基于二代转录组数据的可变剪接和可变多聚腺苷酸化分析

Analysis of Alternative Splicing and Alternative Polyadenylation Patterns Based on Transcriptome Data by Next Generation Sequencing

黄雪娜¹, 战爱斌^{1,2,*}

¹中国科学院生态环境研究中心, 北京; ²中国科学院大学, 北京

*通讯作者邮箱: azhan@rcees.ac.cn

引用格式: 黄雪娜, 战爱斌. (2021). 基于二代转录组数据的可变剪接和可变多聚腺苷酸化分析. *Bio-101* e1010658. Doi: 10.21769/BioProtoc.1010658.

How to cite: Huang, X.N. and Zhan, A.B.. (2021). **Analysis of alternative splicing and alternative polyadenylation patterns based on transcriptome data by next generation sequencing**. *Bio-101* e1010658. Doi: 10.21769/BioProtoc.1010658. (in Chinese)

摘要: 可变剪接 (Alternative splicing, AS) 和可变多聚腺苷酸化 (Alternative polyadenylation, APA) 是指同一 mRNA 前体通过不同的剪接方式产生不同转录本异构体的过程。通过 AS 或 APA 调控机制, 一个基因可在不同时空状态下产生多个结构、功能、细胞定位或稳定性不同的 mRNA 和蛋白质, 进而产生不同的生物学功能和性状, 最终驱动环境适应、物种分化和生物多样性等多种生物学过程。近年来高通量测序和生物信息学技术的快速发展为深入研究可变剪接调控机制提供了丰富的转录组数据资源和逐渐完善的分析方法。基于二代测序技术得到的 RNA-Seq 数据, rMATS 软件是目前最常用的分析可变剪接事件的工具, APAtrap 软件是常用来分析可变多聚腺苷酸化的工具。这两种方法均可通过对剪接位点的鉴定和定量, 得到两组样品比较的差异可变剪接/腺苷酸化位点和基因, 进而通过基因功能注释得到可变剪接调控机制在不同生物过程中的作用。本文将介绍利用 rMATS 和 APAtrap 软件分别进行差异可变剪接和可变多聚腺苷酸化研究的数据分析流程, 以期今后相关软件的使用提供方法借鉴。

关键词: 可变剪接, 可变多聚腺苷酸化, 转录组, rMATS, APAtrap

研究背景

可变剪接是驱动蛋白质多样性的重要调控机制，在生长发育、环境适应、物种分化等多种生物学过程中发挥重要作用，成为目前基因调控机制研究中的热点 (Bush *et al.*, 2017; Schaefke *et al.*, 2018)。同时，可变多聚腺苷酸化作为一种重要的转录后调控机制，通过选择前体 mRNA 3'末端不同的多聚腺苷酸化位点 (PAS)，产生具有不同 3'UTR 长度的转录本，进而影响转录本的稳定性、表达量和翻译效率 (Pereira-Castro and Moreira, 2021)，在疾病发生和各种胁迫响应过程中发挥重要作用 (Sadek *et al.*, 2019; Zheng *et al.*, 2018)，逐渐成为分子生态研究领域的新兴关注热点。全基因组水平可变剪接事件/基因的鉴定和定量是研究可变剪接调控机制的前提，高通量测序技术为可变剪接相关研究提供了良好的技术平台并积累了大量的数据基础。传统的二代测序以通量高、读长短为特点，难以获得完整的转录本全长序列，而转录本重新拼接易造成转录本结构不完整或拼接错误，因此在可变剪接研究中存在一定的局限性。最新的三代测序技术可获得长达 25-30 kb 的超长读长序列，能够直接获得完整的转录本，进而更准确地开展转录本多样性研究。例如，利用 MinION 纳米孔测序仪对非小细胞肺癌组织进行全长转录本测序，成功鉴定到癌细胞中的异常剪接转录本，对评估中肿瘤免疫反应具有重要作用，然而这些异常转录本在之前利用短读长高通量测序的相关研究中则处于被忽视状态 (Oka *et al.*, 2021)。然而，由于三代测序相较于二代测序通量低且成本高的特点，三代测序技术尚未全面开展应用，尤其是在生态进化领域中以群体为研究对象的大样本研究中。因此，基于大量积累的二代转录组测序数据，选择合适的分析工具进行可变剪接研究是目前较为常见且可行性较高的研究手段，例如，基于二代测序转录组分析结果表明可变剪接是北极红点鲑 (*Salvelinus alpinus*) 快速适应不同栖息环境的重要分子调控机制 (Jacobs and Elmer, 2021)，且可变剪接是豌豆蚜虫在不同季节呈现表型多样性的重要分子基础之一 (Grantham and Brisson, 2018)。同样基于二代转录组测序分析结果表明，砷胁迫能够诱导小鼠 NIH3T3 细胞在全基因组范围内更倾向于选择近端 PAS 位点造成大量基因 3'UTR 长度变短，通过逃避 RNA 降解以保持表达量稳定 (Zheng *et al.*, 2018)。

目前利用 RNA-Seq 测序数据进行可变剪接分析的软件有很多，总体来讲分为三大类：基于转录本异构体 (transcript isoform) 的软件如 Cufflinks (Trapnell *et al.*, 2013)，基于外显子为统计单元如 DEXSeq (Anders *et al.*, 2012)，和利用跨越剪接位点的连接区域 (junction region) 为统计单元的方法如 rMATS (Shen *et al.*, 2014) 等。由于二代

测序短读长在组装上的劣势造成转录本组装的不准确性，本文选取基于剪接位点的 rMATS 软件和 APAtrap 软件分别就其在可变剪接和可变多聚腺苷酸化研究中的常规分析流程进行阐述。

软件版本信息及下载地址

文中所有软件的运行均在 Linux 操作系统 Ubuntu 14 环境上进行，数据处理和分析则在个人普通电脑上进行。

1. rMATS v4.1.1 (<http://rnaseq-mats.sourceforge.net/rmats4.0.2/>; Shen *et al.*, 2014)
2. HISAT2 2.2.1 (<https://daehwankimlab.github.io/hisat2/>; Kim *et al.*, 2015)
3. Samtools v.1.13 (<https://sourceforge.net/projects/samtools/files/samtools/>)
4. APAtrap (<https://sourceforge.net/projects/apatrap/>; Ye *et al.*, 2018)
5. rmats2sashimiplot (<https://github.com/Xinglab/rmats2sashimiplot>)
6. R v.4.1.0 (<https://cran.r-project.org/src/base/R-4/>)

实验步骤

一、数据准备

由于本可变剪接数据分析流程重点关注两组样品比较得到的差异可变剪接情况，因此需提前准备两组具有生物学比较意义的、含有生物学重复的 RNA-Seq 数据。本教程中测试数据集包含两组双端测序数据 fastq 格式文件：处理组 T 和对照组 C，每组数据包含 3 个生物学重复，分别命名为 T1、T2、T3 和 C1、C2、C3。

注：原始转录组数据应首先进行质量控制和数据清洗，去处低质量碱基和接头序列，具体流程参考请参考阅读 Trimmomatic 软件的官方说明进行 (<http://www.usadellab.org/cms/index.php?page=trimmomatic>)，本文不做赘述。

二、rMATS 软件进行差异可变剪接分析

1. 软件安装

- 1.1 安装 Python 2.7.x 和 Python 软件包 NumPy; 并安装软件 BLAS, LAPACK, GSL 和 gfortran:

```
pip install numpy
sudo apt-get install libblas-dev liblapack-dev
sudo apt-get install libgsl0ldbl
sudo apt-get install gfortran
```

- 1.2 由于 rMATS v4.0.2 包含对应两种不同 Python 编译器的 rMATS-turbo-Linux-UCS4 和 rMATS-turbo-Linux-UCS2 两个版本，因此用户首先需要确认自己要用的版本。打开 python 控制台并输入：

```
>>> import sys
>>> print sys.maxunicode
```

输出内容为 1114111 或 65535 分别代表应选择 rMATS-turbo-Linux-UCS4 或 rMATS-turbo-Linux-UCS2 版本。

1.3 下载安装 rMATS

从官网 (http://rnaseq-mats.sourceforge.net/rmats4.0.2/user_guide.htm) 下载软件，解压缩，根据步骤 1.2 选择相应版本文件夹进入，然后将所需用的数据都转移进该文件夹：

```
tar -xzf rMATS.4.0.2.tgz
cd rMATS.4.0.2/
```

2. 输入文件格式准备

rMATS 软件支持两种格式文件的输入，一种是 fastq 序列格式，需要同时安装序列比对软件如 STAR 并提供比对的索引文件，另一种是序列比对结果 bam 格式，rMATS 软件支持多种比对软件如 Tophat 或 HISAT2 软件的输出结果。由于使用 fastq 格式文件时候比对索引文件较大且软件运行较慢，因此推荐使用 HISAT2 软件输出的 bam 格式文件。具体 HISAT2 比对流程及参数设置请参考 HISAT2 软件的官方说明进行 (<https://daehwankimlab.github.io/hisat2/>)，本文不做赘述。另外还需准备用于序列比对的基因组注释文件 annotation.gtf。

HISAT2 软件输出文件为 sam 格式，利用 samtools 软件将 sam 格式转换为 bam 格式：

```
samtools sort -@ 8 -o T1.bam T1.sam
```

注: **-@** 设置线程数量, 默认为 1; **-o** 设置最终排序后的输出文件名; 最后是输入 **sam** 格式文件。

然后将需要比较的两组数据信息整理成 **txt** 文本格式, 文件内以逗号分隔重复样本的 **bam** 文件名。例如处理组和对照组三个生物学重复样品的 **bam** 格式比对文件分别整理为 **b1.txt** 和 **b2.txt** 文件:

b1.txt 内容: T1.bam, T2.bam, T3.bam

b2.txt 内容: C1.bam, C2.bam, C3.bam

3. rMATS 软件运行

在 **rMATS-turbo-Linux-UCS4** 文件路径下输入:

```
python rMATS-turbo-Linux-UCS4/rmats.py --b1 b1.txt --b2 b2.txt --gtf
annotation.gtf --od AS -t paired --readLength 150 --nthread 10
```

rMATS 软件运行必须设置的参数包括以下几项:

--b1 b1.txt: 文本文件包含处理组各样品 **bam** 格式比对文件信息

--b2 b2.txt: 文本文件包含对照组各样品 **bam** 格式比对文件信息

--gtf: 基因组的注释文件

--od: 输出文件夹

-t: 序列类型 **readType**, 双端测序则 **readType** 为 **paired**, 单端测序则为 **single**

--readLength: 测序 **reads** 的长度

--nthread: 根据 CPU 处理器设置线程数

4. 结果解读与数据整理

运行 **rMATS** 后所有的输出文件都在 **--od** 参数设置的文件夹中, 主要包含以下五种类型的结果文件, 下游分析通常仅需考虑第一种文件类型即可:

AS_Event.MATS.JC.txt: 仅考虑跨越剪接位点 **reads** (Junction count) 的最终可变剪接结果;

AS_Event.MATS.JCEC.txt: 不仅考虑跨越剪接位点的 **reads**, 同时考虑没有跨域剪接位点的 **reads** (exon counts) 的最终可变剪接结果(图 1 中条纹区域);

FromGTF.AS_Event.txt: 从 GTF 注释文件中衍生出的可变剪接事件;

跨域剪接位点的 reads 的剪接事件数;

4.1 rMATS 识别五种类型的可变剪接事件 (如图 1 所示), 包括外显子跳跃 (skipped exon, SE)、内含子保留 (retained intron, RI)、可变 5'剪接位点 (alternative 5' splice site, A5SS)、可变 3'剪接位点 (alternative 3' splice site, A3SS)、外显子选择性跳跃 (mutually exclusive exons, MXE)。

图 1. 可变剪接类型 (图片来源于 Park *et al.* (2013)中图 1)

4.2 现以 SE 可变剪接类型为例阐述结果文件 SE.MATS.JC.txt, 结果共包含 23 列信息:

第 1-5 列按照表头即可理解;

第 6-11 列记录了发生跳跃的外显子 (skipped exon) 及其上下游外显子的位置信息;

第 13 列表示在样品 1 中外显子被保留的 read 数目, 第 14 列在样品 1 中外显子被剪切的 read 数目, 第 15 和 16 列分别为样品 2 中的相关信息;

第 17 和 18 列中 IncFormLen、SkipFormLen 分别表示外显子是否被剪切产生转录本异构体的有效长度;

第 19 列 Pvalue 由 likelihood-ratio test 计算得到表示两组样品在外显子包含水平 (Inclusion level) 上的差异;

第 20 列 FDR 为利用 Benjamini Hochberg 算法对 P-value 进行校正得到;

第 21 和 22 列 IncLevel1 和 IncLevel2 分别表示样品 1 和样品 2 中 exon inclusion level (PSI);

第 23 列 IncLevelDifference 代表两组样品 exon inclusion level 的差异。

4.3 数据整理分析

根据 IncLevelDifference 和 FDR 两个参数筛选处理组和对照组比较中显著变化的可变剪接事件 (differential alternative splicing events, DASE), 参数设置为 $|\text{IncLevelDifference}| > 0.1$, 同时 $\text{FDR} < 0.05$ 。根据 DASE 发生的基因得到差异可变剪接基因 (differential alternative splicing genes, DASG), 然后利用 GO 和 KEGG 功能富集得到 DASG 涉及的生物学功能。

IncLevelDifference 的分布: 当 IncLevelDifference < -0.1 时, 代表处理组外显子包含水平低于对照组, 暗示着处理后该外显子在成熟转录本中倾向于被剪切掉; 当 IncLevelDifference > 0.1 时, 代表处理组外显子包含水平高于对照组, 暗示着处理后该外显子倾向于出现在成熟转录本中。例如, Huang *et al.* (2021) 研究发现温度胁迫后海鞘基因组范围内大量外显子倾向于被剪切掉 (图 2)。

图 2. 海鞘温度胁迫后外显子倾向于被剪切掉

为进一步研究外显子跳跃类型可变剪接对后续蛋白编码的潜在影响, 可利用结果中外显子的长度信息进行预测。若可变外显子长度是 3 的倍数, 那么该可变剪接事件不会造成读码框移码 (ORF shift), 仅造成蛋白序列缺失或增加; 若可变外显子长度不是 3 的倍数, 则会造成读码框移码, 改变蛋白序列或提早

终止。

5. 关键可变剪接事件的可视化

为了缩短可视化软件运行时间并提高分析针对性，用户可根据 **rMATS** 输出结果中 **IncLevelDifference** 参数大小、或关键生物学功能过程中的关键基因自行挑选重要的可变剪接事件。将候选可变剪接事件从 **rMATS** 输出结果文件中挑选出来，重新形成新的 **SE.MATS.JC.txt** 文件。

5.1 软件下载与安装

从官方网站 (<https://github.com/Xinglab/rmats2sashimiplot>) 下载软件 **rmats2sashimiplot**，该软件可免安装或安装使用：

免安装使用方法：

```
python ./src/rmats2sashimiplot/rmats2sashimiplot.py
```

安装方法：

```
python ./setup.py install
```

5.2 数据准备

提供用于 **rMATS** 分析用的 **bam** 格式文件，以及 **rMATS** 输出结果中挑选出的关键可变剪接事件形成的新 **SE.MATS.JC.txt** 文件。

5.3 软件运行及结果解析

首先需要先对 **bam** 文件进行默认情况下的排序，并进行 **index**：

```
samtools sort T1.bam -o T1_sorted.bam
samtools index A1_2_sorted.bam
```

然后输入以下代码运行 **rmats2sashimiplot**：

```
rmats2sashimiplot --b1 T1.bam, T2.bam, T3.bam --b2 C1.bam, C2.bam,
C3.bam -t SE -e SE.MATS.JC.txt --l1 T --l2 C --exon_s 1 --intron_s 1 -o
SE_plot
```

注：其中 **--b1** 和 **--b2** 参数指定样本对应的 **bam** 文件，**--l1** 和 **--l2** 参数指定出现在最终图片中的样本名字，**-t** 参数指定可变剪切的类型包括 **SE**、**A5SS**、**A3SS**、**MXE** 或 **RI**，**-e** 参数指定 **rmats** 产生的可变剪切结果文件，**exon_s** 和 **intron_s**

参数指定图片中 **exon** 和 **intron** 长度缩小的比例, **-o** 参数指定输出文件夹名称。

通过以上代码, 每个可变剪接事件在每个生物学重复样品中的剪接情况单独输出一张图片, 如图 3A 即表示该外显子 (基因组位置为 **regtig_537:37099:37171**) 在处理组 T 组和对照组 C 组三个生物学重复样品中的 **inclusion level**。

如果需要分组, 将某个外显子在三个生物学重复之间的平均 **inclusion level** 进行重新展示, 需添加 **—group-info** 参数指定组别名称以及该组中包含的样品个数, 输入代码为:

```
rmats2sashimiplot --b1 T1.bam, T2.bam, T3.bam --b2 C1.bam, C2.bam, C3.bam -t SE -e SE.MATS.JC.txt --l1 T --l2 C --exon_s 1 --intron_s 1 -o SE_plot --group-info grouping.gf
```

注: 其中 **grouping.gf** 文件即为分组信息文件, 包含以下信息: 组名和该组包含的样品信息。例如: 文件中第一行内容 “**T1: 1-3**” 则代表第一组名称为 **T1**, 包含的样品为前三个 (**T1.bam, T2.bam, T3.bam**); 第二行内容 “**C1: 4-6**” 则代表第二组名称为 **C1**, 包含的样品为后三个 (**C1.bam, C2.bam, C3.bam**), 输出图片如图 3B 所示:

图 3. 外显子跳跃可变剪接事件在两组样品中的可视化。(A)在每个样品中单独展示；(B) 分组展示。

三、APAtap 软件进行差异可变多聚腺苷酸化分析

近年来研究发现基因在转录过程中存在 poly(A) 位点的选择，造成具有不同长度 3'UTR 序列的转录异构体。虽然 APA 调控一般情况下不会改变编码区的序列，然而 3'UTR 常作为 miRNA 调控的靶区域，同时也会被特定的 RNA 结合蛋白调控，影响 mRNA 的稳定性进而影响蛋白质翻译水平。因此研究 3'UTR 长度调控具有重要的生物学意义，而首要任务是从转录组数据中对加尾信号进行准确的鉴定。APAtap 是一款鉴定 APA 位点的软件，同时能够检测两组数据中差异 APA 事件。

1. 软件安装

从官方网站下载软件，解压缩，文件夹 **APAtap** 中包含两个由 Perl 编译的可执行文件 "identifyDistal3UTR" 和 "predctAPA"，还包含一个名为 "deAPA_1.0.tar.gz" 的 R 包。打开 R 程序并进入 deAPA_1.0.tar.gz 文件所在路径，输入以下命令行安装软件：

```
>install.packages("deAPA_1.0.tar.gz", repos = NULL, type = "source")
```

2. 数据准备

2.1 首先利用 bedtools (<https://bedtools.readthedocs.io/en/latest/index.html>) 工具 genomeCoverageBed 将所有序列比对结果 bam 文件转化为 bedgraph 格式，输入以下命令：

```
genomeCoverageBed -bg -ibam T1.bam -split > T1.bedgraph
```

2.2 其次需要提供包含 12 列信息的 bed 格式基因模型文件，大部分动物物种的 bed 格式文件可由 UCSC (<http://genome.ucsc.edu/cgi-bin/hgTables?command=start>) 工具获得，其他物种则通常可由基因组注释文件 annotation.gtf 文件转化而来，按照以下步骤进行转化：

首先下载 UCSC 工具：

```
$ wget -c http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86\_64/gtfToGenePred
$ wget -c http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86\_64/genePredToBed
```

修改 Linux 文件权限：

```
$ chmod 755 gtfToGenePred genePredToBed
```

最后将 gtf 文件转化为 bed12 格式文件：

```
$ gtfToGenePred annotation.gtf test.genePhred
$ genePredToBed test.genePhred annotation.bed
$ rm test.genePhred
```

3. 软件使用

软件使用共包括三个步骤：

3.1 运行 identifyDistal3UTR 提取已注释的 3'UTR 并鉴定新的 3'UTR

针对含有短 3'UTR 的基因组，使用以下代码：

```
identifyDistal3UTR -i T1.bedgraph T2.bedgraph -m annotation.bed -o
novel.utr.bed -w 50 -e 5000
```

针对含有长 3'UTR 的基因组，使用以下代码：

```
identifyDistal3UTR -i T1.bedgraph T2.bedgraph -m annotation.bed -o
novel.utr.bed
```

其中-i参数输入 bedgraph 或 wig 格式的比对文件，可输入单个或多个文件；
-m 参数输入 bed12 格式的基因模型文件；-o 参数为输出文件名称。

3.2 运行 predictAPA 推断所有潜在 APA 位点并计算相应使用率：

针对含有短 3'UTR 的基因组，使用以下代码：

```
predictAPA -i T1.bedgraph T2.bedgraph -g 2 -n 1 1 -u novel.utr.bed -o
output.txt -a 50
```

针对含有长 3'UTR 的基因组，使用以下代码：

```
predictAPA -i T1.bedgraph T2.bedgraph -g 2 -n 1 1 -u novel.utr.bed -o
output.txt
```

其中-i参数输入 bedgraph 或 wig 格式的比对文件，可输入单个或多个文件；
-g 参数输入样品的分组数目；-n 参数表示每组包含的样品个数；-a 参数表示预测 APA 位点之间最小的距离。

3.3 运行 deAPA 鉴定两组样品中发生差异 APA 的基因：

在 R 中利用 deAPA 进行差异 APA 分析，deAPA 包的使用：

```
deAPA(input_file, output_file, group1, group2, least_qualified_num_in_group1, least_qualified_num_in_group2, coverage_cutoff)
```

其中 `input_file` 即为步骤 (2) 中的输出文件；`output_file` 即为 deAPA 分析的输出文件名称；`group1` 和 `group2` 即为要进行比较的组别名称，默认值分别为 1 和 2；`least_qualified_num_in_group1` 和 `least_qualified_num_in_group2` 即为组 1 和组 2 中最小的生物学重复数量，默认值为 1；`coverage_cutoff` 即为每个样品所需的最低覆盖深度。

4. 结果解读

最终 deAPA 输出结果文件中最为重要的三个参数分别为 `perc_diff`、`r` 和 `p.adjust` (Ye *et al.*, 2018):

`perc_diff` (percentage difference, PD), 用来对两组样品间 APA 位点使用率的差异百分比进行定量, 范围为 0-1, 数值越高代表差异越大; `r` (Pearson product moment correlation coefficient) 皮尔森积差相关系数, 范围为-1 到 1, 正值代表组 2 相比于组 1 而言, 更倾向于使用远端 poly(A)位点或具有更长的 3'UTR, 负值则代表组 2 更倾向于使用近端 poly(A)位点或具有更短的 3'UTR; `p.adjust` 是使用 Benjamini-Hochberg 方法对 P 值进行校正得到的结果。首先根据 $PD \geq 0.2$ 且 $p.adjust < 0.05$ 将显著差异 APA 事件筛选出来, 然后再根据 `r` 值的正负将两组比较中使用远端、近端的 poly(A)位点的 APA 事件进行区分。例如, Ye *et al.* (2019)研究表明: 水稻在干旱胁迫 (Drought) 下, 发生差异 APA 的基因更倾向于使用远端 poly(A)位点, 造成更长的 3'UTR (Lengthening); 然而温度胁迫 (Heat) 下不同品系的水稻表现出不同的趋势; 不同浓度重金属 Cd 胁迫 (cadmium) 也分别造成不同的 APA 使用, 低浓度 Cd 胁迫下差异 APA 基因更倾向于使用近端 poly(A)位点 (Shortening), 而高浓度 Cd 胁迫下差异 APA 基因更倾向于使用远端 poly(A)位点 (Lengthening) (图 4)。对于生物胁迫而言, 白叶枯病原(Bacterial Blight)和稻瘟病原(Rice Blast)感染后使用远端 poly(A)位点的基因数目大于使用近端 poly(A)位点的基因数目, 即更倾向于造成更长的 3'UTR (Lengthening), 而病毒感染后 (Rice Stripe Virus) 则呈现相反趋势(图 4)。

图 4. 水稻不同胁迫下的差异 APA 基因 (图片来源于 Ye *et al.*, 2019)

致谢

我们的研究受到国家自然科学基金委员会面上项目 (31772449)、国际合作项目 (320611430121) 的资助支持。

竞争性利益声明

作者声明没有利益冲突。

参考文献

- Anders, S., Reyes, A. and Huber, W. (2012). [Detecting differential usage of exons from RNAseq data](#). *Genome Res* 22(10): 2008–2017.
- Bush, S. J., Chen, L., Tovar-Corona, J. M. and Urrutia, A. O. (2017). [Alternative splicing and the evolution of phenotypic novelty](#). *Philos Trans R Soc Lond B Biol Sci* 372(1713): 20150474.
- Grantham, M. E. and Brisson, J. A. (2018). [Extensive differential splicing underlies phenotypically plastic aphid morphs](#). *Mol Biol Evol* 35(8): 1934–1946.
- Huang, X. and Zhan, A. (2021). [Highly dynamic transcriptional reprogramming and shorter isoform shifts under acute stresses during biological invasions](#). *RNA Biol*

- 18(3): 340-353.
5. Jacobs, A. and Elmer, K. R. (2021). [Alternative splicing and gene expression play contrasting roles in the parallel phenotypic evolution of a salmonid fish](#). *Mol Ecol* 00: 1-15.
6. Kim, D., Langmead, B. and Salzberg, S. L. (2015). [HISAT: a fast spliced aligner with low memory requirements](#). *Nat Methods* 12: 357-360.
7. Oka, M., Liu, X., Suzuki, T., Yoshikawa, T. and Seki, M. (2021). [Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer](#). *Genome Biology* 22(1): 9.
8. Park, J. W., Tokheim, C., Shen, S. and Xing, Y. (2013). [Identifying differential alternative splicing events from RNA sequencing data using RNASEq-MATS](#). In: Shomron N. (Ed.). In: *Methods in molecular biology: deep sequencing data analysis*. Springer, 171-179.
9. Pereira-Castro, I. and Moreira, A. (2021). [On the function and relevance of alternative 3'-UTRs in gene expression regulation](#). *Wiley Interdiscip Rev RNA*, e1653.
10. Schaefer, B., Sun, W., Li, Y. S., Fang, L. and Chen, W. (2018). [The evolution of posttranscriptional regulation](#). *Wiley Interdiscip Rev RNA* e1485.
11. Sadek, J., Omer, A., Hall, D., Ashour, K. and Gallouzi, I. E. (2019). [Alternative polyadenylation and the stress response](#). *Wiley Interdiscip Rev RNA* 10(5): e1540.
12. Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q. and Xing, Y. (2014). [rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data](#). *Proc Natl Acad Sci U S A* 111: E5593-5601.
13. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013). [Differential analysis of gene regulation at transcript resolution with RNA-seq](#). *Nat Biotechnol* 31(1), 46-53.
14. Ye, C., Long, Y., Ji, G., Li, Q. Q. and Wu, X. (2018). [APAtap: identification and quantification of alternative polyadenylation sites from RNA-seq data](#). *Bioinformatics* 34: 1841-1849.
15. Ye, C., Zhou, Q., Wu, X., Ji, G. and Li, Q. Q. (2019). [Genome-wide alternative polyadenylation dynamics in response to biotic and abiotic stresses in rice](#). *Ecotoxicol Environ Saf* 183: 109485.
16. Zheng, D., Wang, R., Ding, Q., Wang, T., Xie, B., Wei, L., et al. (2018). [Cellular](#)

[stress alters 3'UTR landscape through alternative polyadenylation and isoform-specific degradation](#). *Nat Commun* 9(1), 2268.