

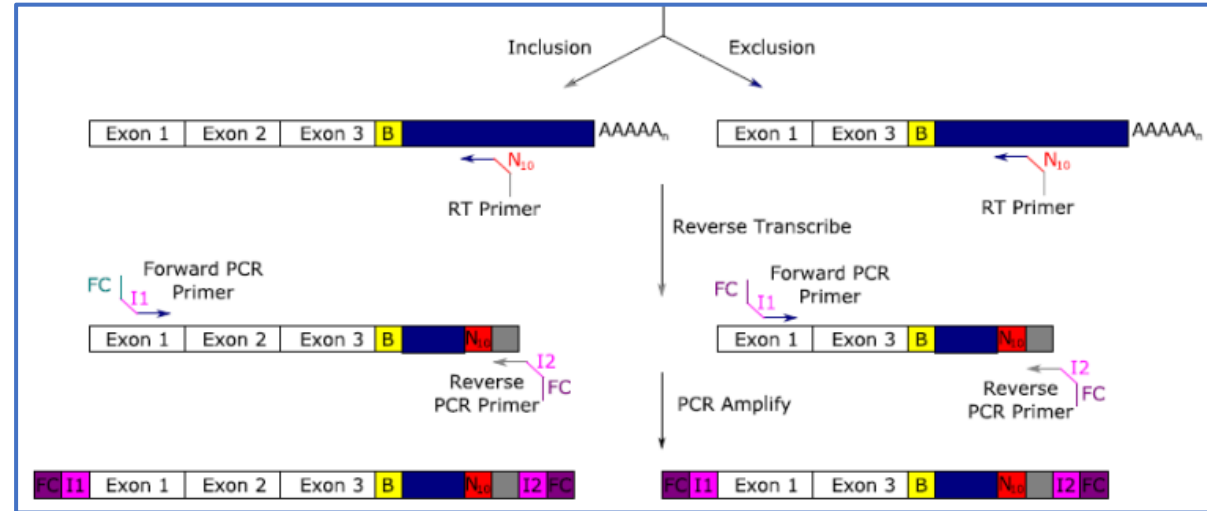
CAGI slide pack guidelines

I agree to my slides being used according to the following terms: *Steve Mount*

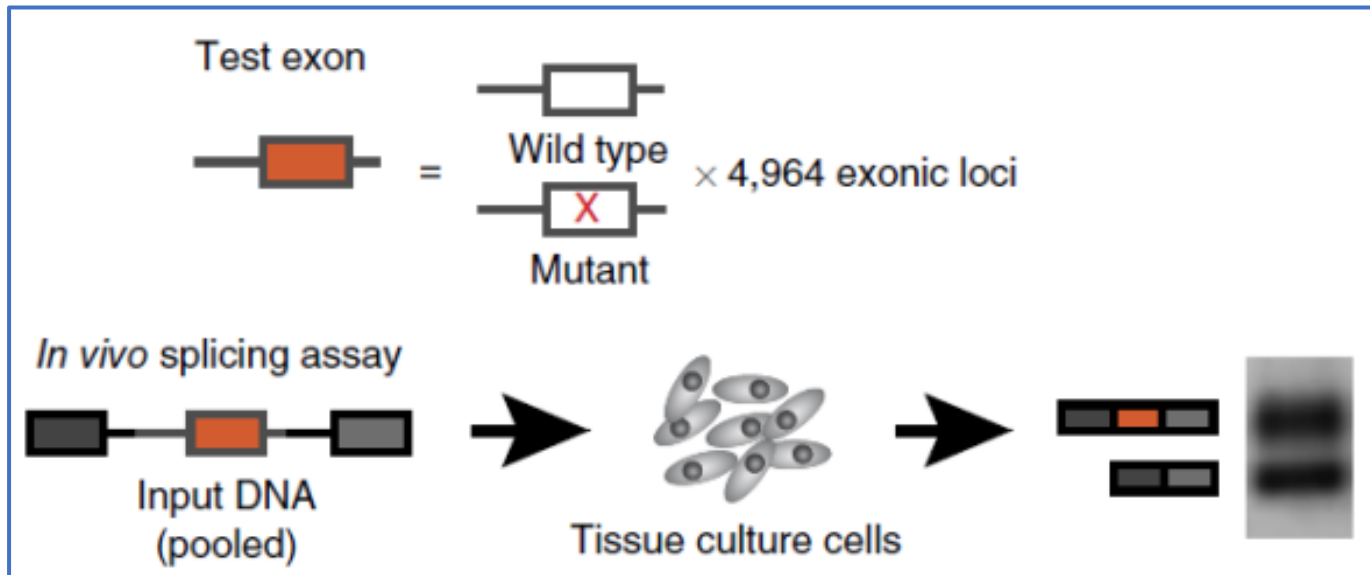
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 5 (2018). Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



CAGI Splicing challenges: Vex-seq and MaPSy

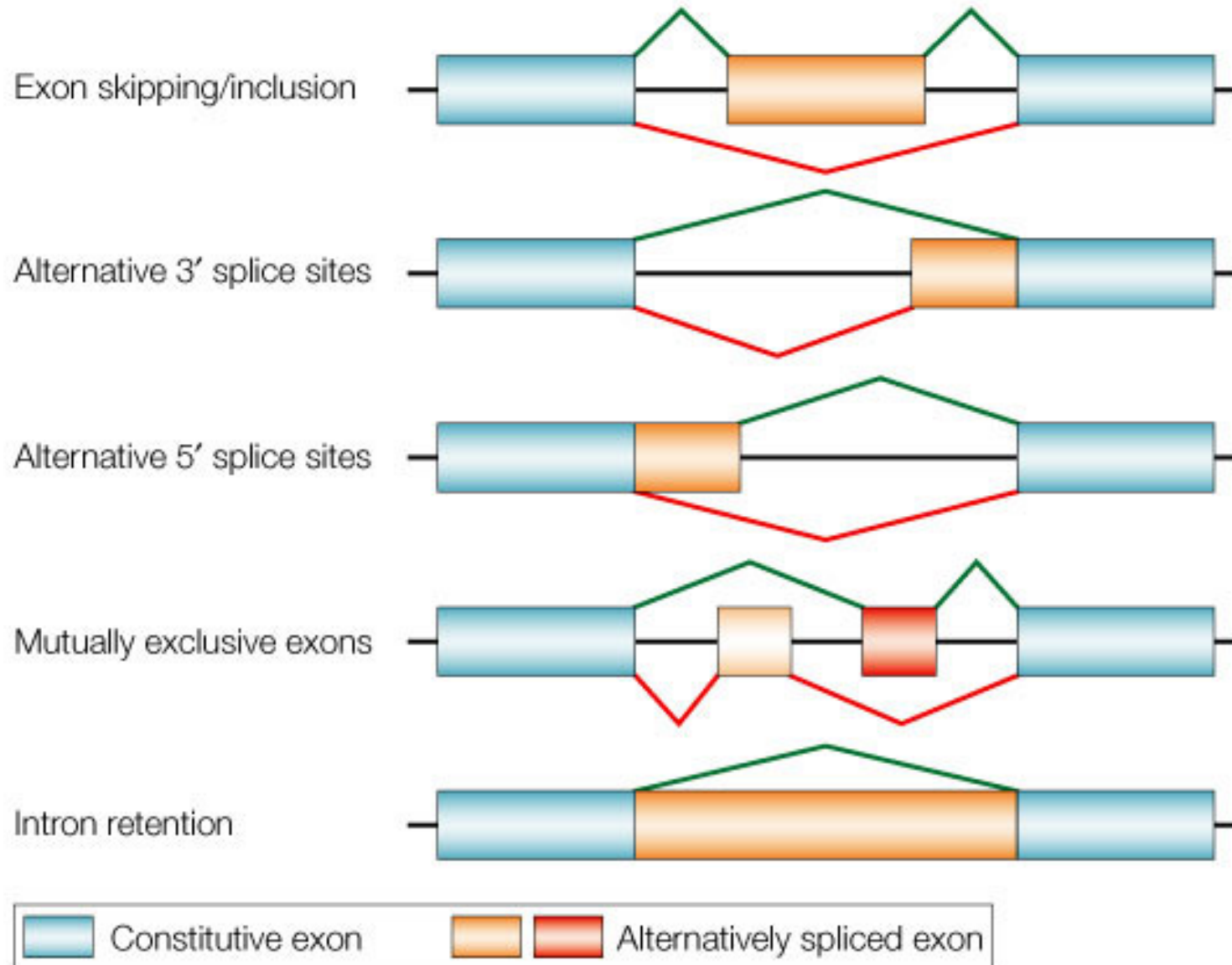


Vex-seq



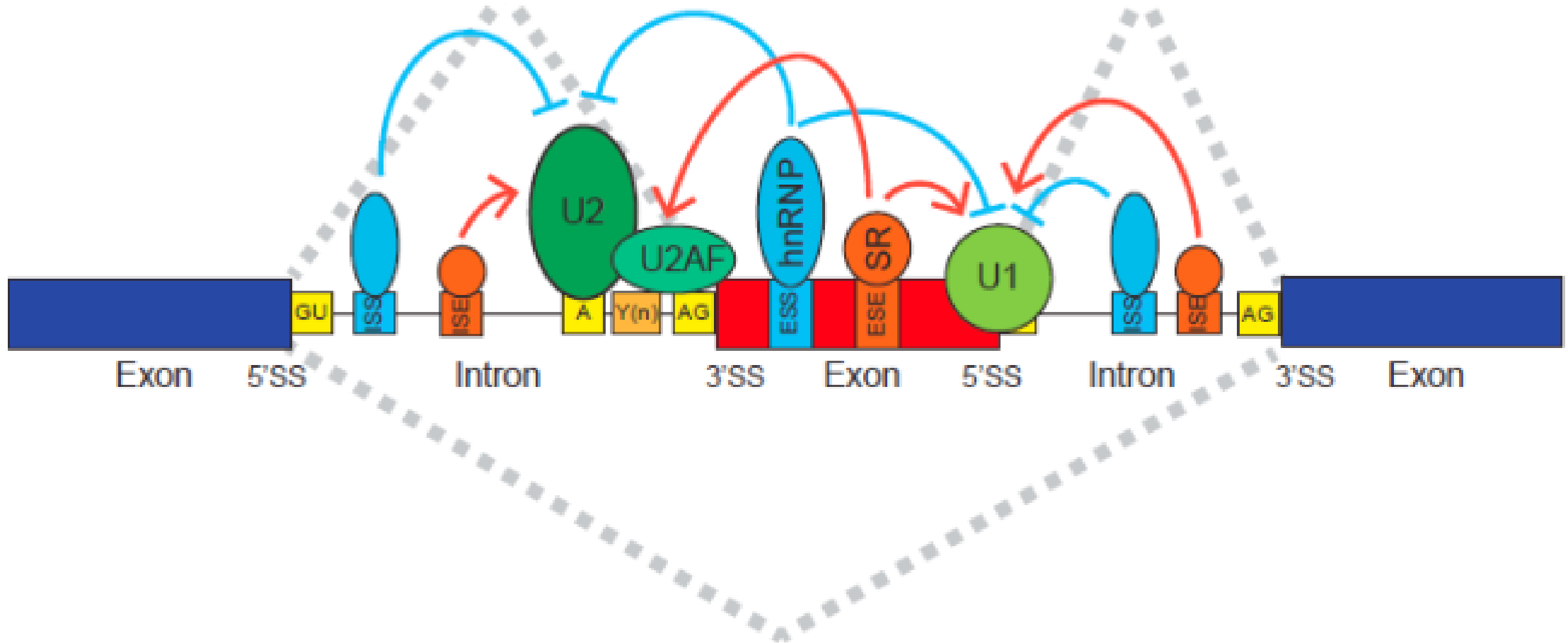
MaPSy

Alternative splicing



Reproduced from Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.* **3**, 285-298 (2002) © Macmillan Magazines Ltd.

Splice site recognition



From **The Expanding Landscape of Alternative Splicing Variation in Human Populations.**

by: [Eddie Park](#), [Zhicheng Pan](#), [Zijun Zhang](#), [Lan Lin](#), [Yi Xing](#)

American journal of human genetics, Vol. 102, No. 1. (04 January 2018), pp. 11-26

Splicing mutation outcomes



Intron retention



Exon skipping



Cryptic site activation (or choice)



Splicing mutation outcomes



Intron retention



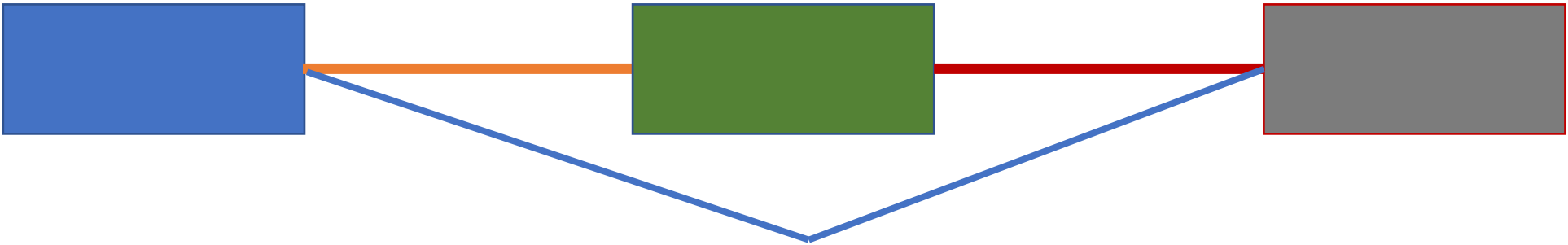
Exon skipping



Cryptic site activation (or choice)



Splicing mutation outcomes



Intron retention



Exon skipping



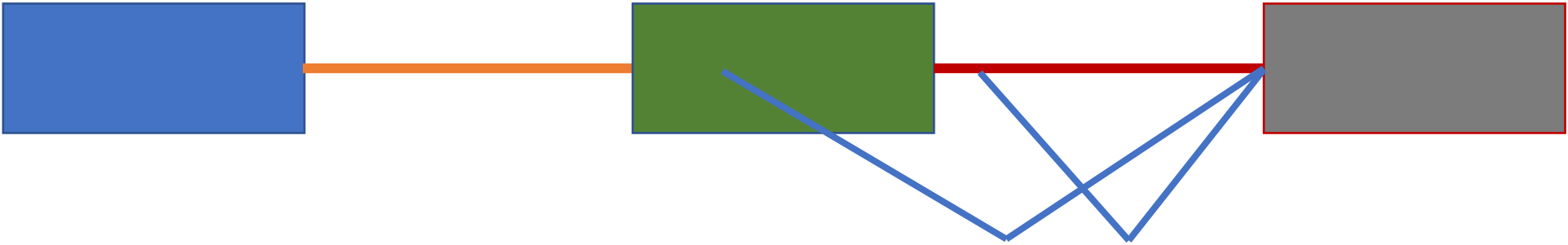
Cryptic site activation (or choice)



or



Splicing mutation outcomes



Intron retention



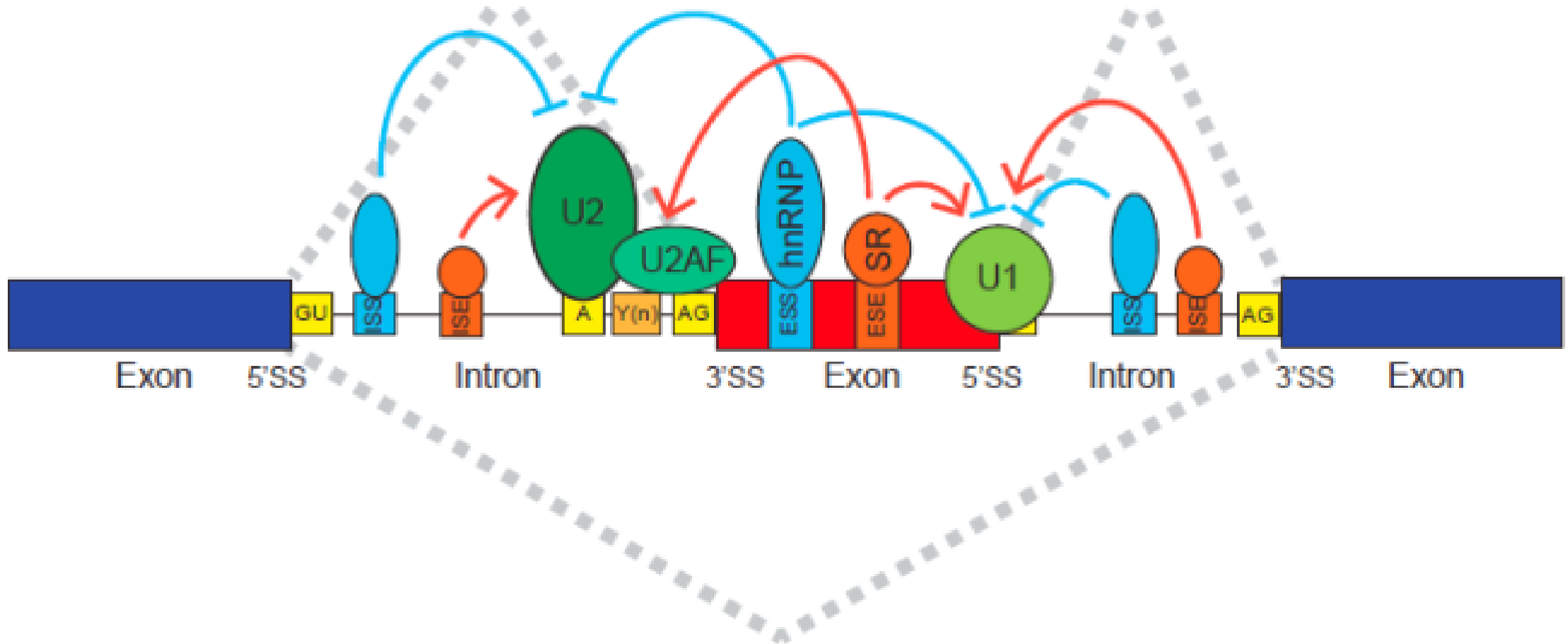
Exon skipping



Cryptic site activation (or choice)



Splice site recognition



From **The Expanding Landscape of Alternative Splicing Variation in Human Populations.**

by: [Eddie Park](#), [Zhicheng Pan](#), [Zijun Zhang](#), [Lan Lin](#), [Yi Xing](#)

American journal of human genetics, Vol. 102, No. 1. (04 January 2018), pp. 11-26

Missense and synonymous mutations are often ESE or ESS or cryptic splice site activation mutations.

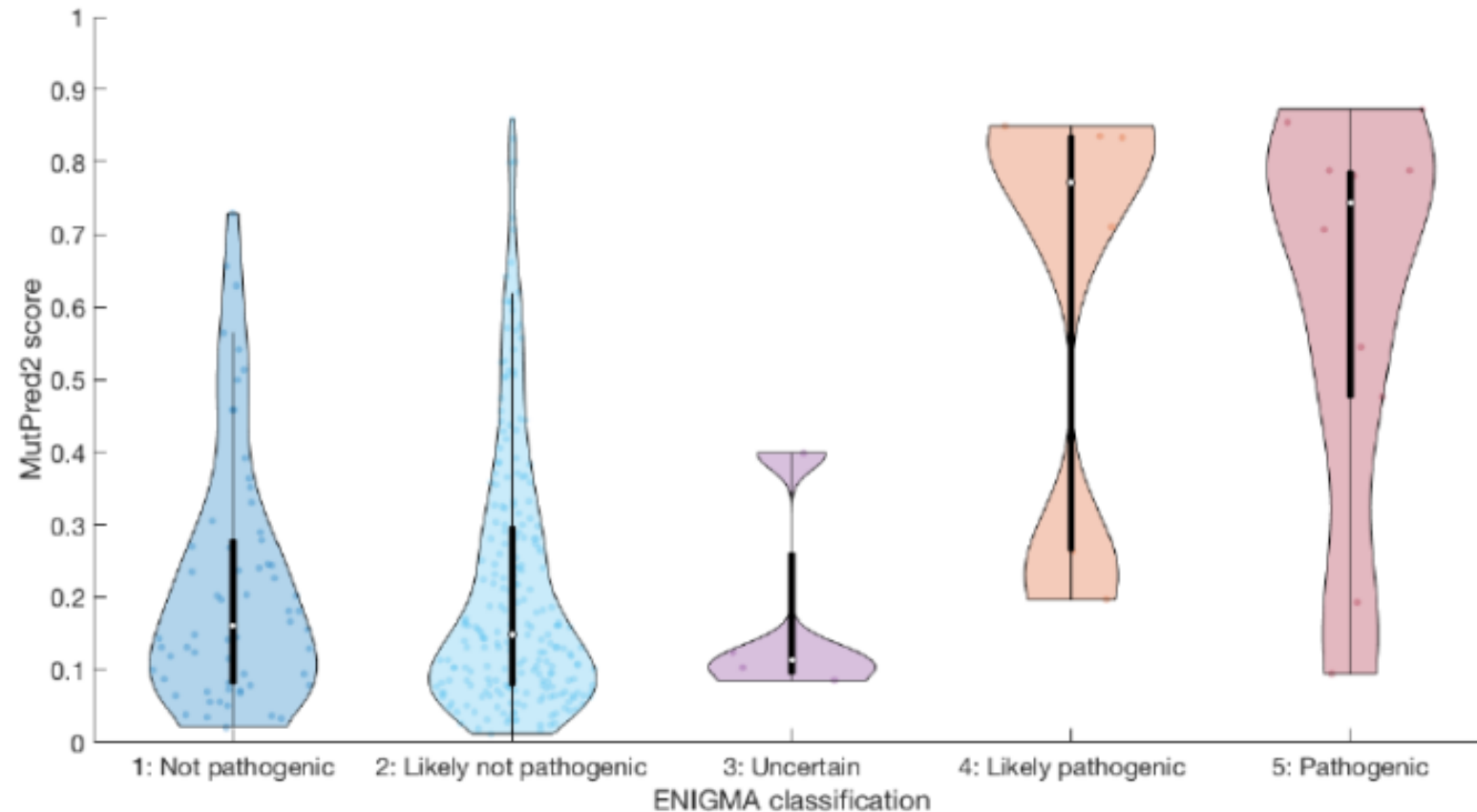
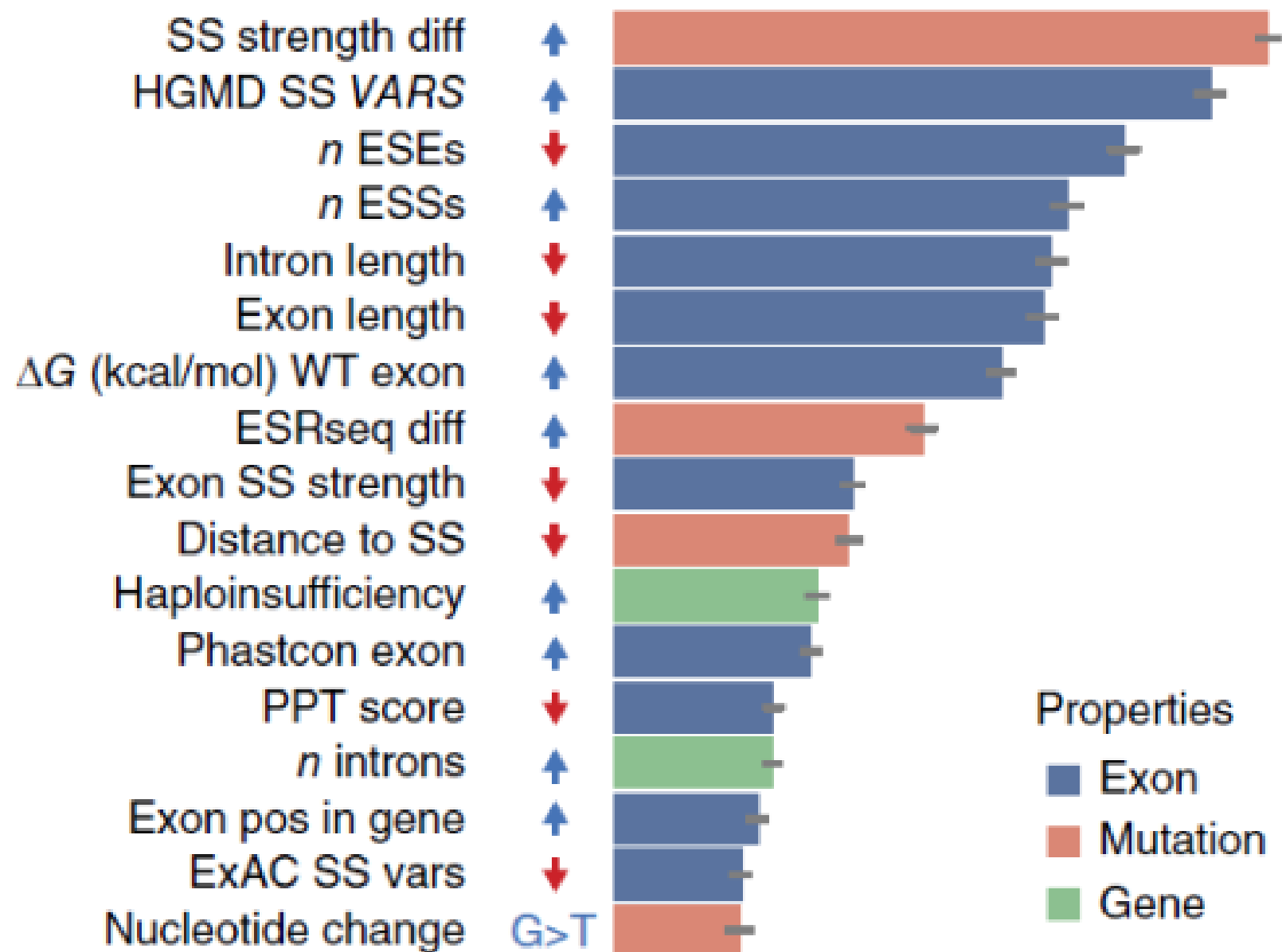
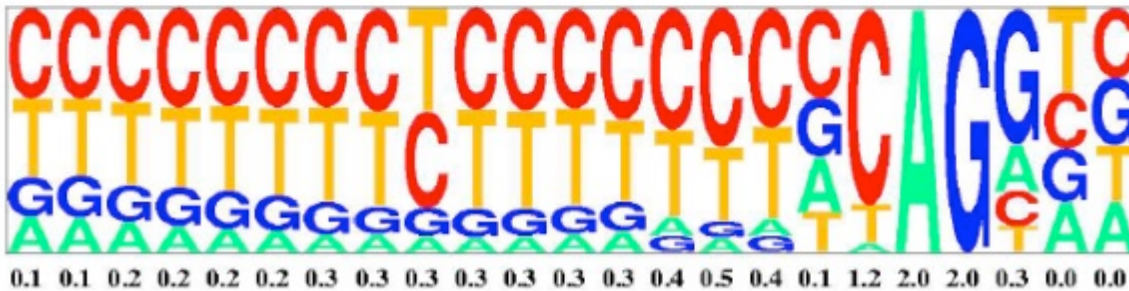
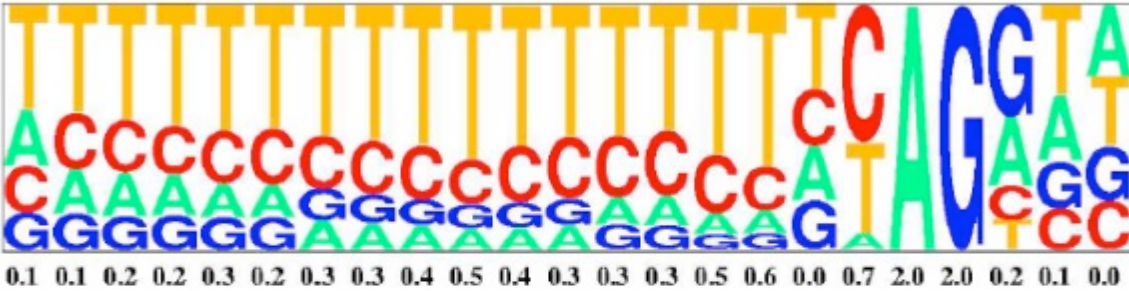
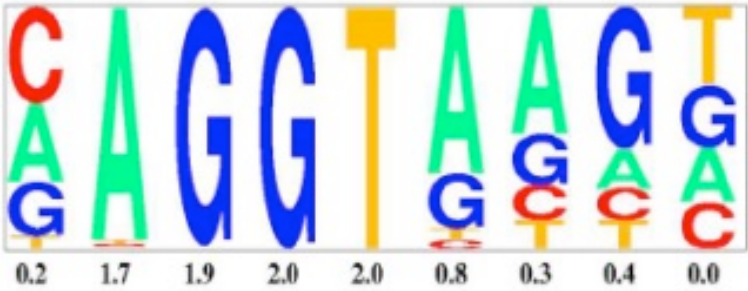
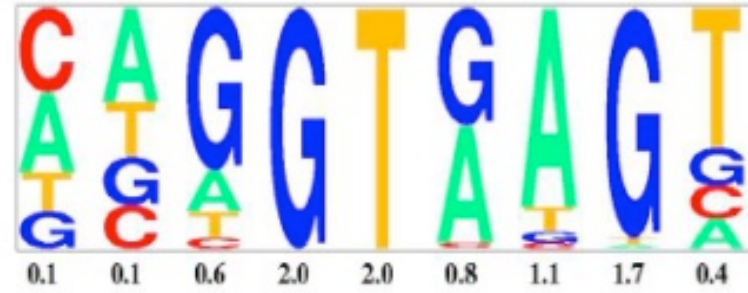


Figure 1. Violin plot showing MutPred2 score distributions over the five classes assigned by the ENIGMA Consortium. Lower MutPred2 scores are expected to be enriched among benign variants and higher scores are expected to be enriched among pathogenic variants.



Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals



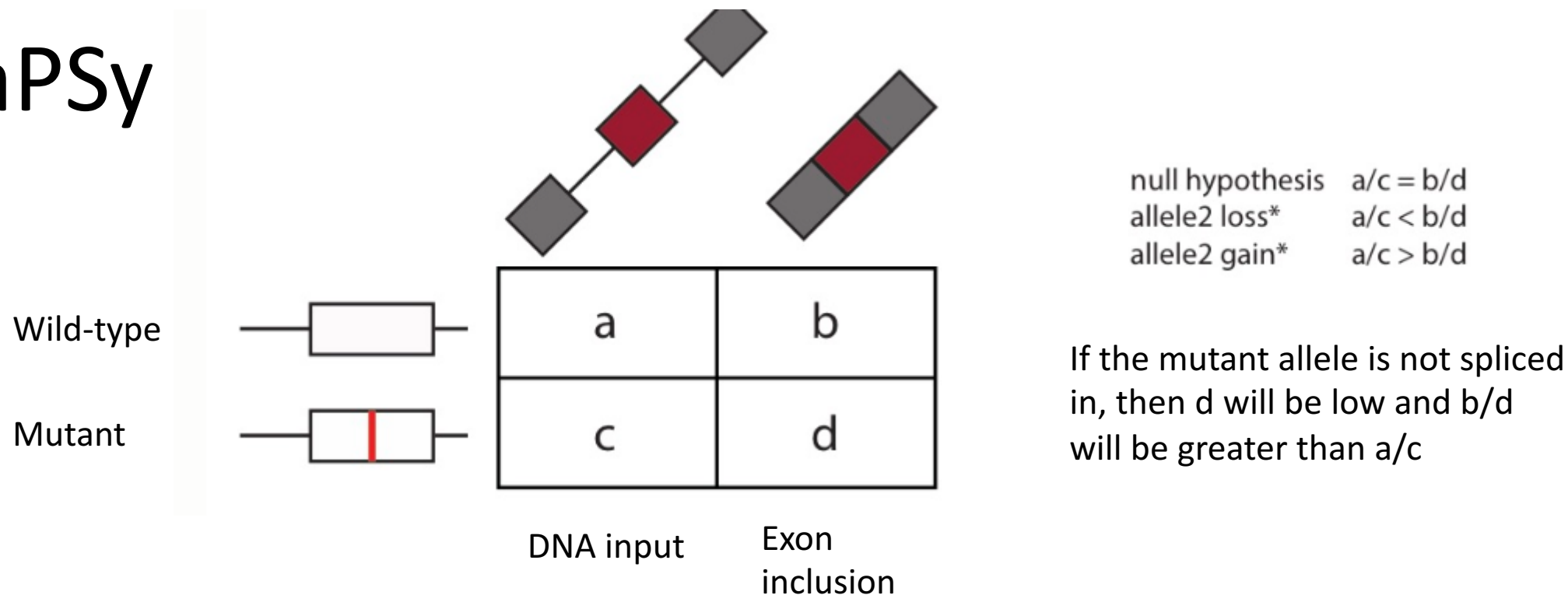
Gene Yeo^{*}
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
geneyeo@mit.edu

Christopher B. Burge^{† ‡}
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
cburge@mit.edu

The principle of minimum cross-entropy (or minimum relative entropy, MRE), first introduced by Kullback, is a generalization of the MEP that applies in cases when a prior background distribution q is known in addition to the set of constraints. Of the distributions that satisfy the constraints, the MRE distribution is the one with the lowest relative entropy (or KL-divergence), D , relative to this background distribution.

$$D_{KL}(\hat{p}) = \sum \hat{p}(x) \log \frac{\hat{p}(x)}{q(x)} \quad (2)$$

MaPSy



Participants should provide the **probability** that each variant is an Exonic Splicing Mutation (that passed the 1.5 fold change and a two-sided Fisher's exact test adjusted with 5% FDR both in vitro and in vivo).

In addition, given the input read counts of the WT input (a) and MT input (c) the participants should predict the log2 **allelic skew ratio** for in vivo and in vitro panels for each pair in the test set.

Exonic Splicing Mutation Probability

Team	Average P(TP) and 1-P(TN)	Per sample agreement	Approach
Bar-Ilan	0.563	0.817	Literature features (LIT) plus k-mers (DB, database). This Prediction was done by training a decision tree (J48) on all features of both LIT and DB.
Munich	0.519	0.916	Compositional modeling
Bologna	0.446	0.715	“All features” (Soemodi, MaxEnt), Random Forest, CNN
Tokyo	0.550	0.653	An ensemble model of convolutional neural networks (CNNs), random forests algorithm (RFs), and logistic regression (LR). Features similar to the Soemodi list.
Toronto	0.514	0.914	MaxEnt, a NN version of MaxEnt, Exon definion, hexamer scores, ESE and ESS motifs.

Bar-Ilon features

1. Exon length.
2. Position of exon in gene (1st, 2nd etc.).
3. Number of exons in gene.
4. Delta G of the wt exon, the mt exon and the difference between them.
5. Delta G of the intron + wt exon, the intron + mt exon and the difference between them.
6. Delta G of the wt exon + intron, the mt exon + intron and the difference between them.
7. Delta G of the intron + wt exon + intron, the intron + mt exon + intron and the difference between them.
8. Number of ese and ess motifs per exon (wt and mt).
9. Sum of ese and ess values for the wt and mt exons and the difference between them.
10. Length of the previous intron.
11. Length of the next intron.
12. Distance of mutation from the beginning of the exon.
13. Distance of mutation from the end of the exon.
14. Position of mutation according to the exon (in the previous intron, in the exon or in the next intron).
15. Output from the Ex – skip website that for two given wt and mt exons returns which one has a higher chance of exon skipping.
16. Haploinsufficiency.
17. Conservation.
18. GC – content difference between the exon and the beginning of the next intron.
19. GC – content difference between the exon and the end of the previous intron.
20. 5' Splice site strength.
21. 3' Splice site strength.

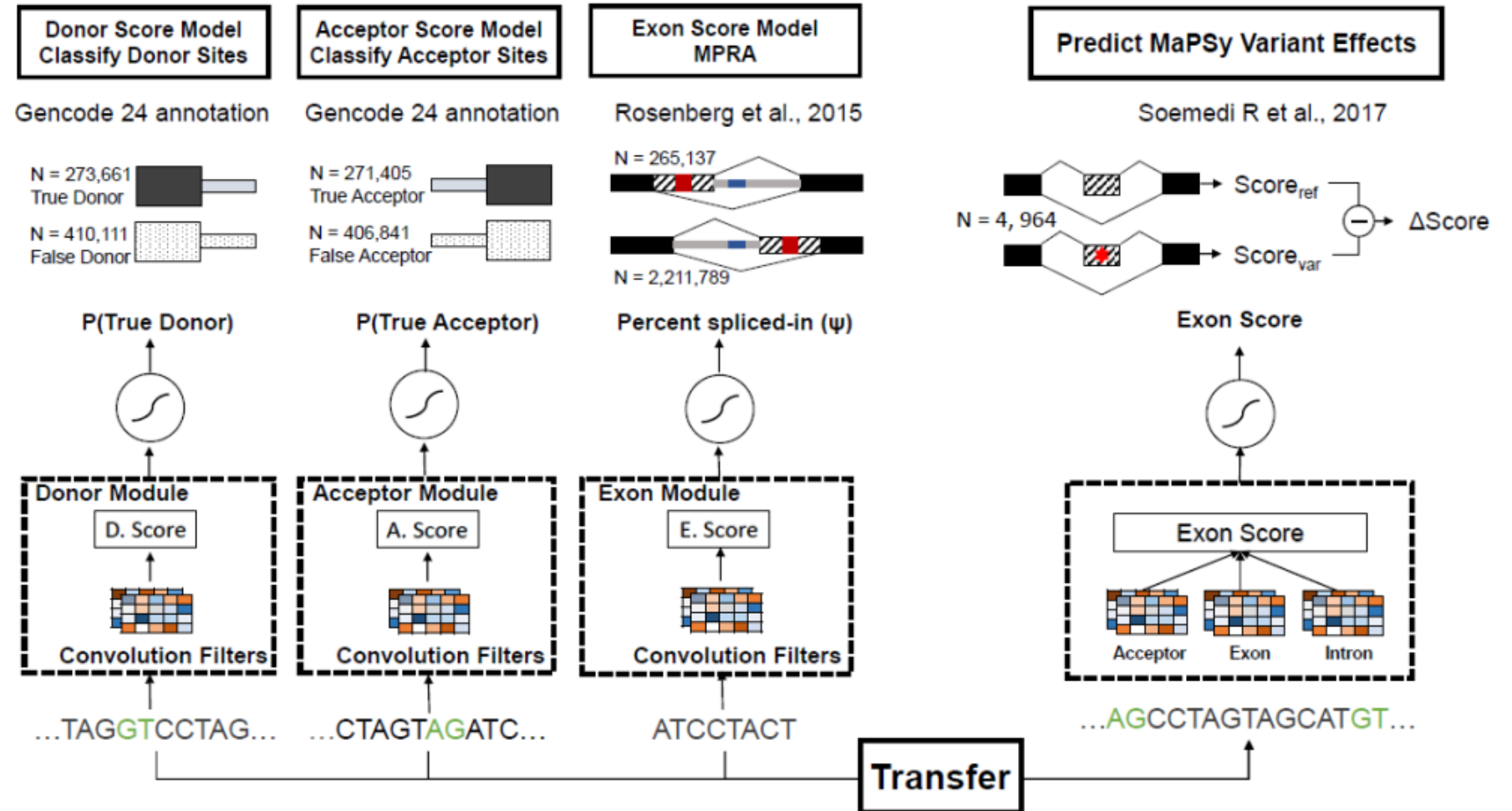
Munich approach

Compositional modeling of core splicing regulation improves
prediction of variant effects on splicing

Jun Cheng¹, Nguyen Thi Yen Duong¹, Žiga Avsec¹, and Julien Gagneur^{*1}

¹Department of Computer Science, Technical University of Munich, Germany

April 2018



Munich approach

2.3 Model training

We summarize all models we trained in this study in (Table 1). All neural network models were trained with binary cross-entropy loss (Eq.1). We used Adam optimizer, which is a variation of stochastic gradient descent optimization algorithm that has momentum to facilitate convergence [9]. Optimal learning rates were determined with hyper-parameter optimization.

Module	Training Data	Strategy
Acceptor-model	Gencode 24 annotation	Classify annotated acceptors and random sequence
A5SS-exon-model	MPRA exonic random sequence	Share conv layer, dense layer specific for alternative 5' exon
A3SS-exon-model	MPRA exonic random sequence	Share conv layer, dense layer specific for alternative 3' exon
Joint-exon-model	MPRA exonic random sequence	Share conv and dense layer, use library specific bias term
Donor-model	Gencode 24 annotation	Classify annotated donors and random sequence
Module ssassembly	MaPSy	Robust regression with Huber loss

Table 1: Summary of trained models and their training data.

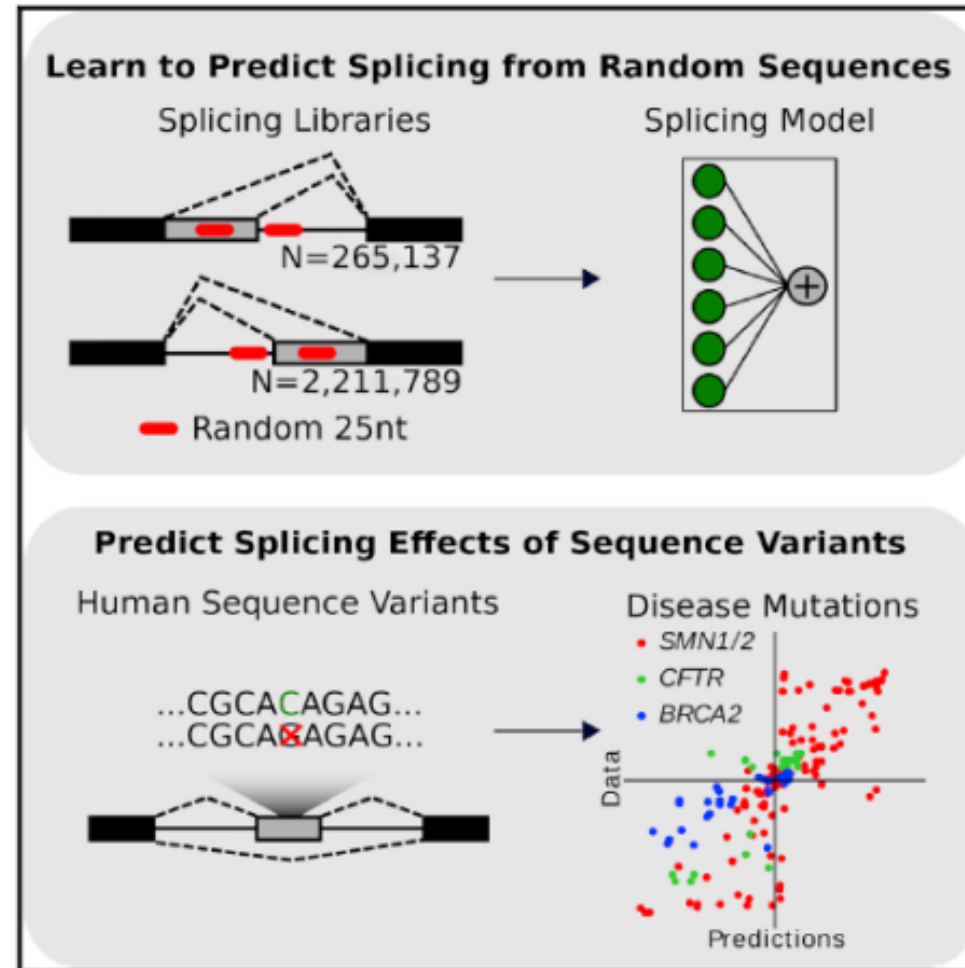
Munich approach

Cell

ARTICLE

Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences

Graphical Abstract



Authors

Alexander B. Rosenberg, Rupali P. Patwardhan, Jay Shendure, Georg Seelig

Correspondence

gseelig@uw.edu

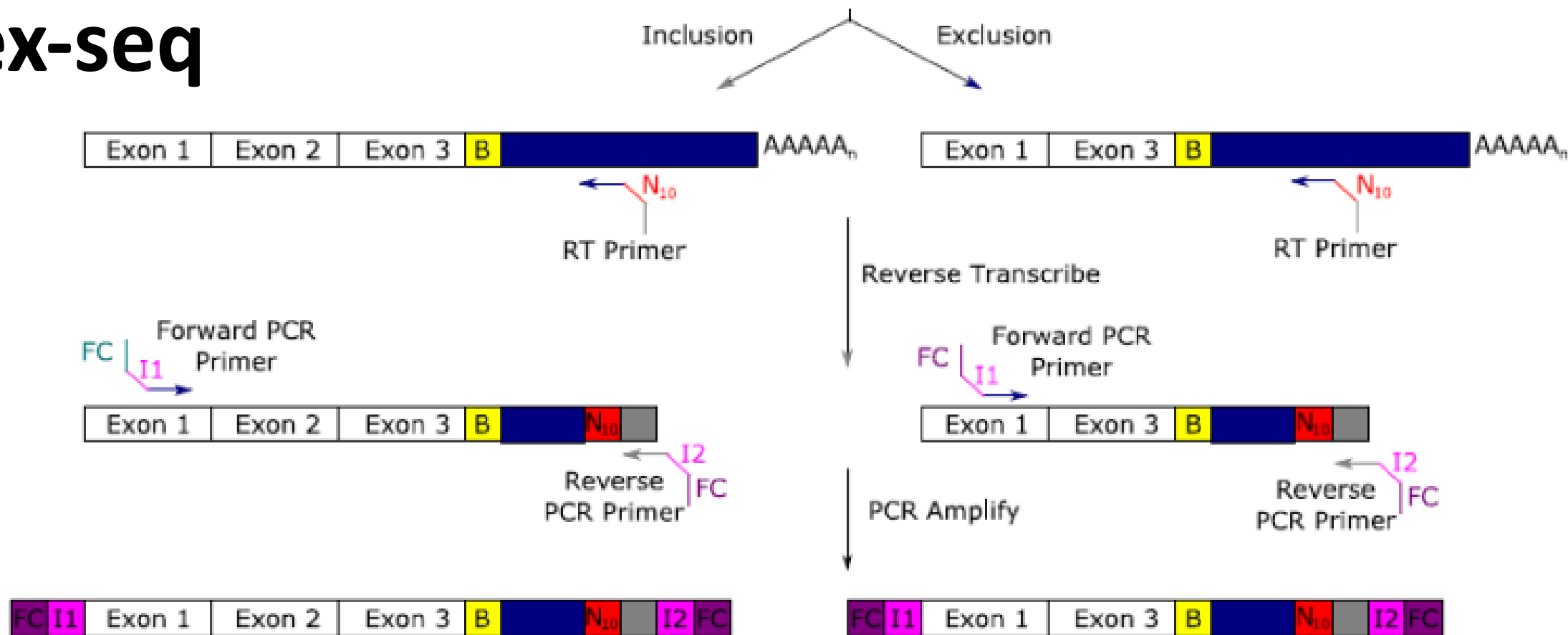
In Brief

A combination of synthetic biology and machine learning approaches identifies universal rules of RNA splicing and enables the accurate prediction of the effects of disease-related human SNPs on isoform levels.

MaPSy Allelic Skew Ratio - correlations

Team	Pearson In vitro/vivo	Spearman In vitro/vivo	Approach
Munich	0.435 0.435	0.602 0.602	Compositional modeling
Bar-Ilan	NA	NA	Literature features (LIT) plus k-mers (DB, database). This Prediction was done by training a decision tree (J48) on all features of both LIT and DB.
Bologna	0.038 0.038	0.058 0.058	“All features” (Soemodi, MaxEnt), Random Forest, CNN
Tokyo	0.354 0.354	0.524 0.524	An ensemble model of convolutional neural networks (CNNs), random forests algorithm (RFs), and logistic regression (LR). Features similar to the Soemodi list.
Toronto	0.349 0.349	0.500 0.500	Two-layer NN using 1) MaxEnt, 2) a NN version of MaxEnt, 3) Exon definition, 4) hexamer scores, 5) ESE and ESS motifs.

Vex-seq



$$\Psi = \frac{\text{inclusion reads}}{(\text{inclusion reads} + \text{exclusion reads})} * 100$$

CAGI Participants are asked to predict the outcome of the Vex-seq experiment. This is expressed in terms of $\Delta\Psi$, (delta PSI, or Percent Spliced In), between the variant Ψ and the reference Ψ

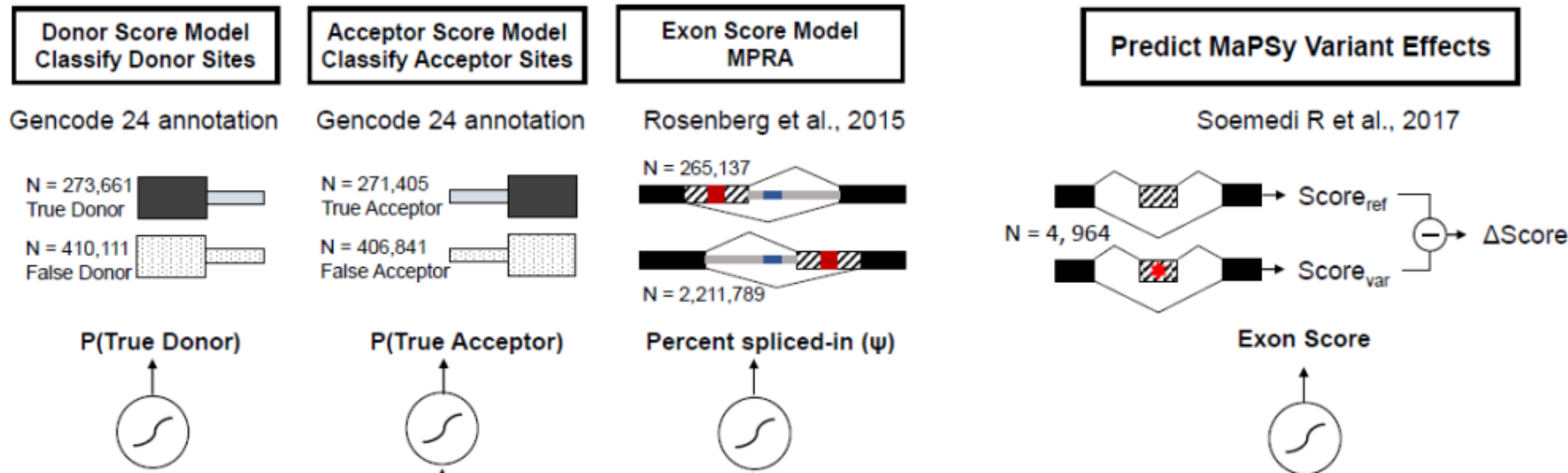
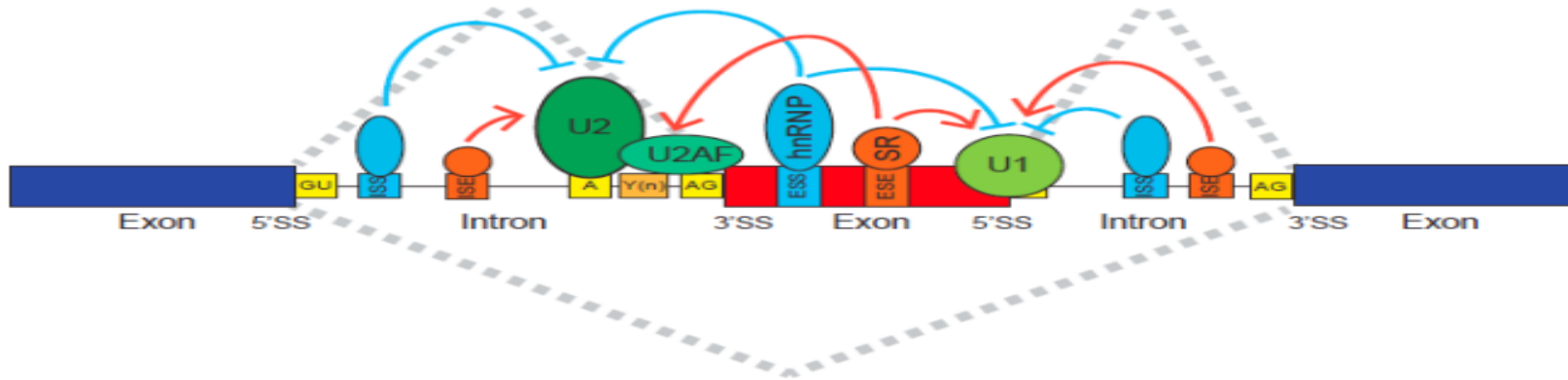
Vex-seq Ψ - correlations

Team	Pearson	Spearman	Approach
Munich	0.675	0.377	Compositional modeling
Sun Yat-Sen	0.566	0.199	Sequence -> (ANNOVAR) -> Features -> (SVM) -> Prediction Features from Livingstone et al. (DDIG-SN), supplemented with MaxEnt and splicing regulatory features from multiple sources
NIH	0.500	0.321	SplicePort and SPANR. The best performance used logit Psi.
Bar-Ilan (entry 3)	0.490	0.140	Literature features (LIT) plus k-mers (DB, database). Random Forest
Berkeley	0.430	0.072	See Poster! Many features, including ESS, ESE and ESR features and many individual binding sites.
Jilin Univ.	0.294	0.151	Methods are entirely code. It looks like they only used the test data and an SVM.

Ψ Correlations improve with 2SD outliers

Team	Pearson	Spearman	Approach
Munich	0.675 0.782	0.377 0.838	Compositional modeling
Sun Yat-Sen	0.566 0.674	0.199 0.821	Sequence -> (ANNOVAR) -> Features -> (SVM) -> Prediction Features from Livingstone et al. (DDIG-SN), supplemented with MaxEnt and splicing regulatory features from multiple sources
NIH	0.500 0.625	0.321 0.762	SplicePort and SPANR. The best performance used logit Psi.
Bar-Ilan (entry 3)	0.490 0.640	0.140 0.719	Literature features (LIT) plus k-mers (DB, database). Random Forest
Berkeley	0.430 0.586	0.072 0.763	See Poster! Many features, including ESS, ESE and ESR features and many individual binding sites.
Jilin Univ.	0.294 0.466	0.151 0.512	Methods are entirely code. It looks like they only used the test data and an SVM.

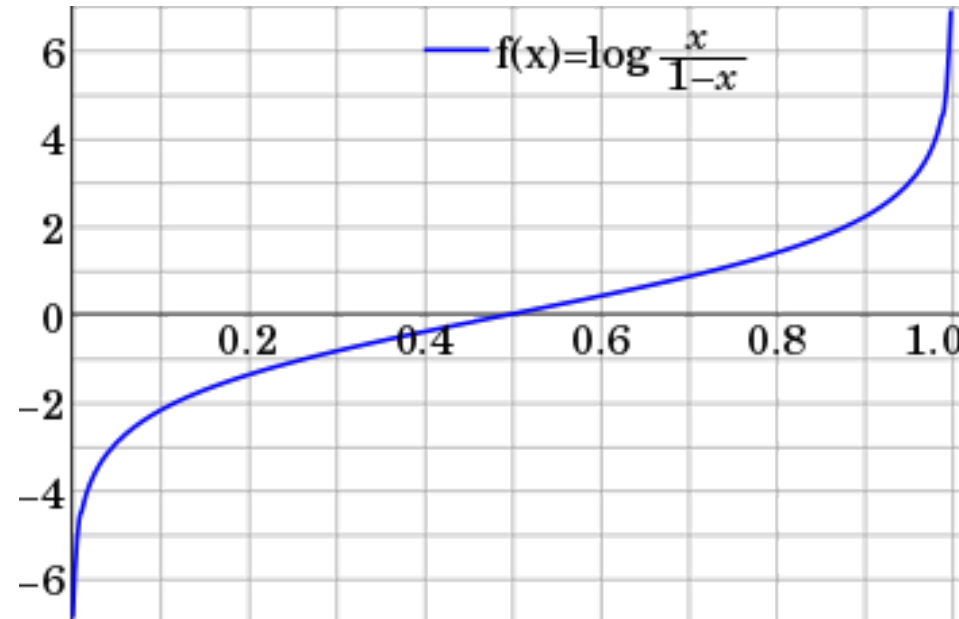
Conclusion: Modular predictions work

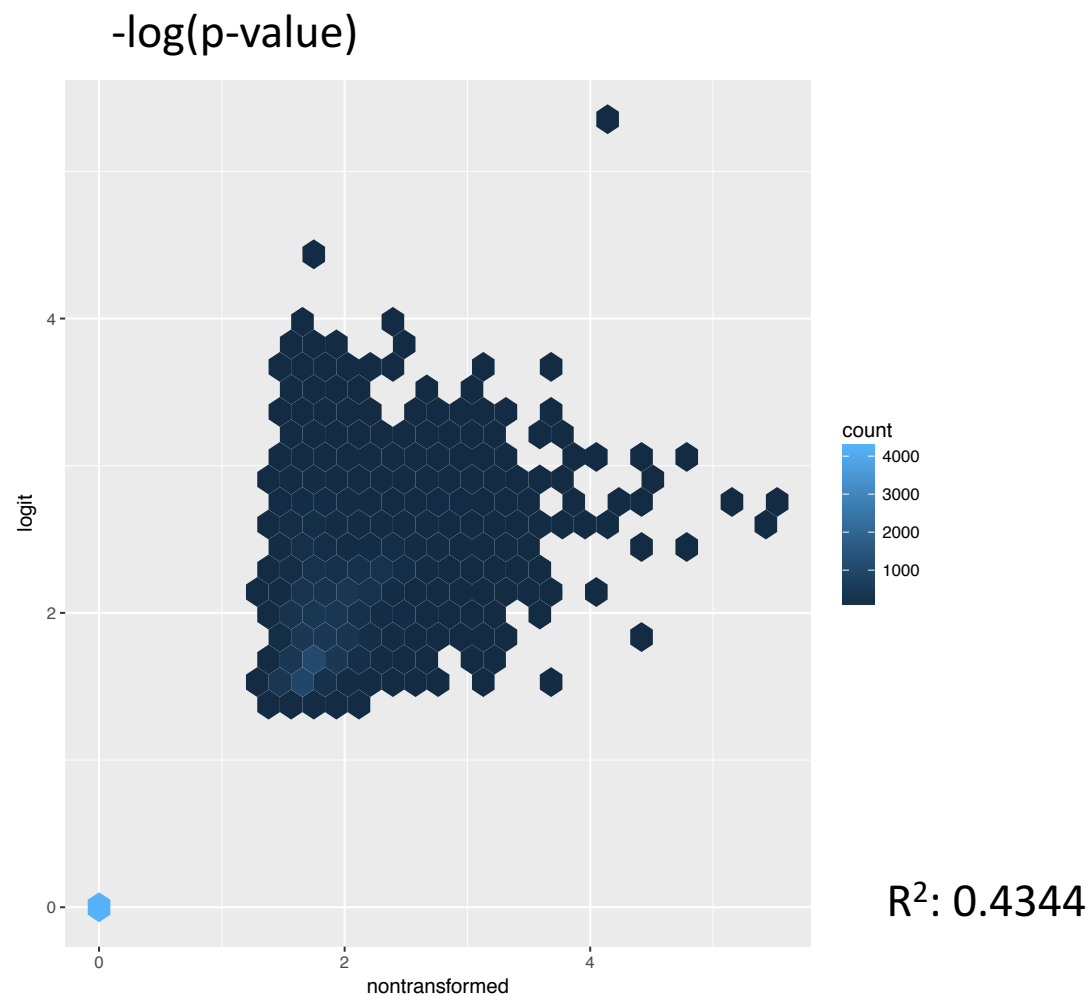


Conclusion

Consider using logit Ψ

Munich, Bar-Ilan, NIH groups
all used this





When ranked by p-value, there is little correlation between logit and nontransformed significant cases of differential alternative splicing.

This is my own data (Nazario and Wu) analyzed in my lab (by Haley Wight).