rojas, M.A. Peinado, M. Perucho, R. Najera and C. Lopez-Galindez (1994). "Natural occurrence of drug resistance mutations in the reverse transcriptase of human immunodeficiency virus type 1 isolates." AIDS Res Hum Retroviruses 10(11): 1479-1488.

Nasioulas G., D. Paraskevis, E. Magiorkinis, M. Theodoridou and A. Hatzakis (1999). "Molecular analysis of the full-length genome of HIV type 1 subtype I: evidence of A/G/I recombination." AIDS Res Hum Retroviruses 15(8): 745-758.

Ndung'u T., B. Renjifo, M. Essex, V.A. Novitsky, M.F. Mclane and S. Gaolekwe (2001). "Construction and analysis of an infectious human Immunodeficiency virus type 1 subtype C molecular clone

Ndung'u T., B. Renjifo, V.A. Novitsky, M.F. Mclane, S. Gaolekwe and M. Essex (2000). "Molecular cloning and biological characterization of full-length HIV-1 subtype C from Botswana." Virology 278(2): 390-399.Molecular cloning and biological characterization of full-length HIV-1 subtype C from Botswana." J Virol 75(11): 4964-4972.

Nee S., E.C. Holmes, A. Rambaut and P.H. Harvey (1995). "Inferring population history from molecular phylogenies." Philos Trans R Soc Lond B Biol Sci 349(1327): 25-31.

Negroni M. and H. Buc (2001). “Mechanisms of retroviral recombination.” Annu Rev Genet 35: 275-302.

Nijhuis M., R. Schuurman, D. de Jong, J. Erickson, E. Gustchina, J. Albert, P.

Schipper, S. Gulnik, and C. A. Boucher. 1999. “Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy.” AIDS 13(17): 2349-2359.

Njouom R., C. Pasquier, K. Sandres-Saune, A. Harter, C. Souyris and J. Izopet (2003). "Assessment of HIV-1 subtyping for Cameroon strains using phylogenetic analysis of pol gene sequences." J Virol Methods 110(1): 1-8.

Nkengasong J.N., M. Peeters, P. Zhong, B. Willems, W. Janssens, L. Heyndrickx, K. Fransen, P.M. Ndumbe, G.M. Gershy-Damet, P. Nys and Et Al. (1995). "Biological phenotypes of HIV-1 subtypes A and B strains of diverse origins." J Med Virol 47(3): 278-284.

Nkengasong J.N., W. Janssens, L. Heyndrickx, K. Fransen, P.M. Ndumbe, J. Motte, A. Leonaers, M. Ngolle, J. Ayuk, P. Piot and Et Al. (1994). "Genotypic subtypes of HIV-1 in Cameroon." Aids 8(10): 1405-1412.

Nkengasong J.N., M. Peeters, M. Vanden Haesevelde, S.S. Musi, B. Willems, P.M. Ndumbe, E. Delaporte, J.L. Perret, P. Piot and G. Van Den Groen (1993). "Antigenic evidence of the presence of the aberrant HIV-1ant70 virus in Cameroon and Gabon." Aids 7(11): 1536-1538.

Nuttall G.H.F. (1904). Blood Immunity and Blood relationship. Cambridge University Press Cambridge.

Nuttall G.H.F. (1902). “Progress report upon the biological test for blood as applied to 500 bloods from various sources, together with a prleminary note upon a method for measuring the degree of reaction.” Brit Med 1:825-827.

Oelrichs R.B., V.A. Lawson, K.M. Coates, C. Chatfield, N.J. Deacon and D.A. Mcphee (2000a). "Rapid full-length genomic sequencing of two cytopathically heterogeneous Australian primary HIV-1 isolates." J Biomed Sci 7(2): 128-135.

Oelrichs R.B., I.L. Shrestha, D.A. Anderson and N.J. Deacon (2000b). "The explosive human immunodeficiency virus type 1 epidemic among injecting drug users of Kathmandu, Nepal, is caused by a subtype C virus of restricted genetic diversity." J Virol 74(3): 1149-1157.

Osmanov S., C. Pattou, N. Walker, B. Schwardlander and J. Esparza (2002). "Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000." J Acquir Immune Defic Syndr 29(2): 184-190.

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 and Et Al. (1992). "Molecular epidemiology of HIV transmission in a dental practice." Science 256(5060): 1165-1171.

Page R.D.M. and E.C. Holmes (1998). Molecular evolution, a phylogenetic approach. Blackwell Science Ltd, Oxford.

Palella F.J., Jr., K.M. Delaney, A.C. Moorman, M.O. Loveless, J. Fuhrer, G.A. Satten, D.J. Aschman and S.D. Holmberg (1998). "Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators." N Engl J Med 338(13): 853-860.

Palmer S., D. Vuitton, M.J. Gonzales, A. Bassignot and R.W. Shafer (2002). "Reverse transcriptase and protease sequence evolution in two HIV-1-infected couples." J Acquir Immune Defic Syndr 31(3): 285-290.

Pandrea I., D.L. Robertson, R. Onanga, F. Gao, M. Makuwa, P. Ngari, I. Bedjabaga, P. Roques, F. Simon and C. Apetrei (2002). "Analysis of partial pol and env sequences indicates a high prevalence of HIV type 1 recombinant strains circulating in Gabon." AIDS Res Hum Retroviruses 18(15): 1103-1116.

Pantaleo G., C. Graziosi and A.S. Fauci (1993). "New concepts in the immunopathogenesis of human immunodeficiency virus infection." N Engl J Med 328(5): 327-335.

Pao D., M. Fisher, S. Hué, G. Dean, G. Murphy, P.A. Cane, C.A. Sabin, and D. Pillay (2005). “Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections.” AIDS 19(1): 85-90.

Pape J.W., B. Liautaud, F. Thomas, J.R. Mathurin, M.M. St Amand, M. Boncy, V. Pean, M. Pamphile, A.C. Laroche and W.D. Johnson, Jr. (1983). "Characteristics of the acquired immunodeficiency syndrome (AIDS) in Haiti." N Engl J Med 309(16): 945-950.

Park J., and C. D. Morrow (1991). “Overexpression of the *gag-pol* precursor from human immunodeficiency virus type 1 proviral genomes results in efficient proteolytic processing in the absence of virion production.” J Virol 65: 5111-5117.

Parkin N.T., M. Chamorro and H.E. Varmus (1992). “Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on mRNA secondary structure: Demonstration by expression in vivo.” J Virol 66: 5147-5151.

Parry J.V., G. Murphy, K.L. Barlow, K. Lewis, P.A. Rogers, F.J. Belda, A. Nicoll, C. Mcgarrigle, S. Cliffe, P.P. Mortimer and J.P. Clewley (2001). "National surveillance of HIV-1 subtypes for England and Wales: design, methods, and initial findings." J Acquir Immune Defic Syndr 26(4): 381-388.

Pasquier C., N. Millot, R. Njouom, K. Sandres, M. Cazabat, J. Puel and J. Izopet (2001). "HIV-1 subtyping using phylogenetic analysis of pol gene sequences." J Virol Methods 94(1-2): 45-54.

Pathak V.K. and H.M. Temin (1990). “Broad spectrum of in vivo forward mutations, hypermutations, and mutational hotspots in a retroviral shuttle vector after a single replication cycle: substitutions, frameshifts, and hypermutations.” Proc. Natl. Acad. Sci. USA 87(16): 6019–6023.

Pedersen C., E. Dickmeiss, J. Gaub, L.P. Ryder, P. Platz, B.O. Lindhardt and J.D. Lundgren (1990). "T-cell subset alterations and lymphocyte responsiveness to mitogens and antigen during severe primary infection with HIV: a case series of seven consecutive HIV seroconverters." Aids 4(6): 523-526.

Peeters M. and P.M. Sharp (2000). "Genetic diversity of HIV-1: the moving target." Aids 14(Suppl 3): S129-140.

Peeters M., R. Vincent, J.L. Perret, M. Lasky, D. Patrel, F. Liegeois, V. Courgnaud, R. Seng, T. Matton, S. Molinier and E. Delaporte (1999). "Evidence for differences in MT2 cell tropism according to genetic subtypes of HIV-1: syncytium-inducing variants seem rare among subtype C HIV-1 viruses." J Acquir Immune Defic Syndr Hum Retrovirol 20(2): 115-121.

Pela A. V. and J.J. Platt (1989). “AIDS in Africa: emerging trends.” Soc Sci Med 28: 1-8.

Peliska J.A. and S.J. Benkovic (1992). “Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase.” Science 258(5085): 1112–1118.

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(5255): 1582-1586.

Perno C.F., F.M. Newcomb, D.A. Davis, S. Aquaro, R.W. Humphrey, R. Calio and R. Yarchoan (1998). "Relative potency of protease inhibitors in monocytes/macrophages acutely and chronically infected with human immunodeficiency virus." J Infect Dis 178(2): 413-422.

Pesenti E., C. Pastore, F. Lillo, A.G. Siccardi, D. Vercelli and L. Lopalco (1999). "Role of CD4 and CCR5 levels in the susceptibility of primary macrophages to infection by CCR5-dependent HIV type 1 isolates." AIDS Res Hum Retroviruses 15(11): 983-987.

Peterson P.K., B.M. Sharp, G. Gekker, P.S. Portoghese, K. Sannerud and H.H. Balfour, Jr. (1990). "Morphine promotes the growth of HIV-1 in human peripheral blood mononuclear cell cocultures." Aids 4(9): 869-873.

Phillips A.N. (1992). "CD4 lymphocyte depletion prior to the development of AIDS." Aids 6(7): 735-736.

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(5102): 1749-1754.

Pilcher C.D., J.J. Eron, Jr., S. Galvin, C. Gay and M.S. Cohen (2004a). "Acute HIV revisited: new opportunities for treatment and prevention." J Clin Invest 113(7): 937-945.

Pilcher C.D., H.C. Tien, J.J. Eron, Jr., P.L. Vernazza, S.Y. Leu, P.W. Stewart, L.E. Goh and M.S. Cohen (2004b). "Brief but efficient: acute HIV infection and the sexual transmission of HIV." J Infect Dis 189(10): 1785-1792.

Pilcher C.D., J.J. Eron, Jr., P.L. Vemazza, M. Battegay, T. Harr, S. Yerly, S. Vom and L. Perrin (2001). "Sexual transmission during the incubation period of primary HIV infection." Jama 286(14): 1713-1714.

Pillay D., P.A. Cane, J. Shirley and K. Porter (2000a). "Detection of drug resistance associated mutations in HIV primary infection within the UK." Aids 14(7): 906-908.

Pillay D., S. Taylor and D.D. Richman (2000b). "Incidence and impact of resistance against approved antiretroviral drugs." Rev Med Virol 10(4): 231-253.

Pitcher C.J., C. Quittner, D.M. Peterson, M. Connors, R.A. Koup, V.C. Maino, L.J. Picker. (1999). “HIV-1-specific CD4+ T cells are detectable in most individuals with active HIV-1 infection, but decline with prolonged viral suppression.” Nat Med 5(5): 518-25.

Plotkin S.A. (2001). “Untruths and consequences: the false hypothesis linking CHAT type 1 polio vaccination to the origin of human immunodeficiency virus.” Philos Trans R Soc Lond B Biol Sci 356(1410): 815-823.

Plummer F.A., N.J. Nagelkerke, S. Moses, J.O. Ndinya-Achola, J. Bwayo and E. Ngugi (1991). "The importance of core groups in the epidemiology and control of HIV-1 infection." Aids 5(Suppl 1): S169-176.

Pollakis G., A. Abebe, A. Kliphuis, T.F. De Wit, B. Fisseha, B. Tegbaru, G. Tesfaye, H. Negassa, Y. Mengistu, A.L. Fontanet, M. Cornelissen and J. Goudsmit (2003). "Recombination of HIV type 1C (C'/C") in Ethiopia: possible link of EthHIV-1C' to subtype C sequences from the high-prevalence epidemics in India and Southern Africa." AIDS Res Hum Retroviruses 19(11): 999-1008.

Posada D. (2002). "Evaluation of methods for detecting recombination from DNA sequences: empirical data." Mol Biol Evol 19(5): 708-717.

Posada D., K.A. Crandall and E.C. Holmes (2002). “Recombination in evolutionary genomics.” Annu Rev Genet 36: 75-97.

Posada D. and K.A. Crandall (1998). "MODELTEST: testing the model of DNA substitution." Bioinformatics 14(9): 817-818.

Poss M., A.G. Rodrigo, J.J. Gosink, G.H. Learn, D. de Vange Panteleeff, H.L. Martin Jr, J. Bwayo, J.K. Kreiss, J. Overbaugh (1998). “Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1.” J Virol 72(10): 8240-8251.

Potts K.E., M.L. Kalish, C.I. Bandea, G.M. Orloff, M. St Louis, C. Brown, N. Malanda, M. Kavuka, G. Schochetman, C.Y. Ou and Et Al. (1993). "Genetic diversity of human immunodeficiency virus type 1 strains in Kinshasa, Zaire." AIDS Res Hum Retroviruses 9(7): 613-618.

Preston B.D. and J.P. Dougherty (1996). "Mechanisms of retroviral mutation." Trends Microbiol 4(1): 16-21.

Preston B.D., B.J. Poiesz and L.A. Loeb (1988). "Fidelity of HIV-1 reverse transcriptase." Science 242(4882): 1168-1171.

Pybus O.G., A.J. Drummond, T. Nakano, B.H. Robertson and A. Rambaut (2003). "The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach." Mol Biol Evol 20(3): 381-387.

Pybus O.G. and A. Rambaut (2002). "GENIE: estimating demographic history from molecular phylogenies." Bioinformatics 18(10): 1404-1405.

Pybus O.G., M.A. Charleston, S. Gupta, A. Rambaut, E.C. Holmes and P.H. Harvey (2001). "The epidemic behavior of the hepatitis C virus." Science 292(5525): 2323-2325.

Pybus O.G., E.C. Holmes and P.H. Harvey (1999). "The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history." Mol Biol Evol 16(7): 953-959.

Quayle A.J., P. Fidel, Jr. and E.S. Rosenberg (2004). "Sex, alloimmunisation, and susceptibility to HIV infection." Lancet 363(9408): 503-504.

Quinones-Mateu M.E. and E.J. Arts (2002). "Fitness of drug resistant HIV-1: methodology and clinical implications." Drug Resist Updat 5(6): 224-233.

Rambaut A. (2000). "Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies." Bioinformatics 16(4): 395-399.

Ramos A., A. Tanuri, M. Schechter, M.A. Rayfield, D.J. Hu, M.C. Cabral, C.I. Bandea, J. Baggs and D. Pieniazek (1999). "Dual and recombinant infections: an integral part of the HIV-1 epidemic in Brazil." Emerg Infect Dis 5(1): 65-74.

Rannala B. and Z. Yang (1996). "Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference." J Mol Evol 43(3): 304-311.

Ratner L., W. Haseltine, R. Patarca, K.J. Livak, B. Starcich, S.F. Josephs, E.R. Doran, J.A. Rafalski, E.A. Whitehorn, K. Baumeister and Et Al. (1985). "Complete nucleotide sequence of the AIDS virus, HTLV-III." Nature 313(6000): 277-284.

Rausch J.W. and S.F. Le Grice (2004). "'Binding, bending and bonding': polypurine tract-primed initiation of plus-strand DNA synthesis in human immunodeficiency virus." Int J Biochem Cell Biol 36(9): 1752-1766.

Ray S.C. and T.C. Quinn (2000). "Sex and the genetic diversity of HIV-1." Nat Med 6(1): 23-25.

Ribeiro R.M. and S. Bonhoeffer (2000). "Production of resistant HIV mutants during antiretroviral therapy." Proc Natl Acad Sci U S A 97(14): 7681-7686.

Richman D.D., D. Havlir, J. Corbeil, D. Looney, C. Ignacio, S.A. Spector, J. Sullivan, S. Cheeseman, K. Barringer, D. Pauletti and Et Al. (1994). "Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy." J Virol 68(3): 1660-1666.

Robbins K.E., P. Lemey, O.G. Pybus, H.W. Jaffe, A.S. Youngpairoj, T.M. Brown, M. Salemi, A.M. Vandamme and M.L. Kalish (2003). "U.S. Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains." J Virol 77(11): 6359-6366.

Roberts J.D., K. Bebenek and T.A. Kunkel (1988). "The accuracy of reverse transcriptase from HIV-1." Science 242(4882): 1171-1173.

Robertson D.L., J.P. Anderson, J.A. Bradac, J.K. Carr, B. Foley, R.K. Funkhouser, F. Gao, B.H. Hahn, M.L. Kalish, C. Kuiken, G.H. Learn, T. Leitner, F. Mccutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P.M. Sharp,

S. Wolinsky and B. Korber (2000). "HIV-1 nomenclature proposal." Science 288(5463): 55-56.

Robertson D.L., B.H. Hahn and P.M. Sharp (1995a). "Recombination in AIDS viruses." J Mol Evol 40(3): 249-259.

Robertson D.L., P.M. Sharp, F.E. Mccutchan and B.H. Hahn (1995b). "Recombination in HIV-1." Nature 374(6518): 124-126.

Rodrigo A.G., E.G. Shpaer, E.L. Delwart, A.K. Iversen, M.V. Gallo, J. Brojatsch, M.S. Hirsch, B.D. Walker and J.I. Mullins (1999). "Coalescent estimates of HIV-1 generation time in vivo." Proc Natl Acad Sci U S A 96(5): 2187-2191.

Rogers A.S., J.W. Froggatt, 3rd, T. Townsend, T. Gordon, A.J. Brown, E.C. Holmes, L.Q. Zhang and H. Moses, 3rd (1993). "Investigation of potential HIV transmission to the patients of an HIV-infected surgeon." Jama 269(14): 1795-1801.

Rogers M.F., P.A. Thomas, E.T. Starcher, M.C. Noa, T.J. Bush and H.W. and H.W. Jaffe (1987). "Acquired immunodeficiency syndrome in children: report of the Centers for Disease Control National Surveillance, 1982 to 1985." Pediatrics 79(6): 1008-1014.

Roques P., D.L. Robertson, S. Souquiere, C. Apetrei, E. Nerrienet, F. Barre-Sinoussi, M. Muller-Trutwin and F. Simon (2004). 'Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N." AIDS 18(10): 1371-81.

Roques P., D.L. Robertson, S. Souquiere, F. Damond, A. Ayouba, I. Farfara, C. Depienne, E. Nerrienet, D. Dormont, F. Brun-Vezinet, F. Simon and P. Mauclere (2002). "Phylogenetic analysis of 49 newly derived HIV-1 group O strains: high viral diversity but no group M-like subtype structure." Virology 302(2): 259-273.

Ross H.A. and A.G. Rodrigo (2002). "Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration." J Virol 76(22): 11715-11720.

Roudinskii N.I., A.L. Sukhanova, E.V. Kazennova, J.N. Weber, V.V. Pokrovsky, V.M. Mikhailovich and A.F. Bobkov (2004). "Diversity of human immunodeficiency virus type 1 subtype A and CRF03_AB protease in Eastern Europe: selection of the V77I variant and its rapid spread in injecting drug user populations." J Virol 78(20): 11276-11287.

Rous P. and J.B. Murphy (1913). "Variation in chicken sarcoma caused by a filterable agent." J Exp Med 17: 219-231.

Rouzine I.M. and J.M. Coffin (1999). "Search for the mechanism of genetic variation in the pro gene of human immunodeficiency virus." J Virol 73(10): 8167-8178.

Roy S., U. Delling and C.H. Chen (1990). “A bulge structure in HIV-1 TAR RNA is required for Tat binding and Tat-mediated trans-activation.” Genes Dev 4: 1365 (1990)

Royce R.A., A. Seña-Soberano, W. Cates and M.S. Cohen (1997). “Sexual transmission of HIV.” N Engl J Med (336): 1072–1078.

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

Salemi M., K. Strimmer, W.W. Hall, M. Duffy, E. Delaporte, S. Mboup, M. Peeters and A.M. Vandamme (2001). "Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution." Faseb J 15(2): 276-278. Epub 2000 Dec 2008.

Salminen M.O., J.K. Carr, D.S. Burke and F.E. Mccutchan (1995). "Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning." AIDS Res Hum Retroviruses 11(11): 1423-1425.

Sanger F. (1959). “Chemistry of insulin.” Science 129(3359): 1340–1344.

Sarafianos S.G., K. Das, A.D. Clark, Jr., J. Ding, P.L. Boyer, S.H. Hughes and E. Arnold (1999). "Lamivudine (3TC) resistance in HIV-1 reverse transcriptase involves steric hindrance with beta-branched amino acids." Proc Natl Acad Sci U S A 96(18): 10027-10032.

Schatz O., J. Mous and S.F.J. LeGrice (1990). “HIV-1 RT-associated ribonuclease H displays both endonuclease and 3’-5’ exonuclease activity.” EMBO J 9(4): 1171–1176.

Schim Van Der Loeff M.F. and P. Aaby (1999). "Towards a better understanding of the epidemiology of HIV-2." Aids 13(Suppl A): S69-84.

Schreiber G.B., M.P. Busch, S.H. Kleinman and J.J. Korelitz (1996). “The risk of transfusion-transmitted viral infections.” N Engl J Med 334(26): 1685-1690.

Schwartz J.S., P.E. Dans and B.P. Kinosian (1988). "Human immunodeficiency virus test evaluation, performance, and use. Proposals to make good tests better." Jama 259(17): 2574-2579.

Seibert S.A., C.Y. Howell, M.K. Hughes and A.L. Hughes (1995). "Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1)." Mol Biol Evol 12(5): 803-813.

Selik R.M., H.W. Haverkos and J.W. Curran (1984). "Acquired immune deficiency syndrome (AIDS) trends in the United States, 1978-1982." Am J Med 76(3): 493-500.

Service S. and S.M. Blower (1995). "HIV transmission in sexual networks: an empirical analysis." Proc R Soc Lond B Biol Sci 260(1359): 237-244.

Shafer R.W., K. Dupnik, M.A. Winters and S.H. Eschleman (2000a). A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. HIV Sequence Compendium. J. Sodroski, Los Alalmos Nationnal Laboratory, Loa Alamos, NM.

Shafer R.W., D.R. Jung and B.J. Betts (2000b). "Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries." Nat Med 6(11): 1290-1292.

Shankarappa R., J.B. Margolick, S.J. Gange, A.G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C.R. Rinaldo, G.H. Learn, X. He, X.L. Huang and J.I. Mullins (1999). "Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection." J Virol 73(12): 10489-10502.

Sharp P.M., G.M. Shaw and B.H. Hahn (2005). "Simian immunodeficiency virus infection of chimpanzees." J Virol 79(7): 3891-902.

Sharp P.M., E. Bailes, R.R. Chaudhuri, C.M. Rodenburg, M.O. Santiago and B.H. Hahn (2001). "The origins of acquired immune deficiency syndrome viruses: where and when?" Philos Trans R Soc Lond B Biol Sci 356(1410): 867-876.

Sharp P.M., E. Bailes, F. Gao, B.E. Beer, V.M. Hirsch and B.H. Hahn (2000). "Origins and evolution of AIDS viruses: estimating the time-scale." Biochem Soc Trans 28(2): 275-282.

Sharp P.M., E. Bailes, D.L. Robertson, F. Gao, and B.H. Hahn (1999). "Origins and evolution of AIDS viruses." Biol Bull 196: 338–342.

Sheehy A.M., N.C. Gaddis, J.D. Choi and M.H. Malim (2002). "Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein." Nature 418(6898): 646-650. Epub 2002 Jul 2014.

Schierup M.H. and J. Hein (2000). Consequences of recombination on traditional phylogenetic analysis. Genetics 156(2): 879-891.

Shriner D., R. Shankarappa, M.A. Jensen, D.C. Nickle, J.E. Mittler, J.B. Margolick and J.I. Mullins (2004). "Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection." Genetics 166(3): 1155-1164.

Siegel S. and J. Castellan Jr (1988). Nonparametric statistics for the behavioral sciences. NY, McGraw-Hill.

Simon F., P. Mauclere, P. Roques, I. Loussert-Ajaka, M.C. Muller-Trutwin, S. Saragosti, M.C. Georges-Courbot, F. Barre-Sinoussi and F. Brun-Vezinet (1998). "Identification of a new human immunodeficiency virus type 1 distinct from group M and group O." Nat Med 4(9): 1032-1037.

Smith T.F., A. Srinivasan, G. Schochetman, M. Marcus and G. Myers (1988). "The phylogenetic history of immunodeficiency viruses." Nature 333(6173): 573-575.

Smith T.F. and M.S. Waterman (1992). "The continuing case of the Florida dentist." Science 256(5060): 1155-1156.

Sokal R.R. and C.D. Michener (1958). "A statistical method for evaluating systematic relationships." Univ. Kansas Sci. Bull. 28: 1409-1438.

Soto-Ramirez L.E., B. Renjifo, M.F. Mclane, R. Marlink, C. O'hara, R. Sutthent, C. Wasi, P. Vithayasai, V. Vithayasai, C. Apichartpiyakul, P. Auewarakul, V. Pena Cruz, D.S. Chui, R. Osathanondh, K. Mayer, T.H. Lee and M. Essex

(1996). "HIV-1 Langerhans' cell tropism associated with heterosexual transmission of HIV." Science 271(5253): 1291-1293.

Starcich, B.R., Hahn, B.H., Shaw, G.M., McNeely, P.D., Modrow, S., Wolf, H., et al., (1986). “Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS.” Cell 45 (5), 637–648.

Studier J.A. and K.J. Keppler (1988). "A note on the neighbor-joining algorithm of Saitou and Nei." Mol Biol Evol 5(6): 729-731.

Stumpf M.P.H. and G.A.T. McVean (2003). “Estimating recombination rates from population-genetic data.” Nat Rev Genet 4(12): 959-68.

Swofford D.L., G.J. Olson, P.J. Waddell and D.M.C. Hillis (1996). Phylogenetic inference. Molecular systematic. B.K. Mable, Sinauer Associates, Sunderland, MA.

Tajima F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics 123(3): 585-595.

Tang J., C.M. Wilson, S. Meleth, A. Myracle, E. Lobashevsky, M.J. Mulligan, S.D. Douglas, B. Korber, S.H. Vermund and R.A. Kaslow (2002). "Host genetic profiles predict virological and immunological control of HIV-1 infection in adolescents." Aids 16(17): 2275-2284.

Tatt I.D., K.L. Barlow, J.P. Clewley, O.N. Gill and J.V. Parry (2004). "Surveillance of HIV-1 Subtypes Among Heterosexuals in England and Wales, 1997-2000." J Acquir Immune Defic Syndr 36(5): 1092-1099.

Tatt I.D., K.L. Barlow, A. Nicoll and J.P. Clewley (2001). "The public health significance of HIV-1 subtypes." Aids 15(Suppl 5): S59-71.

Tatt I.D., K.L. Barlow and J.P. Clewley (2000). "A gag gene heteroduplex mobility assay for subtyping HIV-1." J Virol Methods 87(1-2): 41-51.

Taylor S., P. Cane, S. Hue, L. Xu, T. Wrin, Y. Lie, N. Hellmann, C. Petropoulos, J. Workman, D. Ratcliffe, B. Choudhury and D. Pillay (2003). "Identification of a Transmission Chain of HIV Type 1 Containing Drug Resistance-Associated Mutations." AIDS Res Hum Retroviruses 19(5): 353-361.

Taylor J.E. and B.E. Korber (2005). “HIV-1 intra-subtype superinfection rates: estimates using a structured coalescent with recombination.” Infect Genet Evol 5(1):85-95.

Thompson J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G. Higgins (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res 25(24): 4876-4882.

Thomson M.M., E. Delgado, I. Herrero, M.L. Villahermosa, E. Vazquez-De Parga, M.T. Cuevas, R. Carmona, L. Medrano, L. Perez-Alvarez, L. Cuevas and R. Najera (2002). "Diversity of mosaic structures and common ancestry of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Argentina revealed by analysis of near full-length genome sequences." J Gen Virol 83(Pt 1): 107-119.

Thomson M.M., E. Delgado, N. Manjon, A. Ocampo, M.L. Villahermosa, A. Marino, I. Herrero, M.T. Cuevas, E. Vazquez-De Parga, L. Perez-Alvarez, L. Medrano, J.A. Taboada and R. Najera (2001). "HIV-1 genetic diversity in Galicia Spain: BG intersubtype recombinant viruses circulating among injecting drug users." Aids 15(4): 509-516.

Thorne J.L., H. Kishino and I.S. Painter (1998). "Estimating the rate of evolution of the rate of molecular evolution." Mol Biol Evol 15(12): 1647-1657.

Tovanabutra S., V. Watanaveeradej, K. Viputtikul, M. De Souza, M.H. Razak, V. Suriyanon, J. Jittiwutikarn, S. Sriplienchan, S. Nitayaphan, M.W. Benenson, N. Sirisopana, P.O. Renzullo, A.E. Brown, M.L. Robb, C. Beyrer, D.D. Celentano, J.G. Mcneil, D.L. Birx, J.K. Carr and F.E. Mccutchan (2003). "A new circulating recombinant form, CRF15_01B, reinforces the linkage between IDU and heterosexual epidemics in Thailand." AIDS Res Hum Retroviruses 19(7): 561-567.

Trachtenberg E., B. Korber, C. Sollars, T.B. Kepler, P.T. Hraber, E. Hayes, R. Funkhouser, M. Fugate, J. Theiler, Y.S. Hsu, K. Kunstman, S. Wu, J. Phair, H. Erlich and S. Wolinsky (2003). "Advantage of rare HLA supertype in HIV disease progression." Nat Med 9(7): 928-935.

Triques K., A. Bourgeois, N. Vidal, E. Mpoudi-Ngole, C. Mulanga-Kabeya, N. Nzilambi, N. Torimiro, E. Saman, E. Delaporte and M. Peeters (2000). "Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K." AIDS Res Hum Retroviruses 16(2): 139-151.

Triques K., A. Bourgeois, S. Saragosti, N. Vidal, E. Mpoudi-Ngole, N. Nzilambi, C. Apetrei, M. Ekwalanga, E. Delaporte and M. Peeters (1999). "High diversity of HIV-1 subtype F strains in Central Africa." Virology 259(1): 99-109.

Tscherning C., A. Alaeus, R. Fredriksson, A. Bjorndal, H. Deng, D.R. Littman, E.M. Fenyo and J. Albert (1998). "Differences in chemokine coreceptor usage between genetic subtypes of HIV-1." Virology 241(2): 181-188.

Turner B. and F. Summers (1999). "Structural biology of HIV-1." J Mol Biol 285: 1-32

Twiddy S.S., O.G. Pybus and E.C. Holmes (2003). "Comparative population dynamics of mosquito-borne flaviviruses." Infect Genet Evol 3(2): 87-95.

Van Haastrecht H.J., J.A. Van Den Hoek, G.H. Mientjes and R.A. Coutinho (1992). "Did the introduction of HIV among homosexual men precede the introduction of HIV among injecting drug users in The Netherlands?" Aids 6(1): 131-132.

Van Harmelen J.H., E. Van Der Ryst, A.S. Loubser, D. York, S. Madurai, S. Lyons, R. Wood and C. Williamson (1999). "A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations." AIDS Res Hum Retroviruses 15(4): 395-398.

Van Harmelen J., R. Wood, M. Lambrick, E.P. Rybicki, A.L. Williamson and C. Williamson (1997). "An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa." Aids 11(1): 81-87.

Vartanian J.P., P. Sommer and S. Wain-Hobson (2003). "Death and the retrovirus." Trends Mol Med 9(10): 409-413.

Vartanian J.P., M. Henry and S. Wain-Hobson (2002). "Sustained G-->A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome." J Gen Virol 83(Pt 4): 801-805.

Vartanian J.P., A. Meyerhans, B. Asjo and S. Wain-Hobson (1991). "Selection, recombination, and G----A hypermutation of human immunodeficiency virus type 1 genomes." J Virol 65(4): 1779-1788.

Vento S., G. Di Perri, T. Garofano, E. Concia and D. Bassetti (1993). "Pneumocystis carinii pneumonia during primary HIV-1 infection." Lancet 342(8862): 24-25.

Vernazza P.L., J.J. Eron, S.A. Fiscus and M.S. Cohen (1999). "Sexual transmission of HIV: infectiousness and prevention." Aids 13(2): 155-166.

Vidal N., D. Koyalta, V. Richard, C. Lechiche, T. Ndinaromtan, A. Djimasngar, E. Delaporte and M. Peeters (2003). "High genetic diversity of HIV-1 strains in Chad, West Central Africa." J Acquir Immune Defic Syndr 33(2): 239-246.

Vidal N., M. Peeters, C. Mulanga-Kabeya, N. Nzilambi, D. Robertson, W. Ilunga, H. Sema, K. Tshimanga, B. Bongo and E. Delaporte (2000). "Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa." J Virol 74(22): 10498-10507.

Vogt P.K. (1971). "Genetically stable reassortment of markers during mixed infection with avian tumor viruses." Virology 46(3): 947-952.

Wahl S.M., T. Greenwell-Wild, G. Peng, H. Hale-Donze and J.M. Orenstein (1999). "Co-infection with opportunistic pathogens promotes human immunodeficiency virus type 1 infection in macrophages." J Infect Dis 179 Suppl 3: S457-460.

Wainberg M.A. (2004). "HIV-1 subtype distribution and the problem of drug resistance." Aids 18(Suppl 3): S63-68.

Watson J.D. and F.H. Crick (1953). “Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.” Nature 171(4356):737-738.

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 and Et Al. (1995). "Viral dynamics in human immunodeficiency virus type 1 infection." Nature 373(6510): 117-122.

Weiss P.J., S.K. Brodine, R.R. Goforth, C.A. Kennedy, M.R. Wallace, P.E. Olson, F.C. Garland, F.W. Hall, S.I. Ito, E.C. Oldfield 3rd (1992). “Initial low CD4 lymphocyte counts in recent human immunodeficiency virus infection and lack of association with identified coinfections.” J Infect Dis 166(5): 1149-1153.

Whelan S., P. Lio and N. Goldman (2001). "Molecular phylogenetics: state-of-the-art methods for looking into the past." Trends Genet 17(5): 262-272.

Wilbe K., C. Casper, J. Albert and T. Leitner (2002). "Identification of two CRF11-cpx genomes and two preliminary representatives of a new circulating recombinant form (CRF13-cpx) of HIV type 1 in Cameroon." AIDS Res Hum Retroviruses 18(12): 849-856.

Wilkins A., D. Ricard, J. Todd, H. Whittle, F. Dias and A. Paulo Da Silva (1993). "The epidemiology of HIV infection in a rural area of Guinea-Bissau." Aids 7(8): 1119-1122.

Williamson S. (2003). "Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression." Mol Biol Evol 20(8): 1318-1325. Epub 2003 May 1330.

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

Wolinsky S.M., B.T. Korber, A.U. Neumann, M. Daniels, K.J. Kunstman, A.J. Whetsell, M.R. Furtado, Y. Cao, D.D. Ho and J.T. Safrit (1996). "Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection." Science 272(5261): 537-542.

Wolinsky S.M., C.M. Wike, B.T. 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(5048): 1134-1137.

Wong-Staal F., G.M. Shaw, B.H. Hahn, S.Z. Salahuddin, M. Popovic, P. Markham, R. Redfield, R.C. Gallo (1985). “Genomic diversity of human T-lymphotropic virus type III (HTLV-III).” Science 229 (4715), 759–762.

Worobey M. (2001). "A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria." Mol Biol Evol 18(8): 1425-1434.

Worobey M., M.L. Santiago, B.F. Keele, J.B. Ndjango, J.B. Joy, B.L. Labama, A.B. Dhed, A. Rambaut, P.M. Sharp, G.M. Shaw and B.H. Hahn (2004). "Origin of AIDS: contaminated polio vaccine theory refuted." Nature 428(6985): 820.

Wu W., B.M. Blumberg, P.J. Fay and R.A. Bambara (1995). "Strand transfer mediated by human immunodeficiency virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation." J Biol Chem 270(1): 325-332.

Xiridou M., R. Geskus, J. De Wit, R. Coutinho and M. Kretzschmar (2004). "Primary HIV infection as source of HIV transmission within steady and casual partnerships among homosexual men." Aids 18(9): 1311-1320.

Yahi N., J. Fantini, C. Tourres, N. Tivoli, N. Koch and C. Tamalet (2001). "Use of drug resistance sequence data for the systematic detection of non-B human immunodeficiency virus type 1 (HIV-1) subtypes: how to create a sentinel site for monitoring the genetic diversity of HIV-1 at a country scale." J Infect Dis 183(9): 1311-1317.

Yamaguchi J., A.S. Vallari, P. Swanson, P. Bodelle, L. Kaptue, C. Ngansop, L. Zekeng, L.G. Gurtler, S.G. Devare and C.A. Brennan (2002). "Evaluation of HIV type 1 group O isolates: identification of five phylogenetic clusters." AIDS Res Hum Retroviruses 18(4): 269-282.

Yamaguchi-Kabata Y. and T. Gojobori (2000). "Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes." J Virol 74(9): 4335-4350.

Yang R., X. Xia, S. Kusagawa, C. Zhang, K. Ben and Y. Takebe (2002). "On-going generation of multiple forms of HIV-1 intersubtype recombinants in the Yunnan Province of China." Aids 16(10): 1401-1407.

Yang Z. (1996). "Among-site variation and its impact on phylogenetic analyses." Trends in Ecology and evolution 11: 367-371.

Yang Z. (1994a). "Estimating the pattern of nucleotide substitution." J Mol Evol 39(1): 105-111.

Yang Z. (1994b). "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods." J Mol Evol 39(3): 306-314.

Yang Z., R. Nielsen, N. Goldman and A.M. Pedersen (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics 155(1): 431-449.

Yang Z., N. Goldman and A. Friday (1994). "Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation." Mol Biol Evol 11(2): 316-324.

Yarchoan R., R.W. Klecker, K.J. Weinhold, P.D. Markham, H.K. Lyerly, D.T. Durack, E. Gelmann, S.N. Lehrman, R.M. Blum, D.W. Barry and Et Al. (1986). "Administration of 3'-azido-3'-deoxythymidine, an inhibitor of HTLV-III/LAV replication, to patients with AIDS or AIDS-related complex." Lancet 1(8481): 575-580.

Yerly S., R. Quadri, F. Negro, K. Posfay Barbe, J-J. Cheseaux, P. Burgisser, C-A. Siegrist and L. Perrin (2001a). "Nosocomial Outbreak of Multiple Bloodborne Viral Infections." J Infect Dis 184(3): 369-72.

Yerly S., S. Vora, P. Rizzardi, J.P. Chave, P.L. Vernazza, M. Flepp, A. Telenti, M. Battegay, A.L. Veuthey, J.P. Bru, M. Rickenbach, B. Hirschel and L. Perrin (2001b). "Acute HIV infection: impact on the spread of HIV and transmission of drug resistance." Aids 15(17): 2287-2292.

Yerly S., A. Rakik, S.K. De Loes, B. Hirschel, D. Descamps, F. Brun-Vezinet and L. Perrin (1998). "Switch to unusual amino acids at codon 215 of the human immunodeficiency virus type 1 reverse transcriptase gene in seroconvertors infected with zidovudine-resistant variants." J Virol 72(5): 3520-3523.

Yirell D.L., S.J. Hutchinson, M. Griffin, S.M. Gore, A.J. Leigh-Brown and D.J. Goldberg (1999). "Completing the molecular investigation into the HIV outbreak at Glenochil prison." Epidemiol Infect 123(2): 277-2782.

Yirrell D.L., P. Robertson, D. J. Goldberg, J. McMenamin, S. Cameron and A J Leigh Brown (1997). "Molecular investigation into outbreak of HIV in a Scottish prison." BMJ 314(7092): 1446-1450.

Yu X., Y. Yu, B. Liu, K. Luo, W. Kong, P. Mao and X.F. Yu (2003). "Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex." Science 302(5647): 1056-1060. Epub 2003 Oct 1016.

Yusim K., M. Peeters, O.G. Pybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler and B. Korber (2001). "Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution." Philos Trans R Soc Lond B Biol Sci 356(1410): 855-866.

Zanotto P.M., E.G. Kallas, R.F. De Souza and E.C. Holmes (1999). "Genealogical evidence for positive selection in the nef gene of HIV-1." Genetics 153(3): 1077-1089.

Zekeng L., L. Gurtler, E. Afane Ze, A. Sam-Abbenyi, G. Mbouni-Essomba, E. Mpoudi-Ngolle, M. Monny-Lobe, J.B. Tapka and L. Kaptue (1994). "Prevalence of HIV-1 subtype O infection in Cameroon: preliminary results." Aids 8(11): 1626-1628.

Zennou V., F. Mammano, S. Paulous, D. Mathez and F. Clavel (1998). "Loss of viral fitness associated with multiple Gag and Gag-Pol processing defects in human immunodeficiency virus type 1 variants selected for resistance to protease inhibitors in vivo." J. Virol. 72(4): 3300-3306.

Zhang L., Y. Huang, T. He, Y. Cao and D.D. Ho (1996). "HIV-1 subtype and second-receptor use." Nature 383(6603): 768.

Zhang L.Q., P. Mackenzie, A. Cleland, E.C. Holmes, A.J. 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(6): 3345-3356.

Zhu T., B.T. Korber, A.J. Nahmias, E. Hooper, P.M. Sharp and D.D. Ho (1998). "An African HIV-1 sequence from 1959 and implications for the origin of the epidemic." Nature 391(6667): 594-597.

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 patients with primary infection." Science 261(5125): 1179-1181.

Zhuang J., A.E. Jetzt, G. Sun, H. Yu, G. Klarmann, Y. Ron, B.D. Preston and J.P. Dougherty (2002). "Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots." J Virol 76(22): 11273-11282.

Appendix I

Publications during the course of PhD study related to work undertaken within this thesis

Hué S., D. Pillay, J.P. Clewley and O.G. Pybus (2005). "Genetic analysis reveals the complex structure of HIV-1 transmission within localised risk groups." Proc Natl Acad Sci USA 102(12): 4425-4429.

Hué S., J.P. Clewley, P.A. Cane and D. Pillay (2005). "Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour." AIDS 19(40): 6-7.

Pao D., M. Fisher, S. **Hué**, G. Dean, G. Murphy, P.A. Cane, C.A. Sabin and D. Pillay (2005). "Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections." AIDS 19(1): 85-90.

Hué S., Clewley J.P., Cane P.A., and Pillay D. 'HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy." AIDS 18(5): 719-28.

Appendix II

Work Not Included in this Thesis

Dumans A.T., M.A. Soares, E.S. Machado, S. **Hué**, R.M. Brindeiro, D. Pillay, and A. Tanuri (2004). "Synonymous Genetic Polymorphisms within Brazilian Human Immunodeficiency Virus Type 1 Subtypes May Influence Mutational Routes to Drug Resistance." J Infect Dis 189(7): 1232-1238.

Gubser C., S. **Hué**, P. Kellam, and G.L. Smith (2004). "Poxvirus genomes: a phylogenetic analysis." J Gen Virol 85: 105-117.

Taylor S., P. Cane, S. **Hué**, L. Xu, T. Wrin, Y. Lie, N. Hellmann, C. Petropoulos, J. Workman, D. Ratcliffe, B. Choudhury and D. Pillay (2003). "Identification of a transmission chain of HIV type 1 containing drug resistance-associated mutations." AIDS Res Hum Retroviruses 19(5): 353-361.

Xu L., S. **Hué**, S. Taylor, D. Ratcliffe, J.A. Workman, S. Jackson, P.A. Cane and D. Pillay (2002). "Minimal variation in T-20 binding domain of different HIV-1 subtypes from antiretroviral-naive and -experienced patients." AIDS 16(12): 1684-1686.

5. Joly V, Descamps D, Peytavin G, Touati F, Mentre F, Duval X, *et al.* **Evolution of human immunodeficiency virus type 1 (HIV-1) resistance mutations in nonnucleoside reverse transcriptase inhibitors (NNRTIs) in HIV-1-infected patients switched to antiretroviral therapy without NNRTIs.** *Antimicrob Agents Chemother* 2004; **48**:172–175.

6. Eshelman SH, Guay LA, Mwatha A, Brown ER, Cunningham SP, Musoke P, *et al.* **Characterization of nevirapine resistance mutations in women with subtype A vs. D HIV-1 6–8 weeks after single-dose nevirapine (HIVNET 012).** *J Acquir Immune Defic Syndr* 2004; **35**:126–130.

Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour

Since the first use of molecular methods to ascertain HIV-1 transmission was published in 1992 [1], controversy in this area has been vivid [2–4]. When approaching the use of HIV-1 sequences to characterize linkage, a key determinant is the choice for the most informative genetic region; in this respect, the use of the *pol* gene has hitherto been unpopular. We nonetheless demonstrated that sequences of this genetic region, widely available since the onset of drug resistance testing, holds sufficient genetic variation to allow phylogenetic analyses [5]. In an opinion paper published in response to our article, Stürmer and colleagues [6] not only challenged the reliability of phylogenies constructed on the basis of the protease and reverse transcriptase genes, but also recommended the blind use of the V3 region of the envelope (*env*) gene. In view of the importance of this issue for epidemiological studies, as well as in the clinical and legal arena, we would like to address their line of argument, in order to widen the debate. In essence, we maintain that the key issue is for a dataset to contain sufficient variability as defined by phylogenetic criteria, regardless of whether the sequence is *gag, pol* or *env.*

There is no such thing as an ultimate gene for evolutionary analyses of HIV-1. Ideally, full-length sequences should be used for the investigation of potential linkages by phylogenetic means; however, practicalities preclude such an approach. Echoing previous and well-established opinions, see for example Leitner *et al.* [7], Stürmer *et al.* [6] recommended the use of *env* gene sequences, the extensive variation of which has made it attractive for such analyses. However, the exploitation of *env* is far from unproblematical. First, convergent evolution (i.e. identical mutational patterns in unlinked sequences) has repeatedly been observed in the V3 loop of the *env* gene [8,9]. More importantly, the rapid genetic diversification of this region is likely to compromise the identification of linked sequences in distantly sampled individuals. Both divergence and diversity of the HIV-1 *env* gene have been shown to increase linearly in the early stages of infection [10]. The latter observations may explain why one of the possible transmission pairs identified in the *pol* tree published by Stürmer *et al.* [6] failed to be supported by their *env* tree. The choice of an appropriate genetic target for such studies must not only be considered in light of the intrinsic variability of the dataset itself, but also of the possible time span separating the samples under comparison. Our own data suggest that the relative stability of the polymerase gene may confer some benefit in this respect. We identified the same clustering patterns in phylogenies independently constructed on the basis of the *pol, gag* and *env* genes of the same HIV-1 samples, supporting the idea of sufficient intrinsic sequence variation in the *pol* region.

Stürmer *et al.* [6] supported their view by identifying sequences that appear linked within a phylogeny based on the *pol* region, but which fail to fulfil the arbitrary criteria of a bootstrap support greater than 70% within a tree based on the partial *env* gene. However, the methodological underpinning of this approach is flawed. First, the authors produced a neighbour-joining tree [11], constructed under the 'Kimura two-parameters' (K2P) model of nucleotide substitution [12], which is, in our view, unsatisfactory. If the choice of a phylogenetic method for reconstructing transmissions is less sensitive than the choice of a genetic region [7], the misuse of a model of evolution can have severe consequences on the accuracy of the reconstruction. As rates of evolution differ across HIV-1 lineages, populations, or genetic regions, the selection of an optimal model must be a prerequisite when estimating HIV-1 phylogenies. The systematic (and often unjustified) use of the K2P model of evolution is unfortunately frequent in HIV molecular analyses. When using an over-simplistic model, features of importance in the context of HIV-1 transmission, such as branch length, may be underestimated [13]. Moreover, when statistically testing different evolutionary models for various HIV-1 genetic regions, Posada and Crandall [14] demonstrated that the K2P model was suboptimal, whichever gene was under study. Therefore, the criteria used to discriminate between potential linked and unlinked sequences must be adapted to the specificities of one's dataset. These criteria should be empirically determined from dataset to dataset, and not dogmatically implemented. The unique criterion used by Stürmer *et al.* [6] (i.e. a bootstrap support above 70%, regardless of the tree topology) is clearly insufficient to draw reasonable inferences regarding the true or false linkage of infections. Such a cut-off value must first consider genes evolving at a different pace, because bootstrap evaluation is sensitive to the degree of polymorphism exhibited by the sequences. Second, bootstrap evaluation is not on its own sufficiently discriminatory, and the branch length supporting suspected transmission pairs is another obvious pattern to take into account. A sensible way to determine such criteria would be to incorporate positive controls within the sequence alignment, such as sequences from known transmission pairs or intrapatient follow-up samples.

Stürmer *et al.* [6] failed to present such controls, compromising the pertinence of their findings. As accession numbers for several of the sequences used by the group have not been provided, we have been unable to re-analyse their data according to the above criteria.

In conclusion, we agree that close attention is required in dealing with HIV-1 sequences for epidemiological, clinical or forensic purposes. Using a rigorous and reproducible methodology, we recently showed that our dataset of *pol* sequences hold enough sequence variation to allow phylogenetic reconstruction, despite the conservation of the gene. Our conclusion was based on a stringent comparative analysis of the respective phylogenetic signals held by the three main genes of HIV-1, namely the *pol*, *gag* and *env* genes.

***Stéphane Hue[a,b], Jonathan P. Clewley[b], Patricia A. Cane[c] and Deenan Pillay[a,b]**, [a]Centre of Virology, Department of Infection, Royal Free and University College Medical School, UCL, London, UK; [b]Health Protection Agency, Colindale, London, UK; and [c]HPA Porton Down, Salisbury, UK.*

Received: 17 November 2004; accepted: 22 December 2004.

References

1. Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT, *et al.* **Molecular epidemiology of HIV transmission in a dental practice.** *Science* 1992; **256**:1165–1171.
2. Smith TF, Waterman MS. **The continuing case of the Florida dentist.** *Science* 1992; **256**:1155–1156.
3. DeBry RW, Abele LG, Weiss SH, Hill MD, Bouzas M, Lorenzo E, *et al.* **Dental HIV transmission?** *Nature* **1993; 361:691.**
4. Crandall KA. **Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus.** *J Virol* 1995; **69**:2351–2356.
5. Hue S, Clewley JP, Cane PA, Pillay D. **HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy.** *AIDS* 2004; **18**:719–728.
6. Stürmer M, Preiser W, Gute P, Nisius G, Doerr HW. **Phylogenetic analysis of HIV-1 transmission: pol gene sequences are unsufficient to clarify true relationships between patient isolates.** *AIDS* 2004; **18**:2109–2113.
7. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. **Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis.** *Proc Natl Acad Sci U S A* 1996; **93**:10864–10869.
8. Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Brown AJ. **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 U S A* 1992; **89**:4835–4839.
9. Zhang LQ, MacKenzie P, Cleland A, Holmes EC, Brown AJ, Simmonds P. **Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection.** *J Virol* 1993; **67**:3345–3356.
10. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, *et al.* **Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection.** *J Virol* 1999; **73**:10489–10502.
11. Saitou N, Nei M. **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987; **4**:406–425.
12. Kimura M. **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980; **16**:111–120.
13. Yang Z, Goldman N, Friday A. **Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation.** *Mol Biol Evol* 1994; **11**:316–324.
14. Posada D, Crandall KA. **Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1).** *Mol Biol Evol* 2001; **18**:897–906.

Successful desensitization of enfuvirtide-induced skin hypersensitivity reaction

A 38-year-old Caucasian woman was diagnosed with HIV in 1996 and was treated with various antiretroviral drug regimens, including different combinations of nucleoside reverse transcriptase inhibitors (NRTI) and protease inhibitors (PI), but not non-nucleoside reverse transcriptase inhibitors (NNRTI). She maintained her CD4 cell count between 500 and 600 cells/μl and had a consistently detectable HIV-RNA viral load of 10 000–30 000 copies/ml plasma (HIV-Monitor; Roche, Mannheim, Germany). The patient developed multiple NRTI mutations (M41L, D67N, K70R, M184V, T215Y, K219E) and PI mutations (L10I, K20R, L24I, L33F, M36I, F53L, I54V, L63P, V82A), but no NNRTI mutations.

During the past year, her CD4 cell count began to decrease down to 300 cells/μl and her HIV-RNA viral load increased gradually to 300 000 copies/ml plasma. We decided to initiate a new regimen consisting of tenofovir (300 mg once a day), epivir (150 mg twice a day), efavirenz (600 mg once a day) and enfuvirtide (90 mg twice a day). After 10 days of treatment, the patient developed a confluent erythematous maculopapular rash starting on the trunk and spreading to her neck, face, upper and lower extremities. At this point, we considered efavirenz to be the cause of the rash, and decided to continue treatment considering that there were no systemic symptoms and that a rash caused by efavirenz usually disappears with continuous therapy [1]. However, the patient's rash worsened and her skin became red and itchy. After six more days, we decided to discontinue all antiretroviral drugs.

Two weeks later, after complete remission of the skin symptoms and considering the limited therapeutic options for this patient, we resumed the former treatment regimen and initiated a desensitization protocol to efavirenz as has previously been described [1]. Our Institution's Ethics Committee approval and the patient's informed consent were obtained. Two hours post-initiation of desensitization with efavirenz, the patient injected a subcutaneous dose of enfuvirtide. Half an hour post-injection, the patient developed a local and generalized maculopapular rash that lasted for the next 10 h. Twelve hours later, she injected the next

CONCISE COMMUNICATION

Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections

David Pao[a], Martin Fisher[a], Stephane Hué[b,e], Gillian Dean[a], Gary Murphy[e], Patricia A. Cane[d,e], Caroline A. Sabin[c] and Deenan Pillay[b,e]

Objective: To study primary HIV-1 infections (PHI) using molecular and epidemiological approaches in order to assess correlates of transmission in this population.

Methods: Individuals with PHI were recruited prospectively from a discrete cohort of 1235 individuals under follow-up in a well-defined geographical area between 1999 and 2003. PHI was diagnosed by one of the following: negative HIV antibody test within 18 months, evolving antibody response, or application of the serological testing algorithm for recent HIV seroconversion. The *pol* gene was sequenced to identify genotypic resistance and facilitate molecular epidemiological analysis. Clinical data were collected and linked in an irretrievable fashion when informed consent was obtained.

Results: A total of 103 individuals with PHI diagnosed between 1999 and 2003 were included in the study; 99 (96%) were male and 90 (91%) were men who have sex with men. Viruses from 35 out of 103 (34%) appeared within 15 phylogenetically related clusters. Significant associations with clustering were: young age, high CD4 cell count, number of sexual contacts, and unprotected anal intercourse (UAI) in the 3 months before diagnosis ($P < 0.05$ for all). High rates of acute sexually transmitted infections (STI) were observed in both groups with a trend towards higher rates in those individuals with viruses within a cluster (42.9 versus 27.9%; $P = 0.13$).

Conclusion: High rates of partner change, UAI and STI are factors that facilitate onward transmission during PHI. More active identification of individuals during PHI, the management of STI and highly active antiretroviral therapy may all be useful methods to break transmission networks.

AIDS 2005, **19**:85–90

Keywords: Acute HIV infection, epidemiology, phylogenetic tree, sexually transmitted diseases

From the [a]Department of GU Medicine, Brighton and Sussex University Hospitals, Brighton, UK; [b]Centre of Virology, Division of Infection and Immunity, and [c]Department of Primary Care and Population Sciences, Royal Free and University College Medical School, University College London, London, UK; [d]HPA Antiviral Susceptibility Reference Unit, University of Birmingham Medical School, and Birmingham HPA, Heartlands Hospital, Birmingham, UK; and [e]Centre for Infections, Health Protection Agency, Colindale, UK.

Correspondence to Deenan Pillay, Centre of Virology, Division of Infection and Immunity, Royal Free and University College Medical School, Windeyer Institute, 46 Cleveland Street, London W1P 6DF, UK.

E-mail: d.pillay@ucl.ac.uk

Received: 22 June 2004; revised: 20 September 2004; accepted: 7 October 2004.

ISSN 0269-9370 © 2005 Lippincott Williams & Wilkins

Introduction

Worldwide, 4.2 million adults were estimated to have new HIV-1 infection in 2003 [1], although it is unclear whether these represent new diagnoses of chronic infection or recently acquired infections; nevertheless it is clear that strategies to interrupt the sexual transmission of HIV-1 are key to reducing the worldwide burden of HIV disease. Within the UK, most new diagnoses now represent imported infections [2]; however, continual incident infections among men who have sex with men (MSM) are evident [3]. Taken as a single disease stage, the overall efficiency of sexual transmission of HIV is low, but numerous biological and mathematical modelling studies predict much higher infectiousness during primary HIV infection (PHI) compared with chronic HIV infection.

Biologically, the high plasma viral load seen during PHI [4–6], which probably parallels semen viral load [7–10], is strongly correlated with the risk of sexual transmission [11] and therefore epidemic growth. Other factors that may increase transmission include sexually transmitted infections (STI) and host susceptibility [12,13]. The recent finding of higher concentrations of HIV-1 RNA in rectal mucosa than in blood or semen is also pertinent [14].

Mathematical models estimate the average probability of male–female transmission of HIV-1 per unprotected coital act to be between 0.0005 and 0.003% during chronic HIV infection [15], which in itself would not sustain an epidemic. By contrast, when the high viral load of PHI is taken into account, men with average semen viral load, without concurrent STI, would be expected to infect 7–24% of susceptible female partners during the first 2 months of infection (an eight to 10-fold increase from chronic HIV infection) [9]. According to male–male models, between 25 and 47% of new HIV infections may be transmitted during this period of initial HIV infection [16,17], possibly within steady as opposed to casual relationships [18]. In addition, these individuals are infectious before symptoms of PHI [19], may not even show symptoms of disease [20] (and therefore be unaware of the risk they pose to partners), and often engage in high-risk sexual practices [21,22] with a higher number of sexual contacts [23].

There is also increasing evidence that any decrease in the per-contact risk as a result of the increased availability of antiretroviral therapy appears to have been counterbalanced or overwhelmed by increases in risky sexual behaviour [24,25]. This is reflected in the transmission of primary resistant HIV strains, the prevalence of which approaches 20% in the UK and elsewhere [26–29].

In order to understand further the role played by PHI in sexual transmission we carried out phylogenetic characterization of PHI and collected relevant epidemiological data regarding sexual behaviour, clinical features and STI.

Methods

Study recruitment

Individuals were recruited from a cohort of 1235 HIV-positive patients attending a single genitourinary medicine unit for follow-up from 1999 to 2003. This prospective cohort included over 2100 patients with HIV infection, with 1235 being seen during the study period. Of these, 86% were caucasian, 89% were men, and the predominant route of transmission was sex between men (79%). The department is the sole local provider of HIV and STI care, and national surveillance data confirm that over 90% of individuals with HIV infection resident in the area attend this institution.

Individuals with PHI were identified by one or more of the following: previous negative HIV antibody test within 18 months, evolving Western blot or HIV antibody response, or application of the serological testing algorithm for recent HIV seroconversion (STARHS) assay. STARHS is a dual testing strategy in which specimens that are confirmed anti-HIV positive after detection by a sensitive screening assay are tested on an assay that has been altered to make it less sensitive. Specimens that are unreactive on this less sensitive assay are deemed to be recent infections, whereas specimens that are reactive in both assays are deemed to come from infections that are long standing [30]. At the time of HIV diagnosis the majority of individuals underwent a full STI screen.

Clinical data collection

In those from whom written informed consent was obtained, information regarding clinical status was collected from clinic case notes: the date of diagnosis, CD4 cell count, CD4 cell percentage, HIV viral load, the presence and nature of STI in the 3 months before the diagnosis of PHI (gonorrhoea, chlamydia, non-specific urethritis, early syphilis, herpes simplex), and the absence or presence of PHI symptoms. Information relating to the individual's HIV acquisition risk group, sexual behaviour (including estimated number and nature of sexual contacts in the 3 months before diagnosis of PHI) was also recorded. These data are routinely collected for all new HIV-1 diagnoses within this clinic.

Serological testing algorithm for recent HIV seroconversion and analysis

STARHS testing was performed using the bioMérieux Vironostika HIV-1 assay (bioMérieux UK Ltd., Basingstoke, UK) as previously described [31]. A standardized optical density for each specimen was determined. For this study a standardized optical density of less than

1.0 was used to identify recent infections, and this cut-off equates to an estimated seroconversion within the previous 4–6 months.

Phylogenetic analysis

The HIV *pol* gene was sequenced from plasma obtained at the time of HIV diagnosis. These sequences were used for phylogenetic analysis, a method previously shown by this group to have utility in reconstructing transmission events [32]. Full-length sequences from the protease gene (295 nt) and the first 230 codons of reverse transcriptase were aligned using the program Clustal X (available from http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html) and then adjusted manually with the software BioEdit (available from http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Sequences that could not be unambiguously aligned or were of insufficient length were excluded from the study. Phylogenetic relationships between the *pol* sequences were reconstructed using the neighbour-joining followed by maximum likelihood methods. An initial neighbour-joining tree was built under the Hasegawa–Kishino–Yang (HKY85) model of evolution with a ratio of transversion to transitions of 2:1 using the tree-building software Paup* (available from http://paup.csit.fsu.edu/about.html).

The best fitting model of nucleotide substitution was estimated on the basis, of the neighbour-joining tree topology using a maximum likelihood ratio test with Modeltest version 3.0 (available from http://bioag.-byu.edu/zoology/crandall_lab/modeltest.htm). The derived parameters of the selected model were then used to perform a heuristic search for a maximum likelihood tree with Paup*. The construction of the tree was done according to the general time reversible (GTR) model of evolution, with a proportion of invariable sites and gamma distribution. An HIV-1 subtype K sequence (Genbank accession number AJ249239) retrieved from the Los Alamos HIV database (http://hiv-web.lanl.gov/) was used as an outgroup and six pairs of follow-up sequences from the same individuals were used as controls. The robustness of the neighbour-joining trees was evaluated by bootstrap analysis, with 1000 rounds of replication.

Statistical analysis

Statistical comparisons of those in a cluster with those not in a cluster were performed using Chi-squared tests, Fisher's exact tests or Mann–Whitney U tests, as appropriate. Multivariable logistic regression was used to identify factors independently associated with belonging to a cluster. All statistical analyses were performed using SAS version 8 (available from http://v8doc.sas.-com/sashtml/). The study was approved by the Brighton and Hove Local Research Ethics Committee and the Health Protection Agency Ethics Committee. Confidentiality and anonymity were protected by irreversibly unlinking clinic and laboratory from the study ID number using a firewall system managed by the local public health laboratory. Written, informed consent was obtained from all participants.

Results

Study population description

A total of 103 individuals with PHI diagnosed between 1999 and 2003 were included in the epidemiological and phylogenetic analysis. Of these, 73 (71%) had a STARHS antibody test suggestive of infection within the previous 4–6 months. Almost all (99, 96.1%) were men and 90 (90.9%) were MSM. All the men and two out of four women were Caucasian with a median age of 36 years (range 21–67). The median age was 36 years (range 21–67). Six individuals (6.1%) reported a history of injecting drug use (two MSM, two heterosexual men and two heterosexual women). The median CD4 cell count

Table 1. Comparison of features associated with patients in the cluster and non-cluster groups.

	In cluster	Not in cluster	*P* value[a]
Number of patients	35	68	
Male sex	35 (100%)	64 (94.1%)	0.30
Age (years): median (range)	34 (23–54)	37 (21–67)	0.05
Number of contacts in 3 months before diagnosis: median (range)	3 (1–100)	2 (1–36)	0.006
Homosexual risk group	32 (97.0%)	58 (85.3%)	0.10
Highest reported risk in the 3 months before diagnosis of PHI			
Unprotected oral intercourse	25 (78.1%)	36 (73.5%)	0.83
Protected anal intercourse	2 (6.3%)	5 (10.2%)	0.70
Unprotected anal intercourse	28 (87.5%)	32 (65.3%)	0.05
Unprotected vaginal intercourse	0 (-)	8 (16.3%)	0.02
STI in 3 months before diagnosis			0.31
Yes	15 (42.9%)	19 (27.9%)	
No	18 (51.4%)	37 (54.4%)	
Not known	2 (5.7%)	12 (17.7%)	0.13
CD4 cell count (cells/mm^3): median (range)	612 (195–1477)	474 (196–1259)	0.005
CD4 cell percentage: median (range)	31 (12–40)	26.5 (7–42)	0.003
Viral load ($\log_{10}$ copies/ml): median (range)	4.97 (2.03–6.00)	4.94 (2.30–6.00)	0.90

[a]Entries in table are *n* (%) unless otherwise specified.

(available in 101/103) was 526 copies/ml (range 195–1477) and the median CD4 cell percentage (available in 81/103) was 28 (7–42). The median HIV viral plasma load was log 4.95 copies/ml (2.03–6.00). Thirteen MSM (12.6% of total patients) were infected with viruses that contained primary antiretroviral resistance-associated mutations. STI were diagnosed concurrently with PHI in 34 of the 89 individuals (38.2%) for whom information was available. Among the 90 MSM, 61 (68%) reported unprotected anal intercourse (UAI) in the 3 months before PHI diagnosis; no information was available regarding sexual practices in the period preceding this.

Cluster comparison

Viruses from 35 out of 103 individuals (34%) appeared within 15 transmission clusters, comprising one cluster of five individuals, two of three and 12 of two (full results shown in Table 1 and Fig. 1). All were men and 32 (97%) were MSM. For individuals within 11 out of 15 clusters, the diagnosis of PHI was made within 12 months of each other, giving supporting evidence that transmission occurred during the PHI period. Those in the cluster group had a higher CD4 cell count ($P = 0.005$), higher CD4 cell percentage ($P = 0.003$), were younger ($P = 0.05$), reported a higher number of different sexual contacts in the previous 3 months ($P = 0.006$), and were more likely to have engaged in UAI in the 3 months before the PHI diagnosis ($P = 0.05$) in comparison to those individuals not within a cluster. High rates of STI at the time of PHI were observed in both groups, with a trend towards higher rates in those individuals with viruses in a cluster (42.9 versus 27.9%, $P = 0.13$). Multivariable logistic regression analyses identified the CD4 cell percentage [odds ratio (OR) 1.14, 95% confidence interval (CI) 1.04–1.23, $P = 0.003$] and having more than five sexual partners (OR 3.38, 95% CI 1.13–10.10, $P = 0.03$) as the only independent predictors of belonging to a cluster. Six individuals (17%) had antiretroviral-associated resistance mutations, of whom two (both T215D in reverse transcriptase) belonged to a linkage pair.

Conclusion

In conclusion, the high rates of clustering observed within our study support the assertion that PHI may be associated with an increased risk of onward transmission. The associations we found with younger age, high rates of UAI, and sexual partner change identify this as a high-risk group for HIV transmission. There was a trend towards higher rates of STI in the cluster group on a background of extremely high STI rates in the study population, supporting the argument for increased STI surveillance, particularly of high-risk groups.

The highly significant correlation with CD4 cell counts may represent the early disease stage, or rapid contact tracing and testing of sexual partners of individuals diagnosed with PHI. The plasma viral load at diagnosis was not predictive of clustering, and it is possible that the seminal viral load in men is a more consistent correlate of infectiousness, particularly in the context of genital tract inflammation, with plasma/genital tract discordance playing an important role [7–10]. The presence of the same antiretroviral resistance mutation in one cluster pair, neither of whom had received antiretroviral therapy, illustrates the potential for the secondary spread of such resistant strains, as we have previously documented [33,34]. Our results do not exclude the possibility of a

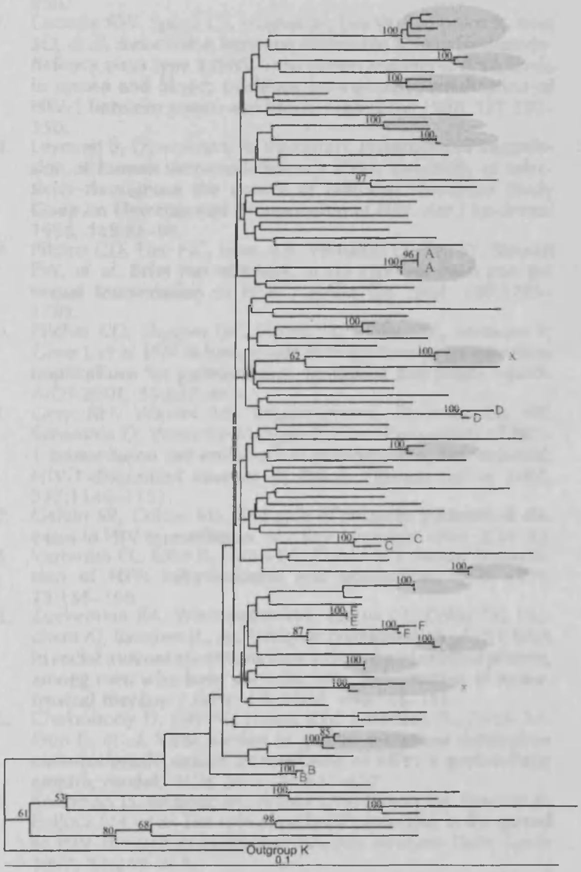


Fig. 1. Maximum likelihood phylogenetic tree based on *pol* sequences from 103 individuals with primary HIV-1 infection. Possible transmission clusters are circled. Linkages confirmed by clinical data are indicated by a red cross. Transmission clusters were identified if the bootstrap value was equal or greater than 99% and the average genetic distance (i.e. branch length) was lower than 0.015 nucleotide substitutions per site. Linkages confirmed by clinical data are indicated by a red cross. Six pairs of multiple sequences from single patients were used as controls for relatedness and are indicated by letters (e.g. A indicates multiples sequences from patient A). Bootstrap values higher than 50% are indicated on the branches.

common source for each cluster, rather than transmission within clusters. However, a phylogenetic tree comprising viruses from these 103 primary infections, together with more than 2000 *pol* sequences from prevalent infections throughout the UK only identified one further potential linkage, and that involved a primary infection case not within an existing cluster (data not shown).

Only 31 of the non-cluster group (64.6%) reported UAI, but it should be noted that this is only in the time window 3 months before diagnosis with PHI. Interestingly, routinely collected data on recent sexual contacts only confirmed three of the linkage pairs that were revealed in the phylogenetic analysis, emphasizing the high rates of anonymous sexual partners and the difficulty in obtaining a reliable sexual history.

Our results provide further evidence that the active management of primary infection will reduce HIV transmission. HIV prevention programmes have been heavily focused on protecting susceptible individuals, but accumulating biological and modelling data suggest that reducing the infectiousness of HIV-positive individuals may also be an effective strategy. A large proportion of PHI remains undiagnosed in the community [35,36], and these findings support the view that as a disease stage PHI represents a major public health threat. Efforts should be re-focused on improving rates of diagnosis of individuals during PHI, timely contact tracing, risk reduction, the management of STI, and possibly early treatment with antiretroviral agents in an effort to break transmission networks during this unique and possibly crucial stage of HIV infection [37]. Furthermore, consideration should be given in information and awareness campaigns to highlight the possible symptoms of PHI in groups with high rates of onward transmission, to encourage such individuals to present to appropriate healthcare providers to enable the timely diagnosis and management of early infection.

Contributors

M.F. and D. Pillay devised the study. D. Pao, M.F. and G.D. recruited patients for the study. D. Pao, M.F., S.H., C.S. and D. Pillay wrote the manuscript. P.A.C. undertook sequencing and curated the sequences. S.H. undertook the phylogenetic analyses. G.M. undertook the STARHS analysis. C.S. undertook statistical analyses.

Acknowledgements

The authors would like to thank the Health Protection Agency for funding, and the patients for agreeing to enter this study.

References

1. Joint United Nations Programme on HIV/AIDS, http://www.unaids.org, copyright 2004. Accessed 30 October 2004.
2. Health Protection Agency, http://www.hpa.org.uk, established 2002. Accessed 30 October 2004.
3. Murphy G, Charlett A, Jordan LF, Osner N, Gill ON, Parry JV. **HIV incidence appears constant in men who have sex with men despite widespread use of effective antiretroviral therapy.** *AIDS* 2004, **18**:265–272.
4. Kaufmann GR, Cunningham P, Kelleher AD, Zaunders J, Carr A, Vizzard J, *et al.* **Patterns of viral dynamics during primary human immunodeficiency virus type 1 infection.** *J Infect Dis* 1998, **178**:1812–1815.
5. Lindback S, Karlsson AC, Mittler J, Blaxhult A, Carlsson M, Briheim G, *et al.* **Viral dynamics in primary HIV-1 infection.** *AIDS* 2000, **14**:2283–2291.
6. Little SJ, McLean AR, Spina CA, Richman DD, Havlir DV. **Viral dynamics of acute HIV-1 infection.** *J Exp Med* 1999, **190**:841–850.
7. Coombs RW, Speck CE, Hughes JP, Lee W, Sampoleo R, Ross SO, *et al.* **Association between culturable human immunodeficiency virus type 1 (HIV-1) in semen and HIV-1 RNA levels in semen and blood: evidence for compartmentalization of HIV-1 between semen and blood.** *J Infect Dis* 1998, **177**:320–330.
8. Leynaert B, Downs AM, de Vincenzi I. **Heterosexual transmission of human immunodeficiency virus: variability of infectivity throughout the course of infection. European Study Goup on Heterosexual Transmission of HIV.** *Am J Epidemiol* 1998, **148**:88–96.
9. Pilcher CD, Tien HC, Eron JJ Jr, Vernazza PL, Leu SY, Stewart PW, *et al.* **Brief but efficient: acute HIV infection and the sexual transmission of HIV.** *J Infect Dis* 2004, **189**:1785–1792.
10. Pilcher CD, Shugars DC, Fiscus SA, Miller WC, Menezes P, Giner J, *et al.* **HIV in body fluids during primary HIV infection: implications for pathogenesis, treatment and public health.** *AIDS* 2001, **15**:837–845.
11. Gray RH, Wawer MJ, Brookmeyer J, Sewankambo NK, Serwadda D, Wabwire-Mangen F, *et al.* **Probablility of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai.** *Uganda Lancet* 2001, **357**:1149–1153.
12. Galvin SR, Cohen MS. **The role of sexually transmitted diseases in HIV transmission.** *Nat Rev Microbiol* 2004, **2**:34–42.
13. Vernazza PL, Eron JJ, Fiscus SA, Cohen MS. **Sexual transmission of HIV: infectiousness and prevention.** *AIDS* 1999, **13**:155–166.
14. Zuckerman RA, Whittington WL, Celum CL, Collis TK, Lucchetti AJ, Sanchez JL, *et al.* **Higher concentration of HIV RNA in rectal mucosa secretions than in blood and seminal plasma, among men who have sex with men, independent of antiretroviral therapy.** *J Infect Dis* 2004, **190**:156–161.
15. Chakraborty H, Sen PK, Helms RW, Vernazza PL, Fiscus SA, Eron JJ, *et al.* **Viral burden in genital secretions determines male-to-female sexual transmission of HIV: a probabilistic empiric model.** *AIDS* 2001, **15**:621–627.
16. Koopman JS, Jacquez JA, Welch GW, Simon CP, Foxman B, Pollock SM, *et al.* **The role of early HIV infection in the spread of HIV through populations.** *J Acquir Immune Defic Syndr* 1997, **14**:249–258.
17. Jacquez JA, Koopman JS, Simon CP, Longini IM Jr. **Role of the primary infection in epidemics of HIV infection in gay cohorts.** *J Acquir Immune Defic Syndr* 1994, 7:1169–1184.
18. Xiridou M, Geskus R, De Wit J, Coutinho R, Kretzschmar M. **The contribution of steady and casual partnerships to the incidence of HIV infection among homosexual men in Amsterdam.** *AIDS* 2003, **17**:1029–1038.
19. Pilcher CD, Eron JJ Jr, Vernazza PL, Battegay M, Harr T, Yerly S, *et al.* **Sexual transmission during the incubation period of primary HIV infection.** *JAMA* 2001, **286**:1713–1714.
20. Kahn JO, Walker BD. **Acute human immunodeficiency virus type 1 infection.** *N Engl J Med* 1998, **339**:33–39.
21. Dodds JP, Nardone A, Mercey DE, Johnson AM. **Increase in high-risk sexual behaviour among homosexual men, London 1996–1998: cross sectional, questionnaire study.** *BMJ* 2000, **320**:1510–1511.
22. Colfax G, Buchbinder SP, Cornelisse PGA, Vittinghoff E, Mayer K, Celum C. **Sexual risk behaviours and implications for**

secondary HIV transmission during and after HIV seroconversion. *AIDS* 2002, 16:1529–1535.

23. Colfax G, Mansergh G, Vittinghoff E, Guzman R, Marks G, Buchbinder S. **Drug use and high-risk sexual behaviour among circuit party participants.** In: *XIIIth International Conference on AIDS*. Durban, 2000 [Abstract TuPeC3422].
24. Katz MH, Schwarcz SK, Kellogg TA, Klausner JD, Dilley JW, Gibson S, *et al.* **Impact of highly active antiretroviral treatment on HIV seroincidence among men who have sex with men: San Francisco.** *Am J Public Health* 2002, **92**:388–394.
25. Clements MS, Prestage G, Grulich A, Van De Ven P, Kippax S, Law MG. **Modeling trends in HIV incidence among homosexual men in Australia 1995–2006.** *J Acquir Immune Defic Syndr* 2004, **35**:401–406.
26. Little SJ, Holte S, Routy JP, Daar ES, Markowitz M, Collier AC, *et al.* **Antiretroviral-drug resistance among patients recently infected with HIV.** *N Engl J Med* 2002, **347**:385–394.
27. Duwe S, Brunn M, Altmann D, Hamouda O, Schmidt B, Walter H, *et al.* **Frequency of genotypic and phenotypic drug-resistant HIV-1 among therapy-naive patients of the German seroconverter study.** *J Acquir Immune Defic Syndr* 2001, **26**:266–273.
28. Grant RM, Hecht FM, Warmerdam M, Liu L, Liegler T, Petropoulos CJ, et *al.* **Time trends in primary HIV-1 drug resistance among recently infected persons.** *JAMA* 2002, **288**:181–188.
29. Pillay D, Cane PA, Shirley J, Porter K. **Detection of drug resistance associated mutations in HIV primary infection within the UK.** *AIDS* 2000, **14**:906–908.
30. Janssen RS, Satten GA, Stramer SL, Rawal BD, O'Brien TR, Weiblen BJ, et *al.* **New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes.** *JAMA* 1998, **280**:42–48. Erratum in *JAMA* 1999; **281**: 1893.
31. Kothe D, Byers RH, Caudill SP, Satten GA, Janssen RS, Hannon WH, *et al.* **Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence.** *J Acquir Immune Defic Syndr* 2003, **33**:625–634.
32. Hue S, Clewley J, Cane P, Pillay D. **HIV-1 *pol* gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy.** *AIDS* 2004, **18**:719–728.
33. Yerly S, Vora S, Rizzardi P, Chave JP, Vernazza PL, Flepp M, *et al.* **Acute HIV infection: impact on the spread of HIV and transmission of drug resistance.** *AIDS* 2001, **15**:2287–2292.
34. Taylor S, Cane P, Hue S, Xu L, Wrin T, Lie Y, *et al.* **Identification of a transmission chain of HIV type 1 containing drug resistance-associated mutations.** *AIDS Res Hum Retroviruses* 2003, **19**:353–361.
35. Melzer M, Brown M, Mullen J, O'Shea S, Chrystie I, Banatvala J. **Undiagnosed symptomatic primary HIV infections in South London [Letter].** *J Infect* 2001, **42**:297–298.
36. Pilcher CD, McPherson JT, Leone PA, Smurzynski M, Owen-O'Dowd J, Peace-Brewer AL, *et al.* **Real-time, universal screening for acute HIV infection in a routine HIV counseling and testing population.** *JAMA* 2002, **288**:216–221.
37. Cates W Jr, Chesney MA, Cohen MS. **Primary HIV infection – a public health opportunity.** *Am J Public Health* 1997, **87**:1928–1930.

HIV-1 *pol* gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy

Stéphane Hué[a,b,c], Jonathan P. Clewley[c], Patricia A. Cane[a] and Deenan Pillay[a,b,c]

Objectives: We wished to assess the potential of using HIV-1 *pol* gene for the identification of transmissions events by phylogenetic means in the era of antiretroviral drug selective pressure.

Design: The relatedness of the viruses within a large database of *pol* sequences generated from HIV-1 infected individuals from the UK was reconstructed by phylogenetic analyses.

Methods: A total of 140 *pol* sequences were selected out of the 2500 database entries, on the basis of a pairwise genetic distance higher than 95%. Neighbour Joining and Maximum Likelihood trees were implemented. Trees were reconstructed after exclusion of codon positions associated with drug resistance from the original *pol* alignment. Trees based on the corresponding *env* and *gag* genes were implemented to confirm the linkages.

Results: Up to 23 transmission clusters were identified, supported by high bootstrap values (> 99), congruent epidemiological data and/or similar drug resistance motifs. The topology of the tree was consistent after exclusion of the drug resistance associated codons. Identical topologies were obtained in trees implemented from *gag* and *env* genes alignments.

Conclusions: Despite its genetic conservation, the HIV-1 *pol* gene holds sufficient variability to permit the phylogenetic reconstruction of transmissions. Identical clusters were obtained whichever of the three principal genes is considered and no bias was induced by the presence of drug resistance mutations. These findings demonstrate the important epidemiological information inherent within routinely collected laboratory data, which can assist in estimating rates of recent HIV-1 transmission within a population.

AIDS 2004, **18**:719–728

Keywords: HIV-1 transmission, pol gene, phylogenetic analyses, drug resistance

Introduction

The reasons for the extensive genetic variability of HIV-1 are several. The absence of proofreading activity of the viral reverse transcriptase (RT) [1], a fast turnover of virions during replication [1–3], and the genomic recombination occurring in infected cells [4] are amongst the factors responsible for the heterogeneity of the HIV-1 genome. Because HIV-1 genetic variation plays a major role in the worldwide AIDS epidemic, molecular techniques have shaped the strategies used in HIV-1 studies such as vaccine development

From the [a]HPA Antiviral Susceptibility Reference Unit, Division of Immunity and Infection, University of Birmingham and Regional Health Protection Agency, Birmingham Heartlands Hospital, [b]Department of Virology, Royal Free and University College Medical School, University College London, [c]Specialist and Reference Microbiology Laboratory, Health Protection Agency, Colindale, London, UK.

Correspondence to S. Hué, Department of Virology, Royal Free and University College Medical School, University College London, Windeyer Building, 46 Cleveland Street, London W1T 4JF, UK.

Received: 8 August 2003; revised: 9 October 2003; accepted: 27 October 2003.

DOI: 10.1097/01.aids.0000111455.61782.c3

ISSN 0269-9370 © 2004 Lippincott Williams & Wilkins

[5–7], monitoring of antiretroviral drug resistance [8,9] and the reconstruction of transmission events [10–13].

To date, phylogenetic reconstruction has been the favoured approach for providing evidence of HIV-1 transmission, not only for epidemiological purposes but also for the resolution of legal cases [14–16]. Hence the choice of the most appropriate genetic region of HIV-1 for phylogenetic analysis is a crucial issue and is still subject to debate [17–19]. Ultimately, complete genome analysis would be applied to transmission studies. However, as there are relatively few full-length sequences available and phylogenetic analyses are restricted by the cost of sequencing, appropriate background material, and by computational power, the sequence length and genetic region of choice need to be carefully considered together in order to guarantee a strong phylogenetic signal. Hence most phylogenetic studies undertaken to date have relied on the V3 loop region of the *env* gene, and to a lesser degree on fragments of the *gag* gene [10,11,13]. Nonetheless it has been argued that fragments covering the V3 loop are too short or too variable to allow robust inferences on the genetic relatedness of specimens [19]. Also the limited number of *gag* sequences in public databases makes use of this gene problematic.

By contrast, the region spanning the protease and RT genes is routinely sequenced in the clinical context of genotypic drug resistance testing and a large body of data is now being generated. Successful attempts to determine HIV-1 subtypes on the basis of the protease and the RT genes have been reported, so long as the fragment used is long and variable enough to counterbalance the lack of genetic constraint [20–22]. However, the *pol* gene remains unpopular for phylogenetic analyses due to its extreme genetic conservation and it is commonly considered suboptimal for the study of HIV-1 transmission histories [17].

The aim of the present work was to determine whether the *pol* gene holds sufficient genetic variability to allow the useful study of potential patterns of transmission. For these purposes, we explored a database containing more than 2500 HIV-1 *pol* sequences from drug-naive and experienced individuals in the UK. The potential linkages identified were compared with clusters obtained from more variable genetic regions of HIV-1 (i.e., the *gag* and *env* genes) and the influence of drug resistance related mutations in the process of phylogenetic reconstruction was assessed.

Material and methods

Study population

The *pol* sequences used for this study were generated from plasma samples collected from HIV-1 infected people in the UK by the Antiviral Susceptibility Reference Unit (ASRU), Health Protection Agency (HPA), Heartlands Hospital, Birmingham. The laboratory provides a service to clinics serving approximately 4000 treated patients (about 20% of the UK treated population), of which 10–20% are tested for resistance per year. The samples were submitted for routine genotypic resistance testing between 1999 and 2001, and include samples from acute infections, chronic but drug-naive infections and from patients at the time of therapy failure. Clinical information on the patients was available for most samples, including the date of collection, geographic area, reason for analysis and viral load, as well as molecular information such as subtype of the virus or genotypic patterns of drug resistance.

For the purposes of the study, data were anonymized prior to analysis, and the research was approved by the HPA Ethics Committee. However, specific consent was requested from patients appearing within clusters in order to document potential sexual contacts, whilst blinding clinicians and patients to the laboratory data. Such epidemiological information was only obtained from a minority of patients.

PCR and sequencing

pol variability

The region spanning the protease gene and the 235 first codons of the RT gene were amplified from plasma virus by random primed reverse transcription followed by nested PCR with the Qiagen Taq PCR mastermix kit (Qiagen Inc., Hiden, Germany) as described previously [12]. The subsequent amplicons were sequenced using either ABI377 or Beckman CEQ2000 protocols.

gag and env variability

Where cDNA was available, regions spanning the *gag* and *env* genes were amplified and sequenced. Consequently *gag* and *env* sequencing was undertaken for samples involved in clusters of *pol* sequences (n = 23), sequential samples from the same individuals used as controls (n = 6), and randomly selected samples where the *pol* gene was unrelated to other sequences (n = 23).

Two fragments of 690 and 550 base pairs, covering the p17/p24 region of the *gag* gene and the V3 loop region of the *env* gene respectively, were amplified by multiplex nested PCR from cDNA already used for *pol* gene amplification using Qiagen Taq PCR mastermix and the following primers: forward outer primer for *gag* (position 790–812), 5′-ATGGGTGCGAGAGCGTC AGTATT-3′; reverse outer primer for *gag* (position 1818–1844), 5′-CCCTGACATGCTGTCATCATTT CTTCT-3′; forward inner primer for *gag* (position 886–908), 5′-CATCTAGTATGGGCAAGCAGGGA -3′; reverse inner primer for *gag* (position 1609–1634), 5′-ATGCTGACAGGGCTATACATTCTTAC-3′; forward outer primer for *env* (position 6557–6582), 5′-ATGGGATCAAAGCCTAAAGCCATGTG-3′; reverse outer primer for *env* (position 7782–7811), 5′-AGTGCTTCCTGCTGCTCCCAAGAACCC-3′; forward inner primer for *env* (position 6817–6845), 5′-ACCTCAGCCATAACACAAGCCTGTCCA-3′; reverse inner primer for *env* (position 7360–7381), 5′-TTGCAATAGAAAAATTCCCCTC-3′. The fragments were sequenced using either ABI 3100 or Beckman CEQ2000 protocols.

Genetic distances

In order to select a subset of closely related *pol* sequences within the database, the pairwise genetic distance between all sequences was computed under the general reversible time model [23] with invariable sites and gamma distribution (GTR+I+G), using the softwares Modeltest [24] and Paup* [25]. The GTR model of nucleotide substitution allows each possible substitution to have a different rate, with the constraint of being symmetrical, so that a substitution from a nucleotide *i* to *j* has to be the same as a substitution from *j* to *i*.

Phylogenetic reconstruction

In-frame multiple alignments of the *pol, gag* and *env* nucleotide sequences were constructed with the program Clustal X [26], then manually adjusted using the editing software BioEdit [27]. Sequences that could not be unambiguously aligned or were of insufficient length were excluded from the study.

Phylogenetic relationships between the *pol* sequences were estimated using successively the Neighbour Joining (NJ) [28] and Maximum Likelihood (ML) methods [29]. The alignment matrices were imported into the tree building software Paup*, and an initial NJ tree was built under the Hasegawa-Kishino-Yano (HKY85) model of evolution [30,31] with a transversion : transition ratio of 2 : 1. The best fitting model of nucleotide substitution was estimated on the basis of the NJ tree topology using a ML ratio test to compare up to 57 different models, as implemented by the software Modeltest version 3.0. The derived parameters of the selected model, together with the initial NJ tree, were then used to perform a heuristic search for a ML tree under the GTR+I+G model of DNA substitution. The robustness of the NJ trees was evaluated by bootstrap analysis [32], with 1000 rounds of replication. The protocol was repeated for the *gag* and *env* alignments. The proportions of invariable sites within the *pol, gag* and *env* alignments were 47.6%, 20.1% and 20.6% respectively. The shape parameters of the gamma distribution used for the reconstruction of the *pol, gag* and *env* ML trees were 1.04, 0.65 and 0.94 respectively.

In order to assess the potential bias induced by drug resistance associated substitutions on the reconstruction of the samples' relatedness, 46 codon positions known to be related to antiretroviral resistance [33,34] were excluded from the previous *pol* sequences alignment and a ML tree was implemented. Resistance mutation positions known as primary (or major) and secondary (or minor) were excluded. Primary mutations are known to lead to an alteration in drug binding by themselves, whereas secondary mutations do not have a significant effect on phenotype by themselves [34]. The phylogeny estimation, model testing and bootstrap procedures were performed with Paup*, as described above. The proportion of invariable sites and gamma distribution shape parameter used for the tree reconstruction were 51% and 1.08 respectively. The positions excluded from the *pol* alignment and the related drug resistance are listed in Table 1.

Sequence data

The nucleotide sequences used in the study were deposited into Genbank under the accession numbers AY362043–AY362180, AY360862–AY360910 and AY360911–AY360959 for *pol, gag* and *env* sequences respectively.

Results

Of the 2500 *pol* sequences generated on samples dated from 1999 to 2003, 140 were selected on the basis of the closest pairwise genetic distances, each representing a single patient with the exception of the 12 pairs or triplets of multiple sequences used as controls. Hence sequences sharing more than 95% similarity with one or more other entries from the database were selected for the study. Overall, the average inter-patient genetic variation amongst the sequences was 5.1% (range, 0–12.4%). Although several subtypes were represented within the subset of sequences, including subtype A, B, C, D, G and CRF01-AE, the vast majority were of subtype B (88%), reflecting the subtype distribution of prevalent infections in the UK at the time of the study. Each sequence was 963 base pairs long and spanned the entire protease gene and the first 223 codons of the RT gene.

The ML tree derived from the selected *pol* sequences is presented in Fig. 1. The tree was rooted against an HIV-1 subtype K sequence (accession number AJ249239) extracted from the Los Alamos HIV-1 Database (http://hiv-web.lanl.gov/). Twelve pairs or triplets of sequential sequences from a single patient were used as controls. Bootstrap values higher than 50% are indicated on the branches, reflecting the frequency with which a given branch occurred in 1000 bootstrap resampling.

Table 1. Drug resistance mutations in HIV-1 [34].

Amino acid substitutions associated with resistance to					
Protease inhibitor		NRTI		NNRTI	
Mutation	Prevalence in the data	Mutation	Prevalence in the data	Mutation	Prevalence in the data
L10F/V/I/R	15	**M41L**[a]	19	**L100I**	1
K20M/R	11	E44D	2	**K103N**	13
L24I	0	**A62V**	2	**V106A/M**	0
D30N	**0**	**K65R**	0	**V108I**	0
V32I	0	**D67N**	3	**Y181C/I**	2
L33F	1	**T69D**	4	**Y188C/L/H**	2
M36I	28	**K70R**	3	**G190A/S**	3
M46I/L	**0**	**L74V**	1		
I47V	0	**V75I**	0		
G48V	**0**	**F77L**	0		
I50V/L	**0**	Y115F	0		
F53L	0	**F116Y**	0		
I54V/M/L	1	V118I	0		
L63P	68	**Q151M**	0		
A71V/T	21	**M184V/I**	18		
G73S/A	0	L210W	6		
V77I	22	**T215Y/F**	15		
V82A/F/T/S	**0**	**K219Q/E**	2		
I84V	**0**				
N88D/S	0				
L90M	**4**				

[a] Primary mutations are indicated in bold. NRTI, Nucleoside reverse transcriptase inhibitor; NNRTI, non-nucleoside reverse transcriptase inhibitor.

A total of 23 possible transmission clusters were identified from the tree topology shown in Fig. 1. The criteria used were determined by plotting the supporting bootstrap score of each terminal cluster against the within-average branch length calculated from the ML tree topology (Fig. 2). Clusters were highly suspicious for true linkages when fulfilling the following two conditions: a bootstrap value equal or greater to 99%; and an average genetic distance (i.e., branch length) lower than 0.015 nucleotide substitutions per sites within the cluster. There was no significant distinction between intra-patient (i.e., control) and inter-patient (i.e., linked) sequences in terms of genetic distance. All controls conformed to these criteria, with the exception of the multiple sequences belonging to patients 7 and 8, whose clusters were supported by lower bootstrap values (95% and 92% respectively). The reason why these two clusters failed to fit the criteria remains unclear. The relative low bootstrap score attributed to samples from patient 7 could be explained by the presence of an archive sequence, subsequently becoming the majority plasma population within the follow-up samples. For instance, a virus originating many years previously may emerge following a treatment interruption. Unfortunately, matched *gag* and *env* sequences could not be generated for these samples.

All putative transmission events involved subtype B sequences. Since bootstrap scores are known to be influenced by the number of taxonomic units considered in a tree, the robustness of the 'non-B clade' is likely to be artificially high due to under-representation within the dataset, and we therefore excluded these subtypes from our categorization of potential clusters.

Where informed consent was obtained from the patient involved, epidemiological evidence of linkage between individuals was documented in order to corroborate the findings from the initial phylogenetic analysis and drug resistance patterns within clusters. Both primary and secondary mutations associated with antiretroviral resistance were considered [9,34]. Although not essential to prove transmissions, such information is important to verify the approach developed in the present study. These data are listed for each cluster in Table 2. Where appropriate information was obtained, three clusters were supported by evidence of epidemiological linkage (clusters 3, 8 and 14). Within clusters, similar drug resistance associated mutations (including secondary mutations) were observed within 14 out of 23 clusters. Four clusters appear to identify transmission of viruses with key resistance mutations to a drug-naive individual (clusters 6, 10, 18 and 21). In five other clusters (numbers 1, 4, 11, 12 and 16) such mutations in the drug-experienced individual were not reproduced in the drug-naive partner.

Since the *pol* gene is under intense selective pressure by antiviral therapy, it might be expected that the presence of drug resistance mutations biases phylogenetic recon-

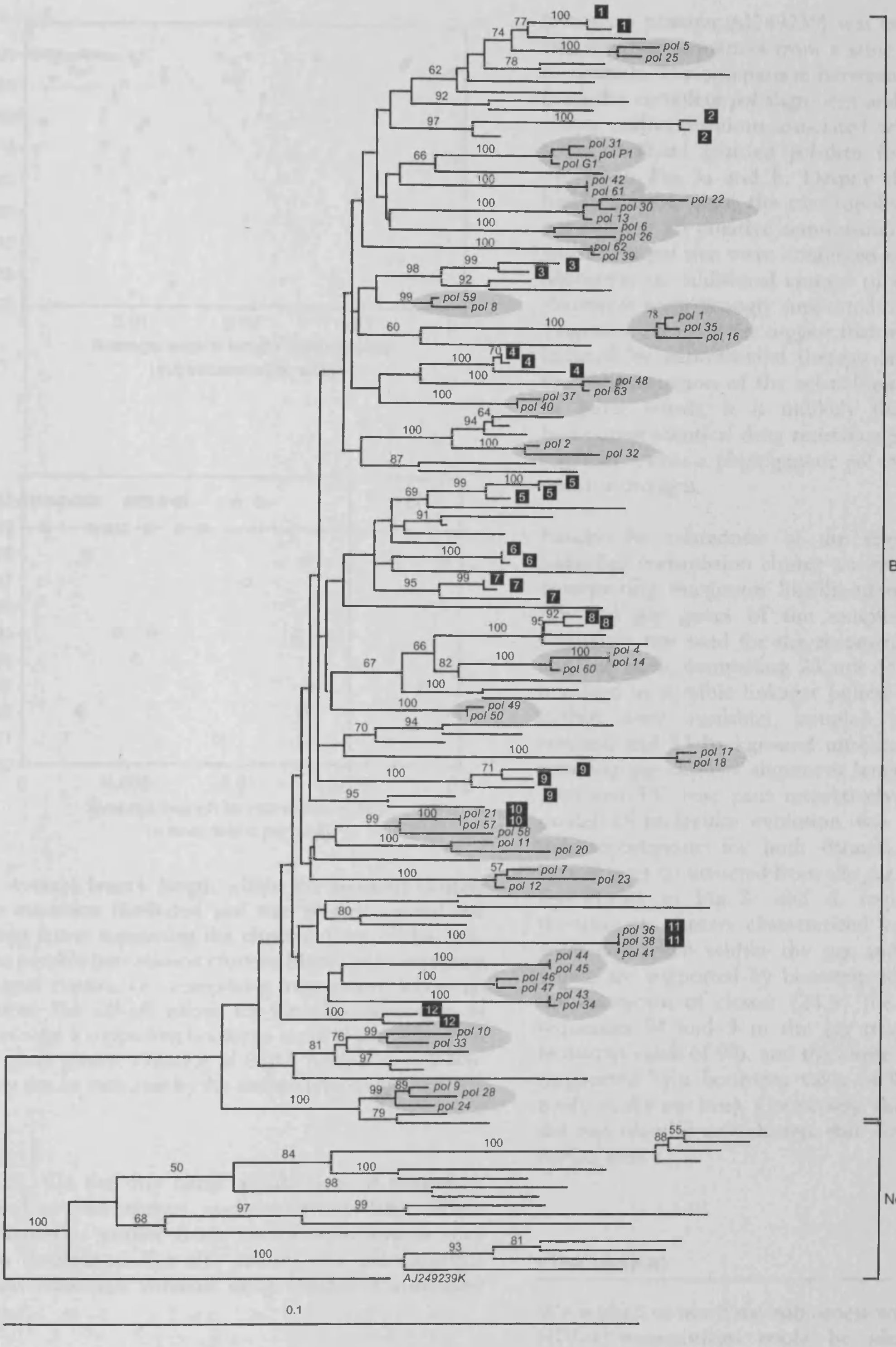


Fig. 1. Maximum likelihood tree representing the phylogenetic relationships between HIV-1 *pol* sequences from the HPA resistance-testing database. The tree was constructed according to the GTR+I+G model of evolution and rooted against a HIV-1 subtype K sequence (AJ249239K) extracted from the Los Alamos HIV database. Bootstrap values higher than 50% are indicated on the branches. Clusters involving potential transmission events are indicated by a circle. Twelve pairs or triplets of multiple sequences from a same patient were used as control. These sequences are tagged by figures in black boxes (e.g., **1** indicates multiples sequences from patient 1). For clarity, only branches involved in possible linkages are labelled.

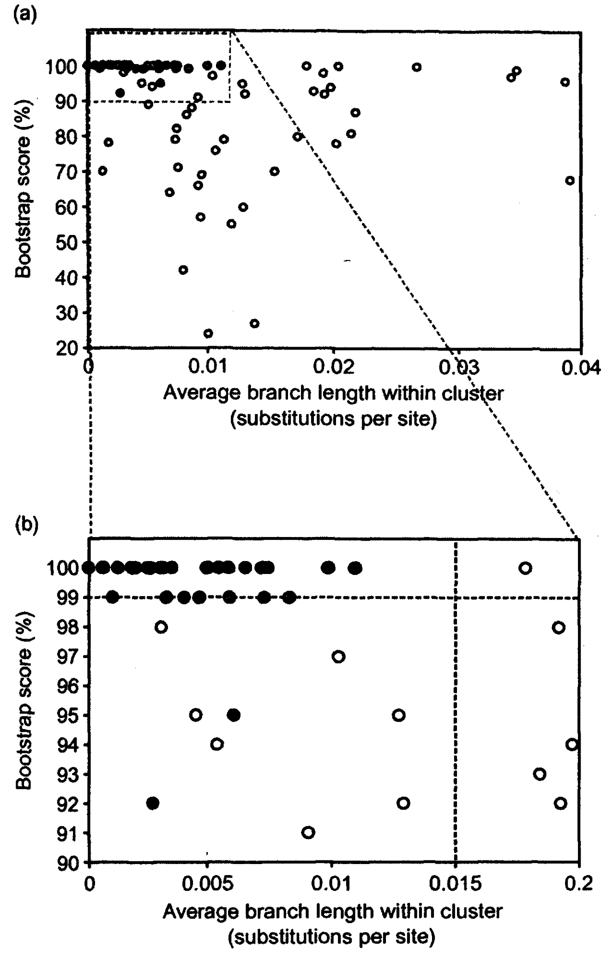


Fig. 2. Average branch length within the terminal clusters of the maximum likelihood *pol* tree plotted against the bootstrap scores supporting the clusters. Grey circles indicate the possible transmission clusters; black circles represent the control clusters, i.e., comprising intra-patient follow up sequences. The cut-off values for the characterization of linkages were a supporting bootstrap score higher than 99% and a mean genetic distance of 0.015 nucleotide substitutions per site, as indicated by the dashed lines in (b).

struction. On the one hand, similar sets of mutations may lead to convergence, and conversely, large differences between viruses from transmission events may lead to divergence. For this reason, the *pol* sequence tree was reassessed without drug resistance associated codons.

The ML tree derived from the *pol* alignment after removal of 46 codon positions most commonly associated with drug resistance is presented in Fig. 3b. The tree was implemented according to the GTR+I+G model of nucleotide substitutions. As with the previous reconstruction, an HIV-1 subtype K *pol* sequence (accession number AJ249239) was used as an outgroup and multiple sequences from a same patient were used as controls. The comparison between the trees obtained from the complete *pol* alignment and the *pol* alignment where codon positions associated with drug resistance were excluded (named *pol*-drm for convenience) is shown in Fig. 3a and b. Despite the deletion of 46 highly variable sites, the two topologies were congruent and the 23 putative transmission clusters identified within the *pol* tree were conserved in the *pol*-drm tree. Moreover no additional clusters to those based on *pol* sequences were strongly supported by bootstrap scores. Together, these results suggest that resistance mutations induced by antiretroviral therapy are unlikely to bias the reconstruction of the relatedness between samples. In other words, it is unlikely that unrelated virus harbouring identical drug resistance patterns will cluster together within a phylogenetic *pol* tree, leading to false positive linkages.

Finally the relatedness of the sequences within an identified transmission cluster was further confirmed by constructing maximum likelihood trees based on the *env* and *gag* genes of the samples. A total of 49 sequences was used for the reconstruction of both *gag* and *env* trees, comprising 23 out of the 53 sequences involved in possible linkages (where stored samples or cDNA were available), coupled to three pairs of controls and 23 background unrelated sequences. The resulting *gag* and *env* alignment lengths were 747 base pairs and 557 base pairs respectively. The GTR+I+G model of molecular evolution was found to be the most appropriate for both datasets. Maximum likelihood trees constructed from the *gag* and *env* sequences are shown in Fig. 3c and d, respectively. The 11 transmission clusters characterized within the *pol* tree were conserved within the *gag* and *env* trees, all of which are supported by bootstrap scores of 100, with the exception of cluster {24,9} (i.e., that comprising sequences 24 and 9 in the *gag* tree; supported by a bootstrap value of 98), and the clusters 37,40 and 13,22 (supported by a bootstrap value of 96 and 98 respectively in the *env* tree). Conversely, the *gag* and *env* trees did not identify any clusters that were not present in the *pol* tree.

Discussion

We wished to assess the robustness with which possible HIV-1 transmissions could be identified from *pol* sequences, despite the relative conservation of this gene. Since the sequences used in the present study correspond to standard amplicons generated for routine resistance testing, the high availability of such fragments in regional databases may provide useful datasets for molecular epidemiological studies. The relatedness of

Table 2. Epidemiological and drug resistance mutation information from the 23 clusters of *pol* sequences.

				Resistance associated mutations		
Cluster	Sequences	Year of sampling	Drug history	Protease inhibitor	Reverse transcriptase inhibitor	*gag* and *env* linkage
1	*pol* 5	2000	Experienced	L10V, L63T	G190A	Yes
	pol 25	2001	Naive	L10V, L63S	None	Yes
2	*pol* 29	2001	Naive	None	None	n.a.
	pol 31	2001	Naive	None	None	n.a.
	pol P1	2001	Naive	None	None	n.a.
3	*pol* 42	2001	Naive	L63P	None	n.a.
	pol 61	2001	Naive	L63P	None	n.a.
4	*pol* 13	2000	Naive	L10V[a], M36I	None	Yes
	pol 22	2001	Experienced	L10V[a], M36I	**M184V, Y188L**	Yes
	pol 30	2001	Naive	L10V[a], M36I	T69I[a]	Yes
5	*pol* 6	2000	Experienced	None	None	n.a.
	pol 26	2001	Naive	None	None	n.a.
6	*pol* 39	2002	Experienced	M36L, L63P	**T69N**	n.a.
	pol 62	2000	Naive	M36L, L63P	**T69N**	n.a.
7	*pol* 8	2000	Experienced	L63P	None	n.a.
	pol 59	1999	Experienced	L63P	None	n.a.
8	*pol* 1	2000	Experienced	L63T	None	Yes
	pol 16	2001	Naive	L63T	None	Yes
	pol 35	2002	Naive	L63T	None	n.a.
9	*pol* 48	2001	Naive	L63H, A71V, V77I, I93L	None	n.a.
	pol 63	2001	Experienced	L63H, A71V, V77I, I93L	None	n.a.
10	*pol* 37	2002	Naive	L63P	**M41L, T215Y**	Yes
	pol 40	1998	Experienced	L63P	**M41L, T215C**	Yes
11	*pol* 2	2000	Naive	I93L	None	n.a.
	pol 32	2001	Experienced	I93L	**A62V, K65R, L74V, G190S**	n.a.
12	*pol* 4	2000	Naive	L10V, I93L	None	Yes
	pol 14	2000	Naive	L10V, I93L	None	Yes
	pol 60	2001	Experienced	L10V, L63P, A71V, I93L	K103N	n.a.
13	*pol* 49	2000	Naive	M36I, L63P, I93L	None	n.a.
	pol 50	2001	Experienced	M36I, L63P, I93L	None	n.a.
14	*pol* 17	2001	Naive	M36I, L63P, V77I, I93L	None	Yes
	pol 18	2001	Experienced	L63P, V77I, I93L	None	Yes
15	*pol* 21	2001	Experienced	L10I, K20R, M36I, L63S, I93L	None	n.a.
	pol 57	2001	Experienced	L10I, L63C, I93L	None	n.a.
	pol 58	2000	Naive	L10I, K20R, L63S, A71T, I93L	None	n.a.
16	*pol* 11	2000	Naive	L10I, L63C, I93L	None	n.a.
	pol 20	2001	Experienced	L10I, L63C, I93L	**M41L, V118I, L210W, T215Y**	n.a.
17	*pol* 7	2000	Experienced	L10I, L63P, V73I, I93L	None	n.a.
	pol 12	1998	Naive	L10I, L63P, V73I, I93L	None	n.a.
	pol 23	2001	Experienced	L10I, L63P, V73I, I93L	L210F	n.a.
18	*pol* 36	2001	Naive	K20R, M36I, L63A	**M41L, T215E †**	Yes
	pol 41	2001	Naive	K20R, M36I, L63A	**M41L, T215E †**	Yes
19	*pol* 44	2002	Experienced	M36I	**T215D**	Yes
	pol 45	2002	Experienced	M36I	**T215D**	Yes
20	*pol* 46	2002	Experienced	L63P	**T69A**	Yes
	pol 47	2002	Experienced	L63P	**T69A**	Yes
21	*pol* 34	2002	Experienced	L10V, L63P	**T215D**	Yes
	pol 43	2000	Naive	L10V, L63P	**T215D**	Yes
22	*pol* 10	2000	Naive	L63P, I93L	None	n.a.
	pol 33	2001	Naive	L63P, I93L	None	n.a.
23	*pol* 9	2001	Naive	L10I, L33I, L63T, A71T, I93L	A98S	Yes
	pol 24	2000	Naive	L10I, L33I, L63T, A71T, I93L	A98S	Yes
	pol 28	2000	Naive	L10I, L33I, L63T, A71T, I93L	A98S	n.a.

[a] Atypical mutation at the given codon. Primary mutations are indicated in bold.

the sequences in our database was reconstructed by phylogenetic analyses, on the basis of different genetic regions within the *pol*, *gag* and *env* genes. Twenty-three possible transmission clusters were identified within the *pol* ML tree topology, supported by high bootstrap values (> 99), congruent epidemiological data and similar drug resistance patterns. All clusters were conserved when codon positions associated with drug resistance were removed from the original *pol* alignment. Finally, trees constructed with the *env* and the *gag* regions of the samples were consistent with the results obtained with the *pol* region and the same transmission clusters were identified.

It has recently been suggested that the *pol* gene is suboptimal for reconstructing transmission events [17],

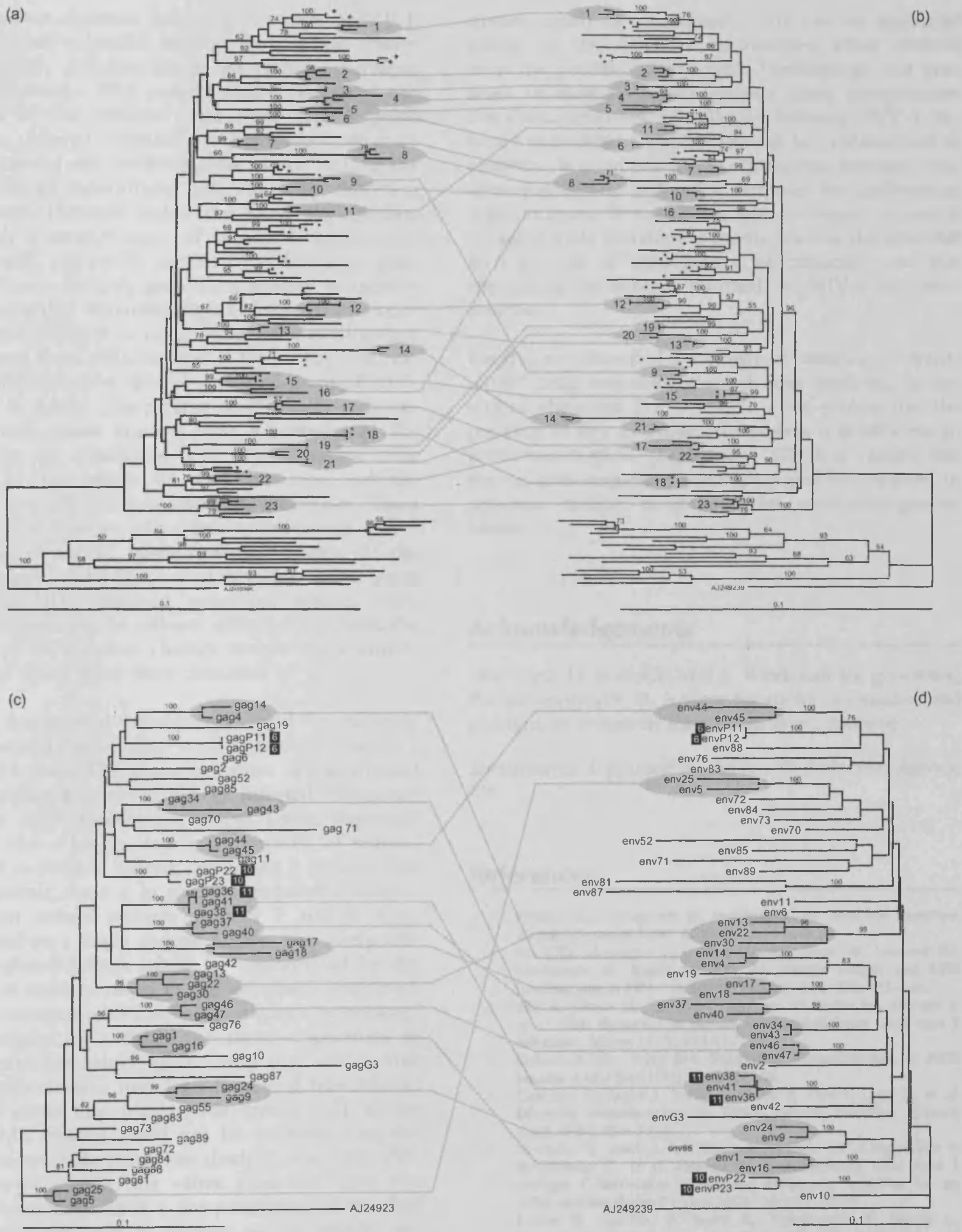


Fig. 3. Phylogenetic trees derived from the HIV-1 *pol* sequences before (a) and after (b) exclusion of codon positions associated with genotypic antiretroviral resistance, as well as derived from the *gag* (c) and *env* (d) regions of the samples. The maximum likelihood method was used to build the trees under the GTM+I+G model of evolution. An HIV-1 subtype K sequence (AJ249239K) was used as the outgroup. Transmission clusters identified within the *pol* tree topologies were circled, numbered (see Table 2) and linked by a line. When present, these clusters were circled and linked between the *gag* and *env* trees. The multiple sequences from a same patient used as controls are indicated by stars (a, b) or by figures in black boxes (c, d). Bootstrap values of 50% or greater are indicated on the branches. For clarity, only branches involved in possible linkages are labelled.

as the genetic distance between protease and RT sequences from unrelated individuals may not always be significantly different from the distance between related individuals. The present study compared the topologies of tree obtained with three HIV-1 genes known to undergo distinctive evolutionary dynamics (i.e., *pol*, *gag* and *env*), *pol* having the lowest and *env* the highest rate of substitution [35,36]. The clustering patterns were identical within the three phylogenetic trees, with a similar range of statistical significance. Consequently, our results suggest that HIV-1 *pol* gene holds sufficient intrinsic genetic variability to permit the reconstruction of transmission histories by phylogenetic means. Whether or not phylogenetic relationships characterized from protease and RT sequences should be confirmed by more variable genetic regions of HIV-1 is open to debate. The present work clearly indicates that identical results are obtained whichever of the three genes are considered, the trees obtained only differing by the length of their branches and the clustering patterns of distant unrelated sequences. These findings could have an immediate consequence in the monitoring of HIV-1 epidemiology. In view of the preponderance of HIV *pol* sequence data consequent on routine HIV resistance genotypic testing, these sequences could also be utilized effectively to track the presence of transmission clusters within the communities from which there were obtained.

We note that most of the sequences used for the study were generated from plasma samples obtained within a period of 3 years. The characterization of transmission patterns within a group of HIV-1 infected individuals might be more problematic when using sequences collected over a longer time span, because of within-individual evolution. Indeed, we noted a greater than average genetic distance in *pol* from sequential samples taken from control patients number 7 and 8. Also, when based on a single genetic region, the interpretation of inferred linkage might be undermined by the presence of recombination in the genomes considered. A further concern relates to the bottleneck represented by transmission of a single, or narrow spectrum of virions, especially when appreciating that within-host compartmentalization may lead to sexual transmission of genital rather than blood virus species [12]. Given that the ML method could not be performed on the whole data set, only sequences sharing more than 95% identity with a least one other sequence from the database were used. Such a pre-processing of the data could potentially have an impact on the results and favoured the presence of strongly supported clusters within the tree.

Although comparison with epidemiological data is important for the validation of the linkages characterized at the molecular level, this information remains hard to obtain and only three of the transmission clusters could be confirmed. This can be attributed mainly to the difficulty encountered when consent from the patients is requested. Furthermore, the presence of multiple sexual partners often compromises the characterization of linkages between HIV-1 infected individuals and networks can be problematical to establish. It is important to distinguish between epidemiological and individual purposes for undertaking these analyses. It is essential that informed consent is obtained from individual patients prior to the potential identification of their source of infection, and that appropriate security is afforded to HIV-1 sequence databases.

Finally, we identified a number of instances of transmitted drug resistance through our analyses, as described elsewhere [12,37,38]. It is self-evident that the presence of key mutations themselves is insufficient to prove transmission virologically. We now suggest that the *pol* gene sequence, itself generated for purposes of resistance testing, is adequate for such phylogenetic studies.

Acknowledgements

We thank D. Ratcliffe and J. Workman for generating the *pol* sequences, P. Kellam for useful discussions, and anonymous reviewers for constructive comments.

Sponsorship: Supported by the Health Protection Agency, UK.

References

1. Preston BD, Dougherty JP. **Mechanisms of retroviral mutation.** *Trends Microbiol* 1996, **4**:16–21.
2. Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M. **Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection.** *Nature* 1995, **373**:123–126.
3. Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, *et al.* **Viral dynamics in human immunodeficiency virus type 1 infection.** *Nature* 1995, **373**:117–122.
4. Robertson DL, Hahn BH, Sharp PM. **Recombination in AIDS viruses.** *J Mol Evol* 1995, **40**:249–259.
5. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, *et al.* **Diversity considerations in HIV-1 vaccine selection.** *Science* 2002, **296**:2354–2360.
6. Novitsky V, Smith UR, Gilbert P, McLane MF, Chigwedere P, Williamson C, *et al.* **Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design?** *J Virol* 2002, **76**:5435–5451.
7. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. **Evolutionary and immunological implications of contemporary HIV-1 variation.** *Br Med Bull* 2001, **58**:19–42.
8. Hanna GJ, D'Aquila RT. **Clinical use of genotypic and phenotypic drug resistance testing to monitor antiretroviral chemotherapy.** *Clin Infect Dis* 2001, 32:774–782.
9. Hirsch MS, Brun-Vezinet F, D'Aquila RT, , Hammer SM, Johnson VA, Kuritzkes DR, *et al.* **Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel.** *JAMA* 2000, **283**:2417–2426.
10. Arnold C, Balfe P, Clewley JP. **Sequence distances between env**

genes of HIV-1 from individuals infected from the same source: implications for the investigation of possible transmission events. *Virology* 1995, **211**:198–203.

11. Hayman A, Moss T, Simmons G, Arnold C, Holmes EC, Naylor-Adamson L, *et al.* **Phylogenetic analysis of multiple heterosexual transmission events involving subtype b of HIV type 1.** *AIDS Res Hum Retroviruses* 2001, **17**:689–695.
12. Taylor S, Cane P, Hue S, Xu L, Wrin T, Lie Y, *et al.* **Identification of a transmission chain of HIV type 1 containing drug resistance-associated mutations.** *AIDS Res Hum Retroviruses* 2003, **19**:353–361.
13. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. **Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis.** *Proc Natl Acad Sci USA* 1996, **93**: 10864–10869.
14. Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT, *et al.* **Molecular epidemiology of HIV transmission in a dental practice.** *Science* 1992, **256**:1165–1171.
15. Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlen M. **Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes.** *J Virol* 1994, **68**:5918–5924.
16. Machuca R, Jorgensen LB, Theilade P, Nielsen C. **Molecular investigation of transmission of human immunodeficiency virus type 1 in a criminal case.** *Clin Diagn Lab Immunol* 2001, **8**: 884–890.
17. Palmer S, Vuitton D, Gonzales MJ, Bassignot A, Shafer RW. **Reverse transcriptase and protease sequence evolution in two HIV-1-infected couples.** *J Acquir Immune Defic Syndr* 2002, **31**:285–290.
18. Smith TF, Waterman MS. **The continuing case of the Florida dentist.** *Science* 1992, **256**:1155–1156.
19. DeBry RW, Abele LG, Weiss SH, Hill MD, Bouzas M, Lorenzo E, *et al.* **Dental HIV transmission?** *Nature* 1993, **361**:691.
20. Pasquier C, Millot N, Njouom R, Sandres K, Cazabat M, Puel J, *et al.* **HIV-1 subtyping using phylogenetic analysis of pol gene sequences.** *J Virol Methods* 2001, **94**:45–54.
21. Kessler HH, Deuretzbacher D, Stelzl E, Daghofer E, Santner BI, Marth E. **Determination of human immunodeficiency virus type 1 subtypes by a rapid method useful for the routine diagnostic laboratory.** *Clin Diagn Lab Immunol* 2001, **8**:1018–1020.
22. Yahi N, Fantini J, Tourres C, Tivoli N, Koch N, Tamalet C. **Use of drug resistance sequence data for the systematic detection of non-B human immunodeficiency virus type 1 (HIV-1) subtypes: how to create a sentinel site for monitoring the genetic diversity of HIV-1 at a country scale.** *J Infect Dis* 2001, **183**: 1311–1317.
23. Rodriguez F, Oliver JL, Marin A, Medina JR. **The general stochastic model of nucleotide substitution.** *J Theor Biol* 1990, **142**:485–501.
24. Posada D, Crandall KA. **MODELTEST: testing the model of DNA substitution.** *Bioinformatics* 1998, **14**:817–818.
25. Swofford DL. *PAUP 4.0: Phylogenetic Analysis Using Parsimony (And Other Methods).* Sunderland, MA: Sinaur Associates; 2001.
26. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25**:4876–4882.
27. Hall TA. **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucleic Acids Symp Series* 2000, **41**:95–98.
28. Saitou N, Nei M. **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406–425.
29. Felsenstein J. **Maximum-likelihood estimation of evolutionary trees from continuous characters.** *Am J Hum Genet* 1973, **25**:471–492.
30. Hasegawa M, Kishino H, Yano T. **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**:160–174.
31. Kimura M. **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16**:111–120.
32. Felsenstein J. **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39**:783–791.
33. Shafer RW, Dupnik K, Winters MA, Eschleman SH. **A guide to HIV-1 reverse transcriptase and protease sequencing for drug resistance studies.** In: *HIV Sequence Compendium.* Edited by Sodroski J. Los Alamos: Los Alamos National Laboratory; 2000.
34. D'Aquila RT, Schapiro JM, Brun-Vezinet F, Clotet B, Conway B, Demeter LM, *et al.* **Drug resistance mutations in HIV-1.** *Top HIV Med* 2003, **11**:92–96.
35. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, *et al.* **Timing the ancestor of the HIV-1 pandemic strains.** *Science* 2000, **288**:1789–1796.
36. Li WH, Tanimura M, Sharp PM. **Rates and dates of divergence between AIDS virus nucleotide sequences.** *Mol Biol Evol* 1988, **5**:313–330.
37. Pillay D, Taylor S, Richman DD. **Incidence and impact of resistance against approved antiretroviral drugs.** *Rev Med Virol* 2000, **10**:231–253.
38. Ammaranond P, Cunningham P, Oelrichs R, Suzuki K, Harris C, Leas L, *et al.* **Rates of transmission of antiretroviral drug resistant strains of HIV-1.** *J Clin Virol* 2003, **26**:153–161.

CONCISE COMMUNICATION

Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections

David Pao[a], Martin Fisher[a], Stephane Hué[b,e], Gillian Dean[a], Gary Murphy[e], Patricia A. Cane[d,e], Caroline A. Sabin[c] and Deenan Pillay[b,e]

Objective: To study primary HIV-1 infections (PHI) using molecular and epidemiological approaches in order to assess correlates of transmission in this population.

Methods: Individuals with PHI were recruited prospectively from a discrete cohort of 1235 individuals under follow-up in a well-defined geographical area between 1999 and 2003. PHI was diagnosed by one of the following: negative HIV antibody test within 18 months, evolving antibody response, or application of the serological testing algorithm for recent HIV seroconversion. The *pol* gene was sequenced to identify genotypic resistance and facilitate molecular epidemiological analysis. Clinical data were collected and linked in an irretrievable fashion when informed consent was obtained.

Results: A total of 103 individuals with PHI diagnosed between 1999 and 2003 were included in the study; 99 (96%) were male and 90 (91%) were men who have sex with men. Viruses from 35 out of 103 (34%) appeared within 15 phylogenetically related clusters. Significant associations with clustering were: young age, high CD4 cell count, number of sexual contacts, and unprotected anal intercourse (UAI) in the 3 months before diagnosis ($P < 0.05$ for all). High rates of acute sexually transmitted infections (STI) were observed in both groups with a trend towards higher rates in those individuals with viruses within a cluster (42.9 versus 27.9%; $P = 0.13$).

Conclusion: High rates of partner change, UAI and STI are factors that facilitate onward transmission during PHI. More active identification of individuals during PHI, the management of STI and highly active antiretroviral therapy may all be useful methods to break transmission networks.

AIDS 2005, **19**:85–90

Keywords: Acute HIV infection, epidemiology, phylogenetic tree, sexually transmitted diseases

From the [a]Department of GU Medicine, Brighton and Sussex University Hospitals, Brighton, UK; [b]Centre of Virology, Division of Infection and Immunity, and [c]Department of Primary Care and Population Sciences, Royal Free and University College Medical School, University College London, London, UK; [d]HPA Antiviral Susceptibility Reference Unit, University of Birmingham Medical School, and Birmingham HPA, Heartlands Hospital, Birmingham, UK; and [e]Centre for Infections, Health Protection Agency, Colindale, UK.

Correspondence to Deenan Pillay, Centre of Virology, Division of Infection and Immunity, Royal Free and University College Medical School, Windeyer Institute, 46 Cleveland Street, London W1P 6DF, UK.

E-mail: d.pillay@ucl.ac.uk

Received: 22 June 2004; revised: 20 September 2004; accepted: 7 October 2004.

ISSN 0269-9370 © 2005 Lippincott Williams & Wilkins

Introduction

Worldwide, 4.2 million adults were estimated to have new HIV-1 infection in 2003 [1], although it is unclear whether these represent new diagnoses of chronic infection or recently acquired infections; nevertheless it is clear that strategies to interrupt the sexual transmission of HIV-1 are key to reducing the worldwide burden of HIV disease. Within the UK, most new diagnoses now represent imported infections [2]; however, continual incident infections among men who have sex with men (MSM) are evident [3]. Taken as a single disease stage, the overall efficiency of sexual transmission of HIV is low, but numerous biological and mathematical modelling studies predict much higher infectiousness during primary HIV infection (PHI) compared with chronic HIV infection.

Biologically, the high plasma viral load seen during PHI [4–6], which probably parallels semen viral load [7–10], is strongly correlated with the risk of sexual transmission [11] and therefore epidemic growth. Other factors that may increase transmission include sexually transmitted infections (STI) and host susceptibility [12,13]. The recent finding of higher concentrations of HIV-1 RNA in rectal mucosa than in blood or semen is also pertinent [14].

Mathematical models estimate the average probability of male–female transmission of HIV-1 per unprotected coital act to be between 0.0005 and 0.003% during chronic HIV infection [15], which in itself would not sustain an epidemic. By contrast, when the high viral load of PHI is taken into account, men with average semen viral load, without concurrent STI, would be expected to infect 7–24% of susceptible female partners during the first 2 months of infection (an eight to 10-fold increase from chronic HIV infection) [9]. According to male–male models, between 25 and 47% of new HIV infections may be transmitted during this period of initial HIV infection [16,17], possibly within steady as opposed to casual relationships [18]. In addition, these individuals are infectious before symptoms of PHI [19], may not even show symptoms of disease [20] (and therefore be unaware of the risk they pose to partners), and often engage in high-risk sexual practices [21,22] with a higher number of sexual contacts [23].

There is also increasing evidence that any decrease in the per-contact risk as a result of the increased availability of antiretroviral therapy appears to have been counterbalanced or overwhelmed by increases in risky sexual behaviour [24,25]. This is reflected in the transmission of primary resistant HIV strains, the prevalence of which approaches 20% in the UK and elsewhere [26–29].

In order to understand further the role played by PHI in sexual transmission we carried out phylogenetic characterization of PHI and collected relevant epidemiological data regarding sexual behaviour, clinical features and STI.

Methods

Study recruitment

Individuals were recruited from a cohort of 1235 HIV-positive patients attending a single genitourinary medicine unit for follow-up from 1999 to 2003. This prospective cohort included over 2100 patients with HIV infection, with 1235 being seen during the study period. Of these, 86% were caucasian, 89% were men, and the predominant route of transmission was sex between men (79%). The department is the sole local provider of HIV and STI care, and national surveillance data confirm that over 90% of individuals with HIV infection resident in the area attend this institution.

Individuals with PHI were identified by one or more of the following: previous negative HIV antibody test within 18 months, evolving Western blot or HIV antibody response, or application of the serological testing algorithm for recent HIV seroconversion (STARHS) assay. STARHS is a dual testing strategy in which specimens that are confirmed anti-HIV positive after detection by a sensitive screening assay are tested on an assay that has been altered to make it less sensitive. Specimens that are unreactive on this less sensitive assay are deemed to be recent infections, whereas specimens that are reactive in both assays are deemed to come from infections that are long standing [30]. At the time of HIV diagnosis the majority of individuals underwent a full STI screen.

Clinical data collection

In those from whom written informed consent was obtained, information regarding clinical status was collected from clinic case notes: the date of diagnosis, CD4 cell count, CD4 cell percentage, HIV viral load, the presence and nature of STI in the 3 months before the diagnosis of PHI (gonorrhoea, chlamydia, non-specific urethritis, early syphilis, herpes simplex), and the absence or presence of PHI symptoms. Information relating to the individual's HIV acquisition risk group, sexual behaviour (including estimated number and nature of sexual contacts in the 3 months before diagnosis of PHI) was also recorded. These data are routinely collected for all new HIV-1 diagnoses within this clinic.

Serological testing algorithm for recent HIV seroconversion and analysis

STARHS testing was performed using the bioMérieux Vironostika HIV-1 assay (bioMérieux UK Ltd., Basingstoke, UK) as previously described [31]. A standardized optical density for each specimen was determined. For this study a standardized optical density of less than

1.0 was used to identify recent infections, and this cut-off equates to an estimated seroconversion within the previous 4–6 months.

Phylogenetic analysis

The HIV *pol* gene was sequenced from plasma obtained at the time of HIV diagnosis. These sequences were used for phylogenetic analysis, a method previously shown by this group to have utility in reconstructing transmission events [32]. Full-length sequences from the protease gene (295 nt) and the first 230 codons of reverse transcriptase were aligned using the program Clustal X (available from http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html) and then adjusted manually with the software BioEdit (available from http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Sequences that could not be unambiguously aligned or were of insufficient length were excluded from the study. Phylogenetic relationships between the *pol* sequences were reconstructed using the neighbour-joining followed by maximum likelihood methods. An initial neighbour-joining tree was built under the Hasegawa–Kishino–Yang (HKY85) model of evolution with a ratio of transversion to transitions of 2:1 using the tree-building software Paup* (available from http://paup.csit.fsu.edu/about.html).

The best fitting model of nucleotide substitution was estimated on the basis of the neighbour-joining tree topology using a maximum likelihood ratio test with Modeltest version 3.0 (available from http://bioag.byu.edu/zoology/crandall_lab/modeltest.htm). The derived parameters of the selected model were then used to perform a heuristic search for a maximum likelihood tree with Paup*. The construction of the tree was done according to the general time reversible (GTR) model of evolution, with a proportion of invariable sites and gamma distribution. An HIV-1 subtype K sequence (Genbank accession number AJ249239) retrieved from the Los Alamos HIV database (http://hiv-web.lanl.gov/) was used as an outgroup and six pairs of follow-up sequences from the same individuals were used as controls. The robustness of the neighbour-joining trees was evaluated by bootstrap analysis, with 1000 rounds of replication.

Statistical analysis

Statistical comparisons of those in a cluster with those not in a cluster were performed using Chi-squared tests, Fisher's exact tests or Mann–Whitney U tests, as appropriate. Multivariable logistic regression was used to identify factors independently associated with belonging to a cluster. All statistical analyses were performed using SAS version 8 (available from http://v8doc.sas.com/sashtml/). The study was approved by the Brighton and Hove Local Research Ethics Committee and the Health Protection Agency Ethics Committee. Confidentiality and anonymity were protected by irreversibly unlinking clinic and laboratory from the study ID number using a firewall system managed by the local public health laboratory. Written, informed consent was obtained from all participants.

Results

Study population description

A total of 103 individuals with PHI diagnosed between 1999 and 2003 were included in the epidemiological and phylogenetic analysis. Of these, 73 (71%) had a STARHS antibody test suggestive of infection within the previous 4–6 months. Almost all (99, 96.1%) were men and 90 (90.9%) were MSM. All the men and two out of four women were Caucasian with a median age of 36 years (range 21–67). The median age was 36 years (range 21–67). Six individuals (6.1%) reported a history of injecting drug use (two MSM, two heterosexual men and two heterosexual women). The median CD4 cell count

Table 1. Comparison of features associated with patients in the cluster and non-cluster groups.

	In cluster	Not in cluster	*P* value[a]
Number of patients	35	68	
Male sex	35 (100%)	64 (94.1%)	0.30
Age (years): median (range)	34 (23–54)	37 (21–67)	0.05
Number of contacts in 3 months before diagnosis: median (range)	3 (1–100)	2 (1–36)	0.006
Homosexual risk group	32 (97.0%)	58 (85.3%)	0.10
Highest reported risk in the 3 months before diagnosis of PHI			
Unprotected oral intercourse	25 (78.1%)	36 (73.5%)	0.83
Protected anal intercourse	2 (6.3%)	5 (10.2%)	0.70
Unprotected anal intercourse	28 (87.5%)	32 (65.3%)	0.05
Unprotected vaginal intercourse	0 (–)	8 (16.3%)	0.02
STI in 3 months before diagnosis			0.31
Yes	15 (42.9%)	19 (27.9%)	
No	18 (51.4%)	37 (54.4%)	
Not known	2 (5.7%)	12 (17.7%)	0.13
CD4 cell count (cells/mm^3): median (range)	612 (195–1477)	474 (196–1259)	0.005
CD4 cell percentage: median (range)	31 (12–40)	26.5 (7–42)	0.003
Viral load ($\log_{10}$ copies/ml): median (range)	4.97 (2.03–6.00)	4.94 (2.30–6.00)	0.90

[a]Entries in table are *n* (%) unless otherwise specified.

(available in 101/103) was 526 copies/ml (range 195–1477) and the median CD4 cell percentage (available in 81/103) was 28 (7–42). The median HIV viral plasma load was log 4.95 copies/ml (2.03–6.00). Thirteen MSM (12.6% of total patients) were infected with viruses that contained primary antiretroviral resistance-associated mutations. STI were diagnosed concurrently with PHI in 34 of the 89 individuals (38.2%) for whom information was available. Among the 90 MSM, 61 (68%) reported unprotected anal intercourse (UAI) in the 3 months before PHI diagnosis; no information was available regarding sexual practices in the period preceding this.

Cluster comparison

Viruses from 35 out of 103 individuals (34%) appeared within 15 transmission clusters, comprising one cluster of five individuals, two of three and 12 of two (full results shown in Table 1 and Fig. 1). All were men and 32 (97%) were MSM. For individuals within 11 out of 15 clusters, the diagnosis of PHI was made within 12 months of each other, giving supporting evidence that transmission occurred during the PHI period. Those in the cluster group had a higher CD4 cell count ($P = 0.005$), higher CD4 cell percentage ($P = 0.003$), were younger ($P = 0.05$), reported a higher number of different sexual contacts in the previous 3 months ($P = 0.006$), and were more likely to have engaged in UAI in the 3 months before the PHI diagnosis ($P = 0.05$) in comparison to those individuals not within a cluster. High rates of STI at the time of PHI were observed in both groups, with a trend towards higher rates in those individuals with viruses in a cluster (42.9 versus 27.9%, $P = 0.13$). Multivariable logistic regression analyses identified the CD4 cell percentage [odds ratio (OR) 1.14, 95% confidence interval (CI) 1.04–1.23, $P = 0.003$] and having more than five sexual partners (OR 3.38, 95% CI 1.13–10.10, $P = 0.03$) as the only independent predictors of belonging to a cluster. Six individuals (17%) had antiretroviral-associated resistance mutations, of whom two (both T215D in reverse transcriptase) belonged to a linkage pair.

Conclusion

In conclusion, the high rates of clustering observed within our study support the assertion that PHI may be associated with an increased risk of onward transmission. The associations we found with younger age, high rates of UAI, and sexual partner change identify this as a high-risk group for HIV transmission. There was a trend towards higher rates of STI in the cluster group on a background of extremely high STI rates in the study population, supporting the argument for increased STI surveillance, particularly of high-risk groups.

The highly significant correlation with CD4 cell counts may represent the early disease stage, or rapid contact tracing and testing of sexual partners of individuals diagnosed with PHI. The plasma viral load at diagnosis was not predictive of clustering, and it is possible that the seminal viral load in men is a more consistent correlate of infectiousness, particularly in the context of genital tract inflammation, with plasma/genital tract discordance playing an important role [7–10]. The presence of the same antiretroviral resistance mutation in one cluster pair, neither of whom had received antiretroviral therapy, illustrates the potential for the secondary spread of such resistant strains, as we have previously documented [33,34]. Our results do not exclude the possibility of a

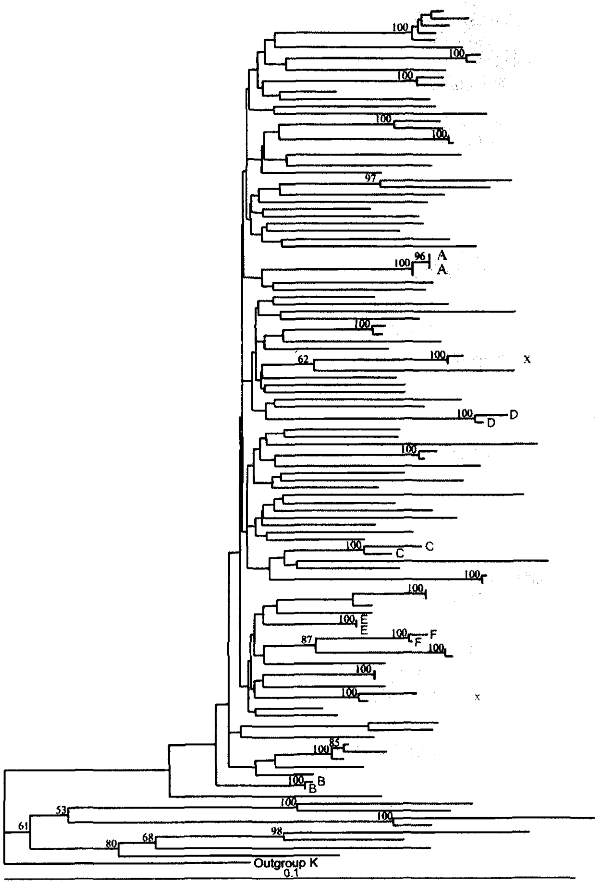


Fig. 1. Maximum likelihood phylogenetic tree based on *pol* sequences from 103 individuals with primary HIV-1 infection. Possible transmission clusters are circled. Linkages confirmed by clinical data are indicated by a red cross. Transmission clusters were identified if the bootstrap value was equal or greater than 99% and the average genetic distance (i.e. branch length) was lower than 0.015 nucleotide substitutions per site. Linkages confirmed by clinical data are indicated by a red cross. Six pairs of multiple sequences from single patients were used as controls for relatedness and are indicated by letters (e.g. A indicates multiples sequences from patient A). Bootstrap values higher than 50% are indicated on the branches.

common source for each cluster, rather than transmission within clusters. However, a phylogenetic tree comprising viruses from these 103 primary infections, together with more than 2000 *pol* sequences from prevalent infections throughout the UK only identified one further potential linkage, and that involved a primary infection case not within an existing cluster (data not shown).

Only 31 of the non-cluster group (64.6%) reported UAI, but it should be noted that this is only in the time window 3 months before diagnosis with PHI. Interestingly, routinely collected data on recent sexual contacts only confirmed three of the linkage pairs that were revealed in the phylogenetic analysis, emphasizing the high rates of anonymous sexual partners and the difficulty in obtaining a reliable sexual history.

Our results provide further evidence that the active management of primary infection will reduce HIV transmission. HIV prevention programmes have been heavily focused on protecting susceptible individuals, but accumulating biological and modelling data suggest that reducing the infectiousness of HIV-positive individuals may also be an effective strategy. A large proportion of PHI remains undiagnosed in the community [35,36], and these findings support the view that as a disease stage PHI represents a major public health threat. Efforts should be re-focused on improving rates of diagnosis of individuals during PHI, timely contact tracing, risk reduction, the management of STI, and possibly early treatment with antiretroviral agents in an effort to break transmission networks during this unique and possibly crucial stage of HIV infection [37]. Furthermore, consideration should be given in information and awareness campaigns to highlight the possible symptoms of PHI in groups with high rates of onward transmission, to encourage such individuals to present to appropriate healthcare providers to enable the timely diagnosis and management of early infection.

Contributors

M.F. and D. Pillay devised the study. D. Pao, M.F. and G.D. recruited patients for the study. D. Pao, M.F., S.H., C.S. and D. Pillay wrote the manuscript. P.A.C. undertook sequencing and curated the sequences. S.H. undertook the phylogenetic analyses. G.M. undertook the STARHS analysis. C.S. undertook statistical analyses.

Acknowledgements

The authors would like to thank the Health Protection Agency for funding, and the patients for agreeing to enter this study.

References

1. Joint United Nations Programme on HIV/AIDS, http://www.unaids.org, copyright 2004. Accessed 30 October 2004.
2. Health Protection Agency, http://www.hpa.org.uk, established 2002. Accessed 30 October 2004.
3. Murphy G, Charlett A, Jordan LF, Osner N, Gill ON, Parry JV. **HIV incidence appears constant in men who have sex with men despite widespread use of effective antiretroviral therapy.** *AIDS* 2004, **18**:265–272.
4. Kaufmann GR, Cunningham P, Kelleher AD, Zaunders J, Carr A, Vizzard J, *et al.* **Patterns of viral dynamics during primary human immunodeficiency virus type 1 infection.** *J Infect Dis* 1998, **178**:1812–1815.
5. Lindback S, Karlsson AC, Mittler J, Blaxhult A, Carlsson M, Briheim G, *et al.* **Viral dynamics in primary HIV-1 infection.** *AIDS* 2000, **14**:2283–2291.
6. Little SJ, McLean AR, Spina CA, Richman DD, Havlir DV. **Viral dynamics of acute HIV-1 infection.** *J Exp Med* 1999, **190**:841–850.
7. Coombs RW, Speck CE, Hughes JP, Lee W, Sampoleo R, Ross SO, *et al.* **Association between culturable human immunodeficiency virus type 1 (HIV-1) in semen and HIV-1 RNA levels in semen and blood: evidence for compartmentalization of HIV-1 between semen and blood.** *J Infect Dis* 1998, 177:320–330.
8. Leynaert B, Downs AM, de Vincenzi I. **Heterosexual transmission of human immunodeficiency virus: variability of infectivity throughout the course of infection. European Study Goup on Heterosexual Transmission of HIV.** *Am J Epidemiol* 1998, **148**:88–96.
9. Pilcher CD, Tien HC, Eron JJ Jr, Vernazza PL, Leu SY, Stewart PW, *et al.* **Brief but efficient: acute HIV infection and the sexual transmission of HIV.** *J Infect Dis* 2004, **189**:1785–1792.
10. Pilcher CD, Shugars DC, Fiscus SA, Miller WC, Menezes P, Giner J, *et al.* **HIV in body fluids during primary HIV infection: implications for pathogenesis, treatment and public health.** *AIDS* 2001, **15**:837–845.
11. Gray RH, Wawer MJ, Brookmeyer J, Sewankambo NK, Serwadda D, Wabwire-Mangen F, *et al.* **Probablility of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai.** *Uganda Lancet* 2001, **357**:1149–1153.
12. Galvin SR, Cohen MS. **The role of sexually transmitted diseases in HIV transmission.** *Nat Rev Microbiol* 2004, **2**:34–42.
13. Vernazza PL, Eron JJ, Fiscus SA, Cohen MS. **Sexual transmission of HIV: infectiousness and prevention.** *AIDS* 1999, **13**:155–166.
14. Zuckerman RA, Whittington WL, Celum CL, Collis TK, Lucchetti AJ, Sanchez JL, *et al.* **Higher concentration of HIV RNA in rectal mucosa secretions than in blood and seminal plasma, among men who have sex with men, independent of antiretroviral therapy.** *J Infect Dis* 2004, **190**:156–161.
15. Chakraborty H, Sen PK, Helms RW, Vernazza PL, Fiscus SA, Eron JJ, *et al.* **Viral burden in genital secretions determines male-to-female sexual transmission of HIV: a probabilistic empiric model.** *AIDS* 2001, **15**:621–627.
16. Koopman JS, Jacquez JA, Welch GW, Simon CP, Foxman B, Pollock SM, *et al.* **The role of early HIV infection in the spread of HIV through populations.** *J Acquir Immune Defic Syndr* 1997, **14**:249–258.
17. Jacquez JA, Koopman JS, Simon CP, Longini IM Jr. **Role of the primary infection in epidemics of HIV infection in gay cohorts.** *J Acquir Immune Defic Syndr* 1994, 7:1169–1184.
18. Xiridou M, Geskus R, De Wit J, Coutinho R, Kretzschmar M. **The contribution of steady and casual partnerships to the incidence of HIV infection among homosexual men in Amsterdam.** *AIDS* 2003, **17**:1029–1038.
19. Pilcher CD, Eron JJ Jr, Vemazza PL, Battegay M, Harr T, Yerly S, *et al.* **Sexual transmission during the incubation period of primary HIV infection.** *JAMA* 2001, **286**:1713–1714.
20. Kahn JO, Walker BD. **Acute human immunodeficiency virus type 1 infection.** *N Engl J Med* 1998, **339**:33–39.
21. Dodds JP, Nardone A, Mercey DE, Johnson AM. **Increase in high-risk sexual behaviour among homosexual men, London 1996–1998: cross sectional, questionnaire study.** *BMJ* 2000, **320**:1510–1511.
22. Colfax G, Buchbinder SP, Cornelisse PGA, Vittinghoff E, Mayer K, Celum C. **Sexual risk behaviours and implications for**

secondary HIV transmission during and after HIV seroconversion. *AIDS* 2002, **16**:1529–1535.

23. Colfax G, Mansergh G, Vittinghoff E, Guzman R, Marks G, Buchbinder S. **Drug use and high-risk sexual behaviour among circuit party participants.** In: *XIIIth International Conference on AIDS*. Durban, 2000 [Abstract TuPeC3422].
24. Katz MH, Schwarcz SK, Kellogg TA, Klausner JD, Dilley JW, Gibson S, *et al.* **Impact of highly active antiretroviral treatment on HIV seroincidence among men who have sex with men: San Francisco.** *Am J Public Health* 2002, **92**:388–394.
25. Clements MS, Prestage G, Grulich A, Van De Ven P, Kippax S, Law MG. **Modeling trends in HIV incidence among homosexual men in Australia 1995–2006.** *J Acquir Immune Defic Syndr* 2004, **35**:401–406.
26. Little SJ, Holte S, Routy JP, Daar ES, Markowitz M, Collier AC, *et al.* **Antiretroviral-drug resistance among patients recently infected with HIV.** *N Engl J Med* 2002, **347**:385–394.
27. Duwe S, Brunn M, Altmann D, Hamouda O, Schmidt B, Walter H, *et al.* **Frequency of genotypic and phenotypic drug-resistant HIV-1 among therapy-naive patients of the German seroconverter study.** *J Acquir Immune Defic Syndr* 2001, **26**:266–273.
28. Grant RM, Hecht FM, Warmerdam M, Liu L, Liegler T, Petropoulos CJ, *et al.* **Time trends in primary HIV-1 drug resistance among recently infected persons.** *JAMA* 2002, **288**:181–188.
29. Pillay D, Cane PA, Shirley J, Porter K. **Detection of drug resistance associated mutations in HIV primary infection within the UK.** *AIDS* 2000, **14**:906–908.
30. Janssen RS, Satten GA, Stramer SL, Rawal BD, O'Brien TR, Weiblen BJ, *et al.* **New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes.** *JAMA* 1998, **280**:42–48. Erratum in *JAMA* 1999; **281**: 1893.
31. Kothe D, Byers RH, Caudill SP, Satten GA, Janssen RS, Hannon WH, *et al.* **Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence.** *J Acquir Immune Defic Syndr* 2003, **33**:625–634.
32. Hue S, Clewley J, Cane P, Pillay D. **HIV-1 *pol* gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy.** *AIDS* 2004, **18**:719–728.
33. Yerly S, Vora S, Rizzardi P, Chave JP, Vernazza PL, Flepp M, *et al.* **Acute HIV infection: impact on the spread of HIV and transmission of drug resistance.** *AIDS* 2001, **15**:2287–2292.
34. Taylor S, Cane P, Hue S, Xu L, Wrin T, Lie Y, *et al.* **Identification of a transmission chain of HIV type 1 containing drug resistance-associated mutations.** *AIDS Res Hum Retroviruses* 2003, **19**:353–361.
35. Melzer M, Brown M, Mullen J, O'Shea S, Chrystie I, Banatvala J. **Undiagnosed symptomatic primary HIV infections in South London [Letter].** *J Infect* 2001, **42**:297–298.
36. Pilcher CD, McPherson JT, Leone PA, Smurzynski M, Owen-O'Dowd J, Peace-Brewer AL, *et al.* **Real-time, universal screening for acute HIV infection in a routine HIV counseling and testing population.** *JAMA* 2002, **288**:216–221.
37. Cates W Jr, Chesney MA, Cohen MS. **Primary HIV infection – a public health opportunity.** *Am J Public Health* 1997, **87**:1928–1930.