On the Relation of Gene Essentiality to Intron Structure: A Computational and Deep Learning Approach

Ethan Schonfeld[1] (Corresponding Author), Edward Vendrow[1], Joshua Vendrow[2], Elan Schonfeld[3]

[1]Stanford University, Stanford, CA, USA
[2]University of California, Los Angeles, Los Angeles, CA, USA
[3]Glenbrook North High School, Northbrook, IL, USA

Abstract

Identification and study of human-essential genes has become of practical importance with the realization that disruption or loss of nearby essential genes can introduce latent-vulnerabilities to cancer cells. Essential genes have been studied by copy-number-variants and deletion events, which are associated with introns. The premise of our work is that introns of essential genes have characteristic properties that are distinct from the introns of nonessential genes. In this paper, we identified several novel properties of introns of essential genes, finding that their structure protects against deletion and intron-loss events, and that these traits are especially dominant in the first intron. We showed that GC density is increased in the first introns of essential genes, allowing for increased enhancer activity, protection against deletions, and improved splice-site recognition. Furthermore, we found that first introns of essential genes are of remarkably smaller size than their nonessential counterparts, and to protect against common 3' end deletion events, essential genes carry an increased number of (smaller) introns. We provided support for our observations by training a deep learning model on introns of essential and nonessential genes and demonstrated that introns alone can be used to classify essential and nonessential genes with high accuracy (AUC of 0.857). We further demonstrated that the accuracy of the same deep-learning model limited to first introns will perform at an increased level, thereby demonstrating the critical importance of introns and particularly first introns in gene essentiality.

Introduction

Essential genes, those where a single-gene-knockout results in lethality or severe loss of fitness, have been well studied in many bacterial genomes to develop therapeutic targets for pathogens. Now, stemming from the discovery that the loss of an essential-nearby gene can introduce latent-vulnerabilities specific to cancer cells, the study of human-essential genes has come of practical importance[1]. This importance is magnified as essential genes for cancer-cell growth are found to be located close to target-deletion genes[1]. Therefore, identifying properties of essential genes can further therapeutic developments.

Older genes, with earlier phyletic origin, are more likely to be essential, as well as genes that are hubs in major protein-protein interaction networks[2,3,4]. Essential genes are highly connected with many protein systems, and thus, consistent transcription timing, maintenance of transcript length, and conservation of gene regulation is of high importance[5]. Identification of human essential genes has been approached through the use of single-gene-knockouts, high-throughput mutagenesis, RNAi, and in most recent work, CRISPR–Cas9 editing[6].

However, moving towards an *in vivo* analysis of gene essentiality, to lend more practical therapeutic insights, studies have focused on the close link between duplication and gene essentiality[7,8]. Duplication is a biological mechanism employed throughout evolution to generate new genetic material[7]. A positive association between singleton, highly-expressed, developmental genes and essentiality is observed, suggesting that essential genes resist duplication events[7,9]. Stemming from these results, copy-number-variants, which result from unequal-crossing-over, retroposition, or chromosomal duplication, were included in efforts to identify essential human genes[1,10]. Intron loss, occurring at an especially greater rate after gene duplication, is the most frequent copy-number-variant in humans, suggesting a likely link between introns and gene-essentiality[11,12].

Introns, which make-up over half of the non-coding genome, have important regulatory and evolutionary functions. Intron losses and deletions can modulate gene expression patterns and even alter gene function[11]. Typically occurring at the 3' end of a gene, losses and deletions arise from mediated recombination of a gene with the reverse-transcribed RNA during duplication events or through irregular splice sites[10,13]. Furthermore, intron deletions are most common to longer introns[12]. Intron 1, typically the longest intron, has frequent intron deletions (30.4% of all known deletions) which are especially serious as the first intron preferentially contains regulatory regions and exhibits the highest density of chromatin marks allowing for gene expression[13,14,15]. GC patterns in intronic sequences are associated with an increase in enhancer activity, correct splice site recognition, and protection from intronic deletions[12,16,17].

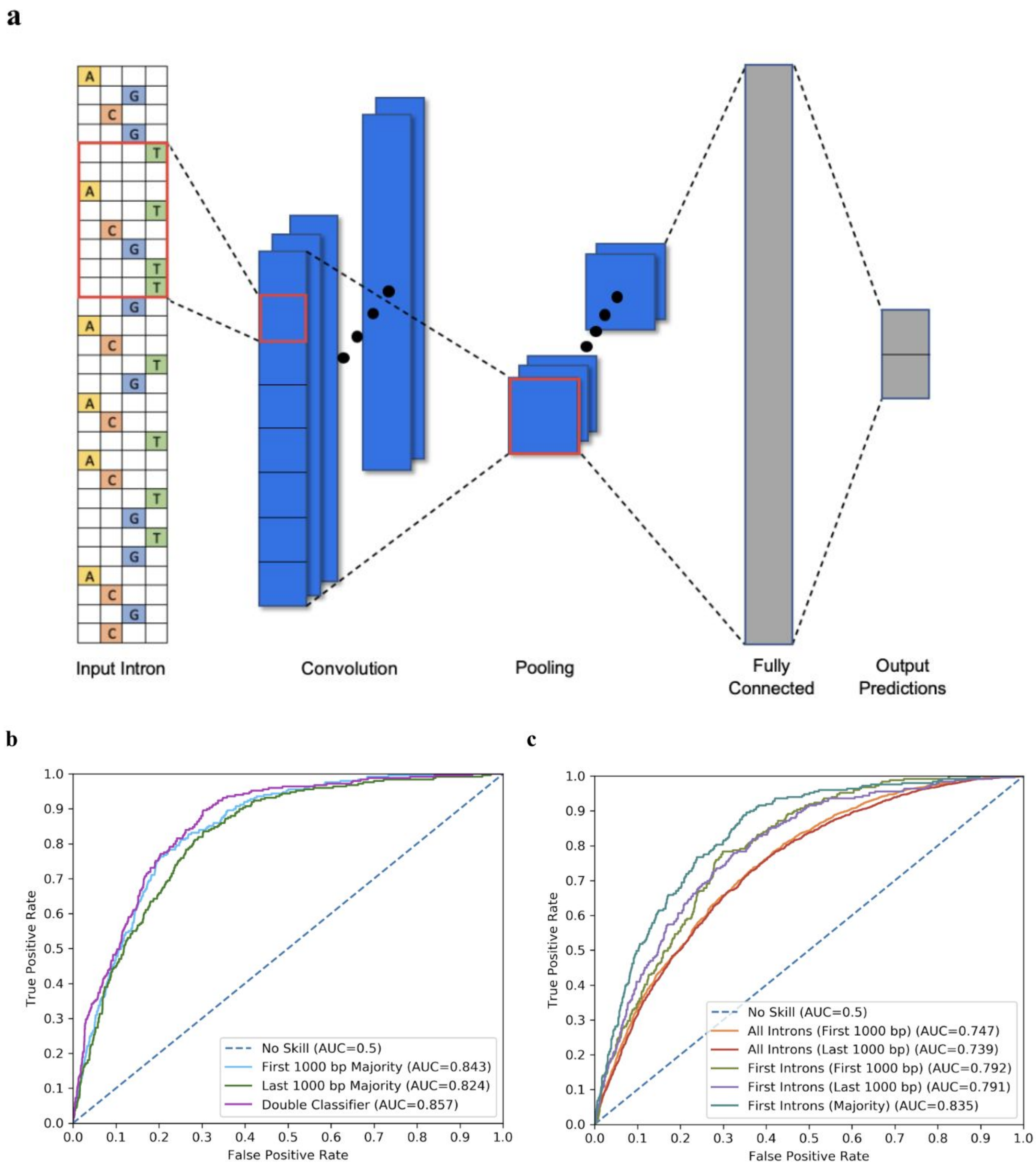
It has been suggested that in highly-expressed-genes, selection has resulted in smaller introns that reduce transcriptional cost, which agrees with reports of shorter introns in essential genes[12,18]. Adding to the seeming importance of introns in essential genes, intron deletions in three-essential-yeast genes drastically decreased RNA levels and caused major growth defects[19].

Owing to the capability of intron losses and deletions to alter gene duplication, expression, and transcription timing, we hypothesize that essential genes, which demand consistency, have developed systems to minimize these events. We thus aim our study to (i) identify whether essential gene introns differ from those of nonessential genes and (ii) characterize the unique properties of essential gene introns to allow for later therapeutic developments.

Results

We extracted 1992 introns of human essential genes, 69371 introns of human conditional genes, and 109640 introns of human nonessential genes from the Ensembl database[20,21]. Human gene essentiality data was gathered from the database of Online Gene Essentiality (OGEE)[3,6]. OGEE gathers data from 18 databases of large-scale experiments; conditional genes are genes where experiments have disagreed on essentiality.

Fig. 1: Details of convolutional neural network and testing results

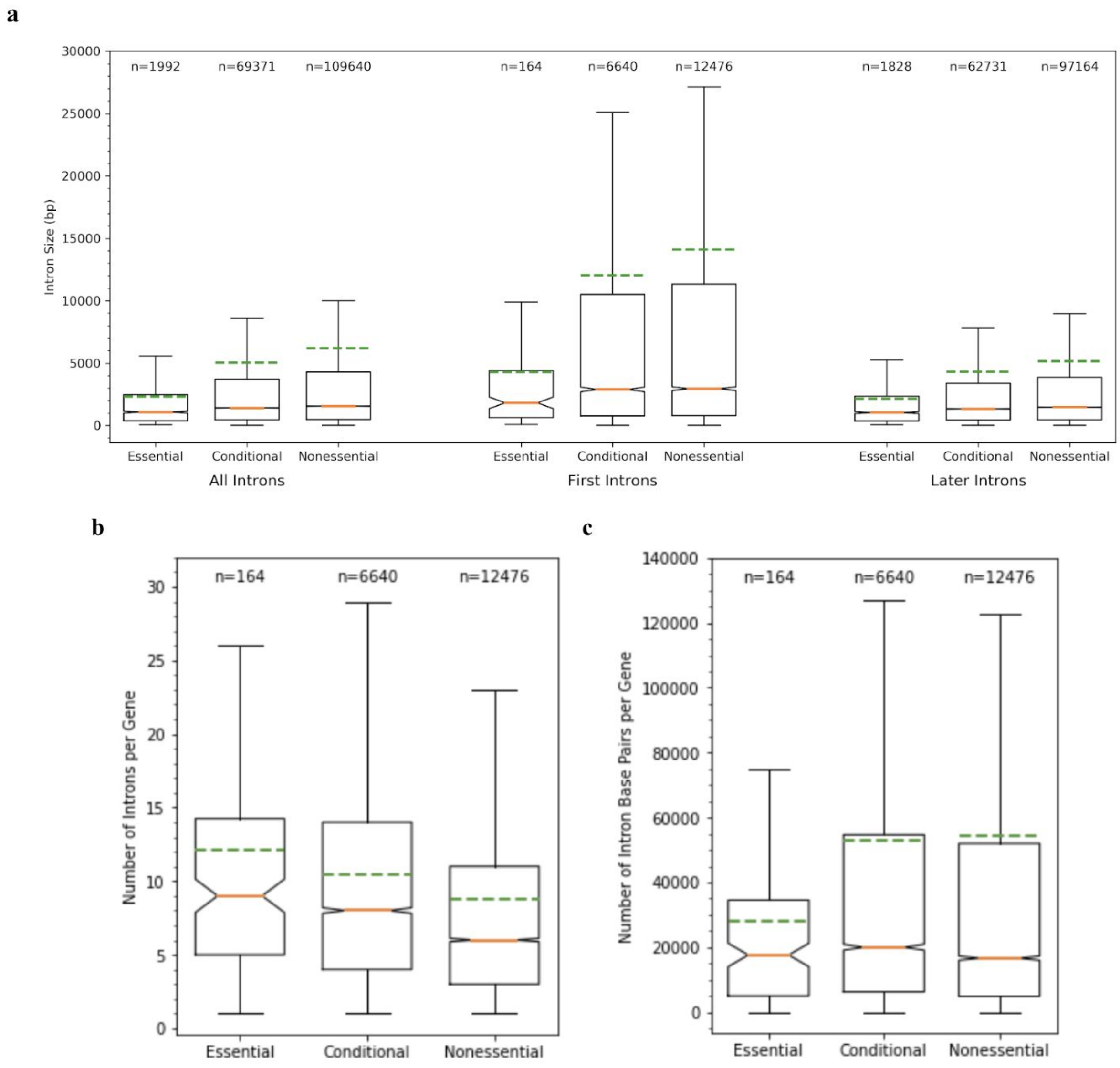


a, Our model uses a convolutional architecture to predict intron essentialities. The convolutional layer contains multiple filters that detect motifs within the intronic sequence. Then, the pooling layer averages each filter's response across the sequence to determine the cumulative presence of motifs. The resulting values are fed into a fully-connected layer followed by a two-value softmax output layer corresponding to the probabilities of the intron being part of an essential or nonessential gene. The best-performing model from our hyperparameter search used 128 convolutional filters with a window size of 24 and a fully connected layer with 128 neurons. We found best results when training with an L2 regularization parameter of 10^{-6} and a dropout rate of 0.2. We trained two models, one on the first 1000 bp of introns and one on the last 1000 bp. This includes the 5' splice site in the first 1000 bp, as well as the 3' splice site and the branch site in the last 1000 bp. In all following results, these models are tested on their respective sections of the intronic sequence. **b,** Our model, trained on the first 1000 bp of introns, had an AUC of 0.747. Our model, trained on the last 1000 bp of introns, had an AUC of 0.739. We predicted gene essentiality using a majority classifier on all introns of a gene. The majority classifier of the model trained on the first 1000 bp of introns saw an AUC of 0.843, and the majority classifier of the model trained on the last 1000 bp of introns saw an AUC of 0.824. We further improved accuracy by averaging the outputs of both majority classifiers. This combined classification strategy achieved an AUC of 0.857. **c,** As the first intron is known to have unique properties, we separately tested the models on only first introns, seeing improved accuracy. On first introns, the model trained on the first 1000 bp of introns had an AUC of 0.792 and the model trained on the last 1000 bp of introns had an AUC of 0.791. We further improved first intron essentiality prediction by averaging the outputs of both models to make a dual average prediction, achieving an AUC of 0.835.

We trained a convolutional neural network, based on DeepBind, to predict gene essentiality based on recurring base-pair motifs of 1000 bp long intronic sequence input[22] (Figure 1). We trained two separate models using the first and last 1000 bp of introns and combined these models by a double classifier which averages essentiality scores from all introns of a gene given by both models. The double classifier optimizes the area under the curve (AUC) of the receiver operating characteristic (ROC) curve used to quantify the diagnostic ability of the model. For the purposes of the neural network, we sought to predict either essentiality or nonessentiality, and thus classified conditional genes from the database as essential if over 50% of experiments agreed on essentiality; we will call these genes conditional–essential. If introns of essential genes and nonessential genes have no markedly-characteristic properties, we would expect an AUC of 0.5. Rather, our double classifier achieved an AUC of 0.838 (Figure 1). We identified the basis used to filter introns, giving a set of 128 sequences of 24 bp long, for each model, that were used to differentiate between essential and nonessential introns (Supplemental Figure 1).

Our results support that introns of essential and nonessential genes have unique properties. To identify unique properties we used a computational approach. We also found that the model performs slightly better at classifying introns of strictly essential or nonessential genes, suggesting that conditional–essential genes do not fit well in either essential or nonessential motifs. Therefore, we now include all OGEE classified 'conditional genes' as separate entities in our computation to characterize properties of introns by essentiality.

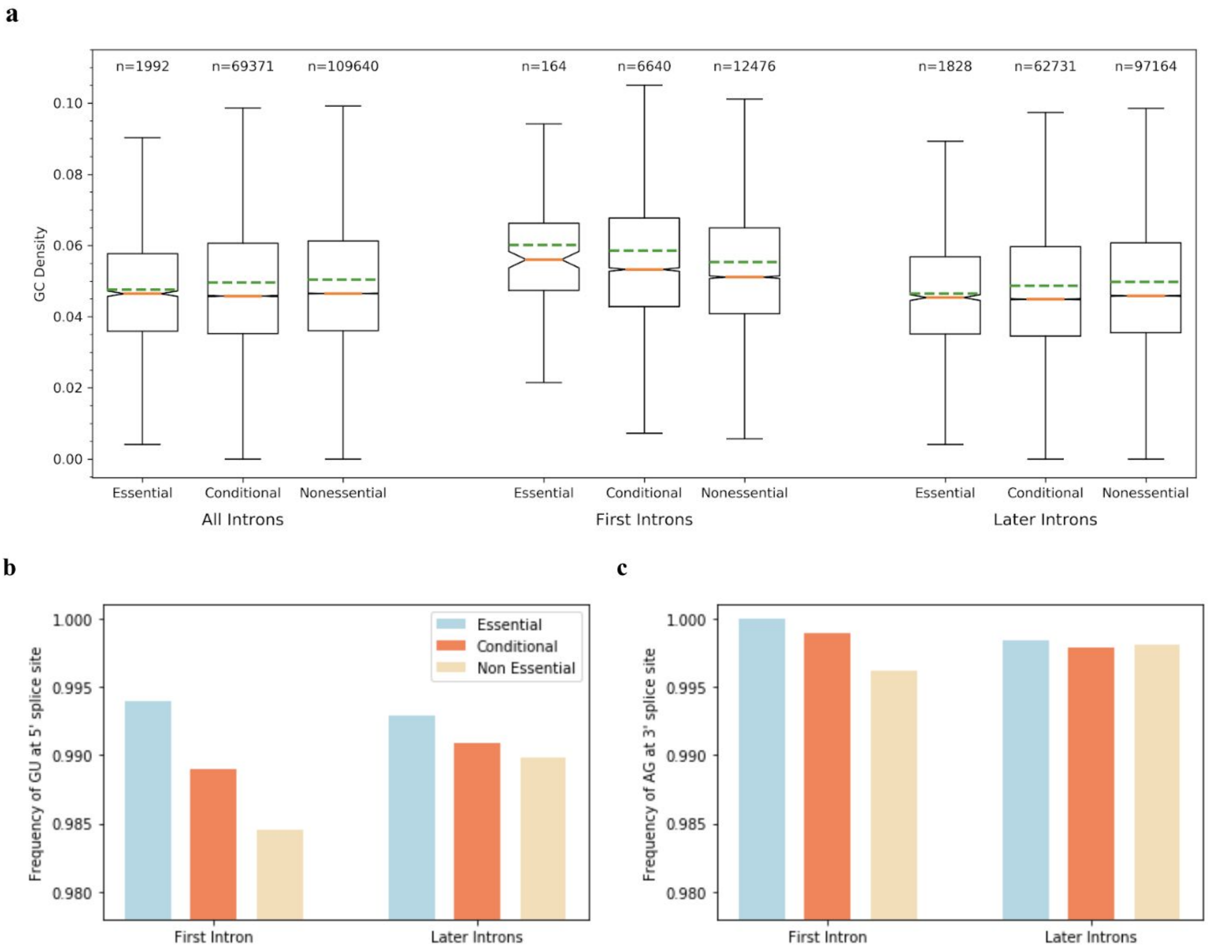
Fig. 2: Introns of essential genes differ from introns of nonessential genes by size, number, and position



a, The dashed-green line represents the mean and the notches are calculated using a gaussian-based-asymptotic approximation to represent confidence intervals around the medians (orange lines). The first introns for essential (p=0.0001), conditional (p<0.00001), and nonessential (p<0.00001) genes are larger than the later introns; however, essential gene first introns are longer than the later introns to a lesser degree than those of nonessential introns. The nonessential first intron is much longer (mean three times greater) than the essential first intron (p<0.00001). For later introns, nonessential are longer than essential (p<0.00001), but these lengths are closer than the disparity between first intron sizes. Conditional introns typically fall within the middle. **b,** Essential genes have a greater number of introns than both conditional (p=0.0383) and nonessential (p=0.0003) genes **c,** However, essential genes have a lesser total length of intronic sequence than both conditional (p<0.00001) and nonessential (p<0.00001) genes.

While introns of essential genes differ from introns of nonessential genes by size and number (Figure 2), they also differ by base specific traits. GC density is significantly greater in the first introns of essential genes (Figure 3). Eukaryotic intron 5' and 3' splice sites for pre-RNA processing, 5'-GU-AG-3' boundaries, are highly conserved, while some minor classes of introns have different boundaries[23]. We report that essential gene first introns are less likely to have an unusual 5' or 3' splice site when compared to first introns of conditional and nonessential genes. The same trend is true of later introns, to a lesser degree (Figure 3).

Fig. 3: Introns of essential genes differ from introns of nonessential genes by GC density and lower frequency of unusual 5' / 3' splice sites



a, The first introns of essential (p<0.00001), conditional (p<0.00001), and nonessential (p<0.00001) genes have a higher GC density than the later introns. Essential (p=0.003) and conditional (p<0.00001) genes have a higher density of GC regions in their first introns than nonessential first introns. The proportion of GC density of the first intron to later introns for nonessential genes is 1.1, for conditional genes is 1.2, and for essential genes is 1.3. GC density is greater in first introns of essential genes. **b,** Essential gene introns less frequently have unusual sequences at the 5' splice site than conditional introns which in turn have less frequent unusual sequences at the 5' splice site than nonessential introns. The first intron of essential genes is less likely to have an unusual 5' splice site than conditional or nonessential first introns. Additionally, essential first introns are less likely to have an unusual 5' splice site than essential later introns. A conditional first intron is less likely to have an unusual 5' splice site than nonessential first introns, so we see that this effect correlates with essentiality. The first intron of nonessential genes is most likely to have an unusual 5' splice site. **c,** The first intron of essential genes is less likely to have an unusual 3' splice site than conditional genes which in turn are less likely to have an unusual 3' splice site than first introns of nonessential genes. We see that this effect again correlates with essentiality.

There exists a minor class of introns, U12, which involves using minor U12-dependant-spliceosomal equipment, and is a rate-limiting-step in gene expression[24]. While only 2.7% of all human genes studied contained at least one U12-type intron, essential genes had an increased frequency of U12-intron-containing genes than conditional genes, which in turn had an increased frequency than nonessential genes. It appears that essential genes that contain a U12 intron generally contain more introns and have a larger amount of total intron bp in that gene. The increase in average number of introns in a U12-containing gene is consistent with both conditional ($p<0.00001$) and nonessential ($p<0.00001$) genes. The increase in average total intron bp in a U12-containing gene is consistent with both conditional ($p=0.0106$) and nonessential ($p<0.00001$) genes. It is possible that essential introns have a greater frequency of U12-introns, especially for genes with many introns, in order to increase the timing of transcript processing by the U12-spliceosome so as to allow for better error correction in long transcripts of essential genes.

Discussion

While essentiality is not wholly an intrinsic property of a gene, the ability of our model to predict essentiality or nonessentiality from just intronic sequences suggests that there exist characteristic motifs unique to introns of essential genes. The model's accuracy for selecting essential introns increases when only testing the first intron as demonstrated by the greater AUC. This suggests that the first intron of essential genes has especially unique motifs when compared to the first introns of nonessential genes. We followed up on these results with computational analysis of intronic sequences of essential, conditional, and nonessential genes with regard to all introns, only first introns, and only later introns. The conclusive findings can be summarized in that (i) first introns of essential genes are much shorter than first introns of unessential genes, (ii) essential genes have more introns per gene but these later introns are markedly shorter than the later introns found in nonessential genes, (iii) essential first introns have a greater GC density than first introns of nonessential genes as well as later essential introns, (iv) essential first introns, with essential later introns slightly less so, infrequently have unusual 5' or 3' splice sites compared to the first introns of nonessential genes.

From these results, essential genes appear to exhibit intronic characteristics that protect their first introns from loss and deletions. The first intron is crucial for regulation of gene expression; for essential genes which are central to PPI hubs, any deletion in the first intron has the potential to disrupt an entire network[3,14]. Because deletions occur in longer introns at much higher frequency, essential first introns are on average over three times less the size of nonessential first introns[12]. First introns of essential genes have a greater GC density which allows for an increase in enhancer activity, correct splice site recognition, and protection from intron deletions[12,16,17]. Similarly, as unusual splice sites can allow for alternative splicing, introns of essential genes, especially the first introns, have the lowest frequency of unusual 5' and 3' splicing sites[25]. Furthermore, as the majority of deletions occur at the 3' end, essential genes have an increased number of introns. These later introns however, are smaller than the average nonessential intron, avoiding long introns in essential genes so as to limit any intron loss or deletions. Because deletions in introns of essential genes would alter transcript length and thus interrupt the timing of a complex molecular network, the unique properties of essential introns appear to have been selected to avoid intron losses and deletions.

Conditional genes are correlated between essential and nonessential genes, suggesting a middle ground for both gene stability and alterations of gene functionality. This middle ground is necessary for successful evolution of the genome. We hypothesize that this reflects the desire of the genome to both innovate its genes as well as to conserve its most essential genes. While selecting for deletion-adverse essential intron systems promotes basic network stability, selecting for long, first introns of nonessential genes allows deletions to alter regulation of nonessential genes and even alter gene function.

The results presented here introduce the concept that essential genes have characteristically unique introns from nonessential genes. These differences, as outlined above, can be exploited to target tumors by disrupting nearby essential genes[1]. Interrupting the complex safety net around the first intron can alter regulation and thereby disrupt a network necessary for tumor growth. Similarly, using targeted CRISPR–Cas9 therapies to force deletions of introns within carefully selected essential genes could likewise stunt cancers. We further identify sequences that characterize the motifs used to differentiate between essential and nonessential introns, that can be exploited with future research in this selective targeting (Supplemental Figure 1).

We demonstrate a deep learning model in this paper that can differentiate between essential and nonessential genes with 1000 bp inputs of intronic sequences from the gene-in-question. Essential genes are located close to target-deletion genes in cancer therapies. Using the double-classifier model (AUC of 0.838) to identify essential genes in cancer genomes can accelerate efforts to locate target-deletion genes for cancer therapies.

Methods

Model

Our deep learning model is a convolutional neural network based on DeepBind, a predictive model that has shown state-of-the-art performance in predicting sequence specificities of DNA–and–RNA-binding proteins[1]. Our model predicts the essentiality of the gene of an intronic sequence *'s'* by calculating an essentiality score $f(s) = net(pool(rect(conv(s))))$. Figure 1 depicts our model architecture. Our model accepts 1000 bp sequences encoded as one-hot vectors. The convolutional layer (*conv*) contains multiple filters that detect motifs within the intronic sequence. We apply the ReLU activation function (*rect*), then the pooling layer (*pool*) averages each filter's response across the sequence to determine the cumulative presence of motifs. The resulting values are fed into a small neural network (*net*) consisting of a fully-connected layer followed by a two-value softmax output layer corresponding to the probabilities of the parent gene being essential or nonessential. The fully-connected layer also uses the ReLU activation function, and the softmax function is applied to the output to normalize prediction probabilities. We prevent our model from overfitting by using L1 and L2 regularization as well as dropout[26].

Data

Human DNA sequences and annotations were collected from the Ensembl genome database project[20,21,27]. For each gene, we used the longest transcript so as to avoid alternative splicing products. We thus preferentially used the transcript whose sequence was verified by both Ensembl and Havana. We used the provided annotations to separate out intronic sequences. Before training, the intronic sequences are transformed using one-hot encoding such that each sequence is represented as an *Lx4* matrix for a sequence of length *L*.

We assign labels using gene essentiality information from OGEE, which gathers data from 18 databases of large-scale experiments to provide a reference of how many studies found a gene essential or nonessential[3,6]. For the model, due to the ambiguity of conditional genes, we discard all conditional genes that have been found to be essential in less than half of studies. Genes are assigned binary labels of essential or nonessential, where the remaining conditional genes are grouped with essential genes.

We trained two models, one on the first 1000 bp of introns, and one on the last 1000 bp. This includes the 5' splice site in the first 1000 bp, as well as the 3' splice site and the branch site in the last 1000 bp. These are the three best characterized regions of eukaryotic introns and are the sites that are most directly involved in spliceosomal modification of the transcript to form mRNA.

Training Procedure

We separate the data into training and testing sets by a randomized 80/20 split of introns, ensuring that all the introns of a gene lie in the same set so that no gene-specific information affects the validity of our accuracy on the test set. At training time, we balance our training set by equally sampling from the essential and nonessential classes so that the model does not overfit to a specific class. We selected our model's hyperparameters by performing a grid search of our model's dropout rate, convolutional layer window size, activation function, and L2 regularization strength. We assessed 36 potential models based on three-fold cross-validation, and we chose the hyperparameters of the best performing model. We trained the final model on the entire training set. We trained all models using Adam gradient descent and a cross-entropy loss minimization objective[28]. The model is trained for 30 epochs with a batch size of 64. We implemented our model using the Keras library running on Tensorflow, and trained on an NVIDIA Tesla M60 GPU.

Prediction and Evaluation

We evaluate our model on our test set using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, which measures how well our model distinguishes between essential and nonessential classes. The model produces an essentiality score corresponding to the predicted confidence in the essentiality of the gene of an intron, and the ROC curve is generated by measuring the sensitivity and specificity of the model at varying prediction thresholds of the essentiality score. We also took advantage of both of our models in order to better classify an intron by averaging the scores produced by our two models on the first and last 1000 bp of the intron.

Our model can be extended to classify entire genes with even better accuracy. Rather than classifying the essentiality of individual introns, we classify whether an entire gene is essential or nonessential by combining information from all of its introns. To classify in this manner, we introduce a majority classification method. We accept the list of all intronic sequences of a specific gene and run each individual intron through the model to get the essentiality score of each intron. Then we calculate a gene's essentiality score as the mean of the essentiality scores of its introns.

We attained our highest AUC using a double majority classifier which uses both the first 1000 and last 1000 bp of each intron to classify a gene. We run the first and last 1000 bp from each intron through the models trained on the first and last 1000 bp of each intron, respectively. Then we similarly calculate a gene's essentiality score as the mean of the essentiality scores of its introns from both models. By combining information from multiple parts of multiple introns, the double majority classifier achieves the highest accuracy.

Code

All the code used for data processing, figure generation, and model training, as well as the weights of our final models, are provided at https://github.com/evendrow/Intron-Essentiality/

References

1. Pertesi, M. *et al.* Essential genes shape cancer genomes through linear limitation of homozygous deletions. *Communications Biology* **2,** (2019).
2. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
3. Chen, W.-H., Minguez, P., Lercher, M. J. & Bork, P. OGEE: an online gene essentiality database. *Nucleic Acids Research* **40,** (2011).
4. Chen, W.-H., Trachana, K., Lercher, M. J. & Bork, P. Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age. *Molecular Biology and Evolution* **29,** 1703–1706 (2012).
5. Seoighe, C. & Korir, P. K. Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. *BMC Bioinformatics* **12,** (2011).
6. Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M. & Bork, P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Research* **45,** (2016).
7. Kabir, M., Wenlock, S., Doig, A. J. & Hentges, K. E. The Essentiality Status of Mouse Duplicate Gene Pairs Correlates with Developmental Co-Expression Patterns. *Scientific Reports* **9,** (2019).
8. Bartha, I., Iulio, J. D., Venter, J. C. & Telenti, A. Human gene essentiality. *Nature Reviews Genetics* **19,** 51–62 (2017).
9. Woods, S. *et al.* Duplication and Retention Biases of Essential and Non-Essential Genes Revealed by Systematic Knockdown Analyses. *PLoS Genetics* **9,** (2013).
10. Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics* **10,** 19–31 (2009).
11. Lin, H., Zhu, W., Silva, J. C., Gu, X. & Buell, C. R. *Genome Biology* **7,** (2006).
12. Rigau, M., Juan, D., Valencia, A. & Rico, D. Intronic CNVs and gene expression variation in human populations. *PLOS Genetics* **15,** (2019).
13. Roy, S. W. & Gilbert, W. The pattern of intron loss. *Proceedings of the National Academy of Sciences* **102,** 713–718 (2005).
14. Bradnam, K. R. & Korf, I. Longer First Introns Are a General Property of Eukaryotic Gene Structure. *PLoS ONE* **3,** (2008).
15. Trynka, G. & Raychaudhuri, S. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. *Current Opinion in Genetics & Development* **23,** 635–641 (2013).
16. Chen, L., Fish, A. E. & Capra, J. A. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLOS Computational Biology* **14,** (2018).
17. Wang, D. & Yu, J. Both Size and GC-Content of Minimal Introns Are Selected in Human Populations. *PLoS ONE* **6,** (2011).
18. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nature Genetics* **31,** 415–418 (2002).
19. Juneau, K., Miranda, M., Hillenmeyer, M. E., Nislow, C. & Davis, R. W. Introns Regulate RNA and Protein Abundance in Yeast. *Genetics* **174,** 511–518 (2006).

20. Hunt, S. E. *et al.* Ensembl variation resources. *Database* **2018,** (2018).
21. Embl-Ebi. *EBI* Available at: https://www.ebi.ac.uk/ena/data/view/GCA_000001405.28. (Accessed: 31st March 2020)
22. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33,** 831–838 (2015).
23. Wu, Q. & Krainer, A. R. U1-Mediated Exon Definition Interactions Between AT-AC and GT-AG Introns. *Science* **274,** 1005–1008 (1996).
24. Patel, A. A. The splicing of U12-type introns can be a rate-limiting step in gene expression. *The EMBO Journal* **21,** 3804–3815 (2002).
25. Ast, G. How did alternative splicing evolve? *Nature Reviews Genetics* **5,** 773–782 (2004).
26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15,** (2014).
27. Cunningham, F. et al, Ensembl 2019, *Nucleic Acids Research*, **47**, D745–D751 (2019).
28. Kingma, D. & Ba, J. L. Adam: A Method For Stochastic Optimization. *arXiv* (2017).

Author Information

Stanford University, Stanford, CA, USA
Ethan Schonfeld
Edward Vendrow

University of California, Los Angeles, Los Angeles, CA, USA
Joshua Vendrow

Glenbrook North High School, Northbrook, IL, USA
Elan Schonfeld

Contributions

Ethan S. and Elan S. conceived the project. E.V. was responsible for data pre-processing. J.V. and E.V. performed the deep–learning and computation. Ethan S., Elan S., J.V., and E.V. analyzed the data. Ethan S. and Elan S. generated conclusions. E.V. and Elan S. created the figures. Ethan S. wrote the manuscript with input from all authors. J.V. and E.V. wrote methods with input from Ethan S.

Corresponding Author

Correspondence to Ethan Schonfeld
eschon22@stanford.edu

Ethics Declaration

Competing Interests

The authors declare no competing interests.

Supplementary Information

Supp. Fig. 1: Optimized 24 base pair sequences for each of the 128 filters in convolution (left trained on first 1000 bp of intron, right trained on last 1000 bp of intron)

AGTATATAGAGTGGTTTAGTCCGG	CCCCCCCCCCAAAAAAAAAAAAAC
CCCCCCCCCCCCCCCCCCCCCCCC	TTCTCCATGGAACGATGAGAGCGT
TACCGTGTAAAGCCTGTACCGGCC	AAGGGGGGGACAAGGAGCCCCAAC
CGATTCCCACGATCATATCCGATG	CCCCCCCCCCCCCCCCCCCCTCCC
CCGGGGGGGTCCTACACAAACCCA	TAACCGTACCGAGAAACGGTTTTT
GTAGTACCGTGCTCGCGTGTCGTA	TACTTTAGGGTCGTGAGGTACTAG
AGCCTCCGCTGAGAGTATGGGAAA	ACGGTTTTTTACAGTTAGCAGTTG
AGCGCCAGGAAAGCAGCGAAGCAG	TCTCCATCCACTGAAGAATGCGAC
GCTTAGGTCTGGTGATACGAAACT	CCCCCCCCCCCCCTCCCCCCCCCC
GTCGTCTTAGACGTTCTCGCTGTG	AAAAAAAAAAATAATAAAAAAAAA
CCCACAACGGGAAGCGACCGACAA	CCTGGGTAGCGGCGCCTAGTAGTG
CCTTTAGTACATACTAGGTACGGT	GGCCGTGTTACCGTTTAGTTGTTC
CAATCAATGGAATGCTCAATGGCC	TGAAACCATACCTAATGTTGCCTT
CACCCCCACACACACCAACGCGAA	AAATTATCCTTCACTATCTAAGGA
GTGCGGTGGGGGTGGTGGGGGGGT	AGTCCCCTACAATGTCCCAGACAC
GATCCCGAATCCATCTTCAATGGC	AATGTGAGTGCGGGACCGCGGGGG
CAATTACTTCCAATGACTCACCCC	CGCTTTTAAGGCGTCGTTGTTTAG
AAAAAAAAAAAAAAAAAAAAAAAA	GGGGGGGGGGGGGGGGGGGGGGCC
CACAATATCCCCCCGGGAACCTCA	ACGGGGAGGGACCAAAAGAGAGCA
CCCCTTTTCCTATCTCCTTTCCCT	GGTACGGAGCCCTTGTACTGTATG
CAAGGAAAGTTTCCCTGGGAGTCC	ACAATCCCTCTATCACTACCCCGG
TACGTCGTCTTTGTCGGTAGTGCG	AAACCCTCCATCACTCATCCCCCC
GGGGGGGGGGGGGGGGGGGGGGGG	ATCCGACGGTCAATGTCATTCCCC
TTTACGTAGCGAGTAGACAGTGTA	GCAGCAGTGCCGAGTTAGGCCTGC
CTGTAGCGCCGTTACGTTGCGTTA	AACCATCTATCAATGGCCAAATCG
CCCCCCCCCCCCCCCCCCCCCCCC	ATTCCTCCTATGGGACCCGAGAAG
ATTATACTGTAGTGCCGTCGCAGC	CCGCGAGCCGGCCTAGTCGCTTAA
GTTTAGTACCGTAAACGGACCGTT	CGATGAATCTAAGCGCCCAATGGC
CCCAACACTGAATCCTTATTCAAT	AGGTGTGTAGGGAGTAGTGTAGTA
ATTGTAGAGTAGTAGCAGCGTACC	TCAATGAAATGCTCCCTTGGACAA
CCCCCGGGCCCCCGCGCCCCCGCG	TTAGTAGCCGTTCGTAGTAGTCGT
GACTAACTCTATGGAAATCTCCAC	ACAATTAAATCACTGCCAATTCCA
GGCATCAAAGGGATGCTCCATGGA	TCCGCCCCCCAAAAAAAAATTCCG
GGGGGGTGCTGTGGGTGGGGGGTT	AATCCTCCGATTAATATTGCTCCA
GTGGCTCCTGTGGGTCTGGGGAAG	CCTTTGTGGTGGTAGTGTCTTGCG
TATAATTTTTCTCTAGATATATAT	CTCAATCTCCCCCCAATGACGCCA
TCCAAATGGCCATGCAAACCACAA	TACTTGAGTCGCCAGCCCCAGTAG
GACACGGTGGCGCGTTGCAGTTCG	CCCCCCCAAAAAATTTTTTATCCC

```
AGAATCATCACTCTATGTCTCAGA
GTTGCGGTCAGTAGAGACCAGTCG
GGGGGTGCCTATGGGTGGGGGGTT
TGTATTACGTGGTAGTCGTTAGGT
AGGCGTTAACGCTGTTTAGGCGTT
GGGTGTTGGGGGTGGGGGCGGGGT
TTTTTTTTTTTTTTTTTTTTTTTT
TCGATTATCGCAATCAATCCAGGC
CGGGTCGACAATCAGAGTGAGGAC
TTACGGCGTAACGTAGTCGTCTGT
CATTCAATATCACTGACTAGAGGC
CGTTATTCCTTCCTTTTCGTCCGG
GTAAATCGCTTGTAGGCCTTTAGT
CCCGCCCCCGCGGGGGCGCCCGCC
GCGTCATCAATGCATAGTCCGACA
CAATCAATCTCATCCTAGCACAAG
GGGGCAAATCCTCCAATCAATAGC
TTTGGTGGCTGGTTGTTTGGTGGG
CGTCGCGTCTGAGGCCTGCAGCCC
CAGGCACTGGAGGCACCTGCAAAA
CGTAGTAGTCGTTACCCGTTTGCC
AACGCATCAATGGCTCCATATCAA
GGGGGGGGGGGGGGGGGGGGGGGG
TCACTCCCACCATAACTCCCATAA
CCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCGGGGGG
GCACCGTAGTCGTAAGTCGTCGTG
TGTAAGTACGTAGTTATCGCGGTA
AGGATCCCACCGCGGCGGGCCCCC
GTTACCGTTCAGCGTAGTAACGGT
AGCCAATGATCTGATAAATTCCAA
TAGGCCGTACGGTTTACCCTAGGT
CCTGTAACAGACTTAAGGTCGTTG
GTAATAGCCGTAACGTCGTCTTAA
GGGCTGAATCCACGTCCAATCCTC
CCAATTCCCAACGAGGAGCCCCAT
GGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGG
AGTACTTAAGATTTTGTTGACTTG
CCAACTCTAACGATCACAATCCCT
TGAGTCGTTAGACAATCGAGGCTG
AAGTCGCCACGGACATCCCCTCTC
TTTTTTTTTTTTTTTTTTTTTTTT
GTAAAGAAATAGGGTTTAGTTTTT
GTTAACCCGTCTTAATAGACTTGT
GTTAGTAGTTAAACCCGTTGGCGG
GAGCCGTTGAGTTAACGGGTTGTT
CCCAAACCAAAACGTGGTAACTGC
TAAAAACTTGAGTGTATGCTTGTT
AAGGGGCTTTCAAGGGGCCAACCC
GGGGGCGGGGGGGGGGGGCCGCG
AGCTTGTCTATCCTCACGAGAGGT
GGCGTAGTACTTTATACCTAGCGT
GAATAGTAGTGAGGCAGTTGTGTT
TTTTGTGGCGTTGGTGGTGTTGTT
ACATGCGACAAATGCGACGACCAA
GCAGTATAGCCGACTCGTGGTAGT
ACTACCACACTTTATTATTCCCCA
GTTATCGCTCCGTAGTAGTCGCGA
CACCAACGATGGATCAACGAGAAA
CCCTGCCTGGGCCCTGGGTCGCGG
GCAGTACGCTCCAGCCGAGCTAGC
CATTTCATTATCAATATCCCAGCC
TGTGCGCGCTATTGGGGGTGTGCG
GAACAGTAGACGTAAACCGGGTAT
CCTCCCCCCTCCATCACTCCCCCC
AATGACTTATCATTCTCTATTAAC
AGCTTTAAACAGTTTAGGTACCGG
CTTGTTTGGGCCGTAGTAGCGAGC
ATTACTCTCCATTATTCATGACCG
ACTTCAAGTCAATTTCTATGGCCA
CCCAAATGTCAATGCCAAATGCAC
TCAATCAATCCAATGGGCACTCTG
CAATCACTTACCATTCAATCACCA
ATGCAAGGACTGAAAGACATTCCC
GTGGTGGTGGTTGTTGGTGGTGGG
AATTCCCCTGCCGGATTCCAAGGC
GGGTGGGGGGGTGTGGGTGGGGGG
CCCCCCCCCCCCACACACACCCAC
TTTTTTGTGGTTGTTGTTGTTTTT
ATTTCAAACACGCAACTTCAATGC
GTAGTTGTTGGACGCCGTACTTAG
GTCGTGTCGTTCCAGGCTTTCGCT
ACAAAGTTCTCCATGTCAACTCGC
GAAAAGTAGTAGATAGTAGTTTAA
ACCAAAACCTTTCAATGCCAAAAG
```

```
CTATCCCATCTGGGGGTGCCAATG
CCGGTAGTACGTAGTACGCGTATA
GTAACTAGTCTTGCGGGGTCCGAG
ATTTCCTCGGTGGATGTACGGAGG
GGCCCGCGGCCCCCGCGCCCCCCC
CGTAATGAGTGTACAGTGTAAGCG
CCCCACCACCCACAACAACGCCCA
CTGTATCGCTTAAGTTCGAAAGGG
GAGTCTGGTATGGTTACGTACCGT
TAAATCAATCACTACTGCTTCAAT
GTAAGACCTTAAAAACCCTTTTTT
AAATTTGCCCCGGAGGACATCTAT
TAACTGGCTTACCTCGTTCAGATC
ATCAATCAATCAATGCTACGTCAA
GTTAGCGTACCTGCTAGTGGCGGG
AAAAAAAAAAAAAAAAAAAAAAAA
AGCCCCGCAAAATTTCCCCGGGGG
GTCGTTCAGTAGCAGTACTCCTGT
CCCCGCCCCCCCCCCCCCCCCCCC
GCTGCTGCGTTCTCCGCTGCGCTC
TAGGTAGTGTGCCCGGATCGTACG
CACTAATAGTCGTAGCCGGTCAGT
AATGTCCACCAGGCTATCTCATGG
TGTAAGCCAGACCGGCCTACAGTC
CAGCATGAATCCCTGGACATTCAC
GCAGTGTAGTACTTAAGGAGGTAG
CTTCAATTCCTATGCACGCAAAGC
AAACTCTATCATCATTTTCATATG
CGTTATTGCCACGTCGTAGACGTG
GTTCGTAGGCTTGTCGTATATAGA
CCCCCCCCCCCCCCCCCCCCCCCC
CCCTTGACGCTTTAGTAGGGCTTA
CAATTCCAATGACTCAATTTCTAG
GGGGGGCCCGGGGCGCGGGGCCCC
CGACGACATATCCAAAGCCCCATA
AAGAGGCGGGGACAGAAAAAAAGA
GTACGTAACGGTAGCGTATACTTT
ATCATCAATGACGTGCGATCTATC
AGGTGAGCGTTTTTTTGAGAAACG
TTGTCGCTTAGGCACTAGTACCTT
ACGTCAGTGTTGTACGTTAGTCGC
ATCAAATTCCAACCACAACATCAA
AAACCAAAAAAAAAAAAAAAAAAA
TCCAATCACCCTCCATGGAGAAGA
AGAAATGGGCGACGGGAATCGGGA
CAGTCCTGTCGCAGTAAGCCTTCT
CGGCACTCTCCTGGCGGTCGGGTT
TACGTAGCCGTTACCCTGTTGTAG
TTTCTCTTCTTCTTCTTGTGTTTT
AAACGCCCTCAAGGACGACCCCGC
CCGGCACCCAGTAGTGATCGTGTT
GTTCGAGTATAGTAAAACTGTACC
GATCTCAGACATACGAGTCCCGAA
GCATAGTTGGTAGTCGCTTAGGCT
CCGTAGTCTTTCCGCCCGGTACTT
CTTAACCTCGTATAGAGACGTGGT
CCTACCGTAAACCTTGTTTTGTAA
GGGGGGGGGGGGGGGGGGGGGGGG
CTTTGGTATTTAAACAGGAAAGGG
ACAAACGACCAATAGGAAGCCCAA
CAGTTCTTTGCTACCAGCCCCACC
GTAAGGATTGCAAACGCCTTTGTG
CGTGGTAGTCGTTACCGACTTGGC
GTATTGCTAGCGAAACCTTGAGGT
GGGGGGGGGGGGGGGGGGGGGGGG
AAATGCCCCCATCCAAGCGATGAC
GTAAACCCTTGTAAAGAACCCTGT
GAGAGGGAGGGAGGGAGGGAGAAA
CCGTCACAGTATCGTGGTAGTACC
TAAGGAGTAGGGGTTGGTTTTTGT
GTTTTTTGGACGTTGGACCGTTTA
TCATGGAGTCGTTCCAGTTCAGTT
AATGTGAGTGCAAGGACGTGGGGG
CTGGCAAGCAATGCATTCCCCGAC
CCTTGCTCCAAAAGATAGACGTCC
TCTCACTCCATATATCCACGGGCA
GCCCAGTGTTGTGTCGCTGTTTAT
TCAATATCCCTCAGGCAACTTCAA
CATTCCCTATTCAATAGCCTCAAA
ACAGTAGCAAAACTGTTAGGTTGT
CGCGGTAGCGGAACCGTTTAGTAG
GTAGACTCGTAAGGTTTAGCGGTT
TCGGCCGCCCCAGCCGCGCCAGCC
TTTTTAAAAAAAAAAAAAAAAAAA
GATCATACTTTCACTATCGATTCC
TAGGGAAGCTGGGTAGCGGTTTTT
```

```
CGGTATCGGTGTTGTACCTTAGCC              CCCCCCCCCCCCCCCCCCCCCCCC
CTCATAATTACTATAATACTCAAA              AGAAAAGGCGGACCCAAAACGCAG
CAATGGAGCTCAAGGGCCACCAAT              GGAGGGGCACACAGAGATCAACAC
GCTGCACATGTGGGCACAGAAAGA              ACAGGGCCGATAAAACAGAGTGGG
```