2008-2010
Master's Thesis

Immigrants to the Nucleus; Analysis of Mitochondrially Derived Nuclear Genomic Regions (NUMT)

Submitted: March, 2010

47-086914 Junko Tsuji

Advisor: Professor Paul Horton

Department of Computational Biology, Graduate School of Frontier Science
The University of Tokyo

Abstract

From the origin of eukaryotic cells to the present, mitochondrial DNA (mtDNA) fragments continuously have transferred to the nucleus and become nuclear mitochondrial-like DNAs (NUMT). Previous studies concluded that NUMT candidate mtDNA sequences and nuclear NUMT insertion sites are randomly chosen. However, from analyzing 310 human NUMT, we found specific mtDNA transferred pattern and NUMT preference features of nuclear genome. Our result suggests that the mitochondrial promoter region and its peripheral domains (548bp-1142bp counted by D-loop as the origin) were seldom transferred. In NUMT preferentially integrated sites of human genome, AT-rich oligomers appeared in all NUMTs flank and 90% of NUMTs contained retrotransposons in their flanks (more than expected by change, P-value = 0.001). The retrotransposon-encoded endonuclease recognizes AT-rich oligomers (5′-TTTTAA-3′). This suggests retrotransposon-encoded endonuclease may be involved in NUMT insertion. We also crosschecked our NUMT dataset against annotation databases. Almost all NUMTs were non-coding (nc) regions or introns. Only 1~2% of NUMTs have annotated functional roles in human and mouse genome. They were mainly ncRNAs, and those functional human and mouse NUMTs were inserted relatively recently in evolutionary time. Interestingly, one of the human specific functional NUMTs is an ncRNA which is expressed during fetal brain development. Hence, it is conceivable that this element might contribute to the difference between human and chimpanzee brain structure.

Keywords

Nuclear genome, mitochondrial DNA, NUMT, mitochondrial pseudogene, retrotransposon, exogenous DNA element

Preface

Poster presentation arising from this thesis

This research was presented as a poster presentation at the 32nd Annual Meeting of Molecular Biology Society of Japan (MBSJ2009) [a], and then subsequently expanded as a poster presentation at the 20th International Conference on Genome Informatics (GIW2009) [b]

[a] Junko Tsuji and Paul Horton, **Analysis of Mitochondrially Derived Nuclear Regions in Human Genome**, *The 32nd Annual Meeting of Molecular Biology Society of Japan*, Yokohama, Japan (2009)
[b] Junko Tsuji and Paul Horton, **Immigrants to the Nucleus; Analysis of Mitochondrial-like Fragments in Human Genome**, *The 20th International Conference on Genome Informatics*, Yokohama, Japan (2009)

Acknowledgements

As expected, this study was not completed without many supports. I greatly appreciate my supervisor Prof. Paul Horton. The provided environment by my supervisor in CBRC, AIST was really the excellent place. The spent time in the Sequence Analysis Team was not only for this master course, but it also changed my view, narrower to broader. Because of my background: experimental biology, despite the littleness of the knowledge about the computational biology, Prof. Horton and the team members gave me a lot of informative lectures. I appreciate the members of the Sequence Analysis Team; Martin Frith, Edward Wijaya, Kenichiro Imai, Raymond Wan and Sachiyo Abratani, and Hajime Harada who used to be a member. Martin told another/new ideas concerned the study, Edward showed me the way of mapping tags to genomes, Kenichiro provided the knowledge about mitochondria, and Raymond and Sachiyo support programming knowledge. Furthermore, the members of Horton lab.: Fu Szu-chin and Yoshinori Fukasawa helped my day-to-day student life, and I spent great time with all people in 7th floor of CBRC. And last, I am grateful to my parents for their support and caring.

Contents

Chapter 1

Introduction

Mitochondria are cellular organelles which were originally diverged from α-proteobacteria [1]. They contain their own genomes aside from the nuclear genome. In mammals, each double-stranded circular mtDNA consists of approximately 16,000 base pairs. The two strands of mitochondrial DNA (mtDNA) are discriminated by their nucleotide content with the guanine rich strand called the heavy chain, and the cytosine rich strand called the light chain.

In the endosymbiotic evolutionary process, many mtDNA fragments were transferred to the nucleus. This early phase of mtDNA transfer was thought to provide massive relocation of mitochondrial genes to nuclear chromosomes [2]. This conception comes from that the extant mitochondrial proteome is now overwhelmingly encoded by the nuclear genome. However, in the present day, the transfer of genetic element is extremely rare or has ceased in most eukaryotes [3]. Despite this fact, mtDNA are still continuously transferred to the nucleus, producing mtDNA-like nuclear domains called "NUMTs"; NUclear MiTochondrial-like DNAs [4]. NUMTs have been identified over 70 eukaryotes, and their numbers widely differ in each species [5]; some species retain several hundred NUMTs, but others have no detectable NUMTs at all. It was hypothesized that this variation of NUMTs population arises from the difference of mtDNA copy number and its length in each species [6]. However, this hypothesis does not explain some cases, so the reasons behind NUMTs population difference still remain unclear.

Recent studies indicate that the creation of NUMTs is mediated by non-homologous end joining repair [7-8]; mtDNA fragments are inserted and joined with nuclear break ends when nuclear double strand breaks (DSB) occur (Figure 1.1). In previous NUMTs studies, the patterns mtDNA contributing to NUMTs and features of NUMTs insertion site were investigated. Most studies indicated transferred mtDNA and

nuclear NUMTs insertion site were randomly chosen [9-11]. Moreover, in the functional NUMTs survey, it was suggested that most NUMTs exists in introns or intergenic regions [12]. Because of the difference of genetic codes between mitochondria and the nucleus, NUMTs tend to be "dead-on-arrival" pseudogenes. However, several studies provided the evidence of functional NUMTs in particular species (e.g. plants, yeasts, and flies) [13]. In humans, only one functional human NUMT has been proposed in the literature: Christos *et al* [14]. Discovered a potential functional NUMT in the 3'UTR of the nuclear receptor coactivator2 mRNA (NCOA2). Additionally, the role of somatic cell NUMTs in disease has been demonstrated [15].

In this study, we mainly focused on human NUMTs. Additionally, also rhesus, mouse and rat NUMTs were investigated for testing and confirming our observations. From analyzing our original datasets, we discovered a specific pattern of mtDNA migration and characteristics of nuclear NUMT integration sites. An understanding of mtDNA-transferred and nuclear NUMT-insertion features would help to understand genomic evolution and diversification among living organisms, and also contribute to identify the positions prone to produce DSBs (perhaps also in somatic cells) which are deleterious for organisms. Furthermore, by crosschecking our NUMT dataset against annotation databases, we observed new functional NUMTs candidates in human genome; three non-coding transcripts, and one secreted protein, R-spondin homolog 1 (RSPO1). Applying phylogenetic analysis, we found that these functional NUMTs were inserted; after humans and gorillas diverged and three of them after humans and chimpanzees diverged. Interestingly, the human specific functional NUMTs were non-coding RNAs which are expressed during fetal brain development. Hence, it is conceivable that this element might contribute to the difference between human and chimpanzee brain structure.

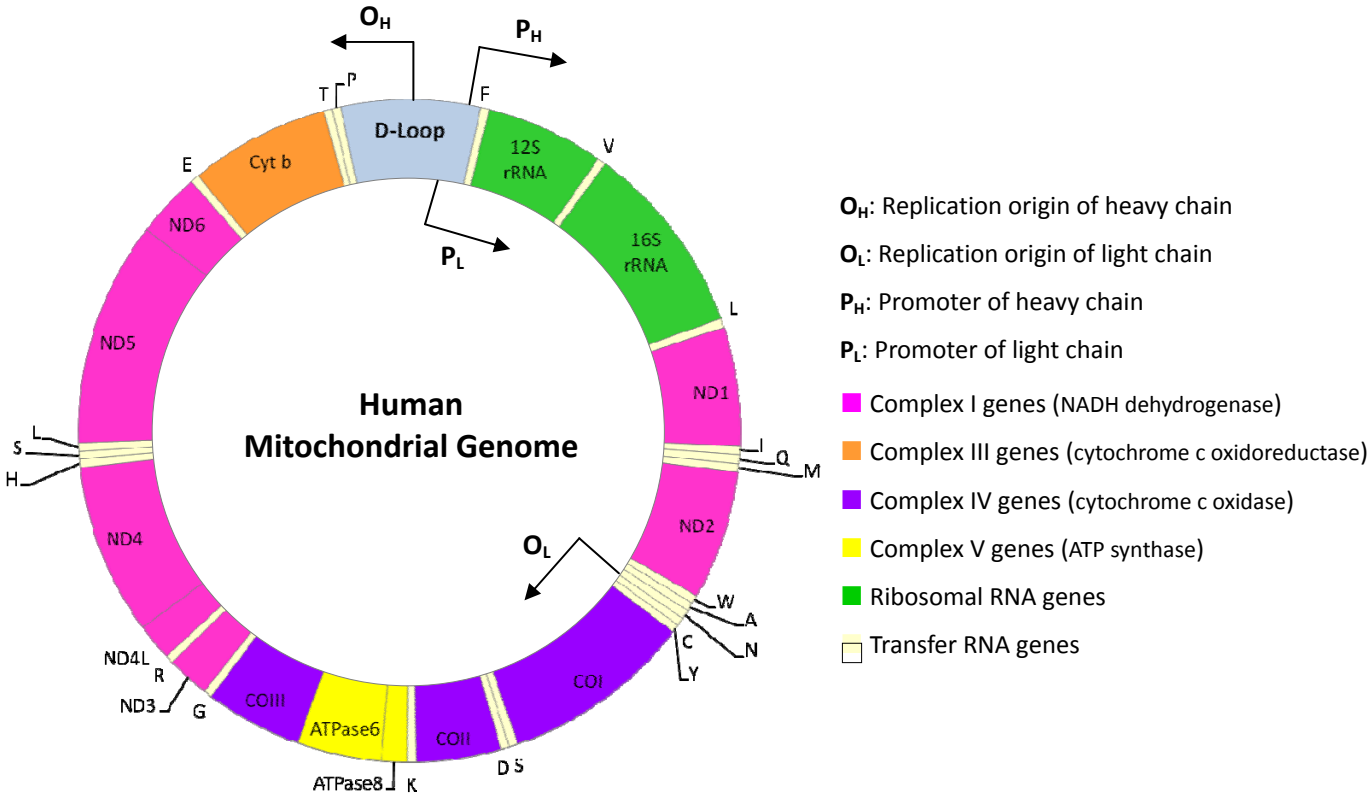

Figure 1.1: The diagram of human mitochondrial genome

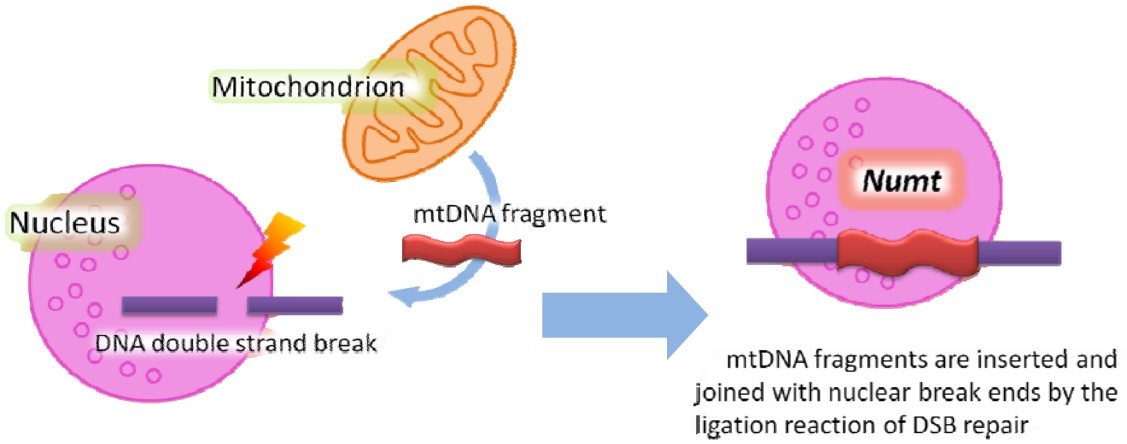

Figure 1.2: The rough sketch of NUMT integration mechanism.

Chapter 2

Results and Discussion

2.1 *Characteristics of NUMTs*

2.1.1 Number of NUMTs and their insertion age

In this investigation, we mainly focused on human and mouse. First, we analyzed the total number of NUMTs, and then we classified the inserted NUMTs into specific evolutionary ages with maximum-likelihood estimation.

2.1.1.1 Human NUMTs

From homology search between the human nuclear and mitochondrial genome, we obtained 821 nuclear segments which share homology with mtDNA. After that, we merged overlap nuclear hits and concatenated linear segments which probably come from the same mtDNA insertion event. Then, the number of NUMTs was reduced to 391. Next, we filtered duplicated/unduplicated NUMTs by crosschecking the information of nuclear duplicated regions, and 81 duplicated NUMTs were removed. Finally, we identified 310 unique (unduplicated) NUMTs (Table S1 in Supplementary Information), and those inserted ages were estimated by maximum-likelihood method (see also: Chapter 4 Methods). We investigated six periods defined by common ancestor divergence of a species with human; (a) before rhesus, (b) after rhesus and before gibbon, (c) after gibbon and before orangutan, (d) after orangutan and before gorilla, (e) after gorilla and before

chimpanzee, and (f) after chimpanzee. The result is in Figure 2.1. From the phylogenetic analysis of each NUMT and primate mtDNA, 138 (approximately 45%) NUMTs were integrated in age (a). In other words, most of human NUMTs insertion events occurred in the Old World monkey and the New World monkey diverged age. Other NUMTs were accumulated at a rate of 25~40 per age (b)-(f). Most detectable NUMTs predate the divergence of human and squirrel monkey but NUMT insertion events continued throughout primate evolution.

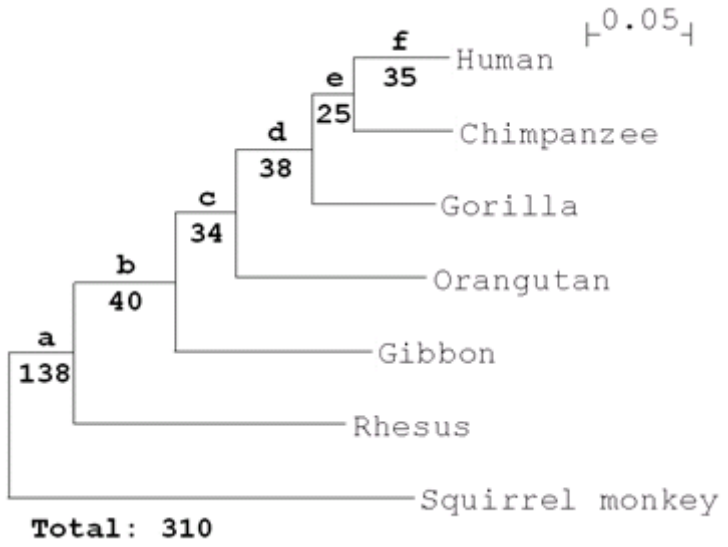

Figure 2.1: Human NUMTs on the mitochondrial phylogenic tree of seven primates.

Next, we calculated the NUMT-accumulation ratio of all six ages. The ratio was derived from a simple division of each branch distance and each number of NUMTs. The detail of each branch length and the calculated ratio are in Table 2.1. With the value of ratio in each age, the transition of NUMT-accumulation ratio in each age was plotted (Figure 2.2). In Figure 2.2, we did not further consider the NUMTs integration ratio of the oldest age (a), because the ratio contains NUMTs which were inserted in the age immediately after squirrel monkey and the human lineage diverged, but also far older NUMTs. In the other ages (b)-(f), accumulation ratios were nearly constant.

Table 2.1: Branch distance and NUMTs accumulation ratio in human.

Age	Distance	Number of NUMTs	Ratio
a	0.03123	138	4419
b	0.0488	40	819.7
c	0.02961	34	1148
d	0.03675	38	1034
e	0.01964	25	1273
f	0.04608	35	759.5

Distance is the span until a new species diverges from an ancestor. NUMT accumulation ratio was calculated from a distance in each age divided by the number of NUMTs.

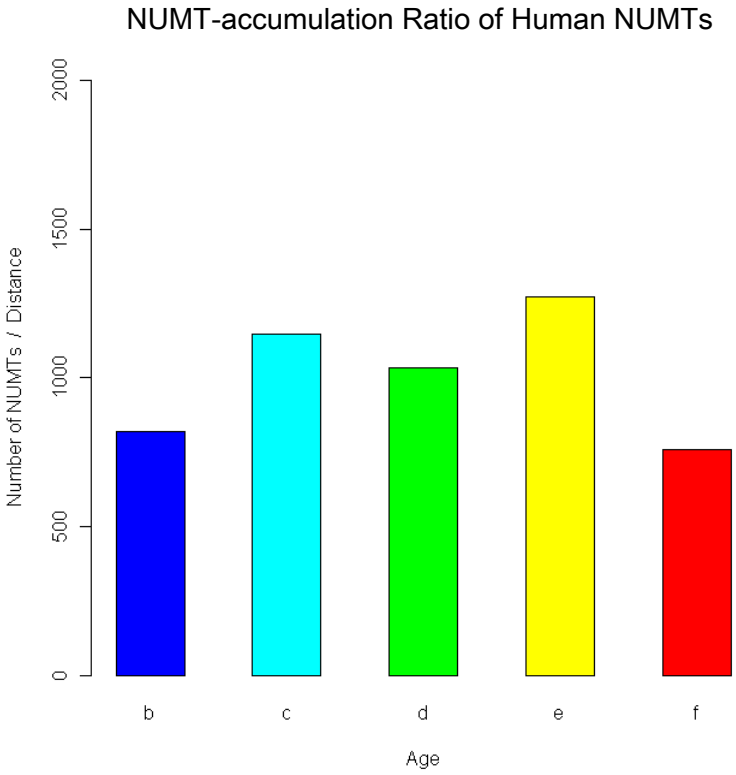

Figure 2.2: Transition of NUMT-accumulation ratio through primate evolution.

2.1.1.2 Mouse NUMTs

We applied same method to the mouse nuclear genome and its mitochondrial genome. 126 initial hits were detected, and same insertion events were merged. The number of mouse NUMTs were reduced to 93. Interestingly, in mouse genome, the number of nuclear duplicated NUMTs was only one fragment. Finally we identified 92 NUMTs (Table S2 in Supplementary Information), and calculated their NUMT-insertion ages. In our mouse NUMTs phylogenetic analysis, we investigated five periods defined by common ancestor divergence of a species with mouse; (g) before hedgehog, (h) after hedgehog and before rabbit, (i) after rabbit and before squirrel, (j) after squirrel and before guinea pig, and (k) after guinea pig. From the results in Figure 2.3, mouse NUMTs insertion events also continuously have occurred from primeval time to present, like human NUMTs accumulation. In addition, the number of NUMTs in each age was roughly allocated in even shares.

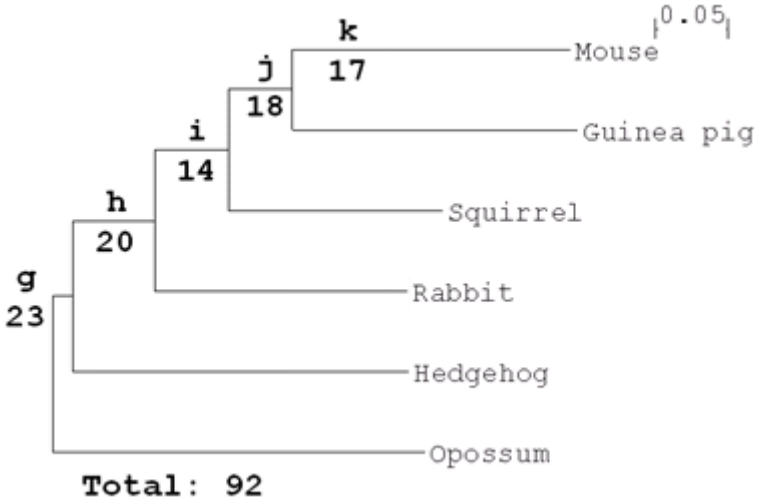

Figure 2.3: Mouse NUMTs on the mitochondrial phylogenic tree of five mammals.

In the same way, we calculated NUMT-accumulation ratio of five ages (Table 2.2), and we plotted the transition of NUMT-accumulation ratio in each age (Figure 2.4). The accumulation ratios in the other ages (h)-(k) were nearly constant.

Table 2.2: Branch distance and NUMTs accumulation ratio in mouse.

Age	Distance	Number of NUMTs	Ratio
g	0.01338	23	1719
h	0.05650	20	354.0
i	0.05014	14	279.2
j	0.04361	18	412.7
k	0.19035	17	89.31

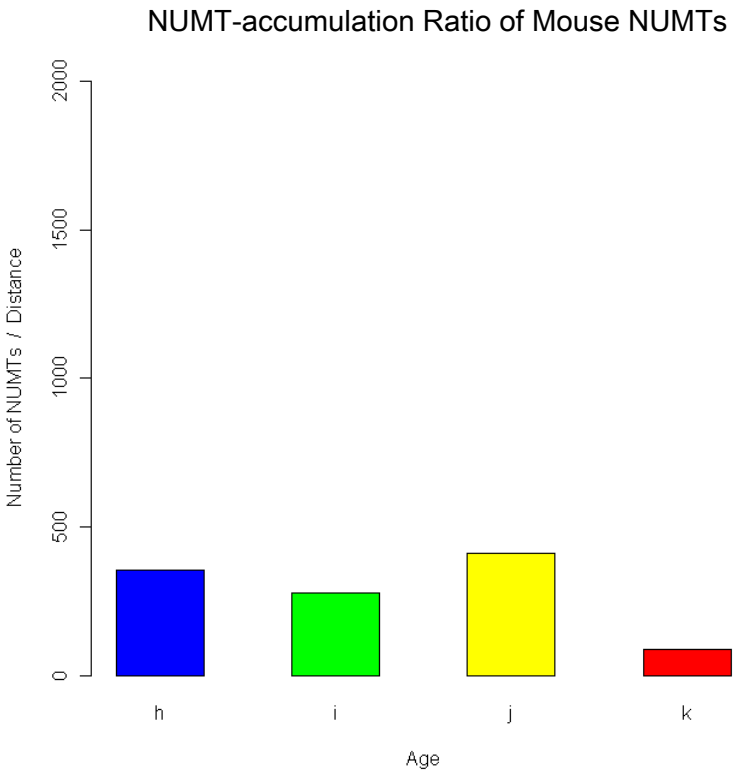

Figure 2.4: Transition of NUMT-accumulation ratio through mammalian evolution.

2.1.2 The distribution of NUMT-insertion length in each age

First, we investigated the human NUMT length in each age (Figure 2.5 A). In human NUMTs, we found the oldest age (a) NUMTs tend to be longer than NUMTs in other ages. This indicates various length of mtDNA inserted to nuclear genome in old age. Moreover, interestingly, human NUMTs length became shorter as the age becomes younger. Next, mouse NUMTs-insertion length was also investigated (Figure 2.5 B). A change in average the insertion length was observed, however the degree of change was quite smaller than the one of human NUMTs insertion. Although the change of the NUMT length was smaller, like human NUMTs, the insertion length tended to be shorter in young ages. In addition, the median of human and mouse NUMT length in each age was shown in Table 2.3.

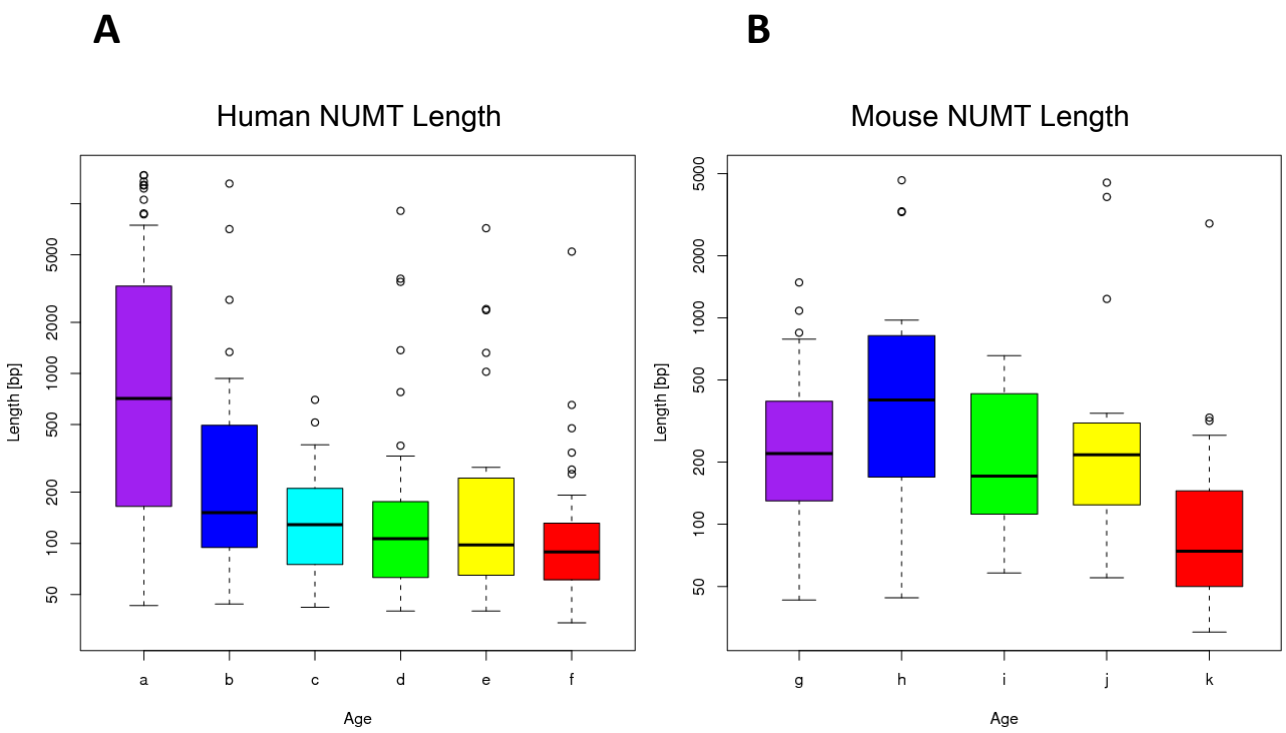

Figure 2.5: The transition of NUMT length through ages.

A: Human NUMT length. Age (a) is the oldest (purple) and age (f) is the youngest (red).
B: Mouse NUMT length. Age (g) is the oldest (purple) and age (k) is the youngest (red).
Ages become younger in alphabetical order.

Table 2.3: Median of NUMT length in each age

Human		Mouse	
Age	**Median of NUMT length [bp]**	**Age**	**Median of NUMT length [bp]**
a	714.5	**g**	225
b	151	**h**	396.5
c	129	**i**	171
d	106.5	**j**	216
e	98	**k**	74
f	89		
All	174.5	**All**	184

2.1.3 Sequence evolution of NUMTs and their mtDNA counterparts

The GC content of Human and mouse NUMTs and their mtDNA source fragments are plotted in Figure 2.6. The change in of human GC content between NUMTs and mtDNA counterparts was bigger than the change in mouse GC content. The common feature of the change in both species was transition from GC to AT. In other words, GC content of NUMTs tends to be lower than the GC content of mtDNA counterparts. Moreover, interestingly the alteration level of GC content between NUMTs and the nuclear genome in human and mouse shows a difference. In mouse, NUMT GC content was lower than nuclear GC content; however human NUMT GC content was higher than nuclear GC content.

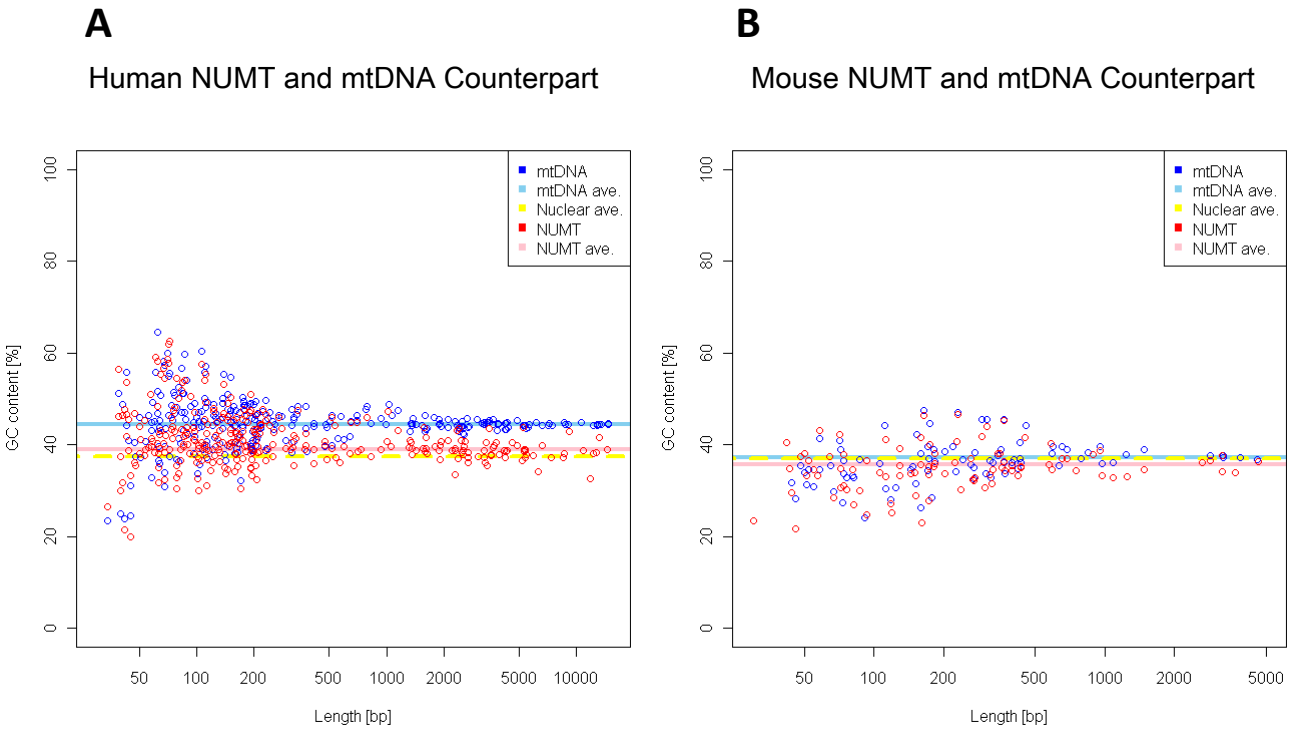

Figure 2.6: GC content alteration between NUMTs and their mtDNA counterparts.

Red dots indicate the GC content of NUMTs. Blue dots indicate the GC content of mtDNA counterparts. The pink and blue lines represent the average GC content of all NUMTs and average GC content of all mtDNA counterparts. Yellow line corresponds to the average GC content of the nuclear genome.

2.1.4 Discussion

Here we summarize the above results. From section 2.1.1, the total population of mouse NUMTs was about a third of the human NUMTs. In the comparison of NUMT-accumulation speed, we found mouse NUMT-accumulation ratios in all ages was smaller than human's one, although the each distance of both species between ages was almost the same length. This observation implies the human genome tends to accumulate mtDNA fragments more actively than the mouse genome. Moreover, from the result of section 2.1.2, the human NUMTs were longest in the oldest age, the longest is almost the whole length of mtDNA (14,720bp). However, the change of the mouse NUMT length was relatively small; the longest mouse NUMTs is 4,649bp which is just one third of the human longest NUMTs length. In section 2.1.3, the human NUMT GC content divergence from original mtDNA was larger than the mouse NUMT divergence (even though in general the evolution rate among mammals is highly conserved). This may simply be due to the larger difference in GC content between nuclear and mitochondrial genomes in human versus mouse (Figure 2.6.), but it is also tempting to speculate that some human NUMTs could be evolving under positive selection.

Although the results between human and mouse were different overall, they shared common features: the accumulation rate was approximately constant over the evolutionary time span investigated and in the oldest age, longer mtDNA fragments were engulfed in the nuclear genome.

However, from Figure 2.5, we must consider the possibility that longer NUMTs population in older age is a result of the difficulty of detecting older, short NUMTs, because older NUMTs contain many mutations: indels, transitions, transversions, etc. So, more easily detectable, longer NUMTs might be overrepresented in our analysis. To test this problem, we randomly extracted fixed length segments of the oldest human NUMT in age (a) from 1000bp to 50bp at intervals of 50bp, and then investigated the detectable length of these segments by BLASTing them against their corresponding mtDNA. For each fixed length, we performed two tests, one with 1000 trials and one with 10000 trials. The result is shown in Figure 2.7.

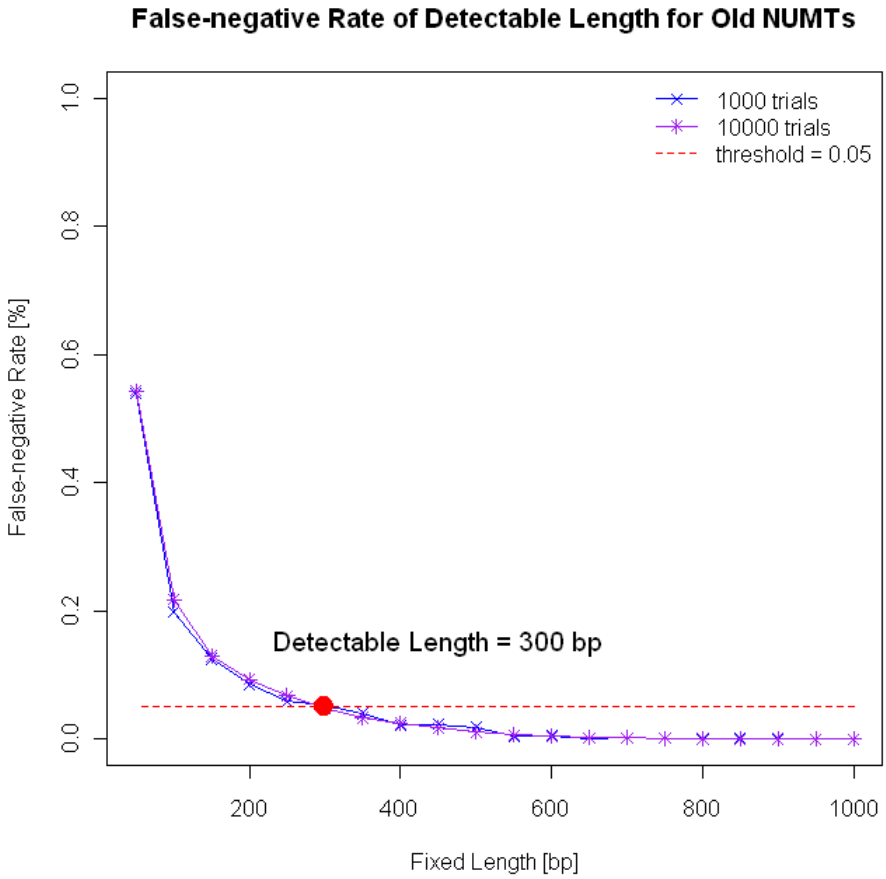

Figure 2.7: False-negative rate vs. NUMT segment length for age (a)

More that 95% of NUMT segments of lengths 1000bp to 350bp were detectable, however this percentage dropped quickly for segments shorter than 300bp. Therefore, it would appear that the distribution of NUMT length in older age (Figure 2.5) misses some short NUMTs, however we can still conclude that younger NUMTs tend to be short.

Overall, considering above outcomes through analyses, we might say that NUMTs affect genomic diversity and complexity in the evolution process.

2.2 *The pattern of the NUMT-candidate mtDNA*

2.2.1 The result of the pattern of transferred mtDNA fragments

By crosschecking NUMT positions and their source of mtDNA positions, we identified the displacement of human NUMT-source mtDNA. Until now, "randomly chosen" NUMT-source mtDNA was thought to migrate to the nucleus [8-10]. However, our result suggests that the mitochondrial promoter region and its peripheral domains (548bp-1142bp from D-loop start point) in the D-loop were seldom transferred (Figure 2.8A). For confirming the significance of our result, NUMT-candidate mtDNA of rhesus monkey was also analyzed. As a result, rhesus NUMT-source mtDNA also shows the same pattern as human. We also applied same method to investigate the pattern of the migration of mouse mtDNA fragments (Figure 2.8B). Because of the few population and small length of mouse NUMT samples, we could not accurately estimate the frequency of mtDNA displacement. However, the tendency of immovability of the promoter region is also observed in mouse mtDNA migration. **A**s the similar example of rodents, we analyzed rat NUMT-candidate mtDNA. Rat NUMT-source mtDNA exhibits the similar pattern as mouse.

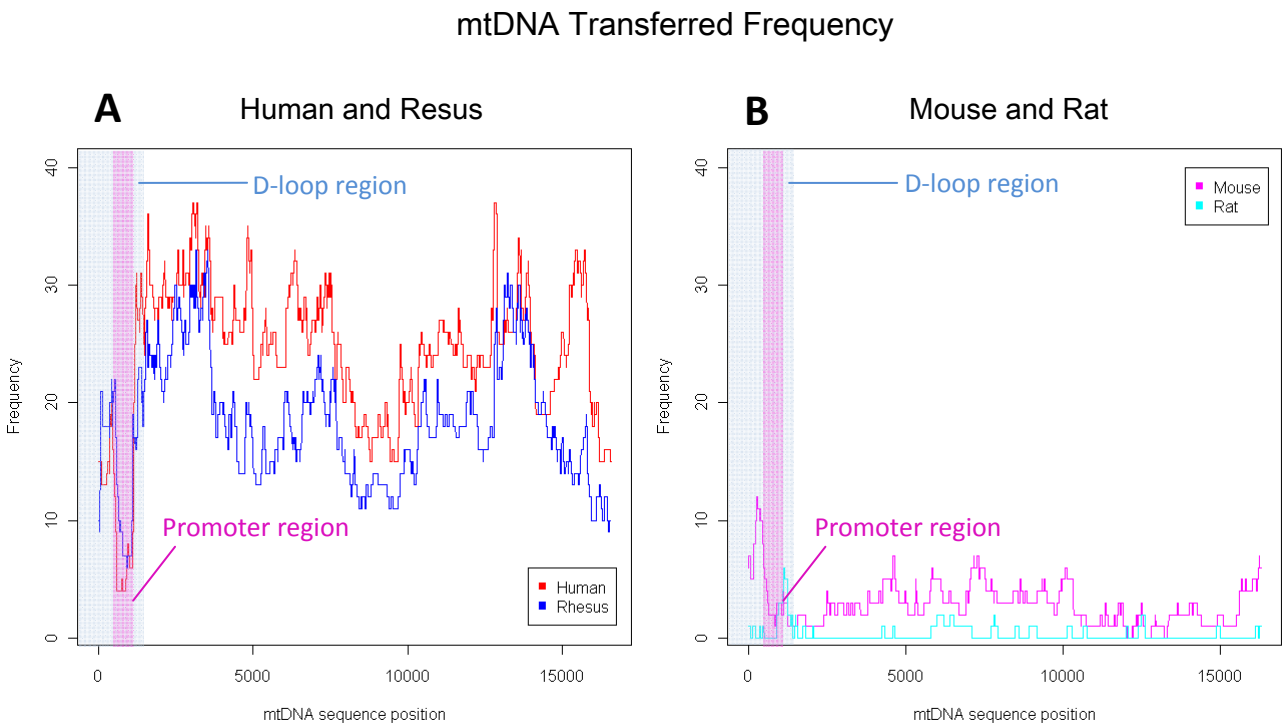

Figure 2.8: The migration pattern of NUMT-candidate mtDNA.
Pink highlighted region: promoter region. Blue highlighted region: D-loop region.

Next, we investigate age-specific mtDNA migration pattern in human (Figure 2.9). In age (e), seemingly the domain of 1bp-2500bp in mtDNA was often transferred. However, overall, transferred NUMT-source mtDNA fragments distributed uniformly through all ages except on promoter region, and there are no obvious age-specific patterns.

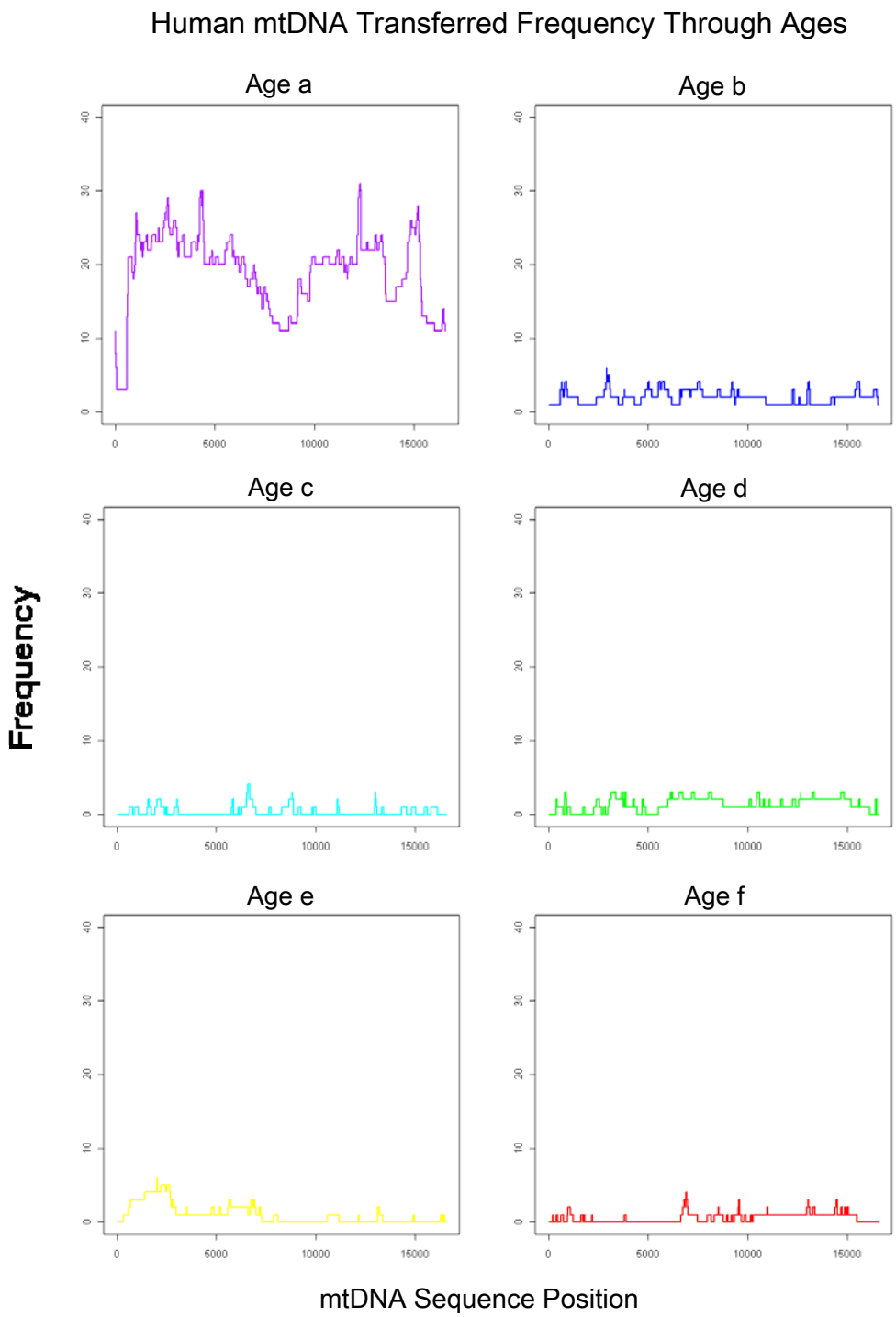

Figure 2.9: Human mtDNA transferred frequency in each age.

2.2.2 Discussion

2.2.2.1 Hypothesis about the region seldom found in NUMTs

As seen above, we shed light on the pattern of transferred NUMT-source mtDNA fragments. Then, why was the promoter region seldom transferred? Because of the several control elements in the mitochondrial D-loop (Table 2.3, and see also Figure 1.2 in Introduction), this region is the center of various proteins binding; the mtDNA polymerase binds to the replication origin, the mitochondrial transcription factor attaches to the promoter region , and the anchor protein bundles mtDNA to the mitochondrial inner membrane [16,17].

Table 2.3: Control elements in D-loop

Name of Elements	Start	End	Ref.
D-loop region	**1**	**1122**	**[18]**
Light chain control element	473	480	[19]
Heavy chain origin	653	986	[20]
Transcription factor binding site	776	803	[21]
	819	847	
	963	990	
	1068	1095	
Heavy chain control element	916	924	[19]
	929	936	
Light chain promoter	937	990	[22]
Heavy chain promoter	1090	1112	[23]
Membrane attachment site*	1	1044	[18]
	(16469	16571)	

* mtDNA is circular, so the 1-1,044bp region and parenthetic 16,469-16,571bp region are continuous.

Furthermore, the recent study mentioned that D-loops (especially promoter regions) of multiple mtDNAs in a single mitochondrion are tightly bundled by some types of proteins (Figure 2.10A) [24-26], and mitochondrial nucleoids are formed. Mitochondrial

transcription factor A (TFAM), the protein which binds the promoter region, is also the component of the nucleoids [27-29]. From those observations from previous studies, we hypothesize that this observation is due to proteins which bind the mitochondrial promoter and its adjacent region, interrupting mtDNA fragmentation and immigration to the nucleus (Figure 2.10B).

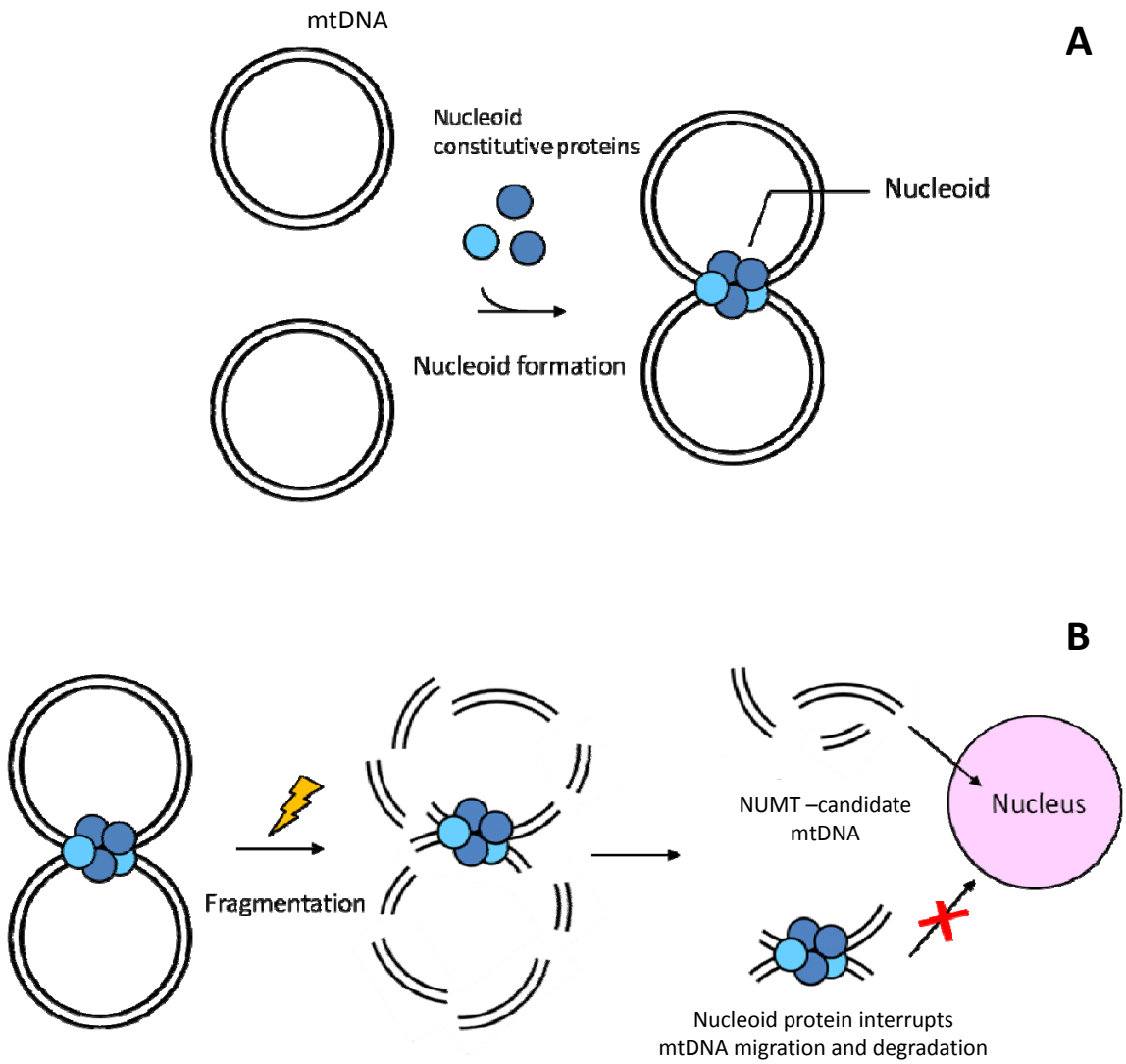

Figure 2.10: mtDNA transfer and nucleoid consisted proteins at promoter region

A: The promoter regions are the center of forming mitochondrial nucleoids.
B: The hypothesized mechanism of the transfer interruption of mtDNA fragments by binding proteins.

2.2.2.2 The impact of under and over counting of mtDNA insertion in previous studies

Some previous studies analyzed the features of mtDNA migration. However they did not report the paucity of D-loop derived NUMTs discovered here. We speculate this discrepancy is due to mis-counting of authentic mtDNA insertion events. Older NUMTs in the nuclear genome are highly diverged from the original mitochondrial fragments by nuclear duplications, indels or retrotranspositions. Because of those events, in NUMT identification, no consideration of duplicated regions gives overestimated results (i.e. over count; Figure 2.11A). Moreover, indels and retrotranspositions in NUMTs generate the separated hits between mtDNA and the nuclear DNA, so the outcome of homology search gives partial hits of authentic inserted fragments (i.e. under count; Figure 2.11B, C). Furthermore, there is another factor which produces under counting hits. This is the difference between the shape of real mtDNA and the shape of mtDNA in the sequence file. Despite mtDNA is circular, it is represented in linear shape in the sequence file (e.g. FASTA file). When true hits lay in the edges of linear mtDNA sequence file, they are not collected if their length is not enough for detection (Figure 2.11D).

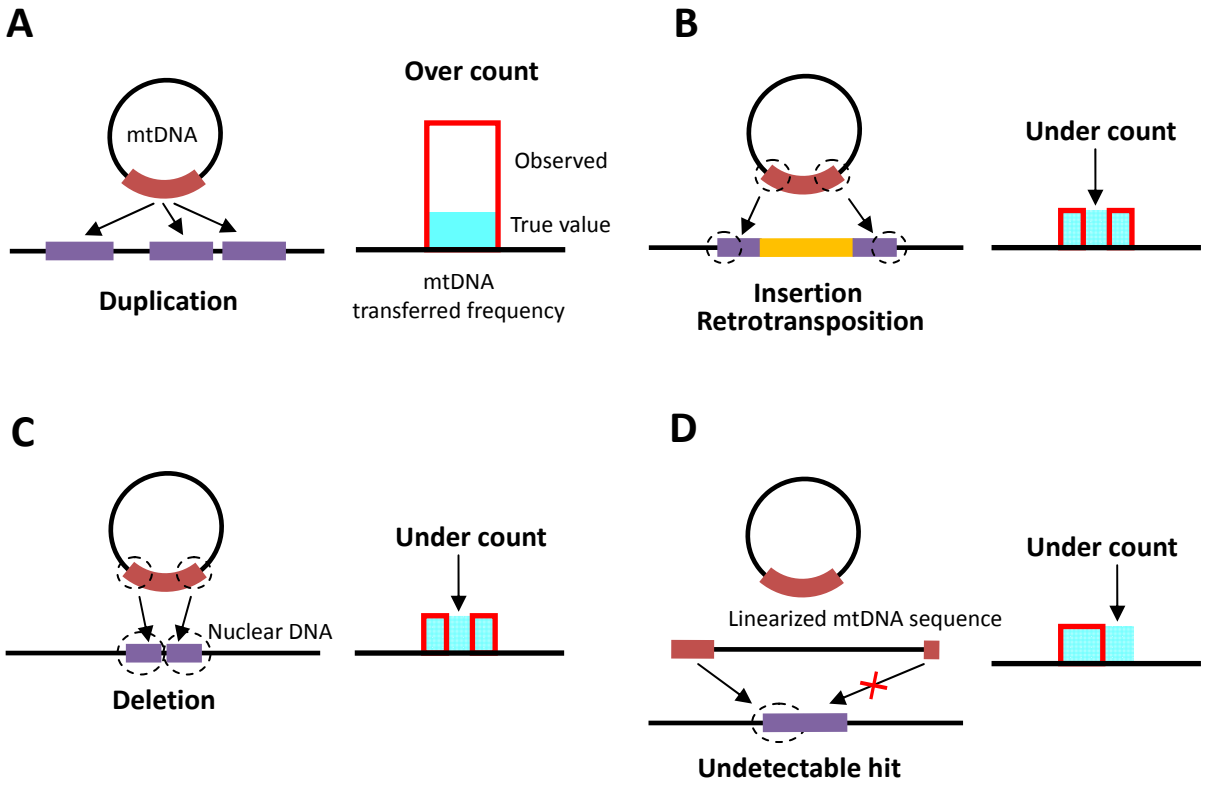

Figure 2.11: Reasons for under and over counting of NUMT insertion events

In previous studies, the possibility of under and over counting was not considered. This is the key point of the difference from our study. The under and over counts lead to false positions of NUMTs and their source of mtDNA fragments. When the under and over counts were not considered, we got the random-shape mtDNA displacement frequency (Figure 2.12).

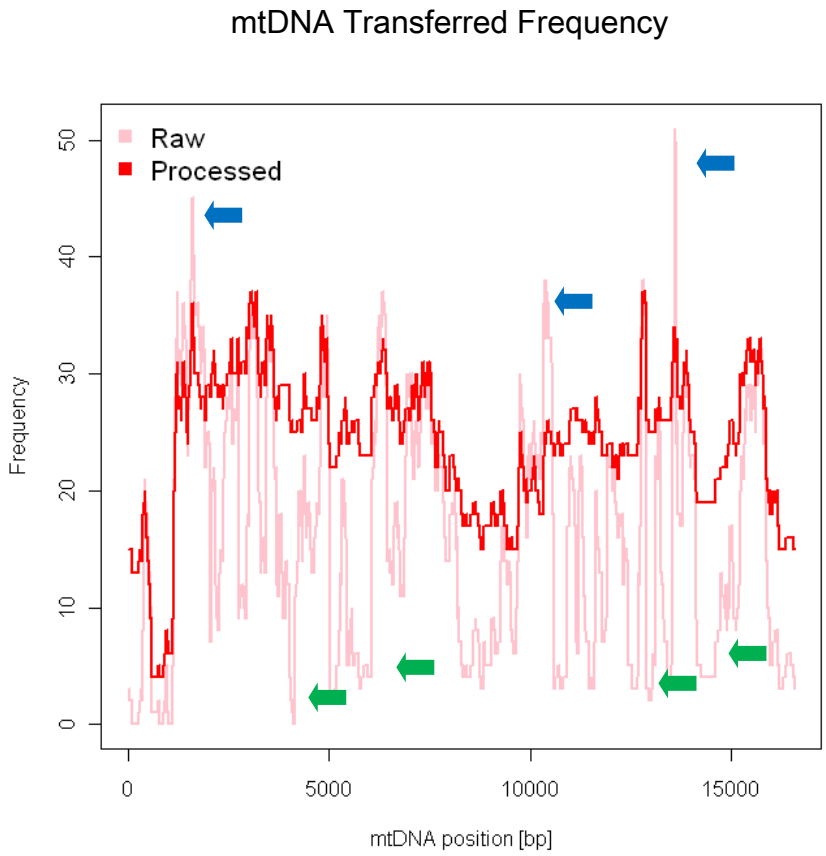

Figure 2.12: Difference between raw data and processed data.

Pink line: raw data which is contained under and over counts
Red line: processed data which is complemented under counts and reduced over counts
Blue arrow ⬅ : over counting regions
Green arrow ⬅ : under counting regions

2.3 *The feature of NUMTs insertion site in nuclear genome*

2.3.1 NUMT distribution on chromosomes

The aim of this section is to find the feature of NUMT insertion site. First we inspect the NUMT chromosomal distribution in mouse and human. In subsequent analyses, we mainly focused on human NUMT.

2.3.1.1 Human NUMT distribution

Whether NUMTs form clusters or exhibit insertion hotspots on specific chromosomes, chromosomal NUMT distribution in human was examined. As you can see in Figure 2.13, there were no obvious clusters and no clear hotspots. However, gently a small correlation with age might exist (P-value = 0.43 by random sampling with 1000 trials). NUMTs, which are in same age or in the close age, tend to occur on the same chromosome; especially chromosome 16 is a good example for observing age depending insertion manner. NUMTs in chromosome 16 are only consisted by age (a) NUMTs.

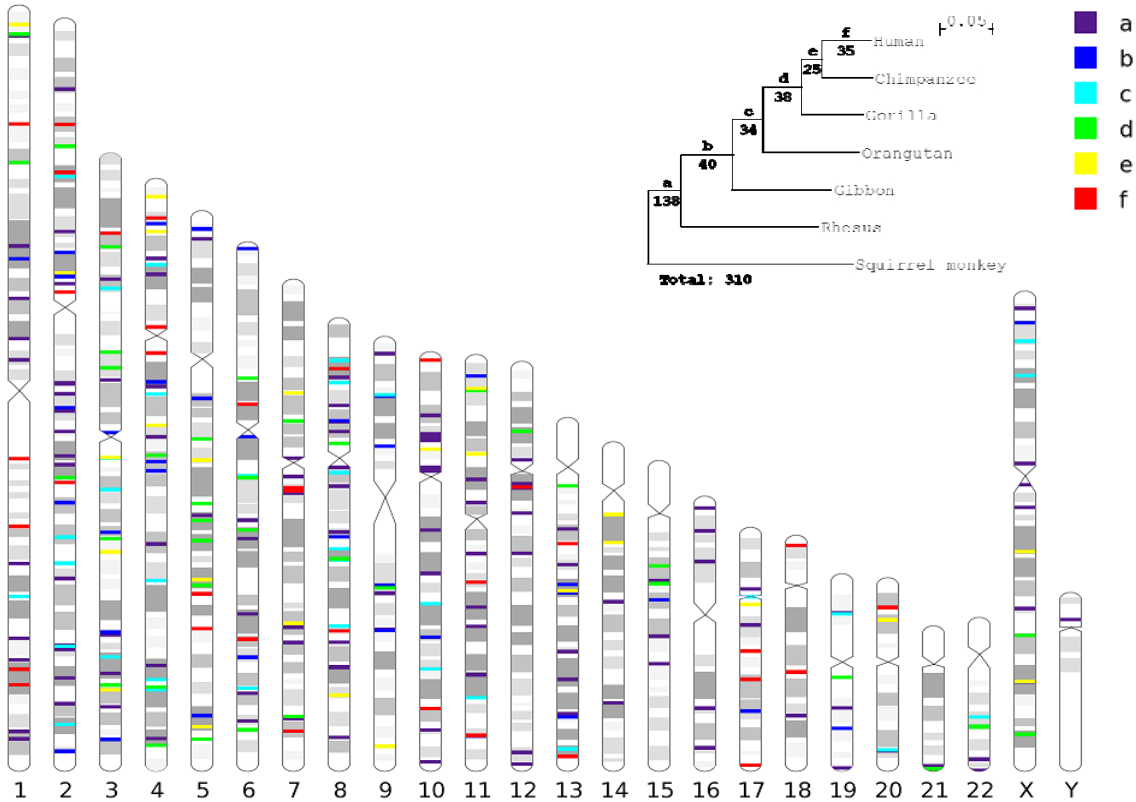

Figure 2.13: Human NUMT distribution on chromosomes

2.3.1.2 Mouse NUMT distribution

Next, we also investigated mouse NUMT distribution on its chromosomes (Figure 2.14). Like human's, mouse NUMT distribution did not exhibit any concrete clusters or hotspots. Similarly to human, the mouse NUMTs exhibited a slight, but statistically insignificant (P-value = 0.49 by random sampling with 1000 trials) correlation between insertion age and chromosome. Although the sample number is very few, the NUMTs in chromosome 16 were only age (j). The NUMTs in chromosome 2 shows also age depending manner; almost they are age (i), and the one of them is age (h) which is just previous age of age (i).

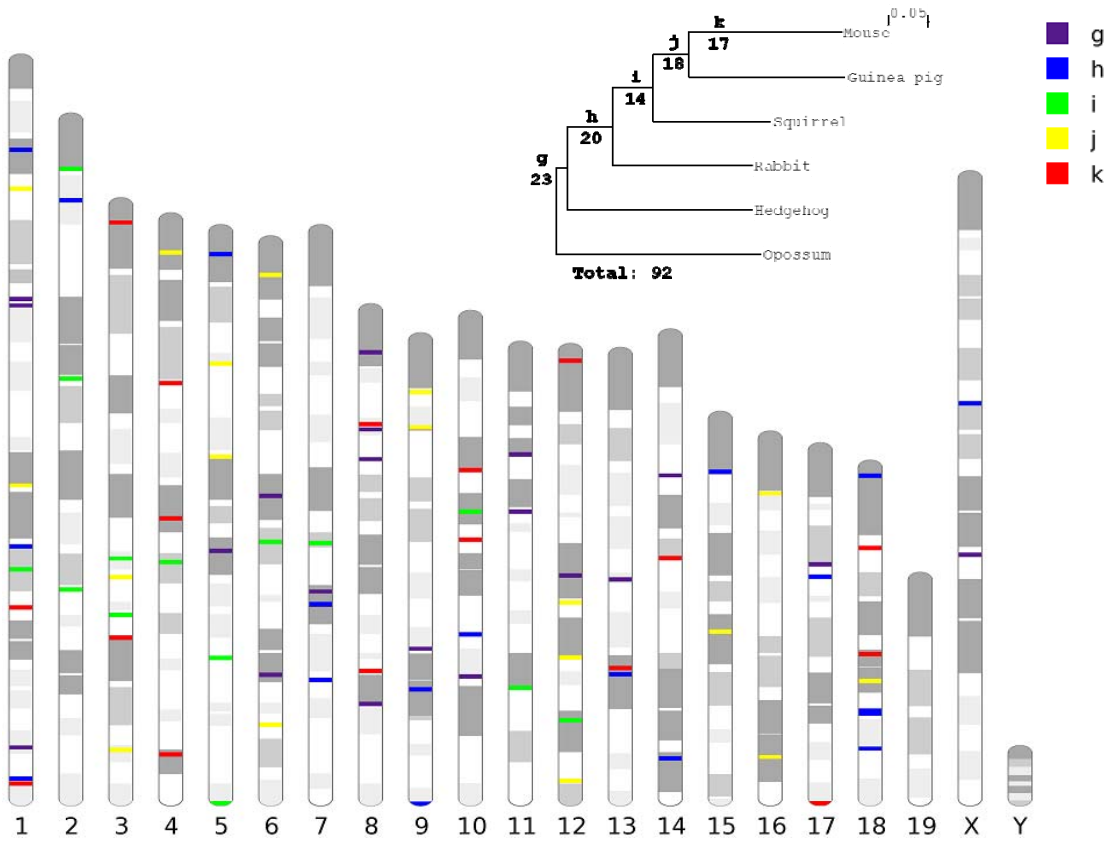

Figure 2.14: Mouse NUMT distribution on chromosomes

2.3.2 NUMT insertion sites and chromosomal fragile sites

Next, considering the proposed DSB-mediated NUMT insertion mechanism, we focused on chromosomal fragile sites (FSs). FSs are the regions which tend to produce breaks and gaps [30-32]. The correlation of NUMTs distribution and FS were investigated. For the availability of FS information [33, 34], we mainly analyzed human NUMTs and FSs. As a result, about 30% of NUMTs were within or adjacent to an FS region, however, the majority of them (approximately 70%) were located in non-FS regions (Table 2.4). From the calculation of P-values, we observed NUMTs and FSs were not correlated (P-value≈0.67) although FSs are the region which easily produce DBSs which are the NUMT insertion clues.

Table 2.4: The number of NUMTs in FSs and not in FSs

Age	Number of NUMTs	NUMTs in FSs		NUMTs not in FSs		P-value
		Sample number (%)		Sample number (%)		
a	138	41	(29.71)	97	(70.29)	0.6667
b	40	12	(30.00)	28	(70.00)	0.6667
c	34	7	(20.59)	27	(79.41)	0.6668
d	38	10	(26.32)	28	(73.68)	0.6667
e	25	7	(28.00)	18	(72.00)	0.6667
f	35	11	(31.43)	24	(68.57)	0.6667
total	310	88	(28.39)	222	(71.61)	0.6667

2.3.3 Flanking gene length

As another nuclear feature of producing DSBs, the effect of gene length is proposed in a recent research [35]. The research mentioned the probability of arising DSBs increases when long genes are transcribed. This is because, single strand DNAs, which are weaker than double strand DNAs, of longer genes are more exposed to surrounding environment for long time in their transcription. For testing the fact that NUMTs are more inserted in longer genes, first, we picked up the genes in human genome which lay within 100bp/10kbp upstream or downstream in NUMTs. The length on the genome (including introns) of all human genes and the length of the genes in NUMT flank are in Figure 2.15.

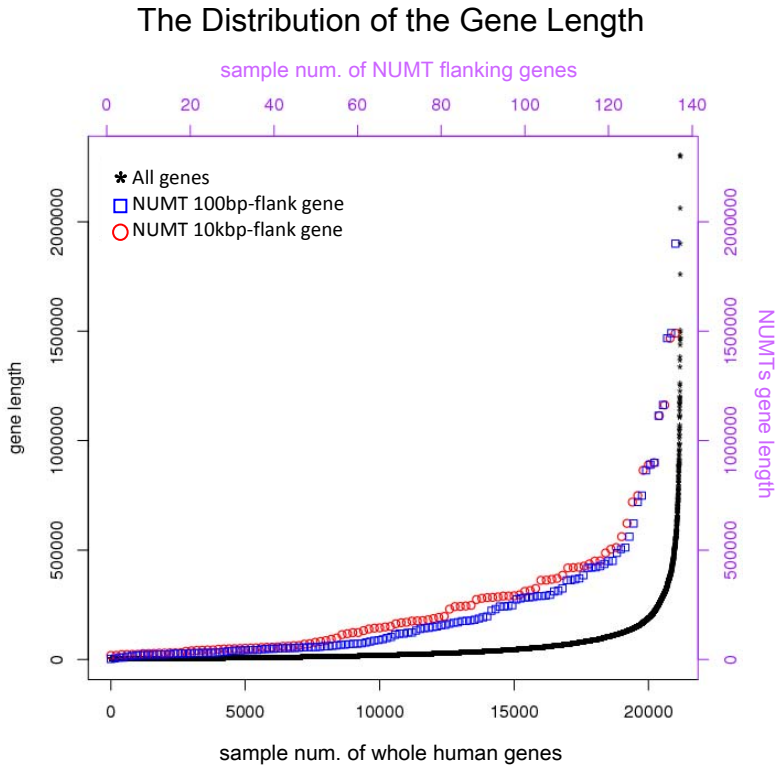

Figure 2.15: Overview of the length of whole human genes and NUMT flanking genes

Black asterisks: the length of whole human genes
Red circles: the length of the genes in 100bp flank of NUMT
Blue squares: the length of the genes in 10kbp flank of NUMT

Superficially, the length of the NUMT-flank genes seems to be longer as compared to the distribution of whole gene length. However, Figure 2.15 is just the plot of the gene length; namely insertion events are not considered. In other words, there is the potential that longer genes have more space to adopt new fragments than shorter genes, so this concept raises a question about the correlation of NUMT insertion events and its flanking gene length. To confirm whether this correlation exists or not, we simulated the distribution of randomly inserting "NUMTs" by randomly choosing genomic positions covered by some gene. Next, considering the difference of NUMT ages, we separated the dataset of NUMT flanking genes by each age. After that, we tested with Kolmogorov-Sminov test to see the difference of the distribution of gene length between sampled genes and NUMT flanking genes, and we also tested with Wilcoxon-Mann-Whitney test to observe the difference of the average of gene length between the two objects. The threshold of P-value was set as 0.05. The calculated result is shown in Table 2.5.

Table 2.5: The average of tested P-value about gene length and NUMT insertion

Age	The average of P-value (Threshold < 0.05)	
	Kolmogorov-Sminov	Wilcoxon-Mann-Whitney
a	0.5991	0.5873
b	0.2121	0.0534
c	0.3878	0.3149
d	0.3457	0.4455
e	0.5102	0.4175
f	0.7312	0.5408
all	0.2387	0.2145

From Table 2.5, the majority of P-value in all ages distributed more than the threshold, 0.05. Therefore, we concluded that there was no significance about the relation between NUMT insertion events and gene length.

2.3.5 GC content of upstream and downstream of NUMTs

We investigated the GC content of upstream and downstream of human NUMTs to look for the feature of NUMT insertion sites. First, we picked up sequences which are 500bp upstream and downstream regions of NUMTs. Then GC content of those extracted sequences was calculated using a sliding window of 50 nucleotides, shifted 20 nucleotides at a time. Moreover, we expected the potential of age specific evolution of NUMTs, so the result was filtered by each age. The final result in each age was plotted in Figure 2.16. From 2.15, the GC content of NUMT flanking sequences in age (a), which is the oldest age, were most stable (median: 38%). In age (b)~age (e), the median of the nearest sections from NUMTs is approximately 38% although the transit variation of GC content in each age was slightly different through the positions. NUMTs in age (f), the youngest age, have a lot of outliers which are extremely high GC content, so its median of the nearest sections from NUMTs is 42%. However, it is possible that those slight differences of GC content might be due to chance.

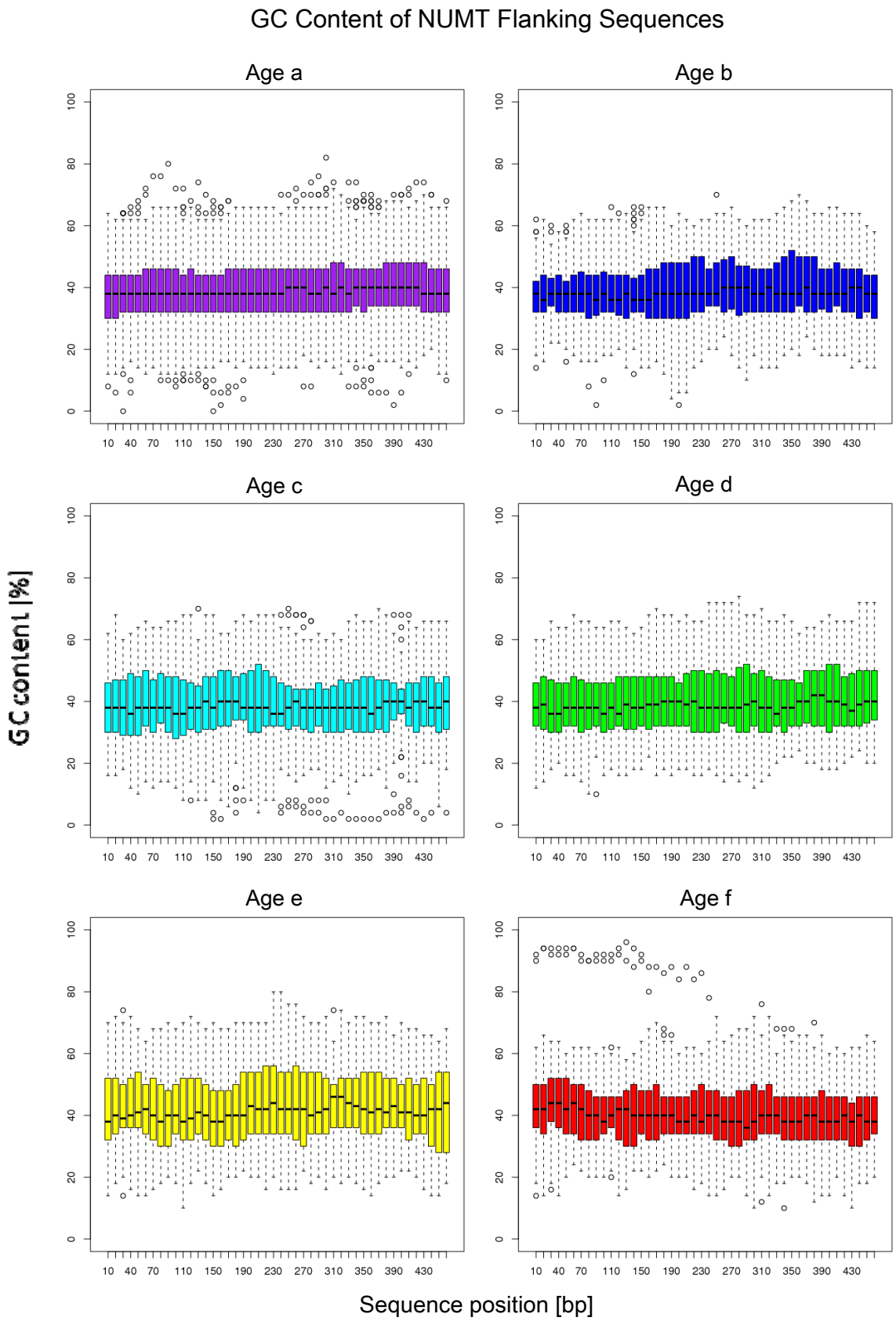

Figure 2.16: GC content of 500bp upstream and downstream of NUMTs.
Bigger positions indicate more outer regions from NUMTs.

For additional information for the nearest sections of NUMTs, we investigated the entire GC content of immediate 100bp upstream and downstream from NUMT sequences (Figure 2.17). As the result, the median of GC percentage seems to increase in younger ages; especially in age (f), however it might be also due to chance when each distributional deviation was considered.

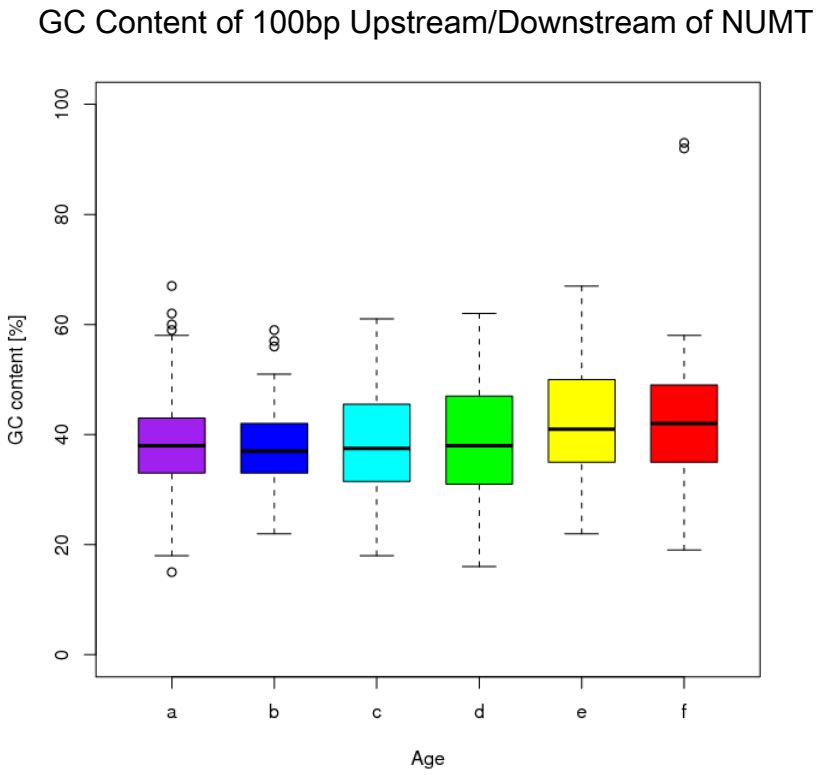

Figure 2.17: GC content of immediate 100bp up- and downstream of NUMTs through ages

2.3.4 Motif survey of NUMT-insertion sites

For finding some motifs in NUMT insertion sites, we extracted 10bp of upstream and downstream of human NUMT (-10bp) and 11bp of human NUMT sequence (+11bp). In addition, we also separated NUMT dataset per each age as the same reason as the previous section. Then -10bp/+11bp sample sequences were analyzed by seqLogo version 1.10.0 which is an R package for plotting DNA sequence logos. From the result in Figure 2.18, adenine and thymine meagerly appeared in NUMT flank. However the value of their bits is extremely small, so we can conclude there is no clear motif in NUMT flank.

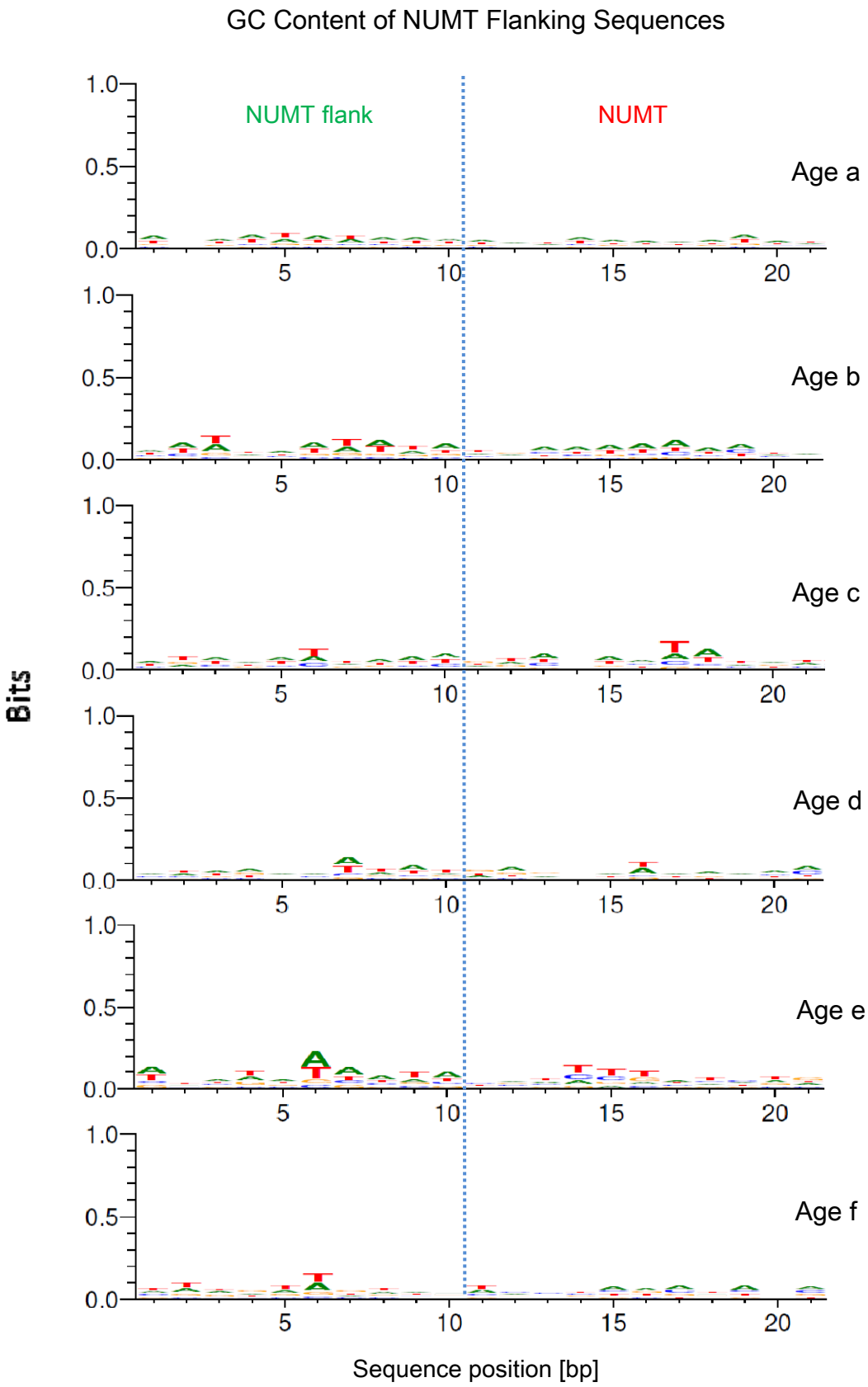

Figure 2.18: The outputs of seqLogo.

2.3.6 The oligonucleotide frequency of human NUMT flank

The genomic nucleotide proportions as the background was not taken in previous analyses, motif search by seqLogo and GC content investigation. In this section, we investigate oligonucreotide frequency of human NUMT flank per each age, which is adjusted by the frequency of genomic background. The calculated oligonucleotide frequency was converted to the entropy-like value (see also: Chapter 4 Methods).

2.3.6.1 Dinucleotide frequency

First of all, we checked the dinucleotide frequency of NUMT flank. Through all the ages, the dinucleotides constituted by A and T (i.e. AA, AT, TA and TT) appeared frequently in NUMT flank (Figure 2.20). However, the frequency of AT in age (c) and age (d) showed weaker peaks around the bound of NUMT and its flank than the frequency of other ages. It would appear that these unclear peaks in age (c) and age (d) came from the smallness of the sample numbers. Except age (a), the number of NUMTs in age (b)~age (f) is about 25~40. Therefore, it can be considered the contiguous value of the calculated nucleotide frequency became rough because of the fewness of samples. So we merged calculation results of all ages because the frequency in each age showed the tendency of AT richness. Then we plotted the merged dinucleotide frequency (Figure 2.19). Finally, the dinucletide frequency of entire NUMT flank strongly exhibit AT richness (P-value < 10^{-7}).

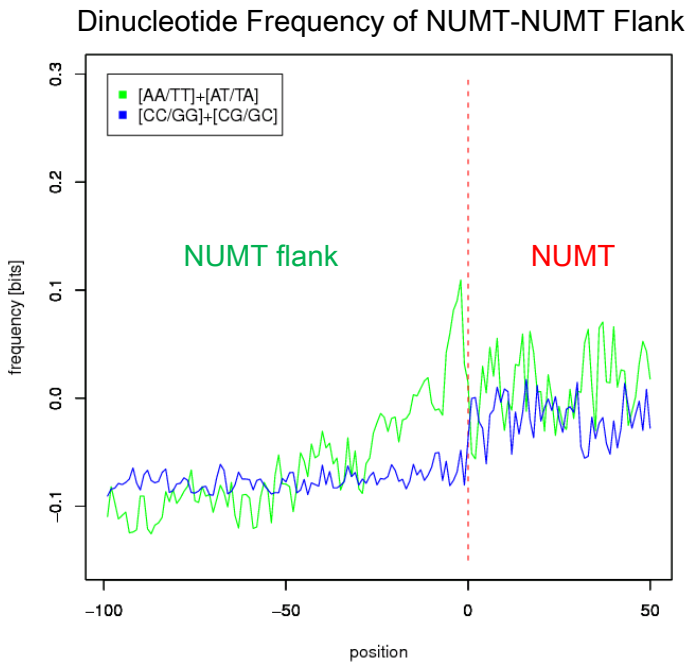

Figure 2.19: The common dinucleotide frequency of all NUMT flank.

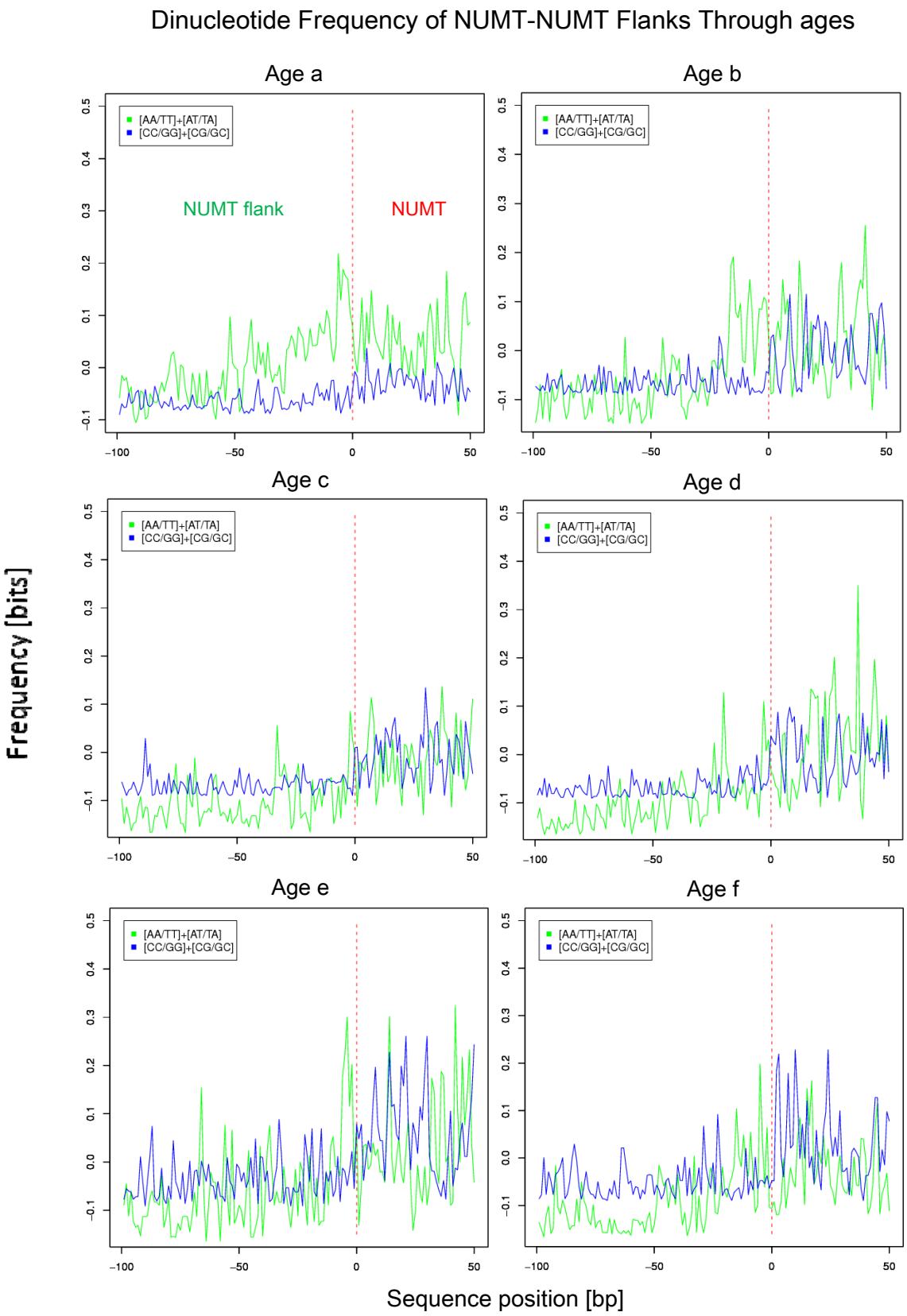

Figure 2.20: The dinucleotide frequency of NUMT flanks through ages
The position 0bp indicates the border between NUMT and NUMT flanks.

Next, we examined the trinucleotide frequency of NUMT flanks through all ages. Four classes of trinucleotides; triplets of A/T and G/C, [A/T]{2}[G/C] and [G/C]{2}[A/T], [G/C][A/T]{2} and [A/T][G/C]{2}, [A/T][G/C][A/T] and [G/C][A/T][G/C], were plotted in Figure 2.21. Incidentally, 12 trinucleotides were not used in this analysis. The detail of the trinucleotides is in Table 2.6.

Table 2.6: Trinucleotides used in this analysis.

[AAA/TTT]	[GGG/CCC]	[A/T]{2}[G/C]	[G/C]{2}[A/T]
		AAC	GGA
		AAG	GGT
		ATC	GCA
AAA	GGG	ATG	GCT
TTT	CCC	TAC	CGA
		TAG	CGT
		TTC	CCA
		TTG	CCT

[G/C][A/T]{2}	[A/T][G/C]{2}	[A/T][G/C][A/T]	[G/C][A/T][G/C]
ACC	GAA	ACA	GAG
AGG	GTT	ACT	GAC
ACG	GAT	AGA	GTG
AGC	GTA	AGT	GTC
TCG	CAT	TCA	CAG
TGC	CTA	TCT	CAC
TCC	CAA	TGA	CTG
TGG	CTT	TGT	CTC

Not considered	
AAT	GGC
ATT	GCC
ATA	GCG
TAT	CGC
TAA	CGG
TTA	CCG

From Figure 2.21, despite the smaller number of included trinucleotides than other classes, the peak of AAA/TTT appeared beside the border of NUMT-NUMT flanks (P-value < 0.001). Meanwhile, the frequency of other trinucleotides did not show any obvious peaks from NUMT flanking sequences to NUMTs. Therefore, those observations were summarized as following: AT triplets tend to assigned to NUMT insertion sites.

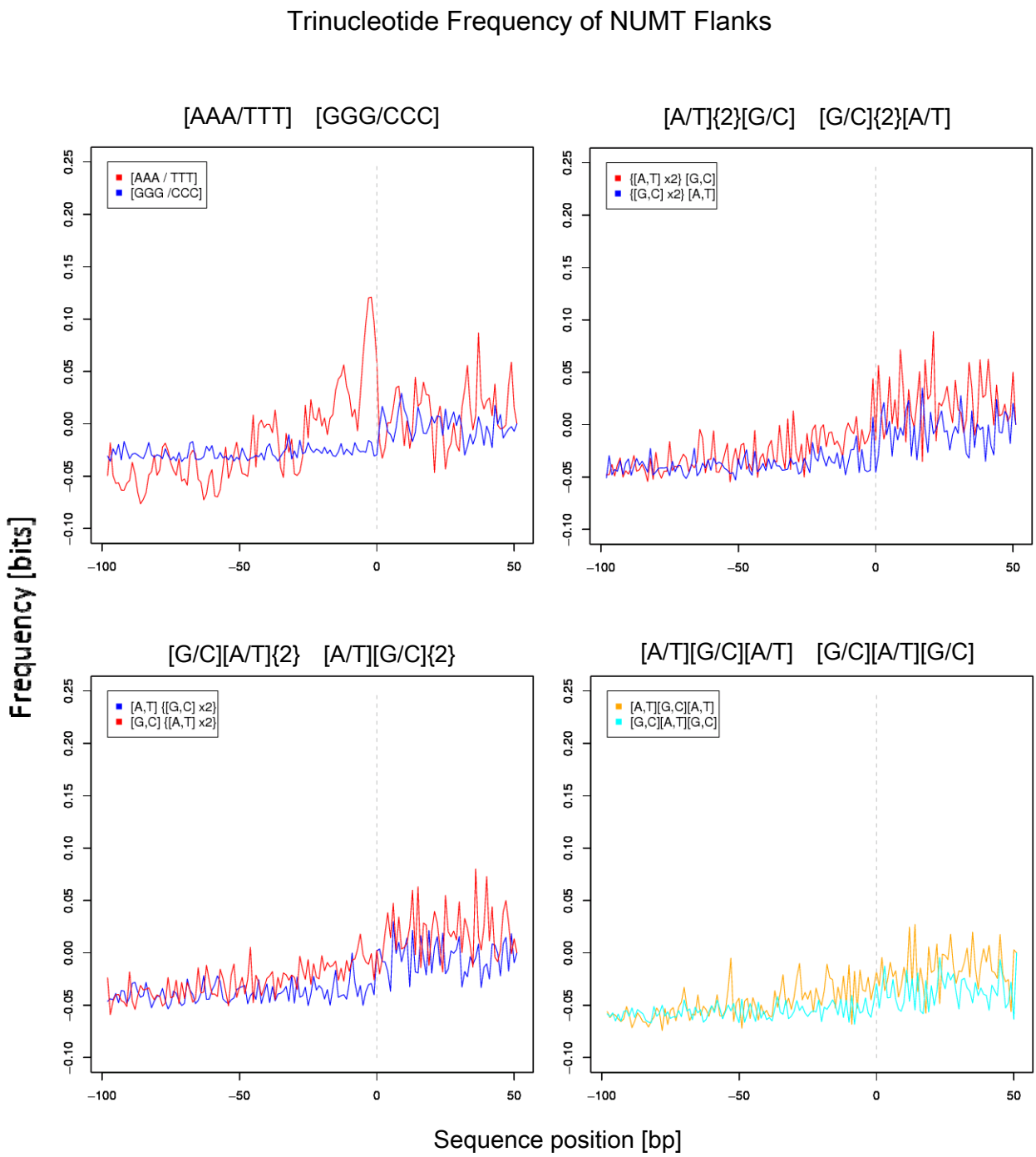

Figure 2.21: The trinucleotide frequency of NUMT flanks

Red line: AT rich trinucleotide constituted by more than two contiguous AT nucleotides
Blue line: GC rich trinucleotide constituted by more than two contiguous GC nucleotides
Yellow line: AT rich trinculeotide but not contiguous
Cyan line: GC rich trinucleotide but not contiguous

To confirm AT triplets tendency, we also investigate the trinucleotide frequency in mouse NUMT flank. From Figure 2.22, the pattern of trinucleotide frequency of mouse NUMT flank was very similar to the human's one. Mouse NUMT flank also exhibits AT triprets peak although the peak slightly leaked over the border of NUMT and NUMT flanks. Therefore, from those observations, we concluded that NUMTs tend to be integrated in such AT oligomers.

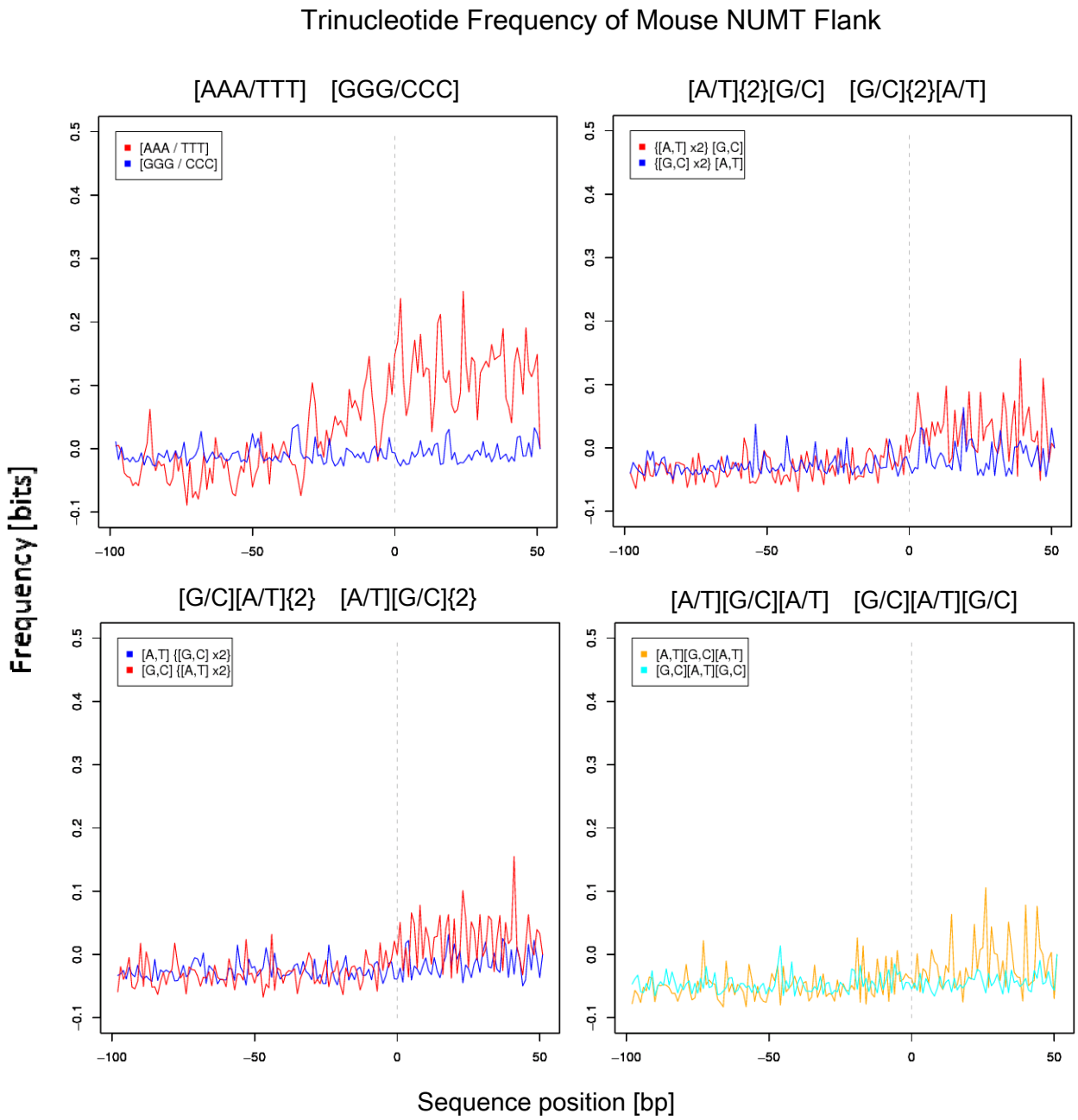

Figure 2.22: The trinucleotide frequency of mouse NUMT flank

2.3.7 Retrotransposons and NUMT-insertion sites

The retrotransposon encoded endonuclease is known to create DSBs in AT oligomer motif, 5′-TTTTAA-3′. Interestingly, the flanking regions of NUMTs are rich in such AT oligomers [36-38]. A recent study shows LINE-encoded endonuclease creates DSBs not only in the locale of retrotransposition, but also in the other retrotransposition-free regions by the free diffusion to other nuclear spaces [39]. Other studies also mentioned the number of LINE-induced DSBs was greater than the predicted numbers of successful LINE insertions [40-42]. Futhermore, poly A tails of retrotransposons become the staging point of another retrotransposition because of the similarity in the AT oligomers of endonuclease recognition sites. From those facts, it has the potential that the truncation sites by retrotransposon encoded endonucleases become NUMT insertion sites. To corroborate this possibility, we investigated the density of retrotransposons in 500bp of human NUMT flanking regions by crosschecking the position information of various sorts of repeats with the output of RepeatMasker and Tandem Repeats Finder. We also checked in NUMT flanking regions of mouse and rhesus monkey. The result is in Figure 2.23.

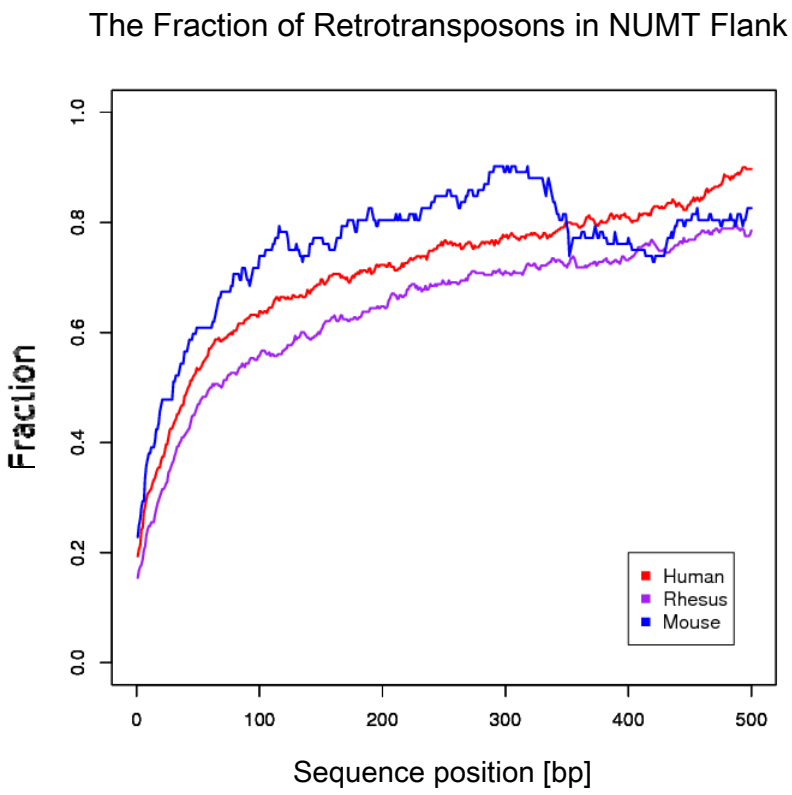

Figure 2.23: The proportion of retrotransposon

The occurrence of retrotransposons is statistically enriched in NUMT flanking regions. These trends were also observed in mouse and rhesus monkey. The detail is also in Table 2.7 in the next page. This suggests that many NUMTs are inserted via endonuclease created DSBs. Just for the record, other repeats in NUMT flank was plotted in Figure 2.24.

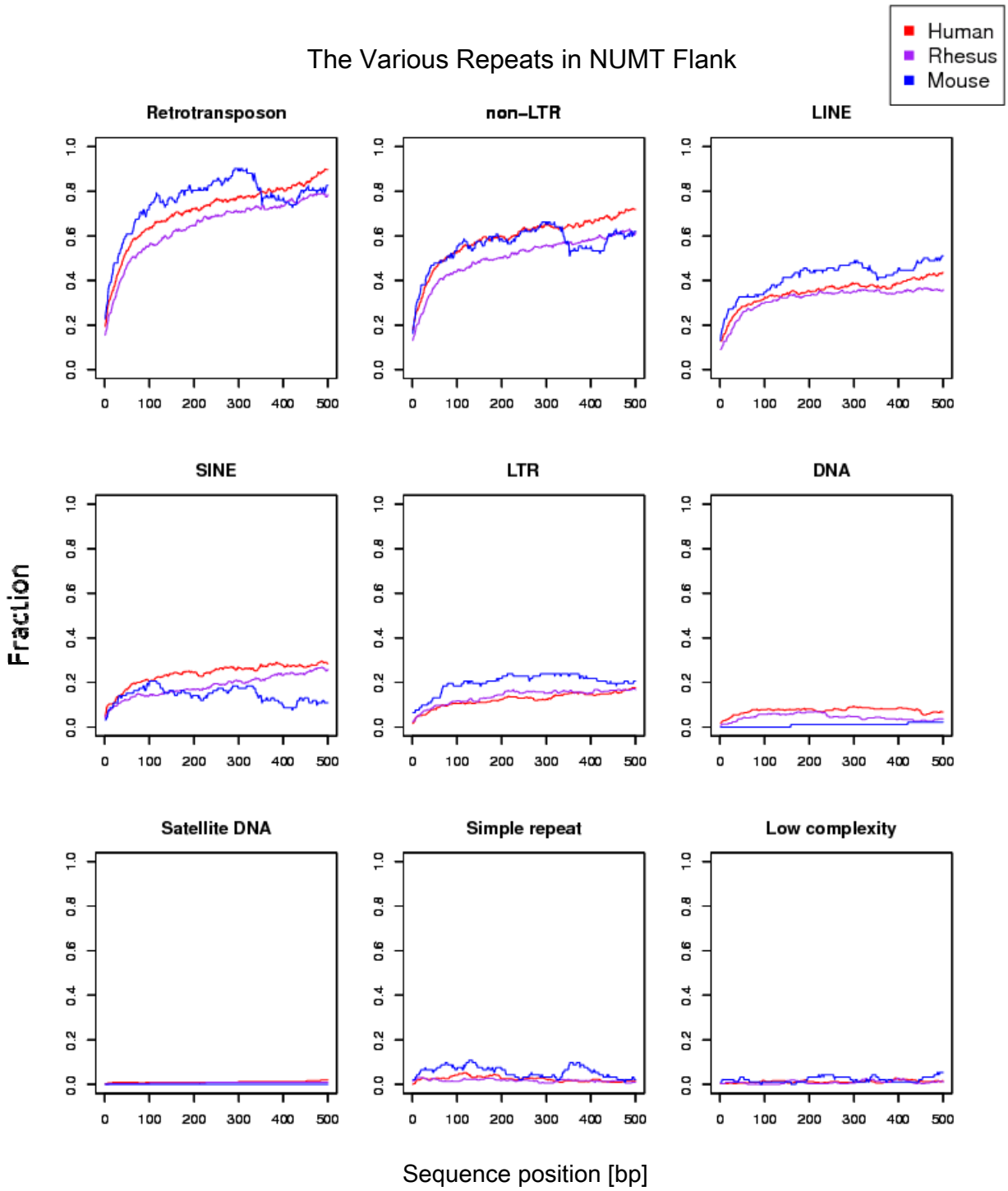

Figure 2.24: Various repeats within NUMT flanking regions
Bigger positions indicate more exterior regions from NUMTs.

Table 2.7: The percentage of the NUMT containing retrotransposons in its flank

Species	Number of NUMTs	Percentage of retrotransposon containing positions
Human	310	90.00
Rhesus	298	79.53
Mouse	92	90.22

2.3.8 Discussion

From the series of the analysis in this section, we can conclude that a feature of NUMT insertion sites in the nuclear genome is AT-oligomer richness. AT-oligomers are used as the recognition site of retrotransposon encoding endonuclease creating DSBs, and it becomes a primer for the accumulation of retrotransposition (i.e. target-primed reverse transcription; TPRT). According to a recent study, the retrotransposon coded endonuclease makes DSBs in other retrotransposon-free nuclear regions as well as in retrotransposition ocurring positions [39]. Moreover, this enzyme is related in the pathway of DSB repair [39]. NUMT insertions were mediated in DBS repair. Therefore, putting these observation together, we hypothesize that NUMT insertion sites may be related to the recognition site of retrotransposon encoding endonuclease.

In this paragraph, we focused on the difference of the number of accumulated NUMTs in species. As mentioned in Chapter 1 Introduction, NUMTs have been identified in many eukaryotes, however the numbers of NUMTs differ widely between species; some species possess several hundred NUMTs, but others have no detectable NUMTs at all. Clifton S. et. al. suggested that this variation of NUMTs population arises from the difference of mtDNA copy number and its length in each species [6]. The species retaining fewer copies and the shorter length of mtDNA have fewer chances of NUMT accumulations. However, there are several exceptions to this rule, so the reason behind the NUMTs population difference between species is still an open question. Now, as an additional new hypothesis, we propose that the NUMT population difference is the product of the influence of retrotransposon activity. This hypothesis is consistent with the abundance of sequences similar to the retrotransposon encoding endonuclease recognition site adjacent to NUMTs. As the example, we can explain with human, mouse, rhesus monkey, fruit fly and worm NUMTs. The number of NUMTs, the density of retrotransposon and genome length for those four species are shown in Table 2.8.

Table 2.8: The number of NUMTs in human, rhesus monkey, mouse, fruit fly and worm.

Species	Number of NUMTs	Retrotransposon density [%]	Genome length [Mb]
Human	310	40.58193	3200
Rhesus	298	39.82778	2800
Mouse	92	36.56292	2700
Fruit fly	6 [43]	2.2	170
Worm	1 [43]	0.4	100

From Table 2.8, we can see the number of NUMTs in fruit fly and worm is extremely few. The retrotransposon density in those two organisms was also less than other NUMT-rich species, which suggests the frequency of retrotransposition was possibly related to the accumulation rate of NUMTs. This does not explain the difference between mouse and the two primate genomes however; in those mammals, the retrotransposon density was comparatively similar, but the number of mouse NUMTs was approximately one third of the number of human and rhesus NUMTs. This observation might be explained by the retrotransposon burst in the ancestor of human and rhesus [44]. As mentioned above, the retrotransposon encoded endonuclease creates DSBs not only in retrotransposition sites, but also in retrotransposon-free sites in the nuclear genome. Therefore, we may speculate that the burst of retrotransposon activity explains the observed number of NUMTs.

2.4 *Survey of transcribed and functional NUMTs*

In this section, we focused on the behavior of NUMTs within the nuclear environment. Especially, the information analysis by referring to annotation databases was done for scoping out functional NUMTs in human and mouse. In addition, the NUMT transcribed frequency in mouse was calculated with the 5′-CAGE data.

2.4.1 Functional NUMTs in nuclear genome and those accumulated age

2.4.1.1 Human NUMTs

We crosschecked our human NUMT dataset against annotation databases (see also: Chapter 4 Methods). Almost all NUMTs (99%) were non-coding (nc) regions or introns. Only 1% of NUMTs have annotated functional roles in human genome (Figure 2.25A). Moreover, as an interesting point, those functional NUMTs were inserted in relatively-recent period; age (e)~age (f) (Figure 2.25B).

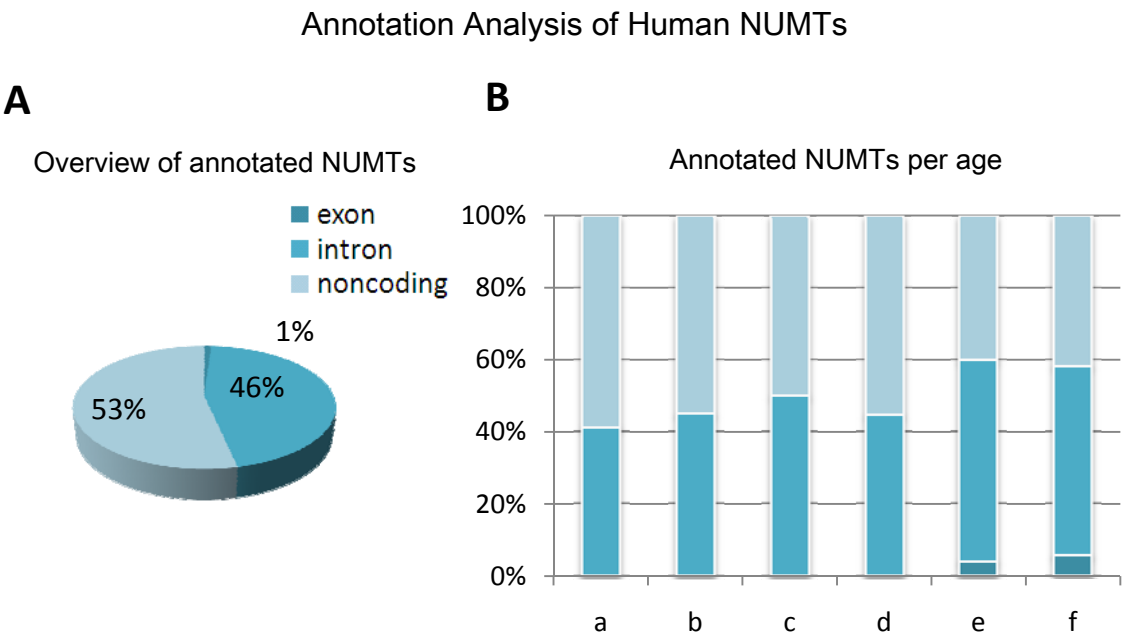

Figure 2.25: The annotated human NUMT in each age

They were mainly ncRNAs or 3'UTRs of mRNA. The human and chimpanzee shared NUMTs are transcribed as ncRNA, and those ncRNA express in several tissue (Table 2.9). Especially, the human specific functional NUMT; age (f) was an ncRNA which expresses only at brain, and it concerns the fetal brain development.

Table 2.9: The detail of annotated NUMTs in each age

Age	Transcripts	Length [bp]	Subcellular location	Accession no. [database]
e	ncRNA	1284	liver, fetal liver	uc001miq.1 [UCSC Genes)
	ncRNA	1252	brain	CR595123 [GenBank]
	ncRNA	1288	brain	CR599987 [GenBank]
	ncRNA	1258	brain, fetal brain	CR606764 [GenBank]
	ncRNA	1285	thyms	CR621961 [GenBank]
f	RSPO1; 3'UTR	74	secreted	NM_001038633 [RefSeq]
	ncRNA	1636	brain, fetal brain	uc003laf.1 [UCSC Genes]
	ncRNA	730	brain	CR616105 [GenBank]

2.4.1.2 Mouse NUMTs

We analyzed mouse NUMT dataset in the same way. Like human NUMTs, the most of NUMTs (98%) were non-coding regions or introns. However the fraction of non-coding regions in mouse was larger than the non-coding regions in human. 2% of NUMTs was annotated as functional elements in mouse genome (Figure 2.26A). Surprisingly, in the same fashion of human NUMTs, the insertion of the functional NUMTs occurred in recent age; age (j)~age (k) (Figure 2.26B).

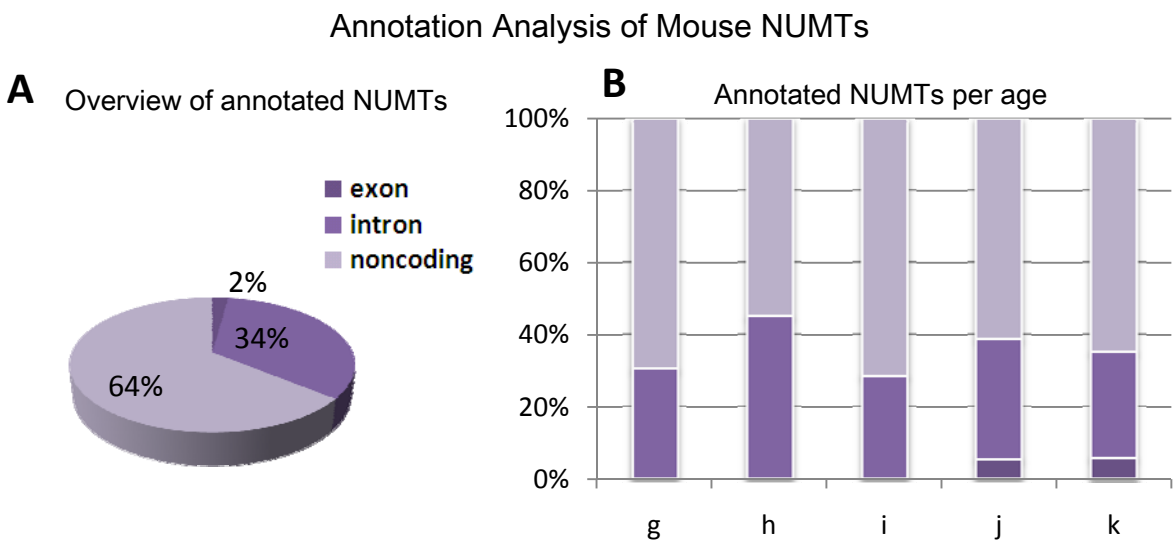

Figure 2.26: The annotated mouse NUMT in each age

The detail of the functional NUMTs is in Table 2.10. The majority of the functional NUMTs in mouse are also related nerves system like human NUMTs. The mouse and rabbit shared NUMTs in age (j) are transcribed in various elements; they are ncRNA, an unnamed protein, and Kiaa1731 which is an alternative spliced protein. The mouse specific functional NUMT in age (k) was an ncRNA which is expressed on fetal spinal cord.

Table 2.10: The detail of annotated mouse NUMTs

Age	Transcripts	Length [bp]	Subcellular location	Accession no. [database]
j	Kiaa1731	3092 12300	n/a*	uc009ofx.1 (UCSC Genes) uc009ofy.1 (UCSC Genes)
	Unnamed protein	3064	fetal head	AK134513 (GenBank)
	ncRNA	4592	fetal cerebellum	AK163204 (GenBank)
	ncRNA	2584	fetal body between diaphragm and neck	AK035174 (GenBank)
k	ncRNA	2022	fetal spinal cord	AK083088 (GenBank)

* n/a: Not available

2.4.2 Transcription start sites in mouse NUMT regions

In this analysis, we used mouse 5′ CAGE data, for investigating transcriptional level of NUMTs in mouse. In living cells, there are a variety of transcripts: many functional transcripts used in biological process are annotated in databases such as the databases we referred to in the above analysis. Of course, some unannotated transcripts may have specific, yet unknown, function. Moreover, the existence of "junk" transcripts, which are the nonfunctional transcripts produced by "noisy transcription", has been proposed in the previous studies [45]. NUMTs often occur in the introns of genes and in this case the fact that they are transcribed is to be expected. In this analysis, we focused mainly on NUMTs which are included in non-coding regions, and observed their transcription frequency. The mouse embryonic CAGE data, used in this analysis (provided by Prof. Yutaka Suzuki) consists of sequenced 5′ ends of transcripts. This data was provided for several points in mouse embryo development: day7, day11, day15 and day17. The tag counts are shown in Figure 2.27. The NUMT ID numbers in the X-axis were numbered by the transcribed frequency in day7.

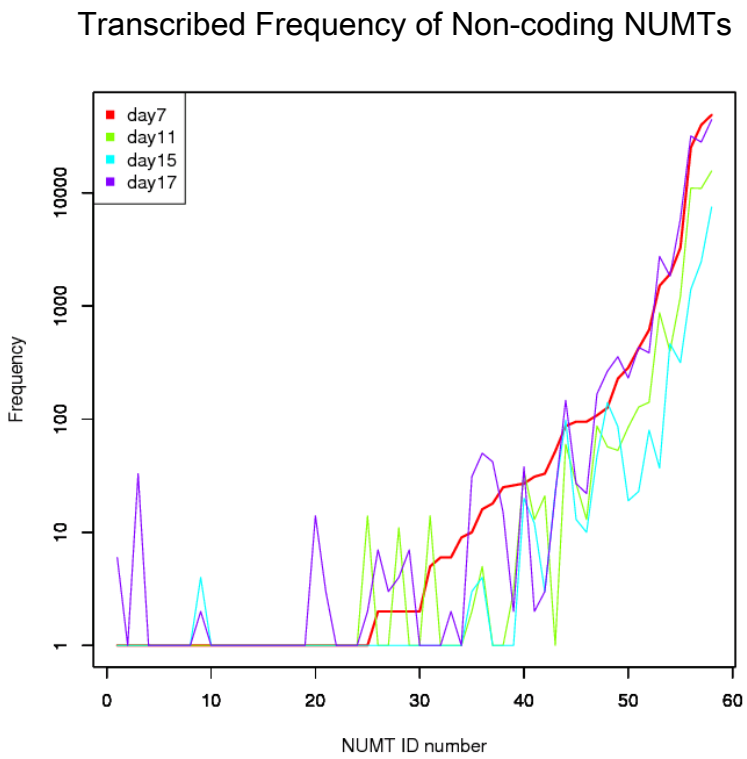

Figure 2.27: The transcribed frequency of non-coding NUMTs in mouse

From Figure 2.27, the transcribed frequency of non-coding NUMT differed between developmental stages, although functional genes were not annotated in the immediate vicinity of these NUMTs. 48,824 tags uniquely mapped to the most highly expressed non-coding NUMT, which is comparable to the median value (20,911) of the number of tags which mapped to NUMT-size region of known functional genes. So the level of the transcription of this non-coding NUMT is similar to that of known functional elements. Furthermore, we separated the data by evolutionary ages, and investigated the existence of the age-specific transcribed frequency (Figure 2.28). The non-coding NUMTs in age (i) and age (h); the ages halfway between the oldest age (g) and younger age (j) ~ age (k), were not so often transcribed (note: the Y-axis is log scale). However, overall, there was no big difference in the transcribed frequency of non-coding NUMTs through evolutionary ages.

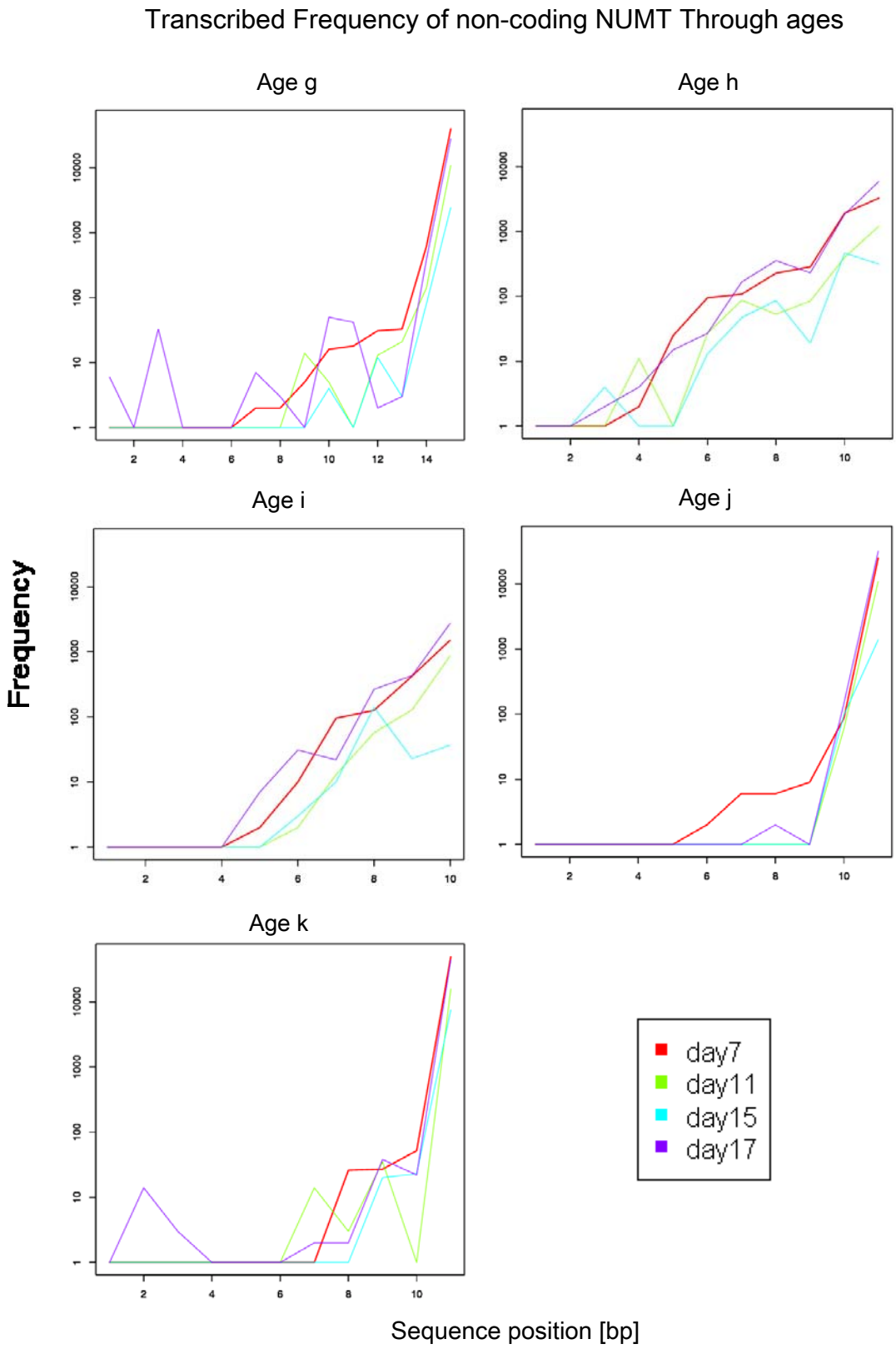

Figure 2.28: The transcribed frequency of non-coding NUMT through evolutional ages.

2.4.3 Discussion

In human NUMT annotation analysis, the human and chimpanzee shared NUMTs age (e) are transcribed as ncRNA, expressed in several tissue. However, interestingly, the human specific functional NUMT; age (f) was an ncRNA, expressed specifically in brain, and known to affect fetal brain development. Hence, it is possible that this element might contribute to the difference between human and chimpanzee brain structure. Moreover, similarity to human, mouse non-coding NUMTs were also mainly ncRNAs, expressed in the fetal brain. Several studies mentioned that NUMTs were seldom transcribed or functional, nonetheless, we discovered some functional NUMTs from the series of comprehensive NUMT annotation analysis.

Moreover, we identified, currently unannotated, transcripts of "non-coding" NUMTs from mouse 5'CAGE data. The level of the transcription is some cases comparable to known functional gene transcribed in mouse development. Thus it would appear that those transcribed non-coding NUMTs might be functional.

Chapter 3

Conclusion

As the conclusion of our study, we revealed various sorts of the characteristics of NUMTs across species. We found specific pattern of mtDNA migration to consider under and over counting hits: duplications, mtDNA circularity, indels and retrotransposition. Furthermore, from the observation of AT-oligomers and retrotransposon in NUMT flank, we also uncovered the common feature of nuclear NUMT integration sites in human, mouse and rhesus monkey. Those two characteristics concerning the formation of NUMTs were previously unknown for long time although several studies tried to reveal them.

Additionally, we extracted new functional NUMTs in human and mouse genome by comprehensive annotation referring analysis. Those functional NUMTs were not mentioned in previous studies. Moreover, applying phylogenetic analysis, we succeed to view all inserted NUMTs through the evolutional time. The knowledge in this study would help to understand genomic evolution and diversification among living organisms, and it poses an issue about the mitochondria involvement in the evolutionary process.

Chapter 4

Methods

5.1 *The list of sequence data source*

In this study, following sequence data was used. The nuclear genome data was downloaded from UCSC Genome Browser [http://genome.ucsc.edu/]. Human, chimpanzee, orangutan, rhesus, mouse and opossum mitochondrial genomes were also downloaded from the UCSC website. In the case of the mtDNAs of gorilla, gibbon, squirrel monkey, guinea pig, squirrel, rabbit and hedgehog, the data was downloaded from the NCBI Genome database [http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome].

5.1.1 Nuclear genome

Species	Identifier of UCSC	Identifier of NCBI Genome	Assembly
Human	hg18	NC_000001 - NC_000024	NCBI Build 36.1
Rhesus	rheMac2	NC_007858 - NC_007878	BCM HGSC Mmul_051212
Mouse	mm9	NC_000067 - NC_000087	NCBI Build 37.1
Rat	rn4	NC_005100 - NC_005120	BCM HGSC Rnor 3.4

5.1.2. Mitochondrial genome

Species	Identifier of UCSC	Identifier of NCBI Genome
Human	hg18	NC_001807
Chimpanzee	panTro2	NC_001643
Gorilla	---	NC_001645
Orangutan	ponAve2	NC_002083
Gibbon	---	NC_002082
Rhesus	rheMac2	NC_005943
Squirrel monkey	---	NC_012775
Mouse	mm9	NC_005089
Guinea pig	---	NC_000884
Squirrel	---	NC_002369
Rabbit	---	NC_001913
Hedgehog	---	NC_002080
Opossum	monDom5	NC_006299

5.2 *NUMTs data collection*

mtDNA is circular. In order to use BLAST with mtDNA, we linearized the mtDNA from 10 different starting points, and those 10 mtDNAs were BLASTed as queries against nuclear DNA (E-value<10^{-4}). BLAST-hits segmentalized by retrotransposons and indels were concatenated. Then duplicated hits were removed by comparing 100bp flanking sequence similarities (E-value<10^{-10}) and referencing the Segmental Duplication Database [46]. Through those procedures, NUMT position and corresponding mtDNA position were identified. These non-redundant hits were used as our NUMT dataset in our analysis.

5.3 *Phylogenic analysis of NUMT inserted age estimation*

Age specific patterns of NUMT insertion were tested. First, mitochondrial sequences were converted to a single linear sequence starting at the D-loop origin. Next, mtDNAs drawn from the same origin were aligned as multiple alignments by ClustalW version 2.0.11 [47]. Then we constructed a mitochondrial evolutionary tree by maximum-likelihood method with Phylip version 3.68 [48]. NUMT phylogenic trees were similarly obtained from each NUMT sequence and its corresponding mitochondrial partial sequences (E-value<10^{-10}). All tree producing procedures were bootstrapped 1000 times. Finally, each NUMT tree was compared to the mitochondrial tree, each NUMT leaf node location was checked by eye, and then NUMT-insertion age was estimated.

5.4 *The test for detectable length of short and old NUMTs*

We tested the detectable length of short and old NUMTs, by randomly extracting NUMT segments to simulate changing the length of old NUMT sequence. The length *l* was decreased from 1000bp to 50bp at intervals of 50bp. For this analysis, we use the oldest human NUMTs in age (a), because this NUMT dataset was, of course the oldest, and it contained sequences with various lengths. For the test, according to the each fixed length from 1000bp to 50bp at intervals of 50bp, we filtered out NUMT sequences shorter than the fixed length. Then, we randomly selected one NUMT and from it one length *l* segment, and masked the remainder of the NUMTs with "n" (hard-mask). After masking, we used BLAST to detect masked NUMT sequences queried with their corresponding mitochondrial sequence (E-value<10^{-10}). 1000 and 10000 trials where performed for each length *l*. From these trials we computed the false negative rate for detection for each length.

5.5 *The analysis of NUMTs in Chromosomal fragile sites*

The information of FSs was taken from the review of Michal S. et al [33]. However, this information was described as chromosome bands (e.g. 10q23.3, Xp22.31). So we converted the band information to the position based information by referreing to the annotation file of "cytoBand" in the UCSC website [49]. After that, we computed NUMTs

placed within FSs. Then, as the pseudo NUMTs, we randomly chose nuclear genome positions in the same number of NUMTs to test the relation of FSs underling NUMT insertion events, and the number of pseudo NUMTs was counted. We repeated this simulation 1000 times. Finally, 1000-time simulated data was filtered and counted using the true NUMT insertion number in FSs as the bound. From this, we calculated P-values of FSs.

5.6 *The significance test for Gene length in NUMT insertion sites*

We investigated the length of the genes in NUMT flanking regions (100bp upstream and downstream). As the significant test, we randomly selected nuclear positions as many as the number of NUMTs. Then next, we took a look around the 100bp of upstream and downstream, and checked the length of the genes around the selected pseudo NUMT-insertion positions. Those steps were duplicated to 1000 trials. The datasets of gene length from 1000-time trials were compared to true NUMT flanking gene length, with Kolmogorov-Sminov test and Wilcoxon-Mann-Whitney test. Applying those tests, P-values were calculated.

5.7 *The search for frequently occurring oligonucleotides in NUMT flank*

First of all, dinucleotide/trinucleotide frequency of the nuclear genome was calculated as the background; window sizes were of course 2bp and 3bp, shifted step was 1bp. Then the dinucleotide/trinucleotide frequency of 150bp NUMT-NUMT flank sequences (50bp of NUMT sequences and 100bp of the upstream/downstream) were computed. We normalized calculated frequency of NUMT-NUMT flank using the following formula.

$$Relative\ Entropy = \sum Observed \log_2 \frac{Observed}{Expected}$$

For confirming statistically significance of the peaks around NUMT-NUMT flank, we randomly selected nuclear regions as the same number of NUMTs, and then the dinucleotide/trinucleotide frequency was calculated. After 1000-time run, there were 1000 simulated frequencies in each position. So those simulated values and the value of the frequency of NUMT-NUMT flank was compared, and P-values were calculated.

5.7 *The investigation of repeats in NUMT flanking regions*

We surveyed repeats in 500bp of upstream/downstream from NUMTs. Using position information, we referred the output of the RepeatMasker and the Tandem Repeats Finder; this output file, "rmsk" was downloaded from the UCSC website [50], and found the repeats.

5.8 *NUMTs annotation analysis*

The NUMT position information was crosschecked against annotation databases. However some databases contain the information of processed pseudogenes as well as the information of genes or functional elements. Therefore, we carefully evaluated the crosschecked results by eye. The following databases were used in this analysis. The URL was also added in the list.

5.4.1 Annotation databases

Database	URL	Ref.
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	[51]
UCSC Known Genes	http://genome.ucsc.edu/	[52]
Ensembl	http://uswest.ensembl.org/index.html	[53]
Vega Genome Browser	http://vega.sanger.ac.uk/index.html	[54]
CCDS Database	http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi	[55]
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/index.html	[56]
fRNAdb	http://www.ncrna.org/frnadb/index.html	[57]
miRBase	http://www.mirbase.org/	[58]

5.8 *The extraction of transcribed NUMTs with mouse 5′ CAGE data*

We mapped 5′ CAGE data, provided by Prof. Yutaka Suzuki in the University of Tokyo, to mouse genome (E-value<10^{-25}) with LAST which is BLAST-like alignment software [http://last.cbrc.jp/]. We only take the tags uniquely mapped to NUMT regions, and the tag appearance frequency (i.e. transcribed frequency) was counted.

Reference

[1] Ralph Bock, Jeremy N. Timmis, Reconstructing evolution: gene transfer from plastids to the nucleus, *BioEssays*, 30(6):556-566, 2008

[2] Dario Leister, Origin, Evolution and genetic effects of nuclear insertions of organelle DNA, *TRENDS in Genetics*, 21(12): 655-663, 2005

[3] Keith L. Adams, Jeffrey D. Palmer, Evolution of mitochondrial gene content: gene loss and transfer to the nucleus, *Molecular Phylogenetics and Evolution*, 29(3): 380-395, 2003

[4] Jeffrey L. Blanchard, Gregory W. Schmidt, Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns, *Molecular Biology and Evolution*, 13(2): 537-548, 1996

[5] Adrian Gherman, Peter E. Chen, Tanya M. Teslovich, Pawel Stankiewicz, Marjorie Withers, et. al., Population bottlenecks as a potential major shaping force of human genome architecture, *PLOS Genetics*, 3(7):1223-1231, 2007

[6] Sandra W. Clifton, Patrick Minx, Christiane M. R. Fauron, Michael Gibson, James O. Allen, et. al., Sequence and comparative analysis of the maize NB mitochondrial genome, *Plant Physiology*, 136: 3486-3503, 2004

[7] Markus Woischnik, Carls T. Moraes, Pattern of organization of human mitochondrial pseudogenes in the nuclear genome, *Genome Research*, 12:885-893, 2002

[8] Einat Hazkani-Covo, Shay Covo, Numt-mediated double-strand break repair mitigates deletions during primate genome evolution, *PLOS Genetics*. 4(10):1-12, 2008

[9] Tobias Mourier, Anders J. Hansen, Eske Willerslev, Peter Arctander, The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus, *Molecular Biology and Evolution*, 18(9):1833-1837, 2001

[10] Yves Tourmen, Olivier Baris, Philippe Dessen, Caroline Jacques, Yves Malthiery, et. al., Structure and chromosomal distribution of human mitochondrial pseudogenes, *Genomics*, 80(1): 71-77, 2002

[11] Einat Hazkani-Covo, Dan Graur, A comparative analysis of *numt* evolution in human and chimpanzee, M*olecular Biology and Evolution*, 24(1): 13-18, 2007

[12] Miria Ricchetti, Fredj Tekaia, Bernard Dujon, Continued colonization of the human genome by mitochondrial DNA, *PLOS Biology*, 2(9): 1313-1324 ,2004

[13] Keith L. Adams, Yin-Long Qiu, Mark Stoutemyer, Jeffrey D. Palmer, Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution, *Proceedings of the National Academy of Sciences of the United States of America*, 99(7): 9905-9912, 2002

[14] Christos Noutsos, Tatjana Keine, Ute Armbruster, Giovanni DalCorso, Dario Leister, Nuclear insertions of organellar DNA can create novel patches of functional exon sequences, *TRENDS in Genetics*, 23(12): 597-601, 2007

[15] Clesson Turner, Christina Killoran, Nick S. T. Thomas, Marjorie Rosenberg, Nadia A. Chuzhanova, et. al., Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer, *Human Genetics*, 112(3): 303-309, 2003

[16] Chihiro Rakamatsu, Shuyo Umeda, Takashi Ohsato, Tetsuji Ohno, Yoshito Abe, et. al., Regulation of mitochondrial D-loops by transcription factior A and single-stranded DNA-binding protein, *EMBO reports*, 3(5): 451-456, 2002

[17] Frederic Legros, Florence Malka, Paule Frachon, Anne Lombes, Mauel Rojo, Organization and dynamics of human mitochondrial DNA, *Journal of Cell Science*, 117:2653-2662, 2004

[18] Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., et. al., Sequence and organization of the human mitochondrial genome, *Nature*, 290(5806): 457-465, 1981

[19] Hiroshi Suzuki, Yoshitaka Hosokawa, Morimitsu Nishikimi, Takayuki Ozawa, Existence of common homologous elements in the transcriptional regulatory regions of human nuclear genes and mitochondrial gene for the oxidative phosphorylation system, *Journal of Biological Chemistry*, 266(4): 2333-2338, 1991

[20] David D. Chang, David A. Clayton, Priming of human mitochondrial DNA replication occurs at the light-strand promoter, *Proceedings of the National Academy of Sciences of the United States of America*, 82(2): 351-355, 1985

[21] Robert P. Fisher, James N. Topper, David A. Clayton, Promoter selection in human mitochondria involves binding of a transcription factor to orientation-independent upstream regulatory elements, *Cell*, 50(2): 247-258, 1987

[22] David D. Chang, David A. Clayton, Precise identification of individual promoter for transcription of each strand of human mitochondrial DNA, *Cell*, 36(3): 635-643, 1984

[23] Daniel F. Bogenhagen, Enid F. Applegate, Barbara K. Yoza, Identification of a promoter for transcription of the heavy strand of human mtDNA: in vitro transcription and deletion mutagenesis, *Cell*, 36(4): 1105-1113, 1894

[24] Ian J. Holt, Jiuya He, Chih-Chieh Mao, Jerome D. Boyd-Kirkup, Peter Martinsson, et. al., Mammalian mitochondrial nucleoids: Organizing an independently minded genome, *Mitochondrion*, 7:311-32, 2007

[25] Jiuya He, Chih-Chieh Mao, Aurelio Reyes, Hiroshi Sembongi, Miriam Di Re, et. al., The AAA^+ protein ATAD3 has displacement loop binding properties and is involved in mitochondrial nucleoid organization, *The Journal of Cell Biology*, 176(2): 141-146, 2007

[26] Martin Kucej, Ronald A. Butow, Evolutionary tinkering with mitochondrial nucleoids, *TRENDS in Cell Biology*, 17(12):586-592, 2007

[27] Dongchon Kang, Sang Ho Kim, Naotaka Hamasaki, Mitochondrial transcription factor A (TFAM): Roles in maintenance of mtDNA and cellular functions, *Mitochondrion*, 7: 39-44, 2007

[28] Tomotake Kanki, Kippei Ohgaki, Martina Gaspari, Claes M. Gustafsson, Atsushi Fukuoh, et. al., Architectural role of mitochondrial transcription factor A in maintenance of human mitochondrial DNA, *Molecular and Cellular Biology*, 24(22): 9823-9834, 2004

[29] Tanfis Istiaq Alam, Tomotake Kanki, Tsuyoshi Muta, Koutarou Ukaji, Yoshito Abe, et. al., Human mitochondrial DNA is packaged with TFAM, *Nucleic Acids Research*, 31(6):1640-1645, 2003

[30] Anne Helmrich, Karen Stout-Weider, Klaus Hermann, Evelin Schrock, Thomas Heiden, Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes, *Genome Research*, 16:1222-1230, 2006

[31] Thomas W. Glover, Martin F. Arlt, Anne M. Casper, Sandra G. Durkin, Mechanisms of common fragile site instability, *Human Molecular Genetics*, 14(2):197-205, 2005

[32] Eitan Zlotorynski, Ayelet Rahat, Jennifer Skaug, Neta Ben-Pprat, Efrat Ozeri, et. al., Molecular basis for expression of common and rare fragile sites, *Molecular and Cellular Biology*, 23(20): 7143-7151, 2003

[33] Michal Schwartz, Eitan Zlotorynski, Batsheva Kerem, The molecular basis of common and rare fragile sites, *Cancer Letters*, 232: 13-26, 2006

[34] T. Lukusa, J. P. Fryns, Human chromosome fragility, *Biochimica et Biophysica Acta*, 1779:3-16, 2008

[35] Andres Aguilera, Belen Gomez-Gonzalez, Genome instability: a mechanistic view of its causes and consequences, *Nature Reviews Genetics*, 9(3): 204-217, 2008

[36] Yoshimi Toda, Rintaro Saito, Masaru Tomita, Characteristic sequence pattern in the 5- to 20-bp upstream region of primate *Alu* elements, *The Journal of Molecular Evolution*, 50-232-237 ,2000

[37] Kostas Repanas, Nora Zingler, Liliana E. Layer, Gerald G. Schumann, Anastassis Perrakis, et. al., Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease, *Nucleic Acids Research*, 35(14):4914-4926, 2007

[38] Kenji Ichiyanagi, Hidenori Nishihara, David D. Duvernell, Norihiro Okada, Acquisition of endonuclease specificity during evolution of L1 retrotransposon, *Molecular Biology and Evolution*, 24(9): 2009-2015, 2007

[39] Stephen L. Gasior, Timothy P. Wakeman, Bo Xu, Prescott L. Deininger, The human LINE-1 retrotransposon creates DNA double-strand breaks, *The Journal of Molecular Biology*, 357: 1383-1393, 2006

[40] Jun Suzuki, Katsumi Yamaguchi, Masaki Kajikawa, Kenji Ichiyanagi, Noritaka Adachi, et. al., Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition, *PLOS Genetics*, 5(4): 1-13, 2009

[41] Shurjo K. Sen, Charles T. Huang, Kyudong Han, Mark A. Batzer, Endonuclease-independnet insertion provides an alternative pathway for L1 retrotransposition in the human genome, *Nucleic Acids Research*, 35(11): 3741-3751, 2007

[42] Deepa Srikanta, Shurjo K. Sen, Charles T. Huang, Erin M. Conlin, Ryan M. Rhodes, et. al., *Genomics*, 93:205-212, 2009

[43] Erik Richly, Dario Leister, NUMTs in sequenced eukaryotic genomes, *Molecular Biology and Evolution*, 21(6): 1081-1084, 2004

[44] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature*, 409(2): 860-921, 2001

[45] Arjun Raj, Alexander van Oudenaarden, Nature, nurthre, or chance: stochastic gene expresstion and its consequences, *Cell*, 135(2): 216-226, 2008

[46] Jeffrey A. Bailey, Zhiping Gu, Royden A. Clark, Knut Reinert, Rhea V. Samonte, et. al., Recent segmental duplications in the human genome, *Science*, 297(5583):1003-1007, 2002

[47] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et. al., Clustal W and Clustal X version 2.0, Bioinformatics, 23: 2947-2948, 2007

[48] Joseph Felsenstein, PYHLIP: Phylogeny Inference Package, Version 3.68, 2009, Department of Genetics, University of Washington, Seattle

[49] Terrence S. Furey, David Haussler, Integration of the cytogenetic map with the draft human genome sequence, *Human Molecular Genetics*, 12(9): 1037-1044, 2003

[50] Jerzy Jurka, Repbase Update: a database and an electronic journal of repetitive elements, *TRENDS in Genetics*, 16(9): 418-420, 2000

[51] Kim D. Pruitt, Tatiana Tatusova, Donna R. Maglott, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research*, 33: 501-504, 2005

[52] Fan Hsu, W. James Kent, Hiram Clawson, Robert M. Kuhn, Mark Diekhans, et. al., The UCSC Known Genes, *Bioinformatics*, 22(9): 1036-1046, 2006

[53] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, et. al., The Ensembl gneome database project, *Nucleic Acids Research*. 30(1): 38-41, 2002

[54] L. G. Wilming, J. G. R. Gilbert, K. Howe, S. Trevanion, T. Hubbard, et. al., The vertebrate genome annotation (Vega) database, *Nucleic Acids Research*, (10): 1-8, 2007

[55] Kim D. Pruitt, Jennifer Harrow, Rachel A. Harte, Craig Wallin, Mark Diekhans, et. al., The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes, *Genome Research*, [Epub ahead of print] Jun 4, 2009

[56] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, David L. Wheeler, GenBank: update, *Nucleic Acids Research*, 32:23-26, 2004

[57] Taishin Kin, Kouichirou Yamada, Goro Terai, Hiroaki Okada, Yasuhiko Yoshinari, et. al., fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences, *Nucleic Acids Research*, 35:145-148, 2007

[58] Sam Griffiths-Jones, Harpreet Kur Saini, Stijn van Dongen, Anton J. Enright, miRBase: tools for micro RNA genomics, *Nucleic Acids Research*, 36:154-158,2008

Supplementary information

Table S1: Human NUMT positions

Chromosome	Chr. start	Chr. end	Mt. start	Mt. end	Age	Annotation
chr1	107146786	107150029	6061	9317	a	---
chr1	113920927	113921018	13337	13428	a	intron
chr1	145799428	145799539	6871	6982	f	intron
chr1	167709925	167709972	8511	8558	f	intron
chr1	179658544	179658737	12144	12337	a	intron
chr1	190197613	190197655	1051	1093	c	intron
chr1	203711167	203711255	11154	11242	a	intron
chr1	210744964	210745037	2628	2701	a	---
chr1	213739762	213739800	9564	9602	f	---
chr1	218695123	218695220	9820	9917	f	---
chr1	233768514	233772288	9783	13594	a	intron
chr1	236170699	236181582	12219,1	16571,6190	a	---
chr1	37849935	37850008	8935	9008	f	exon
chr1	50255456	50255631	4714	4889	d	intron
chr1	5832905	5833095	2487	2677	e	---
chr1	77209497	77209688	16403,1	16571,23	a	intron
chr1	81319084	81319151	4981	5048	b	---
chr1	8892389	8892551	8041	8203	d	intron
chr1	94174550	94174744	2914	3108	a	---
chr1	9557323	9557462	968	1107	a	intron
chr10	101807134	101807648	8296	8810	c	intron
chr10	114644327	114644384	404	461	f	intron
chr10	121587532	121587634	3382	3483	a	intron
chr10	131819396	131819458	4362	4424	a	---
chr10	20075681	20076014	2419	2757	a	intron
chr10	20076201	20076451	3227	3483	a	intron
chr10	2267889	2267999	13019	13129	f	---
chr10	25967452	25967526	2608	2682	a	---
chr10	27202196	27202340	2517	2661	a	---
chr10	28205683	28205744	1554	1615	a	intron
chr10	30894207	30894256	7165	7214	e	---
chr10	36762181	36764084	11670	13578	a	---
chr10	37930127	37931761	1847	3483	a	---

chr10	57027643	57030440	638	3107	a	intron
chr10	71020912	71025687	3822	7699	a	---
chr10	80840991	80841055	5809	5872	c	intron
chr10	91536416	91536469	12563	12616	b	---
chr11	102778067	102786628	1021	9666	a	intron
chr11	10486010	10488403	579	2974	e	exon
chr11	110252926	110253062	15476	15612	c	---
chr11	11218183	11218297	13242	13356	d	---
chr11	122379524	122379595	14661	14732	f	---
chr11	122515984	122516161	15476	15653	b	intron
chr11	31533232	31533316	16304	16389	e	intron
chr11	39745009	39745200	7451	7642	a	---
chr11	47302111	47302308	13256	13453	a	intron
chr11	63711383	63711521	16436,1	16571,2	a	intron
chr11	6479715	6479777	2921	2983	b	intron
chr11	72899354	72899516	6643	6805	f	intron
chr11	80940264	80945683	9821	15244	a	---
chr11	87202142	87204752	14686,1	16571,1107	a	---
chr12	125634809	125634891	4382	4464	a	---
chr12	129366104	129366299	12143	12338	a	---
chr12	22050037	22050137	1728	1828	d	---
chr12	31291864	31292988	14422	15244	a	---
chr12	38966489	38966724	6772	7007	a	intron
chr12	40043704	40043792	3792	3880	f	intron
chr12	40379271	40379477	4247	4453	a	---
chr12	48497408	48497552	4279	4424	a	intron
chr12	61454057	61454124	4242	4309	a	intron
chr13	106509180	106509221	11078	11119	c	---
chr13	108874473	108874728	984	1239	f	---
chr13	21585352	21585454	10085	10187	d	intron
chr13	35537628	35537833	4211	4416	a	intron
chr13	40240488	40240558	9524	9594	f	intron
chr13	47043142	47043185	2175	2218	a	---
chr13	53363732	53363843	13022	13133	b	---
chr13	55443769	55443891	5109	5231	e	---
chr13	56160611	56160784	2900	3073	b	---
chr13	75064376	75064533	9883	10040	a	intron
chr13	83993021	83995318	11535	13416	a	---
chr13	95142796	95146649	13052	16524	a	intron
chr13	96147944	96148085	7477	7618	b	intron
chr14	22965536	22965580	2013	2057	e	intron
chr14	32023055	32024075	5584	6607	e	intron
chr14	51123889	51124214	598	922	a	intron
chr14	83707449	83713093	1837	15326	a	---
chr15	33475765	33475841	3830	3906	d	intron
chr15	38213263	38214685	11707	13135	a	---
chr15	39236690	39236787	11664	11761	d	---
chr15	44420893	44421031	12993	13131	b	---

chr15	56229853	56235023	9787	15319	a	intron
chr15	65120303	65120345	2176	2218	a	intron
chr16	10720552	10726472	8711	15319	a	---
chr16	20639871	20641224	12221	13575	a	---
chr16	3357809	3362068	2470	7354	a	---
chr16	67950099	67950214	12223	12338	a	intron
chr16	80711971	80713031	14592	15653	a	---
chr17	19442485	19449425	598	5980	a	---
chr17	21942648	21955772	14366,1	16571,10914	b	---
chr17	21955968	21956219	14328	14580	c	---
chr17	24325492	24325570	13123	13201	e	intron
chr17	31006003	31006178	2420	2596	a	intron
chr17	39430610	39430677	10144	10211	f	intron
chr17	48538093	48538745	6819	7471	f	---
chr17	58824504	58824678	9396	9570	b	intron
chr17	76205977	76206017	6904	6944	f	intron
chr18	2832230	2832352	14382	14504	f	---
chr18	43633615	43633806	7976	8167	f	intron
chr18	57692784	57693098	965	1283	a	intron
chr19	12067916	12068109	4256	4449	a	intron
chr19	12479785	12479910	9149	9274	c	intron
chr19	32924557	32924623	834	904	d	intron
chr19	42755388	42755506	12219	12339	a	intron
chr19	49342421	49342502	3745	3826	b	intron
chr19	62125407	62125601	12144	12338	a	---
chr2	117226231	117226415	16446,1	16571,60	a	---
chr2	117495259	117500547	598	5893	a	---
chr2	120685762	120690928	9197	13575	a	---
chr2	125154971	125155049	12258	12336	b	intron
chr2	126286241	126286454	651	865	a	---
chr2	132899143	132899200	4149	4206	a	intron
chr2	140691291	140698242	598	5892	a	---
chr2	143566320	143574013	9167,1	16571,60	a	intron
chr2	147739231	147739310	2771	2850	d	---
chr2	149355765	149355896	613	744	f	intron
chr2	155828223	155829560	4861	6197	b	---
chr2	155875844	155879111	11802	15068	a	---
chr2	166979251	166979469	6592	6824	c	intron
chr2	175354012	175354070	9803	9861	c	---
chr2	180312319	180312534	1014	1230	a	intron
chr2	201785264	201787949	10441	13132	a	intron
chr2	202130784	202130905	13298	13419	c	intron
chr2	203187200	203191742	6967	11241	a	intron
chr2	203192214	203193002	5770	6554	a	intron
chr2	212346765	212349578	598	3107	a	intron
chr2	212350179	212352647	4855	7354	a	intron
chr2	220621821	220621953	6337	6470	a	---
chr2	22393436	22393595	15264	15423	a	intron

chr2	227295229	227295386	892	1050	c	---
chr2	236029399	236029551	799	950	b	---
chr2	33846042	33846094	1768	1820	f	intron
chr2	40865601	40865761	958	1126	d	---
chr2	49310271	49310542	6742	7013	f	---
chr2	50669334	50670032	6296	6995	c	intron
chr2	68341376	68341540	5023	5187	a	intron
chr2	75144383	75144488	3000	3105	b	intron
chr2	81747119	81747360	7863	8104	e	---
chr2	82896241	82900557	12221	16527	a	---
chr2	82901528	82901640	598	710	b	---
chr2	85149463	85149664	6857	7058	a	---
chr2	87905524	87905999	8315	8790	f	intron
chr20	13095959	13096001	3501	3543	e	---
chr20	55072517	55072586	12963	13032	c	---
chr20	55366111	55369449	651	4039	a	intron
chr20	9097571	9097612	2182	2223	f	intron
chr21	44719405	44719483	12262	12341	a	---
chr21	45620549	45620727	5996	6174	d	---
chr22	31620948	31621151	14834	15037	c	intron
chr22	34611665	34611711	6182	6228	d	intron
chr22	34905387	34905426	4265	4304	d	intron
chr22	45244827	45244894	12644	12711	a	intron
chr22	48866724	48866905	673	854	a	---
chr3	108095676	108098627	9788	12341	a	intron
chr3	108100197	108101514	641	13537	a	intron
chr3	108102431	108102512	6089	6170	c	intron
chr3	108103491	108103688	4253	4450	a	intron
chr3	108104060	108105060	1056	2050	a	intron
chr3	121923570	121924164	7162	7756	b	intron
chr3	123890275	123890332	4706	4763	d	intron
chr3	128207009	128207098	6747	6836	e	intron
chr3	154120187	154120313	5689	5811	b	---
chr3	154859364	154859466	7456	7558	a	---
chr3	162148136	162148429	1927	2220	c	intron
chr3	167360887	167362489	9167	10769	a	---
chr3	167432964	167433283	1059	1375	a	---
chr3	171137279	171137350	16406	16477	d	intron
chr3	172734924	172735131	6724	6932	e	---
chr3	178166808	178166907	4236	4339	a	---
chr3	188666690	188666733	9206	9249	b	---
chr3	25483999	25484037	10986	11024	f	intron
chr3	29814192	29814362	10429	10599	d	intron
chr3	40268642	40270262	1420	3041	a	intron
chr3	43245910	43246227	15810	16125	c	---
chr3	63807749	63807810	15567	15628	d	intron
chr3	68790897	68791128	3060	4430	d	---
chr3	72715141	72715370	9138	9367	a	---

chr3	89718693	89721366	6605	9317	b	---
chr3	97818722	97820044	1398	2720	e	---
chr3	97966674	97966818	6536	6680	c	---
chr4	117438367	117440910	598	3162	a	---
chr4	12251016	12251357	9339	9680	f	---
chr4	129222010	129222381	2010	2390	c	intron
chr4	14116628	14117171	4645	5196	b	intron
chr4	156592474	156607061	674	15328	a	---
chr4	161185212	161185273	13022	13083	c	---
chr4	163561976	163562143	12251	12418	d	---
chr4	164369641	164369751	5760	5866	c	---
chr4	16672606	16672885	13107	13386	e	---
chr4	180227588	180227736	7626	7774	a	---
chr4	182395550	182395687	7195	7332	d	---
chr4	25328634	25331437	9782	12302	a	---
chr4	27341144	27341295	2901	3052	c	---
chr4	30495717	30495899	15144	15326	a	intron
chr4	47469046	47469138	14982	15074	f	intron
chr4	5457275	5457345	16404	16474	e	intron
chr4	55889084	55889214	964	1094	f	---
chr4	65154700	65154900	9177	9377	b	---
chr4	65155336	65160181	9486	16563	b	---
chr4	66613657	66613769	12229	12341	a	---
chr4	69120816	69120912	7651	7747	c	intron
chr4	79148708	79148939	2227	2458	e	intron
chr4	82873861	82874025	15117	15284	a	---
chr4	88816286	88816371	3659	3744	d	---
chr4	90872030	90872180	2918	3068	b	intron
chr4	93842002	93842704	2795	3496	b	intron
chr5	105916963	105917122	14671	14830	d	---
chr5	118490993	118491042	2005	2054	e	intron
chr5	120394581	120394911	385	710	d	---
chr5	123123872	123123905	205	238	f	---
chr5	123124398	123125331	579	1512	b	---
chr5	134286898	134292116	10270	15488	f	exon/intron
chr5	162369094	162369203	812	920	b	---
chr5	165890005	165890044	12148	12187	e	---
chr5	169995143	169995205	816	878	d	intron
chr5	5448535	5448947	6701	7112	b	---
chr5	5449435	5450074	6962	7600	a	---
chr5	60093123	60093608	3823	4309	b	intron
chr5	73107473	73107513	10803	10843	d	intron
chr5	79981597	79983943	343	2699	e	intron
chr5	8672214	8674124	14672	16564	a	---
chr5	8675212	8675685	609	1089	a	---
chr5	93928917	93932379	12663	16125	d	intron
chr5	97773394	97774943	6361	7897	a	---
chr5	99409541	99418648	6118	15184	d	---

chr6	119522203	119522285	5956	6038	a	---
chr6	127674976	127675031	6793	6848	f	intron
chr6	133513403	133513626	15396	15619	b	---
chr6	143440639	143440770	9869	10000	c	intron
chr6	145091766	145091970	1418	1619	a	intron
chr6	154028400	154032608	7452	11650	a	---
chr6	156910662	156910879	3164	3381	d	---
chr6	1651297	1651398	8506	8607	b	intron
chr6	43566883	43566943	16409	16469	d	intron
chr6	51959736	51959823	13278	13365	f	intron
chr6	62341977	62342493	2420	2936	b	---
chr6	74991919	74991978	3033	3092	c	intron
chr6	75585644	75585690	400	446	d	---
chr6	89321743	89321855	8717	8829	a	---
chr6	92493232	92493962	5522,10475,6533	8775,10623,6746	d	---
chr6	95213557	95213881	4151	4475	a	---
chr7	110512693	110512772	5653	5732	e	intron
chr7	111799937	111802234	13066	15370	a	---
chr7	116691253	116691664	11078	11493	a	intron
chr7	140519373	140519482	12563	12672	d	intron
chr7	141147677	141151647	2895	6554	a	---
chr7	145325359	145325454	1615	1710	f	---
chr7	36234877	36234986	4779	4888	e	intron
chr7	45258095	45258251	2420	2576	d	---
chr7	57257414	57270157	3820,1	16571,39	a	---
chr7	63201998	63210482	3118	11881	a	intron
chr7	66730406	66730469	9170	9233	f	---
chr7	67200158	67200283	7182	7307	a	---
chr7	67369188	67369316	11364	11492	a	---
chr7	67839451	67839556	12962	13067	f	---
chr7	68433640	68436831	5609	8247	a	---
chr8	100577274	100577357	14862	14945	f	intron
chr8	104164459	104171823	1015	7115	a	---
chr8	112014736	112016352	5512	7109	a	---
chr8	112330062	112330210	15077	15226	a	---
chr8	121305733	121305830	6913	7010	e	intron
chr8	13255277	13255401	11065	11189	c	intron
chr8	134836878	134838023	5512	6659	a	---
chr8	15958887	15958955	14432	14500	f	---
chr8	18751092	18751468	14834	15210	a	intron
chr8	20452987	20453197	8667	8877	c	---
chr8	27816796	27816909	14770	14883	a	intron
chr8	32782949	32783027	12258	12337	b	---
chr8	32988565	32992739	638	4889	a	intron
chr8	36254678	36256619	14075	16017	a	---
chr8	40047266	40047363	808	907	d	---
chr8	47858273	47861837	658	4881	a	---
chr8	47867623	47869417	14823,1	16571,35	a	---

chr8	49475808	49475874	12990	13056	c	---
chr8	53826866	53827070	15176	15380	a	---
chr8	68655593	68662552	9177,1	16571,60	a	intron
chr8	70177762	70177868	5506	5613	b	intron
chr8	74060498	74060643	639	784	c	---
chr8	77276553	77276929	2288	2663	d	---
chr8	77720250	77720459	1524	1734	c	intron
chr8	98989484	98989559	8807	8882	c	intron
chr9	131680960	131681016	2772	2828	e	intron
chr9	18326655	18326731	2466	2542	c	---
chr9	18816076	18816126	6605	6655	b	intron
chr9	34989141	34989291	16330	16481	b	---
chr9	5082095	5100699	1296	13575	a	intron
chr9	79769920	79770080	14193	14353	b	intron
chr9	80545747	80548229	598	6215	a	---
chr9	80615638	80615690	3830	3882	d	---
chr9	82368404	82370509	4766	6873	a	---
chr9	93911111	93913772	9203	11599	a	intron
chr9	94341172	94341704	5513	6017	b	intron
chrX	101914775	101951469	3221,1	16571,1369	a	intron
chrX	110546882	110547045	2987	3763	d	---
chrX	125433368	125434948	689,10607	7303,11160	e	---
chrX	125690712	125692598	14686,1	16571,2	a	---
chrX	125693450	125693513	642	705	b	---
chrX	142345841	142349570	1056	4416	a	intron
chrX	142459310	142459354	11076	11120	d	intron
chrX	15655775	15655953	6502	6680	c	---
chrX	26752972	26753046	1559	1633	c	---
chrX	5096881	5098240	14075	15430	a	---
chrX	55081711	55081796	2615	2701	a	---
chrX	55221910	55227180	583	5893	a	---
chrX	61976282	61978565	1051	3162	a	---
chrX	69257463	69264188	12137	15337	a	---
chrX	83545594	83545658	14886	14950	e	intron
chrX	9733688	9733774	12990	13076	b	intron
chrY	8291998	8292566	14672	15244	a	---
chrY	8294669	8300289	598	4478	a	---

Table S2: Mouse NUMT positions

Chromosome	Chr. Start	Chr. End	Mt. start	Mt. end	Age	Annotation	Transcribed frequency			
							Day7	Day11	Day15	Day17
chr1	112649910	112649990	11600	11680	j	---	0	0	0	0
chr1	128612173	128612236	6853	6917	h	intron	0	0	0	2
chr1	134687483	134687935	15601	16054	i	---	2	0	0	7
chr1	144626542	144626587	10749	10794	k	---	0	0	0	0
chr1	181320368	181320592	708	933	g	intron	4691	1249	335	3609
chr1	189536479	189536545	3721	3787	h	intron	0	0	0	0
chr1	190802216	190802302	4097	4183	k	intron	556	227	56	481
chr1	24618374	24623023	6394	11042	h	intron	167909	62356	21654	184525
chr1	34809671	34809974	11445	11749	j	intron	18	10	0	19
chr1	63718054	63718096	14784	14826	g	intron	80	28	0	81
chr1	65467456	65468539	7764	8847	g	---	1	1	1	6
chr10	41405591	41405860	13266	13535	k	---	26	3	0	2
chr10	52238100	52238281	6123	6304	i	---	1	0	0	0
chr10	59625112	59625169	12746	12803	k	---	1	0	0	14
chr10	84488698	84489385	15321	16006	h	---	108	87	47	167
chr10	95540159	95540528	1512	1881	g	---	615	141	80	386
chr11	29212303	29212374	2243	2314	g	---	0	0	0	0
chr11	44197716	44200761	7548	7867	g	---	16	5	4	50
chr11	90399796	90400451	6216	6871	i	intron	260	24	3	230
chr12	114172539	114172702	15605	15768	j	---	6	0	0	0
chr12	4077303	4077354	4409	4460	k	---	27	35	20	38
chr12	4077359	4077627	4719	4989	g	---	18	0	0	42
chr12	60363664	60364083	3397	3816	g	---	39817	10940	2463	28084
chr12	67493169	67493280	15773	15883	j	---	0	0	0	0
chr12	81909609	81909847	3402	3643	j	intron	3870	1518	212	3763
chr12	98299638	98300002	15675	16039	i	---	95	13	10	22
chr13	60410744	60410916	6295	6466	g	---	0	0	0	33
chr13	83603584	83603676	13131	13223	k	---	0	0	0	3
chr13	85266167	85267120	12446	13400	h	---	3268	1209	316	5875
chr14	112109370	112109802	9009	9440	h	---	0	0	0	0
chr14	37948908	37949037	236	365	g	---	31	13	12	2
chr14	59618591	59618641	10462	10512	k	---	0	0	0	1
chr15	15385337	15385380	4440	4483	h	---	0	0	0	0
chr15	57319223	57323499	15687,1	162993914	j	---	6	0	0	2
chr16	15885594	15885712	1602	1725	j	---	0	0	0	0
chr16	85022146	85022263	10458	10575	j	intron	2	0	0	0
chr17	31429865	31430899	14651	15498	g	---	6	1	1	4
chr17	34704679	34705075	2932	3334	h	intron	343	188	12	189
chr17	94080029	94080355	9112	9440	k	---	0	0	0	0
chr18	22531621	22531765	11356	11500	k	intron	1127	523	15	1521
chr18	3638873	3639147	12074	12348	h	---	228	53	86	356
chr18	50394154	50394227	7913	7986	k	---	0	0	0	0
chr18	57474580	57474781	10963	11164	j	---	0	0	0	0
chr18	65279258	65279653	15424	15821	h	intron	34	33	68	193

chr18	65996941	65997124	8290	8473	h	---	95	27	13	27
chr18	75164512	75165220	9062	9694	h	---	2	11	0	4
chr2	124512220	124512312	6942	7032	i	---	0	0	0	0
chr2	14235217	14235366	15356	15503	i	intron	14	4	18	9
chr2	22442808	22446054	4441	7699	h	intron	28948	9169	3396	24649
chr2	69193595	69193723	11514	11642	i	---	126	57	140	265
chr3	108879644	108879905	16298,1	16299261	i	---	1509	873	37	2742
chr3	114888053	114888102	2937	2986	k	---	0	14	0	2
chr3	144345341	144345570	15645	15876	j	---	0	0	0	0
chr3	6049169	6049487	15932	16248	k	---	52	0	23	22
chr3	6049607	6050197	15329	15921	h	---	1	0	4	2
chr3	94138497	94138608	8456	8567	i	---	10	2	3	31
chr3	99003134	99003317	20	191	j	intron	55	31	21	98
chr4	141532843	141532890	10316	10363	k	intron	0	0	0	0
chr4	44219240	44219321	10858	10939	k	exon	0	0	0	0
chr4	79648234	79651104	12488	15356	k	---	48824	15597	7493	44349
chr4	91169889	91169945	6045	6102	i	---	0	0	0	0
chr4	9885849	9887097	3279	4514	j	---	25095	11012	1401	31758
chr5	113185588	113186172	6045	6642	i	intron	4	1	0	5
chr5	151244644	151244803	10885	11044	i	intron	0	0	0	0
chr5	35976811	35977104	15652	15946	j	---	9	0	0	0
chr5	60433994	60434338	4887	5231	j	---	2	1	0	1
chr5	7276324	7277300	14739	15714	h	intron	279	97	66	418
chr5	85047360	85048841	4955	6440	g	---	2	0	1	7
chr6	114701468	114701573	8527	8632	g	intron	0	0	0	0
chr6	127713778	127713952	6501	6675	j	intron	0	0	0	0
chr6	67711046	67711119	9183	9256	g	intron	0	0	0	0
chr6	79767910	79768341	1738	2166	i	---	424	128	23	431
chr6	9839893	9840202	15607	15916	j	---	87	60	98	147
chr7	118941027	118941097	2377	2447	h	---	25	0	0	15
chr7	83029159	83029230	3703	3775	i	---	0	0	0	0
chr7	95761372	95761523	11923	12074	g	---	0	0	0	1
chr7	99109870	99110030	9717	9876	h	intron	0	0	0	0
chr7	99250648	99251175	3695	4122	g	intron	1	0	0	18
chr8	104547848	104548601	15138	15927	g	---	5	14	0	1
chr8	12277013	12277171	8928	9081	g	---	0	0	0	0
chr8	31075891	31075920	11313	11342	k	intron	239	89	1	215
chr8	32508524	32508875	14925	15279	g	---	33	21	3	3
chr8	40309134	40309209	5499	5574	g	intron	3	0	0	3
chr8	95786459	95786500	1609	1650	k	intron	0	0	0	0
chr9	123016270	123016635	4934	5299	h	intron	0	0	0	0
chr9	15123583	15126521	1967	5824	j	exon/intron	44	52	29	147
chr9	24388416	24388470	7750	7804	j	intron	2	0	0	5
chr9	24549532	24549704	2511	2683	g	intron	180	49	41	315
chr9	82361246	82361375	236	365	g	---	0	0	0	0
chr9	93067783	93071178	6175	9463	h	---	285	85	19	231
chrX	100193382	100193609	14723	14942	g	---	2	0	0	3
chrX	60568295	60568467	14033	14211	h	---	1910	399	464	1841