MAYO CLINIC

Identification of aberrant splicing events in rare genetic disease patients

A. Collin Osborne
Senior Bioinformatician
Mayo Clinic Quantitative Health Sciences
Mayo Clinic Center for Individualized Medicine
Authored by: **Gavin Oliver *and Garrett Jenkinson PhD***

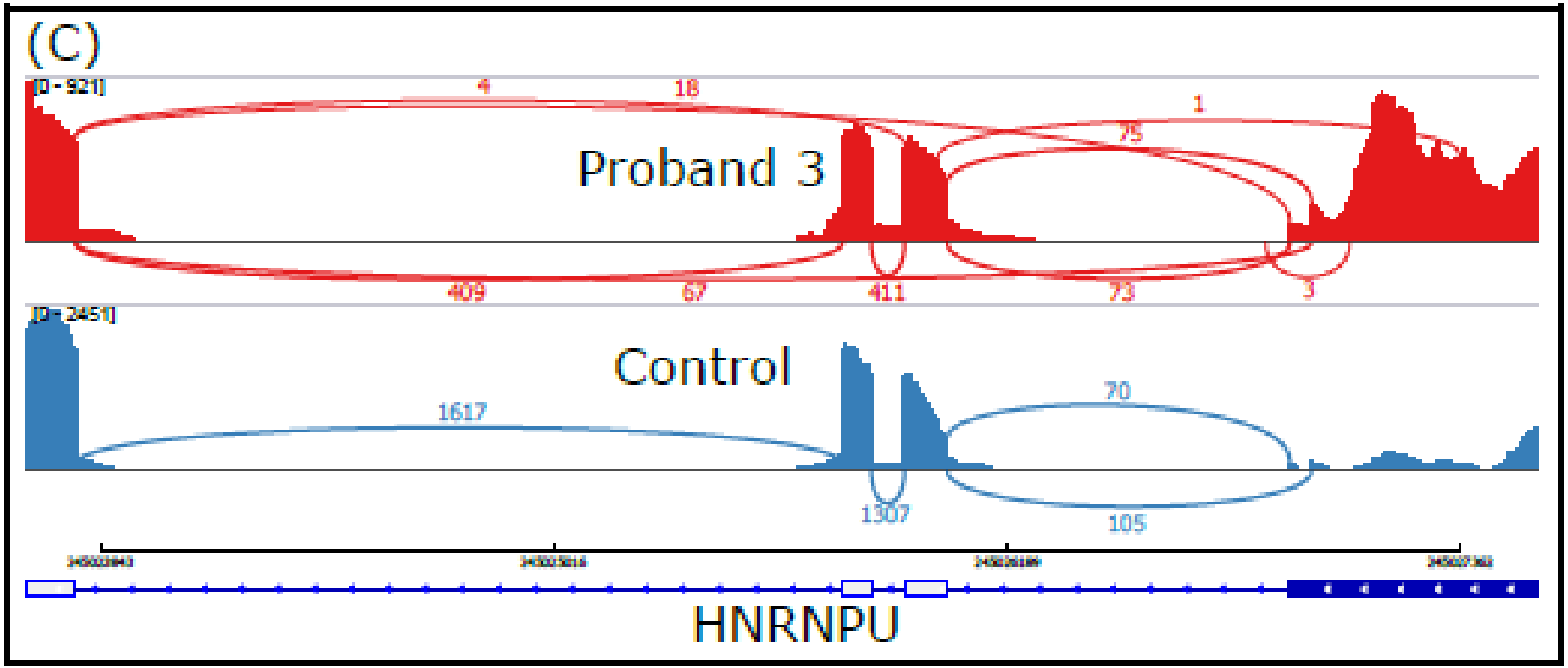
Splicing in rare disease

- 15-60% of rare-disease causing variants are estimated to affect normal splicing
- Splice effects hard to predict from genome since can be affected by distal/intronic functional units
  - SpliceAI is state-of-the-art in predicting DNA variant impact on splicing, but still misses a lot
- RNA-seq provides a direct window into the variable use of isoforms in mature mRNA

What data do we have?

- Split reads
  - aligning partially to one exon and then partially to another
  - evidence of an intron having been excised
- "Junctions" define start/end position of introns
- Number of split reads with identical junctions give junction counts
- Junction counts form starting point for outlier analysis

Splicing Case Study 3

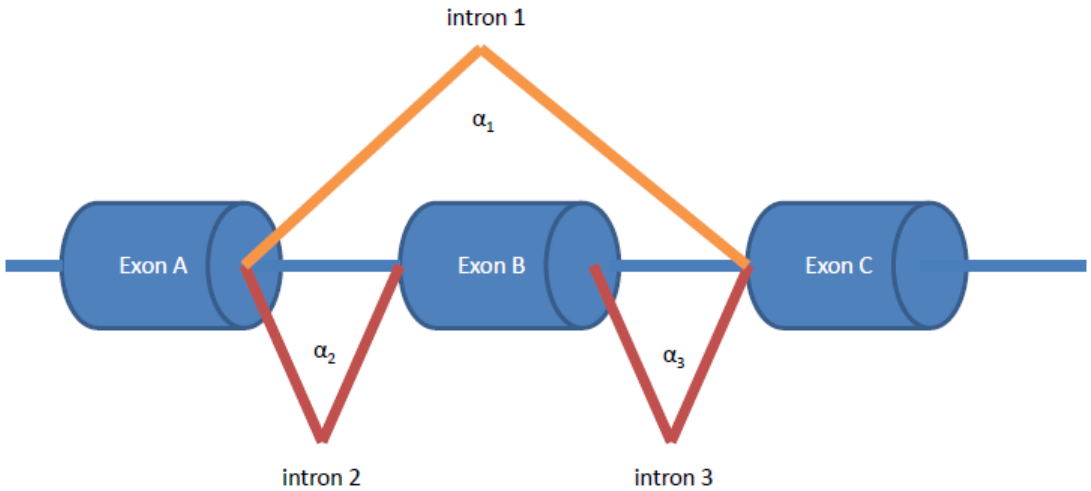


How to normalize?

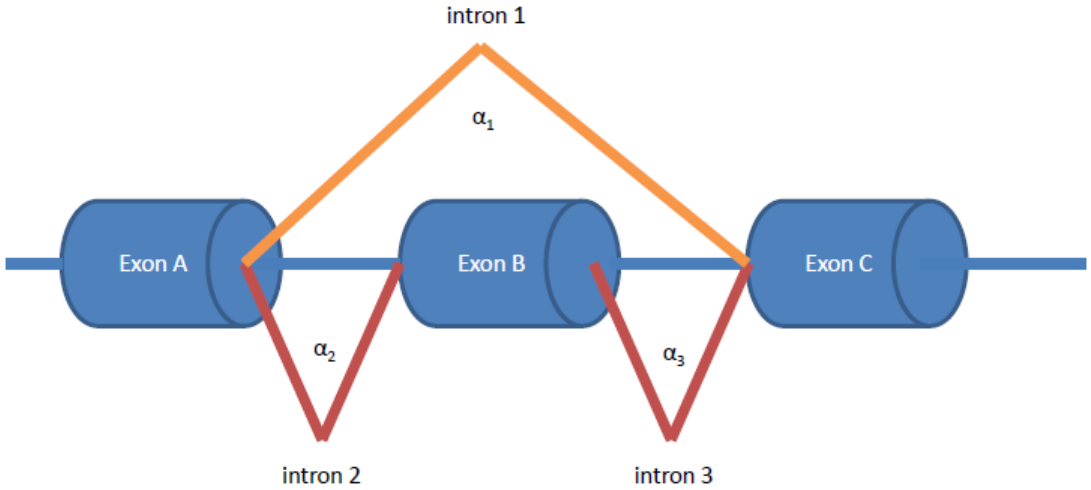
- Open question in field
  - We published LeafcutterMD based on the "LeafCutter" algorithm (Li *et al.*, 2018) that uses graph-theoretic methods to build clusters of introns
  - The OUTRIDER authors recently published "FRASER" using autoencoders to automatically normalize

- While more sophisticated, the underlying principles are directly analogous to outlier expression, with a shift to focus on introns
  - Counts that need normalizing
  - Cohort model determines what is an outlier

Splicing As Intron Use (Leafcutter)

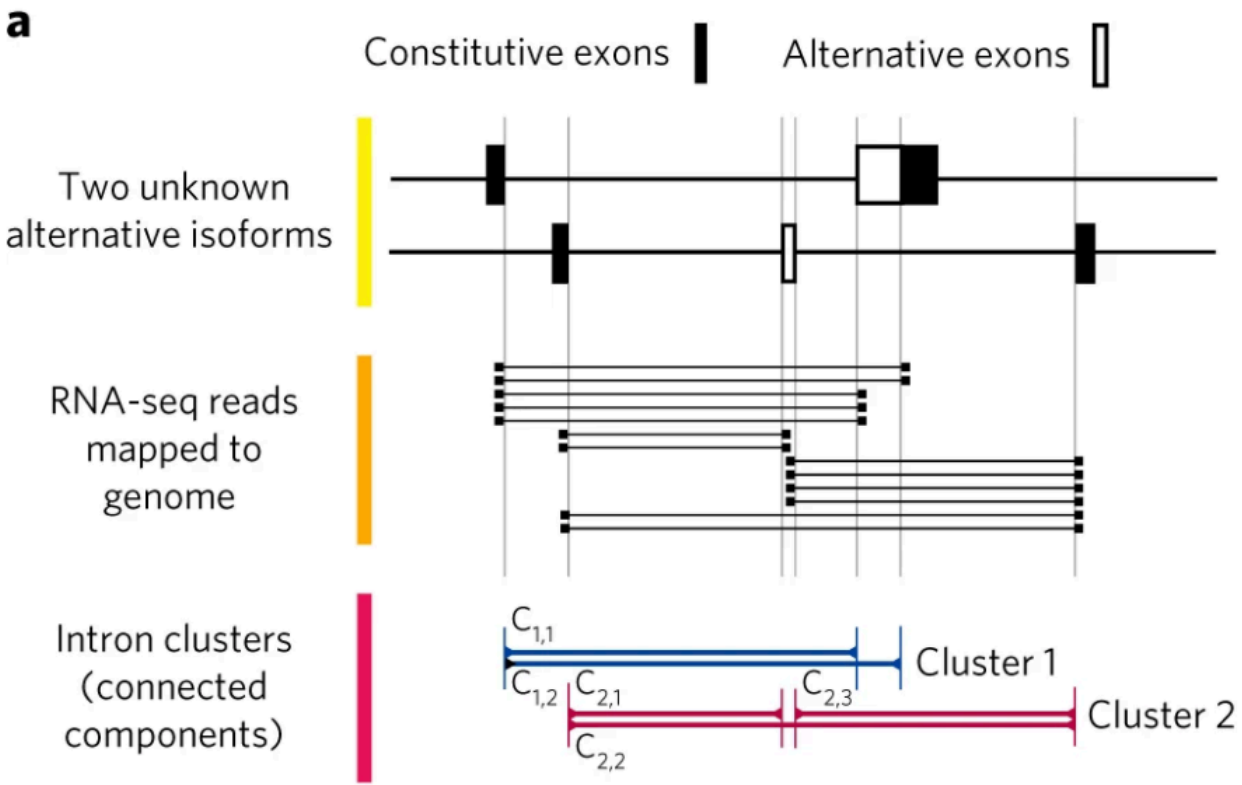


Splicing As Intron Use (Leafcutter)



- If red introns commonly used and orange uncommonly, increased orange = skipping
- Conversely when orange is common, increased red intron use = cryptic exon

Intron Clustering (Leafcutter)



LeafcutterMD

- Uses intron clustering approach from LeafCutter
- Biological differences in splicing are thus captured by differing measurements of intron usage/excision
  - Split reads used to specify intron boundaries and and quantify usage/excision counts
  - Construct a graph whose nodes are introns connected by edges representing a shared splice junction between two introns
  - An iterative filtering and graph building approach is followed until convergence, forming ‘clusters’ of introns

LeafcutterMD

- This procedure results in intron cluster *c* having *I* possible introns in a total of *S* proband samples
  - Each cluster within our total clusters is then considered independently
  - LeafCutter outputs the count n for intron *i* and sample *s*, which can be viewed as an $I \times S$ matrix.

- We formulate the outlier splicing detection problem as identifying n_{is} that indicate abnormally high or low usage of intron *i* in a sample *s* versus all other samples

LeafcutterMD

- If all samples were expected to use an intron with same probability, then counts would be drawn from a Multinomial distribution
- In general, however, we do not expect all samples to have identical usage probabilities
- Biological variability across individuals will result in variable intron excision rates, and modeling the probability distribution for each individual has been shown to work well
- A flexible and computationally convenient choice for modeling this distribution over Dirichlet distribution

LeafcutterMD

- Statistical test to determine how unlikely the count of an intron in a sample comes from the distribution must estimate the parameters that capture the variability in the proportional intron usage within the population that does not include our sample of interest

- Null hypothesis underlying the statistical test in LeafCutterMD is that it is drawn from the same distribution as the rest of the samples, whereas the rejection of the null hypothesis indicates that is an outlier from this population

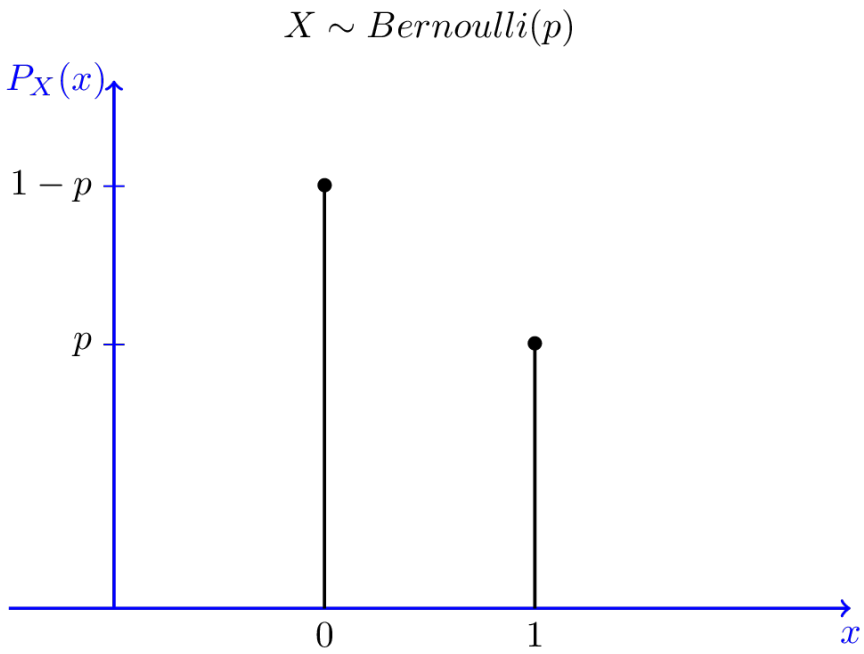
- To compute the *P*-value, we first marginalize the Dirichlet-Multinomial to find the distribution of this particular intron count under the null hypothesis

LeafCutterMD

- Clustering builds a graph of introns that share a common start or end
- Uses clustering of introns and a Dirichlet-Multinomial/Beta-binomial model for intron counts
- We explore next the Dirichlet-Multinomial because it is illustrative of important modeling characteristics
  - Biological variability (Dirichlet)
  - Statistical sampling variability (Multinomial)

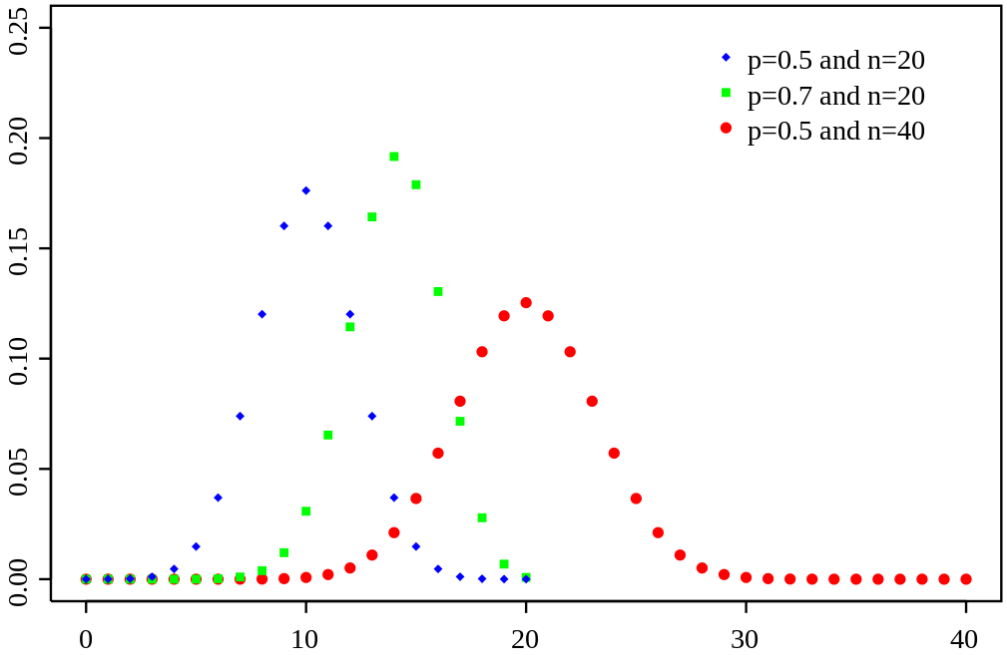
Important distributions

- Bernoulli: Single weighted coin flip
  - Single observation where the option is splice from position A to position B, or to some other position C



Binomial Distribution

- Counts after repeated Bernoulli trials/coin flips that have the same probability of heads



"Over dispersion"

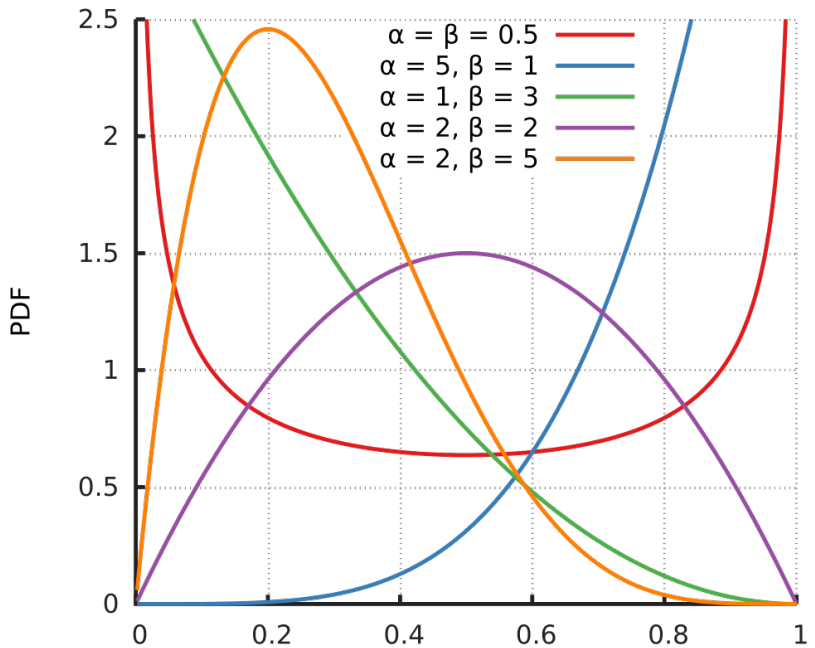
- Biology is messy
  - Cells are not identical replicates
  - Data has technical artifacts/noise other than statistical sampling errors
- Binomial "identical coin flip" model is over-optimistic about noise/variation
  - Need a model that can have variation larger than expected under perfectly identical trials

Compound distribution solution

- What if each trial didn't have same p (probability of heads)? What if each cell sampled had its own probability of splicing to position A versus other sites?
- Need to introduce two step (compound) model
  - For each cluster
    - First cell is selected, which determines p (which we don't know i.e. latent variable)
    - Next Bernoulli trial is held with probability of heads given by p for this cell
    - Repeated for all reads/cells within cluster with each cell having its own p

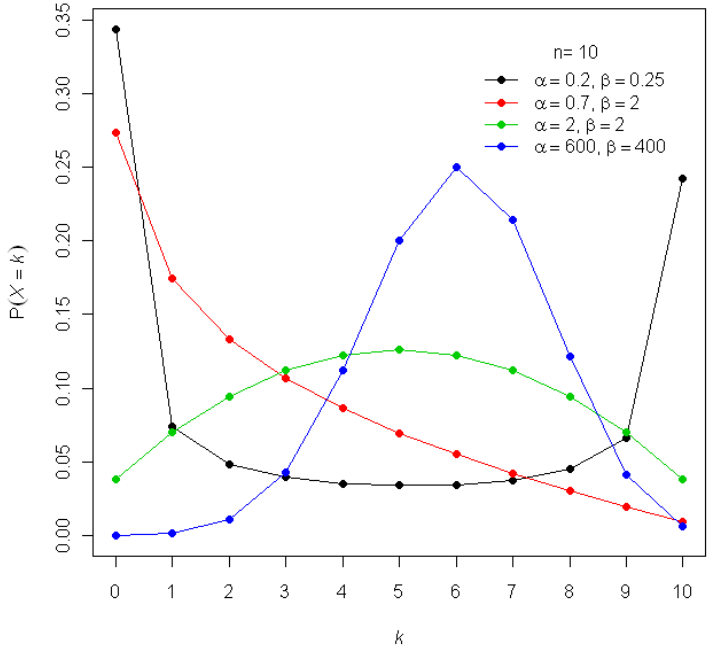
Need distribution for p in population

- Beta distribution is mathematically/computationally convenient and flexible choice



Beta-Binomial

- Distribution of counts you get when p is selected from a beta prior to each bernoulli trial



Dirichlet-Multinomial

- Concepts the same, math slightly more technical
- Multinomial is generalization of binomial to more than two outcomes
  - Count occurrences on weighted dice (not necessarily 6 sides)
- Dirichlet distribution is generalization of beta to draw probability vectors (rather than single probability) that sum to one
  - Assigns continuous distribution to the multidimensional standard simplex
- Each intron in cluster has a probability (drawn from Dirichlet) and count (drawn from multinomial conditioned on probability)
  - Once model fit, do outlier analysis as usual

Summary

- Outlier expression has been described where we model normal variability in gene expression within a cohort
- Splicing is a similar problem but utilizes distinct methods
  - LeafcutterMD uses Dirichlet Multinomial to account for biological and statistical sampling uncertainty
- You will apply these methods and gain greater intuition in the laboratory session

MAYO CLINIC

Questions?

CENTER FOR INDIVIDUALIZED MEDICINE