Introduction to RNA-seq Analysis II

Gunnar Rätsch

Computational Biology Center
Memorial Sloan-Kettering Cancer Center



Memorial Sloan-Kettering
Cancer Center

cBio@MSKCC

Overview

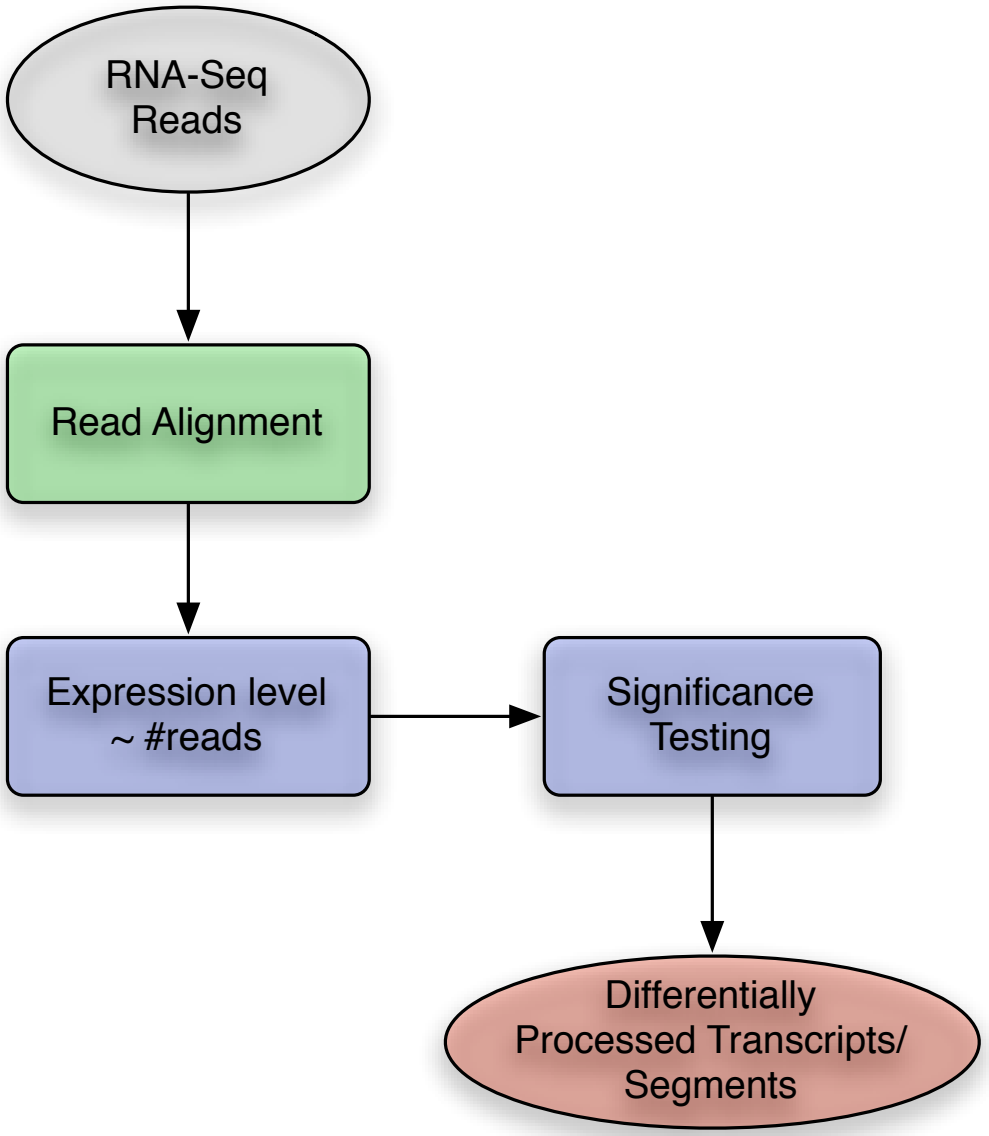
Part I (yesterday)

- RNAs, RNA-seq & research questions
- Common analysis steps
- Read mapping
  - The problem
  - String data structures
  - Programs
  - Extensions

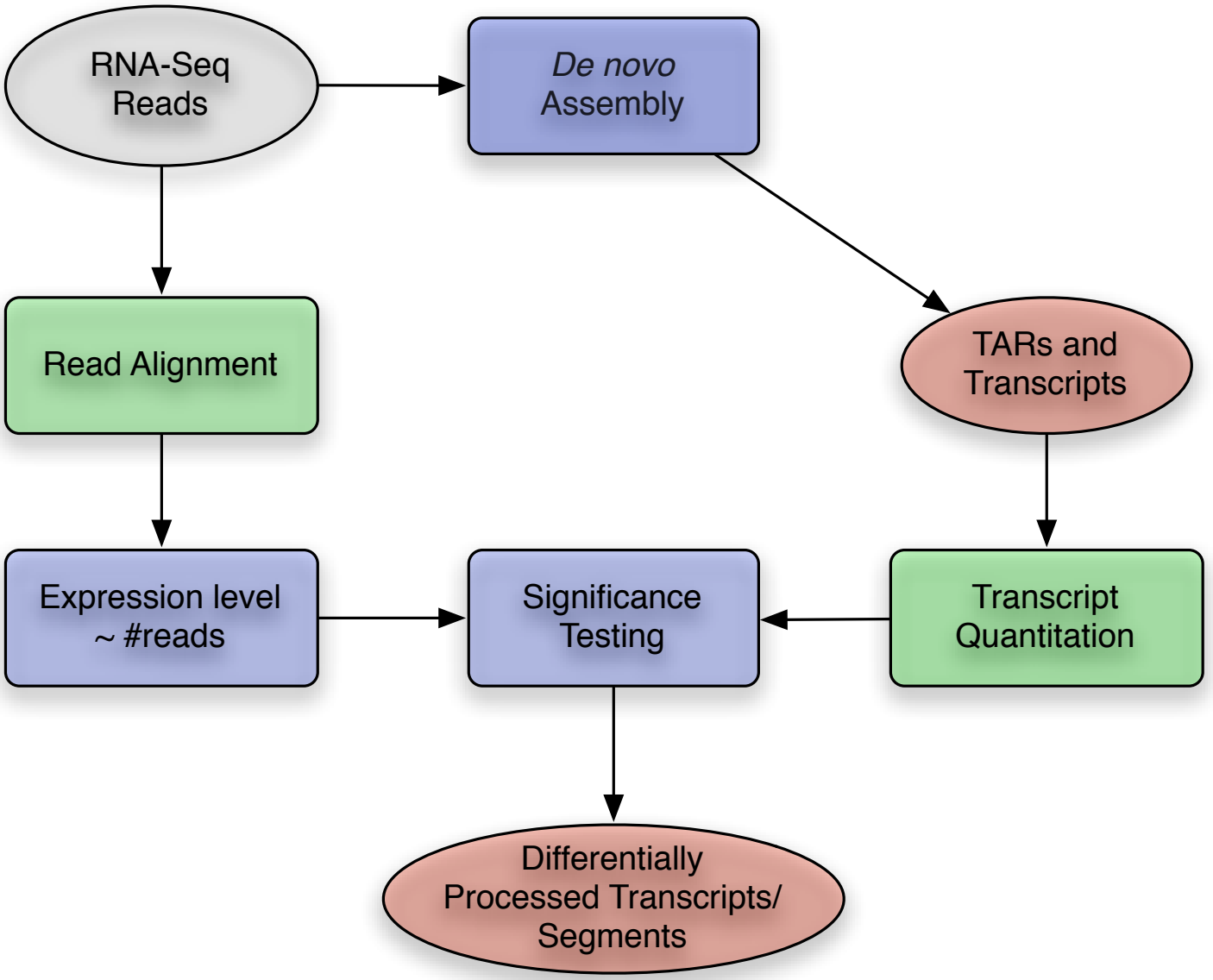
Part II (today)

- Differences between read mappers (RGASP)
- Gene/Transcript quantification & biases
- Differential "expression"
- Transcript reconstruction
- Genome annotation

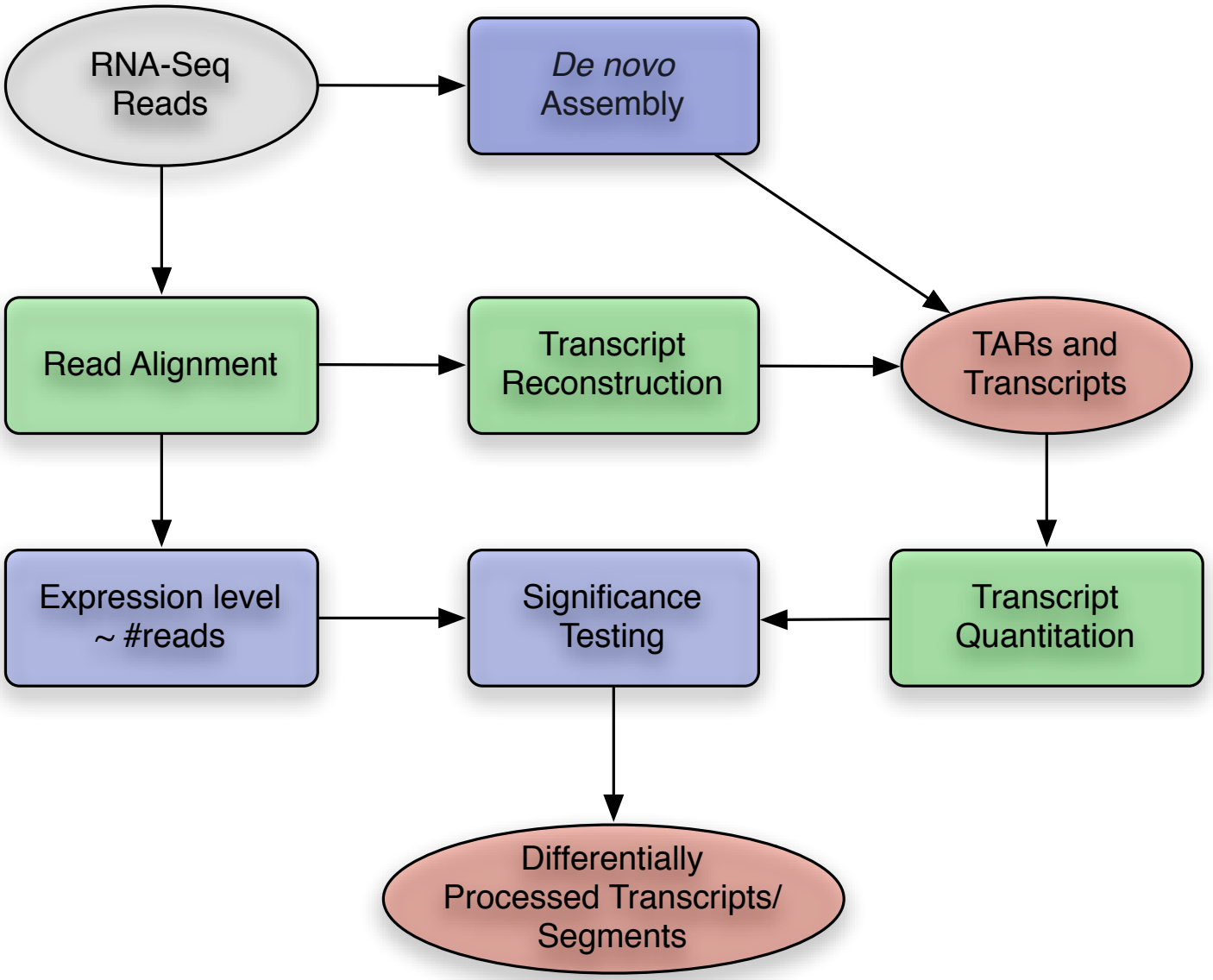
Common RNA-Seq Analysis Steps



Common RNA-Seq Analysis Steps



Common RNA-Seq Analysis Steps



Estimate Gene Expression

Idea: Use the number of reads mapping to a gene as an estimate for the gene expression.
Problem: Read number scales with the total number of reads and transcript length

Approach: Normalize read count, by the

- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in millions)

$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

Alternative quantity for paired end sequencing (2 reads/fragment):

$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

Estimate Gene Expression

Idea: Use the number of reads mapping to a gene as an estimate for the gene expression.
Problem: Read number scales with the total number of reads and transcript length

Approach: Normalize read count, by the

- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in millions)

$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

Alternative quantity for paired end sequencing (2 reads/fragment):

$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

Estimate Gene Expression

Idea: Use the number of reads mapping to a gene as an estimate for the gene expression.
Problem: Read number scales with the total number of reads and transcript length

Approach: Normalize read count, by the

- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in millions)

$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

Alternative quantity for paired end sequencing (2 reads/fragment):

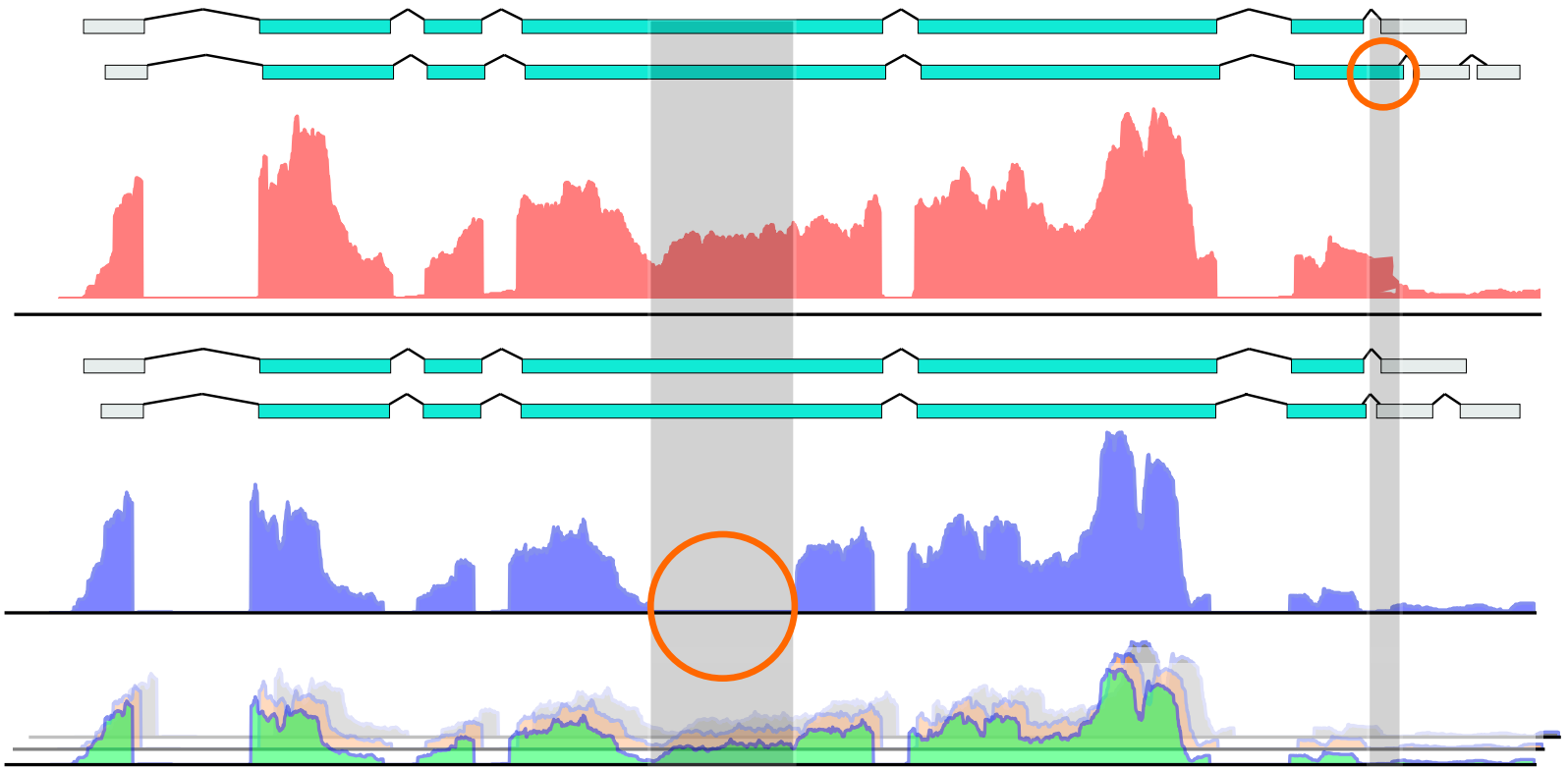
$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

Estimate Gene Expression: Caveats

- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation
- Alternative transcripts/RNA-processing may lead to differential read counts

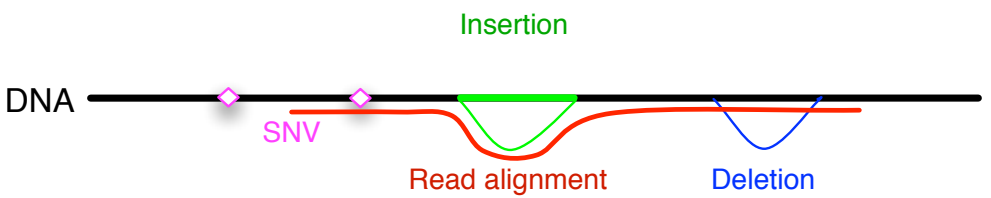
Estimate Gene Expression: Caveats

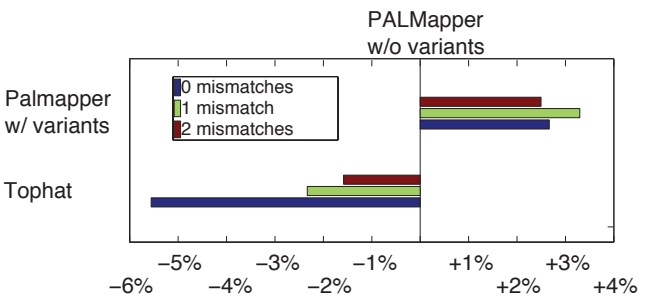
- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation



Estimate Gene Expression: Caveats

- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation

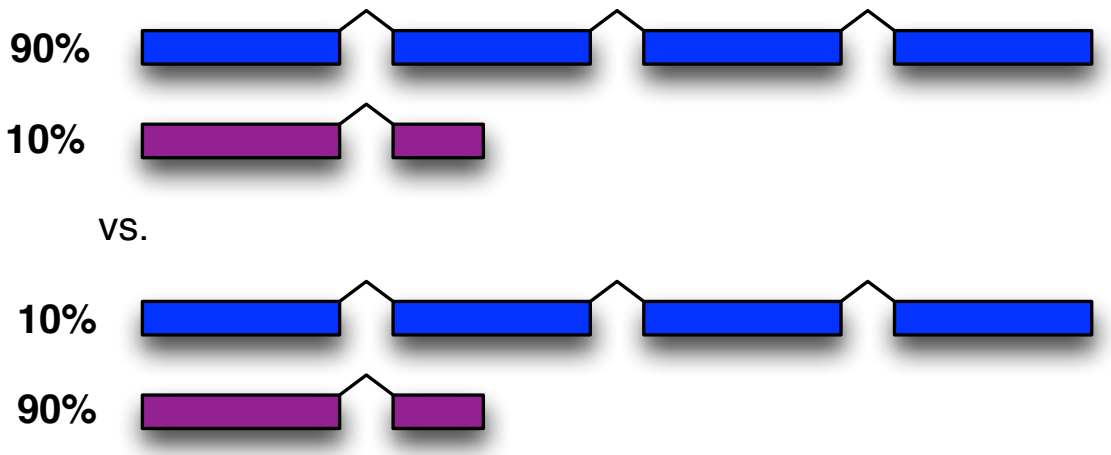




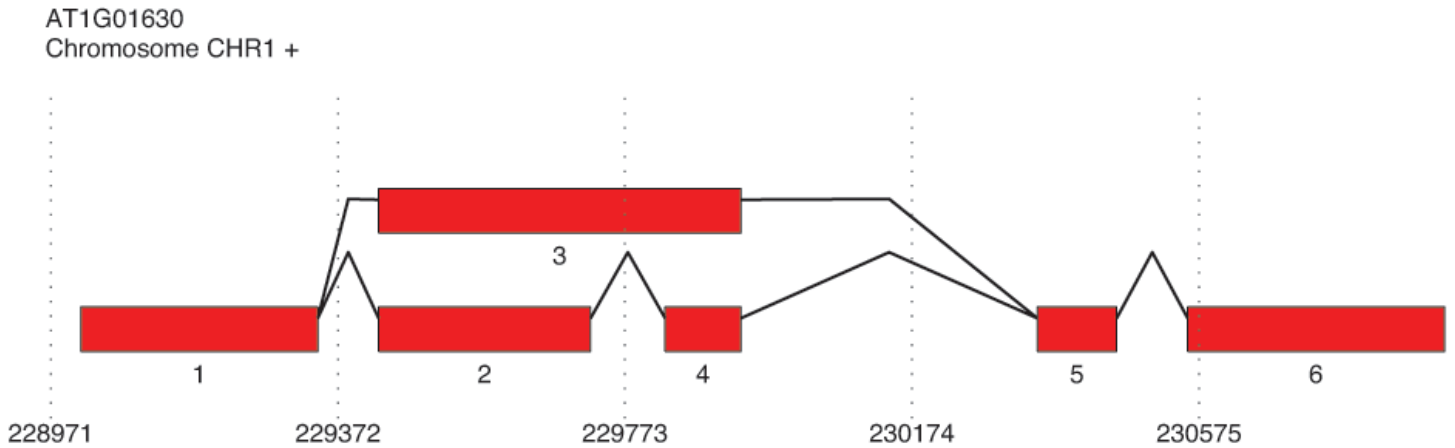
- Alternative transcripts/RNA-processing may lead to differential read counts

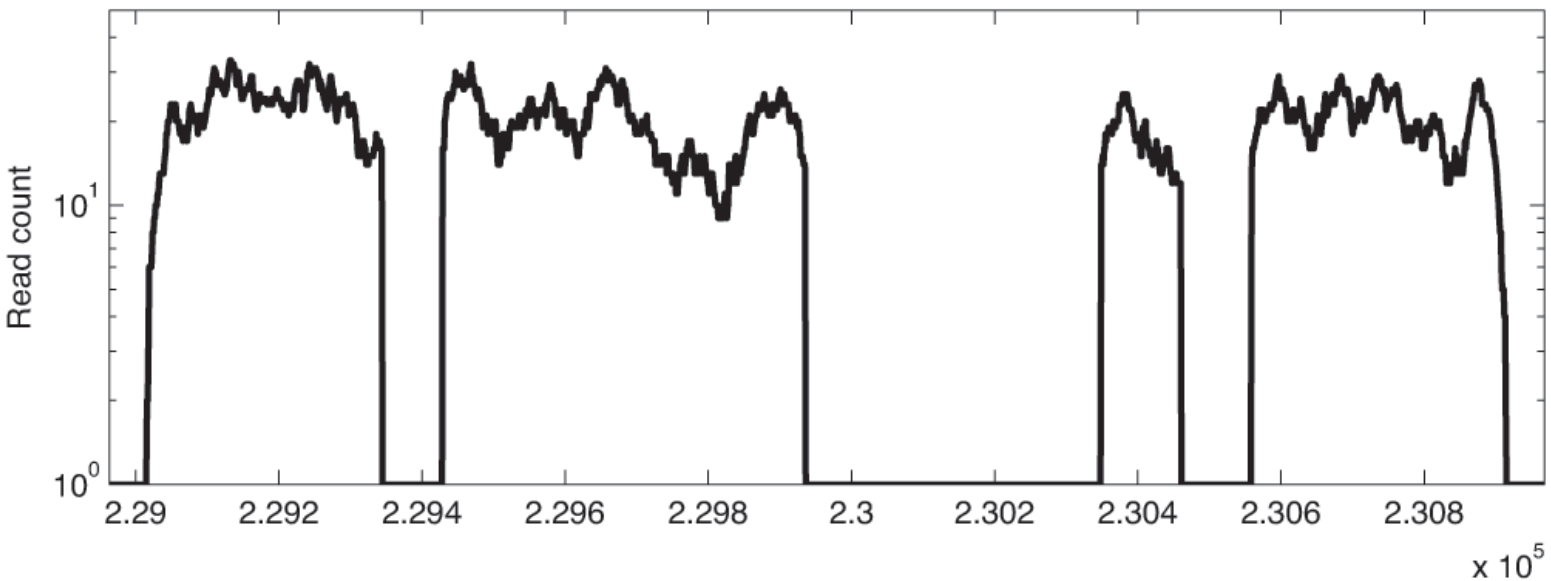
Estimate Gene Expression: Caveats

- RPKM/FPKM values are strongly dependent on the expression level of the highest expressed genes (largest fraction of reads, e.g. rRNA contamination)
- Effect of genomic variation
- Alternative transcripts/RNA-processing may lead to differential read counts

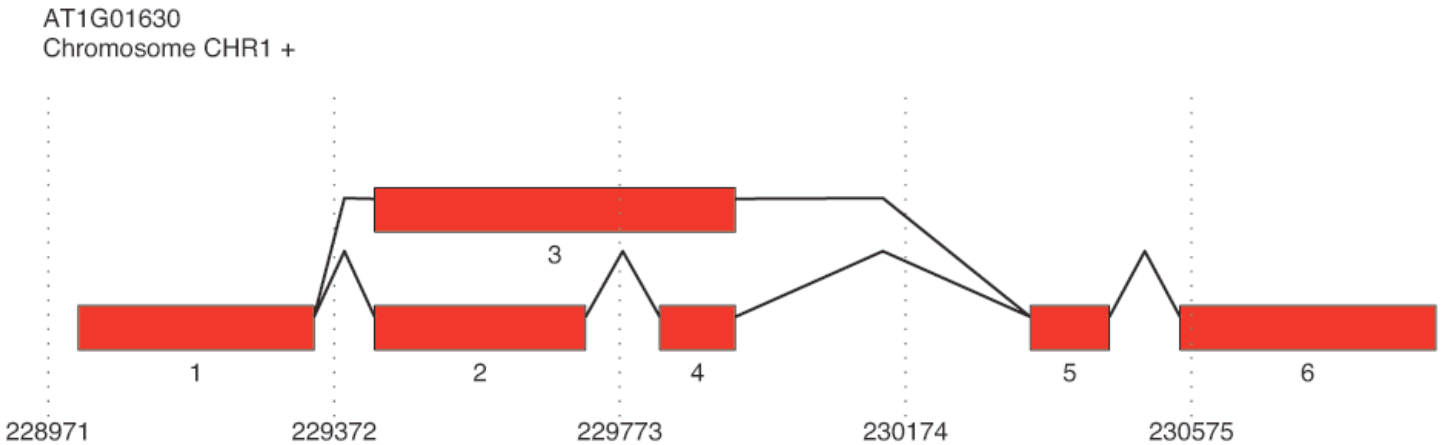


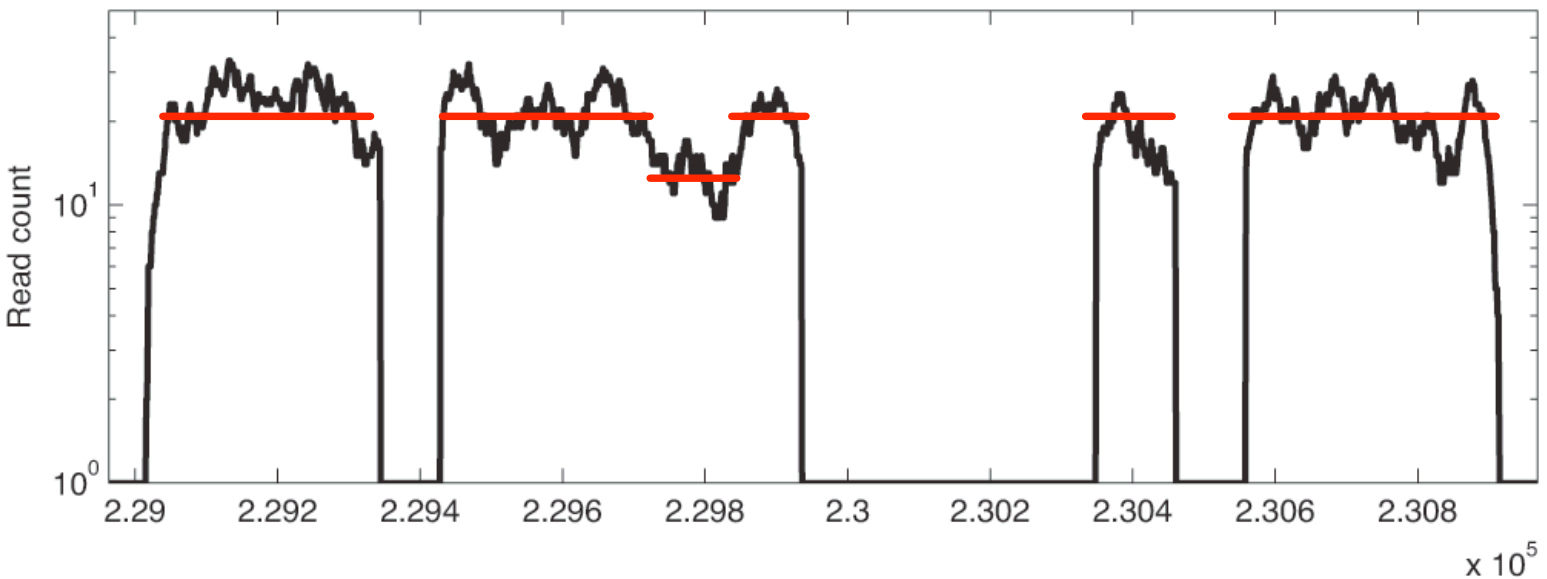
Quantitation of Transcripts





Quantitation of Transcripts





Quantitation of Transcripts

Given short read alignments and a set of known transcripts, can we disentangle transcript abundances?

Solve an optimization problem:

- Optimizing weights w_t for each transcript $t = 1, \ldots, T$
- Exploiting the additive nature of the read coverage
- Minimizing residual error (e.g., squared error)

$$(w_1, \ldots, w_T) = \underset{w_1,\ldots,w_T \geq 0}{\operatorname{argmin}} \sum_{p \in P} \left(R_p - \sum_{t=1}^{T} w_t D_{t,p} \right)^2 ,$$

with

- P: set of considered genomic positions
- R_p: observed read coverage (number of reads covering pos. p)
- $D_{t,p}$: expected read coverage for transcript t at position p

Quantitation of Transcripts

Different approaches rely on similar basic ideas with different models of how to use read count differences and optimization techniques:

- Poisson distributions [Jiang and Wong, 2009]
- Absolute differences using a flow-network [Sammeth, 2009a]
- Squared differences using quadratic programming [Bohnert et al., 2009]
- (approximate) Negative Binomial distribution [Behr et al., 2013]

Other methods: [Li et al., 2010a], [Richard et al., 2010a], [Trapnell et al., 2010b]

Problems:

- Abundances cannot be unambiguously determined with single-end reads; better chance with paired-end reads [Lacroix et al., 2008]
- Solution may not be stable: a few reads more or less may completely change abundance estimates
- Read coverage is not uniform over the transcript

Quantitation of Transcripts

Different approaches rely on similar basic ideas with different models of how to use read count differences and optimization techniques:

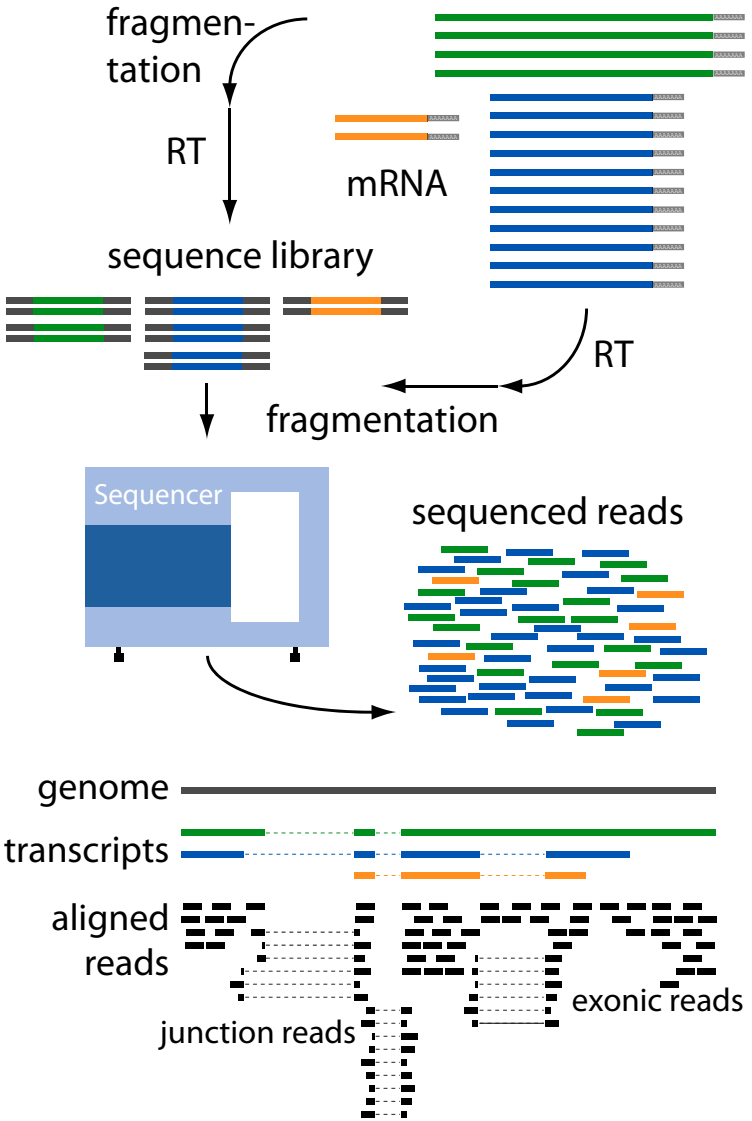
- Poisson distributions [Jiang and Wong, 2009]
- Absolute differences using a flow-network [Sammeth, 2009a]
- Squared differences using quadratic programming [Bohnert et al., 2009]
- (approximate) Negative Binomial distribution [Behr et al., 2013]

Other methods: [Li et al., 2010a], [Richard et al., 2010a], [Trapnell et al., 2010b]

Problems:

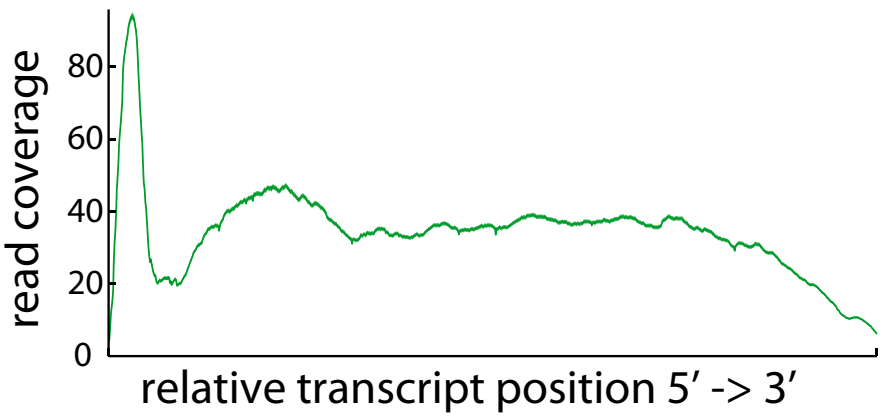
- Abundances cannot be unambiguously determined with single-end reads; better chance with paired-end reads [Lacroix et al., 2008]
- Solution may not be stable: a few reads more or less may completely change abundance estimates
- Read coverage is not uniform over the transcript

RNA-Seq Biases and Quantitation



Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping

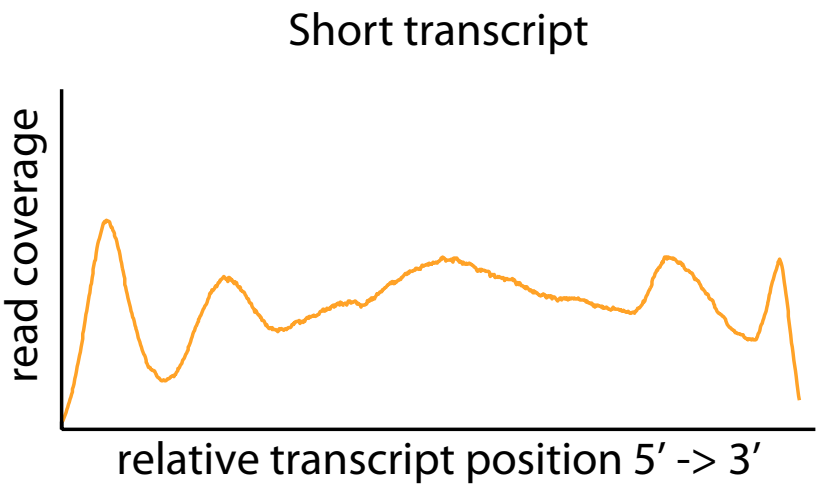


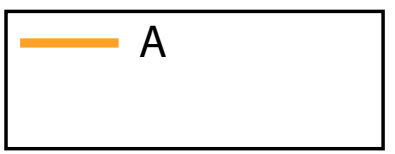
(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

RNA-Seq Biases – Publications

- Transcript length bias [Oshlack and Wakefield, 2009]
- Bias in small RNA sequencing [Linsen et al., 2009]
- Sequence bias in Illumina sequencer [Dohm et al., 2008]
- Biases in nucleotide composition induced by random hexamer priming [Hansen et al., 2010]
- Modeling non-uniformity [Li et al., 2010b, Bohnert et al., 2009]

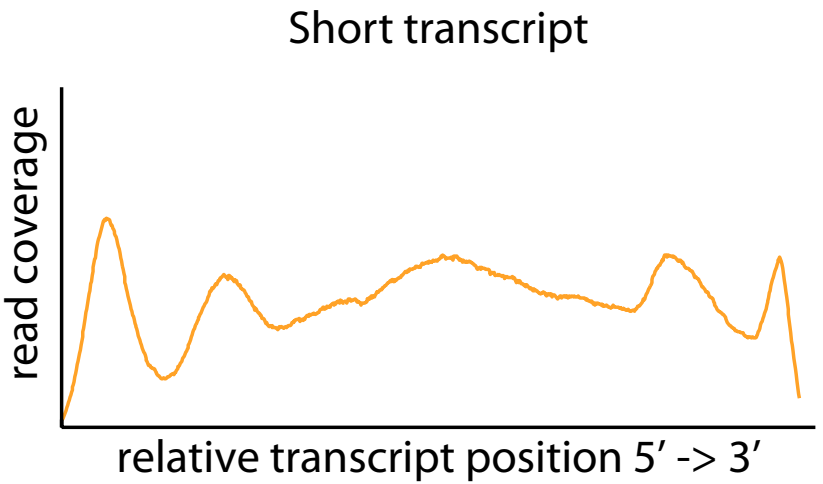
rQuant – Basic Idea

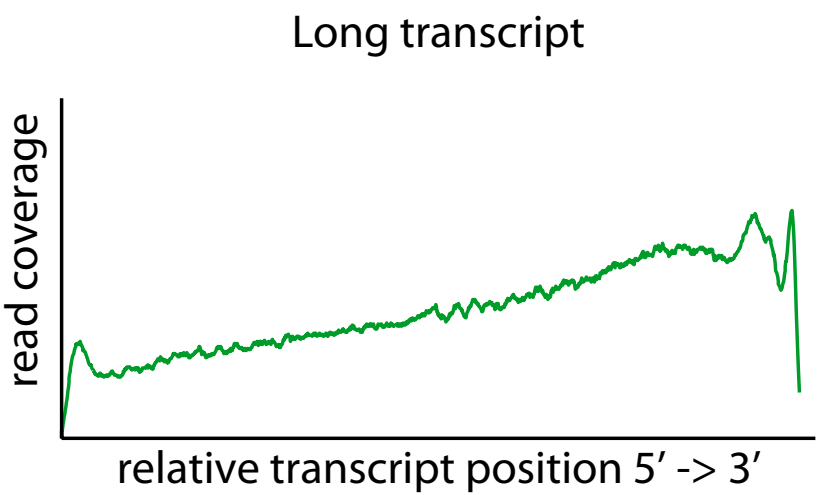


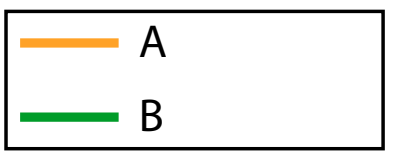


$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea

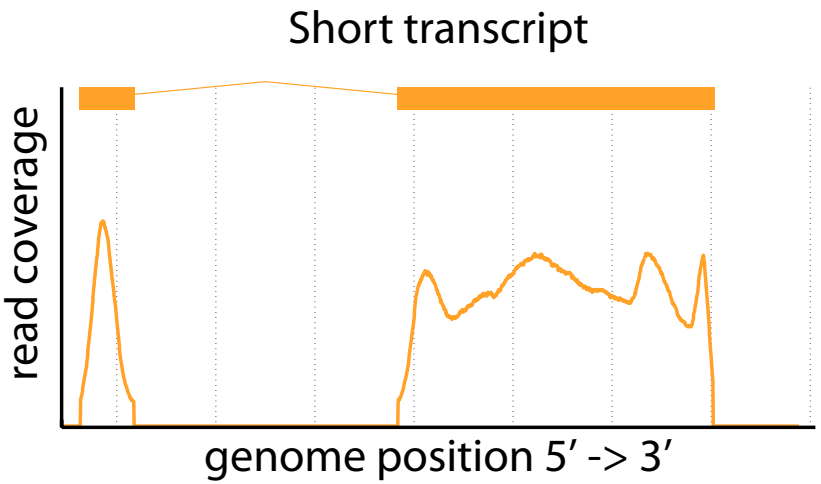


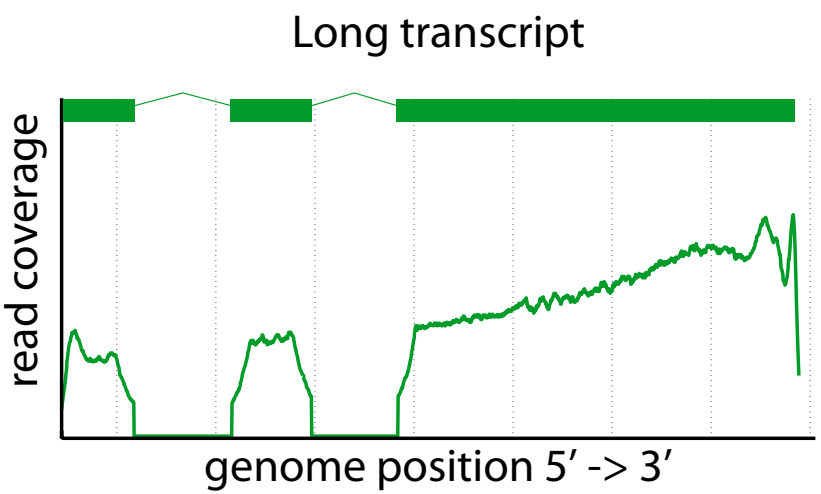


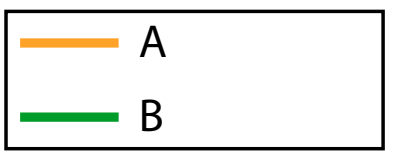


$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea

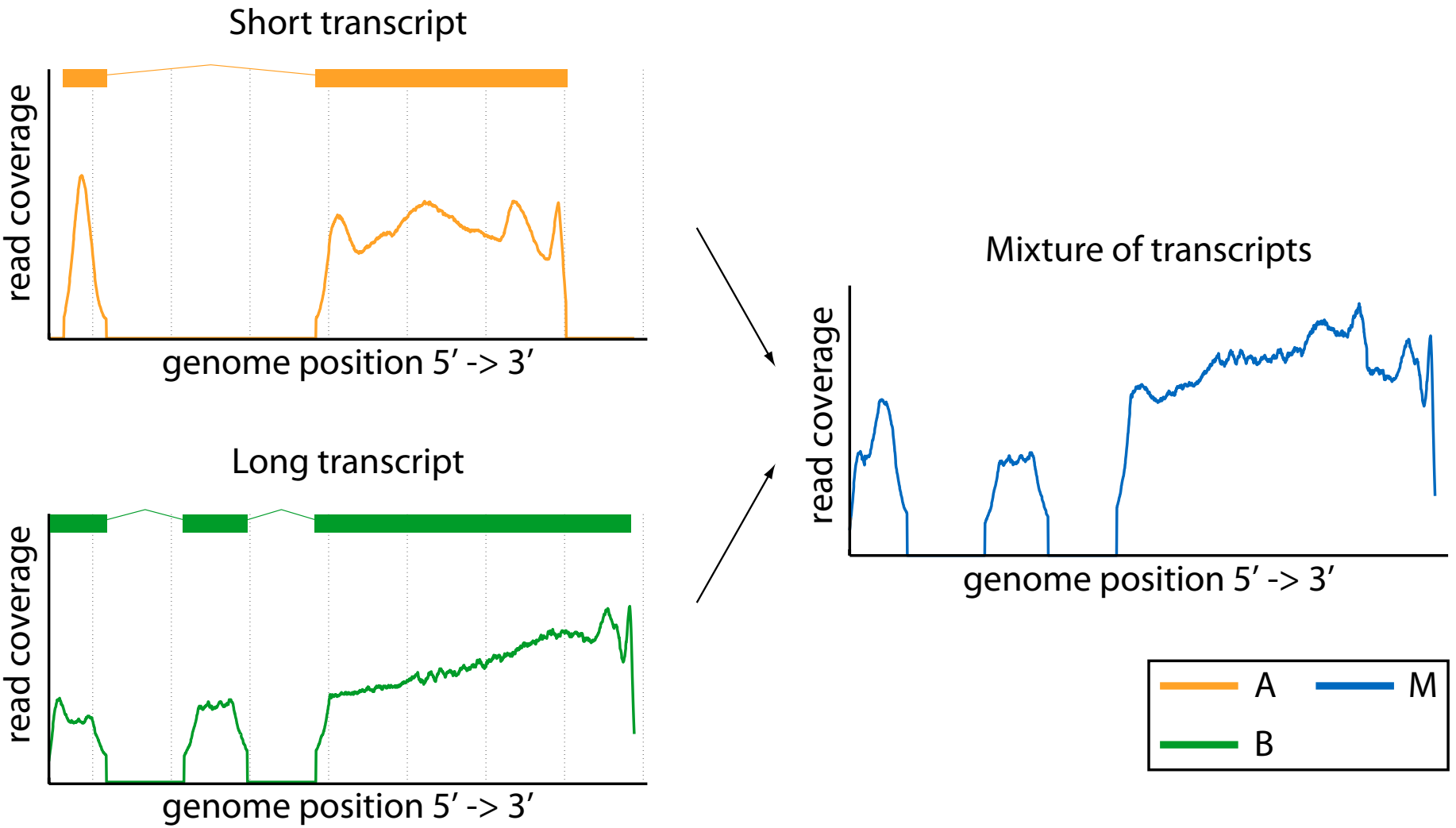






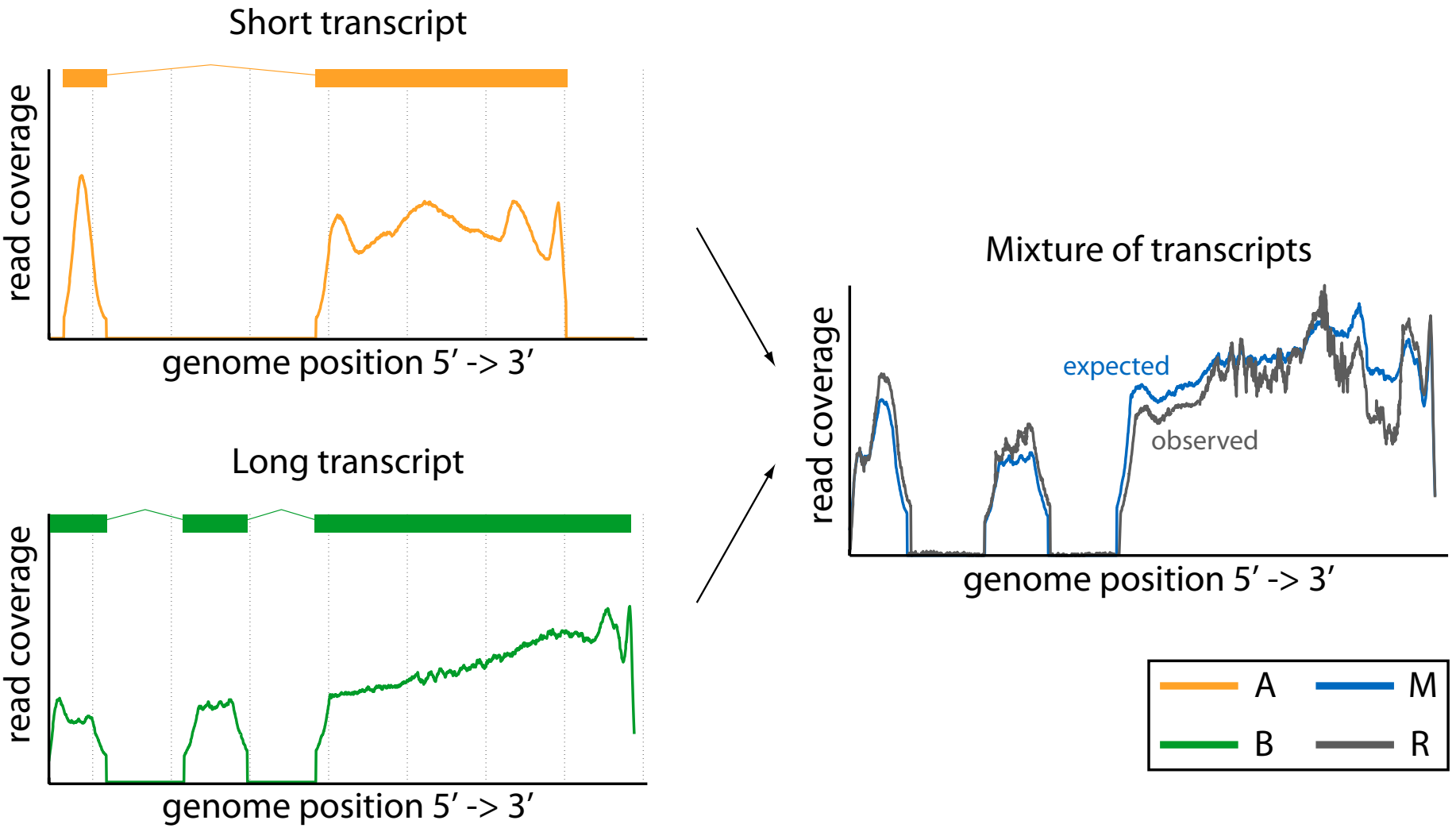
$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea



$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

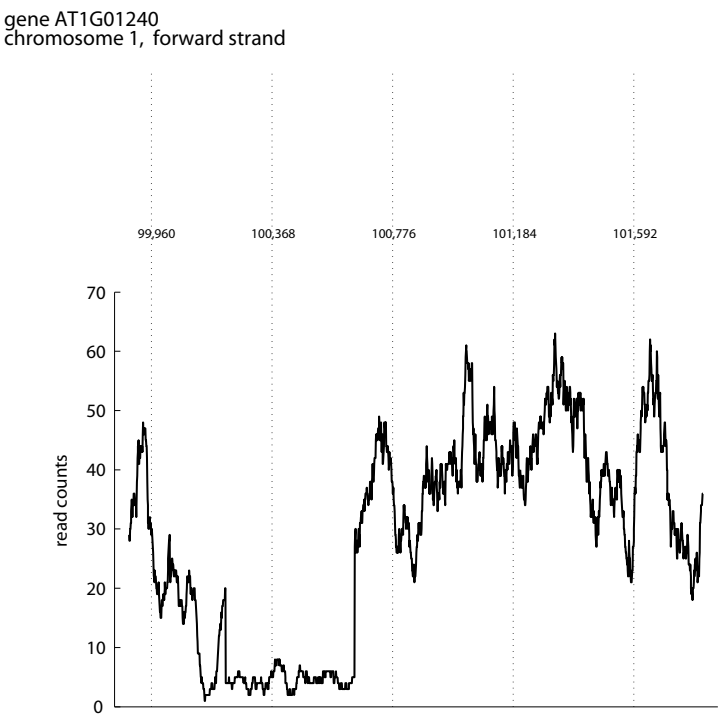
rQuant – Basic Idea



$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell\left(M_p, R_p\right)$$

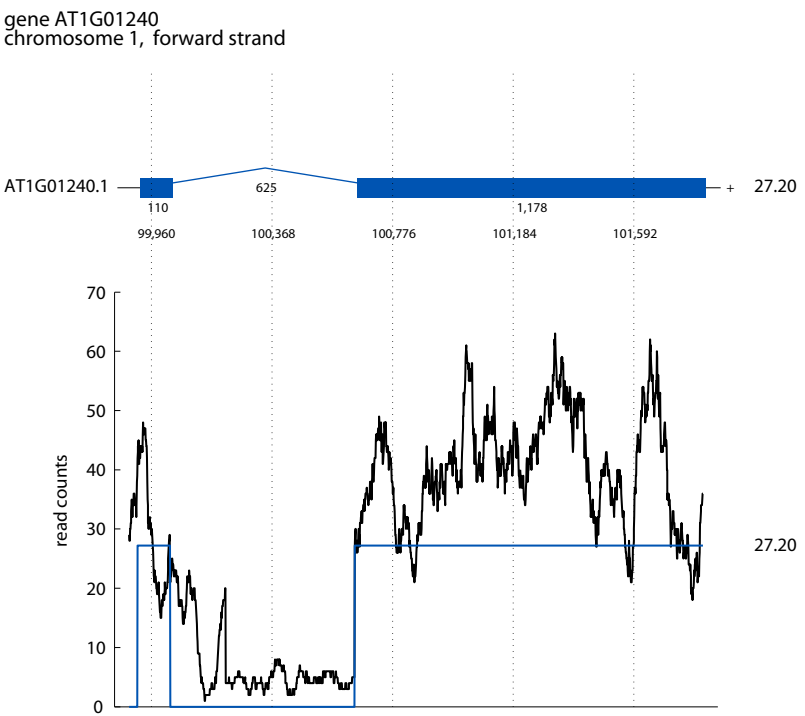
rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence



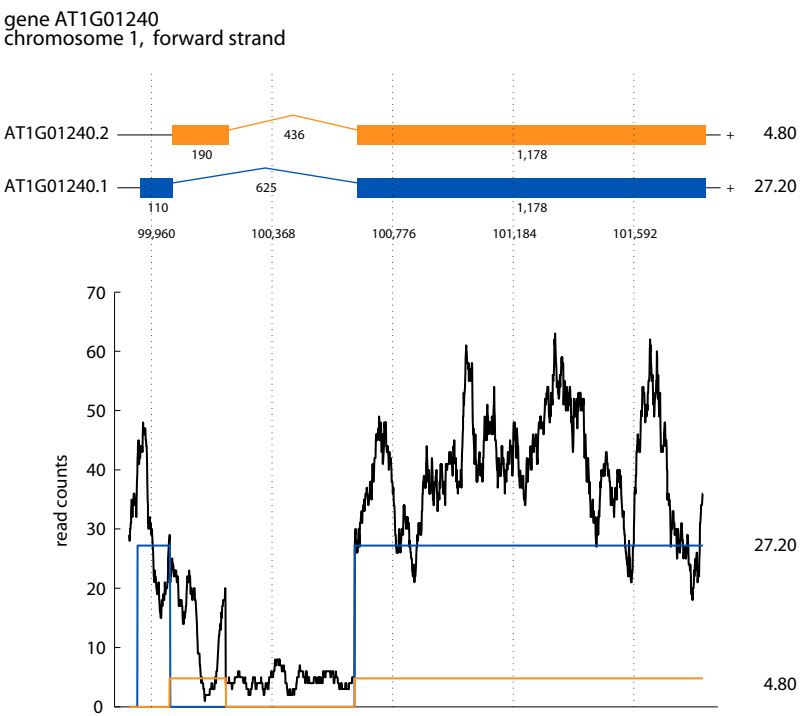
rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence



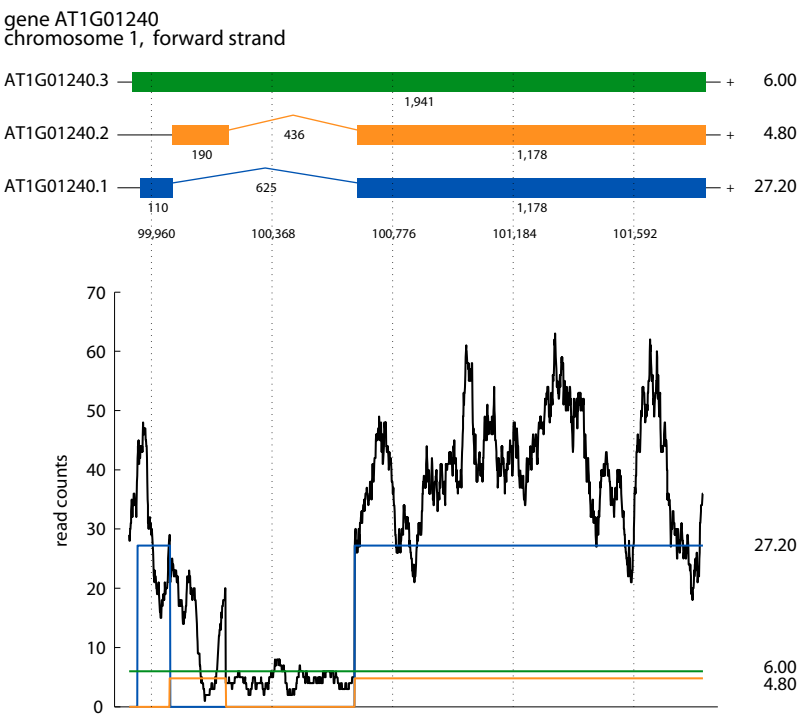
rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence



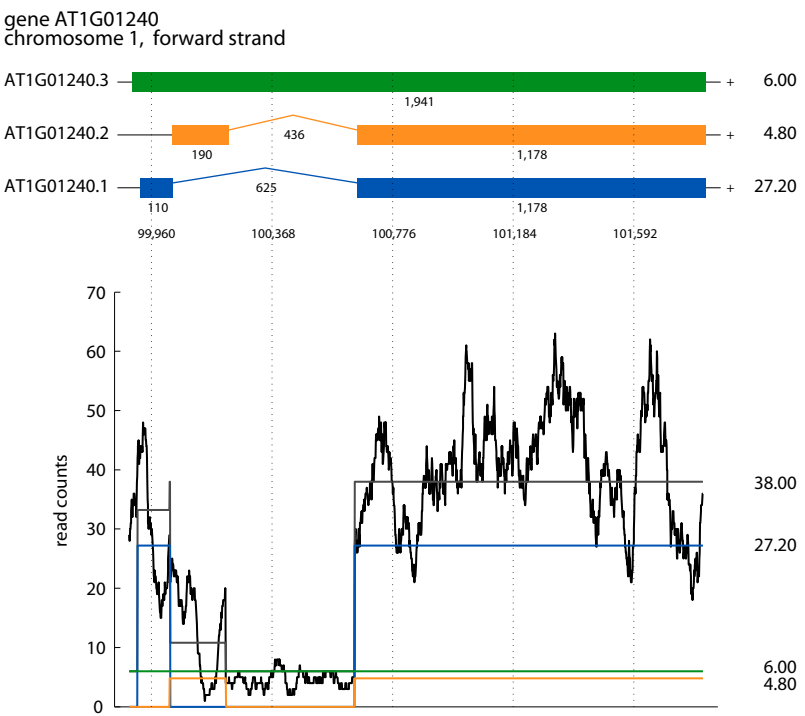
rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence



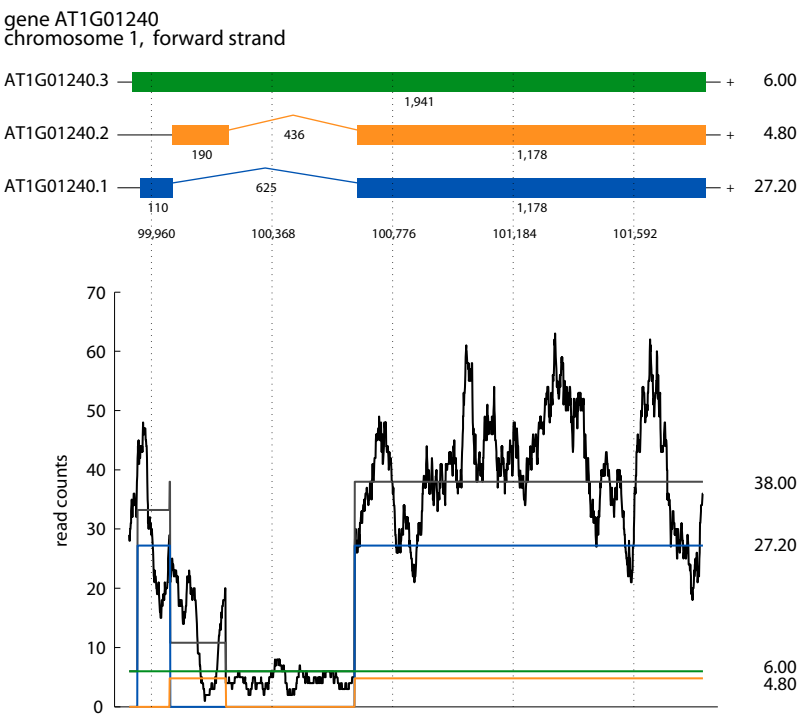
rQuant – Iterative Algorithm

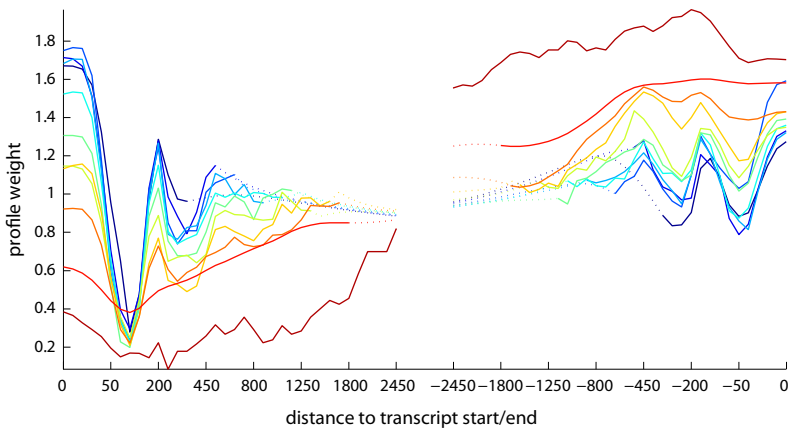
1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence



rQuant – Iterative Algorithm

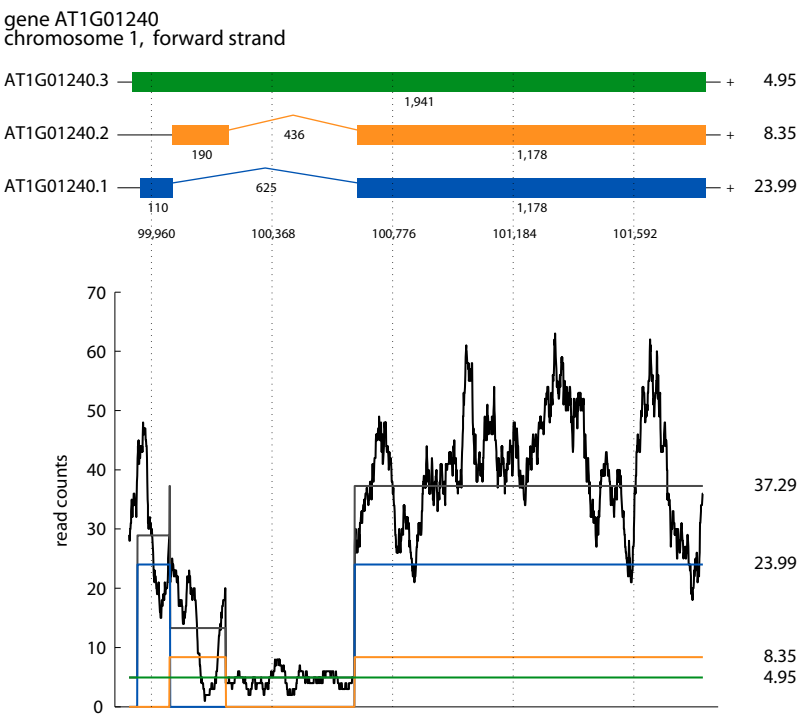
1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence

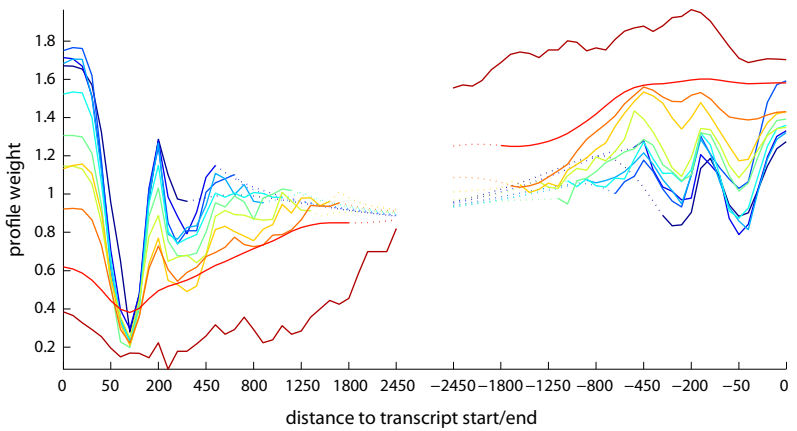




rQuant – Iterative Algorithm

1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence





rQuant Results

Simulated RNA-seq libraries (75 bp) using Flux Simulator for *C. elegans*

[Bohnert & Rätsch, 2010; Bohnert, 2011]

Approach		Pearson's correlation across genes		Pearson's correlation per gene	
rQuant baseline		**0.940**	[10,180]	**0.805**	[3,022]
	By expression				
	low	0.109	[3,770]	0.702	[720]
	medium	0.479	[3,279]	0.822	[1,310]
	high	0.938	[3,131]	0.856	[992]
	By transcript length				
	short	0.996	[3,272]	0.782	[818]
	medium	0.990	[4,680]	0.866	[1,489]
	long	0.706	[2,228]	0.702	[715]
	By number of transcripts				
	1	0.996	[2,945]	n/a	
	2	0.984	[4,500]	0.787	[2,249]
	3	0.723	[1,515]	0.852	[505]
	4 and more	0.614	[1,220]	0.868	[268]

rQuant Results

Simulated RNA-seq libraries (75 bp) using Flux Simulator for *C. elegans*

[Bohnert & Rätsch, 2010; Bohnert, 2011]

Approach		Pearson's correlation across genes		Pearson's correlation per gene	
rQuant baseline		**0.940**	[10,180]	**0.805**	[3,022]
rQuant profiles					
	empirical	**0.957**	[10,180]	0.832	[3,022]
	optimal	0.946	[10,180]	**0.836**	[3,022]

rQuant Results

Simulated RNA-seq libraries (75 bp) using Flux Simulator for *C. elegans*

[Bohnert & Rätsch, 2010; Bohnert, 2011]

Approach		Pearson's correlation across genes		Pearson's correlation per gene	
rQuant baseline		**0.940**	[10,180]	**0.805**	[3,022]
rQuant profiles					
	empirical	**0.957**	[10,180]	0.832	[3,022]
	optimized	0.946	[10,180]	**0.836**	[3,022]
Other methods					
	Cufflinks	0.932	[10,180]	0.758	[3,022]
	Cufflinks bias	0.936	[10,180]	0.793	[3,022]
	MISO	n/a		0.693	[2,903]

Cufflinks: [Trapnell et al., 2010a] MISO: [Katz et al., 2010]

Dependence on Alignments

Idealized simulation:

- Simulate reads from annotated transcripts (Flux Simulator)
- Use reads' origins as alignments
- Use alignments for quantitation

More realistic setting:

- Simulate reads from annotated transcripts (Flux Simulator)
- Add errors to reads
- Align reads to genome with 0-8 mismatches (many alignments)
- Use alignments for quantitation

False alignments, multi-mappers etc. lead to weaker results

Dependence on Alignments

Idealized simulation:

- Simulate reads from annotated transcripts (Flux Simulator)
- Use reads' origins as alignments
- Use alignments for quantitation

More realistic setting:

- Simulate reads from annotated transcripts (Flux Simulator)
- Add errors to reads
- Align reads to genome with 0-8 mismatches (many alignments)
- Use alignments for quantitation

False alignments, multi-mappers etc. lead to weaker results

Dependence on Alignments

Idealized simulation:

- Simulate reads from annotated transcripts (Flux Simulator)
- Use reads' origins as alignments
- Use alignments for quantitation

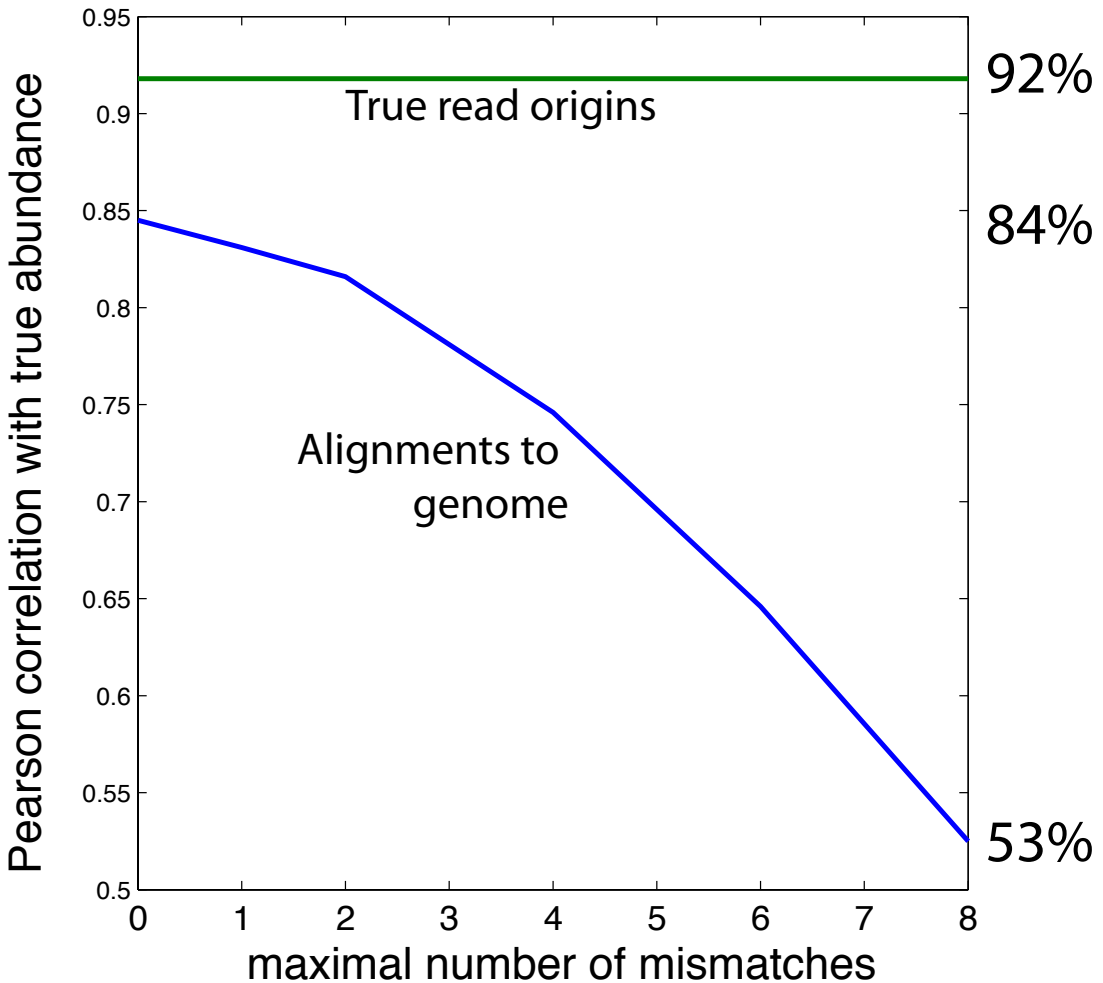
More realistic setting:

- Simulate reads from annotated transcripts (Flux Simulator)
- Add errors to reads
- Align reads to genome with 0-8 mismatches (many alignments)
- Use alignments for quantitation

False alignments, multi-mappers etc. lead to weaker results

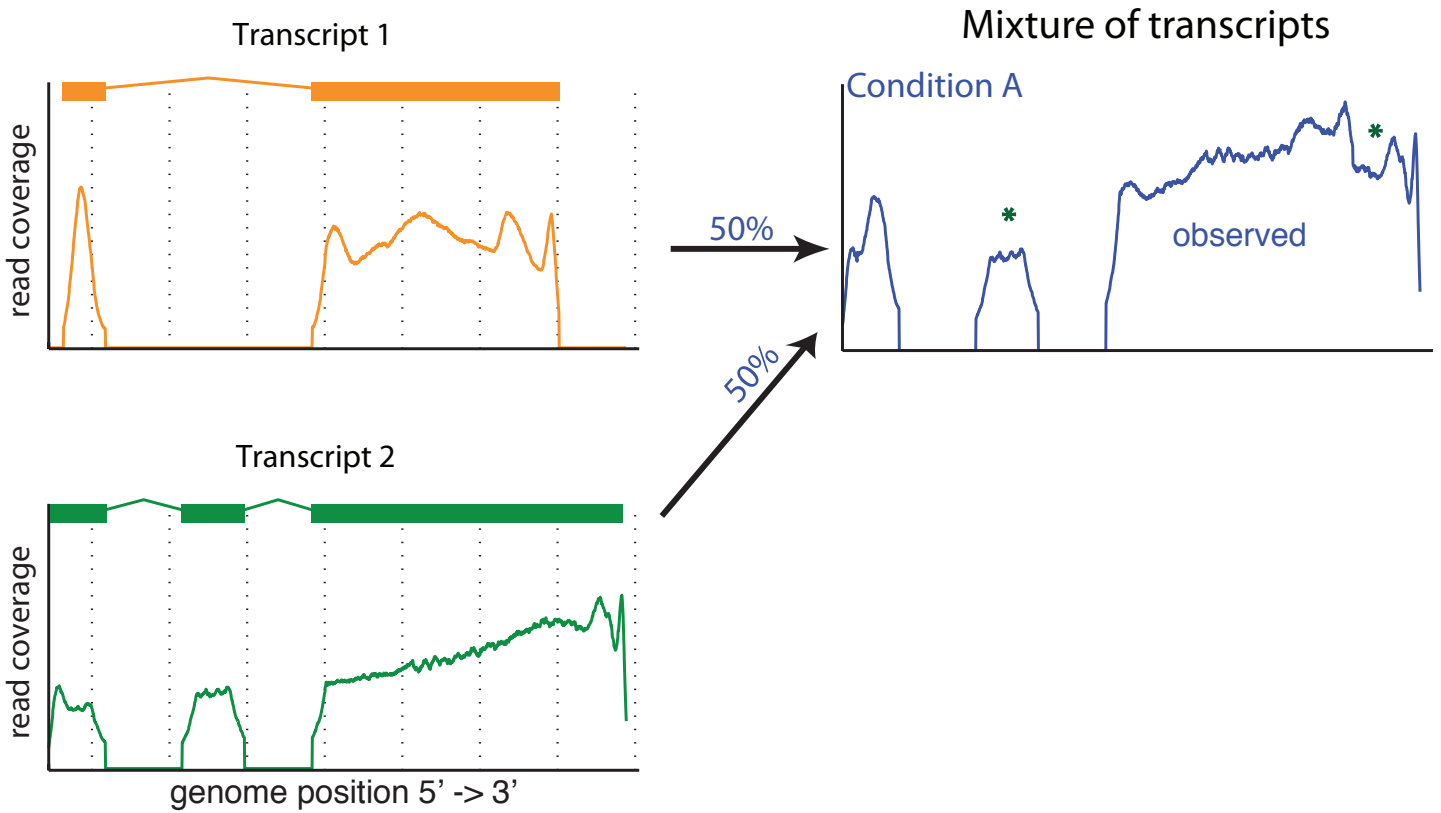
Dependence on Alignments

False alignments, multi-mappers etc. lead to weaker results

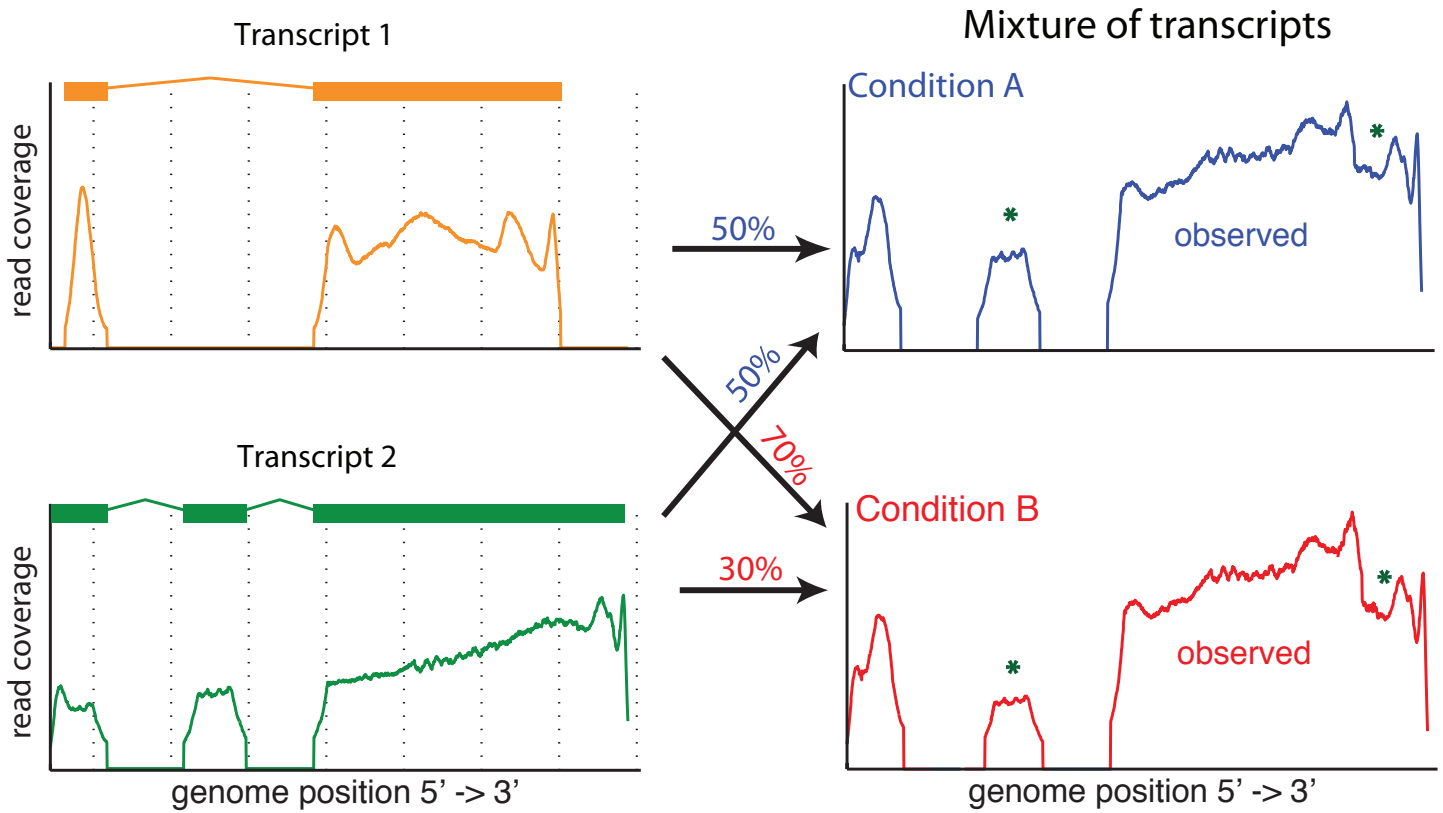


Detecting differential RNA processing

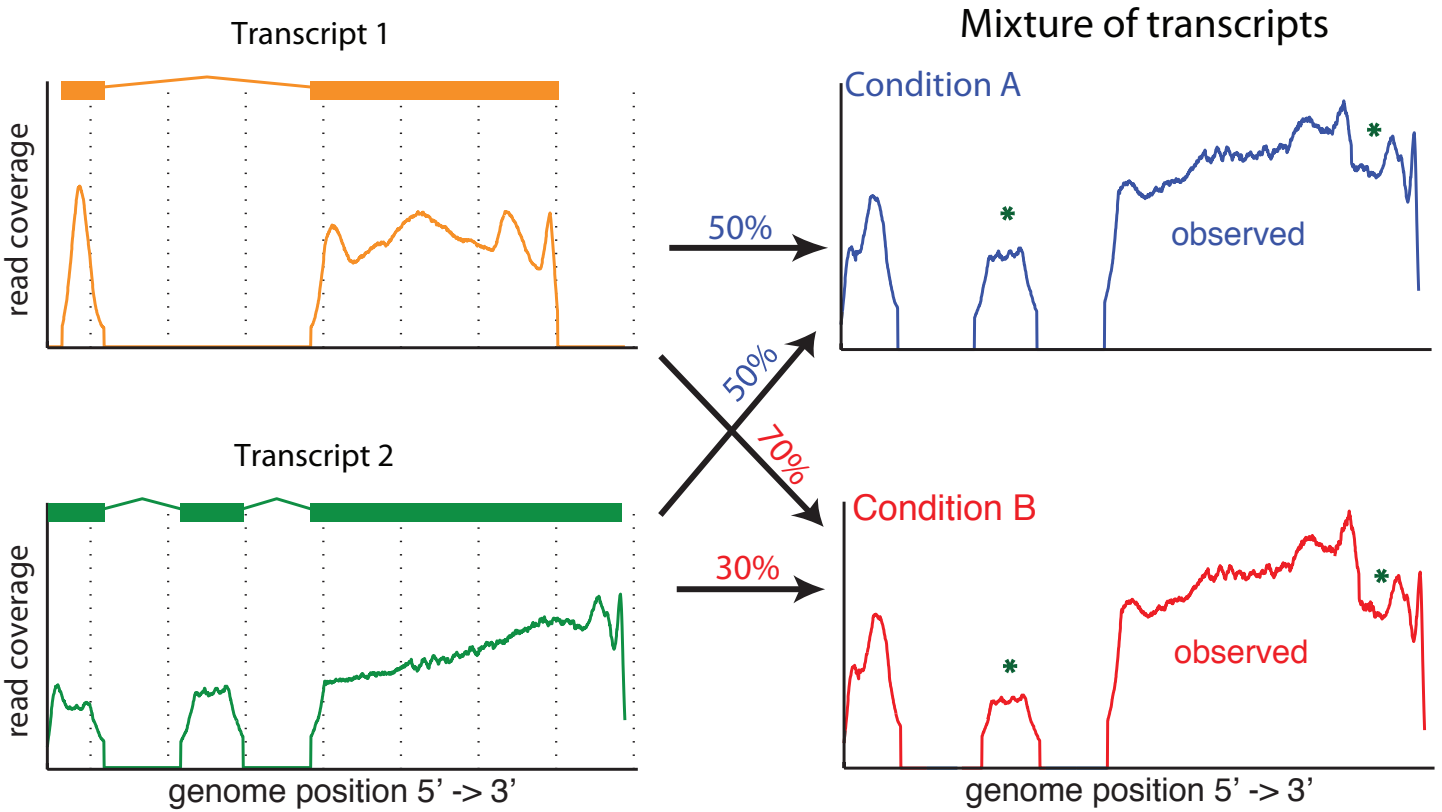
Detecting differential RNA processing



Detecting differential RNA processing

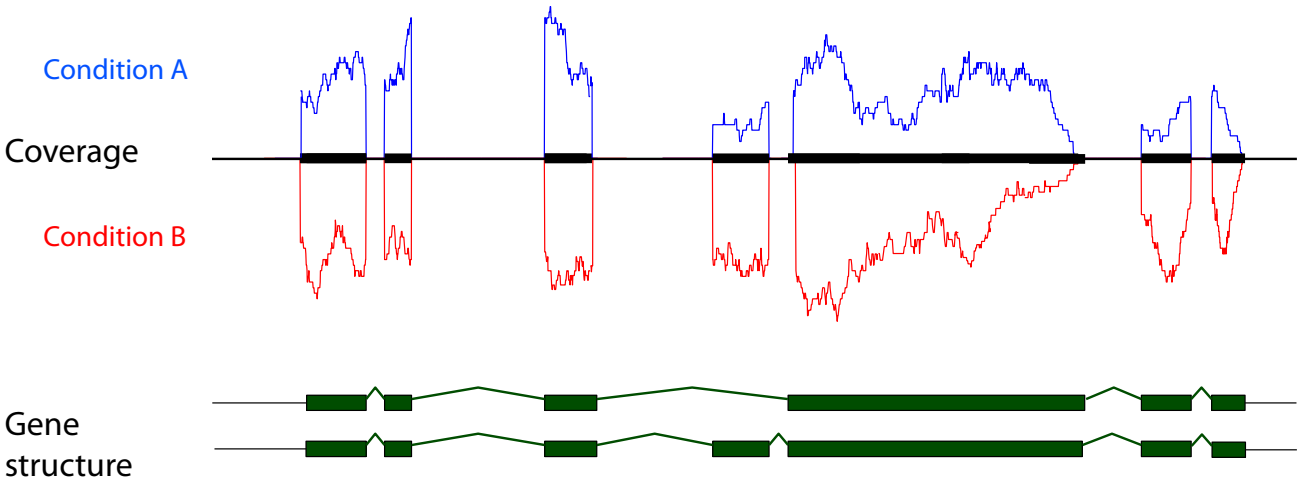


Detecting differential RNA processing

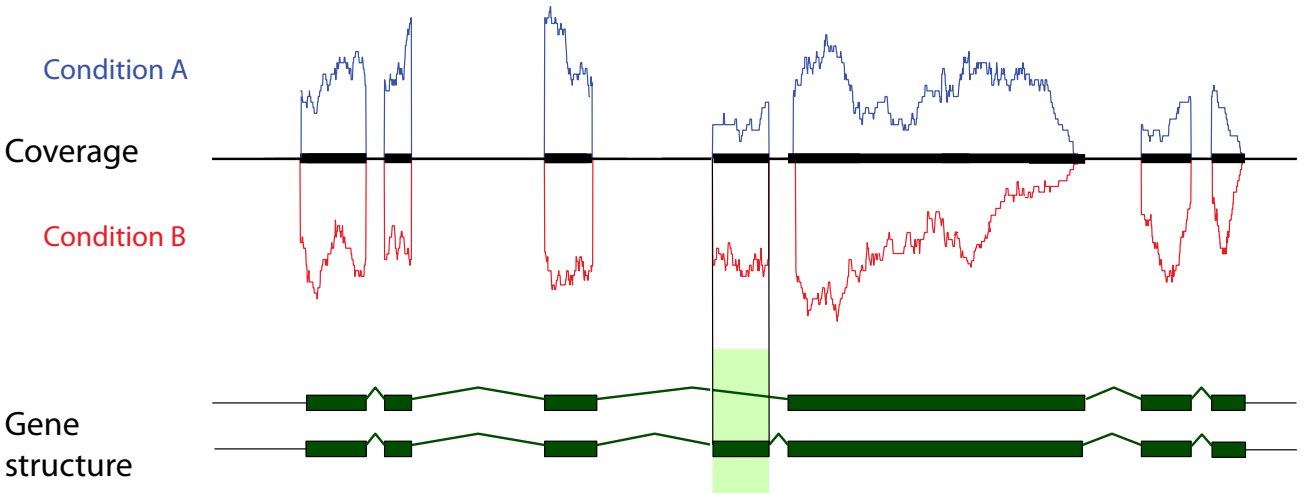


Goal: Design a test to detect differential RNA processing

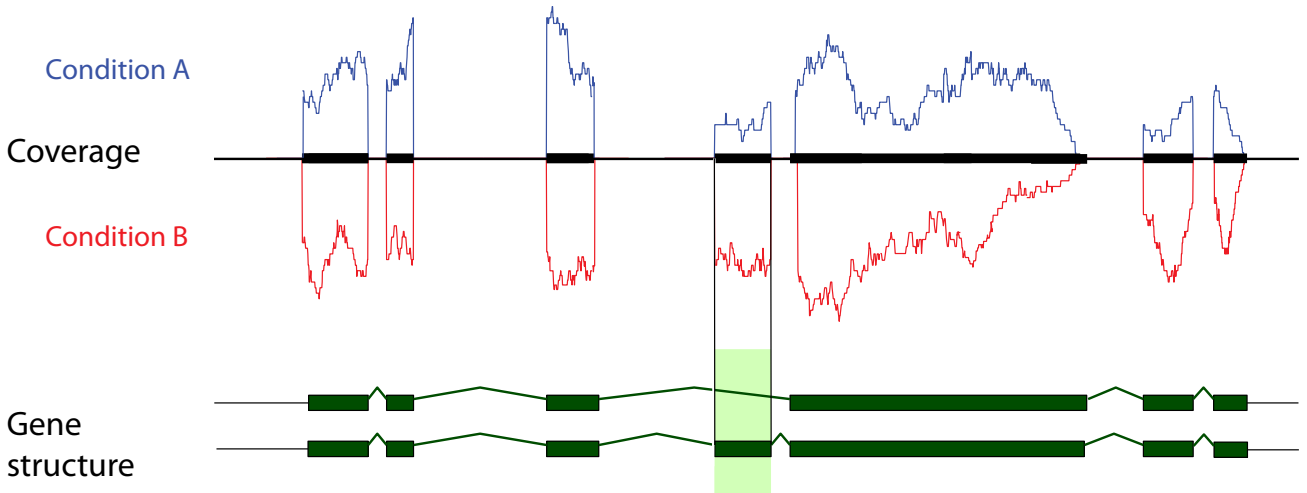
Detecting differential RNA processing



Detecting differential RNA processing



Detecting differential RNA processing



Test whether the relative number of reads than span the exon changes significantly between two conditions

$$PSI = \frac{\text{reads confirming exon}}{\text{total number of reads}}$$

[Wang et al., 2008]

Limits

Limits of this approach

1. **Experimental setup**, such as coverage and the degree of differentiality
2. **Changes in the genome**
3. **Structure of the events** that can be studied

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed

[Drewe et al., 2013]

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed

[Drewe et al., 2013]

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed

[Drewe et al., 2013]

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

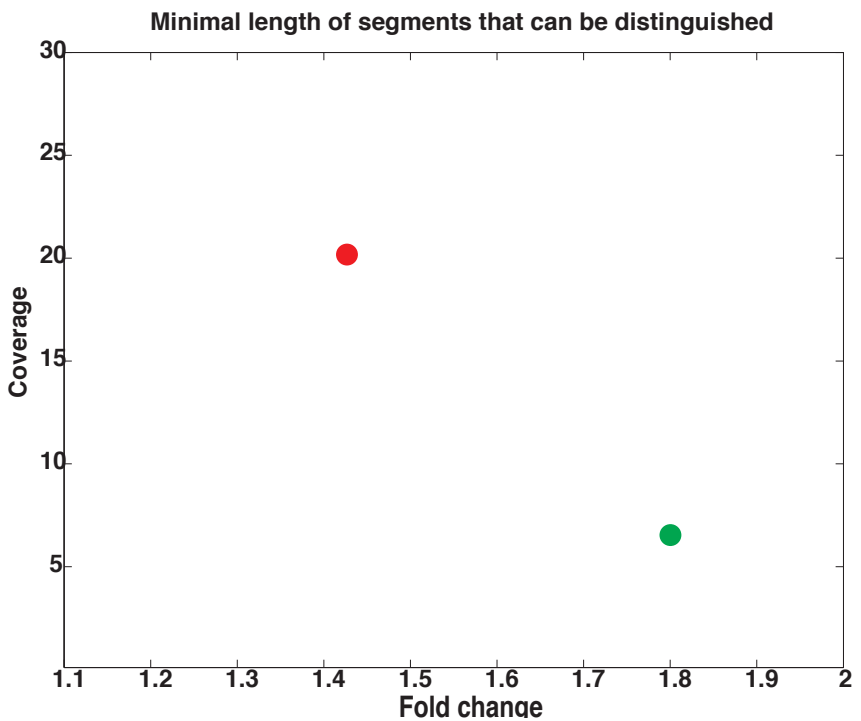
- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed

[Drewe et al., 2013]

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed

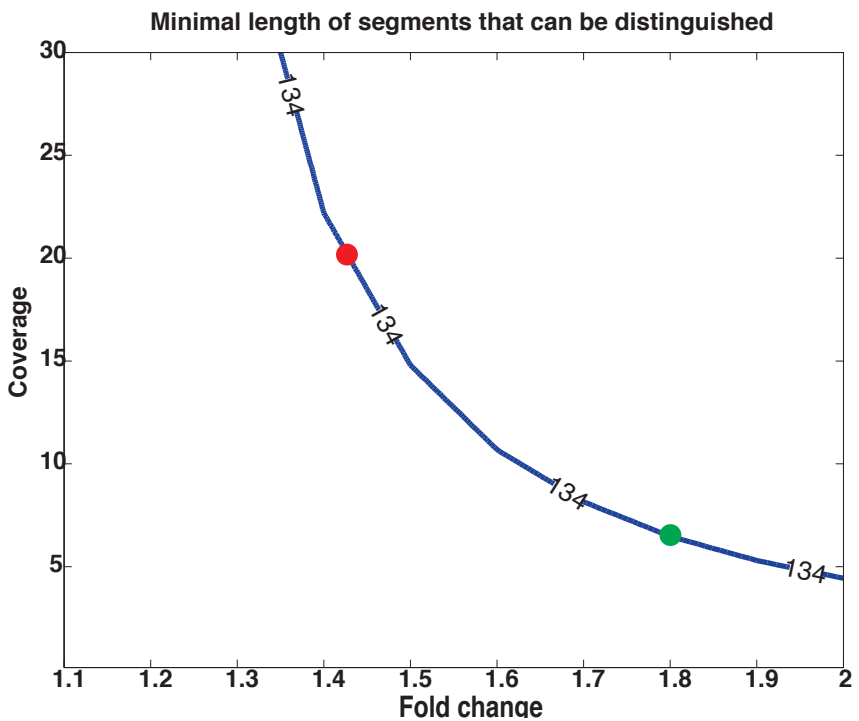


[Drewe et al., 2013]

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed

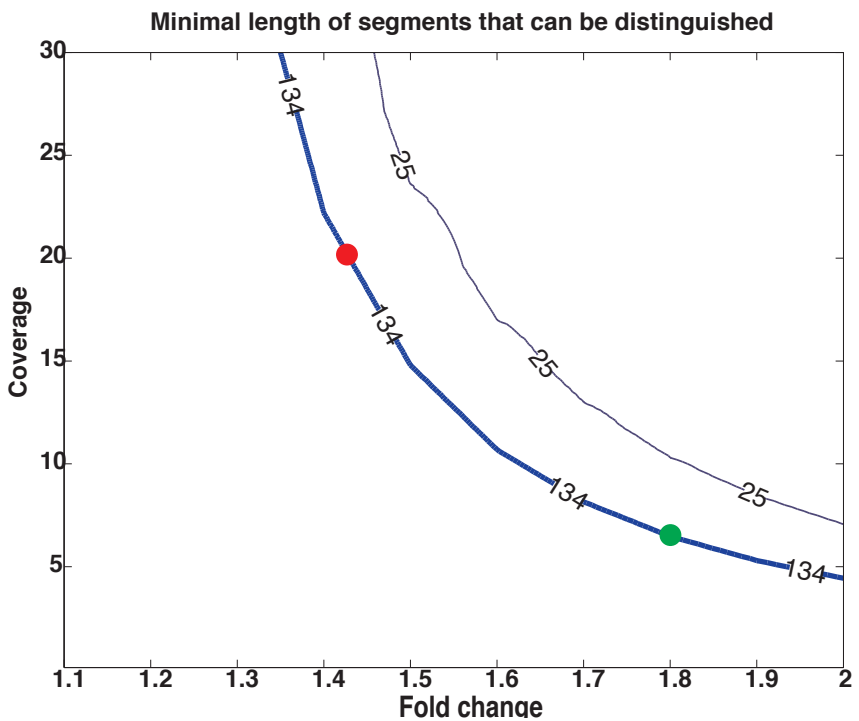


[Drewe et al., 2013]

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed

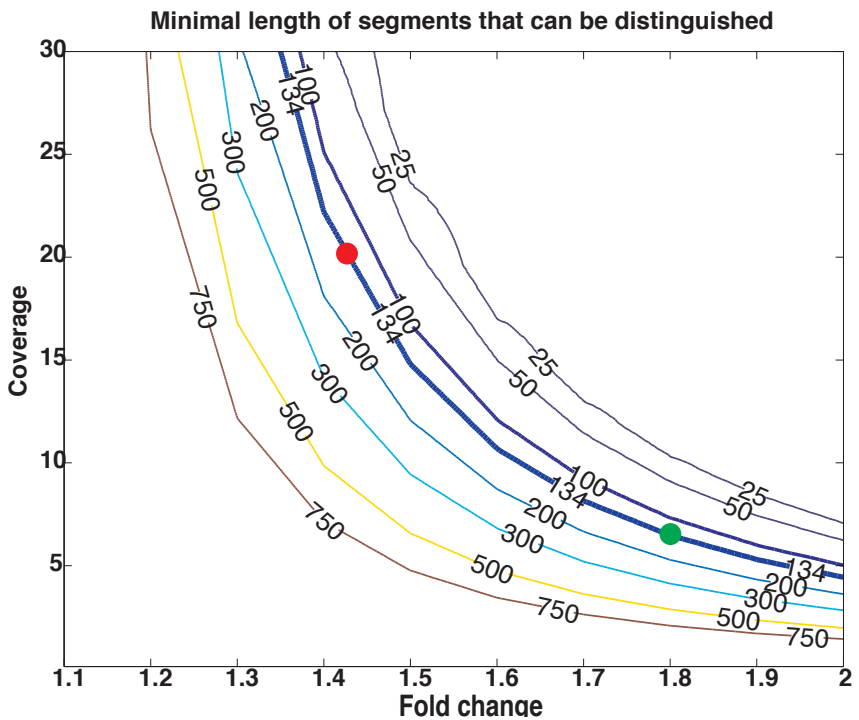


[Drewe et al., 2013]

Limits: Experimental setup

The smallest change in expression that is detectable for a splicing event such as a skipped exon with 75 bp long reads:

- for an exon that is 134 bp long
- when it is from a lowly and a highly expressed gene
- when we assume that reads are Poisson distributed



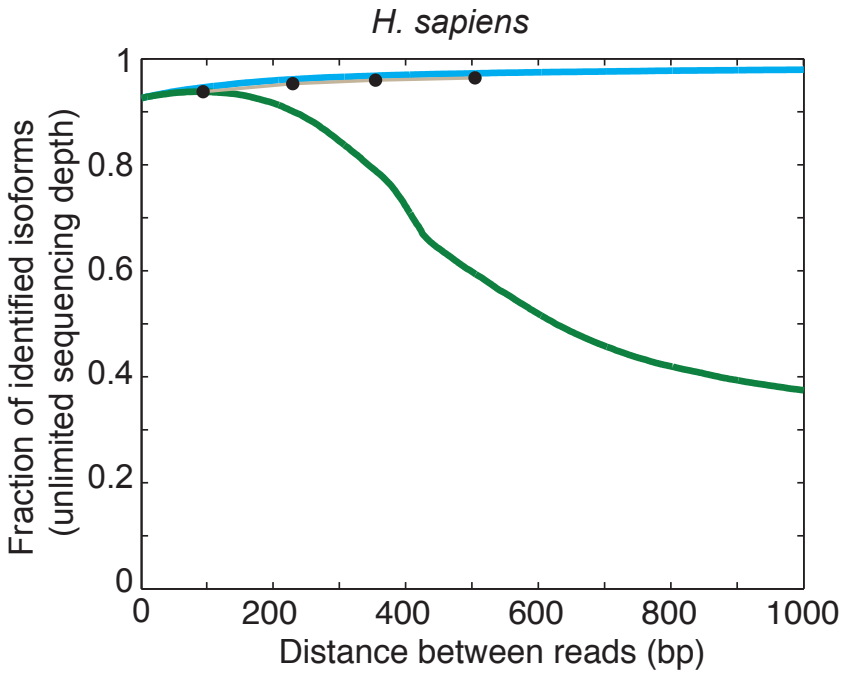
[Drewe et al., 2013]

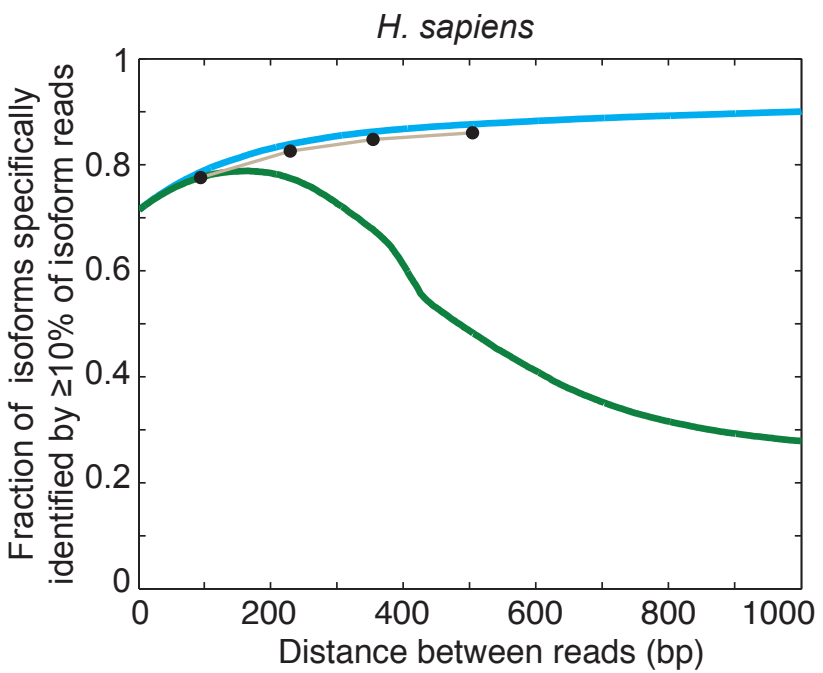
Limits: Experimental setup

What is the effect of insert size in paired-end reads? [Drewe et al., 2013]

Limits: Experimental setup

What is the effect of insert size in paired-end reads?





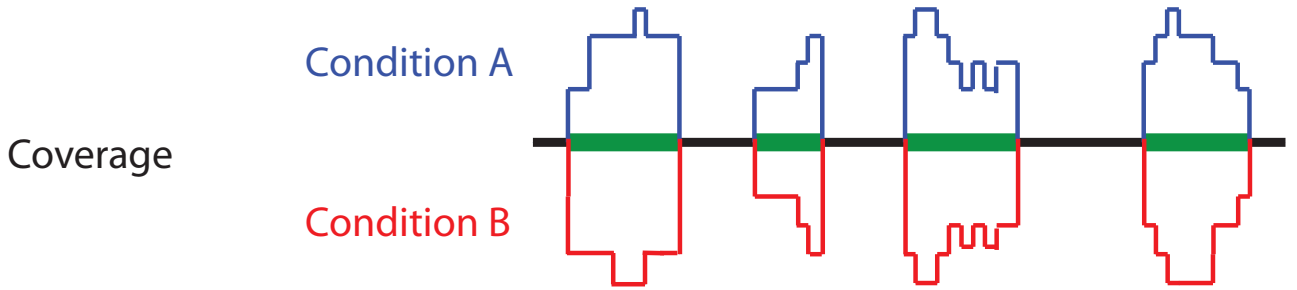
[Drewe et al., 2013]

Limits: Changes in the genome

What happens if the genome or gene structure changes?

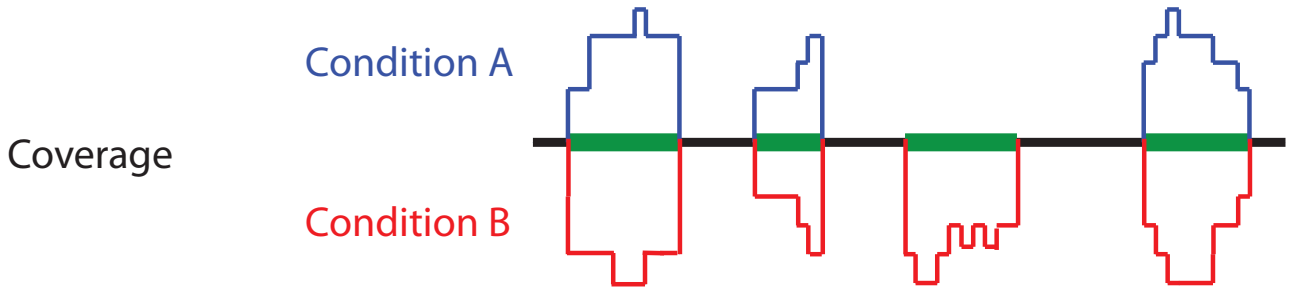
Limits: Changes in the genome

What happens if the genome or gene structure changes?



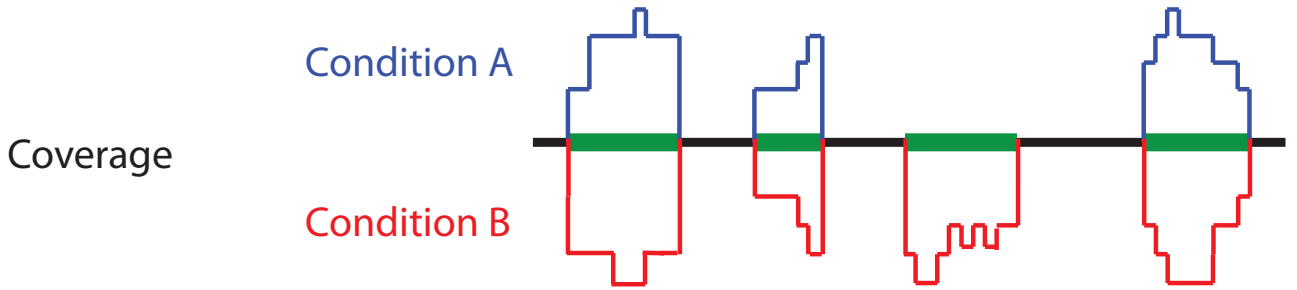
Limits: Changes in the genome

What happens if the genome or gene structure changes?



Limits: Changes in the genome

What happens if the genome or gene structure changes?

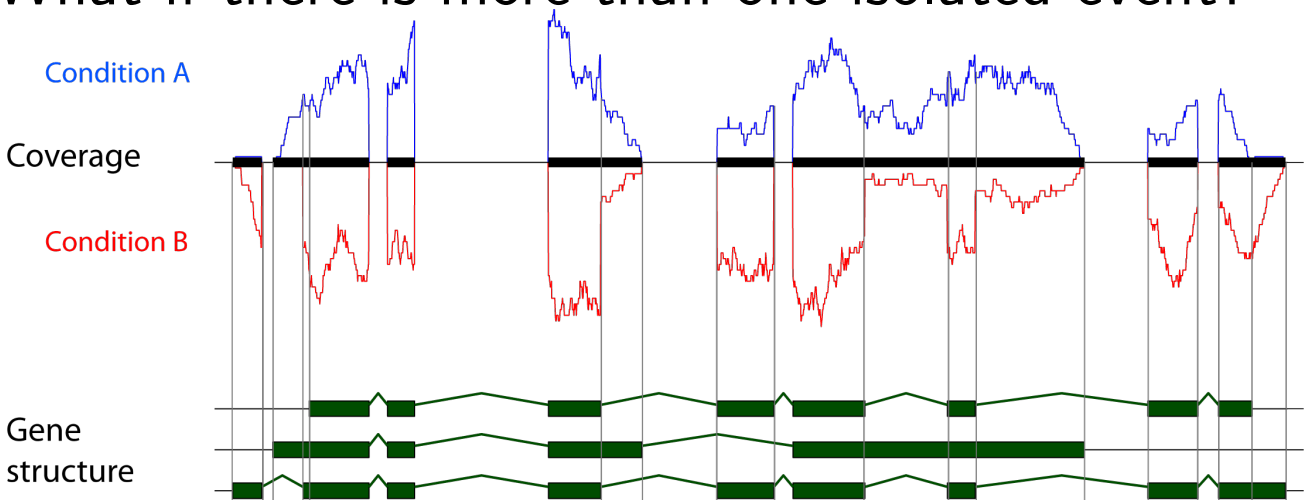


Can lead to many false positives:
45% (64 out of 142) of genes with more than 100-fold differences

[Gan et al., 2011]

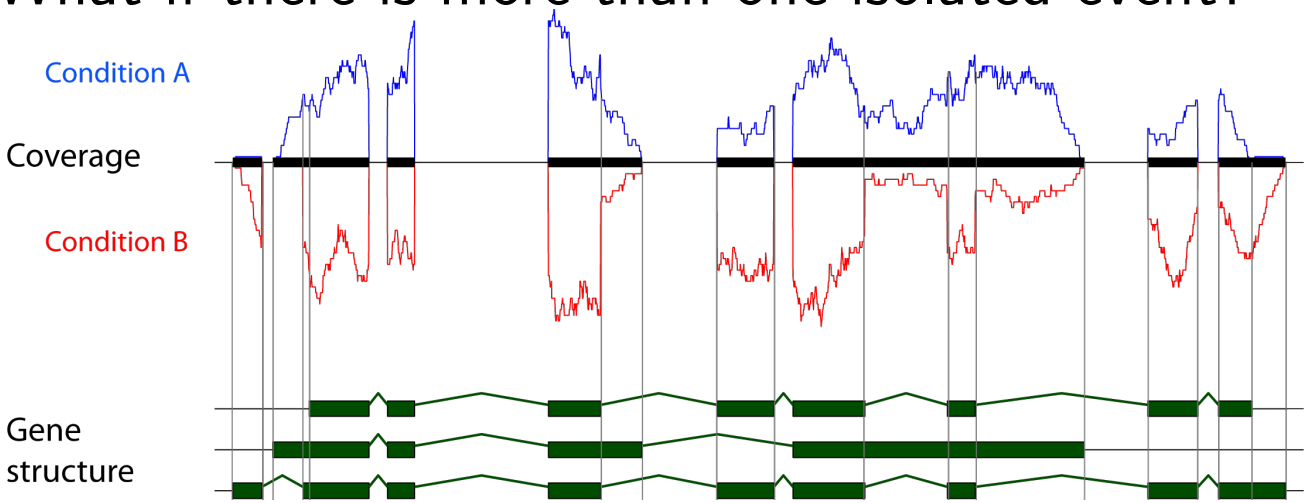
Limits: Structure of events

What if there is more than one isolated event?



Limits: Structure of events

What if there is more than one isolated event?

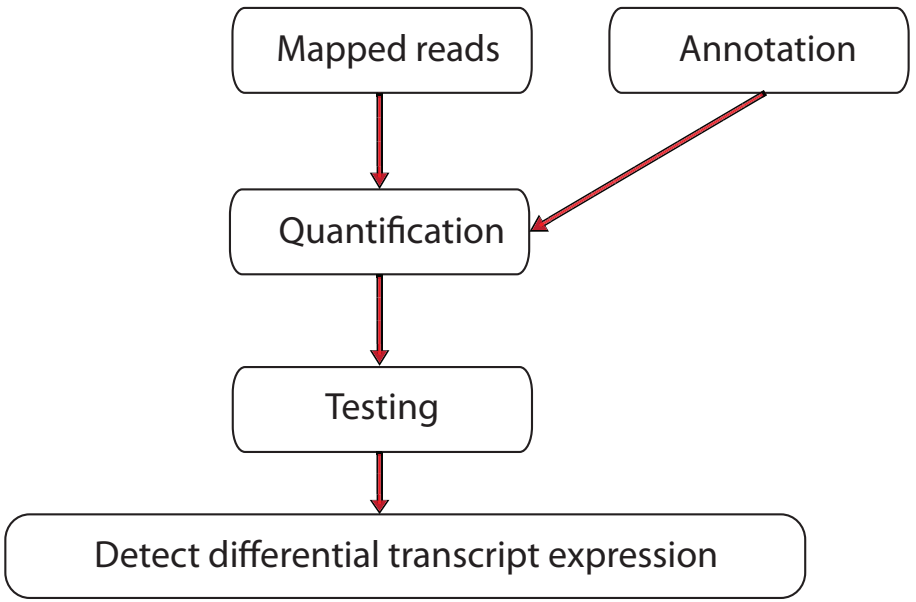


1. Quantification
2. Test on the estimated transcript abundance

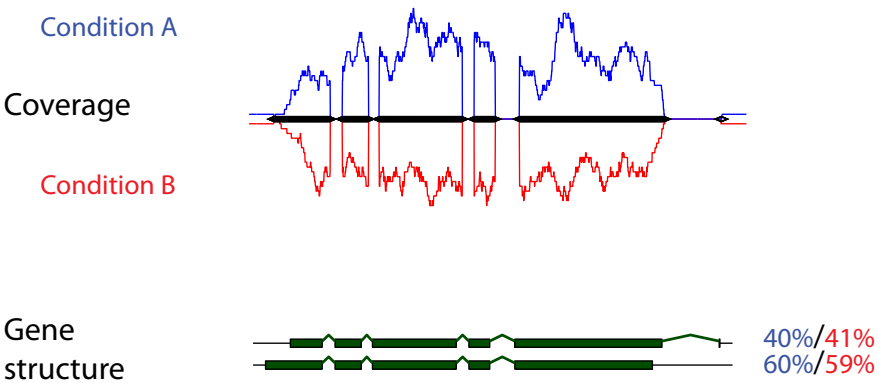
[Richard et al., 2010b, Jiang and Wong, 2009]

Quantification-based methods

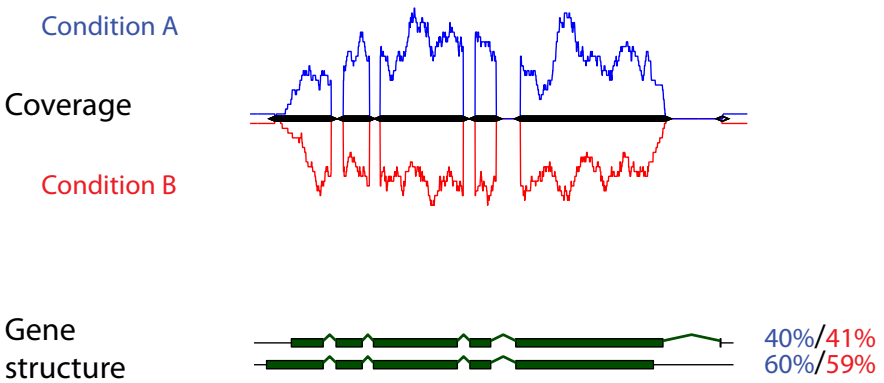
- **Two-step** approach
  1. Quantification
  2. Testing

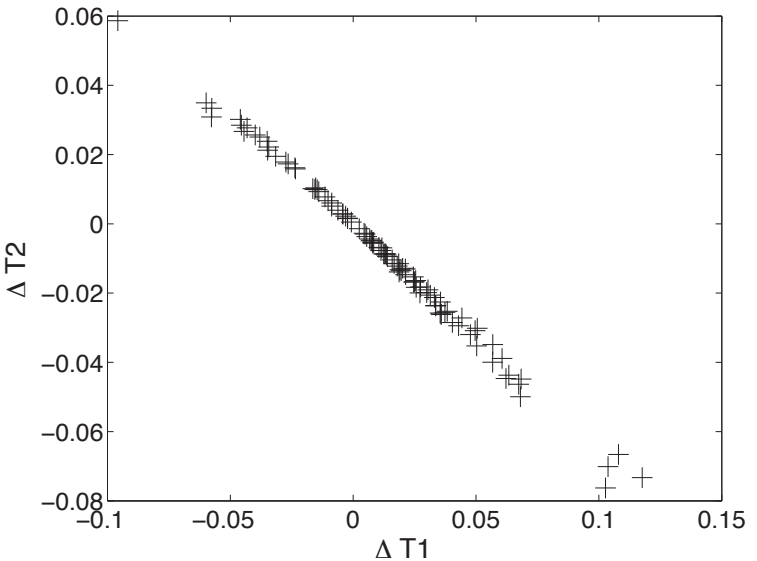


Problems with quantification



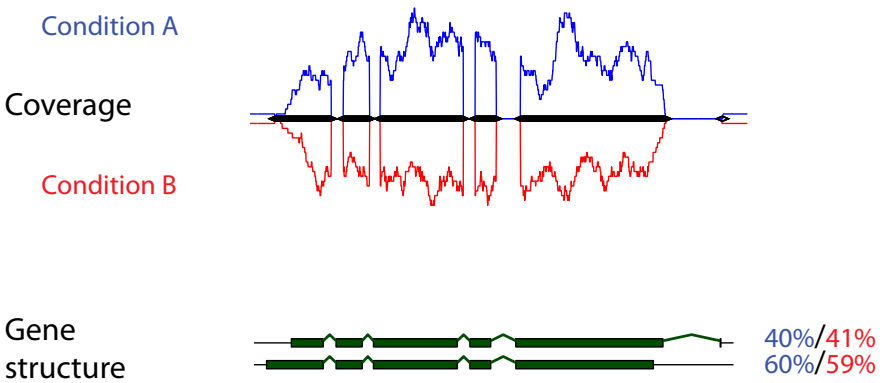
Problems with quantification

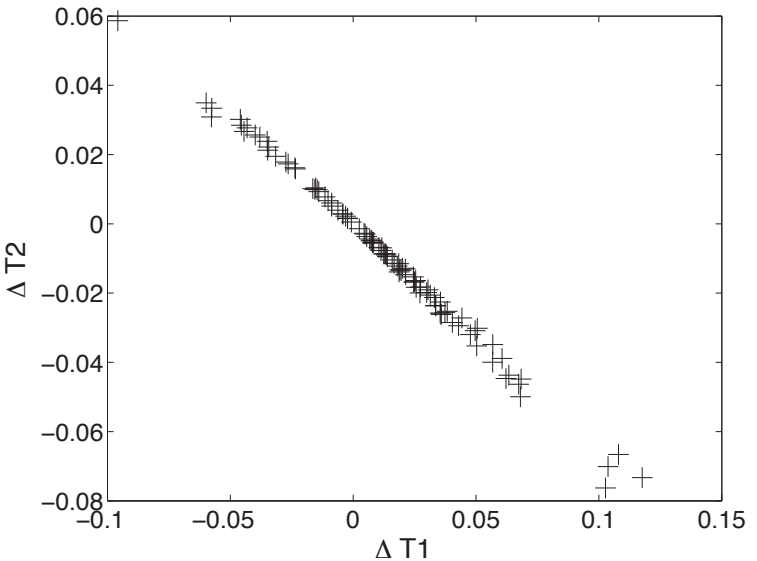




100 perturbed quantification runs

Problems with quantification



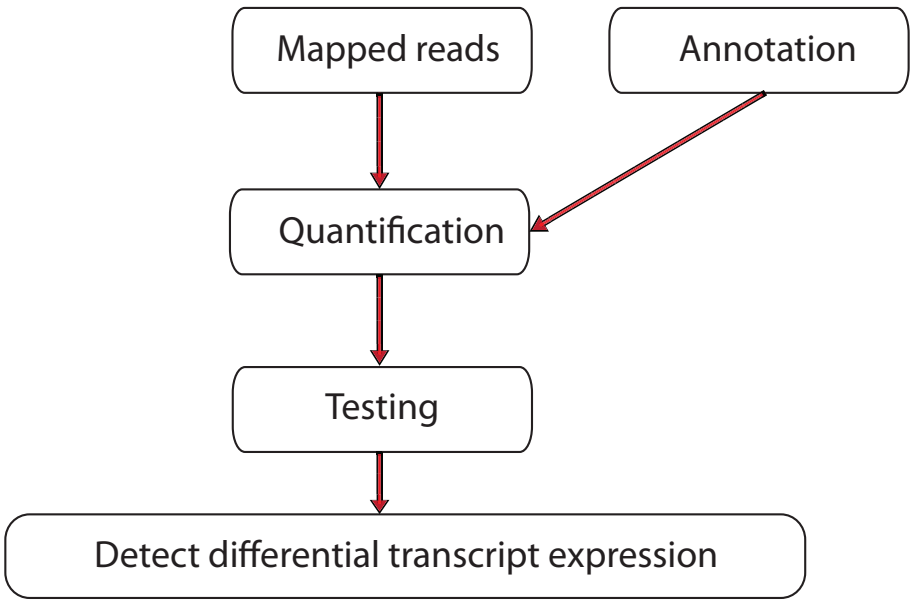


100 perturbed quantification runs

Quantification is challenging

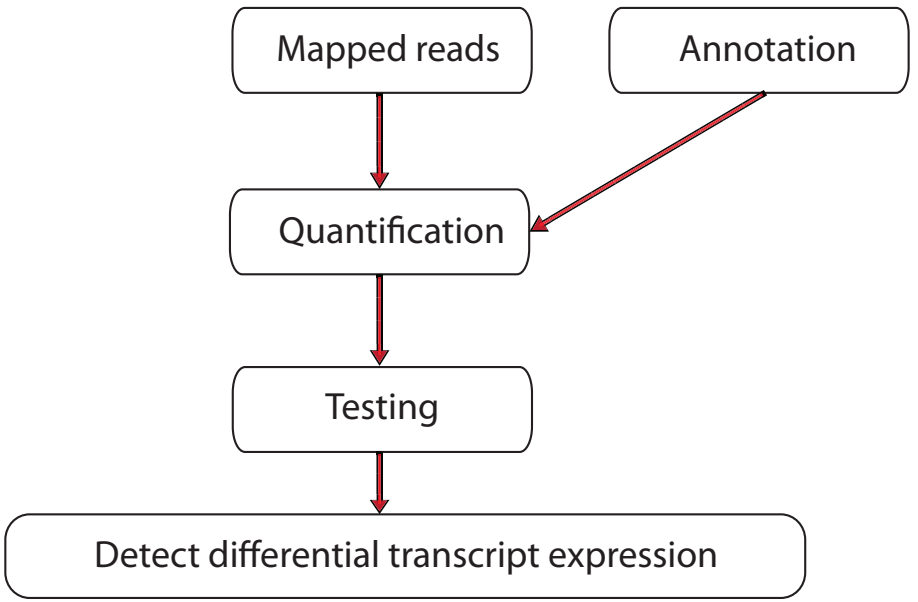
Quantification-based methods

- **Two-step** approach
  1. Quantification
  2. Testing
- Avoid quantification altogether?
  - One-step testing on regions



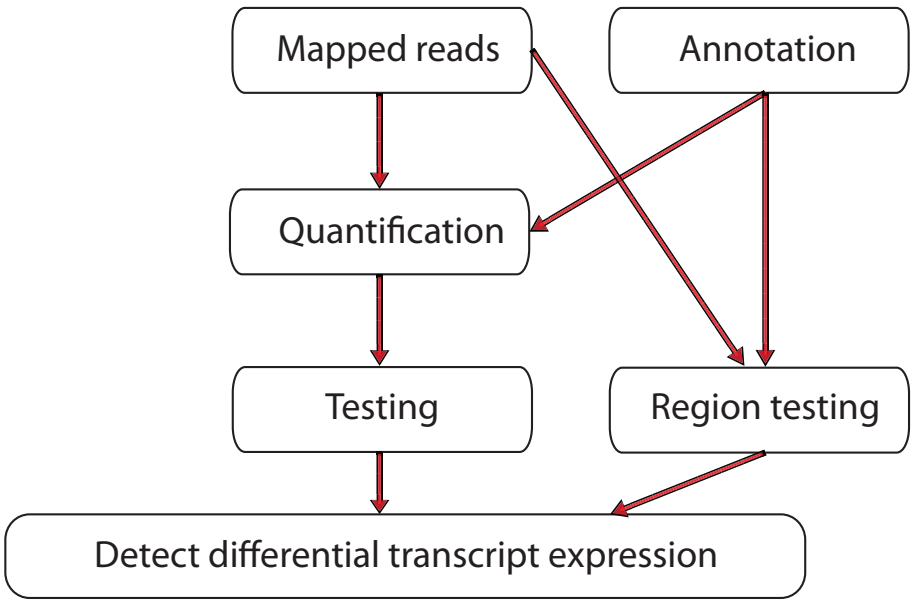
Quantification-based methods

- **Two-step** approach
  1. Quantification
  2. Testing
- Avoid quantification altogether?
  - **One-step** testing on regions

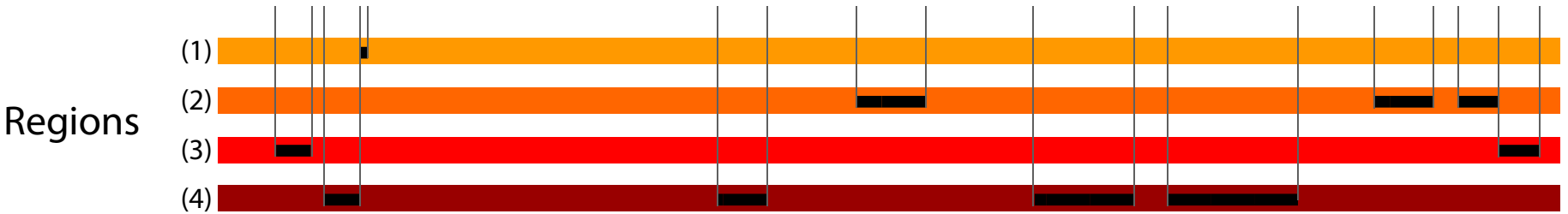


Quantification-based methods

- **Two-step** approach
  1. Quantification
  2. Testing
- Avoid quantification altogether?
  - **One-step** testing on regions



Testing on discriminative regions

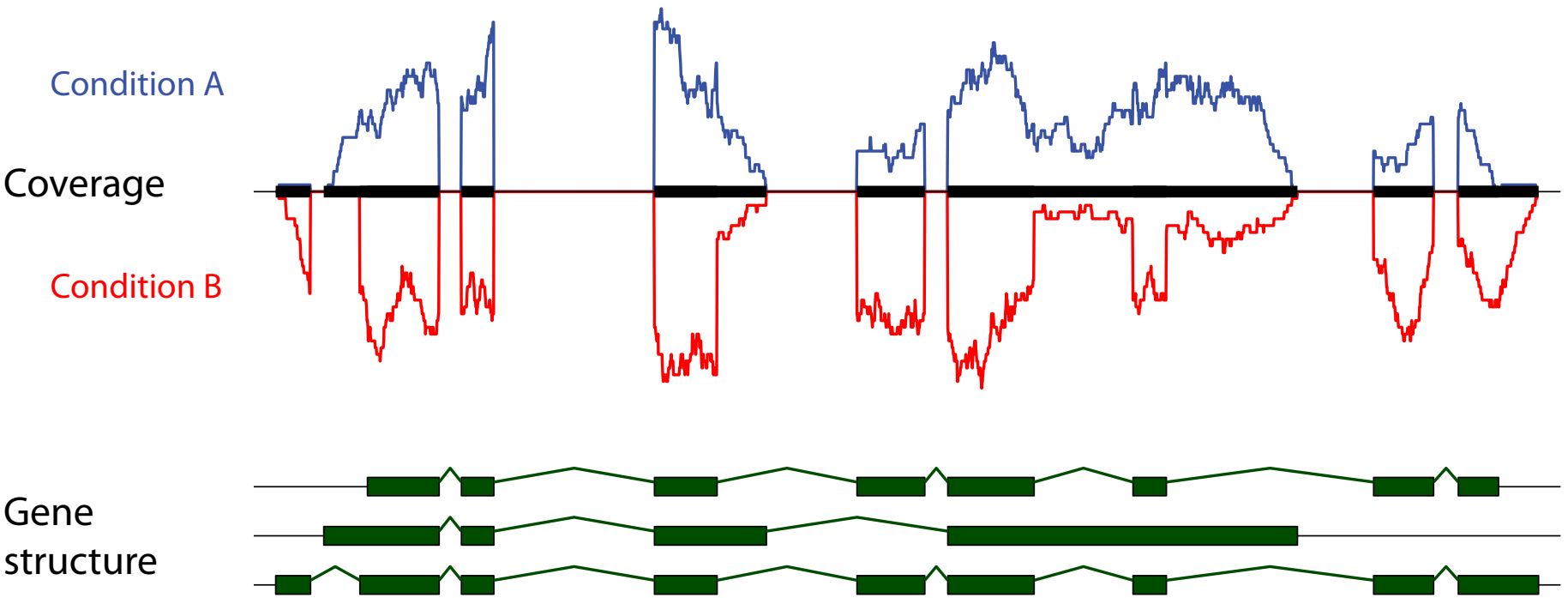


- Use read counts per discriminative region r_1, r_2, r_3, r_4 for testing
- The number of reads falling into a certain region r in condition A and B are **Negative binomial distributed**: $N_r^{A/B} \sim \text{NB}\left(\lambda_r^{A/B}\right)$.
- **Test-statistic** can be derived from

$$P\left(\lambda_r^A = \lambda_r^B | N_r^A, N_r^B\right) = \frac{\sum_{k+l=N_r} p(k,l) I_{p(k,l) \leq p(N_r^A, N_r^B)}}{\sum_{k+l=N_r} p(k,l)}$$

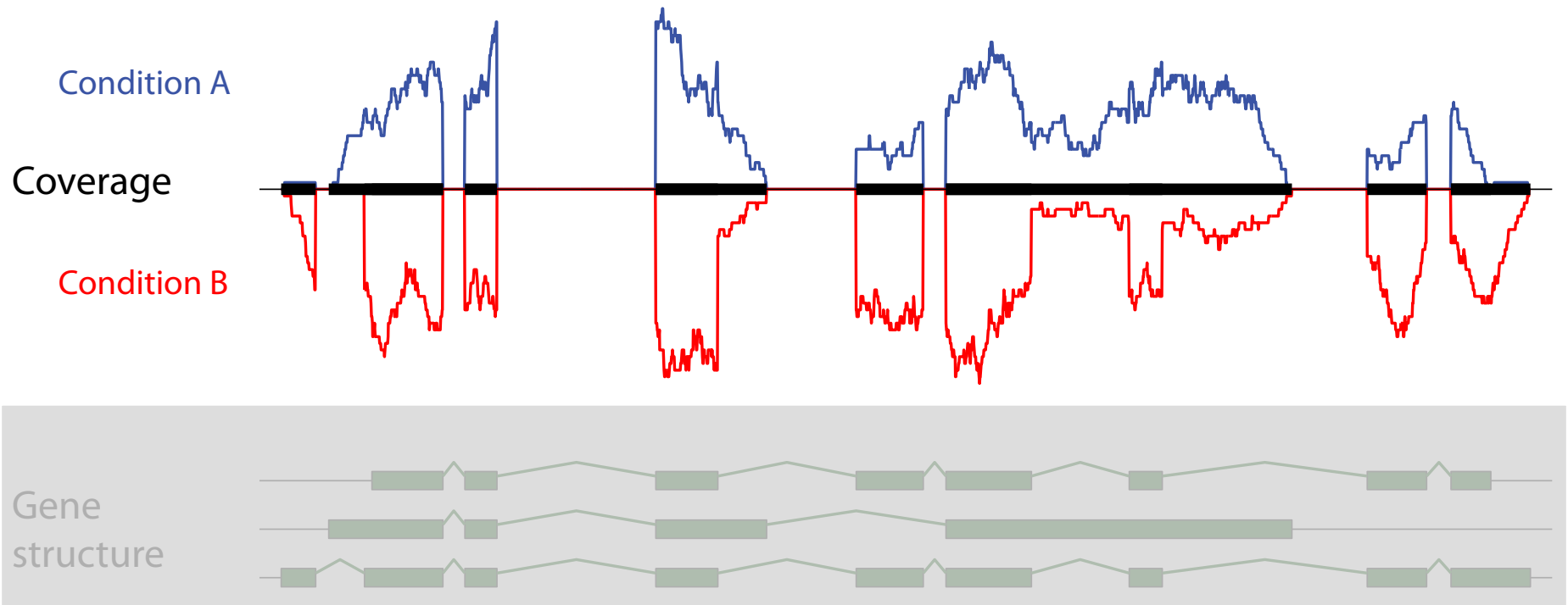
- Correction for multiple testing using **Bonferroni correction**
- Implemented as **rDiff.parametric**

Testing without gene structure



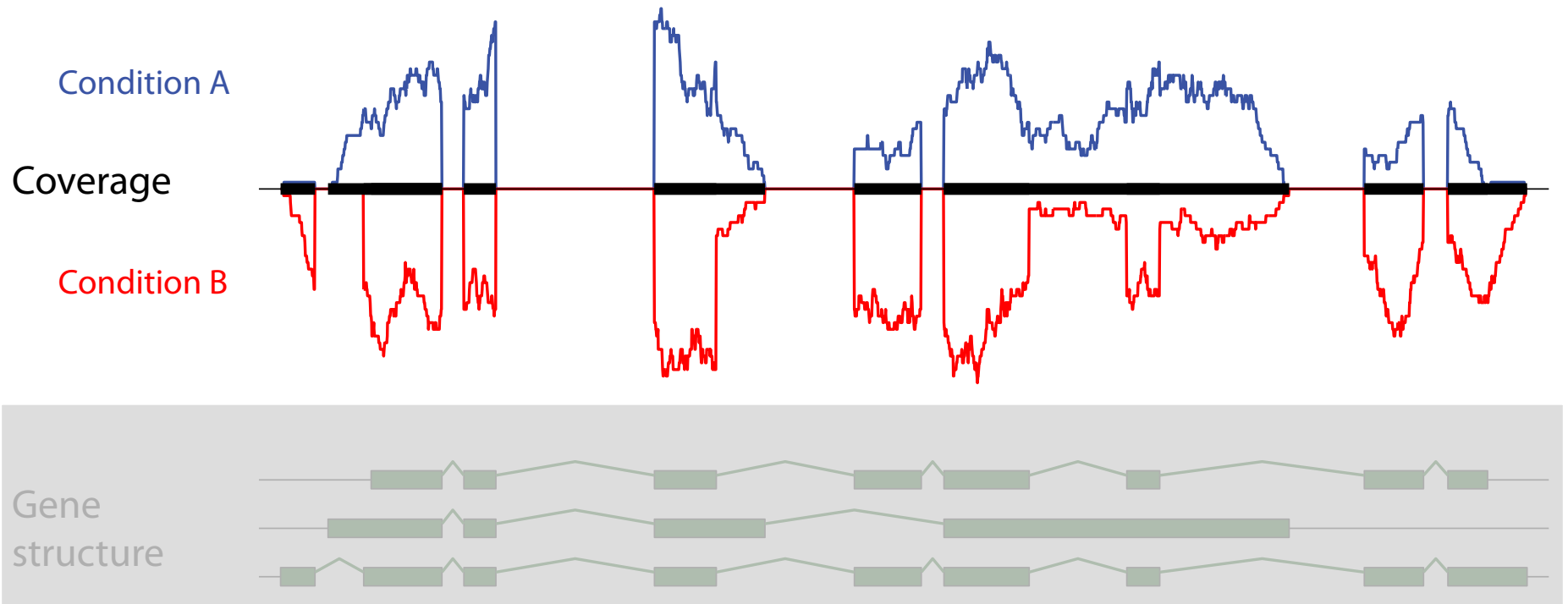
- Complete gene annotation is not always available
  - Antisense transcription
  - Unknown splice forms
- How can we test in the absence or with incomplete annotation?

Testing without gene structure



- Complete gene annotation is not always available
  - Antisense transcription
  - Unknown splice forms
- How can we test in the absence or with incomplete annotation?

Testing without gene structure

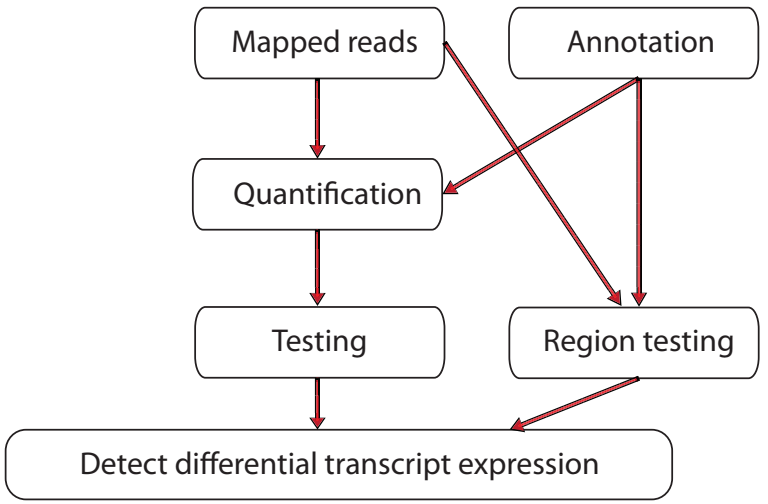


- Complete gene annotation is not always available
  - Antisense transcription
  - Unknown splice forms
- How can we test in the absence or with incomplete annotation?

Testing without gene structure

Possible approaches

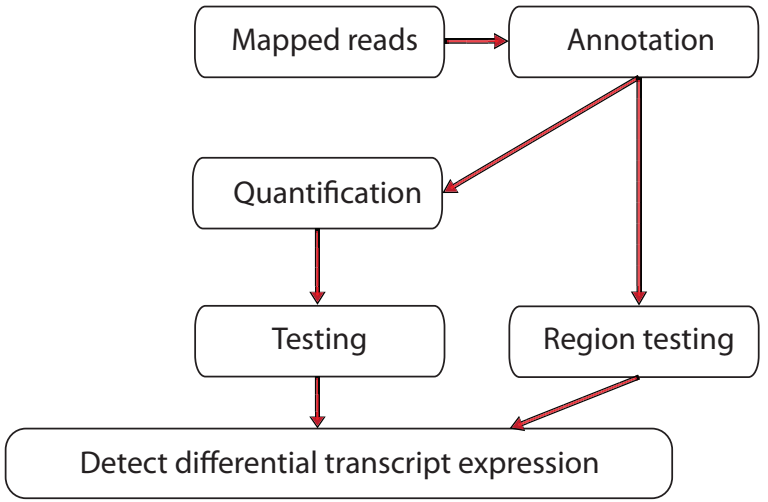
1. Include transcript identification in the analysis
   - Quantification
   - Region testing
   - Complex!
2. **One-step** testing on the read densities.



Testing without gene structure

Possible approaches

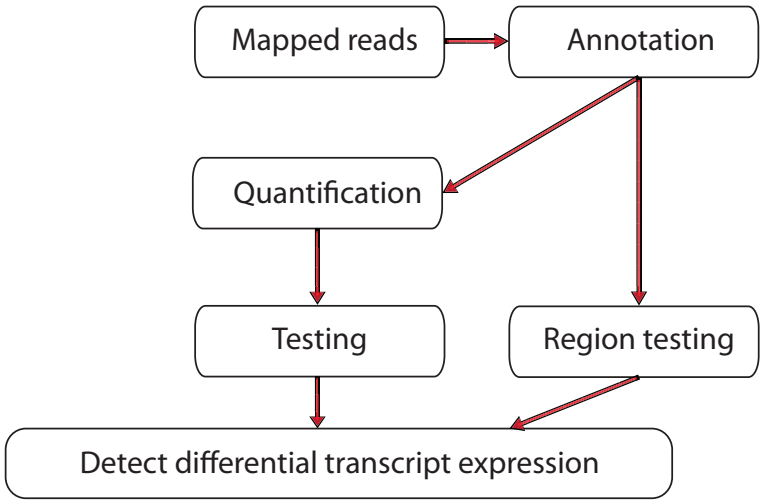
1. Include transcript identification in the analysis
   - Quantification
   - Region testing
   - Complex!
2. **One-step** testing on the read densities.



Testing without gene structure

Possible approaches

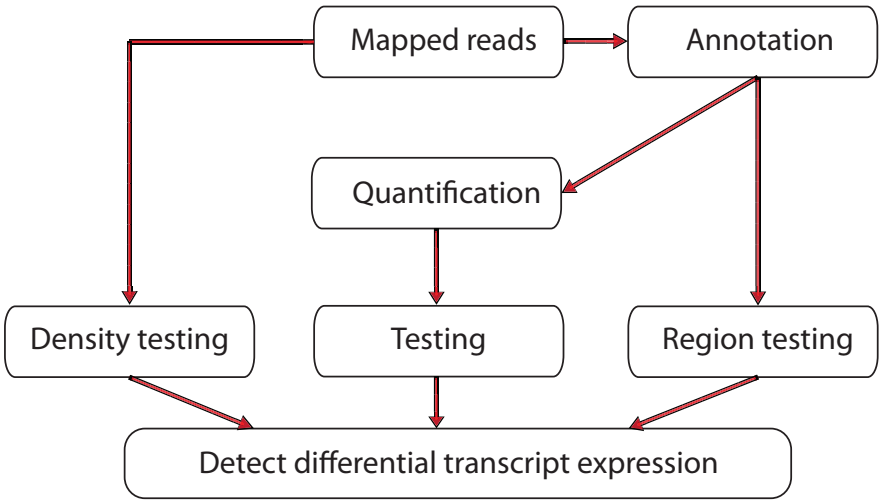
1. Include transcript identification in the analysis
   - Quantification
   - Region testing
   - Complex!
2. **One-step** testing on the read densities.



Testing without gene structure

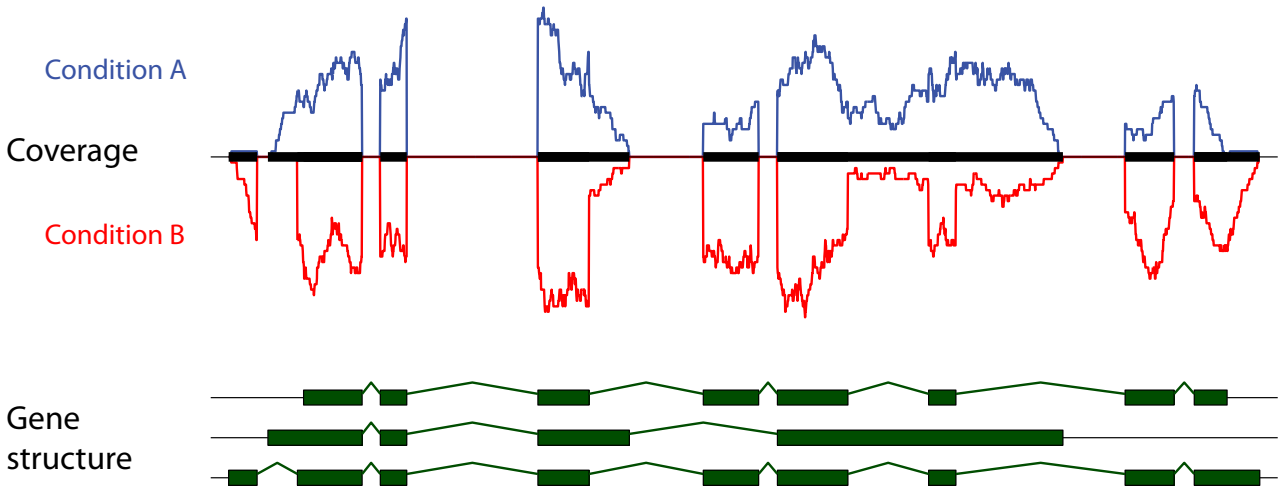
Possible approaches

1. Include transcript identification in the analysis
   - Quantification
   - Region testing
   - Complex!
2. **One-step** testing on the read densities.



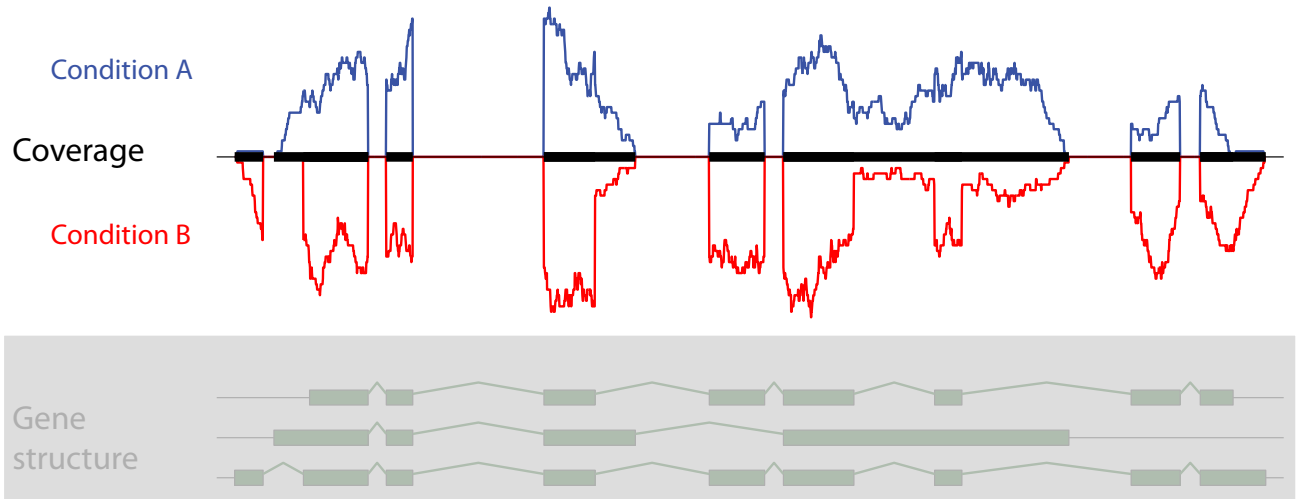
Testing without gene structure

What if the gene structure is **incomplete or unknown**?



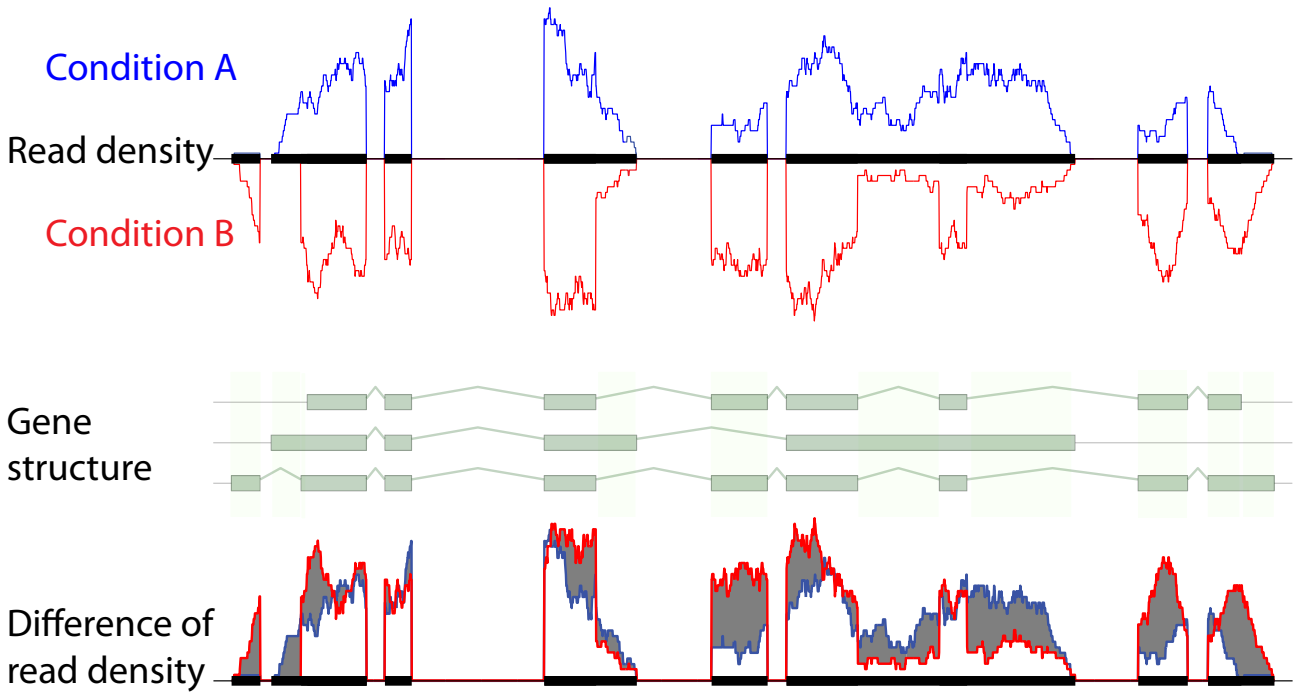
Testing without gene structure

What if the gene structure is **incomplete or unknown**?



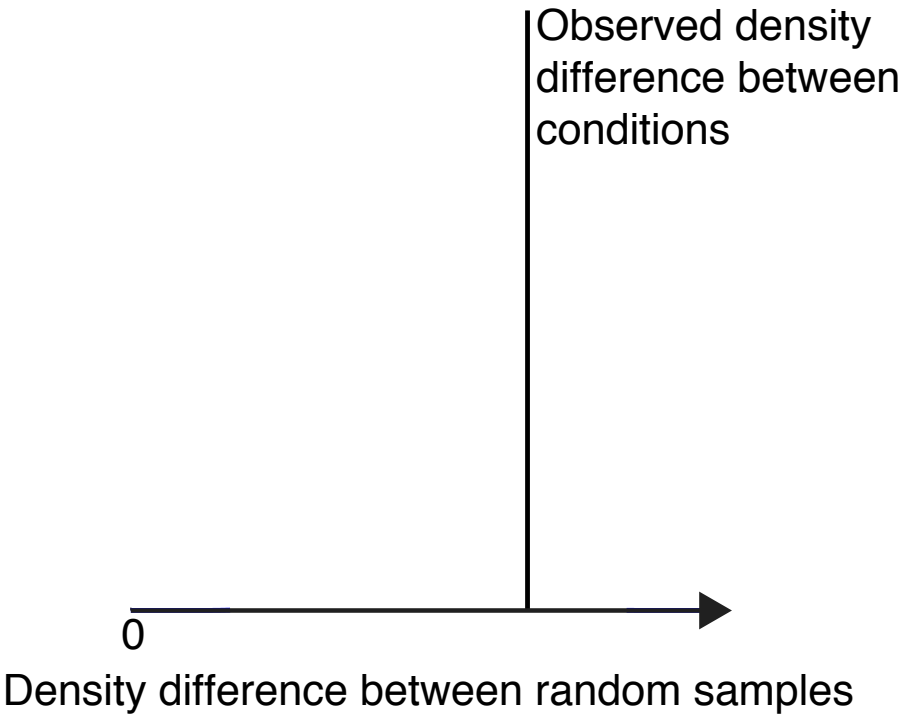
Testing without gene structure

What if the gene structure is **incomplete or unknown**?



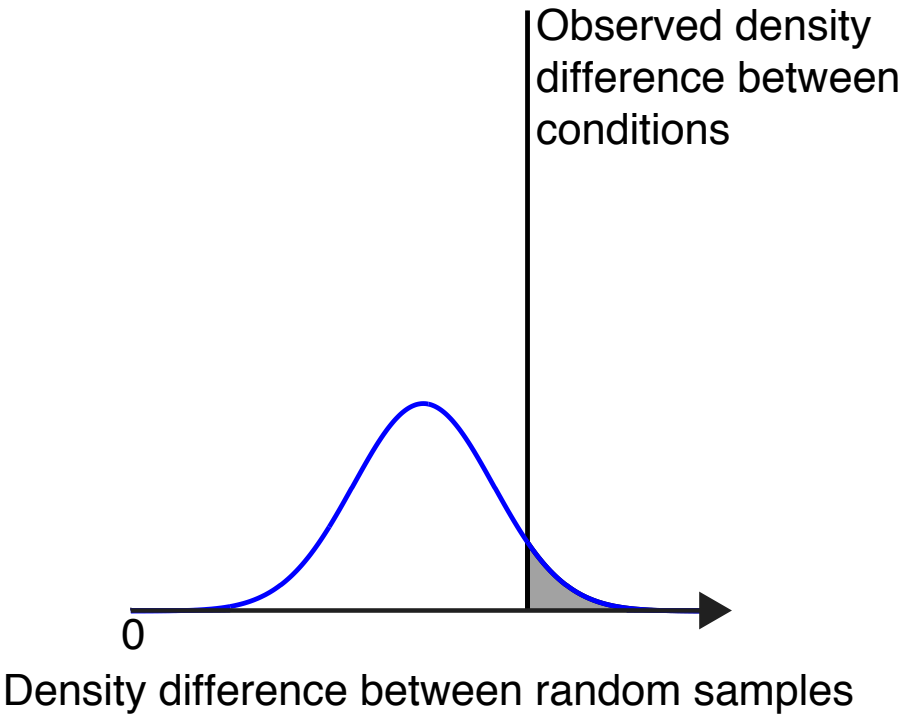
Testing without gene structure

What if the gene structure is **incomplete or unknown**?



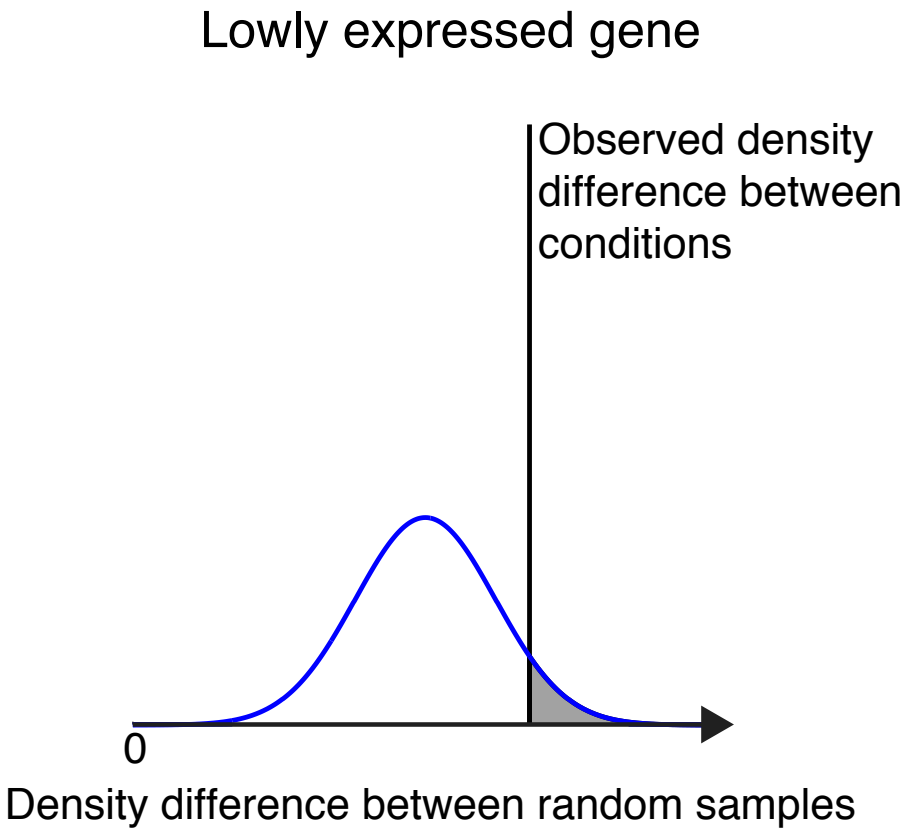
Testing without gene structure

What if the gene structure is **incomplete or unknown**?



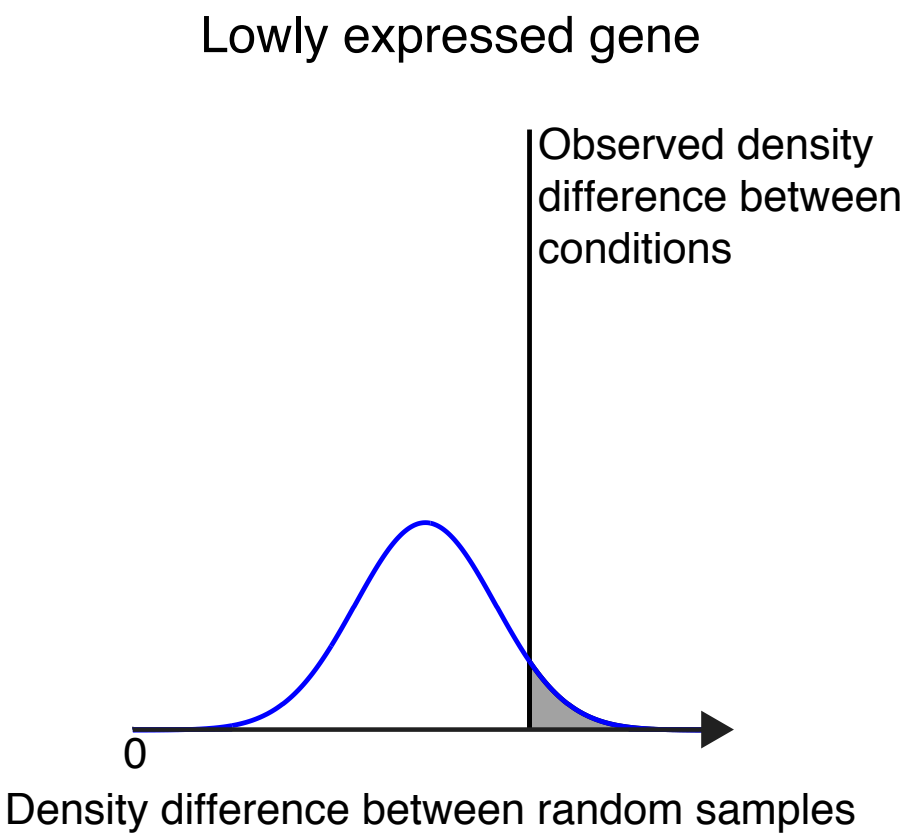
Testing without gene structure

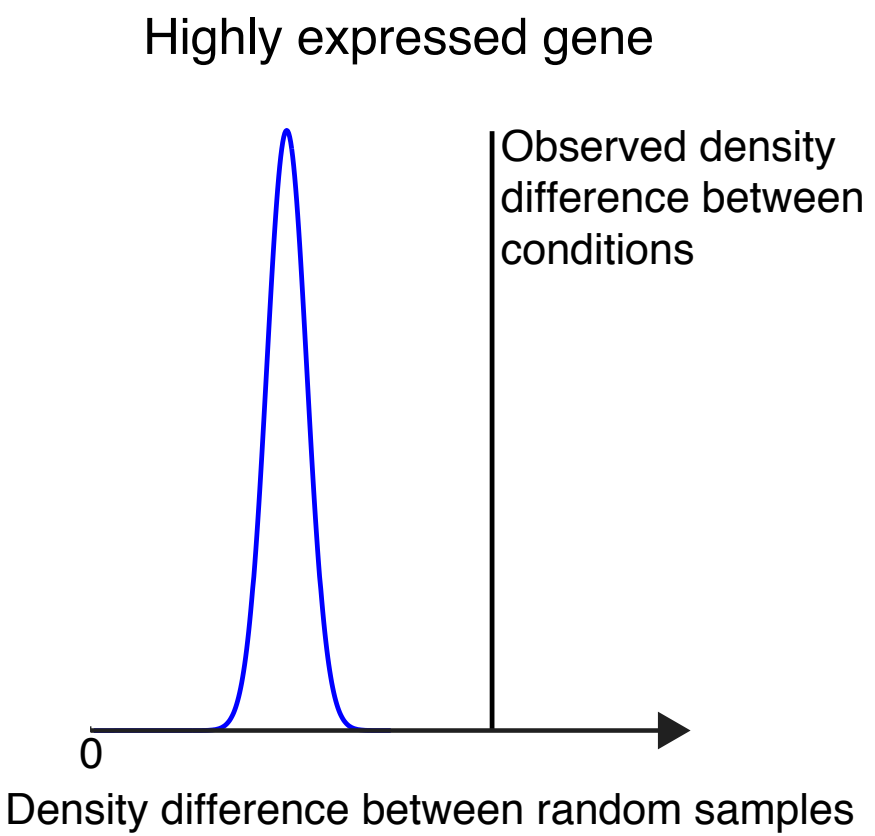
What if the gene structure is **incomplete or unknown**?



Testing without gene structure

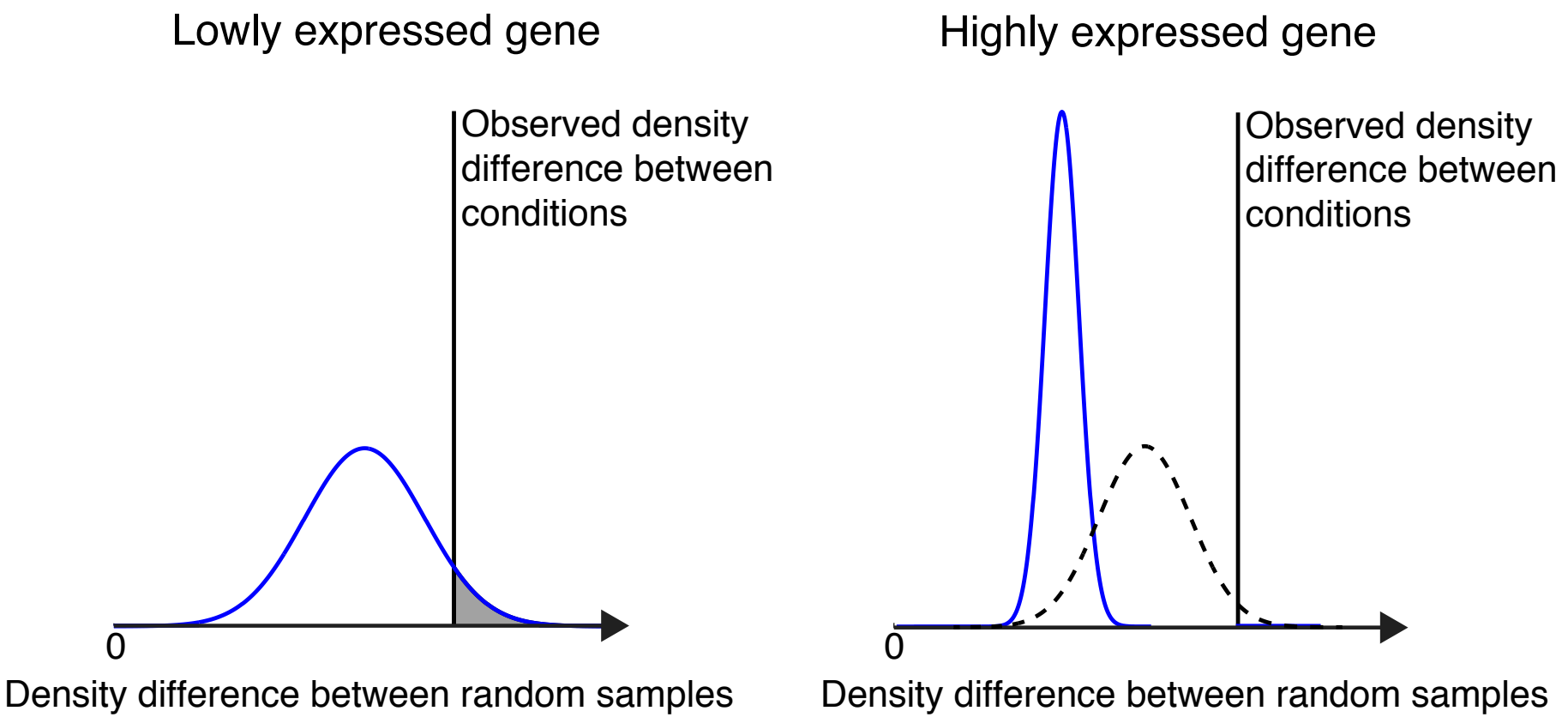
What if the gene structure is **incomplete or unknown**?





Testing without gene structure

What if the gene structure is **incomplete or unknown**?



Testing without gene structure

What if the gene structure is **incomplete or unknown**?

- The distribution of the coverage C_p at a position p follows a random sample with a hypergeometric distribution.

Testing without gene structure

What if the gene structure is **incomplete or unknown**?

- The distribution of the coverage C_p at a position p follows a random sample with a hypergeometric distribution.
- The variance $\sigma^2_{C_P}(n)$ is a function of the subsample size n

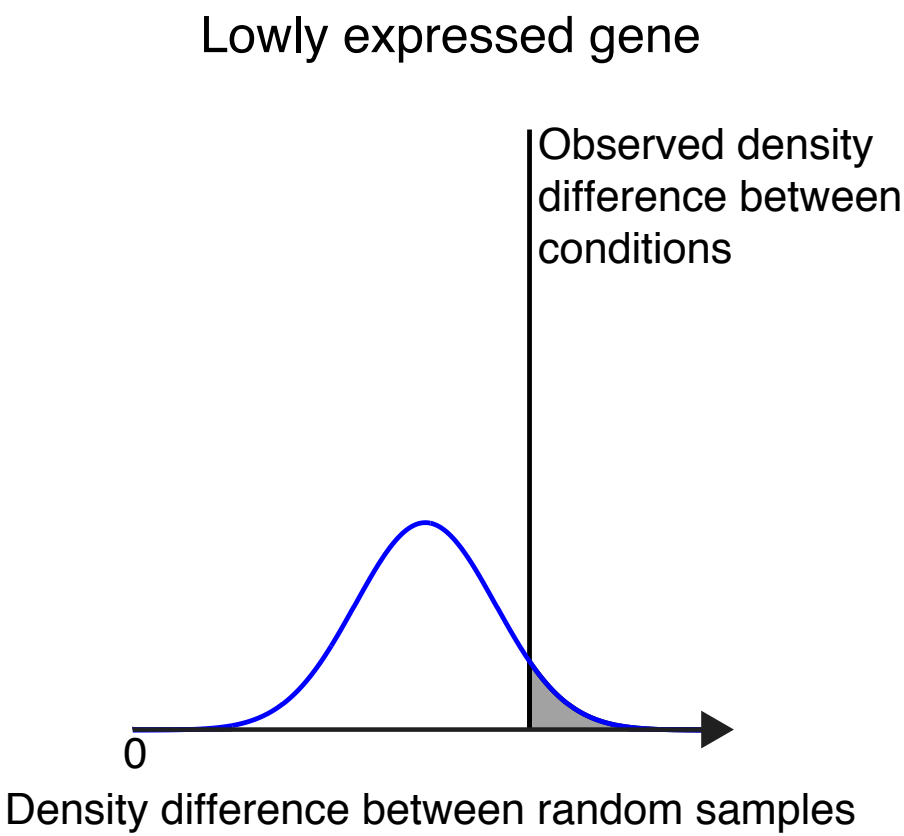
Testing without gene structure

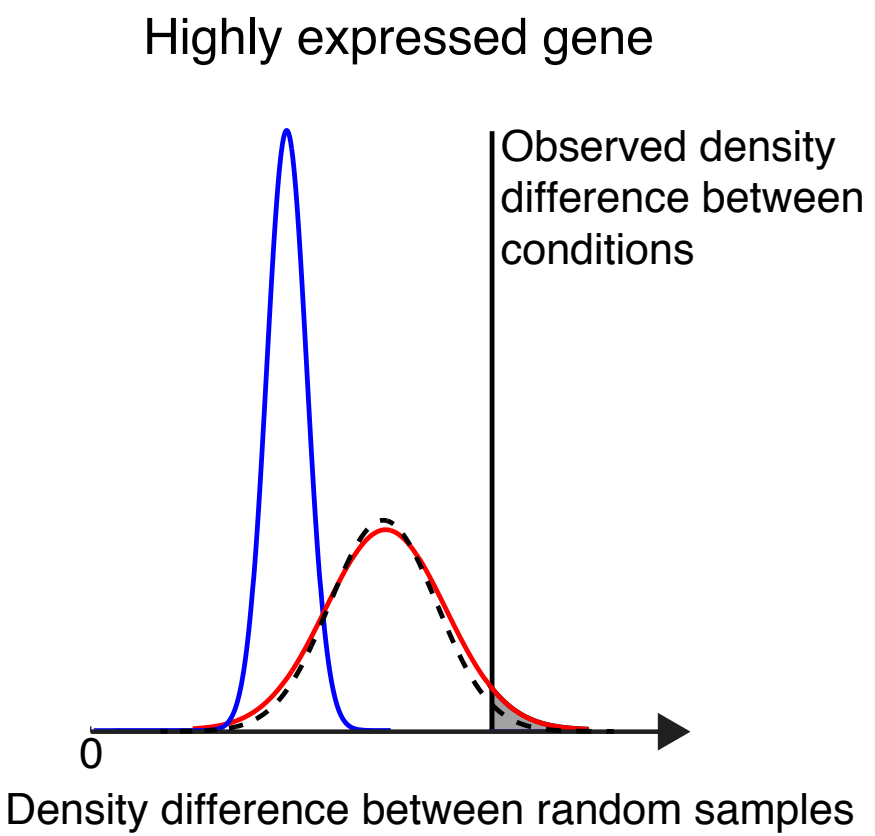
What if the gene structure is **incomplete or unknown**?

- The distribution of the coverage C_p at a position p follows a random sample with a hypergeometric distribution.
- The variance $\sigma^2_{C_P}(n)$ is a function of the subsample size n
- We can choose n such that: $\sigma^2_{C_P}(n) = \sigma_{biol}$

Testing without gene structure

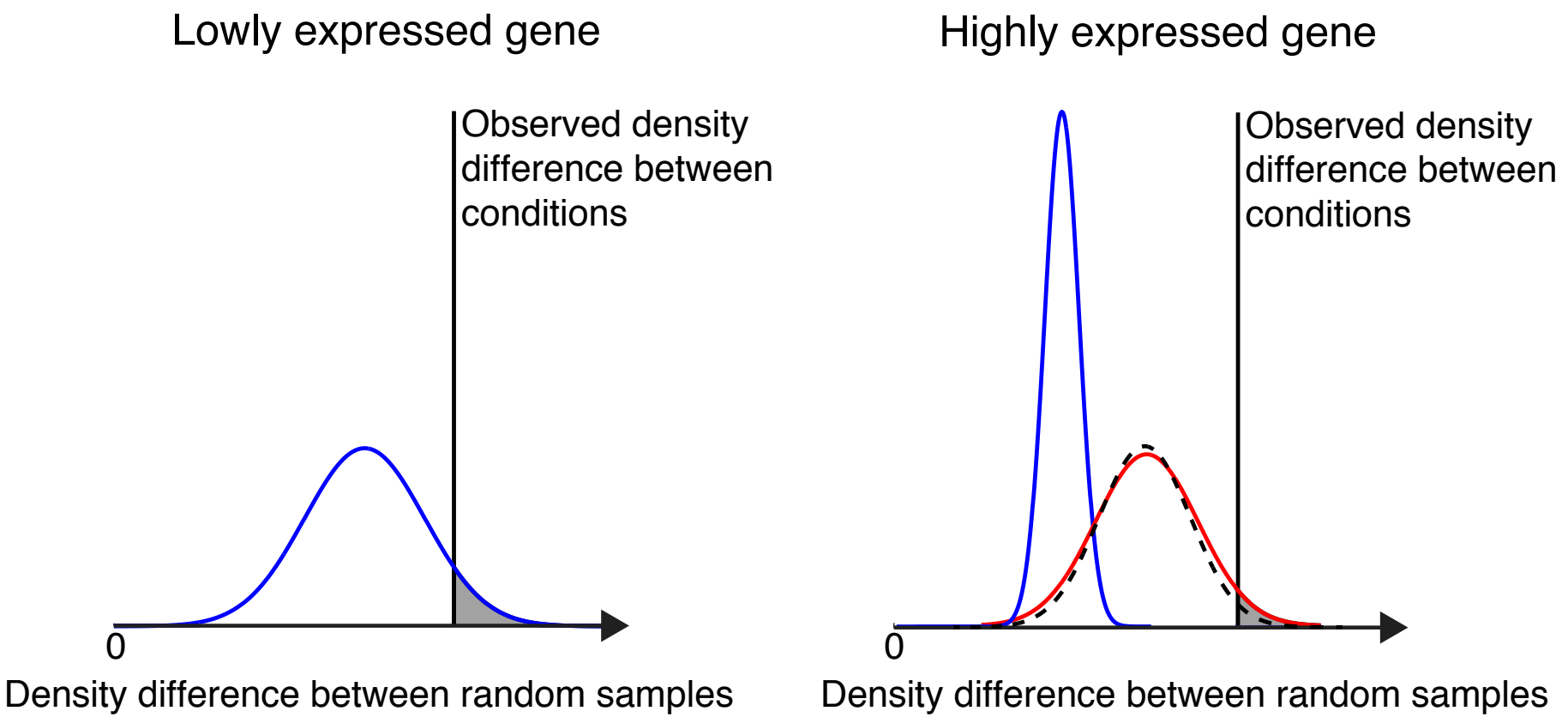
What if the gene structure is **incomplete or unknown**?





Testing without gene structure

What if the gene structure is **incomplete or unknown**?



Implemented as **rDiff.mmd**

Evaluation: Simulated reads

Two datasets:

- 5,875 mRNA coding genes from *A.thaliana* with **multiple** annotated alternative transcripts (TAIR10)
- Measured biological variance and gene expression from real data:
  - Between very similar samples (**small biol. variance**)
  - Between close time points (**large biol. variance**)
- Simulated transcript abundance levels in **two conditions**: differential and non-differentially expressed, with biological replicates
- Reads simulation based on Flux Simulator

[Sammeth, 2009b]

Evaluation: Simulated reads

Two datasets:

- 5,875 mRNA coding genes from *A.thaliana* with **multiple** annotated alternative transcripts (TAIR10)
- Measured biological variance and gene expression from real data:
  - Between very similar samples (**small biol. variance**)
  - Between close time points (**large biol. variance**)
- Simulated transcript abundance levels in **two conditions**: differential and non-differentially expressed, with biological replicates
- Reads simulation based on Flux Simulator

[Sammeth, 2009b]

Evaluation: Simulated reads

Two datasets:

- 5,875 mRNA coding genes from *A.thaliana* with **multiple** annotated alternative transcripts (TAIR10)
- Measured biological variance and gene expression from real data:
  - Between very similar samples (**small biol. variance**)
  - Between close time points (**large biol. variance**)
- Simulated transcript abundance levels in **two conditions**: differential and non-differentially expressed, with biological replicates
- Reads simulation based on Flux Simulator

[Sammeth, 2009b]

Evaluation: Simulated reads

Two datasets:

- 5,875 mRNA coding genes from *A.thaliana* with **multiple** annotated alternative transcripts (TAIR10)
- Measured biological variance and gene expression from real data:
  - Between very similar samples (**small biol. variance**)
  - Between close time points (**large biol. variance**)
- Simulated transcript abundance levels in **two conditions**: differential and non-differentially expressed, with biological replicates
- Reads simulation based on Flux Simulator

[Sammeth, 2009b]

Evaluation: Simulated reads

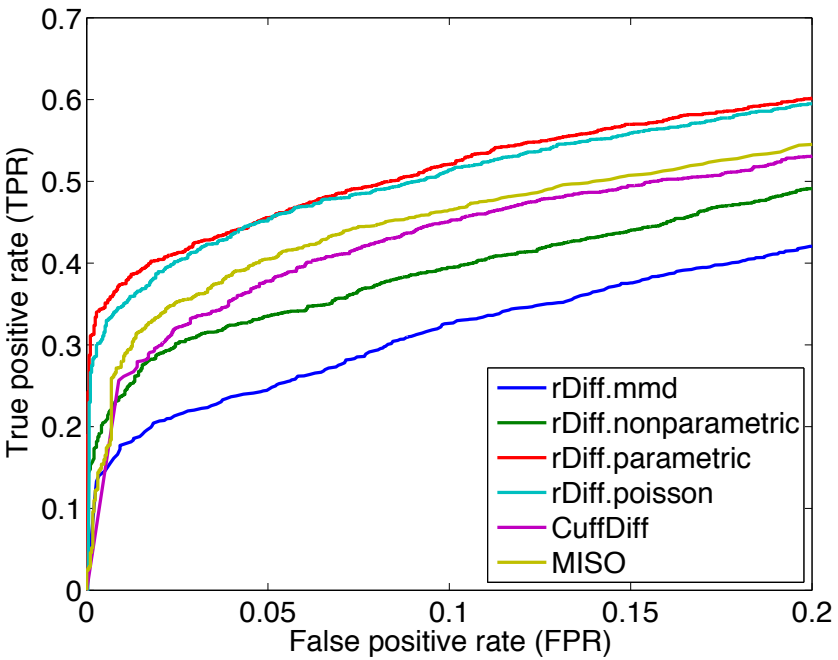
Compared **rDiff.parametric** test and **rDiff.nonparametric** with:

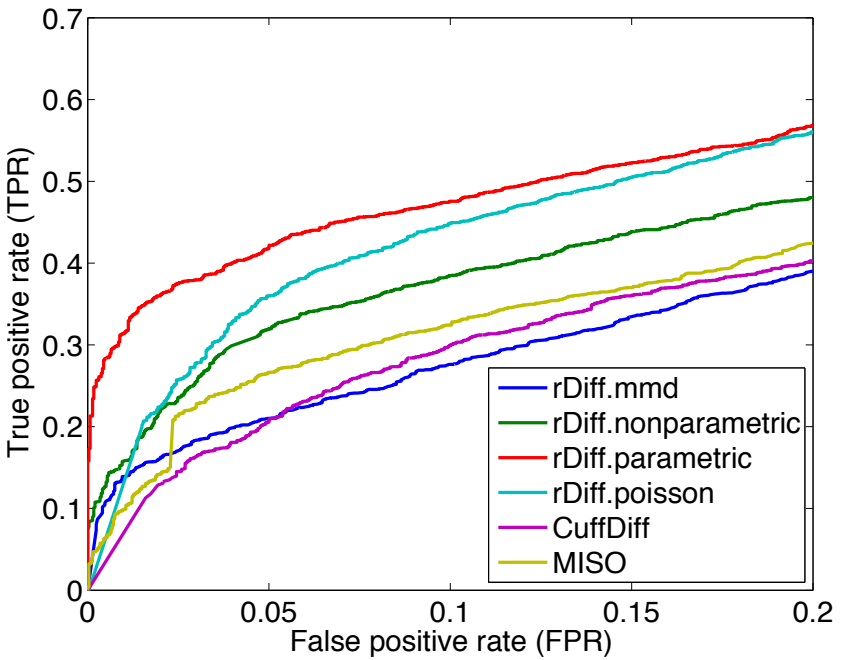
- **MISO** (Katz et al., Nature Methods 2011)
- **CuffDiff** (Trapnell et al., Nature Biotech. 2010)

Evaluation: Simulated reads

Compared **rDiff.parametric** test and **rDiff.nonparametric** with:

- **MISO** (Katz et al., Nature Methods 2011)
- **CuffDiff** (Trapnell et al., Nature Biotech. 2010)





Evaluation: Experimental data

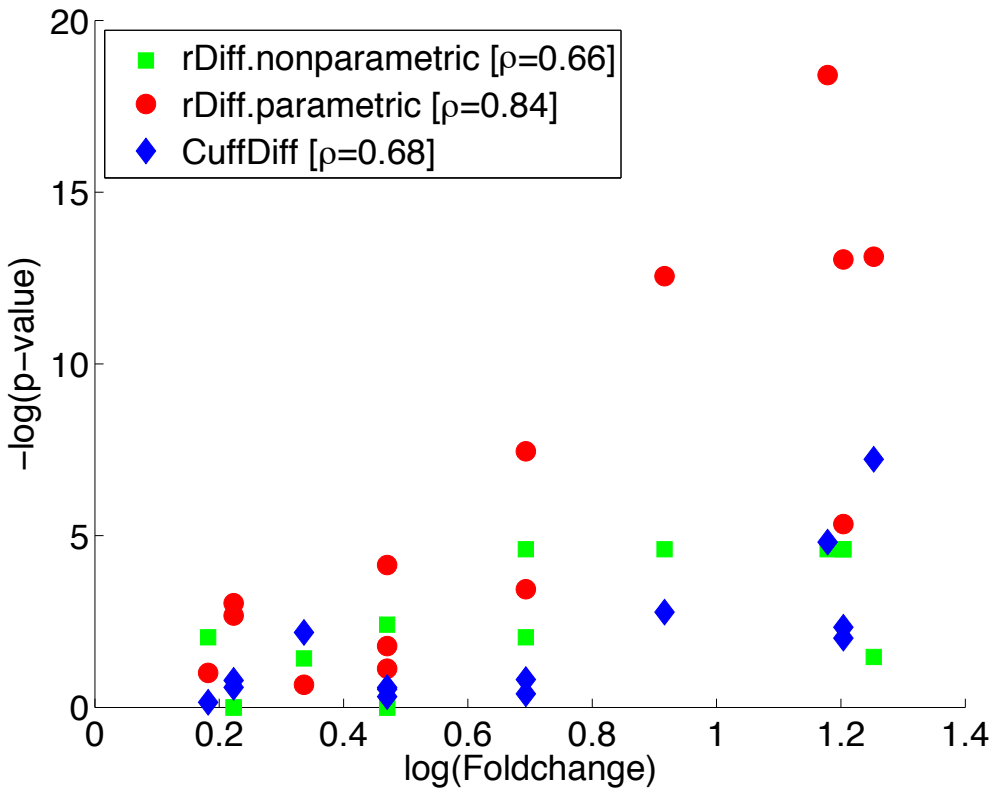
Evaluation on three RNA-Seq libraries from *A.thaliana*.

Evaluation: Experimental data

Evaluation on three RNA-Seq libraries from *A.thaliana*.
Measured transcript abundances using RT-qPCR.

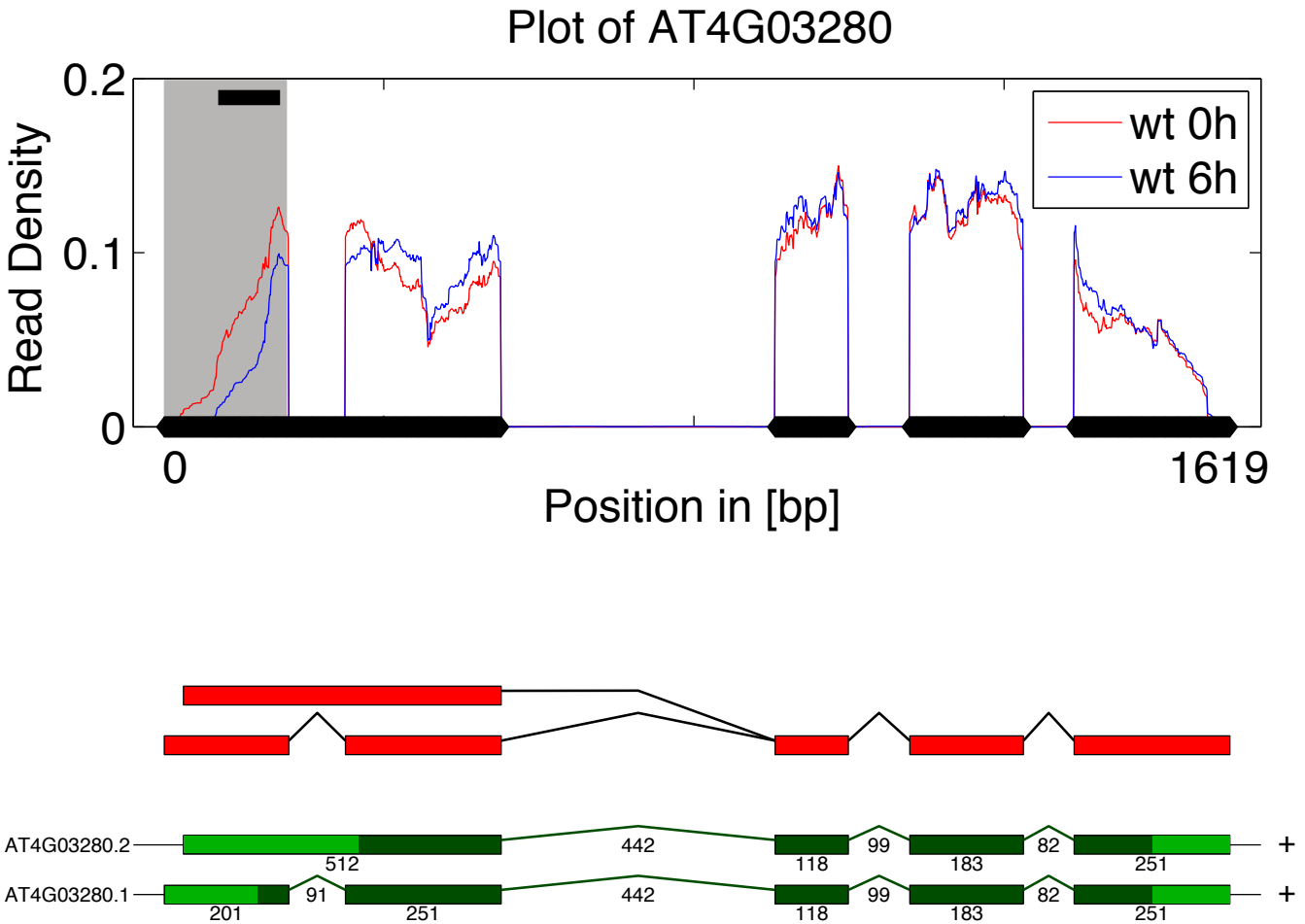
Evaluation: Experimental data

Evaluation on three RNA-Seq libraries from *A.thaliana*.
Measured transcript abundances using RT-qPCR.



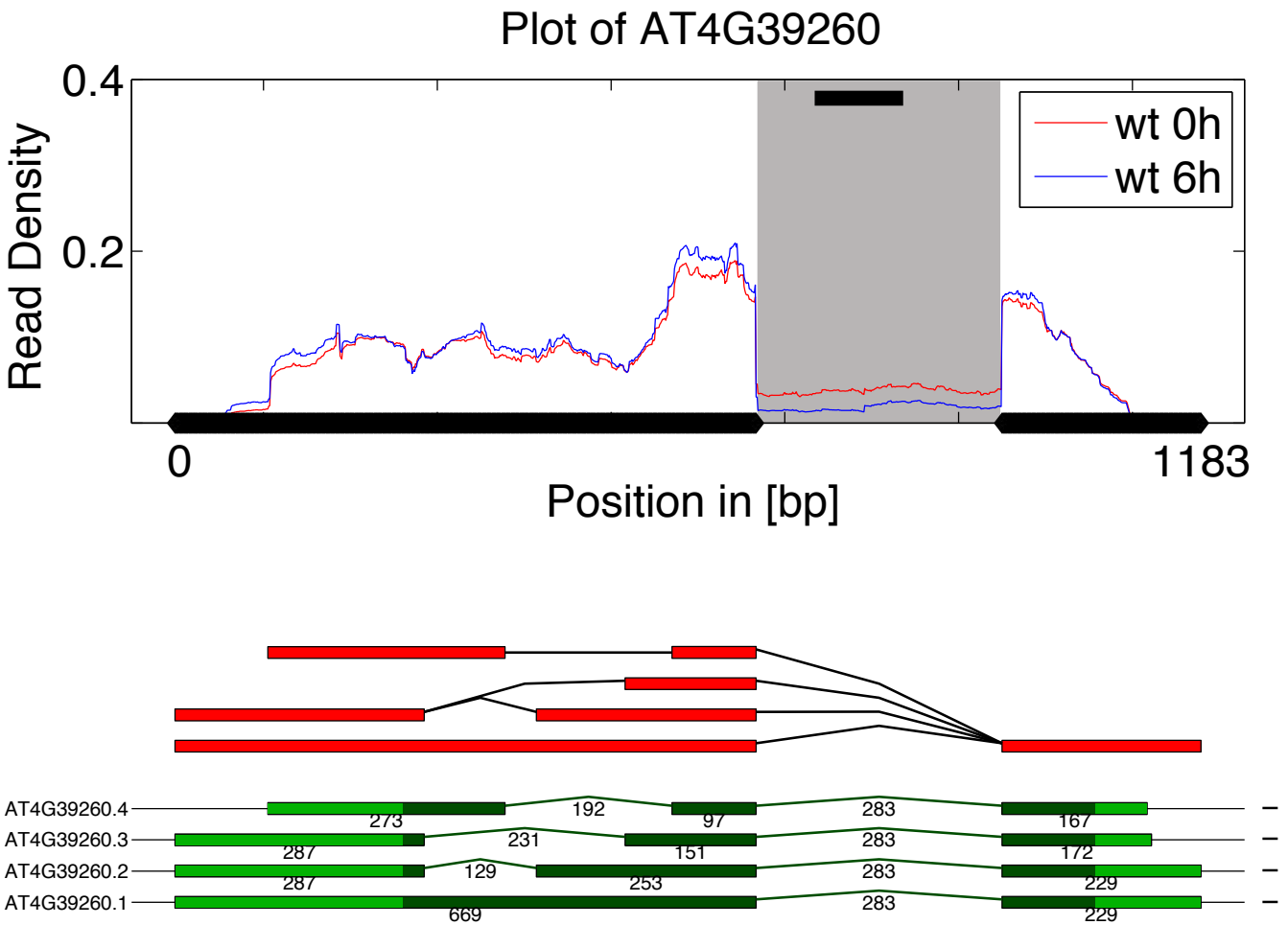
Evaluation: Experimental data

Examples of genes found by rDiff.nonparametric I:



Evaluation: Experimental data

Examples of genes found by rDiff.nonparametric II:



rDiff Summary

- Doing the quantification right is hard
- When the gene structure is **known**: **rDiff.parametric**
- When the gene structure is not completely known: **rDiff.nonparametric**
- Implementation can be found on our Galaxy webserver as part of the **rdiff** package: `http://galaxy.cbio.mskcc.org`
- For more information: `http://bioweb.me/rdiff`

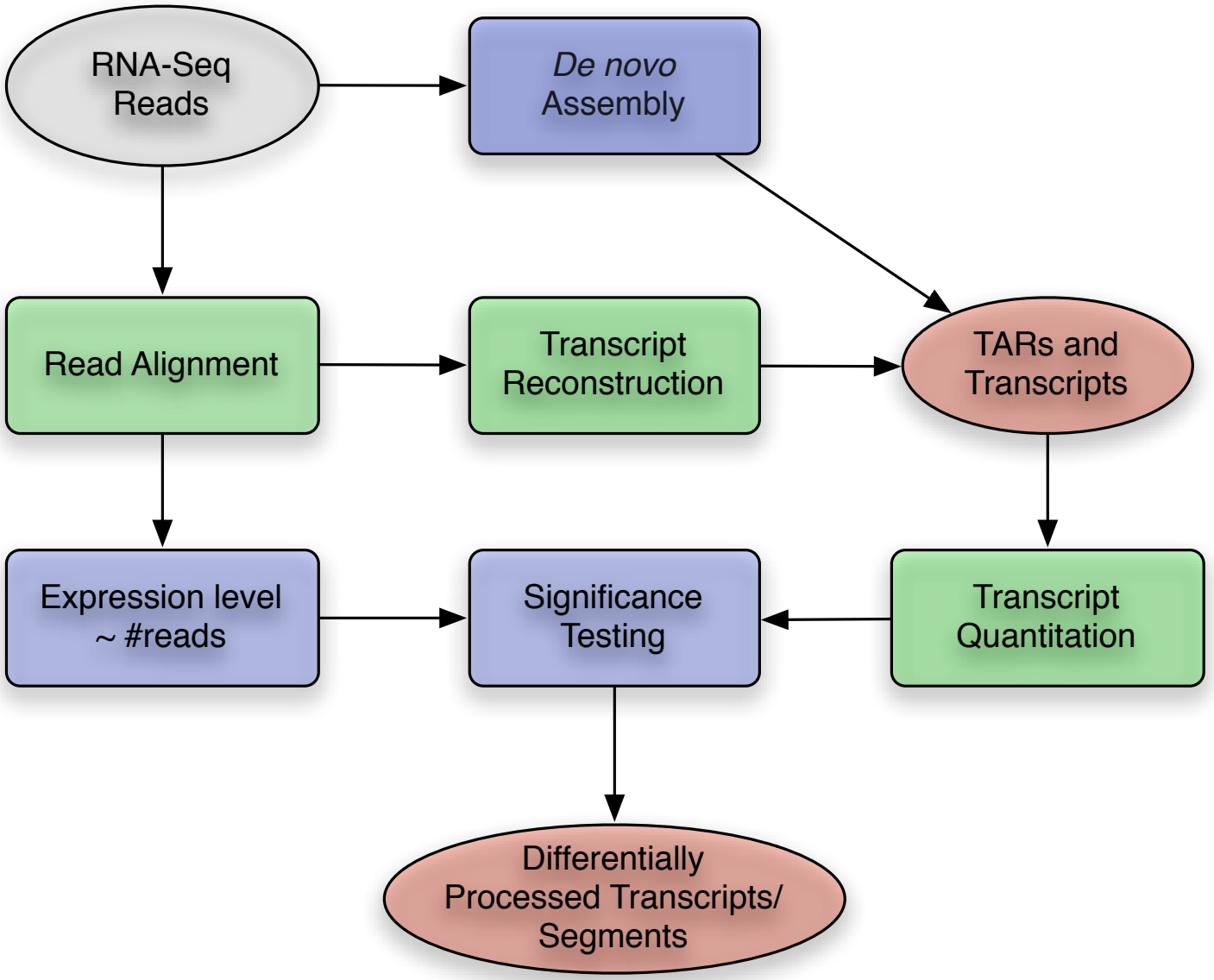
Published online 12 April 2013

Nucleic Acids Research, 2013, Vol. 41, No. 10 ***5189–5198***
doi:10.1093/nar/gkt211

Accurate detection of differential RNA processing

Philipp Drewe[1,2,*], Oliver Stegle[3,4], Lisa Hartmann[2,5], André Kahles[1,2], Regina Bohnert[2], Andreas Wachter[5], Karsten Borgwardt[3,4,6] and Gunnar Rätsch[1,2,*]

Common RNA-Seq Analysis Steps



Transcript Identification Strategies

- Segmentation strategies: Based on read coverage only
  → Little assumptions
  → Find non-coding transcripts
  → Varies with expression
- Trade-off between read information with genomic sequence features
  → Find a comprehensive set of genes
  → More complete due to additional information

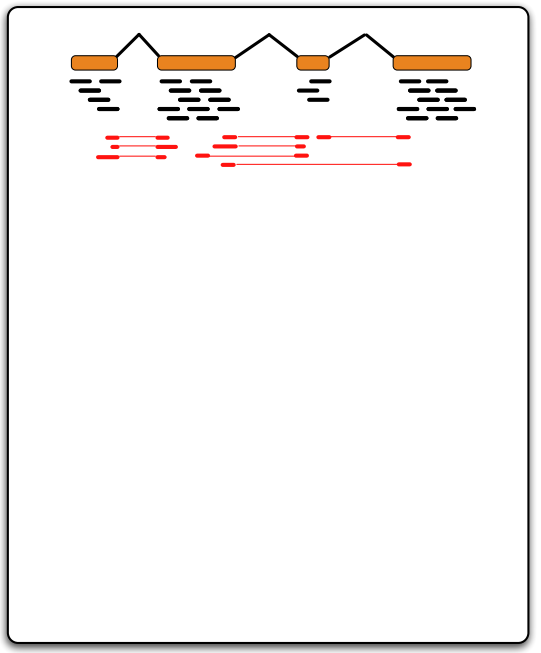
Methods:

- G.Mo.R-Se (threshold based exon identification)
- mGene.NGS (gene finding with RNA-seq)
- Cufflinks, Scripture, MiTie (genome-guided assembly)

Alignments & Transcript Reconstruction

Transcript reconstruction:

- Identify *exons* and
- Relationship in transcripts



Most approaches:

- Build a transcript/splicing graph
- Choose paths as transcripts
- Quantify chosen transcripts

Strategy fail if transcript is incorrect

Exon level errors or confusion of different alternatives is likely

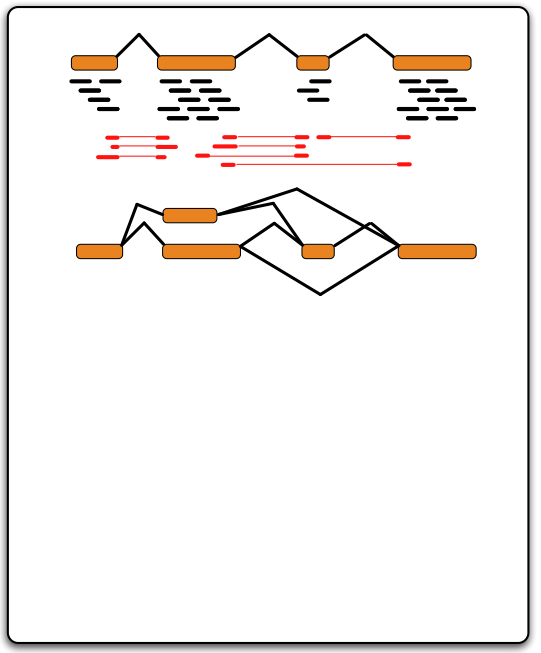
Alignments & Transcript Reconstruction

Transcript reconstruction:

- Identify *exons* and
- Relationship in transcripts

Most approaches:

- Build a transcript/splicing graph
- Choose paths as transcripts
- Quantify choosen transcripts



Strategy fails if graph is incorrect

Exon-level F-Score of Cufflinks for different alignments & filter steps.

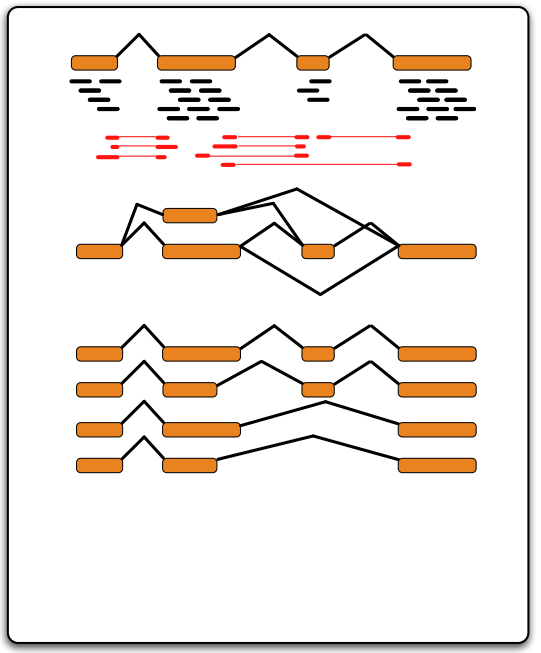
Alignments & Transcript Reconstruction

Transcript reconstruction:

- Identify *exons* and
- Relationship in transcripts

Most approaches:

- Build a transcript/splicing graph
- Choose paths as transcripts
- Quantify choosen transcripts



Strategy fails if graph is incorrect

Exon-level F-Score of Cufflinks for different alignments & filter steps.

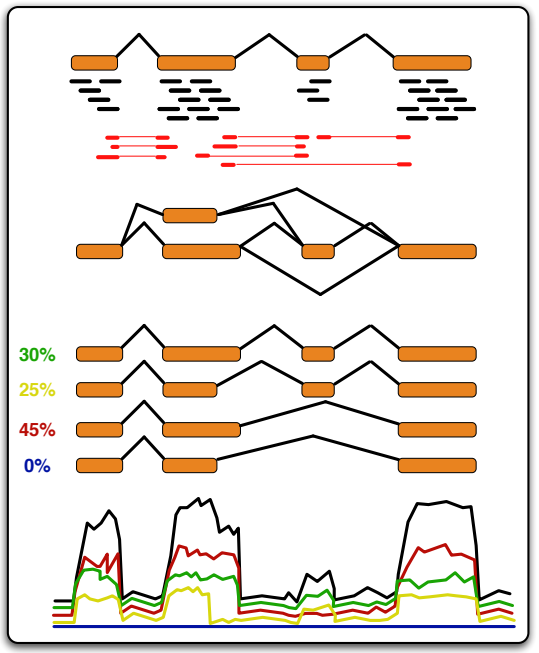
Alignments & Transcript Reconstruction

Transcript reconstruction:

- Identify *exons* and
- Relationship in transcripts

Most approaches:

- Build a transcript/splicing graph
- Choose paths as transcripts
- Quantify choosen transcripts



Strategy fails if graph is incorrect

Exon-level F-Score of Cufflinks for different alignments & filter steps.

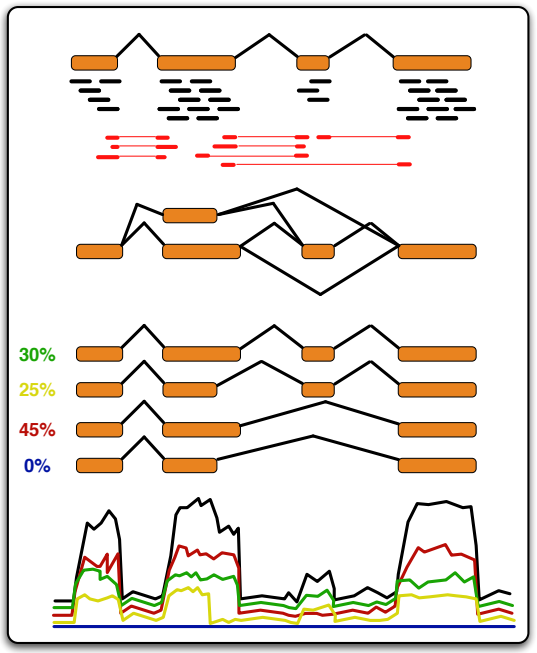
Alignments & Transcript Reconstruction

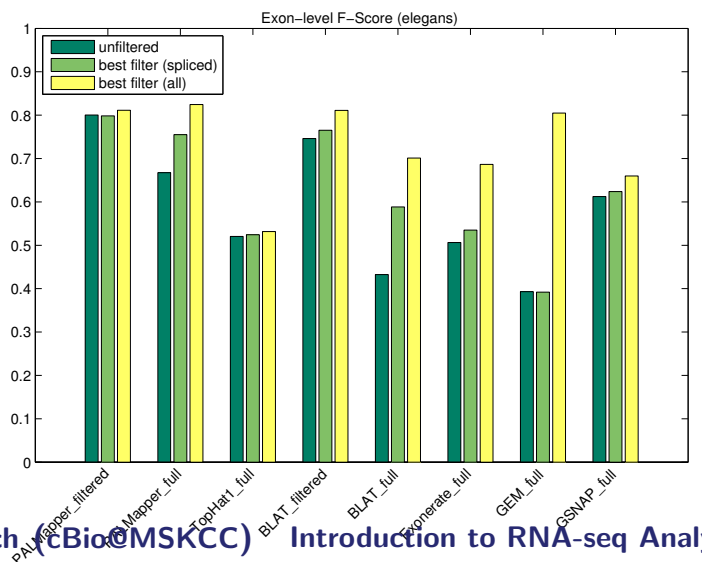
Transcript reconstruction:

- Identify *exons* and
- Relationship in transcripts

Most approaches:

- Build a transcript/splicing graph
- Choose paths as transcripts
- Quantify choosen transcripts

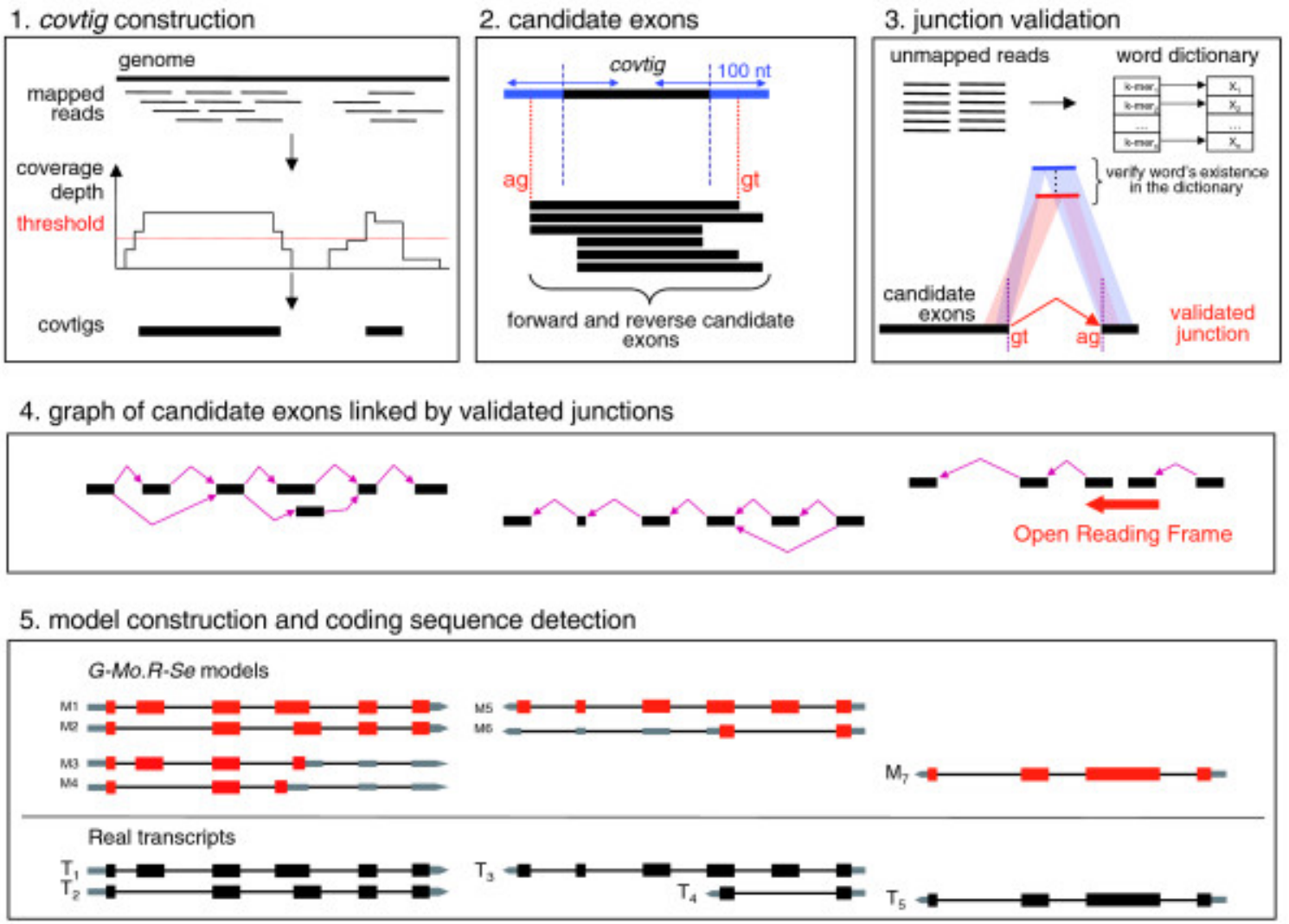




Strategy fails if graph is incorrect

Exon-level F-Score of Cufflinks for different alignments & filter steps.

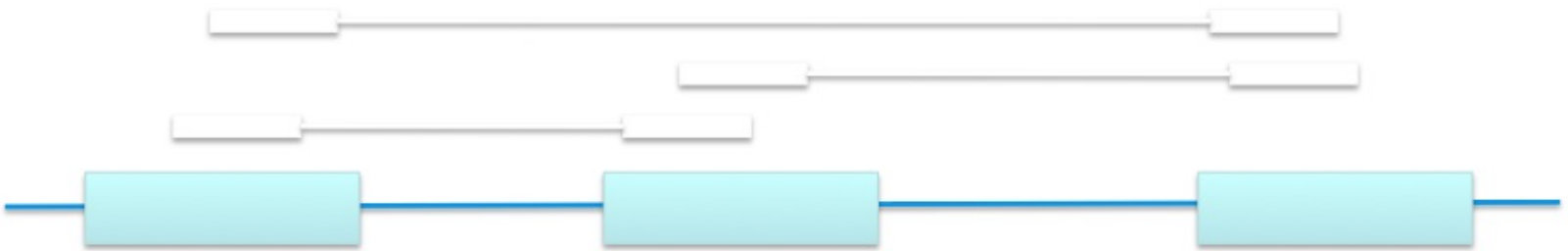
G.Mo.R-Se Splice Graph Reconstruction



Denoeud F. et al. Annotating genomes with massive-scale RNA sequencing. Genome Biology(2008) 9:R175

Cufflinks

Assemble as many reads into as few (highly probable) transcripts as possible.

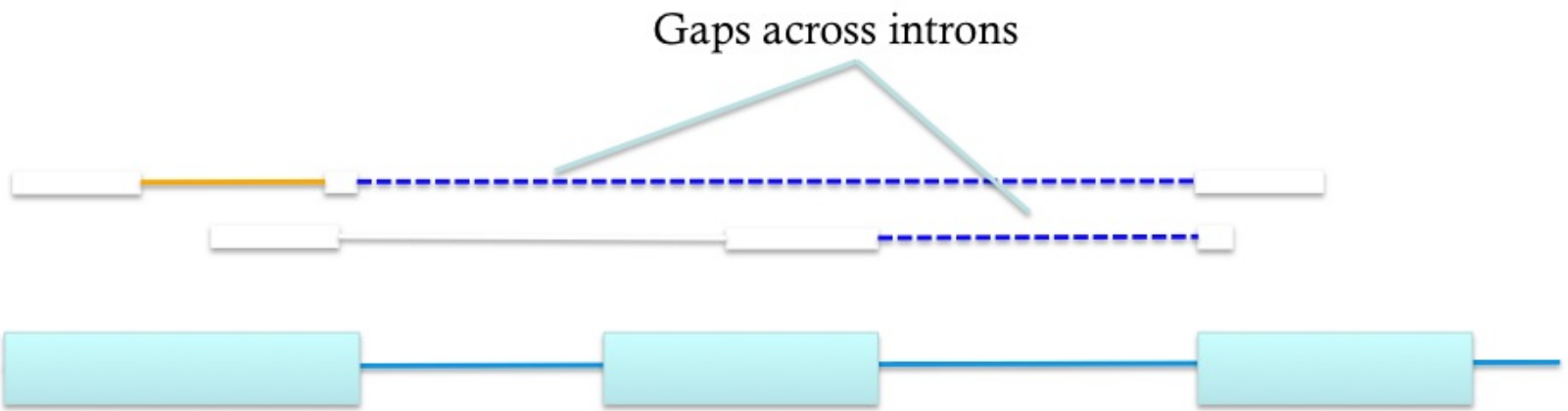


How many transcripts?

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Compatible reads

Don't know that two reads came from the same transcript, but sometimes we know that they came from **different** transcripts.

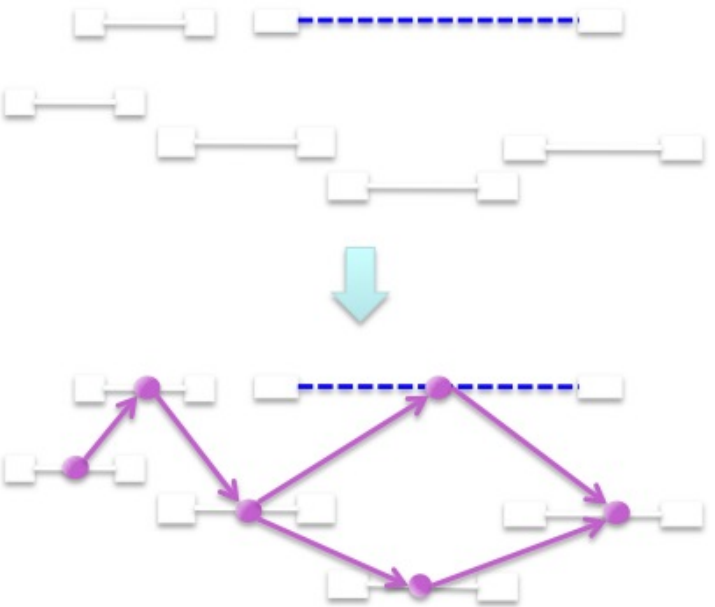


How many transcripts?

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Partially ordering mates

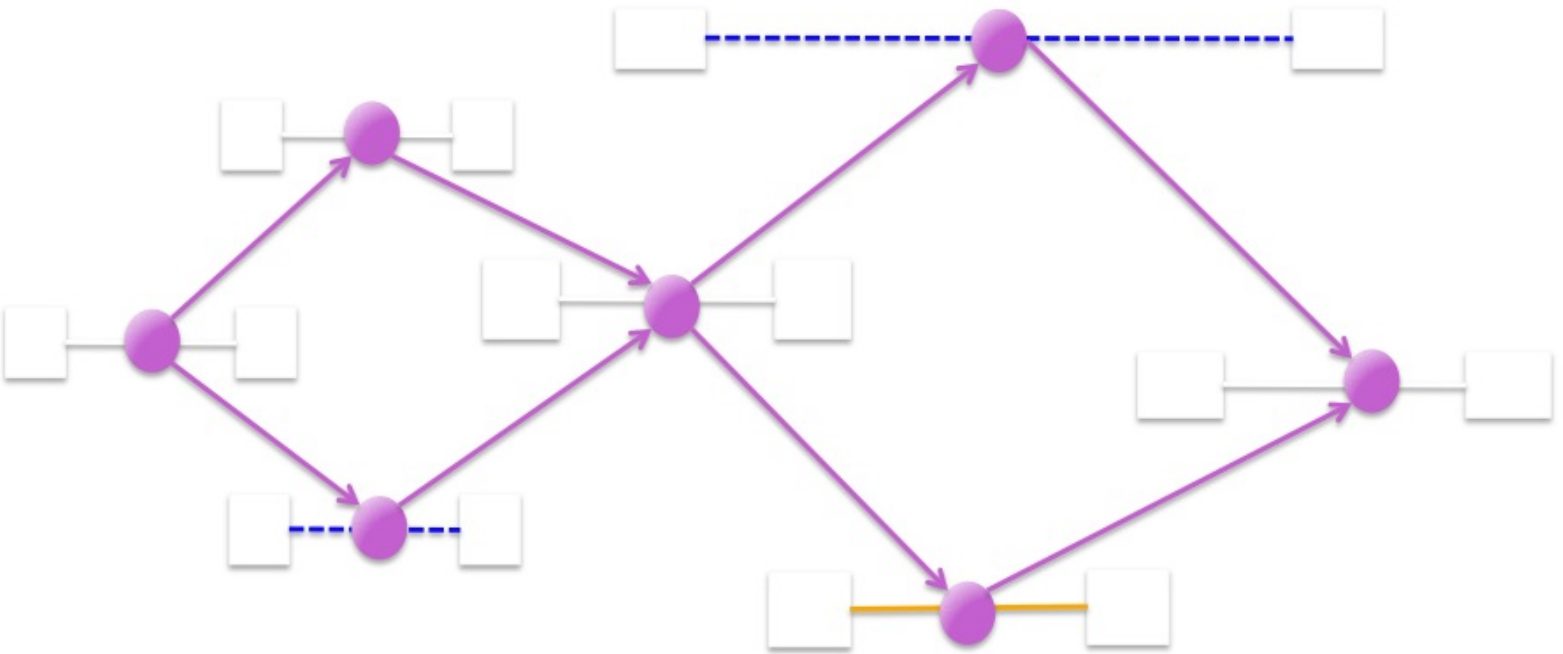
- alignment $x \leq y$ when
  - x starts to the left of y in the reference
  - x and y overlap consistently
  - y is not contained in x
- In other words, $x \leq y$ when they could have come from the same transcript



Building a DAG for the partial order

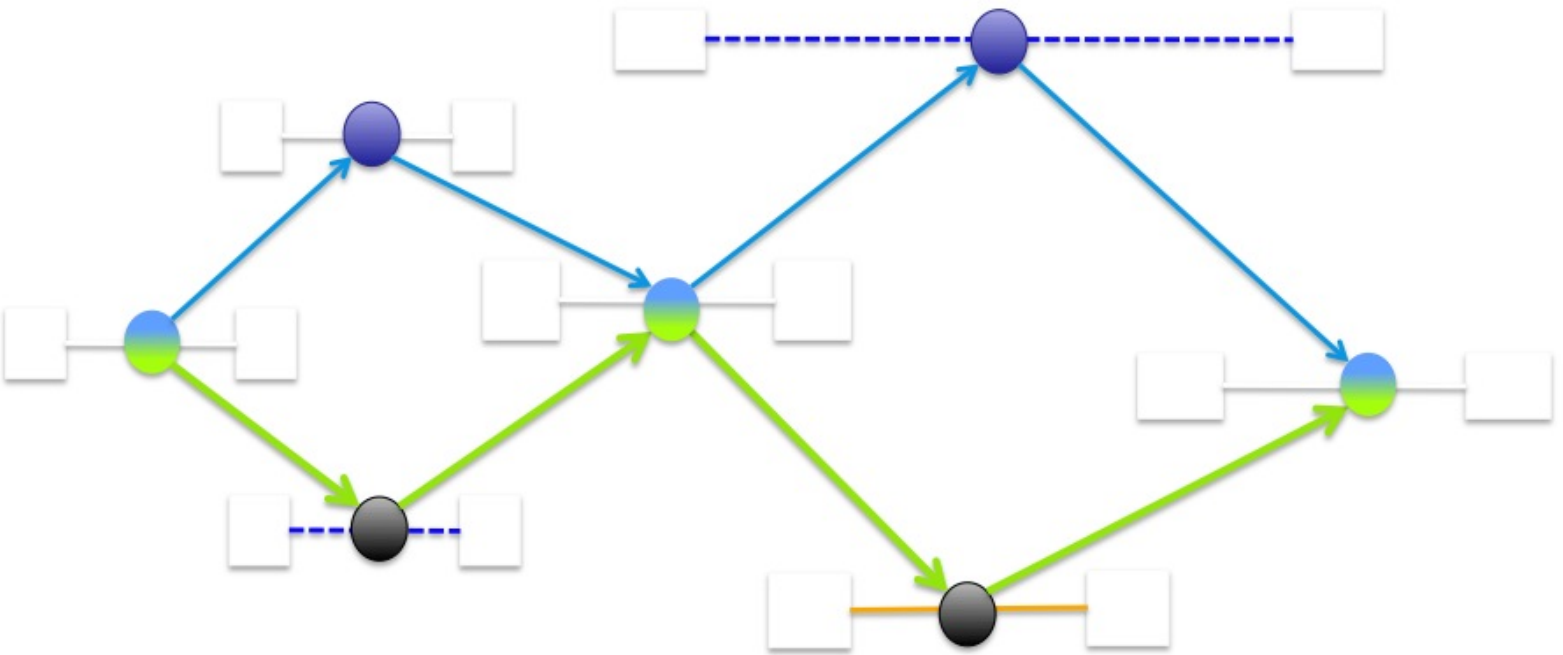
(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Path cover of the DAG



(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

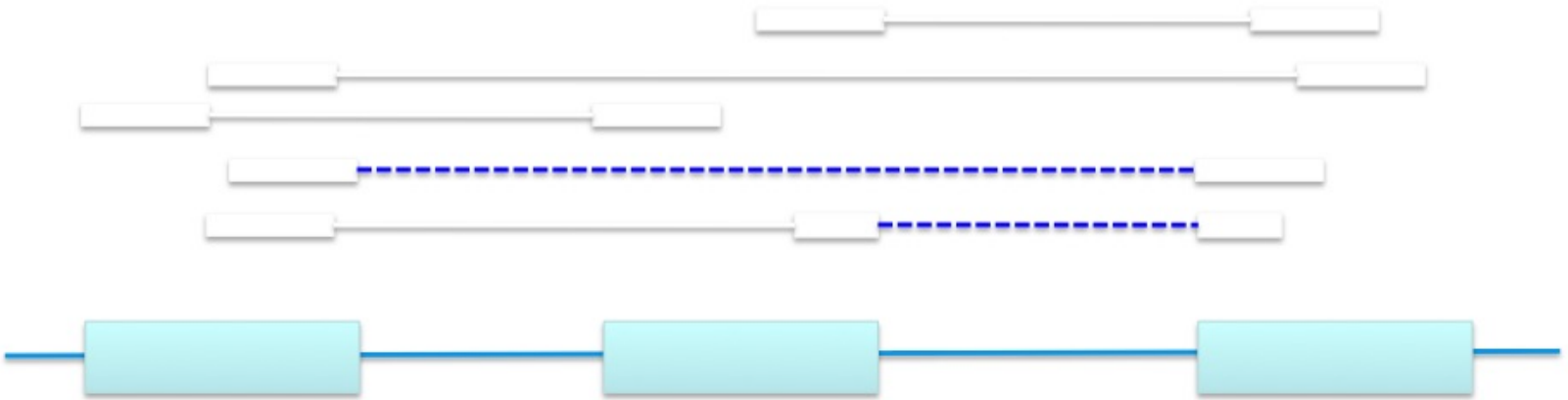
Path cover of the DAG



(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Antichains

Don't know that two reads came from the same transcript, but sometimes we know that they came from **different** transcripts.

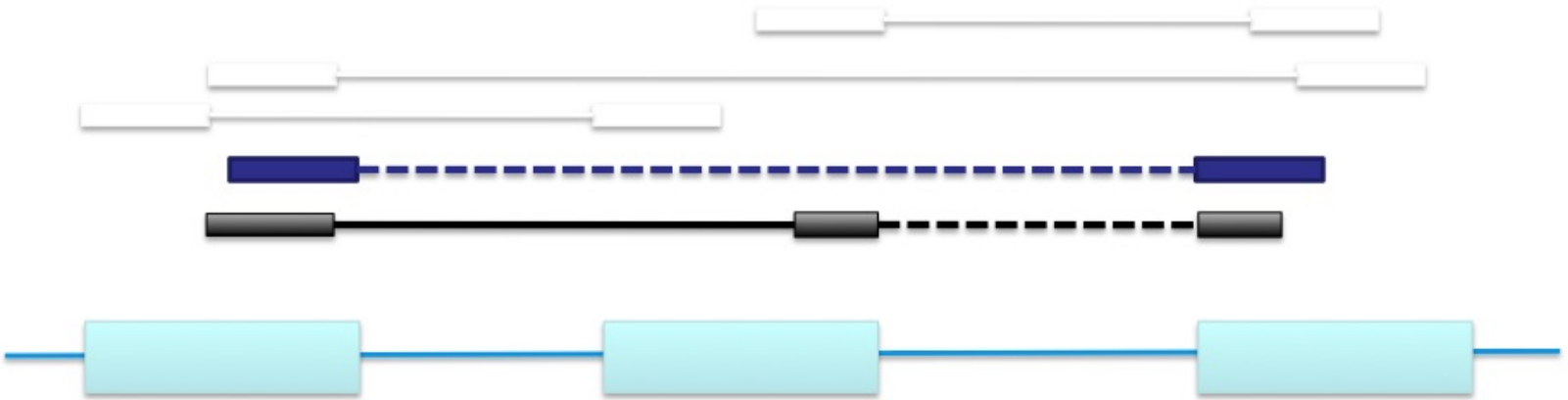


How many transcripts?

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Antichains

Don't know that two reads came from the same transcript, but sometimes we know that they came from **different** transcripts.



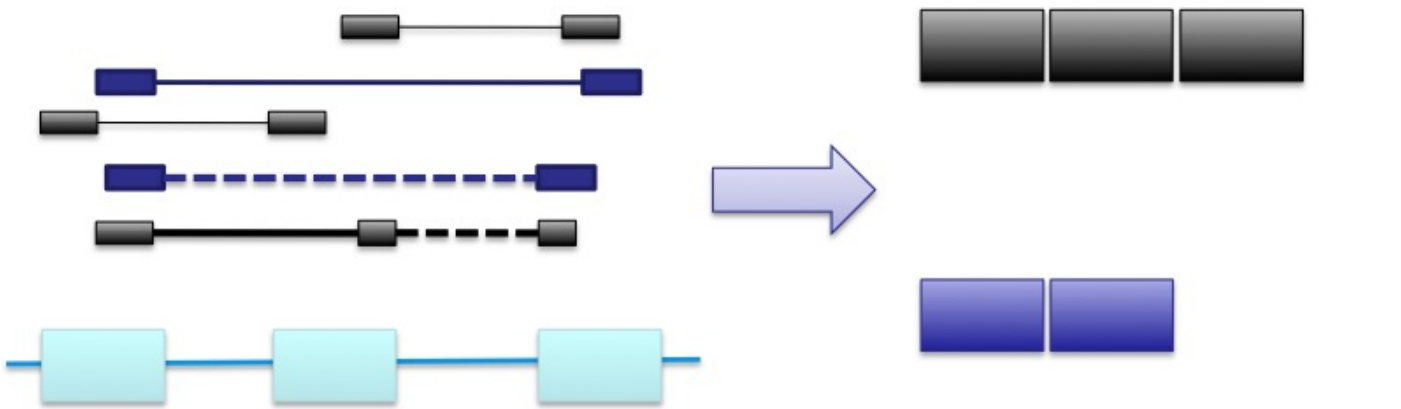
How many transcripts?
Two!

(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Chains

From chain decomposition of alignments…

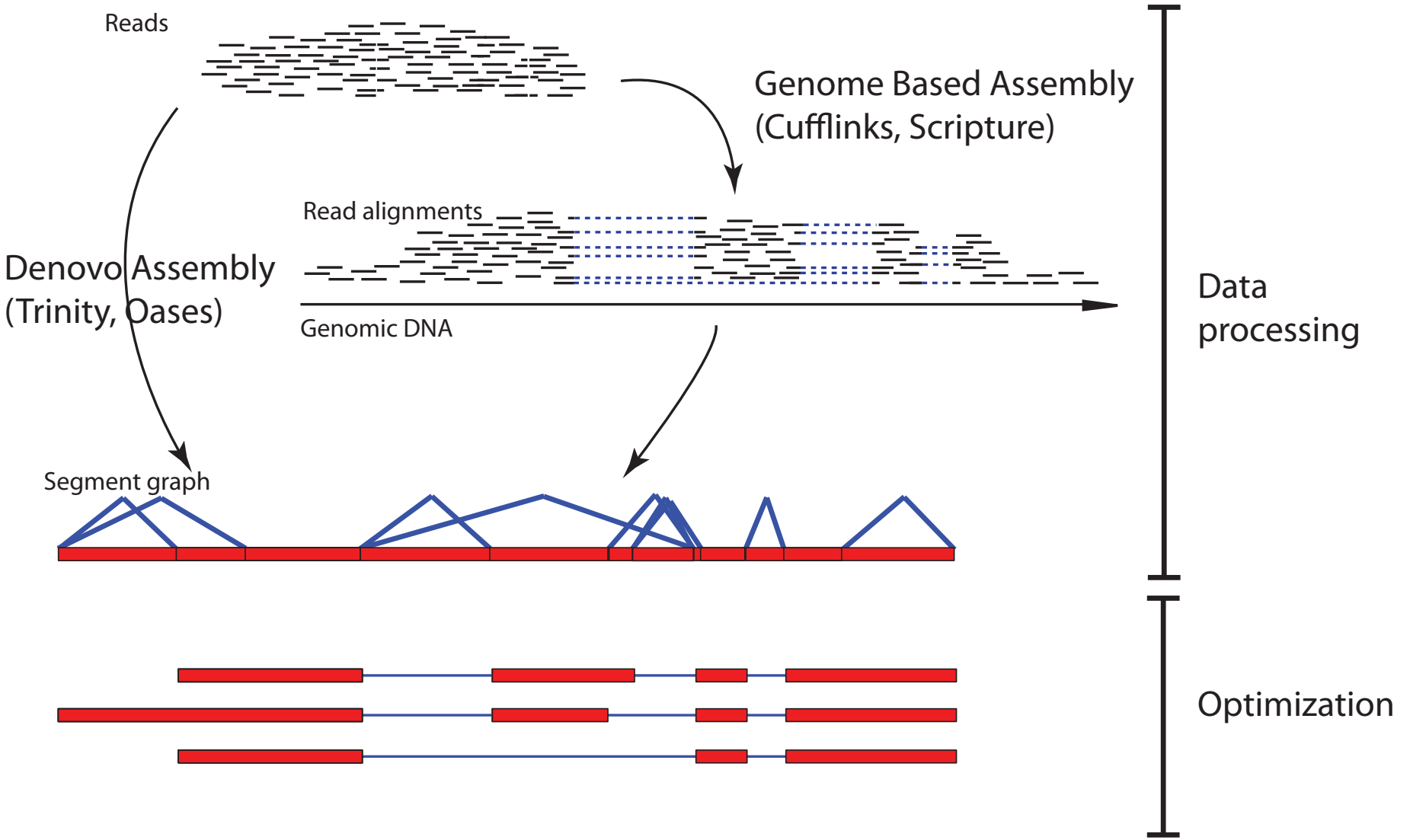
… build transcript assemblies



The chains are sets of mutually compatible alignments, so we can smash them together into transcripts at will.

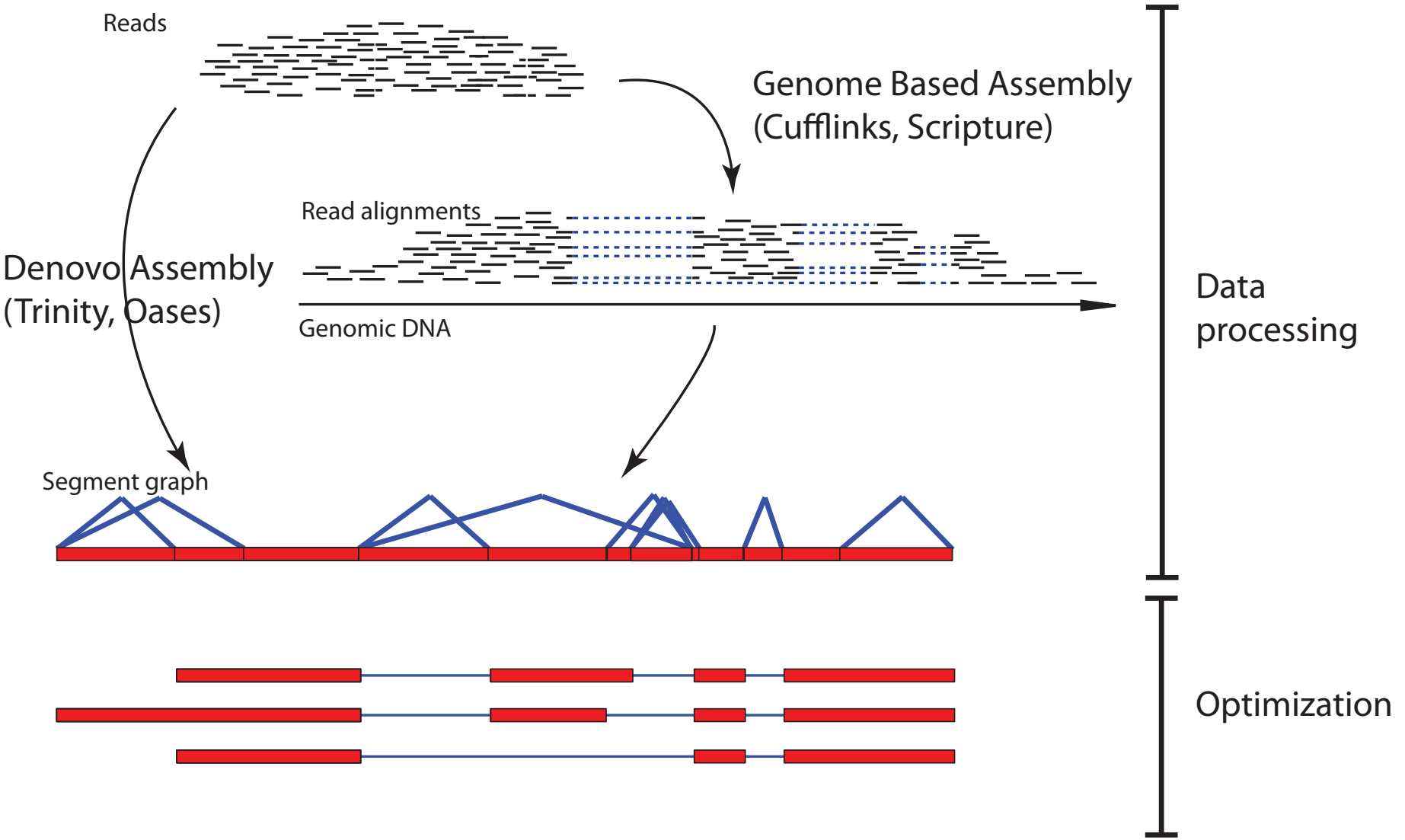
(Slides provided by Cole Trapnell, U Maryland & UC Berkeley)

Transcript Reconstruction with RNA-seq



108 possible transcripts, 10^{28} possible subsets of transcripts

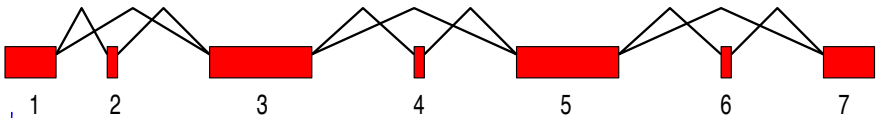
Transcript Reconstruction with RNA-seq



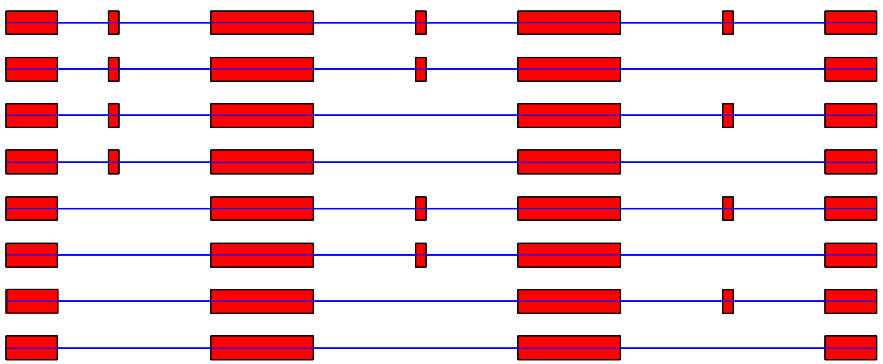
108 **possible transcripts,** 10^{28} **possible subsets of transcripts**

Toy example

Splicegraph

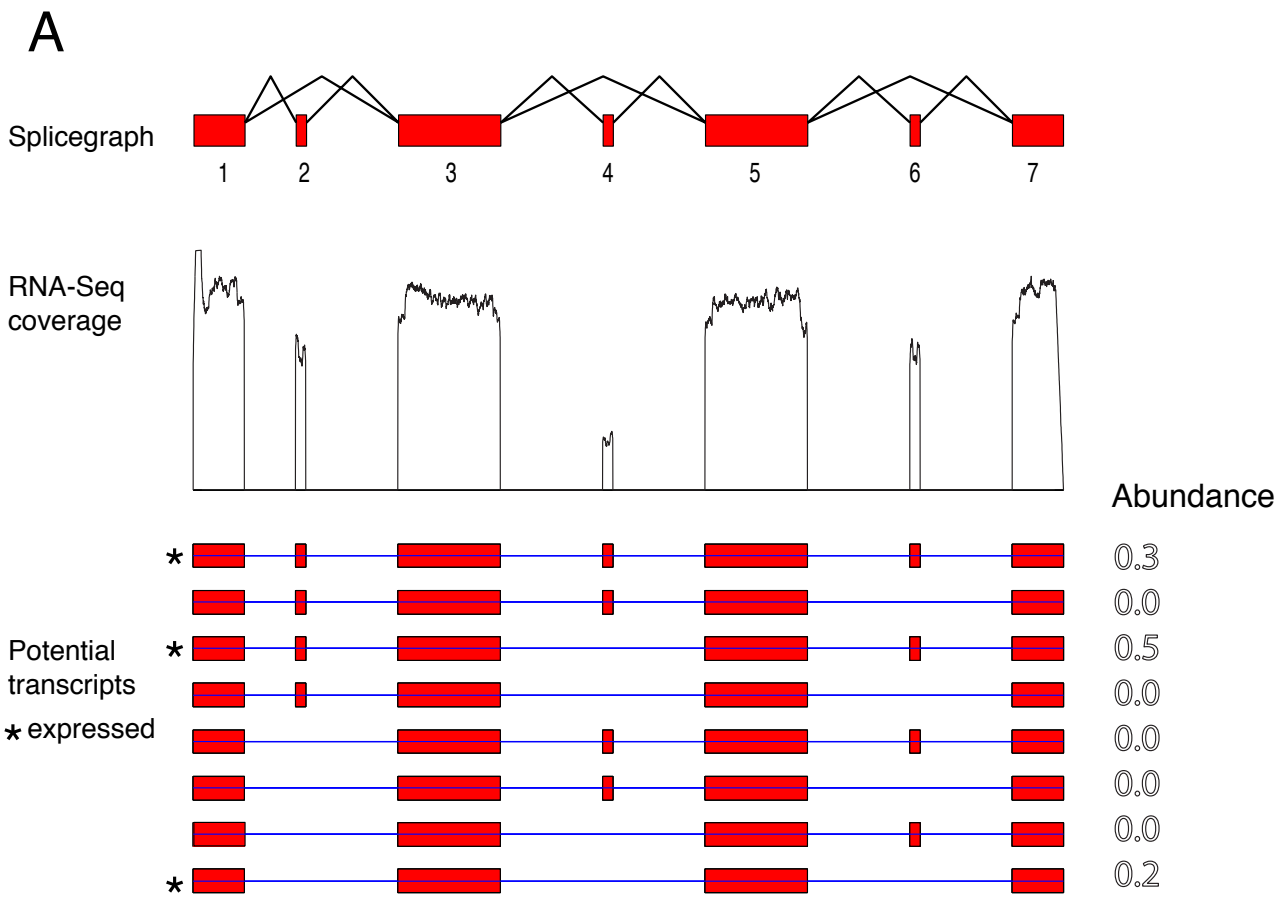


Potential transcripts



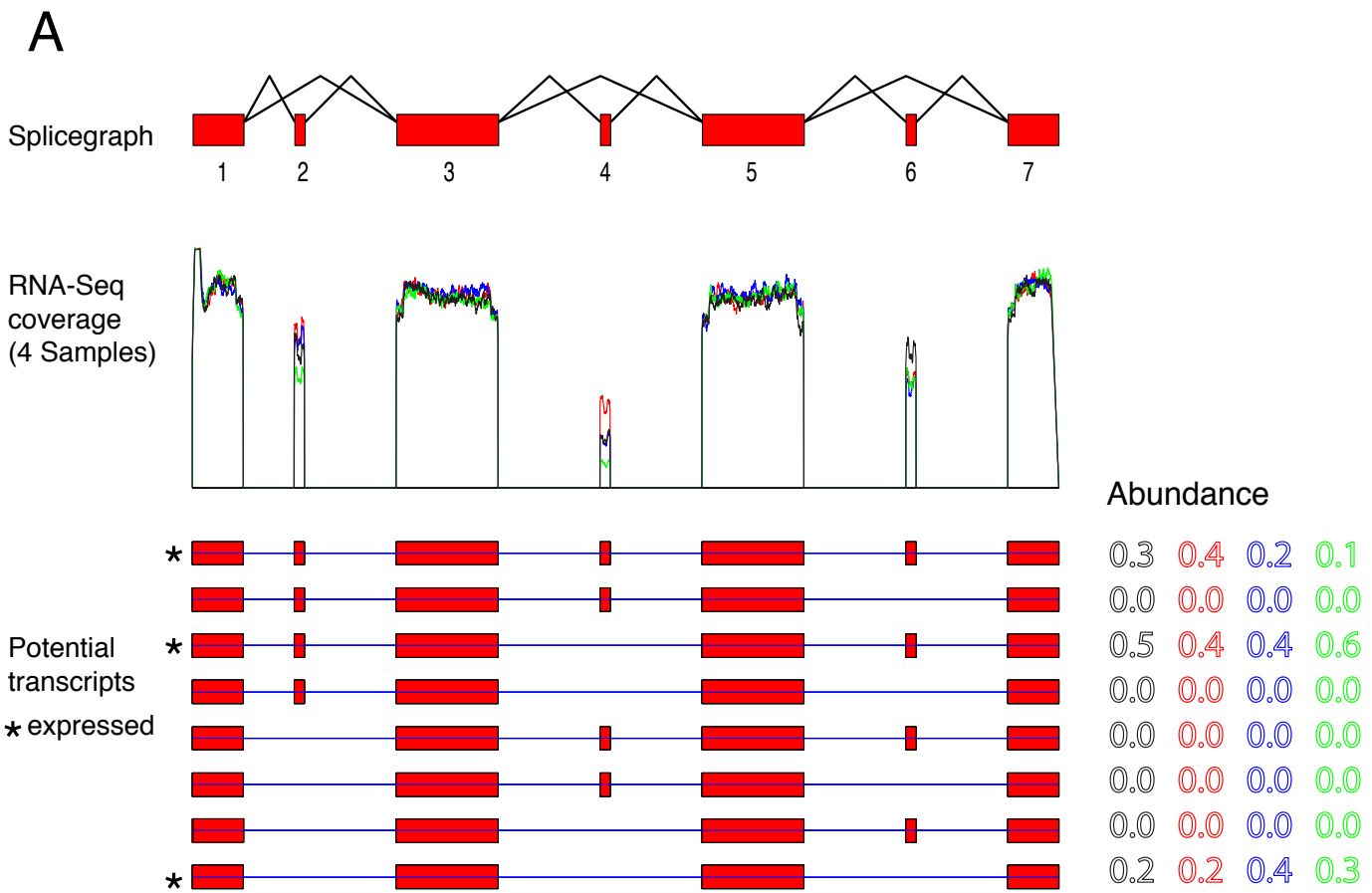
V. Lacroix et al., *WABI '08* (2008)

Toy example



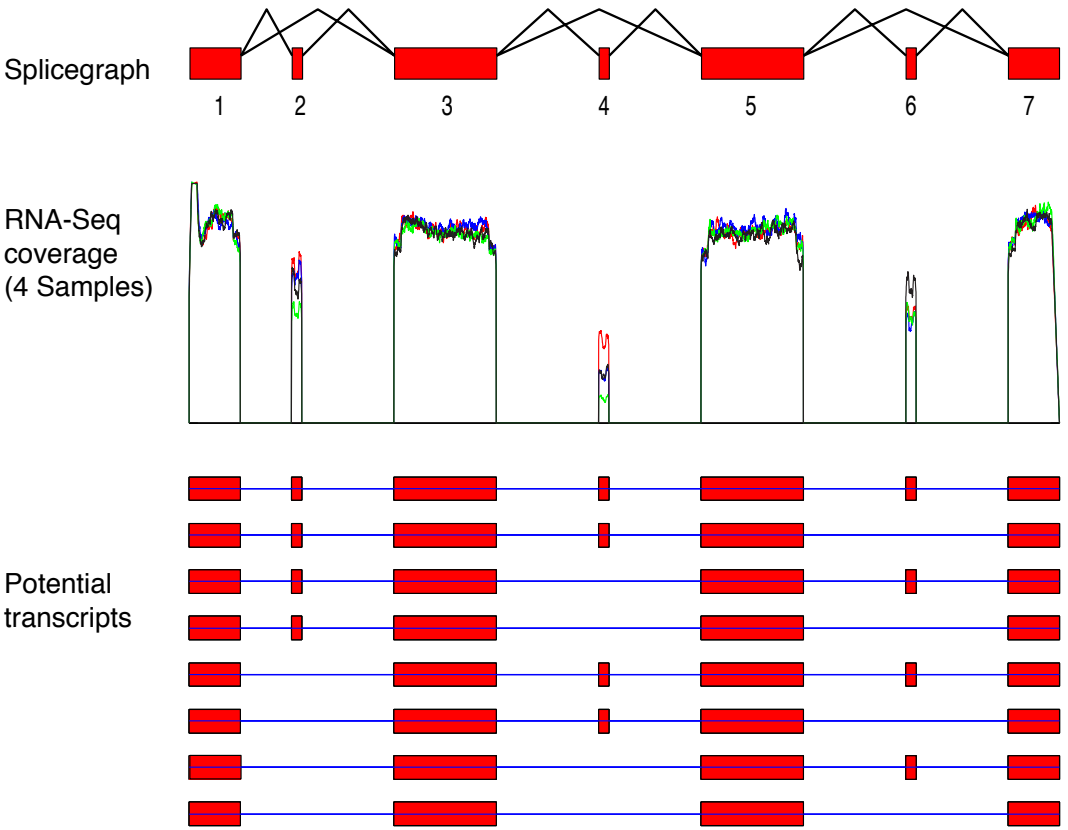
V. Lacroix et al., *WABI '08* (2008)

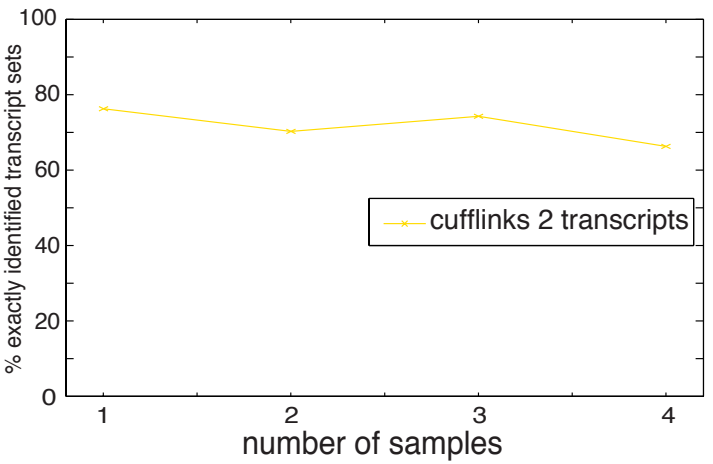
Toy example



V. Lacroix et al., *WABI '08* (2008)

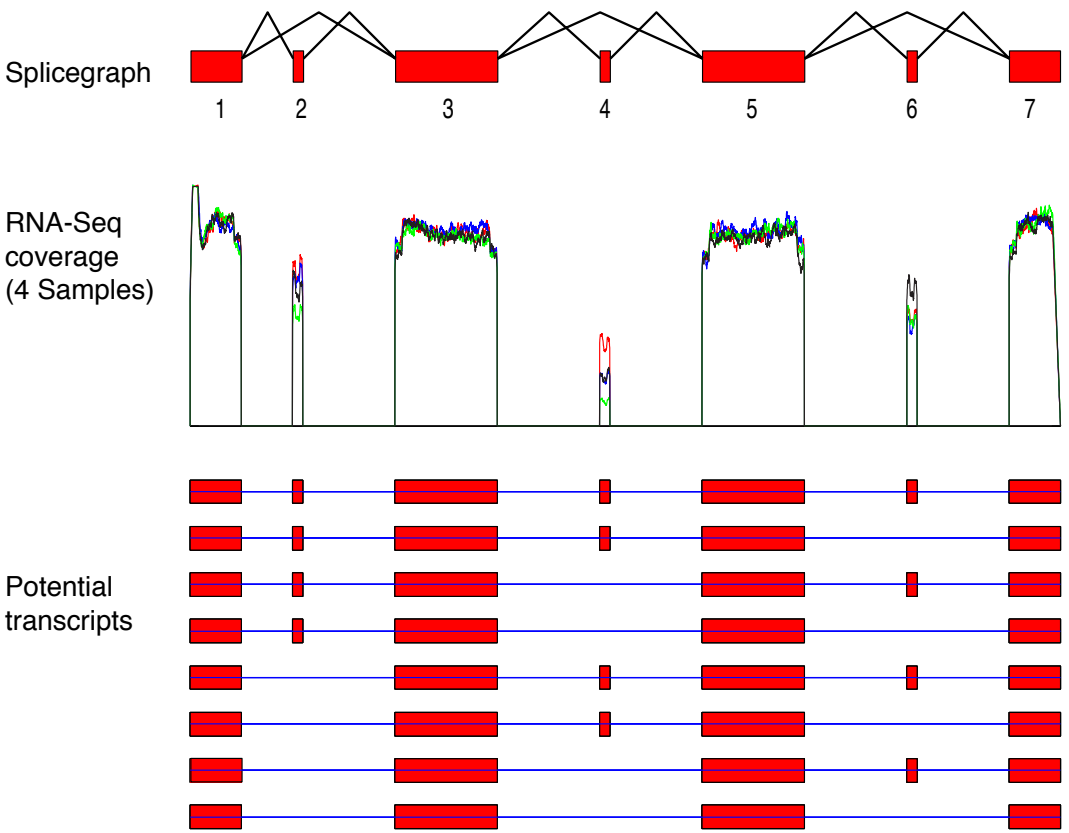
Toy example

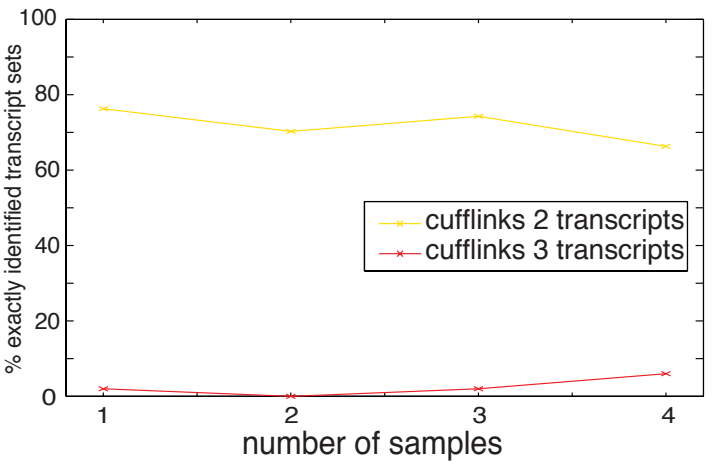




V. Lacroix et al., *WABI '08* (2008)

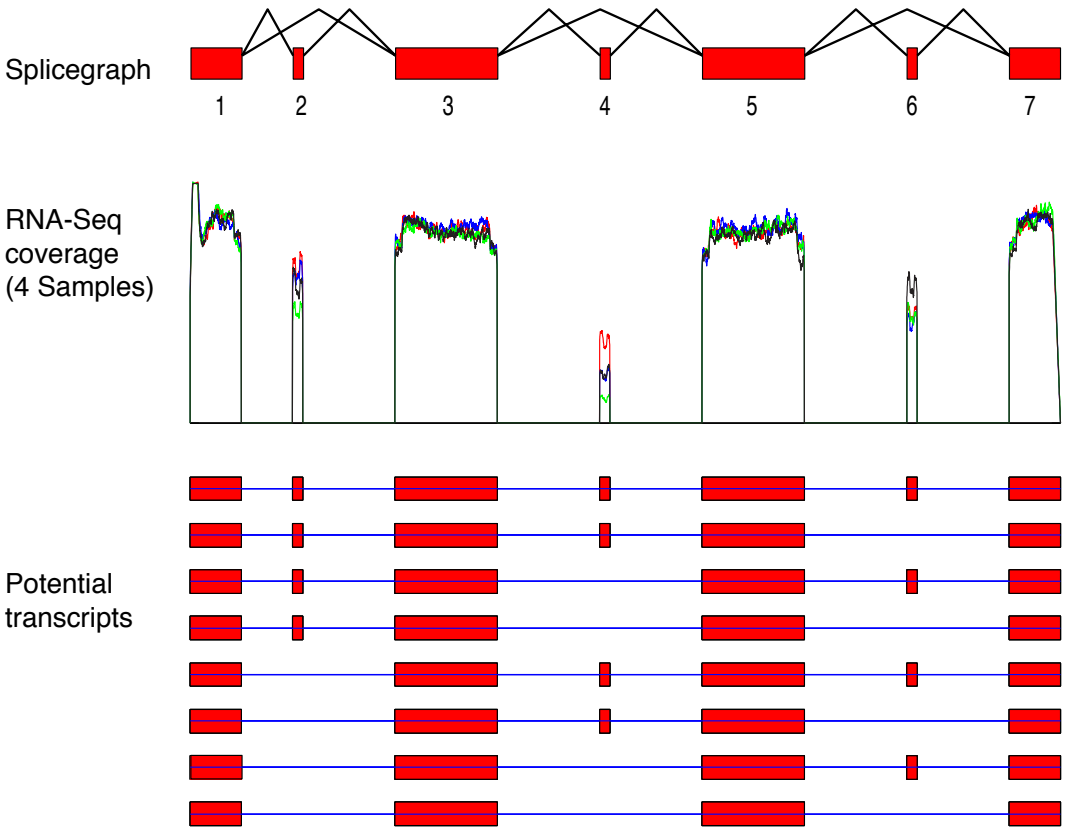
Toy example



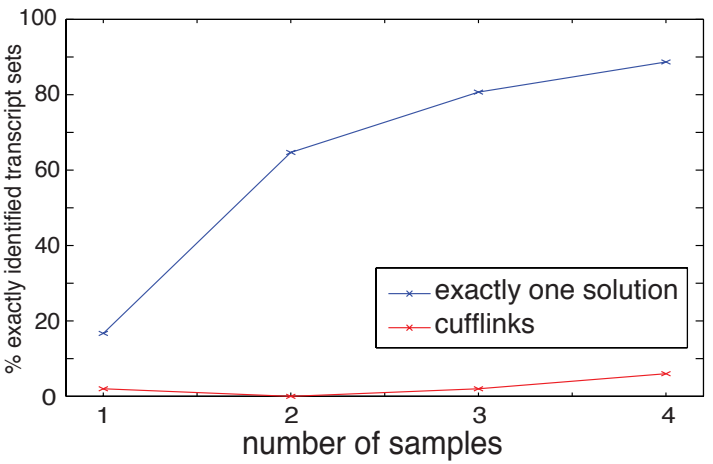


V. Lacroix et al., *WABI '08* (2008)

Toy example



3 transcripts expressed



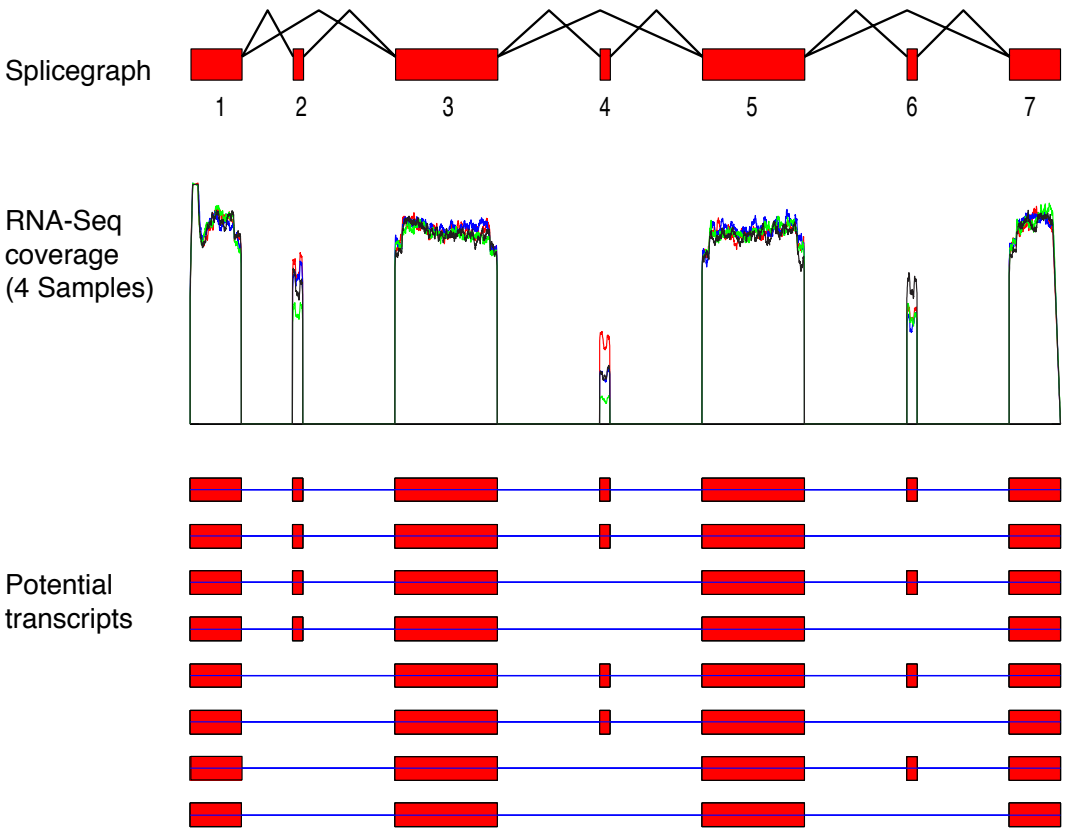
Example:

{{1,3,5},
{3,4,5},
{3,4,6}}

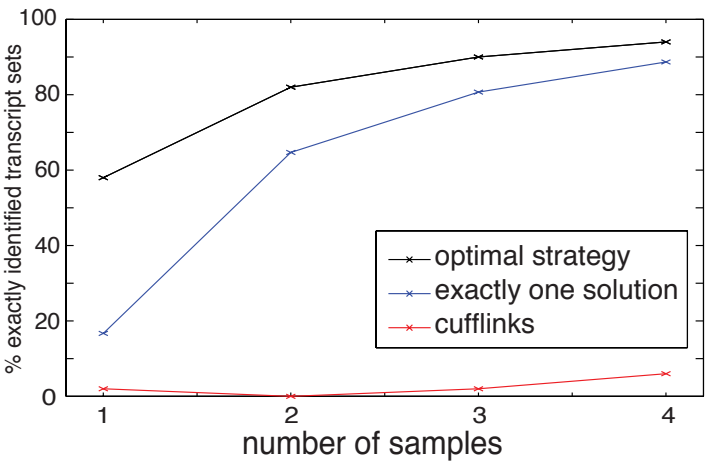
{{3,4,6},
{3,4,8}}

V. Lacroix et al., *WABI '08* (2008)

Toy example



3 transcripts expressed



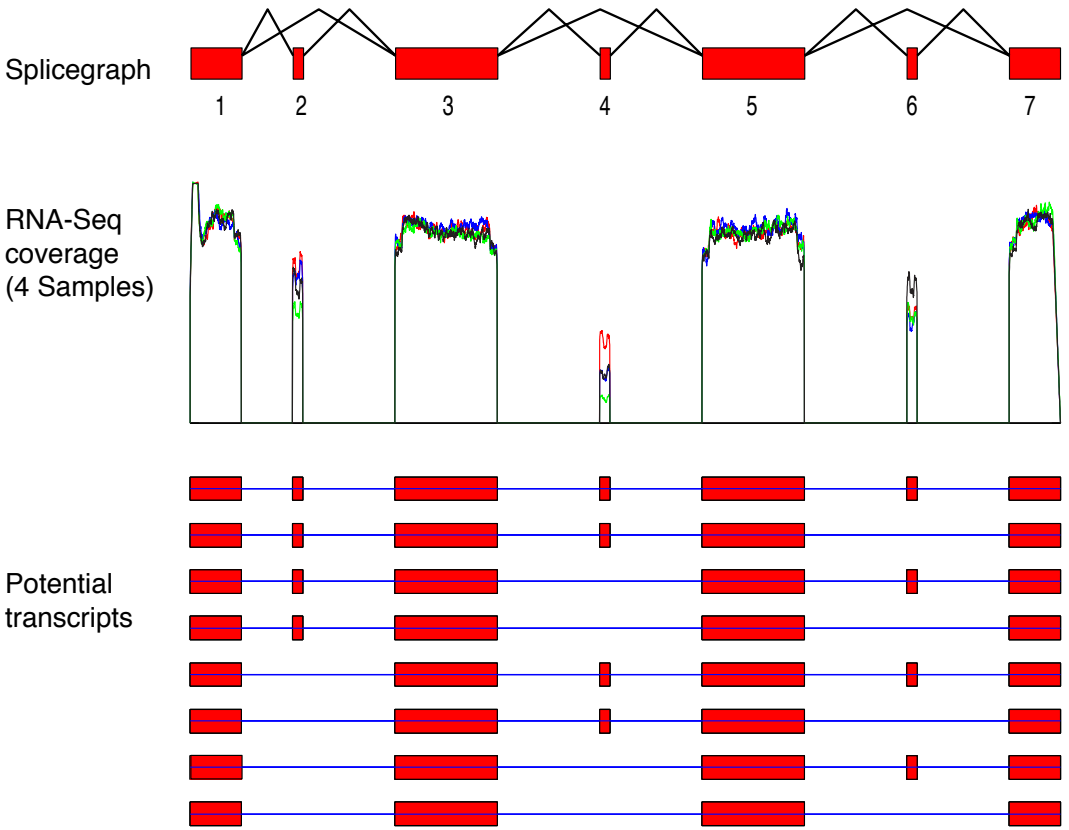
Example:

{{1,3,5},
{3,4,5},
{3,4,6}}

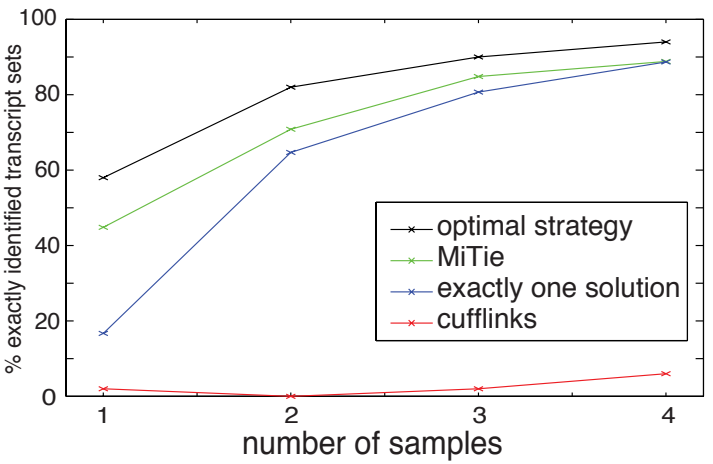
{{3,4,6},
{3,4,8}}

V. Lacroix et al., *WABI '08* (2008)

Toy example



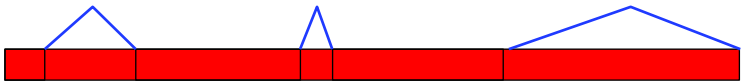
3 transcripts expressed



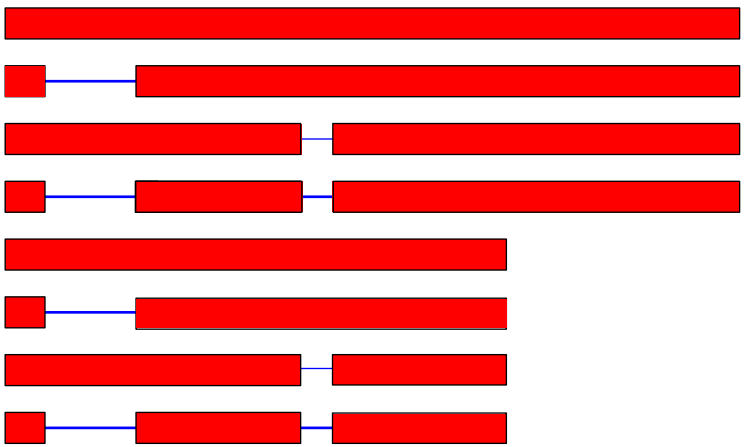
V. Lacroix et al., *WABI '08* (2008)

Quantification

Segment Graph

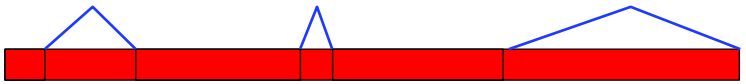


Potential Transcripts

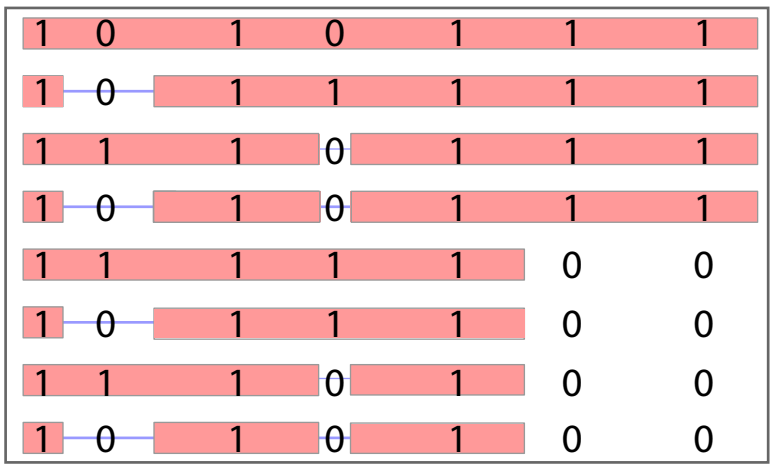


Quantification

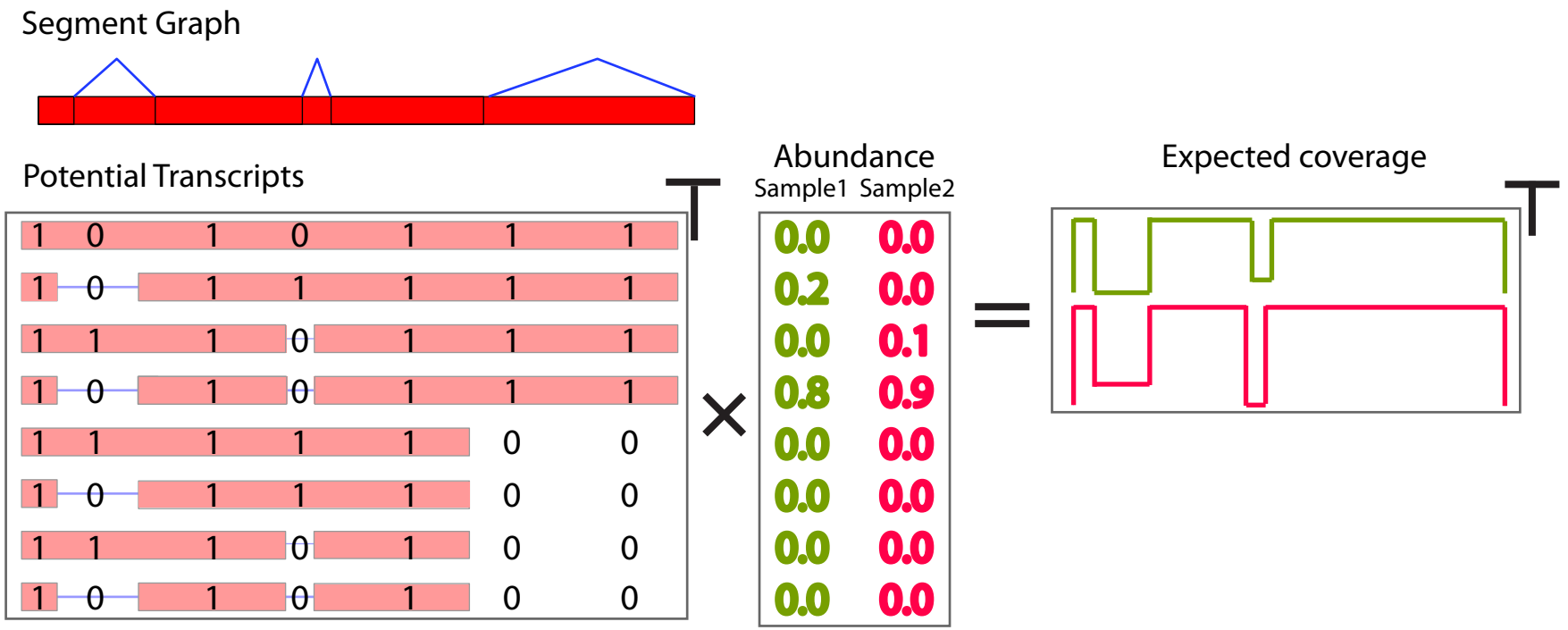
Segment Graph



Potential Transcripts

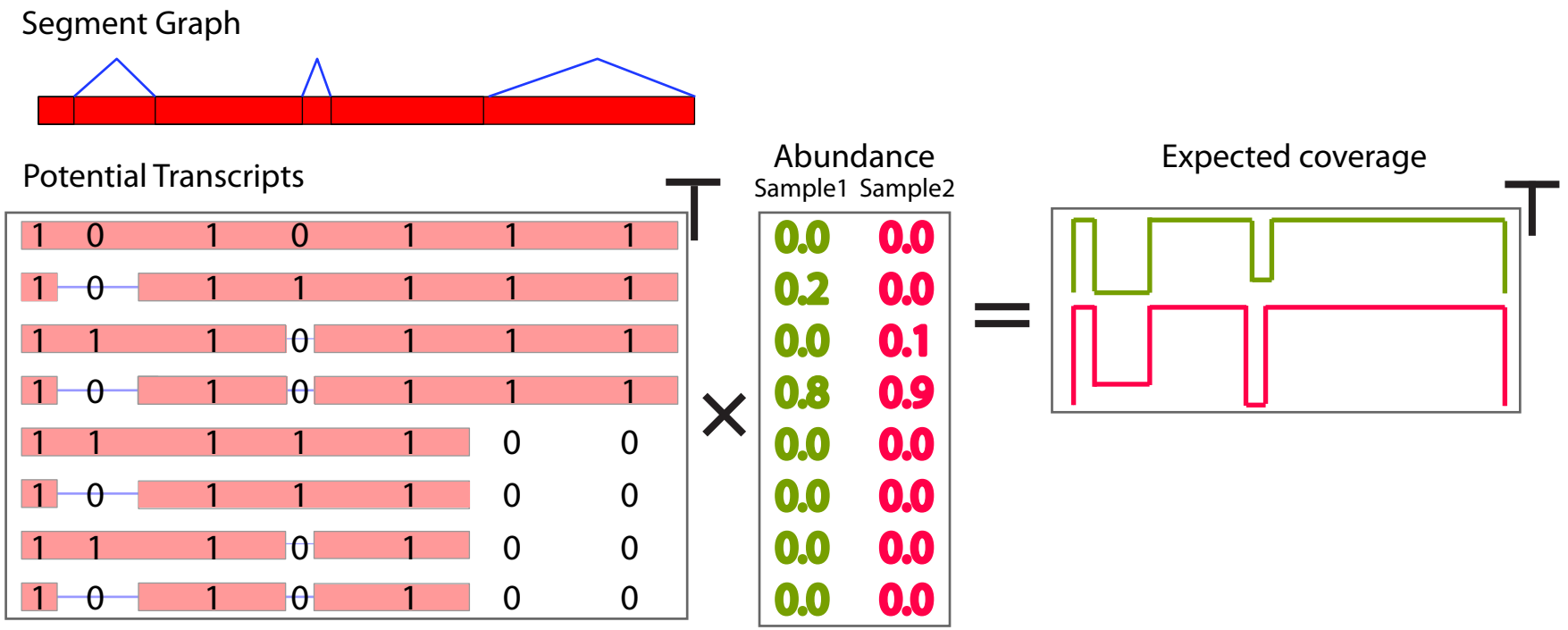


Quantification



R. Bohnert and G. Rätsch, *NAR* (2010)

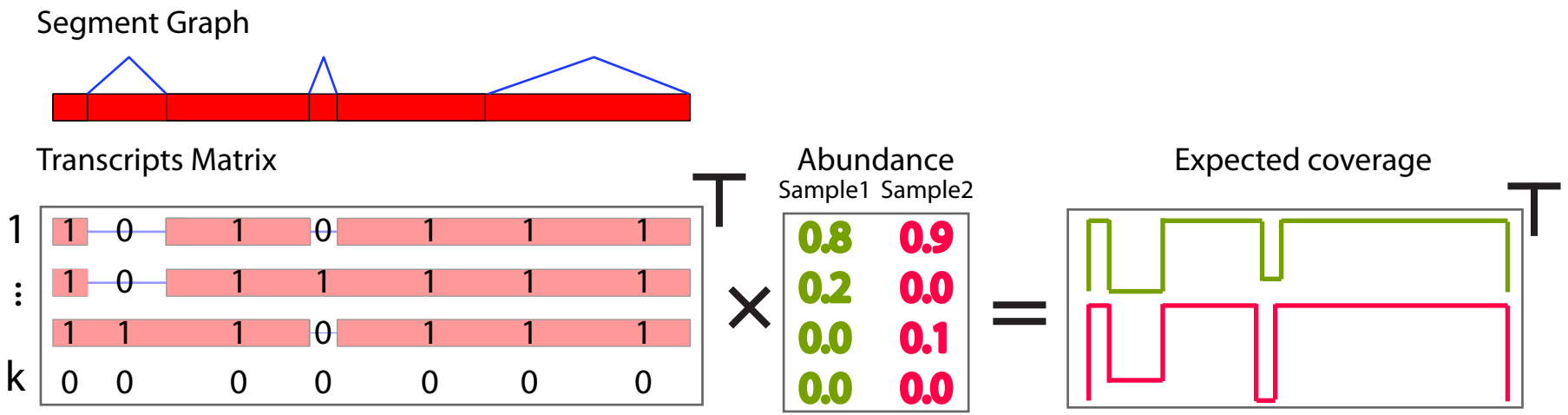
Quantification



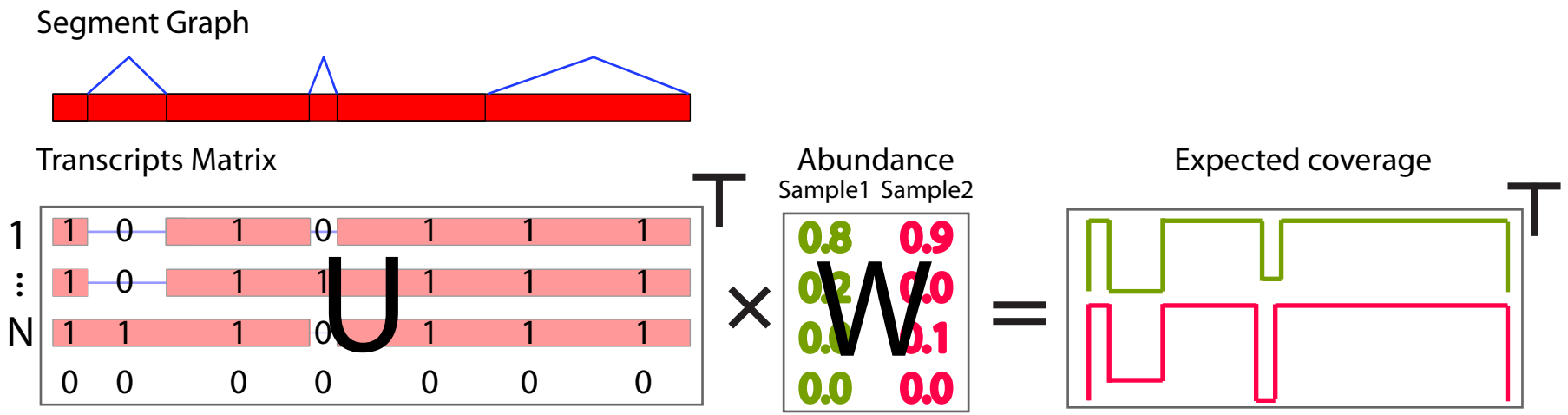
$$\min_W L(\underbrace{U^T \times W}_{\text{expected coverage}}, \underbrace{C}_{\text{observed coverage}}) + \gamma \times \|W\|_1$$

R. Bohnert and G. Rätsch, *NAR* (2010)

Mixed integer optimization

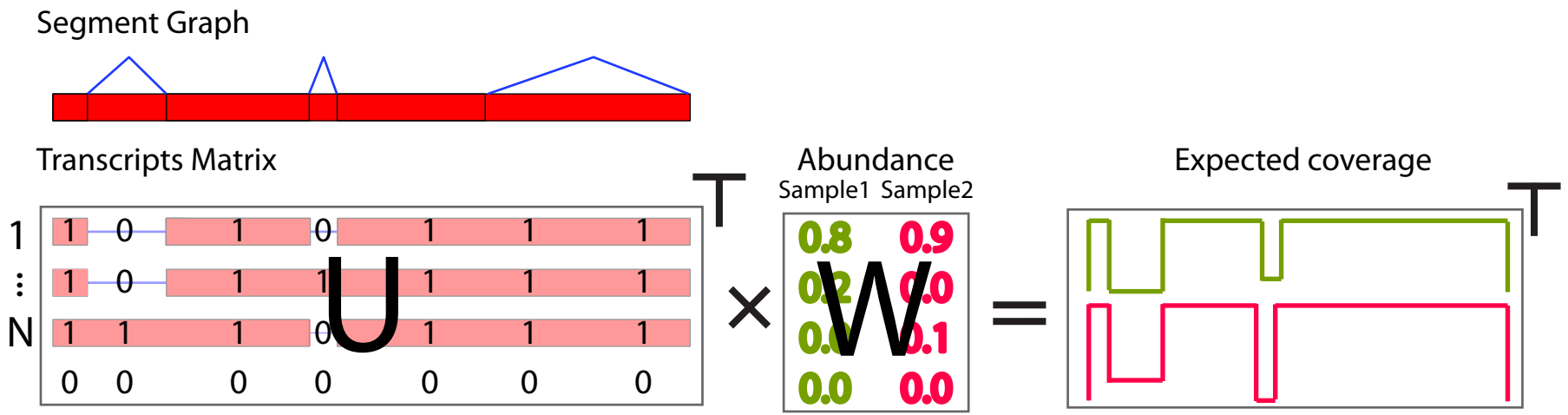


Mixed integer optimization



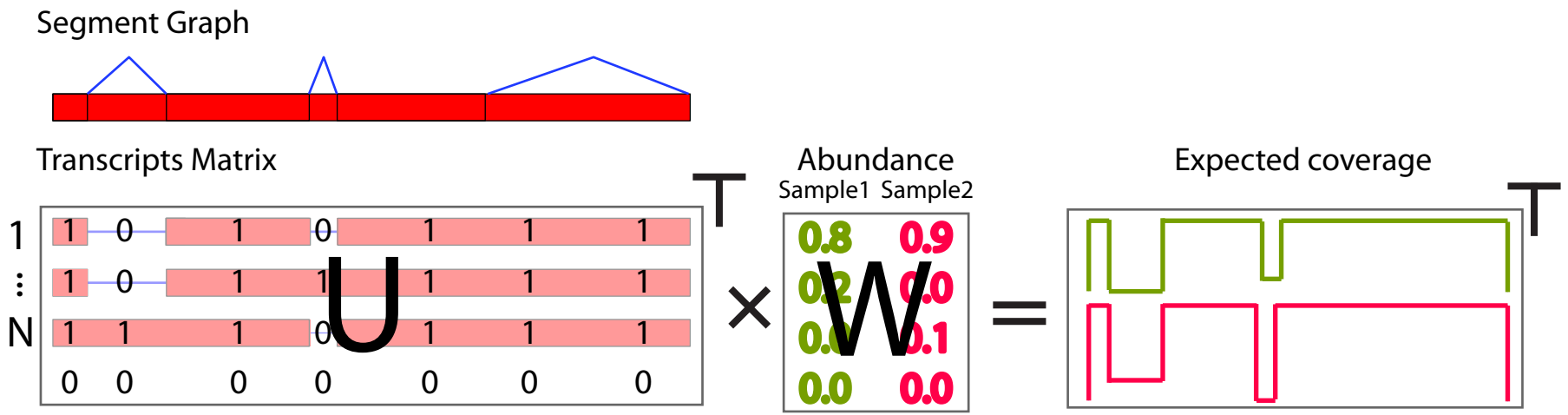
$$\min_{U,W} L(\underbrace{U^T \times W}_{\text{expected coverage}}, \underbrace{C}_{\text{observed coverage}}) + \gamma \times N$$

Mixed integer optimization



$$\min_{U,W} L(U^T \times W, C) + \gamma \times N$$

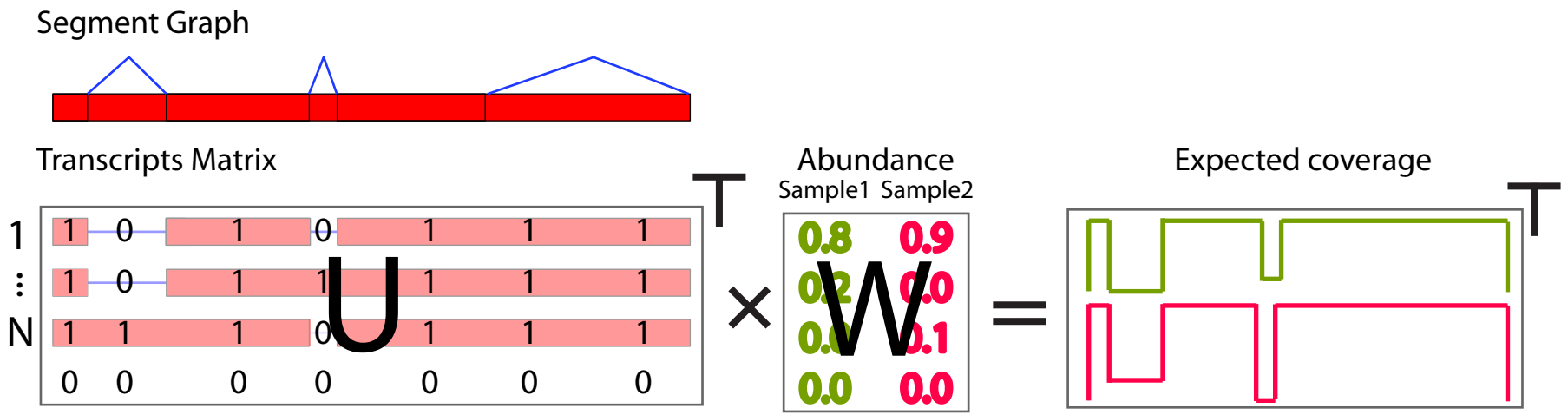
Mixed integer optimization



$$\min_{U,W} L(U^T \times W, C) + \gamma \times N$$

$$s.t. \quad U \text{ is valid}$$

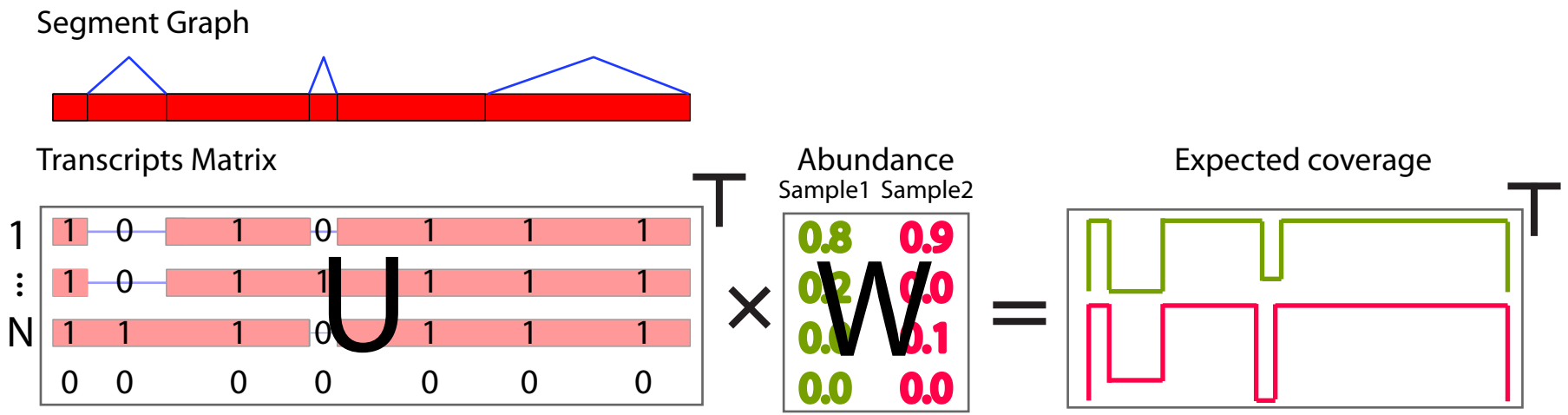
Mixed integer optimization



$$\min_{U,W} L(U^T \times W, C) + \gamma \times N$$

$$s.t. \quad U \text{ is valid}$$

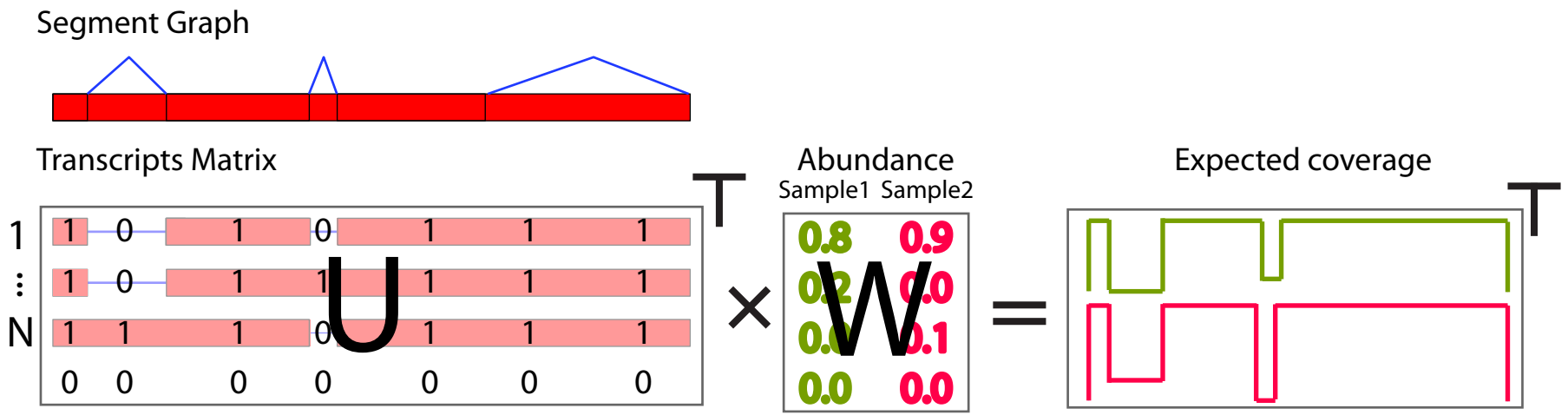
Mixed integer optimization



$$\min_{U,W} L(U^T \times W, C) + \gamma \times N$$

$$s.t. \quad U \text{ is valid}$$

Mixed integer optimization



$$\min_{U,W} L(U^T \times W, C) + \gamma \times N$$

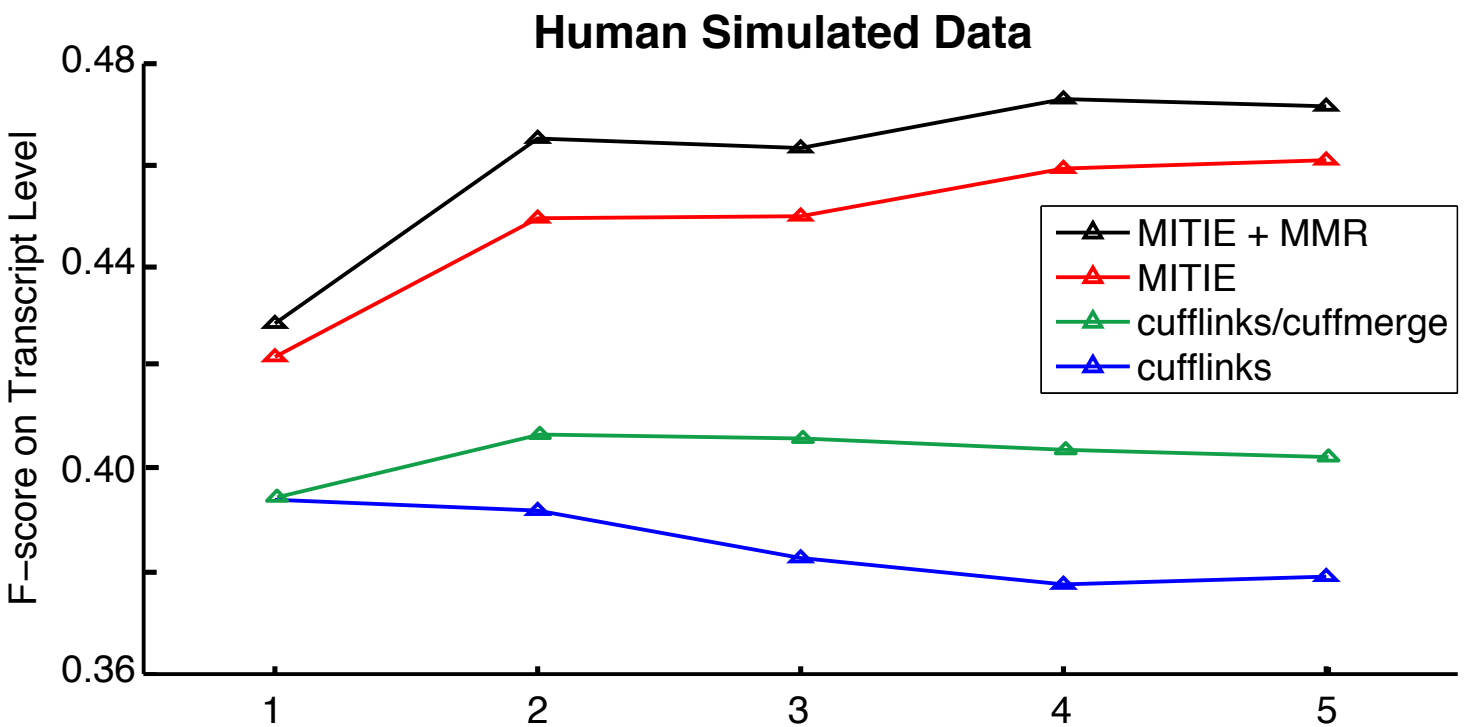
$$s.t. \quad U \text{ is valid}$$

Results: Human artificial data

- Read simulation with Flux Simulator: paired-end, 75 bp
- Transcript with highest expression given
- Evaluation: transcript is correct, if all introns are correct

Results: Human artificial data

- Read simulation with Flux Simulator: paired-end, 75 bp
- Transcript with highest expression given
- Evaluation: transcript is correct, if all introns are correct



Transcript Prediction in Multiple Samples

MiTie: `github.com/ratschlab/mitie`

Characteristics:

- 1-5 seconds for most genes (per sample)
- Uses a proprietary optimization package (CPLEX)

Applications:

- 11 *C. elegans* genomes project [w/ M. De Bono]
- 200 *A. thaliana* genomes MAGIC project [w/ R. Mott & R. Clark]
- ≈5,000 PanCancer genomes transcript prediction

BIOINFORMATICS ORIGINAL PAPER 2013, pages 1–10
doi:10.1093/bioinformatics/btt442

Genome analysis

MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples

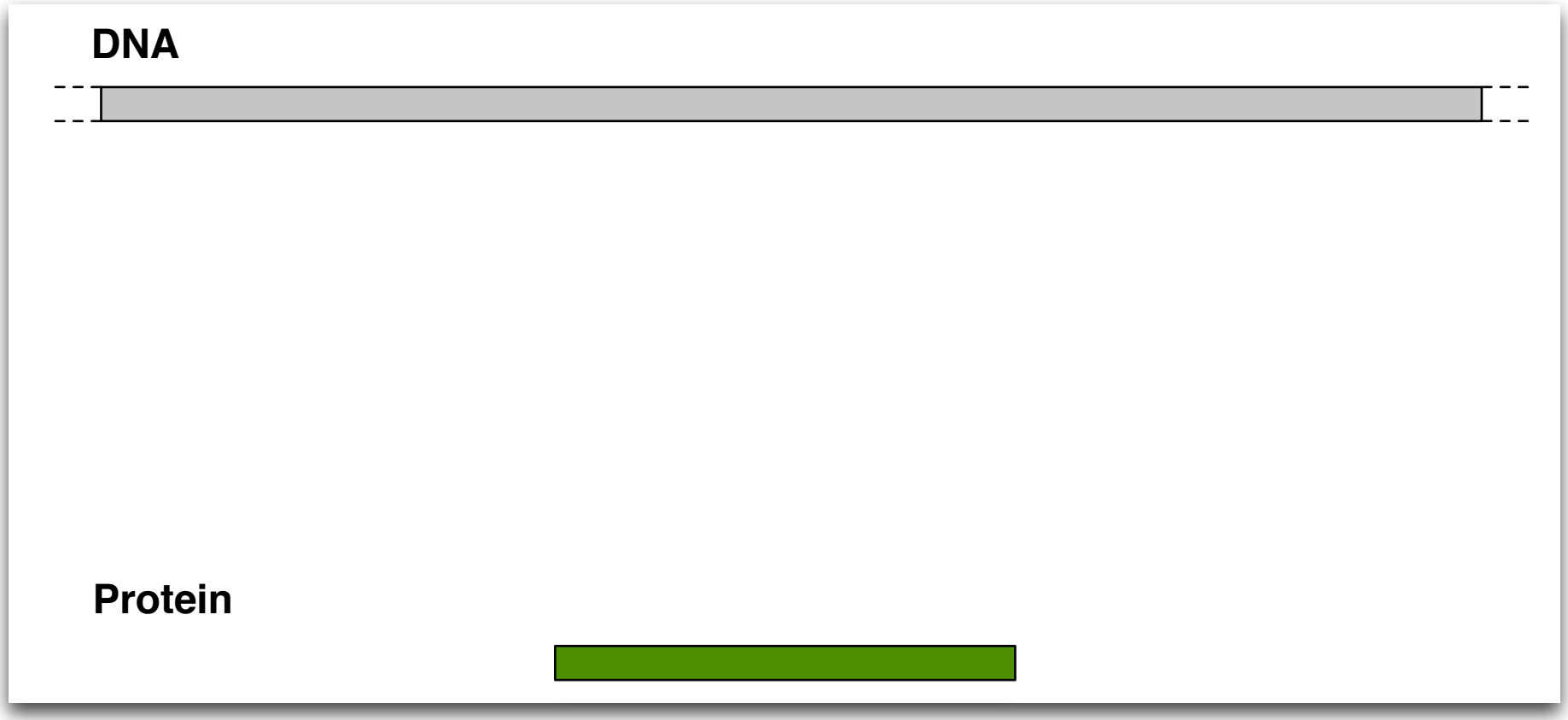
Jonas Behr[1,2,*], André Kahles[1], Yi Zhong[1], Vipin T. Sreedharan[1], Philipp Drewe[1] and Gunnar Rätsch[1,*]

[1]Computational Biology Center, Sloan-Kettering Institute, 1275 York Avenue, New York, NY 10065, USA and [2]Friedrich Miescher Laboratory, Max Planck Society, Spemannstr. 39, 72076 Tu ¨bingen, Germany

Associate Editor: Ivo Hofacker

RNA-seq-based Gene Finding

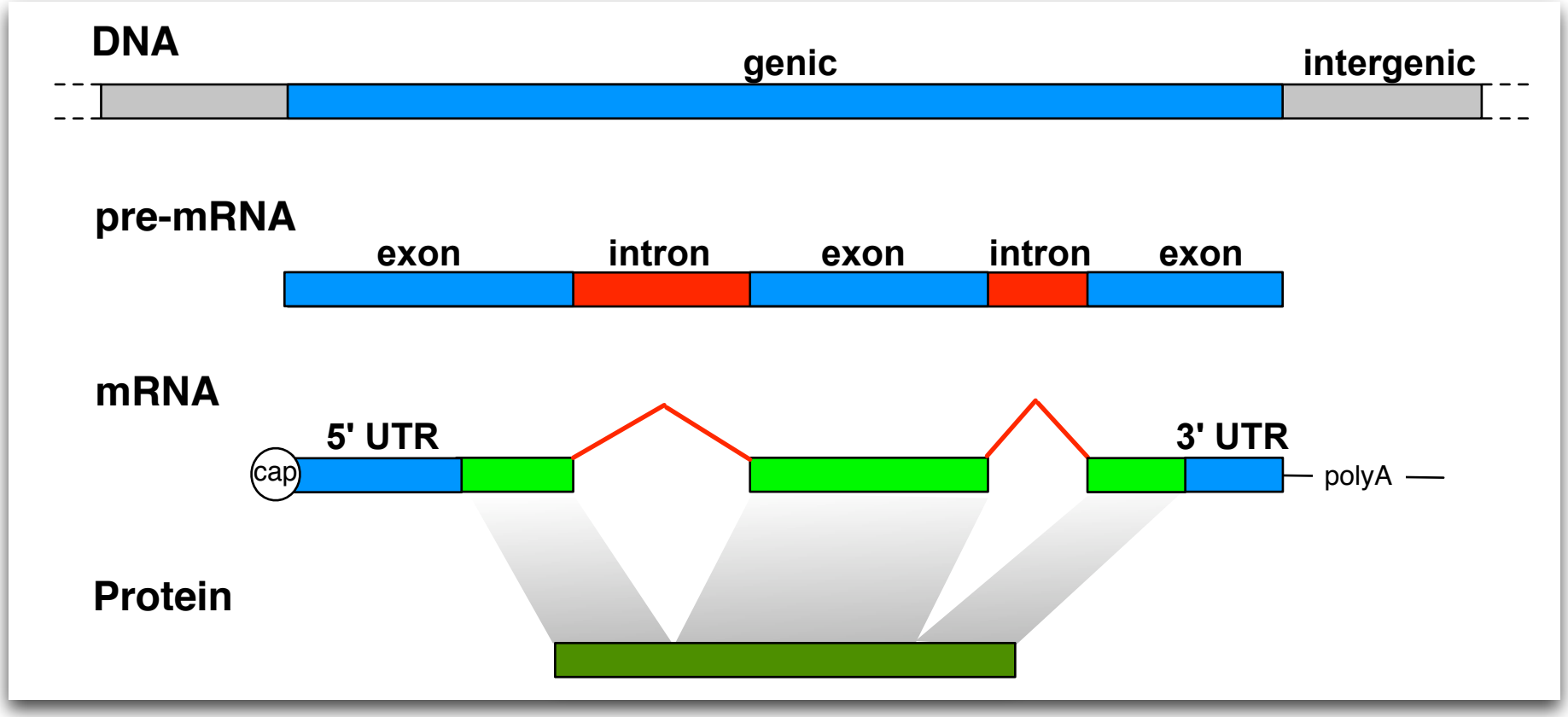
↝ Labeling the Genome



- Given a piece of DNA sequence
- Predict proteins (or non-coding RNAs)

RNA-seq-based Gene Finding

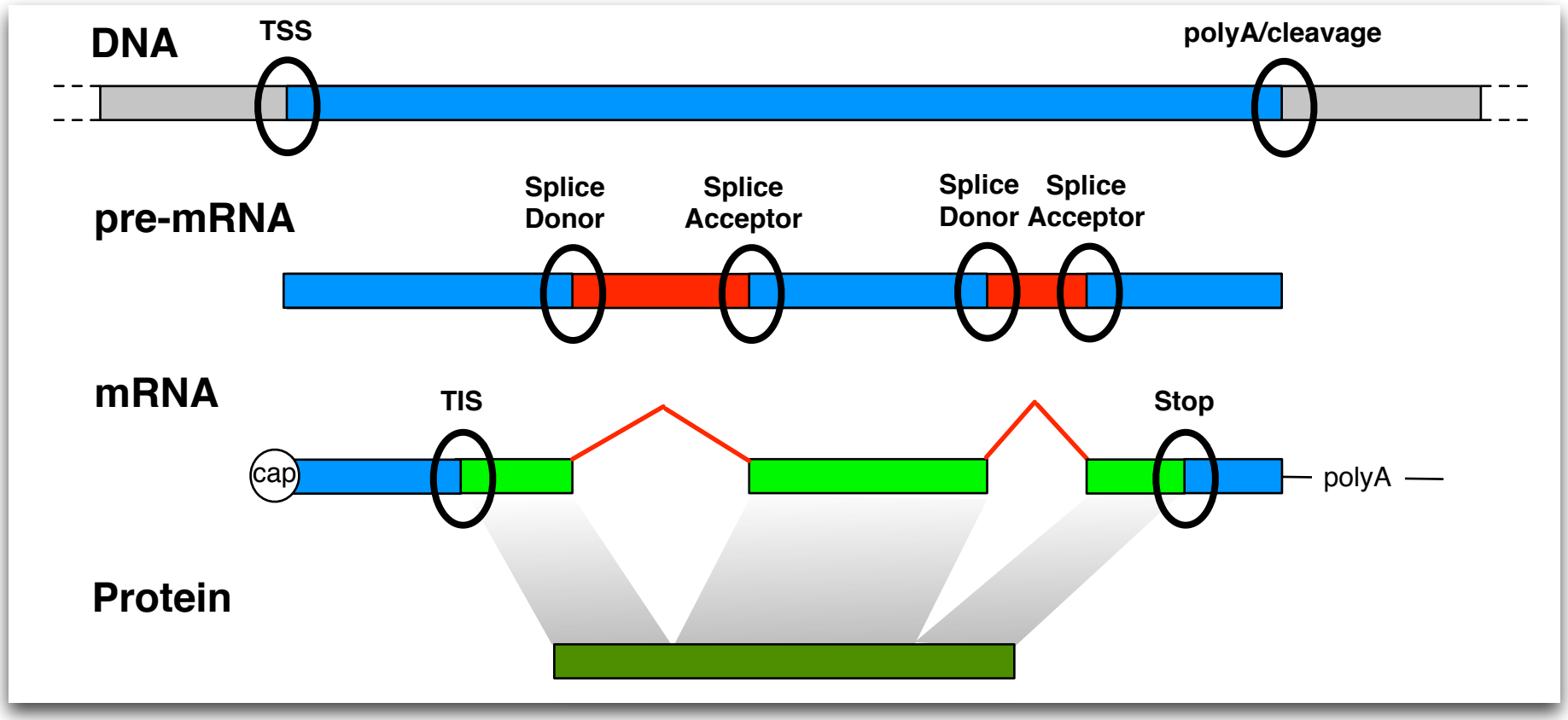
⇝ Labeling the Genome



- Given a piece of DNA sequence
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

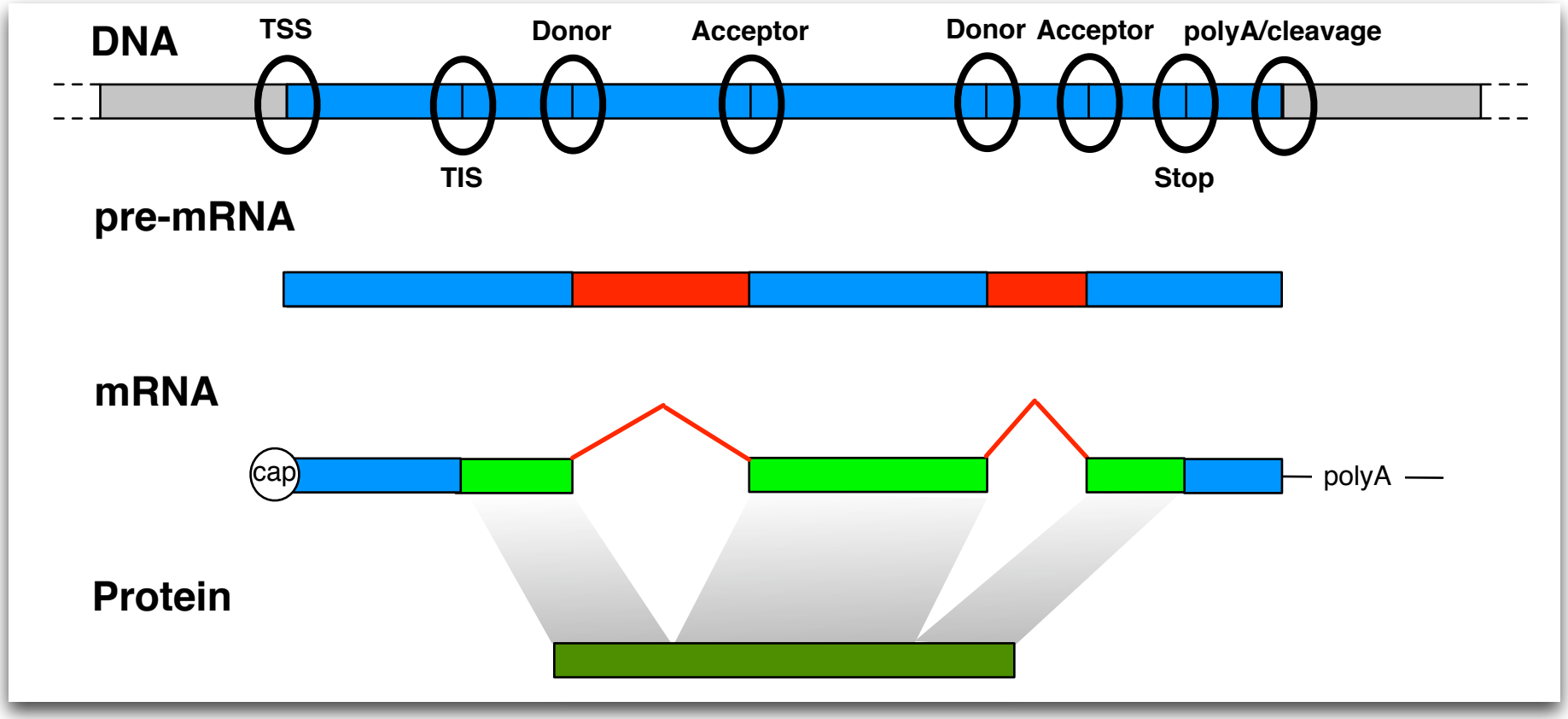
RNA-seq-based Gene Finding

⇝ Labeling the Genome



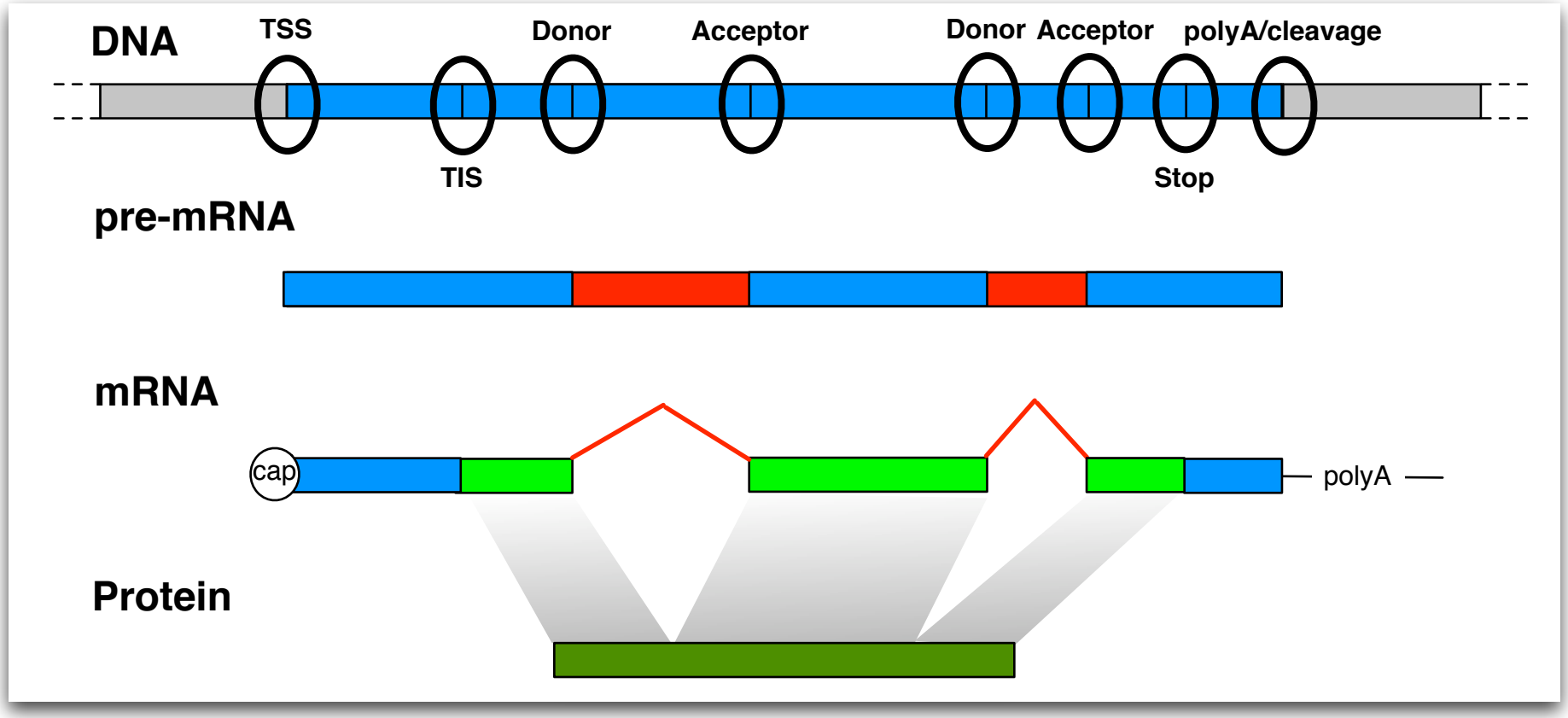
RNA-seq-based Gene Finding

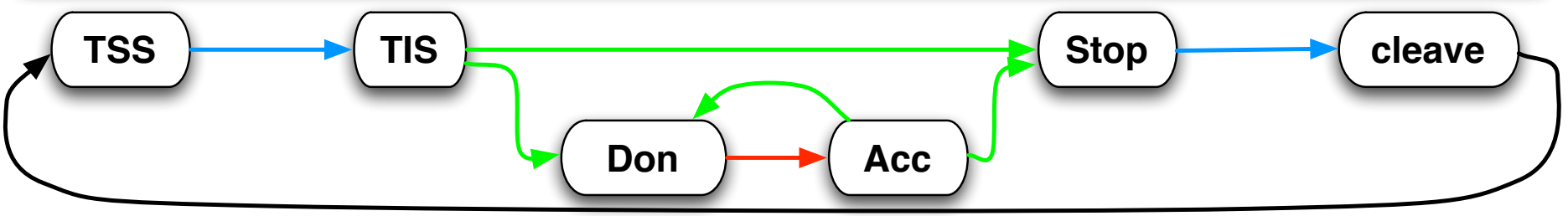
⇝ Labeling the Genome



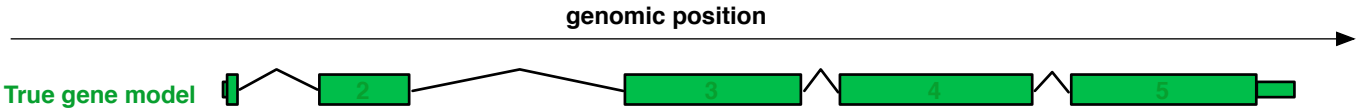
RNA-seq-based Gene Finding

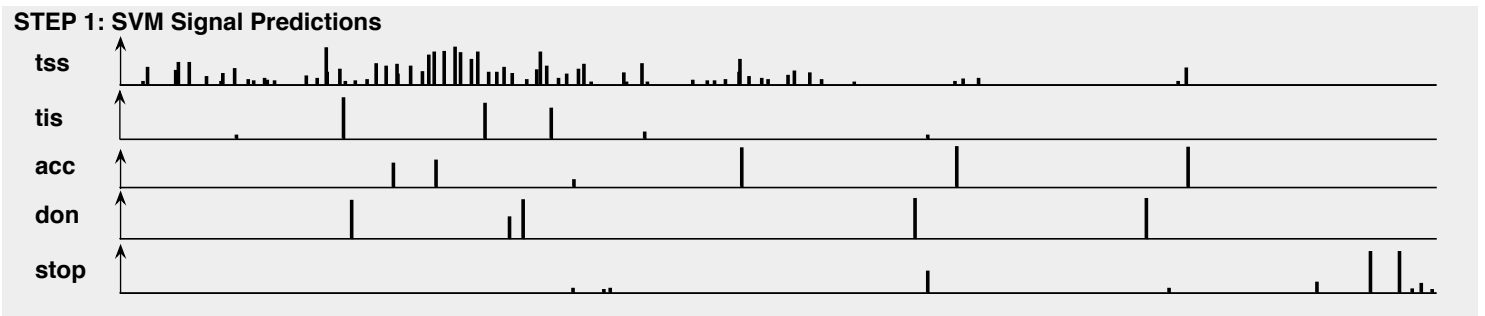
⇝ Labeling the Genome



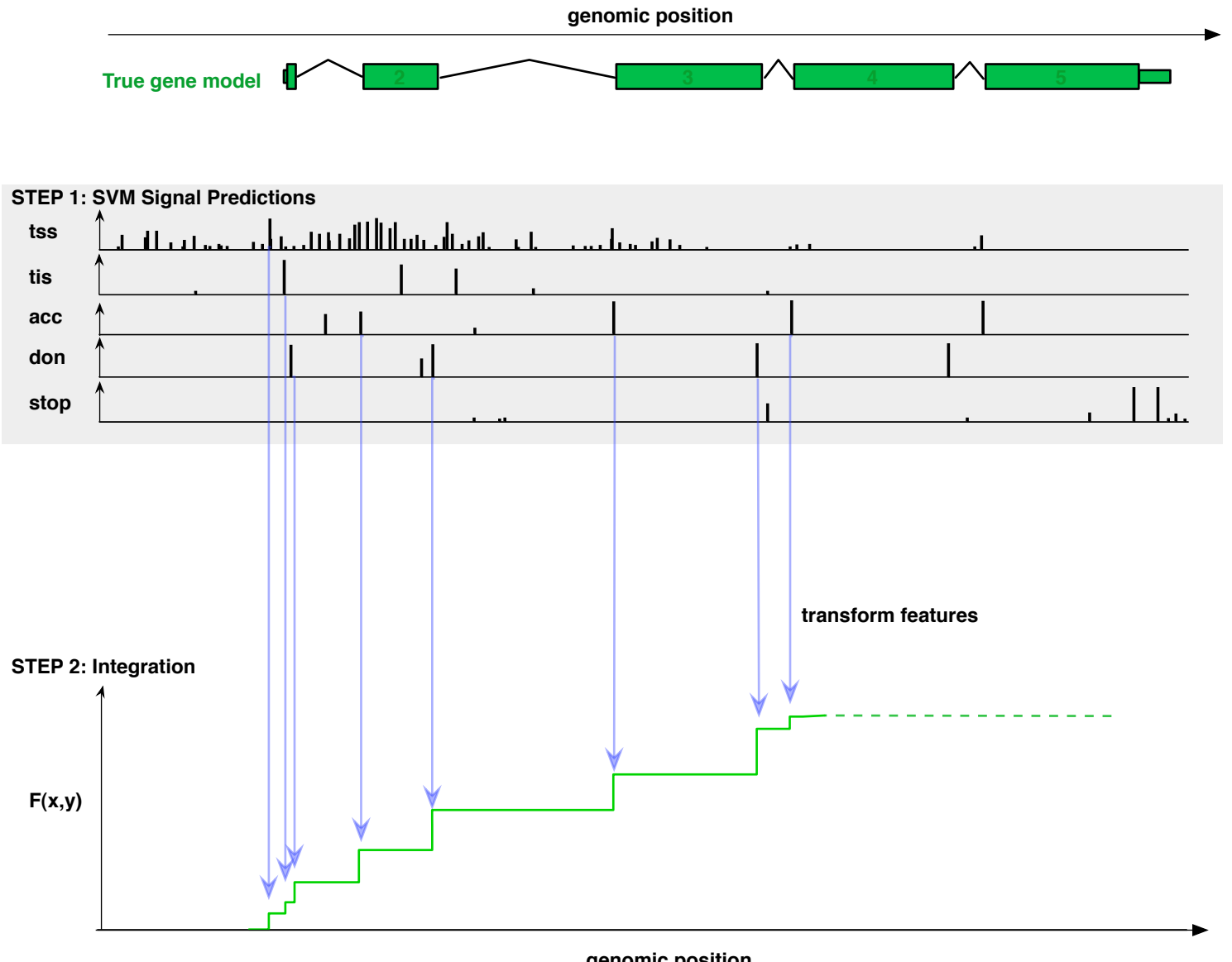


mGene-based Transcript Prediction

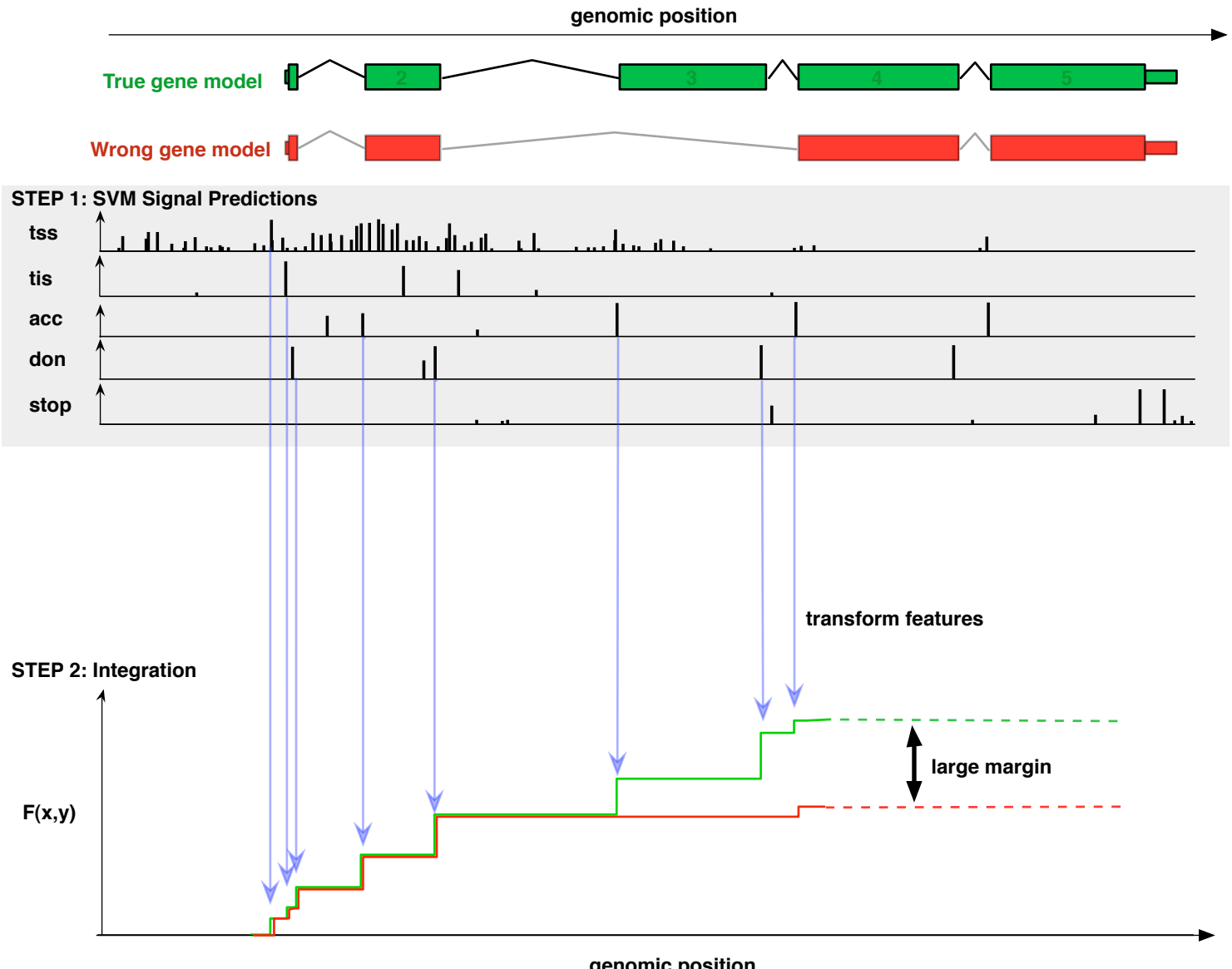




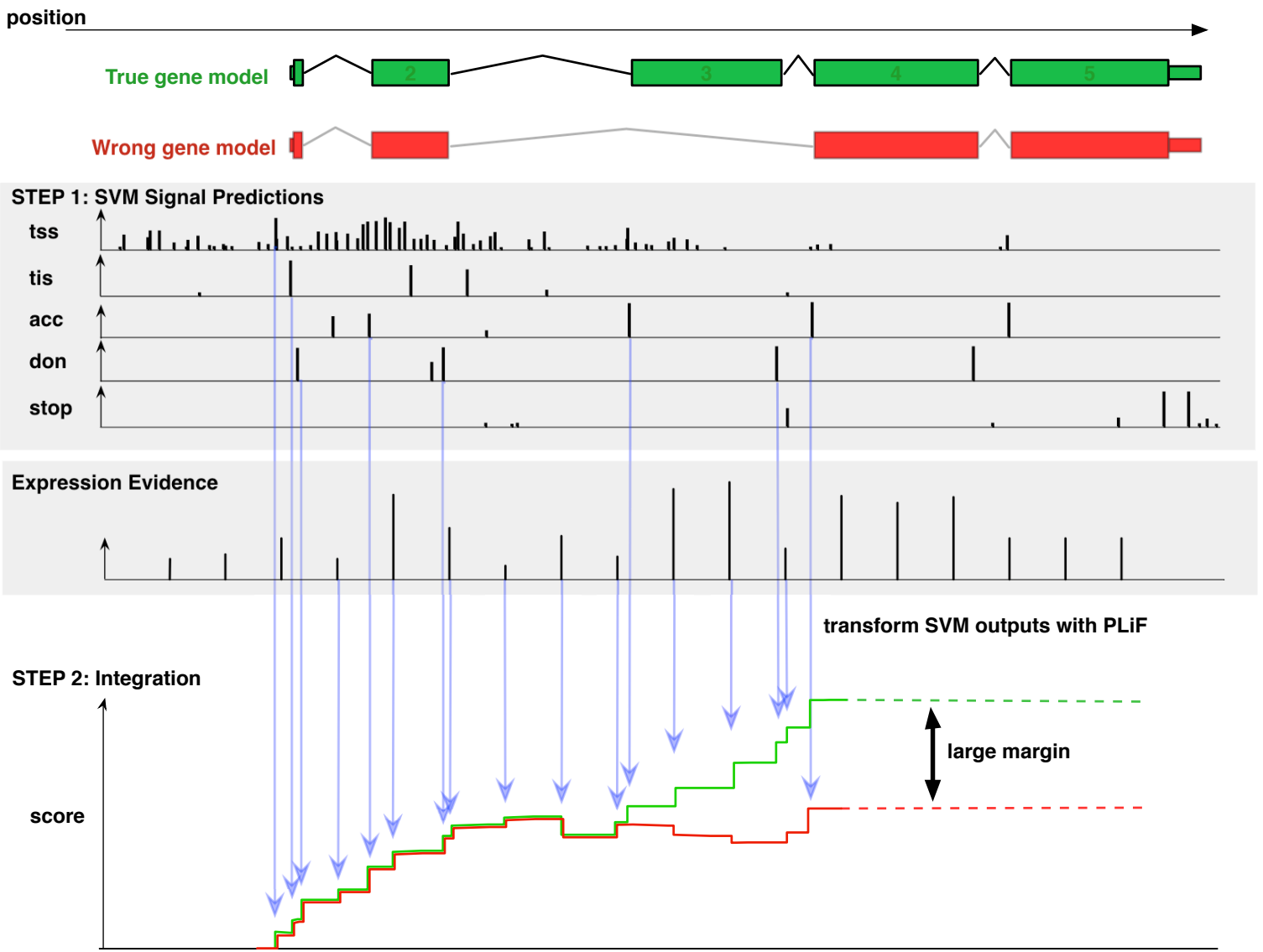
mGene-based Transcript Prediction



mGene-based Transcript Prediction



mGene-based Transcript Prediction



Digestion

- Accurate isoform quantification is non-trivial even if read mapping is accurate
- Most methods for transcript prediction crucially depend on accurate alignments
- Major advances in RNA-seq read mapping, but not solved yet:
  - Multi-mapping reads
  - Effect of polymorphisms
  - Biases by library prep method/fragment size
- Practical advice for quantitative studies:
  - Use exactly the same library prep protocol (same person, same fragment size, sequenced ideally from the same flowcell)
  - Use same read mapping program (and version) for all libraries
  - Care about multi-mappers and polymorphisms
- Use methods capable of taking uncertainty into account

Digestion

- Accurate isoform quantification is non-trivial even if read mapping is accurate
- Most methods for transcript prediction crucially depend on accurate alignments
- Major advances in RNA-seq read mapping, but not solved yet:
  - Multi-mapping reads
  - Effect of polymorphisms
  - Biases by library prep method/fragment size
- Practical advice for quantitative studies:
  - Use exactly the same library prep protocol (same person, same fragment size, sequenced ideally from the same flowcell)
  - Use same read mapping program (and version) for all libraries
  - Care about multi-mappers and polymorphisms
- Use methods capable of taking uncertainty into account

Digestion

- Accurate isoform quantification is non-trivial even if read mapping is accurate
- Most methods for transcript prediction crucially depend on accurate alignments
- Major advances in RNA-seq read mapping, but not solved yet:
  - Multi-mapping reads
  - Effect of polymorphisms
  - Biases by library prep method/fragment size
- Practical advice for quantitative studies:
  - Use exactly the same library prep protocol (same person, same fragment size, sequenced ideally from the same flowcell)
  - Use same read mapping program (and version) for all libraries
  - Care about multi-mappers and polymorphisms
- Use methods capable of taking uncertainty into account

Digestion

- Accurate isoform quantification is non-trivial even if read mapping is accurate
- Most methods for transcript prediction crucially depend on accurate alignments
- Major advances in RNA-seq read mapping, but not solved yet:
  - Multi-mapping reads
  - Effect of polymorphisms
  - Biases by library prep method/fragment size
- Practical advice for quantitative studies:
  - Use exactly the same library prep protocol (same person, same fragment size, sequenced ideally from the same flowcell)
  - Use same read mapping program (and version) for all libraries
  - Care about multi-mappers and polymorphisms
- Use methods capable of taking uncertainty into account

Digestion

- Accurate isoform quantification is non-trivial even if read mapping is accurate
- Most methods for transcript prediction crucially depend on accurate alignments
- Major advances in RNA-seq read mapping, but not solved yet:
  - Multi-mapping reads
  - Effect of polymorphisms
  - Biases by library prep method/fragment size
- Practical advice for quantitative studies:
  - Use exactly the same library prep protocol (same person, same fragment size, sequenced ideally from the same flowcell)
  - Use same read mapping program (and version) for all libraries
  - Care about multi-mappers and polymorphisms
- Use methods capable of taking uncertainty into account

Acknowledgments

We gratefully acknowledge help with the material:

- **Jonas Behr**, FML & Sloan-Kettering Institute
- Regina Bohnert, FML & Molecular Health
- **Philipp Drewe**, FML & Sloan-Kettering Institute
- Fabio De Bona, FML & Google Research
- **André Kahles**, FML & Sloan-Kettering Institute
- Stefan Henz, MPI for Developmental Biology
- Cole Trapnell, Broad Institute
- Ali Mortazavi, UC Irvine
- Georg Zeller, FML & EMBL Heidelberg
- David Kuo, Sloan-Kettering Institute & Weill-Cornell
- Katherine Redfield Chan, Sloan-Kettering Institute

The slides and additional material will be available online at

`http://raetschlab.org/lectures/ows2013/`

References I

J. Behr, G. Schweikert, J. Cao, F. De Bona, G. Zeller, S. Laubinger, S. Ossowski, K. Schneeberger, D. Weigel, and G. Rätsch. Rna-seq and tiling arrays for improved gene finding. Oral presentation at the CSHL Genome Informatics Meeting, September 2008. URL `http://www.fml.tuebingen.mpg.de/raetsch/lectures/RaetschGenomeInformatics08.pdf`.

J. Behr, A. Kahles, Y. Zhong, and G. Rätsch. Mitie: Accurate transcript prediction with integer programming. *Bioinformatics*, 2013. under revision.

R. Bohnert, J. Behr, and G Rätsch. Transcript quantification with RNA-Seq data. *BMC Bioinformatics*, 10(S13):P5, 2009. URL `http://www.biomedcentral.com/1471-2105/10/S13/P5`.

RM Clark, G Schweikert, C Toomajian, S Ossowski, G Zeller, P Shinn, N Warthmann, TT Hu, G Fu, DA Hinds, H Chen, KA Frazer, DH Huson, B Schölkopf, M Nordborg, G Rätsch, JR Ecker, and D Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–342, 2007. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1138632.

France Denoeud, Jean-Marc Aury, Corinne Da Silva, Benjamin Noel, Odile Rogier, Massimo Delledonne, Michele Morgante, Giorgio Valle, Patrick Wincker, Claude Scarpelli, Olivier Jaillon, and François Artiguenave. Annotating genomes with massive-scale rna sequencing. *Genome Biol*, 9(12):R175, 2008. doi: 10.1186/gb-2008-9-12-r175.

References II

Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, Sep 2008. doi: 10.1093/nar/gkn425.

Philipp Drewe, Oliver Stegle, Lisa Hartmann, André Kahles, Regina Bohnert, Andreas Wachter, Karsten Borgwardt, and Gunnar Rätsch. Accurate detection of differential rna processing. *Nucleic Acids Res*, 41(10):5189–98, May 2013. doi: 10.1093/nar/gkt211.

Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G Steffen, Philipp Drewe, Katie L Hildebrand, Rune Lyngsoe, Sebastian J Schultheiss, Edward J Osborne, Vipin T Sreedharan, André Kahles, Regina Bohnert, Géraldine Jean, Paul Derwent, Paul Kersey, Eric J Belfield, Nicholas P Harberd, Eric Kemen, Christopher Toomajian, Paula X Kover, Richard M Clark, Gunnar Rätsch, and Richard Mott. Multiple reference genomes and transcriptomes for arabidopsis thaliana. *Nature*, 477(7365):419–23, Sep 2011. doi: 10.1038/nature10414.

Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol*, 28(5): 503–10, May 2010. doi: 10.1038/nbt.1633.

Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, Apr 2010. doi: 10.1093/nar/gkq224.

References III

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, April 2009.

Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12): 1009–15, Dec 2010. doi: 10.1038/nmeth.1528.

Vincent Lacroix, Michael Sammeth, Roderic Guigó, and Anne Bergeron. Exact transcriptome reconstruction from short sequence reads. In *WABI*, pages 50–63, 2008.

Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, February 2010a. doi: 10.1093/bioinformatics/btp692. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/4/493`.

Jun Li, Hui Jiang, and Wing Hung Wong. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):R50, 2010b. doi: 10.1186/gb-2010-11-5-r50.

Sam E V Linsen, Elzo de Wit, Georges Janssens, Sheila Heater, Laura Chapman, Rachael K Parkin, Brian Fritz, Stacia K Wyman, Ewart de Bruijn, Emile E Voest, Scott Kuersten, Muneesh Tewari, and Edwin Cuppen. Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods*, 6(7):474–476, July 2009. ISSN 1548-7091. doi: 10.1038/nmeth0709-474. URL `http://dx.doi.org/10.1038/nmeth0709-474`.

Alicia Oshlack and Matthew Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(1):14, 2009. doi: 10.1186/1745-6150-4-14.

References IV

G. Rätsch and S. Sonnenburg. Accurate splice site detection for *Caenorhabditis elegans*. In K. Tsuda B. Schoelkopf and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21(Suppl. 1):i369–i377, June 2005.

Hugues Richard, Marcel H Schulz, Marc Sultan, Asja Nürnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A Haas, and Marie-Laure Yaspo. Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic Acids Res*, 38(10):e112, Jun 2010b. doi: 10.1093/nar/gkq041.

Hugues Richard, Marcel H. Schulz, Marc Sultan, Asja Nurnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A. Haas, and Marie-Laure Yaspo. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, page gkq041, February 2010a. doi: 10.1093/nar/gkq041. URL `http://nar.oxfordjournals.org/cgi/content/abstract/gkq041v1`.

M. Sammeth. The Flux Capacitor. *Website*, 2009a. http://flux.sammeth.net/capacitor.html.

M. Sammeth. The Flux Simulator. *Website*, 2009b. http://flux.sammeth.net/simulator.html.

References V

Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krüger, Sören Sonnenburg, and Gunnar Rätsch. mgene: Accurate svm-based gene finding with an application to nematode genomes. *Genome Research*, 2009. URL `http://genome.cshlp.org/content/early/2009/06/29/gr.090597.108.full.pdf+html`. Advance access June 29, 2009.

S. Sonnenburg, G. Rätsch, A. Jagota, and K.-R. Müller. New methods for splice-site recognition. In *Proc. International Conference on Artificial Neural Networks*, 2002.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):e472–480, 2006.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–5, May 2010a. doi: 10.1038/nbt.1621.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotech*, advance online publication, May 2010b. doi: 10.1038/nbt.1621. URL `http://dx.doi.org/10.1038/nbt.1621`.

E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008. ISSN 1476-4687 (Electronic). doi: 10.1038/nature07509.

References VI

G Zeller, RM Clark, K Schneeberger, A Bohlen, D Weigel, and G Ratsch. Detecting polymorphic regions in arabidopsis thaliana with resequencing microarrays. *Genome Res*, 18 (6):918–929, 2008. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *BioInformatics*, 16(9): 799–807, September 2000.