Machine learning predicts translation initiation sites in neurologic diseases with expanded repeats

Short title: Machine learning, translation initiation, repeat expansion disorders

Alec C. Gleason[1], Ghanashyam Ghadge[1,2], Jin Chen[3], Yoshifumi Sonobe[1,2], Raymond P. Roos[1,2*]

[1]University of Chicago, Chicago, Illinois, United States of America

[2]Department of Neurology, University of Chicago, Chicago, Illinois, United States of America

[3]Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

*Corresponding author

E-mail: rroos@neurology.bsd.uchicago.edu

Abstract

A number of neurologic diseases, including a form of amyotrophic lateral sclerosis and others associated with expanded nucleotide repeats have an unconventional form of translation called repeat-associated non-AUG (RAN) translation. Repeat protein products accumulate and are hypothesized to contribute to disease pathogenesis. It has been speculated that the repeat regions in the RNA fold into secondary structures in a length-dependent manner, promoting RAN translation. Additionally, nucleotides that flank the repeat region, especially ones closest to the initiation site, are believed to enhance translation initiation. Recently, a machine learning model based on a large number of flanking nucleotides has been proposed for identifying translation initiation sites. However, most likely due to its extensive feature selection and limited training data, the model has diminished predictive power. Here, we overcome this limitation and increase prediction accuracy by a) capturing the effect of nucleotides most critical for translation initiation via feature reduction, b) implementing an alternative machine learning algorithm better suited for limited data, c) building comprehensive and balanced training data (via sampling *without* replacement) that includes previously unavailable sequences, and, d) splitting ATG and near-cognate translation initiation codon data to train two separate models. We also design a supplementary scoring system to provide an additional prognostic assessment of model predictions. The resultant models have high performance, with 85.00-87.79% accuracy exceeding that of the previously published model by >18%. The models presented here are then used to identify translation initiation sites in genes associated with a number of neurologic repeat expansion disorders. The results confirm a number of experimentally discovered sites of

translation initiation upstream of the expanded repeats and predict many sites that are not yet established.

Abbreviations

RAN	Repeat-associated non-AUG
RLI	Repeat length-independent
KCS	Kozak consensus sequence
KSS	Kozak similarity score
AUROC	Area under receiver operating characteristic
ROC	Receiver operating characteristic
RFC	Random forest classifier

Introduction

More than 40 neurologic diseases are caused by expansions of repeat nucleotide sequences in causative genes. The repeats range from three nucleotides, such as ‘CTG’ associated with myotonic dystrophy Types I and II, to up to 12 nucleotides, such as ‘CCCCGCCCCGCG’, associated with progressive myoclonus epilepsy. Protein products translated from expanded repeat sequences tend to accumulate and aggregate, and have been proposed to contribute to disease [1-9]. Interestingly, in some cases, the repeats have been shown to be translated in all three reading frames from both the plus and minus strands of the RNA [10] by a process termed repeat-associated non-AUG (RAN) translation. It is believed that an affinity of translational machinery to folded regions of the RNA may underlie translation of the repeat sequences. Translation may occur from sequences in a repeat length-independent (RLI) mechanism. Regardless of repeat length, sequences may be ordered in such a way that they naturally increase the affinity of translational machinery to initiate at a particular codon. In such a process, translation may initiate not only within the repeat region, but also from sites upstream of the repeat sequences. In this case, repeat peptides will be produced if a stop codon is not encountered by the translational machinery before encountering the repeats. The large number of nucleotides that comprise and precede repeat sequences make the identification of RLI translation initiation sites challenging without proper laboratory evidence or computational methods.

A machine learning model called TITER has been proposed to predict all translation initiation sites in a given sequence. It addresses multiple limitations of the only other such model (to our best knowledge) [11, 12] and remains an important predictive tool. It appears, however, that the large feature selection of TITER and limited training data impair its predictive accuracy. The predictive models described in our investigations have 85.00-87.79% accuracy that exceeds that of TITER. Our models reduce the feature selection to capture the effect of ten critical nucleotides that flank both sides of a putative translation initiation codon since a number of studies have demonstrated a strong impact of nucleotides within this range on translation initiation [13-20]. We also introduce two models tailored for ATG or near-cognate codons because of their differences in initiating translation [21, 22]. The models described here use an alternative machine learning algorithm better suited for limited data [23]. We also present unbiased training data through sampling techniques *without* replacement, using gene sequences that have been unavailable to TITER. Finally, we generate a scoring metric to supplement model predictions. The models confirm nearly all experimentally established translation initiation sites upstream of repeats and, importantly, predict multiple sites that have not yet been investigated.

Results

Kozak similarity score algorithm

Before applying machine learning, we evaluated the performance of a more straightforward algorithm that uses a limited number of nucleotides as predictors of translation initiation. This algorithm was designed to predict the ability of a codon to initiate translation based on the similarity of its surrounding sequence profile to the Kozak consensus sequence (KCS). The KCS is a nucleotide motif, identified to most frequently border the canonical translation initiation codon (ATG) and optimize translation initiation at the site. Although there exist slight variations, this motif is typically accepted as the conserved pattern of the following underlined nucleotides bordering the AUG codon: CCRCCAUGG. The nucleotide designated by R is a purine, most typically adenine [13].

The sequence logo of the KCS (Fig 1) has been used to produce weighted scorings of identified translation initiation codons and observe notable trends. The sequence logo illustrates conserved nucleotides that tend to border ATG codons that initiate translation. The vertical length of each letter in the sequence logo is related to the observed probability for a particular nucleotide to be at a certain position, as well as the impact of the position on the efficiency of translation initiation. It is formulated by the Shannon method [24].

Fig 1. Schematic of the Kozak Similarity Score Algorithm. Based on the sequences flanking an input codon, the algorithm references the KCS Sequence Logo to assign the codon a score.

We designed a weighted scoring algorithm based on the KCS sequence logo and the ten bases preceding and following a codon. Each nucleotide of the 23-base sequence has a value assigned equal to the height of the nucleotide at its respective position, as illustrated in Fig 1. If a nucleotide is not present in a position, it is assigned a value of zero. These values are then summated, and the total divided by the maximal possible summated score (had each nucleotide in the sequence been assigned the largest possible value for its position). This division serves to make final values more feasible for interpretation. As opposed to the pre-normalized score range of about 0 to 0.5990, scores derived from the normalization procedure more conveniently range from 0 to 1. Overall, the final output score, referenced as Kozak similarity score (KSS), of a codon is deduced by expression:

$$KSS(codon) = \frac{1}{KSS_{max}} \sum_{p=1}^{20} bits(nucleotide_p)$$

In this expression, *p* denotes the position of a nucleotide bordering the codon. Values *p=1, 2, 3, …, 10* designate the positions of the ten nucleotides (from left to right) on the left side of the codon, whereas values *p=11, 12, 13, …, 20* designate the positions of ten nucleotides (from left to right) on the right side of the codon. Furthermore, *bits(nucleotide)* is the assigned height of a particular nucleotide with reference to the KCS sequence logo (Fig 1). KSS_{max} is the maximum possible KSS that can be calculated for a codon.

We then used this algorithm on the sequences flanking known instances of ATG translation initiation and produced a histogram distribution of the resulting scores (Fig 2). We created two baselines to compare the scoring of ATG translation initiation codons against ATG codons that do not initiate translation. For the first baseline, we ran the algorithm on one hundred thousand ‘dummy’ ATG codons that had completely randomized sequences without missing nucleotides (a randomized adenine, cytosine, thymine, or guanine in every position flanking the codons) and graphed the resulting score distribution. For the second, we ran the algorithm on a series of ATG codons derived from the human genome that are believed not to initiate translation.

Fig 2. Kozak Similarity Scores of ATG Translation Initiation Codons Against Baseline.

As these histograms were generated from large datasets, they could more accurately serve as representations of algorithm scoring for respective codon classifications: codons that initiate translation, mixture of codons that initiate translation and do not initiate translation, and codons that do not initiate translation, respectively.

Of note, the histogram in Fig 2 representing a randomized combination of codons that initiate and do not initiate translation, is centered at about 0.59 for both the mean and median. In contrast, in the histogram representing ATG codons that initiate translation, we observed a left-skewed distribution, with mean and median scores of about 0.73 and 0.74, respectively. In the histogram representing ATG codons expected not to initiate translation, we observed a slightly right-skewed distribution, with mean and median scores of about 0.52 and 0.53, respectively.

Although exact sequences bordering near-cognate initiation codons have not been identified, as has been carried out for the canonical ATG initiation codon (the KCS), current literature points out similarities between the two sequences. For instance, in a bioinformatics study that analyzed sequences bordering forty-five mammalian near-cognate initiation codons (including CUG, GUG, UUG, AUA, and ACG), a guanine or cytosine has been shown to frequent the -6 position (6 bases upstream of the codon) [25]. As shown in Fig 1, a guanine or cytosine is also most prevalent in the KCS at this position. The same study also noted the presence of a purine (adenine or guanine) in the -3 position from the codon, which are the two most likely nucleotides to occur in the same position of the KCS [25]. In a study of CUG near-cognate codons, those that most frequently initiated translation had an adenine in the -3 position [26]. Although the frequencies of adenine and guanine in the -3 position of the KCS are similar, analysis suggests that adenine is more conserved. For example, if the nucleotide weightings in the KCS are analyzed, adenine is conserved in about 47% of cases at the position versus that of guanine, with about 37% conservation. Both the bioinformatics study as well as a publication analyzing peptide translation from CUG-initiating mRNA constructs show enhanced translation when guanine is at the +4 position (1 base downstream of the initiation codon) [18, 25]. In the KCS, guanine is most conserved at the +4 position as well.

Because of these similarities, we decided to apply the algorithm to score known near-cognate codons that have been shown to initiate translation (Fig 3). Interestingly, distributions of all results are left-skewed, visibly differing from results derived from scoring of ‘dummy’ codons with randomized flanking sequences, as well as codons expected not to initiate translation. In particular, the distribution of scores for known CTG codons has mean and median of about 0.69. The distribution of scores for known GTG codons has mean and median of about 0.69 and 0.70, respectively. And the distribution of scores for known TTG codons has mean and median of about 0.65. These results are an indication that the KSS of near-cognate codons can be used to predict their ability to initiate translation.

Fig 3. Kozak Similarity Scores of Near-Cognate Translation Initiation Codons Against Baseline.

To use the KSS as a predictor of translation initiation ability, a threshold score has to first be determined. In this way, an algorithm could classify codons with a score above the threshold as initiating translation, and below it, not initiating translation. To find the best threshold, virtual simulations were run using different score cutoffs to classify already known ATG initiation codons and ATG codons expected not to initiate translation. Since there are at least 12,603 cases of known ATG initiation codons in contrast to at least 34,097 ATG codons believed not to initiate translation, the data were first balanced. In this way, the cut-off derived would not bias classifications of codons in favor of not initiating translation. Next, all possible cutoff values were set, ranging from 0.580 to 0.700 by increments of 0.001. This range was determined by contrasting distributions in Fig 2. For each of these cutoff values, one thousand simulations were run classifying the data of 12,603 known ATG translation initiation codons on a randomized subset containing 12,603 of the total 34,097 non-initiating ATGs. Errors were averaged for the one thousand runs at each cutoff value. A cutoff of about 0.64 had the most minimized error. When tested on data containing the 12,603 known ATG-initiating codons and randomized 12,603 instances of non-initiating ATGs, the average accuracy of the model was about 79.85%.

The area under receiver operating characteristic (AUROC) score from one of the thousand model simulations (selected at random) was calculated to be 0.876. This score is a useful metric as it indicates the model's discriminatory ability. In the model context, it would correctly assign a greater prediction value for a codon to initiate translation if it indeed were a translation initiation codon 87.6% of the time [27]. A random classifier has a score of 0.5, whereas a perfect classifier has a score of 1.0 [28]. This score is calculated as the area under the ROC curve. This is a graphical illustration of the model's ability to correctly categorize positives (true positive rate) against decreased discrimination (increased false positive rate).

As carried out in the case of ATG, the cumulative data of the CTG, GTG, and TTG codons was used to deduce a cutoff value for the algorithm's scoring of all near-cognate codons. To identify the best cutoff for near-cognate codons, the same simulation process was used as was carried out for ATG codons. Using this simulation method, with balanced near-cognate codon data consisting of equal numbers of positives (near-cognate initiation codons) and negatives (near-cognate codons that do not initiate translation), the best cutoff of the algorithm classification was about 0.61 for near-cognate codons. After a thousand simulations, the algorithm revealed an average accuracy of about 75.60% for classifying near-cognate codons as initiating translation or not initiating translation. The AUROC score calculated from one randomly selected simulation was 0.835.

Fig 4. Error Classifying ATG and Near-Cognate Codon Ability to Initiate Translation Using Kozak Similarity Score.

Fig 5. ROC Curves of the ATG and Near-Cognate Kozak Similarity Score Classifiers. The AUROC score (area under the curve) of the ATG classifier is equal to 0.876. The AUROC score of the near-cognate RFC is equal to 0.835.

KSS as a reference for likelihood of translation initiation

In the previous section, the weighted scoring algorithm based on the KCS was used as a model to classify whether codons could initiate translation. However, one could question whether the scores of the weighting system could also be used as a metric. To investigate this issue, 12,603 instances of ATGs that initiate translation and 34,097 ATGs believed not to initiate translation were compiled. One thousand balanced test datasets, containing the 12,603 positive ATG instances along with randomly sampled negative ATG instances of the same number, were gathered. The average proportion of codons that initiate translation with a KSS exceeding particular values, across all test datasets was determined. These KSS thresholds ranged from zero to one by increments of 0.02. The proportion of ATGs that initiate translation had a positive correlation with the KSS. In other words, a greater proportion of ATGs would initiate translation with an increased score. This score appeared useful since one could approximate the proportion of ATG codons that initiate translation with equal KSSs to a particular codon encountered.

The same evaluation was conducted for near-cognate codons to deduce if there was a similar trend. The procedures previously applied to the ATG data were used for the cumulative total of 2,413 instances of near-cognate codons that initiate translation, and 141,071 instances of near-cognate codons believed not to initiate translation. There was a positive correlation between the proportion of near-cognate codons that initiate translation and the KSS. In fact, the trend was quite similar to that obtained for ATG data. The KSS was not limited as a metric for ATG codons, but could be used to estimate the likelihood of a near-cognate codon to initiate translation as well.

The results of the analysis for ATG and near-cognate codons is shown in the graph and table of Fig 6.

Fig 6. Proportion of ATG and Near-Cognate Codons that Initiate Translation with KSSs Above Certain Values. The graph and table were both generated to depict the same results, evaluated from balanced data, i.e., an equal background proportion of positives and negatives.

Random forest classifiers

A strong and practical approach for identifying translation initiation codons also includes the application of a machine learning model. Machine learning models are powerful, as they can analyze large amounts of complex data, determine patterns and codependences that are difficult to process by a human, and learn from mistakes to improve over time [29]. Although biological pathways are often sophisticated and produce remarkably diverse data, machine learning models can provide direction for such processes that are not completely understood.

We decided to implement a random forest classifier (RFC). This machine learning algorithm typically produces good results with partly missing data, bears little impact from outliers, and mitigates overfitting. Furthermore, the RFC is a highly preferred model in contemporary genomics [30]. The RFC is based on many decision trees, typically generated from large subsets of data. As each decision tree may split data differently in the classification process, the averaging of many such trees reduces variance and helps avoid overfitting. With an overfit model, data inputs that vary slightly from trained data could have volatile classifications that are not reliable. The RFC, which implements the averaging process, may produce greater accuracy than any one decision tree alone [31].

Accordingly, an RFC was implemented as a separate algorithm to elucidate whether codons initiate translation. To create such an algorithm, the feature variables of codons for the RFC to be trained on were first assigned. For an ATG classifier, these variables designated the ten nucleotides that preceded the codon, and ten that followed it. This range was chosen as studies suggest that alterations of bases in some of these positions are highly impactful, and may define whether a flanked codon is an “optimal, strong, [or] moderate” translation initiation site [13-20]. Although secondary structures can influence translation, which are dependent on a number of nucleotides that may far exceed our incorporated range, successful identification of feature patterns may require exceptionally large amounts of training data that are currently unavailable. This is because the number of training samples required to differentiate data increases exponentially as the number of attributes in a model increases [32]. Since five features are needed to designate whether a nucleotide at each position, *n*, is either adenine, guanine, cytosine, thymine, or missing, 5^n distinct data (enough to cover all possible data variations) may be required for a model to best approximate the impact of each nucleotide, for every position that is considered. By having our models trained on a narrowed scope of nucleotides known to influence translation initiation, we sought to optimize predictive power with limited data. For a near-cognate codon classifier, we included additional features to designate the nucleotide in the first base position of the codons (i.e., the underlined: $\underline{C}$TG, $\underline{G}$TG, $\underline{T}$TG). This is because the nucleotide at this position may significantly impact translation initiation from these codons [21, 22].

Using the package, imbalanced-learn, in Python, we created the RFC models [33]. The ATG RFC was trained using an imbalanced set of 12,603 ATG codons known to initiate translation (positives), and 3,433 of 34,097 generated distinct ATG codons that are believed not to initiate translation (negatives). The set of 3,433 negatives consisted of the total of 1,805 sequences that were not missing nucleotides, and 1,628 (i.e., ten percent fewer) randomly sampled negatives of the remaining 31,697 that were missing nucleotides. We left out five percent of the total 3,433 negatives used (172 ATGs that do not initiate translation), as well as the same number of positives (172 ATG translation initiation codons) from the training data to constitute our test dataset. In this way, accuracy would be based on unbiased data that was balanced with 344 combined cases of equally occurring positives and negatives.

The accuracy of the RFC model on the balanced 344 cases was 87.79%. In other words, the algorithm correctly categorized 302 of the 344 ATGs, based on the sequences flanking each codon. This accuracy is high in comparison to the 79.85% accuracy achieved using the KSS-based classifier. We also calculated the area under receiver operating characteristic (AUROC) score of the model to be 0.948, which is high as well. Increasing the parameter value designating the total number of decision trees included in the RFC had no visible effect on model performance. Other parameters were also best left unchanged for optimal predictions.

The same procedure was used to create an RFC for near-cognate codons as carried out for ATG codons, using data available for all near-cognate codons. To prevent imbalanced data bias in the accuracy measurement for the near-cognate RFC, data that was equally representative of all near-cognate codons was set aside to form the test dataset. As the model was trained on CTG, GTG, and TTG initiation codons, twenty positives and negatives were randomly isolated for each of these codons prior to training. When run on this separated, balanced set of 120 data points, the trained near-cognate RFC performed with 85.00% accuracy. The AUROC score of the near-cognate classifier was calculated to be 0.938.

Fig 7. ROC Curves of the ATG and Near-Cognate Random Forest Classifiers. The AUROC score (area under the curve) of the ATG RFC is equal to 0.948. The AUROC score of the near-cognate RFC is equal to 0.938.

Analysis of the TITER neural network as a benchmark

To our knowledge, there exist only two other models for predicting both ATG and near-cognate translation initiation codons. The latest is the TITER machine learning algorithm [11], which addresses limitations of the first model. We analyzed TITER as a benchmark to compare it with the performance of our presented models.

TITER is a deep learning-based framework that predicts whether a codon initiates translation based on the product of two calculations, which is termed TISScore. One constituent is based on the frequency of the codon of interest (e.g., ATG, CTG, GTG, etc.) in the dataset to initiate translation. The second involves the averaging of calculated scores for a codon with flanking sequences across thirty-two neural networks. A large number of neural networks was used as part of a bootstrapping technique to account for training data imbalance.

Although TITER has a high AUROC score of 0.891 [11], ROC curves can present an “overly optimistic” evaluation of a model’s performance “if there is a large skew in the class distribution” [27, 28]. This evaluation is based on the true positive and false positive rates of the model – and an imbalance of positives and negatives may distort its calculation [34]. One questions whether the test sample of the model is skewed as it consists of 767 positive and 9,914 negative samples in total [11]. Although the authors noted special procedures to account for the data imbalance of the training dataset, it is not clear if such procedures were used for the test dataset.

Since TITER was open-source, TITER’s accuracy was averaged across a hundred balanced subsets from its test dataset. Using all 767 positive samples, 767 negatives were randomly sampled from the 9,914 total negatives, across the hundred runs to account for the data imbalance. Through this technique, the unbiased average of the model accuracy was calculated to be 66.94%. This was the accuracy achieved by the best cutoff, 0.5, of the TISScore for classification. When run on the same sequences comprising the RFC test datasets (with sequences extended to include the additional features TITER was trained with), TITER demonstrated 62.21% and 58.33% accuracy for ATG and near-cognate codons, respectively. These values were lower than the 75.60% and 79.85% accuracy achieved using the KSS scoring system for ATG or near-cognate codons, or the 85.00% and 87.79% accuracy achieved using RFC models. The fact that TITER was trained with less data than the RFC models presented here could account for reduced predictive power. Specifically, it was generated using 9,776 positive samples and 94,899 negatives compared to the total 15,016 positives and 175,168 negatives used for the RFCs.

The performance of TITER may also be a result of the large number of features that this machine learning model incorporated. Although contemporary research suggests a few bases that flank a codon greatly influence translation initiation from the site [13-20], TITER analyzes a total of two hundred bases that flank each codon. Compared to our approach of analyzing ten preceding and proceeding nucleotides, TITER may implement up to 180*5 = 900 additional features. The expression ‘180*5’ is used because any one base at the 180 extra positions is represented by five features to designate whether the base is adenine, guanine, cytosine, tyrosine, or is missing. Although the TITER publication mentions feature reduction in the hidden layer of the neural networks, it is not clear how much feature reduction occurred and whether features with significant correlations were inadvertently reduced. An excess of features may decrease effectiveness in machine learning because the number of training samples required to differentiate the data increases exponentially as the number of attributes in a model increases. Thus, predictive power is lost. In fact, this phenomenon is termed the “curse of dimensionality” in Data Science [32].

In addition to feature reduction, our implementation of the random forest classifier, which is more robust to outliers and erroneous instances (especially when data is limited), creation of two models to account for properties of different data types (i.e., ATG codons versus near-cognate codons), and use of sampling *without* replacement which preserves natural variations found in data (in place of bootstrapping) could explain our improved model performance.

Fig 8. ROC Curves of All ATG and Near-Cognate Classifiers Derived from Same Test Data. All classifiers were run on the ATG and Near-cognate RFC test datasets, and their ROC curves were superimposed. The AUROC scores of the ATG and near-cognate RFCs are 0.948 and 0.938, respectively. The AUROC scores of the ATG and near-cognate KSS classifiers are 0.857 and 0.787 on these test datasets. TITER's AUROC scores are 0.622 and 0.603 for ATG and near-cognate codons, respectively.

Model selection and integration into software

Of the two types of models created, the RFCs appeared the best model to use for predicting translation initiation sites. With accuracy determined from the balanced test dataset for ATGs at 87.79% and for near-cognate codons at 85.00%, their performance exceeds that of the straightforward KSS-based classifiers. To our best knowledge, the RFCs also outperform all other models designed for the same function, including TITER, which they exceed by more than 18% in accuracy. As a next step, we decided to use the RFCs to identify repeat-length-independent (RLI) translation initiation associated with neurologic diseases.

To do this, the RFC models were implemented into software. Developed in Python, the program could be used to evaluate a total sequence consisting of the upstream region, followed by ten nucleotide sequence repeats to represent the repeat expansion. Ten sequence repeats may be adequate to capture the repeat expansion effect on translation initiation from upstream codons as well as codons within the repeat expansion itself because ten nucleotide sequence repeats are at minimum thirty bases long, and the integrated model only uses the ten bases that flank each side of a codon for analysis. Nucleotides within this range have been shown to strongly impact translation initiation [13-20].

The model can scan through each codon in the sequence and return a prediction from the implemented RFCs. If a codon encountered is 'ATG,' then the ATG RFC with 87.79% accuracy predicts whether it initiates translation based on the ten sequences flanking each side of the codon. Otherwise, if the codon encountered is a near-cognate codon, then the near-cognate RFC with 85.00% accuracy predicts whether it initiates translation via the same procedure. Next, the program virtually simulates translation from each predicted codon and filters out those instances in which a stop codon (TAG, TGA, or TAA) is encountered upstream of the repeat expansion. This feature was implemented to remove codons from consideration if their initiated translation would not reach the repeat expansion and produce the pathogenic repeat proteins that are associated with neurologic disease. Then, the program would determine the repeated nucleotide sequence that would be translated from each predicted initiation codon, as well as the associated translation product. Finally, the program outputs a visualization of the input sequence, with predicted codons color-coded to distinguish the associated product translated.

In the figures that follow, nucleotides have a bold font to distinguish initiation codons that the software models were trained on. These codons include canonical start codon ATG, and near-cognate codons CTG, GTG, and TTG. Because the features of the three near-cognate codons were used to extrapolate classifications of the other, less researched near-cognate codons (AAG, AGG, ACG, ATC, ATT, and ATA), it is possible to incur false predictions for these less studied instances. Thus, these six near-cognate codons are designated only with color-coding without bolding to denote that they should be acknowledged with less confidence. If there is an overlap between predicted initiation codons (i.e., one or two nucleotides overlap between predicted codons), the color of the overlapped region is the same as that of the next predicted codon to prevent confusion. The overlapped region may or may not be bolded depending on whether the software was trained on this next codon. We also output the KSSs of each predicted codon to two decimal points, as the score could be a useful metric to evaluate translation initiation likelihood. This may be approximated through comparison of KSSs of a codon to the reference table and graph (Fig 6).

Fig 9. An Example of the Formatting Scheme in Software Output. This example shows predicted codons that are color-coded based on their reading frame: 'ATT,' 'TTG,' 'CTG,' 'AGG,' 'GTG,' and 'CTG.' Codons that the models were trained on show up with bold formatting. If there is an overlap between predicted initiation codons (i.e., one or two nucleotides overlap between predicted codons), the color of the overlapped region is the same as the color of the next predicted codon.

Software ability to identify known RLI translation initiation sites

After the software was completed, its ability to distinguish RLI translation initiation sites was analyzed. We first identified translation initiation codons upstream of repeats in the following genes in which RAN translation is known to occur: *C9orf72* (associated with amyotrophic lateral sclerosis and frontotemporal dementia), *FMR1* (associated with fragile X and fragile X-associated tremor/ataxia syndrome), *DM1* (associated with myotonic dystrophy type 1), and

HDL2 (associated with Huntington disease-like 2) genes. These examples were used as references for software performance. It should be noted that translation initiation codons identified for DM1 were obtained from an experiment that implemented a slightly modified version of the conventional DM1 antisense strand. The strand had been experimentally modified to determine whether changes in its sequence could induce translation initiation from particular codons [35]. Next, the associated upstream regions and repeat expansion sequences for each gene, as recorded in the National Center for Biotechnology Information database, were input into the software. Predictions were generated in order to determine whether they corresponded to experimentally confirmed translation initiation codons (Table 1).

Table 1. Previously identified RLI translation initiation sites from publications.

Gene	Codon	Number of Bases Upstream of Repeat	Peptide Repeat Translated	Kozak Similarity Score
***C9orf72* (Sense) [4]**	AGG	1	Poly-GR	0.66
	CTG	24	Poly-GA	0.69
***C9orf72* (Antisense) [4]**	ATG	194	Poly-PG	0.61
***FMR1* (Sense) [36, 37]**	GTG	11	Poly-G	0.70
	ACG	35	Poly-G	0.80
	ACG	60	Poly-G	0.71
***DM1* (Antisense)** with slightly modified sequence **[35]**	ATC	7	Poly-A	0.61
	ATG	17	Poly-S	0.66
	ATT	23	Poly-S	0.74
***HDL2* (Antisense) [35]**	ATC	6	Poly-Q	0.74

Comparison between the predictions and experimentally identified translation initiation codons demonstrated high performance of the software. In fact, all translation initiation sites previously identified across existing publications were correctly identified by the RFCs with one exception: ATC, which was found experimentally to initiate translation in the modified *DM1* antisense strand seven bases upstream of the repeat [35]. However, the near-cognate RFC model successfully predicted all other instances of translation initiation from less researched near-cognate codons. This accuracy is surprising considering that the near-cognate RFC model was only trained on instances of CTG, GTG and TTG translation. As there was insufficient data to train the model on less used near-cognate codons (ATA, ATC, ATT, AGG, ACG, and AAG), predictions for these codons were extrapolated based on recognized patterns from CTG, GTG, and TTG examples. However, for the same reason that they were not included in model training, near-cognate codons that are not CTG, GTG, or TTG should be acknowledged with less confidence in predictions, out of concern they may be false positives.

Predicted Translation Initiation Sites Associated with Neurologic Diseases

Experimentally identified translation initiation codons for *C9orf72*, *FMR1*, *DM1,* and *HDL2* were confirmed by the model presented here (Table 1, Figs 10 and 11). As the software performed well, it was then used to predict translation initiation codons associated with repeats in neurologic diseases that have not been experimentally identified. The software was also used to make predictions for translation initiation codons for other genes with repeats associated with neurologic repeat diseases, *HTT*, and *DM2* (Fig 12). Predicted translation initiation codons with relatively high KSSs were noted for all analyzed genes (Table 2). In all cases, predicted translation initiation sites are not shown if they have a downstream stop codon located in the same reading frame before the repeat.

Fig 10. Predicted Translation Initiation Codons for *C9orf72 and FMR1.* Predicted codons that the models were trained on show up with bold formatting. Numbers indicate the number of bases upstream of the repeat.

* A predicted translation initiation codon overlaps with the repeat (AGG, located 1 base upstream).

Fig 11. Predicted Translation Initiation Codons for *DM1 and HDL2*. Predicted codons that the models were trained on show up with bold formatting. Numbers indicate the number of bases upstream of the repeat.

* Every CTG within the repeat is predicted to possibly initiate translation.

† Every CTG within repeat, aside from the first one, is predicted to possibly initiate translation.

Fig 12. Predicted Translation Initiation Codons for *HTT and DM2.* Predicted codons that the models were trained on show up with bold formatting. Numbers indicate the number of bases upstream of the repeat.

* Every CTG within the repeat, aside from the first one, is predicted to possibly initiate translation.

† Two predicted translation initiation codons are within repeat.

Table 2. Translation Initiation Codons with High Kozak Similarity Scores per Translated Polypeptide Repeat*

Codon	**Number of Bases Upstream of Repeat**	**Kozak Similarity Score**	**Translated Polypeptide Repeat**
C9orf72 (Sense)			
CTG	24	0.66	Poly-GA
AGG	1	0.69	Poly-GR
C9orf72 (Antisense)			
ATG†	113	0.75	Poly-PG
AAG	350	0.84	Poly-PG
ACG	3	0.79	Poly-PR
AAG	288	0.73	Poly-PR
AAG	384	0.77	Poly-PR
FMR1 (Sense)			
AGG	18	0.83	Poly-R
ACG	60	0.71	Poly-R
ACG	35	0.79	Poly-G
GTG	38	0.76	Poly-G
AAG	332	0.83	Poly-G
FMR1 (Antisense)			
AGG	28	0.71	Poly-A
GTG	26	0.73	Poly-R
CTG	56	0.70	Poly-R
ATT	105	0.81	Poly-P
AAG	156	0.78	Poly-P
AAG	177	0.85	Poly-P
CTG	195	0.74	Poly-P
AGG	207	0.84	Poly-P
ATC	252	0.80	Poly-P
AGG	318	0.74	Poly-P
_DM1_ (Sense)			
AAG	23	0.62	Poly-C
AGG	61	0.77	Poly-A
CTG	-1	0.67	Poly-L
_DM1_ (Antisense)			
CTG	34	0.87	Poly-A
AGG	169	0.85	Poly-A
ATC	193	0.81	Poly-A
ACG	98	0.86	Poly-S
HDL2 (Sense)			
ATC	72	0.71	Poly-L
ATC	68	0.52	Poly-C
AGG	10	0.84	Poly-A
HDL2 (Antisense)			
ATC	6	0.74	Poly-Q
AAG	27	0.80	Poly-Q
ATT	261	0.81	Poly-Q
GTG	372	0.83	Poly-Q

	GTG	378	0.71	Poly-Q
	CTG	122	0.68	Poly-S
	ATC	67	0.69	Poly-A
***HTT* (Sense)**				
	AAG	27	0.76	Poly-Q
	CTG	33	0.72	Poly-Q
	CTG	42	0.87	Poly-Q
	ATG	51	0.89	Poly-Q
	AAG	210	0.72	Poly-Q
	CTG	348	0.74	Poly-Q
	ACG	187	0.75	Poly-A
	GTG	202	0.85	Poly-A
***HTT* (Antisense)**				
	ATC	213	0.76	Poly-L
	AGG	225	0.70	Poly-L
	AAG	330	0.73	Poly-L
	CTG	342	0.70	Poly-L
	AGG	369	0.76	Poly-L
	GTG	13	0.84	Poly-A
	GTG	118	0.72	Poly-A
	CTG	199	0.81	Poly-A
	CTG	229	0.71	Poly-A
	GTG	337	0.73	Poly-A
***DM2* (Sense)**				
	CTG	7	0.50	Poly-CLPA
	CTG	-5	0.61	Poly-LPAC
	ATT	87	0.66	Poly-PACL
***DM2* (Antisense)**				
	AGG	7	0.72	Poly-GRQA
	GTG	58	0.70	Poly-GRQA
	ATA	88	0.75	Poly-GRQA
	AGG	47	0.71	Poly-RQAG
	AGG	113	0.74	Poly-RQAG
	AGG	15	0.72	Poly-QAGR

*Predicted codons are displayed that have KSSs above 0.70. If no KSSs within a reading frame are above 0.70, then the codon with the highest KSS is presented – as in the case of the *C9orf72* sense strand.

† Bolded codons represent codons that the RFCs were trained on.

Results displayed in the figures and table indicate translation initiation sites for proteins translated from the repeat. Of note, the average KSS of all upstream predicted codons is about 0.66. With reference to the table in Fig 6, approximately 80% of ATG and near-cognate codons with a score above 0.65 are estimated to initiate translation from a background population of equally occurring translation initiation codons (positives) and codons believed not to initiate translation (negatives).

With respect to the *C9orf72* sense strand upstream from the repeat, the software predicts a codon to initiate translation of poly-GA, and another to translate poly-GR. Both of these codons have been confirmed through experimentation [4]. In the antisense strand, there are ten codons that could initiate translation of poly-PR, and six predicted with respect to poly-PG. The ATG located 194 bases upstream of the repeat expansion has been confirmed [4].

Predictions for translation initiation codons from the *FMR1* sense strand upstream from the repeat identify nine codons that could be used to initiate translation of poly-G, and two for poly-R. The predicted GTG located 11 bases upstream, the ACG located 35 bases upstream, and ACG located 60 bases upstream, have been confirmed experimentally [36]. The antisense upstream region has a total of sixteen codons predicted to initiate translation of poly-P, three for poly-R, and one for poly-A.

For the *DM1* sense strand upstream from the repeat, the software predicts three codons that initiate translation of poly-C, and two that initiate translation of poly-A. Interestingly, every CTG within the CTG repeat expansion is predicted to initiate translation of poly-L; however, only the first has a relatively high KSS (0.67). Predictions for the DM1 antisense strand are different from those produced for the experimentally modified DM1 antisense strand (Table 1). Namely, there is no predicted ATG located 17 bases upstream of the repeat expansion, nor a predicted ATT located 23 bases upstream of the repeat expansion, since sequences that border the predicted codons in the modified strand differ from those bordering the same codons in the unmodified version. In the unmodified antisense strand, there are seven codons predicted to initiate poly-A translation, and one to initiate translation of poly-S. Also, there are no predicted translation initiation codons in the reading frame of poly-Q which suggests that this polypeptide might be initiated from the repeat expansion, possibly by repeat length-dependent folding.

With respect to the *HDL2* sense strand upstream from the repeat, the software predicts seven codons to initiate translation of poly-L, one to initiate translation of poly-C, and two to initiate translation of poly-A. Furthermore, the software suggests that every CTG of the CTG repeat expansion, aside from the first one in the sense strand, can initiate translation of poly-L. In the antisense strand, there are seventeen codons predicted to initiate translation of poly-Q, three for poly-S, and two for poly-A. The predicted ATC located 6 bases upstream of the repeat expansion in the antisense strand has been confirmed [35].

Predictions for translation initiation codons from the *HTT* sense strand upstream from the repeat identify seventeen codons that initiate translation of poly-Q, and four for poly-A. From the antisense upstream region, sixteen codons are predicted to initiate translation of poly-L, and nine for poly-A. The software also suggests that every CTG of the CTG repeat expansion, aside from the first one in the antisense strand can initiate translation of poly-L.

Predictions for the *DM2* sense strand upstream from the repeat identify five codons used for translation initiation of poly-PACL, two for poly-CLPA, and three for poly-LPAC. Moreover, the software predicts the first two CTGs of the CCTG repeat expansion to initiate translation of poly-LPAC. In the antisense strand, there are three codons predicted to initiate poly-RQAG translation, five to initiate translation of poly-GRQA, and one to initiate translation of poly-QAGR.

Discussion

As shown here, RFCs were able to successfully predict most translation initiation codons associated with neurologic repeat expansion diseases that were experimentally identified. The same models also predicted other codons to initiate translation of repeat expansions for neurologic diseases, that have not been identified. Of note, this software predicted translation initiation sites with more than 18% accuracy than the TITER neural network.

Regardless of the quality of a model, its predictions should not be interpreted as evidence. Instead, predictions should be recognized as likely possibilities that warrant further investigation. The significance of the algorithm's identification of translation initiation codons, however, should not be understated. For example, these data may be important to use to guide treatment of these repeat diseases.

Although the machine learning models show promise in understanding of the pathogenesis of repeat expansion neurologic disorders, their use may be extended to other applications as well. For example, they may be used to predict the translation initiation that are not involved in repeat expansion disorders. One benefit of this implementation includes the ability to speculate protein products from a nucleotide sequence, quickly and easily and without laboratory procedures. In order to accelerate the use of the RFCs, a version of the machine learning software that can predict translation initiation codons in any provided sequence is available (at www.tispredictor.com/tis).

Enhancing Performance

Like other machine learning models, RFC performance is determined by the amount of available training data. Because of this constraint, collecting more examples to train the machine learning models could prove especially useful. In the case of the near-cognate RFC, obtaining sufficient data to account for all near-cognate types could lessen uncertainty in predictions involving these codons. Training the two RFCs discussed here with more of the codon types that have been used would be beneficial since feeding a model with more data will verify existing trends, and introduce variations that the algorithm can recognize and link to a particular classification, thereby improving accuracy.

Materials and Methods

Data acquisition

Examples of translation initiation were mostly obtained from ribosome profiling, mass spectroscopy, and CRISPR-based techniques across different human cell types and under different conditions [38]. These data include sequences of 12,094 examples of translation initiated from ATG, as well as 2,180 examples of translation initiated from near-cognate codons. Translation initiation sites were also captured by quantitative translation initiation sequencing of genes in cultured human kidney cells [39]. Their annotated sequences were collected from the

Ensembl gene annotation system (version 84) [40]. These methods procured 509 and 203 more examples of ATG and near-cognate initiation codons, respectively. In all, we collected 12,603 instances of translation initiation from ATG, and 2,413 instances of translation initiation from near-cognate codons to use in this study.

To obtain examples in which translation does not initiate from ATG (negatives), we used the same transcripts from which positives were derived and recorded nucleotides that flanked ATG codons. Then, we eliminated all instances in which flanking sequences matched any of the 12,603 sequences bordering the known ATG translation initiation sites, leaving 34,097 negatives. We repeated the same procedure to identify negatives for near-cognate codons that do not initiate translation. We found examples of CTG, GTG, and TTG codons in which flanking sequences did not match any of that of the known near-cognate initiation codons, leaving 141,071 negatives.

Random Sampling

All random sampling was conducted without replacement. This method is preferred for KSS evaluations of ATG and near-cognate codons, as the precision of population estimates is higher than that produced by sampling with replacement [41]. Furthermore, sampling without replacement to generate training datasets introduces greater variation for model training.

Random forest classifiers

Using the open-source package, imbalanced-learn, in Python, we created the RFC models [33]. The ATG RFC was trained on an imbalanced set of 12,432 ATG codons known to initiate translation (positives), and 3,261 ATG codons that are believed not to initiate translation (negatives). The set of 3,261 negatives consisted of 1,716 sequences that were not missing nucleotides, and 1,545 (ten percent fewer) randomly sampled negatives of the remaining 31,697 that were missing nucleotides. To clarify, missing nucleotides are registered in the case that a recorded codon is located exceedingly close to the 5' or 3' terminus of an mRNA construct. In such a circumstance, there may not be a full ten bases both preceding and following the codon. The sampling technique was performed to slightly offset the proportion of negatives with and without missing bases in the opposite direction. In this way, more negatives without missing bases would be used for model training. Using the original imbalanced set of negatives, with the majority missing bases, would cause the model to inaccurately assess the effect of missing nucleotides on a codon's ability to initiate translation. Furthermore, using a slightly larger proportion of negatives that had a complete sequence profile resulted in improved accuracy for distinguishing codons that were not missing nucleotides. This is useful, as sequences are less often encountered with missing nucleotides in real-world applications.

To account for the imbalance of positives and negatives, the RFC had decision trees generated from 3,576 negatives, and the same number of randomly sampled positives. One thousand such trees were used, since this number is generally recommended as a starting point for the generation of an RFC [42]. Of the total number of features, n, a total of $\sqrt{n}$ features were used to classify the data in order to optimize predictive power. Training with too many or too few features could have prevented the model from recognizing the best indicators for classification [42]. Each decision tree also had the requirement of grouping at least two codon instances to a certain classification. This constraint reduced the risk of overfitting, yet still allowed tree capacity to differentiate between subtly differing codons. Thus, the trees could better identify precise feature patterns to associate with a particular classification, and remain reliable in face of new, unencountered data.

We evaluated the accuracy of the RFC model with the above configurations. Parameters such as the minimum number of codons to group for classification could then be adjusted to improve predictive power, as necessary. However, parameters were best left unchanged for optimal predictions. To create a separate classifier for near-cognate codons, we repeated the same procedures to create an RFC for near-cognate codons as we had carried out for ATG codons, this time using data available for all near-cognate codons.

Accessibility and implementation

The software is publicly accessible as an interactive website at www.tispredictor.com.

[Availability Statement for Open Access Models and Data]

References

1. La Spada AR, Taylor JP. Repeat expansion disease: Progress and puzzles in disease pathogenesis. Nat Rev Genet. 2010;11(4):247-58.

2. Davis M, Stroud C, editors. Neurodegeneration: Exploring Commonalities Across Diseases: Workshop Summary. Forum on Neuroscience and Nervous System Disorders; 2013; Washington (DC): National Academies Press (US).

3. Rudich PD, Watkins S, Lamitina T. PolyQ-independent toxicity associated with novel translational products from CAG repeat expansions. PLoS One. 2020;15(4).

4. Boivin M, Pfister V, Gaucherot A, Ruffenach F, Luc N, Sellier C, et al. Reduced autophagy upon C9ORF72 loss synergizes with dipeptide repeat protein toxicity in G4C2 repeat expansion disorders. EMBO J. 2020;39(4):e100574.

5. Boivin M, Deng J, Pfister V, Grandgirard E, Oulad-Abdelghani M, Mortlet B, et al. Translation of GGC repeat expansions into a toxic polyglycine protein in NIID defines a novel class of human genetic disorders: The polyG diseases. Neuron. 2021;109(11):1825-35.

6. Lee S-J, Hee-Sun L, Masliah E, Lee H-J. Protein aggregate spreading in neurodegenerative diseases: Problems and perspectives. Neurosci Res. 2011;70(4):339-48.

7. Monaco A, Fraldi A. Protein Aggregation and Dysfunction of Autophagy-Lysosomal Pathway: A Vicious Cycle in Lysosomal Storage Diseases. Frontiers in Molecular Neuroscience. 2020.

8. Ross CA, Poirier MA. Protein aggregation and neurodegenerative disease. Nat Med. 2004;10 Suppl:S10-7. Epub 2004/07/24. doi: 10.1038/nm1066. PubMed PMID: 15272267.

9. Chung CG, Lee H, Lee SB. Mechanisms of protein toxicity in neurodegenerative diseases. Cellular and Molecular Life Sciences. 2018;75(17):3159-80. doi: 10.1007/s00018-018-2854-4.

10. Krans A, Skariah G, Zhang Y, Bayly B, Todd PK. Neuropathology of RAN translation proteins in fragile X-associated tremor/ataxia syndrome. Acta Neuropathologica Communications. 2019;7(1).

11. Zhang S, Hu H, Jiang T, Zhang L, Zeng J. TITER : predicting translation initiation sites by deep learning. Bioinformatics. 2017:33:i234–i42.

12. Reuter K, Biehl A, Koch L, Helms V. PreTIS: A Tool to Predict Non-canonical 5' UTR Translational Initiation Sites in Human and Mouse. PLOS Computational Biology. 2016;12(10):e1005170. doi: 10.1371/journal.pcbi.1005170.

13. Hernández G, Osnaya VG, Pérez-Martínez X. Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. Trends in Biochemical Sciences. 2019;44(12):1009-21.

14. Meijer HA, Thoma AAM. Control of eukaryotic protein synthesis by upstream open reading frames in 5′-untranslated region of an mRNA. Biochem. 2002;367:1-11.

15. Pisarev AV, Kolupaeva VG, Pisareva VP, Merrick WC, Hellen CUT, Pestova TV. Specific functional interactions of nucleotides at key −3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. Genes Dev. 2006;20:624-36.

16. Lütcke HA, Chow KC, Mickel FS, Moss KA, Kern HF, Scheele GA. Selection of AUG initiation codons differs in plants and animals. Embo J. 1987;6(1):43-8.

17. Kozak M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Bio. 1987;196(4):947-50.

18. Kozak M. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. Embo J. 1997;16(9):2482-92.

19. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 1986;44(2):283-92.

20. Kozak M. Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. Mol Cell Biol. 1989;9(11):5073-80.

21. Wei J, Zhang Y, Ivanov IP, Sachs MS. The stringency of start codon selection in the filamentous fungus Neurospora crassa. J Biol Chem. 2013;288(13):9549-62. Epub 2013/02/08. doi: 10.1074/jbc.M112.447177. PubMed PMID: 23396971.

22. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 2017;31(17):1717-31. Epub 2017/10/07. doi: 10.1101/gad.305250.117. PubMed PMID: 28982758; PubMed Central PMCID: PMCPMC5666671.

23. Libbrecht MW, Noble WS. Machine learning in genetics and genomics. Nat Rev Genet. 2015;16(6):321-32.

24. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Research. 1990;18(20):6097-100.

25. Wegrzyn J. L. DTM, Valafar F., Hook V. Bioinformatic analyses of mammalian 5'-UTR sequence properties of mRNAs predicts alternative translation initiation sites. BMC Bioinformatics. 2008.

26. Schwab SR, Shugart JA, Horng T, Malarkannan S, Shastri N. Unanticipated Antigens: Translation Initiation at CUG with Leucine. PLoS Biology. 2004;2(11):e366.

27. Davis J, Goadric M. The relationship between Precision-Recall and ROC curves. ICML. 2006:233-40.

28. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015.

29. Libbrecht M. W. NWS. Machine learning in genetics and genomics. Nat Rev Genet. 2015;16(6):321-32.

30. Goldstein BA, Polley EC, Briggs FBS. Random Forests for Genetic Association Studies. Stat Appl Genet Mol Biol. 2011;10(1):32.

31. Liu Y, Wang Y, Zhang J. New Machine Learning Algorithm: Random Forest. ICICA. 2012;7473:246-52.

32. Friedman JH. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. Data Mining and Knowledge Discovery volume. 1997;1(1):55–77.

33. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. JMLR. 2017;18(17):1-5.

34. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Pattern Recognition Letters. 2004;31(8):1-38.

35. Zu T, Gibbens B, Doty NS, Gomes-Pereira M, Huguet A, Stone MD, et al. Non-ATG-initiated translation directed by microsatellite expansions. Proc Natl Acad Sci U S A. 2010;108(1):260-5.

36. Rodriguez CM, Wright SE, Kearse MG, Haenfler JM, Flores BN, Liu Y, et al. A native function for RAN translation and CGG repeats in regulating fragile X protein synthesis. Nature Neuroscience. 2020;23:386–97.

37. Kearse MG, Green KM, Krans A, Rodriguez CM, Linsalata AE, Goldstrohm AC, et al. CGG Repeat associated non-AUG translation utilizes a cap-dependent, scanning mechanism of initiation to produce toxic proteins. Mol Cell. 2016;62(2):314-22.

38. Chen J, Brunner A-D, Cogan JZ, Nunez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367:1140–6.

39. Gao X, Wan J, Liu B, Ma M, Shen B, Qian S-B. Quantitative profiling of initiating ribosomes in vivo. Nature Methods. 2015;12(2):147-53.

40. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. Database (Oxford). 2016.

41. Seth GR, Rao JNK. On the Comparison between Simple Random Sampling with and without Replacement. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002). 1964;26(1):85-6.

42. Johnson K, Kuhn M. Applied Predictive Modeling. New York: Springer; 2013.

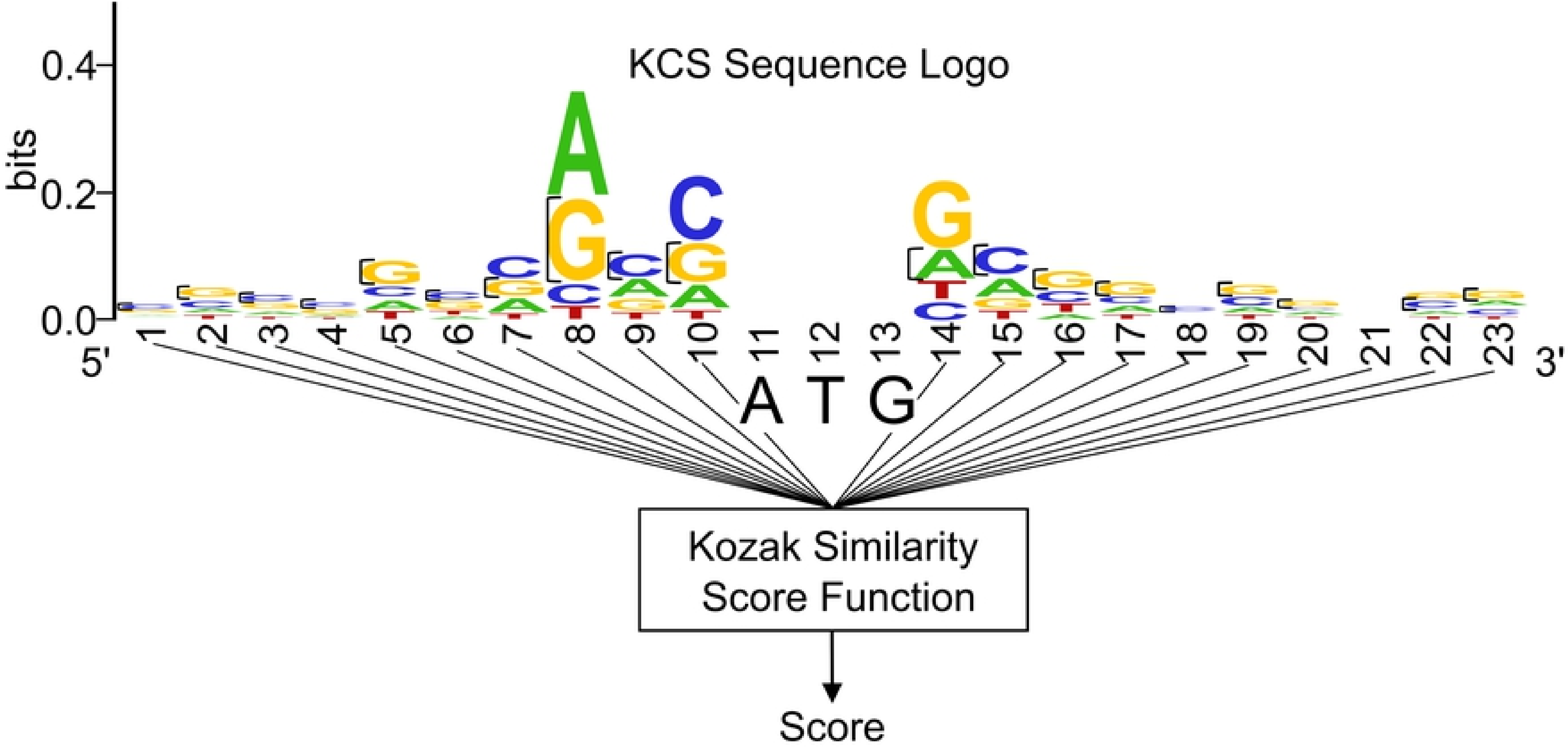

KCS Sequence Logo
bits
0.4
0.2
0.0
5'
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
3'
A T G
Kozak Similarity Score Function
Score

Figure 1

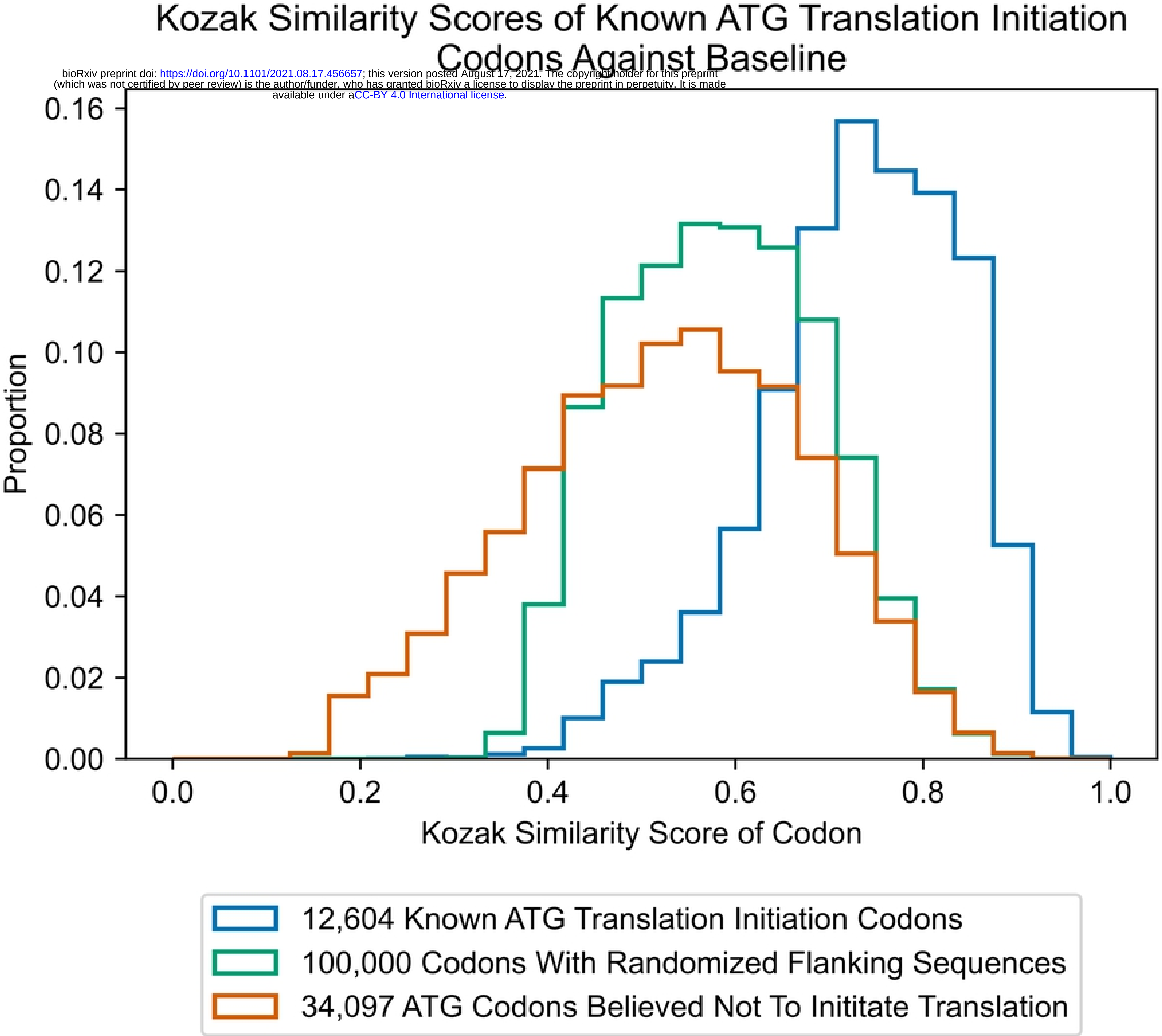

Kozak Similarity Scores of Known ATG Translation Initiation Codons Against Baseline
Proportion
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Kozak Similarity Score of Codon
12,604 Known ATG Translation Initiation Codons
100,000 Codons With Randomized Flanking Sequences
34,097 ATG Codons Believed Not To Inititate Translation

Figure 2

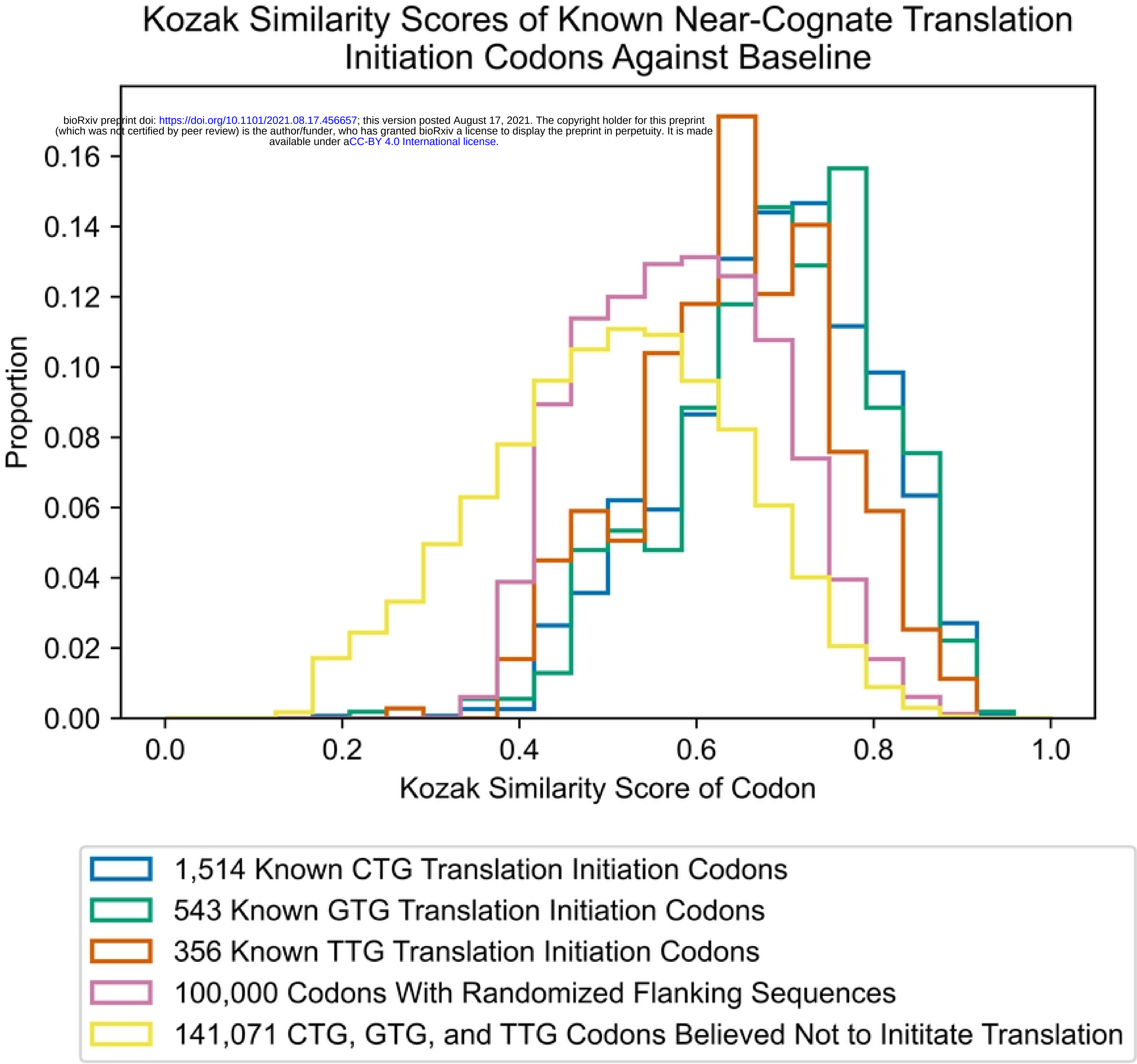

Kozak Similarity Scores of Known Near-Cognate Translation Initiation Codons Against Baseline
Proportion
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.0
0.2
0.4
0.6
0.8
1.0
Kozak Similarity Score of Codon
1,514 Known CTG Translation Initiation Codons
543 Known GTG Translation Initiation Codons
356 Known TTG Translation Initiation Codons
100,000 Codons With Randomized Flanking Sequences
141,071 CTG, GTG, and TTG Codons Believed Not to Inititate Translation

Figure 3

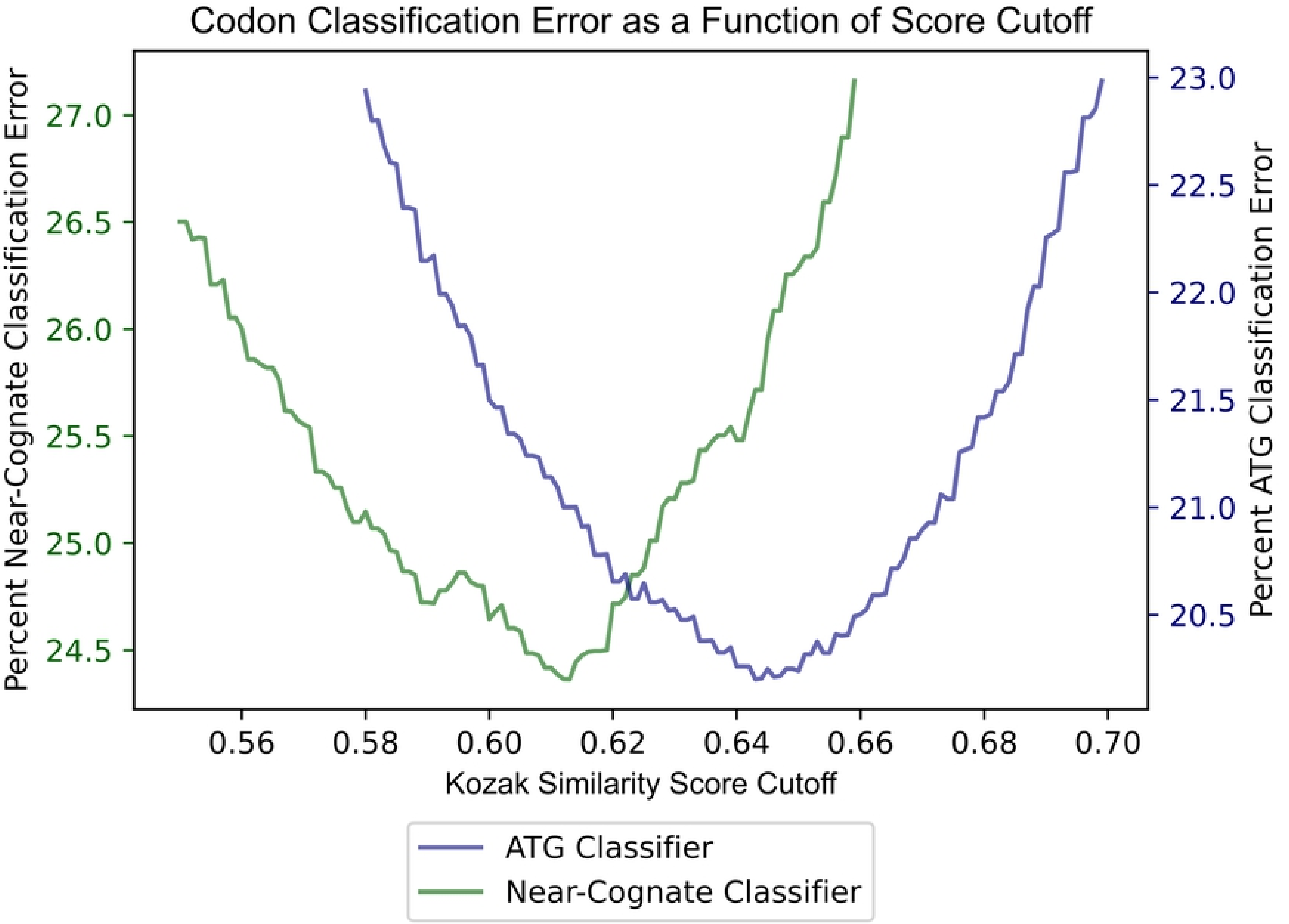

Codon Classification Error as a Function of Score Cutoff
Percent Near-Cognate Classification Error
27.0
26.5
26.0
25.5
25.0
24.5
Percent ATG Classification Error
23.0
22.5
22.0
21.5
21.0
20.5
0.56
0.58
0.60
0.62
0.64
0.66
0.68
0.70
Kozak Similarity Score Cutoff
ATG Classifier
Near-Cognate Classifier

Figure 4

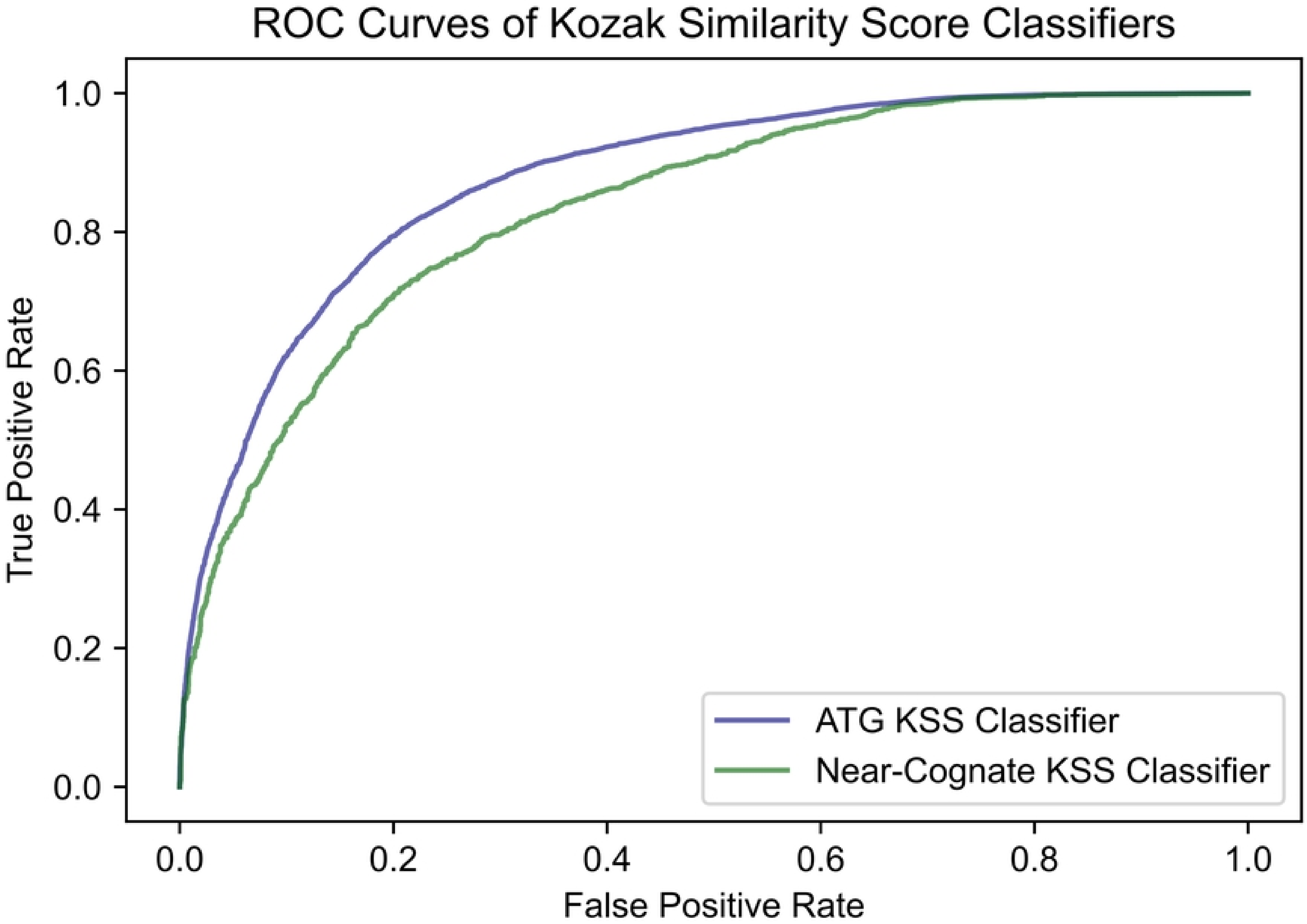
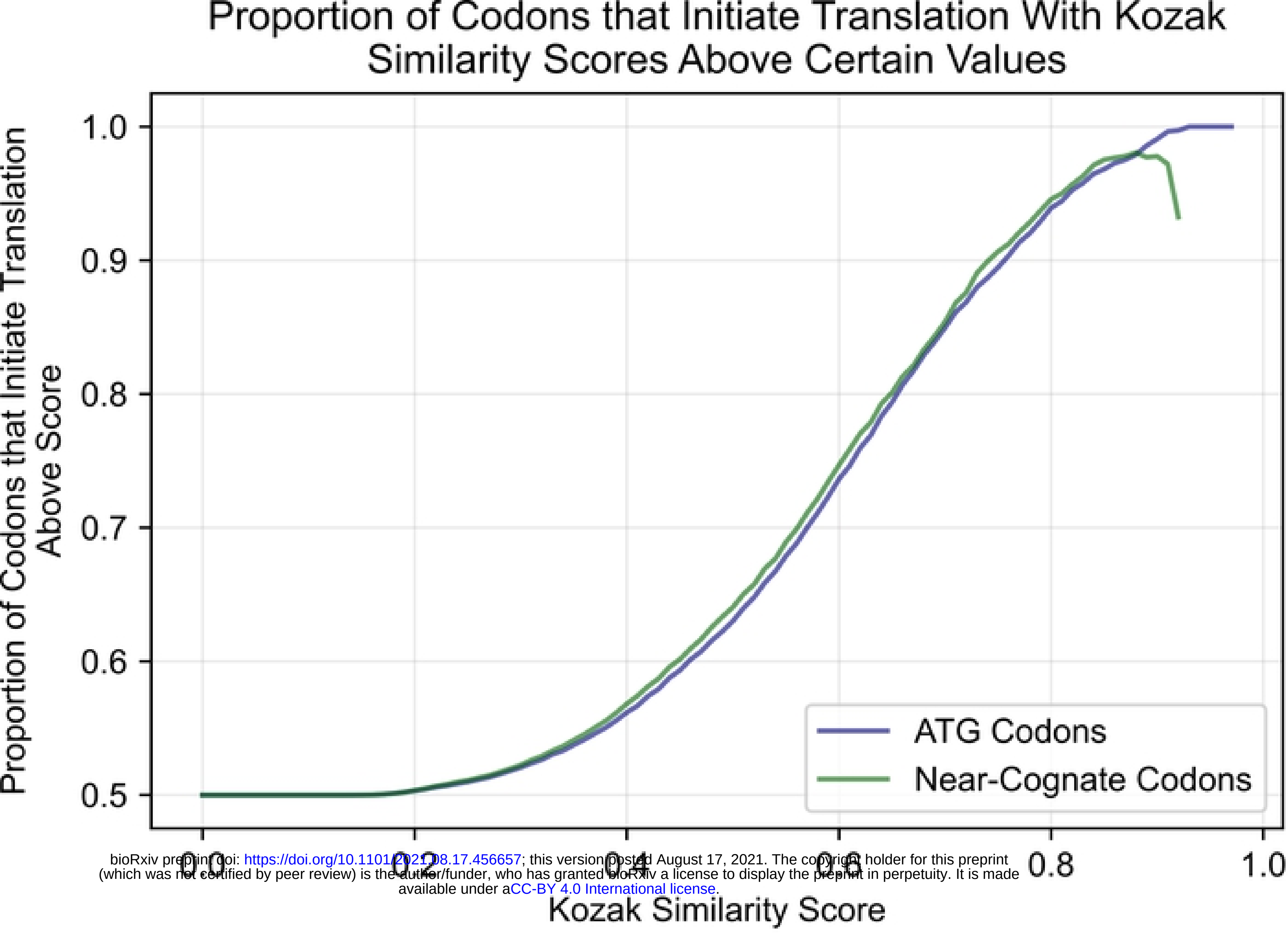

ROC Curves of Kozak Similarity Score Classifiers
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ATG KSS Classifier
Near-Cognate KSS Classifier

Figure 5



Kozak Similarity Score	**Proportion of ATGs that Initiate Translation above Kozak Similarity Score**	**Proportion of Near-Cognate Codons that Initiate Translation Above Kozak Similarity Score**
0.00	0.5000	0.5000
0.05	0.5000	0.5000
0.10	0.5000	0.5000
0.15	0.5001	0.5000
0.20	0.5033	0.5035
0.25	0.5098	0.5110
0.30	0.5206	0.5224
0.35	0.5375	0.5412
0.40	0.5618	0.5683
0.45	0.5933	0.6015
0.50	0.6302	0.6406
0.55	0.6789	0.6893
0.60	0.7364	0.7469
0.65	0.7938	0.8011
0.70	0.8496	0.8541
0.75	0.8948	0.9067
0.80	0.9393	0.9458
0.85	0.9682	0.9753
0.90	0.9908	0.9779
0.95	1.0000	No codons above score

Figure 6

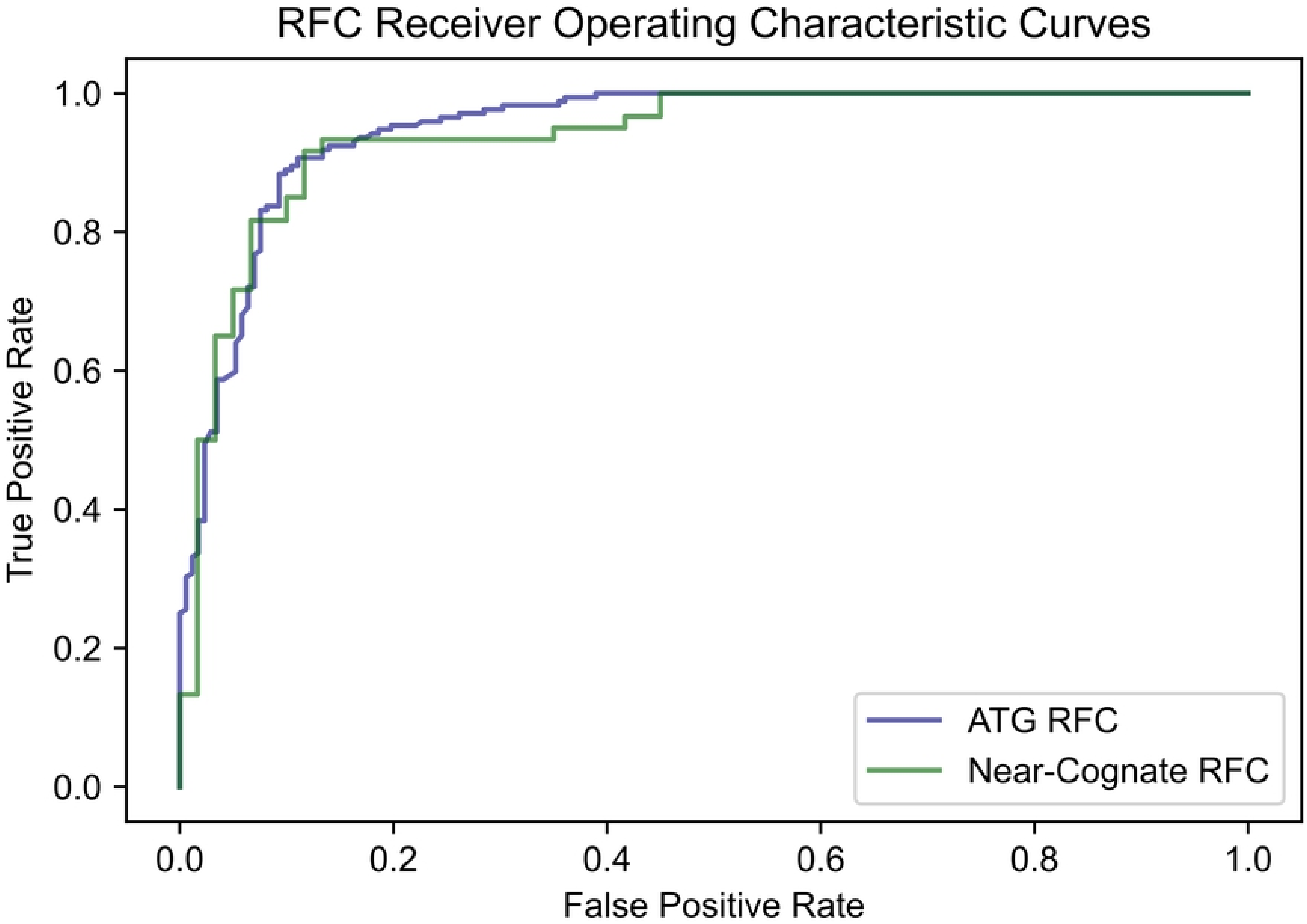

RFC Receiver Operating Characteristic Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ATG RFC
Near-Cognate RFC

Figure 7

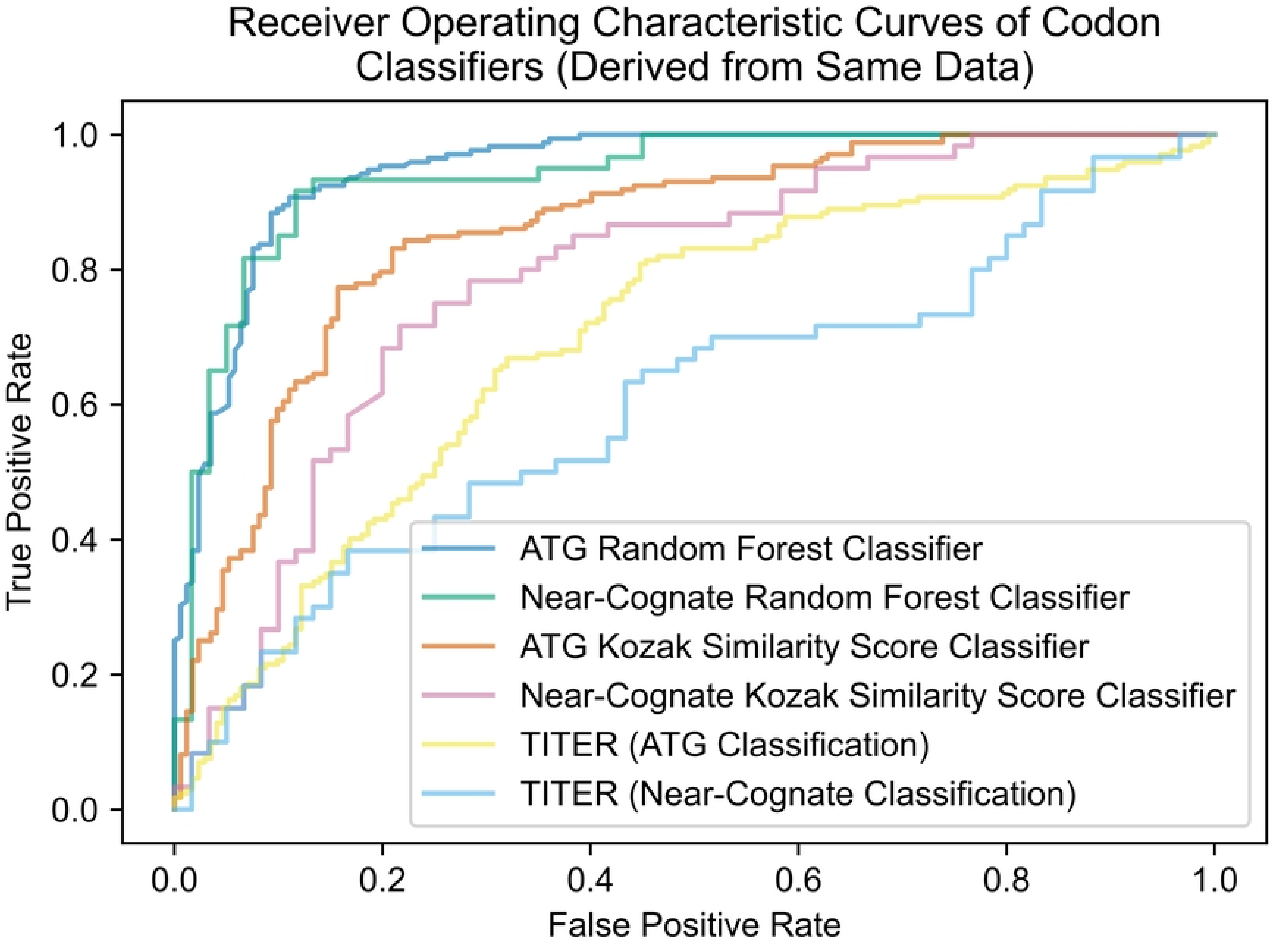
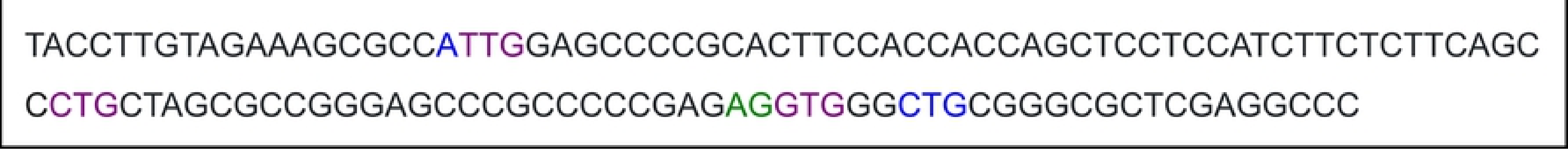

Receiver Operating Characteristic Curves of Codon Classifiers (Derived from Same Data)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ATG Random Forest Classifier
Near-Cognate Random Forest Classifier
ATG Kozak Similarity Score Classifier
Near-Cognate Kozak Similarity Score Classifier
TITER (ATG Classification)
TITER (Near-Cognate Classification)

Figure 8



TACCTTGTAGAAAGCGCCATTGGAGCCCCGCACTTCCACCACCAGCTCCTCCATCTTCTCTTCAGC
CCTGCTAGCGCCGGGAGCCCGCCCCCGAGAGGTGGGCTGCGGGCGCTCGAGGCCC

Figure 9

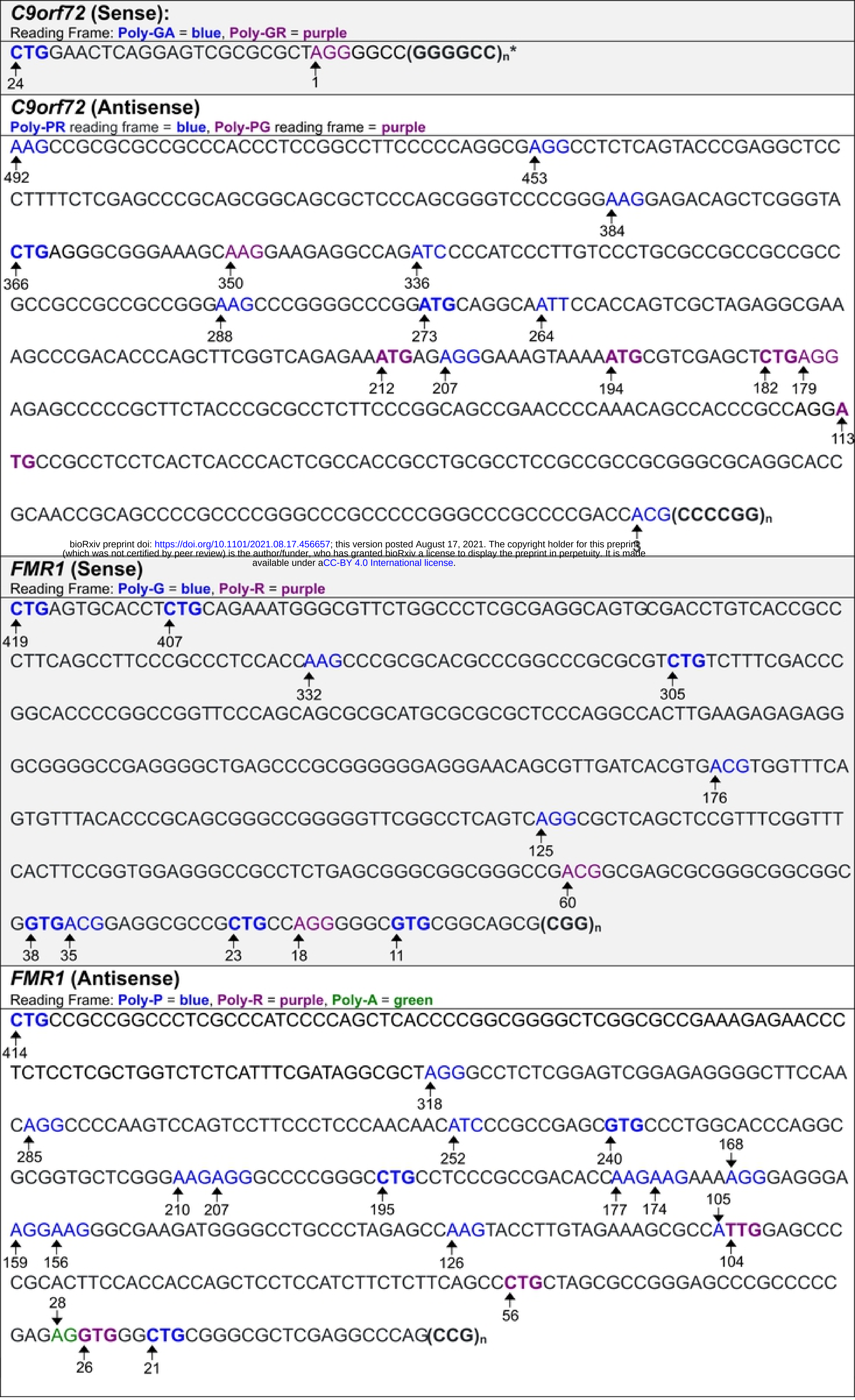

C9orf72 (Sense):
Reading Frame: Poly-GA = blue, Poly-GR = purple
CTGGAACTCAGGAGTCGCGCGCTAGGGGCC(GGGGCC)n*
24
1
C9orf72 (Antisense)
Poly-PR reading frame = blue, Poly-PG reading frame = purple
AAGCCGCGCGCCGCCCACCCTCCGGCCTTCCCCCAGGCGAGGCCTCTCAGTACCCGAGGCTCC
492
453
CTTTTCTCGAGCCCGCAGCGGCAGCGCTCCCAGCGGGTCCCCGGGAAGGAGACAGCTCGGGTA
384
CTGAGGGCGGGAAAGCAAGGAAGAGGCCAGATCCCCATCCCTTGTCCCTGCGCCGCCGCCGCC
366
350
336
GCCGCCGCCGCCGGGAAGCCCGGGGCCCGGATGCAGGCAATTCCACCAGTCGCTAGAGGCGAA
288
273
264
AGCCCGACACCCAGCTTCGGTCAGAGAAATGAGAGGGAAAGTAAAAATGCGTCGAGCTCTGAGG
212
207
194
182
179
AGAGCCCCCGCTTCTACCCGCGCCTCTTCCCGGCAGCCGAACCCCAAACAGCCACCCGCCAGGA
113
TGCCGCCTCCTCACTCACCCACTCGCCACCGCCTGCGCCTCCGCCGCCGCGGGCGCAGGCACC
GCAACCGCAGCCCCGCCCCGGGCCCGCCCCCGGGCCCGCCCCGACCACG(CCCCGG)n
3
FMR1 (Sense)
Reading Frame: Poly-G = blue, Poly-R = purple
CTGAGTGCACCTCTGCAGAAATGGGCGTTCTGGCCCTCGCGAGGCAGTGCGACCTGTCACCGCC
419
407
CTTCAGCCTTCCCGCCCTCCACCAAGCCCGCGCACGCCCGGCCCGCGCGTCTGTCTTTCGACCC
332
305
GGCACCCCGGCCGGTTCCCAGCAGCGCGCATGCGCGCGCTCCCAGGCCACTTGAAGAGAGAGG
GCGGGGCCGAGGGGCTGAGCCCGCGGGGGGAGGGAACAGCGTTGATCACGTGACGTGGTTTCA
176
GTGTTTACACCCGCAGCGGGCCGGGGGTTCGGCCTCAGTCAGGCGCTCAGCTCCGTTTCGGTTT
125
CACTTCCGGTGGAGGGCCGCCTCTGAGCGGGCGGCGGGCCGACGGCGAGCGCGGGCGGCGGC
60
GGTGACGGAGGCGCCGCTGCCAGGGGGCGTGCGGCAGCG(CGG)n
38
35
23
18
11
FMR1 (Antisense)
Reading Frame: Poly-P = blue, Poly-R = purple, Poly-A = green
CTGCCGCCGGCCCTCGCCCATCCCCAGCTCACCCCGGCGGGGCTCGGCGCCGAAAGAGAACCC
414
TCTCCTCGCTGGTCTCTCATTTCGATAGGCGCTAGGGCCTCTCGGAGTCGGAGAGGGGCTTCCAA
318
CAGGCCCCAAGTCCAGTCCTTCCCTCCCAACAACATCCCGCCGAGCGTGCCCTGGCACCCAGGC
285
252
240
168
GCGGTGCTCGGGAAGAGGGCCCCGGGCCTGCCTCCCGCCGACACCAAGAAGAAAAGGGAGGGA
210
207
195
177
174
105
AGGAAGGGCGAAGATGGGGCCTGCCCTAGAGCCAAGTACCTTGTAGAAAGCGCCATTGGAGCCC
159
156
126
104
CGCACTTCCACCACCAGCTCCTCCATCTTCTCTTCAGCCCTGCTAGCGCCGGGAGCCCGCCCCC
28
56
GAGAGGTGGGCTGCGGGCGCTCGAGGCCCAG(CCG)n
26
21

Figure 10

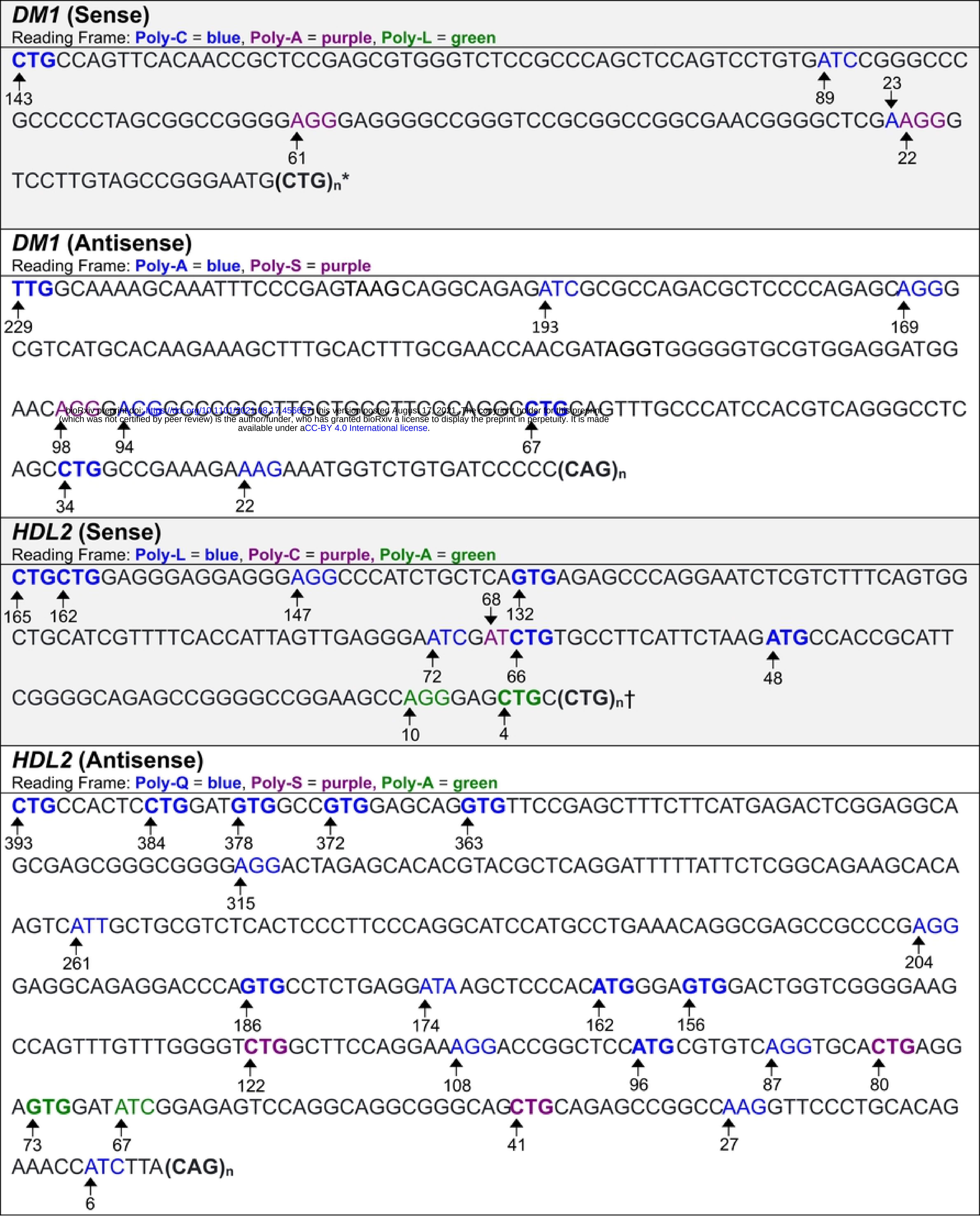

DM1 (Sense)
Reading Frame: Poly-C = blue, Poly-A = purple, Poly-L = green
CTGCCAGTTCACAACCGCTCCGAGCGTGGGTCTCCGCCCAGCTCCAGTCCTGTGATCCGGGCCC
143
89
23
GCCCCCTAGCGGCCGGGGAGGGAGGGGCCGGGTCCGCGGCCGGCGAACGGGGCTCGAAGGG
61
22
TCCTTGTAGCCGGGAATG(CTG)n*
DM1 (Antisense)
Reading Frame: Poly-A = blue, Poly-S = purple
TTGGCAAAAGCAAATTTCCCGAGTAAGCAGGCAGAGATCGCGCCAGACGCTCCCCAGAGCAGGG
229
193
169
CGTCATGCACAAGAAAGCTTTGCACTTTGCGAACCAACGATAGGTGGGGGTGCGTGGAGGATGG
AACAGGGACGGCCCCGGCTTGCTGCCTTGCCCAGGCCTGCAGTTTGCCCATCCACGTCAGGGCCTC
98
94
67
AGCCTGGCCGAAAGAAAGAAATGGTCTGTGATCCCCC(CAG)n
34
22
HDL2 (Sense)
Reading Frame: Poly-L = blue, Poly-C = purple, Poly-A = green
CTGCTGGAGGGAGGAGGGAGGCCCATCTGCTCAGTGAGAGCCCAGGAATCTCGTCTTTCAGTGG
165
162
147
68
132
CTGCATCGTTTTCACCATTAGTTGAGGGAATCGATCTGTGCCTTCATTCTAAGATGCCACCGCATT
72
66
48
CGGGGCAGAGCCGGGGCCGGAAGCCAGGGAGCTGC(CTG)n†
10
4
HDL2 (Antisense)
Reading Frame: Poly-Q = blue, Poly-S = purple, Poly-A = green
CTGCCACTCCTGGATGTGGCCGTGGAGCAGGTGTTCCGAGCTTTCTTCATGAGACTCGGAGGCA
393
384
378
372
363
GCGAGCGGGCGGGGAGGACTAGAGCACACGTACGCTCAGGATTTTTATTCTCGGCAGAAGCACA
315
AGTCATTGCTGCGTCTCACTCCCTTCCCAGGCATCCATGCCTGAAACAGGCGAGCCGCCCGAGG
261
204
GAGGCAGAGGACCCAGTGCCTCTGAGGATA AGCTCCCACATGGGAGTGGACTGGTCGGGGAAG
186
174
162
156
CCAGTTTGTTTGGGGTCTGGCTTCCAGGAAAGGACCGGCTCCATGCGTGTCAGGTGCACTGAGG
122
108
96
87
80
AGTGGATATCGGAGAGTCCAGGCAGGCGGGCAGCTGCAGAGCCGGCCAAGGTTCCCTGCACAG
73
67
41
27
AAACCATCTTA(CAG)n
6

Figure 11

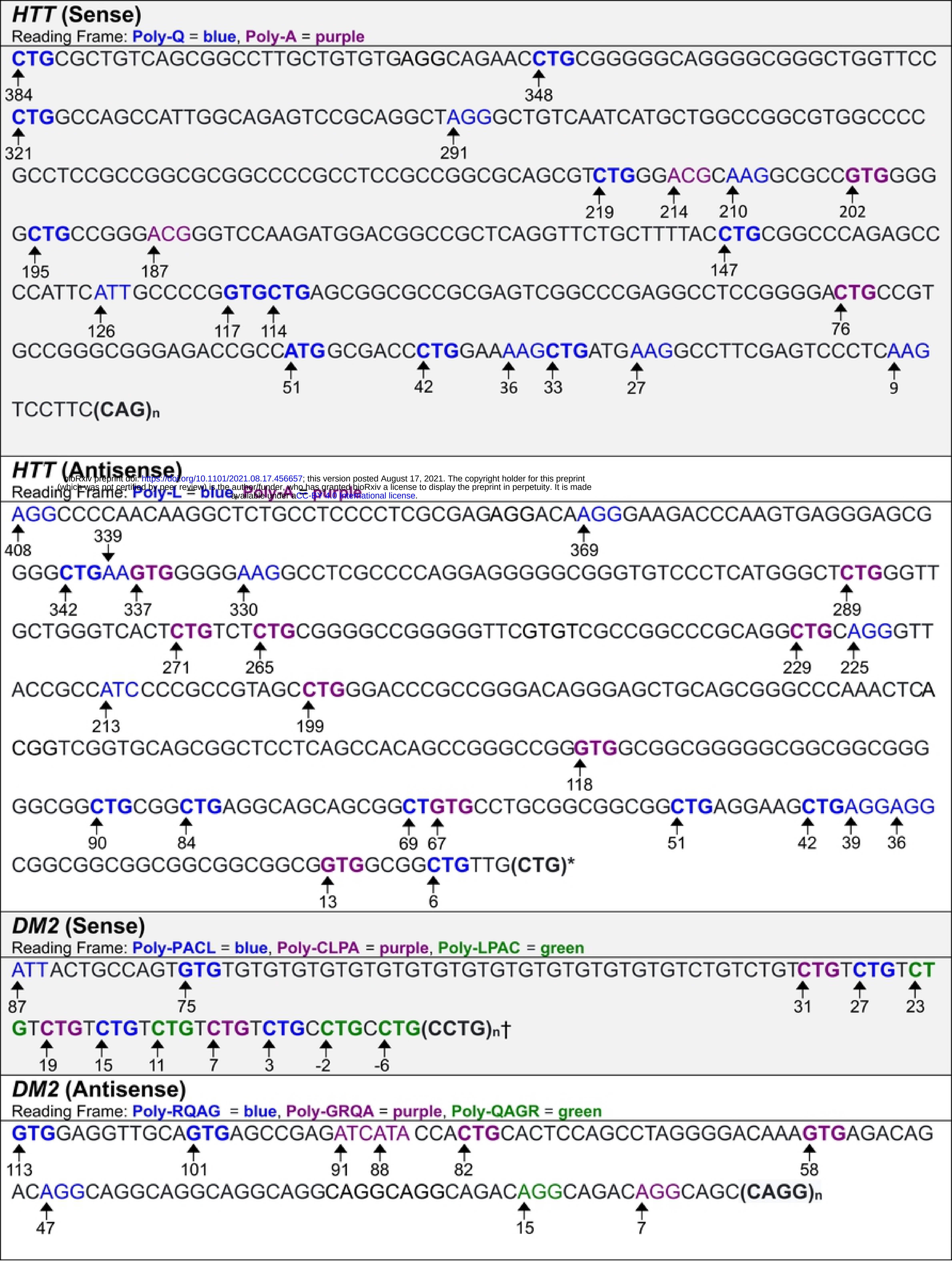

HTT (Sense)
Reading Frame: Poly-Q = blue, Poly-A = purple
CTGCGCTGTCAGCGGCCTTGCTGTGTGAGGCAGAACCTGCGGGGGCAGGGGCGGGCTGGTTCC
384
348
CTGGCCAGCCATTGGCAGAGTCCGCAGGCTAGGGCTGTCAATCATGCTGGCCGGCGTGGCCCC
321
291
GCCTCCGCCGGCGCGGCCCCGCCTCCGCCGGCGCAGCGTCTGGACGCAAGGCGCCGTGGGG
219
214
210
202
GCTGCCGGGACGGGTCCAAGATGGACGGCCGCTCAGGTTCTGCTTTTACCTGCGGCCCAGAGCC
195
187
147
CCATTCATTGCCCCGGTGCTGAGCGGCGCCGCGAGTCGGCCCGAGGCCTCCGGGGACTGCCGT
126
117
114
76
GCCGGGCGGGAGACCGCCATGGCGACCCTGGAAAAGCTGATGAAGGCCTTCGAGTCCCTCAAG
51
42
36
33
27
9
TCCTTC(CAG)n
HTT (Antisense)
Reading Frame: Poly-L = blue, Poly-A = purple
AGGCCCCAACAAGGCTCTGCCTCCCCTCGCGAGAGGACAAGGGAAGACCCAAGTGAGGGAGCG
408
339
369
GGGCTGAAGTGGGGGAAGGCCTCGCCCCAGGAGGGGGCGGGTGTCCCTCATGGGCTCTGGGTT
342
337
330
289
GCTGGGTCACTCTGTCTCTGCGGGGCCGGGGGTTCGTGTCGCCGGCCCGCAGGCTGCAGGGTT
271
265
229
225
ACCGCCATCCCCGCCGTAGCCTGGGACCCGCCGGGACAGGGAGCTGCAGCGGGCCCAAACTCA
213
199
CGGTCGGTGCAGCGGCTCCTCAGCCACAGCCGGGCCGGGTGGCGGCGGGGGCGGCGGCGGG
118
GGCGGCTGCGGCTGAGGCAGCAGCGGCTGTGCCTGCGGCGGCGGCTGAGGAAGCTGAGGAGG
90
84
69
67
51
42
39
36
CGGCGGCGGCGGCGGCGGCGGTGGCGGCTGTTG(CTG)*
13
6
DM2 (Sense)
Reading Frame: Poly-PACL = blue, Poly-CLPA = purple, Poly-LPAC = green
ATTACTGCCAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTCTGTCTGTCTGTCTGTCT
87
75
31
27
23
GTCTGTCTGTCTGTCTGTCTGCCTGCCTG(CCTG)n†
19
15
11
7
3
-2
-6
DM2 (Antisense)
Reading Frame: Poly-RQAG = blue, Poly-GRQA = purple, Poly-QAGR = green
GTGGAGGTTGCAGTGAGCCGAGATCATA CCACTGCACTCCAGCCTAGGGACAAAGTGAGACAG
113
101
91
88
82
58
ACAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGACAGGCAGACAGGCAGC(CAGG)n
47
15
7

Figure 12