1. My long term career goal is to lead a lab that focuses on using high-throughput techniques and computational methods to address biological problems. I value the ability to pursue fundamental questions but also want to see the answers to those questions make a difference in the lives of people. Because of this, I aim to work at a university or research institute where I can study both the fundamental and the applied sides of problems. I am broadly interested in asking questions about how information encoded in the sequence of biological molecules controls the phenotypes we observe and in working on applications in human health or climate. Through these topics, I hope to be able to both contribute to our understanding of biology and to apply that knowledge to real problems. However these topics are broad ones, so a short term career goal of mine is to refine this vision into a compelling lab focus. During my postdoc, I hope to accomplish this goal by exploring applications of high throughput screening and modeling in studying splicing regulation and how this regulation goes awry in cancer. Splicing is a good fit for my longer term goals because there are many fundamental questions to be answered about how regulators interact to tune the outcomes of splicing but treatments that modulate these processes to cure diseases are fast becoming a reality, opening the real prospect of new knowledge affecting the outcomes of patients. CSHL is the perfect place to pursue these goals as I can draw from the expertise of the Kinney Lab on high-throughput screens and computational modeling, the deep knowledge of the Krainer lab on the mechanics of splicing, and the wider environment of the cancer center to provide context and focus.

2. My research proposal relies on high throughput screens and modeling to probe how splicing is regulated in cancer. I have extensive experience with both of these techniques. During my Ph.D. I used high throughput screens to identify genes involved in the CO_2 concentrating mechanism in a chemoautotroph. I also used a mixture of high throughput screens and modeling to map the fitness landscape of dihydrofolate reductase and predict new functional variants. Finally, I applied kinetic modeling to design signal amplification strategies for cas13-based detection of RNA viruses. These experiences prepared me well to use high throughput screens and modeling to characterize splicing regulation. Additionally, I have taken courses directly applicable to the work I will be performing. During my Ph.D. I took the graduate level course "Fundamentals of Molecular and Cell Biology" which had a unit covering splicing, and many sections that touched on different aspects of cancer biology. Further, I took the upper division seminar "The Genetics of Cancer" during my undergraduate. This course included a deep dive into how regulation goes astray to cause cancer, as well as a lab component that included work with cancer cell lines. These courses will provide a firm base for my work on splicing in cancer and a foundation in the cell culture techniques I will need to perform screens in human cells. I also took a wide array of biology and computer science courses in my undergraduate as part of my molecular biology and biochemistry major and my computer science minor that will provide base knowledge for my work. Finally, if the project ends up requiring confirmatory experiments in mouse models, I have ample experience with mouse work from my time studying mouse spermatogenesis in the Ward lab in undergraduate and my time studying mouse photopigment expression in the Neitz lab while I was in high school.

3. I have three major training objectives in applying for the Cancer Gene Discovery and Cancer Biology Postdoctoral Training Program. These are to learn applications of my existing skills in human health and cancer, to strengthen techniques which I believe will be an important part of my future studies, and to develop the new skills I will need as an independent lab head. While I already have deep experience with creating libraries and performing screens, I hope to use this training program to learn new applications of these skills. I am particularly interested in exploring screens utilizing CRISPR based perturbations, human cell contexts, and single-cell long-read sequencing. Adding experience with these tools will help me to use my expertise in screening in new applications. I also want to strengthen techniques that will be important for my future work. This includes expanding my experience with human cell culture. The Kinney lab already has extensive experience with cell culture making this a great place for me to sharpen my skills and building this skill set will be hugely helpful for my ability to perform health related research. I also want to broaden my experience with modeling and more powerful data analysis techniques. While I have experience with these techniques, I haven't had the opportunity to work in a setting where this was the primary focus for my colleagues before. By working in the Kinney lab and in the CSHL Simons Center for Quantitative Biology I will be surrounded by researchers who focus primarily on computation, an optimal environment for me to grow my knowledge of modeling and data analysis techniques. Finally, I hope to use the CSHL Cancer Gene Discovery and Cancer Biology Postdoctoral Training Program to build the skills I will need to head a laboratory of my own. These skills include building my ability to fundraise for a lab by applying for fellowships and other grants. I will use the time I am in the program to apply for grants to fund the work and fellowships to extend my postdoctoral training after the end of the program.

4. One aspect of cancer research that I find particularly interesting is studying how regulation is altered in cancerous cells. This provides a window into how regulatory mechanisms work by demonstrating how they can be subverted. Since these insights are generated in the context of differences between cancer cells and healthy cells, there can be opportunities to apply what is learned to helping patients. In particular, I find the differential regulation of splicing interesting. Most human multi-exon genes have multiple isoforms,[1] and the isoform produced can change in cancer cells.[2–4] Moreover, these changes in splicing are important for disease progression.[2–4] While splice modifying drugs are beginning to be used to treat genetic disease, there is still a lot that is not known about how splicing is regulated. Understanding how cancer cells alter splicing outcomes to drive disease offers the potential to both learn fascinating fundamental biology and to learn how these changes can be disrupted to treat disease. In addition to satisfying my curiosity and providing windows into possible treatments for cancer, studying these aspects of cancer biology will help me develop a career as an independent cancer researcher by allowing me to explore applications of my expertise in high throughput screening and computational modeling to cancer research. Through this exploration, I can begin to build the research plan that will form the core of my lab in the future. Additionally, it will allow me to strengthen my experience with splicing, cancer genetics, and cell culture. Building these skills will help me create a strong foundation for continuing to research splicing in cancer. Finally, by working in the environment of the cancer center at CSHL surrounded by deep expertise on a wide array of topics in cancer, I can build an intuition for the open questions in cancer and splicing. This intuition will form the heart of a successful independent cancer research career.

5. When human genes are transcribed, the pre-mRNA produced contains both introns and exons.This pre-mRNA is spliced to remove the introns and stitch together the exons to produce mRNA. 95% of human multi-exon genes are spliced differentially to generate different mRNAs.[1] This process is regulated to produce programmed changes across cell types and developmental stages,[5] however, this regulation is subverted in cancer.[2–4] The conversion of Pyruvate Kinase M (PKM) from the adult isoform to the embryonic isoform in cancer drives the Warburg effect. This change is seen in a wide variety of cancers and has been shown to be important for lung cancer progression in xenograft studies.[2,4] Similarly, ketohexokinase (KHK) undergoes a switch from the C to A isoform in hepatocellular carcinoma driving lower fructose consumption and increased nucleotide synthesis.[3] This change is important in xenograft models and correlates with poor patient outcomes.[3]

Much of our current knowledge about splicing is from low throughput experiments interrogating small numbers of variants or from RNA-seq datasets which only report on natural changes in isoform abundance. However, new techniques in high throughput variant screening make it possible to assay many tens or hundreds of thousands of perturbations in parallel. This allows gathering information about the regulation of splicing events much more quickly. I am interested in using high throughput screening to study and model the mechanisms of splicing regulation in PKM and KHK. I seek to address the following questions: What sequences are important for regulating splicing in these genes? How do factors like splice site strength and splice enhancers or splice suppressors combine to control splicing? Finally, do these insights suggest mechanisms for altering splicing to treat cancer?

In order to answer these questions I propose a series of three experiments. First, I will expand the Kinney lab multiplex assay of variant effect (MAVE) for splicing outcomes to use single-cell, single-gene, long-read sequencing to quantify splicing isoforms. Second, I will use dCas13 guide tilling to map splicing regulatory sequences. Finally, I will use mini-gene mutant libraries to measure the influence of regulatory elements at much finer grain.

In past MAVE experiments, splicing efficiency was measured by amplification with junction specific primers followed by short read sequencing.[6] However, this technique is only capable of assaying for known splice variants. In recent years, single-cell long read techniques have been developed to enable splicing isoform characterization on the single cell level.[7,8] However these systems are unfocused, spending reads on all of the RNAs present in the cell. To make these methods amenable to the MAVE context, where much deeper sequencing is needed but the focus is on a single transcript, I will use primers designed to just amplify the relevant transcript instead of universal primers. This will allow unbiased quantification of the isoforms produced from a single gene in response to thousands of different perturbations.

With this new technique in hand, I will turn to mapping the regulatory elements that control splicing in PKM and KHK. One way to map regulatory elements is to systematically disrupt different segments of a gene while observing the process of interest. However, mutating a gene may introduce a new regulatory element instead of just removing an old one, complicating analysis. Stericaly occluding regulatory elements avoids this issue by maintaining the underlying

sequence. Steric occlusion of regulatory sites is used in the antisense oligonucleotide (ASO) class of splice modifying drugs. However, ASOs cannot be genetically encoded. This makes it difficult to use large libraries of ASOs to map regulatory elements. Cas13 offers a way around these issues. Cas13 is an RNA-guided nuclease, and it's nuclease inactivated form, dCas13, is an RNA-guided RNA-binding protein.[9] Targeting dCas13 to a transcript can alter its splicing by occluding regulatory sequences.[9,10] I will generate tiling libraries, where dCas13 is targeted to a sliding window along the gene sequence to map splicing regulatory elements in PKM and KHK. Change in isoform production will be measured using single-cell single-gene long read RNA-seq. I will repeat these experiments in both cancerous and noncancerous cells and across cancer types to look at how the regulatory landscape changes in these different contexts. The results of these assays can also be validated using synthesized ASOs. If the observed results correlate with those produced by ASO treatment, this could suggest that dCas13 based screens could be used as an easier alternative to ASO screens for ASO candidate nomination.

While guide tilling screens will provide a global look at splicing regulation, they don't allow measuring multiple different alleles of a single regulatory element. Further, they have limited resolution due to the length of the guide and homologous sequences in exons. To ask more targeted questions about the effects of particular regions on splicing regulation I will use a mini-gene assay. In this assay, a region of a gene containing the differentially spliced exons and their neighbors is cloned into a plasmid so it can be directly manipulated. This plasmid can then be transfected into the appropriate cell types and the splicing outcomes observed. I will use minigenes to make libraries with mutagenized regulatory elements such as 5'-splice sites, 3'-splice sites, branch points, or new regulatory elements identified through dCas13 tilling. By mutagenizing these elements, I will be able to see how their strengths affect splice isoform production. To gain even more precise estimates, I can assay minigene libraries in the context of a risdaplam or branaplam sensitive 5' splice site and measure changes in splicing outcomes for each library member as a function of the concentration of the small molecule. This will allow precise measurements of how different elements interact with smooth changes in 5' splice site strength. Similarly, the expression of known splicing factors can be changed or experiments can be performed across healthy and cancer cell lines to derive information about how these regulatory sequences interact with the cellular complement of splicing factors. This fine grained data on how regulatory elements and context control splicing outcomes will be used to make mechanistic models of splicing regulation.

In this proposal, I have set out a plan for using single-cell single-gene RNA-seq, dCas13 tilling screens, and minigene libraries to understand the regulation of splicing in PKM and KHK. Understanding the regulation of these genes offers a window into a deeper understanding of splicing in general and an opportunity to understand a unique aspect of cancer biology. By building a better understanding of how these genes are spliced we may gain insight into how to target these splicing events with splice modifying drugs to improve patient outcomes in a wide variety of cancers.

References

1. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
2. 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).
3. Li, X. *et al.* A splicing switch from ketohexokinase-C to ketohexokinase-A drives hepatocellular carcinoma formation. *Nat. Cell Biol.* **18**, 561–571 (2016).
4. Wang, Z. *et al.* Exon-centric regulation of pyruvate kinase M alternative splicing via mutually exclusive exons. *Journal of molecular* (2012).
5. Hatje, K. *et al.* The landscape of human mutually exclusive splicing. *Mol. Syst. Biol.* **13**, 959 (2017).
6. 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).
7. Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmable cDNA concatenation. *bioRxiv* 2021.10.01.462818 (2021) doi:10.1101/2021.10.01.462818.
8. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4259.
9. Konermann, S. *et al.* Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell* **173**, 665–676.e14 (2018).
10. Leclair, N. K. *et al.* Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Mol. Cell* **80**, 648–665.e9 (2020).