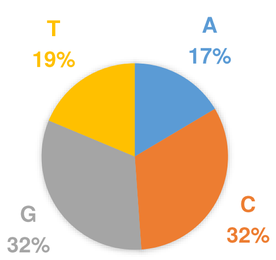
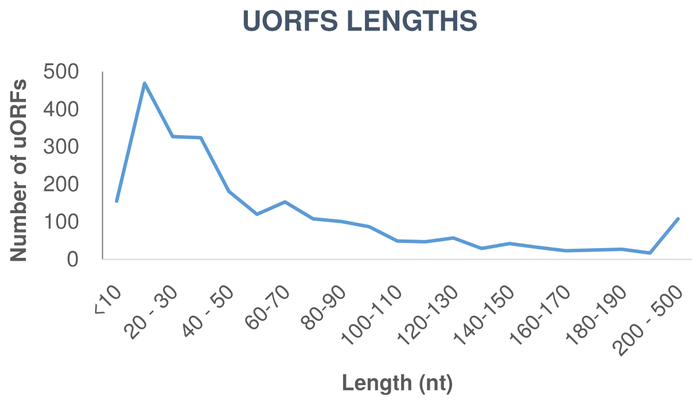
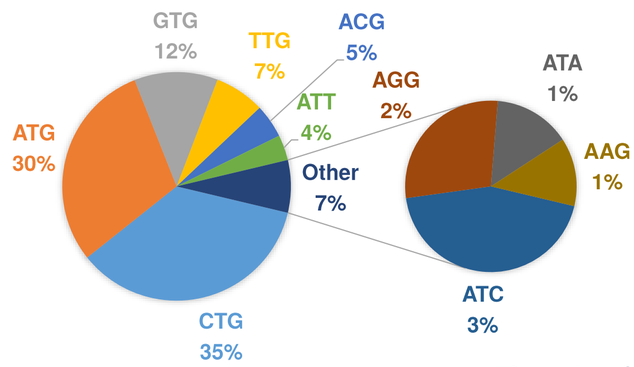
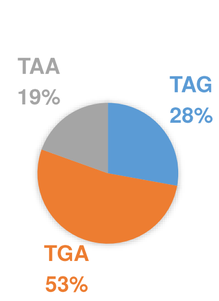




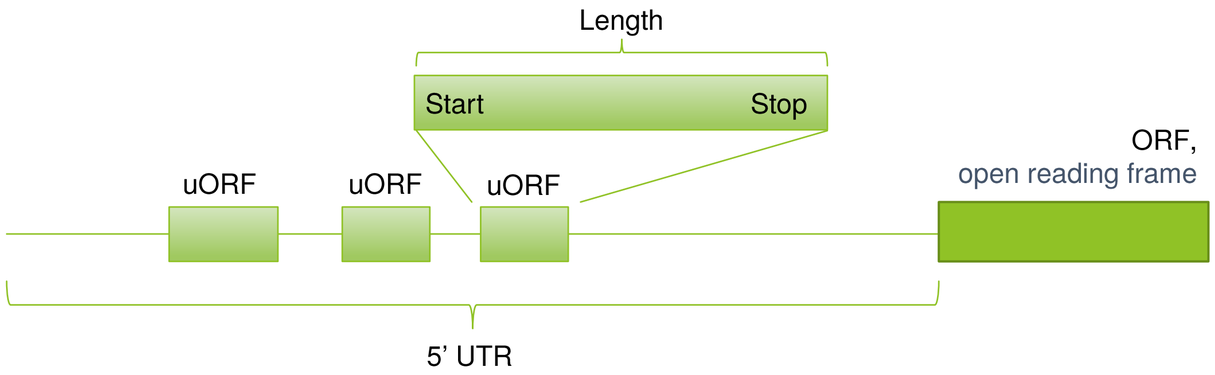




Basic features of uORFs

uORFs from a bird's eye view

K. Zaruba, A. Burenin, I. Eliseeva, I. Kulakovskiy
Laboratory of sequence analysis



Aim: Reveal specific features of upstream open reading frames in 5'UTRs of human mRNAs.

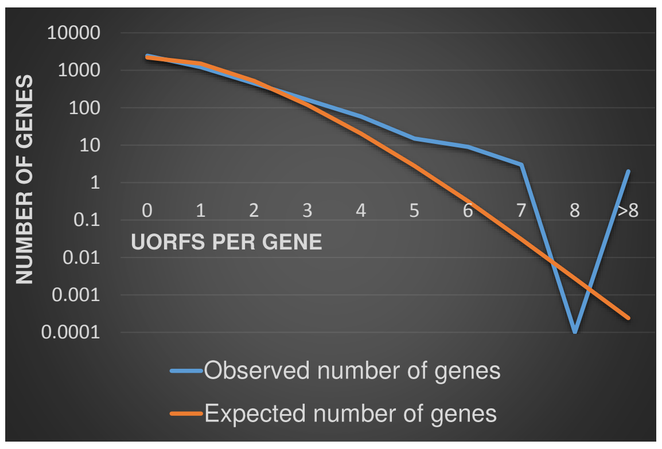
Methods: Ruby programming, Monte-Carlo simulations.

Initial data: 2994 uORFs in 5'UTRs of 4365 genes detected by ribosome footprinting in monocytes.

Sequence data: 2560 uORFs based on HGNC gene names and UCSC transcripts.

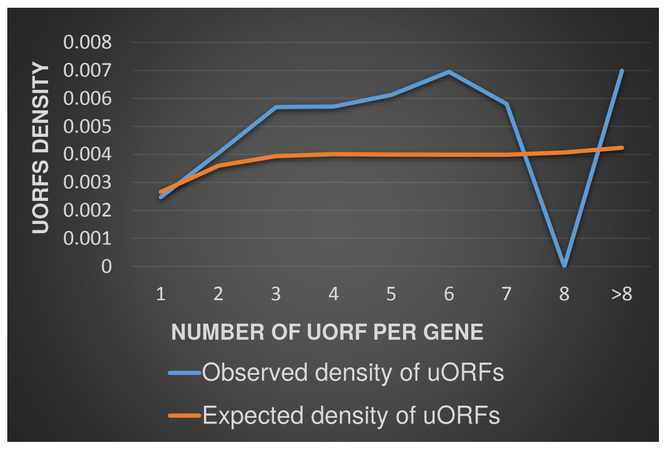
Start+Stop: 2481 sequences.

Results: uORFs form dense clusters. uORFs are enriched with GC-rich repeats.



Hypothesis: uORFs are uniformly distributed among genes.

Results: observed number of genes with a given number of uORFs is significantly higher than expected by chance (P-value << 0.05 on 10000 MC simulations) for genes with 3 and more uORFs.



Hypothesis: uORFs have uniform density within UTRs (i.e. uORFs are distributed uniformly along UTRs).

Results: observed density of uORFs is significantly higher than expected by chance (P-value << 0.05 on 1000 MC simulations) for genes with 2 and more uORFs.

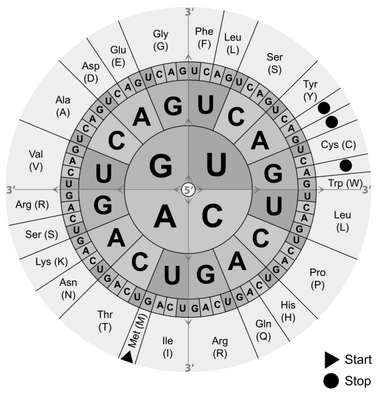
Triplet frequencies (all frames):

underrepresented;

overrepresented;

major triplets.

	Observed	Expected	Observed / Expected
ACG	1084	2334.91	0.46
ATA	353	683.20	0.52
ATG	620	1346.57	0.46
CAT	857	1345.33	0.64
CGT	1308	2651.60	0.49
CTA	712	1345.33	0.53
GTA	547	1346.57	0.41
TAA	414	683.20	0.61
TAC	705	1345.33	0.52
TAG	580	1346.57	0.43
TAT	486	775.87	0.63
TCG	1673	2651.60	0.63
AAA	1216	601.61	2.02
AGA	1989	1185.75	1.68
GAA	1839	1185.75	1.55
GAG	3665	2337.08	1.57
TTT	2042	881.10	2.32
CCC	4983	4593.52	1.08
CCG	4364	4597.78	0.95
CGC	3996	4597.78	0.87
CGG	4515	4602.04	0.98
GCC	4999	4597.78	1.09
GCG	4411	4602.04	0.96
GGC	4913	4602.04	1.07
GGG	4467	4606.31	0.97



Amino acid frequencies:

(blue) overrepresented compared to the genome;

(yellow) underrepresented compared to the genome, overrepresented compared to the expected;

(green) underrepresented.

Amino acid		uORFs	Expected (based on GC%)	Genome (ORFs)	Genome (ORFs >50% GC)	*ratio 1 / 2*	*1 / 3*	*1 / 4*
Arginine	R	**0.12**	0.13	0.06	0.06	0.91	2.11	2.00
Alanine	A	**0.11**	0.11	0.07	0.08	1.08	1.64	1.44
Proline	P	**0.11**	0.11	0.06	0.07	1.04	1.79	1.59
Glycine	G	**0.11**	0.11	0.07	0.08	1.00	1.60	1.27
Serine	S	**0.10**	0.09	0.08	0.07	1.15	1.24	1.44
Leucine	L	**0.09**	0.08	0.10	0.10	1.20	0.93	0.92
Valine	V	**0.05**	0.06	0.06	0.06	0.79	0.78	0.76
Glutamicacid	E	**0.04**	0.03	0.07	0.07	1.64	0.62	0.64
Threonine	T	**0.04**	0.05	0.05	0.05	0.77	0.77	0.78
Glutamine	Q	**0.03**	0.03	0.05	0.05	1.26	0.70	0.70
Phenylalanine	F	**0.03**	0.02	0.04	0.04	1.78	0.84	0.91
Cysteine	C	**0.03**	0.03	0.02	0.02	0.95	1.26	1.27
Lysine	K	**0.02**	0.01	0.06	0.05	1.82	0.43	0.52
Asparticacid	D	**0.02**	0.03	0.05	0.05	0.83	0.48	0.49
Tryptophan	W	**0.02**	0.02	0.01	0.01	1.01	1.50	1.53
Histidine	H	**0.02**	0.03	0.03	0.02	0.69	0.72	0.85
Isoleucine	I	**0.02**	0.02	0.04	0.04	0.84	0.39	0.47
Asparagine	N	**0.01**	0.01	0.04	0.03	0.98	0.37	0.44
Tyrosine	Y	**0.01**	0.02	0.03	0.03	0.56	0.32	0.32
Methionine	M	**0.01**	0.01	0.02	0.02	0.57	0.26	0.26

Major highly overrepresented 6nt motifs:

cyclic GC-rich repeats.

6nt motif	Observed	Expected	Observed / Expected
GCGGCG	407	148.71	2.74
GGCGGC	389	148.71	2.62
CGGCGG	336	148.71	2.26
GCCGCC	331	148.44	2.23
CGCCGC	309	148.44	2.08
CCGCCG	278	148.44	1.87
GCGGCC	253	148.57	1.70
CCTCCC	235	85.45	2.75
CCCGGC	222	148.44	1.50
CCCGGG	218	148.57	1.47
CCGCCC	217	148.30	1.46
GCCGGG	215	148.71	1.45
GGCGGG	213	148.85	1.43
GCCCGG	212	148.57	1.43
CCGGGC	211	148.57	1.42
CCCGCC	208	148.30	1.40
GGCCGC	207	148.57	1.39
GGGCGG	206	148.85	1.38
CCCTCC	205	85.45	2.40
CCGCGG	203	148.57	1.37
CGGCCG	202	148.57	1.36
CGGGGC	202	148.71	1.36
CGCGGC	200	148.57	1.35

Major highly overrepresented 3AA motifs:

poly* tracts and derivatives; frequencies are stable relative to frame shifts.

Motif	Observed	Expected (uORFs)	Expected (genome)	Expected (genome GC-rich)	*ratio*	RF1 (counts)	RF2 (counts)	RF3 (counts)
AAA	0.00626	0.00147	0.00034	0.00049	*4.2*	262	220	243
RRR	0.00449	0.00172	0.00018	0.00022	*2.6*	188	195	188
PPP	0.00351	0.00132	0.00023	0.00033	*2.7*	147	154	130
GGG	0.00325	0.00118	0.00029	0.00057	*2.8*	136	136	141
GRR	0.00277	0.00152	0.00021	0.00030	*1.8*	116	118	78
GRG	0.00249	0.00134	0.00025	0.00041	*1.9*	104	97	92
PGP	0.00244	0.00127	0.00025	0.00040	*1.9*	102	86	112
RRG	0.00239	0.00152	0.00021	0.00030	*1.6*	100	113	89
AAG	0.00237	0.00137	0.00032	0.00052	*1.7*	99	93	95
RRP	0.00229	0.00158	0.00020	0.00025	*1.5*	96	91	83
GPG	0.00225	0.00122	0.00027	0.00048	*1.8*	94	87	93
PAA	0.00225	0.00142	0.00030	0.00043	*1.6*	94	85	100
AAP	0.00213	0.00142	0.00030	0.00043	*1.5*	89	90	78
PLP	0.00210	0.00112	0.00038	0.00049	*1.9*	88	76	82
PGR	0.00210	0.00139	0.00023	0.00034	*1.5*	88	72	71

Speculation: peptide sequence is not important; regulation on the RNA level: G-quadruplexes, GC-rich hairpins?