Published in final edited form as:
J Mol Evol. 2009 June ; 68(6): 688–699. doi:10.1007/s00239-009-9243-4.

A comparative approach reveals differences in patterns of numt insertion during hominoid evolution

M.I. Jensen-Seaman[1,*], **J.H. Wildschutte**[1,2], **I.D. Soto-Calderón**[3,4], and **N.M. Anthony**[3]

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

[2]Department of Molecular Biology & Microbiology, Tufts University, 136 Harrison Ave., Boston MA 02111, USA.

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

[4]Biology Institute, University of Antioquia, AA.1226, Medellín, Colombia.

Abstract

Nuclear integrations of mitochondrial DNA (numts) are widespread among eukaryotes although their prevalence differs greatly among taxa. Most knowledge of numt evolution comes from analyses of whole genome sequences of single species, or more recently from genomic comparisons across vast phylogenetic distances. Here, we employ a comparative approach using human and chimpanzee genome sequence data to infer differences in the patterns and processes underlying numt integrations. We identified 66 numts that have integrated into the chimpanzee nuclear genome since the human-chimp divergence, which is significantly greater than the 37 observed in humans. By comparing these closely related species, we accurately reconstructed the pre-integration target site sequence, and deduced nucleotide changes associated with numt integration. From over 100 species-specific numts, we quantified the frequency of small insertions, deletions, duplications, and instances of microhomology. Most human and chimpanzee numt integrations were accompanied by microhomology and short indels of the kind typically observed in the nonhomologous end-joining pathway of DNA double-strand break repair. Human-specific numts have integrated into regions with a significant deficit of transposable elements, while the same was not seen in chimpanzees. From a separate dataset, we also found evidence for an apparent increase in the rate of numt insertions in the last common ancestor of humans and the great apes using a PCR-based screen. Lastly, phylogenetic analyses indicate that mitochondrial-numt alignments must be at least 500bp, and preferably greater than 1kb in length, in order to accurately reconstruct hominoid phylogeny and recover the correct point of numt insertion.

Keywords

Numt; Hominid; Hominoid; Evolution; Mitochondria; Phylogenetic; Ape

Introduction

Nuclear integrations of mitochondrial DNA (numts) are fragments of the mitochondrial genome that have incorporated into germline nuclear DNA (nDNA), and have been reported in animals, plants, and fungi (Thorsness and Fox 1990; Zischler et al. 1995a; Blanchard and

[*]To whom correspondence should be addressed. seamanm@duq.edu.

Schmidt 1996; Zhang and Hewitt 1996; Bensasson et al. 2001; Leister 2005). They are usually referred to as pseudogenes, although there are a handful of instances where they may have been exonized in a few species (Noutsos et al. 2007), including a human-specific numt that inserted into a 3’-UTR (Ricchetti et al. 2004). Although nonfunctional, numts offer a model for early eukaryotic evolution in which hundreds of genes are believed to have retained function following their migration from the proto-mitochondrial endosymbiont to the nuclear genome (Margulis 1970; Andersson et al. 2003; Lang et al. 1999; Blanchard and Lynch 2000). Numts are commonly believed to “fossilize” following their integration into the nuclear genome; that is, the nuclear translocated mitochondrial copy is more likely to resemble the ancestral mitochondrial haplotype at the time of its insertion than its modern mitochondrial counterpart due to the much lower mutation rate in the nuclear genome (Zischler et al., 1995a). As such, numts offer interesting opportunities for studies of mitochondrial DNA evolution. Numts can also be problematic if mistaken for authentic mitochondrial DNA (mtDNA), potentially confounding interpretations in wildlife genetics, forensics, ancient DNA, or medical studies (van der Kuyl et al. 1995; Zischler et al. 1995b; Wallace et al. 1997; Bensasson et al. 2001; Jensen-Seaman et al. 2004; Anthony et al. 2007).

Although taxonomically widespread, the prevalence of numts varies tremendously among species, suggesting that the processes of numt integration may change over time and/or differ between taxa. Analyses of complete genome sequences have shown numts to be present in all mammals examined so far, with large numbers reported in humans, chimpanzees, and cats, but fewer found in mice and rats (Mourier et al. 2001; Tourmen et al. 2002; Woischnik and Moraes 2002; Richly and Leister 2004; Antunes et al. 2007). Honeybees display an unusually large number of numts (Behura 2007; Pamilo et al. 2007), while there are few to none in *Drosophila* and *Anopheles* (Richly and Leister 2004). There has been some controversy over the presence of numts in any fish species, but for now it appears that at least the fugu genome is devoid of them (Antunes and Ramos 2005; Venkatesh et al. 2006).

Despite this increasing awareness of the prevalence of numts in many taxa, comparisons of numt distributions among relatively closely related species are rare (but see Krampis et al. 2006; Hazkani-Covo and Covo 2008). One approach to identifying differences in insertion rates between species, and especially in locating temporal fluctuations in numt insertion rates, has been to infer the point of insertion using a phylogeny-based approach. In primates this has led to the suggestion that a burst of insertions occurred near the time of the divergence between Old World and New World monkeys (Bensasson et al. 2003; Hazkani-Covo et al. 2003; Gherman et al. 2007). Another study suggested that humans may be experiencing a very recent increase in numt integration (Ricchetti et al. 2004), although that conclusion was based on the assumption that neutral alleles in the human populations are expected to have a coalescence time within the last 100,000 years. Within great apes, it has been suggested that the frequency of numts is elevated in gorillas, although these observations are limited to the mitochondrial D-loop (Jensen-Seaman et al. 2004; Thalmann et al. 2004; Anthony et al. 2007). Whether this claim can be substantiated or not awaits a more detailed comparison of insertion rates between taxa and across the entire mitochondrial genome.

Numts do not appear to show a preference for specific target sequences, as has been found for retrotransposable elements (Cost et al. 2002), although whether they integrate truly randomly remains an open question. Large-scale analyses of human numts, made possible by the complete human genome sequence, suggested that the immediate flanking regions (~15bp) contained fewer transposable elements (TE) than expected by chance while slightly more distal regions (15–150bp) contained more TEs than expected (Mishmar et al. 2004). The opposite conclusion was reached by Gherman et al. (2007) who found a reduction in the

TE content of the first 150bp of sequence flanking human numts, with this reduced TE content extending across at least 1kb away from the numt. With respect to any sequence composition preference for human numt insertions, most studies have not found a strong pattern. Exceptions include the suggestion that numts are preferentially found in regions with a different GC-content than that of the surrounding chromosomal G-band (Mishmar et al. 2004), another being the recent observation that human numts prefer low GC-content isochores (Lascaro et al. 2008). Finally, it has been reported that older numts tend to be found outside genic regions, while more recent human-specific numts preferentially insert within introns (Ricchetti et al. 2004).

Numts are believed to integrate predominantly, or perhaps exclusively, via DNA-mediated transfer during the use of nonhomologous end-joining (NHEJ) repair of double-stranded breaks (Blanchard and Schmidt 1996; Ricchetti et al. 1999; Hazkani-Covo and Covo 2008). In contrast to repair via homologous recombination, classical NHEJ does not seek out truly homologous sequences as templates for repair, but typically uses short (1–4bp) stretches of sequence identity, or "microhomology", to facilitate end-joining. NHEJ is inherently error-prone, commonly involving short insertions and deletions at the repair site, with some of these insertions deriving from the fill-in of staggered double-strand breaks (Roth et al. 1985). The inference of microhomology, insertions, and deletions requires knowledge of both the target sequence prior to numt integration as well as the post-integration sequence. For this reason, only a few naturally occurring numt junctions had been examined (Zischler et al. 1995a; Ricchetti et al. 2004), prior to the recent work by Hazkani-Covo and Covo (2008), who demonstrated the preponderance of NHEJ repair with microhomology, although with a reduced frequency of deletions relative to experimental systems, indicating that numts may help reduce the deleterious effects of deletions during double-stranded break repair.

Here, we employ a comparative genomic approach to address several questions related to numt integrations in primates. A complete genome assembly exists for humans and chimpanzees, whose nuclear DNA differs by approximately 1% (CSAC 2005), making these species ideal for the comparative study of numts in closely related taxa. This study is divided into four main objectives. Firstly, we use closely related species as proxies for the pre-integration site to accurately determine the extent of the numt insertion, and to quantitatively infer the presence of microhomology and indels in over 100 species-specific numts. Secondly, we use these comparative data to test for an overabundance or deficit of transposable elements near numt insertions. Thirdly, we use a PCR-based assay to determine the likely time of insertion of a range of numts across the entire ape phylogeny to test whether numt insertions are uniform through time. Finally, we assess how well these insertion times can be recovered by computationally based phylogenetic methods.

Materials and Methods

Identification of human and chimpanzee numts

The human mitochondrial genome sequence (NC_001807; Ingman et al. 2001) was aligned to the human nuclear genome (March 2006 assembly; NCBI build 36.1) using the *blastn* program of locally installed *BLAST* (Altschul et al. 1990), with an e (Expect) value of 10. Similarly, the chimpanzee mtDNA sequence (NC_001643; Horai et al. 1995) was *BLAST*ed to the chimpanzee genome (March 2006 assembly; build 2, v.1). Hits to "chromosome unknown" or "random" chromosome contigs were ignored. Since *BLAST* tends to fragment contiguous matches interrupted by more diverged sequences (Jareborg et al. 1999) hits within 1kb were automatically grouped into a single hit. No attempt was made to exhaustively identify all, particularly highly divergent, numts. For each hit, the candidate human numt, the corresponding portion of the human mtDNA, and 500bp of the left and right flanking human nuclear sequence were compared to the chimpanzee genome using

BLAT (Kent 2002), followed by visual inspection to determine whether the numt was human-specific or present in both species. The same process was carried out for putative chimpanzee numts, using *BLAT* to compare to the human genome in order to identify chimpanzee-specific numts (Supplemental Figure 1). Comparing the putative numt region to a closely related species to confirm that an insertion indeed took place permits use of a low stringency *BLAST* search, increasing the chance of identifying short and/or more diverged numts, while eliminating false positive *BLAST* hits. Differences in the number of species-specific numts between humans and chimpanzees were tested with a G-test (Sokal and Rolf 1995). For species-specific numts, the homologous nuclear genomic regions including the numt and 50bp on either side from both species, along with the homologous mitochondrial region plus 50bp on either side, were aligned locally with *MUSCLE* (Edgar 2004) and examined by eye to determine the extent of microhomology and to identify small insertions, deletions, and duplications that apparently occurred concomitantly with the integration of mtDNA.

Transposable element content of pre-integration sites

We used the chimpanzee genomic sequence as the proxy for the pre-integration site for human-specific numts. We specified the point of numt insertion in the homologous chimpanzee sequence as the base pair where the two homologous human flanking sequences meet in the absence of the numt, or the midpoint of the short intervening gap if the human homologous flanking sequences did not quite meet when aligned to the chimpanzee genome with *BLAT* (see Supplemental Figure 1). The same criteria were applied to the delimitation of chimpanzee-specific numts, using *BLAT* to determine their presence or absence in the human genome (Supplemental Figure 1). From this point, ten nonoverlapping windows of 100bp at increasing distances (0–1kb) from either side of each numt insertion point were extracted from the pre-integration cross-species proxy sequence and analyzed for repetitive element content using locally installed *RepeatMasker*, with the -e wublast option (Smit et al.). To test for significance we generated a distribution of similar data taken from throughout the genome. For this, we randomly selected a number of genomic locations equivalent to the number of species-specific numts (i.e., 37 random locations for human, 66 for chimpanzee), and again extracted 100bp windows for analysis with *RepeatMasker* as described above. This random selection was repeated 10,000 times using a custom *perl* script to create a distribution with which to compare our observed values.

Determining the point of insertion of hominoid numts using cross-species PCR

In order to empirically determine the point of numt insertions more broadly throughout hominoid evolution, we *BLAT*ed all putative human numts identified in the original mtDNA-to-nDNA *BLAST* search and their flanking sequences to the rhesus macaque genome sequence assembly (Jan. 2006 assembly) to identify those numts not present in the macaque genome (i.e., numts that integrated since the cercopithecoid-hominoid split). We also excluded numts shown to be human-specific since we already determined their time of insertion computationally (see above). From the human-macaque sequence alignment, PCR primers were then designed from conserved regions flanking each numt. We also excluded very large numts or those flanked by too many repetitive elements to be able to design reliable primers. PCR amplification was attempted for 50 of these loci using genomic DNA samples from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), gibbon (*Hylobates lar*), and macaque (*Macaca mulatta*). PCR primer sequences and reaction conditions are available upon request. Presence or absence of each numt for each species was scored based on the expected size of the amplification product with and without the numt. Of the 34 numts successfully placed (out of 50 attempted) in this manner, eight PCR products were sequenced in whichever of the above species is most closely related to human, but found to lack that particular numt, in order to confirm its absence. The observed phylogenetic distribution of nuclear integrations was compared to an expected distribution assuming equal rates of numt insertion across all branches leading to humans, and evaluated with the χ^2 test. For this we used hominoid divergence dates from Raaum et al. (2005) and Stauffer et al. (2001).

Phylogenetic analysis

We assessed the ability of tree-based methods to correctly infer the point of insertion of 40 numts in the hominoid phylogeny by conducting phylogenetic analysis on each of the 34 loci that had been placed empirically by cross-species PCR (see above), along with six more determined to be human specific through comparison to the chimp genome (see above). To do this, mtDNA sequences from human, chimp, gorilla, orangutan, gibbon, and macaque were first aligned without their corresponding nuclear copy using *Clustal X* (Larkin et al. 2007). Maximum parsimony (MP) and maximum likelihood (ML) phylogenetic analysis were then carried out to identify those mitochondrial datasets (without the numt) with sufficient signal to recover the accepted primate phylogeny (Goodman et al. 1998). MP and ML analyses were carried out in *PAUP* v4.4* (Swofford 2002) using starting trees obtained from the stepwise addition and neighbor-joining method options, respectively. Heuristic searches were conducted with the tree-bisection-reconnection (TBR) method and branch support was obtained from 100 bootstrap replicates of the data. Starting evolutionary parameters for ML analyses were obtained from the Akaike Information Criterion option in *MODELTEST* (Posada and Crandall 1998). The eight mtDNA datasets that correctly recovered the recognized hominoid phylogeny were then also subject to Bayesian phylogenetic analysis to check for consistency across methods. Bayesian analysis was carried out using the *BEAUTi/BEAST v1.4.6* package (Drummond and Rambaut 2007) using tree priors based on the accepted tree topology and approximate divergence times between taxa using an uncorrelated, lognormal clock to allow for among lineage rate variation, and with starting evolutionary parameters from *MODELTEST*. A Monte Carlo Markov Chain (MCMC) of 50 million steps in length was run with a sampling interval of every 1000 steps. The appropriate burn-in period (10%) was determined from visual inspection of output in the program *TRACER v1.4* (Drummond and Rambaut 2007). Using these search options, all parameter values could be estimated from effective sample sizes of 100 or more. For those eight cases where mtDNA sequences were able to recover the accepted primate phylogeny, the point of numt insertion in the phylogeny was then tested in a phylogenetic framework by imposing a backbone constraint on the underlying primate phylogeny and using ML to find the most likely placement of the numt insertion.

Results

Chimpanzees have more recent numts than human

BLAST searches of the human genome assembly using the human mitochondrial genome sequence as query and grouping hits less than 1kb apart, yielded a total of 519 putative loci. Similarly, BLAST searches using the chimpanzee mtDNA sequence against the chimpanzee genome assembly yielded 579 hits. We compared each putative numt, left and right flanking regions, and the homologous mtDNA domain from the human to the chimpanzee nuclear genome or vice-versa using BLAT, followed by manual inspection of the BLAT visualization to score each numt as either species-specific or shared between human and chimpanzee (Supplemental Figure 1). This approach identified 37 human-specific numts, and 66 chimpanzee-specific numts (Supplemental Tables 1 and 2). The greater number of chimpanzee numts is significant (G_{adj}=8.237; p<0.005, df=1). The mean length of the chimpanzee-specific numts is larger than that of human (chimp mean = 554bp, human mean = 321bp), although this difference is not significant (p=0.37, t-test with Welch's correction for unequal variances).

Use of outgroup to define junctions and infer pre-integration state

The use of BLAST identifies numts by sequence similarity to the mitochondrial genome, but does not necessarily accurately delineate the exact boundaries of the numt. Here, we define the numt as the stretch of nucleotides that was inserted into the genome compared to the inferred pre-integration site, which in many cases is slightly different from the stretch of nucleotides that share significant similarity with mtDNA in a BLAST search for two main reasons. First, several additional bases are commonly added during the insertion, often from small duplications and direct flanking repeats (Figure 1). Second, short stretches of similarity are frequently found between the mtDNA and the pre-integration sequence, termed “microhomology” (Figure 1). These two features cause underestimation and overestimation, respectively, of the length of the numt. Of the 37 human-specific numts we identified, 23 possess at least some nucleotides inserted without similarity to mtDNA, 18 show deletions of nuclear DNA upon insertion, while nearly all show some microhomology at one or both flanks (see Supplemental Figure 2 for alignments of all human-specific numts). Since nucleotides appear to be commonly deleted from the nuclear target site during the process of integration, as inferred from human-chimp alignments (Figure 1), the flanking sequences, even when properly defined, do not offer the best estimation of the pre-integration sequence (i.e., “target site”). We were unable to accurately align the human and chimpanzee nuclear sequences at the point of insertion for two chimpanzee-specific numts due to complex rearrangements, leaving 64 chimpanzee numts for the integration site analyses below (see Supplemental Figure 3 for alignments of all chimpanzee-specific numts).

Microhomology and indels at integration sites

Comparing the chimpanzee sequence (representing the pre-integration state) to the human mtDNA sequence immediately flanking the numt insertion point reveals that most (35 of 37) of the numt insertions occur in the presence of microhomology at one or both of the numt-nuclear junctions. (Figure 2a–c; Supplemental Figure 2). Similarly, 45 of the 64 chimpanzee-specific numts possess microhomology at one or more of the junctions (Figure 2; Supplemental Figure 3). The distribution of lengths of observed microhomology for both species is very similar to that reported for other types of DNA integration into eukaryotic DNA, with most microhomology being limited to 1–4bp, although one of the human-specific numts did contain a 10bp stretch of identical nucleotides (Figure 2d). These estimates are conservative in that only perfect uninterrupted matches were counted. We are operationally defining microhomology as exact matches regardless of length in order to compare with other data sets (Figure 2d); this does not necessarily imply that the matches were used in the numt integration process and may occur simply by chance.

The majority (23 of 37) of human-specific numts contain insertions of nucleotides that do not appear to have been derived from mtDNA, nor were present prior to insertion—again using the chimpanzee nuclear sequence as an estimate of the pre-integration site. These insertions are between 1–37bp (mean = 7.2bp; median = 5bp), examples of which are shown in Figure 3a–f. Almost all of these insertions can be explained as being derived from one of three sources: i) flanking direct repeats, ii) tandem direct repeats, and iii) tandem inverted repeats. Considering only insertions of at least 4bp, eight of 37 human-specific numts contain flanking direct repeats of 4–14bp, always found precisely at the junction between the numt and flanking DNA (Figure 3a,b). Four cases of tandem direct repeats of 5–22bp were found, only one of which was truly tandem in that the duplicated nucleotides immediately follow the source nucleotides (Figure 3c), while the other three included between two and nine nucleotides spacing the duplication (e.g., Figure 3d). Interestingly, and most certainly anecdotally, the six nucleotides at the left flank of one particular numt insertion that includes a tandem duplication are a perfect complement to the mitochondrial sequence (Figure 3c, underlined). Finally, five of the 37 human-specific numt insertions were accompanied by inverted repeat sequences of between 8–12 nucleotides, spaced by 1–6bp (e.g., Figure 3e,f). Eighteen of the 37 human-specific numts show evidence for deletions of between 1–157bp in the pre-integration sequence using the chimpanzee sequence as proxy. As with humans, the majority (40 of 64) of chimpanzee-specific numts contain insertions of 1–60bp (mean = 5.4bp; median = 1.5bp), including eight flanking direct repeats (4–14bp), nine tandem direct repeats (4–12bp), and four tandem inverted repeats (7–15bp). Alignments of all human-specific and chimpanzee-specific numts with their pre-integration sequence are shown in Supplemental Figures 2 and 3.

Deficit of transposable elements in human numt flanks

Taking the 37 human-specific numts, we used the homologous insertion point in the chimpanzee genome to represent the pre-integration state. There is a reduced density of transposable elements in the first two 100bp windows on either side of the insertion point (Figure 4a). This includes a reduction in all four categories of transposable elements compared to the genome-wide average (1–100bp: SINEs 8.39%, LINEs 9.49%, LTRs 2.09%, DNA-based mobile elements 2.03%; 101–200bp: SINEs 7.97%, LINEs 9.55%, LTRs 3.25%, DNA 1.84%). The total proportion of transposable elements in the 100bp windows immediately flanking the numt insertions (22.00%) falls within the most extreme five percent of a distribution made from 10,000 randomly generated datasets (4.46 percentile; Figure 4b), while the next 100bp (100–200bp away from the insertion point) nearly does (5.32 percentile). When chimpanzee-specific numts are examined in analogous fashion (using the human genome as the outgroup sequence to define the pre-integration state), no significant decrease in transposable element content is seen at or near the point of numt insertion.

Determination of numt insertion time

We determined the time of insertion of 34 numts using cross-species PCR. Most branches of the hominoid phylogenetic tree have about the same or fewer insertions than the expected number based on estimated divergence times between internal nodes, with the exception of the stem hominid (great ape and human) lineage that follows the split with gibbons (Figure 5). This distribution of numts across the primate phylogeny differs significantly than expected (χ^2=9.78; p=0.021; df=3), driven almost entirely by the excess of numt insertions in the stem hominid, using the Bayesian estimates of divergence dates reported by Raaum et al. (2005). Our result of a significant excess of numt insertions in the hominid ancestor is highly dependent on the estimated divergence dates, which vary widely depending on methodology, datasets, and choice of fossil calibration; indeed, the observed uneven phylogenetic distribution is not significantly different than the expected uniform distribution when using dates estimated with ML on the same mitochondrial genome data (Raaum et al. 2005) (χ^2=6.88; p=0.076), nor with dates derived from nuclear data (Stauffer et al. 2001).

Taken together, MP and ML methods recovered the accepted hominoid phylogeny in a total of 8 out of 40 datasets. Of these, MP only recovered the topology correctly in four cases whereas ML recovered the topology successfully in seven cases. Of these eight instances, Bayesian analyses correctly recovered the same topology in seven cases, only one of which was not recovered in ML. The size of the alignments for these eight loci ranged from 149 to 2457 bp in length (mean = 1055.3; median = 676.5). In contrast, alignments that failed to recover the accepted hominoid topology ranged from 50 to 487bp in size (mean = 149.5; median = 135.0), with a significant difference in the two medians of these groups (Mann-Whitney U-test: p < 0.001). ML analysis with the backbone constraint imposed reliably assigned the numt to the correct position in only five of the eight cases where the underlying primate topology was recovered using one or more phylogenetic methods.

Discussion

Nonrandom numt insertion through time

It is well established that some species contain a greater number of numts than others (Bensasson et al. 2001). From available genome sequence data, it appears that *Homo*, *Arabidopsis*, and *Apis* have substantial numbers of numts, while *Drosophila*, *Anopheles*, *Fugu*, and *Gallus* do not (Richly and Leister 2004; Pamilo et al. 2007; Venkatesh et al. 2006; Pereira and Baker 2004). Most of the general conclusions that can be made from these comparisons, however, are across large phylogenetic distances and are seemingly explained by differences between broad taxonomic categories (e.g., fish vs mammals, plants vs animals, dipterans vs hymenopterans).

Here we take advantage of the complete genome sequence of two very closely related species to compare the rates of numt integration and to infer pre-integration target sequences. Chimpanzees have a significantly greater number of recent species-specific numts than humans. Numts can increase in frequency in two ways: greater mitochondria-to-nucleus transfer, or increased intranuclear post-insertion duplications (Hazkani-Covo et al. 2003). None of the chimpanzee-specific numts appear to be post-insertion duplicates of each other, and therefore they likely reflect an increased rate of transfer from the mitochondria.

It is difficult to explain why chimpanzees should have more numts, and we offer no definitive answers here. If numts are considered slightly deleterious mutations, as transposable elements usually are, all else being equal selection should more efficiently remove them from a larger population. However, chimpanzees have historically larger effective population sizes than humans (Jensen-Seaman et al. 2001; Yu et al. 2003; Burgess and Yang 2008), so in the absence of other factors we might expect to see more numts in humans which is not the case. It is also conceivable that chimpanzees and humans differ slightly in their double-strand break repair mechanisms, including overall efficiency or preference for one mechanism over another, that may be related to the likelihood of mtDNA being incorporated during repair. Nonhuman primates have substantially lower rates of cancer than humans, which may be in part due to genetic differences, including those at DNA repair enzymes (Puente et al. 2006). Alternatively, differences may exist between humans and chimpanzees in the cellular availability of degraded mtDNA in germline cells (see Richly and Leister 2004; Willett-Brozick et al. 2001). Chimpanzee sperm cells, and therefore potentially zygotes, may contain more mtDNA than their human counterparts, since the volume of the sperm midpiece—packed with mitochondria—is substantially greater in chimpanzees compared to humans (Anderson and Dixson 2002).

It is important to consider the possibility that the greater number of observed numts in chimpanzees may be an artifact of methodology or data quality. We believe that all of the chimp-specific numts identified here are true numt insertions, not spurious BLAST matches nor deletions in humans, since their identification was not only based on similarity to mtDNA but also the absence of precisely that sequence in the human genome at the homologous location (and vice-versa regarding human-specific numts). Furthermore, we do not believe that we are over-counting chimpanzee numts by including a single numt insertion split into two by the insertion of a TE; no two chimp numts are within 1Mb of each other. Concerning methodology, we note that our approach to identifying human numts did find a nearly identical set of numts as that recently described by others. Specifically, all 34 human-specific numts previously identified using a different approach (Hazkani-Covo and Graur 2007) were identified herein, along with three additional numts. We used a newer version of the human genome assembly, which may account for the additional three numts discovered. Similarly, all 27 human-specific numts described by Ricchetti et al. (2004) were found with our approach. The lower quality of the chimpanzee genome assembly might account for a greater number of identified numts, although it is not clear how. If anything, we might expect a smaller number to be found in a more fragmented, error-prone genome assembly. Indeed, Hazkani-Covo and Covo (2008) found substantially more numts in the build 2 version of the chimpanzee genome than Hazkani-Covo and Graur (2007) found in build 1 version using similar methods, suggesting that increasing assembly quality will only add to the number of recovered numts. We therefore believe that the greater number of observed numts in chimpanzees relative to humans reflects a true biological difference. Nonetheless, we do heed the admonition by Venkatesh et al. (2006) that all numts from shotgun sequenced genomes need to be empirically verified, and anticipate that future research will do so.

Using cross-species PCR on a subset of human numts, we observed an excess of numt insertions in the stem hominid branch leading to the human and great ape clade, with 11 of 34 hominoid-specific numts inserting along this short branch. The approximate time of these insertions, the mid-Miocene, saw an impressive adaptive radiation of apes with a greater species diversity than before or since. However, it should be emphasized that the increases in numt insertions in the stem hominid are strongly dependent on the accuracy of our estimated divergence dates. Although recent years have seen an explosion of sequence data used to date these events with the molecular clock, the fossil record is still woefully inadequate for accurate calibration (Jensen-Seaman and Hooper-Boyd 2008). We also note that this increase could be due in part to an ascertainment bias in that some numt loci failed to amplify in all species and were excluded from analyses. We cannot envision however, how this could have led to an increase the number of observed numt insertions at this point in the tree; if anything, we might expect a bias toward more recent events.

Several previous studies have suggested a non-uniform rate of numt insertions into the human genome, particularly in identifying a burst of insertions near the time of the split between Old World (catarrhine) and New World (platyrrhine) anthropoids, especially along the branch leading to all extant catarrhines (Hazkani-Covo et al. 2003; Bensasson et al. 2003) or even earlier (Gherman et al. 2007). It is, however, difficult to accurately determine the time of insertion of any numt with a purely computational phylogenetic approach for several reasons. First, it requires including mitochondrial and nuclear sequences in the same tree, and as such requires the evolution of these sequences to be modeled with the same parameters. It is widely known that nuclear and mitochondrial sequences have different mutation rates, different transition-transversion biases, and differ in the patterns of among-site rate heterogeneity—all critical variables in modeling sequence evolution. The difficulties with respect to accurately placing numts using phylogenetic methods were explored in depth by Bensasson et al. (2003), who demonstrated not only a strong dependency on which model of sequence evolution was used, but also a consistent reduction in the non-uniformity of numt insertions as increasingly realistic models were used. In this light, the more simplistic models used by Gherman et al. (2007) may explain, at least in part, their observation of an extremely strong temporal burst of numt insertions.

In trying to improve on the ability to phylogenetically place human numt insertions, we applied alternative approaches (MP, ML and Bayesian) to assess which mitochondrial data had sufficient phylogenetic information to recover the accepted primate tree. Only 8 out of 40 numts (20%) contained sufficient phylogenetic signal, indicating that a prior selection of data sets is necessary before attempting to infer numt insertion points. We doubt the reliability of any method to accurately place numts shorter than 500bp, reinforcing the need to complement phylogenetic analyses with wet lab techniques or comparative genomic studies when candidate numt loci are 500 bp or less.

Nonrandom numt insertion in genomic space

Human-specific numts preferentially integrate into regions of low transposable element density. More than 200bp away from the insertion point, this no longer holds true, largely due to the abundant SINEs several hundred nucleotides away. Previous studies examining the flanking regions of all human numts have presented conflicting results. Mishmar et al. (2004) reported a striking deficit of TEs within 15bp of the numt boundary, and an excess of TEs between 15–150bp from the boundary, for 247 human numts. In contrast, Gherman et al. (2007) described a reduced proportion of TEs across the entire first 500bp flanking 266 human numts, with a monotonically increasing TE frequency moving away from the point of insertion for at least 1000bp. Our data differ in being based on a smaller number of numts, but have the advantage of using a closely related species to accurately define the numt insertion point in the genome. In addition, since we are only examining very recent numts, the TE content of the chimpanzee genomic sequence likely represents the state at the time of insertion, since the vast majority of human TEs inserted prior to the human-chimp divergence; in contrast, examining much older numts likely also includes counting TEs that inserted after the numt. To summarize, our observation of a reduced proportion of TEs near the numt insertion point is broadly in agreement with Mishmar et al. (2004) and Gherman et al. (2007). However, we were unable to replicate the results of Gherman et al. (2007) in finding a continuously increasing proportion of TEs with increasing distance from the numt, even when using a dataset composed of the flanking sequences of all identifiable human numts (n=403; data not shown).

Although the pattern of numts inserting into low TE regions is clear, a mechanistic explanation is less forthcoming. TEs may induce conformational changes in DNA, which may make them less susceptible to double-strand breaks, or more likely to lead to correctly repairing such breaks when they occur without incorporating extranuclear DNA. TEs themselves integrate nonrandomly into the primate genome, with LINEs more often found in GC-poor areas of the genome, while SINEs show a preference for GC-rich genic regions (IHGSC 2001; Gasior et al. 2007). As such, it may be likely that the negative association between numts and TEs is due to a co-correlation with some other unknown factor. Finally, it may be that numts that have inserted into TEs may be more frequently removed from the genome via nonhomologous recombination with another copy of that TE, or other mechanisms to excise TEs.

While previous studies have identified microhomology between the target site and the mtDNA, these have been limited to only a handful of occurrences (Zischler et al. 1995a; Blanchard and Schmidt 1996; Ricchetti et al. 1999), prior to the recent study by Hazkani-Covo and Covo (2008). By the use of a closely related outgroup, we were able to accurately define the extent of microhomology present in over 100 recent human- and chimp-specific numt insertions. Nearly all numts contain some degree of microhomology, at least at one end. Our quantification of microhomology is conservative in that for human numts we are comparing modern human mtDNA with chimpanzee nDNA as proxies for the molecules present at the time of integration (and vice-versa for chimp-specific numts), and as such any mutations that have occurred since that time may be obscuring more substantial matches. This is especially relevant considering the high mutation rate of mtDNA. Also, short stretches of microhomology that occur in positions not immediately adjacent to the boundary of numt insertion were not considered in the quantitative analysis. These microhomologies, along with the common occurrence of small deletions and insertions of flanking repeats, tandem repeats, and inverted repeats, are the hallmarks of the NHEJ pathway of double-strand break repair (Varga and Aplan 2005). The occasional presence of longer stretches of microhomology (≥7bp), and recessed microhomology, may indicate the use of an alternative end-joining mechanism (Daley and Wilson 2005; Corneo et al. 2007; Decottignies 2007; Yan et al. 2007), but the data presented here suggest that this is rare, at least for mammalian numts. Close examination of the numt junctions, along with the outgroup for comparison, provides some hints as to the inexactness of this repair pathway, and confirms its description as “dirty” (Odersky et al. 2002). The reliance on short microhomology to complete the double strand break repair with NHEJ further indicates a nonrandom component to numt insertion. While this mechanism does not specifically target genomic locations based on long stretches of homology, it does suggest that there must be some portions of the genome where no sufficient microhomology could be found, leading to a failure to repair DNA damage and to cell death. As such, we only observe numts that originated in germline cells that were able to successfully complete the repair process.

Although the insertion of numts into the genome is often considered random, we show here that primate numts do not insert uniformly through time, nor randomly with respect to genomic location. While mechanistic explanations for these patterns remain elusive, the impending availability of several more hominoid genome sequences may likely provide clues. The use of comparative data to accurately determine the pre-integration site is shown here to be essential, and provides evidence that microhomology and small indels are frequently associated with integration, reinforcing the potential role of NHEJ in numt insertion.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to the National Institutes of Health for funding (R15 GM073682-01 to NMA), with additional funding from the Bayer School of Natural and Environmental Sciences at Duquesne (MJS, JHW). We thank Dr. Charles Bell (University of New Orleans) for his collaboration in the phylogenetic reconstructions. This manuscript has been substantially improved by thoughtful comments from two anonymous reviewers.

References

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

Anderson MJ, Dixson AF. Sperm competition: motility and the midpiece in primates. Nature. 2002; 416:496. [PubMed: 11932733]

Andersson SGE, Karlberg O, Canbäck B, Kurland CG. On the origin of mitochondria: a genomics perspective. Phil Trans R Soc Lond B. 2003; 358:165–179. [PubMed: 12594925]

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. 2007; 43:553–566. [PubMed: 17084645]

Antunes A, Ramos MJ. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. Genomics. 2005; 86:708–717. [PubMed: 16176867]

Antunes A, Pontius J, Ramos MJ, O'Brien SJ, Johnson WE. Mitochondrial introgressions into the nuclear genome of the domestic cat. J Hered. 2007; 98:4141–4420.

Behura SK. Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. Mol Biol Evol. 2007; 24:1492–1505. [PubMed: 17404397]

Bensasson D, Zhang D-X, Hartl DL, Hewitt GM. Mitochondrial pseudogenes: evolution’s misplaced witnesses. Trends Ecol Evol. 2001; 16:314–321. [PubMed: 11369110]

Bensasson D, Feldman MW, Petrov DA. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. J Mol Evol. 2003; 57:343–354. [PubMed: 14629044]

Blanchard JL, Schmidt GW. Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. Mol Biol Evol. 1996; 13:537–548. [PubMed: 8742642]

Blanchard JL, Lynch M. Organellar genes: why do they end up in the nucleus? Trends Genet. 2000; 16:315–320. [PubMed: 10858662]

Burgess R, Yang Z. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. 2008; 25:1979–1994. [PubMed: 18603620]

CSAC: The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005; 437:69–87. [PubMed: 16136131]

Corneo B, Wendland RL, Deriano L, Cui X, Klein IA, Wong SY, Arnal S, Holub AJ, Weller GR, Pancake BA, Shah S, Brandt VL, Meek K, Roth DB. Rag mutations reveal robust alternative end joining. Nature. 2007; 449:483–486. [PubMed: 17898768]

Cost GJ, Feng Q, Jacquier A, Boeke JD. Human L1 element target-primed reverse transcription *in vitro*. EMBO J. 2002; 21:5899–5910. [PubMed: 12411507]

Daley JM, Wilson TE. Rejoining of DNA double-strand breaks as a function of overhang length. Mol Cell Biol. 2005; 25:896–906. [PubMed: 15657419]

Decottignies A. Microhomology-mediated end joining in fission yeast is repressed by Pku70 and relies on genes involved in homologous recombination. Genetics. 2007; 176:1403–1415. [PubMed: 17483423]

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

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nuc Acids Res. 2004; 32:1792–1797.

Gasior SL, Preston G, Hedges DL, Gilbert N, Moran JV, Deininger PL. Characterization of pre-insertion loci of *de novo* L1 insertions. Gene. 2007; 390:190–198. [PubMed: 17067767]

Gherman A, Chen PE, Teslovich TM, Stankiewicz P, Withers M, Kashuk CS, Chakravarti A, Lupski JR, Cutler DJ, Katsanis N. Population bottlenecks as a potential major shaping force of human genome architecture. PLoS Genet. 2007; 3 e119.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol. 1998; 9:585–598. [PubMed: 9668008]

Hazkani-Covo E, Sorek R, Graur D. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J Mol Evol. 2003; 56:169–174. [PubMed: 12574863]

Hazkani-Covo E, Covo S. *Numt*-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet. 2008; 4 e1000237.

Hazkani-Covo E, Graur D. A Comparative analysis of *numt* evolution in human and chimpanzee. Mol Biol Evol. 2007; 24:13–18. [PubMed: 17056643]

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA. 1995; 92:532–536. [PubMed: 7530363]

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000; 408:708–713. [PubMed: 11130070]

IHGSC: International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

Jareborg N, Birney E, Durbin R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. 1999; 9:815–824. [PubMed: 10508839]

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

Jensen-Seaman, MI.; Hooper-Boyd, KA. Encyclopedia of Life Sciences (ELS). Chichester, UK: John Wiley & Sons; 2008. Molecular clocks: determining the age of the human-chimpanzee divergence.

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

Krampis K, Tyler BM, Boore JL. Extensive variation in nuclear mitochondrial DNA content between the genomes of *Phytophthora sojae* and *Phytophthora ramorum*. Mol Plant Microbe Interact. 2006; 19:1329–1336. [PubMed: 17153917]

Lang BF, Gray MW, Burger G. Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet. 1999; 33:351–397. [PubMed: 10690412]

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23:2947–2948. [PubMed: 17846036]

Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C, Attimonelli M. The RHNumtS compilation: features and bioinformatics approaches to locate and quantify human NumtS. BMC Genomics. 2008; 9:267. [PubMed: 18522722]

Leister D. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet. 2005; 21:655–663. [PubMed: 16216380]

Margulis, L. Origin of Eukaryotic Cells. New Haven, CT: Yale University Press; 1970. 349 p.

Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Hum Mutat. 2004; 23:125–133. [PubMed: 14722916]

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]

Noutsos C, Kleine T, Armbruser U, DalCorso G, Leister D. Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends Genet. 2007; 23:597–601. [PubMed: 17981356]

Odersky A, Panyutin IV, Panyutin IG, Schunck C, Feldmann E, Goedecke W, Neumann RD, Obe G, Pfeiffer P. Repair of sequence-specific ^{125}I-induced double-strand breaks by nonhomolougous DNA end joining in mammalian cell-free extracts. J Biol Chem. 2002; 277:11756–11764. [PubMed: 11821407]

Pamilo P, Viljakainen L, Vihavainen A. Exceptionally high density of NUMTs in the honeybee genome. Mol Biol Evol. 2007; 24:1340–1346. [PubMed: 17383971]

Pereira SL, Baker AJ. Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: implications for molecular inference of population history and phylogenetics. BMC Evol Biol. 2004; 4:17. [PubMed: 15219233]

Posada D, Crandall KA. Model Test: Testing the model of DNA substitution. Bioinformatics. 1998; 14:817–818. [PubMed: 9918953]

Puente XS, Velasco G, Gutiérrez-Fernández A, Bertranpetit J, King M-C, López-Otín C. Comparative analysis of cancer genes in the human and chimpanzee genomes. BMC Genomics. 2006; 7:15. [PubMed: 16438707]

Raaum RL, Sterner KN, Noviello CM, Stewart C-B, Disotell TR. Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. J Hum Evol. 2005; 48:237–257. [PubMed: 15737392]

Ricchetti M, Fairhead C, Dujon B. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature. 1999; 402:96–100. [PubMed: 10573425]

Ricchetti M, Tekaia F, Dujon B. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol. 2004; 2 e273.

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

Roth DB, Porter TN, Wilson JH. Mechanisms of nonhomolgous recombination in mammalian cells. Mol Cell Biol. 1985; 5:2599–2607. [PubMed: 3016509]

Smit, AFA.; Hubley, R.; Green, P. RepeatMasker Open-3.0. 1996–2004. <http://www.repeatmasker.org>.

Sokal, RR.; Rohlf, FJ. Biometry. 3d ed.. New York: W.H. Freeman and Company; 1995. 887 p.

Stauffer RL, Walker A, Ryder OA, Lyons-Weiler M, Hedges SB. Human and ape molecular clocks and constraints on paleontological hypotheses. J Hered. 2001; 92:469–474. [PubMed: 11948213]

Swofford, DL. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sunderland MA: Sinauer Associates; 2002.

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; 13:321–325. [PubMed: 14717890]

Thorsness PQ, Tox TD. Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*. Nature. 1990; 346:376–379. [PubMed: 2165219]

Tourmen Y, Baris O, Dessen P, Jacques C, Malthièry Y, Reynier P. Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics. 2002; 80:71–77. [PubMed: 12079285]

van der Kuyl AC, Kuiken CL, Dekker JT, Perizonius WR, Goudsmit J. Nuclear counterparts of the cytoplasmic mitochondrial 12S rRNA gene: a problem of ancient DNA and molecular phylogenies. J Mol Evol. 1995; 40:652–657. [PubMed: 7543951]

Varga T, Aplan PD. Chromosomal aberrations induced by double strand DNA breaks. DNA Repair. 2005; 4:1038–1046. [PubMed: 15935739]

Venkatesh B, Dandona N, Brenner S. *Fugu* genome does not contain mitochondrial pseudogenes. Genomics. 2006; 87:307–310. [PubMed: 16386876]

Wallace DC, Stugard C, Murdock D, Schurr T, Brown MD. Ancient mtDNA sequences in the human nuclear genome: a potential source of errors in identifying pathogenic mutations. Proc Natl Acad Sci USA. 1997; 94:14900–14905. [PubMed: 9405711]

Willett-Brozick JE, Savul SA, Richey LE, Baysal BE. Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation. Hum Genet. 2001; 109:216–223. [PubMed: 11511928]

Woischnik M, Moraes CT. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res. 2002; 12:885–893. [PubMed: 12045142]

Yan CT, Boboila C, Souza EK, Franco S, Hickernell TR, Murphy M, Gumaste S, Geyer M, Zarrin AA, Manis JP, Rajewsky K, Alt FW. IgH class switching and translocations use a robust non-classical end-joining pathway. Nature. 2007; 449:478–482. [PubMed: 17713479]

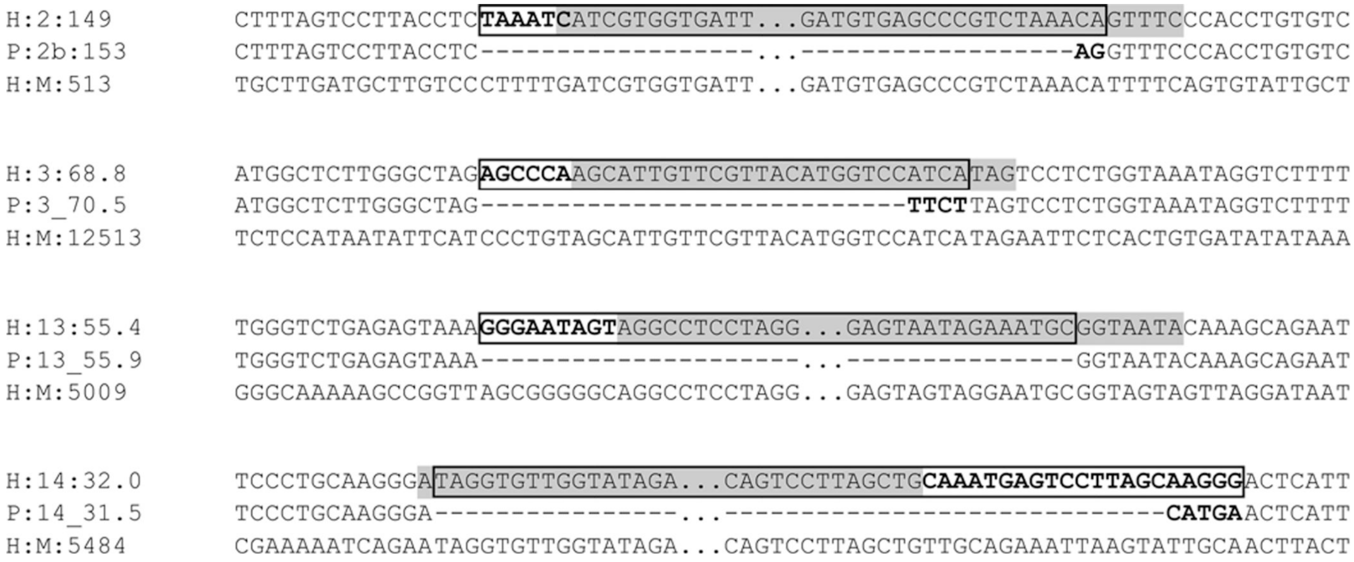
Yu N, Jensen-Seaman MI, Chemnick L, Kidd JR, Deinard AS, Ryder O, Kidd KK, Li WH. Low nucleotide diversity in chimpanzees and bonobos. Genetics. 2003; 164:1511–1518. [PubMed: 12930756]

Zhang D-X, Hewitt GM. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol Evol. 1996; 11:247–252. [PubMed: 21237827]

Zingler N, Willhoeft U, Brose H-P, Schoder V, Jahns T, Hanschmann K-MO, Morrisch TA, Löwer J, Schumann GG. Analysis of 5’ junctions of human LINE-1 and *Alu* retrotransposons suggest an alternative model for 5’-end attachment requiring microhomology-mediated end-joining. Genome Res. 2005; 15:780–789. [PubMed: 15930490]

Zischler H, Geisert H, von Haeseler A, Pääbo S. A nuclear ‘fossil’ of the mitochondrial D-loop and the origin of modern humans. Nature. 1995a; 378:489–492. [PubMed: 7477404]

Zischler H, Höss M, Handt O, von Haeseler A, van der Kuyl AC, Goudsmit J. Detecting dinosaur DNA. Science. 1995b; 266:1229–1232.



```
H:2:149       CTTTAGTCCTTACCTCTAAATCATCGTGGTGATT...GATGTGAGCCCGTCTAAACAGTTTCCCACCTGTGTC
P:2b:153      CTTTAGTCCTTACCTC------------------...------------------AGGTTTCCCACCTGTGTC
H:M:513       TGCTTGATGCTTGTCCCTTTTGATCGTGGTGATT...GATGTGAGCCCGTCTAAACATTTTCAGTGTATTGCT

H:3:68.8      ATGGCTCTTGGGCTAGAGCCCAAGCATTGTTCGTTACATGGTCCATCATAGTCCTCTGGTAAATAGGTCTTTT
P:3_70.5      ATGGCTCTTGGGCTAG----------------------------TTCTTAGTCCTCTGGTAAATAGGTCTTTT
H:M:12513     TCTCCATAATATTCATCCCTGTAGCATTGTTCGTTACATGGTCCATCATAGAATTCTCACTGTGATATATAAA

H:13:55.4     TGGGTCTGAGAGTAAAGGGAATAGTAGGCCTCCTAGG...GAGTAATAGAAATGCGGTAATACAAAGCAGAAT
P:13_55.9     TGGGTCTGAGAGTAAA---------------------...---------------GGTAATACAAAGCAGAAT
H:M:5009      GGGCAAAAAGCCGGTTAGCGGGGGCAGGCCTCCTAGG...GAGTAGTAGGAATGCGGTAGTAGTTAGGATAAT

H:14:32.0     TCCCTGCAAGGGATAGGTGTTGGTATAGA...CAGTCCTTAGCTGCAAATGAGTCCTTAGCAAGGGACTCATT
P:14_31.5     TCCCTGCAAGGGA----------------...-----------------------------CATGAACTCATT
H:M:5484      CGAAAAATCAGAATAGGTGTTGGTATAGA...CAGTCCTTAGCTGTTGCAGAAATTAAGTATTGCAACTTACT
```

Figure 1.

Examples of the use of the chimpanzee genomic sequence to accurately define the boundaries of the inserted numt. Shaded nucleotides represent BLAST-defined homology between human mtDNA and the human numt region. Boxed nucleotides are the actual nucleotides inserted, as inferred from comparison to chimpanzee sequence, the proxy for the pre-integration target site. Bold nucleotides indicate insertions in human, or deletions in chimpanzee. Sequences are labeled by species (H=*Homo*, P=*Pan*), followed by chromosome number (or "M" for the mitochondrial genome), followed by beginning position rounded to the nearest Mb for nuclear DNA or nearest bp for mitochondrial DNA.

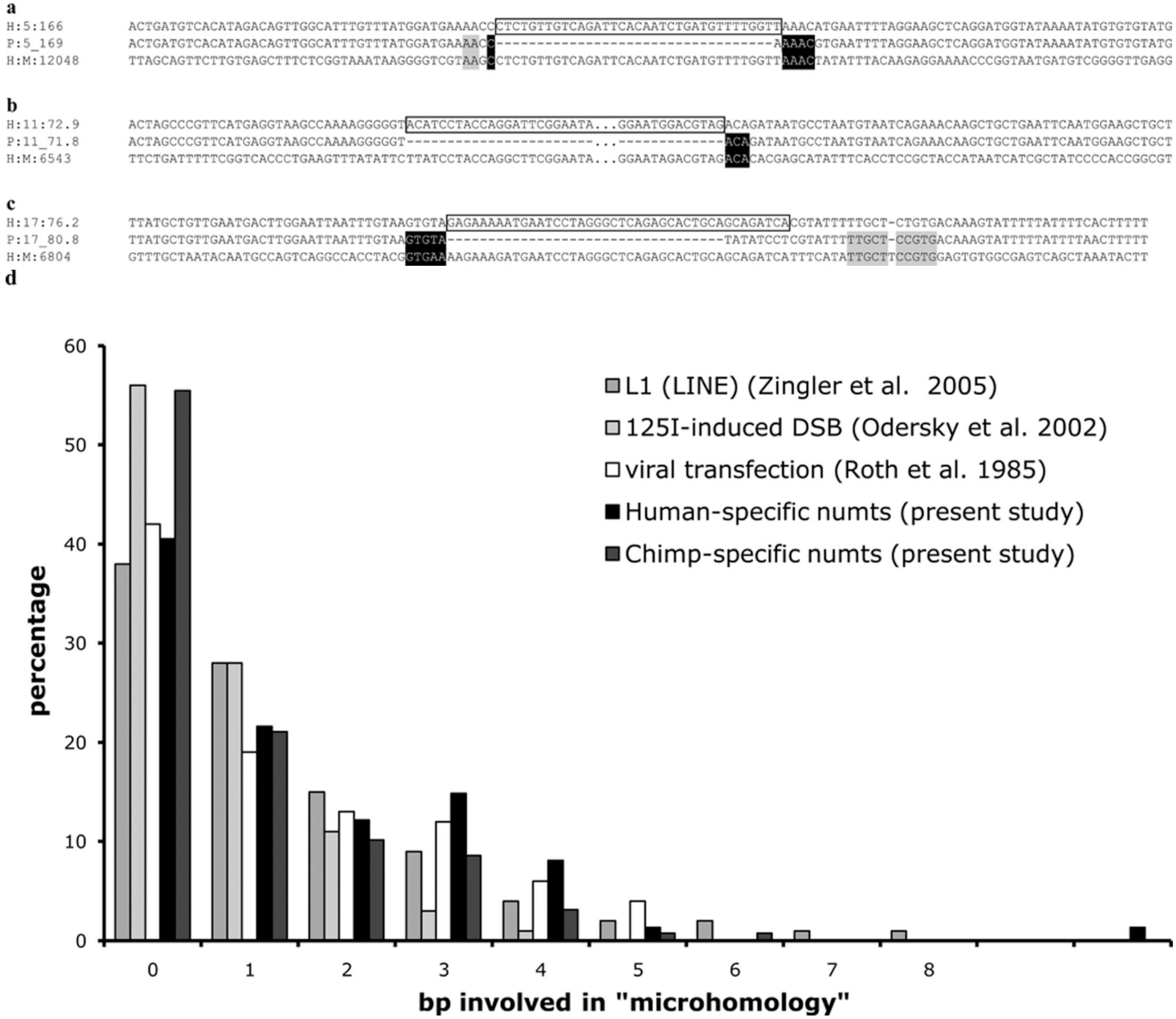
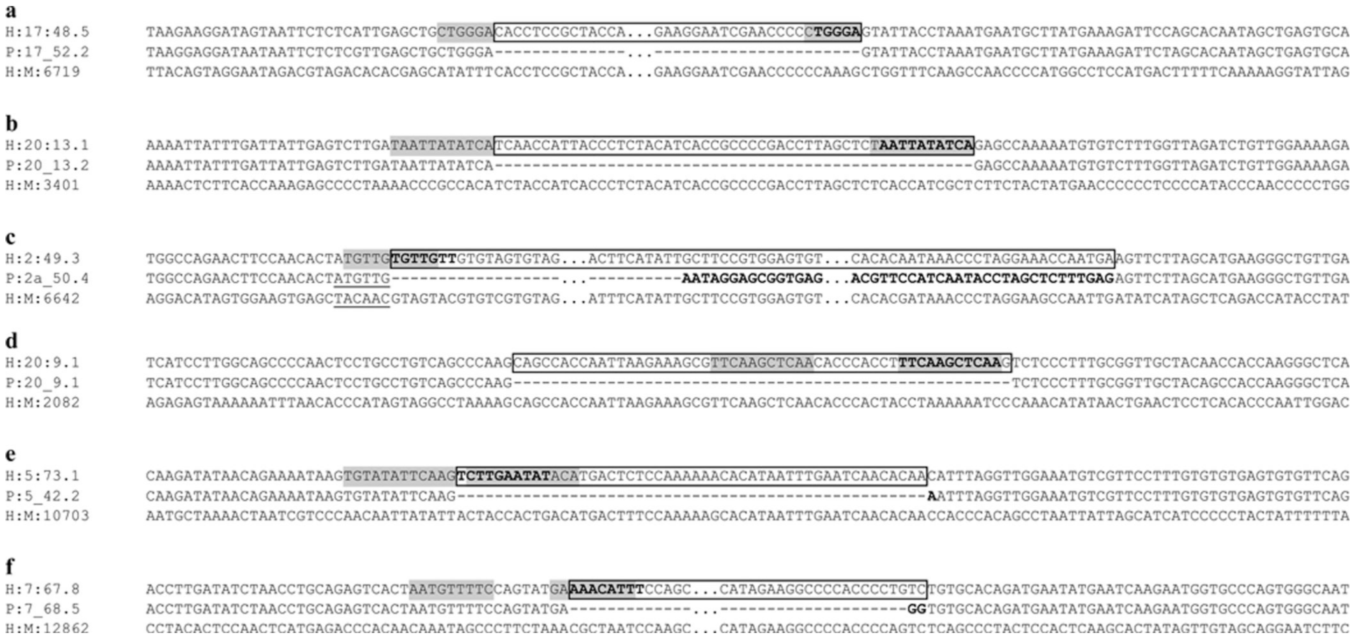


Figure 2.

a–c) Examples of microhomology found at numt-nuclear junctions. Black background with white text shows the conservatively defined microhomology used the quantitative analysis, while the shaded nucleotides show possible additional stretches of microhomology. The six junctions shown here exhibit 0, 0, 1, 3, 4, and 5bp of microhomology. Sequences are labeled as in Figure 1. **d)** Distribution of lengths of microhomology observed at the 37 human-specific numts compared to that described by other for other types of DNA double-strand break repair.

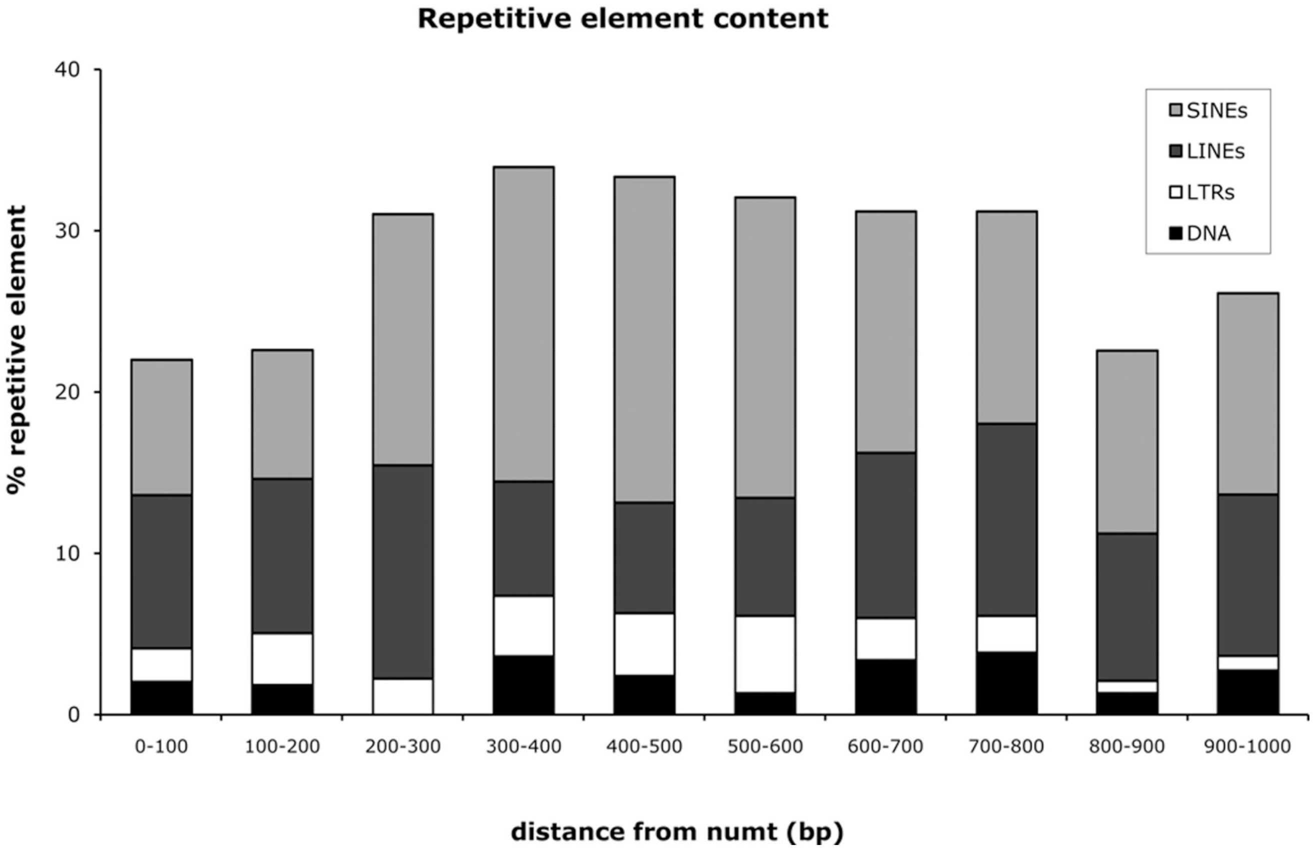


```
a
H:17:48.5    TAAGAAGGATAGTAATTCTCTCATTGAGCTGCTGGGACACCTCCGCTACCA...GAAGGAATCGAACCCCCTGGGAGTATTACCTAAATGAATGCTTATGAAAGATTCCAGCACAATAGCTGAGTGCA
P:17_52.2    TAAGGAGGATAATAATTCTCTCGTTGAGCTGCTGGGA--------------...----------------------GTATTACCTAAATGAATGCTTATGAAAGATTCTAGCACAATAGCTGAGTGCA
H:M:6719     TTACAGTAGGAATAGACGTAGACACACGAGCATATTTCACCTCCGCTACCA...GAAGGAATCGAACCCCCCAAAGCTGGTTTCAAGCCAACCCCATGGCCTCCATGACTTTTTCAAAAAGGTATTAG

b
H:20:13.1    AAAATTATTTGATTATTGAGTCTTGATAATTATATCATCAACCATTACCCTCTACATCACCGCCCCGACCTTAGCTCTAATTATATCAGAGCCAAAAATGTGTCTTTGGTTAGATCTGTTGGAAAAGA
P:20_13.2    AAAATTATTTGATTATTGAGTCTTGATAATTATATCA---------------------------------------------------GAGCCAAAAATGTGTCTTTGGTTAGATCTGTTGGAAAAGA
H:M:3401     AAAACTCTTCACCAAAGAGCCCCTAAAACCCGCCACATCTACCATCACCCTCTACATCACCGCCCCGACCTTAGCTCTCACCATCGCTCTTCTACTATGAACCCCCCTCCCCATACCCAACCCCCTGG

c
H:2:49.3     TGGCCAGAACTTCCAACACTATGTTGTGTTGTTGTGTAGTGTAG...ACTTCATATTGCTTCCGTGGAGTGT...CACACAATAAACCCTAGGAAACCAATGAAGTTCTTAGCATGAAGGGCTGTTGA
P:2a_50.4    TGGCCAGAACTTCCAACACTATGTTG------------------...----------AATAGGAGCGGTGAG...ACGTTCCATCAATACCTAGCTCTTTGAGAGTTCTTAGCATGAAGGGCTGTTGA
H:M:6642     AGGACATAGTGGAAGTGAGCTACAACGTAGTACGTGTCGTGTAG...ATTTCATATTGCTTCCGTGGAGTGT...CACACGATAAACCCTAGGAAGCCAATTGATATCATAGCTCAGACCATACCTAT

d
H:20:9.1     TCATCCTTGGCAGCCCCAACTCCTGCCTGTCAGCCCAAGCAGCCACCAATTAAGAAAGCGTTCAAGCTCAACACCCACCTTTCAAGCTCAAGTCTCCCTTTGCGGTTGCTACAACCACCAAGGGCTCA
P:20_9.1     TCATCCTTGGCAGCCCCAACTCCTGCCTGTCAGCCCAAG-----------------------------------------------------TCTCCCTTTGCGGTTGCTACAGCCACCAAGGGCTCA
H:M:2082     AGAGAGTAAAAAATTTAACACCCATAGTAGGCCTAAAAGCAGCCACCAATTAAGAAAGCGTTCAAGCTCAACACCCACTACCTAAAAAATCCCAAACATATAACTGAACTCCTCACACCCAATTGGAC

e
H:5:73.1     CAAGATATAACAGAAAATAAGTGTATATTCAAGTCTTGAATATACATGACTCTCCAAAAAACACATAATTTGAATCAACACAACATTTAGGTTGGAAATGTCGTTCCTTTGTGTGTGAGTGTGTTCAG
P:5_42.2     CAAGATATAACAGAAAATAAGTGTATATTCAAG--------------------------------------------------AATTTAGGTTGGAAATGTCGTTCCTTTGTGTGTGAGTGTGTTCAG
H:M:10703    AATGCTAAAACTAATCGTCCCAACAATTATATTACTACCACTGACATGACTTTCCAAAAAGCACATAATTTGAATCAACACAACCACCCACAGCCTAATTATTAGCATCATCCCCCTACTATTTTTTA

f
H:7:67.8     ACCTTGATATCTAACCTGCAGAGTCACTAATGTTTTCCAGTATGAAAACATTTCCAGC...CATAGAAGGCCCCACCCCTGTCTGTGCACAGATGAATATGAATCAAGAATGGTGCCCAGTGGGCAAT
P:7_68.5     ACCTTGATATCTAACCTGCAGAGTCACTAATGTTTTCCAGTATGA-------------...--------------------GGTGTGCACAGATGAATATGAATCAAGAATGGTGCCCAGTGGGCAAT
H:M:12862    CCTACACTCCAACTCATGAGACCCACAACAAATAGCCCTTCTAAACGCTAATCCAAGC...CATAGAAGGCCCCACCCCAGTCTCAGCCCTACTCCACTCAAGCACTATAGTTGTAGCAGGAATCTTC
```

Figure 3.

Examples of small duplications accompanying numt integration, derived from flanking direct repeats (**a,b**), tandem direct repeats (**c,d**), and inverted repeats (**e,f**). Shaded nucleotides indicate the duplication while boxed nucleotides indicate the numt and bold nucleotides indicate indels. The underlined nucleotides in **c** show a perfect complement between the preintegration sequence and the mitochondrial DNA.

a


Repetitive element content
% repetitive element
0
10
20
30
40
SINEs
LINEs
LTRs
DNA
0-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000
distance from numt (bp)

b

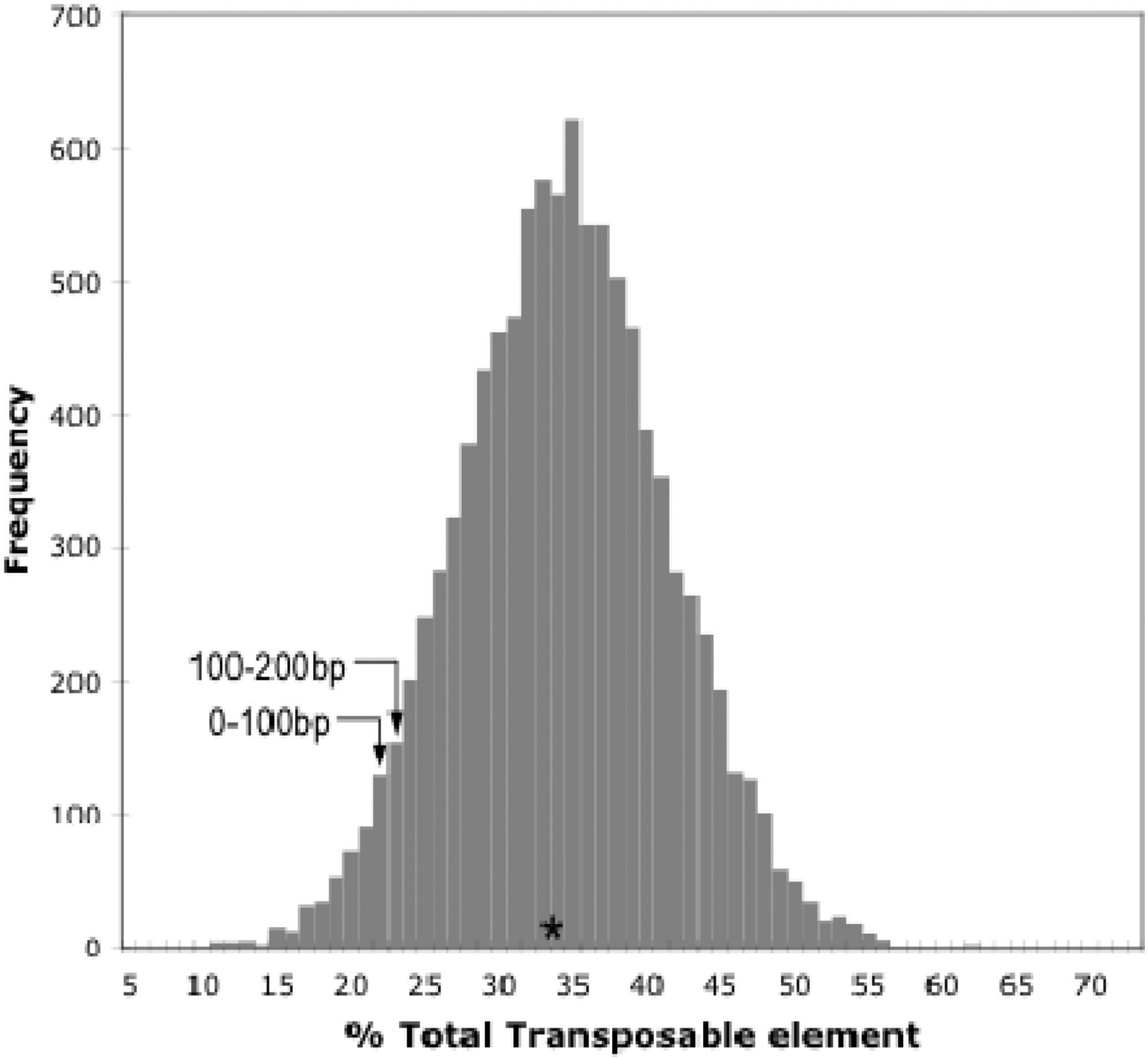


Figure 4.

a) Transposable element content in 100bp windows flanking human-specific numts. Each column shows the major classes of transposable elements estimated from 7400bp (37 numts × 2 flanking regions × 100bp). Dashed line indicates the average (33.8%) of the total transposable element content found in 10,000 randomly generated data sets (each data set consisted of 37 regions × 2 flanking regions × 100bp). **b)** Distribution of the total transposable element content of the 10,000 randomly generated data sets, along with the values from the first 100bp and the second 100bp from the flanking regions of the 37 human-specific numts. Asterisk indicates the average of the distribution.

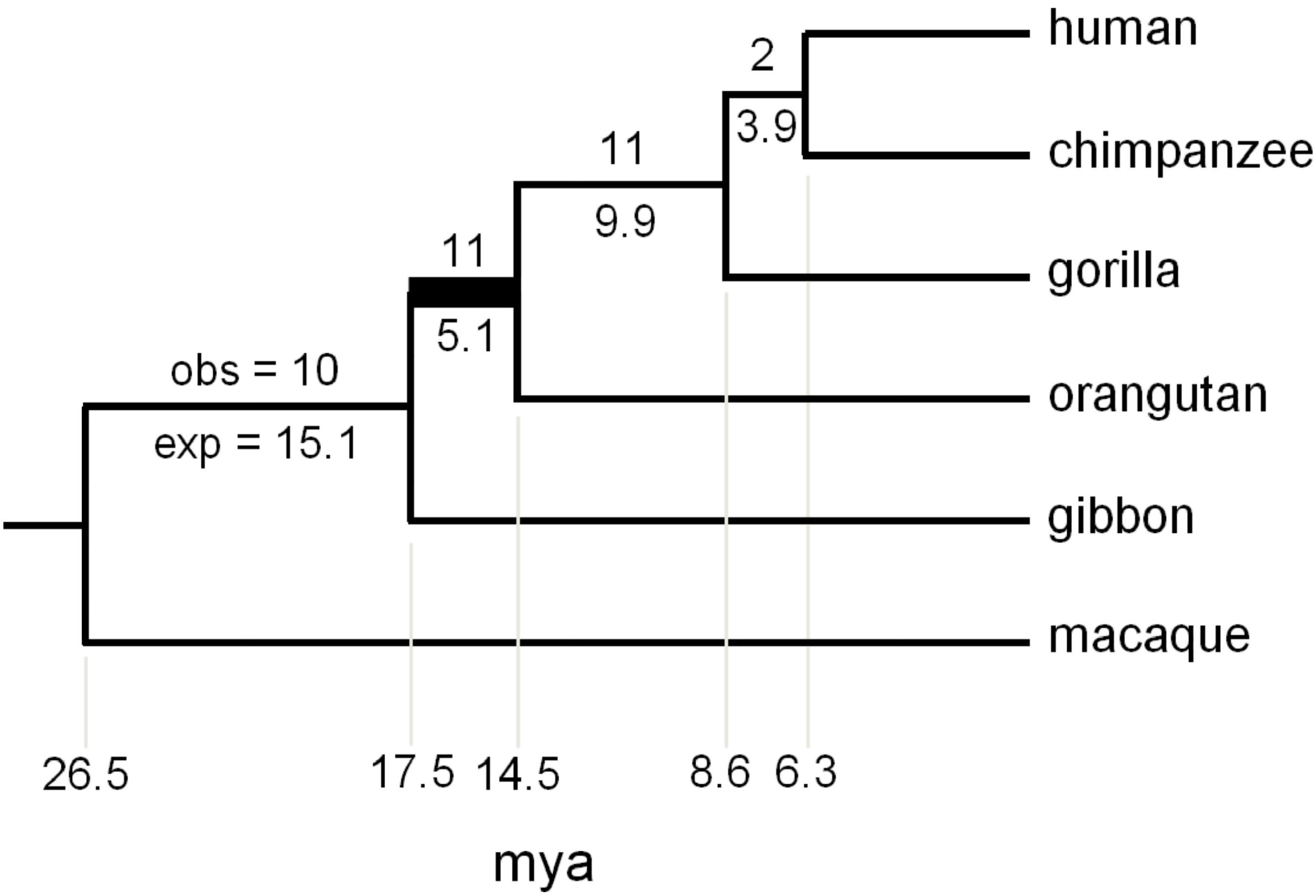


Figure 5.

Observed (above branch) and expected (below branch) distribution of hominoid numt insertions, determined with cross-species PCR, and shown on the universally accepted phylogeny. Bayesian posterior probability estimates of divergence times are given below, taken from Raaum et al. (2005), and used to calculate the expected number of numts on each branch. A significant excess of numts have inserted into the common ancestor of humans and the great apes, following their divergence with gibbons ($p < 0.05$; indicated by a thick branch).