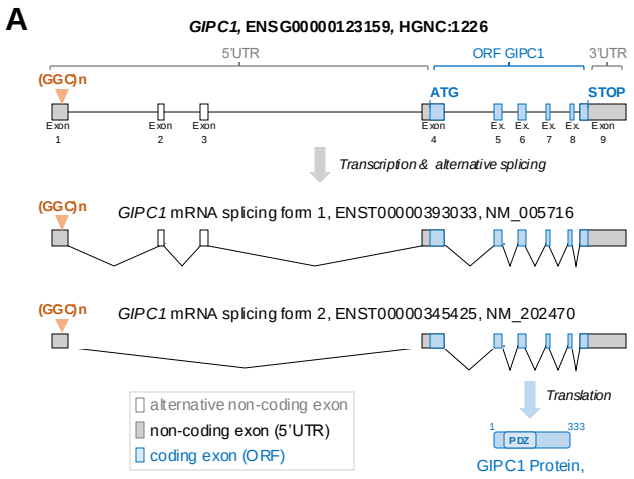
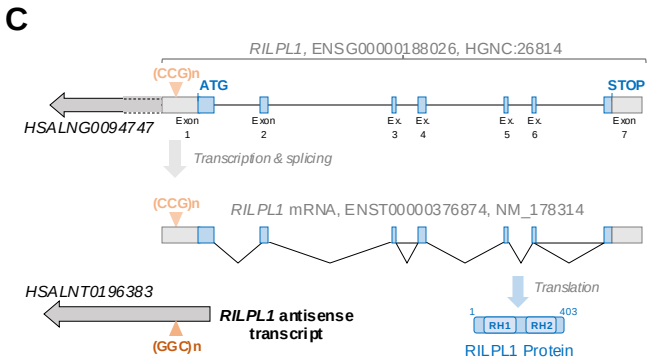
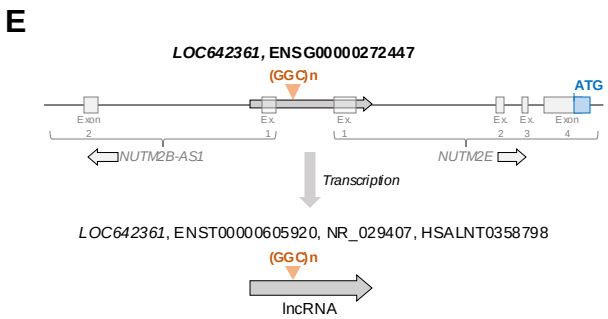
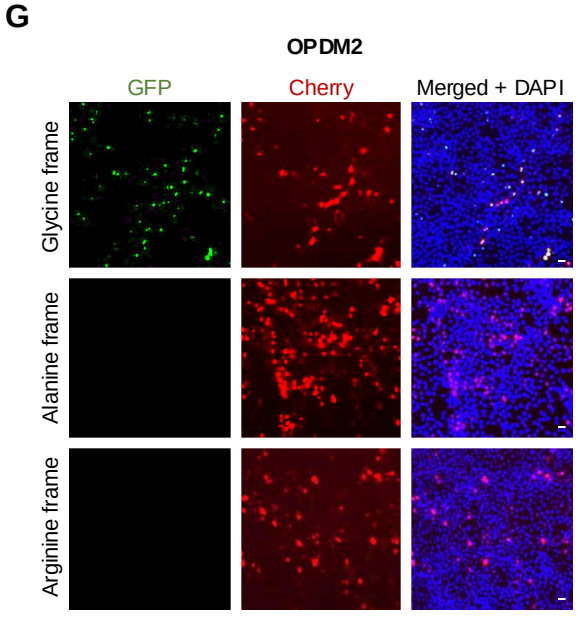
Supplemental Information

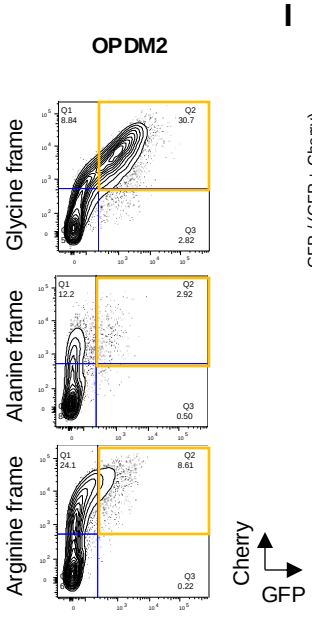
Microsatellite expansions hidden within the human "dark" genome are translated into novel and toxic proteins causing muscle and neurodegenerative diseases

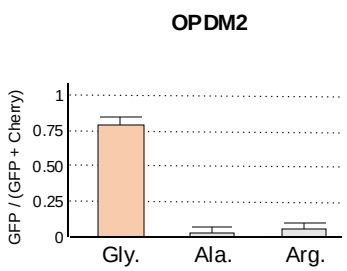












B

GIPC1 exon1 **(GGC)n** *BamHI* - GFP (glycine frame):

GGCGCCGCTGCGGGAGAGGCCCGGTCTGGAGTTTGCGGAGGGCCGAGCCGGGTGCGCACGGGGAGGC
GGAGGCAGC**GGCGGCGGC**...**GGCGGCGGC**G*ggatcc*GCGTCAGCAAAGGCGAGGAGCTTTTCACCGGG
GTTGTTCCCATCCTCGTCGAGCTCGACGGCGACGTAAACGGCCACAAATTCAGCGTTTCCGGCGAGG
GCGAGGGCGATGCCACCTACGGCAAGCTCACCCTCAAGTTCATCTGCACCACCGGCAAGCTCCCCGT
CCCCTGGCCCACCCTCGTCACCACCCTCACCTACGGCGTCCAGTGCTTCAGCCGCTACCCCGACCAC
ATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCT
TCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCG
CATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAAC
TACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGA
TCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGG
CGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCC
AACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGG
ACGAGCTGTACAAG*tga*

D

***SENSE TRANSCRIPT**: RILPL1* exon1 **(CCG)n** *BamHI* - GFP (proline frame):

GGGGAAGCGATCACCCGGCCTCGCAGCCGCCGCCGCCAAAGTTTCCCGGAGGAGCCGCGGGCCGCCC
TTCCTCTCCGGCCCCGCGCTCTGCCGCCCCTCAGGCCCGCACCGCCCCGAGGCGCCCGGCCCGGCTC
CATTGAGCGCAACTCTGATCCCAACTTGGGAGGCGCCGCGGGCCTGCGCACGCGCGCGCCCTCCCCG
CTG**CCGCCGCCG**...**CCGCCGCCGCC**C*ggatcc*GTCAGCAAA...

***ANTISENSE TRANSCRIPT:** asRILPL1* **(GGC)n** *BamHI* - GFP (glycine frame):

CTTCTGGTGCTTGCGCTCCTTCTCGATGCGGTCCATCCTCTCCAGGCGCAGGCGGTCCAGCTCCAGG
CGCAGCTCGTCCAGCTCGGGCGCGACGTGGTGGCGGCTGACCAGCACCTCCAGGATCTCCAGGACGC
GCACGACCTTGGGCATGAGGCGCGCGATGGCCTCGCAGCCGTGCTGGTCAATGACCCGCTCGAACTC
GTGGCCCACAAGCGACGCGATGTCGTACACGTCCATGACGGTCAGCTCGGCCACGTTCTTCTCCAGC
GCCGACTCGGCCGCCAGCGCCGACCCCCGCTCCTCCTCCATGGCCACCCTCCTGGCCTGTCCCCGC
CCCGCAAACTCGTGCAACTCCCAAACTTGCCGCTGTCGAGGGCCGGGCCGGCCGGGCCCAGCCTGGG
CCGCGGGCGGGCGCGCTCAGCGGGCGCTGGGGCGAGGGCGCAGCGGGCAGCGGGCGGCGGGCGCGGG
CGCGGGCACGGGCGGCGGCGCGGGGTGTGCGGGCCCGGGTCTGGGCGCCCGGCTCGGCCCGGAGCT
GCTCCCGAGTG**GGCGGCGGC**...**GGCGGCGGC**AGC*ggatcc*GTCAGCAAAGGCGAGGAGCTTTTCACCG
GGGTTGTTCCCATCCTCGTCGAGCTCGACGGCGACGTAAA..

Figure S1. Related to Figure 1.

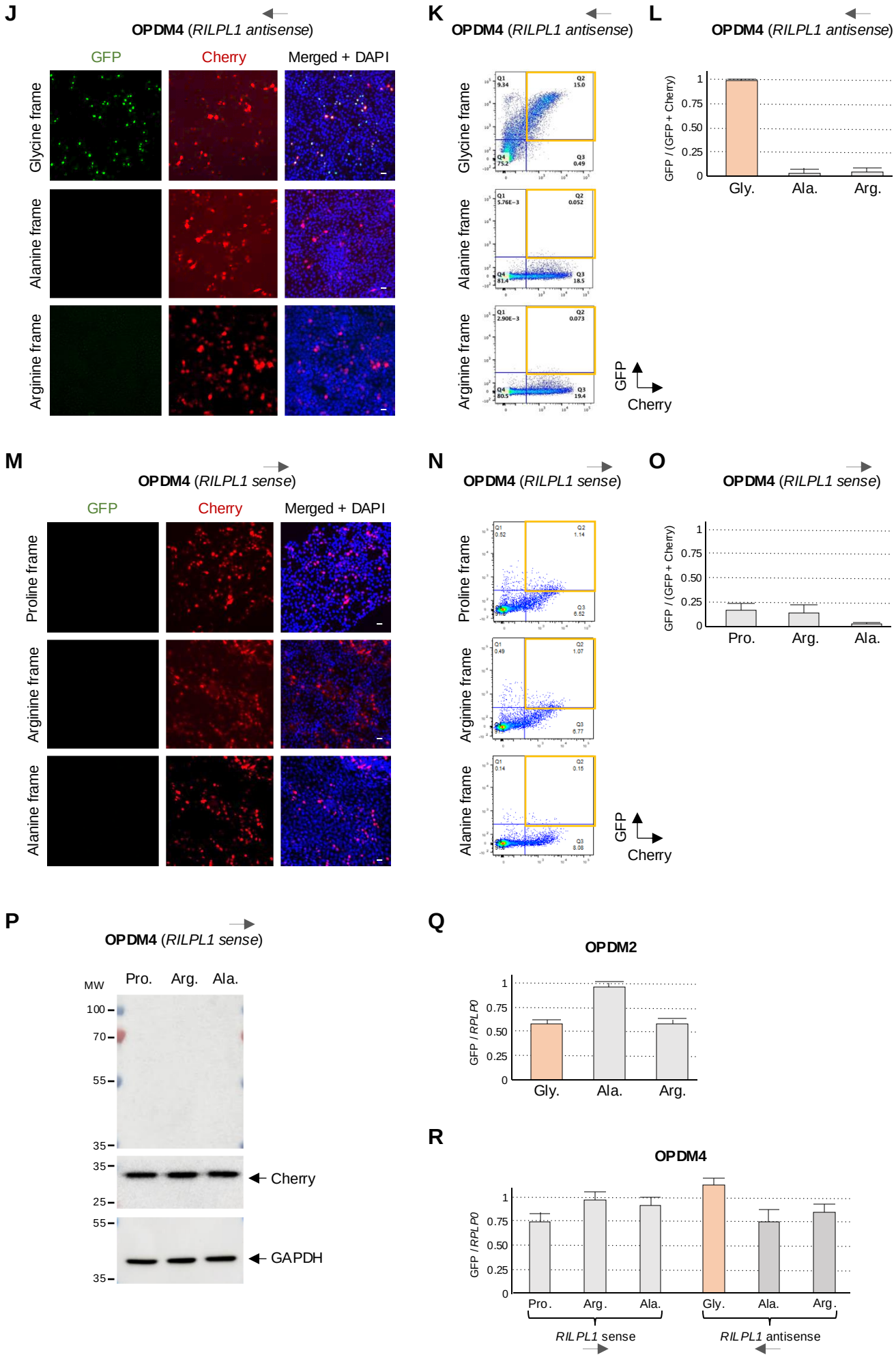
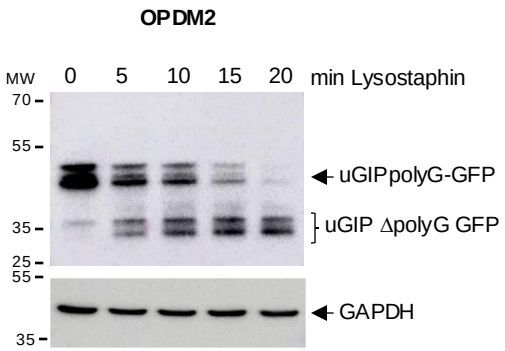
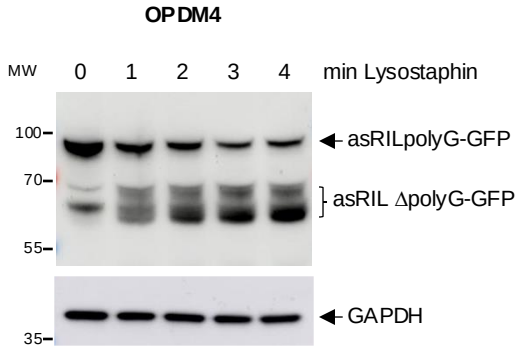


Figure S1. Related to Figure 1.

S



T



U

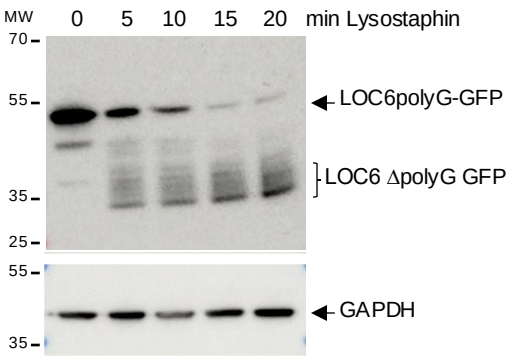


Figure S1. Related to Figure 1.

Figure S1. Related to Figure 1.
OPDM and OPML GGC repeats are translated into novel polyglycine proteins.

(**A**) OPDM2 is caused by an expansion of GGC repeats in the 1st exon/ 5'UTR of the *GIPC1* gene (HGNC:1226), composed of 9 exons, with its exons 4 to 9 encoding GIPC1, an adaptor protein that regulates cell surface receptor trafficking. *GIPC1* pre-mRNA is subject to alternative splicing with facultative inclusion of its non-coding exons 2 and 3, resulting in two mRNA isoforms (NM_005716, NM_202470) where *GIPC1* exon 1 is either bridged to exon 2 or to exon 4.

(**B**) Sequence of the human *GIPC1* exon 1 / 5'UTR with GGC repeats fused to the GFP in the glycine frame and cloned into pcDNA3.

(**C**) OPDM4 is caused by an expansion of CCG repeats located in the 1st exon/ 5'UTR of the *RILPL1* gene (HGNC:26814). This gene is composed of 7 exons, where the CCG repeats are located 237 nucleotides upstream of the ATG start codon of RILPL1, a RAB GTPase interacting protein that regulates ciliogenesis. Of interest, the RILPL1 gene is bidirectionally transcribed resulting in expression of an antisense transcript, asRILPL1 (HSALNT0196383), containing an expansion of GGC repeats in that direction.

(**D**) Sequences of the human sense and antisense *RILPL1 transcripts* with their CCG/GGC repeats fused to the GFP in the proline (sense) or glycine (antisense) frames.

(**E**) OPML is caused by an expansion of GGC repeats located in *LOC642361*, a 1.7 kb long non-coding RNA located between the *NUTM2B-AS1* and *NUTM2E* genes. According to what database is exploited, LOC642361 corresponds to the RefSeq NR_029407, to the Gencode: ENST00000605920 or to the LncBook HSALNT0358798, but with an identical sequence and consistently annotated as non-coding.

(**F**) Sequences of the human *LOC642361* lncRNA with its GGC repeats fused to the GFP in the glycine frame and cloned into pcDNA3.

(**G-I**) GFP and Cherry direct fluorescence (G), FACS analysis (H) and FACS quantification (I) of HEK293 cells transfected for 24 hours with a plasmid expressing GGC repeats embedded in the OPDM2 *GIPC1* 5'UTR sequence fused to the GFP in the glycine, alanine or arginine frame, while the Cherry is expressed independently.

(**J-L**) GFP and Cherry direct fluorescence (J), FACS analysis (K) and quantification (L) of HEK293 cells transfected for 24 hours with a plasmid expressing GGC repeats embedded in the OPDM4 *RILPL1* antisense transcript fused to the GFP in the glycine, alanine or arginine frame. The Cherry protein is expressed independently.

(**M-P**) GFP and Cherry direct fluorescence (M), FACS analysis (N), FACS quantification (O) and immunoblot (P) of HEK293 cells transfected for 24 hours with a plasmid expressing GGC repeats embedded in the OPDM4 *RILPL1* 1st exon/ 5'UTR fused to the GFP in the proline, alanine or arginine frame. The Cherry protein is expressed independently.

(**Q-R**) RT-qPCR of the GFP and *RPLP0* RNA expression of HEK293 cells transfected for 24 hours with GGC repeats embedded in the OPDM2 *GIPC1* 5'UTR sequence (Q), or CCG/ GGC repeats in the OPDM4 *RILPL1* sense/ antisense transcripts (R) and fused to the GFP in their three possible frames.

(**S-U**) Immunoblot against the GFP or the GAPDH of lysostaphin-digested proteins extracted from HEK293 cells transfected for 24 hours with GGC repeats embedded within *GIPC1* exon 1 / 5'UTR (S), OPDM4 *RILPL1* antisense transcript (T) or OPML *LOC642361* (U) fused to the GFP in the glycine frame.

A

OPDM2 (*GIPC1* 5'UTR CTG start codon)

#1	b⁺	b²⁺	Seq.	y⁺	y²⁺
1	174.06	87.53	M-Acetyl		
2	303.10	152.05	E	708.33	354.67
3	450.17	225.59	F	579.29	290.15
4	521.21	261.11	A	432.22	216.61
5	650.25	325.63	E	361.18	181.10
6	707.27	354.14	G	232.14	116.57
7			R	175.12	88.06

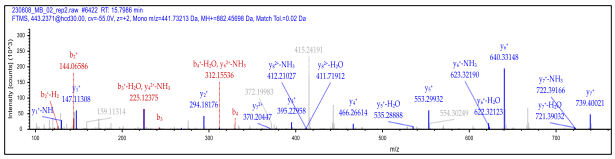
B

OPDM4 OPDM2 (antisense *RILPL1* ATG #1 start codon)

#1	b⁺	b²⁺	b³⁺	Seq.	y⁺	y²⁺	y³⁺
1	174.06	87.53	58.69	M-Acetyl			
2	330.16	165.58	110.72	R	859.55	430.28	287.19
3	417.19	209.10	139.74	S	703.45	352.23	235.15
4	530.28	265.64	177.43	I	616.41	308.71	206.14
5	643.36	322.18	215.12	L	503.33	252.17	168.45
6	730.39	365.70	244.14	S	390.25	195.63	130.75
7	886.49	443.75	296.17	R	303.21	152.11	101.74
8				K	147.11	74.06	49.71

C

OPDM4 OPDM2 (antisense *RILPL1* ATG #2 start codon)



acgtccATGACGGTCAGCTCGGCCACGTTCAAATCCAGCGCCGACTCG…

M[ac] T V S S A T F K S S A D S …

#1	b⁺	b²⁺	Seq.	y⁺	y²⁺
1	43.02	22.01	M-Acetyl		
2	144.07	72.54	T	840.45	420.73
3	243.13	122.07	V	739.40	370.20
4	330.17	165.59	S	640.33	320.67
5	417.20	209.10	S	553.30	277.15
6	488.24	244.62	A	466.27	233.64
7	589.28	295.15	T	395.23	198.12
8	736.35	368.68	F	294.18	147.59
9			K	147.11	74.06

D

OPDML (*LOC642361* lncRNA ATG start codon)

#1	b⁺	b²⁺	Seq.	y⁺	y²⁺
1	174.06	87.53	M-Acetyl		
2	303.10	152.05	E	1271.58	636.30
3	400.15	200.58	P	1142.54	571.77
4	560.18	280.60	C-Carbamidomethyl	1045.49	523.25
5	631.22	316.11	A	885.46	443.23
6	817.30	409.15	W	814.42	407.71
7	916.37	458.69	V	628.34	314.67
8	973.39	487.20	G	529.27	265.14
9	1044.43	522.72	A	472.25	236.63
10	1141.48	571.24	P	401.21	201.11
11	1270.52	635.77	E	304.16	152.58
12			R	175.12	88.06

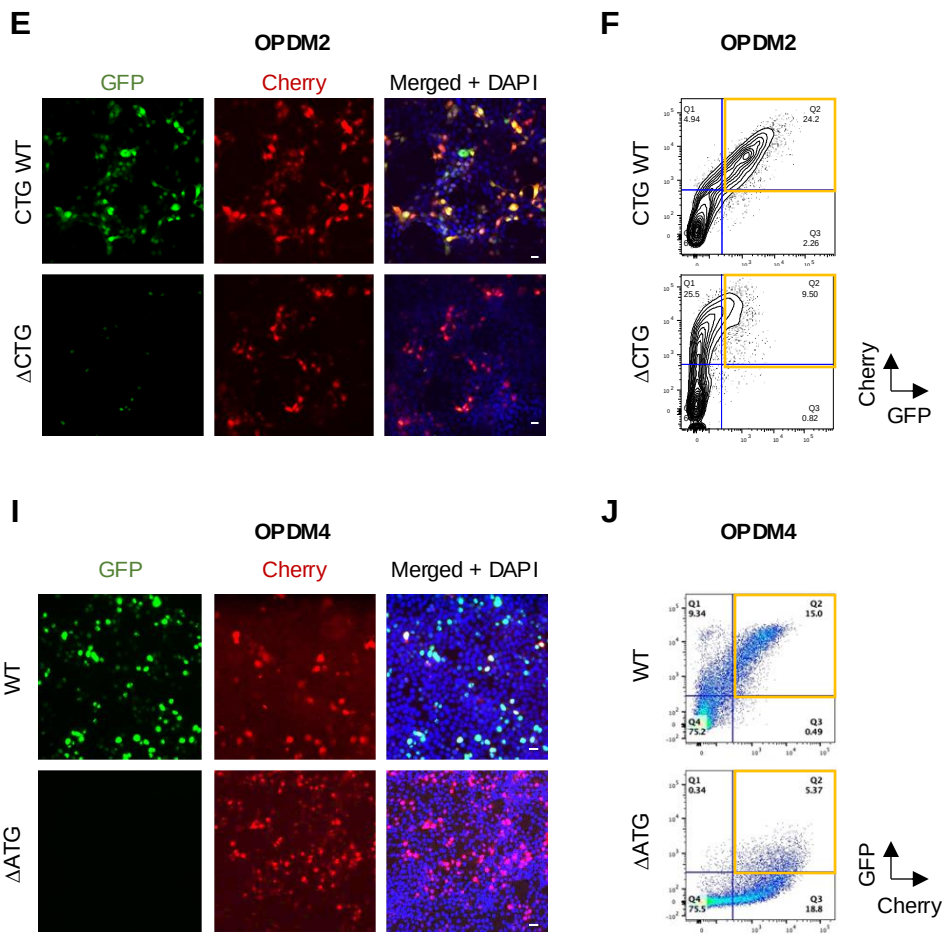
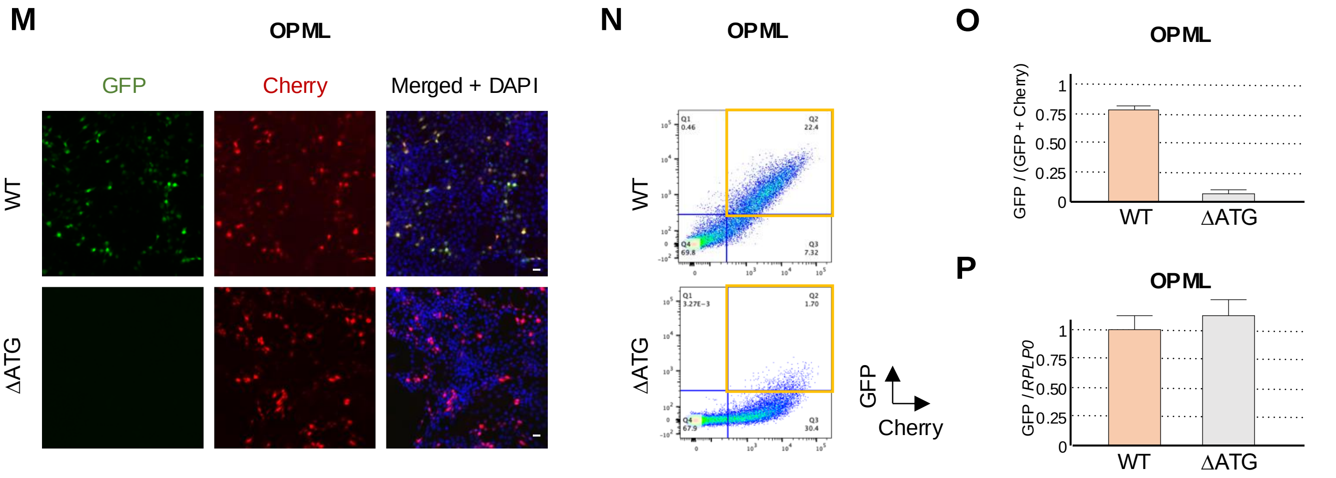


Figure S2. Related to Figure 2.



Q

***GIPC1* EXON 1 to start of intron 4:**

GGCGCCGCTGCGGGAGAGGCCCGGTCTGGAGTTTGCGGAGGGCCGAGCCGGGTGCGCACGGGGAGGCGGAGGCAGCGGCGGCGGC...GGCGGCGGCGGAGCAG*gtgagtggcctc...cccattttttcag*ATGAAG
AAACTGAGGCCCTGTGATGTGAAGTGACTTGCCCCCCAGCCACACAGCTG*gtaagtggcag...tatgttttacag*GACCATTCTGGCTGCTGTCTGGACAAGAAGTCGTAGGGGGTGAGGGTGGAAGCTGGGAA
ACCCACAGGAGGCAACCACACTAGTTTAG*gtaagtgatcat...cctgtcctgcag*ATCTTCTGGTGACCCCACTTCTCGCTGCTCATGCCGCTGGGACTGGGGCGGCGGAAAAAGGCGCCCCCTCTAGTGGAAA
ATGAGGAGGCTGAGCCAGGCCGTGGAGGGCTGGGCGTGGGGGAGCCAGGGCCTCTGGGCGGAGGTGGGTCGGGGGGCCCCCAAATGGGCTTGCCCCCCCCTCCCCCAGCCCTGCGGCCCCGCCTCGTGTTCCA
CACCCAGCTGGCCCATGGCAGTCCCACTGGCCGCATCGAGGGCTTCACCAACGTCAAGGAGCTGTATGGCAAGATCGCCGAGGCCTTCCGCCTGCCAACTGCCGAG*gtatccactggg...*

uGIPpolyG exons 1-2:

CTGGAGTTTGCGGAGGGCCGAGCCGGGTGCGCACGGGGAGGCGGAGGCAGCGGCGGC...GGCGGCGGAGCAG/ATGAAGAAACTGAGGCCCTGTGA
M E F A E G R A G C A R G G G G S G G ... G G G A D E E T E A L .

uGIPpolyG exons 1-4:

CTGGAGTTTGCGGAGGGCCGAGCCGGGTGCGCACGGGGAGGCGGAGGCAGCGGCGGC...GGCGGCGGAGCAG/ATCTTCTGGTGACCCCACTTCTCGCTGCTCATGCCGCTGGGACTGGGGCGGCGGAAAAAGGCGCCCCCTCTAGTGGAAAATGA
M E F A E G R A G C A R G G G G S G G ... G G G A D L L V T P L L A A H A A G T G A A E K G A P S S G K .

R

***asRILPL1*:**

acttacctaaagtcacacagctattaaacggtaaaggtgggattggaactccagagaacagctcttaatctctcgggtggtatctgccagccacccctggtcgggcaaaagaaggccagggcctc cctccctc
cccacgtcgggtctcctctggtccggtccctgaacacacacacgtgcgcacccacagctgcccaatgcggaagagcccttgggtccccgcggtcccactgcccggacggacagaccgaggccgcc gcccacct
tctggtgcttgcgctccttctcgATGCGGTCCATCCTCTCCAGGCGCAGGCGGTCCAGCTCCAGGCGCAGCTCGTCCAGCTCGGGCGCGACGTGGTGGCGGCTGACCAGCACCTCCAGGATCTCCAGGACGCG
CACGACCTTGGGCATGAGGCGCGCGATGGCCTCGCAGCCGTGCTGGTCAATGACCCGCTCGAACTCGTGGCCCACAAGCGACGCGATGTCGTACACGTCCATGACGGTCAGCTCGGCCACGTTCTTCTCCAGC
GCCGACTCGGCCGCCAGCGCCGACCCCCGCTCCTCCTCCATGGCCACCCTCCTGGCCTGTCCCCCGCCCCGCAAACTCGTGCAACTCCCAAACTTGCCGCTGTCGAGGGCCGGGCCGGCCGGGCC CAGCCTGG
GCCGCGGGCGGGCGCGCTCAGCGGGCGCTGGGGCGAGGGCGCAGCGGGCAGCGGGCGGCGGGCGCGGGCGCGGGCACGGGCGGCGGCGCGGGGTGTGCGGGCCCGGGGTCTGGGCGCCCGGCTCG GCCCGGAG
CTGCTCCCGAGTGGGCGGCGGCGGC...GGCGGCGGCAGCGGGGAGGGCGCGCGCGTGCGCAGGCCCGCGGCGCCTCCCAAGTTGGGATCCGAGTTGCGCTCAATGGAGCCGGGCCGGGCGCCTCGGGGCGGTGC
GGGCCTGAGGGGCGGCAGAGCGCGGGGCCGGAGAGGAAGGGCGGCCCGCGGCTCCTCCGGGAAACTTTGGCGGCGGCGGCTGCGAGGCCGGGTGATCGCTTCCCCACAGGGAGGCCGCGCGCCCC GCCCCAAT
CCCGGCCTGGCCAGGGGCGCGCCCGCTTTCCGCCGCCCGGGGGCGCCCCGGGGTACCCCCCCAAGCGCCGACTCCGCGGCCTGCACTCCGGGCCCCCCTCTCCAGCCGCCGGACCCGCGGCGCCC CTCACCCG
TGCCCCGGCCGCCCCTCCCCCGCCCGGCGCTGACTCCTCCCACCACCCGGCCGCCGGGAGGGGTCCAGGCCGCCCCGGGAGAGGGGGCCGGGGGCGCTGGTCTCGCGGGACGATTCGCTGCTCCG GGTAGCGG
CCCCGGCCTTGGAGAGGGATGGAGGGGTCACGAGGAGGTCCCCGGGCTGGGGCTCATCCCAGGAAGAGAACGGGGACCCCCCCCGGCCCGCAGACACTCCGCGCAGACCCGGCCACCAGGCGGAC TTGCAGAG
GCGCCCGGAGAGGGGCCGACGTGTCGAGGAATGCTGGATCCCCGGAGCCCCAGGGCGCGCCTCCCCGAGCCACATCCATCACTGGCTCCTTGTGACATGAACGCCACTTCACAG TAGagcgaagcccgcttgg
cattccagcaccacgagttctcttgtcccaaggcccccctgtccatttgggctggccaaatccacttgtccatcagacccccctcccactcccgggagaacgtttcctgaacccagaaccccatcca ccctcccc
gccccccacgcaggagttaaggacccccttgcaaactcacagtcccctgcttgccccatcaggattgctcttagcatcctgagttttcaccgtgtggttgcaggtaggtttccaccgtcagatct tggctcct
tgagggaagaaatcttaatgcccttgtagccataaagcgcaccccaacgtggcacgtattgattaaaaatctaaaagttggccgggcgcggtggctcacgcctgtaatcccagcactttgggaga ctgaggtg
gacgaggtcaggagttcaagaccagcctagccaaggtggtgaaaccccgtctctactaaaaatacaaaaattagccgggcatggtggcgcatgcctgtaatcccagctactcaggaggctgaggc agagagtt
gcttaaacccgggaggcggaggttgcagcaagccgaaatcgcgccactgcactgcaggctgggtgacagagcaagactccatctcaaaataataatataatacAATAAATAAATAAATAAAtagtaaaaaaaa
aaaaattaaggcCA

asRILpolyG:

MRSILSRRRRSSSRRSSSSSGATWWRLTSTSRISRTRTTLGMRRAMASQPCWSMTRSNSWPTSDAMSYTSMTVSSATFFSSADSAASADPRSSSMATLLACPPPRKLVQLPNLPLSRAGPAGPSLGRGRARSA
GAGARAQRAAGGGRGRGHGRRRGVCGPGVWAPGSARSCSRVGGGG...GGGGSGEGARVRRPAAPPKLGSELRSMEPGRAPRGGAGLRGGRARGRRGRAARGSSGKLWRRRLRGRVIASPQGGRAPRPNPGLARG
APAFRRPGAPRGTPPSADSAACTPGPPLQPPDPRRPSPVPRPPLPRPALTPPTTRPPGGVQAAPGEGAGGAGLAGRFAAPGSGPGLGEGWRGHEEVPGLGLIPGRERGPPPARRHSAQTRPPGGL AEAPGEGP
TCRGMLDPRSPRARLPEPHPSLAPCDMNATSQ.

S

LOC642361:

caaggaccaggagcgggagccgagcctcgcggcccggggcacggcgccgcctgcccctccccggcagtgcctcccgacccccctgccctccgctgggacatggcctgaggccccgcccgccgcccg ggacccct
cccccacgctgcccccggaccctcgccccaggaccgcggggtctcttcccggccccgccgcccgccccggcgagaacaggcccggcgggcaaggcggcggcggaccgagggaggcctggcccggc agcgggga
gaagggtgcggcgcagcccgagtttcccacctttttatcctggcccagacgcggctggggcggacgggacctctcgcgctctgcctcctcctcttgcttcATGGAGCCATGCGCCTGGGTGGGGGCTCCCGAGA
GAAGCTGGCCTGCGGGCGGGCCGGACGCGCTGCGCGGACGGACGGGGCGAAGGAGGCCGGCAGGCGGAGGAGGAAGCGGCGG GGCGGCGGC...GGCGGCGGCCGGGAAGAACTAGAGGTATTCCCCGGGCGGC
TGGAGGACTGAgtcgagccgggacccgagtcctccggtatcccagcagccaccggaggcagtgaggtaatggaggaagaatgtagggagtcctccggtggaccacactccttctagcgagccgcggaaaccat
agagatcagggctcagccggagggccggcccactgctgtcacgtggcctccatccgtgcgctttattggctggtgctggctttacggggttgaattttggacgctgcctgccattcgtgggaccg agtttcag
gcagctggaataaagagagatatcgggggaaagccgcgagaaggccagcgtccctcgctgggagcagagctagccgagtagggcgcccggctgtcaaactggccggcgcagtgcacgctgggccg ccccggag
cgatcgcagagcccgtcgggagtcgtagtcccggacgggcccgcggcatcgtccgcgacgcagctcgggatccgggtcggggtgtcggccgggttgctgccgggcaccgtcgagcgttagccacc cagcattg
agctgccagcggctgttctccctaagcaccctcgctcacgtcgcctcgcctcgcctgaccggccgcagccttggataccagccttaaatcgagccgactacggcccagccccgcccgcggcgagg tgcgcggg
gttcggtgcgagcccctcgcccctggccgggccagcctcttgatgcaccggggacgggcgcgcccatccgactcttcgaggggcggcagggccccatctgtgtctttcgctctcgagcccccagc tagagttg
gcttcagcggaatacctactgtgcgggattattcaacaagccgattgatcacattcttcagctctagcagtgcaaaggcttcactgaaaaagaacattacaactttttttgaaagaaatttggggg taaattta
attttaaatactgattttttgataacatttgatcttatgacagcatttgtatttttaattacagataggattacactaaaagtcatttctgacctgctcccttacccttatggctaaAATAAActgctaggccc
tgaagcaaaggcgagccatgttctagagcaatgtctttcttgtattaaaatacagccccatgctcacaaaaaaaaaaaaaaaaaatttggagtgtaaaCA

LOC6polyG:

MEPCAWVGAPERSWPAGGPDALRGRDGAKEAGRRRRKRRGGGG...GGGGREELEVFPGRLED.

Figure S2. Related to Figure 2.

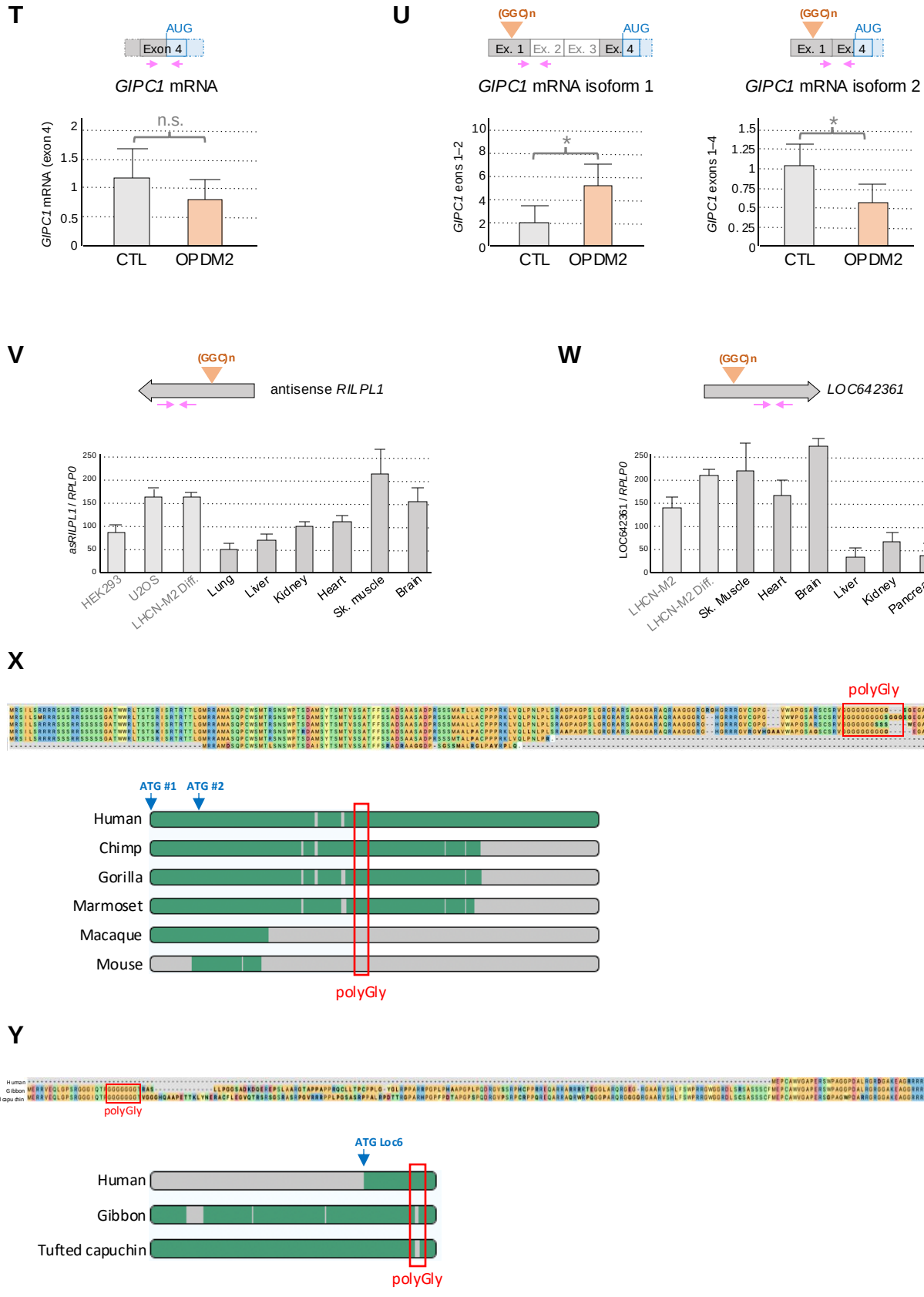


Figure S2. Related to Figure 2.

Figure S2. Related to Figure 2.
OPDM and OPML GGC repeats are translated through initiation at start codons.

(**A**) N-terminal peptide sequence of the LC-MS/MS analysis of GFP-immunoprecipitated proteins extracted from HEK293 cells transfected 24 hours with OPDM2 GGC repeats located in *GIPC1* 5'UTR and fused to the GFP in the glycine frame.

(**B-C**) N-terminal amino acid sequences as in (A) but from OPDM4 *RILPL1* antisense transcript fused to the GFP in the glycine frame. Of interest, translation initiation occurs at two ATG start codons, both located upstream of the GGC repeats.

(**D**) N-terminal amino acid sequences as in (A) but from OPML *LOC642361* lncRNA fused to the GFP in the glycine frame.

(**E-H**) GFP and Cherry direct fluorescence (E), FACS analysis (F), FACS quantification (G) and GFP RNA quantification by RT-qPCR (H) of HEK293 cells transfected for 24 hours with either a wild-type or a ΔCTG start codon plasmid expressing GGC repeats embedded in the OPDM2 *GIPC1* 5'UTR sequence fused to the GFP in the glycine frame. The Cherry protein is expressed independently.

(**I-L**) GFP analyzes as in (E-H) but transfected with either a wild-type or a ΔATGs start codons plasmid expressing GGC repeats embedded in the OPDM4 *RILPL1* antisense transcript fused to the GFP in the glycine frame.

(**M-P**) GFP analyzes as in (E-H) but transfected with either a wild-type or a ΔATG start codon plasmid expressing GGC repeats embedded in the OPML *LOC642361* lncRNA fused to the GFP in the glycine frame.

(**Q**) Upper panel, sequence of the OPDM2 upstream ORF (uORF) embedded in the 5'untranslated region of the *GIPC1* gene and that spans either exons 1 and 2 or exons 1 and 4 according to *GIPC1* mRNA alternative splicing. GIPC1 uORF is indicated in bold upper cases with its exon 1 CTG start codon, and its exon 2 or 4 stop codons annotated in underlined blue. Exon 4 ATG start codon of the main GIPC1 ORF encoding the GIPC1 protein is annotated in underlined green. Exons are annotated in upper cases, while introns are indicated in lower cases. Lower panel, translation of *GIPC1* uORF results in expression of a novel uGIPpolyG protein, which is composed of a 17 amino acids N-terminal part, a variable central stretch of polyglycine corresponding to the number of GGC repeats, and either a short 8 amino acids or a 28 amino acids C-terminus, encoded respectively by GIPC1 exon 2 or exon 4 according to *GIPC1* alternative splicing with its exon 1 bridged either to exon 2 or exon 4.

(**R**) The OPDM4 GGC repeats in the *RILPL1* antisense transcript are located in a novel ORF indicated in bold upper cases with its ATG start codons and its stop codon annotated in underlined blue. Predicted polyadenylation site is indicated as underlined upper cases.

(**S**) The OPML GGC repeats are located in the 1.7 kb *LOC642361* lncRNA, which contains a small open reading frame, indicated in bold upper cases. As confirmation of presence of an ORF in this lncRNA, exploration of ribosome profiling database indicates that the ribosome is engage onto this small ORF, which is annotated as SPROHSA186008.

(**T-U**) RT-qPCR performed on muscle tissue of individuals with OPDM2 shows no significant differences of *GIPC1* mRNA expression compared to control individuals (T). Isoform specific RT-PCR indicates an increase alternative splicing of *GIPC1* exon 1 toward exon 2 in OPDM2 skeletal muscle samples, with concomitant decrease splicing of exon 1 to exon 4 (U).

(**V-W**) RT-qPCR quantification of *RILPL1* antisense transcript (V) or of *LOC642361* lncRNA (W).

(**X-Y**) The ORF contained in *RILPL1* antisense *RNA* is conserved among primates, with presence of a conserved polyglycine stretch in chimpanzee, gorilla and marmoset, but is predicted as a much shorter protein with no polyglycine in macaque and other mammals including mouse

(**Y**) The *LOC642361* ORF is unfound in most mammals, but strikingly identical to the C-terminal part of a longer protein of unknown function found in Gibbon Lesser Apes and Tufted capuchin.

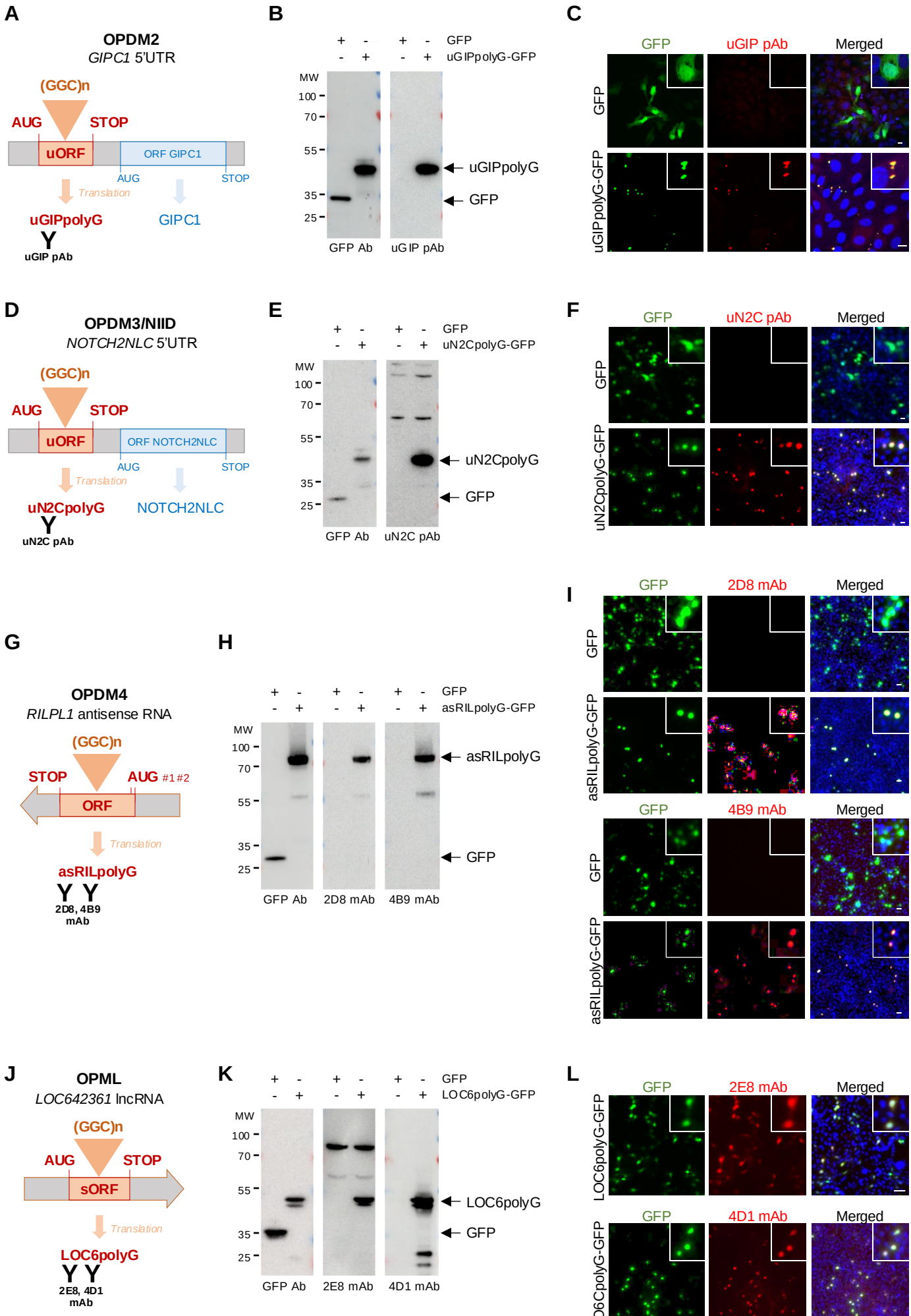


Figure S3. Related to Figure 3.

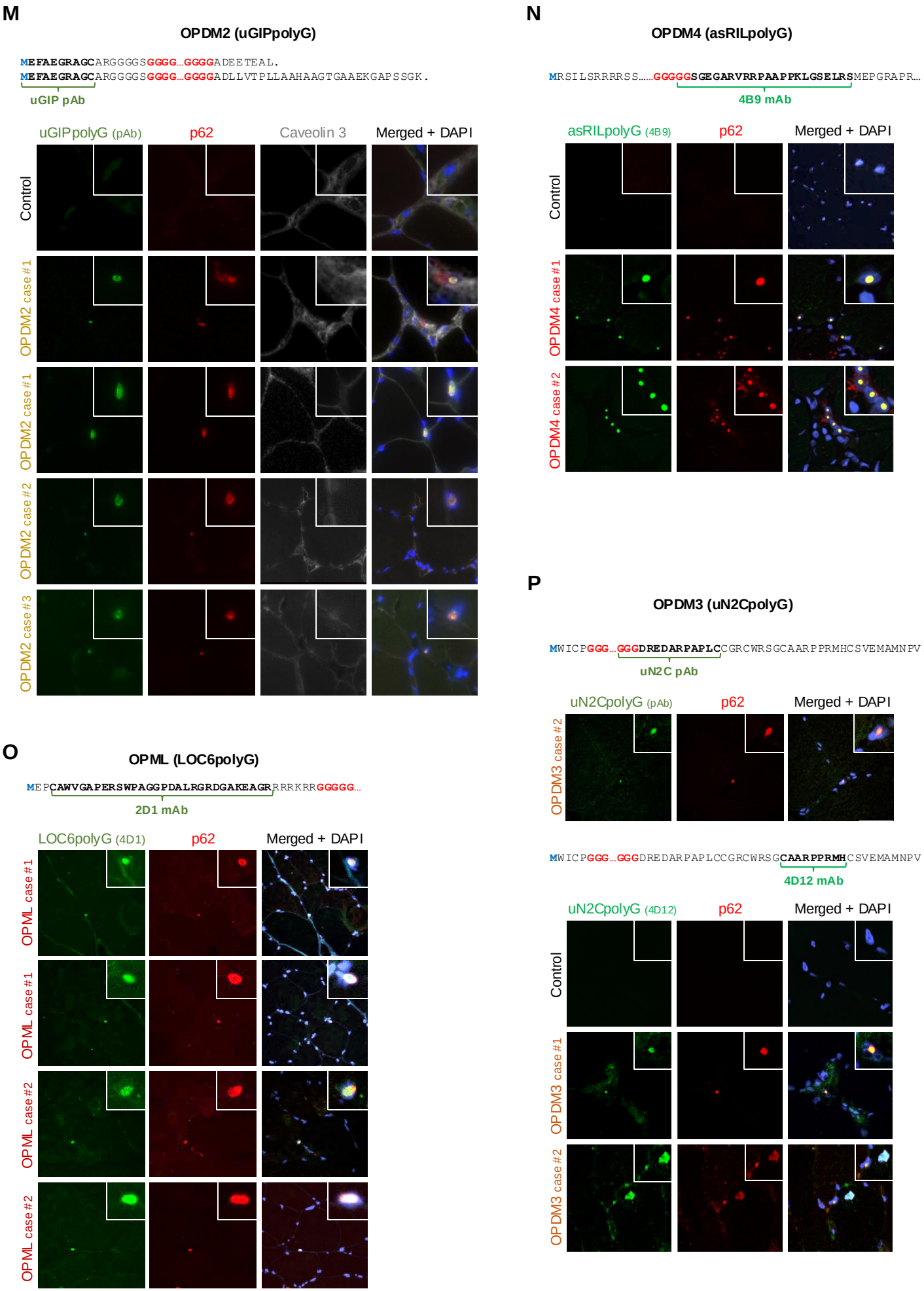


Figure S3. Related to Figure 3.

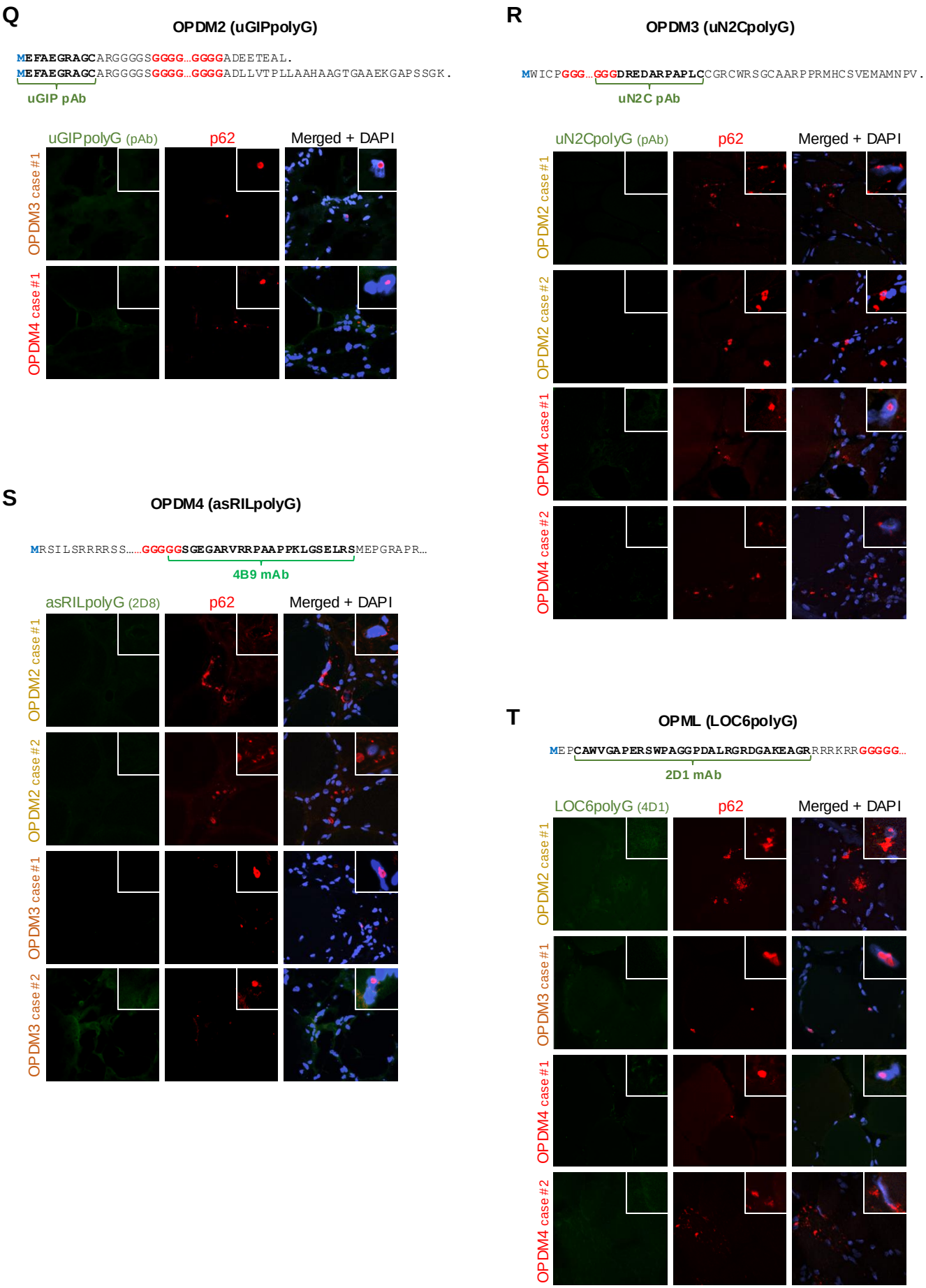


Figure S3. Related to Figure 3.

U

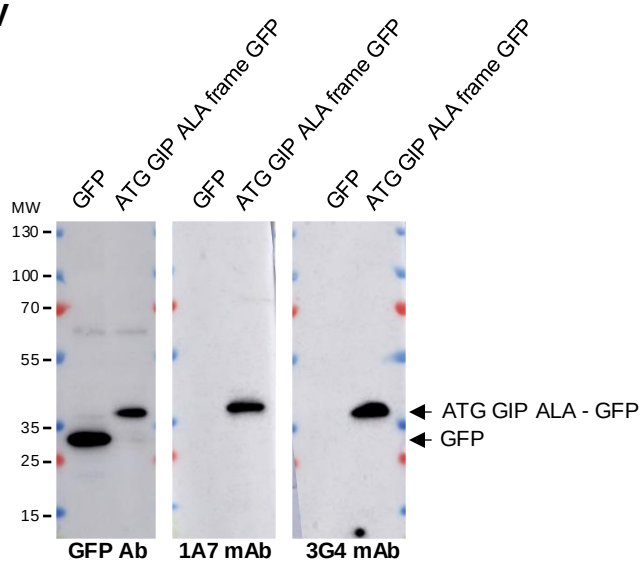
Putative GIPpolyALA protein:

Exons 1-2: …RRAEPGAHGEAEAAA AAAEQIFW.

Exons 1-4: …RRAEPGAHGEAEAAA AAAEQMKKLRPCDVK.

1A7 & 3G4 mAb

V



W

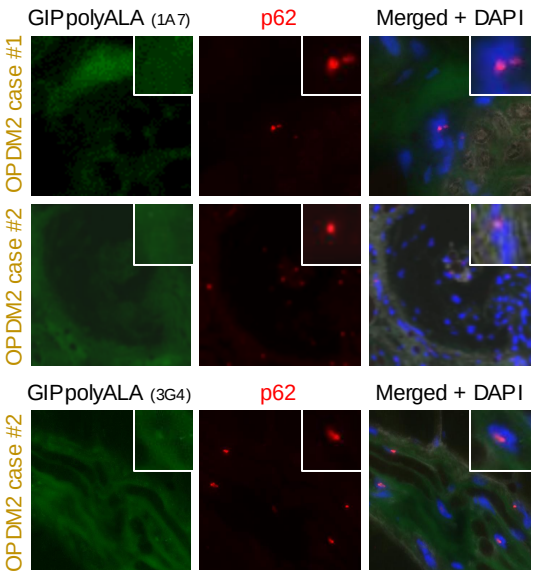


Figure S3. Related to Figure 3.

Figure S3. Related to Figure 3.
Polyglycine proteins are present in the typical OPDM/OPML p62-positive inclusions.

(**A-C**) Scheme of the uGIP antibody directed against the OPDM2 upstream ORF contained in *GIPC1* 5'UTR (A), and its validation by immunoblotting (B) and immunofluorescence (C) on HEK293 cells transfected either with GFP or uGIPpolyG-GFP.

(**D-F**) Scheme of the uN2C antibody directed against the OPDM3/NIID upstream ORF contained in *NOTCH2NLC* 5'UTR (A), and its validation by immunoblotting (B) and immunofluorescence (C) on HEK293 cells transfected either with GFP or uN2CpolyG-GFP.

(**G-I**) Scheme of the 2D8 and 4B9 antibodies directed against the OPDM4 ORF contained in *RILPL1* antisense RNA (A), and their validation by immunoblotting (B) and immunofluorescence (C) on HEK293 cells transfected either with GFP or asRILpolyG-GFP.

(**J-L**) Scheme of the 2E8 and 4D1 antibodies directed against the OPML ORF contained in *LOC642361* lncRNA (A), and their validation by immunoblotting (B) and immunofluorescence (C) on HEK293 cells transfected either with GFP or LOC6polyG-GFP.

(**M**) Upper panels, amino acid sequence of the OPDM2 uGIPpolyG ORF with the peptide target of the uGIP antibody indicated in bold with an underlined bracket. Lower panels, immunofluorescence against p62, caveolin 3 and uGIPpolyG on skeletal muscle sections of individuals with OPDM2 or age-matched control individuals.

(**N**) Upper panels, amino acid sequence of the OPDM4 asRILpolyG ORF with the peptide target of the 4B9 antibody indicated in bold with an underlined bracket. Lower panels, immunofluorescence against p62 and asRILpolyG on skeletal muscle sections of controls or individuals with OPDM4.

(**O**) Upper panels, amino acid sequence of the OPML LOC6polyG ORF with the peptide target of the 2D1 antibody indicated in bold with an underlined bracket. Lower panels, immunofluorescence against p62 and LOC6polyG on skeletal muscle sections of individuals with OPML.

(**P**) Sequence of the OPDM3 uN2CpolyG ORF with the peptides target of the uN2C or 4D12 antibodies indicated in bold with an underlined bracket. Immunofluorescence against p62 and uN2CpolyG on skeletal muscle sections of controls or individuals with OPDM3.

(**Q-T**) Antibodies directed against the OPDM2 uGIP (Q), OPDM3 uN2C (R), OPDM4 asRILPL1 (S) or LOC642361 (T) ORFs are specific to each of these sequence and do not stain other OPDM subtypes caused by GGC repeat expansion in other ORF sequences.

(**U-W**) Scheme of the antibodies directed against a putative GIPC1 5'UTR polyalanine frame (U), with their validation by immunoblotting on HEK293 cells transfected either with GFP or GIPpolyALA-GFP (V) and negative staining on muscle sections of individuals with OPDM2 (W).

A

ATG polyGly GFP :
MEAPLPGGVRQR**(G)100x**GSVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDD
GNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK.

uGIPpolyG ex2 GFP :
MEFAEGRAGCARGGGGS**(G)100x**ADEETEALGSVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPE
GYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKR
DHMVLLEFVTAAGITLGMDELYK.

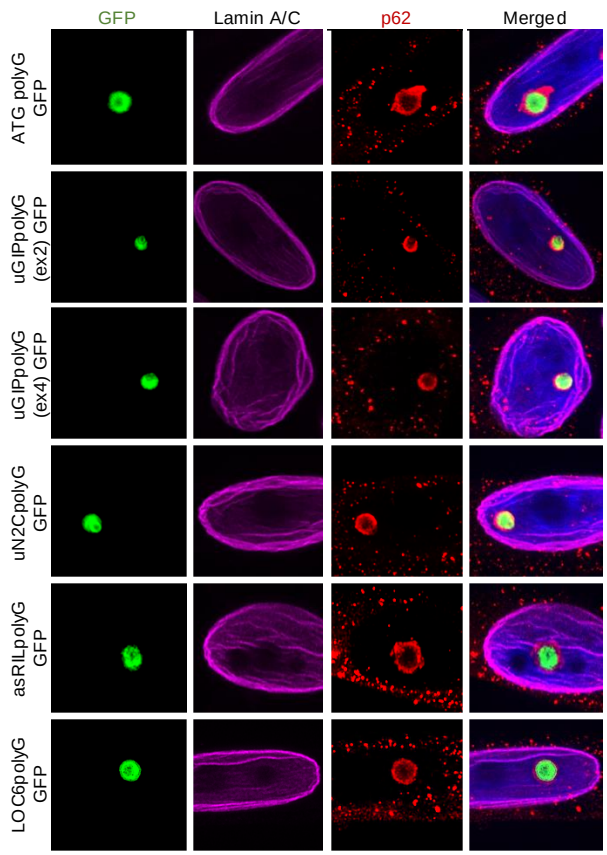
uGIPpolyG ex4 GFP :
MEFAEGRAGCARGGGGS**(G)100x**ADLLVTPLLAAHAAGTGAAEKGAPSSGKGSVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQC
FSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLL
PDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK.

uN2CpolyG GFP :
MWICP**(G)100x**DREDARPAPLCCGRCWRSGCAARPPRMHCSVEMAMNPVGSVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCF
SRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLP
DNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK.

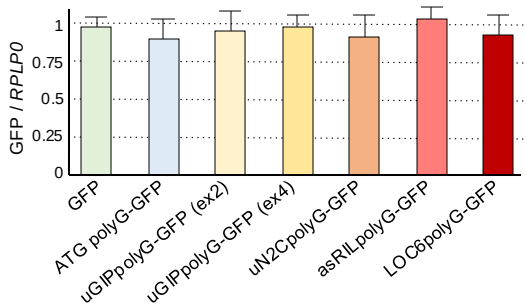
OPDM4 polyGly GFP :
MRSILSRRRRSSSRRSSSSSGATWWRLTSTSRISRTRTTLGMRRAMASQPCWSMTRSNSWPTSDAMSYTSMTVSSATFFSSADSAASADPRSSSMATLLACPPPRKLVQLPNLPLSRAGPA
GPSLGRGRARSAGAGARAQRAAGGGRGRGHGRRRGVCGPGVWAPGSARSCSRV**(G)100x**SGEGARVRRPAAPPKLGSELRSMEPGRAPRGGAGLRGGRARGRRGRAARGSSGKLWR
RRLRGRVIASPQGGRAPRPNPGLARGAPAFRRPGAPRGTPPSADSAACTPGPPLQPPDPRRPSPVPRPPLPRPALTPPTTRPPGGVQAAPGEGAGGAGLAGRFAAPGSGPGLGEGWRGHE
EVPGLGLIPGRERGPPPARRHSAQTRPPGGLAEAPGEGPTCRGMLVPRSPRARLPEPHPSLAPCDMNATSQGSVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICT
TGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDG
SVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK.

OPML polyGly GFP :
MEPCAWVGAPERSWPAGGPDALRGRDGAKEAGRRRRKRR**(G)100x**REELEVFPGRLEDGSVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVT
TLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNT
PIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITLGMDELYK.

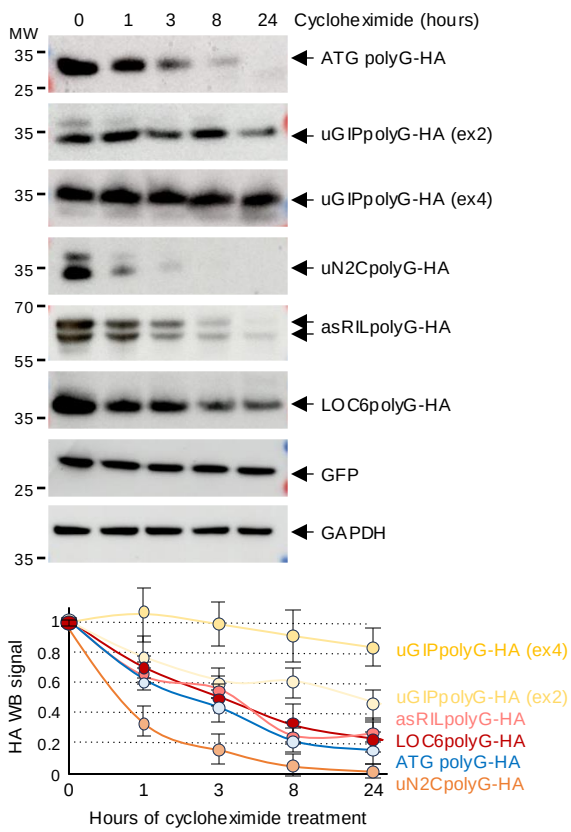
B



C



D



E

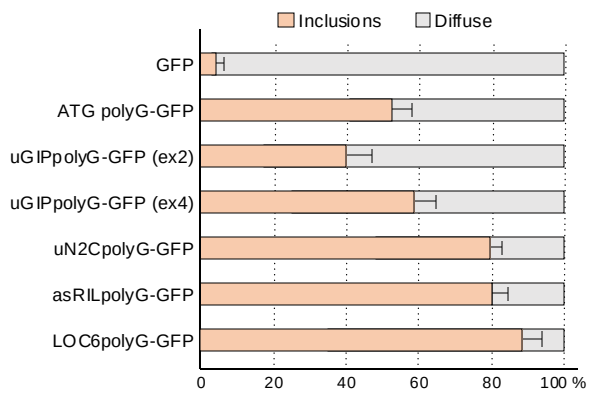


Figure S6. Related to Figure 6.

Figure S4. Related to Figure 4.
Expression of polyG proteins forms inclusions and is pathogenic for muscle cells.

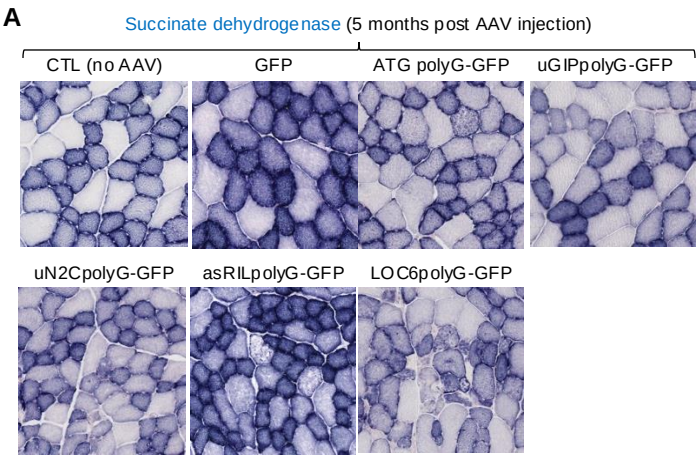
(**A**) Amino acid sequences of the artificial ATG-polyGly and the various OPDM/OPML GFP-tagged polyglycine proteins.

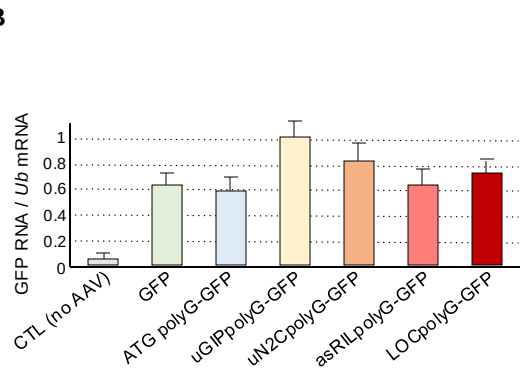
(**B**) GFP fluorescence and immunofluorescence against p62 and lamin A/C proteins of LHCN-M2 muscle cells differentiated for 4 days and expressing either the GFP or GFP-tagged ATG polyG, OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG.

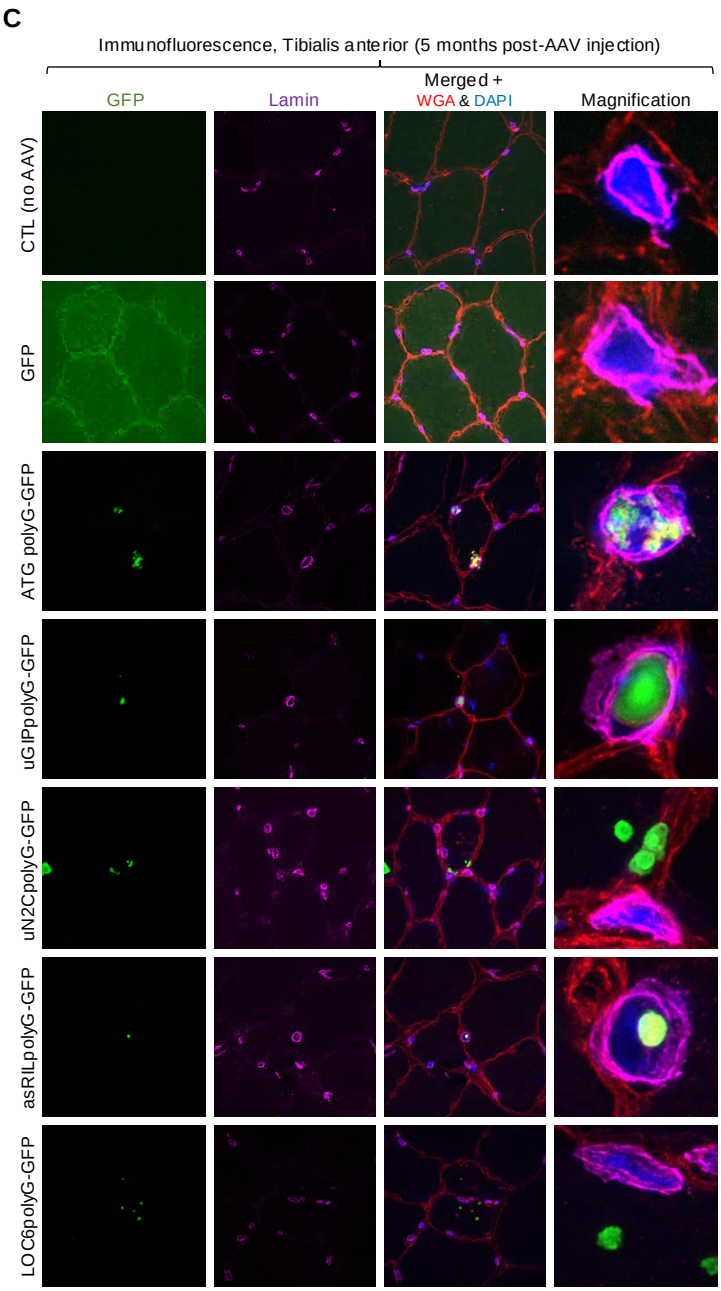
(**C**) RT-qPCR of the *GFP* and *Rplp0* RNA expression of LHCN-M2 muscle cells differentiated for 4 days and expressing controls or the diverse OPDM/OPML polyglycine-GFP tagged constructs.

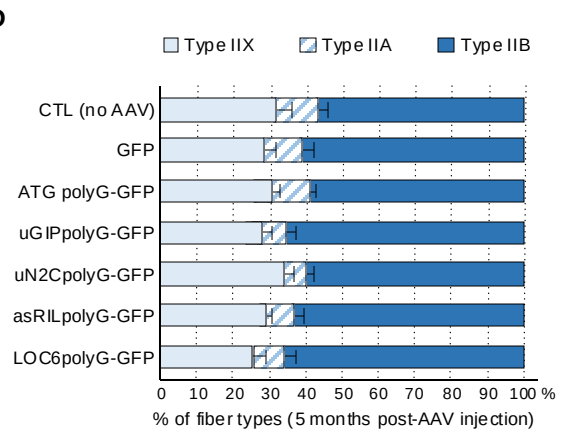
(**D**) Half life of the OPDM/OPML polyglycine proteins determined by cycloheximide chase experiment. Upper panel, immunoblotting against the GAPDH or HA tag of HEK293 cells expressing HA-tagged ATG polyG, OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG and treated with cycloheximide for 1, 3, 8 or 24 hours after 24 hours of transfection. Lower panel, immunoblot signal quantification.

(**E**) Quantification of the GFP signal diffuse vs. localized in inclusions of LHCN-M2 muscle cells differentiated for 2 days and expressing either the GFP or GFP-tagged ATG polyG, OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG.









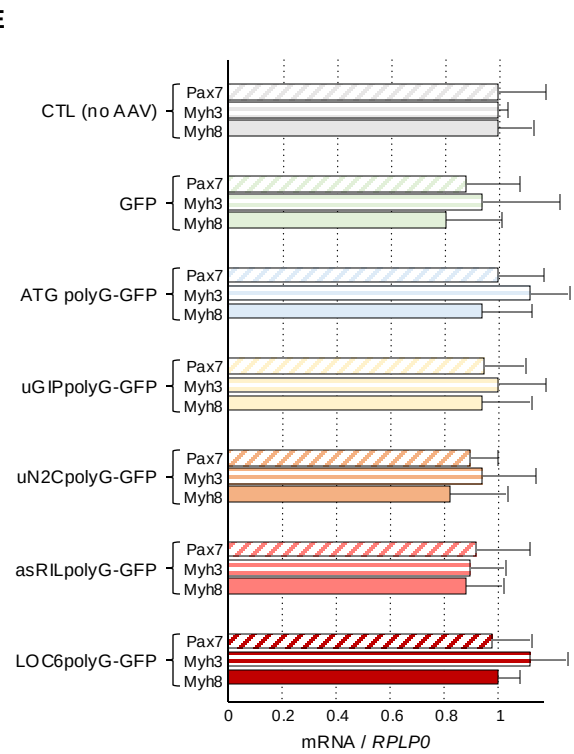


Figure S7. Related to Figure 7.

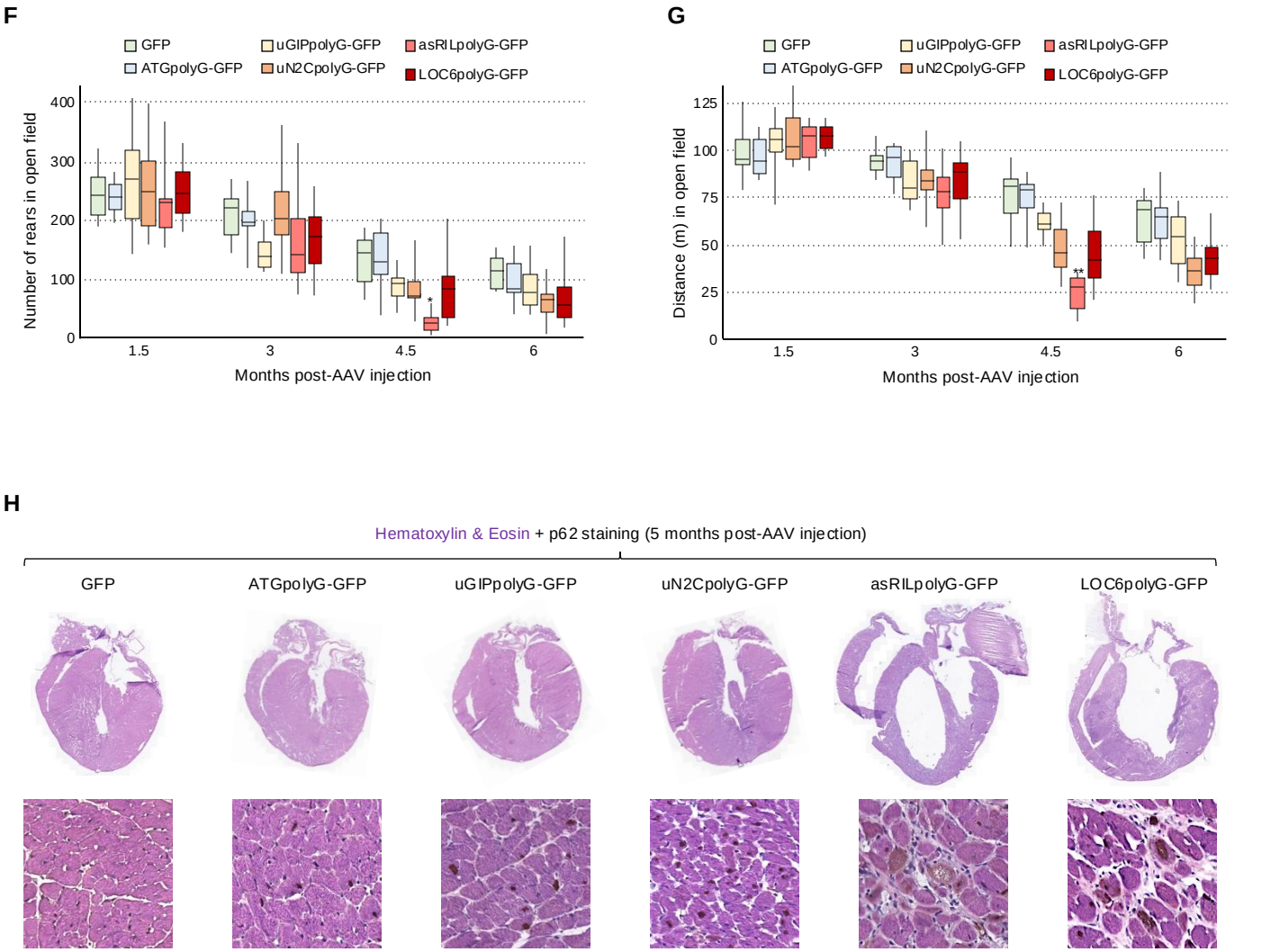


Figure S7. Related to Figure 7.

Figure S5. Related to Figure 5.
Expression of polyG proteins forms inclusions and is pathogenic in animal muscles.

(**A**) Succinate dehydrogenase staining of tibialis anterior (TA) frozen sections of 5-months AAV-injected male mice expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG.

(**B**) RT-qPCR of the GFP and Ubiquitin expression on RNA extracted from TA muscles of 5-months AAV-injected male mice expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG. N=3 mice per condition.

(**C**) GFP fluorescence and immunofluorescence against the lamin B1 and A/C with counterstaining of membranes by fluorescent-conjugated wheat germ agglutinin (WGA) and nuclear DNA by DAPI on frozen TA muscle sections 5-months post-injection of AAV expressing the OPDM/OPML polyglycine proteins.

(**D**) Quantification of muscle fiber types determined by immunofluorescence against MYH7, MYH2 and MYH4 on frozen TA muscle sections of male mice 5-months post-injection of AAV expressing the OPDM/OPML polyglycine proteins. N=3 mice per condition.

(**E**) RT-qPCR of *Pax7*, *Myh3* and *Myh8* expression on RNA extracted from TA muscles of 5-months AAV-injected male mice expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG. N=3 mice per condition.

(**F-G**) Number of rears (F) and maximal distance traveled (G) during 30 min in an open field 1.5-, 3-, 4.5- and 6-months post-injection of AAV expressing either the GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG. N=8 mice per condition. Error bars represent SEM. One-way ANOVA with Tukey post hoc test, * $p<0.05$, ** $p<0.01$.

A) Upper panel, hematoxylin and eosin staining of frozen heart sections of 5-months AAV-injected male mice expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG. Lower panel, corresponding immunohistochemistry against p62 with H&E counterstaining.

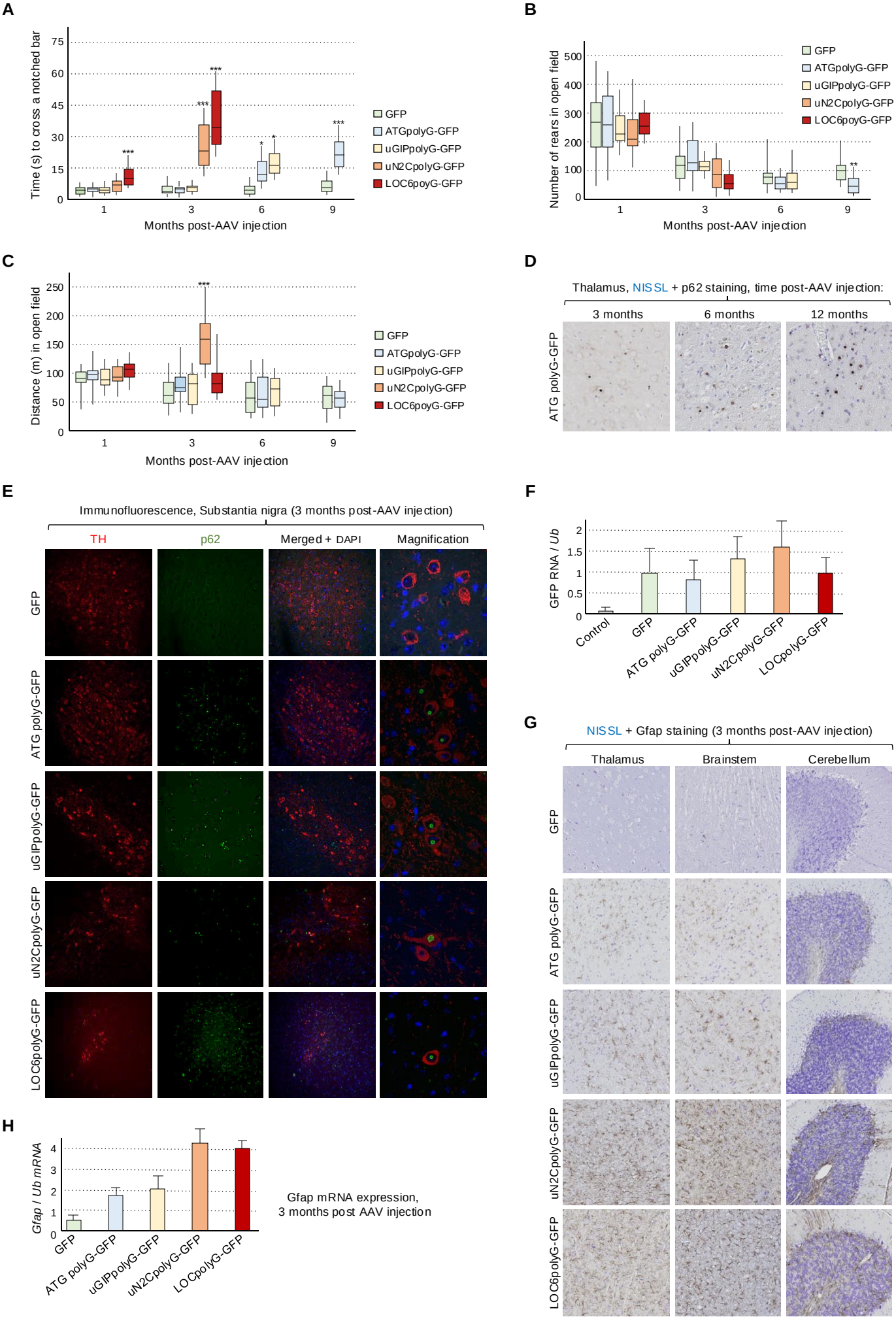


Figure S8. Related to Figure 8.

Figure S6. Related to Figure 6.
Expression of polyglycine proteins forms inclusions and is pathogenic in animal CNS.

(**A-C**) Time to cross a notched bar (A) and number of rears (B) or maximal distance traveled (C) during 30 min in an open field of male mice 1-, 3-, 6- and 9-months post-injection of AAV expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG or OPML LOC6polyG. N=10 AAV injected male mice. Error bars represent SEM. One-way ANOVA with Tukey post hoc test, * $p<0.05$, ** $p<0.01$, *** $p<0.01$.

(**D**) Immunohistochemistry against p62 with cresyl violet (Nissl) counterstaining of mouse brain thalamus area 3-, 6-or 12-months post-injection of AAV expressing ATG polyG-GFP.

(**E**) Immunofluorescence against p62 and tyrosine hydroxylase on the substancia nigra of 3-months post-injection of AAV expressing controls or OPDM/OPML GFP-tagged polyG proteins.

(**F**) RT-qPCR of GFP and Ubiquitin (*Ub*) expression on RNA extracted from the brain of 3-months AAV-injected male mice expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG or OPML LOC6polyG. N=3 mice per condition.

(**G**) Immunohistochemistry against Gfap with cresyl violet (Nissl) counterstaining of various mouse brain areas 3-months post-injection of AAV expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG or OPML LOC6polyG.

(**H**) RT-qPCR of Gfap and Ubiquitin (*Ub*) expression on RNA extracted from the brain of 3-months AAV-injected male mice expressing GFP, ATG polyG-GFP or GFP-tagged OPDM2 uGIPpolyG, OPDM3 uN2CpolyG or OPML LOC6polyG. N=3 mice per condition.

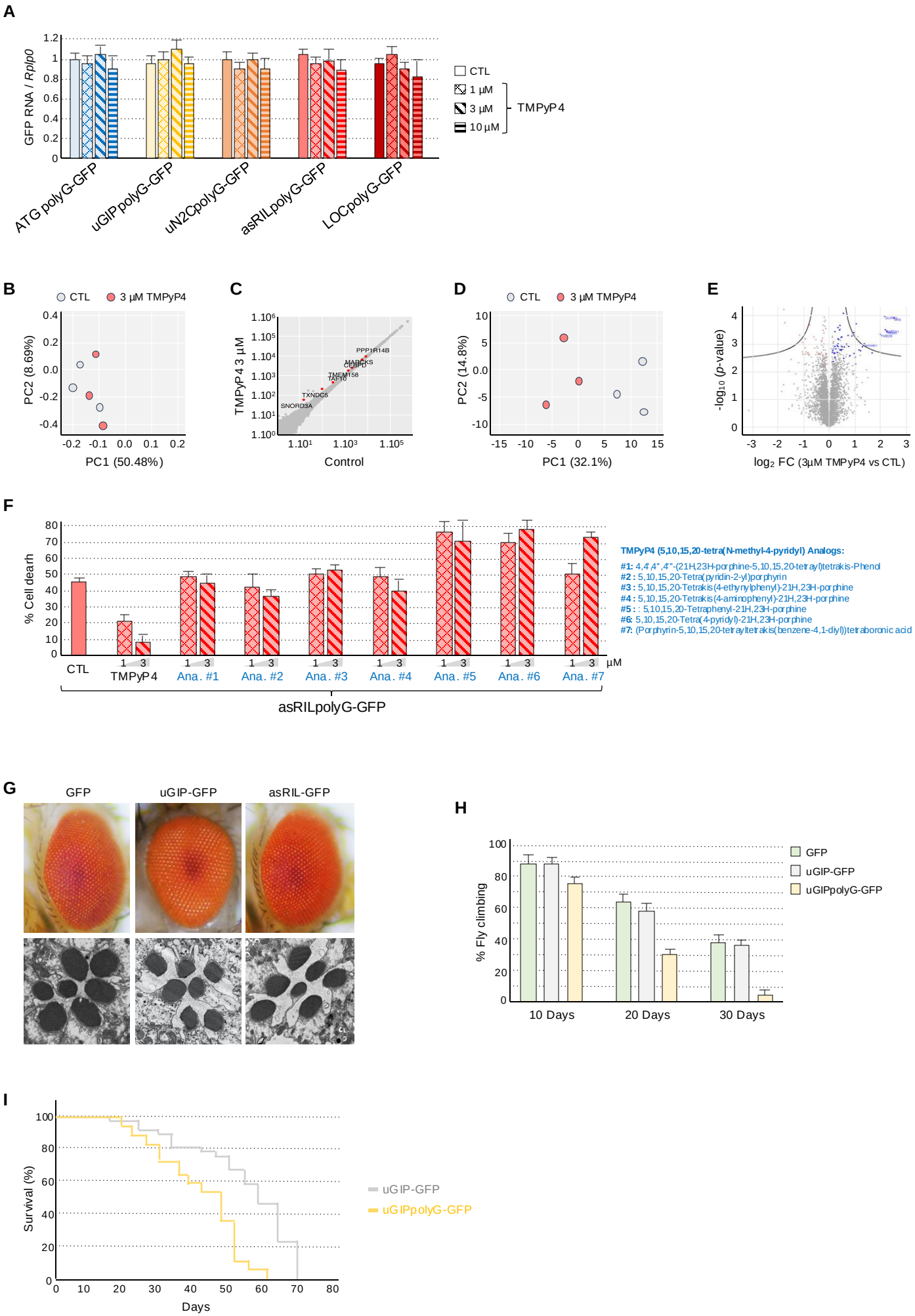


Figure S9. Related to Figure 9.

Figure S7. Related to Figure 7.
The porphyrin TMPYP4 alleviates aggregation and toxicity of polyG proteins.

(**A**) RT-qPCR of GFP and *Rplp0* expression on RNA extracted from 48 hours differentiated LHCN-M2 muscle cells expressing GFP-tagged ATG polyG, OPDM2 uGIPpolyG, OPDM3 uN2CpolyG, OPDM4 asRILpolyG or OPML LOC6polyG and treated overnight with the indicated drug concentration.

(**B-E**) Principal component analysis (B, D) scatter plot (C) and volcano plot (E) of RNA sequencing (B, C) or mass spectrometry (D, E) analysis of control versus overnight TMPyP4-treated LHCN-M2 muscle cells expressing the OPDM4 asRILpolyG.

(**F**) Cell viability of LHCN-M2 muscle cells differentiated for 3 days and expressing OPDM4 asRILpolyG treated overnight with 1 or 3 µM of various analogs of TMPyP4.

(**G**) Representative light microscopy (upper panel) and electron microscopy (lower panel) images of fly eyes expressing control GFP, uGIP-GFP or asRIL-GFP with a control (8x) length of glycine tract.

(**H**) Adult male flies expressing uGIPpolyG-GFP exhibited progressive locomotor deficits compared to GFP- or uGIP-GFP-expressing controls.

(**I**) Survival analysis showed a decreased median lifespan in uGIPpolyG-GFP-expressing male flies compared to uGIP-GFP-expressing controls. Genotypes: Actin5C-Gal4>UAS-GFP (control), Actin5C-Gal4>UAS-uGIP-GFP, Actin5C-Gal4>UAS-uGIPpolyG-GFP. Log-rank Mantel-Cox test; N = 124–139 flies per genotype. ****p<0.0001.