Next Generation Genome Annotation

Gunnar Rätsch

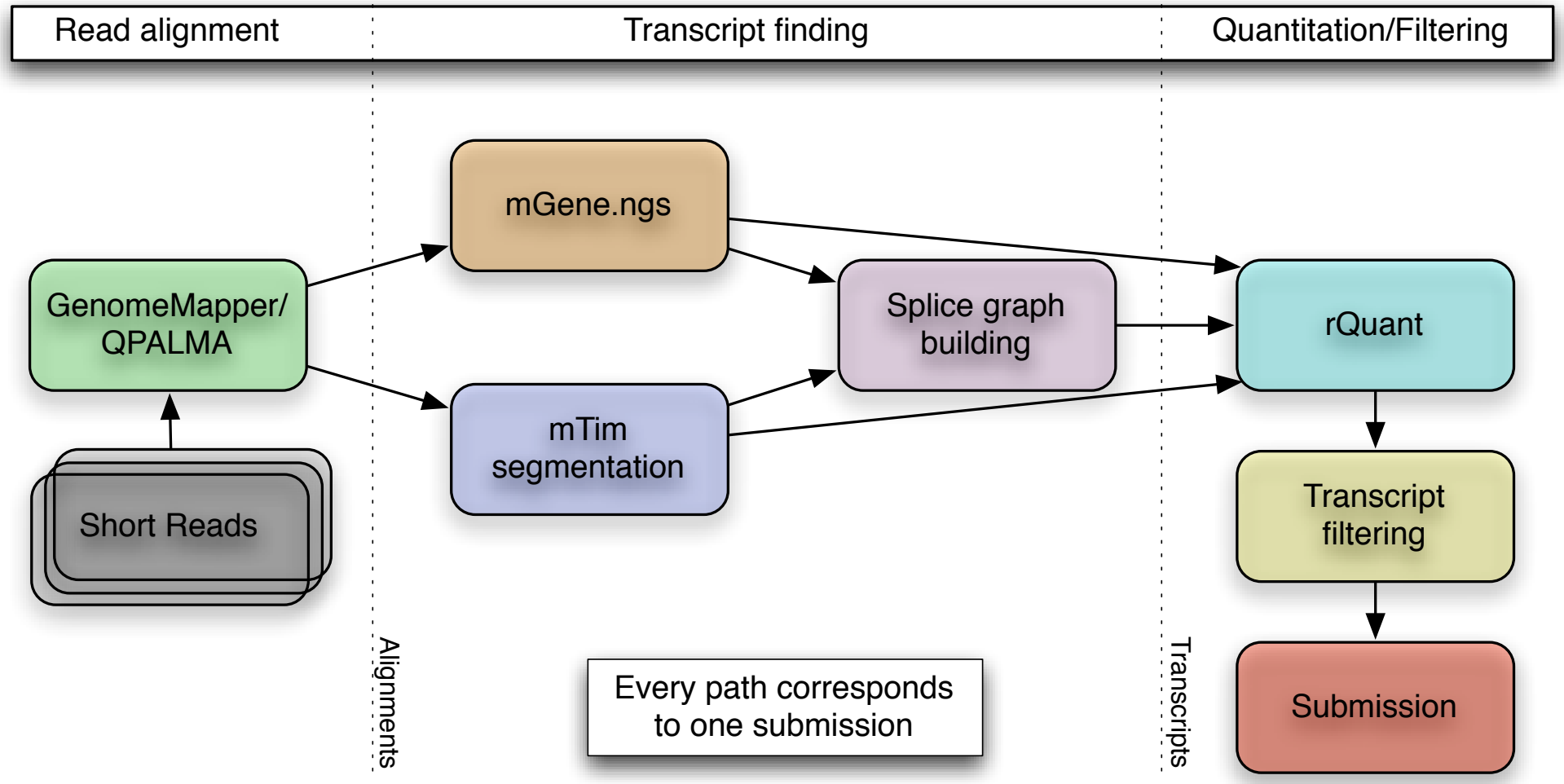
Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

November 10, 2009

The Wellcome Trust Genome Campus, Hinxton

RGASP Overview (Tübingen)



Read Alignment – GenomeMapper/QPALMA

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project (Schneeberger et al., 2009a)
- k-mer based index, well suited for smaller genomes with many mismatches/gaps

Read Alignment – GenomeMapper/QPALMA

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project (Schneeberger et al., 2009a)
- *k*-mer based index, well suited for smaller genomes with many mismatches/gaps

QPALMA for spliced read alignments:

- GenomeMapper identifies seed regions for *spliced alignments*
- Alignments are performed using QPALMA (De Bona et al., 2008)
- QPALMA is individually adapted to every SR dataset

Read Alignment – GenomeMapper/QPALMA

GenomeMapper for (unspliced) read mapping:

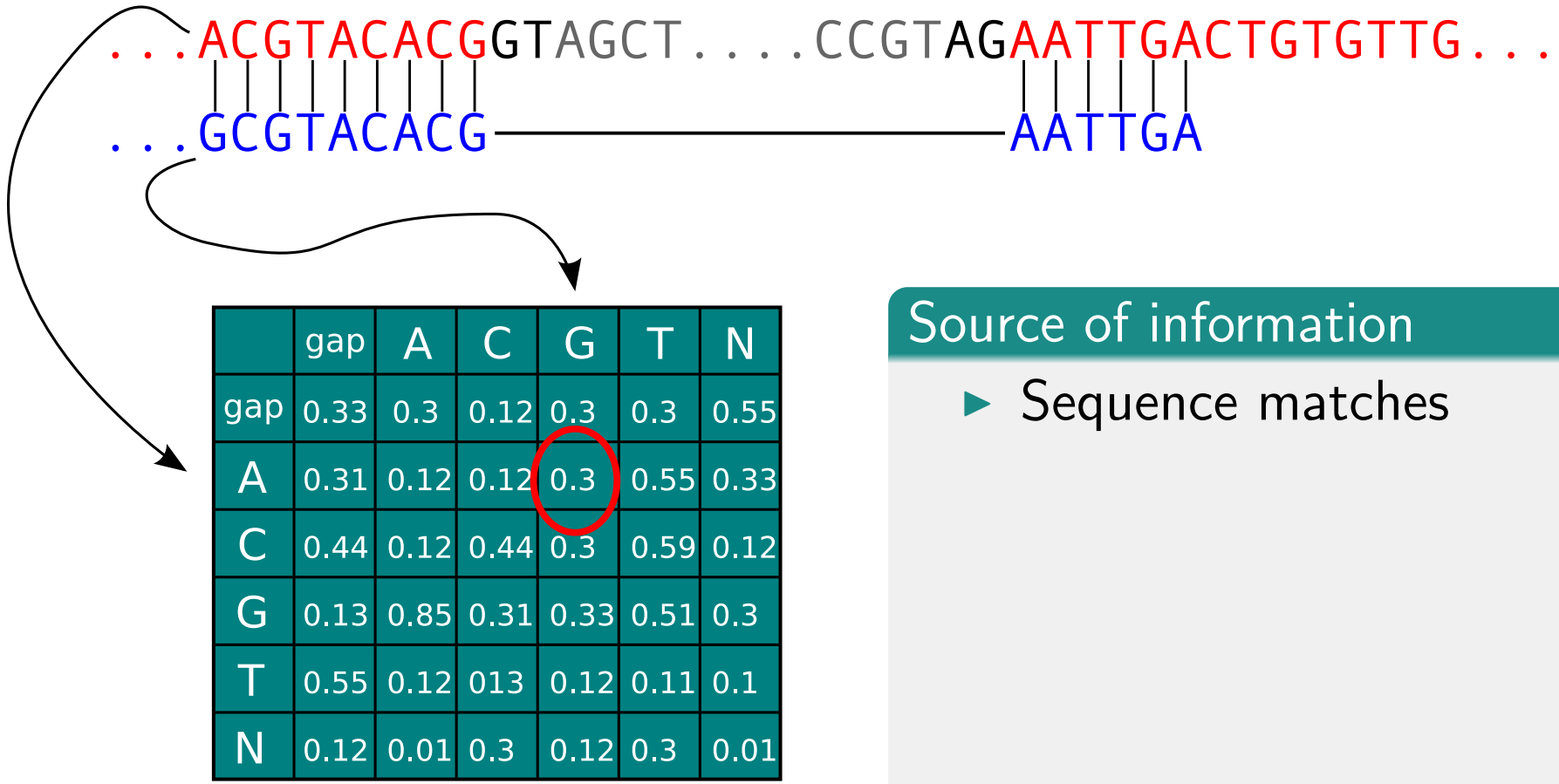
- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project (Schneeberger et al., 2009a)
- *k*-mer based index, well suited for smaller genomes with many mismatches/gaps

QPALMA for spliced read alignments:

- GenomeMapper identifies seed regions for *spliced alignments*
- Alignments are performed using QPALMA (De Bona et al., 2008)
- QPALMA is individually adapted to every SR dataset

Web server available at `http://galaxy.tuebingen.mpg.de`.

Read Alignment – QPALMA



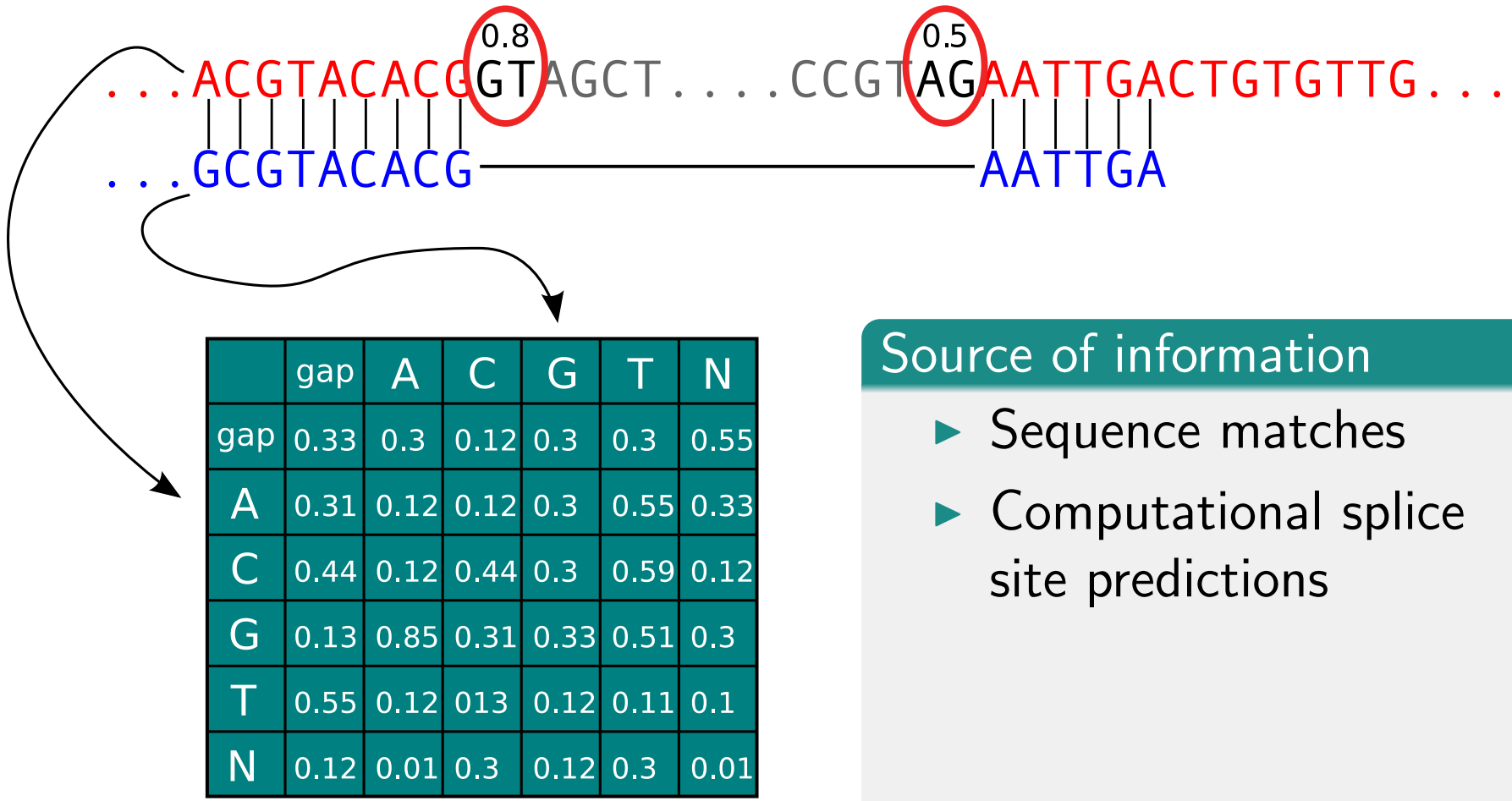
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches

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

Read Alignment – QPALMA



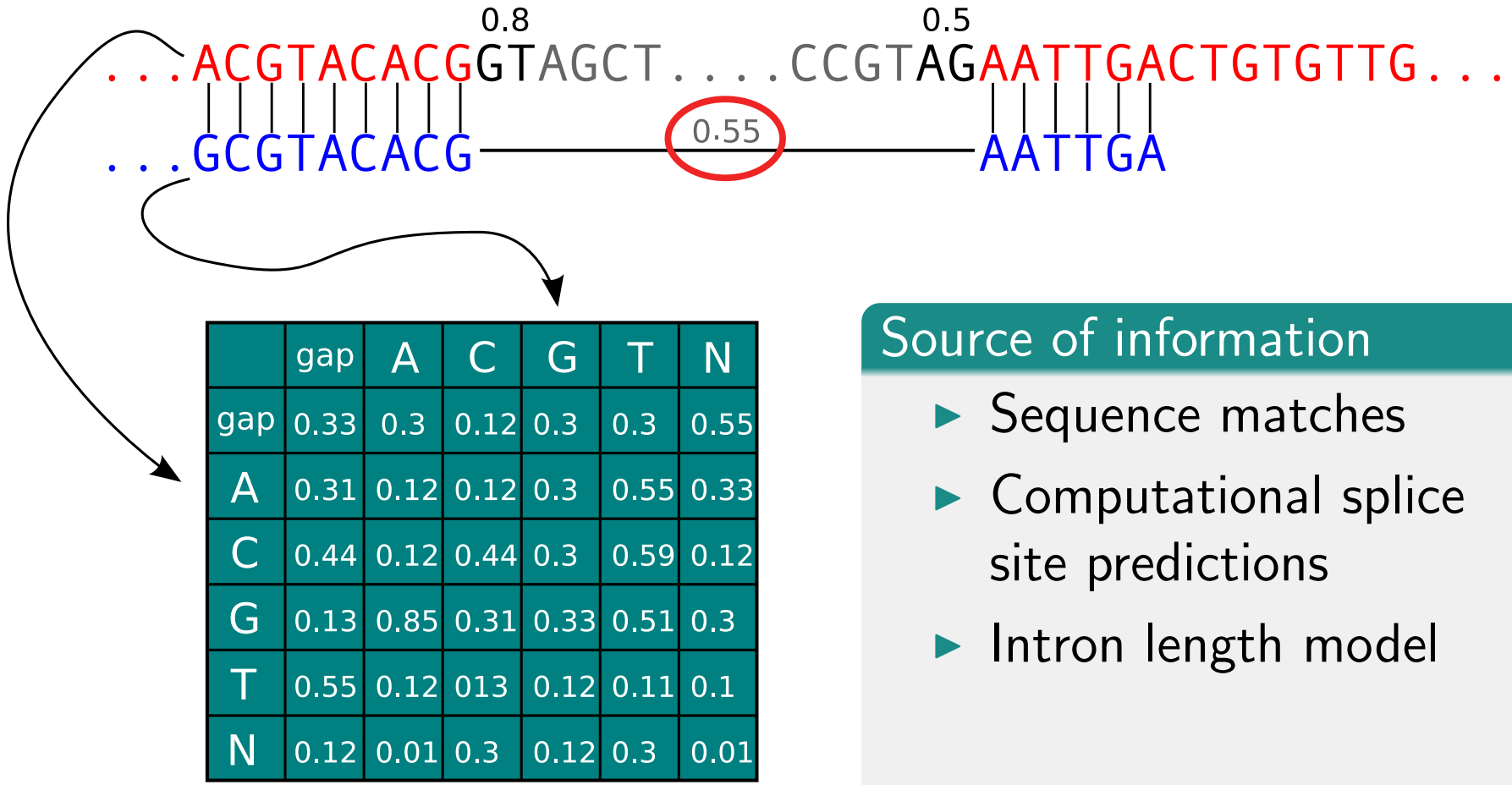
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions

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

Read Alignment – QPALMA



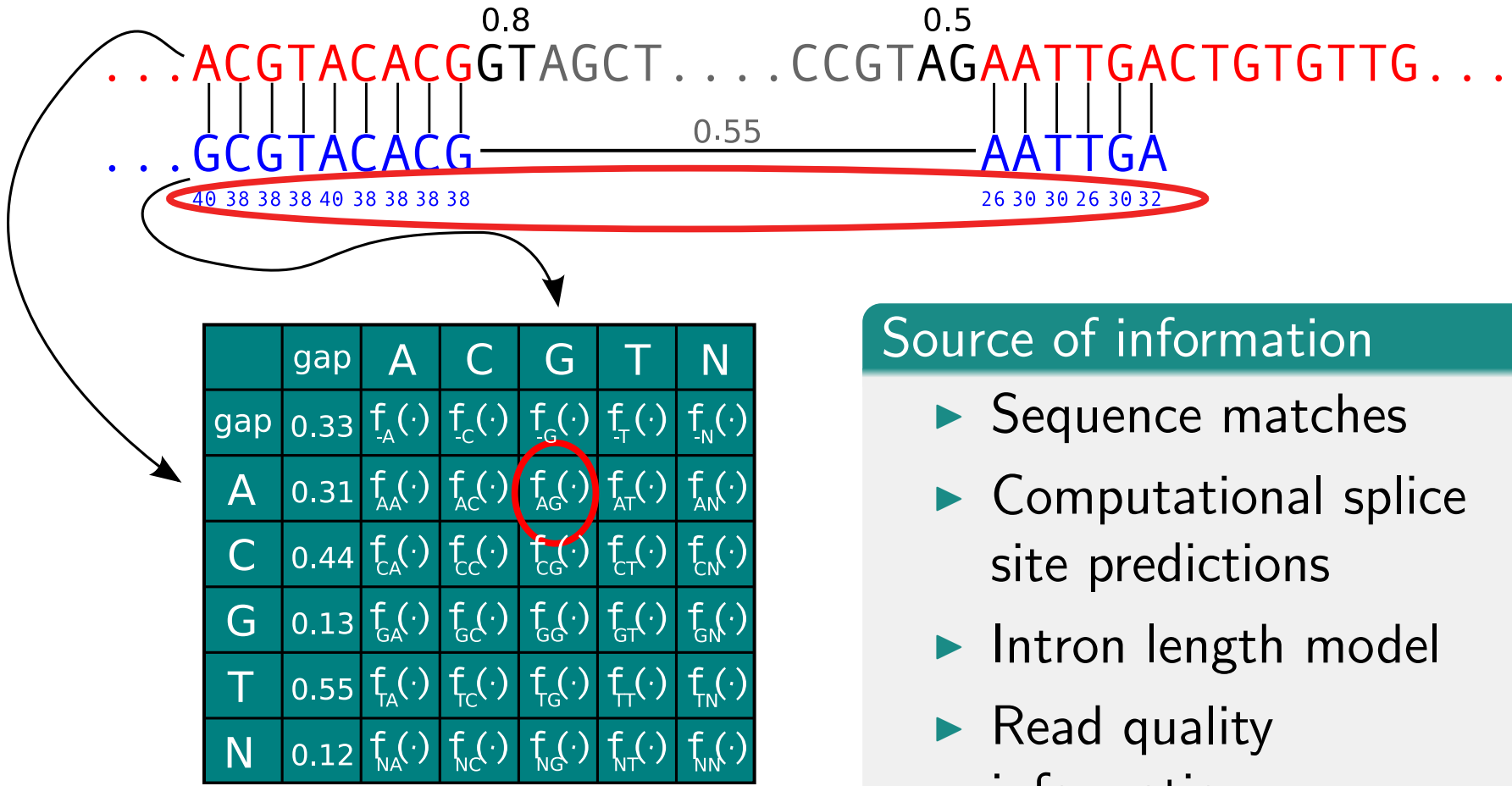
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model

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

Read Alignment – QPALMA



Source of information

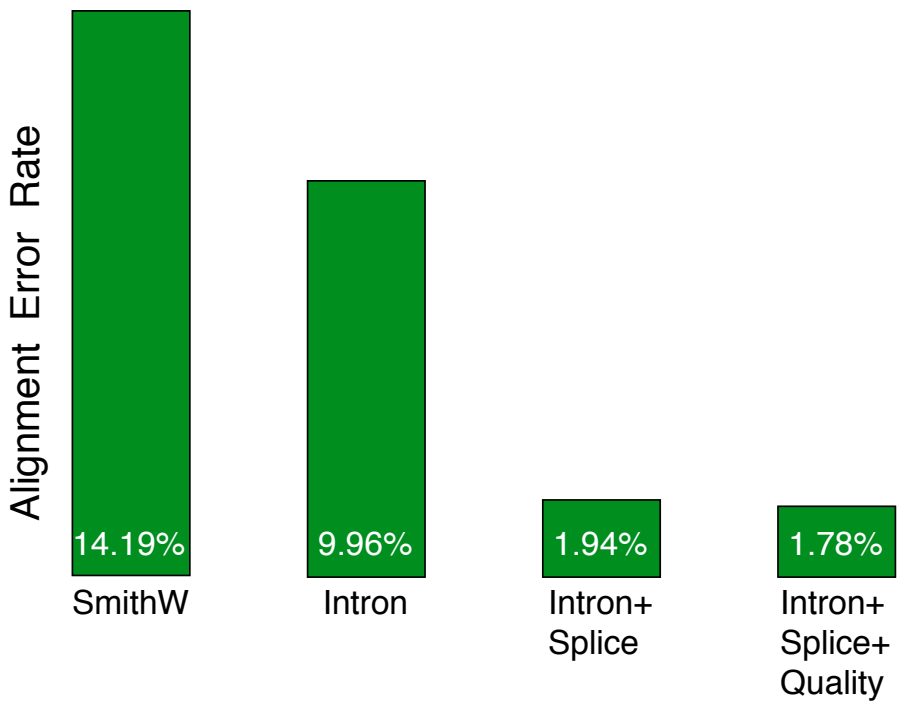
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \longrightarrow \mathbb{R}$ (De Bona et al., 2008)

RNA-Seq Read Alignment – 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

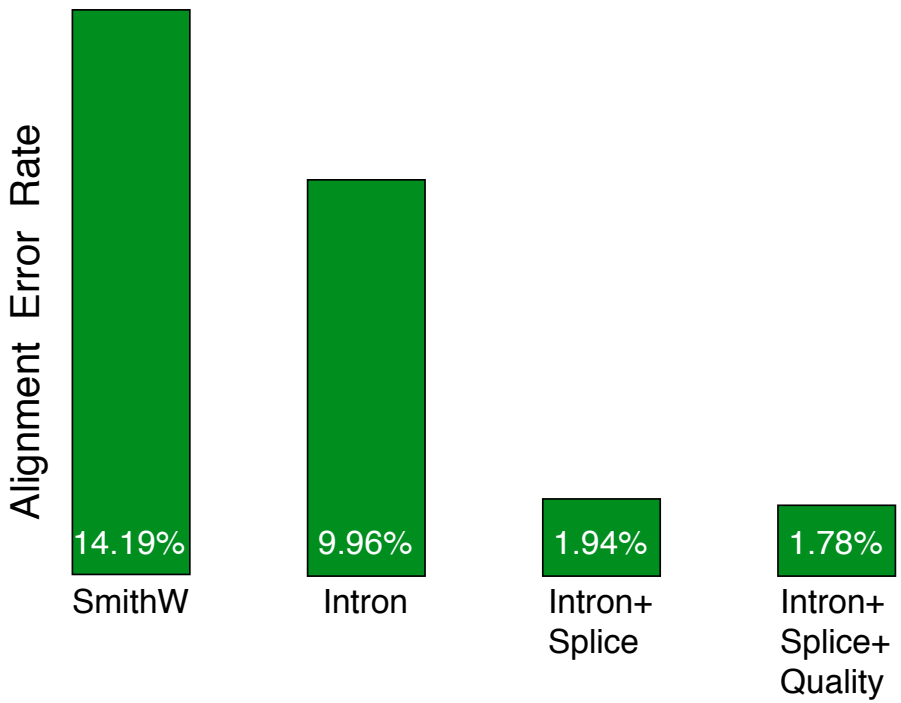


(De Bona et al., 2008)

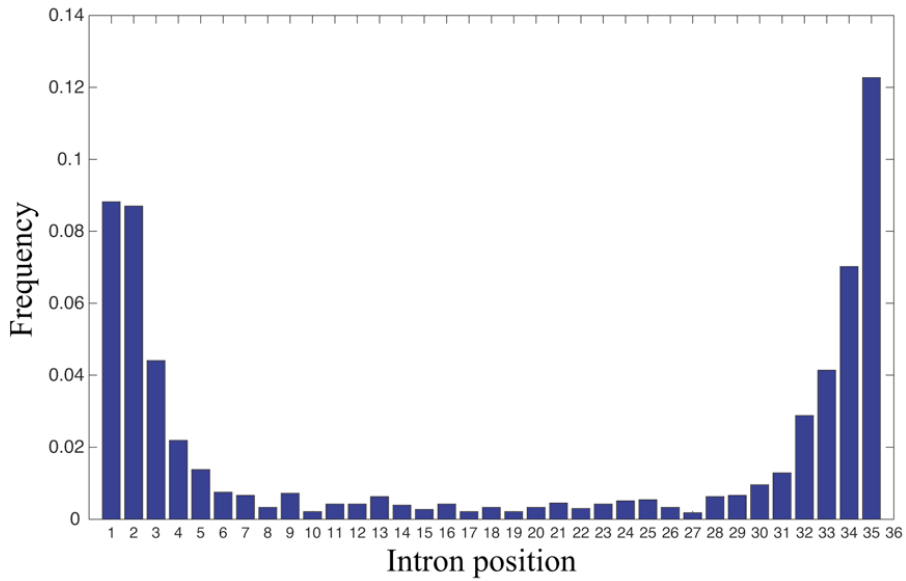
RNA-Seq Read Alignment – QPALMA

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

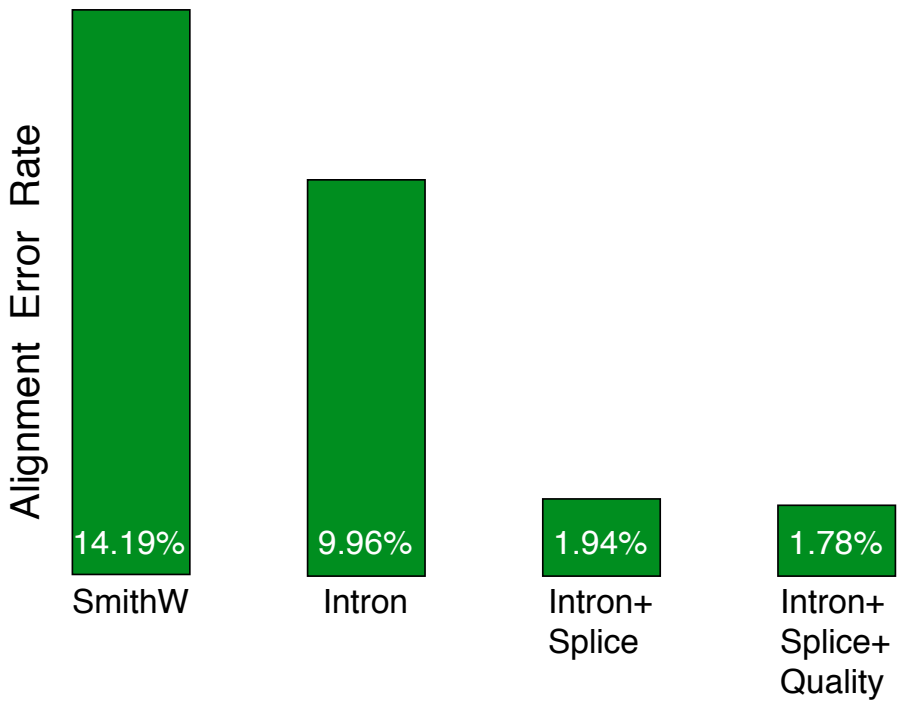


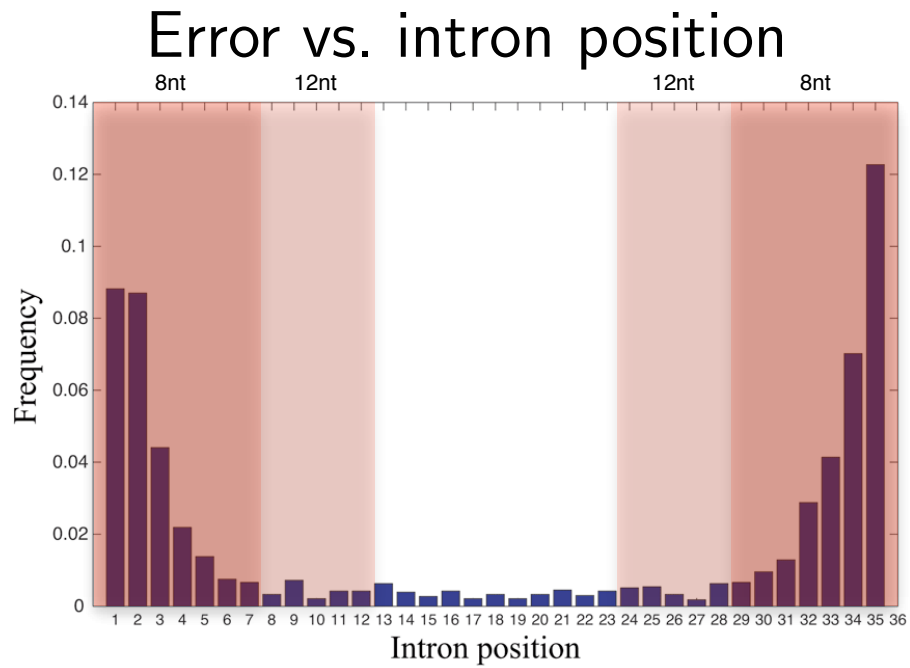
(De Bona et al., 2008)

RNA-Seq Read Alignment – 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





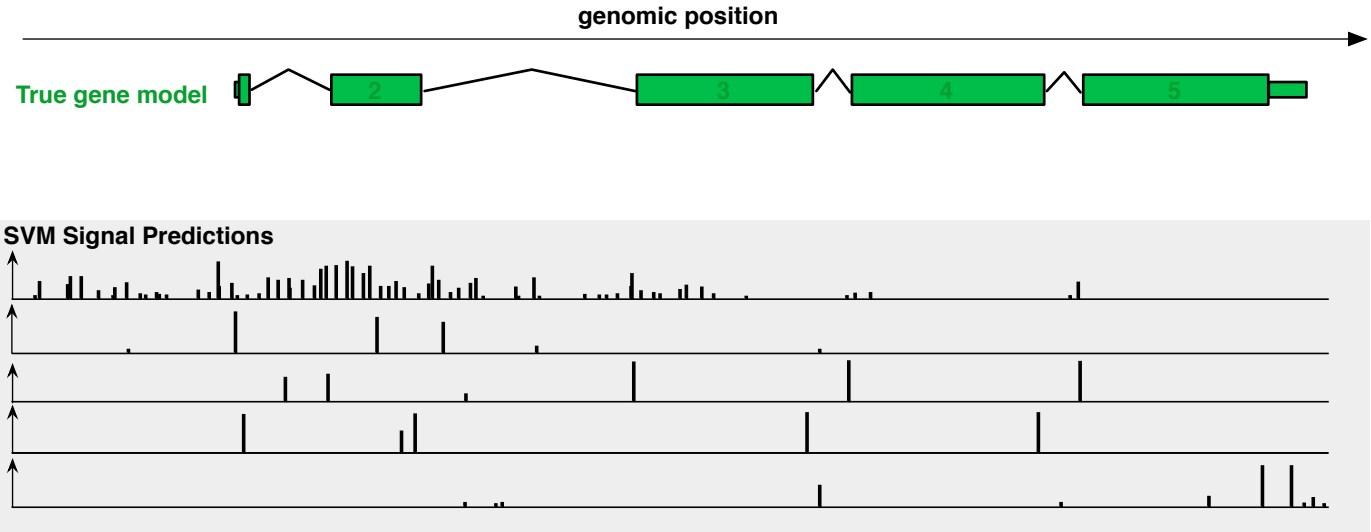
(De Bona et al., 2008)

Step 2: Transcript Prediction

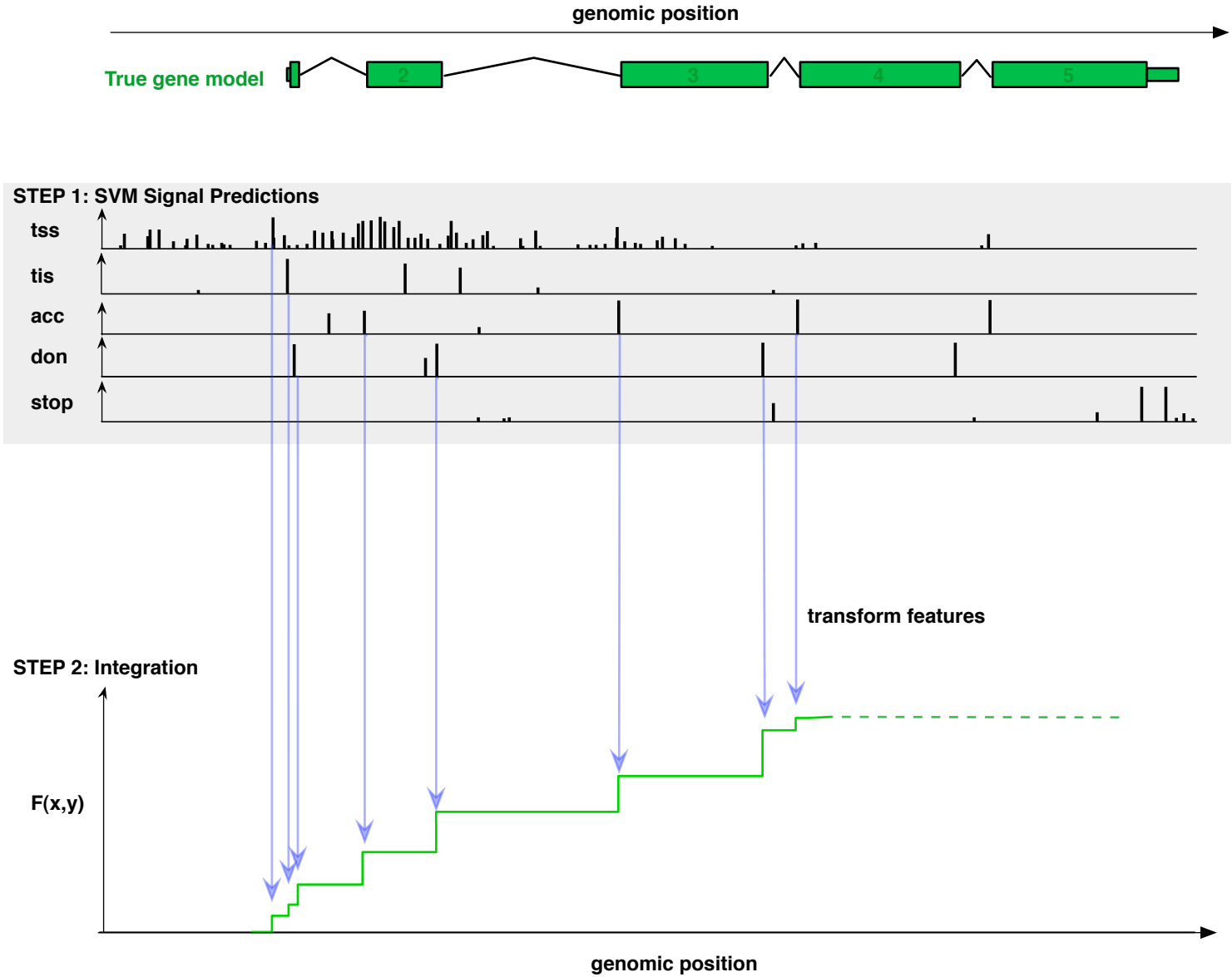
1. Extension of *mGene* gene finding system to use NGS data for protein coding transcript prediction
2. Coverage segmentation algorithm *mTIM* for general transcripts (no coding bias/assumption)
3. Splice graph construction by extending splice graph with spliced reads (connecting exons)

Approaches 1 & 2 use read coverages and spliced reads.
Approach 3 uses existing transcripts and spliced reads.

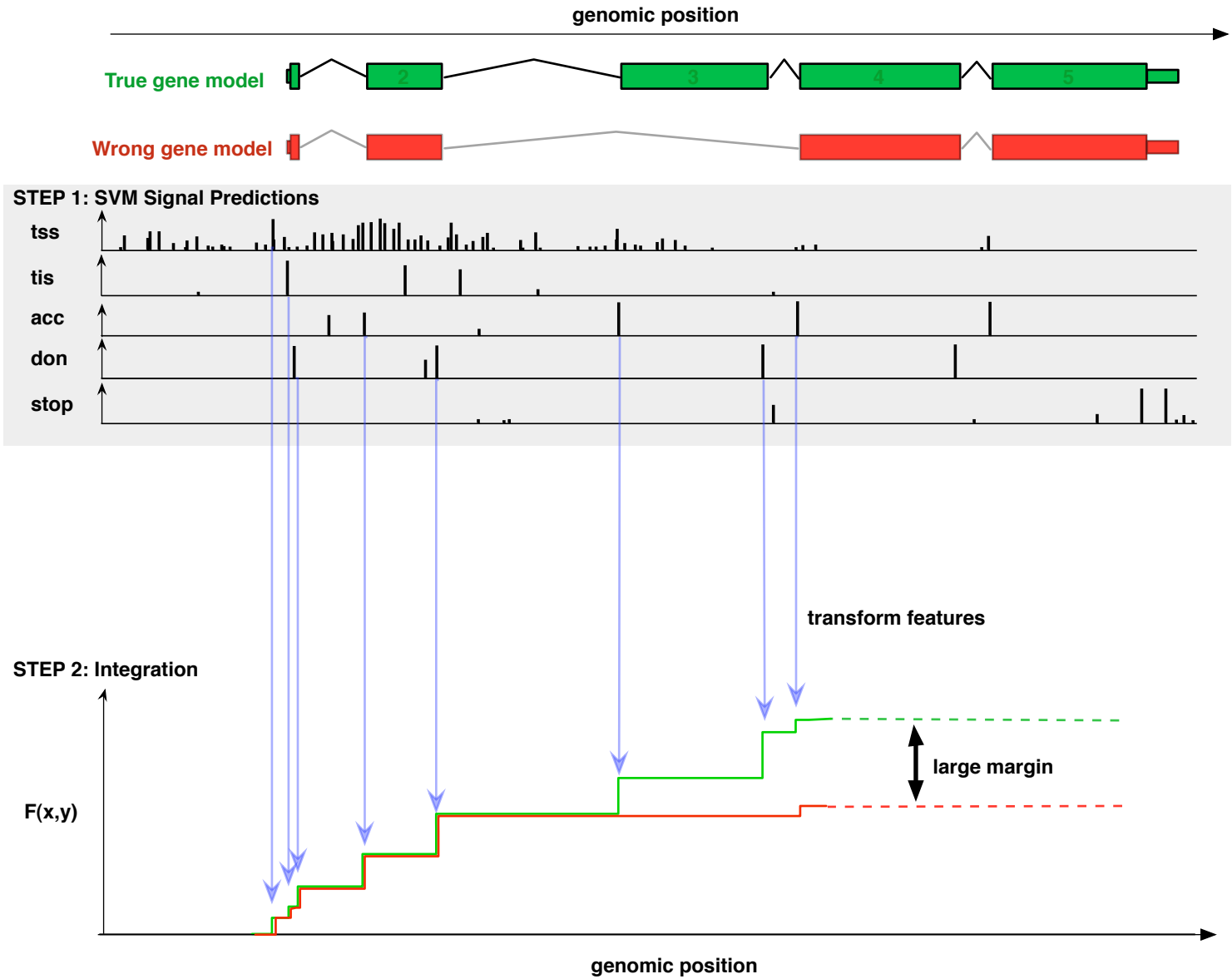
mGene-based Transcript Prediction I



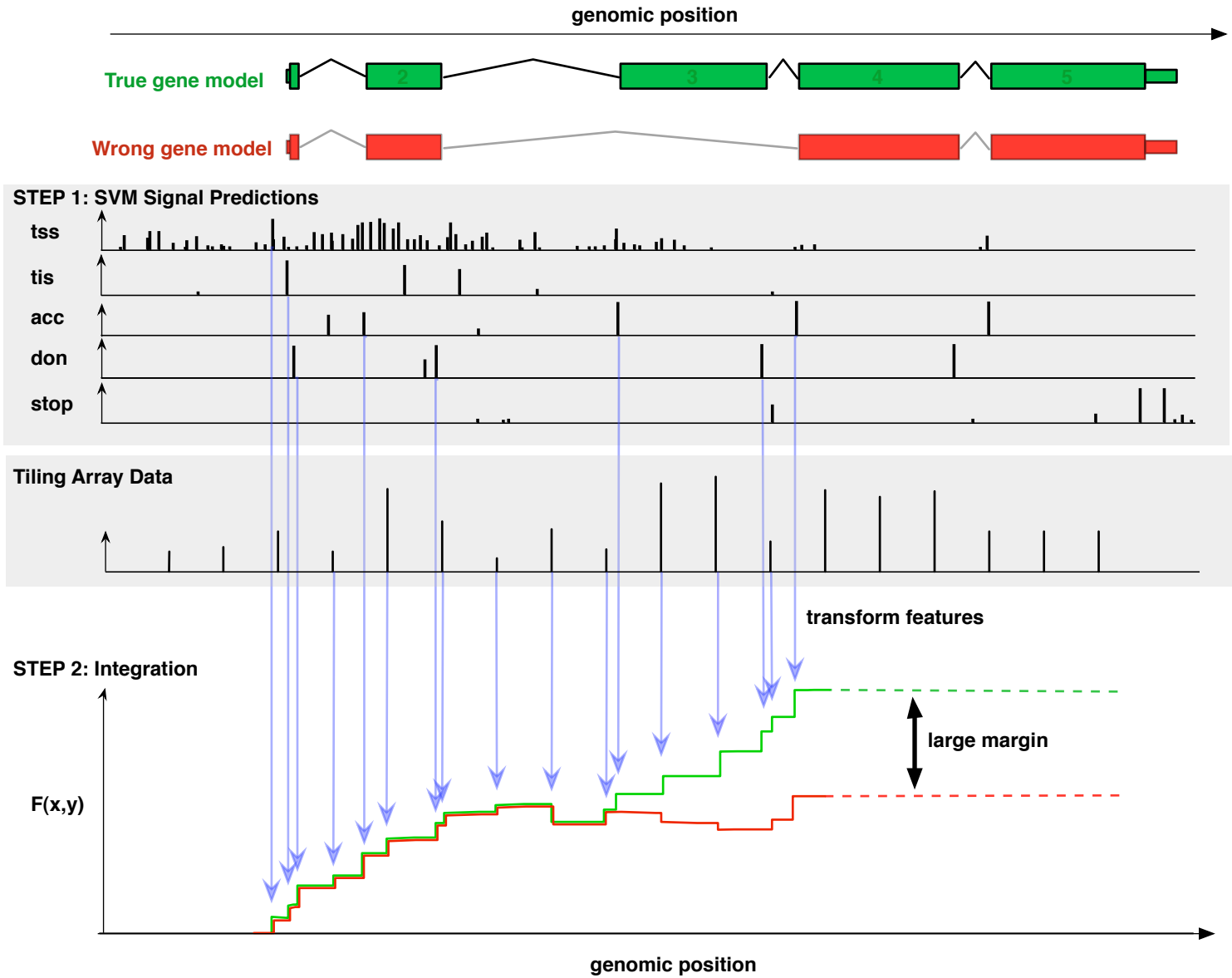
mGene-based Transcript Prediction I



mGene-based Transcript Prediction I



mGene-based Transcript Prediction I



mGene-based Transcript Prediction II

mGene with RNA-Seq (Behr et al., unpublished; Schweikert et al., 2009a,b)

- Use transcriptome measurements to enhance recognition of exonic regions

mGene-based Transcript Prediction II

mGene with RNA-Seq (Behr et al., unpublished; Schweikert et al., 2009a,b)

- Use transcriptome measurements to enhance recognition of exonic regions

Results for A. thaliana: (Comparison with known gene models)

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**

mGene-based Transcript Prediction II

mGene with RNA-Seq (Behr et al., unpublished; Schweikert et al., 2009a,b)

- Use transcriptome measurements to enhance recognition of exonic regions

Results for A. thaliana: (Comparison with known gene models)

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ...with tiling arrays (11 tissues)	**82.1%**
3. ...with mRNA-seq (1 tissue)	**81.1%**

mGene-based Transcript Prediction II

mGene with RNA-Seq (Behr et al., unpublished; Schweikert et al., 2009a,b)

- Use transcriptome measurements to enhance recognition of exonic regions

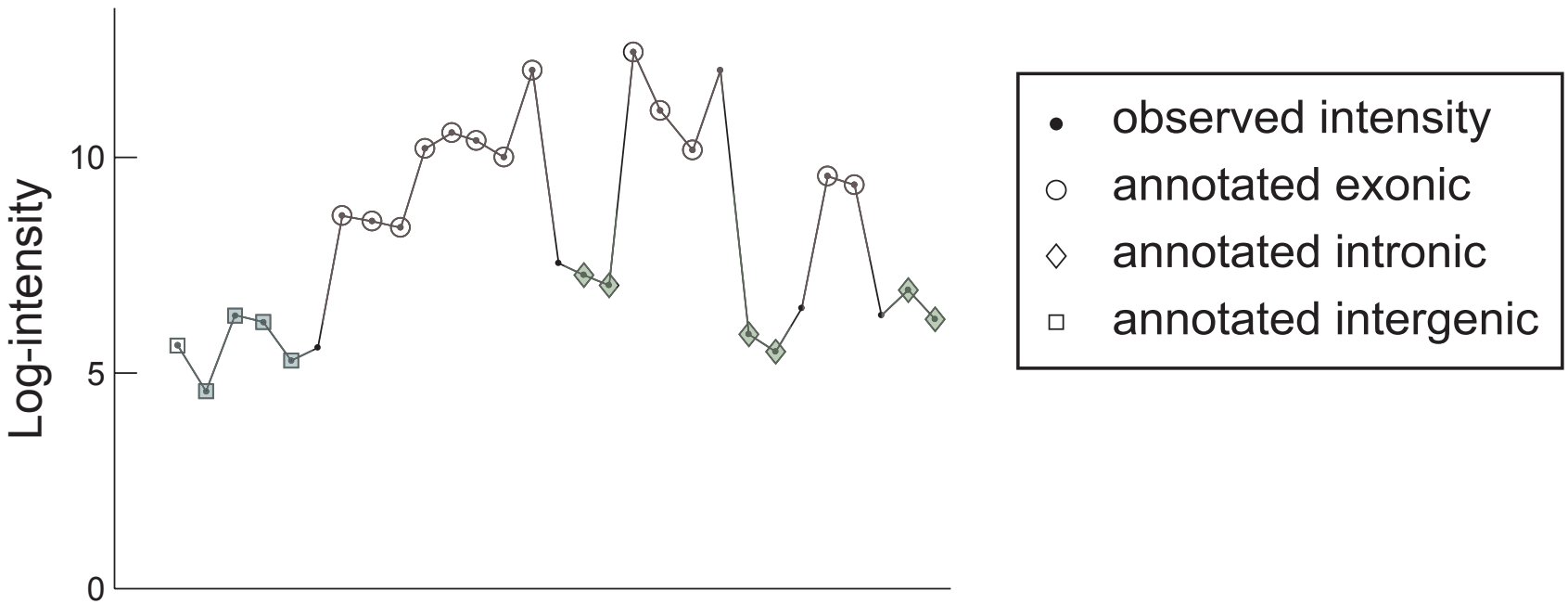
Results for A. thaliana: (Comparison with known gene models)

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ...with tiling arrays (11 tissues)	**82.1%**
3. ...with mRNA-seq (1 tissue)	**81.1%**

Similar observations for RGASP predictions.

Tiling Array/Read Coverage Segmentation

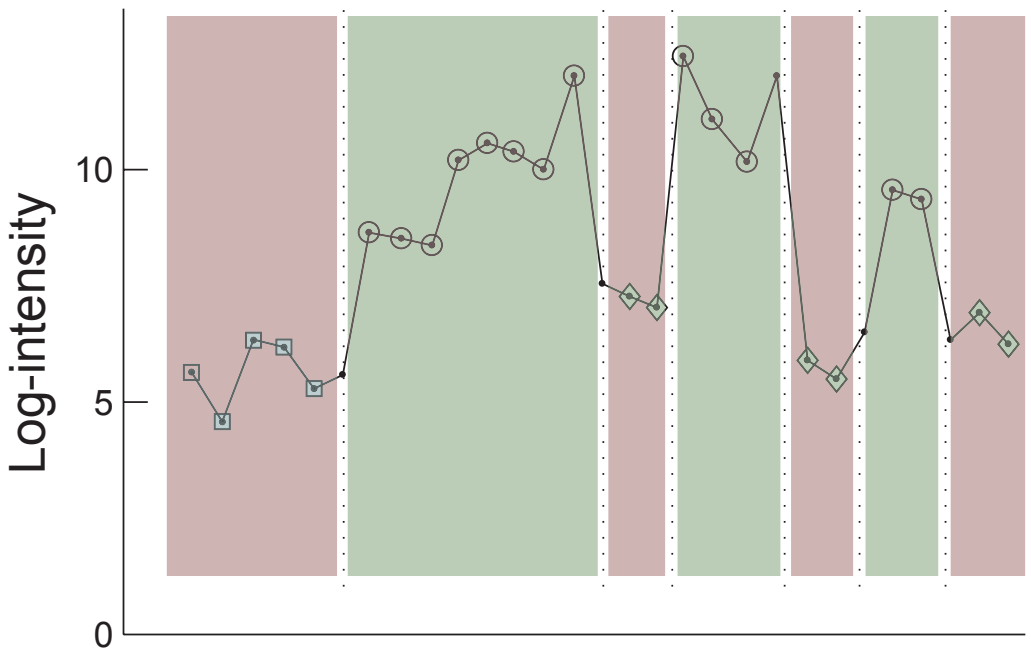
Goal: Characterize each "probe" as either intergenic, exonic or intronic



(Zeller et al., 2008a)

Tiling Array/Read Coverage Segmentation

Goal: Characterize each "probe" as either intergenic, exonic or intronic

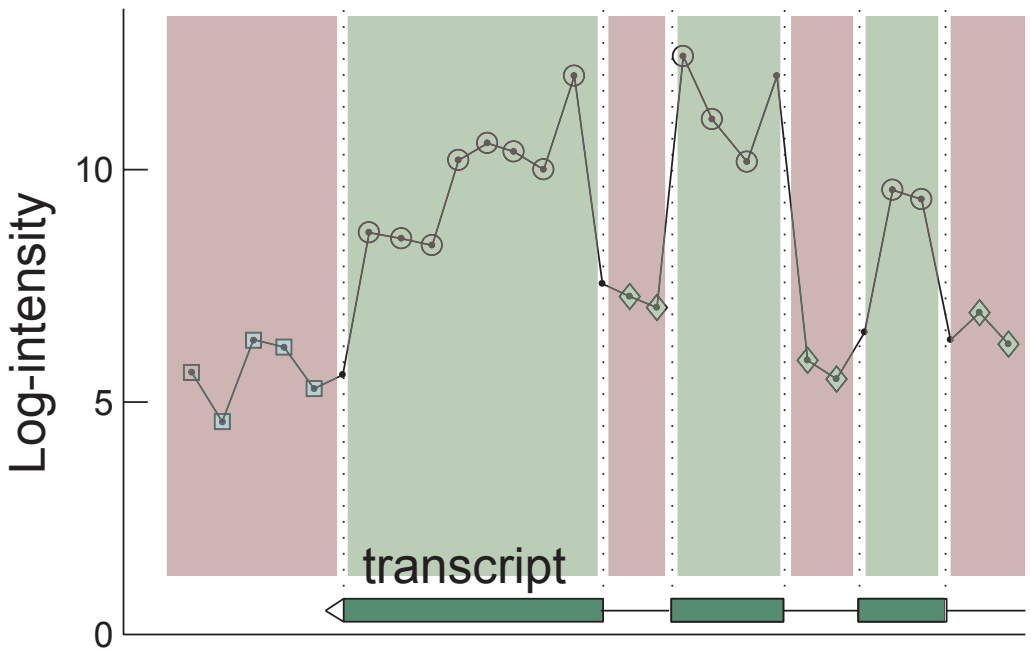


- observed intensity
- annotated exonic
- annotated intronic
- annotated intergenic

(Zeller et al., 2008a)

Tiling Array/Read Coverage Segmentation

Goal: Characterize each "probe" as either intergenic, exonic or intronic

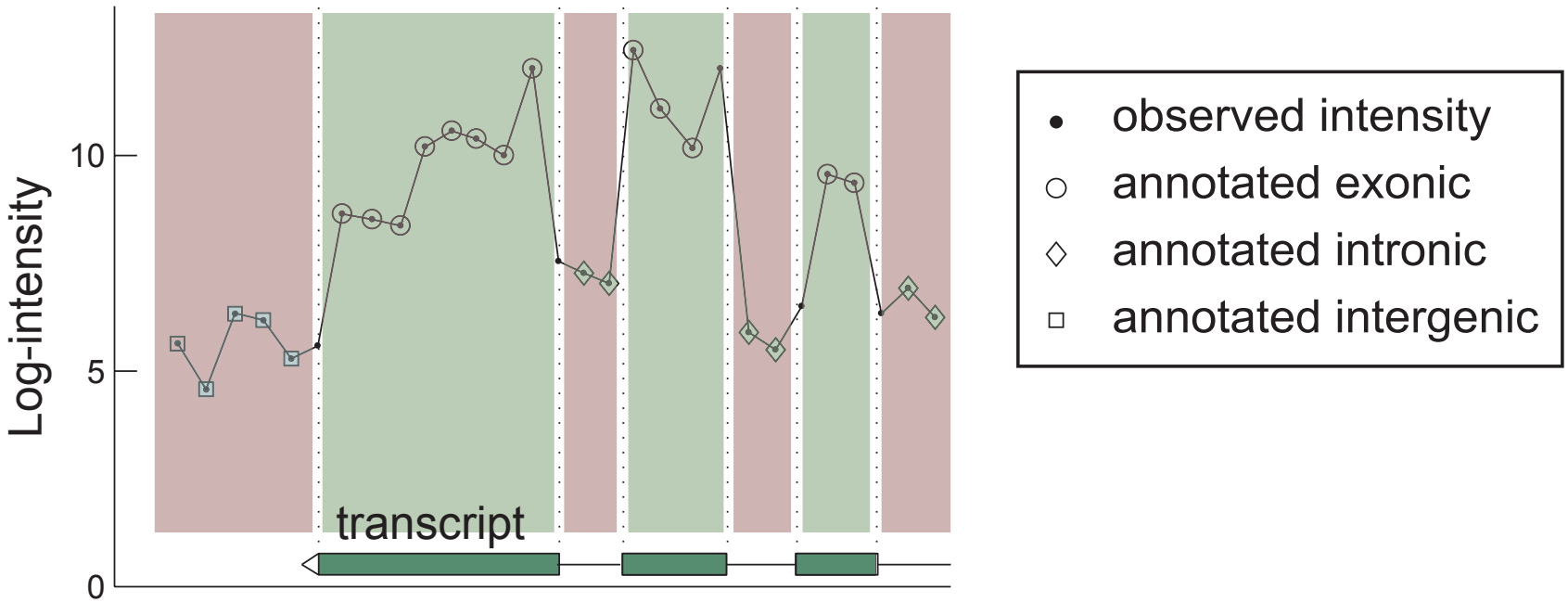


- observed intensity
- annotated exonic
- annotated intronic
- annotated intergenic

(Zeller et al., 2008a)

Tiling Array/Read Coverage Segmentation

Goal: Characterize each "probe" as either intergenic, exonic or intronic

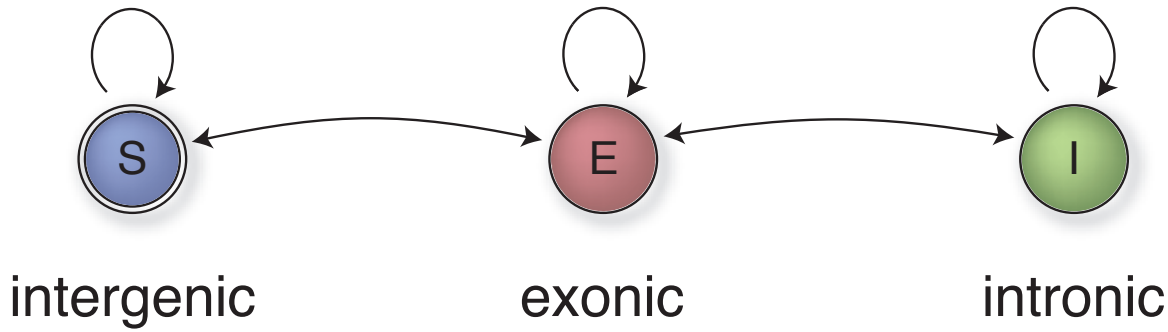


Novel segmentation method ("mSTAD"/"mTIM")

- accounts for spliced transcripts
- provides very accurate predictions

(Zeller et al., 2008a)

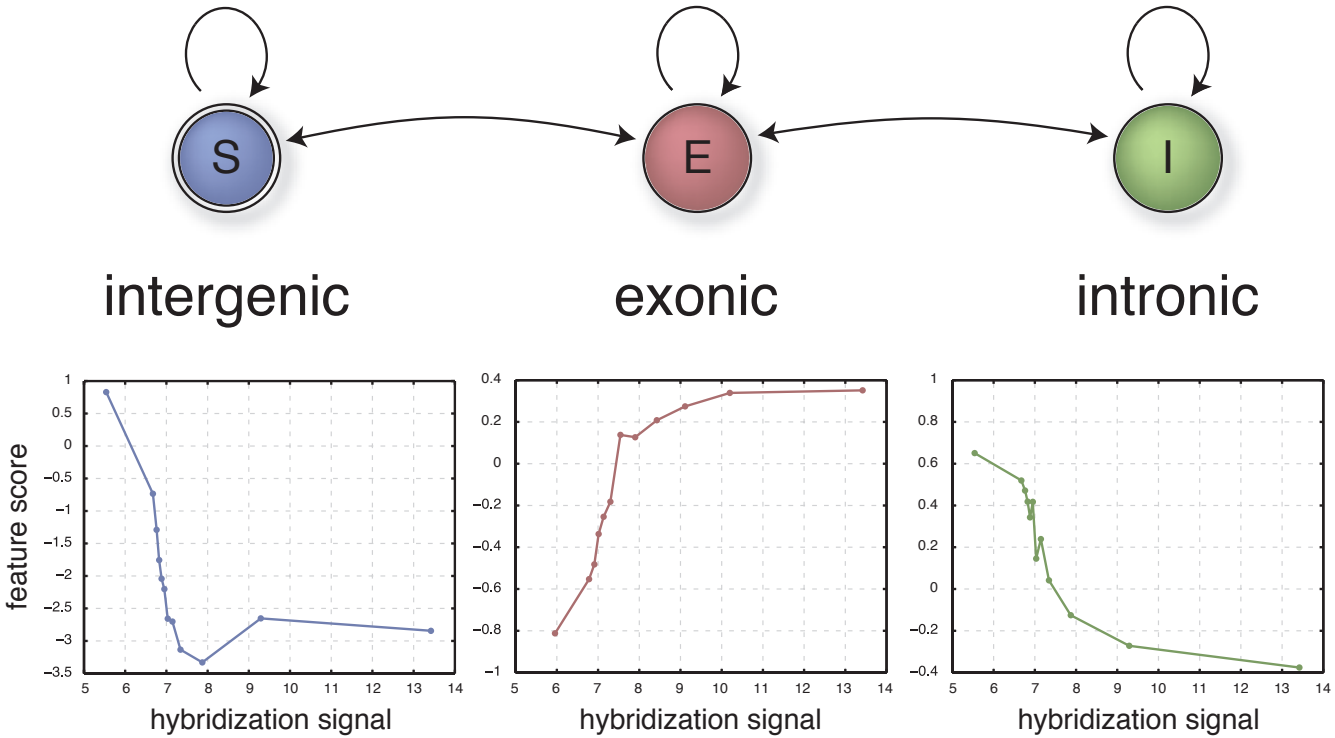
The mSTAD/mTIM Approach



- Learn to associate a state with each probe given its hybridization signal and local context

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

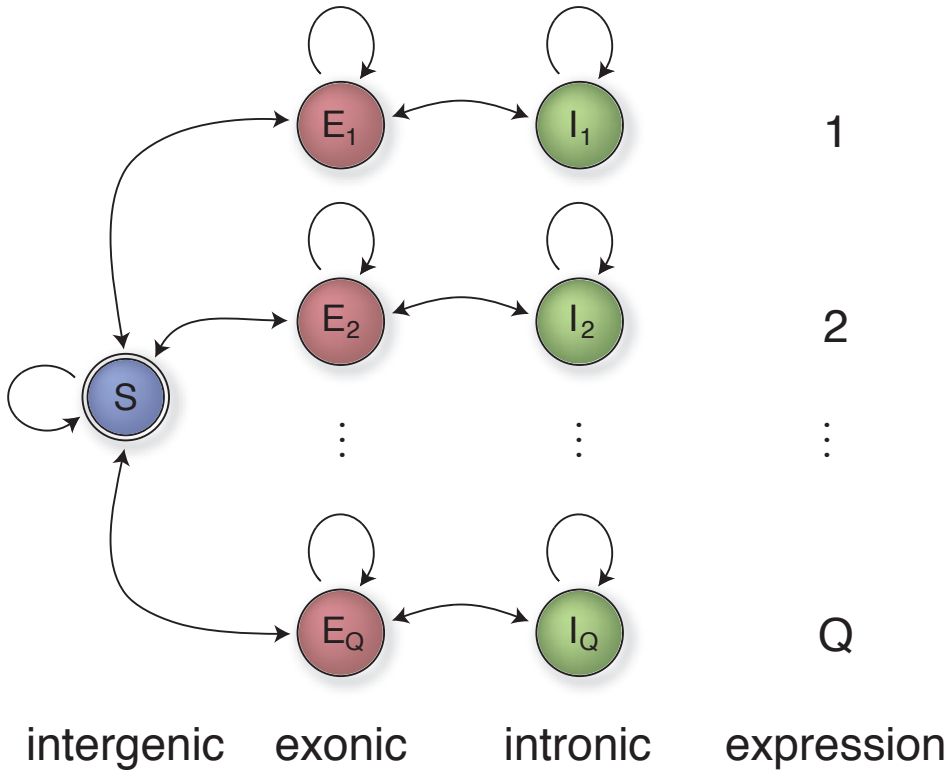
The mSTAD/mTIM Approach



- Learn to associate a state with each probe given its hybridization signal and local context
- For mTIM: also score spliced reads and splice sites

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

The mSTAD/mTIM Approach



- Learn to associate a state with each probe given its hybridization signal and local context
- For mTIM: also score spliced reads and splice sites
- HM-SVM training: Optimize transformations: signal $\longrightarrow$ score

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

Digestion ...

- *mGene* and *mTIM* predict single transcripts (no alternative transcripts)

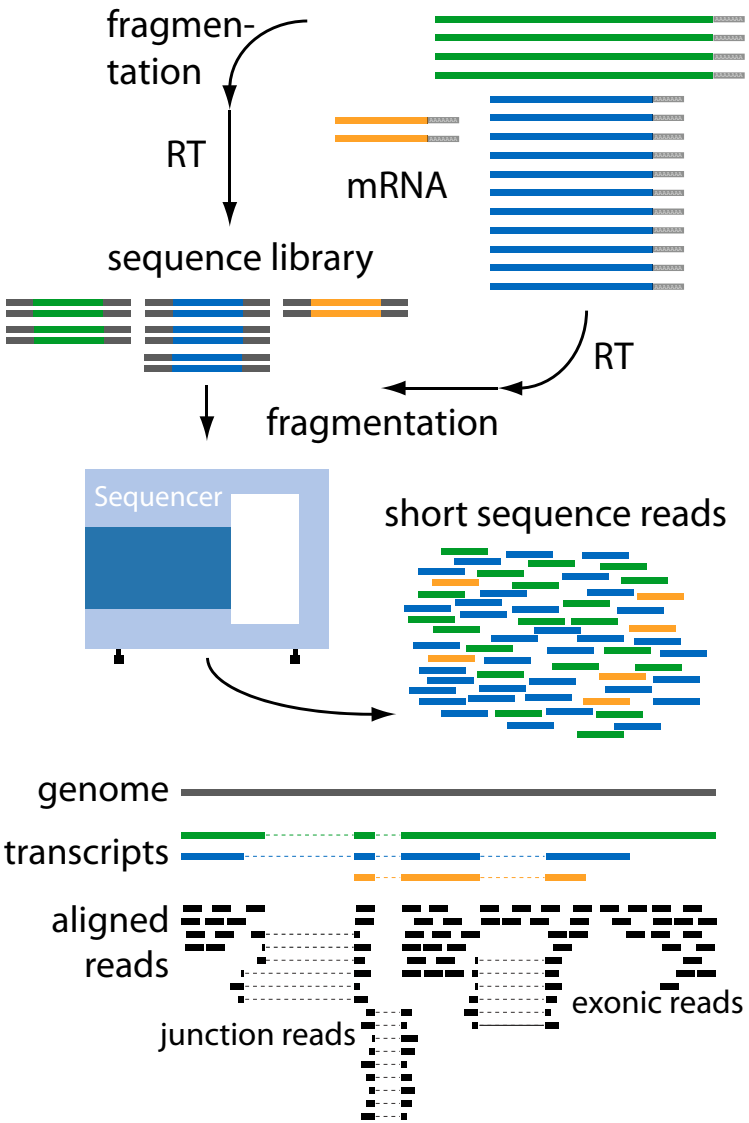
Digestion ...

- *mGene* and *mTIM* predict single transcripts (no alternative transcripts)
- *mGene* uses more assumptions on structure of transcripts
- *mTIM* exploits "uniformity" read coverage among exons of same transcript

Digestion ...

- *mGene* and *mTIM* predict single transcripts (no alternative transcripts)
- *mGene* uses more assumptions on structure of transcripts
- *mTIM* exploits "uniformity" read coverage among exons of same transcript
- Spliced reads used to generate a more complete splicing graph
- Paths through splicing graph define transcripts for quantitation

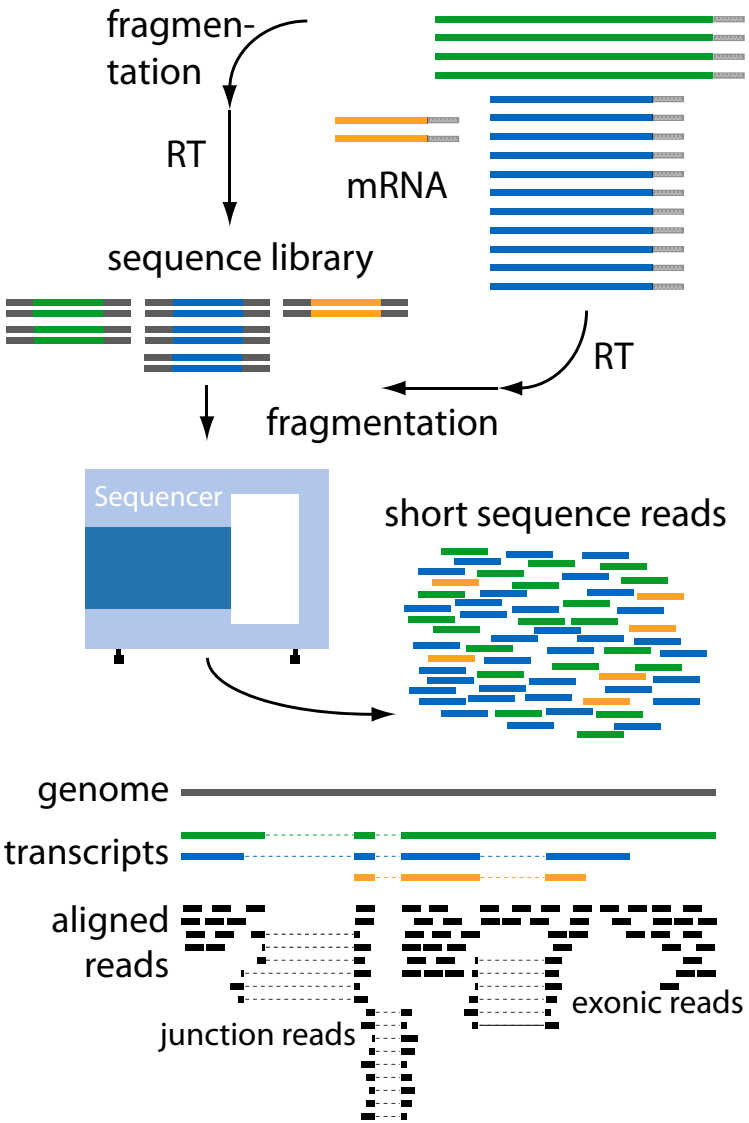
RNA-Seq Biases and Quantitation



Biases due to ...

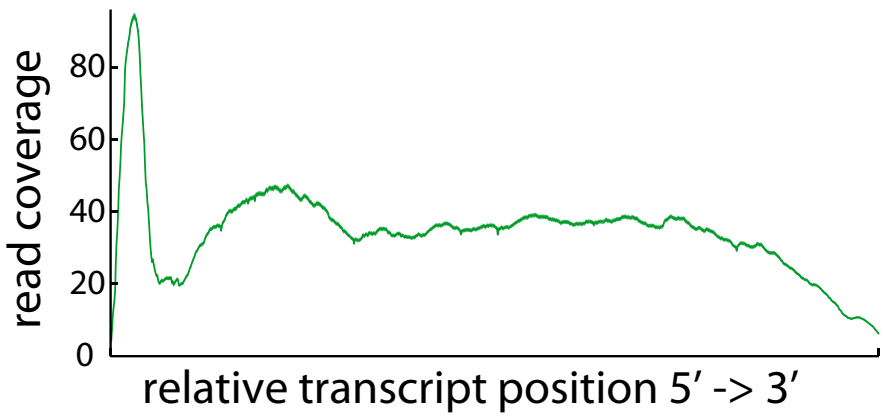
- cDNA library construction
- Sequencing
- Read mapping

RNA-Seq Biases and Quantitation



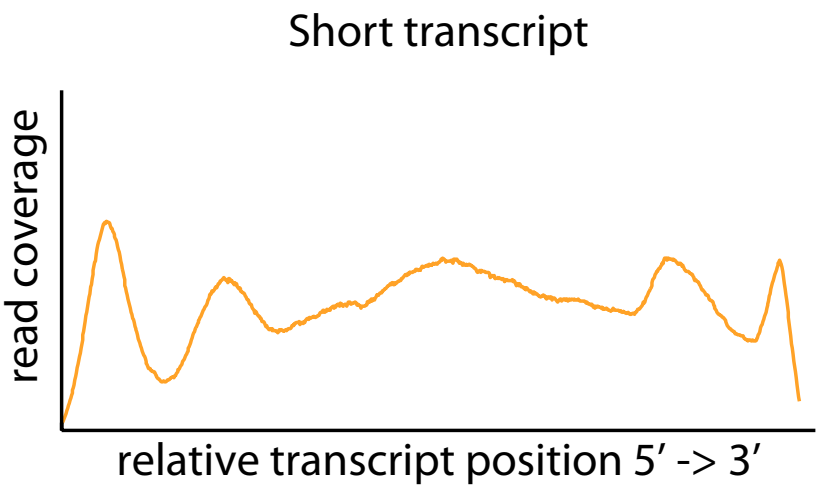
Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping

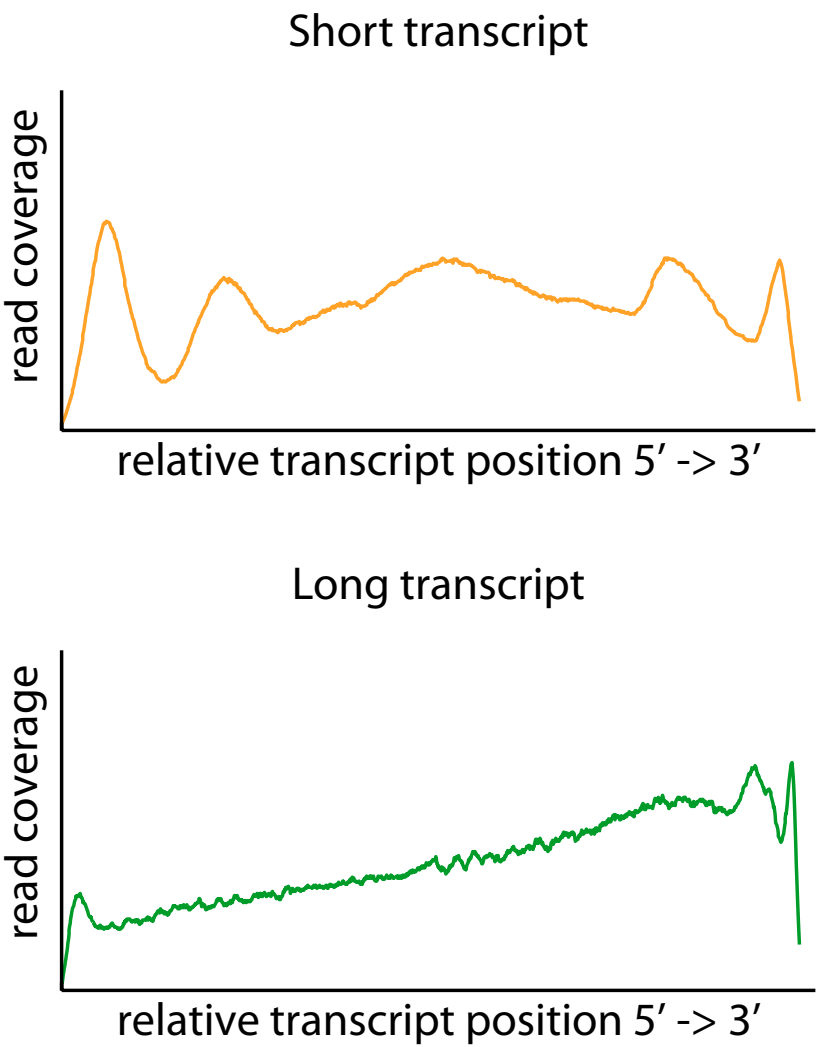


(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

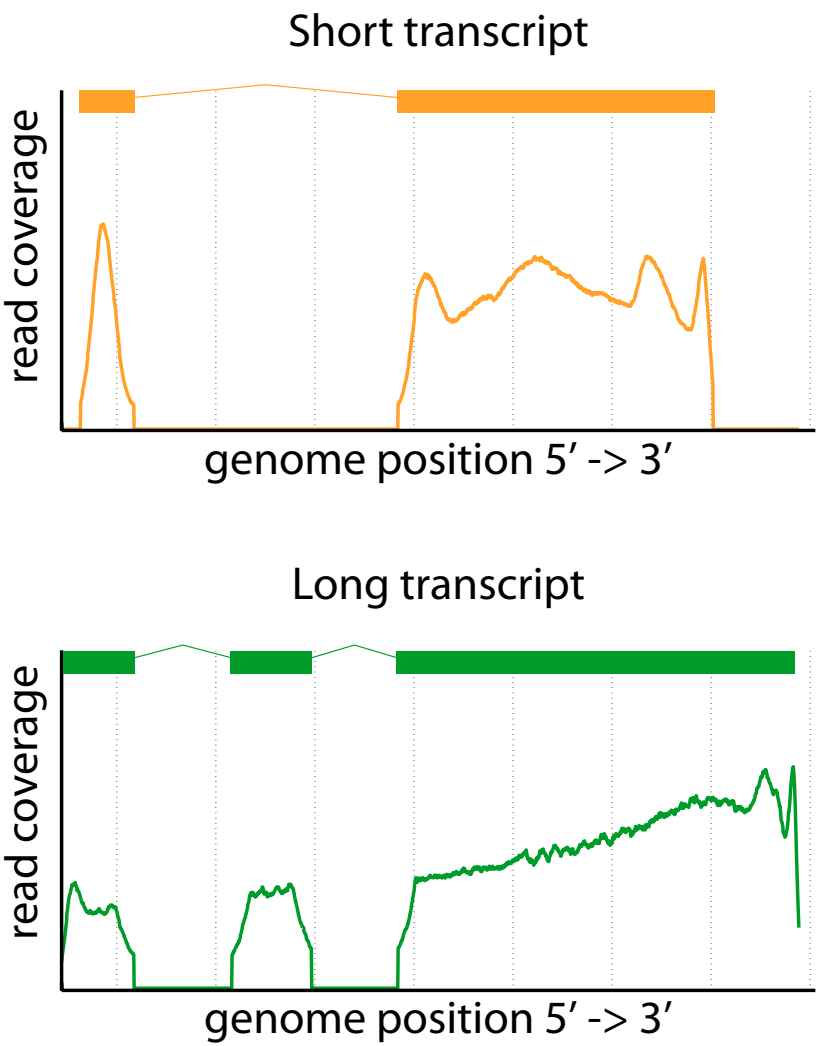
rQuant – Basic Idea



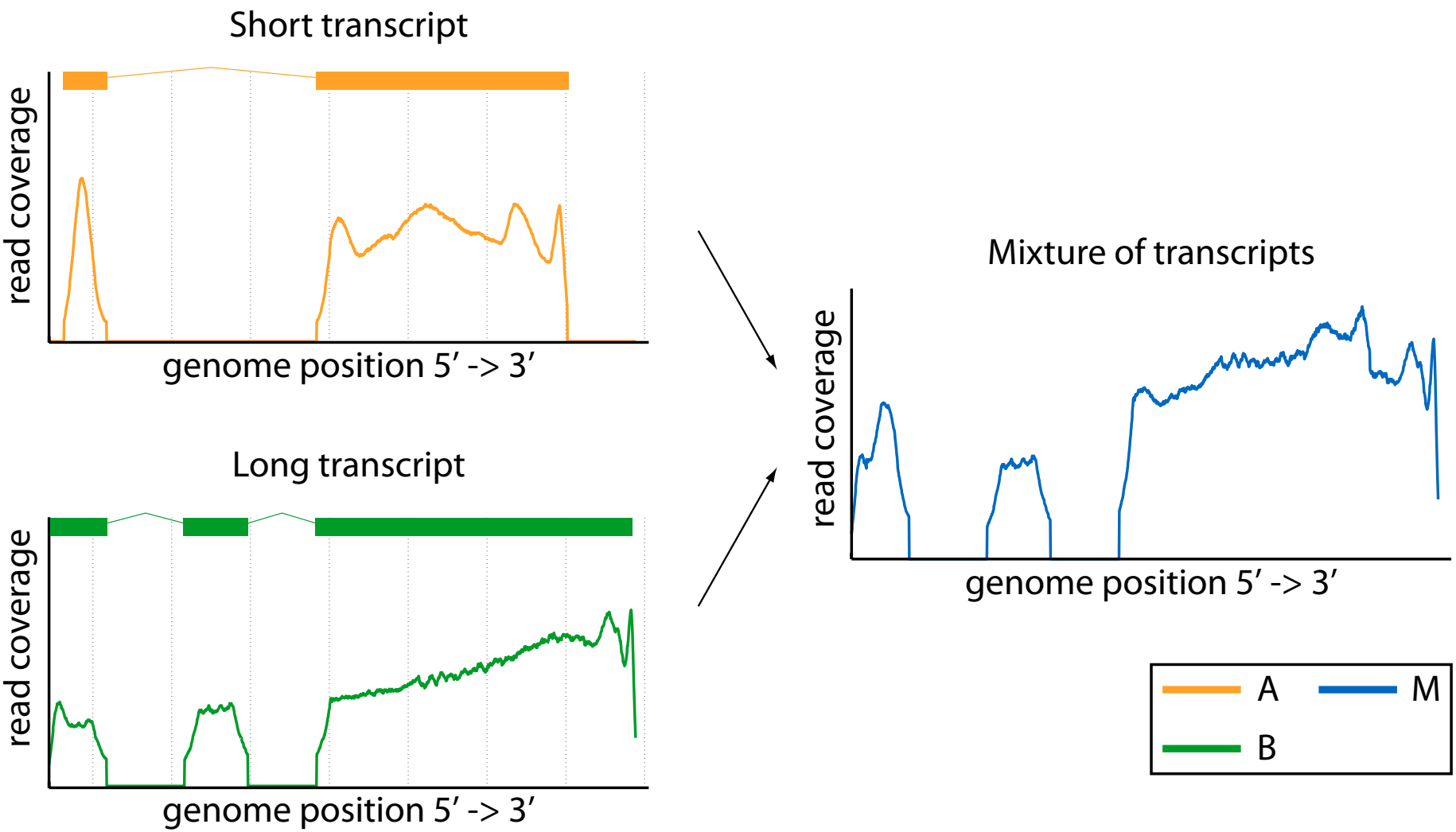
rQuant – Basic Idea



rQuant – Basic Idea

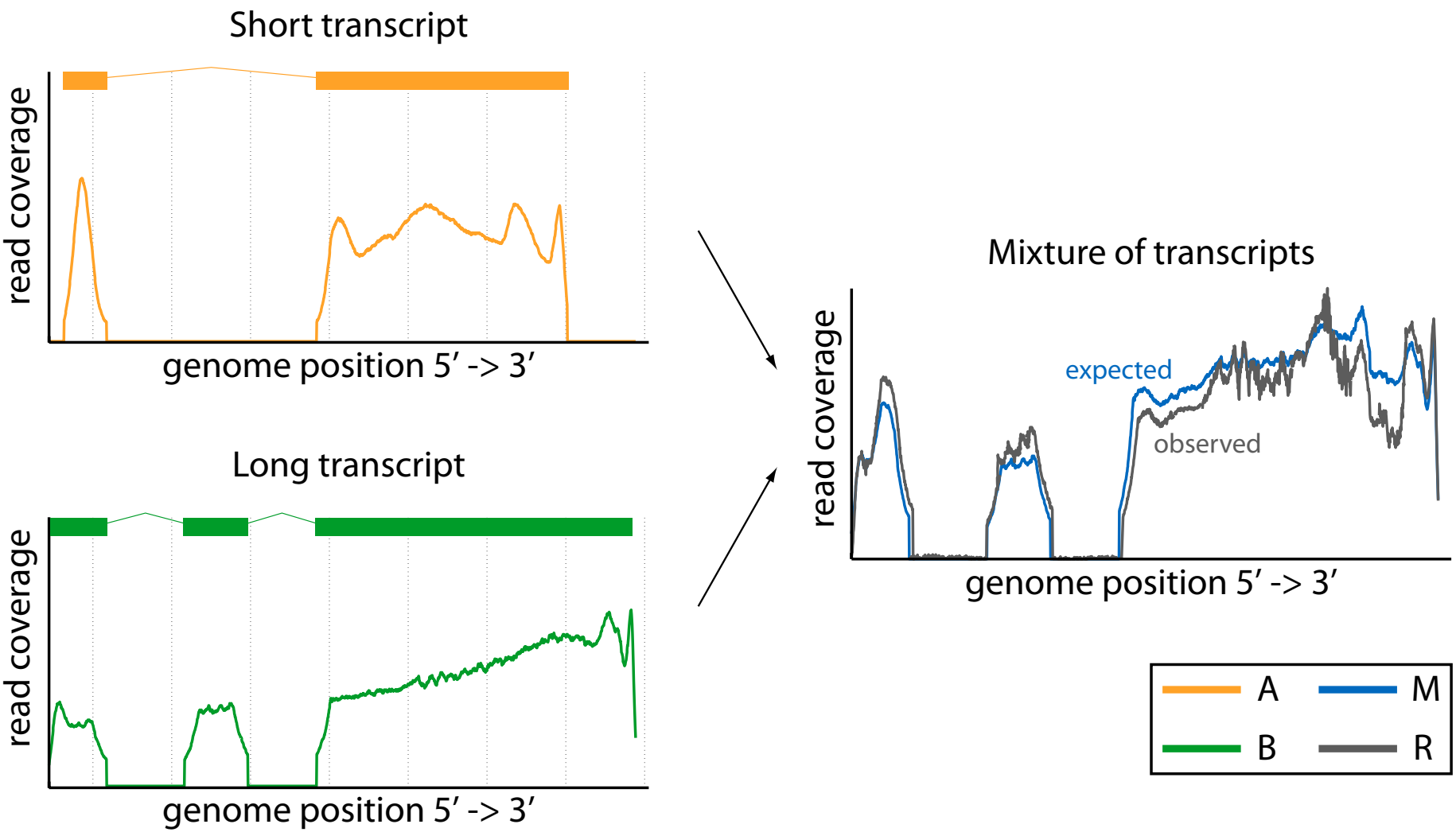


rQuant – Basic Idea



$$M_i = w_A A_i + w_B B_i$$

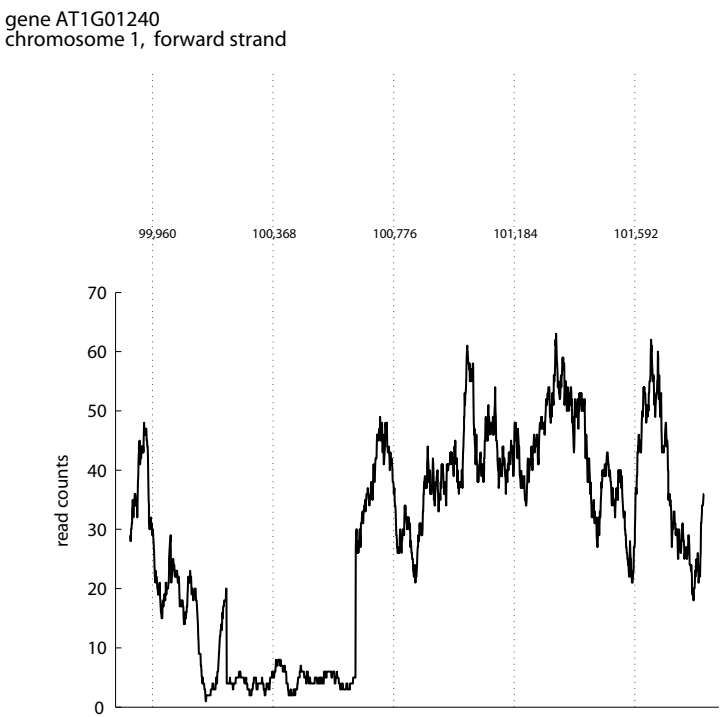
rQuant – Basic Idea



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

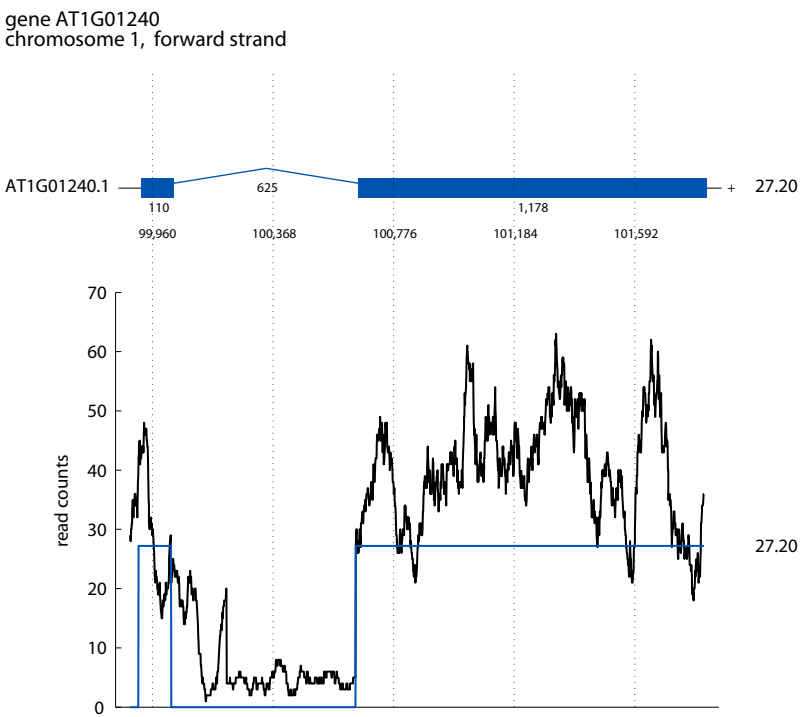
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$



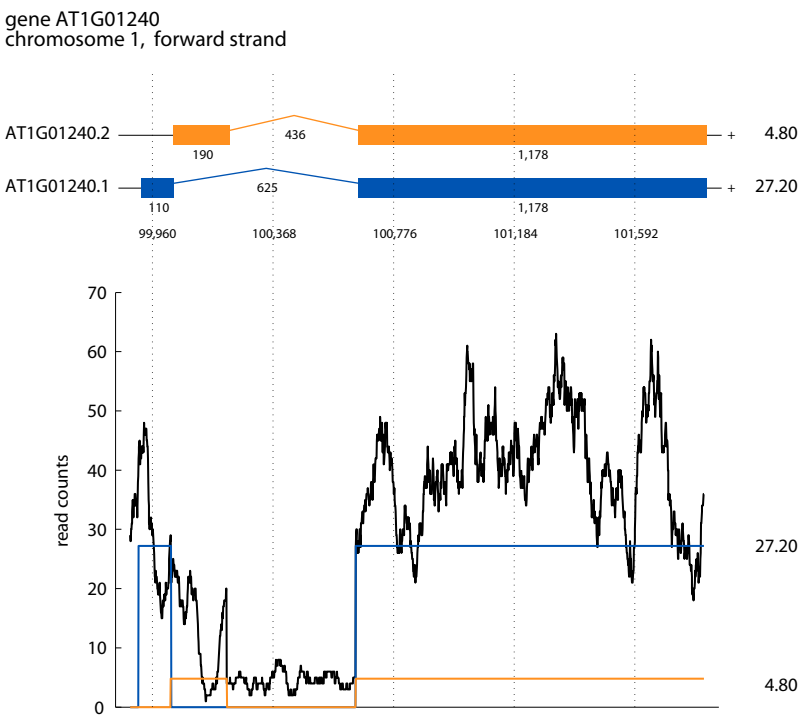
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$



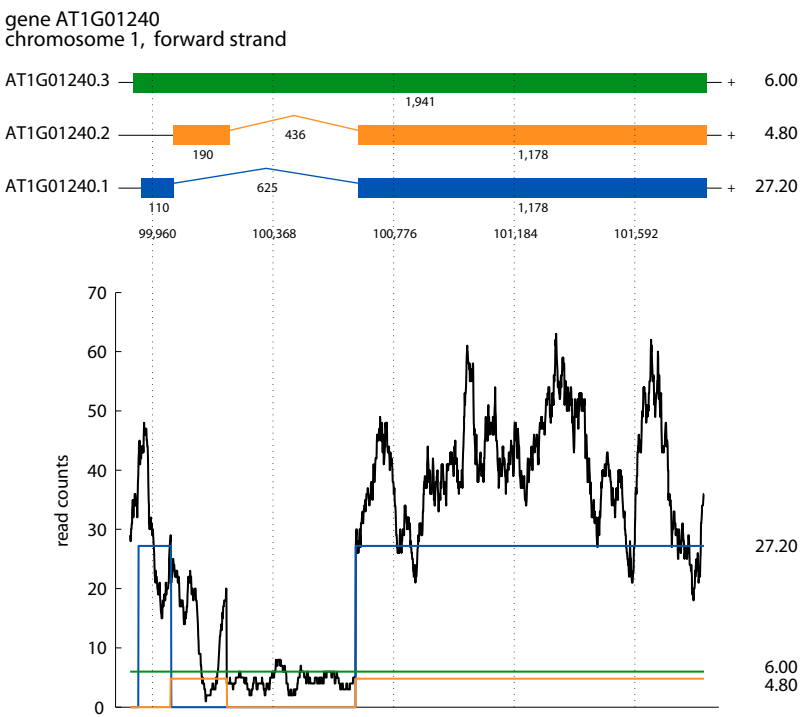
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$



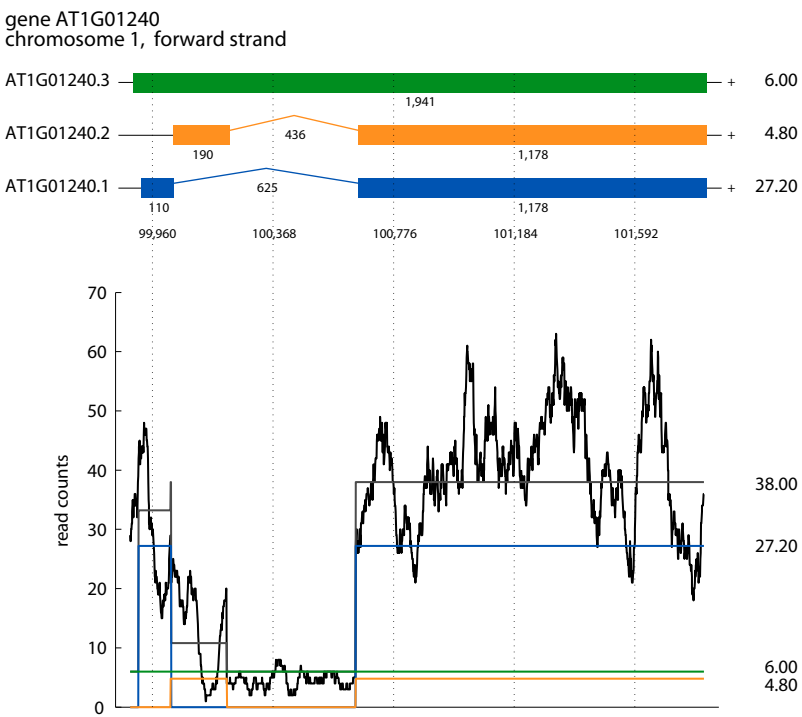
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$



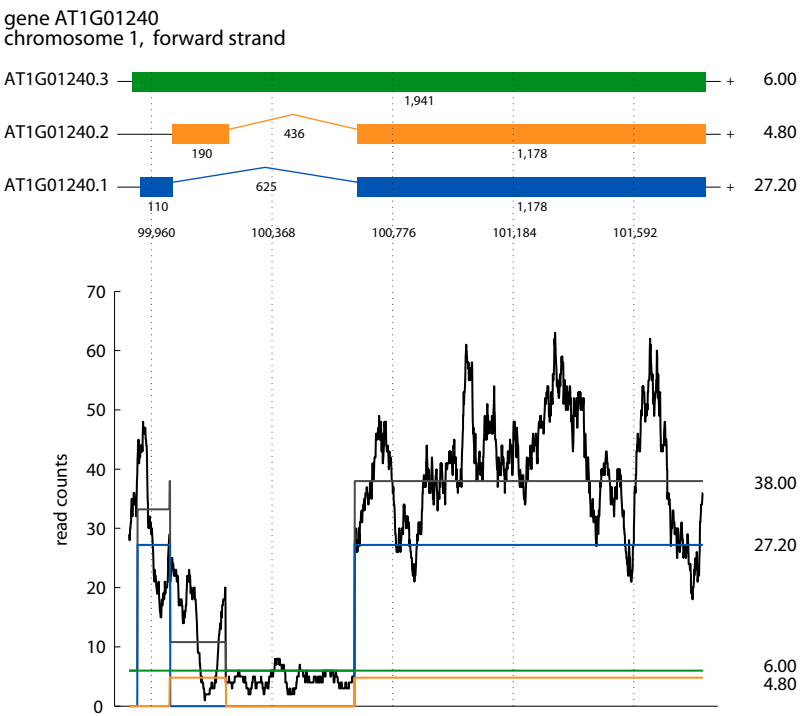
rQuant – Iterative Algorithm

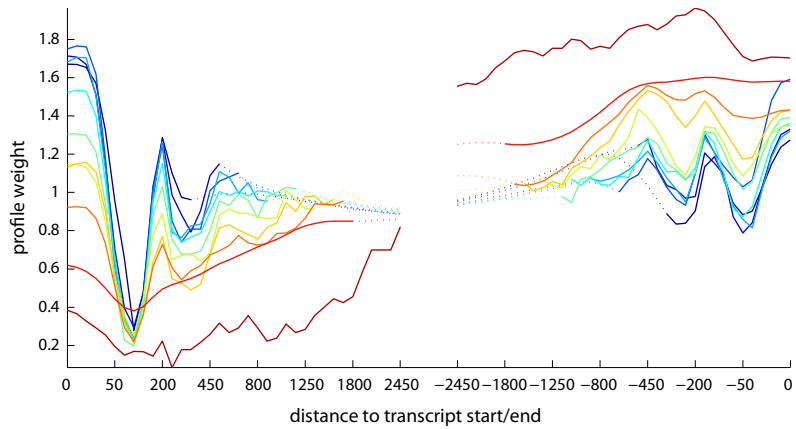
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$



rQuant – Iterative Algorithm

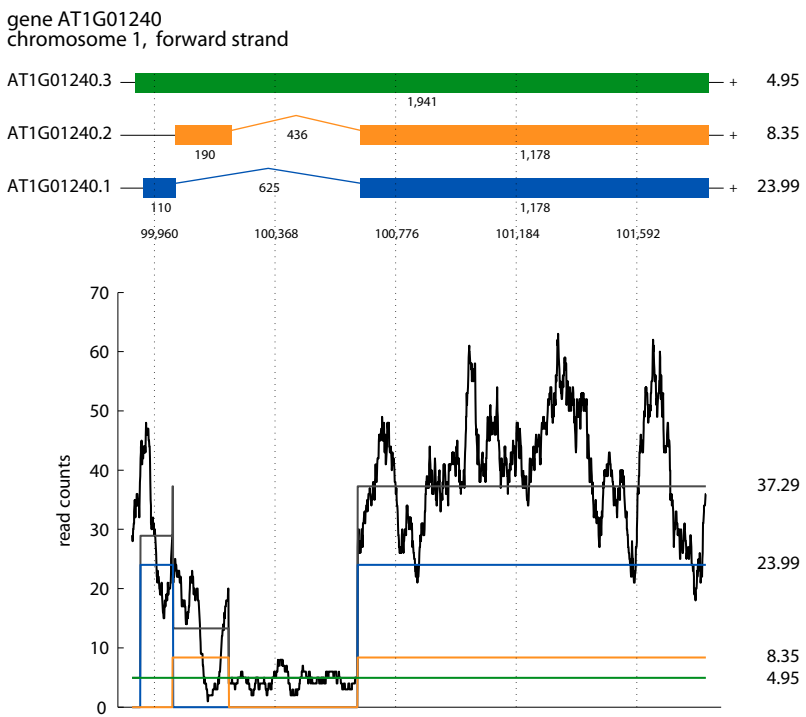
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$

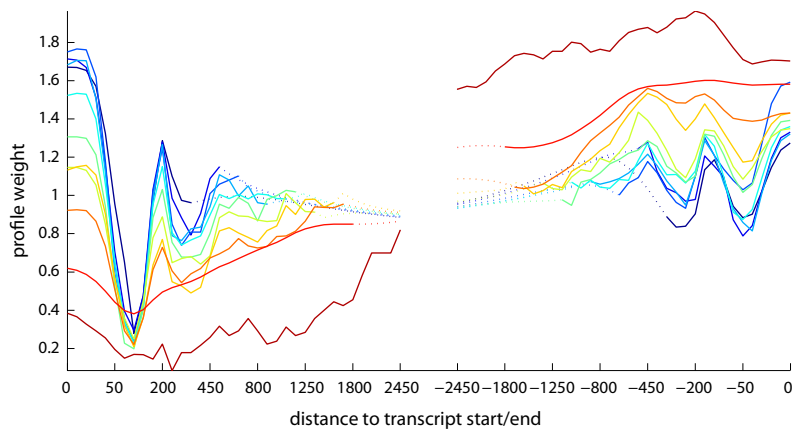




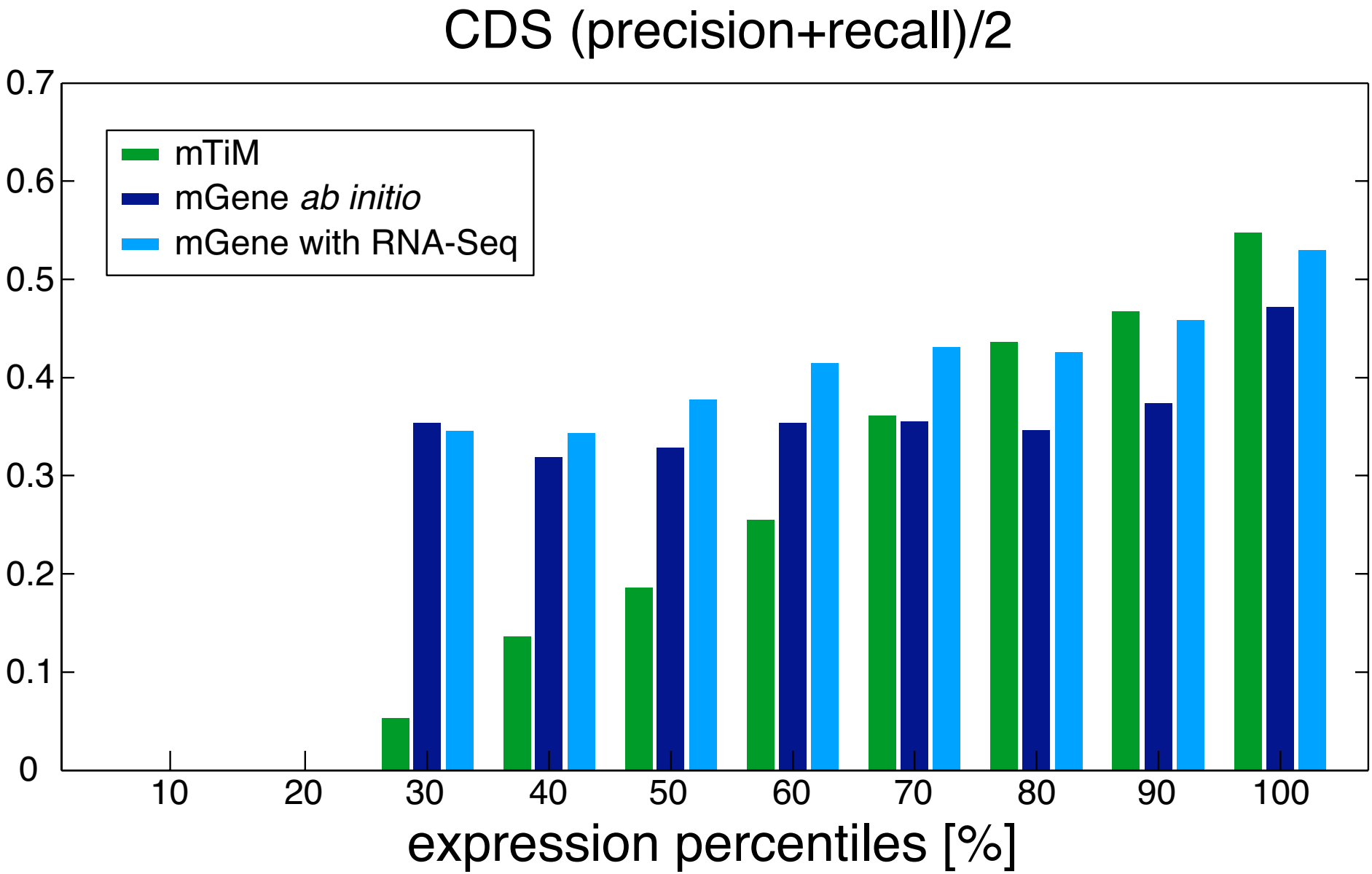
rQuant – Iterative Algorithm

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.

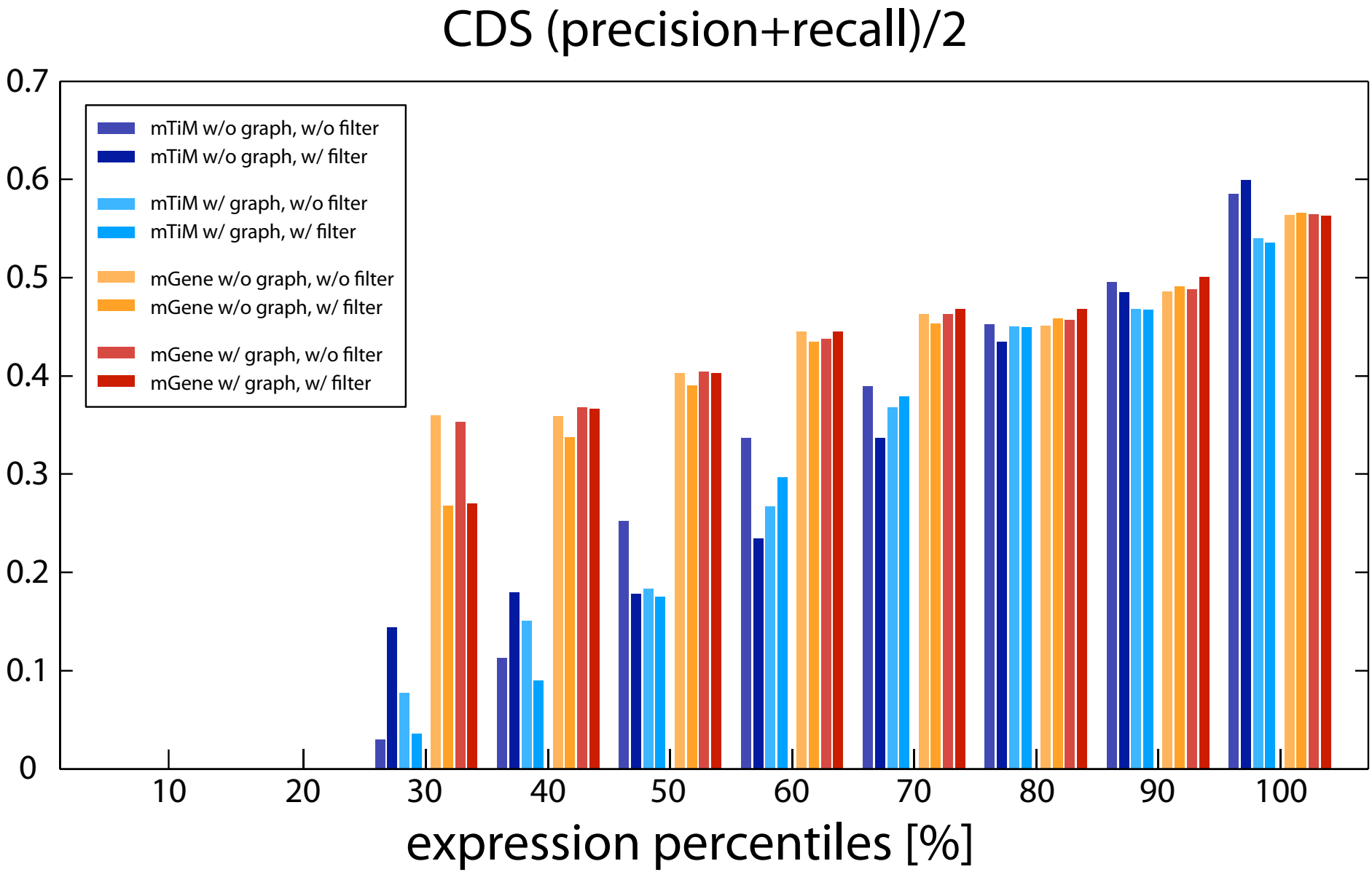




Preliminary Evaluation I



Preliminary Evaluation II



Conclusions

- GenomeMapper/QPalma
  - Splice site predictions improve alignment performance
  - Integrating QPALMA scoring into other read mappers promising

Conclusions

- GenomeMapper/QPalma
  - Splice site predictions improve alignment performance
  - Integrating QPALMA scoring into other read mappers promising
- mGene
  - Higher recall
  - Identifies also non-expressed genes $\Rightarrow$ good for annotation

Conclusions

- GenomeMapper/QPalma
  - Splice site predictions improve alignment performance
  - Integrating QPALMA scoring into other read mappers promising
- mGene
  - Higher recall
  - Identifies also non-expressed genes $\Rightarrow$ good for annotation
- mTIM
  - Higher precision
  - Better for identifying transcripts specific to experimental data

Conclusions

- GenomeMapper/QPalma
  - Splice site predictions improve alignment performance
  - Integrating QPALMA scoring into other read mappers promising
- mGene
  - Higher recall
  - Identifies also non-expressed genes $\Rightarrow$ good for annotation
- mTIM
  - Higher precision
  - Better for identifying transcripts specific to experimental data
- Adding alternative transcripts increases recall

Conclusions

- GenomeMapper/QPalma
  - Splice site predictions improve alignment performance
  - Integrating QPALMA scoring into other read mappers promising
- mGene
  - Higher recall
  - Identifies also non-expressed genes $\Rightarrow$ good for annotation
- mTIM
  - Higher precision
  - Better for identifying transcripts specific to experimental data
- Adding alternative transcripts increases recall
- rQuant-based filtering improves precision

Acknowledgements

RGASP Team

- Jonas Behr (FML)
- Georg Zeller (FML & MPI)
- Regina Bohnert (FML)

Funding by DFG &
Max Planck Society.

Thank you for your attention.

References

Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics (Oxford, England)*, 24(16):i174–180, August 2008.

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

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.

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 2009a. doi: 10.1186/gb-2009-10-9-r98. URL `http://genomebiology.com/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 Biology*, 10(9):R98, 2009b.

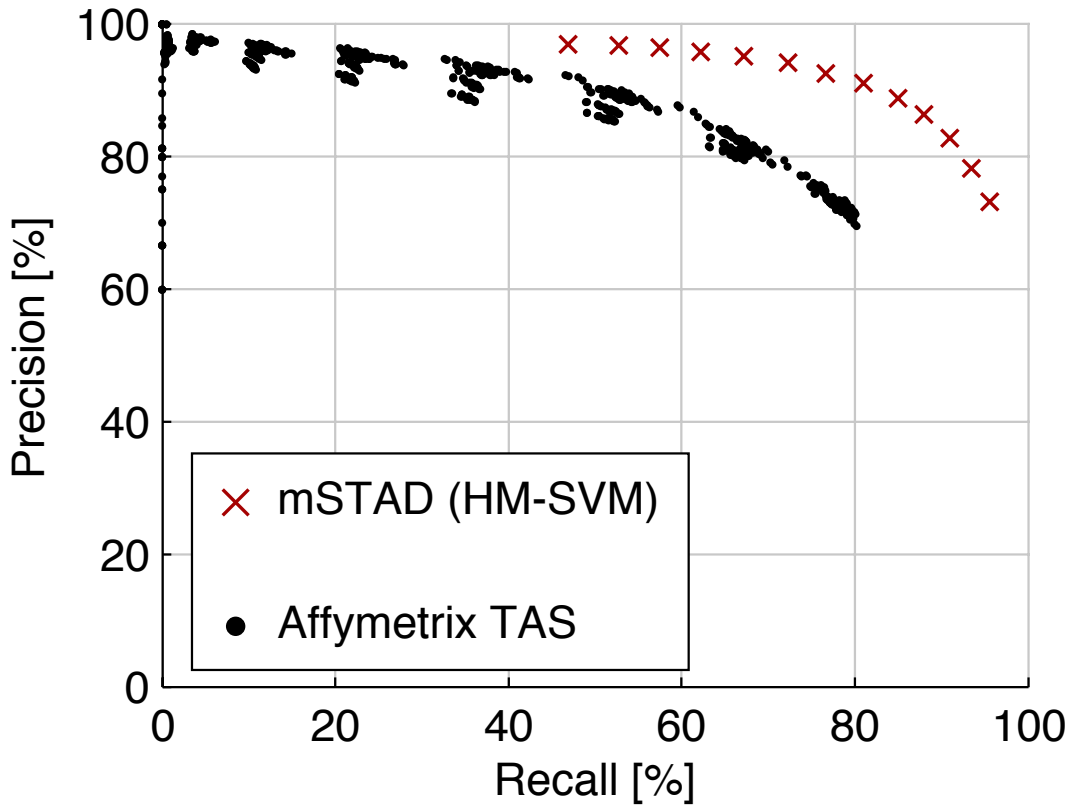
Gabriele Schweikert, Jonas Behr, Alexander Zien, Georg Zeller, Cheng Soon Ong, Sören Sonnenburg, and Gunnar Rätsch. mGene.web: a web service for accurate computational gene finding. *Nucleic Acids Research*, 37(Web Server issue): W312W316, July 2009a.

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*, September 2009b.

G. Zeller, S.R. Henz, S. Laubinger, D. Weigel, and G Rätsch. Transcript normalization and segmentation of tiling array data. In *Proceedings Pac. Symp. on Biocomputing*, pages 527–538, 2008a.

Georg Zeller, Stefan R. Henz, Sascha Laubinger, Detlef Weigel, and Gunnar Rätsch. Transcript normalization and segmentation of tiling array data. In *Proceedings Pac. Symp. on Biocomputing*, pages 527–538, 2008b.

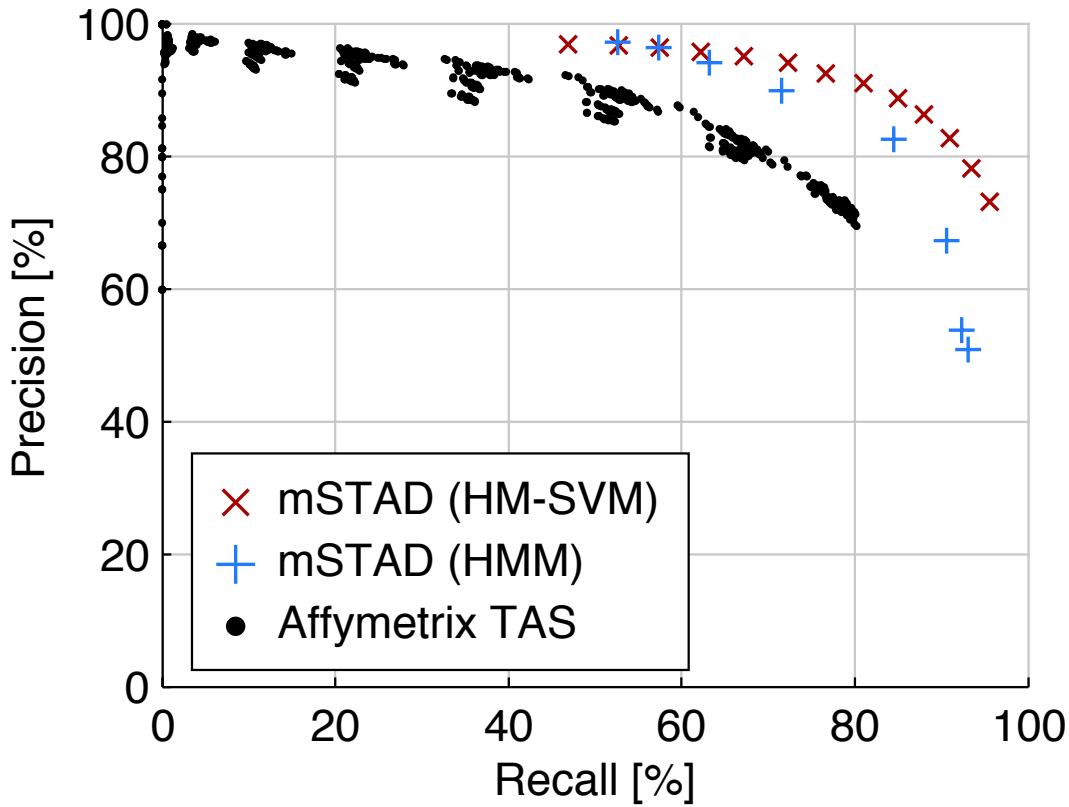
Method Comparison



Substantially improved exon probe recognition over the most widely used "transfrag" method

www.fml.mpg.de/raetsch/research/tiling/

Method Comparison

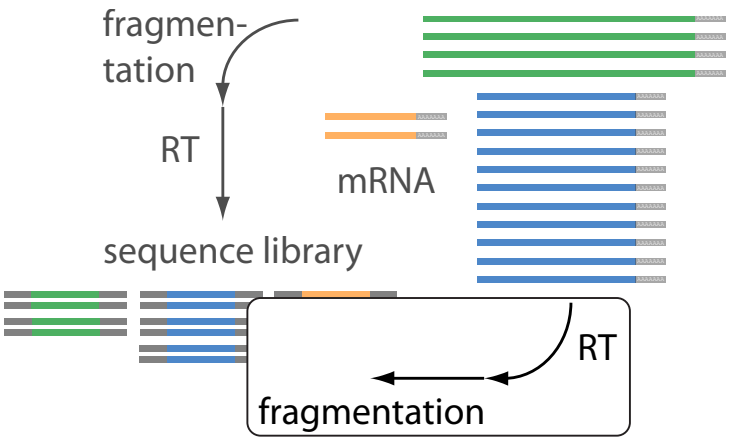


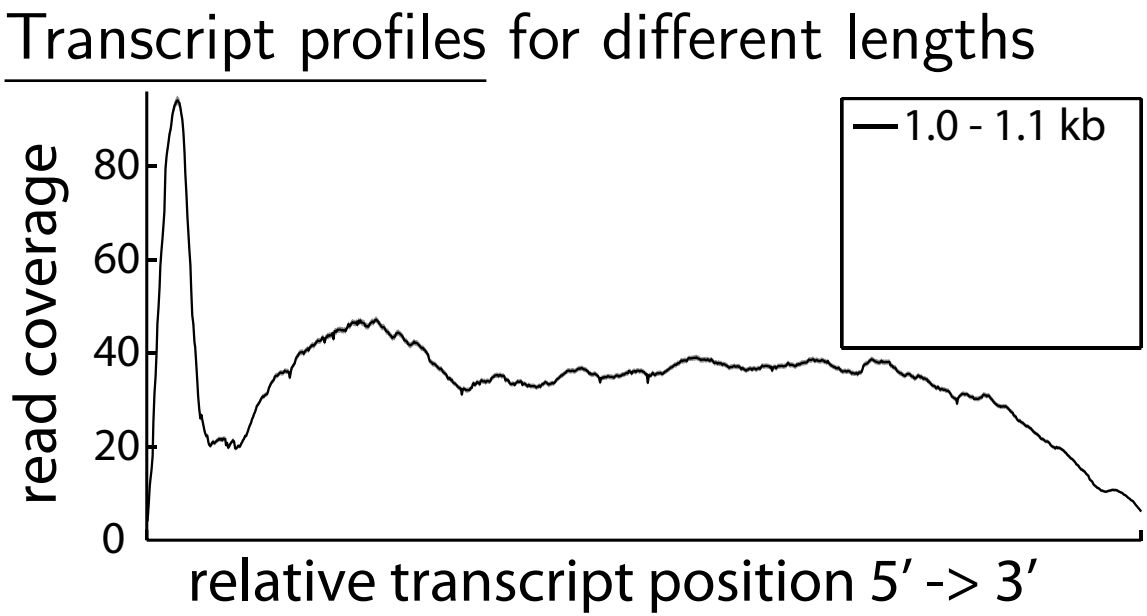
Substantially improved exon probe recognition over the most widely used "transfrag" method

www.fml.mpg.de/raetsch/research/tiling/

Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript



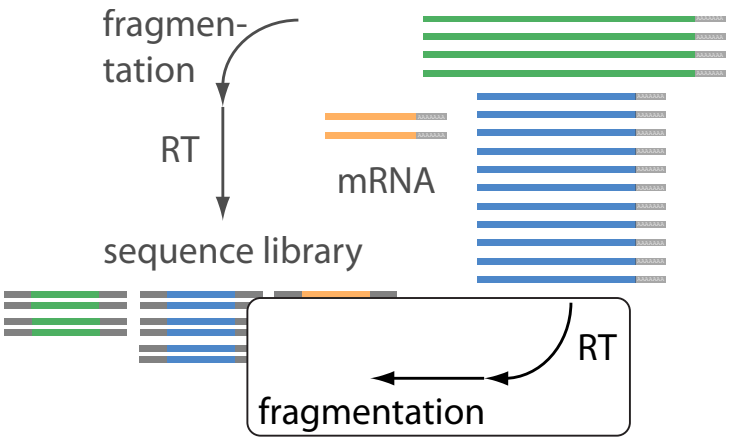


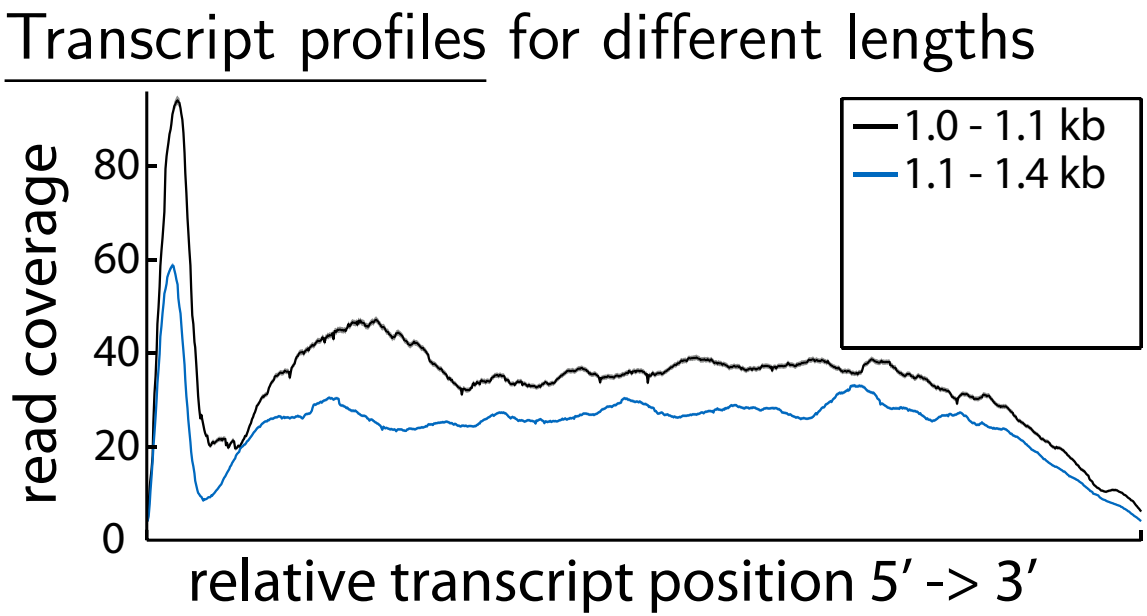
RNA-Seq data (*C. elegans* SRX001872, R. Waterston Lab, University of Washington)

- Random priming
- Physical cDNA fragmentation

Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript



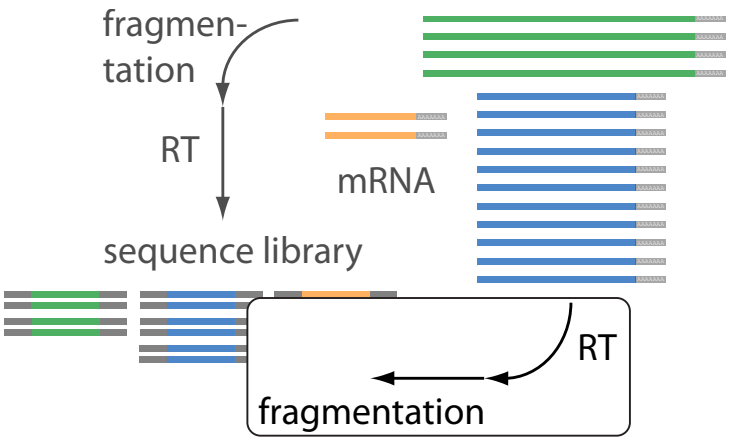


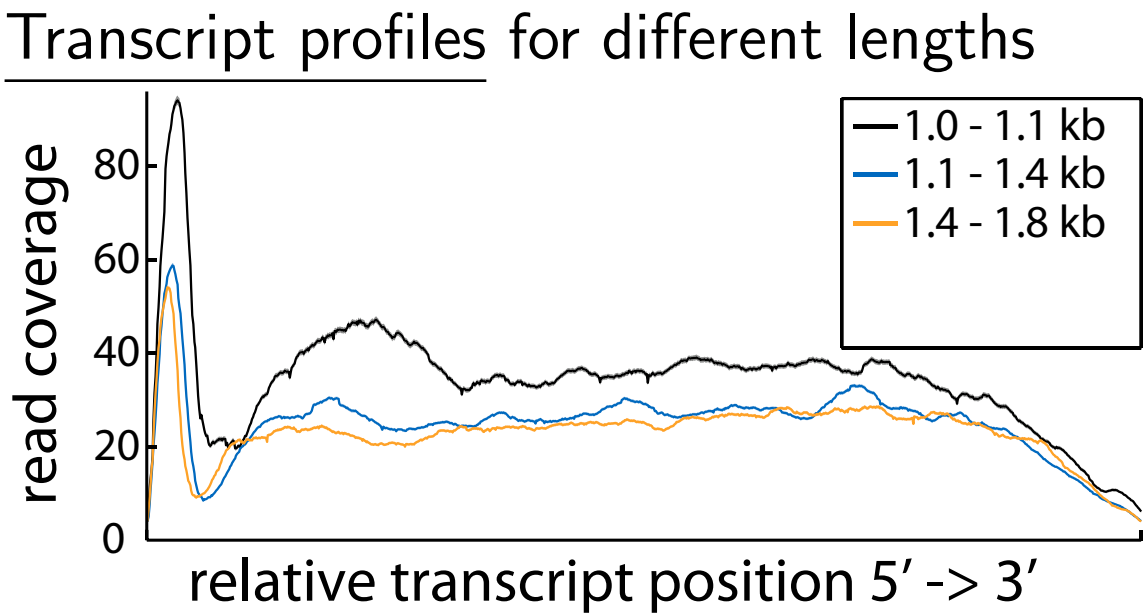
RNA-Seq data (*C. elegans* SRX001872, R. Waterston Lab, University of Washington)

- Random priming
- Physical cDNA fragmentation

Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript



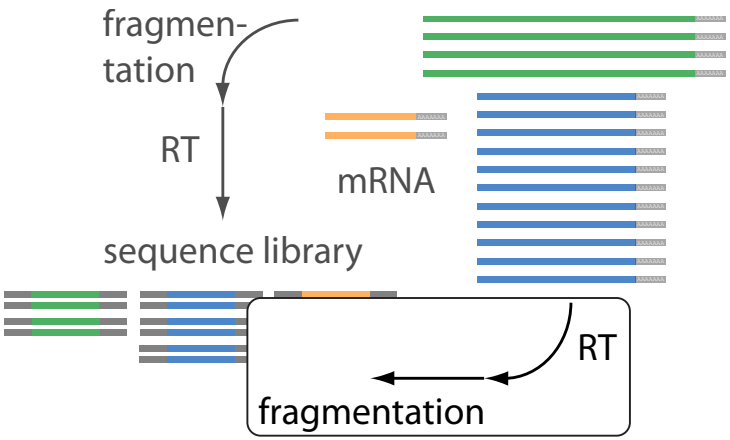


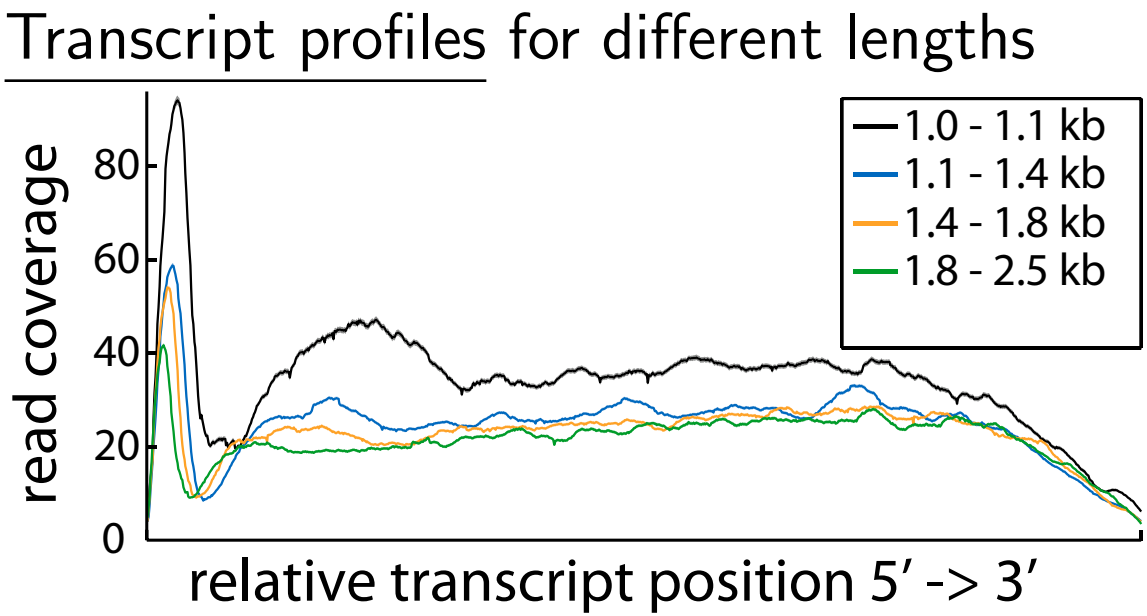
RNA-Seq data (*C. elegans* SRX001872, R. Waterston Lab, University of Washington)

- Random priming
- Physical cDNA fragmentation

Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript



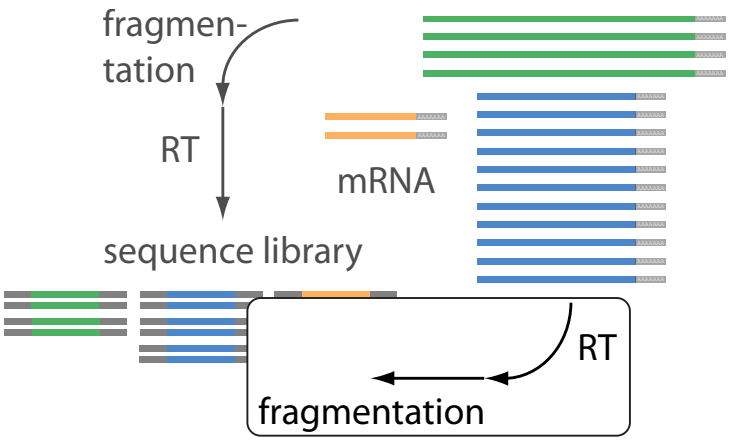


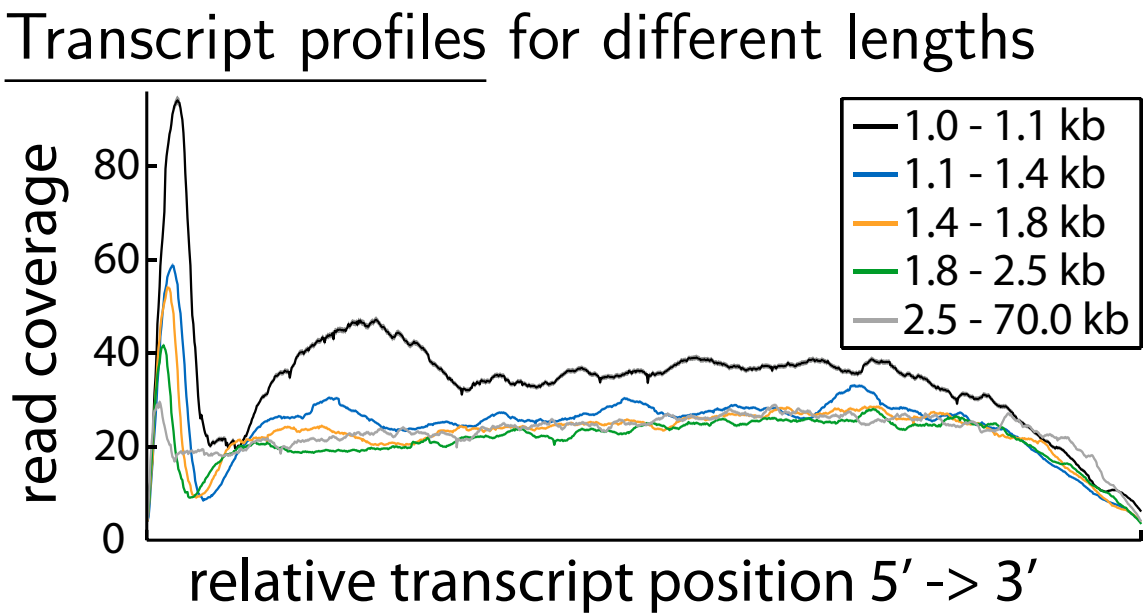
RNA-Seq data (*C. elegans* SRX001872, R. Waterston Lab, University of Washington)

- Random priming
- Physical cDNA fragmentation

Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript



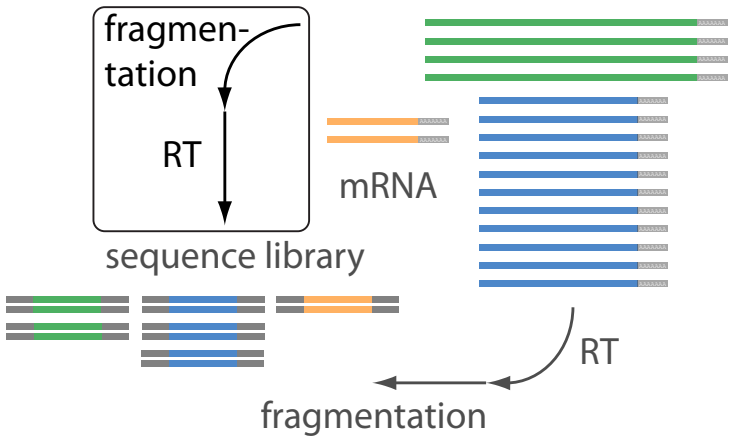


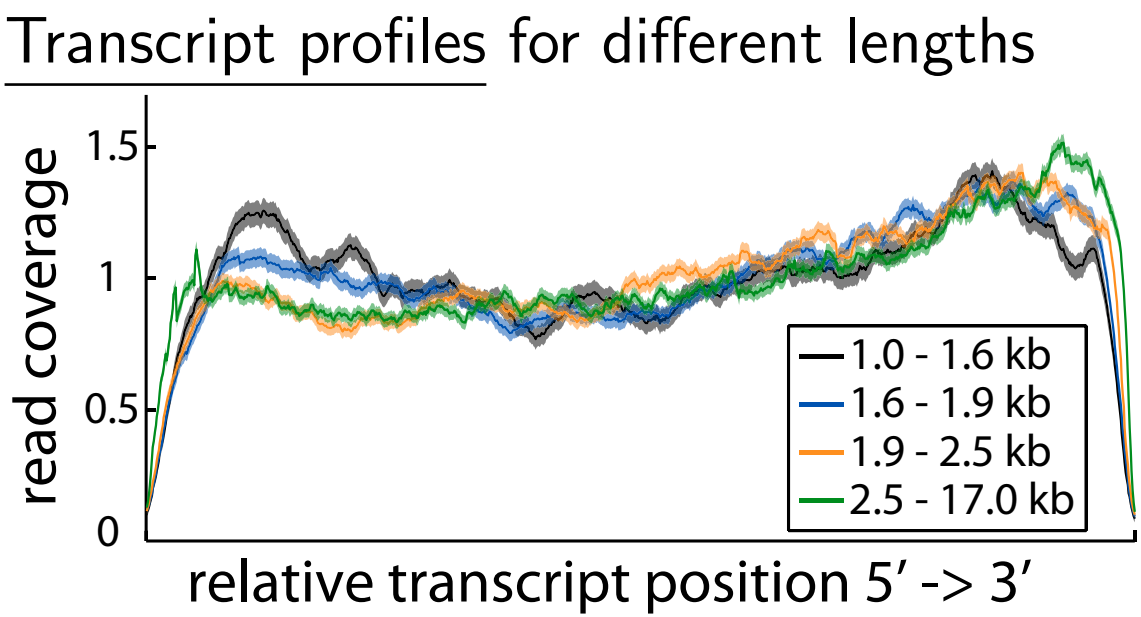
RNA-Seq data (*C. elegans* SRX001872, R. Waterston Lab, University of Washington)

- Random priming
- Physical cDNA fragmentation

Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript

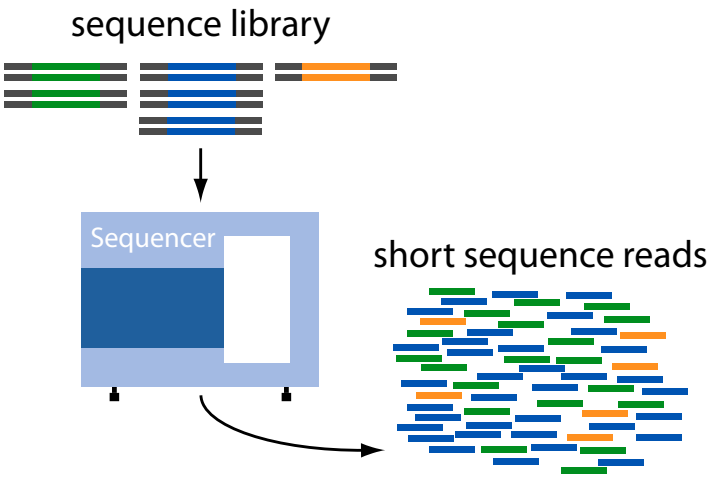




RNA-Seq data (*A. thaliana*, D. Weigel's Lab, MPI Tübingen)

- Chemical RNA fragmentation
- Random priming

Sequence Bias



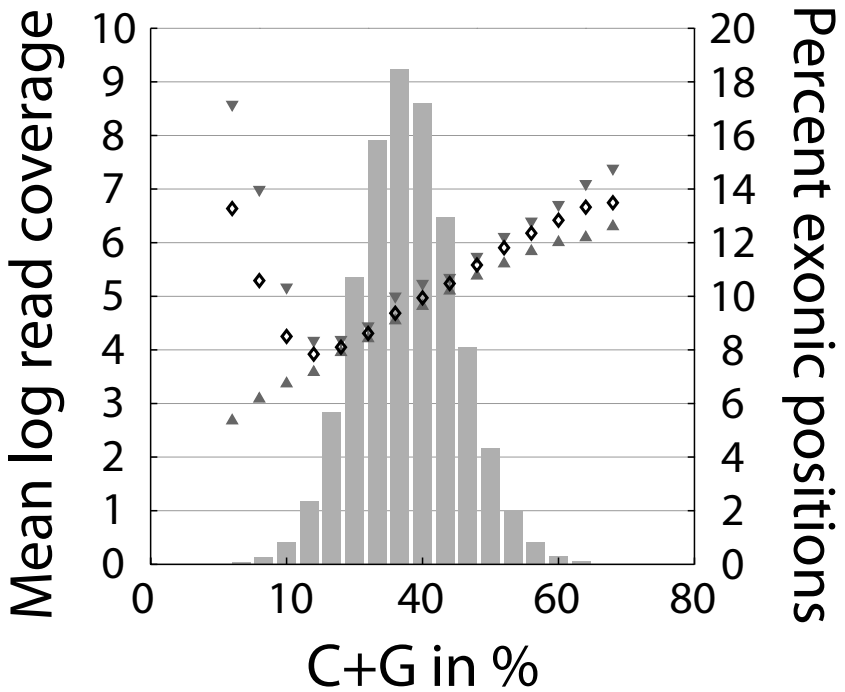
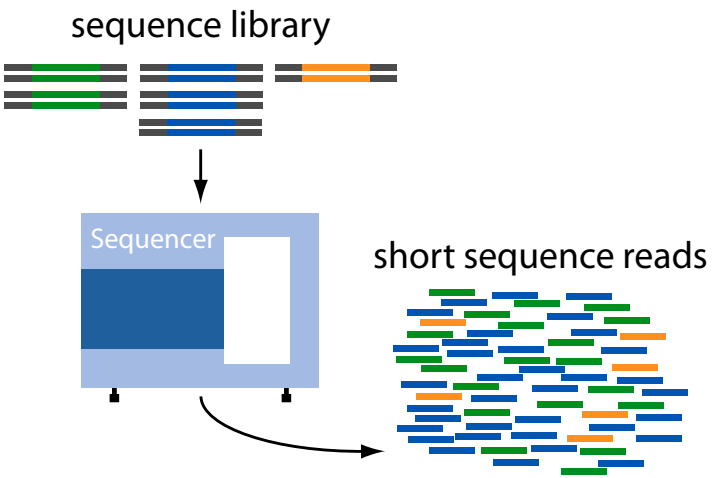


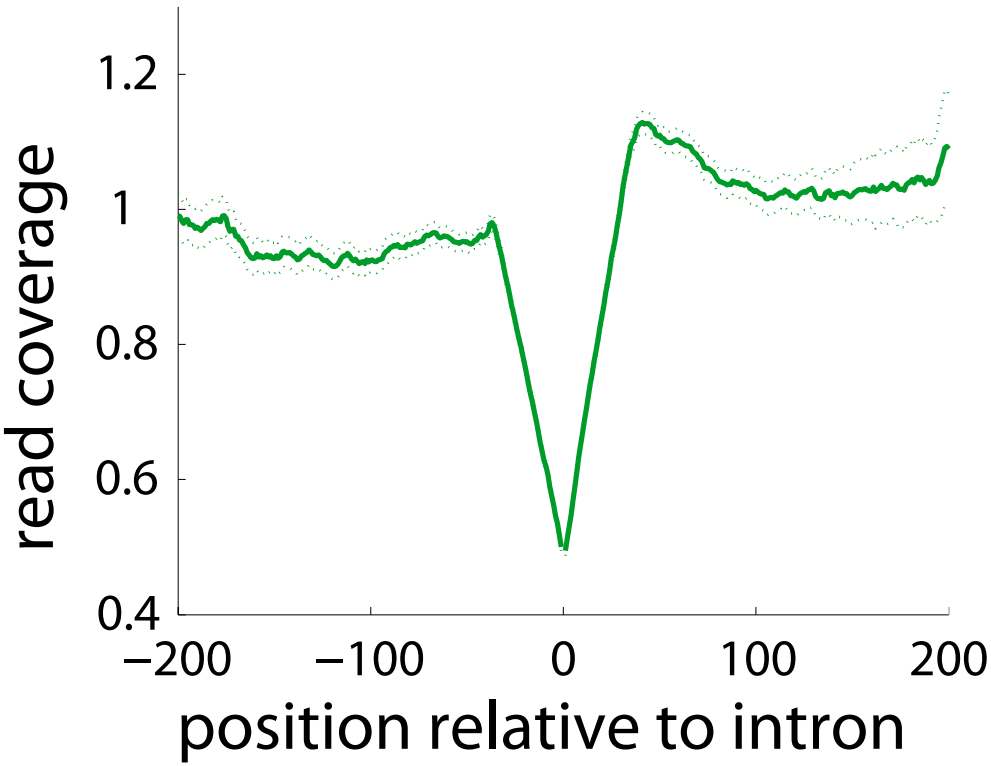
Figure provided by Georg Zeller

RNA-Seq data (*A. thaliana*, D. Weigel's Lab, MPI Tübingen)

- Exonic GC content
- Dinucleotides at the boundaries (Linsen et al., 2009)

Read Mapping Bias





RNA-Seq data (*A. thaliana*, D. Weigel's Lab, MPI Tübingen)

- Exon boundaries

Evaluation I

Our method rQuant: Position-wise, with profiles

(estimating library and mapping bias)

Evaluation I

Our method rQuant: Position-wise, with profiles

(estimating library and mapping bias)

compared to

Evaluation I

Our method rQuant: Position-wise, with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles

Evaluation I

Our 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))

Evaluation I

Our 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 by Sammeth (2009a))

Evaluation I

Our 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 by Sammeth (2009a))

Estimate transcript abundances

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

Evaluation I

Our 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 by Sammeth (2009a))

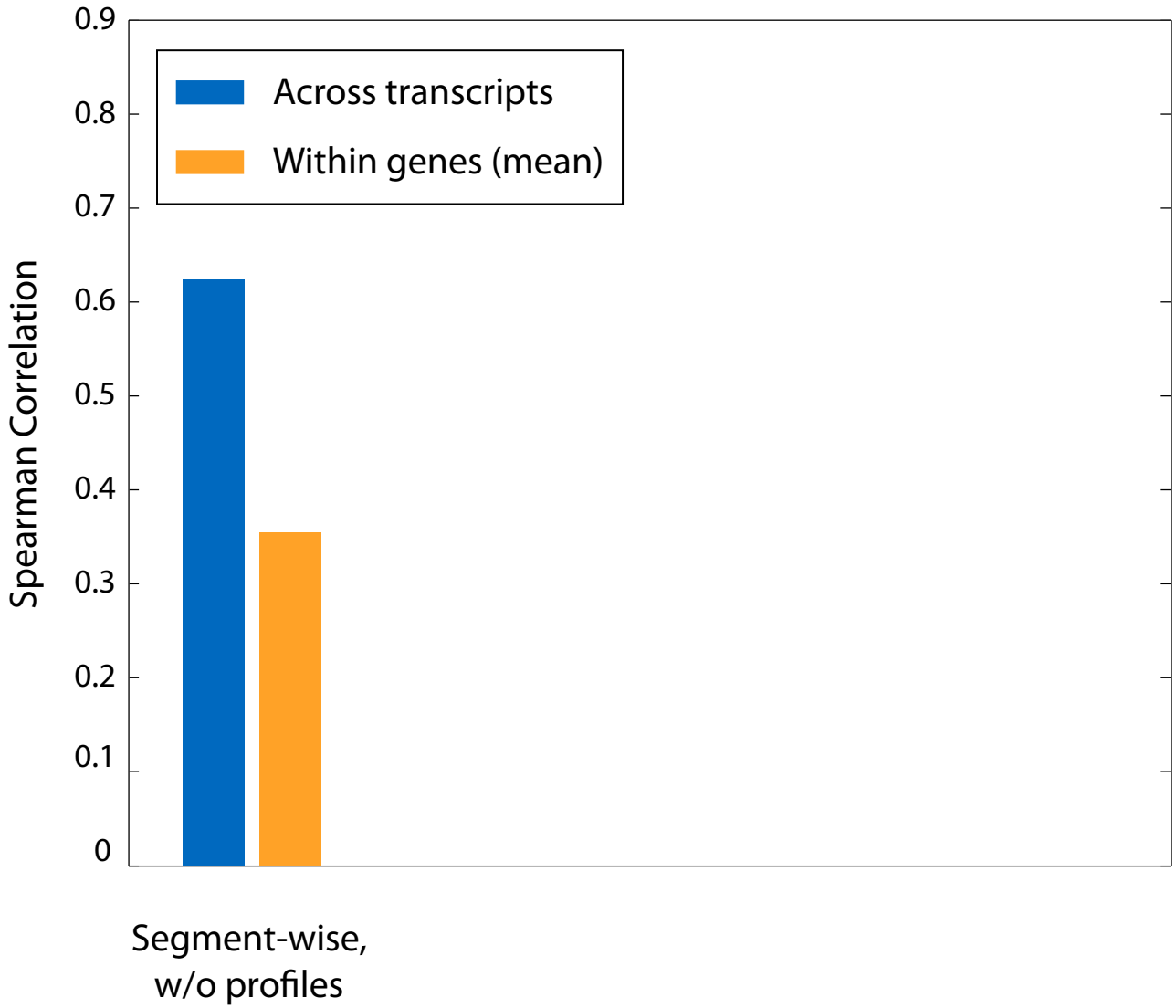
Estimate transcript abundances

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

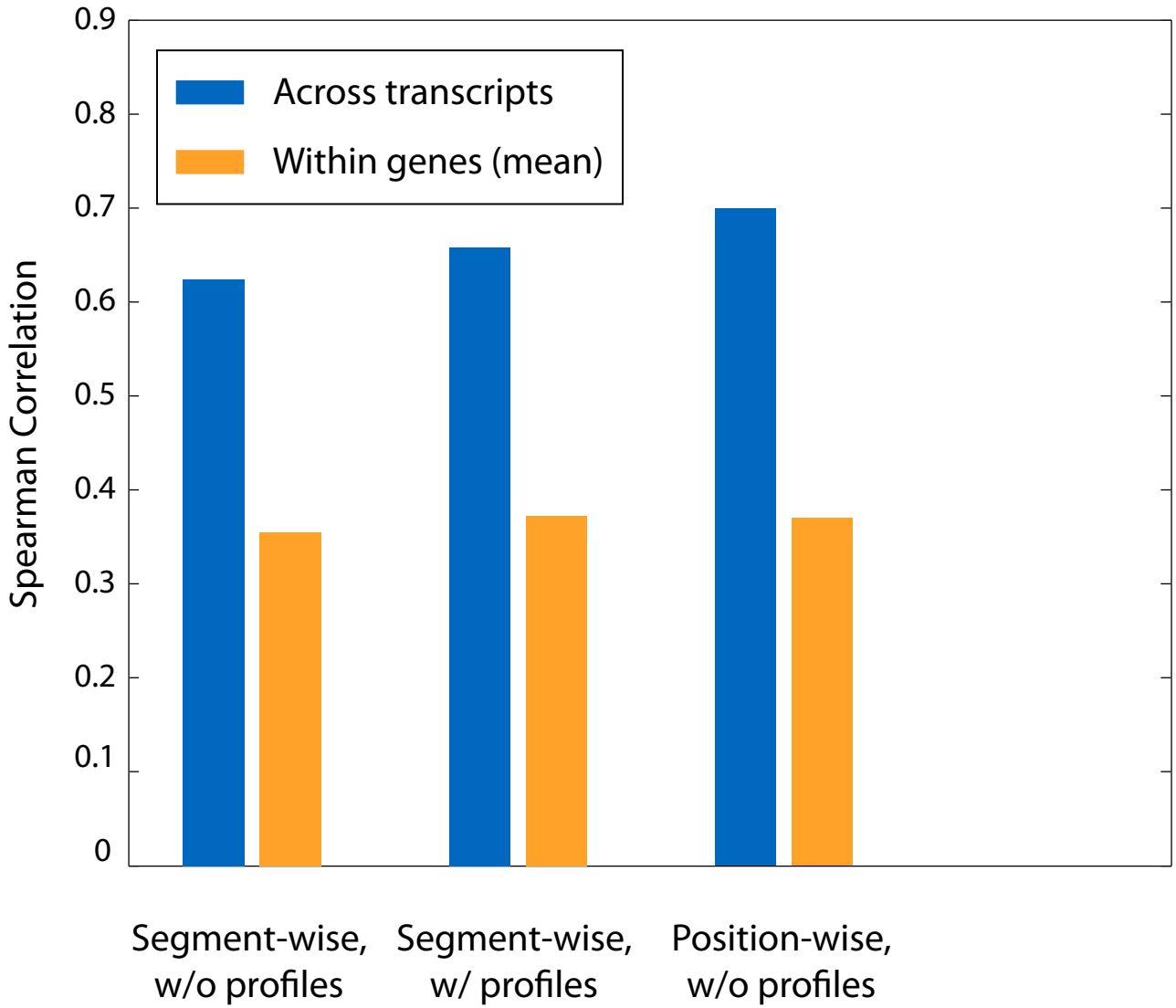
Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

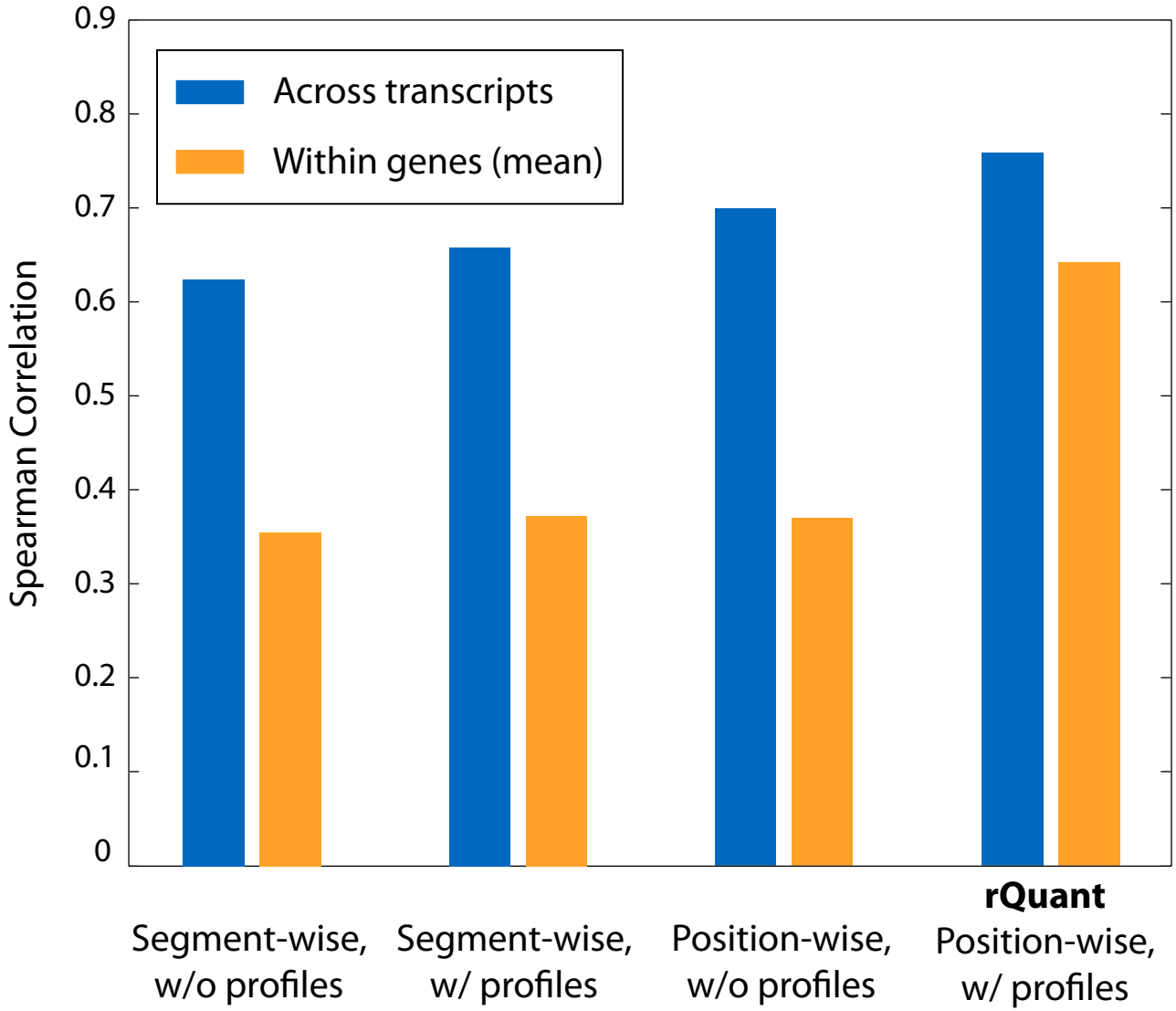
Evaluation II



Evaluation II



Evaluation II



Preliminary Evaluation I