# Yanagi: Transcript Segment Library Construction for RNA-Seq Quantification

**Mohamed Gunady**, Steffen Cornwell, Stephen Mount, Hector Bravo

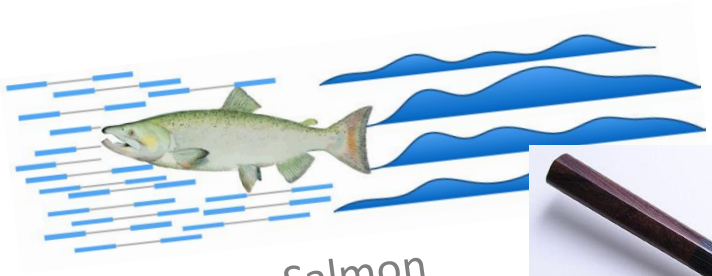
Department of Computer Science

University of Maryland – College Park



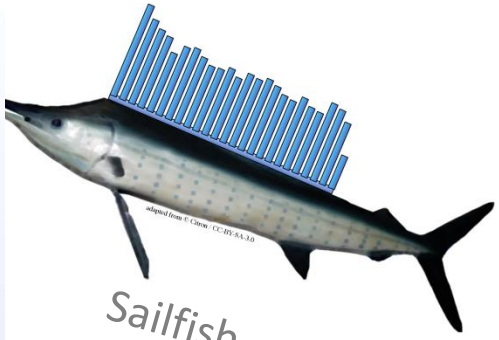
CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

# Yanagi?



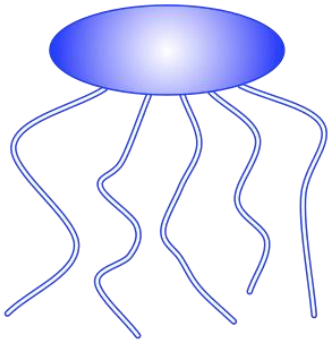
Salmon

Yanagi



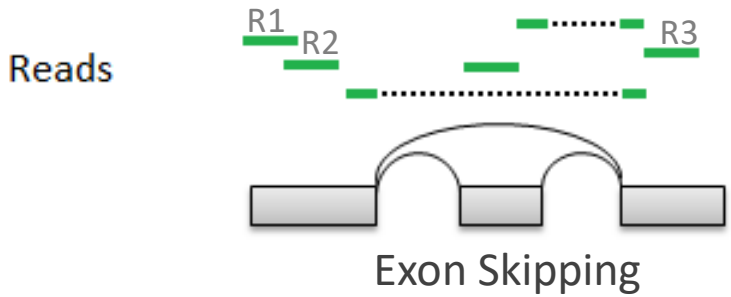
Sailfish

Jellyfish/RapMap



# Motivation

- Challenges with Transcript-level Quantification, e.g.
  - 95% of human genes with multiple exons undergo Alternative Splicing (AS)[1]




  - Ambiguity due to multi-mapped reads
    - Usually resolved using probabilistic models like EM

# Motivation

- Challenges with Transcript-level Quantification, e.g.
  - Considering local splicing variations leads to a combinatorial number of transcripts.[2,3]
    - Standard Annotations list only a minimal subset
  - Short-read sequencing does not provide information for correlation between distant splicing events.[4]

# Motivation

- Our Vision:
  - Eliminate multi-mapping trivially caused by the significant share of genomic regions.
  - Building sufficient statistics describing individual events.
    - Independently from the estimation of transcript abundances.
  - Utilize the graph representation of the transcriptome.
    - Without building a special graph-based aligners.

# Our Approach

## Transcriptome Segmentation

# Yanagi's Approach

- Idea Overview:
  - Segment the transcriptome into a set of disjoint regions.
  - Without losing any possible transcriptome sub-sequences.
    - I.e. Linearizing the splice graph
  - Then use the generated segments as a reference instead of the transcriptome.

# Yanagi's Approach

- A Segment:

$$seg(Exs, loc, w)$$

- Segments are *L-Disjoint*

$$width\left[overlap\left(seg_i, seg_j\right)\right] < L; i \neq j$$

- *L* corresponds to the read length

- No read of length at least *L* can map to both segments
  - Ignoring sequencing errors and paralogs for now!

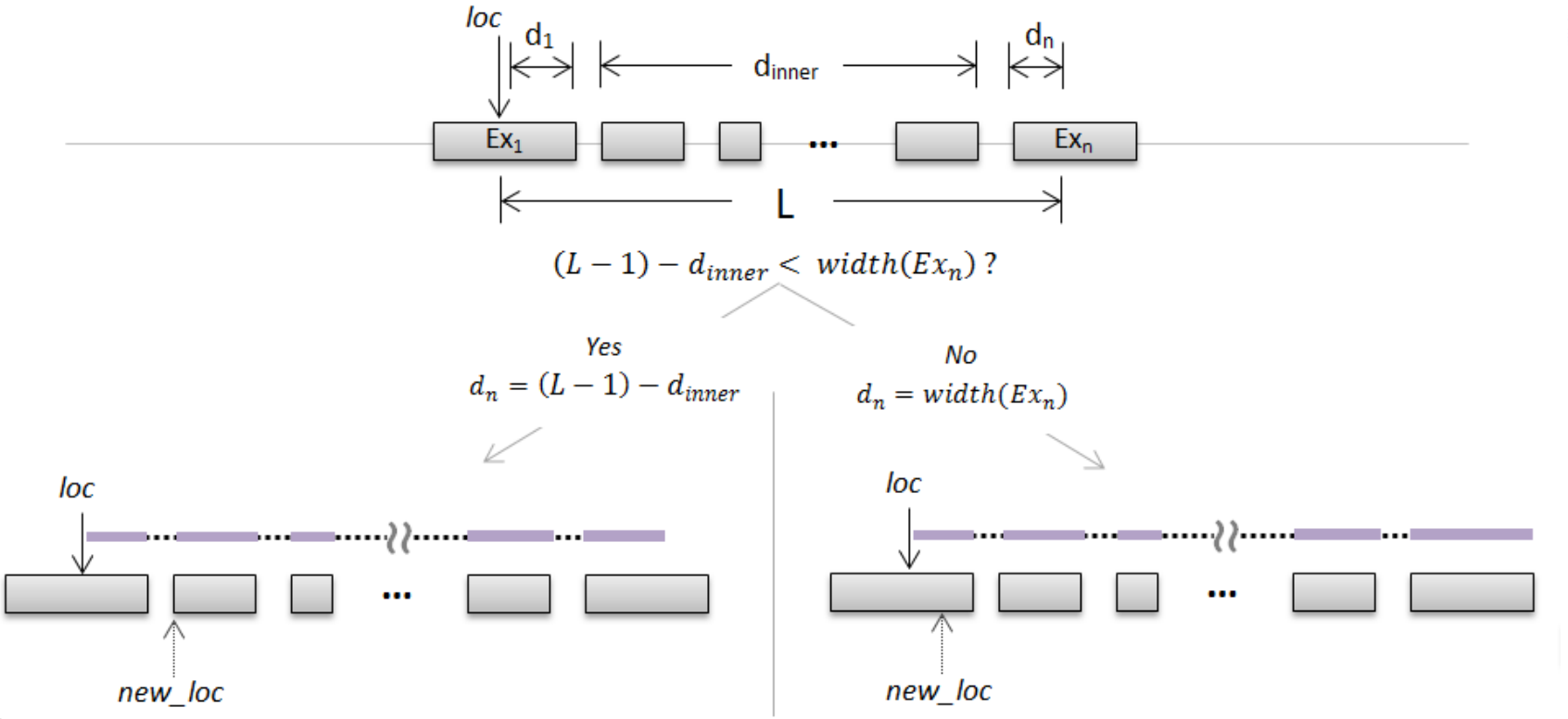
# Yanagi's Approach

- Generating Segments:
  - Naïve Approach


  - ~30% of exons in UCSC hg38 are shorter than 100bp

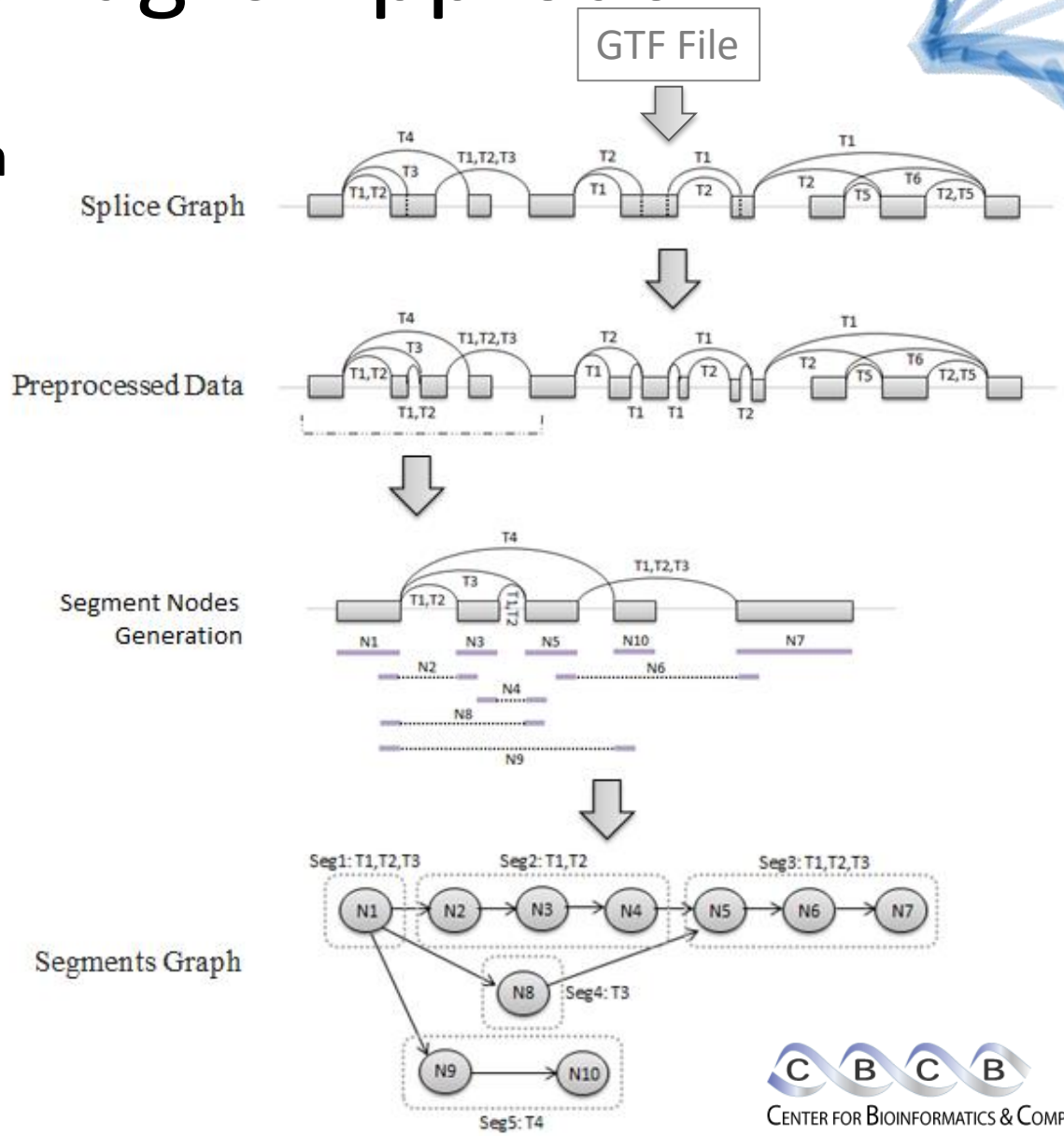
# Yanagi's Approach

- Generating Segments:
  - Yanagi's Approach



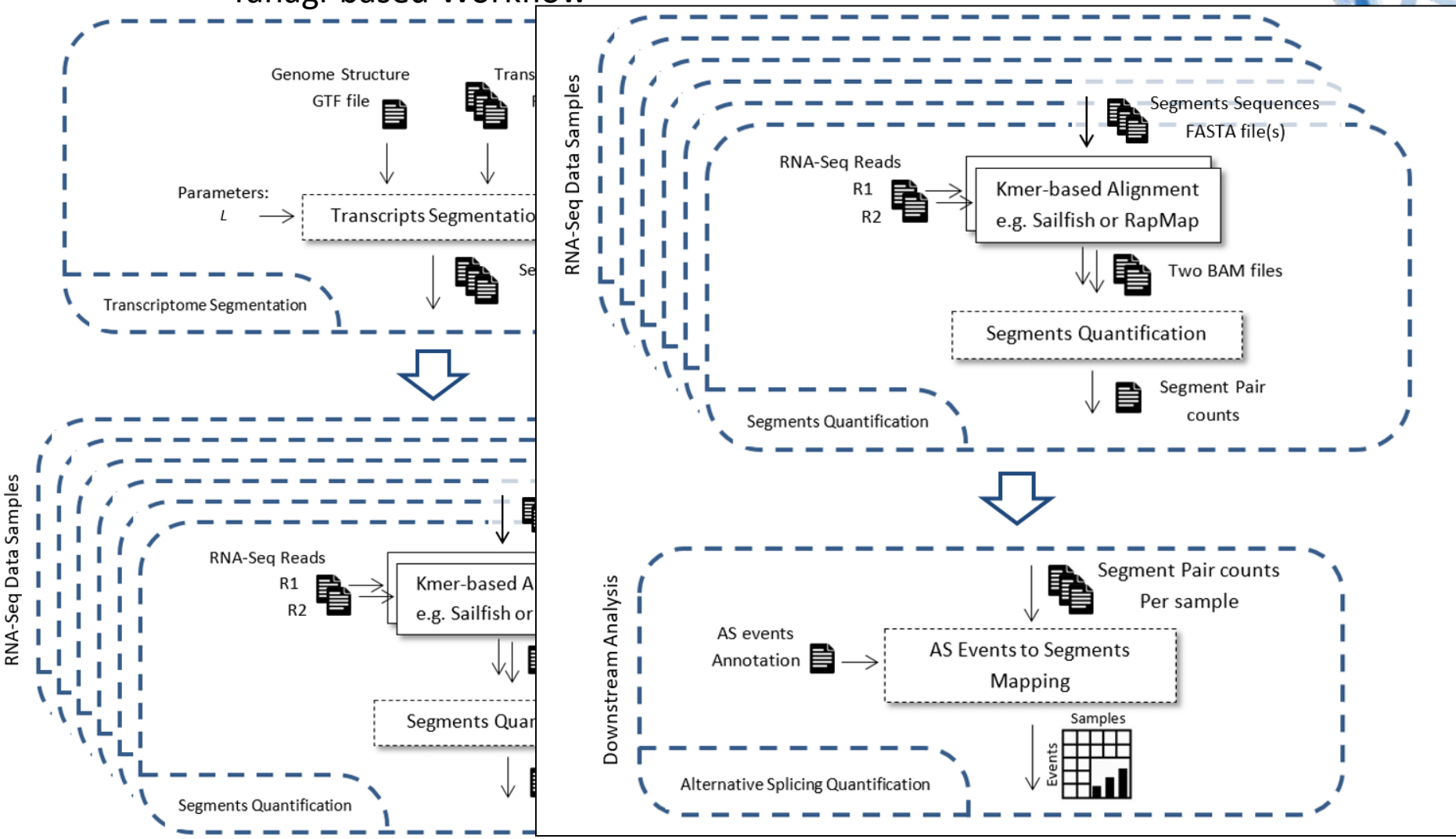
# Yanagi's Approach

- Segments Graph



# Yanagi-based Quantification Workflow

Yanagi-based Workflow

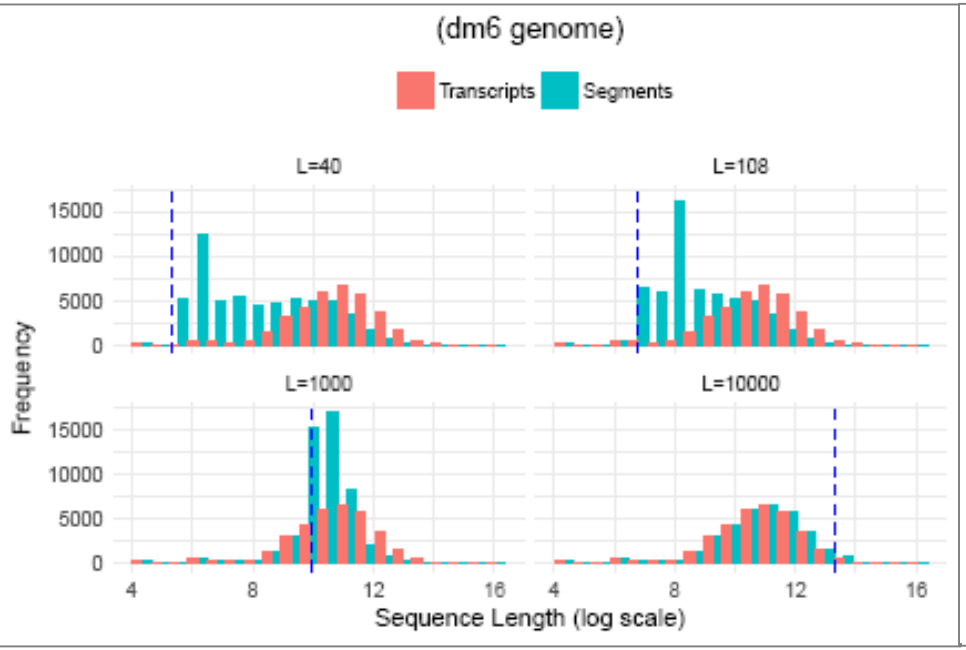


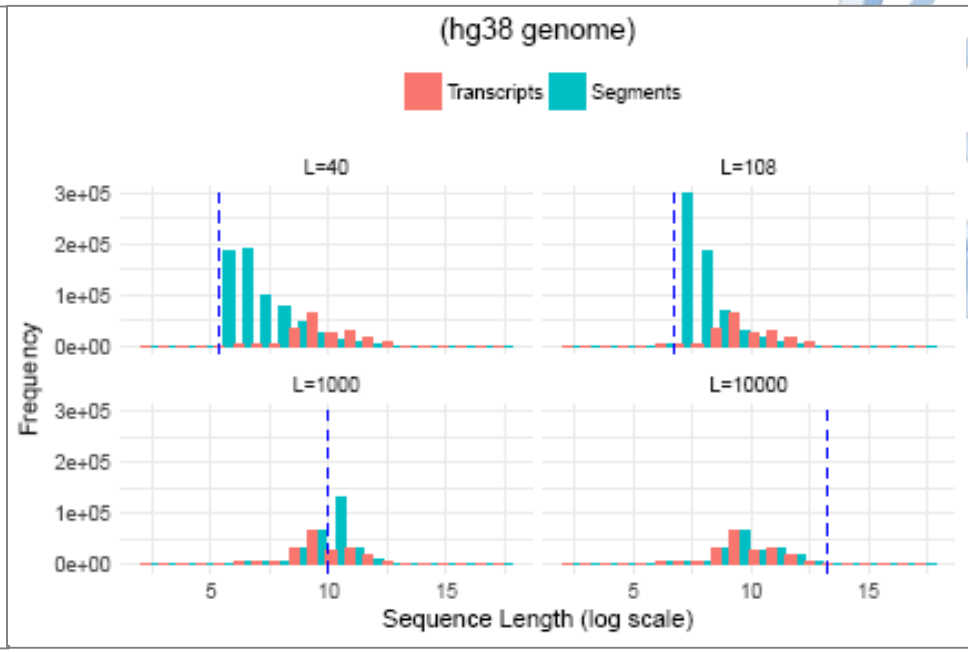
# Experiments

## Segments Analysis

# Segments Analysis

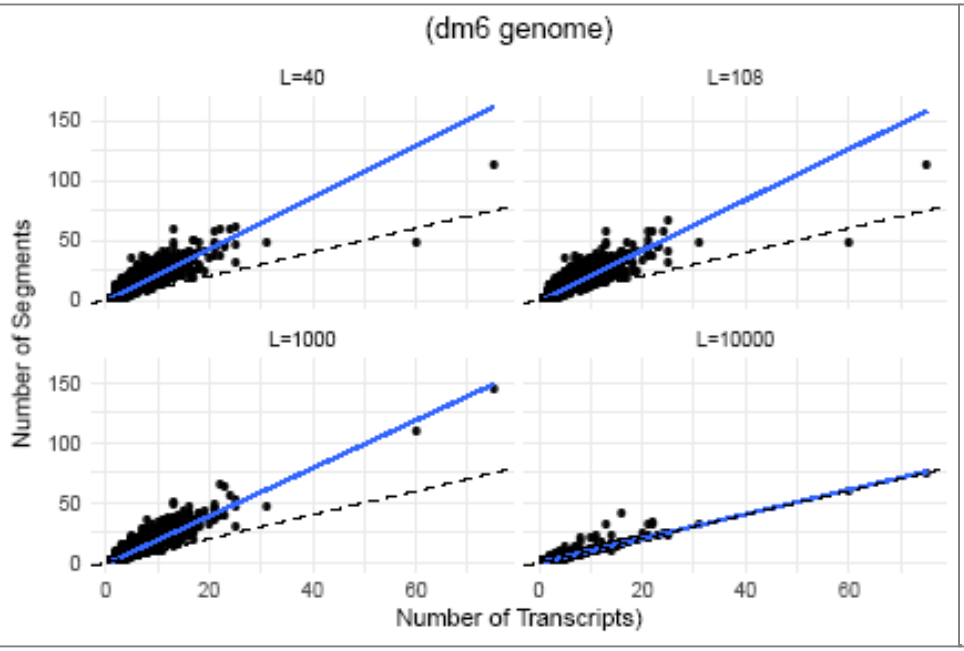
- Segments Length

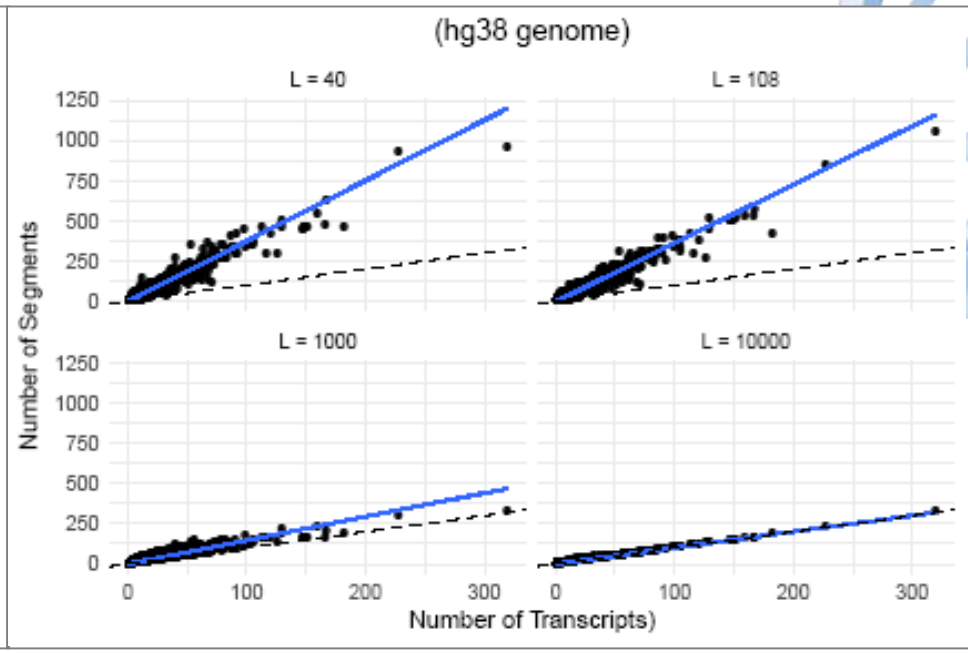


# Segments Analysis

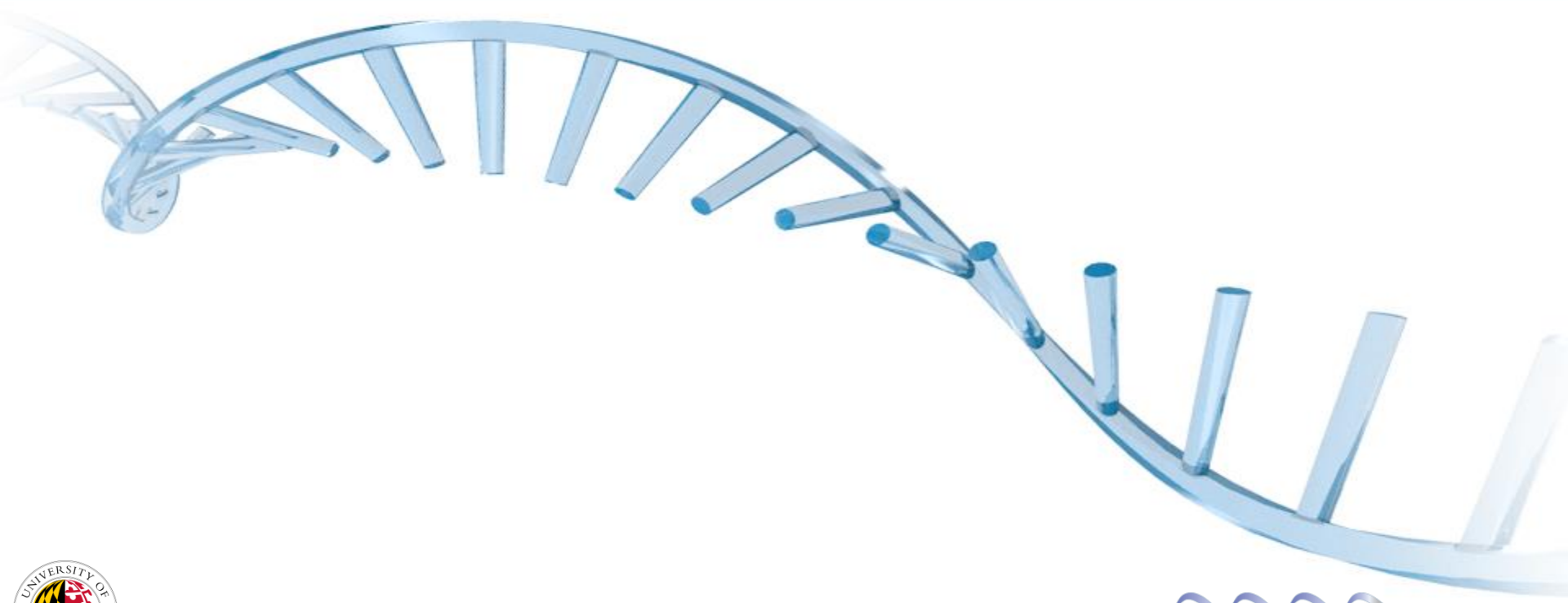
- Number of Segments



# Use Case: Alt. Splicing Quantification

## Differential Exon Skipping



# Differential Exon Skipping

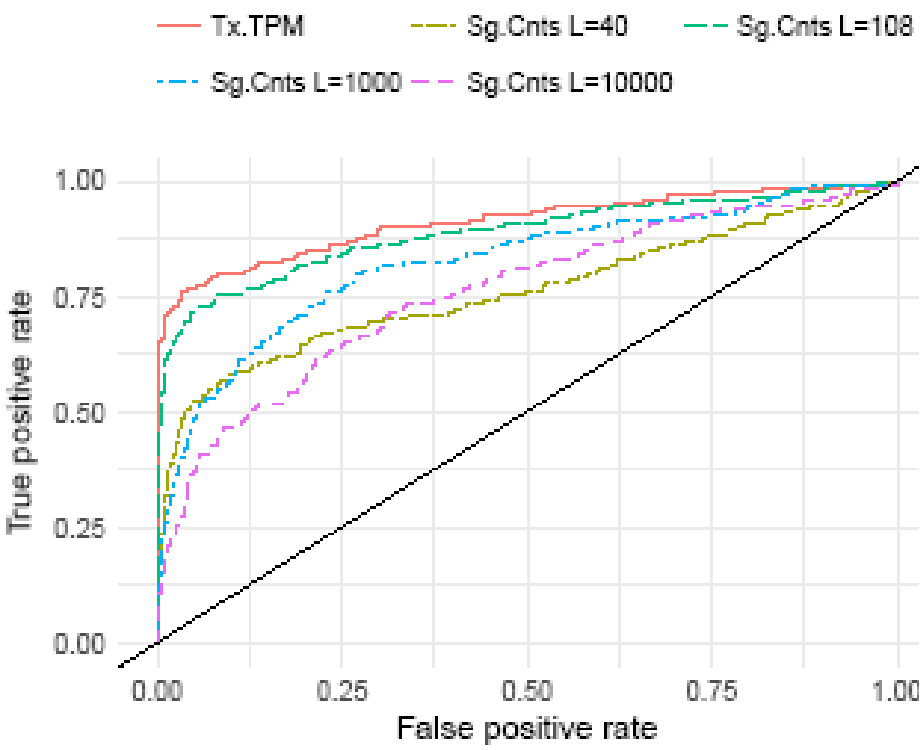
- Synthetic Data:[5]
  - 2 conditions, 3 replicas each.
  - Simulated reads are based on real RNA-Seq data.
  - For 1000 genes with at least two transcripts.
  - Transcription levels of the most abundant two transcripts are switched across conditions.

- Differential Analysis:
  - Exon Skipping events.
  - Linear Model based on the segment counts.
  - Using Limma-Voom.

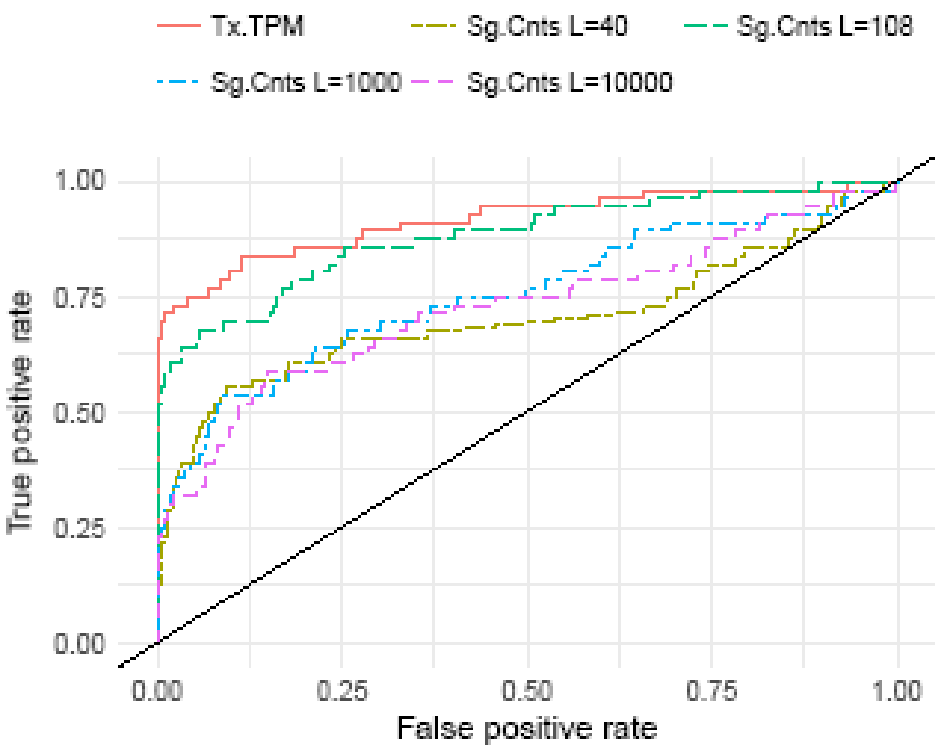
# Differential Exon Skipping

- ROC plots:
  - using RapMap for alignment

(dm3 genome)




(hg37 genome)

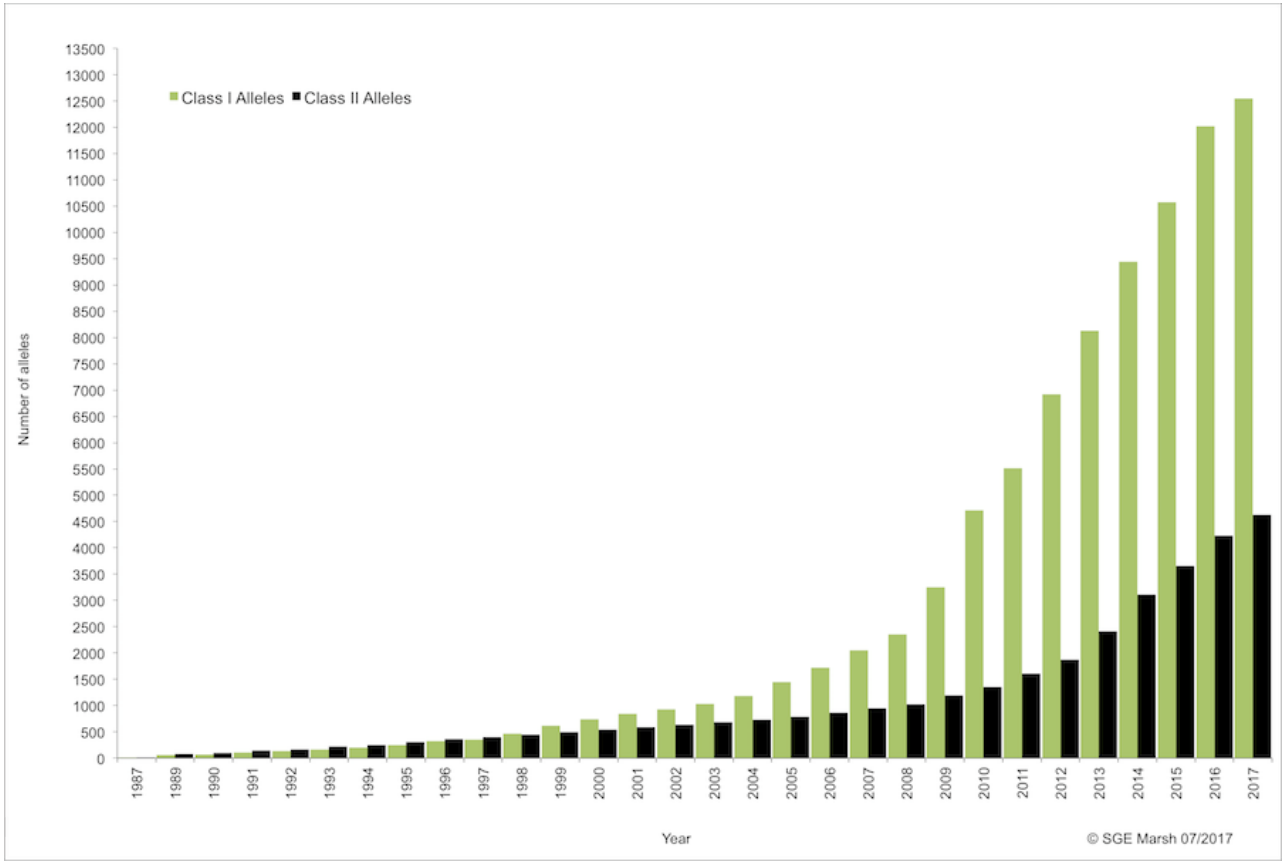


# Use Case: Linearized Population Reference Graph

## Aligning Over Genomic Variants for WGS

# Background

- Several databases of Genomic Variants
  - Rapidly-growing, public archives.
    - E.g. IPD-IMGT/HLA Database currently has 17,344 allele sequences.






illumina

# Background

- In principle, Graphs are reasonable representation.

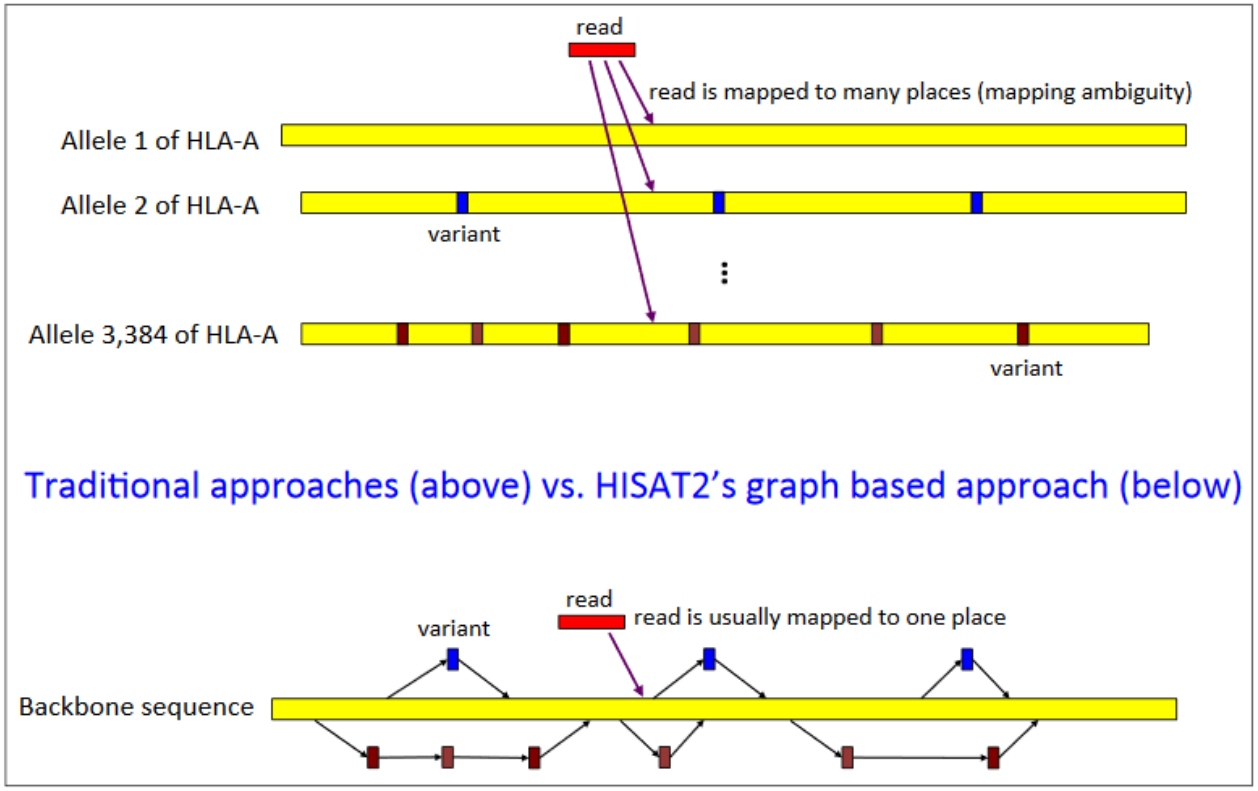


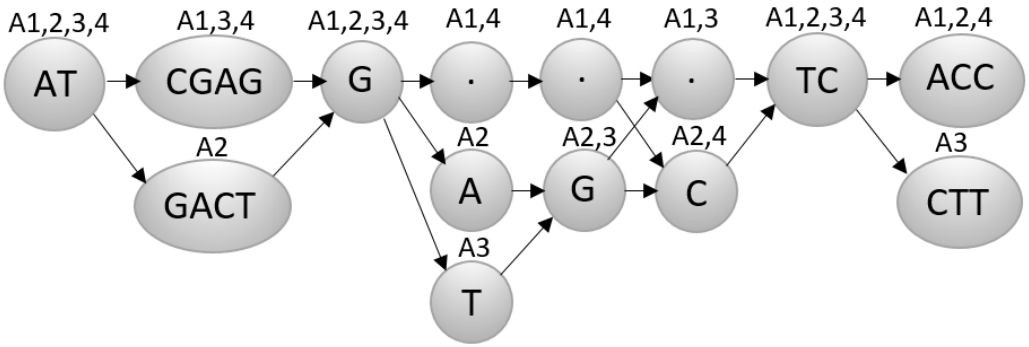

Figure from HISAT-genotype's poster

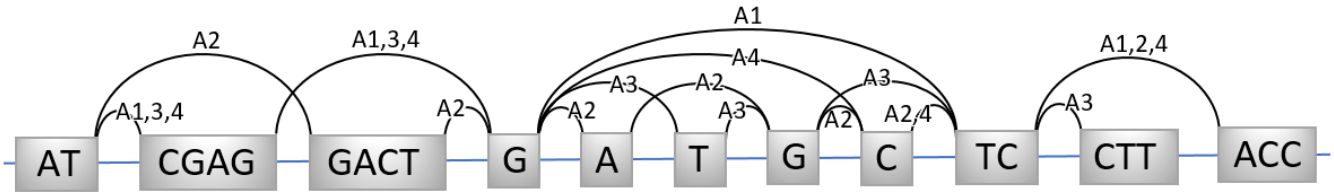
# Yanagi + Genomic Variants

- Project Variants Graphs into Splice Graphs.
  - Start from Multi-Sequence Alignment (MSA) of gene alleles

```
A1: ATC GAG G.. .TC ACC
A2: --G ACT -AG C-- ---
A3: --- --- -TG .-- CTT
A4: --- --- --- C-- ---
```

4 alleles (A1, A2, A3, A4)
( **-** ) same as the first allele
( **.** ) indicate indels



# Yanagi + Genomic Variants

- Preliminary Results: Number of Aligned Reads
  - Simulated reads from 6 HLA genes

| Dataset | Total Reads | HISAT-genotype | bwa-mem | | RapMap | |
|---|---|---|---|---|---|---|
| | | (WG Graph) | Ref | Ref+Segs | Ref | Ref+Segs |
| ClassIEasy | 6,000 | 5,900 | 6,000 | 6,000 | 4,163 | 5,990 |
| ClassIHard | 6,000 | 5,966 | 5,797 | 6,000 | 3,553 | 5,990 |
| ClassIIHard | 14,000 | 13,844 | 12,232 | 13,997 | 7,628 | 13,975 |

# Summary

- Yanagi perform a transcriptome segmentation into *L*-disjoint segments.
- Introduces fine-grained counts as statistics for DE analysis.
- Flexible approach that can be used in different use cases:
  - Alternative Splicing Quantification
  - Variant Calling

# Future Extensions

- Further Analysis, e.g.
  - Experiments on Real Data
  - Comparison with other tools, e.g. SUPPA2 for AS
  - Detailed analysis on multi-mapped reads
- Discovering unannotated transcripts
- Handling paralogs and intersecting genes
- Handling complex repeats and structural variants

# Acknowledgments


- UMD Supervisors
  - Hector Bravo
  - Stephen Mount

- Illumina
  - Sangtae Kim
  - Chris Sanders

- This work was partially supported by:
  - National Science Foundation (NSF)
  - National Institutes of Health (NIH)

# References


1. Jorge Vaquero-Garcia, Alejandro Barrera, Matthew R. Gazzara, Juan Gonzalez-Vallinas, Nicholas F. Lahens, John B. Hogenesch, Kristen W. Lynch, Yoseph Barash, and Juan Valcárcel. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5:e11752+, February 2016.
2. S Lawrence Zipursky, Woj M Wojtowicz, and Daisuke Hattori. Got diversity? wiring the fly brain with dscam. *Trends in biochemical sciences*, 31(10):581–588, 2006.
3. Brian J Haas, Arthur L Delcher, Stephen M Mount, Jennifer RWortman, Roger K Smith Jr, Linda I Hannick, Rama Maiti, Catherine M Ronning, Douglas B Rusch, Christopher D Town, et al. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research*, 31(19):5654–5666, 2003.
4. Hagen Tilgner, Fereshteh Jahanbani, Tim Blauwkamp, Ali Moshrefi, Erich Jaeger, Feng Chen, Itamar Harel, Carlos D Bustamante, Morten Rasmussen, and Michael P Snyder. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature biotechnology*, 33(7):736–742, 2015.
5. Charlotte Soneson, Katarina L Matthes, Malgorzata Nowicka, CharityWLaw, and Mark D Robinson. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome biology*, 2016.

# Thank you!

Questions?