In silico detection and correction of cryptic splice sites created by codon optimisation of an STXBP1 transgene for gene therapy

Gayatri Dhir, BSc Neuroscience, University College London

INTRODUCTION

- STXBP1 developmental and epileptic encephalopathy (STXBP1-DEE) is a paediatric neurological disorder that causes seizures and developmental delay, with an incidence rate of ~ **1:30,000** live births[1,2].
- STXBP1-DEE is caused by haploinsufficiency of **Munc18-1**, a protein essential for synaptic function[2].
- Current mainstream treatments for STXBP1-DEE are symptomatic and do not address the underlying genetic cause[2,3].
- **Thus, restoring gene dosage is an appropriate therapeutic strategy.**

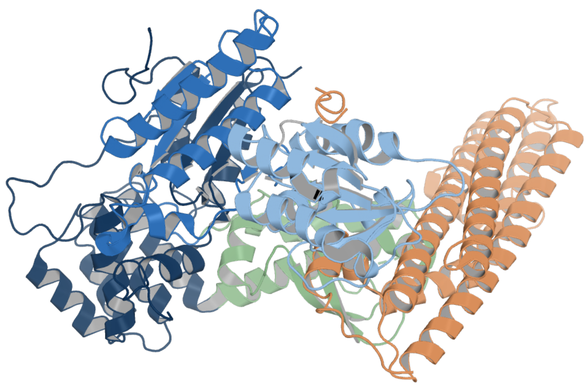

Figure 1. Structure of the Munc18-1 protein, the product of STXBP1, bound to syntaxin-1a (orange). Munc18-1 folds into four domains (domain 1, domain 2, domain 3a, domain 3b) that clamp the syntaxin-1 helix for synaptic vesicle release. PDB: 3C98, rendered in PyMOL.

METHODS

- The wild-type STXBP1 CDS (1,785 nt; 594 aa) was obtained.
- **Five** optimized sequences were generated using GenSmart, IDT (legacy), IDT (pilot), OPTIMIZER and JCat.
- Sequences were aligned to wild-type and assessed for **% identity, GC, GC3 and CpG content** using Biopython[6].
- Cryptic splice sites were screened using two complementary tools: **MaxEntScan[4] and SpliceAI[5]**, and candidate donor and acceptor sites were paired .
- Risk-associated sites were corrected using targeted synonymous substitutions and re-screened with SpliceAI and MaxEntScan.

RESULTS

Variant	GC %	GC3 %	CpG count	Identity to WT %
Wild-type	52.9	68.6	68	100.0
OPTIMIZER	63.8	99.8	124	83.1
JCat	63.8	99.8	133	83.1
IDT-legacy	50.3	63.2	88	78.6
IDT-pilot	55.1	76.8	31	82.1
GenSmart	54.7	76.0	79	80.4

Table 1. Sequence composition of wild-type STXBP1 and the five codon-optimised variants.

- All five codon-optimised sequences encoded the identical 594-aa Munc18-1 protein (Table 1).
- Sequences differed from wild type at **17–21%** of nucleotide positions (Table 1).
- Optimisation produced substantial differences in sequence composition, with GC content ranging from **50.3–63.8%** and GC3 from **63.2–99.8%** (Table 1).
- MaxEntScan identified **10** newly created cryptic splice sites across the five variants, with every optimisation strategy introducing at least one (Fig. 2).
- SpliceAI flagged only one site, in IDT (legacy) **(Δ = 0.30)**, which was also independently detected by MaxEntScan (Fig. 2).

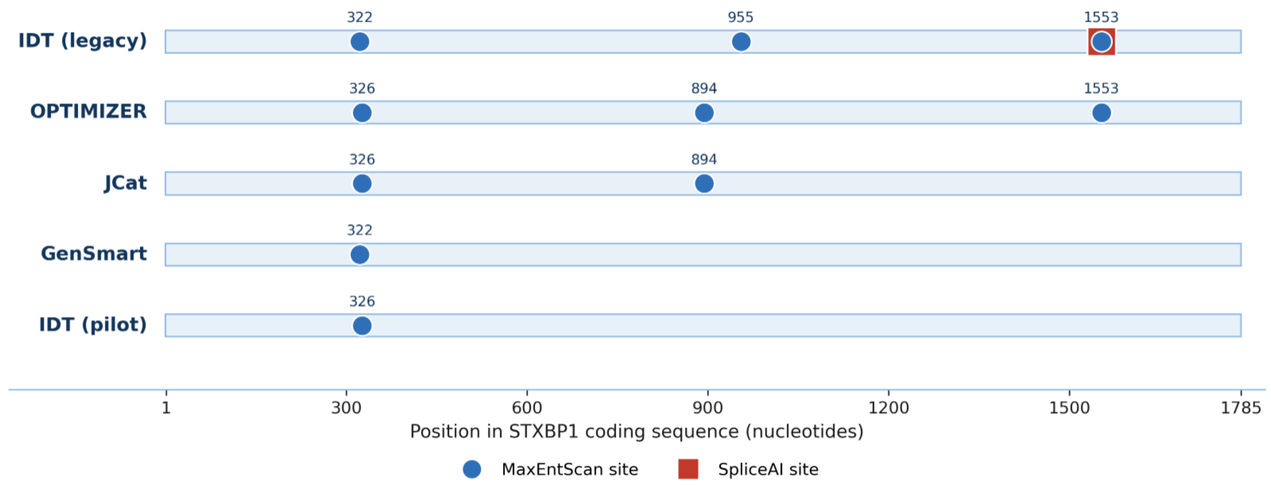

Figure 2. Cryptic splice sites created by codon optimisation softwares across the STXBP1 coding sequence.

Cryptic site	Created by	Reading frame	Consequence to the Munc18-1 protein
Acceptor at 322	IDT (legacy), GenSmart	Frameshift	Stop at residue 44. Domains 1, 2, 3a and 3b are lost.
Acceptor at 326	JCat, OPTIMIZER, IDT (pilot)	Frameshift	Aberrant protein sequence from residue 22. Domains 1, 2, 3a and 3b are lost.
Donor at 895	JCat, OPTIMIZER	In-frame	278 residues deleted (299 to 576). Most of domain 3a and all of domain 3b are lost.
Donor at 956	IDT (legacy)	Frameshift	Stop at residue 329. Domain 3a and domain 3b are lost.
Donor at 1554	IDT (legacy), OPTIMIZER	Frameshift	Aberrant protein sequence from residue 518. The end of domain 3b is lost.

Table 2. Predicted protein consequence to the Munc18-1 protein if each cryptic splice site were used.

- Modelling of the 10 sites predicted **7 frameshift and 2 in-frame deletions** of the protein sequence (Table 2).
- All sites were successfully eliminated where most required only a single synonymous nucleotide change. Re-screening confirmed no new cryptic sites were introduced (Table 3).

Optimisation tool	Rescued site	MaxEntScan score (before to after)	Codon change
IDT (legacy)	Donor at 1554	10.88 → 2.69	AGG → AGA at codon 518
OPTIMIZER	Donor at 1554	9.77 → 1.27	CGG → CGT at codon 518
IDT (legacy)	Donor at 956	8.72 → 0.54	GGT → GGA at codon 319
GenSmart	Acceptor at 322	8.58 → 0.62	ACA → ACG at codon 107
IDT (pilot)	Acceptor at 326	9.67 → −6.76	AGC → TCC at codon 109
JCat and OPTIMIZER	Acceptor at 326	9.03 → −7.40	AGC → TCC at codon 109
JCat and OPTIMIZER	Donor at 895	8.70 → −2.65	GTG → GTC at codon 299; GAG → GAA at codon 298
IDT (legacy)	Acceptor at 321	9.32 → 0.96	ACA → ACT at codon 107

Table 3. Correction of each cryptic splice site by a synonymous codon change.

CONCLUSIONS

- In codon optimisation, tool choice may influence splicing risk, as every tool created at least one cryptic splice site but the number and position varied.
- However, the risk is likely easily checked and easily resolvable with only a single nucleotide change.
- This risk did not correspond to overall GC content, so it is the location of the codon changes rather than the sequence's overall GC-richness that matters.
- **Limitations:** These results are computational predictions that need experimental *in-vitro* and *in-vivo* validation. It is also unclear if these findings generalize across transgenes.

References

1. Lopez-Rivera, J. A., Perez-Palma, E., Symonds, J. et al. A catalogue of new incidence estimates of monogenic neurodevelopmental disorders caused by de novo variants. Brain 143, 1099-1105 (2020).
2. Xian, J. et al. Assessing the landscape of STXBP1-related disorders in 534 individuals. Brain 145, 1668-1683 (2022).
3. Wang, Q.-H. et al. Efficacy of levetiracetam in STXBP1 encephalopathy with different phenotypic and genetic spectra. Seizure 95, 64-74 (2022).
4. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. 11, 377-394 (2004).
5. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. Cell 176, 535-548.e24 (2019).
6. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422-1423 (2009).

Note: All figures and tables were created by the author using BioRender and Matplotlib, or reproduced from sources licensed for reuse with appropriate attribution.

Acknowledgements
Sincere thanks to the Laidlaw Foundation for supporting this research, and to my supervisor, Dr Arianna Fozzato at the Wellcome Sanger Institute, for her guidance and incredible mentorship throughout this project!

Contact: gayatri.dhir.25@ucl.ac.uk