**Significance**

Background

A major outstanding question in the field of splicing is how are the parts of a pre-mRNA to include in the final transcript determined. **Massively parallel splicing assays (MPSAs)**, which use high-throughput sequencing to measure quantitative splicing phenotypes for 10s of thousands of variants in parallel, have revolutionized our ability to understand the regulatory mechanisms of splicing.1–15 These assays link the sequence of the pre-mRNA to a barcode and then either compare barcode abundances between two conditions as a proxy for splicing or directly sequence the splice junction of interest and the barcode to observe the isoform distribution of each variant. These assays have allowed large scale identification of human sequence variants that cause isoform changes,11 in depth study of how drugs interreact with splicing regulation,15 and detailed mechanistic studies on splicing regulations.1,6,7,10,12 However, MPSA studies have been held back by three barriers to progress.

Barrier to progress 1: Current MPSAs cannot report the full isoform diversity of alternative splicing. Many MPSAs rely on counting barcodes across two conditions as a proxy for splicing outcomes, but this forcibly binarizes the isoform data that is collected. It is common for there to be multiple alternative isoforms that can arise from mutations.5,9,16,17 Missing this diversity by forcibly binarizing the data with a proxy assay causes two problems. The first is that unexpected isoforms will be binned into one of the binary conditions causing inaccuracy and noise in the results. The second is that loosing data on the true isoform distributions of each variant misses out on information that is highly mechanistically informative. One solution to overcome this issue is to sequence the splice junction of interest. However, current methods use Illumina short-read sequencing which limits them to isoforms that are small enough to be covered by a short read.3,5–7,9,13 Since the median human intron is 1.7 kb,18 this technique cannot distinguish isoforms that contain some or all of the intron for most human splice junctions unless the introns are cut down or replaced. However, intronic sequences can be highly important for determining splicing outcomes,16,17,19–21 so this strategy can reduce generalizability of the data to native contexts. Creating MPSA methods that are able to measure the full diversity of isoforms produced by alternative splicing will allow both higher quality data in simple splicing contexts and allow extension of MPSA methods to the study of more complicated splicing events.

Barrier to progress 2: Current MPSAs cannot handle mutually exclusive splicing. Recent work has revealed that correlated splicing is far more common than previously understood.22–25 The simplest and best characterized form of correlated splicing is known as mutually exclusive exon splicing. Mutually exclusive exons are clusters of exons where exactly one exon from the cluster is included 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.19 Furthermore, mutations in mutually exclusive exons or changes in their splicing are involved in diseases including Timothy syndrome, cancer, heart disease, and cardiomyopathy.19 Despite their importance to biological processes, more than 75% of human mutually exclusive exon clusters have no known mechanism of maintaining mutual exclusivity.22 Correlated splicing and mutually exclusive exons represent an important class of splicing events that has been largely inaccessible to MPSA techniques. This is because determining the mechanisms maintaining mutual exclusivity between exons relies on tracking the fate of multiple splice junctions that are at a remove for each other. This carries with it all of the difficulties with tracking diverse isoforms that might result from an MPSA of one splice junction (see barrier to progress 1) with the added complication that several splice junctions need to be tracked. To make things worse, in some cases these junctions might not even be adjacent, and in some cases can be at quite a distance and have other exons in between.22–25 This sort of splicing phenomenon where the fate of more than one exon is important can also easily generate isoforms with no included introns that are too large to be covered by an illumina read even in the simplest cases of two short exons close together.16 Creating new MPSA methods that are capable of handling these more complicated systems will open up large numbers of mutually exclusive exon clusters and other correlated splicing events to rapid mechanistic characterization, and help us discover the missing mechanisms of mutual exclusivity.

Barrier to progress 3: There is no community MPSA analysis software. Current MPSA techniques have been independently developed and few experiments have the same experimental design or analysis methods. These factors make it hard to compare findings across multiple experiments and raise barriers to entry because the methods and analysis pipelines both need to be redeveloped by each group. Further, analysis pipelines have not been tested against datasets with known ground truth. Stronger vetting would increase confidence in the robustness and accuracy of these methods. Finally, MPSA data is a rich dataset that holds information useful for modeling the mechanisms of splicing but there is no tool set that integrates MPSA data analysis and modeling. Development of a user-friendly, publicly available, robustly verified, and open-source MPSA toolkit will greatly ease wider use of the method and help unify different techniques.

*This proposal will overcome these barriers by developing new long read MPSA (LR-MPSA) methods and analysis software. These methods will be evaluated by testing on simulated data sets, a well characterized SMN2 5’-splice site library, and new PKM mutually exclusive exons libraries. The analysis software will be made freely available to the research community as software package.*

The LR-MPSA will enable performing splice junction sequencing in constructs with full size introns without ignoring intron inclusion isoforms. The LR-MPSA will also allow analysis of more complicated splicing situations that are expected to produce multiple minor isoforms and 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 key advance of this technique will be to use oxford nanopore sequencing that is capable of sequencing the full length of any isoform or un-spliced transcript.

Our analysis software will provide tools for end-to-end analysis of LR-MPSA data. This will include parsing Nanopore reads, clustering and aligning reads to identify isoforms, quantifying isoforms by library variant, and providing visualization functions and integration with modeling tools like MAVE-NN. The software will be rigorously vetted on simulated data and well ununderstood experimental systems. The key advances of this technique will be to rigorously verify performance, to integrate seamlessly with quantitative modeling packages, and to provide in an easy-to-use, widely-accessible software package.

These new techniques will be demonstrated through investigating the mechanism of producing mutual exclusivity between exons 9 and 10 of pyruvate kinase M (PKM). With the inclusion of exon 9, PKM becomes a constitutively active pyruvate kinase. With the inclusion of exon 10, PKM becomes sensitive to allosteric regulation that switches it from acting as a pyruvate kinase to acting as a protein kinase phosphorylating Stat3 and a transcriptional co-activator of HIF-1 and β-catenin. Increases in the level of the exon 10 included transcript have been associated with cancer,26–30 Alzheimer’s,31 Crohn’s,32,33 and cardiovascular disease.34 However, there has been no systematic analysis of the regulatory mechanisms that produce the mutually exclusive splicing behavior of PKM. This makes PKM a perfect test case for the new LR-MPSA and analysis software.

We propose pursuing these goals through 3 complementary aims. Aim 1 will develop the LR-MPSA experimental techniques. Aim 2 will develop the MPSA analysis software package. Finally, Aim 3 will apply these new tools to understanding the mechanisms of mutually exclusive splicing in PKM.

The goal of my first aim is to create MPSA methods that overcome previous technical limitations. I will do this by creating a long read based MPSA (LR-MPSA) technique. Achieving 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 the dynamic range and accuracy of the assay. The LR-MPSA will also allow analysis of more complicated splicing situations that are expected to produce multiple minor isoforms and 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 but 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 exon clusters.

Innovation

**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 high 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 techniques that are well established in our lab, and optimize library production protocols. We will use an RT-PCR library preparation technique protecting against PCR amplification bias using unique molecular identifiers introduced in the RT step. We will protect against template exchange using the dual barcodes. 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.35

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 capture36–38 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. If barcode tracking with nanopore is difficult because of high error rates we can lengthen the barcodes so that even error filled barcodes can be robustly matched.

**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 set. 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. I will simulate reads for Illumina, PacBio, and ONT sequencing platforms using existing read simulation packages,39–45 adding bias effects from library preparation procedures. I will test data cleaning, read clustering, isoform identification, and alignment pipelines to determine the accuracy and pitfalls of each technique on this data type starting with existing isoform identification tools.46–48 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 different applications, if this is the case, I will write custom versions that are optimized for the MPSA use case.

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

Goal: To make a map of the determinants of PKM exon 9 and exon 10 splicing at nucleotide resolution, with a focus on the elements that maintain the mutual exclusivity of these two exons.

Approach: I will start this map by performing a course grained sweep over the large introns before exon 9 and after exon 10. This course grained sweep will be performed by making a series of large deletions from each of these introns and measuring isoform distributions. Any deletion that causes a change in the isoform distribution will be marked as a region of interest for mapping in finer detail. I will also include the entire region from 100 bp upstream of exon 9 to 100 bp downstream of exon 10 in the fine mapping as previous papers have shown that there are important regulatory elements in these regions.16,17 In order to produce a nucleotide resolution fine mapping of regulatory elements, I will generate libraries containing every single-nucleotide substitution or deletion in the regions of interest. I will then measure quantitative phenotypes for them using the LR-MPSA technique from Aim 1 and the analysis software from Aim 2. This will allow us to identify the isoform distributions associated with each mutant, and to map putative regulatory elements at nucleotide resolution. This may include regulatory elements that affect the ratio of PKM1 to PKM2 as well as mutations that increase the amount of the double included or double skipped isoforms. I will select regulatory elements to follow up with focusing on those that disrupt the mutual exclusivity of exons 9 and 10. Follow up will include low throughput experiments targeting regulatory motifs with 2-O-methoxyethyl modified antisense oligonucleotides. These antisense oligonucleotides bind RNA tightly and sterically block binding by splicing factors without triggering degradation of the transcript. They have worked well in the PKM context in previous studies.17 These experiments will allow us to confirm the effects of blocking identified motifs. If the identified regulatory elements appear to be RNA structural components, follow up will include creating mutant cycles to evaluate the importance of base pairing. If they appear to be protein binding motifs follow up will include pull down experiments to observe binding partners. I will use this information to inform a mechanistic model of splicing regulation in PKM exons 9 and 10. These efforts will allow us to build a high-resolution map of the determinants of PKM splicing and to identify the regulatory elements and mechanisms that enforce the mutual exclusivity of exons 9 and 10.

Pitfalls and alternative approaches: If the high PKM2/PKM1 ratio in HeLa cells makes it hard to distinguish mutant phenotypes, we can perform the experiment in a cell line that normally has an even mixture of the two isoforms such as C2C12 cells differentiated into myotubes.28 Alternatively, we can produce a more even isoform mixture in HeLa cells by strengthening the exon 9 5’-splice site16 or using antisense oligos.17 If the LR-MPSA from Aim 1 proves infeasible, we can proceed with this aim using a short-read sequencing approach which should still be acceptable though we will lose the ability to detect some isoforms. If the isoform analysis software from Aim 2 proves infeasible, we can align reads to a list of expected isoforms. If we end up only identifying known regulatory sites, we will still have produced a higher resolution map of PKM splicing than exists currently and we can refocus onto other mutually exclusive exon clusters such as those in ketohexokinase, CaV1.2, or CD55.

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

2. 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).

3. 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).

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

5. 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).

6. 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).

7. 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).

8. 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).

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

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

11. 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).

12. 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).

13. 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).

14. 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).

15. 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.

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

17. 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).

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

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

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

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

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

23. 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).

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

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

26. 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).

27. 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.

28. 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).

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

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

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

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

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

34. 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.

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

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

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

38. 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).

39. 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).

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

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

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

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

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

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

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

47. 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).

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