INVITED REVIEW

Deciphering host migrations and origins by means of their microbes

THIERRY WIRTH,* AXEL MEYER* and MARK ACHTMAN†

Department of Biology, Lehrstuhl für Zoologie und Evolutionsbiologie, University Konstanz, 78457 Konstanz, Germany, †Max-Planck Institut für Infektionsbiologie, Department of Molecular Biology, Schumannstraße 21/22, 10117 Berlin, Germany

Abstract

Mitochondrial DNA and microsatellite sequences are powerful genetic markers for inferring the genealogy and the population genetic structure of animals but they have only limited resolution for organisms that display low genetic variability due to recent strong bottlenecks. An alternative source of data for deciphering migrations and origins in genetically uniform hosts can be provided by some of their microbes, if their evolutionary history correlates closely with that of the host. In this review, we first discuss how a variety of viruses, and the bacterium *Helicobacter pylori*, can be used as genetic tracers for one of the most intensively studied species, *Homo sapiens*. Then, we review statistical problems and limitations that affect the calculation of particular population genetic parameters for these microbes, such as mutation rates, with particular emphasis on the effects of recombination, selection and mode of transmission. Finally, we extend the discussion to other host–parasite systems and advocate the adoption of an integrative approach to both sampling and analysis.

Introduction

Indirect evidence can be a powerful deductive tool. Consider one of the most famous French dessert wines, the Château d'Yquem from Sauternes, which is made from grapes affected by *Botrytis cinerea*. Under certain climatic conditions this microscopic fungus, gives rise to the 'noble rot'. However, some vintages of Château d'Yquem do not exist, e.g. from the year 1964. Knowing the owners of Château d'Yquem are such perfectionists that they only produce wine under optimal conditions, an amateur wine enthusiast could deduce when unusual climatic conditions occurred in a small area close to the Garonne River in southwestern France. Our wine amateur would use the indirect evidence of a missing vintage to infer that rainfall, temperature and humidity were unusual during the summer of 1964, which was indeed extremely cold and humid.

Evolutionary biologists and population geneticists are sometimes confronted by analogous situations. Imagine an organism that has undergone an extreme bottleneck followed by rapid range expansion and now shows low genetic diversity and weak population genetic structure. The evolutionary relationships of such an organism will be particularly difficult to decipher, if not impossible. However, it may be possible to deduce important aspects of its evolutionary history through the population parameters of parasites or microbes that infect it because they have co-evolved with their host. An eminent illustration of how useful such 'indirect' markers can be is provided by mitochondrial DNA (mtDNA). The mtDNA is extensively used for the phylogenetic analyses of a large spectrum of organisms, although the endosymbiotic origin of mitochondria is generally accepted. Mitochondria probably descended from specialized purple nonsulphur bacteria that became incorporated into the cytoplasm after endocytosis by a eukaryote, a scenario called the endosymbiotic hypothesis (Gray *et al.* 1999). This event happened so long ago, 1 billion years ago, that all eukaryotes

Correspondence: Thierry Wirth, Fax: +49 7531 88 3018, E-mail: thierry.wirth@uni-konstanz.de

possess mitochondria. Furthermore, mtDNA are characterized by mutation rates that are an order of magnitude faster than nuclear genes, recombine rarely and are inherited maternally, features that have promoted the usage of mtDNA to a standard for evolutionary analyses (Cann *et al*. 1987; Cavalli-Sforza 1998).

The population structure of other microbes also reflects the evolutionary histories of their hosts. Two dramatic examples are *Buchnera aphidicola*, which are found in specialized cells (bacteriocytes) within the body cavity of aphids (Funk *et al*. 2000; Moran 2003), and *Wigglesworthia glossinidia* in tsetse flies (Canback *et al*. 2004). Funk *et al*. (2000) showed that the phylogenies of these endosymbiotic Proteobacteria are strictly congruent with those of their hosts. The congruence between the phylogenies relies on over 200 million years (Myr) of co-evolution of microbe with their host, maternal transmission, and obligate ancient mutualism. An increasing number of bacterial endosymbionts have been described that might be exemplary candidates to study host–parasite co-evolution (Wernegreen 2002). In addition to bacteria, the genetic structure of rust fungi is tightly coupled to that of their plant hosts because specific virulence genes in the fungus often match resistance genes in the plant and these fungi have also been used as 'genetic markers' by botanists (Thrall & Burdon 2003).

However, the efficacy of genetic markers from parasites depends on multiple factors. The efficacy depends on the transmission mode of the parasite (vertical vs. horizontal), on the strength of the host–parasite relationship (endosymbiosis, commensalism, or loose unspecific host parasitism) and also on the spread (epidemic vs. endemic) (Holmes 2004). Moreover, many viral and bacterial genomes include genes imported from unrelated organisms by lateral gene transfer (LGT) that can obscure the patterns associated with vertical transmission (Lerat *et al*. 2005). Thus, only some parasites might be expected to yield information about the phylogeny of their host. In this review, we use these considerations to examine how well a selected set of parasites of our own species, *Homo sapiens*, can be expected to provide new insights on its population genetics.

Human origins and migrations: a framework for testing whether microbes can help to decipher the history of their hosts

The origin and spread of modern *Homo sapiens* is one of the most fascinating and hotly debated topics in the field of evolutionary biology. The African replacement model suggests that modern humans left Africa *c*. 150 000 years ago (Cann *et al*. 1987). This major migration was accompanied by a drastic reduction of genetic diversity within modern humans from Asia and Europe. Consequently, the genetic relationships between humans from different geographical sources are difficult to resolve due to extremely low genetic diversity. To overcome this drawback, the use of protein polymorphisms was rapidly replaced by the quickly evolving mitochondrial D-loop in the 1970s (Cavalli-Sforza & Feldman 2003). Microsatellite markers (Goldstein *et al*. 1995; Jorde *et al*. 1997; Ruiz-Linares *et al*. 1999) later became popular as population genetic markers, followed by Y-chromosome (Poloni *et al*. 1997; Semino *et al*. 2000; Su *et al*. 2000; Underhill *et al*. 2000; Ke *et al*. 2001) and autosomal single nucleotide polymorphisms (Collins *et al*. 1997; Wang *et al*. 1998; Zhao *et al*. 2000; Reich *et al*. 2001).

A combination of molecular, archaeological, physical anthropological, and linguistic data has resulted in a convincing concept for human history and evolution (Cavalli-Sforza 1998), according to which African emigrants invaded the Near East, southern Asia and Europe in successive waves. Subsequently, the Americas are thought to have been peopled through multiple migrations of different Asian populations across the Bering Strait (Eshleman *et al*. 2003).

The integration of multiple disciplines has resulted in a near-consensus view regarding human migrations. However, many issues remain unresolved, such as the source of the Polynesians (Murray-McIntosh *et al*. 1998) and the exact number of waves to the Americas (Gibbons 1996; Merriwether & Ferrell 1996; Merriwether *et al*. 1996). Even the number of migratory waves from Africa to Europe and Asia is still being discussed (Templeton 2002; Stringer 2003) and the detailed history of Central Asia remains a topic of continuing investigations (Yao *et al*. 2002a, b). Given that most analyses are based on weak genetic signals, corroborating evidence from other sources, such as human parasites, would be tremendously valuable.

A review of microbial candidates for tracing human migrations

Multiple parasites have accompanied humans during their ancient and recent migrations. These include DNA viruses, such as JCV (JC virus) that infect renal tissue (Agostini *et al*. 1997) and HPV (human papillomavirus) (Chan *et al*. 1992), and RNA viruses, such as HTLV (the human T-cell lymphotropic virus) (Miura *et al*. 1994), GBV-C (GB virus C)/HGV (hepatitis G virus) (Pavesi 2003), and HIV. The most promising eubacterial candidate is *Helicobacter pylori*, a bacterium that colonizes the human stomach (Falush *et al*. 2003b; Wirth *et al*. 2004). Moreover, one fungus seems to mirror the New World pattern of human expansion and migration (Fisher *et al*. 2001). For each of these genetic markers, we will briefly describe the main conclusions obtained from their population structure and will also highlight discrepancies and limits of those population structures with respect to understanding human evolutionary history.

The candidates

Human T-cell lymphotropic virus

Human T-cell lymphotropic virus I (HTLV-I), a member of the Retroviridae family, was first isolated in the early 1980s from cutaneous T-cell lymphomas and leukaemia (Poiesz *et al*. 1980, 1981) and from patients with adult T-cell leukaemia/lymphoma (Miyoshi *et al*. 1981) Later, a new subtype, HTLV-II, was described (Kalyanaraman *et al*. 1982). HTLV-I and HTLV-II display different tropisms, namely for CD4+ and CD8+ T cells. Transmission is either vertical or horizontal: HTLV is transmitted from mother to child via breast milk (Ishak *et al*. 2001; Kashiwagi *et al*. 2004), as well as between adults via drug injection and sexual intercourse. The prevalence of HTLV-I is rather low (3% to 6%), whereas HTLV-II is endemic among intravenous drug users (IDU) and can reach a prevalence of 40%. In a seminal paper, Miura and colleagues analysed a limited number of HTLV-I isolates from native inhabitants of India and South America, as well as from the Ainu, an ancient Japanese ethnic group (Miura *et al*. 1994). Using phylogenetic reconstructions, they showed that the 'cosmopolitan type' is composed of three major lineages (subtypes A, B and C). Interestingly, subtype A was found in Caribbean, South American and Ainu samples, suggesting a migration wave to North and South America via the Bering Strait in the Palaeolithic era. In order to test this hypothesis, ancient HTLV-I DNA was extracted from a 1500-year-old Andean mummy, and as expected, the DNA sequences from the mummy virus were similar to those obtained in contemporary Andeans and Japanese (Li *et al*. 1999; Sonoda *et al*. 2000). This finding confirmed that HTLV-I subtype A accompanied ancient Mongoloids to the American continent prior to the colonial era but has been criticized (Gessain *et al.* 2000; Vandamme *et al.* 2000). Another subtype, subtype C, was isolated from West Africans and Caribbeans (Miura *et al*. 1994), indicating gene flow from West Africa towards America during the slave trade.

HTLV-I also suffers from many drawbacks for elucidating human migrations. One drawback is that the mode of transmission is largely horizontal. Even transmission by breastfeeding is not exclusively vertical within families because babies are nursed by multiple lactating women in South American ethnic groups. Horizontal transmission can dilute signatures of ethnic groups and in areas with multiple ethnicities can reflect recent intercontinental gene flow. Second, unlike human population genetics, the most basal clade in an HTLV-I phylogeny is not found in Africa but instead among isolates from Melanesians in New Guinea, inhabitants of the Solomon Islands and Australian Aborigines (Miura *et al*. 1994). Moreover, the highest genetic diversity was observed in Asia, suggesting that HTLV-I emerged first in this area (Slattery *et al*. 1999).

Even more surprising is the observation that simian T-lymphotrophic type I viruses are more closely related to other human HTLV-I genotypes than they are to the Melanesian HTLV-I genotypes. Such a pattern is most simply explained through multiple viral reservoirs in higher primates from which transmission to humans occurred on independent occasions. This explanation was subsequently substantiated when it was shown that subtype D shares evolutionary history with isolates from wild-born mandrills and baboons (Slattery *et al.* 1999), whereas HTLV-I subtype E is related to isolates from chacma baboons. Thus, HTLV-I population structure seems to reflect multiple, independent transmissions from different ape species to local *H. sapiens* populations, which provides problems for the use of sequence diversity within HTLV to reconstruct prehistoric human migrations.

The population genetic structure of HTLV-II differs from that of HTLV-I. Type II viruses are only isolated from two species, *H. sapiens* and *Pan troglodytes*, and phylogenetic analyses support an African origin of HTLV-II, based on the very deep branching of an HTLV-2d virus isolated from an Efe Bambuti Pygmy (Vandamme *et al.* 1998). However, HTLV-II viruses were initially thought to be native to Amerindian and human Pygmy tribes, which is hard to explain under the traditional model of human migrations. The absence of concordance between phylogenetic signals and geography presents a complex puzzle for evolutionary anthropologists. Most subtypes can be isolated from diverse ethnic groups: subtypes A and B are isolated in Europe, America and Africa while subtype C, whose existence is debated, resembles a variant of subtype A that is restricted to Brazil (Slattery *et al.* 1999). Moreover, *env* and LTR possess different phylogenetic signals from *tax* (Slattery *et al.* 1999). Possibly these discrepancies could be explained by the very recent origin of HTLV-II followed by successive waves of dissemination, or by the retention of ancestral polymorphisms during incomplete lineage sorting.

Still another drawback derives from the nature of the virus. Like other RNA viruses, HTLV viruses display high mutation rates. Based on changes since the introduction of HTLV-II into intravenous drug users in Europe 25 years ago, the nucleotide substitution rate within the long terminal repeat (LTR) was estimated as 1×10^{-4} to 3×10^{-5} (Salemi *et al.* 1998). This is only two orders of magnitude lower than that of HIV, the most rapidly evolving virus discovered so far. The mdecular clock of HTLV-I is lower (Van Dooren *et al.* 2004). Therefore, the LTR sequences that have been so popular for phylogenetic reconstruction may well be saturated by polymorphisms and contain multiple homoplasies, which would reduce the information content. Another partial sequence often used in HTLV phylogenetics is the envelope gene (*env*), favoured because of its high degree of genetic diversity. However, envelope proteins are exposed to strong diversifying selection pressure due to the host

immune system and might be frequently imported. As a result, the evolution of the HTLV-I *env* gene plus the LTR locus does not proceed according to a metric clock (Van Dooren *et al.* 2001) although LTR and third codon positions of the *env* gene did exhibit clocklike behaviour. This observation suggests that the first and second codon positions of the *env* gene are under positive selective pressure.

Although being closely related and despite broad genomic similarities, the evolutionary histories of HTLV-I and HTLV-II are clearly divergent. This is unexpected for such closely related lineages and seems to question the overall utility of HTLV for reconstructing human movement and history.

The GB virus/hepatitis G virus

GBV-C and HGV are alternate designations for a member of the Flaviviridae that initially showed potential as an ethnic marker for tracing prehistoric links between human populations (Pavesi 2001). Its transmission pattern resembles that of HTLV and infection by HGV is widely distributed, with a worldwide prevalence among blood donors of about 2% (Domingo *et al.* 1996). Phylogenetic approaches suggested that this virus has an African ancestry and that strains from Africa and Southeast Asia are closely linked, indicating a major route of ancient human migrations out of Africa (Pavesi 2001). Mutation rates are close to 10^{-5}, unusually low for an RNA virus, possibly because functional constraints generated by RNA secondary structures have subjected the genome to strong selective pressure for silent substitutions (Simmonds & Smith 1999). Different regions of the genome were analysed for functional constraints. Trees based on four slowly evolving regions and the entire coding regions of 8457 bp were compared. Unfortunately, the topologies of the trees differ with the region and are only weakly supported with low bootstrap values, possibly reflecting ongoing LGT. The branching patterns and tree topologies were not tested using sophisticated analyses; thus the interpretation of these data remains unclear and the global interpretation in terms of human migrations is uncertain.

Human immunodeficiency virus

The human immunodeficiency virus (HIV) is possibly the most comprehensively studied retrovirus, but its use in terms of human movements is largely anecdotal and HIV can probably only target events in the last 70 years. Vertical transmission does occur (Cavaco-Silva *et al.* 1997), but is rare in comparison to horizontal transmission. Horizontal transmission coupled with recombination results in mosaic sequences in some HIVs, which contain sequence signatures from subtype A, that is prevalent throughout Africa, as well as from subtype G, which is common in West and Central Africa (Cornelissen *et al.* 2000). Such recombinants indicate HIV has spread between African countries that are thousands of kilometers apart due to human movements.

Recombination poses a problem for phylogenetic analyses in HIV as well as other viruses. Nearly all viruses that have been considered as potential tracers for human migration have been investigated via phylogenetic reconstruction. However, recombination can lead to inaccurate phylogenetic reconstructions and can affect estimates of divergence times by increasing apparent nucleotide substitution rates and decreasing distance measures (Schierup & Hein 2000; Worobey 2001). In the case of HIV, the number of recombination events per genome per replication exceeds the mutation rate per replication (Rambaut *et al.* 2004). Hence, it seems prudent to screen sequences systematically for recombinants prior to phylogenetic analyses, particularly for RNA viruses.

Potentially better viral candidates

The following two DNA viruses may provide excellent indirect evidence for tracing human migrations and should receive more attention in this context.

Human papillomavirus

More than 60 types of HPV have been found so far, of which, two common lineages, type 16 (HPV-16) and type 18 (HPV-18), seem to mirror past human migrations very closely. HPV-16 is notorious due to its association with neoplasias of the genital epithelia and its role in cervical cancer (zur Hausen 1989). HPV-16 is both widespread and is the most predominant type that is isolated (Chan *et al.* 1992). Human papillomaviruses are probably not affected by the human immune response because they are not epidemic and can be maintained for a long time within small populations through dormancy and long latency periods. In a global sampling scheme, Chan *et al.* (1992) tried to decipher the viral phylogeny and its host–parasite relationship using a small genomic segment of the long control region from various biopsies. Among 118 human papillomaviruses from Asia, South America, Africa and Europe, the maximum evolutionary distances between HPV-16 viruses were rather modest and did not exceed 5%. The most striking feature of the resulting UPGMA phylogeny was a split of the genotypes into two discreet groupings that were specific to Africa + Brazil samples vs. Indo-Europeans + Asia, respectively. In a more sophisticated minimal-spanning tree approach, it was possible to discern three lineages rather than two, namely lineages that were specific to Africa, Brazil and Eurasia. Consistent results were obtained when the authors analysed a different DNA fragment (open reading frame

(a)

(b)
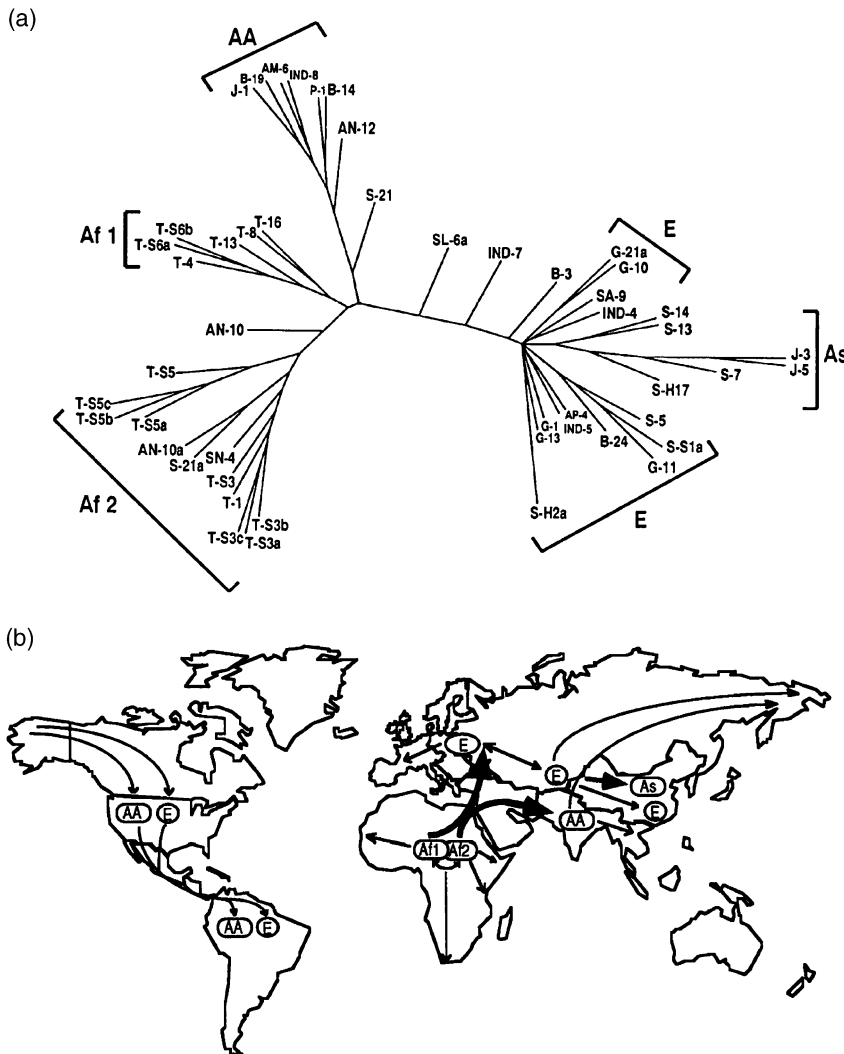
Fig. 1 (a) Neighbour-joining tree of 48 genomic variants of HPV-16 and 25 ethnic groups. Boldface letters indicate the five major lineages of HPV-16, i.e. the two African branches, Af1 and Af2; the Asian/American branch, AA; the European branch, E; and the East-Asian branch, As. The split between the African, Eurasian and AA branches are supported by strong bootstrap values (> 82%). (b) Evolution and spread of HPV-16 variants in prehistoric times. Thin arrows represent spread without or with minor mutational changes; bold arrows represent mutational changes that gave rise to the five major HPV-16 lineages (from Ho *et al*. 1993).

E5), unlike many RNA viruses. The correlations of the partitions with geography were imperfect, possibly reflecting recent colonial history, e.g. Eurasian HPV-16 variants were isolated in Tanzania and European and African variants were found in Brazil. Assuming a constant mutation rate in the range of 10^{-9} to 10^{-7}, the authors estimated that the ancestral sequence present at the root of the tree was between 240 000 and 24 000 000 years old, a time span that encompasses primate and human evolution. Subsequently, the same group produced even more precise results based on 301 virus isolates collected from 25 different human ethnic groups (Ho *et al*. 1993). The larger sample resulted in a total of five major HPV-16 lineages (Fig. 1a): two predominantly present in Africa (Af1 and Af2), two in Asia (AA and As) and one mainly found in Indo-Europe (E). Only 10% of the isolates deviated from these generalizations, which could be easily explained in light of recent human migrations (Fig. 1b). The utility of HPV-16 for tracing human migrations is

illustrated in many ways. For example, two of four isolates from patients in Cape Town were of the E genotype, typical of Europeans, as were three of five patients of mixed racial background. Similarly, nearly all isolates from hosts of Caucasian descent in northern and southern India belonged to the E lineage (Sokal *et al*. 1992). Unfortunately, no appropriate outgroup has been identified and those HPV types that are most similar to HPV-16 (HPV-31 and HPV-35) are so diverse that they could not be aligned with confidence. Therefore, the trees remained unrooted, which precludes the determination of the most ancestral lineage.

The HPV-16 results are supported by analyses of HPV-18. Again, three main lineages were identified that correspond to the three principal human lineages of Africans, Caucasians and East Asians (Ong *et al*. 1993). It was possible to root that tree due to the existence of an acceptable outgroup and the oldest HPV-18 lineage had its roots in Africa. It was also possible to calibrate the mutation rate better with HPV-18 than with HPV-16 by using the close

genetic relationship observed between isolates from Amazonian Indians and Japanese patients: the calculations indicate that one point mutation is fixed every 12 000 years. Consequently, the maximum divergence time between the two basal HPV-18 lineages is estimated to be about 200 000 years, an estimate that fits well with the origin of modern *H. sapiens*.

The ancient human migrations inferred from HPV-18 data and correspond closely to what is known from human data. However, one feature of the HPV-18 data contradicts human genetic data, or is at least somewhat puzzling. The As lineage, isolated primarily in Japan (*n* = 6) and Singapore (*n* = 19), is closely related to the European E cluster (Fig. 1a), although it received a separate designation due to the geographical differences. Based on human phylogenetic reconstruction, one would expect that East Asians should cluster close to the AA lineage, rather than forming a subgroup of the E lineage. A closer look at isolates from Native Americans also revealed unexpected results. Although an association between isolates from Greenland Eskimos and lineage E was not surprising due to the proximity between Greenland and Europe, the same link was observed in many isolates from Navajo, Pueblo and Amazon Indians! These latter ethnic groups originated from North Asiatic emigrants via the Bering Strait and should therefore be associated with the AA lineage. Even more surprising was the isolation of the Af1 lineage from one Inuit and the Af2 lineage from two Navajos. However, despite these discrepancies and possible LGT, the population structure of HPV-18 viruses does largely mirror human migrations reliably and shows reasonable congruence between independent genes within and between subtypes, unlike many other viruses.

JC virus

The last virus to be considered, the human polyomavirus JCV is also a DNA virus and has been extensively investigated in regard to human migrations (Sugimoto *et al.* 2002a; Yogo *et al.* 2004). Although JCV occasionally causes fatal progressive multifocal leukoencephalopathy in patients with compromised immune systems (Astrom *et al.* 1958), it usually results in a harmless, persistent kidney infection that is widespread in humans. After primary infection, JCV persists in renal tissues (Chesters *et al.* 1983) and viral samples can be easily collected from the urine of most subjects. Its lack of virulence may not trigger a host immune response and JCV can be considered to be a human commensal (Fernandez-Cobo *et al.* 2002). Mutation rate estimates for JCV vary between 1 and 5×10^{-7} per site per year (Hatwell & Sharp 2000; Fernandez-Cobo *et al.* 2002), a range that is ideal for tracing human history. Renal infection is normally acquired during adolescence but the transmission route is largely vertical, from parents to children, which

also enhances the utility of JCV for investigating human populations (Kunitake *et al.* 1995; Suzuki *et al.* 2002).

Four main viral types were described in the late 1990s (Agostini *et al.* 1997; Sugimoto *et al.* 1997). Type 1 is common in Europeans, type 2 is predominant in Asia, type 3 is African in origin and type 4 seemed to be restricted to the United States. By using the same sampling scheme as for HPV, Agostini and colleagues demonstrated that two native American ethnic groups, Navajo and Flathead, are predominantly infected by type 2 (Asian) JCV (Agostini *et al.* 1997). More specifically, these two ethnic groups carry the type 2 A subtype, a lineage that is otherwise found only in China and Japan. Based on these results, the authors concluded that JCV infection of Native Americans predated contact with Europeans and might reflect ancestral Amerind migrations across the Bering Strait. Additionally, nine subtypes that differ from each other by about 1–3% of their nucleotide sequence were described in a global sample that excluded the Americas (Sugimoto *et al.* 1997). Although the nomenclature was different, the main lineages can readily be compared between the two reports. As expected, JCV from African Americans is a genetic admixture between type 3 (African) and types 1 and 4 (European) (Stoner *et al.* 2000).

After these initial conclusions were published, JCV became a popular tool for deciphering localized problems concerning human population genetics and gene flow. Studies that focused on New World tribes provided novel insights concerning the number of migratory waves that settled the Americas from northeastern Asia (Fernandez-Cobo *et al.* 2002; Sugimoto *et al.* 2002b; Zheng *et al.* 2003, 2004). JCV has also been used to tackle small-scale questions, like the peopling of the Philippine archipelago by multiple immigration waves (Miranda *et al.* 2003). Large-scale analyses in China and Mongolia discovered different subtypes that varied in abundance (Guo *et al.* 1998; Cui *et al.* 2004) (Fig. 2). The genetic distinction between northern and southern Chinese (Cavalli-Sforza *et al.* 1994) was supported by a latitudinal cline consisting of high frequencies of the genotype CY in the north and high frequencies of another genotype (SC) in the south (Guo *et al.* 1998). Another genotype B1-b was more prevalent in Mongolia than in other regions, in good agreement with human genetic studies in this area (Cavalli-Sforza *et al.* 1994). An identical pattern was unravelled by Cui *et al.* (2004), where type 7A was most prevalent in south China and type 7B was predominant in northeast China and Inner Mongolia. In this later study, another type, type 2D was only found in Mongolia and India but not in South China (Fig. 2).

In a large European study comprising more than 300 individuals from central-eastern Europe and Spain, including Gypsy cohorts and Basques, Agostini *et al.* (2001) detected slight differences in genotype frequency that might correspond to fine genetic substructuring. For example,
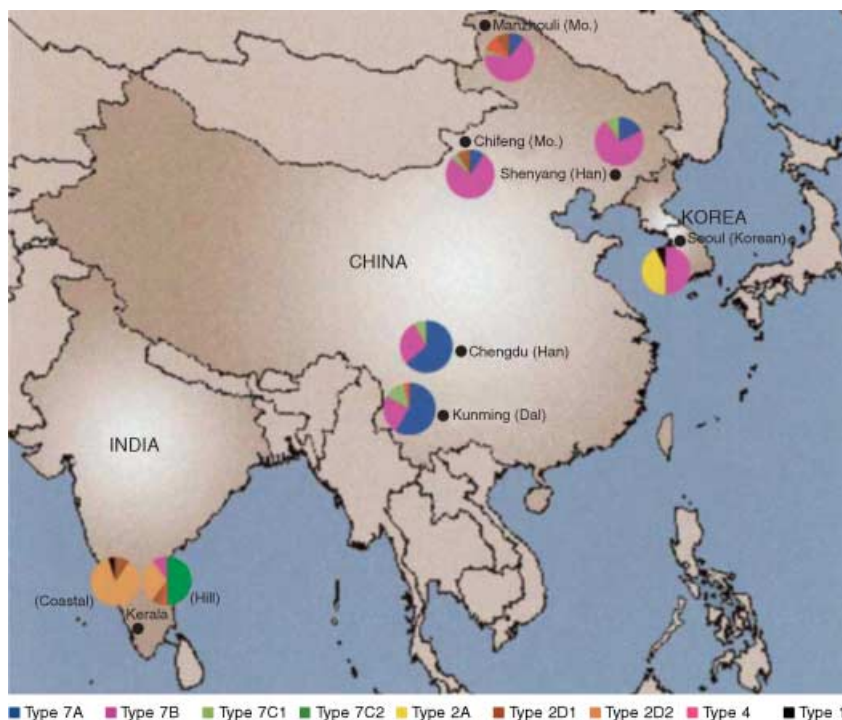
Fig. 2 Geographical regions of JCV sample collection in China, India and Korea. Average proportions of individual JCV genotypes are indicated by different colours in the pie charts (from Cui *et al*. 2004).

■ Type 7A ■ Type 7B ■ Type 7C1 ■ Type 7C2 ■ Type 2A ■ Type 2D1 ■ Type 2D2 ■ Type 4 ■ Type 1

Basques have significantly higher frequencies of a rare genotype (type 4), a genetic signature that confirms the peculiar status of this old European ethnic group. Interestingly, the same pattern was observed for Spanish Gypsies indicating a potential link between these two communities. A cline from southwest Europe to east Europe was also observed for types 1A vs. 1B.

JCV has fewer limitations as a tracer for human migrations than the RNA viruses do. One apparent contradiction was that in phylogenetic trees of JCV, a European lineage is basal (Sugimoto *et al*. 1997). However, this discrepancy was partially resolved by Pavesi, who found that type 6 from West Africa is the ancestral type (Pavesi 2003). Apparently the earlier finding of a basal European type reflected differential GC content in the main lineages; analysis of fast-evolving sites suggests two migrations of JCV out of Africa, one toward Eurasia and one toward Europe. However, this interpretation does not clarify why the genetic diversity of African JCV is less than in European and Asiatic samples (Wooding 2001). Recombination does occur, but it seems to be extremely rare (Stoner *et al*. 2000). One such recombination event resulted in type 5, which is actually a mosaic between types 2 and 6 (Hatwell & Sharp 2000). Thus, JCV fulfils many of the requirements for a suitable marker of human, namely that it mirrors the ethnic and geographical divergence of its host over thousands of years, evolved early with its host, is widespread, and does not seem to be under strong selective pressure. Moreover, it is largely transmitted vertically, or at least does not spread beyond the local community.

Helicobacter pylori

Helicobacter pylori was discovered 20 years ago (Marshall & Warren 1984; Marshall *et al*. 1984) and has recently been added to the list of potential candidates for tracing human migrations. This Gram-negative bacterium chronically infects the stomach of about half of the human population and causes chronic gastritis, peptic and duodenal ulcers, and MALT (mucosa-associated lymphatic tissue lymphomas). However, clinical disease only occurs in a minority of infected individuals and in most humans, *H. pylori* has many of the features of a commensal that colonizes the gastric mucosa. Epidemiological studies have shown that transmission occurs predominantly within families (Tindberg *et al*. 2001; Kivi *et al*. 2003). Since its discovery, *H. pylori* has been the subject of numerous epidemiological and functional investigations (Censini *et al*. 1996; Covacci *et al*. 1999; Suerbaum & Michetti 2002) and a few population genetic analyses (Achtman *et al*. 1999; Achtman 2001). *H. pylori* recombines very frequently and was thought to possess a panmictic population structure (Go *et al*. 1996; Suerbaum *et al*. 1998). Genetic exchange occurs via DNA transformation between multiple strains that colonize a single individual (Suerbaum *et al*. 1998; Kersulyte *et al*. 1999), resulting in a heterogeneous pool of isolates that are closely related but differ in genes and alleles that have been acquired by recombination. Sequences of 10 gene fragments from sequential isolates taken 2 years apart allowed an estimate of nucleotide substitution rates, recombination size, and the minimal age

of *H. pylori* (Falush *et al.* 2001). Using Bayesian statistics, the mean size of recombinant chunks was estimated to be about 500 bp long, the mutation rate was estimated to be $\leq 4 \times 10^{-5}$, and the last common ancestor of all *H. pylori* strains is estimated to have existed at least 2500–11 000 years ago. The authors also extrapolated that the recombination rate in *H. pylori* is 50-fold higher than in *Escherichia coli*.

Despite the evidence for frequent recombination and apparent panmixis, phylogeographical differences exist within *H. pylori* that seem to reflect co-evolution with *H. sapiens.* In the mid-1990s, sequences of virulence genes indicated major differences between isolates from Asia and from Europe (van der Ende *et al.* 1998; Van Doorn *et al.* 1999), which was subsequently shown to reflect genetically different bacterial populations in the two areas (Achtman *et al.* 1999). It was suggested that these differences reflect ancient human migrations (Achtman *et al.* 1999; Covacci *et al.* 1999) but initial analyses of isolates from native Peruvians and Guatemalan Ladinos showed that they carried bacteria with genetic signatures that were also found among isolates from Spain (Kersulyte *et al.* 2000). As a result, it was suggested that *H. pylori* infection of humans was so recent that Native Americans did not carry these bacteria during their migrations across the Bering Strait. Subsequent analyses showed, however, that the *H. pylori* genotypes found in South American Puerto Ayacucho Indians in isolated areas of the Amazon were largely of East Asian origin whereas isolates from South American cities were of European origin (Ghose *et al.* 2002). These studies are consistent with an ancient colonization of human by *H. pylori*, ancient enough that these bacteria were first introduced to the Americas 13 000 years ago from northeast Asia, followed by partial displacement by European isolates after colonization of the Americas. One problem with these analysed is that they rely on phylogenetic reconstructions, which are problematical in the presence of extensive recombination. A second problem is that the sequences analyses were from *vacA* (vacuolating cytotoxin gene) and *cagA* (cytotoxin-associated gene) or neighbouring loci, which are virulence factors that are likely to be under strong positive selection. However, a subsequent global analysis based on sequences of housekeeping genes confirmed that isolates from Ladinos in South American cities did derive their ancestry from Europe or Africa (Falush *et al.* 2003b) and that the East Asian genotypes from Amerinds could only be found in isolated aboriginal communities.

In a global study, Falush *et al.* (2003b) investigated 370 *H. pylori* isolates from 27 geographical and ethnic groups by population genetic techniques based on the Bayesian program STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003a) (see Box 1). STRUCTURE assigns individual genotypes to modern populations and also allows the reconstruction of the properties of the ancestral populations. Based on the concatenated sequences from seven housekeeping gene fragments and one virulence gene, Falush *et al.* (2003b) were able to assign almost all isolates to the populations hpAfrica1, hpAfrica2, hpEastAsia and hpEurope. HpEastAsia contained three subpopulations, called hspAmerind, hspEAsia and hspMaori while hpAfrica1 contained two subpopulations, called hspWAfrica and hspSAfrica. These seven populations and subpopulations were characterized by clear geographical distributions that are reflected in their names. The modern *H. pylori* populations are derived from five ancestral populations called ancestral Africa1, ancestral Africa2, ancestral East Asia, ancestral Europe1 (AE1) and ancestral Europe2 (AE2). Convincingly, many strains isolated from Inuits and Athabaskans in North America were assigned to hspAmerind while most isolates from New Zealand Maoris and Tongans were assigned to hspMaori. Furthermore, within hpEurope, a north-to-south introgression gradient of AE1 vs. AE2 was detected, in agreement with observations based on human genetics (Cavalli-Sforza *et al.* 1994). In the same vein, Spanish strains have the highest proportion of African ancestry, reflecting invasions and intermingled history with North African populations. STRUCTURE also deciphers the presence of recombinant nucleotides from distinct populations (Box 1), providing evidence for extensive admixture in areas such as South Africa, which contained a mixture of hpAfrica1, hpAfrica2 and hpEurope strains.

Although the geographical sources of the *H. pylori* populations were clearly congruent with human migrations, it was unclear whether analysis of *H. pylori* could provide more useful information than can analyses of human markers (Spratt 2003). A preliminary answer was provided by a second project that was designed to test this very question (Wirth *et al.* 2004). This study focused on a remote area in the trans-Himalayan region, the Ladakh province of northern India. There, two major ethnic groups (Buddhists and Muslims) have coexisted for close to 1000 years but remain isolated due to religious and cultural differences. Sequences of the same housekeeping gene fragments as above revealed that the population genetic structure differs between *H. pylori* from Buddhists and Muslims. STRUCTURE detected three ancestral populations, corresponding to ancestral East Asia, ancestral Europe1 and ancestral Europe2. Isolates from Buddhists were mosaics of nucleotides from ancestral East Asia and ancestral Europe1, consistent with the introduction of Buddhism by Tibetans into an ancient Ladakhi population. In contrast the majority of the ancestry of almost all isolates from Muslims was ancestral Europe1, indicating that the Islamic religion was introduced by few missionaries rather than extensive migrations. These conclusions are consistent with the known recent history of this region, whereas

Box 1

Raw data:
Aligned sequences from concatenated housekeeping genes.

Input file:
First column left corresponds to the individual. Second column corresponds to a defined grouping. Variable nucleotide are coded as follow, A = 1111, C = 2222, G = 3333 and T = 4444. Notice that the encoding needs to double the genotype.

Sequences:

A1	ACTCTTTGAAAG
A2	ACTCTTTCAAAG
A3	ACTCATTCAACG
B1	ACTCTTTAAAAG
B2	ACTCTTTAAAAG

STRUCTURE input file:

[Profile]

Title = 'STRUCTURE design'
Nb samples = 2
Genotypic data = 1
Gametic phase = 1
Data type = MICROSAT
Locus separator = TABULATION

[Data]
    [[Samples]]
   Sample name = 'PopB'
    Sample size = 3
    Sample data = {
    A1   1   4444 3333 1111
               4444 3333 1111
    A2   1   4444 2222 1111
               4444 2222 1111
    A3   1   1111 2222 2222
               1111 2222 2222
    }
    Sample name = 'PopB'
    Sample size = 2
    Sample data = {
    B1   1   4444 1111 1111
          1   4444 1111 1111
    B2   1   4444 1111 1111
          1   4444 1111 1111
    }

Ancestry estimates (individuals):
Detailed ancestry estimates for an *H. pylori* strain isolated from a black South African patient. Every single point corresponds to a polymorphic nucleotide which is assigned to each population with a given probability. This isolate reveals that six genes segments are clearly from Africa2 (red) origin, with high probabilities ($P = 1$); whereas two other genes, *efp* and *ppa* seem to be imported from an Africa1 individual. Mosaic patterns observed in *H. pylori* are the direct proof of direct contact and homologous recombination between different microbial populations.
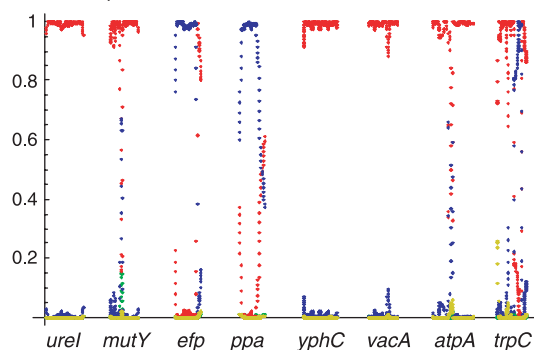
Ancestry estimates (populations):
Ternary plots of proportion of ancestry from each of three ancestral sources (blue, Africa1; yellow, East Asia; and green, Indo-Europe) is represented by its proximity to the corresponding corner of the triangle. Note that Bangladeshi and typical European samples do have similar ancestries. Fins have the highest proportion of East Asiatic genes within European samples and Mestizo who have multiple origins (Black, European and Asiatic) clearly display strong component from all three major *H. pylori* populations.
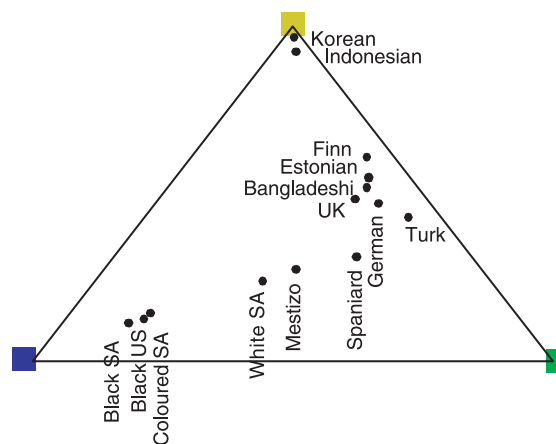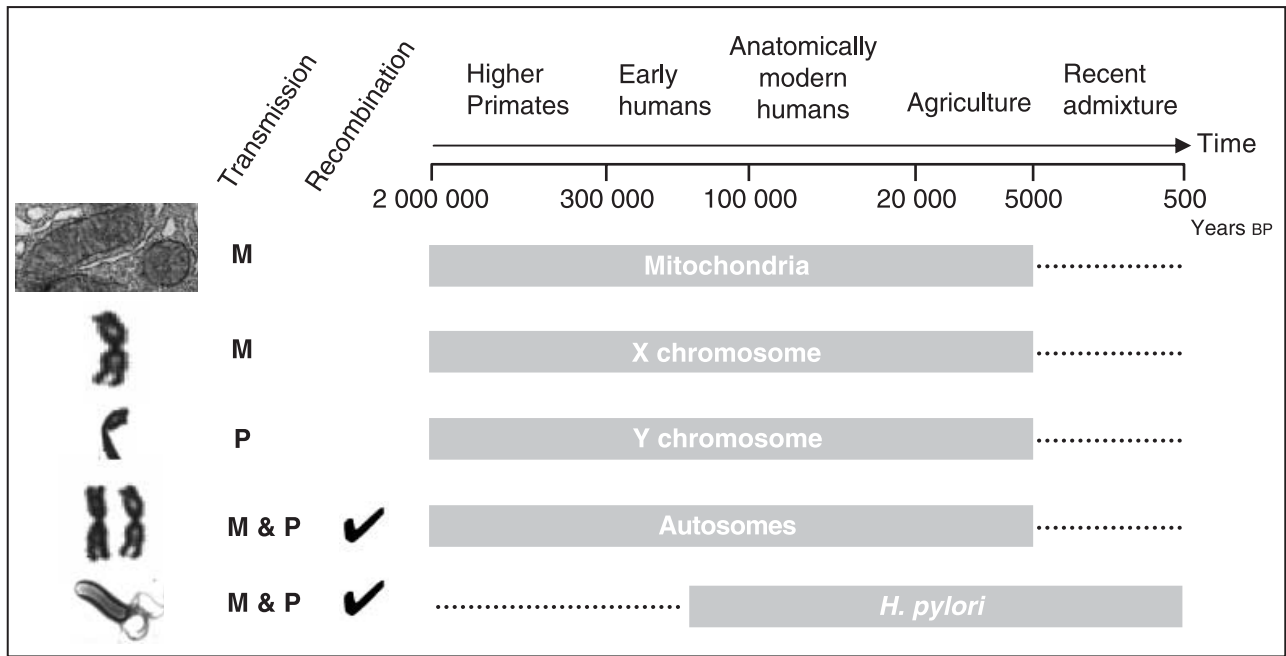
Fig. 3 Synthetic scheme illustrating the suitability and the characteristics of different genetic markers for deducing human population migrations. Transmission: M, maternal; P, paternal.

similar signals were not found with either human microsatellites or mtDNA probably because these markers are not suitable for detecting very recent admixtures (Fig. 3). Thus, *H. pylori* has considerable promise for tracing human population structures and migrations (Disotell 2003).

Helicobacter pylori is largely transmitted within families, at least in industrialized countries. Highly related strains were found in parents and children in Germany (Suerbaum *et al.* 1998), and Estonia (Kersulyte *et al.* 1999) and between older siblings from various countries (Li *et al.* 2002). Horizontal transmission may be more common in other areas because diverse strains were found within Algerian families (Raymond *et al.* 2004) and transmission has been attributed to drinking water in South America (Feldman 2001). Even when transmission occurs between unrelated children, it rarely crosses ethnic barriers, as shown by the rare isolation of European isolates in Okinawa, Japan, despite the extensive presence of the US military since World War II (Azuma *et al.* 2004). The current data indicate that over long time periods *H. pylori* does spread between ethnic groups, as indicated by the presence of hpEurope bacteria among black South Africans (Falush *et al.* 2003b), but this process can take centuries.

Helicobacter pylori also possesses limitations. Sampling relies on microbiological cultivation in a microaerophilic environment from biopsies taken during gastroendoscopy, an invasive procedure. Thus, obtaining samples, particularly from remote regions, can present problems. Moreover, much of current sampling is biased towards isolates from humans with clinical symptoms, which might reduce the frequency of less pathogenic strains, and depends on cultivation, which will ignore organisms that grow poorly or have special growth requirements. Furthermore, *H. pylori* is not universally present and its prevalence is decreasing in industrialized countries.

Crucial parameters: mode of transmission, mutation and recombination rates

One crucial parameter for the utility of microbes as genetic tracers for their hosts is whether the microbial distribution reflects dispersal or vicariance. Many human viruses, such as influenza or HIV, spread epidemically, i.e. horizontally (Fig. 4). Horizontally transmitted pathogens are unlikely to mirror host population genetic structure and instead will possess population structures that will reflect the emergence and rapid spreading of particular genotypes. Tracing human history through microbes needs to focus on nonepidemic microorganisms that are largely transmitted vertically. Low rates of horizontal transmission are acceptable when they reflect genetic exchanges occurring along geographical boundaries. Another issue concerns the rate of the microbial molecular clock. Ancient host history (~500 000 years) can only be reconstructed with a 'genetic tracer' whose rate of accumulation of mutations is not too high, because saturation at informative sites can otherwise lead to the loss of phylogenetic signal. The reverse is also problematic because molecular clocks that are too slow
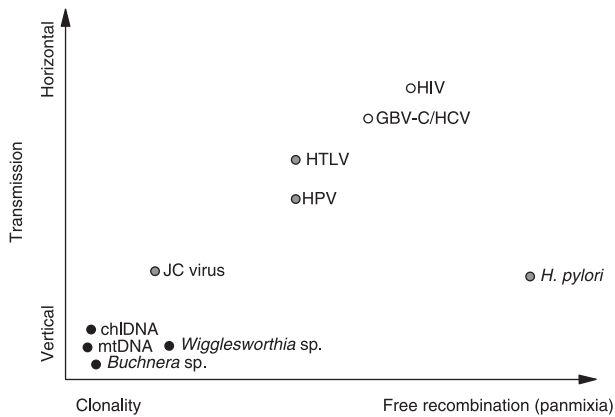
Fig. 4 A comparison of the transmission mode and the recombination rates among different bacterial and viral species used to infer the history of their hosts. The shading corresponds to three classes of mutation rates. White (high mutation rates), grey (intermediate mutation rates) and black (low mutation rates).

prevent the reconstruction of very recent host migration due to a lack of signal. Therefore, the divergence time of the ideal 'microbial tracer' would be similar to that of its host and microbial molecular clock rates need to fit within an appropriate range for the timescale of the investigation.

Recombination causes nontrivial problems because extensive homologous recombination affects data interpretation and can reduce the reliability of molecular phylogenies. Recombination results in long-terminal branches, intermediate genotypes, and leads to underestimates of the divergence times since the most recent common ancestor (Schierup & Hein 2000). When recombination is the main driving force of microevolution, sophisticated population genetic tools should replace tree-based approaches.

Cumulative evidence

Testing for vertical transmission within familial networks

Extensive knowledge of the transmission mode should be available before a microbe should be considered as a tracer of the history of its host. This knowledge can be acquired in many ways, including investigations of transmission from mother to child (Cavaco-Silva *et al.* 1997; Kashiwagi *et al.* 2004), parent to child (Kunitake *et al.* 1995; Kivi *et al.* 2003) and within families over several generations (Suzuki *et al.* 2002). Studies based on identical and nonidentical twins might also provide useful insights.

Multiple gene approaches

A gene tree is not an organismal tree and therefore phylogenies should be based on multiple genes. By doing so, a 'partition homogeneity test' (Farris *et al.* 1995) can be used to detect heterogeneity in phylogenetic signals among the individual genes. If the test reveals insignificant heterogeneity, the different genes can be concatenated for subsequent analyses. The use of multiple genes typically results in stronger statistical support and more accurate phylogenies. The Shimodaira and Hasegawa test (Shimodaira & Hasegawa 1999) can be applied in order to estimate the likelihood of different topologies (Quenouille *et al.* 2004), a question that is particularly appropriate for the possible African origin of human viruses (Vandamme *et al.* 1998; Van Dooren *et al.* 2001; Pavesi 2003). Different genes telling different stories might reflect selection or recombination.
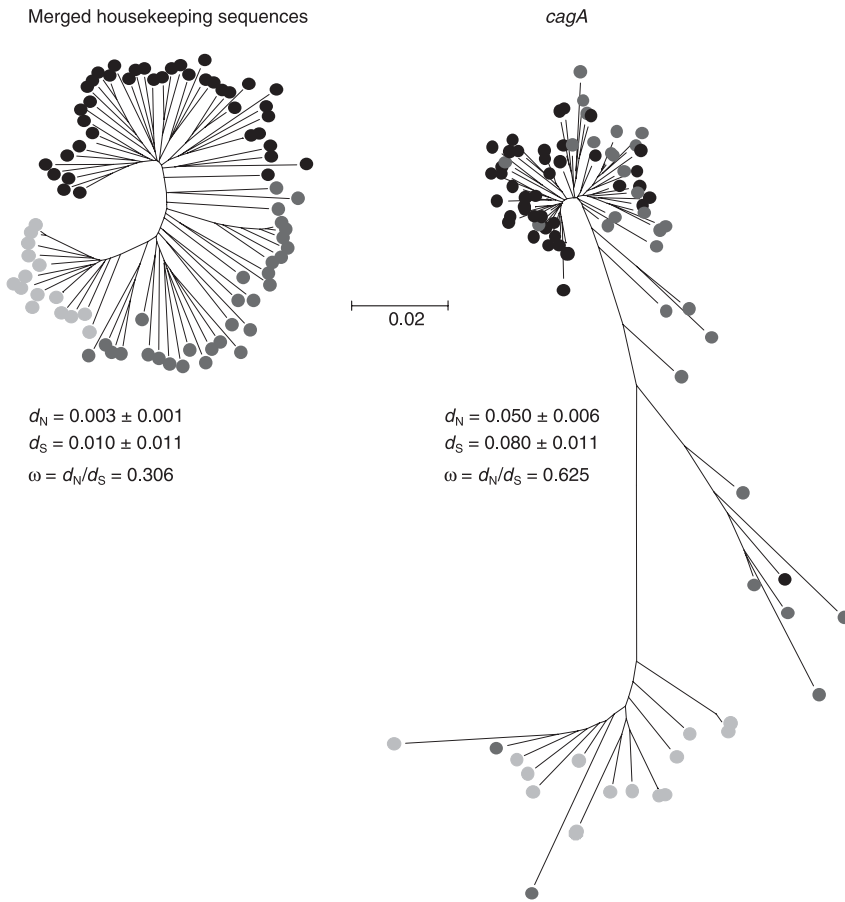
Assessing recombination rates

The spectrum of r/m, the ratio of recombination to mutation, is rather large and can range in bacteria from close to 0 up to 100 (Suerbaum *et al.* 1998; Feil & Spratt 2001; Feil *et al.* 2001). Similar ranges are observed in viruses (Worobey & Holmes 2001; Posada *et al.* 2002). The highest rates of viral recombination are observed in retroviruses, which carry two copies of their genomes within each mature virion, thereby promoting homologous recombination.

Different approaches are available to detect and assess recombination rates; some of which focus on single events between two to four strains (Husmeier & McGuire 2003), whereas others yield a global and synthetic picture of ongoing gene flow (Posada & Crandall 2001). Both coalescent- and Monte Carlo-based simulation methods for detecting and estimating recombination from gene sequences are now available (Haubold *et al.* 1998; Nielsen 2000; Worobey 2001; McVean *et al.* 2002). For example, the population recombination rate can be estimated by a composite likelihood method such as LD_{HAT}. LD_{HAT} employs a parametric approach, based on the neutral coalescent, to estimate the scaled parameters $2N_er$, where N_e is the effective population size and r is the rate at which recombination events separate adjacent nucleotides (McVean *et al.* 2002). An alternative solution is to use the homoplasy ratio (Maynard Smith & Smith 1998; Maynard Smith 1999). Furthermore, once a clear departure from clonality is detected, trees should be scrutinized cautiously and should possibly be replaced by population genetic methods such as STRUC-TURE, which can unravel the amount of recombination occurring between different populations (Box 1).

Testing for selection

The effects of selection are potentially even more detrimental on the reliability of gene trees than are those of recombination. Recombination generally occurs between proximate and only rarely between distant populations. As a result, populations can still be identified as swarms

Merged housekeeping sequences

cagA

0.02

$d_N = 0.003 \pm 0.001$
$d_S = 0.010 \pm 0.011$
$\omega = d_N/d_S = 0.306$

$d_N = 0.050 \pm 0.006$
$d_S = 0.080 \pm 0.011$
$\omega = d_N/d_S = 0.625$

Fig. 5 Comparison of a concatenated housekeeping genes tree with a virulence gene tree (*cagA*) for 87 *Helicobacter pylori* isolates from Africa, East Asia and Indo-Europe. Both neighbour-joining trees rely on the Tamura–Nei + *G* + *I* model. Notice that the trees are unrooted. The different strains are shaded, black, African strains; dark grey, Indo-European strains; and light grey, East Asian strains. d_N and d_S values were calculated using the Kumar method implement in MEGA (Kumar *et al.* 2001).

of strains that are closely related even though their phylogenetic relationships may be obscured. In contrast, selection can result in artificial groupings of populations with different histories and genealogies. This is illustrated in Fig. 5, which contrasts the topologies of trees based on housekeeping genes in *H. pylori* with that of a tree based on the *cagA* virulent factor. The neighbour-joining tree obtained for the housekeeping genes shows one monophyletic East Asian lineage, one nearly monophyletic African lineage and an intermediate paraphyletic European lineage. Nevertheless, the three major lineages do not overlap and coincide well with human population genetic structure. Using the same data set, the *cagA* phylogeny fails to split hpEurope and hpAfrica1 *H. pylori* strains, possibly due to a combination of selection plus selective sweeps of fitter alleles within Africa and Europe. Both trees are on the same scale and it can be observed that the virulence gene tree has longer branches and appears to evolve at a higher rate than the housekeeping genes. Several methods have been developed to test for adaptive molecular evolution. Comparison of nonsynonymous amino acid changes (d_N) and synonymous amino acid changes (d_S) in protein-coding genes provides a somewhat crude measure of selection pressure. The ratio of these two rates is measured by $\omega = d_N/d_S$ with $\omega = 1$, < 1, and > 1 indicating neutral evolution, purifying selection, and directional selection, respectively. Using such an approach, it appears that the merged housekeeping genes (Fig. 5) are under strong purifying selection and that *cagA* is closer to neutral evolution. However, a limitation of lineage-specific models is that they only detect positive selection when the average d_N is greater than the average d_S and do not allow for site-specific analyses. Thus, if diversifying or adaptive selection only affects a few amino acids, they will remain undetected.

Several evolutionary models exist that account for site-specific differences on adaptive selection at the protein level (Yang 1997; Nielsen & Huelsenbeck 2002; Yang & Nielsen 2002). Their utility has been already demonstrated (McClellan *et al.* 2004) and they should be used to screen DNA sequences for signatures of selection.

To illustrate the power of likelihood approaches, we used the CODEML software package in PAML (Yang 1997). Parameter estimates and log-likelihood values under models of variable ω among sites for *cagA* are presented in Table 1. Models of variable ω ratios among sites were used to test for the presence of sites under diversifying selection (ω > 1) and to identify them. Five models were used.

Table 1 Parameter estimates and log-likelihood values under models of variable ω ratios among sites for *cagA*

Model	p	Parameters	l	d_N/d_S	Positively selected sites
M0: one ratio	1	ω = 0.513	−5142.69	= ω	None
M1: neutral	1	$p_0 = 0.534$, $\omega_0 = 0$ $p_1 = 0.466$, $\omega_1 = 1$	−4963.27	0.466	Not allowed
M2: selection	3	$p_0 = 0.420$, $\omega_0 = 0$ $p_1 = 0.302$, $\omega_1 = 1$ $p_2 = 0.278$, $\omega_2 = 0.172$	−4940.00	0.350	No site
M3: discrete	5	$p_0 = 0.651$, $\omega_0 = 0.057$ $p_1 = 0.282$, $\omega_1 = 1.092$ $p_2 = 0.067$, $\omega_2 = 3.950$	−4900.52	0.608	55 sites
M7: beta	2	$p = 0.054$, $q = 0.067$	−4940.66	0.375	Not allowed
M8: beta and ω	4	$p_0 = 0.930$, $p = 0.064$, $q = 0.083$ $p_1 = 0.070$, $\omega = 3.778$	−4899.35	0.610	8 sites

Model M1 (neutral) assume two classes of sites in the protein; conserved sites with ω = 0 and neutral sites with ω = 1. Model M2 (selection) adds a third class of sites with ω as a free parameter, thus allowing for sites under positive selection. Model M3 (discrete) uses a general discrete distribution with three site classes, with the proportions (p_0, p_1 and p_2) and the ω ratios (ω_0, ω_1 and ω_2) estimated from the data. Model M7 (beta) uses a beta distribution B (p, q), which, depending on parameters p and q, can take various shapes in the interval (0, 1). Model M8 (beta and ω) adds an extra class of sites to the beta model, with the proportion and the ω ratio estimated from the data, thus allowing for sites with ω > 1.

Model M0 fits the data poorly when compared to model M3. The latter model involves four more parameters than M0, and the LRT statistic 2ΔL = 484.34 is much greater than the critical $x^2_{1\%} = 13.28$ with d.f. = 4. The results suggest variation in selective pressure among amino acid sites. Moreover, all three models that allow for the presence of sites under selection, i.e. M2 (selection), M3 (discrete) and M8 (beta and ω) fit the data better than alternative models that do not allow selection. A striking feature under the *cagA* 'selection' models M3 and M8 is that numerous codon sites seem to be under positive selection.

Inferring demographic histories

The demographic parameters of host and microbes have only rarely been compared but can provide an important criterion for testing co-evolution. Wooding (2001) tested whether human and JCV genes reflected host–parasite codemography. He showed convincingly that JCV could be grouped into two distinct clusters using a principal component map, one derived from Europe and the other from Africa and Asia. Both clusters have unimodal mismatch distributions, a pattern that is typically observed in expanding populations. Rogers's method of moments

(Rogers & Harpending 1992) gives estimators including $\theta = 2 N_0\mu$, where N_0 is an estimator of ancient effective population size and μ is the per-nucleotide substitution rate times the number of nucleotides considered. Another estimator is $\tau = 2\mu t$, where t corresponds to the elapsed time since the demographic change occurred. Using the molecular clock rate inferred for primate polyomaviruses (Yasunaga & Miyata 1982), Wooding (2001) inferred that the African/Asian cluster expanded 150 000 years ago and the European cluster expanded somewhat later, 115 000 years ago. These expansion times overlap with expansion time estimates for humans inferred from mtDNA. However, these calculations are highly dependent on the mutation rate estimates and must be interpreted with caution. In increasing order of complexity, the following analyses could be made: Tajima's *D*-test (Tajima 1989a, b), Rogers's method of moments (Rogers & Harpending 1992), and finally coalescent-based methods (Kuhner *et al.* 1998, 2000; Pybus & Rambaut 2002; Drummond *et al.* 2003; Wirth & Bernatchez 2003).

Direct comparative methods: testing human and microbial markers from the same samples

A reasonable and intuitive way to check host–parasite co-evolution would consist of screening both hosts and parasites concurrently. This would allow a direct comparison of the inferred genetic structures and avoid sources of noise such as ambiguity of source or origin and absence of genetic information from a sampled area. This strategy has been applied for *H. pylori* by Ghose *et al.* (2002) and Wirth *et al.* (2004). A direct comparison of host and parasite markers from the same material also allows reliable estimates of the resolution and discriminatory power of the microbe vs. the host. Without such an approach, we would still not know whether *H. pylori* is a particularly promising tool for inferring very recent genetic structure in human populations

(Fig. 3) that have been shaped by geographical isolation, drift, and admixture (Wirth *et al*. 2004).

Concluding remarks

Many lessons can be learned from microbes about our evolutionary history. Moreover, many human ethnic groups who have genetic traces of our past are threatened by extinction, and increasing gene flow compromises their genetic uniqueness. Recently, the Human Genome Diversity Project (HGDP) was initiated in cooperation with the Jean Dausset foundation (CEPH); its principal task is to collect blood samples from human populations on a global scale (Cann *et al*. 2002; Rosenberg *et al*. 2002). Because tracers of human migrations such as *H. pylori* and the JCV definitively provide new insights into our recent history, it would be advisable to also establish a parallel sampling scheme for these microbes.

The human–*H. pylori*/JCV systems described here may provide paradigms that will promote the use of microbial species in order to decipher the evolutionary history of diverse target organisms, as exemplified by a recent analysis of the cause of leprosy (Monot *et al*. 2005). To a large extent, other host–parasite systems may share similar virtues and limitations to those presented here. Primate evolutionary history could also be revisited using the hepatitis G virus, a virion, which has been present in New World and Old World apes for more than 10 Myr. New *Helicobacter* species have been described in the stomach of rodents, baboons, and seals and these might provide the same resolution power as genetic tracers that were observed in humans. The genetic changes in crops and livestock during domestication are particularly topical (Harter *et al*. 2004; Larson *et al*. 2005) and might also profit from new genetic tools. New insights might also be gained from the analysis of fungi, endophytes or myccorhiza that undergo interspecific hybridization which leave signatures of contacts between diverse crop species and polyploidg (Schardl & Craven 2003; Arnold 2004). These are only a few examples of how such approaches might provide additional tool for the population geneticist's box that could become increasingly useful, if not predominant, for the study of species with depauperized genetic diversity.

References

Achtman M (2001) Population structure of *Helicobacter pylori* and other pathogenic bacterial species. Horizon Scientific Press, Wymondham, UK.

Achtman M, Azuma T, Berg DE *et al*. (1999) Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Molecular Microbiology*, **32**, 459–470.

Agostini HT, Yanagihara R, Davis V, Ryschkewitsch CF, Stoner GL (1997) Asian genotypes of JC virus in Native Americans and in a Pacific Island population: markers of viral evolution and human migration. *Proceedings of the National Academy of Sciences, USA*, **94**, 14542–14546.

Agostini HT, Deckhut A, Jobes DV *et al*. (2001) Genotypes of JC virus in east, central and southwest Europe. *Journal of General Virology*, **82**, 1221–1331.

Arnold ML (2004) Natural hybridization and the evolution of domesticated, pest and disease organisms. *Molecular Ecology*, **13**, 997–1007.

Astrom KE, Mancall EL, Richardson EP Jr (1958) Progressive multifocal leuko-encephalopathy: a hitherto unrecognized complication of chronic lymphatic leukaemia and Hodgkin's disease. *Brain*, **81**, 93–111.

Azuma T, Kato S, Zhou W *et al*. (2004) Diversity of *vacA* and *cagA* genes of *Helicobacter pylori* in Japanese children. *Alimentary Pharmacology and Therapeutics*, **20**, 1–6.

Canback B, Tamas I, Andersson SG (2004) A phylogenomic study of endosymbiotic bacteria. *Molecular Biology and Evolution*, **21**, 1110–1122.

Cann HM, de Toma C, Cazes L *et al*. (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36.

Cavaco-Silva P, Taveira NC, Lourenco MH, Santos Ferreira MO, Daniels RS (1997) Vertical transmission of HIV-2. *Lancet*, **349**, 177–178.

Cavalli-Sforza LL (1998) The DNA revolution in population genetics. *Trends in Genetics*, **14**, 60–65.

Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, **33**, 266–275.

Cavalli-Sforza L, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton, New Jersey.

Censini S, Lange C, Xiang Z *et al*. (1996) *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proceeding of the National Academy of Sciences, USA*, **93**, 14648–14653.

Chan SY, Ho L, Ong CK *et al*. (1992) Molecular variants of human papillomavirus type 16 from four continents suggests ancient pandemic spread of the virus and its coevolution with humankind. *Journal of Virology*, **66**, 2057–2066.

Chesters PM, Heritage J, McCance DJ (1983) Persistence of DNA sequences of BK virus and JC virus in normal human tissues and in diseased tissues. *Journal of Infectious Diseases*, **147**, 676–684.

Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.

Cornelissen M, van Den Burg R, Zorgdrager F, Goudsmit J (2000) Spread of distinct human immunodeficiency virus type 1 AG recombinant lineages in Africa. *Journal of General Virology*, **81**, 515–523.

Covacci A, Telford JL, Del Giudice G, Parsonnet J, Rappuoli R (1999) *Helicobacter pylori* virulence and genetic geography. *Science*, **284**, 1328–1333.

Cui X, Wang JC, Deckhut A *et al*. (2004) Chinese strains (Type 7) of JC virus are Afro-Asiatic in origin but are phylogenetically distinct from the Mongolian and Indian strains (Type 2D) and the Korean and Japanese strains (Type 2A). *Journal of Molecular Evolution*, **58**, 568–583.

Disotell TR (2003) Discovering human history from stomach bacteria. *Genome Biology*, **4**, 213.

Domingo E, Escarmis C, Sevilla N *et al.* (1996) Basic concepts in RNA virus evolution. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, **10**, 859–864.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003) Measurably evolving populations. *Trends in Ecology & Evolution*, **18**, 481–488.

Eshleman JA, Malhi RS, Smith DG (2003) Mitochondrial DNA studies of Native Americans: conceptions and misconceptions of the population prehistory of the Americas. *Evolutionary Anthropology*, **12**, 7–18.

Falush D, Kraft C, Taylor NS *et al.* (2001) Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences, USA*, **98**, 15056–15061.

Falush D, Stephens M, Pritchard JK (2003a) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Falush D, Wirth T, Linz B *et al.* (2003b) Traces of human migrations in *Helicobacter pylori* populations. *Science*, **299**, 1582–1585.

Farris JS, Kallersjo M, Kluge AG, Bult C (1995) Testing significance of incongruence. *Cladistics*, **10**, 315–319.

Feil EJ, Spratt BG (2001) Recombination and the population structures of bacterial pathogens. *Annual Review of Microbiology*, **55**, 561–590.

Feil EJ, Holmes EC, Bessen DE *et al.* (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences, USA*, **98**, 182–187.

Feldman RA (2001) Epidemiological observations and open questions about disease and infection caused by *Helicobacter pylori* (eds Achtman M, Suerbaum S), pp. 29–51. *Helicobacter pylori*: Molecular and Cellular Biology.

Fernandez-Cobo M, Agostini HT, Britez G, Ryschkewitsch CF, Stoner GL (2002) Strains of JC virus in Amerind-speakers of North America (Salish) and South America (Guarani), Na-Dene-speakers of New Mexico (Navajo), and modern Japanese suggest links through an ancestral Asian population. *American Journal of Physical Anthropology*, **118**, 154–168.

Fisher MC, Koenig GL, White TJ *et al.* (2001) Biogeographic range expansion into South America by *Coccidioides immitis* mirrors New World patterns of human migration. *Proceedings of the National Academy of Sciences, USA*, **98**, 4558–4562.

Funk DJ, Helbling L, Wernegreen JJ, Moran NA (2000) Intraspecific phylogenetic congruence among multiple symbiont genomes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **267**, 2517–2521.

Gessain A, Pecon-Slattery J, Meertens L, Mahieux R (2000) Origins of HTLV-1 in South America. *Nature Medecine*, **6**, 232.

Ghose C, Perez-Perez GI, Dominguez-Bello MG *et al.* (2002) East Asian genotypes of *Helicobacter pylori* strains in Amerindians provide evidence for its ancient human carriage. *Proceedings of the National Academy of Sciences, USA*, **99**, 15107–15111.

Gibbons A (1996) The peopling of the Americas. *Science*, **274**, 31–33.

Go MF, Kapur V, Graham DY, Musser JM (1996) Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *Journal of Bacteriology*, **178**, 3934–3938.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences, USA*, **92**, 6723–6727.

Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. *Science*, **283**, 1476–1481.

Guo J, Sugimoto C, Kitamura T *et al.* (1998) Four geographically distinct genotypes of JC virus are prevalent in China and Mongolia: implications for the racial composition of modern China. *Journal of General Virology*, **79**, 2499–2505.

Harter OAV, Gardner KA, Falush D *et al.* (2004) Origin of extant domesticated sunflowers in eastern North America. *Nature*, **430**, 201–205.

Hatwell JN, Sharp PM (2000) Evolution of human polyomavirus JC. *Journal of General Virology*, **81**, 1191–1200.

Haubold B, Travisano M, Rainey PB, Hudson RR (1998) Detecting linkage disequilibrium in bacterial populations. *Genetics*, **150**, 1341–1348.

Ho L, Chan SY, Burk RD *et al.* (1993) The genetic drift of human papillomavirus type 16 is a means of reconstructing prehistoric viral spread and the movement of ancient human populations. *Journal of Virology*, **67**, 6413–6423.

Holmes EC (2004) The phylogeography of human viruses. *Molecular Ecology*, **13**, 745–756.

Husmeier D, McGuire G (2003) Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution*, **20**, 315–337.

Ishak R, Vallinoto AC, Azevedo VN *et al.* (2001) Molecular evidence of mother-to-child transmission of HTLV-IIc in the Kararao Village (Kayapo) in the Amazon region of Brazil. *Revista Da Sociedade Brasileira de Medicina Tropical (Rio de Janeiro)*, **34**, 519–525.

Jorde LB, Rogers AR, Bamshad M *et al.* (1997) Microsatellite diversity and the demographic history of modern humans. *Proceedings of the National Academy of Sciences, USA*, **94**, 3100–3103.

Kalyanaraman VS, Sarngadharan MG, Robert-Guroff M *et al.* (1982) A new subtype of human T-cell leukemia virus (HTLV-II) associated with a T-cell variant of hairy cell leukemia. *Science*, **218**, 571–573.

Kashiwagi K, Furusyo N, Nakashima H *et al.* (2004) A decrease in mother-to-child transmission of human T lymphotropic virus type I (HTLV-I) in Okinawa, Japan. *American Journal of Tropical Medicine and Hygiene*, **70**, 158–163.

Ke Y, Su B, Song X *et al.* (2001) African origin of modern humans in East Asia: a tale of 12 000 Y chromosomes. *Science*, **292**, 1151–1153.

Kersulyte D, Chalkauskas H, Berg DE (1999) Emergence of recombinant strains of *Helicobacter pylori* during human infection. *Molecular Microbiology*, **31**, 31–43.

Kersulyte D, Mukhopadhyay AK, Velapatino B *et al.* (2000) Differences in genotypes of *Helicobacter pylori* from different human populations. *Journal of Bacteriology*, **182**, 3210–3218.

Kivi M, Tindberg Y, Sorberg M *et al.* (2003) Concordance of *Helicobacter pylori* strains within families. *Journal of Clinical Microbiology*, **41**, 5604–5608.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.

Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, **156**, 439–447.

Kumar S, Tamura K, Jakobsen IB *et al.* (2001) MEGA2: Molecular Evolutionary Genetics Analysis Software. *Bioinformatics*, **17**, 1244–1245.

Kunitake T, Kitamura T, Guo J *et al.* (1995) Parent-to-child transmission is relatively common in the spread of the human polyomavirus JC virus. *Journal of Clinical Microbiology*, **33**, 1448–1451.

Larson G, Dobney K, Albarella U *et al.* (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science*, **307**, 1618–1621.

Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biology*, **3**, e130.

Li HC, Fujiyoshi T, Lou H *et al.* (1999) The presence of ancient human T-cell lymphotropic virus type I provirus DNA in an Andean mummy. *Nature Medicine*, **5**, 1428–1432.

Li OL, Genta RM, Go MF *et al.* (2002) *Helicobacter pylori* strain and the pattern of gastritis among first-degree relatives of patients with gastric carcinoma. *Helicobacter*, **7**, 349–355.

Marshall BJ, Warren JR (1984) Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet*, **1**, 1311–1315.

Marshall BJ, McGechie DB, Francis GJ, Utley PJ (1984) Pyloric campylobacter serology. *Lancet*, **2**, 281.

Maynard Smith J (1999) The detection and measurement of recombination from sequence data. *Genetics*, **153**, 1021–1027.

Maynard Smith J, Smith NH (1998) Detecting recombination from gene trees. *Molecular Biology and Evolution*, **15**, 590–599.

McClellan DA, Palfreyman EJ, Smith MJ *et al.* (2004) Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome *b* proteins. *Molecular Biology and Evolution*, **22**, 437–455.

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

Merriwether DA, Ferrell RE (1996) The four founding lineage hypothesis for the New World: a critical reevaluation. *Molecular Phylogenetics and Evolution*, **5**, 241–246.

Merriwether DA, Hall WW, Vahlne A, Ferrell RE (1996) mtDNA variation indicates Mongolia may have been the source for the founding population for the New World. *American Journal of Human Genetics*, **59**, 204–212.

Miranda JJ, Sugimoto C, Paraguison R *et al.* (2003) Genetic diversity of JC virus in the modern Filipino population: implications for the peopling of the Philippines. *American Journal of Physical Anthropology*, **120**, 125–132.

Miura T, Fukunaga T, Igarashi T *et al.* (1994) Phylogenetic subtypes of human T-lymphotropic virus type I and their relations to the anthropological background. *Proceedings of the National Academy of Sciences, USA*, **91**, 1124–1127.

Miyoshi I, Kubonishi I, Yoshimoto S *et al.* (1981) Type C virus particles in a cord T-cell line derived by co-cultivating normal human cord leukocytes and human leukaemic T cells. *Nature*, **294**, 770–771.

Monot M, Honor ÈN, Garnier T *et al.* (2005) On the origin of leprosy. *Science*, **308**, 1040–1042.

Moran NA (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. *Current Opinion in Microbiology*, **6**, 512–518.

Murray-McIntosh RP, Scrimshaw BJ, Hatfield PJ, Penny D (1998) Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. *Proceedings of the National Academy of Sciences, USA*, **95**, 9047–9052.

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Nielsen R, Huelsenbeck JP (2002) Detecting positively selected amino acid sites using posterior predictive *P*-values. *Pacific Symposium on Biocomputing*, 576–588.

Ong CK, Chan SY, Campo MS *et al.* (1993) Evolution of human papillomavirus type 18: an ancient phylogenetic root in Africa and intratype diversity reflect coevolution with human ethnic groups. *Journal of Virology*, **67**, 6424–6431.

Pavesi A (2001) Origin and evolution of GBV-C/hepatitis G virus and relationships with ancient human migrations. *Journal of Molecular Evolution*, **53**, 104–113.

Pavesi A (2003) African origin of polyomavirus JC and implications for prehistoric human migrations. *Journal of Molecular Evolution*, **56**, 564–572.

Poiesz BJ, Ruscetti FW, Gazdar AF *et al.* (1980) Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proceeding of the National Academy of Sciences, USA*, **77**, 7415–7419.

Poiesz BJ, Ruscetti FW, Reitz MS, Kalyanaraman VS, Gallo RC (1981) Isolation of a new type C retrovirus (HTLV) in primary uncultured cells of a patient with Sezary T-cell leukaemia. *Nature*, **294**, 268–271.

Poloni ES, Semino O, Passarino G *et al.* (1997) Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *American Journal of Human Genetics*, **61**, 1015–1035.

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences, USA*, **98**, 13757–13762.

Posada D, Crandall KA, Holmes EC (2002) Recombination in evolutionary genomics. *Annual Review of Genetics*, **36**, 75–97.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pybus OG, Rambaut A (2002) GENIE: estimating demographic history from molecular phylogenies. *Bioinformatics*, **18**, 1404–1405.

Quenouille B, Bermingham E, Planes S (2004) Molecular systematics of the damselfishes (Teleostei: Pomacentridae): Bayesian phylogenetic analyses of mitochondrial and nuclear DNA sequences. *Molecular Phylogenetics and Evolution*, **31**, 66–88.

Rambaut A, Posada D, Crandall KA, Holmes EC (2004) The causes and consequences of HIV evolution. *Nature Reviews Genetics*, **5**, 52–61.

Raymond OJ, Thiberg JM, Chevalier C *et al.* (2004) Genetic and transmission analysis of *Helicobacter pylori* strains within a family. *Emerging Infectious Diseases*, **10**, 1816–1821.

Reich DE, Cargill M, Bolk S *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**, 552–569.

Rosenberg NA, Pritchard JK, Weber JL *et al.* (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Ruiz-Linares A, Ortiz-Barrientos D, Figueroa M *et al.* (1999) Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. *Proceedings of the National Academy of Sciences, USA*, **96**, 6312–6317.

Salemi M, Vandamme AM, Gradozzi C *et al.* (1998) Evolutionary rate and genetic heterogeneity of human T-cell lymphotropic virus type II (HTLV-II) using isolates from European injecting drug users. *Journal of Molecular Evolution*, **46**, 602–611.

Schardl CL, Craven KD (2003) Interspecific hybridization in plant-associated fungi and oomycetes: a review. *Molecular Ecology*, **12**, 2861–2873.

Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**, 879–891.

Semino O, Passarino G, Oefner PJ *et al.* (2000) The genetic legacy of Paleolithic *Homo sapiens* in extant Europeans: a Y chromosome perspective. *Science*, **290**, 1155–1159.

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, **16**, 1114–1116.

Simmonds P, Smith DB (1999) Structural constraints on RNA virus evolution. *Journal of Virology*, **73**, 5787–5794.

Slattery JP, Franchini G, Gessain A (1999) Genomic evolution, patterns of global dissemination, and interspecies transmission of human and simian T-cell leukemia/lymphotropic viruses. *Genome Research*, **9**, 525–540.

Sokal RR, Oden NL, Thomson BA (1992) Origins of the Indo-Europeans: genetic evidence. *Proceedings of the National Academy of Sciences, USA*, **89**, 7669–7673.

Sonoda S, Li HC, Cartier L, Nunez L, Tajima K (2000) Ancient HTLV type 1 provirus DNA of Andean mummy. *AIDS Research and Human Retroviruses*, **16**, 1753–1756.

Spratt BG (2003) Microbiology. Stomachs out of Africa. *Science*, **299**, 1528–1529.

Stoner GL, Jobes DV, Fernandez Cobo M *et al.* (2000) JC virus as a marker of human migration to the Americas. *Microbes and Infection*, **2**, 1905–1911.

Stringer C (2003) Human evolution: Out of Ethiopia. *Nature*, **423**, 692–695.

Su B, Jin L, Underhill P *et al.* (2000) Polynesian origins: insights from the Y chromosome. *Proceedings of the National Academy of Sciences, USA*, **97**, 8225–8228.

Suerbaum S, Michetti P (2002) *Helicobacter pylori* infection. *New England Journal of Medicine*, **347**, 1175–1186.

Suerbaum S, Smith JM, Bapumia K *et al.* (1998) Free recombination within *Helicobacter pylori*. *Proceedings of the National Academy of Sciences, USA*, **95**, 12619–12624.

Sugimoto C, Kitamura T, Guo J *et al.* (1997) Typing of urinary JC virus DNA offers a novel means of tracing human migrations. *Proceedings of the National Academy of Sciences, USA*, **94**, 9191–9196.

Sugimoto C, Hasegawa M, Kato A *et al.* (2002a) Evolution of human polyomavirus JC: implications for the population history of humans. *Journal of Molecular Evolution*, **54**, 285–297.

Sugimoto C, Hasegawa M, Zheng HY *et al.* (2002b) JC virus strains indigenous to northeastern Siberians and Canadian Inuits are unique but evolutionarily related to those distributed throughout Europe and Mediterranean areas. *Journal of Molecular Evolution*, **55**, 322–335.

Suzuki M, Zheng HY, Takasaka T *et al.* (2002) Asian genotypes of JC virus in Japanese-Americans suggest familial transmission. *Journal of Virology*, **76**, 10074–10078.

Tajima F (1989a) The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 597–601.

Tajima F (1989b) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Templeton A (2002) Out of Africa again and again. *Nature*, **416**, 45–51.

Thrall PH, Burdon JJ (2003) Evolution of virulence in a plant host–pathogen metapopulation. *Science*, **299**, 1735–1737.

Tindberg Y, Bengtsson C, Granath F *et al.* (2001) *Helicobacter pylori* infection in Swedish school children: lack of evidence of child-to-child transmission outside the family. *Gastroenterology*, **121**, 310–316.

Underhill PA, Shen P, Lin AA *et al.* (2000) Y chromosome sequence variation and the history of human populations. *Nature Genetics*, **26**, 358–361.

Vandamme AM, Salemi M, Van Brussel M *et al.* (1998) African origin of human T-lymphotropic virus type 2 (HTLV-2) supported by a potential new HTLV-2d subtype in Congolese Bambuti Efe Pygmies. *Journal of Virology*, **72**, 4327–4340.

Vandamme AM, Hall WW, Lewis MJ *et al.* (2000) Origins of HTLV-1 in South America. *Nature Medecine*, **6**, 232–233.

van der Ende A, Pan ZJ, Bart A *et al.* (1998) *cagA*-positive *Helicobacter pylori* populations in China and the Netherlands are distinct. *Infection and Immunity*, **66**, 1822–1826.

Van Dooren S, Salemi M, Vandamme AM (2001) Dating the origin of the African human T-cell lymphotropic virus type-i (HTLV-I) subtypes. *Molecular Biology and Evolution*, **18**, 661–671.

Van Dooren S, Pybus OG, Salemi M *et al.* (2004) The low evolutionary rate of human T-cell Lymphotropic Virus Type-1 confirmed by the analysis of vertical transmission chains. *Molecular Biology and Evolution*, **21**, 603–611.

Van Doorn LJ, Figueiredo C, Megraud F *et al.* (1999) Geographic distribution of vacA allelic types of *Helicobacter pylori*. *Gastroenterology*, **116**, 823–830.

Wang DG, Fan JB, Siao CJ *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

Wernegreen JJ (2002) Genome evolution in bacterial endosymbionts of insects. *Nature Reviews Genetics*, **3**, 850–861.

Wirth T, Bernatchez L (2003) Decline of North Atlantic eels: a fatal synergy? *Proceedings of the Royal Society London. Series B, Biological Sciences*, **270**, 681–688.

Wirth T, Wang X, Linz B *et al.* (2004) Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. *Proceedings of the National Academy of Sciences, USA*, **101**, 4746–4751.

Wooding S (2001) Do human and JC virus genes show evidence of host-parasite codemography? *Infection Genetics and Evolution*, **1**, 3–12.

Worobey M (2001) A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Molecular Biology and Evolution*, **18**, 1425–1434.

Worobey M, Holmes EC (2001) Homologous recombination in GB virus C/hepatitis G virus. *Molecular Biology and Evolution*, **18**, 254–261.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences: CABIOS*, **13**, 555–556.

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**, 908–917.

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP (2002a) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *American Journal of Human Genetics*, **70**, 635–651.

Yao YG, Nie L, Harpending H *et al.* (2002b) Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *American Journal of Physical Anthropology*, **118**, 63–76.

Yasunaga T, Miyata T (1982) Evolutionary changes of nucleotide sequences of papova viruses BKV and SV40: they are possibly hybrids. *Journal of Molecular Evolution*, **19**, 72–79.

Yogo Y, Sugimoto C, Zheng HY *et al.* (2004) JC virus genotyping offers a new paradigm in the study of human populations. *Reviews of Medical Virology*, **14**, 179–191.

Zhao Z, Jin L, Fu YX *et al.* (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proceedings of the National Academy of Sciences, USA*, **97**, 11354–11358.

Zheng HY, Sugimoto C, Hasegawa M *et al.* (2003) Phylogenetic relationships among JC virus strains in Japanese/Koreans and Native Americans speaking Amerind or Na-Dene. *Journal of Molecular Evolution*, **56**, 18–27.

Zheng HY, Zhao P, Suganami H *et al.* (2004) Regional distribution of two related Northeast Asian genotypes of JC virus, CY-a and -b: implications for the dispersal of Northeast Asians. *Microbes and Infection*, **6**, 596–603.

zur Hausen H (1989) Papillomaviruses in anogenital cancer as a model to understand the role of viruses in human cancers. *Cancer Research*, **49**, 4677–4681.

Thierry Wirth is a population geneticist working on human and bacterial evolution. His main line of research consists of characterizing and measuring the processes that determine bacterial populations, as well as the evolution of sex and virulence. Axel Meyer is an evolutionary geneticist interested in the ecological, developmental and genomic aspects of speciation. Mark Achtman investigates the population structure and evolutionary history of pathogenic bacteria, in particular *H. pylori*, *Yersinia pestis*, *Escherichia coli* and *Salmonella enterica*, by sequence-based techniques. The goals of these projects range from elucidating the phylogenetic history of these species through mapping their phylogeographic diversity to understanding the migrations of their hosts.