Alexander B Rosenberg

Learning Models of Gene Expression from Synthetic DNA Sequences

Alexander B Rosenberg

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Georg Seelig, Chair

Jay Shendure

Eric Klavins

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Learning Models of Gene Expression from Synthetic DNA Sequences

Alexander B Rosenberg

Chair of the Supervisory Committee:
Associate Professor Georg Seelig
Electrical Engineering and Computer Science & Engineering

Over the past decade, new sequencing technologies have enabled the comprehensive cataloging of human genetic variation, but for most DNA sequence variants we do not even understand the impact on molecular phenotype, let alone human traits and disease. Despite recent advances, the throughput of gene editing technologies is orders of magnitude away from allowing the functional testing of all possible variants. As a result, models are desperately needed to accurately predict the impact of variants on gene expression.

In classical genetics, genomes are compared, analyzed, and perturbed to learn the function of the underlying DNA sequences. Here I present a complementary but orthogonal approach, in which fully synthetic DNA sequences are studied. This thesis is a collection of three chapters exploring this approach. In the first chapter, noise buffering at the protein level is explored by constructing feed-forward loops with different engineered miRNA targets. In the second chapter, a model of alternative splicing is learned from millions of random DNA sequences. The resulting model outperforms all other models that we investigated on the task of predicting the effects of human exonic variants on alternative splicing, even though our model was never trained on human DNA sequences. In the third chapter, the same approach is extended to model the impact of the 5' untranslated region on translation in yeast. After training on measurements of ~500,000 synthetic

5’ untranslated regions (5’ UTR), the resulting model accurately predicts the effects of both synthetic and native yeast 5’ UTRs on translational efficiency.

This thesis is by no means a comprehensive exploration of gene expression models that can be learned from synthetic DNA sequences, but rather a starting point. The throughput of both DNA sequencing and DNA synthesis have increased at such a rapid pace that we can now make millions of gene expression measurements in parallel. Given that most genomes only have thousands of genes, we should leverage this additional capacity to test synthetic sequences and improve our models of gene expression.

Acknowledgements

The past several years that I have spent here at the University of Washington have been an absolute blast. I could not have dreamt up a better experience when I started graduate school. I am grateful to so many people who taught me so much and shared with me their passion for science and discovery.

I first want to thank Georg Seelig for surpassing everything I could have expected from an advisor. His enthusiasm for science and engineering has been contagious. He trusted me with the intellectual freedom to find my own research interests, while still providing the guidance and mentorship to help me develop into a more effective researcher and communicator.

I thank Tim Strovas for showing me the ropes and teaching me many of the basic molecular biology techniques. In his own words, he got me "to start thinking like a biologist." I also want to thank Robert Egbert for encouraging me to experiment with new lab techniques.

I thank Jay Shendure for his invaluable guidance and mentorship during our collaboration. Jay played an integral role in shaping the trajectory of my research path towards genomics. I also thank Jay's former student Rupali Patwardhan for guidance on the experimental design of the alternative splicing project. Choli Lee also provided valuable assistance with Illumina sequencing.

I thank Stan Fields for both his contributions on the yeast project and useful advice and insights about building a rewarding research career.

I originally met Josh Cuperus at a conference in Hawaii. We spent a lot of time discussing science while hiking around the big island and our collaboration just naturally grew from there. Working with Josh has been a terrific experience. I hope we continue to collaborate in the future, and I continue to learn from him.

I want to thank the members of my committee: Eric Klavins, Jay Shendure, Bill Noble, and Georg Seelig. They have provided valuable feedback not only on my work, but excellent advice on how to focus my time and effort. I want to give special thanks to Eric and everyone in his lab for providing ideas and making the UW such a fun place to work.

I thank Nebojsa Jojic for providing me the opportunity to work with him at Microsoft Research. Nebojsa pushed me to learn new machine learning methods and think about biology from a different perspective.

I want to thank all of my collaborators and friends in the Seelig lab. All of my direct collaborators have been wonderful to work with (Briana Kuypers, Richard Muscat, Ben Groves, Anna Kuchina, Randolph Lopez, Nick Bogard, Johannes Linder, Sumit Mukherjee, Kali Baker, Will Chen, Yue Zhang, and Charlie Roco). I also want to thank the rest of the lab for advice, support, and great ideas (Yuan-Jyue Chen, Xi Sherry Chen, Sergii Pochekailov, Paul Sample, Alberto Carignano, Gourab Chatterjee, Alexander Baryshev, Sifang Chen, Arjun Khakhar, Sundipta Rao, Ban Wang). Yuan, Sergii, and Ben have been my close friends and workout/ski partners over the last few years and the source of countless laughs. I thank you all for making sure that work never felt like actual work.

Finally, I want to thank my mom, my dad, and my sister Sarah for their love and support. You all are the best. Words cannot express my gratitude for everything.

This thesis is dedicated to my grandmother (mormor) and role model: Ingrid Göthman (1932-2012).

Table of Contents

Chapter 1: Introduction

The primary goal of my thesis work was to develop better ways to build predictive models of gene expression. But why are predictive models of gene expression even important? To answer this question, some context is needed. In this chapter, I will briefly review the current the state of genomics and the barriers towards translating results into advances in human health. This chapter is only meant as a short summary and not as a replacement for some of the excellent reviews of the current state of the field[1-3].

Genome sequencing and genetic variation. The modern era of genomics was born with the announcement of the first human genome draft in 2000. The Human Genome Project (HGP) began 10 years earlier in 1990 and was completed at an estimated cost of $2.8 billion. The HGP sought to reveal the inner workings of our biology as human beings. By creating a genetic reference map, each human disease would be attributed to specific aberrations in the underlying DNA sequence. However, in order to find which sequence variants caused disease, the naturally existing variants in the human population would have to be first catalogued.

Much of the work in the next decade focused on this cataloguing of genetic variants. Inexpensive methods using microarrays provided the means to identify millions of predetermined variants in individuals[4], but came far short of identifying every variant in every individual. To create a complete picture of human genetic variation, whole genome sequencing (WGS) was needed. But considering the 2.8 billion dollar cost of the first genome, a massive drop in sequencing costs was necessary. The goal of making WGS practical for cataloguing variants provided a strong incentive to drive sequencing costs down from the initial price tag of $2.8 billion per genome. In fact, the decrease in DNA sequencing costs even outpaced “Moore’s

Law" (the number of transistors in dense integrated circuits would double every year) with the cost of genome sequencing rapidly approaching $1,000 (Figure 1).

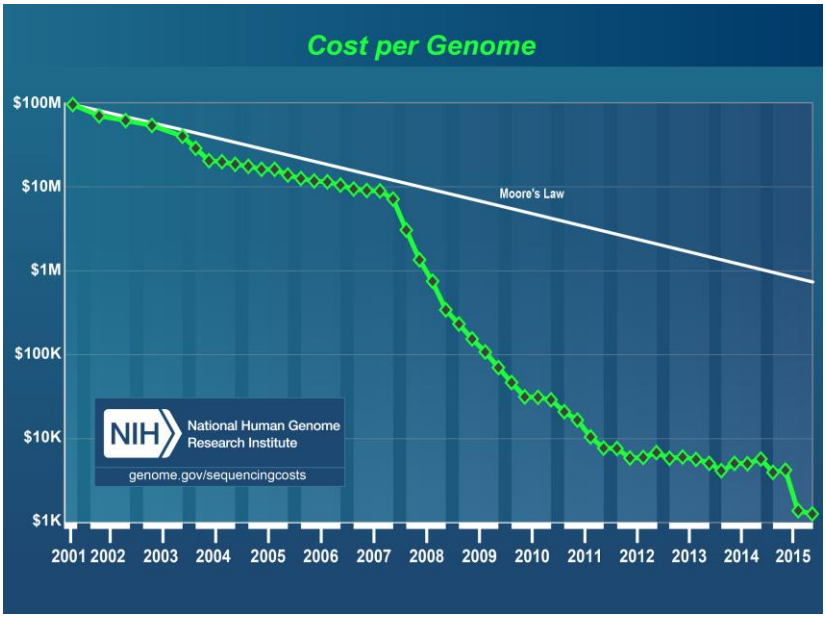


Figure 1 Cost per Genome (https://www.genome.gov/27541954/dna-sequencing-costs/)

Yet when I started my graduate studies in 2010—10 years after the HGP released the first human genome draft—the applications of the Human Genome Project towards health and medicine had not yet delivered on early promise. Journalists were labeling the human genome sequencing project as "The Great DNA Letdown[5]". Both scientists and politicians shared blame for raising the public's expectations to unreasonably high levels. In his White House address celebrating the completion of the first human genome draft, President Clinton certainly did nothing to dampen the expectations[6]:

> Today, we are learning the language in which God created life. We are gaining ever more awe for the complexity, the beauty, the wonder of God's most divine and sacred gift. With this profound new knowledge, humankind is on the verge of gaining immense, new power to heal. Genome science will have a real impact on all our lives -- and even more, on the lives of our children. It will revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases.

> In coming years, doctors increasingly will be able to cure diseases like Alzheimer's, Parkinson's, diabetes and cancer by attacking their genetic roots. Just to offer one example, patients with some forms of leukemia and breast cancer already are being treated in clinical trials with sophisticated new drugs that precisely target the faulty genes and cancer cells, with little or no risk to healthy cells. In fact, it is now conceivable that our children's children will know the term cancer only as a constellation of stars.

While President Clinton hinted at the timescales in which we should expect significant progress ("our children's children"), this was easily overlooked. In reality the HGP was not a failure, but the challenges associated with interpreting the genome had been vastly underestimated.

Genome wide association studies. In the years that followed the HGP, as the catalogue of human sequence variants expanded, it became more and more apparent how few causal variants of disease could actually be identified. As of 2011, Genome-wide association studies (GWAS) had found significant correlations between ~1,300 loci and ~200 diseases or traits[2,7], yet very few studies had identified a single causal variant. Ironically, one of the main obstacles to identifying causal variants was not a lack of variants, but an overabundance of variants within each individual. There were so many variants that some variants would show moderately high associations with disease by random chance without any basis in reality, a textbook example of multiple hypothesis testing. So as the number of variants tested increased, in order to not increase the number of false positive associations, the threshold for a significant association had to be raised. This might have worked if the truly causal variants exhibited large effect sizes, but for most diseases and traits this was not the case. Even when variants with statistically significant associations were identified, there was no guarantee that they were causal variants. Linkage disequilibrium—the nonrandom statistical association between different variants—often made it impossible to verify whether a variant with a significant disease or trait association was truly causal or simply correlated with a causal variant.

Specificity of GWAS variants could be increased by combining evolutionary constraint measurements, but this also resulted in a reduction of sensitivity. While all sequences with evolutionary constraint by definition must be have biological function, not all sequences with biological function exhibit measureable evolutionary constraint[8]. Often in complex diseases, many variants contribute with individually weak effects, and the sensitivity of evolutionary constraint might be too low to identify these variants (only 4.2-8.2% of the human genome has been shown to be under evolutionary constraint[9,10]). Another confounding factor was that multiple mechanisms might contribute to constraint of a given sequence. For example, exons are under constraint from both the amino acid sequence as well as RNA splicing.

This leads to the fundamental limitation of both GWAS and constraint-based measures—the adoption of a purely statistical approach that ignores the underlying biology. Unfortunately most variants identified through GWAS or constraint could not be understood in terms of how they affected gene expression, cell function, much less disease or complex human traits. Ideally, the impact of these variants on biology could be learned by testing the variants with experiments. However when I began graduate school, measuring the effects of all these variants, even just with respect to gene expression, was prohibitive. Each variant could affect RNA expression, protein expression, protein function, even potentially in a cell dependent manner. All of these variants would have to be assayed for changes in each molecular process in every potential cell type.

Massively parallel reporter assays. However, in 2009 the first high-throughput study demonstrated that it was possible to measure the effects of thousands of sequence variants on a single molecular process[11]. In this work, the authors tested the effects of every possible single nucleotide mutation of three bacterial promoters and three mammalian promoters all in a few

highly parallel experiments. To perform the same assay a few years earlier, each promoter variant would have had to be assembled and tested in a separate experiment—a prohibitive amount of work, even for the most determined graduate student.

To create their highly parallel promoter assay, the authors leveraged several technological developments. All promoter variants were synthesized as oligonucleotides in parallel on a programmable microarray. The development of programmable microarrays made it possible to synthesize thousands (eventually hundreds of thousands) of DNA sequences in parallel. In the experiment, each promoter variant was designed to transcribe a barcoded transcript, such that each transcript could be mapped back to the promoter from which it was transcribed. Then a single *in vitro* transcription reaction was performed for each of the six promoter types, with all of the variants pooled together. Using RNA-seq[12], a variant of "Next Generation Sequencing" (NGS) targeted to transcriptomes, it was then possible to count the number of barcoded transcripts and get a digital readout on the activity of each promoter variant. The power of this approach was that everything could be done in a single pooled reaction as opposed to thousands of individual experiments. In the following years massively parallel reporter assays (MPRAs) that combined next-generation sequencing with extensive variation were used to study transcription[11,13-17], translation[18], mRNA stability[19,20], and even alternative splicing[21].

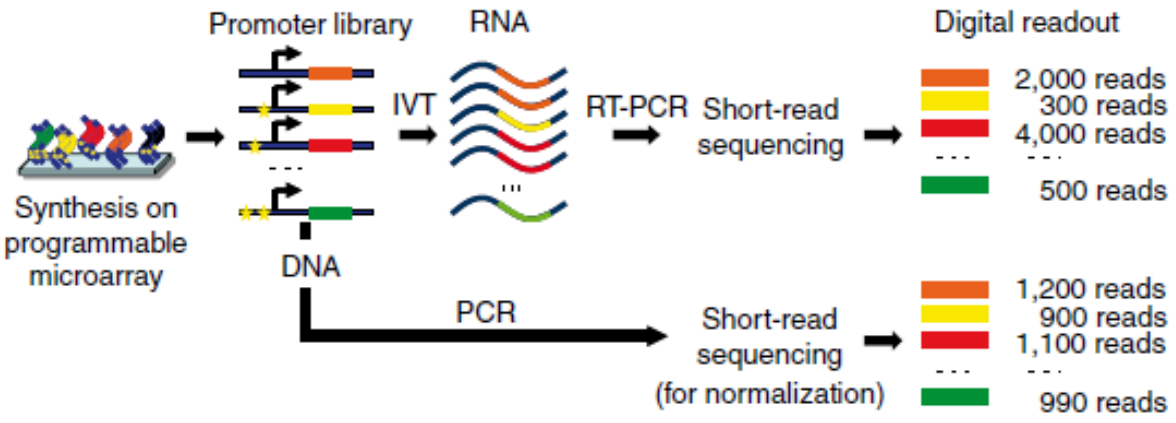


Figure 2. The first massively parallel assay to measure thousands of promoter variants[11]

The need for predictive models of sequence function relationships. While MPRAs provided an excellent method to test the effects of a large number of variants of interest, some quick back-of-the-envelope math made it apparent that they could not be scaled to test *all* possible human variants. Only accounting for substitutions, there are over 9 billion possible variants. Given that each person has approximately 38 *de novo* variants[22] and there are 7 billion humans on earth, every non-embryonic-lethal variant likely exists in at least one human. Therefore it is truly important to understand the effects of all human variants. However, since the effects of all variants could not be measured directly, computational models were needed to predict these effects.

In the field of machine learning, the accuracy of predictive models is fundamentally limited by the training data on which the model was built. Accurate models of complex underlying relationships require huge amounts of training data. For example, Google's Alpha Go algorithm[23] that recently defeated the world champion human Go player, was trained on a dataset of over 30 million Go moves from real human games. However, the human genome only contains ~25,000 genes, a very small dataset by modern machine learning standards. Given the

known complexity of gene regulation and only 25,000 genes to study, learning accurate models is a monumental, if not impossible, challenge.

This is the problem that I have spent the last six years addressing during my PhD. The goal of my work was to learn models of gene expression by training on enormous datasets. When I started graduate school, these datasets did not exist. So I spent my PhD developing MPRAs to measure how cells expressed millions of synthetic sequences. These datasets contained orders of magnitude more sequence measurements than the number of genes in natural genomes. This made it possible to learn richer, more complex models of gene expression that captured more of the underlying rules of regulation, leading to higher accuracy.

There were several other benefits to working with synthetic data. First was the guarantee that the observed effect on gene expression for a given synthetic sequence was truly causal. While this may seem trivial, it is very uncommon in genomics datasets. It is possible to search for motif enrichment, (*e.g.* which codons are most common in highly expressed genes), but there is no guarantee that the motifs are truly affecting the process of interest. Evolutionary selection may have led to enrichment of these motifs for a completely different reason.

Another slightly counterintuitive benefit is that synthetic sequences contain many sequences that have been selected against in the native genome. While purifying selection indicates these sequences are deleterious, there may not be obvious evidence of why. By testing these sequences in MPRAs, it is possible to quantify exactly how they influence gene expression. This is especially important in the analysis of *de novo* variants, since they are much more likely to introduce sequence motifs rarely seen in the rest of the genome.

The results of my PhD work are presented in the next three chapters. The idea to combine MPRAs and machine learning formed over the first two years of my graduate studies. During

this period I was working on a project studying how miRNA targeting impacted noise in gene expression. This project opened my eyes to the potential of using synthetic sequences to learn regulatory rules in biology. The results of this project are presented in the next chapter. The following two chapters present the core work of my PhD; using MPRAs to learn predictive models of alternative splicing in humans and translation in yeast. The project studying translation is an ongoing project, so I expect the final publication to evolve significantly from the results presented here.

References

1 Goldstein, D. B., Allen, A., Keebler, J. & Margulies, E. H. Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics* **14**, 460-470 (2013).

2 Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* **12**, 628-640, doi:10.1038/nrg3046 (2011).

3 Boycott, K. M., Vanstone, M. R. & Bulman, D. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews …*, doi:10.1038/nrg3555 (2013).

4 Lai, J. The Great DNA Letdown. *Forbes* (2010). <http://fortune.com/2010/04/08/the-great-dna-letdown/>.

5 Remarks Made by the President, Prime Minister Tony Blair of England (via satellite), Dr. Francis Collins, Director of the National Human Genome Research Institute ... on the Completion of the First Survey of the Entire Human Genome Project. *genome.gov* (2000). <https://www.genome.gov/10001356/june-2000-white-house-event/>.

6 Lander, E. S. Initial impact of the sequencing of the human genome. *Nature*, doi:10.1038/nature09792 (2011).

7 Consortium, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816, doi:10.1038/nature05874 (2007).

8 Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS genetics* **10**, doi:10.1371/journal.pgen.1004525 (2014).

9 Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature biotechnology* **27**, 1173-1175, doi:10.1038/nbt.1589 (2009).

10 Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-536, doi:10.1016/j.cell.2008.03.029 (2008).

11 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology* **30**, 271-277, doi:10.1038/nbt.2137 (2012).

12 Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology* **30**, 265-270, doi:10.1038/nbt.2136 (2012).
13 Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature genetics*, doi:10.1038/ng.2713 (2013).
14 Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology* **30**, 521-530, doi:10.1038/nbt.2205 (2012).
15 White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 11952-11957, doi:10.1073/pnas.1307449110 (2013).
16 Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Molecular Systems Biology* **10**, 748-748, doi:10.15252/msb.20145136 (2014).
17 Goodarzi, H. *et al.* Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* **485**, 264-268, doi:10.1038/nature11013 (2012).
18 Oikonomou, P., Goodarzi, H. & Tavazoie, S. Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. *Cell reports* (2014).
19 Ke, S. *et al.* Quantitative evaluation of all hexamers as exonic splicing elements. *Genome research* **21**, 1360-1374, doi:10.1101/gr.119628.110 (2011).
20 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475, doi:10.1038/nature11396 (2012).
21 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489, doi:10.1038/nature16961 (2016).

Chapter 2: A microRNA-based single-gene circuit buffers protein synthesis rates against perturbations

This was my first project in graduate school. When I started graduate school, I barely knew the difference between transcription and translation (my undergraduate degree was in electrical engineering and I had not taken a biology class since sophomore year of high school). However, I was very fortunate to have excellent mentors in Tim Strovas and my advisor Georg. Tim helped me cut my teeth in experimental and synthetic biology, and I was hooked from the start.

The project took an elegant but simple concept—that incoherent feedforward loops (iFFL) can buffer noise—and set out to prove it experimentally. We took a "Feynman approach" to the problem; rather than identification and study of natural iFFLs in the human genome, we decided to understand them by building our own iFFLs with miRNAs.

The following work was published as:

Strovas, T. J.*, **Rosenberg, A. B.***, Kuypers, B. E., Muscat, R. A., & Seelig, G. (2014). MicroRNA-based single-gene circuits buffer protein synthesis rates against perturbations. *ACS synthetic biology*, *3*(5), 324-331.

* equal contributors

Achieving precise control of mammalian transgene expression has remained a long-standing, and increasingly urgent, challenge in biomedical science. Despite much work, single cell methods have consistently revealed that mammalian gene expression levels remain susceptible to fluctuations (noise) and external perturbations. Here, we show that precise

control of protein synthesis can be realized using a single-gene microRNA (miRNA)-based feed-forward loop (sgFFL). This minimal autoregulatory gene circuit consists of an intronic miRNA that targets its own transcript. In response to a step-like increase in transcription rate, the network generated a transient protein expression pulse before returning to a lower steady state level, thus exhibiting adaptation. Critically, the steady state protein levels were independent of the size of the stimulus, demonstrating that this simple network architecture effectively buffered protein production against changes in transcription. The single-gene network architecture was also effective in buffering against transcriptional noise, leading to reduced cell-to-cell variability in protein synthesis. Noise was up to 5-fold lower for a sgFFL than for an unregulated control gene with equal mean protein levels. The noise buffering capability varied predictably with the strength of the miRNA-target interaction. Together, these results suggest that the sgFFL single-gene motif provides a general and broadly applicable platform for robust gene expression in synthetic and natural gene circuits.

Keywords: Adaptation, noise, microRNA, feed-forward loop

Gene circuits are subject to sudden changes in their environment and random fluctuations in the numbers of their components ("noise") [1]. For example, a reporter gene integrated into the genome of a mammalian cell was shown to be transcribed in bursts, resulting in a wide range of mRNA and protein molecule numbers across a population of cells [2]. Such randomness is sometimes exploited in cellular decision-making but also poses a challenge to the reliability of biological and engineered gene circuits [3-5]. In order to make synthetic gene circuits practically useful for tissue engineering, cellular reprogramming and related fields that require the long-term stable expression

of engineered genetic programs in mammalian cells, we need to develop methods for reliably buffering transgene expression against both global perturbations and transcriptional noise.

Mounting evidence points to an important role for miRNA, a widespread class of posttranscriptional repressors [6], in buffering biological gene circuits against disturbances [7-9]. For example, a miRNA embedded in an incoherent feed-forward loop (IFFL) has been shown to ensure correct eye development in drosophila embryos exposed to temperature fluctuations [10]. A miRNA-based approach to engineering robust gene circuits is appealing because expression constructs for miRNA of arbitrary sequence are readily available [11], and target sites for any miRNA can be inserted into an mRNA of interest [12]. This flexibility has been exploited in the design of a variety of RNA-based synthetic mammalian gene circuits [13-17]. The high degree of modularity also means that miRNA-based approaches to gene expression buffering may be easily adapted to different regulatory contexts.

Theoretical work supports the notion that the microRNA-based IFFL architecture, in which an upstream transcription factor simultaneously activates expression of a mRNA and of a miRNA targeting that same mRNA, is well suited for limiting variability in gene expression [8,18]. Furthermore, IFFLs can exhibit adaptation [19-21]; that is they respond to a sustained change in the level of a stimulus with a transient gene expression pulse before resetting to the original pre-stimulus expression level. Adaptation provides a mechanism for buffering steady state gene expression against global perturbations while transiently propagating information about that perturbation.

The sgFFL architecture, in which an intronic miRNA targets the mRNA from which it originates, forms a compact, single-gene implementation of an IFFL making it an attractive target for engineering low-noise transgenes [22]. Recent work using transient plasmid transfections showed

that a sgFFL can buffer gene expression against cell-to-cell variability in plasmid copy number [16]. However, it remains unclear how noise buffering in a sgFFL architecture is achieved at the single-copy level, how populations can adapt to time-varying perturbations, or if the steady state protein levels may be predictably tuned.

We engineered a family of sgFFL variants with different biochemical parameters and integrated the constructs into the genome to create stable cell lines. Genomic integration allowed us to quantify miRNA, mRNA, and protein dynamics over extended time periods, and made it possible to characterize steady state gene expression noise without confounding factors due to plasmid copy number variations. We found that this network achieved biochemical adaptation and was effective in buffering against transcriptional noise, leading to reduced cell-to-cell variability in protein synthesis. We varied the number and type of miRNA target sites as well as the miRNA production levels and showed that steady state levels can be tuned with buffering increasing for stronger interactions. Together, our results suggest that the sgFFL mechanism provides a robust and modular mechanism for buffering gene expression against perturbations.

Results and Discussion

Adaptation in a single-gene network. We built a synthetic autoregulatory gene circuit by inserting an intron containing the mouse mir-124-3 gene into a red fluorescent reporter (mCherry). The pre-mRNA is transcribed from a Doxycycline (Dox)-inducible promoter (CMV/TO), leading to coexpression of mir-124 and mCherry [23] (see Fig. 1a). To create a repressive regulatory link between the miRNA and the mCherry transcript, we added a truncated version of the mir-124-regulated 3'UTR of the *Vamp3* gene to the mRNA [24]. This 3'UTR contains one 6-mer and three 7-mer target sites complementary to the mir-124 seed region (nucleotides 2-8 from the 5' end of

the miRNA). To better observe the dynamics of gene expression, we destabilized the mCherry protein using a standard PEST degradation tag. A stable cell line was created by genomic integration of the circuit into a Flp-In™ T-REx™ 293 cell line that expresses a constant background of the TetR repressor protein. By tightly binding to TetR and relieving its repressive activity, Dox acts as an activator for both the mCherry mRNA and mir-124. The corresponding network diagram is shown in Fig. 1b with the negative regulatory link highlighted in red. We also engineered a control cell line containing an expression construct without the target-containing Vamp3 3'UTR, referred to as open-loop control. Plasmid maps for both cell lines are shown in Fig. S1.

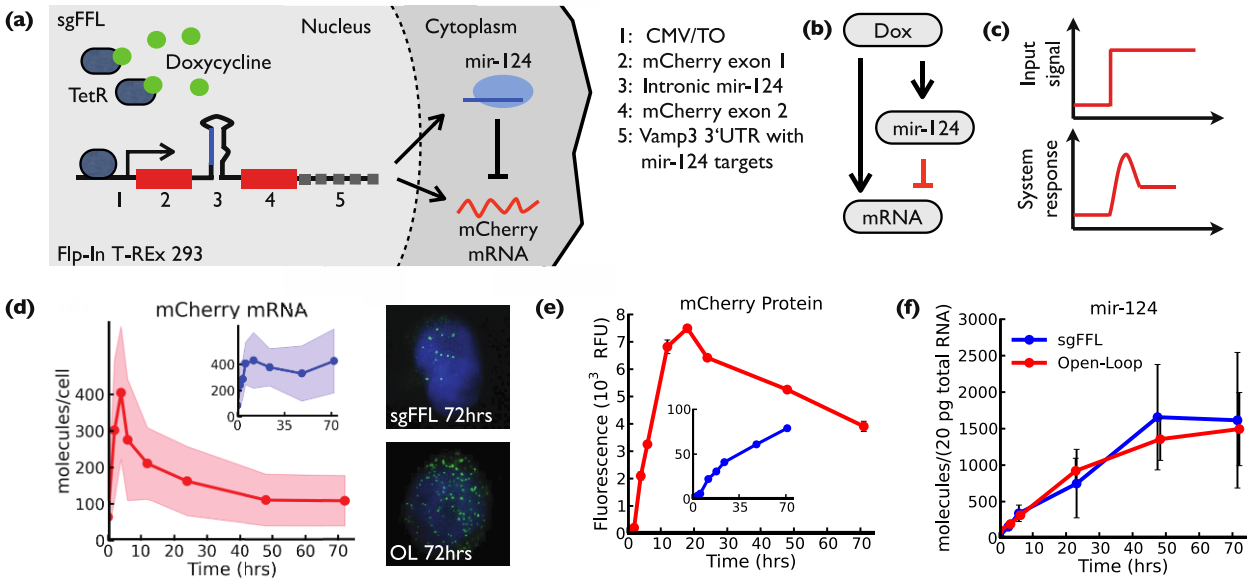


Figure 1: A miRNA-based single-gene circuit can generate transient pulses in gene expression. (a) Induction of the sgFFL cell line by Dox results in transcription of a pre-mRNA containing the primary mir-124 and the mCherry coding sequence. The mature mir-124-RISC complex then the targets 3'UTR of the mCherry transcript. The mir-124 targeted 3'UTR was deleted in the open-loop cell line. (b) The miRNA expression construct implements an incoherent feed-forward motif. (c) The sgFFL exhibits transient pulse generation and adaptation in response to sustained change in the level of the upstream regulator. (d) The sgFFL mCherry mRNA shows a pronounced pulse around t=5hrs (red). No peak is observed in the control cell line (inset, blue). Data was obtained using single molecule FISH and both mean and variance are indicated. Representative images for the two cell lines at the 72hr time point are shown. (e) A pulse is also clearly visible in the expression of the mCherry protein in the sgFFL line but not in the control (inset). Protein expression was assayed using flow cytometry (RFU, relative fluorescence unit). (f) RT-qPCR data show similar accumulation of miRNA for the sgFFL and control cell lines.

Sustained activation of the sgFFL with Dox led to a transient pulse in gene expression (Fig. 1c,d). A sharp peak in mRNA levels as measured by single-cell RNA FISH [2] was observed around 5 hours after induction (Fig. 1d), before mRNA levels (72 hour time point) returned close to their pre-stimulus value. The measured mRNA dynamics display near perfect adaptation, a behavior compatible with feed-forward loop architectures [19-21]. Protein expression was measured by flow cytometry and followed a similar course to the mRNA expression (Fig. 1e). Unlike the mRNA, which is actively targeted by the miRNA, proteins are only slowly cleared from the cell resulting in a slow post-stimulus decay of fluorescence. No pulse was seen in the mRNA or protein levels in the open-loop control cell line. Instead protein and mRNA both monotonically increased towards their respective steady state values (insets, Fig. 1d,e).

We used quantitative PCR to directly measure the levels of mature mir-124. MiRNA levels in the sgFFL and open-loop control cells were almost identical and, over the course of the experiment, gradually approach steady state without pulsing (Fig. 1f). The slow approach to steady state is primarily due to the high stability of the RISC-bound miRNA. Furthermore, the similarity in the miRNA levels for the sgFFL and control cell lines implies that mRNA targeting does not dramatically accelerate miRNA turnover [25]. To confirm that the repression of mCherry is in fact due to the miRNA rather than competition for cellular resources or other non-specific effects, we transfected sgFFL cells with a LNA-modified antisense oligonucleotide complementary to mir-124. Antisense transfection restored red fluorescence confirming direct repression of the target by the miRNA (Fig. S2). We further confirmed this result by engineering a stable sgFFL cell line where we deleted the primary miRNA from the intron. This cell line did not show pulsing and behaved like the original open-loop control cell line (Fig. S3).

Gene expression buffering. The steady state levels in an ideal adaptive system should be insensitive to the size of the stimulus (Fig. 2a). To explore if this is true for the sgFFL, we performed experiments in which we systematically varied the level of induction using Dox (Fig. 2b,c). We identified a regime of Dox concentrations that resulted in intermediate promoter activities as evidenced by the variations in initial rates and peak heights. Traces that were clearly distinct at the peak converged to the same, sub-peak steady state level (Fig. 2b). For example, induction with 1-15 ng/ml Dox led to different initial rates and pulse amplitudes but very similar steady states. Such convergence of traces corresponding to different promoter activities is not seen in the time-course data for the control cell line. Instead, initial differences were amplified over the course of the experiment and resulted in clearly distinct end points (Fig. 2c). We note that the maximal mean fluorescence achievable in the sgFFL cell population is lower than is the case for the open-loop control cell line, because at the highest level of induction the miRNA has a substantial repressive effect.

In most of our experiments we follow the system dynamics over a three-day period. However, we note that the system has not completely reached steady state at the 72 hrs time point and some residual fluorescent protein from the transient peak remains in the cells. We thus also ran an extended 120 hrs time course experiment with the sgFFL cell line and using several different Dox concentrations (Fig. S4). These longer time course experiments agree with the findings shown in the main paper Fig. 2 and suggest that all dynamics relevant to gene expression buffering can be observed within three days.

Figures 2d,e summarize the results on gene expression buffering by comparing the steady state expression levels of the control and sgFFL data as a function of Dox concentration. This analysis again clearly shows that steady state protein expression saturates and becomes insensitive to

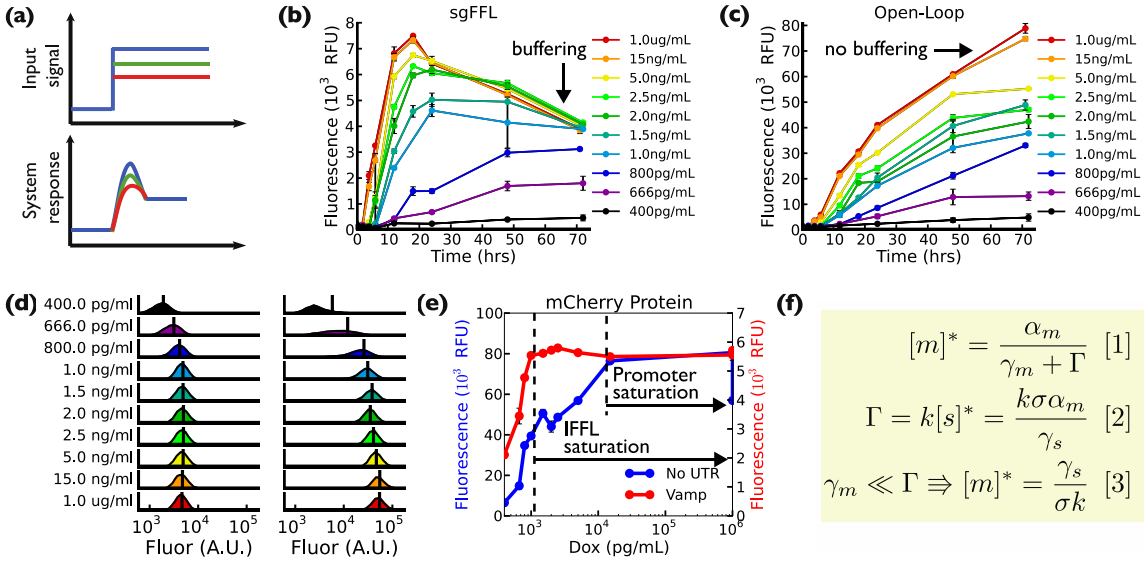


Figure 2: A miRNA-based sgFFL buffers gene expression against a sustained external stimulus. (a) In an adaptive gene circuit steady state protein expression levels are independent of the input amplitude. However, the pulse is proportional to the input. **(b)** Varying levels of Dox induction lead to distinct peak amplitudes that converge to the same steady state level. **(c)** No steady state buffering or pulsing is observed in the open loop control cell line. **(d)** Fluorescence histograms for the sgFFL (left) and control cell line (right) at different levels of induction. Data was taken 72 hours after induction and 30,000 cells were used in each experiment. Means are indicated with a black line. **(e)** Mean fluorescence at 72 hours as a function of Dox. The open loop control cell line can reach higher fluorescence levels at full induction, but the sgFFL cell line reaches its maximal fluorescence at lower promoter activities. **(f)** Buffering can be understood from simple steady state model of gene expression (see Supplementary text for details). [m]* is the steady state levels of the mCherry mRNA (Eq. [1]). The mRNA production rate is α_m and the miRNA production rate is $\sigma\alpha_m$. Here σ accounts for different production efficiencies of the mRNA and miRNA. The native mRNA degradation rate is γ_m, the miRNA degradation rate is γ_s, and Γ is the degradation rate of the mRNA due to the miRNA which is proportional to [s]*, the steady state amount of the miRNA (Eq. [2]). As mRNA degradation due to the miRNA becomes the main source of degradation ($\Gamma >> \gamma_m$), the steady state levels of mRNA become independent of the production rate α_m (Eq. [3]).

promoter activity before the promoter is fully active. The observed gene expression buffering can be understood from a model that compares RNA production and degradation rates (see Fig. 2f, Supplementary text and Fig. S5): the mRNA steady state level is determined by the ratio of the mRNA production and degradation rates, with both miRNA-induced and miRNA-independent processes contributing to the degradation rate. In the sgFFL the production rates of the mRNA and miRNA are proportional to one another such that an increase in the mRNA production is always accompanied by an increase in miRNA production, and consequently in an increase in miRNA-induced degradation. Thus, steady state mRNA levels become independent of the mRNA

production rate if the rate of mRNA degradation due to the miRNA is large compared to the native rate of mRNA degradation.

Noise suppression. How is buffering manifested at the single cell level? Cell-to-cell variability in TetR expression, Dox uptake and other biochemical parameters naturally creates a range of promoter activities even among genetically identical cells in the same environment (Fig. 3a) [4]. This randomness results in a distribution of the experimentally measured fluorescence values in a population of cells (e.g. Fig. 2d). We used the coefficient of variation (CV, standard deviation divided by the mean) of the fluorescence distribution as a measure for the biochemical noise. Fig. 3d shows the CV as a function of the mean fluorescence for the control and sgFFL cell lines. Intriguingly, if we compare populations with the same mean, we find that the noise for the sgFFL line is up to 5-fold lower than noise for the control cell line, suggesting that the sgFFL network architecture acts as a buffer against variability in an upstream regulator. This point is stressed by the scatter plots in Fig. 3c showing two cell populations with similar mean fluorescence: we found that the range of fluorescence values in the sgFFL population is considerably narrower than in the control population at any given cell size.

To achieve the same mean fluorescence in a population of sgFFL and open-loop control cells, it was necessary to more strongly induce the sgFFL. It is tempting then to assume that noise suppression is simply the result of comparing two processes corresponding to different underlying promoter activities. In that case it is expected that noise should be lower for the more transcriptionally active promoter [26]. However, while this mechanism contributes to the observed effect it does not appear to be the only reason for noise reduction in the sgFFL cell line. In fact,

we note that noise was lower in the sgFFL cell line than in the open loop control even when compared at the same level of promoter induction (Fig. 3d).

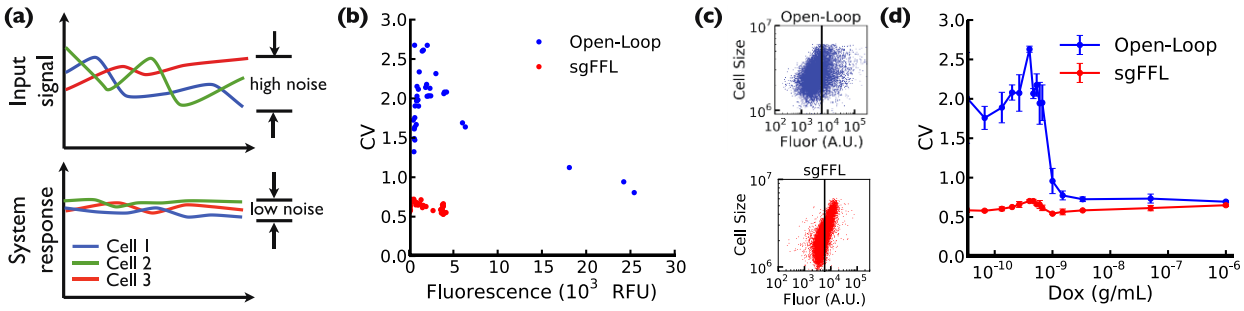


Figure 3: Single-cell analysis reveals noise suppression in a sgFFL. (a) The sgFFL motif is predicted to buffer the expression of the target gene against variability in the upstream regulator. **(b)** Each data point in the plot corresponds to the fluorescence mean and coefficient of variation of the fluorescence distributions shown in Fig. 2d. All data were collected 72 hours after cells were induced with varying amounts of Dox. At the same mean, the noise in the sgFFL cell line is up to 5-fold lower than in the control. **(c)** Each data point corresponds to a single cell from a population of genetically identical cells. Cell size (forward scatter) is plotted against fluorescence. Black lines indicate mean fluorescence. At any given cell size, the distribution of fluorescence values is narrower for the sgFFL cells. **(d)** Noise for sgFFL and open loop control cell lines plotted against the concentration of Dox. Noise is lower in the sgFFL cell line for all Dox levels.

Finally, we observed that for both cell lines noise is highest for Dox concentrations corresponding to intermediate fluorescence values (Fig. S6). This is not surprising since in this regime small differences between cells are amplified by the strong non-linearity of the promoter response function. Similar results have previously been reported in the literature [27]. At all levels of induction, the CV values measured in our experiments are in the range of values previously reported for endogenous proteins in mammalian cells [28].

Tuning network parameters by varying the number of binding sites, interaction type and miRNA number. Next we set out to demonstrate that adaptation and noise suppression can be observed over a range of biochemical network parameter values. Following Ref. [24] we eliminated either one or two 7-mer seed target sites from the Vamp3 3'UTR and then generated sgFFL

expression systems and cell lines sgFFLΔ3 and sgFFLΔ23 based on these modified 3'UTRs (see Fig. 4a and S1). The steady state analysis in Fig. 4b shows that buffering is observable in sgFFLΔ3 and sgFFLΔ23. Fig. 4c demonstrates that noise is suppressed compared to the control in both sgFFL mutant cell lines. Fig. S7 shows time-course flow cytometry data for these constructs; protein expression pulsing becomes weaker with decreasing number of target sites and steady state levels are inversely correlated to the number of targets (Fig. 4b). As expected from our model, buffering becomes less pronounced with decreasing strength of the interactions but is observable over a range of parameter values.

To investigate if adaptation and buffering are observable with different types of miRNA targets and interaction mechanisms, we also created sgFFL constructs with synthetic 3'UTRs that contained either three siRNA-like or seven bulge target sites (Fig. 4d, S1 and S9). The corresponding stable cell lines are sgFFL-3xsiRNA and sgFFL-7xBulge. In the siRNA-like constructs, the sites are fully complementary to the miRNA, which results in cleavage of the mRNA target by the miRNA-RISC complex. In bulge constructs, the targets are fully complementary except for the central three nucleotides of the miRNA, which inhibits catalytic cleavage but still provides very strong interactions between miRNA and mRNA [29]. In agreement with our model (Fig. 2f), these strong interactions lead to very clear signatures of adaptation in the time-course data (Fig. S7) and result in strong noise suppression (Fig. 4b,c). However, we note that noise is higher in the sgFFL-7xBulge cell line than either the sgFFL or sgFFL-3xsiRNA cell lines (Fig. 4c), even though buffering is very pronounced at the level of the population mean (Fig. S7c). In previous work[29], synthetic mRNAs with multiple bulge targets have been used as "sponges" for binding and inhibiting endogenous miRNAs. We speculate, that we observe a similar effect here: given the very large number of potential target sites it is possible that in some

cells all miRNA are bound to only a subset of the available mRNA targets, meaning that other mRNA can escape regulation resulting in increased variability.

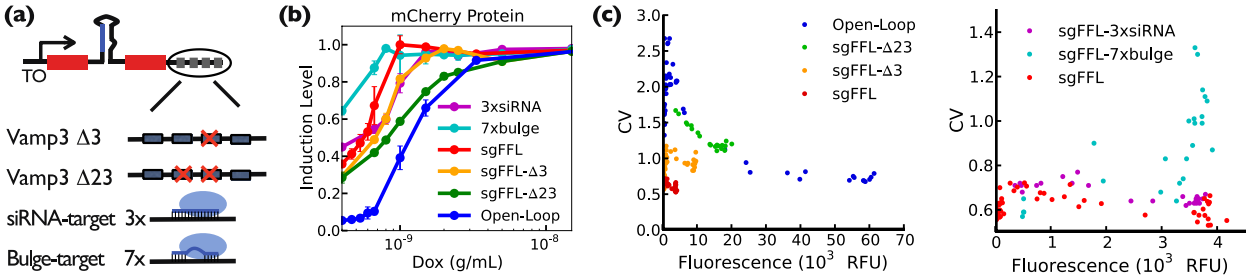


Figure 4: Buffering is tunable by changing the strength of the interaction between miRNA and target. (a) One and two mir-124 seed sites were deleted in the sgFFL-Δ3 and sgFFL-Δ23 cell lines. The 3'UTRs of sgFFL-3xsiRNA and sgFFL-7xbulge were built with three siRNA-like targets (fully complementary to mir-124) and seven bulge targets (fully complementary except central three nucleotides) respectively. (b) Mean fluorescence at 72 hours as a function of Dox for all cell lines including open-loop control and sgFFL. The fluorescence for each cell line is normalized to the respective maximal induction levels. All sgFFL cell lines reach their maximal fluorescence at lower promoter activities than the control, indicative of a buffering effect. The most pronounced buffering is observed with the 7xbulge targets. Buffering becomes less pronounced as interactions become weaker. Full time-course data for all cell lines is shown in the Fig. S7. (c) Noise is lower in any of the sgFFL cell lines than in the control. Noise reduction becomes more pronounced with increasing strength of interactions, except for the sgFFL-7xbulge line, which exhibits comparably high variability.

We next asked if increasing the miRNA production rate could have a similarly pronounced effect on adaptation and buffering as increasing the number of miRNA targets, as would be expected from our model (Fig. 2f). To increase the production rate of the miRNA without changing the production rate of the mRNA, we generated a sgFFL variant where two copies of the same primary microRNA were inserted into the intron (Fig. S8). RT-qPCR data for the miRNA confirm the increase in miRNA production (Fig. S8). The flow cytometry data show an earlier and lower-amplitude peak in protein expression, consistent with increased miRNA production in (Fig. S8). Furthermore, protein steady state levels are reduced as expected given the higher steady state concentration of miRNAs. Together, these results confirm that we can predictably tune steady state levels as well as the degree of buffering.

Multiple naturally occurring instances of the sgFFL motif have been experimentally identified [30-32] and several more have been predicted computationally [22,31], suggesting an important role for this motif in biology. Thus our results on disturbance rejection not only demonstrate that the sgFFL architecture forms a broadly useful tool for buffering transgenes against perturbations but also that it could provide a mechanism for stabilizing protein expression in biological gene cirucits.

We also expect our results on noise suppression to apply to a broader class of endogenous miRNA-based IFFLs where the miRNA and target gene are expressed as independent transcriptional units [10,33-35], if noise affecting the two promoters is correlated [36]. This will be the case, for example, for variations in transcription factor concentration or activity. Conversely, if two promoters are subject to uncorrelated noise we would not expect to observe noise suppression. Importantly, however, adaptation and buffering against global perturbations should be observable independently of the exact arrangement of regulatory elements at the DNA level and may in fact be the most important biological role for this motif.

Given the slow degradation rate of the miRNA [37], the sgFFL has fundamental limits on the types of noise it can filter. Perturbations in a cell that occur over long timescales, for example the accumulation of excessive TetR or global transcription factors, will be effectively filtered since the miRNA has time to compensate. However, perturbations that occur on the timescale of hours rather than tens of hours or days, will not be completely filtered, because the miRNA levels cannot change fast enough to compensate against resulting changes in transcription.

We found that the pulse amplitude for all adaptive sgFFLs is proportional to the size of the perturbation even though the steady state is buffered against that same perturbation. These different behaviors are apparent in Fig. S9, which compares protein expression near the pulse maximum to the steady state. By multiplexing responses across different timeframes, this circuit

can thus buffer gene expression without losing information about the input signal. In spite of its simplicity, similar endogenous network motifs could serve to protect the current cell state against sudden changes in the environment while simultaneously activating new gene expression programs that enable cells to more permanently adapt to large environmental changes [38].

How does our system compare to other mechanisms for noise reduction such as autoregulatory feedback [27,39]? Although feedback provides similar levels of noise suppression, the input-output characteristics of the promoter transfer function are very different. Autoregulatory feedback linearizes the promoter response function, and the rate of transcription is directly proportional to the size of the stimulus. In contrast, the sgFFL shows an all-or-none behavior with a rapid transition from the OFF to the ON state. Furthermore, for a wide range of promoter activities, transcription is independent of the levels of induction. These observations suggest different and complementary regulatory roles for these mechanisms.

In conclusion, we here quantitatively characterized a simple and modular mechanism for buffering transgenes in mammalian cells against perturbations. In combination with recent work on measuring and characterizing miRNA levels and interaction parameters in cells (see e.g. [25,40,41]) our results suggest a path towards the rational design of complex molecular circuits with controlled temporal behaviors that are stably integrated and work reliably in mammalian cells. Engineered molecular circuits that use miRNAs as inputs and network components [13,14,17] could thus eventually become an engineering technology with applications that range from gene therapy to the control of differentiation in stem cells.

Methods

Plasmids. Complete plasmid maps and sgFFL motif details are shown in Figure S1. Plasmids will be made available through Addgene.

Cell Culture. To improve cell adhesion, all culture dishes were coated with Extra Cellular Matrix (ECM) gel from Engelbreth-Holm-Swarm murine sarcoma (SigmaAldrich) diluted with Alpha-MEM media (Mediatech) 1:200 for 16-24 hours then rinsed with 1X Dulbecco's Phosphate Buffered Saline (DPBS; Mediatech) immediately before cells were plated. Cells were cultured in alpha-MEM media supplemented with 10% Tet System Approved fetal bovine serum (FBS; Clontech), penicillin (100 IU/ml; Invitrogen), streptomycin (100 ug/ml; Invitrogen), and L-Glutamine (292 ug/ml; Invitrogen).

Selection of Stable Cell Lines. Transgenic strains were made in the Flp-In™ T-REx™ 293 cell line. The day before transfection, 1.5 million cells per well were seeded into a 6-well plate. 8 ug of pcDNA5 plasmid with 72 ug of pOG44 plasmid were transfected using Lipofectamine 2000™ (Invitrogen) according to the manufacturer's protocol. Growth media was replaced 8 hours post-transfection. Transfected cells were harvested in 1 mL 0.25% Trypsin-EDTA 48 hours post-transfection and re-plated at 1:5, 1:10 and 1:50 dilutions in Alpha-MEM media supplemented as previously described with the addition of Blasticidin (15 ug/ml; Invivogen) and Hygromycin B (100 ug/ml; Invivogen). Media with Blasticidin and Hygromycin was replaced 3 days and 7 days post-transfection. Once visible, 10-17 days post-transfection, individual colonies were dislodged in 250 uL 0.25% Trypsin-EDTA and moved into a 24-well plate, expanded, screened for phenotype, and propagated for this study.

Time-Course Experiments. Cells were seeded into ECM-coated 24-well plates, at a density of 50,000 cells per well, approximately 73 hours prior to collection. During the period between

seeding and collection, cells were maintained in alpha-MEM media supplemented as above. Wells were induced, in duplicate, at 72, 48, 24, 18, 12, 6, 4, 2, and 0 hours before collection by the addition of doxycycline hyclate (Dox, stock concentration: 20ug/ul). For collection, cells were harvested 12 wells at a time by first aspirating growth media, followed by addition of 100 ul 0.25% Trypsin-EDTA (Invitrogen) and resuspension in 250 ul 1X DPBS with 2% (v/v) FBS. Cells were strained through a 40 uM filter before flow cytometry. For each reaction condition, a populations of 30,000 cells was collected and analyzed on an Accuri C6 flow cytometer.

mRNA and miRNA Quantitative RT-PCR. Total RNA was purified using the miRNeasy kit (Qiagen) and 20 units of SUPERase-In^{TM} (Applied Biosystems) was added to both the on column DNase reactions and the final purified total RNA. RNA concentrations were measured using a NanoDrop spectrophotometer (Thermo Scientific). 260/280 nm ratios were consistently greater than 1.8 and RNA integrity was spot-checked using native agarose gels. Reverse transcription reactions were conducted using the Taqman© microRNA Reverse Transcription Kit scaled to a 22.5 ul volume with Taqman© microRNA Assay reverse transcription primers (for miRNA; Applied Biosystems) and Oligod(T)23 VN primers (for mRNA). Quantitative PCR was conducted on a CFX96 real-time PCR machine using SsoFast EvaGreen (mRNA) and Probes (miRNA) Supermixes (Biorad) following manufacturer protocols. RNA was detected using Taqman© microRNA Assays (miRNA; Applied Biosystems) and primers specific to mCherry (Fwd: GGCTTCAAGTGGGAGCGCGT, Rev: GCATTACGGGGCCGTCGGAG; IDT) and TBP (Fwd: CACGAACCACGGCACTGATT, Rev: TTTTCTTGCTGCCAGTCTGGA; IDT). Data were normalized using uninduced samples, TBP [42] and hsa-mir-9* [43] as controls with the ΔΔCt method[44]. hsa-mir-124, hsa-mir-9*, mCherry and TBP qPCR reactions had efficiencies of 95.9%, 101.9%, 87.9% and 88.3%, respectively.

Single-molecule FISH. A FISH probe set [2,45], consisting of 48 oligonucleotides complementary to the exons coding mCherry and H2B, was designed using Stellaris Probe Designer software. The probe set was synthesized by the manufacturer (Biosearch Technologies, CA) labeled with the far-red fluorophore Quasar 670 (similar to Cy5). Cells were prepared for imaging using a modified version of the manufacturers guidelines for cells in suspension. Trypsinized cells were fixed, permeabilized and stored at -20^{o}C. Probes, incubated overnight at a concentration of 125 nM, were hybridized to target mRNA in a 20% formamide hybridization solution. Cells were imaged in Vectashield solution (Vecta Labs, CA) to minimize photobleaching. Z-stacks were taken across the entirety of the cell on a Nikon Ti Eclipse, with 100x objective and CoolSnapEZ camera. Images were analyzed using SpotFinding Suite [46]. A manually curated training set of "true" spots was used to determine the fitting parameters to identify unclassified candidate spots.

LNA Transfections. In a 24-well plate, fully induced sgFFL cells were transfected with 10 nM LNA (Exiqon) using RNAiMAX^{TM} (Invitrogen) according to the manufacturer's protocol.

Acknowledgements. We would like to thank Michael Elowitz, Eric Klavins, Long Cai and Mary Dunlop for their insightful comments on this manuscript. This work was supported by NSF CAREER Award 0954566 and a Burroughs Wellcome Career Award at the Scientific Interface to GS.

Supporting Information Available. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

1 Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226 (2008).

2 Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).

3 Balázsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910-925 (2011).
4 Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962 (2005).
5 Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965-1969 (2005).
6 Carthew, R. W. & Sontheimer, E. J. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642-655 (2009).
7 Ebert, M. S. & Sharp, P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149**, 515-524 (2012).
8 Hornstein, E. & Shomron, N. Canalization of development by microRNAs. *Nat. Genet.* **38**, S20-S24 (2006).
9 Mendell, J. T. & Olson, E. N. MicroRNAs in stress signaling and human disease. *Cell* **148**, 1172-1187 (2012).
10 Li, X., Cassidy, J. J., Reinke, C. A., Fischboeck, S. & Carthew, R. W. A microRNA imparts robustness against environmental fluctuation during development. *Cell* **137**, 273-282 (2009).
11 Chang, K., Elledge, S. J. & Hannon, G. J. Lessons from Nature: microRNA-based shRNA libraries. *Nat. Methods* **3**, 707-714 (2006).
12 Brown, B. D. *et al.* Endogenous microRNA can be broadly exploited to regulate transgene expression according to tissue, lineage and differentiation state. *Nat. Biotechnol.* **25**, 1457-1467 (2007).
13 Deans, T. L., Cantor, C. R. & Collins, J. J. A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells. *Cell* **130**, 363-372 (2007).
14 Tigges, M., Dénervaud, N., Greber, D., Stelling, J. & Fussenegger, M. A synthetic low-frequency mammalian oscillator. *Nucleic Acids Res.* **38**, 2702-2711 (2010).
15 Beisel, C. L., Bayer, T. S., Hoff, K. G. & Smolke, C. D. Model-guided design of ligand-regulated RNAi for programmable control of gene expression. *Mol. Syst. Biol.* **4** (2008).
16 Bleris, L. *et al.* Synthetic incoherent feedforward circuits show adaptation to the amount of their genetic template. *Mol. Syst. Biol.* **7** (2011).
17 Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R. & Benenson, Y. Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science* **333**, 1307-1311 (2011).
18 Osella, M., Bosia, C., Corá, D. & Caselle, M. The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS Comp. Biol.* **7**, e1001101 (2011).
19 Goentoro, L., Shoval, O., Kirschner, M. W. & Alon, U. The incoherent feedforward loop can provide fold-change detection in gene regulation. *Mol. Cell* **36**, 894-899 (2009).
20 Ma, W., Trusina, A., El-Samad, H., Lim, W. A. & Tang, C. Defining network topologies that can achieve biochemical adaptation. *Cell* **138**, 760-773 (2009).
21 Sontag, E. D. Remarks on feedforward circuits, adaptation, and pulse memory. *IET Syst. Biol.* **4**, 39-51 (2010).
22 Bosia, C., Osella, M., Baroudi, M. E., Corà, D. & Caselle, M. Gene autoregulation via intronic microRNAs and its functions. *BMC Syst. Biol.* **6**, 131 (2012).

23 Makeyev, E. V., Zhang, J., Carrasco, M. A. & Maniatis, T. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol. Cell* **27**, 435-448 (2007).
24 Karginov, F. V. *et al.* A biochemical approach to identifying microRNA targets. *Proc. Nat. Acad. Sci. USA* **104**, 19291-19296 (2007).
25 Baccarini, A. *et al.* Kinetic analysis reveals the fate of a microRNA following target regulation in mammalian cells. *Curr. Biol.* **21**, 369-376 (2011).
26 Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415-418 (2004).
27 Nevozhay, D., Adams, R. M., Murphy, K. F., Josić, K. & Balázsi, G. Negative autoregulation linearizes the dose–response and suppresses the heterogeneity of gene expression. *Proc. Nat. Acad. Sci. USA* **106**, 5123-5128 (2009).
28 Sigal, A. *et al.* Variability and memory of protein levels in human cells. *Nature* **444**, 643-646 (2006).
29 Ebert, M. S., Neilson, J. R. & Sharp, P. A. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature methods* **4**, 721-726 (2007).
30 Dill, H., Linder, B., Fehr, A. & Fischer, U. Intronic miR-26b controls neuronal differentiation by repressing its host transcript, ctdsp2. *Genes Dev.* **26**, 25-30 (2012).
31 Megraw, M. *et al.* Isoform specific gene auto-regulation via miRNAs: a case study on miR-128b and ARPP-21. *Theor. Chem. Acc.* **125**, 593-598 (2010).
32 Sun, Y. *et al.* miR-126 inhibits non-small cell lung cancer cells proliferation by targeting EGFL7. *Biochem. Biophys. Res. Commun.* **391**, 1483-1489 (2010).
33 Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-533 (2008).
34 O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V. & Mendell, J. T. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**, 839-843 (2005).
35 Re, A., Corá, D., Taverna, D. & Caselle, M. Genome-wide survey of microRNA–transcription factor feed-forward regulatory circuits in human. *Mol. Biosyst.* **5**, 854-867 (2009).
36 Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183 (2002).
37 Zhang, Z., Qin, Y. W., Brewer, G. & Jing, Q. MicroRNA degradation and turnover: regulating the regulators. *Wiley Interdisciplinary Reviews: RNA* **3**, 593-600 (2012).
38 Yosef, N. & Regev, A. Impulse control: temporal dynamics in gene transcription. *Cell* **144**, 886-896 (2011).
39 Nevozhay, D., Zal, T. & Balázsi, G. Transferring a synthetic gene circuit from yeast to mammalian cells. *Nat. Comm.* **4**, 1451 (2013).
40 Broderick, J. A., Salomon, W. E., Ryder, S. P., Aronin, N. & Zamore, P. D. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA* **17**, 1858-1869 (2011).
41 Béthune, J., Artus-Revel, C. G. & Filipowicz, W. Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. *EMBO Rep.* **13**, 716-723 (2012).
42 Kwon, M. J. *et al.* Identification of novel reference genes using multiplatform expression data and their validation for quantitative gene expression analysis. *PLoS One* **4**, e6162 (2009).

43 Koh, T.-C., Lee, Y.-Y., Chang, S.-Q. & Nissom, P. M. Identification and expression analysis of miRNAs during batch culture of HEK-293 cells. *J. Biotechnol.* **140**, 149-155 (2009).
44 Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2< sup>− ΔΔCT</sup> Method. *Methods* **25**, 402-408 (2001).
45 Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877-879 (2008).
46 Rifkin, S. A. Identifying Fluorescently Labeled Single Molecules in Image Stacks Using Machine Learning. *Methods in Molecular Biology (Clifton, NJ)* **772**, 329 (2011).

Supplementary Materials for

A microRNA-based single-gene circuit buffers protein synthesis rates against perturbations

Timothy J. Strovas, Alexander B. Rosenberg, Brianna E. Kuypers, Richard A. Muscat, and Georg Seelig

Correspondence to: gseelig@uw.edu

This file includes:

Supplementary Text

Figs. S1 to S11

Supplementary Text

Model

The concentrations of mRNA, proteins and miRNA are given by the equations:

[1] $d[m]/dt = \alpha_m - \gamma_m[m] - k[m][s]$

[2] $d[mCh]/dt = \sigma_p[m] - \gamma_p[mCh]$

[3] $d[s]/dt = \sigma\alpha_m - \gamma_s[s]$

Where each parameter is described below:

α_m = mRNA production (mRNA/hr)

α_p = protein production (protein/(mRNA*hr))

γ_m = native mRNA degradation (1/hr)

γ_p = native protein degradation (1/hr)

σ = miRNA efficiency (miRNA/mRNA)

γ_s = native miRNA degradation (1/hr)

k = miRNA-catalyzed mRNA degradation (1/(miRNA*hr))

Solving Equations [1] and [2] in steady state, $d[m]/dt=d[s]/dt=0$ when $k[s] >> \gamma_m$ gives the equation $[m]= \gamma_s / \sigma k$ as shown in Fig. 2f of the main text. Simulation of the model predicts gene expression pulses and the buffering observed in the experiment (Fig. S4). The parameter values used in the simulation are:

α_m = 0-150 (mRNA/hr)

α_p = 10 (protein/(mRNA*hr))

γ_m = 0.1 (1/hr)

γ_p = 0.08 (1/hr)

$\sigma = 0.4$ (miRNA/mRNA)

$\gamma_s = 0.03$ (1/hr)

$k = 0.00075$ (1/(miRNA*hr))

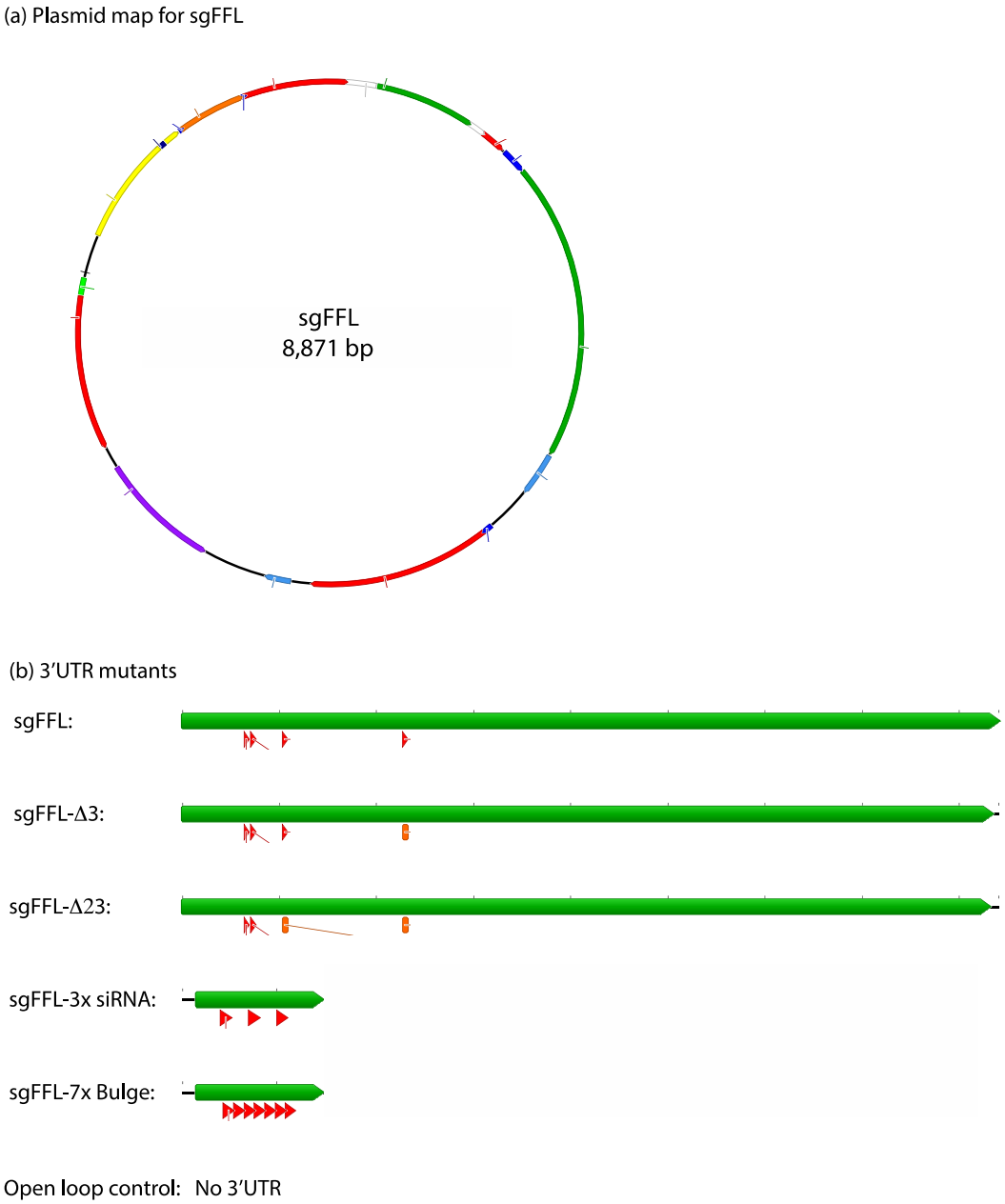


Figure S1: Plasmid map for sgFFL and variants. (a) Complete plasmid map for the sgFFL cell line. **(b**) 3'UTR for all mutants used here. Plasmid maps are otherwise identical to the sgFFL plasmid.

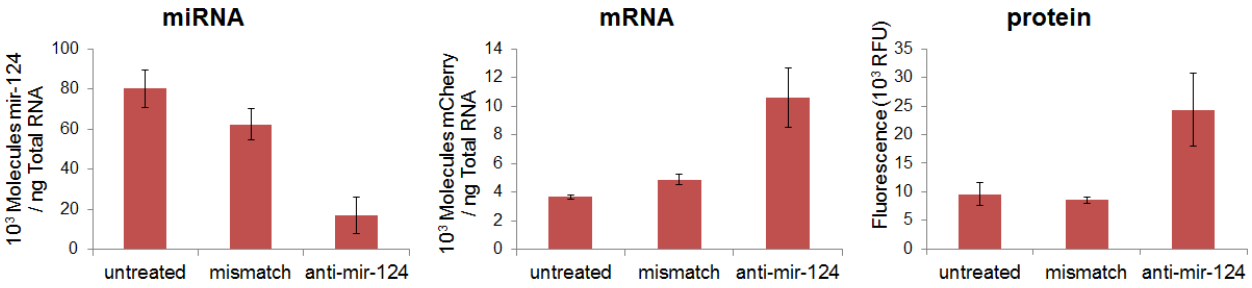


Figure S2: MiRNA knockdown with antisense LNA reduces miRNA levels and increases fluorescence. Fully induced sgFFL cells were transfected with 6 pmol of anti-mir-124 or anti-mir-122 LNA. MiRNA (left panel), mRNA (middle panel) and fluorescence levels (right panel) are compared for untreated cells, cells transfected with a mismatched anti-mir-122 LNA control and an anti-mir-124 LNA. Uninduced cells are shown as an additional control in each panel. The anti-mir-124 antisense LNA reduces the levels of miRNA, which results in increased levels of mRNA and protein. The scrambled control LNA does appear to slightly reduce miRNA levels but this reduction is much less pronounced than for the matching antisense LNA.

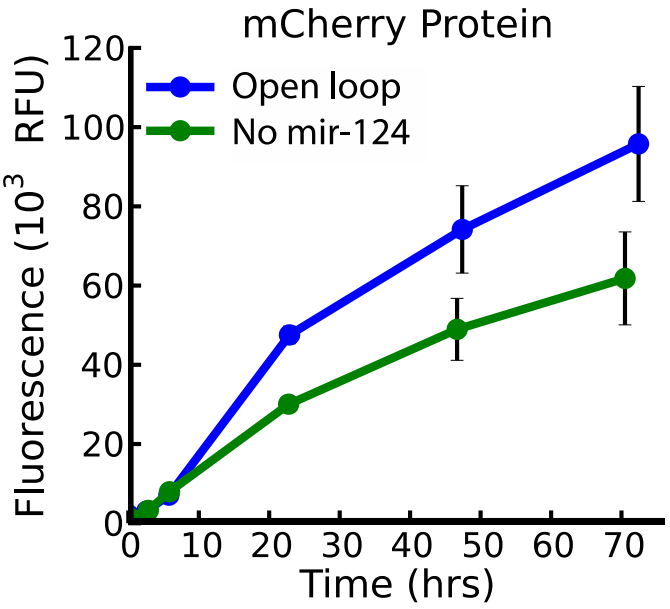


Figure S3: No-miRNA control cell line does not exhibit adaptation. The open loop control line produces miR-124 but lacks a 3'UTR with miRNA targets. The sgFFL-Δmir control contains an intron without the mir-124 gene and consequently does not produce miRNA. This construct contains the same 3'UTR as the sgFFL cell line. Absence of either the intronic miRNA or the 3' UTR removes the interaction between the miRNA and mRNA. Fluorescence for both cell lines is comparable and much higher than that of the sgFFL (not shown here). The slight differences between the two controls are likely due to different mRNA stabilities that result from the much longer 3'UTR of the sgFFL-Δmir line.

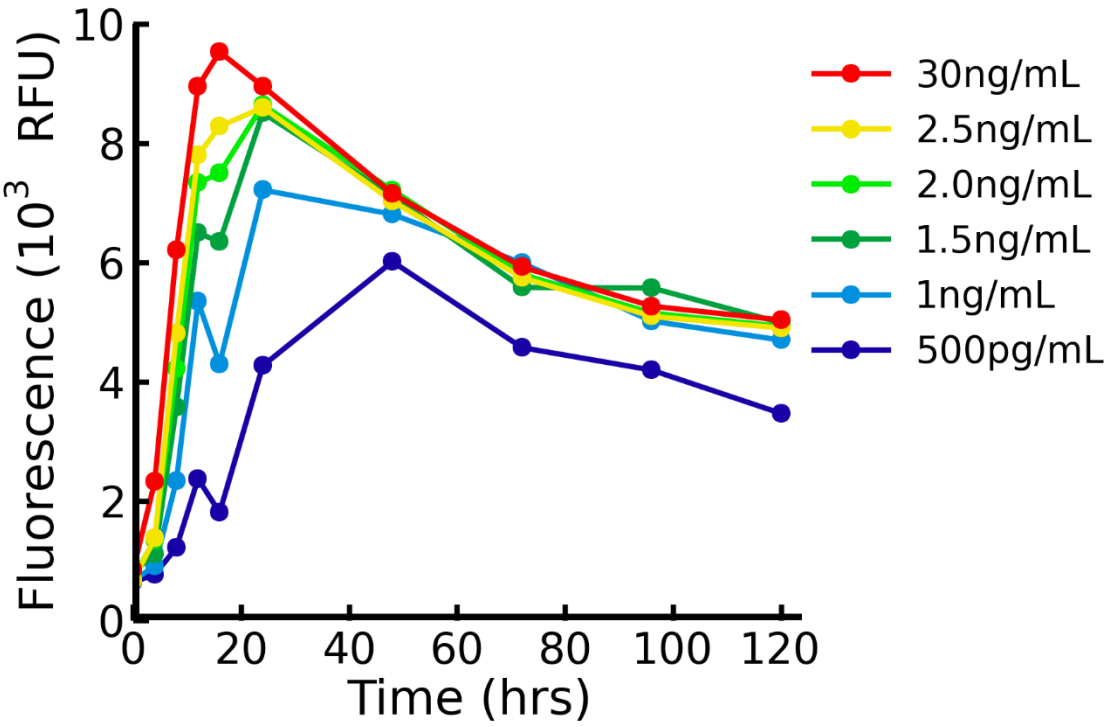


Figure S4: Long-term induction and buffering in the sgFFL cell line. We performed experiments to identify the earliest time point when buffering could be clearly observed. Based on the experiments shown here, we used the 72 hours time point as the end point for all experiments shown in the main paper. Although the system has not fully reached steady state at 72 hours, all relevant buffering dynamics can be observed before then. The data shown are from a single experiment. Cells were passaged after 72 hours and new Dox was added at that point.

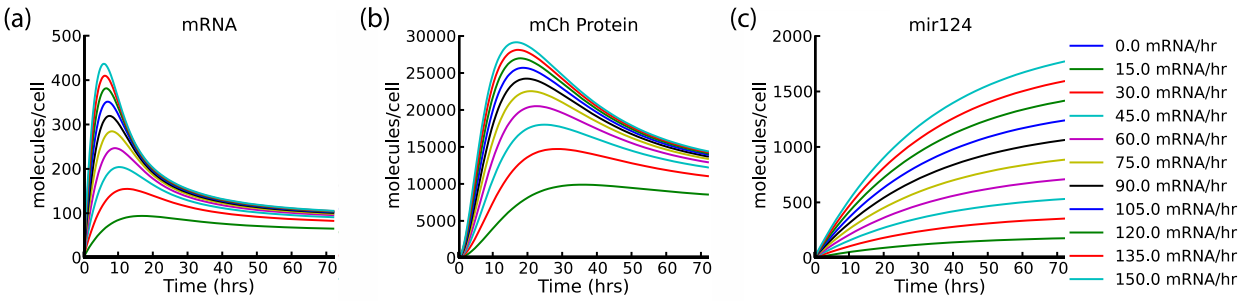


Figure S5: Simulation of the model predicts gene expression pulses observed in the experiment. Parameter values for the simulation are given in the Model section. The degradation rates correspond to half-lives of τ_m = 7 hrs for the mRNA, τ_p = 9 hrs for the protein and γ_s = 23 hrs for the miRNA. MiRNA half-life is consistent with the assumption that miRNA are primarily cleared by dilution during cell division. The production rates were chosen such that the absolute mRNA and miRNA numbers are comparable to the numbers measured experimentally (see Fig. 1 main text).

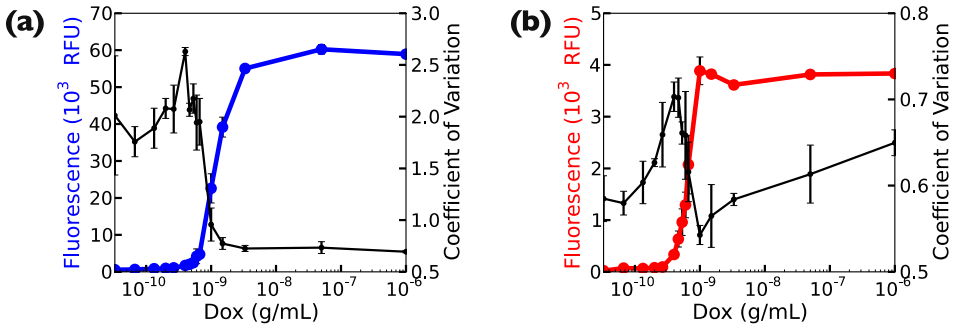


Figure S6: Noise is maximal at the onset of the transition from the OFF to the ON state. Noise and mean fluorescence as a function of promoter activity for **(a)** open loop control and **(b)** sgFFL cell lines.

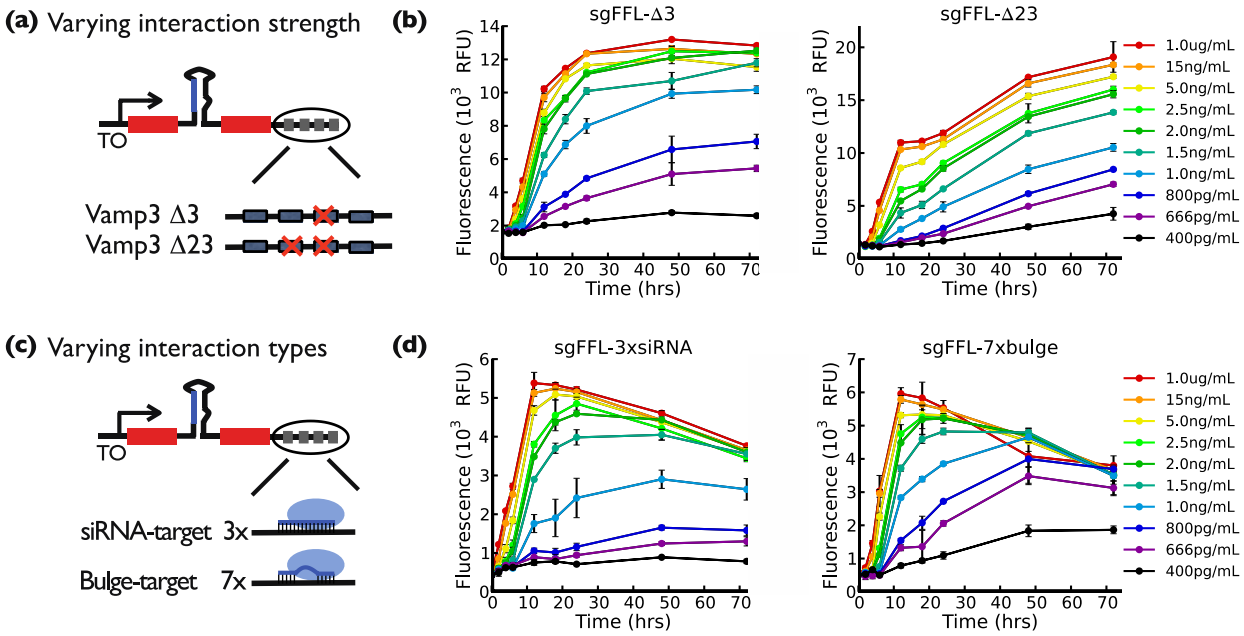


Figure S7: Buffering and steady state levels depend on interaction strength. (a) Cell lines sgFFL-Δ3 and sgFFL-Δ23 contain stably integrated constructs with modified 3'UTRs in which one or two miRNA seed target sites were deleted. (b) Time-course flow cytometry data for the two cell lines show that buffering is observable but becomes less pronounced as interactions become weaker. (c) Cell lines sgFFL-3xsiRNA and sgFFL-7xbulge both contain stably integrated constructs with 3'UTRs that have three siRNA-like targets (fully complementary to mir-124) or seven bulge targets (fully complementary except central three nucleotides). (d) As expected, time-course flow cytometry data show very effective buffering for such strong interactions.

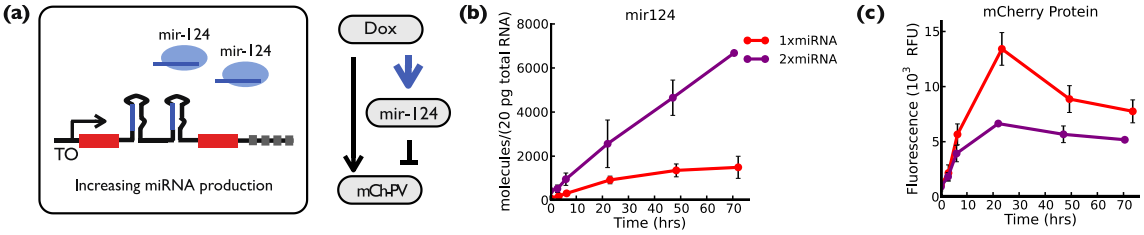


Figure S8. Pulse shape and steady state level depend on miRNA levels. (a) Cell line sgFFL-2xmir contains two copies of the primary mir-124. The blue arrow in the circuit diagram indicates that the microRNA production rate is increased relative to the basic sgFFL cell while mRNA production remains the same. **(b)** RT-qPCR data show increase of miRNA production rate. Time-course flow cytometry data show that increasing miRNA production leads to lower steady state levels (blue trace) compared to the reference cell line (red trace). **(c)** mCherry protein expression. The peak in the sgFFL-2xmir cell line (purple trace) occurs earlier and steady state levels are lower than in the sgFFL (red trace), as expected for a cell line with higher miRNA expression levels.

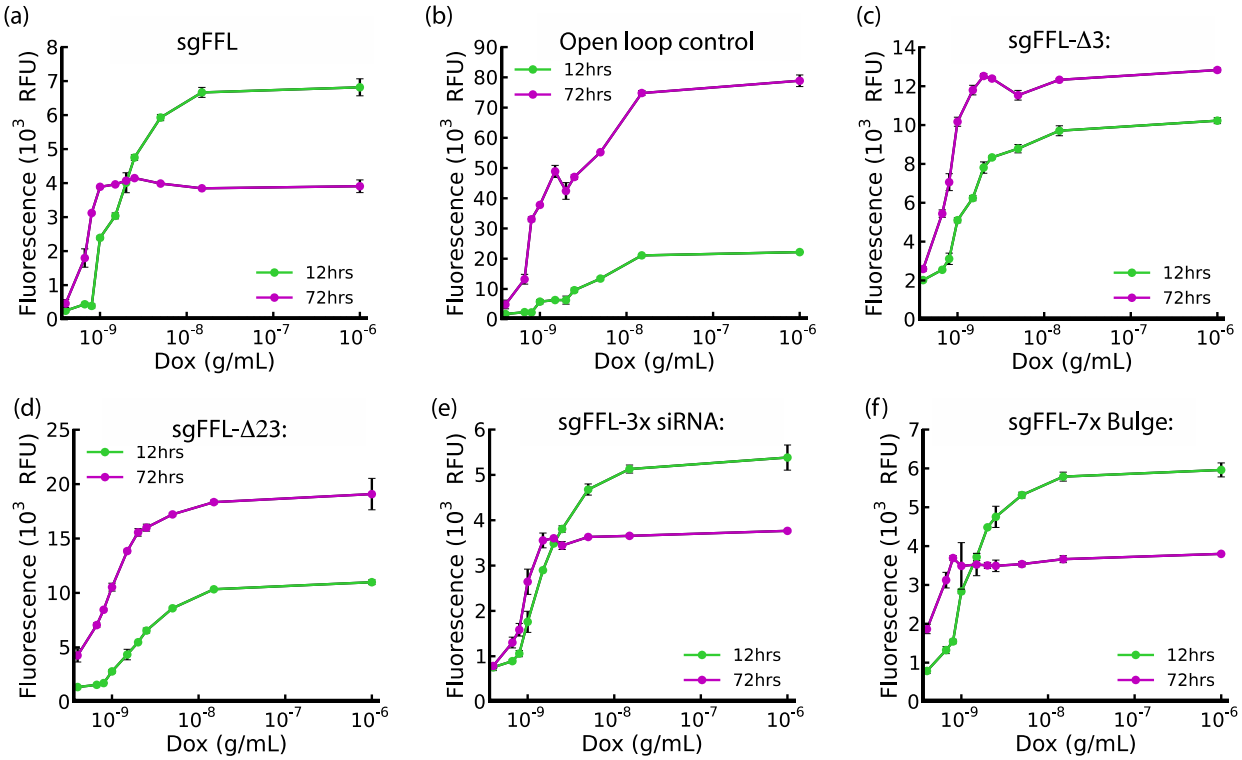


Figure S9. The pulse amplitude encodes information about the size of the perturbation. Protein level as a function of promoter induction (Dox levels) at 12 hrs (green) and 72 hrs (purple) for all cell lines used in this work. **(a)** For the sgFFL cell line, protein levels are clearly distinct at 12 hrs but are buffered at 72 hrs. **(b)** No buffering is seen in the open-loop control. **(c,d)** Limited buffering is observed at 72 hrs for the sgFFL-Δ3 and sgFFL-Δ23 cell lines. **(e,f)** For the sgFFL-3xsiRNA and sgFFL-7xbulge lines, pulse amplitude at 12 hrs is proportional to promoter activity while the steady state protein levels are buffered.

Chapter 3 - Learning the sequence determinants of alternative splicing from millions of synthetic sequences

When I started this project, I knew very little about RNA splicing, RNA-seq, or machine learning. The project actually began with a chance discussion with another graduate student—Robert Egbert. We were discussing oligo synthesis and he mentioned in passing that it was possible to synthesize oligo with completely degenerate bases. At the time, I was studying ways to use RNA splicing in synthetic biology, but designing sequences with desired function was very challenging. So I had the idea of just putting random sequences into alternatively spliced introns and screening for sequences that had the desired output. As a small scale test, I inserted 50 random nucleotides into an alternatively spliced intron and found a large range of variability in how terms of how the different sequences spliced. This piqued both Georg and my interest, but we still had no experience with large scale RNA-seq.

At this point, we met with Jay Shendure and his graduate student Rupali Patwardhan. Jay and Rupali had developed some of the first Massively Parallel Reporter Assays (MPRAs) to measure the effects of variants on both promoters and enhancers. Both Jay and Rupali provided invaluable guidance on both the sequencing end and the experimental design. While getting the assay up and running was no easy task, the real challenge started once we had sequencing data.

Exposing the naïve, young graduate student that I was, I assumed that the data could be analyzed within a few months, and I would have a draft of the paper shortly thereafter. This turned out to be far from the truth. In my initial analysis of the data, I searched primarily for the motifs that were altering splicing. However, Georg, Jay, and I soon realized that we should really be focusing our efforts to build a predictive model of splicing from our data. This starting me down

a long and winding road of which eventually led me shift a large part of my PhD focus towards machine learning.

This project took me about four years to complete, which may sound terrifying to other PIs or students, but those four years were instrumental to my own development and growth as a researcher. I am extremely grateful that neither Georg nor Jay pressed me to publish faster. They encouraged me as I dug deeper and deeper into the data and continued to find new results even years after the data were collected. I understand not every young scientist has this luxury and I am extremely grateful for this opportunity.

The following work was published as:

Rosenberg, A. B., Patwardhan, R. P., Shendure, J., & Seelig, G. (2015). Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell*, *163*(3), 698-711.

Most human transcripts are alternatively spliced, and many disease-causing mutations affect RNA splicing. Towards better modeling the sequence determinants of alternative splicing, we measured the splicing patterns of nearly 2 million (M) synthetic mini-genes, which include degenerate subsequences totaling to nearly 100M bases of variation. The massive size of these training data allowed us to improve upon current models of splicing as well as to gain new mechanistic insights. Our results show that a vast majority of hexamer sequence motifs measurably influence splice site selection when positioned within alternative exons, with multiple motifs acting additively rather than cooperatively. Intriguingly, motifs that enhance (suppress) exon inclusion in alternative 5' splicing also enhance (suppress) exon inclusion in

alternative 3' or cassette exon splicing, suggesting a universal mechanism for alternative exon recognition. Finally, our empirically trained models are highly predictive of the effects of naturally occurring variants on alternative splicing *in vivo*.

Introduction:

Alternative splicing is a major source of proteome diversity in eukaryotes (Nilsen and Graveley, 2010). Regulation of alternative splicing is vital to cellular processes that depend on precise ratios of isoforms. For example, mutations that lead to even subtle changes in the ratio of MAPT isoforms 3R and 4R cause an inherited form of dementia (Garcia-Blanco et al., 2004). While new sequencing technologies have enabled the comprehensive cataloging of human genetic variation, the functional consequences of these variants on even molecular phenotypes such as alternative splicing remain poorly predictable.

Experimentally testing the consequence of every possible genetic variant on endogenous alternative splicing is impractical, motivating the development of predictive models of the "splicing code". The core splicing signals—5' splice donor, 3' splice acceptor, branch point, and polypyrimidine tract—form the basis of the splicing code; they are required for recognition of intron-exon boundaries and for correct intron removal by the splicing machinery. Computational methods have been developed to score the likelihood of splicing at different splice donor and acceptor sequences (Yeo and Burge, 2004). Splice regulatory elements (SREs)—sequence motifs in exons or introns shown to regulate splicing—form the next level of regulatory information. SREs typically regulate alternative splicing by binding *trans*-acting splice factor proteins (Ule et al., 2006; Wang et al., 2013). Depending on their position and mode of action, SREs are classified as exonic splice enhancers (ESEs), exonic splice silencers (ESSs), intronic splice enhancers (ISEs),

or intronic splice silencers (ISSs). Examples of SREs have been identified computationally by analyzing motif enrichment near splice sites (Castle et al., 2008; Fairbrother et al., 2002; Zhang and Chasin, 2004) or sequence conservation between species (Goren et al., 2006). Recently a deep neural network was trained on exon skipping events in the genome to generate a comprehensive model of the splicing code that can be used to predict exon inclusion percentages (Xiong et al., 2014). Despite this progress, current models of alternative splicing do not perform well enough to be used in clinical genetics (e.g. to reclassify 'variants of uncertain significance'), and many machine learning strategies result in 'black boxes' that limit mechanistic insights.

We hypothesized that a model of alternative splicing learned from very large libraries of synthetic sequences could outperform models trained only on the genome. Current technology makes it possible to create and test gene libraries with millions of synthetic sequences—orders of magnitude larger than the number of alternative splice events in the human genome. In other applications of machine learning, such as computer vision, predictive power has increased greatly with access to larger datasets (Le, 2013).

Previous work supports the idea that synthetic gene libraries with extensive and targeted variation can provide mechanistic insights into biological phenomena. *In vivo* (Culler et al., 2010; Wang et al., 2012) and *in vitro* (Yu et al., 2008) randomized selections have identified potential SREs. Massively parallel reporter assays (MPRAs) that combine next generation sequencing with extensive variation have been applied to study transcription (Melnikov et al., 2012; Patwardhan et al., 2012; Patwardhan et al., 2009; Sharon et al., 2012; Smith et al., 2013; White et al., 2013), translation (William et al., 2014), mRNA stability (Oikonomou et al., 2014) and even alternative splicing (Ke et al., 2011). However, MPRA studies to date have overwhelmingly focused on measuring the consequences of variants in endogenous sequences (e.g. saturation mutagenesis) or

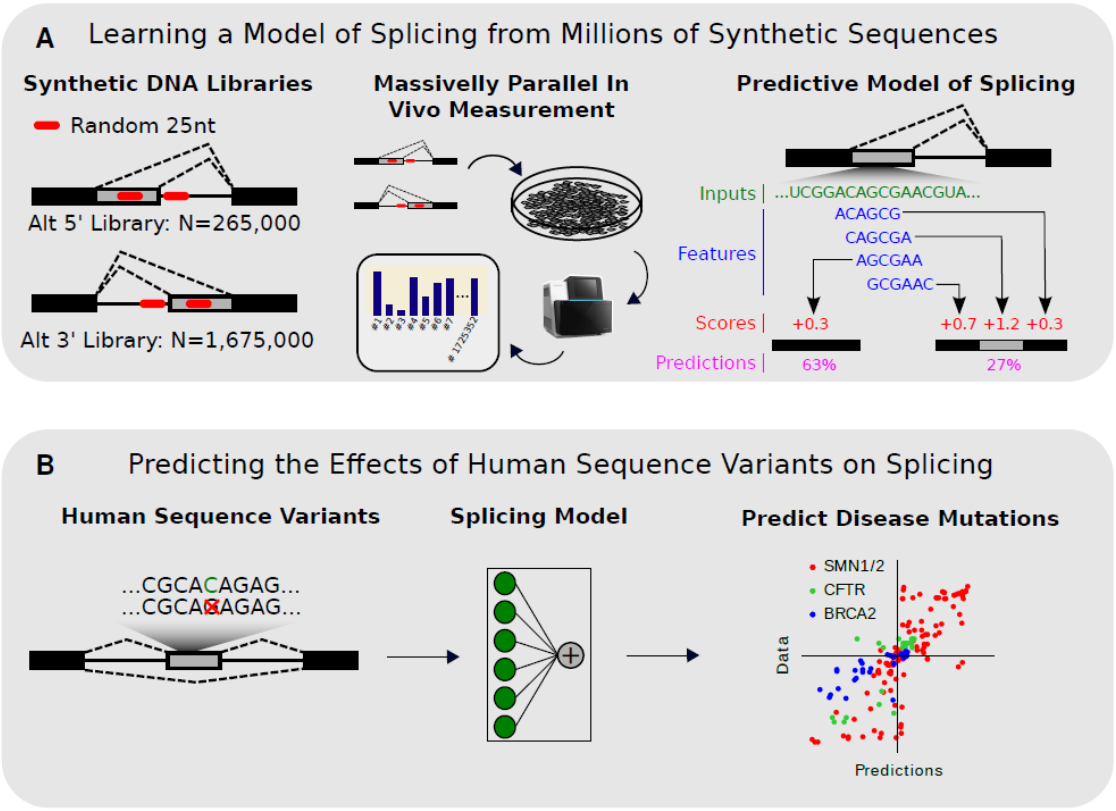


Figure 1. A Predictive Model of Alternative Splicing Learned from Millions of Synthetic Sequences.
(A) Two libraries with either alternative 5' or 3' splice sites were constructed with two 25nt randomized regions. The library was transfected into human cells and massively parallel measurement of isoform ratios was performed with RNA-seq. These two data sets were used to learn a predictive model of alternative splicing. The model takes a sequence as input, which is then converted to 6-mers features. A score for each 6-mer is learned and then used to predict the fractional usage of each splice site.
(B) When human sequence variants are fed to the model as inputs, the model makes more accurate predictions than the current state of the art algorithms.

on validating predicted activities (e.g. enhancers predicted by the ENCODE project). There are thus far few if any examples of *predictive* biological models learned entirely on MPRA data.

To test whether it is possible to learn predictive biological models from synthetic data alone, we developed an MPRA that measures alternative splice site selection in a highly complex library of 'degenerate introns' (**Figure 1A**). We added degenerate regions into an otherwise fixed sequence context, ensuring that any differences in gene expression can be causally attributed to the degenerate region. We created two libraries, one with alternative 5' splice donors consisting of 265,137 members and one with alternative 3' splice acceptors containing 2,211,739 members. We

transfected these libraries to human cells, performed RT-PCR and RNA-seq to quantitatively measure isoform ratio for all minigenes and used the results to learn a predictive model of alternative splicing. To assess the quality of the resulting model, we predicted the effects of human sequence variants on isoform levels and compared our results to available experimental data (**Figure 1B**). We tested variants in alternative 5' splicing events, both within the alternative splice donors themselves and within the alternative exon. Although our MPRA did not include a skipped exon library, our model also predicted the effect of sequence variants in skipped exons with high accuracy.

Results:

Molecular Phenotyping of Millions of Alternatively Spliced Mini-genes Containing Random Sequences

We chose to study both alternative 5' and alternative 3' splice site selection. In the case of alternative 5' splicing, we first generated a complex library by introducing 2 x 25 nucleotide (nt) fully degenerate regions into a single-intron plasmid mini-gene (**Figure 2A**). Specifically, the intron was designed with two competing splice donors separated by 44 nt; one degenerate region was inserted between the splice donors and the other downstream of the second donor. Neither degenerate sequence overlapped a splice donor. The mini-genes contained an additional degenerate 20 nt barcode in the 3'UTR. This barcode was used to create a look-up table linking barcodes and intronic sequences. Thus, even when both degenerate regions were spliced out, their sequences could be recovered from the barcode sequence (**Figure 2A**). To maximize intron sequence variability, we constructed and sequenced a complex library of 265,137 such mini-genes.

Thus, over 13 megabases of unique intronic sequence is represented within the degenerate regions of this library (265,137 x 50 nt).

In the case of alternative 3' splicing, we inserted 2 x 25 nt fully degenerate regions into a single intron system designed to have two alternative 3' splice sites (**Figure 2C**). The degenerate regions did not overlap either splice acceptor, but the upstream degenerate region did overlap the typical position of the first splice acceptor's branchpoint (-44:-19 relative to SA_1). Similarly to the alternative 5' library, we included an additional degenerate 20nt barcode in the 3'UTR. The alternative 3' library contained 2.2 million unique mini-genes encompassing over 110 megabases of unique sequence variation (2,211,739 x 50 nt).

We transfected the pooled libraries of plasmids into HEK293 cells and then quantified isoform ratios with targeted RNA-seq. To identify both the isoform and originating plasmid of each mRNA, we used paired-end sequencing with one read across the exon junction and the other read across the 3' UTR barcode (**Figure 2A and 2C**). We used 13 million reads for the alternative 5' library and 5.4 million reads for the alternative 3' library. We were then able to calculate the isoform ratios for each mini-gene in each library. We averaged 50.0 reads per minigene in the 5' library with reads mapping to 265,044/265,137 (99.96%) of all minigenes. On the other hand, in the 3' library we averaged only 2.47 reads per minigine with reads mapping to 1,686,096/2,211,739 (76.23%) of all minigenes.

Degenerate Sequences in Both Libraries Strongly Influence Isoform Ratios

In the alternative 5' library, isoforms were present from several different splicing events. The most upstream splice donor (SD_1) was used on average 22.4% of the time, while SD_2 was used 50.0% of the time (**Figure 2B**). The remaining transcripts were spliced at new splice donors

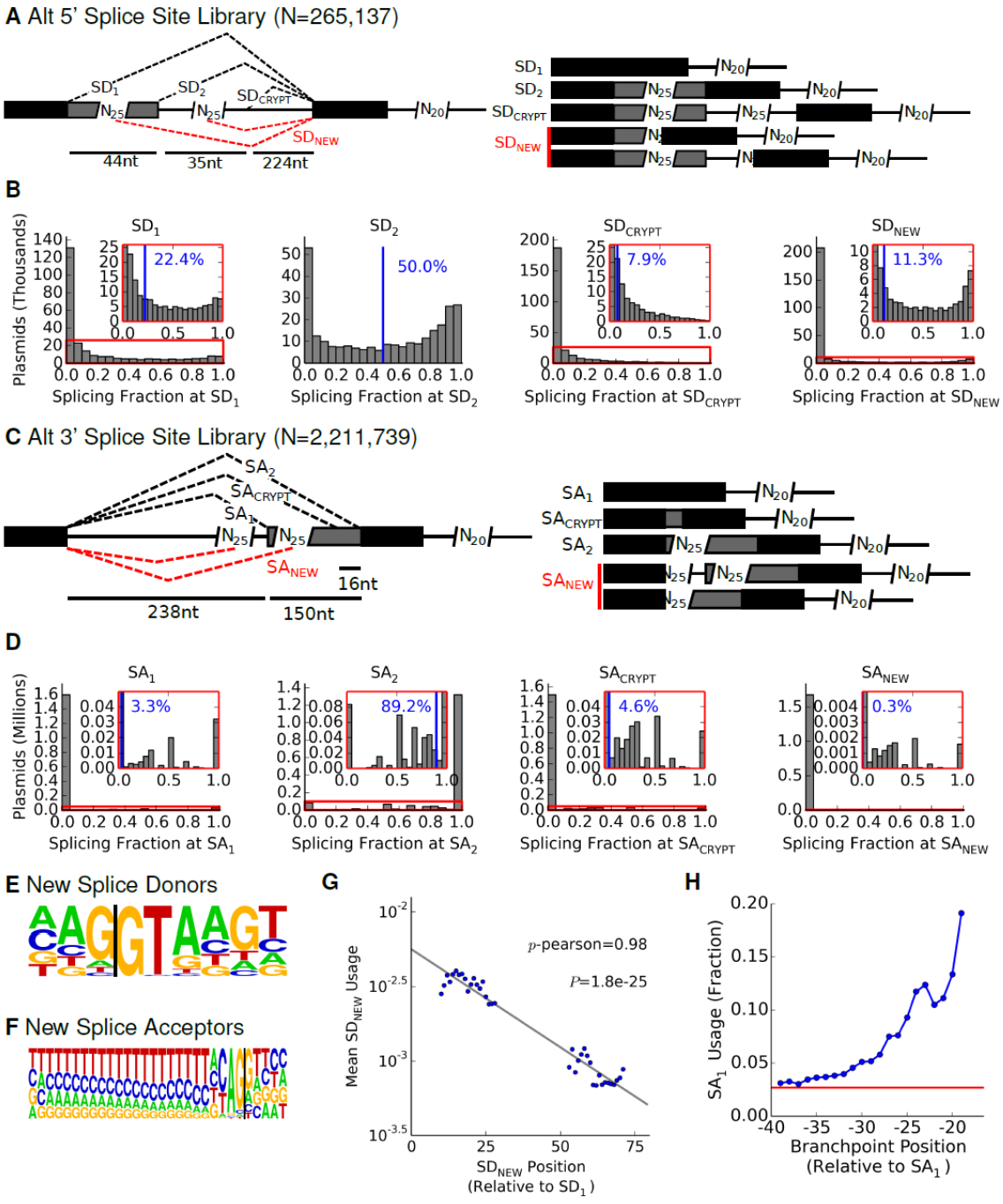


Figure 2. Splice site selection in 2 million alternative 5' and 3' spliced sequences.
(A) A schematic of the alternative 5' library. Spliced reads map to SD_1, SD_2, and a cryptic splice site (SD_{CRYPT}), as well as new splice donors(SD_{NEW}) created in the degenerate regions.
(B) Distributions of splice site usage across library mini-genes. Distributions are shown for SD_1, SD_2, SD_{CRYPT}, and SD_{NEW}. Insets correspond to the framed regions in the main graph. Mean splice site usage is indicated with a blue vertical line.
(C) A schematic of the alternative 3' library. Spliced reads map to SA_1, SA_2, and a cryptic splice site (SA_{CRYPT}), as well as new splice donors(SD_{NEW}) created in the degenerate regions.

(D) Distributions of splice site usage across library mini-genes. Distributions are shown for SD_1, SD_2, SD_{CRYPT}, and SD_{NEW}. Insets correspond to the framed regions in the main graph. Mean splice site usage is indicated with a blue vertical line.
(E) The splice donor motif recovered from new splice alternative 5' library matches the previously known human splice donor site.
(F) The splice acceptor motif recovered from new splice alternative 5' library matches the previously known human splice acceptor site.
(G) The number of spliced reads at each position within the randomized regions shows a strong position dependency. Splicing is more likely to occur at an upstream (5') splice donor than a downstream (3') splice donor. The gray line is a fit that shows the linear relationship between the location of splice donor and the log read count at that location.
(H) Minigenes with a consensus branchpoint near SA_1 are much more likely to use SA_1 than minigenes with a distal branchpoint. The red line indicates the SA_1 usage, when there is no consensus branchpoint.
See also Figure S1.

inserted into the randomized regions (11.3%), a cryptic splice donor site (SD_{CRYPT}) 35nt downstream of SD_2 (7.9%), or not spliced at all (8.4%). However, as evidenced by the broad distributions of usage at each SD (**Figure 2B**), the degenerate regions had a strong influence on splice site selection. For instance, although 49.7% of minigenes spliced at SD_1 with less than 5% frequency, 7705 minigenes (2.9%) spliced at SD_1 with over 95% frequency.

In the alternative 3' library, we also found isoforms from different splicing events, although splice site usage was less evenly balanced than in the 5' library. SA_1 was used an average of 3.3% of the time, while SA_2 was used 89.2% of the time (**Figure 2D**). In this library, new splice sites in the randomized regions were only used with 0.3% frequency, probably reflecting the larger informative footprint of splice acceptors (>20nt) compared to splice donors (9nt), which makes the occurrence of new sites within the degenerate regions less likely. Similarly to the 5' library, we inadvertently inserted a cryptic splice acceptor 16nt upstream of SA_2 that was used with 4.6% frequency. Many other cryptic splice sites were used with very low frequency ($1x10^{-7}$ to $5x10^{-3}$) accounting for a total of 2.3% of transcripts. In contrast with the alternative 5' library, only 0.3% of transcripts were unspliced. Although SA_2 was the dominant splice site, 0.7% of ~1.2M of minigenes represented by multiple reads spliced 100% at SA_1.

With so many transcripts in each library splicing at new splice sites, we asked whether we could rediscover the known motifs for splice donors and splice acceptors from the *de novo* sites alone. When we plotted the relative frequencies of each base at each position for new splice donors (**Figure 2E**) and new splice acceptors (**Figure 2F**), both splice site motifs were nearly identical to the expected motifs for splice donors and splice acceptors. More specifically, the splice donors contained the canonical GT at the +1:+2 positions, while the splice acceptors contain a clear polypyrimidine tract (T and C rich), followed by N[CT]AGG. The ability to fully rediscover canonical signals for splice donors and splice acceptors demonstrates the rich type of information contained in each dataset.

We also asked whether translation might affect the mRNA stability in our libraries. Sequencing of the alternative 5' library yielded fewer median reads on mRNA from minigenes that were primarily spliced out of frame than in frame (**Figure S1A**). However, when the minigenes contained a premature stop codon, the median number of reads per mRNA was similar for all three reading frames (**Figure S1B**). These results indicate that a large string of amino acids translated out of frame will destabilize the mRNA, likely through the no-go decay pathway (Doma and Parker, 2006; Shoemaker et al., 2010) as ribosomes stall due to protein misfolding. We also find evidence of nonsense mediated decay, but only if the premature stop codon occurred >40 nt upstream of the splice donor. This is consistent with previous studies on nonsense mediated decay that suggest the premature stop codon must occur >50 nt upstream of the last exon junction (Lewis et al., 2003).

Splicing Is More Likely to Occur at Upstream Splice Donors

From an analysis of the new splice sites, we found strong evidence that upstream splice donors were favored over downstream splice donors; new splice donors inserted in the first degenerate

region were 4.1 times more likely to be used than new splice donors inserted into the second degenerate region (Region 1: 849,666 spliced reads, Region 2: 208,396 spliced reads). Furthermore, the effect of position of splice donors within each degenerate regions was significant (P<0.005, **Figure S1C**). The number of spliced reads at a new splice site decayed exponentially with the distance from SD_1 (**Figure 2G**). Splicing has been shown to be co-transcriptional, and spliceosome components can begin to assemble at a 5' splice donor before downstream alternative slice sites are transcribed (Listerman et al., 2006), suggesting a potential mechanistic explanation for the observed effect. This strong bias for upstream splice donors is consistent with the typically short length of exons in the human genome (Burge and Karlin, 1997).

Splicing is Less Likely to Occur at Splice Acceptors with Distal Branchpoints

Large scale mapping of human branchpoints with RNA-seq found that 90% of mapped branchpoints occur between 19-37nt upstream of the splice acceptor (Mercer et al., 2015). However, it remains unclear just how detrimental a distal branch point is towards efficient splicing. Consensus branchpoints (CU[AG]A[CU]) occur over 10,000 times at every position between 40 to 19 nucleotide upstream of SA_1 in our dataset, allowing us to answer this question. We found that mini-genes with a consensus branchpoint sequence 19 nt upstream of SA_1 were approximately 6 times more likely to be spliced at SA_1 relative to those with a branchpoint 40 nt upstream of SA_1 (**Figure 2H**). One explanation for this phenomenon could be that distal branchpoints are more likely to contain another AG between the branchpoint and SA_1 that could be used as an alternative splice acceptor. However, we observed a strong distance dependence on branchpoint position for sequences both with and without an AG between the branchpoint and SA_1 (**Figure S1D**). This result suggests that mechanism by which distal branchpoints reduce splicing efficiency is primarily

due to the increased distance between the branchpoint and the splice acceptor and/or polypyrimidine tract.

Sequence Motifs in Alternative Exons Have a Stronger Regulatory Role than Intronic Sequences

We next asked how short sequence motifs affect splice site selection in different contexts. We chose to analyze the effects of 6-mers because each possible 6-mer occurs within an average of 1,294 minigenes for the alternative 5' library, and 8,232 minigenes for the alternative 3' library. Furthermore, most known RNA binding proteins (RBPs) are reported to bind sequences between 4-8 nts (Lunde et al., 2007). In order to estimate the effect of each possible 6-mer in each region, we calculated splice site usage for the subset of minigenes containing the 6-mer and for the much larger subset not containing the motif. We then asked to what extent the odds of splicing at a splice site changed in the presence of the motif relative to the control set. To quantify this "effect size" we use the $\log_2$ odds ratio with and without the 6-mer present (Supplementary Experimental Procedures). For example, we found that minigenes containing the 6-mer GTGGGG in the first degenerate region of the 5' library were spliced at SD_2 only 19.0% of the time while RNA derived from minigenes not containing this motif spliced at SD_1 50.2% of the time resulting in an effect size of -2.1 (**Figure S2A-D**). In other words, the odds of splicing at SD_1 are 4.29 ($2^{2.1}$) times lower in the presence of GTGGGG compared to its absence.

In **Figure 3A**, we plot the empirically measured effect sizes of all hexamers in the first degenerate region on the relative usage of SD_1 and SD_2, with 95% confidence intervals. The strongest enhancers located in the alternative exon (included when splicing occurs at SD_2, but excluded when splicing occurs at SD_1) increased the odds of splicing at SD_2 4.38-fold, while the

strongest silencers decreased the odds 16-fold. Approximately 15% of 6-mers have been previously identified as SREs (Culler et al., 2010; Fairbrother et al., 2002; Wang et al., 2012; Wang et al., 2004) (622/4,096), but here 82.9% of 6-mers (3,396/4,096) exhibited a significant effect on isoform selection (95% confidence interval does not contain zero effect size). Intriguingly, the cumulative effects of previously identified SREs accounted for only 20% of the cumulative effects of all possible 6-mers. The strongest silencers were G-rich, consistent with known binding sites for hnRNPs (Martinez-Contreras et al., 2006). On the other hand, some of the strongest enhancers for SD_2 appear to act by generating secondary structure around SD_1: The 6-mers perfectly complementary to part of SD_1 (-3 to +8) were all in the top 6% of SD_2 enhancers (Percentiles: 97.77, 99.75, 99.97, 94.23, 94.79, 98.92).

We then looked at the effect of 6-mers in the second degenerate region (3' to SD_2). Unlike the first degenerate region, which is located within the alternative exon region, the second degenerate region is intronic to both SD_1 and SD_2. We found that the effect sizes were much smaller than in the first degenerate region (**Figure 3B**). The strongest enhancer and silencer of SD_2 respectively only changed the odds of splicing at SD_2 relative to SD_1 1.95 fold and 1.48 fold. Furthermore, only 36.7% of 6-mers (1505/4096) had a statistically significant effect.

We performed a similar analysis for each degenerate region on the usage of SA_1 in the alternative 3' library (**Figure 3C and 3D**). Again we found that motifs in the alternative exon (3' of SA_1, but 5' of SA_2) had strong effect sizes (statistically significant 6-mer effect sizes: 3500/4096, 85.4%; strongest enhancer: 3.84 fold increase in odds of splicing at SD_2, strongest silencer: 9.87 fold decrease in odds of splicing at SD_2). Unlike in the alternative 5' library, we found that motifs in the intronic degenerate region (5' of SA_1 and SA_2) also have quite strong

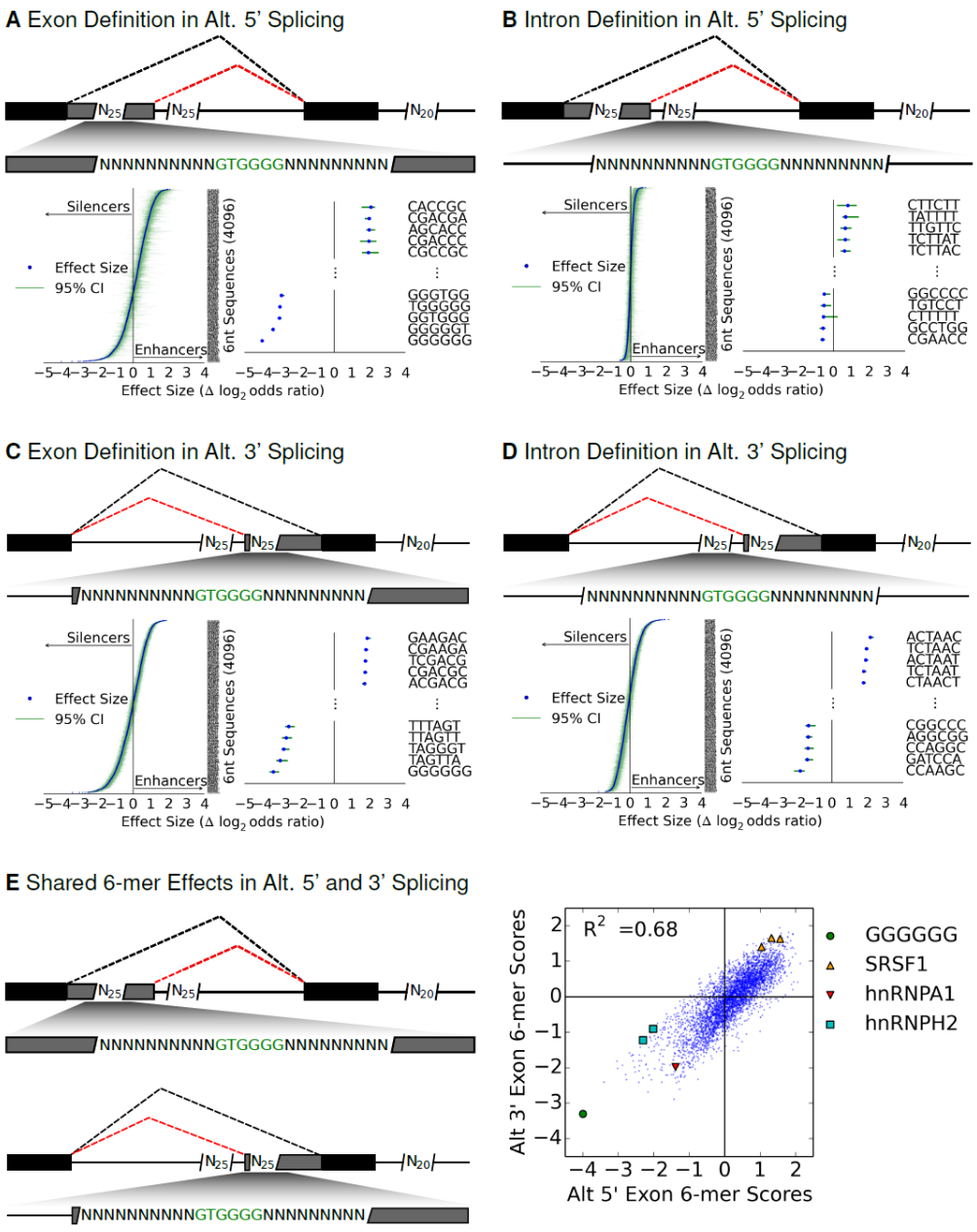


Figure 3. Measured effect sizes of individual 6-mers in each degenerate region.
(A-D) To measure how sequence motifs alter the relative use of SD2/SD1 or SA1/SA2, we calculated effect sizes for every 6-mer (n=4096) within each degenerate region in both libraries. We defined effect sizes as the log odds ratio of SD2 or SA1 usage between minigenes with/without the 6-mer of interest. The 6-mers are ranked by the estimated effect size and plotted with 95% confidence intervals generated by bootstrapping with replacement.
(A) Alternative Exon Region in 5' Library
(B) Intronic Region in 5' Library
(C) Alternative Exon Region in 3' Library
(D) Intronic Region in 3' Library
(E) The 6-mers scores in the alternative exon region in both the 5' and 3' libraries (A and C) are highly similar, suggesting alternative splicing in both libraries is regulated by the same mechanism.
See also Figure S2.

effects (statistically significant 6-mer effect sizes: 3248/4096, 79.3%; strongest enhancer: 3.45 fold increase in odds-ratio, strongest silencer: 4.63 fold decrease in odds-ratio), although still generally smaller in magnitude than the downstream alternative exon region. When we looked at the strongest 6-mers enhancers of SA_1 in this intronic region, we found they all fit the consensus branchpoint sequence CU[AG]A[CU] (**Figure 3D**).

The Same Sequence Motifs Regulate Alternative Exon Inclusion Independent of the Type of Alternative Splicing

Surprisingly, we found that the effect sizes of 6-mers that occur within the alternative exon regions were extremely similar between the alternative 5' and 3' libraries (**Figure 3E**, R^2=0.68). We looked at several motifs known to bind splice factors or that have previously been identified as ESEs/ESSs (G-run, SRSF1, hnRNPA1, hnRNPH2) and found the effect sizes to be highly correlated. In both libraries, GGGGGG was the strongest exonic silencer (5' Library: 16.0 fold change in odds-ratio, 3' library: 9.87 fold reduction in odds-ratio).

We also compared the effect sizes of intronic 6-mers (second randomized region in in Alt. 5 library; first randomized region in the Alt. 3' Library) between the two libraries. We found a significant but weaker correlation between the 6-mer scores (R^2=0.27, **Figure S2E**). The first randomized region in the alternative 3' library overlaps the expected location of the SA_1 branchpoint, which may reduce the effect size correlation. However, the weaker correlation can also be explained by the fact that the effect sizes of intronic 6-mers were much smaller in magnitude compared to 6-mers within the alternative exon regions.

Sequence Motifs Regulate Exon Inclusion Additively Rather than Cooperatively

Although previous studies have observed co-occurrence of conserved sequence motifs around splice sites (Barash et al., 2010), it remains unclear whether such motifs act cooperatively or additively and independently of one another to regulate alternative splicing. In an additive and independent model of regulation, the joint effect size of multiple motifs should simply equal the sum of the individual effect sizes (**Figure 4A**). To assess this, we examined the joint effect sizes of pairs of 4-mers on alternative exon inclusion levels in both the 5' and 3' libraries. We chose 4-mers because pairs of 4-mers occur sufficiently often within each randomized region to allow for robust effect size measurements (Alt. 5' Library: 692 minigenes/4-mer pair; Alt. 5' Library: 4,399 minigenes/4-mer pair).

We first calculated the individual effect size of all 4-mers on exon inclusion in the 5' library. We then calculated the joint effect size of every possible pair of non-overlapping 4-mers. Surprisingly, we found that combinatorial effects were nearly perfectly captured by the sum of the 4-mers' individual effect sizes (R^2=0.913, **Figure 4B**). We did the same analysis for 4-mers located in the second degenerate region of the 3' library. Here we found the linear model fit the experimental data even better (R^2=0.954, **Figure 4C**). Thus, while specific instances of cooperative sequence interactions have been well documented (Huelga et al., 2012; Oberstrass et al., 2005), our results suggests the majority of motifs primarily exert their influence on exon inclusion independently of surrounding motifs.

Predicting Isoform Ratios in Alternative Splicing from Sequence

We then turned to the task of learning a model of alternative splicing to predict isoform levels from sequence information. Because combinatorial regulation of alternative splicing was

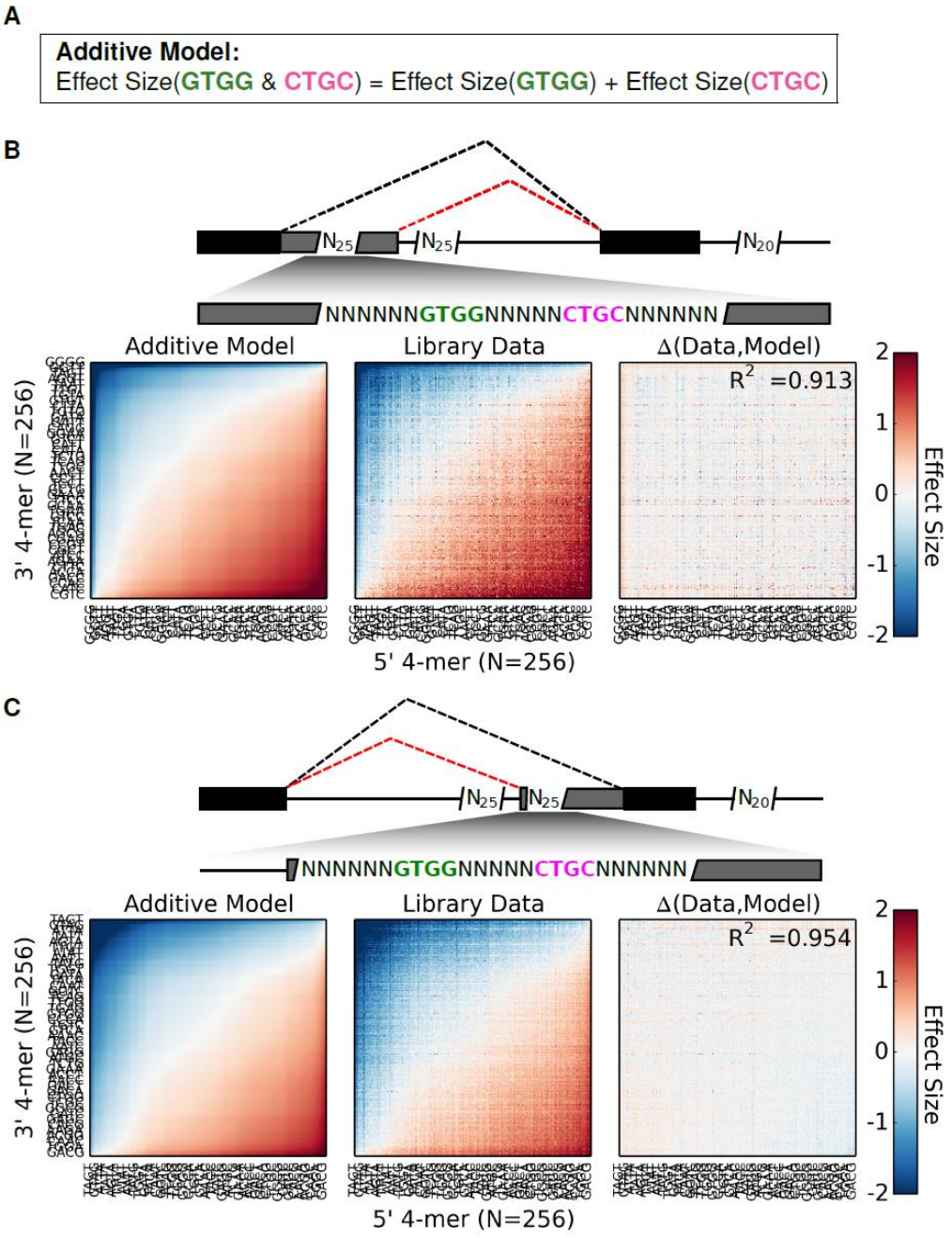


Figure 4. Combinatorial Regulation of Alternative Splicing is Additive

(A) An additive model of alternative splicing regulation: the joint effect size of two 4-mers is equal to the sum the individual 4-mer effects.

(B) Using an additive model, the predicted combinatorial effect size of every pair of 4-mers (N=65,536) is plotted on the left. Each pixel corresponds to a pair of 4-mers with the 5' 4-mer on the x axis and the 3' 4-mer on the y axis. The measured combinatorial effect sizes from the 5' library data are plotted in the middle. On the right the residuals between the additive model and the observed data are plotted. The additive model explains >90% of the combinatorial effect sizes (R^2=0.913).

(C) The same analysis is repeated for the alternative 3' library. In this library the additive model explains over 95% of the combinatorial effect sizes (R^2=0.954)

See also Figure S3.

accurately captured by an additive model, we postulated that an additive model with short sequences as input features would perform well for prediction. Using both the 5' and 3' libraries, we trained a joint model of alternative exon definition in which a score is learned for each of the 4,096 possible 6-mers (**Figure S3A**). The scores learned here are similar to the previously calculated effect sizes, but rather than measuring effects of a single 6-mer one at a time, we learned all the scores together through regression. Given the large number of new splice donors appearing within the 5' library, we also chose to train a model of the splice donor site itself (**Figure S3B**). When we tested the splice donor model using cross validation, we found it accurately predicted the fraction of reads mapping to the three original splice donors, accounting for up to 75% of observed isoform variability (R^2: SD_1=0.75, SD_2=0.75, SD_{CRYPT}=0.54, **Figure 5A**). It also proved accurate in predicting the position and fraction of reads mapping to newly created splice donor sites within the degenerate regions (R^2: 0.83**, Figure 5A and 5B**).

A fundamental advantage of testing synthetic sequences is the ability to learn from larger datasets than were previously available. As an attempt to quantify this advantage, we calculated learning curves on a simple model predicting usage of SD_1 in the alternative 5' library. We split our data into training and test sets (90%/10% split) and trained models using subsets of the training data (between 100 to 177,827 training points). We also trained separate models using 3-mers, 4-mers, 5-mers, 6-mers, or 7-mers. With limited data (1000 or less training points), the simplest model (3-mer) made the most accurate predictions, while the 7-mer model made the least accurate predictions, with the other models ordering between (**Figure 5C**). However, with the largest training subset (177,827 points), the results were reversed with the 7-mer model achieving the highest accuracy. Based on the slopes of the learning curves, the 3-mer to 5-mer models would not benefit significantly from more data points (>177,827), but the 6-mer and especially 7-mer models

seem likely to achieve significantly higher prediction accuracy with larger training sets. These results highlight the intuitive point that richer feature sets can improve predictions accuracy, but require more data to properly train.

Predicting the Effects of Human Genomic SNPs on Alternative Isoform Ratios

We next asked whether we could apply our model (HAL: Hexamer Additive Linear) – developed entirely in the context of synthetic mini-genes – to predict changes in alternative splice donor usage caused by common polymorphisms in human genomes. As a first test case, we focused on 5' alternative splicing. Combining DNA and RNA sequencing data respectively from the 1000 Genomes Project (Genomes Project et al., 2012) and GEUVADIS consortium (Lappalainen et al., 2013), we calculated the percent of splicing at the downstream alternative splice donor (Percent Spliced In, PSI) of wild-type genotypes for 8,546 5' alternative splicing events using the MISO software package (Katz et al., 2010). We separately calculated mean isoform levels for genotypes heterozygous or homozygous for a single SNP in the region between the two competing splice donors or within the splice donors themselves (**Table S1**).

We began by investigating whether the model of the actual 9 nt splice donor sequence—again learned completely from our synthetic mini-genes—could accurately predict the effects of SNPs occurring within splice donor sequences. We also compared our prediction accuracy to a leading splice donor prediction tool trained directly from splice donor usage in the human genome (MaxEnt) (Yeo and Burge, 2004). Among heterozygous SNPs in alternative splice donors occurring in multiple individuals, we found that 93 of 199 SNPs altered PSI by >5% (**Figure 6A and 6B**). Within this set, HAL predicted direction of change with 87.1% accuracy (81/93; binomial P=9.83x10^{-14}), while MaxEnt predicted direction of change with 81.7% accuracy (76/93; binomial

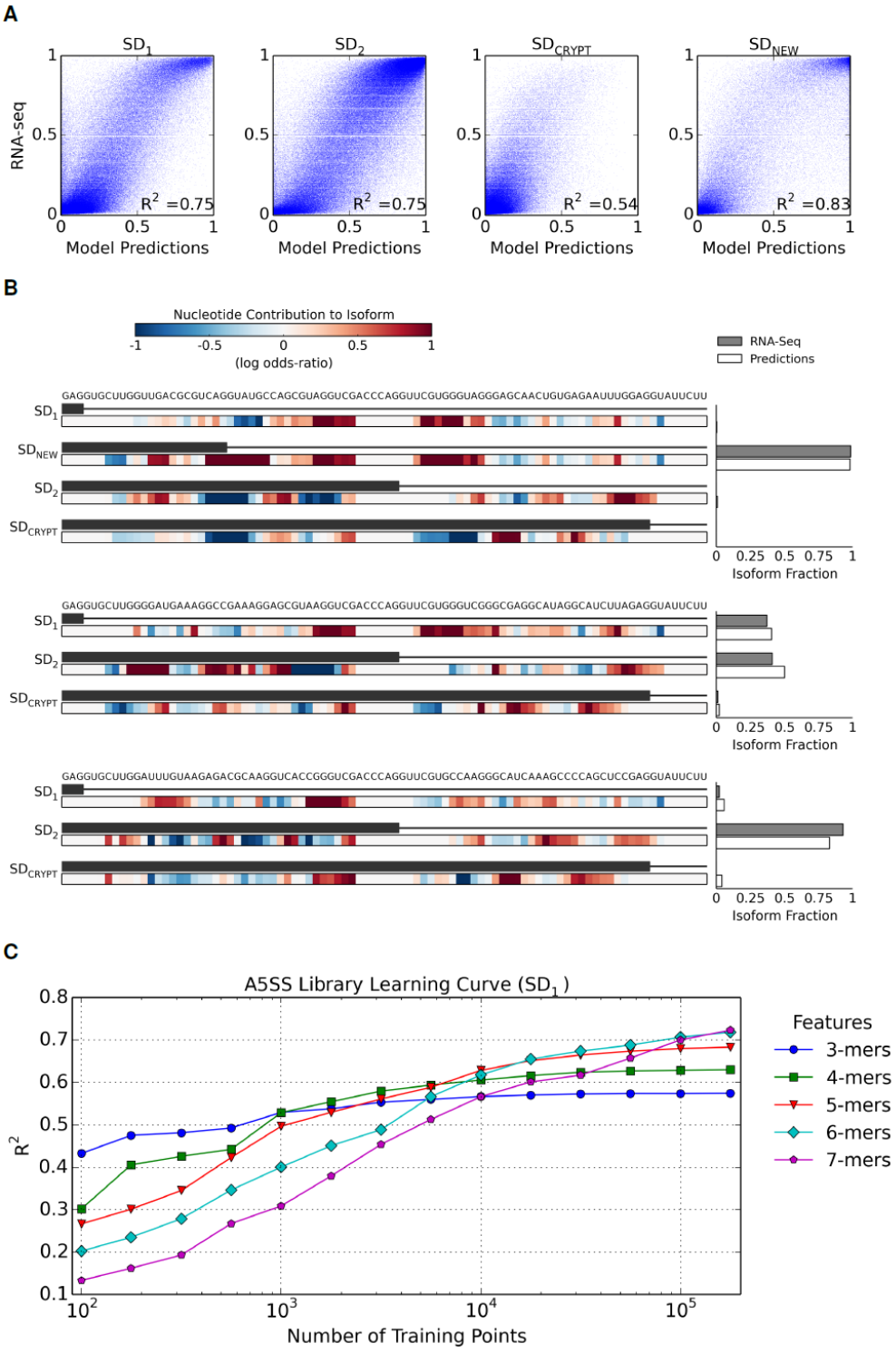


Figure 5. A model accurately predicts alternative 5' splicing and the location of new splice donors
(A) For each splice donor (SD_1, SD_2, SD_{CRYPT}), model predictions are plotted against the observed splice site usage fraction. Each point represents a single test plasmid. The results are also plotted for all new splice sites (SD_{NEW}).
(B) The prediction results for three different minigenes are shown with the associated nucleotide scores for each isoform. Each nucleotide score is calculated by averaging the model weights of all 6-mers overlapping the

nucleotide. In the first example minigene, HAL predicts the usage and position of a new splice donor, which is confirmed by RNA-seq.
(C) A learning curve was generated for different models that predict the fraction of splicing at SD_1. The simplest model (3-mer features) performed the best with small training sets (<1000 data points), but with more data points, richer feature sets offer better performance.
See also Figure S4, Table S1

P=4.45x10^{-10}). Among the 35 homozygous SNPs in splice donors that alter PSI by >5% our model predicted every SNP correctly, while MaxEnt made two mistakes (HAL: 35/35, binomial P=5.82x10^{-14}; MaxEnt: 33/35, binomial P=3.67x10^{-10}). For the set of SNPs within splice donors, our model explained 59.3% of the observed heterozygous effects (R^2=0.593, P=6.38x10^{-8}) and 67.7% of the observed homozygous effects (R^2=0.677, P=4.65x10^{-24}). This is a substantial improvement over MaxEnt, which accounted for 39.8% of the observed heterozygous effects (R^2=0.398, P=1.22x10^{-11}) and 41.1% of the observed homozygous effects (R^2=0.411, P=3.3x10^{-5}). Even when we extended our analysi to all SNPs (including those with less than 5% change in PSI), we found HAL substantially outperformed MaxEnt (HAL: R2=0.48; MaxEnt: R2=0.22, **Figure S5A**).

We then applied the model to predict the effects of human genomic SNPs in the alternative exon region between but not overlapping splice donors. Because most SNPs not occurring in actual splice sites are likely to only have modest effects, we restricted our analysis to SNPs with at least 10 homozygous wild-type or 10 heterozygous samples expressing the relevant mRNA. Moreover, we focused on SNPs that resulted in a change in the PSI of at least 5% to minimize the impact of measurement noise on the validation data set; 43/344 heterozygous and 20/131 homozygous SNPs altered the PSI by >5% (**Figure 6C**). HAL correctly predicted the direction of change for 37/43 heterozygous and 17/20 homozygous SNPs (P: heterozygous=1.63x10^{-6}, homozygous=2.58x10^{-3}, combined=6.11x10^{-9}). Furthermore, our model explained around half of the total observed effects of these SNPs (heterozygous: R^2=0.570, P=9.23x10^{-9}; homozygous: R^2=0.442, P=1.39x10^{-3}). Thus, our model not only outperformed the state of the art splice donor

algorithm (MaxEnt) at predicting the effects of SNPs within splice donors but also successfully predicted the effects of SNPs within the alternative exon region, which to our knowledge, no other tool can do.

Predicting Alternative 5' Isoform Levels from Sequence

To further assess the accuracy of our splice donor model, we predicted the isoform ratios in 6,152 alternative 5' splicing events expressed in lymphoblastoid cell lines and compared our results to four other splice donor prediction algorithms. Our splice donor model significantly outperformed all of the other algorithms (**Figure S4, Table S2**). Interestingly, all of the models (including ours), performed better on events with shorter alternative exon regions (i.e. the region between splice donors). In these events, there is less space for regulation between the splice donors, possibly simplifying the prediction task.

Predicting the Effects of Variants on Exon Skipping in Mendelian Diseases

The most common form of alternative splicing is neither alternative 5' or 3' splicing, but exon skipping. Exon skipping is a highly regulated form of alternative splicing in human cells and misregulation of cassette exon splicing can cause disease (Garcia-Blanco et al., 2004) and cancer (Kim et al., 2008). Given the relatively more complex structure of skipped exons, it might on first sight seem unlikely that a model trained only on 5'and 3' alternative splicing should be able to predict levels of exon inclusion. However, we hypothesized that the similarity between the sequence determinants of alternative exons in alternative 5' and 3' splicing might extend to exon skipping as well. If this were the case we would expect our model to accurately predict the effects of exonic sequence variants on skipped exon inclusion levels, even though it was never trained directly on any exon skipping data. We tested this hypothesis in the context of mutations in several

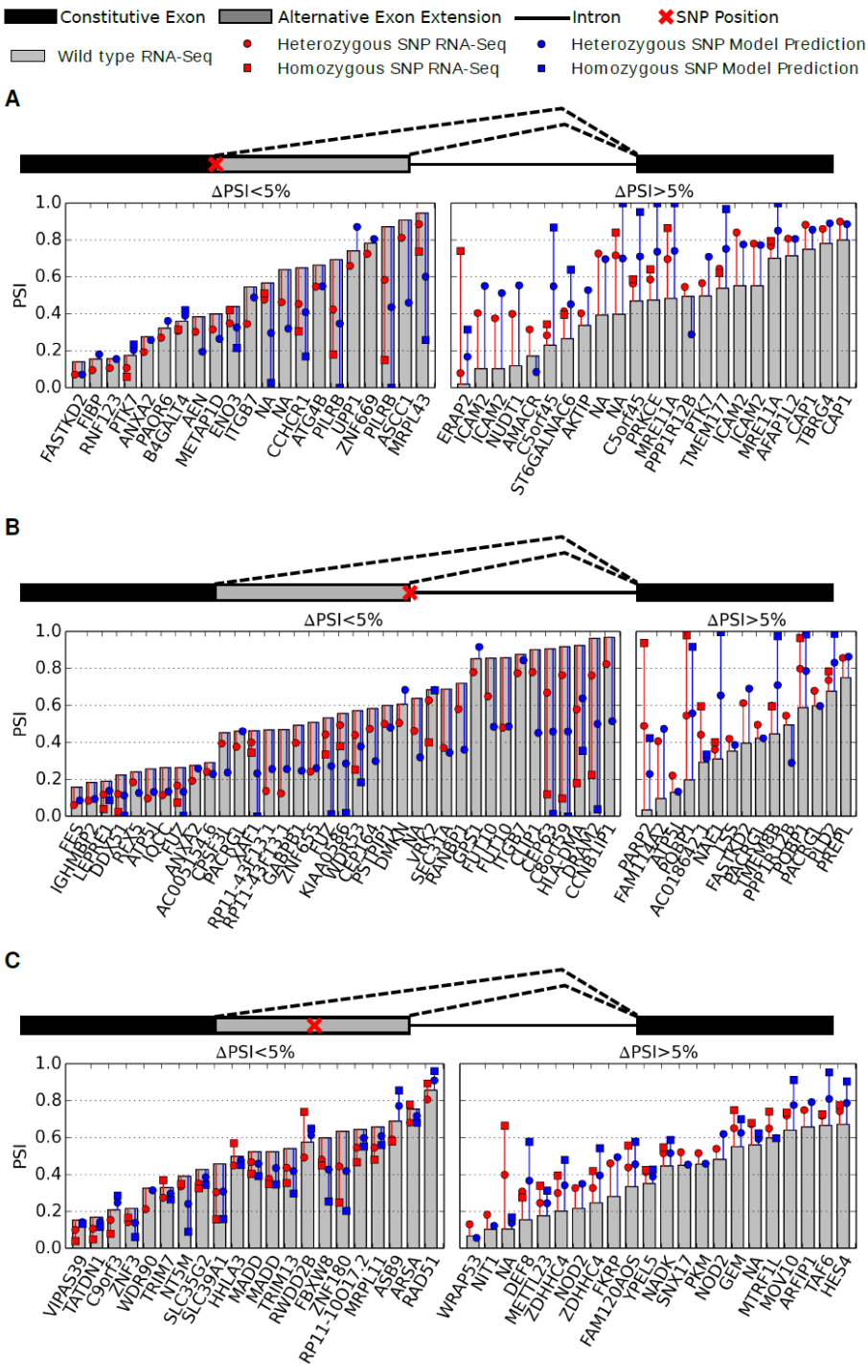


Figure 6. Splicing Model Identifies the Functional Effect of SNPs on Alternative Splicing.
(A-C) Model predictions are plotted with the PSI measured from RNA-seq for SNPs occurring in (A) the upstream splice donor(B) the downstream splice donor, (C) between the competing splice donors that alter the measured PSI by greater than 5 percent. The observed PSI from RNA-seq for the wild type genotype (gray bar) and genotypes containing the SNP (red) are plotted together with the model prediction (blue). The model accurately predicts the direction of change of the heterozygous SNPs in splice donors with 87.1% accuracy (81/93, binomial P=9.83x10^{-14}) and the heterozygous SNPs between splice donors with 86.% accuracy (36/43, binomial P= 8.18x10^{-7}). See also Figure S5, Table S2

distinct genes that are known to cause Mendelian disease by promoting exon skipping (**Figure 7A, Table S3**).

First, we compared model predictions to experimental data for the *SMN1* and *SMN2* genes, whose misregulation can lead to spinal muscular atrophy. Our model correctly predicted increased or decreased exon 7 inclusion in 205/229 (89.5%, **Figure 7D**) variants with experimental data. In **Figure 7B**, we compare predictions (increased or decreased exon inclusion) to experimental data. To make the plot more readable, we only included a single SNP at each position. Our model accurately predicts increased/decreased exon inclusion for 17/19 SNPs. On just the variants with quantitative data (N=131) our model explained 65% of the observed variance (R^2=0.65, **Figure 7E**). The *SMN1/2* variants that we tested included SNPs, indels, as well as combinations of up to 30 nucleotide changes.

We then tested our model on variants in *CFTR*, whose misregulation can lead to cystic fibrosis. Our model correctly predicted increased/decreased exon 12 inclusion in 19/22 variants (**Figure 7D**). When we only looked at the SNP with the largest effect at each position, our model accurately predicted increased/decreased exon inclusion for 11/12 SNPs. Among all the *CFTR* variants, our model explained 60% of the observed variance (**Figure 7E**, R^2=0.60).

We next tested our model predictions on variants in exon 7 of the *BRCA2* gene, a tumor suppressor responsible for DNA damage repair. Mutations in *BRCA2* affecting the ability of the protein to repair DNA lead to such an increased risk of ovarian and breast cancer that patients with these mutations may choose to have prophylactic surgery. However, the effect of many variants on alternative splicing and hence protein function remain unknown, forcing patients and doctors to make clinical decisions with limited information. The ability to identify deleterious variants computationally can provide valuable information to patients with these variants of unknown

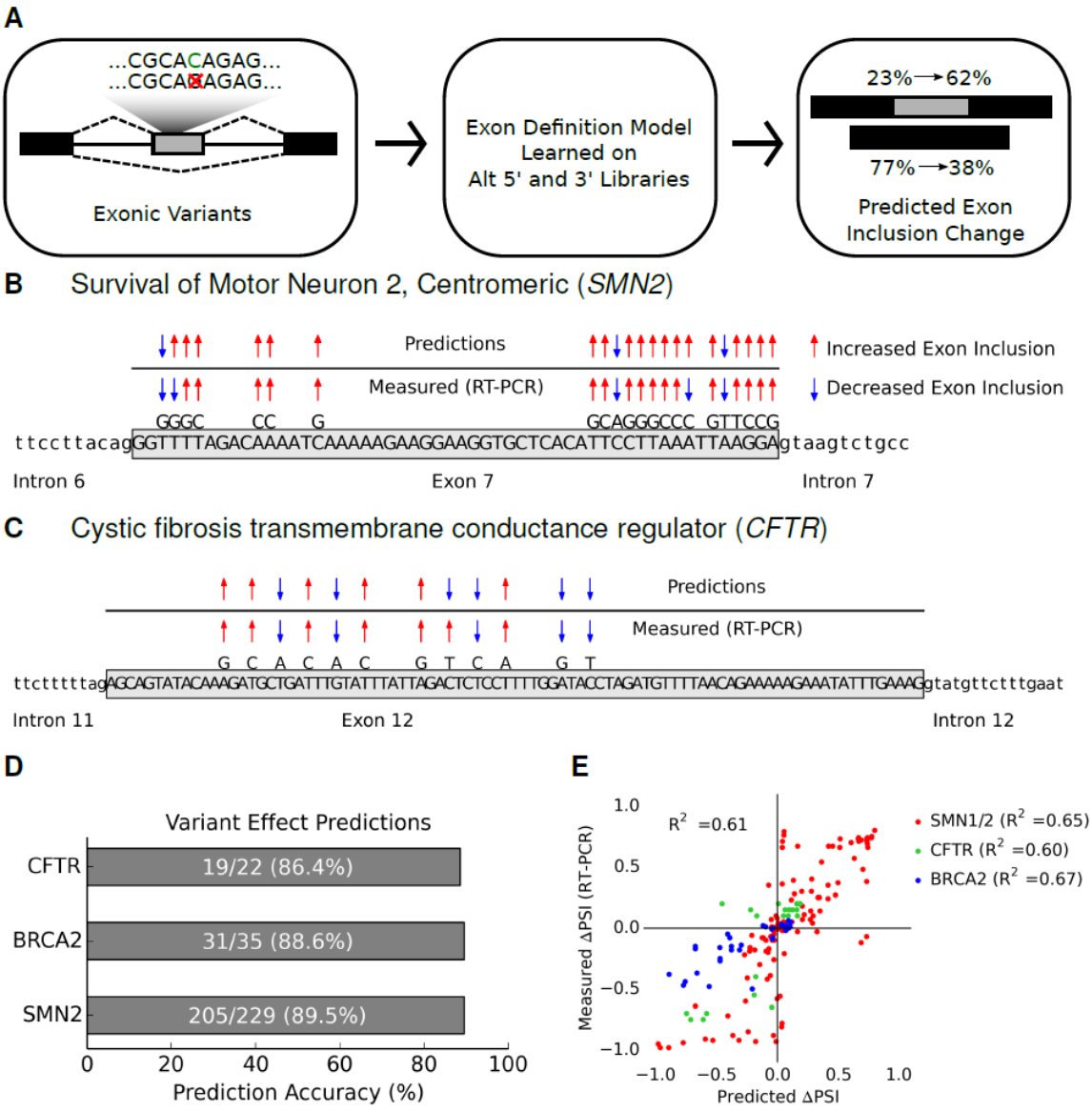


Figure 7. Predicting the Effects of Exonic Variants on Exon Skipping

(A) The inputs to the splicing model can include SNPs, indels, or complex variants within the alternative exon. The splicing model then predicts the exon inclusion levels with the variant present.

(B) Model predictions are compared to experimental results using RT-PCR for SNPs occurring in exon 7 of SMN2. For positions with data for multiple SNPs, the SNP with the largest measured change in PSI was plotted. The model accurately predicted the directional change in PSI (increased exon inclusion/exclusion), for 18/19 SNPs plotted.

(C) Model predictions are compared to experimental results using RT-PCR for SNPs occurring in exon 12 of CFTR. The model accurately predicted the directional change in PSI for 19/22 SNPs plotted.

(D) The prediction accuracy for variants in SMN2, CFTR, and BRCA2 that altered the measured PSI are all between 86 to 90%.

(E) The change in PSI is plotted for every variant with RT-PCR data. The model explains over 60% of the effects of SNPs for variants each gene tested (SMN1/2, CFTR, and BRCA2).

See also Figure S6, Table S3

significance. Our model correctly predicted increased/decreased exon 7 inclusion for 31/35 variants that experimentally altered inclusion levels (**Figure 7D**). The model correctly predicted 19/22 of the SNPs with the largest effect at each position within the exon (**Figure S6B**). Among all the *BRCA2* variants, our model explained 61% of the observed variance (R^2=0.67, **Figure 7E**).

We then compared our results to SPANR (Xiong et al., 2014)—the current state of the art in predicting the effects of SNPs on exon skipping. SPANR consists of a Bayesian deep learning algorithm trained on exon skipping events in the human genome with 1393 carefully hand selected features. As of this paper, SPANR only supports prediction of SNPs, so we were not able to compare our predictions on more complex variants. However, for SNPS in *SMN1/2*, *CFTR*, and *BRCA2*, we found that HAL accounted for three times more of the observed effects than SPANR (HAL: R^2=0.51, SPANR: R^2=0.17, **Figure S6A**).

Discussion

We present a framework based on massively parallel analysis of synthetic sequences to dramatically improve our understanding of alternative splicing and the ability to predict the impact of natural human genetic variation. Our model accurately predicts the effects of sequence variants on alternative 5' splicing that occur both within the alternative exon as well as in the competing splice donors. Even more importantly, our model learned regulatory rules about alternative splicing that generalized to exon skipping—a completely different form of alternative splicing than those on which the model was trained.

Our results suggest that the same regulatory mechanism is shared between all major forms of alternative splicing. Additional evidence for such as common mode of regulation comes from previous smaller-scale studies of ESEs or ESSs that have shown similar effects across different

forms of alternative splicing (Wang et al., 2012; Wang et al., 2006). It is unlikely that regulation occurs during splice site recognition; any exonic splice regulatory element that alters splice donor or splice acceptor recognition should have different effects in alternative 5' and 3' splicing events. It is more likely that alternative exon inclusion is modulated during exon definition, that is the pairing of splice site across exons, which often preceeds the eventual pairing of splice donors and acceptors across introns (Robberson et al., 1990).

Furthermore, our data also show that the exon defining interactions between the upstream splice acceptor and downstream splice donor are regulated additively. In both alternative 5' and 3' splicing, we found the joint effect size of multiple 4-mers to be highly correlated with to the sum of the individual 4-mer effects. This result may indicate that each sequence motif can contribute additively to stabilizing the splice acceptor-splice donor interaction, likely through the *trans*-factors that bind these sites. However, the true mechanistic basis for this additivity will require further investigation. Although, there is evidence supporting specific examples of functional interactions between *cis*-splicing regulatory elements (Oberstrass et al., 2005), our results indicate that these examples are likely uncommon.

A potential limitation of our approach is that mRNAs are transcribed from plasmids rather than directly from the genome, especially considering evidence suggesting that chromatin can influence alternative splicing (Luco et al., 2010). However, advances in high-throughput genome editing may make it possible to perturb the genome in a massively parallel fashion, which will enable extensions of our approach to probe the effects of chromatin on alternative splicing. In fact, recent work demonstrated that small-scale genomic libraries could be created through insertion of degenerate sequences directly into an alternatively spliced gene locus (Findlay et al., 2014). Moreover, our current work focused on mini-genes with short alternative exons and more work

will be necessary to understand to what extent our results generalize to other gene architectures. However, human exons are typically short (average 147 bp for internal exons)(Lander et al., 2001) and, moreover, analysis of sequence conservation suggests that most sequence determinants of alternative splicing can be found within a few hundred nt of intron exon junctions. It is important to emphasize that our approach uncovers only *cis*-regulatory rules. Complementary experiments that connect this *cis* grammar to a repertoire of *trans*-acting splice factor proteins are necessary to fully understand the mechanisms underlying the regulation of alternative splicing.

We have demonstrated that learning the sequence determinants of gene regulation from large libraries of synthetic sequences can be used as a complementary approach to learning directly from the human genome. We assayed over 2 million alternative spliced constructs, nearly two orders of magnitude more events than the 38,000 that are present in the human genome (Wang et al., 2008), and containing nearly 100 megabases of synthetic sequence. Our improved understanding of alternative splicing and performance in predicting the effects of genetic variants is not a result of more sophisticated machine learning algorithms, but simply learning from a larger and more reliable dataset. We anticipate that this general approach will be useful for advancing our biological understanding of diverse forms of gene regulation such as transcription, translation, and polyadenylation.

Experimental Procedures

Cloning of Degenerate Libraries

The libraries were assembled with PCR and standard Gibson assembly (Gibson et al., 2009) using degenerate oligonucleotides (IDTDNA). First citrine was split into two exons and the first exon of the Citrine gene was altered to remove any potential splice donors, without altering the

amino acid sequence. The introns with degenerate sequences were inserted between the two exons of Citrine. The barcode sequence was inserted into the 3'UTR of Citrine.

Cell Culture and Transfection

HEK293 cells were cultured in in DMEM (Cellgro) + 10% FBS and L-glutamine/penicillin/streptomycin on coated plates. Plates were coated for 24 hours with 8mL of 100x diluted extracellular matrix gel (Sigma-Aldrich) before HEK293 cells were added to the plates. For transfection of a complex pool of plasmids, 1.2 million cells were seeded in a 10cm dish 24 hours before transfection. We mixed 10ug of the plasmid library in 1ml of Opti-MEM Reduced Serum Medium (Life Technologies) with 30ul of Lipofectamine LTX and 10ul of Plus Reagent (Life Technologies), before transfecting into the 10cm dish. The DMEM was was replaced 5 hours after transfection.

Isolation of RNA and Generation of cDNA

Total RNA was extracted using RNeasy (Qiagen) kits 24 hours after transfection. The optional on column DNaseI digest was performed with the RNase-Free DNase Set (Qiagen). Total RNA quality and purity was tested by measuring the A260/A280 ratio on a NanoDrop 1000 Spectrophotometer, and in some cases by measuring the ratio of the 18S and 28S rRNA bands on a native 1% agarose gel. mRNA was separated from 35-48ug total RNA using polyA Spin mRNA Isolation Kits (New England Biolabs). Isolated mRNA was again digested by DNaseI for 30 minutes using the Turbo DNA-free kit (Ambion). cDNA was then synthesized from 109-374ng mRNA using MultiScribe Reverse Transcriptase (Ambion) and Oligo d(T)16 (Ambion). cDNA synthesis was performed by holding reactions at 25°C for 10 min, 42°C for 110 min, and 85°C for 5 min. The quality of cDNA and presence of DNA contamination were checked through qPCR: Citrine, mCherry, and TBP were compared using cDNA, No Reverse Transcription Controls

(NRTC) and a No Template Control (NTC). The results indicated that there was no plasmid or genomic DNA carryover into the cDNA reactions.

Generation of Illumina Flow Cell Compatible PCR Products from RNA and DNA Library

The resulting cDNA was then amplified by PCR to generate products compatible with the Illumina HiSeq2000 Flow Cell. PCR reactions were performed in 100uL with 2x Phusion HF Master Mix (New England Biolabs), 50pmol forward primer and 50pmol reverse primer with sample specific barcodes and 20% of each cDNA reaction. Cycling was done on a BioRad T100^{TM} Thermal Cycler with the following protocol: 98°C for 5 min, then 7 cycles of 98°C for 10s, 67.5°C for 15s, 72°C for 30s, and a final extension step at 72°C for 5 min. The necessary number of cycles was determined for each sample by first running PCR reactions with EvaGreen in a Biorad CFX and determining when fluorescence began to plateau. Following PCR, 10% of the products were run on a 2% agarose gel to determine if the expected bands were present. The remainder of the PCR products was purified using the QIAquick PCR Purification kit (Qiagen) and eluted into 30uL of EB. Concentrations as well as A260/280 and A260/230 ratios were measured on a NanoDrop 1000 Spectrophotometer.

Illumina compatible PCR products were also generated from the DNA plasmid library with the same protocol as above except the cDNA template was replaced with 10ng of plasmid library DNA and the PCR reaction was performed with 20 cycles.

Sequencing Plasmid Library and RT-PCR Products

Both the RT-PCR products and plasmid library PCR products were sequenced on an Illumina HiSeq2000 with paired 101 nt reads. The forward read crossed the post-splicing exon-exon junction and the reverse read covered the 3'UTR barcode. A 6nt index read was used to sequence the sample barcode to determine if the read came from a DNA library or a cDNA library.

Associating Degenerate Intronic Regions with 3' UTR Barcode Tags

Using the sequencing results of the DNA plasmid library, we first counted the number of reads for every observed barcode and calculated an average Phred quality score for each position. We discarded any barcode tags with less than two reads or less than an average Phred score of 20 at any position. We then mapped each remaining tag to the associated degenerate sequence with the most reads. If each degenerate sequence had a single read, we chose the sequence with the highest minimum Phred score.

Measuring Isoform Fractions from Sequencing Results

For every read on an RT-PCR product, we recorded the splicing position (or lack of splicing) by aligning the read to the unspliced plasmid. Using the associated barcode read, we were then able to tally the number of reads splicing at each position for every plasmid in our library. With respect to the alternative 5' library, only reads that mapped to a splice donor with GT or GC in the +1 to +2 intronic positions were counted.

Author Contributions: ABR designed and performed experiments, analyzed data, built and tested the splicing model, wrote the manuscript, and developed the web tool; RPP designed experiments; JS and GS designed experiments and wrote the manuscript.

Acknowledgments: This work was supported by National Science Foundation (NSF) CAREER Award 0954566 and a Burroughs Wellcome Career Award at the Scientific Interface to GS.

References

Barash, Y., Calarco, J., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B., and Frey, B. (2010). Deciphering the splicing code. Nature *465*, 53-59.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology *268*.

Castle, J., Zhang, C., Shah, J., Kulkarni, A., Kalsotra, A., Cooper, T., and Johnson, J. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nature genetics *40*, 1416-1425.

Culler, S., Hoff, K., Voelker, R., Berglund, J., and Smolke, C. (2010). Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. Nucleic acids research *38*, 5152-5165.

Doma, M.K., and Parker, R. (2006). Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature *440*, 561-564.

Fairbrother, W., Yeh, R.-F., Sharp, P., and Burge, C. (2002). Predictive identification of exonic splicing enhancers in human genes. Science (New York, NY) *297*, 1007-1013.

Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature.

Garcia-Blanco, M., Baraniak, A., and Lasda, E. (2004). Alternative splicing in disease and therapy. Nature biotechnology *22*, 535-546.

Genomes Project, C., Abecasis, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., Kang, H., Marth, G., and McVean, G. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Gibson, D., Young, L., Chuang, R.-Y., Venter, J., Hutchison, C., and Smith, H. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature methods *6*, 343-345.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. Molecular cell *22*, 769-781.

Huelga, S., Vu, A., Arnold, J., Liang, T., Liu, P., Yan, B., Donohue, J., Shiue, L., Hoon, S., Brenner, S., *et al.* (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell reports *1*, 167-178.

Katz, Y., Wang, E., Airoldi, E., and Burge, C. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods *7*, 1009-1015.

Ke, S., Shang, S., Kalachikov, S., Morozova, I., Yu, L., Russo, J., Ju, J., and Chasin, L. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome research *21*, 1360-1374.

Kim, E., Goren, A., and Ast, G. (2008). Insights into the connection between cancer and alternative splicing. Trends in genetics : TIG *24*, 7-10.

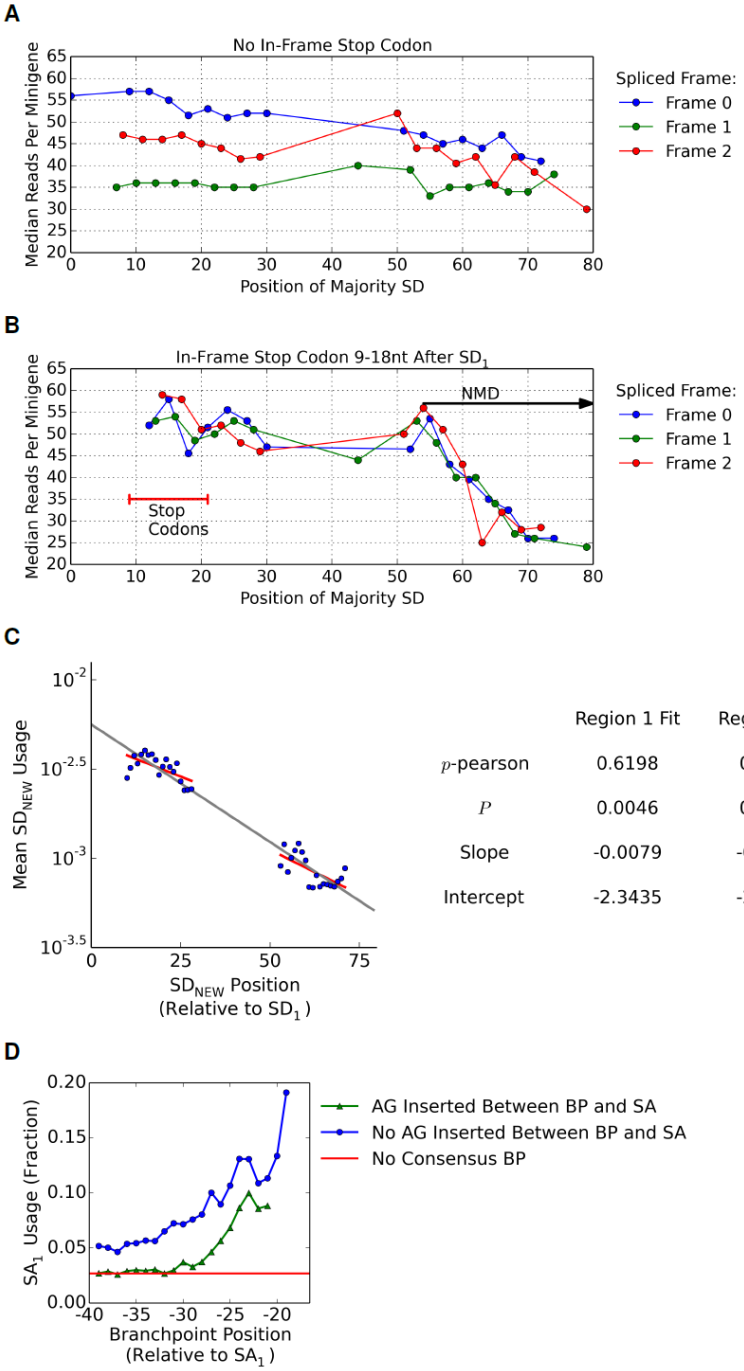
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lappalainen, T., Sammeth, M., Friedländer, M., t Hoen, P., Monlong, J., Rivas, M., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P., *et al.* (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.

Le, Q.V. (2013). Building high-level features using large scale unsupervised learning. Paper presented at: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (IEEE).
Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proceedings of the National Academy of Sciences *100*, 189-192.
Listerman, I., Sapra, A., and Neugebauer, K. (2006). Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nature structural & molecular biology *13*, 815-822.
Luco, R., Pan, Q., Tominaga, K., Blencowe, B., Pereira-Smith, O., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science (New York, NY) *327*, 996-1000.
Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nature reviews Molecular cell biology *8*, 479-490.
Martinez-Contreras, R., Fisette, J.-F., Nasim, F., Madden, R., Cordeau, M., and Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS biology *4*, 172.
Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C., Kinney, J., *et al.* (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nature biotechnology *30*, 271-277.
Mercer, T.R., Clark, M.B., Andersen, S.B., and Brunck, M.E. (2015). Genome-wide discovery of human splicing branchpoints. Genome ….
Nilsen, T., and Graveley, B. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457-463.
Oberstrass, F., Auweter, S., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D., *et al.* (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science (New York, NY) *309*, 2054-2057.
Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. Cell reports.
Patwardhan, R., Hiatt, J., Witten, D., Kim, M., Smith, R., May, D., Lee, C., Andrie, J., Lee, S.-I., Cooper, G., *et al.* (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nature biotechnology *30*, 265-270.
Patwardhan, R., Lee, C., Litvin, O., Young, D., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nature biotechnology *27*, 1173-1175.
Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. Molecular and Cellular Biology *10*, 84-94.
Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nature biotechnology *30*, 521-530.
Shoemaker, C.J., Eyler, D.E., and Green, R. (2010). Dom34: Hbs1 promotes subunit dissociation and peptidyl-tRNA drop-off to initiate no-go decay. Science.

Smith, R., Taher, L., Patwardhan, R., Kim, M., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nature genetics.
Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B., and Darnell, R. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature *444*, 580-586.
Wang, E., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S., Schroth, G., and Burge, C. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.
Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nature structural & molecular biology.
Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C., and Wang, Z. (2013). A complex network of factors with overlapping affinities represses splicing through intronic elements. Nature structural & molecular biology *20*, 36-45.
Wang, Z., Rolish, M., Yeo, G., Tung, V., Mawson, M., and Burge, C. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831-845.
Wang, Z., Xiao, X., Van Nostrand, E., and Burge, C. (2006). General and specific functions of exonic splicing silencers in splicing control. Molecular cell *23*, 61-70.
White, M., Myers, C., Corbo, J., and Cohen, B. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proceedings of the National Academy of Sciences of the United States of America *110*, 11952-11957.
William, L.N., Ross, J.F., Aparna, B., Alexander, J.D.d.A., Jiajing, Z., Paul, A.K., and Clifford, L.W. (2014). Quantitative analysis of mammalian translation initiation sites by FACS-seq. Molecular Systems Biology.
Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., *et al.* (2014). The human splicing code reveals new insights into the genetic determinants of disease. Science (New York, NY).
Yeo, G., and Burge, C. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology : a journal of computational molecular cell biology *11*, 377-394.
Yu, Y., Maroney, P., Denker, J., Zhang, X., Dybkov, O., Lührmann, R., Jankowsky, E., Chasin, L., and Nilsen, T. (2008). Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. Cell *135*, 1224-1236.
Zhang, X., and Chasin, L. (2004). Computational definition of sequence motifs governing constitutive exon splicing. Genes & development *18*, 1241-1250.

Supplemental Figures



	Region 1 Fit	Region 2 Fit	Combined Fit
p-pearson	0.6198	0.6433	0.9761
P	0.0046	0.0030	1.81e-25
Slope	-0.0079	-0.0100	-0.0132
Intercept	-2.3435	-2.4523	-2.2495

Figure S1. Position Dependence of Alternative Splicing and Effects on RNA Stability, Related to Figure 2
(A) Reading frame affects RNA stability. Minigenes were grouped by the position of the SD with the most reads. SDs that shift the reading frame have fewer median reads than those that maintain the same reading frame.
(B) Premature stop codons negate the effects of reading frame on RNA stability. The same analysis as (A) was performed for minigenes with an in-frame stop codon starting 9-18 nt downstream of SD_1. The median number of reads is no longer dependent on the spliced reading frame. We also observe evidence of non-sense mediated decay for SDs more than 40 nt downstream of the stop codons.
(C) The probability of splicing at new positions within the randomized regions shows a statistically significant position dependency.
(D) Distal branchpoints reduce splice acceptor usage, even in the absence of an AG between the branchpoint and splice acceptor.

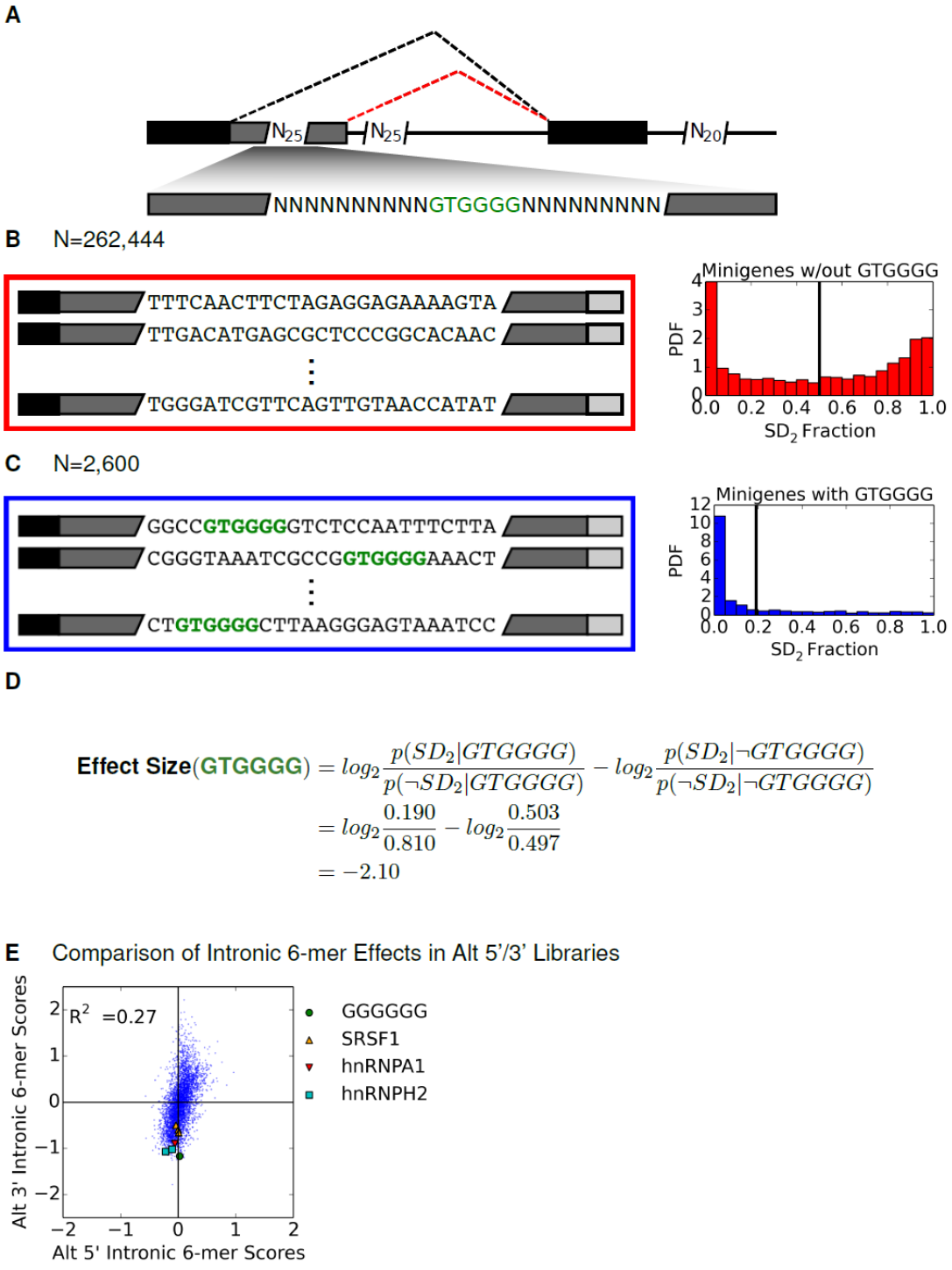


Figure S2. Measuring the Effect Sizes of Motifs on Alternative Splicing, Related to Figure 3

(A) An example of calculating the effect size of GTGGGG on SD_2 usage when located in the first degenerate region.

(B) There are 262,444 minigenes that do not contain GTGGGG in the first degenerate region. On the right a histogram of SD_2 usage is shown with the mean usage indicated with a black vertical line (50.0%).

(C) There are 2,600 minigenes that do contain GTGGGG in the first degenerate region. These minigenes splice an average of 19.0% at SD_2.

(D) To get an effect size we calculate the $\log_2$ odds ratio.

(E) A comparison of intronic 6-mer effects (x-axis: Figure 3B, y-axis: Figure 3D)

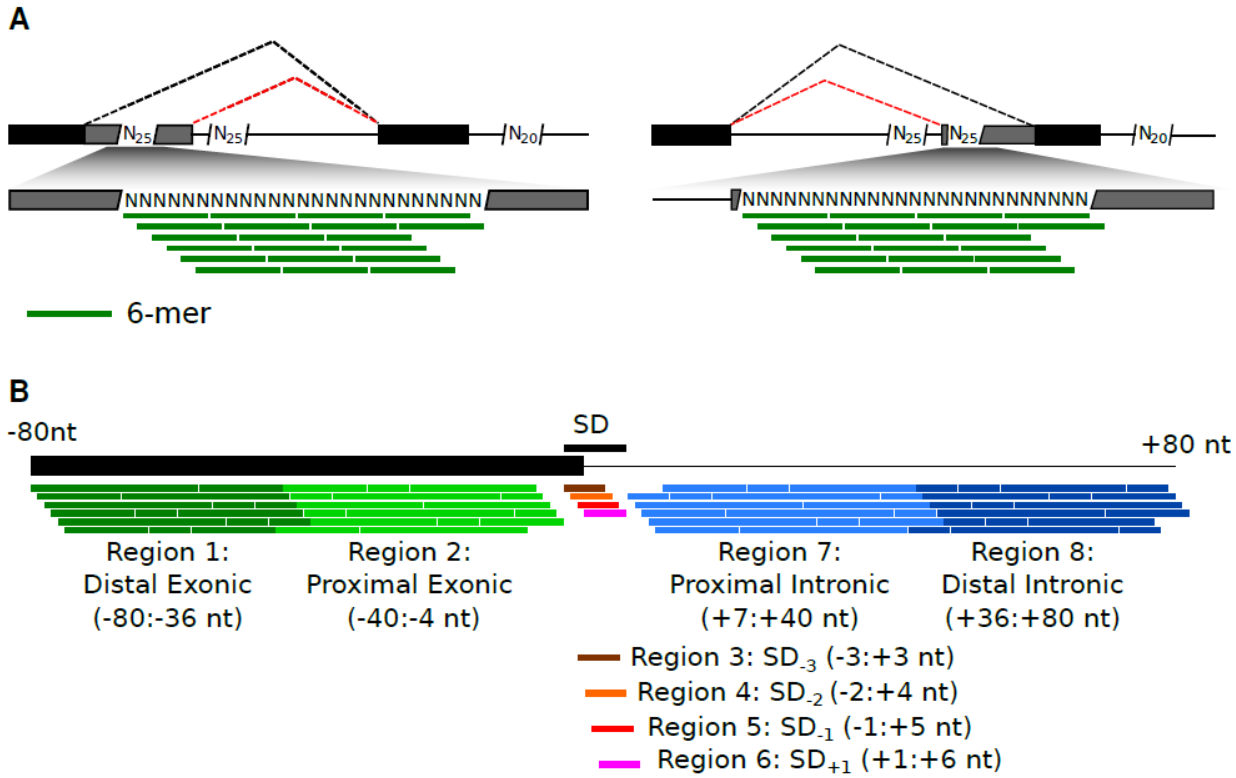


Figure S3. Model Structures, Related to Figure 4

(A) The HAL joint model of alternative exon definition was trained using both the alternative 5' and 3' splicing libraries. The model learns a shared weight for each of the 4,096 possible 6-mers.

(B) The HAL model of splice donors. We split the region surrounding the SD (80 nt upstream and 80 nt downstream) into 8 different regions. Within each region, we use the counts of 6-mers as features. In total, this model has 8 x 4^6 (32,768) features.

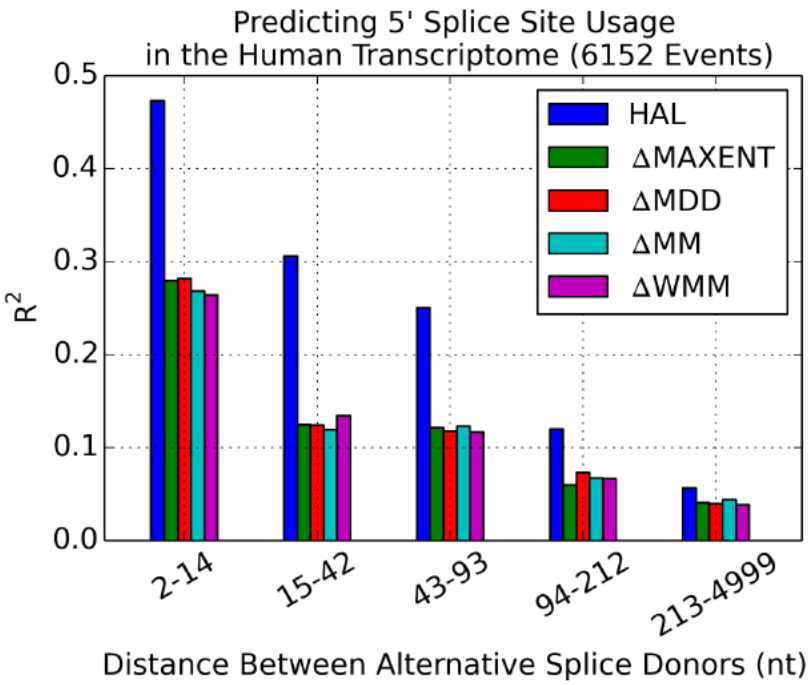


Figure S4. Predicting Isoform Ratios in Alternative 5' Splicing, Related to Figure 5
HAL was used to predict exon inclusion levels in 6,152 alternative 5' splicing events expressed in lymphoblastoid cell lines and compared to four other splice donor algorithms.

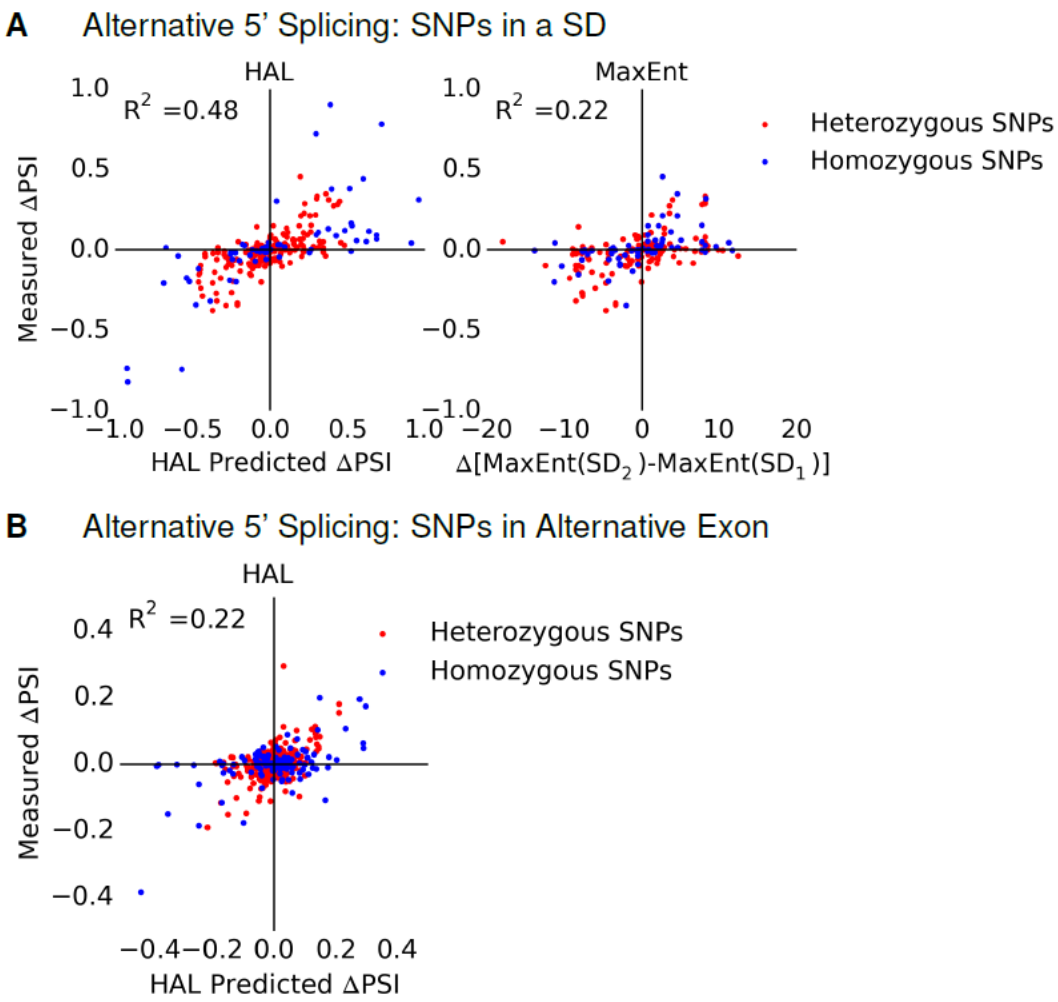


Figure S5. Predicting the Effects of SNPs in Alternative 5' Splicing, Related to Figure 6
(A) We compared the predicted effects of SNPs occurring an alternative splice donor for both HAL and MaxEnt.
(B) The predicted and observed effects of SNPs occurring between to alternative splice donors.

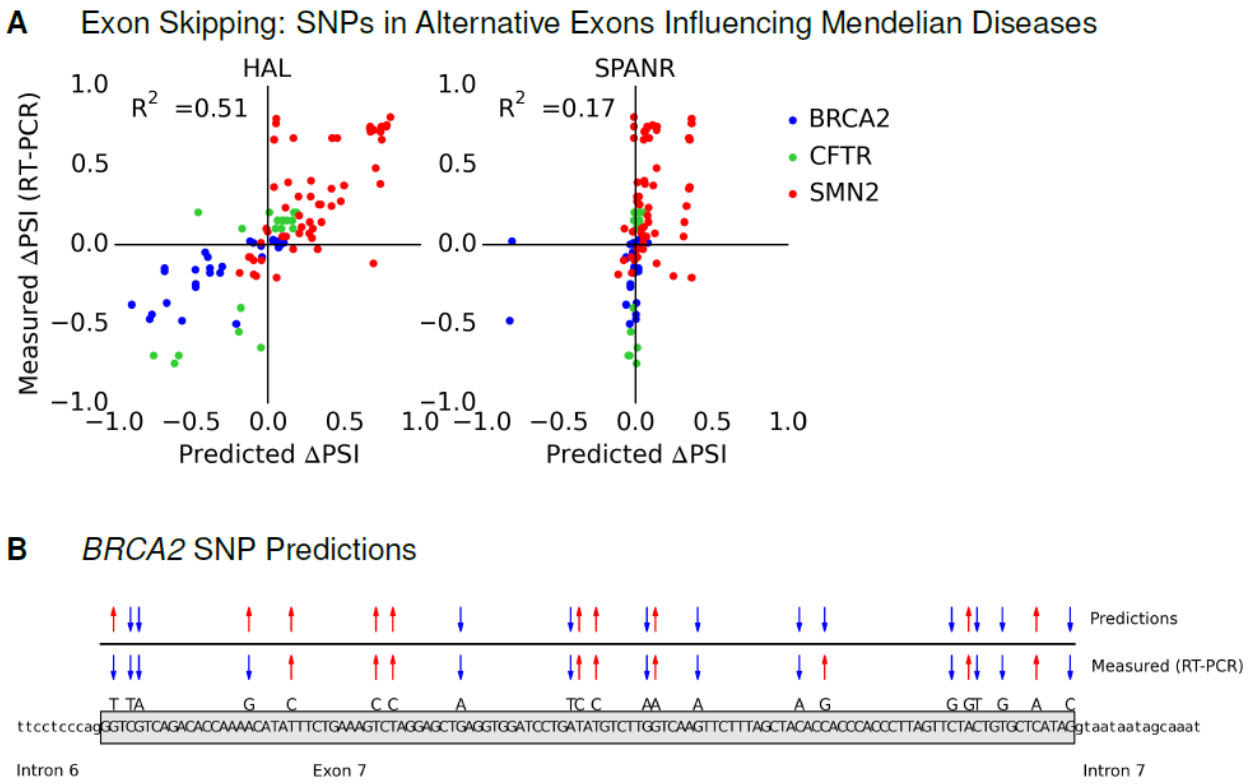


Figure S6. Comparing HAL Exon Skipping Predictions to State of the Art, Related to Figure 7
(A) We compared predictions of HAL to the state of the art (SPANR) for exonic SNPs in three different genes known to cause mendelian disease. We found that HAL has three times more explanatory power than SPANR (HAL: R^2=0.51; SPANR: R^2=0.17).
(B) Model predictions are compared to experimental results using RT-PCR for SNPs occurring in exon 7 of BRCA2. For positions with data for multiple SNPs, the SNP with the largest measured change in PSI was plotted. The model accurately predicted the directional change in PSI (increased exon inclusion/exclusion), for 19/22 of the SNPs plotted.

Supplemental Experimental Procedures

Calculating Sequence Motif Effect Sizes

First, we define the odds of splicing at a given splice site as the ratio of the probability of splicing at that site versus the probability of splicing at other sites. This can be written as:

$$\frac{p(SD_i)}{1-p(SD_i)}$$

When dealing with effect sizes, odds are more useful than probabilities. For example, if the odds of splicing at a splice site are 2:1, we can say that if the odds of splicing at the given site are doubled, the new odds will be 4:1. However, if we use the equivalent probabilities (66.7% chance of splicing at the given site), it makes no sense to say the probability was doubled as the new probability will be 133.3%. So if we want to define the effect size of a motif, an intuitive way to do it is to use the ratio of the odds when the motif is present and the odds when the motif is absent:

$$odds\ ratio = \frac{p(SD_i|motif)/(1-p(SD_i|motif))}{p(SD_i|\neg motif)/(1-p(SD_i|\neg motif))}$$

It can easily be shown that if we assume that the effects of two motif on splicing are conditionally independent, we can calculate the joint odds ratio as the product of the two individual odds ratios. First we write the joint odds ratios as follows:

$$odds\ ratio_{1,2} = \frac{p(SD_i|motif_1, motif_2)/(1-p(SD_i|motif_1, motif_2))}{p(SD_i|\neg(motif_1, motif_2))/(1-p(SD_i|\neg(motif_1, motif_2)))}$$

If we assume conditional independence, we can rewrite the joint odds ratios as follows:

$$odds\ ratio_{1,2} = \frac{p(SD_i|motif_1)/(1-p(SD_i|motif_1))}{p(SD_i|\neg motif_1)/(1-p(SD_i|\neg motif_1))} \cdot \frac{p(SD_i|motif_2)/(1-p(SD_i|motif_2))}{p(SD_i|\neg motif_2)/(1-p(SD_i|\neg motif_2))}$$

Which is simply:

$$odds\ ratio_{1,2} = odds\ ratio_1 \cdot odds\ ratio_2$$

However, if we instead use the log_2 odds ratio, the effect size of conditionally independent motifs becomes simply additive, rather than multiplicative. It is also worth noting that in logistic regression, the coefficients are also log odd ratios. For these reasons, we chose to use the $log2$ odds ratio as the measure of our effect sizes:

$$Effect\ Size = log_2 \frac{p(SD_i|motif)/(1-p(SD_i|motif))}{p(SD_i|\neg motif)/(1-p(SD_i|\neg motif))}$$

In the alternative 5' library analysis in Figure 3 and Figure 4, we only considered how motifs affected the ratio of SD_2 to SD_1, rather than the ratio of SD_2 to all other sites, which would be complicated by the insertion of new splice donors:

$$Effect\ Size = log_2 \frac{p(SD_2|motif)/p(SD_1|motif)}{p(SD_2|\neg motif)/p(SD_1|\neg motif)}$$

Likewise the alternative 3' analysis only considered how motifs affected the ratio of SA_1 to SA_2:

$$Effect\ Size = log_2 \frac{p(SA_1|motif)/p(SA_2|motif)}{p(SA_1|\neg motif)/p(SA_2|\neg motif)}$$

To estimate these effect sizes in practice, we needed to estimate $p(SD_i|motif)$ as well as $p(SD_i|\neg motif)$. To estimate $(SD_i|motif)$, we simply calculated the mean splice site usage in the subset of plasmids containing the motif. Likewise to calculate $p(SD_i|\neg motif)$, we calculated the mean splice site usage in the subset of plasmids not containing the motif. To calculate 95% confidence intervals, we repeated these effect size estimate 200 times with different resampled (with replacement) copies of the splicing libraries. This method was used estimated the effect sizes of both 6-mers and combinations of non-overlapping 4-mers ($4mer_A, N_{0-17}, 4mer_B$), although due to computational constraints, we did not calculate confidence intervals for the combinations of 4-mers.

Learning a Model to Predict Splice Site Usage in the Alternative 5 Library

To predict the usage of splice donors in the alternative 5 library, we set out to develop a model to score splice donors based upon the surrounding sequence (80 nt upstream and 80 nt downstream of the potential splice site). We expected that the position of sequences relative to the splice donor would be important (e.g. upstream vs downstream of the splice donor), so we chose to divide the 160nt window into smaller regions. The regions we chose are as follows (Figure S3B):

1. Distal Exonic: -80 to -36
2. Proximal Exonic: -40 to -4
3. Splice Donor$_{-3}$: -3 to +3
4. Splice Donor$_{-2}$: -2 to +4
5. Splice Donor$_{-1}$: -1 to +5
6. Splice Donor$_{+1}$: +1 to +6
7. Proximal Intronic: +7 to+40
8. Distal Intronic: +36 to +80

Within each region, we used the counts of each possible 6-mer as features. Regions 3-6 are only 6nt, so they have a single 6-mer as a feature. In total there are $4^6 \cdot 8$ (32,768) features, although most feature counts equal zero. To score a single splice donor, we simply took the dot product between the 32,768 vector of 6-mer counts and the 32,768 vector of feature scores. Calculating feature scores will be described shortly. Each minigene can potentially be spliced at many different splice donors, but we can score each potential splice donor based upon the sequence surrounding it (80nt upstream and 80nt downstream). To account for potential positional biases in splicing (e.g. splice donors that are transcribed first might be preferred regardless of sequence), we included a bias term for each position relative to SD_1. For example, the splice donor score of a site 28 downstream of SD_1 would be the dot product of this sites feature vector and the feature scores plus the bias term for splice donors at position +28:

$$\text{Score}(\text{SD}_{pos=28}) = \mathbf{x} \cdot \boldsymbol{\beta} + \beta_{pos=28}$$

To account for the potential that minigenes can be unspliced, we included a single bias term for no splicing that does not depend on the sequence of the minigene. Once we have scores for each splice donor, these scores can be converted to probability estimates using the softmax function:

$$p(\text{SD}_i) = \frac{e^{score(\text{SD}_i)}}{e^{score(no\ splicing)} + \sum_{j=1}^{n} e^{score(\text{SD}_j)}}$$

In order estimate the optimal feature weights, we minimized the Kullback-Leibler divergence between our models probability estimates ($\hat{p}(\text{SD}_i)$) and the observed splicing probabilities from our data ($p(\text{SD}_i)$) across n minigenes and k potential splice donors (and no splicing) with an additional

L1 regularization penalty on the weights (excluding the bias terms) to prevent overfitting:

$$Loss = \lambda|\boldsymbol{\beta}|_1 + \sum_{i=1}^{n}\sum_{c=1}^{k+1} p(y_i = \mathrm{SD}_c) \ln \frac{p(y_i = \mathrm{SD}_c)}{\hat{p}(y_i = \mathrm{SD}_c)}$$

In order to allow for 10-fold cross-validation, we randomly partitioned our data into 10 parts. Cross-validation is performed by splitting data multiple times into a training and test set, and training and testing a new model for each split. In this way predictions can be made for the whole dataset, without testing a model on data that it also learned from. For each test partition, we trained our model on the remaining 9 partitions. Within each training set, we optimized the L1 regularization parameter using 8 partitions for training and 1 partition for validation, before using the full training set with the optimized λ.

Generating a Learning Curve for Different Models Predicting SD_1 Usage

One of the questions we wanted to answer was how more data can improve the performance of different models. Previously, we described a model to predict splicing fractions at any splice donor in the alternative 5 splicing library. However, as this model is computationally expensive to train, we chose to generate learning curves for a similar, but simpler model, predicting only the probability of using SD_1. We trained different models using either 3-mers, 4-mers, 5-mers, 6-mers, or 7-mers as features. More specifically we used the counts of n-mers within the first degenerate region as well as the counts of n-mers within the second degenerate region. So for 3-mers, there are $4^3 \cdot 2$ (128) features, while the 7-mer model contains $4^7 \cdot 2$ (32,768) features. We would naturally expect the models with more features to perform poorly without sufficient data as these models can easily overfit the training data. To some degree this can be alleviated by using L1 regularization, which we did. On the other hand, the models with less features should quickly reach a point at which more training data provides minimal added benefit, because the models are too simple. We split our data into training and test sets (90%/10% split) and trained models using subsets of the training data (between 100 to 177,827 training points). For each model and subset training size, we also tested different amounts of L1 regularization. For each combination of n-mer model and training size, we recorded the results corresponding to the optimal L1 regularization.

Learning a General Model of Alternative Splicing from both the Alternative 5 and 3 Libraries

After finding that 6-mer effect sizes in the alternative exon region of both the alternative 5' and 3' library were so correlated, we chose to learn a single model of using both datasets. For the 5' library, we aimed to predict the relative usage of SD_2 to SD_1 (SD_2/(SD_1+ SD_1)), using only the 6-mer counts in the alternative exon region as features. Similarly, for the 3' library, we aimed to predict the fractional usage of SA_1, again using only the counts of 6-mers in the alternative exon region as features. To learn a single shared score for each 6-mer between the two datasets (Figure S3A), we trained a model to minimize the Kullback-Leibler divergence between our probability estimates and the observed splicing probablilites across both datasets. Our model did include unique bias terms for each library to account for different baseline levels of splice site usage. Since this model was only used for testing the effects of SNPs in other datasets, we trained the model using all of our data. The resulting model contained 4^6 (4,096) parameters for the 6-mers and 2 bias terms.

Predicting Alternative 5' Splicing Ratios in Lymphoblastoid Cell Lines

We first needed to get a list of alternative splicing events with both their sequences and isoform ratios. A list of alternative splicing events was obtained from: `http://genes.mit.edu/burgelab/miso/annotations/ver2/miso_annotations_hg19_v2.zip`. RNA-seq for 462 lymphoblastoid cell lines (LCLs) created from different individuals was obtained from the GEUVADIS (Lappalainen et al., 2013) consortium. Each splicing event has two competing splice donors. For each individual cell line, we used MISO (Katz et al., 2010) to calculate the Percent Spliced In (PSI: fraction of splicing at the downstream splice donor) for every alternative splicing event. To get a global PSI, we averaged the PSIs across all individual LCLs. To reduce the effects of noisy measurements (the coefficient of variation of PSI between LCLs averaged across all events was 0.310), we discarded events that were expressed in less than 10 individual LCLs. We then used our alternative 5' splicing model to predict the PSI for each event. We also tested 4 different splice donor scoring algorithms (Maximum Entropy Model, Maximum Dependence Decomposition, First-order Markov Model, Weight Matrix Model) to get score differentials between the upstream and downstream splice donors. We were interested in how each model performed depending on the length of the alternative exon region. So we grouped the alternative splicing events into 5 equal bins (1230 or 1231 events each) based on the length of the alternative exon region. Within each group, we reported the R^2 between our model's predicted PSI and the measured mean PSI. For the other algorithms, we reported the R^2 between the splice donor score differentials and the measured mean PSI.

Applying the Model to Predict Effects of Human SNPs on Alternative 5' Isoforms

We first needed to compile a list of SNPs and wildtype sequences and their respective PSI in different alternative 5' splicing events. To do this, we used the same list of alternative 5' splicing events described previously. In addition to the RNA-seq data describe above, we obtained genotypes for the same LCLs from the Thousand Genomes Project(Genomes Project et al., 2012). For each alternative splicing event, we grouped individuals based on their genotype. Individuals with no SNPs within 300nt of either alternative splice donor were categorized as wild type. Individuals with a single SNP on one allele and none on the other were categorized as heterozygous, while those with the same single SNP on both alleles were categorized as homozygous. Individuals with multiple SNPs on either allele within 300 nucleotides of either alternative splice donor were discarded. We then used MISO(Katz et al., 2010) to calculate the PSI for every alternative splicing event within all 462 individuals. For many alternative splicing events, low read counts led to unreliable PSI values in specific individuals. However, by averaging PSIs across all the individuals with the same genotype (wild type, heterozygous SNP, or homozygous SNP), we were able to get a much more reliable estimate of the PSI for each genotype. We used these averaged PSIs for our future analysis. We predicted the change in PSI due to SNPs that occurred either within one of the two alternative splice donors themselves or in the alternative exon region. For SNPs that occurred in the alternative exon region and not in one of the two splice donors, we predicted the change in PSI using the model that we trained on both the 5' and 3' libraries. We first calculated the sum of the 6-mer scores overlapping the WT nucleotide in the position of the SNP. We then calculated the sum of the 6-mer scores overlapping the SNP. Then we calculated the difference of these scores and using the original PSI, we calculated a predicted PSI for the sequence with the SNP:

$$\text{PSI}_{SNP\_Pred} = sigmoid(\text{Score}_{SNP} - \text{Score}_{WT} + \ln(\text{PSI}_{WT}) - \ln(1 - \text{PSI}_{WT}))$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

Our predictions make the assumption that individuals with a heterozygous copy of the SNP will express the given transcript equally from both alleles, so the combined PSI should simply be the average of PSI_{SNP} and PSI_{WT}:

$$\text{PSI}_{SNP\_Pred(HOMO)} = \text{PSI}_{SNP\_Pred}$$

$$\text{PSI}_{SNP\_Pred(HETERO)} = (\text{PSI}_{SNP\_Pred} + \text{PSI}_{WT})/2$$

For SNPs that occurred within a splice donor, we calculated score differentials between the splice donor with and without the SNP. To calculate these scores we used the splice donor model described previously, but we only used the features from the regions directly overlapping the splice donor (Splice Donor_{-3}, Splice Donor_{-2}, Splice Donor_{-1}, Splice Donor_{+1}). We then predicted the PSI for the sequence with the SNP in the same manner as above.

Predicting the Effects of Exonic Variants on Exon Skipping Levels

We then wanted to predict the effects of sequence variants on exon skipping. Specifically, we wanted to predict variants within the actual alternative exon. We went through the literature and generated a list of SNPs with measured PSI for both the wildtype and SNP sequences. These PSI values were measured by running gels after RT-PCR and comparing the band intensities between different isoforms[1, 2, 3, 4, 5, 6]. For variants not occurring in either the splice acceptor or splice donor, we could simply apply the model that we learned from both the alternative 5' and 3' datasets. However, we chose to consider variants within the actual splice sites separately. To score variants that occurred within the splice donor of the alternative exon, we used the splice donor model that we trained previously. However, we only used the feature scores of the regions overlapping the exonic part of the splice donor (Splice Donor_{-3}, Splice Donor_{-2}, Splice Donor_{-1}). To account for variants that occurred in the splice acceptor sequence (+1 to +3 in the exon), we scored every 3-mer using the alternative 3' library. Specifically, we calculated effect size of each 3-mer on the odds of splicing at a new splice acceptor in the second degenerate region in the alternative 3 library, when the 3-mer was located in the +1 to +3 position of the splice acceptor candidate. To predict the effects of a variant (or variants), we scored the sequence with and without the variant(s) and calculated the difference in scores:

$$\Delta\text{Score} = \text{Score}_{\text{VAR}} - \text{Score}_{\text{WT}}$$

To then predict PSI_{VAR} and ΔPSI we could use the following equations:

$$\text{PSI}_{VAR} = sigmoid(log(\text{PSI}_{WT}) - log(1 - \text{PSI}_{WT}) + \Delta\text{Score})$$

$$\Delta\text{PSI} = \text{PSI}_{VAR} - \text{PSI}_{WT}$$

While the correlation of these predicted ΔPSI with the experimental ΔPSI was good, the magnitudes of the predictions were smaller than the experimental data. This indicated that while exonic variants had the same type of effect in exon skipping as alternative 5' or 3' splicing, the magnitude of the effects were larger in exon skipping. To account for this in our predictions, we simply included a scaling term for our scores:

$$\Delta\text{Score}' = \alpha \cdot \Delta\text{Score}$$

To choose the best value of α, we used 10-fold cross validation on our dataset (variants in BRCA2, CFTR, SMN1/2). In all of the folds, the optimal α was between 2.9 and 3.1. The predictions that

we report in Figure 7 and Figure S6 are the results of this 10-fold cross validation. To compare our predictions to SPANR, we entered the SNPs in our BRCA2/CFTR/SMN dataset into the online webtool (http://tools.genes.toronto.edu).We downloaded the results and used the ΔPSI reported for our comparisons. At the time of this manuscript, SPANR only supported SNPs and not more complex variants.

References

[1] Luca Cartegni, Michelle L Hastings, John A Calarco, Elisa de Stanchina, and Adrian R Krainer. Determinants of exon 7 splicing in the spinal muscular atrophy genes, smn1 and smn2. *The American Journal of Human Genetics*, 78(1):63–77, 2006.

[2] Daniela Di Giacomo, Pascaline Gaildrat, Anna Abuli, Julie Abdat, Thierry Frébourg, Mario Tosi, and Alexandra Martins. Functional analysis of a large set of brca2 exon 7 variants highlights the predictive value of hexamer scores in detecting alterations of exonic splicing regulatory elements. *Human mutation*, 34(11):1547–1557, 2013.

[3] Franco Pagani, Michela Raponi, and Francisco E Baralle. Synonymous mutations in cftr exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18):6368–6372, 2005.

[4] Natalia N Singh, Elliot J Androphy, and Ravindra N Singh. An extended inhibitory context causes skipping of exon 7 of smn2 in spinal muscular atrophy. *Biochemical and biophysical research communications*, 315(2):381–388, 2004.

[5] Natalia N Singh, Elliot J Androphy, and Ravindra N Singh. In vivo selection reveals combinatorial controls that define a critical exon in the spinal muscular atrophy genes. *Rna*, 10(8):1291–1305, 2004.

[6] Natalia N Singh, Ravindra N Singh, and Elliot J Androphy. Modulating role of rna structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic acids research*, 35(2):371–389, 2007.

Chapter 4: A massively parallel approach to learning the impact of 5'UTR sequence on translation

The final project in my thesis has been a collaboration between our lab (Anna Kuchina, Ben Groves, Georg Seelig, and myself) and Stan Field's lab (Josh Cuperus and Stan Fields), using MPRAs and machine learning to model the effects of 5'UTRs in yeast on protein production. Measuring protein production in an MPRA presents unique challenges from the alternative splicing project. While RNA can be quantified with RNA-seq, protein levels cannot be quantified directly through high-throughput sequencing. Other groups had previously demonstrated that fluorescent activated cell sorting (FACS) could be combined with high-throughput sequencing to measure translation of thousands of different sequence variants in parallel. We began this project with the intent to use that method, but became skeptical that it would scale to hundreds of thousands or millions of variants.

We then switched to a competitive growth assay, in which growth of each cell is proportional to expression of the HIS3 gene. Using this method, we were able to measure protein production of hundreds of thousands of HIS3 variants, with different 5'UTR sequences. We then used this enormous dataset to improve our current understanding and ability to model translation in yeast. The project is still ongoing and the work presented here will likely differ from the final paper as we continue to gain insight into the regulatory mechanisms of the 5'UTR. I debated whether or not to include this work since the manuscript is not in its final form, but the work here constitutes a major part of my PhD, so I chose to include it.

Rosenberg, A. B.*, Cuperus, J.*, Kuchina, A.*, Groves, B.*, Fields, S., & Seelig, G. (2016). A massively parallel approach to learning the impact of 5'UTR sequence on translation (In preparation)

* equal contributors

The relationship between DNA sequence and protein expression is at the heart of the central dogma. And yet, our ability to predict protein expression from DNA sequence alone remains poor, hampering our understanding of both evolution and strain engineering. Here we address this challenge by learning to predict the translational efficiency of 5' untranslated regions. We developed a competitive growth assay to measure the translational efficiency of over 500,000 fully degenerate, synthetic UTR sequences, two orders of magnitude more UTRs than exist in the native *Saccharomyces cerevisiae* genome. While mRNA expression measurements are easily attained through RNA-seq, our approach makes it possible to quantitatively measure protein expression in a single high-throughput experiment. Our data allowed us to quantitate with unprecedented precision the impact on translation of Kozak sequence composition, upstream start codons, secondary structure, and even amino acid composition following a start codon. We then trained regression and convolutional neural network models on our data and identified models of translation that result in highly accurate predictions of 5' UTR efficiency independently of the coding sequence. We anticipate that our model can be used in strain engineering to predictably alter protein expression through introduced 5' UTR nucleotide variants. The model can also be used to examine the role of 5' UTR variants in strain evolution.

Predicting protein expression from DNA sequence is a fundamental challenge in genomics. For example, sequencing of many human genomes has resulted in a catalogue of millions of genetic variants, many with unknown roles in protein synthesis. The complexity of biological systems makes it forbidding to experimentally and systematically explore all variants of potential interest. Similarly, in metabolic engineering, pathways with 10 or more enzymes are often needed to produce a molecule of interest. Achieving a high yield requires the expression levels of many of these components to be tuned; for example, to minimize the steady state levels of toxic intermediates without limiting the productive flow through the pathway. Again, it would be impossible to exhaustively search the space of all possible combinations of regulatory elements to identify those for optimal pathway performance. A comprehensive and predictive model of sequence–function relationships would be of enormous utility in both of these applications, making it possible to predict the effect of mutations in the human genome or to suggest which sequence elements would result in optimal production from a metabolic pathway. Most importantly, such models would also expand our understanding of the "grammar" of gene regulation. However, although progress towards such models is occurring, we are still far from being able to quantitatively predict the impact of *cis*-regulatory sequence elements on gene expression.

Cis-regulatory elements encoded in the DNA sequence determine the amplitude, timing and signal responsiveness of transcription; the start and endpoints of RNAs and proteins; the stability and localization of mRNAs; the inclusion or exclusion of exon-encoded sequences in proteins; the translation rates of mRNAs; and other critical processes that affect the activities of RNAs and proteins. Here we focus specifically on how the 5' UTR regulates translation of the mRNA into protein in *Saccharomyces cerevisiae*. Eran Segal's laboratory recently analyzed the effects of polymorphisms and short sequence motifs in yeast UTRs using a moderately high-throughput

massively parallel reporter assay (MPRA)[1-3]. With respect to 5' UTRs, they identified several features influencing translation, including upstream start codons and secondary structure[1]. It has been also well established that the Kozak sequence (the region directly surrounding the start codon)[4,5] regulates translational efficiency in eukaryotes—including yeast. However, these studies only assayed the impact of short sequences (<10 nt) or point mutations in the yeast 5' UTR on translational efficiency. Given that the median yeast 5' UTR is 53 nt, an accurate model of translation should account for cis-regulatory elements throughout the whole UTR.

Rather than trying to learn the regulatory rules of translation from a limited set of native *S. cerevisiae* 5' UTRs (~5000), we chose to combine machine learning with a MPRA of ~500,000 synthetic 5' UTRs to learn a *de novo* model of translation. We show that the model learned on this fully synthetic dataset recapitulates many of the results of previous studies, while further deepening our understanding of the cis-regulatory elements in 5' UTRs. Furthermore, the model we learn from this data accurately predicts the translational efficiency of both native *S. cerevisiae* and synthetic 5' UTRs. We find that the model even accurately predicts the translational efficiency of 5' UTRs upstream of different coding sequences.

Experimental Assay

In order to better understand how sequences in the 5' UTR affect translation, we turned to synthetic, randomized sequences. By assaying approximately two orders of magnitude more 5' UTR sequences than exist in the yeast genome, we anticipated that we could achieve a high resolution view of the effects of *cis*-regulatory elements controlling translation. While previous groups have had success measuring transcriptional and translational efficiency of sequence libraries with Fluorescence-Activated-Cell-Sorting (FACS) methods such as Flow-seq[6], these

methods are not scalable to libraries with hundreds of thousands or millions of variants. In these methods, cells are separated into fluorescence-based bins using FACS, and then the cells within each bin are sequenced with associated barcodes. However, the FACS step imposes a bottleneck on the number of total cells that can be measured over a reasonable period of time (~1 million) and thus the total library complexity.

In order to overcome these limitations, we turned to a competitive growth assay. The assay is based upon the principle that yeast cells grown in media lacking histidine exhibit growth rates proportional to expression of the *HIS3* gene. The *HIS3* gene product—the imidazoleglycerol-phosphate dehydratase enzyme—catalyzes the sixth step in histidine biosynthesis and becomes the growth-limiting factor for cells under histidine (-) selection. This can be taken advantage of to study the effects of sequence variants (*e.g.* different 5'UTRs) on *HIS3* gene expression by measuring growth rates in media lacking histidine. One limitation of the assay is that at high *HIS3* expression levels, histidine biosynthesis may reach saturating levels. At these saturating levels growth is no longer be dependent on *HIS3* expression. To avoid this situation, cells can be grown in the presence of 3-amino-1,2,4-triazole (3-AT), a competitive inhibitor of the His3 protein[7]. By titrating the concentration of 3-AT, dynamic range of the growth assay can be maximized. Typically each sequence variant is grown separately to measure growth, but our assay is based on competitive growth with all variants pooled into a single culture.

However before testing the pooled, competitive growth assay, we wanted to verify the histidine selection with by testing individual *HIS3* variants in separate cultures. To do this, we constructed two control yeast strains, one harboring the *CYC1* 5' UTR and the other containing a strong hairpin known to impair translation[1,8]. Under various concentrations of 3-AT, the competitive inhibitor of *HIS3*, we found that the *CYC1* 5'UTR exhibited significantly faster doubling times than the strong

hairpin 5'UTR. We saw the maximal separation of growth rate at 1.5 mM (Figure 1A) and thus used this concentration for future experiments.

We then proceeded to construct a library of 489,348 5' UTR variants by inserting 50 nt of degenerate nucleotides directly upstream of the *HIS3* coding sequence, on a single-copy centromeric plasmid with the *CYC1* promoter driving expression (Figure 1A) and the *CYC1* terminator following *HIS3*. We chose this promoter and terminator as they are among the best-characterized genetic elements in *S. cerevisiae*[9-13]. The *CYC1* promoter is relatively short (298 nucleotides), with well-established TATA-binding protein sites and transcriptional start sites. There are only two upstream activating sequences (UASs), one for *HAP1*[14] and one for *MIG1*[15,16].

We also assembled a separate library of native *S. cerevisiae* 5' UTRs. For 5' UTRs less than 50 nt, we included the whole UTR upstream of the *HIS3* gene. For native 5' UTRs greater than 50 nt, we assembled 50 nt fragments that tiled the native sequence with 25-50 nt intervals. This native library included 11855 unique sequences.

Both libraries were separately transformed into an auxotrophic yeast strain with the *HIS3* gene inactivated. During and after transformation, cells were grown in media with histidine. For both the native and random 5'UTR libraries, we then performed three replicate large-batch selections in media lacking histidine with 1.5 mM 3-AT, collecting cells after ~6.2 doublings. We purified plasmid DNA for each replicate both before and after selection. We then used high-throughput sequencing to measure enrichment or depletion of each 5'UTR after selection in media lacking histidine compared to before selection. As our measure of translational strength, we then used log enrichment of each 5' UTR (log growth rate).

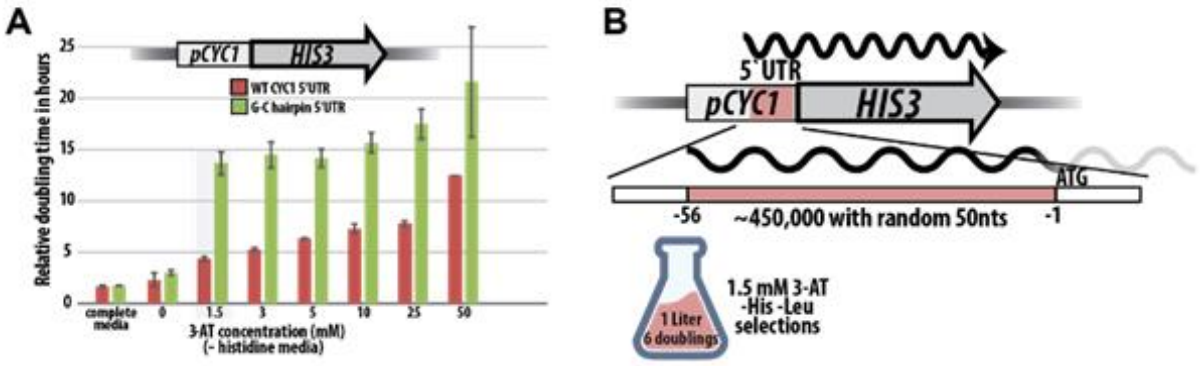


Figure 1. A growth-based massively parallel assay of 5'UTR sequence effects on translation
(A) Yeast with either the wild-type CYC1 promoter, or a CYC1 promoter with a strong hairpin structure in the 5' UTR with the HIS3 coding sequence were grown under increasing concentrations of the HIS3 competitive inhibitor 3-amino-triazole. 1.5 mM 3-AT was chosen for further assays, having the best separation of growth rates.
(B) Experimental design of a liquid-based growth assay of ~450,000 5' UTR variants, competed in 1.5 mM 3-AT.

Nucleotide Effects on Translational Efficiency

To understand how the 5' UTR sequence controls translation we first looked at the impact on growth rate of nucleotide identity in each position of the 5' UTR (Figure 2A). To obtain this relative per-position growth rate, we averaged growth rates for all library members with the same nucleotide in a specific position (e.g. "A" in position -50), excluding library members with an upstream ATG start codon. Our analysis shows that adenosine is the preferred nucleotide in every position of the 5' UTR, followed by thymine. An analogous per-position growth rate analysis performed with our library of native 5' UTRs produced a similar result, though with higher variability due to the much smaller library size.

The relative per-position growth rate measured in our assays closely mirrors the frequency of bases found at positions -50 to -1 in native yeast 5' UTRs (Figure 2B). This finding suggests that natural yeast 5' UTR regulatory sequences evolved to approach optimal expression levels, and therefore suggests that the vast majority of 5' UTRs are positioned for efficient translation. This strong preference for adenosines is unique to *S. cerevisiae*, while other fungi, like

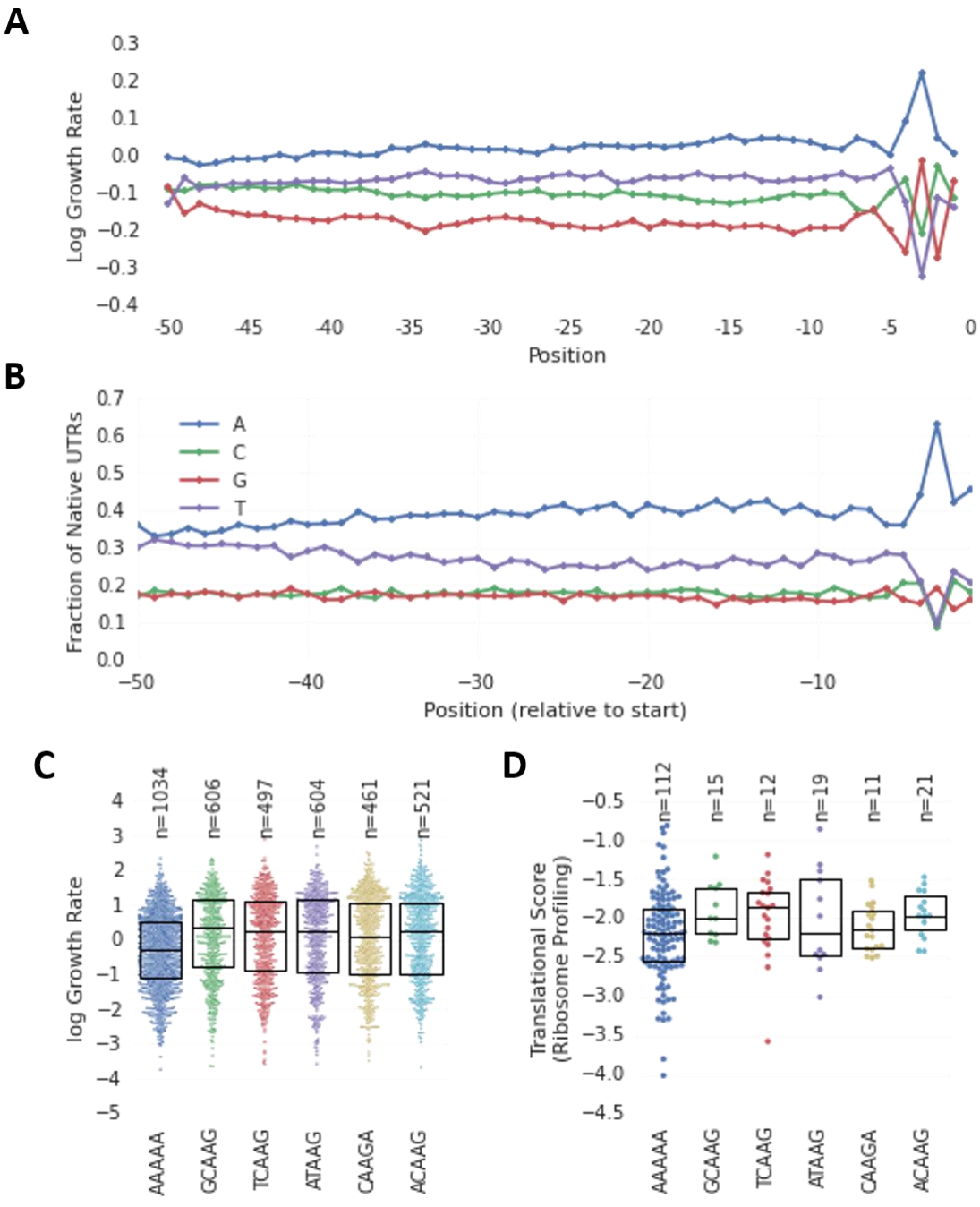


Figure 2. The influence of 5' UTR sequence on translational efficiency
(A) The mean growth rate of UTRs with the each nucleotide at each position
(B) The fraction of native UTRs with each nucleotide at each position.
(C) The most efficiently translated Kozak sequences (-5 to -1 position) in our *HIS3* selection compared to the traditionally recognized yeast Kozak sequence (AAAAA). The 25th, 50th and 75th percentiles are indicated.
(D) Ribosome profiling data supports our data the AAAAA is not the most efficiently translated Kozak sequence in *S. cerevisiae*. The 25th, 50th and 75th percentiles are indicated.

Schizosaccharomyces pombe, *Aspergillus nidulans* and *Neurospora crassa* have no significant enrichment across the 5' UTR (Figure S1).

Our data reflect the importance of nucleotide identity at positions -3 to -1. The strongest positive effect on translation is caused by an adenosine in the -3 position, consistent with prior reports[1,17,18]. This -3 adenosine preference is shared across many fungi and even into plants[19]. If only single base effects are taken into account, adenosine is the optimal choice for translation at each position within the Kozak sequence (positions -5 to -1). By extension, an all-A Kozak sequence should result in the highest translation. Five adenosines is both the consensus and most common Kozak sequence found in the yeast genome[20].

Given the size of our library, we can directly measure the translation efficiency of each of 1,024 possible 5-mers in position -5 to -1. Our random library contained on average 478 plasmids for each possible Kozak sequences and an effective growth rate can be obtained by averaging individual growth rates for all plasmids with the same Kozak sequence. This analysis revealed that the top 5 most efficiently translated Kozak sequences in our data are different from the all-A consensus motif, although they all have an adenosine at the -3 position (Figure 2C). While all adenine Kozak sequences are the most prevalent in the genome (Figure 2D), our data corroborates ribosome profiling data sets with transcripts containing the same 5 Kozak sequences having higher ribosome occupancy than an all adenosine Kozak sequence[21] (Figure 2D).

The effect of UTR length on translational efficiency

We then asked how the length of the UTR can affect translation. Since all of our randomized UTRs were exactly 50nt, we used the library of native UTRs for this analysis. Previous studies have found 5' UTR length is much more evolutionarily conserved in fungi than 3' UTR length[22].

We grouped all of the native yeast UTRs by length and then compared mean growth rates. We found that the optimal length UTR for translation was 25-45 nt (Figure 3B). Shorter UTRs as well as longer UTRs were translated less efficiently. However, the variability in translational strength increased with the length of the UTR. Presumably, longer UTRs have more potential for regulatory sites as well as the ability to form structure. The optimal UTR length of 25-45 nt also corresponds to the peak in the distribution of native yeast 5' UTR lengths (Figure 3A), suggesting that many 5' UTRs evolved to be this length in order to translate efficiently.

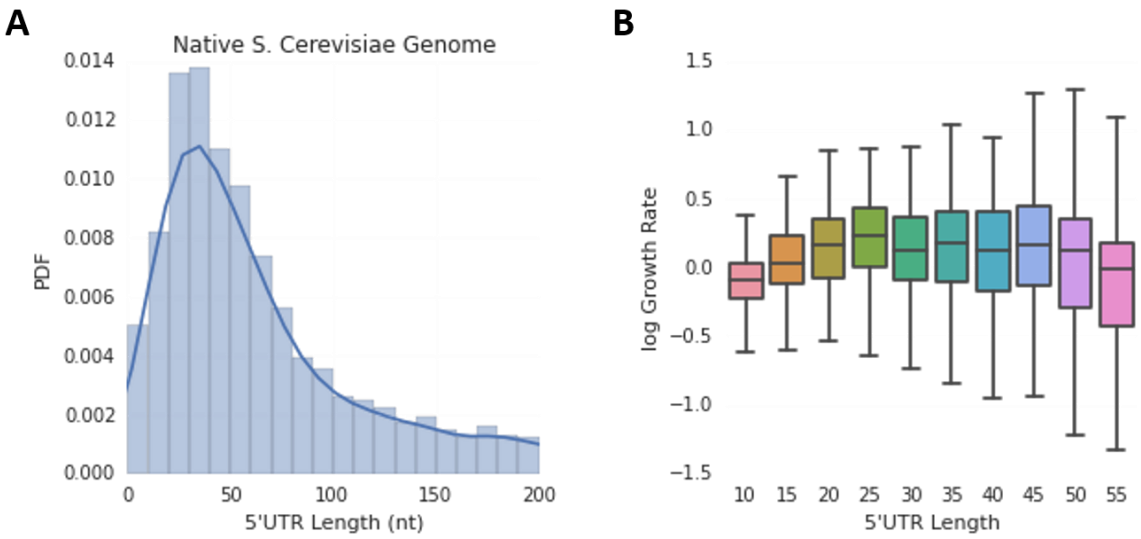


Figure 3. The effect of 5' UTR length on translation
(A) The distribution of 5' UTR lengths in the native *S. cerevisiae* genome with a peak around 30-40 nt.
(B) The measured growth rates of native *S. cerevisiae* 5' UTRs grouped by 5' UTR length also display a peak around 25-45 nt.

Upstream ATGs and uORFs

Upstream open reading frames (uORFs) are short translated reading frames upstream of the start codon which constitute well-known regulatory elements in the 5' UTR[1,23,24] (Figure 4A). About a thousand uORFs were predicted in the native yeast genome based on the presence of an upstream ATG codon (uATG)[25]. Because translation is thought to be initiated by a scanning mechanism where the 40S ribosome subunit binds at the 5’ cap and moves along the 5' UTR until

it reaches the first ATG codon, uATGs are expected to reduce translational efficiency by limiting the number of ribosomes reaching the CDS start codon[26]. The fraction of the ribosomes that do initiate translation at downstream ATGs do so by either a reinitiation or a context-dependent leaky scanning mechanism which are both relatively inefficient[27]. While uATGs in frame with the CDS introduce additional amino acids to the N-terminus of the translated peptide, uATGs that are out of frame shift the reading frame of the protein rendering it nonfunctional and thus should have a more profound effect. For our synthetic UTRs containing a uATG, we observe a minor effect on translation for uATGs in frame that is stronger when the uATG resides farther towards the 5' end of the UTR (Figure 4B left plot). This might represent the cumulative effect on translation, protein function and protein stability of adding a larger peptide chain to the N-terminus of His3. Consistent with previous findings, an in-frame stop codon occurring after an in or out of frame uATG, as well as out of frame uATGs all confer a significant decrease in growth rates (Figure 4B).

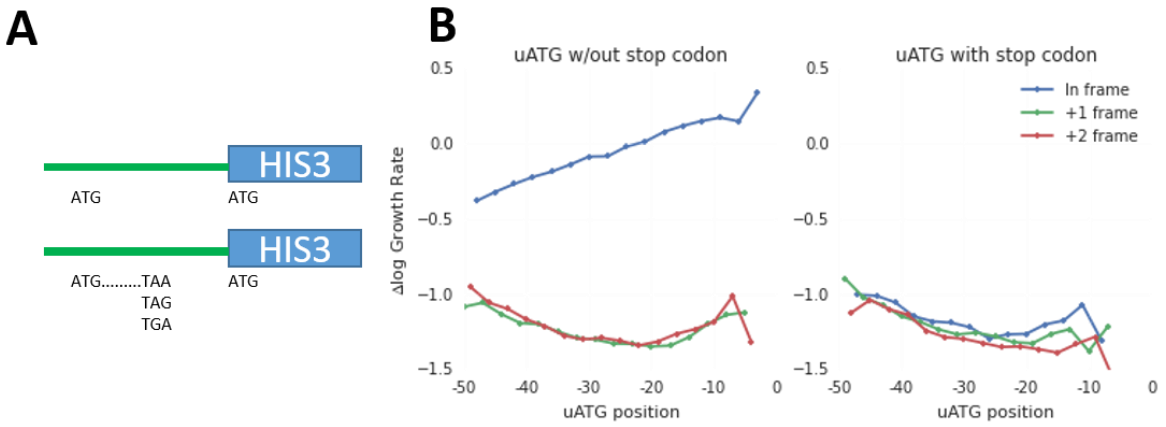


Figure 4. The effects of upstream ATGs
(A) Out of frame uATGs have a strong negative effect on growth rate, while the effect of in-frame uATGs is dependent on the number of additional peptides added to the N-terminus of *HIS3*
(B) All uATGs that introduce a uORF have a strong negative influence on growth rate, regardless of frame.

Effect of secondary structure on translation

A key question in the study of translation both in eukaryotes and prokaryotes is the impact of secondary structure on ribosome initiation, scanning, and elongation. First we looked at the correlation between the predicted minimum free energy (MFE) of the UTRs and the measured translational score. To calculate the predicted MFE for each UTR, we folded the UTR sequence with an additional 70nt in the coding region of *HIS3* using RNAfold[28,29]. When we binned the UTRs by MFE, we found that lower MFE bins corresponds to lower translation, with a general trend of increasing translation rates with higher MFE, but each bin displays significant variability (Figure 5A).

We then investigated the effect of hairpins of varying lengths on translation rates. Hairpins with 4-7 nt loops had a much stronger negative influence on translation than shorter or longer loops (Figure 4B). Longer loops may be more accessible to the natural helicase activity of the scanning ribosomes[30,31] and short loops may stabilize the hairpin structure through formation of tetraloops with exceptionally high thermodynamic stability[32-34]. In fact, when we examined the influence of the loop sequence in short hairpins (4nt stem, 4nt hairpin), we found that the top three inhibitory sequences to translation were exact matches to the previously described GNRA tetraloop motif[35]. We also found that secondary structure was the most detrimental to translation when it occurred near the 5' cap or near the start codon (Figure 5C). The negative influence of secondary structure seems to be roughly proportional to the log probability of a nucleotide being accessible. We speculate that this may be the case because the log probability of a nucleotide being free may be a good proxy for the local minimum free energy of the surrounding bases.

Although we found a correlation between structure and decreased translation, the MFE of the 5' UTR only accounts for a small fraction of the total observed variability in translation (R^2: 0.08).

While secondary structure does not seem to be a main determinant of translation, it is possible that the low correlation between MFE and translation may partially be a function of the poor accuracy of the RNA folding algorithms. Regardless, our results are consistent with previous results in eukaryotes, that indicate only very stable secondary structures (<30kcal/mol) markedly decreased

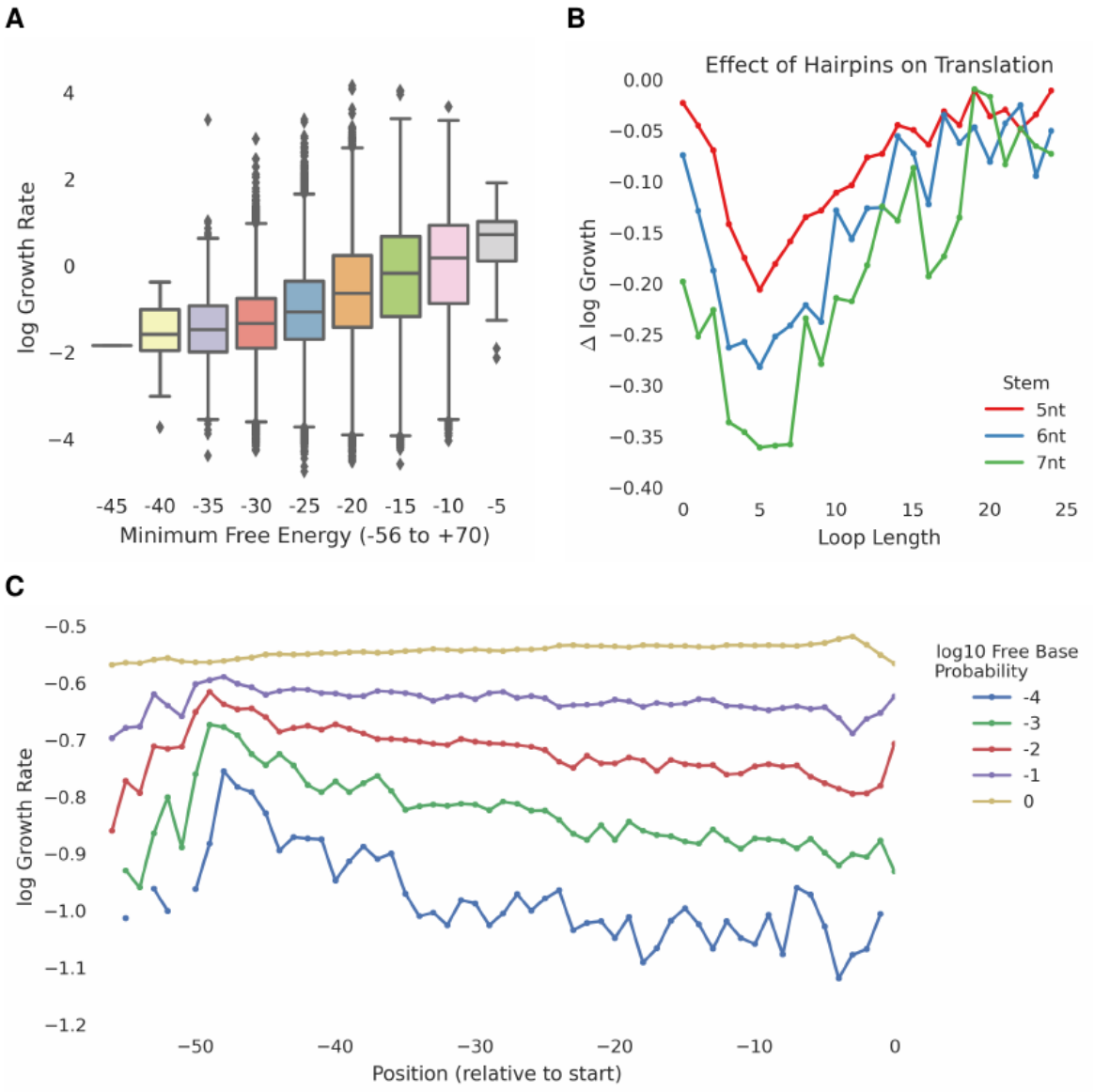


Figure 5. Impact of secondary structure on translation.
(A) The predicted minimum free energy of 5' UTR sequences correlates with growth rates
(B) Hairpins in the 5' UTR only have a strong inhibitory effect on translation if the loop is around 5nt.
(C) Bases with low probabilities of being free negatively influence translational efficiency, especially near the 5' cap and closer to the start codon.

translation rates[36]. The ribosomal initiation factor eIF4A is known to exhibit RNA helicase activity and likely can unwind less stable secondary structures, reducing their impact on translation[37].

Learning Predictive Models of 5' UTR Translational Efficiency with Linear Regression

We then asked whether we could learn a predictive model of translational efficiency based upon 5'UTR sequence. While we trained all our models on our library of ~500,000 randomized 5'UTRs, we wanted to show that they could accurately predict the translational efficiency of both random and native *S. cerevisiae* 5'UTRs. We started with simple models using linear regression, but then used "deep learning" to train non-linear models capable of learning more complex relationships between *cis*-regulatory motifs.

We first tested linear regression models with k-mers as input features. All models used cross-validation to choose optimal L2 regularization parameters. This first preliminary model included the counts of k-mers with disregard for the position. In order to get a sense for how different model types benefit from an increase in data points, we generated learning curves for each mode (Figure S2, left plot). We found that the best performance (performance metric: R^2) occurred with 6-mers. 7-mers led to overfitting due to the large number of parameters, while 5-mers and shorter led to underfitting and worse performance. However, even the 6-mer model could only account for 22% of the observed variability in growth rate. Given that certain positions within the 5'UTR are known to influence translation more strongly than others (adenosine in the -3 position), we speculated that a model that takes the position of k-mers into account might increase the performance of the model. To incorporate this, we then tested different k-mer models in which k-mers at each position within the 5'UTR sequence are assigned different weights (e.g. for a 3-mer model there are 64 possible k-mer sequences and 48 positions, leading to 3072 model weights; $3072=4^3 \times 48$). We found that

the 3-mer model with positional information outperformed all the other models, including the 6-mer model with no positional information (Figure S2, right plot).

We then proceeded to train a 3-mer model, with positional information, on the whole training set of random 5'UTRs. We first tested this model on a test set of randomized 5'UTRs. We found that the model explained 41% of the observed variability in translation (R2: 0.41, Figure 6A). We also found that the model could account for 40% of the observed variability in native 5'UTRs (R1: 0.40, Figure 6B). While the model explained a large fraction of the regulatory effects of both synthetic and native 5'UTRs, we asked whether we could improve upon these results by using more flexible models.

Improving predictions of protein production with deep learning

Linear regression models cannot model interactions between different sequence motifs that may occur through secondary structure or protein intermediates that bind the DNA sequence. Furthermore, the simple model is incapable of accurately modeling the effects of introducing uORFs. Clearly, the effect of a stop codon in the UTR is strongly dependent on whether an uATG exists in the same reading frame. To overcome these limitations of linear regression, we chose to model translation using a type of neural network common in computer vision—a convolutional neural network. Convolutional neural networks (CNNs) are able to incorporate nonlinearities that allow the model to represent these more complex relationships between sequence motifs. The lower layers of CNNs share properties with positional weight matrices (PWM)[38], the main difference being that these lower layers of CNNs can model nonlinearities (*e.g.* dependencies between specific nucleotides). Convolutional neural networks have been previously used to predict transcription factor binding, DNase hypersensitivity sites, enhancers, and DNA methylation [39-46].

When we trained a convolutional neural network on our data, we found that it substantially outperformed the linear regression models. We did not perform exhaustive grid search on network structure or hyperparameters, we simply trained one model (however we will use grid search in the future). Our model contained 64 filters (width: 9) in the first convolutional layer, 64 filters (width: 9) in the second convolutional layer, and 50 units in the following dense layer. The final linear layer produces real valued predictions. All layers used rectified linear units (ReLU) as nonlinearities, except the last linear layer. To prevent overfitting, we used dropout after the second convolutional layer and early stopping. The CNN was able to account for 58% and 61% of the

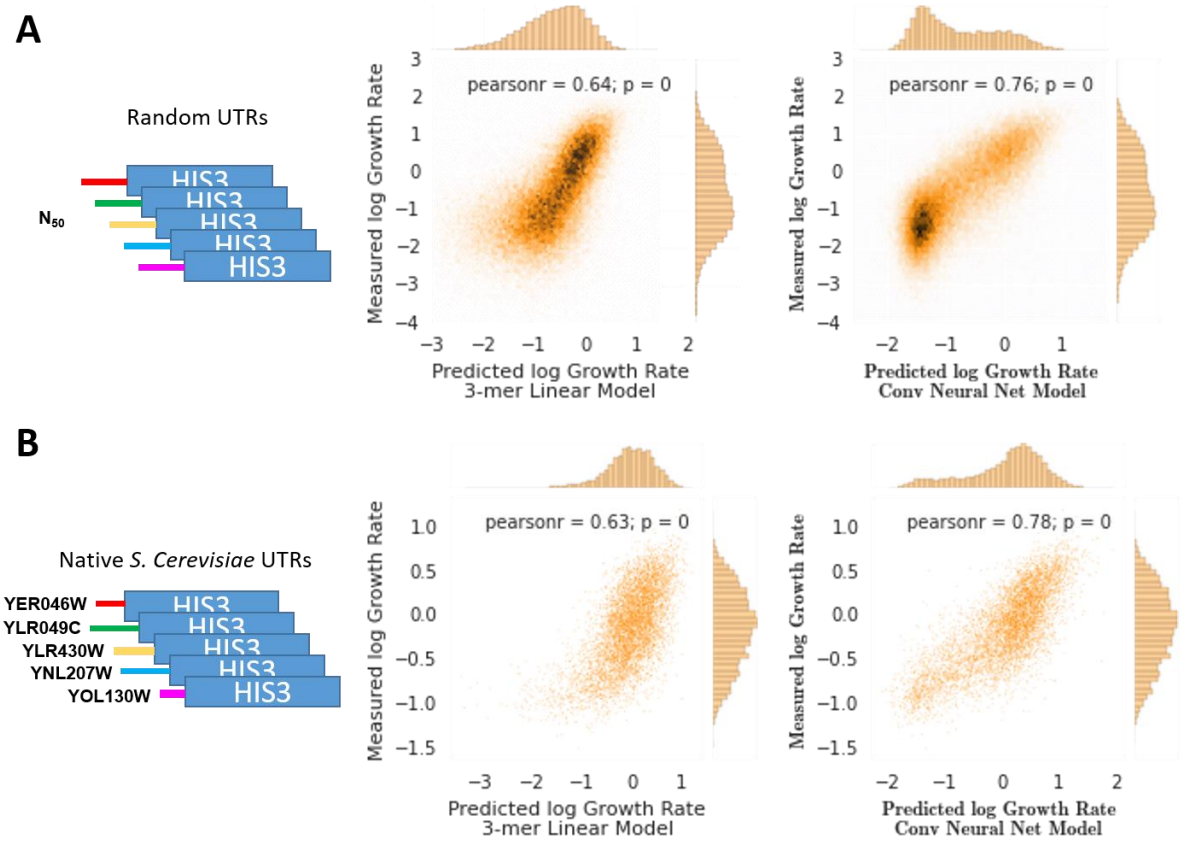


Figure 6. Deep learning predicts growth rates based on sequence more accurately than 3-mer linear regression
(A) We tested our models trained on random UTRs on the heldout test set of random UTRs. The performance of the CNN (right) is substantially better than the 3-mer linear regression model.
(B) When we tested the models on native UTRs, we again found that the CNN performed better.

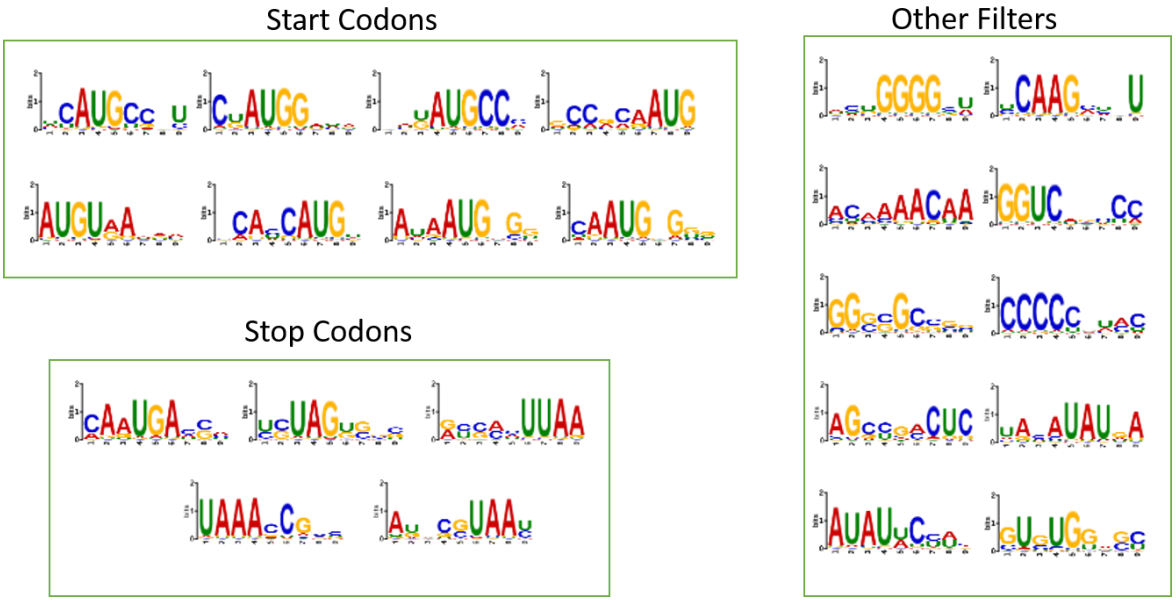


Figure 7. Filter motifs in the convolutional neural network
Many of the filters in the convolutional neural network converge to motifs with ATG or the three stop codons (TAA/TGA/TAG). Other features may be converging to RNA protein binding motifs or motifs likely to form secondary structures.

observed variability in the random UTRs and native UTRs respectively (Figure 6A/B, right plots). We then examined the motifs that the network included in the first layer of features. Many of the features included ATG, indicating that the model is using the features to learn about the impact of upstream ATGs (Figure 7). Five other filters learned motifs corresponding to stop codons. We suspect that many of the other motifs identified by the model correspond to the binding motifs of RNA binding proteins in yeast—unfortunately, the binding motifs of only 4 RNA binding proteins have been characterized in *S. cerevisiae.*

However, we were interested in the generality of these predictions. Specifically, if we changed the gene of interest to a fluorescent reporter, could the convolutional neural network model still make accurate predictions? In order to test this, we cloned 12 UTRs from our random library and 13 native yeast UTRs upstream of the fluorescent reporter Venus (Figure 8A). Upstream ATGs were absent from all 25 of the cloned UTRs. We then asked whether our models' predictions would

correlate with the measured levels of Venus. We found that CNN accounted for 57% of the variability in fluorescence (R2=0.57), comparable to the model's performance in the context of the HIS3 growth assay (R2=0.58, 0.61). This indicated that the model does in fact generalize to predicting translational efficiency of 5'UTRs regardless of the following coding sequence.

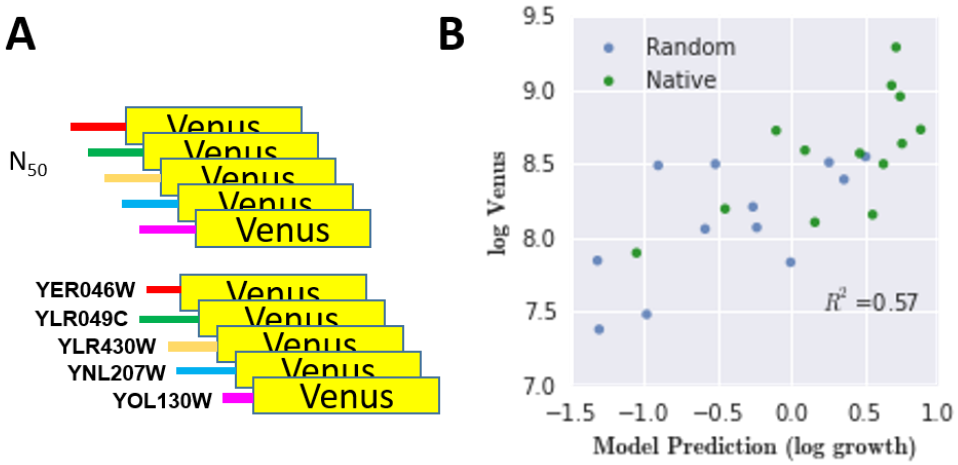


Figure 8. The convolutional neural network predictions generalize to different coding sequences
(A) In order to test whether our models were learning general rules of translation, we constructed plasmids with either random 50 nt 5'UTRs or native yeast 5'UTRs inserted upstream of a fluorescent Venus reporter.
(B) The predictions by the convolutional neural network showed high correspondence with the measured fluorescence.

Discussion

Machine learning algorithms provide tools for building accurate and predictive models, but learning high quality models requires large amounts of training data. For many biological phenomena such large training sets are not available. For example, the set of ~5000 native 5' UTRs in *S. cerevisiae* and associated ribosome profiling data form a relatively small data set on which to train a predictive model of the impact of 5' UTR sequence on translation. In this work, we overcame this issue by training models on very large libraries of synthetic expression constructs with millions of nucleotides of degenerate sequence.

We leveraged a massively parallel growth assay combined with machine learning to create a predictive model of translation in yeast. Our initial library contained ~500,000 fully degenerate

5'UTR sequences as well as approximately 11,000 50-nt fragments covering all native 5'UTR sequences from *S. cerevisiae*. We measured the fitness of each 5' UTR sequence in a competitive growth assay and then trained a set of models that could relate the 5' UTR genotype to this growth phenotype. Although all models were trained exclusively on the data from the library of fully degenerate sequences, they performed well on the task of predicting the growth rates associated with the native 5'UTR fragments.

A comparison between different modeling approaches revealed several interesting trade-offs. We first evaluated several linear regression models with short k-mer features and found that a model with position-dependent 3-mer features outperformed models with more complex but position-independent sequence features. The 3-mer model with position contains a total of 3072 (4^3 x 48) distinct features and is thus comparable in size to a 6-mer model without position (4^6=4096 features). However, given that many key features of translational efficiency in yeast have a clear position dependence—e.g. the identity of the nucleotide at position -3 or the frame of an upstream start codon—it is maybe not surprising that a model that captures such position dependence can outperform a model that does not even at the expense of using relatively simple features. Still, it is likely that a model using 4-, 5-, or 6-mer features together with position could outperform the 3-mer model used here, but training such models without overfitting would require even larger datasets than the one generated in this work.

To further improve the predictive power of our model, we next turned to CNNs which, unlike linear regression, can capture nonlinear interactions between sequences. When tested on the library of native 5'UTR fragments, the CNN approach indeed considerably improved our ability to predict translational efficiency from sequence. Moreover, although neural networks can be difficult to

interpret, it was possible to extract biological meaning from the model. The model learned features corresponding to start codons, stop codons, and even putative RNA binding protein binding motifs.

It is encouraging that when we swapped the *HIS3* coding sequence for a fluorescent reporter, our convolutional neural network performed almost equally as well. Clearly the ability of our model to generalize across different gene contexts will be important to its utility in metabolic engineering. A logical follow-up experiment would be to retest the library with a different promotor than *CYC1*. This may offer insights into the relationship between mRNA expression and translational efficiency in yeast.

It has even been suggested that protein expression in yeast is dominated the abundance of mRNA, and rather less by the efficiency with which that mRNA is translated[47,48]. One current shortcoming of our data is our inability to distinguish between these two processes. Though we have targeted translation by replacing the 5' UTR of the mRNA, since by its nature this region is juxtaposed to the promoter, we have possibly also influenced transcription of some of our mRNA species as well. By measuring the mRNA levels of the members of our library, we can either account for the transcriptional impact of each sequence and/or learn a model that takes into account both transcription and translation-based regulation.

In summary, we have developed predictive models for translation in yeast. Our approach can be used to predict the impact of sequence variants between related species and to tune the protein expression of synthetic constructs.

Supplementary Information

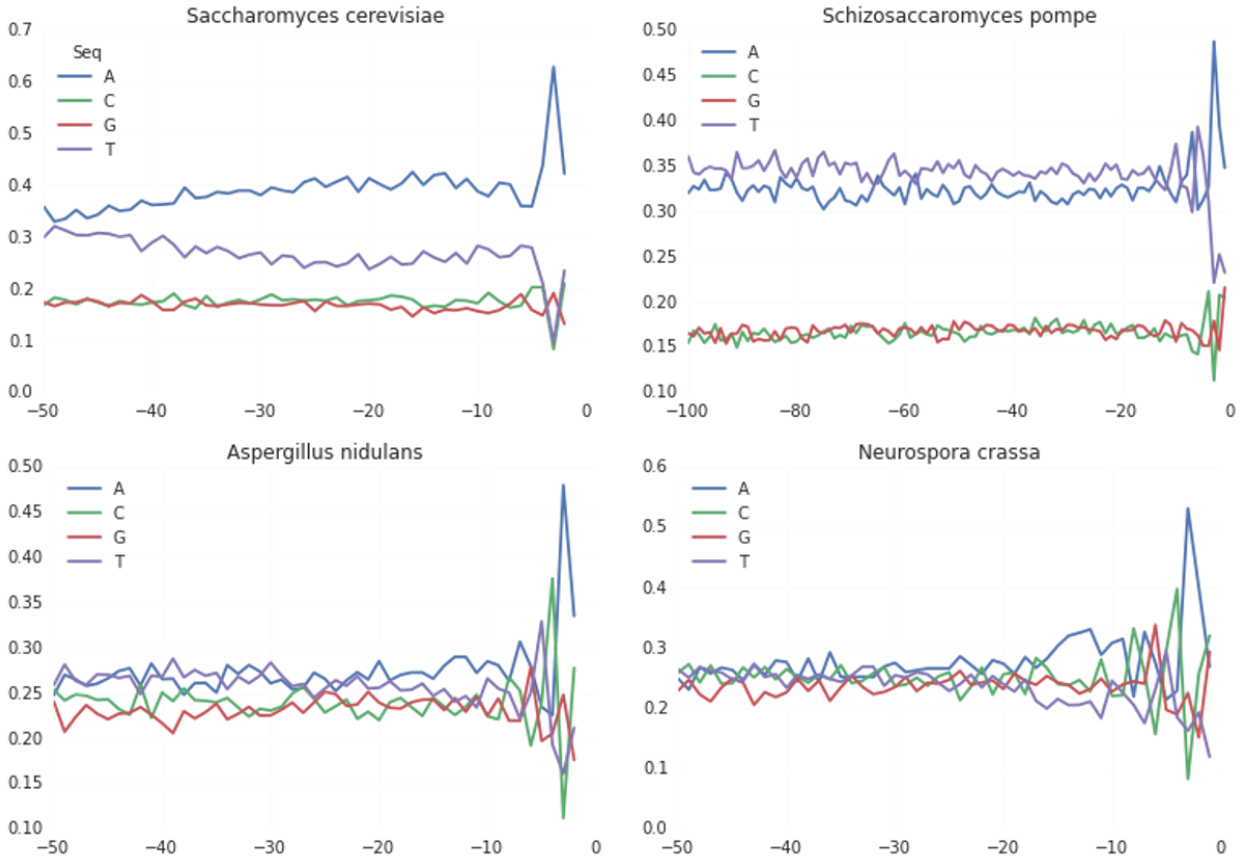


Figure S1. The nucleotide distributions in four different fungi
All fungi species exhibit a preference for an Adenosine at the -3 position, but *S. Cerevisiase* are unique in their strong preference for Adenosine throughout the UTR.

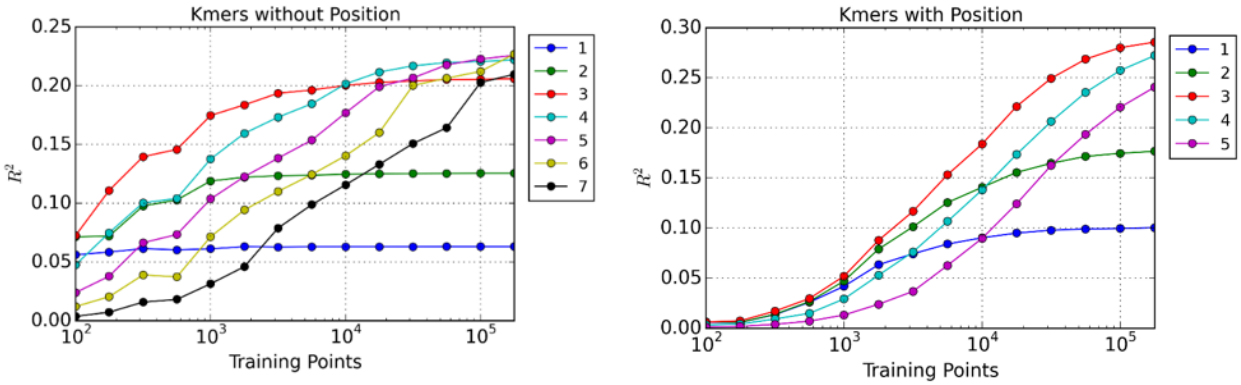


Figure S2. Learning curves for k-mer models
Left: Different learning curves for models using k-mers counts with no positional information
Right: Different learning curves for models using one-hot encoding of k-mers at each position

METHODS

Library construction

Synthetic 5'UTR library

We chose to replace a 56 bp CYC1 5'UTR fragment upstream of His3 ATG on a p415-pCyc1 plasmid with a library of 50 bp synthetic 5'UTR fragments. The synthetic 5'UTR fragments were constructed by annealing primers 126 and 127 containing an overlap region (ggacctttgcagca) and filling the sequence to double stranded using Klenow fragment (NEB). The resulting library fragment had a 50bp random base region and 60bp and 33bp 5' and 3' overlaps with the pCYC1 promoter and the HIS coding sequence, respectively, including the ATG start codon. We PCR-amplified the p415-pCyc1 plasmid fragment with primers 132 and 133 using KAPA Hi-Fi polymerase (Kapa Biosystems) and leaving out the ATG start codon. Moving the start codon into the library fragment served to prevent background plasmids not containing a library fragment from growing in –HIS selective medium. The final library (YTLR200) was assembled using Gibson assembly[49] and electroporated[50] into 40ul of 5-alpha electrocompetent E.coli (NEB).

Native 5'UTR library

For the native library, we constructed 11,855 sequences representing native 5'UTRs from the yeast genome[51] arranged in 50bp fragments with 25bp overlap if the UTR exceeded 50bp in length, and smaller fragments for UTRs shorter than 50bp. 20bp overhangs were added to both 5' (acattaggacctttgcagca) and 3' (ATGacagagcagaaagccct) ends of these sequences, again, overlapping the pCYC1 promoter and HIS gene on the p415-pCYC1 plasmid. The library

sequences were procured from CustomArray Inc. as a mixed oligo pool and amplified by qPCR using primers 126 and 142 in 15 cycles. The resulting fragment was used in the native library plasmid (YTLN200) assembly via Gibson reaction and electroporation as described above.

Yeast transformation

For the library transformation into yeast, we followed the electroporation protocol described previously[52]. For the generation of the large synthetic 5'UTR library (YTLR), we used an overnight culture of BY4741 diluted 1:50 into 50ml of YPAD media, grown to OD1.6, prepared 400ul of electroporation-competent cells as instructed by protocol and transformed it with a mixture of 3.66ug library plasmid YTLR200 linearized with EcoRI and 11.2ug of DNA fragment PCR-amplified from YTLR200 with primers 134 and 135 containing regions of overlap both upstream and downstream of EcoRI restriction site. We grew the transformed library in 500ml of synthetic dextrose media without leucine (SD-Leu) overnight and used colony counts from serial dilutions plated on SD-Leu to estimate library size. Using a longer fragment (2.3kb) resulted in improved transformation efficiency of ca. $2x10^6$. For the generation of the native 5'UTR library (YTLN), the same protocol was followed. 6.7ug of EcoRI-digested library plasmid YTLN200 and 15.55ug of PCR-amplified fragment (primers 134 and 135) were transformed into 800ul of electrocompetent BY4741 yeast cells with similar efficiency as YTLR library described above.

For the transformation of individual plasmids into yeast strains, we followed a lithium acetate method described previously[53].

Growth rates measurements

Yeast cultures were grown overnight at 30 °C until saturated. In 96-well plates, cultures were diluted 1:20 in 200 µl volume of minimal selective media. 96 well plates were continuously

shaken at 30° in media lacking histidine and leucine and with 1.5 mM 3-AT in a synergy H1 hybrid reader (Biotek). Mean (n=6) maximum doubling rate was determined by measuring the largest slope of O.D. 660 measurements over a 2-hour period +/- standard deviation.

Oligonucleotides and DNA sequencing

Oligonucleotides were obtained from Integrated DNA Technologies with standard desalting purification.

Primer name or number	**Sequence**
126	ACTCTTGTTTTCTTCTTTTCTCTAAATATTCTTTCCTTATACATTAGGACCTTTGCAGCA
127	TGTAATACGCTTTACTAGGGCTTTCTGCTCTGTCATNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTGCTGCAAAGGTCC
132	CTAGTACAGAGCAGAAAGCCCTAG
133	tgctgcaaaggtcctaatgtataag
134	cggcatcagagcagattgtac
135	ggtatttcacaccgcatatcgac
142	cttggtttcatttgtaatacgctttactagggctttctgctctgtcat

Sanger sequence and analysis was performed as previously described[54]. Deep sequencing of plasmid DNA was performed on an Illumina Nextseq by purifying DNA using the Zymoprep yeast plasmid prep II (Zymo Research), and PCR amplification for 12 to 20 cycles.

Library selection

Cells from the input population were collected for sequencing and for back dilution into the selection medium (SD–His–Leu + 1.5 mM 3-amino-1,2,4 triazole (Sigma)) in triplicate adding

108 million cells to 1 Liter medium. Each replicate was cultured to logarithmic phase 20 hours (O.D. A660 = 1.0, 6.1 billion cells), after which 300 million cells were collected for sequencing.

Strains and media

Yeast experiments used the BY4741 strain created by Giaever *et al.*[55]. Plasmid-based pCYC1-HIS3 were cloned into the pRS series[56] of yeast vectors with the *LEU2* nutrient marker (pRS415). To construct the plasmids harboring the individual synthetic and native 5'UTRs, we designed a set of one forward and two reverse primers each 30bp long with a 10bp overlap in the middle of the sequence for each sequence listed above. We added a 5' acattaggacctttgcagca overhang to the forward primer (overlapping pCYC1 promoter), and either agggctttctgctctgtcat 3' overhang (overlapping HIS3 gene) or attcttcacctttagacat 3' overhang (overlapping Venus gene) to the reverse primers. We obtained the oligos in a 96-well array (IDT), annealed them, filled in with Klenow fragment and cloned into either p415-pCYC1 backbone or p415-pCYC1-Venus backbone as described above. The p415-pCYC1-Venus plasmid was constructed by replacing the HIS3 sequence in the p415-pCYC1 plasmid used in our library construction with Venus via Gibson assembly.

Training the convolutional neural network:

All models were trained using the python packages Lasagne and Theano. The following network structure was used:

Layer 1: Convolutional, 64 filters (4x9), relu activation
Layer 2: 64 filters (1x9), relu activation
Layer 3: Dropout (0.25 probability of "dropping")
Layer 4: Fully connected layer, 50 hidden units, relu activation
Layer 5: Linear output layer, 1 output unit

The model was trained with the Adam optimizer[57] and early stopping was used to prevent overfitting to the training data.

References

1 Dvir, S. *et al.* Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 801, doi:10.1073/pnas.1222534110 (2013).

2 Lubliner, S. *et al.* Core promoter sequence in yeast is a major determinant of expression level. *Genome research* (2015).

3 Shalem, O. *et al.* Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control. *PLoS Genet* **11**, doi:10.1371/journal.pgen.1005147 (2015).

4 Hamilton, R., Watanabe, C. K. & Boer, H. A. Compilation and comparison of the sequence context around the AUG startcodons in Saccharomyces cerevisiae mRNAs. *Nucleic Acids Research* **15**, 3581-3593, doi:10.1093/nar/15.8.3581 (1987).

5 Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* (1986).

6 Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 14024-14029, doi:10.1073/pnas.1301301110 (2013).

7 Brennan, M. B. & Struhl, K. Mechanisms of increasing expression of a yeast gene in Escherichia coli. *Journal of molecular biology* **136**, 333-338 (1980).

8 Lamping, E., Niimi, M. & Cannon, R. D. Small, synthetic, GC-rich mRNA stem-loop modules 5' proximal to the AUG start-codon predictably tune gene expression in yeast. *Microbial cell factories* **12**, 74, doi:10.1186/1475-2859-12-74 (2013).

9 Chen, J., Ding, M. & Pederson, D. S. Binding of TFIID to the CYC1 TATA boxes in yeast occurs independently of upstream activating sequences. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 11909-11913 (1994).

10 Guo, Z., Russo, P., Yun, D. F., Butler, J. S. & Sherman, F. Redundant 3' end-forming signals for the yeast CYC1 mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 4211-4214 (1995).

11 Martens, C., Krett, B. & Laybourn, P. J. RNA polymerase II and TBP occupy the repressed CYC1 promoter. *Molecular microbiology* **40**, 1009-1019 (2001).

12 Watanabe, K., Yabe, M., Kasahara, K. & Kokubo, T. A Random Screen Using a Novel Reporter Assay System Reveals a Set of Sequences That Are Preferred as the TATA or TATA-Like Elements in the CYC1 Promoter of Saccharomyces cerevisiae. *PloS one* **10**, doi:10.1371/journal.pone.0129357 (2015).

13 Yagil, G., Shimron, F. & Tal, M. DNA unwinding in the CYC1 and DED1 yeast promoters. *Gene* **225**, 153-162 (1998).

14 Pfeifer, K., Arcangioli, B. & Guarente, L. Yeast HAP1 activator competes with the factor RC2 for binding to the upstream activation site UAS1 of the CYC1 gene. *Cell* **49**, 9-18 (1987).

15 Olesen, J., Hahn, S. & Guarente, L. Yeast HAP2 and HAP3 activators both bind to the CYC1 upstream activation site, UAS2, in an interdependent manner. *Cell* (1987).

16 Treitel, M. A., Kuchin, S. & Carlson, M. Snf1 protein kinase regulates phosphorylation of the Mig1 repressor in Saccharomyces cerevisiae. *Molecular and cellular biology* **18**, 6273-6280 (1998).

17 Looman, A. C. & Kuivenhoven, J. Influence of the three nucleotides upstream of the initiation codon on expression of the Escherichia coli lacz gene in Saccharomyces Cerevisiae. *Nucleic Acids Research* **21**, 4268-4271, doi:10.1093/nar/21.18.4268 (1993).

18 Baim, S. B. & Sherman, F. mRNA structures influencing translation in the yeast Saccharomyces cerevisiae. *Molecular and Cellular Biology* **8**, 1591-1601, doi:10.1128/MCB.8.4.1591 (1988).
19 Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H. & Miura, K. i. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Research* **36**, 861-871, doi:10.1093/nar/gkm1102 (2007).
20 Cavener, D. R. & Ray, S. C. Eukaryotic start and stop translation sites. *Nucleic Acids Research* **19**, 3185-3192, doi:10.1093/nar/19.12.3185 (1991).
21 Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Molecular systems biology* **10**, 770 (2014).
22 Lin, Z. & Li, W.-H. H. Evolution of 5' untranslated region length and gene expression reprogramming in yeasts. *Molecular biology and evolution* **29**, 81-89, doi:10.1093/molbev/msr143 (2012).
23 Wang, X.-Q. Q. & Rothnagel, J. A. 5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic acids research* **32**, 1382-1391, doi:10.1093/nar/gkh305 (2004).
24 Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Molecular and cellular biology*, doi:10.1128/MCB.20.23.8635-8642.2000 (2000).
25 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)* **324**, 218-223, doi:10.1126/science.1168978 (2009).
26 Meijer, H. A. & Thomas, A. A. M. Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. *Biochemical Journal* (2002).
27 Kozak, M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene* (2002).
28 Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA Websuite. *Nucleic Acids Research* **36**, doi:10.1093/nar/gkn188 (2008).
29 Lorenz, R. & Bernhart, S. H. ViennaRNA Package 2.0. *Algorithms for …* (2011).
30 Rogers, G. W. Modulation of the Helicase Activity of eIF4A by eIF4B, eIF4H, and eIF4F. *Journal of Biological Chemistry* **276**, 30914-30922, doi:10.1074/jbc.M100157200 (2001).
31 Rozen, F. *et al.* Bidirectional RNA helicase activity of eucaryotic translation initiation factors 4A and 4F. *Molecular and Cellular Biology* **10**, 1134-1144, doi:10.1128/MCB.10.3.1134 (1990).
32 Antao, V. P. & Tinoco, I. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic acids research* **20**, 819-824 (1992).
33 Wolters, J. The nature of preferred hairpin structures in 16S-like rRNA variable regions. *Nucleic acids research* **20**, 1843-1850 (1992).
34 Sheehy, J. P., Davis, A. R. & Znosko, B. M. Thermodynamic characterization of naturally occurring RNA tetraloops. *RNA (New York, N.Y.)* **16**, 417-429, doi:10.1261/rna.1773110 (2010).
35 Woese, C. R., Winker, S. & Gutell, R. R. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proceedings of the National Academy of Sciences of the United States of America* **87**, 8467-8471 (1990).

36 Babendure, J. R., Babendure, J. L., Ding, J.-H. H. & Tsien, R. Y. Control of mammalian translation by mRNA structure near caps. *RNA (New York, N.Y.)* **12**, 851-861, doi:10.1261/rna.2309906 (2006).
37 Svitkin, Y. V. *et al.* The requirement for eukaryotic initiation factor 4A (elF4A) in translation is in direct proportion to the degree of mRNA 5' secondary structure. *RNA (New York, N.Y.)* **7**, 382-394 (2001).
38 Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic acids research* **10**, 2997-3011 (1982).
39 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, doi:10.1038/nbt.3300 (2015).
40 Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods* **12**, 931-934, doi:10.1038/nmeth.3547 (2015).
41 Kelley, D. R., Snoek, J. & Rinn, J. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* **0**, 0, doi:10.1101/gr.200535.115 (2016).
42 Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *bioRxiv*, 32821 (2015).
43 Lanchantin, J., Singh, R., Lin, Z. & Qi, Y. Deep motif: Visualizing genomic sequence classifications. *arXiv preprint arXiv:1605.01133* (2016).
44 Liu, F., Li, H., Ren, C., Bo, X. & Shu, W. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *bioRxiv* (2016).
45 Kleftogiannis, D., Kalnis, P. & Bajic, V. B. DEEP: a general computational framework for predicting enhancers. *Nucleic acids research*, doi:10.1093/nar/gku1058 (2015).
46 Wang, Y. *et al.* Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. *Scientific Reports*, doi:10.1038/srep19598 (2016).
47 Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS genetics* **11**, doi:10.1371/journal.pgen.1005206 (2015).
48 Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports* **14**, 1787-1799, doi:10.1016/j.celrep.2016.01.043 (2016).
49 Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6**, 343-345, doi:10.1038/nmeth.1318 (2009).
50 Dower, W. J., Miller, J. F. & Ragsdale, C. W. High efficiency transformation of E.coli by high voltage electroporation. *Nucleic Acids Research* **16**, 6127-6145, doi:10.1093/nar/16.13.6127 (1988).
51 Park, D., Morris, A. R., Battenhouse, A. & Iyer, V. R. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic acids research* **42**, 3736-3749, doi:10.1093/nar/gkt1366 (2014).

52 Gietz, D., Jean, S. A. & Woods, R. A. Improved method for high efficiency transformation of intact yeast cells. *Nucleic acids research* (1992).
53 Gietz, R. D. & Woods, R. A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods in enzymology* **350**, 87-96 (2002).
54 Sanger, F. & Nicklen, S. DNA sequencing with chain-terminating inhibitors. *Proceedings of the …* (1977).
55 Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *nature* **418**, 387-391, doi:10.1038/nature00935 (2002).
56 Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* (1989).
57 Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).