Title
Learning Quantitative Sequence-Function Relationships using Massively Parallel Reporter Assays

Author
Insigne, Kimberly Danielle

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Learning Quantitative Sequence-Function Relationships using Massively Parallel Reporter Assays

A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy in Bioinformatics

by

Kimberly Danielle Insigne

2019

ABSTRACT OF THE DISSERTATION

Learning Quantitative Sequence-Function Relationships using

Massively Parallel Reporter Assays

by

Kimberly Danielle Insigne

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2019

Professor Sriram Kosuri, Chair

The field of genomics has grown rapidly over the past decade due to the advent of high-throughput sequencing technologies. Genomics relies on this wealth of information to draw biological inferences, but using inference to establish causality can be challenging as many genetic factors correlate with one another. Due to the declining cost of both reading and writing DNA, new techniques known as massively parallel reporter assays (MPRAs) provide the ability to test the function of a large library of tens to hundreds of thousands of designed DNA sequences simultaneously in a single experiment. Testing designed libraries allows us to explore beyond natural sequence variation to directly test thousands of sequence-function hypotheses simultaneously. In this dissertation I discuss two projects that explore sequence-function relationships in different biological systems.

The first project is focused on how human genetic variation affects exon recognition, as mis-splicing is a major mechanism through which variants exert their influence. We developed a Multiplexed Functional Assay of Splicing using Sort-seq (MFASS) and assayed 27,333 variants

in the Exome Aggregation Consortium within or adjacent to 2,198 human exons. We found that 3.8% (1,050) led to large splicing disruptions, many of which are extremely rare, located outside of canonical splice sites, distributed evenly across intronic and exonic regions, and are difficult to predict. MFASS enables direct functional measurement of large-effect splicing defects at scale.

The second project is focused on promoters and transcriptional regulation in *Escherichia coli*. Promoter sequence space in bacteria is vast and difficult to study genome-wide due to external factors that influence transcription. We developed a genomically-encoded MPRA to characterize the global promoter landscape and dissect active promoters for regulatory motifs. We measure promoter activity of over 300,000 sequences spanning the entire genome and identify 3,321 active promoter regions in glucose minimal media and 3,477 in rich LB media. Furthermore, we perform a scanning mutagenesis of 2,057 *E. coli* promoters to identify regulatory sequences. Lastly, we implement a variety of machine learning models to classify promoters and quantitatively predict their activity. We present a series of approaches to rapidly characterize promoter sequences within the *E. coli* genome.

The dissertation of Kimberly Danielle Insigne is approved.

Jason Ernst

Leonid Kruglyak

Xinshu Grace Xiao

Sriram Kosuri, Committee Chair

University of California, Los Angeles

2019

To my family, the boyfriend, and our dog.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Sriram Kosuri, for his excellent mentorship throughout my graduate career. None of the work presented here would be possible without his invaluable support and guidance. I could not have asked for a better mentor and feel very lucky that he accepted me as a student. He taught me how to become a true scientist by setting a constant example for critical thinking and tireless pursuit of the truth. He pushed me when I most needed it because he always believed in me and gave me confidence when I needed him most, especially when transitioning out of academia into the field of data science. He has shaped me into the person I am today and I am forever grateful.

Furthermore, I would like to thank Professors Jason Ernst, Leonid Kruglyak, and Xinshu Grace Xiao for serving as my committee members. Their continual guidance, suggestion, and comments are greatly appreciated.

Other current and past members of the Kosuri lab have played a pivotal role during my PhD career. I would like to thank Rocky Cheung, a postdoc, for his mentorship during both my rotation and as co-first author on one of my main projects in the lab. The project would have been impossible without his experimental expertise, bioinformatics skills, and Adobe Illustrator mastery. Also, David Yao made significant experimental contributions to the splicing work, and Christina Burghard on the computational side. Next, I would like to thank Guillaume Urtecho, a graduate student in the lab, and co-first author on the *E. coli* work. He was one of my first friends in graduate school and we have worked on this project together since we first joined the lab. It has been such a pleasure to work with him and we have become close friends in the process. This would not be possible without his synthetic biology expertise, his willingness to whip up an analysis in search of a biological hunch, and excellent writing skills. Two undergraduates, Arielle Tripp and Marcia Brinck, made substantial experimental contributions to

this work. I would also like to thank the current and former postdocs, students, and members of the Kosuri Lab for the countless hours of scientific insight and socializing, which kept me sane: Dr. Hwangbeom Kim, Dr. Calin Plesa, Jessica Davis, Clifford Boldridge, Angus Sidore, Jeff Wang, Joyce Samson, and Johnny Lee.

I would like to thank Professor Mete Civelek at the University of Virginia, a former postdoc in the Lusis Lab at UCLA, who served as my mentor during a summer of research after my junior year of undergraduate. He was my first truly great mentor and he helped me develop the first research project I felt I could call my own. Under his guidance, I learned the R programming language and executed a project from start to finish that I felt proud of. At his encouragement, I applied for graduate school that coming fall. Without his constant support and inspiration, I may never have pursued a PhD in bioinformatics. I cannot say thank you enough for giving me that crucial initial push.

I would like to thank my parents and my sister for their support. Thank you for thinking I was smart enough when I did not always believe it. None of this would be possible without you.

I would like to thank the friends and colleagues I met here at UCLA. I am lucky to be surrounded by such intelligent people, and every person has impacted my life positively. In particular, I would like to thank my good friends, R.W., C.L., J.S., A.C, T.G.

Finally, I would like to thank my partner C.B. for his endless support. You have been there from the first months when all I did was read papers, to years later when I was interviewing for my first job. Thank you for always listening to my frustrations, feelings of imposter syndrome, and existential crises. Thank you for all the times you took out the dog, prepared food, and helped me de-stress. Thank you for standing by my side at the frontlines of this journey.

VITA

Education

B.S. Bioengineering: Bioinformatics (cum laude), University of California San Diego

June 2014

Research Experience

Graduate Student Researcher, Kosuri Lab, UCLA

June 2015 – August 2019

Undergraduate Research Assistant, Lusis Lab, UCLA

June 2013 – March 2014

Undergraduate Research Assistant, McCammon Lab, UCSD

October 2012 – June 2013

Undergraduate Research Assistant, Abagyan Lab, UCSD

June 2012 – August 2012

Memberships

University of California Leadership Excellence Through Advanced Degrees (UC LEADS)

April 2011 – June 2014

PUBLICATIONS

Comprehensive functional characterization of *Escherichia coli* promoters reveals key components of transcriptional regulation. Guillaume Urtecho*, **Kimberly D. Insigne***, Arielle D. Tripp, Marcia Brinck, Nathan B. Lubock, Hwangbeom Kim, Tracey Chan, Sriram Kosuri. In preparation.

A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. Rocky Cheung*, **Kimberly D. Insigne***, David Yao, Christina P. Burghard, Jeff Wang, Yun-Hua E. Hsiao, Eric M. Jones, Daniel B. Goodman, Xinshu Xiao, Sriram Kosuri. January 3 2019 Molecular Cell (2018) doi : https://doi.org/10.1016/j.molcel.2018.10.037

Systematic dissection of sequence elements controlling σ70 promoters using a genomically-encoded multiplexed reporter assay in E. coli. Guillaume Urtecho, Arielle D. Tripp, **Kimberly D. Insigne**, Hwangbeom Kim, Sriram Kosuri. February 1 2018 Biochemistry, 2019, 58 (11), pp 1539–1551 doi: 10.1021/acs.biochem.7b01069

CHAPTER ONE

Introduction

Sequence-function relationships are key to understanding many diverse biological problems

An organism's genome encodes all the necessary functions of life within its sequence, including the ability to regulate the expression of tens of thousands of genes in a finely-tuned and complex manner. One of the major goals in biology is to precisely understand how these sequences encode function. Sequence-function relationships arise in many different contexts – single nucleotide variants identified from genome-wide association studies and their influence on a phenotype of interest, variants in non-coding regions and splice sites and their effect on splicing, and more broadly sequence variants in *cis*-regulatory elements (CREs) and their impact on gene regulation. A deeper and more quantitative understanding of sequence-function relationships would not only advance genomics in general, but potentially lead to algorithms that more accurately predict effects of sequence variants on gene expression, new mechanisms and targets for therapeutics, improved variant interpretation in a clinical setting, and a more precise and finely-tuned way to engineer biology.

Learning sequence-function relationships from high-throughput genomics datasets remains challenging

Prior to the advent of high-throughput techniques, sequence-function relationships were commonly probed by testing a limited number of sequence variants[1,2] or by "knocking out" sequences of interest and studying their effect on a biological function of interest. Indeed, many consensus CREs (TFs, splice sites, promoter motifs, etc.) and fundamental principles of gene expression and regulation were discovered in this manner. The decreasing cost of high-throughput sequencing technologies has made it routine to study many different facets of gene expression and regulation at a genome-wide scale and across many conditions. Armed with this massive amount of information, it is relatively simple to discover genetic variants associated with perturbed function, yet we are limited by the ability to interpret these variants. There is a

proliferation of computational approaches that attempt to learn genotype-phenotype relationships from these datasets, but accurate and quantitative predictions remain elusive. In addition, conventional attempts to biologically validate predictions requires laborious, low-throughput assays, limiting our ability to effectively improve prediction algorithms.

Generally speaking, genomics attempts to solve the inverse problem of understanding biological mechanism from observations, a type of problem that is very difficult if not impossible to solve[3]. Genomics relies on a wealth of information to draw biological inferences, but using inference to establish causality can be challenging as many genetic factors correlate with one another. To add another layer of complexity, learning on natural sequence space may confound prediction since it is typically evolutionarily constrained and very sparse compared to the vast potential sequence space. Finally, learning sequence-function relationships may be limited by the nature of the data - it remains unclear how well indirect measures of *cis*-regulatory function predict actual function.

Massively parallel reporter assays enable functional testing of a large library of sequences in a single experiment

Due to the declining cost of both reading and writing DNA, a recently developed class of techniques known as massively parallel reporter assays (MPRAs) has emerged, enabling functional testing of tens to hundreds of thousands of sequences simultaneously in a single experiment. Prior to MPRAs, there existed effective methods for studying *cis*-regulatory function such as classic saturation mutagenesis[4] and combinatorial promoter shuffling[5], but they had only been applied at low-throughput. MPRAs are superior to traditional gene reporter assays because they utilize programmable microarrays[6] and next-generation sequencing to synthesize and quantify large libraries of sequences of interest, respectively. Each MPRA is tailored to the question of interest and comes in many varieties, but each follows the same basic framework:

creation of a large variant library of interest, delivery into an organism of interest, a functional assay, sequencing to quantify variant levels, and calculation of functional scores for each variant.

There are two different approaches commonly used to quantify reporter gene activity. In the first approach, each variant is identified by a short barcode sequence which is placed in the 3' UTR and is co-transcribed with the reporter gene. RNA-seq is used to quantify the levels of the co-transcribed barcodes, measuring the activity of thousands of variants simultaneously. An important advantage to this approach is that each variant has multiple barcodes and therefore multiple replicate measurements are taken. The second approach requires a fluorescent reporter gene and sorts a population of cells using flow cytometry, avoiding the need for a co-transcribed barcode. DNA sequencing is performed on each bin and variant expression is calculated based on its distribution across bins. This method is a discrete measurement of expression compared to the continuous RNA-seq readout, but there is typically a sufficient number of sequences to train quantitative sequence-function models.

The MPRA approach was first demonstrated by Patwardhan et al.[7], who used programmable microarrays to synthesize barcoded oligonucleotides containing all possible single-nucleotide mutations in three bacteriophage promoters and three mammalian core promoters in a single experiment per promoter. The library was transcribed *in vitro* to measure activity, although the technique was subsequently adapted for living cells by cloning the library into a plasmid backbone[8]. Many subsequent MPRA approaches choose to focus heavily on only a few sequences of interest, but with great statistical power, measuring the effects of all possible single nucleotide substitutions in a regulatory element (high-throughput saturation mutagenesis)[7,9]. These types of synthetic saturation mutagenesis approaches have proven powerful in developing sequence-function relationships for a sequence of interest, recapitulating known motifs and enabling *de novo* motif discovery.

Massively parallel reporter assays utilizing designed sequence libraries enables testing of thousands of hypotheses simultaneously

In contrast to approaches that leverage the high-throughput nature of MPRAs to deeply dissect a few sequences of interest, others use this capability to test many thousands of regulatory elements at once. A recent study tested over 2,000 candidate enhancers in two human cell lines and synthesized not only the wild-type sequences but also engineered variants that removed, disrupted, or improved predicted causal regulatory motifs of five activators and two repressors[10]. While the first type of MPRA approach discussed is a powerful exploratory tool that can lead to quantitative sequence-function relationships for a few sequences of interest, approaches that utilize designed sequence libraries have the distinct advantage of directly testing thousands of hypotheses simultaneously. A multitude of mechanistic hypotheses can be directly tested and can focus more broadly on how this mechanism operates on a global scale. Furthermore, one can incorporate existing knowledge and design libraries of sequences to differentiate between competing models. MPRAs offer models of sequence-function relationships a unique ability to not only learn on large biological datasets, but to quickly iterate and improve with each successive library design. Each experiment can further refine existing predictive models, act as test sets for previous models, and guide model selection and inform the next iteration of experiments. Here, I propose an outline to leverage these techniques to test existing hypotheses, design the next set of optimal experiments based on previous iterations, and learn better quantitative models of sequence-function relationships in two model systems in gene expression.

Human splicing is an ideal system to study complex and clinically relevant sequence-function relationships

My first project detailed in Chapter 2 is focused on quantifying the effects of sequence variants on exon skipping in the human genome. Towards a deeper understanding of the splicing code, we designed an initial library containing all human exons < 100 bp, as well as designed sequence variants predicted to impact exon skipping.

An increasing proportion of human diseases are associated with aberrant splicing, underscoring the clinical importance of interpreting sequence variants and their potential impact on splicing. Many of the core consensus sequences of splicing are known, but there are many other regulatory elements and *trans*-acting factors that can modulate splicing, making quantitative and predictive models of splicing difficult to achieve.

The splicing code is complex and degenerate

Human protein diversity is primarily due to alternative splicing, a common mechanism that occurs in more than 90% of protein-coding genes to give rise to multiple mRNA isoforms[11]. There are several common types of alternative splicing including cassette exon skipping (the most prevalent type[11]), mutually exclusive exons, alternative 5' splice site, alternative 3' splice site, and intron retention[12]. The most prevalent types of mutations that affect splicing occur in core consensus sequences – the 5' splice site, branch-point adenosine, and the 3' splice site – or *cis*-acting regulatory sequences that occur in exons or introns to enhance (ESEs/ISEs, exonic/intronic splicing enhancers) or suppress (ESSs/ISSs, exonic/intronic splicing silencers) splicing[13]. The sequences that demarcate exon-intron boundaries are short, degenerate signals of varying strength that are necessary but not sufficient for splicing and occur with high frequency in the genome[14]. Additionally, the same splicing regulatory element can have varying and even opposing effects depending on sequence context or cell-type identity, allowing finely-tuned tissue-dependent regulation of splicing.

Splicing is difficult to quantitatively predict

High-throughput RNA sequencing has enabled transcriptomic profiling in many tissues and disease states, providing a basis to deciphering the splicing code. Previous work developed a splicing code, known as SPANR (Splicing-based Analysis of Variants), based on combinations of hundreds of hand-crafted RNA features trained on thousands of exon-skipping events in the

human genome, that predicted tissue-dependent splicing changes from sequence alone for the first time[15]. This work was subsequently expanded to incorporate hundreds of new features and implemented a deep neural network to effectively learn a comprehensive splicing code[16]. Although the code predicts quantitative and tissue-dependent splicing changes, it performs best at predicting only the direction of change and can quantify effects for single nucleotide variants only. Recent work based only on conservation sequence scores and locations of splice sites outperformed state-of-the-art models using hand-crafted biological features, demonstrating that conservation is an unexpectedly powerful indicator of alternative splicing patterns[17]. However, this model cannot predict tissue-differential levels of splicing or the effects of mutations on splicing as the model inherently does not have access to the sequence information. Circumventing a genomic approach, Rosenberg et al[18] learned a splicing model based entirely on millions of synthetic mini-genes containing degenerate regions representing alternative 5' and 3' splicing events. Their model converts input sequences into hexamer features and learns individual effect scores (hexamer additive linear model, HAL), which are used to predict splice site usage between two different sites. Although their model was empirically trained on synthetic libraries from an MPRA, it can predict the quantitative effect of variants on exon skipping in Mendelian diseases. Their model can generate predictions for any combination of exonic variants and is not limited to single nucleotide changes, but it cannot interpret intronic variants, and performs best when only considering direction of change instead of absolute values. This study demonstrates the power of using MPRAs to explore vastly more splicing events than those that naturally occur in the human genome, and further insights are possible with a more designed and focused approach.

Initial library design

Our initial splicing library includes all human exons < 100bp (8.5% of all human exons) flanked with 50bp of surrounding intronic sequence on each side (current OLS technologies are limited to < 200bp). Each natural sequence has 60-80 mutated versions which test different combinations

of splice site strengths, exonic/intronic enhancers/suppressors (ESEs/ISEs/ESSs/ISSs), SNPs, and synonymous codons. Additionally, we included 96 previously characterized splicing efficiency constructs that span a range of splicing efficiencies to give a final library size of 17,290 oligos. There is evidence that alternative exons are associated with longer introns and that intron length may influence exon skipping levels[39]. In order to test our design in a more natural context without artificially short introns, we cloned our exon library into two different constant intron backgrounds, DHFR, (300bp of 5' intron and 700bp of 3' intron) and SMNI (400bp of 5' intron and 500bp of 3' intron), which more closely resemble the average human intron length (~2000bp). However, the original library with short flanking native sequence is still biologically relevant as recent work suggests much of the regulatory information is captured within the first 100bp of flanking intronic sequence[38].

Massively parallel reporter assay to quantify exon skipping in human cells

The library is cloned into a split GFP reporter with a downstream constitutive RFP reporter gene in HEK-2943 cells (**Figure 2.1A**). We use flow cytometry to measure the GFP/RFP ratio to quantify the level of exon skipping (**Figure 2.1B**). If exon skipping occurs, the GFP will be reconstituted, resulting in a high GFP/RFP ratio. If exon inclusion occurs, the GFP will remain split and only the constitutive RFP will be expressed, leading to a low GFP/RFP ratio. Intermediate levels of exon skipping are easily quantified with the continuous readout of GFP/RFP ratio. The population of cells is sorted using fluorescence activated cell sorting (FACS) and each bin is sequenced, giving the quantitative variant levels in bins of different GFP/RFP intensity.

Studying global promoter architecture in *E. coli* is an ideal system to quickly iterate through experimental designs

My second project detailed in Chapter 3 is focused on global gene regulation in *E. coli*. This long-studied model organism is an ideal system to test and develop models of sequence-function relationships because it is relatively fast to test designed libraries, enabling relatively rapid improvements and refinements to models. Prokaryotic gene regulation is a model system that has been extensively studied for decades, and many of the basic mechanistic details are well understood. Promoters are the principal drivers in gene regulatory networks and although much is known about prokaryotic transcription, we are still unable to accurately predict the level of gene expression from promoter sequence alone.

E. coli promoters are degenerate, modular, and difficult to predict

Initiation of RNA-transcript formation is a key regulation point in transcriptional control. Transcript initiation requires the interaction of the RNA polymerase (RNAP) with the promoter DNA, which is mediated by a σ factor to form an active holoenzyme. The σ factor ensures promoter specificity, correct positioning of the polymerase at target promoters, and facilitates unwinding of the DNA near the TSS[19]. Most bacteria contain multiple σ factors, all of which share common features, that allow regulation of basal gene expression as well as regulation in response to altered environmental conditions[20]. The main step in initiation is promoter recognition, which is facilitated by four different sequence elements, each of which bind to different subunits of the σ factor. The two primary elements are the -35 and -10 elements, two hexamer elements with known consensus sequences which are located 35 and 10 bp upstream of the TSS, respectively[21]. The two other important elements are the extended -10 element, a 3-4bp motif immediately upstream of the -10 element[22], and the UP element, a ~20bp sequence located upstream of the -35 element[23]. These four elements together specify the initial binding of RNAP to the promoter, but the relative contribution of each element differs from promoter to promoter. The primarily role of

these elements seems to be docking the polymerase to the promoter DNA for subsequent open complex formation, therefore deficiencies in one element can be compensated for by another. Indeed, there is no naturally occurring promoter which has all four consensus elements present, as this promoter would bind the polymerase too strongly and inhibit transcription. The degeneracy and modularity of these core promoter elements complicate efforts to develop universal rules governing promoter activity.

Previous work predicted the strength of full-length *E. coli* promoters, a set of 60 promoters dependent on the alternative factor, using an UP-element contribution score in combination with a PWM-based core promoter model[24,25]. The model was able to distinguish between active and weak promoters, but is not applicable to the majority of promoters under control of the housekeeping factor. A recent study by Cox et al.[26] created a library of more than 200,000 variants of the *E. coli lacI* promoter fused upstream of a GFP reporter and used flow cytometry to characterize function. They learned a sequence-function map of the *lacI* promoter which recapitulated the known binding sites of the CAP responsive protein (CRP) and RNAP. They used this sequence-function map to explicitly model each protein's sequence-dependent binding energy and how their interaction affects transcription. They subsequently used this model to design promoters with a range of expression by tuning the strength of the RNAP binding site[27]. This approach is generally applicable to biophysically characterize transcriptional regulation by a specific sequence of interest, but is inherently limited in scope to only a few regulatory sequences at one time.

Thousands of putative transcription start sites have been identified in genome-wide studies

Recent efforts have been made to comprehensively characterize the *E. coli* genome using RNA-seq for global transcription start site mapping. A study by Conway et al.[28] analyzed the transcriptome of *E. coli* K-12 using strand-specific RNA-seq at single nucleotide resolution during

log-phase and stationary growth in glucose minimal medium and initially identified ~11,000 putative TSSs. Upon further curation, they report ~2,000 precisely annotated promoters. Another study by Thomason et al.[29] also analyzed the transcriptome of *E. coli* K-12 in three different growth conditions using differential RNA-seq, which can distinguish between primary and processed transcripts, and predicted ~14,000 TSS candidates. Both studies noted several novel antisense transcripts and functions. The TSS numbers reported in both of these studies far exceed the number of annotated genes in *E. coli* (~4,000) and the inflated numbers and reports of prevalent anti-sense transcription could be partially due to the imprecision of the computational predictions or artifacts of deep RNA-seq coverage[30]. This discrepancy between recent findings and known gene annotations highlights the complexity and remaining uncertainty still present in one of our most basic and longest studied model organisms. In order to provide a strong foundation for future study of promoter sequence-function relationships, we must begin with a confident set of TSSs and promoter sequences.

Initial library design

In a first step to understanding promoter sequence-function relationships in *E. coli*, our initial library design tests every reported putative TSS in its local sequence context and will determine which are functionally active *in vivo*. This initial design will provide a firm foundation that will enable us to build more quantitative and accurate sequence-function relationships. Our initial library design incorporates all TSSs from the RegulonDB database[31] and those identified in two recent genome-wide TSS mapping studies[28,29] described above. We synthesize each TSS embedded in its local sequence context -120 to +30 relative the TSS, capturing most of the *cis*--regulatory elements - most regulatory motifs fall within 100bp upstream of the TSS[26] and it has been shown that the initial transcribed region (+1 to +20) can also influence gene expression. Additionally, we included 500 negative controls from the *E. coli* genome that are not expected to have regulatory activity. These 150bp sequences are more than 200bp from a TSS (on either

strand) and fall mostly within coding sequences. Incorporating all three sources, there were 23,798 unique TSSs - for simplicity we reduced this set so that each TSS is at least 20bp apart, yielding a final library size of 17,836.

Massively parallel reporter assay to quantify promoter strength in E. coli

In order to efficiently test our large synthesized library, we designed a massively parallel reporter assay in *E. coli* (Figure 3.1B). Due to the current nature of oligonucleotide (oligo) synthesis (OLS) chemistry, only 30-40% of synthesized sequences are error-free and chip-synthesized barcodes would be attached to perfect as well as imperfect sequences. If barcode readout is the primary measurement (as is the case in our assay), there is no way to remove the potential effects of imperfect sequences – to do so you must sequence the entire construct. To avoid this issue, each variant is tagged with a unique random 20bp barcode added using polymerase chain reaction (PCR) during library amplification instead of using designed barcodes synthesized on the chip.

Next, a self-cleaving ribozyme (RiboJ), constant ribosome binding site, and GFP reporter are cloned immediately downstream – each variant drives the same construct with a constant 5' initially transcribed region to reduce any variable effects not due to promoter sequence. The GFP reporter stabilizes expression of the short barcode and provides an alternate method of functional characterization using flow cytometry. Prior to this cloning step, we perform an initial mapping step using DNA sequencing to identify barcodes for each variant, which allows us to efficiently sequence only the short barcode in downstream quantification. The library is integrated into a genomic "landing pad", an intergenic locus whose expression is not influenced by nearby genetic factors, using a Cre-Lox recombination system to induce a cassette exchange. Landing pads enable singly integrated, stable, and consistent expression across variants so that variations in expression are mainly due to variations in promoter sequence only. Genomic integration more faithfully represents an *in* vivo context compared to libraries expressed on plasmids, the dominant

method for library delivery in MPRAs[27]. Plasmids can have variable copy number between cells, necessitating the need for DNA sequencing for normalization in each tested condition. More importantly, variable plasmid copy number can artificially increase the presence of competing binding sites and change the effective amount of TFs, which can produce sharp changes in the input-output relationship between TFs and gene expression, an observation known as the transcription factor titration effect[32]. Following library integration, we perform a one-time barcode DNA sequencing to quantify construct levels post-integration. The initial stage of the workflow only has to be performed once per library, enabling us to quantify promoter strength in various conditions using only RNA-sequencing of the barcode. Our assay provides a quantitative level of promoter strength that is distinct from transcriptomic measurements – our readout depends only on the *cis*-sequence information and is not influenced by post-transcriptional mechanisms that affect expression. While these are important, training models based on a quantitative readout that is more representative of *cis*-sequence function will provide a solid foundation upon which to build more complicated models that incorporate additional layers of regulation.

REFERENCES

1. Peter Ruhdal Jensen, and K. H. The Sequence of Spacers between the Consensus Sequences.pdf. *Appl. enviromental Microbiol.* **64,** 82–87 (1998)
2. Miroslavova, N. S. & Busby, S. Investigations of the modular structure of bacterial promoters. *Biochem. Soc. Symp.* **10,** 1–10 (2006).
3. Brenner, S. Sequences and consequences. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365,** 207–12 (2010)
4. Baliga, N. S. Promoter analysis by saturation mutagenesis. *Biol Proced Online* **3,** 64–69 (2001).
5. Kinkhabwala, A. & Guet, C. C. Uncovering cis regulatory codes using synthetic promoter shuffling. *PLoS One* **3,** (2008).
6. Schleifer, A. & Tom-Moy, M. Method of producing oligonucleotide arrays with features of high purity. (2000).
7. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27,** 1173–1175 (2009).
8. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30,** 271–277 (2012).
9. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30,** 271–277 (2012).
10. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23,** 800–811 (2013).
11. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–6 (2008).

12. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6,** 386–98 (2005).
13. Lin, G. G. & Scott, J. G. Pre-mRNA splicing in disease and therapeutics. *Cell Trends Mol. Med.* **100,** 130–134 (2012).
14. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11,** 345–55 (2010).
15. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
16. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science (80-.).* **347,** 1254806 (2015)
17. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science (80-.).* **347,** 1254806 (2015)
18. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163,** 698–711 (2015).
19. Feklistov, A., Sharon, B. D., Darst, S. a & Gross, C. a. Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. *Annu. Rev. Microbiol.* 357–376 (2014). doi:10.1146/annurev-micro-092412-155737
20. Wösten, M. Eubacterial sigma-factors. *FEMS Microbiol Rev* **22,** 127–150 (1998).
21. Siebenlist, U., Simpson, R. B. & Gilbert, W. E. coli RNA polymerase interacts homologously with two different promoters. *Cell* **20,** 269–281 (1980).
22. Barne, K. a, Bown, J. a, Busby, S. J. & Minchin, S. D. Region 2.5 of the Escherichia coli RNA polymerase sigma70 subunit is responsible for the recognition of the 'extended-10' motif at promoters. *EMBO J.* **16,** 4034–4040 (1997).
23. Gourse, R. L., Ross, W. & Gaal, T. UPs and downs in bacterial transcription initiation: The role of the alpha subunit of RNA polymerase in promoter recognition. *Mol. Microbiol.* **37,** 687–695 (2000).

24. Rhodius, V. a. & Mutalik, V. K. Predicting strength and function for promoters of the Escherichia coli alternative sigma factor, sigmaE. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 2854–2859 (2010).
25. Rhodius, V. A., Mutalik, V. K. & Gross, C. A. Predicting the strength of UP-elements and full-length E. coli sigma-e promoters. *Nucleic Acids Res.* **40,** 2907–2924 (2012).
26. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9158–63 (2010).
27. Brewster, R. C., Jones, D. L. & Phillips, R. Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. *PLoS Comput. Biol.* **8,** (2012).
28. Conway, T. *et al.* Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. *MBio* **5,** 1–12 (2014).
29. Thomason, M. K. *et al.* Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli. *J. Bacteriol.* **197,** 18–28 (2015).
30. Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W. & Livny, J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* **13,** 734 (2012).
31. Gama-Castro, S. *et al.* RegulonDB (version 6.0): Gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* **36,** 120–124 (2008).
32. Brewster, R. C. *et al.* The transcription factor titration effect dictates level of gene expression. *Cell* **156,** 1312–1323 (2014).

CHAPTER TWO

A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Disruptions to Splicing

Title: A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect disruptions to splicing

Authors:
Rocky Cheung[1†], Kimberly D. Insigne[2†], David Yao[3], Christina P. Burghard[2], Jeffrey Wang[1], Yun-Hua E. Hsiao[4], Eric M. Jones[1], Daniel B. Goodman[5], Xinshu Xiao[2,6,7], Sriram Kosuri[1,7,8*]

Affiliations:
[1] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA
[2] Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, CA 90095, USA
[3] Department of Genetics, Stanford University, Stanford, CA 94035, USA
[4] Department of Bioengineering, University of California, Los Angeles, CA 90095, USA
[5] Department of Microbiology and Immunology, University of California, San Francisco, CA 94143, USA
[6] Department of Integrative Biology and Physiology, University of California, Los Angeles, CA 90095, USA
[7] Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA
[8] UCLA-DOE Institute for Genomics and Proteomics, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed. Tel: +1 310 825 8931; Email: sri@ucla.edu

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

SUMMARY

Mutations that cause exon skipping can have severe consequences on gene function and cause disease. Here we explore how human genetic variation affects exon recognition by developing a Multiplexed Functional Assay of Splicing using Sort-seq (MFASS). We assayed 27,733 variants in the Exome Aggregation Consortium (ExAC) within or adjacent to 2,198 human exons in the MFASS minigene reporter, and found that 3.8% (1,050) of variants, most of which are extremely rare, led to large-effect splice-disrupting variants (SDVs). Importantly, we find that 83% of SDVs are located outside of canonical splice sites, are distributed evenly across distinct exonic and intronic regions, and are difficult to predict *a priori*. Our results indicate extant, rare genetic variants, can have large functional effects at appreciable rates even outside the context of disease, and MFASS enables their empirical assessment for large-effect splicing defects at scale.

Keywords: splicing, exon recognition, rare variation, population variation, variant classification, massively parallel reporter assay

INTRODUCTION

Any individual's genome contains ~4-5 million genetic variants that differ from reference, and understanding how these variants give rise to trait diversity and disease susceptibility is a central goal of human genetics (Auton et al., 2015). A vast majority (96-99%) of an individual's variants are common, though at the population level the overwhelming majority of variants are rare (Montgomery et al., 2011; Nelson et al., 2012; Tennessen et al., 2012; UK10K Consortium et al., 2015). Common variants in the human population usually contribute small, additive effects towards complex traits, as negative selection has removed large-effect deleterious alleles (Altshuler et al., 2008). However, population expansion ~10,000 years ago left humans with an abundance of rare variation, and most Mendelian disease traits are caused by rare alleles with large effect sizes (Keinan and Clark, 2012). Because of their scarcity in an individual's genome, rare variants that play important roles in complex traits are likely to have large functional effects (Bomba et al., 2017; Gibson, 2012), and traditional population or computational genomic methods cannot reliably estimate their contribution (Uricchio et al., 2016).

Recent whole genome and transcriptome sequencing studies of large cohorts indicate that rare variation is playing an important role in shaping global gene expression (GTEx Consortium et al., 2017; Hernandez et al., 2017; Li et al., 2017). However, new comprehensive reverse-genetic studies indicate that individual mutations in promoter and enhancer regions rarely have large effects (Canver et al., 2015; Diao et al., 2016; Gasperini et al., 2017; Rajagopal et al., 2016; Sanjana et al., 2016), which could be the result of functional redundancy between transcription control elements (Frankel et al., 2010; Hong et al., 2008; Osterwalder et al., 2018). How can individual rare variants be broadly shaping gene expression, but at the same time rarely having large effects on transcriptional control? We can expect the mutational profiles of large-effect rare

variants to mirror those that cause Mendelian traits, which are dominated by non-synonymous exonic mutations, structural and copy number variants, or mutations that affect splicing (Bamshad et al., 2011; Chong et al., 2015). While copy number changes and non-synonymous mutations are easy to detect, splicing changes are more difficult to diagnose, as only mutations at canonical splice sites are easy to predict and interpret (Jian et al., 2014).

Recent evidence indicates that splicing is a major mechanism by which genetic variation influences traits. For common variants, large-cohort RNA-Seq studies that examine splicing are finding many splicing quantitative trait loci (sQTL), especially when considering exon-level expression differences (GTEx Consortium, 2015; Ongen and Dermitzakis, 2015; Zhang et al., 2015). Moreover, a majority of eQTLs tend to act on an individual exon level rather than the gene level, indicating that cis-eQTLs might be broadly affecting exon recognition (Ramasamy et al., 2014). In addition, functional genomic measurements of GEUVADIS individuals indicate that common genetic variation influencing splicing is a primary mechanism that confers susceptibility to common diseases (Li et al., 2016). For rare variation, analysis of bottlenecked populations find that many rare variants which segregate with large-effect expression changes are enriched at splice sites (Pala et al., 2017). In addition, prospective transcriptional profiling studies for Mendelian diseases are increasingly finding many rare variants that affect splicing are difficult to predict *a priori* (Cummings et al., 2017; Kremer et al., 2017). More broadly, computational splicing predictors trained on RNA-Seq data and sequence features seem to indicate that many rare and disease variants are predicted to influence splicing levels (Xiong et al., 2015). Finally, mutations that cause an exon to be skipped can have severe functional consequences on gene function, and many known disease-causing mutations reduce or eliminate exon recognition (Baralle and Buratti, 2017). A large-scale functional assay examining ~5,000 exonic disease mutations indicates that ~10% of them have some effect on splicing (Soemedi et al., 2017). Many of these

variants that alter splicing are not located close to the splice sites, suggesting that many splicing defects are likely yet to be discovered.

We developed MFASS as a multiplexed, scalable platform to test the extent to which mutations, both within exons and introns, can lead to large-effect defects in exon recognition. MFASS uses a set of three-exon, two-intron minigene reporters in which skipping of the middle exon leads to reconstitution of fluorescence (**Figures 2.1A and 2.S1A-S1D**). We cloned libraries of microarray-derived oligonucleotides that encoded human exons and surrounding intronic sequences into these reporters *en masse* to construct reporter libraries (LeProust et al., 2010). These libraries are then integrated into HEK293T human cell lines using high-efficiency, serine-integrase based, site-specific integration (**Figure 2.1A**), ensuring one copy of library sequence per cell (Duportet et al., 2014; Matreyek et al., 2017). The pooled sequence library is then separated into bins using fluorescence-activated cell sorting (FACS), and we use DNA-Seq of the constructs to quantify which variants are sorted into which bins. We used MFASS to functionally classify 27,733 exonic and intronic natural genetic variants from ExAC for exon recognition across 1,626 genes in 2,198 exon backgrounds, most of which are extremely rare variation in the human population. Here we show that more than a thousand (3.8%) of these rare genetic variants leads to near complete loss of exon recognition, on par with the prevalence of protein-truncating variants within genomes. Most of the effects of rare variants on splicing are challenging to predict.

RESULTS

Optimization of MFASS

We tested human exons in several reporter designs. Our initial designs relied on the reconstitution of fluorescence using a pair of constant short DHFR introns (~100bp) flanking the exon library (**Figure 2.S1D**). However, we found that much of the library had little to no fluorescence, and even when signal was present, the expression levels were low compared to the longer intron

contexts (**Figures 2.S1E and 2.S1F**). These results were suggestive of intron retention, which is a process that dominates in lower eukaryotic organisms. In humans, due to long intron lengths, exons are first recognized by the splicing machinery in a process called exon definition (Berget, 1995, Black, 2003; De Conti et al., 2012; Keren et al., 2010) and thus mutations that affect exon recognition often result in exon skipping rather than intron retention (Baralle and Buratti, 2017). Due to these concerns, we optimized our reporter designs with longer constant intron backbones. With these backbones, we observed ~20 to 100-fold higher level of fluorescence overall.

In order for MFASS to work in a multiplexed, scalable format, the assay relies on a single copy of the reporter construct per cell before FACS sorting, thereby ensuring that our splicing fluorescence readout corresponds to a single sequence. Each library sequence is integrated once per cell using high efficiency site-specific genome integration (**Figures 2.1A and 2.S1G-S1I**), and expressed at a defined AAVS1 locus to minimize any pleiotropic effects. However, we noticed upon transient transfection of the splicing reporter libraries that each cell contains hundreds of reporter copies on average (**Figure 2.S1J**, **top left**). We characterized the copy number of the reporter library in human cells across culture passages by flow cytometry and RT-PCR and found ~100,000-fold cell dilution to be sufficient (**Figure 2.S1K**). Therefore, we obtained MFASS data only at passage number beyond 1:100,000 from initial transfection, to ensure single copy integration without contaminating plasmids (**Figure 2.S1M**).

While transient assays with splicing minigene reporters are commonly used due to their relative simplicity, it has been reported that splicing outcomes can be more reproducible when sequences are genomically integrated in many cell types (Smith and Lynch, 2014). To compare the effect of genome-integration as opposed to transient expression for our splicing reporters, we constructed reporters corresponding to individual library sequences, and evaluated both fluorescence and RNA splicing under transient expression and site-specific genome integration (**Figures 2.S1N-**

2.S1Q). We selected nine sequence variants that match our reference library for further analysis by flow cytometry (**Figures 2.S1N and 2.S1O; STAR Methods**) and RT-PCR (**Figures 2.S1P and 2.S1Q**) and to validate results from MFASS. Individual controls sorted from the library showed consistent behavior between inclusion rates estimated by RT-PCR and fluorescence output (**Figures 2.S1N-2.S1Q**). While the level of exon inclusion as measured by RT-PCR is consistent between transient and stable expression, reporter fluorescence in stably integrated constructs is more consistent with RT-PCR results because the transient transfections included signals at very high gene dosage (**Figure 2.S1N**, note behaviors of individual constructs that show saturation of fluorescence only at high expression) that is only alleviated when single-copy integration is achieved (**Figure 2.S1O**).

Evaluating MFASS Based on Known Splicing Regulatory Elements

To test and validate MFASS, we first designed, built and assayed a test library of 6,713 mutations aimed at perturbing regulatory elements across a randomly chosen library of 205 natural in-frame human exons and surrounding intronic sequences (Splicing Regulatory Element library). We first developed this test library in order to evaluate the MFASS assay and test the effects of designed mutations in a large set of natural sequence contexts. To mutate sequences iteratively while accounting for the creation of unintentional motifs, we developed a custom software toolkit for the design of *in silico* splicing mutations. In particular, this toolkit incorporates information about splicing regulatory elements from the literature to calculate a composite score for each sequence across different functional classes. We chose natural human exons that are less than 100 bp and begin and end on frame 0, and designed a 170-bp exon library with its surrounding intronic contexts, that includes at least 40 bp of upstream intron and at least 30 bp of downstream intron. Overall, we randomly chose a subset of ~200 human exons and iteratively designed 60-80 perturbations per sequence that weaken, strengthen or destroy splicing motifs focused on three major motif types (**Tables 2.S1 and 2.S2**; **STAR Methods**).

We used MFASS to assay the SRE library with biological replicates across two different intronic backbones (**Figure 2.1A**). We expanded these sorted bins over several passages and observed that the sorted populations remained stable (**Figure 2.1B**). We also performed bulk RT-PCR for each bin, and found that the observed RNA splicing efficiencies corresponded with observed fluorescence of the bins (**Figure 2.1C**). To obtain an exon inclusion index for each sequence, we first considered reads which perfectly matched the SRE library, and normalized based on read depth and weighted by the corresponding bin population percentage from FACS**.** Finally, we computed a weighted average of normalized read counts across all bins using the average exon inclusion level in each bin as measured by the GFP:RFP ratio and confirmed by bulk RT-PCR (**STAR Methods**). Overall, the inclusion indices for our library are bimodal, with most library sequences represented predominantly in one bin, showing either complete exon inclusion or skipping (**Figures 2.1D**).

We measured the replicability of inclusion indices across biological replicates using the tetrachoric correlation (r_t) due to the bimodality in our results (Pearson correlation provided as a comparison). We tested these libraries across two constant intron backbones (SMN1 and DHFR), and found that exon inclusion metrics are highly reproducible within the backbone across biological replicates (**Figures 2.1E and 2.1F**) (r_t = 1.00, $p < 10^{-16}$, tetrachoric; r = 0.94, $p < 10^{-16}$, Pearson, DHFR intron backbone; r_t = 0.97, $P < 10^{-16}$, tetrachoric, r = 0.89, $p < 10^{-16}$, Pearson, SMN1 intron backbone)**,** and between backbones (**Figure 2.1G**) (r_t = 0.96, $p < 10^{-16}$, tetrachoric; r = 0.85, $p < 10^{-16}$, Pearson). We consider 6,713 designed mutations present across both backbones in subsequent analysis and highlight data for the SMN1 intron backbone (**Figure 2.2**).

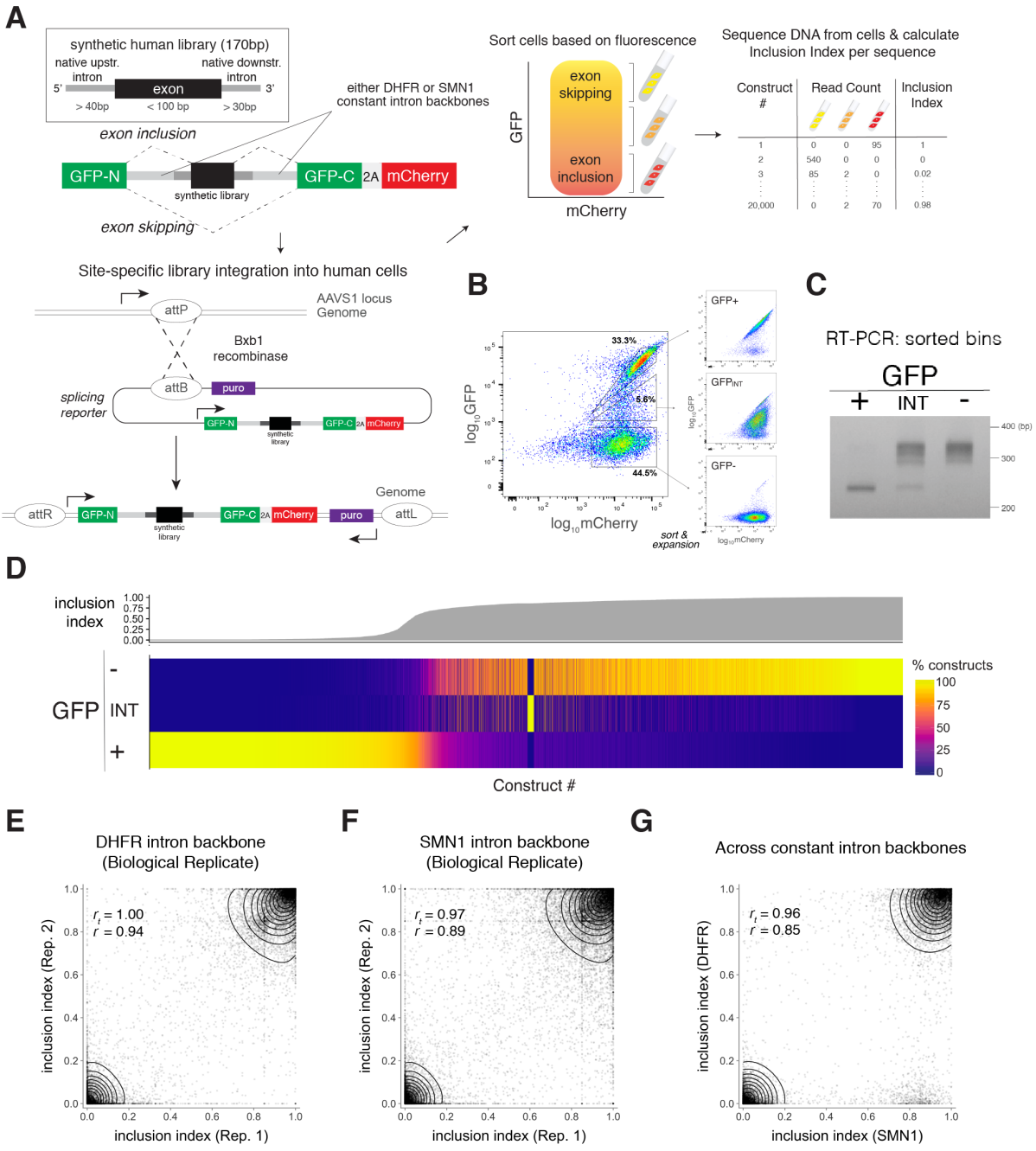


Figure 2.1. Multiplexed Functional Assay of Splicing by Sort-seq (MFASS).
(**A**) We cloned synthetic human exons (*black*) and surrounding intronic sequences (*dark grey*) into our reporter plasmid containing a split-GFP reporter with flanking constant intron backbones (*light grey*), followed by site-specific integration into HEK293T cells using Bxb1 integrase. Cells are sorted into bins based on fluorescence, followed by amplicon sequencing of DNA from cells in each sorted bin. We calculated exon inclusion index for each sequence using a weighted average of normalized read counts based on exon inclusion level from bins (**STAR Methods**).

(**B**) We used FACS to sort the genomically-integrated SRE library into three separate populations (*left*). After expansion, the sorted populations remained stable (*right*). GFP-int, GFP-intermediate. For this library (SMN1 intron backbone), we obtained ~4 million cells for GFP_+ and GFP_{neg} bins, and 4.2 x 10^5 cells for GFP_{int} bin. The percentage of cells sorted is as follows: GFP_+ (33.3%), GFP_{neg} (44.5%), GFP-int (5.6%).
(**C**) The observed RNA splicing efficiencies of the sorted bins as measured by RT-PCR correspond almost directly with observed fluorescence of the bins.
(**D**) We plotted the percentage of reads for each construct in the SRE library containing both natural and mutant exons (n = 10,477). We showed that most sequences fall predominantly into one bin, exhibiting either complete exon skipping or inclusion, allowing for facile classification of exon skipping variants of large effects (Δinclusion index ≤ -0.5). Corresponding exon inclusion indices for each bin are indicated at top panel. The data shown in (**D**) corresponds to the SMN1 backbone.
(**E, F and G**). SRE library splicing behavior replicates between individual biological replicates and across two constant intron backbones. Tetrachoric correlation indicates whether two distinct measurements are concordant in one of the four quadrants, and is more suited to assess large-effect variants. (**E**) Exon inclusion indices show strong correlation between two independent biological replicates for *C. griseus* DHFR intron backbone (r_t = 1.00, $p < 10^{-16}$, tetrachoric; r = 0.94, $p < 10^{-16}$, Pearson). (**F**) Exon inclusion indices show strong correlation between two independent biological replicates for human SMN1 intron backbone (r_t = 0.97, $p < 10^{-16}$, tetrachoric, r = 0.89, $p < 10^{-16}$, Pearson). For (**E**) and (**F**), after calculation of correlation coefficients, sequences for which inclusion indices do not agree within 0.30 (outside the dashed lines) are excluded from subsequent analysis. (**G**) Results are robust across different intron backbones (r_t = 0.96, $p < 10^{-16}$, tetrachoric; r = 0.85, $p < 10^{-16}$, Pearson).
See also **Figure S1**.

Overall, we showed that while the loss of exon recognition is consistent with known splicing motifs, the effects of these perturbations are not easily predicted for 6,713 designed mutations across 205 human exons (**Figure 2.2**). To focus on the mechanisms by which large-effect splicing changes can occur, we defined large-effect variants as Δinclusion index ≤ -0.5 (i.e., mutations to a wild-type exon with an inclusion index of ≥ 0.5, that is reduced by an absolute value of at least 0.5), which we term “splice-disrupting variants” (SDVs). We quantified the percentage of SDVs for designed mutations in each category (**Figure 2.2A**). As expected, we found that splice-site mutations to the nearly invariant dinucleotides cause SDVs at the highest rates (**Figure 2.2A**). Mutations to the splice site (splice acceptor, positions -20 to +3; splice donor, positions -3 to +6) individually result in SDVs 48-73% of the time (**Figure 2.2A,** “acceptor site” and “donor site”), and 96% of the time when mutating simultaneously both splice donor and acceptor (**Figure 2.2A,** “acceptor + donor site”). This is likely an underestimate as mutations eliminating splice site recognition may be utilizing alternative splice acceptors or donors, which cannot be distinguished

from exon inclusion by MFASS. Within exons, mutations can still have strong effects. Encoded synonymous mutations to all putative exonic splicing enhancers (ESEs) lead to SDVs ~72% of the time (**Figure 2.2A,** "all exonic splicing enhancers"). While removing clusters of putative exonic splicing silencers (ESSs) result in increased exon inclusion (**Figure 2.S2A,** "all exonic splicing silencers"), removing the strongest identified ESE alone results in 30% SDVs (**Figure 2.2C,** "strongest exonic splicing enhancer"). More generally, splicing metrics such as MaxEnt for splice site strength (**Figure 2.2B**) or exon hexamer metrics (**Figures 2.2C and 2.S2B**) are consistent with predicted effects on splicing behavior.

Effects of Rare Human Variation on Exon Recognition

While these results indicate that mutations intended to alter previously recognized motifs can commonly lead to loss of exon recognition, we wanted to explore the extent to which natural genetic variation in the human population results in SDVs. We generated the Single Nucleotide Variant library (SNV library) for which we designed and synthesized all cataloged exonic and intronic single nucleotide variants (SNVs) from the Exome Aggregation Consortium (ExAC), for all 2,902 wild-type human exons that demonstrated exon inclusion (inclusion index ≥ 0.8) in the SRE library (**STAR Methods**). From this SNV library, we first tested two reporter constructs that split at distinct positions of GFP to assess how the reading frame affects exon inclusion for either exons that start and end at phase 1, or exons that start and end at phase 0. To evaluate the splicing reporter output across two versions of the SNV dataset from the MFASS assay, we monitored GFP and mCherry fluorescence from the initial library and sorted cells using flow cytometry (**Figures 2.S3A and 2.S3B**). We observe that the fluorescent output from all sorted

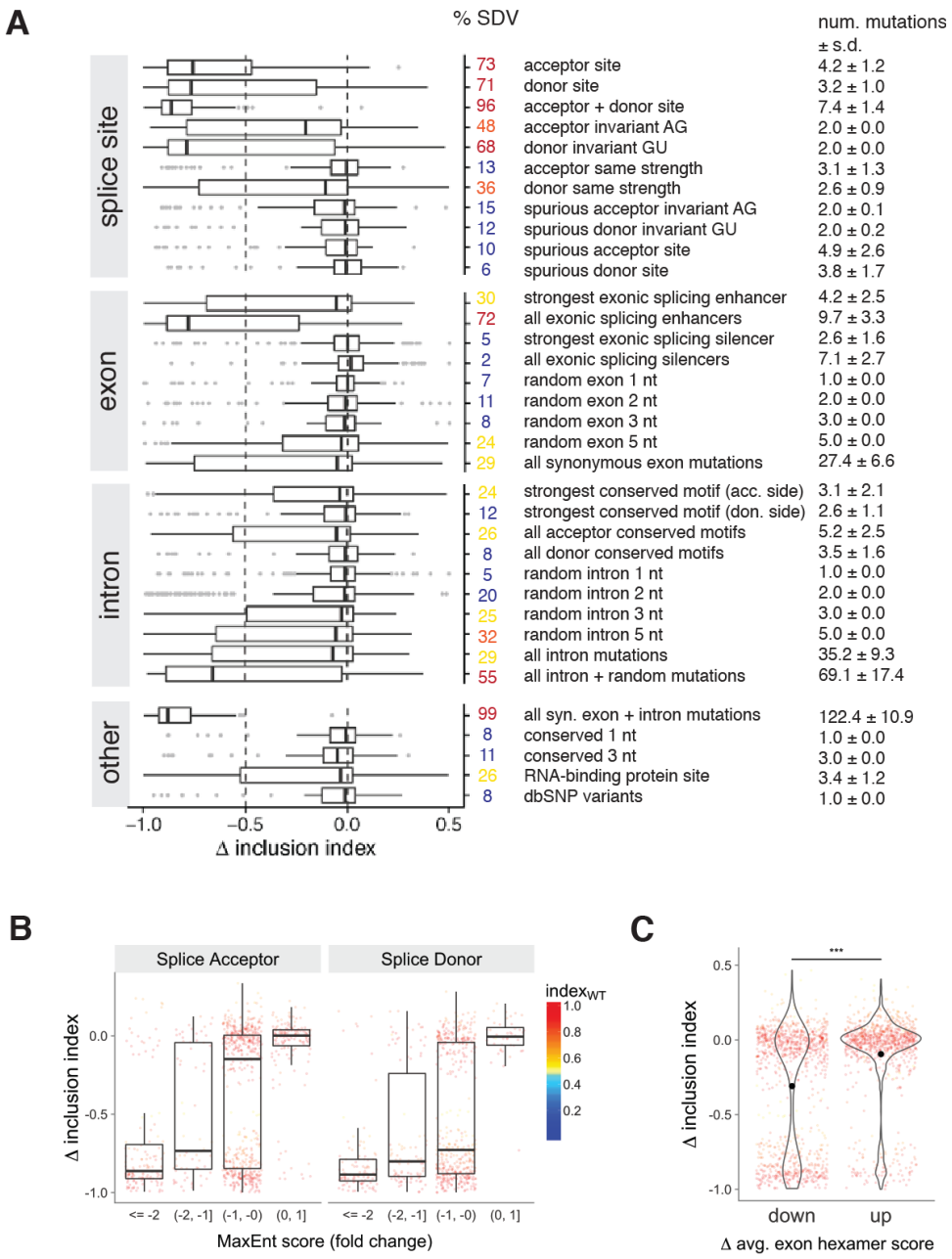


Figure 2.2. Effects on exon recognition are not easily predicted across 6,713 designed mutations in splicing regulatory elements.

(**A**) We quantitatively measured exon inclusion for iteratively-designed mutations (*n* = 6,713) across categories of splicing regulatory elements from 205 human exons (see **Tables 2.S1** and **2.S2** for complete categorical explanations and definitions). For each designed category we show the distribution of inclusion indices, the proportion of splice-disrupting variants (SDVs), and the average number of mutations within one standard deviation. We defined SDVs as variants that result in a Δinclusion index ≤ -0.5 (relative to the wild-type sequence; **STAR Methods**). We only consider SNVs when the corresponding wild-type sequence is also detected, requiring that the wild-type exons demonstrate inclusion in our assay (inclusion index of ≥ 0.5) for variants to be considered an SDV. Dashed lines mark the thresholds for no change in exon inclusion (Δinclusion index = 0) and for SDVs (Δinclusion index = -0.5). Here we highlight the data for the SMN1 intron backbone and detected 21.3% (1,428/6,713) of variants as SDVs across all categories. See also

Figure 2.S2A for mutations to exons that are skipped in MFASS (inclusion index of < 0.5) across designed categories. Splice acceptor, positions -20 to +3; splice donor, positions -3 to +6.
(**B**) Mutating the splice acceptor and splice donor sites adversely affects exon inclusion based on MaxEnt prediction for included exons (inclusion index of ≥ 0.5) (Yeo and Burge, 2004).
(**C**) There is a significant difference in average Δinclusion index between sequences which increase (up) or decrease (down) overall exon hexamer score (Mann-Whitney *U* test, $p < 10^{-16}$) for included exons. Decreasing overall exon hexamer score leads to more exon skipping. Hexamer scores are based on the HAL model (Rosenberg et al., 2015). An alternative score metric is evaluated in **Figure 2.S2B** (Ke et al., 2011).
See also **Figure 2.S2, Tables 2.S1 and 2.S2**.

populations is stable upon passaging. Overall, the two different contexts displayed high correlations for detecting splice-disrupting variants. (**Figure 2.S3C**, n = 5,740, r_t = 1.00, $p < 10^{-16}$, tetrachoric; r = 0.94, $p < 10^{-16}$, Pearson). Since the SNV library was examined in independent reporter constructs testing different frames, this indicates we will be able to use MFASS to screen for exons across in-frame and frameshifting exons for future studies. These results suggest that the exons examined across diverse reporter contexts are functionally consistent and relevant in the context of exon recognition in human cells.

Overall, we quantified the effects of more than half (52.4%, 27,733 of 52,965) of the ExAC SNVs found across 2,198 exons and found that 1,050 of 27,733 (3.8%) ExAC variants assayed led to almost complete loss of exon recognition, are broadly spread across 543 human exon backgrounds from 473 genes (**Figure 2.3A**), correspond to 1,038 distinct genomic positions, and show increased sensitivity at the splice regions (**Figure 2.3B**). Correlations between biological replicates were high (n = 31,583, r_t = 0.94, $p < 10^{-16}$, tetrachoric; r = 0.80, $p < 10^{-16}$, Pearson) (**Figure 2.S3D**). To minimize false positives, we require replicate agreement within 0.20 instead of 0.30 used for the SRE library (**STAR Methods**). To ensure that MFASS-identified SDVs are robust to experimental artifacts, we additionally analyzed a number of controls. First, we tested the SNV library using three control sets (**Figure 2.3C**): (1) scrambled nucleotides, (2) a previously tested set of skipped exons in our SRE library, and (3) systematic mutations of both splice sites from the wild-type sequences. As expected, 24 of 24 (100.0%) scrambled nucleotides, 70 of 71

(98.6%) skipped exons, and 945 of 977 (97.3%) broken splice-signal sequences result in loss of exon recognition (inclusion index < 0.5) (**Figure 2.3C**), noting that alternative 5' and 3' splice site usage result in false negatives for MFASS. In addition, we also analyzed sequences containing synthetic errors resulting in single nucleotide deletions (*n* = 9,801) from our designed sequence library (**STAR Methods**). SDVs are enriched across the exon-intron junction at the splice acceptor and donor for these deletions derived from synthetic errors (**Figure 2.3D**).

Finally, we further validated MFASS results individually for 34 SDVs across multiple functional classes of splicing variation across the original tested context as well as longer intronic contexts in HEK293T cells (**Figure 2.3E; STAR Methods**). Our results suggest that MFASS is robust across a majority of rare genetic variants tested for splicing defects. We individually verified SDVs using transient expression assays and found that nine of 11 (81.8%) showed large-effect splicing defects, with all 11 (100.0%) showing reduced exon inclusion relative to their respective wild-type sequences (**Figure 2.3E**). Furthermore, we tested the effect of longer intronic context on individual SDVs, and found that 17 of 23 (73.9%) showed large defects in splicing, with only one of 23 (4.3%) mutations showing no appreciable exon recognition defect (**Figure 2.3E**). Finally, to examine the cell-type specificity of SDVs, we further picked a subset of 15 SDVs with the strongest changes in exon inclusion, and tested wild-type or matched SDV reporter constructs across three additional cell types in the ENCODE consortium (*n* = 14,15 for HEK293T; *n* = 14,15 for HeLa S3; *n* = 14,15 for HepG2; *n* = 14,15 for K562; *n* = WT, SDV respectively). We found that large-effect splicing disruptions are consistent across four cell types in all 15 of the splice-disrupting variants assayed (15/15, 100.0%) (**Figure 2.3F**). Individually, exon inclusion levels also largely transfer across these cell types for all 29 sequences examined (**Figure 2.S3E**).

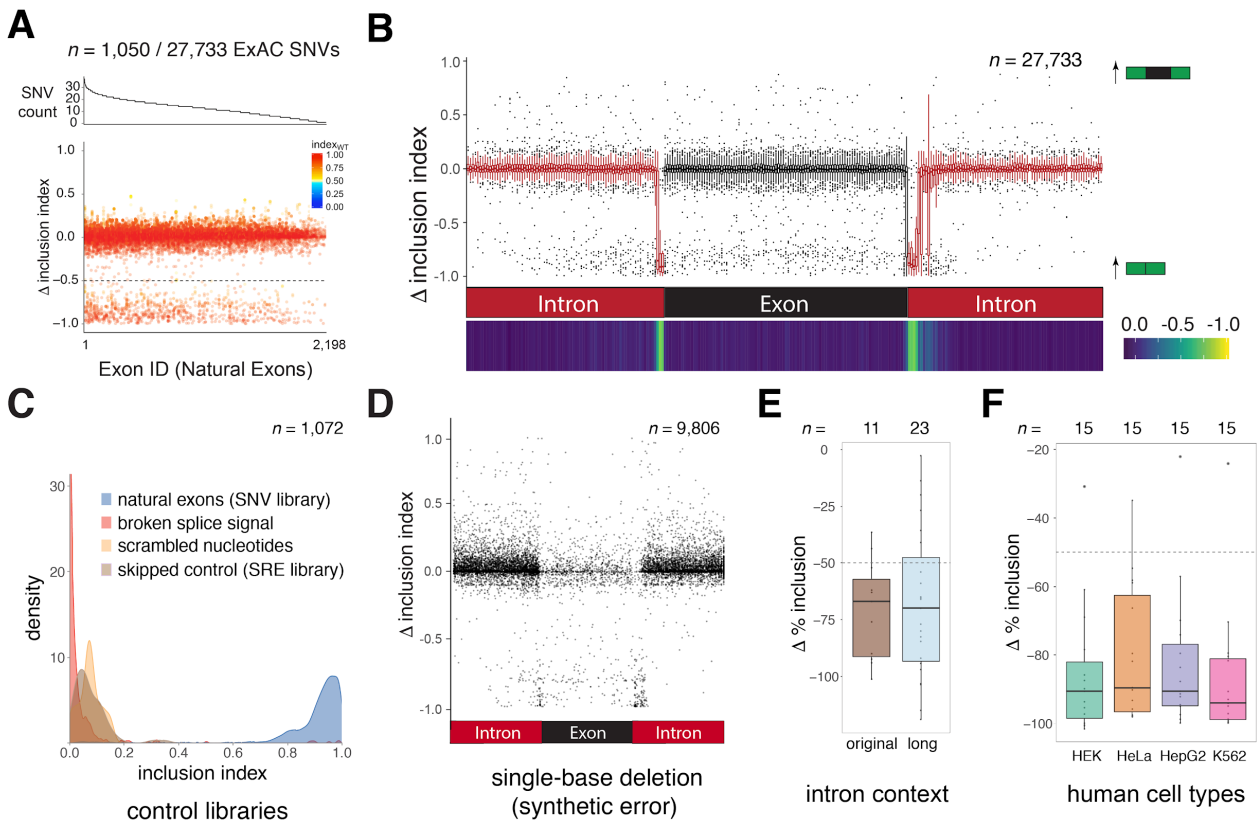


Figure 2.3. MFASS enables functional characterization of variant effect on splicing at scale across libraries of human variants.

(**A**) The number of SNVs per exon sequence (top) and the Δinclusion index (bottom) of the 27,733 ExAC SNVs are plotted against the wild-type exon backgrounds (*n* = 2,198), and colored by the inclusion index of the corresponding WT sequence. Both the top and bottom panels are ordered in decreasing number of variants tested from 44 to 1 per human exon background, with an average of 12.6 human variants and 3.8 SDVs per assayed wild-type exon sequence background. We find 1,050 of 27,733 SNVs tested (3.8%) are SDVs (Δinclusion index ≤ -0.5) and are broadly spread across the 543 human exon backgrounds in 473 genes. Dashed line indicates the threshold (Δinclusion index = -0.5) below which we call SDVs.

(**B**) The change in inclusion index as a function of relative position for our SNV library across 2,198 human exon sequences shows that the splice donor and acceptor sites are most sensitive to mutations. Intron-exon boundary on the left corresponds to the splice acceptor, while the intron-exon boundary on the right corresponds to the splice donor. The splice donor is more sensitive to mutation because its consensus site is longer and more conserved. The bottom panel displays the relative sensitivity of each position. Each bin corresponds to 1-2 nucleotides per position, and locations are relative as we test a range of exon lengths.

(**C**) Three control sets for validating the SNV library (*n* = 1,072). Most control sequences that were designed to cause exon skipping led to almost complete loss of exon recognition. The three control sets were (i) scrambled sequences (*n* = 24), (ii) a previously tested subset of exons that were skipped in the SRE library (*n* = 71), and (iii) breakage of the splice sites (*n* = 977). The broken splice-signal control library mutates 5' splice sites (SD) at the downstream intron from GT to CC, and 3' splice sites (SA) at the upstream intron from AG to TT. SD, splice donor, SA, splice acceptor. We include the distribution of wild-type sequences (i.e., natural exons) (*n* = 2,339, of which 2,198 sequences have relevant SNV data, **STAR Methods**).These exons initially

demonstrated exon inclusion in the SRE library (inclusion index ≥ 0.8), and we subsequently retest them and their associated SNVs in the SNV library.
(**D**) We analyzed the effects of deletions derived from synthetic errors on exon inclusion. We showed the effect of exon inclusion for synthetic deletions (*n* = 9,801) across replicates, with an SDV rate of 3.59%. We observed an enrichment of SDVs at or near the splice sites.
(**E**) We validated large-effect rare variants detected by MFASS (*n* = 34) and their corresponding wild-type sequences. We measured exon inclusion in either the original sequence context examined in MFASS (*n* = 11), or as a more stringent test with an additional 130bp of longer intronic contexts (*n* = 23) in HEK293T cells. For the longer set, we tested SDVs that represent variant classes in Figure 4B: missense variants (*n* = 3), synonymous variants (*n* = 3), intron variants (*n* = 4), splice donor (*n* = 4), splice acceptor (*n* = 5), and splice region variants (*n* = 4). The levels of exon inclusion were calculated for both the individual SDV and its respective wild-type sequence. The change in % exon inclusion is calculated as an absolute difference for that of the mutant and the respective wild-type sequence, with a negative value indicating exon skipping for a variant relative to the wild-type. All mutants were normalized to a no-insert control as a baseline of complete exon skipping for the assessment of change in exon inclusion. Dashed line indicates the threshold (Δ % inclusion = 50%) below which we call splicing-disrupting variants (SDVs).
(**F**) To examine the cell-type specificity of SDVs, we further picked a subset of 15 SDVs from the long intronic context with the strongest change in inclusion levels for testing their effects across 4 cell types, and validated reporter constructs for wild-types (WT) or the corresponding SDVs. *n* = WT, SDV: 14,15 (HEK293T), 14,15 (HeLa S3), 14,15 (HepG2), 14,15 (K562). We found that large-effect splicing disruptions are consistent across 4 cell types in all 15 of the splice-disrupting variants assayed (15 of 15, 100%). The generalizability of *per variant* exon inclusion measurements across cell types is included in **Figure 2.S3E**.
See also **Figure 2.S3**.

Of the 1,050 SDVs detected, we observe almost equal contributions from introns (561, 54%) and exons (489, 46%) among the variants we tested (**Figure 2.4A**). We found that 76% of splice site variants are SDVs (**Figure 2.4B, left**). Compared to the splice site variants, variants in the broader splice region, synonymous exonic variants, non-synonymous exonic variants, and deeper intronic variants disrupt splicing more rarely at 8.5%, 3.0%, 3.1%, and 1.5% respectively (**Figure 2.4B, left, Figures 2.S4A and 2.S4B**). Interestingly, because SNVs are not equally distributed among these categories, splice site SDVs only constitute 17% of all SDVs, whereas intron variants, which are the least sensitive to splicing disruption, comprised 16% of SDVs (**Figure 2.4B, right**). The splice donor and acceptor regions show different patterns of sensitivity to splicing disruptions (**Figure 2.4C**), with splice donor regions being more sensitive than splice acceptor regions. SNVs at the splice sites are rare in our library (**Figure 2.4C,** bottom, SNV density), and also for all ~7.4 million

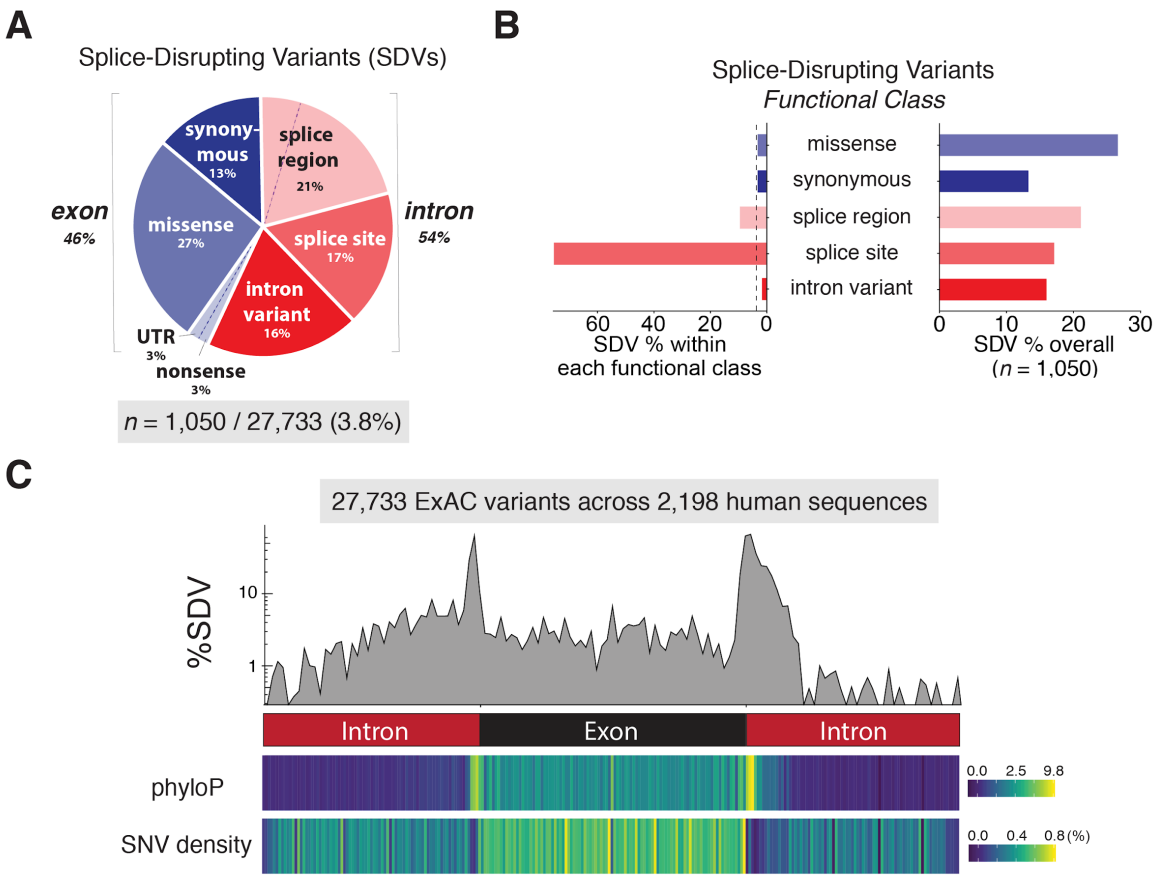


Figure 2.4. Global analysis of splice-disrupting variants across 27,733 ExAC SNVs in or near 2,198 human exons.

(**A**) We functionally classified our variants by variant class from the Ensembl variant effect predictor (**STAR Methods**). SDVs (*n* = 1,050) from natural genetic variation are split almost equally between exonic and intronic regions (*blue* and *red* respectively). Dashed line separates the exonic regions (4%) and intronic regions (17%) of the splice region variants. Splice site variants are defined as those within 2 bp of intron adjacent to exon, whereas splice region variants are located 3 bp into the exon and 8 bp into the intron, excluding splice sites.

(**B**) Splice site mutations are by far the most likely region to result in an SDV (*left*). However, because SNVs at splice sites are relatively rare, SDVs in regions other than the splice site constitute 83% of all SDVs (*right*). The distributions for non-SDVs across variant classes and the distribution of SDV effect sizes are shown in **Figure 2.S4A**.

(**C**) The percentage of SDVs as a function of position along the exon and surrounding intron sequence shows that splice donor regions are more sensitive than splice acceptor regions (*top panel*). Plotted below is the average change in mammalian evolutionary conservation (phyloP score averages) and ExAC SNV density as a function of location. Each bin corresponds to 1 to 2 nucleotides per position, and locations are relative to account for variable exon length.

See also **Figure 2.S4**.

ExAC variants (**Figure 2.S4C**). The larger number of variants in regions away from the splice sites outweighs their reduced sensitivity (**Figure 2.4C,** bottom, SNV density), and contribute 83% of the 1,050 SDVs reported here.

Population Genetic, Evolutionary and Functional Analyses of Splice-Disrupting Variants

A number of population genetic, evolutionary, and functional characterizations indicate that our measured SDVs are relevant. *First*, the proportion of SNVs that are SDVs shows significant reductions as a function of allele frequency (chi-squared test, $p = 1.03 \times 10^{-4}$). Consistent with population genetic theory, a vast majority (98.8%) of our SDVs are extremely rare (allele frequency from the Genome Aggregation Database (gnomAD) < 0.5%) (**Figure 2.5A**). *Second*, we find a significantly lower SDV rate (~2x) within genes that rarely have protein-truncating variants (PTVs) within ExAC indicating strong functional constraint (pLI ≥ 0.9) (Lek et al., 2016) (**Figure 2.5B**) (two-tailed Fisher's exact test, $p = 3.0 \times 10^{-11}$). Considering the rates of SDV and PTV overall, we conclude our SDV rate is at least on par to that of protein-truncating variants from ExAC. *Third*, SNVs that are SDVs show significantly stronger evolutionary conservation, suggesting purifying selection at these sites (Mann-Whitney *U* test, $p < 10^{-16}$) (**Figure 2.5C**). Missense variants alone do not seem to drive the conservation signature, as the difference in mean phyloP conservation score is greater without missense variants ($\text{phyloP}_{\text{non-SDV}} = 0.04$ vs $\text{phyloP}_{\text{SDV}} = 2.7$) than with missense variants ($\text{phyloP}_{\text{non-SDV}} = 1.4$ vs $\text{phyloP}_{\text{non-SDV}} = 3.1$) (Student's two-sample *t*-test, $p < 10^{-16}$, two-sided), suggesting that SDVs are under stronger evolutionary conservation independent of missense variation. *Fourth*, nucleotide positions under strong evolutionary conservation have higher rates of SDVs, and this is especially apparent within introns (two-tailed Fisher's exact test, $p < 10^{-16}$) (**Figure 2.5D**). However, this conservation has limited predictive power, because within introns there are many more SNVs at neutral sites than sites under strong conservation, and within exons most sites are highly conserved (**Figure 2.5E**). *Fifth*, for exonic SNVs, we observed that SDVs significantly reduce exon hexamer scores when

compared with non-SDVs, suggesting that SDVs are disrupting important functional sites for exon recognition (Student's *t* test, $p < 10^{-16}$) (**Figure 2.5F**). *Sixth*, motif enrichments at the splice acceptor suggests that SDVs enriched for T to A mutations disrupt the area near the mechanistically important polypyrimidine tract, while for splice donors we find that guanine-rich motifs are less tolerated (**Figure 2.S5A**). *Seventh,* we found several enriched gene ontology (GO) terms for SDVs (compared to the tested SNV library) comprising of four enriched categories (**Table 2.S3; STAR Methods**). Three of the GO categories contain mostly collagen genes, many of which have large repeated protein domains. In addition, the last category, post-Golgi vesicle-mediated transport, also contained a number of SDVs in genes with other repeat domains such as ankyrin and spectrin repeat domains. Such repeat-expansion genes can often be variable between populations, and in-frame exon skipping events are likely to have fewer severe consequences (Chan et al., 2008).

Cross-validation of Individual SDVs

It is likely that some fraction of SDVs detected by MFASS do not reflect actual changes in humans because minigene reporters are widely used but imperfect models of endogenous exon recognition (Cooper, 2005; Gaildrat et al., 2010). For example, we detect 11 SDVs with a minor allele frequency of greater than 0.5% that correspond to a set of common variants. Since common variants will likely overlap with other datasets, we first cross-referenced our ExAC library with the ClinVar database (Landrum et al., 2013). Only 0.5% (141/27,733) of the ExAC library is present in ClinVar, with eight SDVs and two annotated pathogenic variants in *MTMR2* and *PARN* genes. To look more broadly in the datasets of healthy cohorts other than ExAC, we cross-referenced our assayed SNVs with the Genotype Tissue-Expression (GTEx) project (GTEx Consortium, 2015). Overall, 9 of these 11 common variants have exon inclusion levels from GTEx (∆percent spliced in, ∆PSI, **Figure 2.S5B**), and three had globally significant differences (**Figure 2.S5B, i, ii, and v**). If we extend this analysis to rare variants as well, we were able to determine PSI values

for 1,471 assayed exons (**STAR Methods**), but only 28 are SDVs (including the common SDVs described above). Of these 28, seven (25%) show globally significant difference in exon inclusion levels from RNA-Seq, and five (20%) of which are larger effects. In addition, two additional SNVs have large-effect splicing disruptions in the single tissue they were expressed in (**Figure 2.S5B, viii and vi**). Overall, we consider magnitude instead of sign-concordance, which allows more stringent comparison of splicing changes for specific variants, and that there are some important caveats with this analysis. First, we only use pre-computed PSI values (**STAR Methods**), which cannot account for more complex splicing defects like alternative splice donors or acceptors. Second, the intersection of the two sets of variants are enriched for the most common variants that we call as SDVs, and are likely to be false positives because of the propensity of smaller effect changes in common variants.

To better understand how rare SDVs in ExAC replicate in their full gene context, we assembled 19 SDVs and associated wild-type controls for 12 full-length genes (4 to 13kb in length) using isothermal gene assembly, and examined splicing disruptions using RT-PCR after transient expression of the full gene (**STAR Methods**). We validated that 13 variants in nine genes cause splicing disruptions (**Figures 2.S5C and 2.S5D**) (68.4%, 13/19 variants; or 75.0%, 9/12 genes), with nine of 19 variants (42.1%) having appreciable effects on exon recognition. Interestingly, five of the detected changes involved alternative 5' and 3' splice site usage in the broader full gene context, indicating that many of the identified exon skipping events in MFASS might have different consequences *in vivo*.

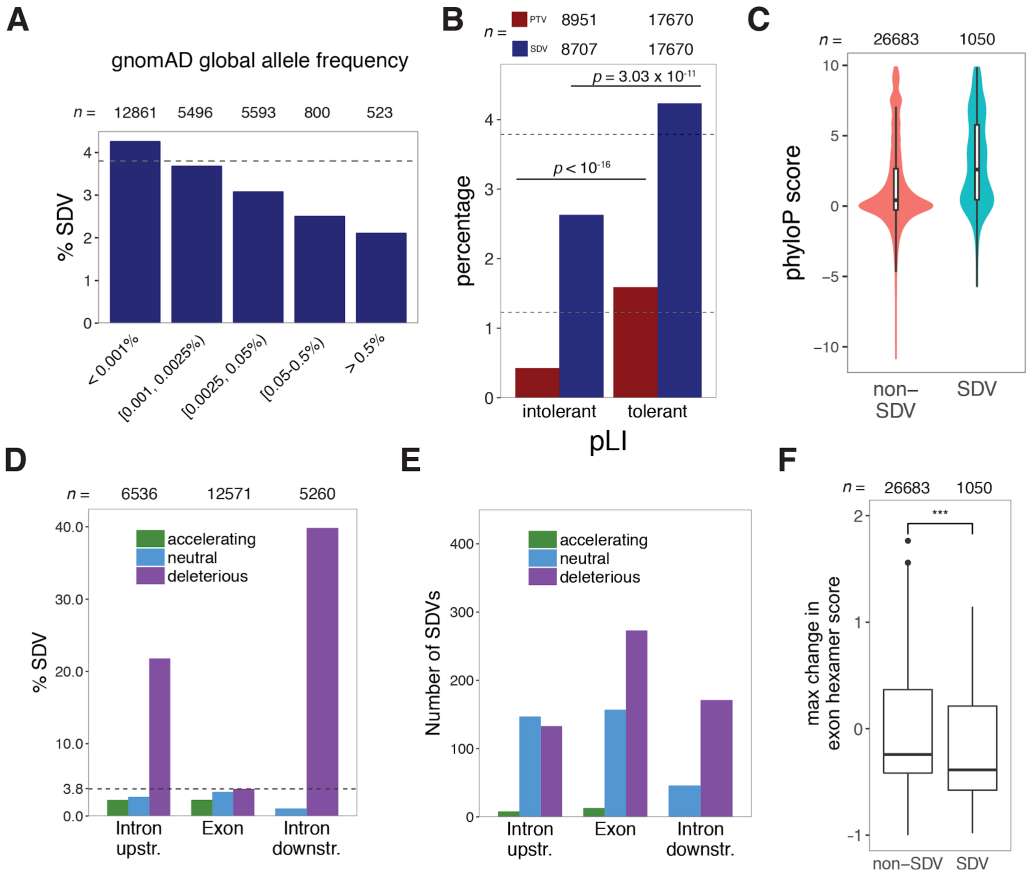


Figure 2.5. Population genetics, evolutionary and functional analyses of splice-disrupting variants (SDVs) across 27,733 ExAC SNVs.

(**A**) The percentage of SDVs as a function of allele frequency shows significant reductions across allele frequencies from the Genome Aggregation Database (gnomAD) (chi-squared test, $p = 1.03 \times 10^{-4}$). A vast majority (97.9%) of the ExAC variants assayed were rare (gnomAD global minor allele frequencies (MAF) ≤ 0.5%). Allele frequencies are not available for 2,460 variants because of insufficient coverage in gnomAD.

(**B**) We analyzed the proportion of SDVs and PTVs in genes predicted to be intolerant to loss-of-function alleles (pLI ≥ 0.9) and tolerant genes. We observe both significantly fewer SDVs (two-tailed Fisher's exact test, $p = 3.03 \times 10^{-11}$) and significant fewer PTVs (two-tailed Fisher's exact test, $p < 10^{-16}$) for exons within intolerant genes. Dashed lines mark the overall percentage of SDVs (3.8%) and PTVs (1.2%) in our dataset without considering the pLI metric.

(**C**) SDVs are under stronger evolutionary conservation as evidenced by higher overall phyloP scores (Mann-Whitney *U* test, $p < 10^{-16}$).

(**D**) Within introns, we find that positions that are evolutionarily conserved (deleterious, phyloP > 2.0, *purple*) have a higher SDV rate than those under neutral (-1.2 ≤ phyloP ≤ 1.2, *blue*) or accelerating selection (phyloP < -2.0, *green*) (two-tailed Fisher's exact test, $p < 10^{-16}$).

(**E**) There are more SNVs outside of regions of high intron conservation, which leads to many SDVs located within nucleotides that display neutral selection.

(**F**) We observed a significantly higher negative maximum change in predicted exonic hexamer scores within exonic SDVs than non-SDVs (Student's *t* test, $p < 10^{-16}$).

See also **Figure 2.S5**

Large-Effect Rare Variants on Splicing are Challenging to Predict

Our results indicate that traditional metrics for assessing how mutations affect splicing are likely to fail, because while it is known that splice site variants are likely deleterious, it has been unclear to what extent rare genetic variation affects splicing outside of these sites. For example, the existing variant effect predictors for missense mutations, such as Polyphen and SIFT, either largely provide no annotation for SDVs or call them benign (**Figures 2.6A and 2.S6A**). Meanwhile, the SDV rate in synonymous mutations, which are usually assumed to be benign, is nearly equivalent to missense variants (3.0% vs 3.1%, **Figure 2.3A**).

We used a number of contemporary variant effect predictors that are capable of predicting the effects of non-coding variation based on both functional genomic and/or evolutionary information (CADD, Kircher et al., 2014), DANN (Quang et al., 2015), FATHMM-MKL (Shihab et al., 2015), fitCons (Huang et al., 2017), LINSIGHT (Gulko et al., 2015), phastCons (Siepel et al., 2005) and phyloP (Pollard et al., 2010), as well as two specifically designed for splicing (SPANR, Xiong et al., 2015) and HAL (Rosenberg et al., 2015) (**Figure 2.6B**). Most predictors have low precision, with several providing no better prediction than random guessing. FATHMM-MKL, CADD, and DANN perform best among those not trained specifically for splicing, but only achieve ~7-8% precision at any appreciable recall. Much of their power is the result of the ability to call intronic SDVs (**Figures 2.6C and 2.S6B**), likely due to increased conservation or molecular function near or at those nucleotides. Not surprisingly, those predictors trained specifically for calling splice defects perform best. At equivalent effect size compared to our assay (>50% splicing disruption), SPANR achieves 44.5% precision, though only a minority of the SDVs are called (11.8%) (**Figure 2.6B**). As we lower the threshold for calling an SDV (i.e., the predicted effect size of an SNV), SPANR can achieve 14.9% precision at 50% recall level, though the predicted effect size is ~2% loss of inclusion. More generally, the SPANR effect sizes poorly predict our observed inclusion

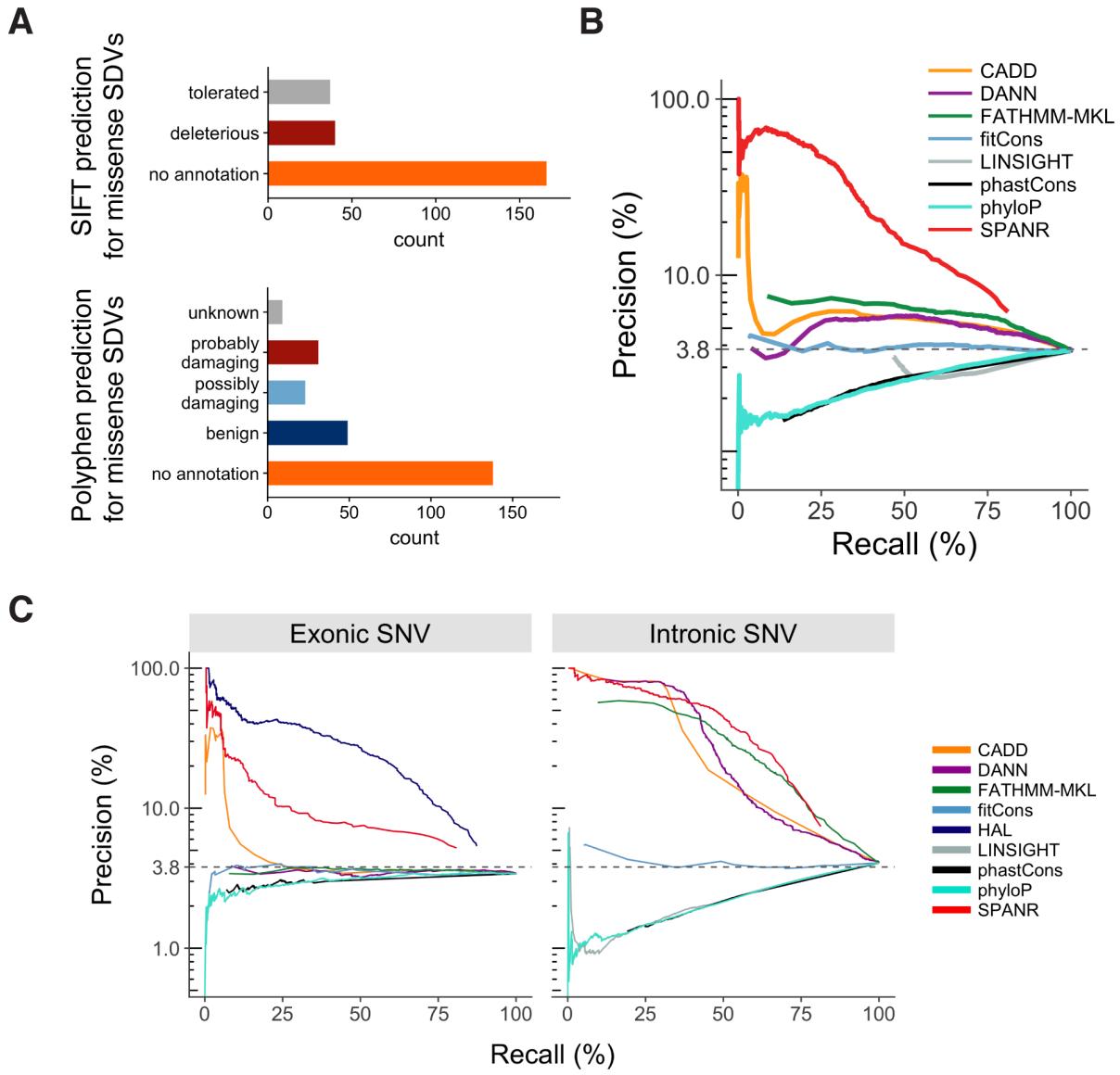


Figure 2.6. Evaluation of genomic and deep-learning predictors for rare variation on splicing. (**A**) Functional prediction from SIFT and Polyphen for missense SDVs (n = 250) show few are predicted to be loss-of-function variants. The distributions for missense non-SDVs for SIFT and Polyphen are shown in **Figure 2.S6A**. (**B**) Precision-recall curves for algorithms that can predict splicing or non-coding genetic variants. Dashed line represents the overall percentage of SDVs (3.8%) from MFASS. Corresponding receiver operating characteristic (ROC) curves are shown in **Figure 2.S6B**. (**C**) Precision-recall curves for algorithms that can predict splicing or non-coding genetic variants, focusing on either intronic or exonic variants only.
See also **Figure 2.S6**.

rates (R^2 = 0.11, **Figure 2.S6C**). The increased power of SPANR over other predictions is largely due to its ability to predict exonic SDVs. HAL provides even better precision in these exonic regions (**Figure 2.6C**), but only calls SNVs within exons.

DISCUSSION

In this work, we tested over half of the variants found in 2,198 human exons across ~60,000 individuals and observed that 3.8% of these variants (1,050 of 27,733) can cause loss of exon recognition. The rate of SDVs we find here is surprisingly high. Our SDV rate (3.8%) is ~73% of the rate of probably damaging variants predicted by PolyPhen for the same set of SNVs (5.2%, 1,437 of 27,733), and ~3-fold higher than the observed rate of protein truncating variants found in ExAC as a whole (1.3%, 121,309 of 7,404,909) (Lek et al., 2016). We would expect such exon skipping events to be detrimental not only to protein function but, if our results generalize to exons that do not preserve frame, also cause large changes to mRNA stability through nonsense-mediated mRNA decay (Lewis et al., 2003). This may help explain why extremely rare variation seems to have large predicted effects on gene expression even though we rarely observe individual mutations with large effects on transcription control elements (Hernandez et al. 2017, Li et al., 2017).

In MFASS, most of the assayed SNVs result in either no effect or near complete exon skipping. These large effect sizes are in contrast with typical effect sizes of sQTLs (GTEx Consortium, 2015, Pala et al., 2017, Takata et al., 2017). We speculate this apparent discrepancy is for several reasons. *First*, MFASS is not well suited to detect small-effect variations due to the limitations of flow cytometry; detecting 10% changes is difficult, unlike in RNA-sequencing. *Second*, we do reproducibly observe smaller differences for some SNVs, but the unnatural context and cell type in MFASS makes it unlikely that small-effect changes we observe actually reflect genuine changes *in vivo*. *Third,* most sQTL studies are done in tissue or blood RNA-seq from heterogeneous cell types. In contrast, in a homogeneous cell type we might expect more bimodal splicing events, such as those revealed in single-cell sequencing studies (Shalek et al., 2013, Faigenbloom et al., 2015). *Fourth,* sQTL studies are usually limited by small sample sizes and

thus are only powered to study common variation, where we would expect few large-effect disruptions to splicing (which we also observe in MFASS). As a vast majority of the variants we assay in MFASS are rare (91% of SNV library, gnomAD MAF <0.5%), we would expect a much larger percentage of large-effect changes. Indeed, many studies of rare variants find large-effect mutations that affect splicing, most notably in GTEx (Li et al., 2017) and in Mendelian diseases (Kremer et al., 2017).

There are a number of technical and biological reasons we may be over- or under-estimating the number of SDVs using MFASS, including the choice of exon set and cell line, ascertainment bias in the experimental workflow, as well as limitations in the reporter assay. *First,* while minigene reporters represent an important standard for the evaluation of clinically relevant splicing mutations, false positives from individual validations of SDVs in their full gene context suggests that minigene reporters do not always capture the necessary context for splicing. *Second*, we only chose exons that are less than 100 bp in length, and start and end at phase 0, due to initial concerns we would only be able to screen in-frame exons using MFASS. We see no appreciable difference in average phastCons conservation in SNVs for in-frame and out-of-frame exons found in ExAC (**Figure 2.S4C**). However, we do find that these constraints do enrich for genes with large repeat expansions such as collagen, where an individual skipped exon is likely to have fewer functional consequences. *Third*, we may not be including enough intronic context to correctly diagnose mutations that will result in SDVs, even though most of the intronic conservation signal was contained within the intron sizes we chose (**Figure 2.S6D**). Because the intronic variation in our genome is on average ~3-fold greater than exonic variation from ExAC, we might be missing a substantial number of SDVs contained within intron regions we do not assay. In addition, because the ExAC consortium is an aggregation of exome sequencing data, surrounding introns have lower coverage and thus fewer covered SNVs. Further development in gene library synthesis may alleviate some, but not all of these issues (Kosuri and Church, 2014; Plesa et al.,

2018). *Fourth*, because the expected number of constructs in each sorted bin is not equal, we may have more power to observe variants for skipped exons. To get an approximate upper-bound on this effect, we found an additional 15,665 SNVs that appear with at least one read in either replicate from our MFASS assay. If we assume these additional SNVs have no effect, we would have an SDV rate of 2.4% (1,050 of 43,398) instead of 3.8% (1,050 of 27,733). *Fifth*, because our reporter can only faithfully report exon skipping when fluorescence is reconstituted, any alternative 5' and 3' splice site usage are false negatives from MFASS, which could still lead to large loss-of-function effects by disrupting protein domains or frame-shifting. Alternatively, SDVs detected by MFASS as exon skipping events might also manifest as alternative splice donors or acceptors *in vivo*, as seen in many of the variants also found in GTEx. *Finally*, we only tested a subset of SDVs for potential cell-type specific splicing regulation. While the SDVs appear to transfer between cell types (**Figure 2.3F**), there may be certain variants that have cell-type specific effects and thus will require MFASS to be conducted in relevant cell types to be detected.

Our results suggest loss of exon recognition by rare human variants may be a major source of functional and expression variation, and their effects are particularly difficult to predict *a priori* using computational prediction. We show most of the large-effect rare variation on splicing would not be easily recognized, as only ~17% of such functional rare variation we found are in canonical splice sites. Compared to other multiplexed splicing reporters, MFASS is unique in that it screens both exonic and intronic variants, uses long constant intron backbones, site-specifically integrates reporters into the same safe-harbor loci at single copy, is applicable to a broad spectrum of human exons, and provides increased power for detecting large-effect loss-of-function variants (Julien et al., 2016; Ke et al., 2011; Rosenberg et al., 2015; Soemedi et al., 2017; Adamson et al., 2018). MFASS is best suited for screening large numbers of large-effect rare variants, which is especially useful for the analysis of mutations in Mendelian diseases, cancer, and population genetics. MFASS is the largest study of splicing defects in SNVs of

natural human exons to date by ~10-fold, and can likely be scaled substantially. More broadly, MFASS, combined with multiplexed assays for variant effects (Gasperini et al., 2016), can help interpret variants found in large exome datasets to obtain a broader understanding for how rare, *de novo*, and somatic variants are shaping complex traits and diseases (MacArthur et al., 2014).

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (5U01HG007912 & DP2GM114829 to S.K., U01HG009417 & R01AG056476 to X.X.), the NIH Biomedical Big Data Training Grant (T32LM012424 to C.P.B.), Searle Scholars Program (to S.K.), Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and Linda and Fred Wudl. We thank Felicia Codrea, Jeffrey Calimlim, and Jessica Scholes (UCLA BSCRC flow cytometry core) and the BSCRC high throughput sequencing core and clinical microarray core for technical assistance; Ron Weiss for the original HEK293T landing pad cell line; Christopher D. Sundberg for the HepG2 cell line; Jason Ernst and Nathan B. Lubock for advice on bioinformatics analysis; Douglas Black and George Church for guidance while developing MFASS. We would also like to thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource.

AUTHOR CONTRIBUTIONS

R.C. and S.K. designed the study. R.C., D.Y. and J.W. developed and validated MFASS. K.D.I. and R.C. developed analysis methods, analyzed and interpreted the data. C.P.B., Y.E.H. and X.X. provided additional bioinformatics analysis. D.B.G. and E.M.J. implemented early prototypes of MFASS. R.C., K.D.I. and S.K. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Sriram Kosuri (sri@ucla.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human cell lines

All cell culture reagents were obtained from Thermo Fisher Scientific. HEK293T chromosomal landing pad cells and derivatives, HepG2 cells, and HeLa S3 cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin, and 0.1 mg/mL streptomycin. K562 cells were cultured and maintained in RPMI supplemented with 10% FBS, 100 U/mL penicillin, and 0.1 mg/mL streptomycin. All cells except K562 cells were passaged using 1x TrypLE Express. All restriction enzymes were obtained from New England Biolabs. Plasmid modifications were performed either by restriction cloning or Gibson assembly (SGI-DNA). Synthesized genes were obtained as sequence fragments from either Gen9 or Twist Biosciences. All oligonucleotides indicated below were obtained from IDT Technologies or Eurofins.

METHOD DETAILS

Splicing reporter design

The organization and key features of our MFASS splicing reporter constructs are as follows: emerald GFP (emGFP) coding sequence is split into two exons that flank a constant intron backbone sequence (**Figures 2.S1A-S1D**). emGFP is split at two different locations for various

reporter designs without disrupting the downstream reading frame. For the SRE library and version 1 of the SNV library, the reporter library contains exons that start and end on phase 1. For version 2 of the SNV library, the reporter library contains exons that start and end on phase 0. The synthetic sequence library is cloned into a pair of restriction sites, AgeI and NheI, or AscI and PacI, in the middle of the backbone. The expression of the splicing reporter module is driven by the CAG-GS promoter. For selection of genomic integrants, we included a Bxb1 attB site and promoterless puromycin such that drug resistance is conferred in the HEK293T cell library following site-specific recombination, due to a CAGGS promoter adjacent to the Bxb1 attP site in the landing pad cell line. We tested two sets of longer constant intron backbones with >250bp of sequence for each intron, which have both been previously characterized as more faithful intron backbones in the context of such three-exon, two-intron reporters (**Figures 2.S1A-S1D**). These two backbones were the *C. griseus* long DHFR intron backbone (Arias et al., 2015) and human SMN1 intron backbone (Cho et al., 2015) (**Figures 2.S1A-S1D**). In particular, the long DHFR introns were the same introns used in previous characterizations of exon definition (Arias et al., 2015).

Microarray-derived oligonucleotide library design

We obtained microarray-derived oligonucleotides of 200 to 212 bp from Agilent Technologies to generate synthetic DNA libraries. We selected human exons that are less than 100 bp and begin and end on frame 0 from the Ensembl mySQL server (Aken et al., 2016). We designed a 170-bp intron-exon-intron sequence library *in silico* containing all 9,634 human exons fulfilling above criteria (Ensembl release 73, hg19 assembly), which includes at least 40 bp of upstream intron and at least 30 bp of downstream intron, with the exon in the middle. We added extra native intronic sequences as length limitations allowed (i.e., if exons were shorter), split between the upstream and downstream equally with an extra base added to the donor side for odd number of bases added. Finally, a pair of 15-mer amplification primer sequences, containing either AscI and

PacI or AgeI and NheI restriction sites, were added to yield 200-mer or 212-mer sequences for DNA synthesis respectively for the SRE or SNV libraries.

Design of SRE library

For the SRE library, we obtained 9,634 human exons that are less than 100 bp and begin and end on frame 0, and designed a 170-bp exon library with its surrounding intronic contexts, that includes at least 40 bp of upstream intron and at least 30 bp of downstream intron. Overall, we randomly chose 230 exons from this set and designed 60-80 synonymous mutations per sequence that correspond to specific functional classes of regulatory elements governing splicing using a toolkit of custom Python scripts we developed for scoring these mutations using defined scoring criteria as detailed below. We focused on three major motif types related to splicing in our custom scoring algorithm (**Table 2.S1**). The first major motif type is the splice acceptors and donors. These sequences are scored with MaxEntScan (Yeo and Burge, 2004), an algorithm based on the maximum entropy principle that learns splice site motif strength. The second major motif type is the exonic splicing enhancers/silencers (ESEs/ESSs) based on the results from Ke *et al.* (Ke et al., 2011). The third major motif type is the conserved intronic sequences that affect splicing in either the acceptor or donor side of the intron (Voelker and Berglund, 2007). Next, we iteratively designed synonymous mutations in exons and/or introns that affect splicing (**Table 2.S2**). Mutations made to sequences were scored in the same fashion as wild-type sequences, with a higher score as a proxy for increased exon inclusion. Mutations were scored and generated to weaken, strengthen or destroy splicing motifs. We define functional classes of mutants that differ in score requirements, minimum base separation between mutants and the number of mutants per class. Mutations were made iteratively until we generate the desired number of mutants or reach the maximum number of iterations. For splice sites or splice regions, the invariant positions of the splice donor or acceptor are not mutated, with the exception for the "weaken splice site" category. In addition, we tested 53 RNA-binding protein motifs obtained from

the RNA-binding protein database (RBPDB) (Cook et al., 2011) as position frequency matrices and thresholded at 1% false positive rate, and 109 human single-nucleotide polymorphisms (SNPs) obtained from dbSNP (build 133) (Smigielski, 2000).

Design of SNV library

For the SNV library, we started with a library of 2,920 natural exons that exhibited exon inclusion using MFASS (inclusion index ≥ 0.8; SRE library, DHFR intron backbone). We designed single nucleotide variants (SNVs) from the Exome Aggregation Consortium (Lek et al., 2016) (ExAC, version 0.3.1). We stored hg19 genomic coordinates of each sequence in BED file format, and used bcftools to intersect the ExAC variants with our library of wild-type human exons to subset all relevant SNVs. We only synthesized variants with a filter status of "PASS", and generated all alternate alleles (up to 3) if more than one alternate allele was indicated. These sequences were filtered to (i) exclude sequences containing unique NheI or AgeI restriction sites used for library cloning and (ii) include SNVs only within nucleotides 11 through 160 of each 170 bp library sequence to avoid possible spurious interactions with restriction sites, resulting in 2,902 exons as template with their associated variants that fit above criteria.

We designed two library subpools with redundancy for wild type that enables separate retrieval of sublibraries from the microarray. We transfected these pools at the stage of plasmid reporters at the ratio of 1:3 that enables increased representation of natural sequences. From the initial design carried through to the completion of MFASS, 80.5% of the designed natural sequences (2,339 of 2,902) were represented in the final cell reporter library. 2,198 out of 2,339 natural sequences have at least one corresponding SNV, while an additional 30 sequences represented in the control library. Ultimately, we only report and include SNV data for which data for natural sequences are available, have replicable data across two biological replicates, and have an inclusion index of greater than or equal to 0.5 for wild-type. For these 2,198 exon

backgrounds, we obtained the corresponding paired variant data for 27,733 SNVs, from which 1,050 SDVs are observed (**Figures 2.3A and 2.4A**).

Library amplification and cloning

The splicing regulatory element (SRE) library was amplified with KAPA HiFi HotStart (KK2701) in eight 50 µL reactions, each with 500 pg of oligonucleotide library, and 0.4µM of ORC405 and ORC406 primers. The reaction and cycling conditions are: 95°C for 3 minutes, 5 cycles of 95°C for 3 seconds, 50°C for 20 seconds, 60°C for 10 seconds, 15 cycles of 95°C for 3 seconds, 60°C for 30 seconds, followed by an extension of 60°C for 5 minutes. The SRE library was amplified similarly as above with ORC403 and ORC404 primers, as well as the following cycling conditions: 95°C for 3 minutes, 5 cycles of 95°C for 3 seconds, 50°C for 20 seconds, 60°C for 10 seconds, 11 cycles of 95°C for 3 seconds, 60°C for 30 seconds, followed by an extension of 60°C for 5 minutes. Splicing reporter plasmids and SRE library were digested with AscI and PacI. Reporter plasmid and library were ligated with T4 DNA ligase (New England Biolabs).

For the SNV library, we performed similar procedures as above with the following alterations: we performed emulsion PCR for the two subpools (35 cycles) containing both natural exons and SNVs with biotinylated primers. The second subpool was amplified similarly (40 cycles), with biotinylated ORC513 and ORC514 primers, and both pools were processed with AgeI and NheI at 37°C before ligation-based cloning in *E. coli*.

Generation of landing pad cell lines and integration

For site-specific integration of exon libraries in HEK293T cells, we engineered a chromosomal landing pad cell line which allows stable expression of splicing reporter library at the AAVS1 locus, which is modified from Duportet et al. by CRISPR-Cas9 in order to remove expression of the endogenous YFP gene (Duportet et al., 2014). We characterized 25 clones expanded from

single cells by flow cytometry, microscopy and genomic PCR, and selected a clone (which we termed RCA7) that does not express any YFP or mCherry fluorescence for our current study.

We site-specifically integrated the splicing reporter using Bxb1 integrase into cells containing the chromosomal landing pad (**Figures 2.1 and 2.S1**), first without any exon library sequences between the intron backbones, and later with individual exons and/or synthetic sequence libraries cloned in between. For the SRE library, we transfected HEK293T chromosomal landing pad cells, grown in six T-225 flasks (BD) per biological replicate that were processed in tandem. Each T-225 flask was transfected at 80% confluency with 50 µg of plasmids containing exon library and Bxb1 integrase, and 150 µL Polyethylenimine (Polysciences Inc.) or 75 µL Lipofectamine 3000 (Thermo Fisher Scientific). Cells were transfected for 72 hours, and then selected with 5 µg/mL puromycin (Thermo Fisher Scientific). Cells were subsequently passaged serially for at least 18 days before cell sorting. For the SNV library, we transfected HEK293T chromosomal landing pad cells, grown in sixteen 150 cm^2 plates (45 µg plasmids per plate) for three days, pooled and transferred to two 4500 cm^2 roller bottles (BD Biosciences) or equivalent volume for 150 cm^2 plates per biological replicate, selected for integrants as above, and maintained in eight 150 cm^2 plates per biological replicate for 20 days before cell sorting.

Fluorescence-activated cell sorting

We measured cell samples for GFP and mCherry fluorescence intensities by flow cytometry (BD LSRFortessa or LSRII) across passages. Cells harboring variant libraries were sorted using a FACSAria III (BD Biosciences) into bins based on GFP fluorescence, given a minimal amount of mCherry fluorescence (threshold set using a genome-integrated mCherry driven by the pCAGGS promoter as a positive expression control, **Figure 2.1A**). For the SRE library (DHFR intron backbone), we sorted ~7.5 million cells for GFP_+ and GFP_{neg} bins, and 7.5 x 10^5 cells for GFP_{int} bin. For the SRE library (SMN1 intron backbone), we obtained ~4 million cells for GFP_+

and GFP_{neg} bins, and 4.2 x 10^5 cells for GFP_{int} bin. Sorted sub-libraries for each replicate were grown separately and passaged. We eliminated dead cells, debris, and doublets based on forward and side scatter, and single-color and double-negative controls were used for gating and calibration. For the SNV library (v1), we performed two sorts to ensure purity of the final populations of GFP_{+}, GFP_{int} and GFP_{neg} cells (**Figure 2.S3A**). For the first sort, we obtained 16 million cells for GFP_{neg} library, 2.6 million cells for GFP_{+} library and 2.7 million cells for GFP_{int} library (biological replicate 1), 15 million cells for GFP_{neg} library, 2 million cells for GFP_{+} library and 2.8 million cells for GFP_{int} library (biological replicate 2). For the purifying sort, we further sub-sorted the libraries from the first sort, and obtained ~2 million cells for GFP_{neg} library, 1 million cells for GFP_{+} library and 2.5 million cells for GFP_{int} library (biological replicate 1), and 1 million cells for GFP_{neg} library, 1 million cells for GFP_{+} library and 2.5 million cells for GFP_{int} library (biological replicate 2).

For the SNV library (v2), we sorted cells based on GFP fluorescence into four bins: GFP_{+}, $GFP_{int\text{-}hi}$, $GFP_{int\text{-}lo}$, and GFP_{neg} bins (**Figure 2.S3B**). For both biological replicates, we obtained 16 million cells for GFP_{neg} library, 2 million cells for GFP_{+} library, 2 million cells for $GFP_{int\text{-}hi}$ and $GFP_{int\text{-}lo}$ library.

DNA-Seq of FACS-sorted libraries

To obtain cells containing a single individual reporter construct, we first sorted single cells by FACS from individual bins, with GFP fluorescence gates defined from library sort, and expanded homogeneous clones from single cell sort. For the SRE library, we extracted genomic DNA from 10 million cells for the sorted populations using blood and cell culture DNA midi kit (Qiagen). We amplified each sublibrary for ~300-fold amplicon coverage, and reactions were performed in 96-well format in three to nine 50 µL reactions for each sublibrary proportional to bin size. Per biological replicate, we amplified library variants from genomic DNA with KAPA HiFi HotStart,

using 5 µg of template for GFP_+ and GFP_{neg} sub-libraries, and 2 µg of template for the GFP_{int} sublibrary, with 500 nM of the primers ODY093 and ODY028 for the DHFR intron backbone, or the primers ODY088 and ODY089 for the SMN1 intron backbone. The following cycling conditions were used: for the DHFR intron backbone, 98°C for 45 seconds, 23 cycles for GFP_{int}, or 22 cycles for GFP_+ and GFP_{neg} using: 98°C for 15 seconds, 68°C for 30 seconds, 72°C for 30 seconds, followed by an extension of 72°C for 1 minute; for the SMN1 intron backbone: 98°C for 45 seconds, 24 cycles for GFP_{int}, or 29 cycles for GFP_+ and GFP_{neg} of: 98°C for 15 seconds, 68°C for 30 seconds, 72°C for 30 seconds, followed by an extension of 72°C for 1 minute. The reactions for each population were pooled separately, purified and gel-extracted on 1% agarose gel and quantified using Tapestation 2200 (Agilent).

For the SNV library, procedures were performed similarly to the SRE library in the DHFR intron backbone, with the following optimizations. Library variants was amplified from genomic DNA (ORC515 and ODY028), and genomic DNA was extracted similar to procedures for the SRE library. Sorted libraries were indexed by PCR amplification, in twenty-four 50 µL reactions for GFP_{neg} and eight 50 µL reactions for all other sublibraries, using the forward primer ORC522, and the reverse primers ODY32 through ODY41, and ORC531 through ORC534.

Validation of MFASS using individual exon controls

We performed individual controls to assess the correspondence to sequences in our library and to observe consistent splicing behavior across RNA and fluorescence output. For the data from Figures S1N through S1Q, we characterized more than 20 cell clones expanded from single cells, and only 9 individual sequences that perfectly match the reference SRE library were used for RT-PCR and flow cytometry analysis.

RNA from sorted sub-libraries as well as individual control exons were extracted using RNEasy MiniKit (Qiagen). Reverse transcription-PCR was performed using Superscript III or Superscript IV (Thermo Fisher Scientific) according to manufacturer's protocol using reverse transcription primer (**Table 2.S4**), which binds to a region in exon 2 of emGFP, and PCR was performed with extracted cDNA. The reaction and cycling conditions are optimized as follows: 95°C for 2 minutes, 18 cycles of 98°C for 3 seconds, 62°C for 15 seconds, 72°C for 10 seconds, followed by an extension of 72°C for 2 minutes.

34 SDVs were tested for exon inclusion by transient transfection using Lipofectamine 3000 (Life Technologies) in HEK293T cells for 24 hours. A ratio of GFP:mCherry fluorescence was obtained in linear mode (BD LSRII or BD LSRFortessa) for the comparison of exon inclusion rates across samples. We subtracted background fluorescence based on a transfected empty vector control, and only consider GFP:mCherry fluorescence above the threshold. We tested sequences either exactly in the original sequence context in the reporter construct examined in MFASS, or with an additional 130 bp of endogenous intronic contexts (65 bp upstream and 65 bp downstream). Percent inclusion is calculated for both the individual SDV and its respective wild-type sequence, with the change in percent inclusion calculated as the absolute difference between the mutant and the wild-type sequence. All mutants were normalized to a no-insert control as a baseline for complete exon skipping for assessment of change in exon inclusion.

Cell-type specificity of SDVs across four cell types

We tested 29 human exons with its surrounding intronic contexts (15 SDVs with the 14 corresponding wild-type sequences) across 4 human cell types. The four human cell lines tested are HEK293T (RCA7 cell line established in this study), HeLa S3 (ATCC CCL-2.2), HepG2 (ATCC HB-8065) and K562 (ATCC CCL-243). We validated these constructs across cell types in the same manner that we validated individual exon controls in above section.

Validation of rare SDVs in full genes

We considered rare 61 SNVs in 34 genes that have a change in inclusion index of ≤ -0.50 across both replicates from MFASS (i.e. SDVs) under 15kb. From these, we were able to assemble complete 12 wild-type full genes (up to ~13kb) with at least one corresponding SDV (19 SDVs total, **Figures 2.S5C and 2.S5D**). Using isothermal gene assembly, mutations were introduced in the middle of the oligonucleotide with ~40bp overlap on each overlapping fragment, and assembled without the mutations for the wild-type gene sequences. Genomic sequences with wild-type and matched SNVs were amplified from the same human genomic DNA template (NIST, SRM 2372, or Promega, G1521) using PrimeSTAR GXL polymerase (R050, Takara). Each partial gene fragment was amplified using 25ng of genomic DNA in a single 50μL PCR reaction, and purified with either the DNA Clean and Concentrator Kit (Zymo Research) or Agencourt AmPURE XP beads (Beckman Coulter). The reaction and cycling conditions are optimized as follows: 94°C for 1 minute, 28 to 30 cycles of 98°C for 10 seconds, 68°C for 5 minutes, followed by an extension of 72°C for 5 minutes. A linear plasmid backbone fragment (~5.2kb) was prepared for isothermal assembly using BamHI and SacI, purified and concentrated using DNA Clean and Concentrator Kit (Zymo Research), and further gel purified using Zymoclean Gel Recovery Kit (Zymo Research). We expressed a subset of these fully assembled genes between the BamHI and SacI sites of the splicing reporter plasmid backbone in this study, in place of the MFASS splicing reporter (see Splicing Reporter Design section). We performed isothermal assembly of 3 to 4 gene fragments of interest and the plasmid backbone using the Gibson Assembly Ultra Kit (SGI-DNA), and transformed into electrocompetent DH10B *E. coli* cells (New England Biolabs, or Life Technologies) to select for correct gene assembly. We confirmed the sequence for each gene with or without splice-disrupting variants using Sanger sequencing, before transfection into HEK293T cells for testing of mutation effects. We extracted and performed reverse transcription from RNA using the Cells to cDNA II kit (Thermo Fisher Scientific) and corresponding gene-specific primer for each exon

(**Table 2.S4**) according to manufacturer's protocol. For each tested exon, qPCR was performed with SYBR FAST qPCR Mastermix (Kapa Biosystems), using 1μL of reverse-transcribed cDNA in a 20μL PCR reaction, as well as primers flanking the upstream and downstream exons, and compared RT-PCR gene
products of wild-type and mutant sequences for each gene of interest. Fragments of interest were further PCR purified and verified using Sanger sequencing.

QUANTIFICATION AND STATISTICAL ANALYSIS

DNA-Seq read processing and filtering

SRE library datasets were generated from two Illumina MiSeq 300-bp paired-end sequencing runs and a Illumina HiSeq 2500 150-bp paired-end sequencing run. SNV library version 1 dataset was generated from Illumina MiSeq 300-bp paired-end sequencing. SNV library version 2 dataset was generated from Illumina NextSeq 2500 150-bp paired-end sequencing. We removed read pairs with any ambiguous "N" base calls, followed by read pair merging with *bbmerge* from the BBMap suite (BBtools package version 37). We developed custom Python and bash scripts to filter for perfect reads aligned to our reference, from which we can aggregate read counts for sequences from each sorted bin. We then further process these read counts to calculate inclusion index (see below section on the quantification of inclusion index).

To allow for stringent analysis of replicable data for SNVs, we require a coverage of at least 5 reads for the SRE library and at least 10 reads across all bins for the SNV library for the two biological replicates. Our SRE library size was 16,717 (5,975 wild-type sequences, 10,683 mutants, 59 controls) for the SMN1 intron backbone, and 13,922 (4,920 wild-type sequences, 8,942 mutants, 60 controls) for the DHFR intron backbone. We additionally require that inclusion indices agree between biological replicates within 0.30 (SRE library) and 0.20 (SNV library). For the SNV library, we only analyzed a mutant sequence if its corresponding wild-type sequence

has an inclusion index of ≥ 0.5. The final library size after all filtering steps for the SRE library is 10,482 (3,714 wild-type sequences, 6,713 mutants, 55 controls). The final library size after all filtering steps for the SNV library size (version 1) is 6,768 (1,981 wild-type sequences, 3,853 mutants, 934 controls). The SNV library size (version 2) is 31,144 (2,339 wild-type sequences, 27,733 mutants that correspond to 2,198 wild-type sequences, 1,072 controls).

Exon inclusion quantification

We normalized bin counts based on read depth (reads per million, RPM) and corresponding bin population percentage after FACS using the following formula:

$$Normalized\ read\ count\ GFP_{bin,i} = \frac{percentage\ sorted \times raw\ read\ count\ GFP_{bin,i}}{reads\ per\ million}$$

We calculated exon inclusion index for each sequence based on a weighted average of normalized counts across all bins. Bin weights are assigned based on GFP fluorescence measurements of individual bins that correspond to the extent of exon inclusion or skipping. For the splicing regulatory element (SRE) library and single nucleotide variant (SNV) library, version 1:

$$\frac{(0\times GFP_{+}) + (0.85\times GFP_{int}) + (1\times GFP_{neg})}{GFP_{+} + GFP_{int} + GFP_{neg}}$$

For the SNV library, version 2:

$$\frac{(0\times GFP_{+}) + (0.80\times GFP_{int-hi}) + (0.95\times GFP_{int-lo}) + (1\times GFP_{neg})}{GFP_{+} + GFP_{int-hi} + GFP_{int-lo} + GFP_{neg}}$$

The change in inclusion index for an individual library sequence between wild-type (WT) and mutant is computed as follows:

$$\Delta\ inclusion\ index\ =\ inclusion\ index_{mutant} - inclusion\ index_{WT}$$

A positive Δinclusion index denotes increased exon inclusion for the mutant relative to WT, while a negative Δinclusion index denotes increased exon skipping for the mutant relative to WT.

ExAC and gnomAD data analysis

Annotation of variants for individual human samples in VCF format were obtained from the Exome Aggregation Consortium (Lek et al., 2016) (ExAC, version 0.3.1), including global allele frequencies. We further obtained global allele frequencies of individual variants from the Genome Aggregation Database (gnomAD). We binned gnomAD global allele frequency similar to the ExAC study (Lek et al., 2016), and tested for significant difference between allele frequency bins using chi-squared test of independence. We obtained the rate of protein-truncating variants from ExAC. We also obtained gene level evolutionary constraint estimates from ExAC based on probability of loss-of-function intolerance (pLI), and defined genes that are extremely intolerant of loss-of-function as those with a pLI score ≥ 0.9. We then tested for genes with enrichment in splice-disrupting variants (SDVs) using Fisher's exact test.

Functional genomic analysis of SNVs

We functionally classified our variants using the Ensembl variant effect predictor (McLaren et al., 2016) (VEP v80), and filtered the most severe sequence ontology (SO) term for a given variant. We obtained phyloP 100-way (v1.4) nucleotide conservation for the hg38 genome for the SNV library, and classified quickly evolving regions of the genome (accelerating, phyloP < -2.0), neutral selection (-1.2 ≤ phyloP ≤ 1.2) and highly conserved region of the genome (deleterious, phyloP > 2.0). To compute genome-wide locations of ExAC SNVs by gene regions, we used GENCODE (Harrow et al., 2012) (release 27, GRCh38 reference assembly) for exon annotation, and bedtools (Quinlan, 2014) to annotate intronic regions by subtracting exon coordinates from gene coordinates. To determine the density of SNVs for each genomic position, we determined the number of SNVs averaged at each relative position for the SNV library across exons and upstream/downstream introns, and relative position is set such that the boundary of upstream

intron/5' exon = 0, and the boundary of 3' exon/downstream intron boundary = 1. In addition, we incorporated scaled positions to normalize for variable intron and exon lengths. We performed similar positional SNV density analysis for genome-wide SNVs from the ExAC consortium across gene regions.

Motif analysis

To define potential disruption of *k*-mer motifs by ExAC SNVs, we performed *k*-mer based motif enrichment analysis using *k*pLogo (git/e2fac18) for both splice acceptor (positions -20 to +3, upstream *intron-exon* junction) and splice donor (positions -3 to +6, downstream *exon-intron* junction). Based on our SNV dataset, SDVs are background-corrected against non-SDVs to obtain motif logos that are enriched or depleted at each nucleotide. We used a *p*-value cutoff of $p < 0.01$, gapped *k*-mer length of $k = 1,2,3,4$ and fixation frequency of 0.75 (Wu and Bartel, 2017).

We scored splice acceptors and donors at the consensus positions, same as above, using MaxEntScan (Yeo and Burge, 2004), an algorithm based on the maximum entropy principle than learns splice site motif strength (**Table 2.S1**). We scored exonic splicing enhancers/silencers (ESEs/ESSs) based on work from Ke et al. 2011 and scored conserved acceptor and donor intronic sequences based on work from Voelker and Berglund, 2007.

In addition, we implemented the hexamer additive linear (HAL) model, which estimates a splicing strength score for every possible exon hexamer (Rosenberg et al., 2015). A positive score indicates the hexamer is more likely to activate nearby splice sites, and a negative score indicates the hexamer is more likely to silence nearby splice sites. For each variant, we calculated the change in score at each position relative to the wild-type sequence, and identified the maximum change in score. We compared the distribution of maximum score change between SDVs and non-SDVs using the Mann-Whitney *U* test, which does not assume normality.

Assessment of variant prediction algorithms

To computationally predict the effects of rare genetic variants on splicing, we used various prediction algorithms that are able to assess coding and/or non-coding SNVs in our assay. For the purpose of method comparison, we selected Δinclusion index ≤ -0.5 as the threshold for splice-disrupting variant (SDV) and designate our calls as true positives. We assessed performance by varying the score threshold at which a variant is called splice-disrupting (considering whether the score is positively or negatively correlated to Δinclusion index). We assessed various genomic predictors that use a variety of machine learning methods, annotations, and training sets to predict the functional impact of coding and non-coding variants. These methods incorporate a variety of functional data, including conservation, histone modifications, DNase hypersensitivity, transcription factor binding, transcript abundance, and protein-level scores.

We obtained functional scores of single nucleotide variants from four genomic predictors based on the hg19 assembly: raw CADD scores from CADD v1.3 (r0.3 Exome Aggregation Consortium dataset), DANN whole-genome SNV scores (Nov. 2014), FATHMM-MKL (git/908d865), fitCons multi-cell (i6 dataset, git/20f336d) highly significant scores (p < ~0.003), and LINSIGHT (git/58fe558). For SPANR (splicing-based analysis of variants) (Xiong et al., 2015), we obtained the predicted change in percent spliced in (Δψ, or ΔPSI) for single nucleotide variants in our SNV library across the genome. The hexamer additive linear model (HAL) (Rosenberg et al., 2015) can only assess exonic variants.

To consider the predictive power of conservation alone, we obtained phyloP 100-way (v1.4) nucleotide conservation for the hg19 genome for the SNV library. In addition, we obtained phastCons (v1.4) scores for 100-way eutherian mammalian nucleotide conservation for our SNV library and genome-wide SNVs from the ExAC consortium (Siepel et al., 2005). To assess the

functional effects of missense, exonic single nucleotide variants from the SNV library, we used variant annotations from PolyPhen (v2.2.2) and SIFT (v5.2.2).

We assessed above predictors using receiver operating characteristic and precision-recall analysis. We used the pROC package version 1.10.0 to compute and plot the ROC curves, calculate the 95% confidence interval, and calculate the area under the curve. The precision recall curves were plotted with a custom function which evaluates each method by varying the score threshold at which a sequence is classified as an SDV, and calculating the corresponding precision and recall. The area under the precision recall curve is calculated with the trapz function in R.

Analysis of SDVs from GTEx RNA-Seq

Genotype data (from Illumina SNP arrays, whole exome sequencing, or whole genome sequencing) and RNA-Seq data were obtained from the GTEx database (v6p release, GTEx Consortium, 2015). To get a list of high-quality SNVs for further analyses, we used a quality filter of GQ ≥ 20 for whole-genome sequencing and whole-exome sequencing and a quality filter of IGC ≥ 0.2 for Illumina SNP arrays, all of which were provided by GTEx. These cutoffs are similar as recommended by the GATK package (DePristo et al., 2011; Van der Auwera et al., 2013). In addition to the genotyped SNPs, we also identified dbSNPs (version 146) that are expressed in the RNA-Seq data by requiring a minimum total read coverage of 10 and a minimum read coverage of 3 for the alternative allele (Zhang and Xiao, 2015).

The RNA-Seq data in FASTQ format were first adaptor-trimmed using Cutadapt (Martin, 2011). Subsequently, the reads were aligned to the hg19 genome and transcriptome (Ensembl Release 75) using HISAT2 (Kim et al., 2015) with parameters --mp 6,4 --no-softclip --no-mixed --no-discordant. Only uniquely mapped read pairs were retained for further analyses. Samples

with fewer than 25 million uniquely aligned read pairs were excluded due to low depth for splicing analysis. In total, 7822 RNA-Seq datasets from 47 tissues and 515 donors were retained.

Percent-spliced-in (PSI) values were calculated using the method described in Schafer et al. (Schafer et al., 2015). This analysis was carried out for all internal exons from the GENCODE comprehensive annotation (v24lift37). To ensure the accuracy of PSI estimation, we required the exons to be covered by ≥ 15 total reads (inclusion reads + exclusion reads) or ≥ 2 exclusion reads per sample (Barbosa-Morais et al., 2012; Hsiao et al., 2018).

We compared PSI values from tissues expressing the gene containing an SDV with a cutoff of transcript per million (TPM) ≥ 1 based on median gene TPM values. After filtering on expression, exon PSI for 28 SDVs (out of 1,050 ExAC SDVs in this study) were available in at least one tissue sample. The distribution of PSI values across tissues was compared for individuals with the alternative SDV alleles versus those with the corresponding reference alleles. Comparisons were made with the Mann-Whitney *U* test, and adjusted *p*-values were calculated using the Benjamini-Hochberg procedure at an FDR of 5%.

Gene ontology enrichment analysis

We performed Gene Ontology (GO) enrichment analysis using the topGO package in Bioconductor (Alexa, A. and Rahnenfuhrer, J. 2016) between SDV-containing genes (n = 473, for 1,050 SDVs) and all genes in the ExAC SNV library (n = 1,616, for 27,733 SNVs). We determined over-representation of GO terms for SDV genes based on gene counts using Fisher's exact test. Each GO term is tested independently and only terms with $p < 0.01$ are shown (see **Table 2.S3**).

Software

bbmerge from the BBMap suite (v37) was used to merge raw paired-end sequencing files. Custom python and bash scripts used for read processing, and mapping reference and synthetic error read counts. Further analysis was performed with Python 2.7, using Pandas v0.21.0 and Numpy v1.13.3, and R v3.4.2, using tidyverse including dplyr v0.7.4 and ggplot2 v2.2.1. Variant analyses were performed using Ensembl variant effect predictor (v80), CADD (v1.3), MaxEntScan (Yeo and Burge, 2004), DANN (https://cbcl.ics.uci.edu/public_data/DANN/), FATHMM-MKL (git/908d865), fitCons (i6 dataset, git/20f336d), HAL (git/ca54d11), *k*pLogo (git/e2fac18), LINSIGHT (git/58fe558), phastCons (v1.4), phyloP (v1.4), PolyPhen (v2.2.2), SIFT (v5.2.2), and SPANR/SPIDEX (v1.0) (http://annovar.openbioinformatics.org/).

DATA AND SOFTWARE AVAILABILITY

Sequencing data were deposited into the Gene Expression Omnibus (GEO) under the accession number GEO: GSE120695. Pre-processed data sets are available upon request. All code needed to reproduce the analyses is included in the following repository: https://github.com/KosuriLab/MFASS

SUPPLEMENTAL INFORMATION

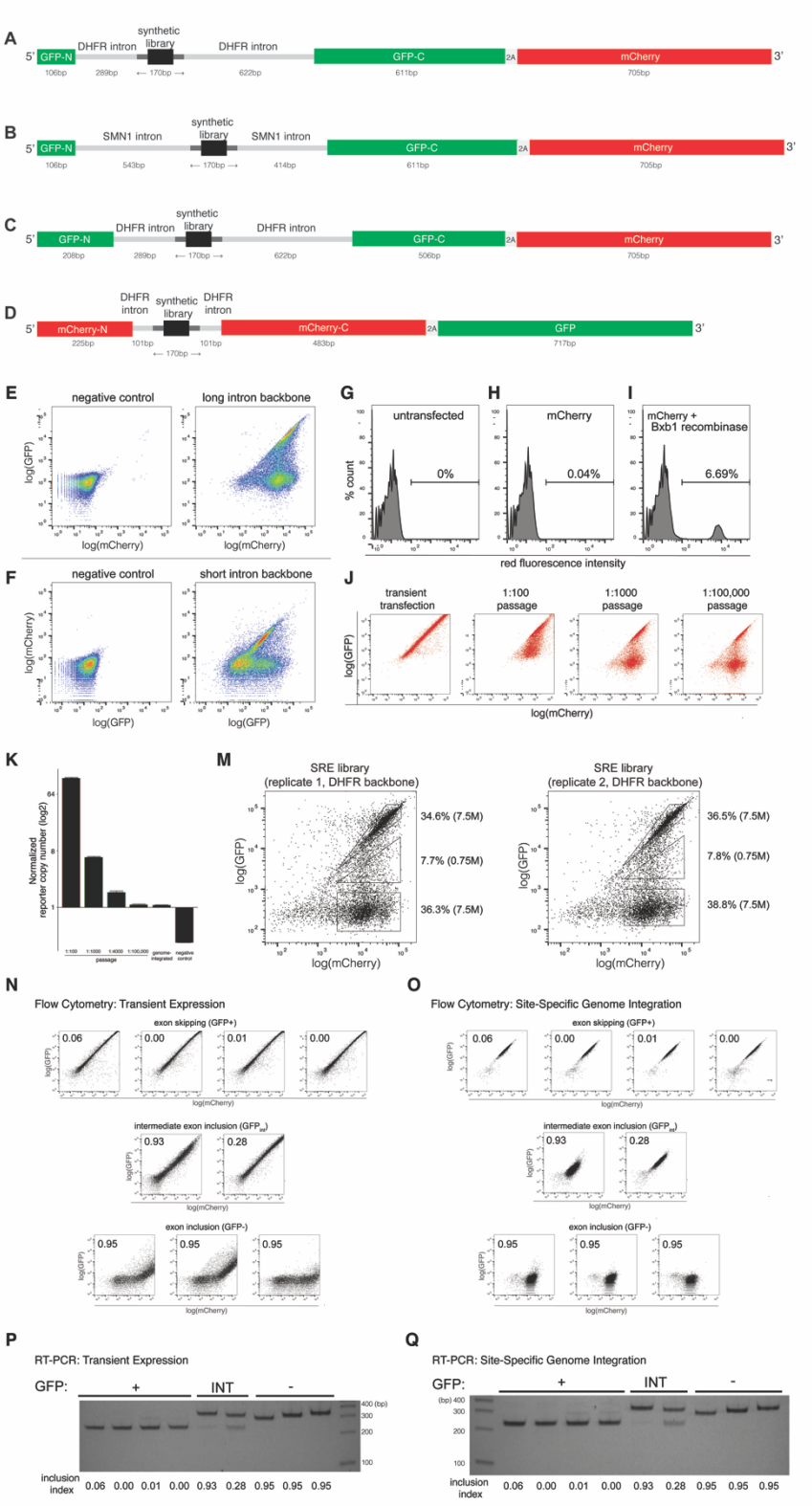


Figure 2.S1. MFASS reporter design, workflow optimization and testing. Related to Figure 2.1.

(**A**) The Split-GFP construct with the DHFR constant intron backbone (289 bp upstream, 622 bp downstream) is used for both the SRE and SNV libraries.

(**B**) The Split-GFP construct with the SMN1 constant intron backbone (543 bp upstream, 414 bp downstream) is used for the SRE library.

(**C**) The Split-GFP construct with DHFR constant intron backbone (289 bp upstream, 622 bp downstream), located in a different position along the split-GFP, is used for the SNV library.

(**D**) The split-mCherry construct with constant DHFR short-intron backbone (101 bp upstream, 101 bp downstream) was a prior iteration of the reporter that displayed significant intron retention (**Figure 2.S2B**). The synthetic library (*black*) contains human exons with surrounding native intron contexts (*dark gray*). The constant intron backbones are colored in light gray.

(**E** and **F**) Flow cytometry of splicing regulatory element (SRE) library expressed in two splicing reporter constructs with different lengths of constant intron backbones. (**E**) Flow cytometry results for the standard DHFR backbone (**Figure 2.S1D**) used in the SRE and SNV datasets of the landing pad cell line before (*left*) and after (*right*) to integration of the SRE library. (**F**) Flow cytometry results of the same library as in a, but integrated into the short-intron DHFR backbone (as in **Figure 2.S1D**). We see far less expression, and much larger double-fluorescence negative population possibly indicating increased intron retention.

(**G**, **H** and **I**) High-efficiency integration of large splicing reporter libraries in human cells. We monitored integration efficiency by comparing genome integration of genetic packages transfected (**H**) without and (**I**) with Bxb1 serine integrase, compared to (**G**) Untransfected control cell line. ~6.7% of mCherry plasmids are site-specifically integrated into the genome. The middle panel includes a promoter-less mCherry plasmid without the Bxb1 integrase, serving as a control for non-specific integration across tissue culture passages.

(**J** and **K**) Evaluation of genome-integrated library across cell culture passages. (**J**) Flow cytometry analysis of the SRE library at various cell culture passages. We characterized reporter copy number per cell in detail, after library transfection, across cell culture passages. Only at 1:100,000 passage do we lose detectable levels of plasmid remaining within the cells.

(**K**) We performed quantitative PCR to determine normalized copy number of reporter constructs per cell across cell culture passages. Copy number for each sample is normalized to that of the genome-integrated cell line containing a single construct. Landing pad HEK293T cell line that does not contain any reporter construct serves as negative control. Error bar indicates standard error of the mean. We see that at the 1:100,000 passage, we fail to detect plasmid remaining over background singly-integrated population. Based on this and the flow cytometry results, we only use the 1:100,000 for Flow-Seq experiments.

(**M**) Verification of library integration by flow cytometry. In particular, flow cytometry analysis for the SRE library in the DHFR intron backbone with MFASS reporter is shown here in biological replicates. The percentage and number of cells sorted (in millions, M) per bin are shown.

(**N-Q**) We measured exon inclusion by using RT-PCR for measurement of splicing at the RNA level, and using flow cytometry for the MFASS splicing readout, from which we calculate an inclusion index for each exon measured. Nine individual single cells from the SRE library were sorted from the library based on fluorescence behavior, expanded to establish individual cell lines, from which we performed individual validation. In addition, we performed transient transfection for the same sequences that we integrated genomically. (**N** and **O**) Flow cytometry analysis of single exons under (**N**) transient expression, and (**O**) site-specific genome-integrated. (**P** and **Q**) RT-PCR analysis of the same individual control exons under (**P**) transient expression, and (**Q**) genome integration. Inclusion indices from MFASS assay are indicated on the upper left hand corner of flow cytometry plots (**N** and **O**) and under each exon (**P** and **Q**). We found the indices correlate strongly with observed splicing levels for individual exons.

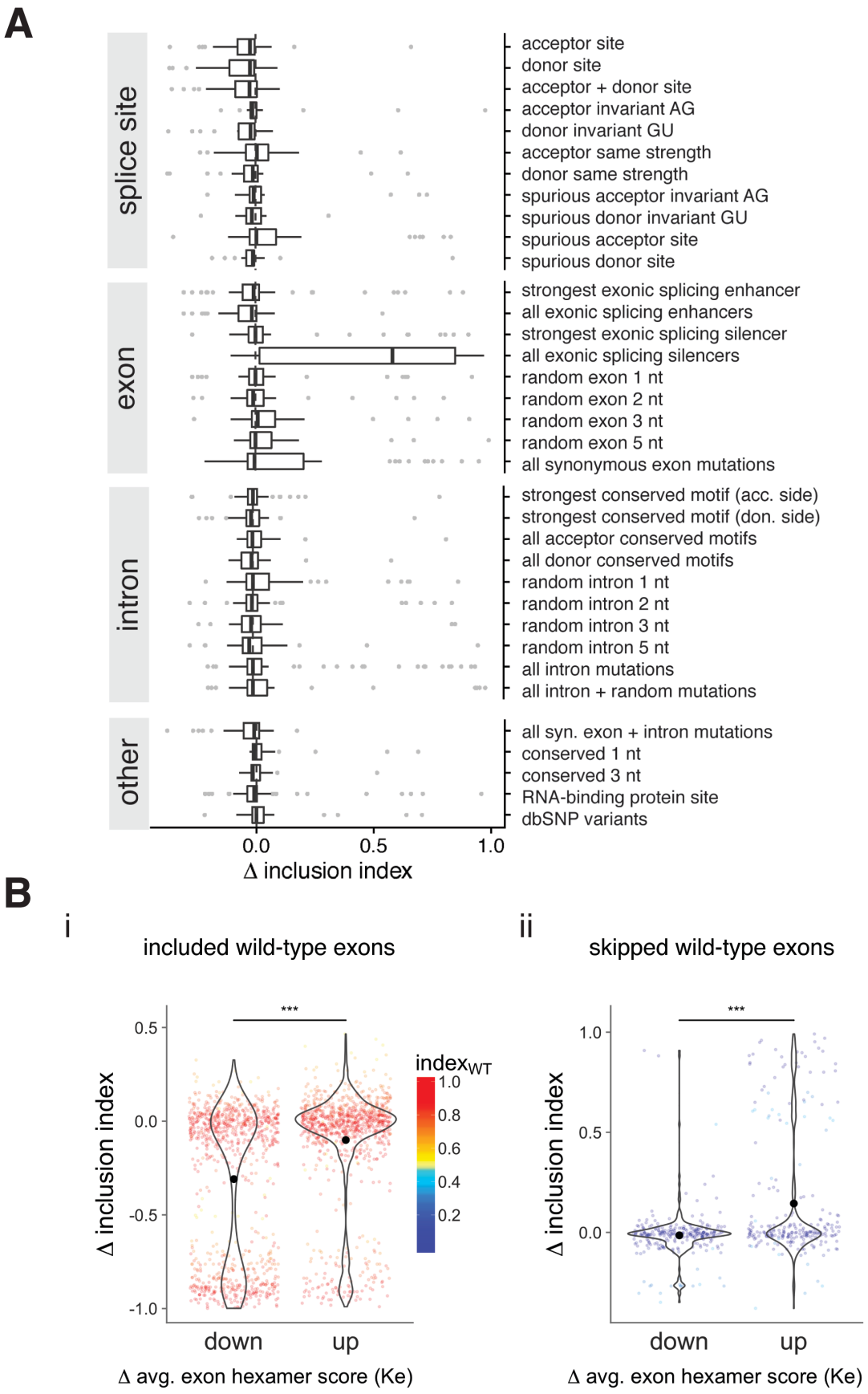

A
splice site
exon
intron
other
acceptor site
donor site
acceptor + donor site
acceptor invariant AG
donor invariant GU
acceptor same strength
donor same strength
spurious acceptor invariant AG
spurious donor invariant GU
spurious acceptor site
spurious donor site
strongest exonic splicing enhancer
all exonic splicing enhancers
strongest exonic splicing silencer
all exonic splicing silencers
random exon 1 nt
random exon 2 nt
random exon 3 nt
random exon 5 nt
all synonymous exon mutations
strongest conserved motif (acc. side)
strongest conserved motif (don. side)
all acceptor conserved motifs
all donor conserved motifs
random intron 1 nt
random intron 2 nt
random intron 3 nt
random intron 5 nt
all intron mutations
all intron + random mutations
all syn. exon + intron mutations
conserved 1 nt
conserved 3 nt
RNA-binding protein site
dbSNP variants
0.0
0.5
1.0
Δ inclusion index
B
i
included wild-type exons
***
0.5
0.0
-0.5
-1.0
Δ inclusion index
index_WT
1.0
0.8
0.6
0.4
0.2
down
up
Δ avg. exon hexamer score (Ke)
ii
skipped wild-type exons
***
1.0
0.5
0.0
Δ inclusion index
down
up
Δ avg. exon hexamer score (Ke)

Figure 2.S2. Exon inclusion rates and alternative hexamer score metrics related to SRE library. Related to Figure 2.2.

(**A**) Exon inclusion rates of designed mutants for skipped natural exons in SRE library. We quantify Δinclusion index for a mutant sequence relative to wild-type (WT) for all skipped exons (inclusion index < 0.50) from the SRE library.
(**B**) Evaluation of an alternative exon hexamer score metric. Average exonic scores defined in Ke et al. for 6,713 designed sequences in (**i**) included wild-type exons and (**ii**) skipped wild-type exons (Ke et al., 2011). Consistent with observations from the HAL model (**Figure 2.2C**), there is a significant difference in average Δinclusion index between sequences that increase (up) or decrease (down) overall exon hexamer score, for both included exons (Mann-Whitney *U* test, $p < 10^{-16}$) and skipped exons (Mann-Whitney *U* test, $p < 10^{-16}$). We quantify Δinclusion index for a mutant sequence relative to wild-type (WT), and colored each point by the inclusion index of the corresponding WT sequence. Included wild-type exons are defined as exons with inclusion index ≥ 0.50, while skipped wild-type exons are defined as exons with inclusion index < 0.50.

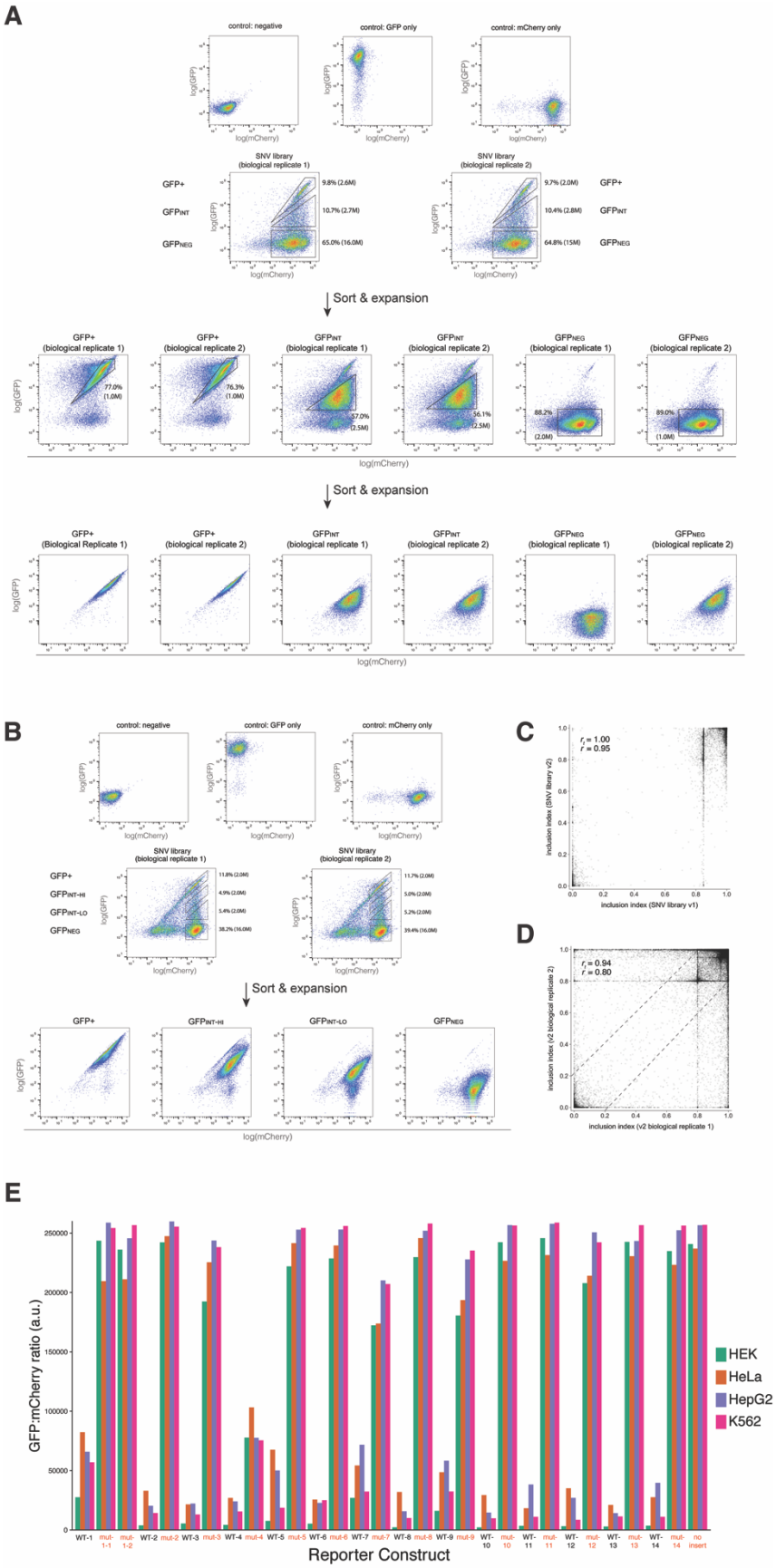

A
control: negative
control: GFP only
control: mCherry only
log(GFP)
log(mCherry)
SNV library (biological replicate 1)
SNV library (biological replicate 2)
GFP+
GFPINT
GFPNEG
9.8% (2.6M)
10.7% (2.7M)
65.0% (16.0M)
9.7% (2.0M)
10.4% (2.8M)
64.8% (15M)
Sort & expansion
GFP+ (biological replicate 1)
GFP+ (biological replicate 2)
GFPINT (biological replicate 1)
GFPINT (biological replicate 2)
GFPNEG (biological replicate 1)
GFPNEG (biological replicate 2)
77.0% (1.0M)
76.3% (1.0M)
57.0% (2.5M)
56.1% (2.5M)
88.2% (2.0M)
89.0% (1.0M)
Sort & expansion
GFP+ (Biological Replicate 1)
GFP+ (biological replicate 2)
GFPINT (biological replicate 1)
GFPINT (biological replicate 2)
GFPNEG (biological replicate 1)
GFPNEG (biological replicate 2)
B
control: negative
control: GFP only
control: mCherry only
SNV library (biological replicate 1)
SNV library (biological replicate 2)
GFP+
GFPINT-HI
GFPINT-LO
GFPNEG
11.8% (2.0M)
4.9% (2.0M)
5.4% (2.0M)
38.2% (16.0M)
11.7% (2.0M)
5.0% (2.0M)
5.2% (2.0M)
39.4% (16.0M)
Sort & expansion
GFP+
GFPINT-HI
GFPINT-LO
GFPNEG
C
r1 = 1.00
r = 0.95
inclusion index (SNV library v2)
inclusion index (SNV library v1)
D
r1 = 0.94
r = 0.80
inclusion index (v2 biological replicate 2)
inclusion index (v2 biological replicate 1)
E
GFP:mCherry ratio (a.u.)
Reporter Construct
HEK
HeLa
HepG2
K562
WT-1
mut-1-1
mut-1-2
WT-2
mut-2
WT-3
mut-3
WT-4
mut-4
WT-5
mut-5
WT-6
mut-6
WT-7
mut-7
WT-8
mut-8
WT-9
mut-9
WT-10
mut-10
WT-11
mut-11
WT-12
mut-12
WT-13
mut-13
WT-14
mut-14
no insert

Figure 2.S3. Related to Figure 2.3.

(**A**) Flow cytometry analysis of ExAC SNV library (version 1) integrated in human cells. For this SNV dataset, the library was sorted and expanded two times to ensure purity of the final population. Three populations, GFP_{+}, GFP_{int}, and GFP_{neg}, were sorted according to GFP fluorescence given a minimal threshold of mCherry expression. The sorted populations were expanded in cell culture and subjected to flow cytometry analysis. Populations remained stable in fluorescence after sort and expansion in culture.Single fluorescence controls (GFP only and mCherry only) and negative control (landing pad cell line only) are shown on the top row. SNV library was evaluated in the reporter constructs from **Figure 2.S1A**. For the first sort, we obtained 16 million cells for GFP_{neg} library, 2.6 million cells for GFP_{+} library and 2.7 million cells for GFP_{int} library (biological replicate 1), 15 million cells for GFP_{neg} library, 2 million cells for GFP_{+} library and 2.8 million cells for GFP_{int} library (biological replicate 2). For the purifying sort, we further sub-sorted the libraries from the first sort, and obtained ~2 million cells for GFP_{neg} library, 1 million cells for GFP_{+} library and 2.5 million cells for GFP_{int} library (biological replicate 1), and 1 million cells for GFP_{neg} library, 1 million cells for GFP_{+} library and 2.5 million cells for GFP_{int} library (biological replicate 2). The percentage and number of cells sorted (in millions, M) per bin are shown.
(**B**) Flow cytometry analysis of ExAC SNV library (version 2) integrated in human cells. For this SNV dataset, four populations, GFP_{+}, $GFP_{int\text{-}hi}$, $GFP_{int\text{-}lo}$, and GFP_{neg}, were FACS-sorted according to GFP fluorescence given a minimal threshold of mCherry expression. In particular, for both biological replicates, we obtained 16 million cells for GFP_{neg} library, 2 million cells for GFP_{+} library, 2 million cells for $GFP_{int\text{-}hi}$ and $GFP_{int\text{-}lo}$ library. These sorted populations were expanded in cell culture and subjected to flow cytometry analysis. Populations remain stable in fluorescence after sort and expansion in culture. Single fluorescence controls (GFP only and mCherry only) and negative control (landing pad cell line only) are shown on the top row. SNV library was evaluated in the reporter construct from **Figure 2.S1C**. The percentage and number of cells sorted (in millions, M) per bin are shown.
(**C**) Exon inclusion metrics across two independent biological replicates for version 2 of the SNV library (n = 31,144, r_t = 0.94, $p < 10^{-16}$, tetrachoric; r = 0.80, $p < 10^{-16}$, Pearson). Dashed line demarcate inclusion indices that agree within 0.20 across biological replicates for the SNV library that we used for subsequent analysis. SNV library were evaluated in the reporter constructs from **Figure 2.S1C**.
(**D**) Exon inclusion metrics across two separate reporter constructs located in different parts of GFP and in different frames. Data shown are sequences appearing both in versions 1 and 2 for the SNV library (n = 5,740, r_t = 1.00, $p < 10^{-16}$, tetrachoric; r = 0.94, $p < 10^{-16}$, Pearson). SNV library was evaluated in the reporter constructs from **Figures 2.S1A and 2.S1C**.
(**E**) Exon inclusion rates of SDVs and corresponding wild-type sequences across four human cell types. We tested 15 of the individual control reporter constructs from longer intronic contexts (**Figure 2.3E**) across 4 human cell types: HEK293T, HeLa S3, HepG2 and K562 cells. A lower GFP-mCherry ratio indicates increased exon inclusion from MFASS. We found that large-effect splicing disruptions are consistent across 4 cell types in all 15 of the splice-disrupting variants assayed (see also **Figure 2.3F**). This is indicated by the increased GFP-mCherry ratio in SDVs (labelled in red as “mut” on the x-axis), as compared to their corresponding wild-type controls (labelled in black as “WT” on the x-axis). Note that the effects of 15 SDVs tested here are highly transferable across the 4 cell types examined (15/15, 100%).

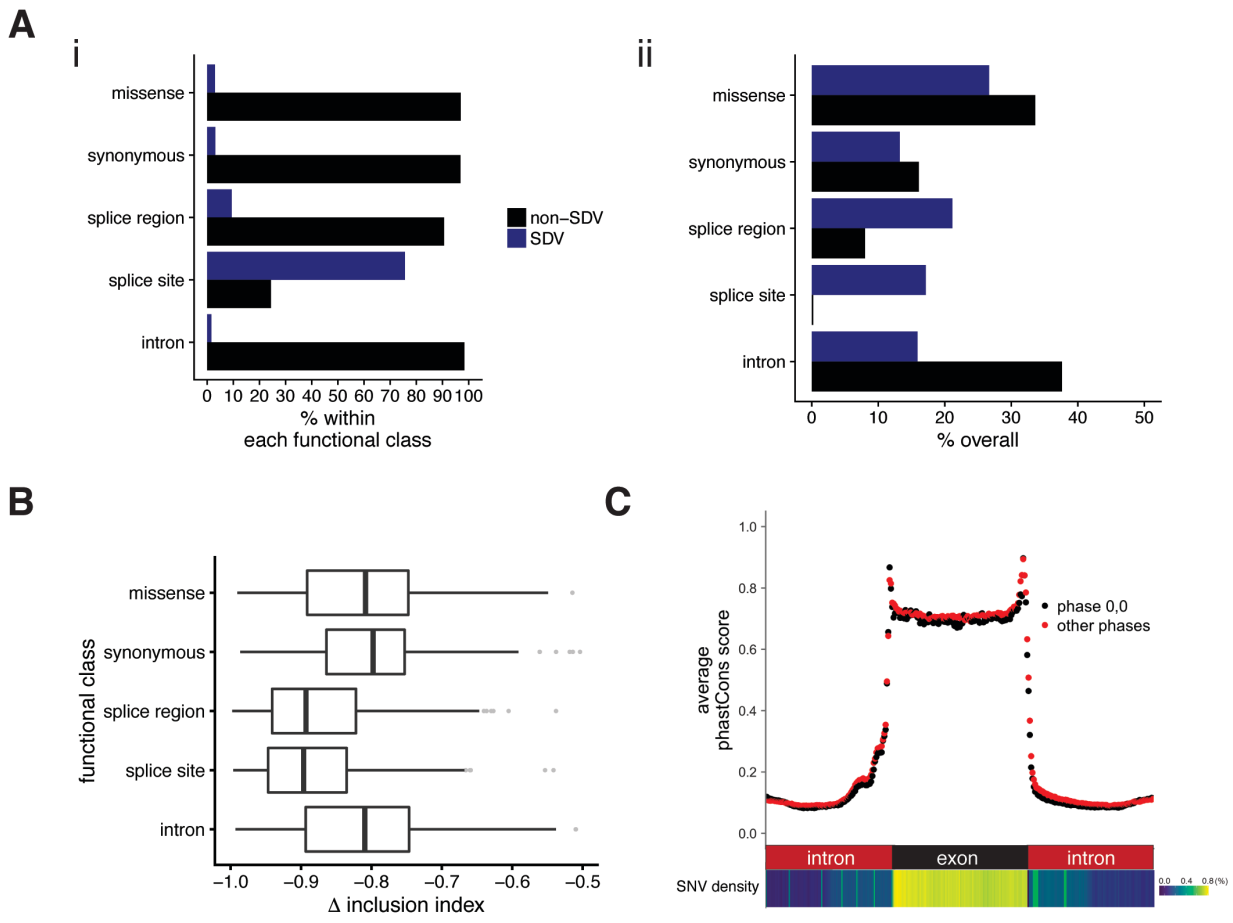
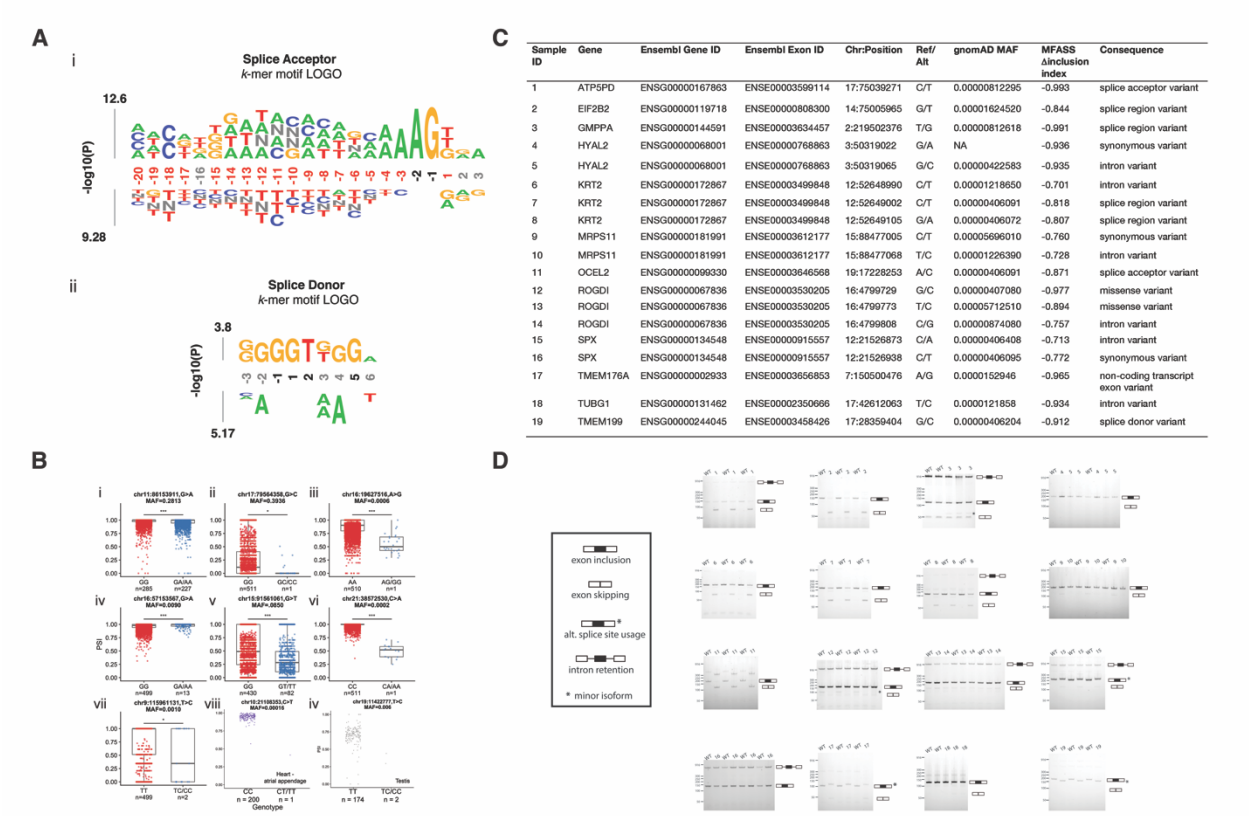


Figure 2.S4. Related to Figure 2.4.

Variants were functionally classified using the Ensembl variant effect predictor (VEP v80) across SDVs (*n* = 1,050) and non-SDVs (*n* = 26,683).
(**A**) (**i**) Proportions of SDVs and non-SDVs within each functional annotation. The proportions of SDV in each category directly correspond to the left panel of Figure 4B, which details SDVs only. (**ii**) Overall percentages of SDVs and non-SDVs by functional annotation.
(**B**) Distributions of change in inclusion index for SDVs by functional annotation. Among these functional classes, splice site variants have the largest mean effect size, followed by splice region variants. Intron variants have roughly the same mean effect size as missense and synonymous variants.
(**C**) Comparison of genome-wide ExAC SNV conservation profiles and SNV density across gene regions. Average phastCons scores for genome-wide ExAC SNVs starting and ending on phase 0 compared to other phases (Siepel et al., 2005), with SNV density illustrated across gene regions (*bottom*). Because exon sizes are variable, graph shows relative positions for all SNVs within 100 bp of the exon-intron boundary for both phastCons and SNV density.



Sample ID	Gene	Ensembl Gene ID	Ensembl Exon ID	Chr:Position	Ref/ Alt	gnomAD MAF	MFASS Δinclusion index	Consequence
1	ATP5PD	ENSG00000167863	ENSE00003599114	17:75039271	C/T	0.00000812295	-0.993	splice acceptor variant
2	EIF2B2	ENSG00000119718	ENSE00000808300	14:75005965	G/T	0.00001624520	-0.844	splice region variant
3	GMPPA	ENSG00000144591	ENSE00003634457	2:219502376	T/G	0.00000812618	-0.991	splice region variant
4	HYAL2	ENSG00000068001	ENSE00000768863	3:50319022	G/A	NA	-0.936	synonymous variant
5	HYAL2	ENSG00000068001	ENSE00000768863	3:50319065	G/C	0.00000422583	-0.935	intron variant
6	KRT2	ENSG00000172867	ENSE00003499848	12:52648990	C/T	0.00001218650	-0.701	intron variant
7	KRT2	ENSG00000172867	ENSE00003499848	12:52649002	C/T	0.00000406091	-0.818	splice region variant
8	KRT2	ENSG00000172867	ENSE00003499848	12:52649105	G/A	0.00000406072	-0.807	splice region variant
9	MRPS11	ENSG00000181991	ENSE00003612177	15:88477005	C/T	0.00005696010	-0.760	synonymous variant
10	MRPS11	ENSG00000181991	ENSE00003612177	15:88477068	T/C	0.00001226390	-0.728	intron variant
11	OCEL2	ENSG00000099330	ENSE00003646568	19:17228253	A/C	0.00000406091	-0.871	splice acceptor variant
12	ROGDI	ENSG00000067836	ENSE00003530205	16:4799729	G/C	0.00000407080	-0.977	missense variant
13	ROGDI	ENSG00000067836	ENSE00003530205	16:4799773	T/C	0.00005712510	-0.894	missense variant
14	ROGDI	ENSG00000067836	ENSE00003530205	16:4799808	C/G	0.00000874080	-0.757	intron variant
15	SPX	ENSG00000134548	ENSE00000915557	12:21526873	C/A	0.00000406408	-0.713	intron variant
16	SPX	ENSG00000134548	ENSE00000915557	12:21526938	C/T	0.00000406095	-0.772	synonymous variant
17	TMEM176A	ENSG00000002933	ENSE00003656853	7:150500476	A/G	0.0000152946	-0.965	non-coding transcript exon variant
18	TUBG1	ENSG00000131462	ENSE00002350666	17:42612063	T/C	0.0000121858	-0.934	intron variant
19	TMEM199	ENSG00000244045	ENSE00003458426	17:28359404	G/C	0.00000406204	-0.912	splice donor variant

Figure 2.S5. Evaluation of the effects of splice-disrupting variants assayed by MFASS. Related to Figure 2.5.

(**A**) Distinct signatures of large-effect rare variants for splicing in splice donor and acceptor regions. We used *k*pLogo (Wu and Bartel, 2017) to generate a motif-based *k*-mer logo that visualizes enriched residues or motifs for SDVs (*top*) and non-SDVs (*bottom*). Stacked residues at a position represent a single most significant motif, starting or ending at a particular position. The total height is scaled relative to the significance of the motif. Motifs are read such that the top position correspond to the stacked position, with the rest of the residues read correspondingly to the right of the stacked position. N indicates positions with no residue preference. For the splice acceptor region (upstream intron-exon boundary), we see that adenine-rich motifs are less tolerated for across positions of around the polypyrimidine tract (-14 to -6), along with an enrichment for thymine-rich motifs for non-SDVs (*left bottom*), consistent with conserved signatures of polypyrimidine tract in our genomes. In contrast, for the splice donor region (downstream exon-intron boundary), we see that guanine-rich motifs are less tolerated (*right top*).
(**B**) Exon percent-spliced-in (PSI) values for splice-disrupting variants (SDVs) with globally significant splicing changes in the Genotype-Tissue Expression (GTEx) project (**Quantification and Data Analysis, STAR Method**). (**i - vii**) First, we assayed 523 SNVs with MAF >0.5% and found 2% (11/523) are SDVs in our assay. 96% of our SNVs with minor allele frequency (MAF) > 0.5% (505/523) overlapped with GTEx across 47 human tissues. Of these 505 with MAF > 0.5% that overlapped with GTEx, 9 were SDVs. Of 1,050 SDVs detected by MFASS, 28 SDVs overlapped with GTEx, from which 9 has MAF > 0.5%, 8 with 0.05% < MAF < 0.5%, 10 with 0.0025% < MAF < 0.5%, 1 with 0.001 < MAF < 0.025% and 1 with MAF < 0.001. 7 of these showed globally significant differences in the distribution of tissue PSI values for alternative alleles as compared to that of reference alleles (Mann-Whitney *U* test). Adjusted *p*-values were calculated using the Benjamini-Hochberg procedure at an FDR of 5% [$p < 0.05$ (*), $p < 0.01$ (**),

$p < 0.001$ (***)]. *n* indicates the number of individuals with the genotype indicated, and each point in the boxplot represents PSI measurement from a single tissue. (**viii, iv**) Exon PSI values for two variants with tissue-specific behavior. While these two variants do not reach global significance, they show reduced exon inclusion in relevant tissue types. (**viii**) A variant in NEBL, a gene that is abundantly expressed in cardiac muscle tissue and potentially involved in cardiac myofibril assembly. (**iv**) A variant in TSPAN16, a member of the tetraspanin protein family, whose members mediate signal transduction events in cell development, activation, growth and motility. This gene is highly expressed only in testis.
(**C and D**) Validation of individual SDVs detected by MFASS within the full gene context.
(**C**) SDVs individually validated for exon recognition in their full gene context. We validated 19 individual SDVs detected by MFASS within their broader gene context by assembling 19 SDV-containing full genes and 12 corresponding wild-types (up to ~13kb) using isothermal assembly, and examined splicing disruptions caused by SDVs using RT-PCR (**STAR Method**). Missing allele frequency indicates insufficient gnomAD coverage for a particular variant.
(**D**) RT-PCR data for 19 SDVs across 12 full genes. The sample number on top of lanes in gel images corresponds to SDV annotations detailed in (**i**). Lanes with the same sample ID represent biological replicates. For the PCR, we used primers that flank the upstream and downstream exons from the exon or intron of interest that contains the SDVs (Table S4, under 'RT-PCR primers' section), thus allowing us to detect exon recognition defects across the exon of interest. We observed 13 of 19 variants (68.4%) cause splicing disruptions in 9 of 12 genes (75%), with 9 of 19 variants (42.1%) having appreciable effects on exon recognition. These disruptions include exon skipping and alternative 5' and 3' splice site usage (see inset, B) in the broader full gene context. Minigene illustrations for each exon (next to each respective gel image) indicate the sizes of genome-annotated major isoforms that correspond to exon inclusion, exon skipping or intron retention events. Asterisk(*) alone indicates minor isoforms. While we do not notice significant changes in intron retention across all 19 SDVs examined, the usage of alternative 5' splice sites (sample IDs 15 & 17) or 3' splice site (sample ID 19) were confirmed (**STAR Method**).

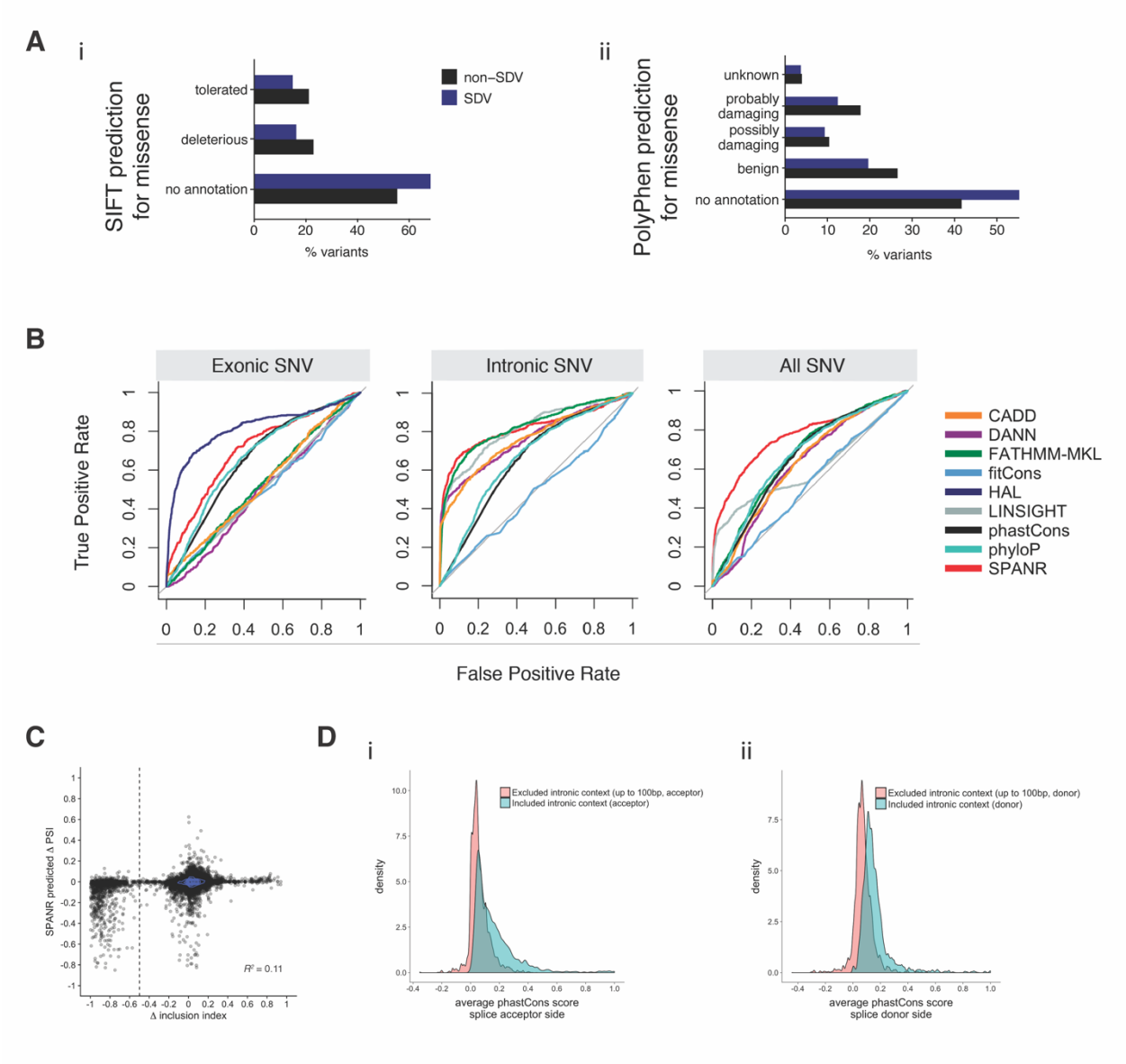


Figure 2.S6. Evaluation of algorithms and metrics for large-effect disruptions to splicing. Related to Figure 2.6.

(**A**) SIFT and PolyPhen predictions of SDVs and non-SDVs for missense variants from ExAC. Distributions of missense SDVs (*n* = 250) and missense non-SDVs (*n* = 8,966) across (**i**) SIFT and (**ii**) PolyPhen predictions. Proportions are shown instead of counts here to allow facile comparison between SDVs and non-SDVs. The proportions of SDV in both (**i**) and (**ii**) directly correspond to **Figure 2.6A**. These functional predictions show that few SDVs are predicted to have functional effect (i.e., benign), and SDVs that do not have any annotations are at a higher proportion than that of non-SDVs.
(**B**) Prediction algorithms which predict splicing and non-coding genetic variation for ExAC SNV library. Receiver operating characteristic (ROC) curves that can predict splicing or non-coding genetic variants. The area under the ROC curve for all SNVs of CADD, DANN, FATHMM-MKL, fitCons, LINSIGHT and SPANR were 0.629 (95%CI: 0.615-0.647), 0.620 (95%CI: 0.604-0.635), 0.670 (95%CI: 0.654-0.685), 0.517 (95%CI: 0.500-0.535), 0.604 (95%CI: 0.583-0.624) and 0.774 (95%CI: 0.752-0.787) respectively. CI, confidence interval. For (**B**), colors for each

algorithm match those in **Figure 2.6**, with the addition of the HAL predictor for (**B**) that evaluates exonic changes only.
(**C**) Predicted percent-spliced-in (PSI) from SPANR plotted against the Δinclusion index (inclusion $index_{mutant}$ - inclusion $index_{WT}$) from MFASS (R^2 = 0.11). The dashed lines indicate threshold (Δinclusion index = -0.50) below which we call splice-disrupting variants (SDVs). Contour lines show a density estimation of the points.
(**D**) Comparison of length-dependent conservation profiles for human introns in ExAC SNV library. Distributions of average phastCons conservation scores for shorter native intron contexts across MFASS library sequences, as compared to intron contexts that are not included, up to 100bp, on either the acceptor or donor side. For the ExAC SNV library, (**i**) for the splice acceptor side, native intron lengths range from 40 to 81 bp, (**ii**) for the splice donor side, native intron lengths range from 30 to 71 bp. Average conservation across sequences is higher for short intron contexts in our library compared to that of the excluded intronic contexts, indicating we are likely capturing a large fraction of the relevant conserved intronic elements.

Table 2.S1. Description of Motif Types used in Splicing Regulatory Element (SRE) Library Design. Related to Figure 2.2.

Motif type	Description
Splice donor (Portales-Casamar et al., 2010; Yeo and Burge, 2004)	5’ splice site (downstream intron), 9 bp, -3 to +6
Splice acceptor (Yeo and Burge, 2004)	3’ splice site (upstream intron), 23 bp, -20 to +3
Exonic splicing enhancer (ESE) (Ke et al., 2011)	Hexamer that is more likely to activate nearby splice sites, leading to enhanced exon inclusion
Exonic splicing silencer (ESS) (Ke et al., 2011)	Hexamer that is more likely to silence nearby splice sites, leading to enhanced exon skipping
Donor intronic conserved sequence (Voelker and Berglund, 2007)	Conserved sequence (CS) defined as a contiguous run of at least 7 nt of identity in a multiple sequence alignment between humans and six eutherian mammals occurring in the donor (downstream intron) or acceptor (upstream intron) region
Acceptor intronic conserved sequence (Voelker and Berglund, 2007)	
RNA-binding protein (RBP) motifs (Cook et al., 2011)	Position frequency matrix representing binding preferences of RBPs which allow for additive probabilistic descriptions

Table 2.S2. Description of Functional Classes in Splicing Regulatory Element (SRE) Library. Related to Figure 2.2.

Category	Description	Criteria to keep mutant	Mutant score
dbSNP variants	Substitute reference allele with alternative allele	Keep all	NA
acceptor site, donor site	Mutate splice acceptor (-20 to +3) / splice donor (-3 to +6) to sequence with lower score, but do not mutate 2bp invariant position	0 < new score < 3, |(original score – new score)| > 0.50	Absolute distance of new score from the midpoint of (0, 3)
acceptor + donor site	Weaken both the splice acceptor and donor as described above	Refer to weaken splice donor/acceptor	
acceptor invariant AG, acceptor invariant GU	Mutate splice donor/acceptor at invariant positions only (last 2 bp of upstream intron, first 2 bp of downstream intron)	Keep all	New score
acceptor same strength, donor same strength	Mutate splice donor/acceptor to sequence with comparable MaxEnt score, except invariant 2 bp	|(original score – new score)| < 0.50	|(original score – new score)|
spurious acceptor site, spurious donor site	Mutate spurious splice donor/acceptor to sequence with lower score, except invariant 2 bp	0 < new score < 3, |(original score – new score)| > 0.50	Absolute distance of new score from the midpoint of (0, 3)
spurious acceptor invariant AG, spurious donor invariant GU	Mutate all sequences called by MaxEntScan as splice sites (but occurring at non-canonical positions)	0 > new score	New score
all exonic splicing enhancers/ exonic splicing silencers/ conserved motifs acceptor side/ conserved motifs donor side	Mutate all occurrences of motif to lower score	|original score| * 0.1 >.new score|	Sum of all mutated motif scores
strongest exonic splicing enhancer/ exonic splicing silencer/ conserved motif acceptor side/ conserved motif donor side	Mutate strongest motif to lower score than original	|original score| * 0.1 >.new score|	New score

RNA-binding protein site	Destroy RBP motifs	New score == 0	PFM score
conserved 1nt, conserved 3nt	Mutate (1,3) random bp of conserved features based on total conservation	Keep all	NA
random exon 1nt, random exon 2nt, random exon 3nt, random exon 5 nt	Randomly mutate 1, 2, 3, or 5 nt of exonic sequence	NA	NA
random intron 1nt, random intron 2nt, random intron 3nt, random intron 5nt	Randomly mutate 1, 2, 3, or 5 nt of intronic sequence	NA	NA
all synonymous exon mutations	Synthesize every synonymous mutation at each position in the exon	NA	NA
all intronic mutations	Synthesize every intronic mutation, except at invariant 2 bp of splice sites	NA	NA
all synonymous exon + all intronic mutations	Combine both all synonymous exon mutations and all intronic mutations, described above	NA	NA
all intronic mutations + all random 1/2/3/5nt mutations	Combines both of the above intronic randomizations (random and aggressive)	Keep all	NA

Table 2.S3: Gene Ontology (GO) Enrichment for ExAC Splice-Disrupting Variants (SDVs, *n* = 1,050). Related to Figure 2.5.

GO ID	Term	Annotated	Significant	Expected	*p*-value (Fisher's test)	Mean number of exons per gene
GO:0030574	collagen catabolic process	26	19	11.16	0.0018	48.8
GO:0044243	multicellular organismal catabolic process	27	19	11.58	0.0036	47.5
GO:0006892	post-Golgi vesicle-mediated transport	14	11	6.01	0.0074	34.4
GO:0032963	collagen metabolic process	32	21	13.73	0.0080	42.7

Table 2.S4. Primers used in this study. Related to Figures 2.3, 2.S1, and 2.S5.

SRE Library	
ODY093	GCAGTGTTTCTCTAACTTTCGGCG
ODY028	GGCATGTACTTGTAATCCTATCAGTGG
ODY088	ACCCGTCCTATATATAGCTATCTATGTCTGGCGCGC
ODY089	GCGACCGTGTACAAAAGTAAATAGCCCGGCTGG
ODY031	AATGATACGGCGACCACCGAGATCTACACGCAGTGTTTCTCTAACTTTCGGCGCGCC
ODY19	GCAGTGTTTCTCTAACTTTCGGCGCGCC
ODY042	AGTTCCAGCCGGGCTATGGCCAGTGAGATCCAAG
ODY106	GCAAGAGTTCCAGCCGGGCTATTTACTTTTGTACACGGTCGC
ODY047	CTTGGATCTCACTGGCCATAGCCCGGCTGGAACTCTTGCTTAATTAA
ODY95	AATGATACGGCGACCACCGAGATCTACACACCCGTCCTATATATAGCTATCTATGTCTGG
ODY96	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGCGACCGTGTACAAAAGTAAATAGCC
ODY097	CAAGCAGAAGACGGCATACGAGATCTAGTACGGCGACCGTGTACAAAAGTAAATAGCC
ODY098	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGCGACCGTGTACAAAAGTAAATAGCC
ODY099	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGCGACCGTGTACAAAAGTAAATAGCC
ODY100	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGCGACCGTGTACAAAAGTAAATAGCC
ODY101	CAAGCAGAAGACGGCATACGAGATCATGCCTAGCGACCGTGTACAAAAGTAAATAGCC
ODY102	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGCGACCGTGTACAAAAGTAAATAGCC
ODY103	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGCGACCGTGTACAAAAGTAAATAGCC
ODY104	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGCGACCGTGTACAAAAGTAAATAGCC
ODY105	CAAGCAGAAGACGGCATACGAGATTCCTCTACGCGACCGTGTACAAAAGTAAATAGCC
ORC403	/5Biosq/CCCTTTAATCAGATGCGTCGTATTATTGGCG
ORC404	/5Biosq/TGGTAGTAATAAGGGCGACCGGGCGGGTTAA
ORC405	/5Biosq/ATATAGATGCCGTCCTAGCGTTTATTTGGCG
ORC406	/5Biosq/AAGTATCTTTCCTGTGCCCACGCGCGCTTAA
SNV Library	
ORC515	GCAGTGTTTCTCTAACTTTCACCG
ODY028	GGCATGTACTTGTAATCCTATCAGTGG
ORC522	AATGATACGGCGACCACCGAGATCTACACAGCACACTGTTAACGCAGTGTTTCTCTAACTTTCACCGGT
ODY032	CAAGCAGAAGACGGCATACGAGATTCGCCTTACTTGGATCTCACTGGCCATAGCC
ODY033	CAAGCAGAAGACGGCATACGAGATCTAGTACGCTTGGATCTCACTGGCCATAGCC
ODY034	CAAGCAGAAGACGGCATACGAGATTTCTGCCTCTTGGATCTCACTGGCCATAGCC

ODY035	CAAGCAGAAGACGGCATACGAGATGCTCAGGACTTGGATCTCACTGGCCATAGCC
ODY036	CAAGCAGAAGACGGCATACGAGATAGGAGTCCCTTGGATCTCACTGGCCATAGCC
ODY037	CAAGCAGAAGACGGCATACGAGATCATGCCTACTTGGATCTCACTGGCCATAGCC
ODY038	CAAGCAGAAGACGGCATACGAGATGTAGAGAGCTTGGATCTCACTGGCCATAGCC
ODY039	CAAGCAGAAGACGGCATACGAGATCAGCCTCGCTTGGATCTCACTGGCCATAGCC
ODY040	CAAGCAGAAGACGGCATACGAGATTGCCTCTTCTTGGATCTCACTGGCCATAGCC
ODY041	CAAGCAGAAGACGGCATACGAGATTCCTCTACCTTGGATCTCACTGGCCATAGCC
ORC517	AGCACACTGTTAACGCAGTGTTTCTCTAACTTTCACCGGT
ODY042	AGTTCCAGCCGGGCTATGGCCAGTGAGATCCAAG
ORC518	CTTGGATCTCACTGGCCATAGCCCGGCTGGAACTCTTGCGCTAGC
RT-PCR (SRE Library)	
ORC505	AATGATACGGCGACCACCGAGATCTACACTATGTAAATGTTGTCACCAGTGTGGGC
ORC527	ATGGTGTCTAAGGGCGAAGAGC
ORC506	AATGATACGGCGACCACCGA
Isothermal assembly of full genes	
A) Gibson assembly primers	
ATP5PD_F1	CGCTCCCAGCGCGGCAGACTTCAATAGTTTGG
ATP5PD_F_ext	CCCCAGAACCAAAAGGCCATTGCTAGTTCCCTGAAATCCTGG
ATP5PD_F2	GAATTTCTCTTTCTTCTTAAAAGTGAAATCTTGTGCTGAGTGGGTGTCTCTCTCAAAGGC
ATP5PD_F2_WT	GAATTTCTCTTTCTTCTTAAAGGTGAAATCTTGTGCTGAGTGGGTGTCTCTCTCAAAGGC
ATP5PD_F3	TCCCGTGCCAGAGGATAAATATACTGC
ATP5PD_R1	ACAGATGGCTGGCAACTAGAAGGCAC
EIF2B2_F1	GGAAGTGCAAACTGTGTGGTCTGGCAGGTGTGG
EIF2B2_F2	GCCATTTTTGCCGTTATGTCAAGAGTCAACAAGGTGGTTATATCTGGAG
EIF2B2_F2_WT	GCCATTTTTGCCGTTATGTCAAGAGTCAACAAGGTGGGTATATCTGGAG
EIF2B2_R1	TGGGCACCCCTGATACCAAGGCTGACAGGTAG
EIF2B2_R2	CTCCAGATATAACCACCTTGTTGACTCTTGACATAACGGCAAAAATGGC
EIF2B2_R2_WT	CTCCAGATATACCCACCTTGTTGACTCTTGACATAACGGCAAAAATGGC
GMPPA_F1	GAGGCCAGGGTTTATTGGACAGAGTCAGTTGTGGGG
GMPPA_F2	CTGTCTCGGGTGTGTCTGTCTGTCAGGCTAACAGGACG

GMPPA_F2_WT	CTGTCTCGGGTGTGTCTGTCTTTCAGGCTAACAGGACG
GMPPA_R1	CAAGGTTCACGGGGTTTATTAGGGAGTCGGGAGGG
GMPPA_R2	CGTCCTGTTAGCCTGACAGACAGACACACCCGAGACAG
GMPPA_R2_WT	CGTCCTGTTAGCCTGAAAGACAGACACACCCGAGACAG
HYAL2_F1	AAACAGGGTCAAGGCGATCTCCTCCCCCACG
HYAL2_F2	GTCTCAGTCTTCCCCAGTGAGCATCCCTTTTCCTGC
HYAL2_F3	GGACCTCATCTCTACCATTGGTGAGAGTGCGGC
HYAL2_F2_WT	GTCTCAGTCTTCCCCAGTGACCATCCCTTTTCCTGC
HYAL_F3_WT	GGACCTCATCTCTACCATTGGCGAGAGTGCGGC
HYAL2_R1	ACTATCTAGGGCAAGGGAGTAGGGTCAGGTCCTCCC
HYAL2_R2	GCAGGAAAAGGGATGCTCACTGGGGAAGACTGAGAC
HYAL2_R3	GCCGCACTCTCACCAATGGTAGAGATGAGGTCC
HYAL2_R2_WT	GCAGGAAAAGGGATGGTCACTGGGGAAGACTGAGAC
HYAL2_R3_WT	GCCGCACTCTCGCCAATGGTAGAGATGAGGTCC
KRT2_F1	GTGTCTGGTGGAAGCCGGAGATCAACTTCCAGC
KRT2_F2	GGCCTCAACCTTTCTTGTAGGACGTGGACAATGC
KRT2_F3	CTCTATGATGCGGTAAAGAGGCTGCTCCGGGACAGTCC
KRT2_F4	CTCTATGATGCGGTAAGGAGGCTGCTCCAGGACAGTCC
KRT2_F2_WT	GGCCTCAACCTTTCTTGCAGGACGTGGACAATGC
KRT2_F3_WT	CTCTATGATGCGGTAAGGAGGCTGCTCCGGGACAGTCC
KRT2_R1	CCCAGCACTGCCAGGCTTAGAGATGAAATCCCTGG
KRT2_R2	GCATTGTCCACGTCCTACAAGAAAGGTTGAGGCC
KRT2_R3	GGACTGTCCCGGAGCAGCCTCTTTACCGCATCATAGAG
KRT2_R4	GGACTGTCCTGGAGCAGCCTCCTTACCGCATCATAGAG
KRT2_R2_WT	GCATTGTCCACGTCCTGCAAGAAAGGTTGAGGCC
KRT2_R3_WT	GGACTGTCCCGGAGCAGCCTCCTTACCGCATCATAGAG
MRPS11_F1	ACTGCTAGAACGAACCATTCGCATATGGAGGGGGTG
MRPS11_F_ext	AACACGGGTATACCTGCAGGGACAAGGAAGATGGGG
MRPS11_F2	CTCCAGAGAGCTAAACAAAAGGGTGTGATCCACATCCG
MRPS11_F5	GCCAGGACGCTTGGTAAGTTACAGTGACTTCCATAGTGTACTTGCC
MRPS11_F2_WT	CTCCAGAGAGCTAAACAAAAGGGCGTGATCCACATCCG
MRPS11_F3_WT	GCCAGGACGCTTGGTAAGTTACAGTGATTTCCATAGTGTACTTGCC
MRPS11_R1	CAGCACCAGCTTTATTGGCCACTCAGAGCCTGG
MRPS11_R_ext	CCCCATCTTCCTTGTCCCTGCAGGTATACCCGTGTT
MRPS11_R2	CGGATGTGGATCACACCCTTTTGTTTAGCTCTCTGGAG
MRPS11_R5	GGCAAGTACACTATGGAAGTCACTGTAACTTACCAAGCGTCCTGGC
MRPS11_R2_WT	CGGATGTGGATCACGCCCTTTTGTTTAGCTCTCTGGAG
MRPS11_R3_WT	GGCAAGTACACTATGGAAATCACTGTAACTTACCAAGCGTCCTGGC
OCEL1_F1	CCGCCAGTCGGGTCCATCCTGCAGTAAATGC

OCEL1_F2	CCTTTCTACTCCTCCCCTTGCCGAAGGAGGCCC
OCEL1_F2_WT	CCTTTCTACTCCTCCCCTTGCAGAAGGAGGCCC
OCEL1_R1	CCCTACTTCCAGGGAACAGGTTGAGATCTGGAGTCCC
OCEL1_R2	GGGCCTCCTTCGGCAAGGGGAGGAGTAGAAAGG
OCEL1_R2_WT	GGGCCTCCTTCTGCAAGGGGAGGAGTAGAAAGG
ROGDI_F1	ACGACGAGGTGCACGCTGTGTTGAAGCAGC
ROGDI_F2	CATGGAAGCTGGCGTCACCCCTCTGGATGACC
ROGDI_F3	GACCCCTGCTTCCAGATCCAGGGTGCCAGAAACC
ROGDI_F4	CCTGCTTACCAGCCGGGACCAAAGCTACCAG
ROGDI_F2_WT	CATGGAAGCTGGCGTGACCCCTCTGGATGACC
ROGDI_F3_WT	GACCCCTGCTTCCAGATCCAGGATGCCAGAAACC
ROGDI_R1	TGGCTCTGTCTGTGGCGTTCCTCACCATCC
ROGDI_R2	GGTCATCCAGAGGGGTGACGCCAGCTTCCATG
ROGDI_R3	GGTTTCTGGCACCCTGGATCTGGAAGCAGGGGTC
ROGDI_R4	CTGGTAGCTTTGGTCCCGGCTGGTAAGCAGG
ROGDI_R2_WT	GGTCATCCAGAGGGGTCACGCCAGCTTCCATG
ROGDI_R3_WT	GGTTTCTGGCATCCTGGATCTGGAAGCAGGGGTC
SPX_F1	GACTGACAAGATGTCCCTGTGGACTCCCAAACTCTACTCC
SPX_F2	CTGGTTGATAGCTTCATATAGGGACTCAGAAGTCTGGCAGC
SPX_F4	CCTTGGCTCTTTTCCTGGTGTTTGTTTTTCTGGGAAACTCC
SPX_F2_WT	CTGGTTGATCGCTTCATATAGGGACTCAGAAGTCTGGCAGC
SPX_F3_WT	CTGGCAGCAACAACCTTGGCTCTTTTCCTGGTGTTTG
SPX_R1	CCAGGCCTTCCTCAGTACCACCTTCTCCTTCAGGG
SPX_R2	GCTGCCAGACTTCTGAGTCCCTATATGAAGCTATCAACCAG
SPX_R4	GGAGTTTCCCAGAAAAACAAACACCAGGAAAAGAGCCAAGG
SPX_R2_WT	GCTGCCAGACTTCTGAGTCCCTATATGAAGCGATCAACCAG
SPX_R3_WT	CAAACACCAGGAAAAGAGCCAAGGTTGTTGCTGCCAG
TMEM176A_F1	CCTGTCCCAGAGCCTGCGGACTGTGGAG
TMEM176A_F2	CCTGCTCCTGGCTCAGTCTCTCCCCAGGC
TMEM176A_F2_WT	CCTGCTCCTGGCTCAATCTCTCCCCAGGC
TMEM176A_R1	ATCACATGACTACTCAGGAGGGGACATGAAGCGGAGC
TMEM176A_R2	GCCTGGGGAGAGACTGAGCCAGGAGCAGG
TMEM176A_R2_WT	GCCTGGGGAGAGATTGAGCCAGGAGCAGG
TUBG1_F1	GAAAGGCGAGACATCCCTGACCCAGTGTCCAC
TUBG1_F2	CCCACTCTGACCCTCCCCTACGTCTGTACAGGGAG
TUBG1_F2_WT	CCCACTCTGACCCTCCCCTATGTCTGTACAGGGAG
TUBG1_R1	GAGATGCGTGAGGTCCCTGATCTGTGCTCTGAGG

TUBG1_R2	CTCCCTGTACAGACGTAGGGGAGGGTCAGAGTGGG
TUBG1_R2_WT	CTCCCTGTACAGACATAGGGGAGGGTCAGAGTGGG
TMEM199_F1	GCAACTTCCGGTGCGCTTAGCGTTACTTCCG
TMEM199_F2	GCAACGTCACTTGTCAGCTAAGGACATGCTCTTCAGTACG
TMEM199_F2_WT	GCAACGTCACTTGTCAGGTAAGGACATGCTCTTCAGTACG
TMEM199_R1	CCTGACAGATCTGGGATGGAGGCAGGTATAGCAGC
TMEM199_R2	CGTACTGAAGAGCATGTCCTTAGCTGACAAGTGACGTTGC
TMEM199_R2_WT	CGTACTGAAGAGCATGTCCTTACCTGACAAGTGACGTTGC
B) Gene-specific RT primers(GSP)	
ATP5PD_GSP_R1	GAAGCTCTGGCCCTTGTATTACACATTCTGGAC
EIF2B2_GSP_R1	TGGAGAAAGTTTGAACATAGGTGCACAGACGATGAGT
GMPPA_GSP_R1	CAAGGTTCACGGGGTTTATTAGGGAGTCGGGAGGG
HYAL2_GSP_R1	AATATTGGGTGGCCCAGGACACATTGACC
KRT2_GSP_R1	CTTGACCTCGGCGATGATGCTATCCAAGTC
MRPS11_GSP_R1	AAGTGCAGGCCTCCTTCCCATCAC
OCEL1_GSP_R1	CCAGAGGAGGAAGTCCTGGAAGGCCTT
ROGDI_GSP_R1	GTGGTGAGCCGGTTTCGGGCTCT
SPX_GSP_R1	GTGCCCCTTTCAGGTAGAGCATAGCTTGA
TMEM176A_GSP_R1	GCATGTCCATGAAGGAGGTACATAGGT
TUBG1_GSP_R1	CTGTCATTCAGCCGTTCTAAGAGGTAGGAACCCA
TMEM199_GSP_R1	GCTTTCCCAGGTCGCTGAGAGTCCCA

C) RT-PCR primers	
ATP5PD_RT_F1 (sample ID 1)	GCCCAGGTGGATGCCGAAG
ATP5PD_RT_R1 (sample ID 1)	TCAAGTCCTCAATGGTCATCTGATCAAATGG
EIF2B2_RT_F1 (sample ID 2)	GCAGAGTGTGCTCCTTTCTGCC
EIF2B2_RT_R1 (sample ID 2)	CAGGATGGTCTTCGTGCCAATGATC
GMPPA_RT_F1 (sample ID 3)	GTCACCCTTTCTTACTCCTTGGCAC
GMPPA_RT_R1 (sample ID 3)	GTGCTGGGTTTCTCCACATAGTGC
HYAL2_RT_F1 (sample IDs 4 & 5)	CCCAGTCTACGTCTTCACACGACC
HYAL2_RT_R1 (sample IDs 4 & 5)	TCAGGTAATCTTTGAGGTACTGGCAGG
KRT2_RT_F1 (sample ID 8)	GCTGCTGAGAATGATTTTGTGACGC

KRT2_RT_R1 (sample ID 8)	TGTCAGTGACACTCTGATGTATCTGGG
KRT2_RT_F3 (sample IDs 6 & 7)	GCGCACAGCTGCTGAGAATGAT
KRT2_RT_R3 (sample IDs 6 & 7)	CGTTGGTGTCAGTGACACTCTGATG
MRPS11_RT_F3 (sample IDs 9 & 10)	AGAAGGGCACAGGCATCGC
MRPS11_RT_R3 (sample IDs 9 & 10)	GTGTTGTCTGTGATTGAGATCACTTCCAG
OCEL_RT_F1 (sample ID 11)	ACAGGCAAAGCTCAGGCAGC
OCEL_RT_R1 (sample ID 11)	AGTTTACCCTTCAGGTAGTGGCAGC
ROGDI_RT_F1 (sample IDs 12-14)	TTCGCCTTCCGGGAGGACAAG
ROGDI_RT_R1 (sample IDs 12-14)	CTCTGGTCAGCTGCAGCATCAC
SPX_RT_F1 (sample ID 15)	ATTTCAGGGTTCTGAAAAGACGCAGAAC
SPX_RT_R1 (sample ID 15)	GGAGTCCAGTTCCTTCTCTCCAACAG
SPX_RT_F2 (sample ID 16)	TTTCAGAGCAAGAGTCGAAAACTCACAG
SPX_RT_R2 (sample ID 16)	GGTAGAGCATAGCTTGAGGAGTCCAG
TMEM176A_RT_F1 (sample ID 17)	TGCCATCTGGACAGGGGCTG
TMEM176A_RT_R1 (sample ID 17)	AGCGTTAGCAGAGTCCTCAGC
TUBG1_RT_F1 (sample ID 18)	CAACAACTGGGCCAGCGGATTC
TUBG1_RT_R1 (sample ID 18)	CCCAGCAATGGAGTGACACAGC
TMEM199_RT_F1 (sample ID 19)	GCTCCTAGAAGGCAGTGAAATCTATCTCC
TMEM199_RT_R1 (sample ID 19)	TCGCTGAGAGTCCCACCATGTC

REFERENCES

Alexa A., and Rahnenfuhrer, J. (2016). topGO: Enrichment Analysis for Gene Ontology. R package version 2.32.0.

Adamson, S.I., Zhan, L., and Graveley, B.R. (2018). Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. Genome Biol. *19*, 71.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., et al. (2016). The Ensembl gene annotation system. Database *2016*.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881–888.

Arias, M.A., Lubkin, A., and Chasin, L.A. (2015). Splicing of designer exons informs a biophysical model for exon definition. RNA *21*, 213–229.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

Baralle, D., and Buratti, E. (2017). RNA splicing in human disease and in the clinic. Clin. Sci. *131*, 355–368.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593.

Berget, S.M. (1995). Exon Recognition in Vertebrate Splicing. J. Biol. Chem. *270*, 2411–2414.

Black, D.L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. Annu. Rev. Biochem. *72*, 291–336.

Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. Genome Biol. *18*, 77.

Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192–197.

Chan, T.-F., Poon, A., Basu, A., Addleman, N.R., Chen, J., Phong, A., Byers, P.H., Klein, T.E., and Kwok, P.-Y. (2008). Natural variation in four human collagen genes across an ethnically diverse population. Genomics *91*, 307–314.

Cho, S., Moon, H., Loh, T.J., Jang, H.N., Liu, Y., Zhou, J., Ohn, T., Zheng, X., and Shen, H. (2015). Splicing inhibition of U2AF65 leads to alternative exon skipping. Proceedings of the National Academy of Sciences *112*, 9926–9931.

Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell,

T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am. J. Hum. Genet. *97*, 199–215.

Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T.R. (2011). RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. *39*, D301–D308.

Cooper, T.A. (2005). Use of minigene systems to dissect alternative splicing elements. Methods *37*, 331–340.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*.

De Conti, L., Baralle, M., and Buratti, E. (2012). Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA *4*, 49–60.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A.Y., Dixon, J., Maliskova, L., Guan, K.-L., Shen, Y., and Ren, B. (2016). A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Res. *26*, 397–405.

Duportet, X., Wroblewska, L., Guye, P., Li, Y., Eyquem, J., Rieders, J., Rimchala, T., Batt, G., and Weiss, R. (2014). A platform for rapid prototyping of synthetic gene networks in mammalian cells. Nucleic Acids Res. *42*, 13440–13451.

Faigenbloom, L., Rubinstein, N.D., Kloog, Y., Mayrose, I., Pupko, T., and Stein, R. (2015). Regulation of alternative splicing at the single-cell level. Mol. Syst. Biol. *11*, 845.

Frankel, N., Davis, G.K., Vargas, D., Wang, S., Payre, F., and Stern, D.L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature *466*, 490–493.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frébourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol. *653*, 249–257.

Gasperini, M., Starita, L., and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. Nat. Protoc. *11*, 1782–1787.

Gasperini, M., Findlay, G.M., McKenna, A., Milbank, J.H., Lee, C., Zhang, M.D., Cusanovich, D.A., and Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am. J. Hum. Genet. *101*, 192–205.

Gibson, G. (2012). Rare and common variants: twenty arguments. Nat. Rev. Genet. *13*, 135–145.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx

(eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. *47*, 276–283.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Hernandez, R.D., Uricchio, L.H., Hartman, K., Ye, J., Dahl, A., and Zaitlen, N. (2017). Singleton Variants Dominate the Genetic Architecture of Human Gene Expression.

Hong, J.-W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.

Hsiao, Y.-H.E., Bahn, J.H., Yang, Y., Lin, X., Tran, S., Yang, E.-W., Quinones-Valdez, G., and Xiao, X. (2018). RNA editing in nascent RNA affects pre-mRNA splicing. Genome Res. *28*, 812–823.

Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624.

Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet. Med. *16*, 497–503.

Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J., and Lehner, B. (2016). The complete local genotype–phenotype landscape for the alternative splicing of a human exon. Nat. Commun. *7*, 11558.

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

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

Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. *11*, 345–355.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Kosuri, S., and Church, G.M. (2014). Large-scale de novo DNA synthesis: technologies and

applications. Nat. Methods *11*, 499–507.

Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research, *42*(D1), D980-D985.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

LeProust, E.M., Peck, B.J., Spirin, K., McCuen, H.B., Moore, B., Namsaraev, E., and Caruthers, M.H. (2010). Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res. *38*, 2522–2540.

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. Proc. Natl. Acad. Sci. U. S. A. *100*, 189–192.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al. (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600–604.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10.

Matreyek, K.A., Stephany, J.J., and Fowler, D.M. (2017). A platform for functional assessment of large variant libraries in mammalian cells. Nucleic Acids Res. *45*, e102.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E.T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. *7*, e1002144.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.

Ongen, H., and Dermitzakis, E.T. (2015). Alternative splicing QTLs in European and African populations using Altrans, a novel method for splice junction quantification.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature *554*, 239.

Pala, M., Zappala, Z., Marongiu, M., Li, X., Davis, J.R., Cusano, R., Crobu, F., Kukurba, K.R., Gloudemans, M.J., Reinier, F., et al. (2017). Population- and individual-specific regulatory variation in Sardinia. Nat. Genet. *49*, 700–707.

Plesa, C., Sidore, A.M., Lubock, N.B., Zhang, D., and Kosuri, S. (2018). Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. Science *359*, 343–347.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. *38*, D105–D110.

Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics *31*, 761–763.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr. Protoc. Bioinformatics *47*, 11.12.1–34.

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M.D., Banerjee, B., Syed, T., Emons, B.J.M., Gifford, D.K., and Sherwood, R.I. (2016). High-throughput mapping of regulatory DNA. Nat. Biotechnol. *34*, 167–174.

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., UK Brain Expression Consortium, North American Brain Expression Consortium, Coin, L., et al. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci. *17*, 1418–1428.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Sanjana, N.E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. Science *353*, 1545–1549.

Schafer, S., Miao, K., Benson, C.C., Heinig, M., Cook, S.A., and Hubner, N. (2015). Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). Curr. Protoc. Hum. Genet. *87*, 11.16.1–14.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.

Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536–1543.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Smigielski, E.M. (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. *28*, 352–355.

Smith, S.A., and Lynch, K.W. (2014). Cell-based splicing of minigenes. Methods Mol. Biol. *1126*, 243–255.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855.

Takata, A., Matsumoto, N., & Kato, T. (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. Nature communications, 8, 14519.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

UK10K Consortium, Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

Uricchio, L.H., Zaitlen, N.A., Ye, C.J., Witte, J.S., and Hernandez, R.D. (2016). Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res. *26*, 863–873.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.1–33.

Voelker, R.B., and Berglund, J.A. (2007). A comprehensive computational characterization of conserved mammalian intronic sequences reveals conserved motifs associated with constitutive and alternative splicing. Genome Res. *17*, 1023–1033.

Wu, X., and Bartel, D.P. (2017). kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Res.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Zhang, X., Joehanes, R., Chen, B.H., Huan, T., Ying, S., Munson, P.J., Johnson, A.D., Levy, D., and O'Donnell, C.J. (2015). Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat. Genet. *47*, 345–352.

CHAPTER THREE

Comprehensive Functional Characterization of *Escherichia coli* Promoters Reveals Key Components of Transcriptional Regulation

Title: Comprehensive Functional Characterization of *Escherichia coli* Promoters Reveals Key Components of Transcriptional Regulation

Authors:
Guillaume Urtecho[1†], Kimberly D. Insigne[2†], Arielle D. Tripp[3], Marcia Brinck[4], Nathan B. Lubock[5], Hwangbeom Kim[5], Tracey Chan[2], & Sriram Kosuri[5,6,7*]

Affiliations:
[1]Molecular Biology Interdepartmental Doctoral Program, University of California, Los Angeles, CA, 90095, USA
[2] Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, CA 90095, USA
[3] Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA, 90095, USA
[4] Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, California, USA
[5] Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA
[6] Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA
[7] UCLA-DOE Institute for Genomics and Proteomics, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed. Tel: +1 310 825 8931; Email: sri@ucla.edu

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

SUMMARY

Promoter sequence space in bacteria is vast and difficult to study genome-wide due to extraneous factors affecting transcript levels. Here, we use a genomically-encoded massively parallel reporter assay (MPRA) to characterize the global *E. coli* promoter landscape and dissect active promoters for motifs encoding promoter regulation. We measure promoter activity of over 300,000 sequences spanning the entire genome and identify 3,321 active promoter regions in glucose minimal media and 3,477 in rich LB media. We show that antisense promoters have a profound effect on global transcript levels and how codon usage has adapted to encode intragenic promoters. Furthermore, we perform a scanning mutagenesis of 2,057 *E. coli* promoters to identify regulatory sequences. Lastly, we implement a variety of machine learning approaches to classify promoters and predict activity. In summary, we present a series of approaches to rapidly characterize promoter sequences within a bacterial genome.

Keywords: Transcription, promoter, RNA Polymerase, promoter landscape, promoter prediction, massively parallel reporter assay, antisense promoter, transcription factors, sequence-function relationships

INTRODUCTION

In 1961, François Jacob and Jacques Monod published a review laying the foundation for what would later be known as the *Escherichia coli* promoter[1]. This seminal work has sparked countless studies delving into the many molecular mechanisms operating at promoters, establishing them as one of the most well-characterized systems in biology. Several key promoters have been the subject of in-depth biochemical and structural studies describing the mechanisms by which the RNA polymerase (RNAP) recognizes promoter sequences, as well as the stepwise process to engage transcription[2–4]. In addition, many transcription factors have been described in similar detail, revealing the multiple mechanisms through which these proteins interact with promoters to modulate the behavior of RNAP and activity of the promoter[5–8]. Lastly, the binding motifs for the majority of these proteins are known and have been studied at high resolution using next-generation sequencing methods[9–12]. In short, the myriad components of this system have been extensively cataloged and characterized, giving the appearance that bacterial promoters are a 'solved' biological phenomenon.

Despite extensive research on individual *E. coli* promoters, when we examine the entire genome we cannot address the fundamental questions of where promoters exist or how these sequences encode transcription. Although the consensus sequences for RNAP recognition motifs have been known for decades, a simple search of the genome based on those motifs yields many false positives. In fact, true promoters within a given region often do not exhibit the greatest similarity to the consensus[13,14]. Experimental efforts to identify promoters using 5' RNA-Seq likely also

suffer from false positives, identifying tens of thousands of putative transcription start sites (TSSs) with little overlap between studies[15,16]. Furthermore, although many *E. coli* promoters have been identified with strong supporting evidence[17], functional annotation of these promoters remains challenging due to the universally degenerate nature of transcription factor binding motifs. As a result, roughly two-thirds of the 2,565 reported *E. coli* operons do not contain any transcription factor binding site annotations[17,18]. Finally, even amongst fully annotated promoter sequences we are still unable to quantitatively predict the activity or behavior of these supposedly well-characterized systems due to a lack of understanding of how the arrangement and sequence composition of binding motifs relate to activity.

There are various confounding factors which complicate a systematic understanding of genome-wide promoter function. Genome-wide characterizations of endogenous promoters in their native contexts are influenced by various mechanisms contributing to perceived levels of transcription. For instance, recent work has shown that promoter activity varies depending on the location in the genome due to factors such as variance in chromosomal copy number[19,20], the distribution of transcription factors within the cell[21], and the accessibility of the chromatin[22]. In addition, RNAs produced by promoters are subject to dynamic degradation processes which alter transcript levels[23,24]. Lastly, mechanisms of transcriptional interference, such as RNAP collisions directed by antisense promoters, have the potential to further modulate activity of promoters, although the impact of this mechanism across the genome has not been fully investigated[25–27]. These extraneous mechanisms confound a deeper understanding of promoter sequences, which will require novel methods to circumvent these factors.

In this work, we dissect promoter regulation in *E. coli* using a massively-parallel reporter assay (MPRA) designed to isolate promoter activity from extraneous mechanisms of genetic regulation. This system can characterize the promoter activity of hundreds of thousands of sequences in a

common genomic location that is insulated from local transcriptional interference and uses a standardized transcript to remove the effects of mRNA degradation on expression levels. We use this powerful assay to identify promoters throughout the *E. coli* genome and systematically dissect their regulatory motifs which encode promoter activity. We measure promoter activity of 17,189 reported transcription start sites (TSSs) and find that a majority of reported TSSs are likely due to transcriptional noise, rather than productive transcription. Furthermore, we measure promoter activity of 321,123 DNA fragments spanning both strands of the *E. coli* genome with 8.5x coverage, allowing us to identify the breadth of endogenous sequences that coordinate expression in rich and minimal media. This genome-wide promoter screen reaffirms the pervasive nature of antisense transcription and its global role in transcriptional interference. To characterize sequence motifs encoding promoter activity in *E. coli*, we implement a scanning mutagenesis approach to systematically dissect the sequences of 2,057 active promoters. With this approach, we characterize the regulatory impact of 568 transcription factor binding sites reported by RegulonDB as well as 2,583 novel sites, thereby providing functionally annotated profiles for promoters driving expression in rich LB media for 1,158 of the 2,565[28] operons in *E. coli*. Lastly, we use this rich dataset to train various machine learning models to identify functional *E. coli* promoter sequences and predict their expression levels. In summary, we present a series of high-throughput approaches to identify and characterize promoters throughout the entire *E. coli* genome.

RESULTS

Functional characterization of 17,635 previously reported *E. coli* promoters

To experimentally characterize endogenous *E. coli* promoters, we measured the transcriptional output of 17,189 previously reported transcriptional start sites (TSSs) from MG1655 cells grown in rich LB media. We assembled these TSSs from three sources: the RegulonDB *E. coli* database[17] (8,486 unique TSSs), a directional RNA-Seq study by Wanner et. al (2,123 unique

TSSs), and a RNA-Seq study by Storz et. al (14,868 unique TSSs). These three sources contain 23,798 unique TSSs, many of which are within a few bases of each other. To minimize redundancy, we collapsed clusters of TSSs within 20 bp of each other into the most upstream TSS, reducing the total number for synthesis to 17,635. Surprisingly, there was little agreement regarding the location of TSSs between studies, with only 94 shared between all three when considering exact matches (**Figure 3.1A**). Furthermore, the identification of 17,635 TSSs is surprising considering the *E. coli* genome contains 4,419 known genes. While this could be the result of an immensely complex transcriptional system, it begs the question of whether or not these transcriptional signals are the result of genuine promoters, transcriptional noise, or experimental artifacts.

Therefore, we sought to determine whether these previously identified TSSs were indicators of true promoter activity and identify the sequence elements that distinguished transcriptionally active from inactive sequences. To this end, we implemented a previously described MPRA[29] to quantitatively measure the individual promoter activity of 17,635 TSSs (**Figure 3.1B**). For each TSS we synthesized oligonucleotides spanning 120 bp upstream to 30 bp downstream of the TSS, which has been reported to encode the majority of promoter activity driving expression at a given TSS[30]. In addition, we synthesized 96 well-characterized promoters from the BioBricks registry that are designed to span a wide range of expression and serve as positive controls. We included 500 negative controls, identical in length to the main library and randomly selected across the genome, that are more than 200 bp from the nearest TSS and are assumed to be transcriptionally inactive. We confidently measured 97.5% (17,767/18,222) of the synthesized library including controls, and 97.4% (17,189/17,635) of reported TSSs, with an average of 69.5 barcodes per library member (**Figure 3.S1A**). Next, we integrated this pooled library of reporter constructs into the *nth-ydgR* intergenic locus within the *E. coli* chromosomal terminus using a recombination-mediated cassette exchange system[31]. We determined expression levels by

quantifying the transcript abundance of each barcode normalized to the DNA-seq abundances, and precisely measured 96.6% of promoters in this library (**Figure 3.1C**). We compare expression of the TSS promoters relative to the negative controls to set a threshold for active and inactive promoters, which was two standard deviations greater than the median negative control (**Figure 3.1D**). Among the 17,635 original TSSs, we identified 2,670 promoters that greater expression than our experimentally determined threshold in rich LB media (**Figure 3.1E**). Notably, this amount is more consistent with the number of operons identified by a recent study using long-read sequencing to characterize full-length *E. coli* transcripts[32]. Amongst these 2,670 confirmed promoters, we recovered expression data for many well-known promoters. In particular, three of the strongest TSS promoters identified corresponded to the 16S and 23S polycistronic operon, the most abundantly expressed operon in the *E. coli* genome.

Location-dependent promoter activity is mostly constant across promoters

An early concern arose regarding whether the genomic location where we integrated our promoter library had a significant effect on the expression values determined by our MPRA. Indeed, several recent studies have shown that promoter expression levels can be highly variable between locations of the genome[21,22,33]. However, these studies have looked at these effects on single promoters, so it is unclear whether location-dependent effects are promoter specific or general. To determine whether genomic position had an effect on our measurements, we also integrated the TSS promoter library in both chromosomal mid-replichores and compared expression measurements between these positions and the *E. coli* chromosomal terminus (**Figure 3.S1B**). We observed that promoter measurements remained highly consistent between integrated locations, although the mid-replichore positions exhibited slightly higher concordance than either to the terminus. It is worth mentioning that RNA-Seq provides us with relative expression between promoters at each location. Therefore, although the relative activity levels between promoters are consistent at different locations, it is likely that the absolute transcription levels differ due to

previously described chromosomal position effects[22]. We conclude that relative promoter activity levels acquired from our assay are consistent between the three locations tested, which suggests that genome-position effects on expression are predominately consistent between promoters.

The promoter architecture of inactive promoters resemble tssRNA-associated promoters

An overwhelming majority of *E. coli* promoters are directly regulated by the housekeeping sigma factor σ70, thus we expected that active promoters would be enriched for the canonical σ70 motifs. Promoters of the σ70 family are well known for containing two hexamer motifs, the -10 and -35 motifs, which recruit RNAP and are named after their position relative to the TSS. We used a broadly accepted σ70 PWM[13] to analyze whether active TSS promoters were enriched for these motifs. Interestingly, although both active and inactive promoters were enriched for the canonical -10 motif compared to our negative controls ($p < 2.2 \times 10^{-16}$, $p < 3.8 \times 10^{-7}$) we found that the -35 scores of inactive promoters were generally no greater than negative controls ($p = 0.21$) (**Figure 3.1E**). Recent work has shown that promoters containing a -10 but lacking a correctly positioned -35 motif allow for the production of short (approximately 35-50 bp) transcripts. These transcripts are known as tssRNAs and do not result in mature, translated products[34,35]. Notably, these short transcripts could be indistinguishable from biologically productive transcripts in 5' RNA-Seq studies, yet would not be detected by our assay, which requires transcription of a barcode within the 3' UTR of the reporter gene. Thus, we find that the sequences encoding inactive TSSs bear a greater resemblance to tssRNA-associated promoters rather than true promoters.

Genome-wide survey of the *E. coli* promoter landscape

Despite functionally screening of 17,635 previously implicated genomic regions for promoter activity, we remained unconvinced that we had captured the entire assortment of endogenous promoters in *E. coli*. Indeed, we discovered several cases where our screening active TSSs did

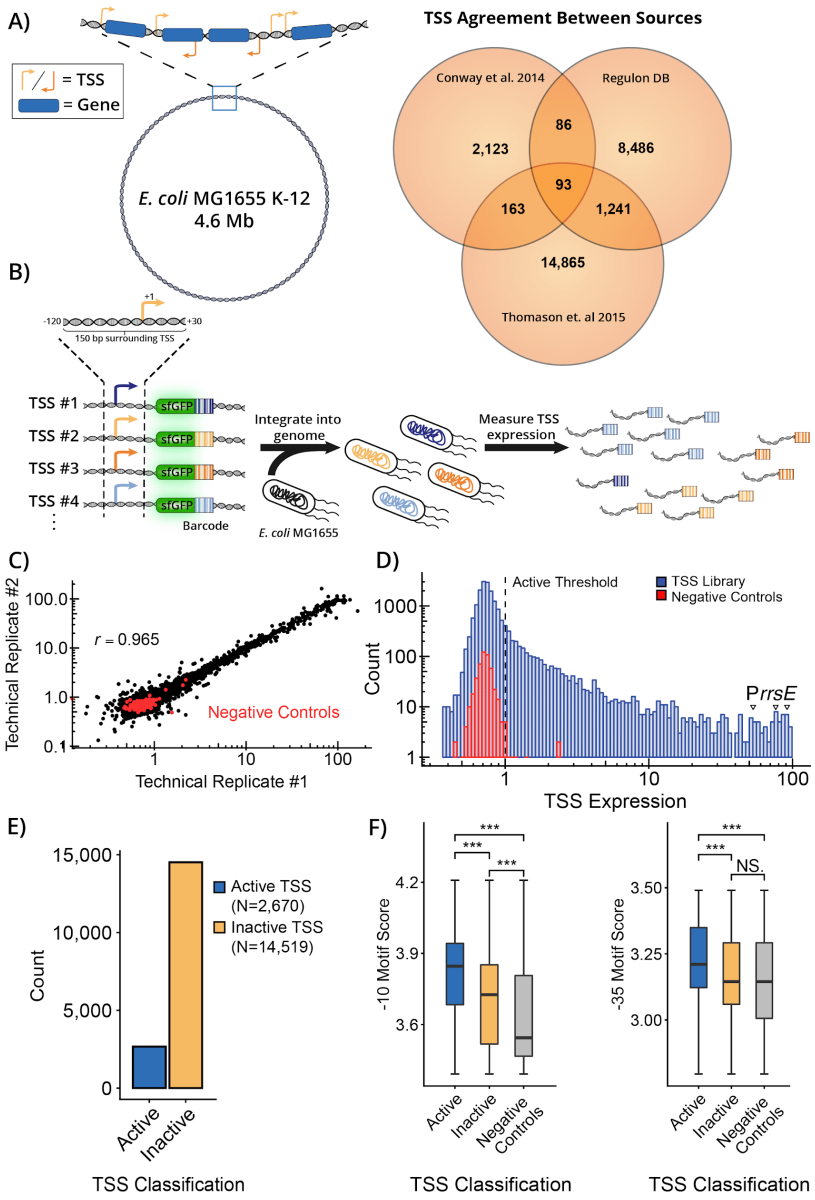


Figure 3.1. Functional characterization of 17,635 previously reported *E. coli* promoters. **A)** We synthesized 17,635 previously identified TSSs and the surrounding -120 to +30 bp context and genomically integrated the construct at a fixed landing pad in the *E. coli* genome (left). There is little agreement between the three different sources at the single nucleotide level (right). **B)** Massively parallel reporter assay (MPRA) (previously described in Urtecho et. al 2018[29]) captures quantitative measurement of transcriptional activity for individual TSSs in multiplex. **C)** MPRA is highly replicable across technical replicates (r = 0.965. **D)** The TSS library spans a wide functional range, over 100-fold, with negative controls exhibiting low levels of expression. **E)** The majority of tested TSSs are inactive in rich LB medium. We report 2,670 active promoters as having expression two standard deviations greater than the median of the negative controls. **F)** Active and inactive TSSs have significantly different mean scores for the two core promoter motifs (Student's t-test, two-sided, "***"=<0.001, "**"=<0.01, "*"=<0.05).

not identify promoters for several essential operons (*data not shown)*, which implied that our initial approach had not completely sampled the entire promoter landscape.

We adapted our MPRA to screen a library of sequences completely spanning the entire *E. coli* genome (**Figure 3.2A**) and discover the full catalog of endogenous promoters. We created this library using sonication to randomly shear the *E. coli* genome, isolate fragments between 200 and 300 bp, and barcode each fragment before passaging them through our MPRA. Using this pipeline, we measured the transcriptional activity of 321,123 fragments with a median size of 244 bp spanning the entire double-stranded *E. coli* genome with an average of 8.5x coverage (**Figure 3.S2A, Figure 3.S2B**). We averaged the expression of all fragments overlapping each genomic position to achieve highly replicable values of promoter activity at single-nucleotide resolution and in a strand-specific manner (**Figure 3.S2C**). We have created a custom visualization for this data, which reveals remarkably defined regions of promoter activity across the entire *E. coli* genome (**Figure 3.2B**). The detected promoter signals are in strong agreement with our previous characterization of TSS-associated promoters, with active TSSs showing greater signals for promoter activity compared to inactive TSSs (**Figure 3.2C**). We identify “promoter regions” considering contiguous regions of at least 60 bp with activity measurements higher than an empirically-derived threshold determined from previously identified TSS-associated promoters (described in methods). Thus, we find that this approach is capable of rapidly screening entire bacterial genomes for regions demonstrating promoter activity.

The *E. coli* promoter landscape is dynamic in response to environmental conditions

We used this approach to explore the complex landscape of promoters in *E. coli* and its dynamic rearrangement in response to environmental changes. We measured expression of our fragment library grown to exponential phase in rich LB media as well as a defined minimal media supplemented with glucose. We predicted that under minimal media conditions global promoter

activity would increase in order to accommodate an increased dependence on endogenous biosynthetic pathways to create necessary resources. Despite our expectations, we identified 3,321 active promoters in minimal media and a comparable 3,477 in rich media (**Figure 3.2D**). However, although the absolute number of promoters between these conditions is similar, the identity of active promoters was variable between conditions, with only 2,466 shared by both conditions.

We reasoned that this dynamic promoter response was mediated by condition-dependent transcription factors and evaluated the TFBS composition of promoters unique to each condition (**Figure 3.2E**). Examining TFBSs reported by RegulonDB, we found 324 TFBS annotations unique to rich media promoters and 370 overlapping promoters unique to glucose minimal media conditions. Upon comparing TFBS content of these promoters, we found that binding sites for several global transcriptional regulators[36], including ArcA, Lrp, and Fis occurred at similar frequencies between these conditions. Conversely, binding sites for CRP, an abundant glucose-inhibited transcription factor, were enriched by roughly three-fold amongst rich media promoters when compared to promoters active in glucose minimal media. Interestingly, we found many TFBSs that appear to be nearly or entirely condition-dependent and these appeared to generally be local regulators. We conclude that the remodeling of the promoter landscape between these conditions is primarily determined by global regulation of CRP as well as many condition-dependent transcription factors with more local effects.

Antisense transcription is a pervasive means of repression

Having identified the arrangement of promoters throughout the genome, we investigated how their positioning influenced local transcription profiles. We first evaluated the position of promoters in the genome and identified 1,465 and 1,623 intergenic promoters in LB and M9, respectively. These were predominantly positioned in the sense orientation relative to the nearest downstream

gene. Surprisingly, 1,998 and 1,660 intragenic promoters were identified in LB and M9, respectively, and nearly half of these promoters were positioned antisense relative to the genes they regulate (**Figure 3.2F**).

The large amount of antisense promoters we discovered encouraged us to explore whether these antisense promoters played a significant role in global expression levels. Although promoters are primarily thought to be positioned upstream of the genes they regulate, many recent studies have focused on their impact when positioned within and downstream of transcriptional units [12,37]. In particular, these alternative positions have generated significant interest in the cases where promoters are poised for *cis*-antisense transcription, a transcriptional interference mechanism where promoters drive transcription against the direction of genes, typically resulting in transcriptional repression[27,37,38]. However, the significance of antisense transcription has recently come under question as being predominantly associated with transcriptional noise that is biologically irrelevant[39].

To explore the impact of antisense transcription, we evaluated whether genes regulated by antisense promoters are associated with a significant reduction in expression. We performed RNA-Seq on MG1655 grown in minimal glucose media and compared the transcript coverage of all genes with sense promoters, antisense promoters, and both sense and antisense promoters. We found that overall, genes with both sense and antisense promoters exhibited a two-fold decrease in expression compared to strictly sense-regulated genes, and this lower expression occurred despite nearly identical levels of sense-promoter activity (**Figure 3.2G**). To explore the quantitative nature of antisense promoter repression, we separated genes by the strength of their sense and antisense promoters into quartiles and determined expression levels of these genes (**Figure 3.2H)**. This analysis revealed an interplay between the strength of a gene's sense

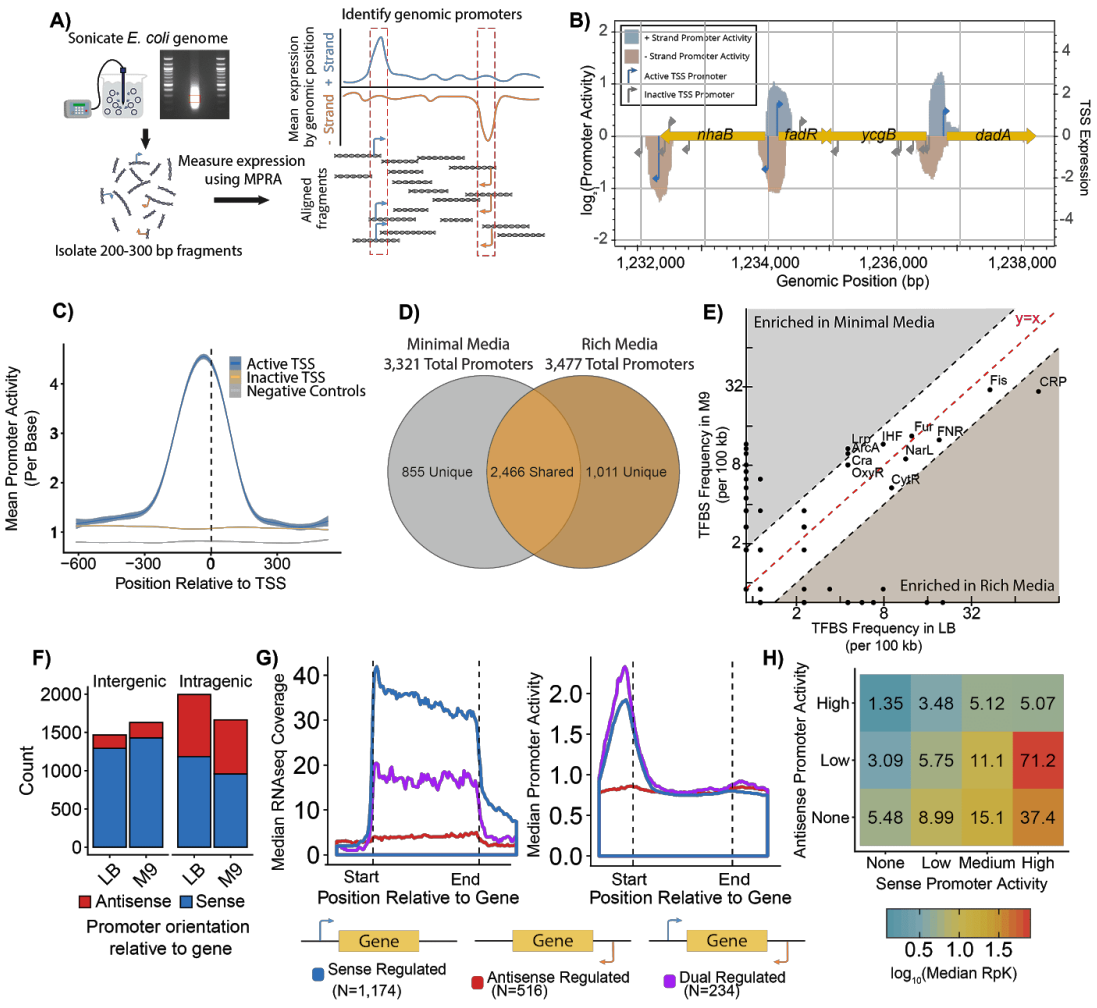


Figure 3.2. Genome-wide survey of the *E. coli* promoter landscape. **A)** We adapted our MPRA to screen genome fragments spanning the entire genome. Fragments were derived using sonication to generate 321,123 fragments of 200 to 300 bp in size with an average of 10x coverage. **B)** We developed custom visualization using Bokeh that enables users to select any genomic position and display multiple layers of information, including our TSS MPRA activity (Figure 1), genome fragmentation screen, and existing gene annotations. **C)** Meta-analysis of mean promoter activity at experimentally validated active TSSs, inactive TSSs, and negative controls. TSSs previously determined to be active are enriched for promoter activity in our genomic fragment screen. **D)** Overlap of promoters active in M9 Minimal media + .2% glucose and rich LB media. **E)** Frequencies of TFBSs overlapping unique, condition-dependent promoters. TFBS locations are reported by RegulonDB and frequencies are normalized to 100 kb of promoter sequence. Dotted-black lines indicate a two-fold enrichment. **F)** Orientation and positioning of identified promoters separated by condition. **G)** Left: Meta-gene analysis showing median RNA-Seq read coverage across all sense, antisense, and dual regulated genes. Middle: Sense promoter activity at sense, antisense, and dual regulated genes. **H)** Median RNA-Seq coverage per kb across genes separated by sense and antisense promoter strength.

promoter and antisense promoter, where increasing sense strength was associated with increased expression, whereas increased antisense strength resulted in an overall decrease in expression. Thus, we conclude that this analysis provides clear evidence in support of the impact of antisense-regulation on global transcript levels.

Fine-mapping of *E. coli* promoters within transcriptionally active regions

Our assay of genomic fragments identified regions of promoter activity that were fairly broad (**Figure 3.S2D**) and well above the expected size of typical promoters[30]. Our next goal was to explore these regions further and reveal concise boundaries for sequence elements encoding promoter activity. To finely map promoter sequence elements, we implemented a synthetic tiling approach to study these promoter regions in greater detail. We used our MPRA assay to measure expression of 46,713 150 bp oligos tiling the length of each of the 3,477 active promoter regions identified in rich media in 10 bp intervals (**Figure 3.3A**). For active promoter regions under 150 bp, we measured expression of a single oligo consisting of the entire region. This approach allowed us to precisely identify the boundaries of sequence elements encoding promoter activity by determining where along the promoter region oligo tiles gained and lost expression.

Interestingly, this analysis showed many of the broad promoter regions we previously identified actually contained multiple discrete promoters (**Figure 3.3B**). As might be expected, the number of promoters within a given region corresponded with the size of the region identified by our genome fragmentation promoter assay (**Figure 3.S4A**). However, we could not identify active oligo tiles for 1,476 of the promoter regions previously identified as active. Regions without promoters were generally less than 150 bp in length, suggesting the entire functional promoter sequence was not captured. These regions without promoters were generally under 150 bp in length, therefore we may not have tested the entire promoter sequence. Nonetheless, we precisely mapped the promoter sequences within 1,599 of the original 3,480 promoter regions

identified in rich media as well as 400 containing multiple discrete promoters (**Figure 3.3C**). Furthermore, with this approach we could infer the minimal sequence necessary for promoter activity at each sub promoter by determining the overlap of all active tiles composing each sub promoter. When comparing the sizes of the minimal sequence necessary for promoter activity, we observed an enrichment of approximately 40 bp which is a typical size for σ70 promoters[40–42], the most abundant class of promoters in *E. coli* (**Figure 3.3D**). We also observed an enrichment for 150 bp minimal promoter regions, although these were generally weak indicating that our resolution had been limited when tiling weaker promoters. Overall, we were able to identify precise boundaries for 2,228 promoters active in rich media.

Having identified the minimal sequences necessary for promoter activity, we sought to explore how these promoters were encoded within intragenic regions. The sequences of intragenic promoters are inherently constrained by the coding regions they overlap and so we were curious how the *E. coli* genome had adapted sequence content to enable promoter activity within these restrictions. After comparing the amino acid composition within intragenic promoters, we found that these sequences were especially enriched for containing STOP codons, and showed a preference for many many other amino acids (**Figure 3.3E**). This enrichment for STOP codons is compelling, as two of the three codons, TGA and TAA match the consensus motifs of sigma70 promoters (the -35 T**TGA**CA motif as well as the -10 TA**TAA** motif). Next, we explored whether codon selection had been influenced to encode intragenic promoters. Indeed, we found that codon usage within intragenic promoters is significantly biased towards certain codons (**Figure 3.3F**). In particular, we found the strongest bias amongst arginine codons, with a strong preference for AGA and AGG. Interestingly, it appears that codons enriched within intragenic promoters are typically rare in the genome. This surprising new role for rare codons may partially explain recent findings that synonymous codons were unequivolent in genome recoding

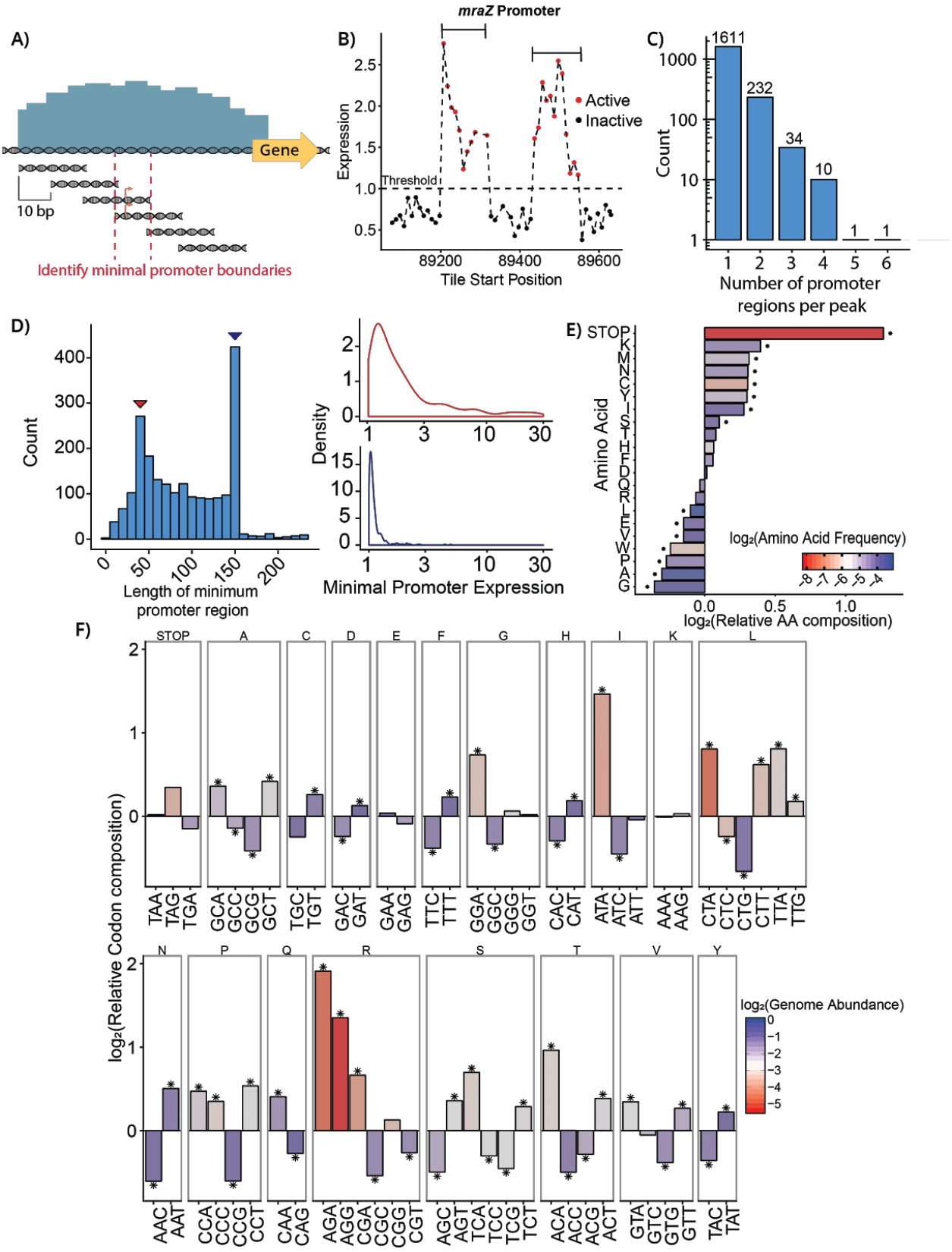

A)
Gene
10 bp
Identify minimal promoter boundaries
B)
mraZ Promoter
Expression
Threshold
Active
Inactive
Tile Start Position
C)
Count
1611
232
34
10
1
1
Number of promoter regions per peak
D)
Count
Length of minimum promoter region
Density
Minimal Promoter Expression
E)
STOP
Amino Acid
log2(Amino Acid Frequency)
log2(Relative AA composition)
F)
log2(Relative Codon composition)
log2(Genome Abundance)

Figure 3.3. High-resolution tiling of promoter regions identifies sequences encoding promoter activity. A) We synthesized 150 bp oligos tiling all promoter regions identified in rich media at 10 bp intervals. We determine minimal promoter boundaries by identifying overlapping segments of transcriptionally active tiles. **B)** Expression of oligo tiles spanning the *mraZ* promoter. Points are shown at the right-most position of their corresponding oligo. Interval bars show two distinct sub promoters within this region. The threshold for active promoters was set to two standard deviations greater than the expression of a set of 500 negative controls. **C)** Number of sub promoters may encode multiple sub promoters. Sub promoters were considered distinct if they were separated by more than 40 bp from other active oligo tiles. **D)** Distribution of the lengths of the minimal sequence encoding promoter activity for each sub promoter. The minimal sequence is acquired from the overlap of all active oligo tiles composing each sub promoter. **E)** Amino acid enrichment within intragenic promoters relative to whole genome amino acid frequencies. **F)** Codon bias within intragenic promoters relative to whole genome. Codons are colored by the relative usage compared to other synonymous codons.

projects[43]. Overall, our findings suggest that the *E. coli* genome has evolved to encode intragenic promoters by manipulating codon usage.

Mutational scanning of 2,057 active promoters in *E. coli*

Our next goal of this work was to develop an approach to identify motifs responsible for transcriptional regulation within the identified promoters. Recent work by Belliveau et al.[18] demonstrated a high-resolution saturation mutagenesis approach for identifying regulatory motifs within entirely uncharacterized promoters on an individual basis. Inspired by this work, we implemented a scanning mutagenesis strategy to explore the sequence features that defined active promoters. For each of the 2,057 active TSS-associated promoters identified in rich LB media, we systematically scrambled 10 bp sequences spanning the -120 to +30 positions at five bp intervals (**Figure 3.4A**). Although single-point mutations can provide detailed information on how individual bases encode regulation at promoters[9,18], we chose 10 bp scrambling mutations to evaluate the contribution of each site as a whole. Furthermore, these scrambled sequences were designed to maximize distance from the original sequence, thereby further guaranteeing that we could obviate any motifs at each position contributing to transcription regulation. Using this approach, we would expect that disrupting a repressor site would increase expression, whereas scrambling an RNAP or activator site would decrease expression. In total, we designed

a library of 59,653 sequences consisting of nearly all 2,057 active TSS-associated promoters and their scrambled variants and measured their expression using our genomically-encoded reporter assay. We recovered expression measurements in rich LB media for 89% (52,900/59,653) of this library, with an average of 8 barcodes per scrambled variant (**Figure 3.S3C, 3.S3D**). By scrambling sequences across active promoters, we could identify regions within these sequences that either increased or reduced expression (**Figure 3.4B**). As expected, these sequences were enriched for regions that increased expression at the -35 and -10 regions in addition to many other regions within these promoters that modulated expression. However, we also identified many sequences throughout these promoters that also contributed to regulation.

Scanning mutagenesis reconfirms previously validated regulatory sequences

We explored if our data could reaffirm known regulatory motifs to determine if we could effectively identify potentially novel regulatory elements. We first examined our scanning mutagenesis of the *lacZYA* promoter, a classic gene regulation model whose regulatory motifs are well characterized. This promoter is an excellent example as it is known to contain a variety of regulatory motifs, including twin LacI repressor sites centered at +11 and -82[44], a CAP activator site centered at -61[45], and a σ70 RNAP binding site. Our scanning mutagenesis of *lacZ* promoter revealed distinct signals corresponding with each of these sites, as well as quantitative measurements for their contribution to expression (**Figure 3.4C**). Scanning mutagenesis of the *relBE* promoter achieved similar results, identifying a previously reported RelBE repressor site at the +1 position[46] as well as the -10 and -35 σ70 recognition motifs[46]. This evidence suggests that our scanning mutagenesis approach is amenable to identifying functional regulatory elements within *E. coli* promoters.

Given that our approach could capture the effects of known binding sites, we next explored whether we could effectively identify novel regulatory sites within promoters. Although we

performed this scanning mutagenesis for 2,057 promoters we choose to highlight a few examples to demonstrate the utility of this method. The cyclopropane fatty acyl phospholipid synthase gene, *cfa,* exhibits dynamic expression[47] and is responsible for a major component of the cell membrane necessary for cell survival in acidic conditions[48]. While there have been several transcription factors implicated in regulation of *cfa*, the motifs responsible for its direct regulation are still unknown. Our scanning mutagenesis approach has identified a candidate σ70 promoter regulating this gene with a -10 motif centered 34 nucleotides upstream of the tested TSS as well as two novel repressor sites located in the spacer region and upstream of the -35 motif. As another example of newly characterized promoter sequences we have also identified novel regulatory motifs for *rpsL*, an essential gene and component of the 30S ribosomal subunit. For this gene, we have found a candidate σ70 RNAP binding site as well as an unknown repressor positioned over the transcription start site which, once obviated, results in a 5-fold increase in expression of the promoter. Although further experiments[18] are necessary to accurately name the transcription factors acting at the regulatory sites acting on these promoters, these results provide clear insights into the regulation of these genes and implicate strong candidate motifs for further dissection.

Global identification of 7,293 *E. coli* promoter regulatory motifs

Next, we expanded the scope of our analysis to characterize regulatory sites at the global level. We used the individual barcode measurements, across four replicates, to find significant differences between the mean expression of the unscrambled sequence and the scrambled sequence (Student's t-test with FDR at 1%). We identified scrambled regions that significantly increased or decreased expression and found 1,885 and 5,408 regions, respectively (**Figure 3.5A**). These sites were located throughout promoters and scrambling these sites resulted in dramatic changes in expression, some over 100-fold (**Figure 3.S5A**). We observed markedly different distributions for the position of regions that either increased or decreased expression

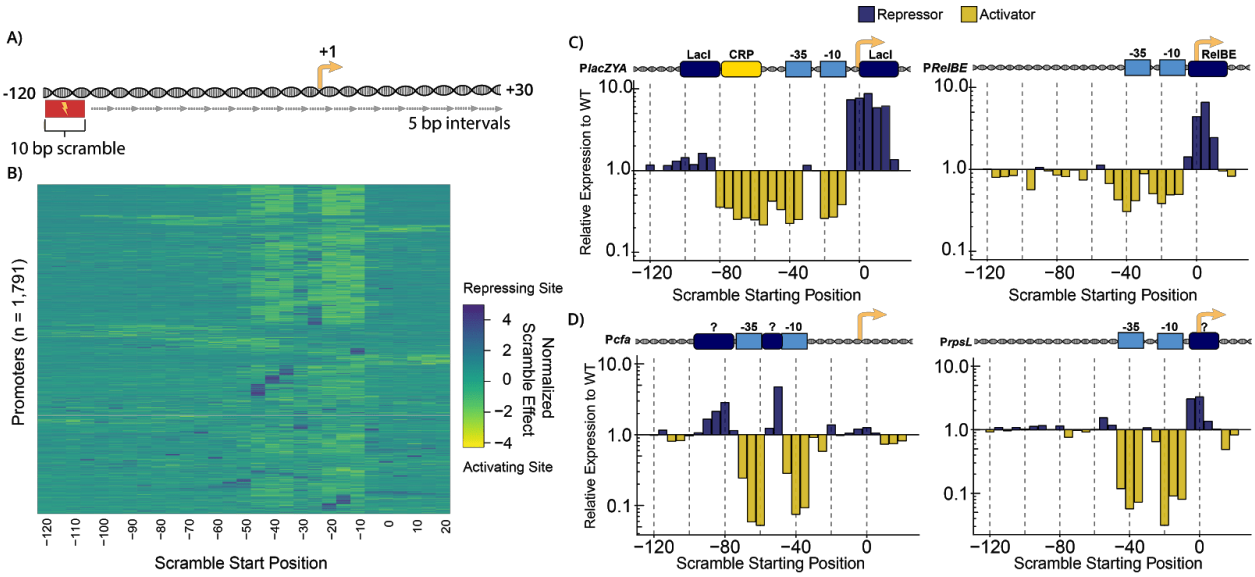


Figure 3.4. Scanning mutagenesis of 2,057 TSS-associated promoters identifies known and novel regulatory motifs A) Each of the 2,057 active TSS-associated promoters were subject to a scanning mutagenesis to identify motifs encoding regulatory activity. We synthesized 59,653 promoter variants in which we scrambled 10 bp sequences for each 5 bp interval across each promoter. **B)** Global promoter scanning mutagenesis profiles. Top: Averaged relative activity of variants scrambled at each position. Bottom: Heatmap representing the impact of mutating each position for 1,826 promoters. Rows are rearranged using hierarchical clustering and the intensities are normalized within each row to accommodate differences in unscrambled promoter activities. **C)** Left: Scanning mutagenesis of the well-characterized *lacZYA* promoter. Right: Scanning mutagenesis of the well-characterized *relBE* promoter. **D)** Left: Scanning mutagenesis of the *cfa* promoter. Right: Scanning mutagenesis of the *rpsL* promoter.

(**Figure 3.5B**). Increased regions were particularly enriched at the -10, -35, and -70 positions, which is consistent with the canonical σ70 RNAP binding motif as well as the typical position of transcriptional activators amongst class I bacterial promoters [49–51]. On the other hand, reduced regions localized to the TSS, spacer, and -35, which is consistent with known mechanisms of RNAP occlusion by steric hindrance[51,52].

Identified motif effects generally agree with reported annotations

Having identified 7,293 regulatory regions throughout the *E. coli* genome, we wanted to cross-reference these with the extensive collection of putative and experimentally determined regulatory sites reported by RegulonDB. Of the 2,453 unique transcription factor binding sites (TFBSs) reported by RegulonDB, 1156 overlap with regulatory sites whose effects were captured by our scanning mutagenesis. We identified at least one scramble that significantly changed expression for 49% (5667/1,156) of these previously annotated sites. After merging contiguous significant scrambled sites into distinct regulatory regions we identified 1,414 and 1,903 merged sites that increase or reduce expression, respectively. Sites were, on average, 20 bp (**Figure 3.S5B**) with effect sizes of these scrambles largely independent of their lengths (**Figure 3.S5C**). Our scrambling results agreed with the reported impact for 65% (185/283) of activators and 43% (196/450) of repressors (**Figure 3.5B**). Our lower concordance for repressors could be due to a scramble disrupting both a repressor and -35 or -10 element, resulting in a decrease in expression which would appear to contradict a reported repressor site. We looked at the distribution of concordance for merged scrambles by position relative to the TSS (**Figure 3.5SD, 3.5SE**) and observed a higher proportion of disagreement near the -35 and -10 elements, suggesting these scrambles are disrupting crucial promoter elements. This may be expected considering that many repressor operate by binding regions proximal to the RNAP binding site. Regardless, we found several examples where the regulatory effects predicted by RegulonDB were contradicted by our results with strong evidence (**Figure 3.5D**). Overall, we were able to generate functional regulatory profiles for promoters driving expression of 1,158 of the 2,565[28] operons in E. coli as well as many other promoters that contribute to global expression levels. Thus, we conclude that this approach is an efficient method to rapidly characterize regulatory motifs within thousands of experimentally verified promoter regions.

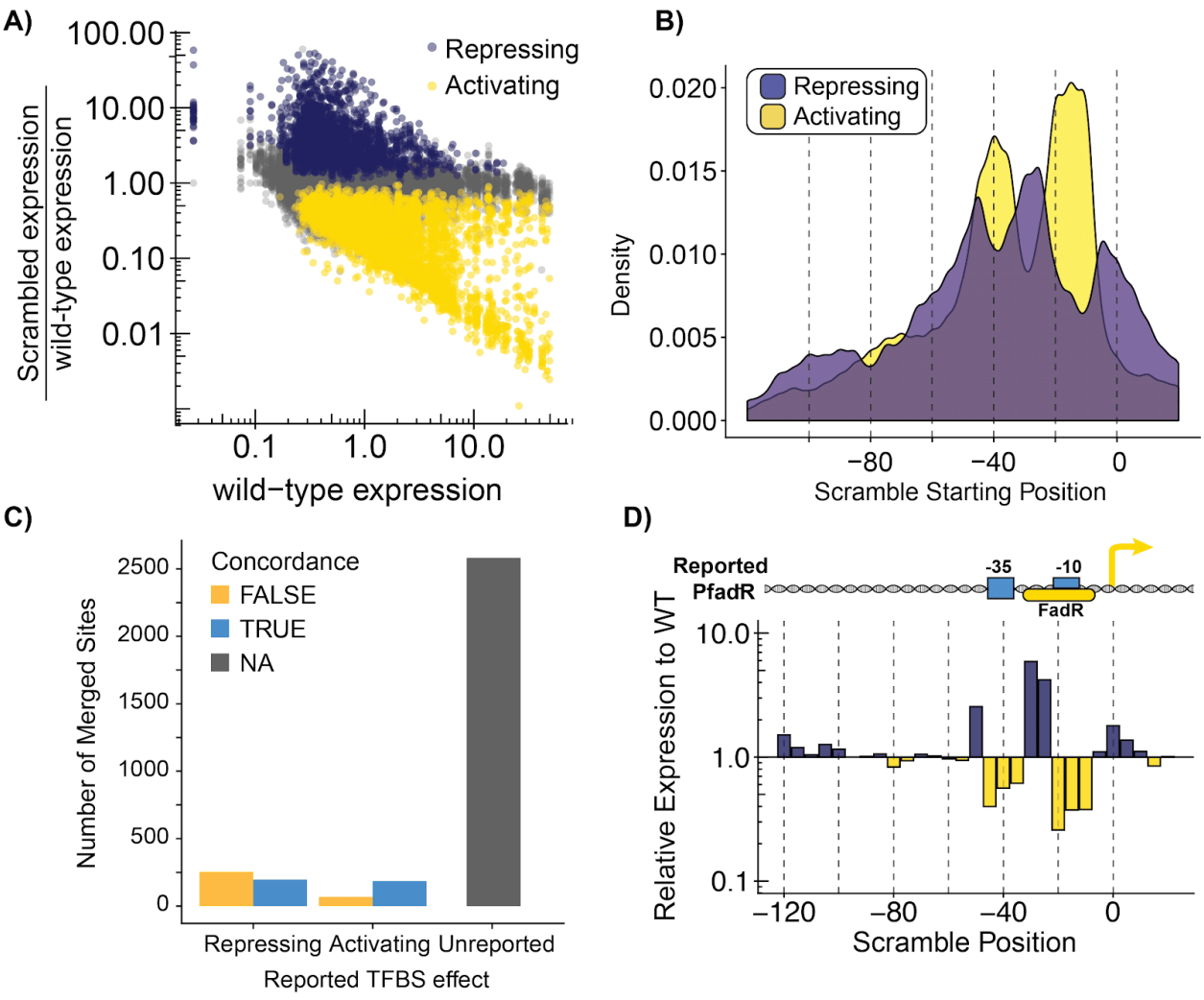


Figure 3.5. Global identification of *E. coli* regulatory motifs by scanning mutagenesis. We selected 2,057 active TSS-associated promoters identified in LB-rich media and systematically scrambled 10 bp sequences spanning the -120 to +30 positions at five bp intervals. **A)** We identified scrambles that significantly increase (N = 1,885) or decrease (N=5,408) expression relative to the unscrambled sequence based on the mean RNA/DNA ratio at the individual barcode level across four replicates (Student's t-test, two-sided, adjusted to 1% FDR). Data are colored whether the scrambled regulatory region activated or repressed transcription of the promoter. **B)** Distribution of the location of significant scrambles relative to TSS. **C)** Comparison of scramble effect to RegulonDB annotation. We compared each significant scramble to all annotated RegulonDB TFBSs that overlapped the genomic coordinates of the scramble. The scrambles are grouped by the effect of the overlapping TFBS as reported by RegulonDB, either repressors, activators, or unreported in the database. 77.8% (2,583/3,317) of significant scrambles are not annotated. **D)** Scanning mutagenesis of the *FadR* promoter (bottom) compared to reported architecture (top).

Promoter activity prediction remains a challenge

In this study we generated a powerful dataset of three distinct library designs, totaling 117,556 unique sequences, that provide a quantitative measure of promoter activity *in vivo*. Using this unique dataset, we sought to evaluate our ability to predict promoter activity from sequence alone, which would be invaluable for annotating promoter sequence space *de novo* as well as designing promoters with designed activities. We leveraged this information and trained several machine learning models of varying complexity, for both classification and regression. Many sequences have high overlap with other sequences, due to library design and close proximity of previously reported TSSs. We split the data into 75% for training (n = 87,164) and 25% (n = 30,392) for testing according to genomic location, ensuring the two sets contain sequences equidistant to the origin (see Methods). For classification, we further filtered the data and only considered expression < 0.75 as negatives and > 1.25 as positives (all datasets have activity threshold normalized to 1). There is some amount of noise in our assay leading us to not evaluate sequences close to the threshold, yielding a training set of 53,326 (12,918 positives and 40,408 negatives) and test set of 18,567 (4,414 positives and 14,153 negatives).

We trained several different classifiers to predict whether a given sequence was active or inactive (**Figure 3.6A**). All classifiers output the predicted probability for each class, rather than directly predicting the class, so they can be compared using precision-recall curves. Further details for all models are included in the methods. We trained a "baseline" logistic regression based on four "mechanistic" features known to be associated with promoter strength: max -10 sigma70 motif position weight matrix (PWM) score, max σ70 -35 motif PWM score, paired -10 and -35 PWM score (PWMs scanned jointly allowing for, 16, 17, or 18 gap), and percent GC content. We trained this model only for the TSS library, split by genome location as described above, because it directly represent endogenous genomic sequence. For comparison, we trained a gapped k-mer SVM (gkm-SVM) model with word-length 10 and 8 informative columns (L= 10, K = 8) on the

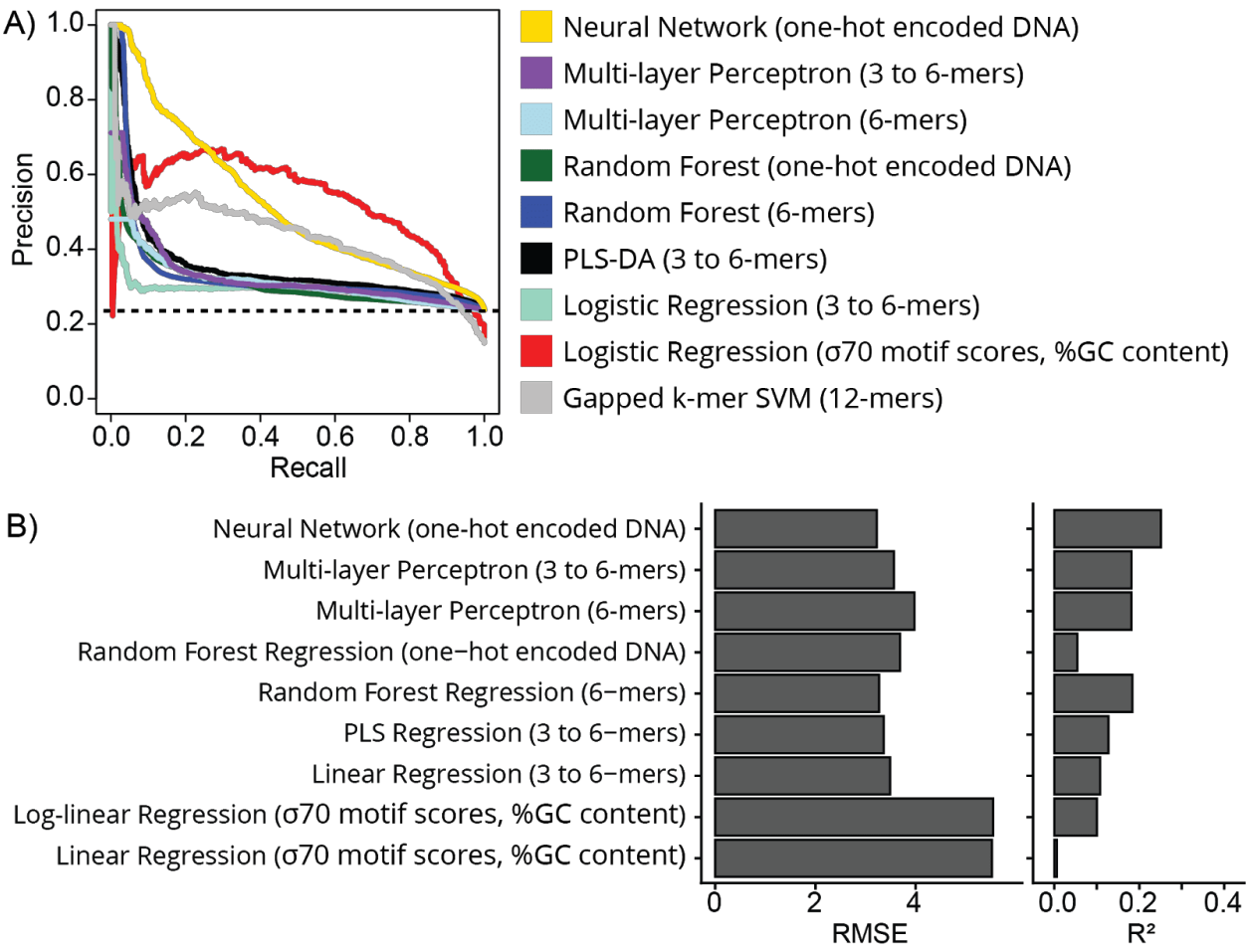


Figure 3.6. Various machine learning models for promoter activity classification and regression. We trained various models to predict promoter activity across all three library designs: endogenous TSS, peak tiling, and scramble. We split the datasets into 75% training (n = 87,164) and 25% testing (n = 30,392) based on their genomic position to prevent overfitting, as many sequences have overlapping sequence content. Classification and regression models were trained independently and details about each model and feature set can be found in Methods. **A)** We trained classification models to predict if a sequence was active or inactive. We considered all samples with activity < 0.75 to be inactive and activity > 1.25 to be active. We created this buffer around the activity threshold of 1 to shift focus away from potentially noisy observations around the threshold. We evaluated classification performance using precision-recall curves because our data is imbalanced. Convolutional neural networks performed best in the lower recall range, while logistic regression based on simple hand-crafted features performs better in the higher recall range. Dashed line represents the expected performance from random prediction. **B)** We trained regression models to predict a quantitative level of promoter activity. We evaluated performance using both root mean squared error (RMSE) and coefficient of determination (R^2) on the held-out test set. Similar to classification, convolutional neural networks performed the best with the lowest RMSE and highest R^2. Simple linear regression based on hand-crafted features performed the worst with the highest RMSE and lowest R^2.

same training set, as this model is best suited for sample sizes under 20,000, and observed decreased performance relative to the logistic regression (AUPRC = 0.43, AUPRC = 0.53, respectively). Furthermore, we created a feature set of all 3 to 6-mer frequencies and trained a logistic regression, partial least squares discriminant analysis (PLS-DA), and multi-layer perceptron (MLP). To observe the effects of reducing dimensionality, we additionally trained on only 6-mer frequencies for the MLP and random forest. For the simpler logistic regression and PLS-DA we performed an additional feature selection step based on the performance of a "random" k-mer (Methods). All of these models performed similarly, with AUPRC ranging from 0.29 to 0.35.

We were interested in a more complex model that could capture more intricate sequence features beyond the core promoter motifs without any *a priori* knowledge, such as RNAP binding motifs or TFBSs. There has been recent work which predicts transcriptional regulatory activity from MPRA data using convolutional neural networks (CNNs)[53]. Inspired by this work, we trained a CNN using the DragoNN toolkit which is built on top of the keras python package. We performed hyperparameter tuning for a three layer CNN and achieved AUPRC = 0.52. Next, we compared the CNN to other machine learning models that require less hyperparameter tuning, construct more interpretable models, and have a faster runtime. For comparison, we trained a random forest on one-hot encoded DNA, although this model is not well suited to categorical input features, and achieved AUPRC = 0.31. To overcome this limitation, we trained the random forest using frequencies of 6-mers and observed a slight increase in performance (AUPRC = 0.34). Overall, the CNN achieved the highest AUPRC, but the logistic regression fit with four features performs better at higher recall. The two models are not directly comparable because they are trained on different datasets, as the logistic regression fails to converge when run on the combined dataset.

We separately trained all of the models described above, with the exception of gkm-SVM, for the more difficult task of regression. Additionally, we included linear regression fit to the four "mechanistic" features, for log-transformed expression. We evaluated each model using root mean squared error (RMSE) and R-squared between predicted and observed values for promoter activity. Many models perform similarly to each other, with the CNN achieving the highest R-squared and lowest RMSE. We observe improvement in the linear regression on log-transformed data compared to linear regression without transformation, suggesting there are non-linear relationships that are presumably captured by more complex models. Random forest on one-hot encoded DNA performs worse than random forest on 6-mer frequencies, in line with the heuristic that random forests are not well suited to categorical features.

Overall, the CNN performs best in both classification and regression, although simpler models have some predictive power and have the benefit of much faster training times. However, all of these models together provide evidence that the problem of prediction, even for a simple model organism, remains challenging. We attempt here to highlight several different approaches to the problem with models of varying complexity. This is a starting point for future work for further optimization and utilizing other types of models.

DISCUSSION

The promoter as described by Jacob and Monod was a distinct unit of the operon with elegant but ultimately simple components. Looking back at over 50 years of research, the definition of a promoter has evolved dramatically to encompass their incredible flexibility and complex roles in genetic regulation. In this work, we use a multiplexed reporter assay to isolate promoter activity from extraneous mechanisms of genetic regulation and characterize hundreds of thousands of sequences spanning the *E. coli* genome. Our genomic screen identifies 3,321 and 3,477 active promoters in rich and minimal glucose media, respectively, and find many intragenic and

antisense promoters. We present evidence that antisense promoters have a profound impact on global transcript levels and shows the quantitative relationship between antisense promoter activity and repression. Our analysis of codon usage within intragenic promoter suggests that codon usage has adapted to enable promoter activity within genes while still navigating the constraints of protein coding space. Amongst these intragenic promoters, there is a significant enrichment of rare codons in these regions. Furthermore, we perform scanning mutagenesis of 2,057 previously identified TSS-associated promoters and identify 7,293 encoding regulation within these promoters, of which 2,583 have not been previously annotated.

A critical question to reflect upon is: Have we finally identified all promoters encoded in the *E. coli* genome? Considering the dramatic rearrangement of promoters between rich and minimal glucose media, it is likely that interrogating other conditions will reveal many other condition-dependent promoters. Furthermore, we are skeptical to make the claim that we've definitively identified all promoters even in the conditions tested in this work. Here we defined sequences as promoters based on empirically derived thresholds, however, this method is somewhat arbitrary and likely underestimates the actual number of promoters. Perhaps promoter sequences cannot be defined in such a binary manner and are actually a quantitative trait inherent to all sequences in the genome. This quantitative perspective is consistent with the incredible amounts of 'spurious' transcription observed by RNA-seq studies. Moreover, it is consistent with the observation that a surprising number of random sequences exhibit promoter activity[54]. To reaffirm this observation, we used our platform to characterize 1,000 random 150 bp sequences, with over 3% of sequences surpassing our empirically derived threshold **(Figure 3.S6),** although as many as 10% of random 100 bp sequences have been previously reported to be active promoters. On the other hand, there may be other mechanisms within *E. coli* that allow the genome to distinguish transcription from promoter sequences vs non-promoter sequences, similarly to what has been observed in the suppression of transcription from tss-

RNA promoters[35]. More sensitive techniques to study promoter activity of sequences near activity thresholds will be necessary to clearly state whether promoter activity is a general trait of all sequences in the genome or a binary identity that is encoded within specific sequences in genomes.

We designed a scanning mutagenesis library of previously identified 2,057 TSS-associated promoters and identified 7,293 regulatory sequences, including 2,583 that have not been previously reported[17]. We identified "scrambled" variants that significantly altered expression (compared to wild-type) and compared these motifs to known TFBS annotations in RegulonDB. The majority of our significant scrambles contained no corresponding annotation in RegulonDB, and for those that overlapped known sites, an appreciable proportion disagreed with the reported effect. There could be several explanations for this disagreement and the discovery of these missing annotations. First, it could be that the predictions of TFBSs in RegulonDB are actually false positives due to promiscuous binding events. Second, some transcription factors may possess condition-dependent behavior and the conditions tested in our study do not capture the full scope of their regulatory program. Finally, it is plausible that a portion of the sites we identify represent true functional sites that are missing from current annotation and should be interesting targets for further dissection, such as identifying which transcription factors operate at these motifs. Further studies to determine which sequences within a promoter contribute to regulation may aid efforts towards predicting promoter sequence-function relationships.

We leveraged our unique and large datasets of quantitative promoter activity across multiple libraries to train machine learning models to predict whether a promoter was active or inactive (classification) and the precise level of activity (regression). We implemented models of varying complexity, from simple linear regression models based on a handful of biological features, to convolutional neural networks trained on raw sequence alone. Even with the large training set,

the performance of these predictive models is limited. Why does this problem remain challenging? First, it is likely quite challenging to develop a single generalizable model for all promoters as there are several families of sigma factors with various consensus motifs. Therefore, models that are sigma-factor specific may be more tractable. Second, although the range of our MPRA is quite dynamic, accurate predictive models may require techniques with even greater quantitative resolution, especially in the noise regime of the assay where most observations fall. Finally, high performance models may require even larger and more narrowly focused training sets. For example, one could create a library design to parameterize the binding motifs for various sigma factors, allowing greater exploration of the vast sequence space than possible with the limited sites present in the genome. In previous work, we designed a "minimal" promoter with various combinations of the core σ70 promoter motifs embedded in a constant background and developed accurate predictive models using the identities of core motifs as features[29]. These types of approaches could be better suited for the prediction task because they use functional promoter sequences as a foundation to expand further into sequence space. There has been recent similar work from other groups[9], and our MPRA would be a powerful tool to further refine biophysical models.

As genome sequencing technologies continue to evolve, our ability to discover genomes has far surpassed our ability to characterize them. While recent technological advances have made it possible to rapidly determine gene functions within bacterial genomes[55], we lack similar genome-scale approaches to dissect the regulation of these genes. Understanding the regulation of genes can help place these functions within their appropriate context, which is paramount to understanding how they contribute to the behavior and plasticity of cells. This work marks a notable leap in our understanding of genetic regulation by promoters in *E. coli*. Although we only characterize the promoter landscape of *E. coli,* we believe this approach can be utilized for any genetically tractable bacteria amenable to the transformation of relatively large libraries of

sequences (~10^6 clones) and RNA-Seq of expressed barcodes. Similar approaches have been utilized in a number of bacterial species to identify functional promoter sequences[56]. Therefore, the approaches we present here will enable researchers to rapidly screen bacterial genomes for active promoter sequences and determine how promoter activity is encoded within genomes.

Acknowledgements

This work was supported by National Science Foundation Graduate Research Fellowship 2015210106 to G.U., National Institutes of Health New Innovator Award DP2GM114829 to S.K., Searle Scholars Program (to S.K.), U.S. Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and Linda and Fred Wudl. We thank the UCLA BSCRC high throughput sequencing core and Technology Center for Genomics and Bioinformatics for technical assistance; Robert B. Phillips, Reid C. Johnson, and Jeffery H. Miller for thoughtful feedback throughout this project; Matteo Pellegrini for computational advice; Christina P. Burghard for advice on bioinformatics analysis; and all past and present members of the Kosuri lab for technical feedback. We would also like to thank the invaluable resources of RegulonDB and EcoCyc as well as all contributors to these collections. Lastly, we thank the UCLA Molecular Biology Interdepartmental Graduate Program and UCLA Bioinformatics Interdepartmental Graduate Program

Author Contributions

G.U., K.D.I., H.K., and S.K. designed the study. G.U., A.D.T., and M.B. developed and performed experimental methods. N.B.L. developed genomic fragmentation isolation method. G.U., K.D.I., and A.D.T. analyzed, and interpreted data. T.C. developed k-mer based multilayer perceptron for promoter prediction. K.D.I. developed and implemented machine learning approaches for promoter prediction. G.U., K.D.I, and S.K. wrote the manuscript.

Declaration of Interests

The authors declare no competing interests

METHODS

Strains

All experiments were performed in the *E. coli* MG1655 background.

TSS library design

The TSS library incorporates all TSSs from the RegulonDB database[28] (Version 8.0, genome version U00096.2) and those identified in two recent genome-wide TSS mapping studies[15,16]. We synthesized each TSS embedded in its local sequence context -120 to +30 relative to the TSS, capturing most of the *cis*-regulatory elements. Recent work provides evidence that most regulatory motifs fall within 100 bp upstream of the TSS[30] and the initial transcribed region (+1 to +20) can also influence gene expression. There were 23,798 unique TSSs across all three sources, many of which were a few base pairs away from each other. We minimized redundancy and collapsed together TSSs within 20 bp and selected the most upstream TSS for our library, yielding 17,635 TSSs for the final synthesized library. Additionally, we included 500 negative controls from the *E. coli* genome that are assumed to have minimal regulatory activity. We randomly selected 150 bp sequences that are more than 200 bp from a TSS (on either strand), and many fall within coding regions. We included a set of 112 short synthetic positive controls that were previously characterized[57,58] and span a wide range of expression.

TSS library barcoding and cloning

The TSS library was synthesized by Twist Biosciences and delivered lyophilized as a 26 pmol pool. The library was resuspended in 100 uL of TE pH 8.0 and 1 uL was amplified for 12 cycles using GU72 and GU116 with NEB Q5 High-Fidelity 2x Master Mix (#M0492L). Unless otherwise stated, all amplifications were performed using this polymerase mixture. This product was then ran on a 2% TAE agarose gel and approximately 200 bp amplicons were extracted using a

Zymoclean Gel DNA Recovery Kit (#D4008). For barcoding, 1 ng of this eluate was amplified for 9 cycles using primers GU72 and GU73. Following cleaning using a Zymo Clean and Concentrator Kit (#D40140), the library was digested using NEB's SbfI-HF and XhoI.

The plasmid backbone, pLibacceptorV2 (Addgene #106250 was digested using SbfI-HF and SalI-HF with the addition of rSAP (NEB #M0371S). The digested library was ligated into pLibacceptorV2 using T7 DNA Ligase (NEB #M0318S), cloned into 5-alpha Electrocompetent *E. coli* (NEB #C2989K), and plated on LB + kanamycin (25 ug/mL) yielding approximately 2.3 million colonies estimated by plating concomitant dilution plates. After allowing for 24 hours of growth on plates, the library was scraped and resuspended in LB, and then 800 million cells (based on OD_{600}) were inoculated in 450 mL LB + kanamycin (25 ug/mL) overnight. Unless stated otherwise, all plasmids were isolated using a Qiagen Plasmid Plus Maxiprep Kit (#12963) and concentrated using a Promega Wizard SV Gel and PCR Clean-up System (#A9281).

In order to clone the RiboJ::sfGFP reporter construct, the library was digested using NEB's BsaI-HF and NheI-HF with the addition of rSAP. The reporter construct was digested using NEB's BsaI-HF and NcoI-HF. Similarly to the previous cloning step, the reporter was cloned into the library using T7 DNA Ligase, cloned into 5-alpha electrocompetent *E. coli*, and plated on LB + kanamycin (25 ug/mL), yielding 6.8 million colonies. The completed plasmid library was isolated as stated above.

Isolation of genomic fragment library

To isolate genomic fragments, 10 ug of *E. coli* MG1655 gDNA was sheared using a Covaris . The settings used were as follows: Duty factor was set to 10%, Intensity was set to 4, cycles/burst was set to 200, and time was 60 seconds. The sheared gDNA was ran on a 3% TAE agarose gel and fragments between 200 and 300 bp were extracted using a Zymoclean

Gel DNA Recovery Kit and eluted in 18 uL water. All 18 uL of the extracted fragments were end repaired using Enzymatics End Repair Mix (Part # Y9140-LC-L) following manufacturers protocols, cleaned using 45 uL (1.8x volume) of Agencourt AMPure XP Beads (#A63880), and eluted in 20 uL of water. The 20 uL eluate was A-tailed following the New England Biolabs protocol:

Reaction:

20 uL End-repaired DNA

5 uL NEB Buffer 2 (10x)\

0.5 uL dATP (10mM)

3 uL Klenow Fragment (3' -> 5' exo-) (Enzymatics #P7010-HC-L)

21.5 uL Nuclease-free water

The reaction was Incubated for 30 minutes at 37ºC, then heat inactivated for 20 minutes at 75ºC before cleaning using 90 uL Agencourt AMPure XP beads and eluting in 20 uL water. Y-adapters to facilitate fragment amplification and barcoding were ligated to the A-tailed fragments using the following reaction mix:

Reaction:

20 uL A-tailed DNA

5 uL NEB T4 DNA Ligase Buffer (10x) (NEB #B0202S)

2 uL Y-adapter GU Y-Frag (25 uM)

1 uL NEB T4 DNA Ligase (NEB #M0202T)

22 uL Nuclease-free water

This reaction was incubated for 20 minutes at 25ºC, heat inactivated for 20 minutes at 65ºC, and subsequently cleaned using 90 uL Agencourt AMPure XP beads and eluting in 12 uL nuclease-free water.

Barcoding and cloning of genomic fragment library

To barcode the genomic fragments, 1 uL of the processed fragments was amplified for 13 cycles using GU72 and GU116. This product was then cleaned using a Zymo Clean and Concentrator Kit and eluted in 12 uL nuclease-free water. For barcoding, 1 ng of this eluate was amplified for 10 cycles using primers GU72 and GU73. Following cleaning using a Zymo Clean and Concentrator Kit (#D40140), the library was digested using NEB's SbfI-HF and XhoI.

This library was cloned following the same protocols as the TSS library. The transformation of the barcoded library yielded approximately 3.3 million colonies and the transformation after addition of the RiboJ::sfGFP yielded approximately 1.25 million colonies.

Genomic promoter tiling library design

We used a custom peak caller on the single-nucleotide resolution strand-specific expression pileup generated from our genomic fragment library to define "peaks" of promoter activity. Our peak calling method is simple and conservative, as we wanted to tile the most active regions and keep the library size reasonable. We defined a peak as a continuous region with expression above an empirically determined threshold. We considered a continuous range of thresholds and for each evaluated the percentage of active TSSs, from our previous library, contained in a peak and determined an expression level of 1.1 was sufficient and captured 90% of active TSSs (data not shown). We required that each peak be at least 60 bp, and merged adjacent peaks that were within 40 bp, yielding 1753 and 1724 peaks for the minus and plus strands, respectively. We tiled each peak by synthesizing 150 bp windows across the region, with no

overlap between adjacent tiles, yielding 48,491 peak tiles. Additionally, we included 1000 randomly generated 150 bp sequences to test what fraction of random sequence can drive expression. We included the same set of positive and negative controls as described in the TSS library design.

Genomic promoter tiling library barcoding and cloning

The active TSS mutagenesis library was synthesized by Agilent and delivered lyophilized as a 10 pmol pool. The library was resuspended in 100 uL of TE pH 8.0 and 1 uL was amplified for 10 cycles using GU120 and GU121. This product was then cleaned using a Zymo Clean and Concentrator Kit and eluted in 12 uL nuclease-free water. For barcoding, 1 ng of this eluate was amplified for 8 cycles using primers GU120 and GU122. Following cleaning using a Zymo Clean and Concentrator Kit (#D40140), the library was digested using NEB's SbfI-HF and XhoI.

This library was cloned following the same protocols as the TSS library. The transformation of the barcoded library yielded approximately 1.5 million colonies and the transformation after addition of the RiboJ::sfGFP yielded approximately 5.2 million colonies.

Active TSS mutagenesis design

We systematically mutagenized all active TSSs from our initial TSS library to design a follow-up library. We used 500 negative controls to classify the TSS library into active and inactive TSSs. We set the active threshold at two standard deviations above the median expression for the negative controls, resulting in 2,670 active TSSs. We mutagenized the active sequence by scrambling 10 bp windows, sliding across the 150 bp at 5 bp intervals, resulting in 5 bp of overlap between adjacent scrambles. We scrambled the sequence using the existing 10 bp to preserve nucleotide content and selected the scramble that was most dissimilar to the original sequence out of 100 scrambling attempts. Our final library included 59,653 scrambled

sequences and 2,057 unscrambled sequences. We also included the same set of negative and positive controls as described above for the TSS library, for a total library size of 62,322.

Active TSS mutagenesis library barcoding

The active TSS mutagenesis library was synthesized by Agilent and delivered lyophilized as a 10 pmol pool. The library was resuspended in 100 uL of TE pH 8.0 and 1 uL was amplified for 12 cycles using GU123 and GU124. This product was then cleaned using a Zymo Clean and Concentrator Kit and eluted in 12 uL nuclease-free water. For barcoding, 1 ng of this eluate was amplified for 10 cycles using primers GU123 and GU125. Following cleaning using a Zymo Clean and Concentrator Kit (#D40140), the library was digested using NEB's SbfI-HF and XhoI.

This library was cloned following the same protocols as the TSS library. The transformation of the barcoded library yielded approximately 3.7 million colonies and the transformation after addition of the RiboJ::sfGFP yielded approximately 5.2 million colonies.

Library Barcode mapping

We used PCR to individually barcode each library sequence to quantitatively measure expression in our MPRA. Prior to genome integration, DNA-sequencing was performed to computationally map barcodes to sequences. A custom barcode mapper developed by Nathan Lubock (manuscript in preparation) was used to collapse reads into a barcode-sequence map. We used two filtering steps for barcode quality. First, we required a minimum number of reads for every barcode, assuming reads that appear once or twice correspond to sequencing errors. Second, BBMap[59] was used to align the reads associated with a given barcode, and discarded barcodes that map to sequences that are too dissimilar to one another. A Levenshtein distance of 30 was used to discard barcodes that map to two very distinct sequences, while still allowing for a small number of sequence errors.

Library integration into specific genomic loci

Library integration was performed as previously described.

The isolated plasmid library was digested with SalI-HF and NheI-HF to eliminate incompletely cloned plasmid before transformation into electrocompetent MG1655 with a landing pad engineered in the nth-ydgR locus and plating on LB + kanamycin (25 ug/mL). Colonies were resuspended in LB and 800 million cells were inoculated into
250 mL LB + kanamycin (25 ug/mL) and grown overnight. Several 2 mL frozen aliquots were made of this overnight culture.

The library was integrated into the nth-ydgR locus as follows. A frozen aliquot of MG1655 with a landing pad engineered in the reverse orientation at the nth:ydgR locus was transformed with the library and grown overnight in 200 mL LB + kanamycin (25 ug/mL). Following overnight growth, 400 million cells of this culture were seeded into 250 mL LB + kanamycin (25 ug/mL) + .2% arabinose (g/mL) and grown for 24 hours. After integration of the library, the plasmid backbone was removed through heat-curing. From the 24 hour induced culture, 800 million cells were inoculated into 80 mL of LB + kanamycin (25 ug/mL) and grown at 42 ºC for approximately 1.5 hours before reaching an OD 600 =.3. Upon reaching exponential growth, 200 million cells from this culture library were plated and grown for 16 hours at 42 ºC. Heat-cured plates were scraped and resuspended in LB and 400 million cells were inoculated into 200 mL LB + kanamycin (25 ug/mL). This culture, consisting of our integrated and heat-cured library, was grown overnight at 37 ºC and several frozen 2 mL aliquots were made.

To test the TSS library in the *essQ-cspB* and *ybbD-ylbG* midreplichore regions, the same protocol was followed using strains engineered with landing pads in these intergenic regions.

Library growth and harvest for expression measurements

To measure expression of all promoter libraries, libraries were grown and harvested as previously described[29] with minor changes to culture conditions.

For each library and biological replicates, a 2 mL frozen aliquot of the library was inoculated in 200 mL LB (Source) with 25 ug/mL of kanamycin and grown at 30 ºC overnight. The overnight cultures were used to seed new cultures at OD_{600} = .0005 and grown for approximately 5.5 hours at 30 ºC until reaching an OD_{600} between = 0.5 and 0.55. The genomic fragment library was also grown in Minimal Media (Source) with .2% glucose (g/mL) and 25 ug/mL of kanamycin for 10 hours at 30 ºC until reaching an OD_{600} between = 0.5 and 0.55. Cultures were rapidly cooled to 0 ºC in an ice slurry for two minutes. Three 50 mL aliquots were pelleted at 4 ºC by centrifugation at 13,000xg for two minutes and the supernatants were poured out before snap-freezing the pellets in liquid nitrogen. Three 5 mL aliquots of each library were harvested using the same approach to be processed for genomic DNA extractions.

RNA and DNA library preparation

RNA was extracted from 50 mL library pellets using a Qiagen RNEasy Midi kit (#75142) and 45 ug of each extract was concentrated using a Qiagen Minelute Cleanup Kit (#74204). Barcoded cDNA was generated from 25 ug of each concentrated RNA extract using Thermo Fisher SuperScript IV (#18090010) primed with GU101. The manufacturer's protocol was followed aside from extending the reaction time to 1 hour at 52 ºC. The cDNA reaction was cleaned using a Zymo Research DNA Clean and Concentrator kit (#D40140) before amplification. Barcoded cDNA was amplified via PCR for 13 cycles using primers GU59 and GU102. This reaction was cleaned using a Zymo Research DNA Clean and Concentrator Kit and 1 uL of this reaction was used in a second PCR for indexing and addition of flow cell adapters. The second

PCR was for 8 cycles and utilized primers GU102 with either GU61, GU62, GU63, or GU64 (which add separate 6 bp indices).

gDNA was extracted from 5 mL cell library pellets using a Qiagen Gentra Puregene kit (#158567). Barcoded DNA was amplified from 1 ug of gDNA via PCR for 12-15 cycles using primers GU59 and GU60. The reaction was subsequently cleaned using a Zymo Research DNA Clean and Concentrator kit. To add sequencing adapters and indices to the library, 1 ng of this reaction was subject to a second PCR for 8 cycles using primers GU70 with either GU63, GU64, GU65, or GU66 (which add separate 6 bp indices). RNA and DNA sequencing libraries were cleaned using a Zymo Research Clean and Concentrator Kit (#D40140) before quantification using an Agilent Tapestation.

For each library, eight separate sequencing libraries were prepared: Four sequencing libraries for each RNA/DNA with two biological replicates and two technical replicates of each biological replicate. Biological replicates originated from separately grown and harvested glycerol stocks of each library. For each biological replicate, two RNA/gDNA extractions and sequencing library preparations (technical replicates) were performed in parallel. Libraries were submitted to the Broad Stem Cell Research Center at UCLA for sequencing on a HiSeq2500 or to the UCLA Translational Pathology Core Laboratory for sequencing on a NextSeq500. Raw sequencing data and promoter expression measurements have been made available on NCBI's Gene Expression Omnibus (GEO Accession no.******).

RNA-Seq of MG1655 in M9 minimal Media

To compare the promoter landscape to local transcriptional levels, RNA-Seq was performed on MG1655 grown in M9 minimal media (BD Difco #248510) supplemented with 0.2% glucose, 2 mM magnesium sulfate, and 0.1 mM calcium chloride. Cells growth and RNA preps were

prepared as previously described (see methods section titled: library growth and harvest for expression measurements). Sample replicates, originating from the same culture, were prepared using an Illumina TruSeq® Stranded mRNA Library Prep (#20020594) following manufacturers protocols to achieve strand-specific coverage. We note that no rRNA depletion was performed to preserve the fully intact transcriptional landscape. Samples were submitted to the UCLA TCGB sequencing core and sequenced on a Hiseq 4000.

Universal Promoter Expression Quantification and Activity Thresholding

We processed all libraries using the same pipeline to facilitate comparisons between libraries and set a consistent activity threshold for each library. First, we use a set of 112 short synthetic positive controls, designed to span a range of activity[57,58], to fit a linear regression with the TSS library as the reference. Each library is compared independently to the TSS library using the set of positive controls present in both libraries. Next, we determined an activity threshold for each library independently based on the distribution of 500 negative controls, 150bp of random genomic sequence at least 200bp away from an annotated TSS (on either strand). We set the threshold at two standard deviations greater than the median negative control. Next, we independently scale each library so the threshold is equal to 1 to facilitate cross-library comparison and modeling. These steps standardize our data so we can train jointly across all datasets.

-10 Motif and -35 Motif characterization

A position weight matrix from bTSSfinder was used to identify and score the best match to the -10 and -35 motifs within active tss-associated promoters, inactive tss-associated promoters, and a set of 500 negative controls. Best scores were reported regardless of position within the sequence. For all pairwise comparisons of active tss-associated promoters, inactive tss-

associated promoters, and the negative controls, the distributions of motif scores were compared and a student's t-test was performed to determine significance.

Genomic fragment alignment and promoter landscape quantification

To identify the coordinates of genomic fragments assayed using the MPRA, fragment sequences were aligned using bowtie2[60] (version 2.3.4.3). To determine nucleotide-resolution calculations for promoter activity, we utilize the script, `frag_expression_pileup.py`. This script outputs WIG files in a strand-specific manner with the number of fragments overlapping each nucleotide position.

Comparison of condition-dependent promoters between rich and minimal media

To identify condition specific promoters, genomic peaks associated with promoter activity that contained no overlaps between conditions were identified. Coordinates of promoter peaks were cross-compared between conditions using the bedtools intersect tool (bedtools v2.27.1) and considered unique to a particular condition if they had no overlap between conditions.

Identification of condition-dependent TFBSs

The TFBS content of promoter peaks unique to each condition was evaluated by cross-referencing with TFBSs reported by (Salgado et al. 2013) (Release 8.8). Unique promoter peaks were assigned TFBSs based on overlapping genomic coordinates using the bedtools intersect tool (bedtools v2.27.1) with default parameters and ignoring strand assignments. Incidents of each TFBS overlap were quantified between conditions and hit frequencies were normalized to incidencies per 100,000 bp of promoter peak sequence.

Determining promoter-gene associations

To assign genomic promoter peaks to their regulated genes, peaks were first assigned specific nucleotide positions by identifying the maximum activity score within a peak. Promoter peaks were considered intragenic if their maximum scoring nucleotide overlapped with a gene coordinate. For peaks whose maximum scoring nucleotides were within intergenic regions, regulated genes were assigned by identifying the first downstream gene within 500 bp. Once gene associations were identified, promoter peaks were labeled sense or antisense depending on whether the regulated gene shared strand orientation with the promoter peak

RNA-Seq alignment and genome transcript coverage

RNA-Seq analysis was performed using the script *RNA-Seq_M9_processing.sh*. This script trims reads using the trimmomatic software (ver. 0.36+dfsg-3) and aligned to the MG1655 reference genome (U00096.2) using Hisat2 (Kim et al. 2015) (ver. 2.1.0-1). Genome nucleotide-resolution coverage was determined using Samtools depth (ver. 1.7-1) and overall gene expression levels were calculated using bedtools multicov (2.26.0+dfsg-5). In all cases, default parameters were used with the exception of allowing for strand-specific quantifications.

Identification of minimal promoter regions

To identify minimal sequences necessary for promoter activity, contiguous stretches of active promoter peak tiles were grouped and the minimal shared overlapping region was identified. Peak tiles above the expression threshold were identified and grouped together if they shared an overlap of at least 110 bp of their 150 bp total length. The minimal region necessary for promoter activity was found by determining the overlap of the outermost sequences within a contiguous stretch of active peak tiles.

Amino acid and codon bias within intragenic promoters

Amino and codon usage was characterized within intragenic promoters and compared to all *E. coli* coding regions. To identify intragenic promoters, minimal regions necessary for promoter activity were identified by cross referencing genomic coordinates to reported genes. Reported gene coordinates were acquired from (Salgado et al. 2013) (Version 8.0). Once intragenic promoters were identified, nucleotide triplets were extracted while conserving the reading frame of the overlapping gene. Similarly, nucleotide triplets were extracted from all reported *E. coli* coding regions after filtering out sequences which did not have lengths of a multiple of three. For these extracted sequences, codon frequencies were normalized to the total number of occurrences of the encoded amino acid. Amino acid frequencies were normalized to the total number of amino acids within each group of sequences. Significantly enriched or depleted codons were identified by performing a chi-squared test within each amino acid group and adjusting the p-value using FDR. Significantly enriched or depleted codons were identified by performing a chi-squared test for each amino acid relative to the total pool of amino acids and adjusting the p-value using FDR.

Identification of statistically significant scrambling promoter variants

We identified scrambling promoter variants that significantly altered expression compared to the wild-type (WT) variant in the script scramble_ttest.Rmd. We considered each scramble and barcode combination as an independent observation, rather than summarizing expression as an average across all barcodes. A two sample two-sided Student's t-test (t.test) was performed to test for a significant difference in mean expression levels between barcodes for a scrambled variant and barcodes for the corresponding WT variant. We performed multiple testing correction and identified 1,885 scrambles that increase expression and 5,408 that decrease expression relative to the WT variant, at a false discovery rate of 1%.

Next, bedtools merge was used to merge overlapping adjacent scramble variants to produce "merged" scrambles. These merged sites correspond to a continuous scrambled region that induced significant changes in expression. We identified 1,414 merged scrambles that increased expression and 1,903 merged scrambles that decreased expression, and scrambles were merged separately based on effect.

Comparison of identified regulatory regions to RegulonDB annotations

We compared our identified merged scramble sites to existing RegulonDB annotations. We used bedtools intersect and required that 10% of the TFBS overlapped with a merged scramble site to count as an overlap. Next, we assessed whether the expression effect seen in our MPRA agreed with the direction of effect of the TFBS as indicated in RegulonDB. A merged scramble site was marked as "concordant" if any of the component scrambles agreed with existing annotation, and not concordant otherwise.

Machine learning models

We implemented several machine learning models, independently trained for both classification and regression. All reproducible code is provided in the Github and we will briefly describe each model and the appropriate parameters or implementation details.

Data processing

We standardize all datasets as detailed above in "Universal Promoter Expression Quantification and Activity Thresholding". Next, we split our data, using custom scripts, into 75%/25% for training/testing based on genomic location, ensuring the splits are equidistant from the origin, to avoid overfitting (define_genome_splits.py). Briefly, we split the genome into eight "chunks", with the first and last chunk adjacent to the origin of replication. We designated the second and

seventh chunk as the test set and remaining chunks as training set. This splitting maintains roughly the same distance from the origin between the training and test sets to avoid any potential effects of genome location. Many of our library designs include high overlap between adjacent positions in the genome. Splitting by genome location mitigates inflated performance due to highly similar sequences present in both train and test sets. Across the three libraries (TSS, peak tiling, scramble) there are 87,164 training samples and 30,392 test samples.

We trained models for both regression and classification. Our data is skewed toward negative examples, with many samples near our determine threshold. For classification, we created a buffer around the threshold and only include sequences with expression <= 0.75 as negatives and >= 1.25 as positives and labeled sequences as active or inactive. Our training set is reduced to 53,326 samples and testing set to 18,567 samples.

We specify classification models to predict probabilities, instead of the class, for precision-recall curves.

Simple model with promoter features

For the models in this section we created features only for the TSS library because it is closest to endogenous sequence and is a smaller dataset. The training and test sets were split by genomic location, as described above, with 13,118 training samples and 4549 testing samples.

We created a simple model which incorporates four features related to promoter function. We calculated the maximum position weight matrix (PWM) score using motifs from bTSSfinder[61] for both the -10 and -35 core promoter motifs. We scanned the -10 and -35 PWM individually and took the max score at any position using scoring functions from the Bioconductor package Biostrings[62]. Next, we scanned the sequence with -10 and -35 PWM jointly, allowing either 16,

17, or 18bp spacing in between the PWMs, reflecting common spacer lengths between core motifs. We assigned the "paired" max score as the max score at any position in the sequence across the three length options. Finally, we calculated the GC content (percentage) as this has been shown to be negatively correlated with promoter strength[56]. We constructed models in R with these four features and fit 1) a linear regression (lm), 2) a linear regression on the log-transformed expression values (lm) , and 3) a logistic regression (glm, family = 'binomial', type = 'response').

We trained the gapped k-mer SVM (gkm-SVM[63]) model on only the TSS dataset because the model is suited for training sets < 20,000. The training and test sets were split by genome position as described above. We specified a word length = 10 with 8 informative columns (L = 10, K = 8).

K-mer frequencies and simple models (linear regression, logistic regression, partial least squares regression, partial least squares discriminant analysis)

All of the models described in the remaining sections were trained using all three combined datasets, as described above.

We created a feature set based on k-mer frequencies, with k-mers ranging in length from 3 to 6-mers. We generated feature sets and trained models in python. For simpler models we performed an additional feature selection step using custom scripts (kmer_feature_generator.py).

We trained four models:

- linear regression (statsmodel.api.OLS)
- logistic regression (sklearn.linear_model.LogisticRegression())

- partial least squares regression (sklearn.cross_decomposition.PLSRegression())
- partial least squares discriminant analysis (sklearn.cross_decomposition.PLSRegression() on binary dependent variable)

For each k-mer, we computed the frequency in a set of random genomic sequences, the same length and size of the training set. We include a k-mer if the absolute correlation with expression is greater than the “random” k-mer frequency, resulting in 4800/5440 filtered k-mers. We chose partial least squares regression because it projects the input features onto a new space and is better equipped to handle a large number of features with high collinearity.

Random forest regression and classification

Next, we trained a random forest, for both regression (sklearn.ensemble.RandomForestRegressor()) and classification (sklearn.ensemble.RandomForestClassifier()). We train on one-hot encoded DNA as a comparison to the neural network model, although random forest is not well suited to categorical input features. To compensate for this, we trained the random forest using frequencies of all 6-mers and observed improved performance.

Multi-layer perceptron and neural networks

We trained a multi-layer perceptron for both regression (sklearn.neural_network.MLPRegressor()) and classification (sklearn.neural_network.MLPClassifier()). MLPs are a class of feedforward artificial networks and are “vanilla” neural networks consisting of an input layer, hidden layer, and output layer. We used two different feature sets: frequency of all 3- to 6-mers and frequency of only 6-mers. Feature sets were standardized with sklearn.preprocessing.StandardScaler() to remove mean and scale to unit variance. We trained all four models with the following parameters: alpha =

0.005, hidden_layer_sizes=(800, 30), solver = ‘lbfgs’, random_state=1, max_iter=10000, early_stopping=True, learning_rate=’adaptive’, tol=1e-8.

We trained a convolutional neural network (CNN) on one-hot encoded DNA sequence for both regression and classification. We performed hyperparameter tuning and training using [53], a toolkit for working with CNNs built on keras. We performed a random hyperparameter search for a three layer CNN for 100 combinations and the optimal parameters are listed below.

Regression:

- Dropout: 0.1340735187802852
- Pooling width: 16
- Convolutional filter width (for each layer): 16, 17, 18
- Number of filters (for each layer): 19, 39, 54

Classification:

- Dropout: 0.45541334972592196
- Pooling width: 7
- Convolutional filter width (for each layer): 8, 29, 29
- Number of filters (for each layer): 99, 87, 60

SUPPLEMENTAL INFORMATION

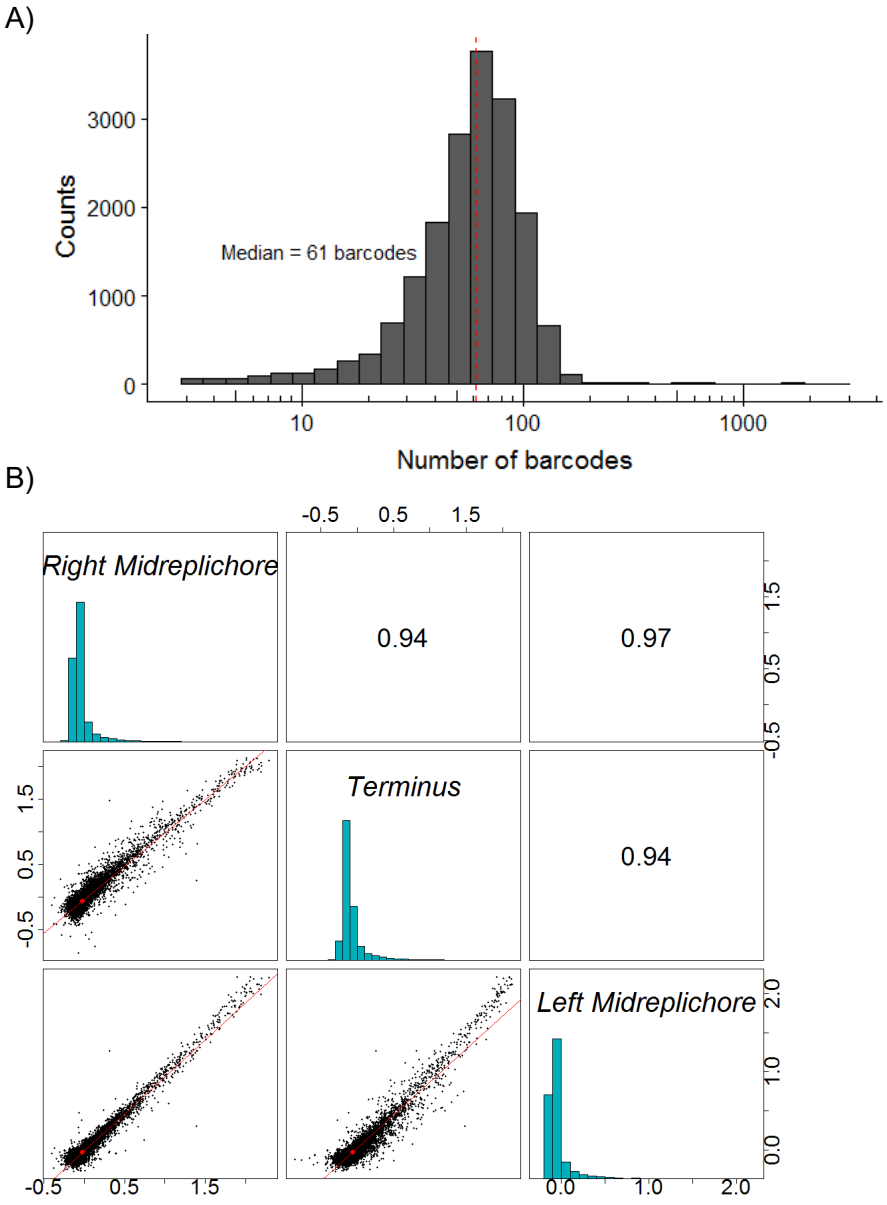


Figure 3.S1, related to Figure 3.1) TSS-associated promoters are represented by multiple barcodes and provide replicable measurements between genomic positions. A) Distribution of the number of barcodes measured per TSS-associated promoter (Median = 61 barcodes). **B)** Comparison of TSS-associated promoter measurements when integrated into distant regions of the *E. coli* chromosome.

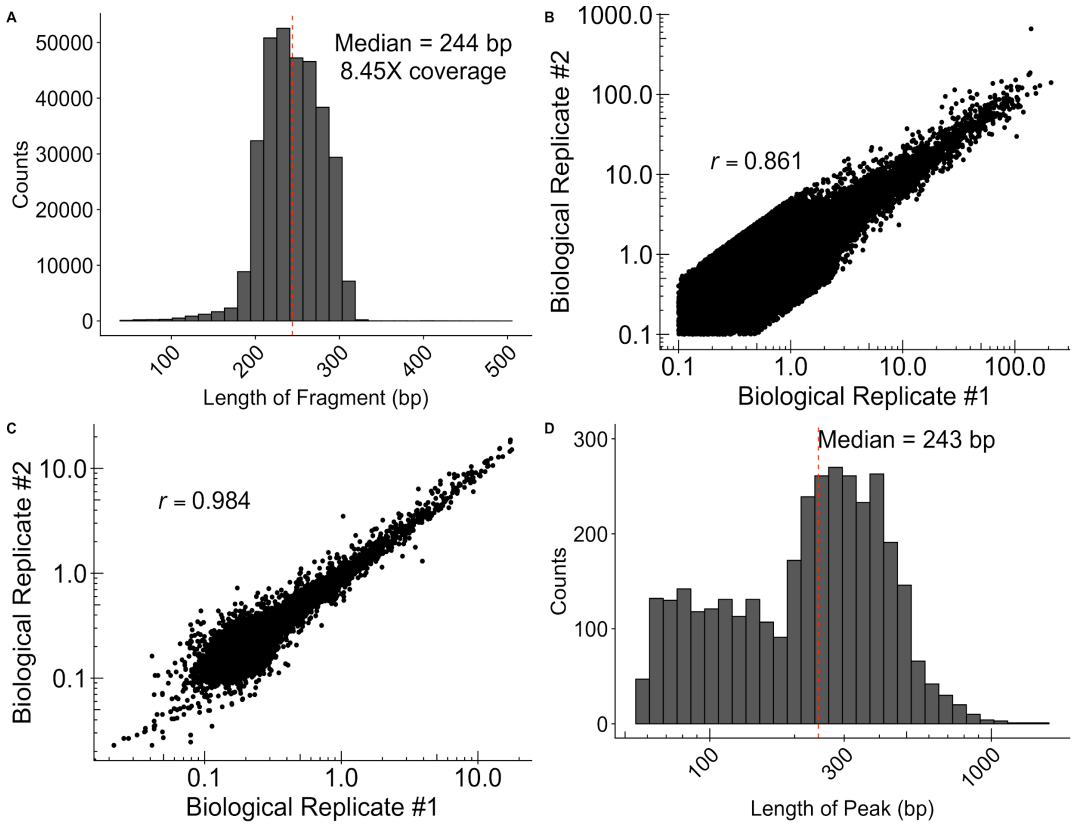


Figure 3.S2, related to Figure 3.2) TSS-associated promoters are represented by multiple barcodes and provide replicable measurements between genomic positions. A) Distribution of the lengths of genomic fragments assayed for promoter activity. **B)** Comparison of fragment expression measurements between biological replicates **C)** Comparison of 50,000 randomly sampled single-nucleotide measurements between biological replicates. **D)** Distribution of the lengths of genomic regions exhibiting significant promoter activity.

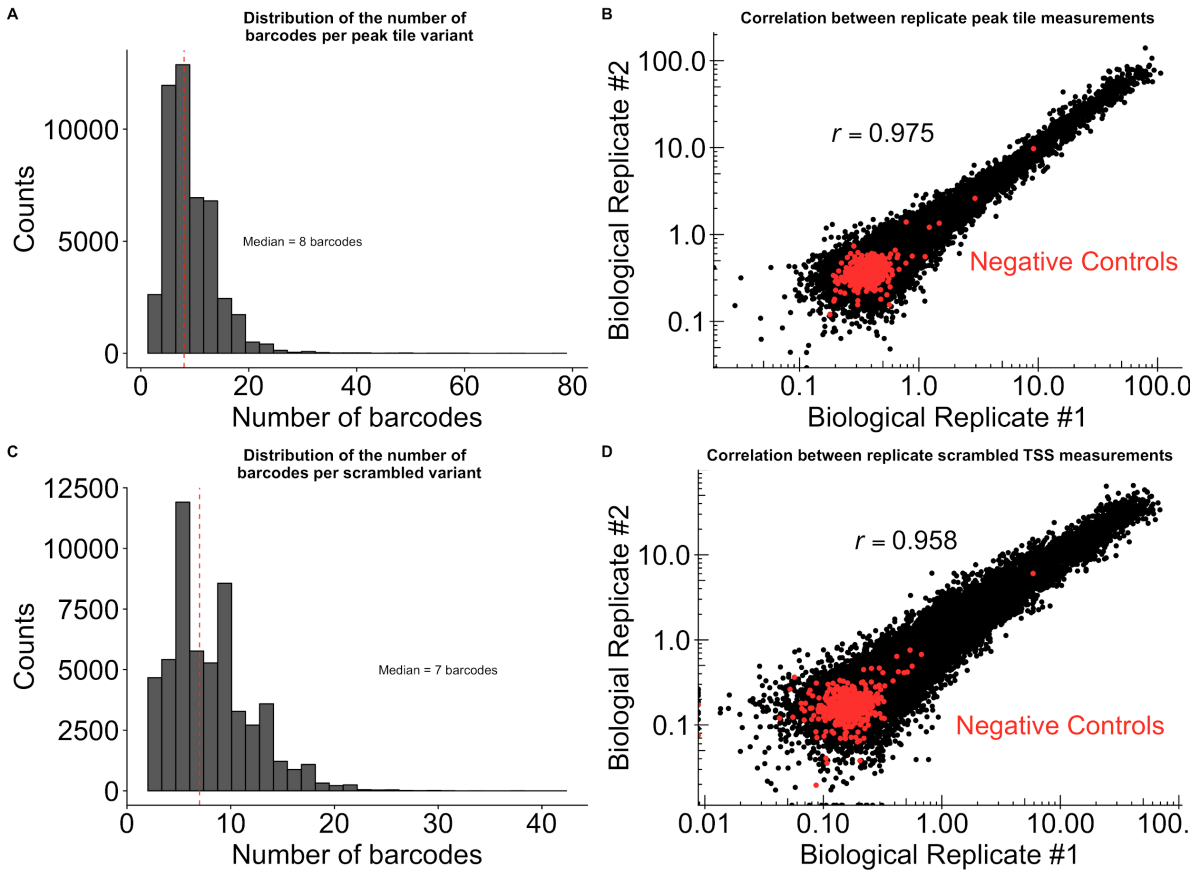


Figure 3.S3, related to Figures 3.3 and 3.4) Quality control for peak tiling and scrambled TSS libraries. A) Distribution of the number of barcodes per variant within the peak tiling library **B)** Comparison of peak tiling variant measurements between biological replicates. **C)** Distribution of the number of barcodes per variant within the scrambled TSS promoter library. **D)** Comparison of scrambled TSS variant measurements between biological replicates.

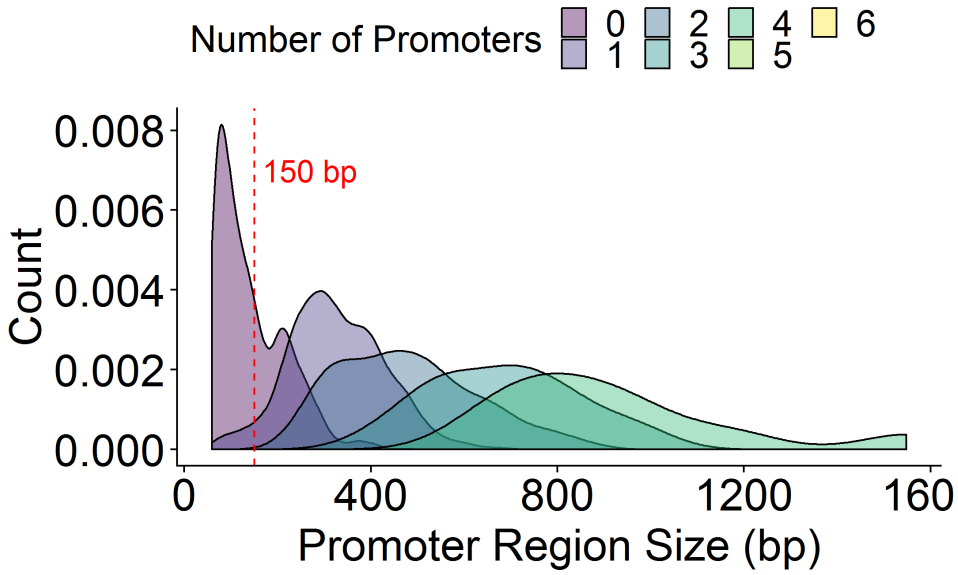


Figure S4, related to Figure 3) A) Distribution of the size of promoter regions identified from the genomic fragment screen separated by their number of distinct minimal promoters.

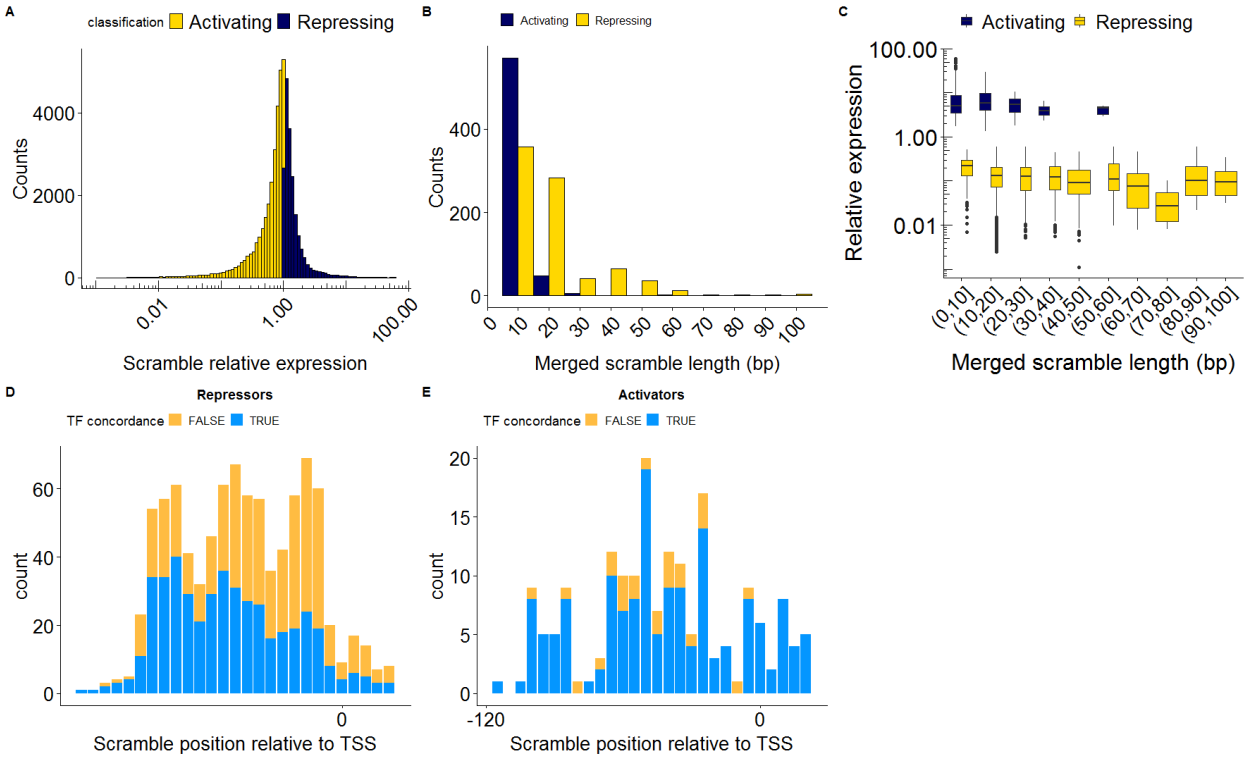


Figure S5, related to Figure 5) **Global identification of *E. coli* regulatory motifs by scanning mutagenesis. A)** Distribution of the effects of scrambling mutations on regulatory regions. **B)** Distribution of significant scramble lengths after merging contiguous regions. **C)** Relative change in expression from merged scrambles by length. **D,E)** Agreement between RegulonDB TFBS annotations and effects of scrambling the overlapping region of the promoter for **D)** Repressors and **E)** Activators

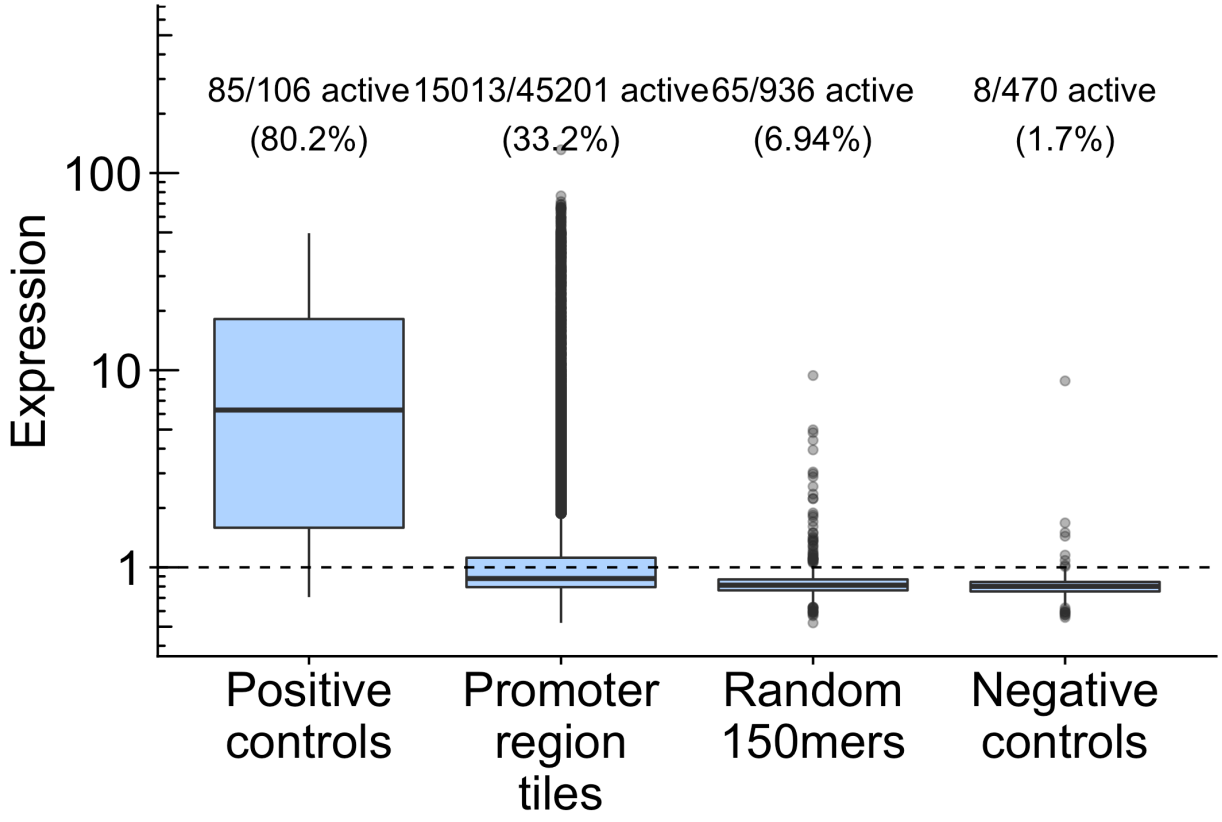


Figure 3.S6) An appreciable number of random 150mer oligos encode promoter activity.

Table 3.S1. Primers used in this study

Primer	Sequence (5' → 3')
GU59	CATGTTGTCCACTCCAATCGGTGATGGTCCTG
GU60	GTAATAGCTAAATCCCACCCGATGCCTGCAGG
GU61	CAAGCAGAAGACGGCATACGAGAT ACTGTG CATGTTGTCCACTCCAATCG
GU62	CAAGCAGAAGACGGCATACGAGAT AGCCAT CATGTTGTCCACTCCAATCG
GU63	CAAGCAGAAGACGGCATACGAGAT ATCTCG CATGTTGTCCACTCCAATCG
GU64	CAAGCAGAAGACGGCATACGAGAT CAGTGT CATGTTGTCCACTCCAATCG
GU70	AATGATACGGCGACCACCGAGATCTACACGTAATAGCTAAATCCCACCCGA TGC
GU72	ACCTGTAATTCCAAGCGTCTCGAG
GU73	TCGTATCCCTGCAGGNNNNNNNNNNNNNNNNNNNGCATGTGAGACCGGATG CTAACTAAACACCGCTAGC
GU79	CGTGCATAGTGCCATGTTATCCCTGAAGTCGAG
GU82	CAAGCAGAAGACGGCATACGAGATATCTCGCGTGCATAGTGCCATGTTATC
GU83	CAAGCAGAAGACGGCATACGAGATAGCCATCGTGCATAGTGCCATGTTATC
GU101	AATGATACGGCGACCACCGAGATCTACACGTAATAGCTAAATCCCACC CGATGCCTGCGG
GU102	AATGATACGGCGACCACCGAGATCTACAC
GU116	GGATGCTAACTAAACACCGCTAGC

REFERENCES

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
2. Johnson, X. B. & Hinton, D. M. Escherichia coli RNA polymerase recognition of a σ70-dependent promoter requiring a- 35 DNA element and an extended- 10 TGn motif. *J. Bacteriol.* **188**, 8352–8359 (2006).
3. Hook-Barnard, I. G. & Hinton, D. M. Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. *Gene Regul. Syst. Bio.* **1**, 275–293 (2007).
4. Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O. & Darst, S. A. Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* **296**, 1285–1290 (2002).
5. Liu, B., Hong, C., Huang, R. K., Yu, Z. & Steitz, T. A. Structural basis of bacterial transcription activation. *Science* **358**, 947–951 (2017).
6. Lee, D. J., Minchin, S. D. & Busby, S. J. W. Activating transcription in bacteria. *Annu. Rev. Microbiol.* **66**, 125–152 (2012).
7. Newberry, K. J. & Brennan, R. G. The structural mechanism for transcription activation by MerR family member multidrug transporter activation, N terminus. *J. Biol. Chem.* **279**, 20356–20362 (2004).
8. Lawson, C. L. *et al.* Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.* **14**, 10–20 (2004).
9. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* **107**, 9158–9163 (2010).
10. Ishihama, A., Shimada, T. & Yamazaki, Y. Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res.* **44**, 2058–

2074 (2016).

11. Peano, C. *et al.* Characterization of the Escherichia coli σ(S) core regulon by Chromatin Immunoprecipitation-sequencing (ChIP-seq) analysis. *Sci. Rep.* **5**, 10469 (2015).
12. Bonocora, R. P., Smith, C., Lapierre, P. & Wade, J. T. Genome-Scale Mapping of Escherichia coli σ54 Reveals Widespread, Conserved Intragenic Binding. *PLoS Genet.* **11**, e1005552 (2015).
13. Huerta, A. M. & Collado-Vides, J. Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.* **333**, 261–278 (2003).
14. Rhodius, V. A., Mutalik, V. K. & Gross, C. A. Predicting the strength of UP-elements and full-length E. coli σE promoters. *Nucleic Acids Res.* **40**, 2907–2924 (2012).
15. Conway, T. *et al.* Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. *mBio* **5**, (2014).
16. Thomason, M. K. *et al.* Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *J. Bacteriol.* **197**, 18–28 (2015).
17. Gama-Castro, S. *et al.* RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.* **36**, D120–D124 (2008).
18. Belliveau, N. M. *et al.* Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4796–E4805 (2018).
19. Block, D. H. S., Hussein, R., Liang, L. W. & Lim, H. N. Regulatory consequences of gene translocation in bacteria. *Nucleic Acids Res.* **40**, 8979–8992 (2012).
20. Sousa, C., de Lorenzo, V. & Cebolla, A. Modulation of gene expression through chromosomal positioning in Escherichia coli. *Microbiology* **143 (Pt 6)**, 2071–2078 (1997).

21. Kuhlman, T. E. & Cox, E. C. Gene location and DNA density determine transcription factor distributions in Escherichia coli. *Mol. Syst. Biol.* **8**, 610 (2012).
22. Scholz, S. A. *et al.* High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. *Cell Syst* **8**, 212–225.e9 (2019).
23. Chen, H., Shiroguchi, K., Ge, H. & Xie, X. S. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Molecular Systems Biology* **11**, 808–808 (2015).
24. Esquerre, T. *et al.* Dual role of transcription and transcript stability in the regulation of gene expression in Escherichia coli cells cultured on glucose at different growth rates. *Nucleic Acids Res.* **42**, 2460–2472 (2013).
25. Shearwin, K., Callen, B. & Egan, J. Transcriptional interference – a crash course. *Trends in Genetics* **21**, 339–345 (2005).
26. Callen, B. P., Shearwin, K. E. & Egan, J. B. Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol. Cell* **14**, 647–656 (2004).
27. Brophy, J. A. N. & Voigt, C. A. Antisense transcription as a tool to tune gene expression. *Mol. Syst. Biol.* **12**, 854 (2016).
28. Salgado, H. *et al.* RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research* **41**, D203–D213 (2013).
29. Urtecho, G., Tripp, A. D., Insigne, K., Kim, H. & Kosuri, S. Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically-Encoded Multiplexed Reporter Assay in E. coli. *Biochemistry* (2018). doi:10.1021/acs.biochem.7b01069
30. Garcia, H. G. *et al.* Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Rep.* **2**, 150–161 (2012).
31. Enyeart, P. J. *et al.* Generalized bacterial genome editing using mobile group II introns and

Cre-lox. *Mol. Syst. Biol.* **9**, 685 (2013).

32. Yan, B., Boitano, M., Clark, T. A. & Ettwiller, L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.* **9**, 3676 (2018).
33. Bryant, J. A., Sellars, L. E., Busby, S. J. W. & Lee, D. J. Chromosome position effects on gene expression in Escherichia coli K-12. *Nucleic Acids Res.* **42**, 11383–11392 (2014).
34. Lloréns-Rico, V., Lluch-Senar, M. & Serrano, L. Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. *Nucleic Acids Res.* **43**, 3442–3453 (2015).
35. Yus, E. *et al.* Transcription start site associated RNAs in bacteria. *Mol. Syst. Biol.* **8**, 585 (2012).
36. Fang, X. *et al.* Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10286–10291 (2017).
37. Dornenburg, J. E., Devita, A. M., Palumbo, M. J. & Wade, J. T. Widespread antisense transcription in Escherichia coli. *MBio* **1**, (2010).
38. Georg, J. & Hess, W. R. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 286–300 (2011).
39. Lloréns-Rico, V. *et al.* Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv* **2**, e1501363 (2016).
40. Krummel, B. & Chamberlin, M. J. RNA chain initiation by Escherichia coli RNA polymerase. Structural transitions of the enzyme in early ternary complexes. *Biochemistry* **28**, 7829–7842 (1989).
41. Hawley, D. K. & McClure, W. R. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res.* **11**, 2237–2255 (1983).
42. He, W., Jia, C., Duan, Y. & Zou, Q. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.* **12**, 44 (2018).

43. Lajoie, M. J. *et al.* Probing the limits of genetic recoding in essential genes. *Science* **342**, 361–363 (2013).

44. Flashner, Y. & Gralla, J. D. Dual mechanism of repression at a distance in the lac operon. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 8968–8972 (1988).

45. Czarniecki, D., Noel, R. J., Jr & Reznikoff, W. S. The -45 region of the Escherichia coli lac promoter: CAP-dependent and CAP-independent transcription. *J. Bacteriol.* **179**, 423–429 (1997).

46. Li, G.-Y., Zhang, Y., Inouye, M. & Ikura, M. Structural mechanism of transcriptional autorepression of the Escherichia coli RelB/RelE antitoxin/toxin module. *J. Mol. Biol.* **380**, 107–119 (2008).

47. Grogan, D. W. & Cronan, J. E., Jr. Cloning and manipulation of the Escherichia coli cyclopropane fatty acid synthase gene: physiological aspects of enzyme overproduction. *J. Bacteriol.* **158**, 286–295 (1984).

48. Chang, Y. Y. & Cronan, J. E., Jr. Membrane cyclopropane fatty acid content is a major factor in acid resistance of Escherichia coli. *Mol. Microbiol.* **33**, 249–259 (1999).

49. Ebright, R. H. Transcription activation at Class I CAP-dependent promoters. *Mol. Microbiol.* **8**, 797–802 (1993).

50. Williams, S. M., Savery, N. J., Busby, S. J. & Wing, H. J. Transcription activation at class I FNR-dependent promoters: identification of the activating surface of FNR and the corresponding contact site in the C-terminal domain of the RNA polymerase alpha subunit. *Nucleic Acids Res.* **25**, 4028–4034 (1997).

51. Browning, D. F. & Busby, S. J. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* **2**, 57–65 (2004).

52. Rojo, F. Repression of transcription initiation in bacteria. *J. Bacteriol.* **181**, 2987–2991 (1999).

53. Paggi, J. *et al.* Predicting Transcriptional Regulatory Activities with Deep Convolutional

Networks. *bioRxiv* 099879 (2017). doi:10.1101/099879

54. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1530 (2018).
55. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
56. Johns, N. I. *et al.* Metagenomic mining of regulatory elements enables programmable species-selective gene expression. *Nat. Methods* **15**, 323–329 (2018).
57. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14024–14029 (2013).
58. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
59. Bushnell, B. BBMap short read aligner. (2016).
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
61. Shahmuradov, I. A., Mohamad Razali, R., Bougouffa, S., Radovanovic, A. & Bajic, V. B. bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. *Bioinformatics* **33**, 334–340 (2017).
62. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. *R package version* **2**, (2017).
63. Ghandi, M. *et al.* gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207 (2016).

CHAPTER FOUR

Conclusion and Future Directions

Summary of Novel Technology and Findings

In this work I discuss two different projects aimed at dissecting sequence-function relationships in two different biological systems. The first project, discussed in Chapter 2, focuses on the impact of sequence variants on exon recognition in human cell lines. The Multiplexed Functional Assay for Splicing using Sort-Seq (MFASS) is a novel technique that enables functional screening of tens of thousands of variants, both intronic and exonic, integrated at a fixed genomic location at single copy. It leverages a intron-exon-intron mini-gene inserted into a split-GFP reporter, followed by flow cytometry sorting into distinct populations, and finally next-generation sequencing to quantitatively measure the level of exon skipping for each designed sequence. This is powerful tool that can be used to assess the functional impact of variants of unknown significance, arising in both clinical applications and computational predictions. We used MFASS to determine that a previously under-appreciated proportion of rare variants cause large-effect splicing changes, and occur at locations other than the canonical splice sites. This suggests an under-studied source of variants that may be implicated in complex diseases caused by aberrations in splicing.

The second project, discussed in Chapter 3, dissects various aspects of transcriptional regulation in *E. coli*. We developed a massively parallel report assay that clones a 150bp designed synthetic DNA sequence upstream of a GFP reporter and short 20bp molecular barcode and use next-generation sequencing to quantitatively measure the level of promoter activity. We functionally annotate the genome in a systematic fashion for regions capable of driving expression in both rich glucose media and minimal media. We designed multiple synthetic libraries to understand which sequences are responsible for regulation and develop machine learning models based on these datasets.

Future Directions

The ultimate goal of sequence-function relationships is to develop a model that can accurately predict the level of activity for a given sequence based on the sequence alone. The work presented here is an important first step in generating training data for future machine learning models. We hope our work enables future research in frameworks such as active learning or Bayesian experimental design, methods which formally maximize the expected utility of an experimental outcome. One can envision designing synthetic libraries that are optimally designed to be the most valuable for predictive modeling. For example, a biophysical model of promoter function could be further refined with libraries that systematically vary the potential sequence space for those parameters, exploring beyond what exists in natural sequence. This feature-driven design can more effectively probe the massive parameter space of individual features at a level of variation that is impossible from natural sequence alone, whether promoters or splice sites. One could also design libraries containing sequences that would behave differently under competing models, potentially guiding model selection and establishing causality. This Bayesian experimental design could be conducted in an iterative fashion – each experiment will further refine previous models, act as test sets for previous models, guide model selection, and inform the next best set of experiments. Using the approach outlined above, it could be possible to distinguish between multiple hypotheses and discover the true underlying biological mechanism. Ultimately, the focus of predictive modeling in biology should not solely be accuracy, but on interpretability and uncovering true biological insights.