Supporting Information

Lang et al. 10.1073/pnas.1322190111

SI Materials and Methods

Yeast Strains and Culture Conditions. Yeast strains (Table S1) were obtained from C. P. Kurtzman and J. Swezey (National Center for Agricultural Utilization Research, Peoria, IL) and R. A. Zvyagilskaya (Bach Institute of Biochemistry, Moscow, Russia). Yeast cells were cultivated at 28 °C in liquid YM [0.3% (wt/vol) yeast extract (Difco), 0.3% (wt/vol) malt extract, 0.5% (wt/vol) peptone (Difco)] or YP [1% (wt/vol) yeast extract (Difco), 2% (wt/vol) peptone (Difco)] media containing either 1% (wt/vol) glucose or 2% (wt/vol) galactose as the sole carbon source. For substrate utilization test, yeasts were grown on synthetic medium [0.67% (wt/vol) Yeast Nitrogen Base without amino acids (Difco), 2% (wt/vol) agar (Difco)] with either 1% (wt/vol) glucose, 3% (vol/vol) glycerol, 3% (vol/vol) ethanol, or, alternatively, 2% (vol/vol) lactate, as the sole carbon source.

DNA Purification and Sequencing. Mitochondrial DNA from the *Magnusiomyces magnusii* strain 270 was prepared from DNase I-treated mitochondria as described previously (1), following which a shotgun library of mtDNA fragments was constructed and sequenced by the dideoxy-chain termination method on double-stranded plasmid templates (2). For the other magnusiomycete strains, total cellular DNA was isolated essentially as described earlier (3), purified using DNeasy Blood and Tissue kit (Qiagen), and sequenced, along with nuclear DNA, by Macrogen using Illumina HiSeq 2000 and paired-end (2 × 100 nt) technology. About 4–13 Gbp of raw data (QV20 > 95%) was obtained per sample, and low-quality reads were filtered and trimmed with tools developed in-house and assembled with Velvet (version 1.1.06) using a k-mer size of 81 (*Magnusiomyces capitatus*) or 61 (remaining species) (4). Contigs corresponding to mtDNAs were manually curated using the Geneious package v5.6.6 (Biomatters). Read mapping to assembled contigs was performed with Geneious at standard settings (medium sensitivity/read mapping).

Mitochondrial Genome Annotation. Gene annotation of *Magnusiomyces* mtDNAs was performed with the automated tool MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl) developed in-house. In brief, MFannot predicts group I and II introns, tRNAs, RNase P-RNA, and 5S rRNA with Erpin as a search engine (5), based on RNA structural profiles established by us. Exons of protein-coding genes are inferred in a first round with Exonerate (6), and then for less well-conserved genes with HMMER (7), based on models for all known mtDNA-encoded proteins. Only sequence positions that are aligned with confidence are retained for model construction. Miniexons (as short as 3 nt) that are not resolved by Exonerate, but inferred by the presence of orphan introns, are detected as missing protein regions in multiple protein alignments. The precise placement of small exons is based on the best fit of HMM protein profiles and on the fit with conserved nucleotide sequence profiles of group I or II exon–intron boundaries. Genes encoding the small and large subunit rRNAs are predicted with HMM profiles covering the most highly conserved domains, allowing precise placement of the small subunit rRNA termini, but only approximate positioning of large subunit rRNA ends. The latter termini, as well as precise exon–intron boundaries of rRNA genes, are predicted manually using comparative-structure modeling and experimentally determined processing sites in *M. magnusii* strain 270 (see *Mapping of rRNA Termini in M. magnusii*). Automated annotations are complemented by manual analyses to account for MFannot warnings (e.g., potential transspliced genes, gene fusions, frameshifts, alternative translation initiation sites, failure to identify mini-introns, etc.), correct potential errors, and find features that are not (yet) recognized by automated procedures. In the case of *M. capitatus*, byp insertions were identified with TFASTA (8), by using protein sequences from other *Magnusiomyces* strains without byps. Only insertions that follow the conserved byp features (matching takeoff and landing sites; avoided codon and RNA hairpin structure following the takeoff site; and reconstitution of bona fide translation of a functional protein sequence without insertion) were annotated as such.

***LacZ* Constructs of Byps and Expression in *Escherichia coli*.** The elements *nad4L*-byp1 and *cox1*-byp1 and mutated versions thereof were directionally inserted into the *lacZ* gene of the *E. coli* cloning vector pUC19. Briefly, PCR primer pairs where designed whose 3′ portions anneal with the vector's multicloning site in divergent orientation and whose 5′ portions cover the 3′ half and the 5′ half, respectively, of the byp element. Primer sequences (5′ to 3′) are as follows: primer puc19-a (control), TCACTGGCCGTCGTTTTACAACG; primer puc19-b (control), ATTCGAGCTCGGTACCCGG; primer puc19-1 (landing), ATCCCTAATAGGGGATTATTATTATAAAAATTTTAAAGGTTCACTGGCCGTCGTTTTACAACG; primer puc19-2 (take-off), TATTATCGTCCTATTAAAAATAATAATAAAAAAATTGTTACCATATTCGAGCTCGGTACCCGG; Primer puc19-3 (CGA -> TGA), TATTATCATCCTATTAAAAATAATAATAAAAAAATTGTTACCATATTCGAGCTCGGTACCCGG; primer puc19-4 (CGA-> TGA, no prepeptide), TATTATCATCCATTCGAGCTCGGTACCCGG; and primer puc19-5 (tetraloop, SD), ATCCCCTTCGGGGGATTATTATTAAAAGAGATTTAAAGGTTCACTGGCCGTCGTTTTACAACG.

Primer pairs were puc19-a plus puc19-b; puc19-1 plus puc19-2; puc19-1 plus puc19-3; puc-19-1 plus Puc19-4; and puc19-5 plus puc19-3. The vector served as amplification template. Ligation of the amplicon generates a plasmid with the byp element inserted in the multicloning site. The control experiment used primers without byp-sequence additions. The ligation protocol involves end repair with a mixture of 2 U each of Klenow DNA polymerase and T7 DNA polymerase, and ligation with 0.6 units T4 DNA Ligase in the presence of 5 U T4 polynucleotide kinase. Competent *E. coli* cells were transformed with ligation products and incubated on agar plates containing isopropyl β-D-1-thiogalactopyranoside (IPTG) and Xgal (9). Purified plasmids isolated from both blue and white recombinant colonies were analyzed by Sanger sequencing.

In Silico Search for Byps. Byp sequences were aligned with Muscle version 3.6 (10). The resulting multiple alignment was manually curated for optimal fit of primary sequence and secondary structure predicted by RNAalifold (11). For manual curation, visualization, editing, and reformatting of sequence alignments, we used the GDE sequence editor (12). A modified GDE version that works in current 64-bit Linux environments, together with the appropriate libraries, is available on request. Byps were searched for with covariance models (CMs), based on the alignment of UCC and CGA byps in regular mitochondrial protein genes of *M. capitatus*. To improve search sensitivity, two models per byp group were constructed, using the cmbuild and cmcalibrate tools that are included in the Infernal package (version 1.1rc1) (7). For a given byp group, the first CM contains all sequences, and the second contains those byps that are not found by the first model

(note that byps are short and poorly conserved at the sequence level). Only confidently aligned nucleotides were used for model building and searching, by applying the –hand option. Searches were performed with a cutoff E-value of 0.01.

Mitochondrial RNA Isolation and Transcript Analysis by RT-PCR. Mitochondria were isolated from a 1-L culture of *M. capitatus* cells grown in YM with 2% (wt/vol) galactose at 28 °C as described in ref. 13 and further purified by a flotation gradient (14), with the modifications described earlier (15). RNA was then extracted using the RNeasy Mini kit (Qiagen) treated with DNase I, and used for cDNA synthesis using the RevertAid H Minus First Strand cDNA synthesis kit (Thermo Scientific) and random hexamer primers. The cDNA was used as a template in PCR reactions using gene-specific primer pairs: for *cob* (primer sequences are given in 5′ to 3′ direction), ATGGCATTACGTAAAAGAATGAA and CTATTTACTTTGTTGTCCAATATAGAA; for *cox1*, ATGGTAAATAGTAAATTTAAACAA and TTATGATTGTAAAGCTGGACT; for *nad2*, ATGTTAGTATTAGGTACATTAA and AAAGTATTGTTAAAATTAAACAT; for *nad3*, ATGTTTAATTTTAACAATACTTT and TTATGAATATAAACTTAAATAATGAT; and for *rps3*, ATGAAAAGAGAAATATTAAAATC and TTATAAATGACCTAATTGAATTT. Primers anneal with the 5′ and 3′ regions of genes. The reactions contained 0.3 μL of cDNA, 0.5 μM upstream and downstream primers, 0.2 mM dNTPs each, 1× reaction buffer containing 2 mM $MgCl_2$, and 1 U of *DreamTaq* DNA polymerase (Fermentas), and were performed using the following cycler profiles: for *cob* [3 min at 95 °C; 29× (45 s at 94 °C, 45 s at 56 °C, 4 min at 72 °C); 3 min at 72 °C]; *cox1* [3 min at 95 °C; 29× (45 s at 94 °C, 45 s at 52 °C, 2 min at 72 °C); 3 min at 72 °C]; *nad2* [3 min at 95 °C; 29× (45 s at 94 °C, 45 s at 49 °C, 2 min at 65 °C); 3 min at 65 °C]; *nad3* [3 min at 95 °C; 29× (45 s at 94 °C, 45 s at 50 °C, 45 s at 72 °C); 3 min at 72 °C]; and *rps3* [3 min at 95 °C; 29× (45 s at 94 °C, 45 s at 52 °C, 1.5 min at 65 °C); 3 min at 65 °C]. Three control PCR reactions were performed using the following templates: (*i*) genomic DNA (positive control), (*ii*) mitochondrial RNA treated with DNase I, and (*iii*) cDNA prepared from mitochondrial RNA treated with DNase I and RNase A (negative control). PCR products were electrophoretically separated in a 1% agarose gel. RT-PCR products of *cob* (1.20 kbp) and *cox1* (1.71 kbp) were sequenced. The other RT-PCR products, which cover regions without introns, were identified by size, which was identical with the size of the PCR product amplified from the genomic DNA: i.e., 2.10 kbp (*nad2*), 0.49 kbp (*nad3*), and 1.35 kbp (*rps3*). Equal size indicates that the byp elements were not spliced out of the mRNA.

Mapping of rRNA Termini in *M. magnusii*. The termini of mitochondrial large and small rRNAs of the *M. magnusii* strain 270 were mapped using an RNA circularization procedure (16). The regions containing ligated rRNA termini were amplified by RT-PCR using the following primer pairs (primer sequences are given in 5′ to 3′ direction): GATTATTCCCACCTACTACCT, TCCTCTAATGAACAAAATCC (*rns*); and ATCTTGGTTATTTTCCTTTCCTTA, CGACTCTACTTATCCTACTGGTG (*rnl*). Amplicons were cloned using the PCR cloning kit (Qiagen) and sequenced. Sequences of the identified rRNA termini are conserved across the analyzed Magnusiomycetes.

Identification of Mitochondrial Proteins by Mass Spectrometry. Mitochondria from a 1-L yeast culture (*M. capitatus* and *Saccharomyces cerevisiae* FY1679; grown on YM medium with galactose; see *Yeast Strains and Culture Conditions*) were solubilized in a buffer containing 20 mM Hepes/KOH, (pH 7.4), 60 mM NH_4Cl, 10 mM $MgCl_2$, 0.5 mM EDTA, 1 mM PMSF, and digitonin [2% (wt/wt) protein], followed by incubation on ice for 30 min and homogenization in a Potter homogenizer. After centrifugation at 18,000 × *g* for 15 min, the supernatant was collected, and ∼150 μg of protein was electrophoretically separated for 30 min (4–14% Blue Native Poly-Acrylamide Gel Electrophoresis; Hoefer apparatus with an 18 × 16-cm electrophoresis chamber; 140 V and 9 mA). The preparation of BN-PAGE gels, electrophoresis buffer, and samples followed previously published procedures (17). The protein-containing zone was cut out of the gel and submitted to a proteomics service (Institute for Research in Immunology and Cancer technology platform, Université de Montréal) for destaining, reduction, alkylation, tryptic digestion, liquid chromatography tandem mass spectrometry (LC-MS/MS) (18, 19), and peptide annotation by Mascot (20).

1. Griac P, Nosek J (1993) Mitochondrial DNA of *Endomyces (Dipodascus) magnusii*. *Curr Genet* 23(5-6):549–552.
2. Burger G, Lavrov DV, Forget L, Lang BF (2007) Sequencing complete mitochondrial and plastid genomes. *Nat Protoc* 2(3):603–614.
3. Philippsen P, Stotz A, Scherf C (1991) DNA of Saccharomyces cerevisiae. *Methods Enzymol* 194:169–182.
4. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821–829.
5. Gautheret D, Lambert A (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol* 313(5): 1003–1011.
6. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
7. Eddy SR (2008) HMMER: Biological Sequence Analysis Using Profile Hidden Markov Models, Version 3.0. Available at http://hmmer.janelia.org.
8. Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* 132:185–219.
9. Sambrook J, Fritsch EF, Maniatis T (1989) *Molceular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY), 2nd Ed.
10. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
11. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9:474.
12. Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994) The genetic data environment an expandable GUI for multiple sequence analysis. *Comput Appl Biosci* 10(6):671–675.
13. Tomáska L, Nosek J, Fukuhara H (1997) Identification of a putative mitochondrial telomere-binding protein of the yeast *Candida parapsilosis*. *J Biol Chem* 272(5):3049–3056.
14. Lambowitz AM (1979) Preparation and analysis of mitochondrial ribosomes. *Methods Enzymol* 59:421–433.
15. Newman SM, Zelenaya-Troitskaya O, Perlman PS, Butow RA (1996) Analysis of mitochondrial DNA nucleoids in wild-type and a mutant strain of *Saccharomyces cerevisiae* that lacks the mitochondrial HMG box protein Abf2p. *Nucleic Acids Res* 24(2):386–393.
16. Yokobori S, Pääbo S (1995) Transfer RNA editing in land snail mitochondria. *Proc Natl Acad Sci USA* 92(22):10432–10435.
17. Daoud R, Forget L, Lang BF (2012) Yeast mitochondrial RNase P, RNase Z and the RNA degradosome are part of a stable supercomplex. *Nucleic Acids Res* 40(4):1728–1736.
18. Wessels HJ, et al. (2009) LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. *Proteomics* 9(17):4221–4228.
19. Fandiño AS, et al. (2005) LC-nanospray-MS/MS analysis of hydrophobic proteins from membrane protein complexes isolated by blue-native electrophoresis. *J Mass Spectrom* 40(9):1223–1231.
20. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18): 3551–3567.

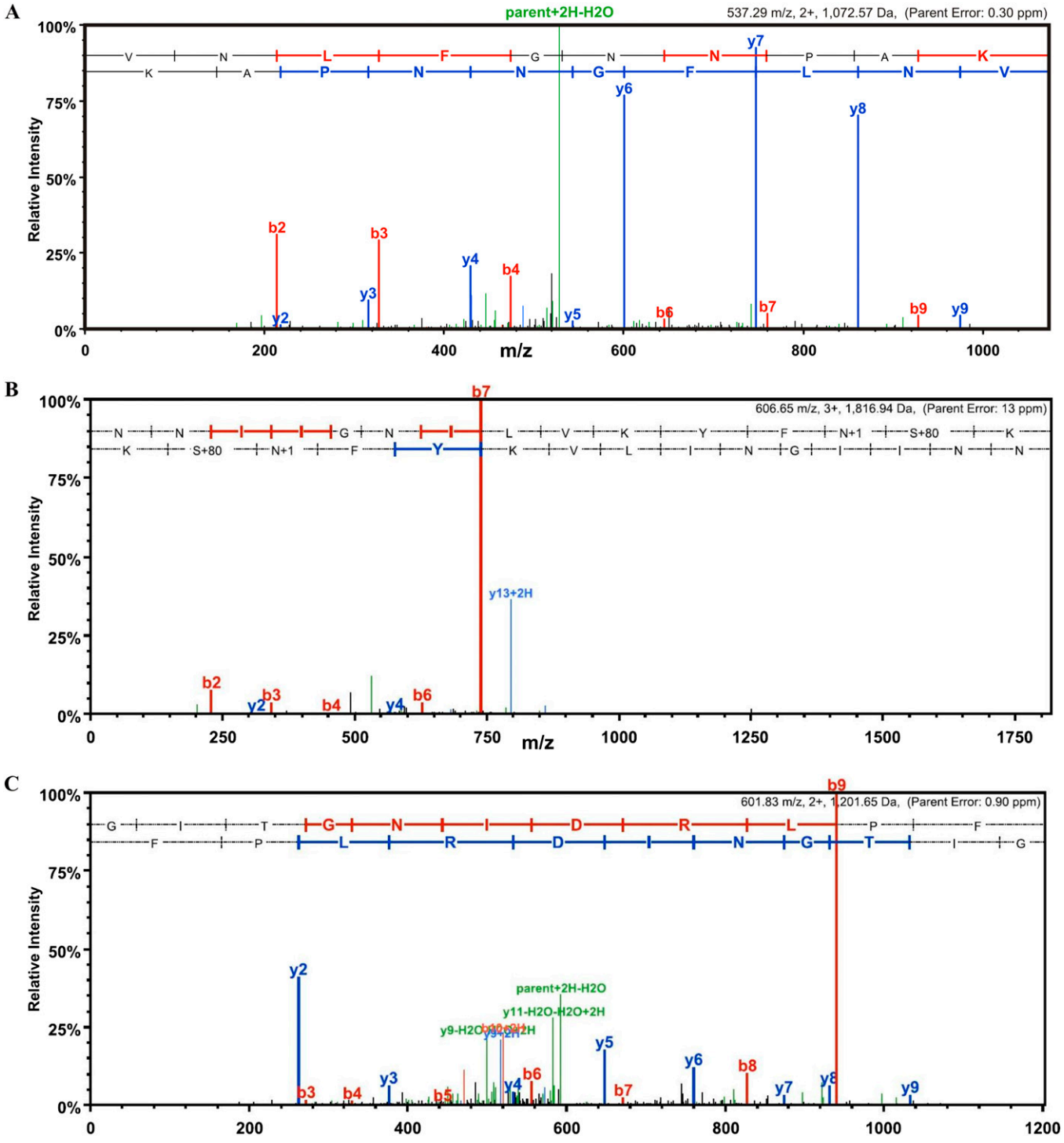

Fig. S1. MS/MS spectra of peptides spanning byp elements. Spectra correspond to tryptic peptides of (*A*) Nad2 (VNLFGNNPAK), (*B*) Rps3 (NNIIGNILVK), and (*C*) Cob (GITGNIDRLPF). Assignment of sequence-specific MS/MS signatures is indicated by fragmentograms. Red and blue bars indicate b and y ion intensities, respectively. The spectra were manually reviewed using Scaffold4 (http://proteomesoftware.com).

```
rps3-byp1                                                                MKREILKSIV...
nad4L-byp1                                                              MVTIFLLLFLIG...
nad4L-byp2                                                             MVTIFLLLFLIGLIG...
nad2-byp1                                                            MLVLGTLILILSTFNLKTIKE...
nad3-byp1                                                            MFNFNNTLYIFIILIPIVGLA...
rps3-byp2                                                          MKREILKSIVIKLIWTKIYQIKE...
nad6-byp1                                                        MWLIELINTHFQMNNTIIILFEFLSILSA...
rps3-byp3                                                       MKREILKSIVIKLIWTKIYQIKEIINKNLG...
nad4-byp1                                                 MGLYIIILGFLSIIGINNKDIITSIWPKLIRQLSLILSV...
cox1-byp1                                                 MVNSKFKQFVTRWLFSTNCKDIAVLYMIFAIFSGLIGTG...
cox3-byp1                                    MNNKSNIMLNIYRQNYQLHPFHLVENSPWPFFSSFSLFGLAMNTALTSHGYIQSS...
nad5-byp1                           MIFITIFLPILGSILSGLLSRSIGSRSYTISWMTPNPKWSGIIATSTIVIAALFSYYLYFDVIV...
```

Fig. S2. Inferred nascent peptide sequences immediately upstream of byp takeoff sites. Shown are sequences inferred from the regions upstream of the most 5′ byp in mtDNA-encoded protein genes of *M. capitatus*. The sequences start at the predicted translation initiation codon and end with the takeoff codon of the corresponding byp. The nascent peptides upstream of byps do not have recognizable patterns of positively charged amino acids or other positional motifs conserved throughout prepeptide sequences.

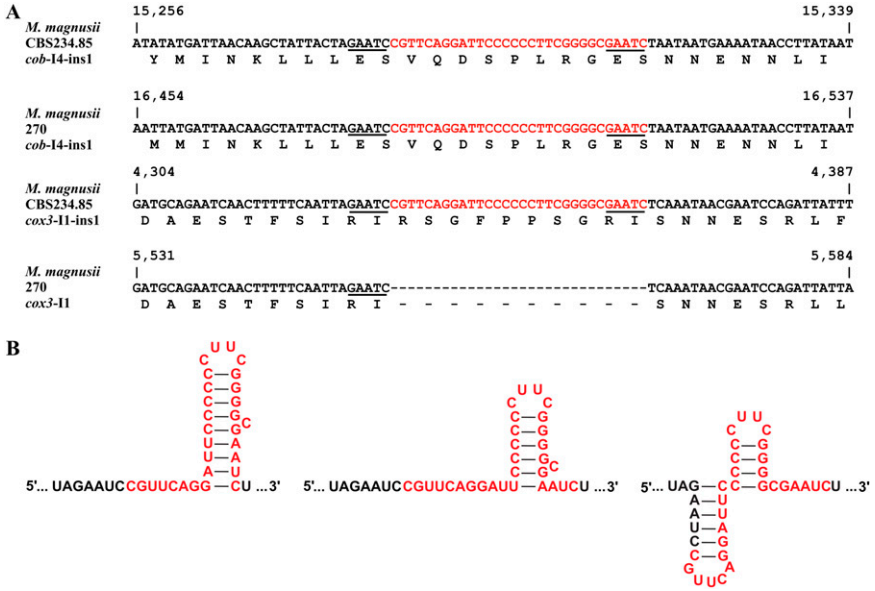

Fig. S3. Byp-like insertions in intron ORFs of *M. magnusii* strains. CM search of byps in the mtDNA sequences of *M. magnusii* strains CBS234.85 and 270 revealed several byp-like elements in intron ORFs. (*A*) Sequence alignment of intron ORFs containing identical byp-like inserts that do not introduce frameshifts or stop codons. For comparison, the *M. magnusii* 270 *cox3*-I1 ORF sequence without a byp-like element is shown, delineating the insertions. Note that *cob*-I4-ins1 and *cox3*-I1-ins1 have identical nucleotide sequences, but that the elements are in different reading frames and therefore translated into different amino acid sequences. A flanking upstream pentanucleotide motif "GAATC" (underlined) is duplicated at the 3′ end of all three byp-like insertions. An analysis of codon use of *M. magnusii* genes indicates that this species has by far a less extreme codon avoidance (Table S7) than *M. capitatus* (Table S3) and that these elements may be translated, although less effectively because several codons are used only rarely. Only strain CBS 234.85, but not strain 270, contains an element in the *cox3*-I1 ORF. Therefore, we assume that this element results from a recent transposition event. (*B*) Alternative RNA folding of the byp-like sequence (red) and flanking nucleotides (black) from *cob*-I4-ins1 and *cox3*-I1-ins1 of the above *M. magnusii* strains. Note that the variant with a double hairpin structure closely resembles double hairpin elements, which are present in fungal mitochondrial genomes, including *Allomyces macrogynus* (1) and *Schizosaccharomyces octosporus* (2).

1. Paquin B, Laforest MJ, Lang BF (2000) Double-hairpin elements in the mitochondrial DNA of *Allomyces*: Evidence for mobility. *Mol Biol Evol* 17(11):1760–1768.
2. Bullerwell CE, Leigh J, Forget L, Lang BF (2003) A comparison of three fission yeast mitochondrial genomes. *Nucleic Acids Res* 31(2):759–768.

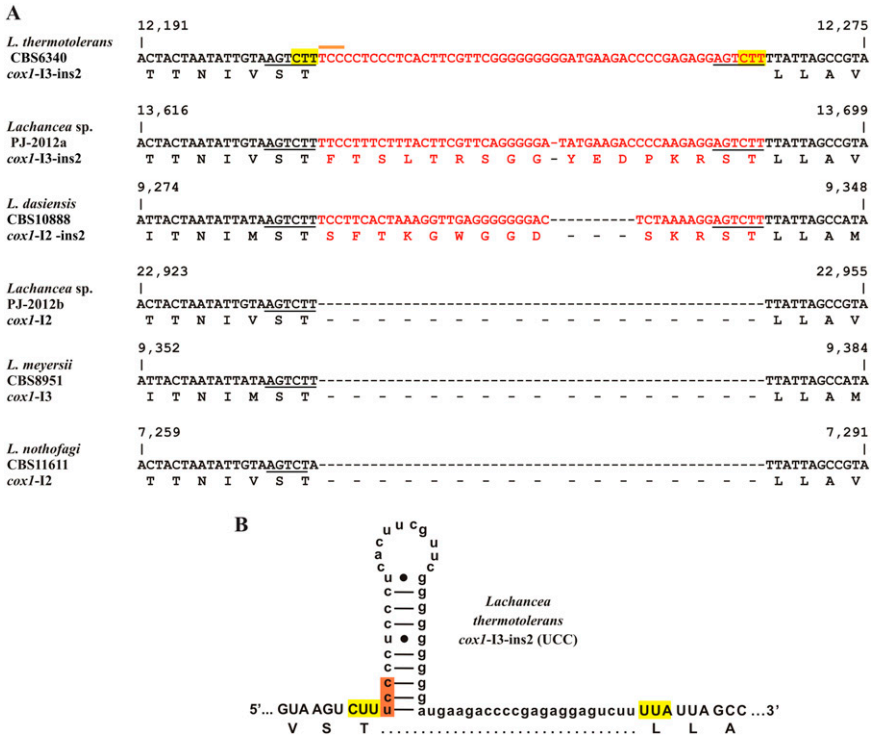

Fig. S4. Variable occurrence of byp-like elements in the same position of *cox1* intron 3 ORF across *Lachancea* species. (*A*) Sequence alignments. The insert in *Lachancea thermotolerans* most closely resembles a byp because it causes a frameshift and has a potential takeoff codon (CTT; labeled yellow), followed by an avoided TCC and a landing triplet that matches the takeoff codon (CTT; labeled yellow). Inserts in *Lachancea* sp. PJ-2012a and *Lachancea dasiensis* CBS1088 resemble byps as well but allow read-through and are substantially different at both the nucleotide and amino acid levels. TCC is an unused codon in mitochondria of all *Lachancea* species. A flanking upstream hexanucleotide motif "AGTCTT" (underlined) is duplicated at the 3′ end of all byp-like elements. (*B*) Secondary structure diagram of the byp-like element in *L. thermotolerans* that closely resembles that of byps in *M. capitatus*.

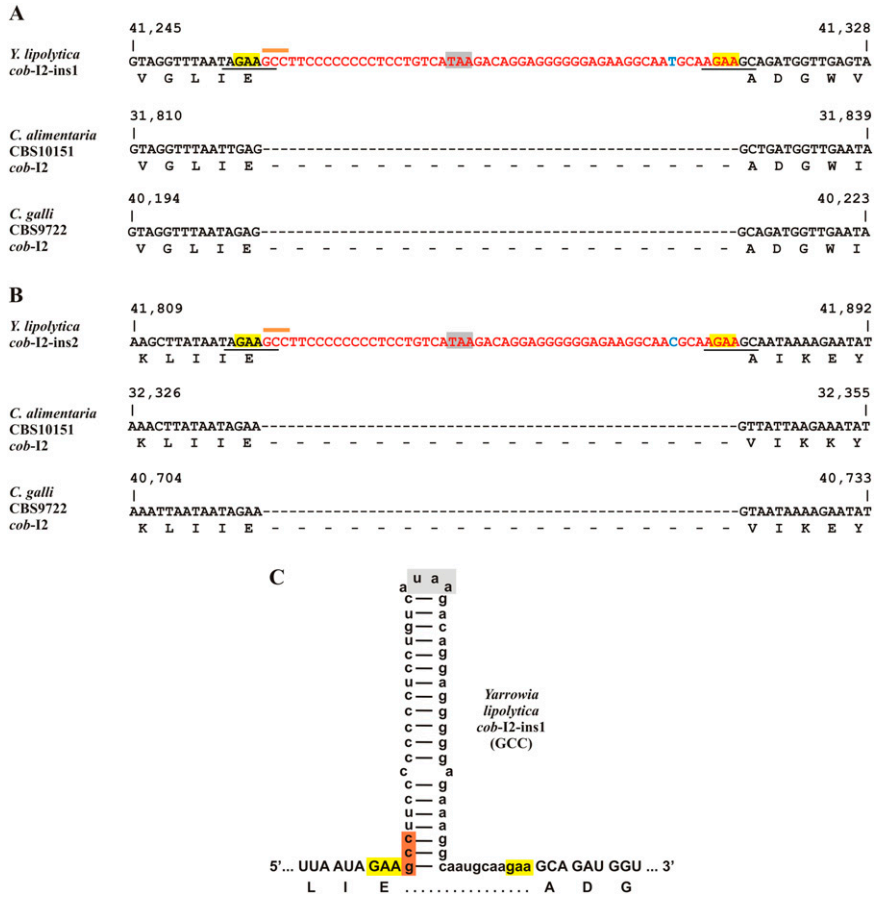

Fig. S5. Byp-like elements in the *cob* intron ORF of *Yarrowia*. Elements are missing in closely related taxa. Nucleotides in red indicate byp-like elements, yellow shading highlights putative takeoff and landing codons, and the rare codon GCC is marked in orange. (*A* and *B*) Sequence alignments around the insertion points of two elements. Underlined hexanucleotides are duplicated at the end of the elements. (*C*) Secondary structure model of the first element in *Yarrowia lipolytica cob*-intron 2 ORF, which closely resembles those of *M. capitatus* byps (see Fig. 2 *A*–*E*).

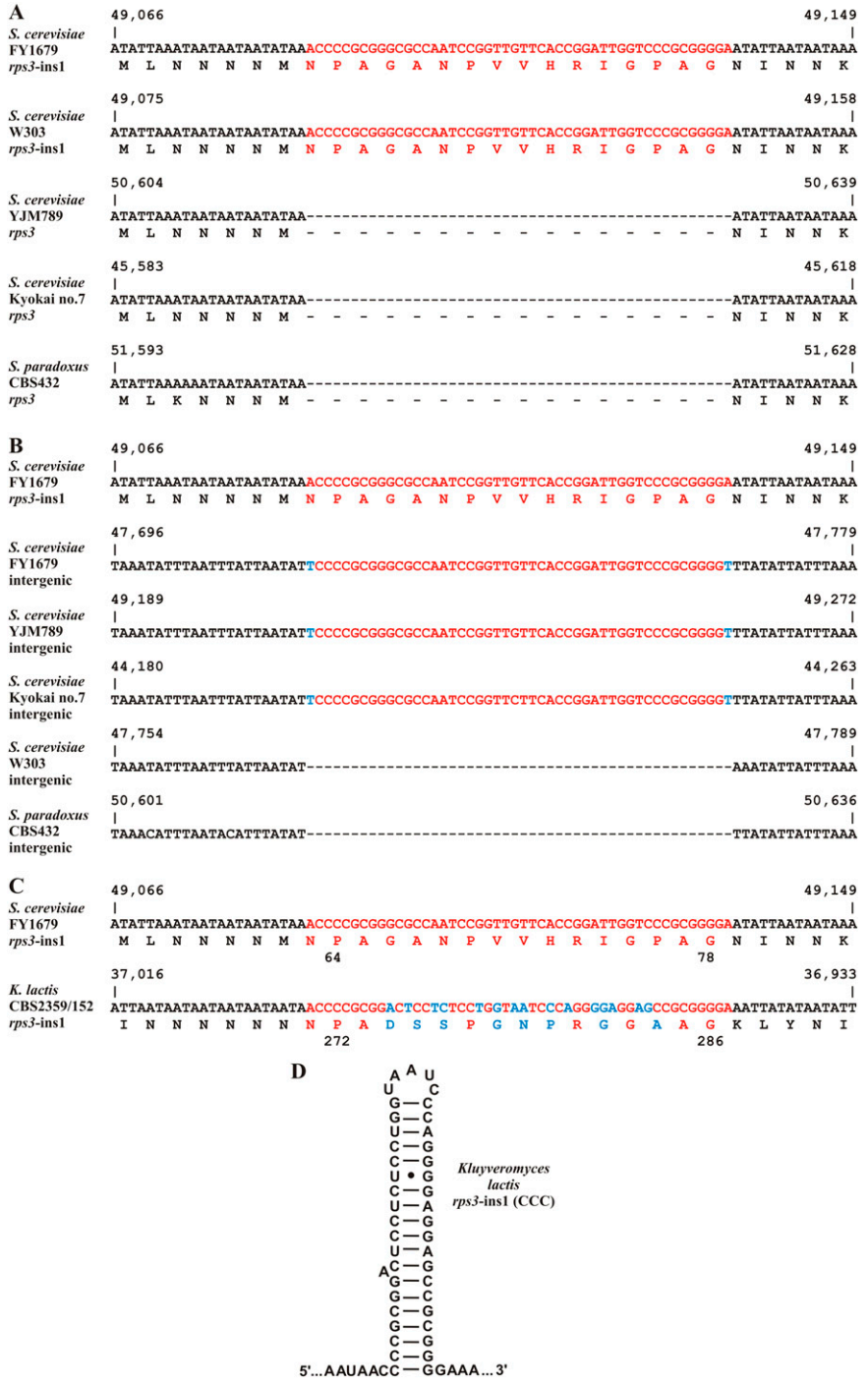

Fig. S6. Byp-like elements in *rps3* and intergenic regions across Saccharomycetales. Nucleotides in red indicate byp-like elements. (*A*) A byp-like insert present in *rps3* of the *S. cerevisiae* reference strain FY1679 and strain W303 is absent in three other, closely related isolates. (*B*) The element in *rps3* of *S. cerevisiae* FY1679 has an almost identical copy in the intergenic region between *trnS1* and *atp9* of the same mtDNA (two sequence differences marked blue). Some *S. cerevisiae* isolates carry the same byp-like element in the corresponding intergenic region whereas others lack an insert in this position. (*C*) Comparison of a byp-like element in *rps3* from *S. cerevisiae* and *Kluyveromyces lactis*. Nucleotides and amino acids in blue highlight positions that differ between the two species. Note that the elements are quite similar at the nucleotide and deduced amino acid levels, and both are inserted in-frame but at different positions of the gene. (*D*) The secondary structure model of the element in *K. lactis rps3* is highly similar to those of *M. capitatus* byps (see Fig. 2 *A*–*E*).

Table S1. *Magnusiomyces* strains used in this study

Species	Strain	Origin* (year; country; substrate)	mtDNA size, bp	G+C content, %	Noncoding intergenic sequence, %	No. of mapped reads	GenBank acc. no.
M. capitatus	CBS 197.35 (NRRL Y-17686)	1935; Sweden; wood pulp	43,486	24.2	21.8	8,920,467	KJ459952
M. ingens	CBS 521.90 (NRRL Y-17630)	1990; South Africa; wine cellar	37,684	20.3	16.9	6,918,467	KJ459950
M. magnusii	CBS 234.85	1985; Pennsylvania; slime flux in *Quercus alba*	42,757	38.2	13.3	10,852,529	JQ236859
M. magnusii	270	Unknown	40,657	38.2	14.0	n/a	KJ459953
M. tetrasperma	CBS 765.70 (NRRL Y-7288)	1970; California; wet conveyor at a prune dehydration plant	44,469	48.7	40.0	4,596,067	KJ459951

*Fungal Biodiversity Centre at www.cbs.knaw.nl/. *M. ingens*, *Magnusiomyces ingens*; *M. tetrasperma*, *Magnusiomyces tetrasperma*.

Table S2. Byps in mitochondrial genes and intron ORFs of *Magnusiomyces capitatus*

Byp location	Byp name	Frameshift	Stop codons	Size, nt	Hairpin	Takeoff codon	Avoided codon	Landing codon	Genomic sequence [TAKEOFF CODON/byp sequence including landing codon/NEXT TRANSLATED CODON]
Genes									
atp6	atp6-byp1^	—	—	33	+	GGA	UCC	GGU	GGAtccccccaactatggtggggtatattcactggtATT
cob	cob-byp1	+	—	41	+	GGA	CGA	GGA	GGAcgatcacctcttatagaggtgattaaaatttaaattggaAAT
cox1	cox1-byp1	—	+	36	+	GGU	CGG	GGA	GGTcggcccctctttggggccgctaaccccacctcggaTTA
	cox1-byp2	+	—	41	+	GGA	CGA	GGA	GGAcgatccaactttgagttggattgttaaattttaaaatggaTTA
cox3	cox3-byp1	+	—	38	+	AGC	CGA	AGU	AGCcgattcacctttatgaggtggattgtttttactttagtATT
	cox3-byp2	—	+	45	+	GGA	CGA	GGA	GGAcgatccaactcttagggagttggattataaaaataaatattaggaATT
	cox3-byp3	—	+	33	+	GGA	UCC	GGA	GGAtccccccttatttgggggtatctcctaaggaTTC
nad1	nad1-byp1	+	+	34	+	GGA	UCC	GGA	GGAtccatctcttagtagatggaaaataattataggaATT
	nad1-byp2	+	—	37	+	GUA	UCC	GUA	GTAtccacccttttttagggtgggatattttatcttgtaAGA
	nad1-byp3	+	—	40	+	UUC	UCC	UUC	TTCtccccctttttagggggtggatgattacctaaaaattcACT
nad2	nad2-byp1	+	+	34	+	GAA	UCC	GAA	GAAtccccacataatagtgggggataatcctcctgaaGTT
	nad2-byp2	+	+	32	+	GAU	CGA	GAU	GATcgaccatttatttaaatggtcataaaaaagatTTT
	nad2-byp3*	+	—	50	+	GCU	UCC	GUU	GCTtccccccccatctcttaaggtgatgtgggaagtttaaaaattataagttAAA
	nad2-byp4	—	+	36	+	GAG	UCC	GAA	GAGtccattatggttaaggccataatggttataatgaaTTA
	nad2-byp5	+	—	37	+	GGG	UCC	GGU	GGGtccccccccctctagggtgggaaatctaaaatggtTTA
	nad2-byp6	—	+	42	+	GGC	UCC	GGU	GGCtcccatatccatattacggatatgagagtttaaataaaaggtTCA
	nad2-byp7	+	—	37	+	GGC	CUC	GGU	GGCctccccctttttagggggagggcaatatccaaggtTTA
	nad2-byp8	+	—	38	+	GGA	CGA	GGU	GGAcgattcacctttataggtgaatttgttatttataaggtTAT
	nad2-byp9	+	+	35	+	GUU	UCC	GUA	GTTtccccccttttttaaaaaggggattataatagtaAAT
	nad2-byp10	—	+	48	+	GGA	CGA	GGU	GGAcgattatccctttatatagggataatttaataataattttaattaggtTTA
	nad2-byp11^	—	—	39	+	GGA	CGA	GGA	GGAcgaccatctttgtagatggtcgtataaaaaaattttggaTTA
nad3	nad3-byp1	+	+	49	+	GCU	UCC	GCA	GCTtccctcttttttattaagtaggggattttaaacttttttttagaatgcaTTA
	nad3-byp2	+	—	40	+	GAA	CGA	GAA	GAAcgattaatccctttaaggggattattcttttttatttgaaGTT
nad4	nad4-byp1	—	+	36	+	GUG	UCG	GUA	GTGtcgccatttttataatggcaaataaaaaatttcagtaATA
	nad4-byp2	—	+	39	+	AGU	CGA	AGU	AGTcgacttaacctaataaggggttaagttgaataaaaaagtATC
	nad4-byp3	—	+	30	+	GGA	UCC	GGA	GGAtccccccttaggggggattttccttaggaATA
	nad4-byp4	+	—	44	+	GGA	CGA	GGA	GGAcgaccaactcttatagagttggattgtaaaaataaaataaaggaATA
	nad4-byp5	+	—	32	+	GGU	UCC	GGU	GGTtccccccctaatgggggggaacaccccaatggtATT
	nad4-byp6	—	+	42	+	GGA	CGA	GGU	GGAcgataatcccttttataaggggggattatctttttatttatggtTTA
	nad4-byp7^	—	—	48	+	GGU	CGA	GGA	GGTcgaccatctttttttaagatggttattccacctttttattctttattggaTTA
nad4L	nad4L-byp1	—	+	48	+	GGA	CGA	GGU	GGAcgataataatccctaataggggattattattataaaaatttttaaaggtTTA
	nad4L-byp2	—	+	42	+	GGU	CGA	GGA	GGTcgacttaatttttaagtcaatcaaaaaattatatataaaggaTTT
	nad4L-byp3	+	—	35	+	GGU	CGA	GGU	GGTcgaatccaattagttggattcgtaaaaaaaatggtTTA
nad5	nad5-byp1	—	+	36	+	GUU	UCC	GUA	GTTtcccatctttttttgaagatggaaataaaattaagtaAAT
	nad5-byp2	+	+	50	+	GGA	CGA	GGU	GGAcgattaatccaccctaagagggtgttgattaattatatataatttaaggtGAT
	nad5-byp3	—	+	42	+	GGU	CGA	GGA	GGTcgatcatcttttttaagatgatcgatataaaaaataaaaggaTTA
	nad5-byp4	+	—	43	+	GCU	UCC	GCA	GCTtcctccacctttttaaaaggtggattgaatttatttttaagcaACA
	nad5-byp5	+	—	38	+	GGA	UCC	GGA	GGAtccccccatttgatgggggggataaattatctcaggaTTA
	nad5-byp6	—	+	39	+	GGU	UCC	GGU	GGTtccacttatttattttaagtggtaataaaattataaggtATA
	nad5-byp7	+	—	38	+	GGA	UCC	GGU	GGAtccactatttttttagtgggttgacatttttttatggtGTT
	nad5-byp8	+	—	41	+	GGU	CGA	GGU	GGTcgatataaccccttttaaaggggttatatctttttatttggtTTT
	nad5-byp9	+	—	34	+	GGA	UCC	GGU	GGAtccccctctattttagagggtatttaatttggtTTT
	nad5-byp10	+	+	34	+	AGU	UCC	AGU	AGTtccccccctttttttaaaggggggattaaaaatagtATC

Table S2. Cont.

Byp location	Byp name	Frameshift	Stop codons	Size, nt	Hairpin	Takeoff codon	Avoided codon	Landing codon	Genomic sequence [TAKEOFF CODON/byp sequence including landing codon/NEXT TRANSLATED CODON]
	nad5-byp11	+	+	32	+	GGA	UCC	GGU	GGAtccccccactattgtggggatcttcataggtACT
	nad5-byp12	+	—	37	+	GCU	UCC	GCU	GCTtccccacccttttttttagagggtgtgattttacttgctATT
nad6	nad6-byp1	+	—	40	+	GCA	UCC	GCA	GCAtcctcctatttatttaggagatggtggataaaaataagcaATT
	nad6-byp2	+	+	41	+	GGA	CGC	GGA	GGAcgccaaccttaaagtagggttggttaaataaaaaatagga TTA
	nad6-byp3	—	+	51	+	GGA	CGA	GGA	GGAcgattaatcactccatagggagtgattaattaaatattaaaaataaaggaTTA
	nad6-byp4	+	+	55	+	GAA	CGA	GAA	GAAcgataaccagtttaactggttattattttattcttacttattagatataaatgaaTTA
	nad6-byp5	+	+	46	+	GCU	CGA	GCA	GCTcgattcacccctttctaaagggaggtgtgaattaaatttataatgcaATA
rps3	rps3-byp1	—	+	33	+	GUA	CGA	GUU	GTAcgattccctaatataaaagggaataaataagttATT
	rps3-byp2	—	+	30	+	GAA	CGA	GAA	GAAcgaggttaatttaacctcaaataatttgaaATA
	rps3-byp3	+	—	38	+	GGA	UCC	GGU	GGAtccccccccctctgggtgggggttatctccagcaggtTCA
	rps3-byp4	+	+	35	+	GGA	UCC	GGU	GGAtccaccttaattaaggtggaataaataataaggtAAT
	rps3-byp5	—	+	39	+	GGC	UCC	GGU	GGCtccaccatttatggtggattagtttataaagtaatggtAAT
	rps3-byp6^	—	—	42	+	GGU	CGA	GGA	GGTcgattcccaccttataaaaggtgggaaaagttatttttatggaTAC
	rps3-byp7	+	—	47	+	GGA	CGA	GGU	GGAcgattatccctttttatatagggataatttatcaatattcttataaggtTCA
Introns									
cob-I1	*cob*-I1-byp1	+	+	40	+	GGG	UCC	GGU	GGGtccccccccccctcatgaggggggggatagttcactaggtTTT
cob-I2	*cob*-I2-byp1^	—	—	30	+	CUA	CCC	CUU	CTAcccccaatcaattattgggggtgggga cttTGT
	cob-I2-byp2	+	—	34	+	GGA	UCC	GGC	GGAtccccccccaccagtggggtaatcttctcaggcTCA
	cob-I2-byp3	+	+	31	+	GCC	UCC	GCU	GCCtccaaccttagattgggga ggataaaggctATT
	cob-I2-byp4	+	—	47	+	GAC	UCC	GAU	GACtccccccaaccttaggttggggagtaggaggtaatattaagtggatAAT
	cob-I2-byp5	+	+	38	+	GGC	UCC	GGA	GGCtcccacaaccaatggttgtgggggatttatttaggaAAT
cox1-I1	*cox1*-I1-byp1	+	—	34	+	GGA	UCC	GGU	GGAtccccccttaggggggggtatctccttaggtACA
	cox1-I1-byp2	+	—	35	+	AUC	UCC	AUU	ATCtccccaccataaacatggtggggatcctaaaattAAT
	cox1-I1-byp3	—	+	54	+	GGC	ACC	GGU	GGCaccaccatctatgatggtggttattaatttaattaccttctgctaaatctaggtAAT
cox1-I2	*cox1*-I2-byp1	—	+	33	+	GAC	UCC	GAU	GACtccaatcatagattgggggttggtataaagatTTA
	cox1-I2-byp2^	—	—	27	+	GCC	UCC	GCU	GCCtccccggccctgggggggaccggagcttTTT
	cox1-I2-byp3	—	+	39	+	GGU	UCC	GGU	GGTtccccccaccatttaatggtggggataattaataaggtCAT
	cox1-I2-byp4	—	+	42	+	GGG	UCC	GGA	GGGtccccccctctttaagagggggggggtggttataattaaggaATT
	cox1-I2-byp5	+	+	35	+	GAU	UCC	GAU	GATtccccccccttataggggggtaatcctaaatttgatACT
cox1-I3	*cox1*-I3-byp1	—	+	39	+	GGA	UCC	GGU	GGAtccccccccataaaaaagggggggggggatattaaggtTTA
	cox1-I3-byp2	+	—	29	+	ACC	ACC	ACU	ACCaccatcatttatgatgggtgggagtactTAT
	cox1-I3-byp3*	+	+	40	+	GCU	UCC	GUU	GCTtccccccaaccaaagattgggaggggtagaatataagttAAA
	cox1-I3-byp4	—	+	36	+	ACU	UCC	ACA	ACTtccccaccactttaagtggtgggggagtagttcacaATG
cox1-I4	*cox1*-I4-byp1	+	—	41	+	AAA	UCC	AAA	AAAtccccccctataaaaaagggggggggtgggtcgtagtttaaaTTT
	cox1-I4-byp2	+	—	29	+	GCU	UCC	GCU	GCTtcccatattcaaaaatatgggtataagctTAT
	cox1-I4-byp3^	—	—	33	+	AGA	ACC	AGA	AGAaccccccccaaccaggggggggtctgtctagaGTT
	cox1-I4-byp4	+	—	42	+	GGA	UCC	GGA	GGAtccccccactttcaagagatggtgggataatttacaaaaggaATG
	cox1-I4-byp5	—	+	37	+	GUU	ACC	GTA	GTTaccccccaacttatggttgggggatttaacaagtaTTT
cox1-I6	*cox1*-I6-byp1	+	+	34	+	AGG	ACC	AGA	AGGaccccaccctttataaagggtggatcaccatagaTTA

Asterisks label byps whose takeoff and landing codons belong to different families. (') indicates byps that are translatable with National Center for Biotechnology Information translation table 4 but otherwise contain codons unused in *M. capitatus* mtDNA. Stop codons are in frame with upstream coding region. Yellow- and green-shaded codons highlight the two predominant avoided codons. Turquoise shading highlights the cases where takeoff and landing codons belong to different codon families. In the column "Genomic sequence," takeoff and landing codons are in red font color, and the avoided codon is shaded. The codon where translation resumes is shown in uppercase. Complementary nucleotides of the hairpin structures are underlined.

Table S3. Mitochondrial codon frequencies in *Magnusiomyces capitatus*

Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count
F	TTT	190	S	TCT	18	Y	TAT	238	C	TGT	29
F	TTC	108	S	TCC	0	Y	TAC	12	C	TGC	0
L	TTA	602	S	TCA	183	*	TAA	11	W	TGA	64
L	TTG	3	S	TCG	0	*	TAG	1	W	TGG	2
L	CTT	12	P	CCT	65	H	CAT	66	R	CGT	23
L	CTC	0	P	CCC	0	H	CAC	5	R	CGC	0
L	CTA	8	P	CCA	70	Q	CAA	72	R	CGA	0
L	CTG	0	P	CCG	0	Q	CAG	0	R	CGG	0
I	ATT	244	T	ACT	113	N	AAT	290	S	AGT	130
I	ATC	37	T	ACC	0	N	AAC	20	S	AGC	5
I	ATA	341	T	ACA	128	K	AAA	137	R	AGA	49
M	ATG	102	T	ACG	0	K	AAG	2	R	AGG	0
V	GTT	73	A	GCT	113	D	GAT	85	G	GGT	150
V	GTC	2	A	GCC	2	D	GAC	2	G	GGC	3
V	GTA	153	A	GCA	87	E	GAA	102	G	GGA	95
V	GTG	5	A	GCG	1	E	GAG	7	G	GGG	5

Numbers are calculated for codon regions across *atp6, -8,* and *-9, cob, cox1, -2,* and *-3, nad1, -2, -3, -4, -4L, -5,* and *-6,* and *rps3*, after exclusion of byps. Red, takeoff codons present in byps; green, other unused codons.

Table S4. Mitochondrial codon frequencies in *Saccharomyces cerevisiae*

Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count
F	TTT	81	S	TCT	39	Y	TAT	106	C	TGT	13
F	TTC	67	S	TCC	0	Y	TAC	11	C	TGC	1
L	TTA	285	S	TCA	78	*	TAA	8	W	TGA	38
L	TTG	2	S	TCG	0	*	TAG	0	W	TGG	1
T	CTT	6	P	CCT	46	H	CAT	53	R	CGT	1
T	CTC	0	P	CCC	3	H	CAC	5	R	CGC	0
T	CTA	19	P	CCA	36	Q	CAA	36	R	CGA	0
T	CTG	0	P	CCG	1	Q	CAG	5	R	CGG	1
I	ATT	193	T	ACT	40	N	AAT	196	S	AGT	25
I	ATC	29	T	ACC	1	N	AAC	12	S	AGC	0
M	ATA	38	T	ACA	53	K	AAA	70	R	AGA	42
M	ATG	72	T	ACG	0	K	AAG	3	R	AGG	0
V	GTT	50	A	GCT	66	D	GAT	53	G	GGT	99
V	GTC	7	A	GCC	7	D	GAC	2	G	GGC	1
V	GTA	83	A	GCA	54	E	GAA	41	G	GGA	31
V	GTG	5	A	GCG	4	E	GAG	3	G	GGG	6

Based on the reduced mitochondrial gene set in this yeast, including *atp6, -8,* and *-9, cob, cox1, -2, and -3*, and *rps3*. Green, unused codons.

Table S5. Mitochondrial codon frequencies in *Kluyveromyces lactis*

Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count
F	TTT	70	S	TCT	35	Y	TAT	100	C	TGT	13
F	TTC	88	S	TCC	1	Y	TAC	12	C	TGC	1
L	TTA	283	S	TCA	96	*	TAA	8	W	TGA	40
L	TTG	2	S	TCG	1	*	TAG	0	W	TGG	0
T	CTT	0	P	CCT	49	H	CAT	54	R	CGT	0
T	CTC	0	P	CCC	4	H	CAC	2	R	CGC	0
T	CTA	0	P	CCA	35	Q	CAA	48	R	CGA	0
T	CTG	0	P	CCG	1	Q	CAG	0	R	CGG	0
I	ATT	212	T	ACT	41	N	AAT	154	S	AGT	27
I	ATC	16	T	ACC	0	N	AAC	13	S	AGC	2
M	ATA	3	T	ACA	65	K	AAA	75	R	AGA	46
M	ATG	63	T	ACG	0	K	AAG	1	R	AGG	1
V	GTT	50	A	GCT	70	D	GAT	49	G	GGT	86
V	GTC	2	A	GCC	4	D	GAC	2	G	GGC	1
V	GTA	101	A	GCA	50	E	GAA	47	G	GGA	53
V	GTG	1	A	GCG	3	E	GAG	2	G	GGG	2

Based on the reduced mitochondrial gene set in this yeast, including *atp6, -8,* and *-9, cob, cox1, -2,* and *-3*, and *rps3*. Green, unused codons.

Table S6. Mitochondrial codon frequencies in *Lachancea thermotolerans*

Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count
F	TTT	85	S	TCT	41	Y	TAT	93	C	TGT	16
F	TTC	72	S	TCC	0	Y	TAC	27	C	TGC	0
L	TTA	303	S	TCA	82	*	TAA	6	W	TGA	41
L	TTG	1	S	TCG	0	*	TAG	2	W	TGG	0
T	CTT	5	P	CCT	56	H	CAT	49	R	CGT	2
T	CTC	0	P	CCC	1	H	CAC	5	R	CGC	0
T	CTA	11	P	CCA	27	Q	CAA	47	R	CGA	0
T	CTG	0	P	CCG	0	Q	CAG	0	R	CGG	0
I	ATT	204	T	ACT	39	N	AAT	142	S	AGT	30
I	ATC	17	T	ACC	0	N	AAC	16	S	AGC	3
M	ATA	2	T	ACA	63	K	AAA	76	R	AGA	44
M	ATG	56	T	ACG	0	K	AAG	1	R	AGG	0
V	GTT	51	A	GCT	88	D	GAT	51	G	GGT	82
V	GTC	1	A	GCC	5	D	GAC	2	G	GGC	0
V	GTA	103	A	GCA	50	E	GAA	49	G	GGA	53
V	GTG	1	A	GCG	0	E	GAG	1	G	GGG	2

Based on the reduced mitochondrial gene set in this yeast, including *atp6, -8,* and *-9, cob, cox1, -2,* and *-3*, and *rps3*. Green, unused codons.

Table S7. Mitochondrial codon frequencies in *Magnusiomyces magnusii* strain 270

Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count
F	TTT	36	S	TCT	102	Y	TAT	48	C	TGT	32
F	TTC	305	S	TCC	20	Y	TAC	198	C	TGC	2
L	TTA	52	S	TCA	86	*	TAA	8	W	TGA	62
L	TTG	3	S	TCG	9	*	TAG	7	W	TGG	2
L	CTT	217	P	CCT	76	H	CAT	29	R	CGT	35
L	CTC	55	P	CCC	2	H	CAC	51	R	CGC	0
L	CTA	220	P	CCA	62	Q	CAA	76	R	CGA	0
L	CTG	152	P	CCG	7	Q	CAG	30	R	CGG	4
I	ATT	95	T	ACT	80	N	AAT	32	S	AGT	87
I	ATC	420	T	ACC	2	N	AAC	214	S	AGC	84
I	ATA	5	T	ACA	169	K	AAA	75	R	AGA	51
M	ATG	108	T	ACG	1	K	AAG	49	R	AGG	0
V	GTT	91	A	GCT	71	D	GAT	38	G	GGT	71
V	GTC	33	A	GCC	81	D	GAC	57	G	GGC	6
V	GTA	113	A	GCA	94	E	GAA	20	G	GGA	148
V	GTG	122	A	GCG	32	E	GAG	95	G	GGG	91

Numbers are calculated for codon regions across *atp6, -8,* and *-9, cob, cox1, -2,* and *-3, nad1, -2, -3, -4, -4L, -5,* and *-6*, and *rps3*. Green, unused codons.

Table S8. Mitochondrial codon frequencies in *Yarrowia lipolytica*

Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count	Amino acid	Codon	Count
F	TTT	160	S	TCT	78	Y	TAT	190	C	TGT	30
F	TTC	160	S	TCC	1	Y	TAC	40	C	TGC	0
L	TTA	610	S	TCA	149	*	TAA	12	W	TGA	55
L	TTG	7	S	TCG	3	*	TAG	2	W	TGG	2
L	CTT	19	P	CCT	65	H	CAT	58	R	CGT	0
L	CTC	0	P	CCC	1	H	CAC	23	R	CGC	0
L	CTA	24	P	CCA	67	Q	CAA	67	R	CGA	0
L	CTG	1	P	CCG	1	Q	CAG	0	R	CGG	0
I	ATT	173	T	ACT	105	N	AAT	186	S	AGT	98
I	ATC	87	T	ACC	0	N	AAC	46	S	AGC	4
I	ATA	277	T	ACA	128	K	AAA	80	R	AGA	75
M	ATG	118	T	ACG	0	K	AAG	5	R	AGG	0
V	GTT	89	A	GCT	162	D	GAT	83	G	GGT	168
V	GTC	4	A	GCC	10	D	GAC	9	G	GGC	0
V	GTA	159	A	GCA	84	E	GAA	61	G	GGA	83
V	GTG	10	A	GCG	3	E	GAG	28	G	GGG	4

Numbers are calculated for codon regions across *atp6, -8,* and *-9, cob, cox1, -2,* and *-3, nad1, -2, -3, -4, -4L, -5,* and *-6*, and *rps3*. Green, unused codons.

Table S9. Expression of *nad4L*-byp1 in *E. coli*

LacZ construct	Colony numbers		% of control
	White	Blue	
pUC19 (control)	0	231	100
Original byp with prepeptide	3	0	1
Modified byp with CGA (avoided codon) replaced by TGA (stop codon)	10	3	5
Modified byp with CGA replaced by TGA; no prepeptide	10	2	5
Modified byp with TTCG tetraloop; SD motif; no prepeptide	5	1	3

Cloning experiment was repeated three times; numbers of colonies are rounded to the nearest integer. Essentially the same results were obtained for *cox1*-byp1. For details, see *SI Materials and Methods*.