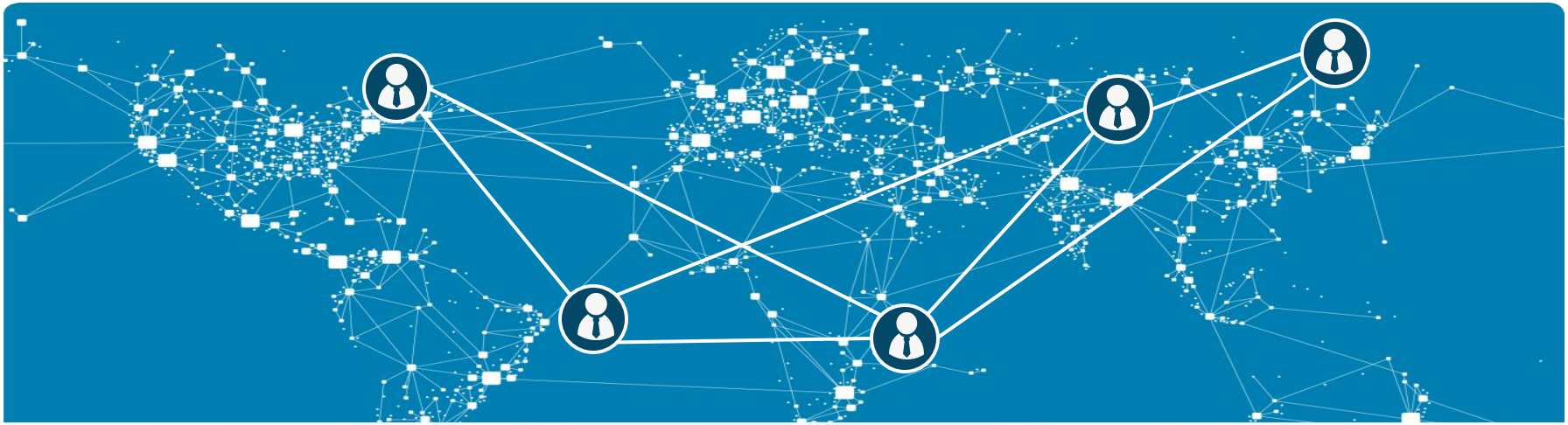
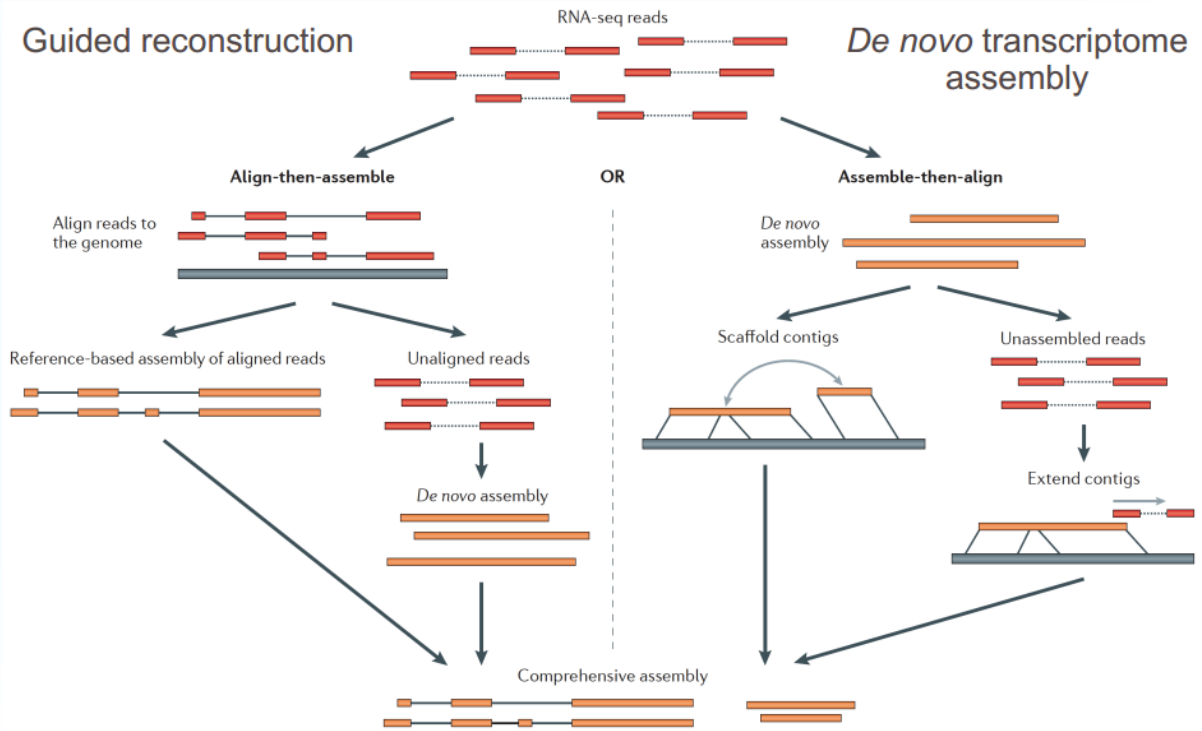


转录本拼装

为什么转录本可以拼接起来？

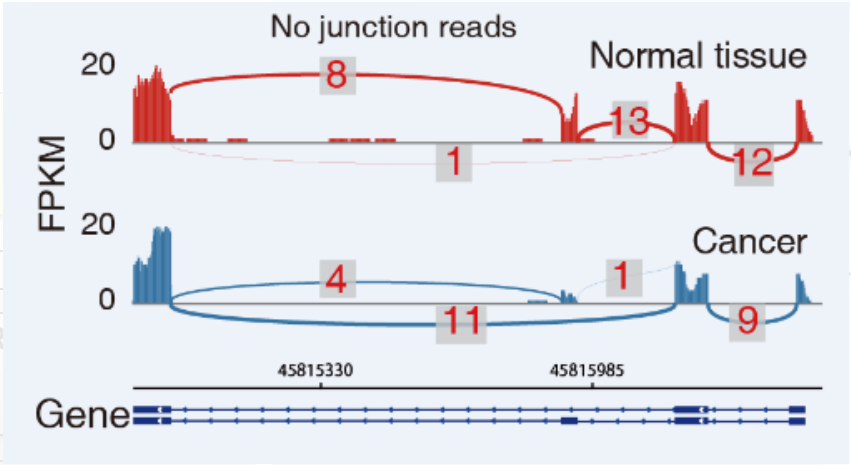
易生信，毕生缘；培训版权所有。

有参组装和无参组装



易生信，毕生缘；培训版权所有。

有参转录组分析流程–可变剪接



可变剪接分析

Gene identification

StringTie

rMATS

Alternative splicing

Differential expression

DESeq2

ImageGP

Visualization

Enrichment analysis

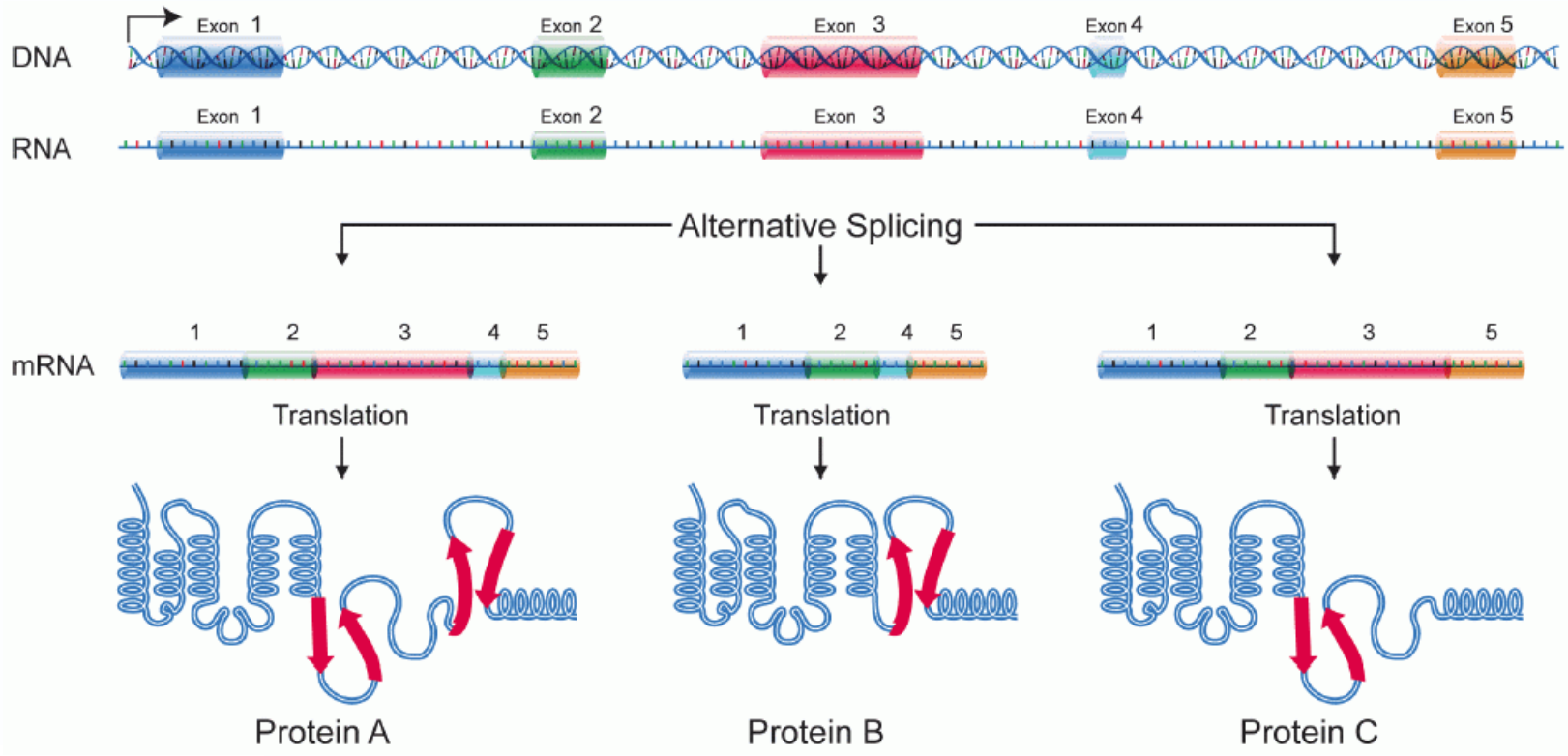
GO/GSEA

WGCNA Cytoscape

Network analysis

易生信，毕生缘；培训版权所有。

选择性剪接丰富了蛋白产物的多样性



选择性剪接WIKI

此件仅供毕生缘；培训版权所有。

卵子特有的Dicer剪切异构体控制RNAi通路

ARTICLE | VOLUME 155, ISSUE 4, P807-816, NOVEMBER 07, 2013

A Retrotransposon-Driven Dicer Isoform Directs Endogenous Small Interfering RNA Production in Mouse Oocytes

Matyas Flemr • Radek Malik • Vedran Franke • ... Radislav Sedlacek • Kristian Vlahovicek • Petr Svoboda • Show all authors

Open Archive • DOI: https://doi.org/10.1016/j.cell.2013.10.001 • Check for updates

Highlights

- Retrotransposition gave rise to an oocyte-specific Dicer isoform (Dicer^{O}) in mice
- Dicer^{O} is N-terminally truncated and has higher activity than somatic Dicer
- Dicer^{O} controls the endogenous RNAi pathway and is essential for mouse oocytes
- Low Dicer activity and low dsRNA abundance constrain endogenous RNAi in mammals

Cell

易生信，毕生缘；培训版权所有。

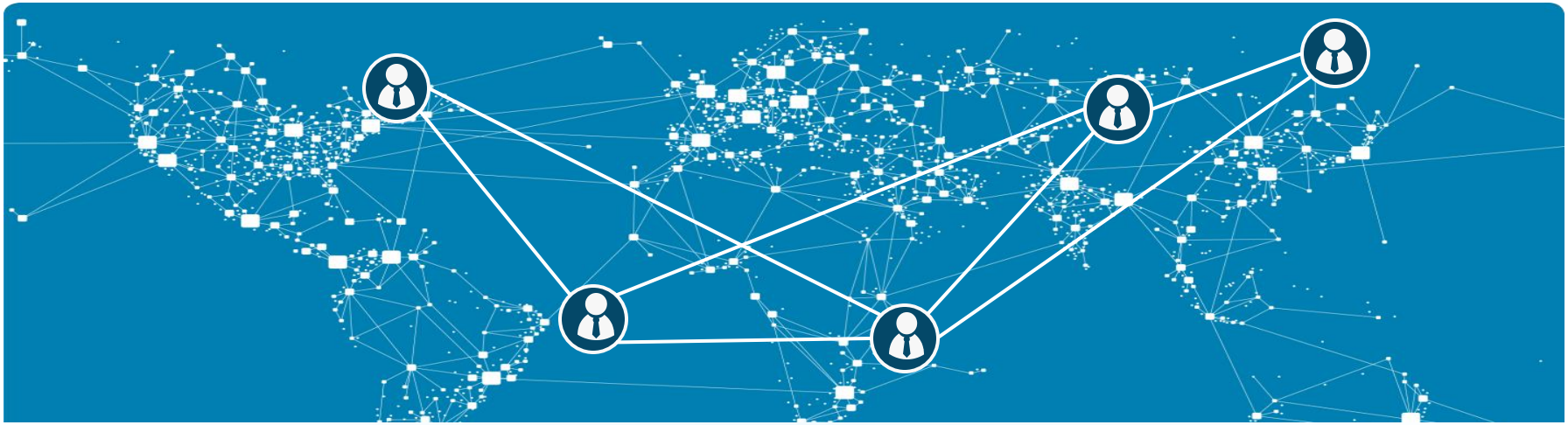
二代测序难获得准确的转录本

- 二代测序难拼出准确转录本，更适合检验单个剪接位点的变化
- 三代测序理论上可以获取全长转录本，能发现更多剪接类型

Transcript classification codes

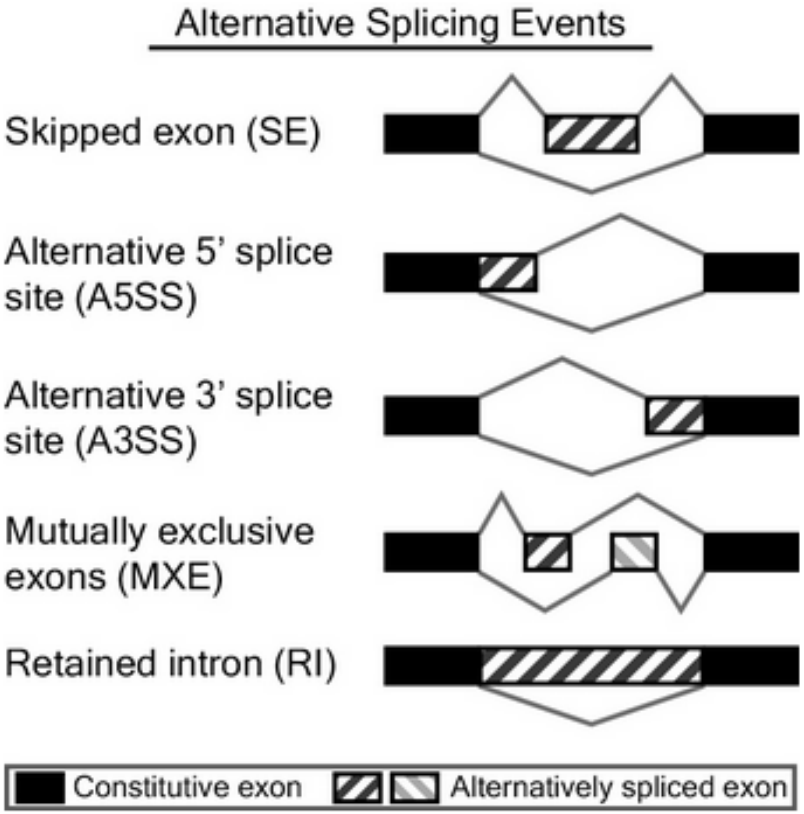
完全匹配	=	complete match of intron chain 
部分包含	c	contained in reference (and intron isoform compatible) 
反向包含参考序列	k	containment of reference (reverse containment) 
潜在新转录本	j	at least one junction match 
潜在pre-mRNA片段	e	single exon, overlapping intron, possibly pre-mRNA fragment (unspliced intron) 
邻近相同链重叠外显子	o	other same strand overlap with reference exons 
相反链匹配内含子	s	intron match on the opposite strand (likely a mapping error) 
相反链重叠外显子	x	exonic overlap on the opposite strand (like 'o' or 'e' but on the opposite strand) 
位于内含子区域	i	fully contained in a reference intron 
内含子包含参考序列	y	contains a reference within its intron(s) 
聚合酶通读产物	p	possible polymerase run-on (no actual overlap) 
重复序列	r	repeat (at least 50% bases soft-masked)
未注释基因间区	u	none of the above (unknown, intergenic)

易汉博基因科技(北京)有限
EMBio Gene Technology (Beijing) co



可变剪接分析

可变剪接的类型



外显子跳跃

可变的5' 剪接位点

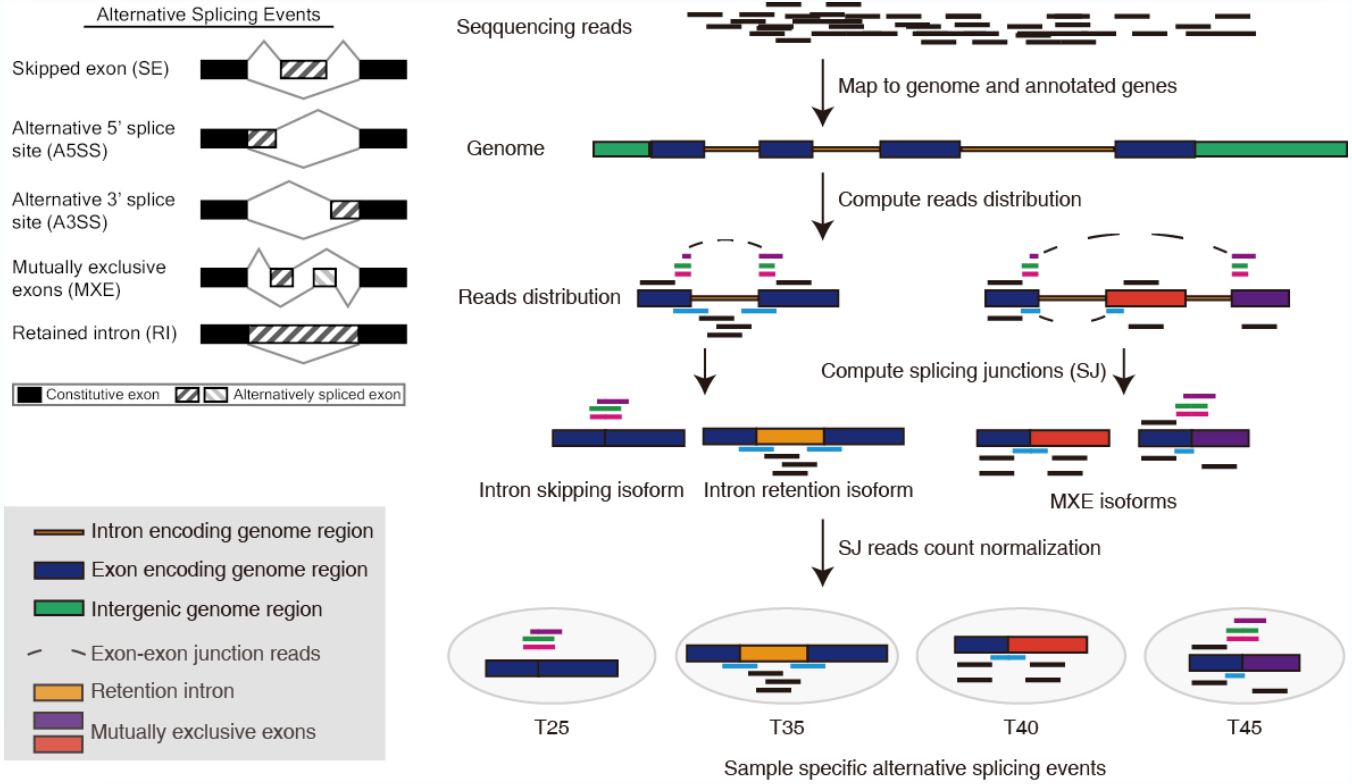
可变的3' 剪接位点

外显子互斥

内含子保留

易生信，毕生缘；培训版权所有。

样品特异的选择性剪接事件的鉴定



易生信，毕生缘；培训版权所有。

rMATS必需参数

Required Parameters:

--s1 s1.txt	A text file contains FASTQ file(s) for the sample_1.(**Only if using fastq**)
--s2 s2.txt	A text file contains FASTQ file(s) for the sample_2.(**Only if using fastq**)
--b1 b1.txt	A text file records mapping results for the sample_1 in bam format. (**Only if using bam**)
--b2 b2.txt	A text file records mapping results for the sample_2 in bam format. (**Only if using bam**)
-t readType	Type of read used in the analysis. readType is either 'paired' or 'single'. 'paired' is for paired-end data and 'single' is for single-end data
--readLength <int>	The length of each read
--gtf gtfFile	An annotation of genes and transcripts in GTF format
--bi STARIndexFolder	The folder name of the STAR binary indexes (i.e., the name of the folder that contains SA file). For example, use ~/STARindex/hg19 for hg19. (**Only if using fastq**)
--od outDir	The output directory

s1.txt和s2.txt是以逗号分隔的样本FASTQ文件

b1.txt和b2.txt是以逗号分隔的样本bam文件

-t readType双端测序为paired，单端测序为single

测序reads的长度

gtf注释文件

输入文件是fastq格式时，指定STAR索引文件位置

输出结果文件夹

易生信，毕生缘；培训版权所有。

rMATS可选参数

Optional:

--tophatAnchor <int>	The "anchor length" or "overhang length" used in the aligner. At least “anchor length” NT must be mapped to each end of a given junction. The default is 6. (**This parameter applies only if using fastq**)	
--nthread <float>	The number of thread. The optimal number of thread should be equal to the number of CPU core.	线程数
--cstat <float>	The cutoff splicing difference. The cutoff used in the null hypothesis test for differential splicing. The default is 0.0001 for 0.01% difference. Valid: 0 ≤ cutoff < 1	
--tstat <float>	The number of thread for statistical model.	统计计算时线程数
--statoff	Turn statistics part off.	
-libType libraryType	Library type. Default is unstranded (**fr-unstranded**). Use **fr-firststrand** or **fr-secondstrand** for strand-specific data.	建议加上

易生信，毕生缘；培训版权所有。

rMATS运行脚本

- python rMATS-turbo-xxx-UCSx/rmats.py --s1 s1.txt --s2 s2.txt --gtf gtfFile --bi STARindexFile --od out_directory -t paired --nthread 2 --tstat 2 --readLength 101 --tophatAnchor 8 --cstat 0.0001

  #输入文件是fastq格式

- python rMATS-turbo-xxx-UCSx/rmats.py --b1 b1.txt --b2 b2.txt -gtf gtfFile --od bam_out_directory -t paired --nthread 2 --tstat 2 --readLength 101 --cstat 0.0001 --libType fr-unstranded

  #输入文件是bam格式

易生信，毕生缘；培训版权所有。

rMATS输出结果解释

Table 11.1: rMATS输出结果解释

V1	V2
AS_Event.MATS.JC.txt	依赖Junction reads定义的剪接位点丰度
AS_Event.MATS.JCEC.txt	依赖Junction reads和对应原件其它reads定义的剪接位点丰度
fromGTF.AS_Event.txt	从GTF中提取的所有的可能的AS(alternative splicing)事件
JC.raw.input.AS_Event.txt	剪接位点的Junction reads count
JCEC.raw.input.AS_Event.txt	剪接位点和剪接元件上的总reads count

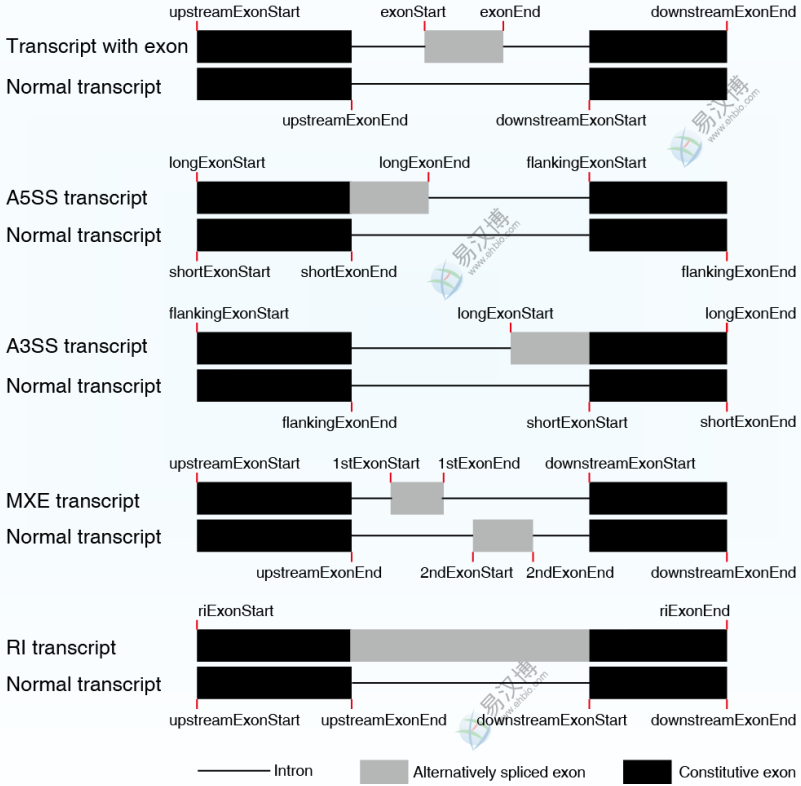
易生信，毕生缘；培训版权所有。

rMATS运行结果

- AS_Event.MATS.JC.txt
  evaluates splicing with only reads that span splicing junctions
- AS_Event.MATS.JCEC.txt
  evaluates splicing with reads that span splicing junctions and reads on target
- fromGTF.AS_Event.txt
  all possible alternative splicing (AS) events derived from GTF and RNA
- JC.raw.input.AS_Event.txt
  evaluates splicing with only reads that span splicing junctions
- JCEC.raw.input.AS_Event.txt
  evaluates splicing with reads that span splicing junctions and reads on target

易生信，毕生缘；培训版权所有。

rMATS输出结果每列含义解释



易生信，毕生缘；培训版权所有。

rMATS输出结果每列含义解释

ID	可变剪接分析编号，标示使用，无特殊含义
GeneID	注释的基因的名字
geneSymbol	基因名字
chr	染色体名字
strand	基因处在染色体的正链还是负链
riExonStart_0base	保留内含子后的新外显子在基因组的起始位置 (0-based, included)
riExonEnd	保留内含子后的新外显子在基因组的终止位置 (0-based, not included)
upstreamES	发生可变剪接的外显子上游外显子起始位置 (0-based, included)
upstreamEE	发生可变剪接的外显子上游外显子终止位置 (0-based, not included)
downstreamES	发生可变剪接的外显子下游外显子起始位置 (0-based, included)
downstreamEE	发生可变剪接的外显子下游外显子终止位置 (0-based, not included)
longExonStart_0base	A3SS中下游外显子最左侧剪接位点或A5SS中上游外显子起始
longExonEnd	A3SS中下游外显子终止位置或A5SS中上游外显子最右侧剪接位点
shortES	A3SS中下游外显子次左侧剪接位点或A5SS中上游外显子起始
shortEE	A3SS中下游外显子终止位置或A5SS中上游外显子次右侧剪接位点
flankingES	A3SS中上游外显子起始位置或A5SS中下游外显子起始位置
flankingEE	A3SS中上游外显子终止位置或A5SS中下游外显子终止位置

1stExonStart_0base	MXE中第一个使用的外显子的起始位置
1stExonEnd	MXE中第一个使用的外显子的终止位置
2ndExonStart_0base	MXE中第二个使用的外显子的起始位置
2ndExonEnd	MXE中第二个使用的外显子的终止位置
IJC_T40	T40样品中包含可变剪接位置处Junction reads数，不同的重复用逗号分开
SJC_T40	T40样品中去除可变剪接位置处Junction reads数，不同的重复用逗号分开
IJC_T45	T45样品中包含可变剪接位置处Junction reads数，不同的重复用逗号分开
SJC_T45	T45样品中去除可变剪接位置处Junction reads数，不同的重复用逗号分开
IncFormLen	包含可变剪接部分的转录本的长度
SkipFormLen	不包含可变剪接部分的转录本长度
Pvalue	差异检验显著性度量
FDR	多重假设检验校正的假阳性率
IncLevel1	样品1（T40)中包含可变剪接部分的转录本的表达，不同的重复用逗号分开
IncLevel2	样品2（T45)中包含可变剪接部分的转录本的表达，不同的重复用逗号分开
IncLevelDifference	样品1(T40)的转录本表达量平均值减去样品2的转录本表达平均值(T45)
Remaining columns	Function annotation

生缘；培训版权所有。

rmats2sashimiplot必需参数

Required Parameters

```
--s1 s1_rep1.sam[,s1_rep2.sam]     Mapping results for the sample_1 in sam format.
                                Replicates must be in a comma separated list
                                (Only if using sam).
--s2 s2.rep1.sam[,s2.rep2.sam]     Mapping results for the sample_2 in sam format.
                                Replicates must be in a comma separated list
                                (Only if using sam).
--b1 s1_rep1.bam[,s1_rep2.bam]     Mapping results for the sample_1 in bam format.
                                Replicates must be in a comma separated list
                                (Only if using bam).
--b2 s2.rep1.bam[,s2.rep2.bam]     Mapping results for the sample_2 in bam format.
                                Replicates must be in a comma separated list
                                (Only if using bam).
-t eventType                       Type of event from rMATS result used in the
                                eventType is 'SE', 'A5SS', 'A3SS', 'MXE' or 'RI'.
                                'SE' is for skipped exon events, 'A5SS' is for
                                alternative 5' splice site events, 'A3SS' is for
                                alternative 3' splice site events, 'MXE' is for
                                mutually exclusive exons events and 'RI' is for
                                retained intron events (Only if using rMATS format
                                result as event file).
-e eventsFile                      The rMATS output event file (Only if using rMATS
                                format result as event file).
-c coordinate:annotaionFile     The coordinate of genome region and an annotation
                                of genes and transcripts in GFF3 format. Coordinate
                                and annotation file must be colon separated
                                (Only if using coordinate and annotaion file).
--l1 SampleLabel1               The label for first sample.
--l2 SampleLabel2               The label for second sample.
-o outDir                       The output directory.
```

--s1和--s2参数：逗号分隔的样本sam文件

--b1和--b2参数：逗号分隔的样本bam文件

-t 可变剪接类型

-e 对应剪接类型的rMATS输出结果

--l1和--l2： 第一组和第二组样本名称

Rmats2sashimiplot可选参数

```
Optional:
--exon_s <int>                  The size of scale down exons. The default is 1.
--intron_s <int>              The size of scale down introns. For example, if
                              -intron_s is 5, it means the size of intron is 5:1
                              (if the real size of intron is 5, the size in the
                              plot will be scaled down to 1). The default is 1.
--group-info                  If user want to divide samples into groups,
                                          they can specify this parameter with a "*.gf" file.
                                          Format specification can be found in following
                              section.
--min-counts                  If the junction count is smaller(<) than this float
                                          number, then this junction would be omitted. The
                                          default value is 3. If you want to display all the
                                          numbers, then set it as 0.
--color                                 User can customerize the colors of the plot using a
                                          sequence of color. The number of the colors are
                              supposed to be corresponding to that of bam_files.
                              eg: --color #FFCC99,#99CC99,#99CC99
--font-size                   Change the default font size which equals to 8.
--no-text-background          Transparent text background.
--hide-number                           Hide the numbers of junction.

-h                                        Print this help message and exit(also --help).
```

rmats2sashimiplot运行脚本

- rmats2sashimiplot --s1 sampleA.R1.sam,sampleA.R2.sam,sampleA.R3.sam --s2 sampleB.R1.sam,sampleB.R2.sam,sampleB.R3.sam -t SE -e /MATS_output/sampleA_sampleB.SE.MATS.events.txt --l1 sampleA --l2 sampleB --exon_s 1 --intron_s 5 -o events_output
  #输入文件是sam格式

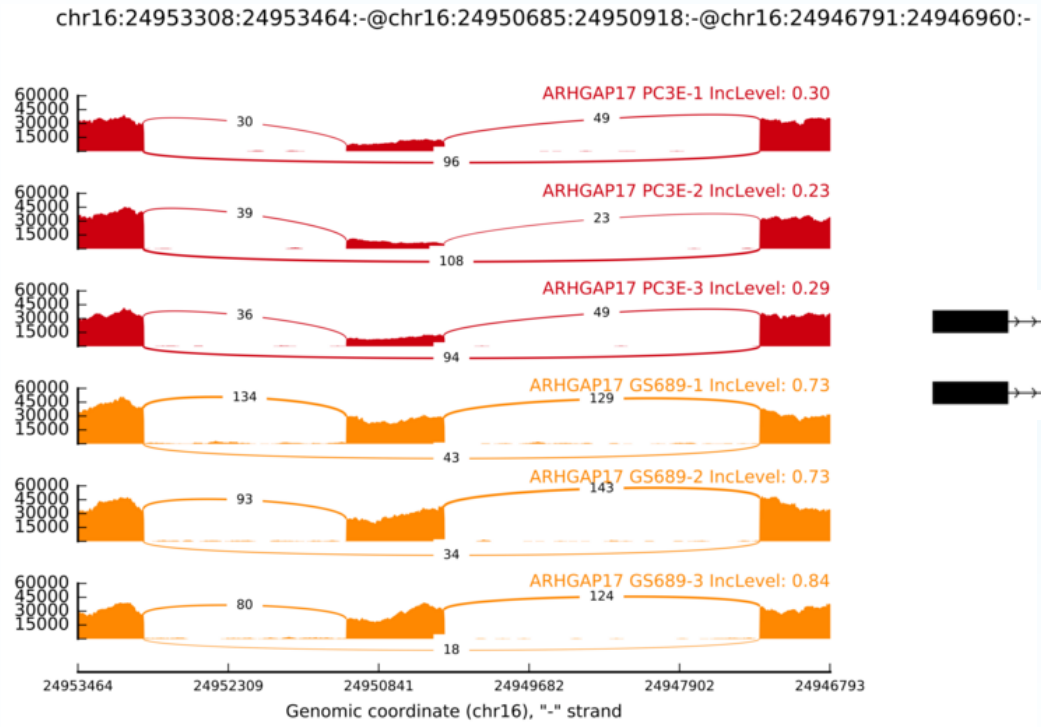
易生信，毕生缘；培训版权所有。

rmats2sashimiplot运行脚本

- rmats2sashimiplot --b1 sampleA.R1.bam,sampleA.R2.bam,sampleA.R3.bam --b2 sampleB.R1.bam,sampleB.R2.bam,sampleB.R3.bam -c chr16:-:24944500:24955500:.gff3File --l1 sampleA --l2 sampleB --exon_s 1 --intron_s 5 -o events_output
  #输入文件是bam格式，用户提供坐标和gff3格式注释文件

易生信，毕生缘；培训版权所有。

sashimiplot

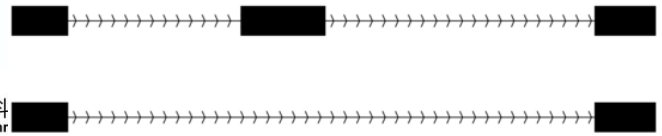


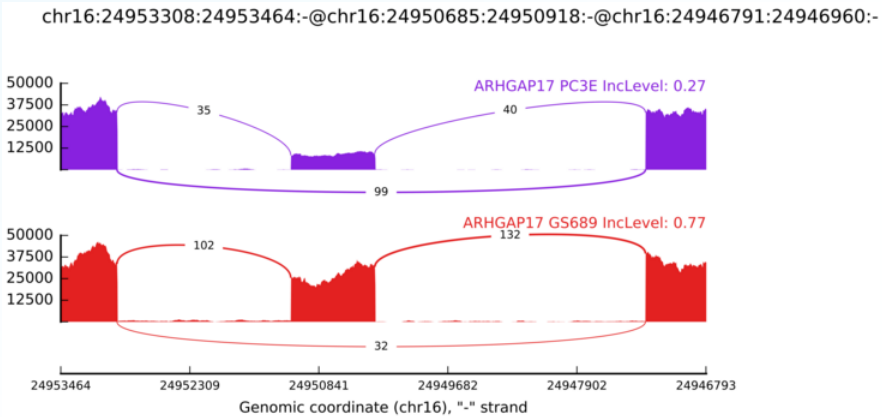
易生信，毕生缘；培训版权所有。

以组的方式展示结果

- rmats2sashimiplot --b1 sampleA.R1.bam,sampleA.R2.bam,sampleA.R3.bam --b2 sampleB.R1.bam,sampleB.R2.bam,sampleB.R3.bam -t SE -e /MATS_output/sampleA_sampleB.SE.MATS.events.txt --l1 sampleA --l2 sampleB --exon_s 1 --intron_s 5 -o events_output --group-info grouping.gf

#grouping.gf:
group1name: 1-3
group2name: 4-6





毕生缘；培训版权所有。

Sequencing costs a lot and gains more

长按关注 生信宝典 ， 简单入门，快速晋级