Molecular Ecology (2005) **14**, 179–188 doi: 10.1111/j.1365-294X.2004.02382.x

Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA

O. THALMANN,* D. SERRE,*‡ M. HOFREITER,* D. LUKAS,* J. ERIKSSON*† and L. VIGILANT*
**Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany, †Department of Animal Ecology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden*

Abstract

Numts are fragments of mitochondrial DNA (mtDNA) that have been translocated to the nucleus, where they can persist while their mitochondrial counterparts continue to rapidly evolve. Thus, numts represent 'molecular fossils' useful for comparison with mitochondrial variation, and are particularly suited for studies of the fast-evolving hypervariable segment of the mitochondrial control region (HV1). Here we used information from numts found in western gorillas (*Gorilla gorilla*) and eastern gorillas (*Gorilla beringei*) to estimate that these two species diverged about 1.3 million years ago (Ma), an estimate similar to recent calculations for the divergence of chimpanzee and bonobo. We also describe the sequence of a gorilla numt still possessing a segment lost from all contemporary gorilla mtDNAs. In contrast to that sequence, many numts of the HV1 are highly similar to authentic mitochondrial organellar sequences, making it difficult to determine whether purported mitochondrial sequences truly derive from that genome. We used all available organellar HV1 and corresponding numt sequences from gorillas in a phylogenetic analysis aimed at distinguishing these two types of sequences. Numts were found in several clades in the tree. This, in combination with the fact that only a limited amount of the extant variation in gorillas has been sampled, suggests that categorization of new sequences by the indirect means of phylogenetic comparison would be prone to uncertainty. We conclude that for taxa such as gorillas that contain numerous numts, direct approaches to the authentication of HV1 sequences, such as amplification strategies relying upon the circularity of the mtDNA molecule, remain necessary.

Keywords: divergence, gorilla, HV1, molecular fossil, numts

Received 9 July 2004; revision received 22 September 2004; accepted 22 September 2004

Introduction

Nuclear insertions of mitochondrial DNA (mtDNA) or 'numts' are commonly found in animal genomes. They are segments of mitochondrial DNA that have become translocated to the nuclear genome (Du Buy & Riley 1967; Lopez *et al*. 1994; Zischler *et al*. 1995; Perna & Kocher 1996; Zhang & Hewitt 1996; Bensasson *et al*. 2001; Ricchetti *et al*. 2004). A mitochondrial segment loses its original function upon arrival in the nuclear genome and so can readily acquire mutations, although the mutation rate is typically slower in the nucleus than in the mitochondria (Brown *et al*. 1982; Zischler *et al*. 1995; Mundy *et al*. 2000; Lü *et al*. 2002). Unfortunately, because recently integrated numts will tend to have a high sequence similarity to genuine organellar mtDNA sequences, inadvertent amplification of numts can be a nuisance in studies of mtDNA variation. In the worst case, unrecognized inclusion of numts in an analysis of mtDNA invalidates the analyses and can lead to false conclusions (van der Kuyl *et al*. 1995; Bensasson *et al*. 2001; Thalmann *et al*. 2004). Therefore, the unambiguous authentication of organellar sequences is a prerequisite for studies inferring evolutionary histories from mtDNA data.

Assuming, however, that one can reliably distinguish numts from authentic mtDNA, the numts themselves can be very informative. Examples of their uses include calibration of the relative rates of nuclear and mitochondrial sequence divergence (Lopez *et al*. 1997), identification of the extent and pattern of information transfer from the

Correspondence: O. Thalmann, Fax: +49-341-3550-299; E-mail: thalmann@eva.mpg.de
‡Current address: Genome Quebec Innovation Center, McGill University, 740 Aven. Dr Penfield, Montreal, Canada.

mitochondrion to the nucleus (Woischnik & Moraes 2002; Bensasson *et al*. 2003), and confirmation of the recent African ancestry of our own species (Zischler *et al*. 1995; Mishmar *et al*. 2004). Here we demonstrate the use of numts to investigate the population history of species, in particular for dating a molecular divergence between West and East African gorillas.

The western gorilla (*Gorilla gorilla*) and eastern gorilla (*Gorilla beringei*) are separated by more than 1000 km, exhibit behavioural and ecological differences (Doran & McNeilage 1998), and are classified as two separate gorilla species as a reflection of these facts and interpretation of morphological, particularly craniometric differences (Groves 2001; Taylor & Groves 2003). Genetic investigations have compared the phylogenetic relationship and degree of divergence found between the two gorilla species with that found between two other recently diverged equatorial African great ape species, the common chimpanzee (*Pan troglodytes*) and the bonobo (*Pan paniscus*). However, analyses of noncoding nuclear loci showed that while chimpanzees and bonobos were reciprocally monophyletic, monophyly was not apparent between western and eastern gorillas, perhaps reflecting a rather recent population split of ancestral gorillas (Kaessmann *et al*. 2001; Jensen-Seaman *et al*. 2003). Any attempts to date recent separation events using such nuclear data in which lineage sorting is not complete are most likely to provide a time estimate for the most recent common ancestor of the particular nuclear segment, but not of the actual population split itself (Edwards & Beerli 2000; Nichols 2001).

Hence, for dating of recent population splits, a genetic marker evolving at a sufficient rate to allow phylogenetic resolution at the level of interest is required. Because of its smaller effective population size and rapid evolution, mtDNA is particularly useful for analysis of the recent evolutionary history of animal populations (Moore 1995; Avise 2000). Analyses of mtDNA revealed reciprocal monophyly within both the *Pan* and the *Gorilla* genera and, as expected for loci completely linked on the same circular molecule, the results were consistent among several segments of the mitochondrial genome (COII, Ruvolo *et al*. 1994; NADH5, Jensen-Seaman *et al*. 2003; 12S, Thalmann unpublished; HV1, Garner & Ryder 1996; Jensen-Seaman 2000). For the two chimpanzee species, estimates based upon either mtDNA control region or COII sequences suggest a divergence time of about 2.5 million years ago (Ma) (Morin *et al*. 1994; Ruvolo 1997), similar to an estimate of 2.7–2.8 Ma based on NADH5 data (Jensen-Seaman 2000), while estimates using restriction enzyme analysis of the entire mitochondrial genome suggested 1.3 Ma for chimpanzee–bonobo divergence (Ferris *et al*. 1981). Similarly, different regions of the mitochondrial genome yielded different estimates of the divergence time between the two gorilla species, ranging from 2.7 Ma for NADH5 data (Jensen-Seaman 2000) and 2.2 Ma for the COII gene (Ruvolo 1997) to highly variable estimates (2.7–7.6 Ma) based upon HV1 data (Jensen-Seaman 2000).

The observed disparities between estimates from different regions in the mtDNA probably reflect high variance resulting from the small number of observed mutations. As a noncoding segment with the highest rate of evolution in the mtDNA genome, the HV1 of the control region provides ample polymorphism information, but rate and hence time estimations are complicated by the variable rates of mutation among nucleotide positions in this segment (Meyer *et al*. 1999). Here we use HV1 sequence information from contemporary gorillas in combination with data from 'molecular fossils' representing recently transposed numts to investigate the timing of a molecular divergence within gorillas. The comparison of the molecular estimates of the dates of divergences within gorillas and chimpanzees is of particular interest for two reasons. The first is the idea that a common biogeographical event in Central Africa affected speciation of some primates and perhaps other animals, and the second is the question of whether the taxonomic designation of the chimpanzee and gorilla species has been applied to lineages that have been independently evolving for similar lengths of time.

While humans and other great apes have been subject to particularly intensive investigation by means of HV1 sequencing in order to address questions such as the origin of modern humans (Vigilant *et al*. 1991) and the relative amounts and geographical pattern of genetic variation within ape species (Morin *et al*. 1994; Gagneux *et al*. 1999; Eriksson *et al*. 2004), it has been noted that for gorillas, there is an apparent tendency for numts to be amplified in addition to, or even instead of, authentic organellar mtDNA (Clifford *et al*. 2004a; Jensen-Seaman *et al*. 2004; Thalmann *et al*. 2004). Several approaches have been proposed to distinguish mitochondrial sequences from numts, including both practical measures to reduce chances of numt amplification and analytical methods to distinguish numts from authentic mtDNA (Bensasson *et al*. 2001 and references therein). We previously showed that successful analysis of authentic organellar HV1 sequences from one great ape — the gorilla — is unlikely without use of long-range PCR (polymerase chain reaction) to directly demonstrate the authenticity of particular HV1 sequences by identifying the same sequence from two overlapping long-range amplifications of mtDNA (Thalmann *et al*. 2004). Unfortunately, the necessary long-range amplification of segments thousands of nucleotides in length is currently not possible from the noninvasive samples obtainable from wild gorillas (Thalmann unpublished data). Therefore, very little reliable HV1 data useful for analysis of the population history of gorillas exist, although other researchers have recently attempted to apply indirect criteria for the authentification of gorilla HV1 sequences (Clifford *et al*. 2004a).

In this paper we present an analysis of multiple gorilla numts, including a sequence still bearing a segment deleted in current gorilla sequences, in comparison to authentic gorilla mtDNA. We derive an estimate of a molecular divergence time of western and eastern gorillas and compare it to recent estimates for the two chimpanzee species. Finally, we use all publicly available data to evaluate proposed schemes for validation of gorilla HV1 sequences and conclude that the plethora of numts means that direct experimental authentication of gorilla HV1 sequences is an unavoidable necessity.

Methods

Sampling methods and laboratory procedures are described in detail in Thalmann *et al*. (2004). In brief, in that study we asked whether, when using typical methods for mtDNA analysis from representatives of all of the great apes, numt sequences were amplified in addition or in preference to the authentic organellar DNA. We first determined the one necessarily authentic organellar HV1 sequence for each individual, including one western gorilla (*Gorilla gorilla gorilla*; Rok) and one eastern gorilla (*Gorilla beringei graueri*; Mukisi) by reamplification of the HV1 segment from two different overlapping long-range amplification products that together covered the entire mitochondrial genome. Then, with primers typically used for HV1 analysis of the gorillas, we directly amplified the HV1 segments from genomic DNAs, cloned the products, and sequenced multiple clones, resulting in identification of both sequences identical to the authentic HV1s and additional sequences defined as numts.

Here we use a subset of the numt sequences found in both gorilla species (Muk5 and Rok8 are herein termed N_1; Muk4 and Rok5 are N_2) to investigate the timing of divergence between the authentic gorilla mitochondrial sequences and the numt sequences. The results of a relative rate test (Tajima 1993), performed on those numts and the authentic mtDNA sequences suggested the application of one single substitution rate for all sequences, namely Tamura and Nei's modal rate of control region evolution 7.5×10^{-8} per site per year (Tamura & Nei 1993) and divergence time was estimated using the equation,

$$\delta = (2\lambda)T$$

in which δ is the sequence divergence between the sequences, λ is the substitution rate and T is the divergence time.

To infer phylogenetic relationships, we used all numt sequences ($N = 8$) obtained from the two gorillas studied, as well as the two gorilla mtDNA HV1 sequences we verified and additional unique sequences from the public database and designated there as either gorilla HV1 sequences ($N = 53$) or numts ($N = 16$). The total length of the sequences used in the phylogenetic analyses was 220 bp and omitted a variable length, cytosine-rich segment of approximately 20–30 bp. Omission of such segments is routine practice in HV1 analyses (Bendall & Sykes 1995) because, as a result of the artefacts generated by slippage in the PCR, it is difficult in practice to determine homopolymer segments with confidence and furthermore because it is substitutional differences that provide the necessary information for a phylogenetic analysis. We also excluded two sequences available in the database (accession numbers: NC 001645, D38114) because of too much missing data.

Phylogenetic analyses using maximum parsimony (MP) and neighbour joining (NJ) methods [applying different substitution models: Kimura 2-parameter (Kimura 1980) and the TN model (Tamura & Nei 1993)] as well as relative rate tests (Tajima 1993) were performed with the software MEGA2 (Kumar *et al*. 2001). Tree analyses using a quartet-based maximum likelihood method (ML) and applying the HKY substitution model (Hasegawa *et al*. 1985) were carried out with TREEPUZZLE 5.0 (Schmidt *et al*. 2002). We performed the analyses with and without the gamma distributed rate heterogeneity model by applying an alpha value that we obtained from TREEPUZZLE 5.0 and implemented into MEGA2 manually. To account for missing sequence information and small indels (insertion and deletions), we either completely deleted those positions or used the pairwise deletion option in MEGA2. We ran 1000 bootstrap or 1000 puzzling steps to assess the statistical support values for the observed branching patterns.

Results

Calculation of divergence time

To verify the authenticity of mitochondrial HV1 sequences obtained from one western and one eastern gorilla (Rokmt and Mukmt, respectively, in Thalmann *et al*. 2004), we used two overlapping long-range PCR products comprising the entire mitochondrial genome. Subsequent PCRs were performed on each amplification product targeting the HV1 in the overlap and revealed one single sequence for each individual. These authentic mitochondrial HV1 sequences were compared to sequences obtained from PCR amplifications of total genomic DNAs of the same individuals. A total of eight numts (NCBI acc.nb. Muk1: AJ812278; Muk4: AJ812279; Muk5: AJ812280; Muk7: AJ812281; Rok4: AJ812282; Rok5: AJ812283; Rok8: AJ812284; Rok11: AJ812285) were identified from the two gorillas analysed, four of which were exclusively found in either one or the other gorilla. There were two instances in which numt sequences from the western gorilla [Rok8 and Rok5 in (Thalmann *et al*. 2004)] were each highly similar (differing by a single substitution over 255 bp) to two numt sequences

from the eastern gorilla (Muk5 and Muk4), suggesting that the two different pairs of sequences (Rok8 and Muk5, now termed N_1; Rok5 and Muk4, now termed N_2) each represent single numts that became integrated at a time before the divergence of western and eastern gorillas. Whereas we attribute the single substitution between the two sequences each representing N_1 and N_2 to be allelic variation, N_1 and N_2 are highly dissimilar, differing by nine to 11 substitutions, and so probably represent two independent transposition events.

Although we are aware of the fact that numts typically show a slower rate of evolution compared to their mitochondrial counterparts, it is not totally clear how one would estimate the mutation rate for a particular numt and so we first assumed that the numts analysed here evolve similarly to mtDNA HV1 sequences. This assumption was supported by the results of relative rate tests (Tajima 1993). Phylogenetic analyses revealed in a topology in which the HV1 sequence derived from the western gorilla separates from the branch leading to the numts and the eastern gorilla HV1 sequence as depicted in the Fig. 1. We examined the constancy of evolutionary rates between the mtDNA sequence from the eastern gorilla in relation to the two numt sequences while using the mtDNA sequence obtained from the western gorilla as an outgroup (Fig. 1). The hypothesis of clock-like evolution could not be rejected for both tests: mtDNA eastern gorilla vs. N_1 ($\chi^2 = 1.06$, 1 df, $P = 0.303$) and mtDNA eastern gorilla vs. N_2 ($\chi^2 = 2.29$, 1 df, $P = 0.131$). Thus, it seems that although likely undergoing a different rate of evolution in the nucleus, the numts have not been in the nuclear genome sufficiently long enough to result in detectable effect in the relative rate test.

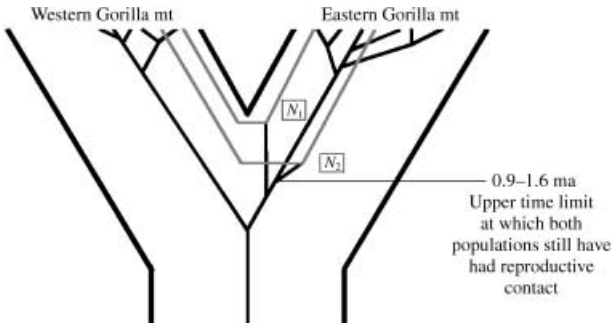

Fig. 1 Schematic diagram illustrating the phylogeny of the authentic gorilla mt HV1 sequences (western Gorilla mt; eastern Gorilla mt) and numt sequences (N_1, N_2), based upon maximum parsimony and neighbour-joining analyses of the substitutional differences between those sequences. The framing bold lines indicate the populations of western and eastern gorillas and bold branching lines show mtDNA lineages. The time that the numts spent in the nuclear genome is illustrated by the grey lines. The horizontal line indicates most recent estimated divergence time of the numt sequences and the mtDNA sequence of the eastern gorilla.

The times at which these sequences diverged from the current mitochondrial lineage of the eastern gorilla were estimated. Using the Tamura & Nei substitution model (Tamura & Nei 1993), the amount of sequence divergence (δ) between the eastern gorilla mtDNA sequence and N_1 and N_2 was calculated as 0.161 (SE. 0.029) and 0.136 (SE. 0.026), respectively, and so we arrived at sequence divergence times of 1.07 Ma and 0.91 Ma by applying a mutation rate of 7.5×10^{-8} per site per year (Tamura & Nei 1993). By comparison, 1.36 Ma was the estimated divergence time of the authentic mitochondrial sequences from the western and eastern gorillas. However, the above calculations do not take into account heterogeneity of mutation rates among sites, therefore a gamma distribution with an alpha value of 0.34 was incorporated to re-estimate the divergence times of the respective sequences. This alpha value was calculated using the two authentic and numt sequences and is comparable to alpha values determined using much larger datasets of other hominoid HV1 sequences (Excoffier & Yang 1999). This resulted in dates of 1.68 Ma and 1.3 Ma for the splits between the authentic HV1 from the eastern gorilla and N_1 and N_2, respectively. Whereas these alpha-corrected divergence times represented increases of 57.0% and 42.8% for N_1 and N_2, respectively, it is noteworthy that the value estimated for the authentic mtDNAs increased by 77.4% (to 2.41 Ma), potentially indicating different evolutionary rates between numts and mitochondrial sequences.

Phylogenetic analyses

One phylogenetic analysis focused upon one of the four numts identified in the analysis of sequenced clones from amplifications of genomic DNA from a western gorilla. This numt, termed Rok11 in Thalmann *et al.* (2004) (NCBI acc.nb.AJ812285), is notable for possessing an approximately 100 bp segment absent from contemporary gorilla HV1 sequences (Foran *et al.* 1988; Xu & Arnason 1996; Thalmann *et al.* 2004). Tree reconstructions using exclusively that 100 bp segment, which is readily alignable to homologous regions from all other great apes, were used to investigate the relationship of this segment to HV1 segments from other hominoids. The analyses yielded the topology expected for hominoid evolution, with the gorilla numt sequence located between the sequences from orangutans and the branch leading to the chimpanzees and humans (Ruvolo *et al.* 1994; Glazko & Nei 2003) (Fig. 2). Note that the length of the sequence analysed is very short, resulting in low statistical support values for some branches.

Because it contained a segment absent from all contemporary authentic gorilla HV1 sequences, the Rok11 sequence could easily be recognized as a nonorganellar, nuclear copy of a gorilla mtDNA representing a molecular fossil. However, this was an exceptionally clear case, and we were interested in assessing whether it is possible to

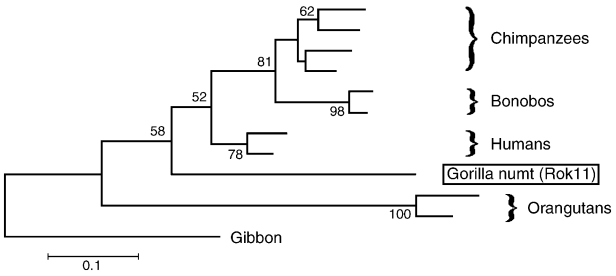

Fig. 2 The phylogenetic arrangement of sequences of an approximately 100 bp (reduced to 91 bp after deletion of gaps) segment of the mitochondrial HV1 from hominoids, which is absent in contemporary gorilla mtDNAs, produced using neighbour-joining tree analysis with the Kimura 2-parameter substitution model. Nodes are labelled with the bootstrap values after 1000 bootstrap steps, values below 50 were omitted. Application of different substitution models or tree building methods revealed similar topologies with variation in the statistical support likely attributable to the fact that the segment analysed here is only 91 bp in length. The different taxa are hominoid sequences taken from the NCBI database with the following accession numbers: humans — AH001262, AJ586554; chimpanzees — AJ586556, AF176732, AJ586557, AF290608; bonobos — AJ586555, AF176756; orangutans — AJ586559, AJ586558 and as an outgroup one gibbon sequence — X99256, and the box indicates the sequence of the ancient gorilla numt.

confidently distinguish other numt sequences from authentic gorilla HV1 sequences by means of phylogenetic analyses. For this purpose, we reanalysed all complete, publicly available sequences derived from studies of gorilla HV1. This data set of 79 sequences thus included proven and putative authentic gorilla HV1 sequences as well as designated numts of gorilla HV1 (see Thalmann *et al*. 2004 and database annotations). Regardless of the tree-building method used (NJ, ML or MP), the use of different substitution models, and of either pairwise or complete deletion of missing sequence information, all approaches yielded a similar topology (Fig. 3). The numt sequence containing the segment missing from contemporary gorilla HV1 is placed outside of all other sequences, which fall into two groupings. Group 1 contains sequences derived from eastern gorillas, whereas group 2 is comprised of sequences derived from western and eastern gorillas. Within group 1 is the only sequence directly demonstrated to represent an authentic organellar eastern gorilla HV1

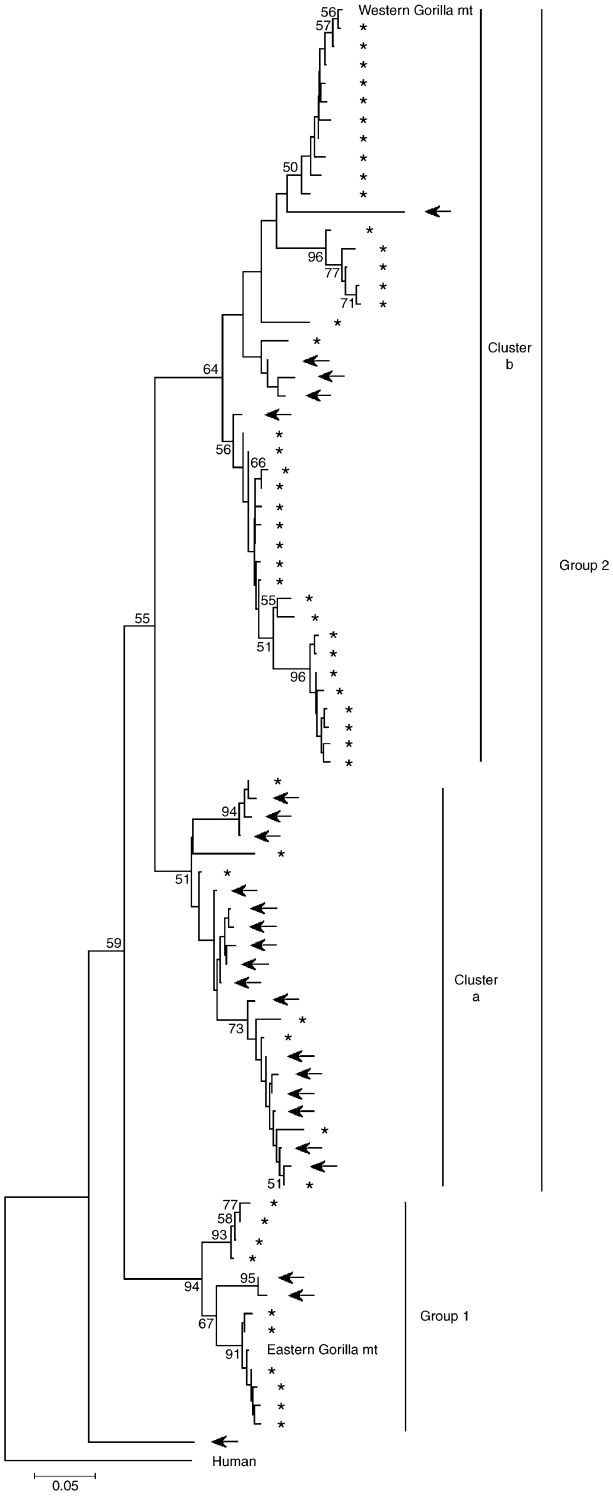

Fig. 3 A neighbour-joining tree produced using 79 gorilla sequences and a human sequence as an outgroup, applying the Tamura–Nei substitution model as well as the assumption of rate heterogeneity among sites (alpha = 0.83) and pairwise deletion of indels. Despite slight differences in the statistical support values, the topology did not change after applying other tree-building methods or changing the assumed substitution model. Only bootstrap support values of 50 or greater are shown. For reasons of readability we used symbols instead of the exact sequence IDs, but the annotated tree is available as supplementary information. The two verified authentic mtDNA sequences are labelled with the respective names (western Gorilla mt and eastern Gorilla mt), whereas other putatively authentic HV1 sequences are indicated by a star. Numt sequences identified in our previous study or designated as such in the GenBank database are marked by arrows.

sequence, as well as two numt sequences and other putative HV1 sequences from eastern gorillas. In group 2, the sequences are arranged in at least two additional clusters with modest bootstrap support (51 and 64, respectively). Cluster 'a' contains putative HV1 sequences from western gorillas and numts derived from both western and eastern gorillas. Within cluster 'b' is the verified HV1 sequence of a western gorilla, a second, very similar authentic western gorilla HV1 sequence (Xu & Arnason 1996), and additional numt and putative HV1 sequences. The nearness of numt and authentic mtDNA sequences in both clusters highlights the complex relationship of putative authentic organellar and numt sequences, as highly similar pairs of putative HV1 and numt sequences are found (Clifford *et al*. 2004a; Thalmann *et al*. 2004). Overall, it does not appear that the gorilla numt sequences fall into discrete, well-supported groups easily distinguishable from the authentic or putatively authentic gorilla HV1 sequences.

Discussion

Numts for dating

We have demonstrated at least two useful features of numts in this study. The first is the preservation of ancestral characters, which provides us with knowledge unobtainable from contemporary samples. Namely, a single gorilla numt contains a portion of the HV1 that is absent from contemporary gorilla sequences (Foran *et al*. 1988; Xu & Arnason 1996; Thalmann *et al*. 2004). This 'missing link' in the evolution of gorilla mitochondrial DNA shows a form of gorilla mtDNA that existed after the divergence from the common mitochondrial ancestor of gorillas, humans, and chimpanzees and prior to the most recent common ancestor of contemporary gorilla mtDNAs.

The second use of numts demonstrated here is in dating of molecular divergences in order to provide insights into the evolutionary history of species (Lopez *et al*. 1997). As was pointed out by Bensasson *et al*. (2001), the estimated divergence time between a numt and contemporary mtDNA sequences does not necessarily represent the time of transposition of the numt into the nuclear genome. This is because after its divergence from the lineage that will evolve to become the contemporary mtDNA sequences, the sequence that will become a numt might have evolved for an additional time as mitochondrial DNA before the actual transposition. However, a particular numt that appears in representatives of what are now two populations must have been integrated into the nucleus at a time when the two populations were still one. We used two such numts (N_1 and N_2) to estimate that western and eastern gorillas diverged no earlier than 0.91–1.68 Ma. Because the common ancestor of the DNA sequences under consideration is always older than the actual common ancestor of the two populations (Nei 1987; Nichols 2001), this time estimate represents a maximal time depth of the split between *Gorilla gorilla* and *Gorilla beringei*.

Compared to other calculations of the divergence time of western and eastern gorillas, such as the date of 2.2 Ma by Ruvolo (1997) [recalculated by Jensen-Seaman (2000) as 1.89 Ma] that were based upon mitochondrial COII data that did not show any indication of the inadvertent inclusion of numts, our estimates suggest a more recent divergence, and are similar to the estimated age of the most recent common ancestor of gorilla noncoding X-chromosomal sequences (mean: 1.28 Ma; Kaessmann *et al*. 2001). The application of information from single genetic loci to infer divergence times usually results in an estimation of an upper time limit of a population split. This is because polymorphism in the ancestral population will lead to different loci having different histories, as well as to genetic divergences that precede population divergences (Nei 1987; Moore 1995; Edwards & Beerli 2000). Thus, amalgamation of information derived from multiple loci would be the best approach towards a more precise estimate of divergence times (Moore 1995; Nichols 2001). First attempts to use analysis of multiple nuclear loci in gorillas to estimate relative diversity as compared to chimpanzees and humans have yielded encouraging results but included few individuals (Jensen-Seaman *et al*. 2003) or only samples from western gorillas (Yu *et al*. 2004). Future investigations should concentrate on such markers and, to obtain a comprehensive view into the evolutionary history of gorillas, should include sampling of individuals of known origin from as many localities as possible.

Consideration of the apparent concordance of the divergence dates within *Pan* (1.8–0.8 Ma: Yu *et al*. 2003; Fischer *et al*. 2004) and within *Gorilla* (1.68–0.91 Ma) suggests that common biogeographical events may have affected both genera. One possible scenario is that the effects of large climatic and geological changes leading into the formation of the Great Rift Valley some 1.5 Ma (Beadle 1981) simultaneously separated the two gorilla species and the two chimpanzee species, a scenario that has also been suggested by Jensen-Seaman (2000).

Validity of gorilla HV1 sequences

The second part of our study focused not upon the useful attributes of numts, but rather on their potential to confuse phylogenetic analyses by masquerading as authentic organellar sequences. Our results are consistent with those we obtained earlier with a smaller data set (Thalmann *et al*. 2004), and showed that numt sequences were scattered throughout the phylogenetic tree composed of both putative and authenticated gorilla HV1 and numt sequences. This lack of a neat separation of numt sequences into a divergent clade implies that for gorillas, phylogenetic

analyses cannot be relied upon to distinguish numts from authentic sequences (see also Jensen-Seaman *et al*. 2004). This is disappointing, because some studies in other taxa have shown instances in which detected numts reliably fall outside the monophyletic arrangement of authentic mitochondrial sequences and can therefore be easily distinguished (Zischler *et al*. 1995; Sorenson & Fleischer 1996). The high similarity of some gorilla numt sequences to authentic gorilla HV1 sequences suggests that some of the numts are of relatively recent origin. This is consistent with conclusions from analyses of the complete human genome, which have shown that the process of transposition of information from the mitochondrion to the nucleus is an ongoing process (Mourier *et al*. 2001; Bensasson *et al*. 2003; Ricchetti *et al*. 2004). However, it is not yet understood why the prevalence of detected numts appears to vary among taxa and be particularly high in gorillas (Bensasson *et al*. 2001; Jensen-Seaman *et al*. 2004; Thalmann *et al*. 2004).

Thus we conclude, as we did earlier based upon analyses of a smaller dataset, that reliable analyses of gorilla HV1 sequences require extraordinary technical measures to ensure validity (Thalmann *et al*. 2004). This contradicts conclusions reached by Clifford and coworkers, who recently presented a phylogeographical analysis of gorillas based upon HV1 sequences (Clifford *et al*. 2004a). The authors relied upon examination of particular nucleotide positions in the sequences as well as phylogenetic analyses as means to distinguish organellar from nuclear-derived sequences. It is worth examining their approach in some detail to see if it might reliably facilitate HV1 analyses from gorillas.

Clifford and colleagues assume that numt sequences should share certain characteristic sequence features and group together in the tree analysis (Clifford *et al*. 2004a). Three key elements in their classification of sequences as authentic organellar HV1 or numts were six single nucleotide positions, a cytosine-rich segment approximately 20–30 bp in length, and the phylogenetic pattern. It appears that an initial phylogenetic analysis of putative organellar and numt sequences was conducted to form sets of sequences with defined variation at these six positions and in the C-rich segment. With regard to the six nucleotide positions, a potential difficulty is that the HV1 is a non-coding, rapidly evolving DNA segment and mutations can occur repeatedly at positions (Meyer *et al*. 1999). In fact, just a single transition, the most common type of substitution, at one of the six diagnostic sites proposed would be sufficient to transform putative authentic HV1 sequences (their category 'Haplogroup D') into a numt ('Numt Class I'), and vice versa.

The second element of the classification scheme used by Clifford and colleagues to distinguish authentic organellar from numt sequences was qualitative assessment of a stretch containing mostly cytosine residues and exhibiting length variation among individuals. A difficulty arising in the application of this criterion is that templates with homopolymer segments tend to pose problems for the polymerase during the PCR, causing sequencing reactions to stutter. As a result of this fact and the occurrence of more than one endogenous sequence in the template, direct sequencing of the gorilla HV1 yields ambiguous, unreadable sequences in most cases (Garner & Ryder 1996; B. Bradley, H. Siedel personal communication). Cloning of PCR products allows determination of the sequences of individual molecules and reveals that in products from the same amplification, different variants can be found within the polycytosine stretch. This is illustrated in Fig. 4, which shows for one individual an alignment of the polycytosine stretch determined from single clone sequences obtained from two reamplifications each of two different long-range PCR

Fig. 4 An alignment of 25 different clones showing the cytosine-rich region (with the C-stretch itself highlighted) and the surrounding nucleotide positions as sequenced from one western gorilla. The clones derive from four independent amplifications, as indicated by the letters A–D in the clone names.

products. In addition to substititutional changes, length variants are also apparent, but convention dictates that such changes be excluded from phylogenetic analyses because they arise through different mutational processes than do substitutions (Saitou & Ueda 1994; Zhang & Gerstein 2003). These concerns notwithstanding, we checked whether the eight numt sequences derived in our previous study (Thalmann *et al*. 2004) presented C-stretch segments consistent with the proposed scheme (Clifford *et al*. 2004a). We found that whereas it would be possible to distinguish the authentic mitochondrial C-stretch of the western gorilla from those of the numt sequences, the authentic mtDNA segment obtained from the eastern gorilla showed only two transitional changes (C–T) compared to one numt sequence (Muk1), making a distinction based on this criterion highly questionable.

The third means used by Clifford and coworkers to distinguish numt and authentic organellar sequences was identification by phylogenetic analysis of clades composed exclusively of numts. The phylogenetic analyses presented here and previously (Fig. 4 in Thalmann *et al*. 2004) suggest that the tree topologies vary depending upon the set of sequences included, and that discrete, well-supported clades containing all numts are not apparent. In fact, Clifford and coworkers recently found it necessary to postulate a third numt clade to account for numts not present in their initial analysis (Clifford *et al*. 2004b).

Finally, it is important to note additional key assumptions underlying their proposed classification scheme (Clifford *et al*. 2004a). One is that representative sets of authentic organellar HV1 and numt sequences have been reliably identified. But the sequences used have been deemed authentic organellar representatives or numts based on *ad hoc* criteria involving phylogenetic analyses, and for a few cases, the assessment of several clone sequences, rather than direct determination of necessarily authentic HV1 sequences. This means that it is not impossible that some have been misclassified. Even assuming that the authentic HV1 sequences have been correctly identified, as the sequences used as the basis of the scheme come from gorillas of captive origin, it is uncertain how well the extent of gorilla HV1 variation in the wild has been sampled. Evidence suggests that integration of mitochondrial sequences into the nucleus is an ongoing process (Mourier *et al*. 2001; Bensasson *et al*. 2003; Ricchetti *et al*. 2004), making recognition of recent numts particularly challenging and suggesting that it would be difficult to determine when all numts present in gorillas have been identified.

In conclusion, the lack of criteria of demonstrated reliability for distinguishing authentic organellar HV1 and numt sequences means that use of the mtDNA molecule to gain insights into the population history of the gorillas remains limited to cases in which high-quality DNA derived from blood or tissue can be subjected to molecular analytical methods for determination of necessarily authentic mtDNA sequences (Thalmann *et al*. 2004). Numt sequences remain a nuisance for studies of mtDNA variation, but as sequencing of entire genomes of various species becomes commonplace, we can look forward to the identification of additional numts appropriate for population genetic analysis, including the identification of numts polymorphic for presence and absence within taxa and so useful for the understanding of intraspecific genetic variation (Ricchetti *et al*. 2004).

Acknowledgements

We thank C. Boesch, B. Bradley, S. Pääbo, H. Siedel and M. Stoneking for helpful discussions and fruitful comments on the manuscript. The study was financed by the Max Planck Society and the Deutsche Forschungsgemeinschaft (VI 229/2–1).

Supplementary material

The following material is available from
http://www.blackwellpublishing.com/products/journals/suppmat/MEC/MEC2382/MEC2382sm.htm

Figure S1. The annotated version of the phylogenetic reconstruction presented in Figure 3.

References

Avise JC (2000) *Phylogeograph*. Harvard University Press, Cambridge, MA.

Beadle LC (1981) *The Inland Waters of Tropical Africa*. Longman, London.

Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *American Journal of Human Genetics*, **57**, 248–256.

Bensasson D, Feldman MW, Petrov DA (2003) Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *Journal of Molecular Evolution*, **57**, 343–354.

Bensasson D, Zhang D-X, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology and Evolution*, **16**, 314–321.

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution*, **18**, 225–239.

Clifford SL, Anthony NM, Bawe-Johnson M *et al*. (2004a) Mitochondrial DNA phylogeography of western lowland gorillas (*Gorilla gorilla gorilla*). *Molecular Ecology*, **13**, 1551–1565.

Clifford SL, Anthony NM, Bawe-Johnson M *et al*. (2004b) Addendum. *Molecular Ecology*, **13**, 1567–1567.

Doran DM, McNeilage A (1998) Gorilla ecology and behaviour. *Evolutionary Anthropology*, **6**, 120–131.

Du Buy HG, Riley FL (1967) Hybridization between the nuclear and kinetoplast DNAs of *Leishmania enriettii* and between nuclear and mitochondrial DNAs of mouse liver. *Proceedings of the National Academy of Sciences USA*, **57**, 790–797.

Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839–1854.

Eriksson J, Hohmann G, Boesch C, Vigilant L (in press) Rivers influence the population genetic structure of bonobos. *Molecular Ecology* **13**, 3425–3435.

Excoffier L, Yang Z (1999) Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Molecular Biology and Evolution*, **16**, 1357–1368.

Ferris SD, Wilson AC, Brown WM (1981) Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proceedings of the National Academy of Sciences USA*, **78**, 2432–2436.

Fischer A, Wiebe V, Pääbo S, Przeworski M (2004) Evidence for a complex demographic history of chimpanzees. *Molecular Biology and Evolution*, **21**, 799–808.

Foran D, Hixson J, Brown W (1988) Comparisons of ape and human sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis. *Nucleic Acids Research*, **16**, 5841–5861.

Gagneux P, Wills C, Gerloff U *et al.* (1999) Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proceedings of the National Academy of Sciences USA*, **96**, 5077–5082.

Garner KJ, Ryder OA (1996) Mitochondrial DNA diversity in gorillas. *Molecular Phylogenetics and Evolution*, **6**, 39–48.

Glazko GV, Nei M (2003) Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution*, **20**, 424–434.

Groves CP (2001) *Primate Taxonomy*. Smithsonian Institution Press, Washington.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Jensen-Seaman M (2000) *Evolutionary Genetics of Gorillas*, PhD Dissertation, Yale University.

Jensen-Seaman MI, Deinard AS, Kidd KK (2003) Mitochondrial and nuclear DNA estimates of divergence between western and eastern gorillas. In: *Gorilla Biology: a Multidisciplinary Perspective* (eds Taylor AB, Goldsmith ML), pp. 247–268. Cambridge University Press, Cambridge.

Jensen-Seaman MI, Sarmiento EE, Deinard AS, Kidd KK (2004) Nuclear integrations of mitochondrial DNA in gorillas. *American Journal of Primatology*, **63**, 139–147.

Kaessmann H, Wiebe V, Weiss G, Pääbo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genetics*, **27**, 155–156.

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 11–120.

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, **17**, 1244–1245.

van der Kuyl AC, Kuiken CL, Dekker JT, Perizonius WRK, Goudsmit J (1995) Nuclear counterparts of the cytoplasmic mitochondrial 12S rRNA gene: a problem of ancient DNA and molecular phylogenies. *Journal of Molecular Evolution*, **40**, 652–657.

Lopez J, Culver M, Stephens J, Johnson W, O'Brien S (1997) Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. *Molecular Biology and Evolution*, **14**, 277–286.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, **39**, 174–190.

Lü X-M, Fu Y-X, Zhang Y-P (2002) Evolution of mitochondrial cytochrome b pseudogene in genus *Nycticebus*. *Molecular Biology and Evolution*, **19**, 2337–2341.

Meyer S, Weiss G, von Haeseler A (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics*, **152**, 1103–1110.

Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. *Human Mutation*, **23**, 125–133.

Moore WS (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution*, **49**, 718–726.

Morin PA, Moore JJ, Chakraborty R *et al.* (1994) Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, **265**, 1193–1201.

Mourier T, Hansen AJ, Willerslev E, Arctander P (2001) The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Molecular Biology and Evolution*, **18**, 1833–1837.

Mundy NI, Pissinatti A, Woodruff DS (2000) Multiple nuclear insertions of mitochondrial cytochrome *b* sequences in callitrichine primates. *Molecular Biology and Evolution*, **17**, 1075–1080.

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

Nichols R (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution*, **16**, 358–364.

Perna NT, Kocher TD (1996) Mitochondrial DNA: molecular fossils in the nucleus. *Current Biology*, **6**, 128–129.

Ricchetti M, Tekaia F, Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS Biology*, **2**, e273.

Ruvolo M (1997) Genetic diversity in hominoid primates. *Annual Reviews of Anthropology*, **26**, 515–540.

Ruvolo M, Pan D, Zehr S *et al.* (1994) Gene trees and hominoid phylogeny. *Proceedings of the National Academy of Sciences USA*, **91**, 8900–8904.

Saitou N, Ueda S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Molecular Biology and Evolution*, **11**, 504–512.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREEPUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Sorenson MD, Fleischer RC (1996) Multiple independent transpositions of mitochondrial DNA control region sequences to the nucleus. *Proceedings of the National Academy of Sciences USA*, **93**, 15239–15243.

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, **135**, 599–607.

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, **10**, 512–526.

Taylor AB, Groves CP (2003) Patterns of mandibular variation in *Pan* and *Gorilla* and implications for African ape taxonomy. *Journal of Human Evolution*, **44**, 529–561.

Thalmann O, Hebler J, Poinar HN, Pääbo S, Vigilant L (2004) Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. *Molecular Ecology*, **13**, 321–335.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science*, **253**, 1503–1507.

Woischnik M, Moraes CT (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Research*, **12**, 885–893.

Xu X, Arnason U (1996) A complete sequence of the mitochondrial genome of the western lowland gorilla. *Molecular Biology and Evolution*, **13**, 691–698.

Yu N, Jensen-Seaman MI, Chemnick L *et al.* (2003) Low nucleotide diversity in chimpanzees and bonobos. *Genetics*, **164**, 1511–1518.

Yu N, Jensen-Seaman MI, Chemnick L, Ryder O, Li W-H (2004) Nucleotide diversity in gorillas. *Genetics*, **166**, 1375–1383.

Zhang Z, Gerstein M (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research*, **31**, 5338–5348.

Zhang D-X, Hewitt GM (1996) Nuclear integrations: challenges for mitochodrial DNA markers. *Trends in Ecology and Evolution*, **11**, 247–251.

Zischler H, Geisert H, von Haeseler A, Pääbo S (1995) A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature*, **378**, 489–492.

This work is part of Olaf Thalmann's dissertation research on the patterns of genetic variation in gorillas. Linda Vigilant is a research scientist leading the genetics group in the Primatology department of the Max Planck Institute for Evolutionary Anthropology, in which Dieter Lukas and Jonas Eriksson also conduct their dissertation research. David Serre and Michael Hofreiter are research scientists in the Evolutionary Genetics department of the MPI. The authors share an interest in understanding the evolutionary history of all great apes including our own species.
