Journal: www.jogg.info
Originally Published: Volume 5, Number 1 (Spring 2009)
Reference Number: 51.005

THE CONTINUING HUNT FOR NUCLEAR MITOCHONDRIAL DNA SEQUENCES (NUMTS) IN THE HUMAN GENOME

Author(s): Ian Logan

The Continuing Hunt for Nuclear Mitochondrial DNA Sequences (NUMTs) in the Human Genome

Ian Logan

Abstract

Hunting for Nuclear Mitochondrial DNA sequences (NUMTs) has attracted the attention of many researchers in the last few years. In most studies there has been an emphasis on identifying the number of NUMTs in the human genome. But the present study describes a process of matching the parts of a NUMT sequence that are similar to the tRNA coding sequences in mitochondrial DNA. Using this method the author reports the discovery of NUMTs that are common to the genomes of the human, the chimpanzee and the rhesus monkey. These NUMTs were therefore formed before the branching off of the rhesus monkey from the human evolutionary line.

Introduction

The 46 chromosomes in the Human genome contain many hundreds of short sequences of bases that match sections of the DNA found in mitochondria (the mtDNA). These chromosomal sequences are known as *nuclear mitochondrial DNA sequences* or more simply as NUMTs, which can be pronounced as "new-mights."

NUMTs are found in the chromosomes of most species (Richly, 2004), and a wide variety of species have been the subject of articles describing their NUMTs, including the domestic cat (Lopez, 1994; Antunes, 2007), and the ant (Martins, 2007).

A NUMT is formed by the incorporation of a fragment of the mtDNA into a chromosome. This type of event is very rare; but over a period of millions of years the number of times this has happened has becomes appreciable. The formation of a NUMT is essentially a random event and the fragment of mtDNA involved can be of any length, from just a few bases to many thousands of bases, and any of the chromosomes can be involved. In many ways NUMTs are considered to be "fossils" preserving the mtDNA sequence as it used to be at various times in our evolutionary past.

After formation a NUMT becomes an ordinary part of the chromosome and the integrity of its DNA is maintained by the chromosomal repair mechanisms—a process that is not available to mtDNA in the mitochondria. But, whereas the chromosomal repair mechanism will tend to preserve a NUMT, its sequence may still be altered by several processes. The bases of a sequence are subject to a very low mutation rate, a NUMT may become split during the process of "recombination," when parts of chromosomes are exchanged between chromosomes, or by an "intrusion" of another piece of DNA, and also the part of a chromosome containing a NUMT may be duplicated completely, or in part, just once or many times.

As a result of these processes, the sequences of most NUMTs differ considerably from the sequence of modern mtDNA and the identifying NUMTs can be considered to be a bit of a "treasure hunt." This has led to different researchers, unsurprisingly, coming to differing conclusions as to whether a particular part of a chromosome represents a NUMT; and, if so, just where that NUMT begins and ends.

It is possible by comparing the sequence of bases in NUMTs against the sequence of modern mtDNA and counting the number of differences in the sequences to suggest a possible order for the formation of NUMTs. So when a sequence matches well against modern DNA the NUMT can be said to be of "recent origin", say, with a date of formation within the last 10 million years. Whereas, NUMT sequences that match less well, will have a "distant origin" - ranging from 10 million to around 50 million years of age (Benasson, 2003). This method of ageing NUMTs is however self-limiting as it becomes more and more difficult to identify a part of a chromosome as being a NUMT as the sequence of modern mtDNA will have diverged further and further from that of a NUMT.

The identification of NUMT sequences is of importance to the study of genetic genealogy for two reasons. Firstly, it allows for suggestions to be made as to which mutations might have occurred in the human mtDNA before the time of 'Mitochondrial Eve", and secondly, during the sequencing of human mtDNA laboratories need to take care so as not to amplify NUMT sequences and mistake them for mitochondrial DNA.

Address for correspondence: Ian Logan, ianlogan@btinternet.com

Received: November 4, 2008; accepted: February 18, 2009.

The study undertaken for this paper is not primarily concerned with the number of NUMTs and their positions in the human genome, something previously considered in detail by Mourier (2001), Tourmen (2002), Woischnik (2002), Hazkani-Covo (2003), Bensasson (2003), Mishmar (2004), Ricchetti (2004), Hazkani-Covo (2007), and most recently by Lascaro (2008). But instead this study concentrates on what can be learnt from looking at the sequences themselves.

In particular, the study concentrates on the NUMT sequences that contain matching sequences to the coding sequences for the 22 Transfer RNA's found in modern mtDNA. In the mtDNA there is one tRNA sequence for each of 18 amino acids and two tRNA sequences for each of the amino acids, leucine and serine.

Each of the tRNAs can be represented as having a two-dimensional "cloverleaf" structure with stems and loops. **Figure 1** shows the suggested structures for two of the tRNAs. All of the tRNA's have a similar structure, but the sequences are sufficiently different from each other that they are easily distinguished.

Methods

Early studies of NUMTs relied on the actual sequencing of chromosomal sequences (for an example of this method, see Herrnstadt, 1999). But with the publication of the Human Genome, and the genomes of several other species, it is now possible to identify NUMTs using computer search programs.

The genome sequences for the human - *Homo sapiens sapiens*, the chimpanzee - *Pan troglodytes,* and the Rhesus monkey - *Macaca mulatta* are to be found on the web site: http://www.ncbi.nlm.nih.gov/mapview/.

For this study the genome sequences were examined for NUMTs using the Basic Local Alignment and Search Tool or BLAST, and in particular the "BLASTN: Compare Nucleotide Sequences" program (Altschul, 1990).

In most instances the searches were made on the *reference only* sequences as they are the sequences that have been shown to be common to the various assemblies and can be assigned to the different chromosomes.

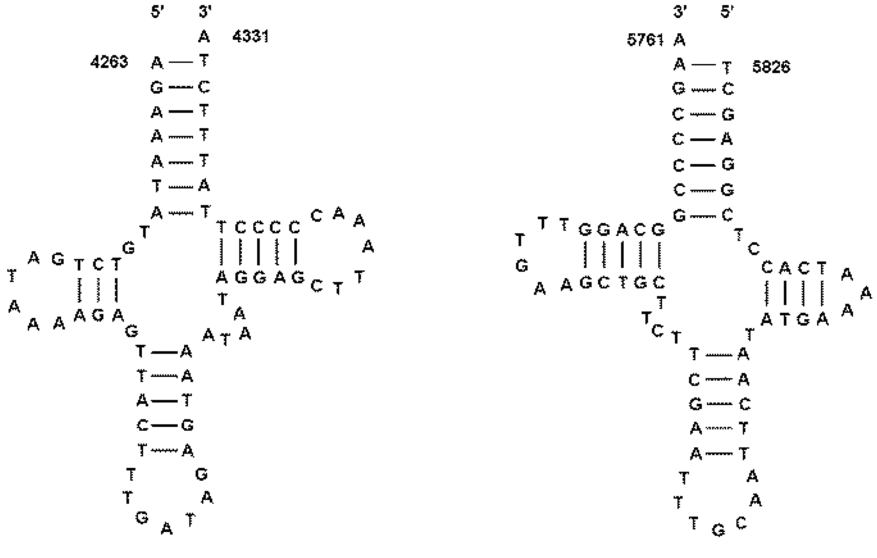
At present *reference only* sequences are available for:

Homo sapiens sapiens – build 36.3 – 368 sequences, covering 2,870,843,926 bases,

Pan troglodytes – build 2.1 – 32,296 sequences, covering 3,010,437,433 bases, and

Macaca mulatta – build 1.1 – 124,049 sequences, covering 3,011,952,279 bases.

The program BLASTN was used to compare nucleotide sequences. Initially the program was used with its default values. However, the default **Expect** value of **0.01** limits the program to reporting only close matches, while using an **Expect** value of **10** can allow chromosomal sequences that match less well to be reported.



(a) Isoleucine

(b) Cysteine

Figure 1 The two-dimensional structures for the t-RNAs isoleucine and cysteine

In the **Advanced options** it is also possible to change the **Word Size** and this makes the matching algorithm less sensitive. The default value is **W11**, but using the parameter at its limit of **W4** can be useful, however this does make the program take a much longer time for each comparison.

Initially, the search string used with BLASTN was the whole sequence of the Cambridge Reference Sequence (CRS) (Anderson, 1981; Andrews, 1999), and this gave a general idea as to how many large and closely matched NUMTs do exist in the human genome. But in practice, it is much better to use only small parts of the mtDNA sequence, and this study concentrates on using as search strings the areas of the mtDNA that code for the 22 Transfer RNA's (tRNA).

Table 1 gives the names of the amino acids, the locations of their corresponding tRNAs in the CRS, and the sequence of bases in the CRS for each of the 22 tRNAs.

Results

The results of the present study are given here in three sections.

NUMTs that match tRNA sequences.
NUMTs of "recent origin"
NUMTs of "distant origin"

NUMTs that Match tRNA Sequences

For each tRNA sequence in the CRS the BLAST search program has been used to find NUMTs that in part match against tRNA sequences.

As an example, **Table 2** shows the results of searching the human genome for NUMTs that match the sequence for the tRNA for the amino acid alanine. The table identifies 32 NUMTs that satisfy the search criteria. The NUMTs vary from having part of their sequence

Table 1
The 22 tRNA Coding Sequences in the CRS

Amino Acid	*Location*	*Sequence of Bases for the Transfer RNAs*
Alanine	5587-5655	TAAGGACT GCAAA ACCCCAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTT
Arginine	10405-10469	TGGTATA TA GTTT AAACA AAAC G AATGA TTTCGAC TCATT AAAT TATGA TAA TCATA TTTACCA A
Asparagine	5657-5729	C TAGACCA ATGGG ACTTAAA CCCAC AAACA CTTAG TTAACAG CTAAG C ACCC TAATCAAC TGGC TT CAATCTA
Aspartic acid	7518-7585	AAGGTAT TA GAAA AACCA TTTC A TAACT TTGTCAA AGTTA AATT ATAGG CTAAAT CCTAT ATATCTT A
Cysteine	5761-5826	A AGCCCCG GCAGG TTTGAA GCTGC TTCT TCGAA TTTGCAA TTCAA T ATGA AAA TCAC CT CGGAGCT
Glutamine	4329-4400	C TAGGACT ATGAG AATCGAA CCCAT CCCT GAGAA TCCAAAA TTCTC C GTGC CACCTATC ACAC CC CATCCTA
Glutamic acid	14674-14742	T ATTCTCG CACGG ACTACAA CCACG ACCA ATGAT ATGAAAA ACCAT C GTTG TATTT CAAC TA CAAGAAC
Glycine	9991-10058	CTCTTT TA GTAT AAATA GTAC C GTTAA CTTCCAA TTAAC TAGT TTTGA CAACAT TCAAA AAAGAGT A
Histidine	12138-12206	GTAAATA TA GTTT AACCA AAAC A TCAGA TTGTGAA TCTGA CAAC AGAGG CTTACGA CCCCT TATTTAC C
Isoleucine	4263-4331	AGAAATA TG TCT GATAAA AGA G TTACT TTGATAG AGTAA ATAAT AGGAG CTTAAAC CCCCT TATTTCT A
Leucine (UUR)	3230-3304	GTTAAGA TG GCAG AGCCCGGTAA TCGC A TAAAA CTTAAAA CTTTA CAGTC AGAGG TTCAATT CCTCT TCTTAAC A
Leucine (CUN)	12266-12336	ACTTTTA AA GGAT AACAGCT ATCC A TTGGT CTTAGGC CCCAA AAAT TTTGG TGCAACT CCAAA TAAAAGT A
Lysine	8295-8364	CACTGTA AA GCTA ACT TAGC A TTAAC CTTTTAA GTTAA AGATT AAGAG AACCAACAC CTCTT TACAGTG A
Methionine	4402-4469	AGTAAGG TC AGCT AAATA AGCT A TCGGG CCCATAC CCCGA AAAT GTTGG TTATAC CCTTC CCGTACT A
Phenylalanine	577-647	GTTTATG TA GCTT ACCTCCTCA AAGC A ATACA CTGAAAA TGTTT AGAC GGGCT CACAT CACCC CATAAAC A
Proline	15956-16023	T CAGAGAA AAAGT CTTTA ACTCC ACCA TTAGC ACCCAAA GCTAA G ATTC TAATTT AAAC TA TTCTCTG
Serine (AGY)	12207-12265	GAGAAAG CTCA CAAGAA CTGCTAA CTCATG CCC CCATG TCTAACAA CATGG CTTTCTC A
Serine (UCN)	7446-7514	C AAAAAAG GAAGG AATCGAA CCCCC CAAA GCTGG TTTCAAG CCAAC CC CATG GCCTC CATG A CTTTTTC
Threonine	15888-15953	GTCCTTG TA GTAT AAACTA ATAC A CCAGT CTTGTAA ACCGG AGAT GAAAA CCT TTTTC CAAGGAC A
Trytophan	5512-5579	AGAAATT TA GGTT AAATACA GACC A AGAGC CTTCAAA GCCCT CAGT AAGTT GCAA TACTT AATTTCT G
Tyrosine	5826-5891	T GGTAAAA AGAGG CCTAA CCCCT GTCT TTAGA TTTACAG TCCAA T GCTT CACT CAGC CA TTTTACC
Valine	1602-1670	CAGAGTG TA GCTT AACACA AAGC A CCCAA CTTACAC TTAGG AGAT TTCAA CTTAAC TTGAC CGCTCTG A

Notes: In this table the amino acids are listed in alphabetic order. Space characters separate the different functional parts of the tRNAs (see Figure 1). The 8 tRNAs for the amino acids, Alanine, Asparagine, Cysteine, Glutamine, Glutamic acid, Proline, Serine (UCN) and Tyrosine appear 'reversed' in the CRS as they are read from the 'light' strand of the mitochondrial DNA.

matching exactly, to having a sequence in which about a fifth of the bases have changed. The table contains only those NUMTs with a sequence that covers the whole of the tRNA sequence. There are other NUMT sequences which match partially, but for the purpose of this paper they have been excluded.

It was found that the BLAST program did not produce the complete set of matches in a single run when the modern mtDNA sequence is used as a search string. However, when these matches were in turn used as search strings it was possible to find further matches. This procedure was then repeated again and again until no more sequences were found.

For the tRNA for alanine there are 2 NUMTs with sequences that do not show any variation from the CRS and these can be considered to be of "recent origin" and are discussed in more detail later. The other NUMTs are considered to be older and therefore in the range 10-50 million years of age.

Table 3 shows a similar pattern of NUMTs was produced for the amino acid arginine. In this instance there

Table 2
NUMTs That Match the tRNA Sequence in CRS for Alanine

Identifier	*Chromosome Location* *Contig Loccation*	*Sequences Found*	*Diff*
CRS Sequence for Alanine		T AAGGACT GCAAA ACCCCAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTT	-
NT_004350.18 |Hs1_4507	chr1:556000..556068 contig: 44769-44837	T AAGGACT GCAAA ACCCCAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTT	0
NT_004487.18 |Hs1_4644	chr1:170946293..170946365 contig:23170096-23170024	T AAGGACT GCATG CAAGACTCTAT CCTAC ATCA ATTGA CTGCAAA TCAAT C ACTT TAATT AAGC TA AGCCCTC	17
NT_004836.17 |Hs1_4993	chr1:236171234..236171302 contig: 2862468-2862400	T AAGGACT GCGAG ACTCTAT TCTGC ATCA ATTGA ATGCAAA TCAAC C ACTT TAATT AAGC TA AGCCCTT	8
NT_032977.8 |Hs1_33153	chr1:94172334..94172401 contig: 64371732-64371665	T AAGGACA -CAAG ACTCTAT CTTAC ATCA GCAGA ATGCAAA TCAAA C ACCT TAATT AAGC TA AATCCTT	16
NT_005403.15 |Hs2_5560	chr2:155828771..155828839 contig: 6330011-6329943	T AAGGACT GCAAA ACCCTAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AACCCTT	2
NT_005403.16 |Hs2_5560	chr2:212350908..212350976 contig: 62852080-62852148	T AAGGGCT GCAAG ACTCTAT TCTGC ATCA GTTGA ACGCAAA TAAAC C ACTT TAATT AAGC TA AGCCCTT	9
NT_022135.15 |Hs2_22291	chr2:117500240..117500308 Contig: 6491692-6491760	T GAGGACT GCAAG ACTCTAT TCTGC ATCA ATTGA ACGCAAA CCAAG C ATTT TAATT GAGC TA AGCCCTT	11
NT_022135.15 |Hs2_22291	chr2:130747629..130747697 contig: 19739081-19739149	T AAGGACT GTAAA ATTCTAC TCTGT ATCA ATTGA ACGCAAA TCAGT C ACTT TAATT AAGC TA AGCCCTT	8
NT_022135.15 |Hs2_22291	chr2:131858358..131858426 contig: 20849878-20849810	T AAGGACT GCAAA ATTCTAC TCTGT ATCA ATTGA ATGCAAA TGAAT C ACTT TAATT AAGC TA AGCCCTT	9
NT_022135.15 |Hs2_22291	chr2:140691531..140691599 config: 29683051-29682983	T AAGGACT GCAAG ACTCTAT TCTGC ATCA ATTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTC	7
NT_016354.18 |Hs4_16510	chr4:156602091..156602159 contig: 80930856-80930788	T AAACACT GCAAG ACTCTAT ACTGC ATCA ATTGA ACGCAAA TCAAC C GCTG TAATT AAGC TA AGCCCTT	11
NT_007758.11 |Hs7_7915	chr7:63208213..63208281 contig: 1603689-1603621	T AAGGACT GAAAA ACTCTAT TCTGT ATCA ATTGA ATGCAAA TCAAT C ACTT TAATT AAGC TA AGCCCTT	9
NT_007758.11 |Hs7_7915	chr7:68436784..68436850 contig: 6832258-6832192	T AGGGATT GCAAG ACT--AT CCTGC ATCG ATTGA ATGCAAA TCAGC C ACTT TAACT AAGC TA GCCCTT	12
NT_007914.14 |Hs7_8071	chr7:141148589..141148657 contig: 2093740-2093672	T AAGGACT GCCAG ACTCTAT TCTGC ATCA GTTGA ATGCAAA TCAAC C ACTT TAACT AAGC TA AACCCTT	11
NT_023629.12 |Hs7_23785	chr7:57259176..57259244 contig: 253757-253825	T AAAGACT GCAAA ACTGTAT TCTGC ATCA ATTGA ATGCAAA TCAAT C ACTT TAATT AAGC TA AGCCATG	10
NT_007995.14 |Hs8_8152	chr8:32994185..32994249 contig: 3195417-3195481	T AAGTACT GCAAG ACTCTAT TCTGC ATCA ATTGA ACGCAAG TGAAC T ACTT TAA-- --GC TA ACCCTTT	16

(Table 2 continued on next page)

Table 2 (continued)

Identifier	*Chromosome Location* *Contig Loccation*	*Sequences Found*	*Diff*
NT_008046.15 |Hs8_8203	chr8:104169887..104169951 contig: 17318884-17318948	T AAGGATT CCAAG ACTCT-- --TAC ATCA ATTGA ATGAAAA AAAAA A ACTT TAATT AAGT GA AATCCGT	22
NT_008046.15 |Hs8_8203	chr8:112016206..112016274 contig: 25165271-25165203	T AAGGACT GCAAG ACTCCAC TCTGC ATCA ATTGA ACGCAAA TCAAC T ACTT TAATT AAGC TA AGCCCTC	6
NT_008046.15 |Hs8_8203	chr8:134836956..134837024 contig: 47985953-47986021	T AAGGAGT GCAAG ACTCTAT TCTGC ATCA ATTGA ACACAAA TCCGC C ACTT TAATT AAGC TA AGCCCTT	8
NT_008413.17 |Hs9_8570	chr9:5086340..5086408 contig 5086340-5086408	C AAGGACT GCAAA ACTCTAT TCTGC ATCA GTTGA ACGCAAA TCAAC C ACTT TAATT AAGC TA AGTCCTT	7
NT_008470.18 |Hs9_8627	chr9:94341562..94341630 contig: 2623014-2622946	C AAGGACT GCAAA ACTCTAC TCTGC ATCA ACTGA ACGCAAA TCAAT C ACTT TAATT AAGC TG AGCCCTT	6
NT_023935.17 |Hs9_24091	chr9:80546308..80546376 contig: 10521088-0521020	T AAGGACT GCAAG ACTCTAT TCTGC ATCA ATTGA ACACAAA TCAAC C ACTT TAATT AAGC TA AGCTCTT	8
NT_023935.17 |Hs9_24091	chr9:82369617..82369685 contig: 12344397-12344329	T AAGGACT GCAAG ACTCTGT TCTGC ATCA ATTGA ACACAAA TCAAC C ACTT TAATT AAGC TA AGCCCTT	8
NT_008583.16 |Hs10_8740	chr10:71022935..71023003 contig: 19904152-19904084	T AAGGACT GCAAG ACTCTGT CCTAC ATCA ATTGT ATGCAAA TCAAT T GCTT TACTT AAGC TA AGCCCTT	15
NT_033899.7 |Hs11_34054	chr11:102782009..102782075 contig: 6839281-6839215	T AAGGACT GCAAG ACT--AT TCTGC ATCA ATTGA ATGGCAAA TCAAT C ACTT TAATT AACC TA AGCCCTT	10
NT_009714.16 |Hs12_9871	chr12:7670420..7670488 contig: 538127-538195	T AAGGACT GTAAA ACTTTAT CCCAC ATTA ATTGA ATGAAAA TTAAA C ACTT TTATT AAGC TA AAACCTC	19
NT_009714.16 |Hs12_9871	chr12:26616819..26616887 contig: 19484594-19484526	T AAGGACT GCAAG ATCTTAT CTTAC ATCA ACTGA ATGCAAA TCAAT C ACTT TAATT GAGC TA ACTCCTT	14
NT_026437.11 |Hs14_26604	chr14:32024003..32024071 contig: 13954071-13954003	T AAGGACT GCAAA ACCCCAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTT	0
NT_010718.15 |Hs17_10875	chr17:19449031..19449098 contig: 19105655-19105722	T AAGGACT GCAAG ACTCTCT TCTGC ATCA -TTGA ACGCAAA TCAAC C ACTT TAATG AAGC TA AGCCCTG	10
NT_024862.13 |Hs17_25018	chr17:21950464..21950532 contig: 343255-343323	C AAGGACT GCAAA ACCCTAC TTTGC ATCT ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTT	4
NT_011512.10 |Hs21_11669	chr21:36185198..36185268 contig: 22925268-22925198	T AAGGACT GCAAC ACTCTCTAT CTTAC ATCA ATTGA ATGCAAA TCAAA C ATTT TAATT AAAC TA AATCCTC	16
NT_011630.14 |HsX_11787	chrX:55222147..55222215 contig: 2759576-2759508	T GAGGACT GCAAG ACTCTAT TCTGC ATCA ATTGA ATGCAAA TCAAC C ACCT TAATT AAGC TA AGCCCTT	9

Notes:
32 NUMT sequences that match the tRNA for Alanine have been identified using the BLAST program.
The Contig. identifiers are those used in Build 36.3 of the Human Genome.
For each NUMT sequence the Contig., chromosome, and position (in the Contig.), and the chromosomal coordinates are given.
Mutational differences from the CRS are shown in Red.
The 'Diff' column give the number of differences from the CRS found in each sequence.

are 27 NUMTs that have been identified, but none is of a "recent origin."

NUMTs of "Recent Origin"

In the human genome there is only one large NUMT of "recent origin" and this was first identified by Herrnstadt (1999). The NUMT was presumably formed after the split with the chimpanzee as it is only to be found in the human genome, and is not in the genomes of either the chimpanzee or the rhesus monkey. The hominid in whom this occurred lived prior to "Mitochondrial Eve," since this NUMT is more divergent from CRS than is any modern human. The NUMT is 5,841 bases in length and matches against the CRS from location 3915 to 9756. **Figure 2** shows that this NUMT matches against about 3/8 of the mtDNA and is located very close to the tip of chromosome 1.

Table 4 shows there are 85 differences between this NUMT and the CRS. The differences result mostly from mutations in the mtDNA along the maternal line leading to modern humans, but a few may have occurred in the NUMT, and a few may have been present in the original mtDNA that was captured in the NUMT. The differences from CRS are shown for the entire NUMT as a conventional mutation list in **Table 4a.** Six of the mutations occurred in tRNA sequences and these are shown in **Table 4b.**

Table 3
NUMTs That Match the tRNA Sequence in CRS for Arginine

Identifier	*Chromosome Location* *Contig Location*	*Sequences Found*	*Diff*
CRS Sequence for Arginine		TGGTATA TA GTTT AAA-CA AAAC G AATGA TTTCGAC TCATT AAAT TATGA TA-A TCATA TTTACCA A	-
NT_004836.17 |Hs1_4993	chr1:233771612..233771678 Contig: 462844-462778	TGGTAAA TA GTTT AAGCCA AAAT A AATGA TTTTTAC TCATT AGAT TATGA TAGA CCATG TTTACCA A	11
NT_005403.16 |Hs2_5560	chr2:203190596..203190662 Contig: 53691768-53691834	TTGTAAA TA GTTT AAGTCA AAAT A AATGA TTTTGAC TCATT AGAT TATGG TAGA CCATA TTTACCA A	11
NT_022135.15 |Hs2_22291	chr2:120687323..120687389 Contig: 9678775-9678841	TGGTAAG TA GTTT AAGTCA AAAT A AGTGA TTTTGAC GCATT AGAT TATGA TAGG CCATA TTTGCCA A	14
NT_022135.15 |Hs2_22291	chr2:130752421..130752486 Contig: 19743873-19743938	TGGTAAG TA GTTT AAGCCA AAAT A AATGA TTTTGAC TCATT AG-T TATGA CAGA ACATG TTTACCA A	15
NT_022135.15 |Hs2_22291	chr2:143572455..143572521 Contig: 32563973-32563907	TGGTAAA TA GTTT AAATTA AAAT G AATGA TTTTGAC TCATT AGAT TATGA TAGA CCATA TTTACCA A	8
NT_005612.15 |Hs3_5769	chr3:108097946..108098012 Contig: 13110468-13110402	TGGTGAA TA GTTT AAGTCA AAAT A AATGA TTTCGAT TCTTT AGAT TATGA TATA TCATA ATTACCA A	11
NT_005612.15 |Hs3_5769	chr3:167361187..167361253 Contig: 72373709-72373643	TGGTAAG TA GTTT AAGCAA AAAT A AATGA TTTCAAC TCATT AGAT TATGA TA-C CCATA CCTACCA A	12
NT_022517.17 |Hs3_22673	chr3:29814321..29814386 Contig: 29779386-29779321	TGGTAAA TA GTTT AAGTCA AAAT A AATGA TTTCAAC TCATT AGAT TATGA TAGA CCGTG TTTACCA A	12
NT_006316.15 |Hs4_6473	chr4:25330751..25330817 Contig: 16397077-16397011	TGGTAAA TA GTTT AAGTCA AAAT A AATGA TTTTGAC TCATT AGAT TATGA TAGA CCACA TTTACCA A	11
NT_016354.18 |Hs4_16510	chr4:156597300..156597366 Contig: 80926063-80925997	TGGTAAA TA GTTT AAGTCA AAAT A AATGA TTTTGAC TCATT AGAT TATGA TAGA CCACA TTACCA A	10
NT_022778.15 |Hs4_22934	chr4:65159223..65159288 Contig: 5679962-5679897	TGGTAGG TA GTTT AAA.CAA AAAT A AATGA TTTTGAC TCATT AGAT TATGG TA-G TCATA CTTACCA A	10
NT_034772.5 | Hs5_34934	chr5:99414256..99414320 Contig: 1801434-1801370	TGGTACA TA GTTT AAA-TA AAAC G AATGA TTTCGAC TGATT AAAT TATGA TA-G TCATA TTTACCA A	4
NT_034772.5 | Hs5_34934	chr5:134291916..134291980 Contig: 36679094-36679030	TGGTATA TA GTTC AAA.CA AAAC G AATGA TTTCGAC TCATT AAAT TATGA TA-A TCATA TTTACCA A	1
NT_025741.14 |Hs6_25897	chr6:154031350..154031416 Contig: 58094086-58094152	TGATAAT TA GTTT AAGTCA AAAT A AATGA TTTCGAC TCATT AAAT TATGA TAGA TTATA ATTACCA A	10
NT_007758.11 |Hs7_7915	chr7:63203399..63203465 Contig: 1598873-1598807	TGGTAGA TA GTTT AAGCCA AAAT A AATGA TTCTGAC TCATT AGAT TATAA TAGA ACATA TTTACCA A	11
NT_023629.12 |Hs7_23785	chr7:57240725..57240791 Contig: 235306-235372	TGGTAGA TA GTTT AAGCCA AAAT A AATGA TTTTGAC TCATT AGAT TATGA TACA ACATA TTTACCA A	9
NT_023629.12 |Hs7_23785	chr7:57263987..57264053 Contig: 258568-258634	TGGTAGA TA GTTT AAGTCA AAAT A AATGA TTTTGAC TCATT AGAT TATGA TAGA ACATA TTAACCA A	9
NT_008046.15 |Hs8_8203	chr8:134597167..134597232 Contig: 47746164-47746229	TGGTACT CA GTTA AA-CCA AAAC A AATGA TTTCAAC TCAGT AGAT TGTGA TAAA TCATA ATTACCA A	12
NT_030737.9 | Hs8_30993	chr8:18102995..18103061 Contig: 5903709-5903643	TGGTAAT GA GTTT AAACCA AAAC A AATGA TTTTGAC TCATT AAAT TATGA TTAC TCATA ATTACTA A	11
NT_008413.17 |Hs9_8570	chr9:5097570..5097636 Contig: 5097570-5097636	TGGTAAA TA GTTT AAGTCA AAAG A AATGA TTTCGAT TCATT AGAT TATAA TAAA CCATA TTTACCA A	10
NT_008470.18 |Hs9_8627	chr9:93912616..93912677 Contig: 2194000-2194061	TGGTAAA TA GTTT AAA-TT AAAT G A---- TTTCGAC TCATT AGAT TATGA TAGA CCATA TTTACCA A	11
NT_009237.17 |Hs11_9394	chr11:38564496..38564564 Contig: 37395228-37395162	TGGTAAT TA GTTT AAATCA AAAT A AATGA A TTTTGAC TCATT AGAT TATGG CAGA GCATA AC TACACCA A	13
NT_033927.7 | Hs11_34082	chr11:80940849..80940915 Contig: 11485950-11486016	TGGTAGT TT AAGT CAAAAT AAAT A AATGA TTTTGAC TCATT AGAT TATGA TAGA CCATA TTTACCA A	16
NT_010194.16 |Hs15_10351	chr15:56230457..56230523 Contig: 29233722-29233788	TGGTAGA TA GTTT AAGTTA AAAT A AATGA TTTCGAC TCATT AGAT TATGA TAGA CTATA TTTACCA A	10
NT_010393.15 |Hs16_10550	chr16:10724714..10724780 Contig: 2130358-2130292	TGATAAA TA GTTT AAGTTA AAAT A AATGA TTTTGAC TCATT AGAT TATGA TAGA CCATA TTTACTA A	12
NT_010393.15 |Hs16_10550	chr16:13993309..13993375 Contig: 5398887-5398953	TGGTAAT TA GTTT TAATCA AAAT A AATGA TTTCGAC TCATT AGAT TATGG CAAA CTGAT CAAACTC T	20
NT_024862.13 |Hs17_25018	chr17:21955261..21955328 Contig: 348052-348119	TGGTAAG TA GTTT AACAGACA AAAA C AATGA TTTTGAC TTGTT AGAT TATGA TA-G TCATA CTTACAA A	13

On chromosome 14 there is a second, but much smaller, NUMT of "recent origin." This NUMT is 1,021 bases in length and matches against the CRS from 5583-6606. **Table 5a** shows the 71 mutational differences between this NUMT and the CRS. The mutations that have occurred in the tRNAs are shown in **Table 5b.**

The recent paper by Hazkani-Covo and Covo (2008) gives a list of NUMTs of "recent origin" - most of which are very short in length and do not match against a complete tRNA sequence. But for reasons that are not totally clear, the two NUMTs discussed above are not on the list.

NUMTs of "Distant Origin"

The sequence of bases in a NUMT of "recent origin" matches the CRS very well; but as described above there are very few NUMTs of that type. The majority of NUMTs are much older - possibly in the range of 10 - 50 million years of age.

Tables 2 & 3 show the details of NUMTs with sequences that match against the tRNAs of alanine and arginine; and it is possible to prepare a detailed analysis for any individual NUMTs. However, there are some NUMTs of particular interest as it has been possible to show that there are NUMTs that can be found in the genome of *Homo sapiens* AND ALSO in the genomes of the Chimpanzee, *Pan troglodytes*, and the Rhesus Monkey, *Macaca mulatta*. This fact suggests that these NUMTs were incorporated into the genome of an ancestor common to all three species.

The best example of this type of NUMT that is common to the Human, Chimpanzee and Rhesus Monkey has been found on Chromosome 21. This NUMT of length 1851 bases corresponds to the part of the mtDNA containing the tRNAs for tryptophan, alanine, asparagine, cysteine and tyrosine. In the Chimpanzee, *Pan troglodytes*, the whole of the NUMT is also found on Chromosome 21. However in the Rhesus Monkey, *Macaca Mulatta* where there is no Chromosome 21, it is found on Chromosome 3.

The sequence from the genome of Homo sapiens shows a considerable number of differences from the CRS. Nevertheless, the three NUMT sequences from the genomes of Homo sapiens, *Pan troglodytes* and *Macaca mulatta* are almost identical to each other suggesting that they had a common formation.

The details of this NUMT are shown in **Table 6.**

Whereas the NUMT on chromosome 21 has been found to be the largest NUMT that is common to the Human, Chimpanzee and Rhesus Monkey, there are several others smaller NUMTs of this type.

Table 7 gives the details of a further 5 NUMTs that are found on the Human chromosomes 3, 4, 8, and X.

Discussion

This paper has concentrated on identifying NUMTs in the human genome by using the BLAST program to find matches against tRNA sequences in modern mtDNA. This technique has led to the identification of several

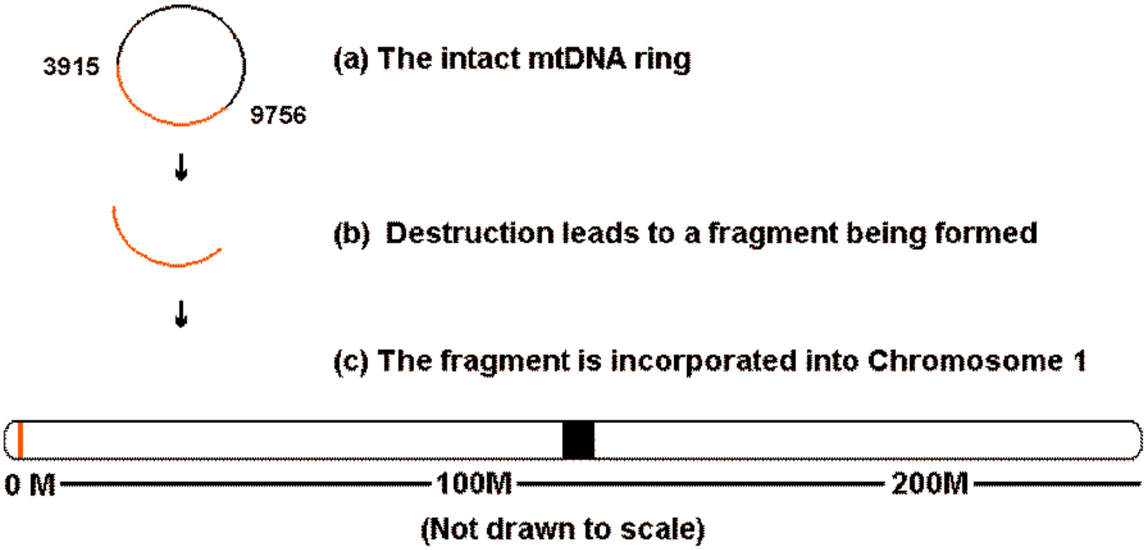

3915
9756
(a) The intact mtDNA ring
(b) Destruction leads to a fragment being formed
(c) The fragment is incorporated into Chromosome 1
0 M
100M
200M
(Not drawn to scale)

Figure 2. Formation of the "Herrnstadt" NUMT. Initially, the mtDNA was only found in mitochondria, but the partial destruction of a mtDNA ring led to the passage of a fragment into the nucleus where it became incorporated into chromosome 1.

NUMTs which are common to the genomes of the Human, the Chimpanzee and the Rhesus Monkey. But developing these ideas has only been possible by considering the published findings in various papers that have appeared over the last few years. Actual quotations from the papers are shown in *italics*.

The early researchers used a laboratory system which involved using bacterial clones, specially prepared primers and direct sequencing. This method was very laborious, but nevertheless, was quite successful.

For example, Nomiyama (1985) used this system to identify 2 NUMTs, subsequently shown to be located on chromosome 3 (GenBank numbers X2226, M12298); and even then it was clear that NUMTs were old as the author suggested these 2 NUMTS *"were transferred from mitochondria into nuclei about 12 and 15 millions of years ago, respectively."*

Later Herrnstadt (1999) used a similar method to identify a NUMT on chromosome 1 (GenBank number AF134583). This NUMT was shown to have a length 5,841 nucleotide bases. The authors were able to link the NUMT to "*a very distal portion of Chromosome 1" and in their discussion they recognise that their NUMT was of a very recent origin and said " it is estimated that this sequence was transferred to the nucleus during evolution long after the divergence of humans from other nonhuman primates.*" Although only the single NUMT was identified in the study, the paper did suggest the possibility of there being other "*hitherto unidentified numtDNA sequences.*"

By 2001, the method of identifying NUMTs by searching Human DNA databases had begun to replace laboratory methods; and Mourier (2001) published "*the first extensive analysis of NUMTs in the human nuclear genome.*" This study found "*296 numts ranging between 106 and 14,654 bp in size.*" The paper is also important as it discusses the possibility of NUMTs being formed at different stages of mammalian evolution. However, whilst this paper is very useful, the identification of the NUMTs was based on early Human Genome

Table 4a
The Mutation List for the "Herrnstadt" NUMT

```
Location: Chromosome 1 (coords chr1: 554327..560167)
Contig: NT_004350.18    43096..48936 Corresponds to: mtDNA (rCRS) 3915-9756
```

G4048A	A4104G	C4312T	C4318T	C4456T	T4736C	A4769G	T4856C	C4904T	C4914T	C4940T	A4958G
G4991A	T5041C	G5147A	C5320T	A5351G	C5387T	T5426C	G5471A	A5474G	A5498G	T5580C	G5821A
C5840T	G6023A	T6221C	C6242T	A6266C	A6299G	G6366A	G6383A	C6410T	C6452T	C6483T	T6512C
C6542T	C6569A	T6641C	A6692G	6698.1A	C6935T	C6938T	A7146G	C7232T	C7256T	G7316A	G7521A
C7650T	T7705C	C7810T	C7868T	C7891T	G7912A	A8021G	G8065A	C8140T	G8152A	T8167C	C8197-
A8198-	C8203T	G8392A	C8455T	C8461T	T8503C	G8545A	C8655T	A8677C	A8701G	A8718G	A8860G
C8943T	C9060A	C9075T	C9168T	A9254G	T9325C	G9329C	A9434G	C9527T	T9530C	T9540C	A9545G
G9548A	A9629G										

Notes: The "Herrnstadt" NUMT has 85 mutational differences as compared to the CRS.
Six of these mutations occur in the coding areas for tRNAs as shown in Table 4b.
The above table differs from Table 1 in Herrnstadt (1999) as the CRS was revised late in 1999 (Andrews, 1999).

Table 4(b)
An Analysis of the Complete tRNAs Found in the "Herrnstadt" NUMT

Isoleucine	4263-4331	AGAAATA TG TCT GATAAA AGA G TTACT TTGATAG AGTAA ATAAT AGGAG TTTAAAT CCCCT TATTTCT A
Glutamine	4329-4400	C TAGGACT ATGAG AATCGAA CCCAT CCCT GAGAA TCCAAAA TTCTC C GTGC CACCTATC ACAC CC CATCCTA
Methionine	4402-4469	AGTAAGG TC AGCT AAATA AGCT A TCGGG CCCATAC CCCGA AAAT GTTGG TTATAT CCTTC CCGTACT A
Trytophan	5512-5579	AGAAATT TA GGTT AAATACA GACC A AGAGC CTTCAAA GCCCT CAGT AAGTT GCAA TACTT AATTTCT G
Alanine	5587-5655	T AAGGACT GCAAA ACCCCAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTT
Asparagine	5657-5729	C TAGACCA ATGGG ACTTAAA CCCAC AAACA CTTAG TTAACAG CTAAG C ACCC TAATCAAC TGGC TT CAATCTA
Cysteine	5761-5826	A AGCCCCG GCAGG TTTGAA GCTGC TTCT TCGAA TTTGCAA TTCAA T ATGA AAA TCAC CT CAGAGCT
Tyrosine	5826-5891	T GGTAAAA AGAGG CTTAA CCCCT GTCT TTAGA TTTACAG TCCAA T GCTT CACT CAGC CA TTTTACC
Serine(UCN)	7446-7514	C AAAAAAG GAAGG AATCGAA CCCCC CAAA GCTGG TTTCAAG CCAAC CC CATG GCCTC CATG A CTTTTTC
Aspartic Acid	7518-7585	AAGATAT TA GAAA AACCA TTTC A TAACT TTGTCAA AGTTA AATT ATAGG CTAAAT CCTAT ATATCTT A
Lysine	8295-8364	CACTGTA AA GCTA ACT TAGC A TTAAC CTTTTAA GTTAA AGATT AAGAG AACCAACAC CTCTT TACAGTG A

Notes: The 6 mutational differences between the "Herrnstadt" NUMT and the CRS are shown in Red.

Table 5(a)
The Mutation list for a NUMT on Chromosome 14

Location: Chromosome 14 (coords chr14:32023055..32024075)
Contig: NT_026437.11 13954075-13953055 Corresponds to: mtDNA (rCRS) 5583-6606

C5662T	G5667T	C5681T	G5703A	T5717-	G5718-	C5743A	G5746A	G5769C	G5821A	C5840T	A5843G
C5893T	C5895T	C5899d	C5922T	A5924G	A5957G	T5964C	A5990G	C6005T	C6015T	A6017G	C6020T
G6026A	G6032A	C6068T	T6071C	A6113G	C6119T	T6160A	G6182A	T6185C	T6216C	T6221C	C6224T
C6236T	C6242T	T6251C	G6260A	A6266C	A6269C	A6281G	C6326T	C6335T	A6353G	C6356T	T6365C
G6366A	T6378C	G6383A	C6389T	T6392C	C6398T	T6407C	C6410T	G6446A	C6452T	C6483T	T6497C
T6524C	A6527G	A6530G	C6531T	G6541A	C6542T	C6569A	G6573A	A6575G	A6581G	C6587T	

Notes: The NUMT on chromosome 14 has 85 mutational differences as compared to the CRS.
Ten of these mutations occur in the coding areas for tRNAs as shown in Table 5b.

Table 5(b)
An Analysis of the Complete tRNAs Found in the NUMT on Chromosome 14

Alanine	5587-5655	T AAGGACT GCAAA ACCCCAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA
Asparagine	5657-5729	C TAGATCA ATTGG ACTTAAA CCCAT AAACA CTTAG TTAACAG CTAAA C ACCC TAATCAAC --GC TT
Cysteine	5761-5826	A AGCCCCG CCAGG TTTGAA GCTGC TTCT TCGAA TTTGCAA TTCAA T ATGA AAA TCAC CT
Tyrosine	5826-5891	T GGTAAAA AGAGG CTTAG CCCCT GTCT TTAGA TTTACAG TCCAA T GCTT CACT CAGC CA

Notes: The 10 mutational differences between the NUMT on chromosome 14 and the CRS are shown in Red.

Project data and it is now quite difficult to correlate the results with the latest analyses.

In 2002 a paper from France (Tourmen, 2002) suggested there were 286 NUMTs and stressed "*Some pseudogenes [NUMTs] appeared highly modified, containing inversions, deletions, duplications, and displaced sequences.*"

Later, a paper from the USA (Woischnik, 2002) identified 612 NUMTs and showed that NUMTs can be found on every chromosome.

In 2003, a paper from Israel (Hazkani-Covo, 2003) discussed the features of 82 large NUMTs; and in particular the workers concluded "*only about a third of all the numt repertoire in the human nuclear genome is due to insertions … the rest originated as duplications of preexisting numts.*"

In a paper from the USA (Benasson, 2003), 348 NUMTs with a length greater than 500 bp are discussed. The paper suggests an age of 25-40 million years for the majority of the NUMTs, and considers that *"numts arose continuously over the last 58 million years."*

Mishmar (2004) was able to identify 247 NUMTs and discusses how it is possible by looking for selected mutations to determine if one NUMT is more ancient than another. The author suggests *"nuclear mtDNA pseudogenes are genetic fossils that reflect our past."*

Later, Richetti (2004) was able to identify 211 NUMTs. The paper is also interesting as the author made the suggestion that *"NUMT integrations preferentially target coding or regulatory sequences."*

The paper of Schmitz, et al. (2005) is rather different to the earlier papers as it discusses *"the evolutionary pathway of a pseudogene which separated from the corresponding mitochondrial gene more than 40 mya [million years ago]."* Their study concentrated on the larger of the 'Nomiyama' NUMTs (GenBank number X02226). The authors suggest that *"numt sequences provide a much more reliable base for dating"* [than] *"molecular dating based on primate mtDNA."*

More recently, Hazkani-Covo (2007), produced a survey of NUMTs common to both human and the chimpanzee. But, the researchers did not report any NUMTs found also in the rhesus monkey.

Lascaro (2008) has produced a compilation of the 90 longest NUMTs found in the human genome. But in the present author's opinion the actual figures given for the start and finishing points for the NUMTs are still inaccurate. In particular, the data from Lascaro has not taken note of the parts of NUMT sequences that match to tRNA sequences and this has resulted in many of the NUMTs being reported as having lengths which are much less than they really are. Nevertheless, Lascaro's compilation is far more accurate than earlier attempts.

Table 6
A NUMT of “Distant Origin” on Human Chromosome 21

(a)Sequence Taken from Locations 5512-5891 of the CRS, Covering Five tRNAs

Tryptophan	5512-5579	AGAAATT TA GGTT AAATACA GACC A AGAGC CTTCAAA GCCCT CAGT AAGTT GCAA TACTT AATTTCT G
(noncoding)	5580-5586	TAACAGC
Alanine	5587-5655	T AAGGACT GCAAA ACCCCAC TCTGC ATCA ACTGA ACGCAAA TCAGC C ACTT TAATT AAGC TA AGCCCTT
(noncoding)	5656	A
Asparagine	5657-5729	C TAGACCA ATCGG ACTTAAA CCCAC AAACA CTTAG TTAACAG CTAAG C ACCC TAATCAAC TGGC TT CAATCTA
(noncoding)	5730-5760	CTTCTCCCGCCGCCGGGAAAAAAGGCGGGAG
Cysteine	5761-5826	A AGCCCCG GCAGG TTTGAA GCTGC TTCT TCGAA TTTGCAA TTCAA T ATGA AAA TCAC CT CGGAGCT
Tyrosine	5827-5891	GGTAAAA AGAGG CCTAA CCCCT GTCT TTAGA TTTACAG TCCAA T GCTT CACT CAGC CA TTTTACC

(b) Corresponding Sequence from a Modern Human (Homo sapiens sapiens) NUMT

Tryptophan	AGGAATT TA GGTT AGG--CA GACC A AAAGC CTTCAAA GCCCT AAGC AATAT TTTA TATTT TATTCCT G
(noncoding)	AAAAAT-
Alanine	T AAGGACT GCAAC ACTCTCTAT CTTAC ATCA ATTGA ATGCAAA TCAAA C ATTT TAATT AAAC TA AATCCTC
(noncoding)	A
Asparagine	T TAGATTG GTAGT ATCCAAC CTCAA GAAAA TTTCA TTAACAG TGAAA T ACCC TAATCACC TGTC TT CAGTCTA
(noncoding)	CTTCTGCTGTTGAGAG-AAAA-GGGCAGGGG
Cysteine	A AGCCCTG GCAGA ATTGAA GCTGC ATCT TTGAG TTTGCAA TTTGA T GTGA CTAT TCAC CT TGAGGCA
(noncoding)	C
Tyrosine	AGTAAAA AGAGG GTTCAA CCTCT GTCT TTAGA -TTACGG ATTAA G GCTT C-CT CAGC CA TTTCACT

Location: Chromosome 21 (coords for this part chr21:36184963..36185340, coords for the whole NUMT chr21:36184442..36186292), Contig: NT_011512.10. The 98 differences from CRS are shown in Red.

(c) Corresponding Sequence from a Chimpanzee *(Pan troglodytes)* NUMT

Tryptophan	AGGAATT TA GGTT AGG--CA GACC A AAAGC CTTCAAA GCCCT AAGC AATAT TTTA TATTT TATTCCT G
(noncoding)	AAAAAT
Alanine	T AAGGACT GCAAC ACTCTCTAT CTTAC ATCA ACTGA ATGCAAA TCAAA C ATTT TAATT AAAC TA AATCCTC
(noncoding)	A
Asparagine	T TAGACTG GTAGT ATCCAAC CTCAA GAAAA TTTCA TTAACAG TGAAA T ACCC TAATCACC TGTC TT CAGTCTA
(noncoding)	CTTCTGCTGTTGAGAG-AAAA-GGGCAGGGG
Cysteine	A AGCCCTG GCAGA ATTGAA GCTGC ATCT TTGAG TTTGCAA TTTAA T GTGA CTAT TCAC CT TGAGGCA
(noncoding)	C
Tyrosine	AGTAAAA AGAGG GTTCAA CCTCT GTCT TTAGA -TTACGG ATTAA G GCTT C-CT CAGC CA TTTCACT

Location: Chromosome 21 (coords for this part chr21:21997766..21998144), Contig: NT_106996.1. The three differences from the human NUMT are shown in Blue.

(d) Corresponding Sequence from a Rhesus Monkey *(Macaca mulatta)* NUMT

Tryptophan	AGGAATT TA GGTT AGG--CA GACC A AAAGC CTTCAAA ACCCT AAGC AATAT TTTA TATTT TGTTCCT G
(noncoding)	AAAAAT
Alanine	T AAGGACT GCAAC ----TCTAT CTTAC ATCA ATTGA ATGCAAA TCAAA C ATTT TAATT AAAC TA AATTCTC
(noncoding)	A
Asparagine	T TAGATTG GTAGT ATCCAAC CTCAA GAAAA TTTCA TTAACAG TGAAA T ACTC TAATCACC TGTC TT CAGTCTA
(noncoding)	CTTCTGCTGTTGAGAG-AAAA-GGGCAGGGG
Cysteine	A AGCCCTG GCAGA ATTGAA GCTGC ATCT TTGAG TTTGCAA TTTGA T GTGA CTAT TCAC CT TGAGGCT
(noncoding)	C
Tyrosine	AGTAAAA AGAGG GTTCAA CCTCT GTCT TTAGA TTTACGG ATGAA G GCTT T-CT CAGG CA TTTCACT

Location: Chromosome 3 (coords for this part chr3:2549029-2549403), Contig: NT_001114167.1. The 14 differences from the human NUMT are shown in Blue.

Table 7
Examples of other NUMTs of *'Distant Origin'* Common to the Human, Chimpanzee and Rhesus Monkey

a Arginine	
Homo sapiens Chr8	TGGTACT CA GTTA AACCA AAAC A AATGA TTTCAAC TCAGT AGAT TGTGA TAAA TCATA ATTACCA A
Pan troglodytes Chr8	TGGTACT TA GTTA AACCA AAAC A AATGA TTTCAAC TCAGT AGAC TGTGA TAAA TCATA ATTACCA A
Macaca mulatta Chr8	TGGTAAT TA GTTA AACCA AAAC A AATGA TTTCAAC CCATT AGAT TATGA TAAA TCATA ATTACCA A
Notes: Human NT_008046.158 Chr8:47746164..47746229, Pan troglodytes NW_001240369.1 Chr8:2660739..2660804, Macaca mulatta NW_001122918.1 Chr8:624397..624462	
b Cysteine	
Homo sapiens ChrX	A AGCCCCA GCAGG ACTGAA GCTTC TCCT TTGAA TTTGCAA TTCAA C ATGA GAAA TCCC CT CAGGGCT
Pan troglodytes ChrX	A AGCCCCA GCAGG ACTGAA GCTTC TCCT TTGAA TTTGCAA TTCAA C ATGA GAAA TCCC CT CAGGGCT
Macaca mulatta Chr14	A AGCCCCG GCAGG ATTGAA GCTGC TCCT TTGAA TTTGCAA CTCAA C ATGA GAAA TCAC CT CAGGGCT
Notes: Human NT_011630.14 ChrX:2759403..2759337,Pan troglodytes NW_001251894.1 ChrX:243536..243470, Macaca Mulatta NW_001100391.1 Chr14:4167244..4167178	
c Lycine	
Homo sapiens Chr4	CACTGTA AA ACTA TC TAGC A TTAAA CTTTTAA GTTAA AGACT GAGCG GATCTACAC CTCTC TGCAGTG A
Pan troglodytes Chr4	CACTGTA AA GCTA TC TAGC A TTAAA CTTTTAA GTTAA AGACT GAGCG GATCTACAC CTCTC TGCAGTG A
Macaca mulatta Chr5	CACTGTA AA GCTA TC TAGC A TTAAT CTTTTAA GTTAA AGACT GAGGG GATCTACAC TTCTC TGCAGTG A
Notes: Human NT_016354.18 Chr4:80928165..80928097, Pan troglodytes NW_001234090.1 Chr4:1346097..1346029, Macaca mulatta NW_001118162.1 Chr5:6371225..6371157	
d Valine	
Homo sapiens Chr3	CAAAATG TA GCTT AACCCA AAGC A TCCGG CTTACAC CCAGA AGAT TTCAT CATGAC CTAAT CACTTTG A
Pan troglodytes Chr3	CAAAATG TA GCTT AACCCA AAGC A TCCGG CTTACAC CCGGA AGAT TTCAT CATGAT CTAAT CACTTTG A
Macaca mulatta Chr2	CAAAATG TA GCTT AACCCA AAAC A TTCGG CTTATAC CCAGA AGAT TTCAT CATGAC CTGAT CACTTTG A
Human NT_022517.17 Chr3:40233828..40233896, Pan troglodytes NW_001232821.1 Chr3:684423..686034, Macaca mulatta NW_001112552.1 Chr2:8033441..8031857	
e Phenylalanine	
Homo sapiens ChrX	CTAAGTG TG GCTC GGGGCCT GCAC A AGGCA TTGAAAA TGCCT AGAT GAGTT CATGT AACTC CATAAAC A
Pan troglodytes ChrX	CTATGTG TG GCTC GGGGCCT GCGC A AGGCA CTGAAAA TGCCT AGAT GAGTT CATGT AACTC CATAAAC A
Macaca mulatta ChrX	CTATGTG TG ACTT GGGGCCT TCAC A AGGCA CTGAAAA TGCCT AGAT GAGTT CATGT AACTC CATAAAC A
Human NT_011757.15 Chr X:28757556..28757488, Pan troglodytes NW_001251729.1 ChrX: 853495..853427, Macaca mulatta NW_001218104.1 ChrX:1909090..1909022	

Notes: In each human sequence the differences between the NUMT sequence and the modern human (CRS) mtDNA are shown in Red. The differences in the Pan and Macaca sequences from the human NUMT are shown in Blue.
Examples (a), (b) and (c) have 11-12 differences in the human NUMT from the CRS per sequence, whereas examples (d) and (e), with 17 mutations and 25 mutations respectively, are presumably very much older. Example (d) is part of the 'Nomiyama' NUMT, with GenBank accession number X02226 (Nomiyama, 1985).

Finally, Covo (2008) discusses just how NUMTs might be formed by the inclusion of mtDNA material following breaks in chromosomal DNA.

Conclusions

The present study reports the result of carefully matching the respective parts of NUMT sequences against the coding area of tRNA sequences in modern mtDNA.

This has shown that there are a few NUMTs of "recent origin"—that is of NUMTs formed since the branching off of the human evolutionary line from the rhesus monkey and the chimpanzee.

But more importantly the study has shown that there is a small number of NUMTs that are common to the genomes of the human, chimpanzee and the rhesus monkey. These NUMTs have a date of formation which predates the branching of these primates from the human evolutionary line.

The study also shows that there is not as yet a consensus view as to which parts of the human genome are NUMTs, and thereby have an origin in the mitochondrial DNA.

However, the search for NUMTs continues and the results presented in this paper are based on an analysis of the genomes that are currently available. There is a lot more yet to be discovered about NUMTs in the human genome.

References

Anderson S., Bankier AT, Barrell BG, de Bruijn MHL, Coulson AC, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature*, 290:457-465.

Andrews RM, Hubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 23:147.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol*, 215:403-410.

Antunes A, Pontius J, Ramos MJ, O'Brien SJ, Johnson WE (2007) Mitochondrial introgressions into the nuclear genome of the domestic cat. *J Hered*, 98:414-420.

Bensasson D, Feldman MW, Petrov DA (2003) Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J Mol Evol*, 57:343-354.

Hazkani-Covo E, Sorek R, Graur D (2003) Evolutionary dynamics of large *Numts* in the human genome: Rarity of independent insertions and abundance of post-insertion duplications. *J Mol Evol*, 56:169-174.

Hazkani-Covo E, Graur D (2007) A comparative analysis of *numt* evolution in human and chimpanzee. *Mol Biol Evol*, 24:13-18.

Hazkani-Covo E, Covo S (2008) Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet*, Oct;4(10).

Herrnstadt C, Clevenger W, Ghosh SS, Anderson C, Fahy E, Miller S, Howell N, Davis RE (1999) A novel mitochondrial DNA-like sequence in the human nuclear genome. *Genomics*, 60:67-77.

Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C, Attimonelli M (2008) The RHNumtS compilation: features and bioinformatics approaches to locate and quantify human NumtS. *BMC Genomics*, 9:267.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ(1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol*, 39:174-190. Erratum in: *J Mol Evol*, 39:544.

Martins J Jr, Solomon SE, Mikheyev AS, Mueller UG, Ortiz A, Bacci M Jr (2007) Nuclear mitochondrial-like sequences in ants: evidence from Atta cephalotes (Formicidae: Attini). *Insect Mol Biol*, 16:777-784.

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

Mourier T, Hansen AJ, Willerslev E, Arctander P (2001) The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Mol Biol Evol*, 18:1833-1837.

Nomiyama H, Fukuda M, Wakasugi S, Tsuzuki T, Shimada K (1985) Molecular structures of mitochondrial-DNA-like sequences in human nuclear DNA. *Nucleic Acids Res*, 13:1649-1658.

Ricchetti M, Tekaia F, Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol*, 2:E273.

Richly E, Leister D (2004) NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol*, 21:1081-1084.

Schmitz J, Piskurek O, Zischler H (2005) Forty million years of independent evolution: a mitochondrial gene and its corresponding nuclear pseudogene in primates. *J Mol Evol*, 61:1-11.

Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P (2002) Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics*, 80:71-77.

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