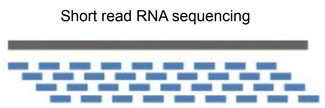
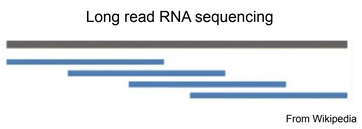
A Unified View of Transcriptome Complexity by Combining Transcriptome Annotation with Long and Short Reads RNA Sequencing

Seong Woo Han [1], Paul Jewell [2], Yoseph Barash [1,2]
[1]Department of Computer and Information Science, University of Pennsylvania
[2] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania

Motivation

- Better understand the underlying isoforms and splicing changes in a given biological condition
- Compare and contrast short-read and long-read RNA sequencing with the annotation of known isoforms





+ Detect and quantify splicing variations

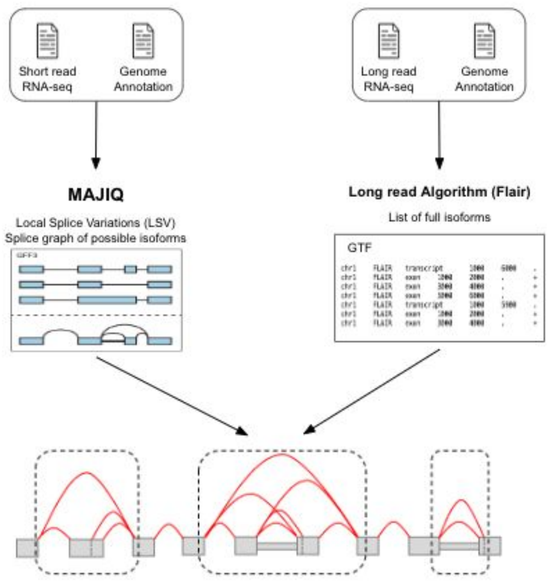
− Difficult to identify full isoforms

+ Offers improved detect of partial or full isoforms

− High error rates, difficult to fully detect and quantify splicing variations

Approach

- Dataset: Human cell line sequenced by Illumina (short read) and Pacbio (long read) in Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Consortium [1]
- Break down gene into module and identify which source contributes to which components [2], [3]

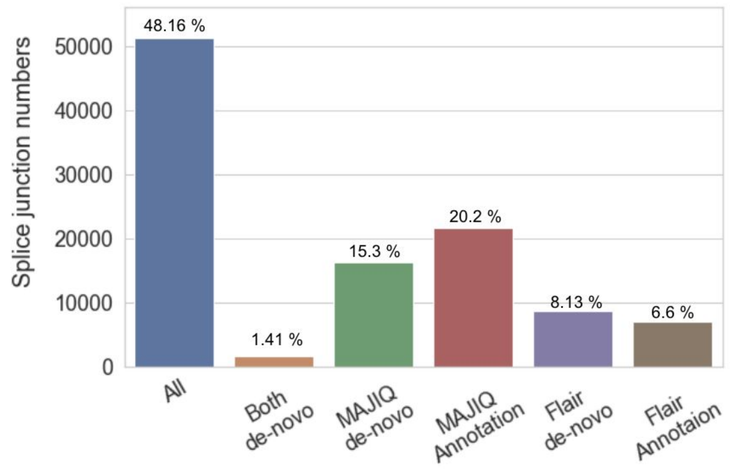


- Modules: distinct segments of a gene splice graph involving overlapping splice junctions which are contained between a single source and single target exon
- Each splice junction in each module can fall into one of six categories

	MAJIQ (short read)	Flair (long read)	Annotation
All	✓	✓	✓
Both de-novo	✓	✓	✗
MAJIQ de-novo	✓	✗	✗
MAJIQ / Annotation	✓	✗	✓
Flair de-novo	✗	✓	✗
Flair / Annotation	✗	✓	✓

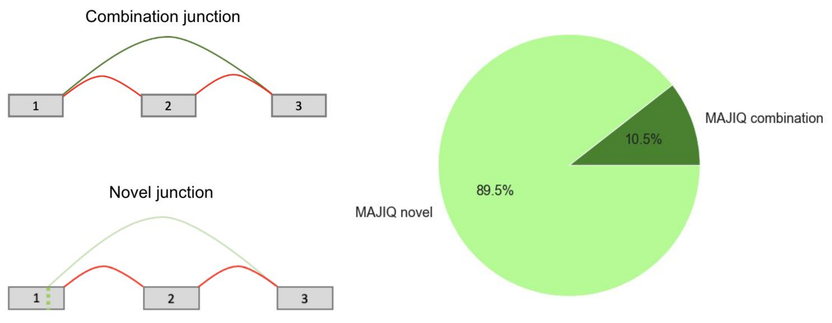
Splice sites identification

- 106,270 splice junctions in 56,127 modules in 15,956 genes

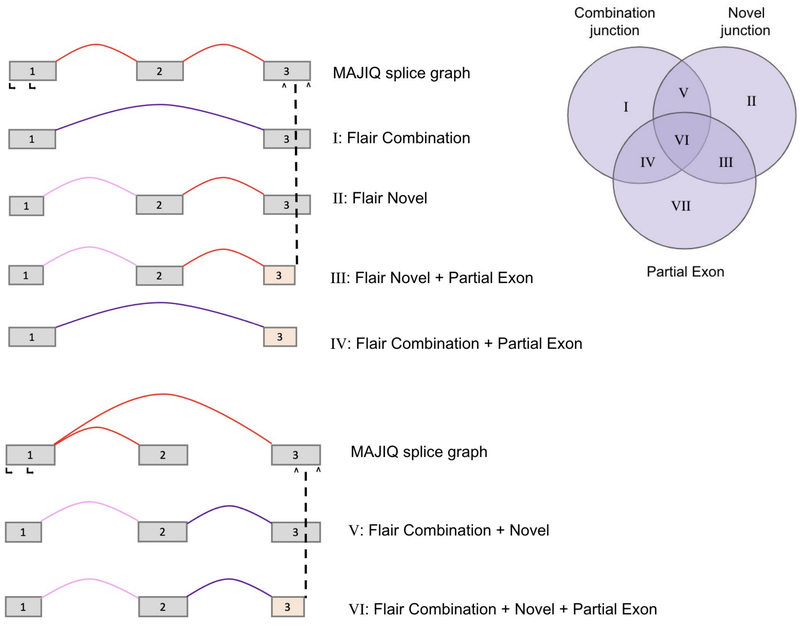


- Approximately 50% are supported by all source of information
- Over 35% are only detected by MAJIQ with or without annotation
  → vast majority of those expected to be true positive.
- Only 6.6% are detected by long read and annotation, over just 8% are only reported by long read
  → potential high rate of false positives.

Two types of MAJIQ de-novo



Break down Flair de-novo transcripts

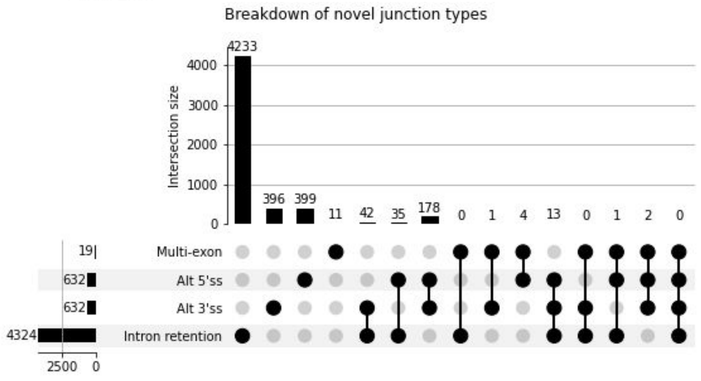


Flair de-novo transcripts analysis

I: Flair Combination	99
II: Flair Novel	454
III: Flair Novel + Partial Exon	4,722
IV: Flair Combination + Partial Exon	1,050
V: Flair Combination + Novel	22
VI: Flair Novel + Novel + Partial Exon	117
In progress categories	2,188

- VII: Partial exon exists as part of MAJIQ, annotation, or combination of these two, which means there cannot be pure partial exon
- Most cases fall in to III: Flair Novel + Partial Exon
- In progress of investigating few categories which don't match any of the established definitions

- Break down novel junction into four subcategories
- Most of the novel junctions fall into intron retention



Future Work

- Sub-categorize de-novo cases within short read and long read
- Investigate the splice sites disagreement in long read technology
- Test the effect of different long read algorithms and technology
- Create a unified visualization and analysis package of the three sources of information

References

1. Pardo-Palacios, Francisco, et al. "Systematic assessment of long-read RNA-seq methods for transcript identification and quantification." (2021).
2. Vaquero-Garcia, Jorge, et al. "A new view of transcriptome complexity and regulation through the lens of local splicing variations." *elife* 5 (2016): e11752.
3. Tang, Alison D., et al. "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns." *Nature communications* 11.1 (2020): 1-12.