Alternative-splicing detection by NGS

Wen-Dar Lin

Bioinformatics core, IPMB

wdlin@gate.sinica.edu.tw

Preface

- In addition to gene expressions, alternative splicing isoforms provide diversity of RNAs and protein products.
- In this presentation, we will go through theories of three programs for alternative splicing analyses,
  - as well as a section of a way of doing corresponding motif discovery.
- Files: PowerPoints, walk-through logs, and example data
  - https://maccu.project.sinica.edu.tw/20211007/
    - would have some update by noon of 20211007

Disclaimer

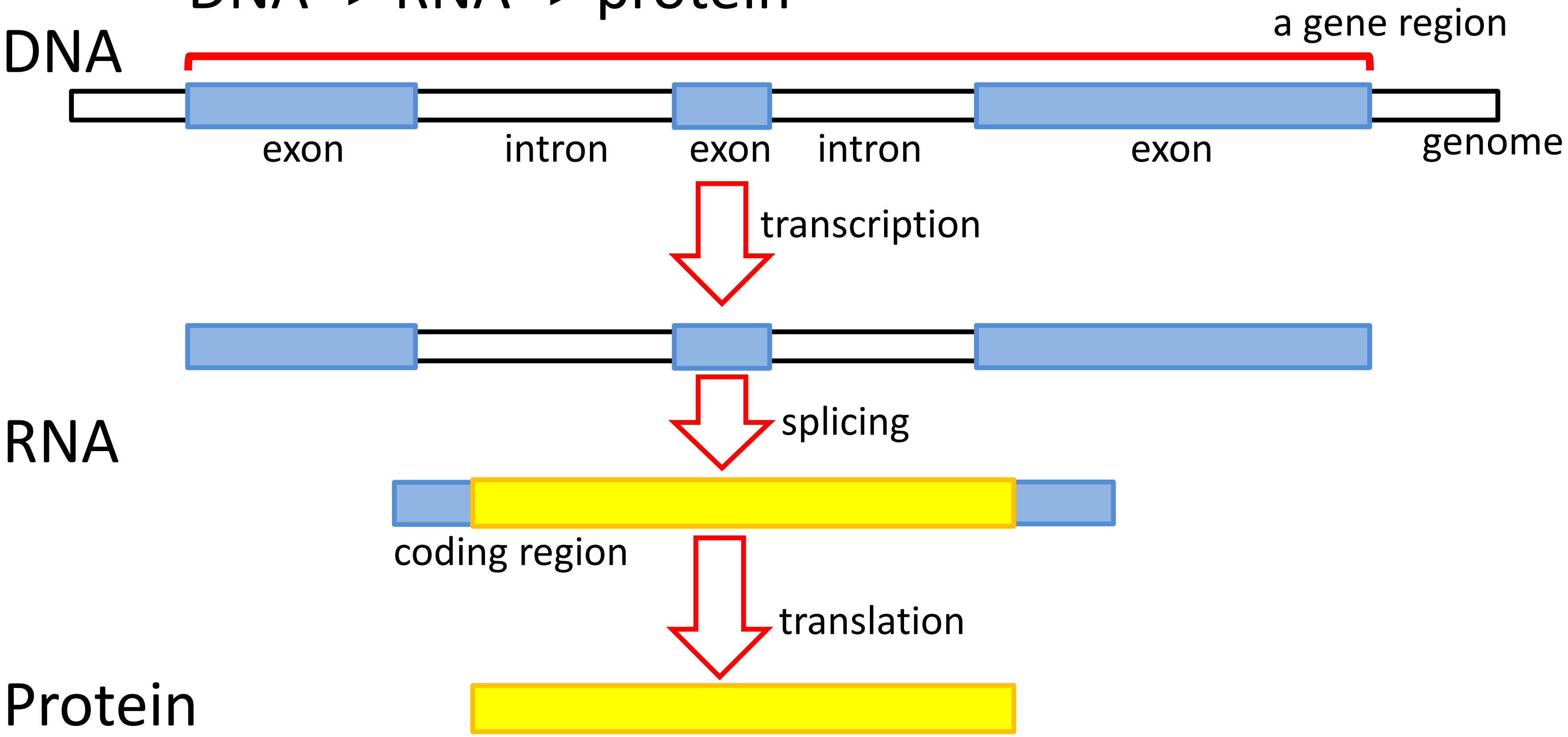
- This presentation was made based on my work experiences
  - mainly for plants.
- This presentation is *not* intended to cover related biology knowledge.
- In this presentation, the words “transcript” and “isoform” have the same meaning.
  - In some context, isoforms mean protein variants

Topics

1. Detecting alternative splicing (AS),
2. Theories of isoform-based algorithms,
3. Theories of event-based algorithms,
5. AS-related motif discovery,
4. Walk-throughs of AS computation programs, and
6. Discussions.

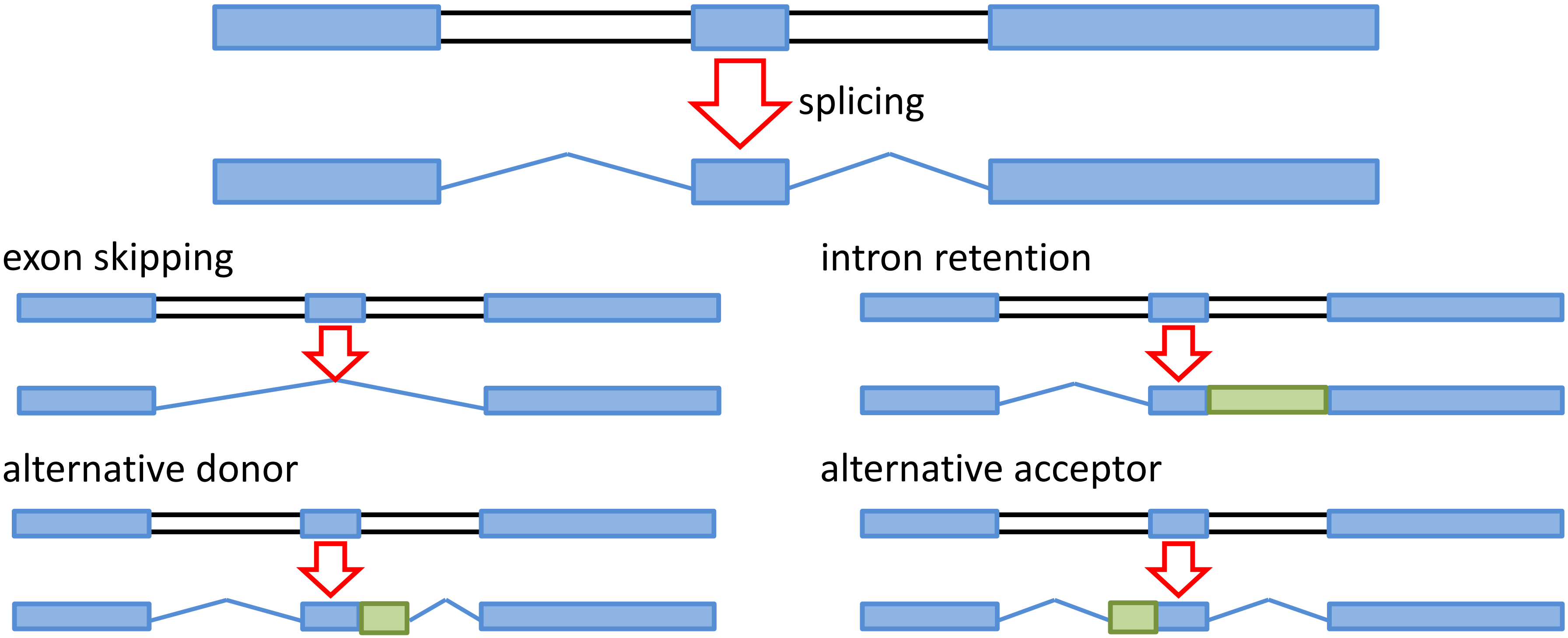
Detecting alternative splicing

- The central dogma
  - DNA -> RNA -> protein



Detecting alternative splicing

- Splicing events
  - Types of splicing junction variation



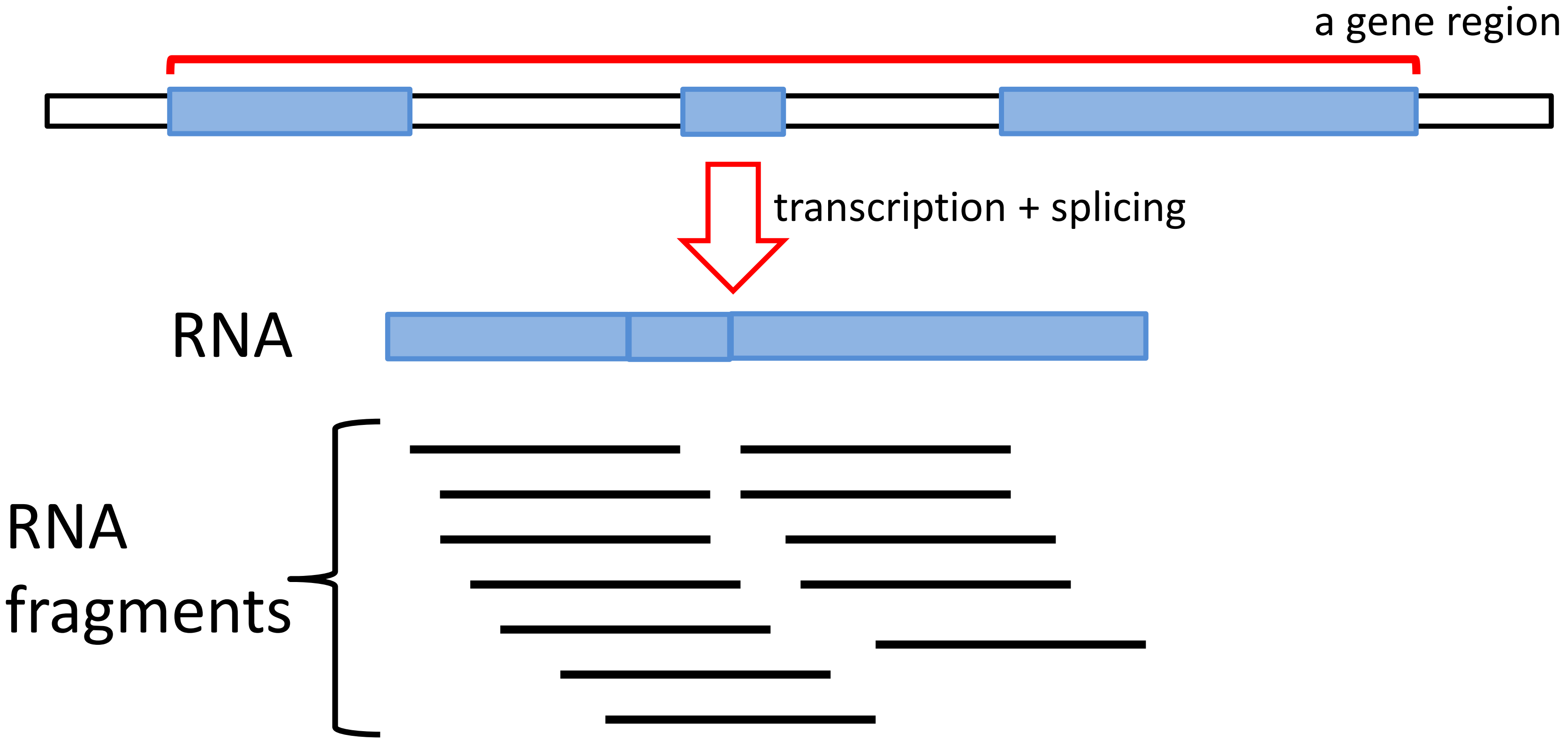
Various combinations of splicing events => various isoforms

Detecting alternative splicing

- Currently, algorithms said to be detecting alternative splicing can be *roughly* classified into two categories
  - Isoform-based
    - Predict expressed isoforms (*combinations of splicing events*)
    - Predict expression levels of isoforms => differential expressed isoforms
  - Event-based
    - Collect read counts related to *splicing events* and do corresponding computation

Detecting alternative splicing

- RNAseq
  - Sequencing of RNA fragments



a gene region

transcription + splicing

RNA

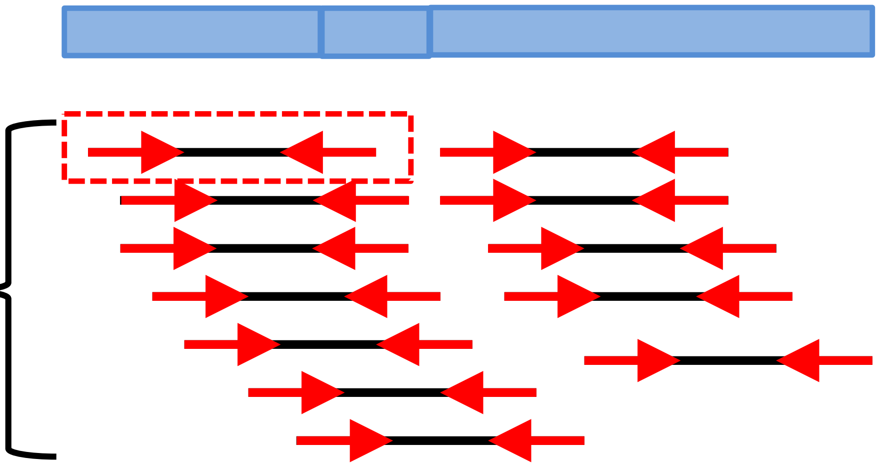
RNA fragments

Detecting alternative splicing

- Illumina YouTube video
  - https://youtu.be/fCd6B5HRaZ8
  - Keywords
    - fragment
    - lane / tile
    - amplification / cluster
    - read 1 / read 2
    - fluorescently tagged nucleotides

Detecting alternative splicing

- Read pairs in RNAseq data

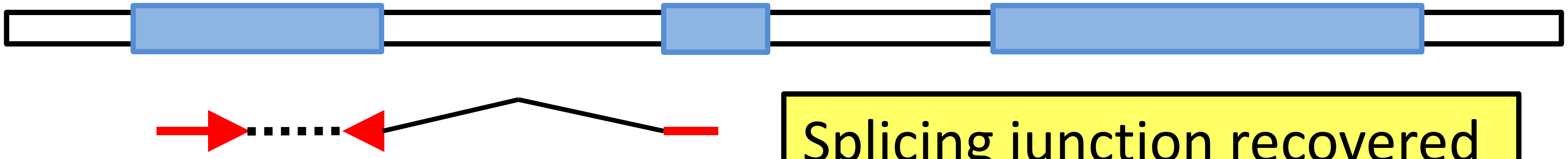


RNA

RNA fragments

Every two arrows of one fragment would be read1 and read2 in RNAseq data

When we mapping reads back to the genome



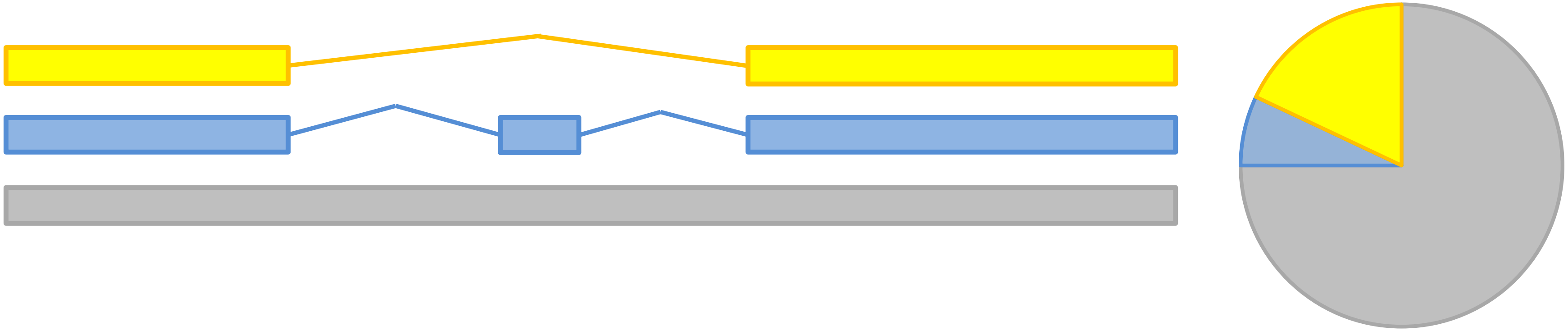
Splicing junction recovered if reads are *spanning* it

Detecting alternative splicing

- Short conclusions
  - Different isoforms were made by different combination of splicing junctions (events)
  - Splicing junctions could be recovered by RNAseq reads
  - Isoform-based methods are computing differentially expressed isoforms (*combination of splicing junctions*)
  - Event-based methods are computing differentially expressed *splicing junctions*
  - NOTE: the word “alternative” should refer to some “change of preference” from one to some other

Theories of isoform-based algorithms

- What isoform-based algorithms do?
  - Predict transcripts
  - Predict expression level of transcripts
  - Predict Differentially expressed isoforms

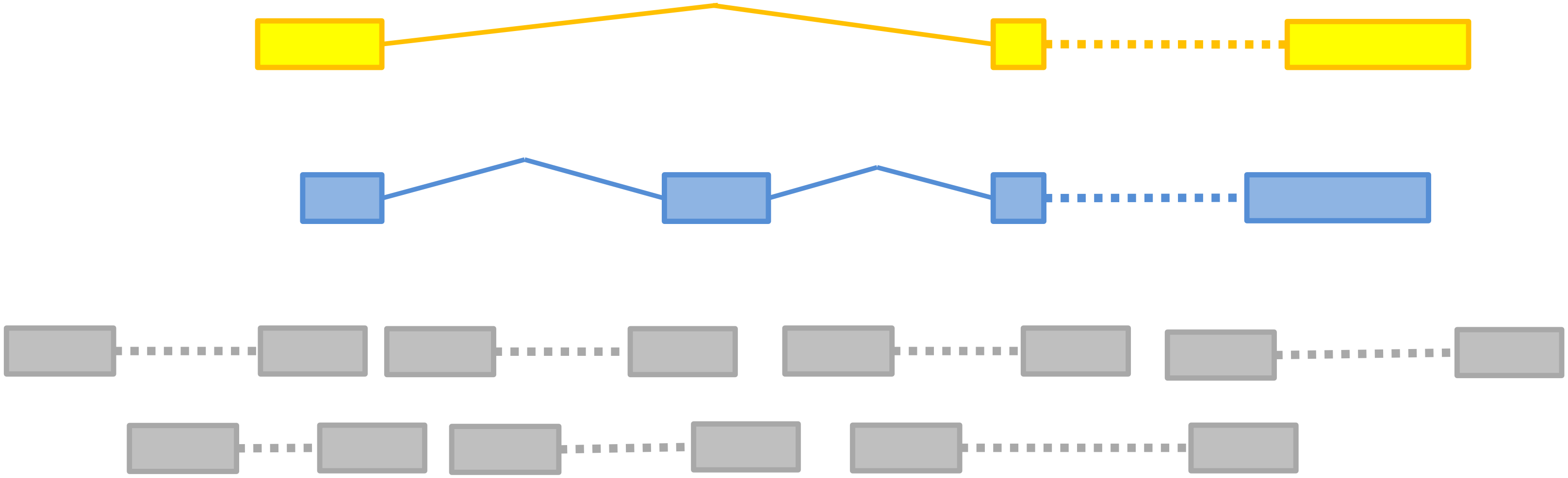


Theories of isoform-based algorithms

- In this tutorial, we will go through underlying theories of two of best isoform-based algorithms
  - Cufflinks
    - Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation
      - Trapnell *et al.*, Nat Biotechnol. 2010
  - StringTie
    - StringTie enables improved reconstruction of a transcriptome from RNA-seq reads
      - Pertea *et al.*, Nat Biotechnol. 2015

Underlying theories of Cufflinks

- Consider the following read pairs been mapped to the reference genome



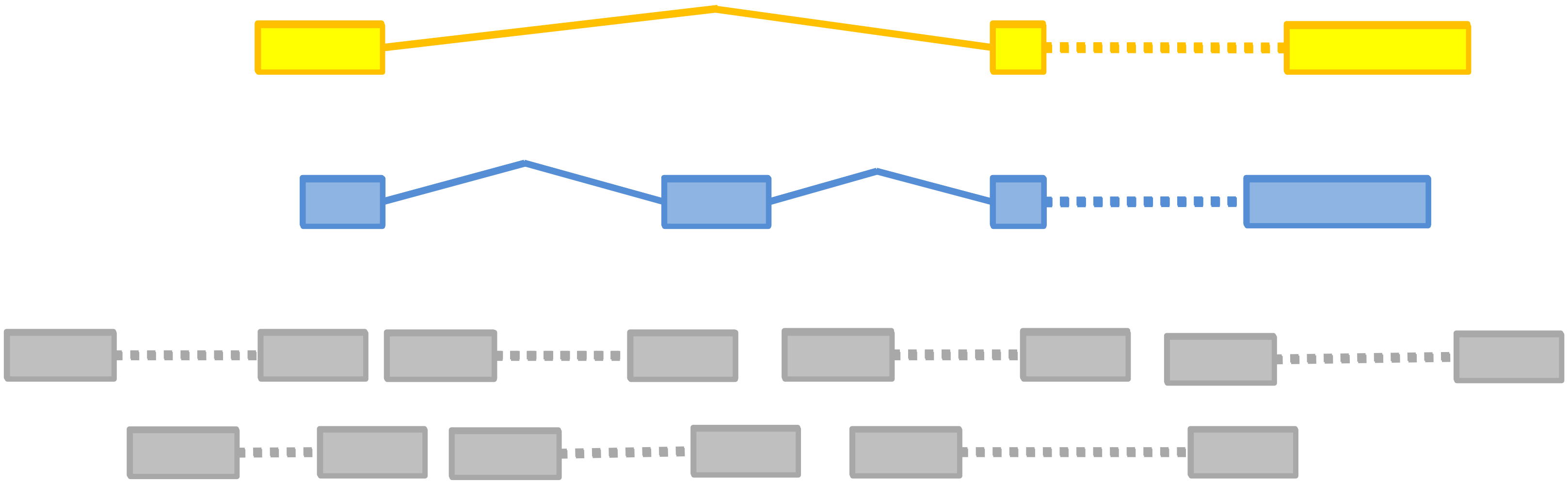
∧ splicing junction

⋯⋯ read pair

a remake of Fig1 of the Cufflinks 2010 paper

Underlying theories of Cufflinks

- For every two *overlapping* read pairs, identify whether they are *compatible* or *incompatible*



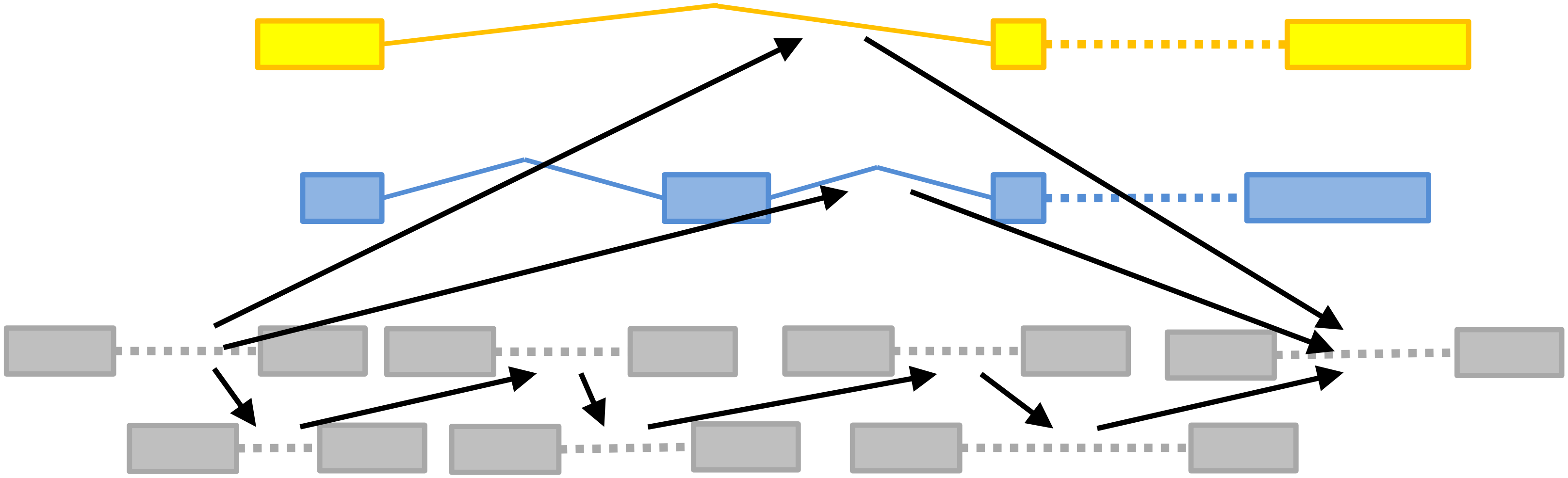
splicing junction
read pair

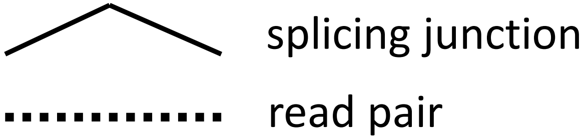
“Incompatible” means overlapping read pairs must *not* from the same isoform

a remake of Fig1 of the Cufflinks 2010 paper

Underlying theories of Cufflinks

- For every two *compatible* read pairs, define *orders* by their positions



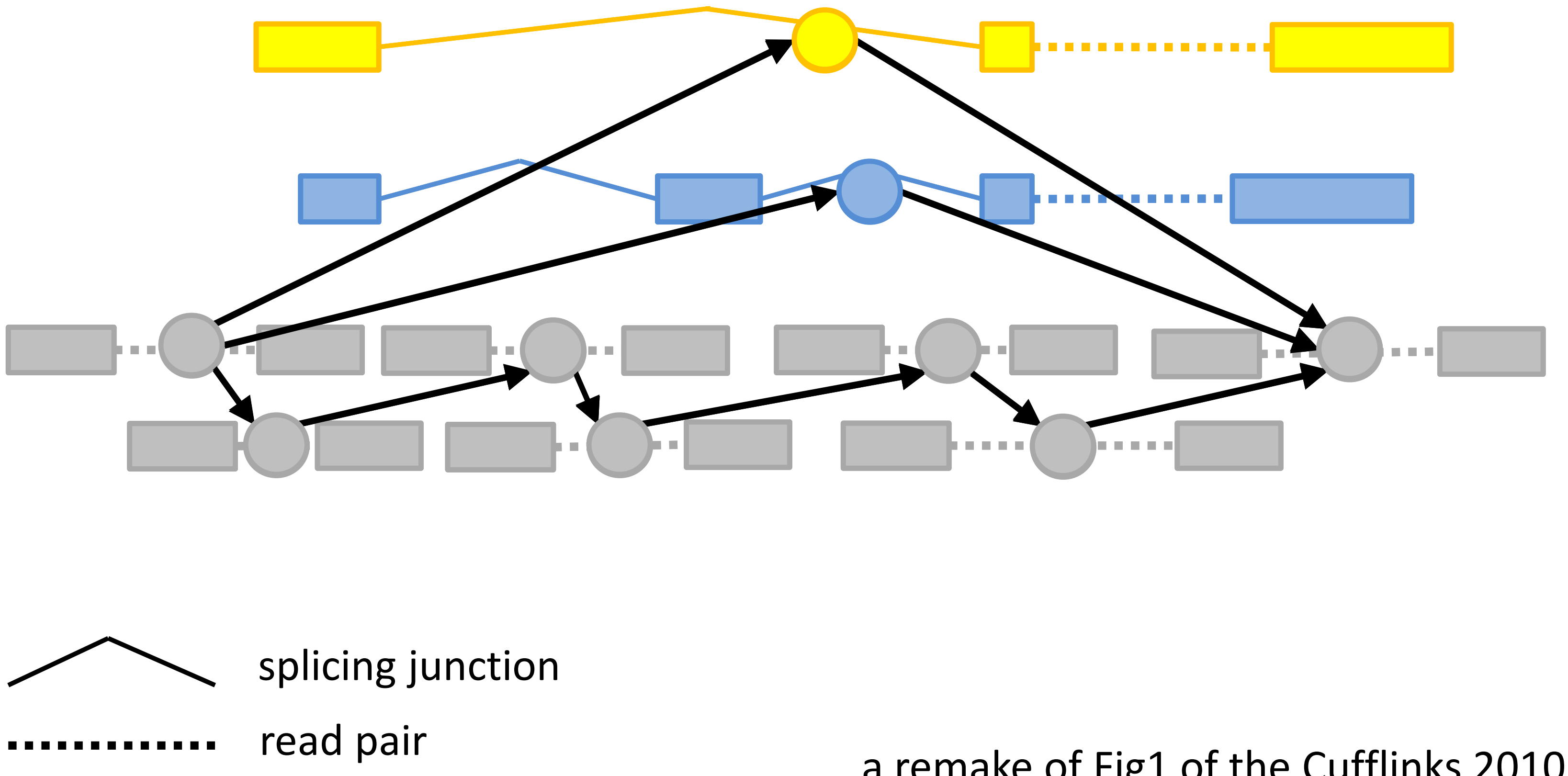


“order” in math: a<b AND b<c => a<c

a remake of Fig1 of the Cufflinks 2010 paper

Underlying theories of Cufflinks

- The Dilworth's theorem (1950) ensures the minimum number of *fully* ordered partitions



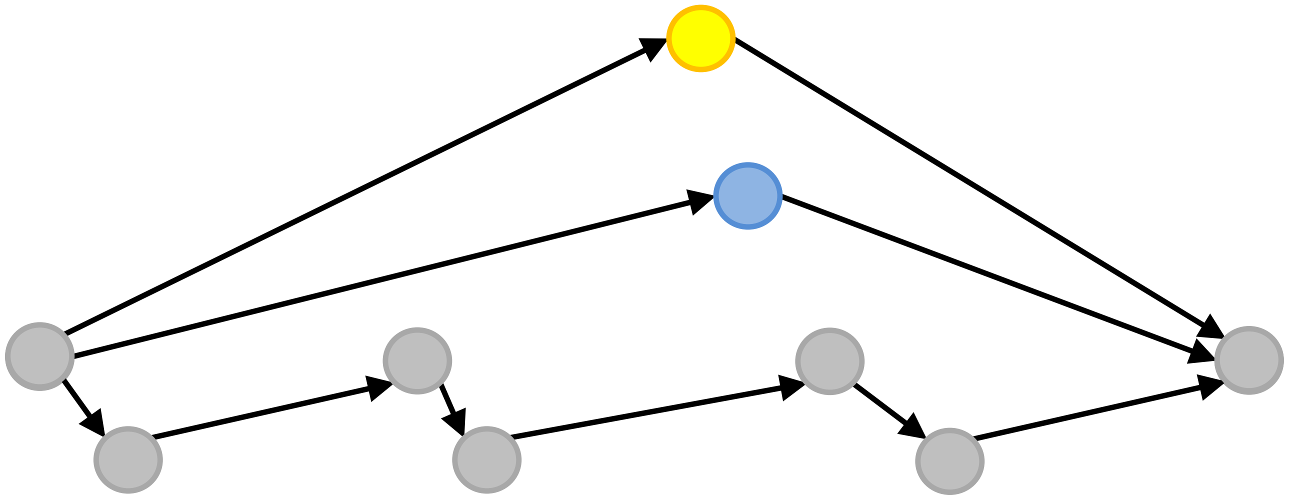
a remake of Fig1 of the Cufflinks 2010 paper

Underlying theories of Cufflinks

- The Dilworth's theorem (1950) ensures the minimum number of *fully* ordered partitions
- In English, “the minimum number of transcripts”
- The LOGIC
  - In a fully ordered partition, every two nodes can be compared => not incompatible => not “must not from the same isoform”

Underlying theories of Cufflinks

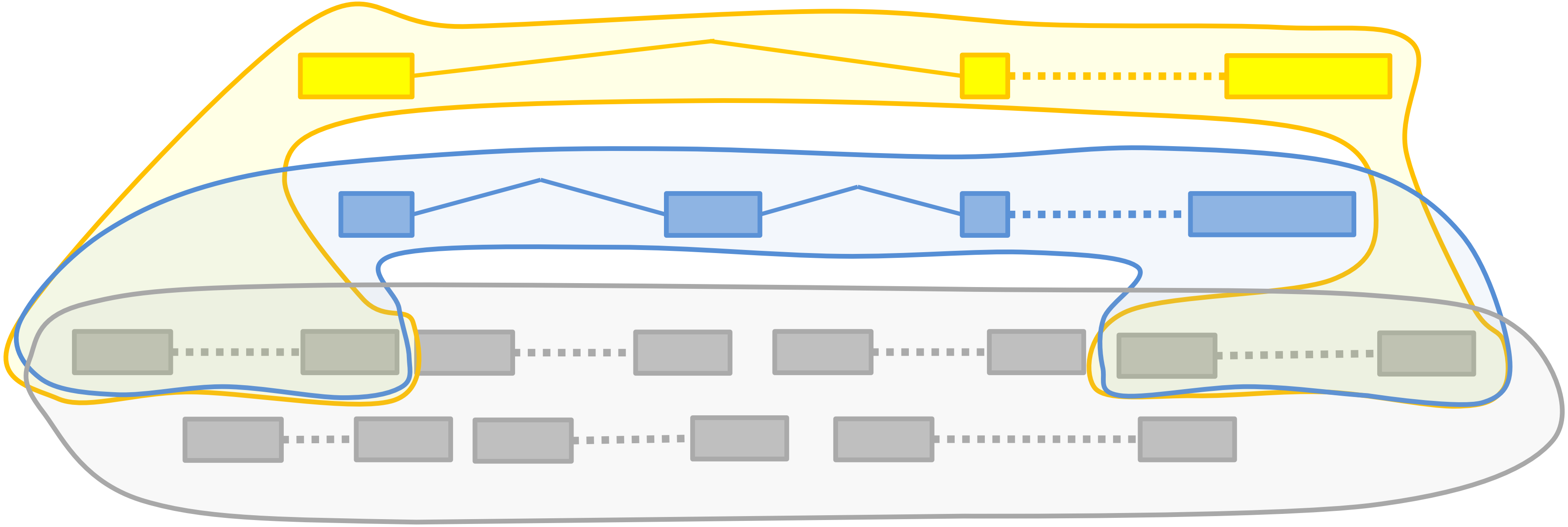
- Potential transcripts were inferred by reads from the same fully ordered partitions (1)



a remake of Fig1 of the Cufflinks 2010 paper

Underlying theories of Cufflinks

- Potential transcripts were inferred by reads from the same fully ordered partitions (2)



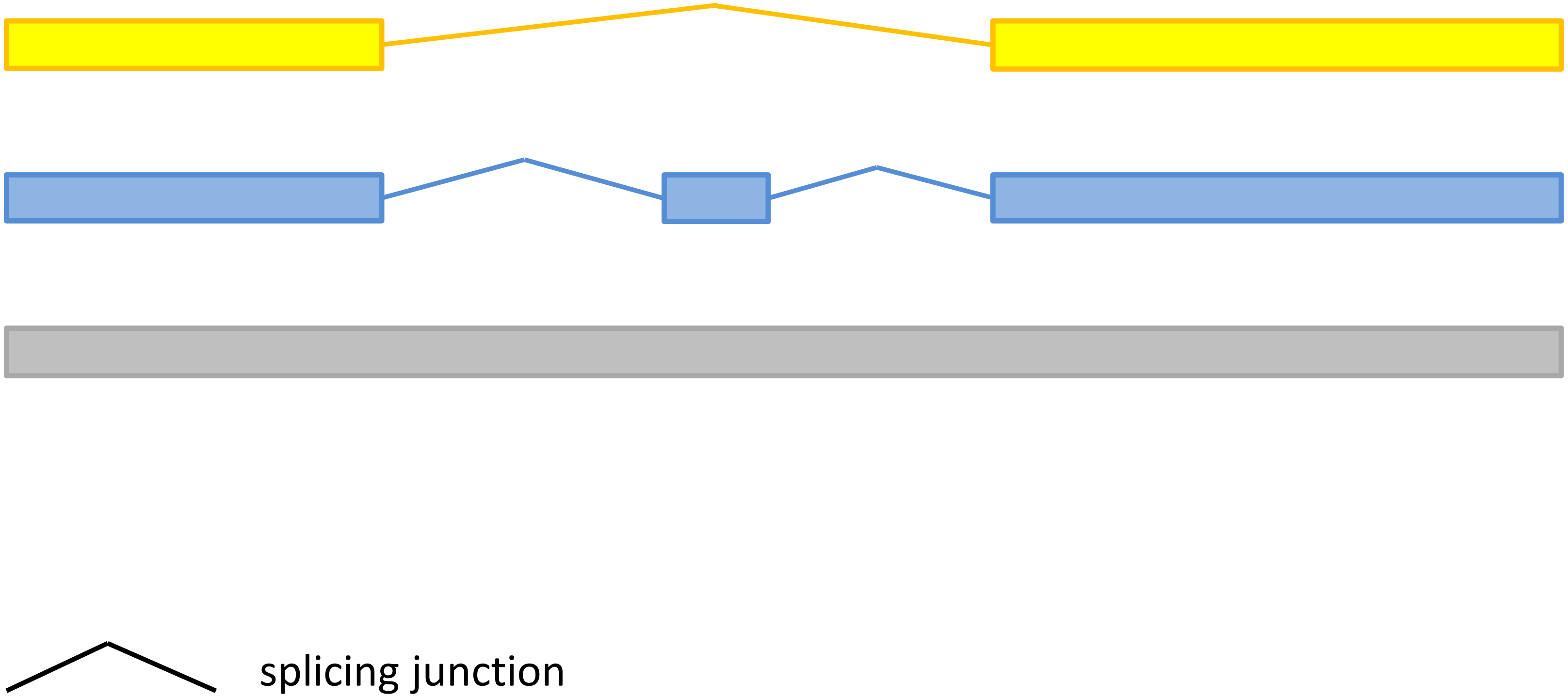
∧ splicing junction

⋯⋯ read pair

a remake of Fig1 of the Cufflinks 2010 paper

Underlying theories of Cufflinks

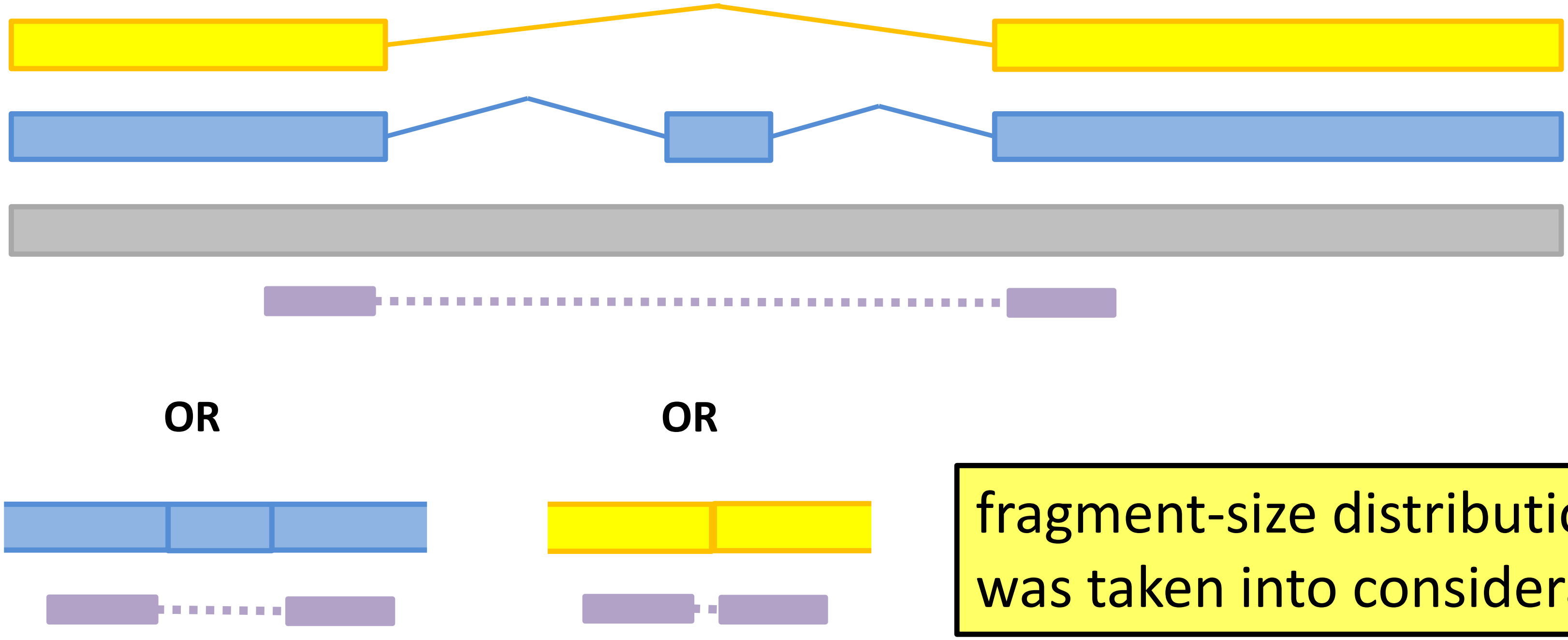
- Potential transcripts were inferred by reads from the same fully ordered partitions (3)



a remake of Fig1 of the Cufflinks 2010 paper

Underlying theories of Cufflinks

- Transcript abundance estimation was done by incorporating guesses of “which read pair is from which transcript”



fragment-size distribution was taken into consideration!

Underlying theories of Cufflinks

- Transcript abundance estimation was done by incorporating guesses of “which read pair is from which transcript” and
- finding best compositions of transcript percentages on a likelihood function

fragment-size distribution

$$\prod_{r\in R:r\in g} \sum_{t\in g} \gamma_t \frac{F(I_t(r))}{l(t) - I_t(r) + 1}$$

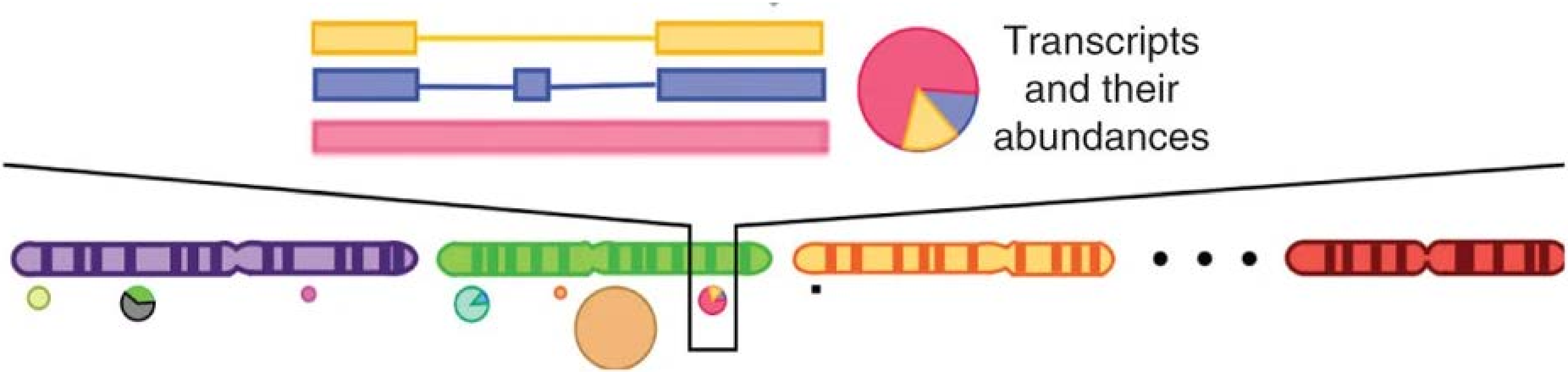
even distribution along transcript length *l*(*t*)

transcript percentages

Eq26 from supplement of Cufflinks paper

Underlying theories of Cufflinks

- A short conclusion
  - For each gene, Cufflinks generates all possible transcripts and then
  - predicts their percentages of expressions of this gene.



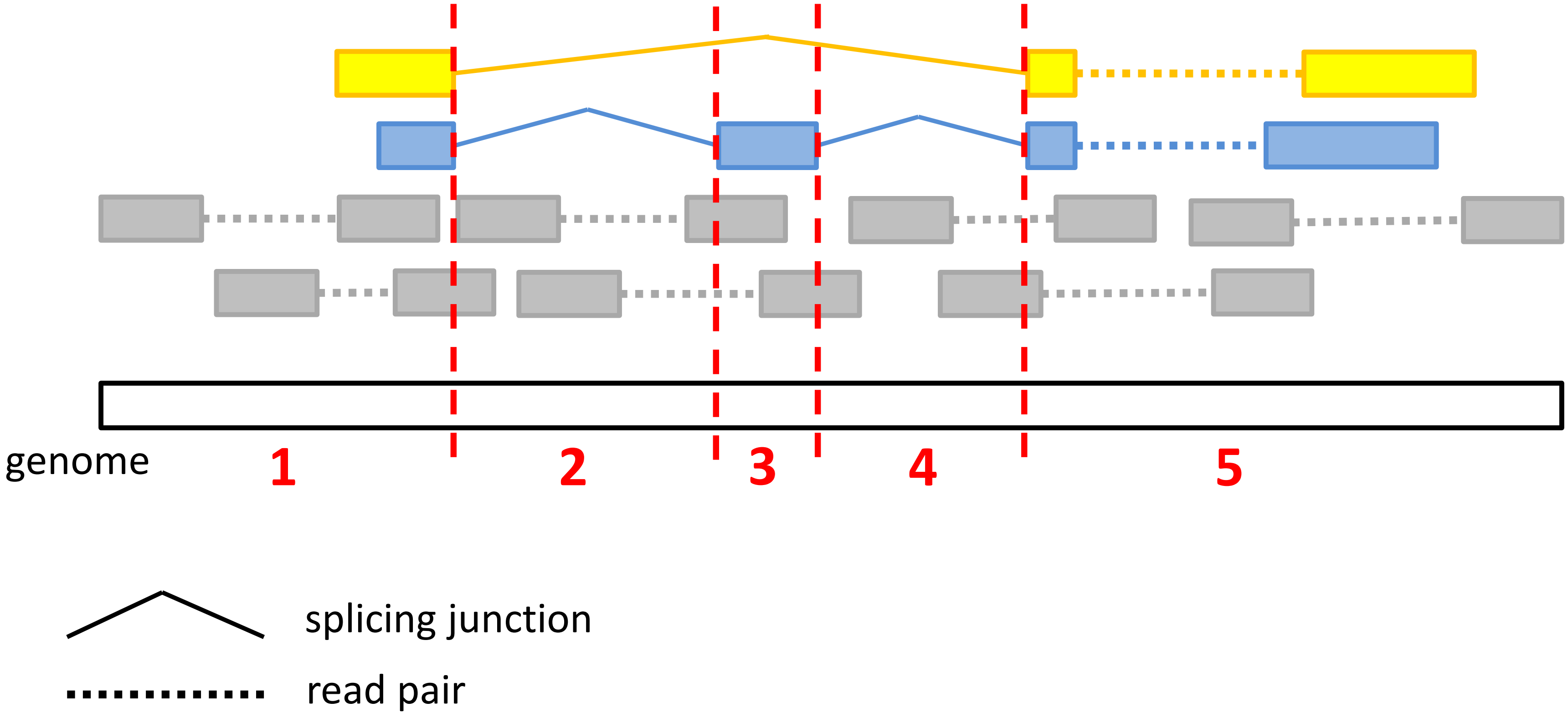
from Fig1 of Cufflinks paper

Underlying theories of StringTie

- Unlike Cufflinks, treating reads (or read-pairs) as nodes to build graphs
- StringTie
  - divides a gene region into segments (as nodes) based on splicing junctions expressed by reads
  - connect two nodes (genomic segments) if some reads are spanning them
  - treat the resulted graph as a graph of the maximum flow problem

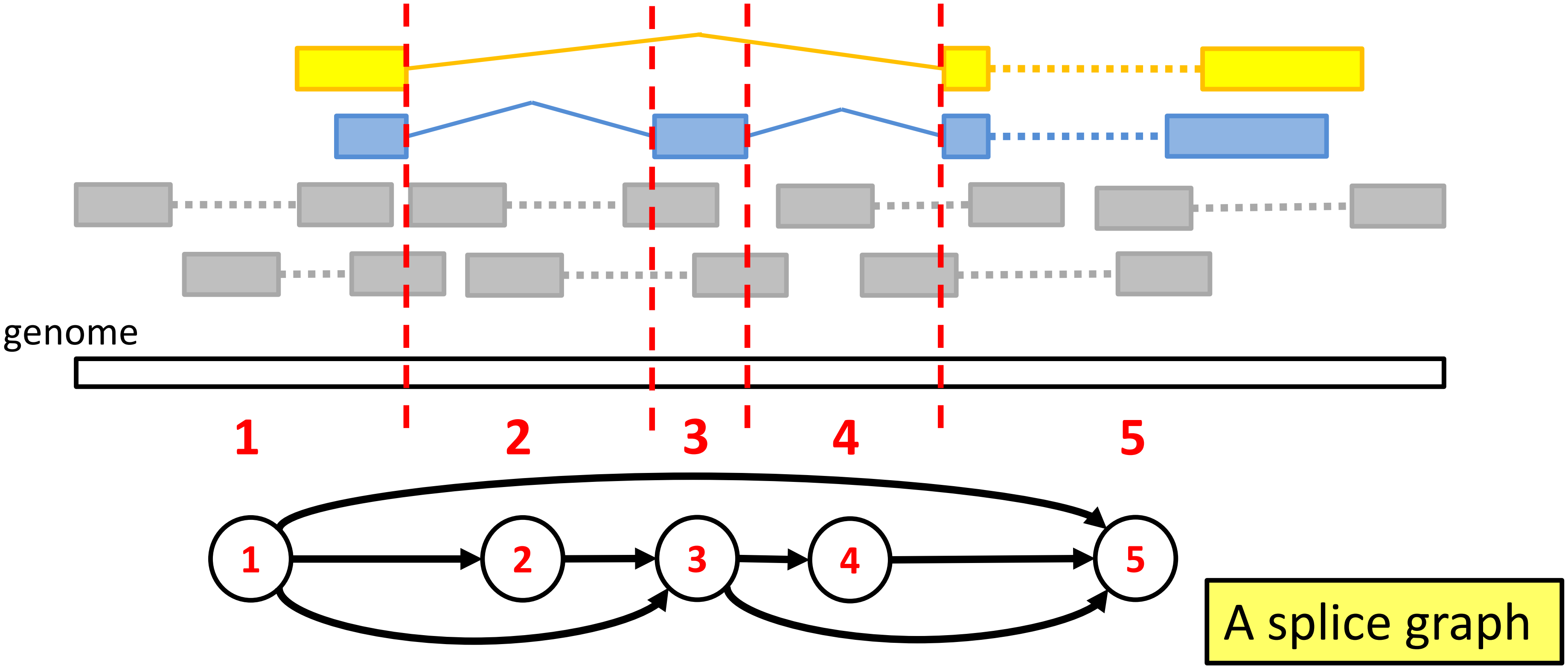
Underlying theories of StringTie

- Consider the same sets of read pairs, the first step is to divide the gene region into segments



Underlying theories of StringTie

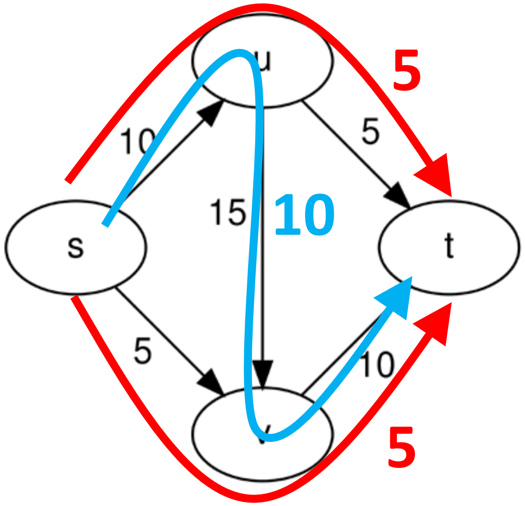
- By treating segments as nodes, connect two nodes if some reads are spanning them



A splice graph

Underlying theories of StringTie

- The next step is to transform the problem into a maximum flow problem
- What is a maximum flow problem?
  - "finding a *feasible* flow through a flow network that obtains the maximum possible flow rate"

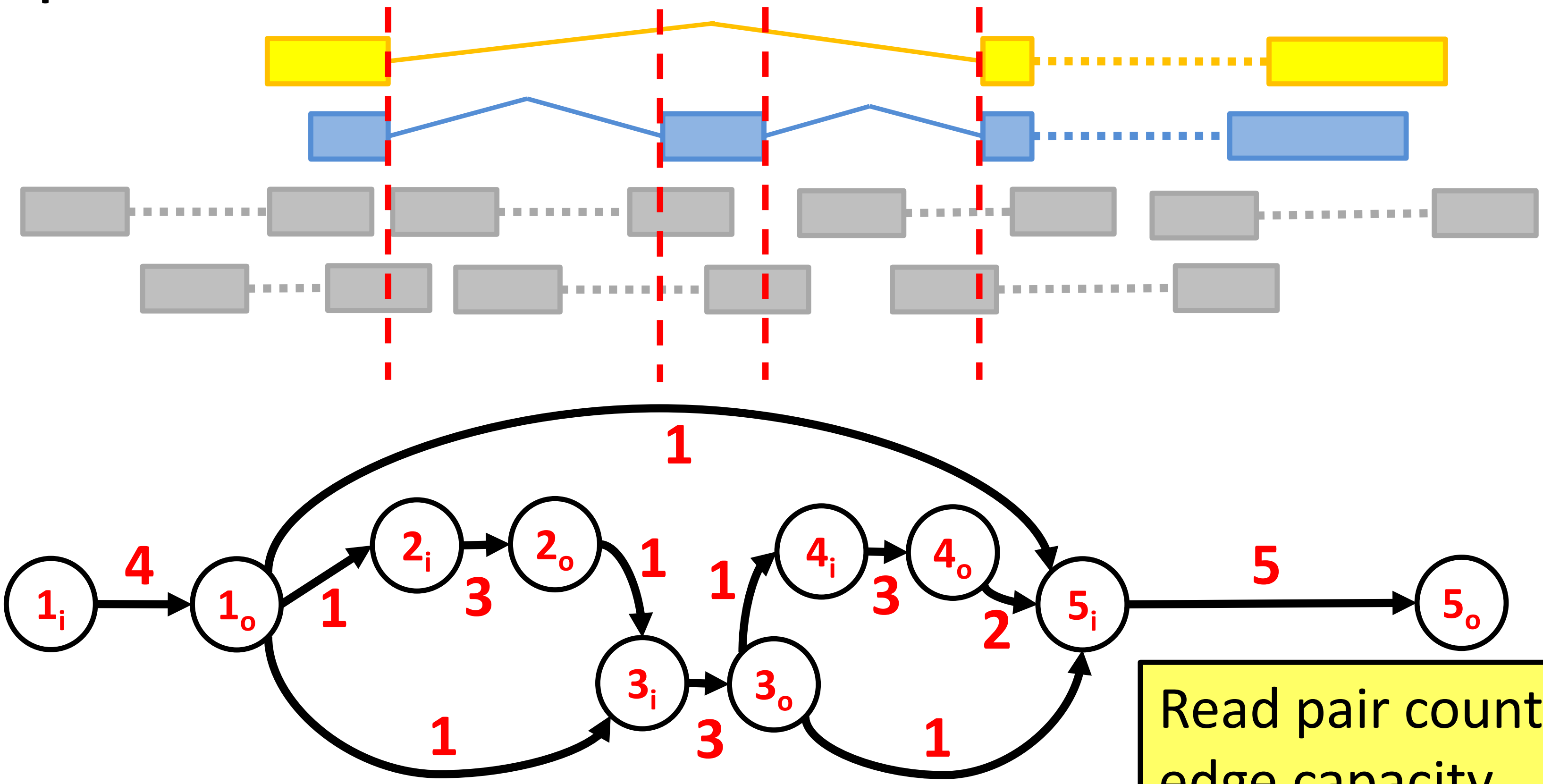


How much flow can be obtained from source to terminal?
(black numbers as *capacity*)

Source: wikipedia

Underlying theories of StringTie

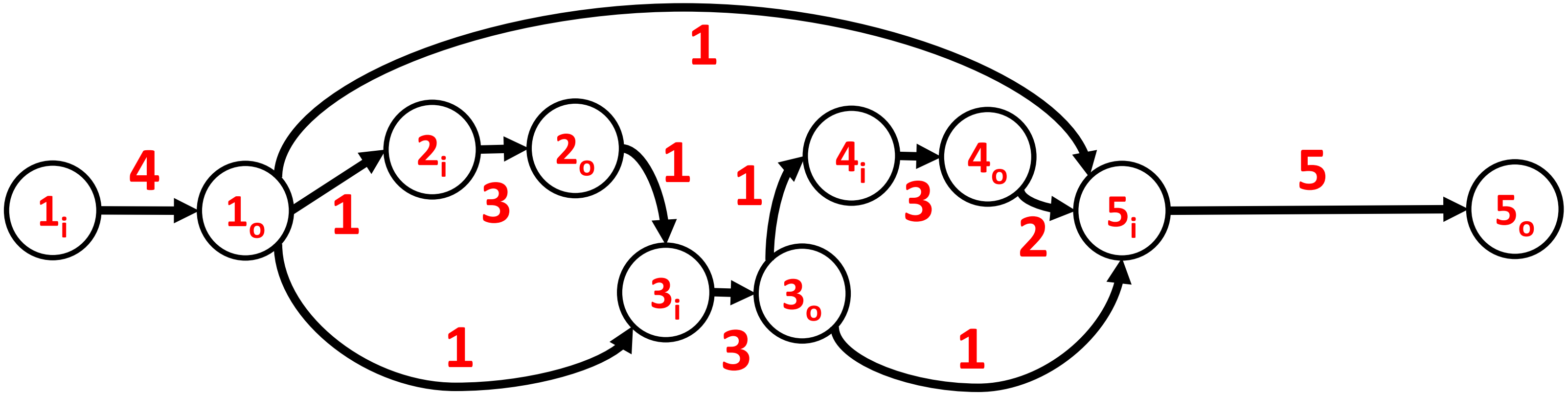
- Transform the graph into a maximum flow problem



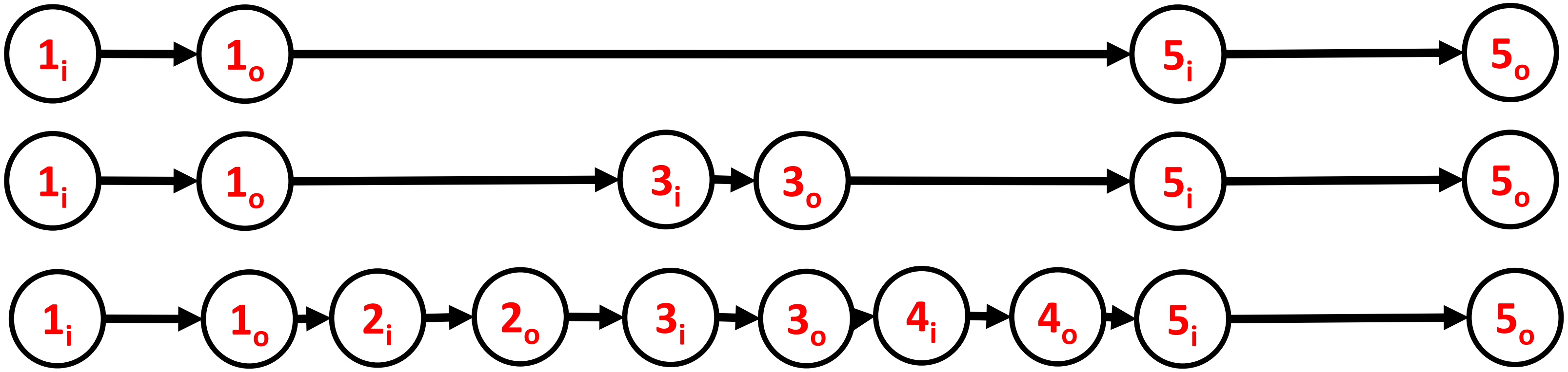
Read pair counts as edge capacity

Underlying theories of StringTie

- The maximum flow?



- Three paths, each with flow 1



Underlying theories of StringTie

- By treating each path as an isoform, we would obtain the same three isoforms as what we have by the Cufflinks algorithm
  - For each isoform, StringTie counts reads when computing the corresponding flow
  - => expression of the isoform

Theories of isoform-based algorithms

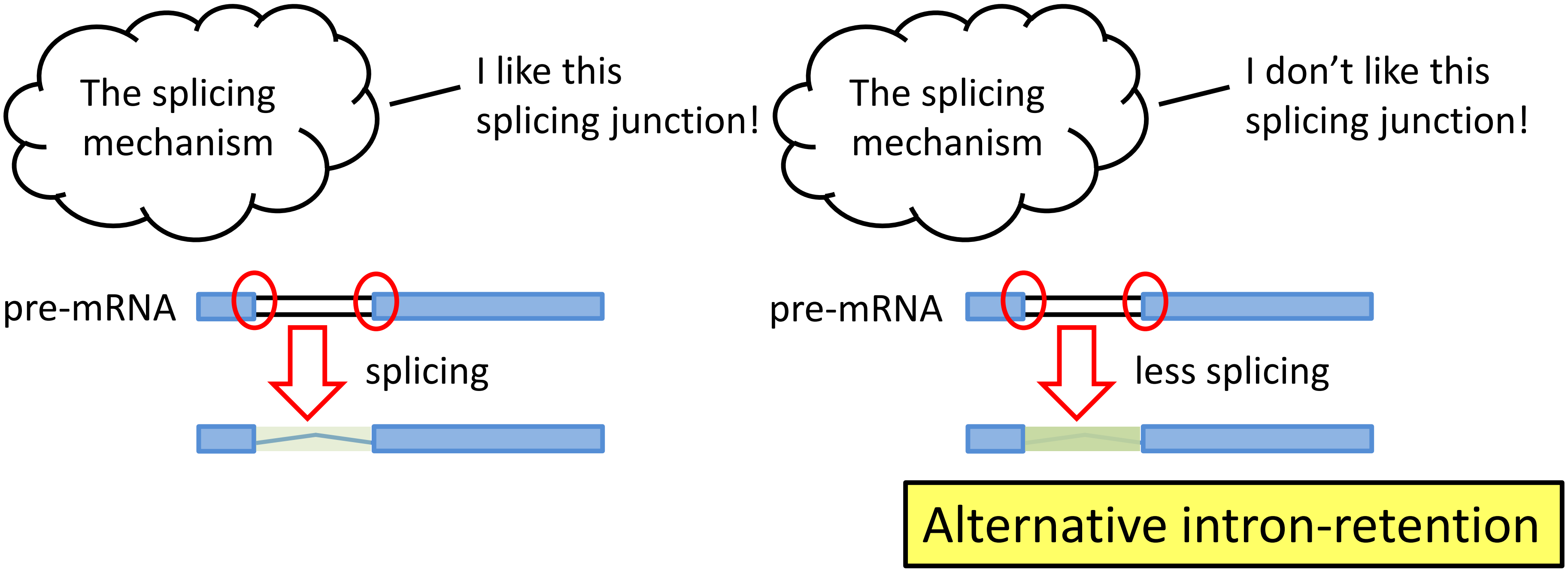
- Short conclusions
  - Reasonably transforming questions into some mathematical models could be helpful for solving problems.

Theories of event-based algorithms

- Cautions
  - This part contains methods that I have been applying for years in my works
    - But not general descriptions of event-based algorithms
  - All mentioned methods have been incorporated in a (few) number of papers
  - Software repository: RackJ
    - https://sourceforge.net/projects/rackj/
    - Direct binary download: https://downloads.sourceforge.net/project/rackj/0.99a/rackJ.tar.gz
    - subversion command for source code:
      - svn checkout svn://svn.code.sf.net/p/rackj/code/tags/trunk YourDir
      - need apache ant to compile

Theories of event-based algorithms

- The underlying thinking of the methods to be described is
  - to taking *preference* of the splicing mechanism into consideration



The splicing mechanism

I like this splicing junction!

pre-mRNA

splicing

The splicing mechanism

I don't like this splicing junction!

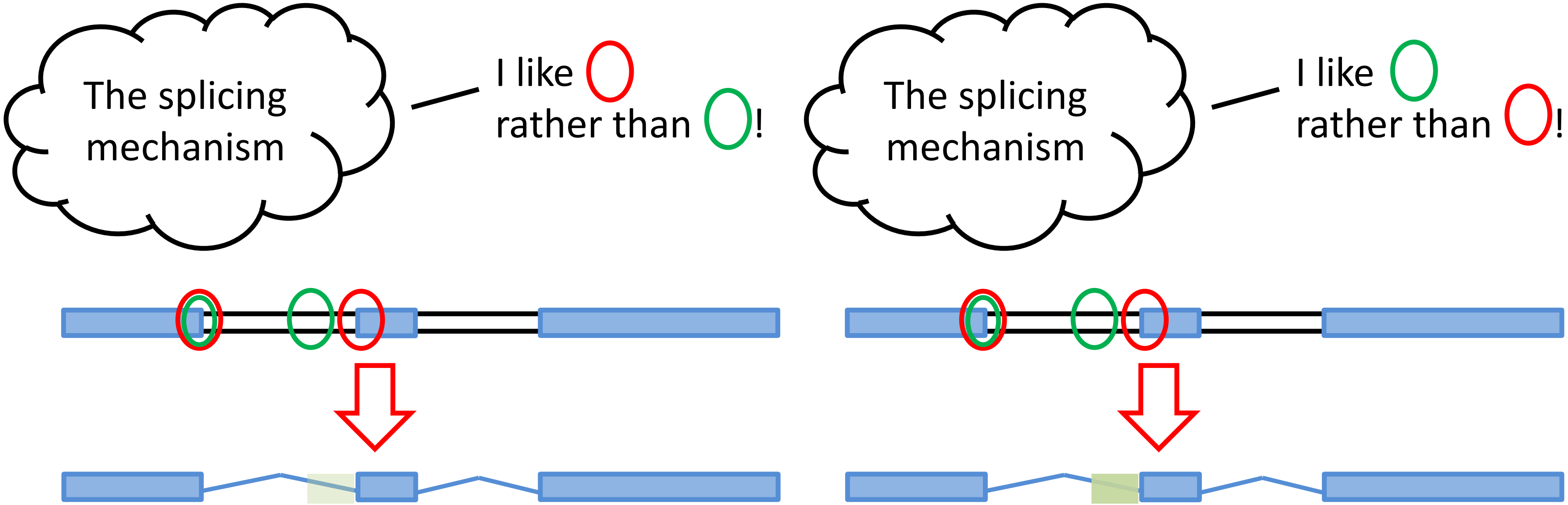
pre-mRNA

less splicing

Alternative intron-retention

Theories of event-based algorithms

- Taking *preference* of the splicing mechanism into consideration.
  - another example on alternative accepter



The splicing mechanism

I like ○ rather than ○!

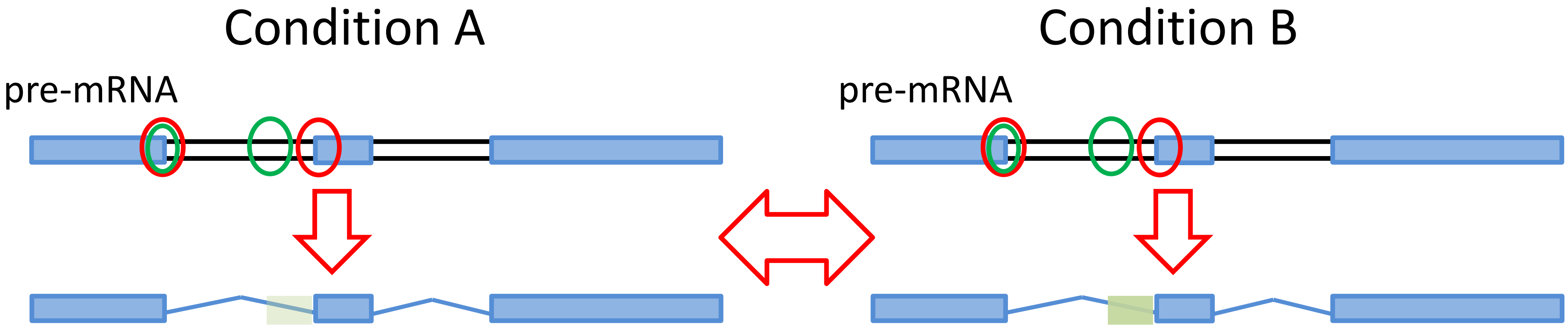
The splicing mechanism

I like ○ rather than ○!

Preference of ○ (or ○) changed

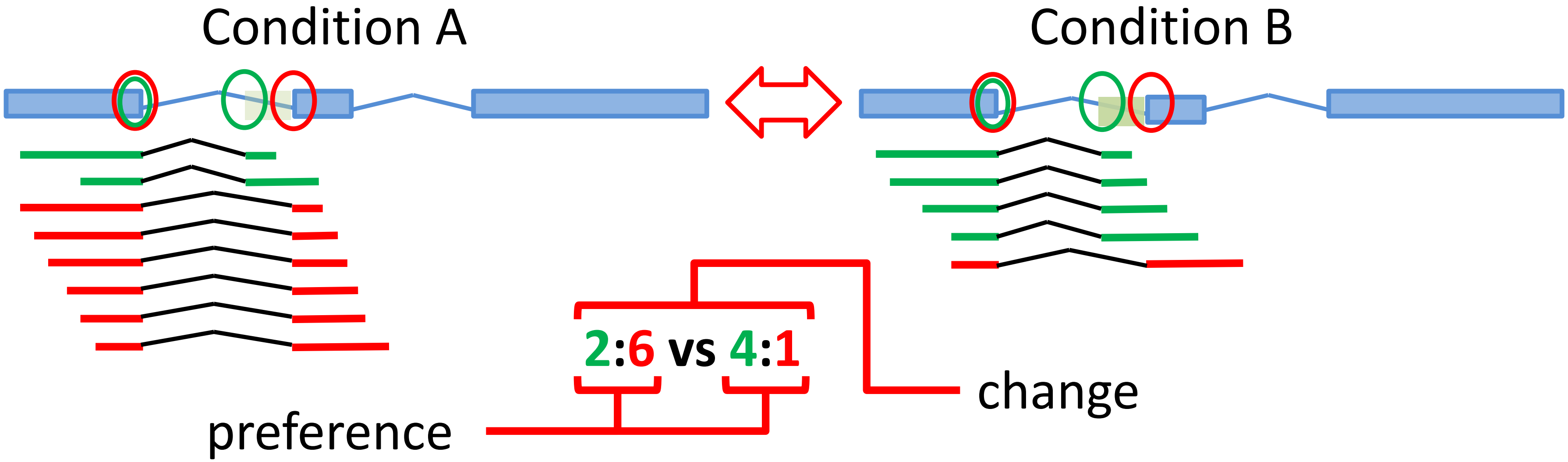
Theories of event-based algorithms

- Revisit the term “alternative”
  - change of splicing preference between two conditions
- The term “preference” means
  - the possibility of choosing something against some *background*.



Theories of event-based algorithms

- Take alternative donor/acceptor events as an example
  - The *preference* can be somehow measured by read counts
  - The *change of preference* can be measured by some statistical tests



Theories of event-based algorithms

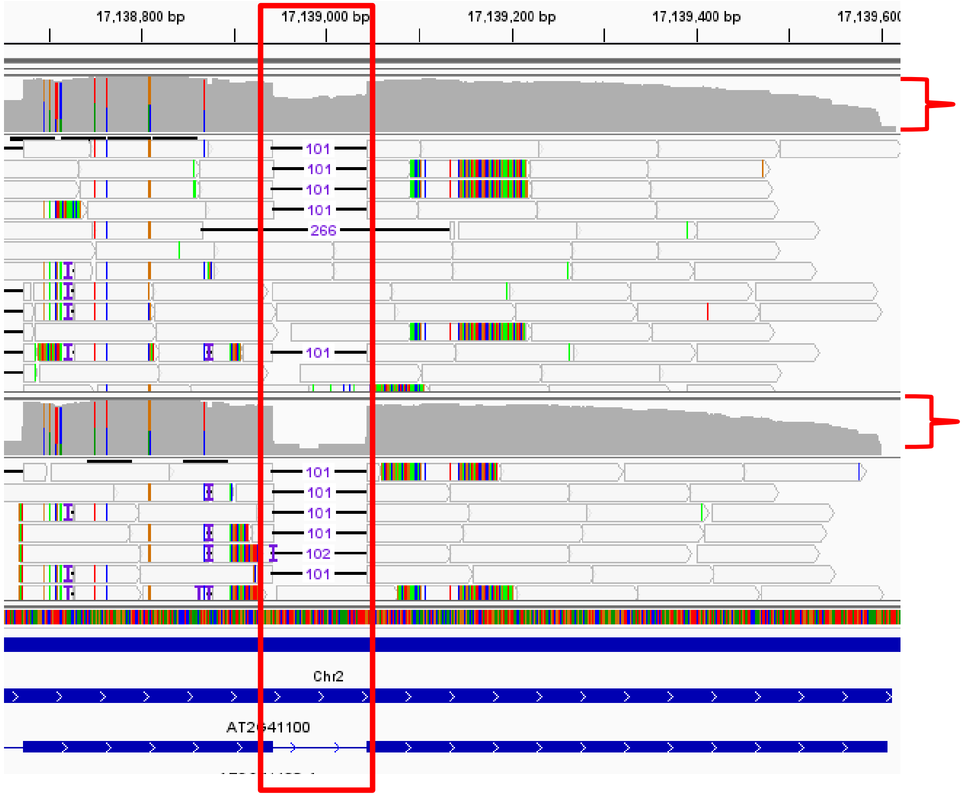
- In next slides
  - We show cases of alternative splicing comparisons of the example data
  - with visualization and explanation

Theories of event-based algorithms

- Alternative intron-retention

#GeneID	intronNo	intronLen	intronC	intronT	exonC	exonT	chiSquared	P-value
AT2G41100	3	101	28.1	1.85	205.7	148.4	15.7	0.00007

intron 3



log-scaled coverage

control

log-scaled coverage

treatment

Theories of event-based algorithms

- Alternative intron-retention

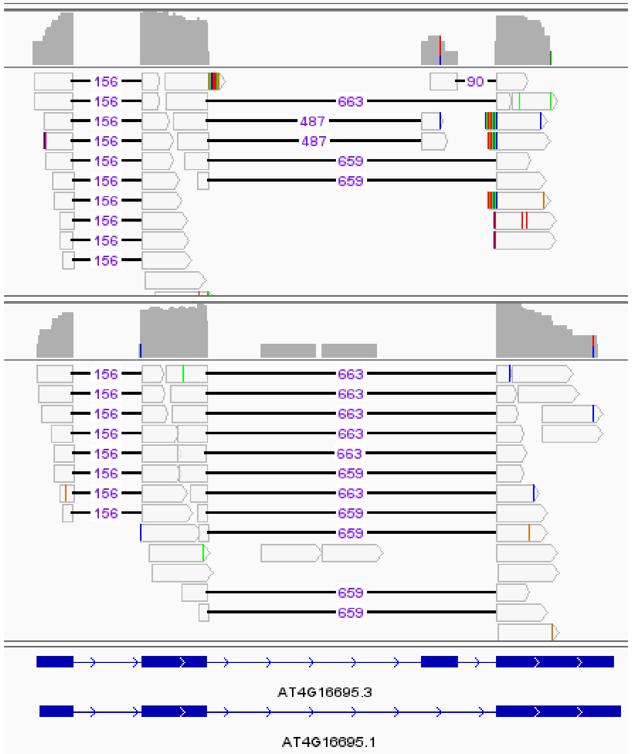
#GeneID	intronNo	intronLen	intronC	intronT	exonC	exonT	chiSquared	P-value
AT2G41100	3	101	28.1	1.85	205.7	148.4	15.7	0.00007

  - We computed read depths of an intron region (28.1 & 1.85) and took read depths of neighboring exons (205.7 & 148.4) as the background
  - Chi-squared test of *goodness of fit* was used to see if intron read depths are following the background
  - In English, to see if the chance of retaining the intron was changed between the two conditions.

Theories of event-based algorithms

- Alternative exon-skipping

#GeneID	exonPair	control	treatment	xControl	xTreatment	xChiSquared	P-value
AT4G16695	2<=>4	3	11	3	0	10.45249	0.001225



Theories of event-based algorithms

- Alternative exon-skipping

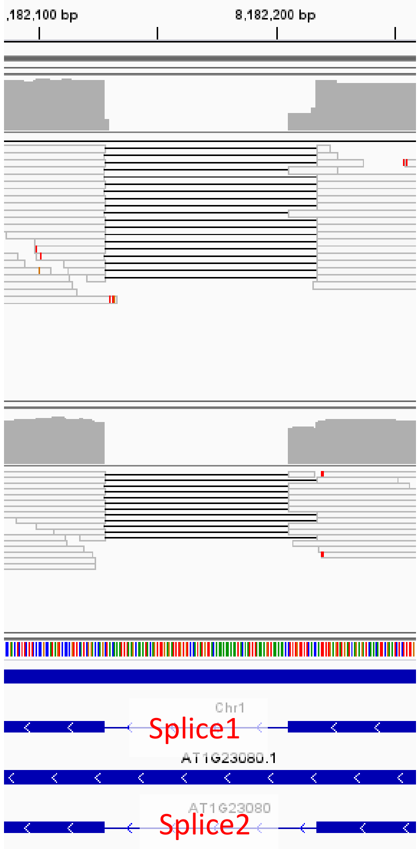
#GeneID	exonPair	control	treatment	xControl	xTreatment	xChiSquared	P-value
AT4G16695	2<=>4	3	11	3	0	10.45249	0.001225

  - We counted reads that are supporting the exon-skipping event (3 & 11) and reads not supporting the event (3 & 0)
  - Chi-squared test of *goodness of fit* was used to see if any of the two sets of numbers are not following the other
  - In English, to see if the chance of skipping (or not skipping) an exon was changed between the two conditions.

Theories of event-based algorithms

- Alternative donor/accepter change

#Genec	Splice1	Splice2	Ctr Splice1	Trt Splice1	Ctr SpliceO	Trt SpliceO	p-value
AT1G23080	2(0)-3(0)	2(0)-3(-12)	2	8	17	4	0.002004



Theories of event-based algorithms

- Alternative donor/accepter change

#Genec	Splice1	Splice2	Ctr Splice1	Trt Splice1	Ctr SpliceO	Trt SpliceO	p-value
AT1G23080	2(0)-3(0)	2(0)-3(-12)	2	8	17	4	0.002004

  - We counted reads that are supporting junction *splice1* “2(0)-3(0)”(2 & 8) and splice reads from the same exon pairs but not supporting *splice1* (17 & 4)
  - Fisher exact test was used to see if any of the two sets of numbers are not following the other
  - In English, to see if the chance of picking *splice1* as the splicing junction was changed between the two conditions.

Theories of event-based algorithms

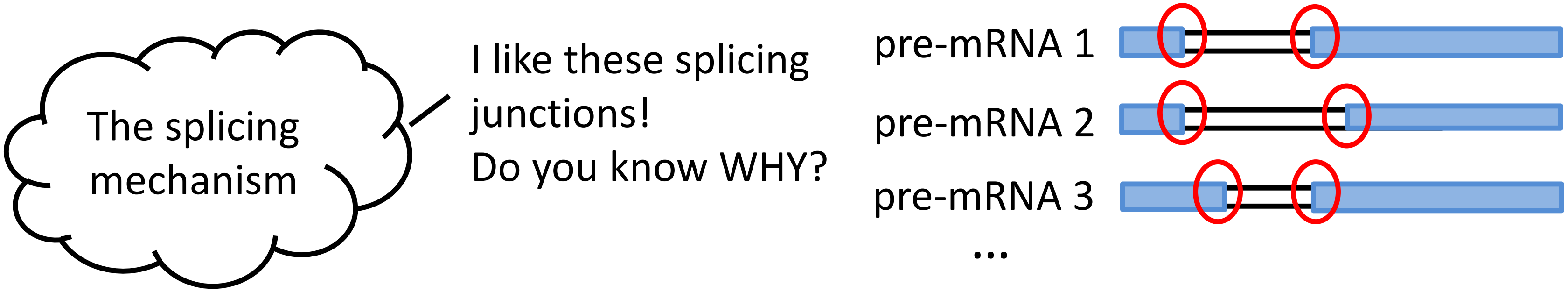
- A short note
  - For the three types of AS comparisons
    - Intron retention
    - Exon skipping
    - Alternative donor/accepter
  - The applied statistical tests hold *the same null hypothesis*
    - the preference of the splicing event is the same between the two conditions
    - A literal interpretation on a significant P-value: it is *unlikely* the preference is the same between the two conditions

Theories of event-based algorithms

- Short conclusions
  - Event-based algorithms, at least as we presented, take RNAseq evidences *directly* for statistical comparisons
  - The presented event-based methods take the preference of the splicing mechanism into consideration
  - Our recent development also enables comparisons between sample groups
    - A choice of not merging biological replicates and taking replication into consideration

AS-related motif discovery

- Considering that we have a list of splicing junctions that were differentially preferred between two conditions (alternatively spliced)
- An interesting topic would be to discover the rationale why the splicing mechanism has different preference on these splicing sites.



The splicing mechanism

I like these splicing junctions!
Do you know WHY?

pre-mRNA 1

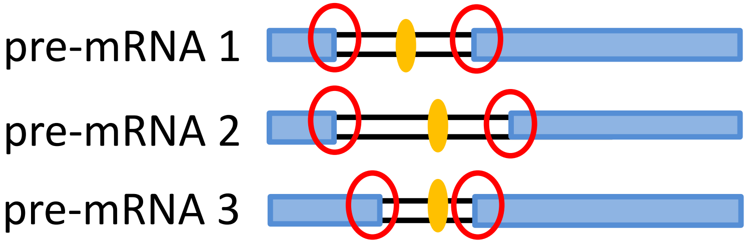
pre-mRNA 2

pre-mRNA 3

…

AS-related motif discovery

- A way to study this question is to find *cis* elements nearby these splicing sites.
- Applying motif database searches or *de novo* motif discovery on regions around these splicing sites may help
  - – and we can do better



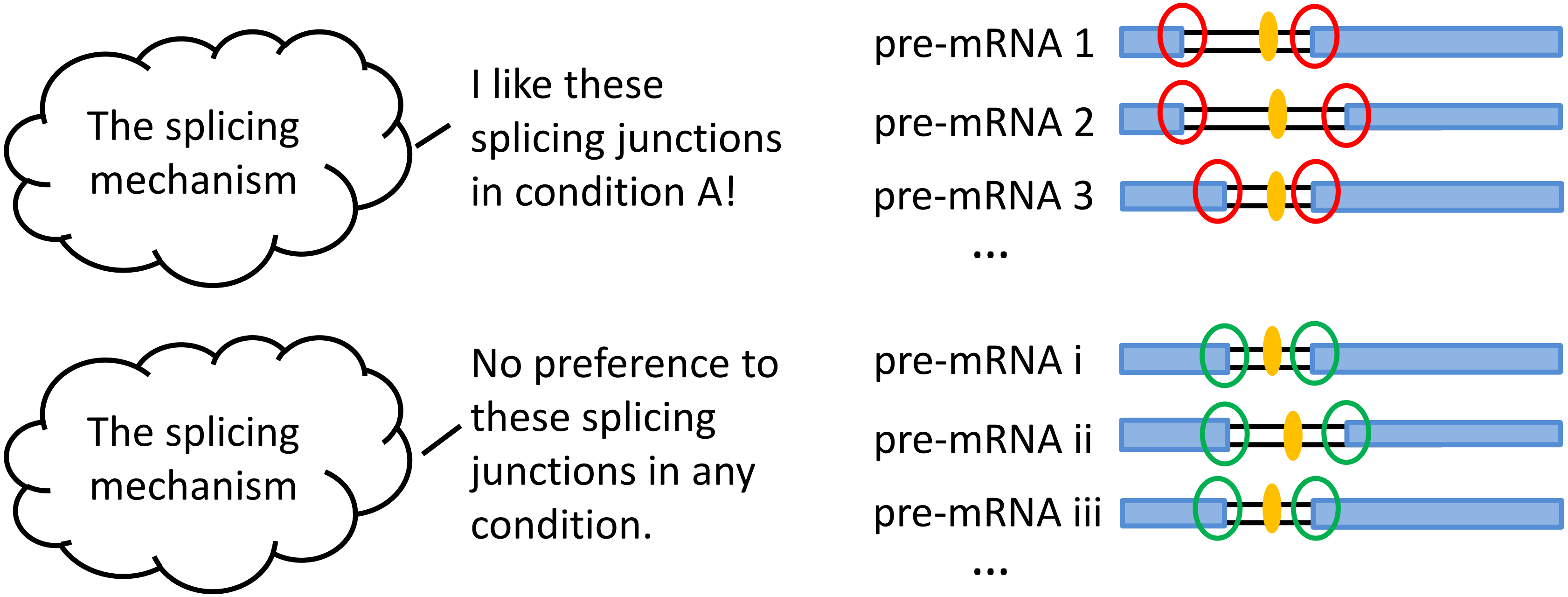
Some motif there?

AS-related motif discovery

- Considering that *de novo* motif discovery is actually a multiple-sequence local alignment problem
- Existing methods from the very first Gibbs sampling to currently popular tools like MEME are actually *heuristics*
  - and tend to report motifs whose appearance numbers are higher than *expected*.

AS-related motif discovery

- If a motif was reported simply because it appears every where, it may not be the key to the difference we are looking for.



The splicing mechanism

I like these splicing junctions in condition A!

pre-mRNA 1

pre-mRNA 2

pre-mRNA 3

…

The splicing mechanism

No preference to these splicing junctions in any condition.

pre-mRNA i

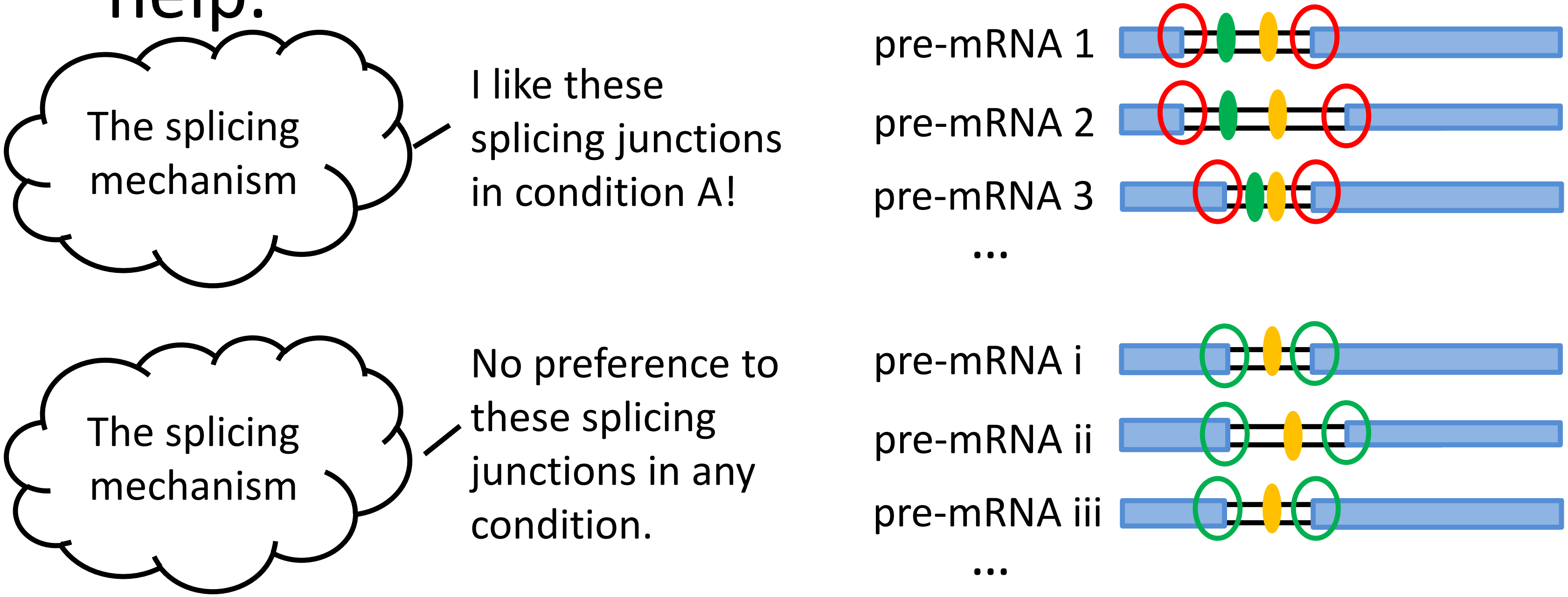
pre-mRNA ii

pre-mRNA iii

…

AS-related motif discovery

- Once we have a motif candidate, comparing its number of appearances in target regions against an appropriate *background* would help.



The splicing mechanism

I like these splicing junctions in condition A!

pre-mRNA 1

pre-mRNA 2

pre-mRNA 3

…

The splicing mechanism

No preference to these splicing junctions in any condition.

pre-mRNA i

pre-mRNA ii

pre-mRNA iii

…

AS-related motif discovery

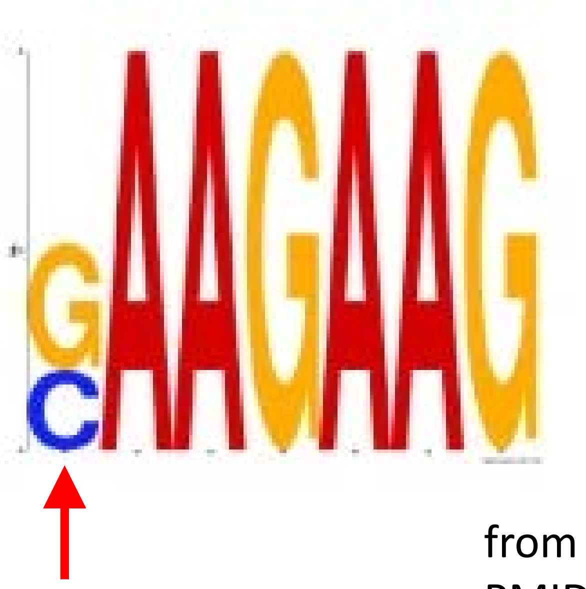
- Another question: what do we mean by “appearances in target regions”?
- Motif discovery tools and motif database used to give a position weight matrix
  - Similarity between a sequence of *hit* and a motif is usually measured by P-values
    - Any appropriate P-value threshold to define appearances?
    - We always need definition for computation.

AS-related motif discovery

- Considering an extreme case
  - Assuming uniform random background of {A,C,G,T} in sequences
  - An exact match to motif “ACGT” means
    - P-value = 4^-4
  - An Exact match to motif “ACGTACGT” means
    - P-value = 4^-8
  - => the same sequence matching identity but P-value decided by motif length

AS-related motif discovery

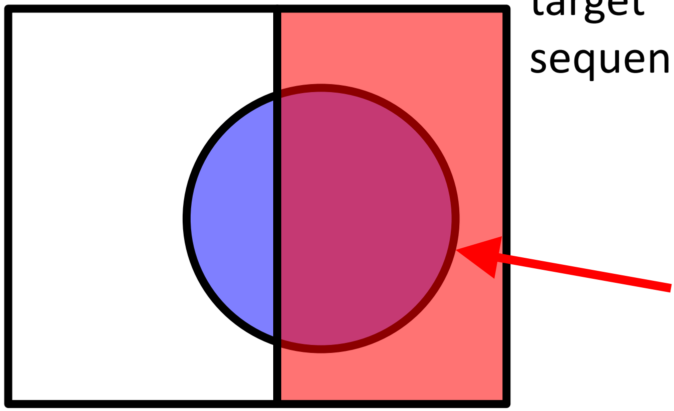
- Also considering that a PWM used to have more than one nucleotide(protein) at on position, the way to decide an appropriate P-value threshold would be complicated.



from Wu *et al*. Genome Biology 2014.
PMID: 24398233

AS-related motif discovery

- Think this question conversely
  - If there is an appropriate P-value threshold for a motif that is actually related with our study
  - With this appropriate P-value threshold, thus defined appearances should show a preference to our target sequences



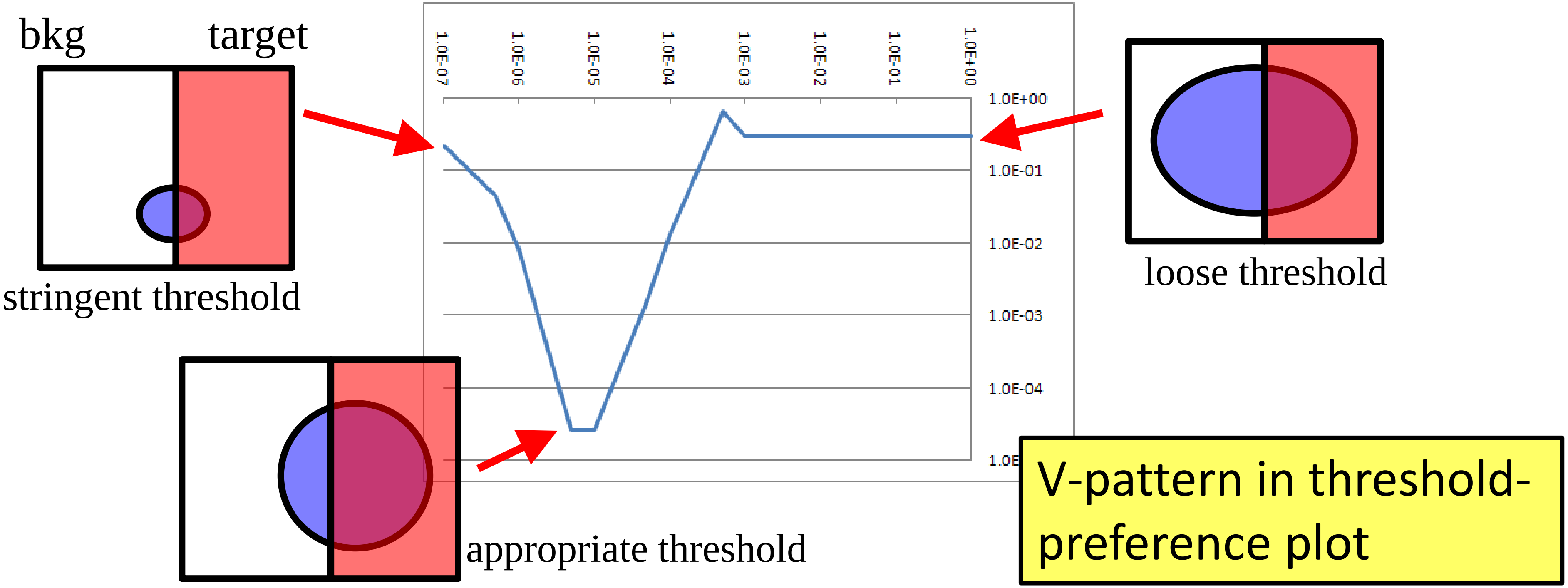
background sequences

target sequences

Sequences hit by the motif under some threshold

AS-related motif discovery

- So the first step is to use a motif discovery tool to report a certain number of motifs (ex: 30)
- For each motif, we examine its *preference* of appearances (target against background) under *various thresholds*



bkg

target

stringent threshold

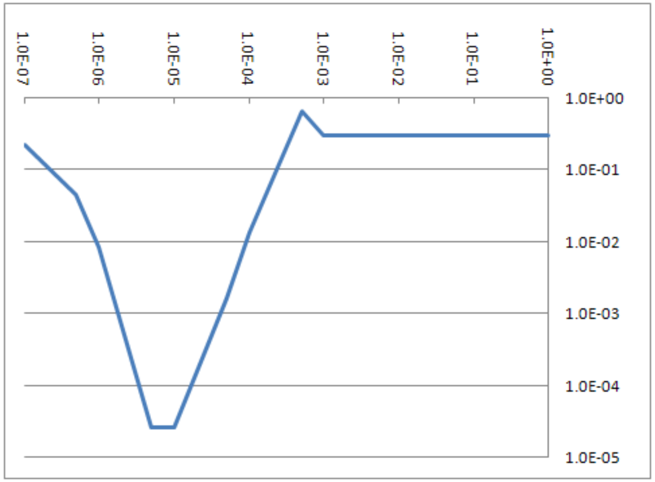
loose threshold

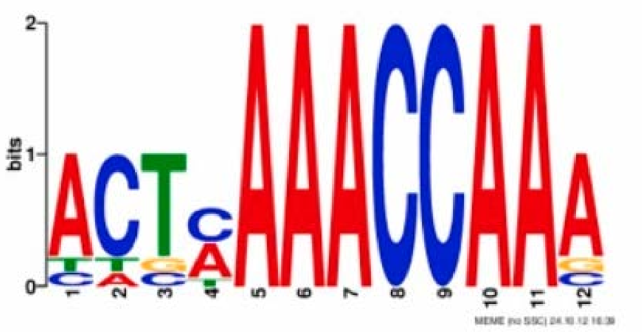
appropriate threshold

V-pattern in threshold-preference plot

AS-related motif discovery

- Above described idea can be applied to not only AS-related motifs but also promoter motifs
  - This threshold-preference plot is corresponding to a 21th motif with E-value 14000 by MEME





from Rodríguez-Celma *et al*. Plant Physiology 2013. PMID: 23735511

AS-related motif discovery

- In the rackj package, we have a set of programs dealing with AS- and promoter-related sequence extraction and MEME/MAST output.
  - They were exactly designed for the aforementioned motif discovery strategy.
  - http://rackj.sourceforge.net/SpecialScripts/index.html

Discussions

- Isoform-based algorithms vs event-based algorithms, which kind of method to use?
  - This depends on your research purpose
    - Isoform-based algorithms predicts expression levels of transcripts
      - Overall results of splicing events per gene
    - Event-based algorithms should report changes that focus on splicing events
    - There should be no problem to do both of them at the same time
      - Always study the results carefully

Discussions

- Can we incorporate technologies like nanopore or PacBio in alternative splicing analyses?
  - The key should be the quality of results.
  - *Currently*, sequencing *error rates* of nanopore & PacBio were considered higher than that of Illumina
  - This may affect fitting of mapping records to exon boundaries
    - => alternative donor/accepter detection, and may be small exons

Discussions

- Which *background* should I choose for the described motif discovery?
  - Choose different backgrounds may result in different answers.
  - Take promoter motif discovery as an example, given differentially expressed genes as the *target*
    - Choose non-expressed genes as the background
      - The difference could be *expressed or not*.
    - Choose expressed genes not in DEGs as the background
      - The difference could be differential expression or not.

Finally

- Thank you for your attentions.
- I am willing to answer and/or discuss questions via email or in some other interactive form.
  - Please don’t hesitate to let me know if you have any questions.