Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts

Christine Sacerdot[1], Serge Casaregola[2], Ingrid Lafontaine[1], Fredj Tekaia[1], Bernard Dujon[1] & Odile Ozier-Kalogeropoulos[1]

[1]Institut Pasteur, Unité de Génétique Moléculaire des Levures, CNRS, URA 2171, Université Pierre et Marie Curie-Paris 06, Paris, France; and [2]INRA CNRS AgroParisTech, UMR Microbiologie et Génétique Moléculaire, Thiverval-Grignon, France

Abstract

Transfer of fragments of mtDNA to the nuclear genome is a general phenomenon that gives rise to NUMTs (NUclear sequences of MiTochondrial origin). We present here the first comparative analysis of the NUMT content of entirely sequenced species belonging to a monophyletic group, the hemiascomycetous yeasts (*Candida glabrata*, *Kluyveromyces lactis*, *Kluyveromyces thermotolerans*, *Debaryomyces hansenii* and *Yarrowia lipolytica*, along with the updated NUMT content of *Saccharomyces cerevisiae*). This study revealed a huge diversity in NUMT number and organization across the six species. *Debaryomyces hansenii* harbors the highest number of NUMTs (145), half of which are distributed in numerous large mosaics of up to eight NUMTs arising from multiple noncontiguous mtDNA fragments inserted at the same chromosomal locus. Most NUMTs, in all species, are found within intergenic regions including seven NUMTs in pseudogenes. However, five NUMTs overlap a gene, suggesting a positive impact of NUMTs on protein evolution. Contrary to the other species, *K. lactis* and *K. thermotolerans* harbor only a few diverged NUMTs, suggesting that mitochondrial transfer to the nuclear genome has decreased or ceased in these phylogenetic branches. The dynamics of NUMT acquisition and loss are illustrated here by their species-specific distribution.

Introduction

The term 'promiscuous DNA' was coined by Ellis (1982) to denote DNA mobility among the genetic compartments of eukaryotic cells, a phenomenon evidenced by many eukaryotic genome sequences (Richly & Leister, 2004). The transfer of mitochondrial (mt) sequences to the nuclear genome gives rise to the so-called NUMTs (NUclear sequences of MiTochondrial origin) [see Leister (2005) for a review].

NUMTs are being discovered in an increasing number of eukaryotes (Lopez *et al*., 1994; Bensasson *et al*., 2001; Richly & Leister, 2004), with recent reports in species from the *Panthera* genus (Kim *et al*., 2006), in chimpanzee (Hazkani-Covo & Graur, 2007), in the rodent *Microtus rossiaemeridionalis* (Triant & DeWoody, 2007) and in honeybee (*Apis mellifera*) (Behura, 2007; Pamilo *et al*., 2007). The human genome sequence has provided a broad view of the extent of mtDNA transfer that contributed to elucidate the evolutionary dynamics of NUMTs (Mourier *et al*., 2001; Tourmen *et al*., 2002; Woischnik & Moraes, 2002; Hazkani-Covo *et al*., 2003; Ricchetti *et al*., 2004).

In the yeast *Saccharomyces cerevisiae*, fragments of mtDNA were found to integrate into chromosomes during the repair of DNA double-strand breaks (DSB) by nonhomologous end joining (NHEJ) (Ricchetti *et al*., 1999; Yu & Gabriel, 1999). *De novo* integrations of mtDNA have been detected recently in yeasts and human, suggesting that migration of mtDNA to the nucleus is a continuous and dynamic evolutionary process (Ricchetti *et al*., 1999, 2004; Yu & Gabriel, 1999; Turner *et al*., 2003; Decottignies, 2005).

The presence of NUMTs in the nuclear genome is a ubiquitous phenomenon observed in dozens of eukaryotes with great variations across species as diverse as human and plants. Few comparative studies have been performed to

characterize NUMTs in related species or subspecies (Pons & Vogler, 2005; Krampis *et al.*, 2006; Behura, 2007; Hazkani-Covo & Graur, 2007). The hemiascomycetous yeasts offer the largest number of complete and annotated genome sequences among a monophyletic group of eukaryotes. Ricchetti *et al.* (1999) have characterized the NUMT content of *S. cerevisiae*. In the present study, we performed a detailed comparative analysis of the NUMT content in five other hemiascomycetes: *Candida glabrata*, *Kluyveromyces lactis*, *Kluyveromyces thermotolerans*, *Debaryomyces hansenii* and *Yarrowia lipolytica* (Dujon *et al.*, 2004; Dujon, 2006). *Candida glabrata* is the second causative agent of human candidiasis, phylogenetically closer to *S. cerevisiae* than to *Candida albicans*, the major human fungal pathogen. *Kluyveromyces lactis* is a lactose-utilizing yeast found in cheese and commonly used in genetic studies and biotechnologies. *Kluyveromyces thermotolerans* is mainly found on grapes. *Debaryomyces hansenii* is a halotolerant yeast closely related to *C. albicans* and *Y. lipolytica* is an alkane-utilizing yeast, distantly related to all other hemiascomycetes studied so far. All five nuclear genomes are completely sequenced and annotated, as well as the cognate mitochondrial genomes, and some major steps of their evolutionary history have been identified (Dujon, 2005). These yeasts cover a broad evolutionary range that is comparable to the entire phylum of chordates (Dujon, 2006).

This work revealed unexpected interspecies variations in the NUMT abundance and organization across the hemiascomycete phylum. The presence of numerous large mosaics of NUMTs in the genome of *D. hansenii* illustrates this diversity.

Materials and methods

Genomes analyzed in this work

Candida glabrata, *K. lactis*, *D. hansenii* and *Y. lipolytica* nuclear genome sequences and annotations were retrieved from the 'Génolevures' web site (http://cbi.labri.fr/Genolevures/). The 'Génolevures' consortium has also sequenced the genome of *K. thermotolerans* (accession numbers CNS12AB9, CNS12ABA to CNS12ABG).

Full-length mitochondrial sequences were retrieved from the 'Génolevures' web site. The mitochondrial genome of *D. hansenii* was also sequenced by the Génolevures Consortium and annotated by us (EMBL accession number DQ508940). For all species, the sequenced mitochondrial and nuclear genomes come from the same strain: CBS138, CLIB210, CBS6340, CBS767 and CLIB99 for *C. glabrata*, *K. lactis*, *K. thermotolerans*, *D. hansenii* and *Y. lipolytica*, respectively.

NUMT detection

BLASTN searches were conducted locally using the complete nuclear genome sequence of each species as the database and the cognate mitochondrial genome sequence as the query. The default parameters were used (BLASTN matrix [1-3], gap penalties [existence 5; extension 2], without filtering for low complexity).

In order to select HSPs ['High-scoring Segment Pair' (Altschul *et al.*, 1997)], we proceeded as follows: (1) we identified *a posteriori* the segments of low compositional complexity contained in the query sequence using the formula of Wootton & Federhen (1996). (2) We deduced the length of high-complexity sequence of each HSP, defined as the sum of the lengths of all high-complexity subsequences. (3) We selected HSPs having a predefined minimal length of high complexity, except if their score was higher than a 'higher cut-off,' above which we accepted all HSPs, in order to validate low-complexity sequences with very high scores, as true NUMTs. (4) We compared systematically the output of this algorithm with the results of a filtered BLASTN search. Most HSPs from the filtered BLASTN were already obtained with our algorithm and some of them were extended. However, our algorithm excluded a few HSPs present in the filtered BLASTN search, which we recovered. The algorithm will be published elsewhere in more detail.

Taking the NUMT content of *S. cerevisiae* published by Ricchetti *et al.* (1999) as a reference, we chose the following parameters: minimal score = 20, minimal length of high complexity = 22 (which was also the minimal length of an HSP recovered from filtered BLASTN), higher cut-off of low-complexity tolerance = 55.

Out of the 34 published *S. cerevisiae* NUMTs, seven were extended by 6–36 nt, using our algorithm. This algorithm also found three new NUMTs and we recovered one new NUMT from a filtered BLASTN search. This algorithm did not retain five published NUMTs of small size (22–25 nt) and we did not validate them, as they were absent from a filtered BLASTN search. Then we discarded a 30-nt-long HSP as it consisted in a repeated low-complexity sequence, although it came from a filtered BLASTN search. Altogether, we updated the NUMT content of *S. cerevisiae* to 32 NUMTs (see Supporting Information, Table S1).

From *K. lactis*, we obtained 53 HSPs, out of which 40 were made of repeated sequences from two regions of the mitochondrial genome: a TAATAG repeat present in the *COX1* gene (20 HSPs) and a TATG repeat present in a mitochondrial intergene (20 HSPs). These 40 HSPs, made of periodic low-complexity sequences, were discarded. A small HSP (24 nt) mainly made of TA repeats on the nuclear sequence was also discarded. Finally, a 24-nt sequence of mt tRNA-Thr matched with four different copies of nuclear tRNA-Lys, with a single mismatch. We considered that this

low score (20) alignment, which was found four times, resulted from sequence similarity between tRNAs and probably not from mtDNA insertion. Consequently, the four HSPs were discarded, leaving eight authentic NUMTs in *K. lactis*. Exactly the same four alignments of fragments of tRNA sequences were also found in the genome of *K. thermotolerans* and discarded, leaving only one NUMT in the latter genome.

NUMT classification

Some NUMTs appear to be clustered on the chromosomes. In order to define clusters, we calculated the probability for two NUMTs to be closer than a given distance D by chance.

Theoretically, if NUMTs are considered as a set of N points distributed randomly along a straight line of length L (the chromosome), the probability P that the distance of a NUMT to its closest neighbor is lower than D can be shown to be equal to KD where $K = 2(N-1)/L$. This result is obtained by stating that this probability is equal to one minus the probability to have all the $(N-1)$ other NUMTS at a distance larger than D, i.e. outside a segment of length $2D$ centered on the NUMT. Therefore, $P = 1 - (1 - 2D/L)^{N-1}$, which is approximately equal to $2(N-1)D/L$, if we assume the distance D to be much shorter than the length of the chromosome L (i.e. $D/L \ll 1$).

For each species and for every chromosome containing at least two NUMTs, we calculated the value of D associated with a probability of 10^{-2} and compared this value with the actual distances separating every NUMT from its closest neighbor. The value of D was taken as a cut-off to define the clusters on each chromosome (Table 1).

Depending on the succession of the mitochondrial segments of the NUMTs found in a cluster, two situations were distinguished: (1) when the NUMTs in the cluster were in the same order and orientation as the corresponding mtDNA segments on the mitochondrial map and separated by intervening sequences of similar sizes, the cluster was classified as a 'procession.' Processions correspond to a single mtDNA segment transferred to the nucleus, followed by mutational decay. (2) Otherwise, the cluster was classified as a 'mosaic.' Mosaics correspond to multiple mtDNA segments transferred to the same chromosomal locus.

Note that in this work, the term 'NUMT' always refers to the segment of nuclear sequence homologous to a segment of mitochondrial genome, identified in practice as the hit sequence of a validated HSP; it does not refer to the whole cluster.

Detection of duplicated NUMTs within a species

Within each species, we searched for NUMTs that would be duplicated along with their chromosomal flanking regions. We selected pairs of NUMTs that originated from the same

Table 1. Calculation of the distances that define a cluster of NUMTs

Species	Chromosome	D	Longest distance $< D$	Shortest distance $> D$
S. cerevisiae	I	576	6	None
	IV	7660	None	385436
	VII	909	12	162845
	IX	1100	None	183343
	X	3728	None	409190
	XI	3332	9	None
	XII	2695	None	38592
	XIII	1156	69	3104
C. glabrata	A	491	18	1360
	H	5252	None	692938
	K	6514	None	1186123
K. lactis	C	2923	276	None
	D	8578	5	None
	F	13011	104	None
D. hansenii	A	568	16	6548
	B	320	298	2748
	C	1134	120	10507
	D	335	216	12219
	E	478	328	4923
	F	398	162	736
	G	427	107	838
Y. lipolytica	A	1280	88	23565
	B	3833	None	206799
	C	3273	None	288157
	D	2595	None	23870
	E	3017	None	21511
	F	2224	18	126541

The distance D associated to a probability of 0.01 was calculated for each chromosome containing at least two NUMTs (see Materials and methods). The absence of distance $< D$ ('none' in column 4) is characteristic of a chromosome devoid of clustered NUMTs. The absence of distance $> D$ ('none' in column 5) is characteristic of a chromosome containing only clustered NUMTs.

mitochondrial sequence or from largely overlapping mitochondrial segments (overlap longer than 70% of the longer NUMT). Then we looked for a possible similarity between the flanking nuclear DNA sequences of the two NUMTs by BLASTN search using a 2-kb nuclear region encompassing each NUMT as a query. We used the same BLASTN parameters as for NUMT detection.

Search for orthologous NUMTs between species

Note that the yeast species analyzed here cover a large evolutionary distance. Unlike primates, their genomes have undergone extensive reshuffling of genetic maps. When comparing their genomic sequences, mosaics of short conserved syntenic blocks separated by numerous breakpoints and often containing internal inversions of a few genes are found between all chromosomes (Dujon, 2005). In order to find orthologous NUMTs, we searched first for NUMTs that would appear in the same syntenic block between two

species. In a first step, we defined the gene environment of each NUMT as the 10 surrounding genes (the five upstream and five downstream annotated protein coding genes) and searched for NUMTs that shared at least two homologous genes in their gene environment: we found seven pairs of NUMTs or NUMT loci satisfying this condition. However, these regions did not correspond to a syntenic block involving the corresponding chromosomes, according to the syntenic blocks calculated by D. Sherman (unpublished results), with the smallest blocks containing at least three conserved genes in the same order along the chromosomes.

Identification of orthologous proteins from *Pichia stipitis* and protein alignments

Pichia stipitis genomic data were retrieved from the JGI website (http://genome.jgi-psf.org/Picst3/Picst3.home.html). We considered the best bi-directional BLASTP hit as the orthologue. In order to compare two protein sequences, we performed global alignments using the Needleman–Wunsch algorithm (Needleman & Wunsch, 1970).

PCR analyses

Total genomic DNA was extracted from *D. hansenii* sequenced strain CBS767 using the same procedure as for *S. cerevisiae*. Primer sequences are available upon request.

Results

NUMT content in six hemiascomycetous yeast genomes

We identified all NUMTs present in six hemiascomycetous yeasts: *S. cerevisiae*, *C. glabrata*, *K. lactis*, *K. thermotolerans*, *D. hansenii* and *Y. lipolytica* (see Tables S1–S6), using a new algorithm that selects HSPs from a BLASTN search upon their score and the amount of high-complexity sequence they contain (see Materials and methods). We found 32 NUMTs in *S. cerevisiae*, in agreement with the publication of Ricchetti *et al.* (1999) (34 NUMTs), with six small (22–30 nt) low-complexity sequences that were not retained by our algorithm and four new NUMTs. Moreover, the size of seven NUMTs was significantly extended (by 6–36 nt). A striking variation in the NUMT number was observed across the species, with 145 NUMTs in *D. hansenii*, by far the largest number observed here, and only one low-score NUMT in *K. thermotolerans* (length 25 nt, score 21, *e*-value of 5.11×10^{-2}). *Candida glabrata*, *K. lactis* and *Y. lipolytica* genomes contain, respectively, 14, 8 and 47 NUMTs. Altogether, the number of NUMTs varies by two orders of magnitude across the six species (Fig. 1a), without correlation with the size of the corresponding mitochondrial genomes (Table 2).

The percentage of sequence identity between the NUMTs and their mitochondrial counterpart varies from 82 % to 100% under our conditions of detection. Two species lack NUMTs with 100% sequence identity: *K. thermotolerans* (one 96% identity NUMT) and *K. lactis* (87–97% identity NUMTs) (Fig. 1b).

All NUMTs are in the same size range, with the largest NUMT being found in the *C. glabrata* genome (388 nt). The average (and median) lengths of NUMTs are 71 (60), 102 (80.5), 50 (47.5), 65 (54) and 43 (38) nt, respectively, in *S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii* and *Y. lipolytica* genomes. We verified that the NUMTs do not result from sequence assembly errors by PCR analysis on the nuclear DNA of *D. hansenii* and confirmed the presence of 28 NUMTs in this species. Figure 2 shows the amplification of 10 chromosomal loci.

No orthologous NUMTs were detected between the species analyzed here, suggesting that each NUMT arose from a novel insertion that occurred in each species lineage. This result is consistent with the broad evolutionary range covered by these yeasts (see Materials and methods) and with the fact that probably NUMTs disappear by mutational decay at a higher rate than the one that would leave recognizable traces at such distances.

In most cases, a given mitochondrial segment was present only once in the genome. However, we detected one duplication of NUMTs in *D. hansenii*, one NUMT was present in six identical copies in *Y. lipolytica*, and we confirmed the four duplicated NUMTs described previously in *S. cerevisiae* (Ricchetti *et al.*, 1999). In all cases, the flanking sequences of the NUMTs are identical or highly similar, suggesting that the duplication of the NUMTs results from the duplication of a larger chromosomal region after insertion of the mtDNA segment. The *D. hansenii* paralogous NUMTs (31 nt long, 100% sequence identity with mtDNA) were found within an 8.9 kilobase (kb) segmental duplication of subtelomeres of chromosomes D and F. The six paralogous NUMTs of *Y. lipolytica* were all in subtelomeric regions. The sequence identity between the subtelomeric regions containing the NUMT is much higher (96–100%) than the sequence identity found between the NUMTs and their relative mitochondrial segment (89–91%), indicating that mtDNA inserted first into one chromosome and spread later by subtelomere shuffling. Louis & Haber (1991) had already suggested that NUMTs could propagate by subtelomeric shuffling in *S. cerevisiae*.

NUMT locations on the chromosomes

Using present annotations of the genomes, we examined the position of each NUMT in its genetic environment. All NUMTs detected in *C. glabrata*, *K. lactis*, *K. thermotolerans* and *Y. lipolytica* were intergenic. All updated NUMTs of
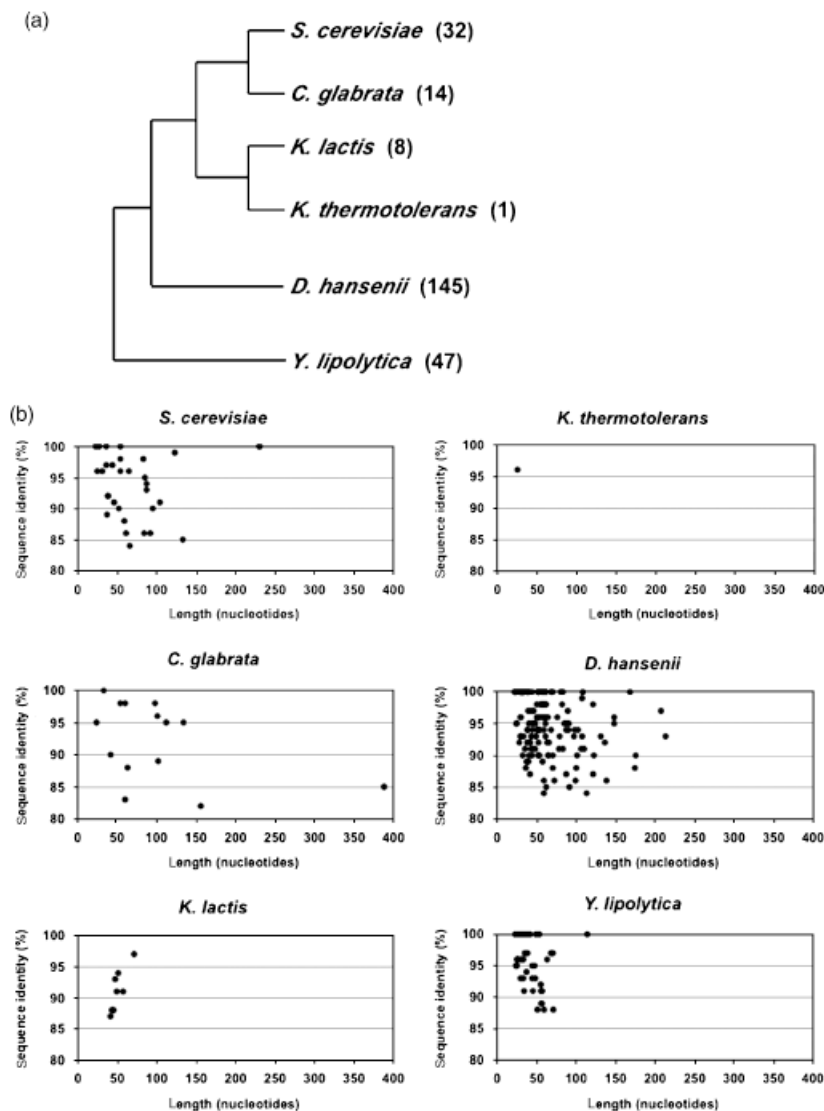
Fig. 1. Diversity of the NUMT content in six hemiascomycetous yeasts. (a) The phylogeny of the species is adapted from Dujon (2006). Only the general topology of the tree is illustrated and the NUMT number in each species is added within parentheses. (b) Distribution of sequence length and conservation of the various NUMTs identified in each species. The percentage of sequence identity is expressed as the percentage of nucleotides conserved between the NUMT and the mtDNA of the same strain.

Table 2. NUMT occurrence in six hemiascomycetous yeast species

Yeast species	S. cerevisiae*	C. glabrata	K. thermotolerans	K. lactis	D. hansenii	Y. lipolytica
Nuclear genome size (Mb)	12.1	12.3	10.4	10.6	12.2	20.5
Mitochondrial genome size (kb)	85.7	20	23.5	40.2	29.4	47.9
NUMT number	32	14	1	8	145	47
Total transferred mitochondrial DNA (bp)	2356	1423	25	403	9377	2005
Transferred mitochondrial DNA (%)	2.7	7.1	0.1	1	31.8	4.2

*The NUMT content of the genome of *Saccharomyces cerevisiae*, previously published by Ricchetti *et al.* (1999), has been updated using our method of detection.

S. cerevisiae were also within intergenic regions. The two NUMTs of *S. cerevisiae* described previously within an ORF (Ricchetti *et al.*, 1999) are composed of low-complexity sequences and were not validated as NUMTs by our criteria (see Materials and methods). In *D. hansenii*, 92% of the NUMTs (133/145) were found within intergenic

regions. The other 12 NUMTs overlap one (in one case, two) annotated gene(s), and can be classified into five categories.

(1) Four NUMTs (D6, D86, D97, D129 in Table S5) overlap by 8 to 31 nt the end of a short annotated ORF (117 nt to 219 nt). In all cases, the ORF harbors no similarity
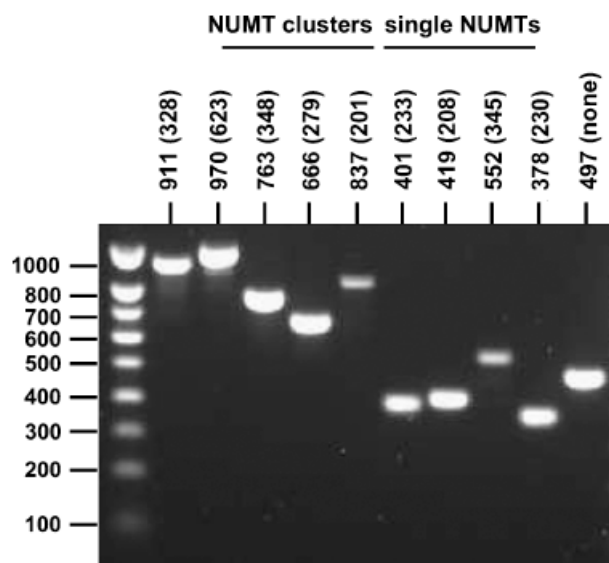
Fig. 2. PCR analysis of 10 NUMT loci of *Debaryomyces hansenii*. The presence of five NUMT clusters and five single NUMTs was verified by PCR amplification of 10 *D. hansenii* chromosomal loci and electrophoresis of the PCR products on a 1.2% agarose gel. PCR reactions were performed on genomic DNA of strain CBS767. The two numbers shown on top of each lane indicate the expected size of the PCR product carrying and not carrying (within parentheses) the NUMT or cluster of NUMTs. All oligonucleotides were designed outside the NUMT (or NUMT cluster) sequence, except for the last lane (right), where one of the two oligonucleotides was chosen within the NUMT.

to known genes. Either a true gene has been interrupted by a NUMT, or the ORF is spurious and the NUMT is actually intergenic. Similarly, one NUMT (D47 in Table S5) overlaps by 10 nt, the beginning of a short ORF (216 nt).

(2) Five NUMTs (D8, D37, D47, D91, D138 in Table S5) overlap by 1 to 39 nt significantly longer genes (804 nt to 1520 nt), including either start (3 NUMTs) or stop (2 NUMTs) codons (Fig. 3a). The length of these ORFs and their sequence similarities to known genes suggest that they are true genes. It is noteworthy that all NUMTs of this category are low-identity NUMTs (87–93% sequence identity with the mitochondrial segment), suggesting that they evolved to fit with the function of the gene.

(3) One case concerns a NUMT within an intron (Fig. 4a and b). NUMT D105 (Table S5) is located within the S1 region situated in between the 5′-splice site and the branchpoint of the third intron of *DEHA2F06314g*, which contains five introns. This part of spliceosomal introns is variable in length, which makes possible the insertion of a 51-nt-long NUMT without a threat on splicing. Note that this is one of the only two genes of *D. hansenii* with five introns, the other genes having much less ntrons (mostly one) or no intron at all (the majority of the genes).

(4) One NUMT (D64 in Table S5) was found within a highly conserved gene (*DEHA2D14124g*). A multiple alignment between orthologues from other yeast species revealed highly similar protein sequences in this locus, without a gap (data not shown). However, no NUMT was detected within the orthologous genes from the other species, suggesting that this low-score NUMT (29 nt, 93% sequence identity) is a fortuitous match rather than a real NUMT.

(5) One 148 nt long NUMT (D74 in Table S5) contains a spurious ORF (antisense of the mitochondrial gene *ATP9*); however, the mitochondrial insertion did not occur within a preexisting gene.

In order to gain more insight into the five genes overlapping a NUMT (Fig. 3a), we compared their protein sequence with their possible orthologue from *P. stipitis*, a yeast that is phylogenetically close to *D. hansenii* (Jeffries *et al.*, 2007). Significant alignments could be generated for three genes (Fig. 3b and Supporting Information). *DEHA2C03388g* protein differs from its orthologue from *P. stipitis* mainly by an extended N-terminal domain provided by the NUMT, insofar as its coding sequence actually starts on the first ATG (Fig. 3b). On the contrary, alignment shows that *DEHA2F00198g* protein sequence starts 114 amino acids (aa) downstream of the *P. stipitis* protein orthologue, suggesting that the insertion of the NUMT interrupted the gene of *D. hansenii*, followed by loss of the 5′ coding sequence (see alignment in Appendix S1). Finally, despite the presence of a NUMT overlapping the stop codon of the target gene, *DEHA2G19470g* protein does not appear to be significantly truncated, because the alignment with its orthologue from *P. stipitis* shows just one amino acid (aa) truncation (see alignment in Appendix S1).

As NUMTs are mainly located in intergenic regions, we asked whether we could find NUMTs overlapping pseudogenes. The pseudogenes were identified by protein sequence similarity of the translated intergenic regions in the six reading frames, using the proteome of the six yeast species, including mitochondrial proteins (I. Lafontaine, unpublished data). Two kinds of pseudogenes were found, which correspond to different situations: (1) the mitochondrial pseudogenes are traces of mitochondrial coding sequences and were all found overlapping a NUMT detected by our method (based on DNA sequence similarity). NUMTs associated with a mitochondrial pseudogene were found in all species, except *K. lactis* and *K. thermotolerans* (Table 3). (2) The nuclear pseudogenes are traces of nuclear genes whose sequences have been degraded over evolutionary time. Among the mutations, the insertion of a NUMT might be the event that triggered pseudogenization. Alternatively, the NUMT inserted within an intergene already containing a pseudogene. NUMTs overlapping (or within) nuclear pseudogenes were found only in *D. hansenii* (Table 4). In
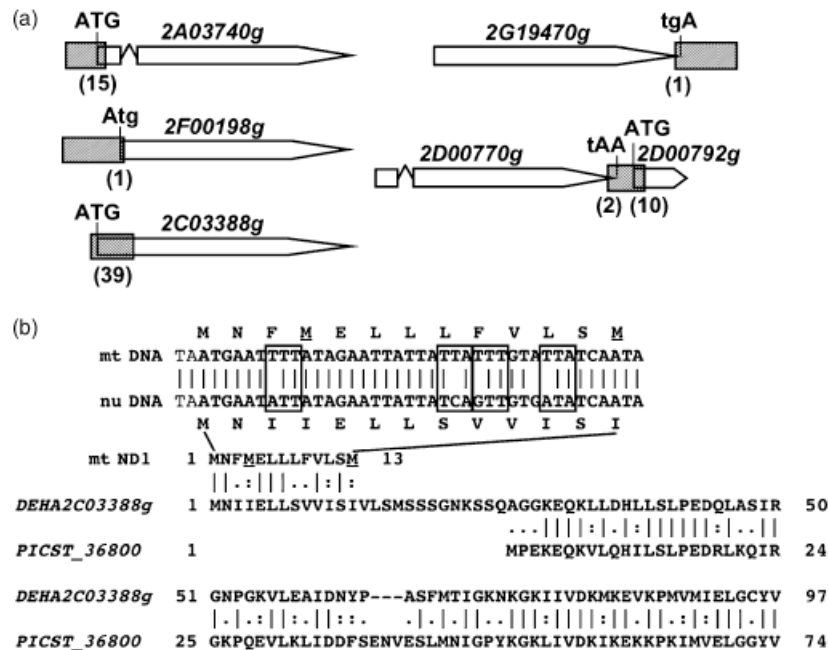
Fig. 3. Cases of NUMTs overlapping protein-coding genes. (a) Five NUMTs of *Debaryomyces hansenii* overlap a long gene (> 800 nt). The genes are depicted as large arrows below their name without the species code ('*2A03740g*' instead of '*DEHA2A03740g*'). The broken lines symbolize the introns. A hatched box represents the NUMT. The number in parentheses below the hatched box indicates the length of the overlap (nt). Start codons (ATG) and stop codons (TGA, TAA) of the genes are indicated if they are located within the NUMT sequence or if they overlap it: upper case letters indicate nucleotides located inside the NUMT sequence, and lower case letters, indicate nucleotides outside the NUMT sequence. The end of gene '*2D00770g*' is overlapped by a NUMT that also overlaps the beginning of a short ORF (216 nt, annotated as '*2D00792g*'). The five genes are annotated as follows: *DEHA2A03740g*: 'some similarities with uniprot|P43634 *Saccharomyces cerevisiae* YLR098C CHA4 Zinc-finger protein with Zn[2]-Cys[6] fungal-type binuclear cluster domain. *DEHA2C03388g*: 'similar to CA5283|IPF1912 *Candida albicans* IPF1912,' *DEHA2D00770g* : 'weakly similar to uniprot|O93999 *C. albicans* Ca20C1 Possible zinc-finger protein,' *DEHA2F00198g* : 'weakly similar to uniprot|P53388 *S. cerevisiae* YPL265W DIP5 dicarboxylic amino acid permease mediates high-affinity and high-capacity transport of L-glutamate and L-aspartate,' *DEHA2G19470g* : 'similar to uniprot|O13871 *Schizosaccharomyces pombe* SPAC1B3 SPAC1B3.06c protein.' (b) The beginning of the alignments of gene *DEHA2C03388g* product with its orthologue from *Pichia stipitis* (*PICST_36800*) and with the N-terminal part of *ND1* mitochondrial gene is shown. The first 13 aa that are homologous to *ND1* correspond to the NUMT whose DNA sequence (nu DNA) is represented above the protein alignment, aligned with its mitochondrial counterpart (mtDNA), along with their translation. The two internal methionines of *ND1* (codons ATA that encode an isoleucine in the nucleus) are underlined. Boxes surround the four nonsynonymous mutations in the NUMT sequence.

most cases (four of five), only a small portion of the original gene shows detectable similarity to the pseudogene (12–25% of the length of the homologous protein). NUMTs associated with these pseudogenes either overlap one extremity of the pseudogene or are located within it. Three NUMTs clustered on chromosome D were found within the same pseudogene (Fig. 5a). One full-length pseudogene containing a NUMT was also detected. The low degradation of the pseudogene allowed the identification of the main mutations that degraded the sequence of the original gene (Fig. 5b).

Genomic organization of NUMTs

Analysis of the distances separating the NUMTs on chromosomes revealed single NUMTs and NUMTs grouped into clusters. We defined clusters as sets of NUMTs separated by a distance associated with a probability lower than 0.01 to

occur by chance (see Materials and methods). The proportion of single NUMTs varies considerably across the species analyzed, from 0% in *K. lactis* to 86% in *C. glabrata* – if we exclude *K. thermotolerans* with its single NUMT (Table 5). We distinguished two main types of clusters, depending on the order and orientation of the fragments in the mitochondrial genome compared with the nuclear cluster. The 'procession' type is characterized by a succession of NUMTs that are collinear to their associated mitochondrial segments. Processions probably result from a single primary insertion of mtDNA, whose sequence has degenerated over time, leaving small detectable regions homologous to mtDNA (the NUMTs) separated by more diverged sequences. We identified five processions altogether, including one in *S. cerevisiae*, one in *C. glabrata*, one in *K. lactis* and two in *D. hansenii*. Dot-matrix analyses showed the extent of homology between the chromosomal and mitochondrial sequences, revealing less conserved homologous fragments

(Fig. 6). The smaller procession of *D. hansenii* (cluster 3 in Table S5, 'DEHA 3' in Fig. 6) differs from the others as it contains a small intact NUMT (23 nt, 100% sequence
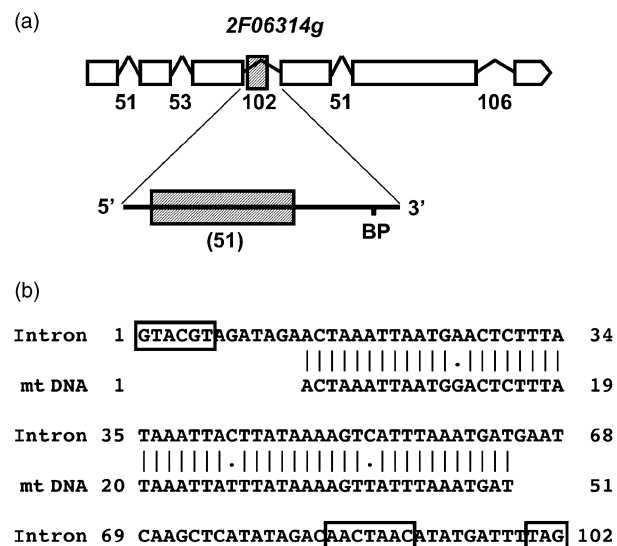

Fig. 4. Case of a NUMT present in an intron. (a) Representation of gene *DEHA2F06314g* with its five introns, one of which contains a NUMT. This gene is annotated as 'similar to uniprot|Q08271 *Saccharomyces cerevisiae* YOL132W GAS4 Putative 1 3-β-glucanosyltransferase.' The gene is depicted as an arrow interrupted by broken lines symbolizing the introns. The numbers below the introns indicate their size (nt). The intron that contains the NUMT is depicted at a larger scale; 'BP' is the branchpoint; the length of the NUMT is indicated within parentheses. (b) Sequence of the intron of *DEHA2F06314g* that contains the NUMT. The NUMT is represented by its alignment with the mtDNA. Boxes surround the motifs of 5', branchpoint and 3' splice sites.

identity), suggesting that this cluster may not have arisen from a single insertion event.

The second type of cluster consists in 'mosaics,' where NUMTs originate from disparate mitochondrial fragments. Mosaics are by far the most common type of cluster, with a total of 32 mosaics (vs. five processions) across the six species (Table 5). However, no mosaic was found in *C. glabrata*, while all NUMTs of *K. lactis* are distributed in three mosaics, one of which contains an internal procession of two NUMTs. In *S. cerevisiae*, three mosaics had been reported previously (Ricchetti *et al.*, 1999) and we found an additional NUMT in one of them; moreover, we identified another mosaic due to a new NUMT joined to a previously described NUMT. *Debaryomyces hansenii* contains 23 mosaics, by far the largest number, accounting for 54% of its NUMTs. Besides nine mosaics of two NUMTs and five mosaics of three NUMTs, this species harbors the largest mosaics: six mosaics of four NUMTs, one mosaic of six NUMTs, one of seven NUMTs and one of eight NUMTs, with an average number of 3.4 NUMTs per mosaic (Fig. 7). Most mosaics contain NUMTs from both orientations, suggesting that mtDNA pieces inserted in random orientation.

In the yeast *S. cerevisiae*, mosaics were associated with multiple insertions of noncontiguous mitochondrial fragments as observed during the repair of a chromosomal DSB (Ricchetti *et al.*, 1999). In the *in vivo* double insertion events, the two fragments shared a terminal homology of 2–3 nt. Similarly, we found NUMTs overlapping their neighbor on a few nucleotides. Out of 55 mosaic junctions in the genome of *D. hansenii*, 24 correspond to overlapping NUMTs (on 1–7 nt), seven to exact junctions, four to nearly exact junctions (1 nt gaps) and 12 are gaps of 3–32 nt. These

Table 3. Number of pseudogenes overlapped by a NUMT in the six species

Yeast species	*S. cerevisiae*	*C. glabrata*	*K. lactis*	*K. thermotolerans*	*D. hansenii*	*Y. lipolytica*	All species
Mitochondrial pseudogenes	3	4	0	0	9	1	17
Nuclear pseudogenes	0	0	0	0	5	0	5
Total pseudogenes	3	4	0	0	14	1	22

Pseudogenes were identified by I. Lafontaine (unpublished data), using the nuclear and mitochondrial proteomes of all species.

Table 4. NUMTs overlapping (or included in) nuclear pseudogenes of *Debaryomyces hansenii*

Chromosome	NUMT start	NUMT end	Pseudogene start	Pseudogene end	Homologous gene	% Length of protein	% Protein sequence identity
Deha2B	15293	15315	15309	15689	*DEHA2C00132g*	22	47
Deha2C	19512	19643	19209	19802	*DEHA2E00176g*	25	33
Deha2D	6098	6186	5852*	6522	*DEHA2E13244g*	20	27
Deha2D	6403	6447	5852*	6522	*DEHA2E13244g*	20	27
Deha2D	6479	6583	5852*	6522	*DEHA2E13244g*	20	27
Deha2E	42195	42256	42012	42195	*DEHA2F01012g*	12	62
Deha2F	20393	20453	18637	20600	*DEHA2B16676g*	98	81

*Same pseudogene containing three clustered NUMTs.

47 junctions (85%) can be considered as 'tight junctions.' The remaining cases (8/55, 15%) concern NUMTs that are separated by 63–328 nt from their neighbor, and can be considered as 'loose junctions.' Both tight and loose junctions coexist in three mosaics (clusters 7, 18, 20 in Table S5, see Fig. 7b for mosaic cluster 18).
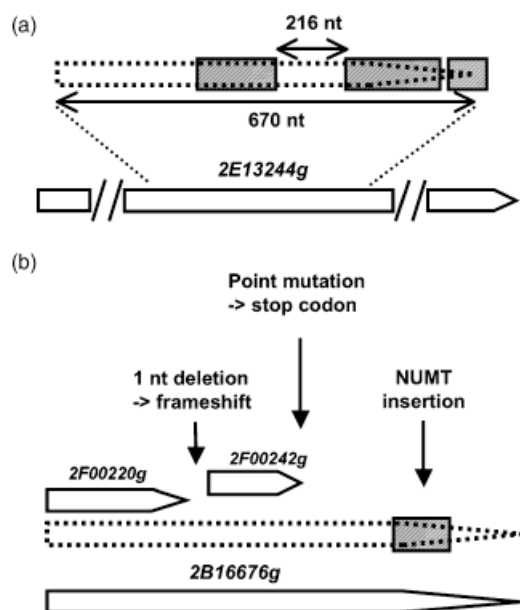

Fig. 5. Cases of NUMTs in nuclear pseudogenes of *Debaryomyces hansenii*. Annotated genes are depicted as large arrows below their name (see the legend of Fig. 3). Pseudogenes are depicted as dotted arrows. Hatched boxes represent NUMTs. (a) A mosaic cluster of three NUMTs (D43, D44, and D45 building cluster 7, Table S5) is found within a highly degraded pseudogene of chromosome D. The paralogous gene (from chromosome E) is depicted as a large interrupted arrow below the pseudogene. The size of the detected part of the pseudogene and that of the larger gap of the mosaic are indicated, respectively, below and above a bidirectional arrow. (b) NUMT D91 (Table S5) is present within a full-length pseudogene of chromosome F. The paralogous gene (from chromosome B) is represented below the pseudogene. Two annotated ORFs (located within the pseudogene) are indicated. Vertical arrows indicate the major mutations that degraded the sequence of the original gene.

Discussion

In the present study, we report the diversity of the NUMT content across the hemiascomycete phylum and provide a detailed genome-wide investigation of NUMT organization within six species.

The size of the NUMTs in all six yeast species does not exceed 400 nt, which is small in comparison with other eukaryotes where they can reach several thousands nucleotides (Richly & Leister, 2004). The longest NUMT was found in the genome of *C. glabrata* (388 nt) and *Y. lipolytica* harbors the smallest average and median length of NUMTs (43 and 38 nt, respectively). Traces of longer insertions were detected as processions of NUMTs, which reveals that mtDNA fragments of > 520 nt (in *C. glabrata*) and 800 nt (in *D. hansenii*) have integrated into these nuclear genomes. The surprising absence of recent insertions as long as processions (the longest NUMT with 100% sequence identity is 230 nt, found in *S. cerevisiae*) suggests that long insertions were rare in recent times. The small size of all yeast NUMTs could be linked to the compactness of the yeast genomes. However, this is not a general rule because the compact genome of *Neurospora crassa* harbors numerous long NUMTs (up to 4136 nt, average size 647 nt) while the large genome of *Rattus norvegicus* contains only small NUMTs (Richly & Leister, 2004).

All NUMT sequences in this study were 82–100% identical to the mtDNA. Mismatches between NUMTs and mtDNA result either from nuclear mutations after transfer or from mutations in the mitochondrial genome since the transfer occurred.

The most striking results are the high number of NUMTs and the presence of numerous large mosaics in the genome of *D. hansenii*, which contains up to eight mitochondrial fragments at the same locus. The fragments originate from different regions of the mtDNA, inserted independently of their orientation and order along the mitochondrial genome. Tight mosaics (when the NUMTs are very close to each other or overlap) are likely to result from the capture of several fragments of mtDNA during the repair of a chromosomal DSB, as reported experimentally in *S. cerevisiae*

Table 5. Genomic organization of the NUMTs in the different species

Species	Number of single NUMTs	Number of NUMTs in a cluster	Number of clusters	Number of mosaics	Number of processions	Specific cases
S. cerevisiae	19 (59%)	13 (41%)	5	4	1	
C. glabrata	12 (86%)	2 (14%)	1	0	1	
K. lactis	0 (0%)	8 (100%)	3	3	1	The procession is within a mosaic
K. thermotolerans	1 (100%)	0 (0%)	0	0	0	
D. hansenii	61 (42%)	84 (58%)	25	23	2	
Y. lipolytica	40 (85%)	7 (15%)	3	2	0	One cluster is a tandem repeat of NUMTs

Fig. 6. Dot-matrix comparisons of the nuclear and mtDNA sequences encompassing processions of NUMTs. Nuclear sequences are on the horizontal axis and mitochondrial sequences are on the vertical axis. The processions of *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii* (clusters 3 and 6) and *Saccharomyces cerevisiae* are indicated, respectively, by CAGL, KLLA, DEHA3, DEHA6, and SACE. The DNA dot matrices were performed with DNA Strider software, using a 15/23 nt threshold (15 identical nucleotides over a sliding window of 23 nts) for all matrices, except DEHA6, where a threshold of 15/21 was used. Black boxes below the matrices represent the detected NUMTs. The NUMT procession of *C. glabrata* consists of two NUMTs separated by a low-complexity sequence. The NUMT procession of *K. lactis* consists of two NUMTs whose mitochondrial segments' intervening sequence contains a mobile GC cluster (gray box) that did not integrate into the chromosome.
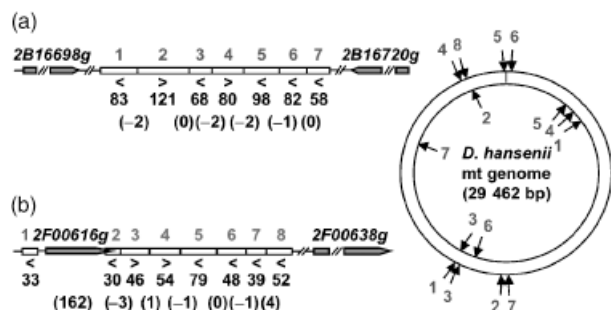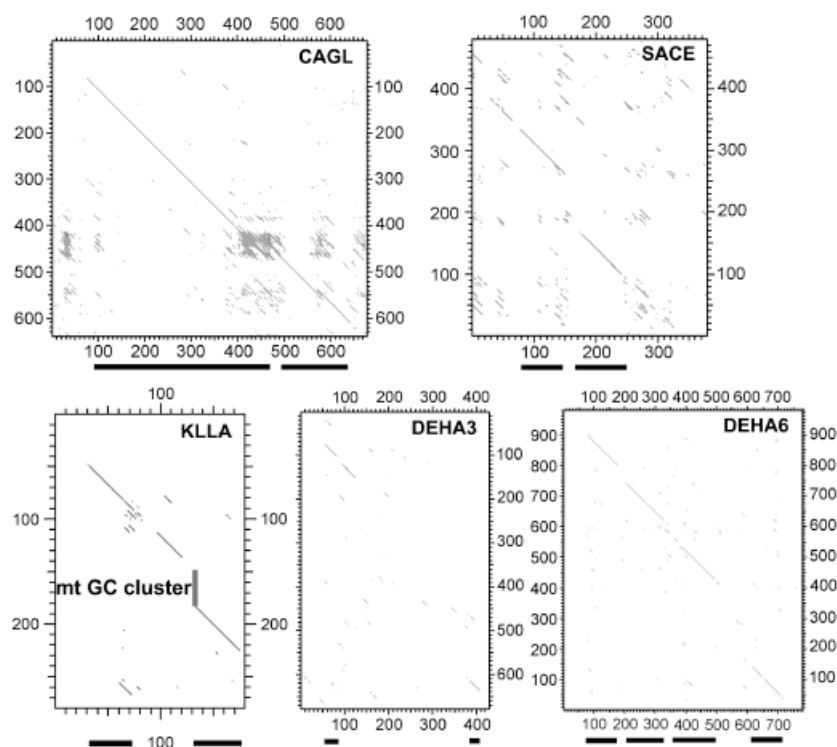

Fig. 7. The two largest mosaics of NUMTs of *Debaryomyces hansenii*. Each NUMT is represented as a white box and numbered above the box according to its order on the chromosome. Below each box, the character ' > ' (=same) or ' < ' (=reverse) indicates the orientation of the NUMT, refering to the orientation of the nontranscribed strand of the mitochondrial genome relative to the arbitrarily oriented nuclear genome. The number below the symbol for the orientation indicates the size of the NUMT (nt) and the numbers in parentheses between two neighbor NUMTs indicate the distance separating them [a negative number ($-k$) means that the two NUMTs overlap on k nt]. The flanking genes are depicted as gray arrows that are interrupted when the gene cannot be drawn to scale. The name of these genes is given above the arrows, as in Fig. 3. The double circle represents the mitochondrial genome of *D. hansenii*. The internal arrows indicate the regions where NUMTs of mosaic depicted in (a) come from. Similarly, the external arrows indicate the origin of the NUMTs of mosaic depicted in (b). (a) Mosaic cluster 5 (Table S5) on chromosome B. (b) Mosaic cluster 18 (Table S5) on chromosome F.

(Ricchetti *et al.*, 1999). It is not excluded that some mosaics result from more than one insertion event. For example, mosaic cluster 16 (Table S5) combines a diverged NUMT (91% sequence identity over 106 nt) and an intact NUMT (100% sequence identity over 40 nt), which are likely to be the result of two successive insertions. However, due to the small size of the NUMTs, their sequence conservation compared with the mtDNA does not allow accurate estimation of the time of insertion. Loose mosaics (when the NUMTs are more distant to each other) could be the result of successive insertions in a fragile region of the chromosome. The high number of mosaics in the genome of *D. hansenii* accounts for 54% of the NUMTs and 27% of the NUMT loci and suggests that an important reservoir of mtDNA fragments is available to be captured during the repair of DSBs. This yeast might undergo a continuously high rate of mtDNA fragmentation, or bursts of massive mtDNA fragmentation under some conditions, like stress.

Another striking result is the quasi absence of NUMTs in the genome of *K. thermotolerans*. The only validated HSP harbors an *e*-value of 5.1×10^{-2}, below the threshold chosen by Richly & Leister (2004). Interestingly, *K. lactis*, which is phylogenetically close to *K. thermotolerans*, harbors only degraded NUMTs (average and median % sequence identity=91%, and no NUMT with 100% sequence

identity), as if mitochondrial transfer had decreased or ceased in this evolutionary branch.

Insertion of mtDNA into nuclear genomes is a potentially mutagenic phenomenon. Recent insertions in the human genome have been associated with diseases (Willett-Brozick *et al.*, 2001; Borensztajn *et al.*, 2002; Turner *et al.*, 2003; Goldin *et al.*, 2004). As already observed in *S. cerevisiae*, we found most NUMTs in intergenic regions (92% in *D. hansenii* and 100% in the other species). However, we discovered new situations. We found a NUMT within an intron the preferential location of recent human-specific NUMTs (Ricchetti *et al.*, 2004). This NUMT is located in a part of the intron, where it should not interfere with splicing. We also found five NUMTs of *D. hansenii* that overlap one extremity of a gene, potentially changing gene expression and/or resulting in protein truncation or extension (Fig. 3a). All five NUMTs are degraded, signing old insertions; nevertheless, the open reading frames have not been interrupted by nonsense mutations, suggesting that the genes are still functional and have adapted to the presence of the NUMT. Moreover, one of these NUMTs could have acted positively on the evolution of the overlapped gene by addition of a new protein sequence (*DEHA2C03388g*, Fig. 3a and b). This NUMT contains four non-synonymous mutations out of five, suggesting that it may undergo positive selection, as observed for the evolution of new genes (Long *et al.*, 2003). This hypothesis is consistent with the recent study of Noutsos *et al.* (2007), who found a positive impact of NUMTs on evolution.

The diversity in the NUMT content of the six genomes analyzed here may be explained by differences in mtDNA fragmentation or escape rate, differences in transfer to the nucleus or efficiency of integration into chromosomes or different rates of mutational decay over evolutionary time. The causes of mtDNA fragmentation and the mechanisms of transfer to the nucleus are largely unknown. Experimental evidences showed that NUMT acquisition ultimately relies on the DNA repair machinery in *S. cerevisiae*. Most genes involved in DSB repair are conserved across the six hemiascomycetes analyzed here (Richard *et al.*, 2005), including *K. thermotolerans* (unpublished). The NHEJ pathway has been investigated experimentally in *K. lactis*, where integration of nonhomologous DNA molecules into the genome has been found 1000-fold more frequently than in *S. cerevisiae* (Kegel *et al.*, 2006). This observation shows that differences in DNA repair efficiency do exist among the species. It also indicates that the low number of NUMTs in the genome of *K. lactis* is not due to low efficiency of NHEJ. The genomes of *D. hansenii* and *K. thermotolerans*, which are the 'exceptions' for the NUMT content, are not an exception in terms of presence or absence of a gene involved in DSB repair. However, even orthologous genes may possess different regulation and functional properties. Moreover, the comparative analyses were performed with *S. cerevisiae* taken as a reference, and so the existence of *D. hansenii*- or *K. thermotolerans*-specific genes that would modulate the DSB repair machinery in these species cannot be excluded.

Acknowledgements

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Behura SK (2007) Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. *Mol Biol Evol* **24**: 1492–1505.

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

Borensztajn K, Chafa O, Alhenc-Gelas M, Salha S, Reghis A, Fischer AM & Tapon-Bretaudiere J (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.

Decottignies A (2005) Capture of extranuclear DNA at fission yeast double-strand breaks. *Genetics* **171**: 1535–1548.

Dujon B (2005) Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev* **15**: 614–620.

Dujon B (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* **22**: 375–387.

Dujon B, Sherman D, Fischer G *et al.* (2004) Genome evolution in yeasts. *Nature* **430**: 35–44.

Ellis J (1982) Promiscuous DNA – chloroplast genes inside plant mitochondria. *Nature* **299**: 678–679.

Goldin E, Stahl S, Cooney AM, Kaneski CR, Gupta S, Brady RO, Ellis JR & Schiffmann R (2004) Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV. *Hum Mutat* **24**: 460–465.

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.

Jeffries TW, Grigoriev IV, Grimwood J *et al.* (2007) Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat Biotechnol* **25**: 319–326.

Kegel A, Martinez P, Carter SD & Astrom SU (2006) Genome wide distribution of illegitimate recombination events in *Kluyveromyces lactis*. *Nucleic Acids Res* **34**: 1633–1645.

Kim JH, Antunes A, Luo SJ, Menninger J, Nash WG, O'Brien SJ & Johnson WE (2006) Evolutionary analysis of a large mtDNA translocation (NUMT) into the nuclear genome of the Panthera genus species. *Gene* **366**: 292–302.

Krampis K, Tyler BM & Boore JL (2006) Extensive variation in nuclear mitochondrial DNA content between the genomes of *Phytophthora sojae* and *Phytophthora ramorum*. *Mol Plant Microbe Interact* **19**: 1329–1336.

Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet* **21**: 655–663.

Long M, Betran E, Thornton K & Wang W (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–875.

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.

Louis EJ & Haber JE (1991) Evolutionarily recent transfer of a group I mitochondrial intron to telomere regions in *Saccharomyces cerevisiae*. *Curr Genet* **20**: 411–415.

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.

Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.

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.

Pamilo P, Viljakainen L & Vihavainen A (2007) Exceptionally high density of NUMTs in the honeybee genome. *Mol Biol Evol* **24**: 1340–1346.

Pons J & Vogler AP (2005) Complex pattern of coalescence and fast evolution of a mitochondrial rRNA pseudogene in a recent radiation of tiger beetles. *Mol Biol Evol* **22**: 991–1000.

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.

Richard GF, Kerrest A, Lafontaine I & Dujon B (2005) Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. *Mol Biol Evol* **22**: 1011–1023.

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

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

Triant DA & DeWoody JA (2007) Extensive mitochondrial DNA transfer in a rapidly evolving rodent has been mediated by independent insertion events and by duplications. *Gene* **401**: 61–70.

Turner C, Killoran C, Thomas NS, Rosenberg M, Chuzhanova NA, Johnston J, Kemel Y, Cooper DN & Biesecker LG (2003) Human genetic disease caused by *de novo* mitochondrial–nuclear DNA transfer. *Hum Genet* **112**: 303–309.

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. *Genome Res* **12**: 885–893.

Wootton JC & Federhen S (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266**: 554–571.

Yu X & Gabriel A (1999) Patching broken chromosomes with extranuclear cellular DNA. *Mol Cell* **4**: 873–881.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Tables S1–S6. NUMT data; one table per yeast species.

Appendix S1. Protein alignments of genes *DEHA2C03388g*, *DEHA2F00198g* and *DEHA2G19470g* of *Debaryomyces hansenii* with their putative orthologue from *Pichia stipitis*.

Please note: Blackwell Publishing is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.