Published in final edited form as:
Mol Phylogenet Evol. 2014 December ; 0: 61–70. doi:10.1016/j.ympev.2014.08.018.

Identification of species-specific nuclear insertions of mitochondrial DNA (numts) in gorillas and their potential as population genetic markers

Iván Darío Soto-Calderón[a,b,1], **Jonathan Clark Nicholas**[a], **Vera Halo Wildschutte Julia**[c], **Kelly DiMattio**[c], **Michael Ignatius Jensen-Seaman**[c], and **Nicola Mary Anthony**[a]

[a]Department of Biological Sciences, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA

[b]Genética, Mejoramiento y Modelación Animal (GaMMA), University of Antioquia, AA1226, Medellín, Colombia

[c]Department of Biological Sciences, Duquesne University, 600 Forbes Ave., Pittsburgh, PA 15282, USA

Abstract

The first hyper-variable region (HV1) of the mitochondrial control region (MCR) has been widely used as a molecular tool in population genetics, but inadvertent amplification of nuclear translocated copies of mitochondrial DNA (numts) in gorillas has compromised the use of mitochondrial DNA in population genetic studies. At least three putative classes (I, II, III) of gorilla-specific HV1 MCR numts have been uncovered over the past decade. However, the number, size and location of numt loci in gorillas and other apes are completely unknown. Furthermore, little work to date has assessed the utility of numts as candidate population genetic markers. In the present study, we screened Bacterial Artificial Chromosome (BAC) genomic libraries in the chimpanzee and gorilla to compare patterns of mitochondrial-wide insertion in both taxa. We conducted an intensive BLAST search for numts in the gorilla genome and compared the prevalence of numt loci originating from the MCR with other great ape taxa. Additional gorilla-specific MCR numts were retrieved either through BAC library screens or using an anchored-PCR (A-PCR) amplification using genomic DNA from five unrelated gorillas. Locus-specific primers were designed to identify numt insertional polymorphisms and evaluate their potential as population genetic markers. Mitochondrial-wide surveys of chimpanzee and gorilla BACs showed that the number of numts does not differ between these two taxa. However, MCR numts are more abundant in chimpanzees than in other great apes. We identified and mapped 67 putative gorilla-specific numts, including two that contain the entire HV1 domain, cluster with sequences from

Correspondence to: Iván Darío Soto-Calderón.

[1]Present Address: Sede de Investigación Universitaria. Torre 2. Laboratorio 430. University of Antioquia. Carrera 53 #61-30. Medellín, Antioquia, COLOMBIA. ivandariosoto@hotmail.com Phone number: (+57) 313-7091801.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

two numt classes (I, IIb) and will likely co-amplify with mitochondrial sequences using most published HV1 primers. However, phylogenetic analysis coupled with *post-hoc* analysis of mitochondrial variation can successfully differentiate nuclear sequences. Insertional polymorphisms were evident in three out of five numts examined, indicating their potential utility as molecular markers. Taken together, these findings demonstrate the potentially powerful insight that numts could make in uncovering population history in gorillas and other mammals.

Keywords

Mitochondrial DNA; numt; polymorphism; gorilla; great ape; BAC screens

1. Introduction

For decades, mitochondrial DNA (mtDNA) has been one of the most popular markers in population genetics and systematics. In particular, the mitochondrial control region (MCR) has proven to be especially useful at uncovering population structure and resolving intra-specific units for conservation. However, the unintentional amplification of nuclear integrations of mtDNA (numts) can inflate estimates of genetic diversity (see Garner and Ryder, 1996) and lead to erroneous phylogenetic inference (Song et al., 2008). Assembly of available genomic databases from a handful of eukaryotes has shown that the number of numts varies among taxa (Richly and Leister, 2004; Triant and DeWoody, 2007; Hazkani-Covo, et al., 2010). However, the abundance of numts in many species remains poorly understood, despite the increasing number of whole genome sequences becoming available.

It has been hypothesized that nowhere are MCR numts more abundant than in the genome of gorillas (Calvignac et al., 2011; Jensen-Seaman et al., 2004), leading some to question the reliability of mitochondrial data in this genus (Thalmann et al., 2004; 2005). Since many numts are small in size (< 500b), their inadvertent amplification might be avoided through the use of long-range PCR (Thalmann et al., 2004). However, this approach is not always feasible as degraded samples from feces or museum specimens are frequently the only source of DNA for genetic studies of wild primates such as gorillas. Furthermore, amplification of *in vitro* recombinants has also been a risk in gorillas that if ignored could artificially inflate estimates of mitochondrial diversity and mislead phylogenetic analysis (Anthony et al., 2007a). Despite these empirical evidences and an observed difference in the prevalence of MCR numts across great ape taxa (Soto-Calderón et al., 2012) the abundance of MCR numts in the gorilla genome remains to be assessed. Fortunately, the increasing availability of genomic resources for many taxa now presents new opportunities to characterize numt loci and rigorously assess patterns of abundance across closely related taxa.

Previous phylogeographic studies of western (*Gorilla gorilla*) and eastern (*G. beringei*) gorilla species have thus far identified four (A – D) geographically restricted MCR haplogroups (Anthony et al., 2007b) and three corresponding numt classes (I – III). To date, class I numts have only been found in western gorillas whereas class II and class III numts have been reported in both eastern lowland and western gorilla populations (Anthony et al., 2007a). These findings contrast with the reciprocal monophyly observed between eastern and western gorilla mitochondrial clades and strongly suggest a history of recent hybridization between eastern and western gorillas (Ackermann and Bishop, 2009; Thalmann et al., 2007).

Although numts have the potential to confound mtDNA analyses, these loci can also be of considerable use as population genetic or cladistics markers. Like transposable elements (TE), numts are considered homoplasy-free markers since they are rarely excised from the genome, allowing the ancestral (absence) and derived (presence) states to be inferred (Batzer and Deininger, 2002). Insertional polymorphisms in numts and TEs have been effectively used to infer historical demographic processes in humans (Batzer et al., 1994; Batzer and Deininger, 2002; Lang et al., 2011; Perna et al., 1992; Thomas et al., 1996) and in the case of TE loci, reconstruct comprehensive primate phylogenies (Herke et al., 2007; Ray et al., 2005). Yet no studies to date have systematically evaluated the utility of numts as population genetic markers in non-human primates, despite the many advantages inherent to such a marker system (see Batzer and Deininger, 2002; Herke et al., 2007; Stewart et al., 2011).

Here we screened genomic Bacterial Artificial Chromosome (BAC) libraries of both gorilla and chimpanzee with probes from across the mitochondrial genome in order to compare the abundance of mitochondrial fragments in the nuclear genome of two of the most commonly studied great apes. We also complemented this strategy with BLAST surveys of the gorilla genome using the entire mitochondrial genome as a query sequence. Secondly, we conducted genomic BLAST searches of the current and previous gorilla scaffolds using the entire mitochondrial genome as a query sequence. Thirdly, we carried out wet lab screens of a gorilla BAC library in combination with an anchored PCR (A-PCR) survey of nuclear-enriched genomic DNA from five cell lines to capture additional MCR numt loci that may be absent in the gorilla scaffold and the BAC library. The use of complementary strategies is especially important as recent integrations may have not yet gone to fixation. These data were then used to: (1) compare the abundance of mitochondrial-wide numts in gorillas and chimpanzee genomes; (2) compare the prevalence of MCR numts in gorillas and other great ape genomes; (3) provide the first comprehensive list and map of gorilla numts; (4) determine whether gorilla MCR numts uncovered in the present study fall into any of the previously described numt classes (I–III); (5) determine the risk of amplifying gorilla numts with traditional mitochondrial primers and; (6) assess whether polymorphic numt loci can be used to infer gorilla evolutionary history within a phylogenetic and population genetic framework.

2. Methods

2.1. Gorilla and chimpanzee BAC library screens

In order to design probes encompassing the entire mtDNA genome, whole mitochondrial sequences from the chimpanzee (NC_001643) and western lowland gorilla (NC_001645) were aligned with ClustalW (Thompson et al., 2002). A custom *perl* script was then used to identify all 40-mer sequences with at least 90% identity between species and 45–55% GC content. From these, sixty-four 40-mer dsDNA probes were chosen from locations spaced at intervals of approximately 250bp around the ~16kb genome (Table S1), ^{32}P-radiolabeled as previously described (Ross et al., 1999), and purified with G-50 Sephadex columns (GE Healthcare, Piscataway, NJ, USA). Probes were pooled into 17 groups of adjacent loci, along with a control probe derived from *Caenorhabditis briggsae*, and hybridized overnight at 61°C in Express Hyb hybridization solution (Clontech, Mountain View, CA, USA) to an arrayed genomic BAC library (CHORI-255) derived from the male western gorilla “Frank” and to the male chimpanzee “Clint” library (CHORI-251) (Children’s Hospital Oakland Research Institute). Library filters were washed with 1XSSC, 0.1% SDS at 61°C as per Ross et al. (1999), and exposed to autoradiographic film for 48 to 96 hrs.

To identify gorilla BAC clones specifically containing MCR numts, nine different 40-mer dsDNA probes were designed from the gorilla MCR and flanking tRNA sequences using paired 24-mer ssDNA oligos with an 8bp overlap (Table S1). ^{32}P-radiolabeled overgo hybridization probes were generated from these oligos and hybridized to the gorilla genomic library (CHORI-255) as above. Thirty-two positive BACs were subsequently grown up in 500ml Luria Broth cultures and purified with NucleoBond BAC maxi-prep kits (Clontech). The BAC-ends were sequenced with the primers SP6 and T7 (Table S2), using the BigDye v3.1 kit (Applied Biosystems, Foster City, CA, USA), following manufacturer’s instructions but modified to include 75 PCR cycles. Sequences were mapped via BLAT to the human genome sequence (build hg19) (Kent, 2002; Kent et al., 2002). BACs that overlapped with a known human numt were not further characterized. BACs containing candidate loci likely to contain a gorilla-specific numt were sequenced using a combination of internal overgo oligos and mtDNA primers used to “walk out” to the corresponding nuclear flanks.

2.2. BLAST surveys of the gorilla genome using the mitochondrial genome as a query

We conducted a BLASTn search (Altschul et al., 1990) of the current partially assembled version of the reference genome of the female western gorilla Kamilah (*Gorilla gorilla gorilla*, build gorGor3.1 May 2011), using the entire gorilla mitochondrial genome (X93347) as a query sequence. The filters and mask options of BLAST searches were clicked off, word size was relaxed to a value of 7 and, match/mismatch scores and gap creation/extension penalties were set to 1/−1 and 3/1, respectively. All putative numts with expect-values of 0.39 or less and an overall identity of 60% or more were retained for future analysis. Flanking nuclear sequences from each candidate numt were mapped to the reference genome of humans (build 36.3) and chimpanzees (build 2.1) using the BLAT tool (Kent, 2002). A numt was considered to be gorilla-specific when its sequence coincided with an integration observed only in gorillas. In some cases, numt duplications in the gorilla genome were diagnosed when found nested within a duplicated larger chromosomal region. Additional searches of the original trace files (Gorilla_gorilla_WGS) and the former genomic assembly CABD00000000.2 were also conducted in order to recover additional MCR numts not yet assembled in the current genome version. Mapping and determination of gorilla-specificity of each candidate numt was carried out as described above.

2.3. Relative abundance of MCR numts in gorillas and other great apes

The prevalence of numts specifically derived from each of the four MCR sub-domains (the hyper-variable 1 [HV1], the Central Conserved Domain [CCD], the hyper-variable 2 [HV2] and the sub-domain proximal to the phenylalanine tRNA gene [MCR_F]) and the 500bp flanking regions in gorillas was compared with findings from other great apes (humans, chimpanzees and orangutans). In order to make data derived from the genomic survey comparable with previous studies, the database of gorilla numts was constrained to only hits of either i) an identity greater than 70% and size between 50 and 99bp or ii) at least 60% identity and 100bp in length (see Soto-Calderón et al., 2012). The number of numts per site was estimated for each nucleotide position in the region of interest for each taxon. A Kruskal-Wallis test was conducted to test the hypothesis that the distribution of the absolute number of numts per nucleotide derived from the four MCR sub-domains and the 500bp flanking regions does not differ among the four great apes (gorillas, humans, chimpanzees and orangutans). Wilcoxon signed-rank tests were then used to test for significant pair-wise differences between taxa.

2.4. Isolation of nuclear-enriched gorilla DNA and anchored PCR

As part of an anchored-PCR strategy designed to upwardly bias the amplification of nuclear copies over mtDNA sequences, we adopted a number of steps to enrich gorilla genomic DNA extractions for nuclear DNA (Soto-Calderón, 2012). As a first step, we selected five different gorilla fibroblast cell lines (Coriell Institute for Medical Research, Camden, NJ, USA) whose mitochondrial genomes lacked a *Bgl*II (A↓GATC↑T) recognition site, as determined by restriction enzyme digestion of three overlapping fragments amplified from across the entire mtDNA genome using the long-range *LA Taq* polymerase (Takara Bio Inc., Mountain View, CA, USA). Primers used to amplify these three overlapping fragments are listed in Table S2. After digesting with *Bgl*II, the mtDNA of these individuals would remain circularized during the Y-linker ligation described below, thus making it impossible to amplify.

The five selected cell lines lacking *Bgl*II sites were subsequently grown for four weeks in the presence of 2’,3’-dideoxycytidine (ddC) to selectively inhibit mtDNA replication, leading to a progressive dilution and virtual loss of mtDNA in cultured cells (Ashley et al., 2005). Nuclear DNA (nDNA) was then extracted from the ddC-treated cell lines using the Blood & Cell Culture DNA Maxi Kit (Qiagen, Valencia, CA, USA), which preferentially recovers high-molecular-weight DNA averaging between 50–100kb, further reducing the abundance of any remaining mtDNA (~16.5 Kbp). The copy number of mtDNA relative to nDNA was then estimated by quantitative PCR of two reporter genes, the mitochondrial cytochrome *b* gene and the nuclear tumor suppressor gene p53, using the primer pairs CytbGor F / CytbGor R and p53iiPrim F / p53ii-R, respectively (Table S2).

The nuclear-enriched DNA samples obtained from the five ddC-treated cell lines were subsequently pooled in equal concentrations, completely digested with *Bgl*II to generate fragments with 5’-GATC overhangs and ligated to a compatible Y-linker made up of the two partially complementary oligos *Bgl*II-top (5’-GATCGAAGGAGAGGACGCTGTCTGTCGAAGG-3’; modified from Ray et al., 2005) and *Bgl*II-bottom (Ray et al., 2005). Partial sequences of putative MCR numt loci and the corresponding flanking regions were then amplified from the treated samples through an anchored PCR strategy using one of several MCR primers in combination with the primer LNP (see Ray et al., 2005 for details). In this method, the MCR primer binds to a given numt fragment leading to the extension of the first strand (Figure S1). This creates a binding site for the primer LNP at the 3' end of the nascent strand, allowing the amplification of the complementary reverse strand during the second round of PCR. This is followed by a semi-nested PCR using the same LNP primer and an internal MCR primer to increase the specificity of the amplified products. A-PCR amplifications were carried out in 20μl reactions containing 1U *LA Taq* polymerase TaKaRa (Bio Inc.), 0.4mM dNTPs, 0.2μM of each primer, 1× buffer and 20–30ng DNA. Cycling consisted of initial denaturation at 94°C, followed by 35 cycles of 94°C for 15 s and 68°C for 15 min, with a final extension at 72°C for 2 min. PCR products were then cloned into the pCR®2.1 vector using the TOPO TA-cloning kit (Invitrogen, Carlsbad, CA, USA) and sequenced with the BigDye v1.1 (Applied Biosystems). These sequences were then aligned with the gorilla mitochondrial genome to determine the location of the 5' portion of the numt and its adjacent nuclear flank. The identified flank was then BLATed (Kent, 2002) against the human and chimpanzee reference genomes as described above to identify the genomic location of the orthologous locus, infer the sequence of the second flank and determine whether the corresponding numt was unique to gorillas. Lastly, locus-specific primers were designed to amplify across the second nuclear flank of each numt locus.

2.5. Sequencing and phylogenetic analysis of the HV1 region

Genomic DNA samples from 41 western gorillas from USA zoos were extracted using the DNeasy Blood & Tissue Kit (Qiagen) from peripheral blood samples kindly donated by participating institutions and collaborators (Table S3). A section of 6,880bp of the mitochondrial region containing the HV1 domain was amplified with *LA Taq* polymerase using specific primers (mt10261 Fa / mt726 R; Table S2) and sequenced using the BigDye v1.1 and primers flanking the HV1 (mt15365 F / mt15888 R).

In order to determine whether gorilla HV1 numts identified in the study belonged to any of the previously defined numt classes I – III (Anthony et al., 2007a), we first aligned the HV1 portion of four new gorilla-specific numt loci identified in the present study with a database of existing gorilla HV1 mitochondrial and numt sequences using the ClustalW algorithm implemented in MEGA v5.1 (Tamura et al., 2011). This database was comprised of 41 new HV1 sequences obtained in this study, 10 publicly available HV1 sequences (Table S3) and a set of 207 sequences previously compiled by Anthony et al. (2007b) comprising 166 mitochondrial HV1 sequences of free-range western gorillas (Haplogroups A–D) and 41 gorilla-specific HV1 numt sequences (classes I, II and III). A stretch of 90bp unique to the gorilla-specific HV1 numt Go9_5746 and a poly-cytosine stretch of 26bp that is prone to error during polymerase amplification were removed from the sequence matrix leaving a final alignment length of 235bp.

Redundant sequences were identified using the program Collapse v1.2 (Posada, 2006), leaving 102 mitochondrial and 42 unique numt sequences. A Bayesian phylogeny of this alignment was then conducted in Beauti/Beast v1.7.5 (Drummond and Rambaut, 2007), using a Monte Carlo Markov Chain (MCMC) of 50 million steps in length with a sampling interval every 1000 steps. A GTR+G model of nucleotide substitution was employed and an exponential relaxed molecular clock was assumed with default parameters in order to allow for among lineage rate variation. A burn-in period of 10% was determined from visual inspection of output in Tracer v1.5 and a final tree was summarized in TreeAnnotator v1.7.5 (Drummond and Rambaut, 2007).

2.6. Specificity of mitochondrial primers for mitochondrial and nuclear targets

The two gorilla-specific numts that contain an entire copy of the HV1 sub-domain (Go1_1300 and Go2b_2500) exhibit high identity with mitochondrial sequences. A number of primers have been previously designed to amplify the MCR HV1 mitochondrial region in gorillas but recent nuclear copies could be inadvertently co-amplified. In order to assess the extent to which existing mitochondrial primers would be expected to co-amplify nuclear copies we carried out a detailed comparison of the primer regions of two numt sequences with a high identity to the mitochondrial sequence (Go1_1300 and Go2b_2500) to determine the number of mismatches between the primer sequence and nuclear copies. We assessed the following primer pairs: MidRev4/ProFor2 and D-441/D88 (Jensen-Seaman et al., 2004); H402/L91 (Garner and Ryder, 1996); H16498/L15926 (Thalmann et al., 2004); and MTD1AS/MTD1S (LaCoste et al., 2001), where MTD1AS is also known as H16498.

2.7. Analysis of insertional polymorphisms

Once flanking sequences were obtained for all gorilla MCR numts, their presence/absence was assessed in a panel of 67 DNA samples of captive gorillas whose mitochondrial haplogroup affiliation (A – D) was determined through HV1 sequencing or through pedigree analysis of the gorilla studbook (Wharton, 2007). Where possible, primers were designed to amplify the entire region containing the numt and infer individual co-dominant genotypes, i.e. a single larger product with the numt insertion (+/+), a smaller product without the insertion (−/−) or two different products consistent with the heterozygous state (+/−). Amplification of the entire region containing a numt was not always possible due to either the large size of the target region or potential disparities within the primer region. In such cases, primers were designed to amplify a portion of the numt and one of its flanks (Figure S2) and only the presence or absence of the insertion could be determined. To safeguard against false negatives (i.e. failure of the PCR reaction), all PCR reactions were carried out using an internal standard based on co-amplification of a conserved region in the housekeeping tumor-suppressor gene p53 using the primer P53ii R in combination with either TP53 F or p53 3F (Table S2). A human DNA sample was also included in each PCR reaction as a negative control. Each 20μl PCR reaction contained 0.5U of *Taq* DNA polymerase (Invitrogen), 2.5mM $MgCl_2$, 200μM dNTPs, 250μM each primer and 20–30ng DNA and 1× PCR buffer. Cycling consisted of 2 min of initial denaturation at 94°C followed by 35 cycles of 94°C for 30s, 58–64°C for 30s and 72°C for 50s −2 min, with a final extension at 72°C for 2 min.

The mitochondrial haplotype affiliation of a given gorilla has been shown to be a general indicator of geographical origin (Anthony et al., 2007a) and can provide important information on the genealogical history of captive individuals where the complete pedigree can be inferred. In order to assess the impact of regional genetic structure on the distribution of insertional polymorphisms, we genotyped five numt loci in 19 zoo animals born in the wild and one captive-born gorilla (Kwanza) whose parents shared the same mitochondrial haplogroup.

3. Results

3.1. BAC surveys of mitochondrial DNA-wide numts in gorillas and chimpanzees

A total of 64 radiolabeled probes designed from mtDNA sequence were used to screen the gorilla and chimpanzee BAC library. Because of the inherent variation among hybridization experiments due to activity of the isotope, labeling efficiency, and reuse of library filters, all labeled probes were hybridized to chimpanzee and gorilla nuclear genomic BAC libraries together. BAC hits from the developed autoradiographs of the hybridizations using the 64 mtDNA-wide probes were scored if the signal was as dark, or darker than the *C. briggsae* control probe signal (see Methods). A total of 542 BACs were hit in gorilla, and 438 BACs in chimpanzee. It is important to remember that the genomic coverage differs between these libraries, estimated at ~7× for gorilla and 5.2× for chimpanzee (www.bacpac.chori.org). As a result, the normalized number of hits is similar with ~77 hits per genome equivalent in the gorilla, compared to ~84 in the chimpanzee (Figure 1). Moreover, while there is variation among mtDNA loci, there is little difference between the gorilla and the chimpanzee at any given locus, with 8 of 16 probe pools detecting more BAC hits in the gorilla, and eight detecting more in chimpanzee. The region of the mtDNA genome producing the most hits to the nuclear genomic library codes for the 12S and 16S ribosomal RNA genes.

3.2. Mitochondrial-wide BLAST surveys of the gorilla genome

The BLAST search of the current partially assembled version of the reference genome using the complete gorilla mtDNA genome as a query sequence retrieved 46 gorilla-specific numts with an average size of 251bp, 11 of which exist as duplications within major chromosomal segments. The logarithm of numt size of these 46 numts was inversely related to frequency ($y = -11.41\log(x) + 110.7$; Pearson = −0.38; d.f. = 44; $p = 0.009$; $R^2 = 0.1440$). The average numt size is likely to be an underestimation since sequencing gaps remain to be filled for 16 of these numts (34.8%) in the current genome assembly. The average percentage of sequence identity between retrieved numts and their corresponding mitochondrial sequences was 85.4%. The shortest numts (≤ 100bp) represented over one third (37%) of all numts. The longest numt sequence detected from the BLAST search was 1,038bp in length and shared a sequence identity of 69% with the corresponding mitochondrial fragment. However, the length of two numts with internal sequence gaps could be as much as 2,400bp (Table 1) after alignment of their flanks with their paralogous mitochondrial sequences.

3.3. Prevalence of MCR numts in the four major great ape taxa

Using BLAST search criteria common to all species, we found four gorilla MCR numts in the current genome assembly (Go9_5746, Go1_1300, Go5_30 and Go2b_43). This number contrasts with the amount of species-specific numts found in orangutans (23), chimpanzees (25) and humans (2) (see Soto-Calderón et al., 2012). The pattern of numt prevalence across the four sub-domains is similar in all four great apes (Figure 2), with a general underrepresentation of numts from the sub-domains HV2 and MCR_F. However, the number of numts differed significantly among taxa ($\chi^2 = 1{,}679.0$, d.f. = 3, $p < 0.0001$), with a significantly higher number of numts in chimpanzees. In addition, a significant deficit of HV1 numts was also detected in gorillas relative to the other three taxa.

3.4. Complementary searches for gorilla-specific MCR numts

We identified a total of 67 putative gorilla-specific numts and successfully mapped them to their location in the gorilla genome (Table S4). Fifty of these numts were validated by their presence in the *Gorilla gorilla* build (gorGor3.1) and/or experimentally verified through amplification from the gorilla genomic DNA panel (Table S4). Seventeen of the remaining numt hits identified through the A-PCR or BLAST searches of unassembled contigs could not be validated either through PCR amplification of available genomic DNA samples or by their presence in the assembled genome database. One and three putative numts were identified through complementary BLAST searches of the former gorilla genomic assembly CABD00000000.2 and trace files, respectively (Table 1). Twenty-five gorilla genomic BAC clones identified through hybridization with MCR probes were mapped to the human genome assembly via BLAT of their end-sequences (Table S5). These BACs mapped to 11 unique locations, consistent with the redundancy of BAC libraries. Numts were clearly present in the human genome at five of these 11 loci. From the remaining six loci containing putatively gorilla-specific numts, we prioritized BACs with multiple hybridization hits, and characterized two large gorilla specific numts (Go1_1300 and Go2b_2500) by sequencing out from the insertion into the flanking nuclear DNA.

Combining the results from these various search methods, we mapped seven numts partially or entirely derived from MCR (Table 1). These loci comprise two found in both the BAC library screens and BLAST searches of the current gorilla build, one identified through the A-PCR approach alone and four only found through BLAST searches. Four MCR numt loci encompassed all or part of the HV1 sub-domain and the remaining three contained other MCR sub-domains. These seven MCR numts specific to the gorilla genome are described below (See Table 1 for details).

Go1_1300—This numt is an integration of 1,347bp in length that contains most of the MCR including HV1, CCD and an extensive portion of HV2, as well as nuclear pseudogenes of cytochrome b, and the tRNAs of threonine and proline.

Go2b_2500—This is the longest gorilla-specific numt found in this study. It corresponds to an integration of 2,497bp that contains the entire MCR region as well as copies of the NADH dehydrogenase subunit 6, cytochrome *b*, 12S ribosomal RNA and the tRNAs of glutamic acid, threonine, proline and phenylalanine.

Go5_40—The numt Go5_40 is a 456bp integration that encompasses the entire CCD along with a section of both HV1 and HV2.

Go9_5746—This numt consists of an integration of 447bp made up of two non-contiguous mitochondrial fragments, the first containing a portion of HV1 and the CCD and the second containing 37bp of the ATP synthase - subunit 6 that is 7,100bp apart in the mitochondrial genome. Although Go9_5746 is exclusive to gorillas, it contains a ~90bp section of mtDNA that is no longer present in the mitochondrial genome of contemporary gorillas but is still found in all the other great ape taxa.

Go11_188—After Go2b_2500, Go11_188 is the longest gorilla-specific numt found in this study. This mitochondrial integration is only partially sequenced in the gorilla genome assembly but an alignment with the reference mitochondrial genome suggests that its full size would be around 2,420bp. This numt lacks the HV1 although it contains all the other MCR sub-domains as well as copies of the 12S and 16S rRNAs and the tRNAs of phenylanine and valine.

Go5_30—A partial sequence of the numt Go5_30 was recovered from the current gorilla genome assembly. This sequence is comprised of two sequence fragments totaling 375bp and an internal gap in the sequence. The left and right hand fragments of this numt contain partial sequences of a cytochrome *b* pseudogene and CCD, respectively, suggesting that the full size of this mitochondrial integration might be approximately 1,560bp. This means that in addition to cytochrome *b* and the CCD MCR sub-domain, the numt Go5_30 might also contain the HV1 and pseudogenes of the tRNAs for threonine and proline.

Go2b_43—This numts is a 54bp integration entirely derived from HV2 and the shortest gorillaspecific MCR numt found in this study.

3.5. Phylogenetic analysis of gorilla-specific HV1 numts

Phylogenetic analysis of mitochondrial HV1 and numt sequences recovered all previously defined mitochondrial haplogroups (A – D) and numt classes (I – III) as well as a previously undescribed mitochondrial lineage identified in four captive gorillas referred to as C3 (Figures 3 and S3). Of the four HV1 numts characterized here, Go1_1300 and Go2b_2500 clearly clustered with the numt sub-classes IIb and I respectively, whereas Go5_40 clustered with the numt sub-class IIa (Figure S3). Unlike the three other mapped numts, Go9_5746 could not be assigned to any predefined numt class and is substantially different from previously reported HV1 numts (<85% identity).

The HV1 section of Go1_1300 shares ≥ 99% similarity to representatives of class IIb numts and exhibits all the diagnostic sites (see Figures 3 and S3) that are characteristic of this numt group, as defined by Anthony et al. (2007a). Sequences showing the highest identity with this numt include AY530149, amplified from a wild gorilla in Lobéké, Cameroon (Clifford et al., 2004); L76766, from the captive gorilla Carolyn, captured in the Congo region (Garner and Ryder, 1996) and the Rok8 sequence from the captive gorilla Rok (Thalmann et al., 2004), which with the exception of a 4bp gap showed perfect identity with Go1_1300. Go1_1300 was also identical to two numts found in two different eastern lowland gorillas: AF240455 (LUT2DTA9) (Lutunguru, Democratic Republic of Congo; Jensen-Seaman et al., 2004) and Muk5 (Thalmann et al., 2004).

The HV1 portion of Go2b_2500 clustered with class I numts and exhibited a sequence identity of 99% with the eastern lowland gorilla numts AF240456 (LUT2DTA10) and AF240448 (LUT2DTA1). This sequence also contains the diagnostic T79 site that characterizes numt classes I, II and III in gorillas (Figures 3 and S3) (Jensen-Seaman et al., 2004). This numt also exhibited high sequence identity (97%) with the putative numt L76760, from the western gorilla Jojo (Garner and Ryder, 1996). Curiously, Go2b_2500 also shared elevated sequence identity (99%) with a western lowland gorilla sequence AY530145 (BEL1a), considered to be an *in vitro* recombinant between mitochondrial and class I numts (see Anthony et al., 2007a).

The numt Go5_40 possesses three diagnostic sites (T79, A92 and G156) that define the numt class II group of sequences (Figures 3 and S3). However, phylogenetic analysis failed to assign Go5_40 to a particular numt sub-class within this group (IIa – IIc). Although this numt possesses one of the several diagnostic sites that define the IIa subgroup, it also shares high sequence identity (<95%) with class IIc numts (such as Muk4, Muk6 and Muk7 recovered from the eastern lowland gorilla Muk), as well as Rok5 (amplified from the western lowland gorilla Rok) (Thalmann et al., 2004).

3.6. Lack of specificity of mitochondrial primers

Regions that bind previously published PCR primers of the mitochondrial genome and corresponding numt loci Go1_1300 and Go2b_2500 are virtually identical in sequence composition and are unlikely to favor amplification of the mitochondrial over the nuclear copy (Figure 4). In four cases, the primer sequences were identical to both the mitochondrial and the nuclear copies (H16498, MidRev4, D-441 and ProFor2). In two cases there was only one internal mismatch between the primer and both mitochondrial and nuclear copies (H402 and MTD1S). In the remaining three cases (D-88, L91 and L15926), between three and seven mismatches were observed between primer sequences and both mitochondrial and nuclear targets.

3.7. Numt insertional polymorphisms

From the five gorilla numts genotyped in this study, only Go1_1300, Go2b_2500 and Go5_40 were found to be polymorphic in captive western gorillas. Both Go9_5746 and Go11_188 are fixed, at least in the sample of western gorillas in the present study (See Table 1 and Table S6). Despite the limited sample size, some noticeable differences in numt insertional polymorphisms were detected between samples with different haplogroup affiliations (Table 2). Go1_1300 and Go2b_2500 were found associated with lineages C1, C3 and D3 but absent in C2 and D2. In contrast, Go5_40 was common in sub-haplogroups C1, C2, D2 and D3 but absent in C3.

4. Discussion

This is the first study that combines wet lab methods with bioinformatic tools to characterize numts and determine the size and mitochondrial region of origin of these sequences in gorillas. A combination of survey strategies proved to be the most effective means of recovering numts that might not have been otherwise found using a single strategy approach. Combination of data from BAC library screens of chimpanzee and gorilla genomes and specific surveys of the MCR have shown that although the total number of numts does not appear to vary between these two taxa, there are differences in the representation of particular mitochondrial subdomains in the nuclear genome of these taxa

The contrast of BLAST surveys in different great apes however revealed that the observed number of MCR numts does differ significantly among the great ape taxa examined here. Interestingly, chimpanzees and orangutans seem to have accumulated a much larger number of MCR numts than gorillas and humans (Soto-Calderón et al., 2012). Whether or not differences in the prevalence of MCR numts between taxa can be extrapolated to the rest of the mitochondrial genome is arguable. Future and ongoing sequencing of whole genomes (Lang et al., 2011; Prado-Martinez et al. 2013) and mapping of candidate numts to one or more reference genomes will permit comparisons across a wider range of taxa and identification of other putative MCR numts not found in the present study, such as those in numt classes IIa, IIc and III.

The analysis of abundance of numts across each of the four sub-domains in the MCR and the two 500bp flanks revealed a marked deficit of numts from the sub-domains HV2 and MCR_F, not only in gorillas but in other great apes included in this study (Soto-Calderón et al., 2012). The abrupt change in the prevalence of numts from contiguous sub-domains is unlikely to reflect differences in the rate of translocation across MCR sub-domains (Nugent and Palmer, 1991; Henze and Martin, 2001). Instead, the scarcity of HV2 and MCR_F numts in all great ape species examined in the present study is likely explained by elevated mutation rate of these sub-domains leading to a subsequent rapid loss of sequence identity with nuclear copies (Mourier et al., 2001; Soto-Calderón et al., 2012). It is therefore possible that an unusually high mutation rate of HV1 in gorillas has obliterated the overall sequence similarity between the HV1 sequences and their nuclear copies, leading to an apparent deficit of numts from this region. As the peripheral regions of the MCR in mammals exhibit exceptionally high variation in length and sequence associated with the presence of nucleotide repeats (Saccone et al. 1991; Sbisà et al. 1997), rapid sequence evolution of this region would be expected to lead to loss of sequence identity with nuclear copies and a generalized deficit of numts from this region across mammals, as evidenced so far not only in primates but also in taxa as distant as Perissodactyla (Mourier et al. 2001; Nergadze et al. 2010; Soto-Calderón et al. 2012).

Genomic mapping of numt loci and verification of their nuclear flanks allowed us for the first time to demonstrate the nuclear origin of at least two putative MCR numt classes described in other studies (Thalmann et al., 2004; 2005; Anthony et al., 2007a) and also enabled the design of specific primers for characterization of MCR numt polymorphisms in gorilla populations. Our ability to exhaustively survey numts in the gorilla genome was nonetheless constrained by several factors. Firstly, the relatively small number of gorillas used to retrieve numt loci necessarily limited the scope of our search. Secondly, our A-PCR approach to numt retrieval was limited by the proximity of numt loci to a *Bgl*II restriction site. These restriction sites are only found on average every 5,460bp in the great ape genome (NEB, 2004); therefore, numts contained in genomic fragments lacking nearby *Bgl*II restriction sites might be overlooked. Thirdly, the process of assembling the gorilla genome reads is guided by homology with humans, which may constrain the ability to assemble regions containing gorilla-specific integrations or regions with multiple chromosomal duplications, as might be the case for gorilla numts (Scally et al., 2012). It is likely that the constraints that we faced in the present study are broadly applicable to numt surveys of other taxa, which rarely examine more than a handful of individuals. Future work aimed at identifying recently integrated numts in other taxa should adopt a diversity of approaches and include a broad subset of individuals likely to represent range-wide genetic variation.

Phylogenetic analysis provided evidence for at least two classes of gorilla HV1 numts that have been shown to be prevalent in other PCR-based studies of gorilla genetic variation (Anthony et al., 2007a; Clifford et al., 2004; Garner and Ryder, 1996; Jensen-Seaman et al., 2004; LaCoste et al., 2001; Thalmann et al., 2004; 2005). Our findings also show that numts Go1_1300 and Go2b_2500, which contain the entire MCR, exhibit so few differences with the mitochondrial sequence as to permit co-amplification of mitochondrial and nuclear templates (Anthony et al., 2007a; Clifford et al., 2004; Garner and Ryder, 1996; Jensen-Seaman et al., 2004; LaCoste et al., 2001; Thalmann et al., 2004). The risk of co-amplifying these numts alongside mitochondrial sequences is also exacerbated by the fact that these two numts seem to be widely distributed across the geographic range of western gorillas. This problem might be particularly acute in non-invasive studies where the nuclear-to-mitochondrial ratio tends to increase relative to fresh tissue, increasing the chance of amplifying nuclear templates with conventional mitochondrial primers (Greenwood and Pääbo, 1999; Berger et al., 2001; Clifford et al., 2004; Jensen-Seaman et al., 2004; Foran, 2006). Template switching during PCR amplification can also lead to the generation of *in vitro* recombinants that may further confound correct identification of numt sequences (Anthony et al., 2007a). Given the greater probability of amplifying numts in degraded samples and limited primer specificity of mitochondrial primers, we recommend the use of genetic material from fresh tissue and long-range PCR whenever possible to amplify mitochondrial sequences in gorillas (Thalmann et al., 2004; Triant and DeWoody, 2007; Song et al., 2008; Calvignac et al., 2011). In cases where degraded samples are the only source of genetic material and long-range PCR is not a practical option, we recommend implementing quality control tools such as phylogenetic analysis and recombination detection to differentiate mitochondrial sequences from *in vitro* recombinants and nuclear integrations (Jensen-Seaman et al., 2004; Anthony et al., 2007a).

The phylogenetic placement of numt loci in relation to major mitochondrial haplogroups also provides some interesting insights into the divergence history of eastern and western gorillas. Previous studies have shown that historical gene flow between eastern and western gorillas is much more recent than was previously thought (Jensen-Seaman et al. 2001; Anthony et al., 2007b; Scally et al., 2012; Thalmann et al., 2007; 2011; Mailund et al. 2012). Evidence to support this hypothesis is provided by the phylogenetic placement of the class III numt group which, despite being sister to the eastern gorilla mitochondrial haplogroups A and B, is present in both east lowland and western gorillas (Jensen-Seaman et al. 2004; Thalmann et al., 2004). This hypothesis is also backed up by the presence of class II numts in eastern lowland and western gorillas. Remarkably however, mtDNA haplogroups are never shared between eastern and western gorillas, providing support for male mediated east-to-west gene flow (Jensen-Seaman et al. 2003; Thalmann et al., 2004; 2007) and greater philopatry in female gorillas (Douadi et al., 2007). Combined analyses of craniometrical and mitochondrial variation have also provided compelling evidence of ancestral gene flow between western and eastern gorillas and between eastern gorilla subspecies (Ackermann and Bishop, 2009).

Lastly, this study has provided a general strategy for genotyping insertional polymorphisms in western gorillas that could be applied to other taxa. The absence of a numt locus is generally considered to be the ancestral condition whereas the insertion of a numt locus is considered to correspond to the derived state. Patterns of insertional polymorphism may therefore be useful in inferring the direction and intensity of gene flow between free-range populations of western and eastern gorillas and provide novel insights into the recent evolutionary history of this and other taxa. Insertional polymorphisms in *Alu* elements for instance, have been extensively used in humans to infer relationships between populations (Perna et al. 1992; Thangaraj et al. 1999; Nasidze et al. 2001; Watkins et al. 2003) and reconstruct historical patterns of migration (Batzer et al.1994; 1996). Similarly, polymorphic numts have allowed the reconstruction of migration patterns and more recently patterns of admixture in humans (Thomas et al. 1996; Lang et al. 2012). However, the limited availability of these markers has hindered their use in population genetic studies of non-human primates. Strategies such as those described here could readily be applied to other taxa where genomic resources are now available and provide a complementary set of genetic markers suited to resolving questions at the population level.

5. Conclusion

The present study revealed 67 numts that appear to be specific to gorillas, including seven loci encompassing partial or entire copies of the MCR. Of the gorilla-specific MCR numts identified in the present study, two (Go1_1300 and Go2b_2500) cluster with the previously described numt classes I and IIb, thus providing conclusive evidence for their nuclear origin. Phylogenetic analysis of numt sequences in this study also provides convincing evidence for past hybridization between eastern and western gorillas, as has been hypothesized in earlier studies (Ackermann and Bishop, 2009; Thalmann et al., 2007). These findings also show that numts can provide additional insights into evolutionary history that until now have gone largely underappreciated. Owing to the high sequence similarity between recent numts and their corresponding mitochondrial copies, it appears that the risk for inadvertent amplification of nuclear copies is high. To safeguard against this problem, we recommend using a phylogenetic approach and/or PCR amplification of large mitochondrial fragments in order to distinguish HV1 mitochondrial sequences from nuclear copies and minimize the risk of their inclusion in mitochondrial sequence databases. Finally, we provide a comprehensive list of gorilla-specific numts and demonstrate their potential utility as both cladistic and population genetic markers. As whole genome sequence data becomes increasingly available, future work should seek to inventory numts in other non-model organisms, not only as a means of diagnosing these loci and sorting them from their mitochondrial copies but also for their use as essentially homoplasyfree genetic markers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Dr. Richard A. Bergl at the North Carolina Zoo, Dr. Naida Loskutoff at the Omaha's Henry Doorly Zoo and Dr. Betsy Dresser for donating samples of captive gorillas in USA zoos. We also thank Dr. Christine Drea at Duke University who kindly offered her laboratory for DNA extractions. We are also thankful to graduate committee members Drs. Prescott Deininger at Tulane University and Steve Johnson, Mary Clancy and Charles Bell at the University of New Orleans. This work was supported by the National Institutes of Health (Grant No. R15 GM073682-01).

Abbreviations

MCR	Mitochondrial Control Region
CCD	Central Conserved Domain in the MCR
MCR$_F$	the sub-domain in the MCR proximal to the phenylalanine tRNA gene
HV1 and HV2	Hyper-variable regions 1 and 2 in the MCR, respectively
MT$_P$ and MT$_F$	the two 500bp regions flanking the MCR; numt, a nuclear copy of mitochondrial DNA
BAC	Bacterial Artificial Chromosome
A-PCR	Anchored PCR
TE	Transposable Element

References

Ackermann RR, Bishop JM. Morphological and molecular evidence reveals recent hybridization between gorilla taxa. Evolution. 2009; 64:271–290. [PubMed: 19804402]

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990; 215:403–410. [PubMed: 2231712]

Anthony NM, Clifford SL, Bawe-Johnson M, Abernethy KA, Bruford MW, Wickings EJ. Distinguishing gorilla mitochondrial sequences from nuclear integrations and PCR recombinants: guidelines for their diagnosis in complex sequence databases. Mol. Phylogenet. Evol. 2007a; 43:553–566. [PubMed: 17084645]

Anthony NM, Johnson-Bawe M, Jeffery K, Clifford SL, Abernethy KA, Tutin CE, Lahm SA, White LJ, Utley JF, Wickings EJ, Bruford MW. The role of Pleistocene refugia and rivers in shaping gorilla genetic diversity in central Africa. Proc. Natl. Acad. Sci. USA. 2007b; 104:20432–20436. [PubMed: 18077351]

Ashley N, Harris D, Poulton J. Detection of mitochondrial DNA depletion in living human cells using PicoGreen staining. Exp. Cell Res. 2005; 303:432–446. [PubMed: 15652355]

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ. African origin of human-specific polymorphic Alu insertions. Proc. Natl. Acad. Sci. USA. 1994; 91:12288–12292. [PubMed: 7991620]

Batzer MA, Arcot SS, Phinney JW, Alegria-Hartman M, Kass DH, Milligan SM, Kimpton C, Grill P, Hochmeister M, Ioannou PA, Herrera RJ, Boudreau DA, Scheer D, Keats BJB, Deininger PL, Stoneking M. Genetic variation of recent Alu insertions in human populations. J. Mol. Evol. 1996; 42:22–29. [PubMed: 8576959]

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nature Reviews Genetics. 2002; 3:370–379.

Batzer MA, Deininger PL. Alu repeats and human genetic diversity. Nature Rev Genet. 2002; 3:370–380. [PubMed: 11988762]

Berger A, Bruschek M, Grethen C, Sperl W, Kofler B. Poor storage and handling of tissue mimics mitochondrial DNA depletion. Diagn. Mol. Pathol. 2001; 10:55–59. [PubMed: 11277396]

Calvignac S, Konecny L, Malard F, Douady CJ. Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (numts). Mitochondrion. 2011; 11:246–254. [PubMed: 21047564]

Clifford SL, Anthony NM, Bawe-Johnson M, Abernethy KA, Tutin CEG, White LJT, Bermejo M, Goldsmith ML, Mcfarland K, Jeffery KJ, Bruford MW, Wickings EJ. Mitochondrial DNA phylogeography of western lowland gorillas (*Gorilla gorilla gorilla*). Mol. Ecol. 2004; 13:1551–1565. [PubMed: 15140097]

Douadi MI, Gatti S, Levrero F, Duhamel G, Bermejo M, Vallet D, Menard N, Petit EJ. Sex-biased dispersal in western lowland gorillas (*Gorilla gorilla gorilla*). Mol. Ecol. 2007; 16:2247–2259. [PubMed: 17561888]

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 2007; 7:214. [PubMed: 17996036]

Foran DR. Relative degradation of nuclear and mitochondrial DNA: An experimental approach. J. Forensic. Sci. 2006; 51:766–770. [PubMed: 16882217]

Frey JE, Frey B. Origin of intra-individual variation in PCR-amplified mitochondrial cytochrome oxidase I of *Thrips tabaci* (Thysanoptera: Thripidae): mitochondrial heteroplasmy or nuclear integration? Hereditas. 2004; 140:92–98. [PubMed: 15061785]

Garner KJ, Ryder OA. Mitochondrial DNA diversity in gorillas. Mol. Phylog. Evol. 1996; 6:39–48.

Greenwood AD, Päävo S. Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants. Mol. Ecol. 1999; 8:133–137. [PubMed: 9919702]

Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 2010; 6(2):e1000834. [PubMed: 20168995]

Henze K, Martin W. How do mitochondrial genes get into the nucleus? Trends Genet. 2001; 17:383–387. [PubMed: 11418217]

Herke SW, Xing J, Ray DA, Zimmerman JW, Cordaux R, Batzer MA. A SINE-based dichotomous key for primate identification. Gene. 2007; 390:39–51. [PubMed: 17056208]

Jensen-Seaman MI, Deinard AS, Kidd KK. Modern African ape populations as genetic and demographic models of the last common ancestor of humans, chimpanzees, and gorillas. J. Hered. 2001; 92:475–480. [PubMed: 11948214]

Jensen-Seaman, MI.; Deinard, AS.; Kidd, KK. Mitochondrial and nuclear DNA estimates of divergence between western and eastern gorillas. In: Taylor, AB.; Goldsmith, ML., editors. Gorilla Biology: a multidisciplinary perspective. Cambridge University Press; 2003.

Jensen-Seaman MI, Sarmiento EE, Deinard AS, Kidd KK. Nuclear integrations of mitochondrial DNA in gorillas. Am. J. Primatol. 2004; 63:139–147. [PubMed: 15258958]

Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002; 12:656–664. [PubMed: 11932250]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002; 12:996–1006. [PubMed: 12045153]

Lacoste V, Mauclère P, Dubreuil G, Lewis J, Georges-Courbot MC, Gessain A. A Novel γ2-herpesvirus of the rhadinovirus 2 lineage in chimpanzees. Genome Res. 2001; 11:1511–1519. [PubMed: 11544194]

Lang M, Sazzini M, Calabrese F, Simone D, Boattini A, Romeo G, Luiselli D, Attimonelli M, Gasparre G. Polymorphic NumtS trace human population relationships. Hum. Genet. 2012:1511–1519.

Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, Lunter G, Prüfer K, Scally A, Hobolth A, Schierup MH. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 2012; 8:e1003125. [PubMed: 23284294]

Mourier T, Hansen AJ, Willerslev E, Arctander P. The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol. Biol. Evol. 2001; 18:1833–1837. [PubMed: 11504863]

Nasidze I, Risch GM, Robichaux M, Sherry ST, Batzer MA, Stoneking M. Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. Eur. J. Hum. Genet. 2001; 9:267–272. [PubMed: 11313770]

Nergadze SG, Lupotto M, Pellanda P, Santagostino M, Vitelli V, Giulotto E. Mitochondrial DNA insertions in the nuclear horse genome. Anim. Genet. 2010; 41:176–185. [PubMed: 21070293]

Nugent JM, Palmer JD. RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. Cell. 1991; 66:473–481. [PubMed: 1714355]

Perna NT, Batzer MA, Deininger PL, Stoneking M. Alu insertion polymorphism: A new type of marker for human population studies. Human Biol. 1992; 64:641–648. [PubMed: 1328024]

Posada D. Collapse: describing haplotypes from sequence alignments. 2006 http://darwin.uvigo.es/software.html.

Prado-Martinez J, Sudmant PH, Kidd JM, et al. Great ape genetic diversity and population history. Nature. 2013; 499:471–475. [PubMed: 23823723]

Ray DA, Xing J, Hedges DJ, Hall MA, Laborde ME, Anders BA, White BR, Stoilova N, Fowlkes JD, Landry KE, Chemnick LG, Ryder OA, Batzer MA. Alu insertion loci and platyrrhine primate phylogeny. Mol. Phylogenet. Evol. 2005; 35:117–126. [PubMed: 15737586]

Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. Mol. Biol. Evol. 2004; 21:1081–1084. [PubMed: 15014143]

Ross, MT.; LaBrie, S.; McPherson, J.; Stanton, VP, Jr. Screening large-insert libraries by hybridization. In: Boyl, A., editor. Current Protocols in Human Genetics. New York: Wiley; 1999. p. 5.6.1-5.6.52.

Saccone C, Pesole G, Sbisà E. The main regulatory region of mammalian mitochondrial DNA: Structure-function model and evolutionary pattern. J. Mol. Evol. 1991; 33:83–91. [PubMed: 1909377]

Sbisà E, Tanzariello F, Reyes A, Pesole G, Saccone C. Mammalian mitochondrial D-loop region structural analysis: Identification of new conserved sequences and their functional and evolutionary implications. Gene. 1997; 205:125–140. [PubMed: 9461386]

Scally A, Dutheil JY, et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012; 483:169–175. [PubMed: 22398555]

Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Sci USA. 2008; 5:13486–13491. [PubMed: 18757756]

Soto-Calderón, ID. Ph.D. Dissertation. University of New Orleans; 2012. Evolution of nuclear integrations of the mitochondrial genome in great apes and their potential as molecular markers.

Soto-Calderón ID, Lee EJ, Jensen-Seaman MI, Anthony NM. Factors Affecting the Relative abundance of nuclear copies of mitochondrial DNA (numts) in Hominoids. J. Mol Evol. 2012; 75:102–111. [PubMed: 23053193]

Stewart C, Kural D, Stromberg MP, et al. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genetics. 2011; 7:e1002236. [PubMed: 21876680]

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 2011; 28:2731–2739. [PubMed: 21546353]

Thalmann O, Hebler J, Poinar HN, Pääbo S, Vigilant L. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol. Ecol. 2004; 13:321–335. [PubMed: 14717890]

Thalmann O, Serre D, Hofreiter M, Lukas D, Eriksson J, Vigilant L. Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. Mol. Ecol. 2005; 14:179–188. [PubMed: 15643961]

Thalmann O, Fischer A, Lankester F, Pääbo S, Vigilant L. The complex evolutionary history of gorillas: Insights from genomic data. Mol. Biol. Evol. 2007; 24:146–158. [PubMed: 17065595]

Thalmann O, Wegmann D, Spitzner M, Arandjelovic M, Guschanski K, Leuenberger C, Bergl RA, Vigilant L. Historical sampling reveals dramatic demographic changes in western gorilla populations. BMC Evol. Biol. 2011; 11:85. [PubMed: 21457536]

Thomas R, Zischler H, Päävo S, Stoneking M. Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. Hum Biol. 1996; 68:847–854. [PubMed: 8979460]

Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc. Bioinformatics. 2002; 2(Unit 2.3)

Triant DA, DeWoody JA. The occurrence, detection, and avoidance of mitochondrial DNA translocations in mammalian systematics and phylogeography. J. Mammal. 2007; 88:908–929.

Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AE, Carroll ML, Nguyen SV, Walker JA, Prasad BVR, Reddy G, Das PK, Batzer MA, Jorde LB. Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. Genome Res. 2003; 13:1607–1618. 2003. [PubMed: 12805277]

Wharton, D. Chicago Zoological Society. Brookfield, IL, USA; 2007. North American studbook for the western lowland gorilla (*Gorilla gorilla gorilla*).

Williams ST, Knowlton N. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. Mol. Biol. Evol. 2001; 18:1484–1493. [PubMed: 11470839]

Highlights

- Sixty seven putative gorilla-specific numts were identified and mapped.
- No evidence that numts are more prevalent in gorillas relative to other great apes.
- Three insertional polymorphisms with utility as molecular markers are described.
- Successful differentiation of mitochondrial and numt sequences.
- Evidence of past hybridization of eastern and western gorillas.

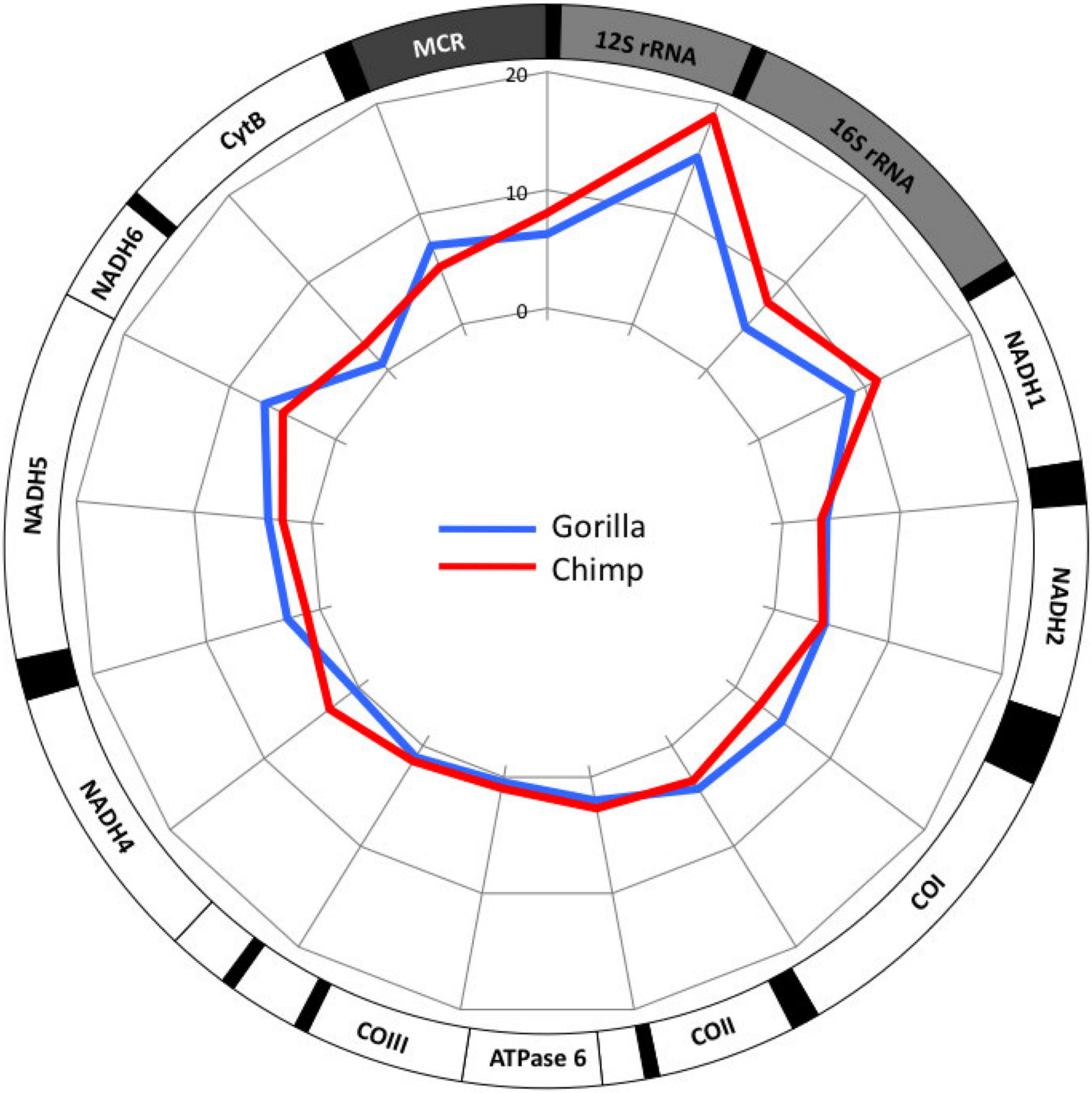


Figure 1.

Number of gorilla and chimpanzee BACs identified with radiolabeled 40bp probes designed from the mitochondrial genome, normalized by the redundancy of each library. Probes were pooled and hybridized in 17 groups; the normalized number of BACs identified is plotted at the midpoint of the location of the four probes. Unlabeled black boxes on the mitochondrial map indicate tRNA genes.

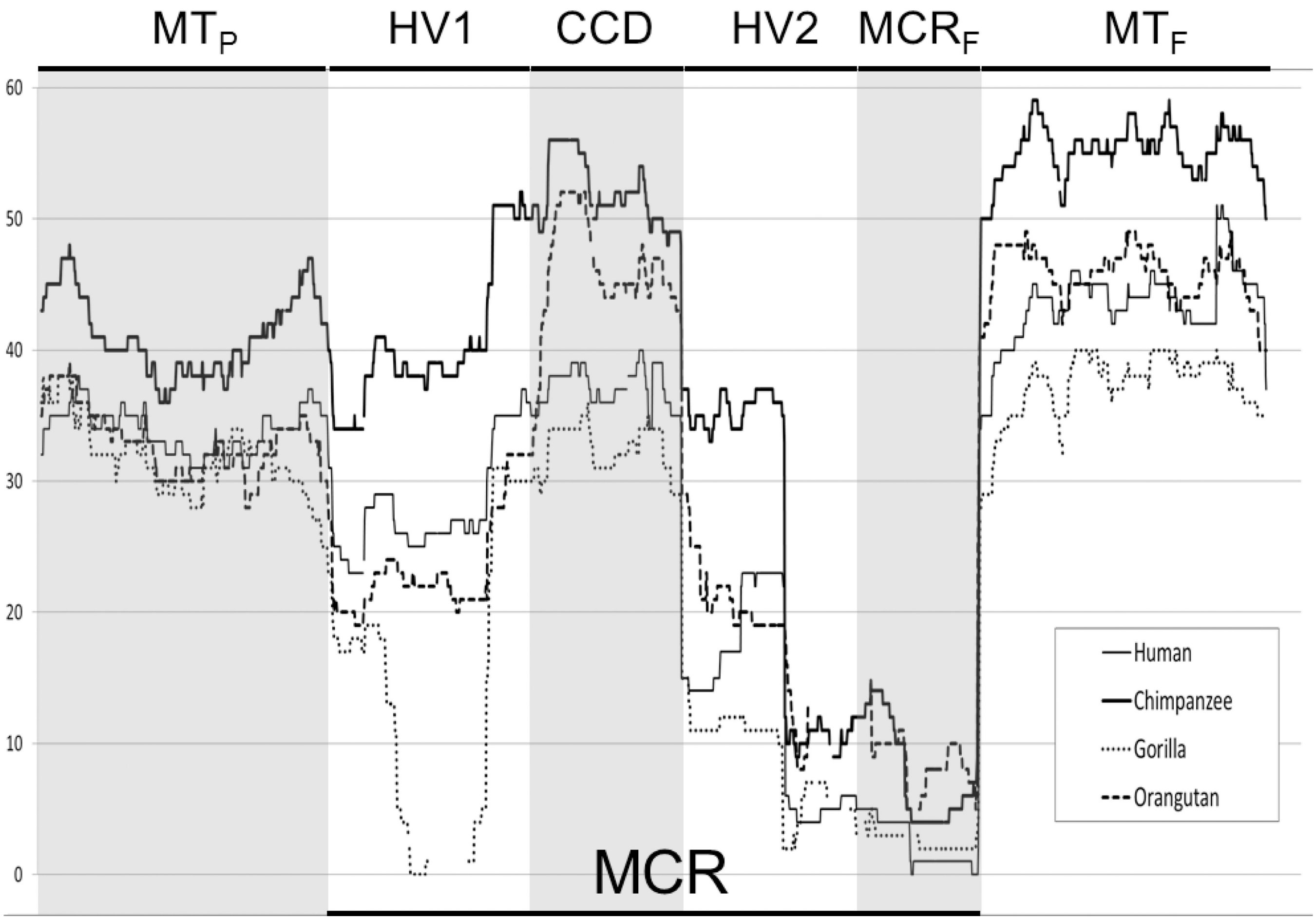


Figure 2.

Absolute number of numts per site in the four MCR sub-domains and 500bp flanking regions derived from BLAST searches in the reference genomes of humans, chimpanzees, orangutans and gorillas. The four MCR sub-domains are defined as HV1 (the first hypervariable region), HV2 (the second hyper-variable region), CCD (the conserved central sub-domain) and MCR$_F$ (the sub-domain proximal to the phenylalanine tRNA gene). The two 500bp flanking regions are defined here as MT$_P$ (the genes for tRNA of proline and threonine and 32% of the cytochrome *b* gene) and MT$_F$ (the gene for phenylalanine tRNA and 45% of the 12S rRNA gene). The observed patterns are significantly different among taxa (Kruskal-Wallis χ^2= 1679.009, d.f. = 3, p < 0.001).

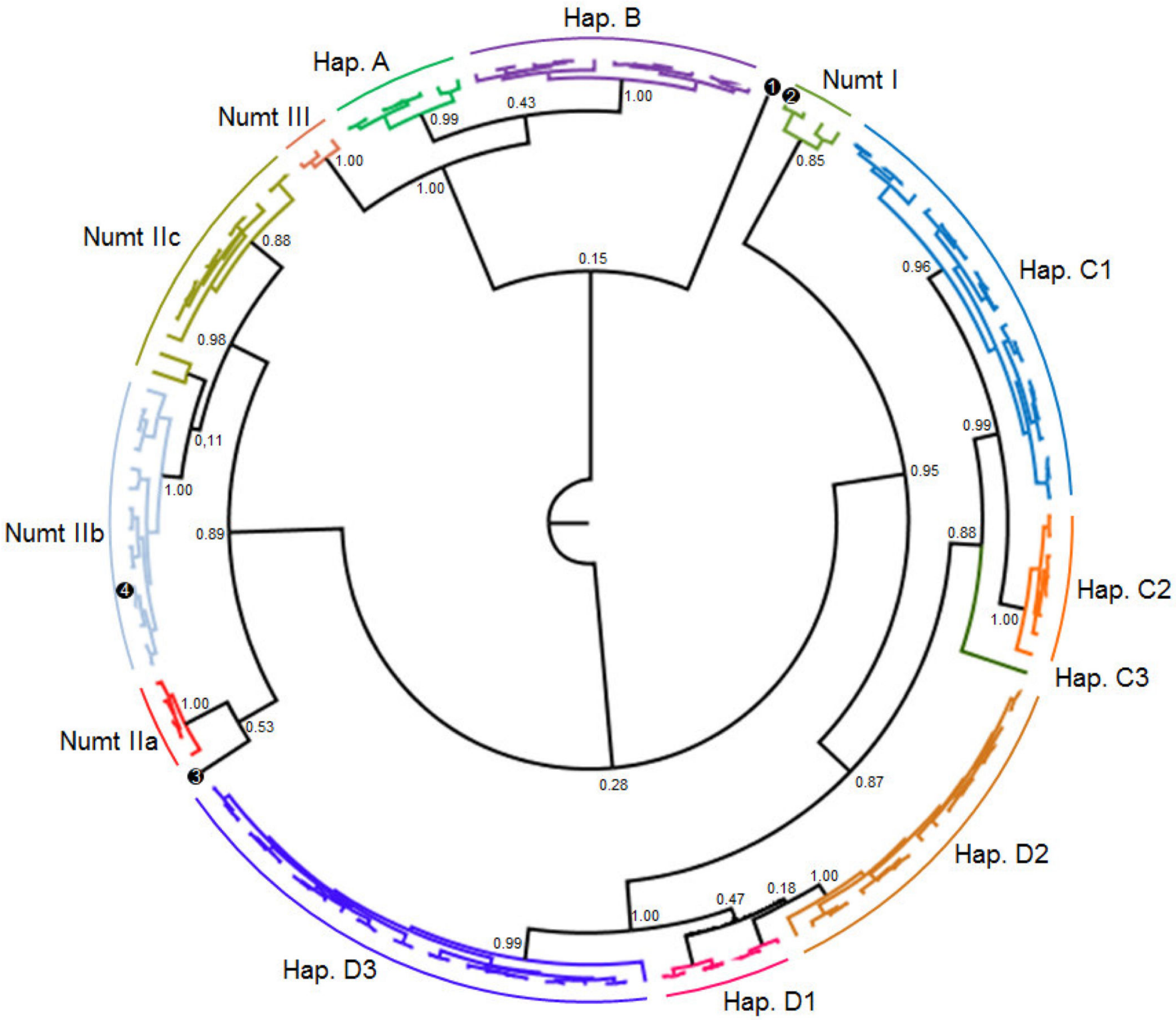
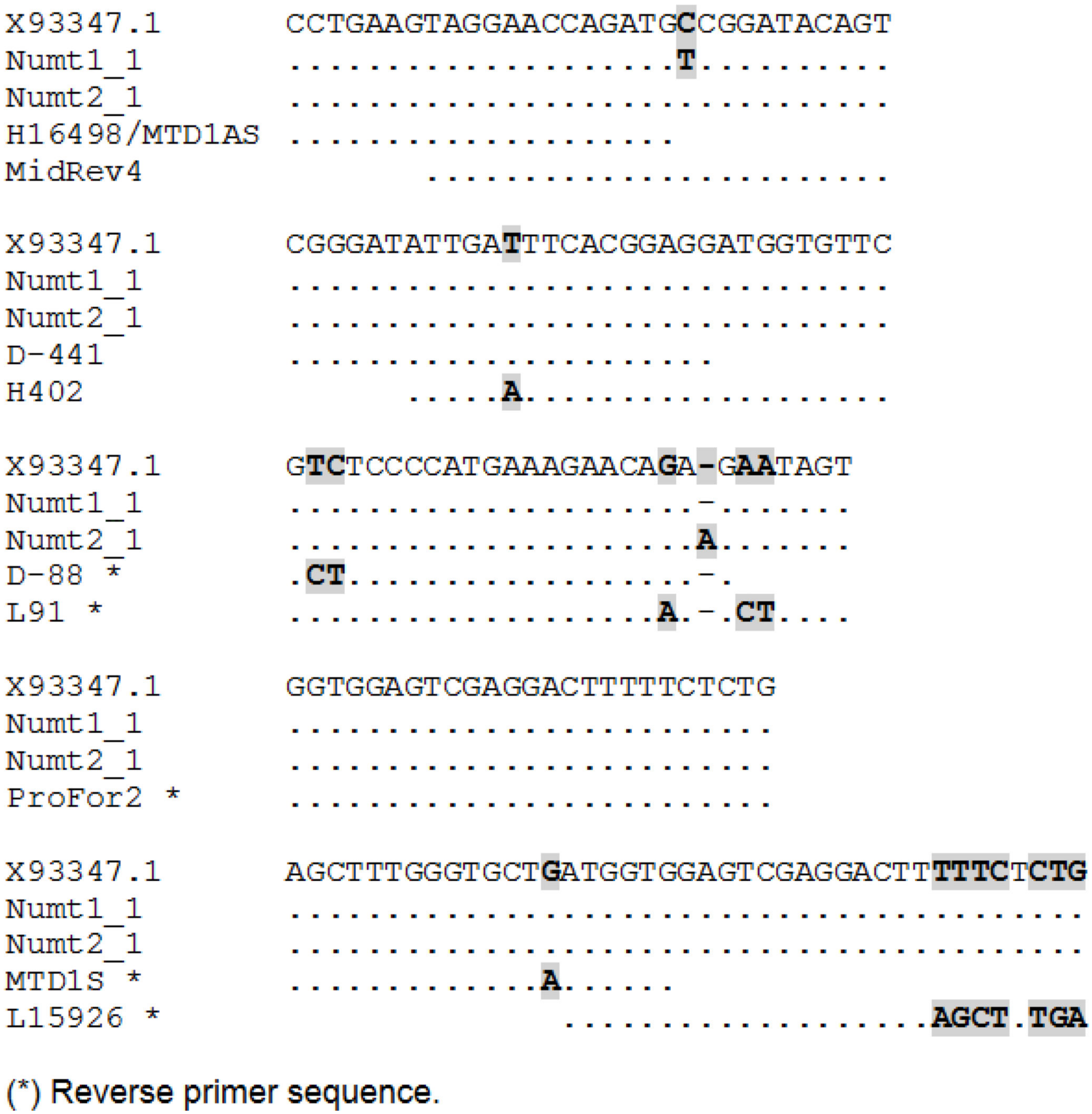


Figure 3.

Bayesian phylogeny of 102 HV1 and 42 numt gorilla sequences assuming a GTR+G model of nucleotide substitution (alpha = 0.5) with an exponential relaxed molecular clock. Posterior probabilities are shown adjacent to each node. Mitochondrial haplogroups A and B are restricted to east gorillas (*G. beringei*), whereas haplogroups C 1–3 and D 1–3 are only found in western gorillas (*G. gorilla*) (Anthony et al., 2007a). Bullets denote the location of the four mapped gorilla-specific numts containing the HV1 (1. Go9_5746; 2. Go2b_2500; 3. Go5_40; 4. Go1_1300).



```
X93347.1        CCTGAAGTAGGAACCAGATGCCGGATACAGT
Numt1_1         ....................T..........
Numt2_1         ...............................
H16498/MTD1AS   ....................
MidRev4                ........................

X93347.1        CGGGATATTGATTTCACGGAGGATGGTGTTC
Numt1_1         ...............................
Numt2_1         ...............................
D-441           ......................
H402                  .....A...................

X93347.1        GTCTCCCCATGAAAGAACAGA-GAATAGT
Numt1_1         .....................-.......
Numt2_1         .....................A.......
D-88 *          .CT..................-.
L91 *           ...................A.-.CT....

X93347.1        GGTGGAGTCGAGGACTTTTTCTCTG
Numt1_1         .........................
Numt2_1         .........................
ProFor2 *       .........................

X93347.1        AGCTTTGGGTGCTGATGGTGGAGTCGAGGACTTTTTCTCTG
Numt1_1         .........................................
Numt2_1         .........................................
MTD1S *         .............A......
L15926 *                      ...................AGCT.TGA

(*) Reverse primer sequence.
```

Figure 4.

Alignment of 5’-3 ’sequences from the gorilla mtDNA genome (X93347.1), gorilla numts containing the entire HV1 region (Go1_1300 and Go2b_2500), and primers used by previous authors to amplify this region in the mitochondrial genome.

Table 1

Description of the seven mapped gorilla-specific numts derived from the MCR.

Numt name	Isolation method	Numt Size	Mt position (X93347)	Sample Size	Polymorphic Status
Go1_1300	BAC screening BLAST (Trace files & current version)	1,347	14806–16150	62	Yes
Go2b_2500	BAC screening BLAST - Trace files	2,497	141–16412; 1–14060	58	Yes
Go5_40	A-PCR	456	15530–15993	56	Yes
Go9_5746	BLAST (Contigs & current version)	447 *	15615–15904; 8430–8466;	10	Fixed
Go11_188	BLAST - Trace files	2,420 **	15788–16412; 1–1720	14	Fixed
Go5_30	BLAST current version	1,560 **	14244–14588; 15780–15808	1	Untested
Go2b_43	BLAST current version	54	16172–16214	1	Untested

[*] This numt contains a fragment of 90bp lost from the gorilla mitochondrial genome.

[**] An internal portion of this numt remains to be sequenced. Size predicted from the alignment with the reference mitochondrial genome.

Table 2

Proportion of individuals (chromosomes in parentheses) possessing gorilla-specific numt polymorphisms.

Numt name	C1	C2	C3	D2	D3	Total
Go1_1300	3/4	0/3	1/3	0/3	0/4	15/17
Go2b_2500	5/5 (7/10)	0/3 (0/6)	1/3 (1/6)	0/3 (0/6)	3/4 (4/8)	9/18 (12/36)
Go5_40	6/6 (7/12)	3/3 (4/6)	0/3 (0/6)	3/3 (4/6)	2/3 (3/6)	14/18 (18/36)

See Table S5 for a detailed description of gorilla haplogroup identity and numt genotypes.