**Significance**

Alternative splicing is important for both understanding fundamental aspects of human biology and treating a wide array of diseases. Splicing is a process where introns are removed from pre-mRNAs and the exons are stitched together. This offers the potential to introduce tremendous diversity into the proteome by changing which sequences from the pre-mRNA are included as exons. Accordingly, more than 90% of human genes are alternatively spliced.1 Errors in splicing are involved in a wide variety of human diseases including familial dysautonomia, early onset Parkinson disease, and cancer.2 Recent efforts have produced drugs that modulate splicing outcomes to treat spinal muscular atrophy, Huntington's disease, Duchenne muscular dystrophy, and cancer.3 However, despite these recent advances, mechanisms of splicing regulation are still poorly understood. Improving our understanding of splicing will deepen our knowledge of development and differentiation and our ability to target these processes with therapeutics.

A particularly interesting type of alternative splicing is mutually exclusive exons. Mutually exclusive exons are clusters of exons whose splicing is anticorrelated such that one exon from the cluster is selected for inclusion in the final transcript. Mutually exclusive exons are important for regulating a wide variety of processes including cell fate determination, neural development, and muscle development.4 Furthermore, mutations in mutually exclusive exons or changes in their splicing are involved with diseases including Timothy syndrome, cancer, heart disease, and cardiomyopathy.4 In fact, mutually exclusive exons are twice as likely to be associated with a parthenogenic SNP in clinvar.5 These SNPs are most frequently associated with neurological diseases, neuromuscular diseases, cardiomyopathies, or cancer.5 While mutually exclusive exon splicing has long been thought to be uncommon, recent studies have uncovered far more instances that were previously thought.5 Research on mutually exclusive exons has uncovered several mechanisms that drive mutual exclusivity, including Spliceosome incompatibility, steric hindrance, and RNA secondary structure.4 However, despite these advances, more than 75% of human mutually exclusive exon clusters have no known mechanism.5

Pyruvate kinase M (PKM) is a model for mutually exclusive exon splicing with deep disease relevance. This gene has 2 primary isoforms characterized by a switch between inclusion of either exon 9 (PKM1) or 10 (PKM2).6 PKM1 is constitutively active and is the dominant form in energy hungry tissues like muscle and brain, while PKM2 is allosterically regulated and is dominant in proliferating or embryonic tissues.7–11 The allosteric regulation of PKM2 causes it to transition from a tetrameric form which is competent as a pyruvate kinase to a dimeric form that is inactive as a pyruvate kinase.8 This dimeric form however, gains activity as a protein kinase and translocates into the nucleus where it phosphorylates stat3, transactivates β-catenin, and acts as a coactivator for HIF-1.12–14 Cancer is associated with increases in the expression of the PKM2 isoform, and with increased levels and nuclear localization of its dimeric form specifically.7–10,12 The presence of PKM2 instead of PKM1 is necessary for the Warburg effect and is important for tumorigenesis.9,10,13–15 A PKM1 to PKM2 transition also drives neuronal fate loss and cell death in Alzheimer’s disease through both metabolic and regulatory pathways.16 PKM2’s activity as a transcriptional regulator is highly important for the normal functioning of the immune system and loss of PKM has been linked to repeated infections while overexpression of PKM2 has been observed Crohn’s.17–19 Further, overexpression of PKM2 relative to PKM1 has been observed in cardiovascular diseases including myocardial infarction, pulmonary arterial hypertension, atherosclerosis and cardiac hypertrophy.20 While much work has been done mapping the regulatory motifs that control the choice between exon 9 and exon 10 in PKM, no mechanism for maintaining mutual exclusivity between the exons has been demonstrated.10,21–23 Understanding these mechanisms would both deepen our understanding of an important gene and provide insight into the many diseases it is linked with.

Massively parallel splicing assays (MPSAs) provide a window into splicing regulation but are limited by read length. In recent years MPSAs have been used to interrogate the mechanisms of splicing.24–37 These methods allow rapid mechanistic characterization by measuring quantitative splicing phenotypes for as many as 10s of thousands of variants in a single assay.24–37 MPSAs have been designed using a variety of different techniques. Many MPSAs are isoform specific and look to read out splicing as a binary choice between two known isoforms. These include methods based on linking outcome to protein expression, using fluorescent proteins30,34 or counter selectable markers.38 These MPSAs suffer from an inability to distinguish unexpected isoforms and they require very non-native contexts which can affect splicing outcomes. SOMETHING HERE ABOUT THE IMPORTANCE OF NATIVE INTRON CONTEXTS. Another class of isoform specific MPSA uses an RT-PCR and amplicon sequencing as a readout. These work either by using isoform specific primers and comparing between primer sets,37 or by gel extracting the isoform of interest and assuming RNA expression level is constant for all variants.24,33,35 These assays can occur in more native contexts than the protein based screens, but still suffer from an inability to account for isoform diversity. The restriction to quantifying known isoforms has two significant downsides that stem from the fact that it is common for there to be multiple alternative isoforms that can arise from mutations. The first is that by ignoring this diversity, a large amount of information is lost that could shed insight into the mechanisms of splicing decisions. The second, is that unexpected alternative isoforms are frequently incorrectly counted as one of the expected isoforms causing noise in the measurements of isoform frequencies. An alternative strategy that avoids this issue is to sequence over the splice junction of interest with short-read sequencing and directly count all isoforms for each variant.26,28–30,32,36 This allows the quantification of all isoforms but requires that all isoforms only differ over a region small enough to fit in a short read. Since the median intron is 1.7 kb39 this means that to apply this technique to most splice junctions the experiment needs to be performed in a non-native, small-intron context or isoforms that retain some or all off the intron cannot be accurately quantified. Additionally, isoforms that loose one of the primer sites can not be quantified. Finally, none of these techniques utilize double sided barcodes, which leaves them open to isoform miscalls caused my PCR or RT template exchange events. SOMETHING HERE ABOUT THE DATA ANALYSIS PIPELINES. Because of these issues, current MPSAs are not well suited to interrogating the mechanisms of complicated splicing decisions in the context of the native introns, where important isoforms may be large and multiple splicing outcomes are expected. New techniques will be needed to fully interrogate the mechanisms of mutually exclusive exon clusters.

The goal of my proposed first aim is to create MPSA methods that overcomes previous technical limitations. I will do this by creating a long read based MPSA (LR-MPSA) technique. The LR-MPSA will enable performing splice junction sequencing in more native contexts by making it possible to sequence over constructs with full size introns without ignoring intron included contexts. Full isoform resolution with native introns will expand the window we have into the biology and reduce noise caused by isoform misassignment. Reducing isoform misassignment noise will increase dynamic range and accuracy of the assay. The LR-MPSA will allow analysis of more complicated splicing situations that are expected to produce multiple minor isoforms or isoforms too large to fit in a short read. This includes the study of systems with intron retention or cryptic splice sites as well as mutually exclusive exons and other forms of correlated splicing outcomes.

The goal of my second aim is to create robust open-source software for analyzing isoform resolution MPSA data including LR-MPSA data. This is significant because it will enable wider use of MPSA methods by automating read processing, isoform identification, isoform quantification, and data visualization. Further, this software will be robustly tested on synthetic datasets to ensure accuracy and robustness, an improvement over previous bespoke analysis scripts.

The goal of my third aim is to apply isoform resolution MPSA techniques to discover the mechanisms of mutually exclusive splicing regulation in PKM. This is significant because it will provide insight into the regulation of PKM which has biological relevance across a wide range of human biology and links to human health in contexts including cancer, Alzheimer’s disease, Crohn’s disease, and cardiovascular disease. Previous studies have investigated the regulation of PKM through low throughput methods and have not focused on identifying mechanisms that maintain mutual exclusivity between exons. The proposed study will be capable of being much more thorough and will detect not just regulatory elements that control exon choice, but also elements that affect mutual exclusivity. This will expand our understanding of a crucial regulatory event that is implicated in a wide array of diseases and may provide insight into new mechanisms for creating mutually exclusive exons.

**Approach**

**Aim 1: Develop a nucleotide-resolution MPSA using long read sequencing.**

Goal: Create a long read MPSA (LR-MPSA) that can produce nucleotide resolution isoform information of complex splicing systems in the context of native introns. This will involve developing sequencing library preparation techniques and sequencing strategies that allow generation of the highest quality long read data while minimizing potential sources of bias. This will also involve benchmarking of our methods.

Approach: We will trial different library preparation techniques for generating high quality long read RNA-seq data on variant libraries. In order to do this, we need test libraries with well characterized behavior. For this we will use a 5’-splice site library that our lab has previously developed in the gene SMN2. If further verification is needed, we can also expand our testing to similar libraries we developed in BRCA2, and IKBKAP. These libraries were originally characterized with a isoform specific RT-PCR and sequencing protocol that utilized 3’ barcodes that had been matched to variants in an earlier DNA sequencing experiment. We will regenerate libraries with both 5’ and 3’ barcodes to allow detection and elimination of any reads that derive from template exchange or recombination events during library preparation. This should allow reduction of noise that derives from incorrect isoform assignment due to crossovers between the splice junction and the barcode. We will match barcode pairs to variant sequences by nanopore direct ligation DNA sequencing of a restriction fragment which contains all three components. This will eliminate any possibility of crossover during the original mapping experiment. We will then generate spliced library RNA in HeLa cells using well established techniques in the lab, and optimize library production. We will use an RT-PCR library preparation technique protecting against PCR amplification bias using barcodes introduced in the RT step and against template exchange using the dual barcoding strategy. We will then use nanopore sequencing to generate long reads. We will bench mark this library preparation technique by comparing variant abundance and isoform distribution against established short read methods of junction sequencing and isoform specific PCR sequencing. We will also benchmark against nanopore direct RNA sequencing. I will also test these methods with defined RNA mixtures with known ratios of different alternate isoforms called RNA sequins.40

Pitfalls and alternative approaches: If we are unable to eliminate significant bias from RT or PCR in our library preparation, we could use oligo based direct RNA capture and nanopore direct RNA sequencing or direct cDNA sequencing. If nanopore sequencing proves to produce error rates that are two high, we can use PacBio sequencing or use combined nanopore and illumina reads.

**Aim 2: Develop software for analyzing data from nucleotide-resolution MPSAs**

Goal: I propose to develop open-source software for analyzing nucleotide resolution MPSA datasets. This will help MPSA techniques become more available to researchers by automating the data analysis procedures for MPSAs

Approach: In order to create MPSA analysis software that is accurate and robust, I will create simulated MPSA data to use as a test data. Since this data is simulated, I will know the true underlying isoform distribution for each variant and can evaluate the performance of different analysis strategies. I will generate simulated data for different genes to ensure that results are generalizable. I will also vary what kind of variant diversity is added and simulated library prep procedures. I will simulate reads for Illumina, PacBio, and ONT sequencing platforms using existing read simulation packages.41–47 I will test different data cleaning, read clustering, isoform identification, and alignment pipelines to determine the accuracy and pitfalls of each technique on this data type. I will test existing isoform identification tools.48 Once an effective MPSA isoform identification strategy is identified, I will write isoform quantification and data visualization scripts and test them on the simulated datasets. I will provide the resultant pipelines as a python package and as independent software to make it available to other scientists who want to use MPSA methods.

Pitfalls and alternative approaches: It is possible that nanopore reads will be too error prone to successfully identify isoforms, in that case we could use PacBio sequencing or explore combined illumina and nanopore reads. It is possible existing read simulators and isoform clustering tools will not work well for MPSA data because they were designed for slightly different purposes, if this is the case, I will write custom versions that are optimized for the MPSA use case.

1. ~~Develop open-source nucleotide resolution MPSA analysis software~~
   1. ~~Simulate reads from libraries of variants with different isoform mixtures~~
   2. ~~Compare different alignment and isoform calling algorithms~~
   3. ~~Create software to process reads, identify isoforms, quantify isoforms, and visualize data~~

**Aim 3: Define the mechanism of mutually exclusive exon inclusion in PKM**

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

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

6. Noguchi, T., Inoue, H. & Tanaka, T. The M1- and M2-type isozymes of rat pyruvate kinase are produced from the same gene by alternative RNA splicing. *J. Biol. Chem.* **261**, 13807–13812 (1986).

7. Taniguchi, K. *et al.* Organ-specific PTB1-associated microRNAs determine expression of pyruvate kinase isoforms. *Sci. Rep.* **5**, 1–8 (2015).

8. Mazurek, S., Boschek, C. B., Hugo, F. & Eigenbrodt, E. Pyruvate kinase type M2 and its role in tumor growth and spreading. *Semin. Cancer Biol.* **15**, 300–308 (2005).

9. 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).

10. 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).

11. Jurica, M. S. *et al.* The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate. *Structure* **6**, 195–210 (1998).

12. Gao, X., Wang, H., Yang, J. J., Liu, X. & Liu, Z.-R. Pyruvate kinase M2 regulates gene transcription by acting as a protein kinase. *Mol. Cell* **45**, 598–609 (2012).

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

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

15. 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.

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

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

18. Burge, P. S., Johnson, W. S. & Hayward, A. R. Neutrophil pyruvate kinase deficiency with recurrent staphylococcal infections: first reported case. *Br. Med. J.* **1**, 742–745 (1976).

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

20. 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.

21. David, C. J., Chen, M., Assanah, M., Canoll, P. & Manley, J. L. HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature* **463**, 364–368 (2010).

22. 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).

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

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

25. 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).

26. 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).

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

28. 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).

29. 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).

30. 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).

31. 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).

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

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

34. 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).

35. 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).

36. 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).

37. 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).

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

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

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

41. 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).

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

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

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

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

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

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

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