SUPPLEMENTARY MATERIALS

Title: Amino Termini of Many Yeast Proteins Map to Downstream Start Codons

Authors: Claire T. Fournier[1,3], Justin J. Cherny[1,3], Kris Truncali[1,2]†, Adam Robbins-Pianka[1,2]‡, Miin S. Lin[1], Danny Krizanc[2] and Michael P. Weir[1]*

Affiliations:

[1]Department of Biology, Wesleyan University Middletown CT 06459.

[2]Department of Mathematics and Computer Science, Wesleyan University Middletown CT 06459.

[3]Contributed equally.

*Correspondence to: mweir@wesleyan.edu.

†Current address: Boehringer Ingelheim Pharmaceuticals, 900 Ridgebury Road, Ridgefield CT 06877.

‡Current address: University of Colorado Boulder, Computer Science, Boulder, CO 80309.

Supplementary Materials

Supplementary Methods and Results

Supplementary Figures

Supplementary Tables

Supplementary Files

Supplementary Methods and Results

TAP-tagged Protein Purification

TAP-tag constructs (1) for selected downPeptide genes were expressed as described (2). 1 or 2 liters of cells were grown to early log phase and collected by centrifugation. Pelleted cells were resuspended in 5 ml of Hepes lysis buffer (100mM Hepes-KOH, pH 8-8.5, 20mM Mg(OAc)2, 10% glycerol, 10mM EGTA, 0.1mM EDTA, 0.4% NP-40, 100mM PMSF in ETOH, and 2 mini-tablet protease inhibitors (Roche)) and vortexed with glass beads for lysis. The lysate was then spun at 6,000 RPM to remove cell debris and the supernatant was collected and incubated with 3 mg (100 µl) of Dynabeads (Invitrogen) coupled to whole molecule rabbit IgG (Invitrogen) for 1 hr at 4˚C. Following a water rinse, the beads were collected and boiled with 50 ul sample buffer to elute the bound proteins. The sample was run on a 10% SDS-PAGE gel, fixed, and stained with colloidal coomassie (Invitrogen). The bands of interest were excised and digested in-gel according to Shevchenko et al. (3). The peptides were resuspended in 0.1% TFA and loaded onto a C18 resin packed column for MS/MS analysis. The enhanced protein coverage of partially purified TAP-tagged proteins allowed confirmation of downPeptide expression (see Results and Discussion)

Translation Relative Individual Information (TRII score) measurement

The sequences flanking the annAUGs of downPeptide and annPeptide mRNAs were compared using a positional weight matrix computed from high-confidence translation initiation sites (4, 5). The weight matrix was used to compute Translation Relative Individual Information (TRII) scores which indicate how well individual sequences conform to the high-confidence sites.

Ribosome profiles

Ribosome profiles (*6*) for aligned dnAUGs showed tag density patterns consistent with translation initiation at dnAUGs. The ribosome profiling results of Ingolia et al. (*6*) are striking in that the profiles of sequence tags protected by ribosomes outline the expressed open reading frames of genes with nucleotide resolution. For example, when tag profiles were aligned to the annAUGs of annPeptide genes detected in this study, high densities of tags were observed centered at positions 2, 3 and 15 (arrowheads, Fig. 4A) with a characteristic depression surrounding position 15 (arrow) mirroring the previously published profiles of aligned translation start sites (*6*). Profile alignment relative to dnAUGs of downPeptide genes revealed ribosome profiles with elevated densities at positions 2 and 15, and depressed densities surrounding 15. This profile was more pronounced for a subset of downPeptide genes with lower quality sequence context surrounding their annAUGs (Fig. 4B; Translation Relative

Individual Information (TRII) score less than 8; (*4, 5*)), suggesting that genes with poor context surrounding their annAUG are more likely to have translation initiation at a dnAUG.

In light of these observations, we examined ribosome profiles of aligned rank 1 dnAUGs for all genes with an annAUG sequence context TRII score < 8 (1267 genes). Only genes with ribosome tag densities above 0.1 tags/nt in the first 200 nt of their annORF were analyzed. The average ribosome profile for these genes (where each gene received equal weight; Fig. 4C) shows elevated ribosome tag densities at positions 2 and 15, suggesting that translation initiation occurs at dnAUGs in many of these genes. In contrast, genes with higher quality sequence context at their annAUG (TRII $\geq$ 10; 1372 genes; Fig. 4D) show less pronounced signals at positions 2 and 15, suggesting lower levels of translation initiation at their dnAUGs. These results, which are independent of our mass spectrometry analysis, also suggest that dnAUG translation initiation is common in yeast, particularly for genes with poorer context annAUGs.

Compared to profiles of multiple gene alignments, the ribosome profiles of individual downPeptide genes tend to have low signal and be noisy; nevertheless, individual gene profiles for several high-expression genes identified in our mass spectroscopy analysis had patterns consistent with leaky scanning (Supplementary Fig. S6A) in which the ribosome density showed a significant increase at the dnAUG. Other individual genes showed background tag densities 5' of the dnAUG (Supplementary Fig. S6B), suggesting that the principal initiation is at the dnAUG.

Protein sequence conservation

Under-utilization of the annAUGs in downPeptide genes might be expected to be associated with poorer conservation across species of the annAUG region. Indeed, assessment of conservation in four species of *Saccharomyces* has been used to refine the annotations of the N-termini of proteins (*7*). Orthologs have been identified (*8*) based on an alignment algorithm that takes into account chromosome synteny as well as conservation of individual gene sequences. Conservation of protein sequences was assessed for the subset of *Saccharomyces cerevisiae* downPeptide genes in which the annAUG and dnAUG are in the same reading frame. The protein sequences of these downPeptide genes were assessed in *Saccharomyces cerevisiae* and the three other yeast species. Following alignment with ClustalW (*9*), protein conservation at the annAUG and dnAUG were scored using widows of 12 amino acids in width (see Supplementary Fig. S3 legend). DownPeptide genes with lower TRII score annAUGs (TRII < 8) have poorer conservation at the annAUG window compared to the dnAUG window (Supplementary Fig. S3), consistent with the interpretation that some of these genes have little or no selection in the annAUG region – presumably because this region is not translated (reflecting possible misannotation), or codes for protein of limited functional importance.

To investigate this further, codon bias was assessed in the vicinity of the implicated dnAUGs of the frame 1 downPeptide genes. The Codon Adaptation Index (CAI (*10*)) provides a quantification of codon bias; the most common codon for each amino acid is assigned a CAI score of 100% and the less common codons are assigned proportionally lower scores. More highly expressed proteins tend to utilize codons with higher CAI index. Our assessment of CAI scores indicated that codons immediately downstream of the implicated dnAUG in downPeptide genes tend to have higher mean CAI scores than codons between the annAUG and the dnAUG (Supplementary Fig. S4B). Bootstrap analysis (Supplementary Fig. S4C) indicated that this increase in CAI score is unlikely to be random (p = 0.05 for 1% false identification set). The increase compares in magnitude with that observed at the annAUGs of annPeptide genes (Supplementary Fig. S4A), and is considerably higher than the increase at control dnAUGs (not shown). This increase in codon bias at dnAUGs is consistent with translation initiation at these sites, and reduced (or no) initiation at annAUGs.

Protein motif prediction and Gene Ontology (GO)

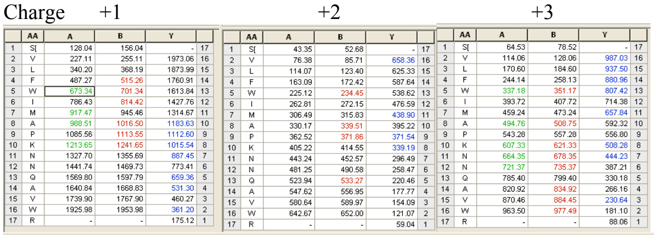
Translation initiation at dnAUGs may be important in orchestrating the cellular functions of protein products. Proteins resulting from initiation at a dnAUG in the same reading frame as the annAUG may result in N-terminal truncated proteins with different targeting signals thereby allowing expression of the protein in alternative cellular compartments (*11, 12*). Indeed, for many downPeptide genes (25), ER signal sequences are predicted by the SignalP algorithm to be located between the annAUG and dnAUG (Supplementary Table S7). In some cases, signal peptide cleavage sites are predicted, whereas in other cases, there is a potential signal sequence, but no cleavage site is predicted, suggesting that the longer protein product could become anchored to the ER membrane, whereas the shorter downPeptide product lacking a signal sequence could instead enter the cytoplasm. Cases of sequences associated with enzyme activities are predicted by the InterProScan algorithm between the annAUG and dnAUG (Supplementary Table S8). The amino acid composition coded between the annAUG and dnAUG of downPeptide genes had a greater tendency towards predicted disorder than the equivalent regions of randomly selected genes (Supplementary Fig. S6), which in some cases could facilitate association with regulatory cofactors (*13*).

We assessed the frequencies with which Gene Ontology (GO) terms (*14*) are associated with downPeptide and annPeptide genes. The frequencies of some GO terms were significantly different for the two gene sets (Supplementary Table S9). However, similar trends were observed when comparing GO term frequencies for high protein expression genes with frequencies for all genes, consistent with our observation that annPeptide genes show a trend towards higher protein expression than downPeptide genes.

Supplementary Figures

Suppl. Fig. S1

Charge +1 +2 +3

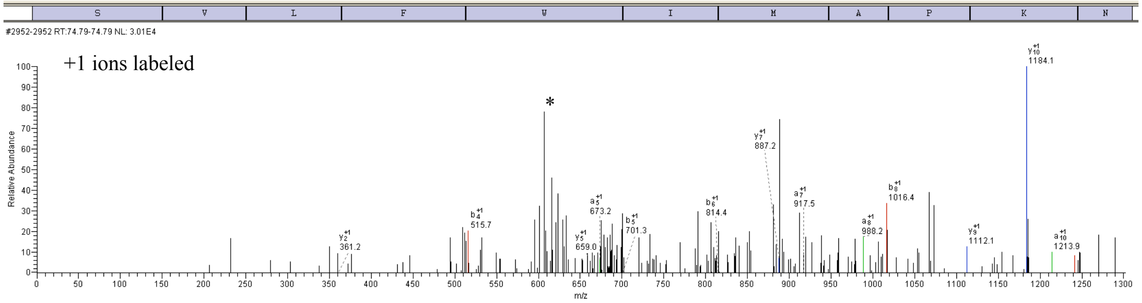


	AA	A	B	Y	
1	S[	128.04	156.04	-	17
2	V	227.11	255.11	1973.06	16
3	L	340.20	368.19	1873.99	15
4	F	487.27	515.26	1760.91	14
5	W	673.34	701.34	1613.84	13
6	I	786.43	814.42	1427.76	12
7	M	917.47	945.46	1314.67	11
8	A	988.51	1016.50	1183.63	10
9	P	1085.56	1113.55	1112.60	9
10	K	1213.65	1241.65	1015.54	8
11	N	1327.70	1355.69	887.45	7
12	N	1441.74	1469.73	773.41	6
13	Q	1569.80	1597.79	659.36	5
14	A	1640.84	1668.83	531.30	4
15	V	1739.90	1767.90	460.27	3
16	W	1925.98	1953.98	361.20	2
17	R	-	-	175.12	1

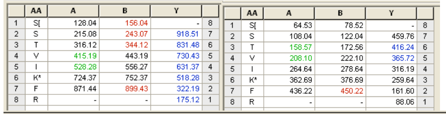
	AA	A	B	Y	
1	S[	43.35	52.68	-	17
2	V	76.38	85.71	658.36	16
3	L	114.07	123.40	625.33	15
4	F	163.09	172.42	587.64	14
5	W	225.12	234.45	538.62	13
6	I	262.81	272.15	476.59	12
7	M	306.49	315.83	438.90	11
8	A	330.17	339.51	395.22	10
9	P	362.52	371.86	371.54	9
10	K	405.22	414.55	339.19	8
11	N	443.24	452.57	296.49	7
12	N	481.25	490.58	258.47	6
13	Q	523.94	533.27	220.46	5
14	A	547.62	556.95	177.77	4
15	V	580.64	589.97	154.09	3
16	W	642.67	652.00	121.07	2
17	R	-	-	59.04	1

	AA	A	B	Y	
1	S[	64.53	78.52	-	17
2	V	114.06	128.06	987.03	16
3	L	170.60	184.60	937.50	15
4	F	244.14	258.13	880.96	14
5	W	337.18	351.17	807.42	13
6	I	393.72	407.72	714.38	12
7	M	459.24	473.24	657.84	11
8	A	494.76	508.75	592.32	10
9	P	543.28	557.28	556.80	9
10	K	607.33	621.33	508.28	8
11	N	664.35	678.35	444.23	7
12	N	721.37	735.37	387.21	6
13	Q	785.40	799.40	330.18	5
14	A	820.92	834.92	266.16	4
15	V	870.46	884.45	230.64	3
16	W	963.50	977.49	181.10	2
17	R	-	-	88.06	1

VMA9
id_28508_1049_to_1220_frame_1_trimmed_1
ions 30/96



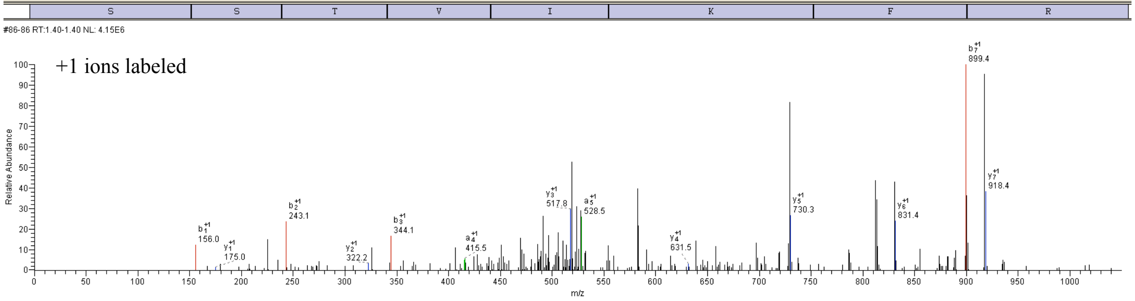
Charge +1 +2



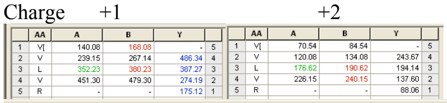
	AA	A	B	Y	
1	S[	128.04	156.04	-	8
2	S	215.08	243.07	918.51	7
3	T	316.12	344.12	831.48	6
4	V	415.19	443.19	730.43	5
5	I	528.28	556.27	631.37	4
6	K*	724.37	752.37	518.28	3
7	F	871.44	899.43	322.19	2
8	R	-	-	175.12	1

	AA	A	B	Y	
1	S[	64.53	78.52	-	8
2	S	108.04	122.04	459.76	7
3	T	158.57	172.56	416.24	6
4	V	208.10	222.10	365.72	5
5	I	264.64	278.64	316.19	4
6	K*	362.69	376.69	259.64	3
7	F	436.22	450.22	161.60	2
8	R	-	-	88.06	1

RPN13
id_4413_1007_to_1469_frame_1_trimmed_1
ions 13/21



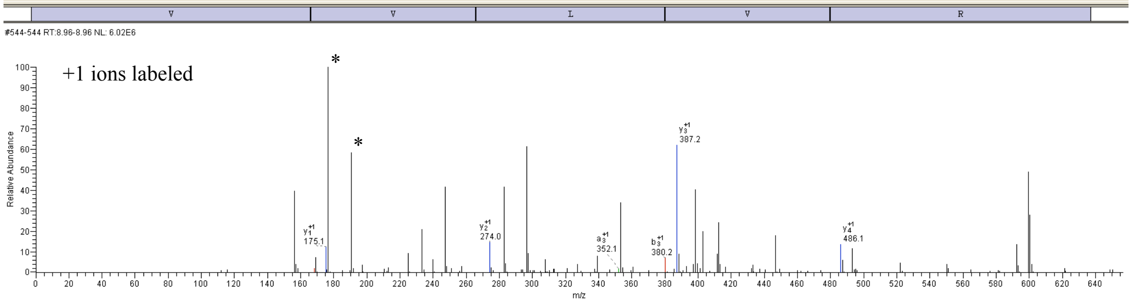
Charge +1 +2



	AA	A	B	Y	
1	V[	140.08	168.08	-	5
2	V	239.15	267.14	486.34	4
3	L	352.23	380.23	387.27	3
4	V	451.30	479.30	274.19	2
5	R	-	-	175.12	1

	AA	A	B	Y	
1	V[	70.54	84.54	-	5
2	V	120.08	134.08	243.67	4
3	L	176.62	190.62	194.14	3
4	V	226.15	240.15	137.60	2
5	R	-	-	88.06	1

GSP1
id_4284_1059_to_1098_frame_2_trimmed_1
ions 9/14



Suppl. Fig. S1 (cont.)

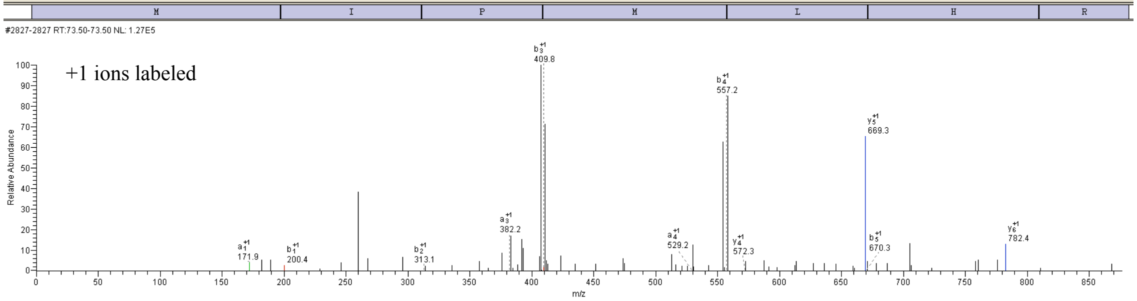
ZRG17
id_5322_1080_to_1116_frame_2
ions 11/14

Charge +1

	AA	A	B	Y	
1	M[	172.05	200.05	-	7
2	I	285.14	313.13	782.43	6
3	P	382.19	410.18	669.35	5
4	M@	529.23	557.22	572.30	4
5	L	642.31	670.30	425.26	3
6	H	779.37	807.36	312.18	2
7	R	-	-	175.12	1

+2

	AA	A	B	Y	
1	M[	86.53	100.53	-	7
2	I	143.07	157.07	391.72	6
3	P	191.60	205.60	335.18	5
4	M@	265.12	279.11	286.65	4
5	L	321.66	335.66	213.13	3
6	H	390.19	404.19	156.59	2
7	R	-	-	88.06	1



YAL026C-A
id_28730_1064_to_1436_frame_1_trimmed_1
ions 18/60

Charge +1

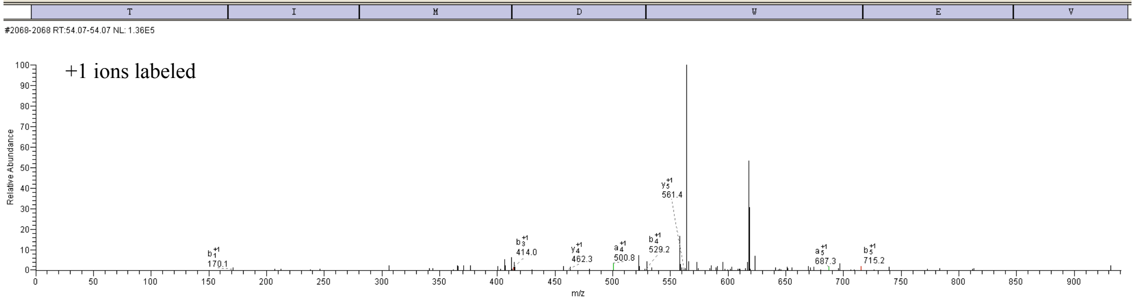
	AA	A	B	Y	
1	T[	142.06	170.05	-	11
2	I	255.14	283.14	1235.63	10
3	M	386.18	414.18	1122.55	9
4	D	501.21	529.21	991.51	8
5	W	687.29	715.29	876.48	7
6	E	816.33	844.33	690.40	6
7	V	915.40	943.40	561.36	5
8	T	1016.45	1044.44	462.29	4
9	T	1117.50	1145.49	361.24	3
10	L	1230.58	1258.58	260.20	2
11	K	-	-	147.11	1

+2

	AA	A	B	Y	
1	T[	71.53	85.53	-	11
2	I	128.08	142.07	618.32	10
3	M	193.60	207.59	561.78	9
4	D	251.11	265.11	496.26	8
5	W	344.15	358.15	438.74	7
6	E	408.67	422.67	345.71	6
7	V	458.20	472.20	281.18	5
8	T	508.73	522.73	231.65	4
9	T	559.25	573.25	181.13	3
10	L	615.79	629.79	130.60	2
11	K	-	-	74.06	1

+3

	AA	A	B	Y	
1	T[	48.02	57.36	-	11
2	I	85.72	95.05	412.55	10
3	M	129.40	138.73	374.85	9
4	D	167.74	177.07	331.17	8
5	W	229.77	239.10	292.83	7
6	E	272.78	282.11	230.81	6
7	V	305.81	315.14	187.79	5
8	T	339.49	348.82	154.77	4
9	T	373.17	382.50	121.09	3
10	L	410.87	420.20	87.40	2
11	K	-	-	49.71	1



AEP1
id_4668_1030_to_1051_frame_3
ions 13/36

Charge +1

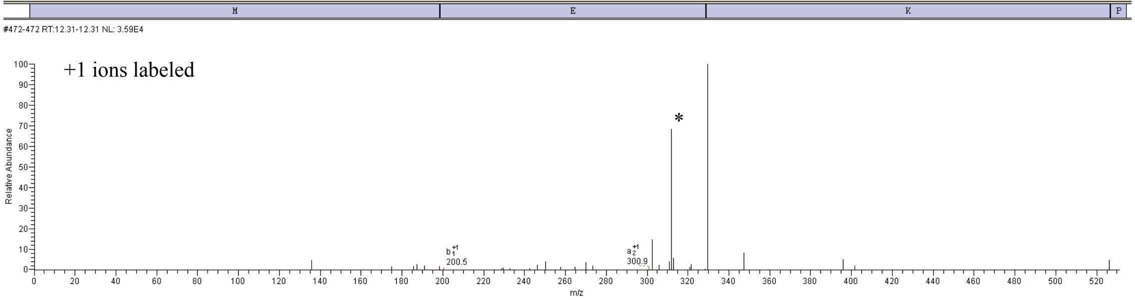
	AA	A	B	Y	
1	M[	172.05	200.05	-	7
2	E	301.10	329.09	790.42	6
3	K*	497.19	525.19	661.38	5
4	P	594.24	622.24	465.28	4
5	V	693.31	721.31	368.23	3
6	L	806.40	834.39	269.16	2
7	H	-	-	156.08	1

+2

	AA	A	B	Y	
1	M[	86.53	100.53	-	7
2	E	151.05	165.05	395.71	6
3	K*	249.10	263.10	331.19	5
4	P	297.63	311.62	233.14	4
5	V	347.16	361.16	184.62	3
6	L	403.70	417.70	135.08	2
7	H	-	-	78.54	1

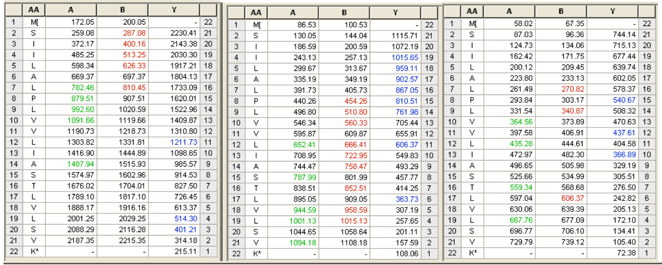
+3

	AA	A	B	Y	
1	M[	58.02	67.35	-	7
2	E	101.04	110.37	264.14	6
3	K*	166.40	175.73	221.13	5
4	P	198.75	208.08	155.77	4
5	V	231.78	241.11	123.41	3
6	L	269.47	278.80	90.39	2
7	H	-	-	52.70	1



Suppl. Fig. S1 (cont.)

Charge +1 +2 +3



+1

	AA	A	B	Y	
1	M[	172.05	200.05	-	22
2	S	259.08	287.08	2230.41	21
3	I	372.17	400.16	2143.38	20
4	I	485.25	513.25	2030.30	19
5	L	598.34	626.33	1917.21	18
6	A	669.37	697.37	1804.13	17
7	L	782.46	810.45	1733.09	16
8	P	879.51	907.51	1620.01	15
9	L	992.60	1020.59	1522.96	14
10	V	1091.66	1119.66	1409.87	13
11	V	1190.73	1218.73	1310.80	12
12	L	1303.82	1331.81	1211.73	11
13	I	1416.90	1444.89	1098.65	10
14	A	1487.94	1515.93	985.57	9
15	S	1574.97	1602.96	914.53	8
16	T	1676.02	1704.01	827.50	7
17	L	1789.10	1817.10	726.45	6
18	V	1888.17	1916.16	613.37	5
19	L	2001.25	2029.25	514.30	4
20	S	2088.29	2116.28	401.21	3
21	V	2187.35	2215.35	314.18	2
22	K*	-	-	215.11	1

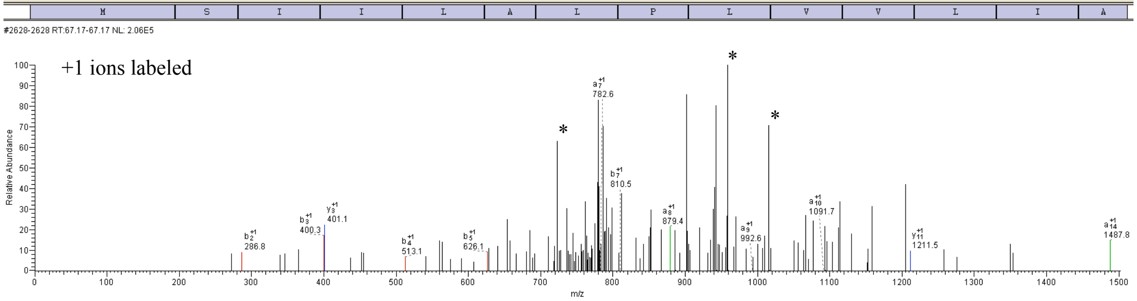
+2

	AA	A	B	Y	
1	M[	86.53	100.53	-	22
2	S	130.05	144.04	1115.71	21
3	I	186.59	200.59	1072.19	20
4	I	243.13	257.13	1015.65	19
5	L	299.67	313.67	959.11	18
6	A	335.19	349.19	902.57	17
7	L	391.73	405.73	867.05	16
8	P	440.26	454.26	810.51	15
9	L	496.80	510.80	761.98	14
10	V	546.34	560.33	705.44	13
11	V	595.87	609.87	655.91	12
12	L	652.41	666.41	606.37	11
13	I	708.95	722.95	549.83	10
14	A	744.47	758.47	493.29	9
15	S	787.99	801.99	457.77	8
16	T	838.51	852.51	414.25	7
17	L	895.05	909.05	363.73	6
18	V	944.59	958.59	307.19	5
19	L	1001.13	1015.13	257.65	4
20	S	1044.65	1058.64	201.11	3
21	V	1094.18	1108.18	157.59	2
22	K*	-	-	108.06	1

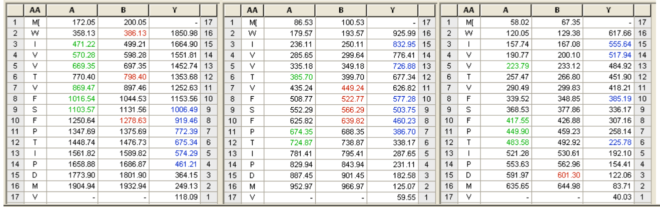
+3

	AA	A	B	Y	
1	M[	58.02	67.35	-	22
2	S	87.03	96.36	744.14	21
3	I	124.73	134.06	715.13	20
4	I	162.42	171.75	677.44	19
5	L	200.12	209.45	639.74	18
6	A	223.80	233.13	602.05	17
7	L	261.49	270.82	578.37	16
8	P	293.84	303.17	540.67	15
9	L	331.54	340.87	508.32	14
10	V	364.56	373.89	470.63	13
11	V	397.58	406.91	437.61	12
12	L	435.28	444.61	404.58	11
13	I	472.97	482.30	366.89	10
14	A	496.65	505.98	329.19	9
15	S	525.66	534.99	305.51	8
16	T	559.34	568.68	276.50	7
17	L	597.04	606.37	242.82	6
18	V	630.06	639.39	205.13	5
19	L	667.76	677.09	172.10	4
20	S	696.77	706.10	134.41	3
21	V	729.79	739.12	105.40	2
22	K*	-	-	72.38	1

IRC9
id_3678_1073_to_1391_frame_1
ions 30/126



Charge +1 +2 +3



+1

	AA	A	B	Y	
1	M[	172.05	200.05	-	17
2	W	358.13	386.13	1850.98	16
3	I	471.22	499.21	1664.90	15
4	V	570.28	598.28	1551.81	14
5	V	669.35	697.35	1452.74	13
6	T	770.40	798.40	1353.68	12
7	V	869.47	897.46	1252.63	11
8	F	1016.54	1044.53	1153.56	10
9	S	1103.57	1131.56	1006.49	9
10	F	1250.64	1278.63	919.46	8
11	P	1347.69	1375.69	772.39	7
12	T	1448.74	1476.73	675.34	6
13	I	1561.82	1589.82	574.29	5
14	P	1658.88	1686.87	461.21	4
15	D	1773.90	1801.90	364.15	3
16	M	1904.94	1932.94	249.13	2
17	V	-	-	118.09	1

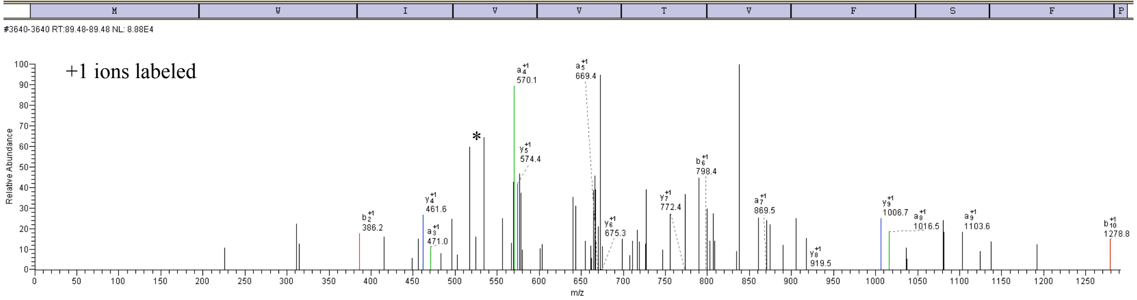
+2

	AA	A	B	Y	
1	M[	86.53	100.53	-	17
2	W	179.57	193.57	925.99	16
3	I	236.11	250.11	832.95	15
4	V	285.65	299.64	776.41	14
5	V	335.18	349.18	726.88	13
6	T	385.70	399.70	677.34	12
7	V	435.24	449.24	626.82	11
8	F	508.77	522.77	577.28	10
9	S	552.29	566.29	503.75	9
10	F	625.82	639.82	460.23	8
11	P	674.35	688.35	386.70	7
12	T	724.87	738.87	338.17	6
13	I	781.41	795.41	287.65	5
14	P	829.94	843.94	231.11	4
15	D	887.45	901.45	182.58	3
16	M	952.97	966.97	125.07	2
17	V	-	-	59.55	1

+3

	AA	A	B	Y	
1	M[	58.02	67.35	-	17
2	W	120.05	129.38	617.66	16
3	I	157.74	167.08	555.64	15
4	V	190.77	200.10	517.94	14
5	V	223.79	233.12	484.92	13
6	T	257.47	266.80	451.90	12
7	V	290.49	299.83	418.21	11
8	F	339.52	348.85	385.19	10
9	S	368.53	377.86	336.17	9
10	F	417.55	426.88	307.16	8
11	P	449.90	459.23	258.14	7
12	T	483.58	492.92	225.78	6
13	I	521.28	530.61	192.10	5
14	P	553.63	562.96	154.41	4
15	D	591.97	601.30	122.06	3
16	M	635.65	644.98	83.71	2
17	V	-	-	40.03	1

BUD16
id_755_1092_to_1143_frame_2
ions 27/96

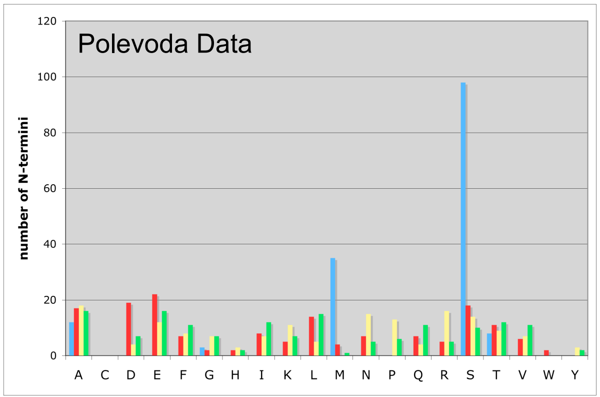


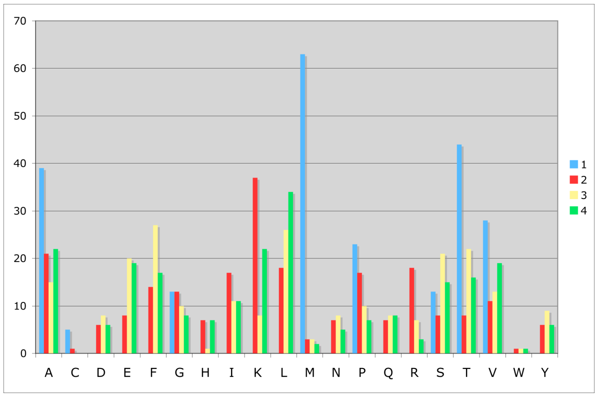
Supplementary Fig. S1.

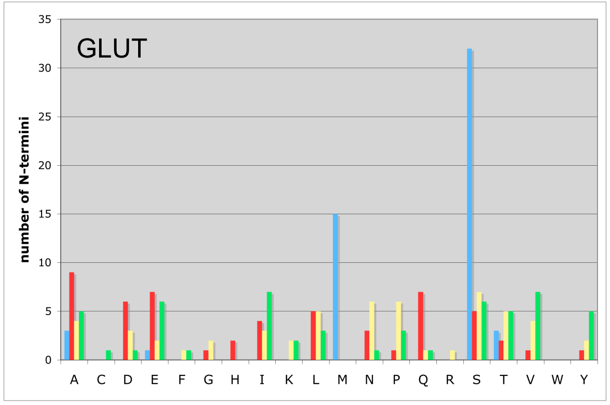
Sequest-display examples of spectra for matches to downPeptides from our glutaraldehyde-treatment experiments. Matched +1 ion peaks are labeled with designations (a, b and y ions labeled with green, red and blue respectively); matched +2 and +3 ion peaks with relative abundance >50% are labeled with *. The trypsin peptide sequences are illustrated in the purple boxes. Additional details of the matches are in supplementary file Fournier_SuppData.xls.

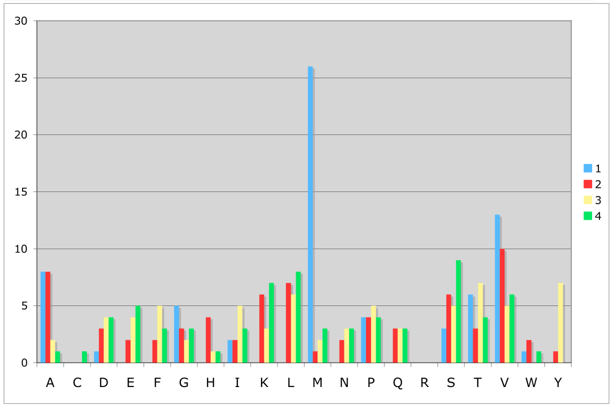
Acetylated

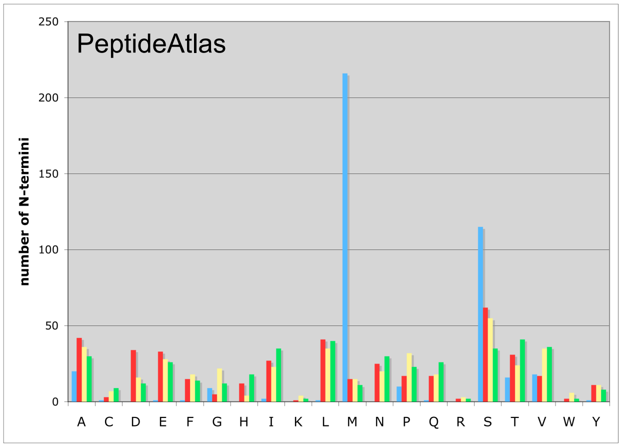
Not Acetylated

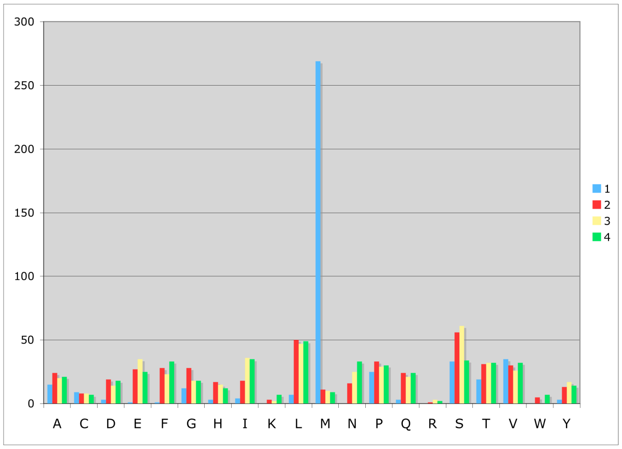


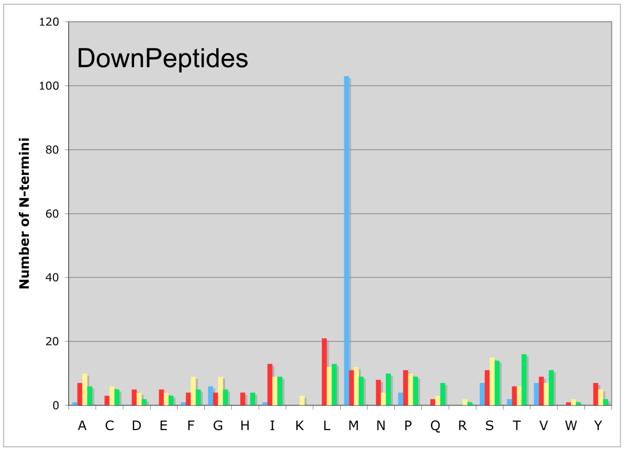


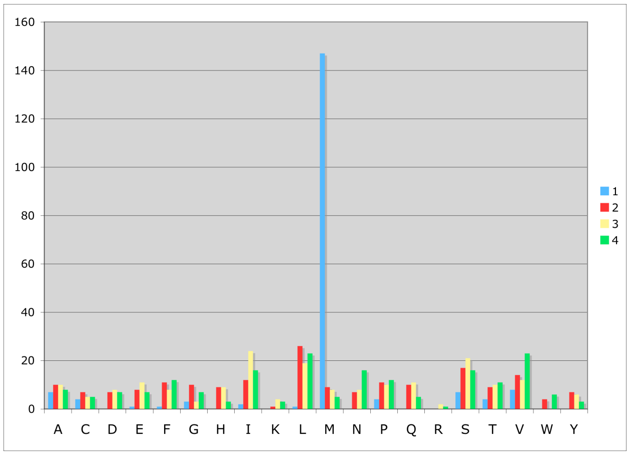




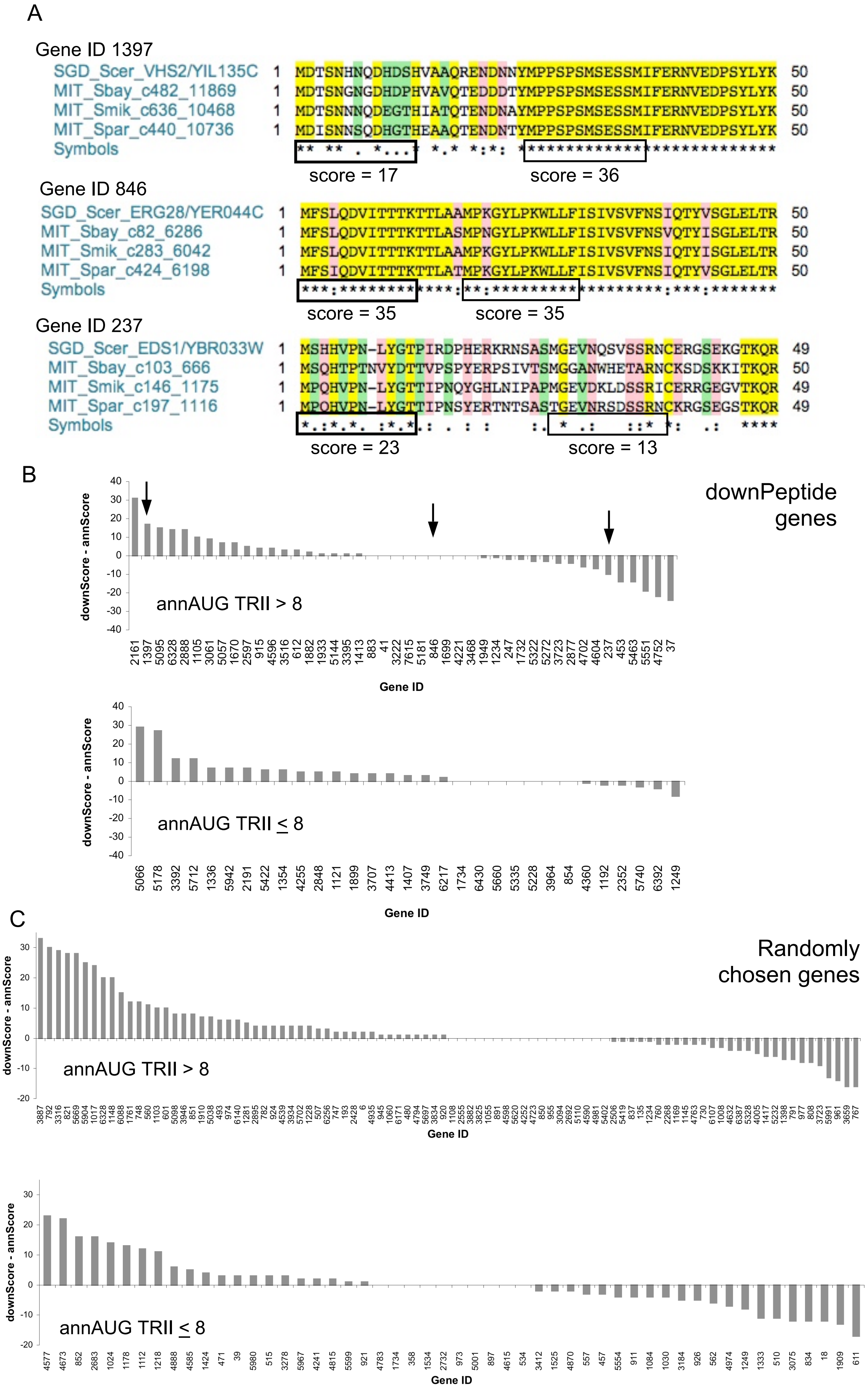






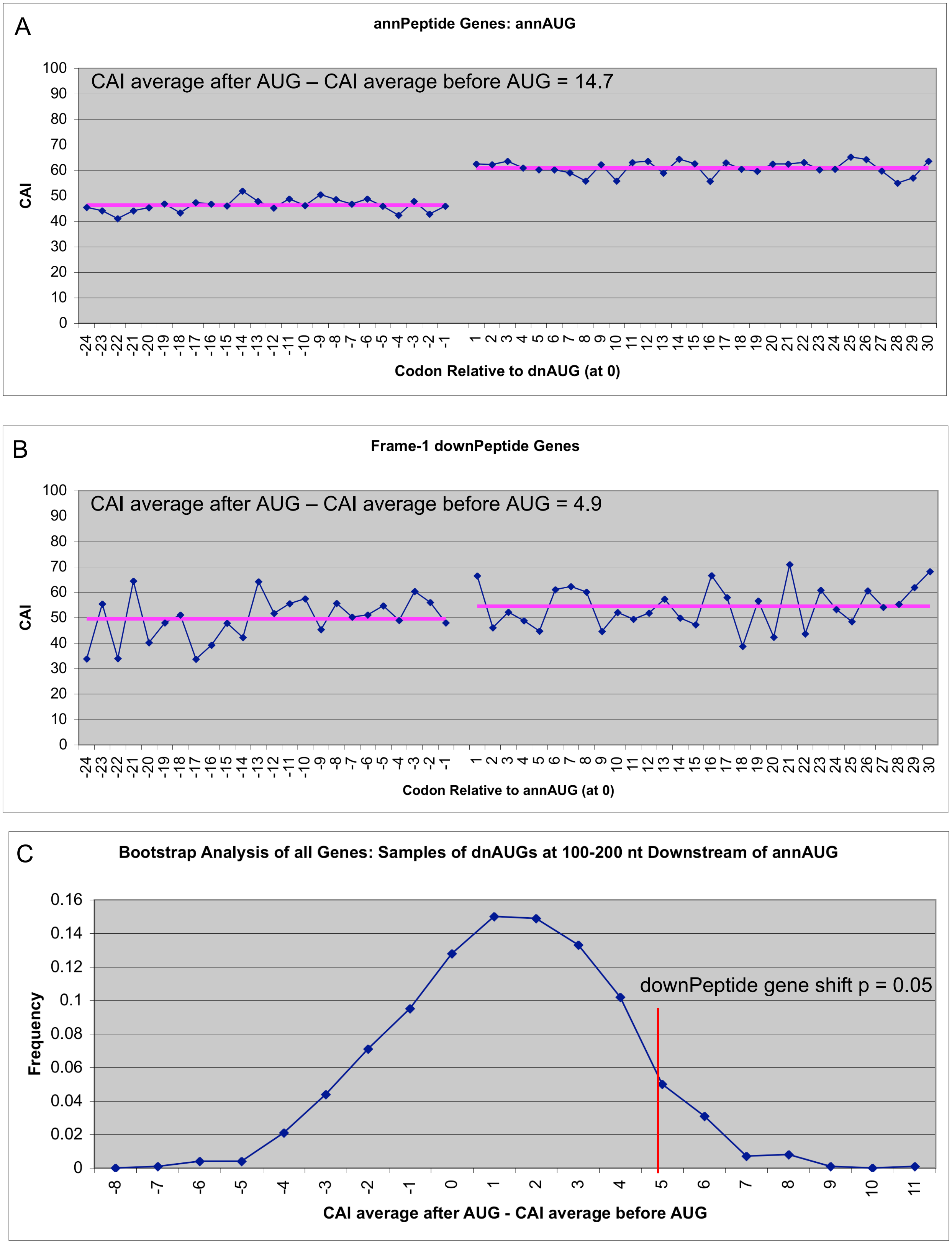


Supplementary Fig. S2. Amino acids at positions 1 – 4 of amino peptides. Data from Polevoda and Sherman (*15*) is compared with amino termini detected in our glutaraldehyde treated cell lystates (GLUT), our screen of Peptide Atlas data, and the amino termini of the proteins initiating at dnAUGs of the downPeptide genes identified in this study.


A
Gene ID 1397
SGD_Scer_VHS2/YIL135C 1 MDTSNHNQDHDSHVAAQRENDNNYMPPSPSMSESSMIFERNVEDPSYLYK 50
MIT_Sbay_c482_11869 1 MDTSNGNGDHDPHVAVQTEDDDTYMPPSPSMSESSMIFERNVEDPSYLYK 50
MIT_Smik_c636_10468 1 MDTSNNNQDEGTHIATQTENDNAYMPPSPSMSESSMIFERNVEDPSYLYK 50
MIT_Spar_c440_10736 1 MDISNNSQDHGTHEAAQTENDNTYMPPSPSMSESSMIFERNVEDPSYLYK 50
Symbols ** ** . *...* *.* *:*: **************************
score = 17
score = 36
Gene ID 846
SGD_Scer_ERG28/YER044C 1 MFSLQDVITTTKTTLAAMPKGYLPKWLLFISIVSVFNSIQTYVSGLELTR 50
MIT_Sbay_c82_6286 1 MFSLQDVITTTKTTLASMPNGYLPKWLLFISIVSVFNSVQTYISGLELTR 50
MIT_Smik_c283_6042 1 MFSLQDVITTTKTTLAAMPKGYLPKWLLFISIVSVFNSIQTYISGLELTR 50
MIT_Spar_c424_6198 1 MFSIQDVITTTKTTLATMPKGYLPKWLLFISIVSVFNSIQTYVSGLELTR 50
Symbols ***:*************:**:*******************:***:*******
score = 35
score = 35
Gene ID 237
SGD_Scer_EDS1/YBR033W 1 MSHHVPN-LYGTPIRDPHERKRNSASMGEVNQSVSSRNCERGSEKGTKQR 49
MIT_Sbay_c103_666 1 MSQHTPTNVYDTTVPSPYERPSIVTSMGGANWHETARNCKSDSKKITKQR 50
MIT_Smik_c146_1175 1 MPQHVPN-LYGTTIPNQYGHLNIPAPMGEVDKLDSSRICERRGEGVTKQR 49
MIT_Spar_c197_1116 1 MPQHVPN-LYGTTIPNSYERTNTSASTGEVNRSDSSRNCKRGSEGSTKQR 49
Symbols *.:*.*. :*.*.: . : :   :. * .:   ::* *:   .:   ****
score = 23
score = 13
B
downPeptide genes
downScore - annScore
40
30
20
10
0
-10
-20
-30
-40
annAUG TRII > 8
2161 1397 5095 6328 2888 1105 3061 5057 1670 2597 915 4596 3516 612 1882 1933 5144 3395 1413 883 41 3222 7615 5181 846 1699 4221 3468 1949 1234 247 1732 5322 5272 3723 2877 4702 4604 237 453 5463 5551 4752 37
Gene ID
downScore - annScore
40
30
20
10
0
-10
-20
-30
-40
annAUG TRII < 8
5066 5178 3392 5712 1336 5942 2191 5422 1354 4255 2848 1121 1899 3707 4413 1407 3749 6217 1734 6430 5660 5335 5228 3964 854 4360 1192 2352 5740 6392 1249
Gene ID
C
Randomly chosen genes
downScore - annScore
30
20
10
0
-10
-20
annAUG TRII > 8
3887 792 3316 821 5669 5904 1017 6328 1148 6088 1761 748 560 1103 601 5098 3946 851 1910 5038 493 974 6140 1281 2895 782 924 4539 3934 5702 1228 507 6256 747 193 2428 6 4935 945 1060 6171 480 4794 5697 3834 920 1108 2555 3882 3825 1055 891 4598 5620 4252 4723 850 955 3094 2692 5110 4590 4981 5402 2506 5419 837 135 1234 760 2268 1169 1145 4763 730 6107 1008 4632 6387 5328 4005 1417 5232 1398 791 977 808 3723 5991 961 3659 767
Gene ID
downScore - annScore
30
20
10
0
-10
-20
annAUG TRII < 8
4577 4673 852 2683 1024 1178 1112 1218 4888 4585 1424 471 39 5980 515 3278 5967 4241 4815 5599 921 4783 1734 358 1534 2732 973 5001 897 4615 534 3412 1525 4870 557 457 5554 911 1084 1030 3184 926 562 4974 1249 1333 510 3075 834 18 1909 611
Gene ID

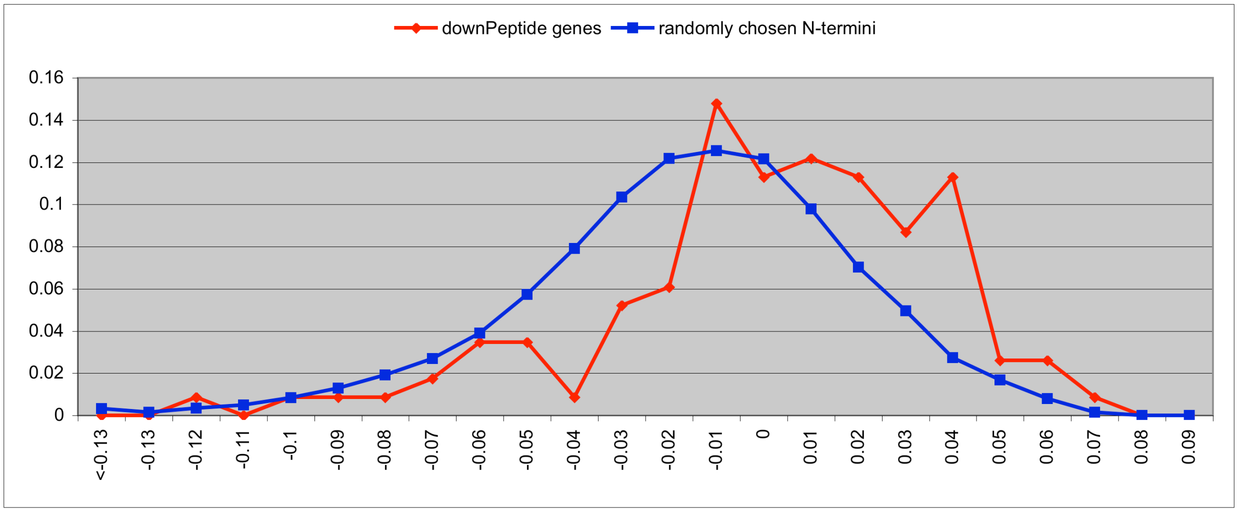
Supplementary Fig. S3.

Sequence conservation. (**A**) DownPeptide gene protein sequences were aligned with homologs from three other *Saccharomyces* species (http://www.yeastgenome.org). Amino acid sequence conservation was compared in the 12 amino acid windows starting at the annAUG and dnAUG (boxed). *VYS2* shows greater conservation at the dnAUG window. *ERG28* has similar conservation at the annAUG and dnAUG windows, whereas *EDS1* has stronger conservation in the annAUG window. (**B**) 75 of the 135 downPeptide genes with frame 1 dnAUGs have homologs reported in the three species and were analyzed here. Unlike controls (see C), downPeptide genes with poor sequence context annAUGs (TRII < 8) have greater conservation at their dnAUG window compared to the annAUG ($p < 0.036$, Mann-Whitney U test). Conservation of each amino acid in the 12 amino acid window was scored as identical (3 points = yellow *), strong similarity (2 points = pink :), or weak similarity (1 point = green .) The sum of scores for the annAUG window was subtracted from the sum for the dnAUG window. The resulting difference scores (DownScore – annScore) are graphed. *VYS2*, *ERG28* and *EDS1* (gene IDs [score] 1397 [9], 846 [0] and 237 [-10], respectively) are marked with arrows. (**c**) Amino acid conservation in randomly selected downORFs. Genes with rank 1, frame 1 dnAUGs located 60 - 100 nt downstream of the annAUG were selected randomly. The genes were divided into two groups based on their annAUG TRII score (≤ 8 and > 8). Amino acid sequence conservation was compared in the 12 amino acid windows starting at the annAUG and dnAUG. Unlike downPeptide genes, the annAUG TRII ≤ 8 set did not have elevated conservation at the downAUG region.

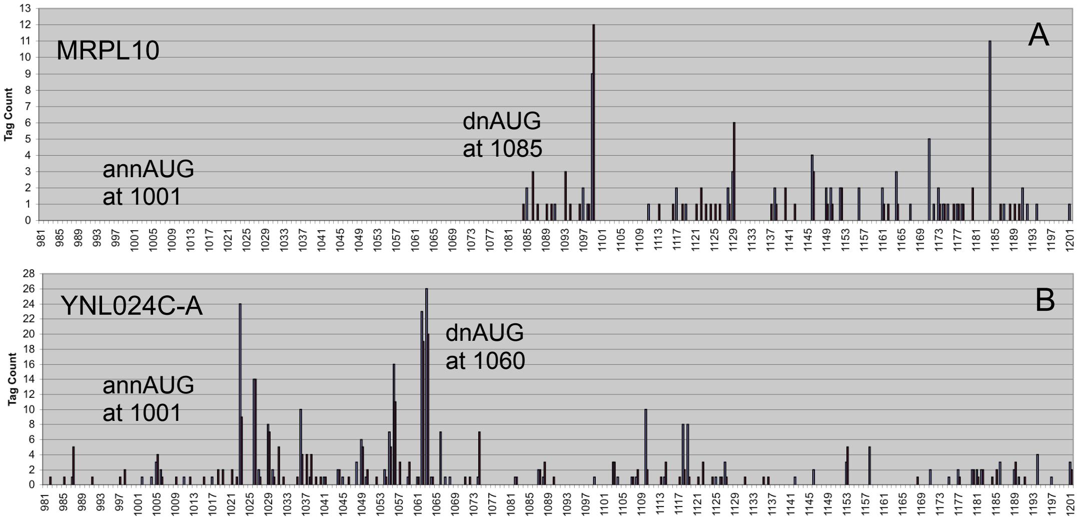


Supplementary Fig. S4. (**A**) Codon bias tests. The average Codon Adaptation Index (CAI percentage) immediately downstream of the annAUG is elevated compared to upstream of the AUG. (**B**) A significant shift of 4.9 in CAI index is also seen at dnAUGs of frame-1 downPeptide genes (p = 0.05). dnAUGs of annPeptide genes do not show a significant increase, as is the case for dnAUGs 100-200 nt or 200-300 nt downstream of annAUGs in all genes (data not shown). Only codons in detected mRNAs are included in this analysis. The mean values before and after the AUG are illustrated in pink (A,B). (**C**) Bootstrap analysis of dnAUGs sampled with replacement from dnAUGs at 100 – 200 nt downstream of the annAUG; the CAI shift of 4.9% observed in (B) has p = 0.05. The analysis in (B) is for downPeptide genes conforming to a 1% false identification rate; genes conforming to a 5% rate show a less pronounced CAI shift of 2.3.

The bootstrap analysis was performed as follows: For each downPeptide gene, the sequence flanking a randomly chosen frame 1 dnAUG from a randomly chosen gene was selected such that the length of the upstream sequence was equal to the length between the annAUG and the dnAUG of a frame 1 downPeptide gene, and the downstream sequence (after the dnAUG) was set to a fixed length (30 codons). This was performed for each of the downPeptide genes resulting in a group of 29 sequences which were aligned at their dnAUGs and the average CAI for each position calculated, provided that there are at least 10 sequences extended to that position. The mean CAI upstream of the AUG was subtracted from the mean CAI downstream of the AUG. This was repeated 1000 times to generate a distribution of CAI differences in (C).



Supplementary Fig. S5. Examination of predicted protein disorder for protein sequence between the annAUG and dnAUG of downPeptide genes. For each downPeptide gene, the mean predicted disorder score for amino acids coded between the annAUG and dnAUG was computed using a scoring weight matrix (*16*) (red curve). For each downPeptide gene, control amino acid sequences of the same length were selected from the annotated amino termini of 500 randomly selected genes (blue). The analysis was limited to downPeptide genes with >5 codons between annAUG and implicated dnAUG. 70 of the 115 downPeptide genes had disorder scores >0 which is significantly higher than the control sequences (chi-square $p < 0.01$), suggesting that the protein segments encoded between the annAUG and dnAUG tend to be disordered.



Supplementary Fig. S6. Ribosome profiles of individual genes.

(**A**) DownPeptide gene *MRPL10* had an amino peptide mapping to the dnAUG at coordinate 1085. The ribosome profile for this gene suggests most translation initiation is at the dnAUG. (**B**) An amino peptide mapped to 1060 in *YNL024C-A*. The gene's ribosome profile suggests translation initiation at this dnAUG as well as the annAUG, although in individual gene profiles, it can be difficult to distinguish initiation from ribosome pausing (*17*).

Supplementary Tables

Supplementary Table S1. TurboSequest probability thresholds

sample accession[1]	forward peptides[2,4]	decoy peptides[2,4]	probability threshold[5]	PeptideAtlas modifications	additional modifications (this study)
PAe000079	233	11	0.0033	C+9 opt., C+227.13 static	M+16 opt., N-term+42 opt.
PAe000084	12097	589	0.0810	C+9 opt., C+227.13 static	M+16 opt., N-term+42 opt.
PAe000086	13901	688	0.1329	C+9 opt., C+227.13 static	M+16 opt., N-term+42 opt.
PAe000088	16262	785	0.2025	C+9 opt., C+227.13 static	M+16 opt., N-term+42 opt.
PAe000089	10239	505	0.1125	C+9 opt., C+227.13 static	M+16 opt., N-term+42 opt.
PAe000120	1720	85	0.4783	none	M+16 opt., N-term+42 opt.
PAe000125	6209	161	0.9113	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000141	5549	274	0.8100	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000142	1193	36	1.000	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000145	7054	328	0.0405	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000149	1916	94	0.1823	none	M+16 opt., N-term+42 opt.
PAe000151	819	37	0.0729	C+8 opt., C+442.2 static	M+16 opt., N-term+42 opt.
PAe000158	8641	432	0.1640	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000160	4894	241	0.1125	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000162	1125	56	0.0133	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000164	9552	467	0.1125	C+9 opt., C+227.13 static	M+16 opt., N-term+42 opt.
PAe000165	10105	499	0.1476	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000166	2239	108	0.0531	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000167	422	20	0.0045	C+57.02 static	M+16 opt., N-term+42 opt.
PAe000324	57583	2834	0.5314	C+57.02 static	M+16 opt., N-term+42 opt.
ACET	250[3,4]	12[3,4]	0.0600	N/A	M+16 opt., N-term+68 opt., K+68 opt., P(N-term)+86 opt.
GLUT	175[3,4]	9[3,4]	0.2800	N/A	M+16 opt., N-term+42 opt., K+68 opt., P(N-term)+86 opt.

[1] http://www.peptideatlas.org/repository/
[2] Matches to N-terminal, internal and C-terminal peptides with no internal trypsin sites, and TurboSequest RSp rank = 1
[3] Matches to N-terminal peptides with no internal trypsin sites, and TurboSequest RSp rank = 1
[4] Peptides detected multiple times are counted only once per MS/MS experiment
[5] TurboSequest probability threshold for false identification rate < 5%

Supplementary Table S2. Gene Status

gene status*	annPeptide genes	downPeptide genes	annotated genes
Verified	424	203	4673
Uncharacterized	75	54	1118
Dubious	50	42	813
Transposable element gene	7	0	89
Pseudogene	2	0	21
Silenced gene	1	0	4

* http://www.yeastgenome.org

Supplementary Table S3.
Frequencies of annPeptides and downPeptides detected for the same gene

gene id	gene name	dnAUG location(s)*	annPeptide count †	downPeptide count †
YDR385W	EFT2	1053, 1064	46	1, 13
YIL053W	RHR2	1053	24	1
YFR053C	HXK1	1046, 1097	22	1, 6
YKL152C	GPM1	1074	12	1
YJR024C	YJR024C	1032	8	6
YLR293C	GSP1	1059	7	1
YLL041C	SDH2	1055	3	2
YER044C	ERG28	1052	2	2
YOL061W	PRS5	1008	2	2
YLR345W	YLR345W	1007	2	1
YPR048W	TAH18	1032	2	1
YBR249C	ARO4	1034	1	3
YGL255W	ZRT1	1059	1	2
YLL053C	YLL053C	1095	1	2
YDL185W	TFP1	1083	1	1
YEL048C	YEL048C	1074	1	1
YGL041C	YGL041C	1046	1	1
YGR097W	ASK10	1049	1	1
YIL074C	SER33	1013	1	1
YIL151C	YIL151C	1031	1	1
YJL171C	YJL171C	1037	1	1
YJL187C	SWE1	1082	1	1
YKL209C	STE6	1079	1	1
YKL216W	URA1	1019	1	1
YKR024C	DBP7	1028	1	1
YLR003C	YLR003C	1026	1	1
YOR200W	YOR200W	1008	1	1
YOR317W	FAA1	1075	1	1

* Nucleotide locations of dnAUG compared to annAUG (at 1001, 1002, 1003)
† Number of independent matches to annPeptides and downPeptides in multiple MS/MS experiments. Illustrated are genes with ≥ 1 matches to both an annPeptide and downPeptide. For most of these genes, the annPeptides and downPeptides were detected at similar frequencies.

Supplementary Table S4. Peptide Reading Frames

		reading frame[1]		
AUG type	**AUG rank[2]**	**1**	**2**	**3**
AnnPeptides	0	583	0	0
Down-Peptides	1	60	76	20
	2	41	49	9
	3	17	17	2
	4	12	8	1
	5	1	5	0
	>5	1	1	0
Annotated Genes[3]	0	6711	0	0
	1	1356	2511	572
	2	816	1416	332
	3	417	654	118
	4	180	240	45
	5	69	84	17

[1] Reading frame relative to annotated ORF (frame 1)
[2] The rank indicates the AUG rank relative to the annotated AUG (rank = 0)
[3] All AUGs within 100 nt downstream of the annotated AUG, the region screened by SEQUEST analysis, and encoding a theoretical peptide $\geq$ 5 amino acids

Supplementary Table 5A. Ribosome densities of downPeptide genes

				Ribosome tag counts[1]		
id	gene name[2]	annORF length	annAUG to dnAUG distance	-30 to annAUG-1	annAUG to annAUG+30	dnAUG to dnAUG+30
28739	YDR524C-B	200	70	36	5084	5646
1635	GPM1	743	73	4	3889	3871
2793	EFT2	2528	52	4	1628	1660
2793	EFT2	2528	63	4	1628	1172
2976	PMA1	2756	91	111	1385	360
2344	TFP1	3215	58	0	617	235
3222	HXK2	1460	45	2	592	258
7524	YMR122W-A	254	73	7	302	109
1315	RHR2	752	52	11	262	273
2634	ADK1	668	91	1	244	261
453	ARO4	1112	33	3	230	70
3882	ATP2	1535	96	0	219	37
4284	GSP1	659	58	1	192	163
4604	MIC17	470	72	3	185	14
1699	URA1	944	78	8	161	161
4276	CTS1	1688	55	31	154	147
4219	CDC42	575	94	0	138	59
5844	FAA1	2102	74	0	130	19
1407	PAN6	929	96	0	104	15
3516	ERV29	932	36	0	89	25
4359	RPS22B	392	34	0	89	189
4182	HCR1	797	99	1	74	36
2391	OST4	110	76	1	73	50
3367	PRE9	776	91	12	71	82
5057	RPC19	428	93	0	70	22
4247	YLR257W	965	58	5	67	71
4511	GSF2	1211	28	1	53	23
1868	STE2	1295	88	8	47	85
2534	ARO1	4766	34	0	47	36
1476	HYR1	491	91	9	46	45
2418	SNQ2	4505	37	12	46	24
28698	YNL024C-A	218	59	1	46	61
3545	CCT8	1706	73	0	45	24
1933	RSC8	1673	33	0	44	29
1949	HXK1	1457	45	2	41	10
1692	STE6	3872	81	0	38	22
2264	PHO2	1679	70	0	37	13
2560	ENT5	1235	40	0	37	20
2161	MCD1	1700	87	2	36	13
2864	NHX1	1901	76	0	33	13
3427	SKI6	740	43	4	33	38
517	GBP2	1283	91	3	30	23
3492	TNA1	1604	46	0	29	51
3492	TNA1	1604	43	0	29	45
4596	RSC9	1745	81	0	29	16
2939	YDR531W	1103	40	0	28	29
113	PEP1	4739	58	12	27	25
1778	YKR070W	1058	61	0	27	18

846	ERG28	446	51	4	25	11
486	MRPL27	440	43	0	24	16
1734	GCN3	917	63	2	24	52
680	TUP1	2141	37	8	23	40
2352	NUS1	1127	76	1	20	11
3171	KEX1	2189	97	0	19	12
5212	LYP1	1835	72	0	17	22
2597	SLY1	2000	48	3	16	19
7615	YKL018C-A	299	78	0	16	11
2553	SWI5	2129	82	0	15	10
4221	BNA5	1361	78	0	14	18
5931	RET3	569	67	0	14	75
1597	APN1	1103	97	0	13	10
3379	NAT2	866	53	0	12	12
3785	YJR024C	734	31	0	12	19
5095	RPC31	755	45	2	12	33
774	YEL048C	458	94	0	11	11
6392	MLC2	491	75	0	11	24
6430	MRP10	287	99	1	10	23
1234	CTF8	401	87	1	9	11
3707	YJL171C	1190	30	0	9	20
5144	YNL200C	740	57	0	8	16
5052	YNL108C	812	31	0	7	24
4694	VBA1	1688	79	4	5	10
5411	GAL11	3245	84	0	5	15
1336	SER33	1409	48	0	4	39
1389	YIL127C	620	52	0	4	13
2818	STE14	719	28	0	4	11
3411	OKP1	1220	55	0	3	11
6283	MRL1	1145	67	1	3	16
3268	CAX4	719	37	0	1	21
5066	YNL122C	345	51	0	0	26
5228	MRPL10	966	84	0	0	14

[1] Ribosome tag densities were compared in three 30-nt windows (i) starting 30 nt upstream of the annAUG, (ii) at the annAUG, and (iii) at the dnAUG. The tags in the dnAUG window are likely due to ribosomes that initiated translation at the dnAUG or, in some cases, the annAUG.

[2] We analyzed 81 of the 320 downPeptide genes with the following properties: (i) > 0.1 tags/nt in the first 200 nt of their annotated ORF, (ii) ORF ≥ 200, (iii) ≥10 tags/30 nt in the 30-nt window starting at the implicated dnAUG, and (iv) annAUG to dnAUG distance ≥ 30. Almost all of these genes have ribosome tags in the window starting at the annAUG, consistent with translation initiation at the annAUG as well as the dnAUG initiation implicated by our MS/MS analysis; only five of these genes have ≤ 3 tags in this window, a density similar to background levels in the 30-nt window upstream of the annAUG.

Supplementary Table S6. DownPeptide genes with truncated 5'UTRs

id[1]	gene id	gene name	max 5'UTR	min 5'UTR	5' cap closest to dnAUG[2]	dnAUG location[3]	dnAUG frame
48	YAL051W	OAF1	8	8	8	1032	2
48	YAL051W	OAF1	8	8	8	1044	2
769	YEL043W	YEL043W	8	8	8	1033	3
774	YEL048C	YEL048C	21	16	16	1095	2
846	YER044C	ERG28	83	8	8	1052	1
1105	YHR063C	PAN5	43	18	18	1049	1
1234	YHR191C	CTF8	42	-3	-3	1088	1
1241	YHR198C	YHR198C	-17	-17	-17	1057	3
1304	YIL042C	PKP1	8	8	8	1021	3
1304	YIL042C	PKP1	8	8	8	1027	3
1315	YIL053W	RHR2	51	13	13	1053	2
1401	YIL139C	REV7	342	-78	-47	1074	2
1447	YIR008C	PRI1	14	14	14	1023	2
1635	YKL152C	GPM1	46	11	11	1074	2
1778	YKR070W	YKR070W	69	-38	-38	1062	2
2391	YDL232W	OST4	52	9	9	1077	2
2671	YDR263C	DIN7	139	3	3	1067	1
2887	YDR479C	PEX29	90	-9	-9	1029	2
3268	YGR036C	CAX4	178	-16	-16	1038	2
3492	YGR260W	TNA1	-3	-3	-3	1044	2
3492	YGR260W	TNA1	-3	-3	-3	1047	2
3725	YJL189W	RPL39	43	14	14	1075	3
3964	YLL041C	SDH2	66	16	16	1055	1
3993	YLR003C	YLR003C	38	18	18	1026	2
4105	YLR115W	CFT2	10	10	10	1026	2
4255	YLR265C	NEJ1	3	3	3	1097	1
4299	YLR308W	CDA2	111	16	16	1033	3
4360	YLR368W	MDM30	6	6	6	1025	1
4449	YLR457C	NBP1	38	-15	-15	1050	2
4702	YMR096W	SNZ1	220	-565	-3	1064	1
4763	YMR154C	RIM13	21	19	19	1005	2
4793	YMR181C	YMR181C	197	-660	-61	1071	2
4849	YMR236W	TAF9	20	16	16	1095	2
5057	YNL113W	RPC19	9	6	6	1094	1
5178	YNL234W	YNL234W	53	-58	-58	1082	1
5178	YNL234W	YNL234W	53	-58	-58	1085	1
5212	YNL268W	LYP1	350	-7	-7	1073	1
5228	YNL284C	MRPL10	-42	-1056	-42	1085	1
5434	YOL073C	YOL073C	28	12	12	1034	1
5489	YOL129W	VPS68	48	-9	-9	1017	2
5551	YOR025W	HST3	70	15	15	1046	1
5931	YPL010W	RET3	83	-29	-29	1068	2
5957	YPL036W	PMA2	556	-9	-9	1071	2

6041	YPL120W	VPS30	16	16	16	1015	3
6286	YPR082C	DIB1	26	-31	-31	1036	3
6392	YPR188C	MLC2	35	4	4	1076	1
6430	YDL045W-A	MRP10	54	-3	-3	1100	1
6432	YIL009C-A	EST3	37	-3	-3	1095	2
7524	YMR122W-A	YMR122W-A	48	19	19	1074	2
28508	YCL005W-A	VMA9	25	10	10	1049	1
28708	YOL085W-A	YOL085W-A	-65	-65	-65	1097	1
28736	YBR056W-A	YBR056W-A	44	14	14	1092	2
28739	YDR524C-B	YDR524C-B	78	1	1	1071	2

[1] All genes with truncated 5'UTRs where the 5'cap is within 20 nt upstream of the annAUG or between the annAUG and implicated dnAUG. Based on datasets of Miura et al. (*18*) and Nagalakshmi et al. (*19*).
[2] 5' cap notated relative to annAUG; caps downstream of the annAUG have negative values
[3] dnAUG location where annAUG is at 1001

Supplementary Table S7A.
Frame 1 DownPeptide genes with predicted signal sequences between annAUG and dnAUG

id	gene name	len	sequence	type	prob	cleavage prob	cleavage location
92	YAR069C	36	MEDHTLVAIVVFFGNG**E**PFHVSLSVE MVFVLLLSST	Signal peptide	0.569	0.325	16-17
846	ERG28	27	MFSLQDVITTTKTTLAA MPKGYLPKWL	Non-secretory	-	0.005	16-17
915	YER113C	22	MRVRPKRSVITL MAIVVVMLIL	Signal anchor	0.992	0.000	-
1105	PAN5	26	MTAPHRSTIHILGLGA MGTVLAVDLL	Signal anchor	0.684	0.052	18-19
1121	IRE1	17	MRLLRRN MLVLTLLVCV	Non-secretory	-	0.000	-
1478	YPS6	23	MQLISILSLLSSL MC**S**LTVLGSS	Signal peptide	0.958	0.523	15-16
1670	YKL187C	35	MRIEKHRTPLSKGIIWTILSVCLLF MFTTLILVIV	Signal anchor	1.000	0.000	-
1699	URA1	36	MTASLTTKFLNNTYENPFMNASGVHC MTTQELDELA	Non-secretory	-	0.000	-
1781	YKR073C	36	MIVFDVSLMIIIIFSFAFNMSQS**N**IL MLYNSPHVLV	Signal peptide	0.803	0.549	23-24
3468	SPG1	33	MKLDSGIYSEAQRVVRTPKFRYI MLGLVGAAVV	Non-secretory	-	0.000	-
3603	YJL067W	30	MSKKRKRKYVLIVFVNTHHF MLHLGTGTLG	Non-secretory	-	0.013	25-26
3707	YJL171C	20	MLQSIVLSVCMF MLHTVAAS	Non-secretory	-	0.002	14-15
3822	YJR061W	19	MMLSLRRFS MYVLRSLRLH	Non-secretory	-	0.000	-
3964	SDH2	28	MLNVLLRRKAFCLVTKKG MATATTAAAT	Non-secretory	-	0.004	22-23
3976	YLL053C	19	MWFPQIIAG MAAGGAASAM	Non-secretory	-	0.003	13-14
4702	SNZ1	31	MTGEDFKIKSGLAQMLKGGVI MDVVTPEQAK	Non-secretory	-	0.000	-
5118	YNL174W	26	MALEFLAATRGMDNLV MSCSVTLLFS	Non-secretory	-	0.007	18-19
5181	YTP1	42	MTAANKNIVFGFSRSISAILLICFFFEKVCG**D** MEHDMGMDDT	Signal peptide	0.927	0.914	31-32
5272	MDJ2	22	MVLPIIIGLGVT MVALSVKSGL	Non-secretory	-	0.156	15-16
5740	YOR214C	24	MLGLYLSSLFFAFF MA**Q**VFATKYS	Signal peptide	0.591	0.552	16-17
6350	YPR146C	16	MQNMLS MHFFSVMASL	Non-secretory	-	0.000	-
6430	MRP10	43	MSGKPPVYRLPPLPRLKVKKPIIRQEANKCLVL MSNLLQCWSS	Non-secretory	-	0.000	-
7615	YKL018C-A	36	MLGMIRWVVEGTLVAMLLSA**I**RRETG MIFFYNQYQL	Non-secretory	-	0.147	20-21
28508	VMA9	26	MSSFYTVVGVFIVVSA MSVLFWIMAP	Signal anchor	0.853	0.072	16-17
28769	YGL006W-A	26	MLIFIIHYHRHLALHL MGAFQKHSNS	Non-secretory	-	0.001	19-20

Supplementary Table S7B.

id	gene name	cell_loc	RC	mTP	SP	other
92	YAR069C	-	4	0.021	0.408	0.659
846	ERG28	-	1	0.096	0.062	0.903
915	YER113C	S	4	0.234	0.627	0.073
1105	PAN5	-	5	0.191	0.327	0.375
1121	IRE1	-	5	0.312	0.158	0.347
1478	YPS6	S	2	0.061	0.849	0.162
1670	YKL187C	S	2	0.025	0.878	0.134
1699	URA1	-	2	0.104	0.052	0.9
1781	YKR073C	S	1	0.017	0.951	0.073
3468	SPG1	-	3	0.193	0.075	0.713
3603	YJL067W	M	5	0.352	0.346	0.095
3707	YJL171C	S	4	0.061	0.669	0.382
3822	YJR061W	M	5	0.539	0.029	0.5
3964	SDH2	M	4	0.597	0.052	0.353
3976	YLL053C	-	2	0.128	0.083	0.827
4702	SNZ1	-	1	0.059	0.087	0.924
5118	YNL174W	-	2	0.114	0.087	0.872
5181	YTP1	S	4	0.133	0.514	0.188
5272	MDJ2	S	5	0.086	0.587	0.45
5740	YOR214C	S	2	0.062	0.864	0.145
6350	YPR146C	-	3	0.234	0.045	0.775
6430	MRP10	-	4	0.374	0.067	0.664
7615	YKL018C-A	S	2	0.06	0.829	0.104
28508	VMA9	S	1	0.039	0.91	0.103
28769	YGL006W-A	-	5	0.249	0.217	0.439

25 of 152 Frame 1 downPeptide genes have predicted signal sequences (SignalP (*20*)) between their annAUG and dnAUG (corresponding methionines in red).

The first amino acid after predicted signal peptide cleavage sites are indicated in large bold (5 proteins).

In many cases, no cleavage site is predicted, suggesting the annPeptide protein may be anchored to the ER membrane, whereas the downPeptide protein would not enter the ER.

Part B illustrates cellular localizations predicted by SignalP; M: mitochondria; S: secretory.

Supplementary Table S8. Motif Predictions for Frame 1 downPeptide Genes.

gene name	id	sequence	motif detected by InterProScan	statistical signif.	algorithm
ERG28	846	MFSLQDVITTTKTTLAA MPKGYLPKWL	Family Not Named	3.10E-07	HMMPanther
PAN5	1105	MTAPHRSTIHILGLGA MGTVLAVDLL	Signal Peptide	NA	SignalPHMM
IRE1	1121	MRLLRRN MLVLTLLVCV	Signal Peptide	NA	SignalPHMM
SER33	1336	MSYSAADNLQDSFQRA MNFSGSPGAV	D-3-phosphoglycerate dehydrogenase	6.90E-05	HMMPanther
		MSYSAADNLQDSFQRA MNFSGSPGAV	2-hydroxyacid dehydrogenase-related	6.90E-05	HMMPanther
PAN6	1407	MKIFHTVEEVVQWRTQELRETRFRETIGFVPT MGCLHSGHAS	Pantoate_ligase	1.60E-12	HMMPfam
		MKIFHTVEEVVQWRTQELRETRFRETIGFVPT MGCLHSGHAS	Nucleotidylyl transferase	3.70E-07	superfamily
		MKIFHTVEEVVQWRTQELRETRFRETIGFVPT MGCLHSGHAS	Rossmann-like alpha/beta/alpha sandwich fold	3.80E-08	Gene3D
YPS6	1478	MQLISILSLLSSL MCSLTVLGSS	Signal Peptide	NA	SignalPHMM
YKL187C	1670	MRIEKHRTPLSKGIIWTILSVCLLF MFTTLILVIV	Signal Peptide	NA	SignalPHMM
		MRIEKHRTPLSKGIIWTILSVCLLF MFTTLILVIV	transmembrane_regions	NA	TMHMM
URA1	1699	MTASLTTKFLNNTYENPFMNASGVHC MTTQELDELA	dihydroorotate dehydrogenase	1.90E-18	HMMPanther
		MTASLTTKFLNNTYENPFMNASGVHC MTTQELDELA	Dihydroorotate dehydrogenase, class 1/ 2	3.90E-06	HMMPfam
YKR073C	1781	MIVFDVSLMIIIIFSFAFNMSQSNIL MLYNSPHVLV	Signal Peptide	NA	SignalPHMM
		MIVFDVSLMIIIIFSFAFNMSQSNIL MLYNSPHVLV	transmembrane_regions	NA	TMHMM
MCD1	2161	MVTENPQRLTVLRLATNKGPLAQIWLASN MSNIPRGSVI	sister chromatid cohesion protein 1	1.00E-12	HMMPanther
DIN7	2671	MGIPGLLPQLKRIQKQVSLKKY MYQTLAIDGY	xpg_n	3.70E-06	HMMPfam
		MGIPGLLPQLKRIQKQVSLKKY MYQTLAIDGY	exonuclease 1	1.70E-07	HMMPanther
EFT2	2793	MVAFTVDQMRSLMDKVTNVRN MSVIAHVDHG	eukaryotic translation elongation factor 2	2.00E-09	HMMPanther
		MVAFTVDQMRSLMDKVTNVRN MSVIAHVDHG	P-loop containing nucleoside triphosphate hydrolases	3.10E-06	superfamily
SPG1	3468	MKLDSGIYSEAQRVVRTPKFRYI MLGLVGAAVV	Signal Peptide	NA	SignalPHMM
YJL067W	3603	MSKKRKRKYVLIVFVNTHHF MLHLGTGTLG	Signal Peptide	NA	SignalPHMM
YJL171C	3707	MLQSIVLSVCMF MLHTVAAS	Signal Peptide	NA	SignalPHMM
YJR061W	3822	MMLSLRRFS MYVLRSLRLH	Signal Peptide	NA	SignalPHMM
SDH2	3964	MLNVLLRRKAFCLVTKKG MATATTAAAT	Signal Peptide	NA	SignalPHMM
YLL053C	3976	MWFPQIIAG MAAGGAASAM	Signal Peptide	NA	SignalPHMM
HCR1	4182	MSWDDEAINGSMGNDDAVLMDSWDAEIGDDEPV MQSWDAEEEE	Translation initiation factor eIF3 subunit	1.80E-07	HMMPfam
SNZ1	4702	MTGEDFKIKSGLAQMLKGGVI MDVVTPEQAK	SOR_SNZ (Vitamin B6 biosynthesis)	1.40E-09	HMMPfam
		MTGEDFKIKSGLAQMLKGGVI MDVVTPEQAK	PDXS_SNZ_2 (Vitamin B6 biosynthesis)	19.276	ProfileScan
YNL174W	5118	MALEFLAATRGMDNLV MSCSVTLLFS	Signal Peptide	NA	SignalPHMM
YNL200C	5144	MSTLKVVSSKLAAEIDKEL MGPQIGFTLQ	n-terminal yjef related	3.80E-09	HMMPanther
YNL234W	5178	MTGEKILHSQLLTNSDMSSGNVHHTKP MMYNVTLPSY	Family Not Named	7.80E-16	HMMPanther
YNL234W	5178	MTGEKILHSQLLTNSDMSSGNVHHTKPM MYNVTLPSYN	Family Not Named	4.20E-16	HMMPanther
YTP1	5181	MTAANKNIVFGFSRSISAILLICFFFEKVCGD MEHDMGMDDT	signal-peptide	NA	SignalPHMM
		MTAANKNIVFGFSRSISAILLICFFFEKVCGD MEHDMGMDDT	transmembrane_regions	NA	TMHMM
MDJ2	5272	MVLPIIIGLGVT MVALSVKSGL	signal-peptide	NA	SignalPHMM
YOR214C	5740	MLGLYLSSLFFAFF MAQVFATKYS	signal-peptide	NA	SignalPHMM
YPR146C	6350	MQNMLSMHFFSV MASL	signal-peptide	NA	SignalPHMM
MLC2	6392	MDHSESLTFNQLTQDYINKLKDAFQ MLDEDEDGLI	MYOSIN LIGHT CHAIN 2	9.20E-17	HMMPanther
		MDHSESLTFNQLTQDYINKLKDAFQ MLDEDEDGLI	EF_HAND_2	10.218	ProfileScan
MRP10	6430	MSGKPPVYRLPPLPRLKVKKPIIRQEANKCLVL MSNLLQCWSS	signal-peptide	NA	SignalPHMM

YKL018C-A	7615	MLGMIRWVVEGTLVAMLLSAIRRETG MIFFYNQYQL	signal-peptide	NA	SignalPHMM
VMA9	28508	MSSFYTVVGVFIVVSA MSVLFWIMAP	signal-peptide	NA	SignalPHMM
		MSSFYTVVGVFIVVSA MSVLFWIMAP	transmembrane_regions	NA	TMHMM
YGL006W-A	28769	MLIFIIHYHRHLALHL MGAFQKHSNS	signal-peptide	NA	SignalPHMM

34 of 152 frame 1 downPeptide genes have motifs predicted by InterProScan (*21*) (http://www.ebi.ac.uk/Tools/InterProScan/) located between their annAUG and downAUG + 10 residues.
Motif regions are underlined in the sequence.
Signal peptides detected by InterProScan or SignalP are illustrated in Supplementary Tables S6 and S7.

Supplementary Table S9.
Gene Ontology terms whose frequencies differ significantly in downPeptide and annPeptide genes

GO id	GO detail	downPeptide Percent	annPeptide Percent	all genes[1]	top 10% protein expression[2]	chi-square[3]
3677	DNA binding	0.09	0.05	0.08	0.05	4.00
3735	structural constituent of ribosome	0.02	0.06	0.04	0.20	6.36
3824	catalytic activity	0.08	0.15	0.08	0.17	8.30
5488	binding	0.02	0.07	0.04	0.10	7.79
5622	intracellular	0.05	0.10	0.06	0.21	5.88
5737	cytoplasm	0.31	0.51	0.33	0.65	31.24
5829	cytosol	0.03	0.07	0.07	0.10	6.22
5840	ribosome	0.03	0.08	0.05	0.24	7.19
6412	translation	0.04	0.10	0.05	0.29	11.81
6417	regulation of translation	0.02	0.08	0.02	0.20	11.65
8150	biological_process	0.21	0.14	0.20	0.02	6.79
16020	membrane	0.29	0.22	0.26	0.18	5.22
16021	integral to membrane	0.22	0.16	0.19	0.11	5.51
16301	kinase activity	0.07	0.04	0.03	0.03	4.69

[1] GO term representation in 6357 annotated yeast genes

[2] top 10 of protein expression based on TAP-epitope tagged expression levels (*1*)

[3] downPeptide and annPeptide frequencies differ significantly by chi-square tests (3.84, p = 0.05; 6.64, p = 0.01; 10.83, p = 0.001)

Supplementary Files

Supplementary File S1: Fournier_SuppData.xls contains data for detected dnPeptide and annPeptide genes, including the downPeptide sequences and where they map relative to the annAUG. For each peptide, the highest scoring TurboSequest matches are illustrated. Note that some peptides are shown more than once if the same peptide sequence was detected with different amino acid modifications.

1. Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003) Global analysis of protein expression in yeast, *Nature 425*, 737-741.
2. Lambert, J. P., Mitchell, L., Rudner, A., Baetz, K., and Figeys, D. (2009) A novel proteomics approach for the discovery of chromatin-associated protein networks, *Mol Cell Proteomics 8*, 870-882.
3. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes, *Nature protocols 1*, 2856-2860.
4. Robbins-Pianka, A., Rice, M. D., and Weir, M. P. (2010) The mRNA landscape at yeast translation initiation sites, *Bioinformatics 26*, 2651-2655.
5. Weir, M. P., and Rice, M. D. (2010) TRII: A Probabilistic Scoring of Drosophila melanogaster Translation Initiation Sites *EURASIP Journal on Bioinformatics and Systems Biology 2010*.
6. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling, *Science 324*, 218-223.
7. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature 423*, 241-254.
8. Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E. S. (2004) Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery, *J Comput Biol 11*, 319-355.
9. Higgins, D. G., and Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, *Gene 73*, 237-244.
10. Sharp, P. M., and Li, W. H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res 15*, 1281-1295.
11. Kochetov, A. V. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs, *Bioessays 30*, 683-691.
12. Kozak, M. (2002) Pushing the limits of the scanning mechanism for initiation of translation, *Gene 299*, 1-34.

13. Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008) Function and structure of inherently disordered proteins, *Curr Opin Struct Biol 18*, 756-764.
14. Lewis, S., Ashburner, M., and Reese, M. G. (2000) Annotating eukaryote genomes, *Curr Opin Struct Biol 10*, 349-354.
15. Polevoda, B., and Sherman, F. (2003) N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins, *J Mol Biol 325*, 595-622.
16. Weathers, E. A., Paulaitis, M. E., Woolf, T. B., and Hoh, J. H. (2007) Insights into protein structure and function from disorder-complexity space, *Proteins 66*, 16-28.
17. Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes, *Cell 147*, 789-802.
18. Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S., and Ito, T. (2006) A large-scale full-length cDNA analysis to explore the budding yeast transcriptome, *Proc Natl Acad Sci U S A 103*, 17846-17851.
19. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science 320*, 1344-1349.
20. Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions, *Nature methods 8*, 785-786.
21. Zdobnov, E. M., and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro, *Bioinformatics 17*, 847-848.