Supplementary material

Complete characterization of the edited transcriptome of the mitochondrion of *Physarum polycephalum* using deep sequencing of RNA

R. Bundschuh, J. Altmüller, C. Becker, P. Nürnberg, and J. M. Gott

Oligonucleotides for experimental verification of new editing sites

12nd2: GAAATAGTCAAAAATAATAAATCAG
cirRTpro1: AACCGAATGCTCTACCAG
Primer sets for PCR:
1nd5: AGCATGTGAGAAAATAACAG / 3nd5: TGTTTTTAGGATGGGAAGCT
7LSU: CAGTAGGTAAACGAGACTG / 24LSU: GTGCCAAACAATTCCGTC
34LSU: TTACCAGTGATTTAAGAGAC / 35LSU: TTACATATAAAGCGGACTAGT
1rpL16: AACGATGTACTTTACGTAGT / 2rpL16: TTAGAGACTGCACGTAGAG
2nd2: GCAACAATATTACCATTCCC / 7nd2: AACTTTATGTTTGCTTTTATAC
1nd3: GTCAATTTGATAAAAAGTAGTTG / 3nd3: AGAAACTAACAATATGGCGAG
1rpS12: AAGGATCCAACCTAATTCTGCAAACGCA / 2rpS12: AAAGTCGACCACACCATATTTACTACAC
2rpS2: AAAGTCGACTAGTATTAGATACTTCAGC / 1rpS2/ndG: CTTAGTTCTTCTTGCAAATAC
1php22: TGCTTAATAAAATAAAAATAAGT / 2php22: AACAAAAATGATAAAGCCGT
1ssu: TCACGTACAGACCGCCC / tRNAK3: TGGTTGGCTCCACAGGACTTGC
cirRTlys1: AAAGCCGATAGCATTACTAT / cirlys2: CGACAGAGTGCAGGTGC

gene	genomic location	status	accession numbers
nad5	17259-19152	annotated	HQ849407
nadG	19300-20316	new	HQ849408
rpS2	20278-21633	predicted	HQ849424
rpS12	21746-22241	annotated	HQ849419
rpS7	22246-23009	predicted	HQ849427
rpL2	23009-23777	predicted	HQ849416
rpS19	23774-24139	predicted	HQ849423
php15	24416-25471	annotated	NC_002508
cox1	27534-25816	known	L14769
nad7	27670-27535	known	AB039844
cox2	29666-28983	known	DQ092489
php22 (rpL11?)	29776-30652	new	HQ849409
nad2	30699-32105	known	DQ092490
rpS16	34704-34988	new	HQ849422
rpL19	34988-35540	predicted	HQ849415
atp8	35567-35788	known	DQ092488
nad4L	35788-36062	known	DQ092491
atp6	36067-36774	known	HQ849400
nad4	38315-36933	annotated	HQ849406
nad3	38808-38435	annotated	HQ849405
rpL14	38985-39338	predicted	HQ849413
php23	39338-39822	new	HQ849410
rpS14	39823-40087	predicted	HQ849421
rpS8	40088-40509	predicted	HQ849428
rpL6	40506-40972	new	HQ849417
rpS13	40978-41517	predicted	HQ849420
nad9	41520-41994	known	S67221
rpS11	41997-42717	new	HQ849418
php24	42721-43372	new	HQ849411
rpS4	44229-43440	new	HQ849426
tRNA-Glu	48231-48163	known	AF059032
tRNA-Met1	48305-48237	known	AF059032
23S rRNA	48432-51149	known	HQ849399
17S rRNA	51353-53166	known	X75592
5S rRNA	53178-53273	known	HQ916349
tRNA-Met2	53273-53344	known	AF059033
tRNA-Lys	53364-53435	known	HQ849429
tRNA-Pro	53454-53524	known	AF059033
php25 (atpB?)	53565-53858	new	HQ849412
atpA	53845-55380	known	HQ849402
cox3	56278-55517	known	AF084527
nad6	56758-56280	known	DQ092492
rpL16	57434-56910	predicted	HQ849414
rpS3	58800-57431	predicted	HQ849425
nad1	58893-59821	annotated	HQ849404
cytb	61038-59903	known	HQ849403
atp9	61225-61467	known	HQ849401

Supplementary Table 1: Genomic locations of the genes identified with our high throughput sequencing approach as well as genes previously known. Genes, for which the genomic position is given in reverse order are located on the reverse strand. The status column indicates if the editing sites in the gene were known before our study, if the location of the gene had been annotated in the mitochondrial genome without the editing sites being known, if the gene was predicted in [Beargie, C., Liu, T., Corriveau, M., Lee, H.Y., Gott, J. and Bundschuh, R. (2008) Genome annotation in the presence of insertional RNA editing. *Bioinformatics*, **24**, 2571-2578] or if the gene is completely new.

genomic region	ORFs	gene before	gene after	coverage
1-6399 61652-62862	1-5 20	atp9	ORF6	occasional short stretches of overlapping reads not covering a complete ORF
6400-17150	6-13	ORF5	nad5	scattered hits
32250-34630	15-16	nad2	rpS16	one single read
44293-47874	17-19	rpS4	tRNA-Glu	one single read

Supplementary Table 2: Coverage of the previously annotated ORFs with the exception of ORF14. The first region covering ORFs 1-5 and 20 is the only region where we find a potential sign of transcription, albeit at a lower level than for the transcripts reported in Figure 1 of the main manuscript. However, this may also be a sign of genomic contamination. ORF14 is well covered by our sequencing reads and is reported as php15 in Figure 1 of the main manuscript and in Supplementary Table 1.

genomic position	gene	our transcript	published transcript	transcript accession	genomic accession
39589	atp6	G	U	FJ154098	-
54679	atpA	C	U	M31718	M31717
54843	atpA	C	U	M31718	M31717
54884	atpA	U	C	M31718	M31717
54903	atpA	U	C	M31718	M31717
60173	cytb	U	C	AF079799	AF079798
61425	atp9	G	C	S67221	S67222
61426	atp9	C	G	S67221	S67222

Supplementary Table 3: Genomic positions in which our transcripts differ from the cDNA sequences deposited in GenBank. In all instances, the nucleotide in our transcript agress with the sequence of the published genome [Takano, H., Abe, T., Sakurai, R., Moriyama, Y., Miyazawa, Y., Nozaki, H., Kawano, S., Sasaki, N., and Kuroiwa, T. (2001) The complete DNA sequence of the mitochondrial genome of *Physarum polycephalum*. *Mol Gen Genet*, **264**, 539-545]. However, the sequences of the published atpA, cytb, and atp9 cDNAs match the genomic GenBank entry from the same publication [Miller, D., Mahendran, R., Spottswood, M., Constandy, H., Wang, S., Ling, M.L., and Yang, N. (1993) Insertional editing in mitochondria of *Physarum*. *Semin Cell Biol*, **4**, 261-266], which predate the publication of the full mitochondrial genome. There is similar agreement between our previously published atp6 cDNA [Gott, J.M., Parimi, N., and Bundschuh, R. (2005) Discovery of new genes and deletion editing in *Physarum* mitochondria enabled by a novel algorithm for finding edited mRNAs. *Nucleic Acids Res*, **33**, 5063-5072] and its genomic DNA (both derived from a different isolate of the M3 strain than the one used in this work, genomic data not published). Thus, these differences are most likely due to genomic variations between strains.

tRNA	codon	# reads	accession number
Ala1	GCU	17334	HQ849430
Ala2	GCU	5587	HQ849431
Ala3	GCU	1172	HQ849432
Ala4	GCA	2917	HQ849433
Arg	CGA	871	HQ849434
Asp1	GAC	2331	HQ849435
Asp2	GAC	1945	HQ849436
Gln1	CAA	7637	HQ849437
Gln2	CAA	1468	HQ849438
Glu2	CAG	1642	HQ849439
Gly	GGA	1617	HQ849440
His	CAC	2325	HQ849441
Leu1	CUU/CUC	931	HQ849442
Leu2	UUA	789	HQ849443
Leu3	CUA	759	HQ849444
Leu4	CUA	709	HQ849445
Leu5	CUG	435	HQ849446
Ser1	UCA	4980	HQ849447
Ser2	UCC	162	HQ849448
Thr	ACA	709	HQ849449
Val1	GUA	5456	HQ849450
Val2	GUU/GUC	591	HQ849451

Supplementary Table 4: Nuclear encoded tRNAs of *Physarum polycephalum* and the number of reads from our mitochondrial RNA preparation that support them.

insertion	**gene**	**genomic position**
G	nad5	17959
G	23S RNA	50099
A	rpL16	57109

Supplementary Table 5: Locations of instances of new types of RNA editing. In order to specify the genomic position of an insertion we by convention quote the genomic position of the base 5' of the last possible site of nucleotide insertion.

genomic position	**location**	**comments**
29766	5' UTR of php22	
30658 30685	between php22 and nad2	both genes on same transcript
38833 38847	5' UTR of nad3	5'UTR of nad3 is 92nt long
53352	between tRNA-Met2 and tRNA-Lys	partial editing site
61493 61524 61584 61607	3' UTR of atp9	possible ORF but no stop codon before end of 3' UTR could be a structural RNA

Supplementary Table 6: Extragenic editing sites. All ten extragenic editing sites are C insertions. The two C insertions within the 44 nucleotides between php22 (putative rpL11) and nad2 have been confirmed by primer extension sequencing of both total mtRNA and cloned mtDNA with an end-labeled primer that anneals to genomic region 30749-30773 (within nad2). Both sequences extend well into php22, confirming that these two genes are on the same transcript (see Supplementary Figure 5). The C insertions within the 5' UTRs of php22 and nad3 have also been confirmed by Sanger sequencing (data not shown). Based on our read coverage, the nad3 transcript starts around 38900 (on the reverse strand), with the start codon of nad3 at 38808. There are no ORFs in this highly AT-rich region (82% A+T), making it unlikely that this 92 nt region encodes another protein. The region of the four C insertions within the long (~240 nt) 3' UTR of the atp9 mRNA could potentially encode a separate protein starting at 61473 (6 nts downstream of the atp9 stop codon), but no in-frame stop codons are encountered before our read coverage ends at 61708, making this the only potential ORF without an identifiable stop codon. Translation from any of the upstream AUGs would end at or before the atp9 stop codon at 61468, ruling out the possibility of a gene overlapping atp9. It is also possible that this region encodes a structural RNA, but BLAST searches yielded no hits (other than *Physarum*) to either nucleotide or protein databases.

genomic position	gene	editing type	edited/total reads	editing rate
20491	rpS2	U	54/70	77%
21906	rpS12	C	14/20	70%
29879	php22	C	101/132	77%
30848	nad2	U	70/88	80%
53352	tRNA-Met2/tRNA-Lys	C	60/103	58%

Supplementary Table 7: Potential sites of partial insertional editing based on high throughput sequencing reads. The first four of these come relatively close to our threshold of 80% for considering an editing site fully edited (the presence of sequencing errors and misalignments requirres a somewhat generous threshold). None of these four was found to be partially edited via Sanger sequencing. Only the fifth site could be confirmed to be partially edited by Sanger sequencing.

codon position	**previously known mRNAs**	**newly discovered mRNAs**	**all mRNAs**
first	67 (25%)	194 (37%)	261 (33%)
second	30 (11%)	117 (22%)	147 (18%)
third	172 (64%)	217 (41%)	389 (49%)

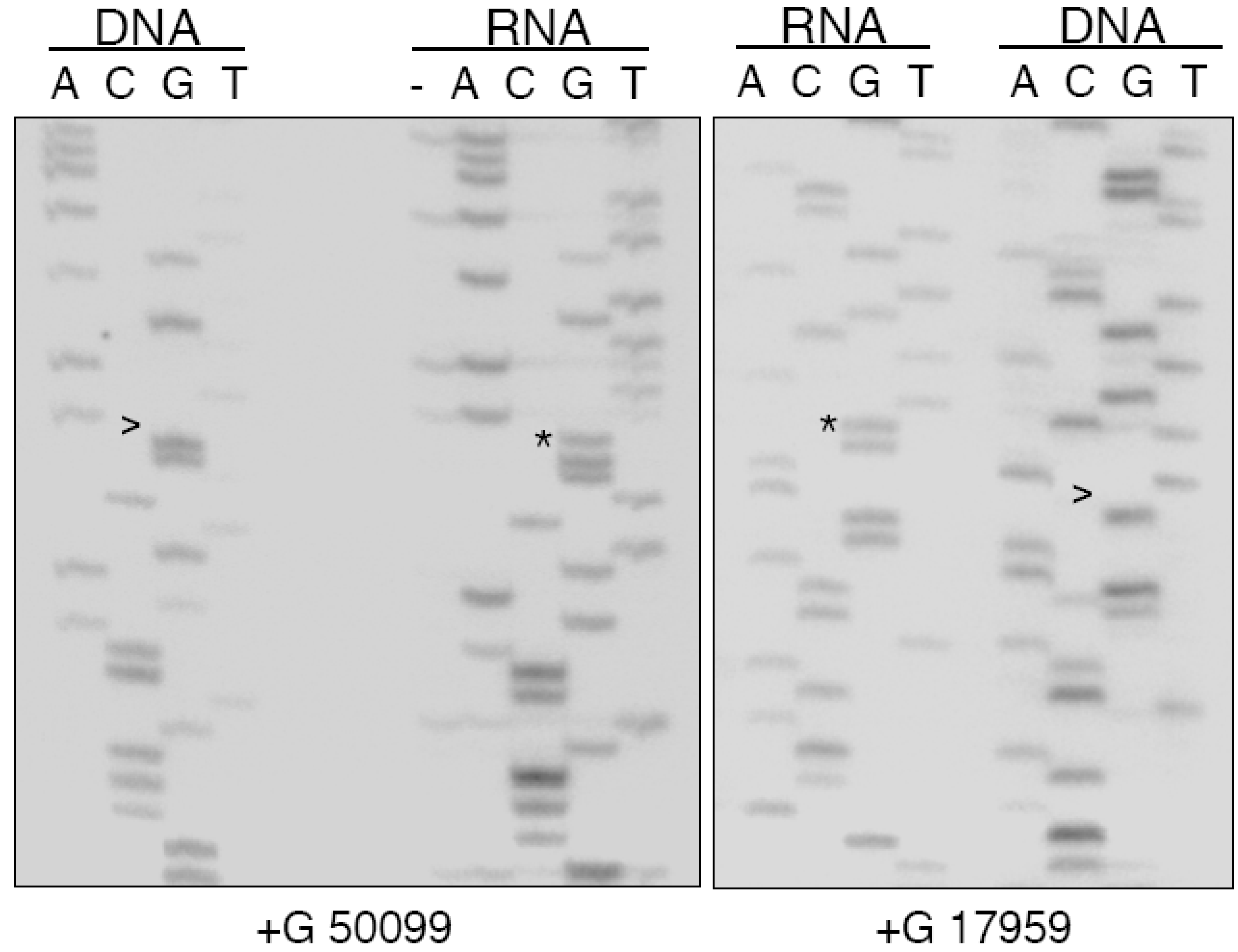
Supplementary Table 8: Codon positions of the unambiguous C insertion sites in the previously known, newly discovered, and total set of mRNAs.

1st base 2nd base	U			C			A			G		
U	U	687 -	6.7% -	U	346 38	3.3% 3.3%	U	510 -	4.9% -	U	71 -	0.7% -
	C	173 37	1.7% 3.2%	C	81 34	0.8% 3.0%	C	52 13	0.5% 1.1%	C	22 5	0.2% 0.4%
	A	758 -	7.3% -	A	156 39	1.5% 3.4%	A	34 -	0.3% -	A	4 -	0.0% -
	G	80 -	0.8% -	G	14 4	0.15% 0.3%	G	1 -	0.0% -	G	84 -	0.8% -
C	U	283 73	2.7% 6.3%	U	214 53	2.1% 4.6%	U	212 34	2.0% 3.0%	U	228 45	2.2% 3.9%
	C	46 28	0.5% 2.4%	C	21 9	0.2% 0.8%	C	17 6	0.2% 0.5%	C	42 10	0.4% 0.9%
	A	148 77	1.5% 6.7%	A	120 40	1.2% 3.5%	A	242 43	2.3% 3.7%	A	98 27	1.0% 2.3%
	G	20 12	0.2% 1.0%	G	14 4	0.2% 0.3%	G	26 2	0.2% 0.2%	G	10 3	0.1% 0.3%
A	U	487 -	4.7% -	U	258 17	2.5% 1.5%	U	448 -	4.3% -	U	137 -	1.3% -
	C	268 188	2.6% 16.3%	C	115 79	1.1% 6.9%	C	69 17	0.7% 1.5%	C	52 21	0.5% 1.8%
	A	293 -	2.8% -	A	171 5	1.7% 0.4%	A	646 -	6.2% -	A	151 -	1.5% -
	G	197 -	1.9% -	G	14 1	0.2% 0.1%	G	92 -	0.9% -	G	14 -	0.2% -
G	U	238 -	2.3% -	U	316 1	3.0% 0.1%	U	275 -	2.7% -	U	272 -	2.6% -
	C	88 63	0.8% 5.5%	C	89 52	0.9% 4.5%	C	40 16	0.4% 1.4%	C	36 10	0.4% 0.9%
	A	132 -	1.3% -	A	133 12	1.3% 1%	A	280 -	2.7% -	A	114 -	1.1% -
	G	30 -	0.3% -	G	17 1	0.2% 0.1%	G	40 -	0.4% -	G	16 -	0.2% -

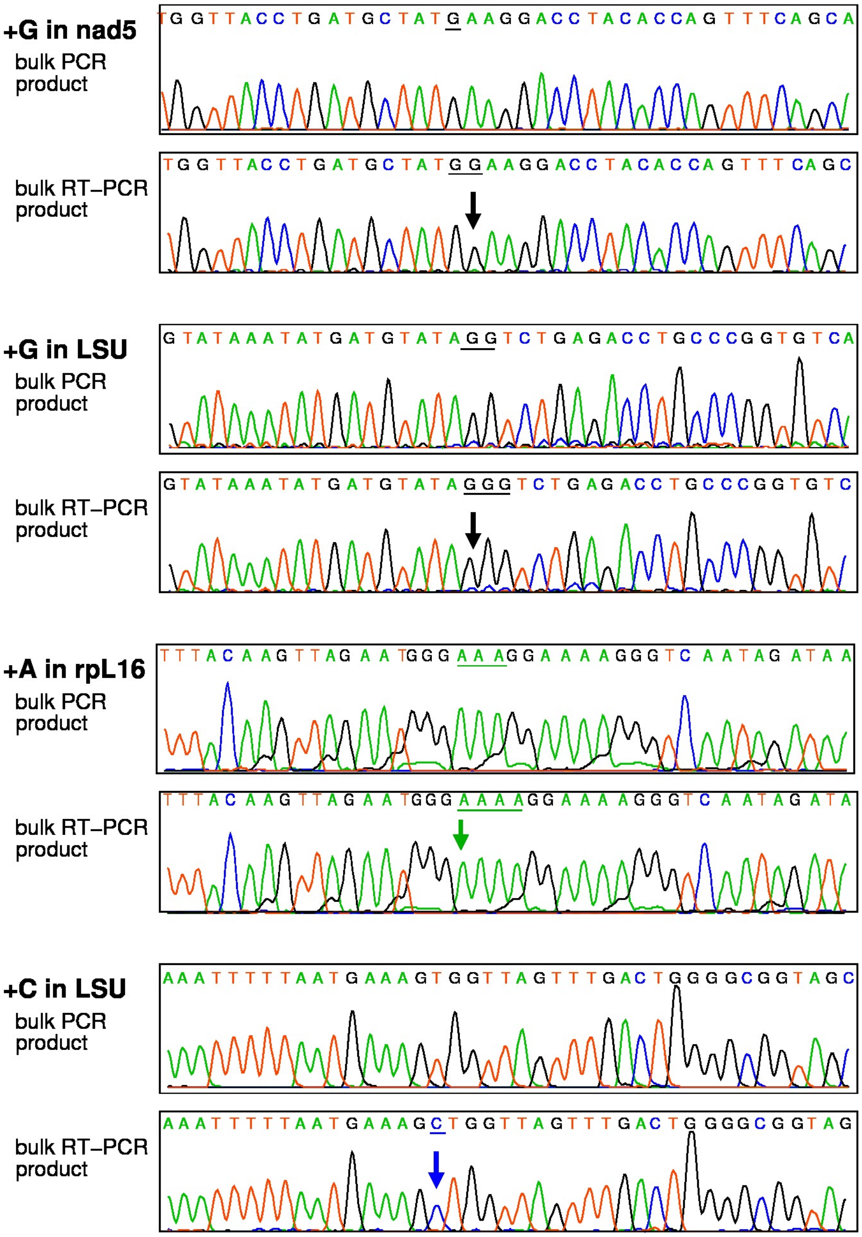
Supplementary Table 9: Codon usage over all 39 mitochondrial protein coding genes of *Physarum polycephalum* found to be transcribed in this study. The first number and percentage in each entry correspond to the total number of codons. The second number and percentage in each entry represents only codons containing a C insertion. In cases where a C is inserted next to an encoded C, the first C was designated as the edited nucleotide.

type	number	contexts
AA	4	AAA(4x)
UU	2	UUUU,UUU
UG/GU	4	GUGU(3x),UGUGU
UC/CU	9	CUC(4x),CUCU(2x),UCUC,UCUCU(2x)
UA	2	UA,UAUA
GC/CG	2	GCGC(2x)

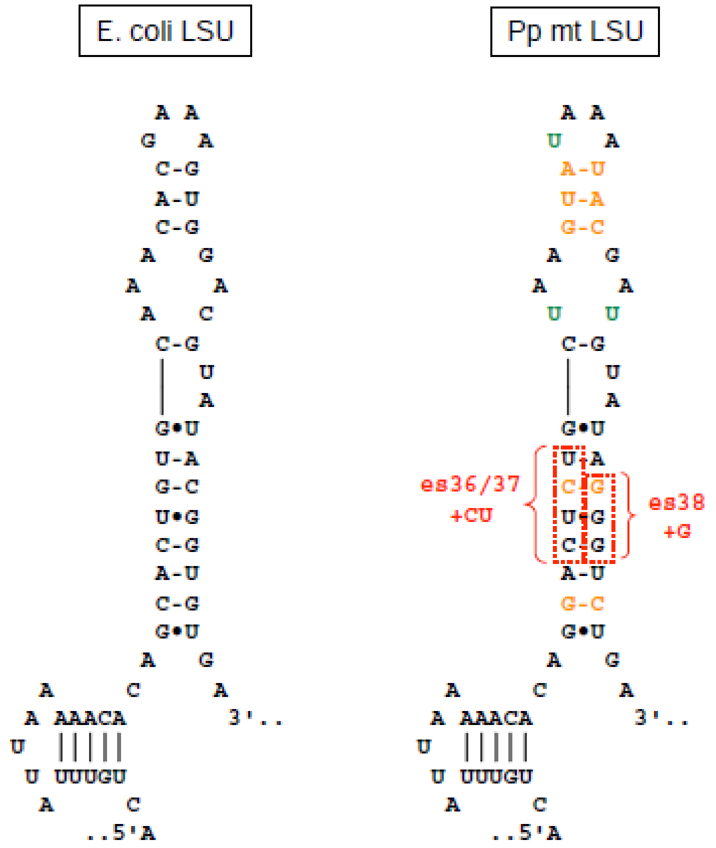
Supplementary Table 10: Dinucleotide insertions observed in the mitochondrion of *Physarum polycephalum*, their frequencies, and sequence contexts.The third column provides the stretches of the mRNA sequence within which the position of the actual dinucleotide insertion is ambiguous.



Supplementary figure 1: Confirmation of G insertions via primer extension sequencing. End-labeled primers specific for 23S rRNA (left) and nad5 (right) were annealed to bulk mitochondrial RNA and fragmented mitochondrial DNA and extended by reverse transcriptase in the presence of ddNTPs. Arrowheads indicate the position of G insertion sites within the genomic DNA; added Gs in the RNA are marked with asterisks.

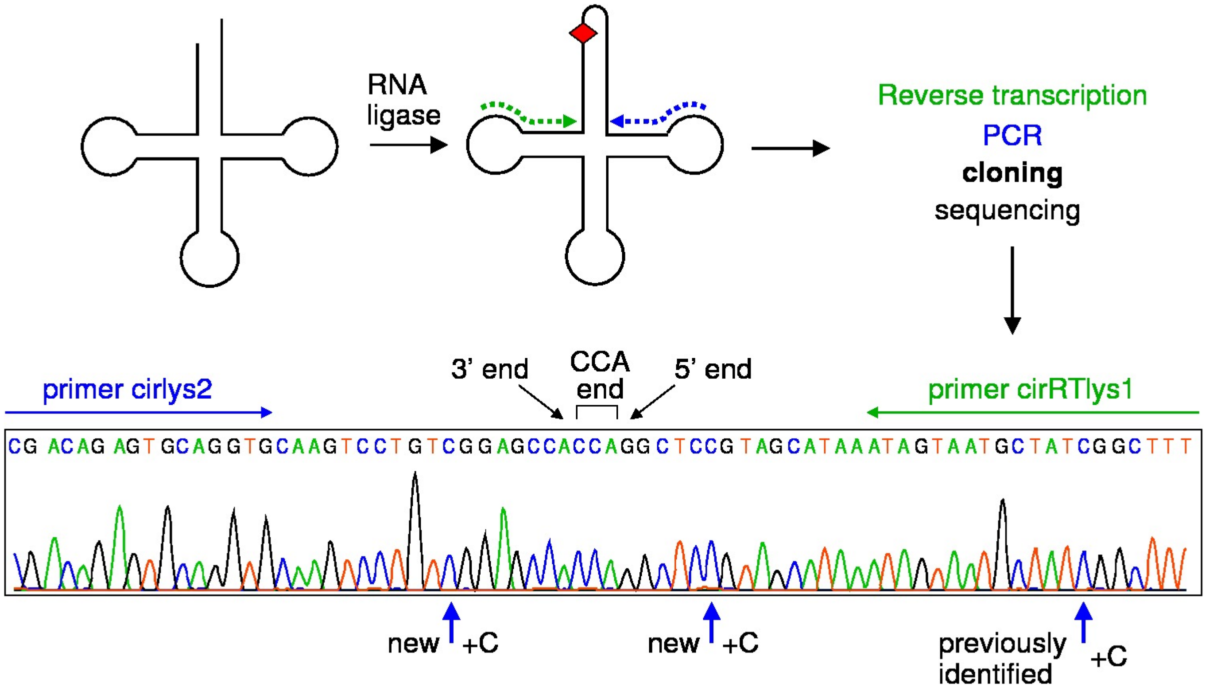


Supplementary figure 2: Sanger sequencing traces confirming the three instances of novel editing types (G and A insertions) as well as the previously not reported C insertion in the large subunit ribosomal RNA. The arrows indicate the added nucleotides in the RT-PCR products.

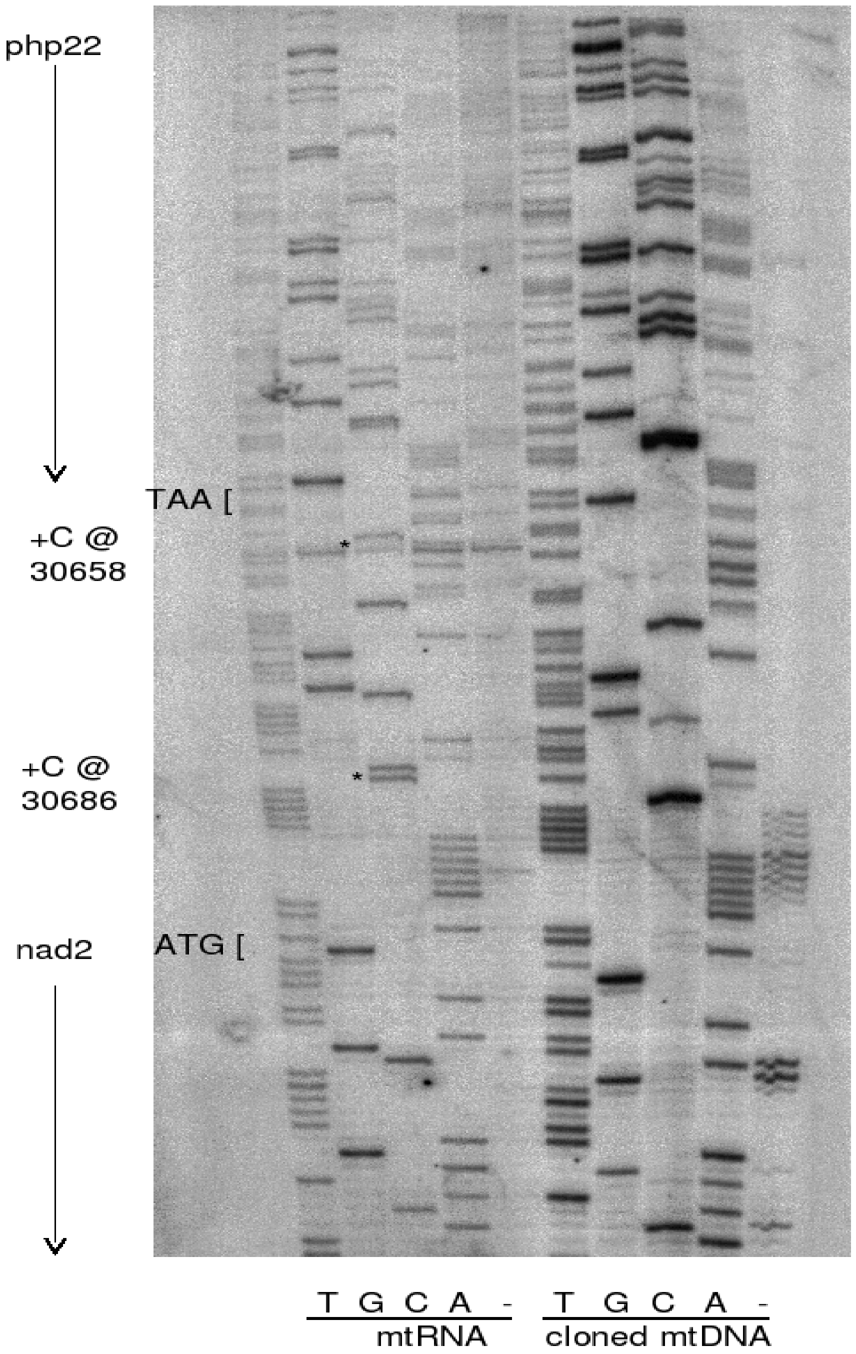


Supplementary figure 3: Predicted secondary structures for nt 1773-1829 of the *E. coli* large subunit rRNA (taken from Gutell, R.R., Gray, M.W. and Schnare, M.N. (1993) A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. Nucleic Acids Research, 21, 3055-3074) and the equivalent region (nt 1655-1710) of the large subunit rRNA from *Physarum polycephalum* mitochondria. Compensatory changes are shown in orange; other differences between the two sequences are shown in green.

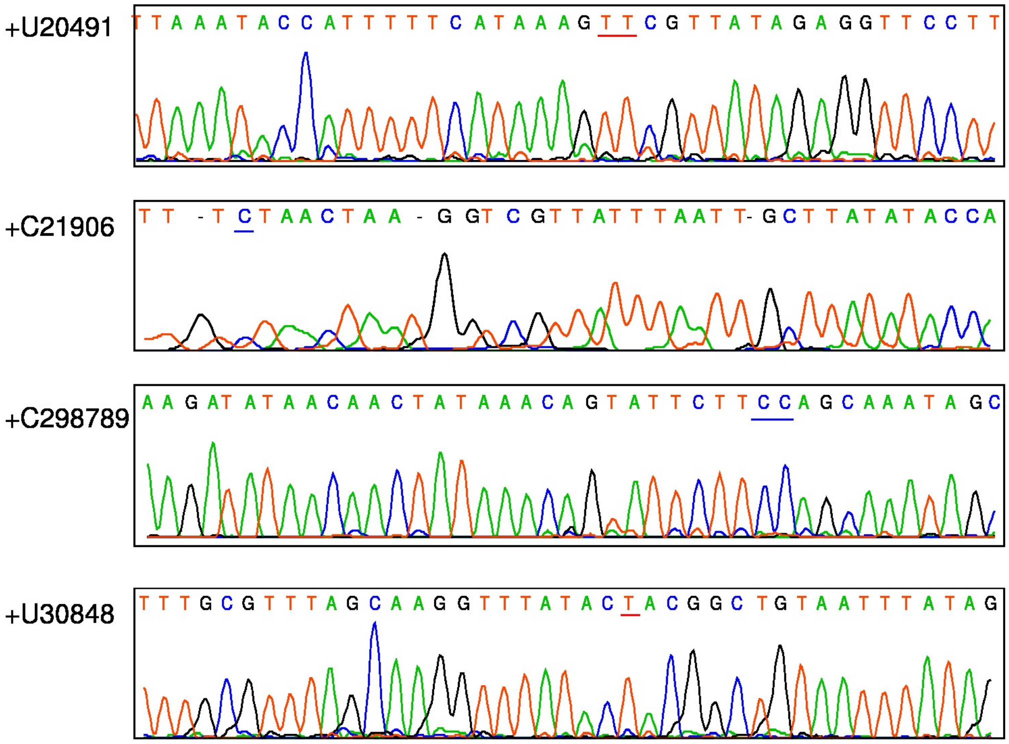
Editing sites are indicated in red. Note that the exact sites of nucleotide insertion are ambiguous, since the added CU (or UC) is inserted next to an encoded CU and the single G insertion is adjacent to encoded Gs. The extent of this ambiguity is indicated by dotted boxes.



Supplementary figure 4: RT-PCR products derived from circularized tRNA-Lys. The ends of tRNA-Lys were ligated and the acceptor stem was reverse transcribed and PCR amplified using the primers indicated in green and blue. The PCR product was cloned and subjected to Sanger sequencing. The resulting sequence trace shows three C insertions, two of which had not been previously identified, as well as the actual 3' and 5' end of tRNA-Lys.



Supplementary figure 5: Confirmation of intergenic C insertions via primer extension sequencing. End-labeled primers specific for the 5' end of nad2 were annealed to bulk mitochondrial RNA and a plasmid containing the php22-nad2 region of mtDNA and extended by reverse transcriptase in the presence of ddNTPs. Added Cs in the RNA are marked with asterisks.



Supplementary figure 6: Sanger sequencing traces of RT-PCR products at editing sites suspected to display partial editing based on high throughput sequencing reads. The absence of double sequence (refer to figure 4(a) in the main text for an example of true partial editing) indicates that these are not partially edited at any significant level.