**Significance**

Background

Alternative splicing is a fundamental aspect of eukaryotic gene regulation and more than 90% of human genes are alternatively spliced.1 Splicing is the process of identifying introns and exons in a pre-mRNA sequence and removing the introns while stitching together the exons to form an mRNA. Errors in this process are involved in a wide variety of diseases including familial dysautonomia, early onset Parkinson disease, and cancer.2 Further, recent efforts have produced drugs that modulate splicing outcomes to treat spinal muscular atrophy, Huntington's disease, Duchenne muscular dystrophy, and cancer.3 There are four essential components that mark a potential new splice junction these are the 5’ splice site, branch point, polypyrimidine tract, and 3’ splice site. The branch point and polypyrimidine tract are contained in the intron while the 5’ and 3’ splice sites mark the boundaries between the intron and the 5’ and 3’ exons respectively. Splicing is initiated by binding of the U1 snRNP to the 5’ splice site, binding of SF1 to the branch point, and binding of U2AF to the polypyrimidine tract and 3’ splice site, these factors interact to form the E complex which then initiates splicing. In humans, the initial interaction of these factors frequently occurs across the exon in a process called exon-definition, but the interaction then transfers to cross the intron that will be removed before splicing occurs. This complex then recruits additional spliceosome factors and undergoes a series of rearrangements to yield an active complex. In this process, the U2 snRNP replaces SF1 and U2AF to bind the branch point and U1 is replaced by the U6 snRNP. The actual splicing reaction then occurs as a sequence of two nucleophilic attacks, severing first the 5’ exon and then the 3’ exon from the intron and sealing the two exons together.1 In humans, however while the strength of a splice junction can be controlled by adjusting the affinity of these regulatory sequences for their respective spliceosomal factors, much finer regulation of splicing is required to allow dynamic changes to isoform populations. This is achieved through the binding of proteins called splicing factors to motifs called splicing enhancers or suppressors that may be present in the introns or the exons of the pre-mRNA. These factors affect the recruitment of the spliceosome components and allow specific isoform changes to be tightly regulated through changes in splicing factor abundance or activity.2 Additionally, RNA secondary structure changes can dramatically affect which isoforms are produced.3,4 How these influences integrate to produce the final splicing outcome is only poorly understood. A major goal of the splicing field has been

Deep learning has

Massively parallel splicing assays (MPSAs) offer the potential to rapidly profile the regulatory mechanisms of splicing in genes like PKM. These assays use high-throughput sequencing to measure quantitative splicing phenotypes for 10s of thousands of variants in parallel.14–28 They work by linking the sequence of the pre-mRNA to a barcode and then either compare barcode abundances between two conditions as a proxy for splicing14,20,23–25,27–29 or directly sequence the splice junction of interest and the barcode to observe the isoform distribution of each variant.16,18–20,22,26 These assays have allowed identification of human sequence variants that cause splicing changes,24 in depth study of how drugs interreact with splicing regulation,28 and detailed mechanistic studies on splicing regulations.14,19,20,23,25 However, MPSAs have been held back by two barriers to progress.

Barrier to progress 1: …

Barrier to progress 2: …

Barrier to progress 3: …

*This proposal will overcome these barriers by …*

THE KEY ADVANCE FOR EACH SECTION (underline the actual key advance in each paragraph)

We propose pursuing these goals through 3 complementary aims. Aim 1…. Aim 2 …. Finally, aim 3 ...

Innovation

Our work will enable broader use of MPSAs to interrogate the mechanisms of splicing by extending the technique to work in the context of complex isoform distributions and correlated splicing outcomes. It will further enable wider use by providing end to end analysis and modeling software for MPSA data. This work is driven by innovations that will change the practices of the MPSA field and the broader splicing community.

**Innovation 1: …** intro **…** **Aim 1** **…** and **aim 2** **…**

**Innovation 2: …** i

**Innovation 3: …** i

**Approach**

**Aim 1: Develop a foundation model for splicing that integrates data from diverse genomics datasets.**

Goal: ….

Approach: ….

Pitfalls and alternative approaches: ….

**Aim 2: Design and perform targeted MPSAs to improve models of splicing.**

Goal: ….

Approach: ….

Pitfalls and alternative approaches: ….

**Aim 3: Computationally extract and experimentally test mechanistic hypothesis suggested by splicing models.**

Goal: ….

Approach: ….

Pitfalls and alternative approaches: ….

1. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

2. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2015).

3. Neil, C. R. *et al.* Reprogramming RNA processing: an emerging therapeutic landscape. *Trends Pharmacol. Sci.* **43**, 437–454 (2022).

4. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).

5. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–813 (2008).

6. Jin, Y., Dong, H., Shi, Y. & Bian, L. Mutually exclusive alternative splicing of pre‐mRNAs. *Wiley Interdiscip. Rev. RNA* **9**, e1468 (2018).

7. Xu, B., Meng, Y. & Jin, Y. RNA structures in alternative splicing and back-splicing. *Wiley Interdiscip. Rev. RNA* **12**, e1626 (2021).

8. Hatje, K. *et al.* The landscape of human mutually exclusive splicing. *Mol. Syst. Biol.* **13**, 959 (2017).

9. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**, 736–742 (2015).

10. Tilgner, H. *et al.* Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* **28**, 231–242 (2018).

11. Zhu, C. *et al.* Single-molecule, full-length transcript isoform sequencing reveals disease-associated RNA isoforms in cardiomyocytes. *Nat. Commun.* **12**, 4203 (2021).

12. Hemani, Y. & Soller, M. Mechanisms of Drosophila Dscam mutually exclusive splicing regulation. *Biochem. Soc. Trans.* **40**, 804–809 (2012).

13. Smith, C. W. J. Alternative splicing--when two’s a crowd. *Cell* **123**, 1–3 (2005).

14. Ke, S. *et al.* Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* **28**, 11–24 (2018).

15. Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 1–8 (2016).

16. Adamson, S. I., Zhan, L. & Graveley, B. R. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol.* **19**, 71 (2018).

17. Soemedi, R. *et al.* Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* **49**, 848–855 (2017).

18. Cortés-López, M. *et al.* High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling CD19 splicing and CART-19 therapy resistance. *Nat. Commun.* **13**, 1–17 (2022).

19. Schirman, D., Yakhini, Z., Pilpel, Y. & Dahan, O. A broad analysis of splicing regulation in yeast using a large library of synthetic introns. *PLoS Genet.* **17**, e1009805 (2021).

20. Mikl, M., Hamburg, A., Pilpel, Y. & Segal, E. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat. Commun.* **10**, 1–14 (2019).

21. Braun, S. *et al.* Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nat. Commun.* **9**, 1–18 (2018).

22. Souček, P. *et al.* High-throughput analysis revealed mutations’ diverging effects on SMN1 exon 7 splicing. *RNA Biol.* **16**, 1364–1376 (2019).

23. Baeza-Centurion, P., Miñana, B., Valcárcel, J. & Lehner, B. Mutations primarily alter the inclusion of alternatively spliced exons. *Elife* **9**, (2020).

24. Cheung, R. *et al.* A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Mol. Cell* **73**, 183-194.e8 (2019).

25. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell* **176**, 549-563.e23 (2019).

26. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).

27. Wong, M. S., Kinney, J. B. & Krainer, A. R. Quantitative Activity Profile and Context Dependence of All Human 5’ Splice Sites. *Mol. Cell* **71**, 1012-1026.e3 (2018).

28. Ishigami, Y. *et al.* Specificity, synergy, and mechanisms of splice-modifying drugs. *bioRxiv* 2022.12.30.522303 (2022) doi:10.1101/2022.12.30.522303.

29. North, K. *et al.* Synthetic introns enable splicing factor mutation-dependent targeting of cancer cells. *Nat. Biotechnol.* **40**, 1103–1113 (2022).

30. Wang, Z. *et al.* Exon-centric regulation of pyruvate kinase M alternative splicing via mutually exclusive exons. *Journal of molecular* (2012).

31. Wang, Z., Jeon, H. Y., Rigo, F., Bennett, C. F. & Krainer, A. R. Manipulation of PK-M mutually exclusive alternative splicing by antisense oligonucleotides. *Open Biol.* **2**, 120133 (2012).

32. Piovesan, A. *et al.* Human protein-coding genes and gene feature statistics in 2019. *BMC Res. Notes* **12**, 315 (2019).

33. Kalinina, M. *et al.* Multiple competing RNA structures dynamically control alternative splicing in the human ATE1 gene. *Nucleic Acids Res.* **49**, 479–490 (2021).

34. Christofk, H. R. *et al.* The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452**, 230–233 (2008).

35. Ma, W. K. *et al.* ASO-based PKM Splice-switching Therapy Inhibits Hepatocellular Carcinoma Cell Growth. *Cancer Res.* 2020.09.01.278580 (2022) doi:10.1158/0008-5472.CAN-20-0948.

36. Clower, C. V. *et al.* The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism. *Proceedings of the National Academy of Sciences* **107**, 1894–1899 (2010).

37. Luo, W. *et al.* Pyruvate kinase M2 is a PHD3-stimulated coactivator for hypoxia-inducible factor 1. *Cell* **145**, 732–744 (2011).

38. Yang, W. *et al.* Nuclear PKM2 regulates β-catenin transactivation upon EGFR activation. *Nature* **480**, 118–122 (2011).

39. Traxler, L. *et al.* Warburg-like metabolic transformation underlies neuronal degeneration in sporadic Alzheimer’s disease. *Cell Metab.* **34**, 1248-1263.e6 (2022).

40. Liu, C., Liu, C. & Fu, R. Research progress on the role of PKM2 in the immune response. *Front. Immunol.* **13**, 936967 (2022).

41. Tang, Q. *et al.* Pyruvate kinase M2 regulates apoptosis of intestinal epithelial cells in Crohn’s disease. *Dig. Dis. Sci.* **60**, 393–404 (2015).

42. Rihan, M. & Sharma, S. S. Role of Pyruvate Kinase M2 (PKM2) in Cardiovascular Diseases. *J. Cardiovasc. Transl. Res.* (2022) doi:10.1007/s12265-022-10321-1.

43. De Paoli-Iseppi, R., Gleeson, J. & Clark, M. B. Isoform Age - Splice Isoform Profiling Using Long-Read Technologies. *Front Mol Biosci* **8**, 711733 (2021).

44. de la Rubia, I. *et al.* RATTLE: reference-free reconstruction and quantification of transcriptomes from Nanopore sequencing. *Genome Biol.* **23**, 153 (2022).

45. Dierckxsens, N., Li, T., Vermeesch, J. R. & Xie, Z. A benchmark of structural variation detection by long reads through a realistic simulated model. *Genome Biol.* **22**, 342 (2021).

46. Stephens, Z. D. *et al.* Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. *PLoS One* **11**, e0167047 (2016).

47. Eliseev, A. *et al.* Evaluation of haplotype callers for next-generation sequencing of viruses. *Infect. Genet. Evol.* **82**, 104277 (2020).

48. Cosma, B.-M. *et al.* When do longer reads matter? A benchmark of long read de novo assembly tools for eukaryotic genomes. *bioRxiv* 2023.01.30.526229 (2023) doi:10.1101/2023.01.30.526229.

49. Voshall, A. & Moriyama, E. N. Next-generation transcriptome assembly and analysis: Impact of ploidy. *Methods* **176**, 14–24 (2020).

50. Walter, C., Schuetzmann, D., Rosenbauer, F. & Dugas, M. Benchmarking of 4C-seq pipelines based on real and simulated data. *Bioinformatics* **35**, 4938–4945 (2019).

51. Hardwick, S. A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* **13**, 792–798 (2016).

52. Hardwick, S. A. *et al.* Targeted, High-Resolution RNA Sequencing of Non-coding Genomic Regions Associated With Neuropsychiatric Functions. *Front. Genet.* **10**, 309 (2019).

53. Sheynkman, G. M. *et al.* ORF Capture-Seq as a versatile method for targeted identification of full-length isoforms. *Nat. Commun.* **11**, 2326 (2020).

54. Tan, C. C. S., Maurer-Stroh, S., Wan, Y., Sessions, O. M. & de Sessions, P. F. A novel method for the capture-based purification of whole viral native RNA genomes. *AMB Express* **9**, 45 (2019).

55. Escalona, M., Rocha, S. & Posada, D. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* **17**, 459–469 (2016).

56. Wick, R. Badread: simulation of error-prone long reads. *J. Open Source Softw.* **4**, 1316 (2019).

57. Hafezqorani, S. *et al.* Trans-NanoSim characterizes and simulates nanopore RNA-sequencing data. *Gigascience* **9**, (2020).

58. Ono, Y., Hamada, M. & Asai, K. PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genom Bioinform* **4**, lqac092 (2022).

59. Lau, B. *et al.* LongISLND: in silico sequencing of lengthy and noisy datatypes. *Bioinformatics* **32**, 3829–3832 (2016).

60. Zhang, W., Jia, B. & Wei, C. PaSS: a sequencing simulator for PacBio sequencing. *BMC Bioinformatics* **20**, 352 (2019).

61. Schmeing, S. & Robinson, M. D. ReSeq simulates realistic Illumina high-throughput sequencing data. *Genome Biol.* **22**, 67 (2021).

62. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).

63. Al Kadi, M. *et al.* UNAGI: an automated pipeline for nanopore full-length cDNA sequencing uncovers novel transcripts and isoforms in yeast. *Funct. Integr. Genomics* **20**, 523–536 (2020).