# Predicting cryptic splice site selection in genetic disorders

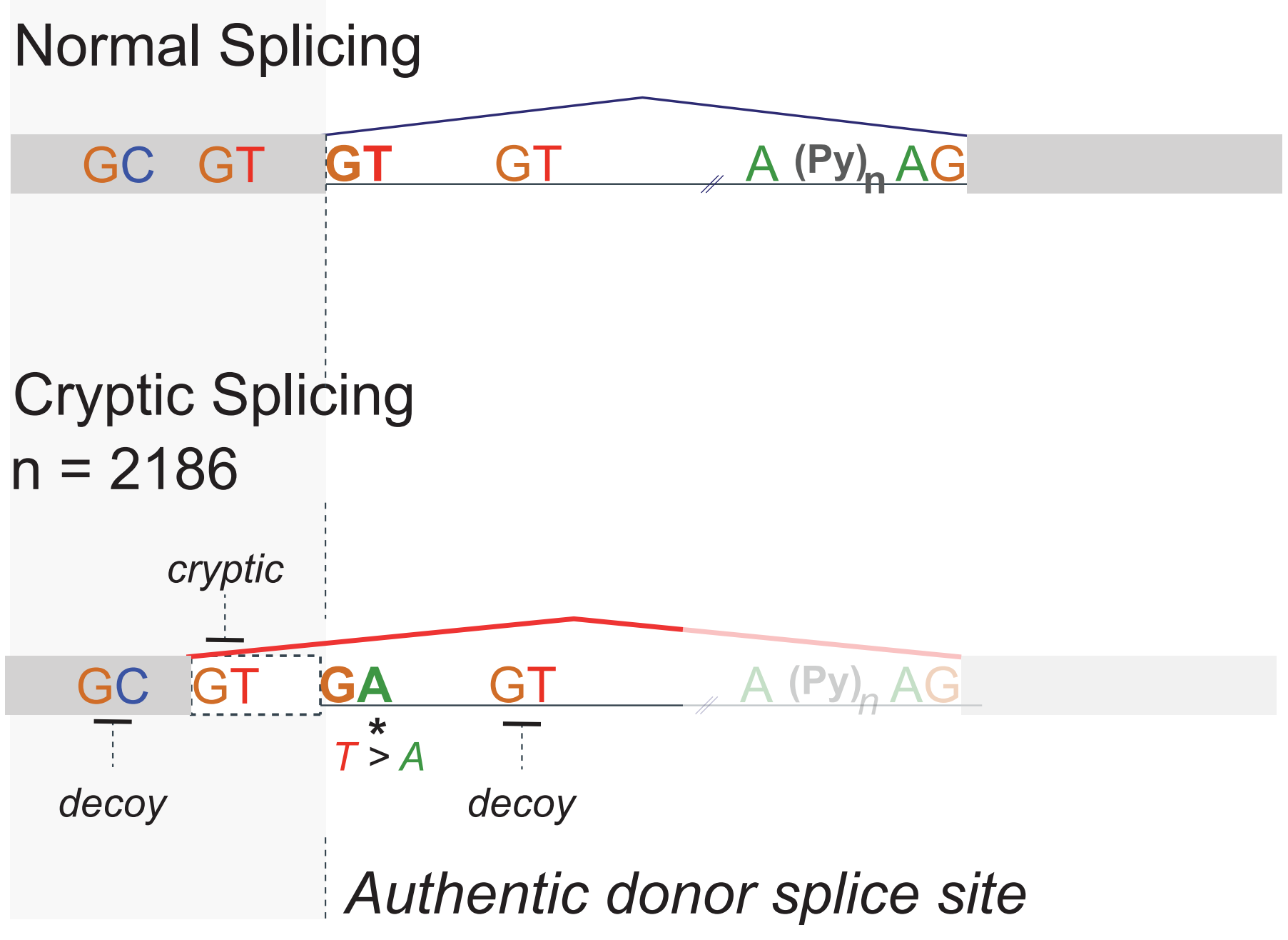
**Ruebena Dawes**[1,2], Himanshu Joshi[1,2],
Samantha Bryen[1,2], Adam Bournazos[1,2] and Sandra T. Cooper[1,2]

[1]Kids Neuroscience Centre, Kids Research, Children's Hospital at Westmead, Sydney, New South Wales, Australia, [2]Discipline of Child and Adolescent Health, Sydney Medical School, University of Sydney,

## Background

DNA variants that alter mRNA splicing are estimated to account for up to half of all disease-associated genetic variation (Baralle & Baralle 2005).

Variants modifying an Authentic donor splice site can activate spliceosomal use of a cryptic donor that may be in-frame or out-of-frame. Predicting which cryptic donor will be activated is notoriously difficult.



## Aim

To define binary features that inform prediction of cryptic donor activation.

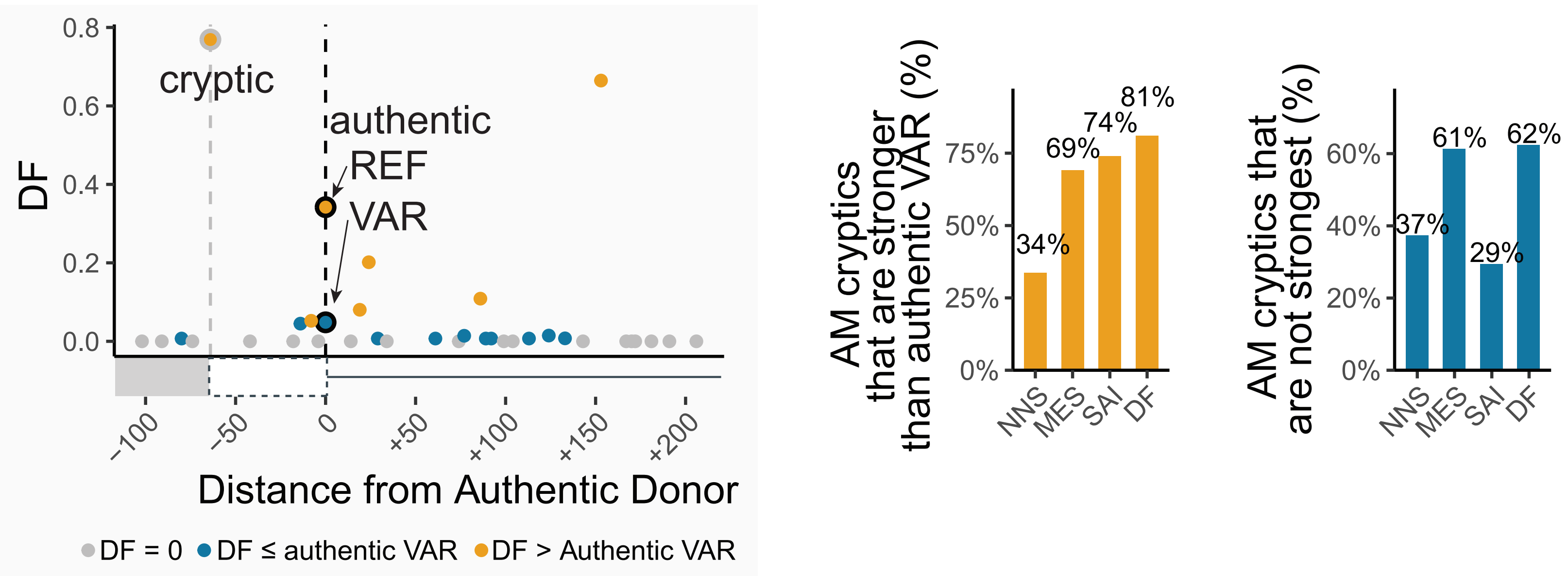
## Methods

1) Analyse features of activated cryptic donors (2,186 variants) in comparison with decoy donors not used.

2) Examine the ability of four measures of 'splice-site strength' to predict cryptic donor activation - NNSplice (NNS), MaxEntScan (MES), SpliceAI (SAI) and our own 'Donor Frequency (DF)'[1,2]

3) Assess predictive value of natural, rare cryptic donor use detected in RNA-seq data.

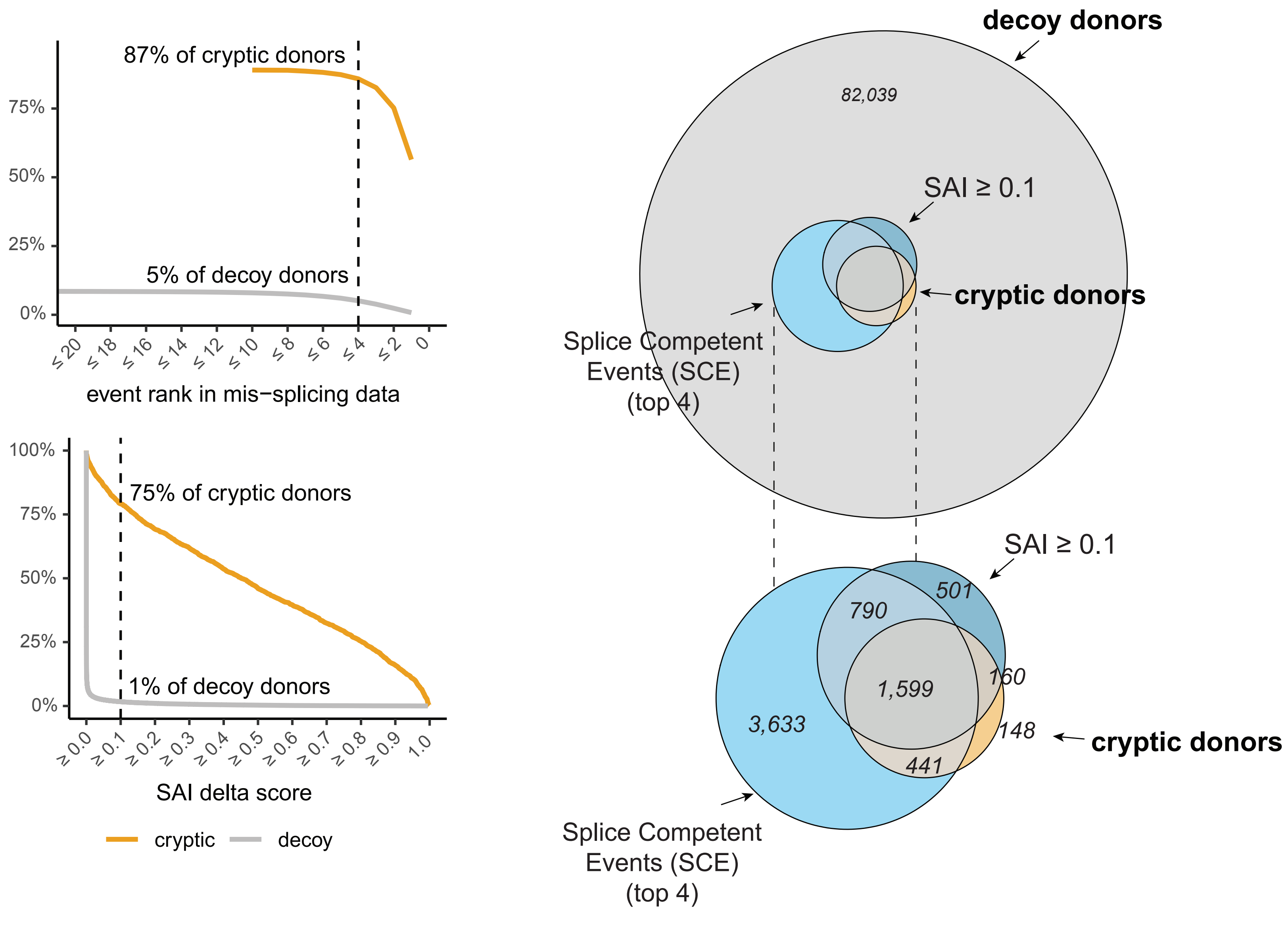
## 3) Measures of splice site strength hold predictive value but are often unreliable



A characteristic example variant shows that the authentic splice site was weakened by the variant and the cryptic then outcompetes it.

However only 34-81% of cryptic donors scored as stronger than the authentic donor after the variant, and 29-62% of cryptic donors were not scored as the strongest decoy donor within 250 nt

## 5) 90% of confirmed cryptic donors are detected as rare splice junctions in RNA-Seq data, and 95% of unused decoy donors are absent



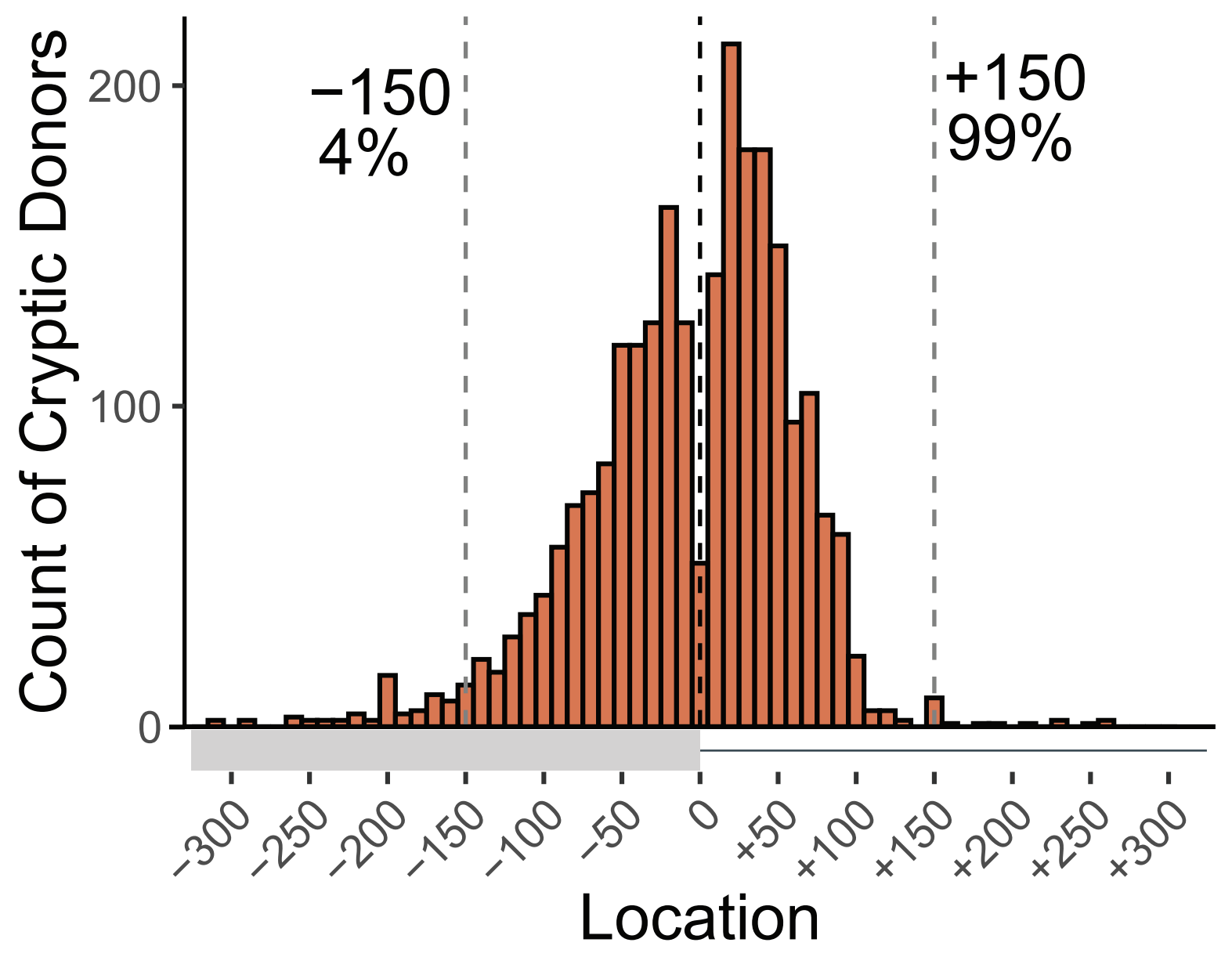
90% of cryptic donor are detected as splice-competent events in RNA-Seq data, with 87% in the top 4 splice competent events. 95% of unused decoy donors show no evidence of splice competence in RNA-seq data.

Using a liberal cut-off of 0.1 for its delta score, spliceAI accurately predicts 75% of cryptic donors and inaccurately predicts only 1% of decoy donors as functional splice sites.

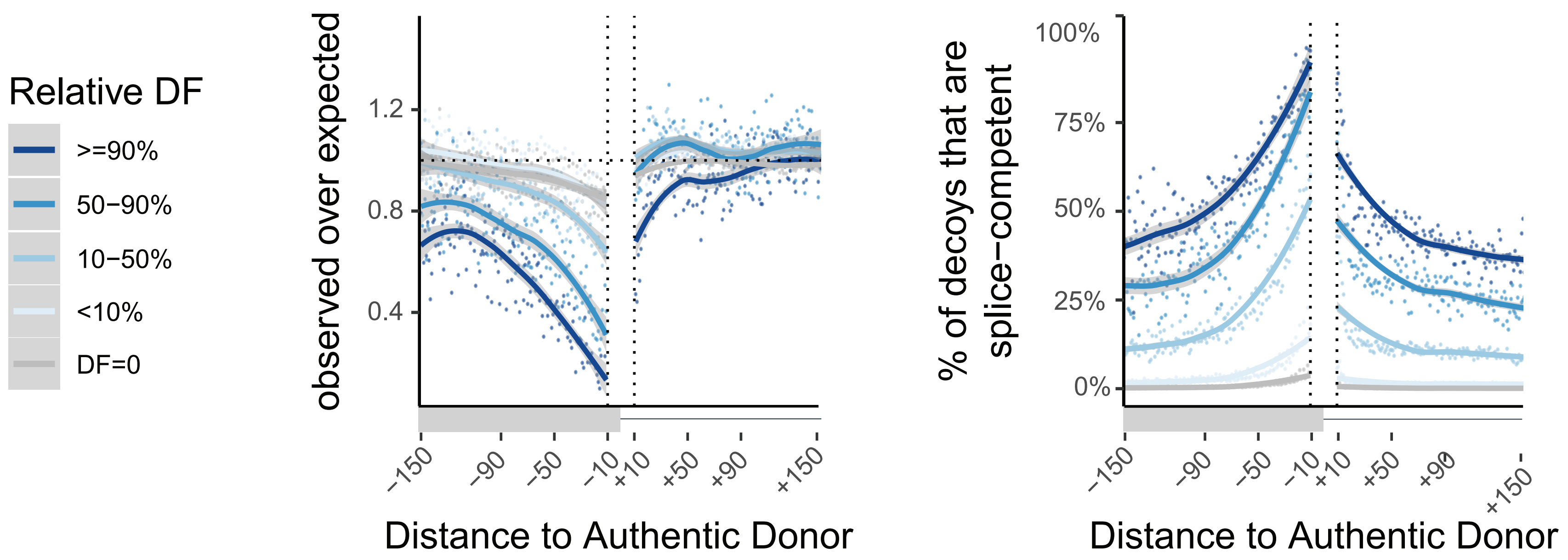
Using both, we predict 2210/2389 (93%) of cryptic donors and inaccurately predict only 6% of unused decoy donors.

## Results

### 1) 95% of cryptic donors lie within 150 nt of the authentic donor



## 2) Strength and proximity are important for spliceability

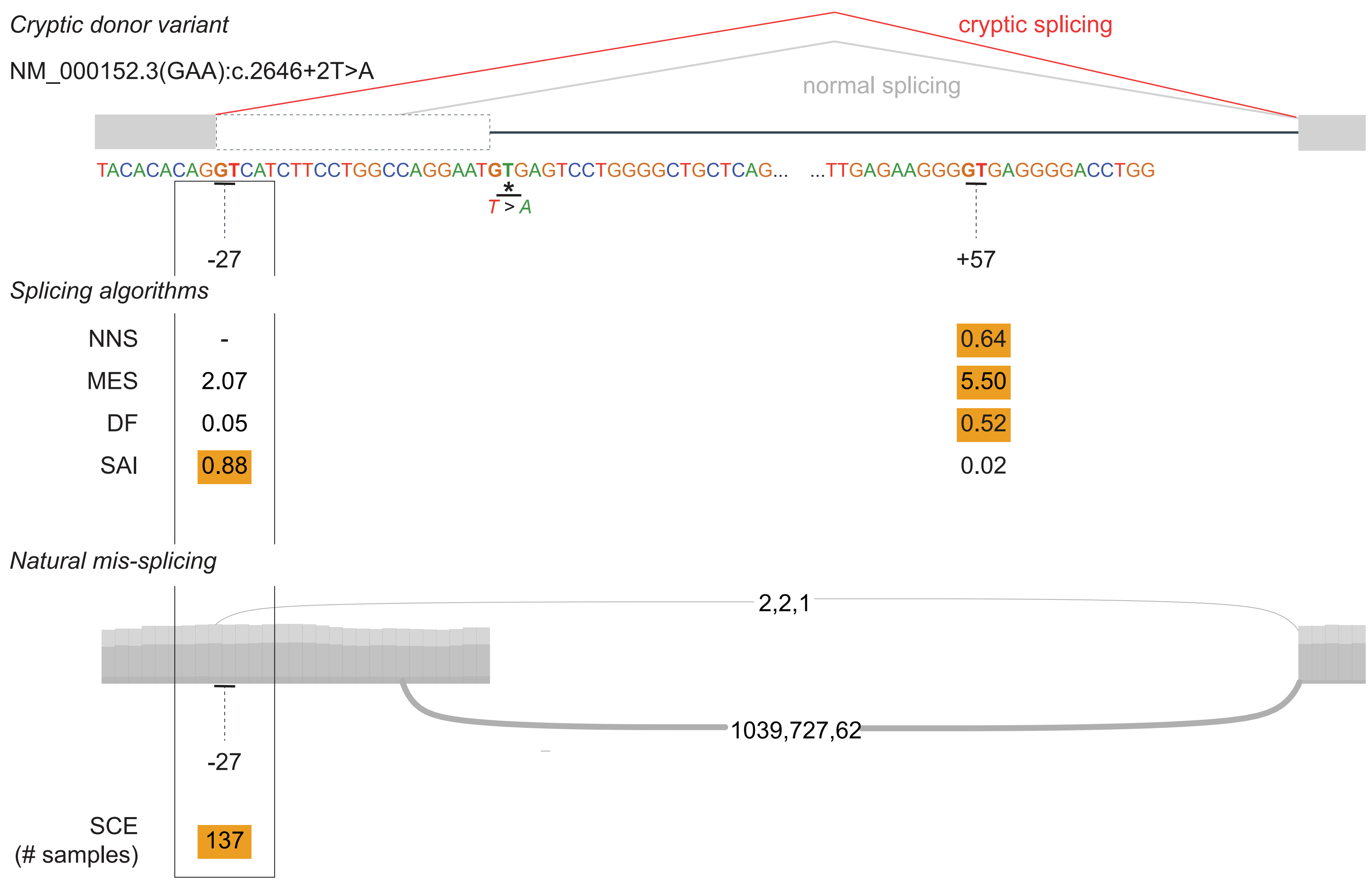


Decoy donors are ubiquitous in the human genome, however the last 50nt of exons and (to a lesser extent) the first 50nt of introns show depletion of decoy donors, increasing with strength relative to the authentic donor.

Inversely, a much higher proportion of all decoys are *splice-competent* (i.e. useable by spliceosome) in the 100 nt surrounding the authentic donor.

Decoys stronger and closer to the authentic donor are inherently more spliceable

## 4) Evidence of 'splice-competence' provides potent predictive value



A previously published variant in GAA (NM_000152.3:C.2646+2T>A) found in a patient with glycogen storage disease type II (Huie et al. 2002). The variant induces cryptic splicing to a donor 27nt upstream of the authentic donor.

SpliceAI is the only algorithm to correctly predict the cryptic splice site at -27, the rest scoring the decoy at +57 as the strongest donor. Despite the intrinsic strength of the decoy at +57, additional sequence features likely render it non splice-competent.

The cryptic at -27 is seen at very low levels in 137 samples in the natural mis-splicing database, providing evidence that it is splice-competent.

## Conclusion

While splice-site motif strength and proximity to the authentic splice site are strong determinants of cryptic donor selection, the most potent predictive information is derived from natural, rare cryptic donor use in RNA-seq data. This provides independently verifiable evidence as predictive information for pathology and is highly accurate.

The ability to accurately predict cryptic donor activation will greatly improve the clinical interpretability of splicing variants.

*Baralle, D, and M Baralle. 2005. "Splicing in Action: Assessing Disease Causing Sequence Changes." Journal of Medical Genetics 42 (10): 737–48.*

*Huie, Maryann L., Kwame Anyane-Yeboa, Edwin Guzman, and Rochelle Hirschhorn. 2002. "Homozygosity for Multiple Contiguous Single-Nucleotide Polymorphisms as an Indicator of Large Heterozygous Deletions: Identification of a Novel Heterozygous 8-Kb IntragenicDeletion (IVS7–19 to IVS15–17) in a Patient with Glycogen Storage Disease Type II." American Journal of Human Genetics 70 (4): 1054–57.*

*1. Methods to measure Donor Frequency constitutes protected IP owned jointly by the University of Sydney and Sydney Children's Hospitals Network.*
*2. Prof Sandra Cooper is Co-Director of Frontier Genomics Pty Ltd with intent to create a commercial software product that leverages invented IP. Prof Cooper does not receive a salary or other payments for services provided to Frontier Genomics Pty Ltd.*