**Figures:**

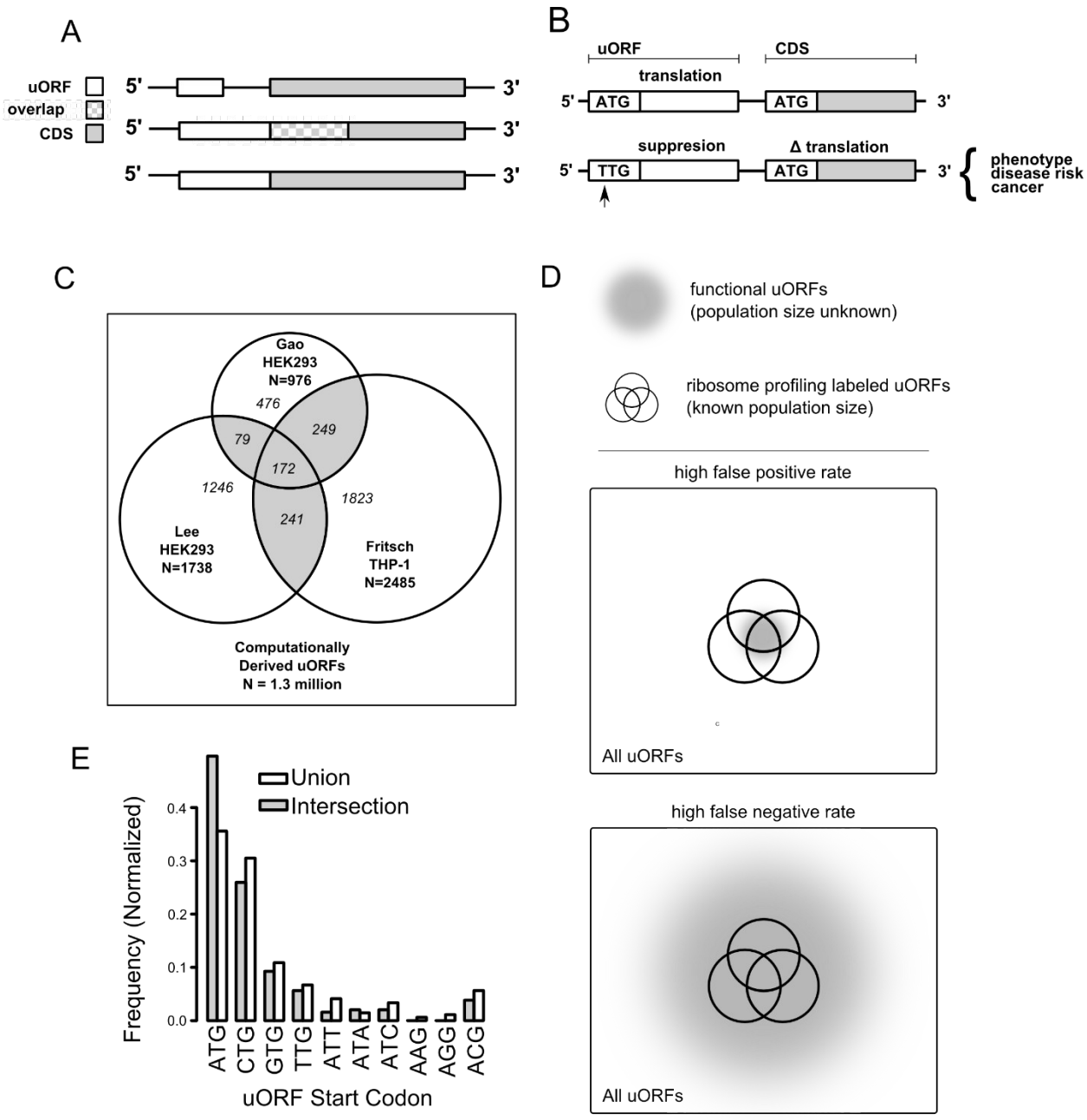



**Figure 1:**

**A. *The structure of upstream open reading frames.*** The stop codon for a uORF may be located before the CDS start codon. It may also be located within the CDS, if the uORF is frame-shifted relative to the CDS (upper and middle, respectively). An open reading frame may also utilize the same stop codon as the CDS, such that the ORF acts as a 5’ extension of the CDS.

**B.** ***The effect of mutation or variation on upstream open reading frames.*** The creation or destruction of an upstream open-reading, may result in downstream effect on the rate of translation of the coding sequence. Change in degree of translation of the coding sequence, may in turn, result in change in phenotype, and disease risk.

**C.** ***The global space of upstream open reading frames, and within that space, the subset of ribosome profiling identified uORFs.*** Ribosome studies by Fritsch et al., Lee et al., and Gao et al., are interpreted as a Venn diagram. Pair-wise and three-way intersections between these experiments are highlighted, as these uORFs consititute our gold standard positive set. The universe of all possible uORFs, is derived from the GENCODE annotation, and numbers 1.3 million. Ribosome profiling positive uORFs, are used to identify a population of predicted positive uORFs, among the set of 1.3 million computationally derived uORFs.

***D. The sensitivity and specificity of ribosome profiling, for identifying upstream open reading frames.*** Ribosome profiling studies identify a known number of translated upstream open reading frames. However, it is unknown how this number compares, to the total number of translated upstream open reading frames. It is possible that ribosome profiling studies have a high false positive rate (top), or a high false negative rate (bottom). We make the assumption that ribosome profiling studies have a high false negative rate for identifying translated upstream open reading frames (high specificity).

***E. The frequence of uORF ATG start codons, and near-cognate start codons, from ribosome profiling experiments.*** Frequency is given both for the overall frequency of start codons (union), and uORFs that are translated in more than one experiment (intersection).

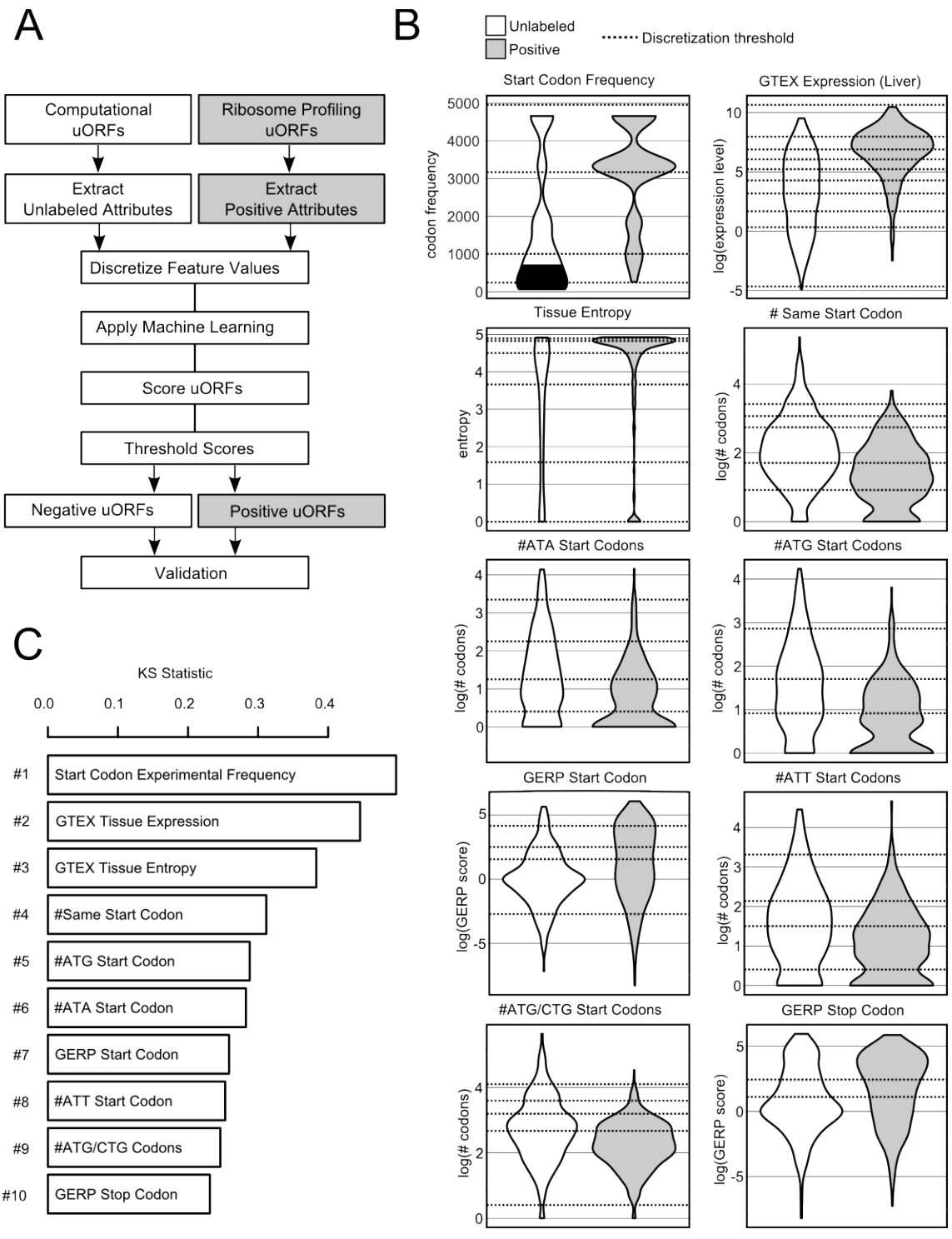

A
Computational uORFs
Ribosome Profiling uORFs
Extract Unlabeled Attributes
Extract Positive Attributes
Discretize Feature Values
Apply Machine Learning
Score uORFs
Threshold Scores
Negative uORFs
Positive uORFs
Validation
B
Unlabeled
Positive
Discretization threshold
Start Codon Frequency
codon frequency
GTEX Expression (Liver)
log(expression level)
Tissue Entropy
entropy
# Same Start Codon
log(# codons)
#ATA Start Codons
#ATG Start Codons
GERP Start Codon
log(GERP score)
#ATT Start Codons
#ATG/CTG Start Codons
GERP Stop Codon
C
KS Statistic
0.0
0.1
0.2
0.3
0.4
#1 Start Codon Experimental Frequency
#2 GTEX Tissue Expression
#3 GTEX Tissue Entropy
#4 #Same Start Codon
#5 #ATG Start Codon
#6 #ATA Start Codon
#7 GERP Start Codon
#8 #ATT Start Codon
#9 #ATG/CTG Codons
#10 GERP Stop Codon

**Figure 2:**

**A. *Methodology for distinguishing positive from unlabeled uORFs.*** Computationally derived uORFs and ribosome profiling identified uORFs, represent unlabeled and positive examples respectively. Attributes of these positive and unlabeled uORFs are extracted. The positive and unlabeled examples are used to train a machine learning algorithm. The machine learning algorithm assigns a score all computationally derived uORFs. A threshold on this score, yields positive and negative uORFs.
**B. *Distributions of attributes for positive and unlabeled uORFs.*** The attributes of uORF, are used to distinguish positive from unlabeled uORFs. Attributes like the sequence conservation (GERP score), and tissue mRNA expression (GTEX Expression - Liver), have different continuous distributions for the unlabeled uORFs, compared to the positive uORFs. These continuous distributions, can be discretized and optimized for machine learning, using the minimum description length principle (MDLP) binning algorithm. Horizontal lines on the plot correspond to these binning intervals. The 10 attributes with the greatest difference in distribution (largest kolmogorov smirnov statistic) between positive and unlabeled uORFs are shown.
***C. Upstream open reading frame attributes as classifiers, ranked.*** The Kolmogorov Smirnov (KS) test provides an index for distinction between positive and unlabeled attributes. Attributes are ranked, according to the difference in distribution between positive and unlabeled attributes, using the KS statistic. The KS statistic thus provides an index for the utility of attributes in distinguishing between positive and unlabeled uORFs.

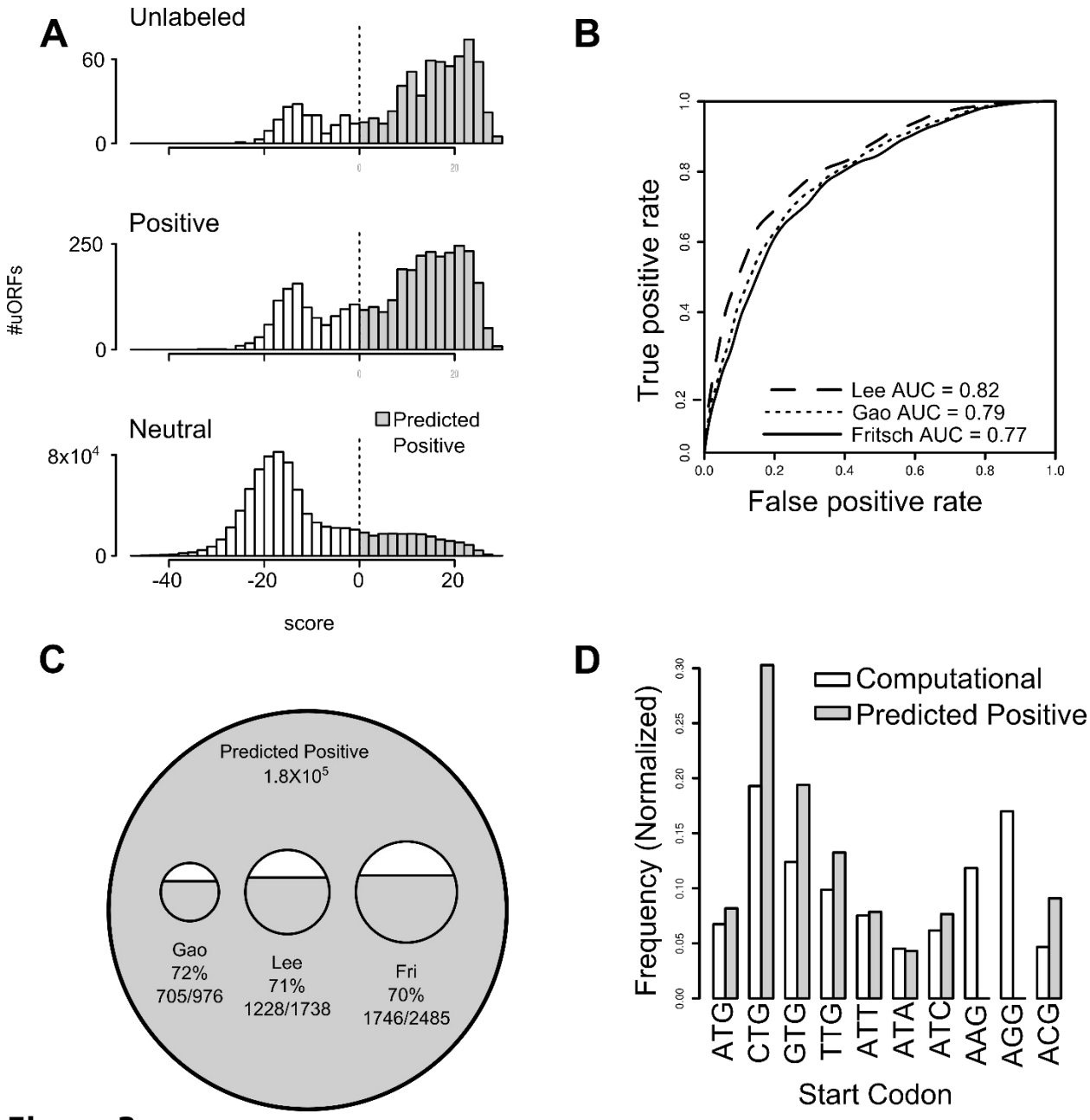



**Figure 3:**

***A. Score distributions for upstream open reading frames, according to category determined via ribosome profiling.*** Score distributions for [a] and unlabeled uORFs, that are identified computationally, through a comprehensive scan of the GENCODE annotation, but are not found translated in any ribosome profiling experiment (top), [b] positive ribosome profiling uORFs, that are positively identified in two or more ribosome profiling experiments (middle), and [c] neutral ribosome profiling uORFs, that are identified in only a single ribosome profiling experiment, and are so witheld from both the positive and the unlabeled sets (bottom).

***B. ROC curves gauge performance of the machine learning algorithm.*** The machine learning algorithm was trained on two of the three ribosome profiling data set, and then used to extract the third data set from the unlabeled data set. The ROC curve is shown for each of the three combinations 1. Train Lee et al. and Fritsch et al. – extract Gao et al. (AUC = 0.79), 2. Train Lee et al. and Gao et al. – extract Fritsch et al. (AUC = 0.77). 3. Train Fristch et al. and Gao et al. - extract Lee et al. (AUC = 0.82).

***C. Positively identifed uORFs from the computational set, and ribosome profiling experiments.*** Of the computationally derived uORFs extracted from the GENCODE annotation, approximately 180 000 are predicted as active upstream open reading frames. These are the upstream open reading frames that are predicted to undergo translation. This large set, includes 72% of the uORFs identified in the ribosome profiling experiment of Gao et al., 71% of the uORFs identified in the experiment of Lee et al., and 70% of the uORFs identified in the experiment of Fritsch et al.

***D. The frequence of uORF ATG start codons, and near-cognate start codons, for predicted positive upstream open reading frames.*** Frequency is given both for the overall frequency of computationally derived uORFs from GENCODE (computational), and for the subset of computationally derived uORFs that are predicted to be translated (predicted positive).

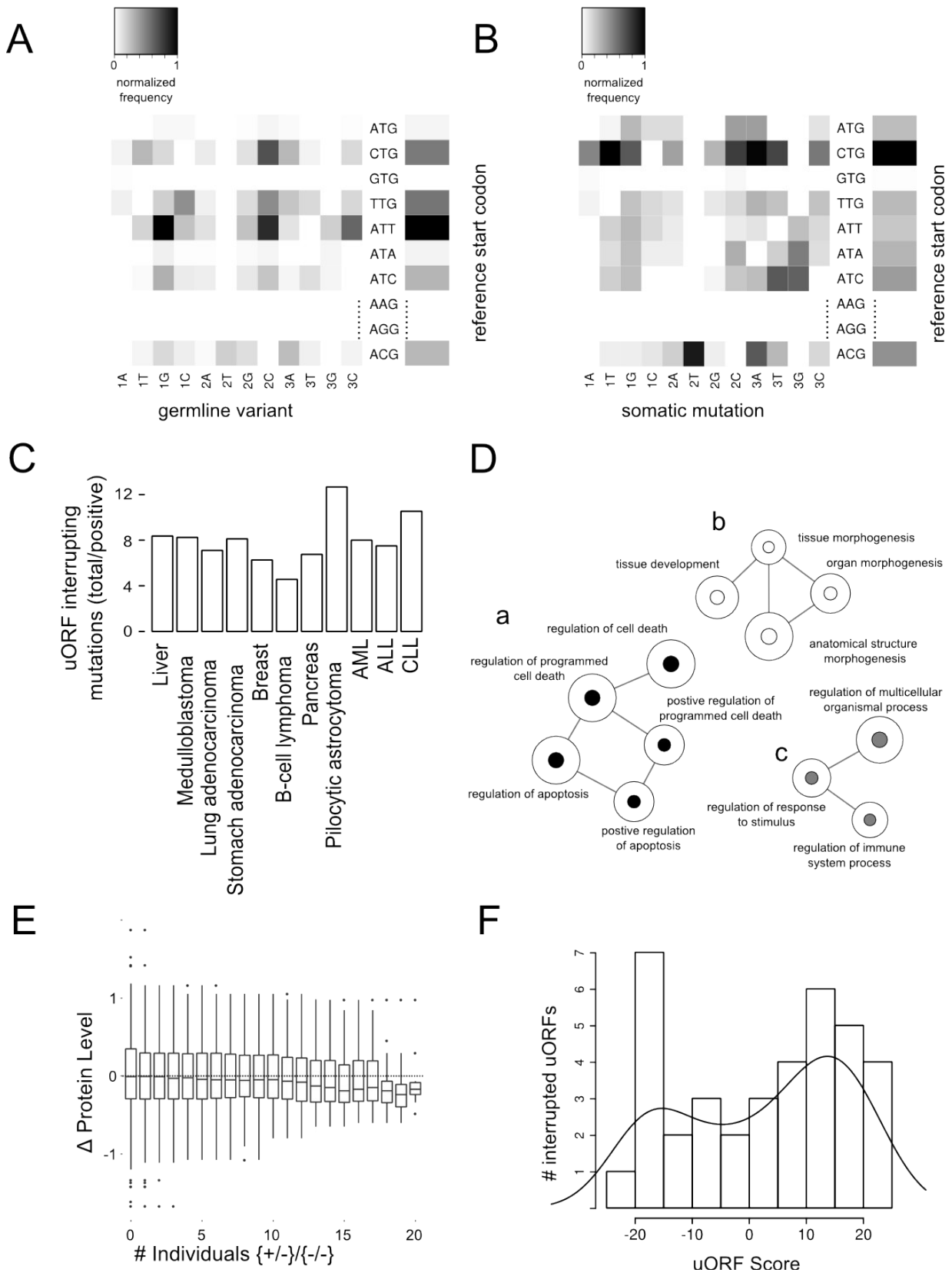

A
normalized frequency
0
1
ATG
CTG
GTG
TTG
ATT
ATA
ATC
AAG
AGG
ACG
reference start codon
1A 1T 1G 1C 2A 2T 2G 2C 3A 3T 3G 3C
germline variant
B
normalized frequency
somatic mutation
C
uORF interrupting mutations (total/positive)
12
8
4
0
Liver
Medulloblastoma
Lung adenocarcinoma
Stomach adenocarcinoma
Breast
B-cell lymphoma
Pancreas
Pilocytic astrocytoma
AML
ALL
CLL
D
a
b
c
tissue morphogenesis
tissue development
organ morphogenesis
anatomical structure morphogenesis
regulation of cell death
regulation of programmed cell death
postive regulation of programmed cell death
regulation of apoptosis
postive regulation of apoptosis
regulation of multicellular organismal process
regulation of response to stimulus
regulation of immune system process
E
Δ Protein Level
1
0
-1
0
5
10
15
20
# Individuals {+/-}/{-/-}
F
# interrupted uORFs
1
2
3
4
5
6
7
-20
-10
0
10
20
uORF Score

**Figure 4:**

***A: Density matrix, showing the distribution of 1000 genomes variants, interrupting postively scored uORF start codons.*** The vertical axis displays the reference start codon, the horizontal axis shows the interrupting variant (position – 1,2,3 – and codon – A,T,G,C).
***B: Density matrix, showing the distribution somatic mutations found in tumor samples (Alexandrov et al.), interrupting positively scored uORF start codons.*** The vertical axis displays the reference start codon, the horizontal axis shows the interrupting variant (position – 1,2,3 – and codon – A,T,G,C).
***C: Ratio of all uORFs interrupted by start-codon destroying mutants (Alexandrov et al.), to positively scored uORFs interrupted by start codon destroying mutants, according to cancer type.***
***D: GO/PANTHER terms, for statistically overrepresented genes with uORF start codons interrupted by somatic variants in tumor samples (Alexandrov et al.).*** The size of each node, corresponds to the number of uORFs associated that GO term. Thresholds were established to eliminate relatively common GO terms (>1250 associated uORFs), and relatively uncommon GO terms (<250 associated uORFs). This was done, in order to produce a network structure that is neither too general, nor too specific. 3 principle networks emerge a) tissue morphogenesis b) immune function c) apoptosis. Networks were developed using the statistical package BiNGO, and include adjustment for multiple testing.
***E: The standardized change in protein level for a given gene, between wild type individuals, and individuals with uORF start codon interrupting variants.*** This difference in protein level is shown for different ratios of variant possessing individuals (+/-, -/-) to wild-type individuals (+/+). Larger numbers of individuals with the variant allele, allow for larger statistical power, in calculating the effect of the variant on protein level.
***F: rQTLs (Battle et al. 2015) interrupting uORF start codons, according to the score of the corresponding uORF.*** rQTLs are more likely to be associated with a positive scoring uORF.

**Title:**

**A comprehensive catalogue of predicted functional upstream open reading frames.**

**Patrick McGillivray, Russell Ault, Mayur Pawashe, Rob Kitchen, Suganthi Balasubramanian, Mark Gerstein**

## Abstract

Upstream open reading frames (uORFs), are associated with translational regulation of downstream coding sequences. The translation of a uORF latent in an mRNA transcript, is thought to modify the translation of coding sequences in that same transcript, by modifying ribosome localization. Not all uORFs are thought to be active in such a process. It represents a challenge to estimate the impact and scope of the role uORFs play in regulation of translation.

We use the GENCODE annotation of the human genome, to circumscribe the universe of all possible translated uORFs. This universe includes over one million unique uORFs. We compare patterns of structure in these uORFs, to the structure of uORFs labeled as translated in experiment. This comparison allows us to catalog a population of uORFs that likely undergo translation. It is a substantially larger catalog of uORFs, than has previously been associated with active translation. It suggests the translation of uORFs, is a widespread phenomenon, with considerable impact on the translational landscape.

Our catalog of uORFs, allows researchers to test their hypotheses regarding the role of upstream open reading frames, in health and disease.

## Intro

Upstream open reading frames (uORFs) consist of a start codon in the 5' untranslated region of a gene (UTR), and an associated stop codon appearing before the stop codon of the main coding sequence (CDS). The uORF may begin and end before the main gene coding sequence. Alternatively, if the upstream reading frame is out of frame with the CDS, it may overlap with the CDS [Figure 1.A]. uORFs are latent in mRNA transcripts, and may undergo partial or complete translation.

Initial survey of the human genome, identified uORFs contained in approximately 10% of mRNA transcripts (1). More recent analyses broaden estimates of prevalence, with identification of uORFs in association with nearly half of all mRNA transcripts (2). The discovery that many uORFs utilize near-cognate start codons, rather than the canonical ATG start codon, has broadened estimates of uORF prevalence still further (3–6).

Study of uORF translation and function, was historically limited to the experimental evaluation of individual uORFs (7,8), with no genome-scale approach to identifying translated uORFs. The advent of ribosome profiling studies, has allowed for the identification of a large population of uORFs known to undergo translation (4,9,10). Ribosome profiling studies that arrest the ribosome at translation initiation, allow for the identification of

translation initiation sites to within a few nucleotides. This mapping of translation initiation is sufficient for association between ribosomes and particular start codons and reading frames (11–13).

At the same time as ribosome profiling studies have allowed for large-scale identification of upstream open reading frames, there has been expansion in knowledge of the functional role of uORFs. Upstream open reading frames, have generally been thought to suppress translation of downstream genes (8,14–18). The molecular mechanisms for modification of translation are varied, and include leaky-scanning of uORFs by ribosomes, translation reinitiation subsequent to uORF translation, and ribosome-stalling on uORFs. These mechanisms have been uncovered in some detail (3,19,20). Apart from increases and decreases in a single protein product, differential translation of multiple protein products may occur in consequence to a uORF (21). There may even be additional direct effect of translated products between uORF and CDS, as has been observed in dual-coding genes (22)

From these studies, it is important to note, that a uORF may increase translation of the downstream CDS or decrease translation of the downstream CDS, according to genomic and epigenomic context. Related to a differential effect of uORFs on CDS translation, depending on context, the study of translation in stress conditions, has revealed a differential function for uORFs in stressed cells, compared with non-stressed controls (23–28).

Interest in the study of the function of upstream open reading frames has also increased related the discovery of short open reading frames, encoding short functional peptides. These functional peptides from short open reading frames, may be differentiated from upstream open reading frames: uORFs are thought to have primarily regulatory control (29–31). However, it is a strong possibility that many upstream open reading frames, once thought to only have regulatory impact, will be re-evaluated for the possibility that they encode functional protein products.

**Discovery of the function of uORFs, is predicated on identification of uORFs that are translated. For this reason, we took interest in the identification of translated uORFs in humans. Ribosome profiling experiments are the current standard for genome-wide identification of translated upstream open reading frames. However, the performance of these experiments at this identification task, is not well characterized.**

**A comparison of the ribosome profiling experiments, of three independent groups of researchers – Lee et al. 2012, Fritsch et al. 2012, and Gao et al. 2014 – reveals that the number of uORFs co-identified between these experiments is relatively low. Pairwise intersections are 12.2% (Gao ∩ Fritsch), 9.2% (Gao ∩ Lee), and 9.8% (Lee ∩ Fritsch). The number of uORFs co-identified between all three**

**sets, represents only 3.3% of uORFs identified in these studies [Figure 1.C.].**

**This result highlights an important ambiguity. The sensitivity and specificity of ribosome profiling for identifying translated upstream open reading frames, is uncertain.**

**It is possible that ribosome profiling experiments are sensitive at identifying translated uORFs. Under this assumption, there are relatively few translated uORFs. Those uORFs identified in multiple studies, most likely represent true positive examples. Many uORFs labeled in ribosome profiling experiments are false positives, and there is a low false negative rate.**

**Alternatively, it is possible that ribosome profiling experiments have high specificity in identifying translated uORFs. From this perspective, the large majority of the uORFs identified through ribosome profiling experiments are truly translated. It follows that there exist a large number of translated uORFs, that are not identified by ribosome profiling experiments – a high false-negative rate.**

**We proceed on the assumption that the total universe of translated upstream open reading frames, is much larger than that identified through ribosome profiling experiments. In other words, we assume ribosome profiling experiments are sensitive in identifying translated uORFs, with a high false-negative rate [Figure 1.D.].**

Researchers have recently explored this assumption in the model organisms A*rabidopsis thaliana* and *Saccharomyces cervisiae.* In these studies, ribosome profiling data is used to predict translated uORFs that are not identified experimentally (32,33). In humans, patterns of ribosome profiling occupancy have been used to maximize the number of inferred translation products identified in ribosome profiling experiments (34,35). These areas of study have proven productive. In these non-human and human studies, likely translated CDSs and uORFs are identified, numbering in the thousands.

For our investigation of the prevalence of translated upstream open reading frames in humans, we began by performing a computational scan of the GENCODE genome annotation (36). We searched for uORFs associated with protein coding genes. All the possible uORFs beginning either with ATG, or a near-cognate start codon, were identified (all single nucleotide variants of the canonical ATG). This scan yields a universe of all possible uORFs, numbering nearly 1.3 million.

We do not expect that all uORFs identified in a genome-wide scan are functional. For this reason, we sought means to separate translated uORFs, from uORFs with a low chance of translation. In order to effect this

identification of functional uORFs, we studied the human ribosome profiling experiments of three research groups – Lee et al. 2012, Fritsch et al. 2012, and Gao et al. 2014 (11–13). Each of these three experiments uses translational inhibitors that arrest the translating ribosome at the first peptide bond. This arrest of translation at initiation, allows for the identification of translated upstream open reading frames to high precision.

DIDN'T WE INTRO

uORFs in our computational set, that displayed considerable similarity to known translated uORFs, we predicted to be translated and functional. We validate our predicted uORFs, using statistical analyses, and by examining the effect of individual genotype, on parameters related to uORF translation: protein level, and ribosome occupancy at the uORF.

MORE

Following examination efficacy of our method, we demonstrate biological applications of our large set of predicted uORFs. Specifically, we use the predictions we generate, to measure the functional impact of somatic mutations affecting uORFs, in tissue-matched tumor samples (37). We also provide a baseline for the functional consequence of uORFs, using the 1000 Genomes project's database of human variation (38) and the NHGRI-EBI GWAS catalog (39).

The set of uORFs that we predict are likely translated and functional, extends scope far beyond those identified in ribosome profiling experiments. Through our study, we predict that there exist many thousands of translated, functional uORFs, that have not yet been annotated accordingly. We provide a resource of predicted translated uORFs, for other scientists to use in their effort to understand uORF function in health and disease.

**Methods:**

*Extracting uORFs from GENCODE:*

uORFs were extracted from the v19 of the GENCODE annotation of the human genome(36). uORFs were defined as a start codon within the 5’UTR, and a downstream stop codon before the end of the CDS. All three possible reading frames were examined. ATG, and near cognate start codons were included in this search [ATG, TTG, GTG, CTG, AAG, AGG, ACG, ATA, ATT, ATC].

*Ribosome profiling experiments as a reference set:*

The ribosome profiling experiments of Lee et al. (2012), Fritsch et al. (2012) and Gao et al. (2014), were used to obtain an experimentally validated set of translated upstream open reading frames [Figure 1.C]. These studies identify translation initiation sites (TIS), through treatment of human cell lines with antibiotic translation inhibitors. These treatments reliably halt translation, in predictable proximity to the start codon (12-13 nucleotides downstream). As such, these experiments provide us with high resolution information about translation initiation sites in the human genome.

We employed the read alignments and identification of the translation initiation sites, as provided by these three groups of researchers. Each group ultimately expressed their results as positional coordinates for uORF start codons, with corresponding transcripts identified by RefSeqID. We mapped the RefSeqIDs provided in these papers, to corresponding GENCODE Ensembl IDs. This mapping provides position information in the global positioning coordinates of the GENCODE annotation.

The cell lines, treatment protocols, and TIS identification mechanism employed by each of these three research groups is summarized in *Methods Supplement*.

*Literature review of translated human uORFs:*

In addition to ribosome profiling studies, confirmed translated uORFs were obtained from the biomedical literature (8,40,41). uORFs studied in humans that displayed functionality (demonstrated regulation of the CDS product) were added to the set of positive uORFs. In total, 33 uORFs, associated with 33 separate genes, were included from this literature review.

*Cleansing the data set, by removal of N-terminal extensions and aTISs, and isolation of unique transcript IDs:*

Reading frames labeled as uORFs in experiment, but without a stop codon before the stop codon of the CDS, contain the complete CDS sequence. These N-terminal extensions of the CDS sequence, may have some of the functional activity of the primary gene protein product, and were removed from the data set. In addition, any uORF start codon that is annotated as an alternative translation initiation sites (aTISs) for the CDS, was also removed from the data set.

Multiple transcript IDs, may share identical chromosomal coordinates. In order to avoid over-counting, only one transcript ID was included for a given set of chromosomal coordinates. This selection was made randomly, from among transcripts with identical chromosomal coordinates.

*Positive, neutral, and unlabeled data sets:*

uORFs were divided into three separate sets, according to their experimental translation status:

*Positive:* uORFs identified as translated in two or more ribosome profiling experiments, or through literature review.
*Neutral:* uORFs identified as translated in not more than one ribosome profiling experiment.
*Unlabeled:* uORFs that were not identified as translated in any ribosome profiling experiment, or through literature review.

*Extraction of attributes associated with uORFs:*

In order to determine what features make a uORF more likely to be translated (classified as positive), feature data was extracted for each uORF. **The features chosen cover a broad range of categories of data, including features associated with uORF structure (e.g. uORF length, % A/T/G/C base content, start codon), uORF evolutionary conservation (e.g. GERP score, SNP content / length), and the genomic context of uORFs (e.g. mRNA expression level, Kozak start codon context, distance between CDS and uORF). 89 features were used. A complete listing of these features, including details relating to the extraction and calculation of each feature, is included in *Methods Supplement*.**

*Feature discretization:*

The minimum description length principle (MDLP) algorithm was used to discretize each of our chosen attributes (42). The MDLP algorithm discretizes data, while optimizing bin size according to an information theoretic principle. In choosing between two locations for a possible cut point in the data, the MDLP process selects the location that minimizes disruption of pattern in the continuous data. Maintaining pattern corresponds to maximizing information content retained in the discretization. The result is optimal number and spacing of bins. MDLP discretization was implemented using the 'discretization' package available for R (http://cran.r-project.org/web/packages/discretization/index.html).

*Prioritization of feature data:*

For each included feature, the distribution for that feature was compared between positive and unlabeled uORFs. This comparison was completed using the kolmogorov-smirnov (KS) statistic. A greater KS statistic, indicates a greater difference between the distributions for that feature. The KS statistic was thus used as a proxy for the ability of that attribute, to distinguish between positive and unlabeled features.

*Classifying uORFs, according to attributes:*

Using discretized feature data, the probability distribution for each attribute was used to distinguish between positive uORFs and unlabeled uORFs. For a given uORF, we determined if the attributes of that uORF were consistent with a translated uORF, according to the following algorithm:

$$P_{pos}\prod_{i=\{1...89\}}p(A_i|pos) == p_{pos}$$
$$P_{neg}\prod_{i=\{1...89\}}p(A_i|unl) == p_{neg}$$

With

$$P_{pos} = 0.61$$
$$P_{neg} = 1 = P_{pos}$$

$P_{pos}$ is the prior probability associated with positive uORFs. $P_{pos}$ is chosen as the f-statistic maximizing value seen in cross-validation (0.61). $P_{neg}$ is the prior probability associated with negative uORFs. $A_i$ is the value of a given attribute, such that $p(A_i|pos)$, and $p(A_i|unl)$ represent the frequency of that attribute value among the positive, and unlabeled sets respectively. $p_{pos}$ represents the probability the uORF is positive. $p_{pos}$ represents the probability the uORF is negative. This formulation corresponds to a Naive-Bayes machine learning algorithm applied to positive and unlabeled examples (43). **We note likely violation of the feature independence requirement of Naive-Bayes. However, empirical and theoretical study has demonstrated optimal classification performance, even where feature independence does not hold (44,45).**

*Model validation:*

To validate our model, we serially trained our model on two of three ribosome profiling data sets. Following this training, we used the model to extract uORFs only identified in the witheld third ribosome profiling data set, from among the unlabeled examples. The accurate classification of ribosome profiling data using this method, would suggest that ribosome profiling experiments have a high false-negative rate, and a low false-positive rate. The success of these differentially trained models, is expressed as ROC curves, with area under the curve (AUC) calculated for each curve.

As a measure of the biological significance of the uORFs we identify, we examined how natural variation affecting our predicted translated uORFs, alters protein level and ribosome localization in humans. Protein levels and local ribosome quantitative trait loci (cis-rQTL), were obtained from the ribosome profiling and proteomic experiments of Battle et al. 2015 (46). Individual genotype information is available from the 1000 Genomes project, for 47 individuals in the Battle et al. study. Changes in protein level and ribosome localization, as a function of natural variation affecting uORFs, suggests the biologic validity of our predictions.

*Natural variation affecting predicted positive uORFs:*

The impact of variation on uORF start codons was of interest, as the impact of variation altering a start codon is relatively predictable. Uncertain significance of uORFs as protein products, makes other structure-function relationships less straightforward. Natural variant SNPs affecting the start codons of predicted positive uORFs, were obtained from the 1000 Genomes

project. The subset of these SNPs, that are associated with differential disease susceptibility, are identified through search of the NHGRI-EBI GWAS database. Also, measurement of comparative frequency of mutation among uORF start codons, is used to examine differential evolutionary conservation and functional significance.

*Cancer mutation affecting predicted positive uORFs:*

The study of Alexandrov et al. 2012 (37) provides a set of exomic somatic mutations according to patient sample, and cancer type. We employed the start codons of our predicted positive uORFs, to uncover possible cancer mutation, altering uORF function. This allowed for estimation of the frequency with which uORFs are impacted in cancer, according to cancer type and according to uORF start codon. Patterns of function in genes affected by mutation of uORFs in cancer, was assessed via the GO genome annotation database (47). Overrepresented GO terms were identified, with overrepresentation assessed via the hypergeometric statistical test, with multiple testing correction via Benjamini & Hochberg's FDR correction (48). Networks between GO terms, were constructed using the Cytoscape package BiNGO (49).

**Results:**

The search of the GENCODE genome annotation, for the universe of all possible uORFs, yielded 1 270 265 unique uORFs. Within this large set, we isolated the subset of uORFs found to be translated in the studies of Lee et al. 2012, Fritsch et al. 2012, and Gao et al. 2014 [Figure 1.C]. We further stratified this set of translated uORFs, according to shared representation of uORFs among the three studies. uORFs identified in the intersection between two or more of these studies, were used as the reference standard for translated uORFs. This intersection also helps to control for possible false positives, and to control for differences in experimental procedure and tissue specificity (HEK293 vs. THP-1). Literature review, yielded 33 additional examples of translated uORFs, that were also included in the set of positive, translated uORFs.

Overlap between the three ribosome profiling experiments was found to be low, with pairwise intersections of 12.2% (Gao ∩ Fritsch), 9.2% (Gao ∩ Lee), and 9.8% (Lee ∩ Fritsch), with the number of uORFs shared between all three sets representing only 3.3% of uORFs identified in these studies.

The relative representation of start codons identified in ribosome profiling experiments, is noteworthy for the prevalence of both CTG (28.2%) and ATG (46.1%) start codons. These start codons represent the majority (74.3%) of start codons found in these ribosome profiling studies. In intersection between ribosome profiling studies, CTG (30.5%) and ATG (34.6%) continue

to represent the majority of start codons (65.1%) [Figure 1.E.]. Representation of every near-cognate start codon was found in intersections between studies, with the exception of AAG and AGG.

We next followed the procedure outlined in Figure 2.A, in an effort to isolate uORFs that are likely to be translated, from those identified via genome-wide scan. Distributions of attributes for positive, translated uORFs, were compared with distributions of those same attributes as observed in the set of unlabeled, computationally derived uORFs [Figure 2.B.]. Differences in these distributions between positive and unlabeled examples, suggest that the attribute is of greater utility in identifying translated uORFs. The KS statistic and corresponding p-value, for each of the 89 attributes assessed in this study, is provided in Table 2. The top 10 attributes, listed according to magnitude of KS statistic, are given in Figure 2.C.

**This relative utility among attributes for distinguishing translated uORFs, reveals several notable structure-function relationships. The particular start codon employed by the uORF is important, with ATG suggesting the greatest functional significance. Interestingly, the clustering of large numbers of start codons in the uORF, appears to increase the likelihood that a uORF is translated, as does a shorter positioning between the uORF and the CDS. High GERP evolutionary conservation scores for the start codon and stop codon of the uORF are also useful predictors. mRNA expression level for the transcripts holding the latent uORF, showed great importance. However, this result must be considered in the context of high expression level transcripts, having proportionately higher representation in ribosome profiling experiments.**

The discretized attributes of positive and unlabeled sets of uORFs, were used to build a statistical classifier, within a Naive-Bayes framework. This classifier predicts translated uORFs, through examination of the totality of attributes associated with an individual uORF.

The result of application of the classifier is shown in figure 3.A. The percentage of positive examples, that are ultimately retained as likely translated is 76.8% [590/768], 67.1% of neutral uORFs are classified as likely translated [2379/3543], and 14.7% of unlabeled uORFs are likely translated [185833/1265954]. The overall number of uORFs classified as likely translated, is 188 802, representing 14.9% of computationally identified uORFs [188802/1270265]. 75.5% of predicted positive uORFs lie entirely upstream of the CDS, throughout their length. 25.5% of predicted translated uORFs are out-of-frame with the CDS, and overlap with the CDS.

Predicted functional uORFs, are ranked according to probability of translation. This ranking allows for provision of a top 10% of likely translated

uORFs. This more demanding threshold, is useful in highlighting the most reliable predictions. A complete list of upstream open reading frames identified as likely translated, is provided in *Results Supplement.* The 10% highest probability examples are also specified.

As validation of our technique for distinguishing between positive and unlabeled upstream open reading frames, we serially excluded one of the three ribosome profiling experiments from the positive training set, including that set among unlabeled examples. Retrieval of the excluded set, then functions as a measure of the accuracy and generalizability of our method. The result of this validation procedure is shown in Figure 3.B. The ROC curve for the retrieval of each ribosome profiling set is given. The AUC for each of these ROC curves, is similar. 0.82, 0.79, and 0.77 for the retrieval of Lee, Gao, and Fritsch uORFs respectively. A subtle but important result related to this validation, is the suggestion of a high false-negative rate for ribosome profiling studies. Namely, predicted positive examples based on a limited set of studies, reflect those examples that additional experiments would ultimately identify to be translated.

The proportion of uORFs ultimately identified as positive from each ribosome profiling study, is shown in Figure 3.C. The results were similar for each of the ribosome profiling experiments, at approximately 70% in each case (72% of Gao, 71% of Lee, 70% of Fritsch).

The distribution of start codons for predicted translated uORFs, in comparison to the computational set, is shown in Figure 3.D. There are a large number of CTG start codons in the computationally derived set (19.3%), and the greatest number of predicted positive uORFs are also initiated with a CTG start codon 11.8%. ATG has a lower comparative prevalence in both the computationally derived set and predicted set (6.7% and 8.2% respectively).

Figure 4.A shows the frequency with which predicted positive uORF start codons are altered by 1000 Genomes Project germline variants. The results are normalized by population start codon frequency. The ATG start codon is relatively conserved among start codons, suggesting functional importance. It is rarely interrupted by human variants (relative rate (RR) 0.03). The CTG start codon, although more prevalent among predicted positive uORFs, is altered relatively frequently by natural human variants (RR 0.52).

GWAS SNPs listed in the NHGRI-EBI GWAS database, that impact our predicted uORFs are listed in Table 3. These disease associated SNPs, may owe their functional consequence to alteration of a translated uORF.

An analysis of the alteration of predicted positive uORFs, was applied to somatic mutations across cancer types. This analysis is shown in figure 4.B.

CTG is the most commonly modified start codon in these combined cancers. ATG is interrupted at a RR of 0.25 in comparison to CTG. The higher RR of interruption of both ATG and CTG in cancer as compared to germline variants – 8 fold higher, and 2 fold higher respectively – further suggests functional consequence attributable to these uORFs. Exomic cancer mutations breaking the highest scored uORFs, are listed in Table 4.

In order to evaluate the frequency with which uORFs are interrupted by mutation in cancer, the proportion of positive uORFs interrupted by mutation was calculated for each cancer type. This analysis is shown in Figure 4.C. This proportion of positively scored uORFs to negative scored uORFs, is near consistent across cancer types, ranging from a low of 8:1 for acute lymphoblastic leukemia, to a high of 20:1 for pancreatic cancer. The between group differences for interrupted predicted translated uORFs are significant (chi-square = 45, p-value = <<0.001). A pilocytic astrocytoma may rely to a greater extent on altered uORFs for survival, than B-cell lymphoma, or breast cancer.

Networks of GO terms were constructed, for genes associated with the mutation of predicted translated uORFs in cancer [Figure 4.D.]. Three networks of overrepresented GO terms remain, following correction for statistical significance and multiple testing. These are networks associated with cellular functions of probable significance in cancer -- cellular death, immune modulation, and tissue morphogenesis. Lack of response to apoptotic signaling, and immune tolerance, are well known mechanisms that cancer cells prolong survival. The alteration of genes involved in tissue morphogenesis, may relate to the poor tissue differentiation exhibited by cancer cells that is integral to tumor grading schemas.

As further validation of the biologic significance of our predictions, we explored the effect of human germline variation, on measured local ribosome occupancy, and on protein level from downstream protein coding genes. The results of Battle et al. 2015, and genotype information from the 1000 Genomes project, provide the basis for this natural study. Both protein level and ribosome occupancy, if affected by alteration of a predicted translated uORF alteration, suggest the functional significance of that uORF.

47 individuals were assessed for the effect on protein level, of variants altering predicted translated uORFs. For those genes where this natural experiment provides close to the ideal assignment ratio of 23 individuals per group, we a see a definite trend to decreased protein levels [Figure 4.E.]. This decrease is statistically significant.

Known cis-rQTLs provide an inventory of variants with statistically significant effect on local ribosome occupancy. There is significant enrichment for rQTLs interrupting positively scored start codons as compared with negatively scored start codons [Figure 4.F.]. While the effect we would expect due to

random mutation is 14.9%, we observe that 48% of these rQTLs (21/44) interrupt positively scored start codons -- a 3x higher rate. This indicates that many rQTLs, may measure the direct effect of disruption of translated uORFs.

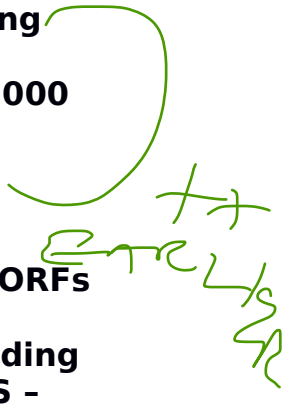
Both our results for protein level, and for ribosome occupancy, suggest that our predictions have validity, and measurable functional impact.

Discussion:

In this study, we are able to identify 188 802 likely translated upstream open reading frames, from a global set of 1 270 265 unique uORFs identified in the human genome. **We highlight the 10% of our predictions that are most likely to be translated, as a high reliability subset.**

**We began with the assumption that ribosome profiling experiments have a high false negative rate for identification of translated upstream open reading frames. The low overlap between ribosome profiling experiments suggests this possibility. Furthermore, the finding that pairs of ribosome profiling experiments, may be used to correctly identify the uORFs translated in a third experiment, also suggests a high false negative rate. The number of uORFs we identify as likely translated is consistent with these observations, but remarkable in comparison to other studies on the topic. This is true even for our 10% highest quality predictions.**

**An estimate for the number of translated uORFs can be made, using the mathematical framework developed for mark and recapture experiments of population ecology. If independent ribosome profiling experiments -- such as those of Fritsch et al. 2012, Lee et al. 2012, and Gao et al. 2014 -- represent a resampling of the same population, we can use the repeat identification of uORFs among these experiments, to estimate the total number of translated uORFs. This procedure yields an estimate of approximately 10 000 functional uORFs in the human genome, using the Schnabel or Schumacher and Eschmeyer equations (50,51).**



**However, these estimates rely on a fixed population, without distinction among members of the population. Translation of uORFs may vary according to cell-type and environmental condition. Furthermore the structure and context of a given uORF -- including start codon, base composition, and relative position to the CDS – likely contribute to varying degrees of affinity to translation, among translated uORFs.**

Related to the differential translation of uORFs among cell-types and environmental conditions, our study applies the intersection of three ribosome profiling studies, to form a reference set of known translated uORFs. This intersection provides some control against tissue specific results (both human embryonic kidney, and human monocytic cell lines were examined). It also provides some control against differences experimental condition and protocol. However, this method may discount the functional significance of uORFs that are translated in cell-type specific fashion, or only under specific cellular conditions.

Just as protein levels vary widely across cell-type (52), it may prove that the translation of uORFs varies considerably across cell types, and cellular conditions. This has been suggested by the large number of studies that have demonstrated differential translation from uORF start codons, in stress conditions compared to control. While outside the scope of our study, the analysis of cell-type specific and condition specific translated uORFs, may expand estimates of the population of uORFs.

Related to variable compatibility of uORFs translation, ATG is the most common uORF start codon in the ribosome profiling studies examined in this study (46.1%). However, it is only the fourth most common uORF start codon identified computationally, and 5$^{th}$ most common predicted positive uORF start codon. If ATG has high representation in ribosome profiling experiments due to its affinity for translation, lower affinity yet still functional near-cognate start codons, may be similarly underrepresented. Lower affinity near cognate-start codons, due to their overall abundance, may ultimately prove to have the greatest functional impact on the landscape of translation.

Perhaps the most convincing validation of our predictions, is our finding that alteration of predicted functional uORFs, as a consequence of germline genetic variation, appears to impact ribosome binding and protein level in humans. Our ability to conduct this natural experiment, is an impressive testament to availability and usefulness of data spanning the information flow from DNA→RNA→protein in individual human subjects.

From this biological validation procedure, it is also of interest, that the overall effect of uORF interruption, appears to be a decrease in downstream protein level. This is contrary to common view that uORFs act as translational repressors. Mechanisms have been studied, where uORFs act to up-regulate the presence of a downstream coding sequence (e.g. leaky-scanning). However, our analysis would appear to suggest that this effect is a more common consequence for upstream open reading frames than is credited.

Identification of human germline variants altering predicted positive uORFs, suggests locations where the creation or destruction of a uORF, is likely to alter protein levels. Some of these alterations have already been

characterized as consequential in GWAS studies. Further study of these locations, could reveal further important disease associations.

The application of our results to exomic cancer somatic mutation data, identifies locations where mutation of uORFs, may contribute to the pathogenesis of cancer. GO terms, associated with the mutation of predicted translated uORFs in cancer, appear to correspond to essential domains of cancer pathogenesis: tissue differentiation, cell survival, and immune response. Mutation of uORFs, could conceivably both up-regulate oncogenes, or down-regulation of tumor suppressor genes. In this way, our work could be used to help broaden knowledge of the role of uORFs in cancer, beyond recently identified individual examples (53).

These applications of our results, suggest exciting avenues for future research. Our results offer a broad and validated catalog of uORFs. We provide that catalog can serve as a point of reference for other researchers, towards investigation of the function of uORFs, in meaningful context to areas of personal expertise and interest.

**Bibliography:**


1. Kozak M. An analysis of 5’-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res [Internet]. 1987 [cited 2016 Aug 16];15(20):8125–48. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/15.20.8125

2. Kochetov A V., Sarai A, Rogozin IB, Shumny VK, Kolchanov NA. The role of alternative translation start sites in the generation of human protein diversity. Mol Genet Genomics [Internet]. 2005 Jul 15 [cited 2016 Aug 16];273(6):491–6. Available from: http://link.springer.com/10.1007/s00438-005-1152-7

3. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell. 2011;147(4):789–802.

4. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science [Internet]. 2009 Apr 10 [cited 2016 Aug 16];324(5924):218–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19213877

5. Ivanov IP, Loughran G, Atkins JF. uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs. Proc Natl Acad Sci [Internet]. 2008 Jul 22 [cited 2016 Aug 23];105(29):10079–84. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0801590105

6. Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov P V. Identification of

evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res [Internet]. 2011 May 1 [cited 2016 Aug 17];39(10):4220–34. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr007

7. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell [Internet]. 1986 Jan [cited 2016 Aug 17];44(2):283–92. Available from: http://linkinghub.elsevier.com/retrieve/pii/0092867486907622

8. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci [Internet]. 2009 May 5 [cited 2016 Aug 16];106(18):7507–12. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0810916106

9. Brar G a, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat Rev Mol Cell Biol [Internet]. 2015;16(11):651–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26465719

10. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet [Internet]. 2014 Jan 28 [cited 2016 Aug 17];15(3):205–13. Available from: http://www.nature.com/doifinder/10.1038/nrg3645

11. Gao X, Wan J, Liu B, Ma M, Shen B, Qian S-B. Quantitative profiling of initiating ribosomes in vivo. Nat Methods [Internet]. 2014 Dec 8 [cited 2016 Aug 17];12(2):147–53. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3208

12. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res [Internet]. 2012 Nov 1 [cited 2016 Aug 17];22(11):2208–18. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.139568.112

13. Lee S, Liu B, Lee S, Huang S-X, Shen B, Qian S-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci [Internet]. 2012 Sep 11 [cited 2016 Aug 17];109(37):E2424–32. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1207846109

14. Johnstone TG, Bazzini AA, Giraldez AJ, Abràmofff M, Magalhães P, Ram S, et al. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J [Internet]. 2016 Apr 1 [cited 2016 Aug 17];35(7):706–23. Available from: http://emboj.embopress.org/lookup/doi/10.15252/embj.201592759

15. Somers J, Pöyry T, Willis AE. A perspective on mammalian upstream

open reading frame function. Int J Biochem Cell Biol. 2013;45(8):1690–700.

16. Meijer HA, Thomas AAM. Control of eukaryotic protein synthesis by upstream open reading frames in the 5’-untranslated region of an mRNA. Biochem J [Internet]. 2002 Oct 1 [cited 2016 Aug 17];367(Pt 1):1–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12117416

17. Barbosa C, Peixeiro I, Romão L, Morris D, Geballe A, Calvo S, et al. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. Fisher EMC, editor. PLoS Genet [Internet]. 2013 Aug 8 [cited 2016 Aug 17];9(8):e1003529. Available from: http://dx.plos.org/10.1371/journal.pgen.1003529

18. Morris DR, Geballe AP. Upstream Open Reading Frames as Regulators of mRNA Translation. Mol Cell Biol [Internet]. 2000 Dec 1 [cited 2016 Aug 17];20(23):8635–42. Available from: http://mcb.asm.org/cgi/doi/10.1128/MCB.20.23.8635-8642.2000

19. Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell [Internet]. 2009 Feb 20 [cited 2016 Aug 30];136(4):731–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19239892

20. Hinnebusch AG, Ivanov IP, Sonenberg N, Hinnebusch AG, Kozak M, Starck SR, et al. Translational control by 5’-untranslated regions of eukaryotic mRNAs. Science [Internet]. 2016 Jun 17 [cited 2016 Aug 17];352(6292):1413–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27313038

21. Chua JJE, Schob C, Rehbein M, Gkogkas CG, Richter D, Kindler S, et al. Synthesis of two SAPAP3 isoforms from a single mRNA is mediated via alternative translational initiation. Sci Rep [Internet]. 2012 Jul 2 [cited 2016 Aug 16];2:277–98. Available from: http://www.nature.com/articles/srep00484

22. Bergeron D, Lapointe C, Bissonnette C, Tremblay G, Motard J, Roucou X. An Out-of-frame Overlapping Reading Frame in the Ataxin-1 Coding Sequence Encodes a Novel Ataxin-1 Interacting Protein. J Biol Chem [Internet]. 2013 Jul 26 [cited 2016 Aug 17];288(30):21824–35. Available from: http://www.jbc.org/cgi/doi/10.1074/jbc.M113.472654

23. Starck SR, Tsai JC, Chen K, Shodiya M, Wang L, Yahiro K, et al. Translation from the 5’ untranslated region shapes the integrated stress response. Science [Internet]. 2016 Jan 29 [cited 2016 Aug 17];351(6272):aad3867. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26823435

24. Andreev DE, O’Connor PBF, Zhdanov A V, Dmitriev RI, Shatsky IN, Papkovsky DB, et al. Oxygen and glucose deprivation induces

widespread alterations in mRNA translation within 20 minutes. Genome Biol [Internet]. 2015 Dec 6 [cited 2016 Aug 17];16(1):90. Available from: http://genomebiology.com/2015/16/1/90

25. Shalgi R, Hurt JA, Krykbaeva I, Taipale M, Lindquist S, Burge CB. Widespread Regulation of Translation by Elongation Pausing in Heat Shock. Mol Cell. 2013;49(3):439–52.

26. Wiita AP, Ziv E, Wiita PJ, Urisman A, Julien O, Burlingame AL, et al. Global cellular response to chemotherapy-induced apoptosis. Elife [Internet]. 2013 Jul [cited 2016 Aug 17];2(1):e01236. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1097276507004005

27. Gerashchenko M V., Lobanov A V., Gladyshev VN. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc Natl Acad Sci [Internet]. 2012 Oct 23 [cited 2016 Aug 17];109(43):17394–9. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1120799109

28. Liu B, Han Y, Qian S-B. Cotranslational Response to Proteotoxic Stress by Elongation Pausing of Ribosomes. Mol Cell. 2013;49(3):453–63.

29. Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, et al. Extensive identification and analysis of conserved small ORFs in animals. Genome Biol [Internet]. 2015 Dec 14 [cited 2016 Aug 23];16(1):179. Available from: http://genomebiology.com/2015/16/1/179

30. Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet [Internet]. 2014 Mar [cited 2016 Aug 17];15(3):193–204. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24514441

31. Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, Natsume T, et al. Analysis of Small Human Proteins Reveals the Translation of Upstream Open Reading Frames of mRNAs. Genome Res [Internet]. 2004 Oct 15 [cited 2016 Aug 17];14(10b):2048–52. Available from: http://www.genome.org/cgi/doi/10.1101/gr.2384604

32. Selpi S, Bryant CH, Kemp GJ, Sarv J, Kristiansson E, Sunnerhagen P, et al. Predicting functional upstream open reading frames in Saccharomyces cerevisiae. BMC Bioinformatics [Internet]. 2009 [cited 2016 Aug 17];10(1):451. Available from: http://www.biomedcentral.com/1471-2105/10/451

33. Hu Q, Merchante C, Stepanova AN, Alonso JM, Heber S. Genome-Wide Search for Translated Upstream Open Reading Frames in Arabidopsis Thaliana. IEEE Trans Nanobioscience [Internet]. 2016 Mar [cited 2016 Aug 31];15(2):148–57. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7404026

34. Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, et al. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol Cell. 2015;60(5):816–27.

35. Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Elife [Internet]. 2016 [cited 2016 Aug 31];5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27232982

36. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res [Internet]. 2012 Sep 1 [cited 2016 Aug 21];22(9):1760–74. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.135350.111

37. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature [Internet]. 2013 Aug 14 [cited 2016 Aug 17];500(7463):415–21. Available from: http://www.nature.com/doifinder/10.1038/nature12477

38. Project Consortium G, Consortium Participants are arranged by project role G, by institution alphabetically then, alphabetically within institutions except for Principal Investigators finally, Leaders P, indicated as, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;490.

39. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L and PH. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;Vol. 42((Database issue)):D1001–6.

40. Wen Y, Liu Y, Xu Y, Zhao Y, Hua R, Wang K, et al. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. Nat Genet [Internet]. 2009 Feb 4 [cited 2016 Aug 31];41(2):228–33. Available from: http://www.nature.com/doifinder/10.1038/ng.276

41. Raveh-Amit H, Maissel A, Poller J, Marom L, Elroy-Stein O, Shapira M, et al. Translational Control of Protein Kinase C by Two Upstream Open Reading Frames. Mol Cell Biol [Internet]. 2009 Nov 15 [cited 2016 Aug 31];29(22):6140–8. Available from: http://mcb.asm.org/cgi/doi/10.1128/MCB.01044-09

42. Fayyad U, Irani K. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. donga.ac.kr [Internet]. [cited 2016 Aug 17]; Available from: http://web.donga.ac.kr/kjunwoo/files/Multi interval discretization of continuous valued attributes for classification learning.pdf

43. Liu B, Dai Y, Li X, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In: Third IEEE International Conference on Data Mining [Internet]. IEEE Comput. Soc; 2003 [cited 2016 Aug 17]. p. 179–86. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1250918

44. Rish I. An empirical study of the naive Bayes classifier. researchgate.net [Internet]. [cited 2016 Sep 19]; Available from: https://www.researchgate.net/profile/Irina_Rish/publication/228845263_An_Empirical_Study_of_the_naive_Bayes_Classifier/links/00b7d52dc3ccd8d692000000.pdf

45. Zhang H. The Optimality of Naive Bayes. AA [Internet]. 2004 [cited 2016 Sep 19]; Available from: http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/Optimality_of_Naive_Bayes.pdf

46. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, et al. Genomic variation. Impact of regulatory variation from RNA to protein. Science [Internet]. 2015 Feb 6 [cited 2016 Aug 31];347(6222):664–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25657249

47. Gene Ontology Consortium T, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene Ontology: tool for the unification of biology.

48. Author T, Benjamini Y, Hochberg Y, Benjaminit Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Source J R Stat Soc Ser B J R Stat Soc Ser B J R Stat Soc B [Internet]. 1995 [cited 2016 Aug 31];57(1):289–300. Available from: http://www.jstor.org/stable/2346101

49. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics. 2005;21(16):3448–9.

50. Schnabel ZE. The Estimation of Total Fish Population of a Lake. Am Math Mon [Internet]. 1938 Jun [cited 2016 Sep 19];45(6):348. Available from: http://www.jstor.org/stable/2304025?origin=crossref

51. Schumacher, F. X. and Eschmeyer RW. The estimation of fish populations in lakes and ponds. J Tennessee Acad Sci . 1943;18(228–249).

52. Pontén F, Gry M, Fagerberg L, Lundberg E, Asplund A, Berglund L, et al. A global view of protein expression in human cells, tissues, and organs. Mol Syst Biol [Internet]. 2009 Dec 22 [cited 2016 Aug 17];5(1):799–816. Available from: http://msb.embopress.org/cgi/doi/10.1038/msb.2009.93

53. Wethmar K, Schulz J, Muro EM, Talyan S, Andrade-Navarro MA, Leutz A. Comprehensive translational control of tyrosine kinase expression by

upstream open reading frames. Oncogene [Internet]. 2016 Mar 31 [cited 2016 Aug 17];35(13):1736–42. Available from: http://www.nature.com/doifinder/10.1038/onc.2015.233

**Tables:**

*Table 1:* uORF features. Features are listed according to the KS statistic for each attribute, measured between positive and unlabeled uORFs.

| ***Rank*** | ***Attribute*** | ***KS statistic*** | ***p value*** | ***Rank*** | ***Attribute*** | ***KS statistic*** | ***p value*** |
|---|---|---|---|---|---|---|---|
| 1 | GTEX Bone Marrow | 0.54 | 0.000 | 46 | #AGG | 0.20 | 0.000 |
| 2 | GTEX Liver | 0.50 | 0.000 | 47 | #CTG | 0.20 | 0.000 |
| 3 | GTEX Lung | 0.49 | 0.000 | 48 | Kozak context | 0.19 | 0.000 |
| 4 | GTEX Pituitary | 0.49 | 0.000 | 49 | % GERP elements | 0.19 | 0.000 |
| 5 | Ribosome profiling uORF start codon frequency | 0.48 | 0.000 | 50 | uORF start codon to CDS start codon distance | 0.19 | 0.000 |
| 6 | GTEX Nerve | 0.48 | 0.000 | 51 | mRNA ΔG uORF start [20-59]BP | 0.18 | 0.000 |
| 7 | GTEX Muscle | 0.47 | 0.000 | 52 | #GTG | 0.18 | 0.000 |
| 8 | GTEX Pancreas | 0.47 | 0.000 | 53 | mRNA ΔG uORF start [40-79]BP | 0.18 | 0.000 |
| 9 | GTEX Adipose Tissue | 0.47 | 0.000 | 54 | %A | 0.17 | 0.000 |
| 10 | GTEX Skin | 0.47 | 0.000 | 55 | 5' cap to uORF start codon distance | 0.16 | 0.000 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11 | GTEX Spleen | 0.47 | 0.000 | 56 | mRNA ΔG uORF stop codon [0,39]BP | 0.16 | 0.000 |
| 12 | GTEX Stomach | 0.46 | 0.000 | 57 | mRNA ΔG uORF stop codon [-20,19]BP | 0.16 | 0.000 |
| 13 | GTEX Cervix Uteri | 0.46 | 0.000 | 58 | uORF stop codon to CDS start codon distance | 0.16 | 0.000 |
| ***14*** | ***GTEX (combined)*** | ***0.46*** | ***0.000*** | 59 | mRNA ΔG CDS start [-20,19]BP | 0.14 | 0.000 |
| 15 | GTEX Salivary Gland | 0.46 | 0.000 | 60 | Noderer context | 0.13 | 0.000 |
| 16 | GTEX Uterus | 0.46 | 0.000 | 61 | %G | 0.13 | 0.000 |
| 17 | GTEX Small Intestine | 0.46 | 0.000 | 62 | mRNA ΔG uORF start [60,99]BP | 0.12 | 0.000 |
| 18 | GTEX Prostate | 0.46 | 0.000 | 63 | mRNA ΔG uORF start [80,119]BP | 0.12 | 0.000 |
| 19 | GTEX Esophagus | 0.46 | 0.000 | 64 | mRNA ΔG uORF start [-20,19]BP | 0.12 | 0.000 |
| 20 | GTEX Heart | 0.46 | 0.000 | 65 | mRNA ΔG uORF end [-40,-1] | 0.11 | 0.000 |
| 21 | GTEX Bladder | 0.46 | 0.000 | 66 | mRNA ΔG uORF start [100,139] | 0.11 | 0.000 |
| 22 | GTEX Brain | 0.45 | 0.000 | 67 | mRNA ΔG CDS start [20,59] | 0.11 | 0.000 |
| 23 | GTEX Breast | 0.45 | 0.000 | 68 | mRNA ΔG uORF start [0,39] | 0.10 | 0.000 |
| 24 | GTEX Blood Vessel | 0.45 | 0.000 | 69 | %C | 0.10 | 0.000 |
| 25 | GTEX | 0.45 | 0.000 | 70 | mRNA ΔG uORF | 0.09 | 0.000 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Fallopian Tube | | | | end [20,59] | | |
| 26 | GTEX Blood | 0.45 | 0.000 | 71 | mRNA ΔG CDS start [40,79] | 0.09 | 0.000 |
| 27 | GTEX Thyroid | 0.44 | 0.000 | 72 | mRNA ΔG uORF end [40,79] | 0.08 | 0.000 |
| 28 | GTEX Vagina | 0.44 | 0.000 | 73 | SNPs/length | 0.07 | 0.001 |
| 29 | GTEX Colon | 0.44 | 0.000 | 74 | mRNA ΔG CDS start [0,39] | 0.06 | 0.004 |
| 30 | GTEX Kidney | 0.43 | 0.000 | 75 | #TCG | 0.06 | 0.011 |
| 31 | GTEX Testis | 0.43 | 0.000 | 76 | mRNA ΔG CDS start 100.139 | 0.05 | 0.027 |
| 32 | GTEX Adrenal Gland | 0.42 | 0.000 | 77 | mRNA ΔG uORF start [80,119] | 0.05 | 0.028 |
| 33 | GTEX Ovary | 0.41 | 0.000 | 78 | %T | 0.05 | 0.053 |
| 34 | GTEX Tissue Entropy | 0.40 | 0.000 | 79 | #ACG | 0.05 | 0.056 |
| 35 | #Same start codon | 0.30 | 0.000 | 80 | #CGA | 0.04 | 0.142 |
| 36 | #ATG | 0.28 | 0.000 | 81 | #CGT | 0.04 | 0.198 |
| 37 | #ATA | 0.28 | 0.000 | 82 | mRNA ΔG CDS start [60,99] | 0.03 | 0.309 |
| 38 | #ATT | 0.26 | 0.000 | 83 | uORF length (BP) | 0.03 | 0.447 |
| 39 | #ATG + CTG | 0.26 | 0.000 | 84-89 | #ACG | | |
| 40 | #AAG | 0.23 | 0.000 | 84-89 | #CTA | | |
| 41 | #ATC | 0.22 | 0.000 | 84-89 | #GTA | | |
| 42 | Size 5'UTR (%) | 0.22 | 0.000 | 84-89 | Heterozygosity/length | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 43 | Start codon GERP score | 0.22 | 0.000 | 84-89 | #1000 Genomes SNPs | | |
| 44 | Stop codon GERP score | 0.21 | 0.000 | 84-89 | Heterozygosity | | |
| 45 | #TTG | 0.21 | 0.000 | | | | |

*Table 3:* Individual genes, with uORFs interrupted by germline human variation. Top 10, with disease assocations.

| ***uORF ID*** | ***SNP*** | ***Score*** | ***VAF*** | ***Gene*** | ***Transcripts Affected*** | ***Disease Process (PMID)*** |
|---|---|---|---|---|---|---|
| ENST00000435422.3.uORF_CTG.11 | rs13170573 | 12.3 | 0.47 | SGCD | 28/80 | OSA (25474115) |
| ENST00000526686.1.uORF_TTG.4 | rs1461496 | 10.3 | 0.68 | HSPA8 | 3/72 | CHF/asthma (20300519, 22370858) |
| ENST00000228872.4.uORF_CTG.8 | rs34330 | 21.5 | 0.66 | CDKN1B | 4/9 | Various cancers (17908995) |
| ENST00000355739.4.uORF_ATG.13 | rs751402 | 15.0 | 0.71 | ERCC5 | 3/45 | Gastric cancer (27228234) |
| ENST00000302418.4.uORF_ACG.1 | rs12251445 | 23.9 | 0.31 | KIF5B | 1/7 | Exercise response (18984674) |
| ENST00000270139.3.uORF_GTG.2 | rs2850015 | 24.2 | 0.78 | IFNAR1 | 2/16 | Malaria susceptibili |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | ty (25445652) |
| ENST00000270139.3.uORF_GTG.4 | rs2850015 | 23.1 | 0.78 | IFNAR1 | 2/16 | |
| ENST00000406438.3.uORF_ATT.1 | rs1563634 | 9.0 | 0.68 | SMCR8 | 1/3 | Cancer risk (19432957) |
| ENST00000462284.1.uORF_ATC.1 | rs937283 | 19.2 | 0.34 | MDM2 | 15/20 | Epithelial cancer (26261649) |
| ENST00000310823.3.uORF_CTG.2 | rs12692386 | 21.5 | 0.58 | ADAM17 | 4/8 | Vascular disease (24853957) |

*Table 4:* Individual genes, with uORFs interrupted by somatic cancer mutations. Top 10 by prediction score.

| uORF ID | Location | Score | Cancer Type | Gene Name | Transcripts Affected |
|---|---|---|---|---|---|
| ENST00000371142.4.uORF_ACG.2 | 98346749 | 28.4 | Lung | TM9SF3 | 2/3 |
| ENST00000371142.4.uORF_ACG.1 | 98346749 | 28.1 | Lung | TM9SF3 | 2/3 |
| ENST00000254480.5.uORF_ACG.2 | 47823347 | 26.9 | Lung | SMARCC1 | 3/8 |
| ENST00000254480.5.uORF_ACG.1 | 47823347 | 26.8 | Lung | SMARCC1 | 3/8 |
| ENST00000000233.5.uORF_ACG.2 | 127228421 | 26.7 | Stad | ARF5 | 1/3 |
| ENST00000250894.4.uORF_ACG.1 | 1756190 | 26.5 | Lung | MAPK8IP3 | 1/3 |
| ENST00000345496.2.uORF | 46221698 | 25.7 | Breast | UBE2G2 | 4/15 |

| | | | | | |
|---|---|---|---|---|---|
| _CTG.3 | | | | | |
| ENST00000358015.3.uORF_GTG.2 | 110045592 | 25.6 | Stad | RAD23B | 2/17 |
| ENST00000258341.4.uORF_ACG.1 | 182992786 | 25.5 | Lung | LAMC1 | 1/8 |
| ENST00000395686.3.uORF_GTG.1 | 53162310 | 25.4 | Breast | ERO1L | 5/20 |