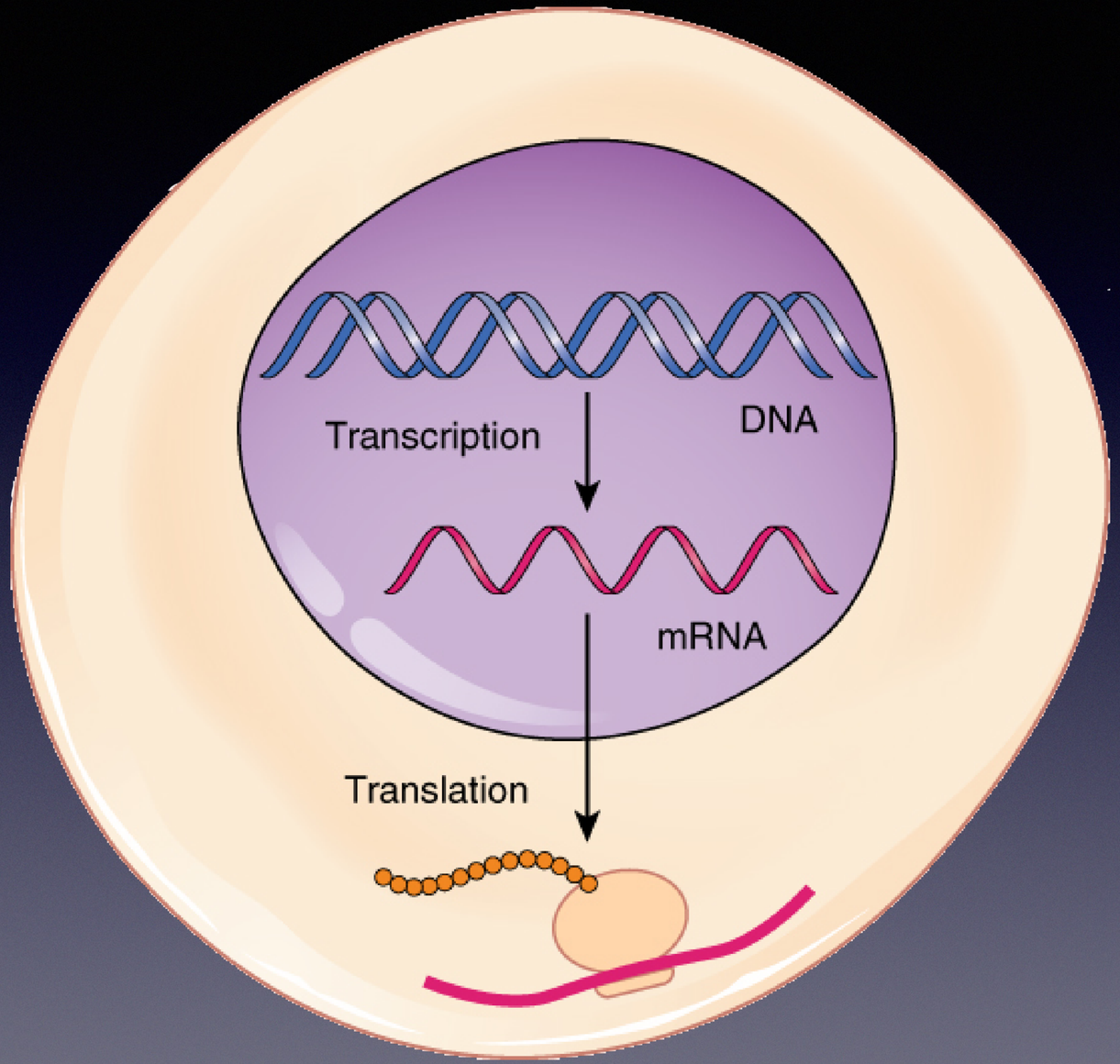
Upstream Open Reading Frames

As Potential Cancer Drivers

PROF. DR. WOJCIECH MAKAŁOWSKI, INSTITUTE OF BIOINFORMATICS, WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

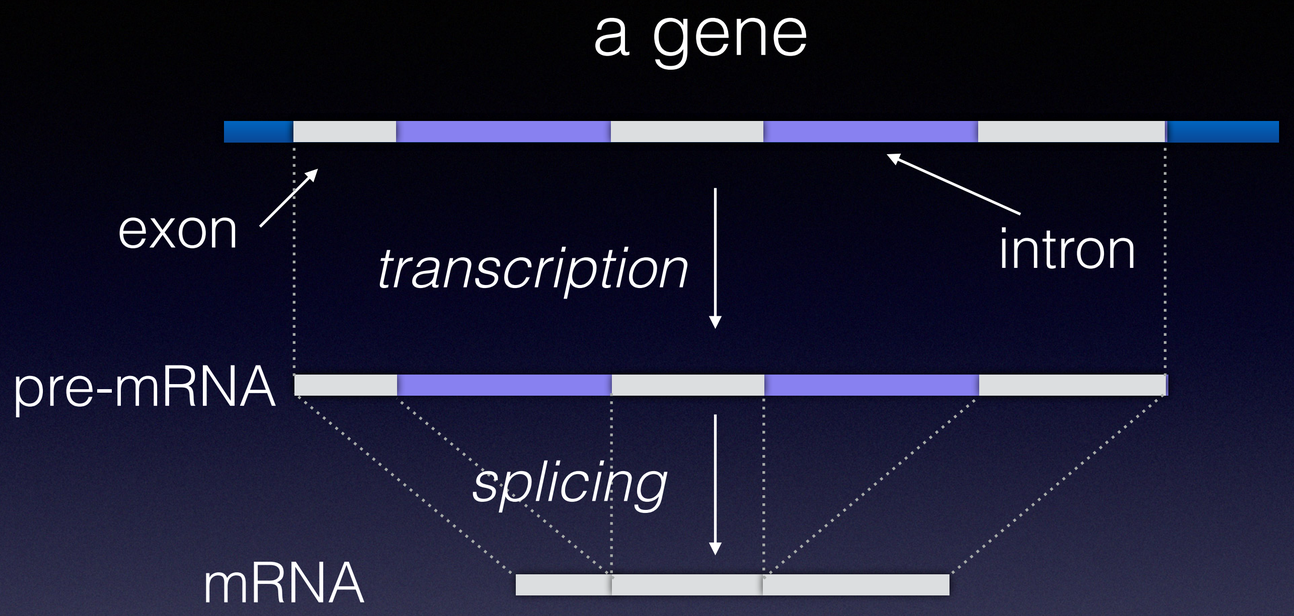
From gene to protein

a textbook version

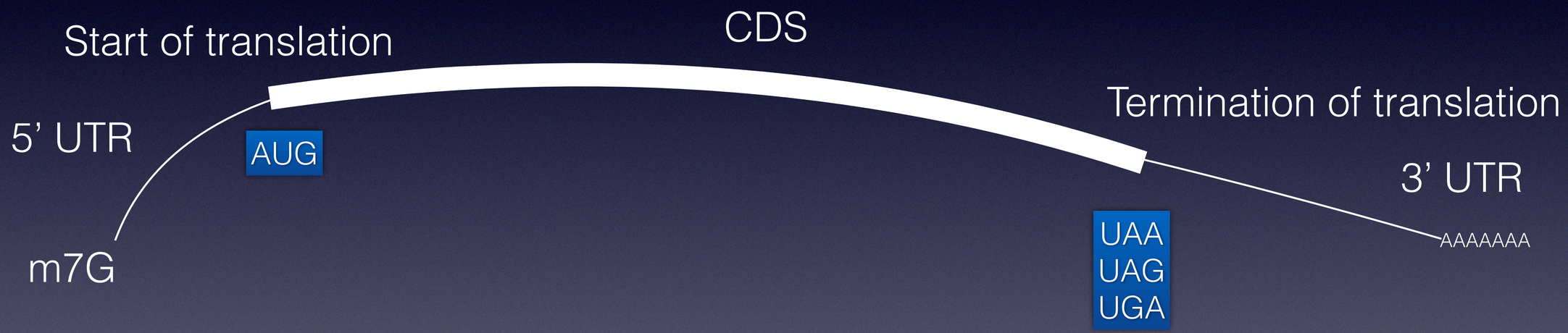


From gene to protein

a cell version



Mature mRNA


Start of translation
CDS
Termination of translation
5’ UTR
AUG
3’ UTR
m7G
UAA
UAG
UGA
AAAAAAA

But wait, there’s more…

But wait, there's more…

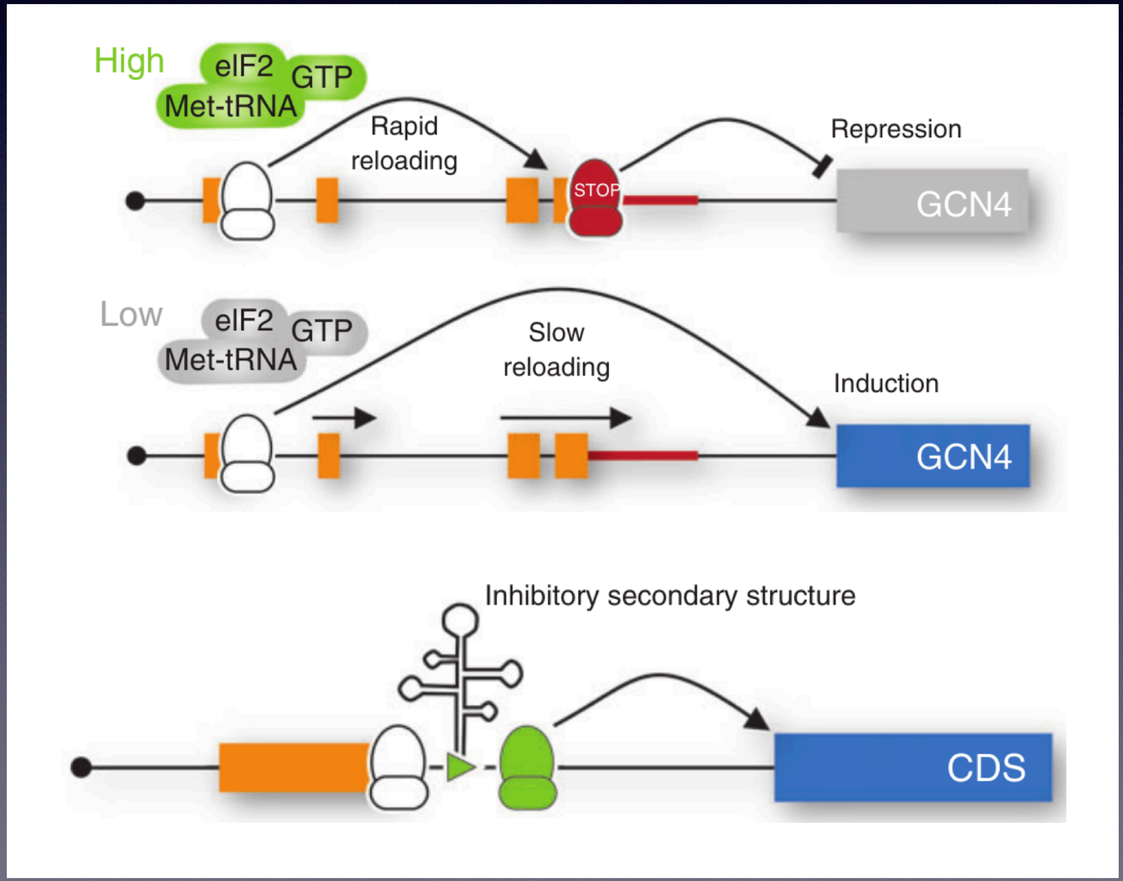
upstream Open Reading Frames (uORFs)

An uORF can start with an AUG with a strong or weak Kozak context
or
with an alternative initiation codon: AAG, AUU, UUG, AGG, AUC, GUG, ACG, CUG, AUA

Post-transcriptional gene expression regulation

Upstream open reading frames in 5’-UTR

- Exhaust the ribosome
- Mostly suppressive
- Adaptations and diseases

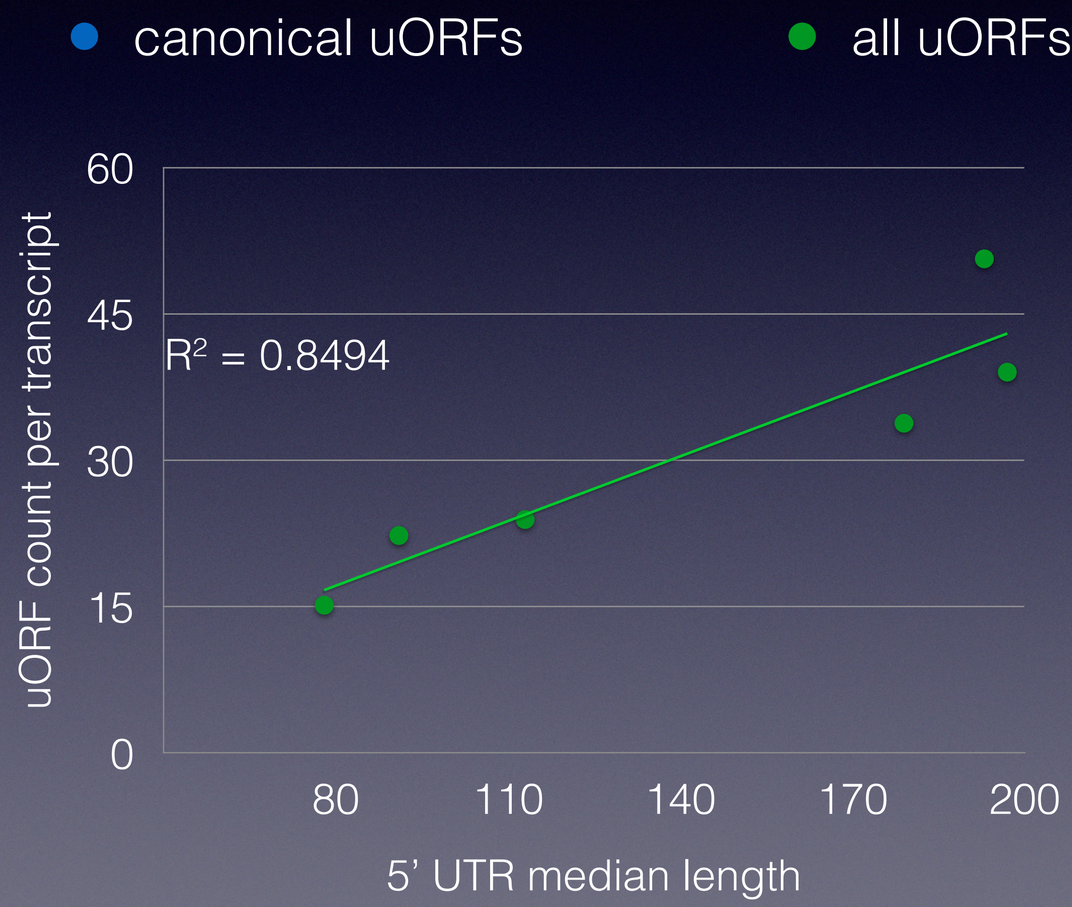


Wethmar, Klaus. 2014. „The Regulatory Potential of Upstream Open Reading Frames in Eukaryotic Gene Expression“. Wiley Interdisciplinary Reviews: RNA 5: 765–68. https://doi.org/10.1002/wrna.1245.

How frequent are uORFs in different species?

	Homo sapiens	*Mus musculus*	*Gallus gallus*	*Xenopus laevis*	*Danio rerio*	*Drosophila melanogaster*
Total number of genes	18,941	19,536	5,319	9,607	13,310	13,445
Total number of transcripts	60,217	36,414	5,665	9,617	13,965	30,085
Fraction of transcripts with canonical uORFs	0.593	0.559	0.332	0.728	0.592	0.615
Average number of canonical uORFs per transcript	2.54	1.93	0.83	2.41	1.76	2.68
Fraction of transcripts with uORFs	0.992	0.991	0.937	0.976	0.98	0.994
Average number of uORFs per transcript	33.82	55.9	15.17	22.13	23.95	50.66

Canonical and alternative start codons

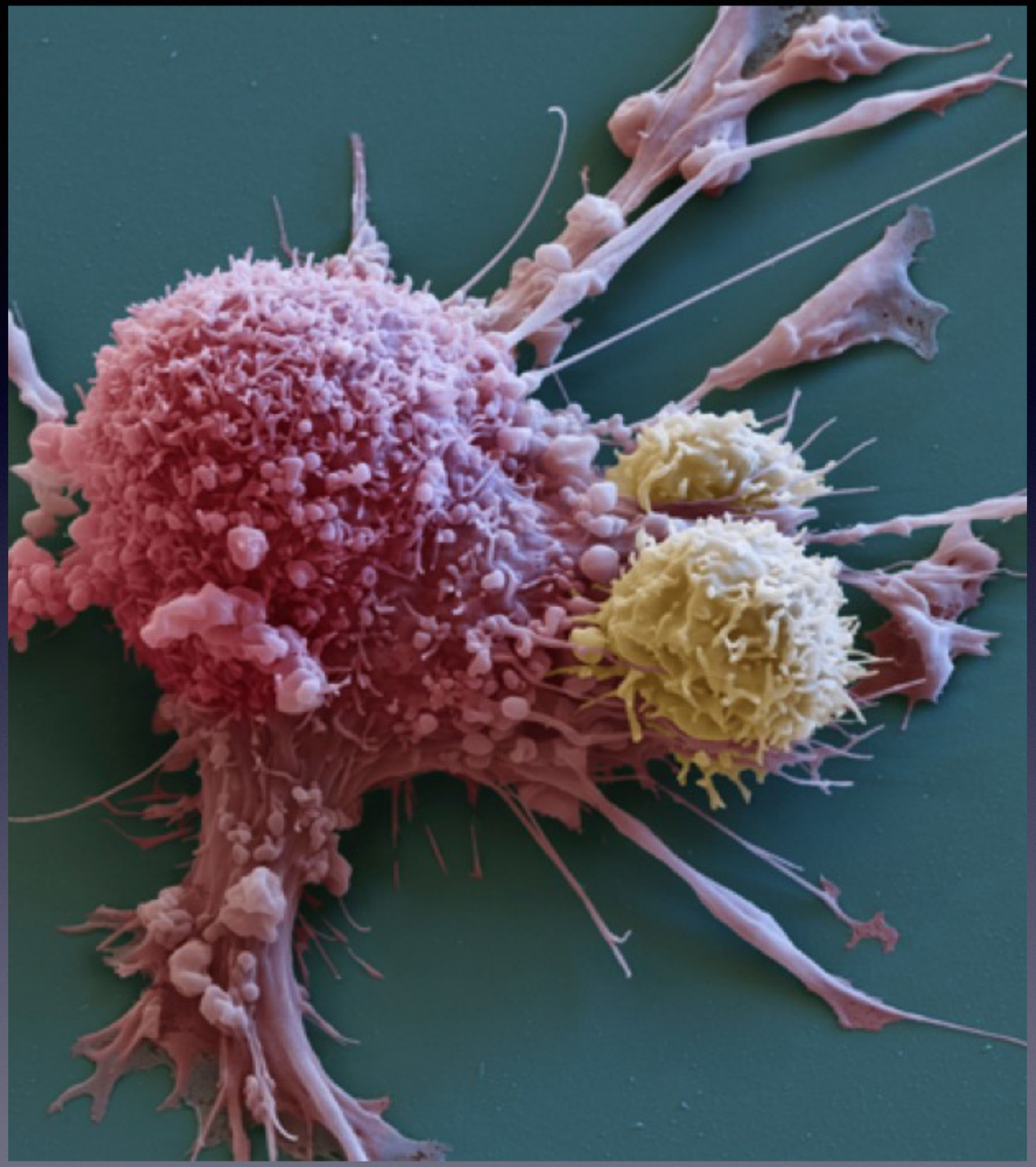


canonical uORFs start with
AUG

non-canonical uORFs start with
AAG, AUU, UUG, AGG, AUC, GUG, ACG, CUG, AUA

Hypothesis

uORF mutations may lead to a gene over-translation and consequent tumorigenesis

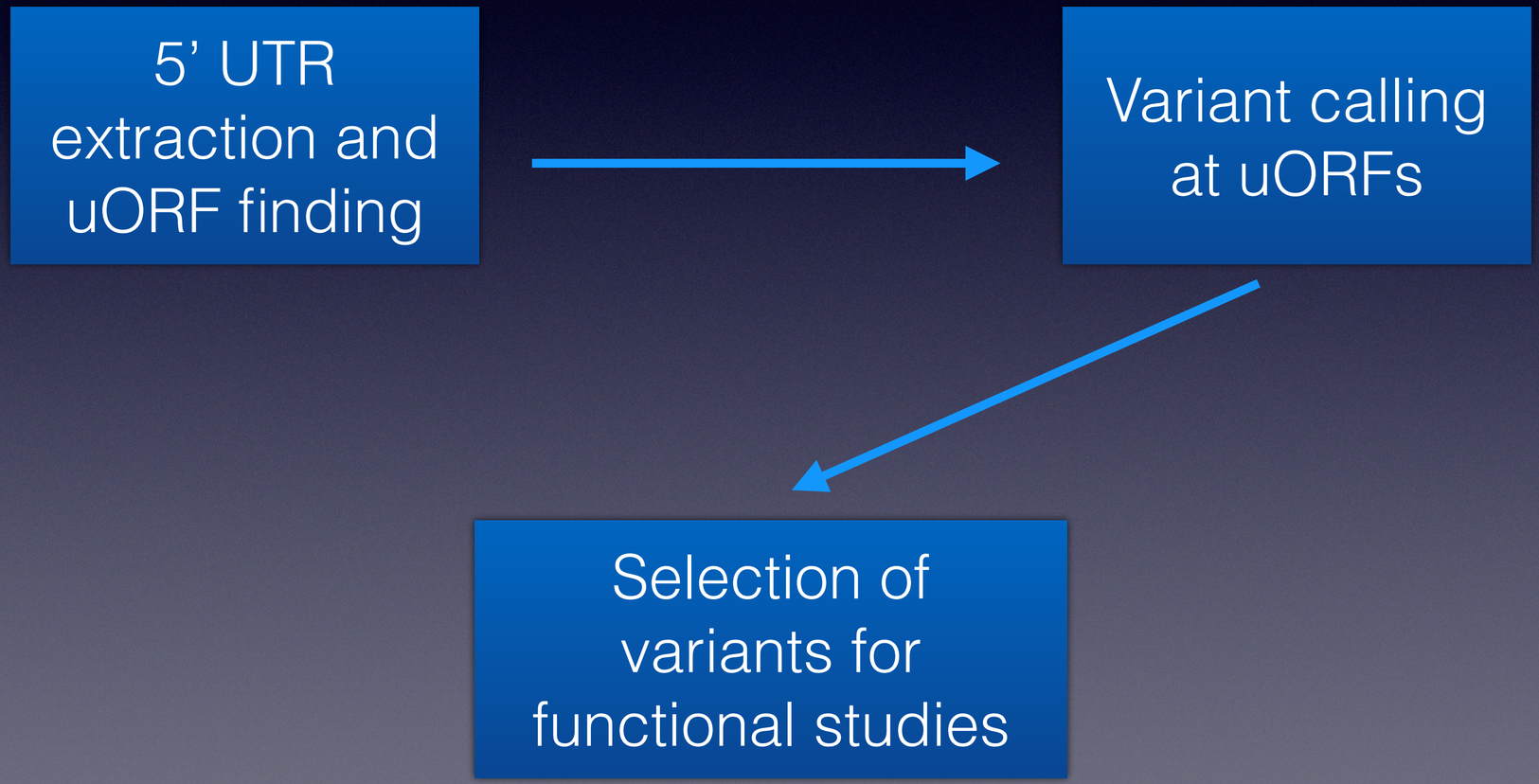


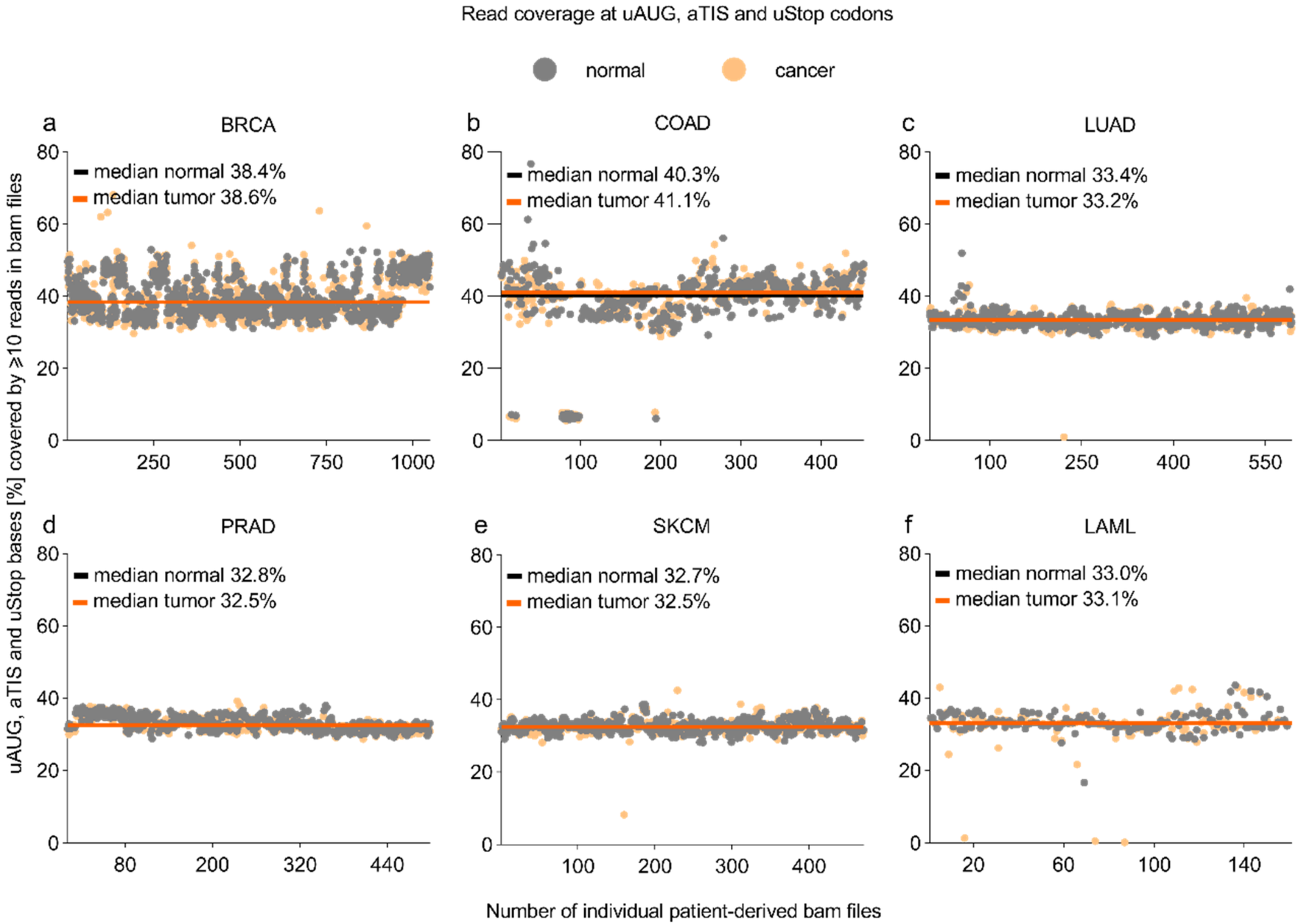
Credit: Eye of Science/SPL

Whole exome sequencing from TCGA

Type of cancer	Number of patients analyzed
Breast invasive carcinoma (BRCA)	1044
Colon adenocarcinoma (COAD)	433
Acute myeloid leukemia (LAML)	149
Lung adenocarcinoma (LUAD)	569
Prostate adenocarcinoma (PRAD)	498
Skin cutaneous melanoma (SKCM)	470
Total	3163

Study set up

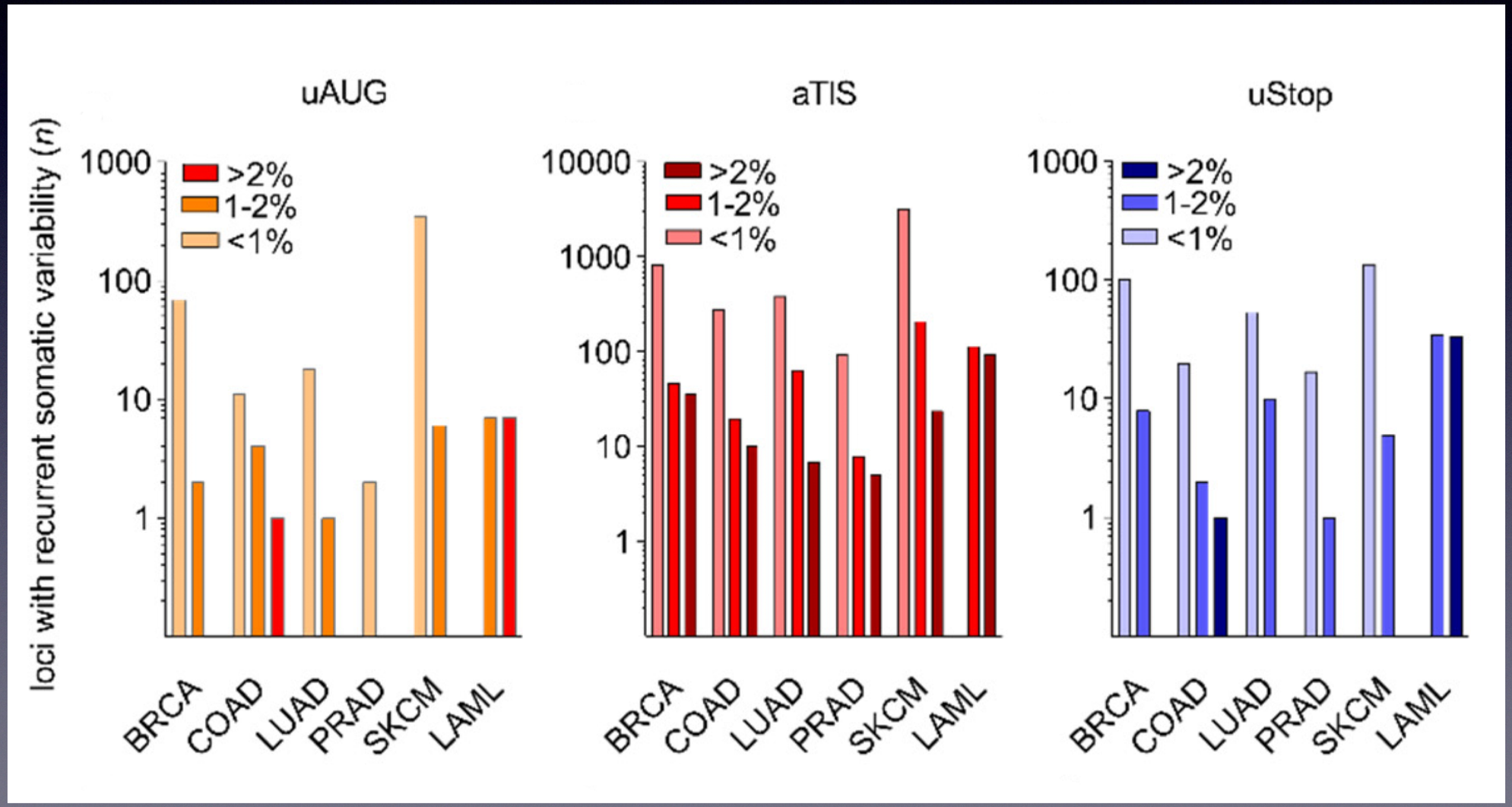

5’ UTR extraction and uORF finding
Variant calling at uORFs
Selection of variants for functional studies


Read coverage at uAUG, aTIS and uStop codons
normal
cancer
a
BRCA
median normal 38.4%
median tumor 38.6%
b
COAD
median normal 40.3%
median tumor 41.1%
c
LUAD
median normal 33.4%
median tumor 33.2%
d
PRAD
median normal 32.8%
median tumor 32.5%
e
SKCM
median normal 32.7%
median tumor 32.5%
f
LAML
median normal 33.0%
median tumor 33.1%
uAUG, aTIS and uStop bases [%] covered by ≥10 reads in bam files
Number of individual patient-derived bam files

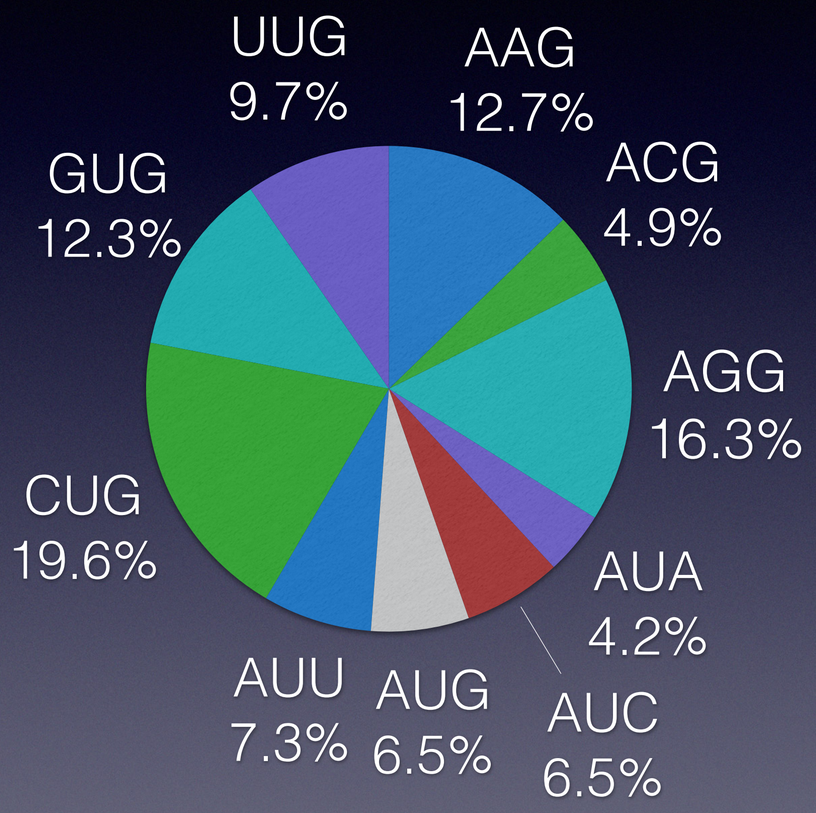
SNVs affecting uStart, and uStop codons

Type of cancer	Recurrent somatic SNVs	Affected patients	More frequent variants*	Affected patients	New SNVs
BRCA	1029	68.0%	94	35.1%	80
COAD	339	67.4%	37	40.4%	45
LUAD	494	68.0%	75	31.1%	56
PRAD	114	38.8%	14	20.9%	3
SKCM	3748	86.8%	241	61.7%	2112
LAML	258	75.2%	258	75.2%	72
Total	5277	66.5%	567	38.7%	2363

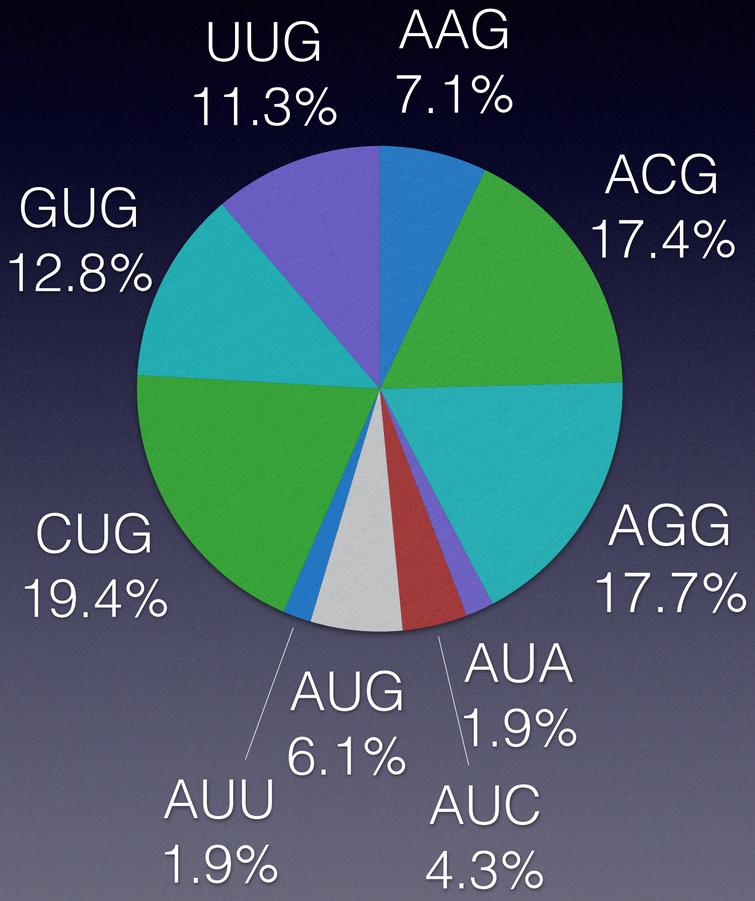
Recurrent somatic uORF-associated SNVs



Variants in Start codons

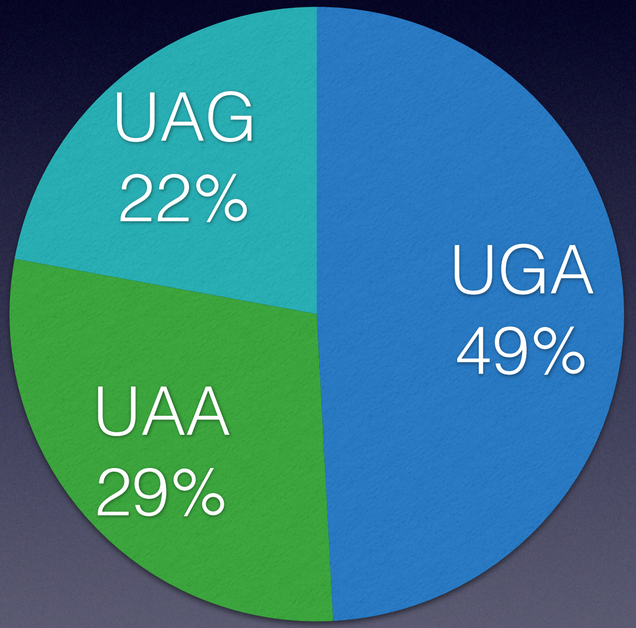


uORF start codons distribution

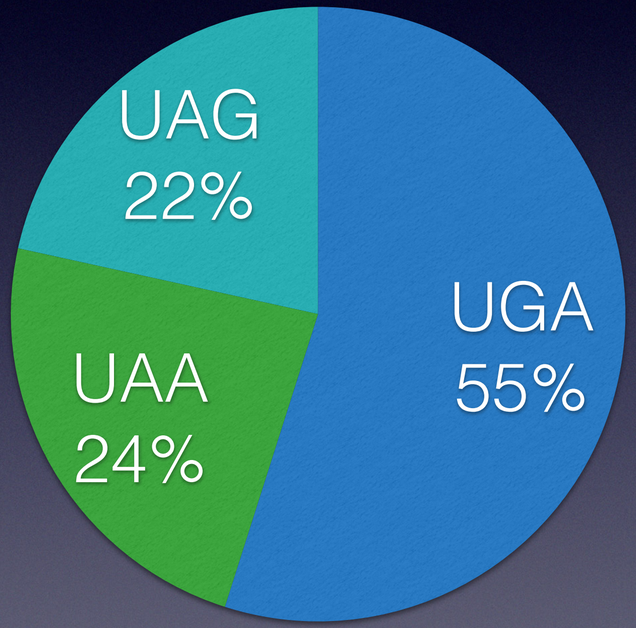


Variant distribution in different start codons

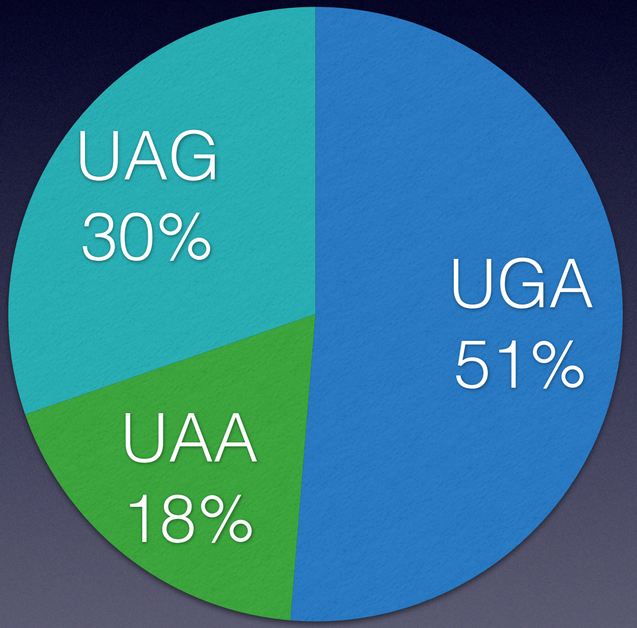
Variants in STOP codons



Human CDSs



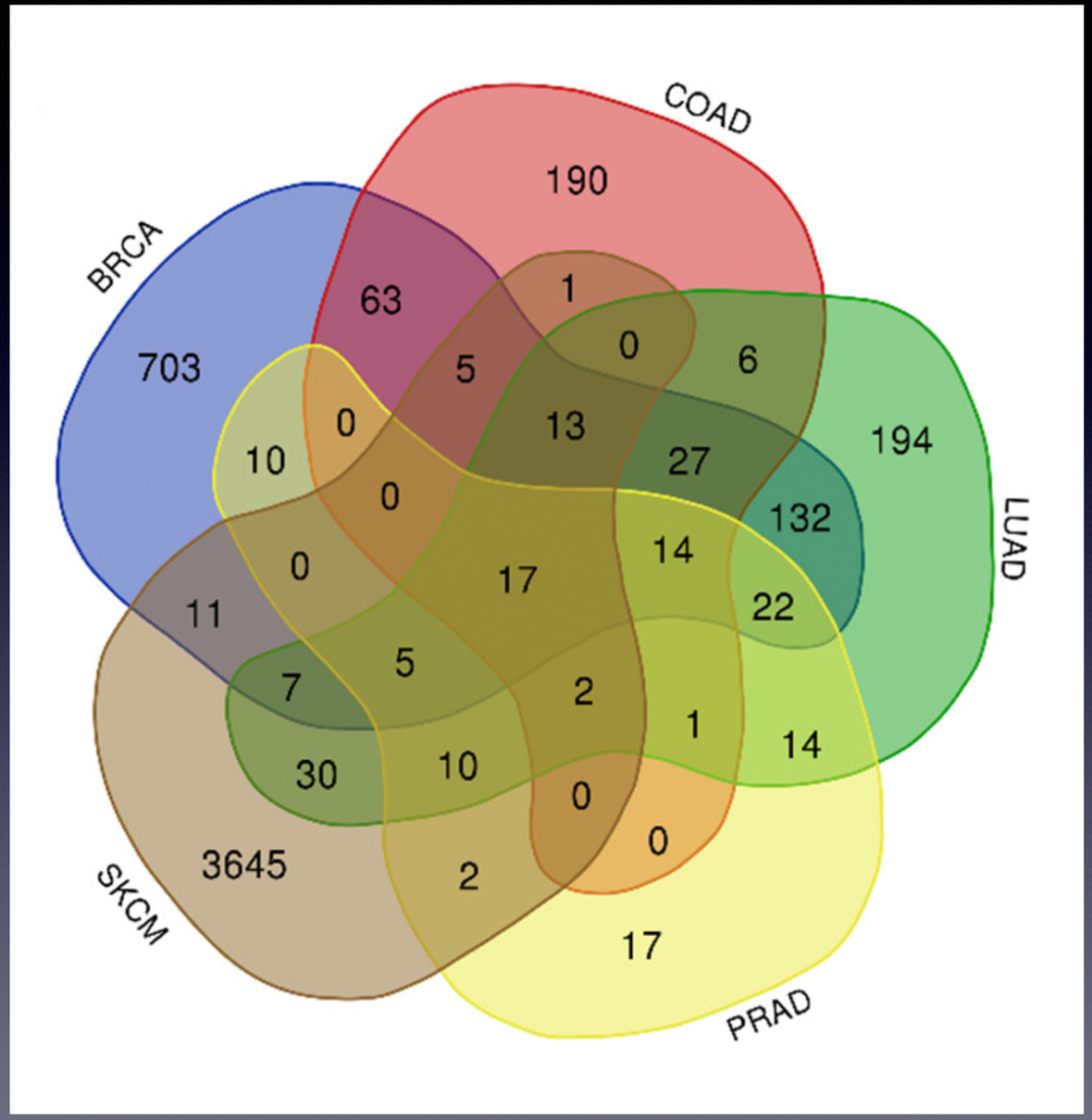
Human uORFs



Varianats in uORFs' stop codons

Jürgens et al. *Biomedicines* 2021, 9(6), 618

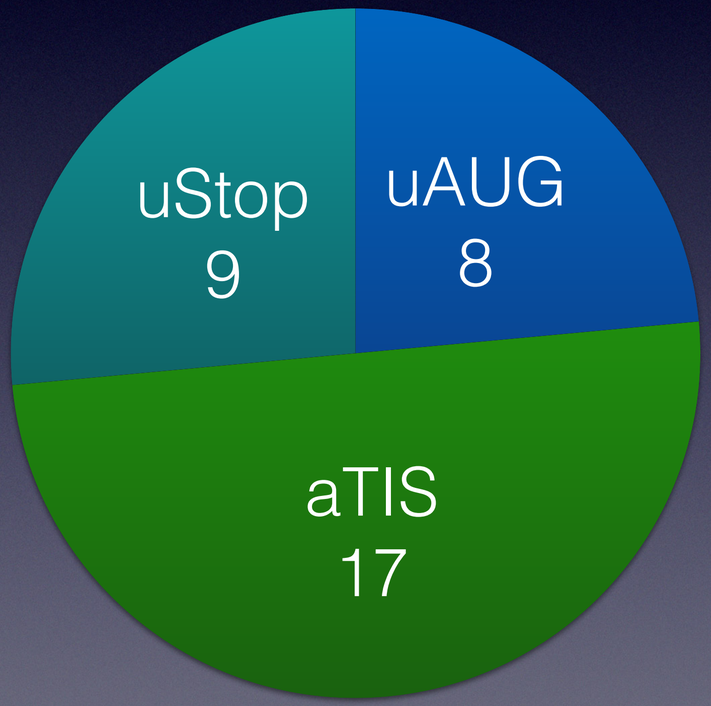
Vast majority of SNVs are cancer specific



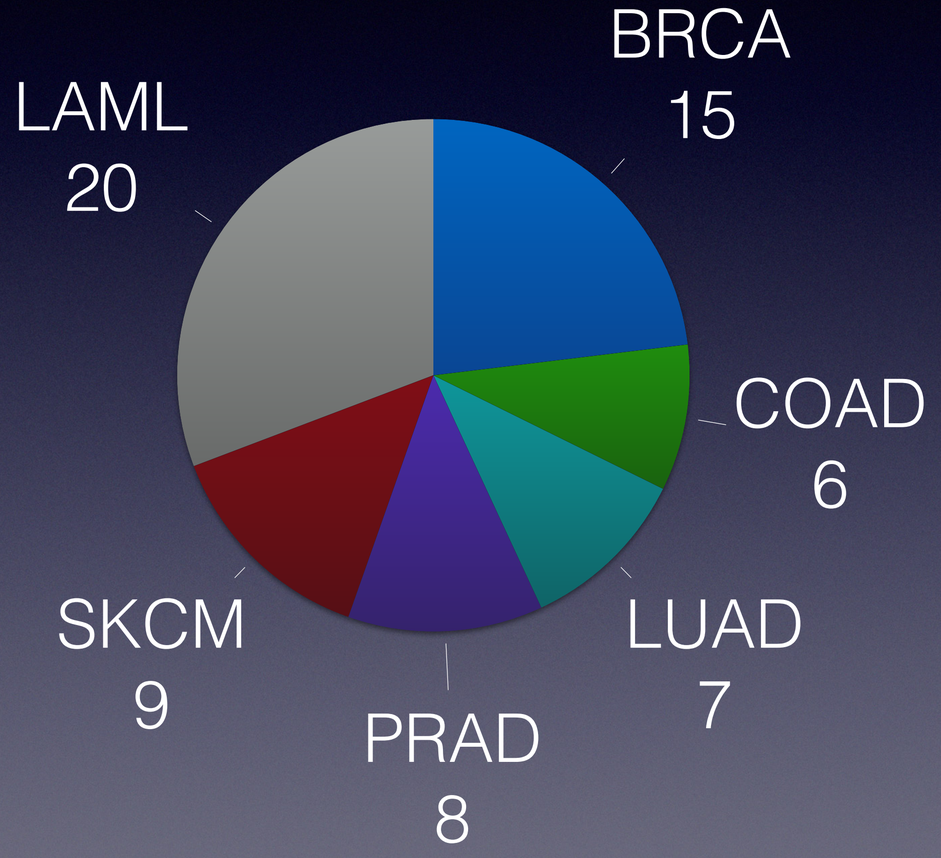
Jürgens et al. *Biomedicines* 2021, 9(6), 618

Wet lab analysis

Dataset selected for luciferase experiments (24 genes)

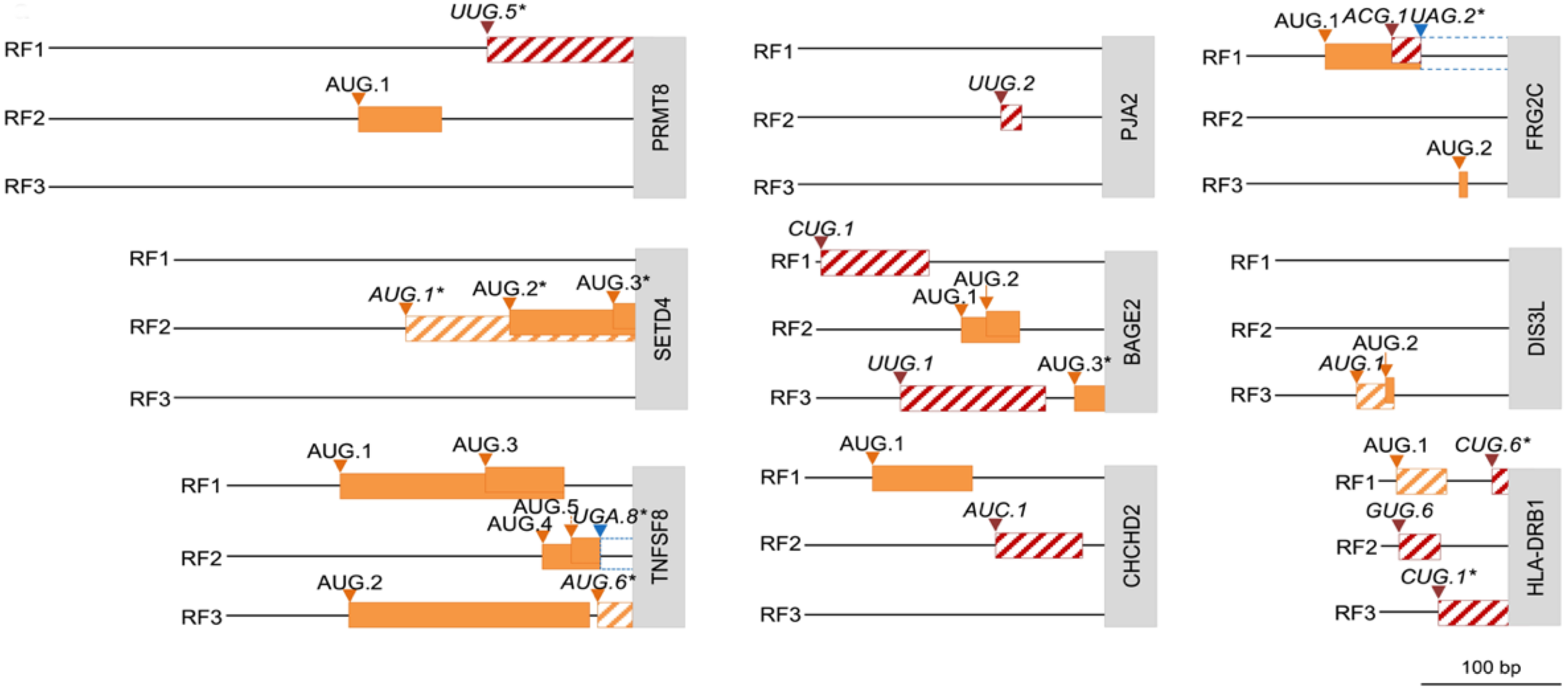


Mutation localization



Cancer distribution

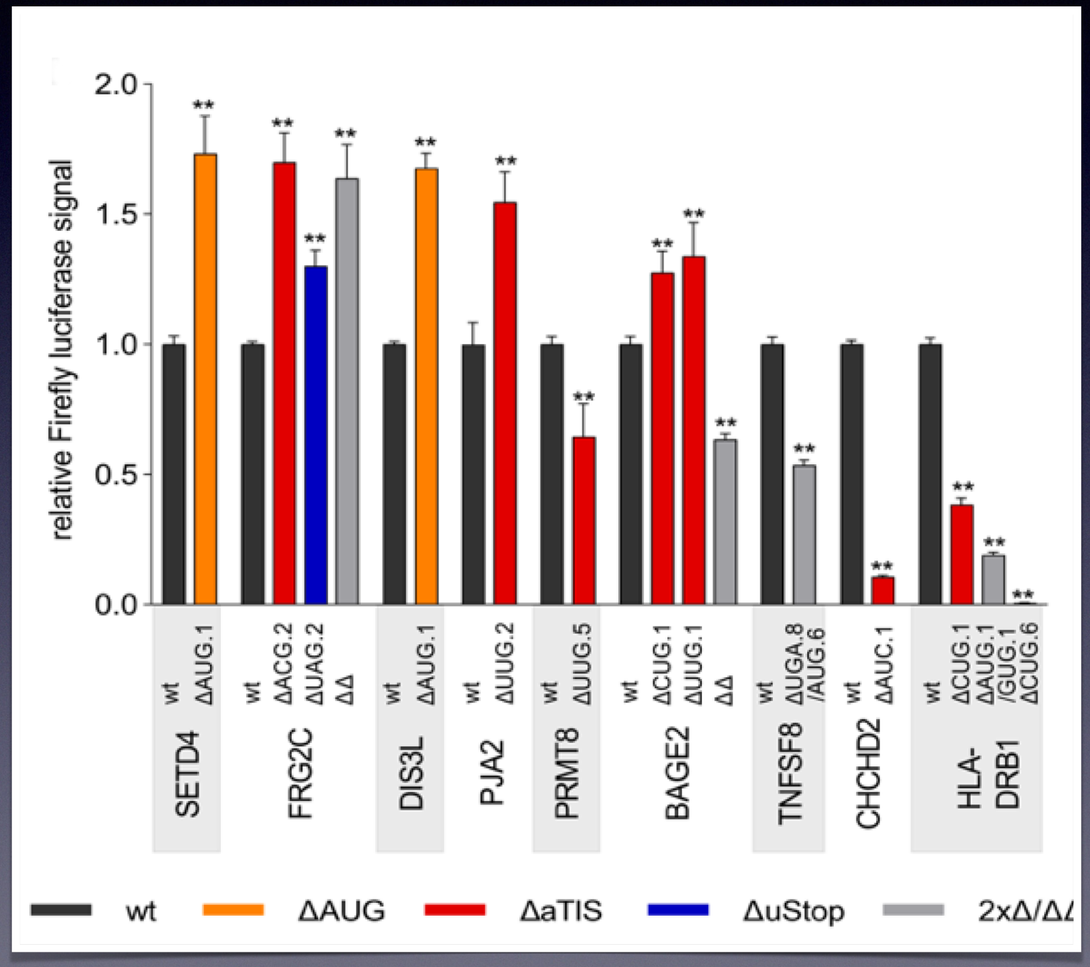
The selected genes possessed the whole spectrum of different uORFs



Nineteen SNVs caused significant alteration of translation level

Different effect for different transcripts and mutations

- 1.7-fold induction for the AUG.1 > UUG variant in SET Domain Containing 4 (SETD4)
- 0.006-fold repression caused by the CUG.6 > GUG variant Major Histocompatibility Complex Class II DR Beta 1 (HLA-DRB1)

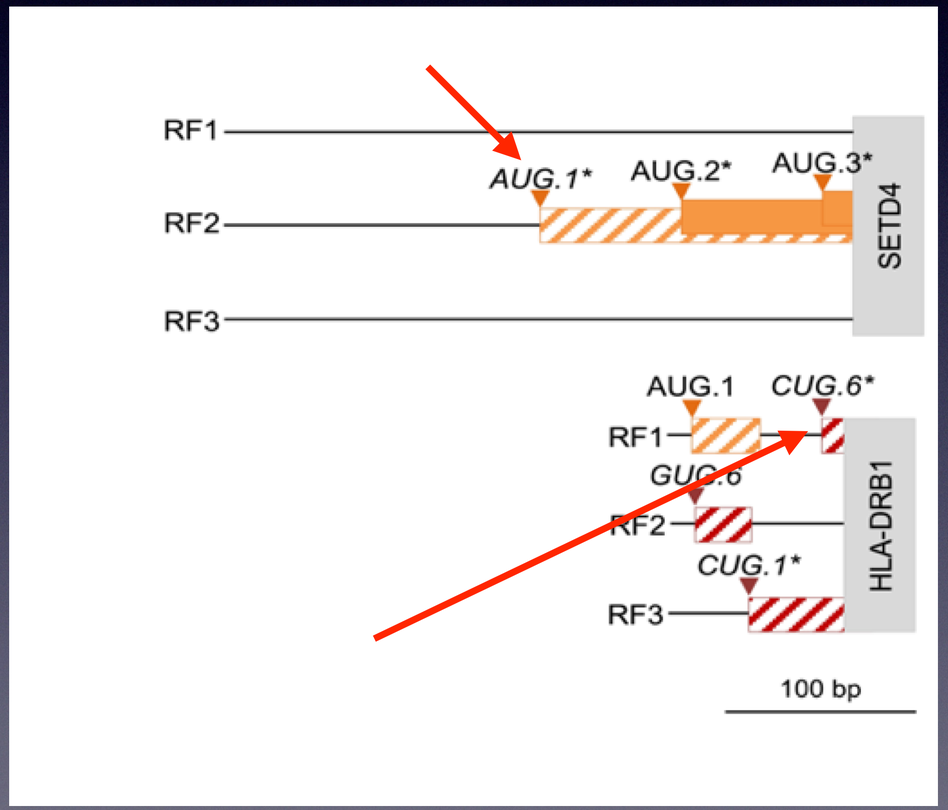


Jürgens et al. *Biomedicines* 2021, 9(6), 618

Nineteen SNVs caused significant alteration of translation level

Different effect for different transcripts and mutations

- 1.7-fold induction for the AUG.1 > UUG variant in SET Domain Containing 4 (SETD4)
- 0.006-fold repression caused by the CUG.6 > GUG variant Major Histocompatibility Complex Class II DR Beta 1 (HLA-DRB1)



Jürgens et al. *Biomedicines* 2021, 9(6), 618

Conclusions

- Our analysis revealed recurrent somatic uAUG, aTIS, or uStop mutations in a large proportion of cancer patients
- Individual uORF variants caused a wide range of activating and repressing effects on downstream translation
- We extended the catalog of translationally active uORFs by 19 somatic variants observed in patient-derived malignant tissues
- Besides the uORF-mediated impact on CDS translation, recent work of others and of our group revealed that a substantial fraction of canonical and non-canonical uORF start sites serve to initiate uORF-encoded peptides. Those uORF-peptides may form direct complexes with the associated main protein and can act in both, cis- and trans-regulatory ways.
- The read coverage analysis of current WES datasets at uORFs underlines that available WES data still cover less than half of all potential uORF-associated initiation and termination codons, leaving room for future genome-wide analyses.

Follow up study - whole genome sequencing data

Type of cancer	Number of patients	Analyzed
Breast invasive carcinoma (BRCA)	119	117
Colon adenocarcinoma (COAD)	118	
Acute myeloid leukemia (LAML)	50	
Lung adenocarcinoma (LUAD)	152	151
Prostate adenocarcinoma (PRAD)	122	122
Skin cutaneous melanoma (SKCM)	138	135
Total	699	525

uORFdb - a gene-centric database of upstream Open Reading Frames

Institute of Bioinformatics Münster

HOME TEACHING PUBLICATIONS THESES TOOLS INTERN

MYWWU UNIVERSITY CLINICS MUENSTER BIOLOGY

uORFdb

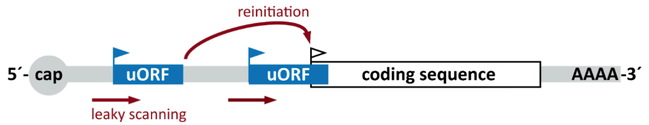
- Query
- About
- Documentation
- Statistics
- Publication Updates
- Downloads
- Cite
- Contact

Selection History

Welcome to uORFdb!

Upstream open reading frames (uORFs) are important regulatory elements in the 5´-transcript leader sequences (TLSs) of eukaryotic messenger RNAs. With uORFdb we aim to merge expert annotations of the latest uORF literature with the crucial sequence context of individual uORFs in humans and 12 other species. At present, the database contains more than 1,040 manually curated publications and more than 6.6 million uORFs of which over 2.4 million are located on human transcripts. For human uORFs, you can access information on genetic variability via external databases (e.g. dbSNP and ClinVar) and at the level of individual studies (e.g. WGS data of cancer patients). Currently, the database contains more than 129,000 somatic variant positions identified in WGS data of patient cohorts from breast, colon, blood, lung, prostate and skin cancer.

If you want to learn more about the database and how to use it, please click the "About" and "Documentation" buttons in the menu on the left or read its publication which includes a thorough example analysis.



Query the database by using the search bar below. The panels below the search bar define which fields will be queried. By default, you are searching in all searchable fields. Click on the blue arrow in one of the panels to go to your view of interest. Numbers show the number of hits for the searchable fields of a given view. For performance reasons, the number of hits per view is limited to 1000.

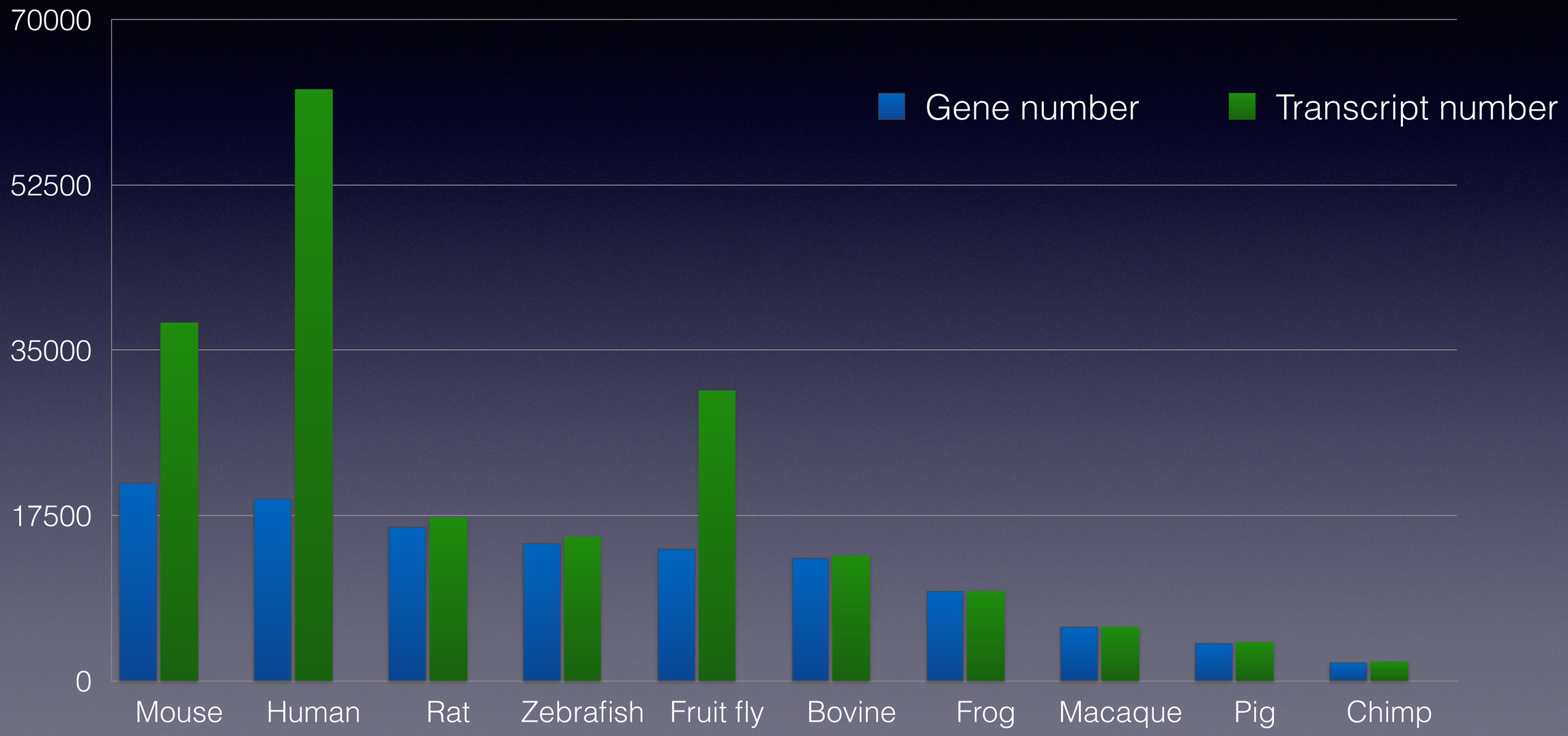
The database contains genomic sequences. Submit your nucleotide queries using "T", instead of "U".

Search (cannot be empty):

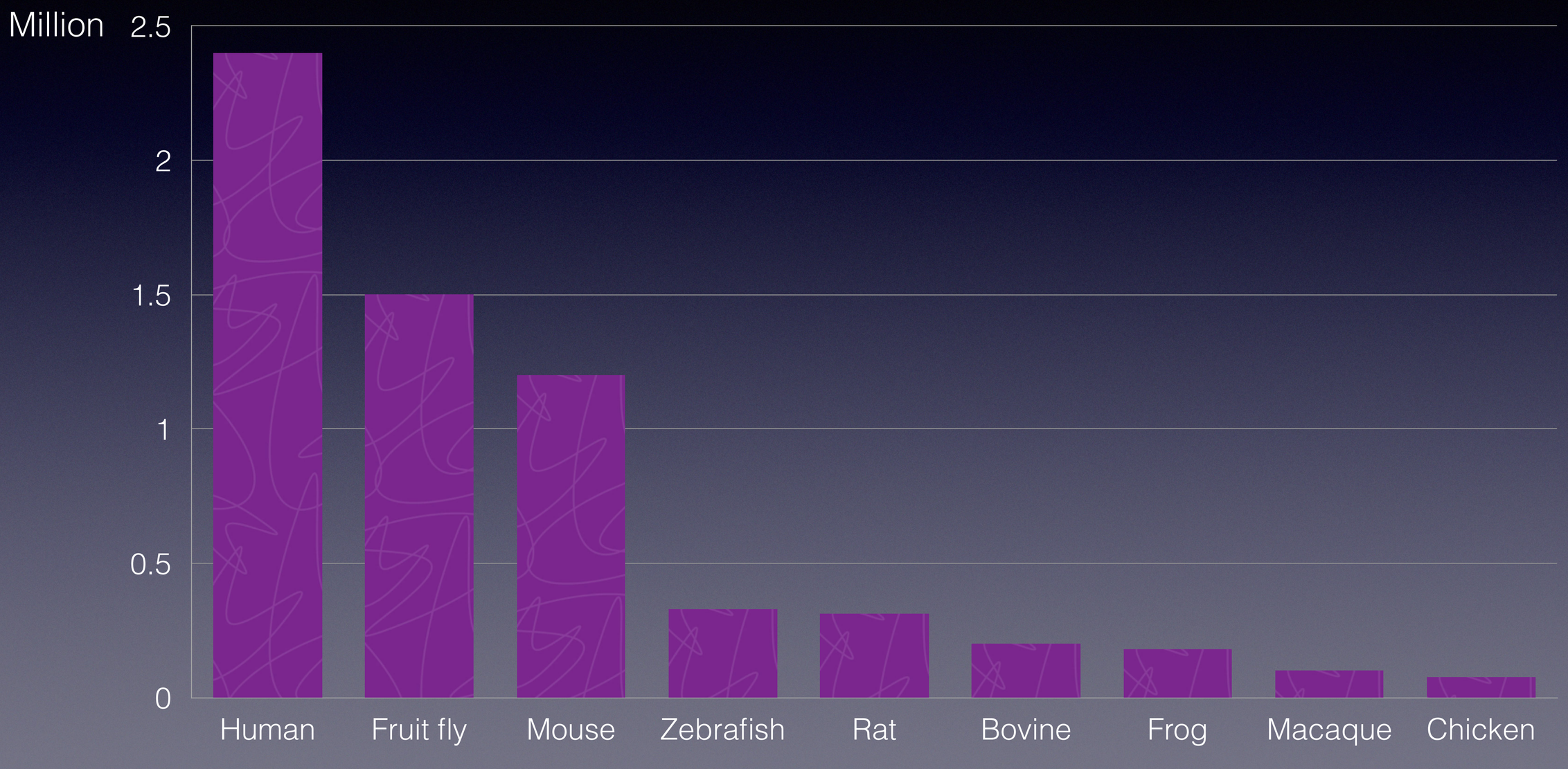
e.g. MIEF1; Kozak; NM_001304564.2 or rs730882029

- **Publications** [0]
  - PubMed ID
  - Authors
  - Taxa
  - Title
  - Gene name in paper
  - Gene symbols
- **Genes** [0]
  - Entrez gene ID
  - NCBI accession
  - Gene symbol
  - Gene names
  - Symbol aliases
- **Transcripts** [0]
  - NCBI ID
  - Kozak context
- **uORFs** [0]
  - Start codon
  - Stop codon
  - Kozak context
- **Variants** [0]
  - ClinVar ID
  - dbSNP ID

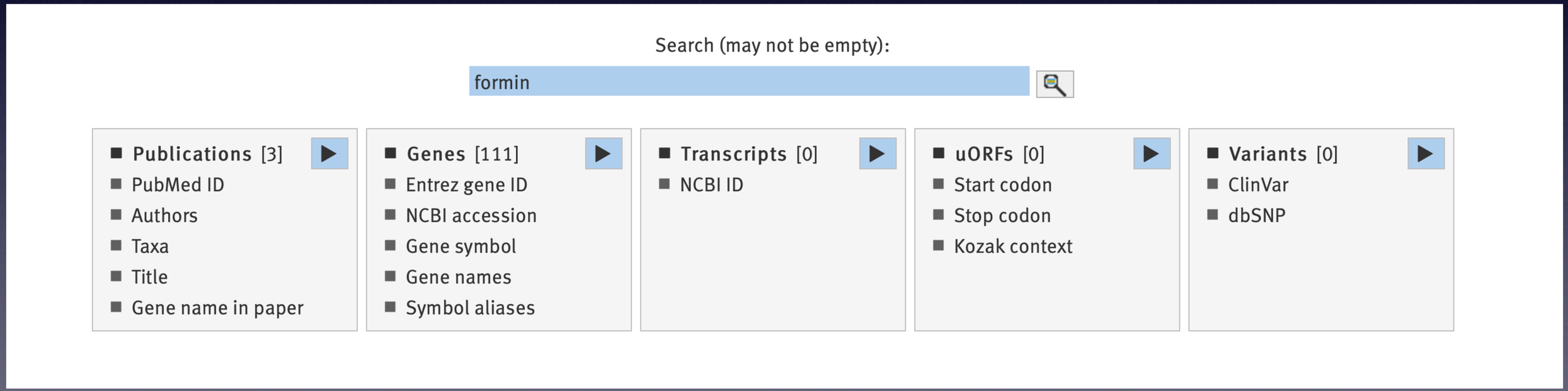
A Bit of Statistics



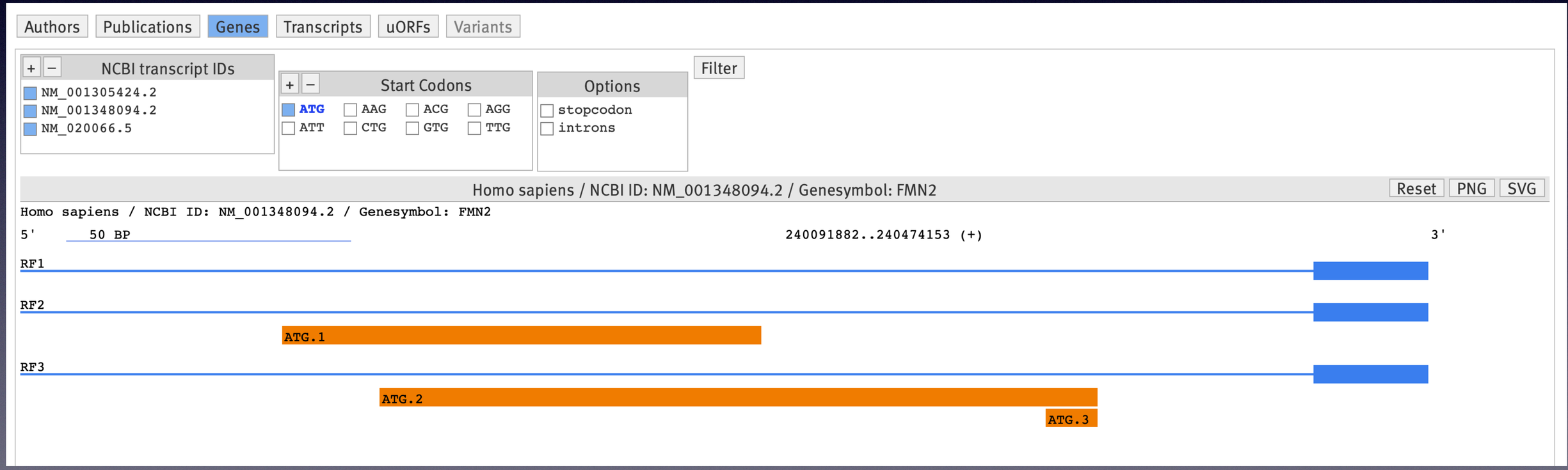
uORFs per taxon



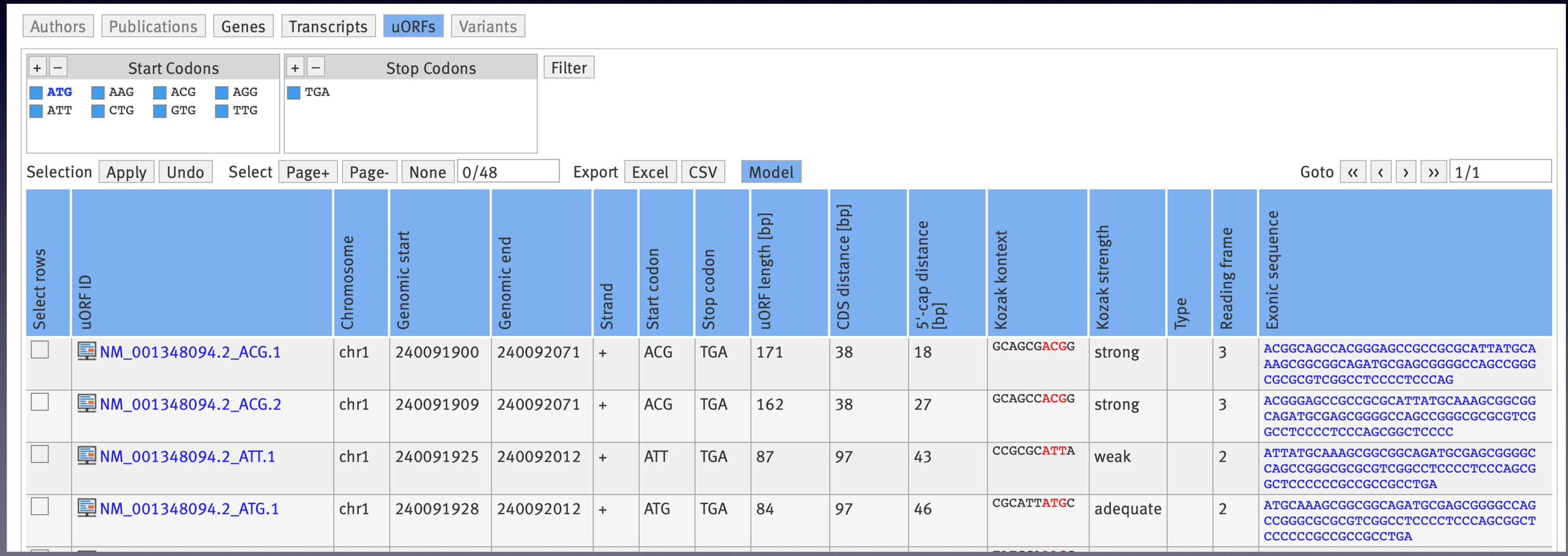
uORFdb - a gene-centric database of upstream open reading frames



uORFdb - a gene-centric database of upstream open reading frames



uORFdb - a gene-centric database of upstream open reading frames



Authors | Publications | Genes | Transcripts | uORFs | Variants

Start Codons: ATG, AAG, ACG, AGG, ATT, CTG, GTG, TTG

Stop Codons: TGA

Filter

Selection: Apply | Undo | Select: Page+ | Page- | None | 0/48 | Export: Excel | CSV | Model | Goto: « ‹ › » | 1/1

Select rows	uORF ID	Chromosome	Genomic start	Genomic end	Strand	Start codon	Stop codon	uORF length [bp]	CDS distance [bp]	5'-cap distance [bp]	Kozak kontext	Kozak strength	Type	Reading frame	Exonic sequence
☐	NM_001348094.2_ACG.1	chr1	240091900	240092071	+	ACG	TGA	171	38	18	GCAGCGACGG	strong		3	ACGGCAGCCACGGGAGCCGCCGCGCATTATGCA AAGCGGCGGCAGATGCGAGCGGGGCCAGCCGGG CGCGCGTCGGCCTCCCCTCCCAG
☐	NM_001348094.2_ACG.2	chr1	240091909	240092071	+	ACG	TGA	162	38	27	GCAGCCACGG	strong		3	ACGGGAGCCGCCGCGCATTATGCAAAGCGGCGG CAGATGCGAGCGGGGCCAGCCGGGCGCGCGTCG GCCTCCCCTCCCAGCGGCTCCCC
☐	NM_001348094.2_ATT.1	chr1	240091925	240092012	+	ATT	TGA	87	97	43	CCGCGCATTA	weak		2	ATTATGCAAAGCGGCGGCAGATGCGAGCGGGGC CAGCCGGGCGCGCGTCGGCCTCCCCTCCCAGCG GCTCCCCCCGCCGCCGCCTGA
☐	NM_001348094.2_ATG.1	chr1	240091928	240092012	+	ATG	TGA	84	97	46	CGCATTATGC	adequate		2	ATGCAAAGCGGCGGCAGATGCGAGCGGGGCCAG CCGGGCGCGCGTCGGCCTCCCCTCCCAGCGGCT CCCCCCGCCGCCGCCTGA

uORFdb - a gene-centric database of upstream open reading frames

uORF ID	NM_001348094.2_ATG.2
Chromosome	chr1
Genomic start	240091945
Genomic end	240092071
Strand	+
Start codon	ATG
Stop codon	TGA
uORF length [bp]	126
CDS distance [bp]	38
5'-cap distance [bp]	63
Kozak kontext	CGGCAGATGC
Kozak strength	weak
Type	
Reading frame	3
Exonic sequence	ATGCGAGCGGGGCCAGCCGGGCGCGCGTCGGCCTCCCCTCCCAGCGGCTCCCCCCGCCGCCGCCTGACTCTCCCGGGAGACTCCCTAGGCCCGGACCTGG GGCCGAGGAGGGCCGGGATGGCCTGA
Exonic amino sequence	MRAGPAGRASASPPSGSPRRRLTLPGDSLGPDLGPRRAGMA*
Exon variants in dbSNP	»
Exon variants in Clinvar	»

Acknowledgments

Lynn Ogoniak

Felix Manske

Norbert Grundmann

Lara Jürgens

Klaus Wethmar