# Outline

1. Intro to splicing
   1. Disease relevance of splicing
   2. Detection of rare variants
2. Splicing datasets and models
   1. Genomics datasets
   2. MPSAs
   3. Benefits of being able to combine data across datasets
3. Aims
   1. Generate a model that can take MPSA and genomics data
      1. Apply continual learning/multitask techniques to adding MPSAs to annotation trained data
      2. Look for improvements in some aspects of performance
   2. Generate a simulator to predict experiment power
      1. Explore different strategies for active learning
      2. different measures of uncertainty
   3. Carry out the experiments in truth
4. Short description of the training plan
   1. Focus on training in DL techniques
      1. Computational experiments
      2. Interactions with the Koo lab
   2. Training in splicing
      1. Bench experiments
      2. Adrian
   3. Conferences
      1. ML-CB
      2. Probabilistic networks in genomics (CSHL)
      3. DL summer school
         1. Ask peter lab ppl
         2. Evan
      4. mRNA processing meeting (CHSL/GRC)

# Specific Aims:

# Alternative splicing is an important factor in gene regulation; however, its regulation is still incompletely understood. Mutations that change splicing patterns cause diverse genetic diseases.1 Current models of splicing regulation are informed by large scale splicing datasets.2–4 These datasets include transcript isoform reference annotations like those found in GENCODE,5 RNA-seq reads covering splicing junctions in different contexts like those found in GTEx and ENCODE,6,7 and massively parallel splicing assays (MPSAs) which collect information on splicing outcomes for particular variants.4,8–14 However, current practice does not efficiently utilize this data to train splicing models because models are not directly trained in a cross-dataset fashion and therefore information learned from one dataset cannot inform learning on another. Further, new data collection is not targeted to maximally improve model performance. These factors combine to limit our ability to predict splicing outcomes.

# I propose to develop methods for improving splicing models by training them across multiple types of datasets and to design experiments to maximally improve model performance. I will use these methods both in the context of improving current state-of-the-art splicing models and for training new model architectures. I will accomplish these goals in 3 complementary aims.

# Aim 1: Expand models of splicing to train across diverse genomics datasets. I will create a framework for applying multi-task learning and continual learning techniques to train splicing models across multiple datasets. Training the same model across multiple splicing datasets can allow information learned in one dataset to be transferred to the other datasets. I will apply this framework to extending state-of-the-art splicing models like SpliceAI and Pangolin as well as to trialing new model architectures. I will determine how training across diverse datasets including reference isoform annotations, RNA-seq splice junction annotations, and MPSAs across species and cell types affects model performance compared to training in a single task context.

# Aim 2: Develop methods for designing targeted MPSAs to improve splicing models. I will use active learning techniques to identify maximally informative datasets for fine tuning splicing model performance. I will identify sequences where model performance is poor or uncertainty is high and use model attribution techniques to map sequence features driving poor performance. I will design maximally informative MPSA libraries, then simulate results *in silico*. By fine tuning the models with these datasets, I will evaluate different model guided library design strategies for improving model performance.

# Aim 3: Generate new MPSA data and evaluate effects on model understanding. I will perform a full model guided MPSA and collect data for model fine tuning to evaluate the effectiveness of these techniques outside of the simulated context. I will evaluate performance differences between models fine-tuned with unguided MPSA data and model guided MPSA data. I will focus on the ability of the model to transfer learning from the MPSA dataset to the previously learned datasets and to specifically improve performance on previously poor contexts and to resolve uncertainty driven by the targeted sequence elements.

# Specific Aims:

Alternative splicing is an important factor in gene regulation; however, its regulation is still incompletely understood. While splicing decisions are driven by information in the local sequence context, outcomes can be correlated over large distances. A particularly common example of this is mutually exclusive exons. These occur when exons in a cluster are anticorrelated so only one is included in the mature transcript. Several mechanisms for producing this behavior have been identified, including spliceosome incompatibility, 5’-splice site/branch point steric occlusion, and RNA secondary structures but most mutually exclusive exon clusters in humans still have no known mechanism. While these long-distance anti-correlations present challenges as a study system because they require measuring multiple exons simultaneously in a single transcript. Work on the function of mutually exclusive exons has borne fruit. A common theme is switch-like changes in protein function that occur in cell differentiation or development. Pyruvate Kinase M (PKM), ketohexokinase, and CaV1.2 are some of the better-known examples of important regulatory events that occur through mutually exclusive exon switches. PKM is converted to a constitutively active form during differentiation, reversion of this change in cancer cells is a major driver of the Warburg effect. Ketohexokinase can be switched between high and low affinity forms to control fructose metabolism, aberrant expression of the high affinity form causes pathological cardiac hypertrophy while the low affinity form is a major driver of hepatocellular carcinoma. Finally, mutations that shift the isoform distribution of CaV1.2 cause timothy syndrome characterized by catastrophic developmental, neurological, and cardiac symptoms with a life expectancy of ~2.5 years. Understanding mutually exclusive splicing will not only deepen our knowledge of development and differentiation but bring us closer to treatments for a wide variety of human diseases.

Massively parallel splicing assays (MPSAs) provide a method for interrogating the mechanisms of alternative splicing but have technical limitations. There are several different types of MPSA, but fundamentally they all rely on introducing a library of variant mini-genes and using short read sequencing to measure how each variant is spliced. These techniques have great promise for dissecting splicing regulation by enabling quantitative phenotype measurement of large numbers of mutants in parallel. However, because these methods use short-read sequencing they cannot confidently quantify all isoforms of minigenes which are longer than a short read. This is especially problematic for measuring exons in the context of their native introns, for measuring clusters of mutually exclusive exons, and for distinguishing isoforms where cryptic splice sites or intron retention events play a role. These issues can also lead to noisy measurements and limit the dynamic range of the assay by collapsing multiple isoforms which cannot be distinguished using only short-read based methods.

I propose to develop tools for studying complicated splicing regulation. These tools are a long read based MPSA (LR-MPSA) and analysis software for nucleotide resolution MPSA data. While these tools will be broadly useful for understanding a diverse set of splicing decisions, here I will apply them to dissecting the mechanisms of mutually exclusive exon splicing in the specific context of pyruvate kinase M (PKM). I will accomplish these goals in three complementary aims.

**Aim 1: Develop a nucleotide-resolution MPSA using long read sequencing***.* I will develop an LR-MPSA optimized for characterizing the mechanisms of complex splicing decisions in the context of full native introns and all isoform outputs. I will adapt MPSA techniques honed by the Kinney lab by creating new sample and library preparation pipelines optimized for unbiased isoform enrichment and nanopore sequencing.

**Aim 2: Develop software for analyzing data from nucleotide-resolution MPSAs***.* Calling low abundance RNA isoforms from nanopore reads faces challenges from the low depth and high noise. I will develop analysis software for identifying and quantifying isoforms that optimizes for power to detect low abundance isoforms and minimizes incorrect isoform assignment. I will simulate nanopore reads given a known ground truth isoform distribution for each variant to test and optimize the software at various depths and error profiles. I will make this software publicly available and open-source.

# Aim 3: Define the mechanism of mutually exclusive exon inclusion in PKM. PKM has a pair of mutually exclusive exons and switching between them drives the Warburg effect, this splicing event is a target for cancer therapies. PKM mutually exclusive exon splicing causes a shift from respiration to fermentation that is a driver of cancer. I will use LR-MPSAs to identify motifs that are important for maintaining mutual exclusivity in PKM mutually exclusive exons by identifying mutants that deviate from mutually exclusive exon behavior. I will verify identified motifs with anti-sense oligo based blocking and low throughput experiments, then model their behavior.

# My background positions me well to execute this proposal, yet this proposal offers a training opportunity that will help me found an independent lab. In my doctorate I used massively parallel assays and computational modeling in microbiology, protein engineering, and CRISPR tool development providing me with the foundational skills I will need for this proposal. This proposal will allow me to explore new applications of these skills in a field with opportunities to do fundamental science with applications in human health. While I pursue these scientific goals, I will also be focused on utilizing my position at Cold Spring Harbor lab to gain wider training and mentorship that will help launch my career as an independent researcher. I will attend CSHL meetings and Gordon conferences on RNA processing and quantitative methods, attend lab meetings and journal clubs in the Kinney, Krainer, Koo and McCandlish labs, attend CSHL grant writing and professional development courses, and hone my skill at lecturing through teaching opportunities.

Together, these research and training opportunities will position me to launch an independent research career focusing on applying massively parallel assays and modeling to understanding deep biological questions in RNA processing.

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

2. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

3. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).

4. Liao, S. E., Sudarshan, M. & Regev, O. Machine learning for discovery: deciphering RNA splicing logic. *bioRxiv* 2022.10.01.510472 (2022) doi:10.1101/2022.10.01.510472.

5. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).

6. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

7. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

8. 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).

9. 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).

10. 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).

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

12. 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).

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

14. 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).