Archaic mitochondrial DNA inserts in modern day nuclear genomes

Robert Bücking (robert_buecking@eva.mpg.de)
Max-Planck-Institut fur evolutionare Anthropologie https://orcid.org/0000-0002-0180-5551
Murray P Cox
Massey University Institute of Fundamental Sciences
Georgi Hudjashov
Massey University Institute of Fundamental Sciences
Lauri Saag
Tartu Ulikool
Herawati Sudoyo
Eijkman Institute for Molecular Biology
Mark Stoneking
Max-Planck-Institut fur evolutionare Anthropologie

Research article

Keywords: Archaic introgression, NUMTs, Denisovans

Posted Date: December 19th, 2019

DOI: https://doi.org/10.21203/rs.2.14881/v3

Version of Record: A version of this preprint was published on December 26th, 2019. See the published version at https://doi.org/10.1186/s12864-019-6392-8.

RESEARCH

Archaic mitochondrial DNA inserts in modern day nuclear genomes

Robert Bücking[1*†], Murray P Cox[2], Georgi Hudjashov[2], Lauri Saag[3], Herawati Sudoyo[4,5,6] and Mark Stoneking[1†]

*Correspondence:
robert_buecking@eva.mpg.de
[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D04103 Leipzig, Germany
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: Traces of interbreeding of Neanderthals and Denisovans with modern humans in the form of archaic DNA have been detected in the genomes of present-day human populations outside sub-Saharan Africa. Up to now, only nuclear archaic DNA has been detected in modern humans; we therefore attempted to identify archaic mitochondrial DNA (mtDNA) residing in modern human nuclear genomes as nuclear inserts of mitochondrial DNA (NUMTs).

Results: We analysed 221 high-coverage genomes from Oceania and Indonesia using an approach which identifies reads that map both to the nuclear and mitochondrial DNA. We then classified reads according to the source of the mtDNA, and found one NUMT of Denisovan mtDNA origin, present in 15 analysed genomes; analysis of the flanking region suggests that this insertion is more likely to have happened in a Denisovan individual and introgressed into modern humans with the Denisovan nuclear DNA, rather than in a descendant of a Denisovan female and a modern human male.

Conclusions: Here we present our pipeline for detecting introgressed NUMTs in next generation sequencing data that can be used on genomes sequenced in the future. Further discovery of such archaic NUMTs in modern humans can be used to detect interbreeding between archaic and modern humans and can reveal new insights into the nature of such interbreeding events.

Keywords: Archaic introgression; NUMTs; Denisovans

Background

The presence of mitochondrial DNA (mtDNA) sequences in nuclear genomes has been widely reported in various eukaryotic organisms [1, 2]. The transfer of such genetic material in humans, reported to still be an ongoing evolutionary process [3, 4, 5, 6], has resulted in various fixed and polymorphic Nuclear Mitochondrial DNA segments (NUMTs) in present day genomes. In the human reference genome, a total of 755 NUMTs have been identified [7]. In addition to these NUMTs, many more polymorphic NUMTs have been detected in various human populations around the world [8] and the analysis of additional populations is expected to reveal many more polymorphic NUMTs. NUMTs vary in size and can consist of almost the entire mitochondrial genome. There is no evidence for a preference for the insertion of particular regions of the mtDNA genome, which is entirely represented by the various NUMTs in the nuclear genome [9, 8].

In general, NUMTs behave as noncoding sequences in the nuclear genome and evolve without any functional constraints [10]. Old insertions tend to get modified

by deletions, duplications, inversions and other mutations over a long period of time until they are no longer recognizable as mtDNA. More recent insertions tend to still preserve the sequence at the time of their insertion due to the lower mutation rate in nuclear than in mitochondrial DNA (as reviewed in [11]). NUMT sequences have been used to reveal processes of molecular evolution in the absence of selection [12] and to date events of species divergence [13].
Moreover, their frequencies and presence-absence patterns have also been used to study the genetic relationships of different human populations [14, 15, 16] .
NUMTs have also been used to detect admixture and hybridisation events between present or extinct species. Interbreeding events in insects [17] and ancient hybridisation between two monkey genera [18] have been detected by combining presence-absence analyses with information about the NUMT sequence. While NUMTs are usually transmitted vertically across generations and thus represent an ancestral state of the mtDNA of that species, they can also be transferred horizontally through interbreeding between different species or diverged populations. A NUMT arising via such interbreeding would not follow the mtDNA phylogeny of the individual it was found in. If the NUMT sequence is more similar to the mtDNA of another species or distant population, it would indicate introgression through interbreeding between these two lineages.
In this study, we screened modern human genomes for fragments of mtDNA in the nuclear genome with a different phylogeny than that of modern humans. Such NUMTs can be used to detect and analyse interbreeding events in the evolutionary history of humans. It is already well-established that the ancestors of various modern human populations interbred with Neanderthals and Denisovans [19, 20, 21, 22]. All genomes outside of sub-Saharan Africa contain $\sim 2\%$ Neanderthal DNA [22], while Denisovan ancestry is less evenly distributed around the world; Eastern Eurasian and Native American populations only contain small amounts, whereas Oceanian populations derive up to $\sim 4\%$ of their genome from Denisovan DNA [21, 23]. More recent studies have shown that interbreeding events between archaic and modern humans occurred several times during the evolutionary history of modern humans [24, 25, 26, 27, 28].

While the introgression of archaic nuclear DNA into modern humans has been widely detected, archaic mtDNA genomes have not been detected in modern humans. However, archaic NUMTs in modern humans have not been systematically investigated. There are two pathways by which archaic NUMTs could be introduced into modern human genomes (Figure 1): (a) the NUMT arose in the archaic humans and was then transferred to modern humans along with other archaic nuclear DNA via interbreeding; (b) the NUMT arose *de novo* in the germ line of an archaic-modern human hybrid with an archaic mtDNA genome, and was subsequently passed on to the modern human population. This latter pathway is perhaps of more interest as it provides information about the sex of the interbreeding individuals that is otherwise not available from the archaic nuclear DNA in modern humans. In this case, it would indicate that the original interbreeding involved an archaic female and a modern human male. These two pathways can be distinguished by examining the genomic region surrounding the archaic NUMT: if the

archaic NUMT had occurred in an archaic human, then the surrounding genomic region should consist of archaic DNA; if the archaic NUMT arose *de novo* in an archaic-modern human hybrid, then the surrounding genomic region should consist of modern human DNA (unless the NUMT happened to insert into a region of archaic DNA in the hybrid).

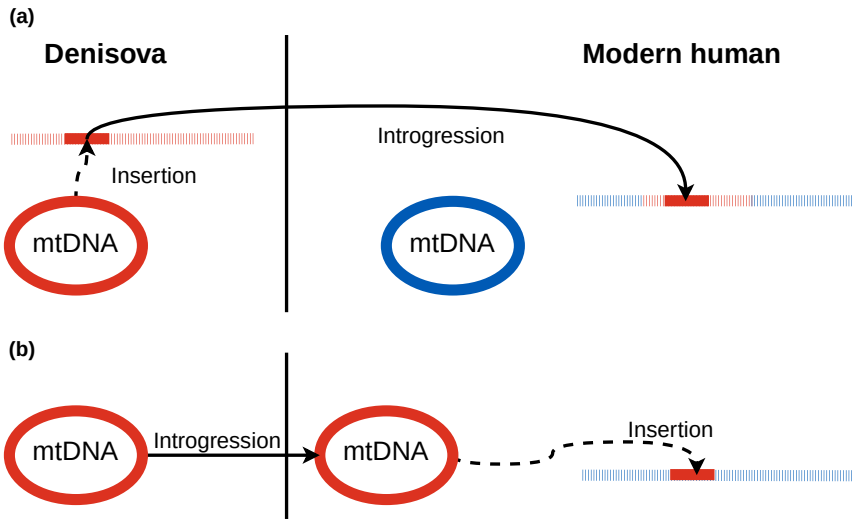

Figure 1 Hypotheses for NUMT introgression. Two possible scenarios for the introgression of Denisovan mitochondrial DNA (mtDNA) in modern human nuclear DNA. (a) The NUMT occurred in a Denisovan and introgressed together with its nuclear DNA. This path would result in a flanking region with Denisovan ancestry. (b) The mtDNA introgressed into a modern human, where it subsequently inserted into the nuclear DNA. mtDNA: solid colours; nuclear DNA: striped; Denisovan DNA: red; modern human DNA: blue

To detect such NUMTs, we scanned 221 genomes from Indonesia and Oceania for NUMTs that arose from archaic mtDNA, as individuals from these regions harbour both Neanderthal and Denisovan ancestry. Such NUMTs can be identified as a deviation from the actual mtDNA phylogeny in which modern humans form a monophyletic group compared to archaic humans [20]. We detected polymorphic mtDNA insertions in next generation sequencing data following the approach of Dayama et al. [8]. Afterwards we reconstructed sequences for these NUMTs, which were then analysed regarding their phylogeny, revealing the archaic ancestry of one NUMT. To discover the pathway of this NUMT into the modern human genome, we further analysed the flanking regions for archaic ancestry. Additionally, we detected population-specific patterns of polymorphic NUMTs which could be useful as markers in phylogenetic studies.

Results

Factors influencing the number of detected NUMTs

To detect polymorphic NUMTs, genomes were scanned for read pairs mapping both to chromosomal DNA and mtDNA as described in [8]. A total of 221 genomes from Indonesia and Oceania, from three studies, the Indonesian Genome Diversity

Table 1 Summary of discovered NUMTs in the different studies analysed; results for the 1000 GP dataset are taken from Dayama et al. [8].

Study	1000 GP [8]	SGDP (Oceanian samples)	Vernot et al.	IGDP
Populations	20	8	13	27
Analysed samples	946 (1000 GP) 53 (HGDP)	25	35	161
Coverage	$\sim 4-6$ (1000 GP) $\sim 5-20$ (HGDP)	$\sim 30-40$	$\sim 30-40$	$\sim 30-40$
NUMTs per sample	~ 1.5	16.0 ± 6	13.4 ± 3.3	16.8 ± 10.2
Total NUMTs	~ 1499	399	470	4035
Distinct NUMTs	141	171	121	1062
Study-specific NUMTs	115	128	81	975

Distinct NUMTs: amount of different NUMTs found in each study.
Study-specific NUMTs: amount of distinct NUMTs not found in any other study.

Project (IGDP) [28], Simons Genome Diversity Project (SGDP) [29] and Vernot et al. [26] were analysed. A summary of the results for each study is shown in Table 1. In total, 1222 distinct NUMTs not annotated in the human reference genome were found (Supplementary File 2), with 1197 not detected in the 1000 Genomes Project (1000 GP) [30] and Human Genome Diversity Project (HGDP) samples analysed previously by Dayama et al. [8]. The 25 NUMTs already detected by Dayama et al. [8] were all present in samples from Europe, East Asia, America and Sub-Saharan Africa. The majority of NUMTs found in one study were not found in either of the other studies. These differences in NUMT patterns, even between populations within Oceania, supports previous observations of ongoing NUMT insertions in humans [4]. On average, 16.3 NUMTs were detected in each sample from the three high-coverage studies, compared to the average of 1.5 per sample found in the low to medium-coverage 1000 GP dataset. In addition, the ratio of distinct NUMTs to the total number found is very low in the 1000 GP dataset compared to others. Higher coverage thus seems to strongly increase the detection rate of this approach.

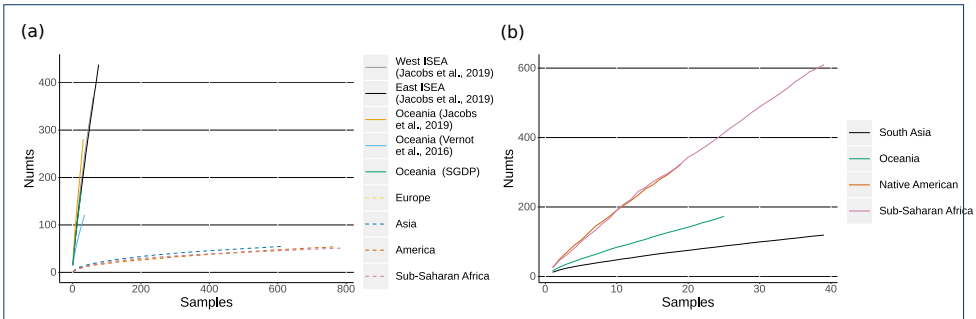

Figure 2 Comparison of rarefaction curves. Rarefaction curves plotted as the mean out of 100 repetitions. (a) Results from Island South East Asia (ISEA) and Oceania (continuous lines) are compared with worldwide 1000 GP data (dashed lines). (b) Comparison of different geographic regions within the Simons Genome Diversity Project (SGDP) dataset.

We further analysed differences in NUMT diversity between different studies and larger geographical regions, using rarefaction analysis. Figure 2a shows a comparison of rarefaction curves for the different studies. For the 1000 GP dataset the

rarefaction curves saturate at a low level, compared to the curves for the high-coverage studies, which do not reach saturation. Coverage therefore has a significant impact on the discovery of NUMTs. We also investigated NUMT diversity in different geographic regions, controlling for coverage by focusing on samples from the SGDP study. NUMT diversity here is higher in sub-Saharan Africa than in Oceania and South Asia (Figure 2b). Additionally, we downsampled high coverage genomes to lower coverages and screened them for NUMTs. This analysis showed a moderate correlation between coverage and the amount of NUMTs detected ($r^2 = 0.35, p \leq 2.2e^{-19}$, Supplementary Figure S1), suggesting that many NUMTs may be missed in genomes with a coverage below 10.

Ancestral and archaic NUMTs in Indonesian and Oceanian genomes

A sequence for each NUMT was reconstructed from the supporting reads. Those with a total sequence length below 20 base pairs (bp) were discarded. This cutoff was chosen to keep sequences that might contain enough phylogenetic information for further classification, but still filter out sequences that can not be classified. In most cases, one single sequence was generated. For some, multiple fragments were obtained and concatenated. These NUMTs were too long to be covered completely by short split reads or did not have sufficient read coverage in all positions. From one side of an insertion, a split read pair could span a part of the NUMT with a length of the sum of the mitochondrial read length, plus the insert size and the length of the clipped half of the chromosomal read. With an average read length of 150 bp and a median insert size up to 400 bp in the analysed samples, no single fragment longer than 1250 bp would be expected. The length of the fragments generated ranged between 26 and 774 bp (median: 70 bp), with the majority being shorter than 100 bp (Supplementary Figure S2a). For the high-coverage genomes used in our study, NUMTs with an overall mean coverage below 5x might result from sequencing artefacts. Sequences generated from such a low coverage are also more likely to be influenced by sequencing errors and therefore were discarded. As the coverage distribution for the NUMTs (up to 59x, median 14x, Supplementary Figure S2b) was not different from the rest of the genome sequence, filtering for excessive coverage was not required. GC-content for the NUMTs varied between 0.27 and 0.64 (median: 0.46) (Supplementary Figure S2c), similar to the the GC-content for the mtDNA genome (0.45), so no filtering based on GC-content was applied.

After filtering, a total of 2041 assembled fragments were obtained from all samples combined. These fragments belong to 172 distinct NUMTs. For each of these fragments the phylogeny was analysed by building a tree with the corresponding mtDNA fragments from various humans and hominins, using chimpanzee as an outgroup.

The rooted trees were used to infer the origins of the NUMTs according to where their sequences fell within the hominin mtDNA phylogeny. To be able to classify a NUMT as either archaic, modern human or ancestral, at least some phylogenetic information is required, but the partial mtDNA sequences of some of the NUMTs

are too short or conserved for accurate placement on the tree. Therefore only those trees for which either Denisovans, Neanderthals, modern humans, or Neanderthals and modern humans formed a monophyletic clade were considered for further analysis. A tree with a monophyletic clade of Neanderthals and Denisovans was also allowed as they both might represent the ancestral state for a region where modern humans show derived alleles. Additionally, trees placing the NUMT outside of all humans were used to classify it as ancestral to all humans.

As this study aimed to detect archaic NUMT insertions, only those classified as not arising from modern humans were further analysed. To exclude that these are still part of modern human variation, pairwise nucleotide distances within and between modern humans, Neanderthals and Denisovans were calculated as in Figure 4. For 98 NUMTs no useful tree could be generated. Either the alignments did not contain more than four distinct sequences or the trees did not contain any reasonable clades, as the sequences were too short or from mtDNA regions that were too conserved to contain enough phylogenetic information.

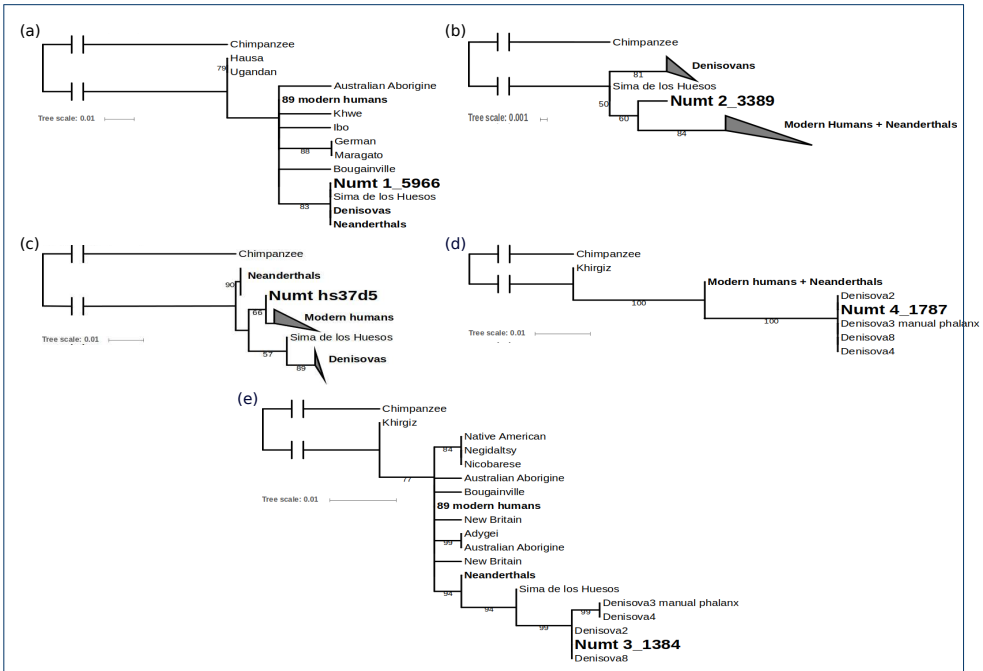

Figure 3 Phylogenetic trees for NUMTs. Maximum likelihood trees for putative ancestral NUMTs (a, b, c) and putative Denisovan NUMTs (d, e) in relationship with other hominin mtDNA sequences using distances based on nucleotide substitution rates. Ancestral NUMTs form a sister clade to at least all modern humans and are present in 1000 Genomes Project (1000 GP) samples from around the world, except for (c). Denisovan NUMTs are more similar to Denisovan mtDNA than to modern human mtDNA and are not present in 1000 GP samples. Bootstrap values over 50 are indicated at branch locations.

Most of the remaining NUMTs could be classified as modern human mtDNA according to their trees. For five NUMTs, the trees suggest an origin other than modern human mtDNA. Out of these five, two were also found by Dayama et al. [8] and are present in genomes from sub-Saharan Africa, Europe, Asia and America. The corresponding names from Dayama et al. [8] are given in brackets below.

Putative ancestral NUMTs

For NUMT 1_5966 (Poly_NumtS_67) the 58 bp sequence is identical with all corresponding Neanderthal and Denisovan mtDNA sequences, but differs at only one position from most modern humans (Figure 3a). Therefore it cannot be confidently classified as archaic mtDNA. Taking into account its presence in worldwide modern human genomes, this NUMT was likely inserted in an ancestor of all modern humans, possibly even before the split with archaic humans, and therefore might resemble the ancestral state of this mtDNA region (Supplementary Figure S3).

NUMT 2_3389 (Poly_NumtS_1239) forms a sister clade to all modern humans and Neanderthals (Figure 3b). The 246 bp sequence inferred here is identical to that obtained by Dayama et al. [8] through Sanger sequencing. It is equally distant to modern humans, Neanderthals and Denisovans, but does not completely fall outside of modern human variation (Figure 4a). Based on the comparison with an inferred ancestral human mtDNA, its age of insertion is estimated to be around 720,000 years ago, although this comparison does not allow a precise estimation [8]. This estimation falls into the proposed time range of the population split between archaic and modern humans of between 550,000 and 765,000 years ago [31] (Supplementary Figure S3). The presence of these two NUMTs in genomes from various populations around the world, including sub-Saharan Africa, suggests an insertion before the worldwide expansion of modern humans.

NUMT hs37d5_2745 is 446 bp and was detected in the decoy sequences. These sequences are found in several *de novo* assemblies of the human genome, but are not present in the hg19 reference [32]. A BLAST search (https://blast.ncbi.nlm.nih.gov/Blast.cgi) for a 240 bp region flanking this NUMT in the National Center for Biotechnology Information database (https://www.ncbi.nlm.nih.gov/) showed the best hit for a human Bacterial Artificial Chromosome (BAC) clone from chromosome 4 (AC124864.3). The exact location of this BAC clone on chromosome 4 is not given, therefore the exact insertion site of the NUMT can not be localised. Its sequence forms a sister clade to all modern humans (Figure 3c), but it does not fall outside of their variation (Figure 4b). The alignment contains only 15 positions with more than one allele in hominins. For one of these positions, the NUMT shares the ancestral state with chimpanzee and all archaic humans, whereas all modern humans share the derived allele (Supplementary Figure S4). This could indicate that the NUMT originated from an ancestral extinct or unsampled mtDNA lineage, but could also be due to a convergent mutation in the NUMT. Therefore it cannot be classified confidentially as ancestral.

Putative Denisovan NUMTs

The sequences for two NUMTs were identical to Denisovan mtDNA. NUMT 4_1787 was detected in five samples from west Indonesian populations speaking Austronesian languages (Supplementary Table S1) and is identical to the mtDNA sequence of all Denisovan individuals (Figure 3d). However, the sequence obtained is only 43 bp long and differs from most other humans at just one position; thus, it cannot be confidently identified as Denisovan mtDNA.

NUMT 3_1384 is present in 15 samples from eastern Indonesia and New Guinea (Supplementary Table S1). A sequence of 251 bp was generated, which is identical to

two Denisovan mtDNAs. It forms a clade with Denisovans and Sima de los Huesos, distinct from all other humans (Figure 3e) and falls outside of all modern human and Neanderthal variation (Figure 4c). The alignment contains 13 variable positions within hominins (Supplementary File 3). For five of these positions, Denisovans and the NUMT share an allele which differs from all modern humans. This suggests that it originated from Denisovan mtDNA rather than from mtDNA of a modern human or an ancestor of Denisovans and modern humans (Supplementary Figure S3). The phase of this NUMT was inferred in each sample using five phased genotypes (Supplementary Table S2).

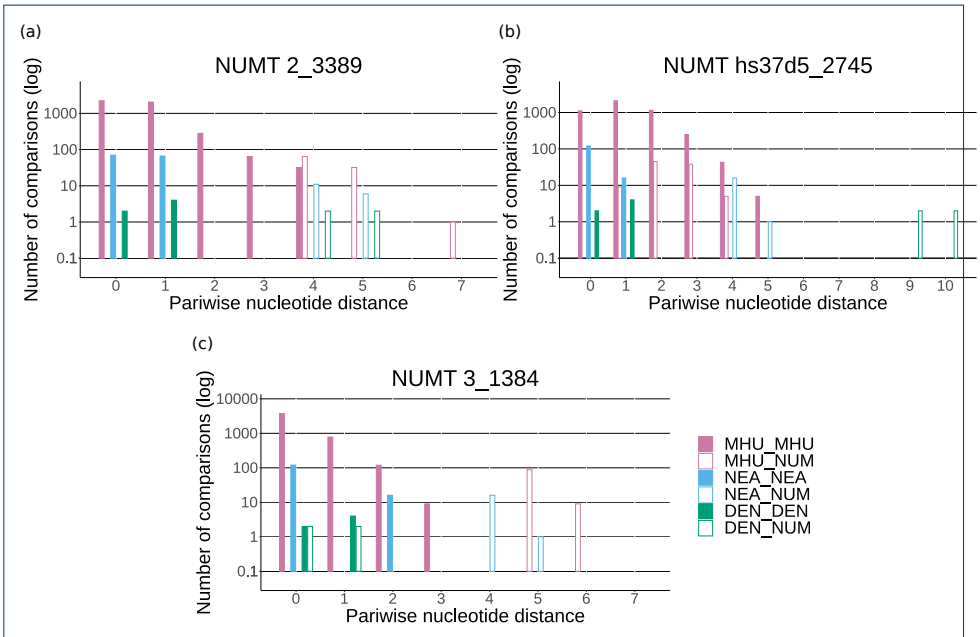

Figure 4 Pairwise nucleotide distances between NUMTs and mtDNA. Pairwise nucleotide distances vs. frequency (in logarithmic scale) within and between 97 modern humans (MHU, purple), 17 Neanderthals (NEA, blue), four Denisovans (DEN, green) and a specific NUMT (NUM, empty bars) for two ancestral NUMTs (a, b) and one Denisovan NUMT (c).

A Denisovan NUMT introgressed as nuclear DNA

To determine if the archaic NUMT introgressed within Denisovan nuclear DNA as explained in Figure 1, the flanking regions were analysed for Denisovan ancestry using phased genotype data. For NUMT 3_1384, alleles shared between modern humans and Denisovans were identified in the flanking region. These alleles could be shared due to an introgression event, incomplete lineage sorting or homoplasy. As Denisovan ancestry is low or absent in populations outside of Oceania [23, 27], any Denisovan alleles in these populations presumably reflect the latter two cases, and hence can be used to distinguish them from true introgression. Figure 5a shows the frequencies of alleles in the SGDP and 1000 GP dataset shared between a Denisovan and an Oceanian sample containing NUMT 3_1384. Shared alleles with low frequencies in these worldwide datasets are abundant before and around the NUMT, suggesting the presence of an introgressed haplotype. Shared alleles further

away from the insertion point mostly show higher frequencies in the worldwide datasets, suggesting that these alleles are either shared due to incomplete lineage sorting or homoplasy. Similar distributions of low and high frequency shared alleles were observed for all haplotypes containing NUMT 3_1384, indicating the end of that putatively introgressed haplotype around position chr3:13851000. In contrast, Figure 5b shows the frequencies of shared alleles in the same region for a sample from sub-Saharan Africa, containing very few low-frequency shared alleles.

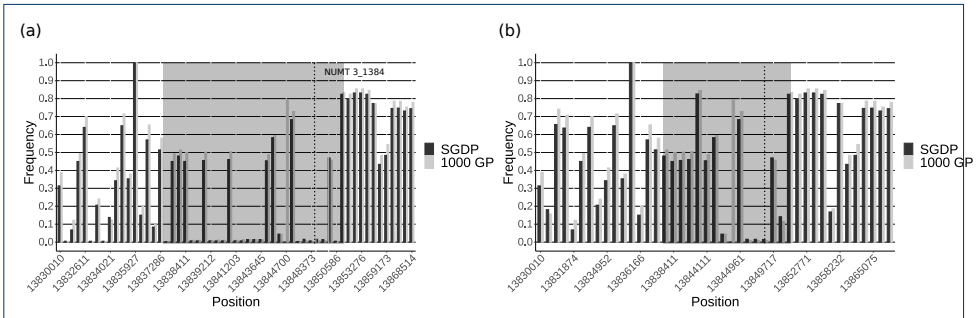

Figure 5 Worldwide frequencies of alleles shared with the Denisovan genome. Worldwide frequencies of alleles shared with the Denisovan genome in the region flanking NUMT 3_1384 in (a) Oceanian sample UV925, and (b) the same region in the sub-Saharan African sample NA19309. The alleles are within 20 kbp before or after the insertion site (vertical dashed line) and on the same phase. Frequencies were calculated in the 1000 Genomes Project (1000 GP) (black) and Simons Genome Diversity Project (SGDP) (grey) datasets. X-axis intervals are not linear, but indicate positions of shared alleles. High-frequency shared alleles reflect either homoplasy or incomplete lineage sorting; the greater abundance of low-frequency alleles shared with the Denisovan genome close to the insertion point (grey area) in the Oceanian sample vs. the sub-Saharan African sample suggests a Denisovan-introgressed haplotype in the Oceanian.

For the region around the insertion site of NUMT 3_1384, match ratios with the Denisovan genome were calculated as described above. Introgressed haplotypes are expected to show a higher match ratio than non-introgressed haplotypes. Due to the absence of Denisovan ancestry in European and sub-Saharan African populations, these populations can be used to estimate the expected distribution of match ratios for the haplotype of interest due to incomplete lineage sorting or chance; values higher than expected from this distribution suggest Denisovan ancestry. On average, 53 phased sites contained a non-reference allele, and the distribution of match ratios is shown in Figure 6 (min = 0, Q1 = 0.13, mean = 0.25, Q3 = 0.35, max = 0.83,). Match ratios for haplotypes containing NUMT 3_1384 range from 0.59 to 0.83 with the exception of sample S_Papuan-5 (0.48). Except for this sample, they represent the highest 0.4 % match ratios together with eight other haplotypes from two Europeans, two Southeast Asians, three Oceanians and one Indonesian. The regions flanking NUMT 3_1384 show higher similarity with the Denisovan genome than the same regions in all other analysed samples, suggesting a Denisovan origin of these flanking regions. The exception for sample S_Papuan-5 may reflect depletion of Denisovan ancestry through recent recombination around the insertion site.

For samples containing a heterozygous Denisovan NUMT insertion, the average match ratio of haplotypes containing the insertion (mean = 0.74) was significantly higher (one tailed paired t-test, $p = 2 * 10^{-9}$) than the average match ratio of

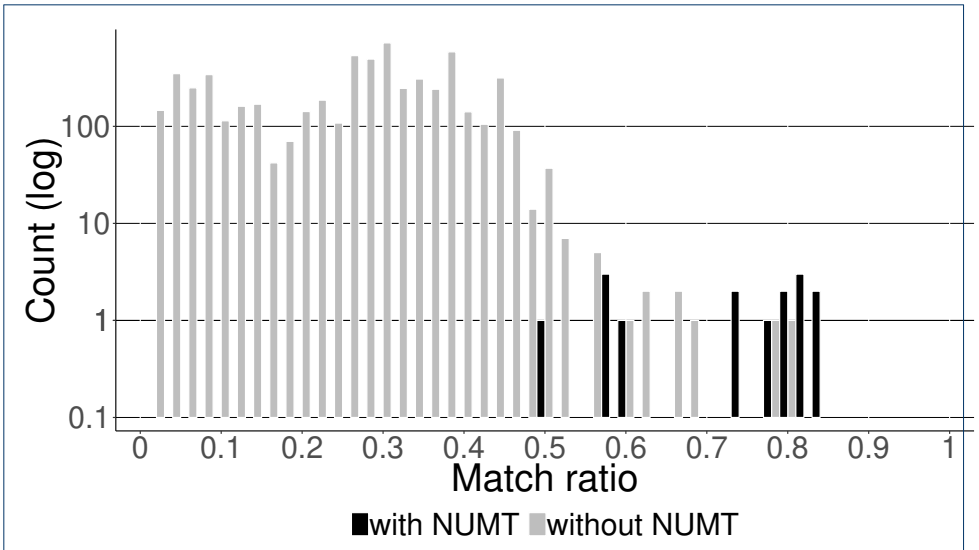

Figure 6 Match ratios of modern human genomes with the Denisovan Genome. Distributions of match ratios between modern humans and Denisovans for a 20 kbp region around NUMT 3_1384 in all samples from the 1000 GP, SGDP, IGDP and Verot et al. [26]. Match ratios were calculated for each phase individually counting shared non-reference alleles. Black bars represent haplotypes containing the NUMT insertion, grey bars represent haplotypes without the NUMT insertion.

haplotypes lacking the insertion (mean = 0.25). This further suggests that the NUMT insertion is part of a haplotype introgressed from Denisovans, whereas haplotypes for the same region but lacking the NUMT are not derived from Denisovans.

The reference genome influences the detection of archaic NUMTs

NUMTs originating from Denisovan mtDNA might not be detected in genomes mapped against modern human mtDNA due to their diverged sequences. To test the effect of this potential reference bias, we remapped ten samples from Vernot et al. [26] (Supplementary Table S3) against a reference genome containing Denisovan mtDNA instead of modern human mtDNA. No additional NUMTs were detected. On average only 1.9 NUMTs were detected per sample, many fewer than the 16 NUMTs per sample in the original mapping files (Table 1). Calculating the coverage along the Denisovan reference mtDNA reveals gaps where only few or no modern human reads could be aligned (Supplementary Figure S5). Many modern human reads containing mtDNA sequences could not be aligned to the Denisovan reference, which reduces the number of detectable NUMTs. This also suggests that some Denisovan NUMTs might not be detected due to the use of a modern human reference for all analysed samples; however the regions that differ greatly between Denisovan and modern human mtDNA are short enough that reliable detection of NUMTs involving only those regions would be difficult anyway.

Short read length complicates the identification of archaic NUMTs

Read pairs originating from NUMTs longer than 600 bp might not overlap the insertion site. Such read pairs will not be detected by our approach, but might con-

tain important phylogenetic information. Therefore we analysed reads overlapping diagnostic mtDNA positions for different hominin lineages from Meyer at al. [31]. For the Neanderthal and Denisovan lineages, we could not detect more than five reads supporting the same diagnostic allele. For the Sima de los Huesos lineage, five reads supported a diagnostic allele at position 13391 in sample UV1134. An average of seven diagnostic positions for the Denisovan-Sima de los Huesos lineage were supported by 5-47 reads in each sample. This range of coverage is closer to the range of the nuclear chromosomal coverage of up to 50 x than to the mitochondrial coverage of around 1000 x.

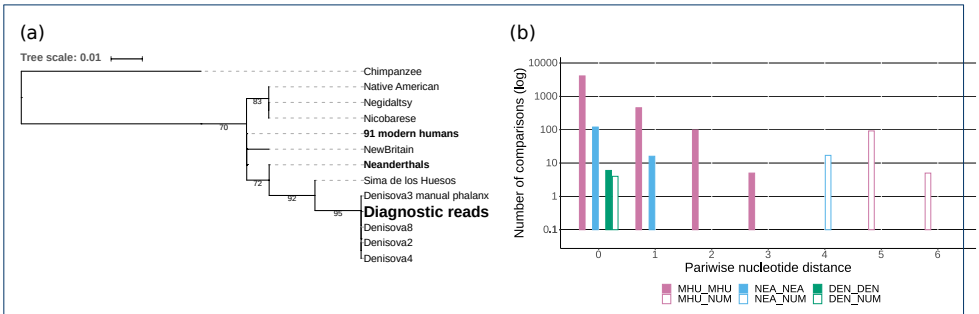

Figure 7 Phylogenetic analysis of diagnostic reads. A potential Denisovan NUMT inferred from diagnostic alleles. (a) Maximum likelihood tree for a 142 bp sequence generated from reads that contain the diagnostic allele for the Denisovan-Sima de los Huesos branch at position 9884. Bootstrap values over 50 are indicated at branch locations. (b) Pairwise nucleotide distances vs. frequency (in logarithmic scale) within and between 97 modern humans (MHU, blue), 17 Neanderthals (NEA, yellow), four Denisovans (DEN, red) and a nuclear insert of mitochondrial DNA (NUMT) (NUM, empty bars).

For some of these positions, the sequence generated for the surrounding region falls outside of modern human variation and shows a phylogeny more similar to Denisovans and Sima de los Huesos than to modern humans (Figure 7). This phylogeny, and the fact that the coverage is similar to the chromosomal coverage, suggests that these reads might originate from NUMTs that introgressed from an archaic hominin. They might belong to a detected NUMT which is too long to be covered by split reads, and the end of the NUMT might consist of conserved mtDNA regions that therefore do not distinguish this NUMT from NUMTs arising from other hominins.

Discussion

In the human reference genome 755 NUMTs are annotated [7]. Most of these are fixed, with only 14 polymorphic in modern humans [16]. Our study focused solely on polymorphic insertions not present in the reference, using an approach that detects NUMTs in next generation sequencing data by analysing split reads [8]. We were able to discover a variety of polymorphic NUMTs in Indonesian and Oceanian genomes. Of these NUMTs, only 25 are also present in the 1000 GP samples. In addition to the 741 fixed NUMTs in the reference genome, our results suggest the presence of $\sim$ 16 polymorphic NUMTs on average in each human genome. This number might be even higher, as *dinumt* is not able to detect insertions in highly

repetitive genomic regions. Most of these polymorphic NUMTs were only found in one study (Table 1), and therefore seem to be population-specific, further supporting their use as phylogenetic markers in modern human populations [15, 16]. The NUMTs that are specific to only some populations are most likely to have inserted around or after modern human dispersal from sub-Saharan Africa. These results provide further evidence for the ongoing transfer of mtDNA into the human nuclear genome even during recent human history [4, 6, 8].

A previous study of NUMTs in low to medium-coverage genomes could not find a strong correlation between average depth and detection rate [8]; however we find that with high-coverage genomes there is a 10-fold increase in the number of detected NUMTs. This is presumably because coverage depth is not uniform for all positions across the genome, but follows a Poisson distribution [33]. To detect a NUMT at a certain position, *dinumt* needs a minimum of 5 clipped reads. For the high-coverage samples we used, more than 95 % of the genome is expected to be covered sufficiently to detect a NUMT. For a medium coverage of around 10x this fraction would be around 80 %, and for low coverage less than 50 % [34].
For the 1000 GP dataset, the rarefaction curves in Figure 2a show that analysing more samples in a population does not necessarily lead to the discovery of many more NUMTs. This could mean that for some NUMTs to be detected, more samples cannot compensate for the lack of coverage. Merging samples from one population might be a solution to obtain sufficient read support to detect more NUMTs. In addition, the populations analysed differ among studies and thus might also influence a comparison of NUMT diversities on a smaller scale. In the 1000 GP many individuals were sampled from a few populations. The SGDP contains many more populations from around the world but many fewer individuals per population, and the Indonesian Genome Diversity Project (IGDP) [28] and the study of Vernot et al. [26] both sample much more intensively over a more focused geographic scale.

We identified three putative ancestral polymorphic NUMTs, two of which were also detected by Dayama et al. [8]. Their presence in sub-Saharan Africa and other populations around the world suggests their insertion in an ancestor of all modern humans, rather than as a result of archaic introgression. Their sequences place them outside of modern humans; however, in one case this classification was based on a difference at only one position. Even for these older polymorphic insertions, exact classification is not always possible due to the lack of sequence variation. We found an additional 23 NUMTs which were also detected by Dayama et al. [8] in samples from each continental group of the 1000 GP and thus these might also be of ancestral origin. Most of these are either not classifiable or were classified as of modern human origin.

The identification of ancient and archaic mtDNA in modern human genomes is constrained by two factors. On the one hand, conserved mtDNA is difficult to classify. On the other hand, strongly diverged sequences might not be detectable with the method we used. To be detected, their reads have to align to the mtDNA reference sequence, but if the NUMT sequence is too diverged from the reference,

it might not align. Additionally, a read spanning the NUMT and the nuclear DNA has to be clipped, which will further decrease its mappability. Mapping approaches are in general optimized to map reads to a quite similar reference and not a highly diverged one. Changing the strategy by reducing the mismatch and clipping penalties might facilitate the mapping of such reads, and therefore enable more diverged sequences to be discovered [35, 21].
Another possibility would be to use another reference more similar to the target sequences. The remapping of some genomes to a reference genome containing Denisovan mtDNA instead of the revised Cambridge Reference Sequence (rCRS) demonstrates the impact of the reference on the detection of NUMTs. Although Denisovan and modern human mtDNA only show 3.5 mismatches on average for a read length of 150 bp [36], many modern human reads do not align to a Denisovan reference. No new NUMT could be detected, but the ability to detect modern human NUMTs strongly decreased. It might therefore be the case that there are more Denisovan NUMTs in the analysed genomes which could not be detected, even though we did not find any in 10 samples that were re-analysed with the Denisovan mtDNA as a reference. It could therefore be useful to explore the impact of reference bias in future studies.

Using short read sequencing technology restricts the maximum length for generated NUMT sequences to around 1,000 bp. Among the detected NUMTs, some were too long to be covered completely by split reads. For these we could only analyse the ends, which might not contain enough phylogenetic information to confidently classify them as to the hominin lineage of origin. By analysing reads which map to diagnostic positions, we were able to identify additional reads that might originate from Denisovan mtDNA. The presence of such reads suggests that there are additional Denisovan NUMTs present in the analysed genomes, but with our methods we could not confidently identify them as of Denisovan origin. These reads probably belong to longer NUMTs which are not fully covered by split reads. Conserved regions between the ends of these reads make it difficult to detect them, as they would also allow modern human reads to overlap. The advances in long-read sequencing technologies might enable the detection of such proposed long archaic NUMTs in the future.

Despite constraints in detecting NUMTs and the difficulties in classification of short mtDNA subsets, we were able to detect one NUMT that probably originated from Denisovan mtDNA. Its sequence contains five informative positions which classify it as Denisovan mtDNA, as opposed to originating from a common ancestor of modern humans and Denisovans. To further investigate how this NUMT ended up in a modern human genome, we considered two potential explanations. The first is that the insertion happened in a Denisovan individual (Figure 1a) and that the NUMT later introgressed into modern humans within the nuclear Denisovan genome. Accordingly, the genomic region flanking the insert would also be introgressed and therefore should exhibit Denisovan ancestry. The other possibility is that a Denisovan female interbred with a modern human male (Figure 1b). The maternal lineage of descendants of this interbreeding event would carry Denisovan

mtDNA, from which a piece was inserted into the nuclear genome. Here, the flanking region could be of either modern human origin, or it could be an introgressed region, as the descendants of such an interbreeding event would also contain introgressed Denisovan nuclear DNA where such an insertion could happen. This becomes more and more unlikely with every subsequent interbreeding of this lineage's offspring with modern humans, as the fragments of Denisova nuclear DNA would decline in length and number through segregation and recombination.
Analysing the flanking region of the NUMT confidently classified as of Denisovan origin indicates a Denisovan ancestry. Although the second hypothesis cannot be fully rejected, it seems to be more likely, that the archaic NUMT is part of an introgressed haplotype.

Our results indicate that there are potentially many more NUMTs to be discovered via sequencing of additional populations to high coverage. Moreover, application of long-read technologies should also increase the number of detected NUMTs, promising to provide more insights into the introgression history of archaic and modern humans.

Conclusions

We modified an existing method to detect NUMTs in next-generation sequence data, and applied the method to whole genome sequences from Indonesians and Papuans, in order to detect NUMTs arising from archaic human mtDNA. In high coverage genomes, an average of 16 NUMTs per individual is detectable. Most of these NUMTs seem to be population specific, indicating their insertion in recent human history. This finding further supports previous findings of an ongoing transfer of mtDNA to the nucleus in humans and suggests that the analysis of additional populations would lead to the discovery of many more NUMTs. A Denisovan NUMT could be identified in 16 samples from Indonesia and Oceania. Analyses of the flanking region of this NUMT reveals that it is part of a Denisovan haplotype. This suggests that the insertion of the NUMT most likely happened in a Denisovan individual and then introgressed into modern humans within nuclear DNA. Our pipeline can be applied to newly sequenced genomes in the future, which could reveal additional archaic NUMT insertions and new insights into the nature of interbreeding events.

Methods

Data analysis

To detect NUMTs, paired-end whole genome sequence reads aligned against the human reference genome version hg19 in BAM-format were used. In total, 221 Oceanian and Indonesian genomes were analysed for archaic NUMTs. 35 Papuan genomes with a median sequencing depth of 38 (min=33, Q1=35, Q3=39, max=43) were obtained from Vernot et al. [26], 25 Oceanian genomes with a median sequencing depth of 44 (min=34, Q1=42, Q3=45, max=51) from the SGDP [29] and 161 Indonesian and Papuan genomes with a median sequencing depth of 38 (min=18,

Q1=35, median=38, Q3=43, max=48) from the IGDP [28] (Supplementary Table S4). For rarefaction analysis, an additional 40 sub-Saharan African, 19 native North American and 39 South Asian genomes from the SGDP were obtained (Supplementary Table S5).
To determine the phase of the NUMTs and to analyse the ancestry of flanking regions, phased genotypes were obtained for all available samples from the 1000 GP [30], the SGDP, Vernot et al. [26] and the IGDP.
If not mentioned otherwise, custom Python scripts were used to conduct all analyses. These scripts are available at https://github.com/robbueck/arcnumt.

NUMT detection and analysis

For each sample, mean insert size and standard deviation of the read pairs were calculated using Picard CollectInsertSizeMetrics version: 2.17.10 (http://broadinstitute.github.io/picard/). NUMTs were detected using the *dinumt* software package [8]. The program detects NUMTs that are not annotated in the human reference genome by identifying read pairs where one end aligns to the mtDNA and its mate aligns to the nuclear genome. The amount of mismatches, gaps and clipping allowed depended on the mapping processes used by the three different studies. Each sample was analysed for NUMTs individually following the procedure described in Dayama et al. [8].

Rarefaction analysis and downsampling

Rarefaction curves were compared between studies to estimate the effect of coverage on the detection of NUMTs, and compared between regions within the SGDP dataset to estimate the difference between geographic regions. To determine if the number of detected NUMTs per genome had reached a threshold, which would indicate that increased sequencing would not reveal more NUMTs, NUMT rarefaction analysis was performed for each larger geographic region in each analysed dataset, including results for the 1000 GP dataset obtained from Dayama et al. [8]. Additional NUMT detection and rarefaction was done for all publicly-available sub-Saharan African, American and South Asian samples from the SGDP dataset (Supplementary Table S5). Samples were grouped into larger geographic regions, which were analysed individually for each study (SGDP: sub-Saharan Africa, America, Oceania, South Asia; 1000 GP: sub-Saharan Africa, America, Asia, Europe; IGDP: West Island South East Asia (ISEA), East ISEA, Oceania; Vernot et al. [26]: Oceania). For each group from each study, resampling was performed by successively adding all samples to the dataset. For each sample size, the number of different NUMTs in these samples were counted. The mean of 100 repetitions was calculated for each sample size and plotted in Figure 2. A downsampling analysis was performed to further investigate the effect of coverage on the detection of NUMTS: the 25 Oceanian genomes from the SGDP were downsampled to lower coverages (Supplementary Figure S1) using samtools version 1.3.1 [37]. NUMT detection was performed as described above.

Reconstruction of NUMT sequences

A SAM-file containing the insertion supporting reads for each sample was obtained from *dinumt*. These reads were clustered according to the NUMT insertion they

supported. As no NUMT insertions closer than 50 kbp were observed in our data, mitochondrial reads with mates mapping on the same chromosome within 2 kbp of each other were considered as supporting the same NUMT insertion and clustered together. The chromosome and the first four digits of the position of a NUMT were used to name it, e.g. a NUMT inserted at position 13848625 on chromosome 3 is named NUMT 3_1384. These mitochondrial reads were then mapped to the Reconstructed Sapiens Reference Sequence (RSRS) [38] of the mitochondrial genome, which represents the ancestral modern human mtDNA sequence, using BWA-MEM [39]. As the mitochondiral genome is circular, reads originating from the parts that are located at the beginning or the end of the reference genome would not map properly. Therefore nucleotide positions 1-1000 of the RSRS were copied and inserted after position 16569, to allow unbiased mapping to a circular genome. From the mapping output the RSRS coordinates for regions with a coverage of at least five reads were extracted. Positions with lower coverage were excluded to minimize the effect of sequencing errors. The NUMT sequence for these regions was obtained using GATKs HaplotypeCaller version 4.0.0 and FastaAlternativeReferenceMaker version 3.8.0 [40]. HaplotypeCaller was used to call variants for the mapping output. For each NUMT within one sample no heterozygous alleles were called, thus enabling the construction of one unambiguous consensus sequence for each NUMT within one sample using FastaAlternativeReferenceMaker. For some NUMTs, the sequence was broken down to multiple parts as the NUMT was too long to be fully covered by short reads, or not enough read coverage was available for each position of the NUMTs. In these cases, multiple sequences were obtained for one NUMT, which were concatenated according to their order on the RSRS, taking into account the circular nature of mtDNA.

Phylogeny of NUMT insertions

An alignment of mitochondrial genomes of 87 present day modern humans, 17 Neanderthals, ten ancient modern humans, four Denisovans, the Sima de los Huesos fossil, chimpanzee, the rCRS and the RSRS (Supplementary Table S6) was produced using MUSCLE version 3.8 [41]. For each NUMT, the corresponding mtDNA region was extracted from the alignment by using the RSRS coordinates adjusted for gaps in the aligned RSRS. The NUMT sequence was aligned with the corresponding mtDNA regions using MUSCLE version 3.8 [41]. Alignments were cleaned by removing identical sequences. If two or more sequences were identical, only one copy was kept along with the taxonomic information of the removed sequences. For each cleaned alignment containing more than three sequences, trees were built using RAxML version 8.2 [42]. Pairwise nucleotide distances within and between modern humans, Neanderthals, Denisovans and the NUMTs were calculated. In each possible sequence pair, the number of sites where the two sequences differed from each other were counted. A graphical overview of the pipeline is shown in Supplementary Figure S6.

Determining the phase of NUMT insertions

We determined the position of NUMTs within available phased chromosomal data by analysing NUMT reads that covered informative heterozygous positions that

flanked the insertion point. The genotypes and the NUMT-reads were visualized together using the IGV Browser [43, 44]. The phase of a genotype supported by more than two thirds of all reads, and at least three NUMT reads, was assumed to be the phase of the NUMT insertion.

Flanking region analysis

The flanking regions of the Denisovan NUMT were analysed for Denisovan ancestry in each sample with the NUMT. Therefore phased genotype data was obtained for all samples analysed for archaic NUMTs and additionally for all available 1000 GP and SGDP samples. Each phase in each sample was separately compared with the published Denisovan genome sequence [21]. Within a window of 20 kbp before and after the insertion site, all positions where a modern human shares a non-reference allele with the Denisovan genome were identified.

To detect alleles shared with Denisovans due to introgression, we calculated the frequency of all shared alleles in the 1000 GP and SGDP datasets. Alleles with a frequency below 0.05 in at least one of the two studies were considered to be shared likely due to introgression. The distribution of these low-frequency alleles around the insertion site of an archaic NUMT was used to estimate the boundary of a potentially introgressed haplotype as shown in Figure 5.

We further analysed for Denisovan ancestry using a match ratio. For this analysis in a 20 kbp region around the insertion site each phase for each sample in all four datasets was individually compared to the Denisovan genome. In total, 2973 phased genomes were analysed. Unphased sites were excluded from the analysis. A match ratio was calculated as the proportion of sites where a phased genome and Denisovan shared a non-reference allele, compared to all sites where the phased genome or Denisovan contained a non-reference allele. For samples which are heterozygous for a Denisovan NUMT insertion, the match ratios of both phased regions were compared using a one tailed paired t-test. Under the assumption that the NUMT is part of an introgressed haplotype, the phased region without the Denisovan NUMT should show a significantly lower match ratio than the phased region with the Denisovan NUMT.

Evaluation of potential biases

To examine the influence of the reference genome on the detection of archaic NUMTs, ten Oceanian genomes (Supplementary Table S3) were mapped against the hg19 reference genome with the mtDNA replaced by the mtDNA of Denisova8 (accession number: KT780370.1) using BWA-MEM [39]. NUMT detection and analysis were performed as described above. For each mapping file, coverage per position along the mtDNA was calculated using GATK DepthOfCoverage version 3.8 [40].

Analysis of diagnostic alleles

As an alternative approach for detecting NUMTs from archaic humans, we screened genomes for reads mapping to the mitochondrial genome containing alleles specific to different branches of the phylogenetic tree of hominin mtDNA sequences. Diagnostic alleles on the mitochondrial genome for the branches in the mtDNA phylogenetic tree of Neanderthals, Denisovans, the Sima de los Huesos fossile and Denisovans-Sima de los Huesos were obtained from Meyer et al. [31]. For all samples from Vernot et al. [26], the reads were remapped to the rCRS using BWA-MEM [39]. For each set of diagnostic alleles, a pileup was performed at the position of each allele. All reads containing a diagnostic allele with a minimum base quality of 15 were extracted. Positions where less than five reads supported the diagnostic allele were discarded. For the region around each remaining position, a consensus sequence was constructed using the extracted reads and its phylogeny was analysed as described above for NUMTs.

List of abbreviations

1000 GP	1000 Genomes Project
BAC	Bacterial Artificial Chromosome
bp	base pairs
HGDP	Human Genome Diversity Project
ISEA	Island South East Asia
mtDNA	mitochondrial DNA
NUMT	Nuclear Mitochondiral DNA Segment
rCRS	revised Cambridge Reference Sequence
RSRS	Reconstructed Sapiens Reference Sequence
SGDP	Simons Genome Diversity Project

Declarations

Ethics approval and consent to participate
No new samples or data were collected for this study; all of the sequence information came from previous studies that have made the data publicly-available, and that obtained relevant ethical permission and informed consent for producing and analyzing genomic data.

Consent for publication
Not applicable

Availability of data and materials
The genomes from the Simons Genome Diversity Project analysed in this study were taken from the European Nucleotide Archive, accession number ERP010710 [29]. The genomes from Vernot et al. [26] and the Indonesian Genome Diversity Project [28] were obtained from the authors on request and are available from the Database of Genotypes and Phenotypes (dbGAP) accession number phs001085.v1.p1 (Vernot et. al) and the European Genome-phenome Archive (EGA), accession number EGAS00001003054 (Indonesian Genome Diversity Project). NUMT data for the 1000 Genomes Project was taken from Dayama et al., supplementary data (https://doi.org/10.1093/nar/gku1038) [8].
The NUMT-dataset generated during the current study and an alignment of the Denisovan NUMT 3_1384 with other human mtDNA are available in Supplementary Files 2 and 3. All scripts used in this analysis including a description of usage are available at https://github.com/robbueck/arcnumt All other datasets generated during this study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Funding
MS and RB were supported by funds from the Max Planck Society; MPC, GH, LS and HS by the Estonian Research Council grant PUT (PRG243), by the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.15-0012), and a Royal Society of New Zealand Marsden Grant 17-MAU-040; MPC was

additionally supported by a German Alexander von Humboldt Foundation fellowship. Computational resources for the analysis of the IGDP samples were provided by the High Performance Computing Center, University of Tartu, Estonia. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author's contributions

MS supervised the study. MPC, GH, LS and HS provided the Indonesian Genome Diversity Project data and helped with their analysis. RB performed the analyses. RB and MS wrote the manuscript with help from all co-authors. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Alexander Hübner and Benjamin Vernot for technical support, helpful discussions and feedback concerning this work.

Author details

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D04103 Leipzig, Germany. [2]School of Fundamental Sciences, Massey University, 4442 Palmerston North, New Zealand. [3]Institute of Genomics, University of Tartu, 51010 Tartu, Estonia. [4]Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, 10430 Jakarta, Indonesia. [5]Department of Medical Biology, Faculty of Medicine, University of Indonesia, 10430 Jakarta, Indonesia. [6]Sydney Medical School, University of Sydney, NSW 2006 Sydney, Australia.

References

1. Bensasson, D., Zhang, D.-X., Hartl, D.L., Hewitt, G.M.: Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends in Ecology & Evolution **16**(6), 314–321 (2001). doi:10.1016/S0169-5347(01)02151-6. Accessed 2018-11-05
2. Richly, E., Leister, D.: NUMTs in sequenced eukaryotic genomes. Molecular biology and evolution **21**(6), 1081–1084 (2004)
3. Mourier, T., Hansen, A.J., Willerslev, E., Arctander, P.: The Human Genome Project Reveals a Continuous Transfer of Large Mitochondrial Fragments to the Nucleus. Molecular Biology and Evolution **18**(9), 1833–1837 (2001). doi:10.1093/oxfordjournals.molbev.a003971. Accessed 2018-11-05
4. Ricchetti, M., Tekaia, F., Dujon, B.: Continued Colonization of the Human Genome by Mitochondrial DNA. PLOS Biology **2**(9), 273 (2004). doi:10.1371/journal.pbio.0020273. Accessed 2018-02-01
5. Hazkani-Covo, E., Covo, S.: Numt-Mediated Double-Strand Break Repair Mitigates Deletions during Primate Genome Evolution. PLOS Genetics **4**(10), 1000237 (2008). doi:10.1371/journal.pgen.1000237. Accessed 2018-02-12
6. Hazkani-Covo, E., Zeller, R.M., Martin, W.: Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. PLOS Genetics **6**(2), 1000834 (2010). doi:10.1371/journal.pgen.1000834. Accessed 2018-02-12
7. Calabrese, F.M., Simone, D., Attimonelli, M.: Primates and mouse NumtS in the UCSC Genome Browser. BMC Bioinformatics **13**(4), 15 (2012). doi:10.1186/1471-2105-13-S4-S15. Accessed 2018-11-01
8. Dayama, G., Emery, S.B., Kidd, J.M., Mills, R.E.: The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Research **42**(20), 12640–12649 (2014). doi:10.1093/nar/gku1038. Accessed 2017-11-20
9. Tsuji, J., Frith, M.C., Tomii, K., Horton, P.: Mammalian NUMT insertion is non-random. Nucleic Acids Research **40**(18), 9073–9088 (2012). doi:10.1093/nar/gks424. Accessed 2018-10-27
10. Perna, N.T., Kocher, T.D.: Mitochondrial DNA: molecular fossils in the nucleus. Current Biology **6**(2), 128–129 (1996)
11. Pakendorf, B., Stoneking, M.: Mitochondrial DNA and human evolution. Annual Review of Genomics and Human Genetics **6**, 165–183 (2005). doi:10.1146/annurev.genom.6.080604.162249
12. Keller, I., Bensasson, D., Nichols, R.A.: Transition-Transversion Bias Is Not Universal: A Counter Example from Grasshopper Pseudogenes. PLOS Genetics **3**(2), 22 (2007). doi:10.1371/journal.pgen.0030022. Accessed 2018-11-05
13. Schmitz, J., Piskurek, O., Zischler, H.: Forty Million Years of Independent Evolution: A Mitochondrial Gene and Its Corresponding Nuclear Pseudogene in Primates. Journal of Molecular Evolution **61**(1), 1–11 (2005). doi:10.1007/s00239-004-0293-3. Accessed 2018-11-05
14. Zischler, H., Geisert, H., von Haeseler, A., Pääbo, S.: A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. Nature **378**(6556), 489–492 (1995). doi:10.1038/378489a0. Accessed 2018-03-04
15. THOMAS, R., ZISCHLER, H., PÄÄBO, S., STONEKING, M.: Novel Mitochondrial DNA Insertion Polymorphism and Its Usefulness for Human Population Studies. Human Biology **68**(6), 847–854 (1996). Accessed 2018-03-04
16. Lang, M., Sazzini, M., Calabrese, F.M., Simone, D., Boattini, A., Romeo, G., Luiselli, D., Attimonelli, M., Gasparre, G.: Polymorphic NumtS trace human population relationships. Human Genetics **131**(5), 757–771 (2012). doi:10.1007/s00439-011-1125-3. Accessed 2018-03-04
17. Baldo, L., de Queiroz, A., Hedin, M., Hayashi, C.Y., Gatesy, J.: Nuclear–Mitochondrial Sequences as Witnesses of Past Interbreeding and Population Diversity in the Jumping Bristletail Mesomachilis. Molecular Biology and Evolution **28**(1), 195–210 (2011). doi:10.1093/molbev/msq193. Accessed 2018-11-05
18. Wang, B., Zhou, X., Shi, F., Liu, Z., Roos, C., Garber, P.A., Li, M., Pan, H.: Full-length Numt analysis provides evidence for hybridization between the Asian colobine genera Trachypithecus and Semnopithecus. American Journal of Primatology **77**(8), 901–910 (2015). doi:10.1002/ajp.22419. Accessed 2018-11-05
19. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B.,

Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., Rasilla, M.d.l., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., Pääbo, S.: A Draft Sequence of the Neandertal Genome. Science **328**(5979), 710–722 (2010). doi:10.1126/science.1188021. Accessed 2018-11-06

20. Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.-J., Kelso, J., Slatkin, M., Pääbo, S.: Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature **468**(7327), 1053–1060 (2010). doi:10.1038/nature09710. Accessed 2018-02-12
21. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., Filippo, C.d., Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andrés, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J., Pääbo, S.: A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science **338**(6104), 222–226 (2012). doi:10.1126/science.1224344. Accessed 2018-08-30
22. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J.C., Vohr, S.H., Green, R.E., Hellmann, I., Johnson, P.L.F., Blanche, H., Cann, H., Kitzman, J.O., Shendure, J., Eichler, E.E., Lein, E.S., Bakken, T.E., Golovanova, L.V., Doronichev, V.B., Shunkov, M.V., Derevianko, A.P., Viola, B., Slatkin, M., Reich, D., Kelso, J., Pääbo, S.: The complete genome sequence of a Neanderthal from the Altai Mountains. Nature **505**(7481), 43–49 (2014). doi:10.1038/nature12886. Accessed 2018-11-03
23. Qin, P., Stoneking, M.: Denisovan Ancestry in East Eurasian and Native American Populations. Molecular Biology and Evolution **32**(10), 2665–2674 (2015). doi:10.1093/molbev/msv141. Accessed 2018-10-25
24. Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., Nickel, B., Peltzer, A., Rohland, N., Slon, V., Talamo, S., Lazaridis, I., Lipson, M., Mathieson, I., Schiffels, S., Skoglund, P., Derevianko, A.P., Drozdov, N., Slavinsky, V., Tsybankov, A., Cremonesi, R.G., Mallegni, F., Gély, B., Vacca, E., Morales, M.R.G., Straus, L.G., Neugebauer-Maresch, C., Teschler-Nicola, M., Constantin, S., Moldovan, O.T., Benazzi, S., Peresani, M., Coppola, D., Lari, M., Ricci, S., Ronchitelli, A., Valentin, F., Thevenet, C., Wehrberger, K., Grigorescu, D., Rougier, H., Crevecoeur, I., Flas, D., Semal, P., Mannino, M.A., Cupillard, C., Bocherens, H., Conard, N.J., Harvati, K., Moiseyev, V., Drucker, D.G., Svoboda, J., Richards, M.P., Caramelli, D., Pinhasi, R., Kelso, J., Patterson, N., Krause, J., Pääbo, S., Reich, D.: The genetic history of Ice Age Europe. Nature **534**(7606), 200–205 (2016). doi:10.1038/nature17993. Accessed 2018-12-24
25. Kuhlwilm, M., Gronau, I., Hubisz, M.J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H.A., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Rudan, P., Brajkovic, D., Kucan, Z., Gušic, I., Marques-Bonet, T., Andrés, A.M., Viola, B., Pääbo, S., Meyer, M., Siepel, A., Castellano, S.: Ancient gene flow from early modern humans into Eastern Neanderthals. Nature **530**(7591), 429–433 (2016). doi:10.1038/nature16544. Accessed 2018-12-24
26. Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H., Scheinfeldt, L.B., Merriwether, D.A., Koki, G., Friedlaender, J.S., Wakefield, J., Pääbo, S., Akey, J.M.: Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science **352**(6282), 235–239 (2016). doi:10.1126/science.aad9416. Accessed 2017-12-06
27. Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S., Akey, J.M.: Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell **173**(1), 53–619 (2018). doi:10.1016/j.cell.2018.02.031. Accessed 2018-12-24
28. Jacobs, G.S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C.C., Lawson, D.J., Mondal, M., Pagani, L., Ricaut, F.-X., Stoneking, M., Metspalu, M., Sudoyo, H., Lansing, J.S., Cox, M.P.: Multiple deeply divergent Denisovan ancestries in Papuans. Cell (2019)
29. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J.P., Song, Y.S., Poletti, G., Balloux, F., Driem, G.v., Knijff, P.d., Romero, I.G., Jha, A.R., Behar, D.M., Bravi, C.M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O.L., Balanovska, E., Balanovsky, O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah, M.S., Ruiz-Linares, A., Beall, C.M., Rienzo, A.D., Jeong, C., Starikovskaya, E.B., Metspalu, E., Parik, J., Villems, R., Henn, B.M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J.T.S., Khusainova, R., Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M.F., Kivisild, T., Klitz, W., Winkler, C.A., Labuda, D., Bamshad, M., Jorde, L.B., Tishkoff, S.A., Watkins, W.S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K., Pääbo, S., Kelso, J., Patterson, N., Reich, D.: The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature **538**(7624), 201–206 (2016). doi:10.1038/nature18964. Accessed 2018-07-10
30. The 1000 Genomes Project Consortium, T.: A global reference for human genetic variation. Nature **526**(7571), 68–74 (2015). doi:10.1038/nature15393. Accessed 2018-07-11
31. Meyer, M., Arsuaga, J.-L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., de Castro, J.M.B., Carbonell, E., Viola, B., Kelso, J., Prüfer, K., Pääbo, S.: Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature **531**(7595), 504–507 (2016). doi:10.1038/nature17405. Accessed 2018-02-12
32. Li, H.: Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics **30**(20), 2843–2851 (2014). doi:10.1093/bioinformatics/btu356. Accessed 2018-11-08
33. Lander, E.S., Waterman, M.S.: Genomic mapping by fingerprinting random clones: A mathematical analysis.

Genomics **2**(3), 231–239 (1988). doi:10.1016/0888-7543(88)90007-9. Accessed 2018-11-05
34. Ajay, S.S., Parker, S.C.J., Abaan, H.O., Fajardo, K.V.F., Margulies, E.H.: Accurate and comprehensive sequencing of personal genomes. Genome Research **21**(9), 1498–1505 (2011). doi:10.1101/gr.123638.111. Accessed 2018-11-05
35. Green, R.E., Malaspinas, A.-S., Krause, J., Briggs, A.W., Johnson, P.L.F., Uhler, C., Meyer, M., Good, J.M., Maricic, T., Stenzel, U., Prüfer, K., Siebauer, M., Burbano, H.A., Ronan, M., Rothberg, J.M., Egholm, M., Rudan, P., Brajković, D., Kućan, Z., Gušić, I., Wikström, M., Laakkonen, L., Kelso, J., Slatkin, M., Pääbo, S.: A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. Cell **134**(3), 416–426 (2008). doi:10.1016/j.cell.2008.06.021. Accessed 2018-11-04
36. Krause, J., Fu, Q., Good, J.M., Viola, B., Shunkov, M.V., Derevianko, A.P., Pääbo, S.: The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature **464**(7290), 894–897 (2010). doi:10.1038/nature08976. Accessed 2018-10-17
37. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) **25**(16), 2078–2079 (2009). doi:10.1093/bioinformatics/btp352
38. Behar, D., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N., Kivisild, T., Torroni, A., Villems, R.: A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. The American Journal of Human Genetics **90**(4), 675–684 (2012). doi:10.1016/j.ajhg.2012.03.002. Accessed 2018-08-01
39. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). doi:10.1093/bioinformatics/btp324. Accessed 2018-07-28
40. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**(9), 1297–1303 (2010). doi:10.1101/gr.107524.110. Accessed 2018-07-28
41. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32**(5), 1792–1797 (2004). doi:10.1093/nar/gkh340. Accessed 2018-07-28
42. Stamatakis, A.: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9), 1312–1313 (2014). doi:10.1093/bioinformatics/btu033. Accessed 2018-07-28
43. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer (2011). doi:10.1038/nbt.1754. `https://www.nature.com/articles/nbt.1754` Accessed 2018-08-28
44. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P.: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics **14**(2), 178–192 (2013). doi:10.1093/bib/bbs017. Accessed 2018-08-28

Supplementary Files

Supplementary file 1 — Supplementary Tables and Figures

Table S1. Summary of Oceanian and Indonesian samples analysed for archaic NUMTs. **Table S2.** Additional SGDP samples from Africa, America and south Asia used for rarefaction analysis. **Table S3.** Sequences in the alignment used for phylogenetic analysis of the NUMTs. **Table S4.** Samples from Vernot et al., 2016 mapped against a custom reference genome containing Denisovan mtDNA. **Table S5.** Summary of samples containing a putative Denisovan NUMT. **Table S6.** Results for determining the phase of NUMT 3_1384 in all containing samples. **Figure S1.** Number of NUMTs detected in downsampled genomes. **Figure S2.** Mean coverage (a), GC-content (b) and sequence length (c) for reconstructed NUMT sequences. **Figure S3.** Schematic tree for the hominin mitochondrial genome with putative NUMT insertions. **Figure S4.** Alignment of NUMT hs37d5_2745 with Denisovans, Neanderthals and modern humans. **Figure S5.** Per base coverage along Denisovan (mtDNA). **Figure S6.** Workflow for NUMT and flanking region analysis.

Supplementary file 2 — NUMTs detected in Indonesian and Oceanian samples

A VCF-file containing the insertion points of all NUMTs detected in Indonesian and Oceanian samples.

Supplementary file 3 — Alignment of NUMT 3_1384

Alignment of the Denisovan NUMT 3_1348 with 87 present day modern humans, 17 Neanderthals, ten ancient modern humans, four Denisovans, the Sima de los Huesos fossil, chimpanzee, the rCRS and the RSRS. GenBank accession numbers are give for each sample.

Figures

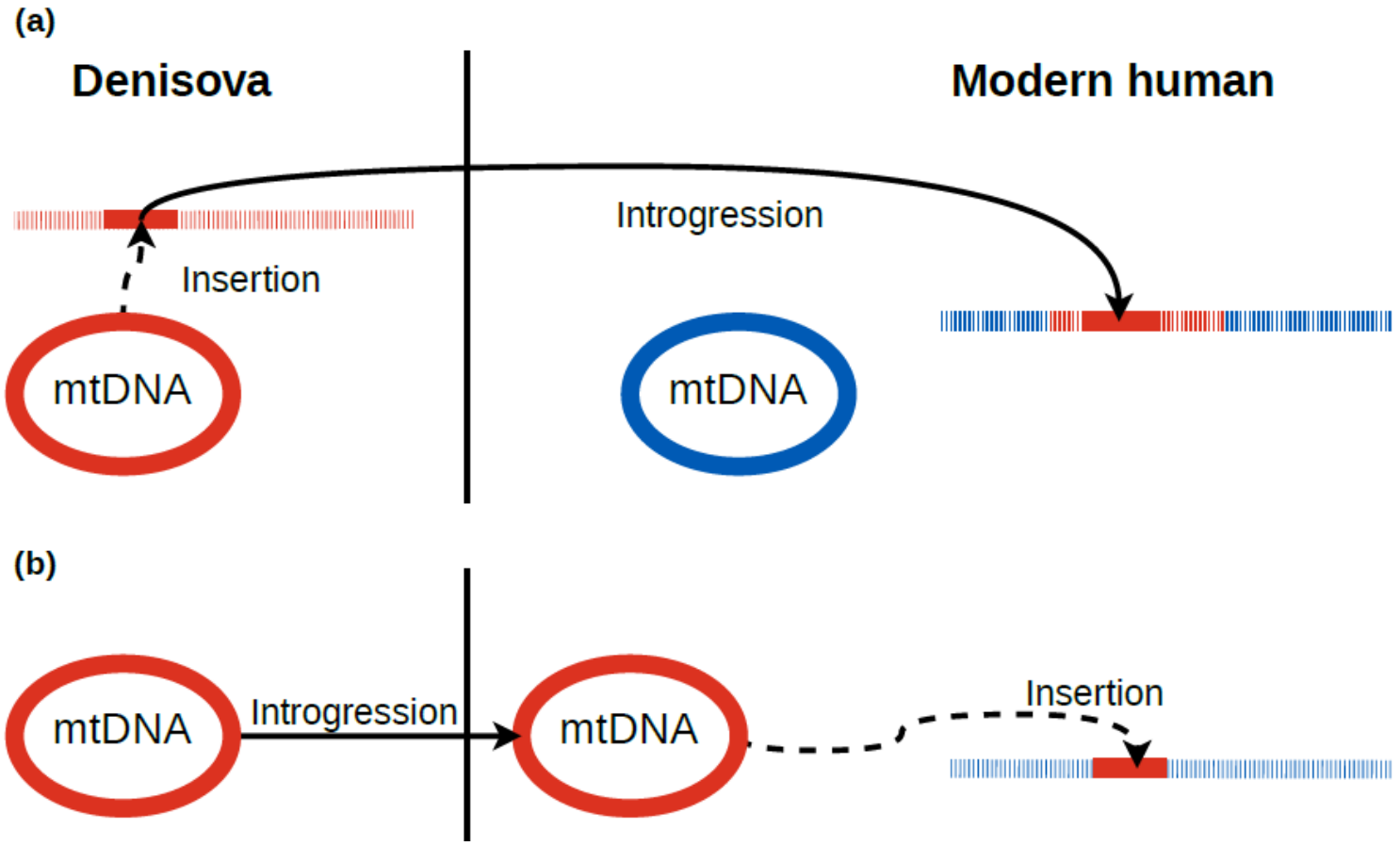

Figure 1

Hypotheses for NUMT introgression. Two possible scenarios for the introgression of Denisovan mitochondrial DNA (mtDNA) in modern human nuclear DNA. (a) The NUMT occurred in a Denisovan and introgressed together with its nuclear DNA. This path would result in a anking region with Denisovan ancestry. (b) The mtDNA introgressed into a modern human, where it subsequently inserted into the nuclear DNA. mtDNA: solid colours; nuclear DNA: striped; Denisovan DNA: red; modern human DNA: blue

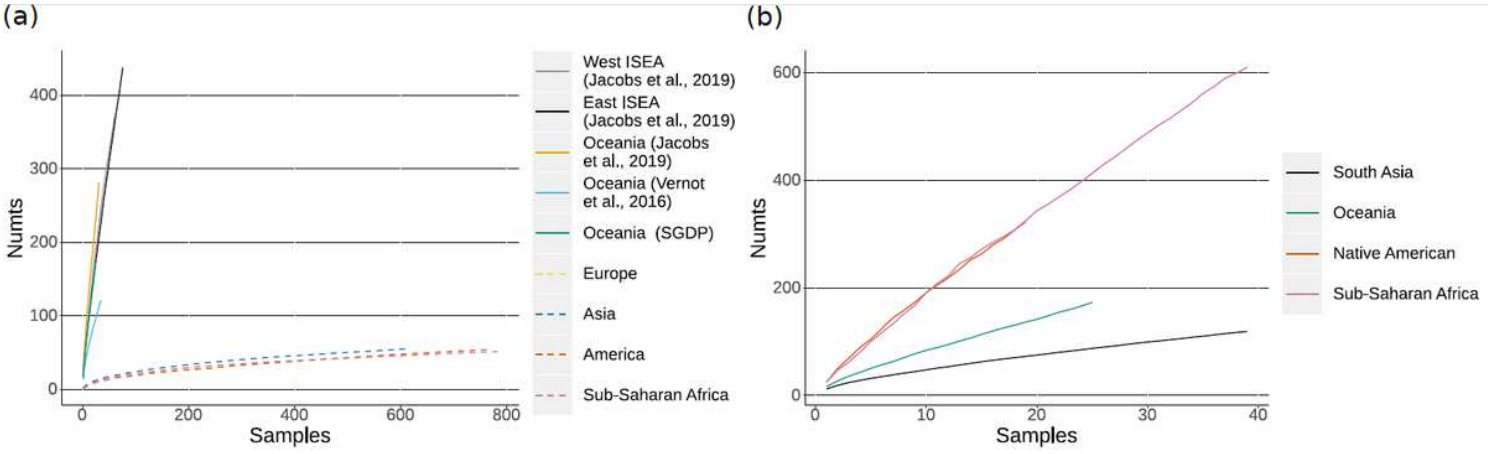

Figure 2

Comparison of rarefaction curves. Rarefaction curves plotted as the mean out of 100 repetitions. (a) Results from Island South East Asia (ISEA) and Oceania (continuous lines) are compared with worldwide 1000 GP data (dashed lines). (b) Comparison of di□erent geographic regions within the Simons Genome Diversity Project (SGDP) dataset.

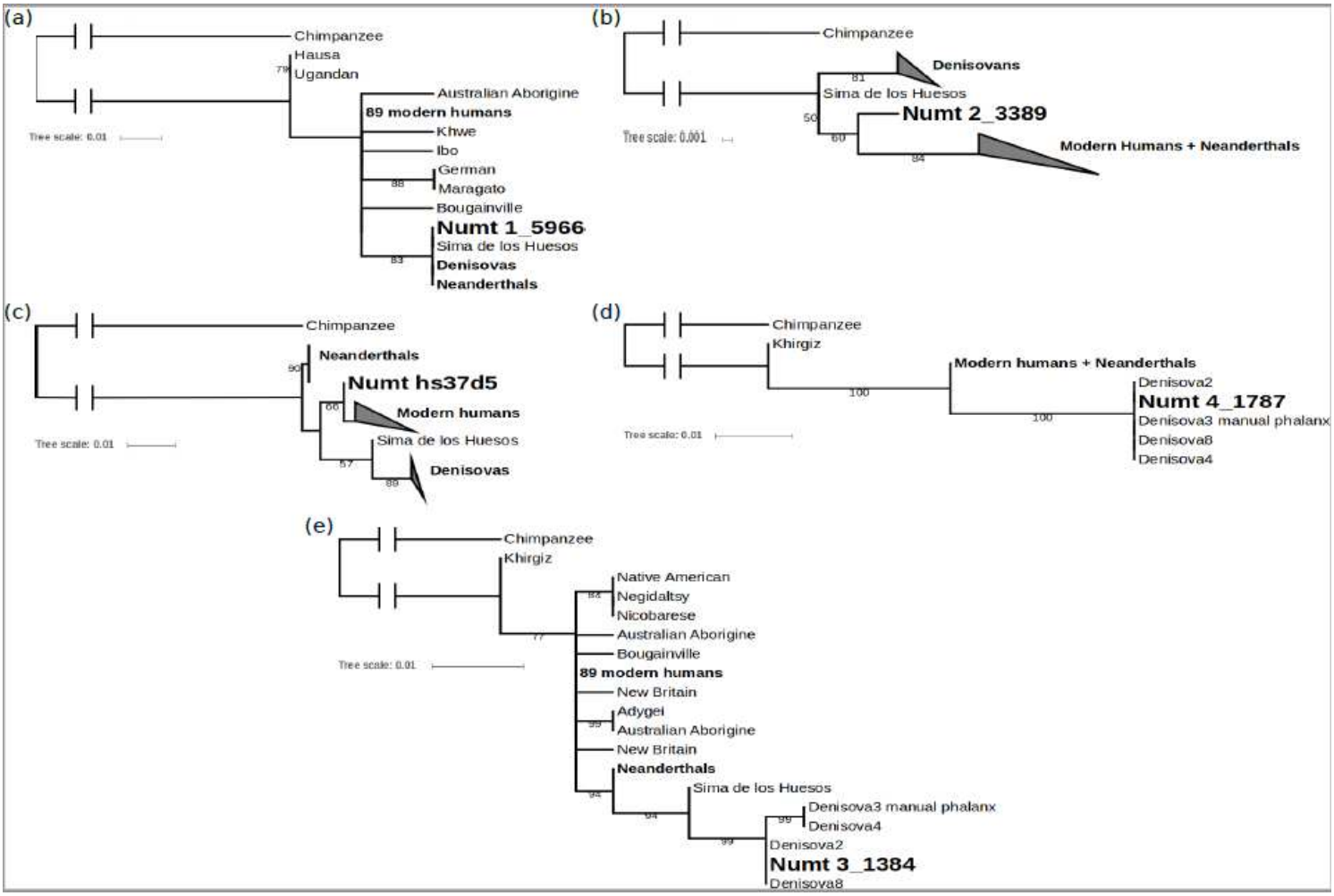

Figure 3

Phylogenetic trees for NUMTs. Maximum likelihood trees for putative ancestral NUMTs (a, b, c) and putative Denisovan NUMTs (d, e) in relationship with other hominin mtDNA sequences using distances based on nucleotide substitution rates. Ancestral NUMTs form a sister clade to at least all modern humans and are present in 1000 Genome Project (1000 GP) samples from around the world, except for (c). Denisovan NUMTs are more similar to Denisovan mtDNA than to modern human mtDNA and are not present in 1000 GP samples. Bootstrap values over 50 are indicated at branch locations.

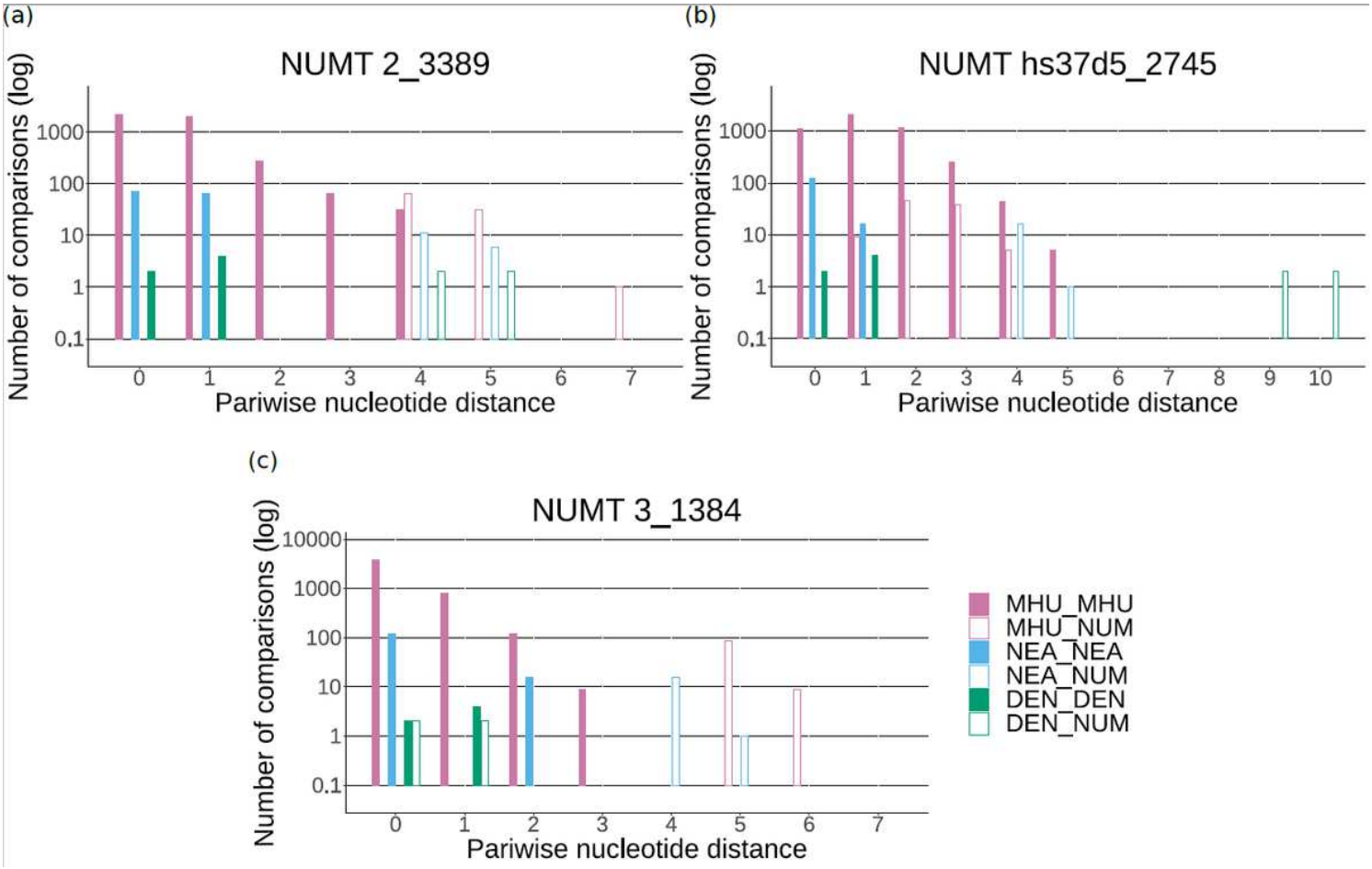

Figure 4

Pairwise nucleotide distances between NUMTs and mtDNA. Pairwise nucleotide distances vs. frequency (in logarithmic scale) within and between 97 modern humans (MHU, blue), 17 Neanderthals (NEA, yellow), four Denisovans (DEN, red) and a specic NUMT (NUM, empty bars) for two ancestral NUMTs (a, b) and one Denisovan NUMT (c).

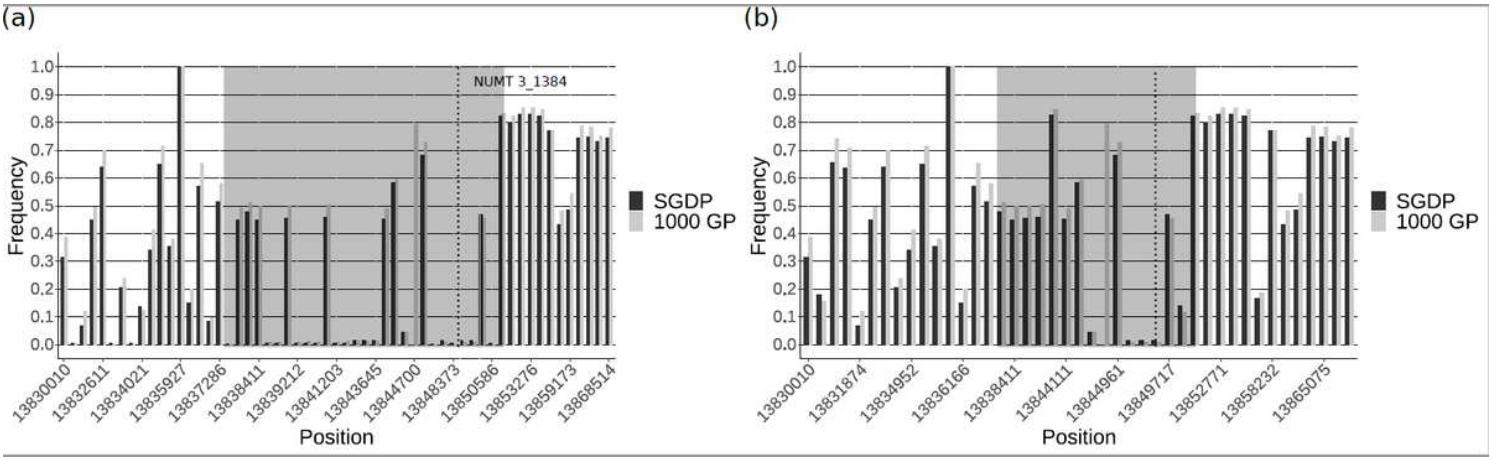

Figure 5

Worldwide frequencies of alleles shared with the Denisovan genome. Worldwide frequencies of alleles shared with the Denisovan genome in the region anking NUMT 3 1384 in (a) Oceanian sample UV925, and (b) the same region in the sub-Saharan African sample NA19309. The alleles are within 20 kbp before or after the insertion site (vertical dashed line) and on the same phase. Frequencies were calculated in the 1000 Genome Project (1000 GP) (black) and Simons Genome Diversity Project (SGDP) (grey) datasets. X-axis intervals are not linear, but indicate positions of shared alleles. High-frequency shared alleles re ect either homoplasy or incomplete lineage sorting; the greater abundance of low-frequency alleles shared with the Denisovan genome close to the insertion point (grey area) in the Oceanian sample vs. the sub-Saharan African sample suggests a Denisovan-introgressed haplotype in the Oceanian.

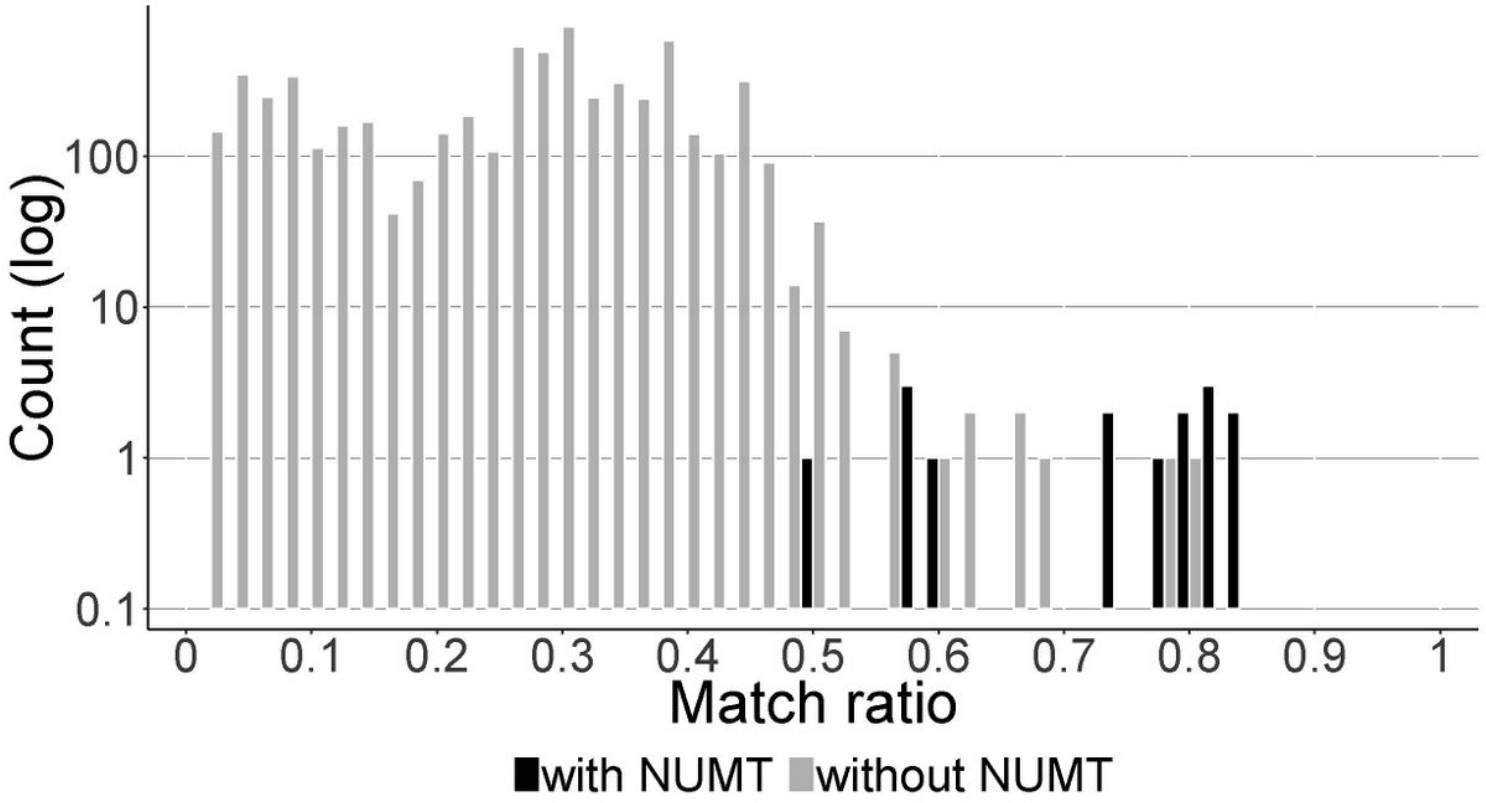

Figure 6

atch ratios of modern human genomes with the Denisovan Genome. Distributions of match ratios between modern humans and Denisovans for a 20 kbp region around NUMT 3 1384 in all samples from the 1000 GP, SGDP, IGDP and Verot et al. [26]. Match ratios were calculated for each phase individually counting shared non-reference alleles. Black bars represent haplotypes containing the NUMT insertion, grey bars represent haplotypes without the NUMT insertion.

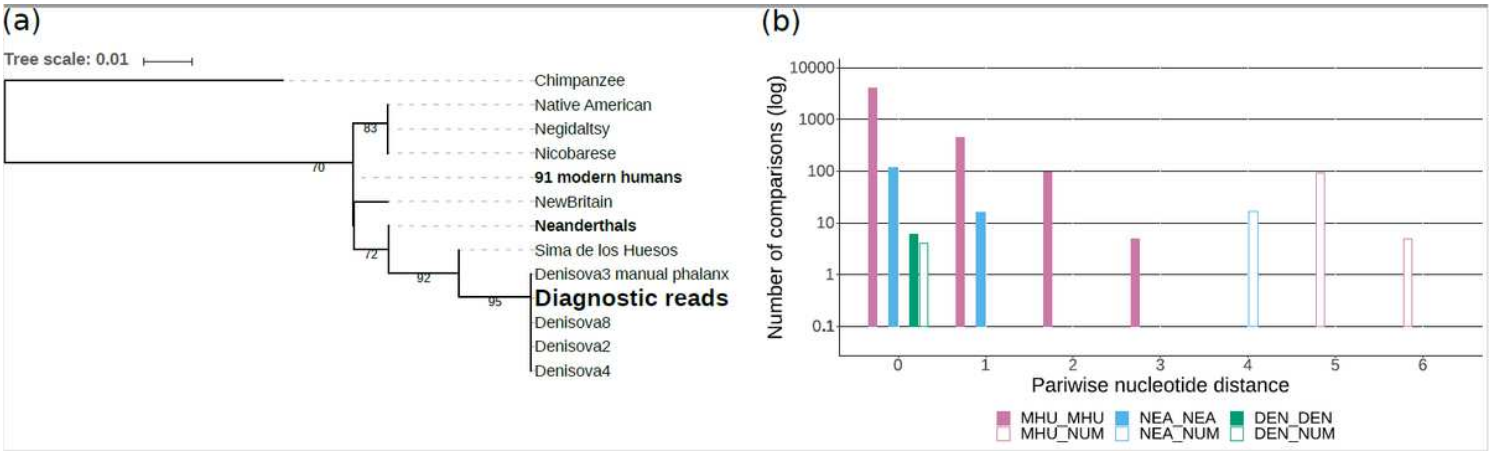

Figure 7

Phylogenetic analysis of diagnostic reads. A potential Denisovan NUMT inferred from diagnostic alleles. (a) Maximum likelihood tree for a 142 bp sequence generated from reads that contain the diagnostic allele for the Denisovan-Sima de los Huesos branch at position 9884. Bootstrap values over 50 are indicated at branch locations. (b) Pairwise nucleotide distances vs. frequency (in logarithmic scale) within and between 97 modern humans (MHU, blue), 17 Neanderthals (NEA, yellow), four Denisovans (DEN, red) and a nuclear insert of mitochondrial DNA (NUMT) (NUM, empty bars).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement.pdf