Functional Genomics by Deep Sequencing

Gunnar Rätsch

Friedrich Miescher Laboratory, Max Planck Society

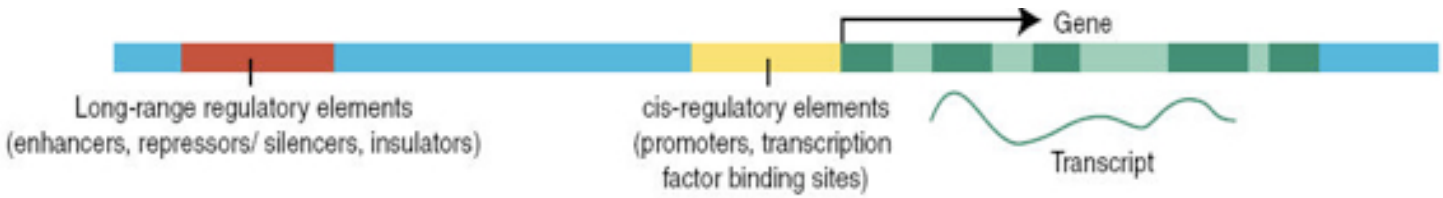
Tübingen, Germany

Erice Course, September 14, 2010

The Encyclopedia of DNA Elements



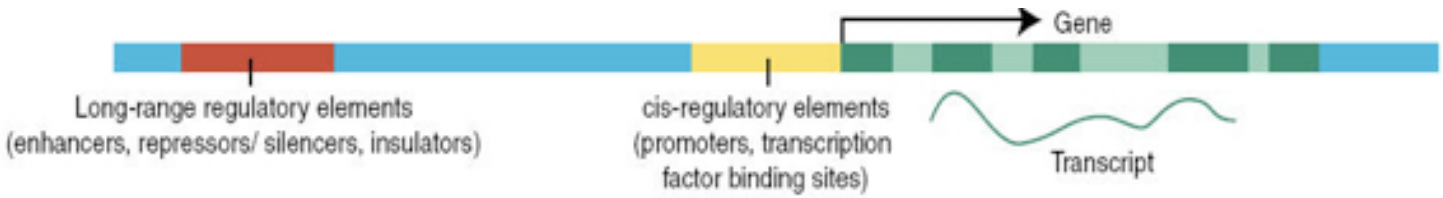
- How to make sense of the genome?
  - Which nucleotides are functional?
  - What is their function?



- National Human Genome Research Institute started the Encode project to annotate human and model organism genomes:
  - 2004: Human 1%
  - 2007: Human whole-genome
  - 2008: *Drosophila* and *C. elegans*
  - 2010: Mouse

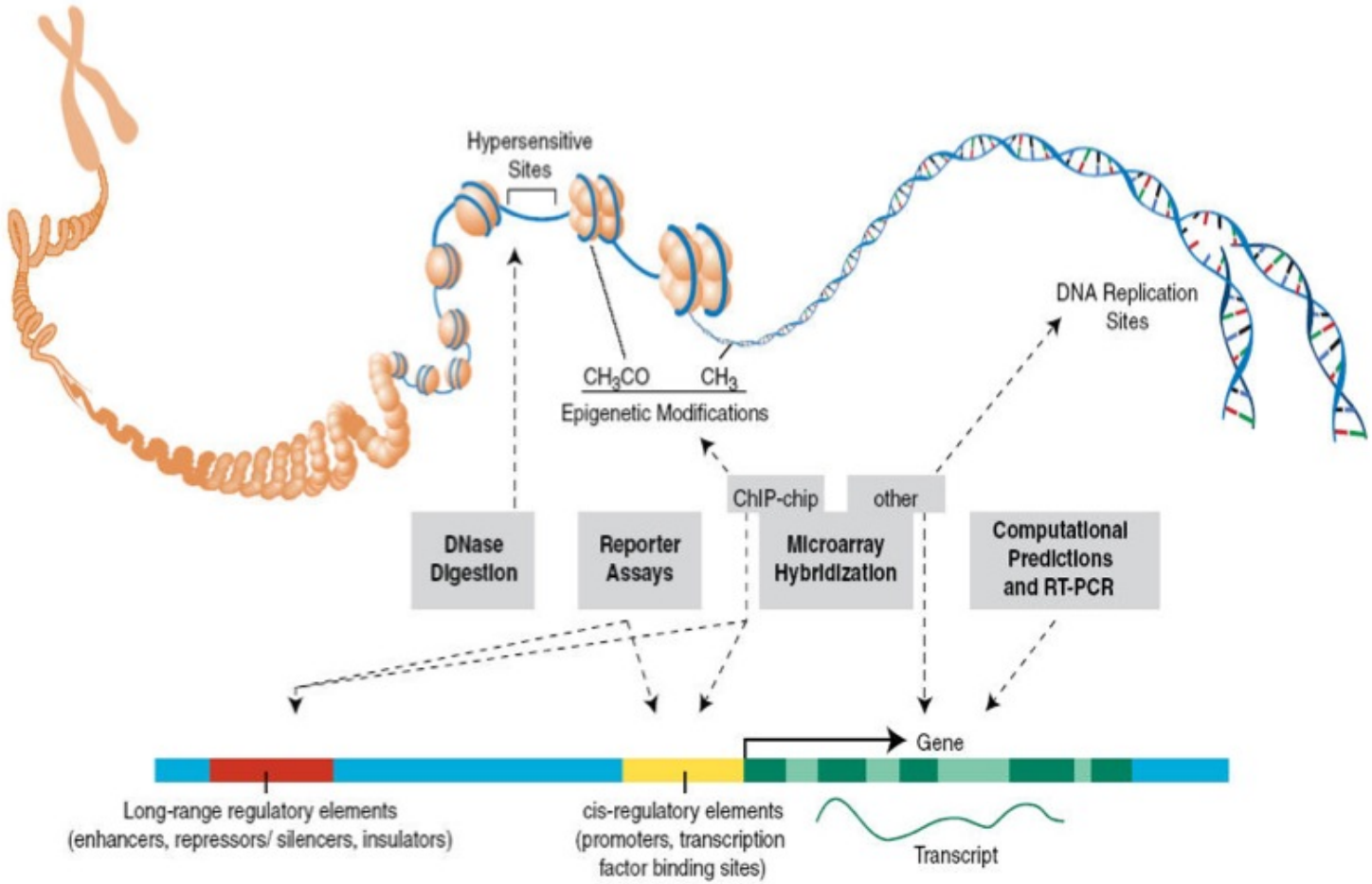
The Encyclopedia of DNA Elements

- How to make sense of the genome?
  - Which nucleotides are functional?
  - What is their function?



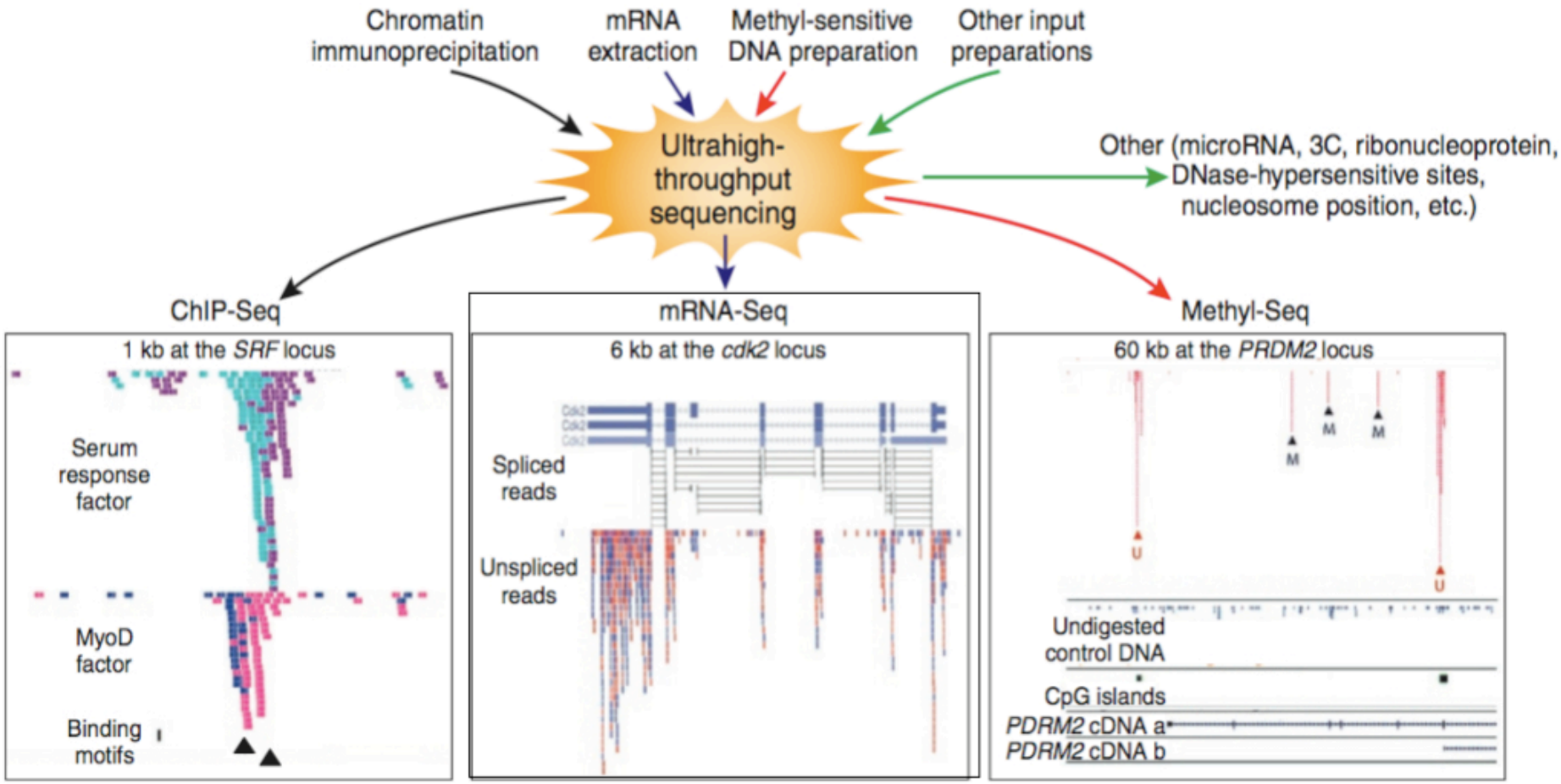
- National Human Genome Research Institute started the Encode project to annotate human and model organism genomes:
  - 2004: Human 1%
  - 2007: Human whole-genome
  - 2008: *Drosophila* and *C. elegans*
  - 2010: Mouse

Originally, microarrays were used to read out genome-wide functional assays



ENCODE Project Consortium (2004). Science 306: 636.

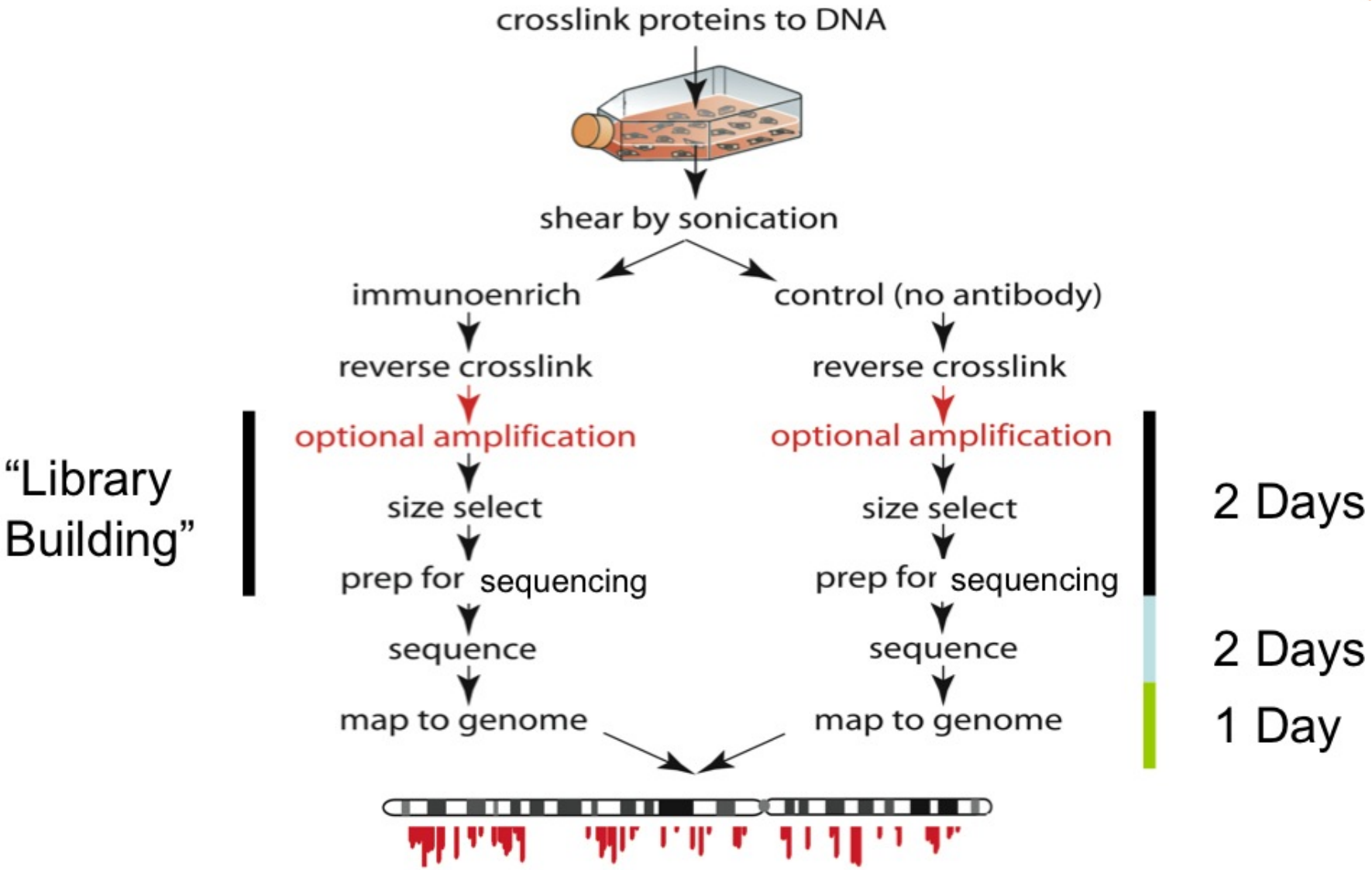
Sequencing for Functional Genomics



[Wold & Myers, Nature Methods, 2007]

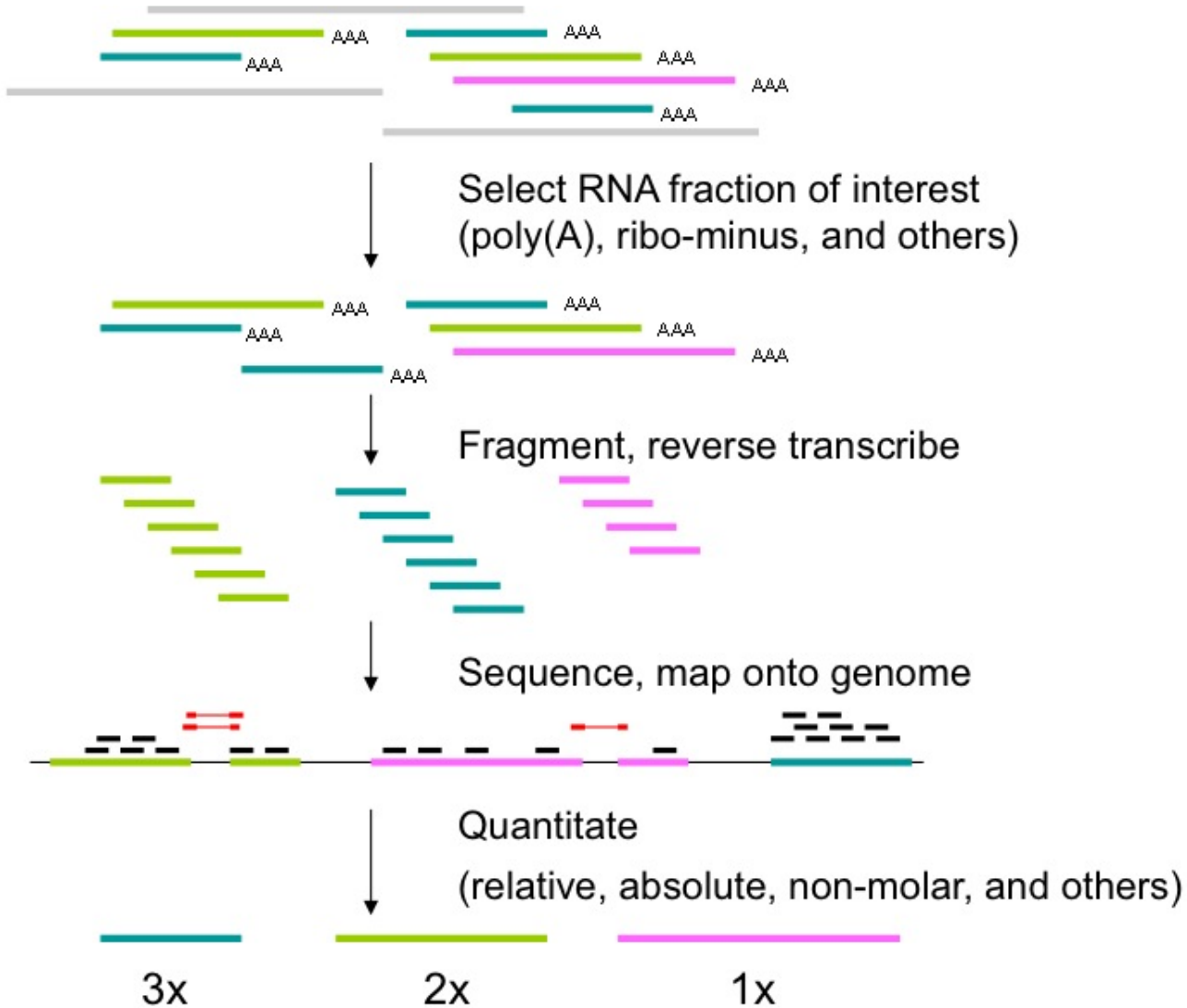
ChIP-seq

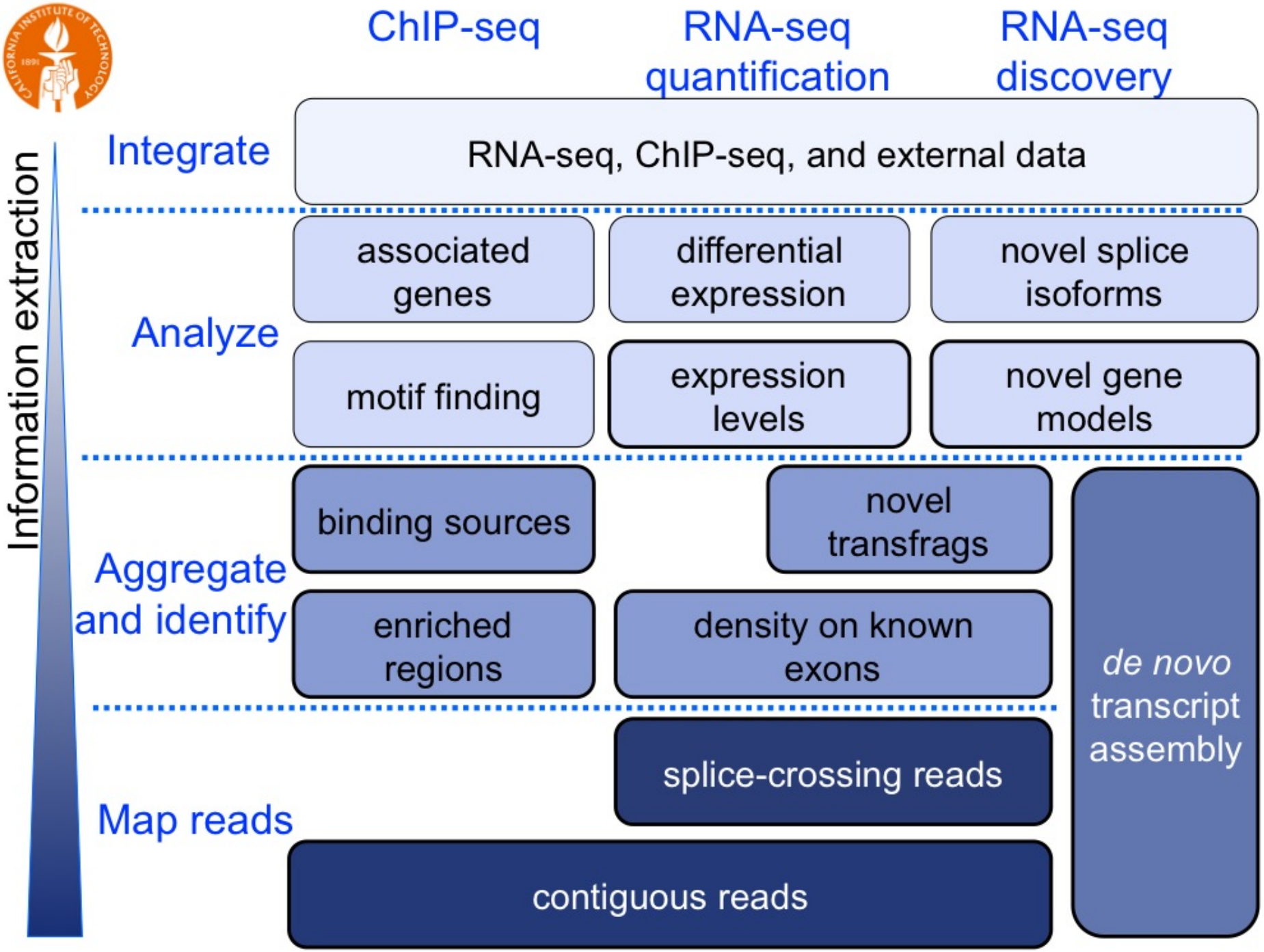
A digital counting method to score site occupancy by DNA binding proteins



RNA-seq

A digital counting method for transcriptome discovery and quantification




ChIP-seq
RNA-seq quantification
RNA-seq discovery
Information extraction
Integrate
RNA-seq, ChIP-seq, and external data
Analyze
associated genes
differential expression
novel splice isoforms
motif finding
expression levels
novel gene models
Aggregate and identify
binding sources
novel transfrags
enriched regions
density on known exons
de novo transcript assembly
Map reads
splice-crossing reads
contiguous reads

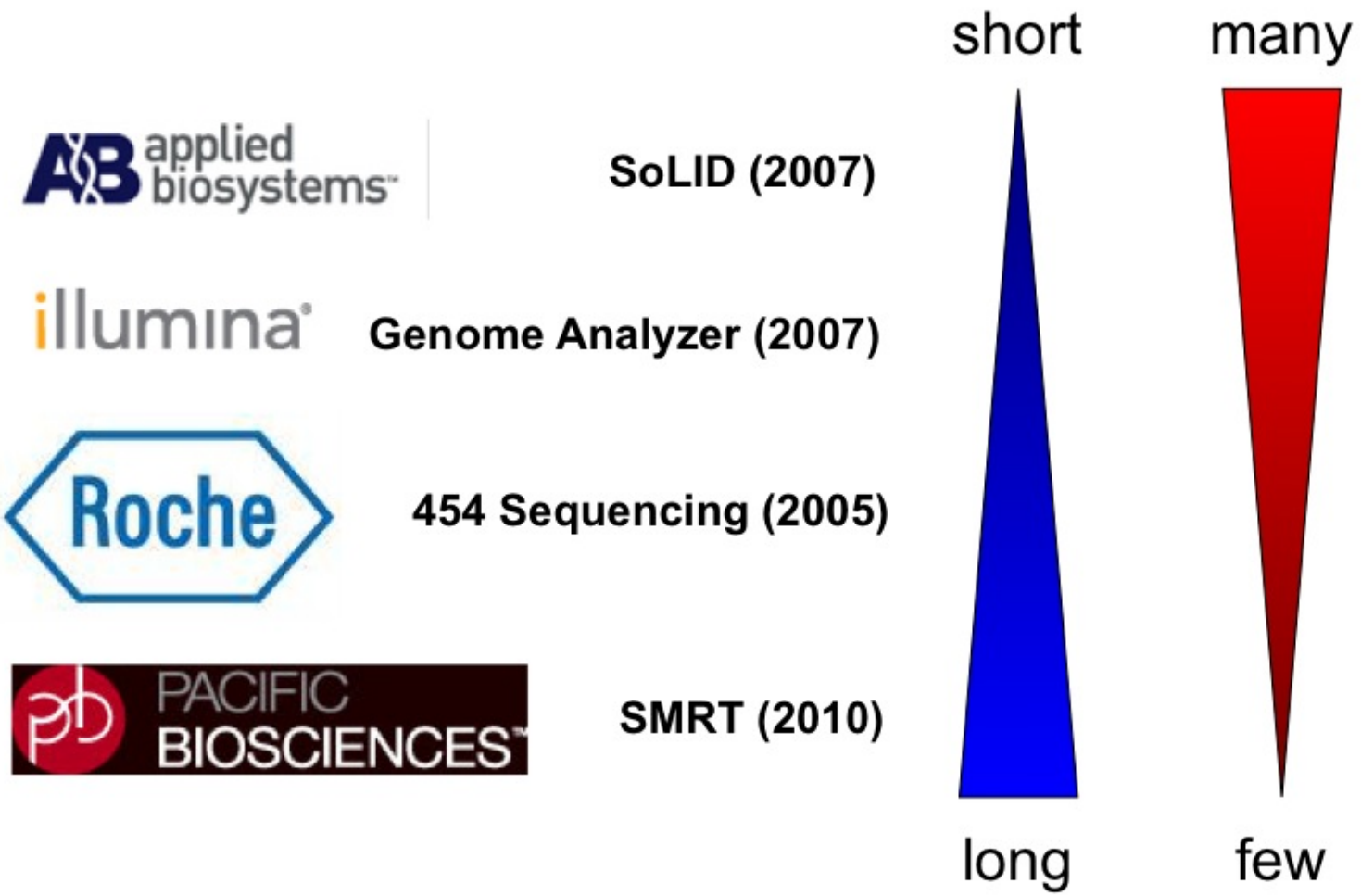
Sequencing considerations



- Amount of sequence:
  - 6-15 million reads for ChIP-seq of sequence specific factor in mammals
  - 20-40 million reads for RNA-seq quantitation, more for transcript assembly
- Type of sequence:
  - 1x36bp for ChIP-seq, RNA-seq quantitation
  - 2x75bp or more for RNA-seq transcript assembly
- For RNA-seq, longer reads translate into more splice events observed

A variety of current technologies are available with different tradeoffs





RNA Transcripts and Library Preparation



There are many different kinds of RNAs:

- Protein-coding mRNAs
- Noncoding RNAs
  - Structural RNAs (e.g. rRNAs, tRNAs, ...)
  - Small RNAs (e.g. miRNAs, endogenous siRNAs, ...)
  - Antisense / promoter-associated transcripts
  - ...

Analysis of biological sample starts with sample/library preparation.

Depending on which RNAs should be targeted, different preparation strategies have to be used.

Sample/Library Preparation Choices



Directly sequencing total-RNA is suboptimal in most cases:

- rRNA, tRNAs constitute the largest fraction of RNA ($> 90\%$)

Sample preparation choices:

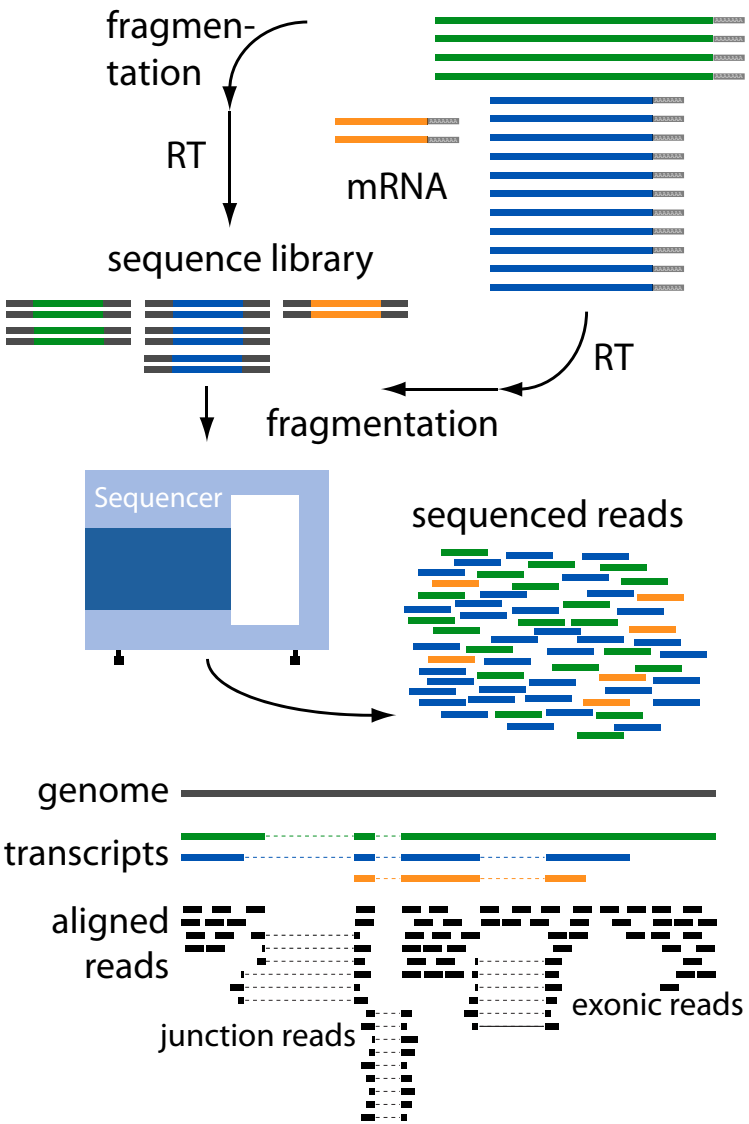
- ribo-minus (rRNA depletion, if it works)
- oligo-dT (selection of poly-adenylated transcripts),
- exonuclease treatment (degrade 5'-P RNAs)

Library preparation options (depend on sequencing technology)

- Strand information
- paired end, mate-pair sequencing

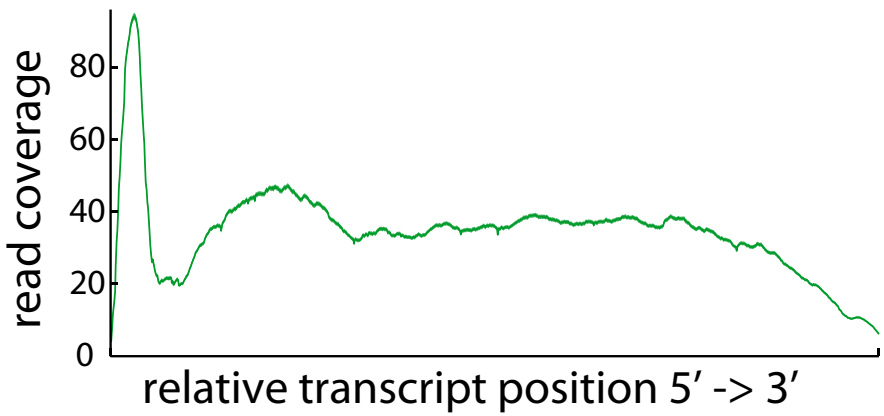
Most of these steps distort RNA transcript concentrations.

RNA-Seq Biases



Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping



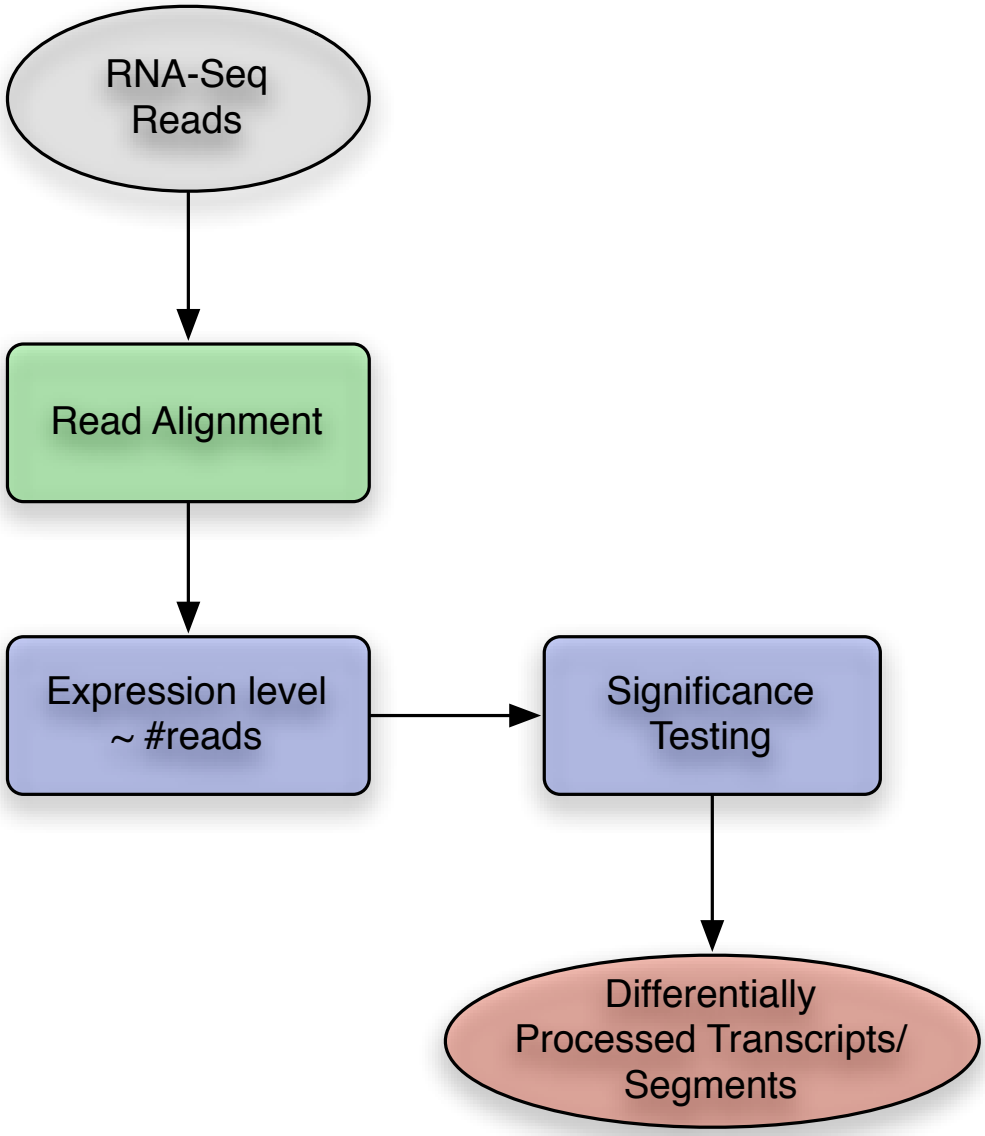
(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

RNA-Seq Biases – Publications

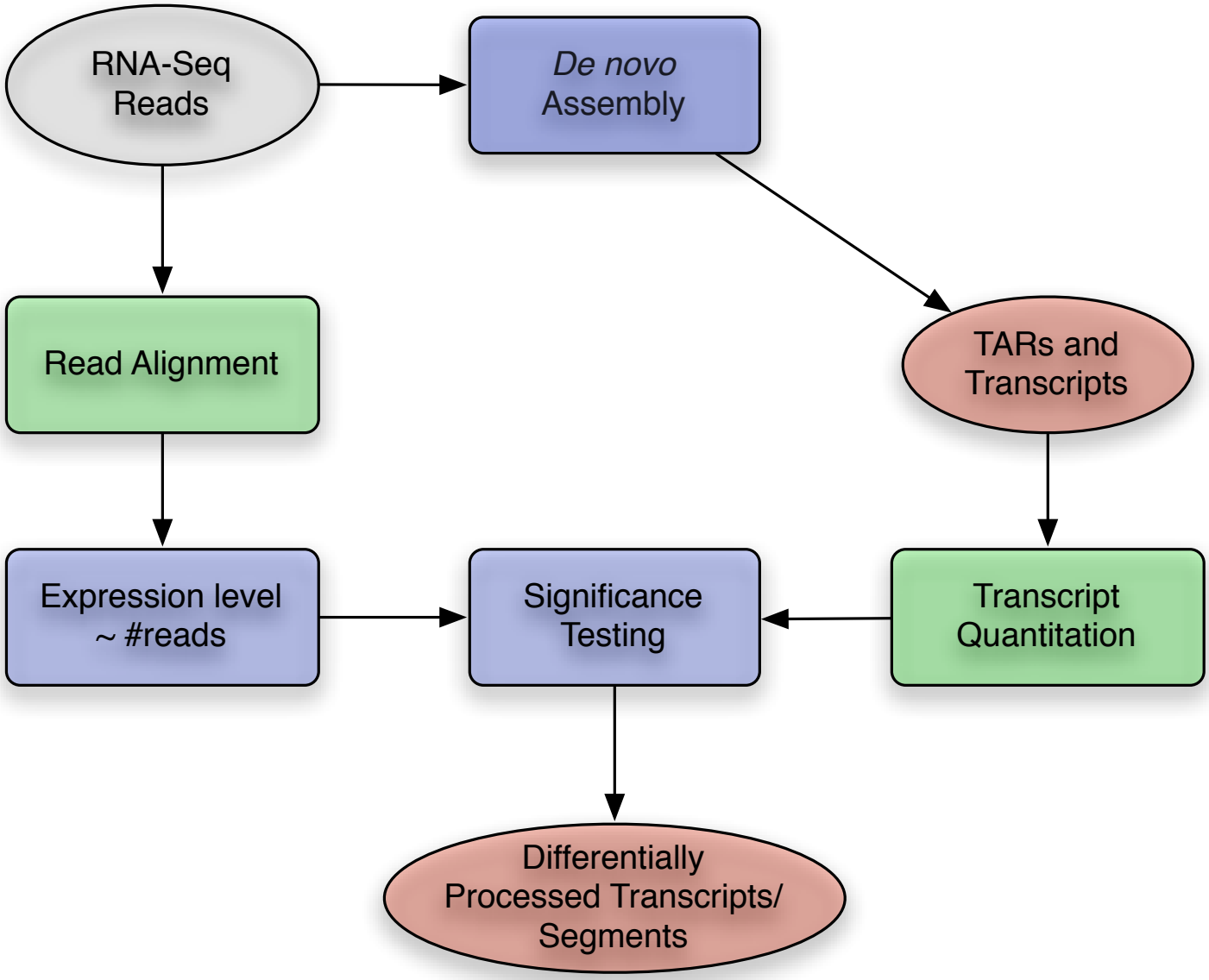


- Transcript length bias [Oshlack and Wakefield, 2009]
- Bias in small RNA sequencing [Linsen et al., 2009]
- Sequence bias in Illumina sequencer [Dohm et al., 2008]
- Biases in nucleotide composition induced by random hexamer priming [Hansen et al., 2010]
- Modeling non-uniformity [Li et al., 2010,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

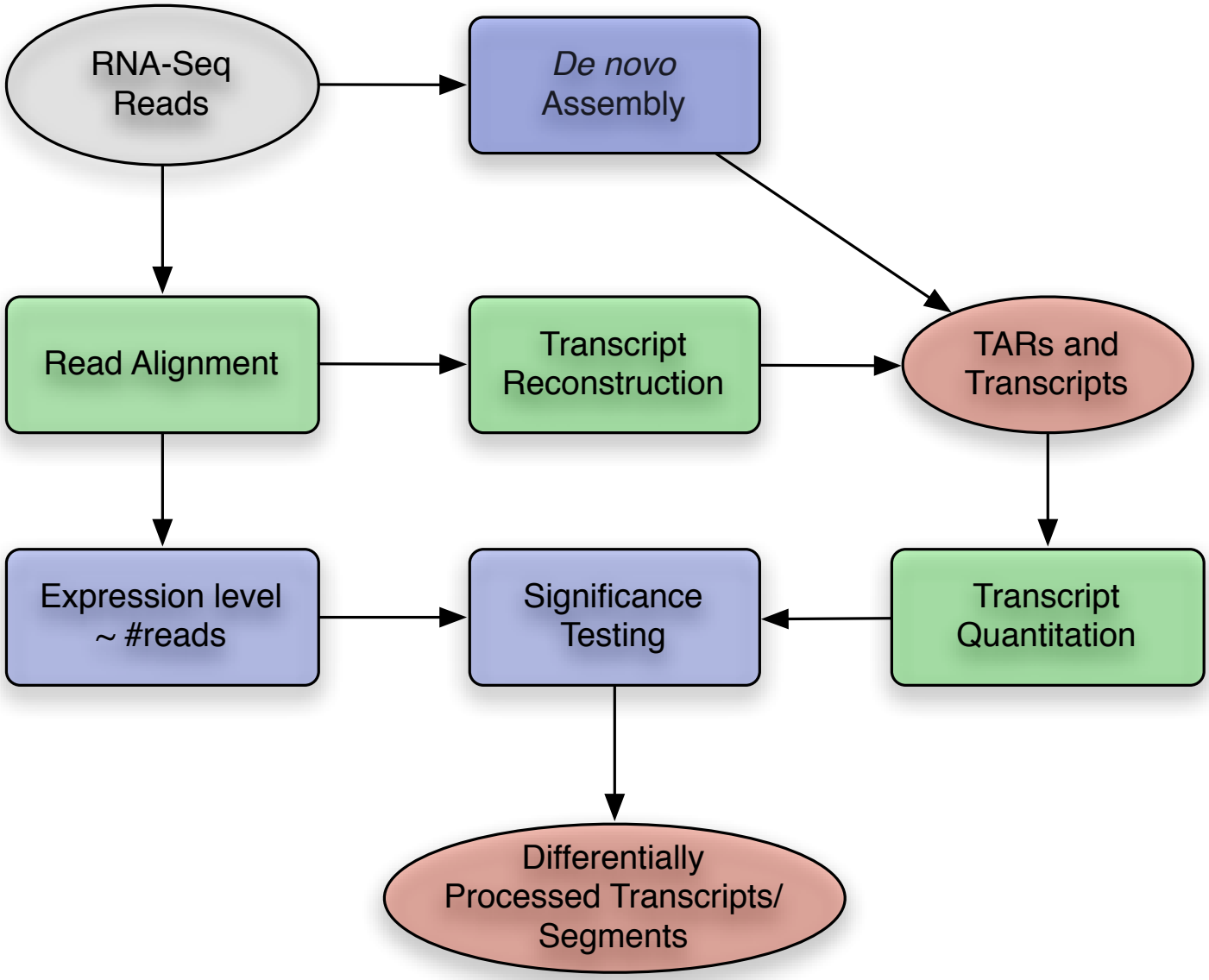
Common RNA-Seq Analysis Steps



Common RNA-Seq Analysis Steps



Common RNA-Seq Analysis Steps



RNA-seq Reads Analysis - Mapping

Reads mapping problem

For each read find its target regions on the reference genome such that are at most k mismatches between read and target.

- Global/local alignment of all reads prohibitive
- A read stems from a certain small region
- Find this region and then do an alignment
  - (Spaced) seeds
  - Suffix trees/arrays
  - Burrows-Wheeler
- Common tools: bowtie [Langmead et al., 2009], bwa [Li and Durbin, 2009, 2010], *GenomeMapper* [Schneeberger et al., 2009a], *Shrimp* [Rumble et al., 2009], *SOAP(2)* [Li et al., 2009], *VMATCH*, MAQ [Li et al., 2008], ELAND, segemehl [Hoffmann et al., 2009], . . . (≈50 more)
- Main issues:
  - Accuracy
  - Speed
  - Memory Consumption

RNA-seq Reads Analysis - Mapping



Reads mapping problem

For each read find its target regions on the reference genome such that are at most k mismatches between read and target.

- Global/local alignment of all reads prohibitive
- A read stems from a certain small region
- Find this region and then do an alignment
  - (Spaced) seeds
  - Suffix trees/arrays
  - Burrows-Wheeler
- Common tools: bowtie [Langmead et al., 2009], bwa [Li and Durbin, 2009, 2010], *GenomeMapper* [Schneeberger et al., 2009a], *Shrimp* [Rumble et al., 2009], *SOAP(2)* [Li et al., 2009], *VMATCH*, MAQ [Li et al., 2008], ELAND, segemehl [Hoffmann et al., 2009], . . . (≈50 more)
- Main issues:
  - Accuracy
  - Speed
  - Memory Consumption

RNA-seq Reads Analysis - Mapping



Reads mapping problem

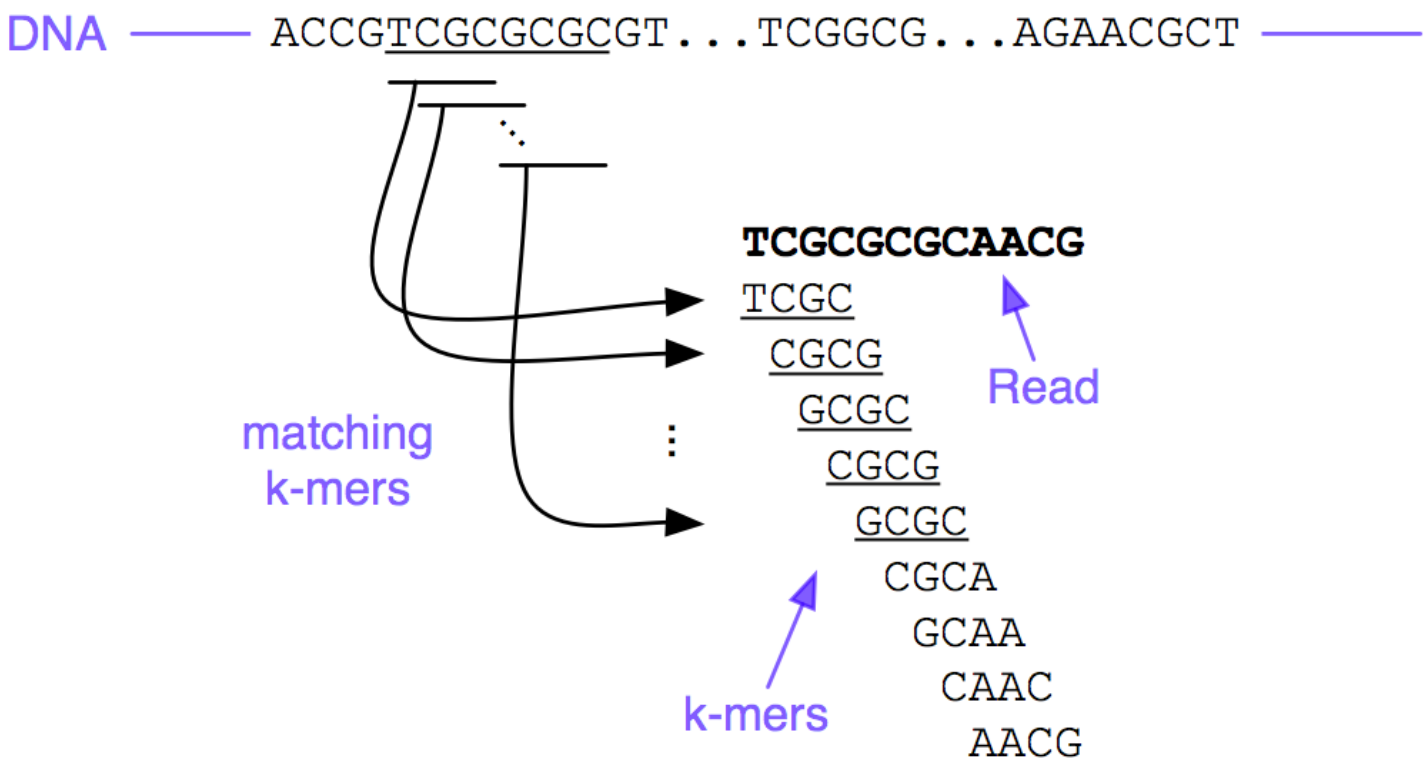
For each read find its target regions on the reference genome such that are at most k mismatches between read and target.

- Global/local alignment of all reads prohibitive
- A read stems from a certain small region
- Find this region and then do an alignment
  - (Spaced) seeds
  - Suffix trees/arrays
  - Burrows-Wheeler
- Common tools: bowtie [Langmead et al., 2009], bwa [Li and Durbin, 2009, 2010], *GenomeMapper* [Schneeberger et al., 2009a], *Shrimp* [Rumble et al., 2009], *SOAP(2)* [Li et al., 2009], *VMATCH*, MAQ [Li et al., 2008], ELAND, segemehl [Hoffmann et al., 2009], . . . (≈50 more)
- Main issues:
  - Accuracy
  - Speed
  - Memory Consumption

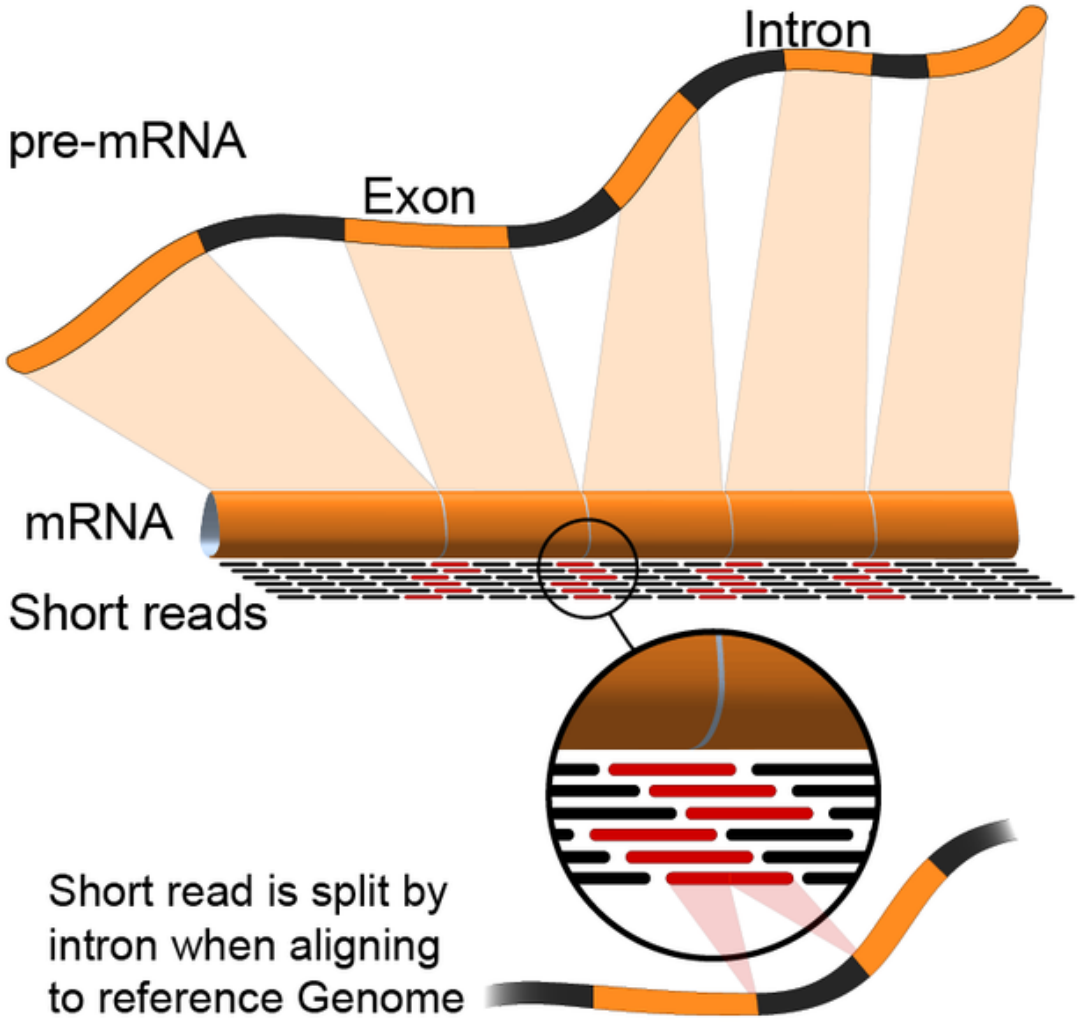
Read Alignment with GenomeMapper

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

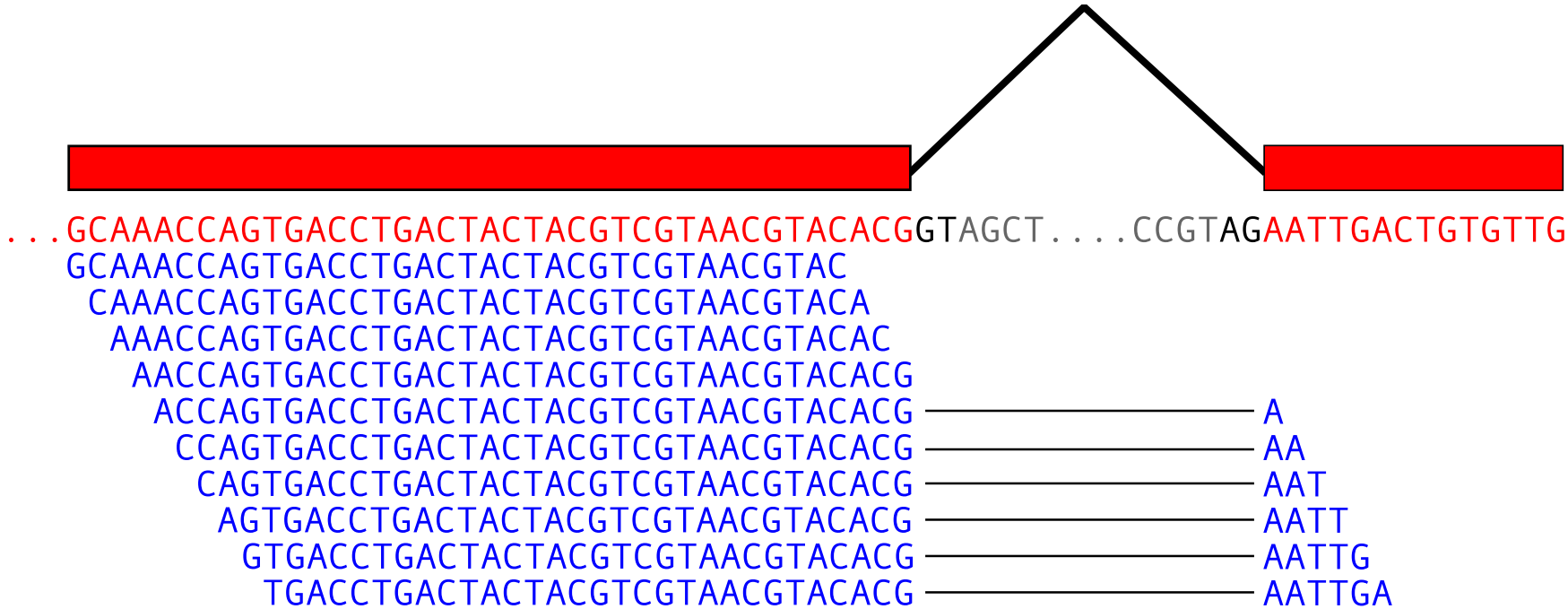


Spliced vs. Unspliced Alignments



[Wikipedia]

Spliced vs. Unspliced Alignments



- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

Read Alignment with PALMapper



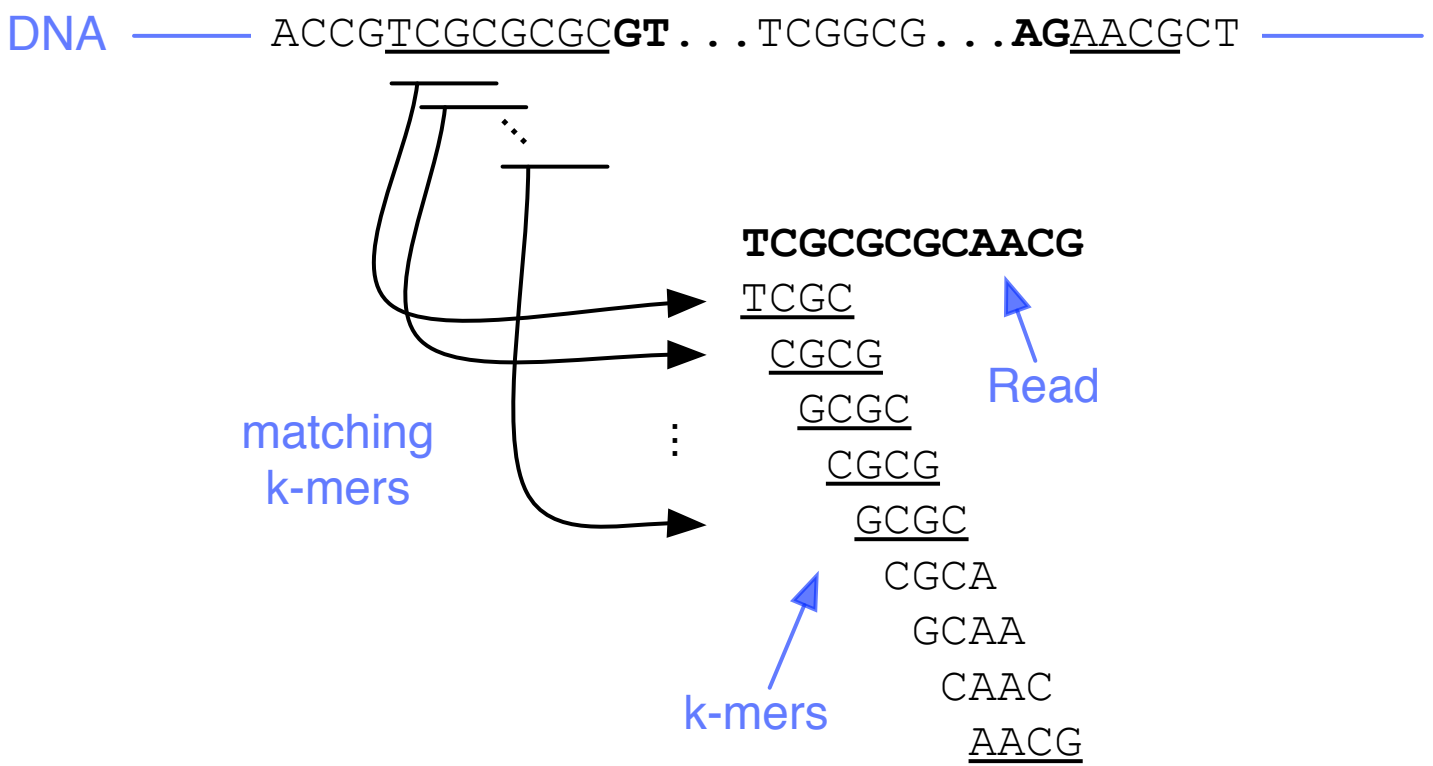
GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

Read Alignment with PALMapper

GenomeMapper for (unspliced) read mapping:

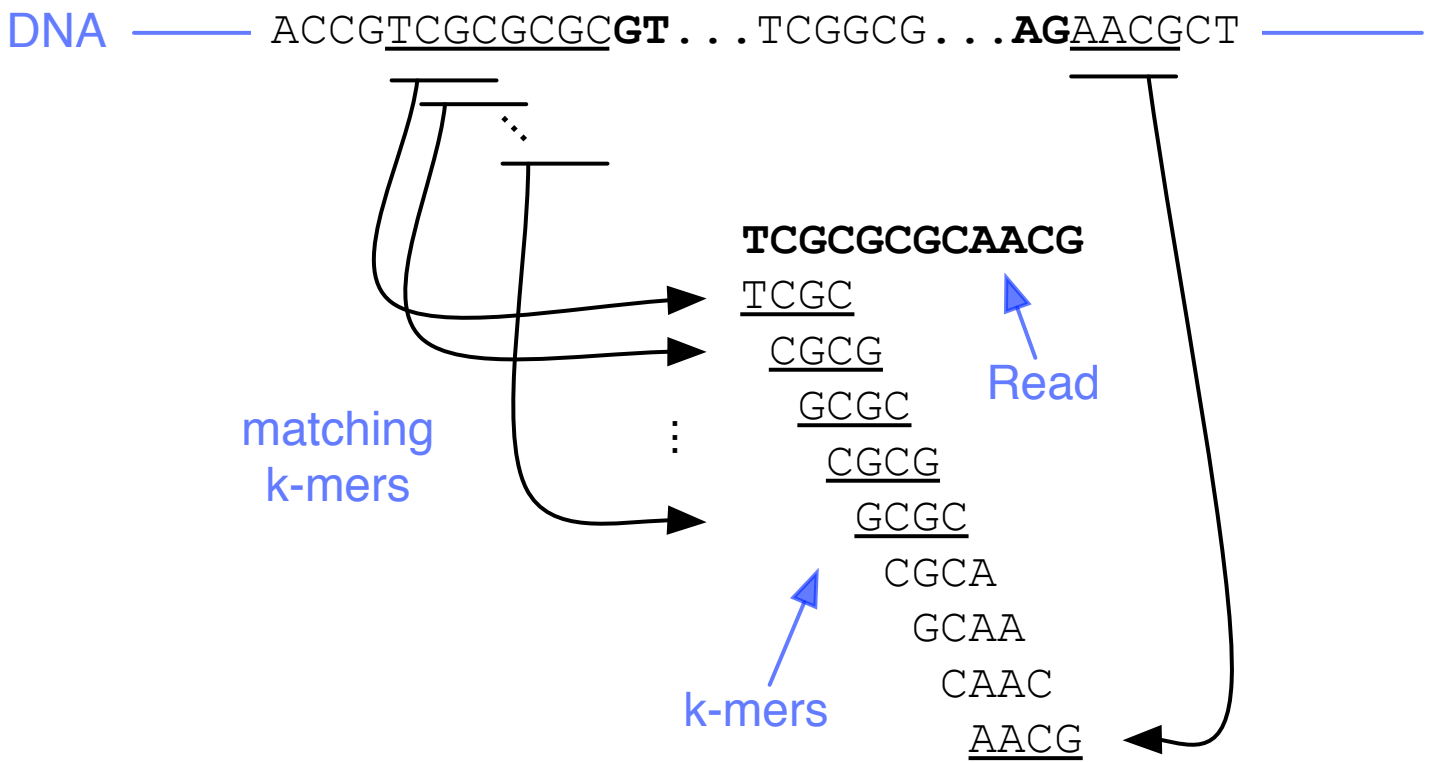
- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes



Read Alignment with PALMapper

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes



Read Alignment with PALMapper



GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:

- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA [De Bona et al., 2008]

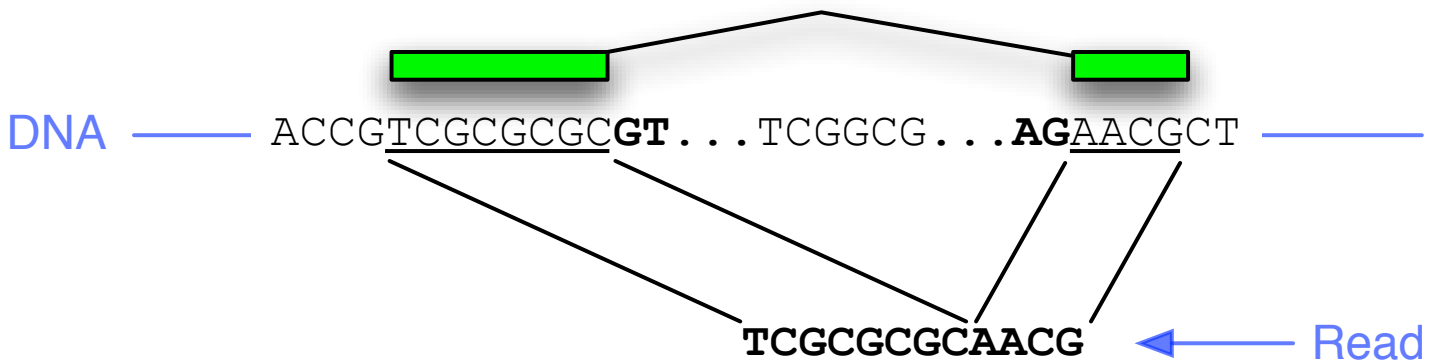
Read Alignment with PALMapper

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009b]
- *k*-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:

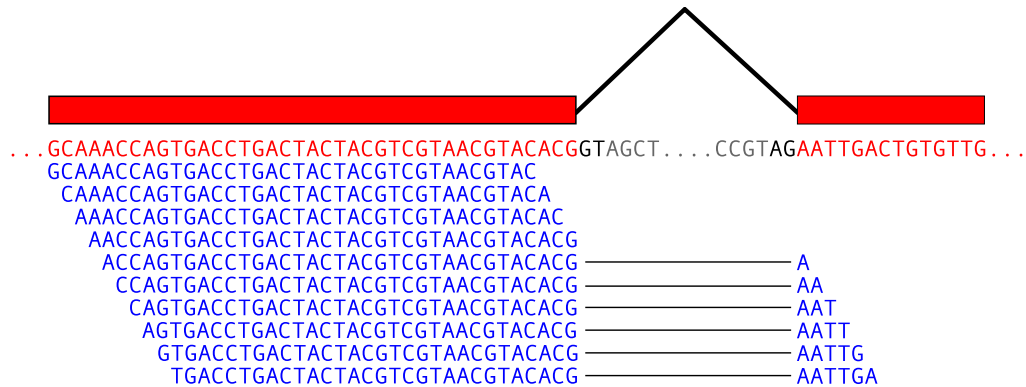
- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA [De Bona et al., 2008]



RNA-seq Read Analysis - Spliced Alignment



- Genomic read mapping already challenging
- Transcriptome read mapping is even more difficult:



Improve alignments by using more information:

- Accurate splice site models
- Intron length model
- Quality scores model

Idea: Use a **machine learning** method to infer optimal parameters

QPALMA: Extended Smith-Waterman Alg.



	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Extended Smith-Waterman Alg.



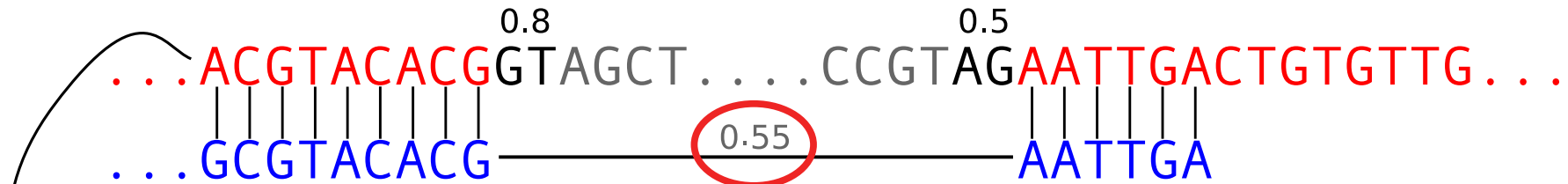
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Extended Smith-Waterman Alg.



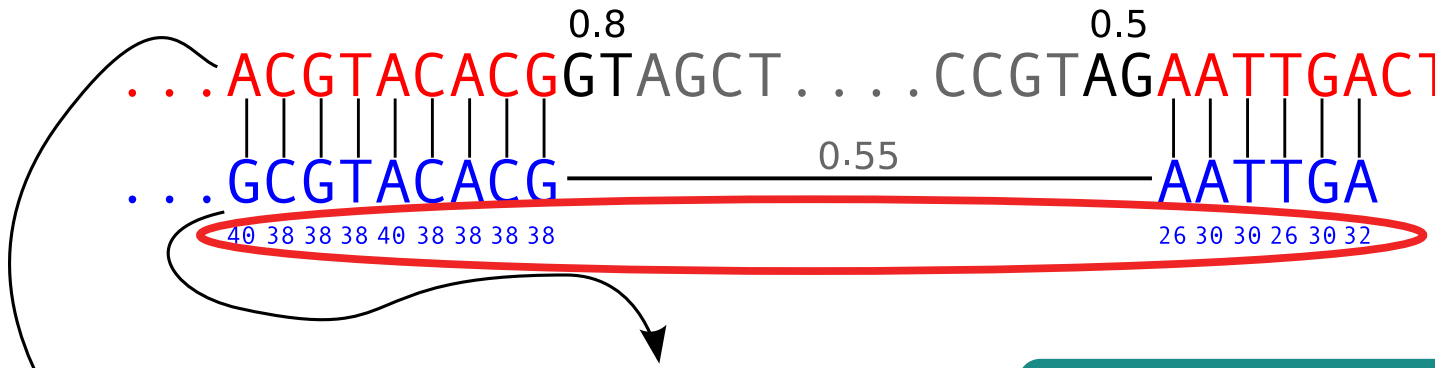
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Extended Smith-Waterman Alg.



	gap	A	C	G	T	N
gap	0.33	$f_{-A}(\cdot)$	$f_{-C}(\cdot)$	$f_{-G}(\cdot)$	$f_{-T}(\cdot)$	$f_{-N}(\cdot)$
A	0.31	$f_{AA}(\cdot)$	$f_{AC}(\cdot)$	$f_{AG}(\cdot)$	$f_{AT}(\cdot)$	$f_{AN}(\cdot)$
C	0.44	$f_{CA}(\cdot)$	$f_{CC}(\cdot)$	$f_{CG}(\cdot)$	$f_{CT}(\cdot)$	$f_{CN}(\cdot)$
G	0.13	$f_{GA}(\cdot)$	$f_{GC}(\cdot)$	$f_{GG}(\cdot)$	$f_{GT}(\cdot)$	$f_{GN}(\cdot)$
T	0.55	$f_{TA}(\cdot)$	$f_{TC}(\cdot)$	$f_{TG}(\cdot)$	$f_{TT}(\cdot)$	$f_{TN}(\cdot)$
N	0.12	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$

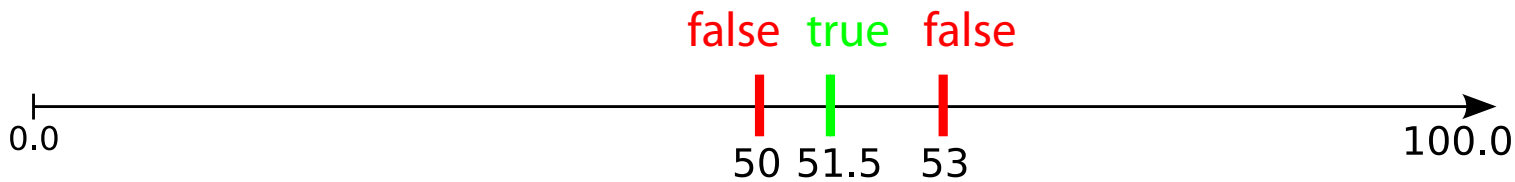
Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

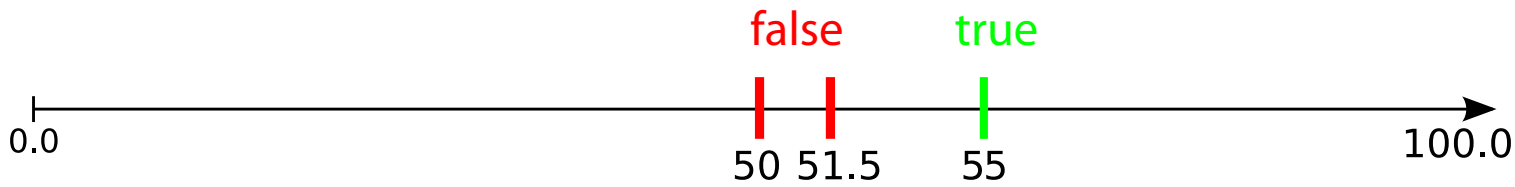
[De Bona et al., 2008]

Cartoon: Maximize the Margin



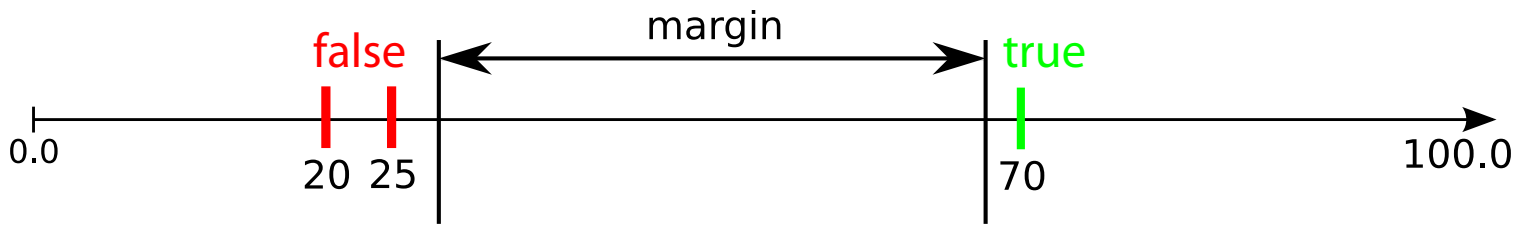
- Correct alignment is **not** highest scoring one
- Correct alignment is highest scoring one
- Can we do better?

Cartoon: Maximize the Margin



- Correct alignment is **not** highest scoring one
- Correct alignment is highest scoring one
- Can we do better?

Cartoon: Maximize the Margin

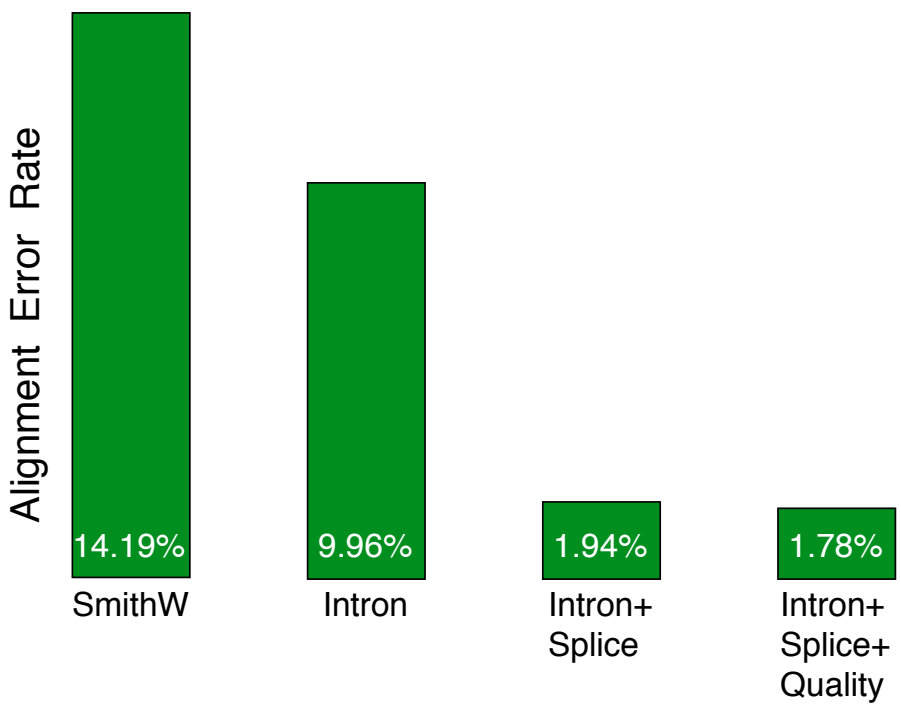


- Technique motivated by SVMs ("large-margin")
- Enforce a margin between correct and incorrect examples
- Application of Structured Output Learning

RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use $10,000$ reads for training and $30,000$ for testing

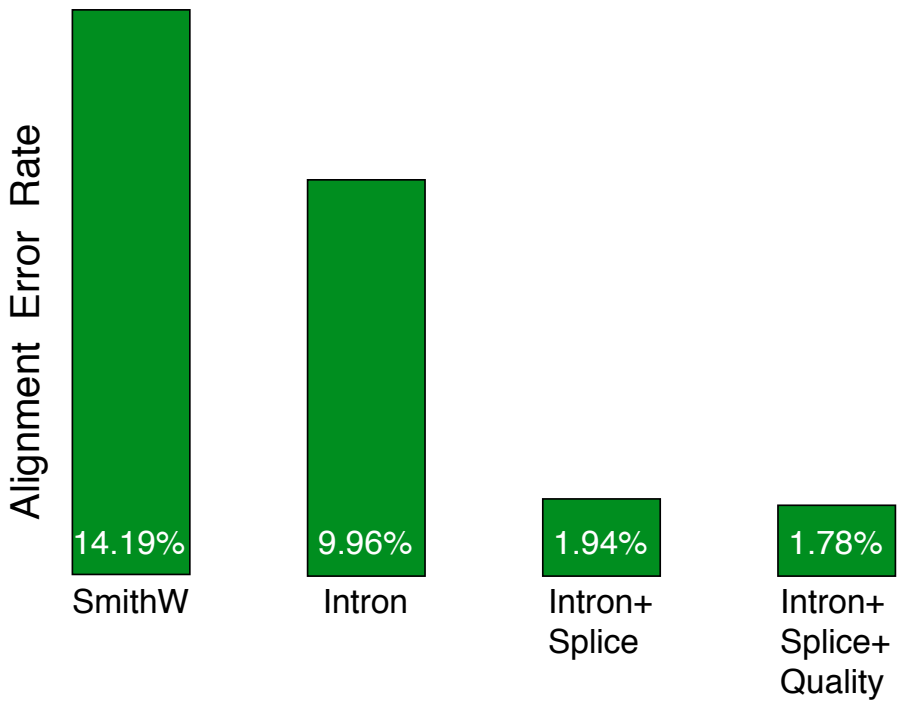


[De Bona et al., 2008]

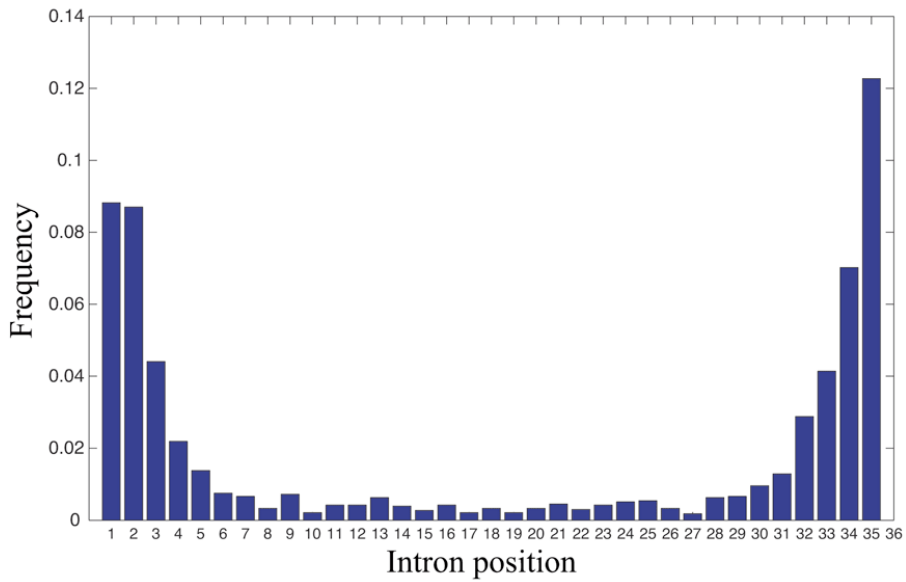
RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing



Error vs. intron position

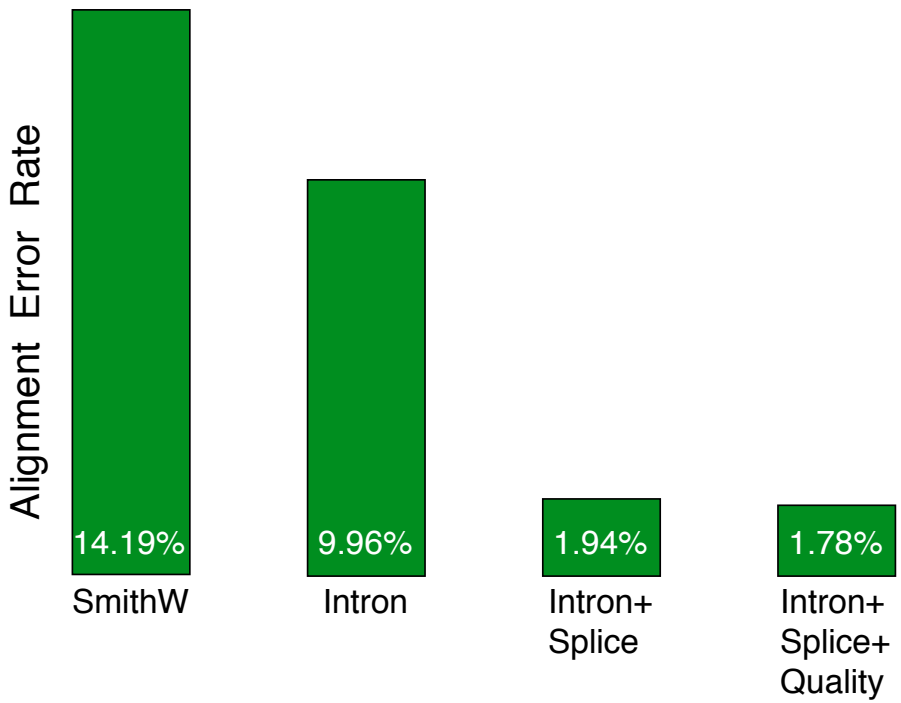


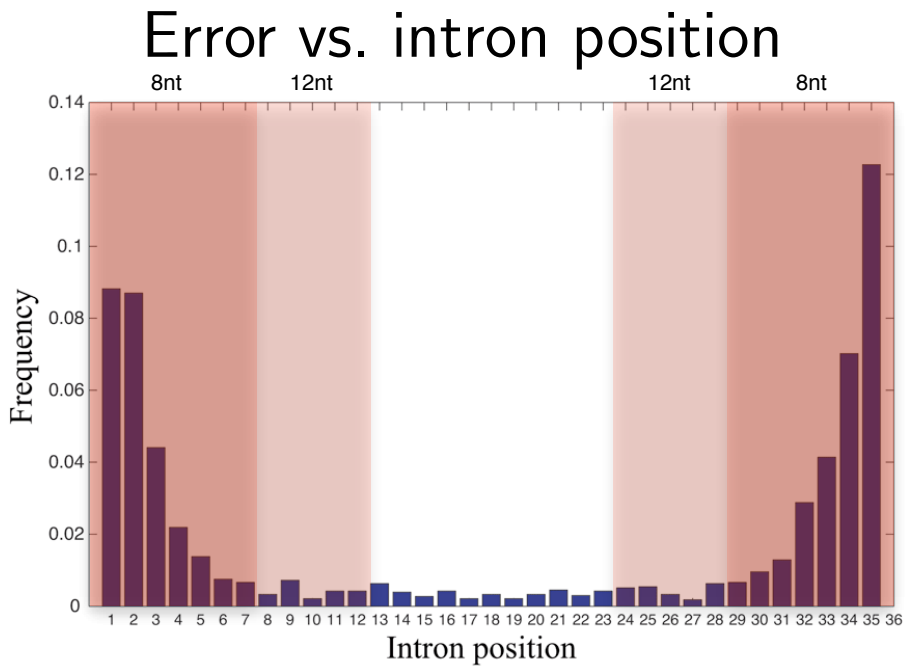
[De Bona et al., 2008]

RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing





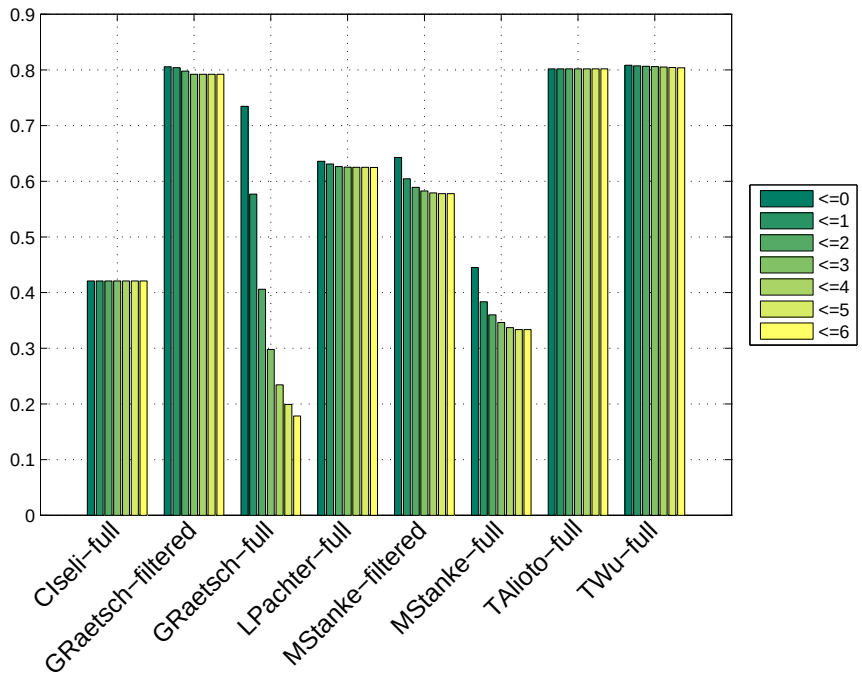
[De Bona et al., 2008]

Agreement of Intron Predictions with Annotation (*C. elegans*)

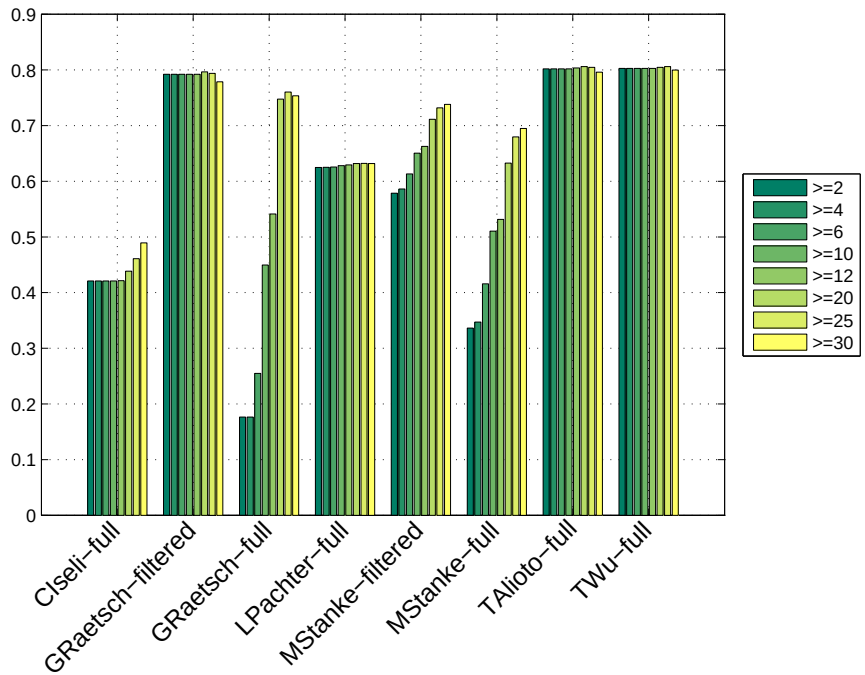
Strategy:

- Consider annotated introns as gold standard.
- Compute F-score of agreement of predicted introns in different submissions with gold standard.

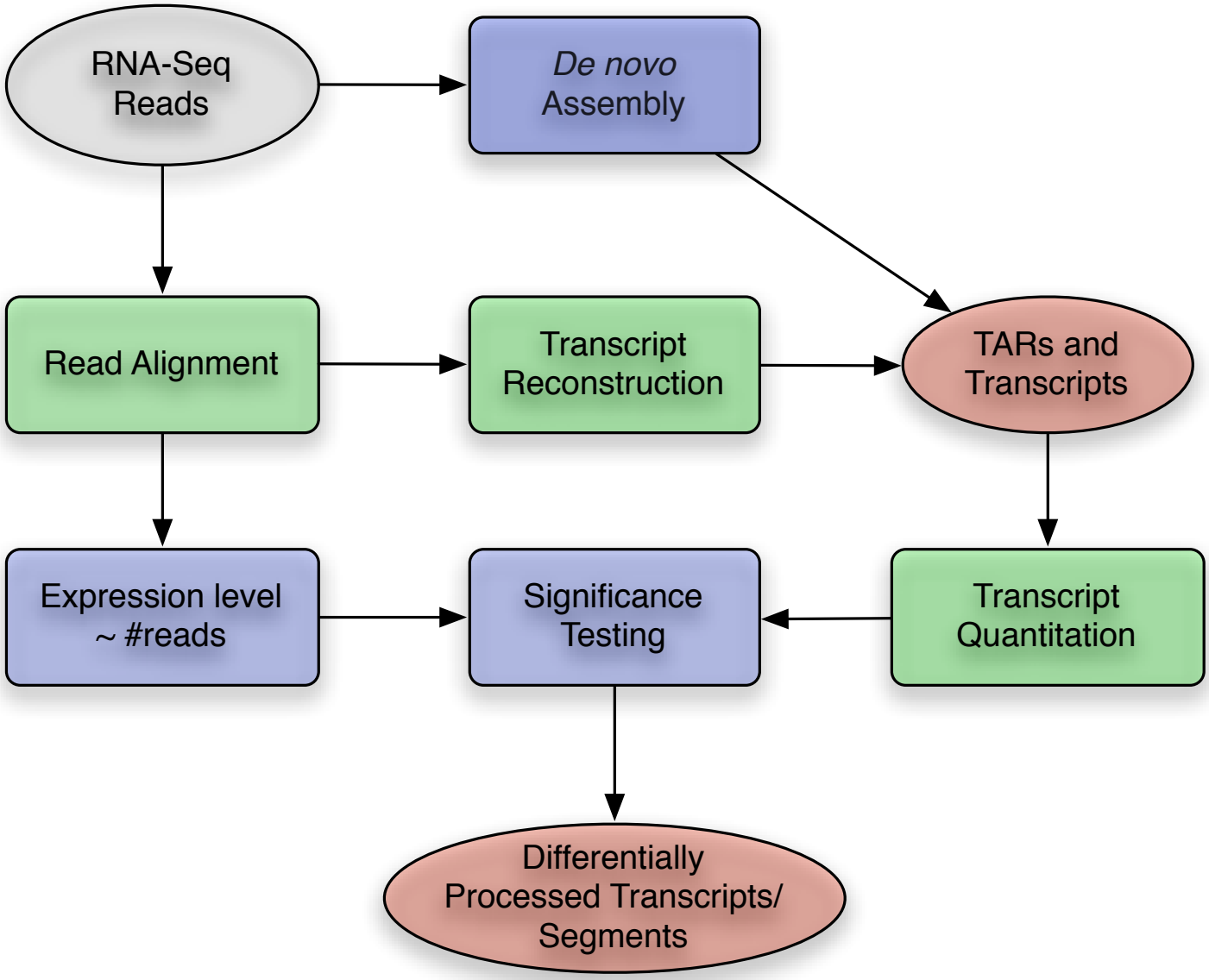
Filtered by mismatches



Filtered by segment length



Common RNA-Seq Analysis Steps



Estimate Gene Expression



Idea: Use the number of reads mapping to a gene as estimate for the gene expression.

Problem: Read number scales with total number of reads and transcript length

Approach: Normalize read count, by

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

$\Rightarrow$ **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 estimate for the gene expression.
Problem: Read number scales with total number of reads and transcript length

Approach: Normalize read count, by

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

$\Rightarrow$ **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 estimate for the gene expression.
Problem: Read number scales with total number of reads and transcript length

Approach: Normalize read count, by

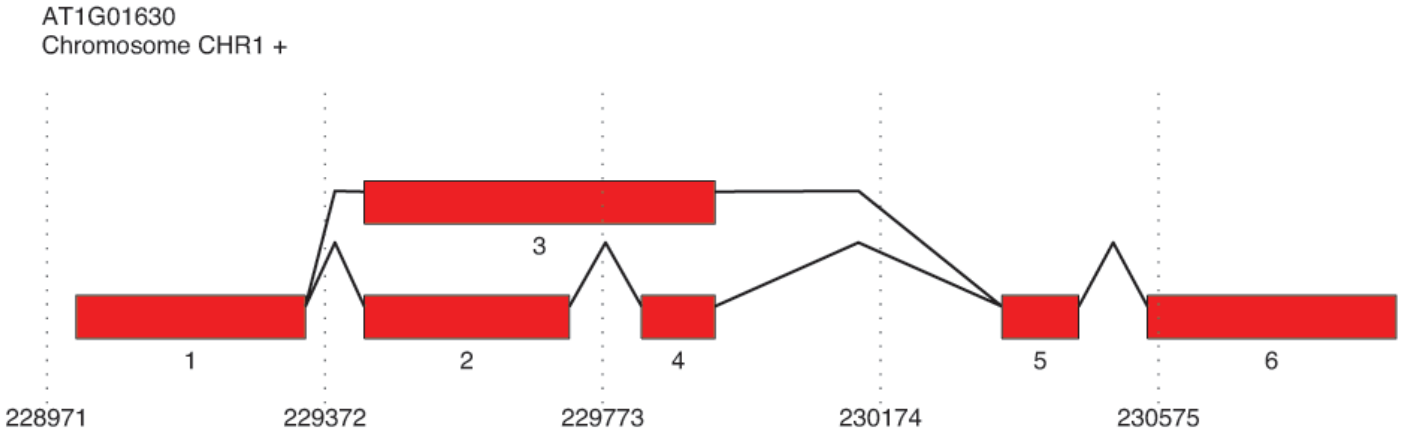
- Length of the transcript (sum of exonic regions in kilobases)
- Total number of reads (in million)

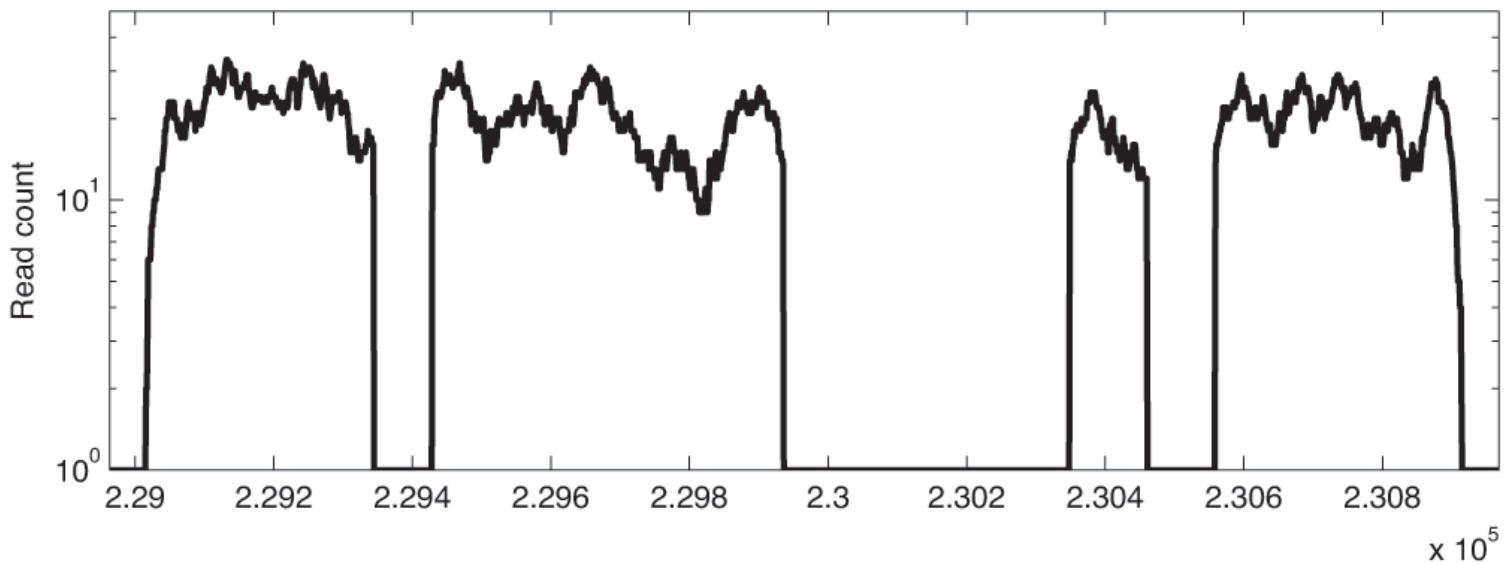
$\Rightarrow$ **Reads per kilobase per million mapped reads (RPKM)**

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

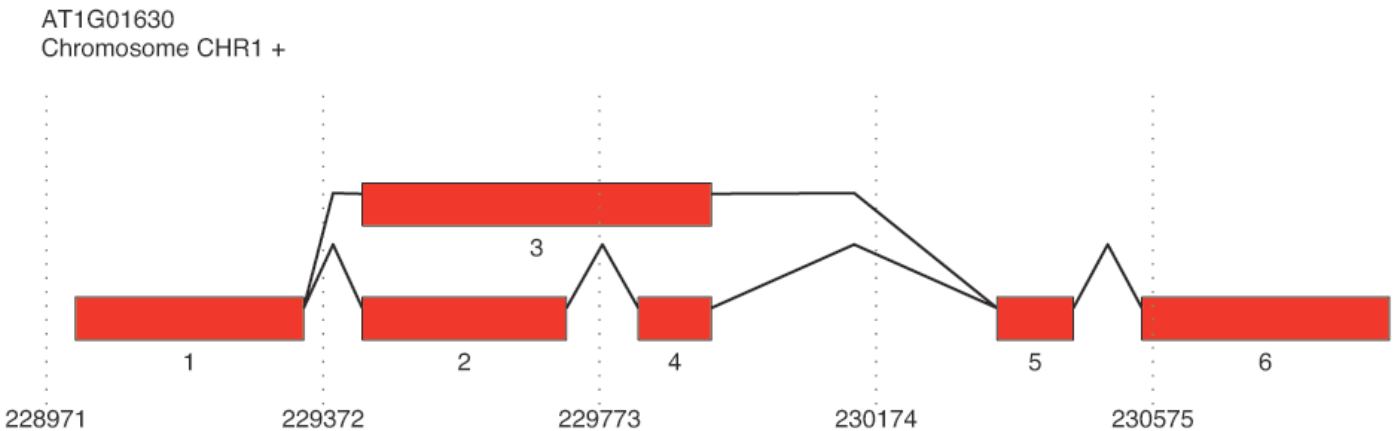
$\Rightarrow$ **Fragments per kilobase per million mapped reads (FPKM)**

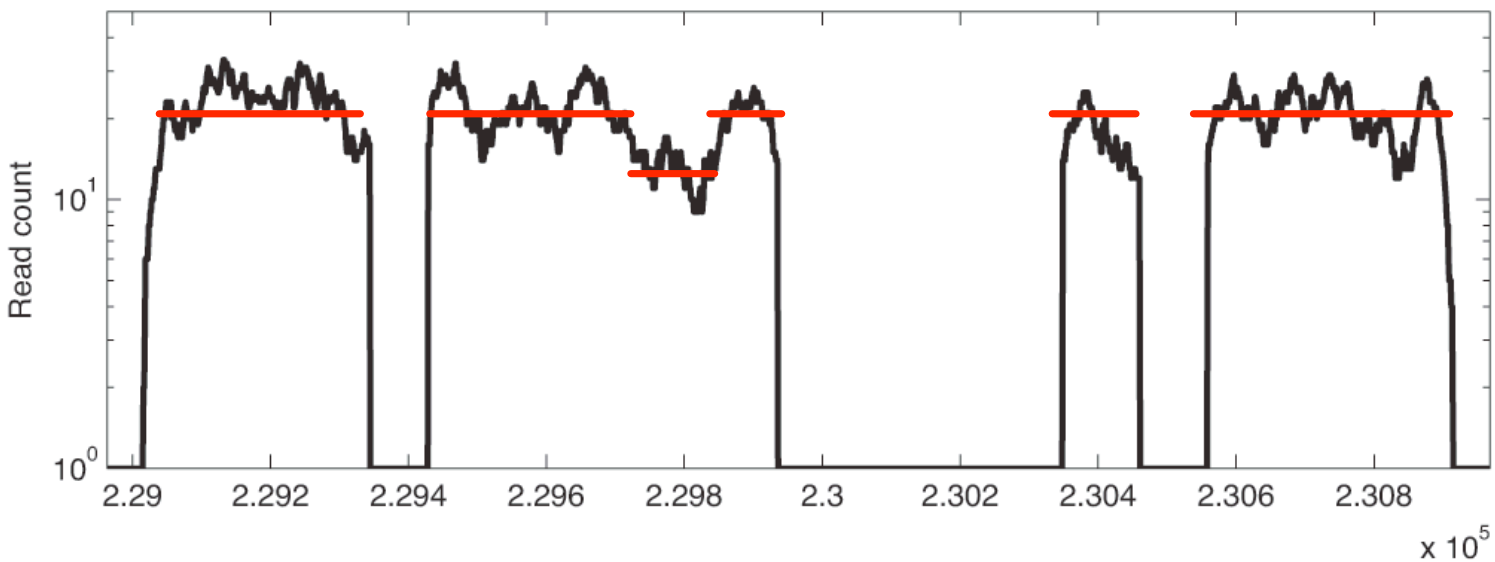
Quantitation of Transcripts





Quantitation of Transcripts





Quantitation of Transcripts

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

Solve an optimization problem:

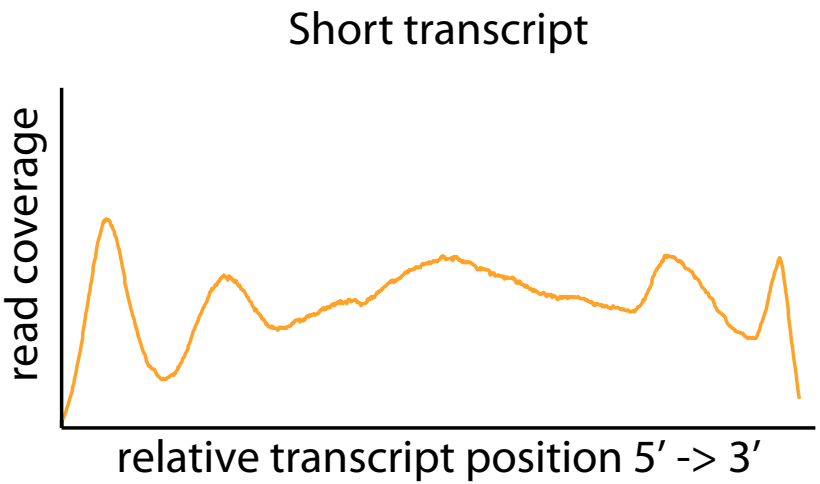
- Optimizing weights w_t for each transcript $t = 1, \ldots, T$
- Exploiting additive nature of the read coverage
- Minimizing residual error (e.g., squared error)

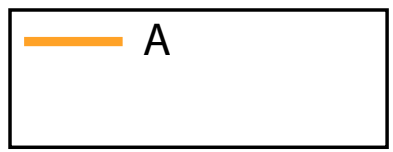
$$(w_1, \ldots, w_T) = \underset{w_1,\ldots,w_T \geq 0}{\operatorname{argmin}} \sum_{p \in P} \left(R_p - \sum_{t=1}^{T} w_t D_{t,p} \right)^2 ,$$

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

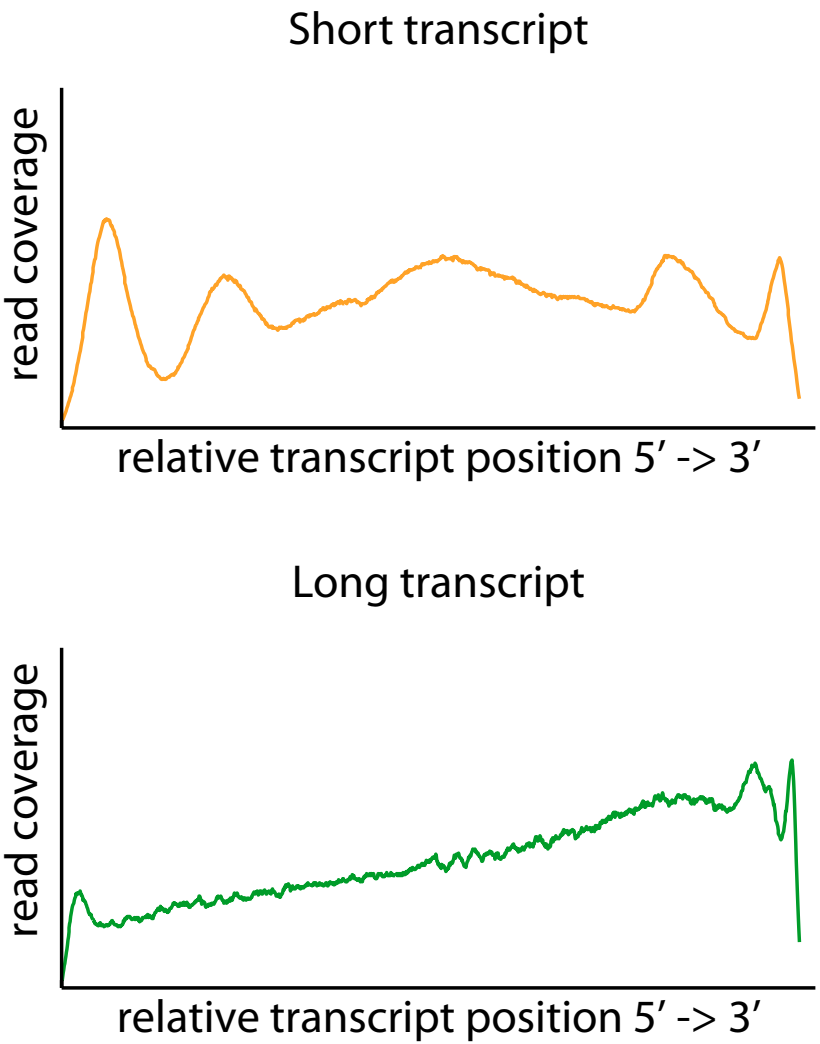
rQuant – Basic Idea





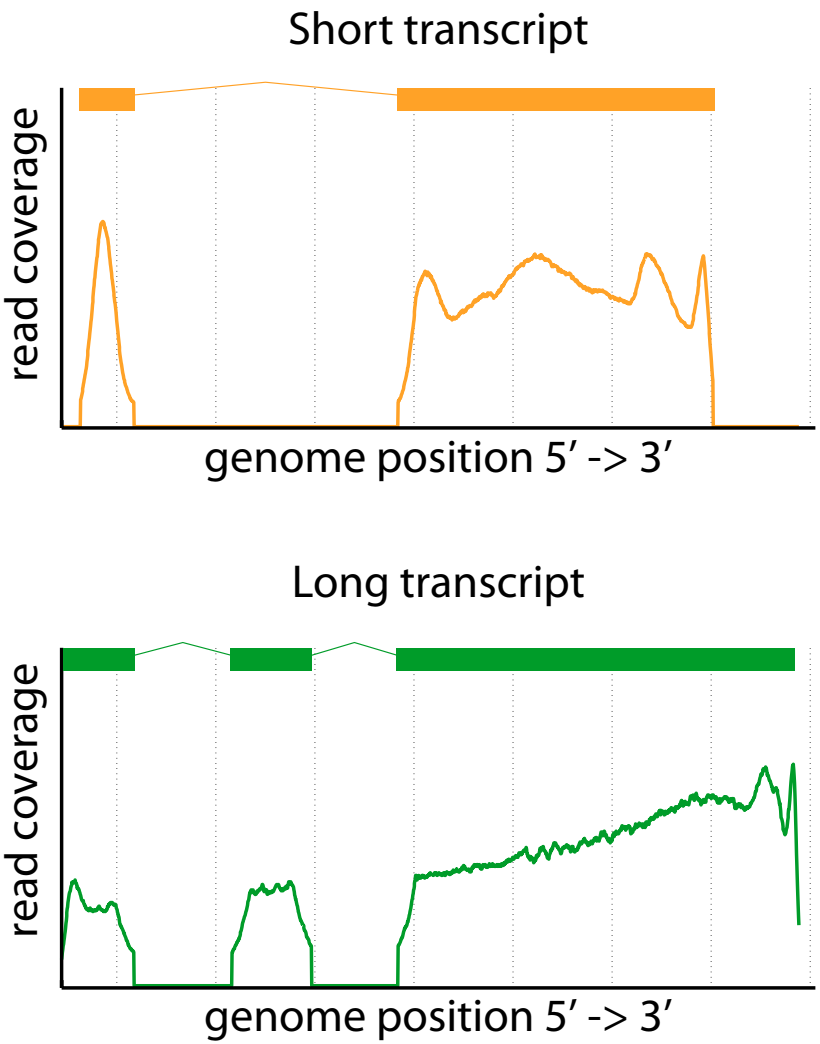
$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea



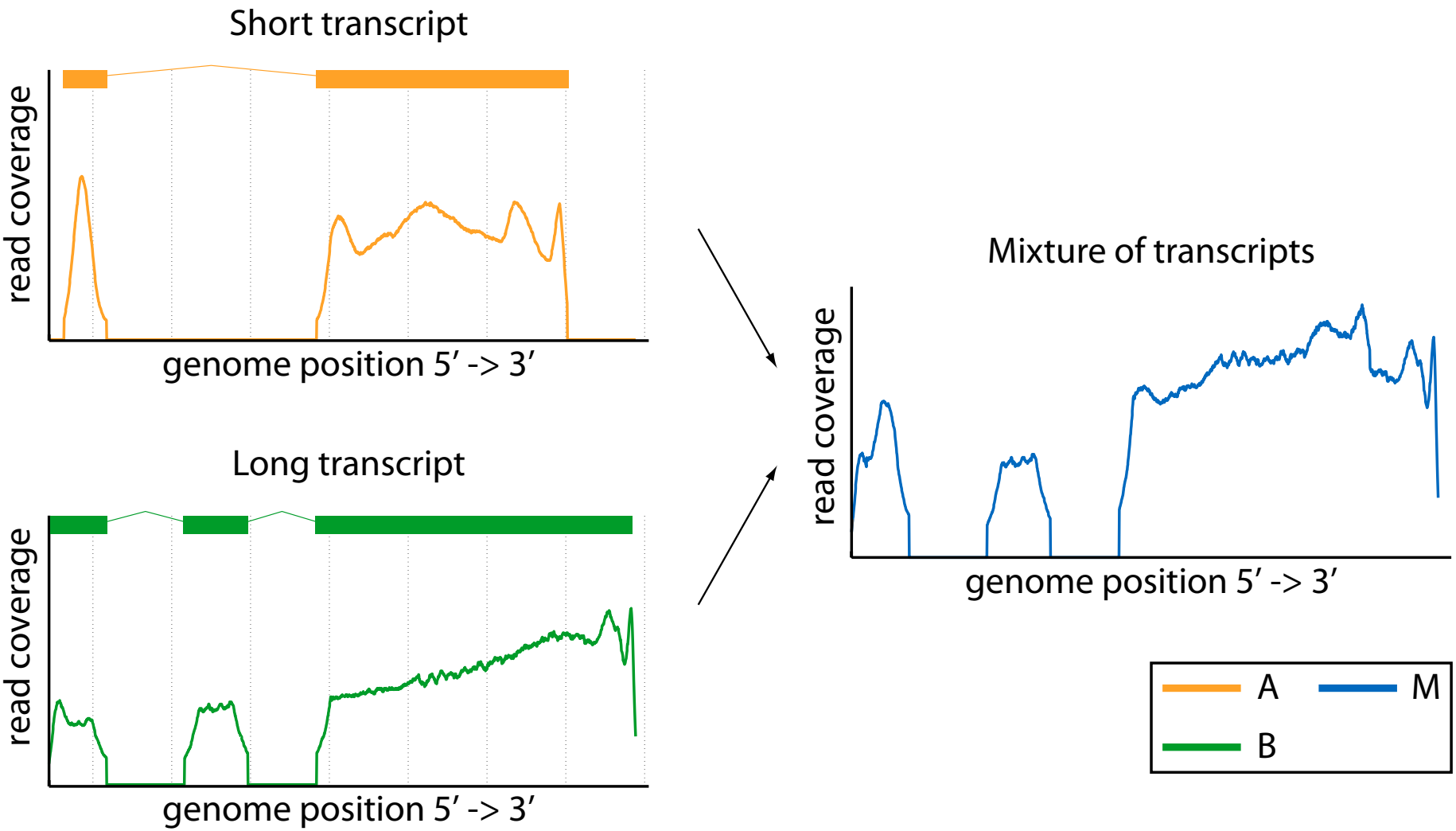
$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea



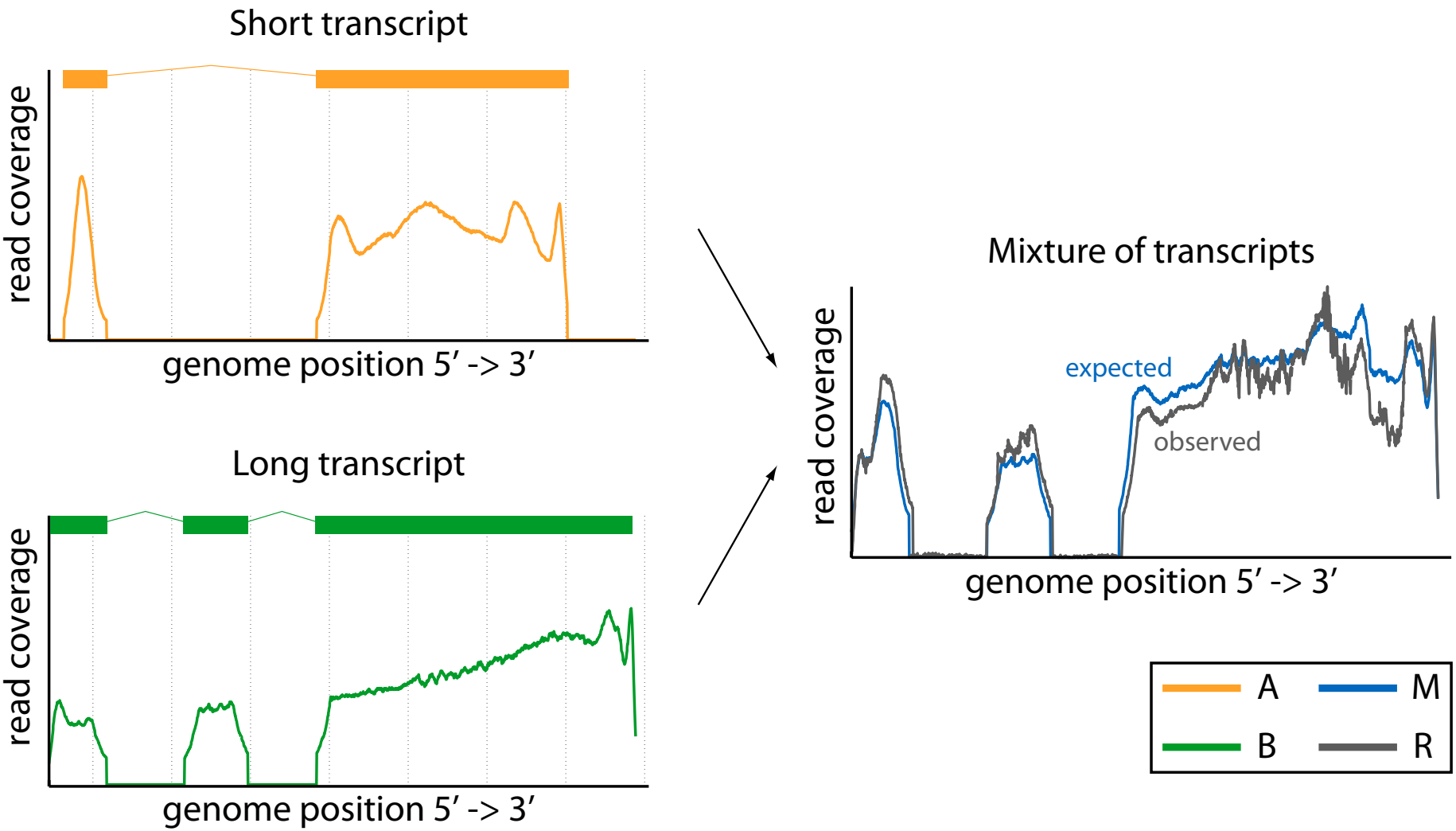
$$M_p = w_A D_{A,p} + w_B D_{B,p} \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, R_p)$$

rQuant – Basic Idea



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

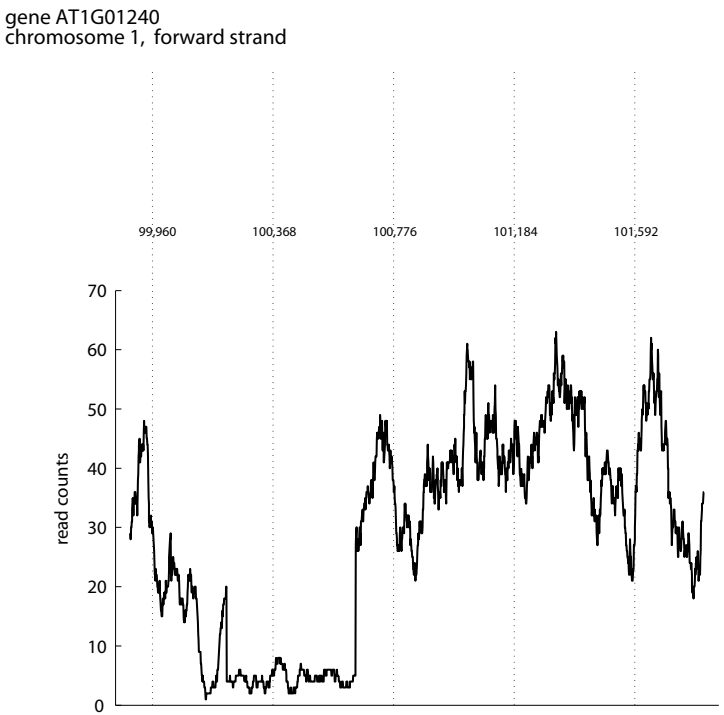
rQuant – Basic Idea



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

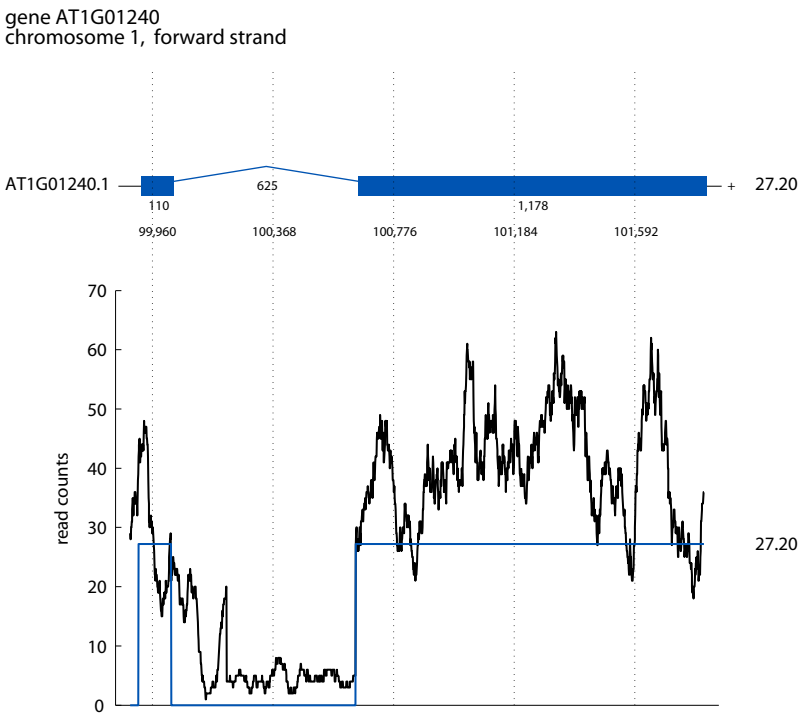
rQuant – Iterative Algorithm

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.



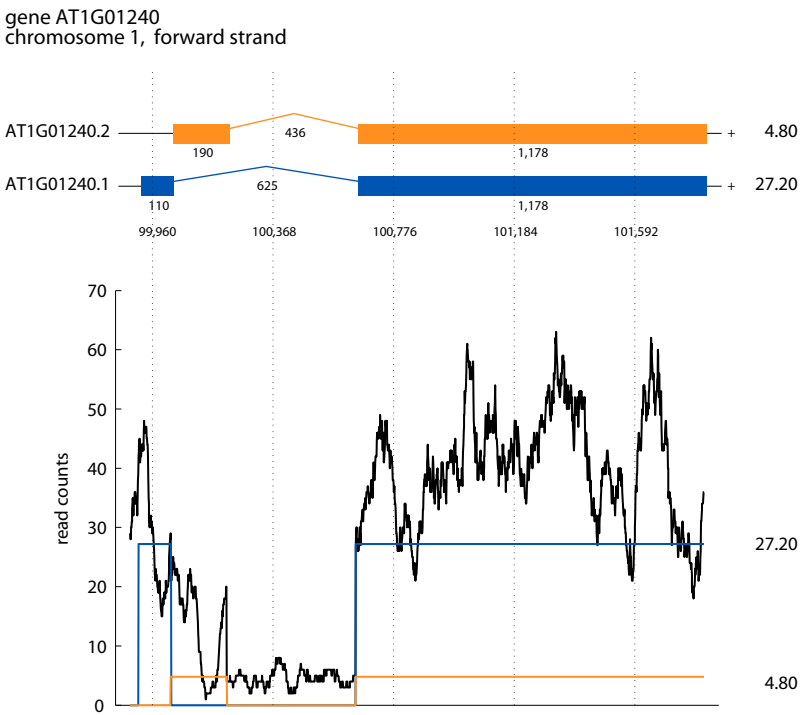
rQuant – Iterative Algorithm

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.



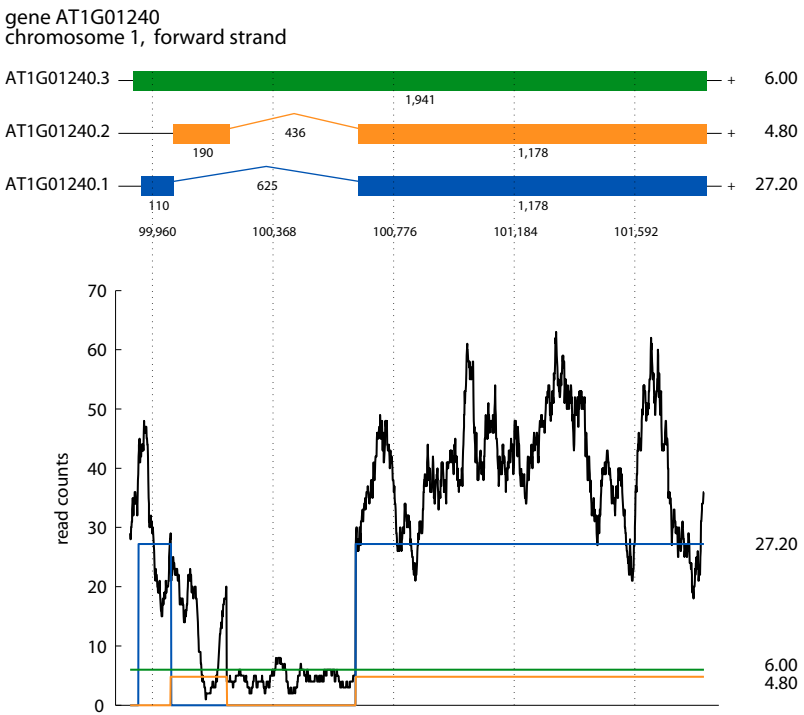
rQuant – Iterative Algorithm

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.



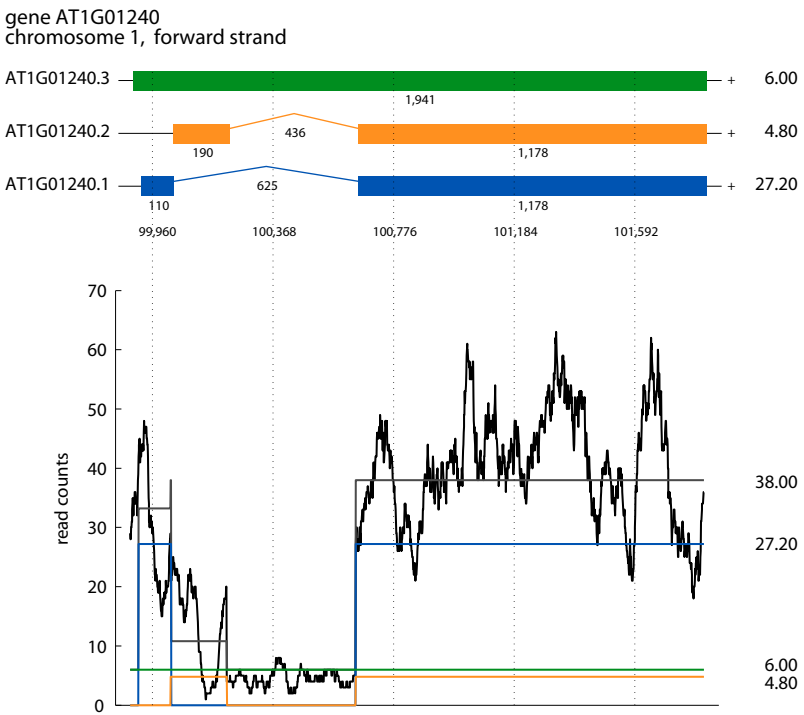
rQuant – Iterative Algorithm

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.



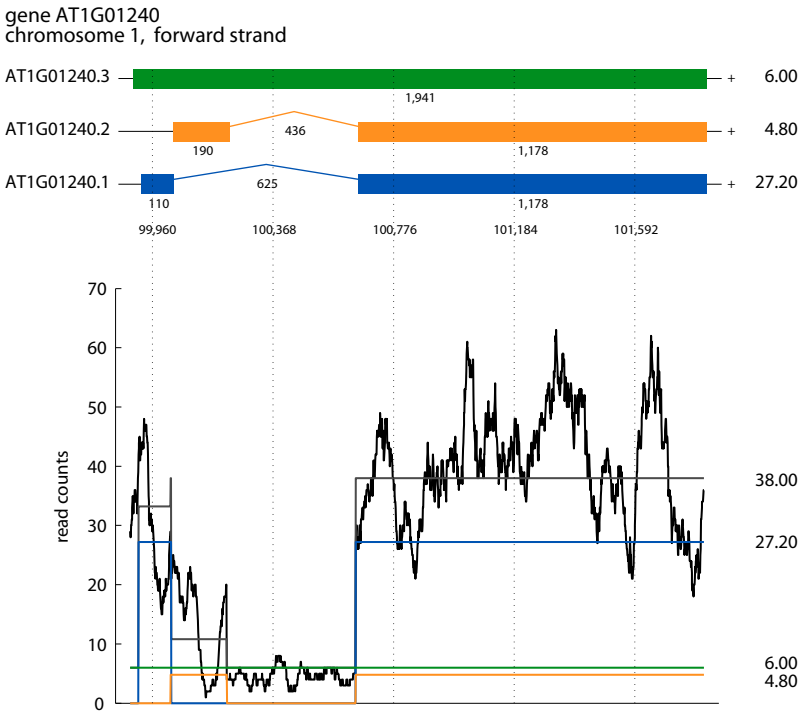
rQuant – Iterative Algorithm

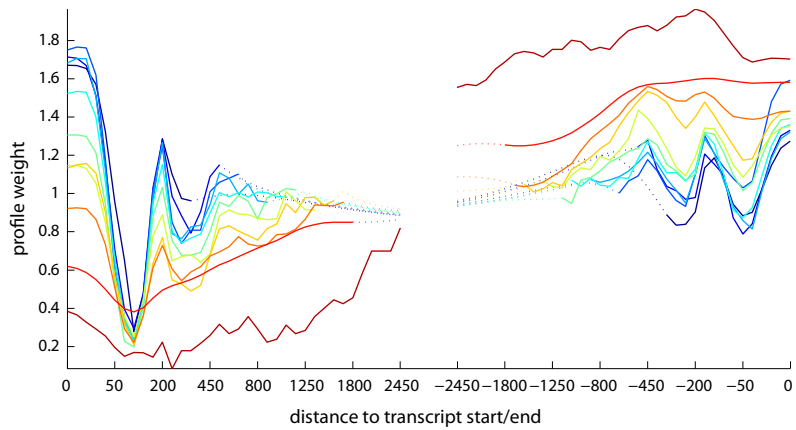
1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.



rQuant – Iterative Algorithm

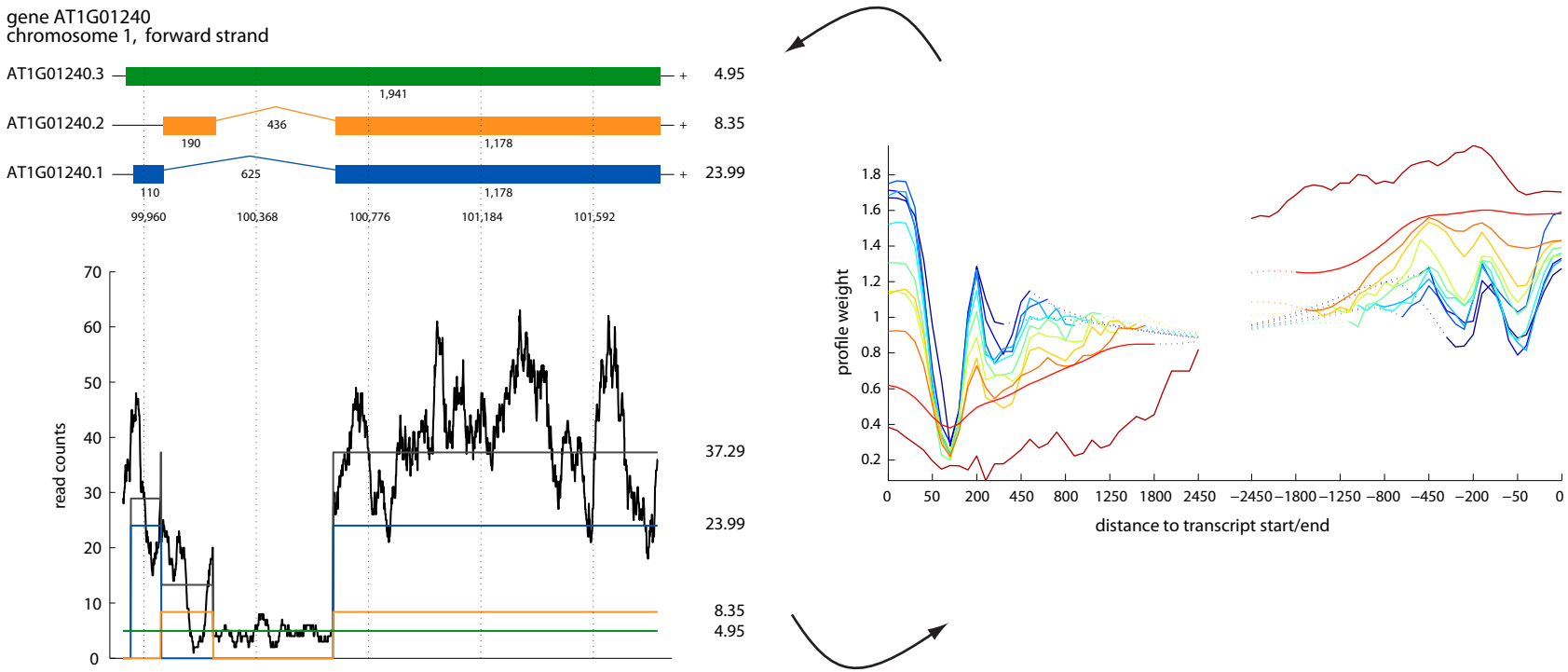
1. Optimize transcript weights: $\min_{\mathbf{w}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
2. Optimize profile weights: $\min_{\mathbf{p}} \sum_p \ell\left(\sum_t w_t D_{t,p}, R_p\right)$
3. Repeat 1. and 2. until convergence.





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 Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I



Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

Method rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

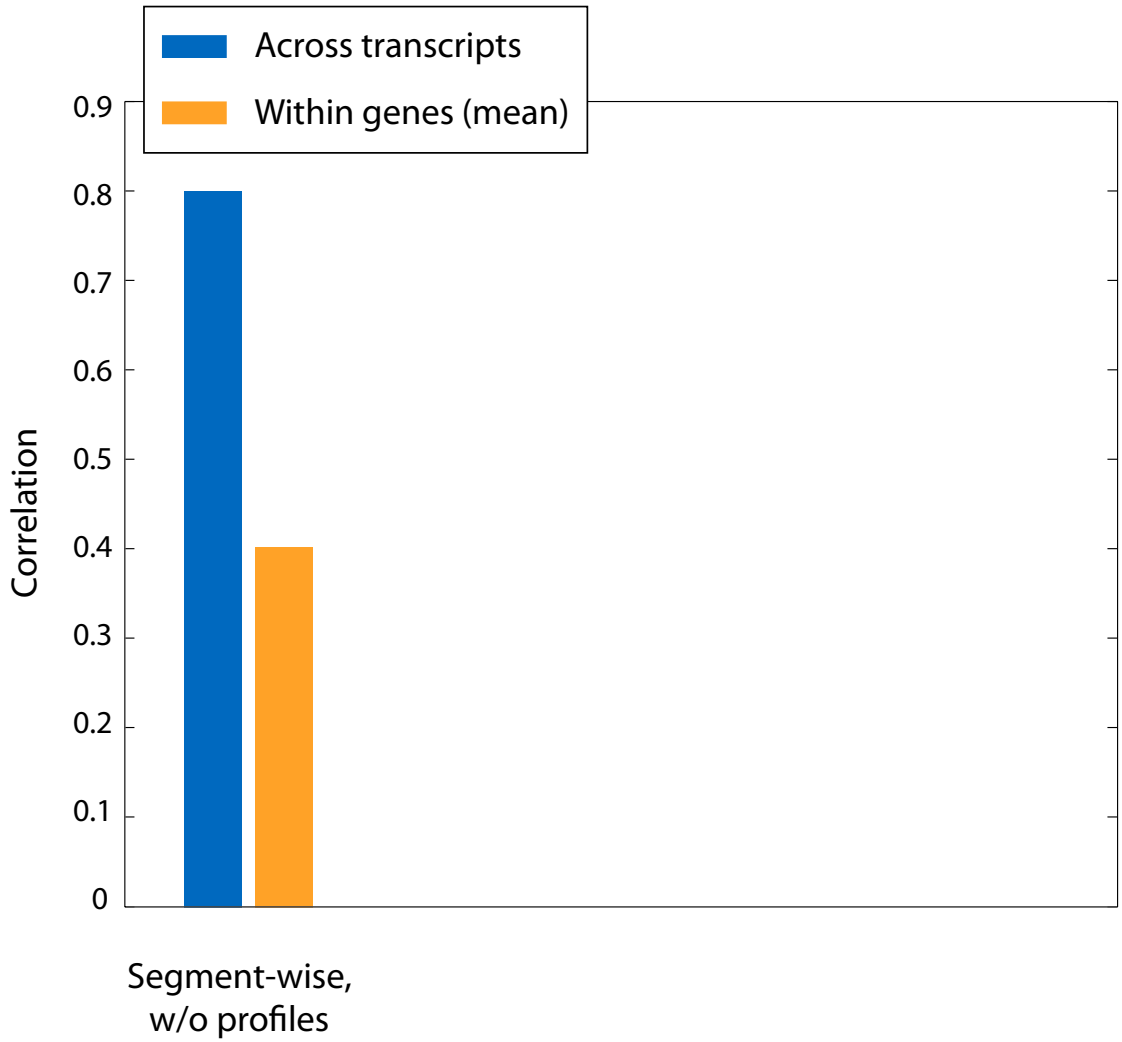
Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

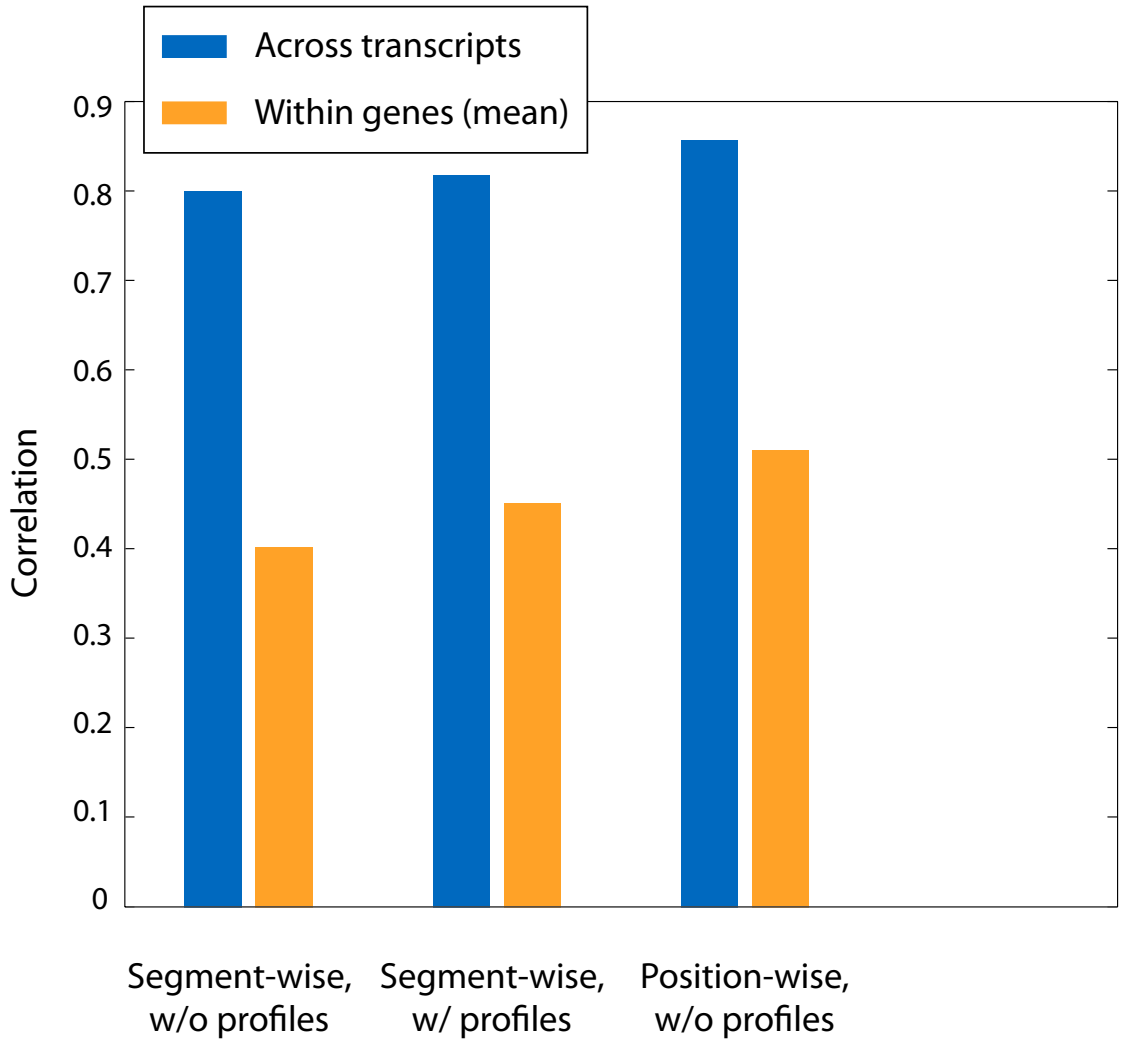
- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation II



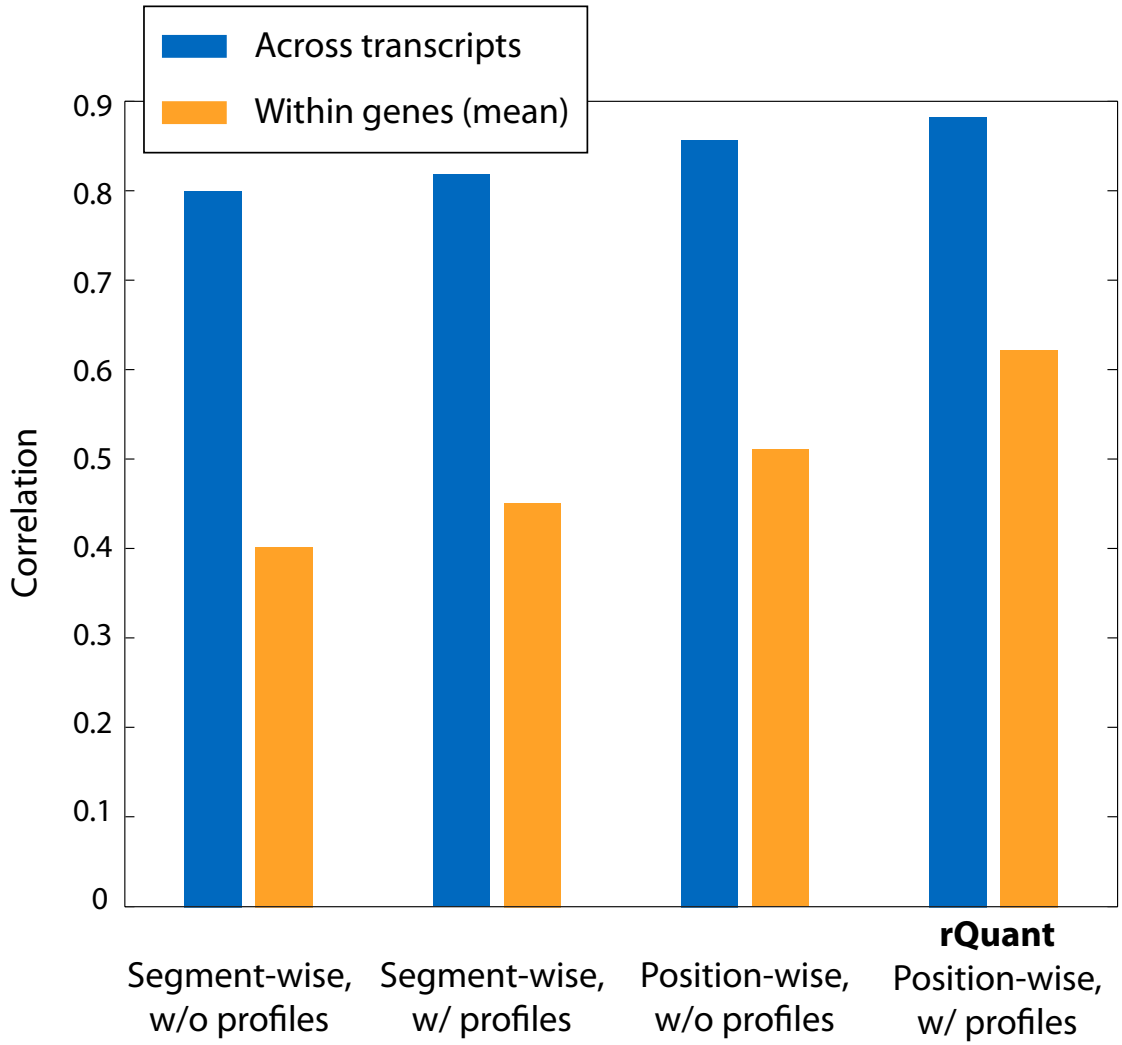
(Bohnert et al., in preparation, 2010)

rQuant Evaluation II



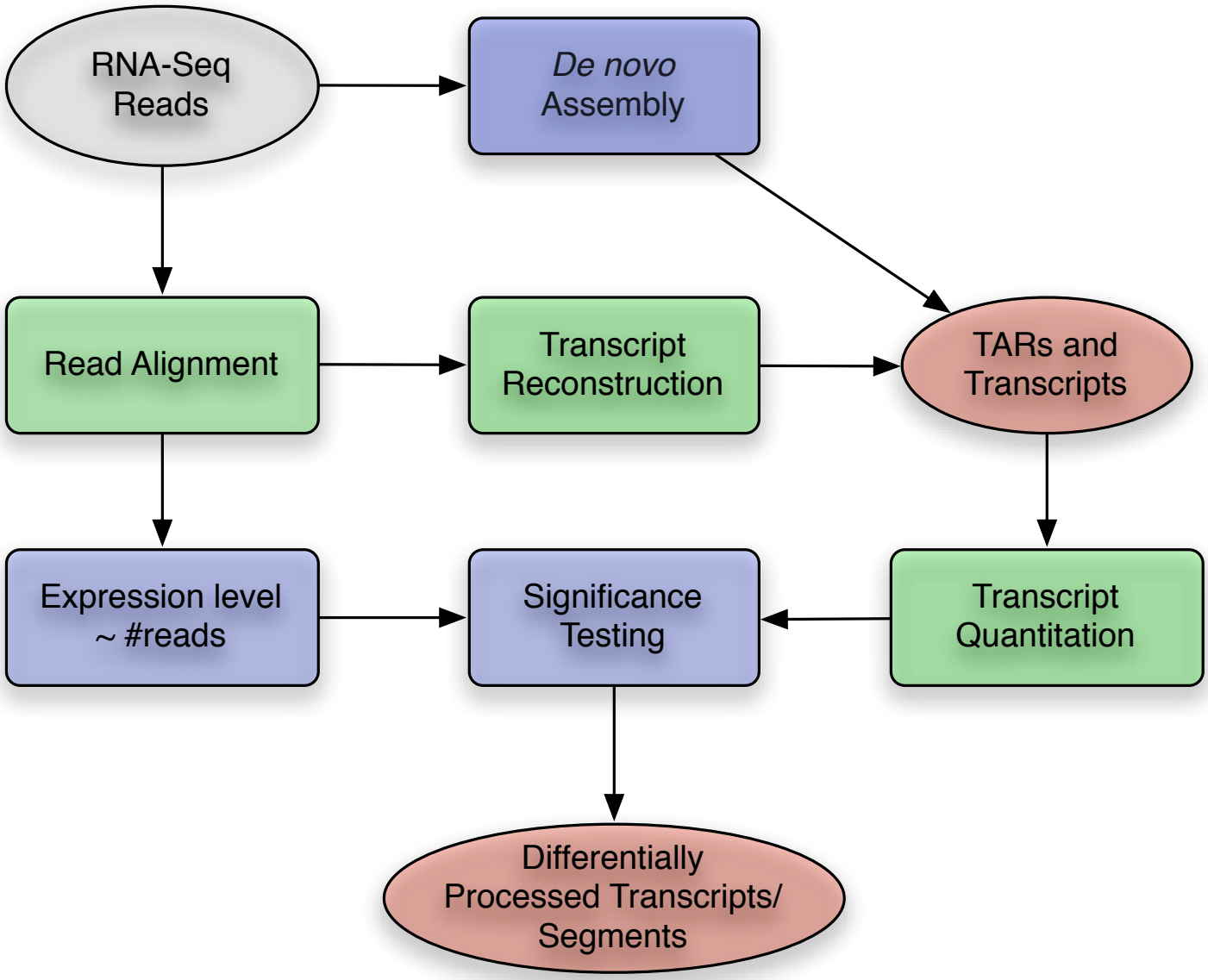
(Bohnert et al., in preparation, 2010)

rQuant Evaluation II



(Bohnert et al., in preparation, 2010)

Common RNA-Seq Analysis Steps



Transcript Identification Strategies



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

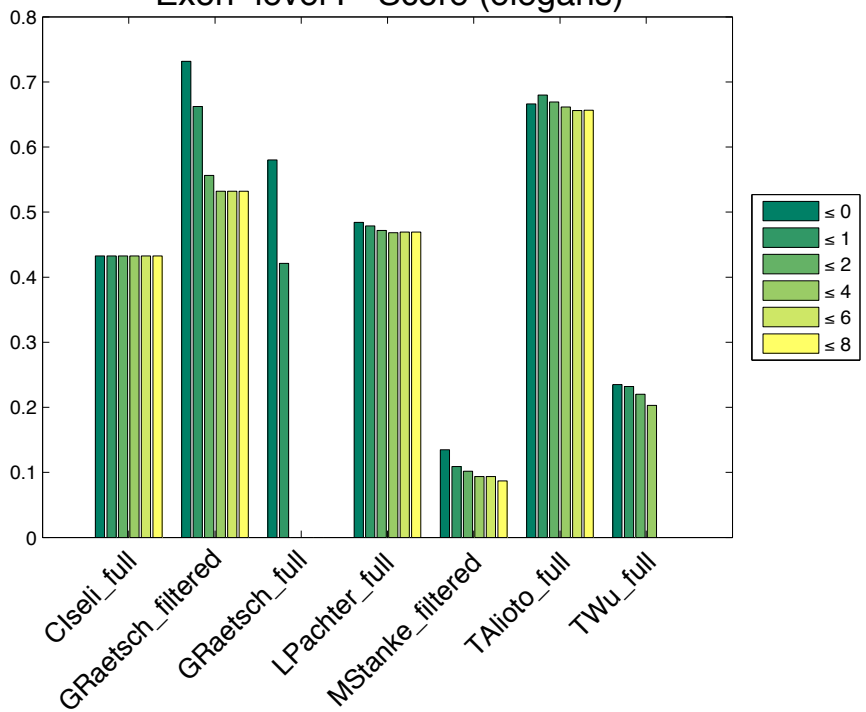
Methods:

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

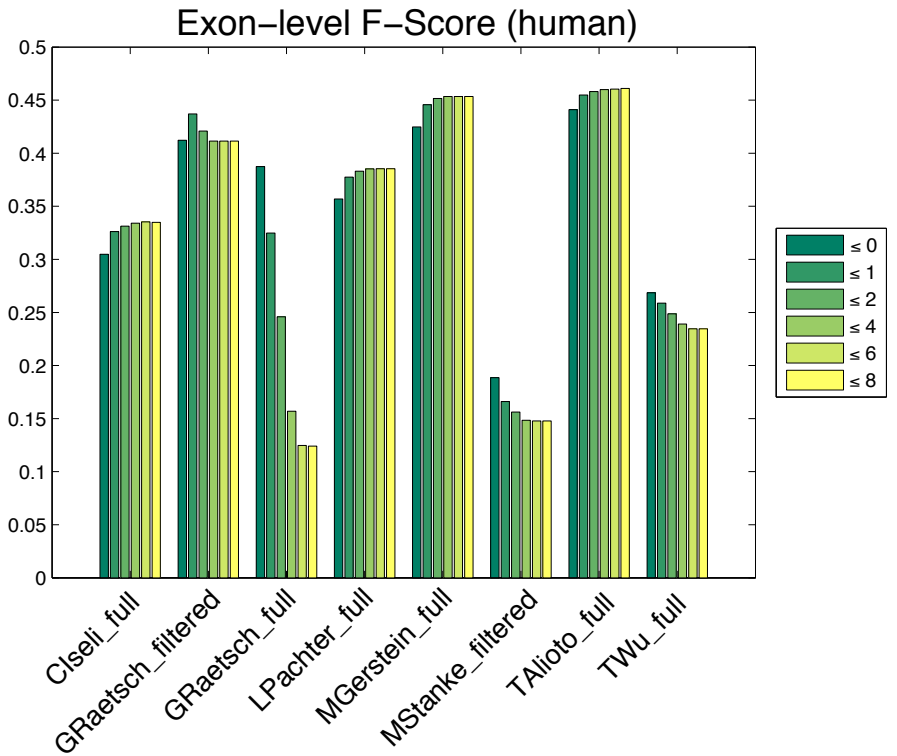
Influence of Read Mapping: here Scripture

Exon-level F-Score of Scripture transcript predictions relative to annotation.

C. elegans



human

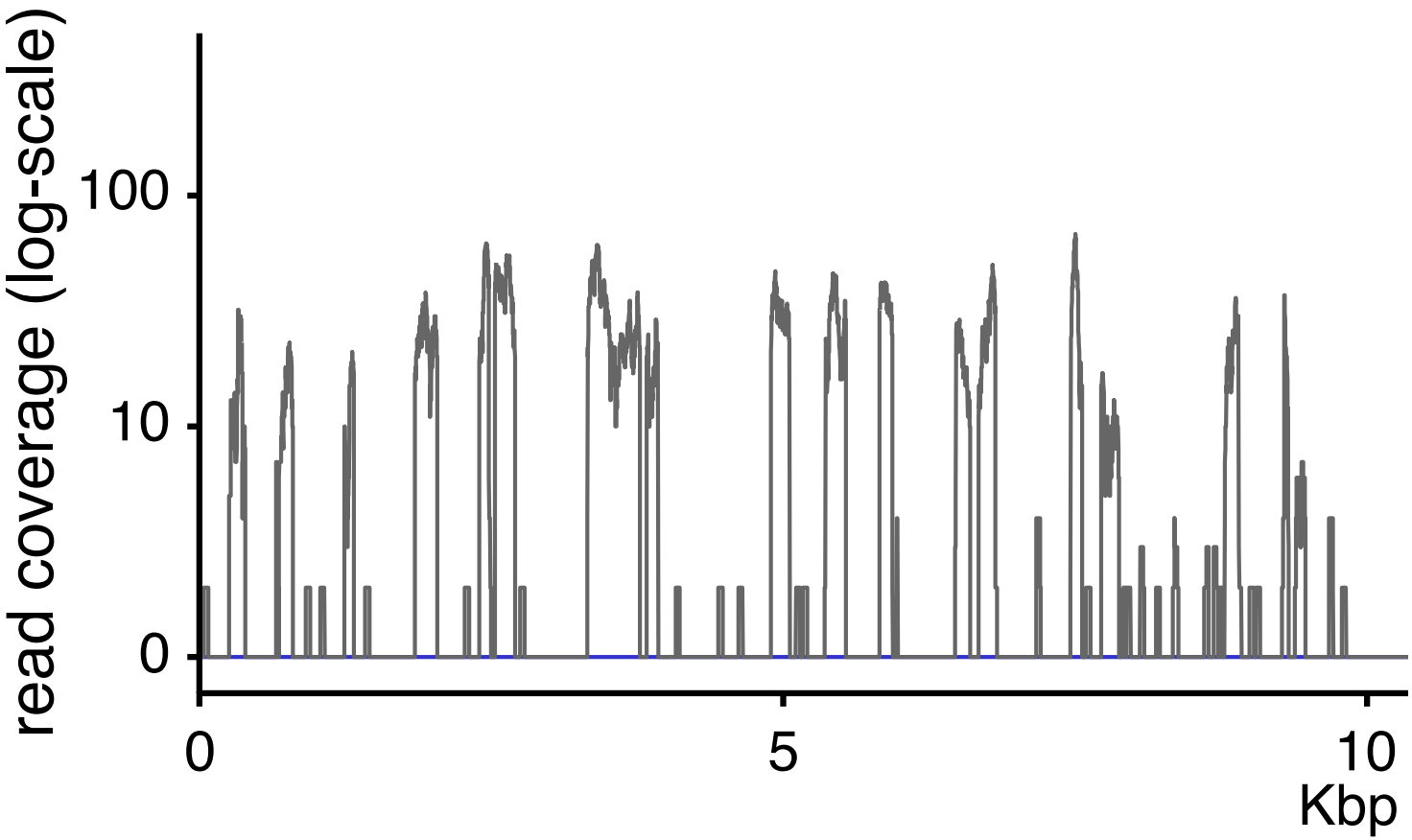


Filtering can significantly increase accuracy.

Scripture on less-filtered alignments may not terminate (GRaetsch_full, TAlioto, TWu).

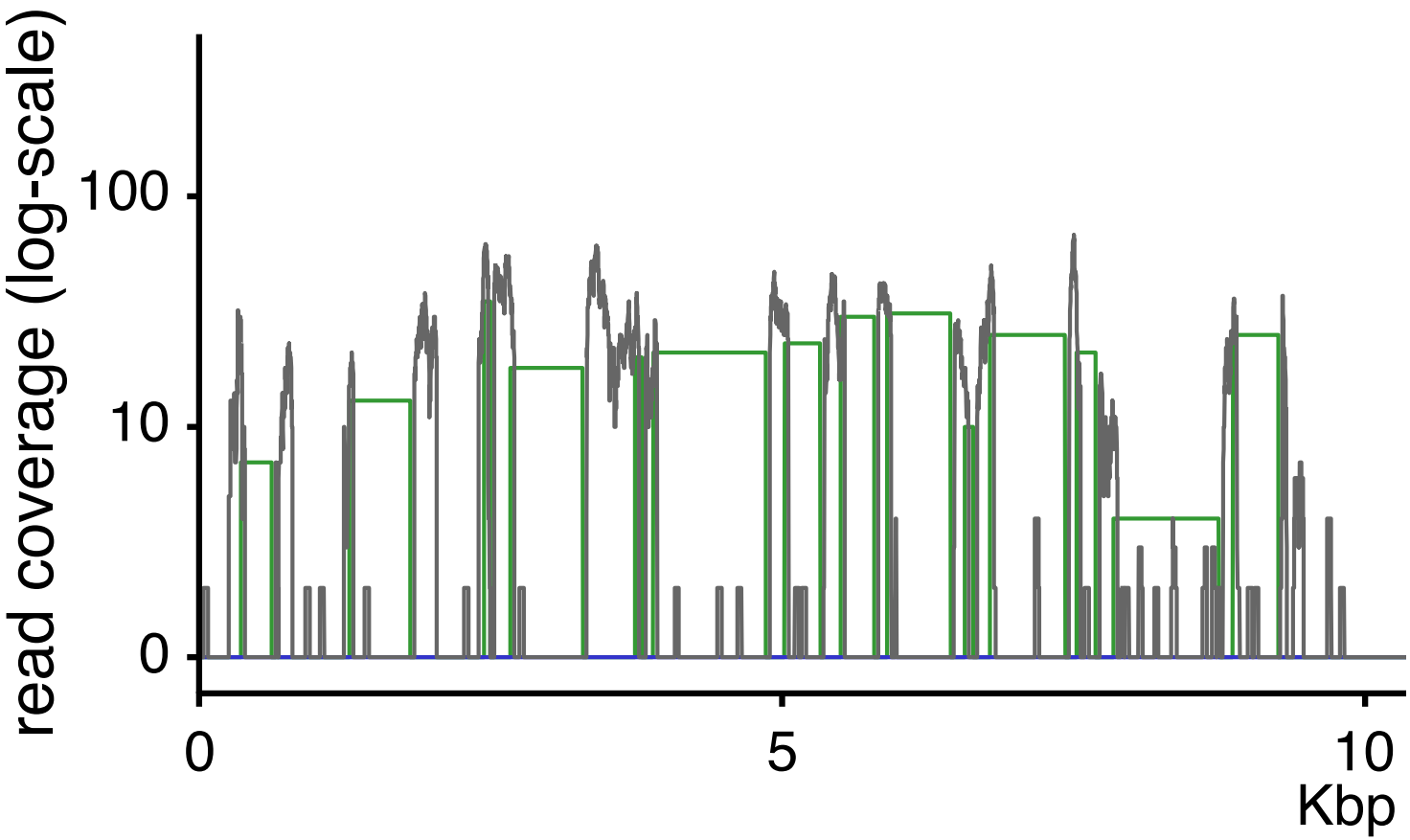
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



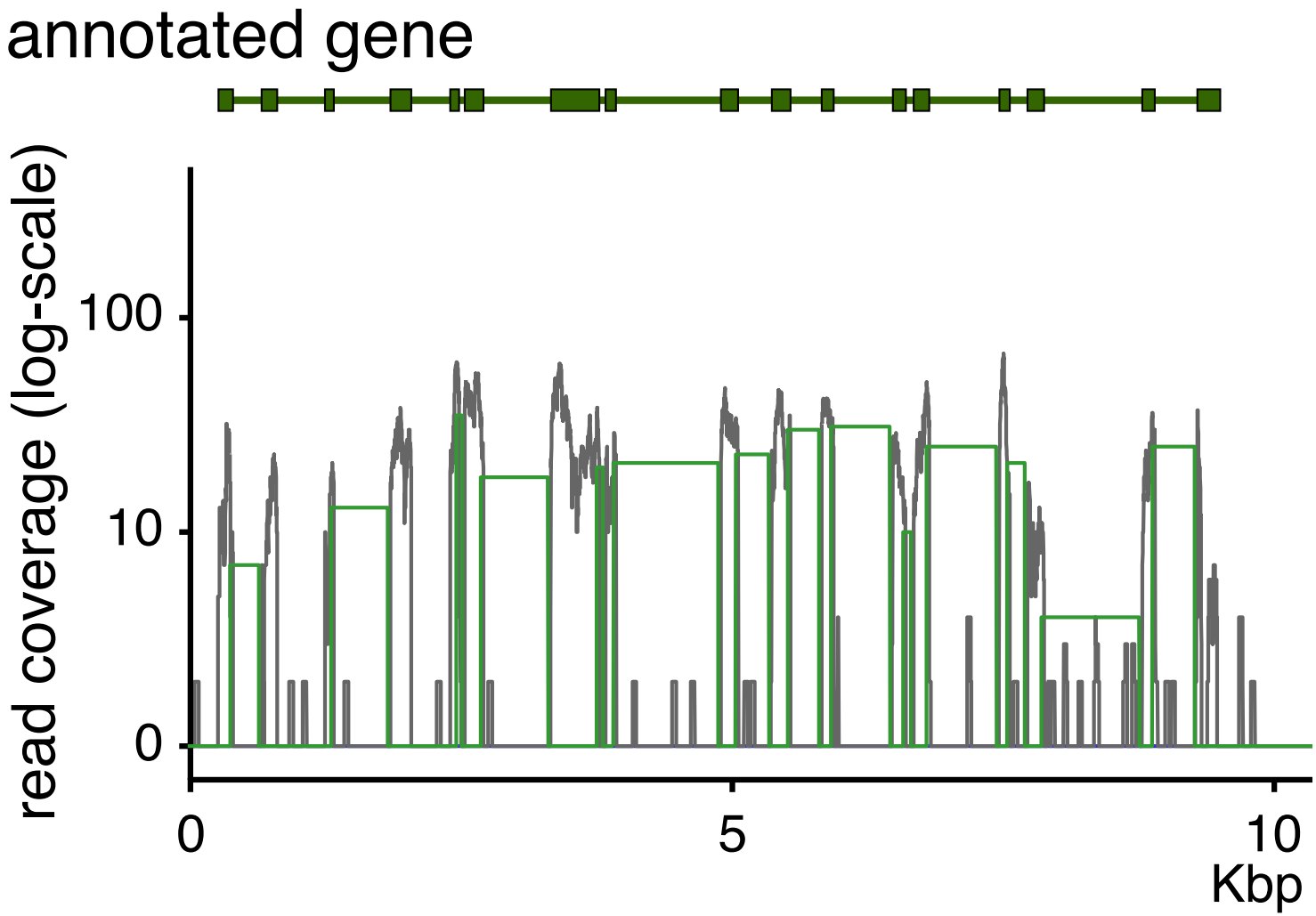
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*

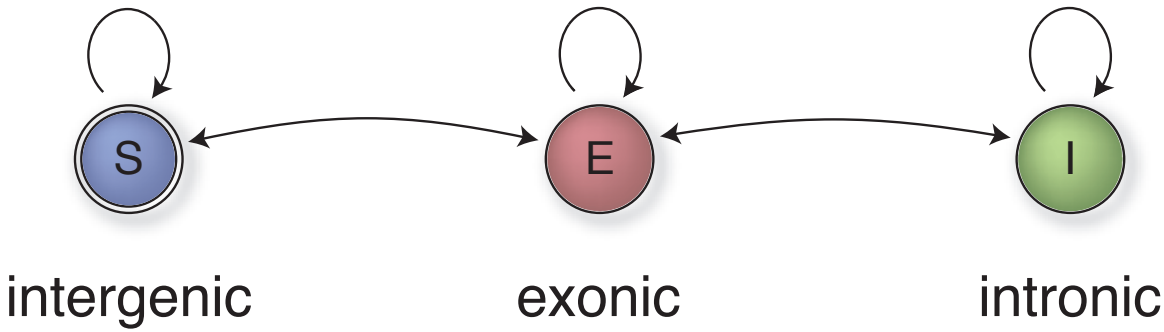


mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



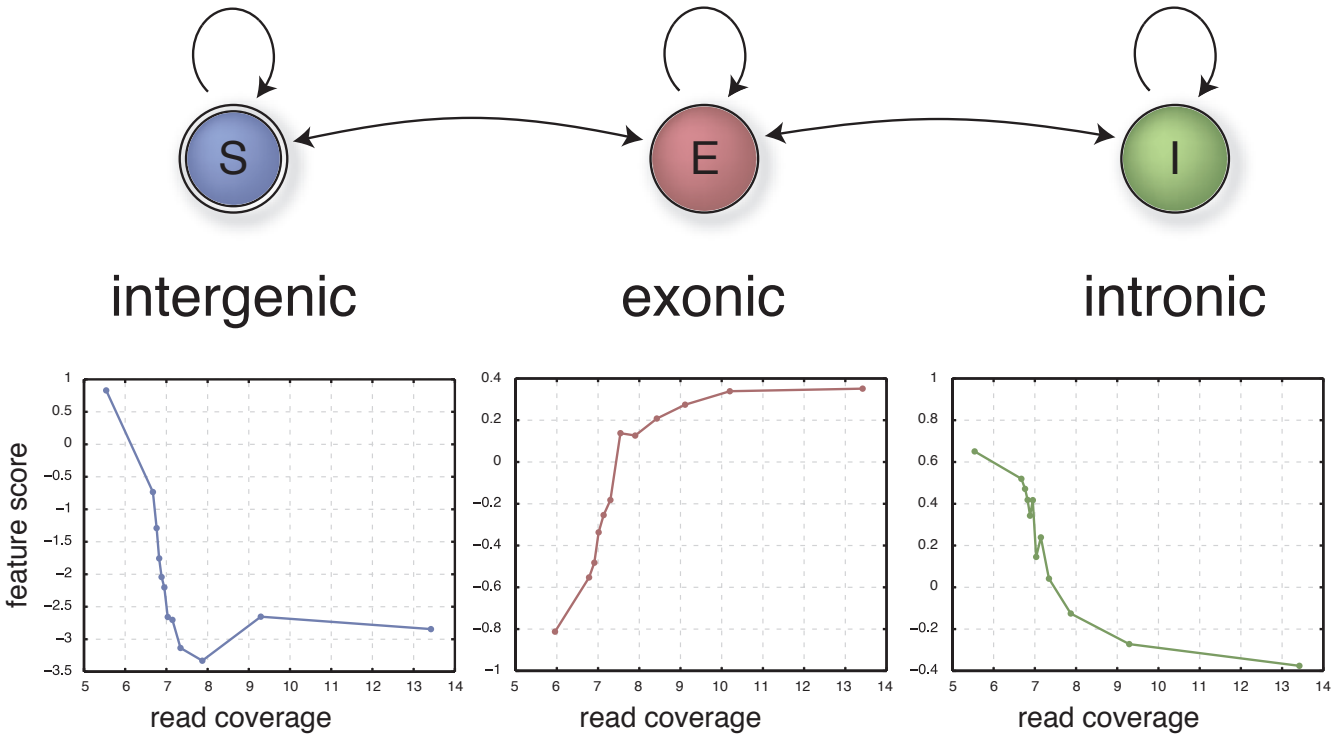
The mTiM Segmentation Approach



- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal → score
- Extension: Score spliced reads and splice sites

(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

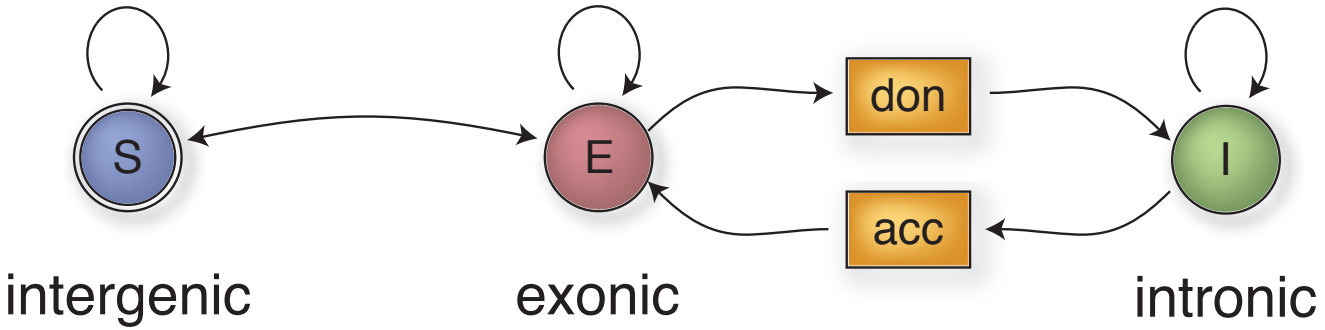
The mTiM Segmentation Approach

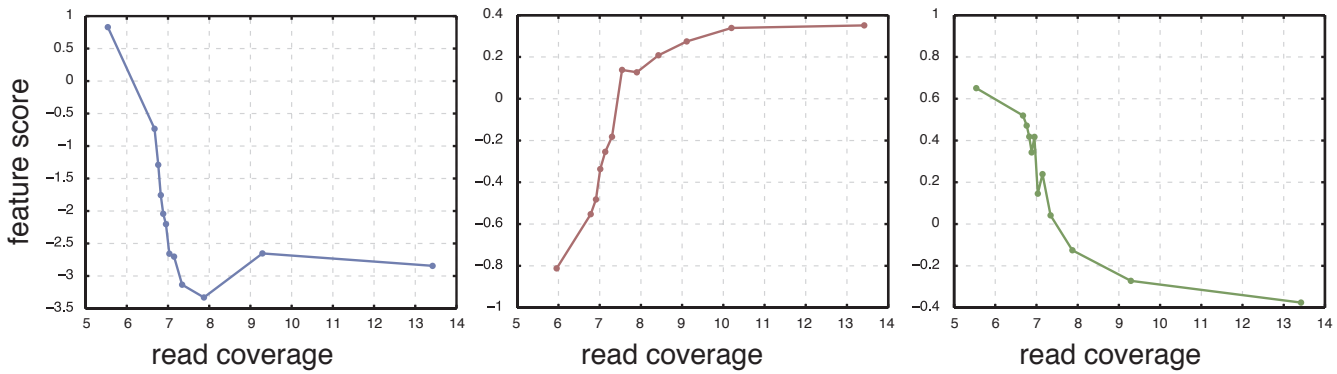


- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal $\longrightarrow$ score
- Extension: Score spliced reads and splice sites

(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

The mTiM Segmentation Approach



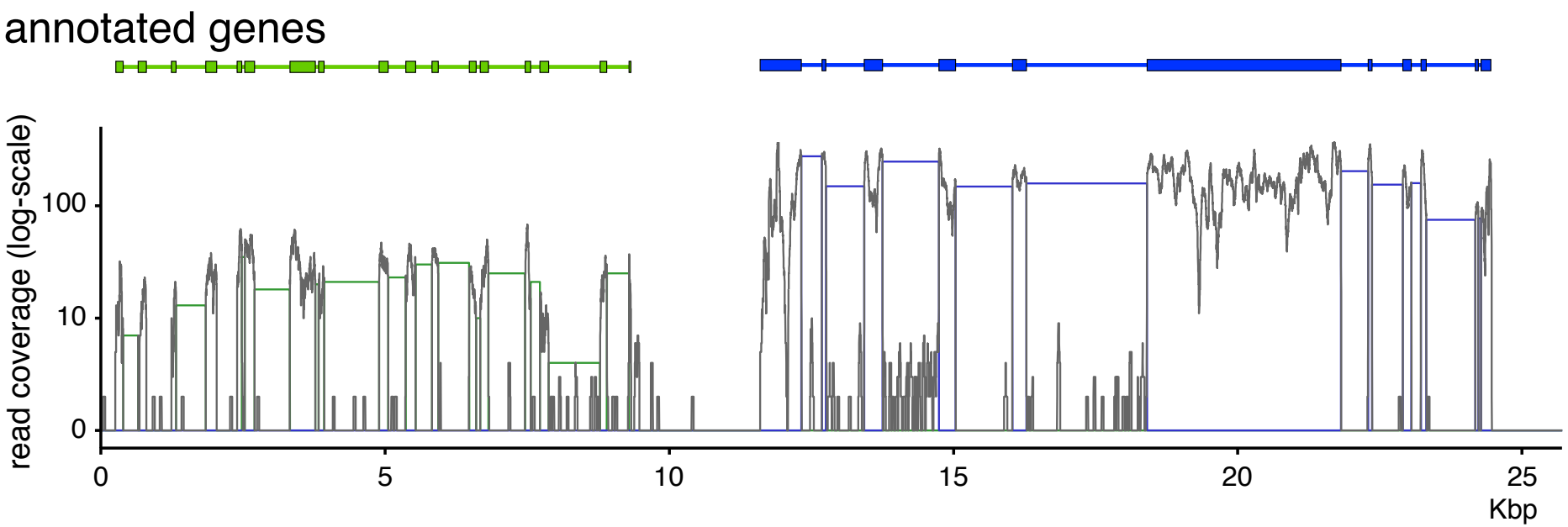


- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal $\longrightarrow$ score
- Extension: Score spliced reads and splice sites

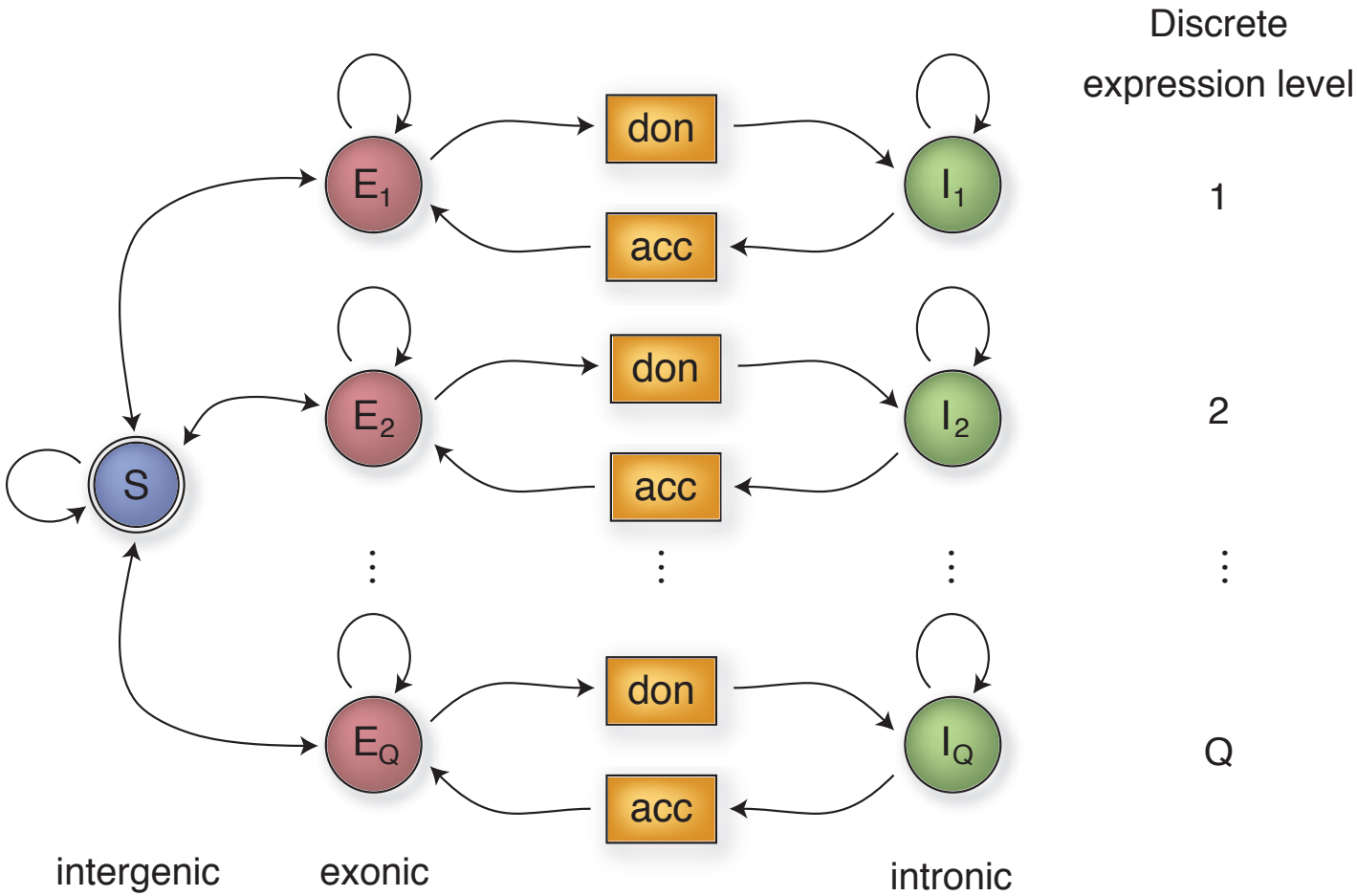
(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

The mTiM Segmentation Approach

Idea: Assume uniform read coverage within exons of same transcript



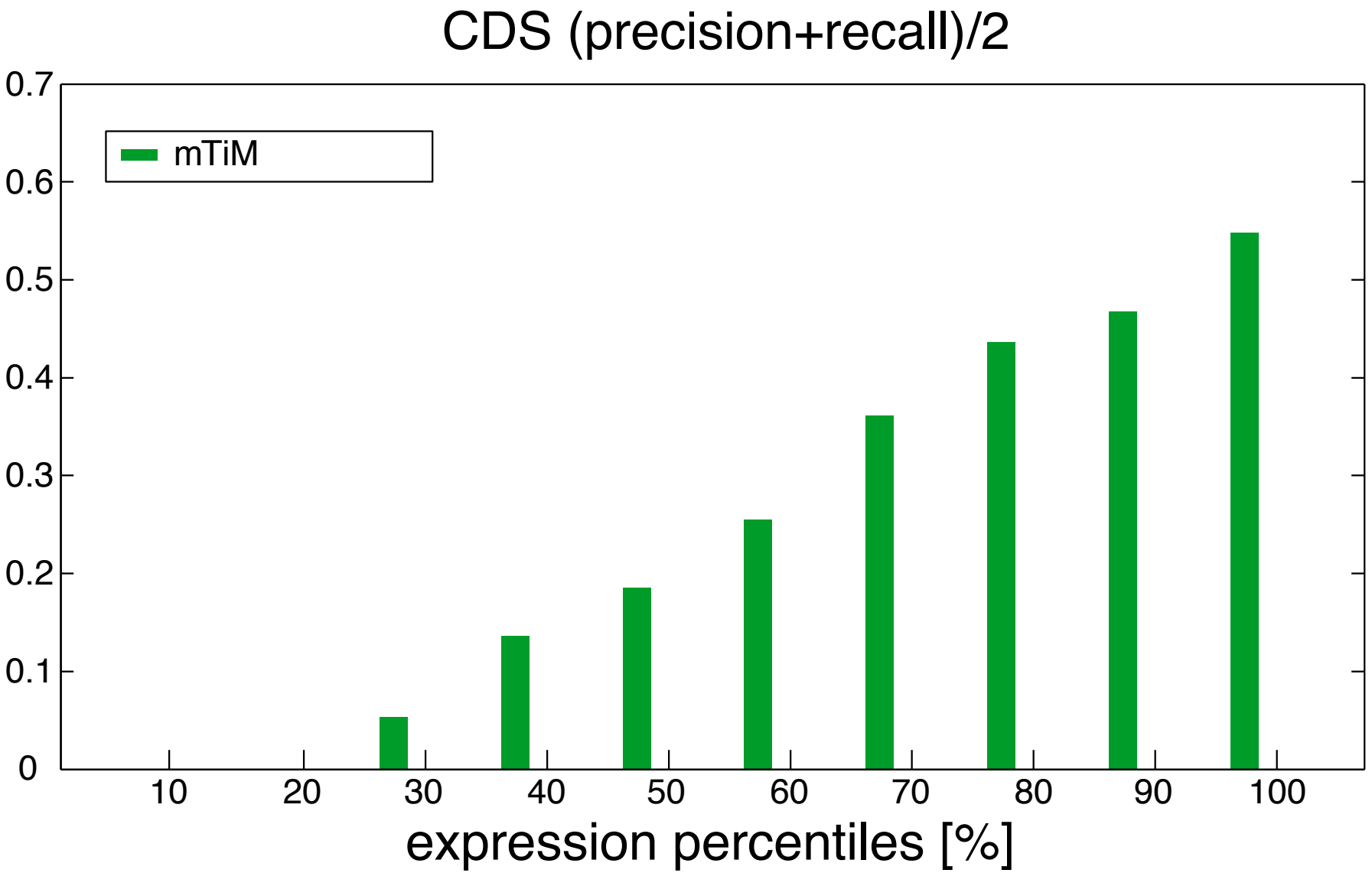
The mTiM Segmentation Approach



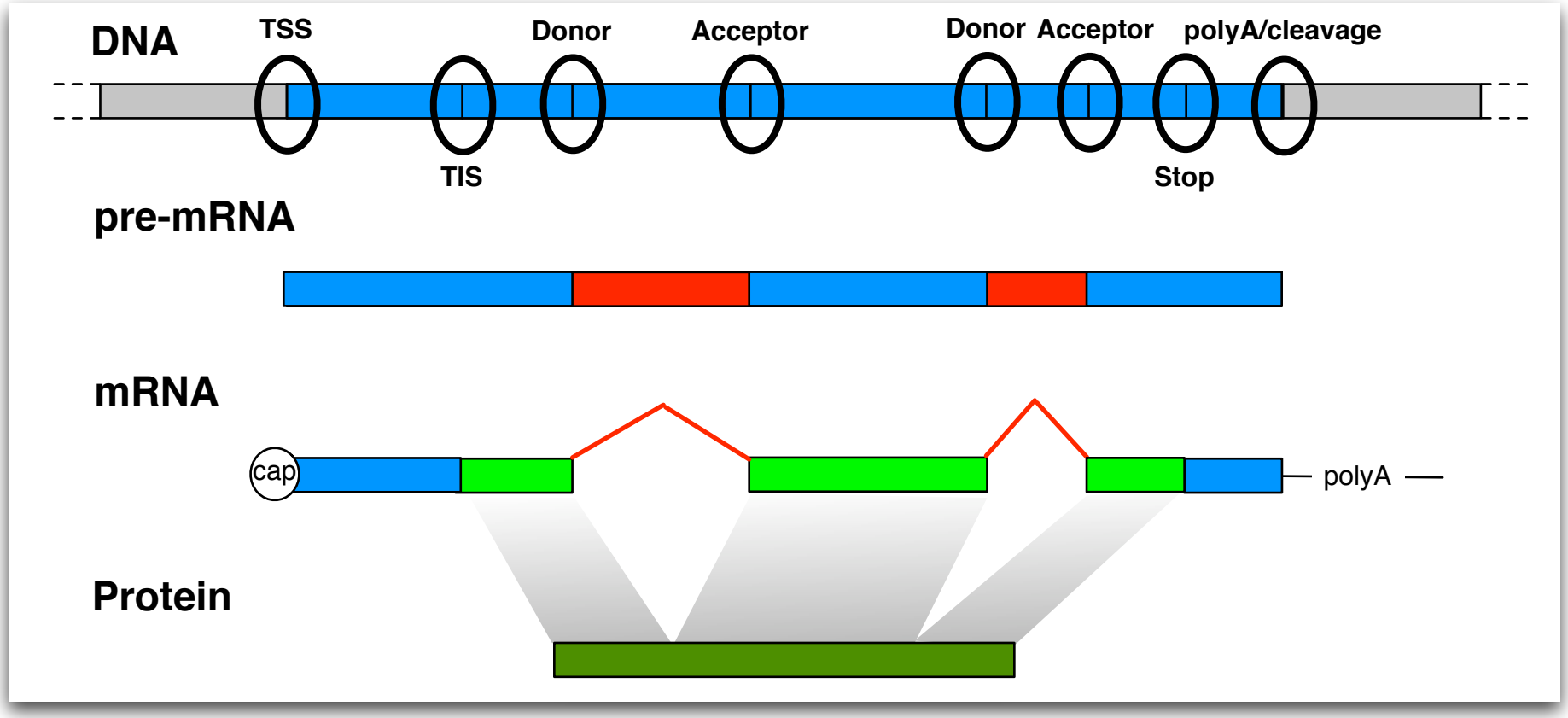
Carry "expression level" information between exons of same transcript

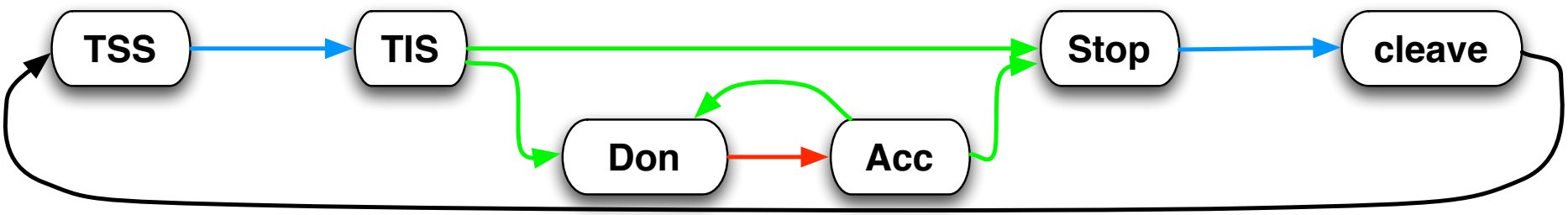
(G. Zeller et al., Pac. Symp. Biocomp., 2008; G. Zeller et al., unpublished, 2010)

Preliminary Evaluation (*C. elegans*)

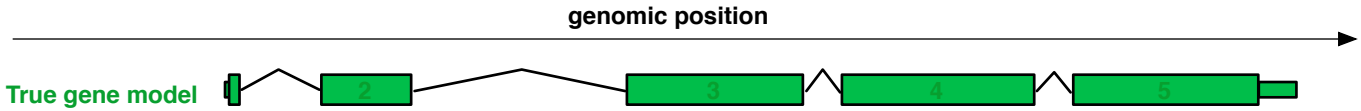


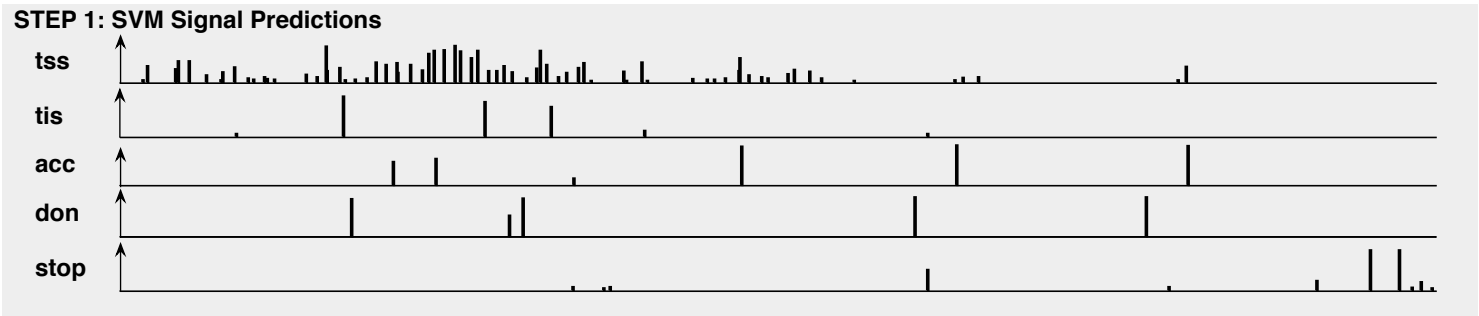
RNA-seq + Gene Finding

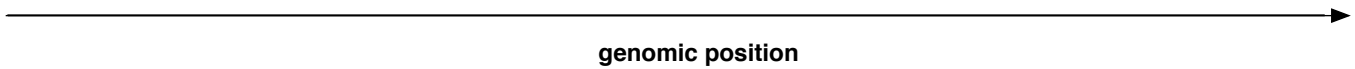




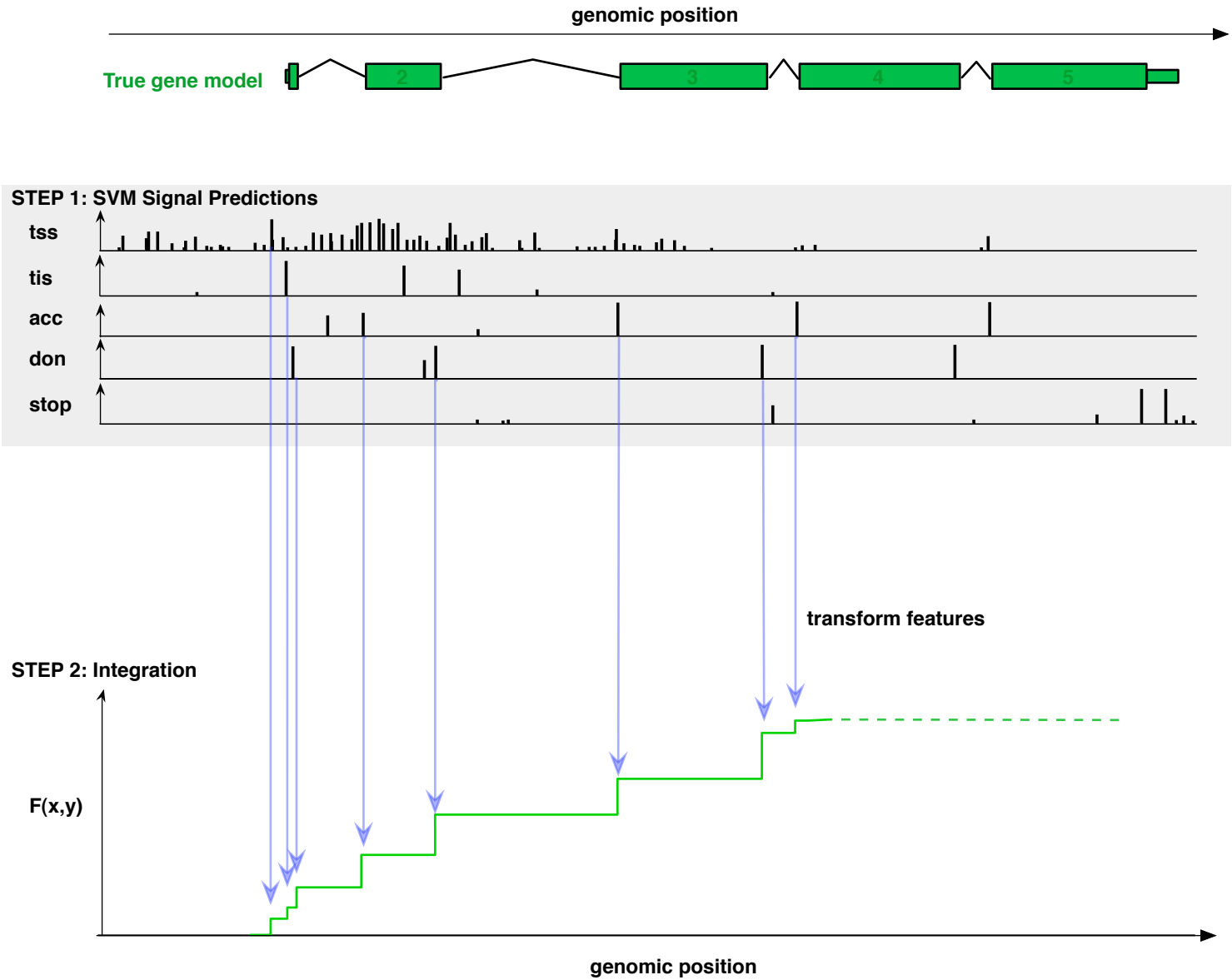
mGene-based Transcript Prediction I



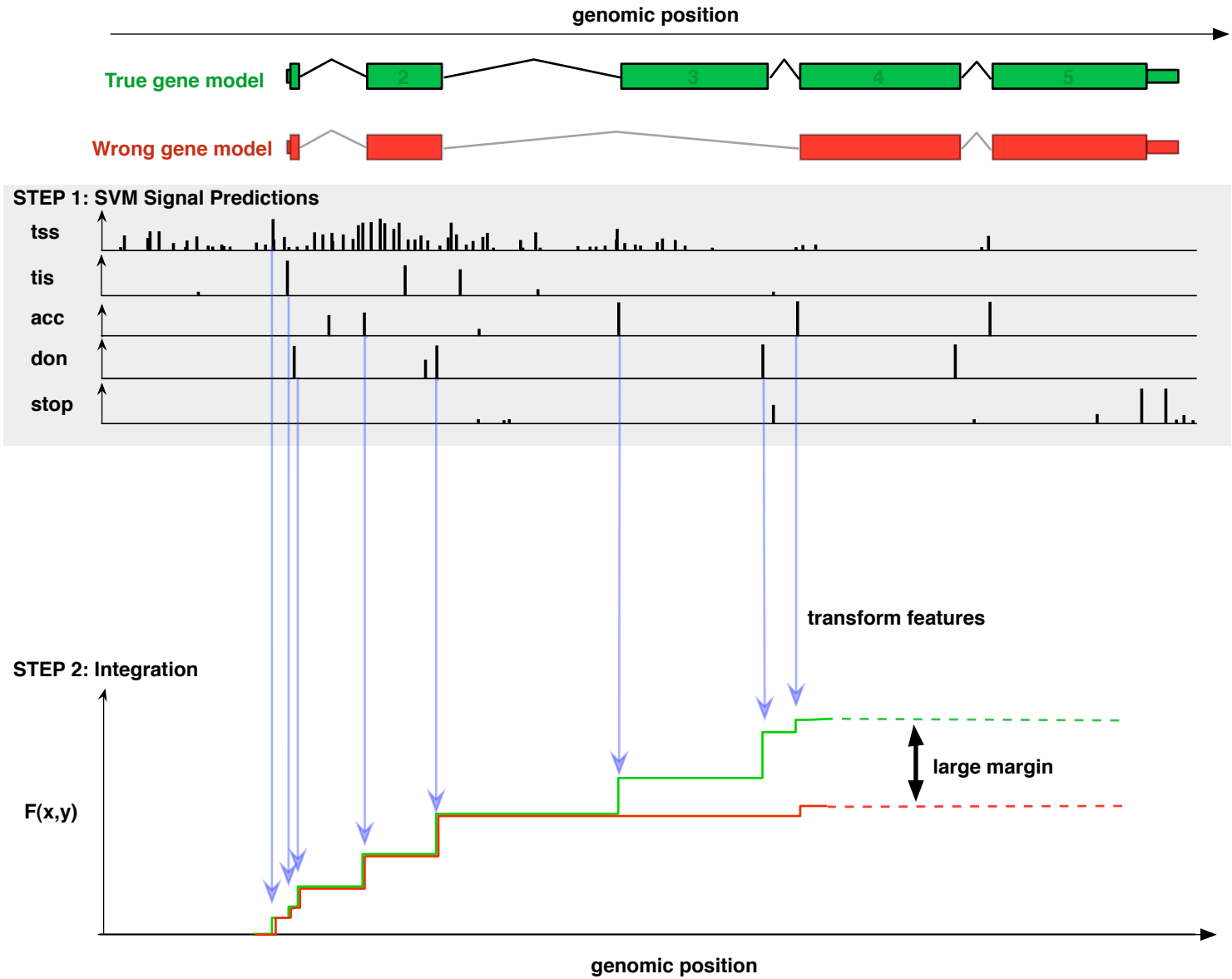




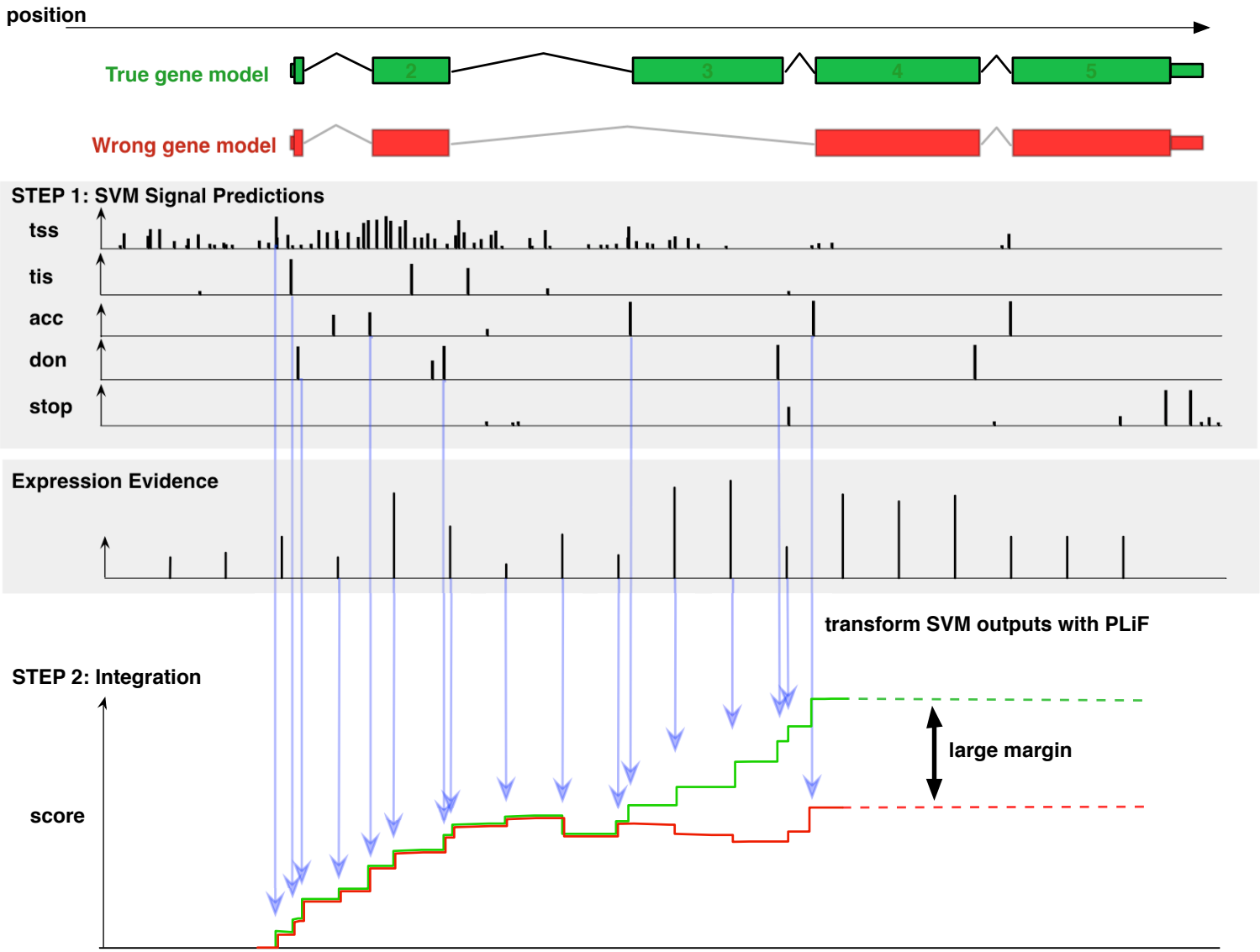
mGene-based Transcript Prediction I



mGene-based Transcript Prediction I



mGene-based Transcript Prediction I

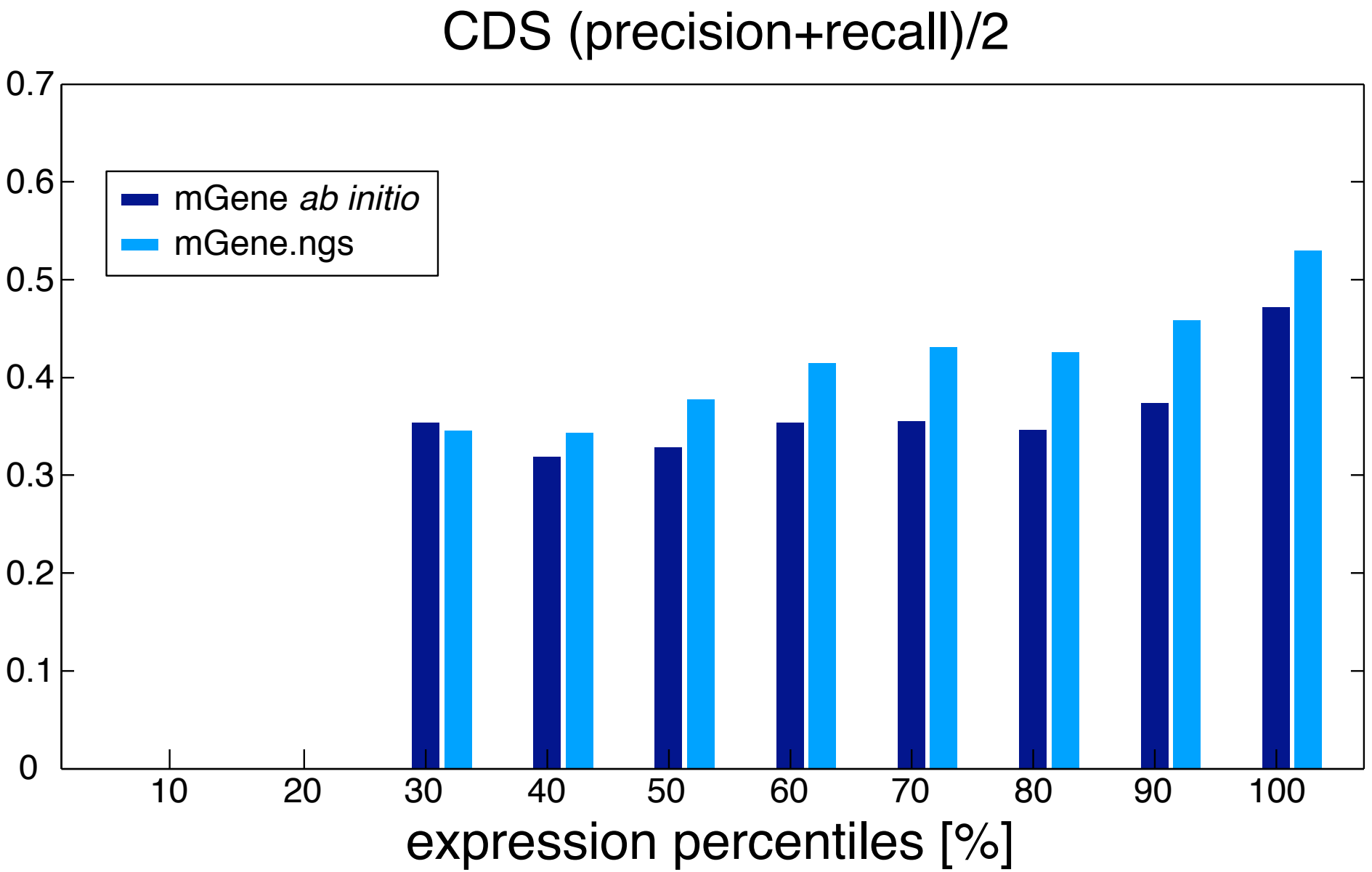


Learning the uncertainty helps!

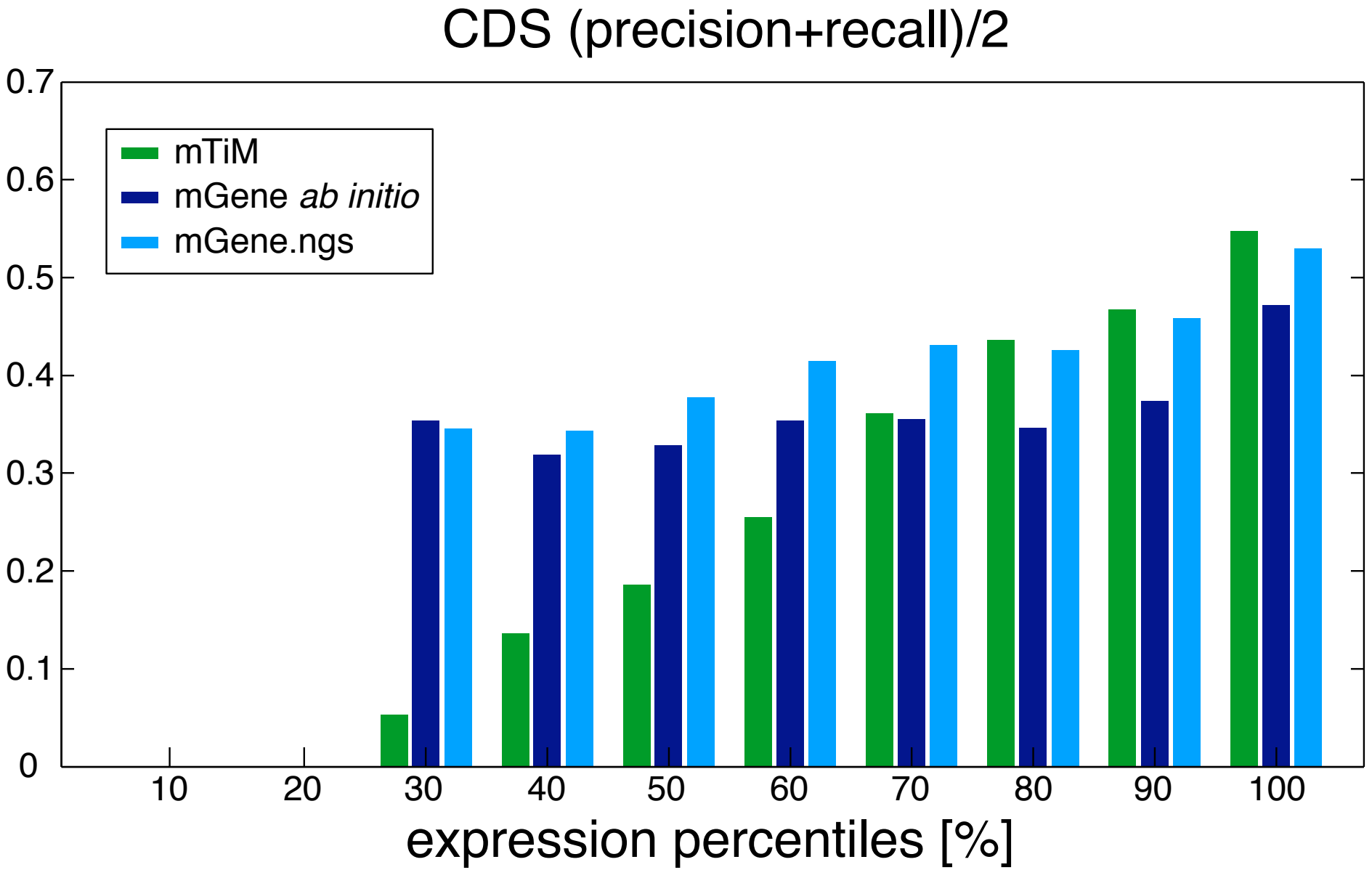


	Exon Level			Transcript Level		
	SN	**SP**	**F**	**SN**	**SP**	**F**
ab initio	82.3	82.6	82.5	43.1	49.5	46.1
ESTs heuristic	85.3	84.7	85.0	49.5	56.4	52.7
ESTs trained	84.8	85.8	85.3	50.5	57.8	53.9
RNA-Seq and ESTs	84.7	86.9	85.8	50.3	60.5	54.9
RNA-Seq data	84.6	84.9	84.8	49.1	55.2	52.0

Preliminary Evaluation (*C. elegans*)



Preliminary Evaluation (*C. elegans*)



Digestion

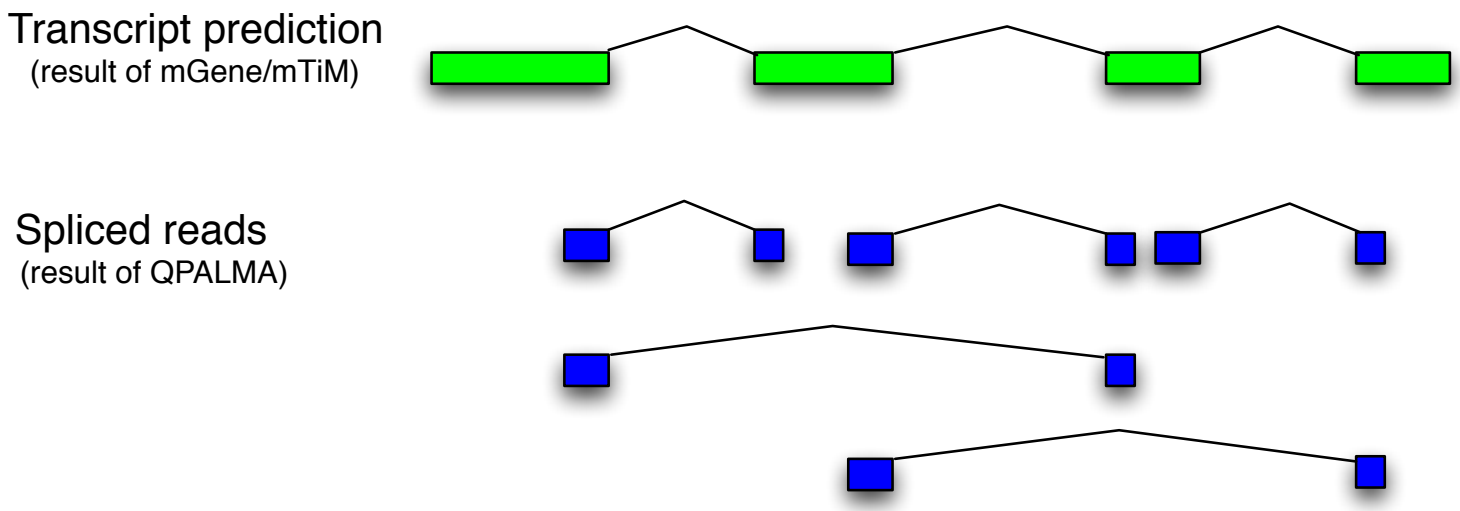
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

Digestion

- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

Digestion

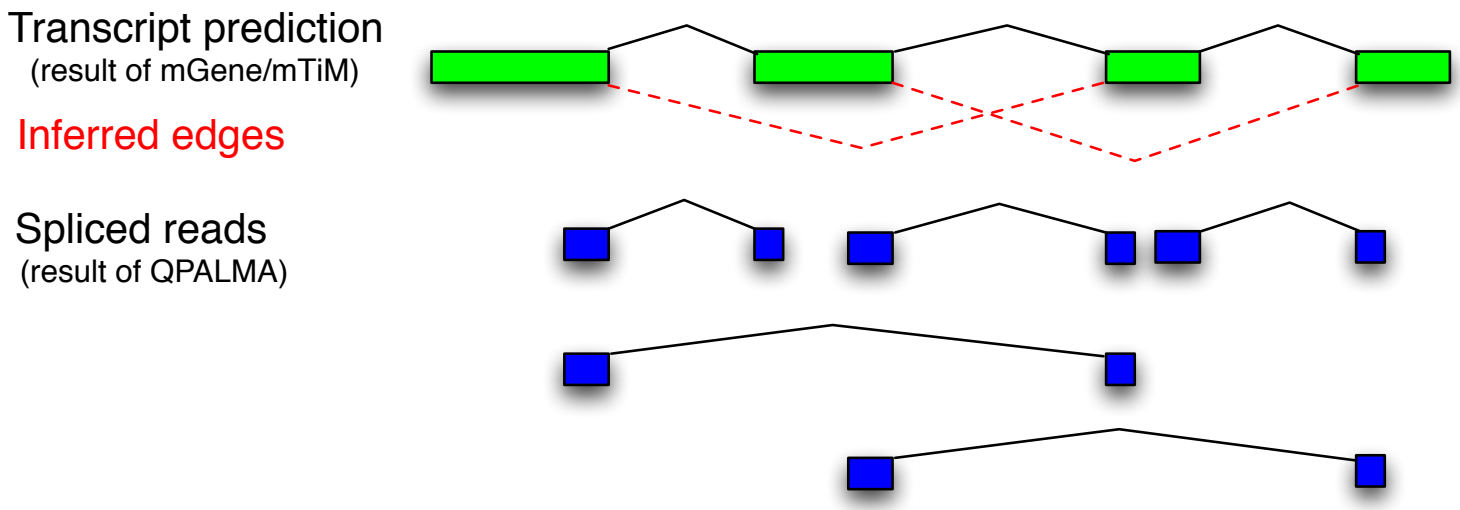
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



- Paths through splicing graph define *alternative transcripts*

Digestion

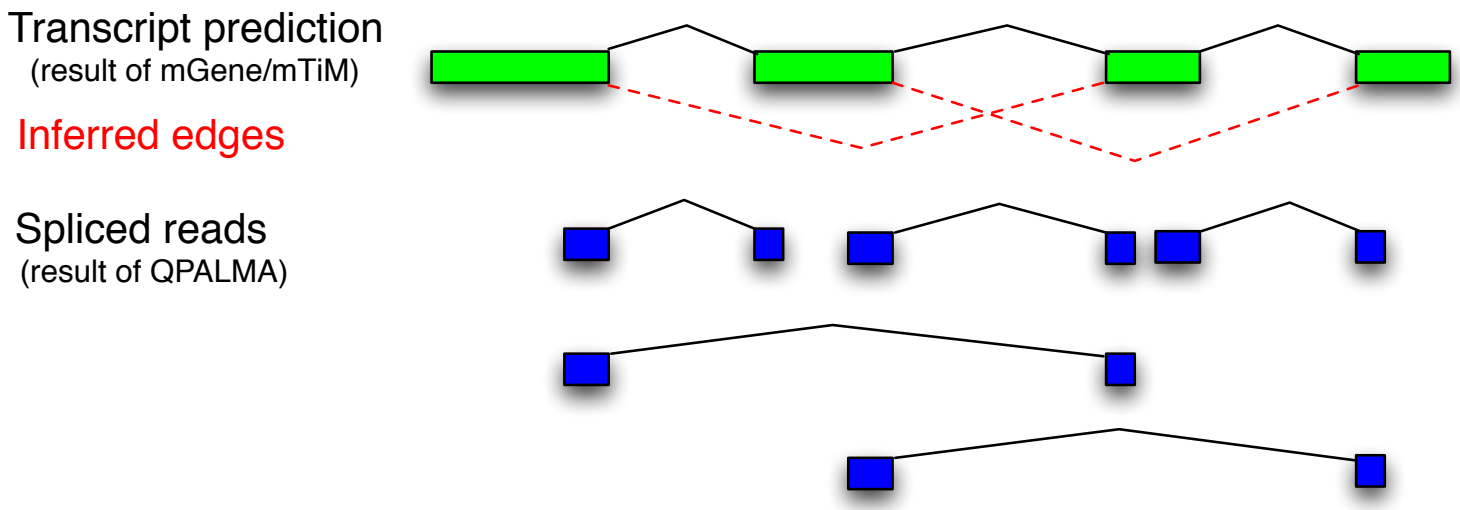
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



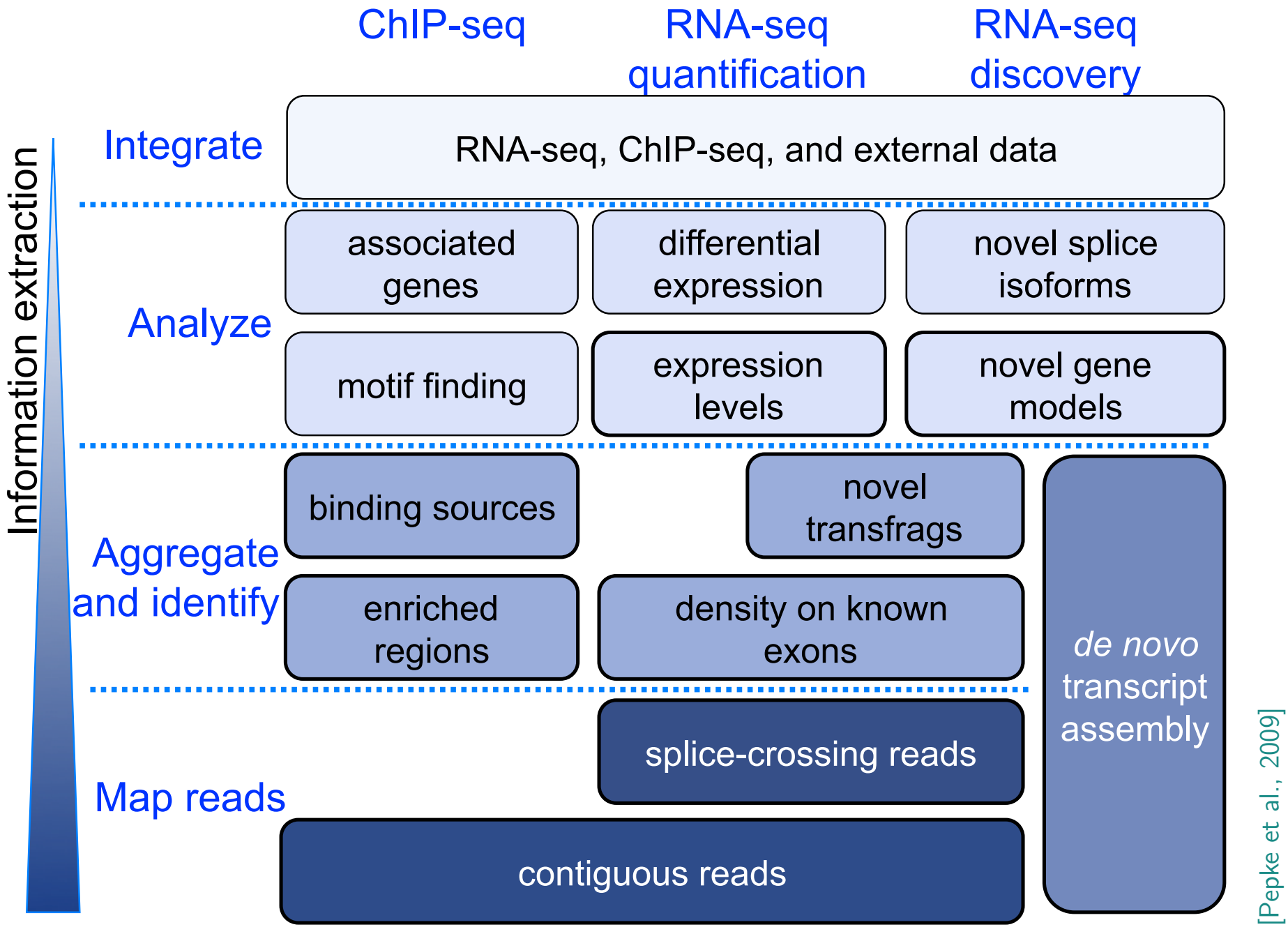
- Paths through splicing graph define *alternative transcripts*

Digestion

- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



- Paths through splicing graph define *alternative transcripts*


ChIP-seq
RNA-seq quantification
RNA-seq discovery
Information extraction
Integrate
RNA-seq, ChIP-seq, and external data
Analyze
associated genes
differential expression
novel splice isoforms
motif finding
expression levels
novel gene models
Aggregate and identify
binding sources
novel transfrags
enriched regions
density on known exons
Map reads
splice-crossing reads
contiguous reads
de novo transcript assembly
[Pepke et al., 2009]

Gene Finding + Auxiliary Data

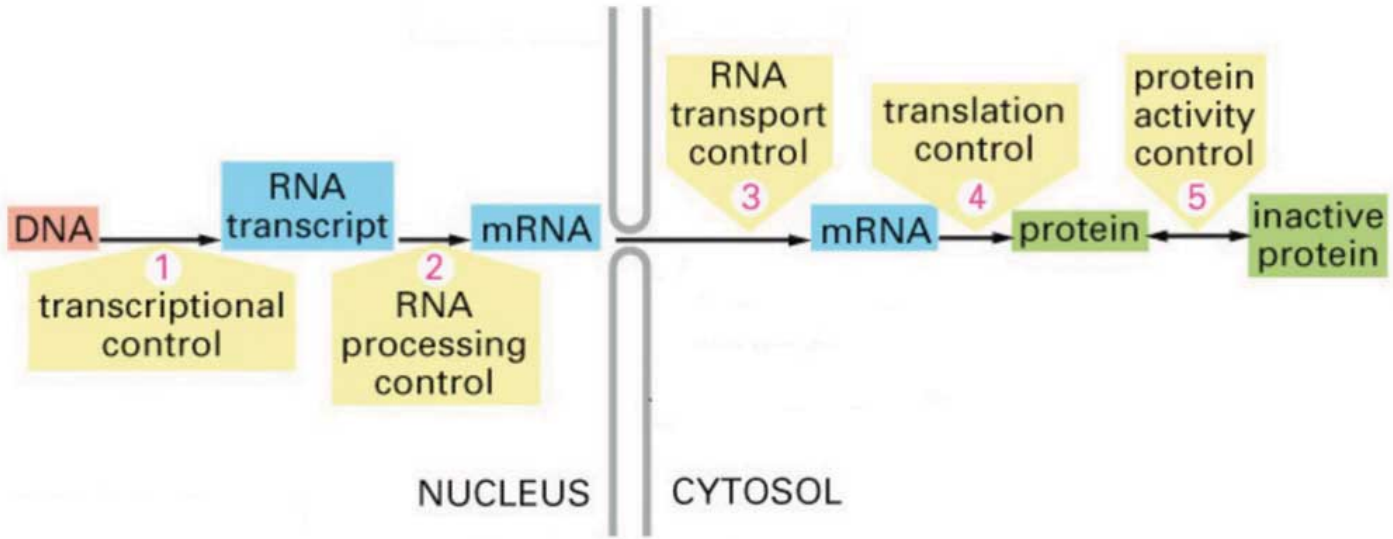


Is there other information that may be used by cellular processes that can improve prediction results?

Preliminary study: (*A. thaliana*) transcript level $(SN + SP)/2$

1. mGene (*ab initio*) . . . **73.3%**
2. . . . with DNA methylation (1 tissue) **76.1%**
3. . . . with Nucleosome position predictions **78.0%**
4. . . . with RNA secondary structure predictions **76.7%**

Learning about the Transcriptome



Summary and Conclusions



Methods:

1. PALMapper: Splice site predictions improve alignments
2. mTiM: Identifies transcripts specific to experiment
3. mGene: Best for annotation; also finds lowly expressed genes
4. rQuant: Models library prep., sequencing, alignment biases

Unsolved problems:

1. Better prediction of alternative transcripts (without RNA-seq?)
2. Differential expression detection [Stegle et al., 2010]
3. Multi-modal modelling to handle the many different ways genes are processed
4. Data integration; very promising for detecting missing components, for instance, in gene prediction systems

Summary and Conclusions



Methods:

1. PALMapper: Splice site predictions improve alignments
2. mTiM: Identifies transcripts specific to experiment
3. mGene: Best for annotation; also finds lowly expressed genes
4. rQuant: Models library prep., sequencing, alignment biases

Unsolved problems:

1. Better prediction of alternative transcripts (without RNA-seq?)
2. Differential expression detection [Stegle et al., 2010]
3. Multi-modal modelling to handle the many different ways genes are processed
4. Data integration; very promising for detecting missing components, for instance, in gene prediction systems

Summary and Conclusions



Methods:

1. PALMapper: Splice site predictions improve alignments
2. mTiM: Identifies transcripts specific to experiment
3. mGene: Best for annotation; also finds lowly expressed genes
4. rQuant: Models library prep., sequencing, alignment biases

Unsolved problems:

1. Better prediction of alternative transcripts (without RNA-seq?)
2. Differential expression detection [Stegle et al., 2010]
3. Multi-modal modelling to handle the many different ways genes are processed
4. Data integration; very promising for detecting missing components, for instance, in gene prediction systems

Summary and Conclusions



Methods:

1. PALMapper: Splice site predictions improve alignments
2. mTiM: Identifies transcripts specific to experiment
3. mGene: Best for annotation; also finds lowly expressed genes
4. rQuant: Models library prep., sequencing, alignment biases

Unsolved problems:

1. Better prediction of alternative transcripts (without RNA-seq?)
2. Differential expression detection [Stegle et al., 2010]
3. Multi-modal modelling to handle the many different ways genes are processed
4. Data integration; very promising for detecting missing components, for instance, in gene prediction systems

Summary and Conclusions



Methods:

1. PALMapper: Splice site predictions improve alignments
2. mTiM: Identifies transcripts specific to experiment
3. mGene: Best for annotation; also finds lowly expressed genes
4. rQuant: Models library prep., sequencing, alignment biases

Unsolved problems:

1. Better prediction of alternative transcripts (without RNA-seq?)
2. Differential expression detection [Stegle et al., 2010]
3. Multi-modal modelling to handle the many different ways genes are processed
4. Data integration; very promising for detecting missing components, for instance, in gene prediction systems

Acknowledgments

We gratefully acknowledge help with the material:

- Ali Mortazavi, Caltech
- Jonas Behr, Friedrich Miescher Laboratory
- Regina Bohnert, Friedrich Miescher Laboratory
- Fabio De Bona, Friedrich Miescher Laboratory & Google Research
- André Kahles, Friedrich Miescher Laboratory
- Stefan Henz, MPI for Developmental Biology

The slides and additional material will be available online at

`http://www.fml.mpg.de/raetsch/lectures/erice10a`

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`.

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.

F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Qpalma: Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24:i174–i180, 2008.

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.

References II

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.

Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9): e1000502, Sep 2009. doi: 10.1371/journal.pcbi.1000502.

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

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10 (3):R25, 2009. doi: 10.1186/gb-2009-10-3-r25.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, Jul 2009. doi: 10.1093/bioinformatics/btp324.

References III

Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–95, Mar 2010. doi: 10.1093/bioinformatics/btp698.

Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–8, Nov 2008. doi: 10.1101/gr.078212.108.

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

Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15): 1966–7, Aug 2009. doi: 10.1093/bioinformatics/btp336.

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.

Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for chip-seq and rna-seq studies. *Nat Methods*, 6(11 Suppl):S22–32, Nov 2009. doi: 10.1038/nmeth.1371.

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.

Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5): e1000386, May 2009. doi: 10.1371/journal.pcbi.1000386.

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.

Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biol*, 10(9):R98, 2009a. doi: 10.1186/gb-2009-10-9-r98.

Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biol*, 10(9):R98, Jan 2009b. doi: 10.1186/gb-2009-10-9-r98. URL `http://genomebiology.com/2009/10/9/R98`.

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.

O. Stegle, P. Drewe, R. Bohnert, K. Borgwardt, and G. Rätsch. Statistical tests for detecting differential rna-transcript expression from read counts. Technical report, Nature Preceedings, 2010.

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 2010. doi: 10.1038/nbt.1621. URL `http://dx.doi.org/10.1038/nbt.1621`.

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.

References VI

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.