StringTie

Transcript assembly and quantification for RNA-Seq

Johns Hopkins University
Center for Computational Biology
CCB

Home Manual FAQ

CCB » Software » StringTie

StringTie1+2
RNA-seq assembly methods

Chanhee Lee

Img : https://ccb.jhu.edu/software/stringtie/

Resources that you might want to revist after class

Publications

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M **Transcriptome assembly from long-read RNA-seq alignments with StringTie2**, *Genome Biology* 20, 278 (2019), doi:10.1186/s13059-019-1910-1

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL **Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown**, *Nature Protocols* 11, 1650-1667 (2016), doi:10.1038/nprot.2016.095

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT & Salzberg SL. **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads** *Nature Biotechnology* 2015, doi:10.1038/nbt.3122

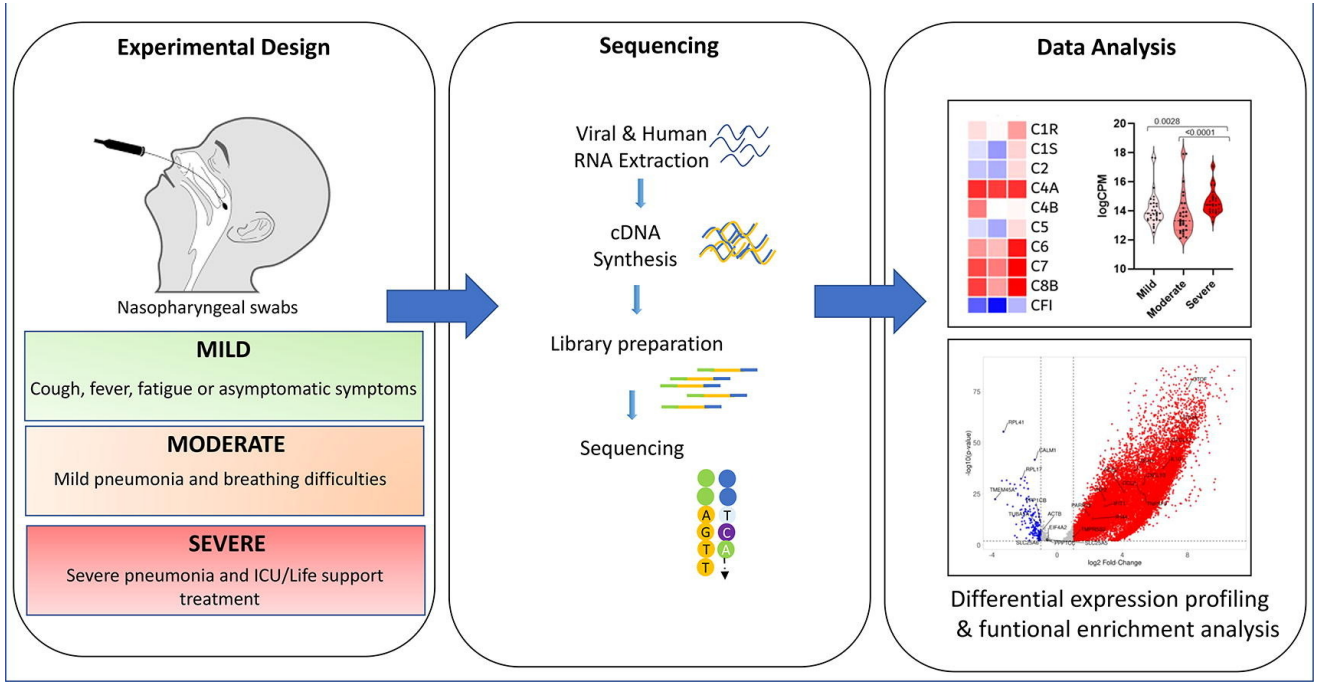
Youtube videp from Salzberg : https://www.youtube.com/watch?v=2qGiw4MRK3c&t=2501s

Img : https://ccb.jhu.edu/software/stringtie/

What you will learn after the presentation

- Why do we perfrom RNA-seq analysis?
- What are the challenges in RNA-seq analyis?
- What main idea did StringTie use to solve the problems?
- How much better is StringTie compared to other methods?

Why do we perform RNA-seq analysis?

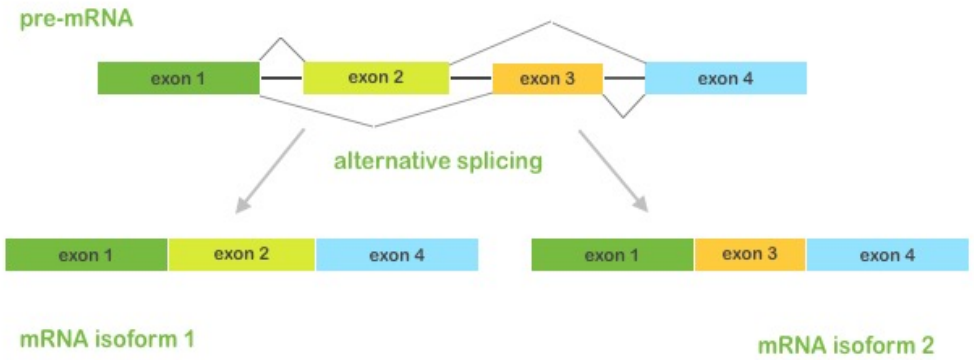


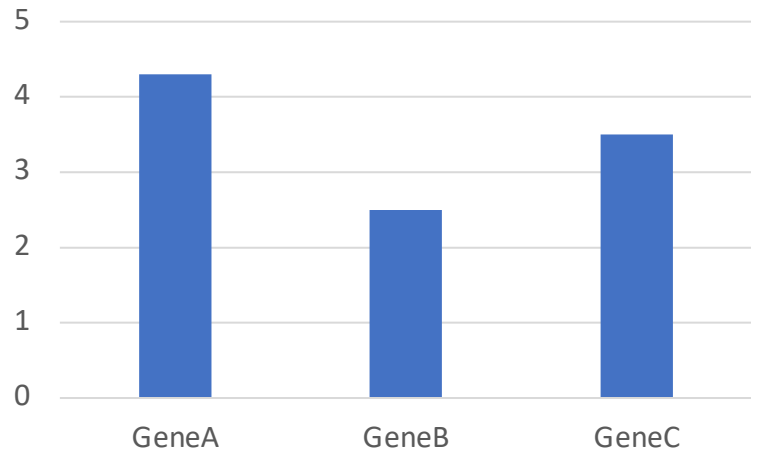
- For DNA-seq analysis, we are usually interested in variant! (GATK pipeline)
- DNA doesn't change much => SNP, INDEL, PRS, GWAS …

- For RNA-seq analysis, what are we interested in?
- Transcription varies significantly across different conditions!
- Compare gene(transcripts) expresseion between case vs control => DEG!

https://www.frontiersin.org/files/Articles/341949/fcell-06-00028-HTML/image_m/fcell-06-00028-g002.jpg

Why do we perform RNA-seq analysis?





1) Who are being expressed?
Identification

+

2) How much are they being expressed?
Quantification

=> Comparison between conditions!

=> What caused that difference?

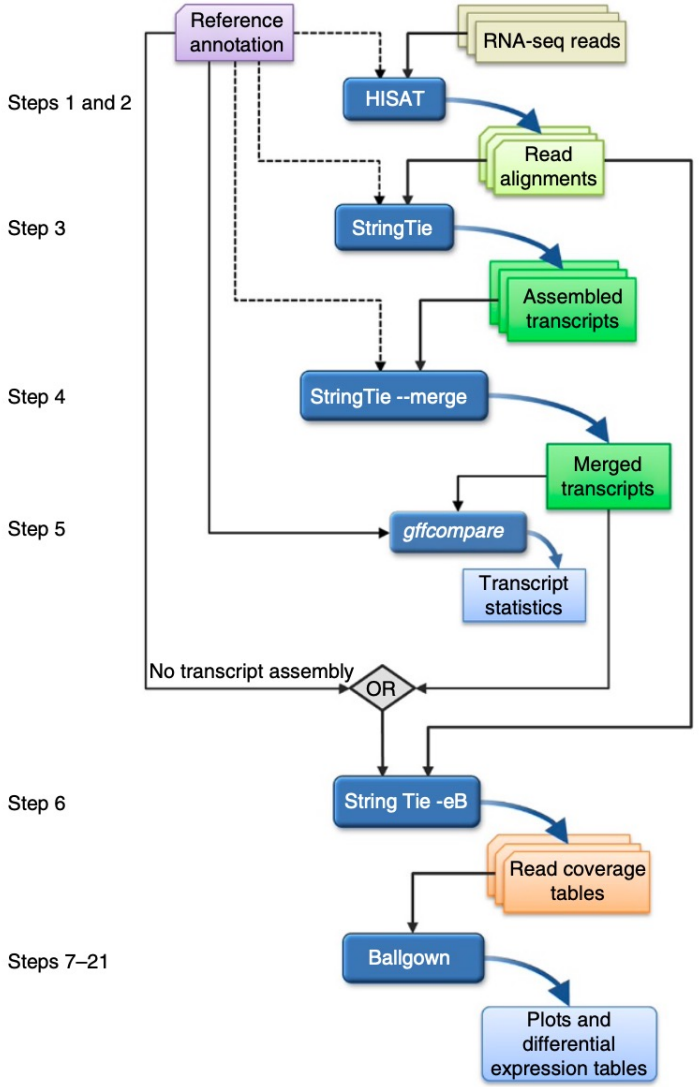
=> How does it affect translation?

⋮

https://genestack.com/blog/2015/04/02/testing-differential-isoform-expression-on-genestack-platform/

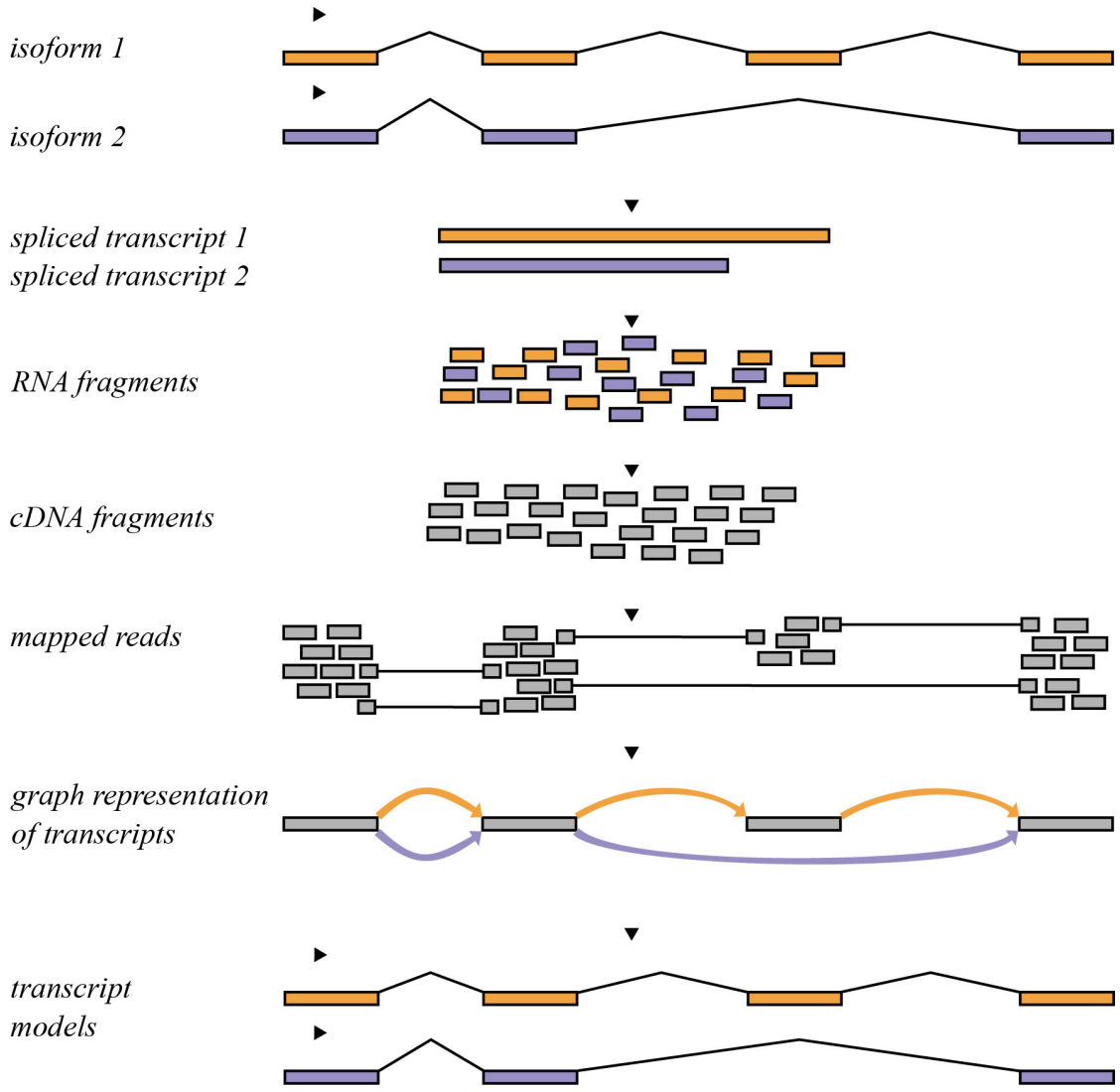
RNA-seq analysis pipeline : new tuxedo protocol

1. Alignment of the reads to the genome; (HISAT)
2. Assembly of the alignments into full-length transcripts; (StringTie)
3. Quantification of the expression levels of each gene and transcript; (StringTie)
4. Calculation of the differences in expression for all genes among the different experimental conditions; (Ballgown)



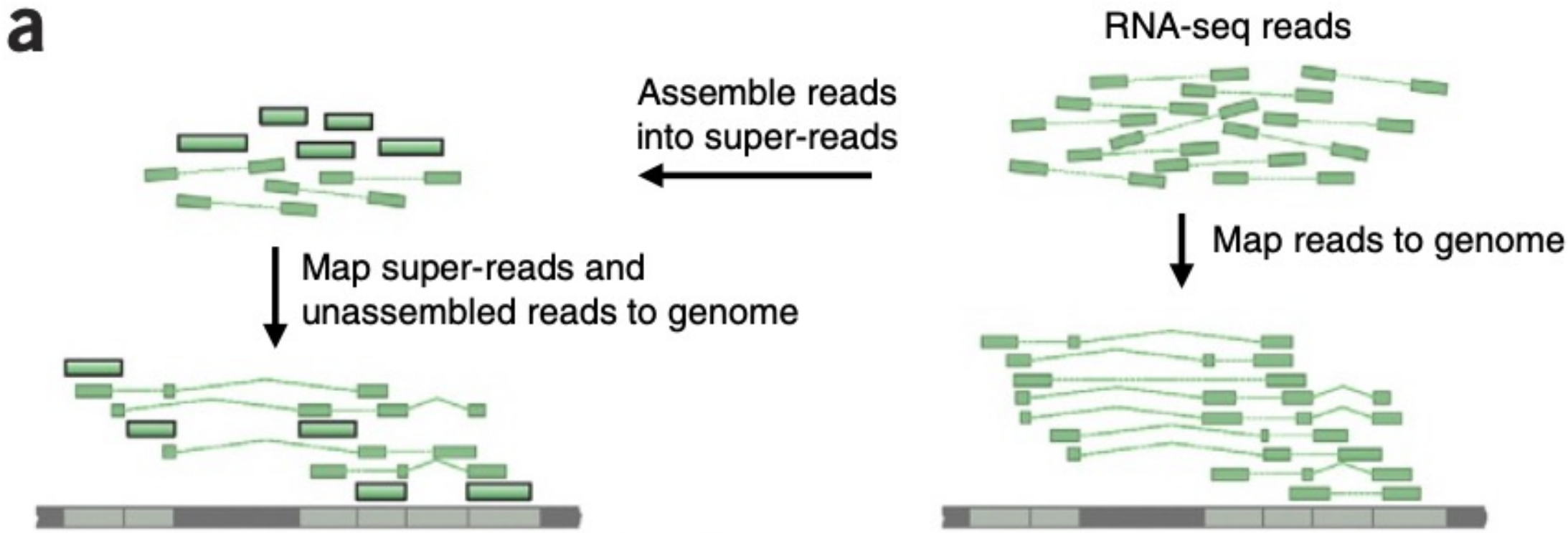
RNA-seq analysis pipeline : new tuxedo protocol

1. Alignment of the reads to the genome; (HISAT)
2. Assembly of the alignments into full-length transcripts; (StringTie)
3. Quantification of the expression levels of each gene and transcript; (StringTie)
4. Calculation of the differences in expression for all genes among the different experimental conditions; (Ballgown)



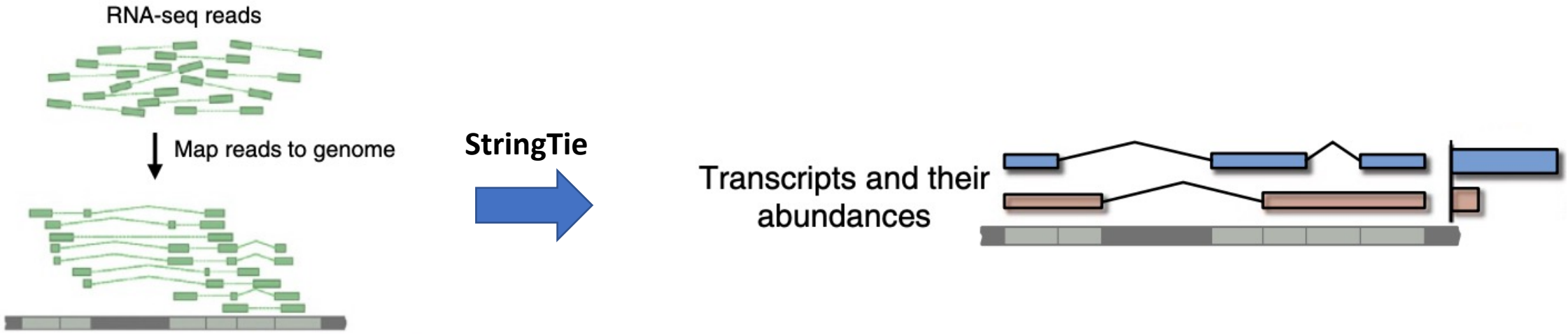
https://open.oregonstate.education/appliedbioinformatics/chapter/chapter-6/

RNA-seq analysis pipeline : new tuxedo protocol



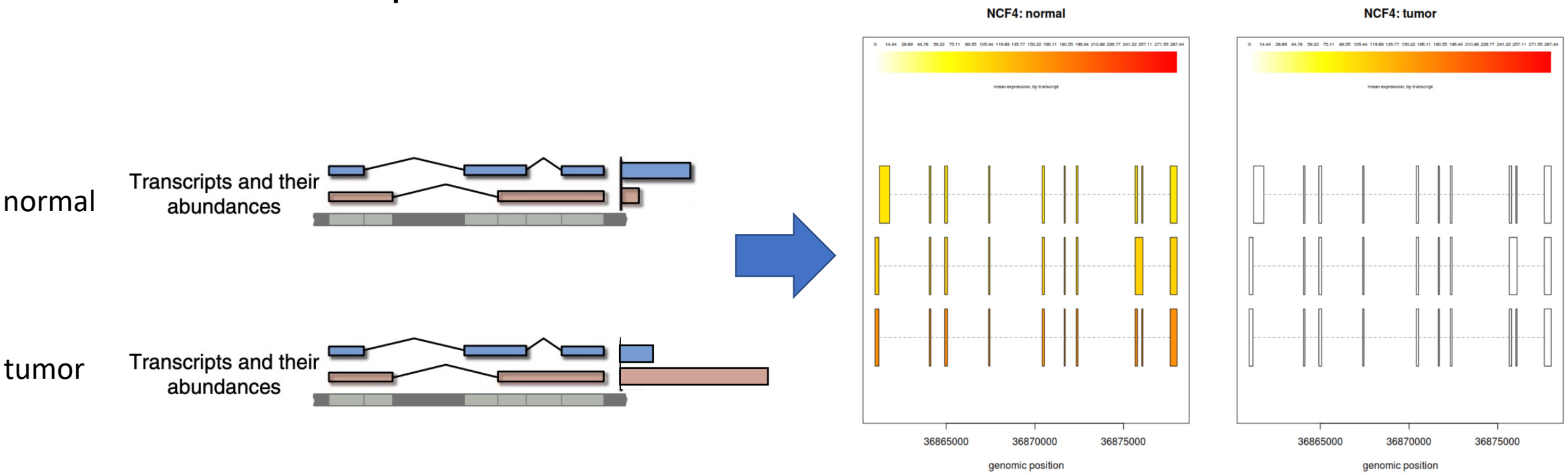
1. Alignment of the reads to the genome; (HISAT)

RNA-seq analysis pipeline
: new tuxedo protocol



2. Assembly of the alignments into full-length transcripts; (StringTie)
3. Quantification of the expression levels of each gene and transcript; (StringTie)

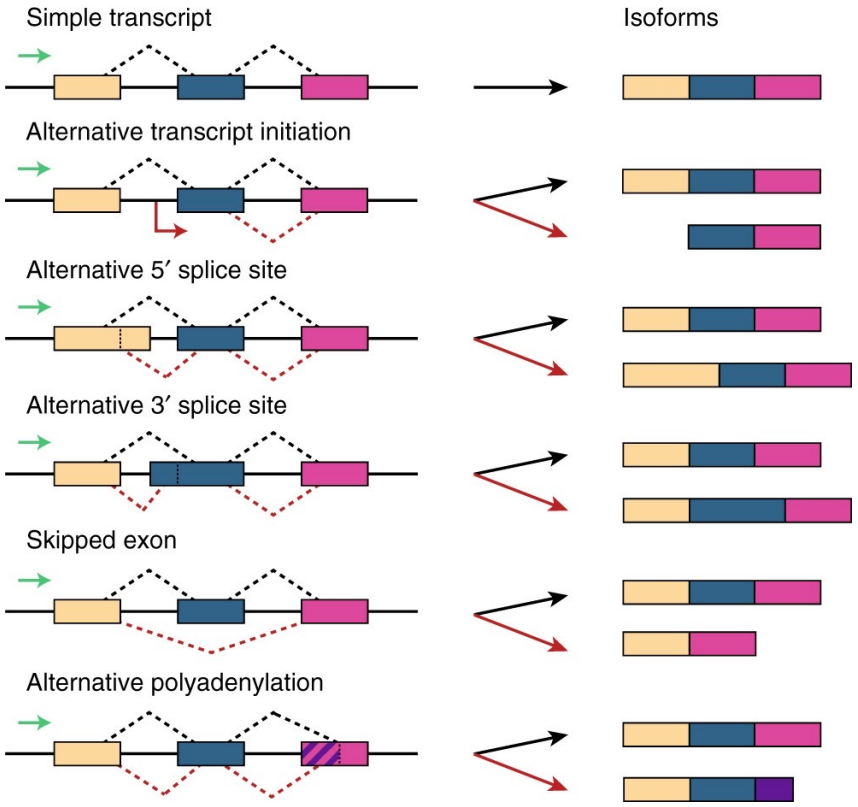
RNA-seq analysis pipeline : new tuxedo protocol



4. Calculation of the differences in expression for all genes among the different experimental conditions; (Ballgown)

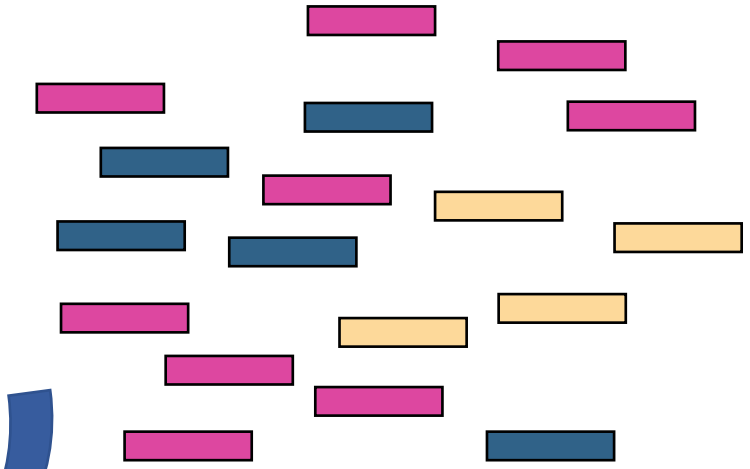
https://blog.daum.net/kimuks/7541198

What are the challenges in RNA-seq analysis?



A single gene may be transcribed into several distinct mRNA variants called isoforms

200 million reads (100-150bp)



Going back is hard…

1. High variable sequence coverage for different transcripts
2. Alternative transcripts from the same locus can share exons
3. Determining the quantities is also challenging,
   even if we assume that the transcript structures are known.

Aguiar, D., Cheng, LF., Dumitrascu, B. et al. Bayesian nonparametric discovery of isoforms and individual specific quantification. Nat Commun 9, 1681 (2018

What are the challenges in RNA-seq analysis?

transcript identification problem + expression quantification problem

How to solve the challenges : Previous solution

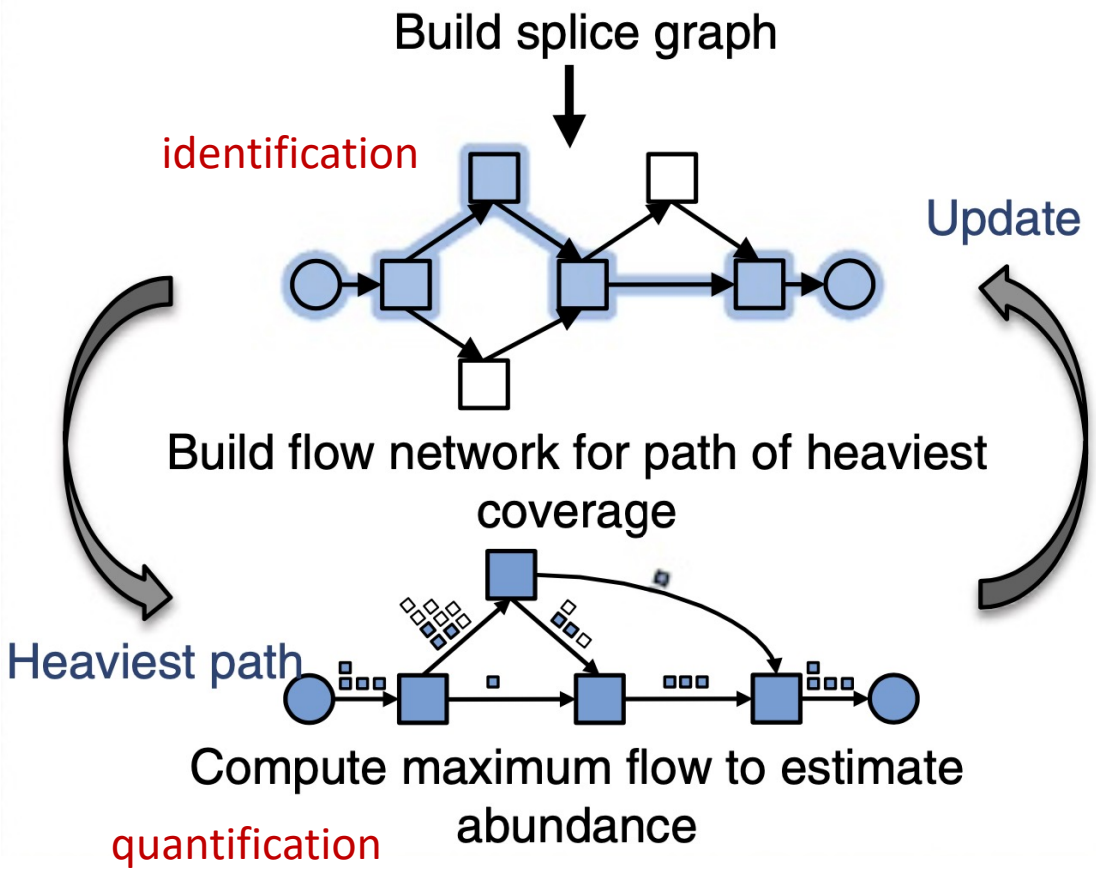
- Transcript identification problem : Trinity, Oases
- Expression quantification problem : RSEM, eXpress
- Solve both : IsoInfer, Scripture, **Cufflinks**, SLIDE, IsoLasso, iReckon, Traph

Still, much work remains to produce consistent, highly accurate solutions

How to solve the challenges : StringTie

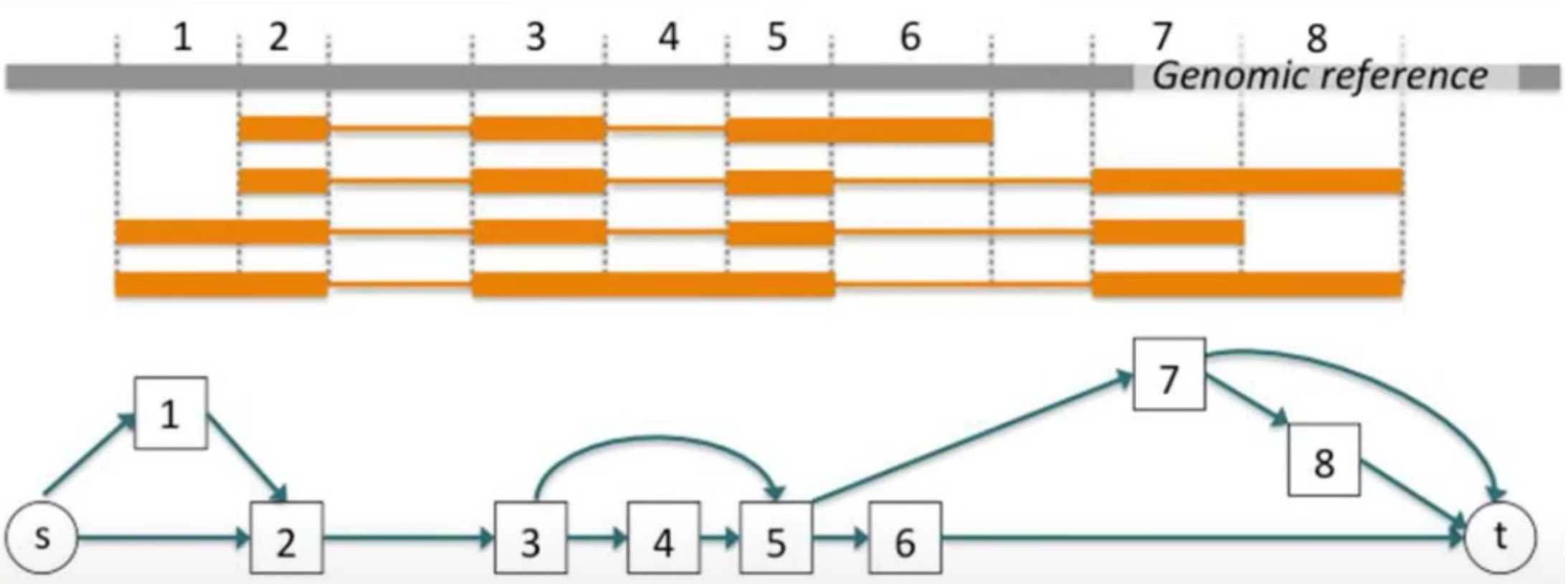
- StringTie uses a genome-guided transcriptome assembly approach along with concepts from *de novo* genome assembly to improve transcript assembly.
- StringTie is RNA-seq assembly method that tackles identification and quantification simultaneously.
- StringTie correctly identified 36–60% more transcripts than the next best assembler (Cufflinks) on multiple real and simulated data sets!!

Main idea used in StringTie



- StringTie assembles transcripts and estimates their expression levels simultaneously.

1. StringTie first groups the reads into clusters,(by gene locus)
2. creates a **splice graph** from which it identifies transcripts,
3. creates a separate **flow network** to estimate its expression level using a maximum flow algorithm.



After alignment… we have reads mapped to the reference genome

- Splice variant1 : 2=>3=>5=>6
- Splice variant2 : 2=>3=>5=>7=>8
- Splice variant3 : 1=>2=>3=>5=>7

Using Splice Graph, we can represent all the combinations of different isoforms.(s=>t)

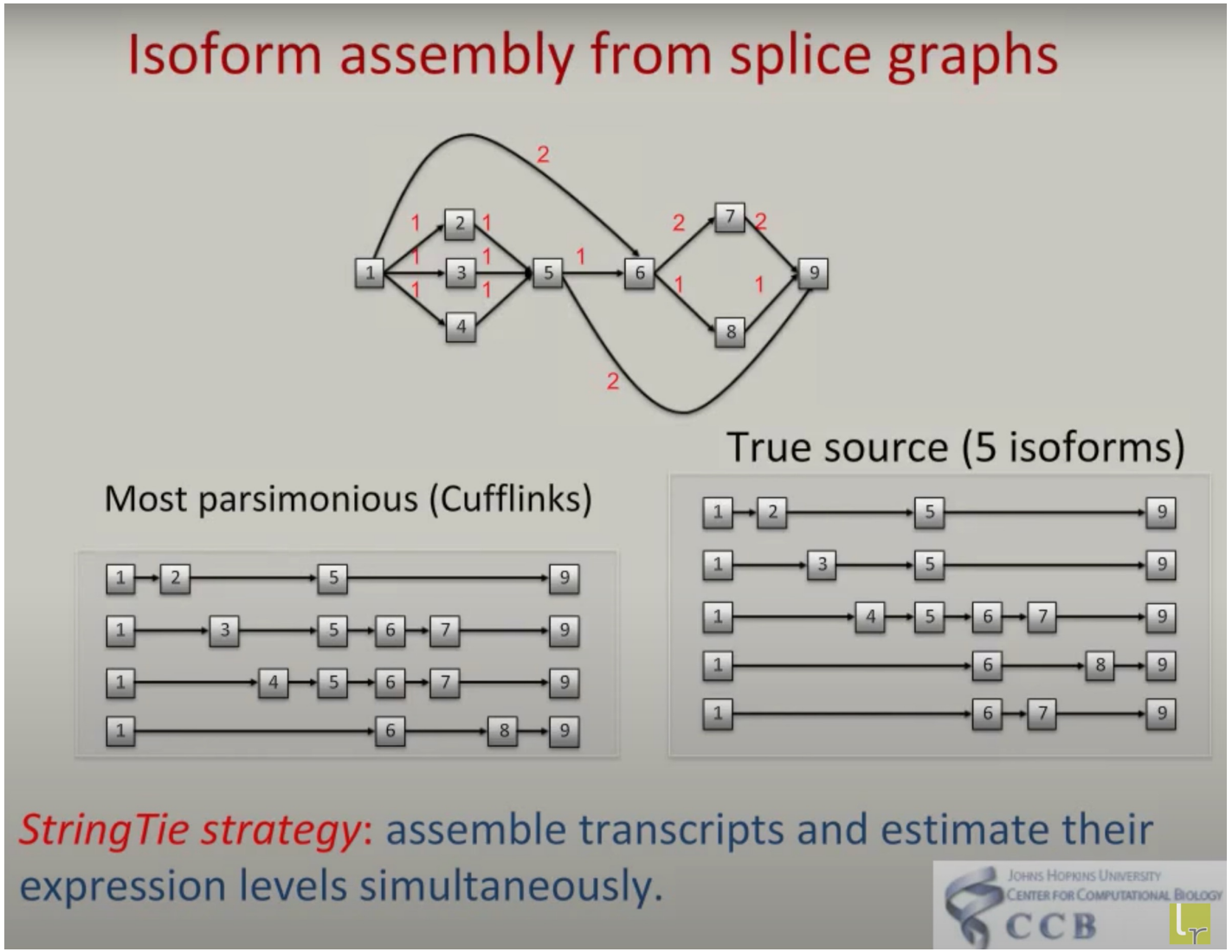
* Node: exon or part of exon
* Edge : possible path

StringTie Algo : https://www.youtube.com/watch?v=2qGiw4MRK3c

From the splice graph, we must figure out which isoforms the reads were from - Identification problem

Cufflinks: Most parsimonious methods

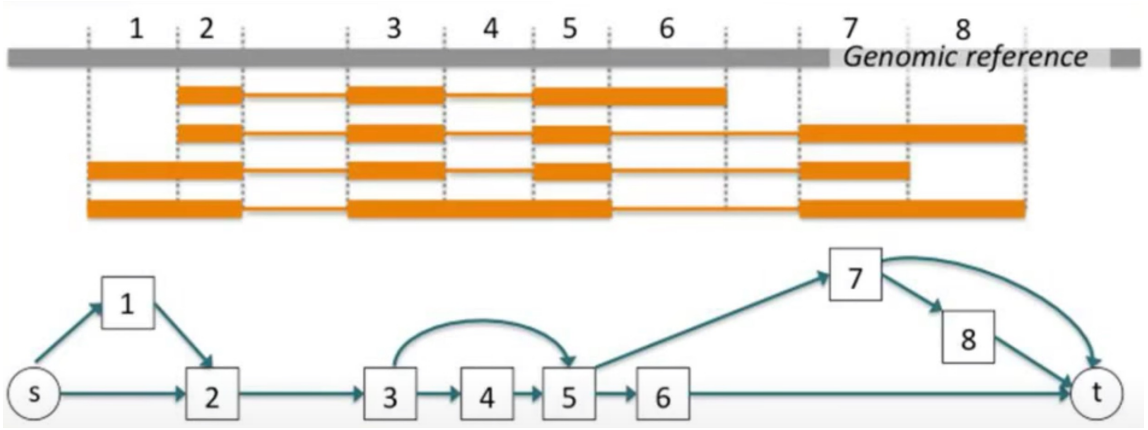
- generates the minimal number of transcripts that will explain all reads in the graph
- Can't explain right example.
- **Does not account for expression level in identification of isoforms**



StringTie assembles transcripts and estimates their expression levels simultaneously.

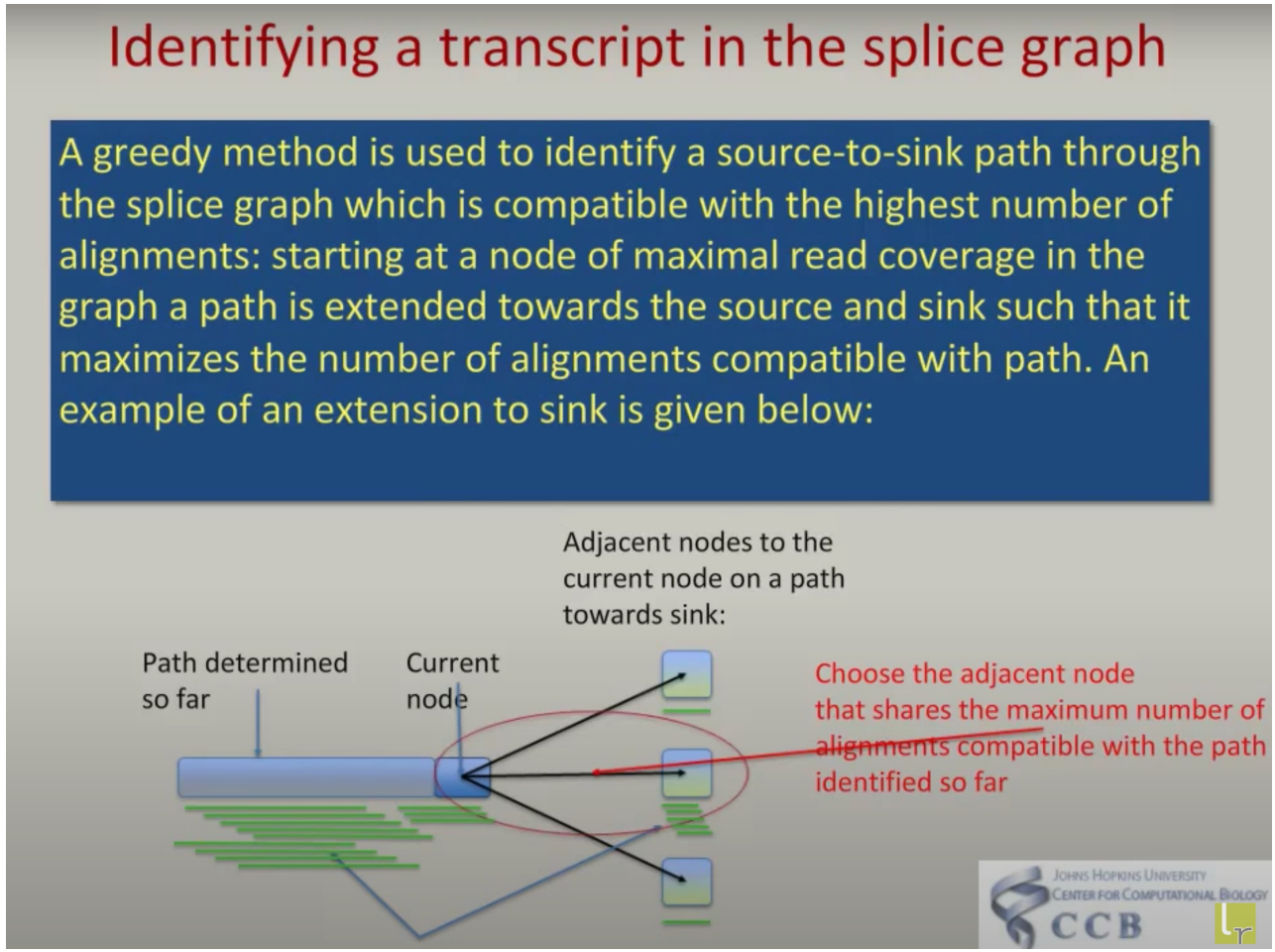
StringTie Algo : https://www.youtube.com/watch?v=2qGiw4MRK3c

Main idea used in StringTie (Splice Graph)



Heaviest path : 2=>3=>5=>7=>8

=> **Isoform is identified, using coverage information**

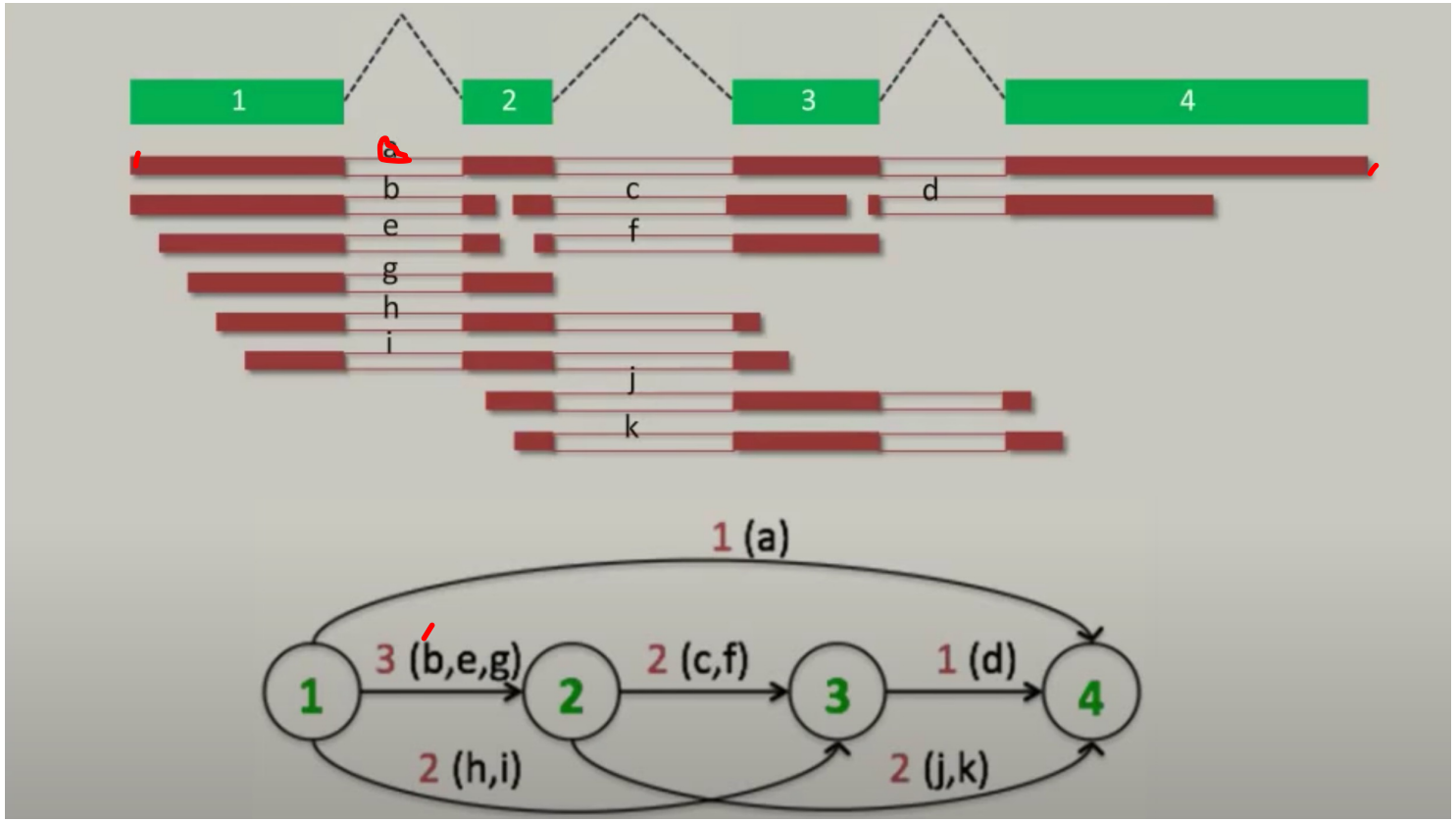


Splice Graph : https://www.coursera.org/lecture/algorithms-part2/introduction-to-maxflow-jZVUm

Main idea used in StringTie (Maxflow algorithm)

Maxflow Network(Quantification)

- Assume 1=>2=>3=>4 is the isoform we found using heaviest path
- Edge is generated at where reads start and ends, +1 to the edge (first read : starts at 1, ends at 4)



StringTie Algo : https://www.youtube.com/watch?v=2qGiw4MRK3c

Main idea used in StringTie (Maxflow algorithm)

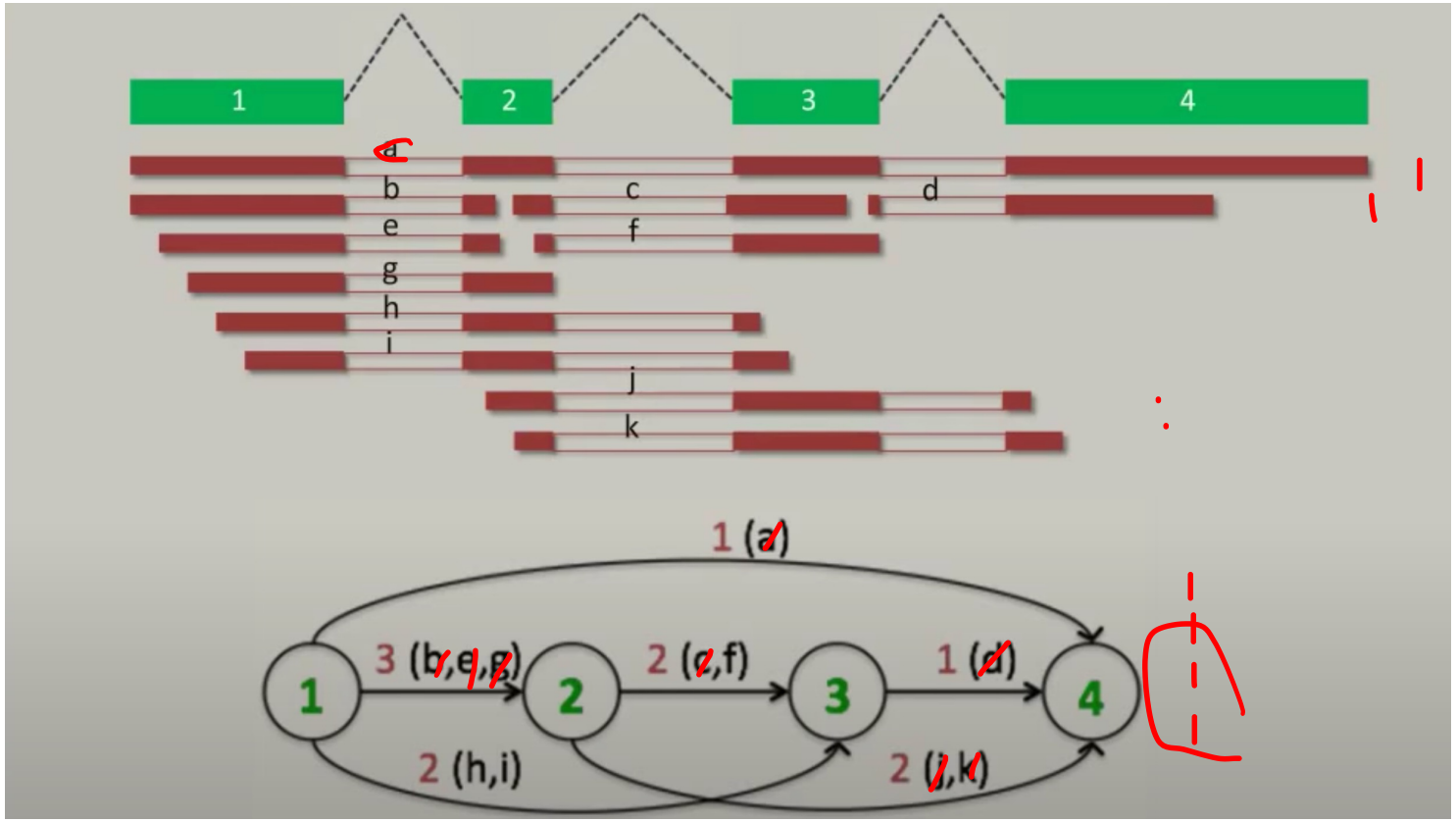
Maxflow algorithm(Quantification)

- What is the maximum number of water that can flow through the pipe

- What is the maximum number of the isoform that can be built from the reads

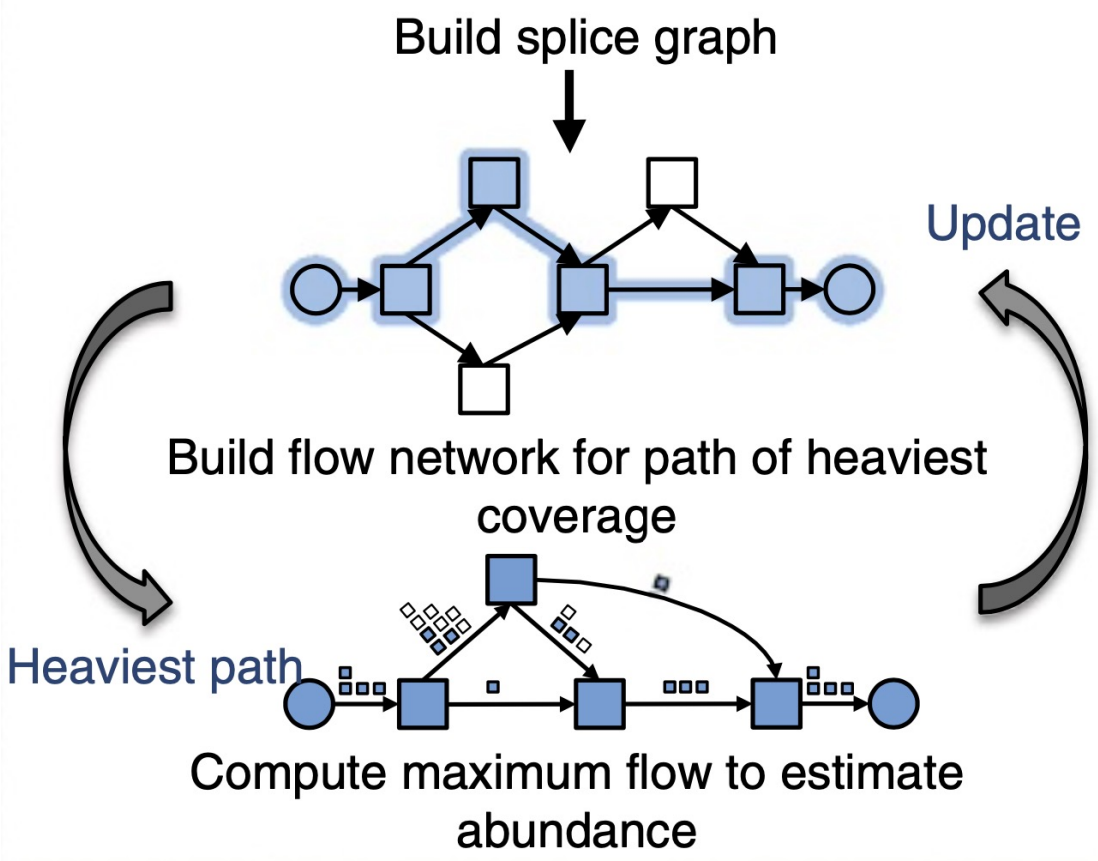
  1) a
  2) b, c, d
  3) e, j
  4) g, k

  => max fragments : 4 (Quantification)



Max flow : https://www.coursera.org/lecture/algorithms-part2/introduction-to-maxflow-jZVUm
StringTie Algo : https://www.youtube.com/watch?v=2qGiw4MRK3c

Main idea used in StringTie



- StringTie first groups the reads into clusters,(by gene locus)
- splice graph for each cluster is create and **identifies** transcripts(heaviest path)
- then for each transcript it creates a separate flow network to **estimate** its expression level using a maximum flow algorithm.

Using coverage information in identification & max-flow algorithm in quantification is the main idea!

How accurate is StringTie? (Experiment settings)

Simulation1

- using the exact parameters specified for a directional human RNA-seq protocol provided on the **Flux Simulator** web page

Simulation2

- we replaced the empirical fragment size distribution from Sim-I with a parameterized normal *N(250,20)* distribution, because some transcriptome assemblers13 assume a normal distribution.

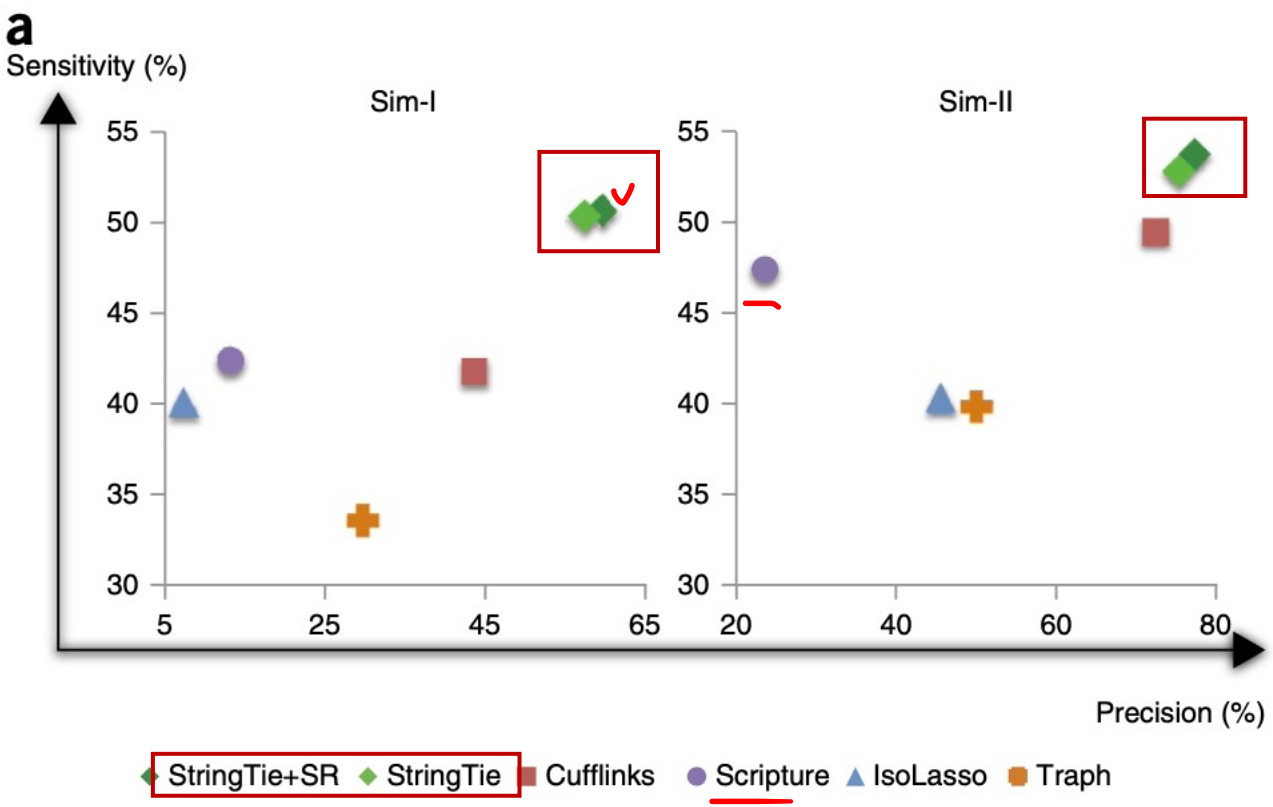
How accurate is StringTie? (Experiment settings)

Real Data (ENCODE Project from UCSC genome browser + Own data)

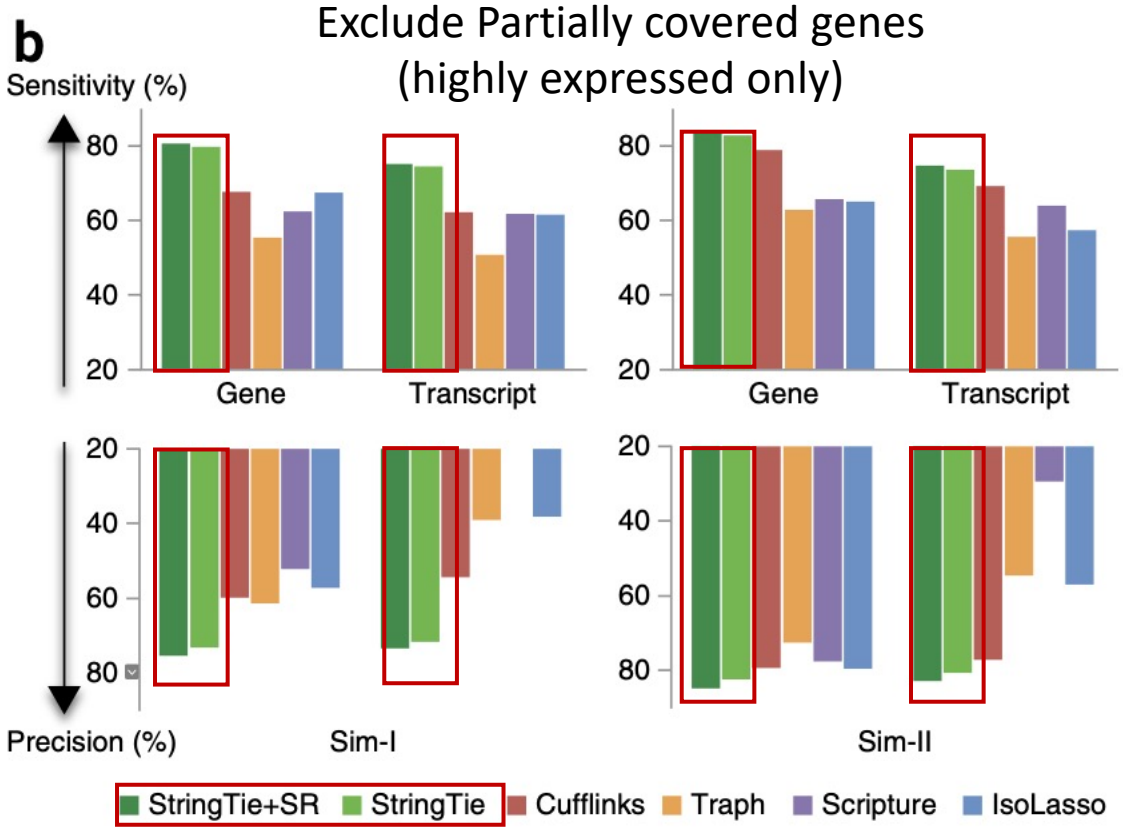
- whole B cells in blood : 90 million 76-bp paired-end reads
- Cytosol of fetal lung fibroblasts : 145 million 101-bp paired-end reads
- CD14-positive monocytes : 120 million 76-bp paired-end

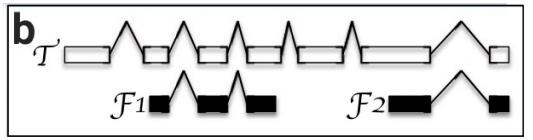
- Kidney cell line (own): 180 million 100-bp paired-end reads

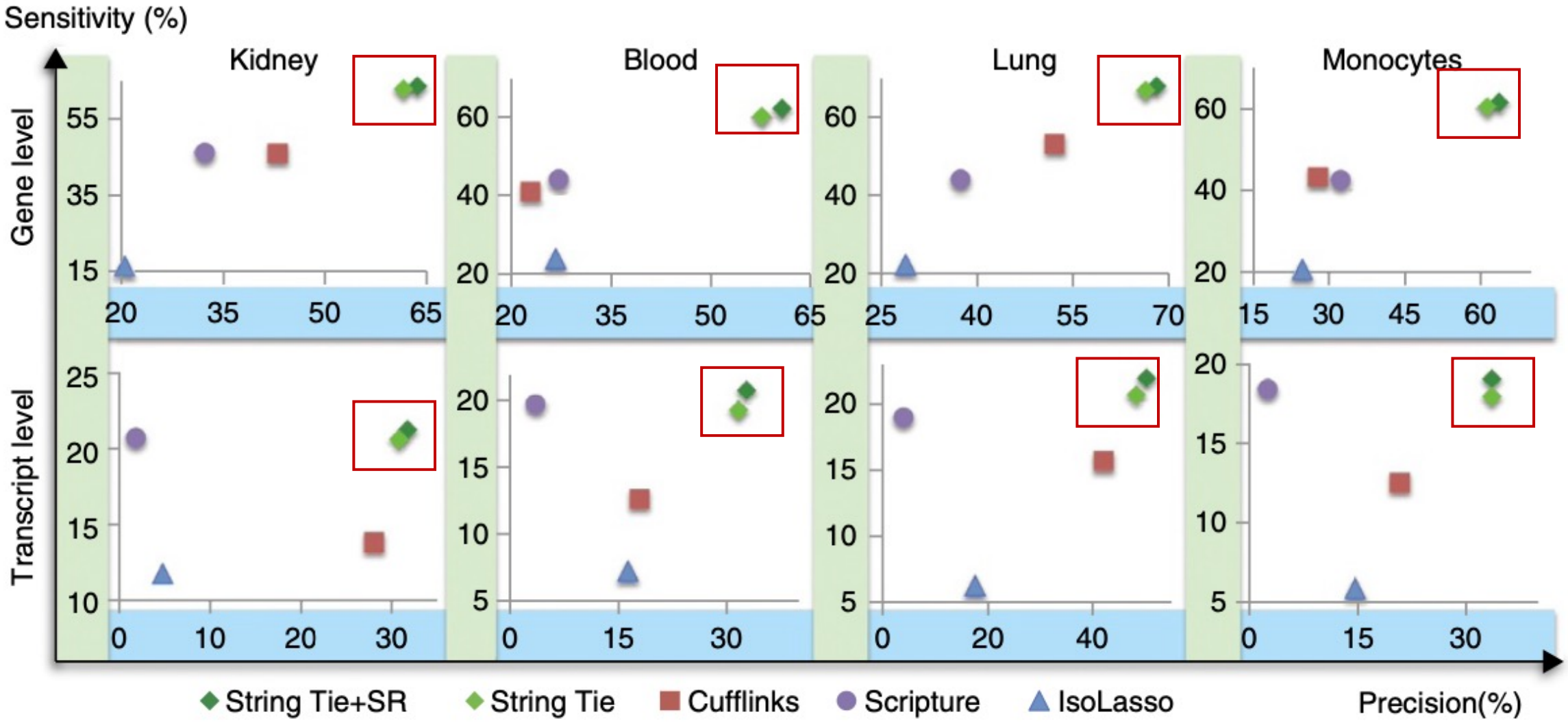
How accurate is StringTie? (Identification)



Exclude Partially covered genes
(highly expressed only)






Sensitivity (%)
Kidney
Blood
Lung
Monocytes
Gene level
Transcript level
55
35
15
60
40
20
60
40
20
60
40
20
25
20
15
10
20
15
10
5
20
15
10
5
20
35
50
65
20
35
50
65
25
40
55
70
15
30
45
60
0
10
20
30
0
15
30
0
20
40
0
15
30
String Tie+SR
String Tie
Cufflinks
Scripture
IsoLasso
Precision(%)

How accurate is StringTie? (Quantification)

Table 1 Transcriptome assemblers' performances on simulated and real data

Data set	Measure	StringTie+SR	StringTie	Cufflinks	Traph	Scripture	IsoLasso
Sim-I	ρ_{all}	0.648	0.646	0.551	0.080	–0.361	0.162
	$\rho_{predicted}$	0.871	0.878	0.826	0.432	–0.228	0.500
Sim–II	ρ_{all}	0.799	0.787	0.720	0.310	–0.435	0.000
	$\rho_{predicted}$	0.913	0.907	0.883	0.524	–0.301	0.258
Kidney	Genes	10,773	10,659	7,774	n/a	7,813	2,785
	Transcripts	13,900	13,720	9,245	n/a	13,833	3,191
Blood	Genes	9,198	8,938	6,073	n/a	6,533	3,526
	Transcripts	11,489	10,990	7,187	n/a	11,213	4,124
Lung	Genes	10,913	10,779	8,566	n/a	7,070	3,590
	Transcripts	14,055	13,706	10,370	n/a	12,559	4,187
Monocytes	Genes	9,005	8,859	6,351	n/a	6,244	3,020
	Transcripts	11,059	10,748	7,502	n/a	1,1046	3,528

FPKM => f1, f2, f3 => pred_rank1, pred_rank2, pred_rank3 => compare it with real rank => spearman correlation

Is it computationally better method?

- On the four real data sets used here, the maximum memory used by
  - StringTie : **1.6 - 12 GB.**
  - Cufflinks, IsoLasso and Scripture : 6.4 - 26.6 GB

- Computation time comparison on four real data sets
  - StringTie required from 35–76 min
  - StringTie was more than three times faster than the fastest of the other four programs, and in some cases over **50 times faster**

Why StringTie is good?

- Analyzed the genes that were found by StringTie but not by Cufflinks, and vice versa, on all four of the real data sets
- StringTie was better at reconstructing at least three types of genes:

(i) those at low abundance;

(ii) those with more exons and

(iii) those with multiple isoforms.

Summary : StringTie

StringTie *is a RNA-seq assembly methods that*
1) applies a network flow algorithm,
2) together with optional de novo assembly,
to assemble these complex data sets into transcripts.

- more complete and accurate reconstructions of genes
- better estimates of expression levels
- faster on all data sets tested to date

What about StringTie2? : updated version

The new StringTie2 focuses on extending StringTie1's capacity to handle long-read data

- Implements much more efficient data structures that overall lead to faster run times and much lower memory usage
- To handle the high error rates in the long reads,
    1. Correct potentially wrong splice sites by checking all the splice sites present in the alignment of a read with a high-error alignment rate
    2. Designed and implemented a pruning algorithm that reduces the size of the splicing graph to a more realistic size

Q & A