**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 as well as improve 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 where 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.4 Furthermore, mutations in mutually exclusive exons or changes in their splicing are involved in diseases including Timothy syndrome, cancer, heart disease, and cardiomyopathy.4 In fact, mutually exclusive exons are twice as likely to be associated with parthenogenic variants in clinvar as exons generally.5 These variants are most frequently associated with neurological diseases, neuromuscular diseases, cardiomyopathies, or cancer.5 While mutually exclusive exon splicing has long been considered 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 PKM2 regulation acts through changes in oligomerization state between a tetrameric form which is competent as a pyruvate kinase and 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, the mechanism for maintaining mutual exclusivity between the exons has not been demonstrated and a comprehensive screen for splicing regulatory elements is lacking.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.

Current massively parallel splicing assays (MPSAs) provide a window into splicing regulation but cannot capture the full isoform diversity of variants in their regulatory context. In recent years, MPSAs have been used to interrogate the mechanisms of splicing.24–38 These methods allow rapid mechanistic characterization by measuring quantitative splicing phenotypes for as many as 10s of thousands of variants in a single assay.24–38 MPSAs have been designed using a variety of different techniques. These broadly split into two categories, methods that measure a proxy of splicing and methods that directly sequence the splice junction of interest. However, neither strategy is capable of capturing the full diversity of splicing outcomes in the context of the native introns. It is common for there to be multiple alternative isoforms that can arise from mutations.22,23,28,32 Missing this diversity can cause two problems for an MPSA. The first is that a large amount of information that could provide insight into the mechanisms of splicing regulation is lost. The second, is that unexpected isoforms can be incorrectly counted as an expected isoform causing increasing error in the data. Loss of the regulatory context is also problematic because sequences in both the introns and exons can be important regulators of alternative splicing both affecting exon inclusion probability and establishing mutual exclusivity between exons.4,22,23,39,40

Proxies of splicing allow performing an MPSA without needing to sequence over the splicing junction but loose isoform information. This form of MPSA includes techniques based on engineering an intron so that splicing restores the reading frame of a fluorescent protein30,34 or counter selectable marker.41 Splicing can then be measured by using massively parallel variant assay techniques that measure protein function, and assuming that there are only two isoforms spliced and un-spliced. In addition to compressing all isoform diversity into a binary choice, this requires embedding the splice junction in a marker protein and the loss of the local regulatory environment. Proxy based MPSAs also include strategies that use RT-PCR and variant counting of known isoforms either based on isoform specific primers,37,38 or gel extracting the isoform of interest.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.

Direct sequencing of splicing junctions offers a potential solution, but is hindered by short reads. Direct sequencing of splicing junctions uses RT-PCR to amplify the region including the splice junction of interest and a variant barcode. Sequencing is then used to identify and quantify the observed isoforms and match them to the variant that produced them using the barcode.26,28–30,32,36 This technique allows unbiased quantification of isoforms. However, current methods use Illumina short-read sequencing which limits them to isoforms that are small enough to be covered by a short read.26,28–30,32,36 Since the median human intron is 1.7 kb,42 this technique cannot distinguish isoforms that contain some or all of the intron for most human splice junctions unless the native introns are replaced with smaller non-native introns. Additionally, in mutually exclusive exon contexts even isoforms with all introns spliced out but with multiple exons included can become too large to be covered by an Illumina read. For example, the PKM mini gene used to study mutually exclusive splicing of exons 9 and 10 is more than 6 kb including the introns and the isoform containing both exons 9 and 10 is too long to fit on an illumina read if you include a variant barcode and an RT-PCR primer binding site. Both double exon inclusion and cryptic intronic splice sites have been observed in experiments with this construct.23 These space constraints get even tighter if you consider using barcoding based techniques to help reduce artifacts from library preparation such as including 5’ and 3’ barcodes to eliminate PCR recombination events or unique molecular identifiers to remove amplification bias and accordingly, current MPSAs have not used these techniques.26,28–30,32,36

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 methods. They also raise barriers to entry for scientists looking to use the method because the methods and analysis pipelines both need to be redeveloped. Further, analysis pipelines have not been verified against datasets with known ground truth. Development of a user-friendly, publicly available, robustly verified, and open-source MPSA analysis toolkit will greatly ease wider use of the method and help unify different techniques.

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.43

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. 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 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. I will simulate reads for Illumina, PacBio, and ONT sequencing platforms using existing read simulation packages,44–50 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.51–53 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, specifically the elements that maintain mutual exclusivity of these two exons, at nucleotide resolution.

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 mapped in finer detail in the fine mapping step. 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.22,23 I will generate libraries containing every single nucleotide substitution or deletion in the regions of interest and 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 create a nucleotide resolution map of the isoform distributions caused by each mutant, and to map putative regulatory elements. 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 be look to specifically select regulatory elements to follow up with that disrupt the mutual exclusivity of exons 9 and 10. I will follow up on identified regulatory elements with low throughput mechanistic experiments. This includes targeting them with antisense oligonucleotides with 2-O-methoxyethyl modifications. These antisense oligonucleotides bind RNA tightly and sterically block binding by splicing factors without triggering degradation of the transcript, and have worked well in the PKM context in previous studies.22 It will also include making low throughput mutations, for instance to perform mutant cycle analysis on putative RNA structural elements discovered. It may also include pull down experiments to confirm the binding of particular RNA binding proteins to identified motifs. I will build models using the MAVE-NN framework developed by our lab54 and mechanistic modeling frameworks. Together these efforts will allow us to build a high resolution map of the determinants of PKM splicing and to identify the regulatory elements 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 myotubes10 alternatively, we can produce a more even isoform mixture in HeLa cells by strengthening the exon 9 5’-splice site to a consensus sequence23 or using antisense oligos.22 If the LR-MPSA form Aim 1 proves infeasible, we can proceed with this aim using a short read junction sequencing approach which should still be acceptable though we will lose the ability to detect some isoforms. If the all isoform detecting software from Aim 2 proves infeasible, we can aligning reads to a list of expected isoforms instead of trying to use clustering to identify isoforms. If we end up only identifying already 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 ketohexokinase, CaV1.2, or CD55.

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. 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.

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

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

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

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

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

44. 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).

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

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

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

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

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

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

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

52. 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).

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

54. Tareen, A. *et al.* MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biol.* **23**, 98 (2022).