Applying edgeR for alternative splicing analysis

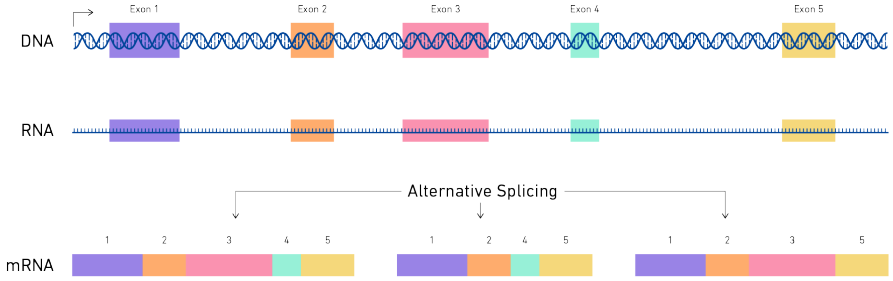
Lizhong Chen

Smyth Lab, Bioinformatics Division

Walter and Eliza Hall Institute of Medical Research

April 23, 2024

What is alternative splicing



https://www.technologynetworks.com/genomics/articles/alternative-splicing-importance-and-definition-351813

Alternative splicing using differential exon usage

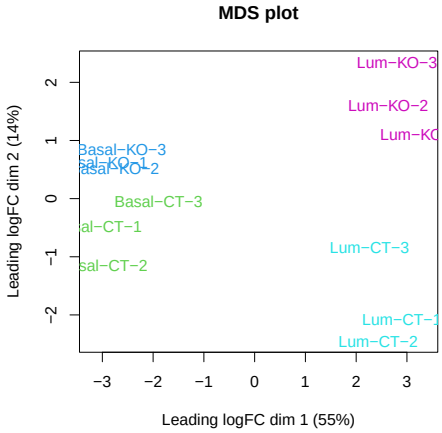
- Only able to find differential isoform usage
- Exon proportion change implies isoform proportion change

$$\text{Exon} \Rightarrow \text{Isoform}$$

- Isoform proportion change does not mean exon proportion change

$$\text{Isoform} \not\Rightarrow \text{Exon}$$

Foxp1 data set




```
> y$samples
           group lib.size norm.factors
Basal-CT-1 BasalCT  1.1e+07          0.99
Basal-CT-2 BasalCT  1.2e+07          1.02
Basal-CT-3 BasalCT  1.1e+07          1.09
Basal-KO-1 BasalKO  1.1e+07          0.89
Basal-KO-2 BasalKO  1.1e+07          1.02
Basal-KO-3 BasalKO  1.1e+07          0.95
Lum-CT-1     LumCT  1.2e+07          0.97
Lum-CT-2     LumCT  1.0e+07          1.09
Lum-CT-3     LumCT  1.1e+07          0.97
Lum-KO-1     LumKO  1.1e+07          0.98
Lum-KO-2     LumKO  1.1e+07          1.08
Lum-KO-3     LumKO  1.6e+07          0.98

> dim(y)
[1] 285931     12
```

Rsubread: align and featureCounts

```
# align
align(index        = indir,
      readfile1    = paste0("./FASTQ/", file),
      input_format = "gzFASTQ",
      output_file  = paste0("./BAM/", bam),
      nthreads     = 14)

# featureCounts
counts <- featureCounts(files = paste0("./BAM/", bam),
                        annot.inbuilt     = "mm39",
                        useMetaFeatures   = FALSE,
                        allowMultiOverlap = TRUE,
                        isPairedEnd       = FALSE,
                        nthreads          = 14)

# red: reference genome (RefSeq) or index
#      for annotation keep updating
```

```
# data
y <- featureCounts2DGEList(counts)

# filtering
keep <- filterByExpr(y)
y    <- y[keep,,keep.lib.sizes=FALSE]

> table(keep)
keep
 FALSE   TRUE
179116 106815

# normalization
y <- normLibSizes(y)

# orange: alternative splicing is sensitive to filtering
```

diffSplice: exon counts

 # Agtrap

	Basal-CT-1	Basal-CT-2	Basal-CT-3	Basal-KO-1	Basal-KO-2	Basal-KO-3	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
148161518	143	186	82	157	316	98	398	337	273	420	648	561
148166020	7	18	6	7	19	3	27	17	21	31	36	37
148166728	3	8	4	5	8	7	15	15	9	9	18	23
148168416	1	0	0	0	0	0	1	4	0	1	4	1
148172374	3	5	4	4	7	2	4	0	3	7	15	8

 # Acads

	Basal-CT-1	Basal-CT-2	Basal-CT-3	Basal-KO-1	Basal-KO-2	Basal-KO-3	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
115248358	56	76	43	60	111	42	176	220	92	143	201	229
115249153	6	11	8	6	15	4	24	24	13	16	15	20
115249378	10	4	7	10	17	1	26	26	10	16	25	20
115249696	9	15	5	17	23	6	32	46	15	30	48	37
115249916	26	22	12	21	34	13	47	64	26	39	60	51
115250299	18	28	8	24	38	8	50	70	17	38	51	55
115250824	9	15	12	11	23	8	36	31	23	32	31	40
115251123	10	25	11	22	36	7	47	52	22	32	44	58
115255624	15	21	6	15	41	8	40	44	34	32	54	49
115257284	6	7	5	3	12	3	18	16	6	11	24	21

 # red: large counts, a gene with less exons seems to have one or two large exon counts

diffSplice: exon counts

```
# head(cadvl)
```

	Basal-CT-1	Basal-CT-2	Basal-CT-3	Basal-KO-1	Basal-KO-2	Basal-KO-3	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
69901009	96	143	96	150	117	80	120	163	131	150	181	181
69901351	47	58	34	62	53	38	56	77	40	64	81	85
69901524	41	45	32	48	53	30	45	67	39	37	77	53
69901675	60	63	35	55	70	44	60	50	43	55	92	88
69901832	57	82	64	94	74	60	83	92	76	95	119	107
69901994	92	120	86	124	139	105	113	135	109	128	165	170
69902171	83	117	74	123	124	106	117	120	87	122	141	176
69902362	57	49	37	74	56	38	65	56	44	76	67	84
69902510	60	70	32	58	53	46	69	62	47	58	75	79
69903012	91	87	68	104	92	68	98	119	82	98	125	136
69903274	124	185	115	154	133	145	121	153	121	165	203	212
69903784	104	137	79	104	124	100	140	139	82	120	166	173
69903998	82	99	66	90	111	83	115	92	78	118	116	133
69904260	97	112	78	115	100	86	98	99	83	112	160	140
69905071	79	127	66	101	126	81	100	131	79	109	140	135
69905305	36	30	29	25	27	29	39	37	26	45	56	59
69905457	60	76	47	83	80	61	87	61	47	73	102	96
69905619	44	48	30	52	55	46	58	65	35	62	64	71
69905943	32	55	36	49	43	37	54	45	38	52	77	63
69906089	40	36	39	77	42	52	64	58	29	50	77	76

```
# red: relatively large counts, a gene with more exons seems to express randomly
```

diffSplice: pipeline using edgeR and limma

```r
# edgeR
y   <- estimateDisp(y, design)
fit <- glmQLFit(y, design, legacy = TRUE)

# diffsplice
dsp <- diffSplice(fit, contrast=contr,
       geneid = "GeneID", exonid = "Start")

# summarize results
# exon proportion change
topSplice(dsp, test = "exon")
# simes method for gene level
topSplice(dsp, test = "simes")
# F test for gene level
topSplice(dsp, test = "gene")

# visualization
plotSplice(dsp, geneid = "Foxp1", genecolname = "Symbol")
```

```r
# limma
vfit <- voomLmFit(y, design)
vfit <- contrasts.fit(vfit, contr)
vfit <- eBayes(vfit)

# diffsplice
vsp  <- diffSplice(vfit,
       geneid = "GeneID", exonid = "Start")

# summarize results
# exon proportion change
topSplice(vsp, test = "t")
# simes method for gene level
topSplice(vsp, test = "simes")
# F test for gene level
topSplice(vsp, test = "f")

# visualization
plotSplice(vsp, geneid = "Foxp1", genecolname = "Symbol")
```

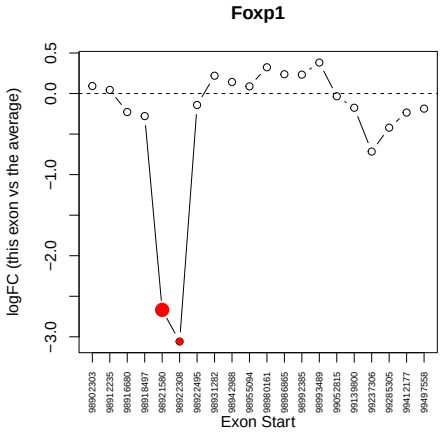
```r
# we are working on the S3 method for diffSplice, topSplice, and plotSplice
```

diffSplice: results using simes p-value

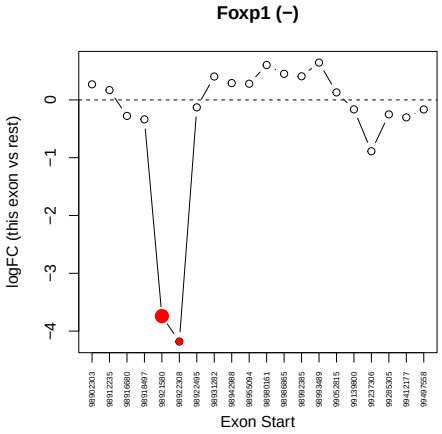
```
# edgeR diffsplice
> topSplice(dsp, test = "simes")
       GeneID   Chr Strand   Symbol NExons P.Value      FDR
49079  108655  chr6      -    Foxp1     20 4.2e-14 4.7e-10
122571 208263  chr1      - Tor1aip1     11 1.4e-11 8.0e-08
9811   102436  chr9      +    Lars2      9 4.0e-06 1.5e-02
9555   102103  chr8      -    Mtus1     11 3.0e-05 8.3e-02
62164   11687 chr11      -   Alox15     13 1.2e-04 2.7e-01
79298   13518  chr1      +      Dst     93 4.8e-04 7.9e-01
69092   12095  chr3      -   Bglap3      6 4.9e-04 7.9e-01
213798  56460  chr7      +     Pkp3     14 9.0e-04 1.0e+00
184804 319448 chr14      -   Fndc3a     27 1.0e-03 1.0e+00
140303  22228  chr7      +     Ucp2      9 1.4e-03 1.0e+00

# teal: expected results
# red : could be an outier for edgeR
```

```
# limma diffsplice
> topSplice(vsp, test = "simes")
       GeneID   Chr Strand   Symbol NExons P.Value      FDR
49079  108655  chr6      -    Foxp1     20 4.4e-11 4.9e-07
122571 208263  chr1      - Tor1aip1     11 5.7e-07 3.2e-03
62164   11687 chr11      -   Alox15     13 2.4e-06 8.9e-03
100783  17758  chr9      +     Map4     21 4.6e-05 1.3e-01
177764  26942 chr15      +    Spag1      4 9.7e-05 2.2e-01
273944  76866  chr4      +    Morn1      4 2.8e-04 5.1e-01
86220   14683  chr2      +     Gnas     14 3.2e-04 5.1e-01
17583  102634756 chrX     + Gm32262      2 4.4e-04 6.2e-01
9555   102103  chr8      -    Mtus1     11 5.4e-04 6.6e-01
138111 219249 chr14      +    Tdrd3     13 7.3e-04 8.1e-01
```

diffSplice: visualization for Foxp1



edgeR

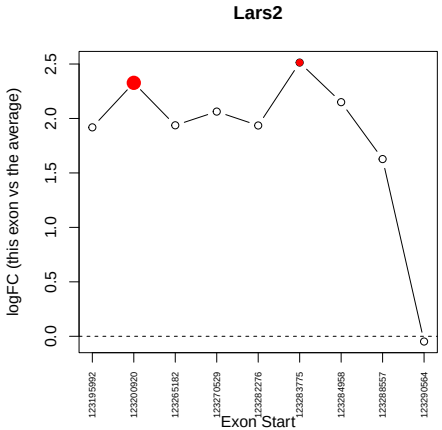


limma

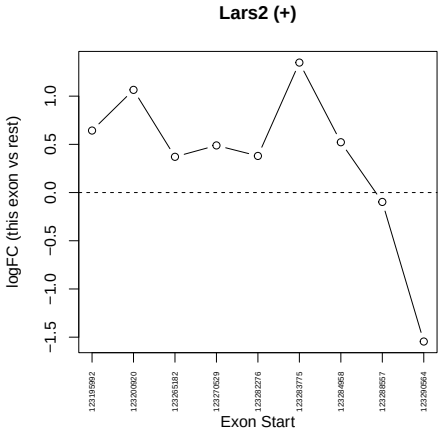
diffSplice: edgeR vs limma

- edgeR and limma share the same idea and results are similar
- edgeR performs one vs the average on the exon level test
- limma performs one vs the rest on the exon level test
- limma is more robust to the potential outliers, e.g Lars2

diffSplice: visualization for Lars2



edgeR



limma

diffSplice: exon counts for Lars2

```
# Lars2
```

	Basal-CT-1	Basal-CT-2	Basal-CT-3	Basal-KO-1	Basal-KO-2	Basal-KO-3	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
123195992	3	3	4	2	3	6	12	14	2	7	10	12
123200920	5	6	8	8	4	7	20	10	18	17	28	31
123206766	3	8	3	2	5	3	10	11	7	9	6	4
123221848	2	4	8	2	7	4	2	4	4	1	8	7
123224016	3	0	1	0	1	0	1	3	0	1	0	4
123238679	2	0	1	0	0	0	2	3	4	3	3	6
123240938	2	2	5	4	5	0	5	9	6	8	7	11
123242082	0	0	3	0	0	0	0	0	1	0	0	0
123247244	1	5	1	0	2	3	4	5	9	4	5	12
123247700	4	3	6	4	5	3	5	6	2	8	6	6
123256512	0	0	1	1	1	3	5	2	3	6	6	5
123260951	4	0	2	0	2	2	12	7	6	7	11	4
123265182	8	11	6	1	5	6	15	21	13	9	19	25
123267200	0	3	2	1	4	1	7	6	7	5	5	10
123270529	5	5	5	2	4	1	11	11	11	12	15	12
123281810	2	3	6	6	4	5	13	2	11	5	7	10
123282276	7	5	8	4	10	4	13	17	15	11	17	20
123283775	2	8	5	3	7	3	9	7	15	17	19	23
123284025	2	1	3	2	4	1	10	3	11	6	8	9
123284958	5	5	2	2	2	5	7	10	9	3	16	16
123288557	8	7	9	4	5	5	18	24	21	10	16	24
123290564	2205	1379	1353	7522	1024	1066	1822	43934	2038	2093	1663	3142

```
# orange: filtered by filterByExpr
# blue: line of interest; red: outlier
```

diffSplice: aim

- Assume a gene g has m exons
- Let Y_{ij} and $Z_{ij'}$ be the counts of exon i for sample j and j' in two groups k and k'.
- Let p_i and q_i be the relative proportion of exon i in two groups k and k'

$$p_i = \frac{\mu_{ik}}{\sum_i \mu_{ik}} \quad \text{and} \quad q_i = \frac{\mu_{ik'}}{\sum_i \mu_{ik'}}$$

  where μ_{ik} and $\mu_{ik'}$ are the group means for exon i

- The aim is test

$$p_i = q_i$$

diffSplice: Exon level

- Let β_i be the log fold change for exon i
- Let β_g be the log fold change for gene g
- β_g is the same as the average log fold change for all exons
- Then testing

$$p_i = q_i$$

  is equivalent to testing

$$\beta_i = \beta_g$$

- The EB process is performed on gene level σ_g^2, estimated using all exons

diffSplice: Gene level Simes p-value

- Let p_i be the p-values for m tests of exons $H_i : \beta_i = \beta_g$
- The m tests for exons are correlated and $m-1$ of m tests are independent
- The Simes p-value p_g^s for the gene g is

$$p_g^s = \min_j \{(m-j)p_j'\}$$

  where p_j' is the ordered p-values

- p_g^s is sensitive to the extremely small p-values from one test
- p_g^s is sensitive to the significant log fold change for one exon

diffSplice: Gene level F-test

- Let H_g be the test for the gene g

$$H_g : \beta_1 = \beta_2 = \cdots = \beta_m = \beta_g$$

- Let p_g^f be the p-value for the F test of the gene g

$$F = (s_g^2)^{-1} \Delta \ell \sim F_{m-1, d_g + d_0}$$

  where s_g^2 is the posterior of σ_g^2 and d_0 is the prior df

- p_g^f is sensitive to the sum of log fold change among exons

diffSplice: expression vs usage

- Expression analysis only cares about the absolute counts
- Usage analysis cares about the relative proportion
- Filtering affects expression analysis through FDR and the EB process indirectly
- Filtering affects usage analysis by direct proportion calculation
- Usage analysis depends on a complete reference set

diffSplice: filtering

- `filterByExpr` may filter exons causing alternative splicing
- Approach I: filter those genes with zero counts
- Approach II. filter those exons with zero counts
- Both keep the same exons with non-zero counts
- Approach II may have slightly smaller simes p-value

diffSplice: edgeR-v4 QL method

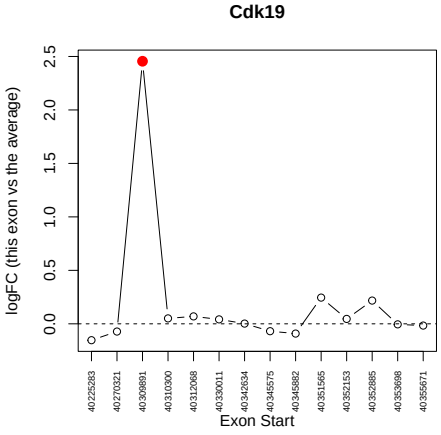
```
# edgeR-v4: filtering exon with zero counts
> topSplice(dsp, test = "simes")
       GeneID   Chr Strand   Symbol NExons P.Value      FDR
49079  108655  chr6      -    Foxp1     32 1.7e-15 3.8e-11
122571 208263  chr1      - Tor1aip1     13 1.4e-11 1.6e-07
9811   102436  chr9      +    Lars2     22 1.1e-08 8.6e-05
217841  59013 chr11      +  Hnrnph1     14 4.0e-08 2.2e-04
277722  78334 chr10      +    Cdk19     14 1.0e-07 4.5e-04
100783  17758  chr9      +     Map4     21 4.9e-06 1.9e-02
140303  22228  chr7      +     Ucp2      9 7.7e-06 2.5e-02
184805 319448 chr14      -   Fndc3a     31 1.1e-05 3.0e-02
88907   15257  chr3      -    Hipk1     18 1.8e-05 4.4e-02
9558   102103  chr8      -    Mtus1     20 2.6e-05 5.9e-02

# teal: expected results
# red : new results using simes method
# blue: new results using F test
```

```
# edgeR-v4: filtering exon with zero counts
> topSplice(dsp, test = "gene")
       GeneID   Chr Strand   Symbol NExons gene.F P.Value      FDR
9811   102436  chr9      +    Lars2     22   18.3 1.2e-36 2.8e-32
49079  108655  chr6      -    Foxp1     32    6.5 2.6e-19 2.9e-15
216231  57738 chr16      - Slc15a2     23    4.7 6.3e-10 4.8e-06
122571 208263  chr1      - Tor1aip1     13    6.6 5.5e-09 3.1e-05
69092   12095  chr3      -   Bglap3      7    8.9 4.2e-07 1.9e-03
9558   102103  chr8      -    Mtus1     20    3.8 7.9e-07 3.0e-03
150669 229487  chr3      +     Gatb     15    4.2 3.9e-06 1.3e-02
192361 329977  chr4      -    Fhad1     34    2.6 8.5e-06 2.4e-02
136170 218194 chr13      + Phactr1     18    3.2 4.2e-05 1.1e-01
193001 330450  chr6      +     Far2     12    3.9 9.0e-05 2.0e-01
```

```
# filter zero counts
> keep <- rowSums(y$counts) > 0
> table(keep)
keep
 FALSE   TRUE
 74741 211190
```

```
# filter by filterByExpr
> keep <- filterByExpr(y)
> table(keep)
keep
 FALSE   TRUE
179116 106815
```

edgeR-v4: Cdk19

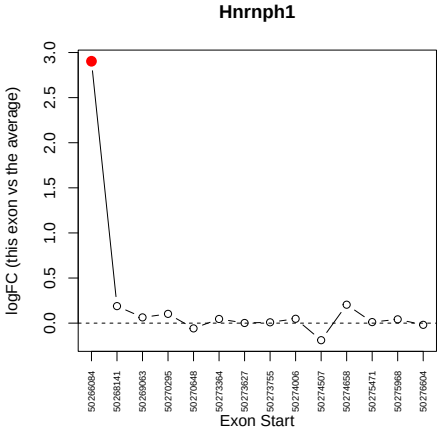



```
# Cdk19
```

	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
40225283	57	44	40	40	79	73
40270321	27	25	26	27	37	53
40309891	1	4	0	9	24	73
40310300	39	39	47	47	70	96
40312068	51	65	61	67	98	143
40330011	10	7	11	2	16	31
40342634	65	62	70	65	99	160
40345575	55	48	58	55	72	118
40345882	26	23	37	28	35	66
40351565	17	20	28	32	44	58
40352153	42	45	64	48	76	139
40352885	20	28	33	31	54	80
40353698	82	88	96	72	158	204
40355671	1159	1270	1715	1280	2163	3305

```
# orange: filtered by filterByExpr
```

edgeR-v4: Hurnph1

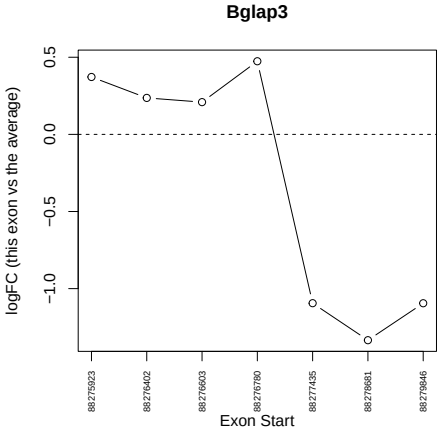



```
# Hnrnph1
```

	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
50266084	0	2	0	13	10	10
50268141	11	17	8	13	8	13
50269063	331	268	260	236	169	322
50270295	517	485	490	434	331	538
50270648	2642	2862	2434	1926	1514	2460
50273364	854	783	761	637	512	836
50273627	322	330	290	234	191	323
50273755	630	530	503	410	316	616
50274006	534	431	499	401	298	518
50274507	308	207	314	184	131	229
50274658	297	343	294	307	221	373
50275471	536	481	420	369	280	503
50275968	321	305	269	227	183	334
50276604	2733	2410	2594	1900	1442	2721

```
# orange: filtered by filterByExpr
```

edgeR-v4: Bglap3




```
# Bglap3
```

	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
88275923	426	339	228	577	208	325
88276402	28	29	24	39	25	16
88276603	6	8	2	6	5	5
88276780	345	255	224	562	206	236
88277435	228	286	112	65	60	37
88278681	136	184	77	29	24	31
88279846	130	209	87	41	41	30

edgeR-v4 vs edgeR-v3

- edgeR-v4 extends to low counts without filtering
- edgeR-v4 is able to find those exon filtered before, causing alternative splicing
- edgeR-v4 is sensitive to significant change in low counts, especially

  Zeros *vs.* Non-zeros

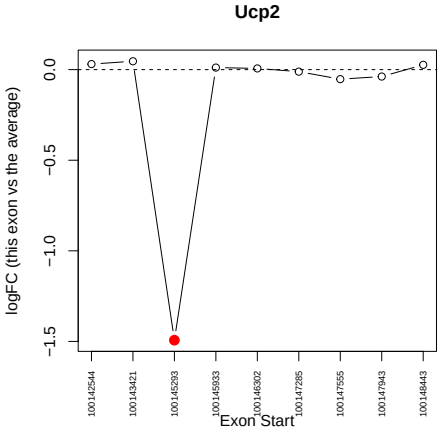
- edgeR-v3 fails to perform differential expression analysis for low counts
- edgeR-v3 performs better in usage than expression analysis for low counts

edgeR-v4 vs edgeR-v3: results

```
# edgeR-v4
> topSplice(dsp, test = "simes")
       GeneID   Chr Strand   Symbol NExons P.Value      FDR
49079  108655  chr6      -    Foxp1     32 1.7e-15 3.8e-11
122571 208263  chr1      - Tor1aip1     13 1.4e-11 1.6e-07
9811   102436  chr9      +    Lars2     22 1.1e-08 8.6e-05
217841  59013 chr11      +  Hnrnph1     14 4.0e-08 2.2e-04
277722  78334 chr10      +    Cdk19     14 1.0e-07 4.5e-04
100783  17758  chr9      +     Map4     21 4.9e-06 1.9e-02
140303  22228  chr7      +     Ucp2      9 7.7e-06 2.5e-02
184805 319448 chr14      -   Fndc3a     31 1.1e-05 3.0e-02
88907   15257  chr3      -    Hipk1     18 1.8e-05 4.4e-02
9558   102103  chr8      -    Mtus1     20 2.6e-05 5.9e-02

# teal: expected results
# red : new result identified by edgeR-v4

# edgeR-v4 has more power than edgeR-v3
```

```
# edgeR-v3
> topSplice(dsp, test = "simes")
       GeneID   Chr Strand   Symbol NExons P.Value      FDR
49079  108655  chr6      -    Foxp1     32 9.3e-14 2.1e-09
122571 208263  chr1      - Tor1aip1     13 2.2e-11 2.4e-07
9811   102436  chr9      +    Lars2     22 5.9e-08 4.5e-04
277722  78334 chr10      +    Cdk19     14 3.7e-07 2.1e-03
217841  59013 chr11      +  Hnrnph1     14 7.3e-06 3.3e-02
9558   102103  chr8      -    Mtus1     20 4.5e-05 1.7e-01
62164   11687 chr11      -   Alox15     15 1.5e-04 4.7e-01
69092   12095  chr3      -   Bglap3      7 2.7e-04 7.5e-01
79298   13518  chr1      +      Dst    107 3.9e-04 9.7e-01
229432  66793 chr16      +     Clxn     13 5.3e-04 1.0e+00
```

edgeR-v4: Ucp2




```
# Ucp2
          Lum-CT-1 Lum-CT-2 Lum-CT-3 Lum-KO-1 Lum-KO-2 Lum-KO-3
100142544      113      150       99      121      158      209
100143421      275      314      241      273      366      505
100145293       18       13        8        5        2        4
100145933      347      377      303      319      442      609
100146302      244      295      213      214      337      450
100147285      287      345      241      265      383      483
100147555      253      262      214      216      280      419
100147943      319      409      309      293      430      598
100148443      590      723      593      592      795     1215

# red: edgeR-v4 is more sensitive to significant change
#      in small counts because of the biological variation
#      estimation, especially with many zeros
```

diffSplice: reference genome (RefSeq annotation)

- The reference genomes (RefSeq annotation) like mm39, hg38, keep updating
- More exons may be annotated for some genes
- The exon count matrix may change, causing different results
- More exons are counted using updated mm39

diffSplice: updated mm39 reference genome

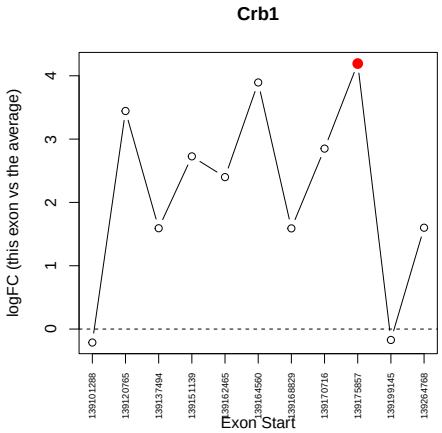
```
# edgeR-v4: updated reference "mm39"
> topSplice(dsp, test = "simes")
 GeneID    Chr Strand        Symbol NExons P.Value      FDR
 108655  chr6      -         Foxp1     33 1.2e-15 2.9e-11
 208263  chr1      -      Tor1aip1     15 6.0e-10 7.0e-06
 102436  chr9      +          Lars2     22 1.1e-08 8.8e-05
  59013 chr11      +        Hnrnph1     14 3.7e-08 2.2e-04
  78334 chr10      +          Cdk19     14 1.1e-07 5.2e-04
  17758  chr9      +           Map4     21 5.4e-06 2.1e-02
 319448 chr14      -         Fndc3a     31 1.1e-05 3.6e-02
  15257  chr3      -          Hipk1     18 1.9e-05 5.5e-02
  26894  chr6      -         Cops7a     11 2.3e-05 5.9e-02
 654318  chr4      - C530005A16Rik      3 2.8e-05 6.3e-02

# teal: expected results
# red : new results using F test with updated reference genome
```

```
# edgeR-v4: updated reference "mm39"
> topSplice(dsp, test = "gene")
   GeneID    Chr Strand   Symbol NExons gene.F P.Value      FDR
   102436   chr9      +    Lars2     22   18.3 1.5e-36 3.5e-32
   108655   chr6      -    Foxp1     33    6.4 1.1e-19 1.3e-15
    13411  chr12      -    Dnah11    85    2.6 2.6e-12 2.0e-08
    57738  chr16      -   Slc15a2    23    4.8 5.5e-10 3.2e-06
   170788   chr1      -     Crb1     11    8.3 8.0e-09 3.7e-05
    69707  chr16      -     Iqcg     13    6.0 1.7e-08 6.8e-05
100048534  chr19      -    Cfap43    39    3.1 2.8e-08 9.3e-05
   208263   chr1      - Tor1aip1     15    5.0 1.3e-07 3.7e-04
    12095   chr3      -    Bglap3     7    8.9 4.0e-07 1.0e-03
   102103   chr8      -    Mtus1     24    3.2 3.1e-06 7.3e-03
```

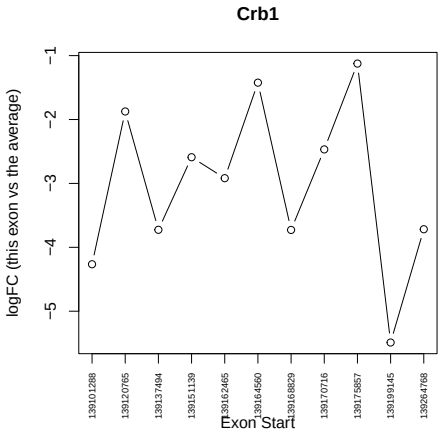
```
# updated reference
> keep <- rowSums(y$counts) > 0
> table(keep)
keep
 FALSE   TRUE
 68986 217522
```

```
# original reference
> keep <- rowSums(y$counts) > 0
> table(keep)
keep
 FALSE   TRUE
 74741 211190
```

edgeR-v4: Crb1



updated reference



originally from Foxp1.Rdata

edgeR-v4: Crb1

```
# Crb1: updated reference
```

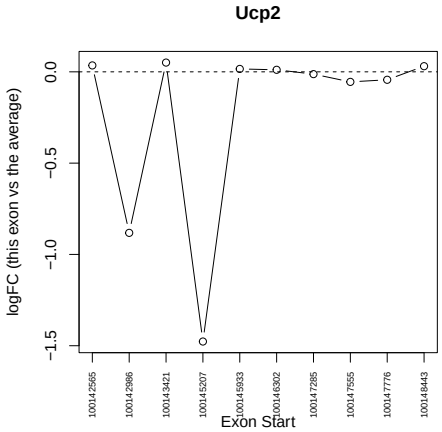
	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
139101288	100	134	144	106	115	212
139120765	0	0	0	2	7	6
139137494	0	0	0	1	0	1
139151139	0	0	0	3	3	1
139162465	0	0	0	1	2	2
139164560	0	0	0	10	9	4
139168829	0	0	0	0	1	1
139170716	0	0	0	3	3	2
139175857	0	0	0	14	11	6
139199145	0	0	0	0	0	0
139264768	0	0	0	2	0	0

```
# Crb1: originally from Foxp1.Rdata online
```

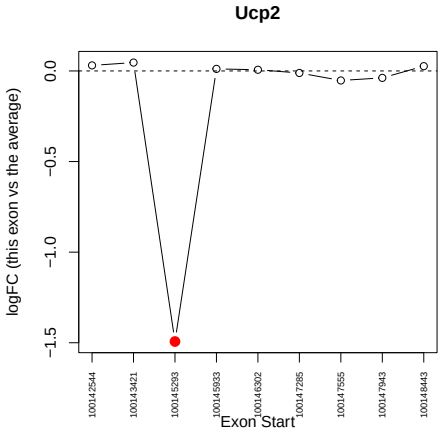
	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
139101288	0	0	0	0	0	1
139120765	0	0	0	2	7	6
139137494	0	0	0	1	0	1
139151139	0	0	0	3	3	1
139162465	0	0	0	1	2	2
139164560	0	0	0	10	9	4
139168829	0	0	0	0	1	1
139170716	0	0	0	3	3	2
139175857	0	0	0	14	11	6
139199145	0	0	0	0	0	0
139264768	0	0	0	2	0	0

```
# red: much more counts for first exon
#      it seems Rsubread misses the first exon using original genome reference

# edgeR will fail in the case that the gene has zero counts in one group, meaning no alternative splicing
```

edgeR-v4: Ucp2 update



updated reference



originally from Foxp1.Rdata

edgeR-v4: Ucp2 update

```
# Ucp2: updated reference
```

	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
100142565	113	150	99	121	158	209
100142986	3	14	0	0	3	6
100143421	275	314	241	273	366	505
100145207	19	14	9	5	3	4
100145933	347	377	303	319	442	609
100146302	244	295	213	214	337	450
100147285	290	349	241	265	385	483
100147555	256	266	214	216	282	419
100147776	329	419	317	297	431	618
100148443	590	723	593	592	795	1215

```
# Ucp2: originally from Foxp1.Rdata online
```

	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
100142544	113	150	99	121	158	209
100143421	275	314	241	273	366	505
100145293	18	13	8	5	2	4
100145933	347	377	303	319	442	609
100146302	244	295	213	214	337	450
100147285	287	345	241	265	383	483
100147555	253	262	214	216	280	419
100147943	319	409	309	293	430	598
100148443	590	723	593	592	795	1215

```
# red : one more exon in updated reference genome
# blue: the exon counts change when updating the reference genome
```

Rsubread: `align` vs `subjunc`

- `align` reports the largest mappable regions for each read and soft-clips the remainder of the read
- `subjunc` reports discovered exon-exon junctions and it also performs full alignments for every read including exon-spanning reads
- `subjunc` requires the presence of donor and receptor sites when calling exon-exon junctions
- `subjunc` is more suitable for the exon level analysis
- `subjunc` may report more exon counts

diffSplice: exon counts from subjunc

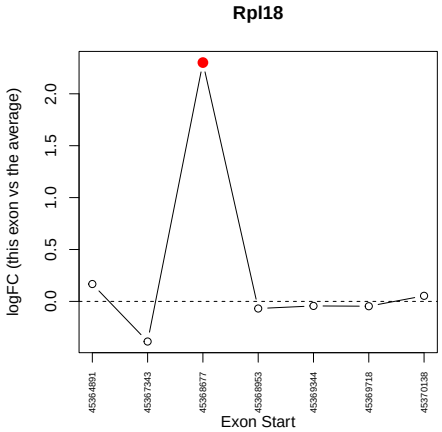
```
# edgeR-v4: exon counts from subjunc
> topSplice(dsp, test = "simes")
 GeneID    Chr Strand         Symbol NExons P.Value      FDR
 108655  chr6      -          Foxp1     34 8.2e-21 1.9e-16
 208263  chr1      -       Tor1aip1     15 3.9e-15 4.6e-11
  19899  chr7      +          Rpl18      7 3.9e-10 3.0e-06
  78334 chr10      +          Cdk19     14 1.6e-09 9.6e-06
 102436  chr9      +          Lars2     22 1.1e-08 5.2e-05
  59013 chr11      +       Hnrnph1     14 3.5e-08 1.4e-04
  17758  chr9      +           Map4     21 5.2e-07 1.8e-03
 101772  chr7      -           Ano1     30 9.6e-07 2.8e-03
 654318  chr4      - C530005A16Rik      3 4.1e-06 1.1e-02
 234663  chr8      -       Dync1li2     13 5.9e-06 1.4e-02

# teal: expected results
# red : new result using simes method
#       with exon counts from subjunc
```

```
# edgeR-v4: exon counts from subjunc
> topSplice(dsp, test = "gene")
    GeneID    Chr Strand   Symbol NExons gene.F P.Value      FDR
    102436   chr9      +    Lars2     22   25.2 4.7e-44 1.1e-39
    108655   chr6      -    Foxp1     34    8.5 1.4e-27 1.6e-23
     13411  chr12      -    Dnah11    85    2.6 2.0e-12 1.6e-08
     69707  chr16      -     Iqcg     13    8.4 6.0e-12 3.6e-08
 100048534  chr19      -   Cfap43     39    4.0 1.6e-11 7.4e-08
    208263   chr1      - Tor1aip1     15    7.1 3.1e-11 1.2e-07
     57738  chr16      -  Slc15a2     23    4.8 7.3e-10 2.5e-06
    102103   chr8      -    Mtus1     24    4.5 1.2e-09 3.4e-06
    170788   chr1      -     Crb1     11    9.1 2.1e-09 5.5e-06
     19106  chr17      -  Eif2ak2     16    5.6 8.7e-09 2.1e-05
```

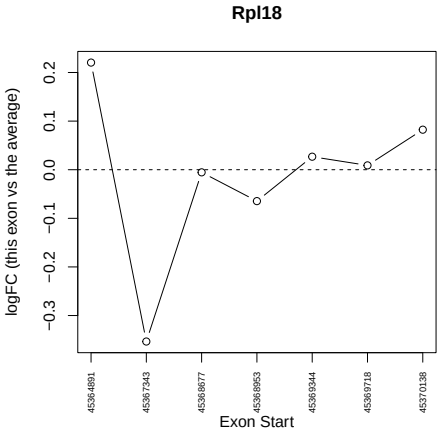
```
# subjunc
> keep <- rowSums(y$counts) > 0
> table(keep)
keep
 FALSE   TRUE
 64659 221849
```

```
# align
> keep <- rowSums(y$counts) > 0
> table(keep)
keep
 FALSE   TRUE
 68986 217522
```

edgeR-v4: Rpl18



subjunc



align

edgeR-v4: Rpl18

```
# Rpl18: subjunc
```

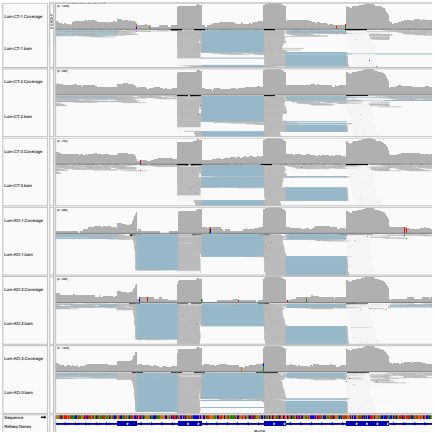
	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
45364891	3	2	5	3	4	8
45367343	47	64	48	36	42	56
45368677	17	22	16	256	159	254
45368953	1057	981	680	862	971	1304
45369344	1051	985	715	907	991	1359
45369718	1372	1261	979	1121	1369	1790
45370138	80	59	46	69	70	103

```
# Rpl18: align
```

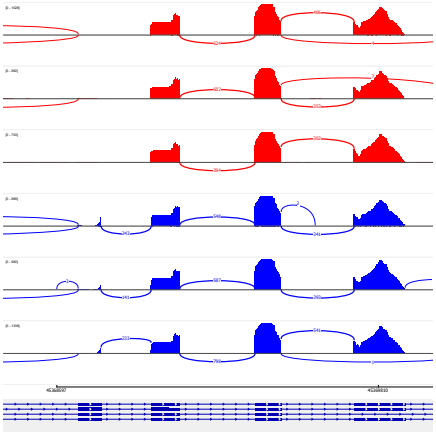
	Lum-CT-1	Lum-CT-2	Lum-CT-3	Lum-KO-1	Lum-KO-2	Lum-KO-3
45364891	3	2	5	3	4	8
45367343	45	63	48	34	41	54
45368677	16	23	15	19	18	26
45368953	434	387	294	328	390	510
45369344	1065	1033	695	924	1038	1401
45369718	986	931	680	795	989	1303
45370138	78	54	45	68	57	102

```
# red: significant change of exon counts between subjunc and align
# it may be due to featureCounts failing to match the correct gene
# as few counts of that exon are observed
```

edgeR-v4: Rpl18



igv coverage plot



igv sushimi plot

diffSplice: exon-exon junction

- Exon counts are correlated because of exon-exon junction
- Exon internal and exon-exon junction are independent
- Exon-exon junctions can be novel and reveal more alternative splicing types
- Exon-exon junction change does not guarantee alternative splicing
- `featureCounts` may fail to locate the exon-exon junction to correct genome
- `featureCounts` may not report strand for exon-exon junctions

diffSplice: exon-exon junction

```
# edgeR-v4: exon-exon junction counts
> topSpliceDGE(dsp, test = "simes", n=15)
```

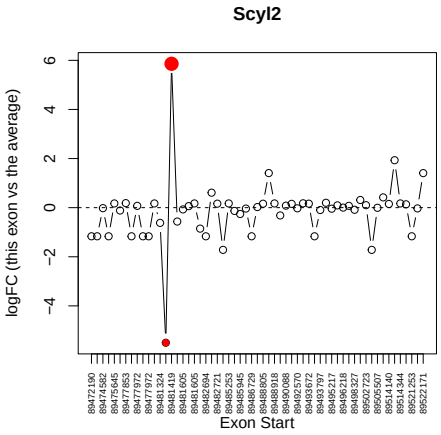
GeneID	Chr	Symbol	NExons	P.Value	FDR
213326	chr10	Scyl2	59	1.5e-17	3.5e-13
19899	chr7	Rpl18	19	9.7e-17	1.2e-12
22388	chr5	Wdr1	52	9.0e-14	7.2e-10
19652	chrX	Rbm3	22	5.0e-13	3.0e-09
51810	chr1	Hnrnpu	53	1.1e-12	5.1e-09
108655	chr6	Foxp1	162	1.3e-12	5.1e-09
208263	chr1	Tor1aip1	41	1.7e-12	5.7e-09
18674	chr10	Slc25a3	41	1.2e-11	3.5e-08
18148	chr11	Npm1	34	1.1e-10	3.0e-07
20005	chr5	Rpl9	14	1.3e-10	3.2e-07
76808	chr8	Rpl18a	14	9.1e-10	2.0e-06
12034	chr6	Phb2	23	1.7e-09	3.3e-06
20733	chr7	Spint2	29	2.9e-09	5.2e-06
55936	chrX	Ctps2	59	5.7e-09	9.7e-06
19231	chr1	Ptma	21	8.5e-09	1.3e-05

```
# edgeR-v4: exon-exon junction counts
> topSpliceDGE(dsp, test = "gene",n=15)
```

GeneID	Chr	Symbol	NExons	gene.F	P.Value	FDR
102436	chr9	Lars2	59	8.9	8.8e-50	2.1e-45
213326	chr10	Scyl2	59	3.6	2.2e-14	2.6e-10
19899	chr7	Rpl18	19	6.6	1.1e-12	8.8e-09
108655	chr6	Foxp1	162	2.1	2.3e-12	1.4e-08
12095	chr3	Bglap3	23	4.9	2.8e-10	1.3e-06
57738	chr16	Slc15a2	62	2.7	2.0e-09	8.1e-06
20005	chr5	Rpl9	14	5.5	3.8e-08	1.3e-04
19652	chrX	Rbm3	22	4.1	5.7e-08	1.7e-04
69707	chr16	Iqcg	24	3.7	1.7e-07	4.4e-04
100302730	chr3	5830417I10Rik	38	2.9	2.0e-07	4.8e-04
13411	chr12	Dnah11	173	1.7	3.2e-07	6.9e-04
20084	chr17	Rps18	12	5.5	4.0e-07	7.9e-04
55936	chrX	Ctps2	59	2.3	1.3e-06	2.4e-03
18044	chr17	Nfya	29	3.0	1.4e-06	2.4e-03
170788	chr1	Crb1	17	4.1	2.2e-06	3.6e-03

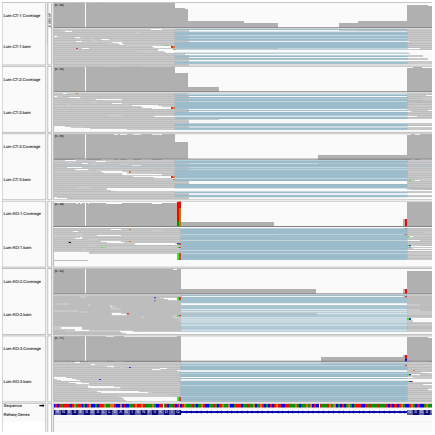
```
# teal: expected results
# red : new result using simes method with exon-exon junction counts, seems new alternative splicing
# blue: new result using simes method with exon-exon junction counts, seems new alternative splicing
# orange: not interesting, might be false discoveries using exon-exon junction counts
```

edgeR-v4: Scyl2

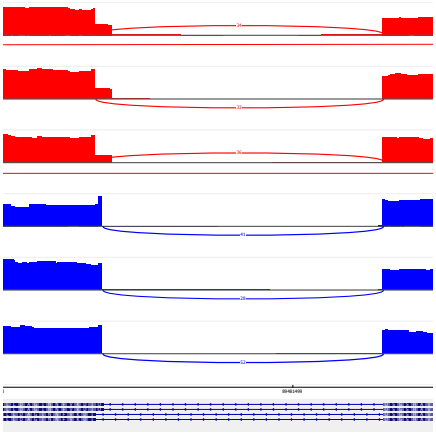



```
# Scyl2
                Start       End LumCT LumCT LumCT LumKO LumKO LumKO
203945 89474582 89476978   684   559   764   545   553   832
203946 89477853 89477972    18     7    11    11    15    16
203947 89481324 89481419    25    17    15     5    10    15
...
203958 89495997 89496218    68    54    42    43    53    81
203959 89498178 89498327    10    12    22    12    17    17
203960 89502579 89502723    11    13    15    13    18    21
203961 89505507 89505664    19    10    20    13    12    23
203962 89514140 89514344    28    28    23    23    27    40
203963 89521900 89522179    46    23    20    33    17    33
...
203968 89476978 89477853    68    52    64    51    39    70
203969 89477972 89479987     0     1     0     0     0     0
203970 89477972 89481324    56    30    42    39    36    60
...
203974 89481416 89481539    34    33    36     0     0     0
203975 89481419 89481539     0     0     0    41    28    52
203976 89481605 89482621    43    44    51    42    26    58
...

# red and blue: two different junctions between same exons
# it shares the same End, slightly different Start
# ...: omitted exon or exon-exon junctions
```

edgeR-v4: Scyl2

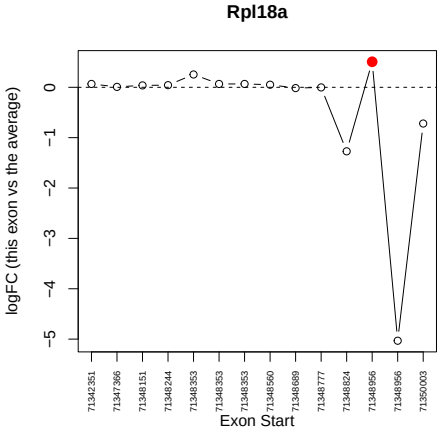


igv coverage plot



igv sushimi plot

edgeR-v4: Rpl18a

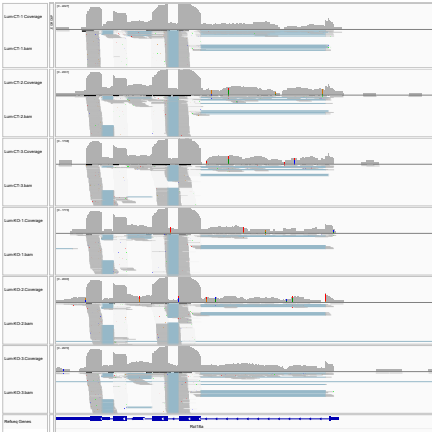



```
# Rpl18a
```

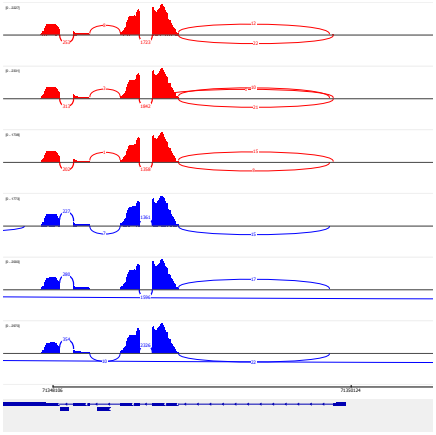
	Start	End	LumCT	LumCT	LumCT	LumKO	LumKO	LumKO
499015	71347366	71348151	576	628	475	506	552	775
499016	71348244	71348353	130	159	112	109	145	203
499017	71348560	71348689	465	446	350	383	440	618
499018	71348777	71348956	1617	1675	1172	1314	1436	2080
499019	71350003	71350087	3	4	3	3	1	1
499020	71342351	71347911	0	0	0	0	0	0
499021	71348151	71348244	253	313	202	227	280	354
499022	71348353	71348560	8	3	1	7	0	10
499023	71348353	71348777	0	0	0	0	0	0
499024	71348353	71349977	0	0	0	0	0	0
499025	71348689	71348777	1723	1842	1358	1361	1596	2326
499026	71348824	71350003	0	1	0	0	0	0
499027	71348956	71349977	12	10	8	15	17	22
499028	71348956	71350003	22	21	15	0	0	0

```
# blue: same junction between last two exons
# red : new junction between last two exons in one group
```

edgeR-v4: Rpl18a

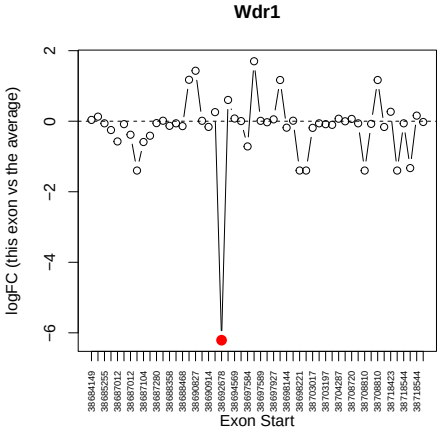


igv coverage plot



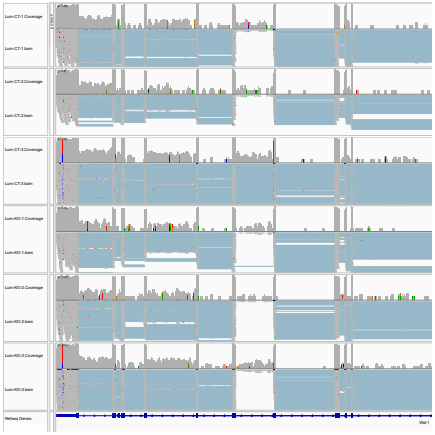
igv sushimi plot

edgeR-v4: Wdr1

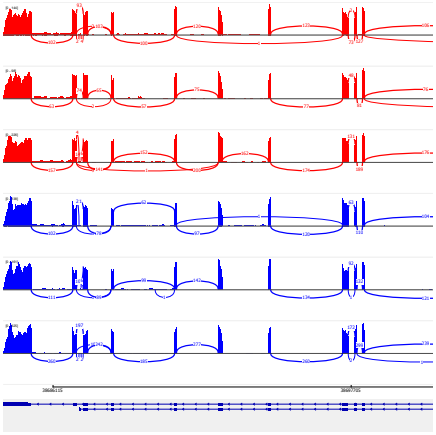



```
# Wdr1
          Start       End LumCT LumCT LumCT LumKO LumKO LumKO
236376 38684149 38685255   889   612  1446   856  1037  2033
236377 38686868 38687012    62    42    93    44    50   101
...
236390 38718423 38718544    20    19    37    28    25    75
236391 38718859 38720265    15    10    18    13    11    29
236392 38685255 38686868   102    63   157   102   111   260
...
236395 38687012 38687280    93    74   145    71   101   197
...
236401 38688468 38690827   100    57   152    62    99   185
236402 38690081 38690827     0     0     0     0     1     0
236403 38690914 38692527   120    75   200    97   142   277
236404 38690914 38697356     1     0     0     1     0     0
236405 38692678 38694477     0     0   162     0     0     0
236406 38694569 38697356   123    77   174   120   134   260
...

# red: the junction observed in only one sample
# ...: omitted exon or exon-exon junction counts
```

edgeR-v4: Wdr1

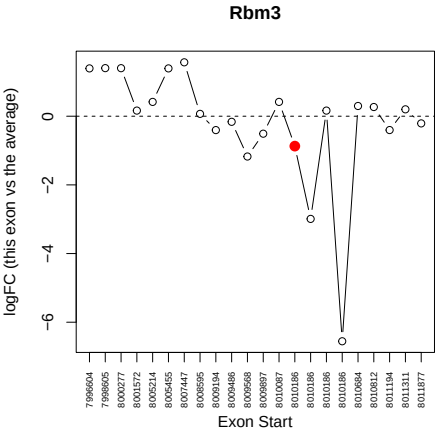


igv coverage plot



igv sushimi plot

edgeR-v4: Rbm3

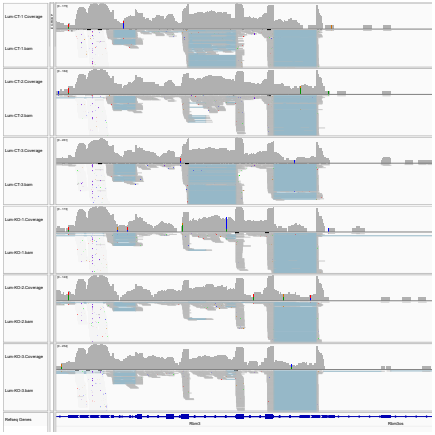



```
# Rbm3
```

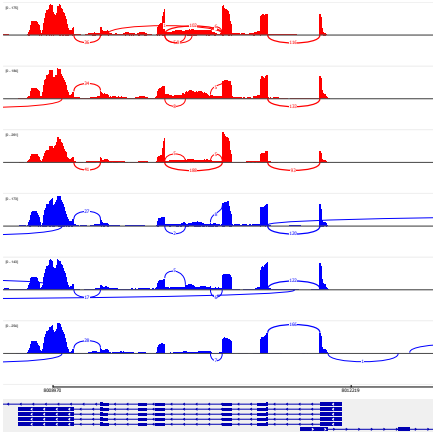
	Start	End	LumCT	LumCT	LumCT	LumKO	LumKO	LumKO
174344	8005214	8007447	55	48	54	91	59	78
174345	8008595	8009194	432	432	433	413	352	607
174346	8009486	8009571	19	23	23	15	12	28
174347	8009897	8009990	22	21	19	6	12	19
174348	8010087	8010186	32	73	19	49	51	86
174349	8010812	8010918	73	118	81	140	72	130
174350	8011194	8011311	24	31	20	14	20	14
...								
174358	8009194	8009486	36	34	41	27	17	28
174359	8009568	8010812	1	0	0	0	0	0
174360	8010186	8010414	5	8	5	2	5	0
174361	8010186	8010484	8	0	0	0	0	0
174362	8010186	8010527	0	0	0	0	0	0
174363	8010186	8010812	102	0	188	0	0	0
174364	8010684	8010812	6	5	5	8	6	7
174365	8011311	8011877	116	133	93	120	122	166

```
# red: the junction observed in only two samples
# ...: omitted exon or exon-exon junction counts
```

edgeR-v4: Rbm3



igv coverage plot



igv sushimi plot

diffSplice: exon vs transcripts

- Exons are well annotated, compared with transcripts
- Exon-exon junctions can be novel
- The reference for transcripts may not be complete
- Transcripts may be novel or un-annotated
- Exon level analysis seems more powerful for alternative splicing

diffSplice: Simulation

- Simulating exon counts directly from NB distribution is impossible
- Simulating exon and exon-exon junction counts is more complicated
- Multinomial distribution is not suitable because of BCV assumption
- Simulating transcript reads is a better approach
- The reference for transcript may not be complete

Summary

- `diffSplice` is able to find the proportion change in count matrix
- `diffSplice` is a powerful tool for alternative splicing analysis
- The exon or exon-exon junction count matrix is complicated
- It is difficult to prepare a biologically meaningful count matrix
- It requires careful align and count on the reads both on exon or transcript
- Novel exon-exon junction or transcript can be essential

Acknowledgement

Smyth Lab

Gordon Smyth

Hannah Coughlan

Pedro Baldoni

Waruni Abeysekera

Mengbo Li

Jinming Cheng

Chen Lab

Andy Chen

Davidson Lab

Alex Yan

Visvader Lab

Thank you