PSI script updates and Limitations of Bedtools

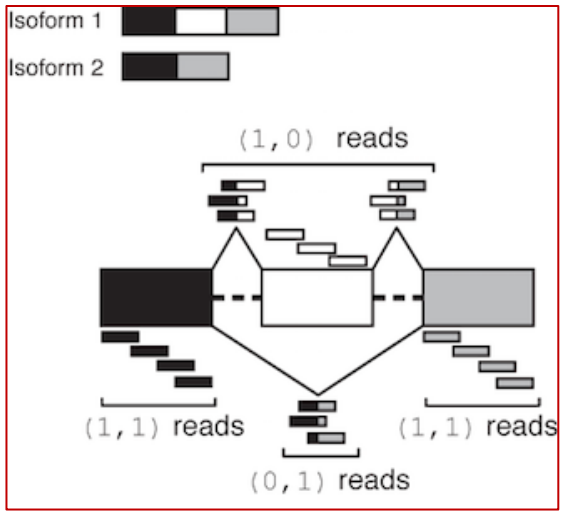
Giovanni Quinones Valdez

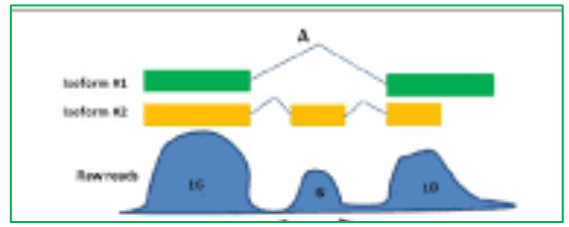
10/25/2018

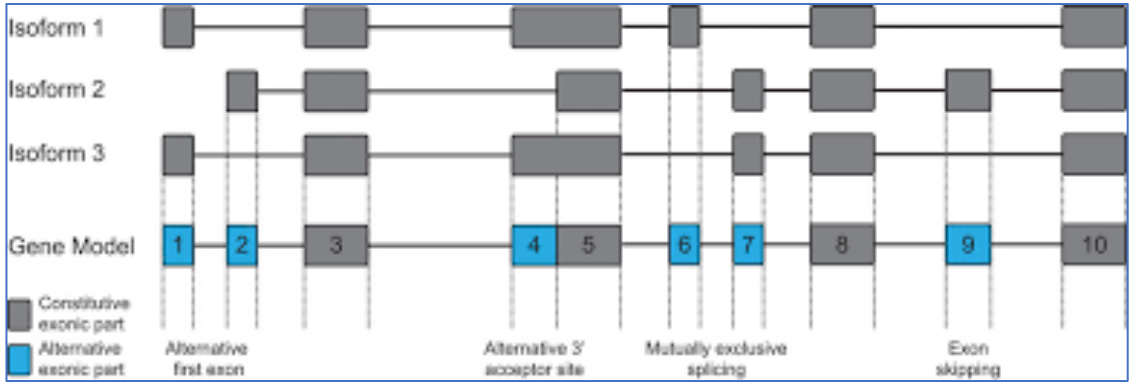
Xiao Lab

Different approaches for splicing quantification

- Isoform based PSI estimation (e.g. MISO):
  - Estimates PSI based on local isoforms, e.g. exon inclusion.
  - Doesn't account for complex splicing events.
- Isoform quantification (e.g. Cufflinks):
  - Estimate abundance of complete isoforms. In principle, the most ideal.
  - Not always reliable because of not enough informative reads, novel isoforms, etc
- Exon Segment PSI estimation (e.g. DEXSeq):
  - Estimates PSI of unique, non-overlapping region.
  - Most unbiased, but does not provide information about the splicing type.

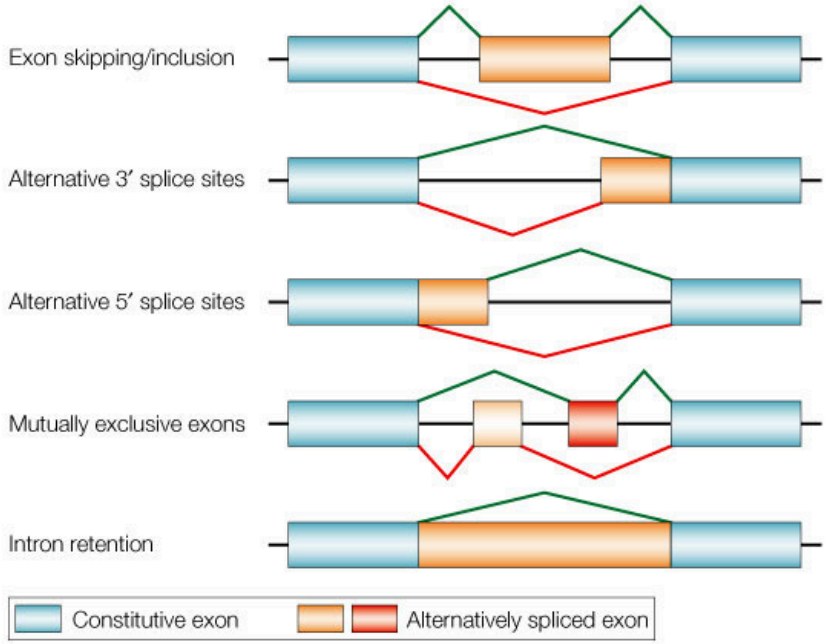


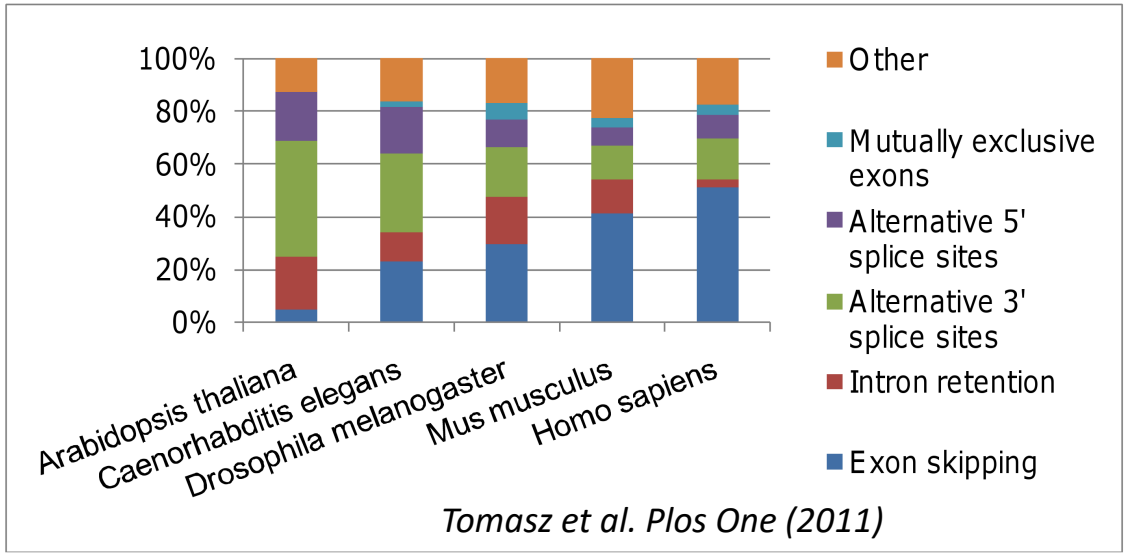




Alternative splicing

- There are five main types of alternative splicing.
- Most splicing software are limited to these types.
- About 20% of all alternative splicing events do not follow any of these patterns.





https://www.rna-seqblog.com/differential-splicing-analysis-with-rna-seq-current-applications-approaches-limitations/

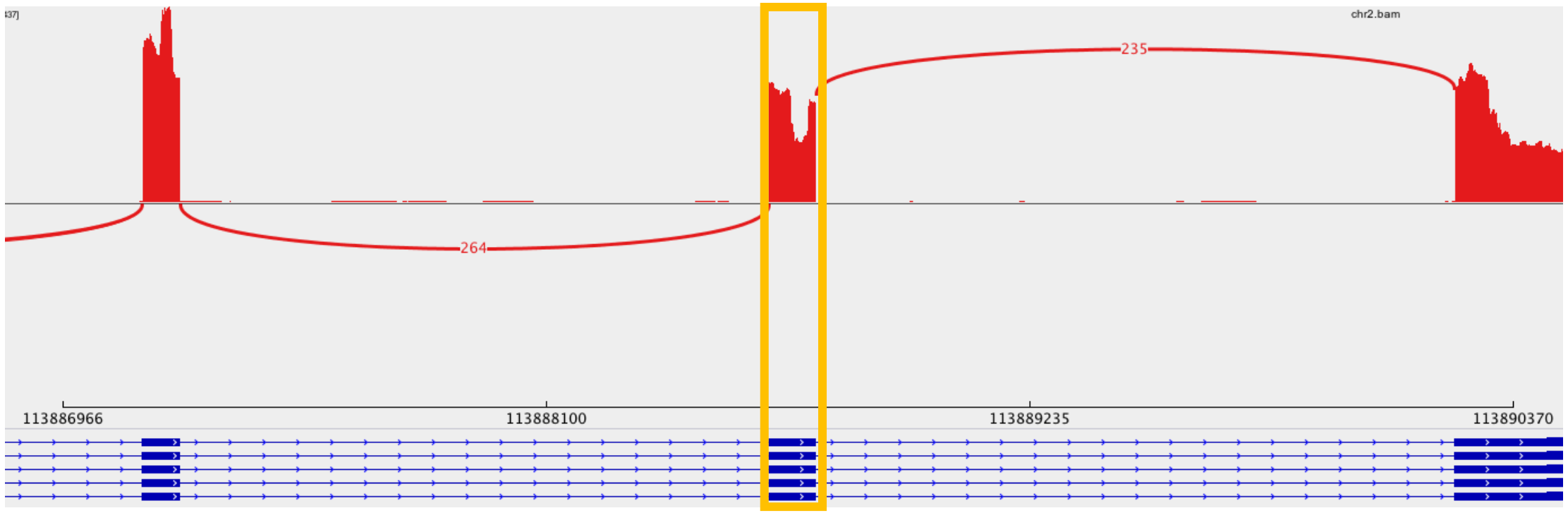
Issues with Bedtools

```
...running: command = bedtools coverage -b chr2.bam -a chr2.sub_annot.gtf -split
        chr2 tmp_annot exon 113888621 113888734 . + . partid 520

...running: command -F 0.08 18

...running: command -f 0.08 481

...running: command -F 8E-9 520

...running: command -f 8E-9 520

...running: command -s 244
```

- Different parameters will yield different coverage results.
- Current pipeline uses –F and –s options to estimate inclusion reads

```
Raw Manual count
        chr2 tmp_annot exon 113888621 113888734 . + . partid 520

Filtered Manual count
        chr2 tmp_annot exon 113888621 113888734 . + . partid 486
```

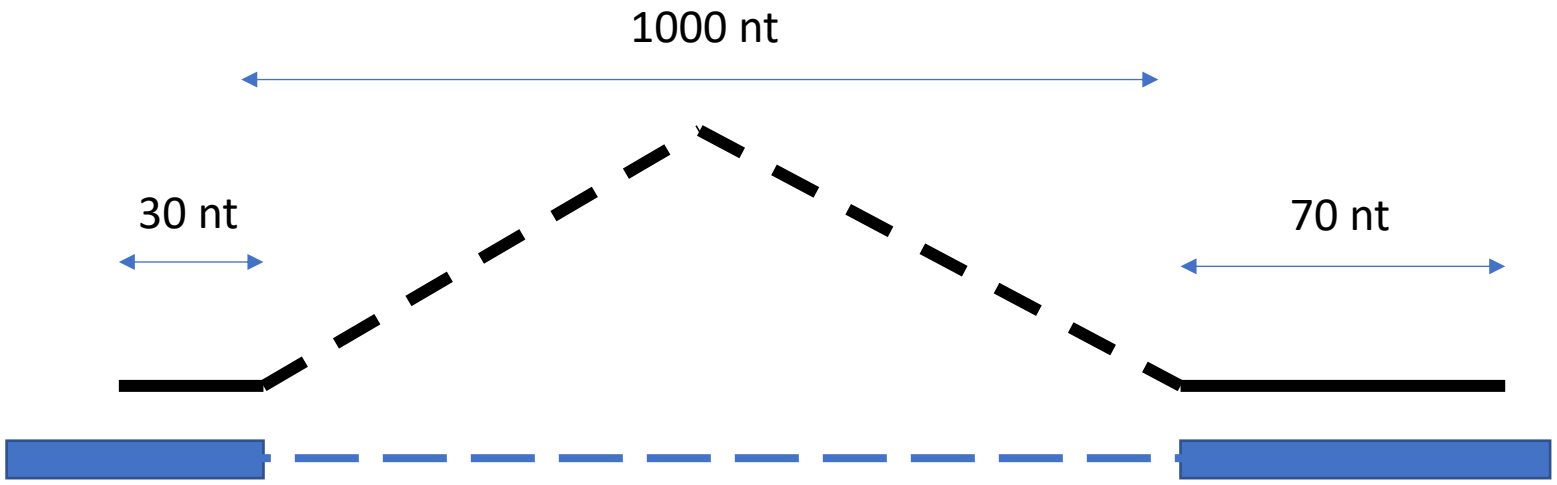
Visual coverage estimation



Exon length = 114 bases

Approximate coverage = 18 (100M) + 264(Junction1) + 235 (Junction2) = 517

```
...running: command = bedtools coverage -b chr2.bam -a chr2.sub_annot.gtf -split
        chr2 tmp_annot exon 113888621 113888734 . + . partid 520

...running: command -F 0.08 18
```



- The –F option (minimum overlap as a fraction of B, bam file) stablishes a minimum overlap.
- For sequencing reads, bedtools considers the whole span of the read, instead of the read length. For example, for an overlap of at least 8 bases in out example it should be, **F = 8/1100 instead of F = 8/100**

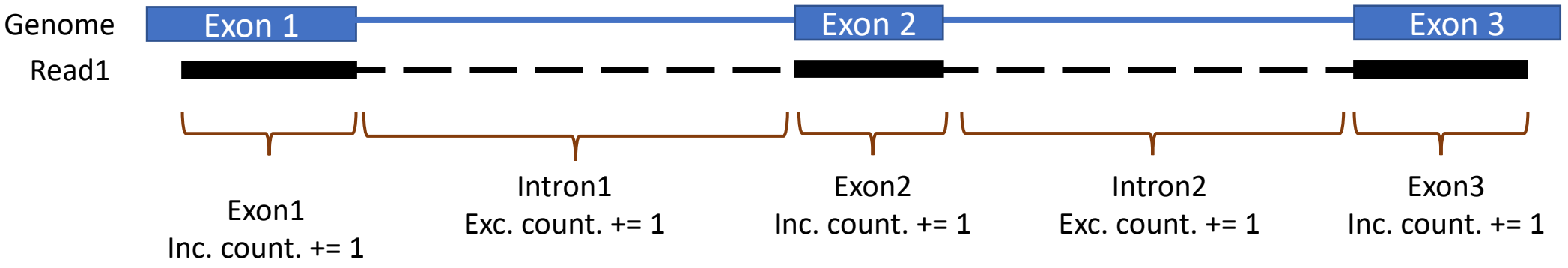
```
...running: command = bedtools coverage -b chr2.bam -a chr2.sub_annot.gtf -split
        chr2 tmp_annot exon 113888621 113888734 . + . partid 520

...running: command -s 244
```

					PYSAM Coverage by flag				Bedtools Coverage
CHROM	**START**	**END**	**STRAND**	**part_id**	**99**	**147**	**83**	**163**	**(-s)**
chr2	177040722	177040886	-	ENSG00000224189:002	276	**337**	0	0	**337**
chr2	242276797	242276848	+	ENSG00000168385:066	0	0	1077	**1001**	**1001**
chr2	55887292	55887328	-	ENSG00000138035:023	69	**91**	0	0	**91**
chr2	9547728	9548130	-	intron.ENSG00000119185:001	24	**30**	**1**	1	**31**

- The –s option in bedtools forces the overlaps to be strand-specific. When dealing with bam files, forward reads are “+” and reverse reads are “-” strand, which is not always the case.
- Bedtools doesn’t know the strandedness of the library. This can’t be directly inferred from the bam file.

Flag	Read Pair	Direction	Strand
99	Read1	Forward	-
147	Read2	Reverse	-
83	Read1	Reverse	+
163	Read2	Forward	+

New pipeline overview

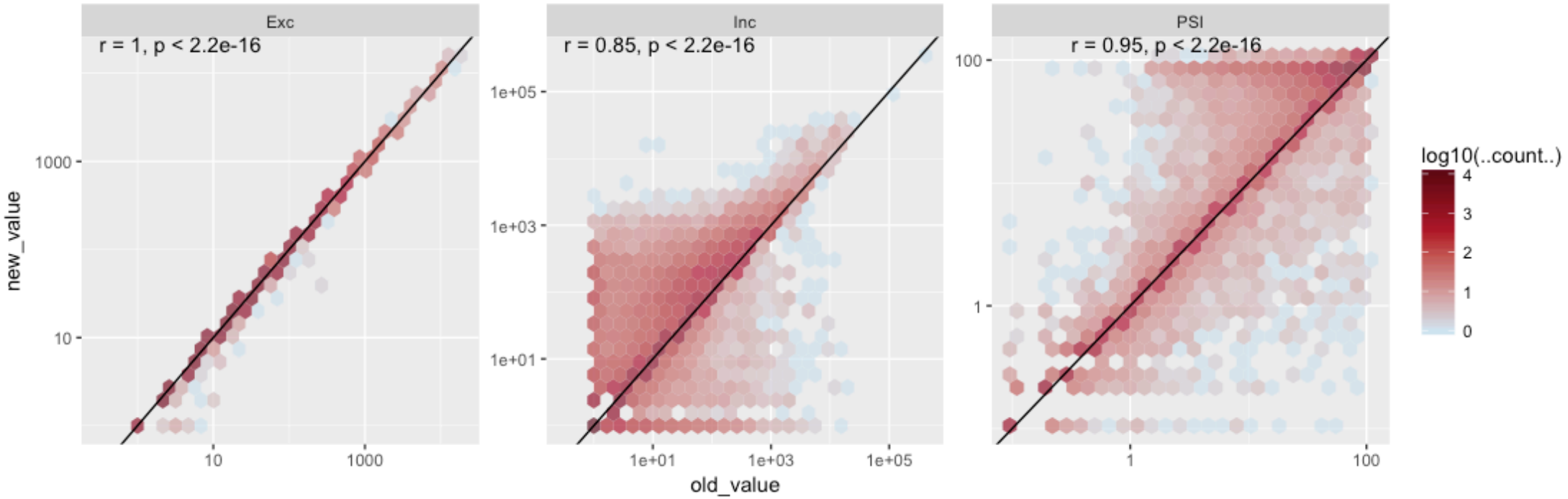


Approach for efficient overlap:

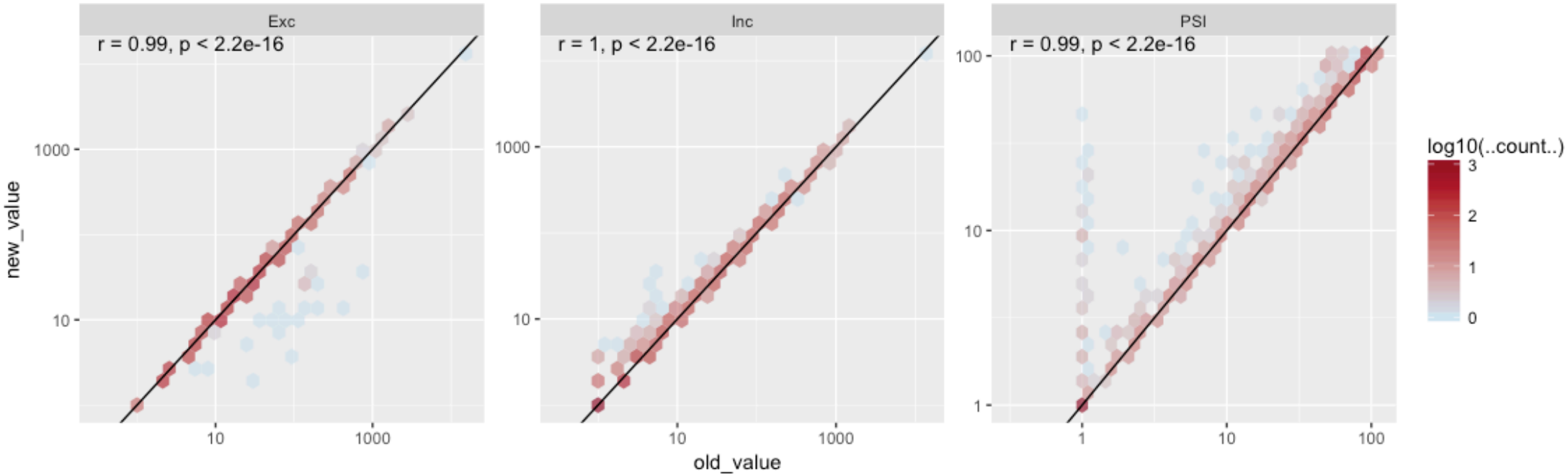
- For every read, identify all features (exonic parts, intronic parts) that overlap that read
- Sorted reads and sorted features do not require any ‘query search time’

- Memory usage (For a bam file with 82M reads):
  - Bedtools: 102 GB
  - Pysam: 0.6 GB
- Run time
  - Bedtools: 25 min.
  - Pysam: 32 min.

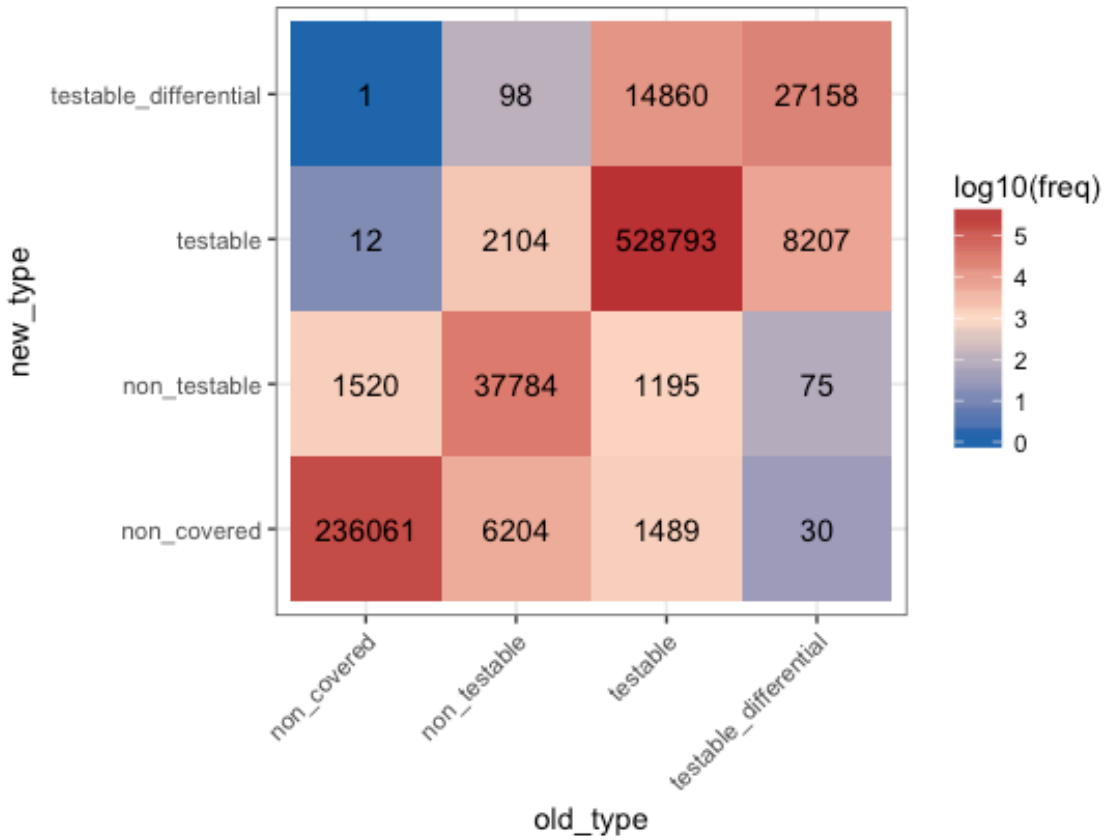
For macro-exons (length > 20 bases)

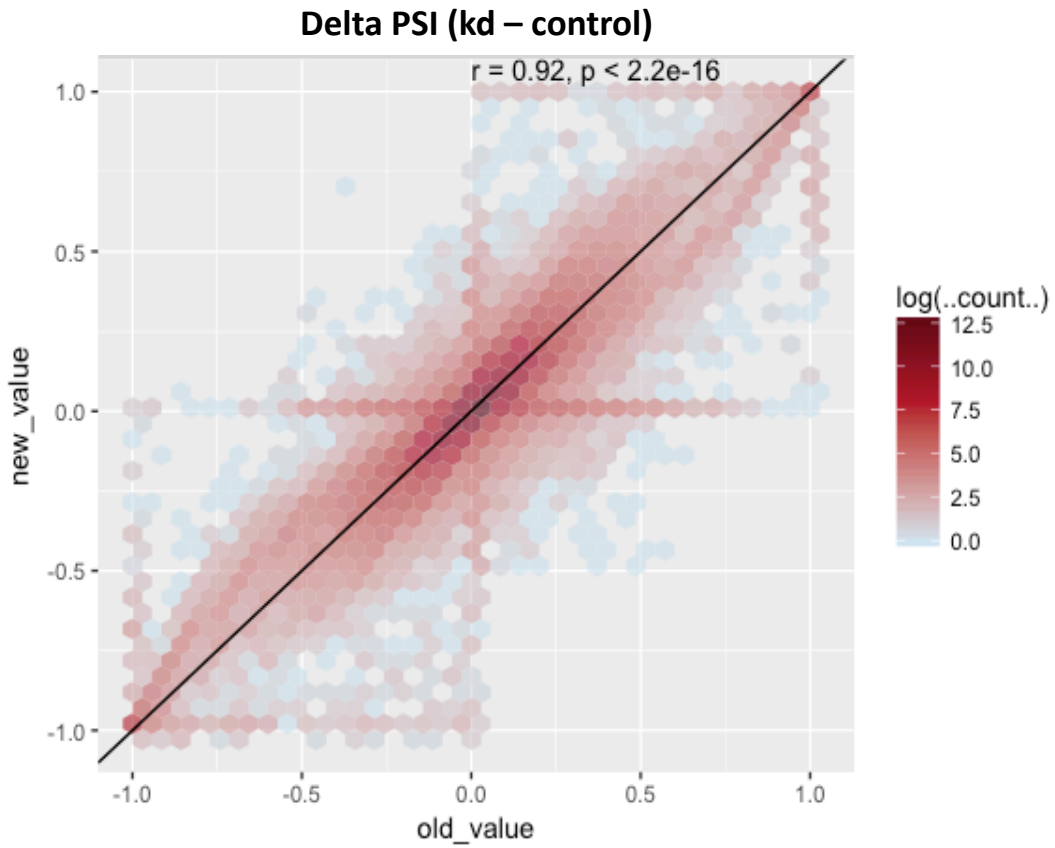


For macro-exons (length $\leq$ 20 bases)



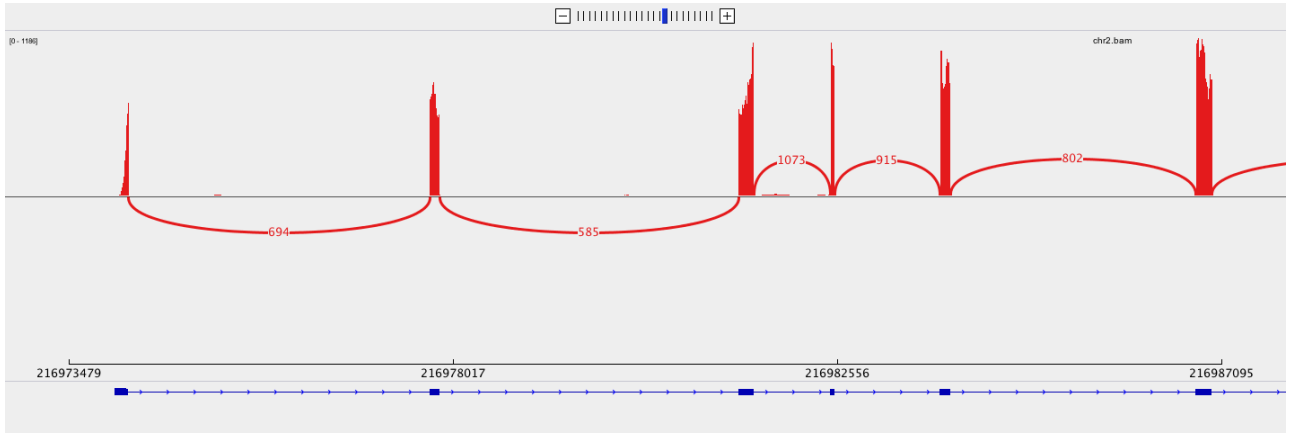
Comparison of delta PSI (old vs new script)





Limitations of our pipeline

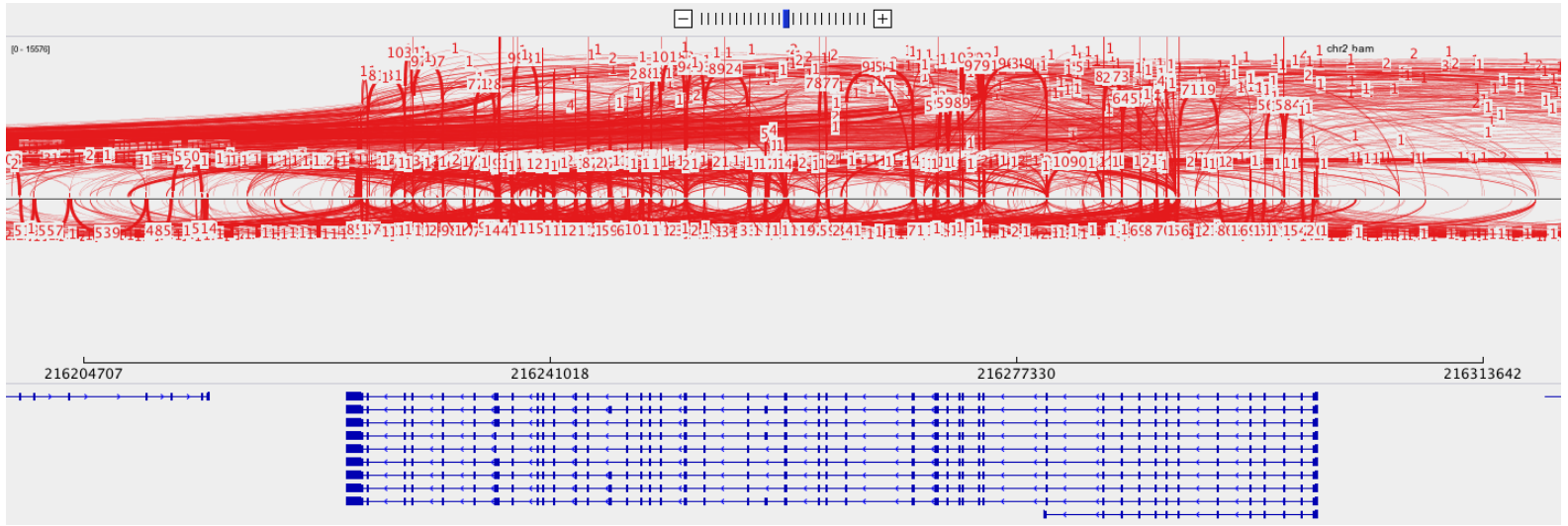
- PSI calculation for first exon
  - First exons do not contain exclusion reads. There is no upstream exon.
  - PSI ~ 1



- Ambiguous/false splice junctions

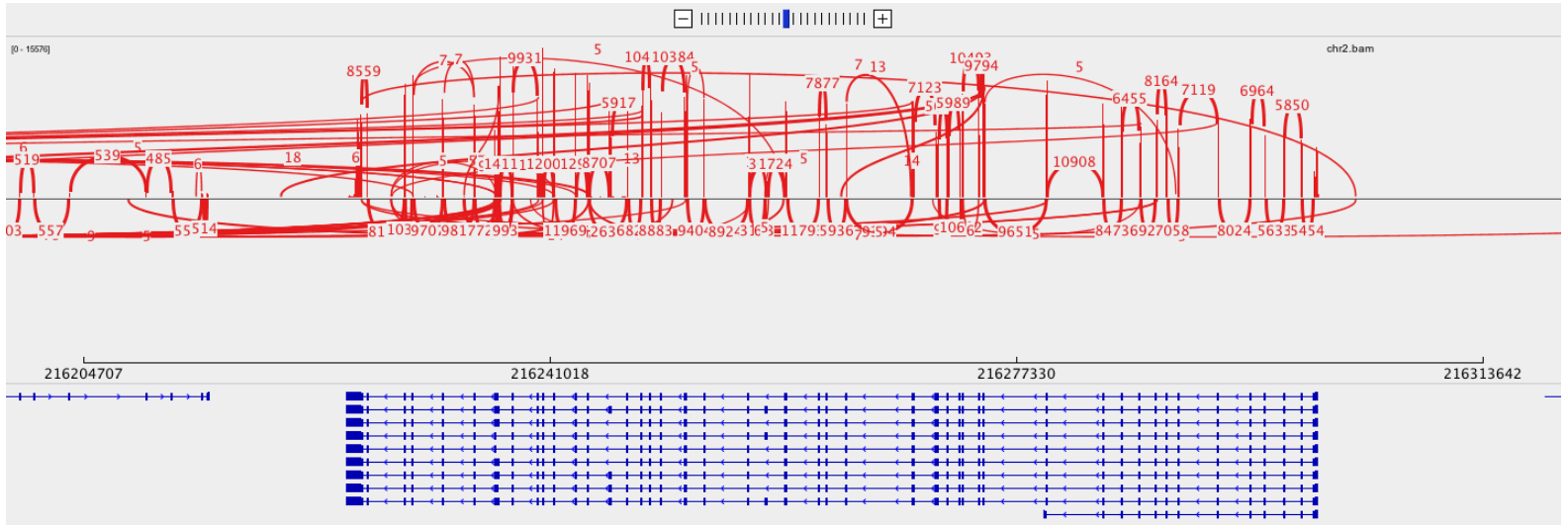
Complexity of splice junctions

Min. number of reads per junction = 1

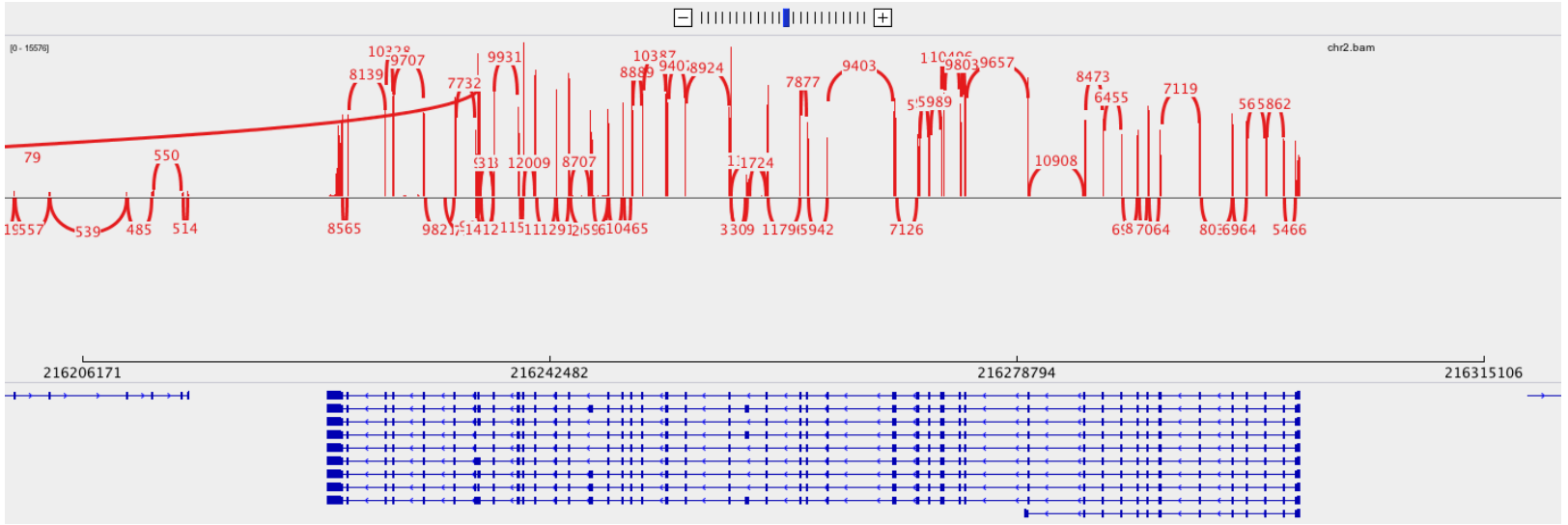


Average exon coverage ~ 10k reads

Min. number of reads per junction = 5



Min. number of reads per junction = 20



Splice-junction mapping ambiguity

```
Read:   CCGTCCCC
```

```
REF:   CCCGT------------------AGGTCCCCCC
Alg1:   CCGT----------------------CCCC (4M21N4M)
Alg2:   CC------------------------GTCCCC (2M21N6M)
```

- Junction starts and ends are not always consistent among reads. Some software deal with this:
  - STAR 2-round mapping
- Mainly due to mapping ambiguity.
- It's not a good decision to discard these reads, since they are not multi-mapped, spurious.