A massively parallel approach to understanding genomic information

Alexander Rosenberg, Rupali Pathwardan, Jay Shendure, Georg Seelig

Electrical Engineering and Computer Science & Engineering, University of Washington

Sequencing genome.

Complete.

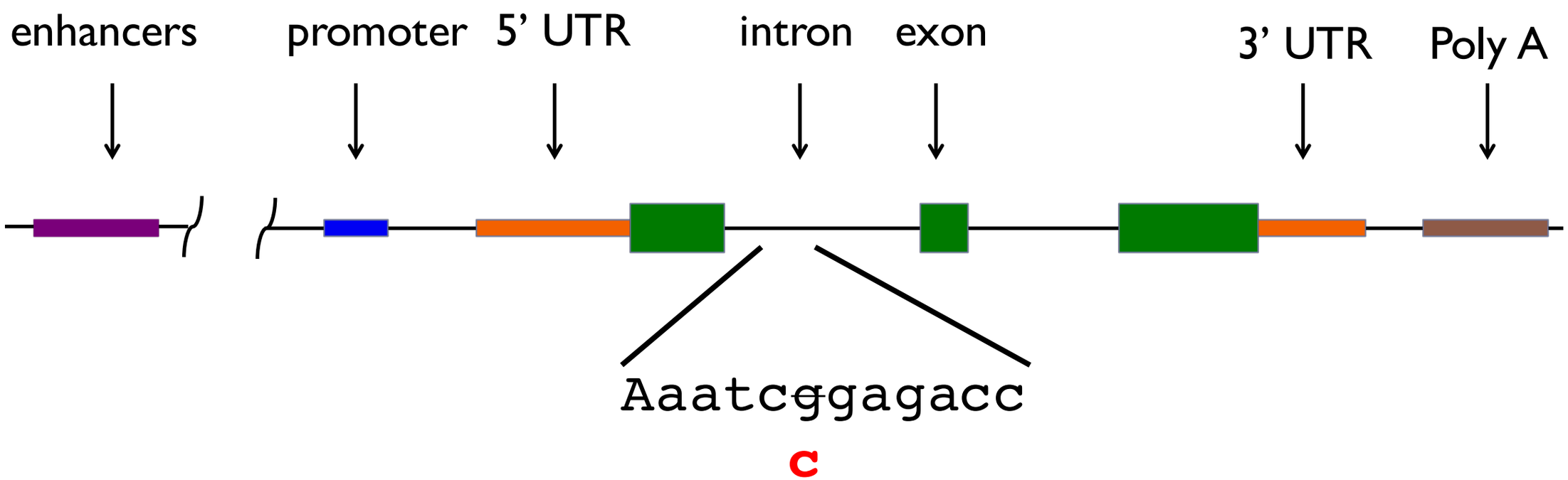
Compiling list of variants.

Complete.

Interpreting genome

Jay Shendure

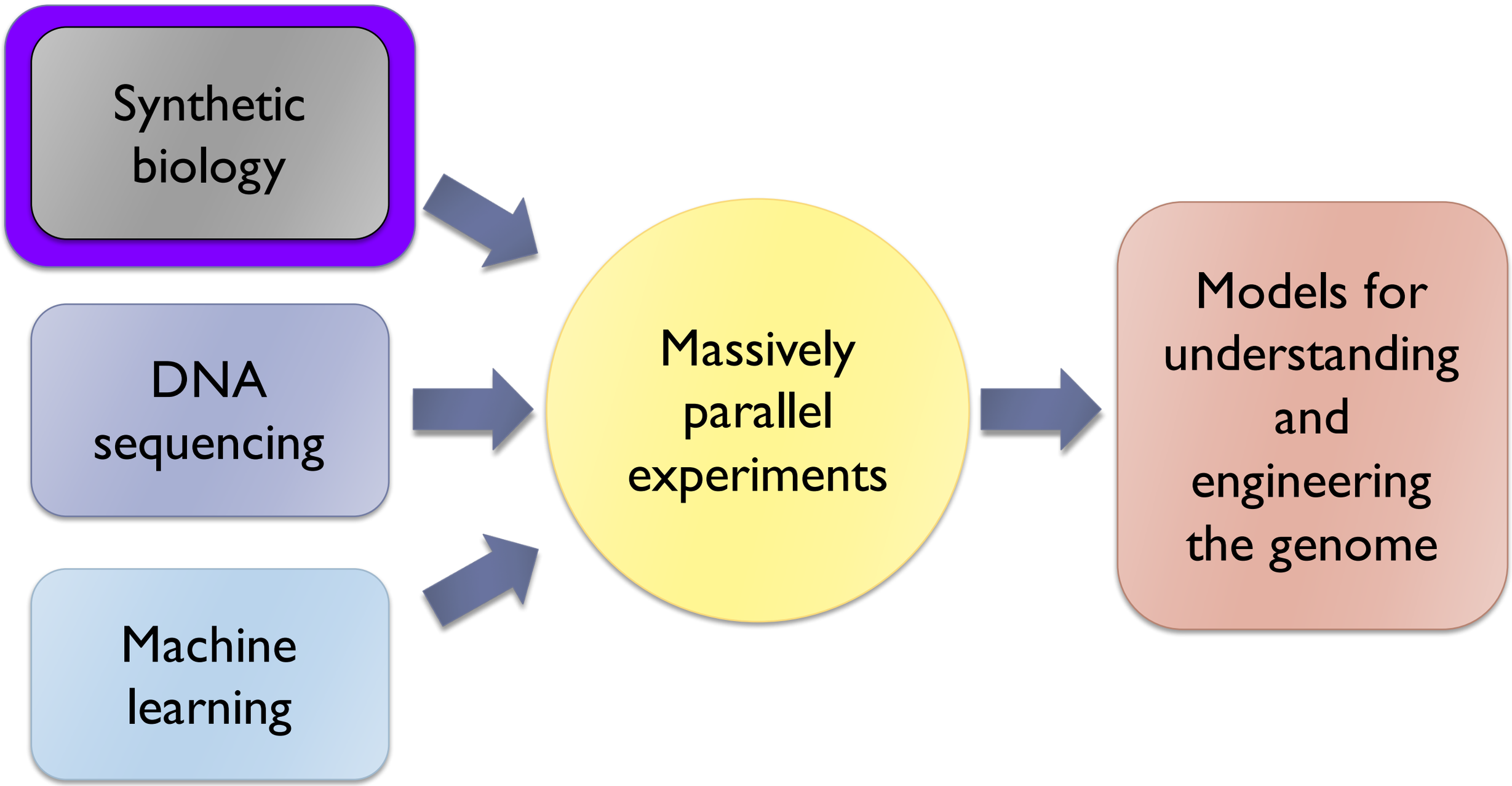
Understanding the impact of variant with machine learning



- Build a sequence-function model using machine learning
- Model are limited by data (e.g. "only" 50K splice events)

More data is better

A massively parallel approach to understanding the genome

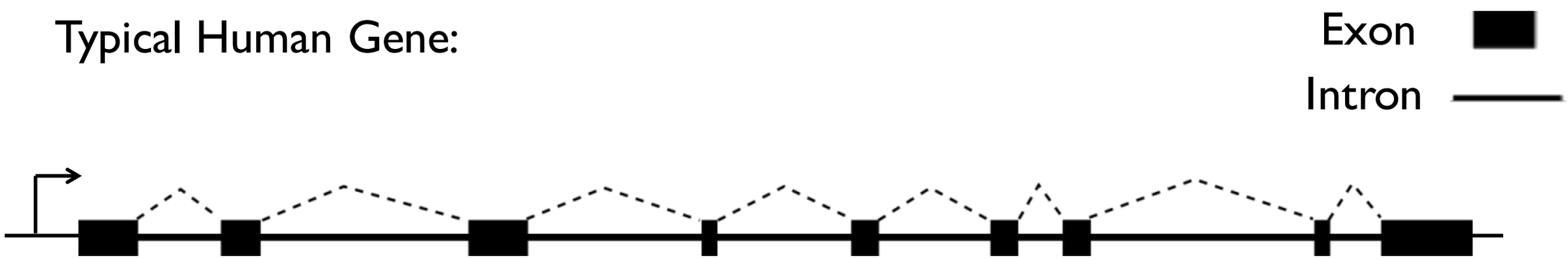


Overview

- **A massively parallel approach to understanding sequence-function relationship: 5'alternative splicing**
- Cell-type specific effects in alternative splicing
- Skipped exons: attempt 1
- Skipped exons and 3' alternative splicing: exon definition

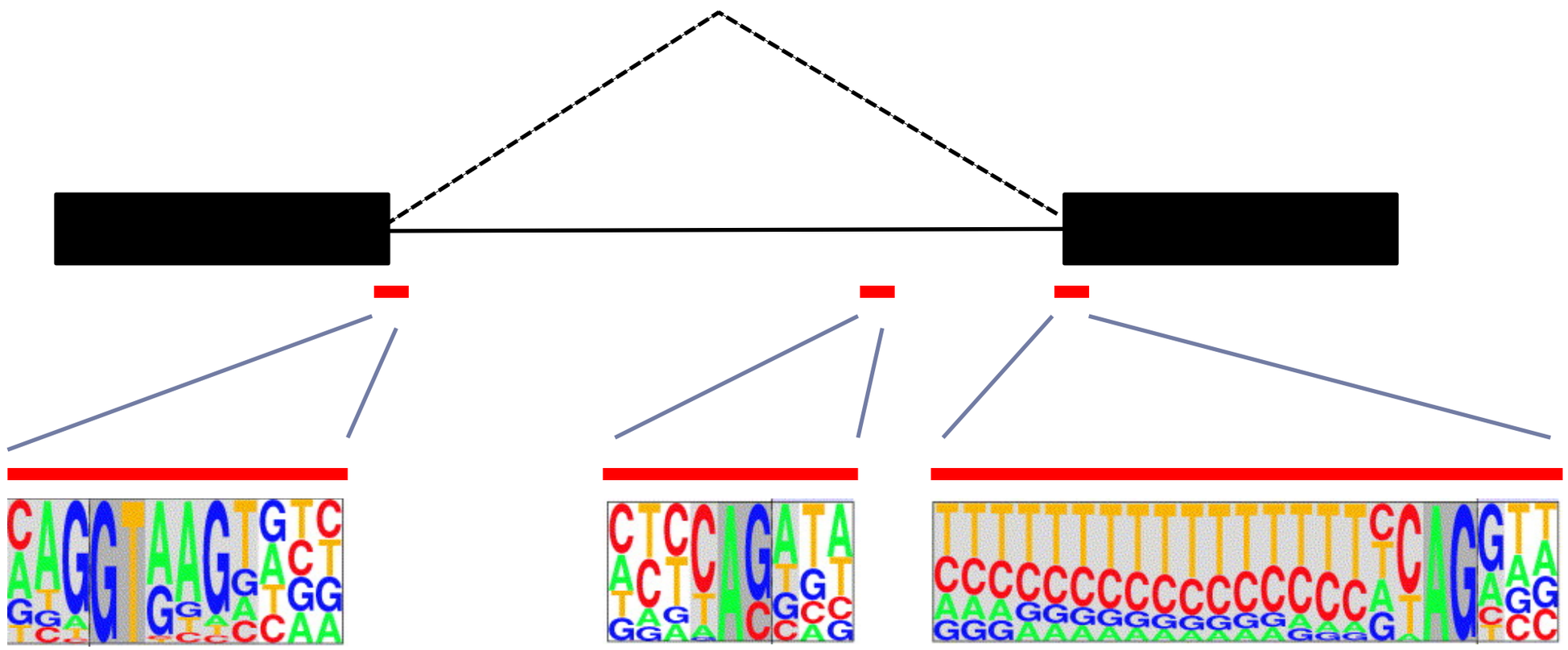
RNA-Splicing

Typical Human Gene:



Core splicing signals

- Splicing is regulated by cis-regulatory sequences motifs and a trans-acting RNA-protein complex, the spliceosome



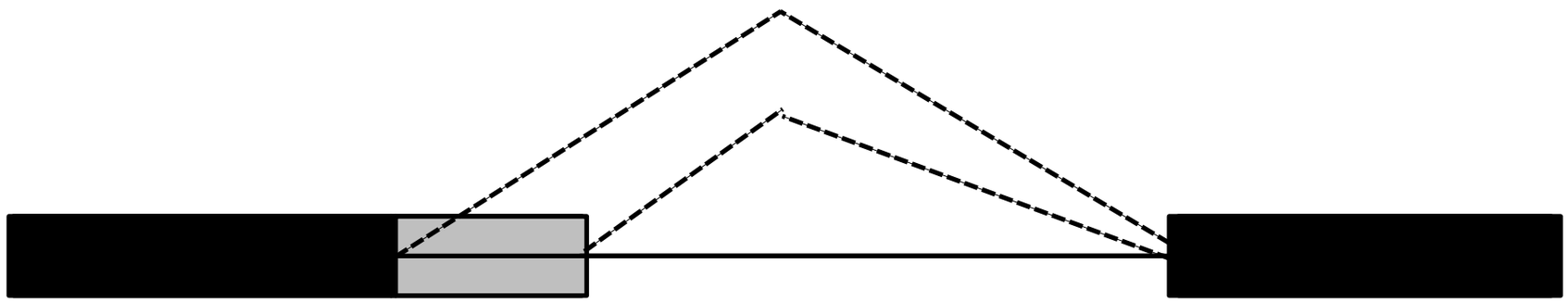
Splice donor

Branch point

PPT + Splice acceptor

Alternative Splicing

- Different isoforms can have distinct protein functions
- 95% of coding genes are alternatively spliced
- Misregulation of splicing can lead to disease and cancer



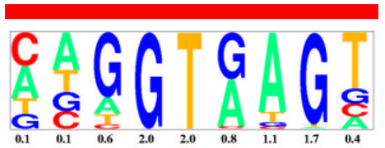
Isoform A

Isoform B

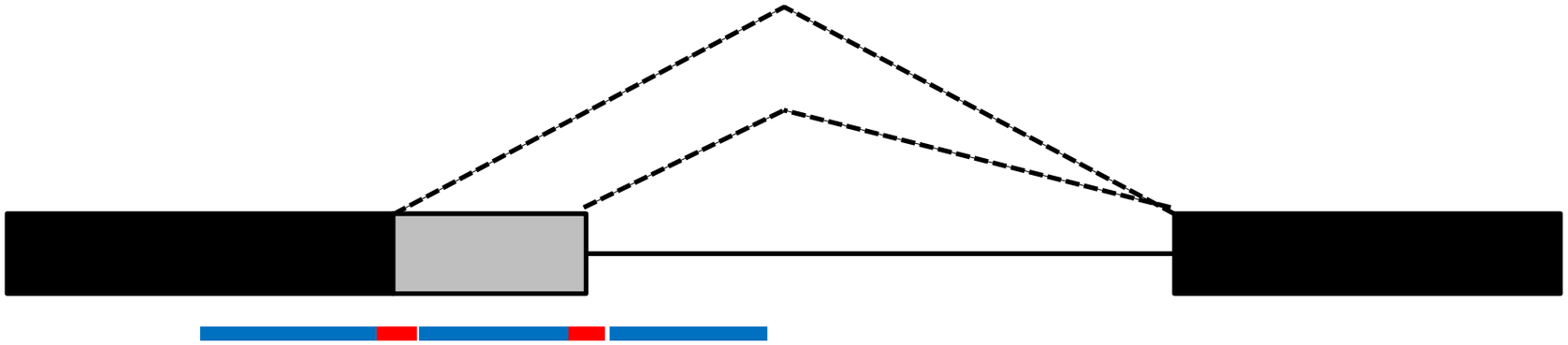
Regulation of Alternative Splicing

What are the sequence determinants of alternative splicing?

- The splice site sequences (splice donors)

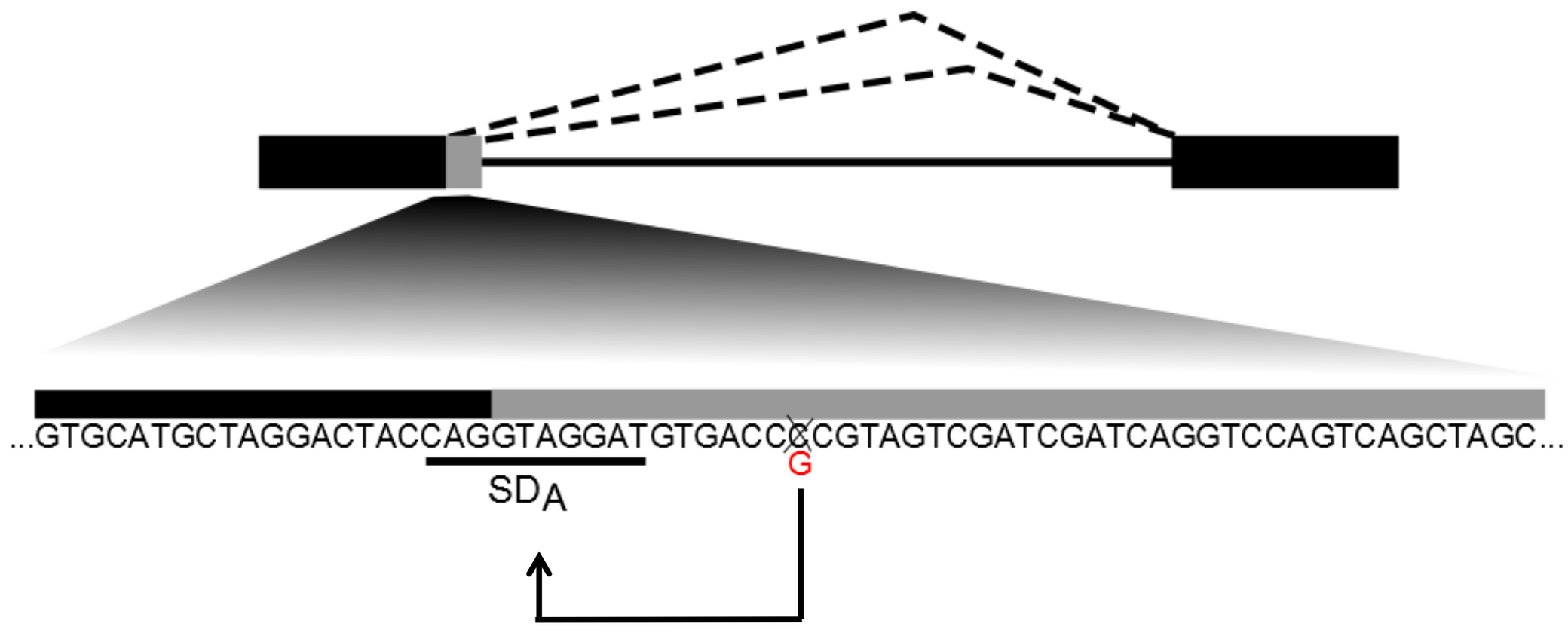


- Sequences around the splice sites



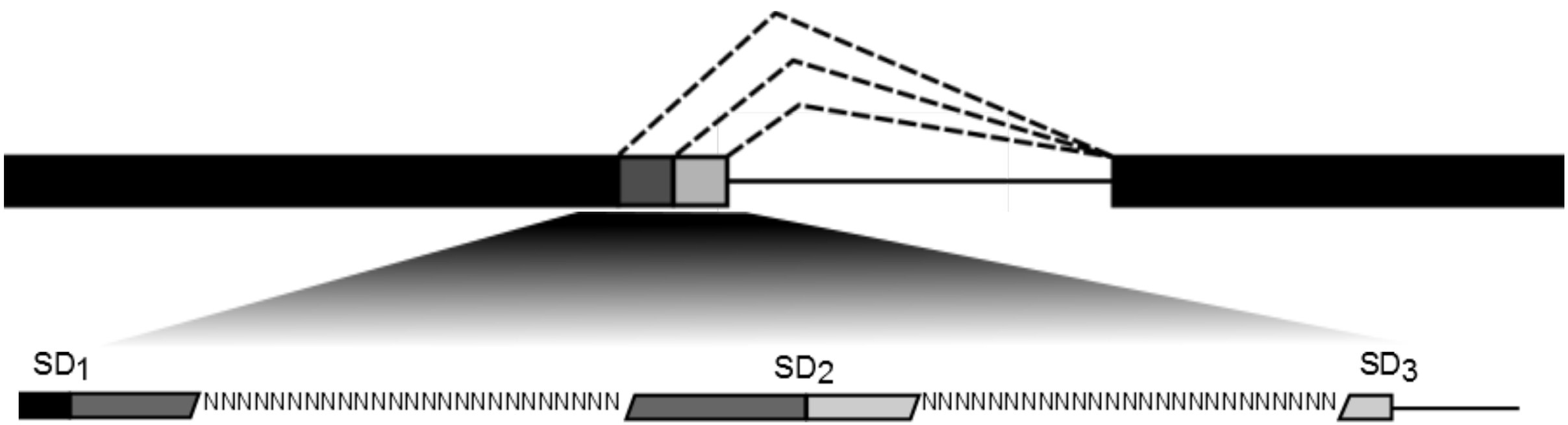
Effects of Single Nucleotide Polymorphisms (SNPs) on Alternative Splicing in Humans

- Can we create a model that predict the effects of nucleotide changes on alternative splicing?

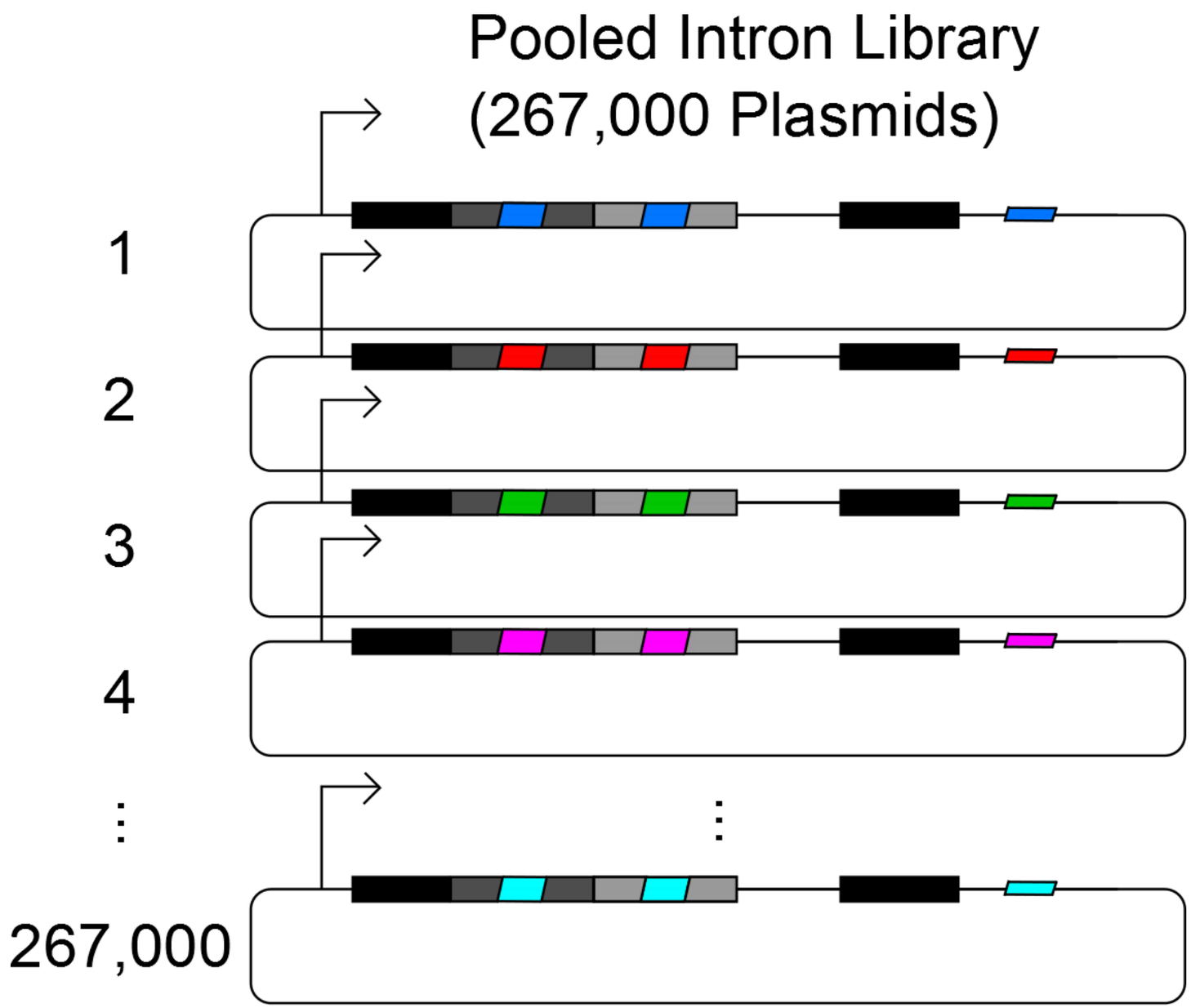


Massively Parallel Splicing Assay

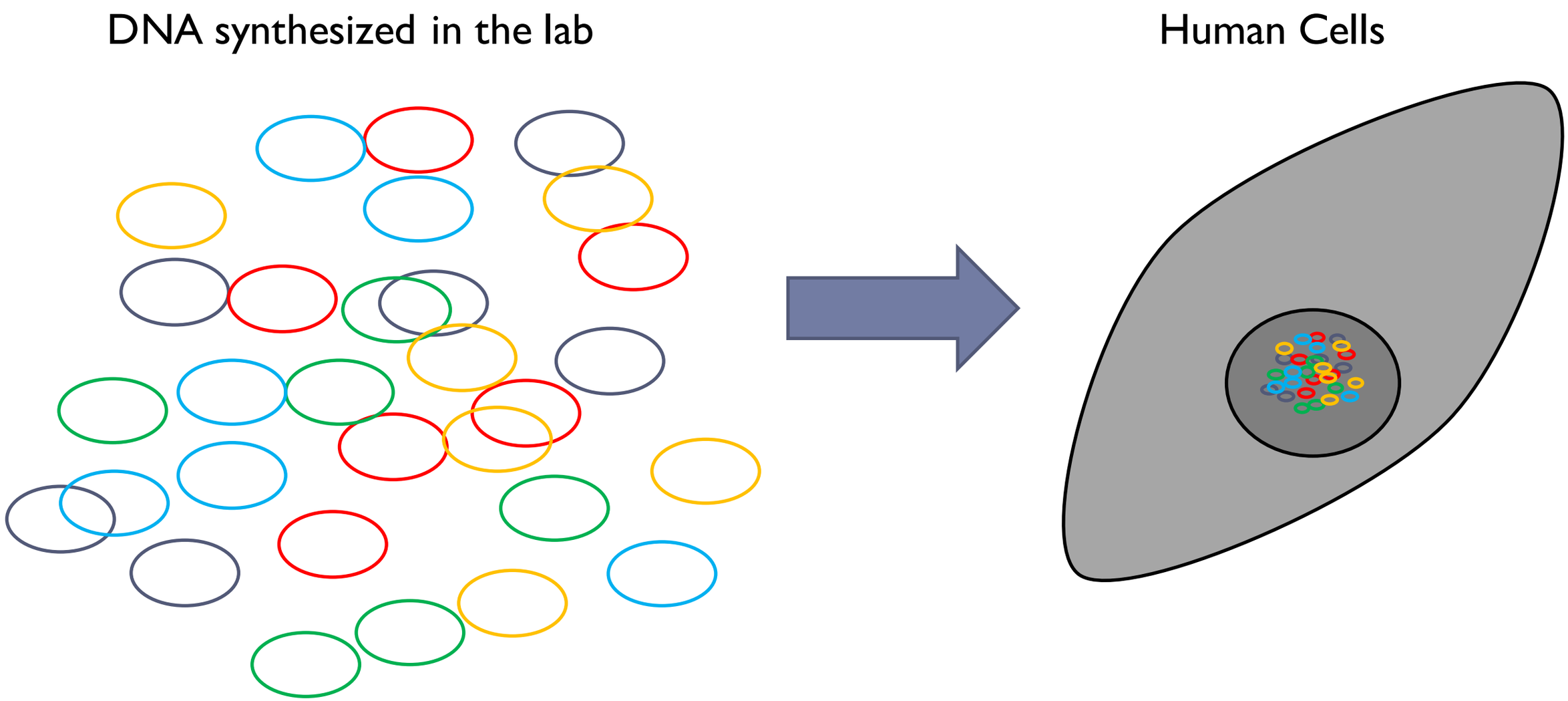
- Alternatively spliced plasmid mini-gene with 3 splice donors
- Introduced degenerate nucleotide sequences between the splice donors
- **How does sequence variation in these positions affect alternative splicing?**



Massively Parallel Splicing Assay

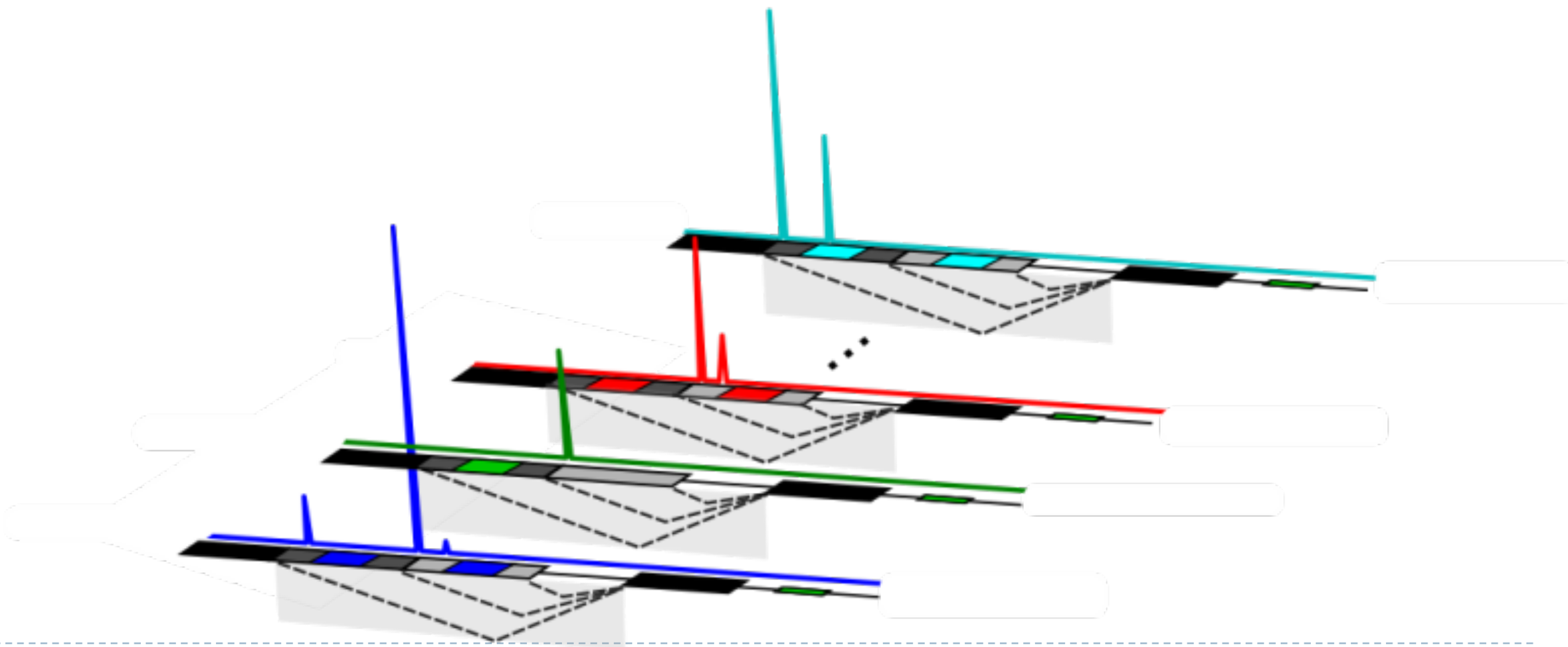


Let's give a cell lots of DNA sequences and record what happens



Massively Parallel Splicing Assay

- Used RNA-seq to quantify isoform levels
- For every mRNA molecule that we sequenced we determined:
  - how it spliced
  - which plasmid variant it was transcribed from (barcode in 3'UTR)

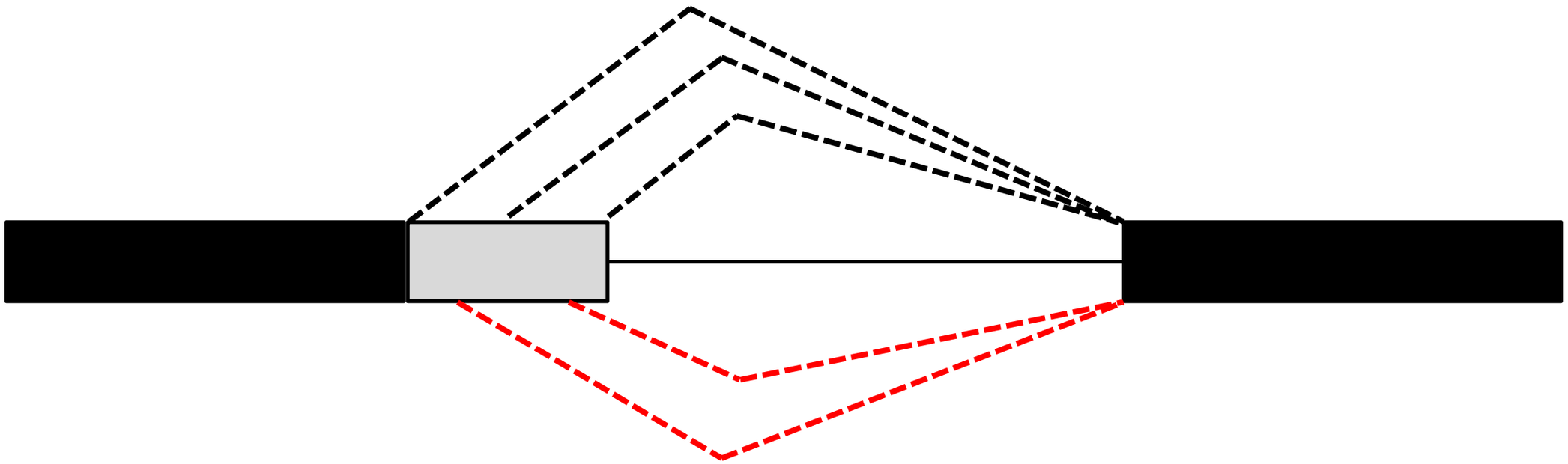


Resulting Data

	SD_1	SD_2	SD_3	SD_{NEW}
	0	26	0	0
	0	2	0	27
	113	4	1	0
⋮			⋮	

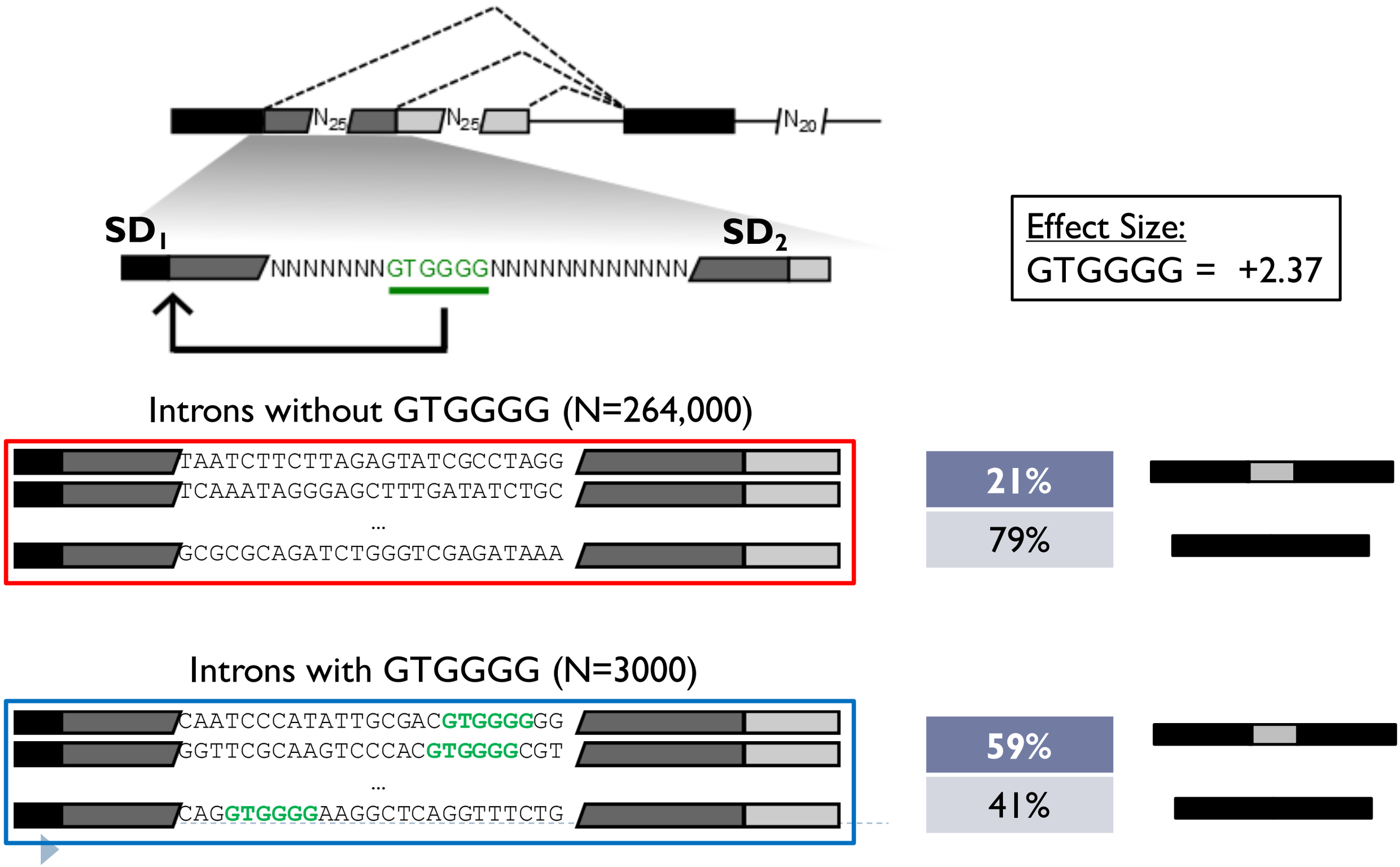
267,000
Different
Sequences

Resulting Data - Summary

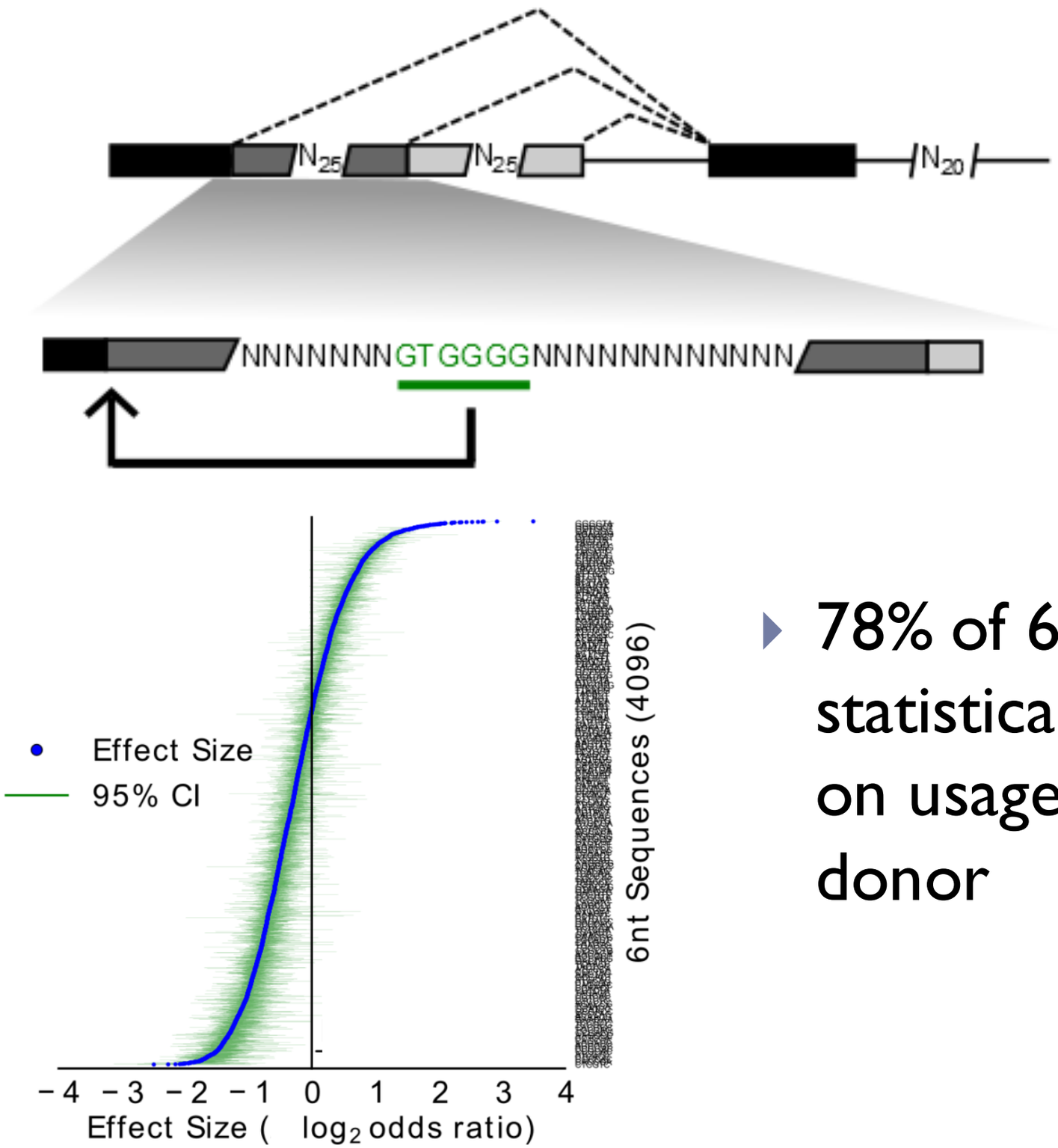


SD_1	SD_2	SD_3	SD_{NEW}
28%	47%	6%	15%

Short Sequence Motif Effect Sizes

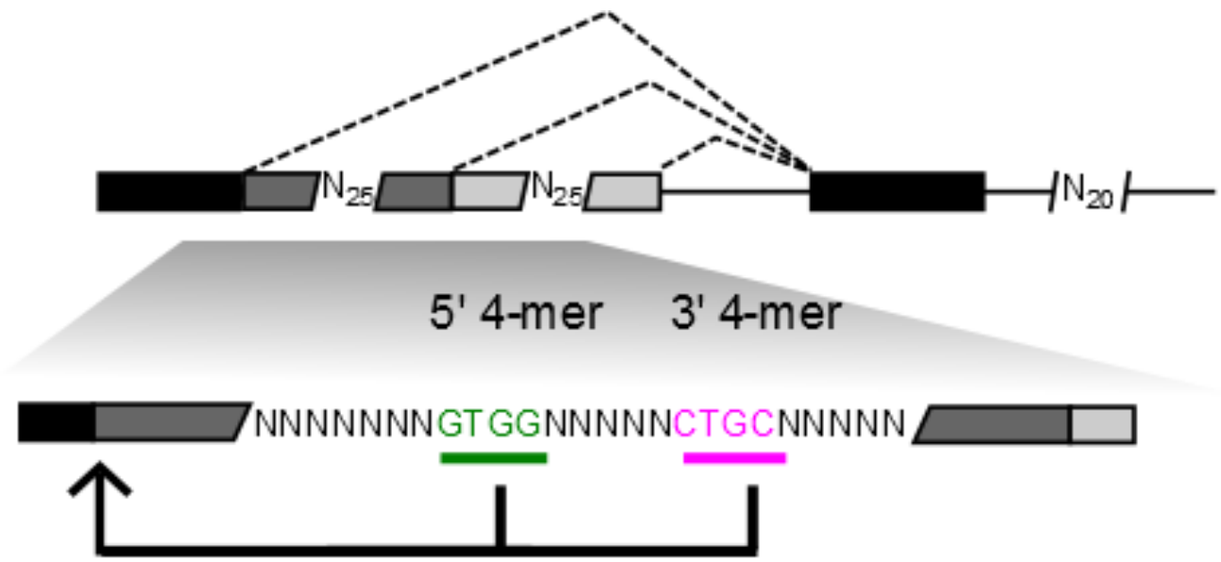


All 6-mer Effect Sizes



- 78% of 6-mers have statistically significant effect on usage of the first splice donor

Combinatorial Regulation of Alternative Splicing

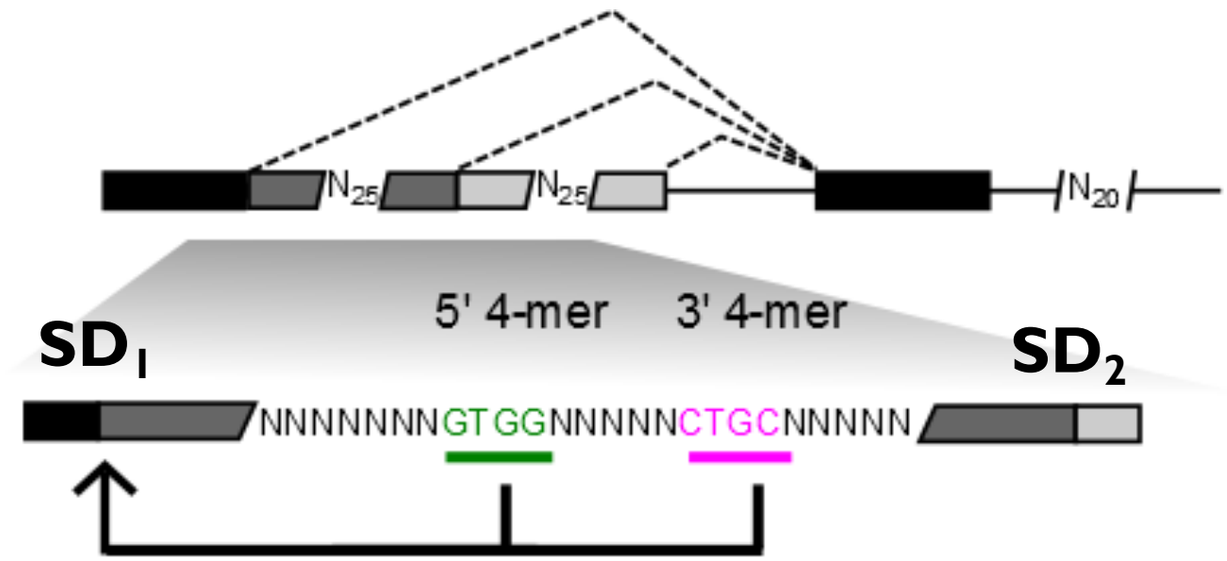


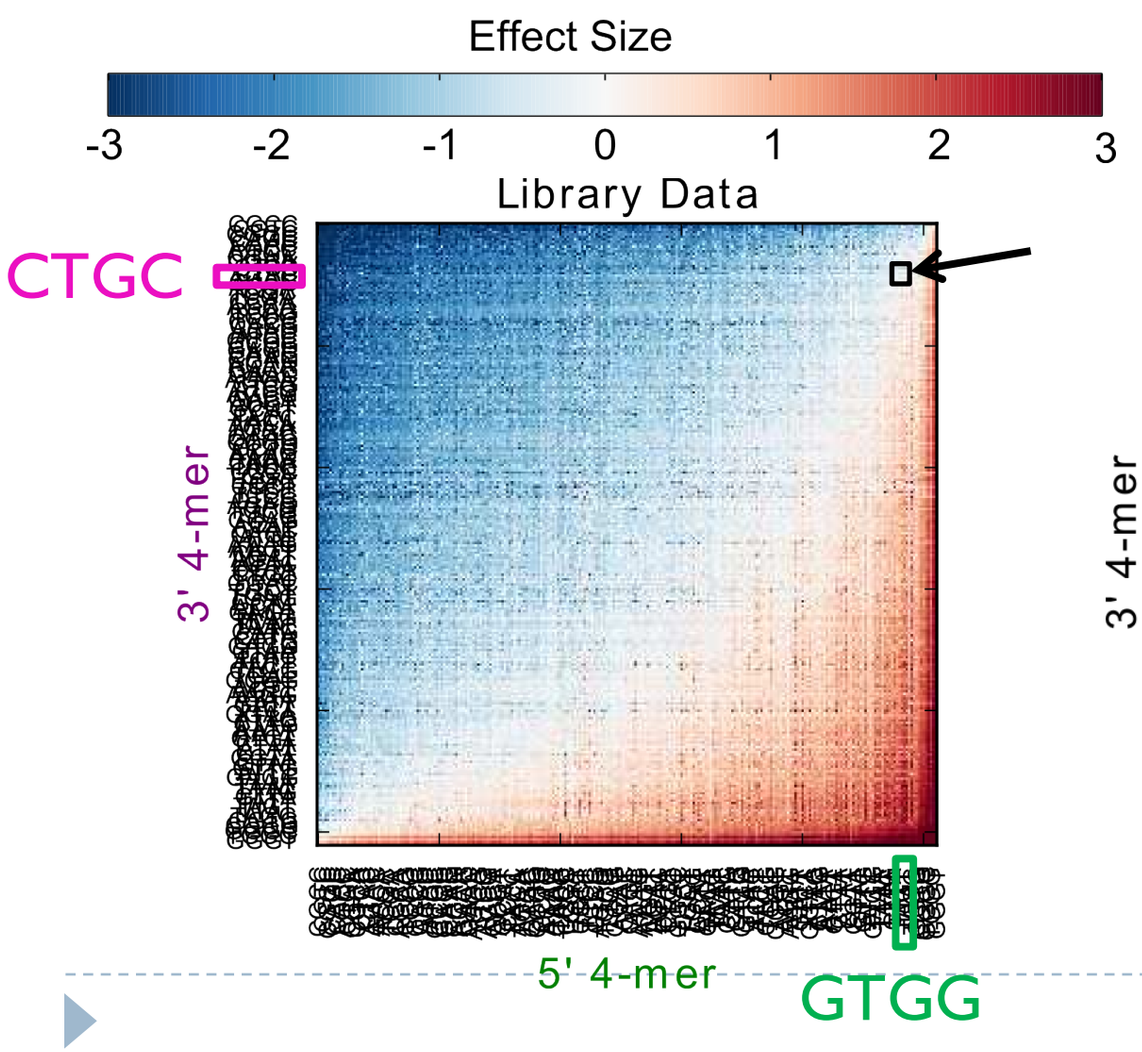
Two Possible Models of Combinatorial Sequence Regulation:

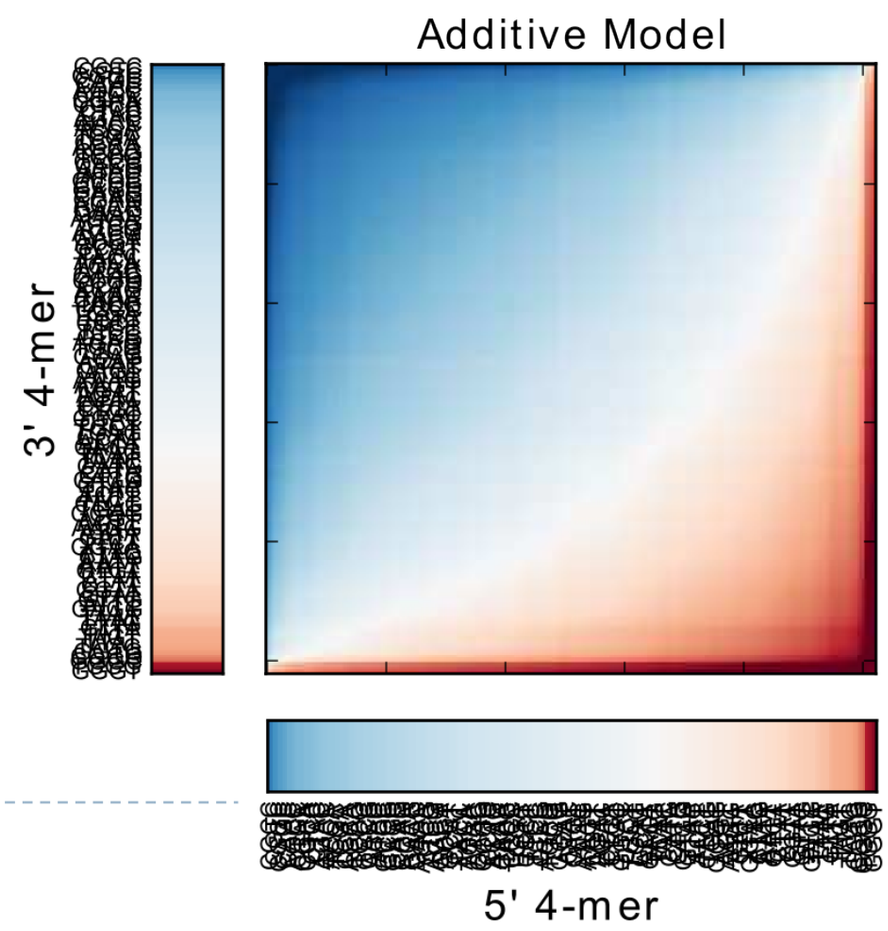
- Additive: Sequence motifs act independently of each other
  - Effect Size(GTGG & CTGC) = Effect Size(GTGG) + Effect Size(CTGC)
- Cooperative: Sequence motifs interact with other motifs

Combinatorial Regulation of Alternative Splicing

- Short motifs act additively and independently of each other

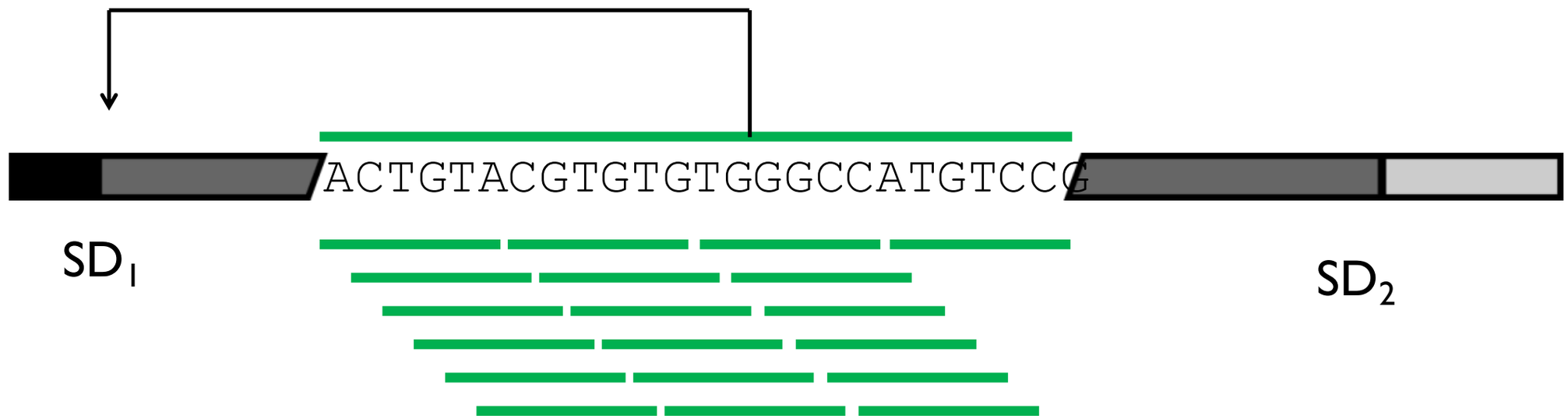






R^2=0.89

Building an Additive Model of Splicing

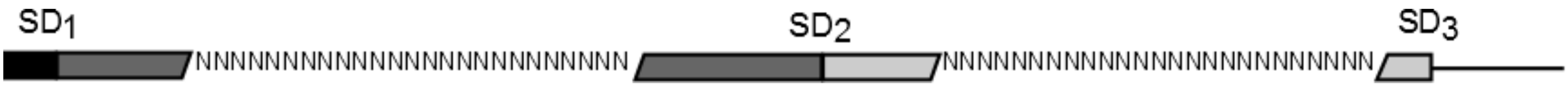


- Effect Size(ACTGTACGTGTGTGGGCCATGTCCG) = Effect Size (ACTGTA)
  + Effect Size (CTGTAC)
  + Effect Size (TGTACG)
  …
  + Effect Size (TGTCCG)

Individual Contribution of a Nucleotide to Splicing

- Effect Size(**G** at position 12) = (Effect Size (CGTGT**G**)
  + Effect Size (GTGT**G**T)
  + Effect Size (TGT**G**TG)
  + Effect Size (GT**G**TGG)
  + Effect Size (T**G**TGGG)
  + Effect Size (**G**TGGGC)) / 6

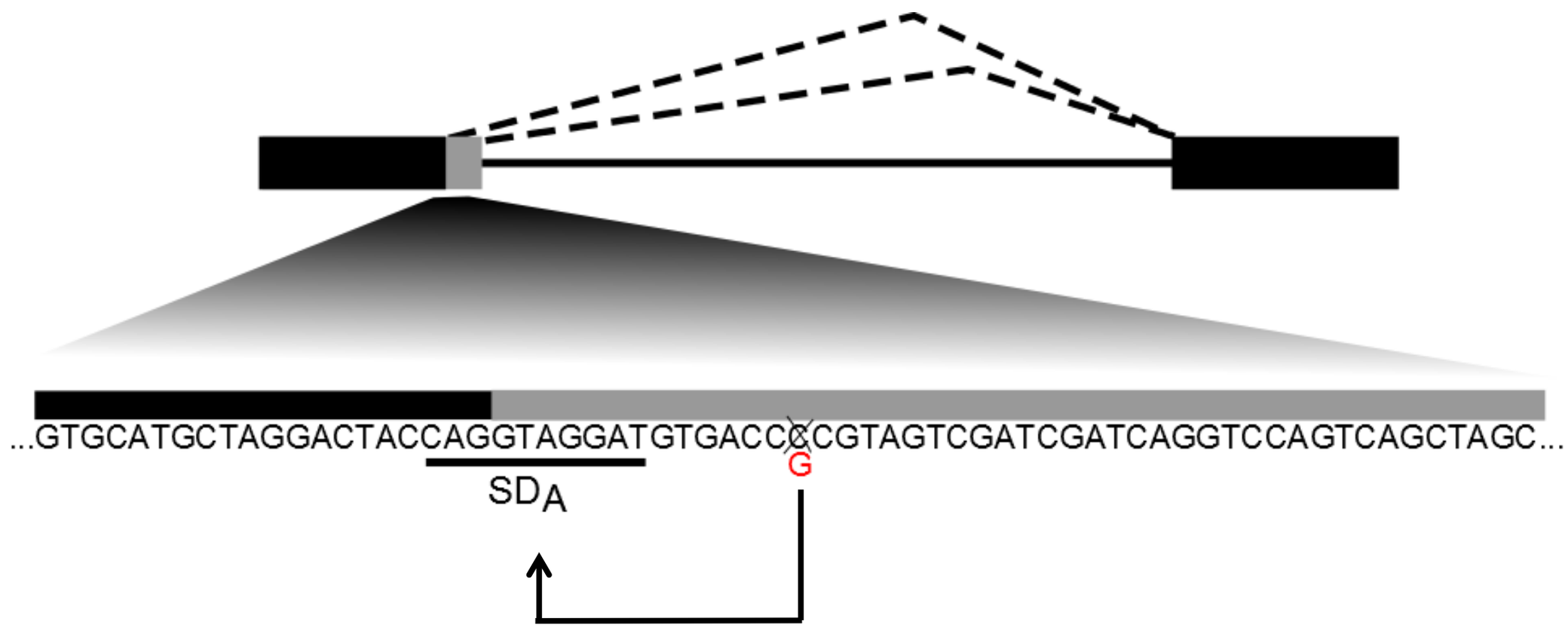
Testing An Additive Model



- Trained model using multinomial logistic regression
- Tested the accuracy of model predictions on a test set
- For each intron variant:
  - Score every potential splice site
  - Convert splice donor scores into splicing probabilities (softmax function)

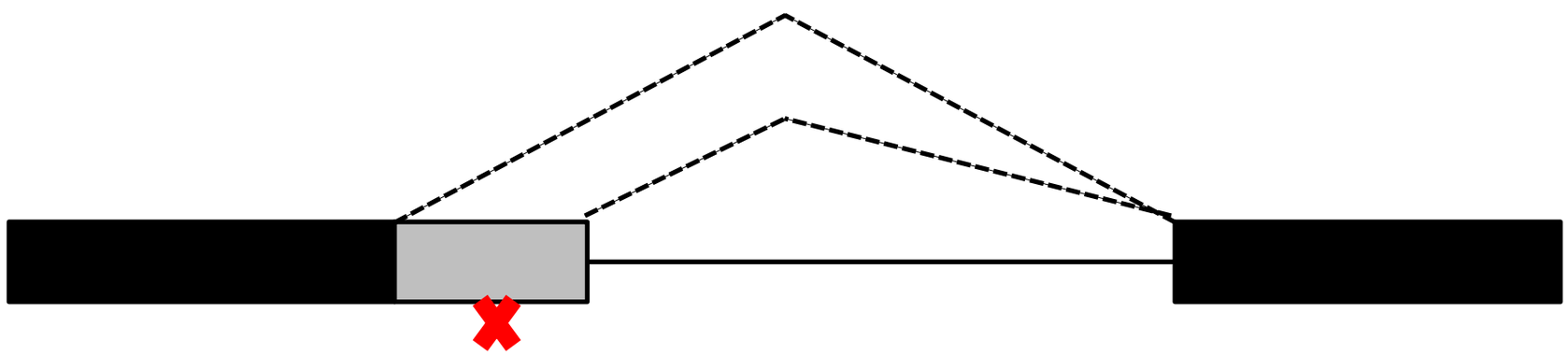
Effects of Single Nucleotide Polymorphisms (SNPs) on Alternative Splicing in Humans

- Can our model predict the effects of nucleotide changes on alternative splicing?

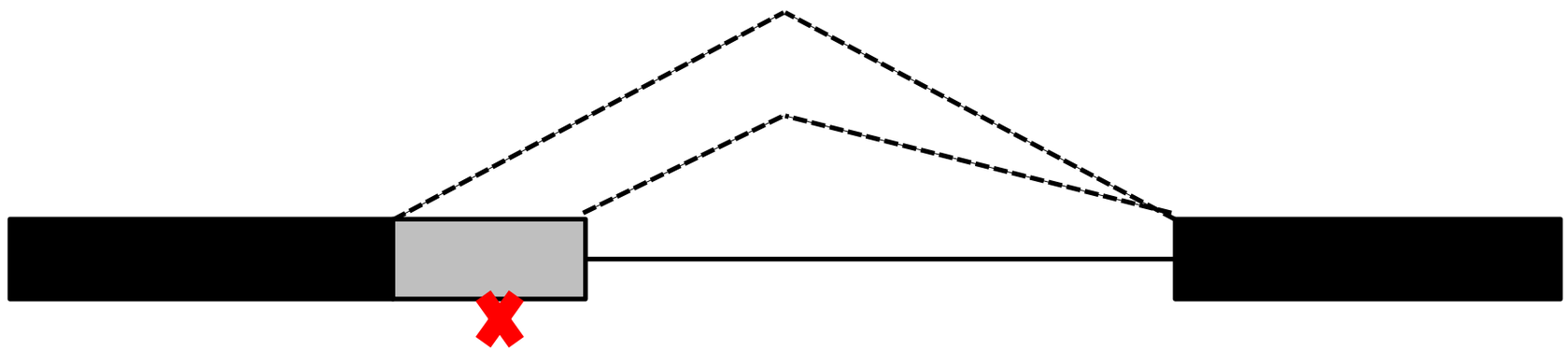


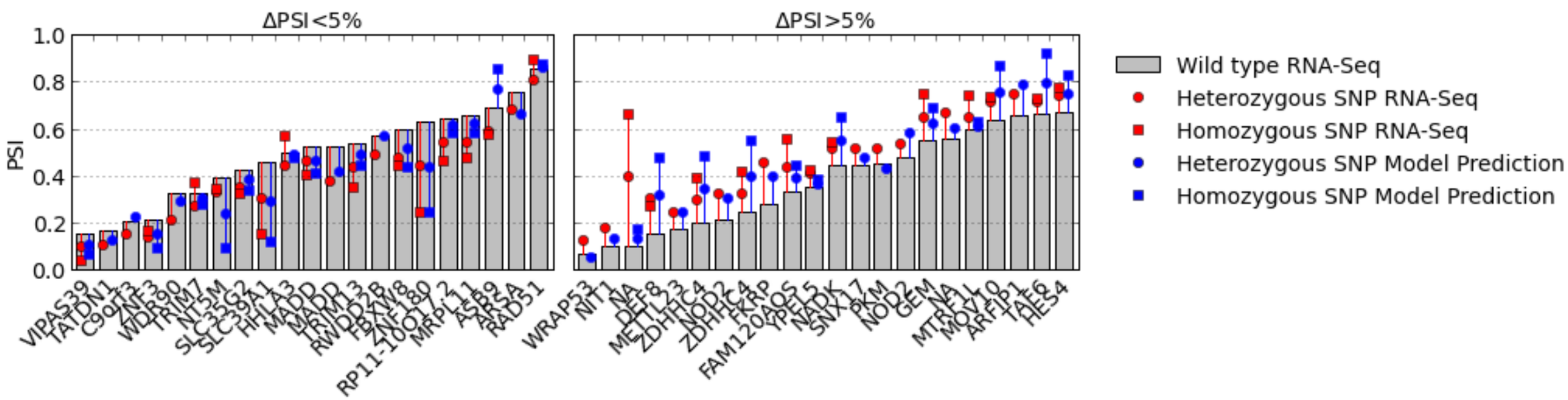
Measuring the Effects of SNPs on Alternative Splicing

- Started with a list of alternatively spliced human genes
- Used Thousand Genomes data and RNA-seq data from GEUVADIS to calculate isoform percentage for:
  - Individuals with a SNP
  - Individuals with no SNP

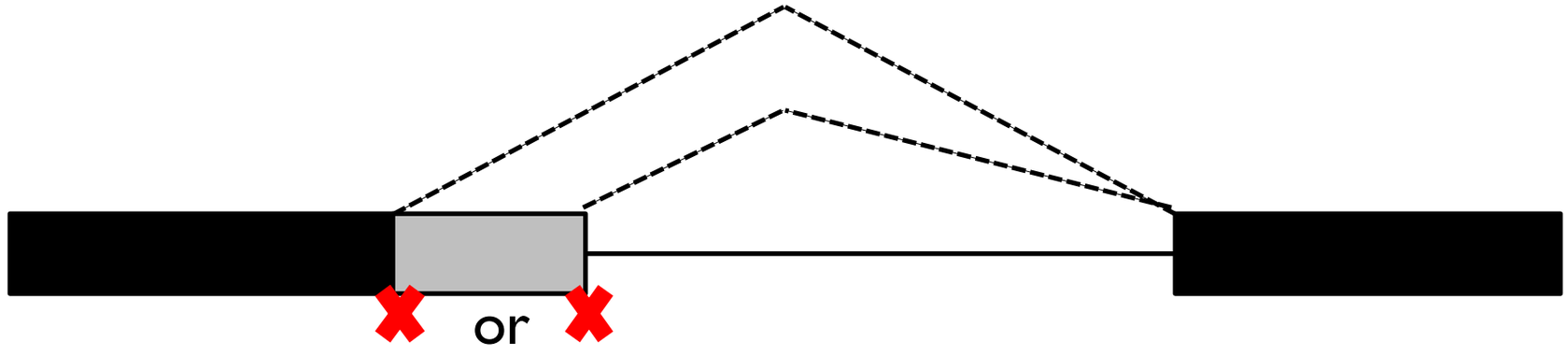


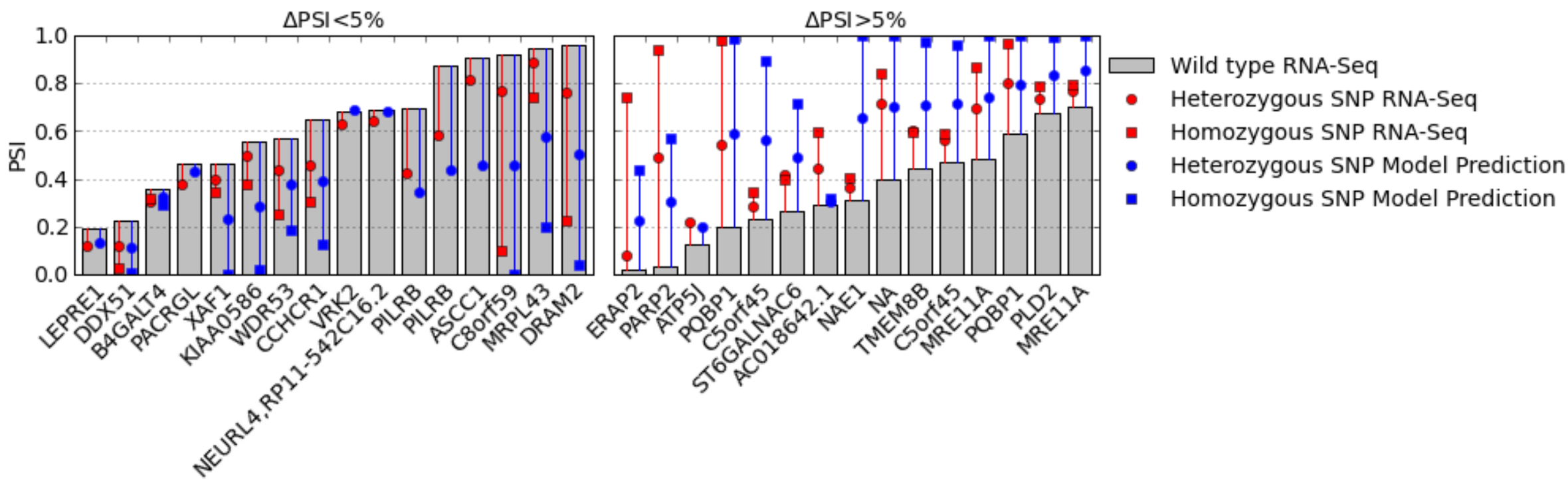
Predicting Effects of SNPs between Alternative Splice Donors





Predicting Effects of SNPs in an Alternative Splice Donor



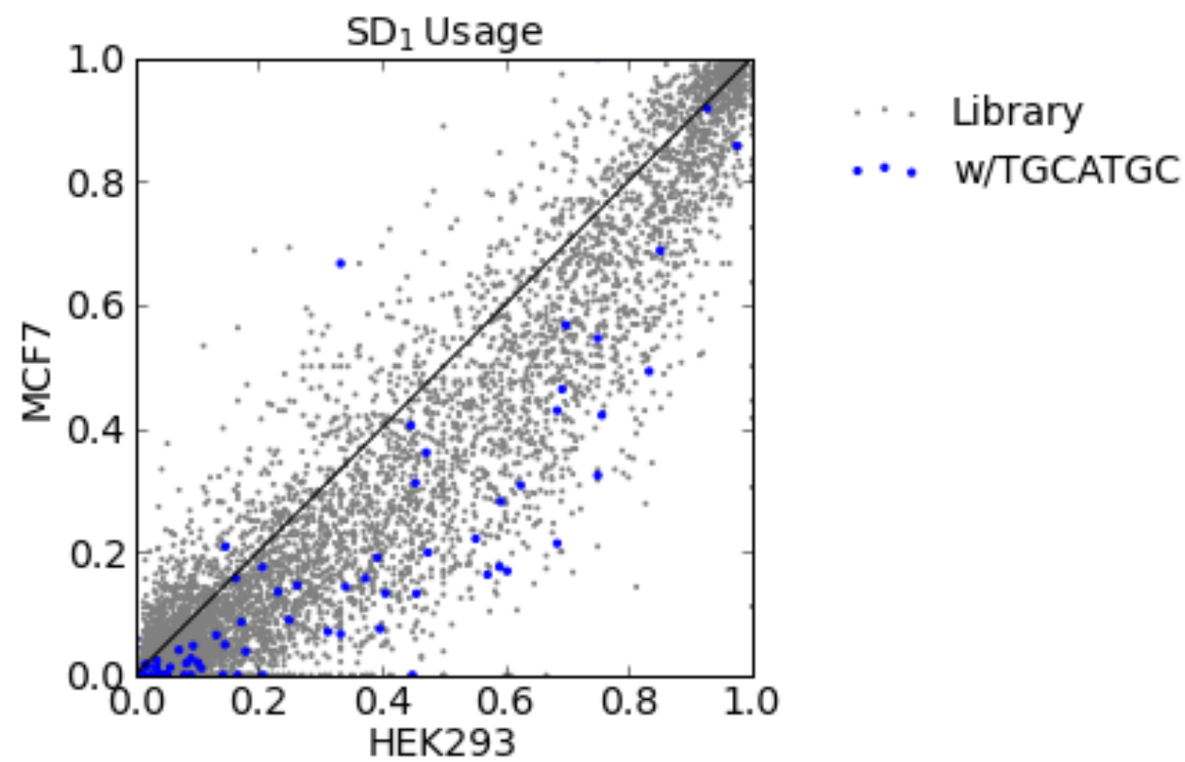


Overview

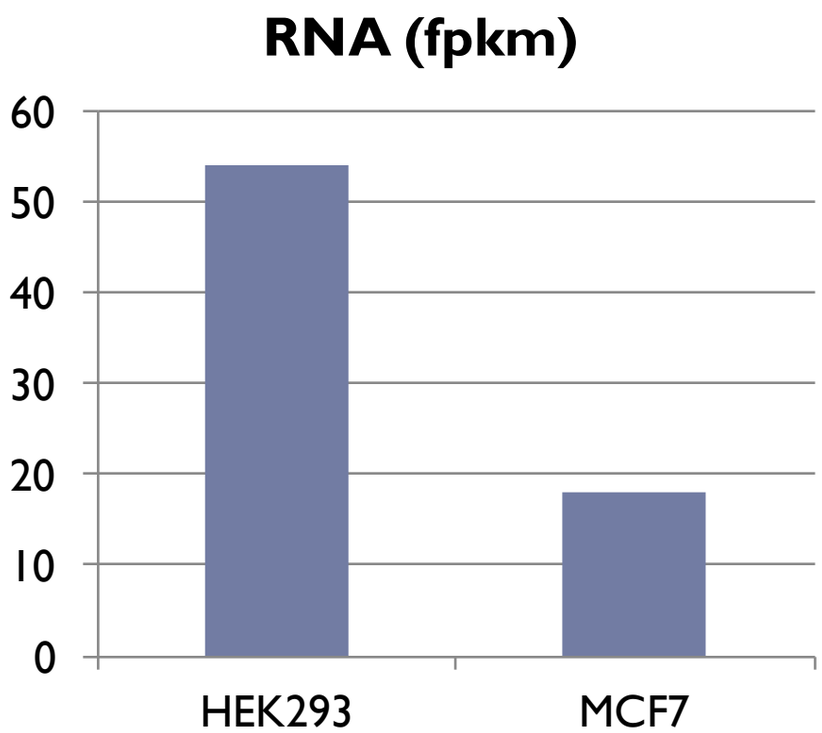
- A massively parallel approach to understanding sequence-function relationship: 5'alternative splicing
- **Cell-type specific effects in alternative splicing**
- Skipped exons: attempt 1
- Skipped exons and 3' alternative splicing: exon definition

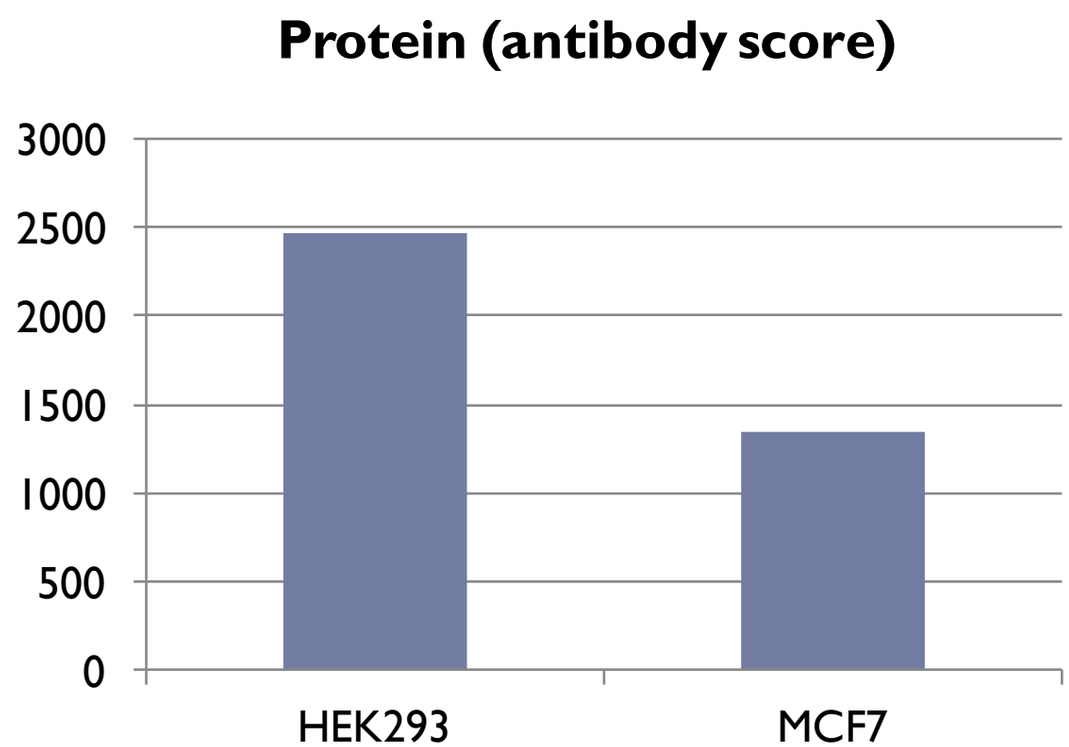
RBFOX1/2 Binding Site Differences in HEK293 and MCF7 Cells

Rank	Motif
1	TGCATG
2	GCATGC
3	CGCATG
4	TCGCCT
5	ATGCAT
6	ACGACA
7	ACGACG
8	AGCCCC
9	CTCGGC
10	CATGCA
11	CCCCAC
12	AGCATG
13	AACGAC



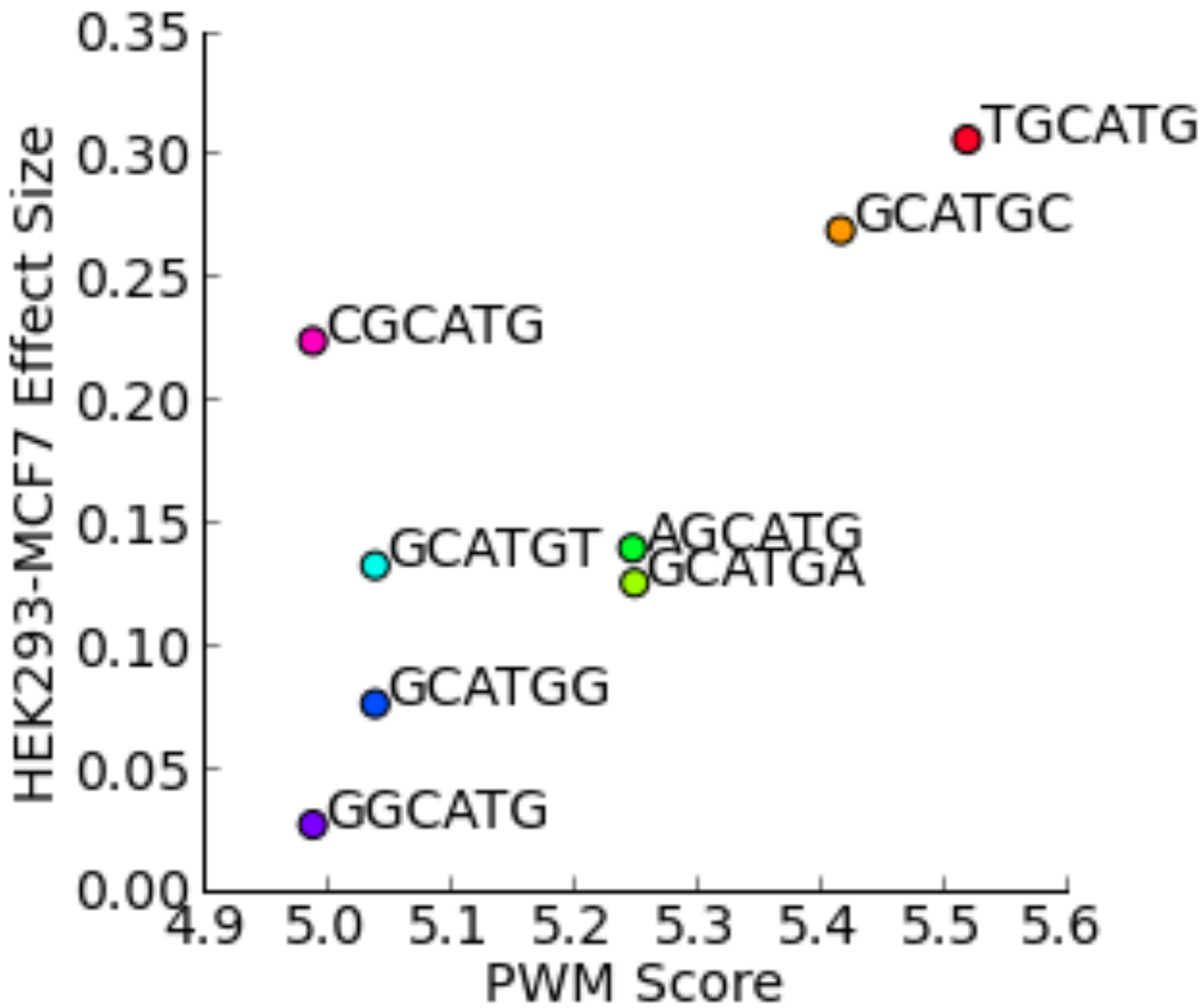
RBFOX2 Expression in HEK293 vs MCF7





The Human Protein Atlas

RBFOX1/2 Binding Site Differences in HEK293 and MCF7 Cells

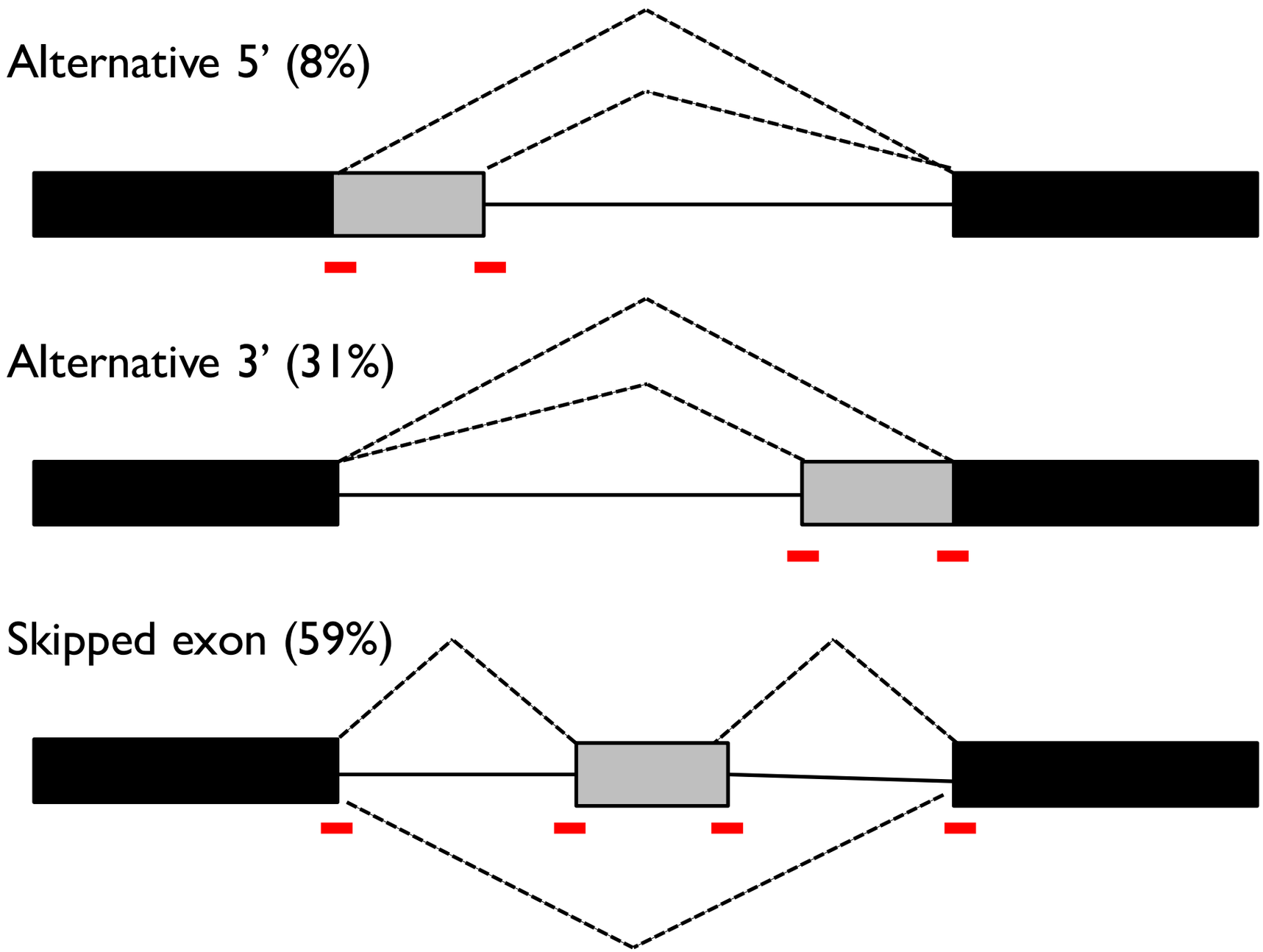


Ray, Debashish, et al. "A compendium of RNA-binding motifs for decoding gene regulation." *Nature* 499.7457 (2013): 172-177.

Overview

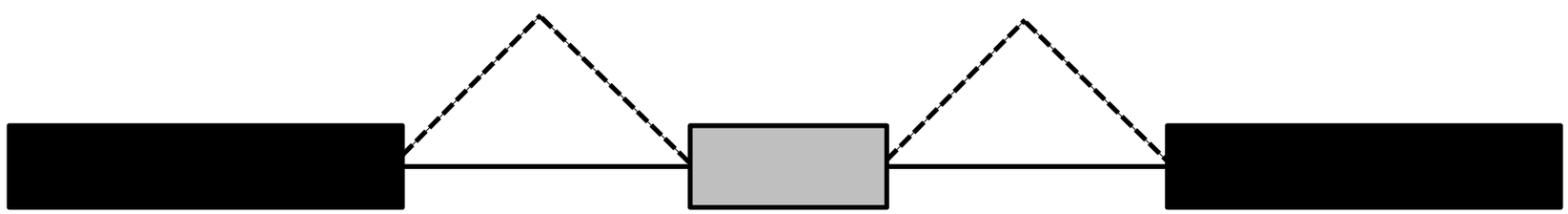
- A massively parallel approach to understanding sequence-function relationship: 5'alternative splicing
- Cell-type specific effects in alternative splicing
- **Skipped exons: attempt 1**
- Skipped exons and 3' alternative splicing: exon definition

Alternative Splicing



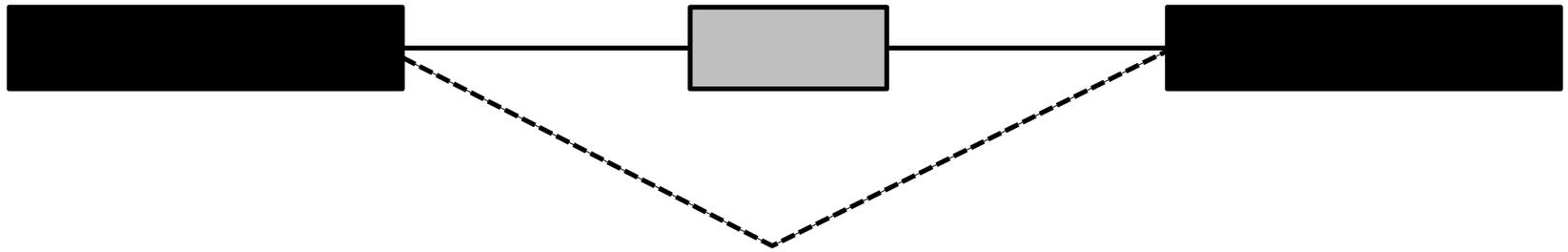
Bradley, R., et al. "Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution." *Plos Biol* 10 (2013): e1001229

Skipped exons



Skipped exons

- Exon skipping



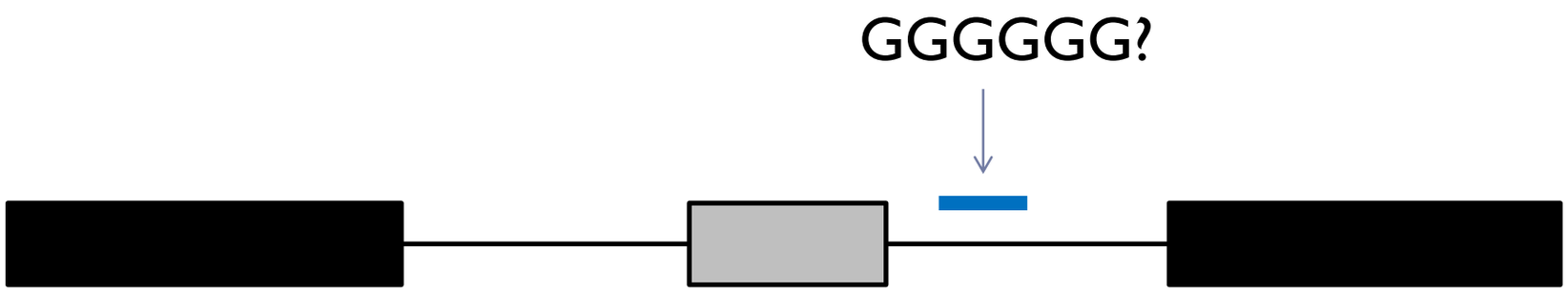
Skipped exons

Massively Parallel Exon Skipping Assay

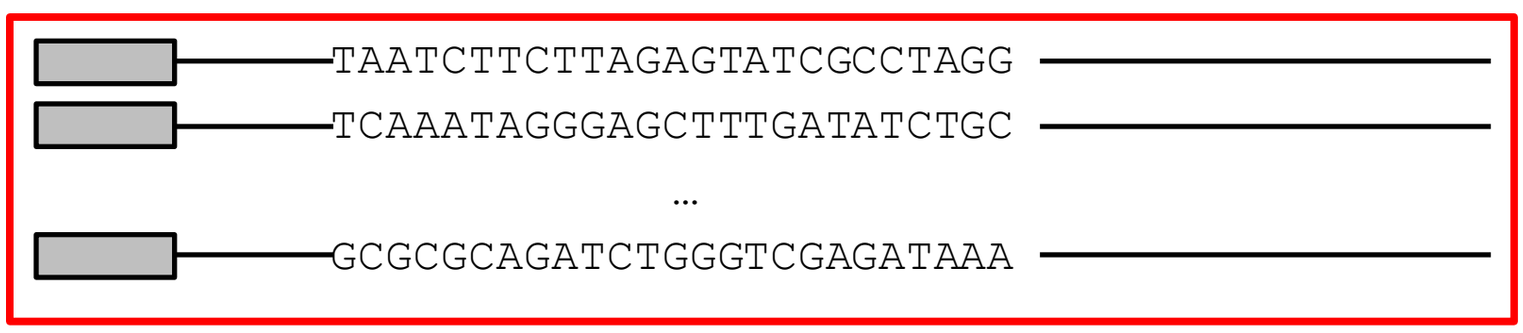
- Exon skipping minigene base on SMN1/2 exon7
- Randomized two intronic 25 nucleotides regions
- Tested ~1 million different sequences (for perspective: ~25,000 genes in the human genome)

SMN1/2 exon 7

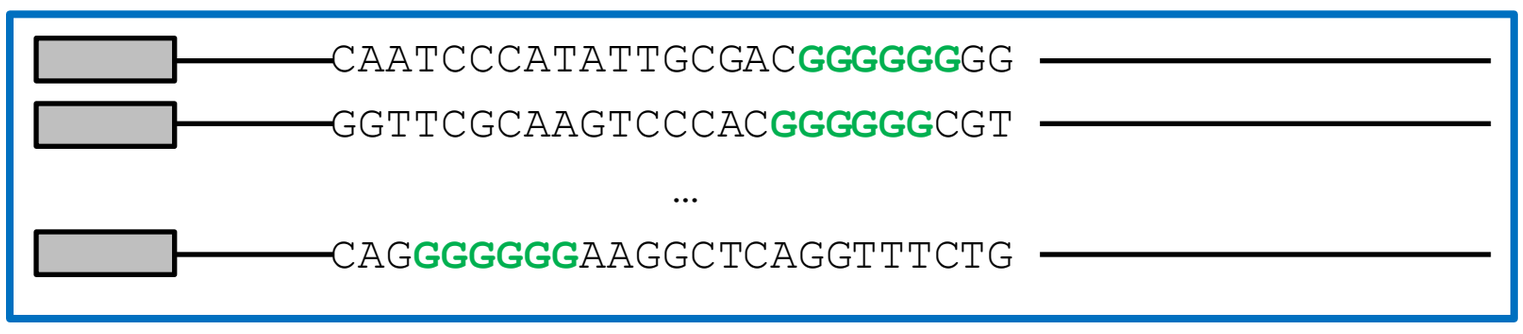
Short Sequence Effects



Introns without GGGGGG (N= 973,471)

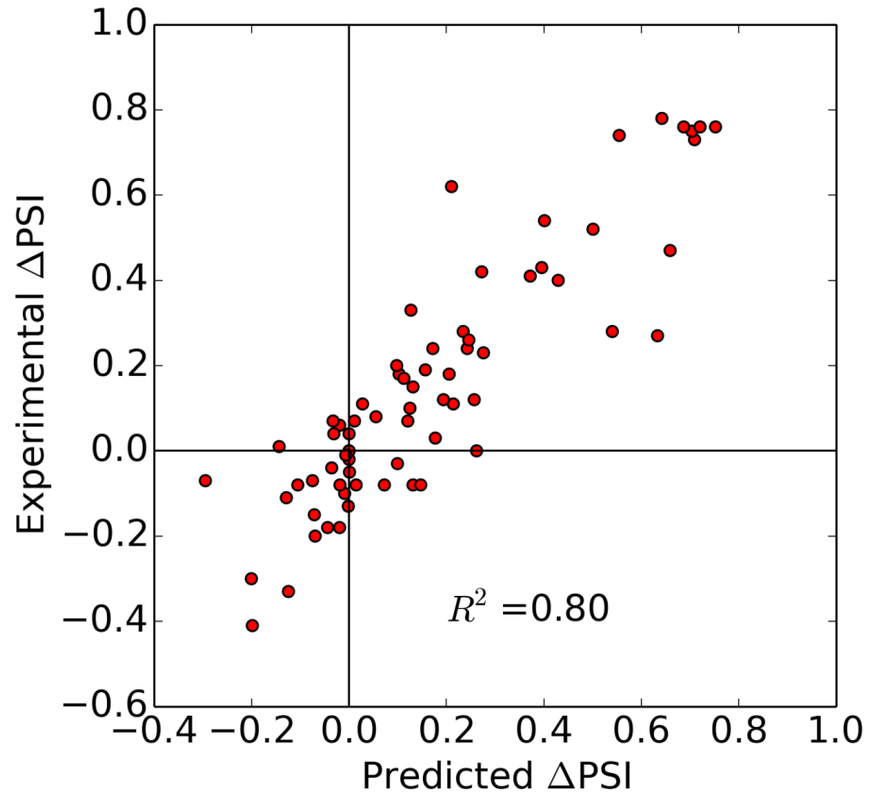


Introns with GGGGGG (N=2,087)



Effects of Genetic Variation on Alternative Splicing in Humans

Predicted Effects of SMN2 Mutations



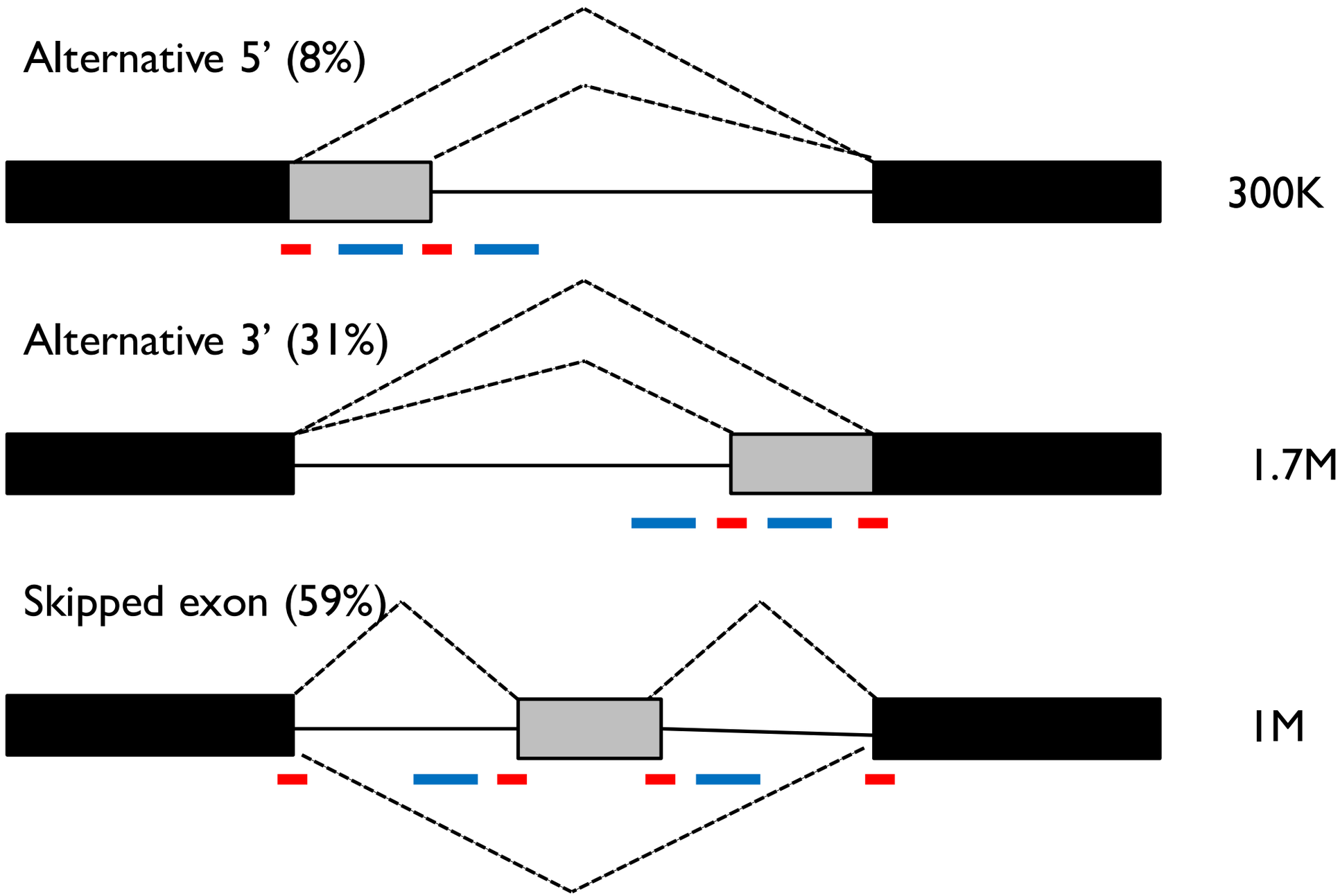
SMN1/2 exon 7

- Works only for intronic mutations
- And works only for SMN1/2

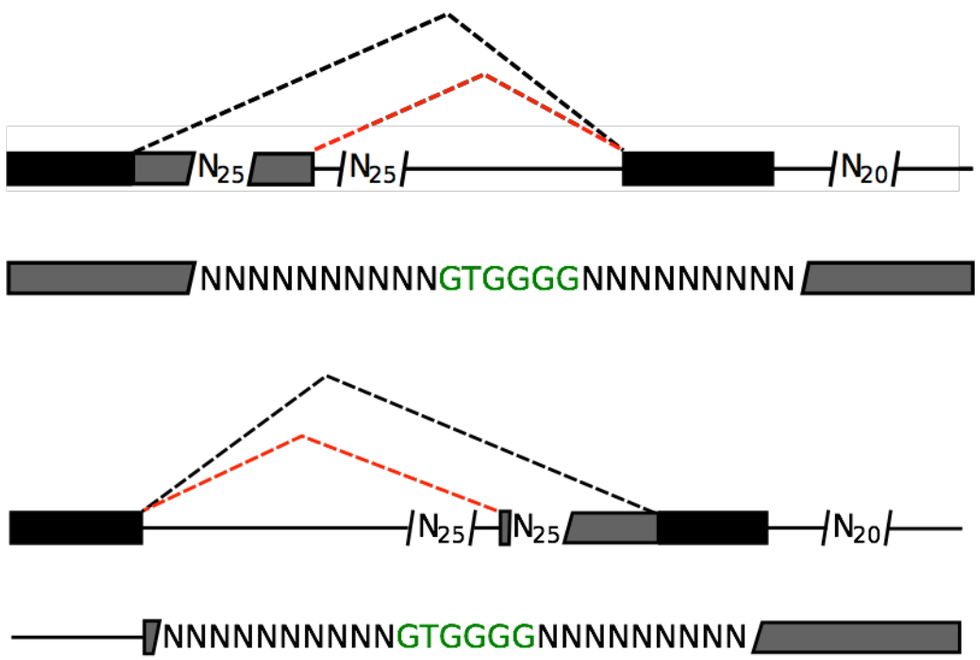
Overview

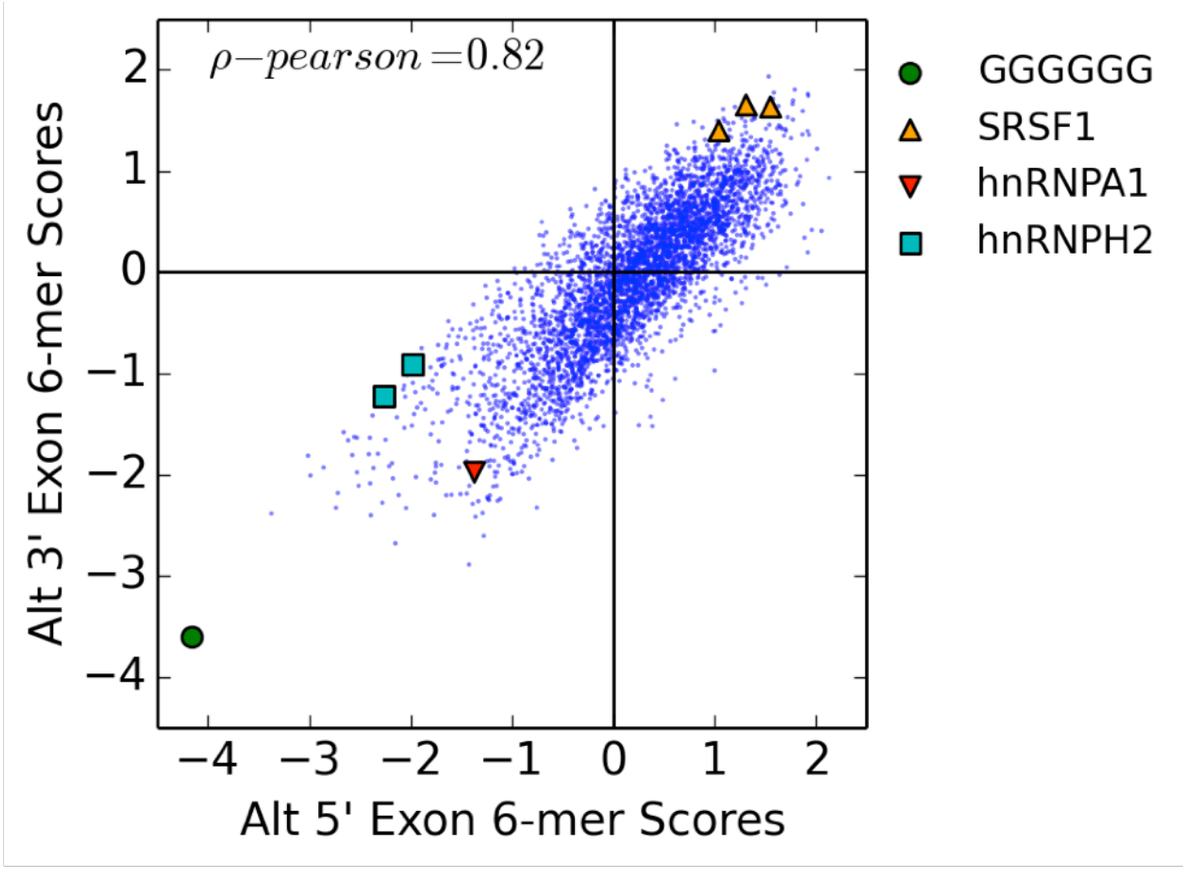
- A massively parallel approach to understanding sequence-function relationship: 5'alternative splicing
- Cell-type specific effects in alternative splicing
- Skipped exons: attempt 1
- **Skipped exons and 3' alternative splicing: exon definition**

Alternative Splicing Libraries



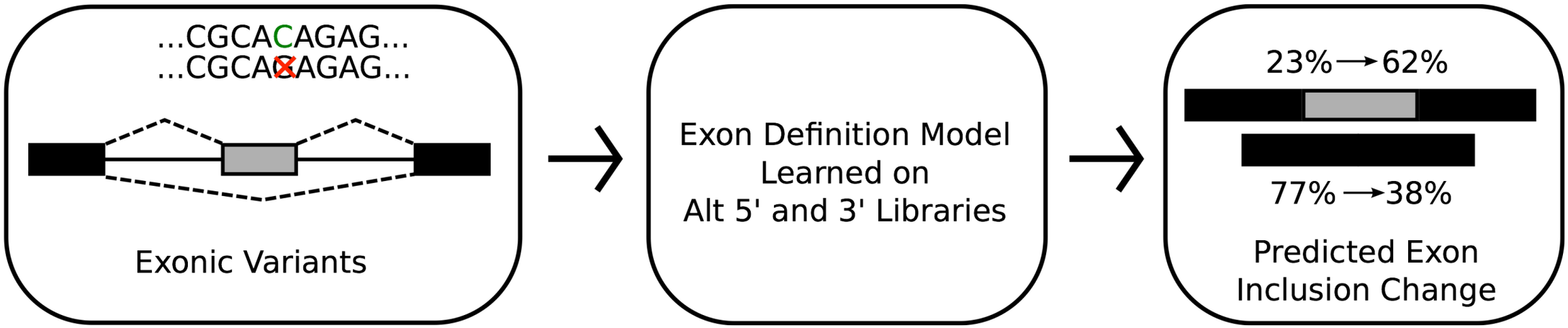
Nearly identical exon definition in 3’ and 5’ alternative splicing

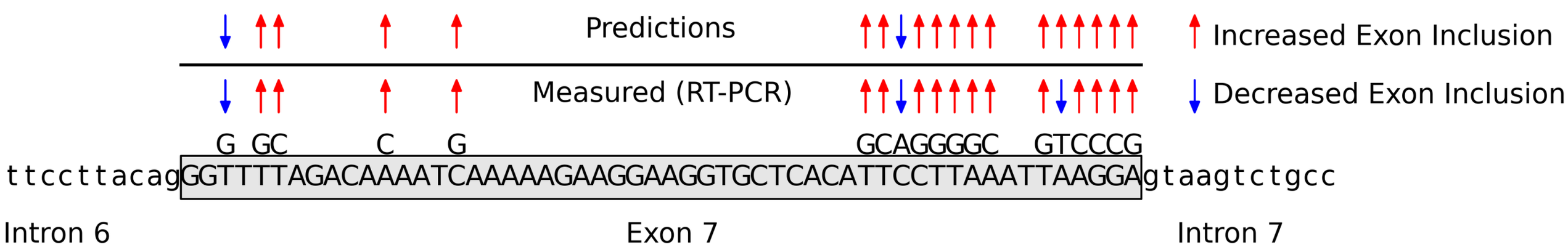




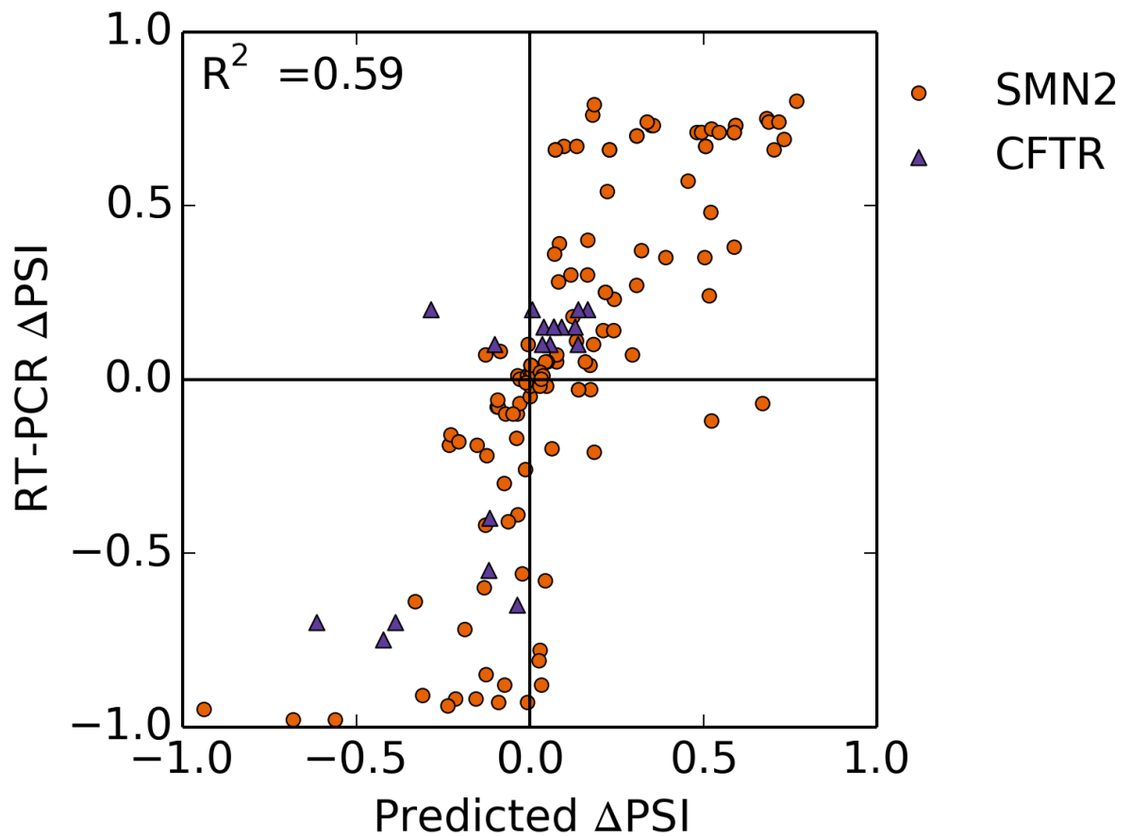
~1.7 million 3’alternative splice events

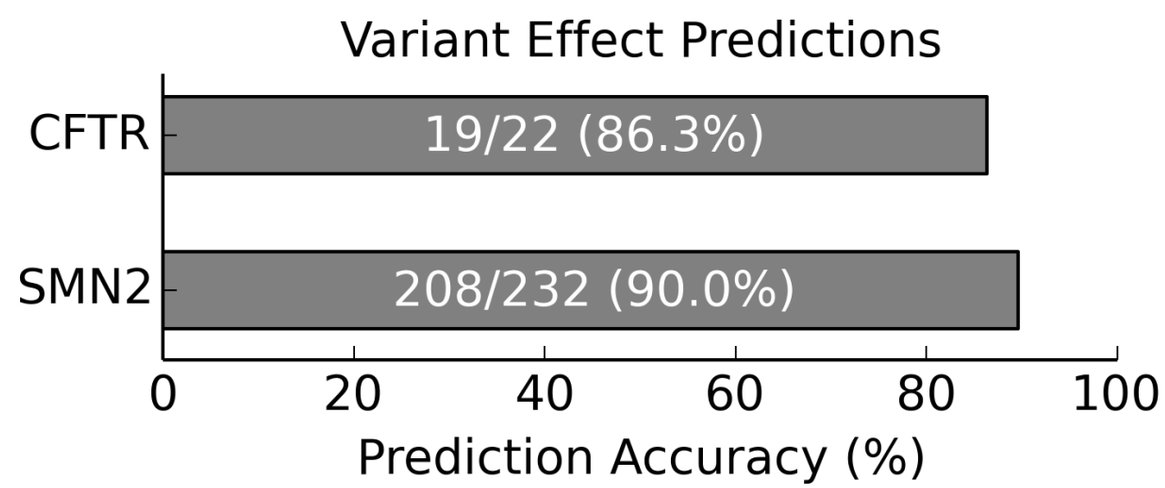
Predicting the Effects of Mutations in Skipped Exons



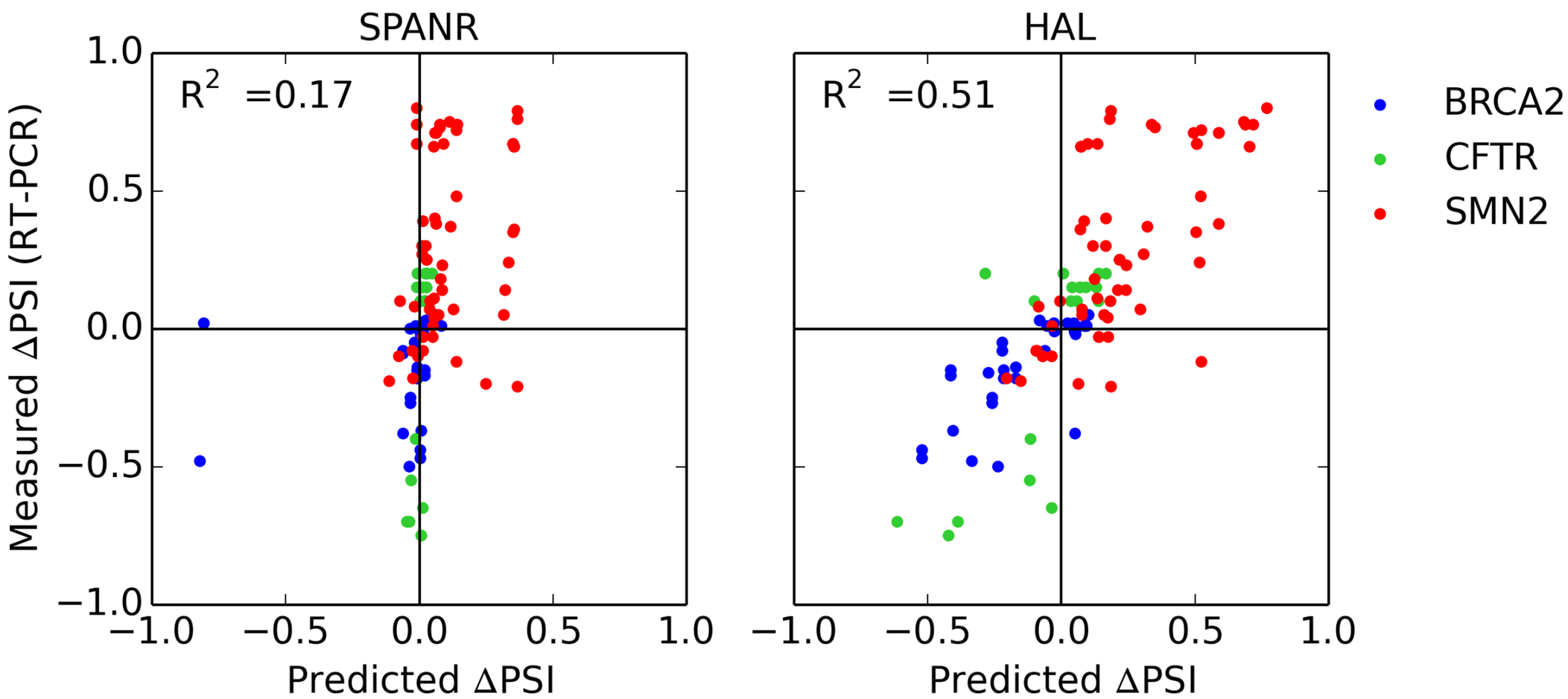


Predicting the Effects of Mutations in SMN and CFTR proteins





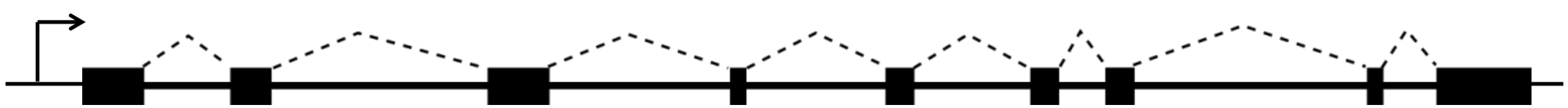
Nearly identical exon definition in 3' and 5' alternative splicing



SPANR: Ailpanahi et al., Science (2015)

Exon definition

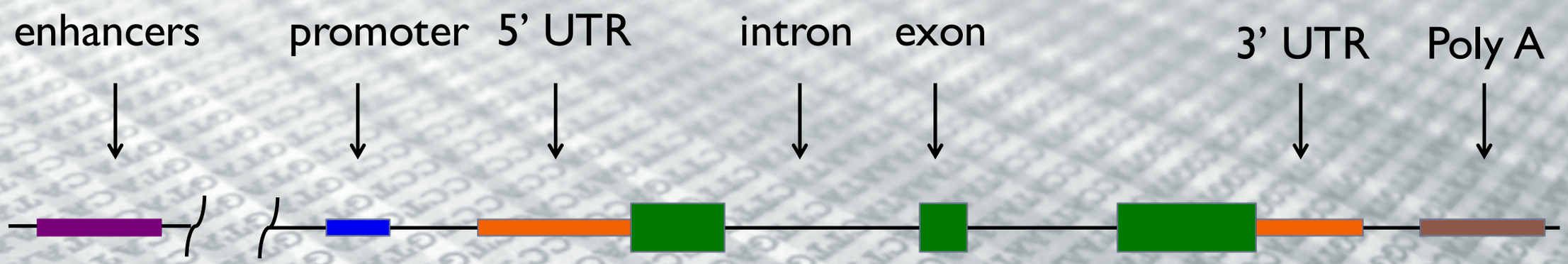
- Human exons are short: typically 50-250 bp
- Human introns are long: often 10^5 bp
- Splice sites are recognized in pairs across exons



Summary

- We presented a new approach to learn the regulatory rules governing alternative splice site selection
- A model that was trained only on synthetic data predicts splice site selection better than any previous model directly trained on the genome
- A model that was not trained on skipped exon can predict the effect of mutations in skipped exons
- Our approach makes it possible to identify cell-types specific differences in splicing

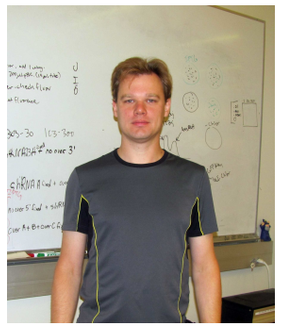
A broadly applicable method for understanding gene regulation



Transcription
Alternative Splicing
Translation
Poly-adenylation
…

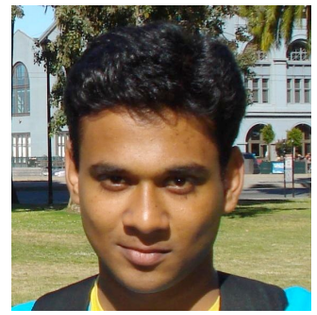
Acknowledgements

Yuan-Jyue Chen



Sergii Pochekailov

Ben Groves



Gourab Chatterjee

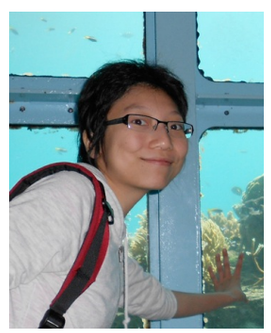
Rebecca Black

Alex Rosenberg

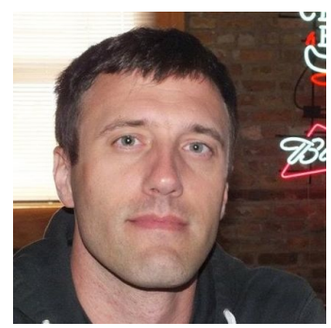
Paul Sample

Alex Baryshev

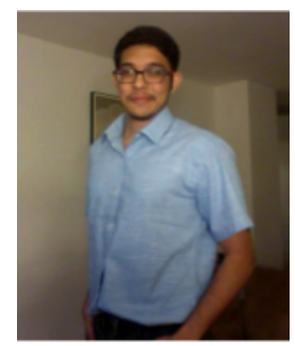
Sumit Mukherjee



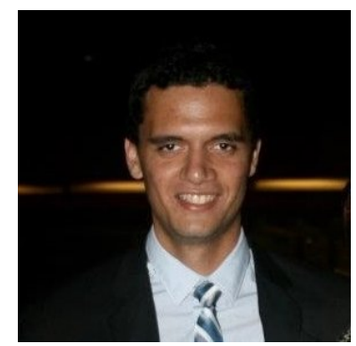
Sifang Chen



Nick Bogard



Arjun Khakhar



Randolph Lopez

BURROUGHS WELLCOME FUND

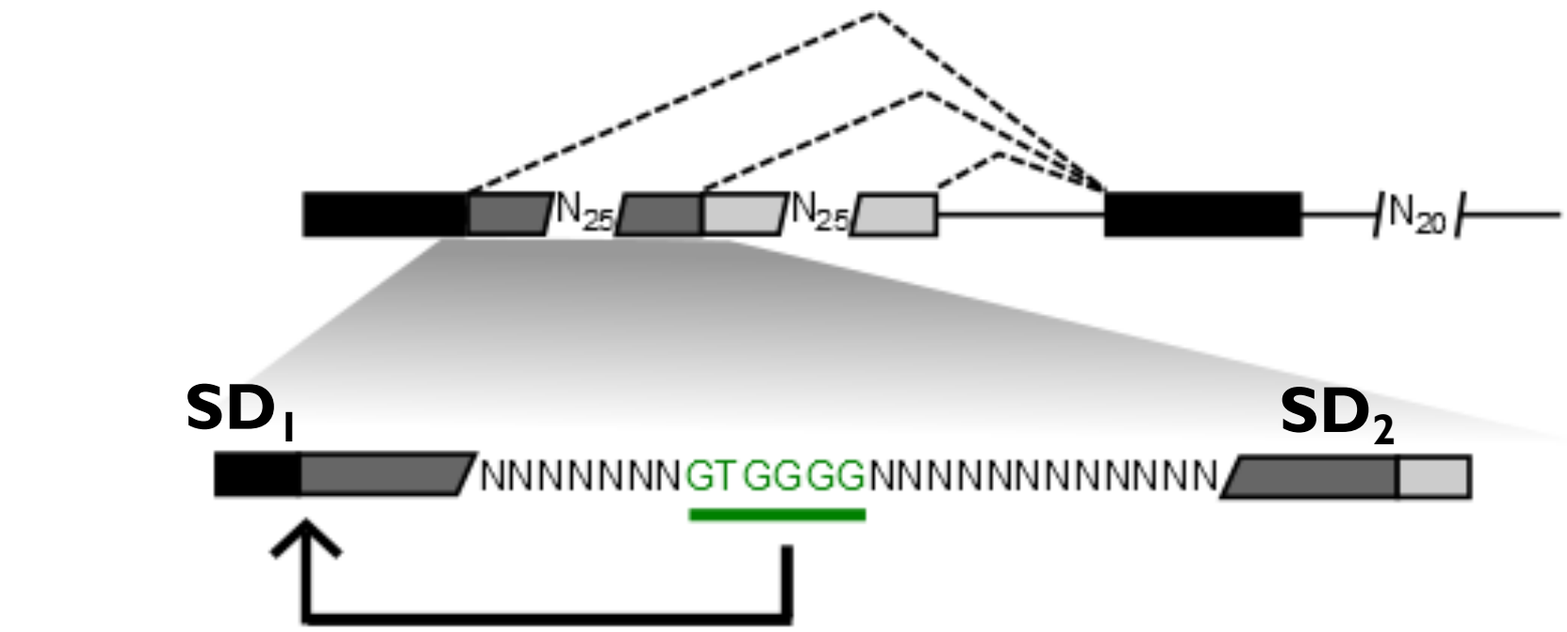
NSF

ALFRED P. SLOAN FOUNDATION 1934

DEPARTMENT OF THE NAVY ONR Science & Technology

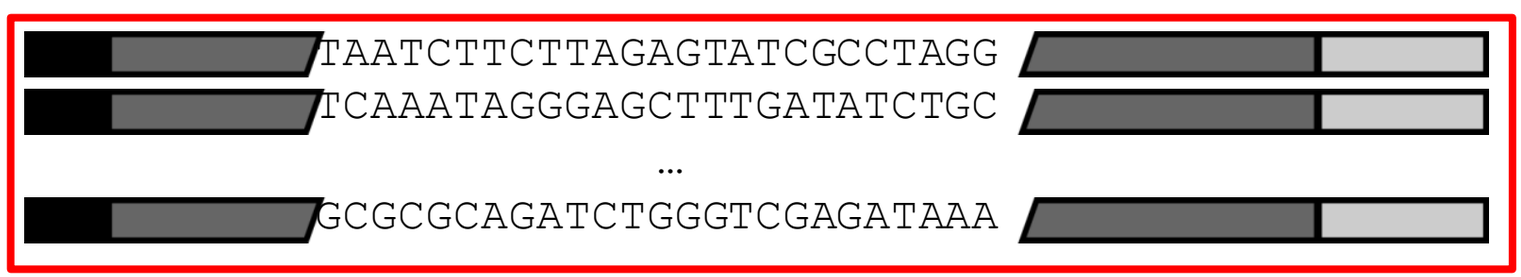
DARPA

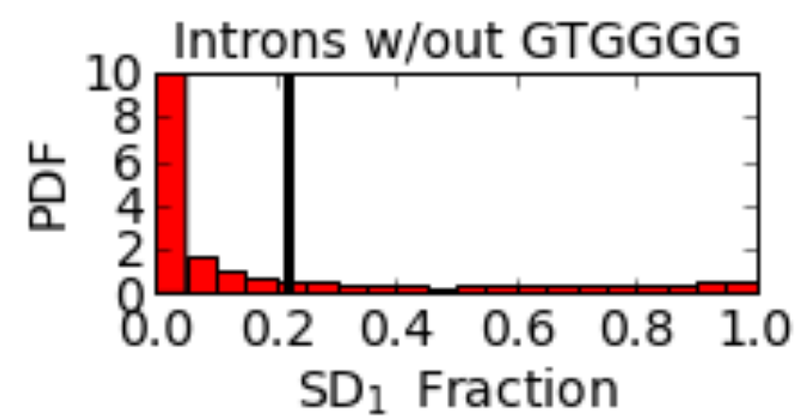
Short Sequence Motif Effect Sizes



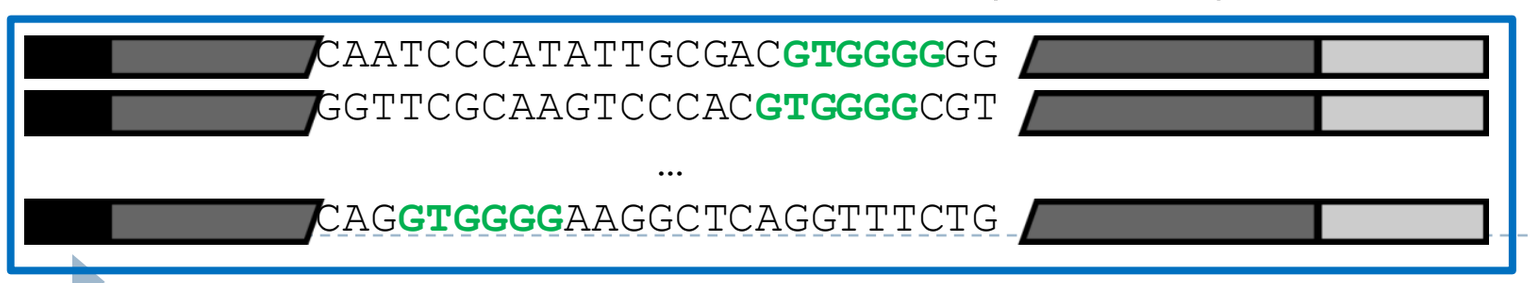
Effect Size:
GTGGGG = +2.37

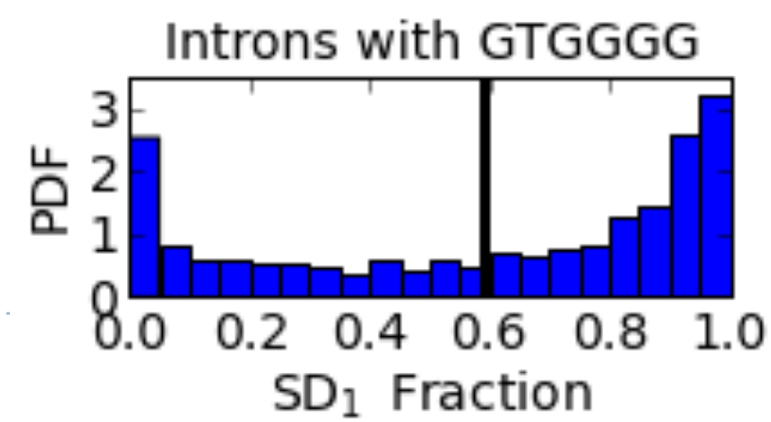
Introns without GTGGGG (N=264,000)





Introns with GTGGGG (N=3000)

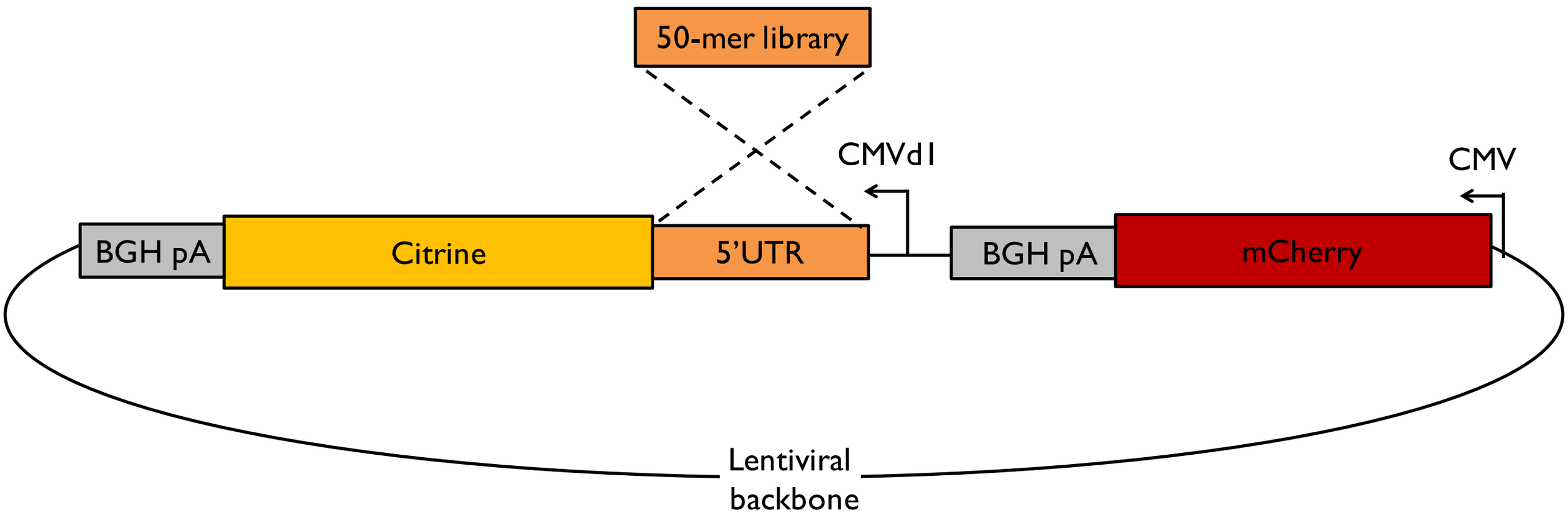




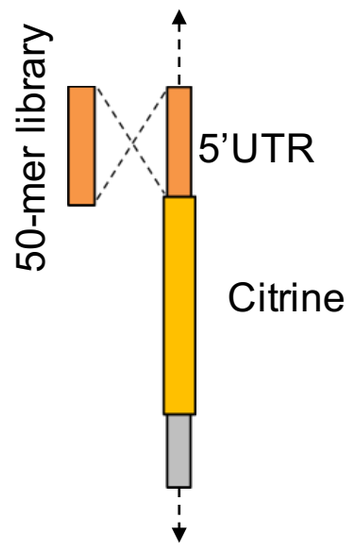
Predicting the Effects of Mutations in Survival Motor Neuron (SMN) protein

- Mutations in SMN proteins alter RNA splicing and cause spinal muscular atrophy (SMA)
- SMA can severely affect muscle control
- SMA affects between 1/6,000 to 1/10,000 people
- Can we predict which mutations will alter splicing of SMN proteins?

A massively parallel approach to studying translation

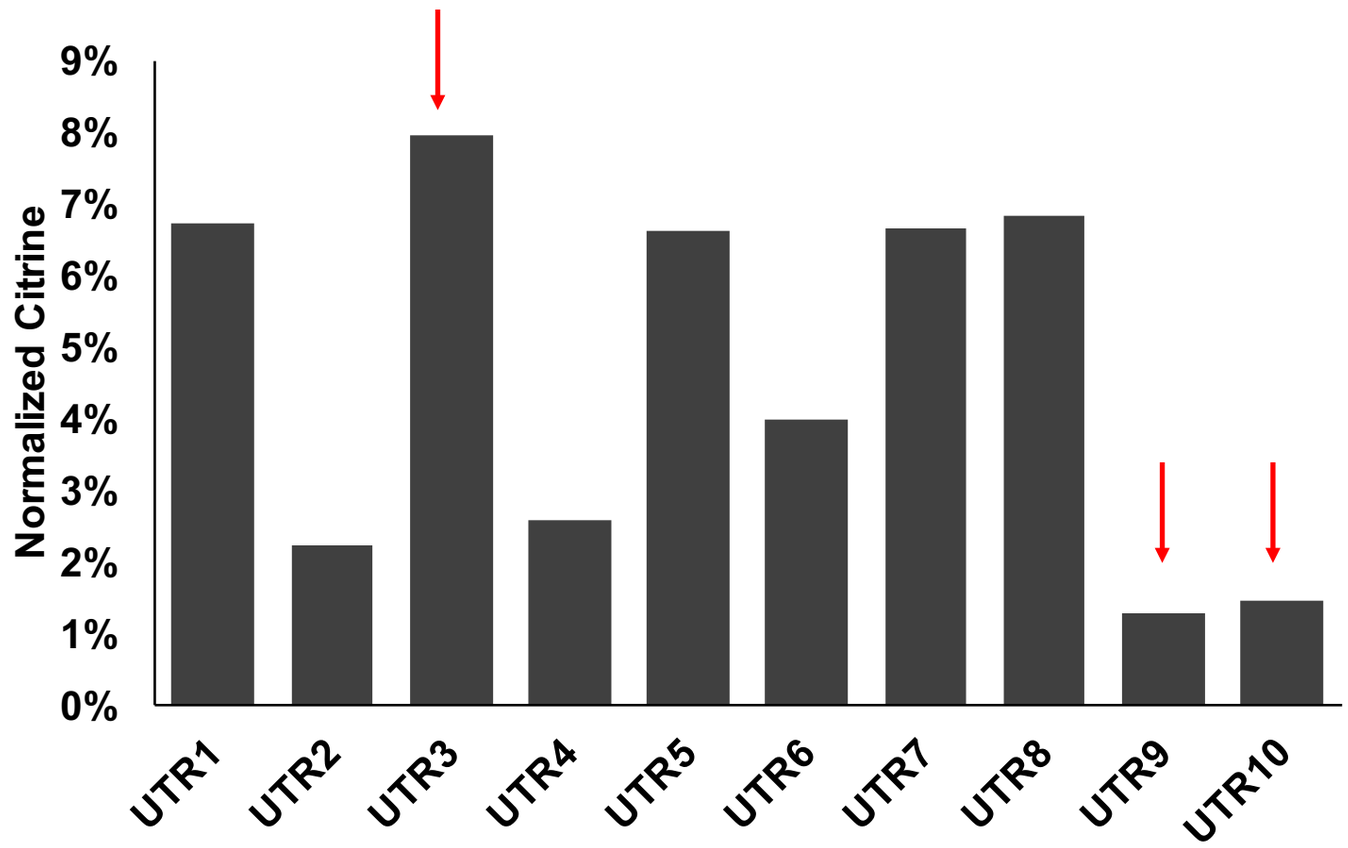


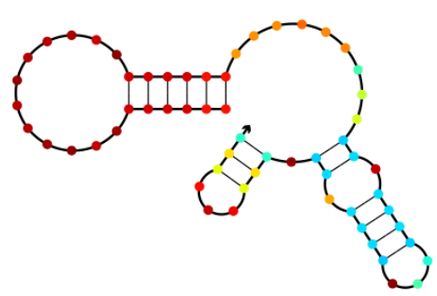
Work flow



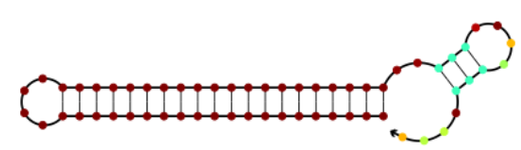
500k random
5'UTRs

Flow cytometry results for 7 random and 3 designed 5'UTRs

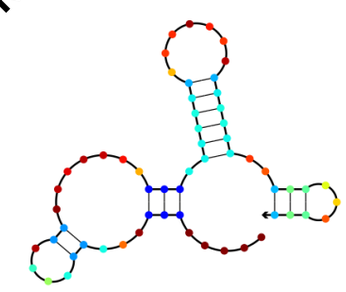




UTR3
- Purine at -3
- No uAUGs
- Moderate-low 2° structure



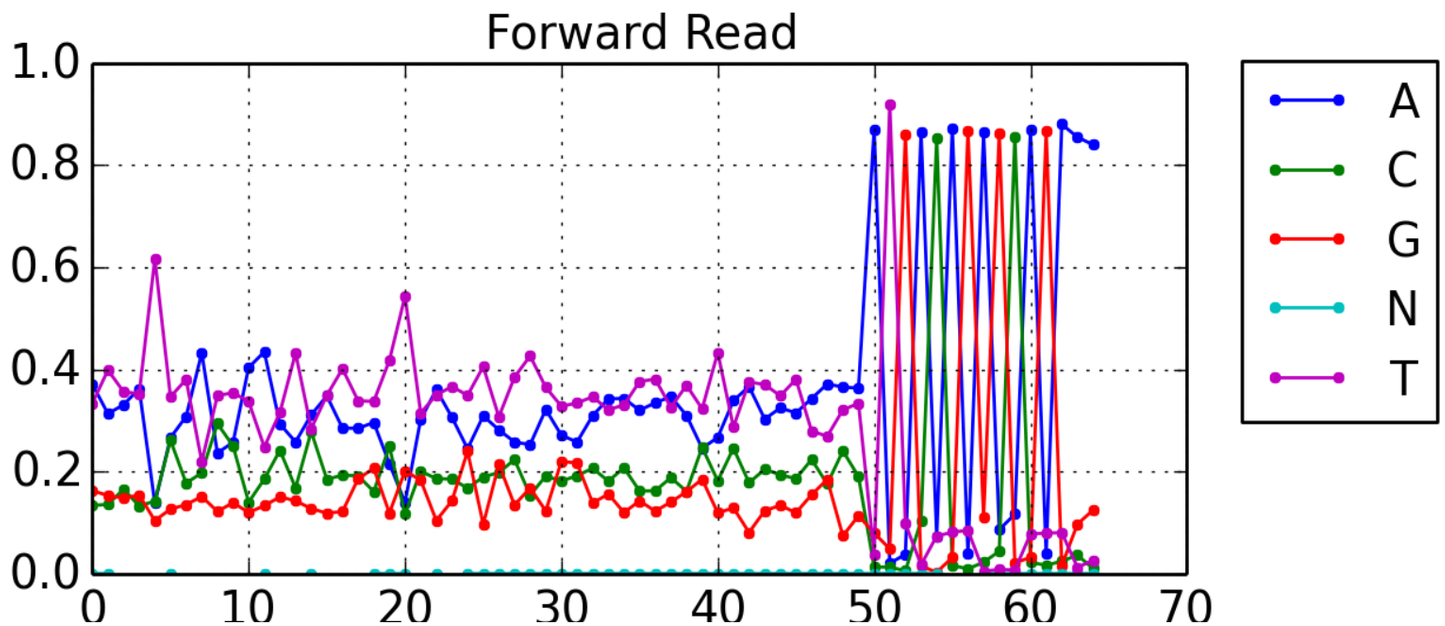
UTR9
- Purine at -3
- No uAUGs
- Very High 2° structure



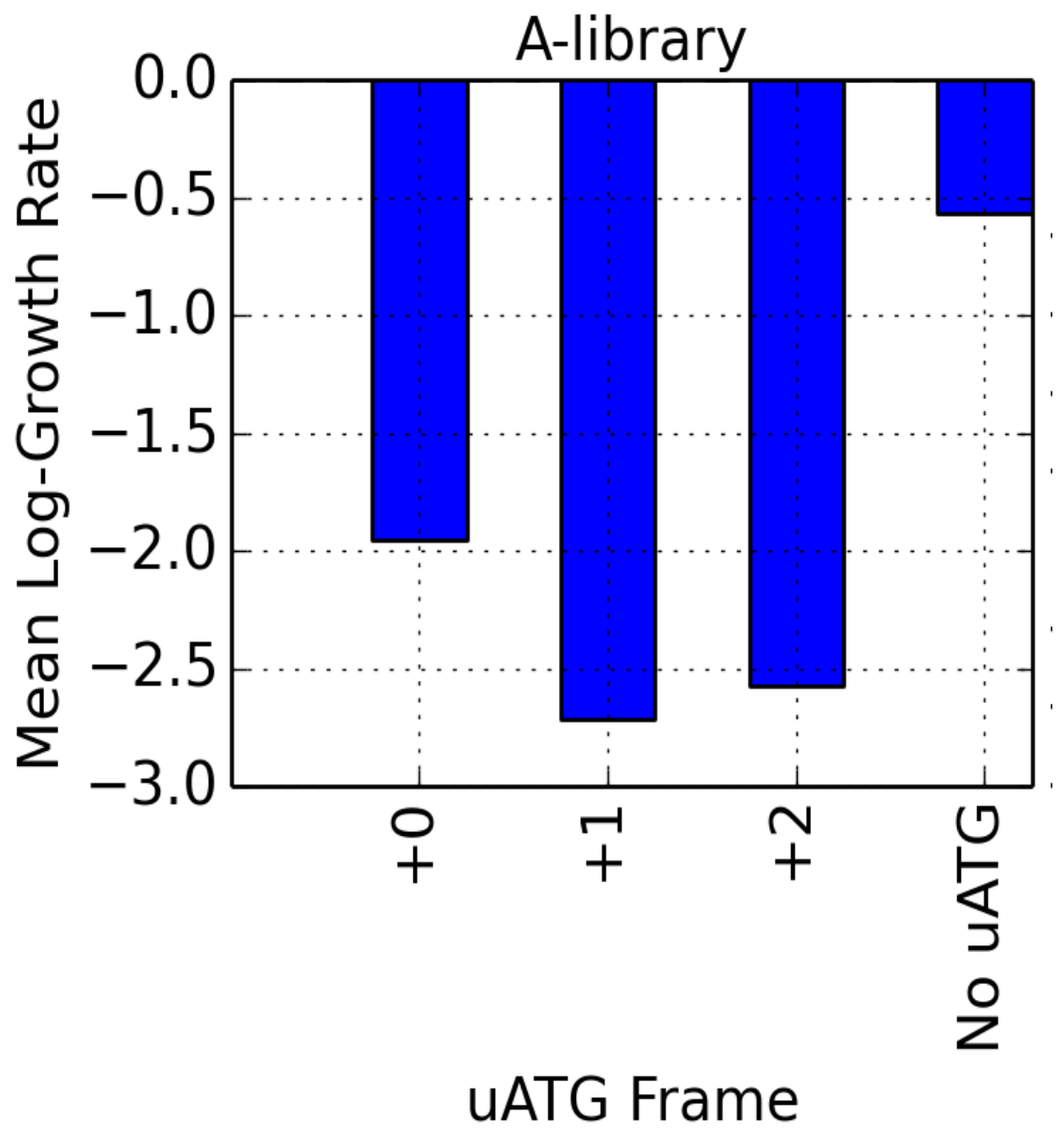
UTR10
- Purine at -3
- Two uAUGs
- Low 2° structure

W. L. Noderer, et al. "Quantitative analysis of mammalian translation initiation sites by FACS-seq." *Mol. Sys. Biol.* 10, 748 (2014).

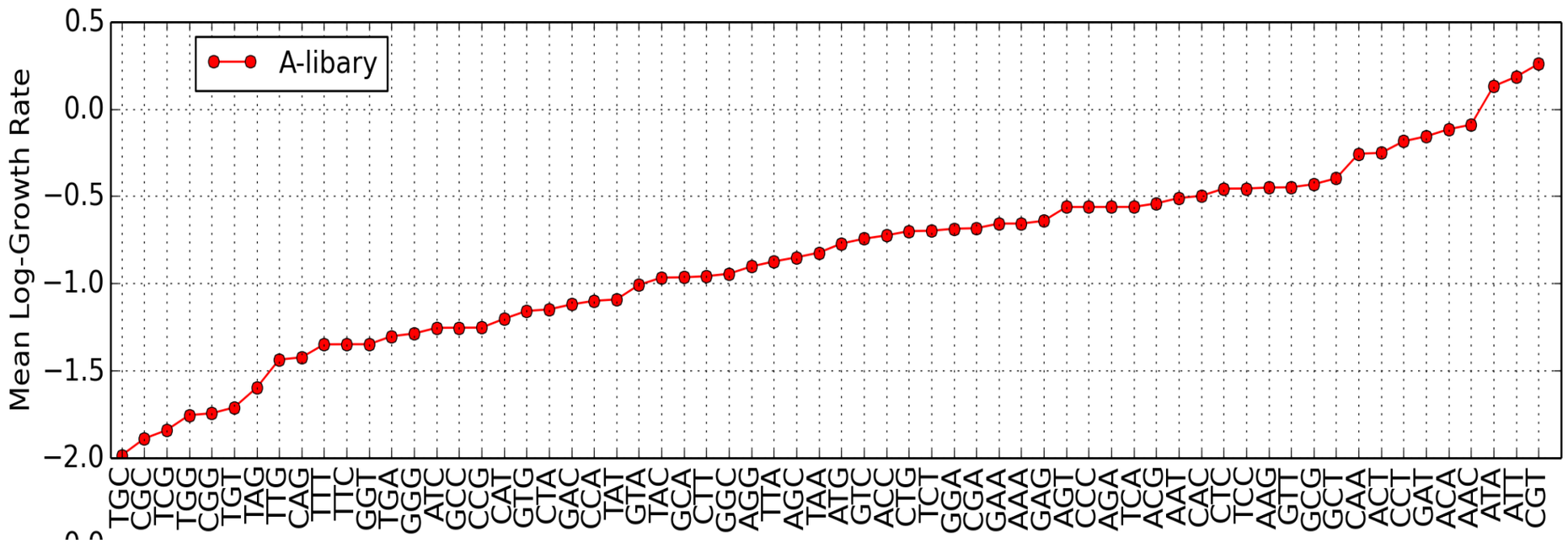
Sequencing confirms random 5'UTR



Upstream ATGs modulate translation



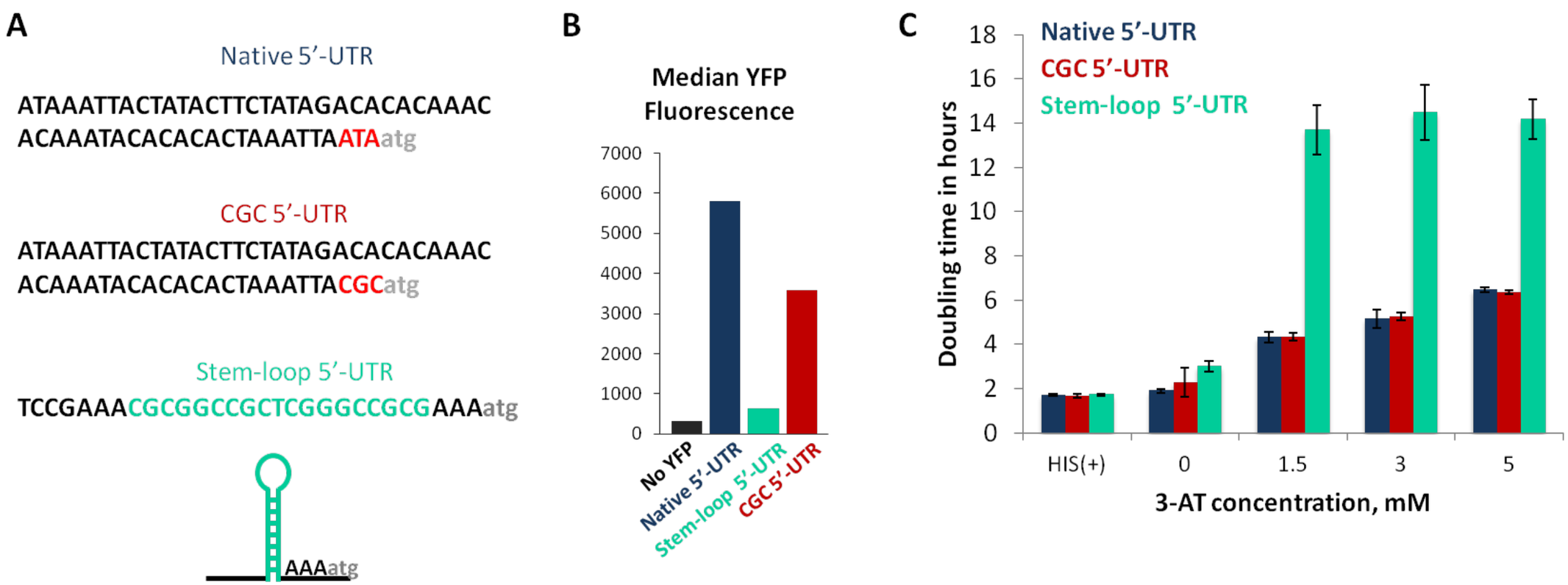
Nucleotides at -3:-1 strongly influence translation



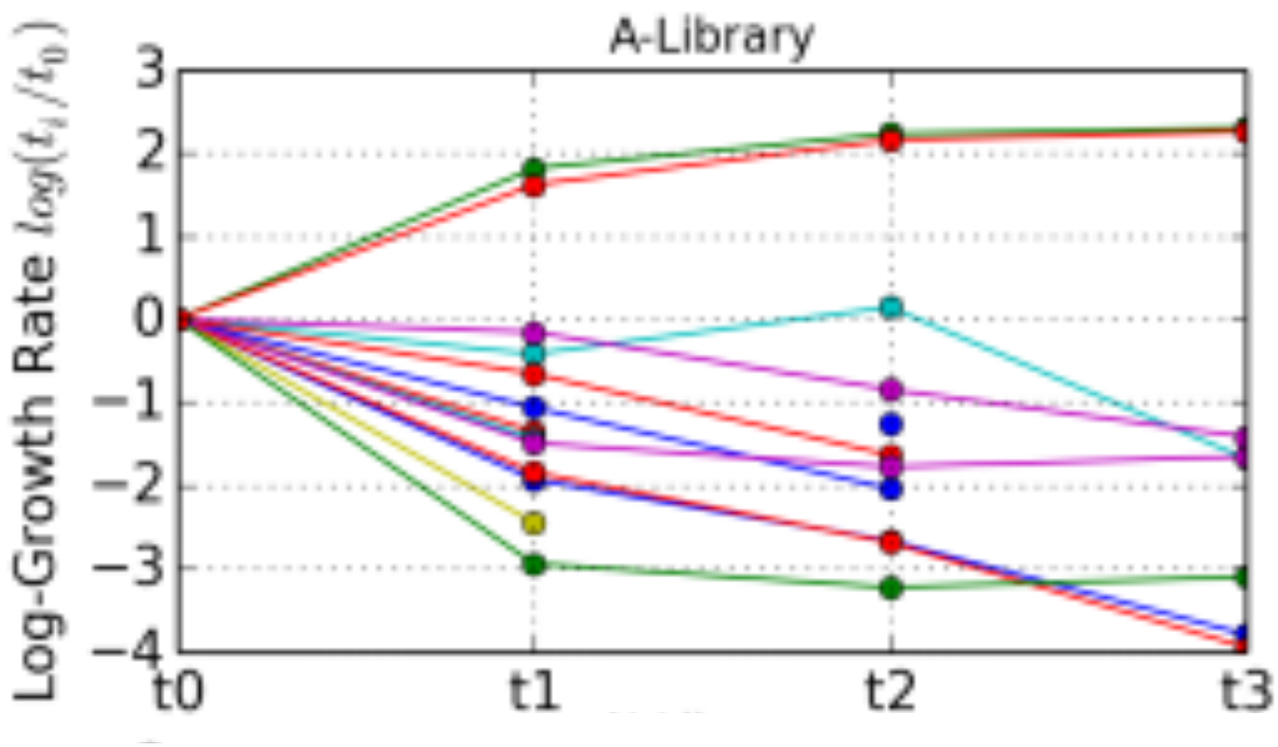
Translation Summary

- We are developing a massively parallel approach to understanding the 5'UTR sequence-function relationship
- Very large "super-biological" data sets enable predictive models
- This approach can in principle be applied in the context of your favorite gene and cell type

Flow cytometry results for 7 random and 3 designed 5'UTRs



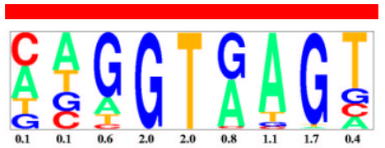
Example Growth Traces for a Few Library Members



Regulation of Alternative Splicing

What are the sequence determinants of alternative splicing?

- The splice site sequences

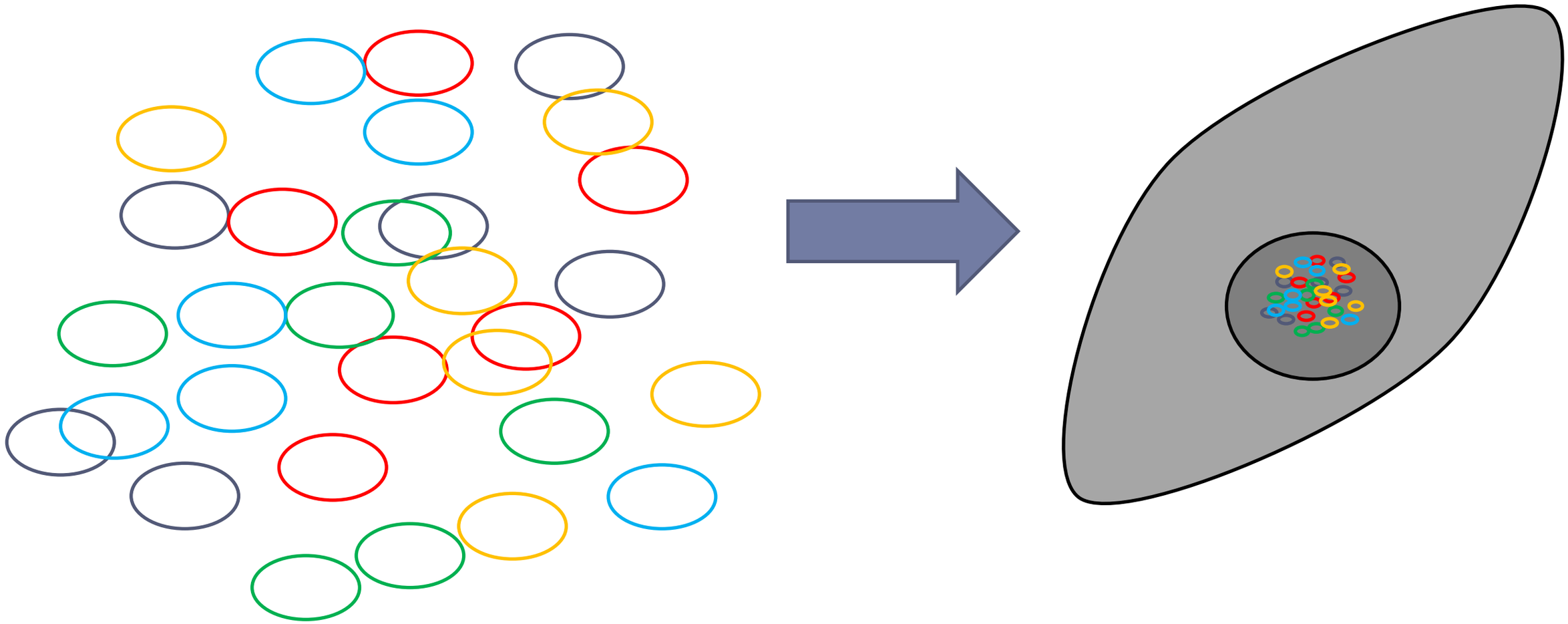


- Sequences in the introns

Experimental Methods

DNA synthesized in the lab

Human Cells



Resulting Data

	0	26
	113	4
	1	12
	⋮	⋮

~1 million
Different
Sequences

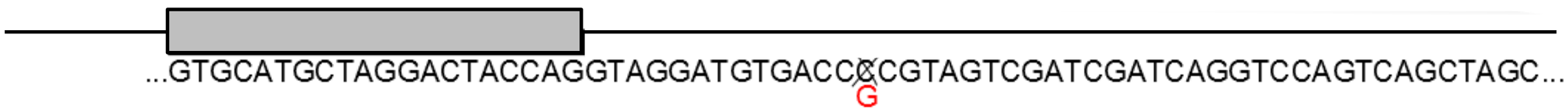
Predicting the Effects of Mutations in Survival Motor Neuron (SMN) protein

- Mutations in SMN proteins alter RNA splicing and cause spinal muscular atrophy (SMA)
- SMA can severely affect muscle control
- SMA affects between 1/6,000 to 1/10,000 people
- Can we predict which mutations will alter splicing of SMN proteins?

Definition: Percent Spliced In

- *Percent Spliced In (PSI)=mRNA↓A /mRNA↓A +mRNA↓B*

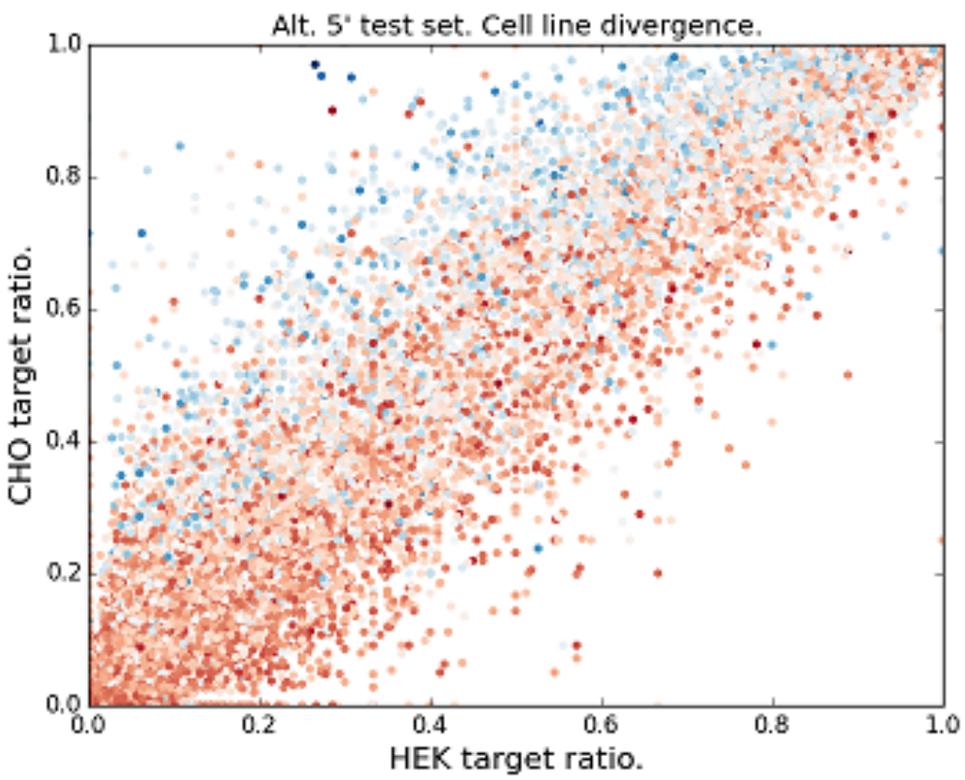
Dataset: Mutations Tested in Studies on SMN2 Splicing



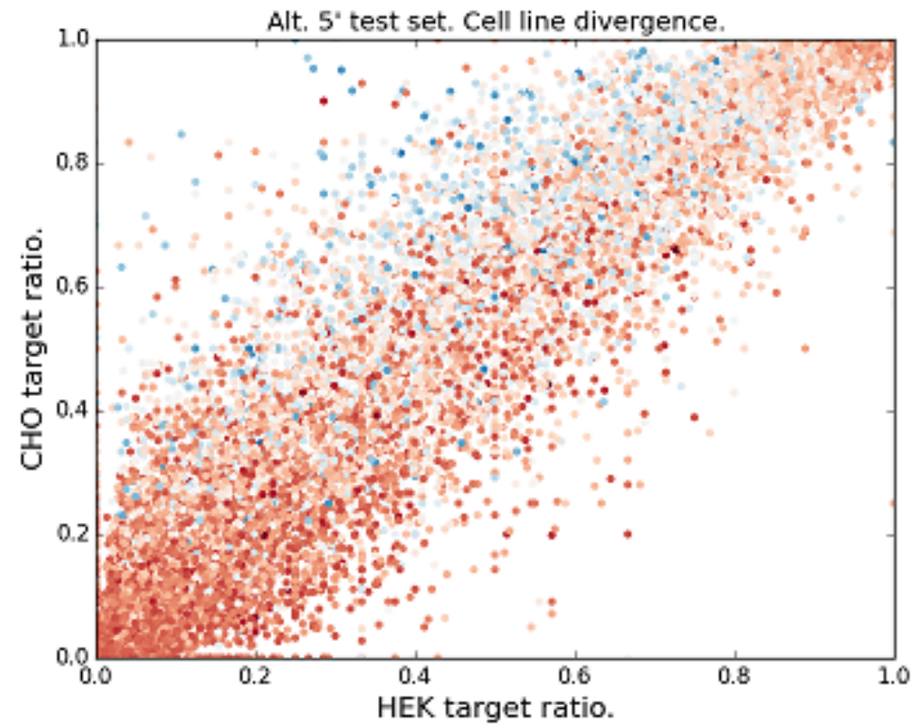
Position	Mutation	ΔPSI
3	C>G	+21.2%
5	A>T	-20.8%
12	G>A	+3.3%
…		
50	A>C	+65.2%

Uncovering cell type specific splicing

Logistic regression: R^2=0.14



Logistic regression: R^2=0.16



Ray, Debashish, et al. "A compendium of RNA-binding motifs for decoding gene regulation." *Nature* 499.7457 (2013): 172-177.