Quantifying the Number of Independent Organelle DNA Insertions in Genome Evolution and Human Health

Einat Hazkani-Covo[1,*] and William F. Martin[2]

[1]Department of Natural and Life Sciences, The Open University of Israel, Ra'anana, Israel

[2]Institute of Molecular Evolution, Heinrich-Heine University, Düsseldorf, Germany

*Corresponding author: E-mail: einatco@openu.ac.il.

Accepted: April 20, 2017

Abstract

Fragments of organelle genomes are often found as insertions in nuclear DNA. These fragments of mitochondrial DNA (numts) and plastid DNA (nupts) are ubiquitous components of eukaryotic genomes. They are, however, often edited out during the genome assembly process, leading to systematic underestimation of their frequency. Numts and nupts, once inserted, can become further fragmented through subsequent insertion of mobile elements or other recombinational events that disrupt the continuity of the inserted sequence relative to the genuine organelle DNA copy. Because numts and nupts are typically identified through sequence comparison tools such as BLAST, disruption of insertions into smaller fragments can lead to systematic overestimation of numt and nupt frequencies. Accurate identification of numts and nupts is important, however, both for better understanding of their role during evolution, and for monitoring their increasingly evident role in human disease. Human populations are polymorphic for 141 numt loci, five numts are causal to genetic disease, and cancer genomic studies are revealing an abundance of numts associated with tumor progression. Here, we report investigation of salient parameters involved in obtaining accurate estimates of numt and nupt numbers in genome sequence data. Numts and nupts from 44 sequenced eukaryotic genomes reveal lineage-specific differences in the number, relative age and frequency of insertional events as well as lineage-specific dynamics of their postinsertional fragmentation. Our findings outline the main technical parameters influencing accurate identification and frequency estimation of numts in genomic studies pertinent to both evolution and human health.

Key words: numts, nupts, organelle insertions, mitochondria, cancer genomics.

Introduction

Organelle DNA in Eukaryotic Chromosomes

Mitochondria and chloroplasts were once free-living proteobacteria and cyanobacteria, but their genomes are highly reduced in comparison to their free-living bacterial relatives (Allen 2015). Plastid genomes encode 23–200 proteins and mitochondrial DNA encodes anywhere from 3 to 60 proteins (Timmis et al. 2004). Both bioenergetic organelles contain roughly 2,000 proteins, however, that are required in order for the organelle to function, the vast majority of which are encoded in the nucleus (Gray 1992; Paul et al. 2013). The discrepancy between the number of proteins that organelles encode and the number of proteins that they contain is explained by a corollary to endosymbiotic theory called endosymbiotic gene transfer (EGT), which posits that during the course of evolution, most genes present in the genome of the ancestral organelle were relinquished to the nucleus where the corresponding gene products could be translated by cytosolic ribosomes and undergo one of five fates. Either 1) the protein became targeted back to the organelle of its origin to allow loss of the organelle gene, 2) the protein became targeted to other compartments within the cell, 3) the genes underwent recombination and accumulated mutations so as to encode or impart new functions, 4) the gene became a pseudogene, or 5) any combination of the above (Martin and Herrmann 1998).

Gene transfer from organelles to the nucleus has been investigated in experimental studies in yeast (Thorsness and Fox 1990; Ricchetti et al. 1999) and in plants (Huang et al. 2003), the latter revealing that the rate of transfer is of the order of one successful event among 250,000 gametes tested. In evolution, gene transfers from organelles have impacted on eukaryotic genomes in two main ways. Early in eukaryote evolution, at the endosymbiotic acquisition of

chloroplasts and mitochondria, massive influx of organelle genes into eukaryotic genomes not only permitted the transition from endosymbiont to organelle, it also served as a source of genetic starting material in nuclear chromosomes for the origin of genes and proteins that are specific to the eukaryotic lineage. At the level of whole sequence analysis, eukaryotic genomes harbor no evidence for evolutionarily more recent or continuous gene acquisitions from bacteria (Ku et al. 2015; Ku and Martin 2016), further underscoring the importance of gene acquisition at organelle origin. In recent evolutionary history, EGT is a widespread and ongoing process that that still occurs across all eukaryotes harboring organelle DNA. This is evidenced by the presence, in nuclear genome sequences, of nuclear copies of organelle DNA. Such recently transferred nuclear copies are witnesses to the gene transfer process, they are known as nuclear sequences of mitochondrial origin (numts) and nuclear sequences of plastid origin (nupts).

Numts and nupts are found in virtually all sequenced eukaryotic genomes (Bensasson et al. 2001; Richly and Leister 2004a, 2004b). The number of insertions and total amount of numts and nupts varies considerably among genomes and is correlated with the size of nuclear genome (Hazkani-Covo et al. 2010). Identifying numts and nupts in eukaryotic genome assemblies is more challenging than it might seem at first sight. In particular, the mere counting of independent numt and nupt insertions in genomes is not an easy task.

Numts and nupts are very pesky insertions. They are ubiquitous in nuclear *DNA*, but they are *not* ubiquitous in nuclear genome *assemblies*. Very often they are removed (edited out) during the genome annotation process (Pesole et al. 2012). As genomes become increasingly important in evolutionary studies and in modern medical applications—be it in diagnostics (Sturm et al. 2016) or in monitoring the progression of cancers (Wang and Wheeler 2014)—it is becoming increasingly important that organelle insertions are properly accounted for in nuclear genome assemblies. Not only do numts and nupts plague studies of molecular systematics (Thalmann et al. 2004; Lorenz et al. 2005); different DNA sequencing and assembly methods deliver genome sequences that harbor very different numbers of numts and nupts for the same organism (Hazkani-Covo and Graur 2007; Dayama et al. 2014). A number of numts are associated with human disease (Chen et al. 2005) and some specific numt insertions have even been linked to lifespan (Cheng and Ivessa 2012). With the increased number of reports revealing the significance of numts in human health and cancer progression studies, it is important that tools are available to ensure that numts and nupts are properly identified and annotated in genome sequences.

Dynamics in Two Genomes Simultaneously

Numts and nupts that occur in genomes today are the result of DNA dynamics both within and between organelles and the nucleus. Mitochondria and plastid genomes vary significantly in size, gene content, and shape (Gould et al. 2008; Gray 2012). One parameter that can significantly influence the number of detected transfers is the amount of repetitive DNA in the organelle. If an organelle DNA includes repeats, and if repeated regions are transferred into the nucleus, then the same nuclear region will be detected multiple times if organelle DNA is BLASTed to the nuclear assembly. The inverted repeat of plastid DNA, which can be up to ten thousand bases long (Kolodner and Tewari 1979) is a classic example of this effect, but other examples of shorter repeats appear, such as the *Arabidopsis thaliana* mitochondrial genome, where they account for ~7% of the 367 kb genome (Unseld et al. 1997). A complete nuclear copy of the 367 kb *Arabidopsis* mitochondrial genome in length was not present in the original sequence of *Arabidopsis* chromosome 2 (Lin et al. 1999); it was edited out during annotation (Stupar et al. 2001), illustrating the difficulties that numts and nupts present in nuclear genome assembly.

Numts and nupts studied so far show no obvious tendency to trace to one particular region of the organelle genome; it appears that all regions of organelle DNA have the same propensity to become integrated into nuclear chromosomes (Tourmen et al. 2002; Woischnik and Moraes 2002; Bensasson et al. 2003; Hazkani-Covo et al. 2003; Mishmar et al. 2004). Older numts were reported to be shorter (Bensasson et al. 2003), which generally suggests that they tend to be fragmented after insertion, not during insertion. This is consistent with the report that very recent large continuous insertions correspond to the sequence of the entire organelle genome (Huang et al. 2005). However, experimental studies have shown there are also cases of insertion in which organelle DNA is rearranged prior to or during the insertion event (Ricchetti et al. 1999; Huang et al. 2003). Thus, during insertion, the inserted DNA can undergo duplications or assume a gene arrangement that is nonlinear with respect to the gene organization in the organelle. Nonlinear organization includes both the polarities of regions adjacent in the organelle as well as the presence of fragments from different regions of the organelle. Within the nuclear genome, recent EGT events are mostly considered dead on arrival pseudogenes (Bensasson et al. 2001), though some exceptions are known (Noutsos et al. 2007). As such, mitochondrial and plastid fragments inserted into the nuclear genome undergo fragmentation (Huang et al. 2005), and translocations as well as tandem duplications.

Examples of complex numt and nupt architecture have been reported. An insertion of three chloroplast fragments at a single nuclear location was demonstrated experimentally (Lloyd and Timmis 2011) by screening nuclear activation of a chloroplast reporter gene *aadA* in plant lines. Ricchetti et al. (1999) reported three events of concatenated numts among the 30 numts in *Saccharomyces cerevisiae*. The nuclear cat genome includes 7.9 kb that is duplicated in tandem 38–76 times (Lopez et al. 1994) and human chromosome 12 includes 18 tandem repeated numts of 47 bp length (Hazkani-Covo

and Graur 2007). Sheppard and Timmis (2009) demonstrated instability of insertion, deletion, and rearrangement of plastid sequences in the nuclear genome. Similar insertion, deletion, and recombination dynamics have been reported in numerous other studies (Matsuo et al. 2005; Noutsos et al. 2005; Michalovova et al. 2013). Additionally, compound loci carrying numts and nupts have also been reported in plants (Noutsos et al. 2005). Such recombinational events have a major impact on efforts to identify organelle DNA insertions via sequence comparison methods, motivating our present efforts to find ways to more effectively quantify EGT events.

Factors Affecting Numt and Nupt Counts

The principle of identifying numts and nupts is simple: one uses BLAST (or similar) comparison of a sequenced organelle DNA to the nuclear genome of the same species. This is the commonly used procedure, it delivers a rough quantitative estimate (Bensasson et al. 2003; Hazkani-Covo et al. 2003; Richly and Leister 2004a; Hazkani-Covo et al. 2010). But staying in this approximate BLAST comfort zone can be misleading. If we examine individual genomes carefully, obtaining good estimates for the number and nature of numts/nupts becomes more complicated. This is because the number of BLAST hits only provides an estimate for the upper bound of the number of recent independent insertions in the nuclear sequence data, for three reasons: 1) organelle genome organization and organelle repeated sequences, 2) structural rearrangements of organelle DNA that can occur during insertion, and 3) the subsequent rearrangement and mutational fate of insertions within the nucleus. During the sequencing process, where genome sequences are evolving from the stage of contigs and scaffolds to carefully annotated and curated assemblies, accurate estimation of the number of numt/nupt insertions in the genome (instead of only the number of BLAST hits) is particularly important.

Surprisingly (or perhaps not), counting of BLAST hits is still a common approach both in single genome and in large scale analysis (Richly and Leister 2004a; Hazkani-Covo et al. 2010; Du and Qin 2015; Sun and Yang 2016), even though methods to better estimate the number of evolutionary independent events in genomes have been reported. These mostly pertain to single genome studies. Some of the factors that can result in overestimation of the number of inferred insertions are discussed in individual genome reports, mostly involving the combination of collinear fragments. The DNA Block Aligner (DBA), which was used in an early human genome numt analysis (Mourier et al. 2001), is one example. A set of rules to join closely related BLAST hits was applied in later studies (Woischnik and Moraes 2002; Pamilo et al. 2007; Hazkani-Covo 2009; Tsuji et al. 2012; Michalovova et al. 2013). The distance used to define two BLAST hits as originating from the same event varies from study to study. The use of closely related nuclear genomes can improve the concatenation and border delineation of orthologous numts (Hazkani-Covo and Graur 2007).

In contrast, events such as duplications and insertion that are not collinear with the organelle DNA are rarely accounted for. Examples include Tourmen et al. (2002), where human hits that are 30 kb apart are concatenated, Rogers and Griffiths-Jones (2012), who manually inspected *Drosophila* numts within 25 kb windows, and Michalovova et al. (2013), who estimated numts and nupts in six plants using concatenation of fragments with mixed orientations from nearby organelle regions. Duplications of numts are also considered in a few cases (Hazkani-Covo and Graur 2007; Tsuji et al. 2012). The screening out of repetitive organelle regions that cause multiple BLAST hits is rarely, if ever, reported even though organelle DNA can harbor repeated sequences. Clearly, there are a number of factors to be considered when combining BLAST hits, yet there is no overview of the parameters used or of their impact when analyzing real genome data. With the number of sequenced genomes steadily increasing, and with the use of genome sequences steadily expanding in the study of human health, such comparison is needed.

Materials and Methods

Data

Complete mitochondrial, plastid and nuclear genomes of 44 eukaryotes were downloaded (for complete list with accession numbers see supplementary table S3, Supplementary Material online). This is a subset of the 55 eukaryotic nuclear genomes analyzed by Ku et al. (2015) representing the cases where in addition to the nuclear genome at least one organelle genome is available. Out of the 44 genomes we were able to analyze 36 nuclear genomes for the existence of numts and 24 nuclear genomes for the existence of nupts.

Counting the Number of BLAST Hits

Each organelle genome was compared with the corresponding nuclear genome using BLAST+ (Camacho et al. 2009) with the blastn search, expect (*E*) value 0.0001, and with the DUST filter for masking of low complexity regions from nucleic acid sequences. The raw number of BLAST hits and the sum of hits were reported for each eukaryotic organism.

Inferring of the Number of Numts and Nupts

Analysis was done in a combination of PERL and MySQL. First, BLAST hits that mapped without a gap in the corresponding nuclear coordinates were concatenated using the BioPerl function disconnected_ranges of BIO::Range. This function finds the minimal set of ranges such that each input range is fully contained by at least one output range, and none of the output ranges overlap. These BLAST hits were concatenated regardless of organelle localization and orientation. The

first step takes care of repeated regions on organelle genomes, tandem duplications, as well as DNA segments that were inserted nonlinearly. Following this first step, concatenations were performed separately by three different stringency protocols. The strict protocol concatenation is only allowed if organelle and nuclear DNA are colinear in a given distance. In the intermediate stringency the distance is only enforced on the nuclear but not on the organellar genome. The permissive estimate gives minimal numbers of independent insertions. In this case nonlinear organelle hits and hits on both nuclear strands were allowed. Thus, BLAST hits were concatenated regardless of the position and orientation of hits on organelles while only distances between nuclear coordinates were scored.

Results

The Method for Counting Numts and Nupts

There are currently no conventions or established protocols for estimating the numbers of independent insertions through the joining of fragmented organelle insertions. Here, we investigated three approaches involving stringencies of joining criteria (strict to permissive) to infer the number of independent insertions from raw BLAST hits in 44 fully sequenced genomes. The genomes span six eukaryotic supergroups: Archaeplastida, Hacrobia, SAR, Excavata, Opisthokonta, and Amoebozoa. Numts were investigated in 36 genomes, 24 genomes were analyzed for nupts. Our approach took the following factors into account: 1) repetitive areas in organelles causing the identification of a nuclear area multiple times, 2) tandem duplication within the nuclear genome, and 3) fragmented collinear organelle bits. An additional factor 4) fragmented noncollinear organelle bits, is treated differently at the different stringencies.

We first ran BLAST of the organelle DNA (mitochondria or plastid) against the nuclear genome with an arbitrary but sensible *E*-score threshold of 10^{-4}. We then calculated the number of independent events by three concatenation approaches. 1) The strict strategy holds on to nuclear and organelle coordinates as well as to their orientation. 2) The intermediate approach holds on to nuclear coordinates and to nuclear and organelle orientation but not to organelle coordinates. 3) The permissive counting holds on to only the nuclear coordinates. We used the following distances between nuclear coordinates in aligned matches for each strategy to combine nearby nuclear BLAST hits: 50 bp, 500 bp, 3 kb, and 10 kb.

In all three approaches, we first merged BLAST hits on the nuclear genome if they overlap or if the distance between BLAST hits is zero. In the strict approach, we then joined fragmented BLAST hits if they are collinear and separated by less than the specified distance (50 bp, 500 bp, etc.) on both the organelle and the nuclear genome.

In the intermediate approach we allowed concatenation of organelle BLAST hits regardless of their distance on the organelle as long as they are in the right orientation. In the third and most permissive approach, the minimal number of insertions is inferred by concatenating BLAST hits based only on the nuclear distance, ignoring mitochondrial (or plastid) order and strand orientation. This permissive approach lumps nearby numts (and nupts) as stemming from the same insertion and gives the lowest estimation, or minimal number, of recent EGT events.

How Many Were Found?

BLAST hits were concatenated for distances of 50 bp, 500 bp, 3 kb, and 10 kb. Table 1 and supplementary table S1, Supplementary Material online (for numts) and table 2 and supplementary table S2, Supplementary Material online (for nupts) show the number of BLAST hits, the coverage of BLAST hits as well as the unique coverage of BLAST hits in the nuclear genomes. In addition, the tables present the number of inferred numts/nupts for the three methods across distances (50 bp to 10 kb). For each inferred number of numts/nupts, the number of numts that are composed of more than one BLAST hit under the specific parameter used is shown. For example, for *Arabidopsis thaliana* 1390 BLAST hits were found that cover 620,480 nuclear bases of which 554,892 bp are unique nuclear bases. If we consider concatenation in the distance of 50 bp, 915 events are inferred using the collinear approach, among which 96 are composed of multiple BLAST hits. As expected, the number of numts and nupts dropped, compared with the number of BLAST hits, for these genomes for all three methods. As shown in figure 1, even when the strictest concatenation was performed, concatenating only hits that are 50 bp apart and collinear on the nuclear and the organelle genomes generates a steep decline (of the order of 34%, 915 instead of 1,390 events) in the estimated number of numts.

Figure 1 summarizes the results for genomes with at least 400 BLAST hits. There is a consistent drop from the number of BLAST hits to the number of inferred numts/nupts but within each concatenating distance the differences in the number of final numts/nupts is moderate. If we continue with the previous example, BLASTing the nuclear genome of *Arabidopsis thaliana* with the corresponding mitochondrial DNA yields 1390 BLAST hits. Concatenating these numts results in anywhere from 626 to 915 (45–66%) inferred numts, depending both on the distance used for concatenation and on the method used to concatenate. Using a concatenation distance of 50 bp yields 915, 883, and 845 inferred numts for the strict, intermediate and permissive criteria, respectively. A distance of 500 bp yields 845, 821, and 773 inferred numts for the strict, intermediate, and permissive criteria, whereas a distance of 3 kb and 10 kb yields 813, 744, 681 and 792, 715, and 626. However, the fraction of inferred independent events compared with the number of BLAST hits changes dramatically across genomes, especially when considering

Table 1

Number of Inferred Numts (Complex Numts in Parentheses) in Different Genomes as Determined by **S**trict, **I**ntermediate, and **P**ermissive Concatenation Criteria and for Different Distances (50 bp up to 10 kb) between Organelle DNA Fragments in Nuclear DNA

	#BLAST Hits	UBB	Distance	S	I	P
Arabidopsis thaliana	1390	620480/554892	50 bp	915 (96)	883 (122)	845 (115)
			500 bp	845(154)	821 (171)	773 (165)
			3 kb	813(181)	744 (217)	681(216)
			10 kb	792(193)	715(234)	626(237)
Chondrus crispus	58	8877/8525	50 bp	49(8)	48(9)	44(12)
			500 bp	48(9)	47(10)	43(13)
			3 kb	47(9)	46(10)	42(13)
			10 kb	46(10)	45(11)	40(14)
Cyanidioschyzon merolae	3	330/330	50 bp	3(0)	3(0)	3(0)
			500 bp	3(0)	3(0)	3(0)
			3 kb	3(0)	3(0)	3(0)
			10 kb	3(0)	3(0)	3(0)
Cyanophora paradoxa	178	39441/31990	50 bp	139(30)	138(31)	138(31)
			500 bp	138(31)	136(31)	135(32)
			3 kb	136(30)	135(31)	134(32)
			10 kb	134(32)	134(32)	133(33)
Chlamydomonas reinhardtii	52	4295/3898	50 bp	45(6)	45(6)	44(7)
			500 bp	45(6)	45(6)	43(8)
			3 kb	45(6)	45(6)	43(8)
			10 kb	45(6)	45(6)	43(8)
Coccomyxa subellipsoidea	829	66090/55833	50 bp	597(75)	590(80)	569(93)
			500 bp	591(77)	578(85)	551(100)
			3 kb	589(78)	571(89)	524(111)
			10 kb	576 (82)	556 (97)	487(130)
Galdieria sulphuraria	7	402/344	50 bp	4(1)	4(1)	4(1)
			500 bp	4(1)	4(1)	4(1)
			3 kb	4(1)	4(1)	4(1)
			10 kb	4(1)	4(1)	4(1)
Micromonas sp. RCC299	18	2298/1143	50 bp	9(9)	9(9)	9(9)
			500 bp	6(6)	6(6)	6(6)
			3 kb	6(6)	6(6)	6(6)
			10 kb	3(3)	3(3)	3(3)
Oryza sativa	6550	1720939/1185113	50 bp	3095 (1538)	2984 (1499)	2660 (1325)
			500 bp	2900(1470)	2768(1431)	2249(1130)
			3 kb	2820 (1468)	2661 (1420)	2072 (1071)
			10 kb	2772 (1476)	2572 (1409)	1914 (1026)
Physcomitrella patens	449	116749/114219	50 bp	405(31)	392(41)	380(48)
			500 bp	393(40)	378(49)	360(57)
			3 kb	376(48)	366(54)	348(62)
			10 kb	364(54)	357(59)	338(68)

NOTE.—The number of hits obtained by BLASTing mitochondrial DNA against the nuclear genome (BLASTN *E* value 0.001) is shown. UBB (Unique Bases by BLAST)—values in the column UBB indicate the sum of bases (before slash) or the unique bases (after slash), respectively, by BLAST. Unless otherwise indicated, the values shown in red (permissive concatenation with a maximum distance of 500 bp) were used in this study. Additional genomes appear in supplementary table S1, Supplementary Material online.

genomes where BLAST hits are few and split across the entire genome. For example, the *Dictyostelium discoideum* genome includes 147 BLAST hits assigned to 122–132 numts, which correspond to 83–90% of the number of BLAST hits. It is essential to concatenate nearby numt/nupt BLAST hits to obtain more accurate estimates, although the impact of each approach used to correct for concatenation or the distance between fragments is only moderate. Below we present the details of analysis with concatenation that are no more than 500 bp apart in the nuclear genome and any distance or orientation in the organelle genome.

Out of the 44 genome analyzed (supplementary table S3, Supplementary Material online), 16 were analyzed for both numts and nupts: *Arabidopsis thaliana*, *Chondrus crispus*, *Cyanidioschyzon merolae*, *Cyanophora paradoxa*, *Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea*, *Galdieria sulphuraria*, *Micromonas sp. RCC299*, *Oryza sativa*, *Physcomitrella patens*, *Emiliania huxleyi*, *Plasmodium*

Table 2

Number of Inferred Nupts (Complex Nupts in Parentheses) in Different Genomes as Determined by **S**trict, **I**ntermediate, and **P**ermissive Concatenation Criteria and for Different Distances (50 bp up to 10 kb) Between Organelle DNA Fragments in Nuclear DNA

	#BLAST hits	UBB	Distance	S	I	P
Arabidopsis thaliana	442	62144/47921	50 bp	302(107)	291(110)	287(110)
			500 bp	282(114)	278(114)	270(114)
			3 kb	271(120)	267(120)	257(118)
			10 kb	257(115)	251(115)	238(111)
Chondrus crispus	48	62028/60350	50 bp	40(4)	40(4)	40(4)
			500 bp	38(6)	38(6)	38(6)
			3 kb	34(6)	34(6)	34(6)
			10 kb	32(6)	32(6)	32(6)
Cyanidioschyzon merolae	19	3622/3604	50 bp	19(0)	19(0)	19(0)
			500 bp	16(3)	16(3)	16(3)
			3 kb	13(6)	13(6)	13(6)
			10 kb	10(3)	10(3)	10(3)
Cyanophora paradoxa	374	144842/122167	50 bp	260(68)	259(68)	257(69)
			500 bp	256(65)	256(65)	254(66)
			3 kb	256(65)	255(66)	250(68)
			10 kb	256(65)	255(66)	250(68)
Chlamydomonas reinhardtii	143	8798/4304	50 bp	63(38)	63(38)	63(38)
			500 bp	56(31)	56(31)	56(31)
			3 kb	50(25)	50(25)	50(25)
			10 kb	48(25)	48(25)	48(25)
Coccomyxa subellipsoidea	3289	326917/108834	50 bp	363(42)	360(44)	345(53)
			500 bp	358(47)	356(48)	338(59)
			3 kb	353(48)	346(51)	324(62)
			10 kb	347(51)	334(59)	306(71)
Chlorella variabilis	29	1829/1811	50 bp	29(0)	29(0)	28(1)
			500 bp	29(0)	29(0)	28(1)
			3 kb	29(0)	29(0)	28(1)
			10 kb	27(1)	26(2)	25(3)
Galdieria sulphuraria	168	190296/176916	50 bp	69(35)	68(34)	64(31)
			500 bp	53(35)	53(35)	47(30)
			3 kb	38(27)	38(27)	30(21)
			10 kb	38(27)	38(27)	30(21)
Micromonas sp. RCC299	22	3775/2586	50 bp	13(9)	13(9)	13(9)
			500 bp	12(10)	12(10)	12(10)
			3 kb	6(4)	6(4)	6(4)
			10 kb	6(4)	6(4)	6(4)
Oryza sativa	2850	1331500/988435	50 bp	1746(666)	1722(674)	1642(651)
			500 bp	1662(673)	1630(681)	1483(638)
			3 kb	1627(674)	1579(681)	1387(619)
			10 kb	1595(684)	1544(691)	1313(619)
Physcomitrella patens	305	58953/48812	50 bp	238(65)	235(66)	233(65)
			500 bp	235(67)	232(67)	229(66)
			3 kb	228(60)	222(57)	219(56)
			10 kb	218(51)	216(51)	212(51)
Porphyridium purpureum	353	290772/215564	50 bp	24(18)	23(17)	23(17)
			500 bp	24(18)	23(17)	23(17)
			3 kb	23(17)	23(17)	23(17)
			10 kb	23(17)	23(17)	23(17)
Selaginella moellendorffii	860	314782/253383	50 bp	505(324)	504(325)	499(328)
			500 bp	492(322)	492(322)	484 (324)
			3 kb	373(207)	372(206)	361(209)
			10 kb	343(183)	343(183)	330(185)

NOTE.—The number of hits obtained by BLASTing mitochondrial DNA against the nuclear genome (BLASTN *E* value 0.001) is shown. UBB (Unique Bases by BLAST)—values in the column UBB indicate the sum of bases (before slash) or the unique bases (after slash), respectively, by BLAST. Unless otherwise indicated, the values shown in red (permissive concatenation with a maximum distance of 500 bp) were used in this study. Additional genomes appear in supplementary table S2, Supplementary Material online.

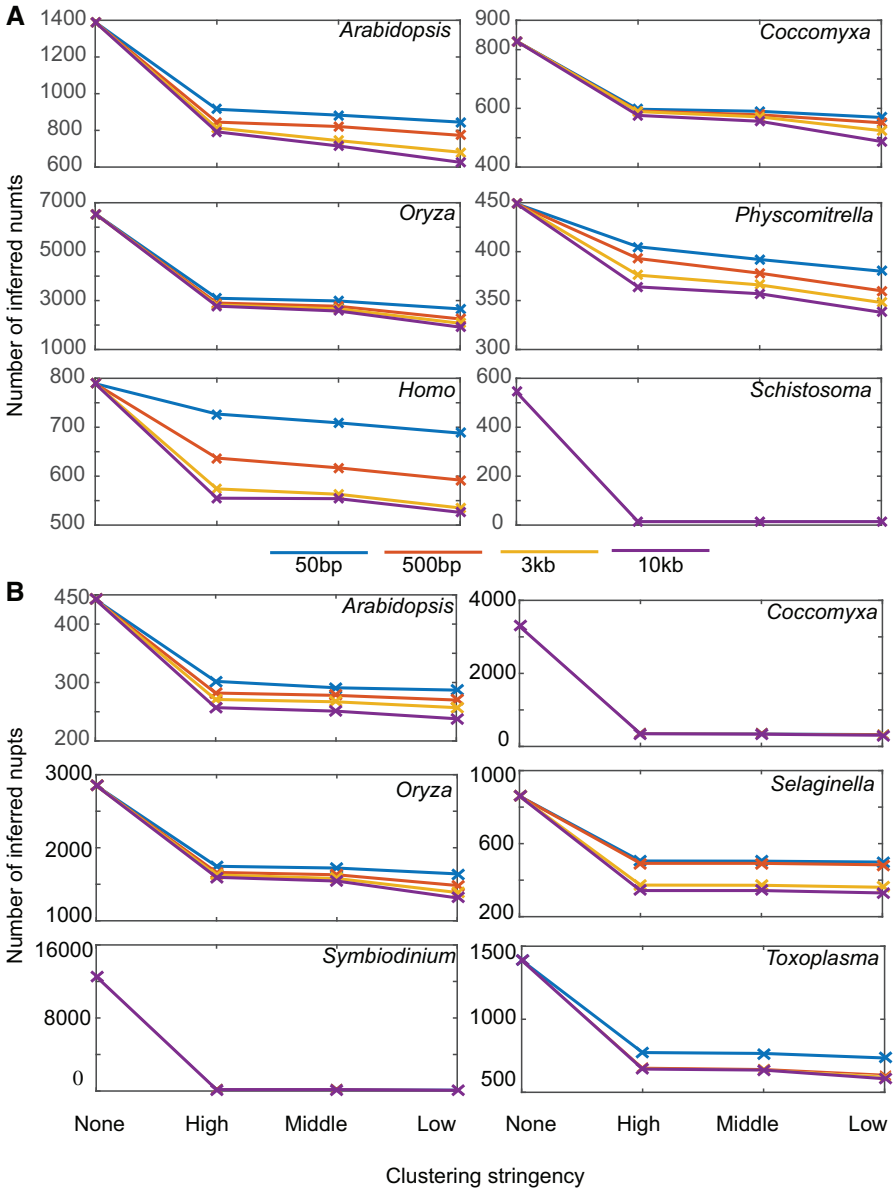

Fig. 1.—Number of inferred (*A*) numts and (*B*) nupts obtained by different clustering stringencies and concatenation distances. Clustering stringency is shown on the *x*-axis, different concatenation distances are depicted in colors (blue 50 bp, red 500 bp, yellow 3 kb, and purple 10 kb). Species with at least 400 BLAST hits are shown.

falciparum, *Aureococcus anophagefferens*, *Nannochloropsis gaditana*, *Phaeodactylum tricornutum*, and *Thalassiosira pseudonana* . The data from these 16 genomes reveal a strong correlation between the number of inferred numts and the number of inferred nupts (Spearman nonparametric rho = 0.723, $P = 0.0015$). That indicates that the tendency to undergo organelle DNA insertion is a characteristic of the nuclear lineage. Examples include the *Arabidopsis thaliana* genome with 773 numts and 270 nupts, the *Coccomyxa subellipsoidea* genome with 551 numts and 338 nupts, and the *Oryza sativa* genome with 2,249 numts and 1,483 nupts.

How Old Are Numts and Nupts?

Using a naïve molecular clock approach, one can obtain estimates for the age of numt and nupt insertions. This entails many assumptions, including the assumption that the mutation rate is constant, such that accumulated mutational differences between the nuclear insertion and the organelle copy reflect roughly the amount of time that has elapsed since the time of nuclear insertion (Hazkani-Covo et al. 2003; Huang et al. 2005). There are a number of caveats to such estimates. For example in animals, most observed mutations have accumulated in the organelle copy, because the

mitochondrial mutation rate is an order of magnitude faster than the nuclear rate. In plants, precisely the opposite is observed, that is, the nuclear rate is higher (Huang et al. 2005). Despite many approximations and uncertainties, such estimates help to illustrate the recent nature of numts and nupts.

The *relative* age distribution of numts and nupts as indicated by percent nucleotide identity extracted from their BLAST alignments is shown in figure 2. The patterns are different for different genomes. Some genomes are enriched in recent insertions such as numts in *Cyanophora*, in others we hardly see evidence for recent insertions, such as numts in *Dictyostelium*. Though we assumed a molecular clock and though we know that age is overestimated, in primates we were able to identify numts that are 50–60 Ma without sequences having reached saturation (Hazkani-Covo et al. 2003).

Figure 2*A* shows a schematic phylogenetic tree for genomes having at least 20 numts or nupts. Figure 2*B* (for numts) and 2*C* (for nupts) show the distribution of BLAST scores or inferred BLAST scores for the permissive BLAST analysis of numts with an allowed concatenation distance of 500 bp. Figure 2*C* shows the same for nupts. The fraction of scores is shown for each species separately for all insertions with 60–100% identity to organelle DNA. The distribution of insertions with high identity to organelle DNA differs across species. Some genomes show a high proportion of very recent (almost identical to organelle DNA) insertions, for example numts in *Phytophthora*, whereas others show a higher proportion of low sequence identity (more ancient) insertions, such as numts in *Oryza*.

For those plastid-bearing eukaryotes with sufficient numbers of numts and nupts to plot a distribution, the relative age

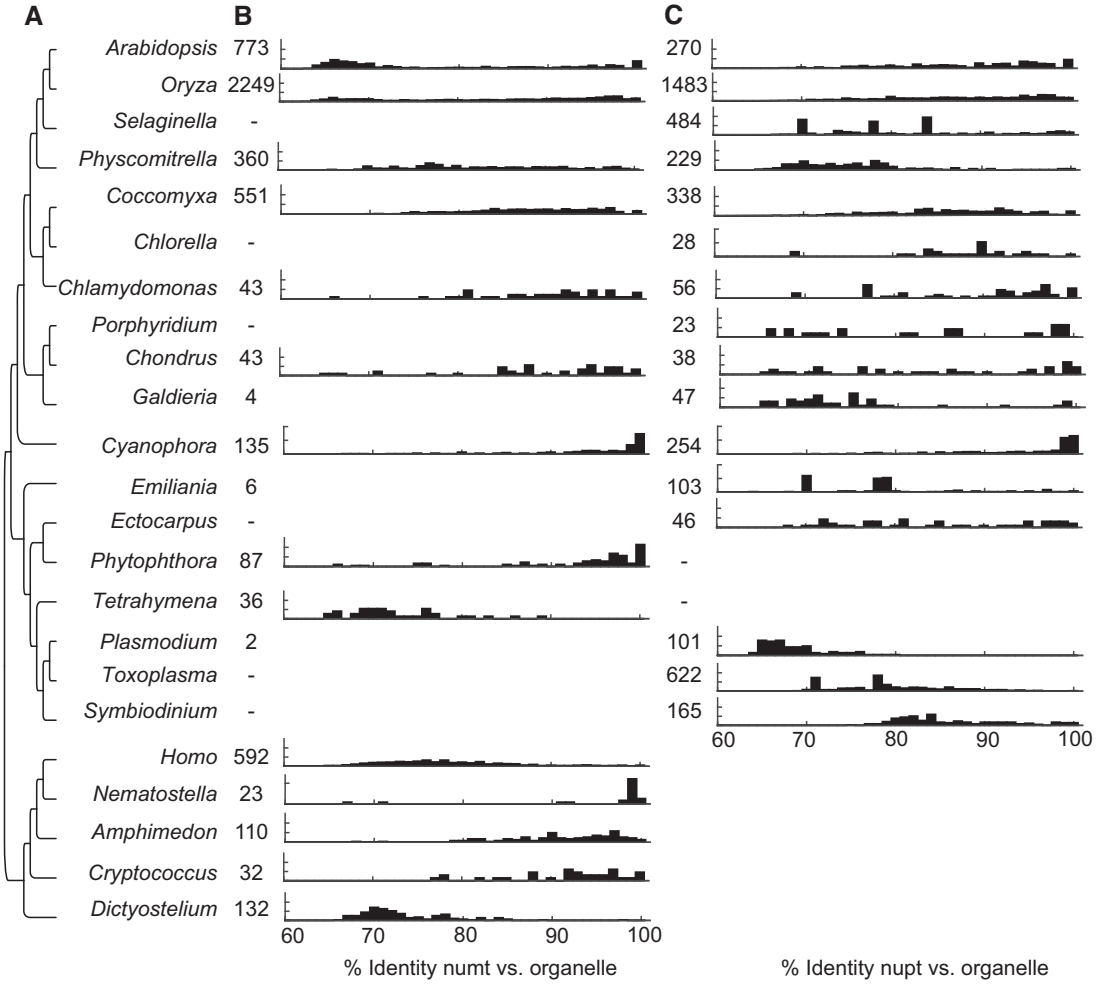

Fig. 2.—(*A*) A schematic phylogenetic tree and distribution of BLAST identity scores for permissive counting showing at least 20 inferred insertions with a concatenation distance of 500 bp for (*B*) numts and for (*C*) nupts. Distribution of BLAST identity scores for permissive counting showing at least 20 inferred insertions with a concatenation distance of 500 bp. Concatenated numt and nupt scores were calculated as a weighted mean. Scales are shown between 60% and 100% identity and the distribution is shown up to 30%. The sum of the cumulative distribution is one. Some species include columns above 30% as *Chlorella* (40%), *Coccomyxa* (40%) and *Emiliania* (40%) for nupts, and *Cyanophora* (40%) and *Nematostella* (70%) for numts.

distribution of the two kinds of insertions, not just their numbers, appears to follow the same trend. This is significant for *Arabidopsis thaliana*, *Oryza sativa*, *Physcomitrella patens*, *Coccomyxa subellipsoidea*, and *Chondrus crispus* but not for *Cyanophora paradoxa* and *Chlamydomonas reinhardtii* (Kolmogorov–Smirnov $P < 0.05$). This suggests that the process of transfer is not organelle-dependent, but again, as with the number of insertions, dependent upon processes specific to the lineage or its nuclear genome. It can furthermore be seen from figure 2 that transfers can be clustered in different lineages, either skewed predominantly towards very recent transfers (near 100% sequences identity) or towards more ancient transfers (near 60% sequences identity). This indicates that although numt and nupt transfers occur continuously throughout evolution, they do not occur at a constant rate, showing evidence across several lineages for periods of higher and lower transfer frequencies, respectively.

Discontinuous or Continuous with Organelle DNA?

When organelle DNA is transferred to the nucleus, it can be integrated as a single copy or fragment of the genome. In such cases, the BLAST result would be sufficient and there would be no need to look for rearrangements, subsequent insertions, or other kinds of recombinational events that would generate discontinuity with organelle DNA. As shown in tables 1 and 2, many of the numts and nupts that we identified in the present study are interrupted in some manner, that is, they are not continuous with organelle DNA. We call these insertions "complex," where the designation as complex makes no statement on the issue of whether the discontinuity arose during the process of insertion or through subsequent recombinational events. We use the word recombination here in a generic sense to designate all kinds of events that could lead to discontinuity between organelle and numt/nupt DNA, that is, insertion, inversion, deletion, and the like, without specifying the mechanism(s) involved.

The number of these complex insertions varies between species. Although only 5% of the numts in *Dictyostelium discoideum* are complex, 50% of the numts in *Oryza sativa* are composed of more than a single BLAST hit. Similarly, complex events in nupts range between 16% in *Chondrus crispus* and 64% in *Coccomyxa subellipsoidea* and include almost all events (98%) in *Symbiodinium minutum*. This raises the question of whether hits that are proximate on the nuclear genome within a distance of 0.05–500 kbp, which are the ranges that we examined and tabulated in tables 1 (numts) and 2 (nupts), are in fact the result of the same insertion, or whether independent insertions occurred at the same site in these cases. The simplest answer to this question is obtained by comparing the sequence identity of adjacent hits to the organelle DNA. Both for numts and for nupts, this difference is usually very low, even for the least stringent concatenation criteria (fig. 3). Hence for the vast majority of complex insertions sampled here, adjacent hits appear to stem from the same insertion event as part of the same numt/nupt, but this does not reveal whether the fragmentation of the numt/nupt occurred during insertion or subsequently.

What Breaks Them Up?

In principle, complex numts/nupts could be the result of fragmentation during the transfer/insertion event or they could result from DNA insertion into the numt/nupt itself. In the latter case, the insertion separating the organelle derived sequences should, in the simplest case, be some kind of mobile DNA. We tested this possibility by looking at the complex numts and nupts that are composed of two or more BLAST hits separated by at least 10 bp of nuclear DNA. For genomes with at least 80 such complex insertions, we used RepeatMasker (Smit et al. 2013) to look for the existence of components of the repetitive DNA in each genome within these complex numts/nupts. The analyzed genomes include *Oryza sativa* for both numts and nupts as well as *Arabidopsis*

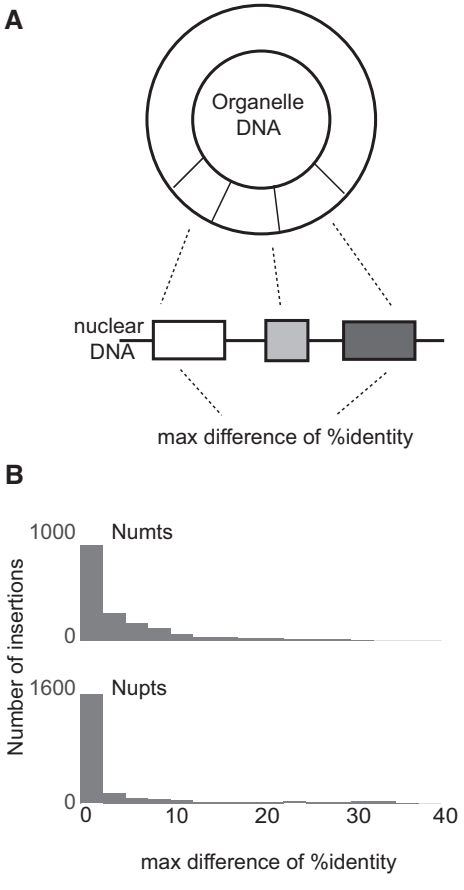

Fig. 3.—Difference in percent identity between BLAST hits belongs to the same insertion. (*A*) For each inferred insertion, the maximum identity score-difference between separated BLAST hits was calculated. (*B*) Histograms of the difference of % identity for numts and nupts. Data from permissive concatenation with a distance up to 500 bp.

thaliana and human for numts, where large numbers of numts/nupts are observed. As seen in table 3, when numts or nupts include nonorganelle DNA, complex numts are enriched in repetitive sequences germane to the nuclear genome. For example, 165 out of 773 *Arabidopsis thaliana* numts are complex, of which 92 include at least 10 bp of interrupting DNA. Within the interrupting DNA, 77 have repeats including 92 retroelements, 12 DNA transposons, 2 small RNAs, 1 satellite, and 7 low complexity regions. Similarly, when considering *Oryza sativa* numts, 1,130 are complex of which 196 include at least one 10 bp stretch of nonnumt DNA. Using RepeatMasker (Smit et al. 2013), we found 64 retroelements, 25 DNA transposons, 8 small RNA, 1 simple repeat, and 9 low complexity regions in 82 numts. In *Oryza sativa*, out of 1,483 nupts 638 are complex, of which 113 include more than 10 bp nonnupt DNA. These 113 nupts include 4 retroelements, 13 DNA transposons, 18 small RNAs, 2 simple repeats, and 13 low complexity regions in 37 nupts. Thus, when there is nonnumt/nupt derived DNA in a numt/nupt region in the nuclear genome, these are components of the nuclear genome's repetitive DNA, and thus most likely correspond to post numt/nupt insertions of mobile DNA.

Discussion

In the present study, we have inferred the number of independent numt and nupt insertions in genomes with a focus on quantifying the degree to which standard BLAST approaches overestimate the numbers of numts and nupts in nuclear genomes. The number of inferred, independent numts and nupts in nuclear genomes is dramatically reduced when organelle repeats, tandem duplication of inserted organelles as well as pre and post insertion fragmentations are taken into account. The stringency of concatenation is of lesser importance. Nevertheless, overestimating the number of insertions is only one part of the problem of inferring the number of numts and nupts in genomes. Because of their similarity to *bona fide* organelle DNA, numt and nupt insertions are often missed during the genome assembly process, or they are found but misidentified as mitochondrial and plastid DNA, or they are otherwise unrecognized.

Methods to help improve quantification of organelle insertions, as we have presented here, should be also relevant in nonevolutionary investigations, for example in the investigation of numt acquisition and dynamics in human health studies, for example, somatic ageing and cancer progression genomics. As new sequencing technologies permit more rapid identification of organelle DNA sequences in healthy and pathogenic cells, more rapid and accurate counting of novel numt insertions becomes a key to understanding their significance.

Missing and Unrecognized Inserts: Where Will It Make a Difference?

The oldest and best known source of problems that numts and nupts create is in the field of molecular systematics using organelle DNA markers (Bensasson et al. 2001). The appearance of numts in mitochondrial DNA studies has long been known (Zhang and Hewitt 1996; Thalmann et al. 2004; Benesh et al. 2006; Sword et al. 2007; Buhay 2009). More attention is now given to this problem. A main source of

Table 3

Repeats Identified by RepeatMasker in Complex Numts and Complex Nupts in Genomes Harboring at Least 80 Inferred Insertions

	Numts/Nupts	Complex Numts/Nupts Out of the Total Number	Number of Numts/Nupts With at Least 10 bp Spacer	Number of Numts/Nupts With Repeats	Repeats Identified
Numts					
Arabidopsis thaliana	773	165	92	77	92 retroelements, 12 DNA transposons, 2 small RNA 1, satellite, 7 low complexity
Oryza sativa	2249	1130	196	82	64 retroelements, 25 DNA transposons, 8 small RNA, 1 simple repeat, 9 low complexity region
Homo sapiens	592	96	82	61	52 SINEs, 4 LTRs, 1 DNA element, 45 small RNA, 9 simple repeats, 25 low complexity regions
Nupts					
Oryza sativa	1483	638	113	37	4 retroelements, 13 DNA transposons, 18 small RNA, 2 simple repeats, 31 low complexity regions

NOTE.—Values apply to interruptions of the numt or nupt by a nonorganelle-DNA spacer of >10 bp.

confusion concerning numts is that DNA barcoding techniques identify organisms with a fragment of mitochondrial cytochrome c oxidase I (COI) (Blaxter 2004; Lorenz et al. 2005). DNA barcoding among arthropods was found to overestimate the number of species when numts are coamplified (Song et al. 2008). Bertheau et al. (2011) identified numts that differed by only 1–3 nucleotides from mitochondrial DNA in European bark beetles. Haran et al. (2015) compared phylogeographic patterns obtained before and after elimination of numts from mtDNA data for the beetle *Monochamus galloprovincialis*: numts increase haplotype numbers with dramatic effects on genetic diversity estimates and phylogeography.

A different type of problem is encountered with next generation sequence (NGS) data. These new methodologies are biased towards underestimation of numt/nupt content. Studies in humans suggest that the rate of numt insertion is $\sim 5.1–5.6 \times 10^{-6}$ per germ cell per generation, or that every two human haploid genomes should be polymorphic for at least two numt loci (Bensasson et al. 2003; Ricchetti et al. 2004). However, this prediction does not seem to hold for next generation sequencing (NGS) data. In a previous study (Hazkani-Covo et al. 2010), we reported 12 numts that are polymorphic in humans, ten of which we identified ourselves in the reference human genome and an additional two insertions that were known previously (Yuan et al. 1999; Giampieri et al. 2004), but that are absent from the human reference genome. Lang et al. (2012) surveyed the 1000 human genome data for numt polymorphism and reported another two numts that are present in the human reference genome. However, analyzing 940 individuals based on the indels reported by the 1000 genome consortium, they only found four additional numts that are missing from the reference genome.

By contrast, Dayama et al. (2014) recently reassembled the 1000 human genome data from reads (note that reassembly and analysis is very different from reanalysis of the existing assembly) and reported 141 new numts that were overlooked in the initial 1000 genome assemblies, probably because of their similarity to mitochondrial DNA. They also found nearly complete mtDNA insertions that had been removed in earlier assemblies (Dayama et al. 2014). In addition to assembly procedures, sequencing technologies newer than NGS should improve the situation, as the lesson from tardigrade genomes reveals. The NGS tardigrade genome was reported to have recently acquired 17% of its protein coding genes from bacteria (Boothby et al. 2015), whereas the same species independently sequenced using long reads revealed almost no evidence for LGT at all (Koutsovoulos et al. 2016), the 17% report largely being an artifact of NGS sequencing data.

Missing and Unrecognized Inserts: Disease and Somatic Insertions

So far we have focused on numt and nupt insertions through evolutionary time. Although numts were mis-reported as mitochondrial mutations in patients (Wallace et al. 1997), there is also evidence for organelle to nucleus transfers both in human disease and during the individual ontogeny. Five numt insertions have been implicated in causing human disease (Chen et al. 2005). One involves a 41-bp mtDNA insertion at the breakpoint junction of a translocation (Willett-Brozick et al. 2001). Four others were insertions in genes, one causing severe plasma factor VII deficiency (Borensztajn et al. 2002), one causing a case of mucolipidosis IV (Goldin et al. 2004), one causing Usher syndrome type IC, and one that was associated with the Chernobyl incident causing a rare case of Pallister–Hall syndrome (Turner et al. 2003).

The case of the Chernobyl disaster (Turner et al. 2003) is of specific interest as numts have long been related to DNA repair and instability. Sporadic reports showed early on that numt insertion occurs in cancer-related cases such as HeLa cell (Shay and Werbin 1992), rat hepatoma cells (Corral et al. 1989), and low grade brain tumor (Liang 1996). The limiting factor in numt/nupt integration is probably the physical introduction of organelle DNA into the nuclear compartment. If so, then increased organelle instability (organelle lysis) might be associated with increased integration. Numts and nupts are inserted into double strand breaks (DSBs) in the nuclear genome via a nonhomologous end joining (NHEJ) mechanism (Blanchard and Schmidt 1996; Ricchetti et al. 1999). Their integration into the nuclear genome involved fewer local deletions than expected (Hazkani-Covo and Covo 2008). Ju et al. (2015) used a split read approach to analyze NGS data from cancer tissues. Their analysis revealed 16 recent, independent somatic insertions in 12 out of 587 comparisons of tumor versus normal cells from the same individuals. Most of the numts observed occur in combination with other events of genomic instability such as inversions, translocations, and large deletions. In a more recent study, a numt insertion event into USP51 gene led to formation of a novel *FOXR2* promoter and induced oncogene expression (Sturm et al. 2016). No reports have yet emerged that indicate a mechanism other than NHEJ in cancer cell numt mutational insertions. *Bona fide* mitochondrial DNA is thought to be commonly mutated in cancer, although it was also suggested (Schon et al. 2012) that some of the mitochondrial mutations common in cancer might be numt artifacts. If so, then the same fundamental kinds of problems that numts cause in systematics might also affect investigations of mtDNA in tumor cells. If numt insertions can affect gene expression in tumor cells, there is reason to suspect that they might be able to impact gene expression in evolutionary terms as well.

Both exogenous and endogenous factors can affect the integration of numts and nupts into genomes. Aging is the first and, for now, the most well-established factor influencing numt (and by inference nupt) integration. Wright and Cummings (1983) showed that in the fungus *Podospora anserina*, integration of mitochondrial DNA into the nuclear genome occurred during senescence. In was later suggested

that mitochondrial fragment integration into the nuclear genome might promote ageing as well as cancer (Richter 1988). Increased numt insertion to the nuclear chromosomes was shown to increase with aging in yeasts and rats (Caro et al. 2010; Cheng and Ivessa 2010, 2012). It was suggested that in *Saccharomyces cerevisiae* oxidative stress increases during the aging process and that this may also lead to increased fragmentation of mtDNA and numt insertion as well as to reduced survival rate (Cheng and Ivessa 2012).

Numt/nupt insertion is also affected by environmental conditions. Wang et al. (2012) showed that only mild heat stress in tobacco plants increases the frequency of both numt and nupt insertions into the nuclear genome. Their results suggest that the frequency of numt/nupt integration might vary over time both between and within lineages, which is exactly what we found in the present genomic investigation (fig. 2). Clearly such transfers are potentially a source of variation during evolution. In the early phases of eukaryote evolution, gene transfers from organelles to the nucleus were a source of new genes. In the later phases of eukaryote evolution, numt/nupt insertions were a source of variation as well. At the scale of individual life spans, the rate of numt and nupt insertion is also a detectable source of variation. Even at the level of tumor cell proliferation, the origin of novel numts can be observed. The fact that cancer cells tend to not need their mitochondria during the transition to less efficient but more rapid glycolytic growth, an observation noted by Otto Warburg (1956), might bear upon the increased rate of numt insertion noted in the comparatively few reports of genomic variation during cancer progression that have been published to date.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgment

We thank the following funding agencies: the German-Israeli Foundation grant I-1321-203.13/2015 (E.H.-C., W.F.M.), the European Research Council grants 666053 (W.F.M.); and the Open University of Israel Research Fund (E.H.-C.).

Literature Cited

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A. 112:201500012.

Benesh DP, Hasu T, Suomalainen L-R, Valtonen ET, Tiirola M. 2006. Reliability of mitochondrial DNA in an acanthocephalan: the problem of pseudogenes. Int J Parasitol. 36:247–254.

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

Bensasson D, Zhang D, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 16:314–321.

Bertheau C, Schuler H, Ck SK, Arthofer W. 2011. Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. Mol Ecol Resour. 11(6):1056–1059.

Blanchard JL, Schmidt GW. 1996. 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. 13:893.

Blaxter ML. 2004. The promise of a DNA taxonomy. Philos Trans R Soc L B Biol Sci. 359:669–679.

Boothby TC, et al. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci U S A. 112:15976–15981.

Borensztajn K, et al. 2002. Characterization of two novel splice site mutations in human factor VII gene causing severe plasma factor VII deficiency and bleeding diathesis. Br J Haematol. 117:168–171.

Buhay JE. 2009. "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. J Crustac Biol. 29:96–110.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Caro P, et al. 2010. Mitochondrial DNA sequences are present inside nuclear DNA in rat tissues and increase with age. Mitochondrion 10:479–486.

Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. 2005. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat. 25:207–221.

Cheng X, Ivessa AS. 2010. The migration of mitochondrial DNA fragments to the nucleus affects the chronological aging process of *Saccharomyces cerevisiae*. Aging Cell 9:919–923.

Cheng X, Ivessa AS. 2012. Accumulation of linear mitochondrial DNA fragments in the nucleus shortens the chronological life span of yeast. Eur J Cell Biol. 91(10):782–788.

Corral M, et al. 1989. DNA sequences homologous to mitochondrial genes in nuclei from normal rat tissues and from rat hepatoma cells. Biochem Biophys Res Commun. 162:258–264.

Dayama G, Emery SB, Kidd JM, Mills RE. 2014. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 42:12640–12649.

Du WX, Qin YC. 2015. Distribution of mitochondrial DNA fragments in the nuclear genome of the honeybee. Genet Mol Res. 14:13375–13379.

Giampieri C, et al. 2004. A novel mitochondrial DNA-like sequence insertion polymorphism in intron I of the FOXO1A gene. Gene 327:215–219.

Goldin E, et al. 2004. Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV. Hum Mutat. 24:460–465.

Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. Annu Rev Plant Biol. 59:491–517.

Gray MW. 1992. The endosymbiont hypothesis revisited. Int Rev Cytol. 141:233–357.

Gray MW. 2012. Mitochondrial evolution. Cold Spring Harb Perspect Biol. 4:a011403–a011403.

Haran J, Koutroumpa F, Magnoux E, Roques A, Roux G. 2015. Ghost mtDNA haplotypes generated by fortuitous NUMTs can deeply disturb infra-specific genetic diversity and phylogeographic pattern. J Zool Syst Evol Res. 53:109–115.

Hazkani-Covo E. 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. Mol Biol Evol. 26:2175–2179.

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

Hazkani-Covo E, Graur D. 2007. A comparative analysis of numt evolution in human and chimpanzee. Mol Biol Evol. 24:13–18.
Hazkani-Covo E, Sorek R, Graur D. 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J Mol Evol. 56:169–174.
Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 6:e1000834.
Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422:72–76.
Huang CY, et al. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 138:1723–1733.
Ju YS, et al. 2015. Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. Genome Res. 25:814–824.
Kolodner R, Tewari KK. 1979. Inverted repeats in chloroplast DNA from higher plants. Proc Natl Acad Sci U S A. 76:41–45.
Koutsovoulos G, et al. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. Proc Natl Acad Sci U S A. 113(18):5053–5058.
Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. BMC Biol. 14:89.
Ku C, et al. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524:427–437.
Lang M, et al. 2012. Polymorphic NumtS trace human population relationships. Hum Genet. 131:757–771.
Liang BC. 1996. Evidence for association of mitochondrial DNA sequence amplification and nuclear localization in human low-grade gliomas. Mutat Res Mol Mech Mutagen 354:27–33.
Lin X, et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761–768.
Lloyd AH, Timmis JN. 2011. The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. Mol. Biol Evol. 28:2019–2028.
Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol. 39:174–190.
Lorenz JG, Jackson WE, Beck JC, Hanner R. 2005. The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. Philos Trans R Soc L B Biol Sci. 360:1869–1877.
Martin W, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why?. Plant Physiol. 118:9–17.
Matsuo M, Ito Y, Yamauchi R, Obokata J. 2005. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. Plant Cell 17:665–675.
Michalovova M, Vyskot B, Kejnovsky E. 2013. 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) 111:314–320.
Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC. 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Hum Mutat. 23:125–133.
Mourier T, Hansen AJ, Willerslev E, Arctander P. 2001. The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol Biol Evol. 18:1833–1837.
Noutsos C, Kleine T, Armbruster U, Dalcorso G, Leister D. 2007. Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends Genet. 23:597–601.
Noutsos C, Richly E, Leister D. 2005. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Res. 15:616–628.
Pamilo P, Viljakainen L, Vihavainen A. 2007. Exceptionally high density of NUMTs in the honeybee genome. Mol Biol Evol. 24:1340–1346.
Paul P, et al. 2013. The protein translocation systems in plants—composition and variability on the example of *Solanum lycopersicum*. BMC Genomics 14:189.
Pesole G, et al. 2012. The neglected genome. EMBO Rep. 13:473–474.
Ricchetti M, Fairhead C, Dujon B. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature 402:96–100.
Ricchetti M, Tekaia F, Dujon B. 2004. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol. 2:E273.
Richly E, Leister D. 2004a. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 21:1081–1084.
Richly E, Leister D. 2004b. NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. Mol Biol Evol. 21:1972–1980.
Richter C. 1988. Do mitochondrial DNA fragments promote cancer and aging?. FEBS Lett. 241:1–5.
Rogers HH, Griffiths-Jones S. 2012. Mitochondrial pseudogenes in the nuclear genomes of Drosophila. PLoS One 7:e32593.
Schon EA, DiMauro S, Hirano M. 2012. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet. 13:878–890.
Shay JW, Werbin H. 1992. New evidence for the insertion of mitochondrial DNA into the human genome: significance for cancer and aging. Mutat Res. 275:227–235.
Sheppard AE, Timmis JN. 2009. Instability of plastid DNA in the nuclear genome. PLoS Genet. 5:e1000323.
Smit AFA, Hubley R, Green P. 2013. No Title. RepeatMasker Open-4.0. Available from: http://www.repeatmasker.org/.
Song H, Buhay JE, Whiting MF, Crandall KA. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Sci U S A. 105:13486–13491.
Stupar RM, et al. 2001. Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci U S A. 98:5099–5103.
Sturm D, et al. 2016. New brain tumor entities emerge from molecular classification of CNS-PNETs. Cell 164:1060–1072.
Sun X, Yang A. 2016. Exceptionally large mitochondrial fragments to the nucleus in sequenced mollusk genomes. Mitochondrial DNA 27:1409–1410.
Sword GA, Senior LB, Gaskin JF, Joern A. 2007. Double trouble for grasshopper molecular systematics: intra-individual heterogeneity of both mitochondrial 12S-valine-16S and nuclear internal transcribed spacer ribosomal DNA sequences in *Hesperotettix viridis* (Orthoptera: Acrididae). Syst Entomol. 32:420–428.
Thalmann O, et al. 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol Ecol. 13:321–335.
Thorsness PE, Fox TD. 1990. Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*. Nature 346:376–379.
Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123–135.
Tourmen Y, et al. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics 80:71–77.
Tsuji J, Frith MC, Tomii K, Horton P. 2012. Mammalian NUMT insertion is non-random. Nucleic Acids Res. 40:9073–9088.
Turner C, et al. 2003. Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. Hum Genet. 112:303–309.
Unseld M, Marienfeld JR, Brandt P, Brennicke a. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat Genet. 15:57–61.
Wallace DC, Stugard C, Murdock D, Schurr T, Brown MD. 1997. Ancient mtDNA sequences in the human nuclear genome: a potential source

of errors in identifying pathogenic mutations. Proc Natl Acad Sci U S A. 94:14900–14905.
Wang D, Lloyd AH, Timmis JN. 2012. Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants. Proc Natl Acad Sci U S A. 109:2444–2448.
Wang L, Wheeler D. a. 2014. Genomic sequencing for cancer diagnosis and therapy. Annu Rev Med. 65:33–48.
Warburg O. 1956. Injuring of respiration the origin of cancer cells. Science 123:309–314.
Willett-Brozick JE, Savul SA, Richey LE, Baysal BE. 2001. Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation. Hum Genet. 109:216–223.
Woischnik M, Moraes CT. 2002. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res. 12:885–893.
Wright RM, Cummings DJ. 1983. Integration of mitochondrial gene sequences within the nuclear genome during senescence in a fungus. Nature 302:86–88.
Yuan JD, Shi JX, Meng GX, An LG, Hu GX. 1999. Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome. Cell Res. 9:281–290.
Zhang D, Hewitt GM. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol Evol. 11(6):247–251.

Associate editor: Dan Graur