G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(3):90-95
http://dx.doi.org/10.5808/GI.2016.14.3.90

ORIGINAL ARTICLE

Analysis of Nuclear Mitochondrial DNA Segments of Nine Plant Species: Size, Distribution, and Insertion Loci

Young-Joon Ko, Sangsoo Kim*

Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Korea

Nuclear mitochondrial DNA segment (Numt) insertion describes a well-known phenomenon of mitochondrial DNA transfer into a eukaryotic nuclear genome. However, it has not been well understood, especially in plants. Numt insertion patterns vary from species to species in different kingdoms. In this study, the patterns were surveyed in nine plant species, and we found some tip-offs. First, when the mitochondrial genome size is relatively large, the portion of the longer Numt is also larger than the short one. Second, the whole genome duplication event increases the ratio of the shorter Numt portion in the size distribution. Third, Numt insertions are enriched in exon regions. This analysis may be helpful for understanding plant evolution.

Keywords: DNA transfering, nuclear mitochondrial DNA, numt, plant mitochondrial DNA

Introduction

From the beginning of endosymbiosis between the origin eukaryote cell and alphaproteobacteria, the phenomenon of mitochondrial gene transfer to the host cell is still an ongoing evolutionary process [1, 2]. It is termed nuclear mitochondrial DNA (Numt), pronounced "new might" [3]. In general, the mutation rate of nuclear DNA is lower than that of the mitochondrial genome. For this reason, Numt is often called a molecular fossil and is used as a molecular marker for speciation events in evolution [4, 5]. While Numt insertion is a well-known phenomenon, the mechanism of DNA insertion into the nuclear genome is not clear. One of the strongly supported hypotheses is that during the process of double-strand break repair, an absorbed mitochondrial DNA fragment is inserted into the nuclear genome via a non-homologous end joining event [6-9].

After whole-genome sequencing was finished, Numt analysis was performed in various species: cat, cattle, dog, fruit fly, gorilla, grasshopper, goose, horse, horseshoe bat, honeybee, human, maize, squirrel, and whale [10-17]. One review paper summarized all existences of Numt in complete genome sequences [18]. In the case of whale, a phylogenetic analysis of Numts with six whale species was carried out, establishing Numt as an evolutionary marker in speciation events [19]. In plant, a recent study discovered that Numt insertion is dispersed throughout the periphery of the centromere [20]. But, there are many barriers in plant Numt analyses. Genomic complexity is a big problem in not only the nuclear genome but also the mitochondrial genome [21-23].

In this article, Numts of two green algae (*Chlamydomonas reinhardtii* and *Coccomyxa subellipsoidea*), three monocots (*Oryza sativa, Sorghum bicolor*, and *Zea mays*), and four eudicots (*Vitis vinifera, Glycine max, Brassica rapa*, and *Arabidopsis thaliana*), for which whole-genome nuclear and mitochondrial sequences are publically available, were detected using the nucleotide-nucleotide Basic Local Alignment Search Tool (BLASTN) searches and subjected to a basic analysis for their fundamental properties, which will be required in further comparative genome analyses in plants.

Received August 2, 2016; Revised August 16, 2016; Accepted August 29, 2016

***Corresponding author:** Tel: +82-2-820-0457, Fax: +82-2-820-0816, E-mail: sskimb@ssu.ac.kr

Methods

Data sources

We downloaded all of the genomic data and gene annotation data (gff3) of two green algae genomes (*C. reinhardtii* and *C. subellipsoidea*) from the Joint Genome Institute, four plant genomes (*O. sativa, Z. mays, V. vinifera*, and *A. thaliana*) from the Ensembl genome database, two plant genomes (*G. max* and *S. bicolor*) from the Plant Genome Database, and the *B. rapa* genome and annotation information data, available online from the *B. rapa* Database. We also collected each mitochondrial genome sequence from the National Center for Biotechnology Information (NCBI). All these data sources are summarized in Table 1.

Detection of Numts and data generation

Plant Numt insertions were identified using BLASTN local alignment tools in the BLAST program package (ver. 2.2.26), with mitochondrial genomic DNA as a query sequence and each genome dataset as a BLAST database. The execution options included an e-value cutoff set to 0.01, filtering switched off (-dust no), a mismatch penalty of −2, and a word size of 9. The neighboring Numt hits within 10 kb were, if necessary, merged into a single event of Numt insertions. All of these analytical processes were carried out with in-house Python codes.

Calculation of odd ratio for Numt insertion loci

To calculate the relative abundance of each genomic feature,

Table 1. Sources of genomic sequences

Taxa	DB	Data file name	Account No. of mitochondrial DNA
Chlamydomonas reinhardtii	http://genome.jgi.doe.gov/	Chlamydomonas_reinhardtii.v3.1.31.dna.genome.fa	NC_001638
Coccomyxa subellipsoidea	http://genome.jgi.doe.gov/	CsubellipsoideaC_169_227_v2.0.softmasked.fa	NC_015316
Oryza sativa	Ensemblgenomes.org	Oryza_sativa.IRGSP-1.0.31.dna.genome.fa	DQ_167400
Sorghum bicolor	plantgdb.org	SBgenome	DQ_984518
Zea mays	Ensemblegenomes.org	Zea_mays.AGPv3.31.dna.genome.fa	NC_007982
Vitis vinifera	Ensemblegenomes.org	Vitis_vinifera.IGGP_12x.31.dna.genome.fa	NC_012119
Glycine max	Plantgdb.org	Gmax_109.fa	NC_020455
Brassica rapa	Brassicadb.org	Brapa_sequence_v1.5.fa	NC_016125
Arabidopsis thaliana	Ensemblegenomes.org	Arabidopsis_thaliana.TAIR10.31.dna.genome.fa	Y08501

Table 2. Number of Numt hits and their sizes

Taxa	Genome size (Mb)	Mitochondrial genome size (kb)	No. of hits	After merging (overlapped)	After merging (within 10 kb)	Maximum length	Minimum length	Total length of Numt (kb)
Chlamydomonas reinhardtii	109	15.8	64	55	49	333	26	3.3
Coccomyxa subellipsoidea	48	65.5	1,003	644	510	4070	25	51.5
Oryza sativa	373	491	6,549	2,878	1,620	40,410	28	980
Sorghum bicolor	659	469	4,333	3,676	3,094	4,166	29	956
Zea mays	2,059	570	10,782	7,485	5,050	106,610	30	2,241
Vitis vinifera	426	773	288,200	14,509	9,022	5,888	29	1,603
Glycine max	950	403	3,105	2,277	1,611	7,430	29	439
Brassica rapa	257	220	1,883	1,531	1,104	1,408	27	128
Arabidopsis thaliana	119	367	1,293	770	562	40,130	27	376
Minke whale	2,440	16	530	-	144	7,771	30	291
Bowhead whale	2,300	16	494	-	136	8,990	34	317
Sperm whale	2,280	16	647	-	218	7,680	31	378
Yangtze river dolphin	2,530	16	829	-	253	6,552	31	471
Killer whale	2,370	16	677	-	170	13,310	30	365
Bottlenose dolphin	2,550	16	1,108	-	549	7,716	31	432

Numt, nuclear mitochondrial DNA segment.

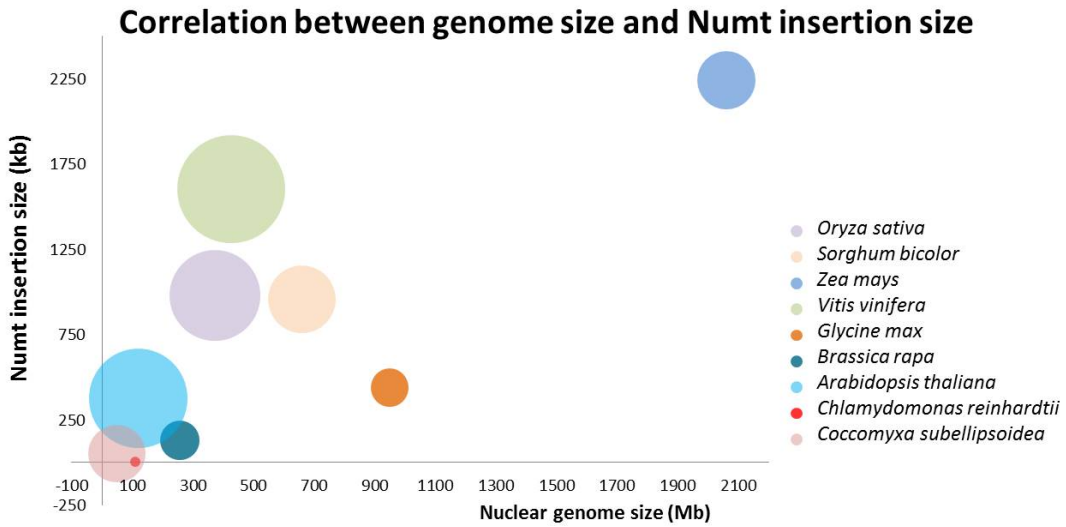

Fig. 1. Nuclear genome size and total length of nuclear mitochondrial DNA (Numts).

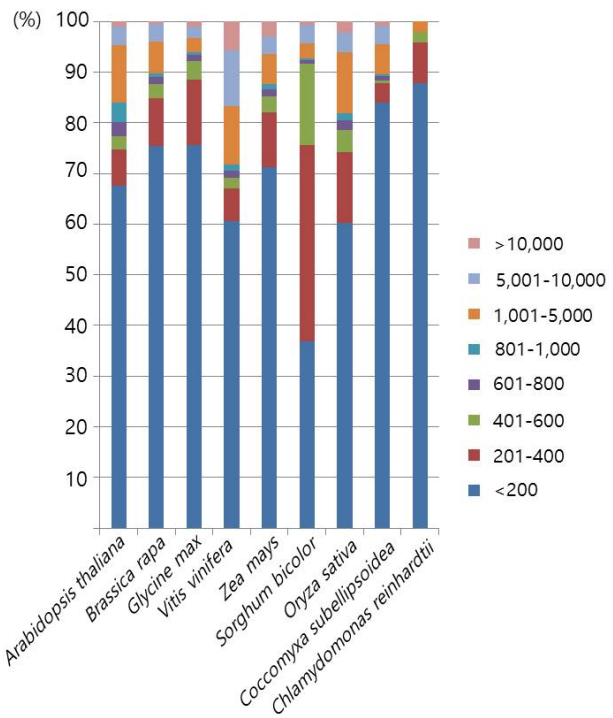

Fig. 2. Nuclear mitochondrial DNA (Numt) size distribution chart by plant species.

we gathered the length information of the categories, such as gene, coding sequence (CDS), exon, pseudogene, and noncoding RNA (ncRNA) (tRNA, rRNA, and long noncoding RNA [lncRNA]), from the gene annotation files of each species (gff3 format). The total length of exons and introns included only the protein-coding genes. The total length of introns was computed by subtracting the total exon length from the sum of all gene lengths. With all of the length information, we estimated the portion of each feature by dividing the total length of each feature (S_i) by each whole-genome length (G). The relative abundance (RA_i) of each feature was then calculated as follows:

$$RA_i = C_i / (S_i/G)$$

, where C_i is the count of the genic feature *i* in a species.

Results and Discussion

All of the data are summarized in Table 2. The genome size of the nine plants species varied from 48 Mb to 2 Gb. The mitochondria genome size also varied from 15.8 kb to 773 kb. There was no correlation between whole-genome size and mitochondrial whole-genome size. We drew a correlation chart between whole nuclear genome length and the sum of the inserted Numt lengths (Fig. 1). The larger the genome size, the more nuclear mitochondrial insertions there were. This confirms a previous study result [18]. The added green algae species also showed this tendency. One of the peculiarities of plant species is their many Numt hits. Except for green algae (*C. reinhardtii* and *C. subellipsoidea*), the number of BLAST hits after merging all overlapping hits ranged from 770 for *A. thaliana* to 14,509 for *V. vinifera*. Furthermore, when integrating all of the neighboring hits within 10 kb into one single event, the hit count ranged from 562 in *A. thaliana* to 9,022 in *V. vinifera*. This implies that the transposition of mitochondrial DNA of plants into chromosomal DNA is more preferable than in whale species [19].

Next, we examined the size distribution of the inserted Numts. Here, we merged the neighboring hits within 10 kb

into single events. The merged hits showed a high degree of variation in size—the shortest and largest being 25 bp and 107 kb, respectively (Table 2). The size distribution of Numt was also quite variable between species (Fig. 2). Over 70% of Numts were less than 400 bp in all of the analyzed plants. Green algae species that had shorter mitochondrial DNA than other species had over 80% in the group with less than 200 bp, especially in *C. reinhardtii* (over 96%). *V. vinifera*, which has a larger mitochondrial genome size than other plants, included 30% of Numts over 1 kb in size, and half of this group was over 5 kb. *Z. mays*, which has the largest genome and the second largest mitochondrial genome, and *B. rapa*, which has a relatively shorter genome than *Z. mays*, showed similar ratio distributions. In general, species having short mitochondrial genomes had a large ratio of short Numts. When comparing monocots and eudicots, there was no clearly shared feature. But, there were some differences when contrasting green algae and land plants. However, it is not a matter of the species group but rather a matter of genomic size variation. There are two kinds of closely related speciation events: one is between *A. thaliana* and *B. rapa*, and the other is between *S. bicolor* and *Z. mays*. In each of the speciation events, there were whole-genome triplication or duplication events, leading to *B. rapa* and *Z. mays* [24, 25]. Because of that, each pair has a similar genomic content, but the within-pair Numt size distribution patterns are different. In general, *B. rapa* and *Z. mays* have lower ratios of long sizes of Numts than *A. thaliana* and *S. bicolor*, respectively. Genome triplication or duplication events may have split the long Numt sequences, such that the number of long Numts was reduced. These patterns were also observed in the speciation between *G. max* and *V. vinifera*.

In the previous whale Numt study [19], they also performed a Numt size distribution analysis. The average whale genome size is 2.5 Gb, and the average mitochondrial genome size is 16 kb. It has a much larger nuclear genome and smaller mitochondrial genome. In whales, the Numt size group over 5 kb is under 2%, but in plants, it is over 4%, and in *V. vinifera*, it is over 17%. It is presumed that as a result of a 20-fold larger mitochondrial genome, even if going through the second whole-genome duplication event, there are longer Numt sequences that still reside in the plant genome.

The next analysis was the classification of Numt insertion loci by genic features (Table 3). In land plants, a substantial portion of Numt hits lay in intergenic regions, except for green algae, where over 70% of the hits were found within genic boundaries. Within genic regions, over 90% of the hits overlapped exons. This is in contrast with the Numt hits in animals, like whales, where the total number of Numt hits was quite low and in which fewer hits were found in exons than in introns [19]. When we calculated the relative abundance of each genic feature after accounting for the total size of each genic feature, the exon was the most enriched in most plants (Fig. 3). Considering the importance of exons in biological processes, it may be tempting to speculate that the numerous Numt insertions into exons may affect the diversity of plant phenotypes.

Table 3. Numt counts by genic features

Taxa	Exon[a]	Intron[a]	Pseudogene	ncRNA	Intergenic	Total
Chlamydomonas reinhardtii	39	5	0	0	11	55
Coccomyxa subellipsoidea	442	69	0	0	133	644
Oryza sativa	454	0	0	288	2,136	2,878
Sorghum bicolor	172	33	0	0	3,471	3,676
Zea mays	608	0	0	0	6,877	7,485
Vitis vinifera	3,206	0	0	0	11,303	14,509
Glycine max	334	51	0	0	1,892	2,277
Brassica rapa	141	18	0	0	1,372	1,531
Arabidopsis thaliana	335	0	12	4	419	770
Plant subtotal	5,731	176	12	292	27,614	33,825
Minke whale	3	36	1	2	102	144
Bowhead whale	12	16	54	0	54	136
Sperm whale	7	28	5	22	156	218
Yangtze river dolphin	4	59	5	0	182	253
Killer whale	3	36	8	0	122	170
Bottlenose dolphin	4	10	6	0	529	549
Whale subtotal	33	185	79	24	1,145	1,470

Numt, nuclear mitochondrial DNA segment; ncRNA, noncoding RNA.
[a]Protein-coding genes.

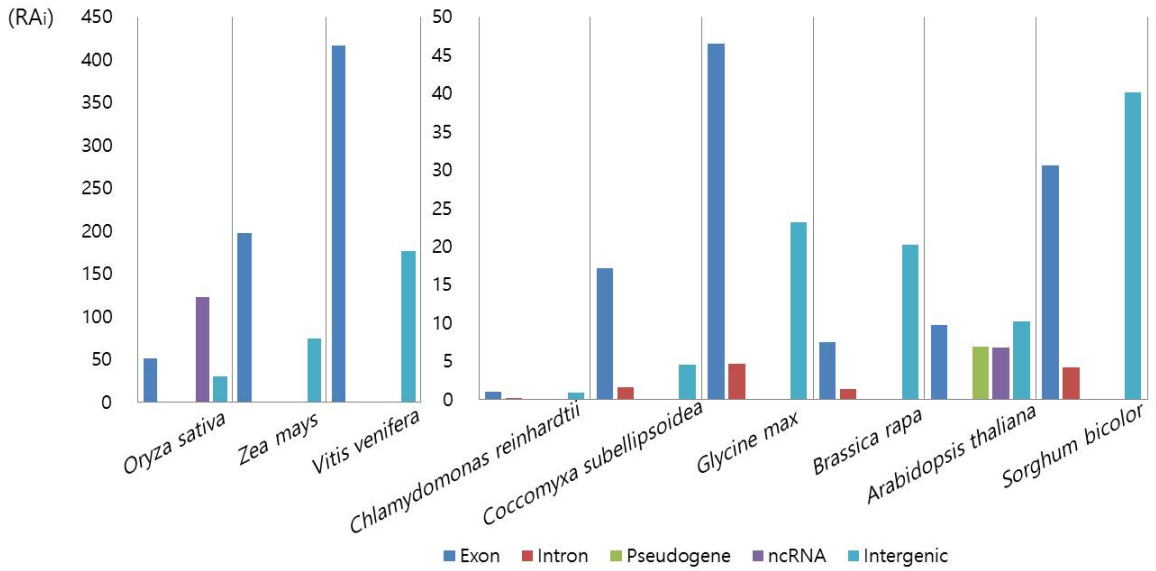

Fig. 3. Genic features of Numt-inserted positions. The Y-axis represents relative abundance of each gene feature (see Methods for definition). Numt, nuclear mitochondrial DNA; ncRNA, noncoding RNA.

Many research studies on Numt analysis have been performed. But, they usually lack details on Numts, such as the correlation between genome size and inserted Numt size, Numt size distribution ratio, loci classification by gene annotation, and so on. Our general basic analysis shows an interesting tendency but is still not enough to infer the biological meaning. Currently, not many plant genomes have been completely sequenced, and furthermore, their accuracy is somewhat compromised due to high repeat contents or high heterozygosity in the genomes. In order to draw a clearer picture of the effect of Numt insertion in the nuclear genome, more population-level genomic data and more accurate genome sequences may be required. Nevertheless, Numts may be one of the key clues of the mysterious biological implications of genomic analysis.

Acknowledgments

This work supported by a program (PJ01167402) from the RDA (Rural Development Administration) and a program (NRF-2012M3A9D1054705) from the NRF (National Research Foundation of Korea).

References

1. Kleine T, Maier UG, Leister D. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol* 2009;60:115-138.
2. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 2004;5:123-135.
3. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol* 1994;39:174-190.
4. Hazkani-Covo E. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. *Mol Biol Evol* 2009;26:2175-2179.
5. Zhang DX, Hewitt GM. Nuclear integrations: challenges for mitochondrial DNA markers. *Trends Ecol Evol* 1996;11:247-251.
6. Ricchetti M, Fairhead C, Dujon B. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 1999;402:96-100.
7. Blanchard JL, Schmidt GW. Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Mol Biol Evol* 1996;13:893.
8. Thorsness PE, Fox TD. Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*. *Nature* 1990;346:376-379.
9. Hazkani-Covo E, Covo S. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet* 2008;4:e1000237.
10. Hassanin A, Bonillo C, Nguyen BX, Cruaud C. Comparisons between mitochondrial genomes of domestic goat (*Capra hircus*) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA* 2010;21:68-76.
11. Behura SK. Analysis of nuclear copies of mitochondrial sequences in honeybee (Apis mellifera) genome. *Mol Biol Evol* 2007;24:1492-1505.
12. Liu Y, Zhao X. Distribution of nuclear mitochondrial DNA in cattle nuclear genome. *J Anim Breed Genet* 2007;124:264-268.
13. Verscheure S, Backeljau T, Desmyter S. *In silico* discovery of a nearly complete mitochondrial genome Numt in the dog

(*Canis lupus familiaris*) nuclear genome. *Genetica* 2015;143: 453-458.
14. Soto-Calderón ID, Clark NJ, Wildschutte JV, DiMattio K, Jensen-Seaman MI, Anthony NM. Identification of species-specific nuclear insertions of mitochondrial DNA (numts) in gorillas and their potential as population genetic markers. *Mol Phylogenet Evol* 2014;81:61-70.
15. Rogers HH, Griffiths-Jones S. Mitochondrial pseudogenes in the nuclear genomes of Drosophila. *PLoS One* 2012;7:e32593.
16. Lough AN, Roark LM, Kato A, Ream TS, Lamb JC, Birchler JA, *et al*. Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize. *Genetics* 2008; 178:47-55.
17. Bensasson D, Zhang DX, Hewitt GM. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Mol Biol Evol* 2000;17:406-415.
18. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* 2010;6:e1000834.
19. Ko YJ, Yang EC, Lee JH, Lee KW, Jeong JY, Park K, *et al*. Characterization of cetacean Numt and its application into cetacean phylogeny. *Genes Genomics* 2015;37:1061-1071.
20. Michalovova M, Vyskot B, Kejnovsky E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity (Edinb)* 2013;111:314-320.
21. Cupp JD, Nielsen BL. Minireview: DNA replication in plant mitochondria. *Mitochondrion* 2014;19 Pt B:231-237.
22. Chang S, Wang Y, Lu J, Gai J, Li J, Chu P, *et al*. The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. *PLoS One* 2013;8:e56502.
23. Michael TP. Plant genome size variation: bloating and purging DNA. *Brief Funct Genomics* 2014;13:308-317.
24. Lysak MA, Koch MA, Pecinka A, Schubert I. Chromosome triplication found across the tribe Brassiceae. *Genome Res* 2005;15:516-525.
25. Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, *et al*. Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* 2007;3:e123.