Extensive Variation in Nuclear Mitochondrial DNA Content Between the Genomes of *Phytophthora sojae* and *Phytophthora ramorum*

Konstantinos Krampis,¹ Brett M. Tyler,¹ and Jeffrey L. Boore²

¹Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0477, U.S.A.; ²Evolutionary Genomics Department, Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, 2800 Mitchell Drive, Walnut Creek, CA 94598, U.S.A.

Submitted 31 March 2006. Accepted 26 July 2006.

Fragments of mitochondrial DNA (mtDNA) transferred to the nuclear genome are called nuclear mitochondrial DNAs (NUMTs). We report here a comparison of NUMT content between genomes from two species of the same genus. Analysis of the genomes of *Phytophthora sojae* and *P. ramorum* revealed large differences in the NUMT content of the two genomes: 16.27×10^{-3} and $2.28 \times 10^{-3}\%$ of each genome, respectively. Substantial differences also exist between the two species in the sizes of the NUMTs found in each genome, with ranges of 20 to 405 bp for *P. sojae* and 19 to 137 bp for *P. ramorum*. Furthermore, in *P. sojae*, fragments from the mitochondrial genes *rns*, *rnl*, *cox1*, and *nad* (various subunits) are found most frequently, whereas *P. ramorum* NUMTs most often originate from the *cox3*, *rps14*, *nad4*, and *nad5* genes. The large differences in the presumptive mtDNA insertions suggest that the insertions occurred subsequent to the divergence of the two species, and this is supported by sequence comparisons among the NUMTs and the mtDNA sequences of the two species. *P. sojae* mtDNA sequences inserted in the nuclear genome appear to have been altered as a result of insertions, deletions, inversions, and translocations and provide insights into active mechanisms of sequence divergence in this plant pathogen. No clear examples were found of NUMTs forming functional nuclear genes or of NUMTs inserted into exons or introns of any nuclear gene.

Additional keywords: diversification of NUMTs, mitochondria, speciation.

In eukaryotes, nuclear DNA sequences that are homologous to mitochondrial DNA (mtDNA) are called nuclear mitochondrial DNAs (NUMTs). They are presumed to result from the insertion of mtDNA into the nuclear DNA (Lopez et al. 1994). It has been shown that the most feasible mechanism for the integration of mitochondrial sequences into the nuclear genome is by nonhomologous recombination of the incoming mtDNA with the nuclear DNA (Mourier et al. 2001; Woischnik and Moraes 2002). It also has been argued that nucleic acid sequences released from mitochondria dur-

ing their disintegration or when transient breaches of the organelle's membrane occur can reach the cellular nucleus (Thorsness and Weber 1996).

The accumulation of NUMTs within a genome is a continuous evolutionary process (Bensasson et al. 2003; Mourier et al. 2001; Woischnik and Moraes 2002), and the duplication of NUMTs already existing in the nuclear genome contributes to an increase in their number (Bensasson et al. 2000; Hazkani-Covo et al. 2003; Lopez et al. 1994; Tourmen et al. 2002). Several studies have demonstrated the heterogeneous nature of NUMTs, with different degrees of similarity to their mitochondrial counterparts, variability in size, and random distribution within the nuclear genome (Bensasson et al. 2001, 2003; Woischnik and Moraes 2002). Substantial differences exist with respect to the content of mtDNA among nuclear genomes of organisms within related genera such as human, rat, and mouse; *Drosophila* spp. and mosquito; *Arabidopsis* and rice (Richly and Leister 2004a). Integration of mtDNA into the nuclear genome may have positive or negative outcomes. A number of cases of mitochondrial gene transfer and functional activation have been reported, especially for plants (Adams et al. 1999, 2001a,b; Figueroa et al. 1999; Kubo et al. 1999, 2000; Palmer et al. 2000). On the other hand, in some human tumors, mtDNA fragments were found to have disrupted tumor suppressor genes (Ricchetti et al. 2004).

The recent completion of sequencing of the nuclear and mitochondrial genomes of the oomycete plant pathogens *Phytophthora sojae* and *P. ramorum* (Tyler et al. 2006) has enabled us to characterize the mtDNA content of the nuclear genomes of two closely related species and to characterize mechanisms of sequence alteration in the inserted mtDNAs.

RESULTS

mtDNA sequence fragments in the *P. sojae* and *P. ramorum* genomes.

Using the *P. sojae* mtDNA sequence as a query for BLAST against the *P. sojae* nuclear genome with an expectation threshold (E value) of 10^{-4} resulted in 78 hits (the term "hit" refers to high-scoring segment pairs [HSP]). Of those, 76 had $\geq 90\%$ identity and 62 were at least 99% identical. The same analysis for *P. ramorum* returned only eight hits, with two at least 99% identical and all at least 90% identical. With an E value of 10^{-4} , the summed length of all BLAST hits for *P. sojae* was 11,982 bp or $13.93 \times 10^{-3}\%$ of the nuclear genome sequence. For *P. ramorum*, the sum was only 768 bp or $1.15 \times 10^{-3}\%$ of the nuclear genome.

Corresponding author: K. Krampis; E-mail: agbiotec@vt.edu; Telephone: +1.001.540.231.7752.

The e-Xtra logo stands for "electronic extra" and indicates the HTML abstract available on-line contains supplemental material not included in the print edition. An additional table appears online.

When the expectation threshold for the BLAST searches was increased to an E value of 10, 105 hits were returned for *P. sojae*, of which 103 had at least 90% identity, whereas 36 hits were returned for *P. ramorum*, with 33 being at least 90% identical. The hits comprised 13,998 bp or $16.27 \times 10^{-3}\%$ of the *P. sojae* genome and 1,516 bp or $2.28 \times 10^{-3}\%$ of the *P. ramorum* genome. Therefore, there is an approximately sevenfold difference in the genome content of NUMTs between the two *Phytophthora* spp. Finally, the ranges of lengths for the BLAST hits were 20 to 405 bp for *P. sojae* (average size: 133 bp) and 19 to 137 bp (average size: 42 bp) for *P. ramorum*.

For both *Phytophthora* spp., the BLAST searches were repeated after the mtDNA sequences had been randomized. In both cases, this did not yield more than five hits, none of which had a length over 20 bp (or a bit score over 40) even with an E value of 10.

Characterization of NUMTs found in the two *Phytophthora* genomes.

Of the NUMTs detected in the BLAST searches for *P. sojae*, we examined in detail the 33 with lengths greater than 80 bp. In addition, we examined 25 others with length less than 80 bp but located close to the other NUMTs (within 50 bp in most cases) because these NUMTs may have originated from the same insertion event of a mtDNA fragment as the nearby NUMTs. For *P. ramorum*, we examined those that were at least 50 bp in length because this genome contained fewer and smaller NUMTs than *P. sojae*. The sequences of the NUMTs were annotated using National Center for Biotechnology Information (NCBI) BLAST searches. As expected, in all cases, the hit with the highest bit score returned by NCBI BLAST was from the *P. infestans* mitochondrial genome. The gene annotations from the *P. infestans* mtDNA were transferred to each NUMT under consideration. It is not surprising, given the high density of gene coding regions in the *Phytophthora* and other mtDNAs, that all NUMTs were from coding regions of mitochondrial genes. The annotations of the NUMTs, together with their sizes and positions in the nuclear and mitochondrial genomes, are summarized for *P. sojae* and *P. ramorum* in Tables 1 and 2, respectively.

The most abundant NUMTs found in the genome of *P. sojae* were fragments of the mitochondrial genes *rns* (small subunit ribosomal RNA), *rnl* (large subunit ribosomal RNA), *cox1* (cytochrome c oxidase subunit I), and *nad7* (NADH dehydrogenase subunit 7). Fragments of other genes were found present at lower frequencies (Table 1). For *P. ramorum*, the most common NUMTs were from the mitochondrial genes *cox3*, *rps14*, *nad4*, and *nad5* (Table 2).

We found no clear evidence that any of the *P. sojae* or *P. ramorum* NUMTs were functional. Several of the *P. sojae* NUMTs overlapped with predicted transcripts (downloaded from the Joint Genome Institute's [JGI] website, *ab initio* predictions from genomic sequence), but the predicted proteins in each case were very short, and no evidence is available to support the prediction of these transcripts (Table 3). Only two *P. sojae* expressed sequence tag (EST) contigs (Virginia Bioinformatics Institute [VBI] microbial genomes database, discussed below) matched NUMTs, which appeared to be derived from mitochondrial mRNA in one case (CL2185Contig1) and rRNA in the other (CL915Contig1). Comparison of the NUMTs' sequences with *P. infestans* ESTs returned only matches to ESTs that, in turn, had perfect matches to mitochondrial DNA and, therefore, presumably were contaminants. We also found no evidence for NUMTs disrupting nuclear genes in either species.

Shuffling of NUMTs within the nuclear genome.

In eight genomic scaffolds of *P. sojae*, we found clusters of NUMTs located no more than 50 bp apart on the nuclear ge-

nome whose collective sequences corresponded to almost continuous regions of mitochondrial genes. We hypothesize that those adjacent fragments are descendants of a single mtDNA fragment insertion. These NUMTs are numbered with suffixes in Table 1.

In *P. ramorum*, only two sites were found in the nuclear genome where adjacent NUMTs appeared to originate from a single insertion (Table 2, r1-1 and r1-2, and r2-1 and r2-2). The NUMT pair r1-1 and r1-2 is identical to r2-1 and r2-2, matching nucleotides 23,639 to 23,588 and 23,577 to 23,441, respectively, in the *P. ramorum* mtDNA in the *cox3* gene. Close inspection suggests that this resulted from a sequence assembly error that has created an apparent duplication in the nuclear genome sequence (see below).

Several types of rearrangements are represented by the NUMT clusters and are shown on Table 1. NUMTs s1-1, s1-2, and s1-3 are contiguous in the nuclear DNA but separated in the mtDNA, indicating that deletions occurred in the mtDNA fragment originating these NUMTs. Likewise, NUMTs s8-2 and s8-3 and s18-1 and s18-2 exhibit deletions in the NUMT. In contrast, NUMTs s9-1 and s9-2 are nearly contiguous in the mtDNA but separated by 9 bp in the nuclear DNA, indicating an insertion of 9 bp. NUMTs s20-1 and s20-2 are separated in both genomes and indicate that both deletions and insertions have occurred during insertion of the mtDNA in the genome and creation of the NUMT.

Inversions are also apparent. On *P. sojae* scaffold 87, NUMTs s8-2 and s8-3 appear to have been separated from s8-1 by an inversion (Table 1), so that they match the order of their corresponding sequences on the original mitochondrial *rnl* gene if placed in the order s8-1, s8-3 reverse, and s8-2 reverse (Fig. 1A).

We also observed more complex rearrangements of the inserted mtDNA. One such example is the *P. sojae* NUMTs s9-1, s9-2 on scaffold 42 and s15-1, s15-2 on scaffold 1,483. Together, they represent an almost continuous region spanning half the *rnl* mitochondrial gene, from 9,849 to 10,729, which is very unlikely to have occurred by chance. This arrangement must have involved two translocations (Fig. 1B); although we cannot rule out the possibility that this resulted from sequence assembly errors because the genome sequence is a draft sequence.

Duplication of NUMTs in the nuclear genome.

Several NUMTs from each species appeared to be duplications of each other, matching similar fragments of the respective mitochondrial genomes. In *P. ramorum*, NUMT r3 (Table 2) from scaffold 40 (mtDNA region 1,226 to 1,284) is nearly identical to NUMT r4 on scaffold 561 (mtDNA region 1,225 to 1,302). Their only difference is 18 nucleotides missing at the 3' end of the NUMT r4 and three base substitutions. Further analysis of the scaffolds of *P. ramorum* containing those two NUMTs showed that the region of scaffold 40, 254,172 to 254,621 containing NUMT r3, is a close duplicate of scaffold 561, 2,365 to 2,834, where NUMT r4 resides, though there are several substitutions and small indels that distinguish the two regions. Thus, it appears likely that a duplication of this region was responsible for creating NUMTs r3 and r4. Alternatively, the two copies might represent different alleles which were assembled as distinct contigs.

In *P. sojae*, NUMT s24 from scaffold 19 (mtDNA region 34,257 to 34,353) and NUMT s27 on scaffold 64 (mtDNA region 34,236 to 34,394) were found to have great similarity. These NUMTs differ by only three substitutions, and NUMT s27 is shorter at each end. In this case, the genomic flanking sequences of the two NUMTs do not share any sequence similarity. However, the two NUMTs are flanked by AT-rich microsatellite repeats that might be responsible for their duplication.

P. sojae NUMTs s3 and s4 and s32 and s33 and *P. ramorum* NUMTs r1-1 and r2-1 and r1-2 and r2-2 also appear in the respective genome sequences as exact duplications. However, close inspection of the sequences suggests that these duplications are more likely assembly artifacts.

Dense clusters of NUMTs in *P. sojae*.

Several locations in the *P. sojae* genome contain dense clusters of NUMTs, which are tandemly arranged in the nuclear genome but originate from disconnected regions of the mtDNA. Curiously, no such clusters occur in *P. ramorum*. One

Table 1. Location and annotations of *Phytophthora sojae* nuclear mitochondrial DNAs (NUMTs)

NUMT index ^a	Scaffold	<i>P. sojae</i> genome		<i>P. sojae</i> mtDNA		NUMT size	Mitochondrial gene ^c	Orientation ^d
		5' ^b	3'	5'	3'			
s1-1	20	136,184	136,331	5,157	5,010	147	<i>rns</i>	–
s1-2	20	136,343	136,408	4,987	4,922	65	<i>rns</i>	–
s1-3	20	136,401	136,748	4,912	4,565	347	<i>rns</i>	–
s2-1	8	274,218	274,389	5,098	5,269	171	<i>rns</i>	+
s2-2	8	275,506	275,591	6,003	5,918	85	<i>rns</i>	–
s3	881	7,319	7,415	5,258	5,162	96	<i>rns</i>	–
s4	580	840	936	5,258	5,162	96	<i>rns</i>	–
s5	39	468,069	468,187	5,351	5,469	118	<i>rns</i>	+
s6	121	5,516	5,594	5,162	5,240	78	<i>rns</i>	+
s7	101	216,955	217,063	9,175	9,067	108	<i>rnl</i>	–
s8-1	87	320,084	320,147	9,286	9,349	63	<i>rnl</i>	+
s8-2	87	320,152	320,290	11,170	11,030	103	<i>rnl</i>	–
s8-3	87	320,291	320,495	10,824	10,621	212	<i>rnl</i>	–
s9-1	42	145,420	145,460	10,656	10,616	40	<i>rnl</i>	–
s9-2	42	145,469	145,615	10,613	10,468	145	<i>rnl</i>	–
s10	1	486,010	486,128	10,155	10,037	118	<i>rnl</i>	–
s11	27	570,692	570,887	10,148	9,953	195	<i>rnl</i>	–
s12	19	99,482	99,543	1,408	1,347	61	<i>rnl</i>	–
s13	59	47,335	47,414	10,019	10,098	79	<i>rnl</i>	+
s14	273	14,596	14,660	1,834	1,770	64	<i>rnl</i>	–
s15-1	1,483	1	72	10,658	10,729	71	<i>rnl</i>	+
s15-2	1,483	74	235	9,849	10,020	171	<i>rnl</i>	+
s16	40	535,135	535,187	33,945	33,893	52	<i>rnl</i>	–
s17	77	210,637	210,683	11,304	11,258	46	<i>rnl</i>	–
s18-1	59	47,123	47,213	1,873	1,783	90	<i>cox1</i>	–
s18-2	59	47,216	47,345	1,621	1,491	129	<i>cox1</i>	–
s19	101	216,690	216,769	1,860	1,939	79	<i>cox1</i>	+
s20-1	62	133,539	133,645	2,074	2,180	106	<i>cox1</i>	+
s20-2	62	133,952	134,006	2,463	2,516	54	<i>cox1</i>	+
s21	67	358,029	358,113	2,370	2,286	84	<i>cox1</i>	–
s22	2	1042,912	1,042,992	39,015	38,935	65	<i>cox1</i>	–
s23	77	210,684	210,778	3,055	2,961	94	<i>cox2</i>	–
s24	19	99,545	99,641	34,257	34,353	96	<i>nad7</i>	+
s25	25	458,854	458,972	34,448	34,330	118	<i>nad7</i>	–
s26	67	357,830	357,967	34,852	34,714	138	<i>nad7</i>	–
s27	64	92,062	91,904	34,394	34,236	158	<i>nad7</i>	–
s28	64	48,249	48,324	34,655	34,580	75	<i>nad7</i>	–
s29	105	112,909	112,982	34,127	34,200	73	<i>nad7</i>	+
s30	101	216,385	216,483	22,414	22,512	98	<i>nad1</i>	+
s31	40	539,770	539,824	22,198	22,144	54	<i>nad1</i>	–
s32	1,088	5,973	6,028	35,609	35,554	55	<i>nad2</i>	–
s33	1,088	1,697	1,752	35,609	35,554	55	<i>nad2</i>	–
s34	25	458,973	459,020	32,598	32,551	47	<i>nad4</i>	–
s35	1	485,836	485,899	21,057	21,120	63	<i>nad11</i>	+
s36	82	313,623	313,727	40,199	40,303	104	<i>nad3</i>	+
s37	273	14,983	15,069	17,379	17,465	86	<i>rpl2</i> (5')	+
s38	64	92,099	92,504	17,123	17,528	405	<i>rpl2</i> (3')	+
s39	559	8,542	8,616	17,370	17,296	74	<i>rpl2</i>	–
s40	390	11,527	11,789	39,392	39,128	264	3' <i>cox3</i> + 5' <i>atp6</i>	–
s41	82	313,428	313,622	30,366	30,172	194	<i>atpF₁-α</i>	–
s42	19	99,325	99,387	40,909	40,847	62	<i>atpF₀</i>	–
s43	16	340,561	340,754	20,039	19,846	193	<i>trnL^e</i>	–
s44	67	356,626	356,727	12,324	12,223	101	<i>rps2</i>	–
s45	67	357,666	357,769	18,348	18,245	103	<i>rps11</i> (3')	–
s46	105	112,331	112,399	13,754	13,722	32	<i>rps14</i>	–
s47	77	210,559	210,617	11,388	11,330	58	<i>orf100</i>	–
s48	121	11,321	11,388	23,875	23,808	67	<i>orf100</i>	–
s49	1	838,942	839,107	29,882	30,047	166	<i>atpF1</i>	+

^a NUMTs that are located within a distance of 50 bp, correspond to an almost continuous region of a single mitochondrial gene, and do not have overlapping sequence were classified as descendants of single-insertion events of mitochondrial mtDNA. They are numbered with suffixes and shown in bold.

^b As 5' and 3' end of the NUMTs, we consider the left and right end from the high-scoring segment pair (HSP) alignment of the BLAST output.

^c Sizes of the mitochondrial gene coding regions: *rns*, 1,502; *rnl*, 2,653; *cox1*, 1,478; *cox2*, 776; *nad7*, 1,178; *nad1*, 980; *nad2*, 1,493; *nad4*, 1,475; *nad11*, 2,006; *nad3*, 353; *rpl2*, 806; *cox3*, 917; *atpF6*, 719; *atpF1*, 1,529; *trnL*, 83; *rps2*, 590; *rps11*, 416; *rps14*, 299; and *orf100*, 302.

^d HSP alignments of the reverse complement of the mtDNA to the genomic sequence from the BLAST output were arbitrarily denoted with the “–” sign, and with “+” otherwise.

^e This NUMT corresponds to the 5' region upstream of the mitochondrial *trnL* gene, and almost the complete transcribed region except 8 bp in the 3' end.

cluster on *P. sojae* scaffold 8, 274,115 to 276,240, contains 30 NUMTs. Other regions of the *P. sojae* genome are densely populated with NUMTs in a similar fashion on the scaffolds 101, 67, 1, 62, 273, and 105.

The clustered NUMTs on scaffold 8 of *P. sojae* also are interesting because they appear to be located within a retrotransposon. The entire cluster is located between a reverse transcriptase gene (scaffold 8, 273,688 to 274,149) and a gypsy/Ty-3 polyprotein gene (scaffold 8, 276,379 to 277,660). Furthermore, this putative retrotransposon is flanked by two identical direct repeats (repeat length 1,319 bp, located on scaffold 8, 271,725 to 273,044 and scaffold 8, 277,913 to 279,232). The two retrotransposon genes and their flanking repeats were found in many copies throughout the genome, but none of the other copies included the NUMT cluster. From the other *P. sojae* NUMT clusters, we found the one located on scaffold 1 to have a putative reverse transcriptase gene in its vicinity (scaffold 1, 487,294 to 489,095) and the one on scaffold 62 to be next to a transposon-like protein (scaffold 62, 130,562 to 131,474). These *P. sojae* NUMT clusters adjacent to retrotransposons on scaffolds 8, 1, and 62 are shown schematically in Figure 2A, B, and C, respectively.

DISCUSSION

We report here an assessment of the mtDNA content of the genomes of two closely related species from the genus *Phytophthora*. We observed a large difference in NUMTs content between the *P. sojae* and *P. ramorum* genomes that likely occurred by recent insertion events in *P. sojae* and that illustrates

one mechanism by which the genomes of sister pathogen species can diversify. We also demonstrated that both species have mechanisms for diversifying the sequences of the inserted mtDNAs. Such mechanisms may act on a broader range of sequences in the genomes, thereby enhancing the genetic flexibility of these pathogens in their co-evolutionary battle with their host species or adaptation to new hosts and environments.

The comparison of the mtDNA content of the genomes of *P. sojae* and *P. ramorum* represents the first for two species from a single genus. Despite the fact that the two species belong to the same genus, they have a sevenfold difference in the NUMT content of their genomes. Dramatic differences in the content of nuclear mtDNA have been reported in other organisms (Richly and Leister 2004a), even in species of related genera. More specifically, NUMTs range in number from 0 in *Anopheles* to less than 10 in *Fugu*, *Drosophila*, *Plasmodium*, *Caenorhabditis* spp., and more than 500 in human, rice, and *Arabidopsis*. Similar variation was observed for the percentages of mtDNA content in the nuclear genomes, based on the summed length of the NUMTs and the size of each genome. So far, NUMTs have been detected in more than 80 species (Antunes and Ramos 2005; Bensasson et al. 2001; Pereira and Baker 2004; Richly and Leister 2004a). Recent studies extended the analysis to NUMTs in genomes of fish (Antunes and Ramos 2005) and chicken (Pereira and Baker 2004); however, there have been no studies of oomycete species. No correlation was found between the content of mtDNA and the ratio of coding to noncoding portions of the genomes studied to date. In other words, species with more noncoding DNA were not necessarily

Table 2. Location and annotations of *Phytophthora ramorum* nuclear mitochondrial DNAs (NUMTs)

NUMT index	Scaffold	<i>P. ramorum</i> genome		<i>P. ramorum</i> mtDNA		Size	Mitochondrial gene ^b	Orientation ^c
		5' ^a	3'	5'	3'			
r1-1 ^d	67	30,157	30,208	23,639	23,588	51	<i>cox3</i>	–
r1-2	67	30,220	30,356	23,577	23,441	136	<i>cox3</i>	–
r2-1 ^d	67	32,786	32,837	23,639	23,588	51	<i>cox3</i>	–
r2-2	67	32,849	32,985	23,577	23,441	136	<i>cox3</i>	–
r3	40	254,204	254,262	1,226	1,284	58	<i>rps14</i>	+
r4	561	2,725	2,802	1,225	1,302	77	<i>rps14</i>	+
r5	17	455,081	455,130	14,369	14,418	50	<i>nad4</i>	+
r6	22	377,082	377,172	14,177	14,267	90	<i>nad5</i>	+
r7	27	262,504	262,570	27,129	27,195	66	<i>cob</i>	+

^a As 5' and 3' end of the NUMTs, we consider the left and right end from the high-scoring segment pair (HSP) alignment of the BLAST output.

^b Sizes of the mitochondrial gene coding regions: *cox3*, 917; *rps14*, 299; *nad4*, 1,475; *nad5*, 1994; and *cob*, 1,151.

^c HSP alignments of the reverse complement of the mitochondrial mtDNA to the genomic sequence from the BLAST output were denoted with the “–” sign, and with “+” otherwise.

^d These two pairs of NUMTs are exact copies of each other.

Table 3. Nuclear mitochondrial DNAs (NUMTs) predicted as transcripts^a

Index ^b	<i>P. sojae</i> genome			mtDNA ^b			BLAST data	
	Scaffold		Gene name	Size		Alignment	Transcript	Number ^c
	No.	Location		Gene	NUMT			
s38	64	92,099–92,504	<i>rpl2</i>	806	405	405	453	119,515
s23	77	210,684–210,778	<i>cox2</i>	776	94	94	94	120,409
s1-1, s1-2, s1-3	20	136,188–136,752	<i>rns</i>	1,502	Note ^d	352	445	114,538
s41	82	313,428–313,622	<i>atpF₁-α</i>	1,529	194	194	226	120,787
s36	82	313,623–313,727	<i>nad3</i>	353	104	104	210	120,789
s40	390	11,527–11,789	3' <i>cox3</i> + 5' <i>atp6</i>	...	264 ^e	72	72	125,782

^a All NUMTs have smaller sizes than the mitochondrial gene from which they originated (Gene size versus NUMT size columns). Most of each NUMT sequence is predicted as transcript (NUMT size versus Alignment size and Transcript size columns).

^b Index corresponding to Table 1.

^c Transcript identification numbers in accordance with the data downloaded from the Joint Genome Institute.

^d Predicted transcript includes all three NUMTs from the same mitochondrial gene on this region.

^e Predicted transcript is part of the NUMT resulted from transfer of the mtDNA region, including the 3' *cox3* and 5' *atp6* mitochondrial genes. These genes are in a head-to-tail orientation in the mtDNA.

ily more prone to accumulate NUMTs in their genomes. In concert with this observation, *P. sojae* and *P. ramorum* have similar percentages of gene-coding regions (57 versus 53%) but very different NUMT contents.

In principle, the large difference in NUMT content between *P. sojae* and *P. ramorum* could have occurred either by preferential loss in *P. ramorum* following the divergence of the two species or by more rapid accumulation in *P. sojae* after the divergence. The near-perfect sequence identity of the NUMTs of the two species with their respective mtDNA strongly argues for the latter hypothesis. The mtDNA sequences of the two species clearly have diverged since speciation. Because the NUMTs match the modern mtDNA sequences, they must have been acquired relatively recently after speciation. For example, the 405-bp *P. sojae* NUMT s38 has 13 substitutions compared with the *P. sojae* mtDNA, but the same region of *P. ramorum* mtDNA has 39 substitutions compared with the *P. sojae* mtDNA. Moreover, only 1 of the 39 substitutions also is found

in the NUMT. Both observations clearly place the origin of the NUMT later than the divergence of the two species.

The most detailed studies involving analysis and characterization of NUMTs have been conducted for the human genome (Hazkani-Covo et al. 2003; Mishmar et al. 2004; Mourrier et al. 2001; Olson and Yoder 2002; Richetti et al. 2004; Tourmen et al. 2002; Woischnik and Moraes 2002). In those studies, the identified NUMTs had a continuous range of homologies from almost identical to very poor, which revealed a continuous process of accumulation of human mtDNA fragments into the genome. Similarly, NUMTs were found in the genomes of other primates, rodents, birds, and invertebrates (Zhang and Hewitt 1996) with varying degrees of sequence similarity to their mitochondrial counterparts.

Despite the recent acquisition of mtDNAs into the genomes of *P. sojae* and *P. ramorum*, the NUMTs are relatively short in length. In neither species did we find any mtDNA fragments longer than 405 bp (*P. sojae* NUMT 38). In contrast, large NUMTs containing more than 70% of the mtDNA sequence were found inserted in three regions of the human genome (mtDNA size 16,571 bp). Similarly, in the genomes of the domestic cat and chimpanzee (mtDNA sizes 17,009 and 16,554 bp, respectively), the NUMTs were composed of fragments of up to half of the mtDNA (Collura and Stewart 1995; Lopez et al. 1994).

Fig. 1. A, Representation of genomic rearrangements of *Phytophthora sojae* nuclear mitochondrial DNAs (NUMTs) s8-1 through s8-3. If placed in the order s8-1, s8-3 reverse, and s8-2 reverse, they match the order of their corresponding sequences on the original mitochondrial *ml* gene. **B**, Possible genomic rearrangements taking place during the integration into the nuclear genome of *P. sojae* NUMTs s9-1, s9-2, s15-1, and s15-2.

Fig. 2. *Phytophthora sojae* nuclear mitochondrial DNA (NUMT) clusters located near putative retrotransposon sequences. The genomic scaffold sequences are represented as continuous bars in the middle of each figure. The vertical lines on top of each figure denote a 500-bp scale for the size of the sequences. The schematics that show the positioning of the NUMTs on the scaffold sequence are a dark shade for alignment of the reverse complement of the mtDNA to the genomic sequence and a light shade otherwise. The figure was adapted from the Joint Genome Institute's genome browser.

Origin and propagation of NUMTs within the *P. sojae* and *P. ramorum* genomes.

Potential pathways for the escape of nucleic acids from mitochondria include transient breaches in the organelle's membrane during fusion, budding, or morphological reorganization. Also, release of mtDNA fragments can take place on terminal digestion of damaged mitochondria by lysosomes or vacuoles and even during the fusion of heterotypic membranes (Thorsness and Weber 1996). Most studies suggest that the mechanism for the generation of NUMTs is based on the integration of mtDNA fragments into the nuclear genome by non-homologous recombination during the repair of double-stranded breaks (Leister 2005).

Mechanisms for NUMT propagation within the genome include their duplication along with that of segments of the genome where they reside (Bensasson et al. 2003) or replication along with LINE repeats after the NUMTs being inserted adjacently to those (Hadler et al. 1998; Mishmar et al. 2004; Tourmen et al. 2002). We found one duplication in the *P. ramorum* genome (NUMTs r3, r4) that could be justified based on the former mechanism, and one in *P. sojae* (NUMTs s24, s27) which would adhere to the latter. Duplications of NUMTs have been reported for many species, including humans (Bensasson 2000; Hazkani-Covo et al. 2003; Hu and Thilly 1995; Mishmar et al. 2004; Tourmen et al. 2002), monkeys (Collura and Stewart 1995), rats (Hadler et al. 1998), domestic cats (Lopez et al. 1994), plants (Blanchard and Schmidt 1995), grasshoppers (Bensasson et al. 2001), and shrimp (Williams and Kwonlton 2001).

Transposons also may be responsible for NUMT propagation. For example, all mtDNA fragments in the NUMT-dense region of genomic scaffold 8 of *P. sojae* were found inserted in a retrotransposon. NUMTs duplicating with the aid of transposons have been found in both the human (Ricchetti et al. 1999) and plant (Mishmar et al. 2004) genomes. No other exact copies of this NUMT-dense region from scaffold 8 were found across the *P. sojae* genome; therefore, we can conclude that this particular retrotransposon has not duplicated following the NUMT insertion into it.

Diversification of NUMT sequences.

In the two *Phytophthora* spp., we observed several clusters of NUMTs that appeared to have originated from a single fragment of mtDNA but subsequently had undergone insertions, deletions, inversions, and translocations. Groups of NUMTs in the human genome also have been hypothesized to be products of single insertion events, and these groups have extensive rearrangements of the NUMTs compared with the original mtDNA sequence (Tourmen et al. 2002). The short lengths of NUMTs in *P. sojae* and *P. ramorum* and the rearrangements we observed in the *P. sojae* NUMTs suggest that these species have active mechanisms for creating diversity in sequences that are not under purifying selection. Such mechanisms may increase the pathogens' genetic flexibility in the face of selection pressure from the host. Furthermore, the presence of the NUMT sequences themselves provides a mechanism for change of the genomic content in both *Phytophthora* spp. Similar rearrangements in other species (Leister 2005) suggest these mechanisms may not be restricted to oomycete pathogens.

A more extreme form of rearrangement is represented by the dense clusters of NUMTs found in *P. sojae*. These "hot spots" for accumulation of mtDNA are densely populated with tandemly arranged NUMTs originating from distant loci of the mtDNA. More specifically, for the mtDNA-dense region of scaffold 8 of *P. sojae*, the change of order and orientation of the NUMTs residing there results in an approximate coverage of the mtDNA region 1,000 to 39,000. As described above,

transfers containing half or even up to 70% of the mtDNA previously have been reported for cats, chimpanzees, humans, and plants (Collura and Stewart 1995; Lopez et al. 1994; Noutsos et al. 2005; Richly and Leister 2004b; Woischnik and Moraes 2002). Therefore, on this region of scaffold 8, we may have a case where almost all of the mitochondrial genome of *P. sojae* (total length 42,975 bp) was extensively rearranged and transferred to the nucleus. In recent studies of *Arabidopsis* and rice (Noutsos et al. 2005; Richly and Leister 2004b), it has been shown that dense regions of NUMTs in genomes result from random end-joining of smaller mtDNA fragments before insertion, whereas replication slippage or insertion of non-NUMT DNA may contribute further to the diversification of NUMTs.

NUMTs and functional nuclear genes in the *P. sojae* and *P. ramorum* genomes.

In *P. sojae*, we found that NUMT s38, corresponding to the 3' end of the mitochondrial *rpl2* gene, was predicted as a putative transcript by the gene-calling software used to annotate the *P. sojae* genome sequence. It has been shown that the 3' end of the mitochondrial *rpl2* gene is a functional nuclear gene in the genomes of various plant species (Adams et al. 2001a, 2002), with different stages in the transfer and activation of *rpl2* to the nucleus (whole *rpl2* gene in either mitochondrion or the nucleus, or partitioned into two fragments with each transcribed in a different cellular compartment). Therefore, NUMT s38 of *P. sojae* might be an intermediate stage during the functional transfer and activation of the mitochondrial *rpl2* gene in the nucleus. Another example is NUMT s43, where almost all of the mitochondrial *trnL* (tRNA-Leucine) gene except for eight bases at the 3' end is inserted in the nuclear genome. Transfer and functional activation of mitochondrial tRNA genes in the nucleus has been documented to greater or lesser extent in all eukaryotes (Adams and Palmer 2003).

We compared the loci where NUMTs are inserted in the *P. sojae* and *P. ramorum* genome with the loci of predicted genes in each genome, and we did not find any NUMTs within gene-coding regions. Only two cases have been reported where NUMTs were found specifically within introns of functional nuclear genes: in a tumor-suppressor gene in human (Richetti et al. 2004) and in the *sdhb* gene in rice (Kubo et al. 1999).

MATERIALS AND METHODS

Sequence comparisons.

Genomic sequences and their predicted transcripts were downloaded from the JGI. The downloaded files contained the genomic sequence scaffolds and predicted transcripts in FASTA format. As documented on the JGI website, transcripts were predicted *ab initio* from the genome sequence, using the "FGENESH" software. The size of the assembled draft sequence of *P. sojae* is 86,027,303 bp, with an estimated actual genome size of 95 Mb. The size of the assembled draft sequence of *P. ramorum* is 66,652,401 bp, with an estimated actual genome size of 65 Mb. After being downloaded, the sequences were compiled with default settings as local BLAST (Altschul et al. 1990) databases. The scaffolds are assemblies of sequence contigs connected by paired end-sequencing data. Their relative positions in the genome are not known. The complete mtDNA sequences of *P. sojae* (42,975 bp) and *P. ramorum* (39,314 bp) were obtained from the assembly of the respective genome sequences (GenBank accession numbers DQ832717 and DQ832718, respectively). The mtDNA was used in a BLAST search against all available nuclear genome sequence scaffolds or predicted transcripts, with "word size" 11, E-value cutoffs of 10 and 10⁻⁴, and all other settings as default.

Bioinformatic analysis.

In order to elucidate possible functions of the NUMTs within the *P. sojae* and *P. ramorum* genomes, we used the VBI's Microbial Genomes Database. This database, based on the genomic unified schema (GUS), is a data warehouse of the annotated genomes of *P. sojae* and *P. ramorum* (including both automated gene predictions and manually curated data) and also contains EST sequences for *P. sojae* (Tripathy et al. 2006) and *P. infestans*. The GUS database was queried directly via a PERL script containing embedded SQL commands. The "BPlite" module from the BioPerl distribution was utilized for parsing the BLAST output for the positions of the NUMTs on the genomic sequence. The PERL script performs the following steps. First, it parses the genomic positions of the discovered NUMTs from the HSPs of the alignments of mtDNA and genomic sequence. Based on those genomic positions, the script queries GUS in respect to the predicted gene regions and compares the position of the NUMT insertion with those of the gene exons and introns. This PERL script is available upon request from the authors. The output of the PERL script was verified using the VBI microbial genomes database browser.

In order to evaluate the significance of sequence matches between the mitochondrial and nuclear genome sequences, we used a PERL script to randomize the mtDNA sequences while keeping their base composition. This process included complete rearrangement of the letters in the text strings that represent the mtDNA.

The analysis and characterization of the regions of the *P. sojae* genome dense in NUMTs was automated by the creation of a bioinformatic workflow using the Taverna software tool (Oinn et al. 2004). The workflow parses the results from the BLAST comparison of the *P. sojae* mtDNA to its genome, submits the NUMTs sequences as queries to BLASTX (XML DNA DataBank of Japan, E value = 10, database "PROTEIN", all other options as standard), and retains the top hit from the results returned. Finally, the workflow outputs a summary table with the NUMT positions in the genome and their corresponding mtDNA fragments and includes the BLASTX results. This workflow is available upon request from the authors.

CLUSTALW (Thompson et al. 1994) was used for aligning small regions within each *Phytophthora* genome in order to assess whether those regions are duplicated or not. The online version of PROSCAN was used to verify whether any eukaryotic polymerase III promoters exist within the vicinity of genomic regions under consideration.

ACKNOWLEDGMENTS

The authors would like to thank S. Tripathy for invaluable help with the preparation of the SQL queries for the VBI Microbial Database, C. Volker and M. C. Chibucos of VBI for assistance with preparing and editing of the manuscript, and M. A. Saghai Maroof of Virginia Tech for useful comments and suggestions. This work was supported by grants to B. M. Tyler from the National Research Initiative of the United States Department of Agriculture Cooperative State Research, Education and Extension Service, grant number 2002-35600-12747; and the United States National Science Foundation, grant numbers MCB-0242131 and EF-0130263; and by funds from the Virginia Bioinformatics Institute. Part of this work also was performed under the auspices of the United States Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract no. DE-AC02-05CH11231.

LITERATURE CITED

Adams, K. L., and Palmer, J. D. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol. Phylogenet. Evol.* 29:380-395.
Adams, K. L., Song, K., Roessler, P. G., Nugent, J. M., Doyle, J. L., Doyle, J. J., and Palmer, J. D. 1999. Intracellular gene transfer in action:

Dual transcription and multiple silencings of nuclear and mitochondrial *cox2* genes in legumes. *Proc. Natl. Acad. Sci. U.S.A.* 96:13863-13868.
Adams, K. L., Ong, H. C., and Palmer, J. D. 2001a. Mitochondrial gene transfer in pieces: Fission of the ribosomal protein gene *rpl2* and partial or complete gene transfer to the nucleus. *Mol. Biol. Evol.* 18:2289-2297.
Adams, K. L., Rosenbluth, M., Qiu, Y. L., and Palmer, J. D. 2001b. Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. *Genetics* 158:1289-1300.
Adams, K. L., Qiu, Y. L., Stoutemyer, M., and Palmer, J. D. 2002. Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc. Natl. Acad. Sci. U.S.A.* 99:9905-9912.
Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
Antunes, A., and Ramos, M. J. 2005. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. *Genomics* 86:708-717.
Bensasson, D., Zhang, D. X., and Hewitt, G. M. 2000. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Mol. Biol. Evol.* 17:406-415.
Bensasson, D., Zhang, D. X., Hartl, D. L., and Hewitt, G. M. 2001. Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends Ecol. Evol.* 16:314-321.
Bensasson, D., Feldman, M. W., and Petrov, D. A. 2003. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J. Mol. Evol.* 57:343-354.
Blanchard, J., and Schmidt, G. W. 1995. Pervasive migration of organellar DNA to the nucleus in plants. *J. Mol. Evol.* 41:397-406.
Collura, R. V., and Stewart, C. B. 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. *Nature* 378:485-489.
Figueroa, P., Gomez, I., Carmona, R., Holuigue, L., Araya, A., and Jordana, X. 1999. The gene for mitochondrial ribosomal protein S14 has been transferred to the nucleus in *Arabidopsis thaliana*. *Mol. Gen. Genet.* 262:139-144.
Hadler, H., Devadas, K., and Mahalingam, R. 1998. Selected nuclear LINE elements with mitochondrial-DNA-like inserts are more plentiful and mobile in tumor than in normal tissue of mouse and rat. *J. Cell. Biochem.* 68:100-109.
Hazkani-Covo, E., Sorek, R., and 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.
Hu, G., and Thilly, W. G. 1995. Multi-copy nuclear pseudogenes of mitochondrial DNA reveal recent acute genetic changes in the human genome. *Curr. Genet.* 28:410-441.
Kubo, N., Kyuya, H., Atsushi, H., and Koh-Ichi, K. 1999. A single nuclear transcript encoding mitochondrial RPS14 and SDHB of rice is processed by alternative splicing: Common use of the same mitochondrial targeting signal. *Proc. Natl. Acad. Sci. U.S.A.* 96:9207-9211.
Kubo, N., Jordana, X., Ozawa, K., Zanlungo, S., Harada, K., Sasaki, T., and Kadowaki, K. 2000. Transfer of mitochondrial *rps10* gene to the nucleus in rice: Acquisition of the 5' untranslated region followed by gene duplication. *Mol. Gen. Genet.* 263:733-739.
Leister, D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet.* 21:655-663.
Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S. J. 1994. NUMT, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evol.* 2:174-190.
Mishmar, D., Ruiz-Pesini, E., Brandon, M., and Wallace, D. 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, A. J., Willerslev, E., and 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., Richly, E., and Leister, D. 2005. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Res.* 15:616-628.
Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M., R., Wipat, A., and Li, P. 2004. Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20:3045-54.
Olson, L., and Yoder, A. 2002. Using secondary structure to identify ribosomal NUMTs: Cautionary examples from the Human Genome. *Mol. Biol. Evol.* 19:93-100.
Palmer, J. D., Adams, K. L., Cho, Y., Parkinson, C. L., Qiu, Y. L., and

- Song, K. 2000. Dynamic evolution of plant mitochondrial genomes: Mobile genes and introns and highly variable mutation rates. *Proc. Natl. Acad. Sci. U.S.A.* 97:6960-6966.
- Pereira, S. L., and Baker, A. J. 2004. Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: Implications for molecular inference of population history and phylogenetics. *BMC Evol. Biol.* 25:4-17.
- Ricchetti, M., Fairhead, C., and Dujon, B. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402:96-100.
- Ricchetti, M., Tekaia, F., and Dujon, B. 2004. Continued colonization of the Human genome by mitochondrial DNA. *PLoS Biol.* 2:1313-1324.
- Richly, E., and Leister, D. 2004a. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21:1081-1084.
- Richly, E., and Leister, D. 2004b. NUMTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Mol. Biol. Evol.* 21:1972-1980.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.
- Thorsness, P. E., and Weber, E. R. 1996. Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. *Int. Rev. Cytol.* 165:207-234.
- Tourmen, Y., Baris, O., Dessen, P., Jacques, C., Malthie, Y., and Reynier, P. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 80:71-77.
- Tripathy, S., Pandey, V. N., Fang, B., Salas, F., and Tyler, B. M. 2006. VMD: A community annotation database for oomycetes and microbial genomes. *Nucleic Acids Res.* 34:379-381.
- Tyler, B. M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R. H. Y., Aerts, A., Arredondo, F., Baxter, L., Bensasson, D., Beynon, J. L., Damasceno, C. M. B., Dickerman, A., Dorrance, A. E., Dou, D., Dubchak, I., Garbelotto, M., Gijzen, M., Gordon, S., Govers, F., Grunwald, N. J., Huang, W., Ivors, K., Jones, R. W., Kamoun, S., Krampis, K., Lamour, K., Lee, M.-K., McDonald, W. H., Medina, M., Meijer, H. J. G., Nordberg, E., Maclean, D. J., Ospina-Giraldo, M. D., Morris, P. F., Phuntumart, V., Putnam, N., Rash, S., Rose, J. K. C., Sakihama, Y., Salamov, A., Savidor, A., Scheuring, C., Smith, B., Sobral, B. W. S., Terry, A., Torto-Alalibo, T., Win, J., Xu, Z., Zhang, H., Grigoriev, I., Rokhsar, D., and Boore, J. 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*. 313:1261-1266.
- Williams, S. T., and Kwonlton, K. N. 2001. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. *Mol. Biol. Evol.* 18:1484-1493.
- Woischnik, M., and Moraes, C. T. 2002. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res.* 12:885-893.
- Zhang, D. X., and Hewitt, G. M. 1996. Nuclear integrations: Challenges for mitochondrial DNA markers. *Trends Ecol. Evol.* 11:247-251.

AUTHOR-RECOMMENDED INTERNET RESOURCES

BioPerl distribution: www.bioperl.org
 A census of nuclear mitochondrial pseudogenes (numts):
www.pseudogene.net
 DNA DataBank of Japan (DDJP): www.xml.nig.ac.jp/index.html
 Genomic Unified Schema browser: www.gusdb.org
 Joint Genome Institute (JGI) Eukaryotic Genomics database:
www.genome.jgi-psf.org/euk_home.html
 National Institutes of Health, Center for Information Technology's
 PROSCAN: thr.cit.nih.gov/molbio/proscan
 The Virginia Bioinformatics Institute (VBI) microbial database:
phytophthora.vbi.vt.edu