The human splicing code reveals new insights into the genetic determinants of disease

Christine Tataru, Nate Stockham,
Greg McInnes, Kelley Paskov

October 24, 2016

Outline

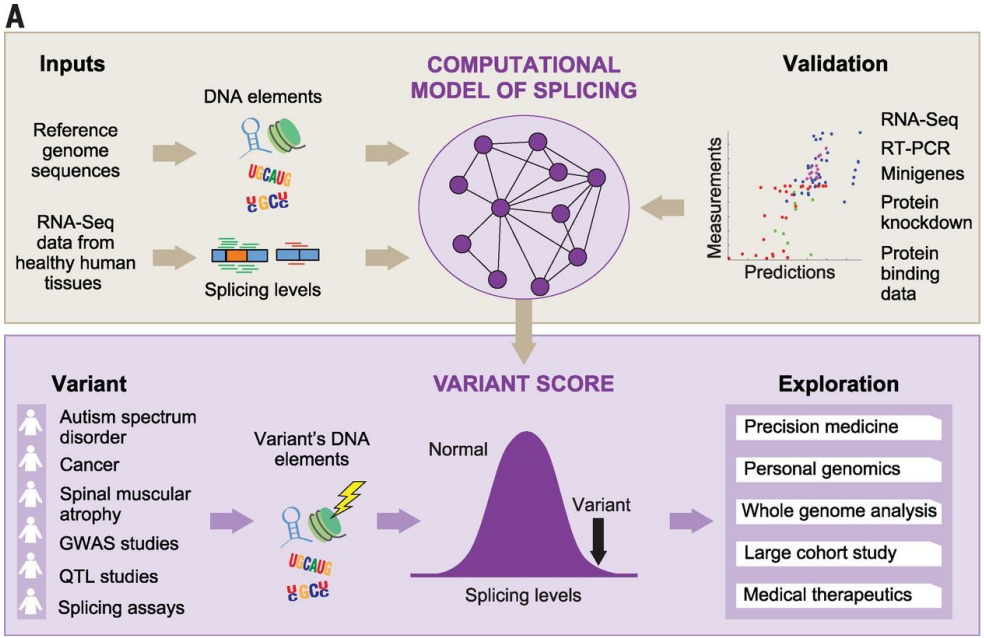
- Introduction
  - What did they do?
  - What is splicing?
  - Why do we care?
- The model
  - What features they use
  - How it compares to what else might be out there
- Analysis
  - General validation
  - Analysis of GWAS studies
- Clinical examples

Introduction

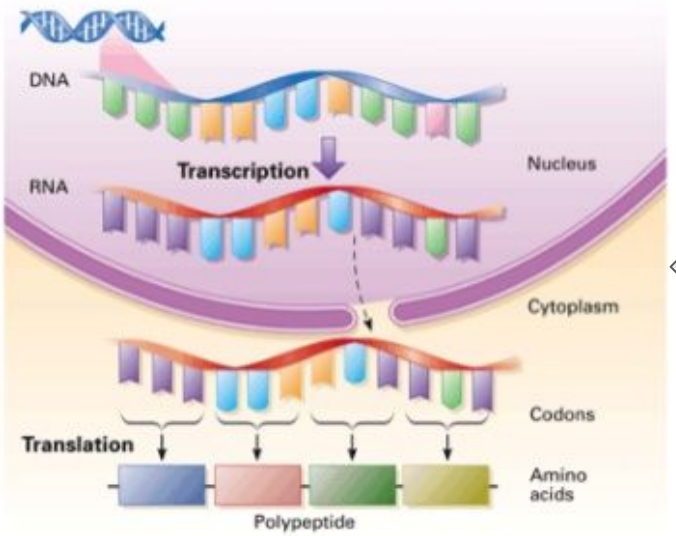
SPANR(splicing based analysis of variants) is a computational technique that scores how strongly genetic variations affect splicing.

Training

Predicting



The Central Dogma of

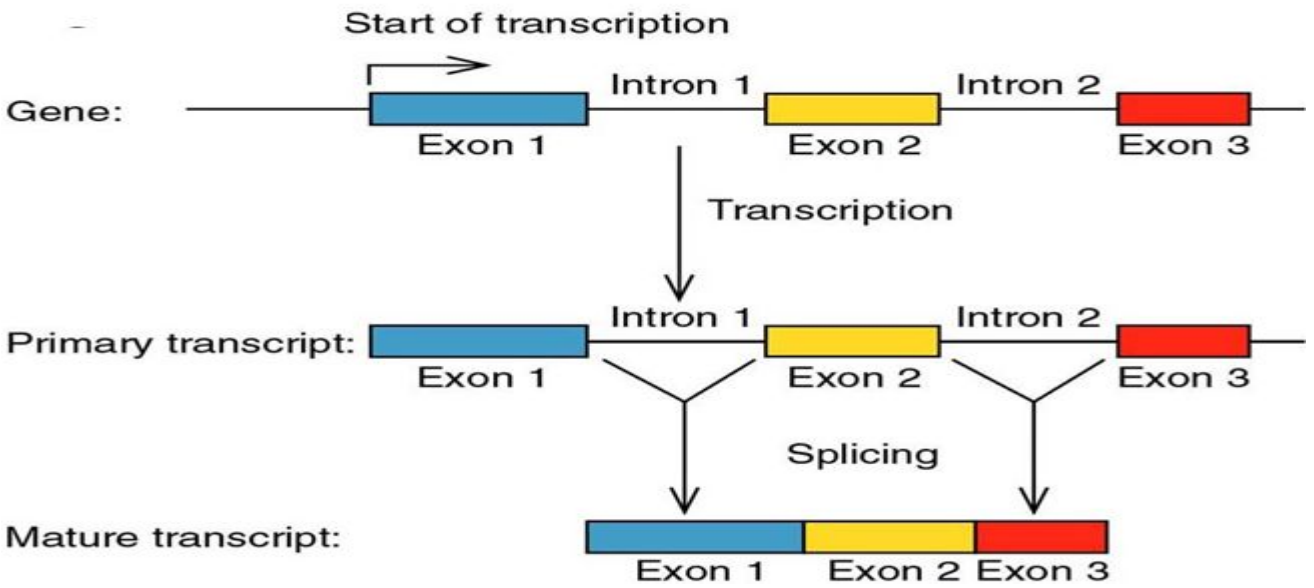


⇦ Splicing

Molecular Biology

Splicing Overview

2. Splicing increases the coding potential of the genome through alternative splicing.



Why splicing

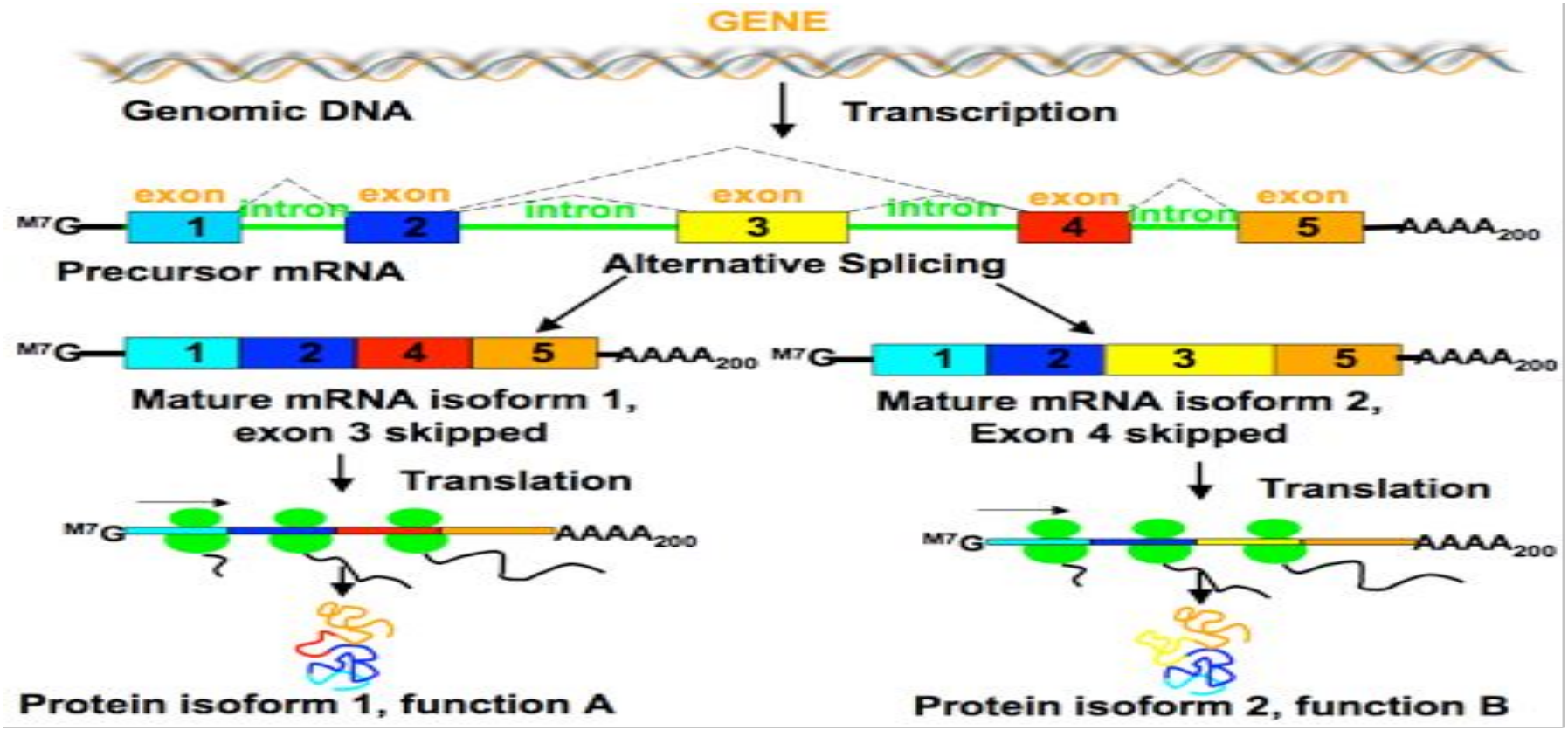
Three main arguments for selection of pre-mRNA splicing:

1. Domain evolution/increase protein functionalities
2. Additional layer of gene expression regulation
3. Alternative Splicing increase expression diversity

20,000-25,000 human protein-coding genes

60% of transcripts in human are spliced in different ways.

Alternative splicing



How sequence elements affect splicing

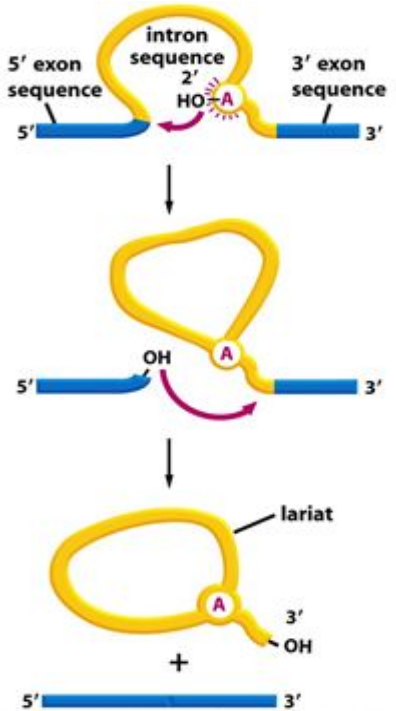


Figure 6-26a Molecular Biology of the Cell 5/e (© Garland Science 2008)

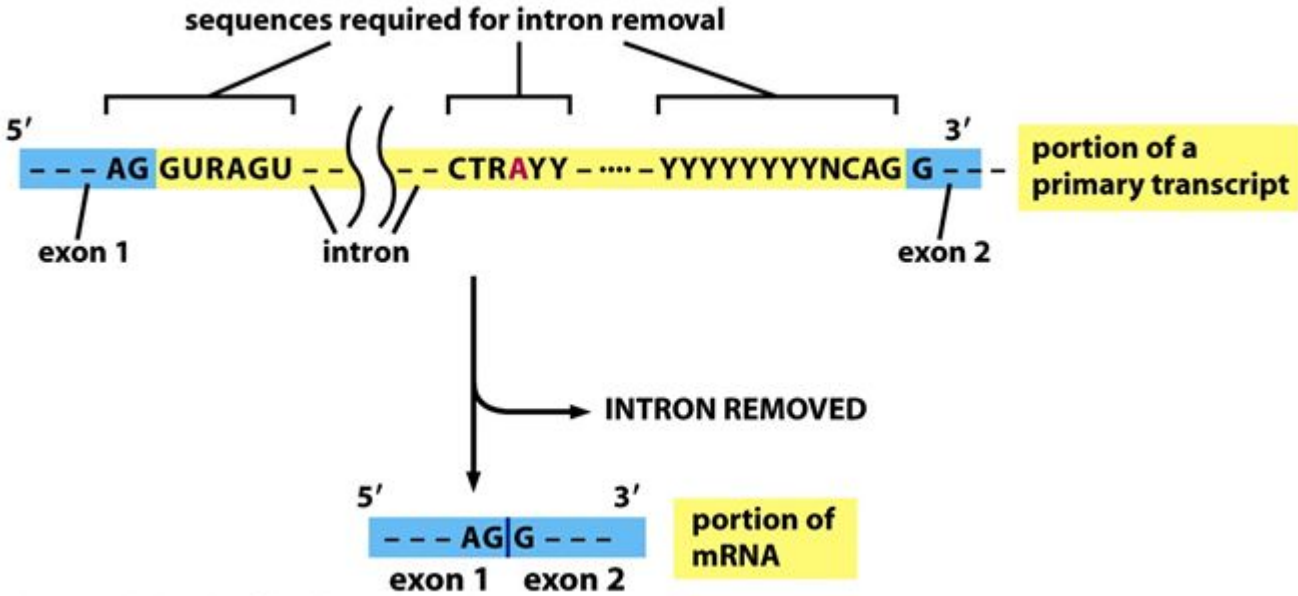
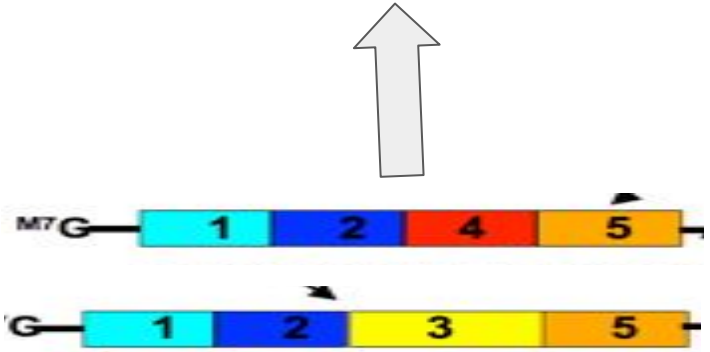


Figure 6-28 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Why do we care: Sequence to disease vs. Sequence to splicing

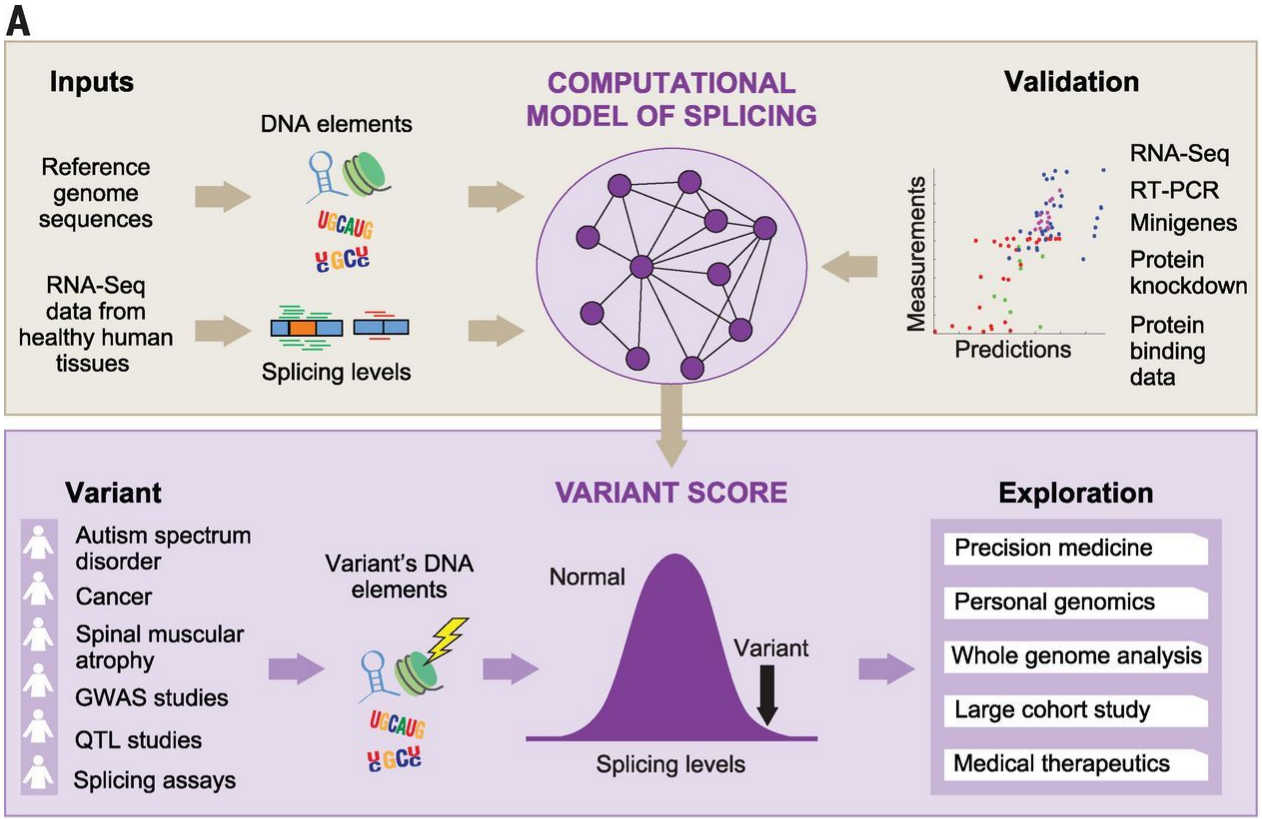
Hard!

Easier!



Model

Predict RNA splicing levels from DNA sequence



In our case

- Input: DNA sequence data with 3 annotated exons
- Output: % of transcripts that have center exon spliced in

We need methods to...

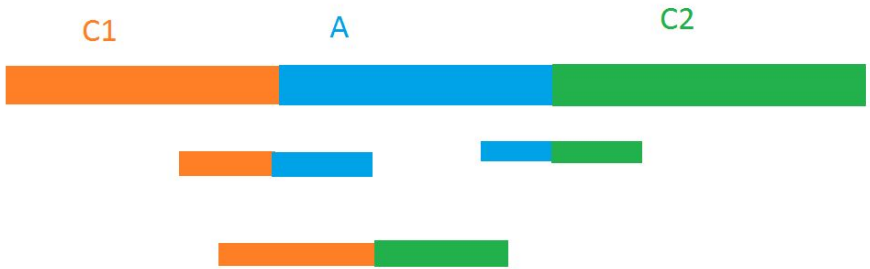
1. Construct inputs
2. Determine expected output
3. Calculate error

Input: DNA Features

1. Lengths of exons and introns
2. Does central exon introduce frameshift
3. Strength of intron acceptor/donor sites
4. 4 nucleosome positioning features
5. Alu related features to account for Alu repeats
6. 350 DNA binding protein binding motif features
7. Translatability features: 1 for a given splicing if exons of that splicing can be translated with no stop codons in at least 1 of 3 reading frames
8. Is sequence TTG present in intron 1

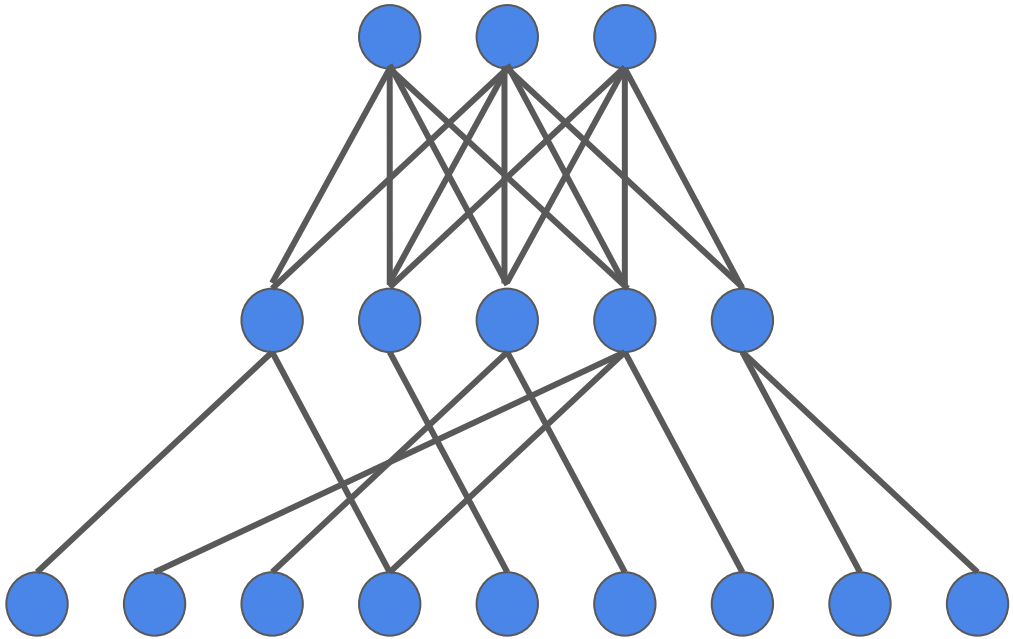
1,358 features in total

Output : Ψ : % of transcripts with exon spliced



Model definition

- Ensemble of 2 layer neural net that shares hidden nodes across tissues: max 30 hidden variables with sigmoidal non-linearities and softmax output
- Each tissue NN trained jointly as separate output units
- Bayesian Markov chain Monte Carlo (MCMC) to avoid overfitting and search billions of potential models with different structure and parameter values
- Maximize amount of information provided by the predictions of the model beyond a naïve guesser:
- 41,820 input to hidden parameters, 960 hidden to output parameter



Tissue (softmax)

Hidden layer

Inputs

Objective Function

$$CQ = \sum_e \sum_t D_{KL}(q_{t,e}|\hat{q}_t) - D_{KL}(q_{t,e}|p_{t,e}) \quad \text{where} \quad D_{KL}(q|p) = \sum_i p(i) \log \frac{p(i)}{q(i)}$$

- $q_{t,e}$ is target splicing pattern for exon e in tissue t
- q_t is the prediction of the optimal guesser that ignores the RNA features
- $p_{t,e}$is the prediction made by the regulatory model not trained on exon e
- D_{KL} is the Kullback-Leibler divergence between two distributions
- $D_{KL}(q_{t,e}|p_{t,e})$ can be interpreted as a likelihood function of predictions $p_{t,e}$ based on partial counts

Method Summary

- a tool that takes in local DNA sequence (3 exons necessary), and predicts, based on specific features of neighboring introns and exons, % of transcripts with central exon spliced in Ψ.

Training:

1. Identify exons
2. Extract features from local DNA seq
3. Detect alternative splicing with RNA seq
4. Run Bayesian inference to weight features in context dependent manner
5. Compare to RNA-seq determined splicing levels.

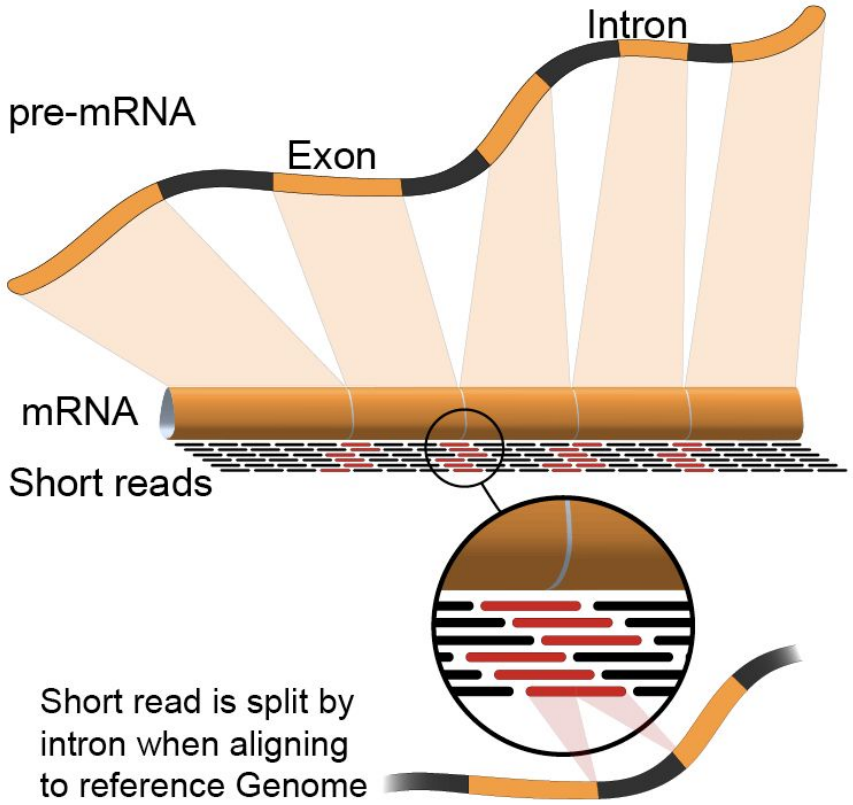
Testing:

1. Identify exons
2. Extract features from local DNA seq
3. Use weighted and dependent features to determine psi, % of transcripts that have center exon spliced in

Results

RNA-Seq

- Method of determining presence of mRNA within cells
- Relatively simple method of counting transcripts
- Also possible to determine
  - **Alternative splicing**
  - Gene fusions
  - Transcription mutations
  - Post-transcriptional modifications



RNA-seq Validation

The model was trained on RNA sequencing data

Data from the Illumina Body Map was used to determine Ψ for 16 tissue types and 10,689 exons that showed evidence of alternative splicing

Good agreement between RNA-seq data and predicted Ψ*:* $R^2 = 0.65$

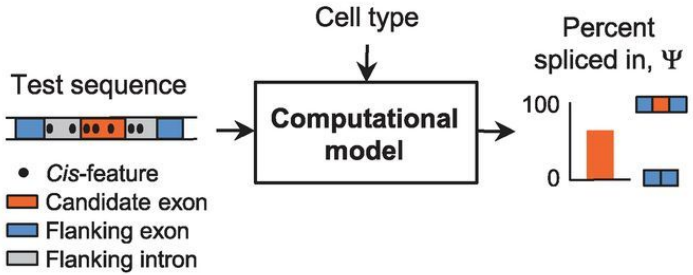


Figure 1B

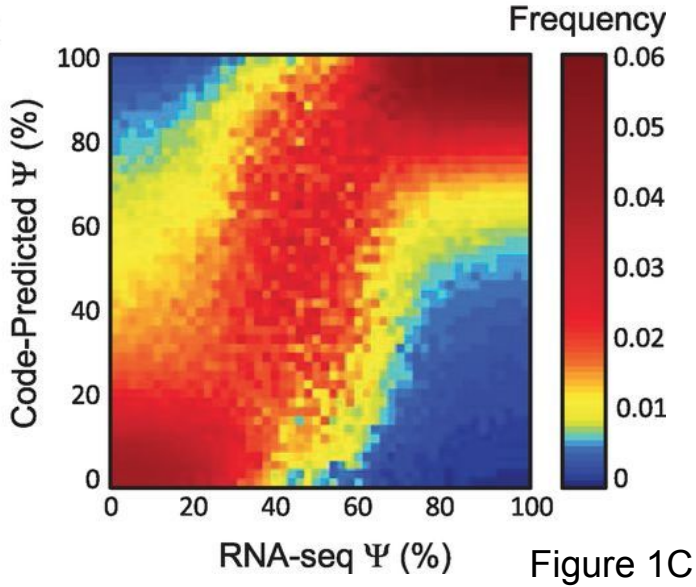
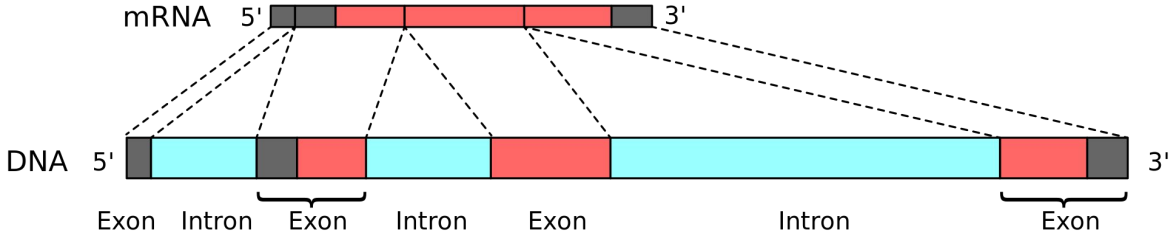


Figure 1C

Splicing regulation assessment

- Validated model using genomic data
  - 658,420 single nucleotide variations
    - 543,525 SNPs, MAF > 1%
    - 114,895 rare, disease linked, MAF < 1%
  - ~120,000 exons
  - ~16,000 genes
- Computed splicing prediction for each variant



Computing ΔΨ

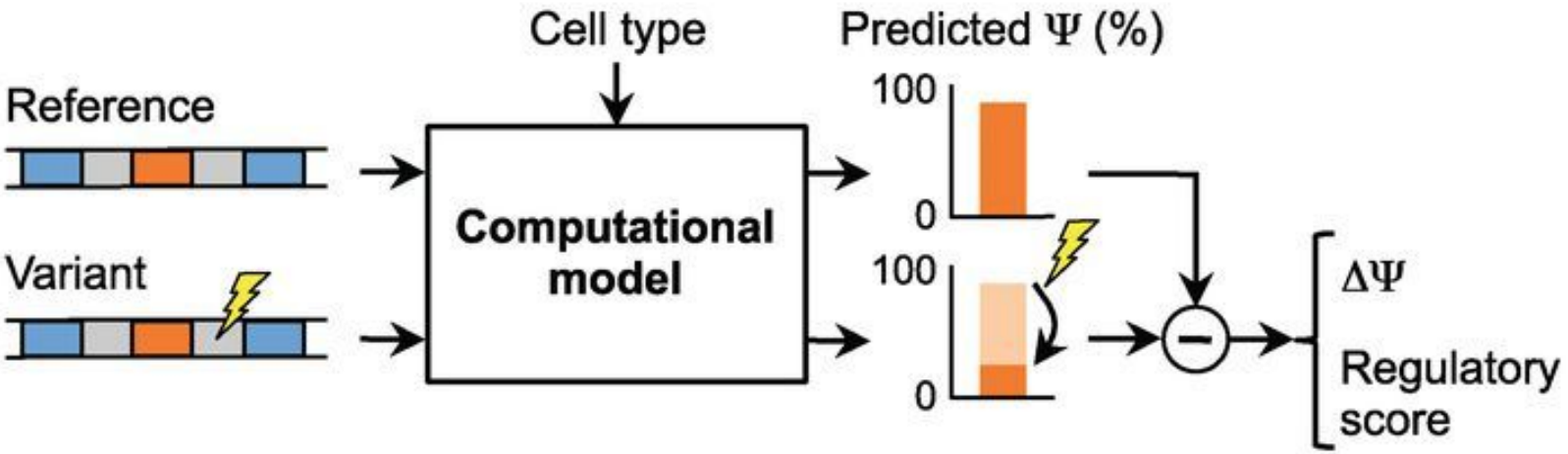


Figure 2A

SNV Effects on Splicing

- Started with more than 650,000 variants
- 20,183 Variants disrupt splicing ($|\Delta\Psi| \geq 5\%$)
- Intronic SNVs near splice sites (30bp)
- 465 intronic SNVs more than 30 bp away
- 9,525 nonsense SNVs
- 1,273 missense
- 579 synonymous SNVs

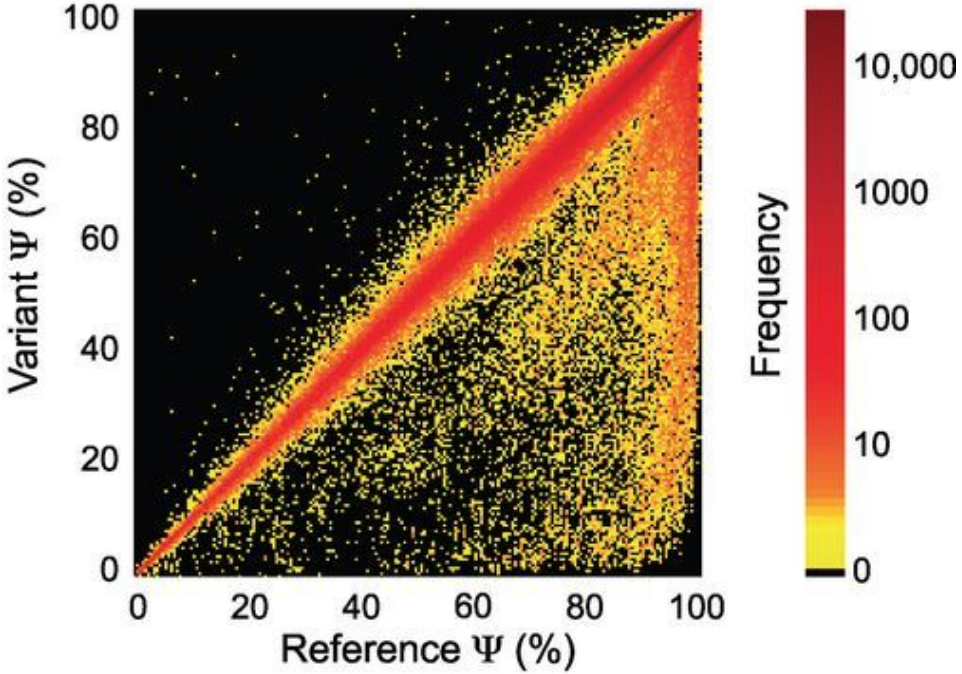


Figure 2B

Disease SNV Analysis

- $\Delta\Psi$ was computed for SNVs associated with disease
- Collected 81,608 SNVs known to be disease related, near splicing juntions
- Exonic and intronic SNPs near splice junctions associated with disease are much more significantly associated with splicing than common variants

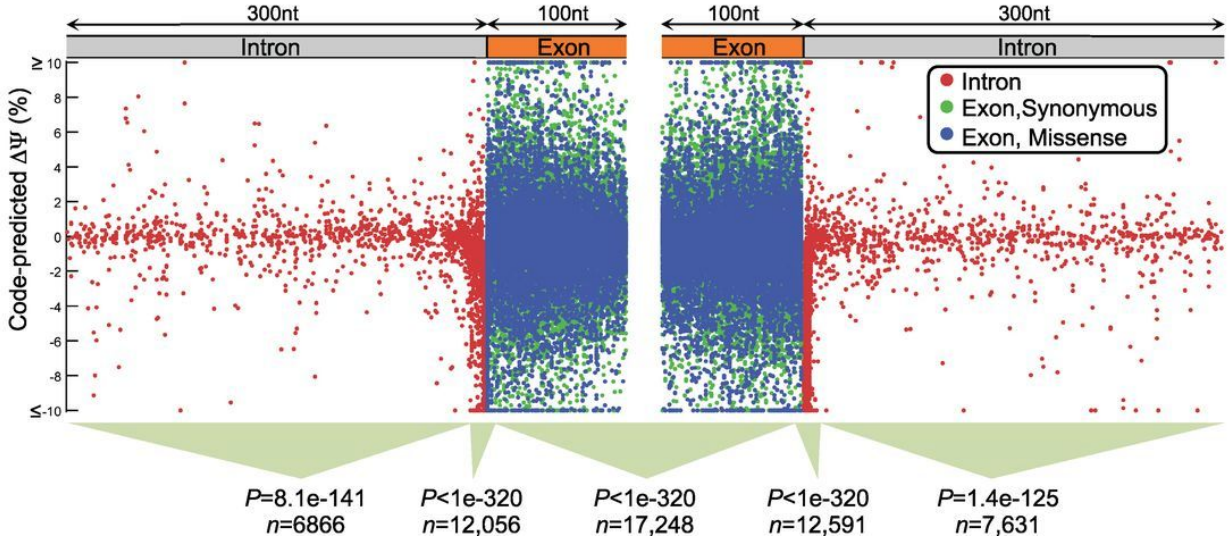


Figure 2C

Disease Investigation

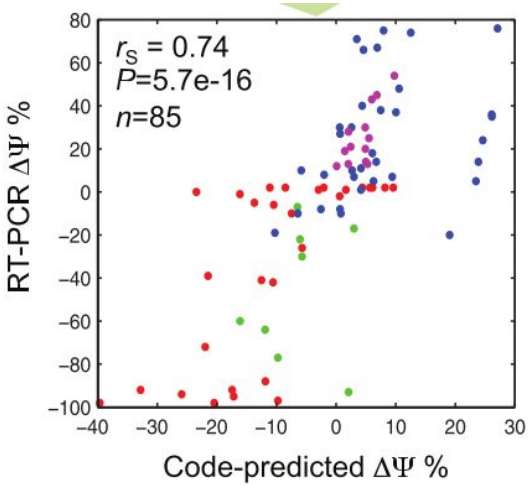
Goal: Validate Model, Demonstrate Use Cases

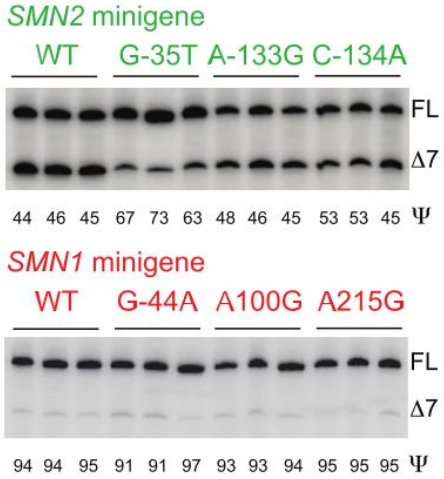
Focused on three diseases

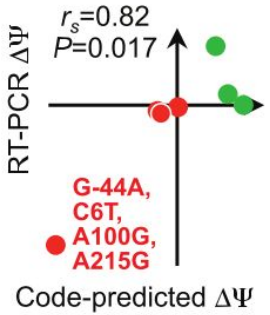
1. Spinal muscular atrophy - autosomal recessive single gene
2. Nonpolyposis colorectal cancer - oligogenic
3. Autism spectrum disorder - multigenic

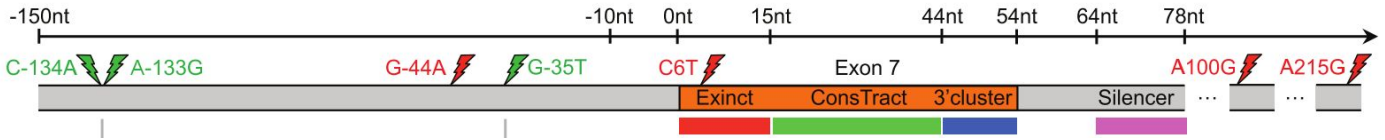
Spinal Muscular Atrophy (SMA)

- Used model to predict regulatory activity of 700 mutations around exon 7 of SMN1/2
- Validated predictions with RT-PCR, minigene, and literature





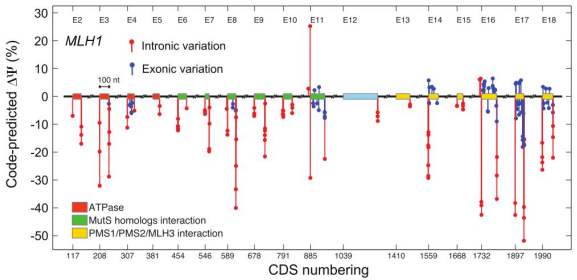


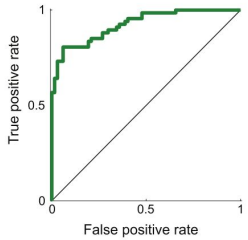


Nonpolyposis Colorectal Cancer (Lynch syndrome)

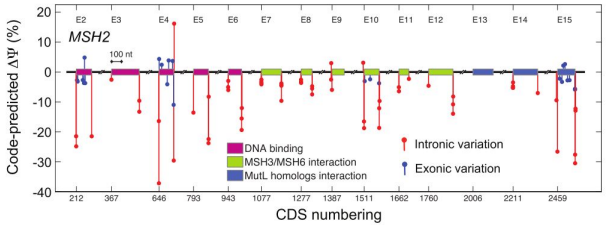
- Used model to predict regulatory activity of 977 SNVs in MLH1 and MSH2
- Validated predictions with RT-PCR, literature
- Predictions for common SNPs had significantly lower scores than predictions for patient SNVs - model is detecting causal variants

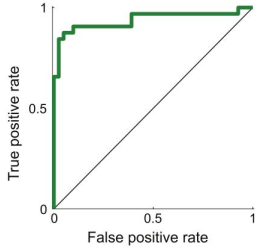
MLH1





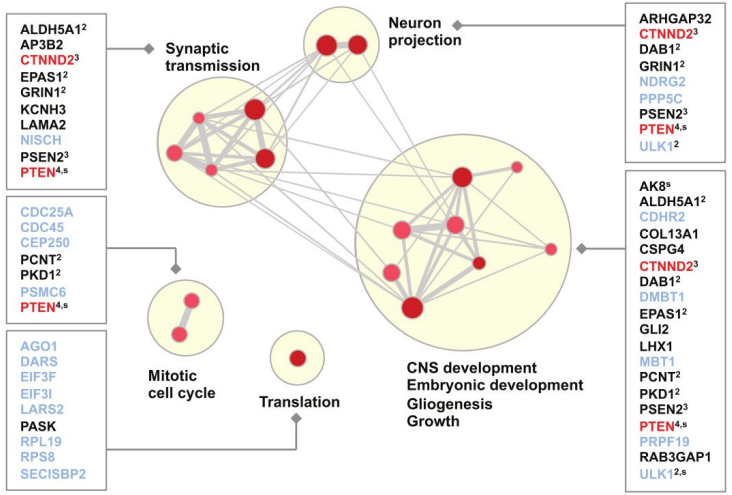
MSH2





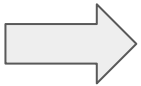
Autism Spectrum Disorder (ASD)

- Sequenced brain samples of 5 ASD cases from Autism Tissue Program, 12 control samples
- Focused on SNVs from genes with high expression in brain tissue
- Compared predicted misregulation of genes in cases and controls
- Looked at GO enrichment
- Validated subset with
  - Human Phenotype Ontology (HP)
  - Online Mendelian Inheritance in Man (OMIM)
  - Mouse Genomics Informatics (MGI/MPO)



Online Tool

- Run this analysis using SPANR at http://tools.genes.toronto.edu/

SNV 

SPANR

- Identifies exons that may be affected
- For each exon, predicts % of transcripts with exon spliced in for both reference and mutated sequence
- Reports maximum change across 16 tissues
- Produces a regulatory score for the SNV
- Shows how this SNV compares to common SNVs