Machine Learning Methods for RNA-seq-based Transcriptome Reconstruction

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society, Tübingen, Germany

StatSeq Workshop, Gent (Mai 20, 2010)



Friedrich Miescher Laboratory
of the Max Planck Society



Discovery of the Nuclein
(Friedrich Miescher, 1869)









Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein" (Miescher, 1874)

Discovery of the Nuclein
(Friedrich Miescher, 1869)









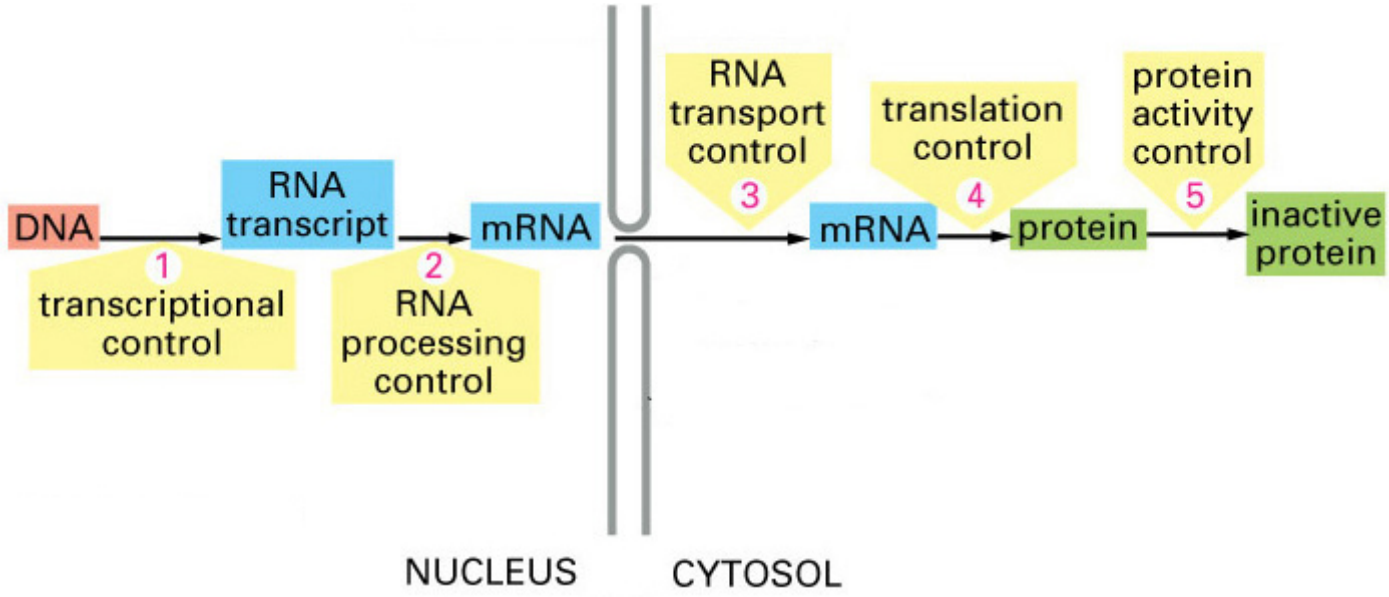
Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein" (Miescher, 1874)

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



Computational Point of View

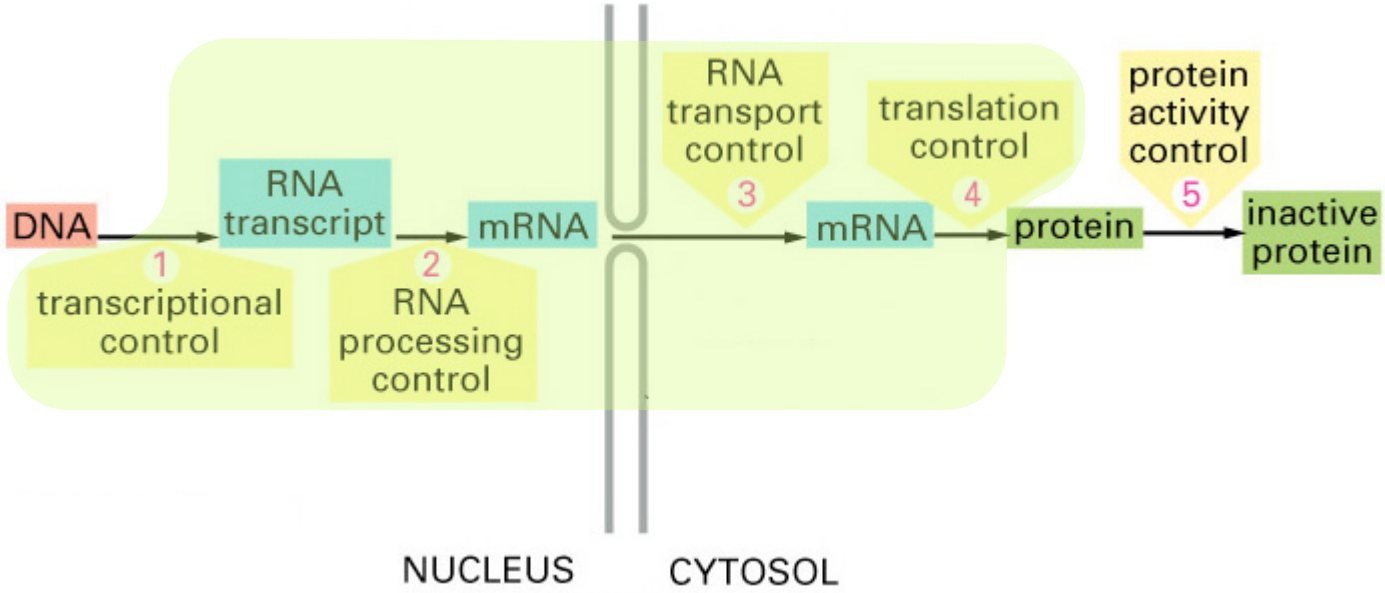
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

- What can we not predict yet? What is missing?
- Can we derive a deeper understanding of these processes?

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



Computational Point of View

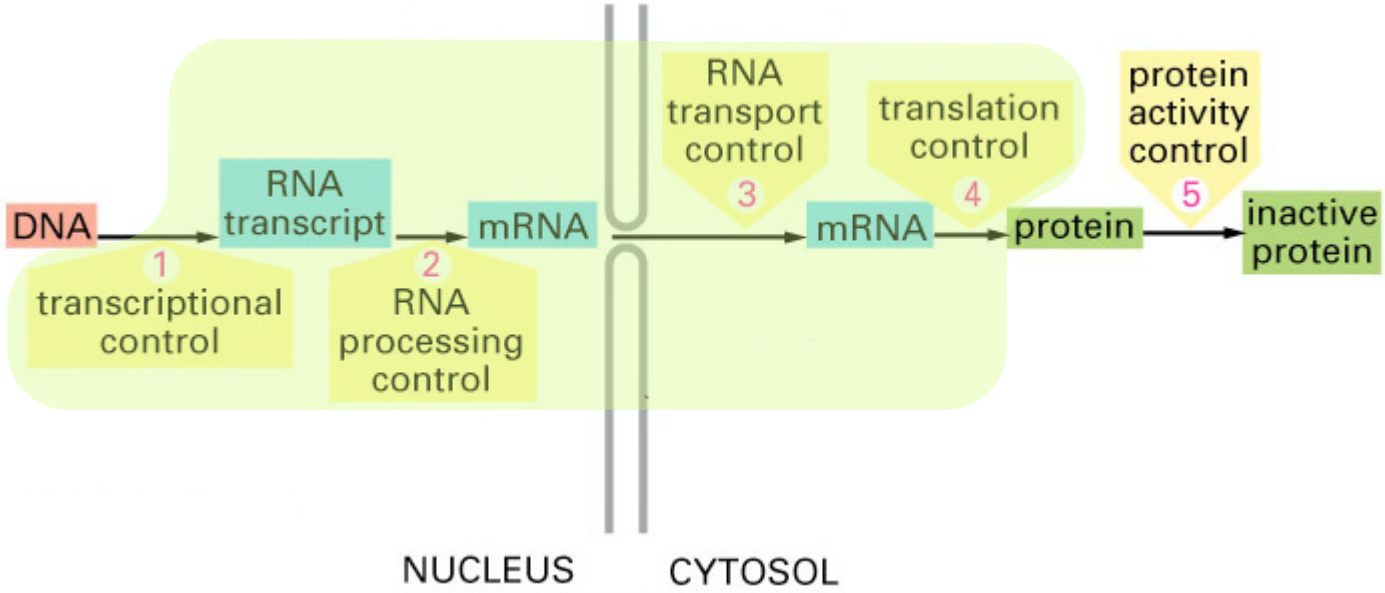
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

- What can we not predict yet? What is missing?
- Can we derive a deeper understanding of these processes?

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



Computational Point of View

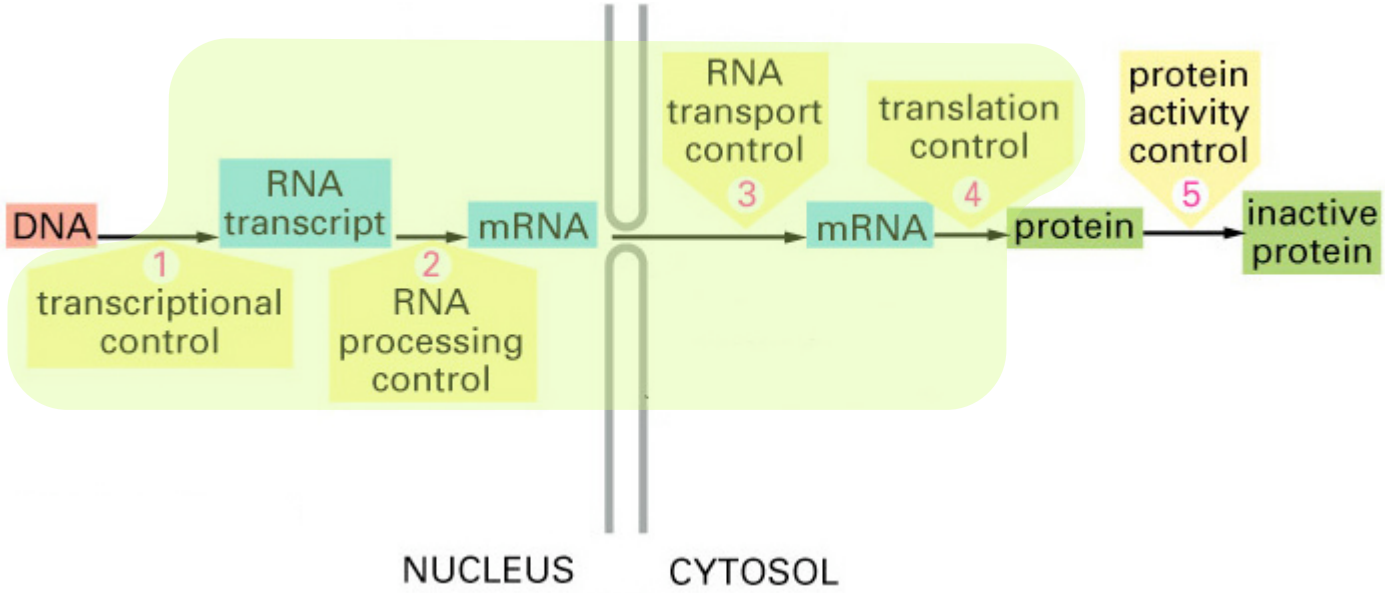
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

- What can we not predict yet? What is missing?
- Can we derive a deeper understanding of these processes?

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



Computational Point of View

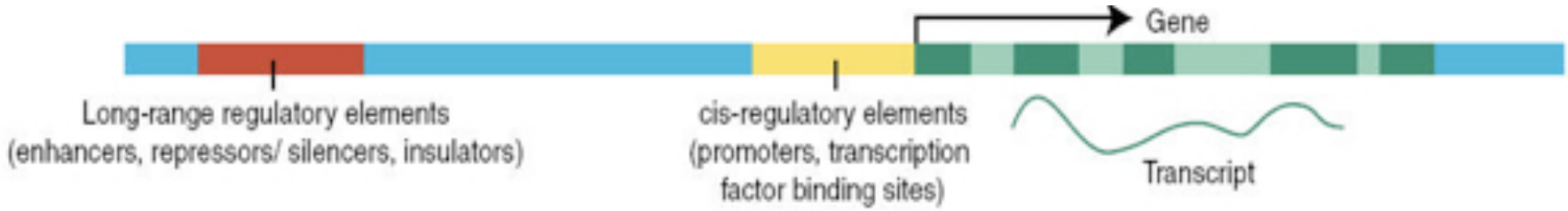
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

- What can we not predict yet? What is missing?
- Can we derive a deeper understanding of these processes?

Gene-by-Gene vs. Genome-wide Analyses

- Cloning individual genes during the last 30 years revealed important hints to the complex structure of eukaryotic genes



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

- Comprehensive understanding of gene regulation necessarily has to include the entire organism's gene collection as well as its intergenic regions.

Sequencing Genomes

- Concerted effort to sequence genomes of model organisms:
  - *E. coli* 4.5 Mb (1997)
  - *S. cerevisiae* 6.0 Mb (1997)
  - *C. elegans* 98 Mb (1998)
  - Human 3 Gb (2000)
    - Estimated cost: $2.7 billion in 1991 dollars
    - Estimated time in 1990: 15 years
- In 2003, NHGRI committed to develop next-generation sequencing technologies to lower the cost of 30x a human genome (∼100 Gbp):
  - $100,000 genome
  - $1,000 genome

Sequencing Genomes

- Concerted effort to sequence genomes of model organisms:
  - *E. coli* 4.5 Mb (1997)
  - *S. cerevisiae* 6.0 Mb (1997)
  - *C. elegans* 98 Mb (1998)
  - Human 3 Gb (2000)
    - Estimated cost: $2.7 billion in 1991 dollars
    - Estimated time in 1990: 15 years
- In 2003, NHGRI committed to develop next-generation sequencing technologies to lower the cost of 30x a human genome (∼100 Gbp):
  - $100,000 genome
  - $1,000 genome

New Sequencing Technologies





SoLID (2007)



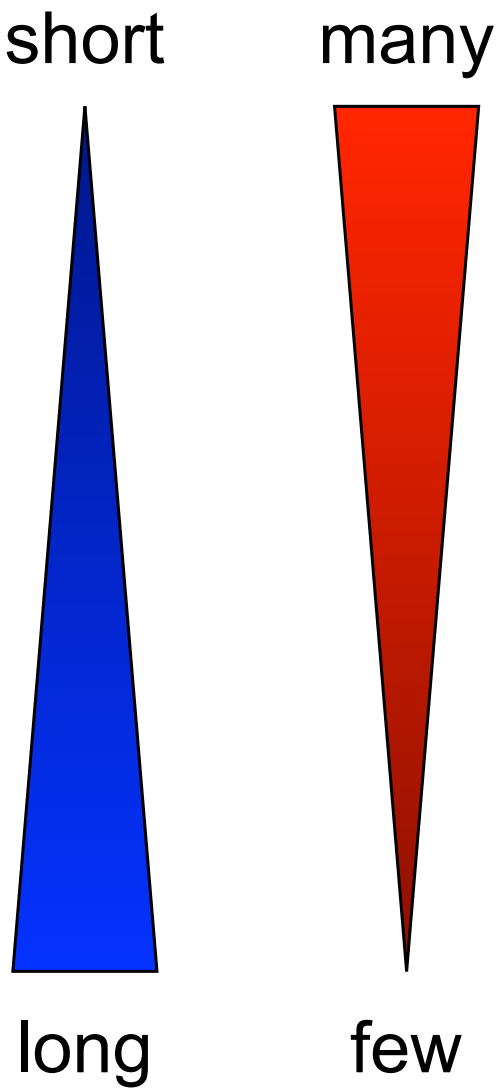
Genome Analyzer (2007)



454 Sequencing (2005)



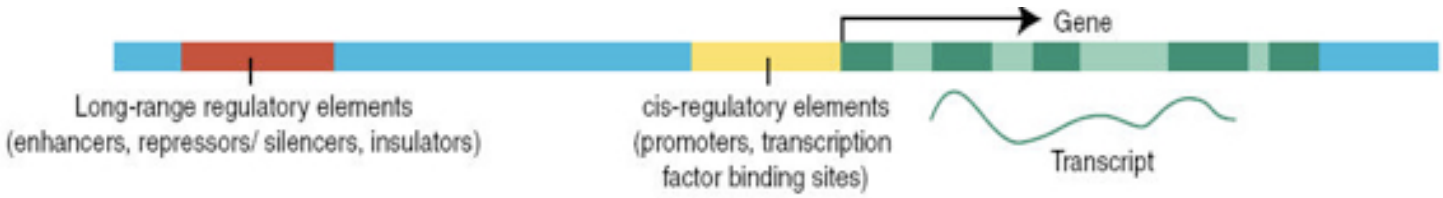
SMRT (2010)



[Slide provided by Ali Mortazavi]

The Encyclopedia of DNA Elements

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

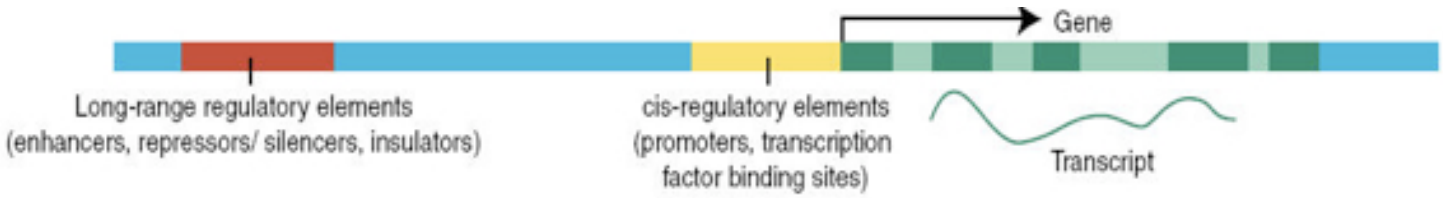


- National Human Genome Research Institute started the Encode project to annotate human and model organism genomes:
  - 2004: Human 1%
  - 2007: Human whole-genome
  - 2008: *Drosophila* and *C. elegans*
  - 2010: Mouse

The Encyclopedia of DNA Elements

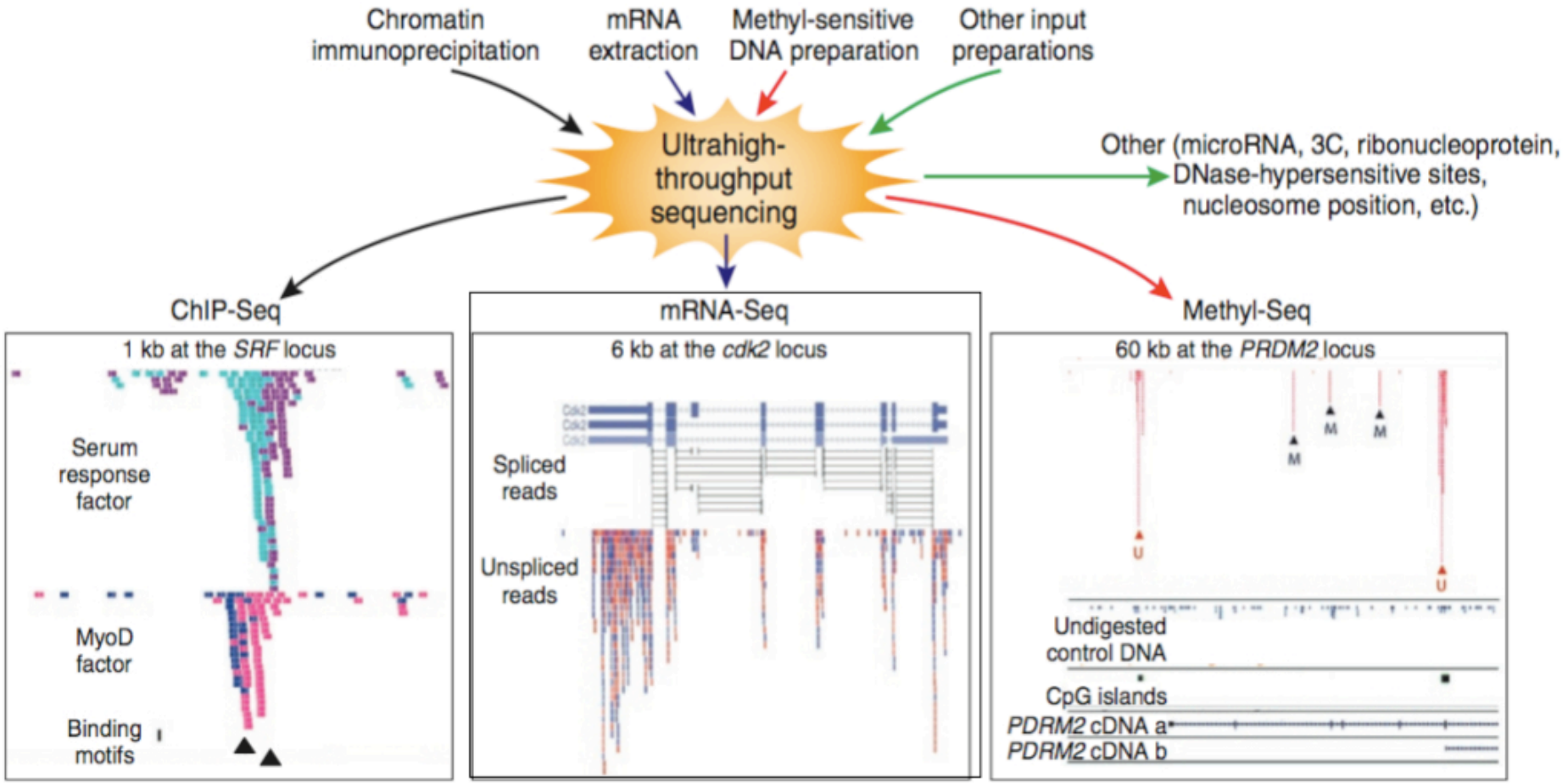


- How to make sense of the genome?
  - Which nucleotides are functional?
  - What is their function?

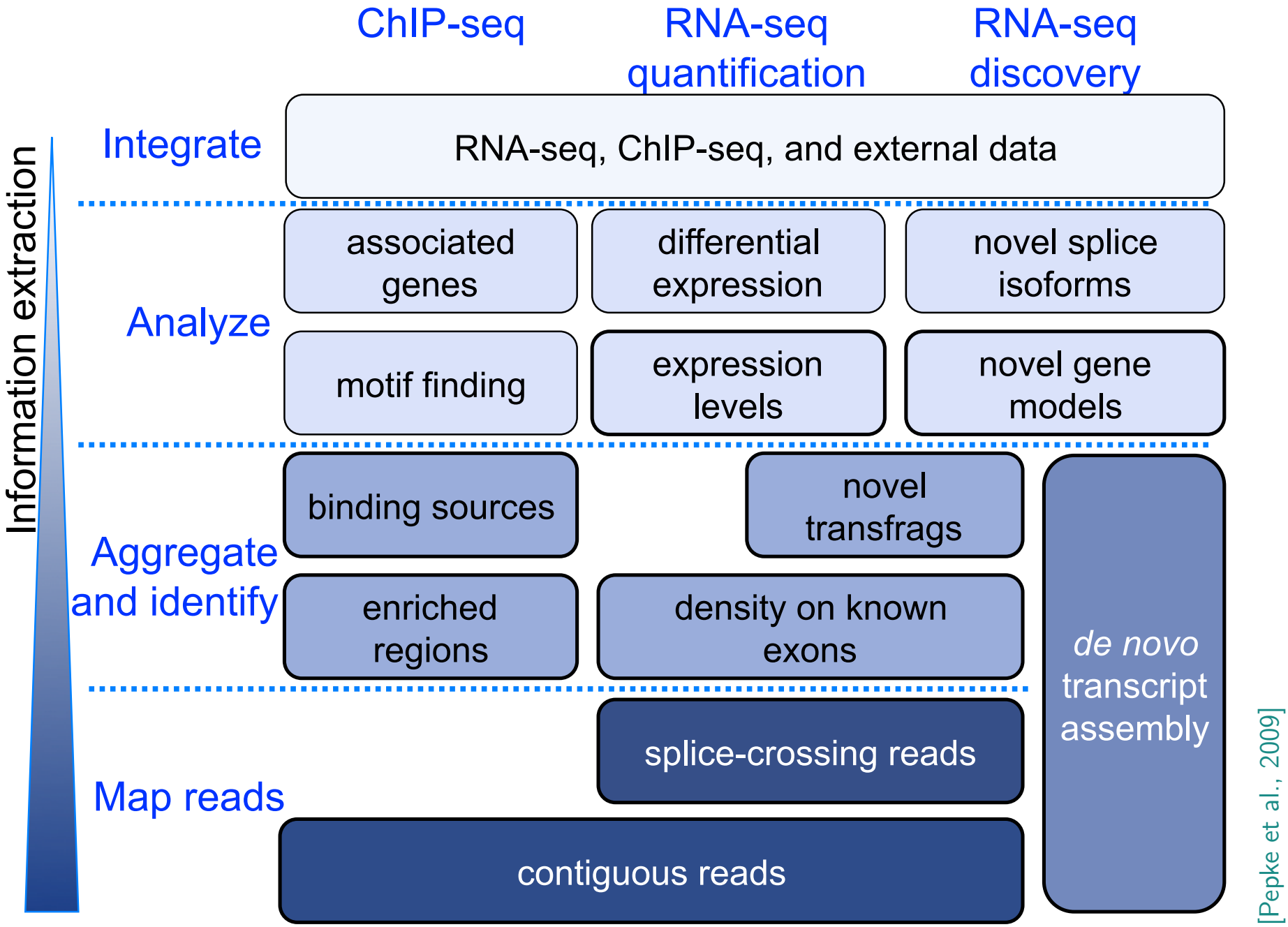


- National Human Genome Research Institute started the Encode project to annotate human and model organism genomes:
  - 2004: Human 1%
  - 2007: Human whole-genome
  - 2008: *Drosophila* and *C. elegans*
  - 2010: Mouse

Sequencing for Functional Genomics



[Wold & Myers, Nature Methods, 2007]


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]

Deep RNA Sequencing (RNA-Seq)

RNA-Seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
- Quantitative studies at high resolution

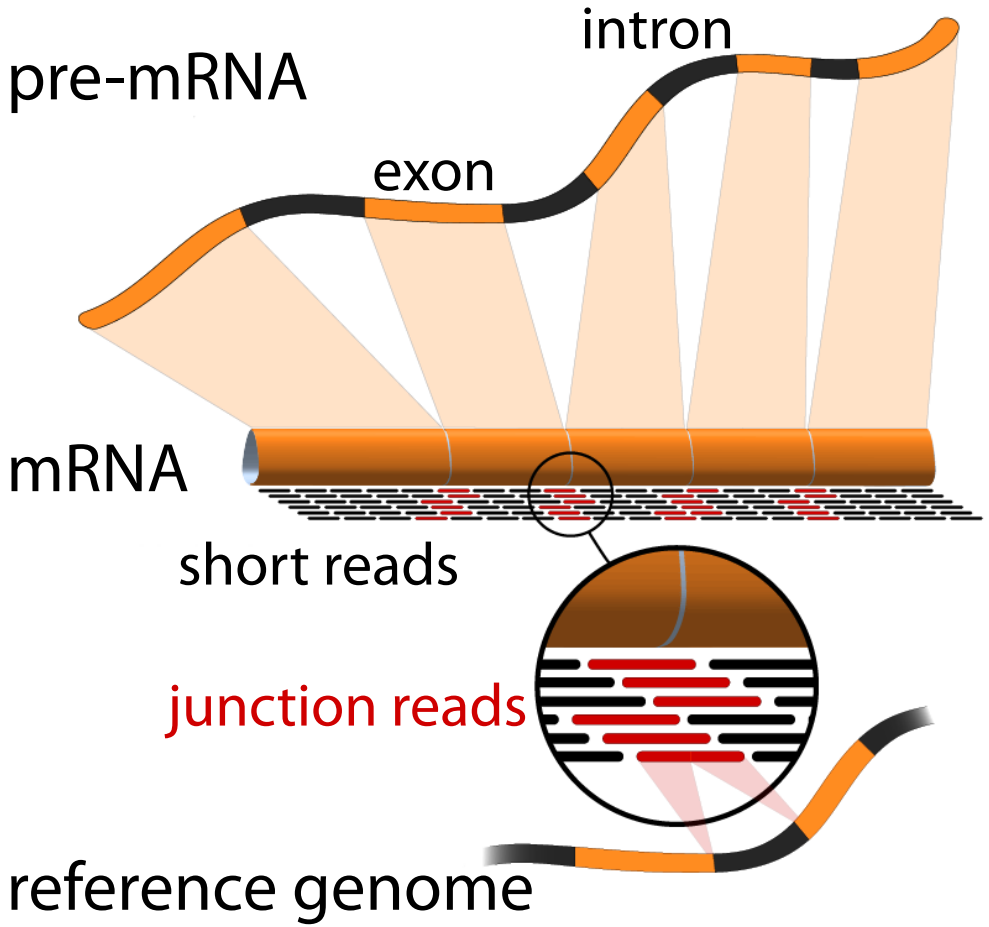


Figure adapted from Wikipedia

Goal: Obtain complete transcriptome for further analyses

Deep RNA Sequencing (RNA-Seq)

RNA-Seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
- Quantitative studies at high resolution

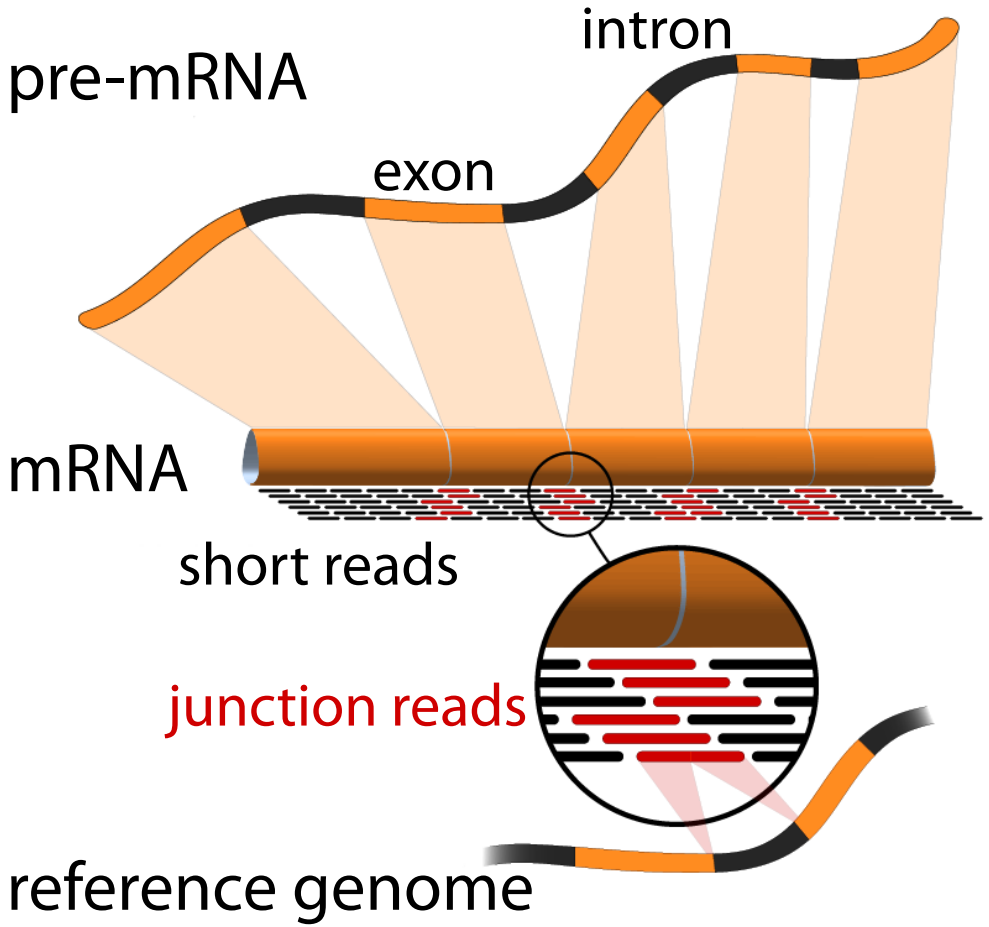
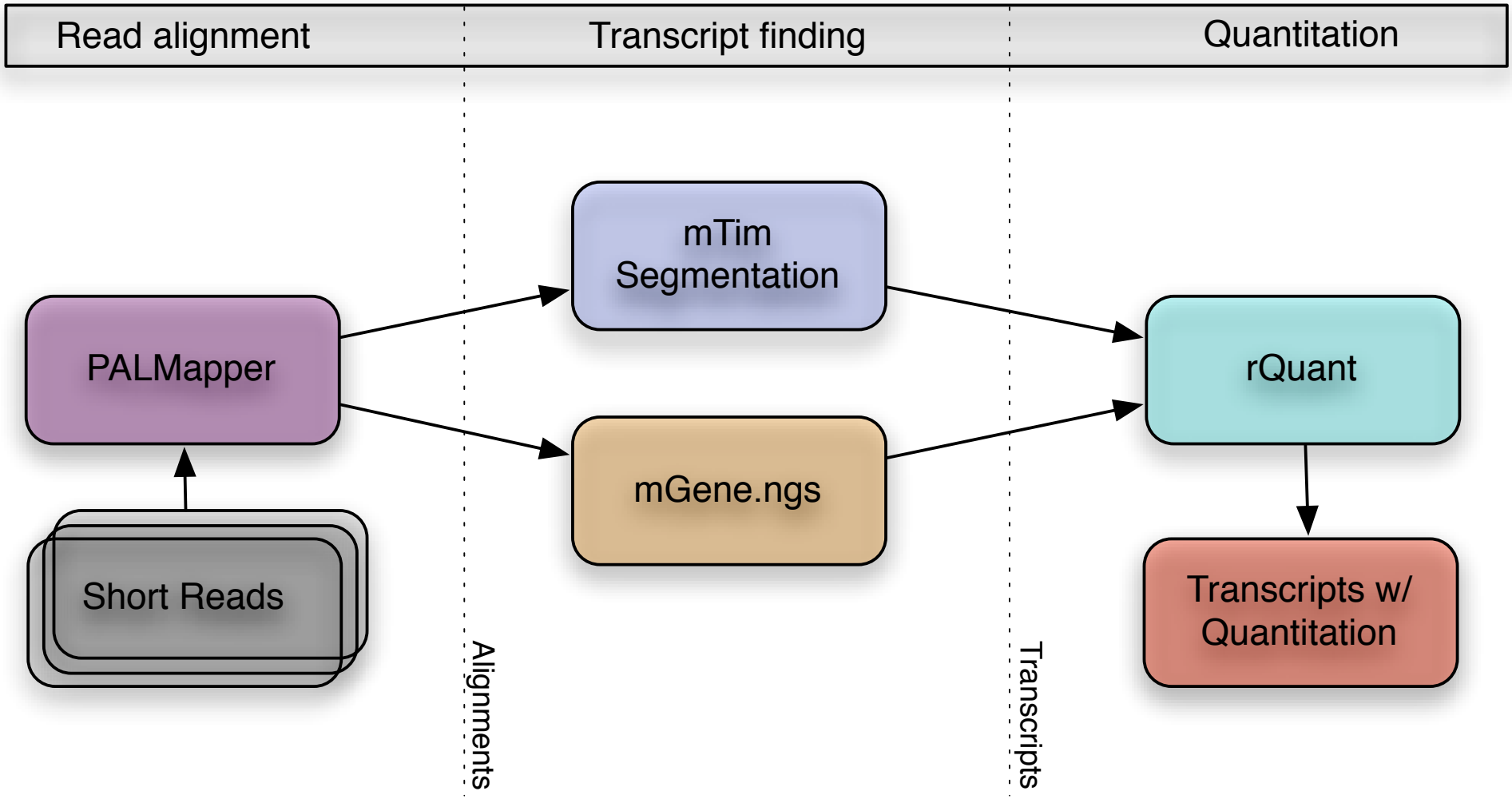


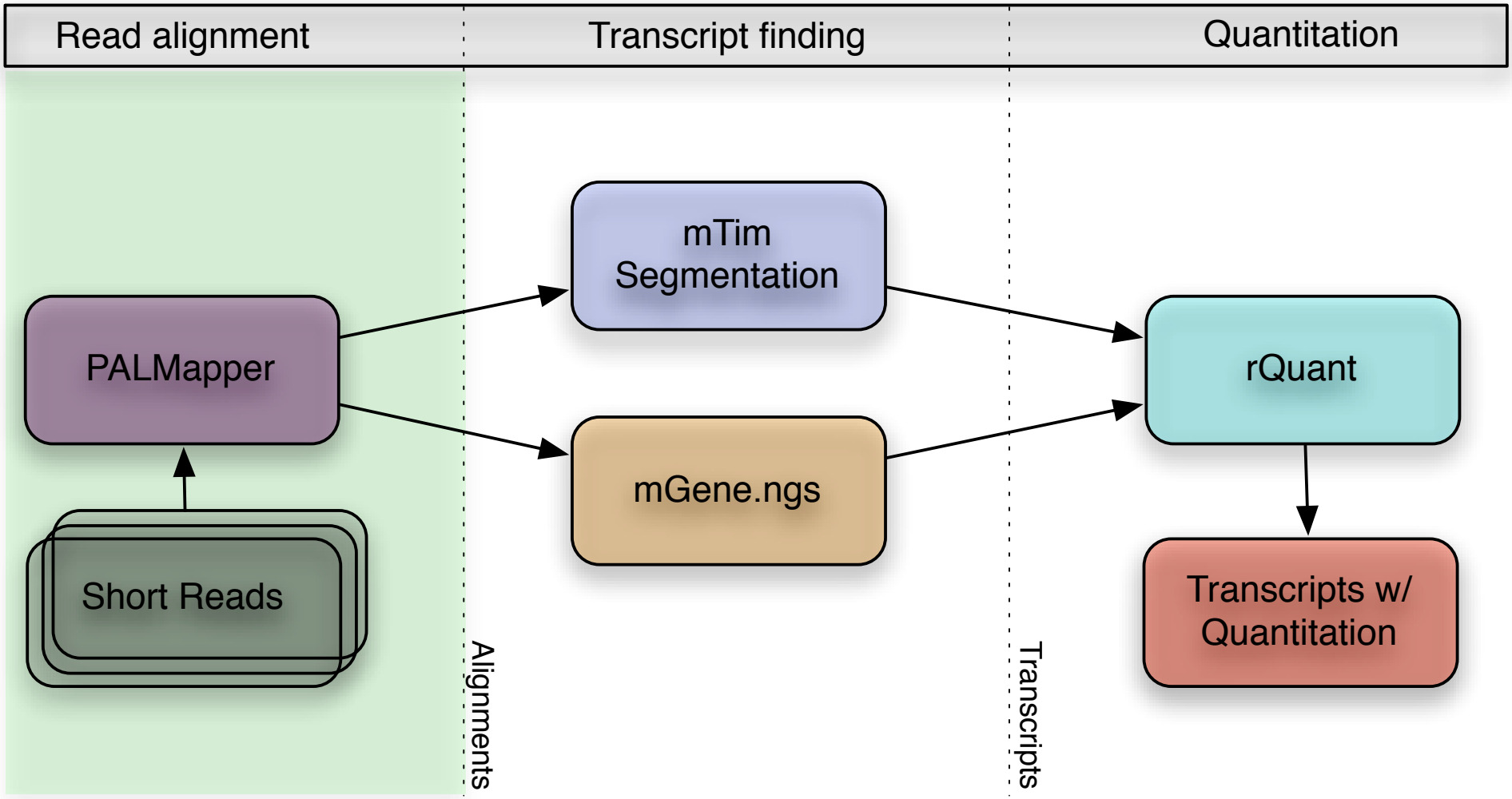
Figure adapted from Wikipedia

Goal: Obtain complete transcriptome for further analyses

RNA-Seq Pipeline Overview



RNA-Seq Pipeline Overview



Step 1: PALMapper Read Alignment

(PALMapper = QPALMA + GenomeMapper)



GenomeMapper for (unspliced) read mapping:

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

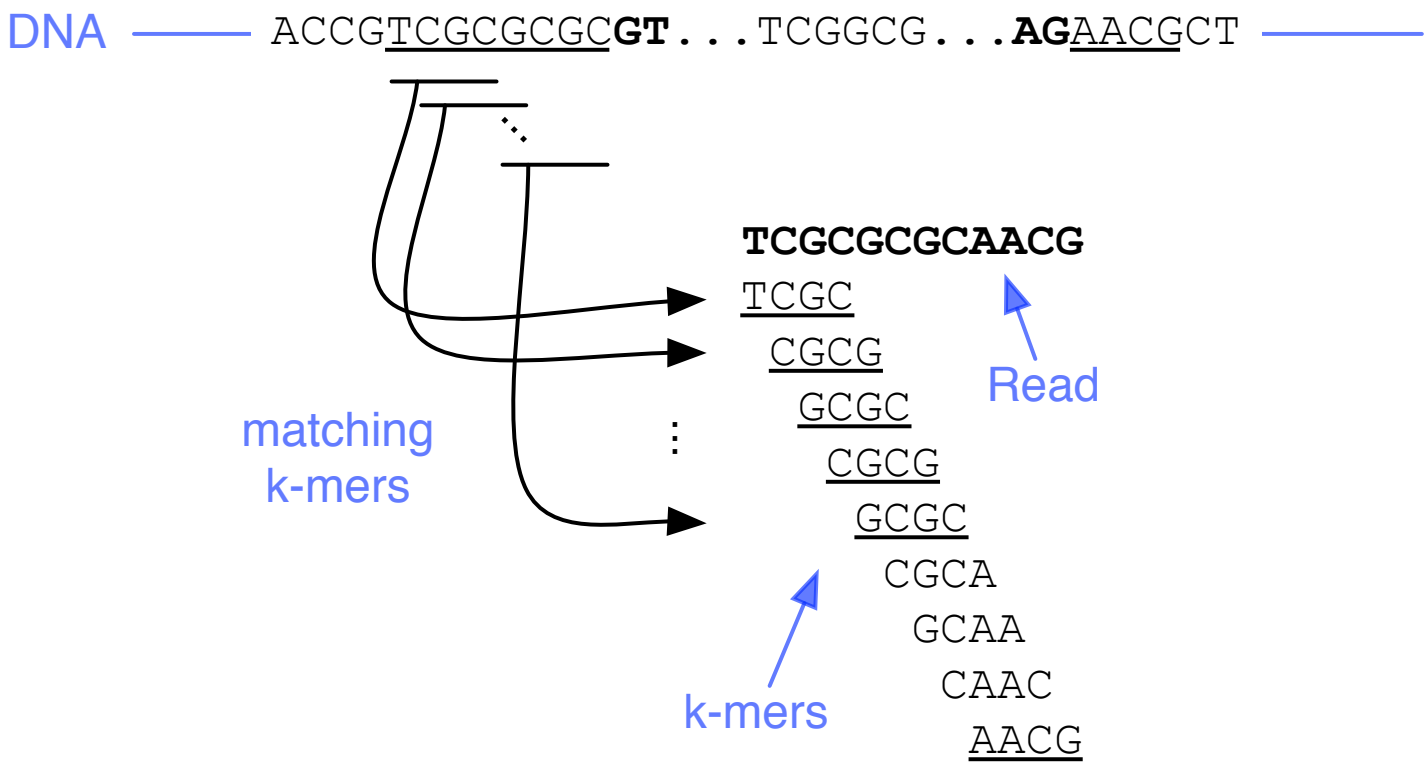
More info: http://fml.mpg.de/raetsch/suppl/palmapper

Step 1: PALMapper Read Alignment

(PALMapper = QPALMA + GenomeMapper)

GenomeMapper for (unspliced) read mapping:

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

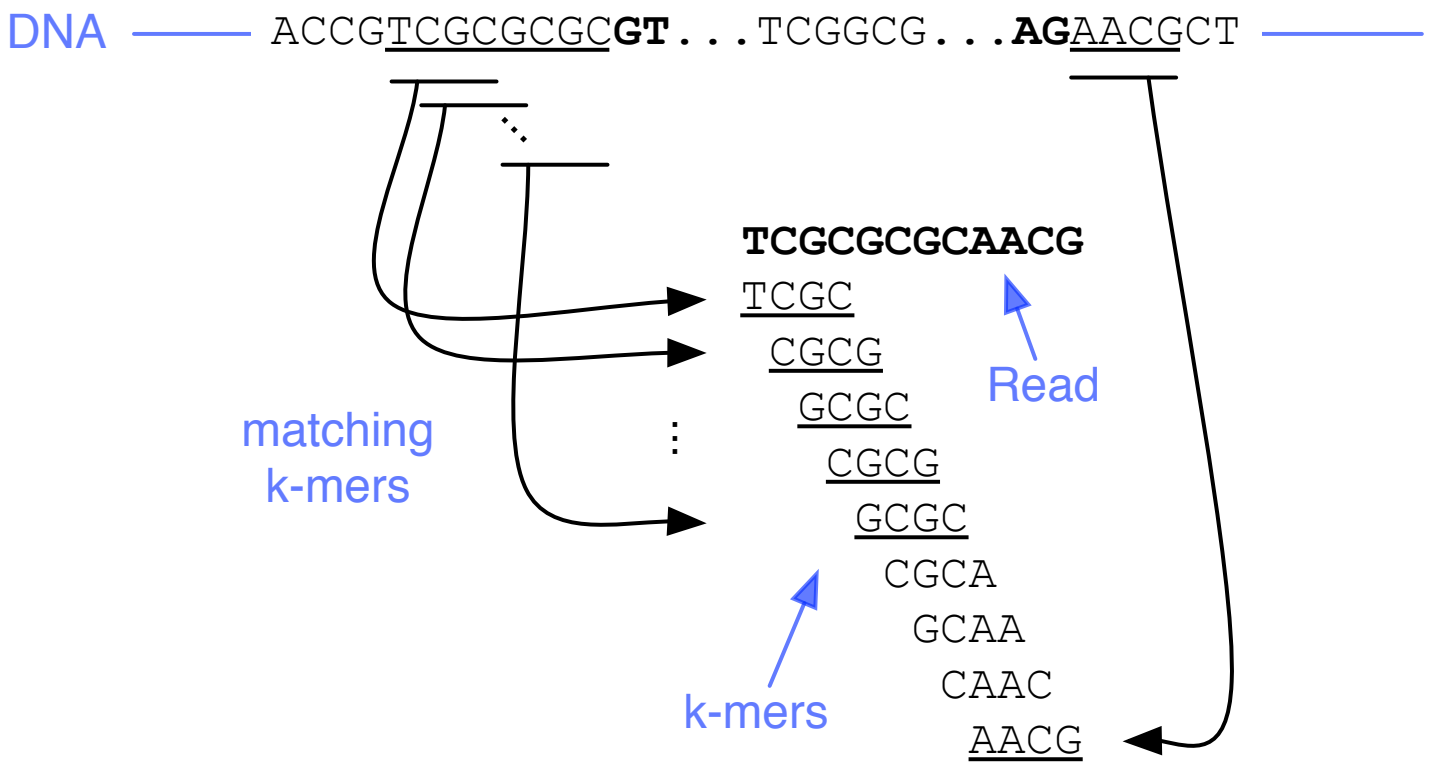


Step 1: PALMapper Read Alignment

(PALMapper = QPALMA + GenomeMapper)

GenomeMapper for (unspliced) read mapping:

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



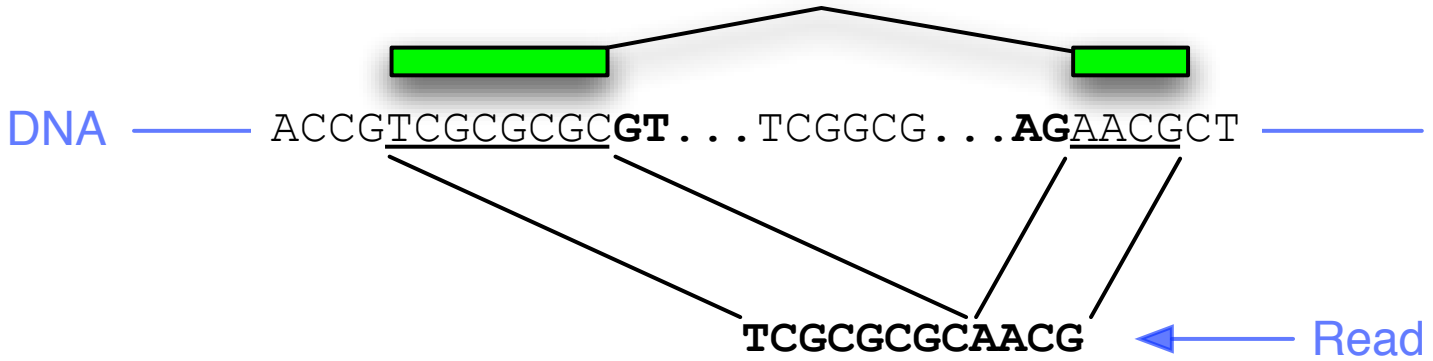
Step 1: PALMapper Read Alignment
(PALMapper = QPALMA + GenomeMapper)

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *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]



More info: `http://fml.mpg.de/raetsch/suppl/palmapper`

QPALMA: Extended Smith-Waterman Scoring



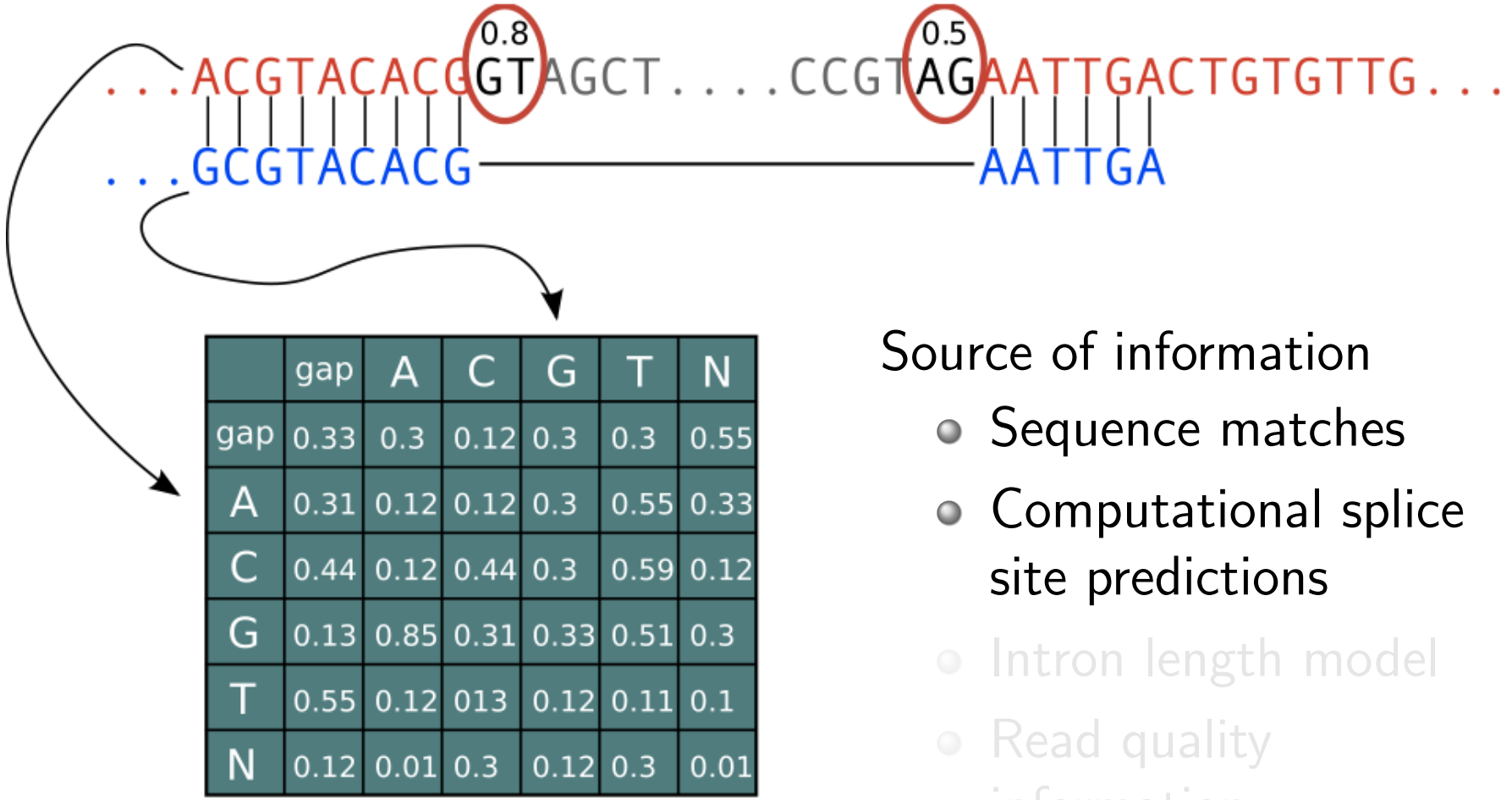
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

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 Scoring



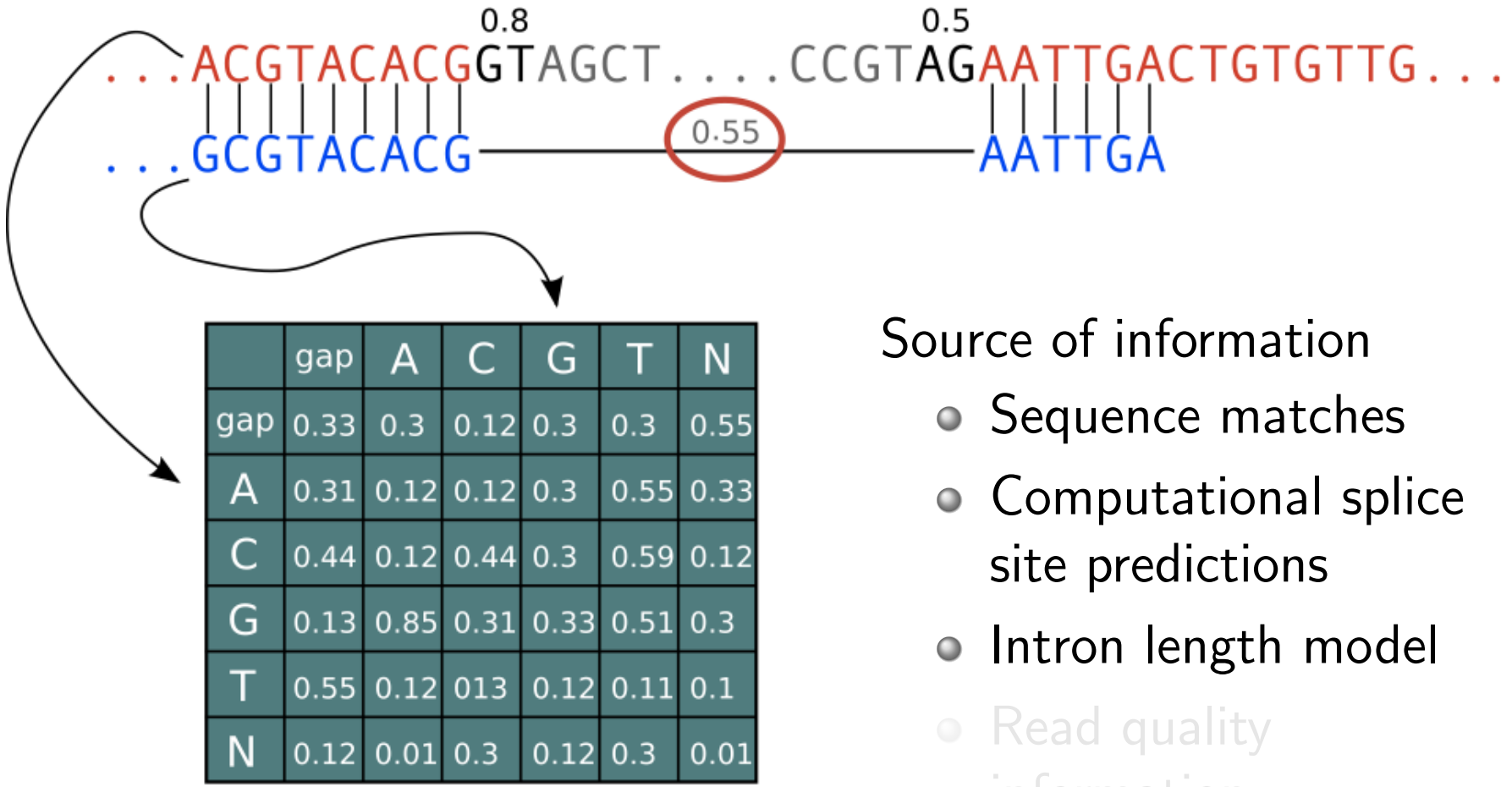
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

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 Scoring



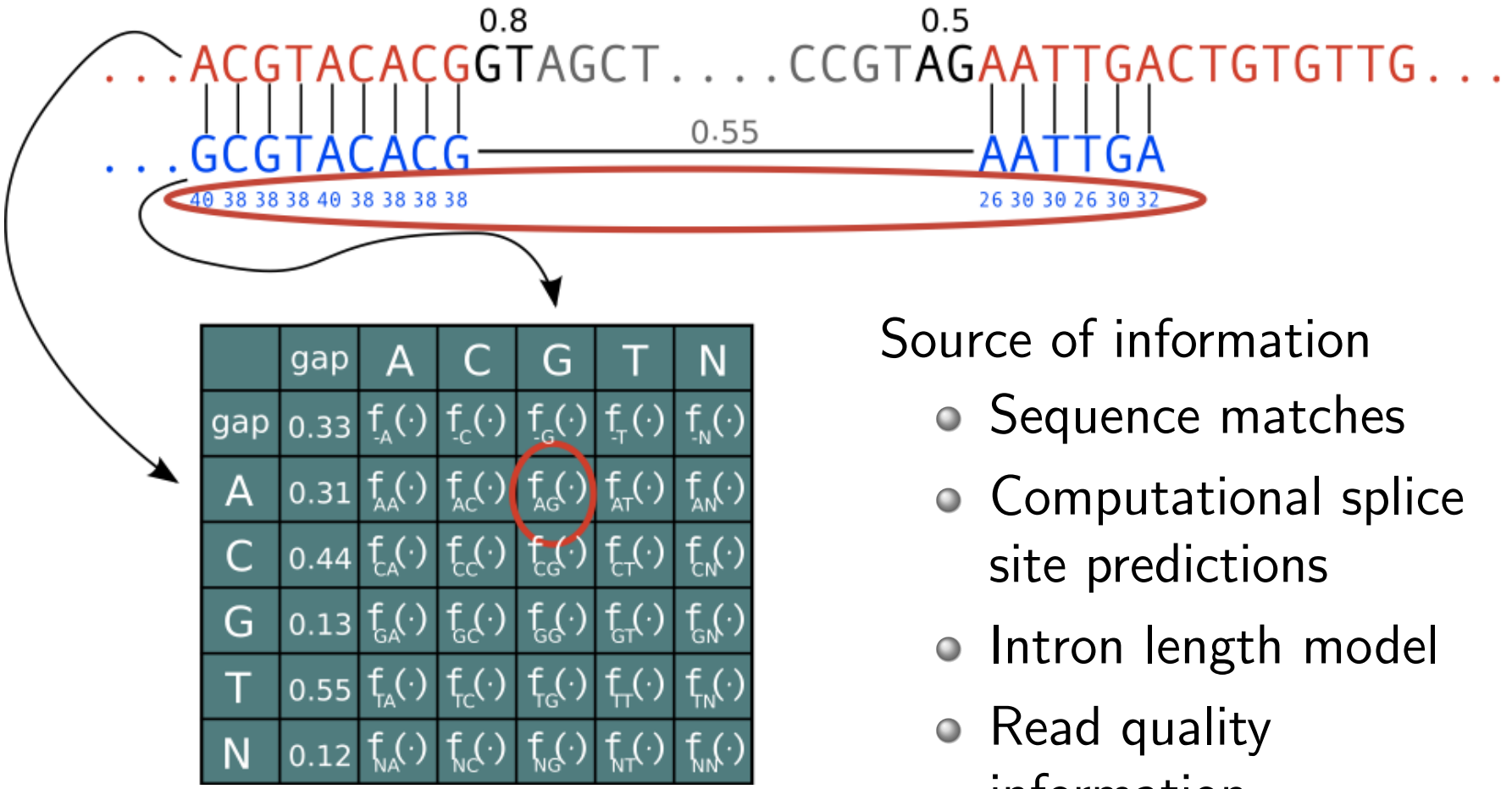
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

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 Scoring



Source of information

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

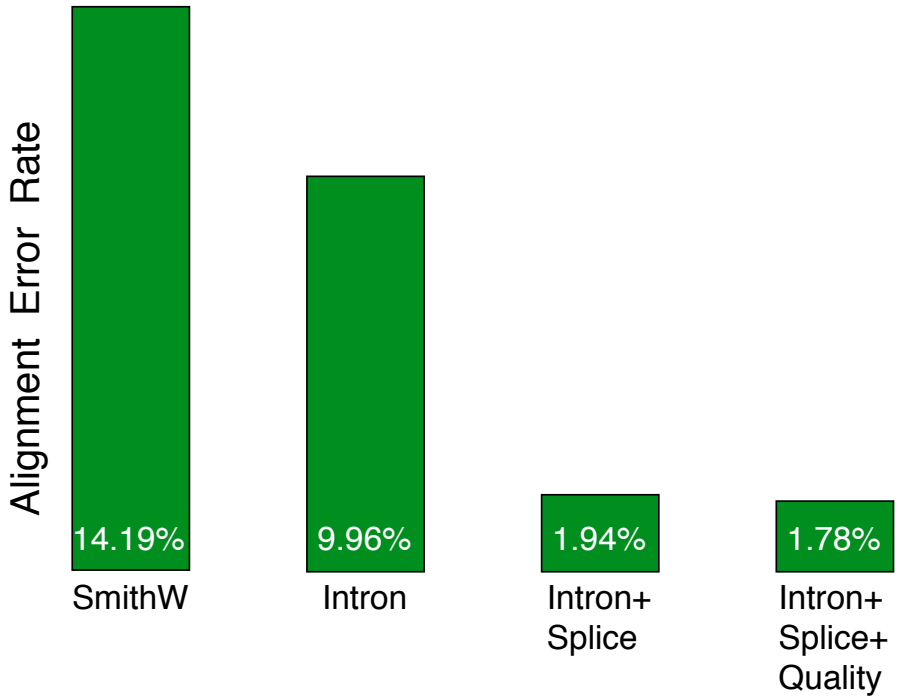
Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

[De Bona et al., 2008]

QPALMA RNA-Seq Read Alignment

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

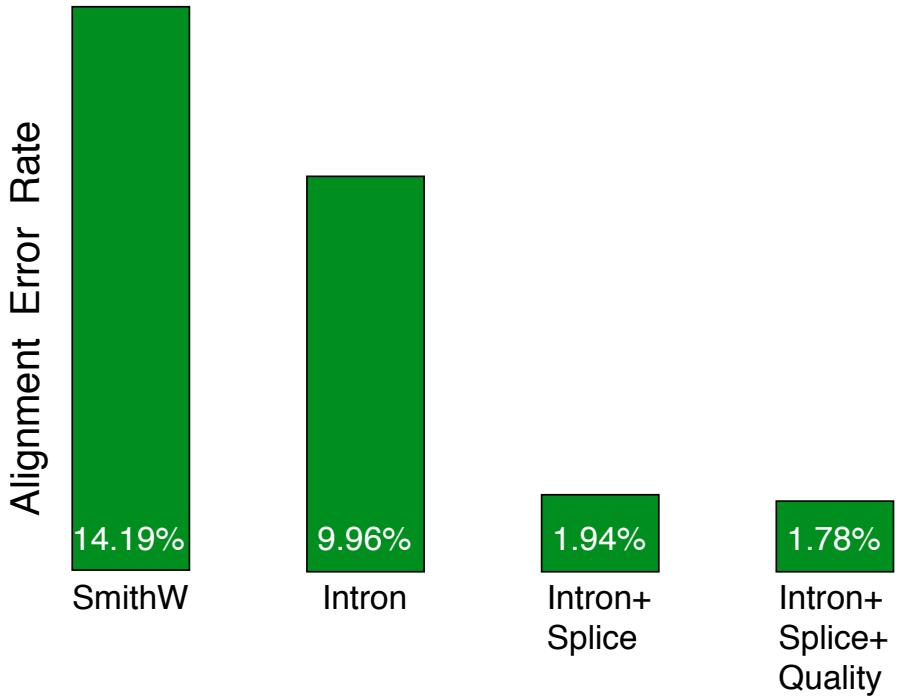


[De Bona et al., 2008]

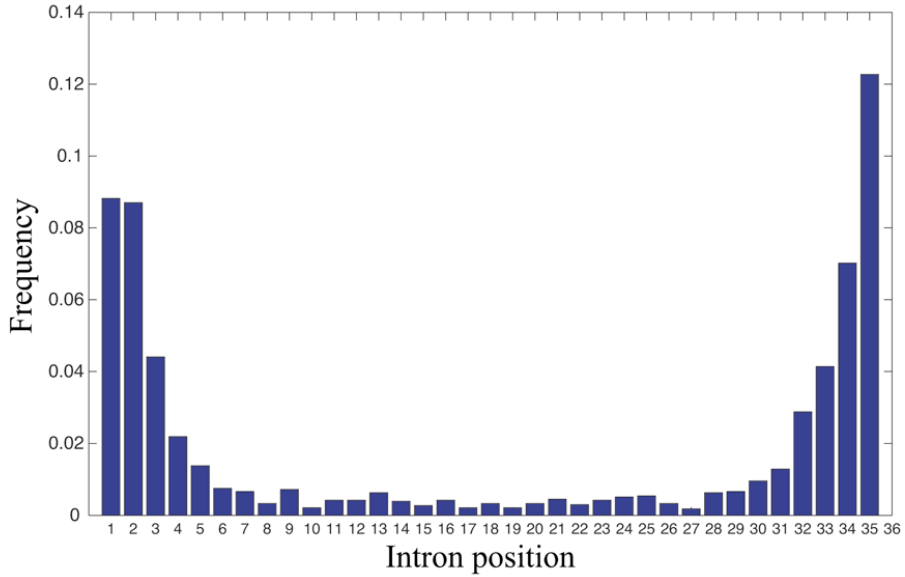
QPALMA RNA-Seq Read Alignment

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

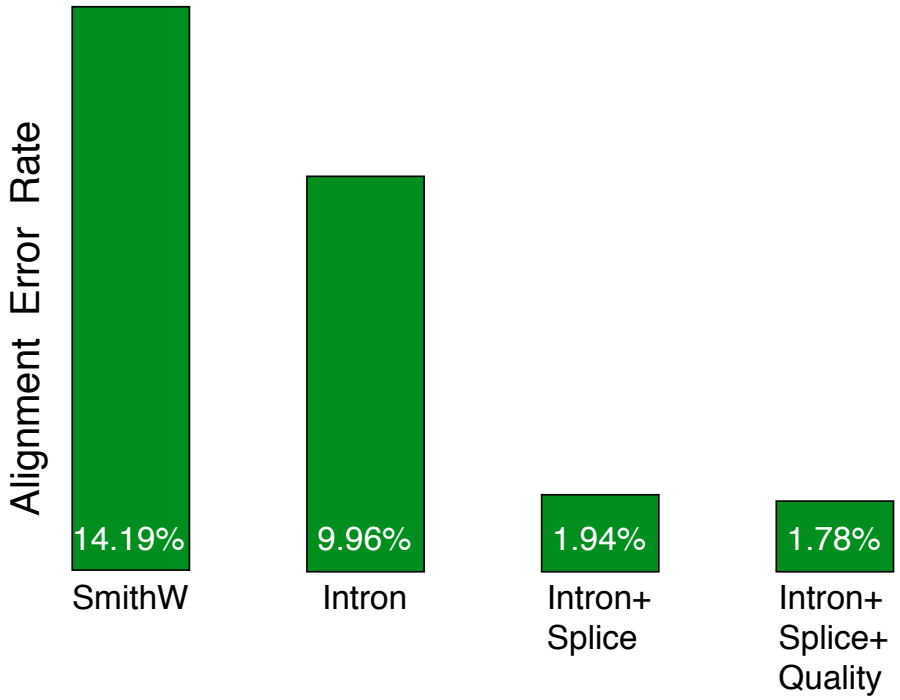


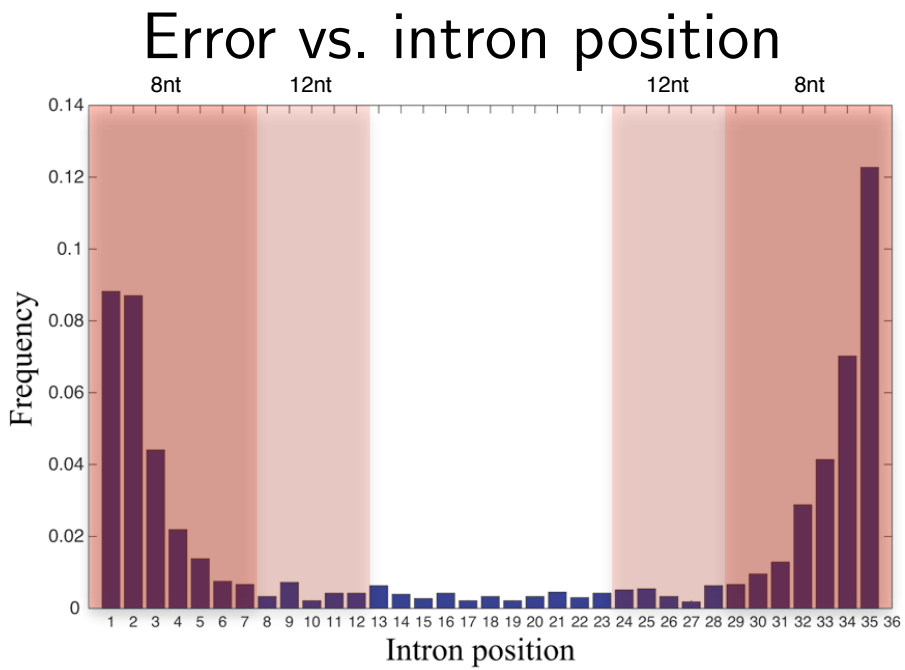
[De Bona et al., 2008]

QPALMA RNA-Seq Read Alignment

Generate set of artificially spliced reads

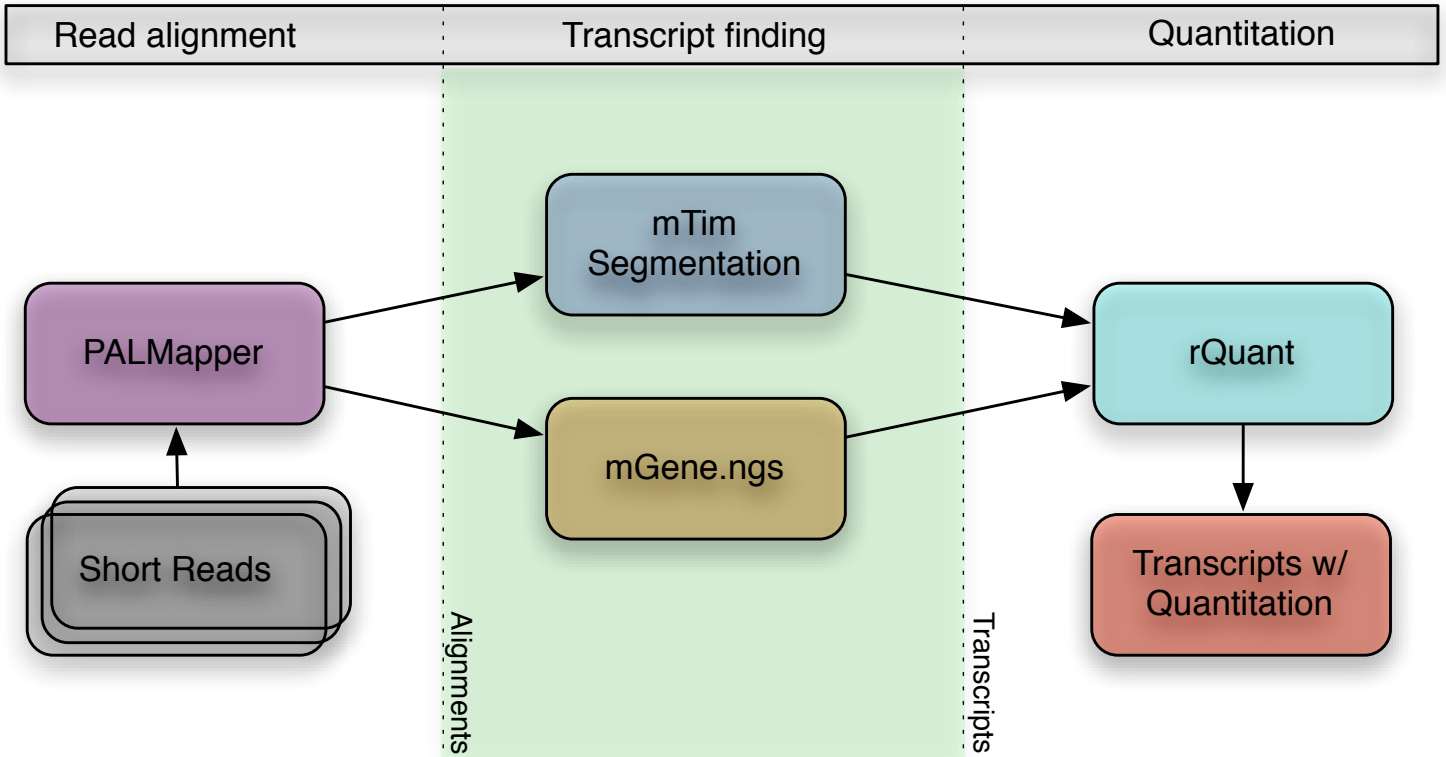
- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing





[De Bona et al., 2008]

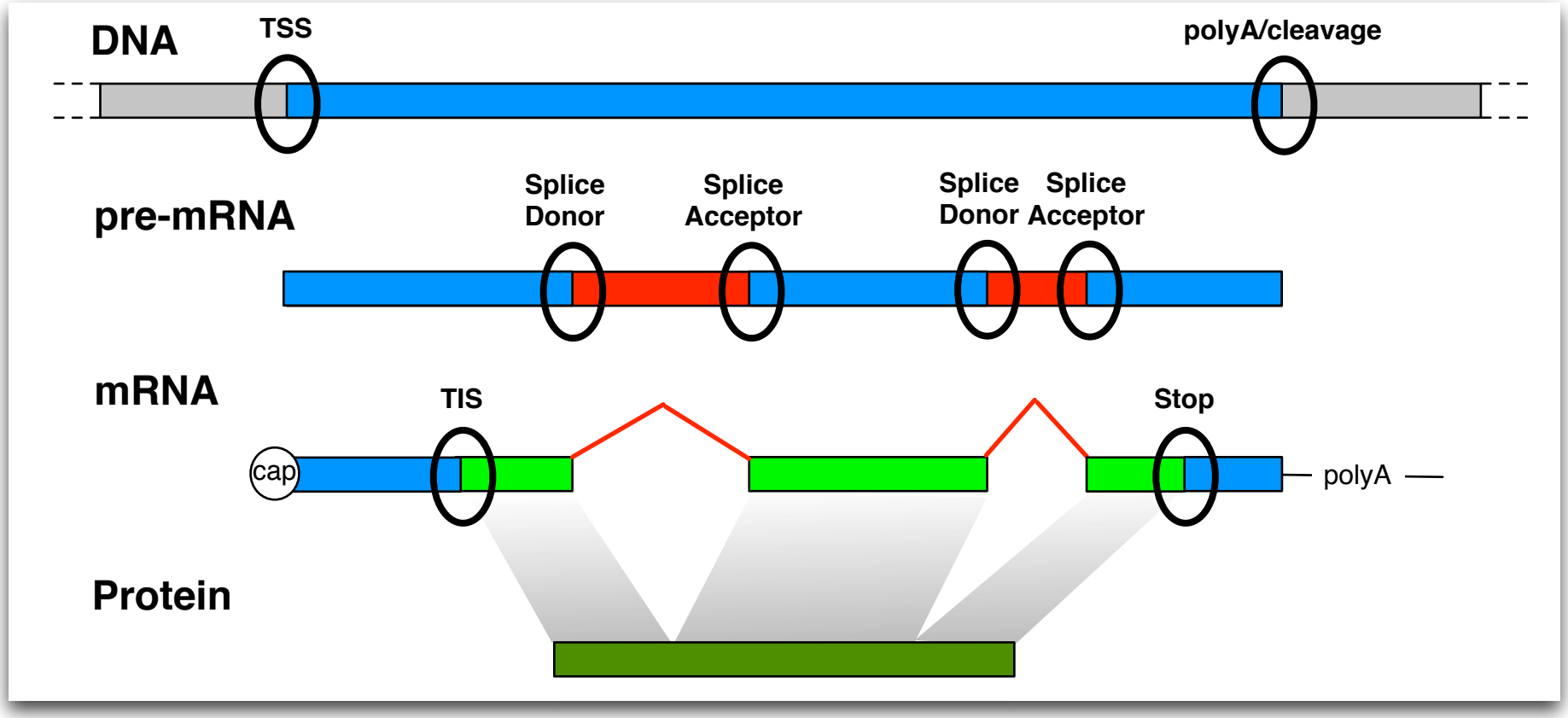
Step 2: Transcript Prediction



A. Coverage segmentation algorithm **mTiM** for general transcripts (no coding bias/assumption)
B. Extension of the mGene gene finding system to use NGS data for protein coding transcript prediction (**mGene.ngs**)

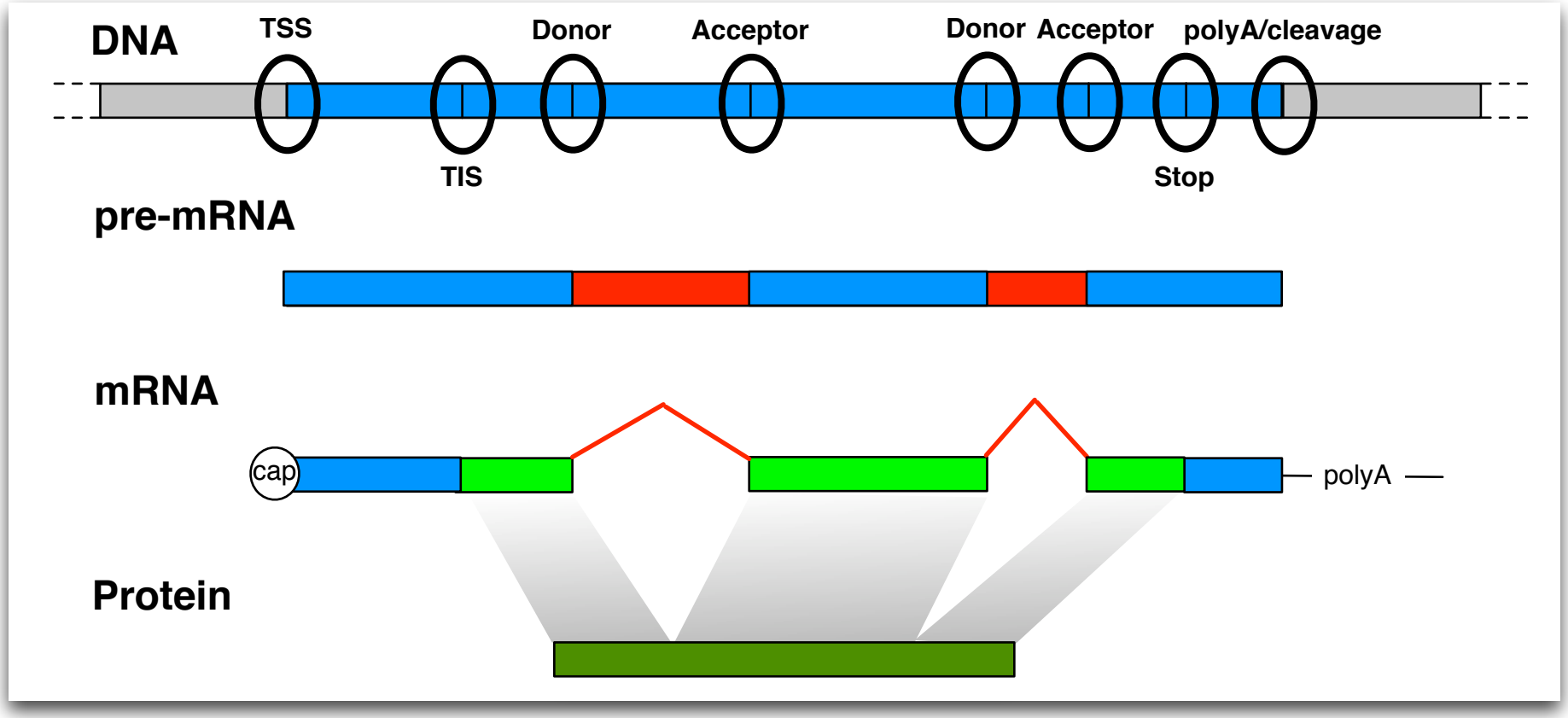
Computational Gene Finding

⇝ Labeling the Genome



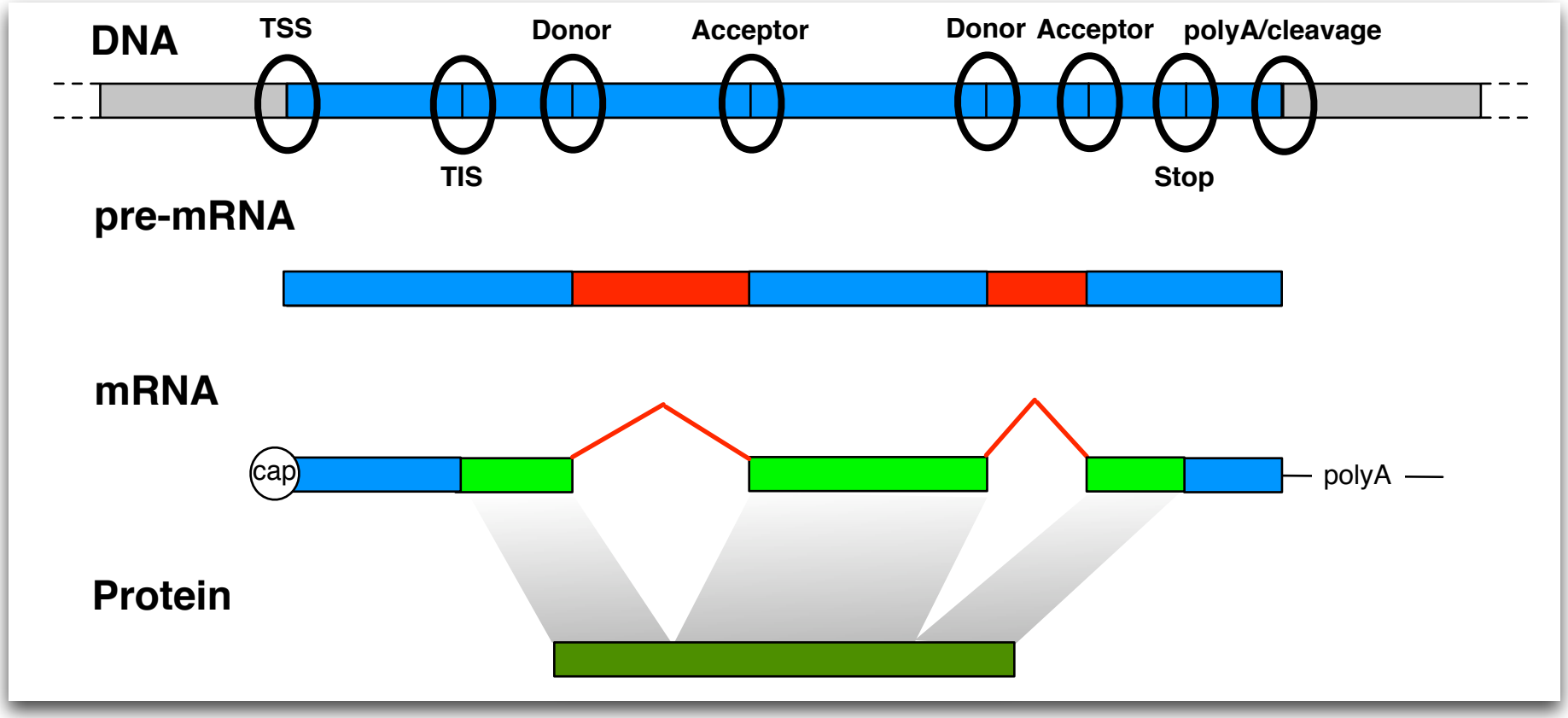
Computational Gene Finding

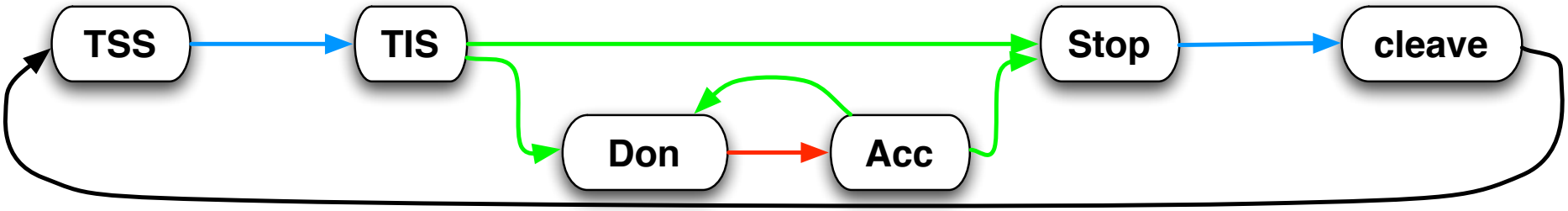
⇝ Labeling the Genome



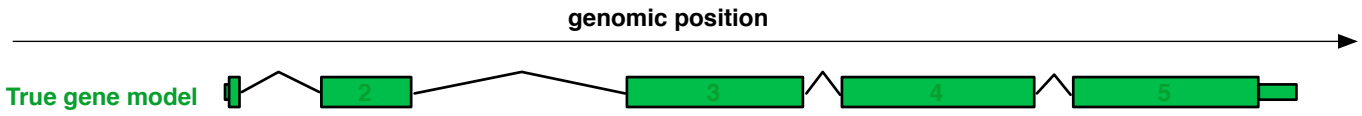
Computational Gene Finding

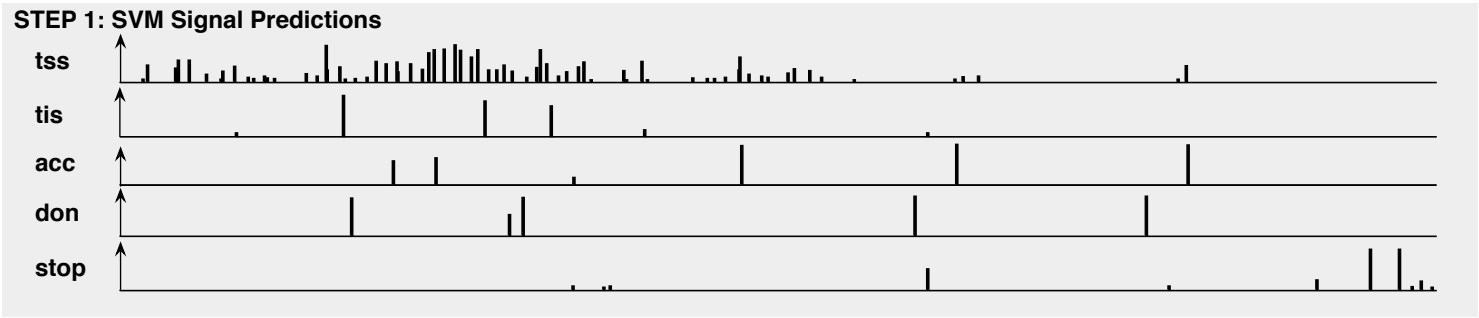
⇝ Labeling the Genome



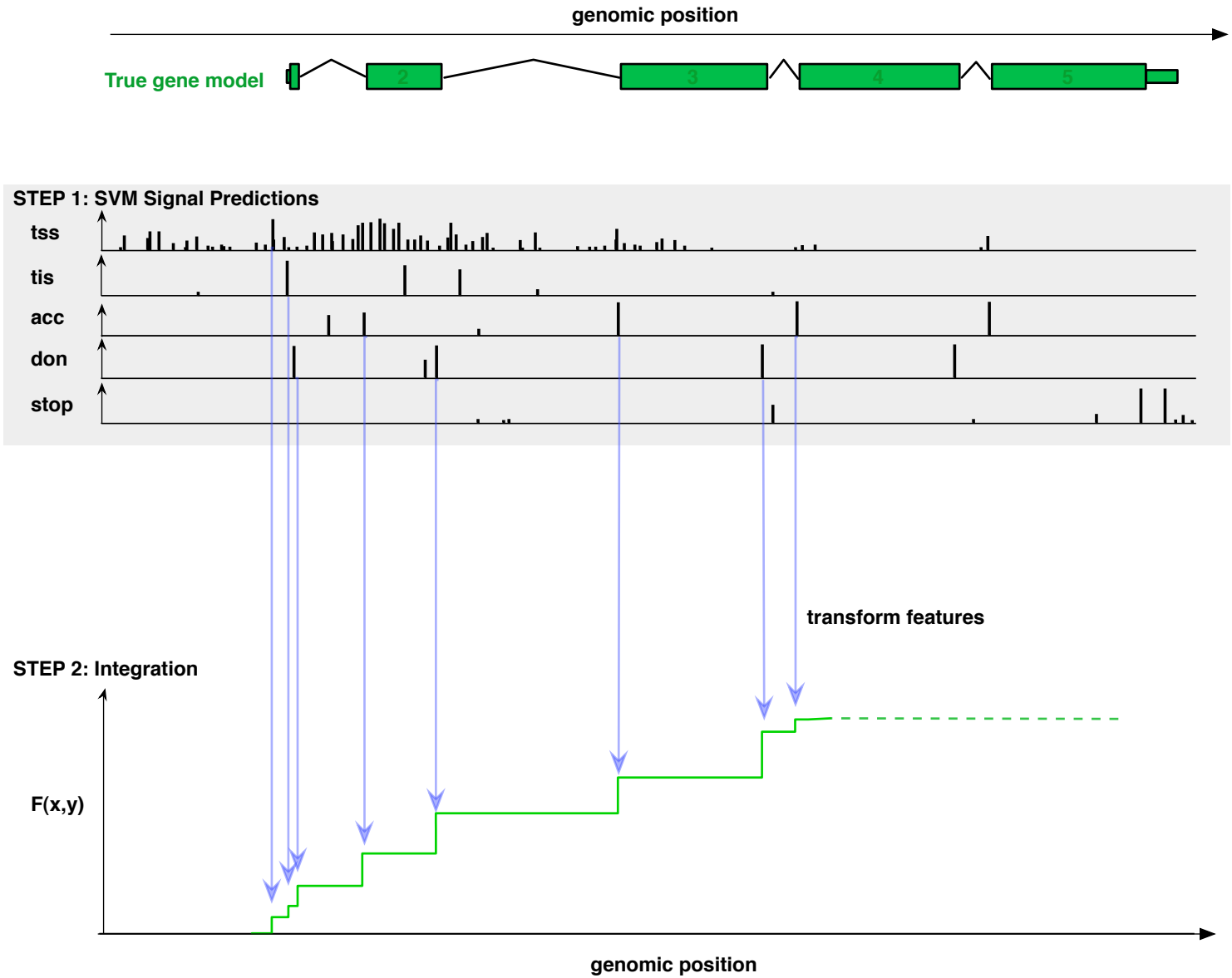


mGene-based Transcript Prediction

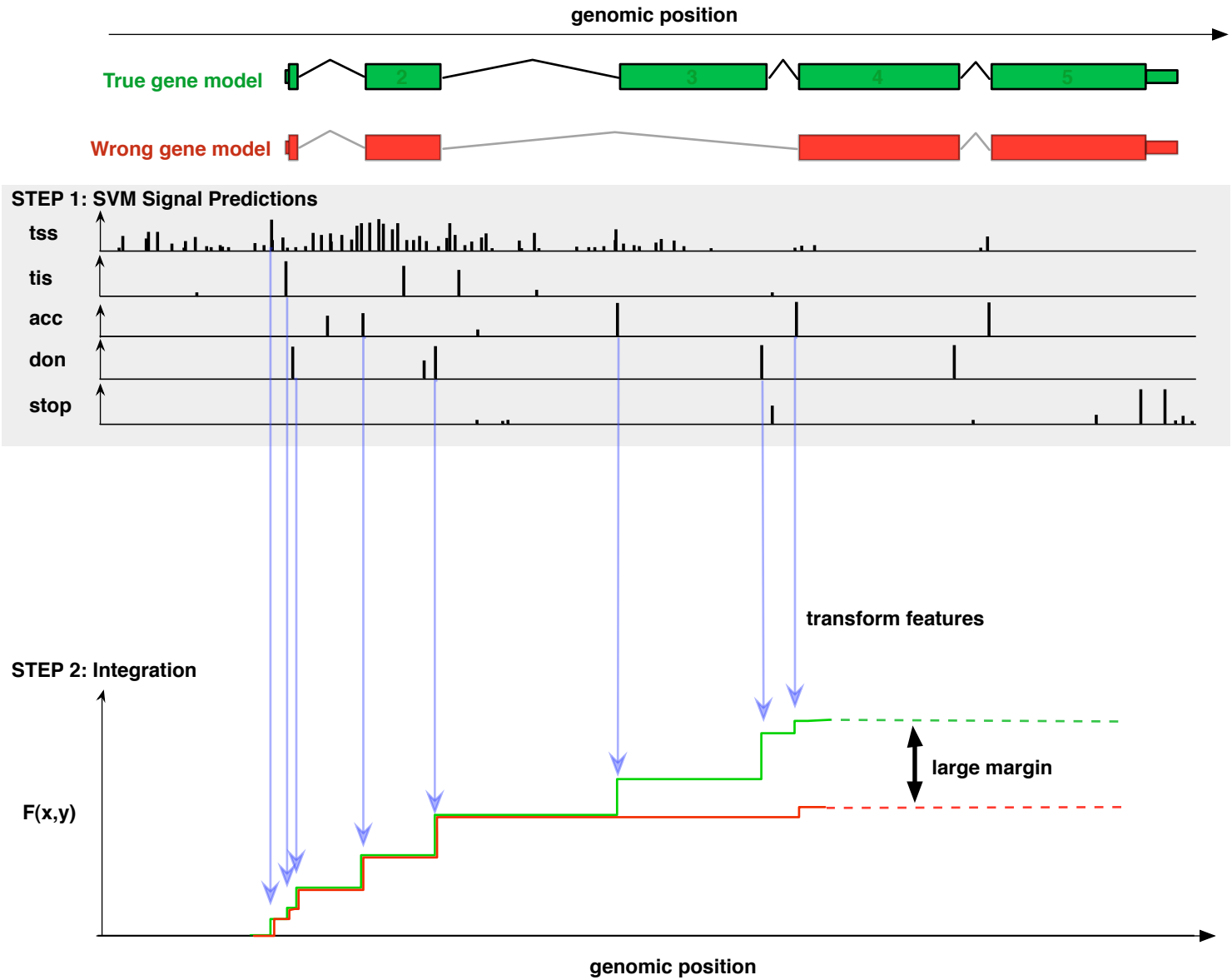




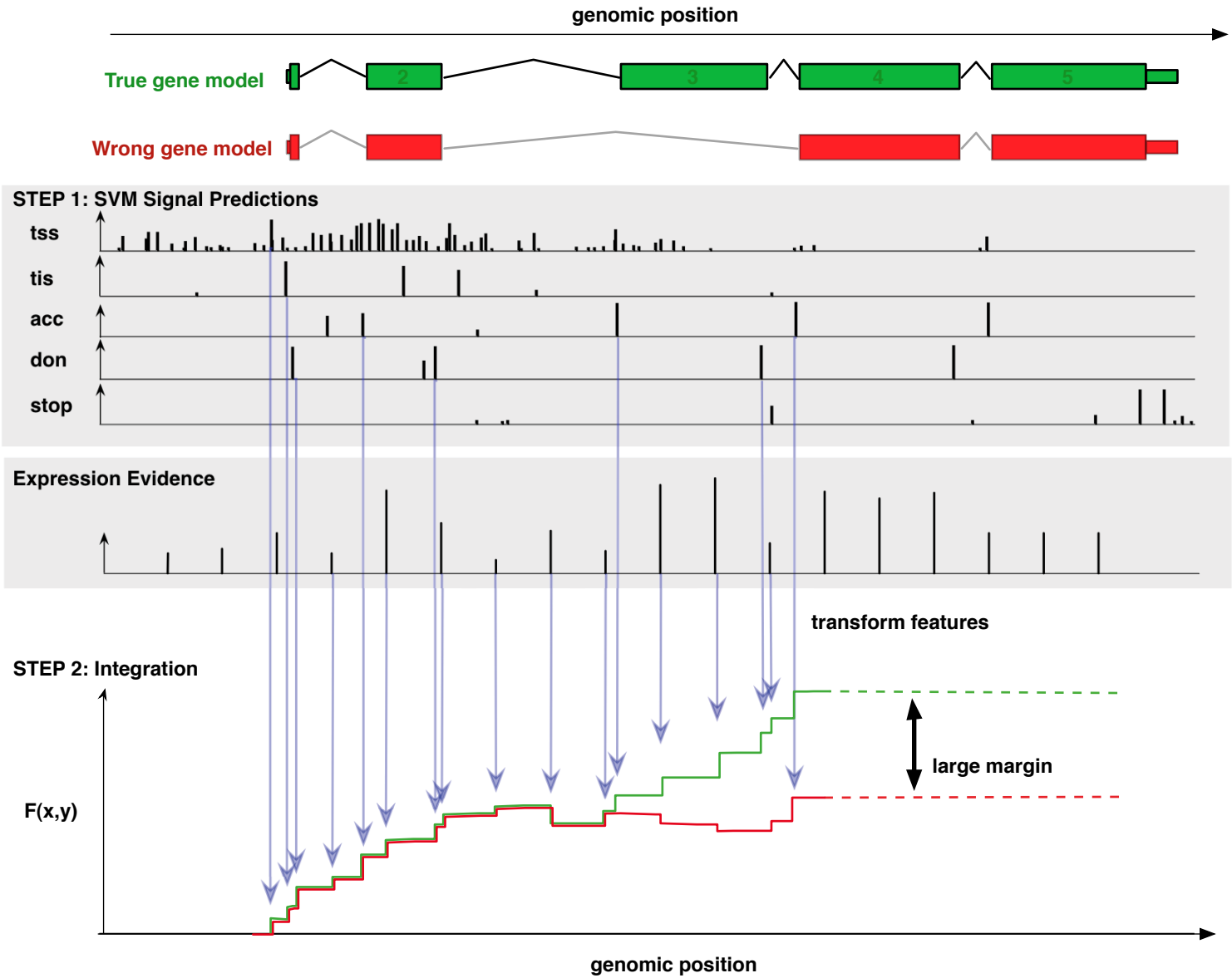
mGene-based Transcript Prediction



mGene-based Transcript Prediction



mGene-based Transcript Prediction



Learning to use Expression Measurements

Two approaches:

- Heuristic to incorporate ESTs/reads/tiling array measurements to *refine predictions*
- Directly *use evidence during learning* to learn to appropriately weight its importance

	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**

Gene prediction in *C. elegans* (CDS evaluation)

Behr et al., in prep., 2010

Learning to use Expression Measurements

Two approaches:

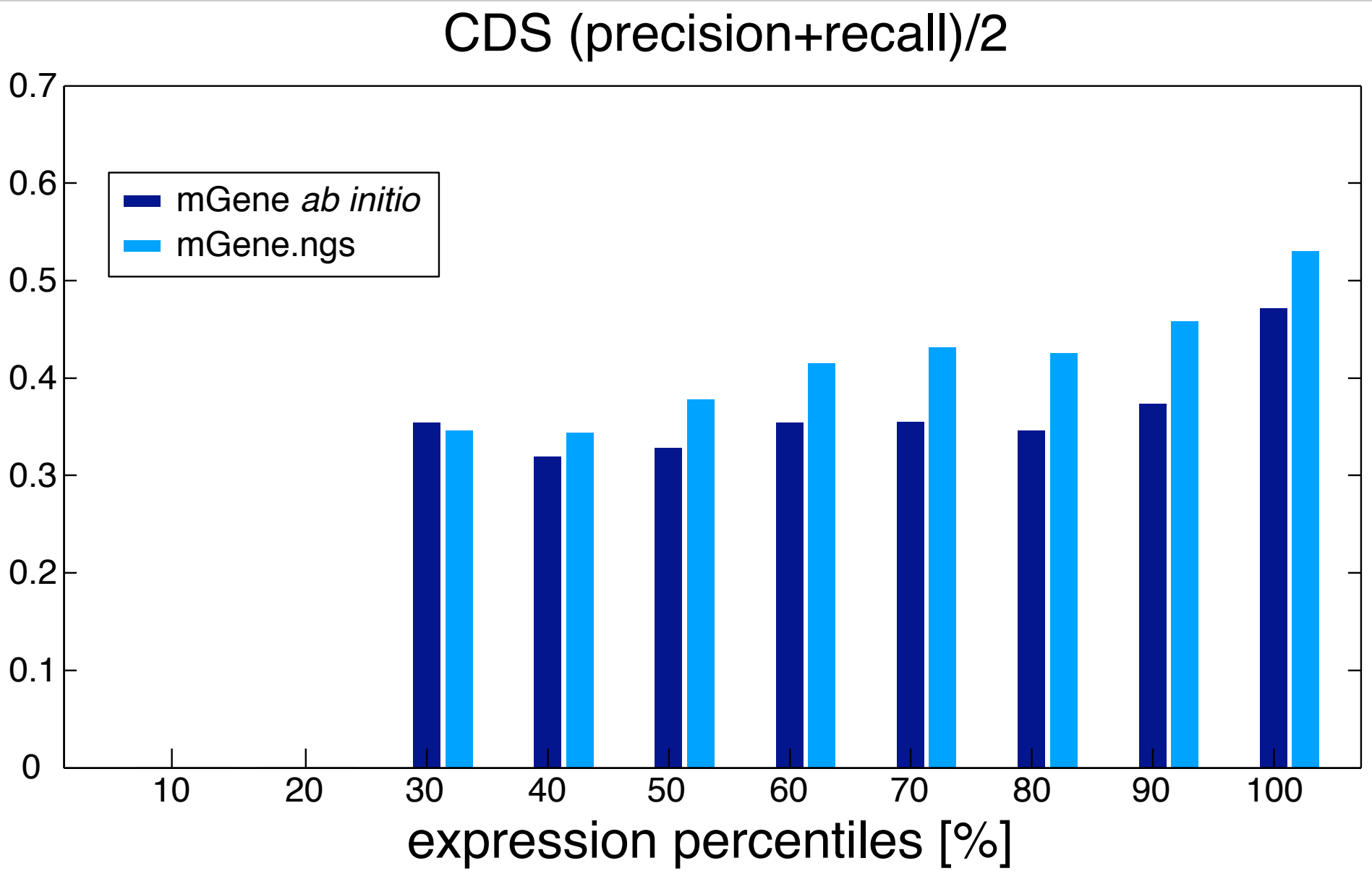
- Heuristic to incorporate ESTs/reads/tiling array measurements to *refine predictions*
- Directly *use evidence during learning* to learn to appropriately weight its importance

	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 trained	84.6	84.9	84.8	49.1	55.2	52.0
RNA-Seq/ESTs trained	84.7	86.9	**85.8**	50.3	60.5	**54.9**

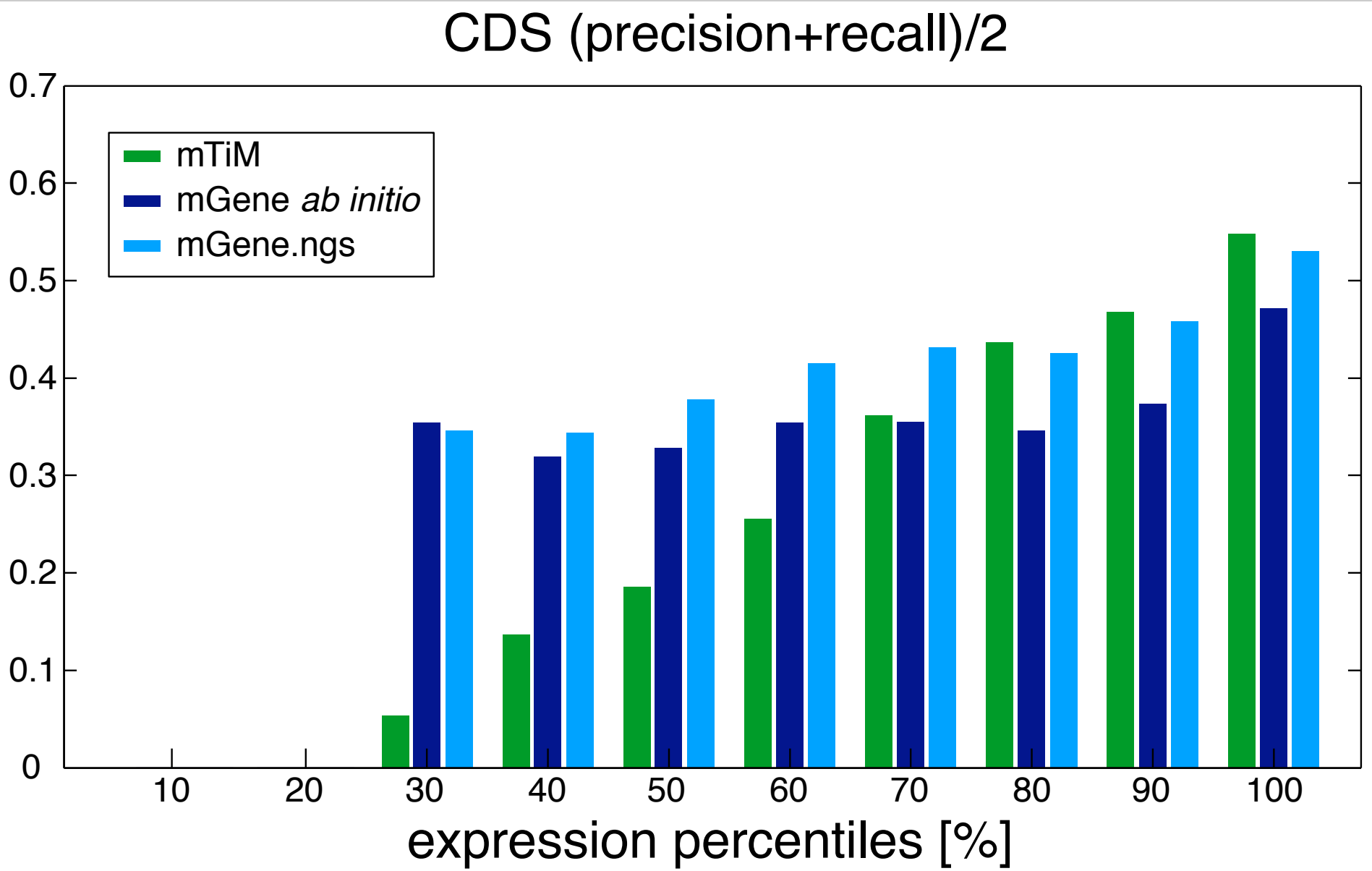
Gene prediction in *C. elegans* (CDS evaluation)

Behr et al., in prep., 2010

Preliminary Evaluation (*C. elegans*)



Preliminary Evaluation (*C. elegans*)



Digestion

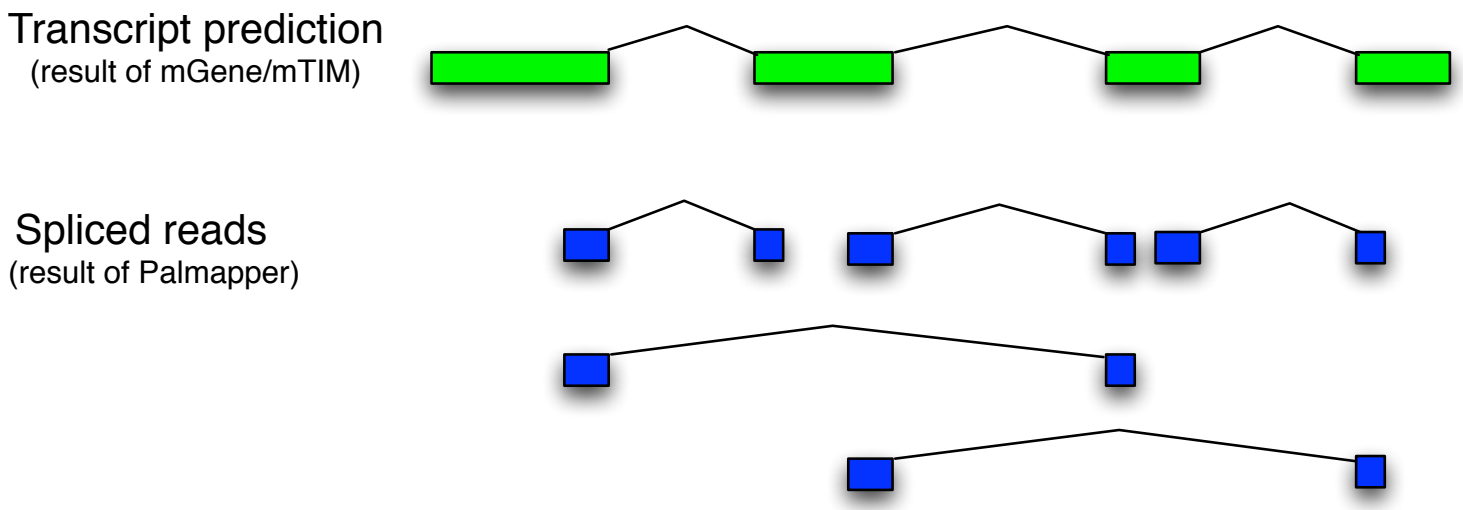
- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alternative Transcripts:** Spliced reads for splicing graphs:
- Paths through splicing graph define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

Digestion

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

Digestion

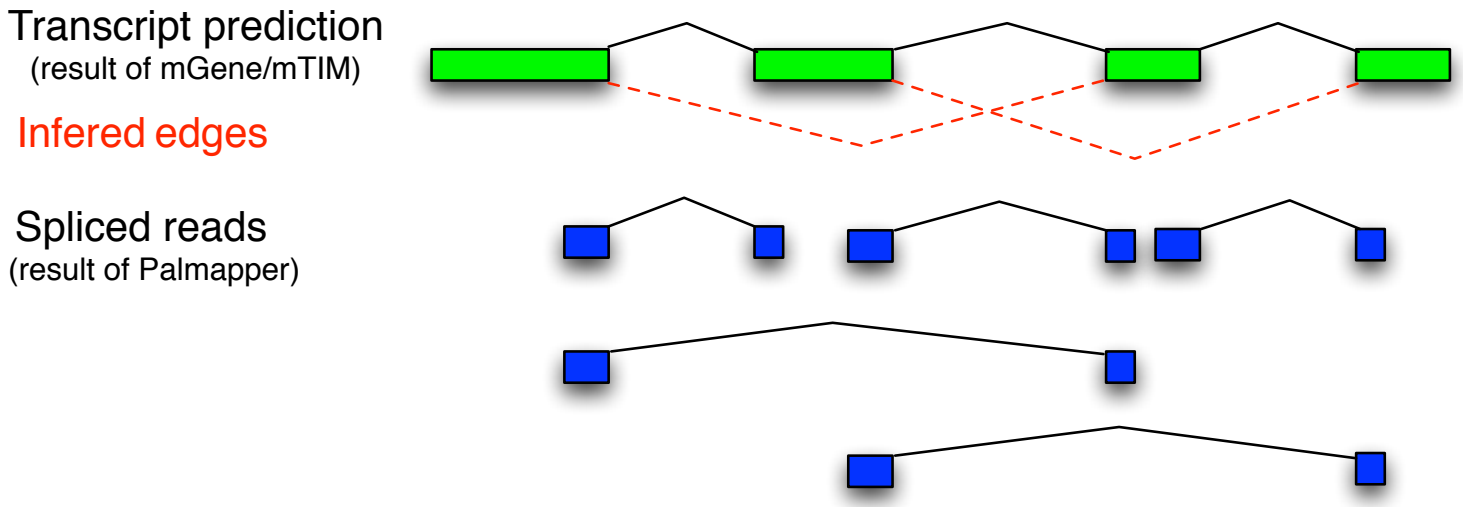
- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alternative Transcripts:** Spliced reads for splicing graphs:



- Paths through splicing graph define *alternative transcripts*

Digestion

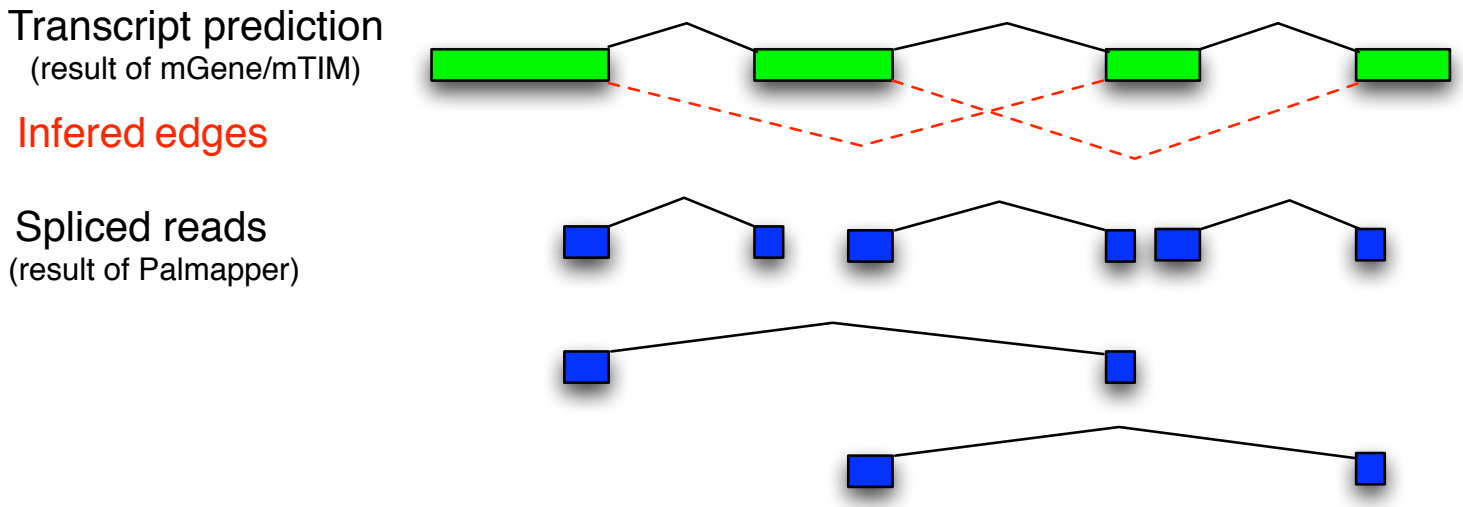
- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alternative Transcripts:** Spliced reads for splicing graphs:



- Paths through splicing graph define *alternative transcripts*

Digestion

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



- Paths through splicing graph define *alternative transcripts*

Prediction of Alternative Transcripts



- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:
    Input: DNA sequence
    Output: All possible transcripts (not just one)
- First step: Predict **"simple splicing graphs"**
- Conceptually works . . .
- But, only with **clean data** we train the model!

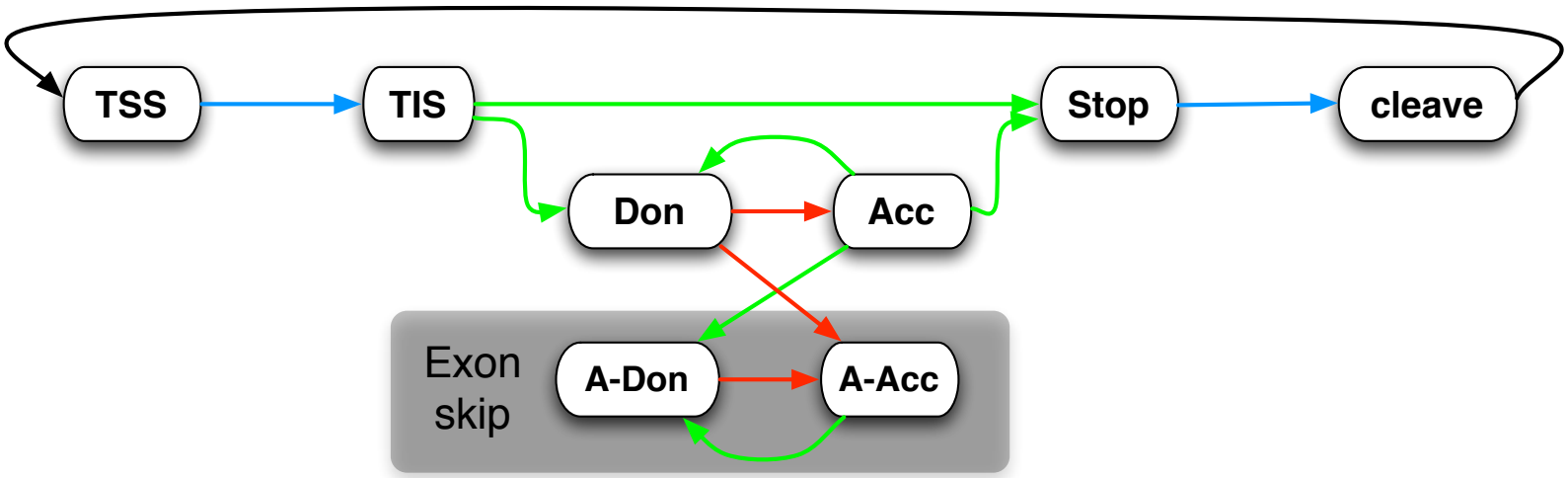
Prediction of Alternative Transcripts

- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:

    Input: DNA sequence

    Output: All possible transcripts (not just one)

- First step: Predict **"simple splicing graphs"**



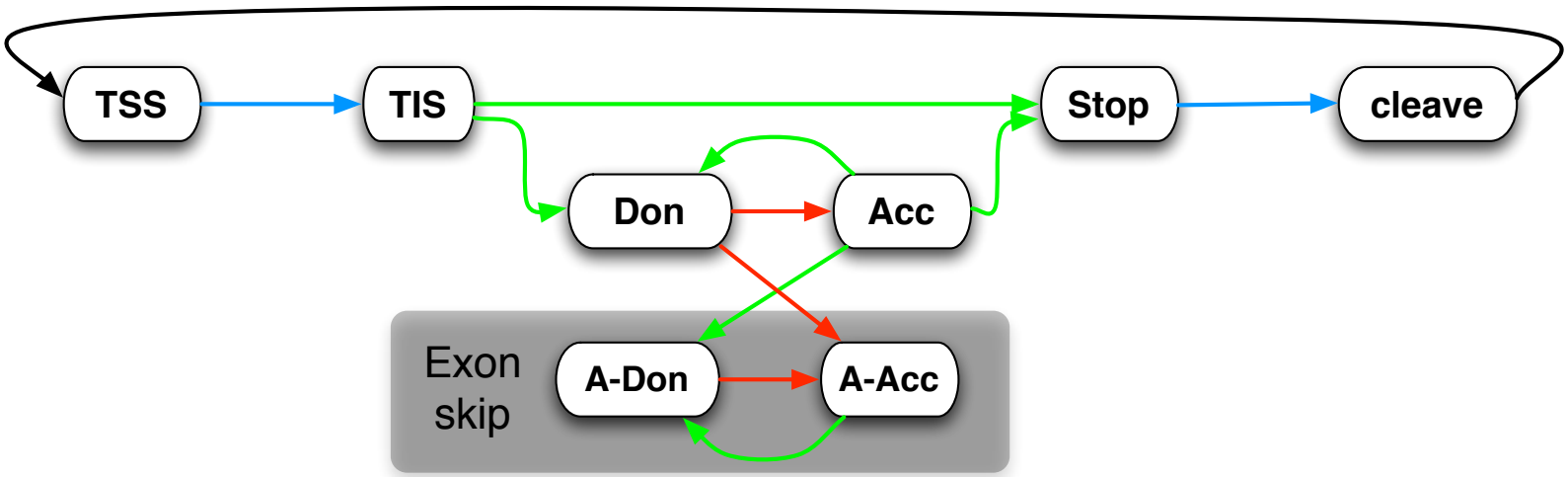
- Conceptually works . . .
- But, only with **clean data** we train the model!

Prediction of Alternative Transcripts

- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:

    Input: DNA sequence

    Output: All possible transcripts (not just one)
- First step: Predict **"simple splicing graphs"**



- Conceptually works ...
- But, only with **clean data** we train the model!

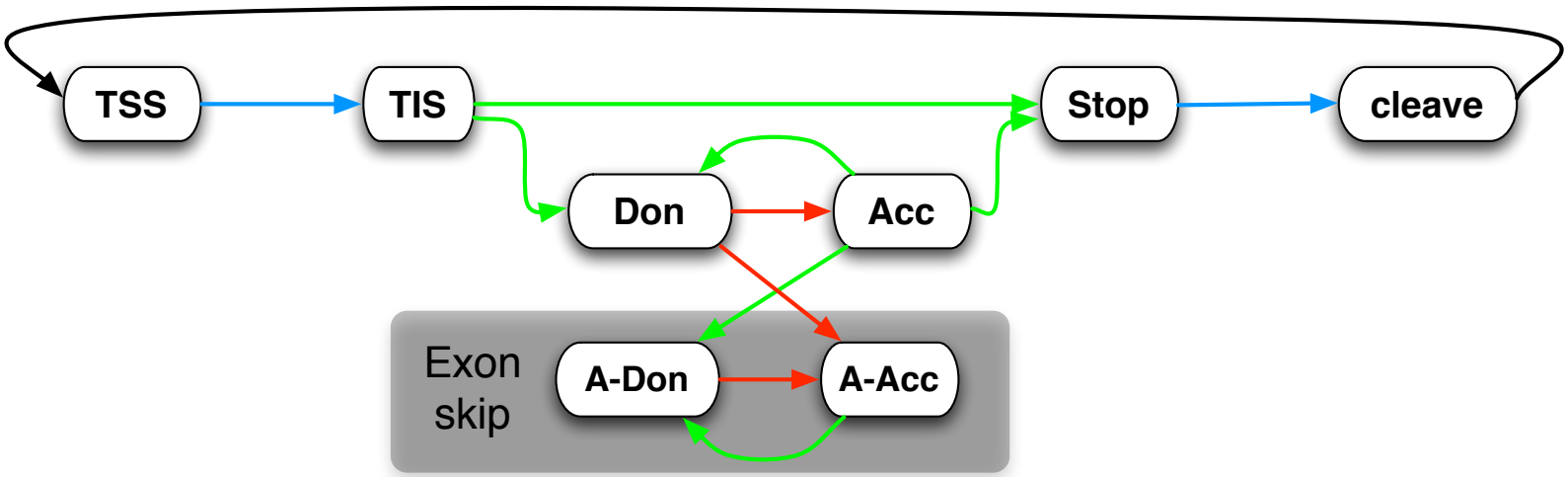
Prediction of Alternative Transcripts

- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:

    Input: DNA sequence

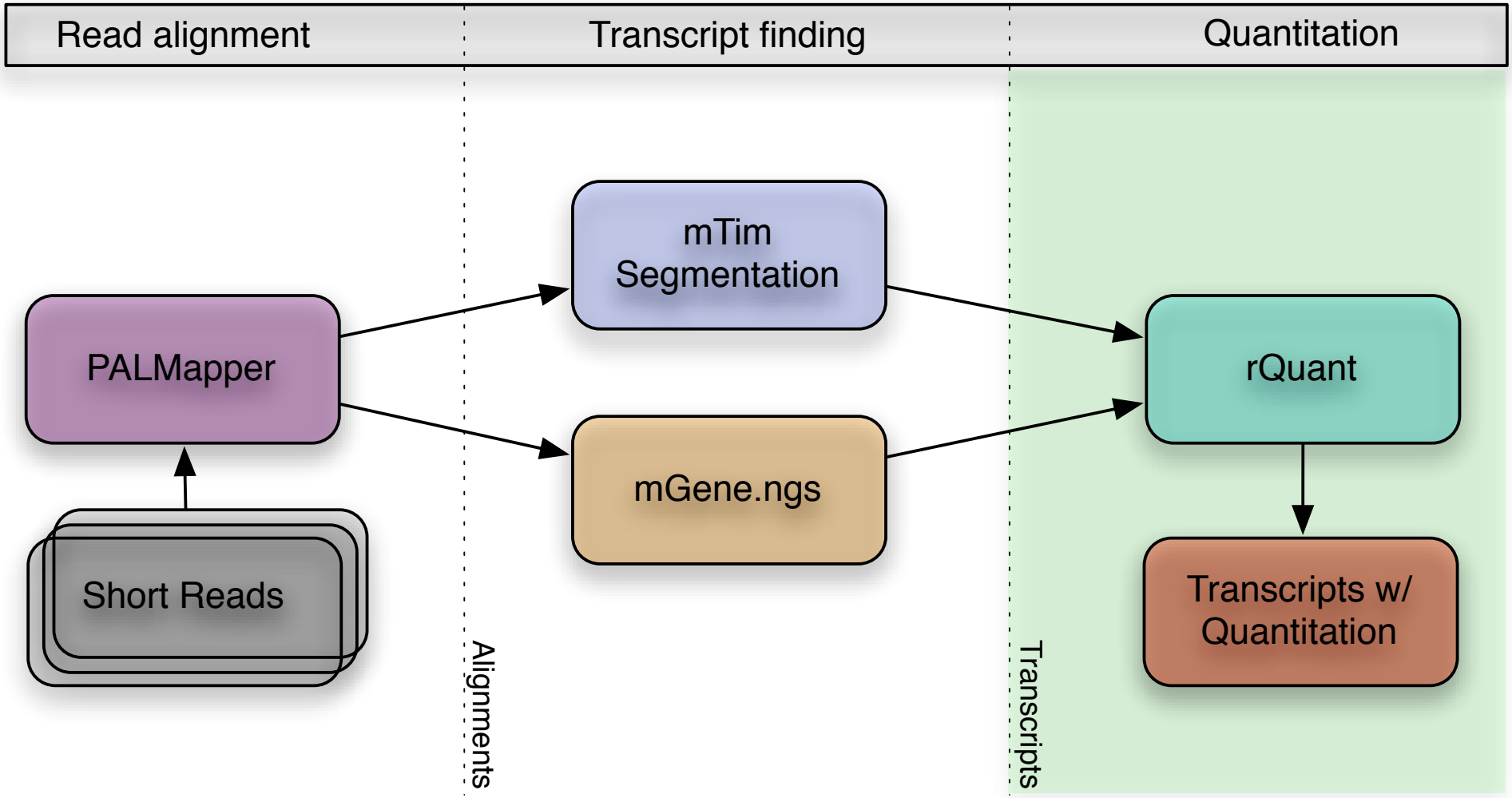
    Output: All possible transcripts (not just one)

- First step: Predict **"simple splicing graphs"**

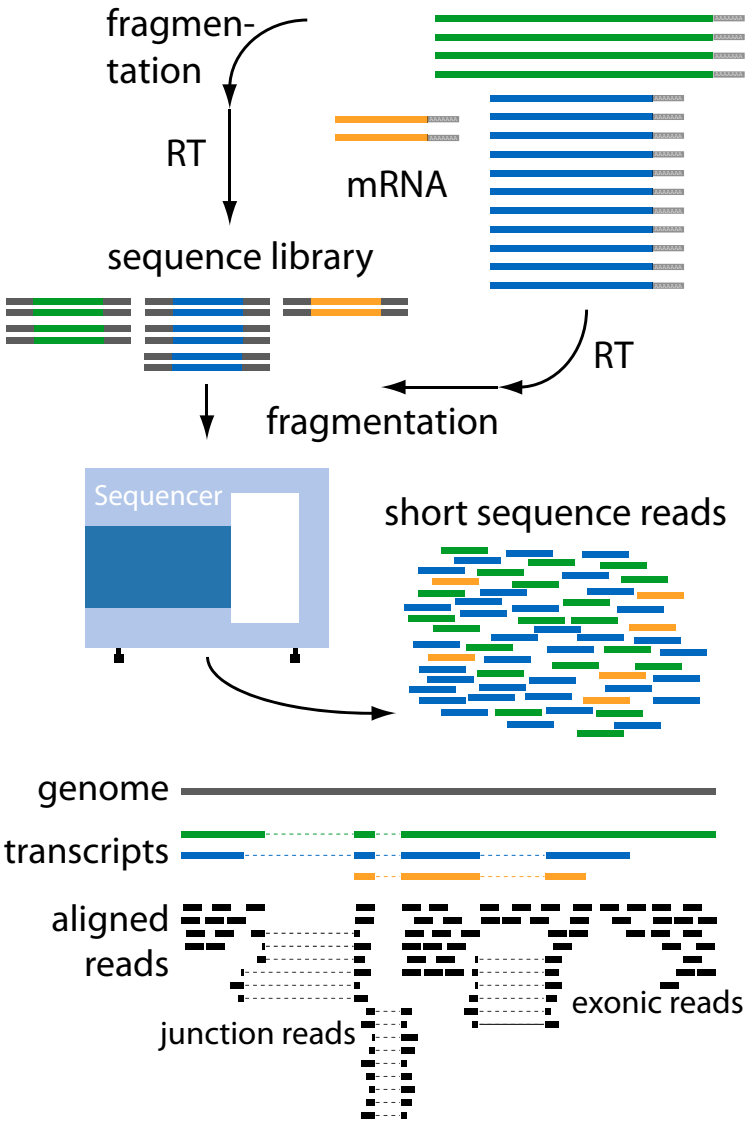


- Conceptually works ...
- But, only with **clean data** we train the model!

RNA-Seq Pipeline Overview

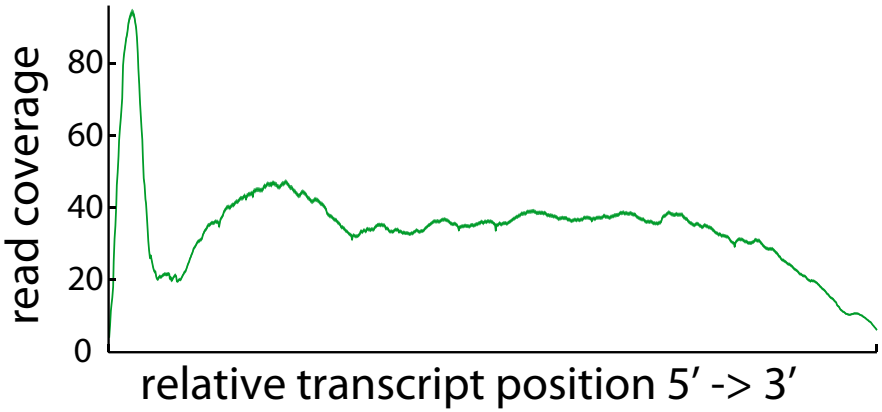


RNA-Seq Biases and Quantitation



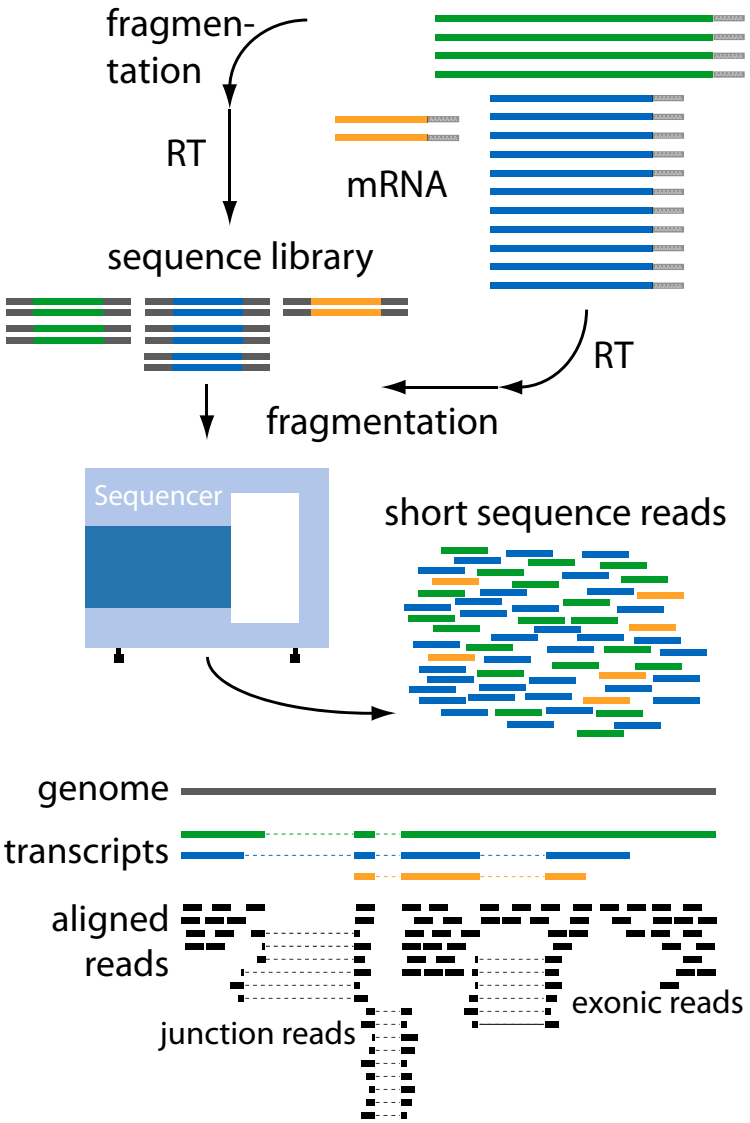
Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping



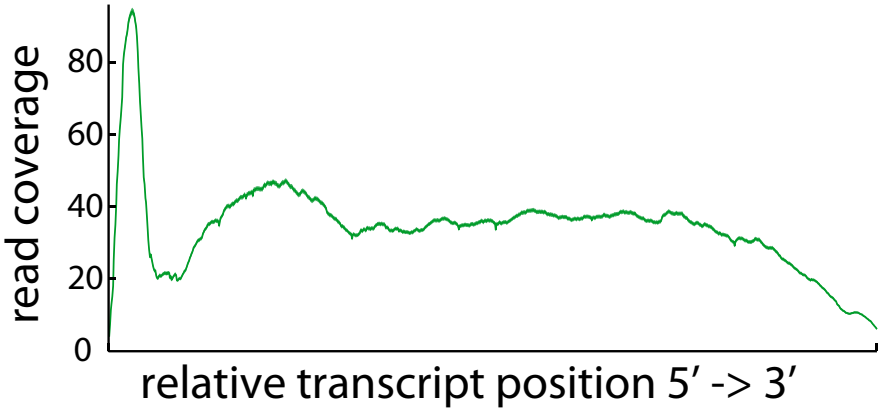
(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

RNA-Seq Biases and Quantitation



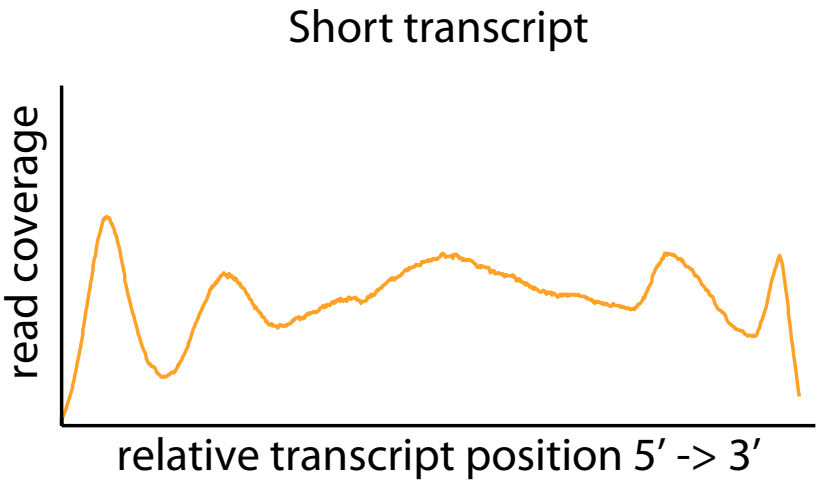
Biases due to ...

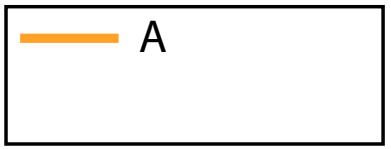
- cDNA library construction
- Sequencing
- Read mapping



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

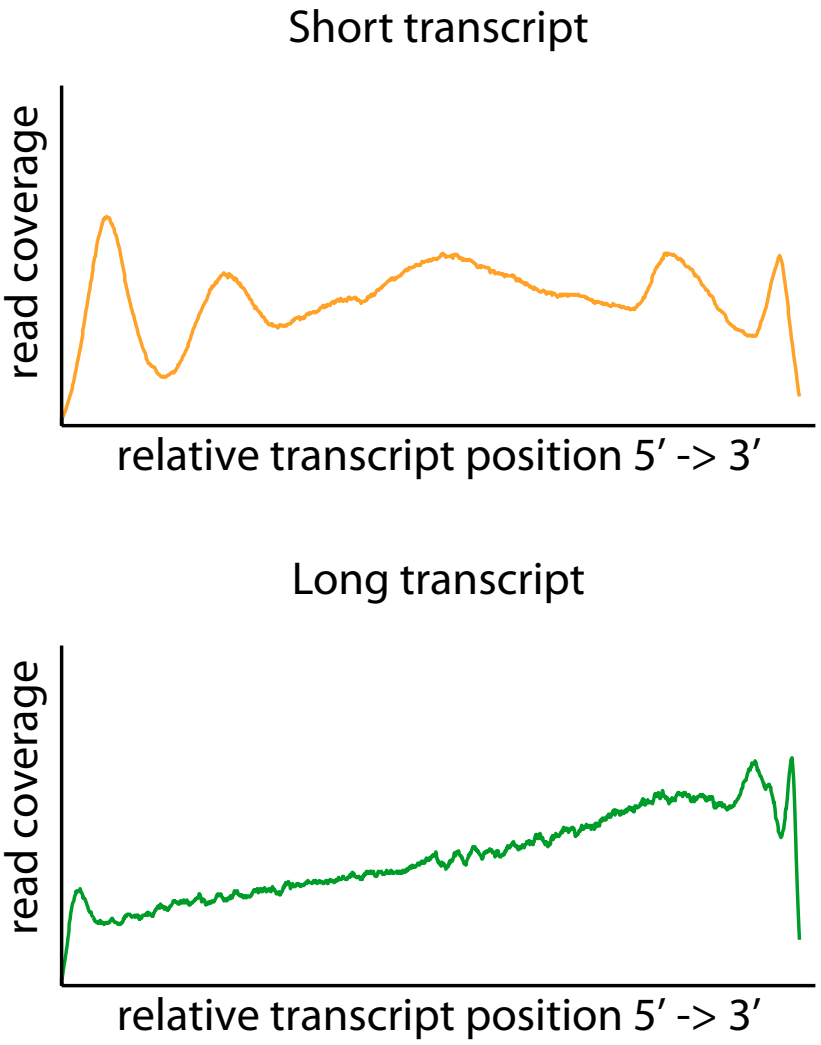
rQuant – Basic Idea





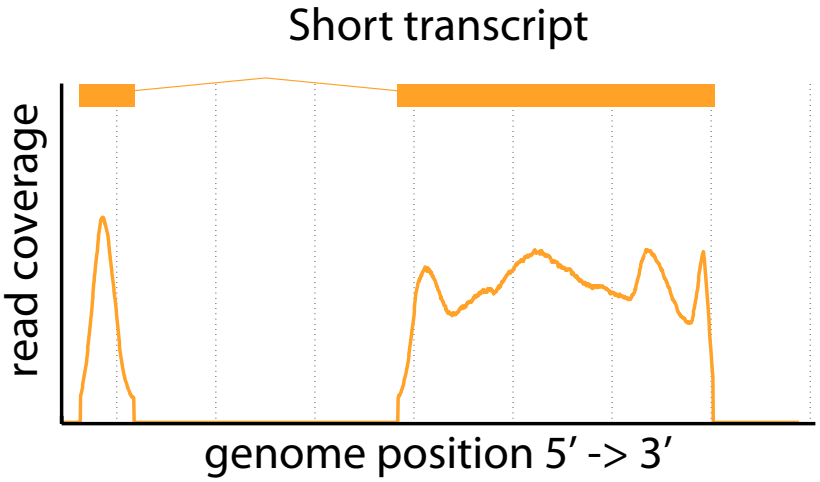
$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

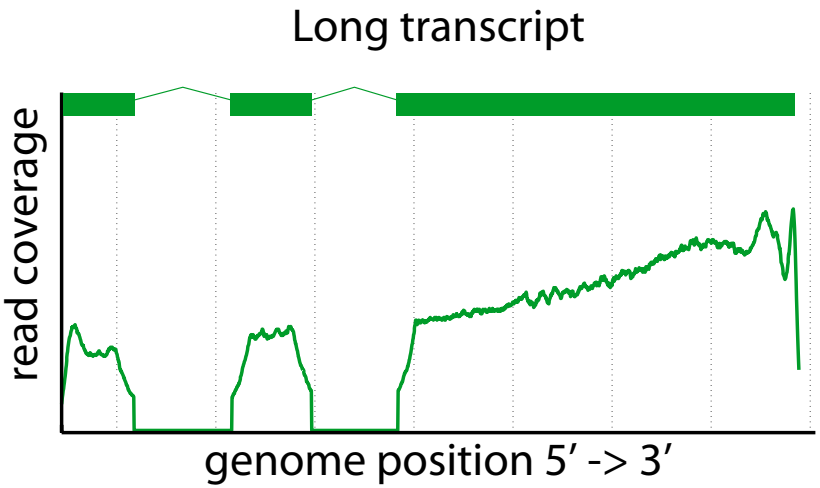
rQuant – Basic Idea

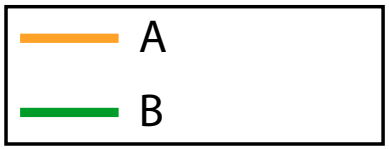


$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea

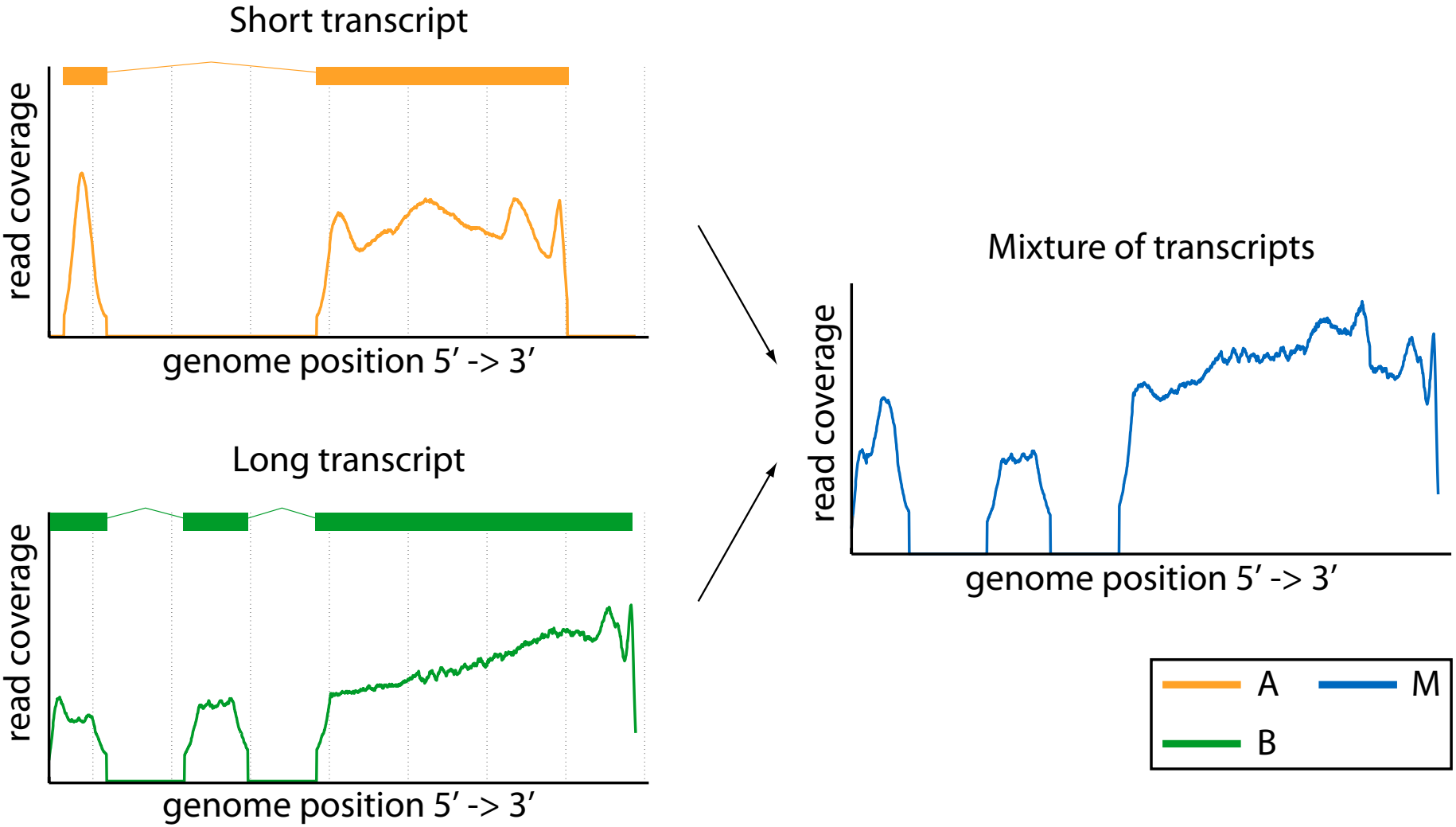






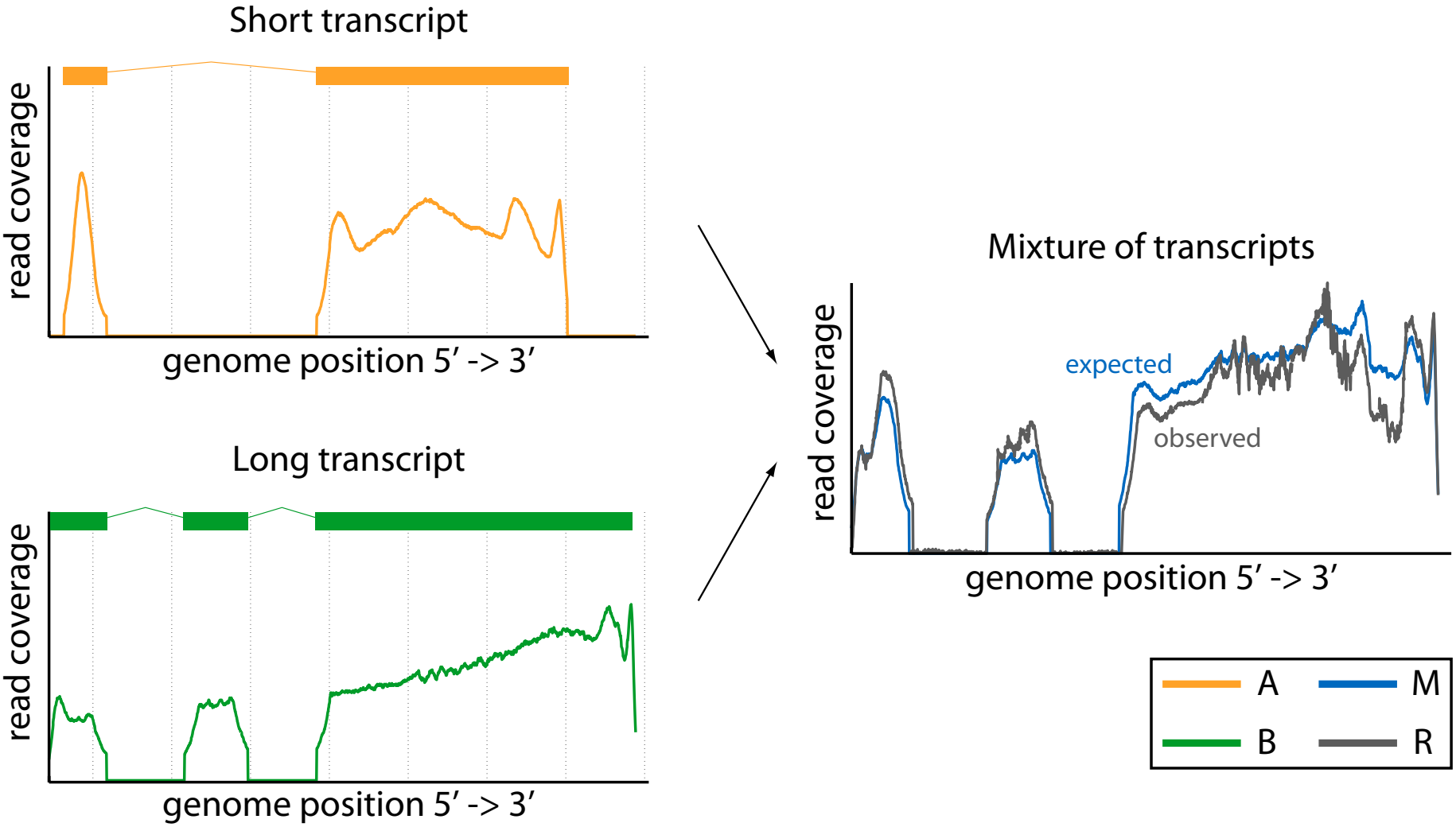
$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea



$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea

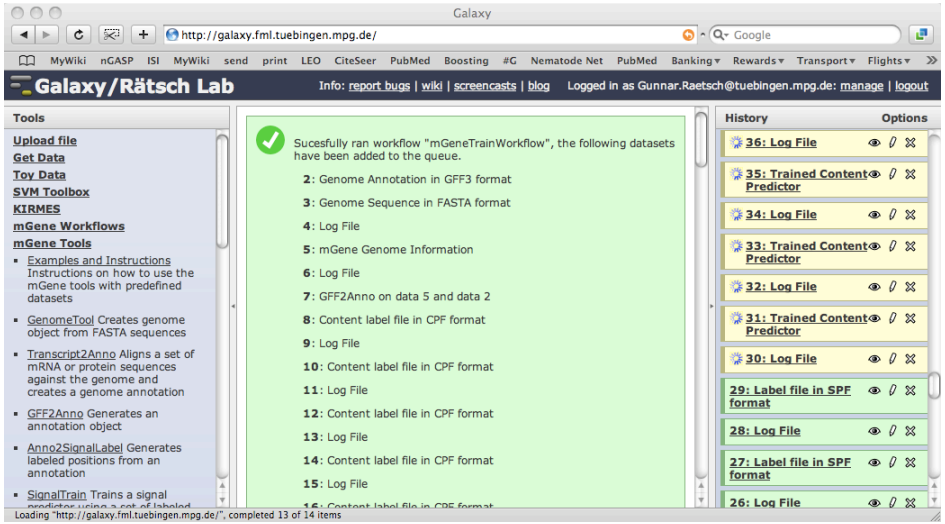


$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

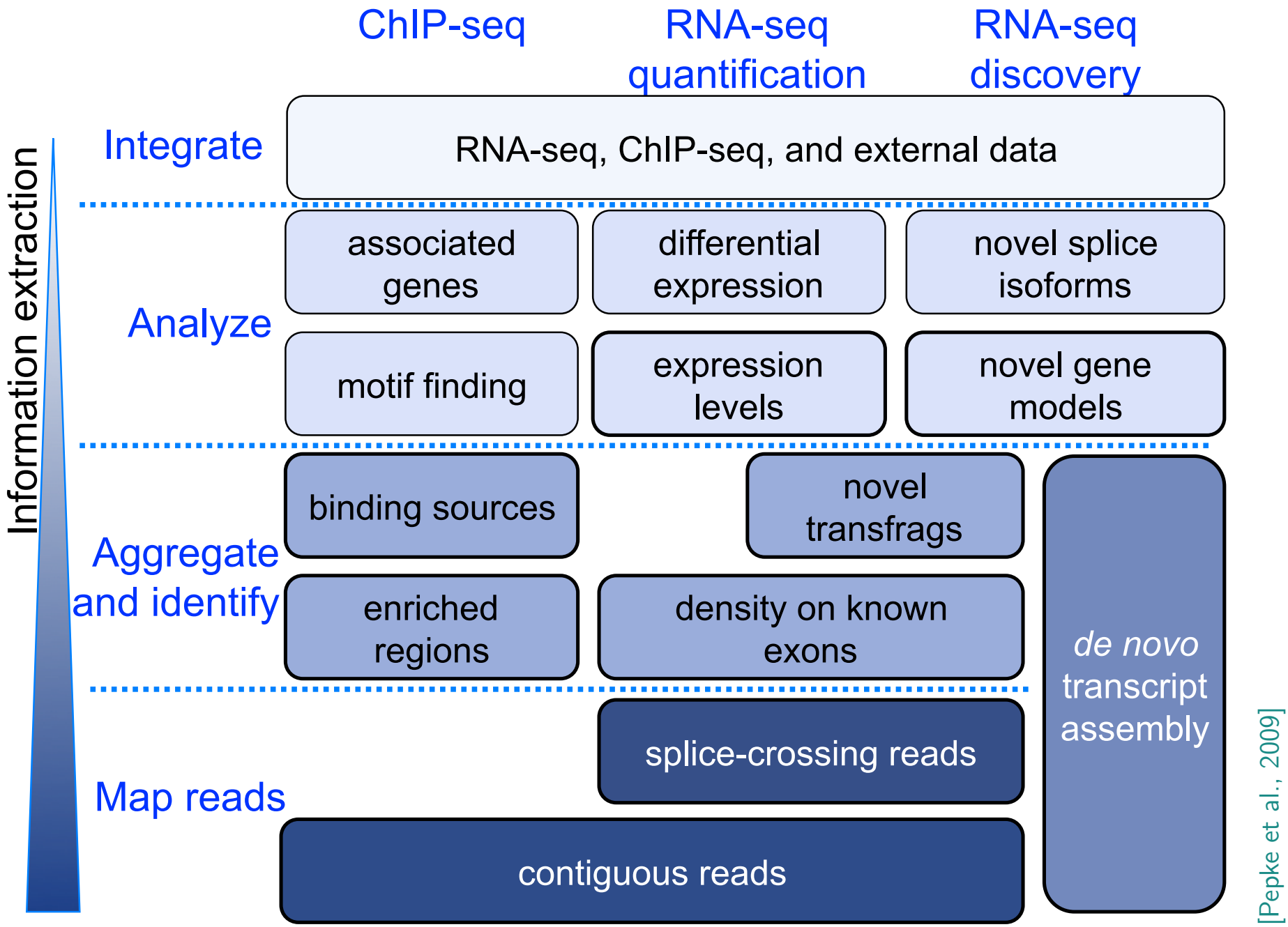
Galaxy-based Web Services for NGS Analyses

Galaxy-based web service http://galaxy.fml.mpg.de

- PALMapper http://fml.mpg.de/raetsch/suppl/palmapper
- mGene http://mgene.org/web
- mTIM http://fml.mpg.de/raetsch/suppl/mtim (in prep.)
- rQuant http://fml.mpg.de/raetsch/suppl/rquant/web



[Rätsch et al., in preparation, 2010]


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

[Behr et al., 2009]

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%**

In progress:

- Study effect of other information sources for gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Data

[Behr et al., 2009]

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%**

In progress:

- Study effect of other information sources for gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Data

[Behr et al., 2009]

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%**

In progress:

- Study effect of other information sources for gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Data

[Behr et al., 2009]

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%**

In progress:

- Study effect of other information sources for gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Data

[Behr et al., 2009]

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%**

In progress:

- Study effect of other information sources for gene prediction
- Ideally, consider measurements from the same sample

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

Acknowledgements



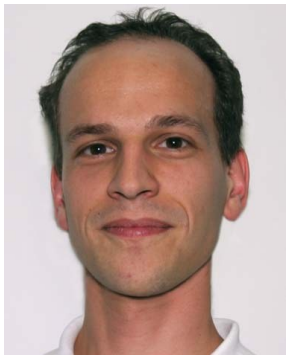
Fabio De Bona

Alignments



Jonas Behr

Gene finding



Georg Zeller

Segmentation



Regina Bohnert

Quantitation

Help with Slides

- Jonas Behr (FML)
- Georg Zeller (EMBL)
- Regina Bohnert (FML)
- Ali Mortazavi (Caltech)

Funding by DFG &
Max Planck Society.

Thank you for your attention!

Acknowledgements



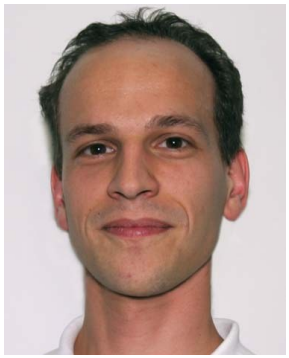
Fabio De Bona

Alignments



Jonas Behr

Gene finding



Georg Zeller

Segmentation



Regina Bohnert

Quantitation

Help with Slides

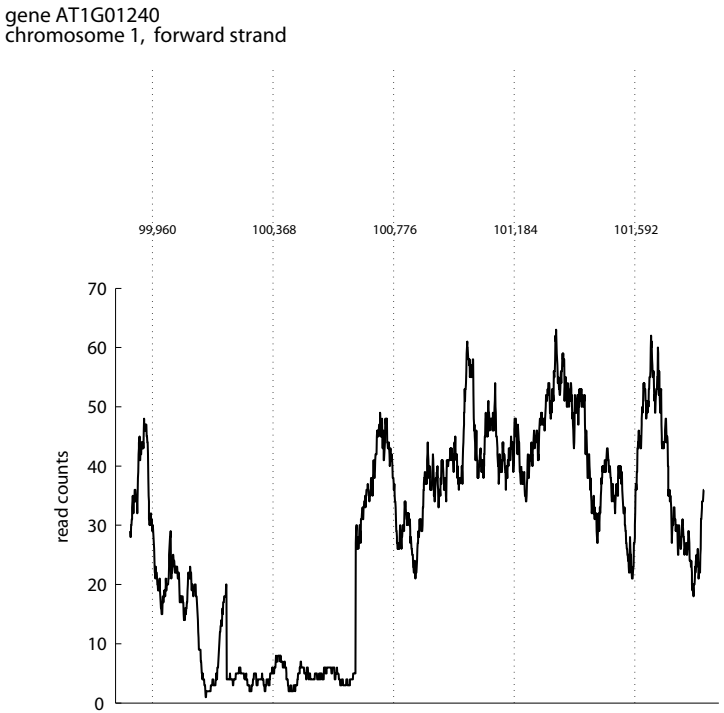
- Jonas Behr (FML)
- Georg Zeller (EMBL)
- Regina Bohnert (FML)
- Ali Mortazavi (Caltech)

Funding by DFG &
Max Planck Society.

Thank you for your attention!

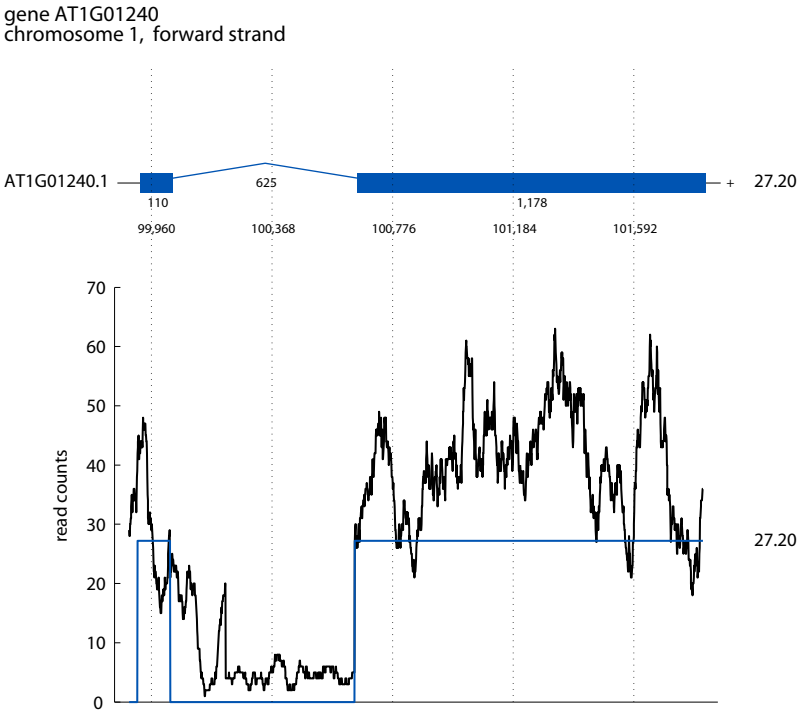
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



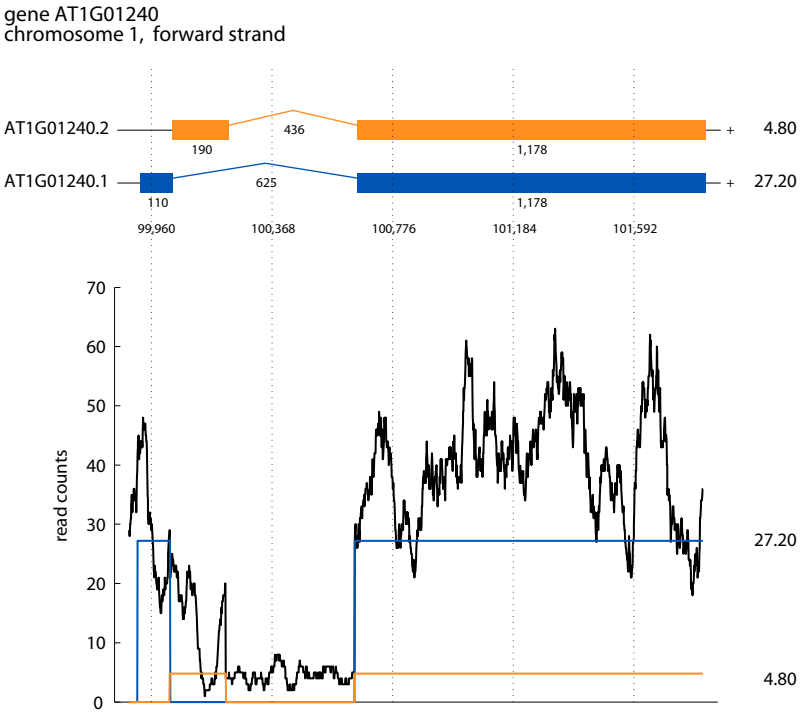
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



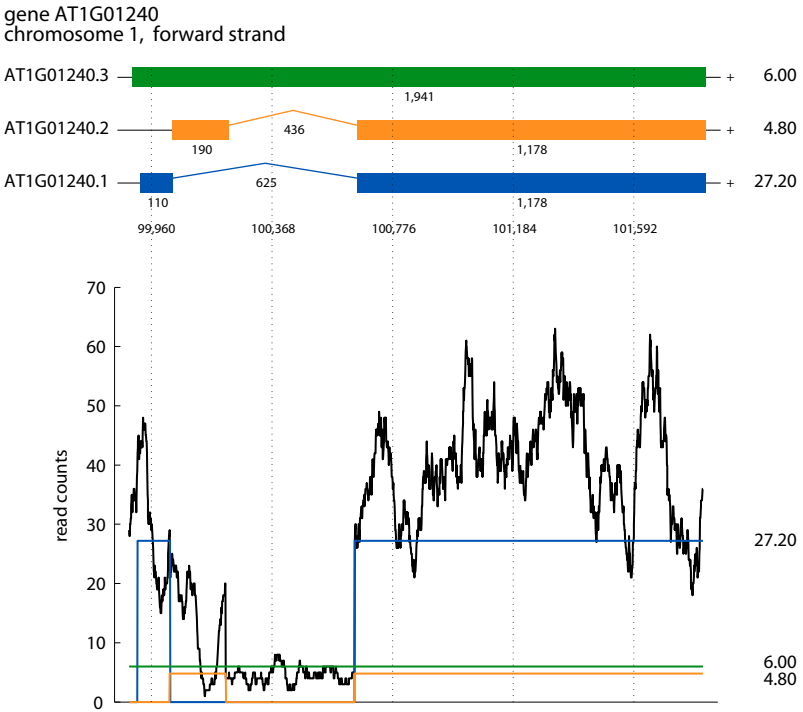
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



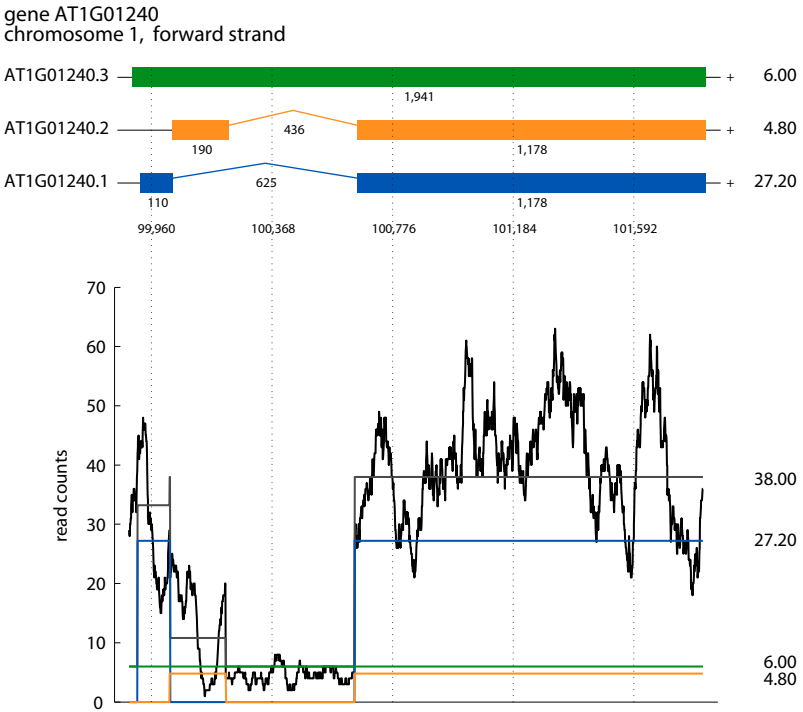
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



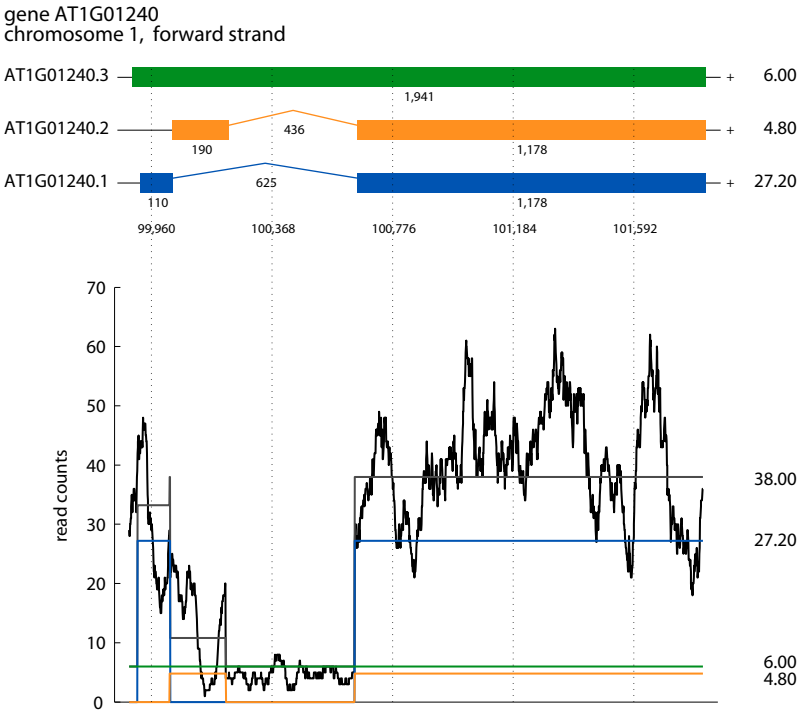
rQuant – Iterative Algorithm

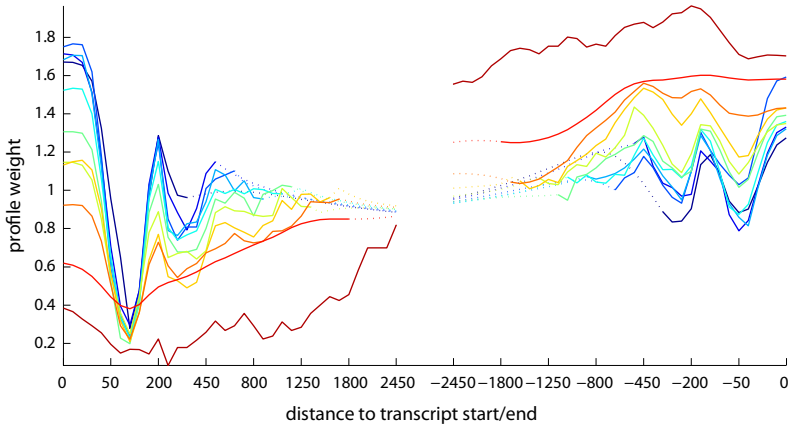
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



rQuant – Iterative Algorithm

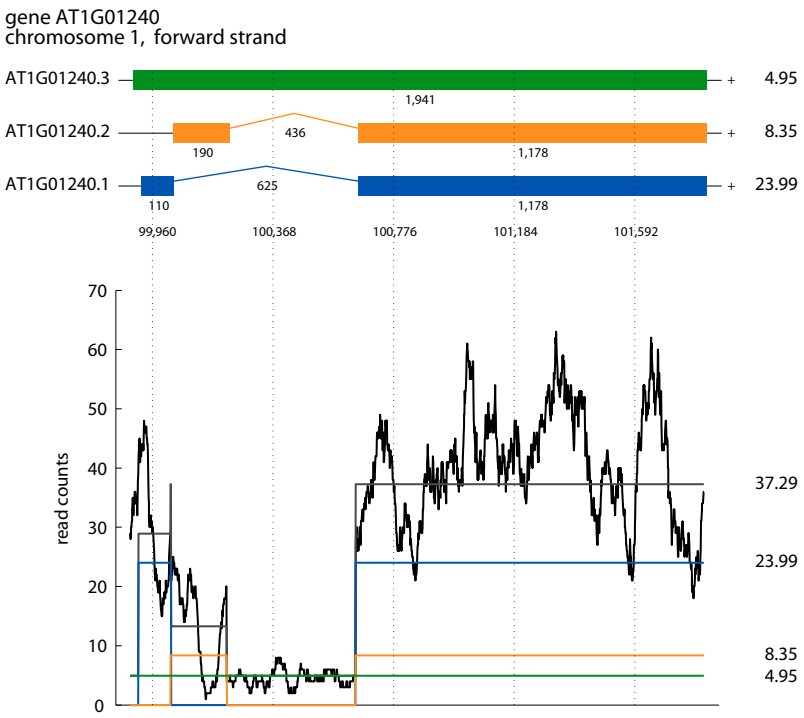
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.

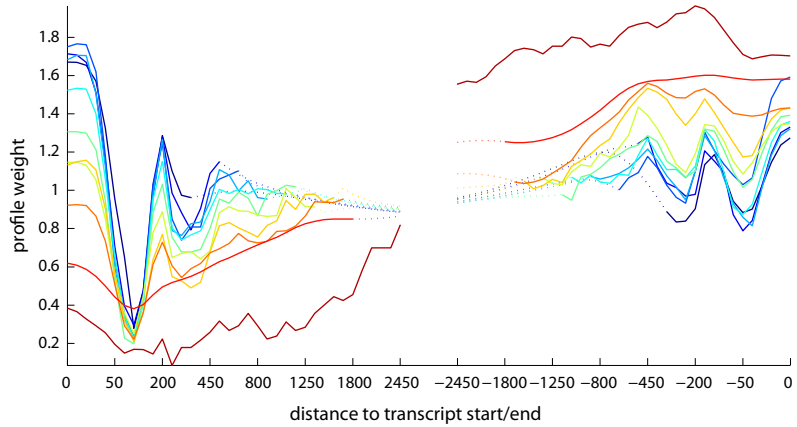




rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.





rQuant Evaluation I

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

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

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

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

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

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])

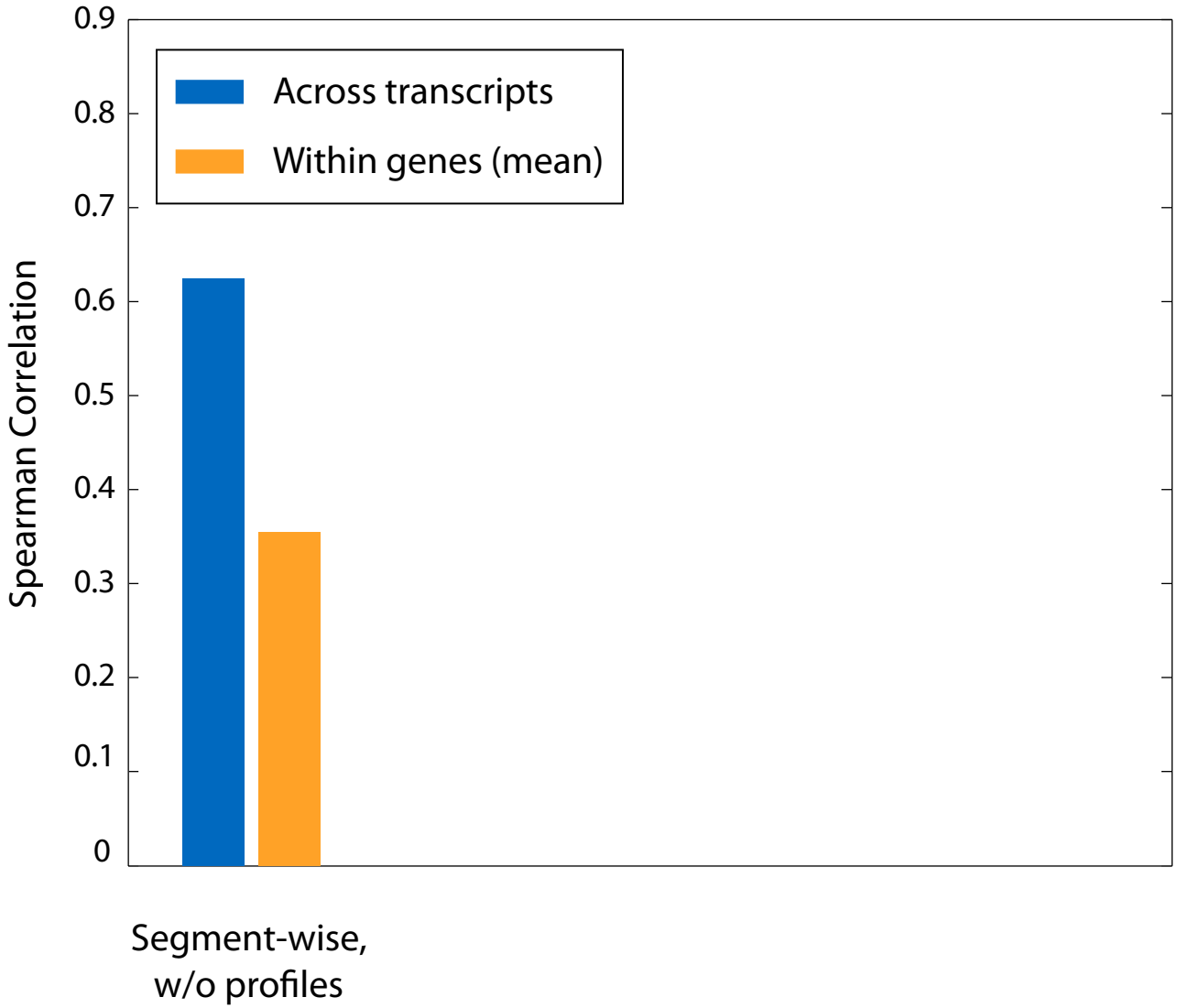
Estimate transcript abundances

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

Evaluation: Spearman correlation between

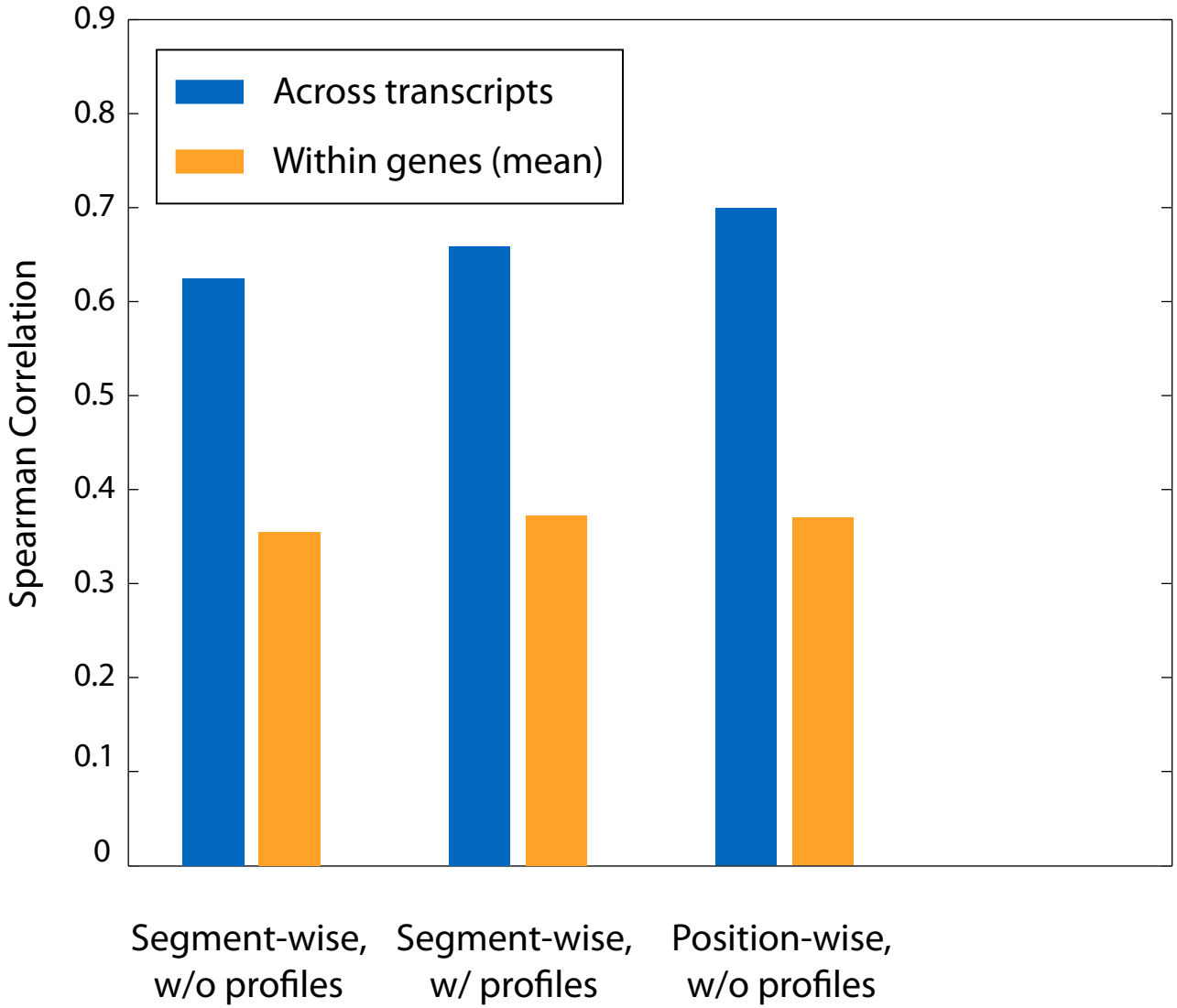
- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation II



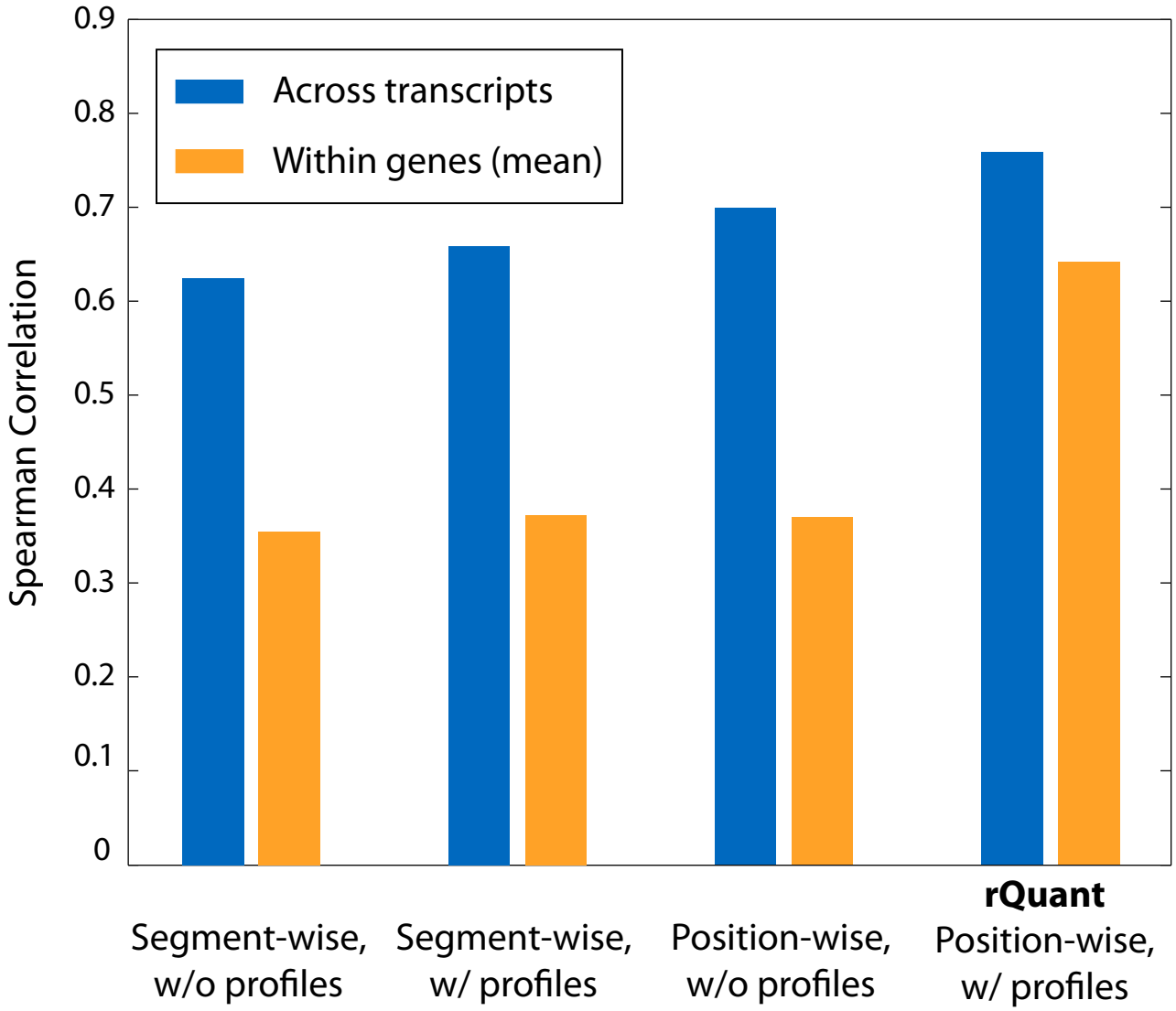
(Bohnert et al., submitted, 2010)

rQuant Evaluation II



(Bohnert et al., submitted, 2010)

rQuant Evaluation II



(Bohnert et al., submitted, 2010)

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, 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. Presented at the CSHL Genome Informatics Meeting, July 2009.

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.

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

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 II

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.

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, Jan 2009. doi: 10.1186/gb-2009-10-9-r98. URL `http://genomebiology.com/2009/10/9/R98`.

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.

References III

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.

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.