# Machine Learning Challenges in Analysing Short Read Sequencing and Tiling Array Data

Computational Statistics Workshop, London, UK

**Gunnar Rätsch**



Friedrich Miescher Laboratory
of the Max Planck Society

July 6, 2009

# Machine Learning in Biology
## Research Topics

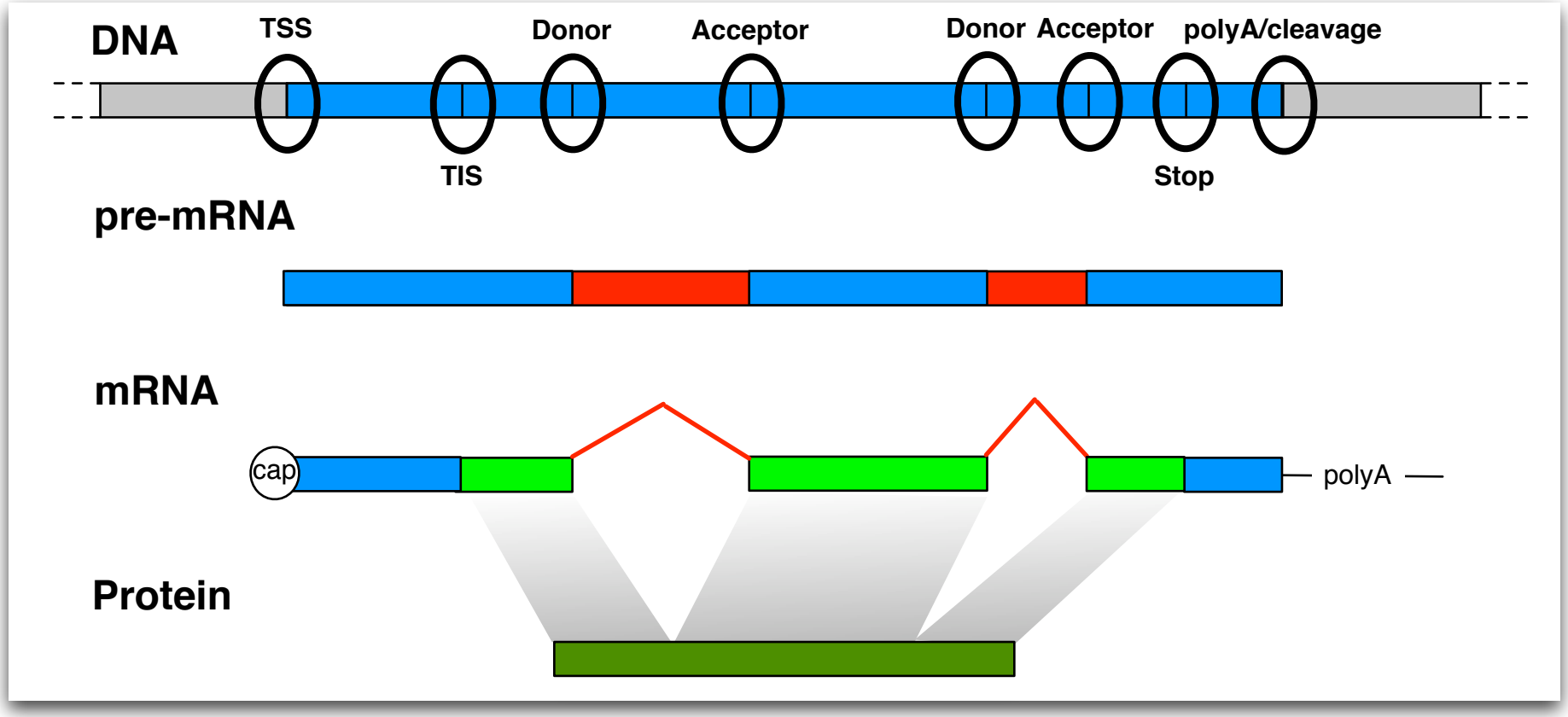
**Machine Learning**

1. Inference methods for structured data
   ⇒ *Develop fast and accurate learning methods*

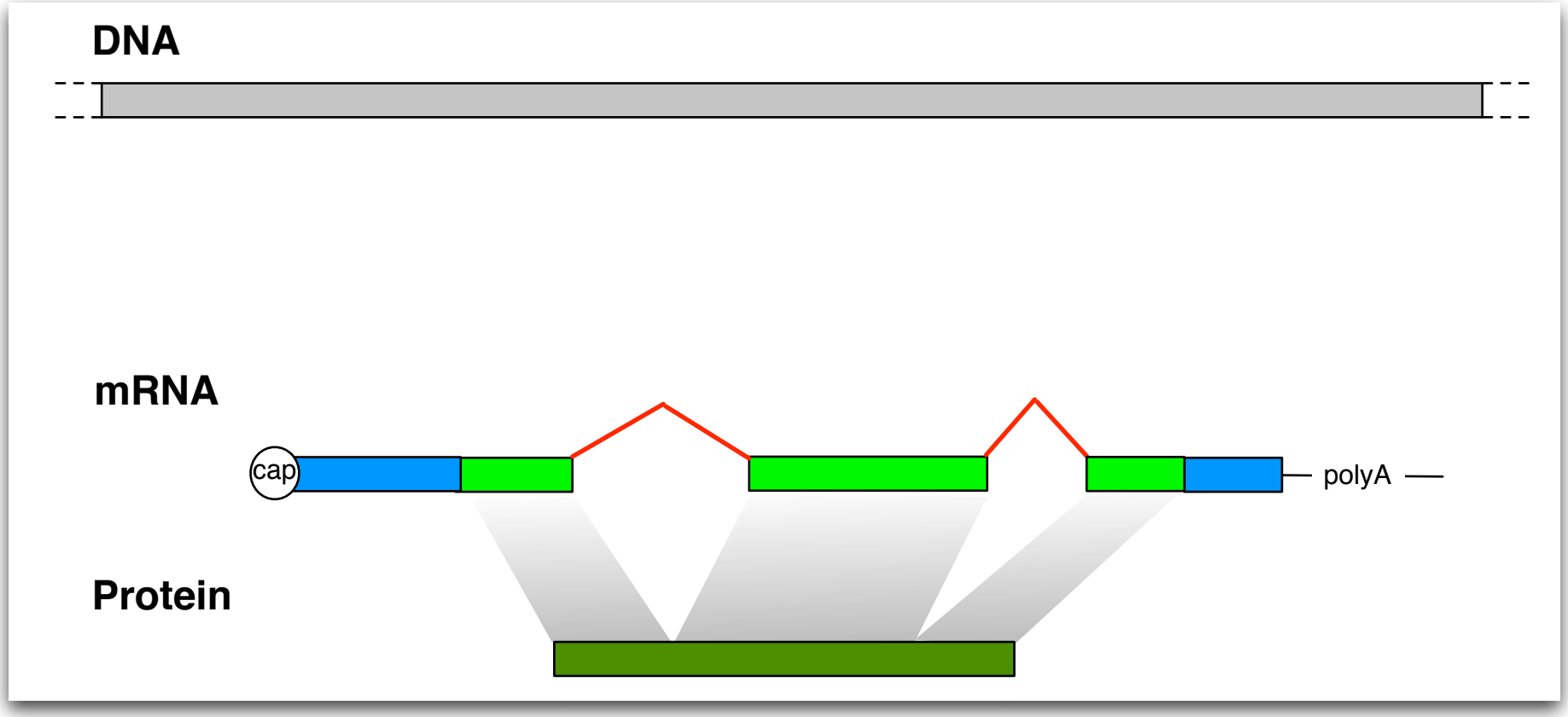
**Molecular Biology**

2. Genome annotation
   ⇒ *Predict features encoded on DNA*
3. Biological networks
   ⇒ *Understand interactions between gene products*
4. Analysis of polymorphisms
   ⇒ *Discover polymorphisms and associate with phenotypes*

# From Genome to Products



# From Measurements to Products




Directly measure the transcriptome

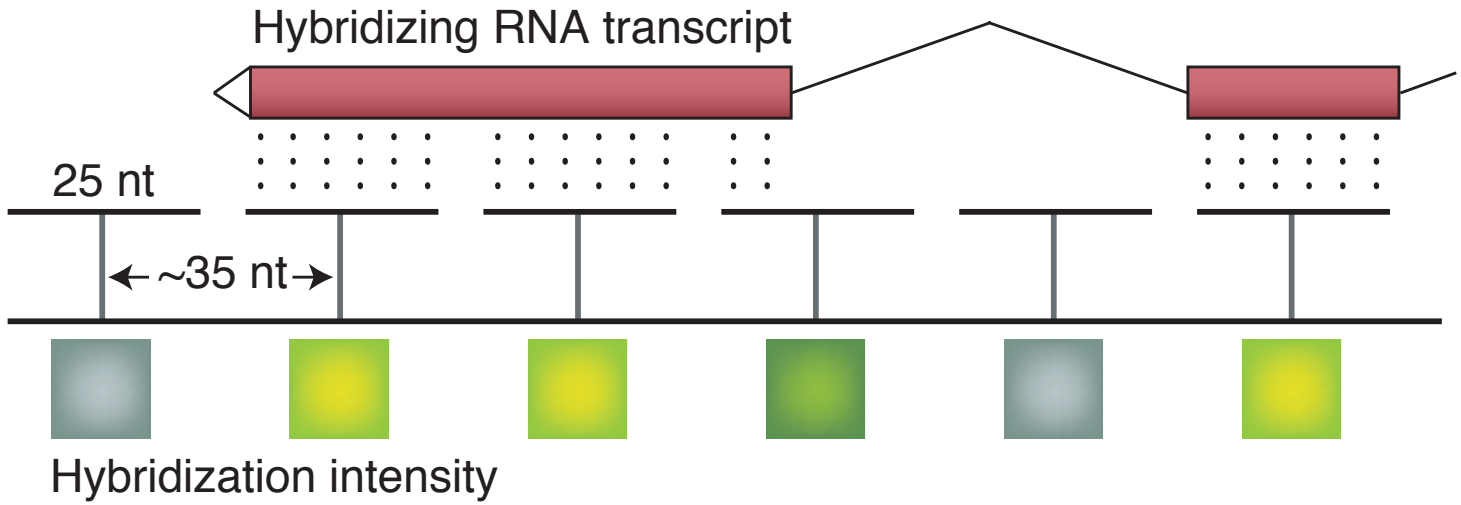
- Whole genome tiling arrays
- Transcriptome sequencing (Sanger or Next Generation Sequencing)

# Whole-genome Tiling Arrays



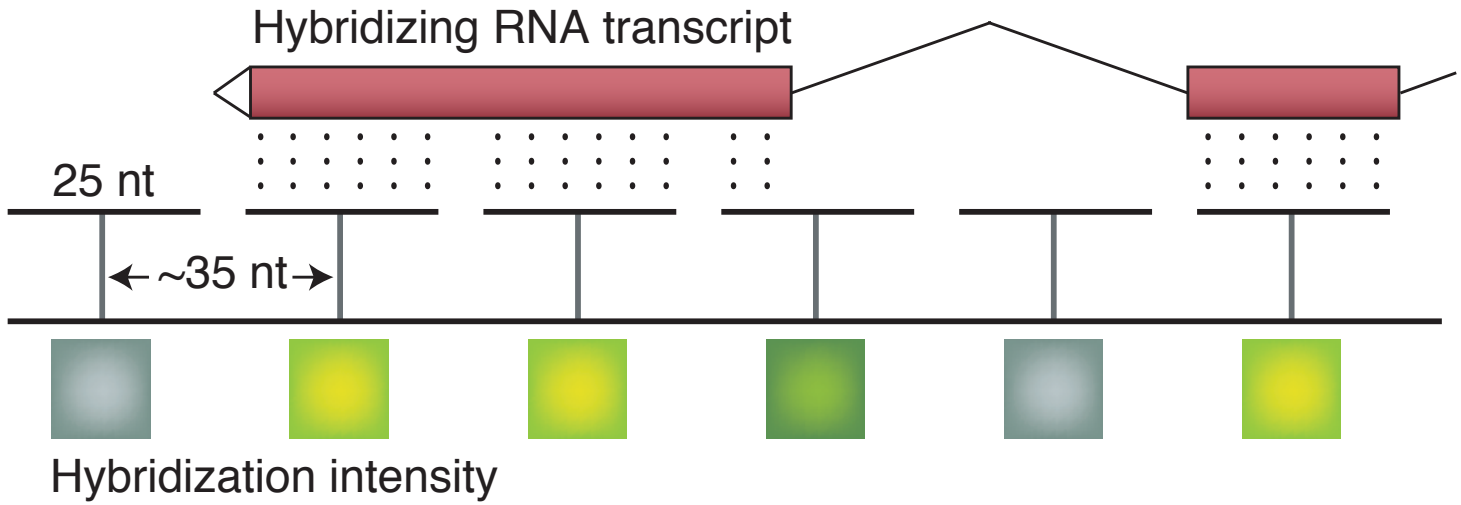

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays




see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays

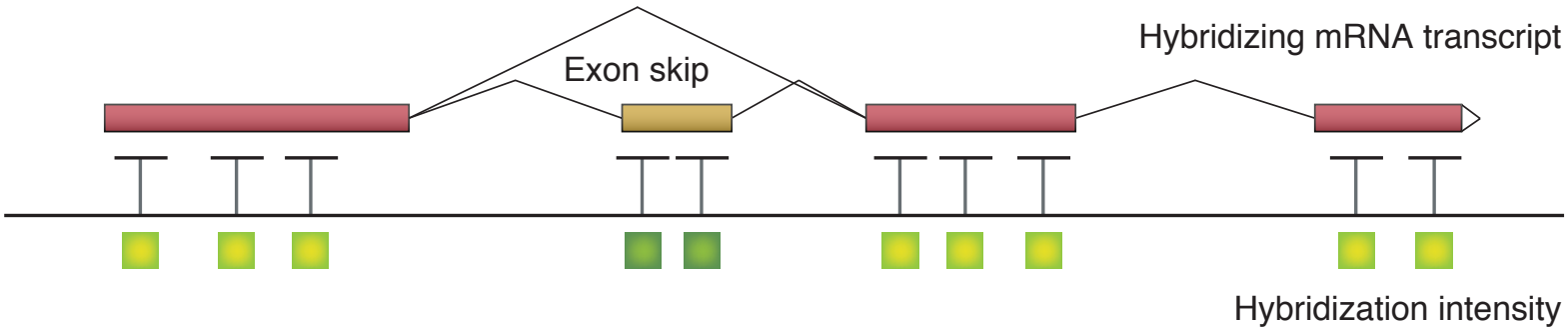



- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays

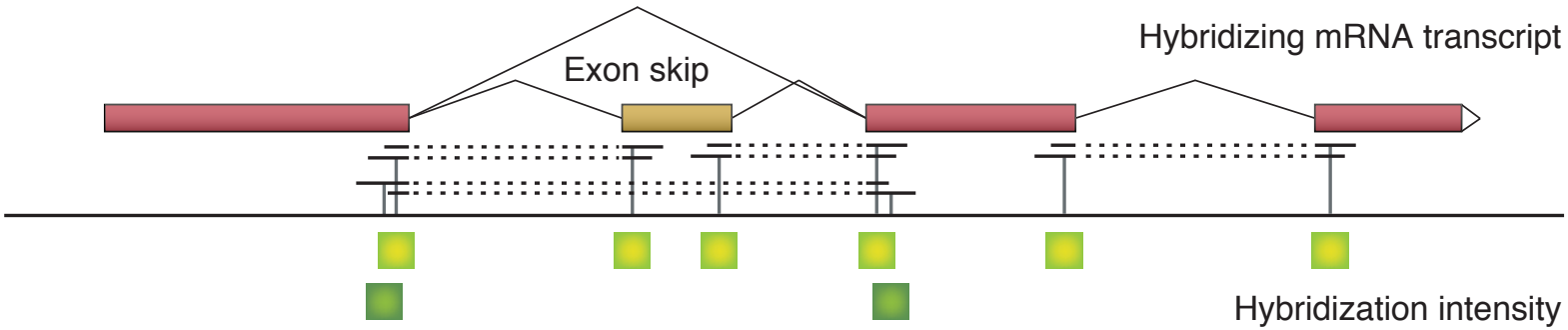



- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Variants: exon arrays,

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays

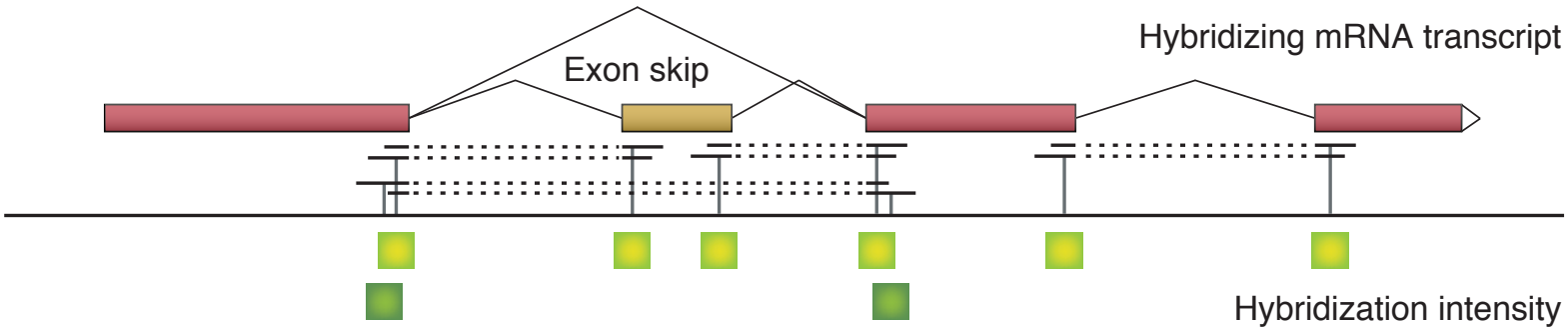



- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Variants: exon arrays, exon junction arrays

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays



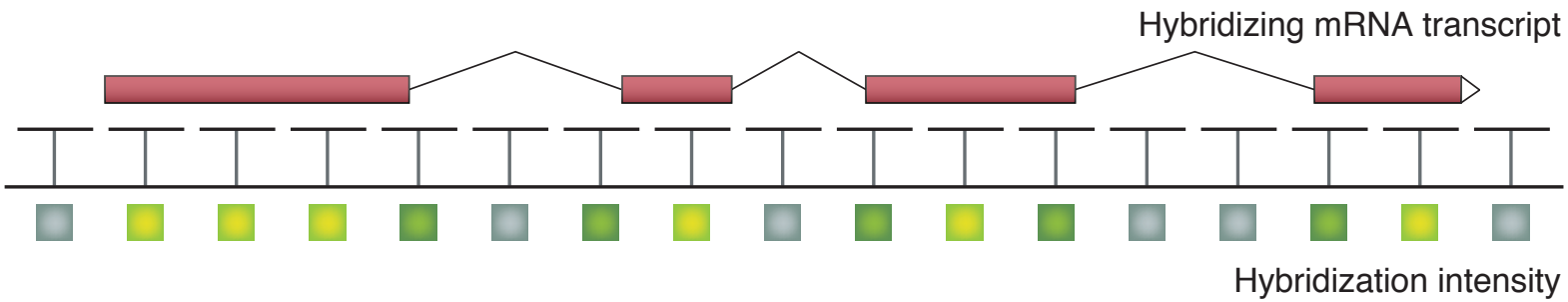

- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Variants: exon arrays, exon junction arrays
- Hybridization data is noisy, analysis challenging

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

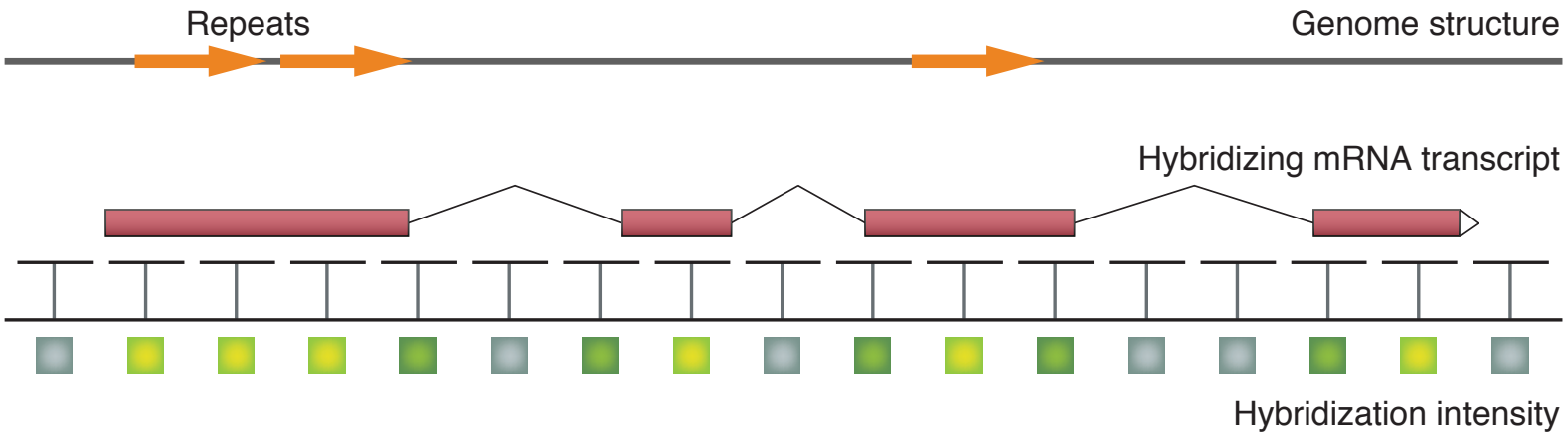
# Tiling Array Analysis Challenges (I)

Repeats cause cross-hybridization



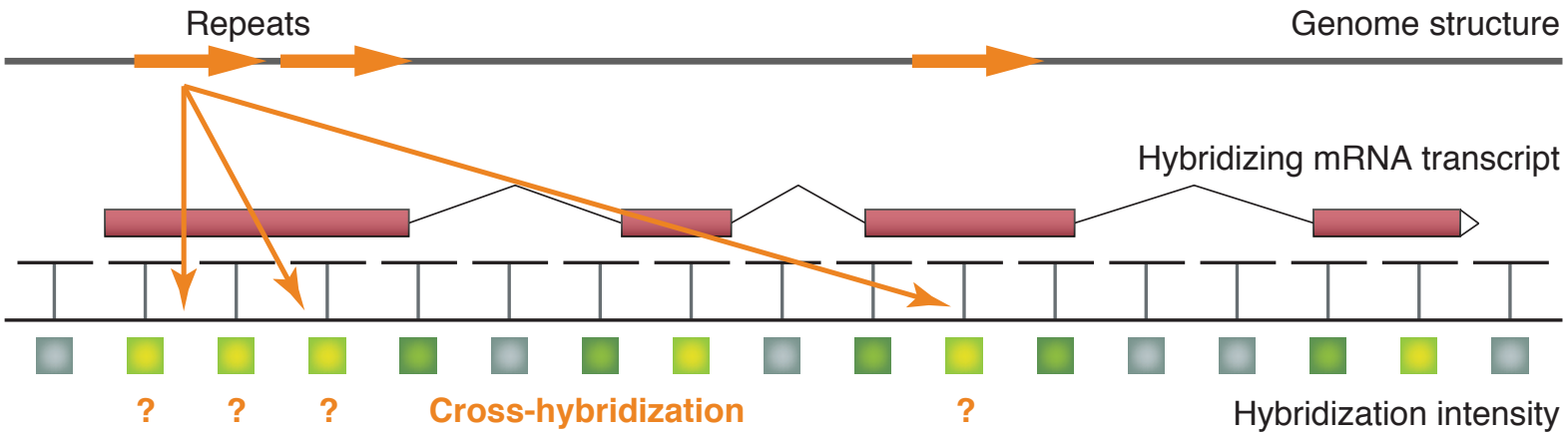
# Tiling Array Analysis Challenges (I)

Repeats cause cross-hybridization



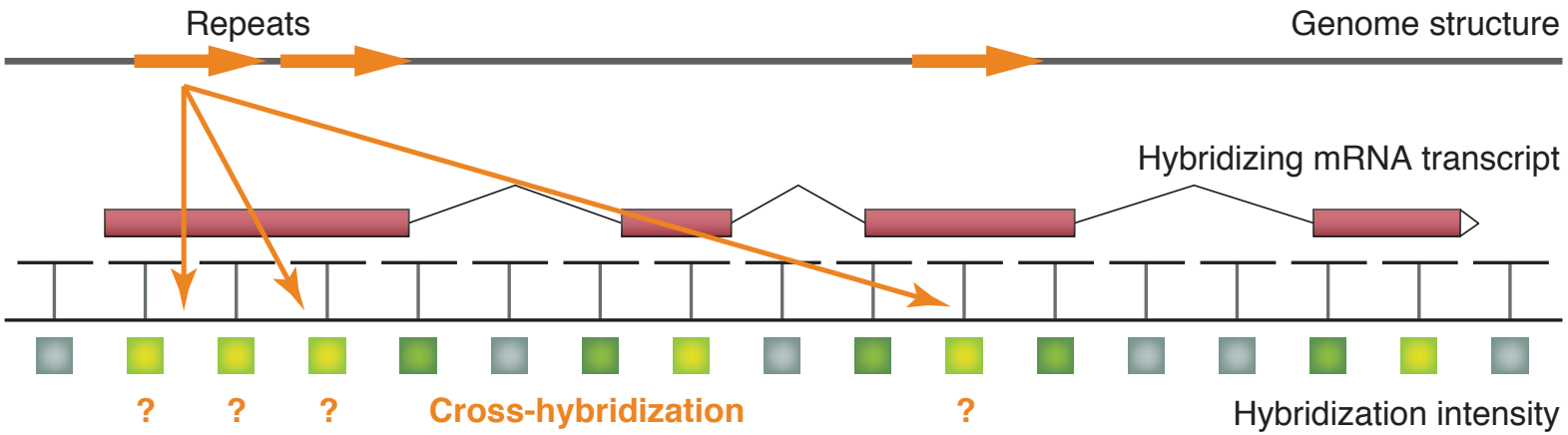
# Tiling Array Analysis Challenges (I)

Repeats cause cross-hybridization



# Tiling Array Analysis Challenges (I)

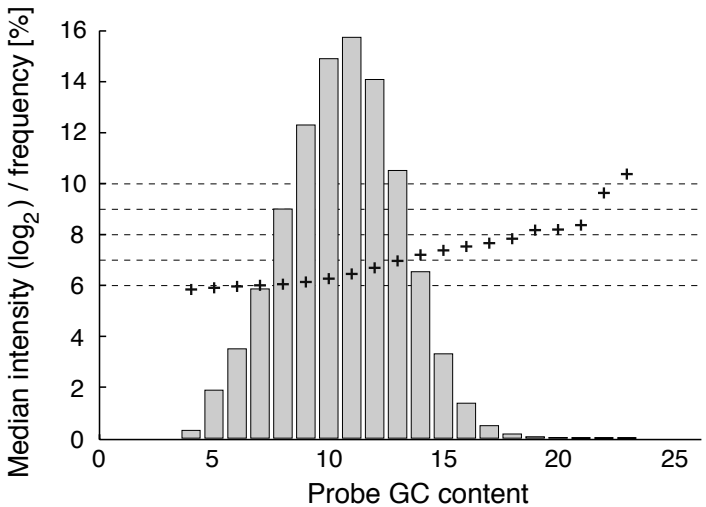
Repeats cause cross-hybridization




⇒ Discard tiling probes with high sequence similarity to >1 location in the genome

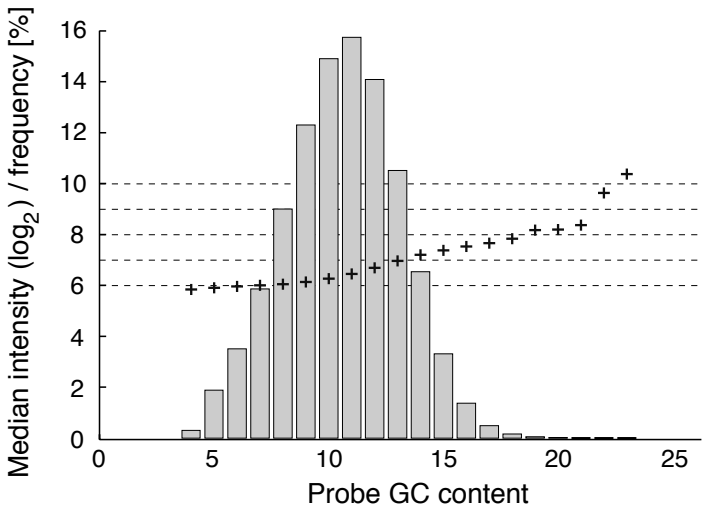
# Tiling Array Analysis Challenges (II)

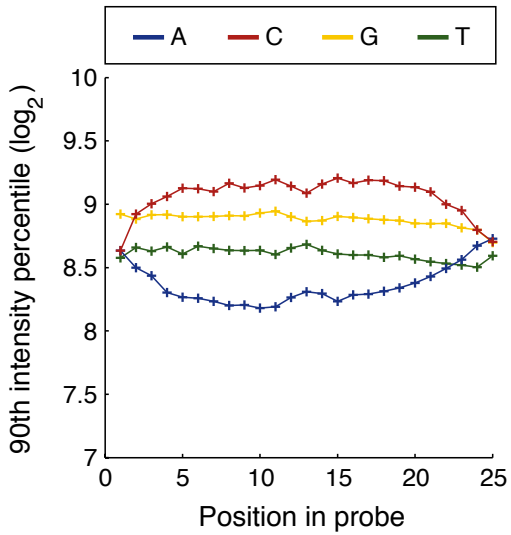
Hybridization intensity exhibits a probe sequence bias



# Tiling Array Analysis Challenges (II)

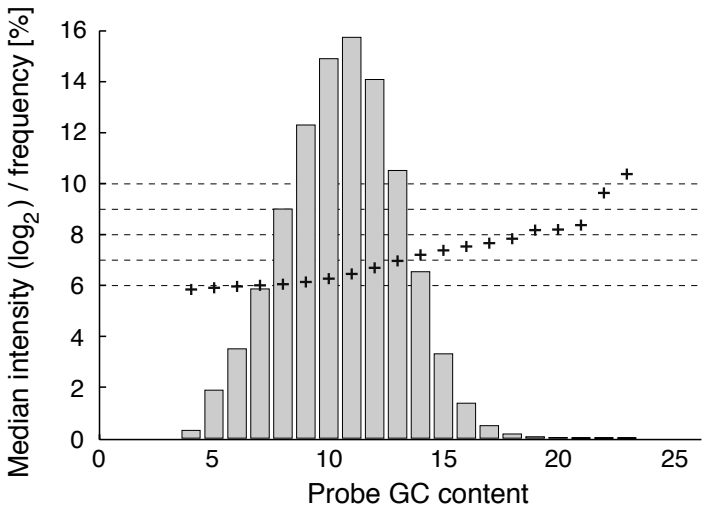
Hybridization intensity exhibits a probe sequence bias

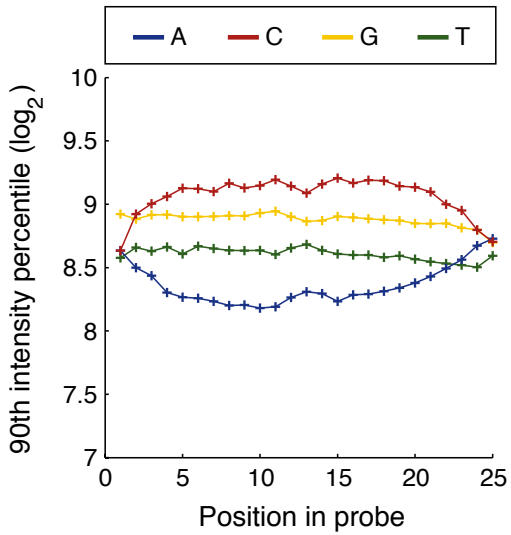


# Tiling Array Analysis Challenges (II)

Hybridization intensity exhibits a probe sequence bias




Sequence-normalization approaches:

- Rescaling by probe GC content [Samanta et al., 2006]
- Rescaling using genomic DNA hybridization [David et al., 2006]

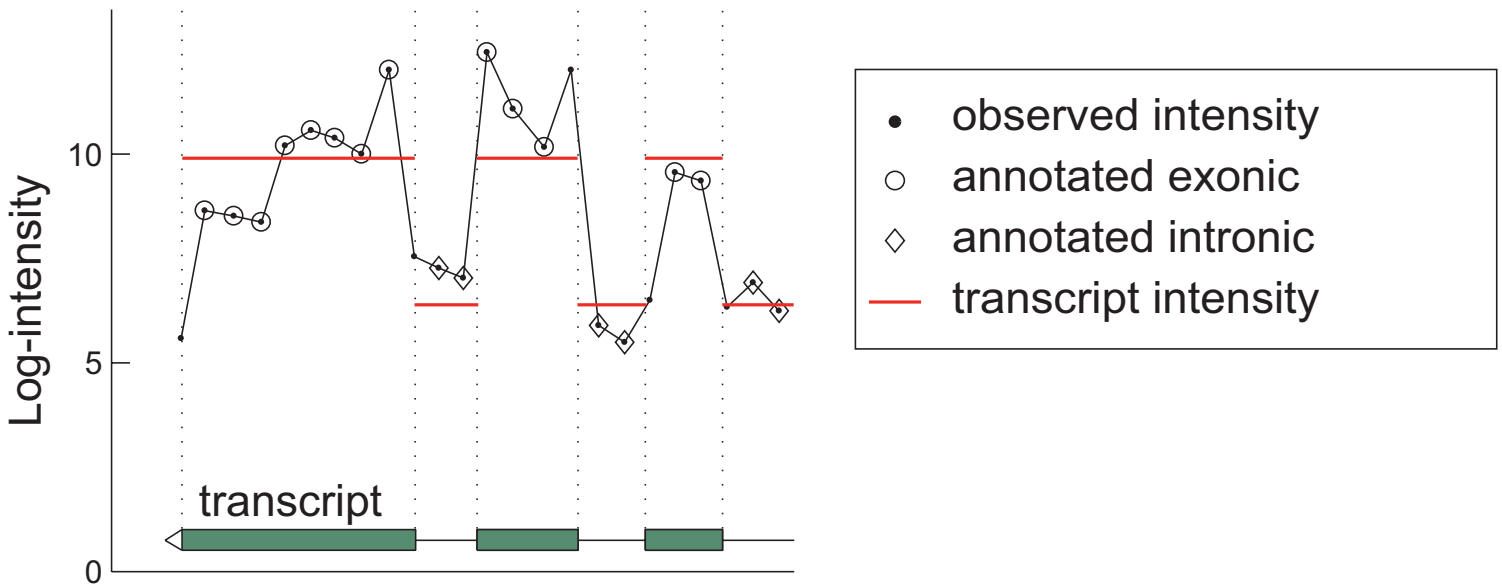
$$n_{ij} = \frac{x_{ij} - b_j(y_i)}{y_i}$$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$ to obtain normalized signal $n_{ij}$; DNA hybridization signal $y_i$ is transformed into RNA background signal by $b_j$ estimated from intergenic probes.

- Regression techniques [Royce et al., 2007, Zeller et al., 2008b]

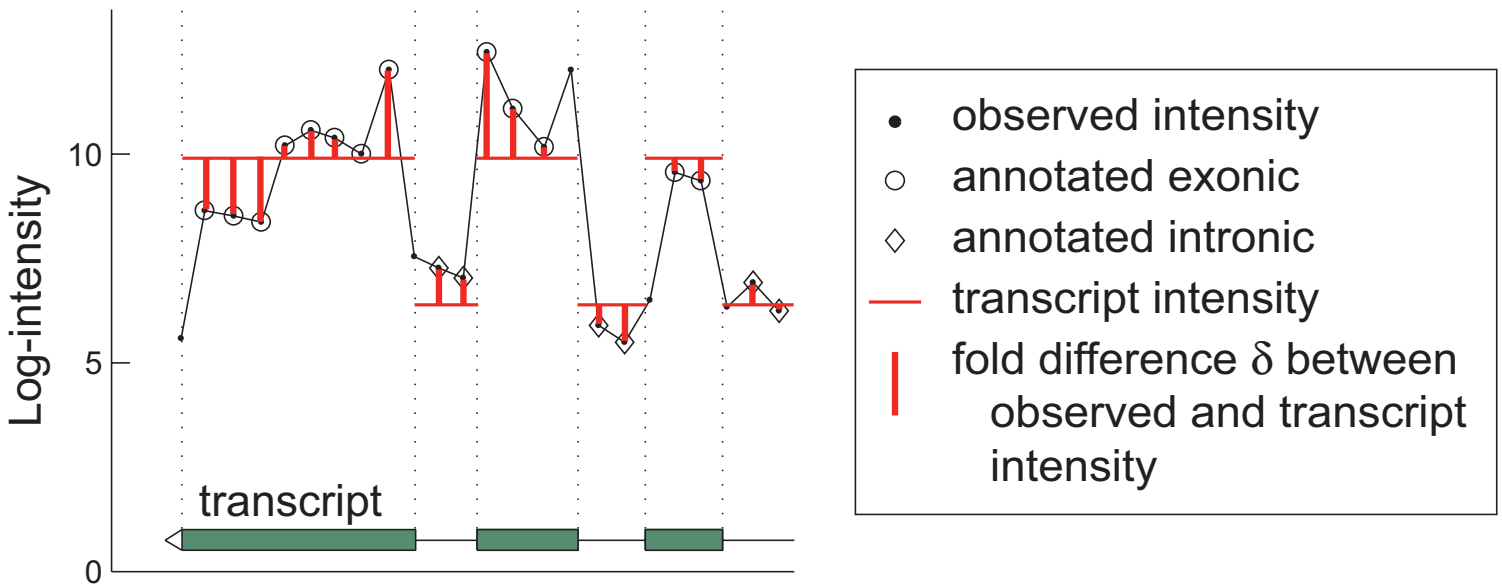
# Tiling Array Normalization

- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008b]



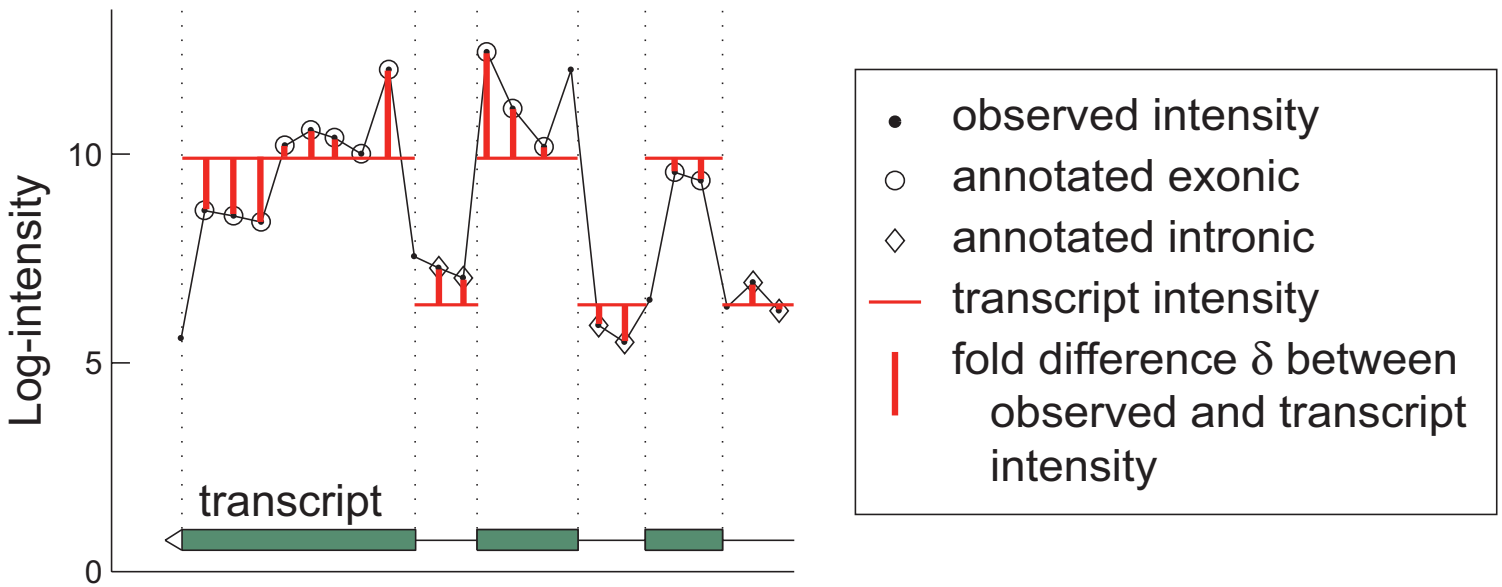
# Tiling Array Normalization

- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008b]
- Learns intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$



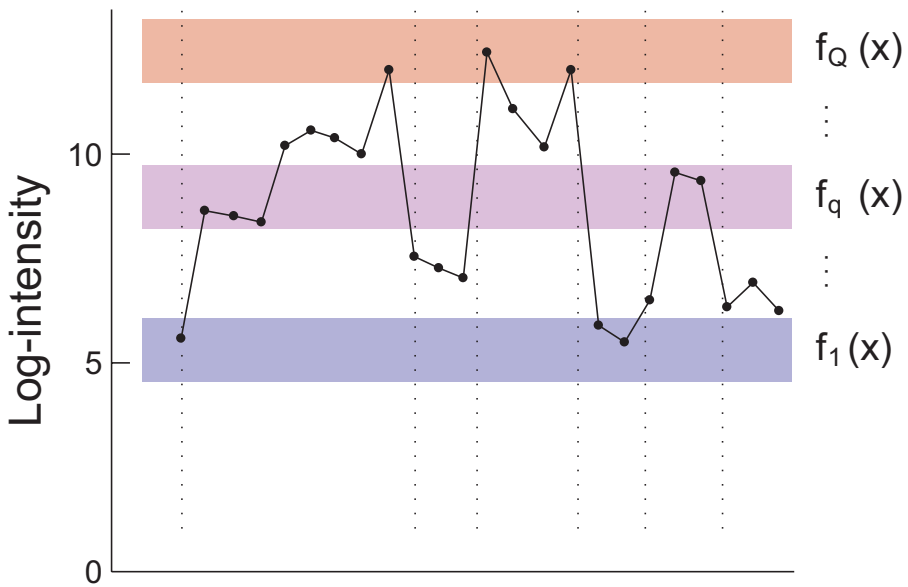
# Tiling Array Normalization

- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008b]
- Learns intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Takes probe sequence $\mathbf{x}_i$ (positional information on mono-, di- and tri-mer occurrence) as input for regression.
  Models probe sequence effect depending on $y_i$: $f(\mathbf{x}_i, y_i) \approx \delta_i$



# Tiling Array Normalization

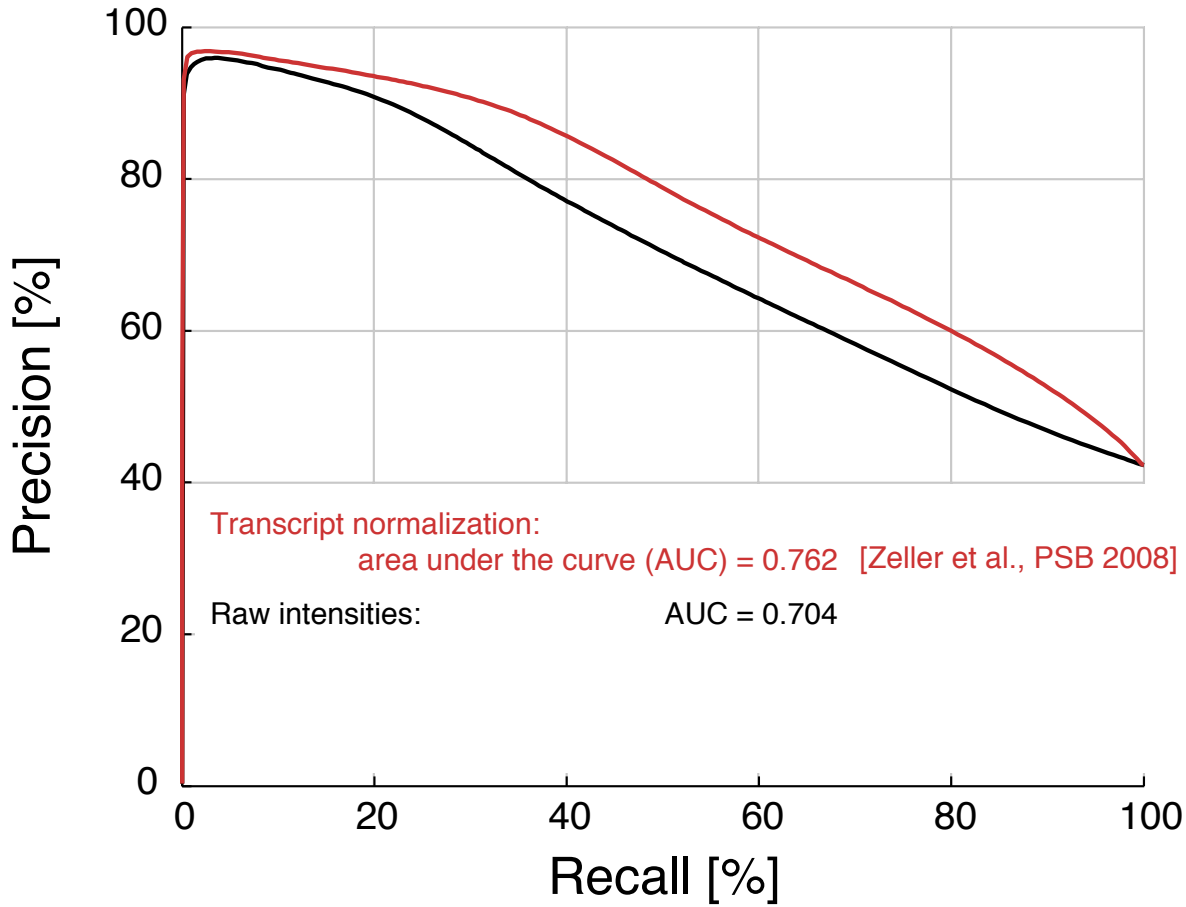
- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008b]
- Learns intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Takes probe sequence $\mathbf{x}_i$ (positional information on mono-, di- and tri-mer occurrence) as input for regression.
  Models probe sequence effect depending on $y_i$: $f(\mathbf{x}_i, y_i) \approx \delta_i$




Discretize $y$ into $Q = 20$ quantiles and estimate $Q$ independent functions $f_1(\mathbf{x}), \ldots, f_Q(\mathbf{x})$

Linear regression
$f_q(\mathbf{x}) = \mathbf{w}_q^T \mathbf{x}$

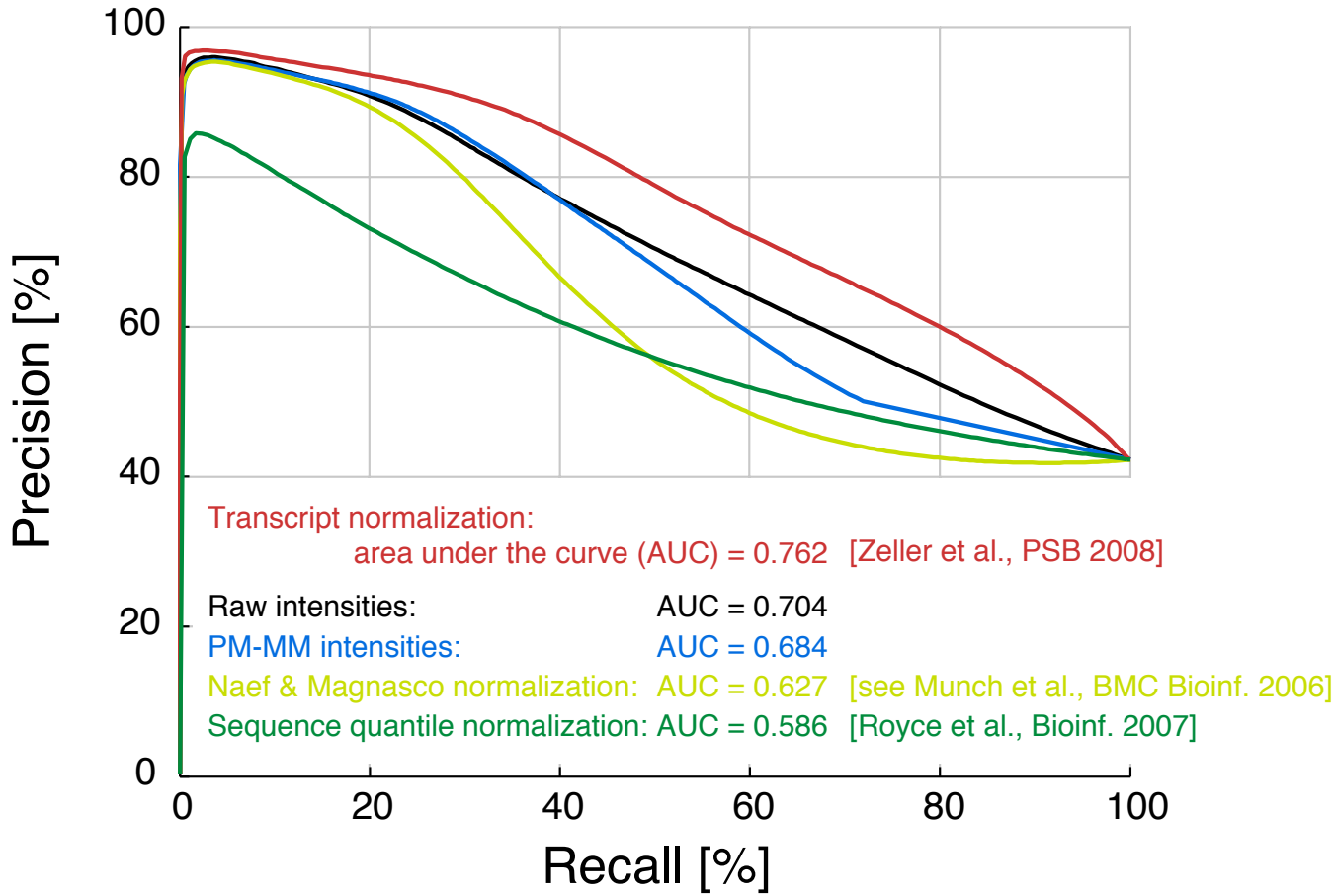
# Transcript Normalization Results




Transcript normalization (TN) results in better separation between background and exon signals → improves transcript identification

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

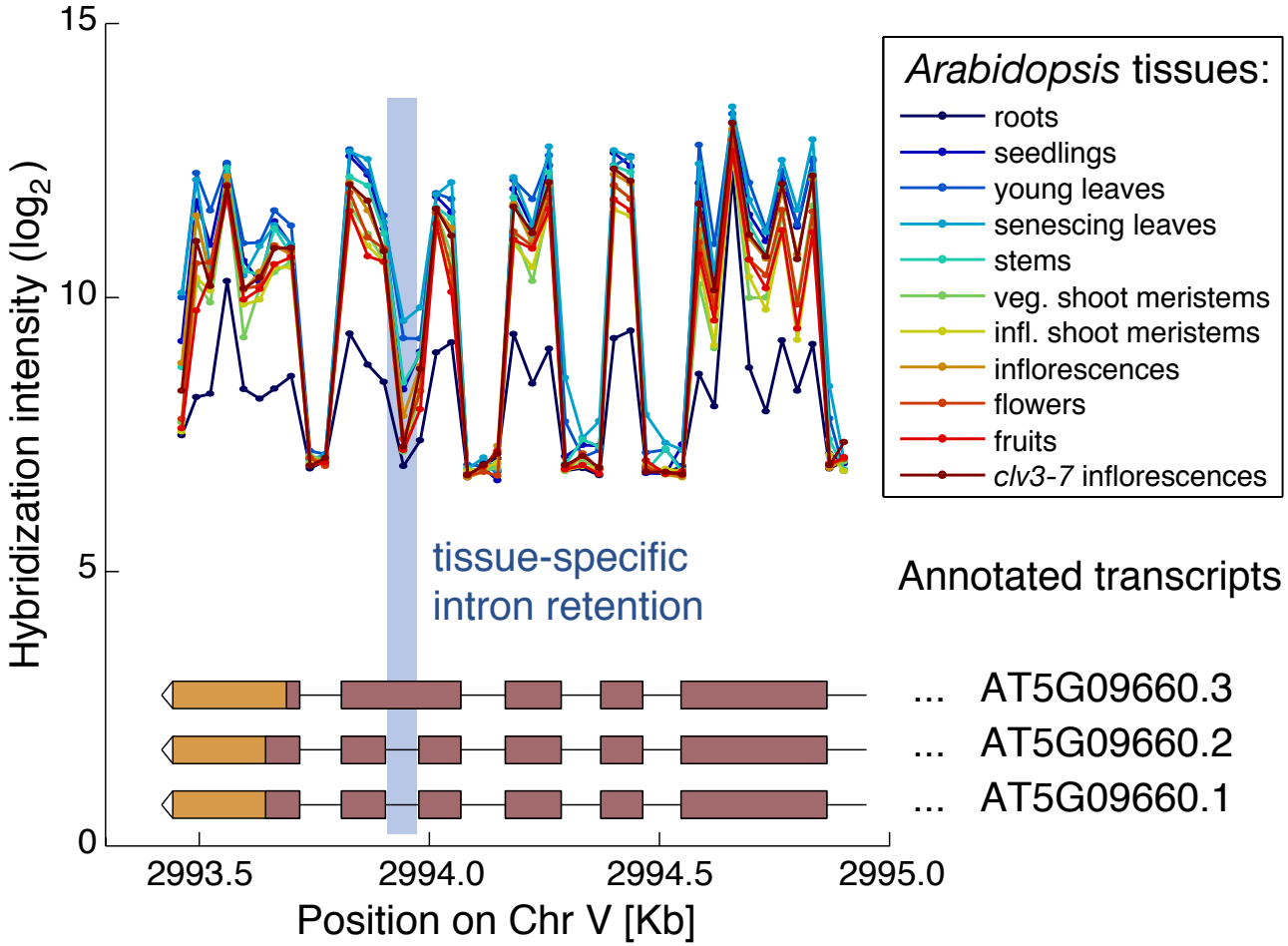
# Transcript Normalization Results




Transcript normalization (TN) results in better separation between background and exon signals → improves transcript identification

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

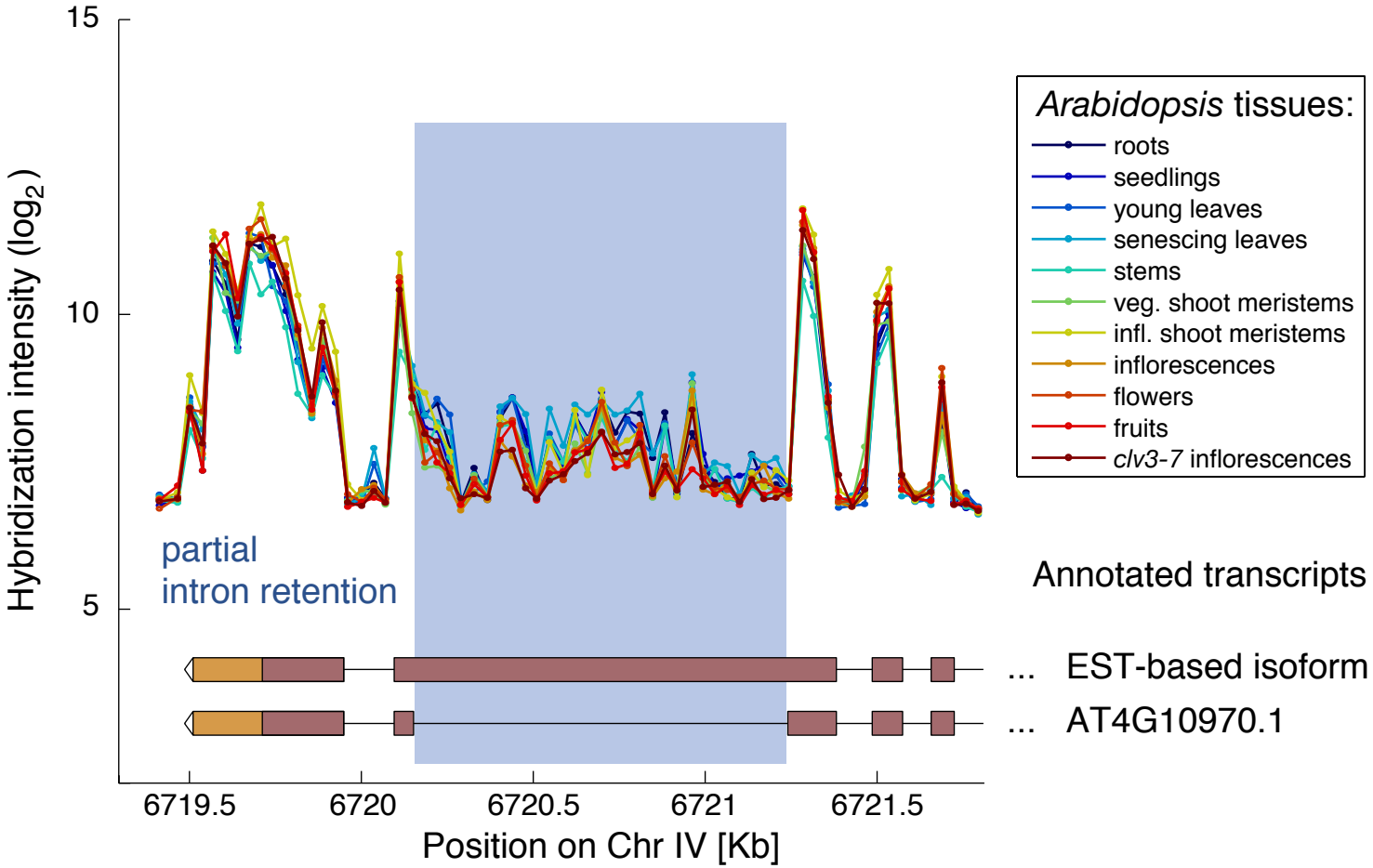
# Detection of Alternative Transcripts (I)




**Goal:** Identify exon/intron segments that are *differentially* spliced in the analyzed samples.

[Eichner, 2008, Eichner et al., 2009]

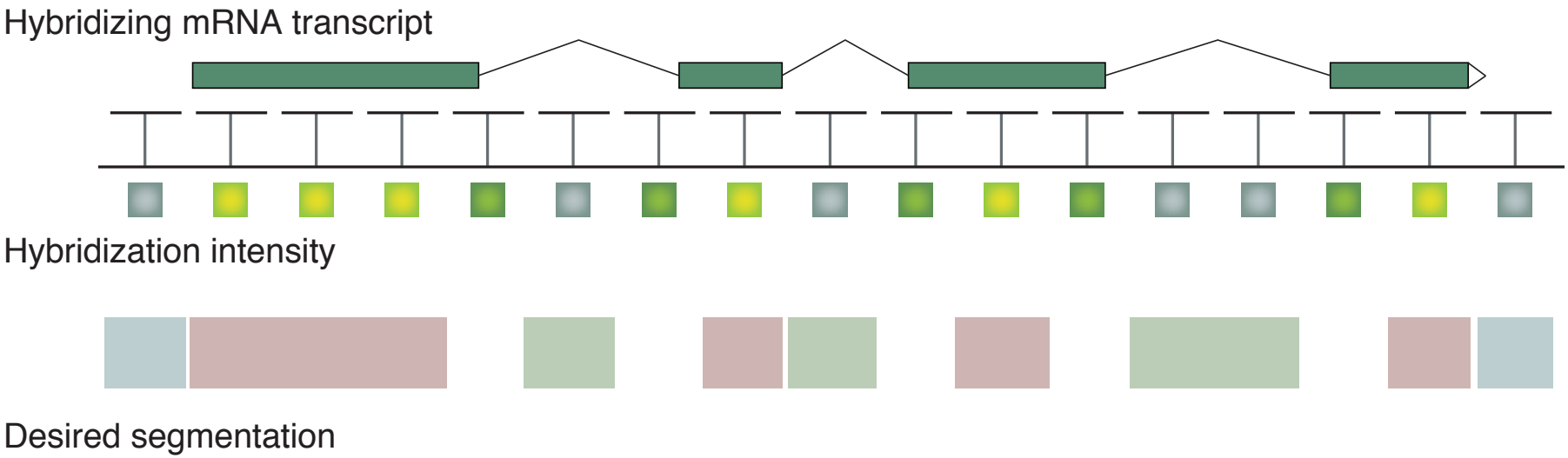
# Detection of Alternative Transcripts (II)




**Goal:** Identify exon/intron segments that show different intensities than *other exons/introns* in at least one analyzed sample.

[Eichner, 2008, Eichner et al., 2009]

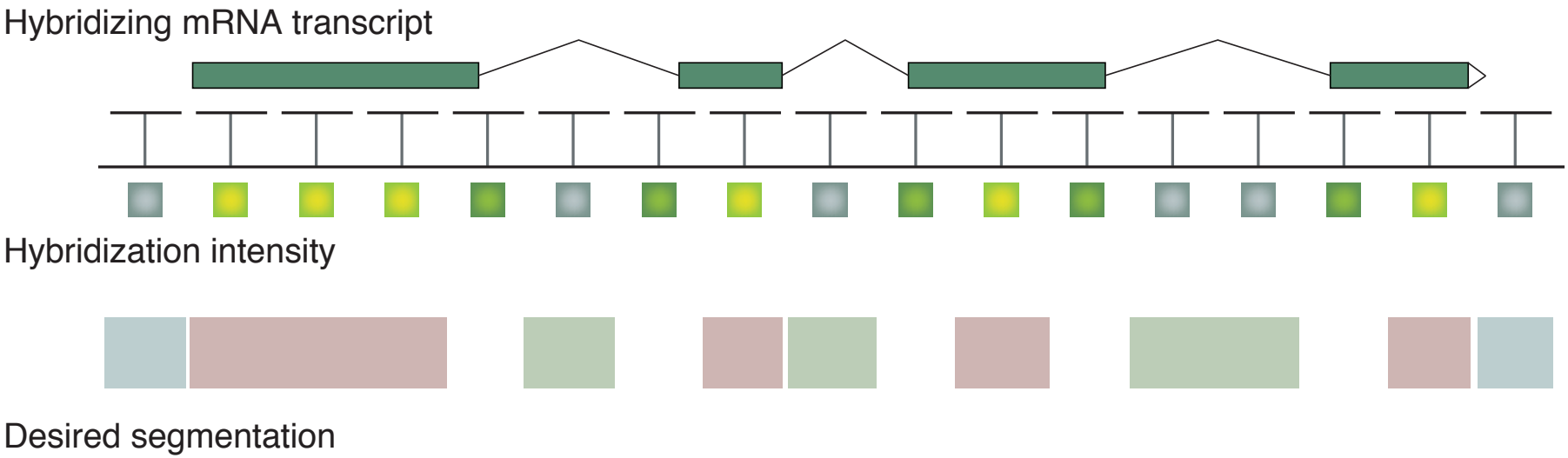
# Tiling Array Segmentation




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

[Zeller et al., 2008b]

# Tiling Array Segmentation



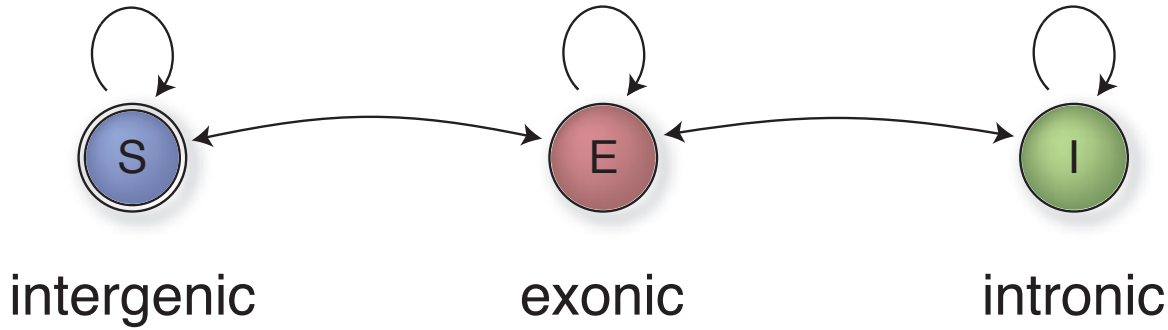

Goal: Characterize tiling probes either as intergenic, exonic or intronic

margin-based segmentation of tiling array data (mSTAD)

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

[Zeller et al., 2008b]

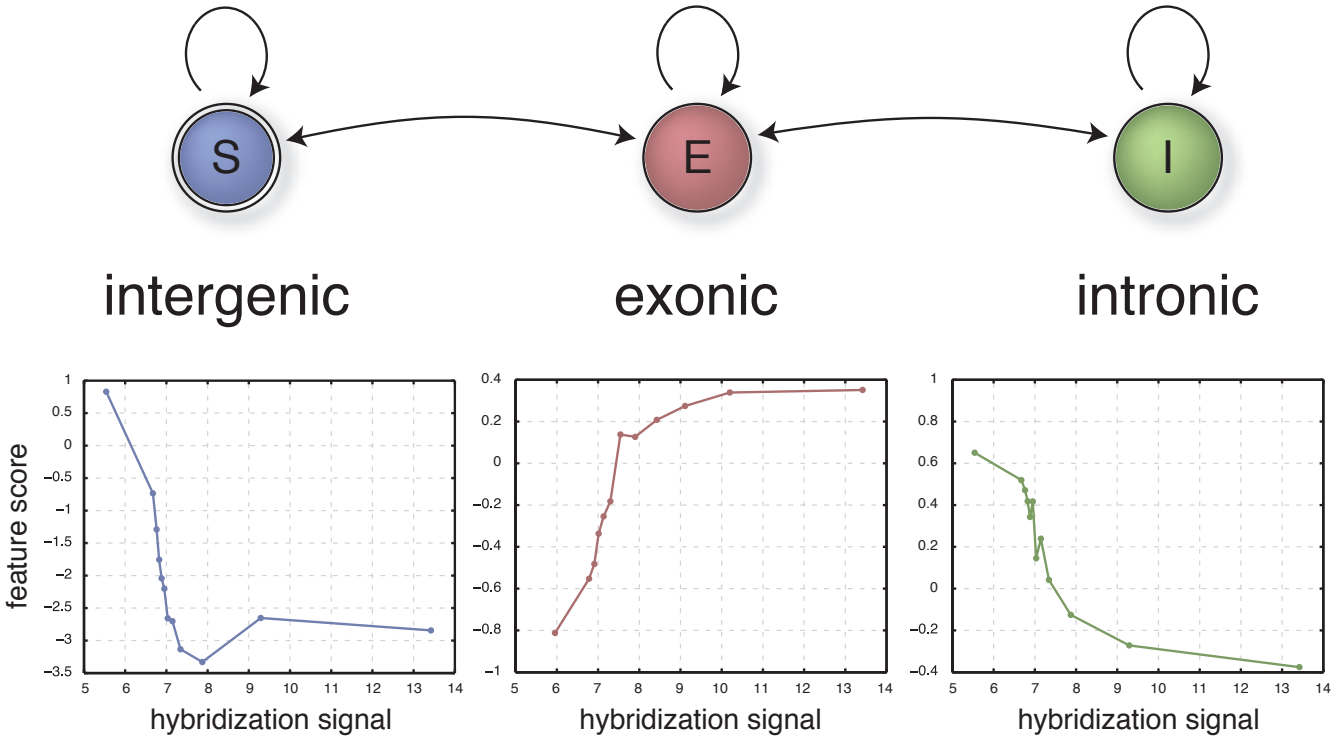
# The mSTAD Approach




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

[Zeller et al., 2008b]

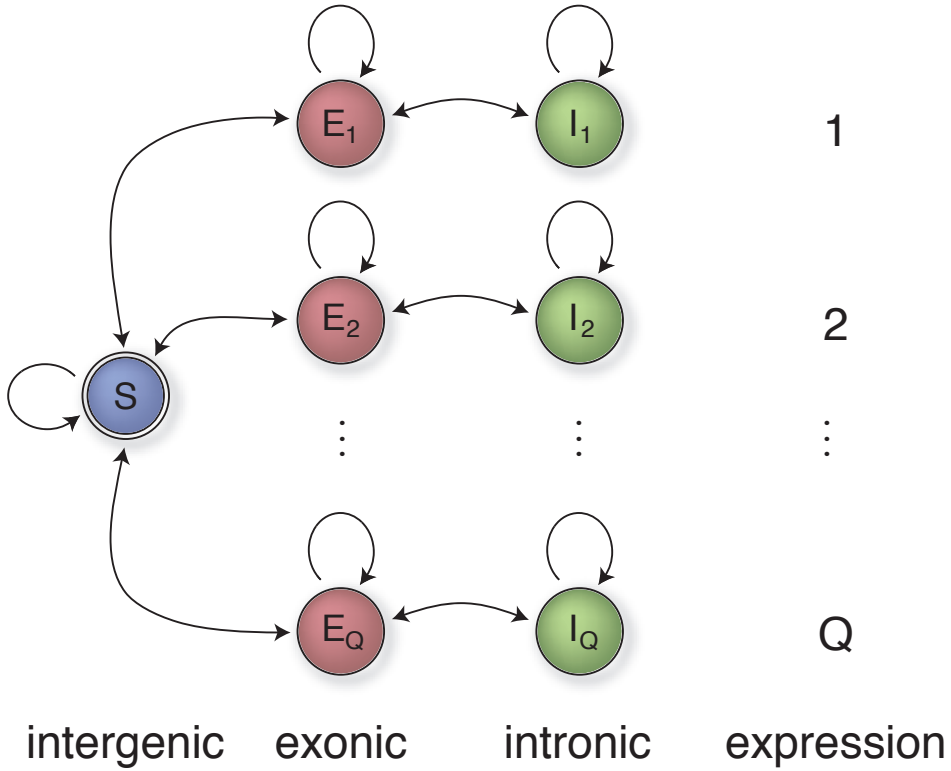
# The mSTAD Approach




- Learn to associate a state with each probe given its hybridization signal and local context
- HM-SVM training: Optimize transformations: signal $\longrightarrow$ score

[Zeller et al., 2008b]

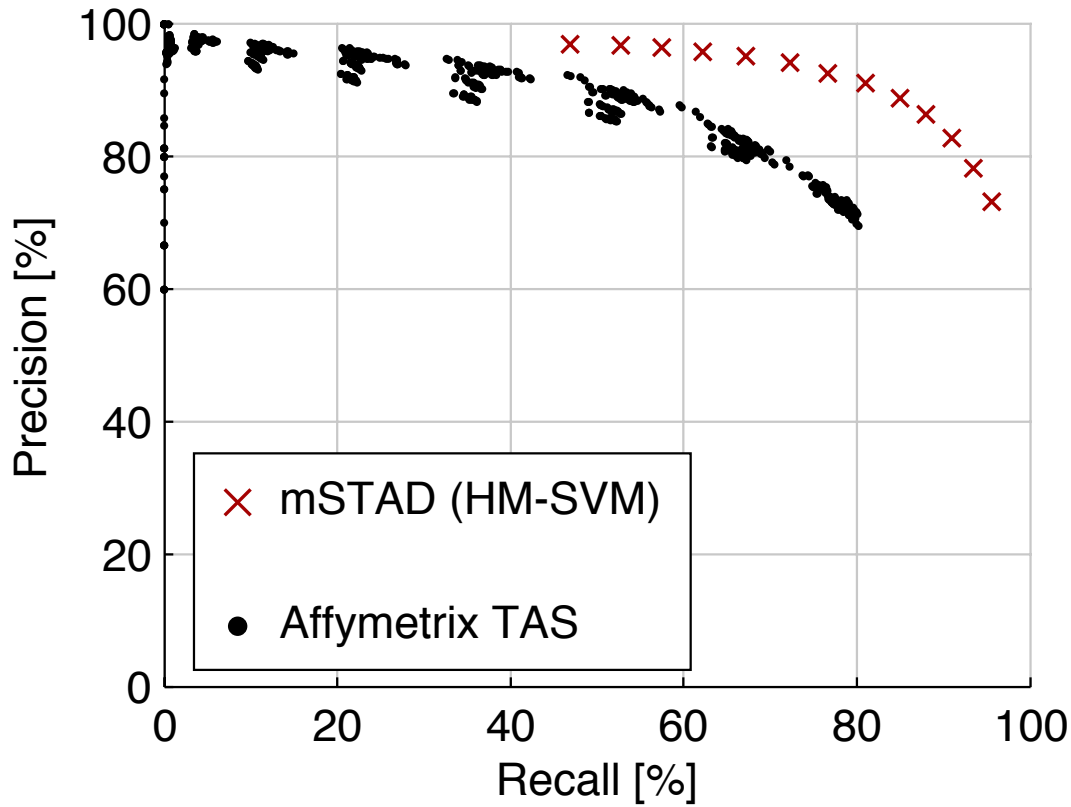
# The mSTAD Approach




- Learn to associate a state with each probe given its hybridization signal and local context
- HM-SVM training: Optimize transformations: signal $\longrightarrow$ score

[Zeller et al., 2008b]

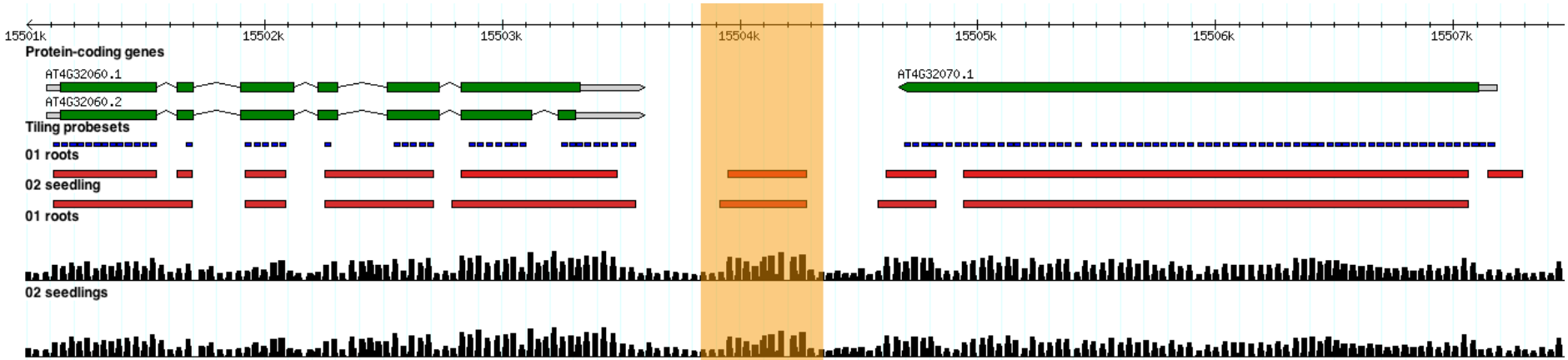
# Method Comparison




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

[Zeller et al., 2008b]

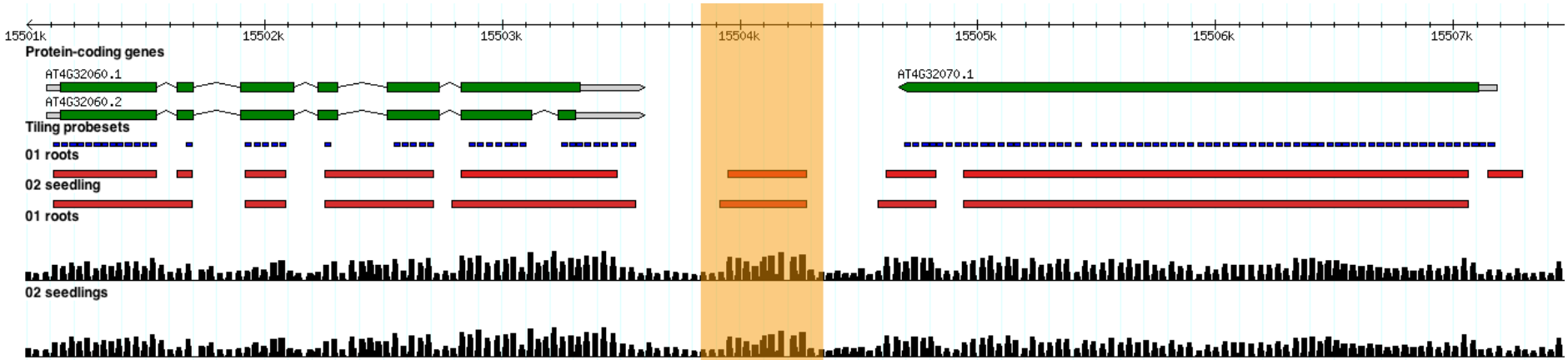
# Transcriptionally Active Regions (TARs)




www.weigelworld.org/resources/microarray/at-tax/

[Laubinger et al., 2008b,a, Zeller et al., 2009]

# Transcriptionally Active Regions (TARs)

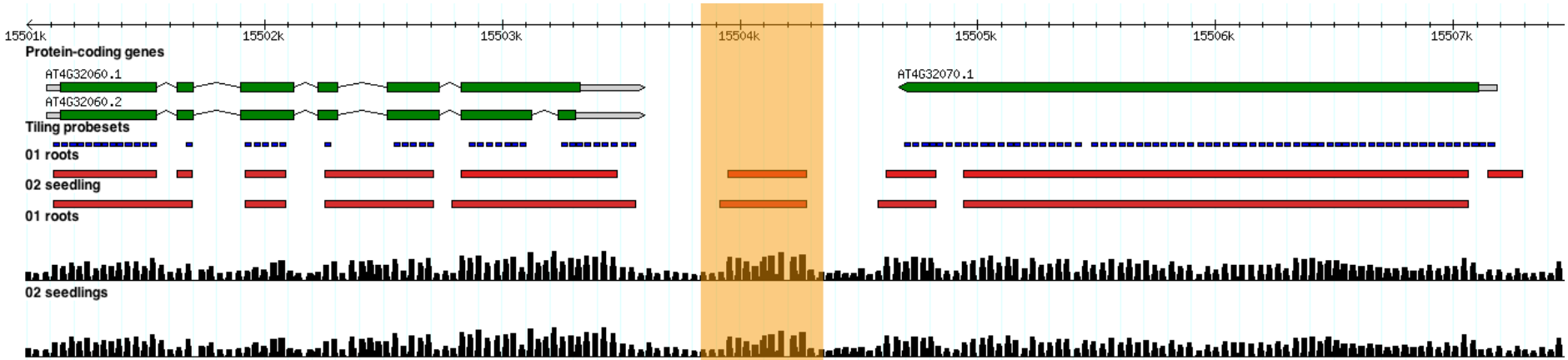



new TAR, not previously apparent from ESTs or cDNAs

`www.weigelworld.org/resources/microarray/at-tax/`

[Laubinger et al., 2008b,a, Zeller et al., 2009]

# Transcriptionally Active Regions (TARs)




www.weigelworld.org/resources/microarray/at-tax/

Applications:

- Discovery of novel transcripts
- Stress induced changes
- Analysis of RNA-processing mutants

[Laubinger et al., 2008b,a, Zeller et al., 2009]

# Sequencing Techniques & Applications




Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

# Sequencing Techniques & Applications




Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

Sequencing Technology

- **Capillary/Sanger sequencing**

# Sequencing Techniques & Applications

Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

Sequencing Technology

- **Capillary/Sanger sequencing**
- Pyrosequencing (Roche/454)
- SOLiD sequencing (ABI)
- Flow cell sequencing (Illumina)

} Next Generation Sequencing

# Sequencing Techniques & Applications

Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

Sequencing Technology

- **Capillary/Sanger sequencing**
- Pyrosequencing (Roche/454)
- SOLiD sequencing (ABI)
- Flow cell sequencing (Illumina)
- Single molecule sequencing (Nano pores, etc.)

} Next Generation Sequencing

# Illumina Sequencing

- *Solexa* released a sequencing machine in 2006
- Fragment sizes from 28 – 100nt
- Probes are fixed to a glass plate "flow cell"
- Reagents are directed through flow cell



(see Movie [Ossowski, 2007])

# Overview / Extensions

| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | 40 – 100 bp | ≈6-20 Gbp | $5 - 8d$ |
| ABI SOLiD | 35 – 50 bp | ≈6-15 Gbp | $6 - 7d$ |
| Roche/454 | 300 – 500 bp | ≈100 Mbp | $7h$ |
| Sanger | 1000 bp | ≈67 kbp | $1h$ |

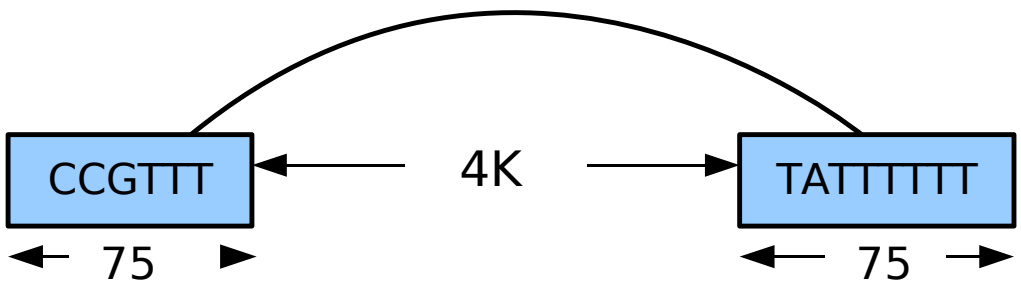
[Sutskever, 2008]

# Overview / Extensions

| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | 40 – 100 bp | ≈6-20 Gbp | 5 – 8*d* |
| ABI SOLiD | 35 – 50 bp | ≈6-15 Gbp | 6 – 7*d* |
| Roche/454 | 300 – 500 bp | ≈100 Mbp | 7*h* |
| Sanger | 1000 bp | ≈67 kbp | 1*h* |

There are several extensions available:

- mate-pair / paired-end




[Sutskever, 2008]

# Overview / Extensions

| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | 40 – 100 bp | ≈6-20 Gbp | $5 - 8d$ |
| ABI SOLiD | 35 – 50 bp | ≈6-15 Gbp | $6 - 7d$ |
| Roche/454 | 300 – 500 bp | ≈100 Mbp | $7h$ |
| Sanger | 1000 bp | ≈67 kbp | $1h$ |

There are several extensions available:

- mate-pair / paired-end
- bisulfite treatment

[Sutskever, 2008]

# Overview / Extensions

| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | $40 - 100$ bp | ≈6-20 Gbp | $5 - 8d$ |
| ABI SOLiD | $35 - 50$ bp | ≈6-15 Gbp | $6 - 7d$ |
| Roche/454 | $300 - 500$ bp | ≈100 Mbp | $7h$ |
| Sanger | 1000 bp | ≈67 kbp | $1h$ |

There are several extensions available:

- mate-pair / paired-end
- bisulfite treatment
- multiplexing $8 \times 12$
  $\rightsquigarrow$ 96 samples per flow cell

[Sutskever, 2008]

# Short Reads Analysis - Methods

After sequencing:

- Base Calling (image analysis & phasing correction)



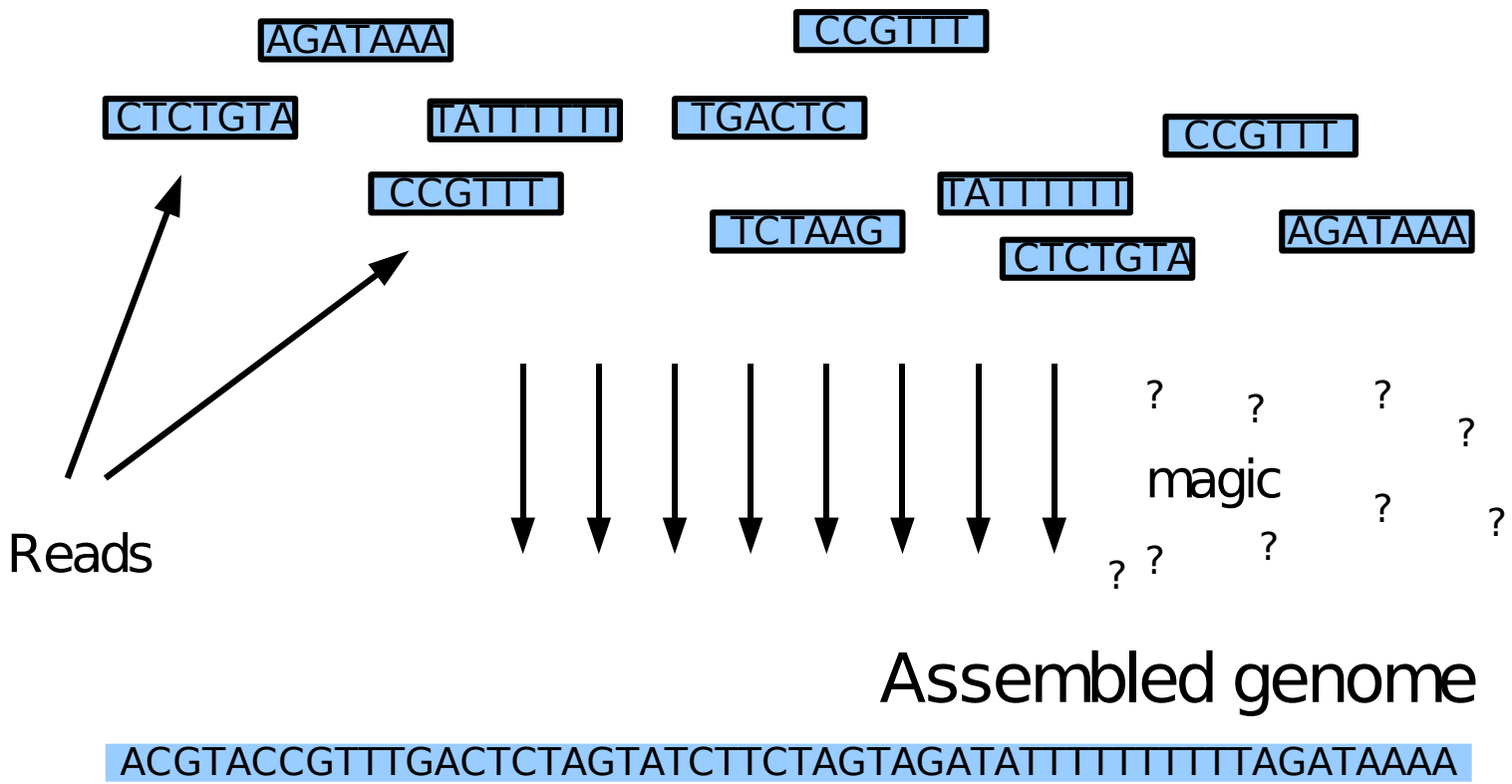
GCTGA

[Erlich et al., 2008]

# Short Reads Analysis - Methods

After sequencing:

- Base Calling (image analysis & phasing correction)
- Assembly (generate larger contigs)




[Zerbino and Birney, 2008, Sutskever, 2008]

# Short Reads Analysis - Methods

After sequencing:

- Base Calling (image analysis & phasing correction)
- Assembly (generate larger contigs)
- Mapping/Alignments (map reads to reference genome)

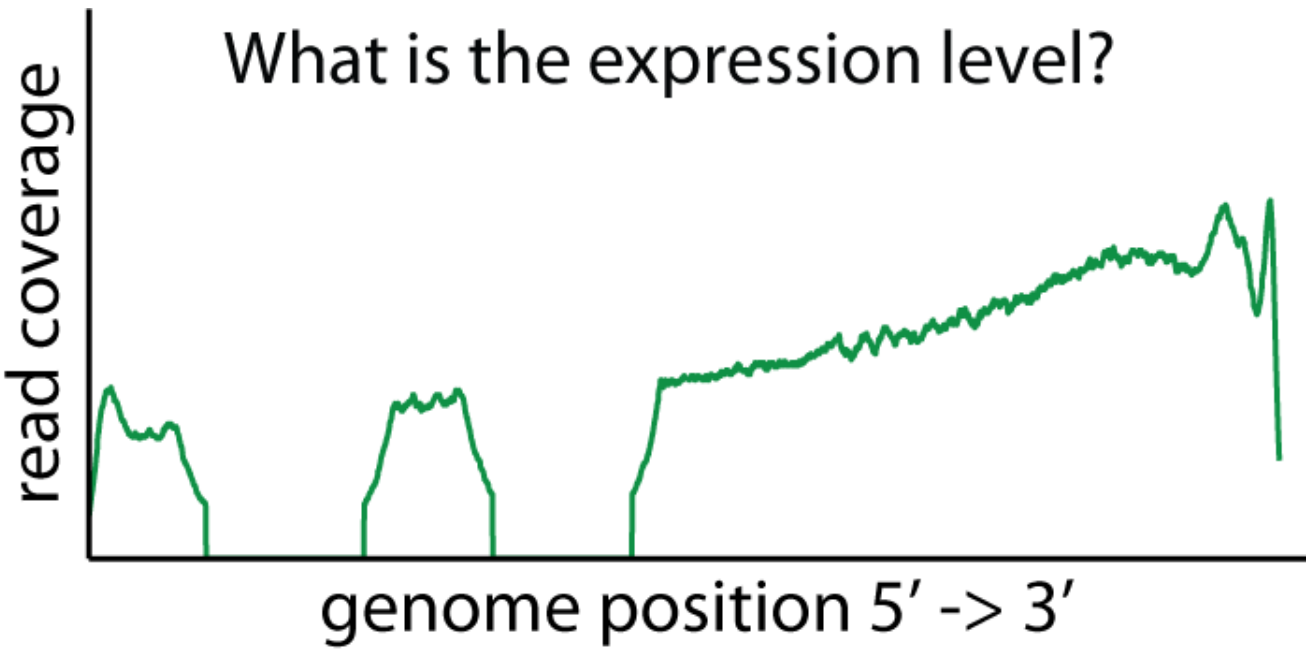
```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
       ACCAGTGACCTGACTACTACGTCGTAACGTACACG
```

[Ossowski et al., 2008, Rumble et al., 2009, Li et al., 2009]

# Short Reads Analysis - Methods

After sequencing:

- Base Calling (image analysis & phasing correction)
- Assembly (generate larger contigs)
- Mapping/Alignments (map reads to reference genome)
- Identification and quantification of transcripts




[Behr et al., 2008, Jiang and Wong, 2009, Bohnert et al., 2009, Sammeth, 2009a]

# Short Reads Analysis - Methods



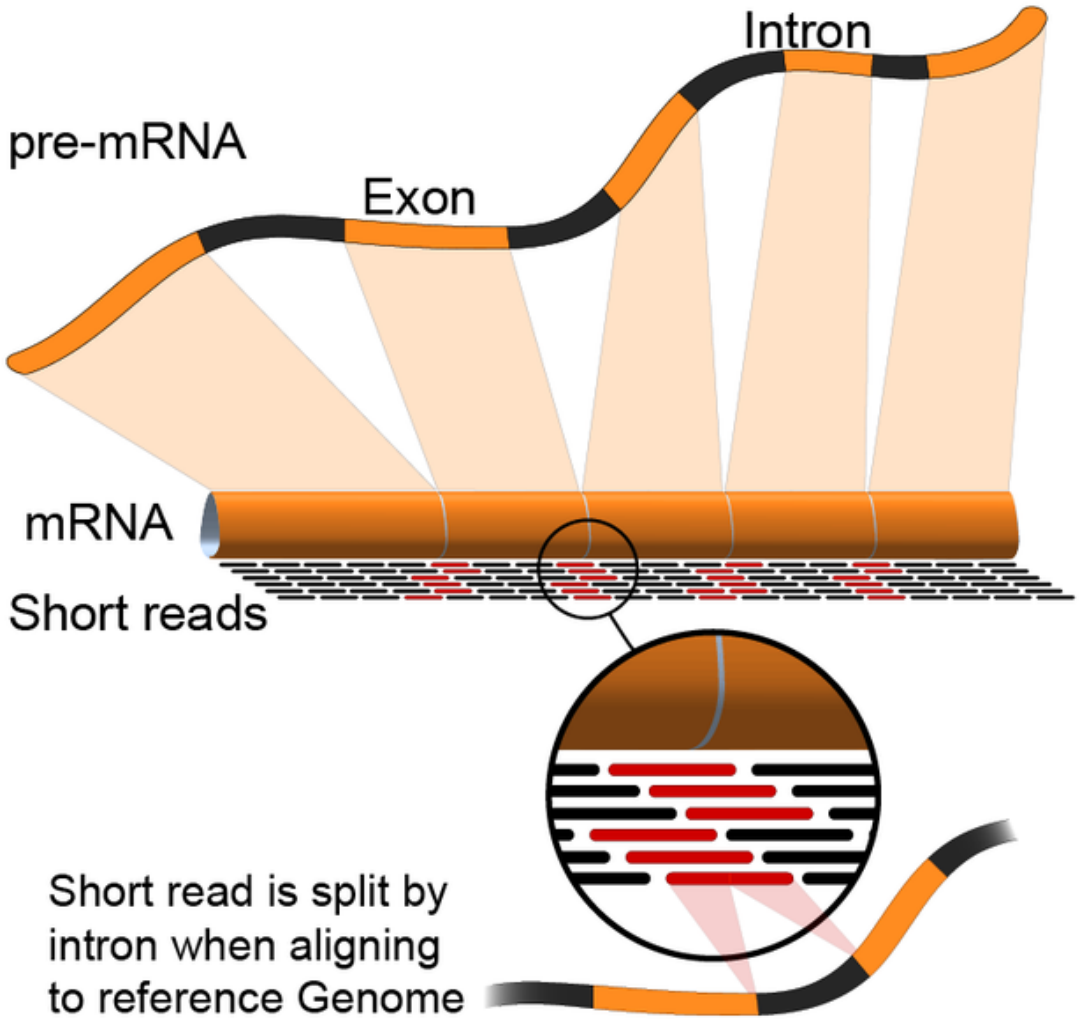

After sequencing:

- Base Calling (image analysis & phasing correction)
- Assembly (generate larger contigs)
- Mapping/Alignments (map reads to reference genome)
- Identification and quantification of transcripts

**Problem:** 100 million reads of short length often with many errors

$\Rightarrow$ Big computational challenge

# Spliced vs. Unspliced Alignments




[Wikipedia]

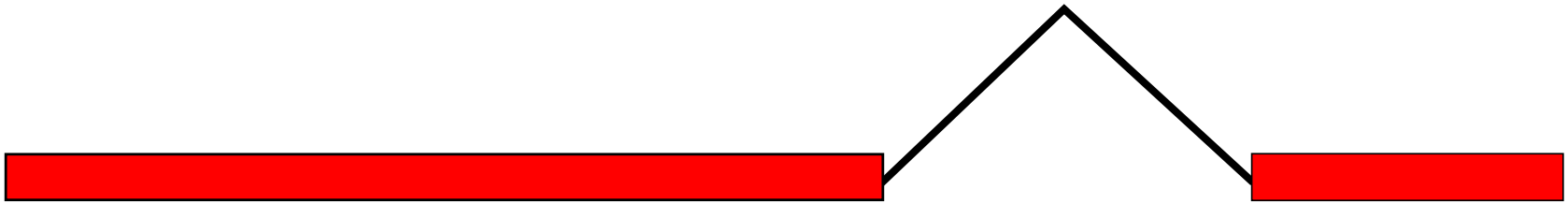
# Spliced vs. Unspliced Alignments



...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC

- Find matching region on genome with a few mismatches

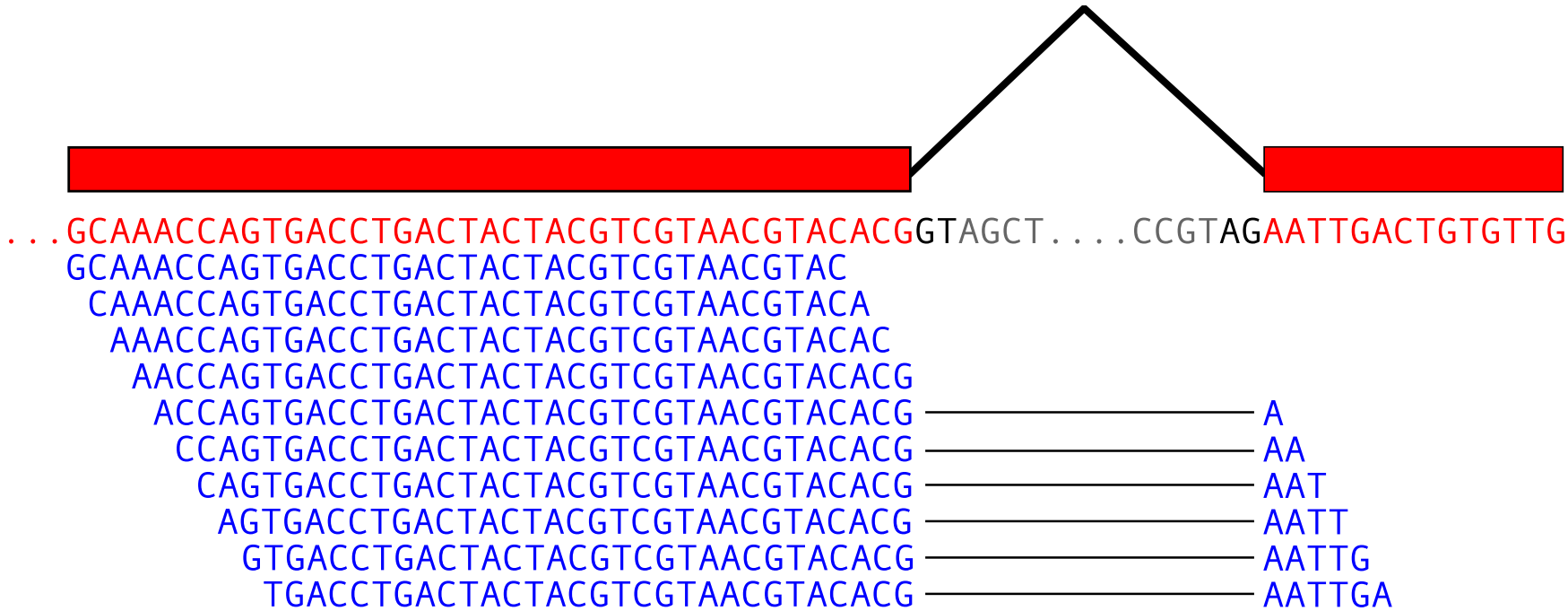
# Spliced vs. Unspliced Alignments



```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
```

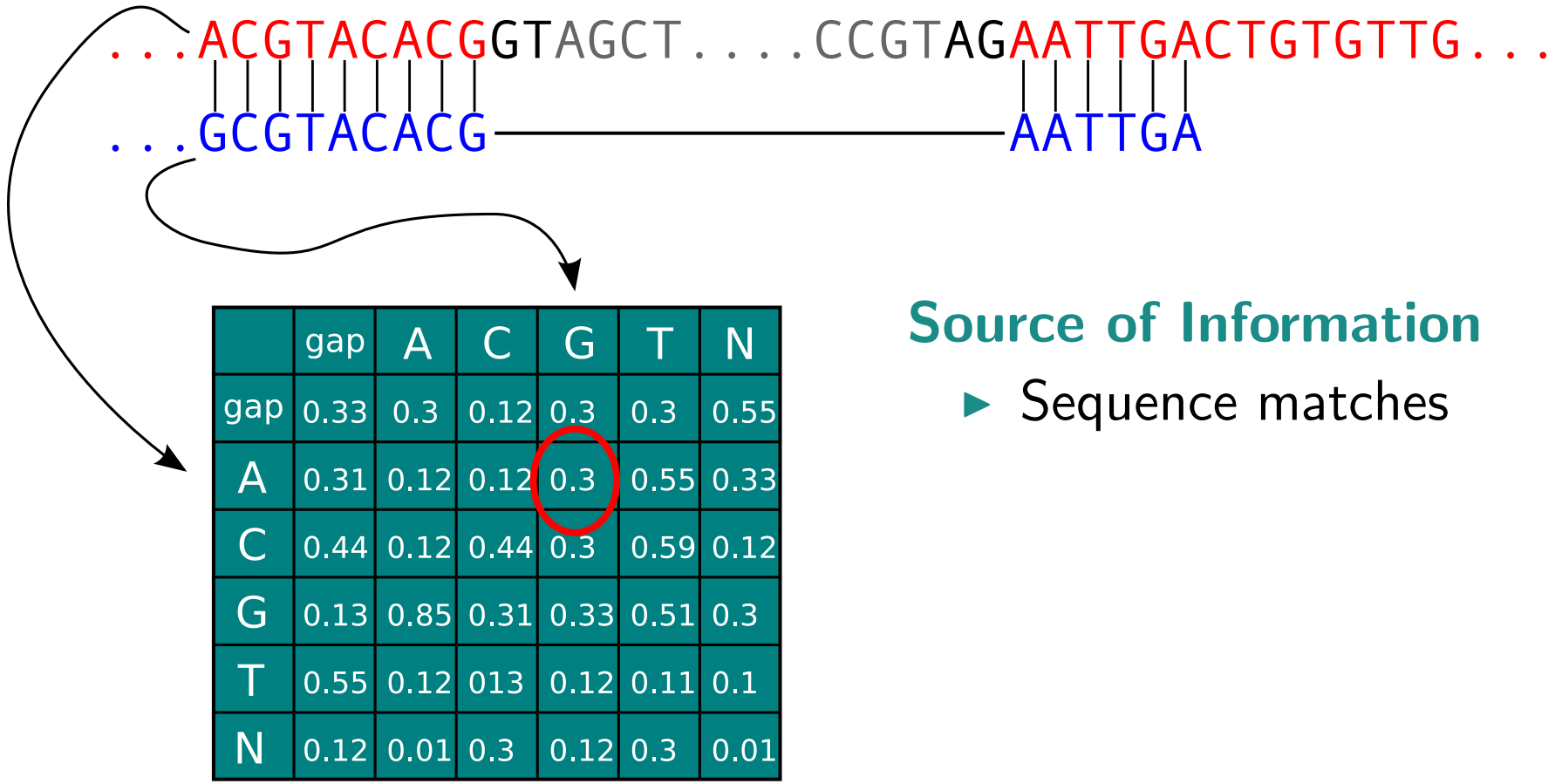
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads

# 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

# Extended Smith-Waterman Algorithm



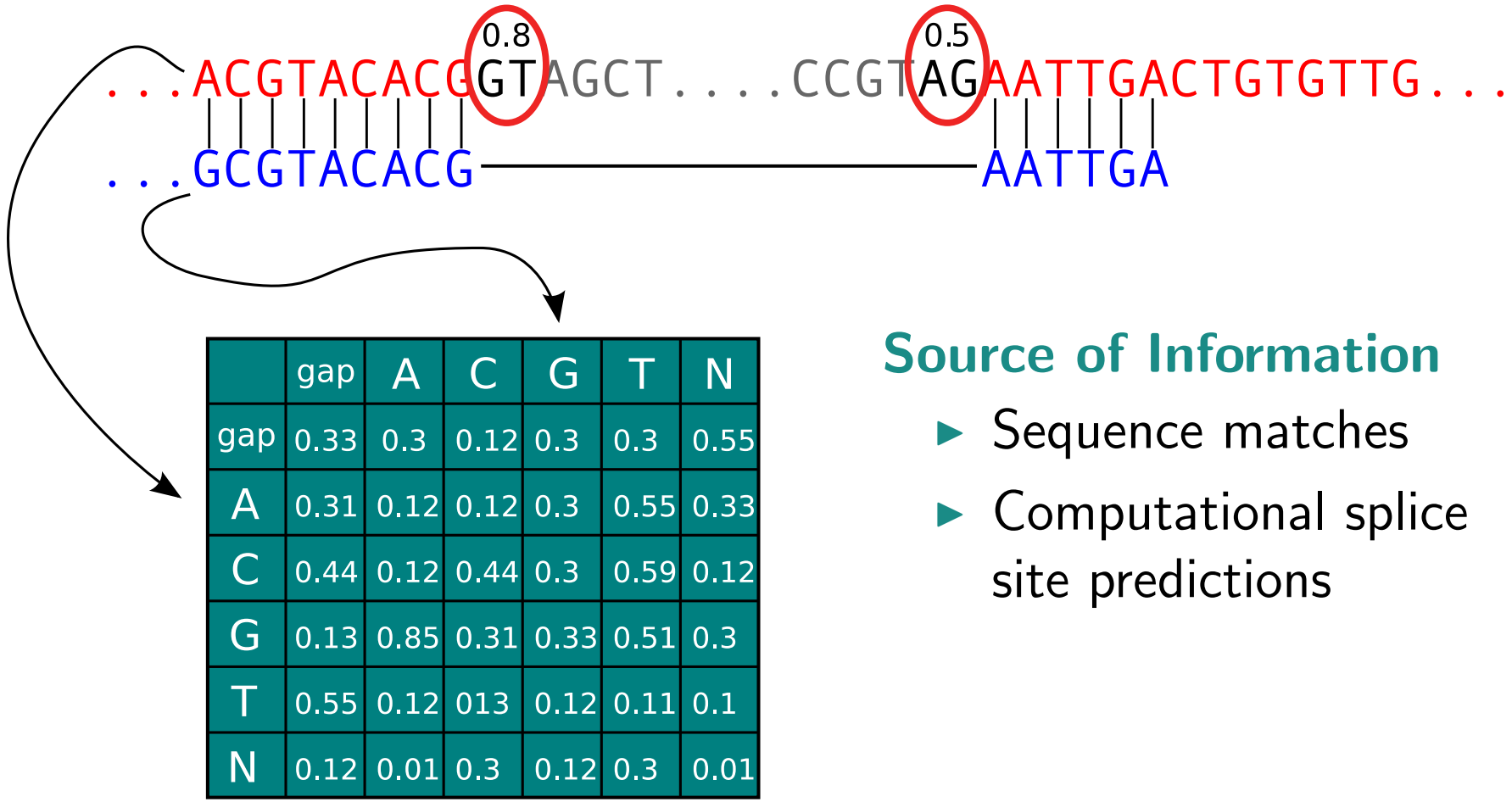

| | gap | A | C | G | T | N |
|---|---|---|---|---|---|---|
| gap | 0.33 | 0.3 | 0.12 | 0.3 | 0.3 | 0.55 |
| A | 0.31 | 0.12 | 0.12 | 0.3 | 0.55 | 0.33 |
| C | 0.44 | 0.12 | 0.44 | 0.3 | 0.59 | 0.12 |
| G | 0.13 | 0.85 | 0.31 | 0.33 | 0.51 | 0.3 |
| T | 0.55 | 0.12 | 013 | 0.12 | 0.11 | 0.1 |
| N | 0.12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

## Source of Information

- Sequence matches

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

# Extended Smith-Waterman Algorithm



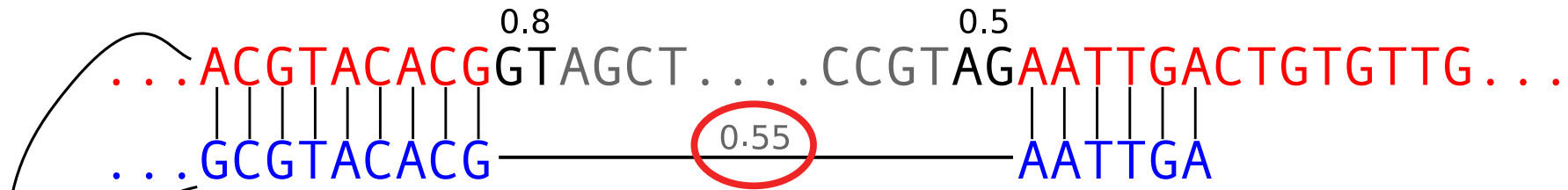

| | gap | A | C | G | T | N |
|---|---|---|---|---|---|---|
| gap | 0.33 | 0.3 | 0.12 | 0.3 | 0.3 | 0.55 |
| A | 0.31 | 0.12 | 0.12 | 0.3 | 0.55 | 0.33 |
| C | 0.44 | 0.12 | 0.44 | 0.3 | 0.59 | 0.12 |
| G | 0.13 | 0.85 | 0.31 | 0.33 | 0.51 | 0.3 |
| T | 0.55 | 0.12 | 013 | 0.12 | 0.11 | 0.1 |
| N | 0.12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

## Source of Information

- Sequence matches
- Computational splice site predictions

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

# Extended Smith-Waterman Algorithm



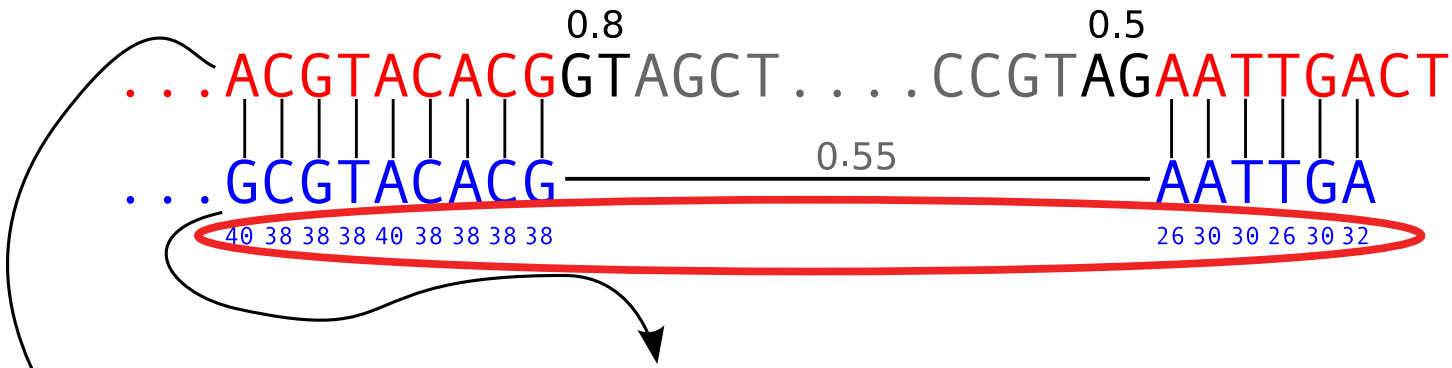

| | gap | A | C | G | T | N |
|---|---|---|---|---|---|---|
| gap | 0.33 | 0.3 | 0.12 | 0.3 | 0.3 | 0.55 |
| A | 0.31 | 0.12 | 0.12 | 0.3 | 0.55 | 0.33 |
| C | 0.44 | 0.12 | 0.44 | 0.3 | 0.59 | 0.12 |
| G | 0.13 | 0.85 | 0.31 | 0.33 | 0.51 | 0.3 |
| T | 0.55 | 0.12 | 013 | 0.12 | 0.11 | 0.1 |
| N | 0.12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

## Source of Information

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

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

# Extended Smith-Waterman Algorithm




| | 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

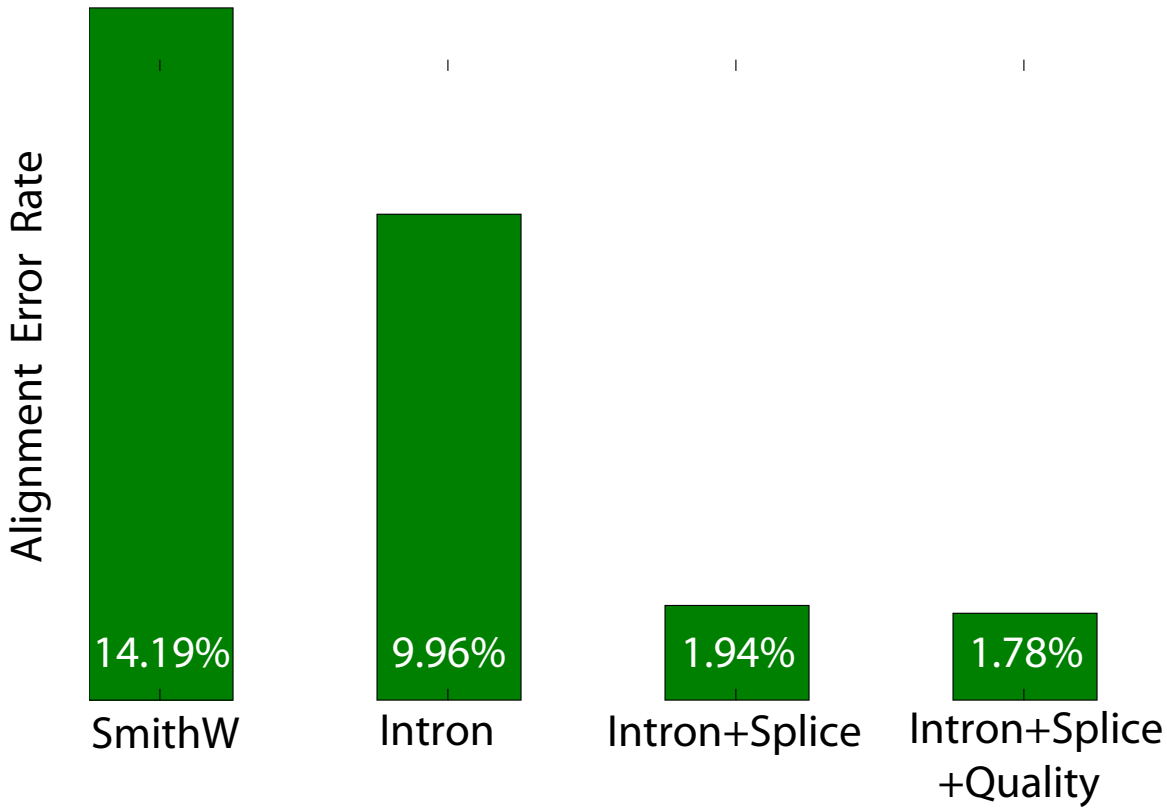
Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

[De Bona et al., 2008]

# QPalma's Accurate Alignments

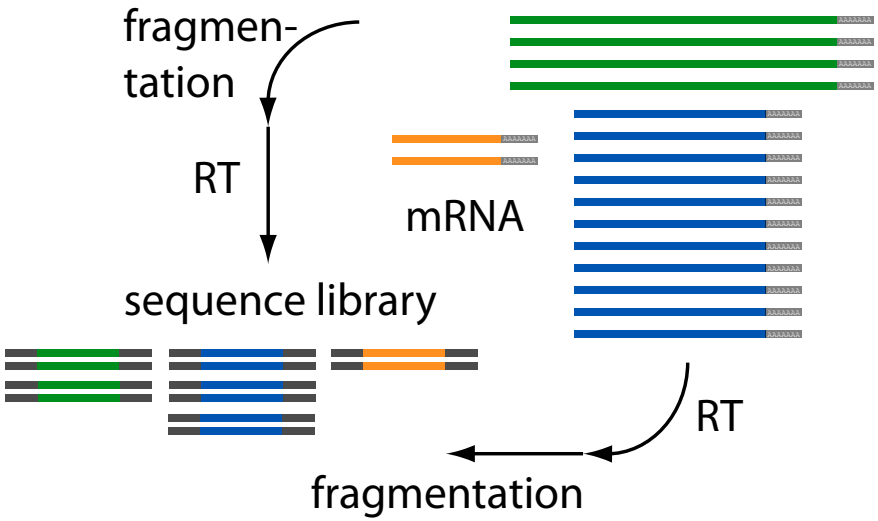
Generate set of artificially spliced reads

- 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]

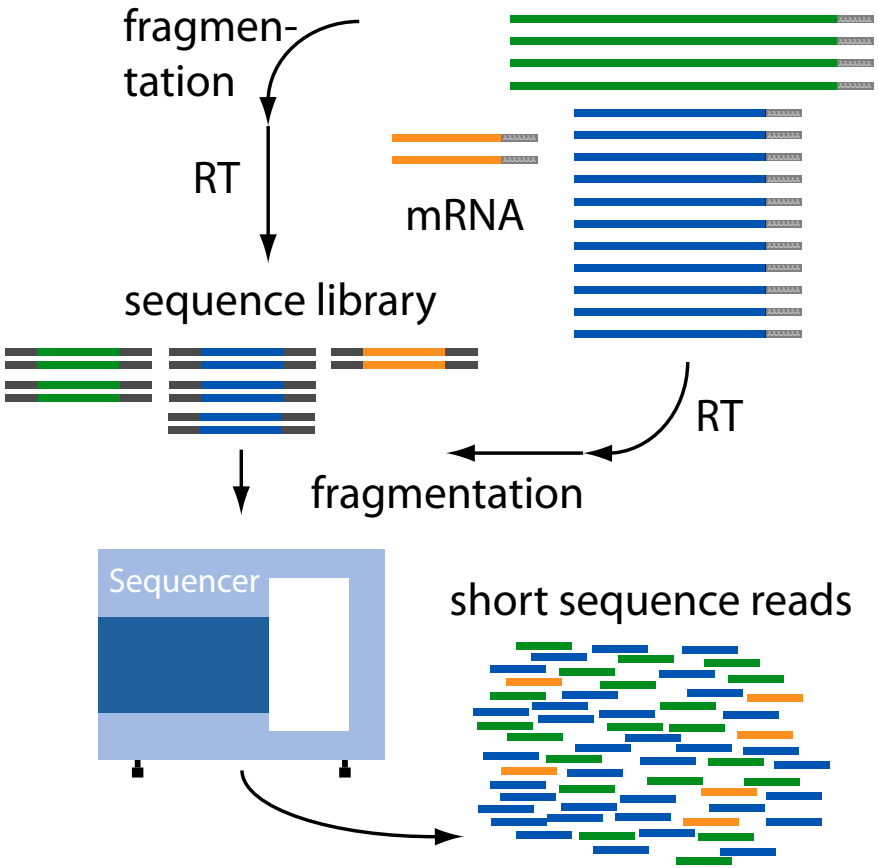
# RNA Sequencing and Biases




## Biases due to ...

- cDNA library construction

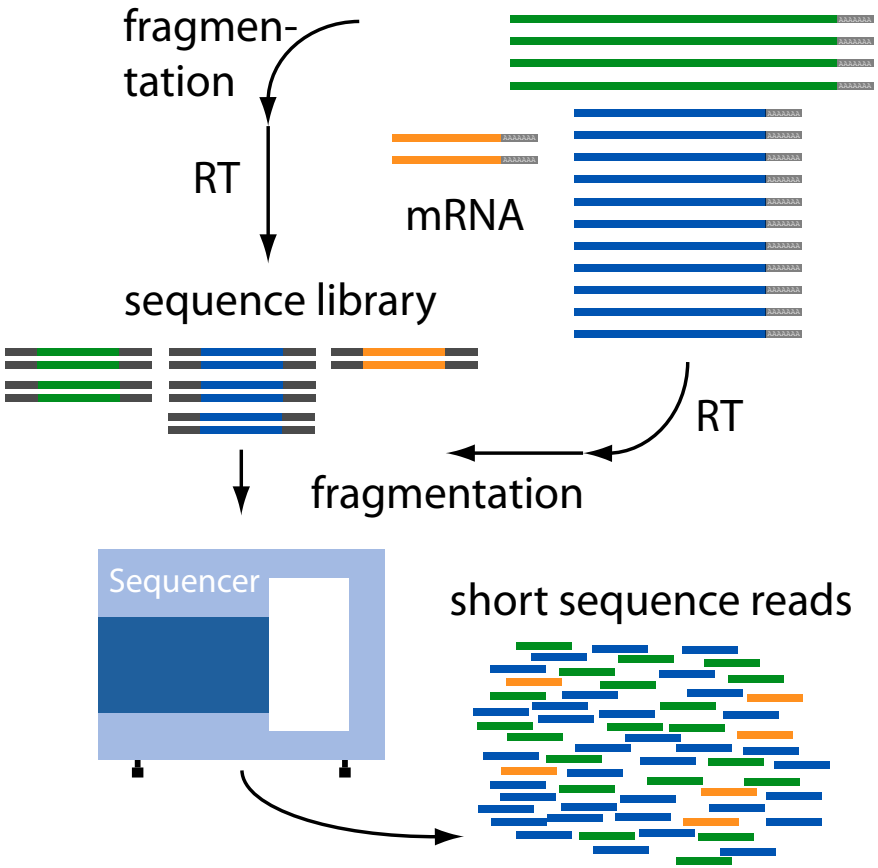
# RNA Sequencing and Biases




## Biases due to ...

- cDNA library construction
- Sequencing

# RNA Sequencing and Biases



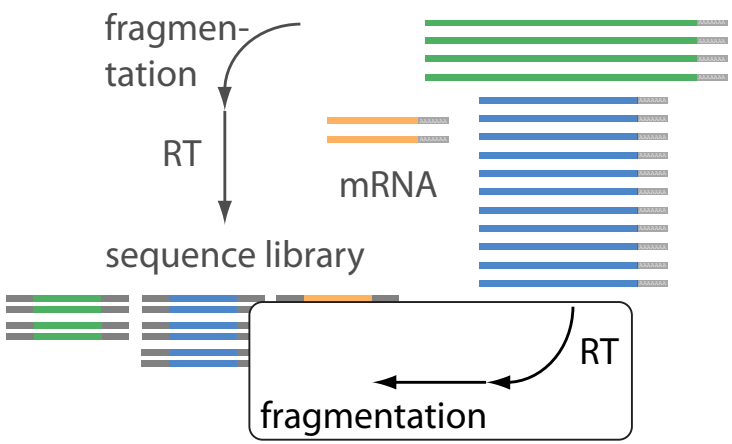

## Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping

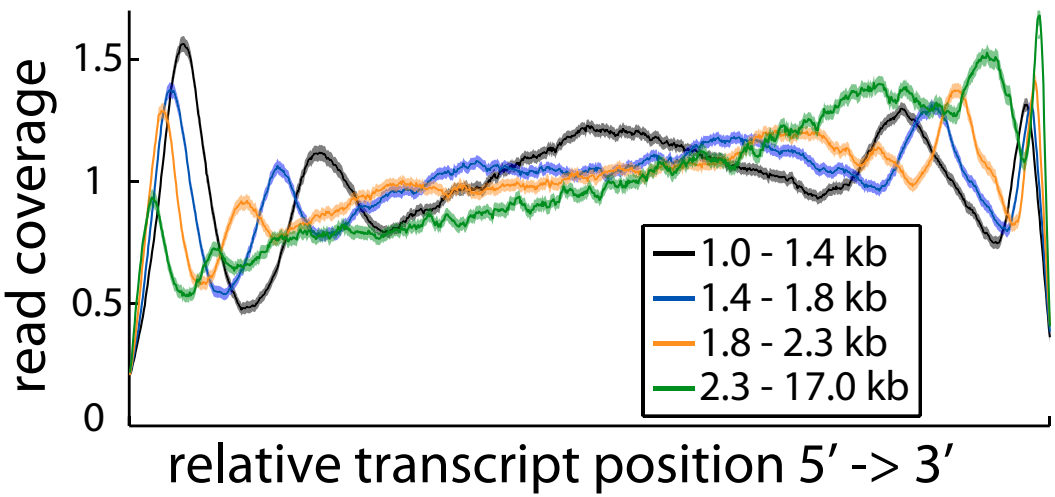


# Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript




Transcript profiles for different lengths

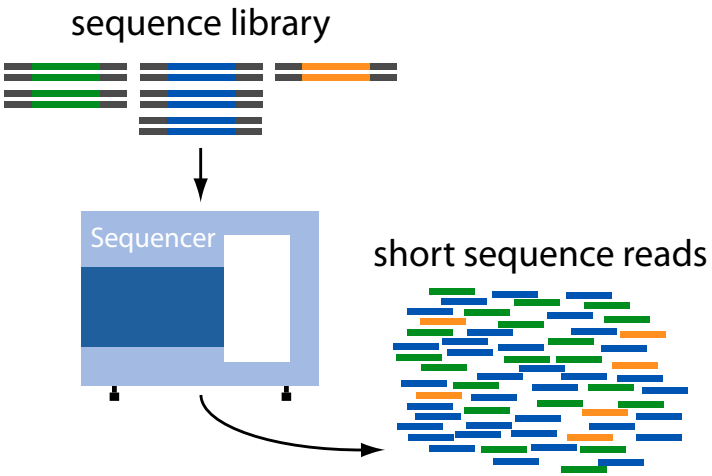



Data generated for *A. thaliana* with Flux Simulator v1.0 [Sammeth, 2009b]

- Oligo-dT priming
- Physical cDNA fragmentation

# Other Biases

## PCR Bias



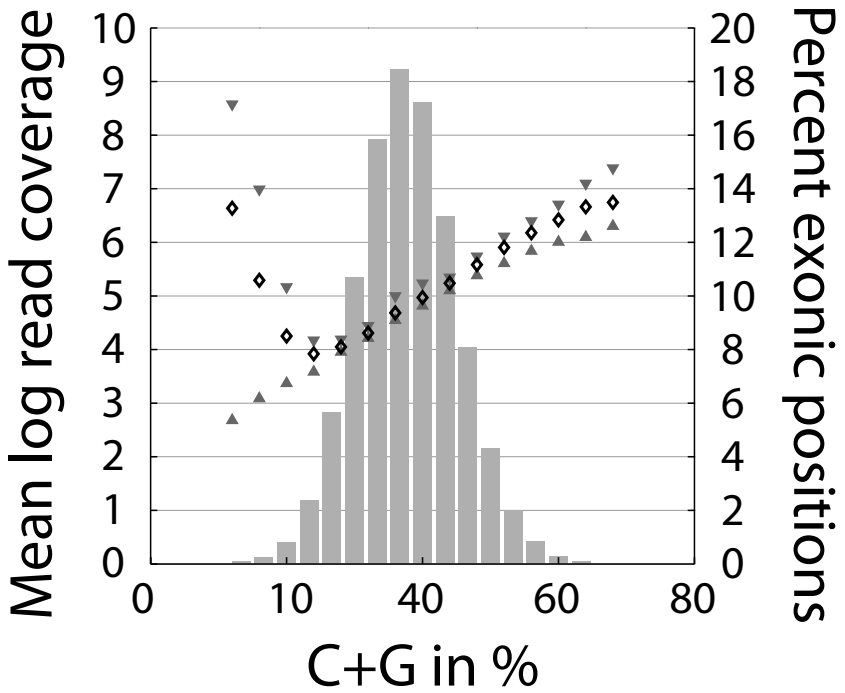




Figure provided by Georg Zeller

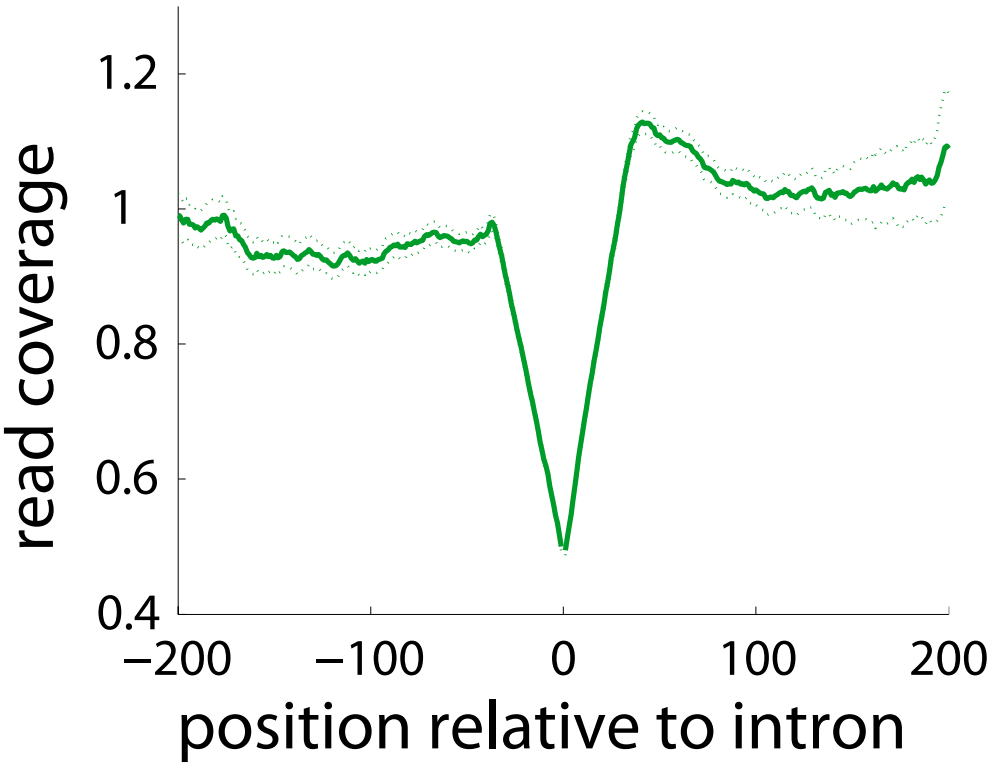
RNA-Seq data (*A. thaliana*)

- Exonic GC content

# Other Biases

## Read Mapping Bias

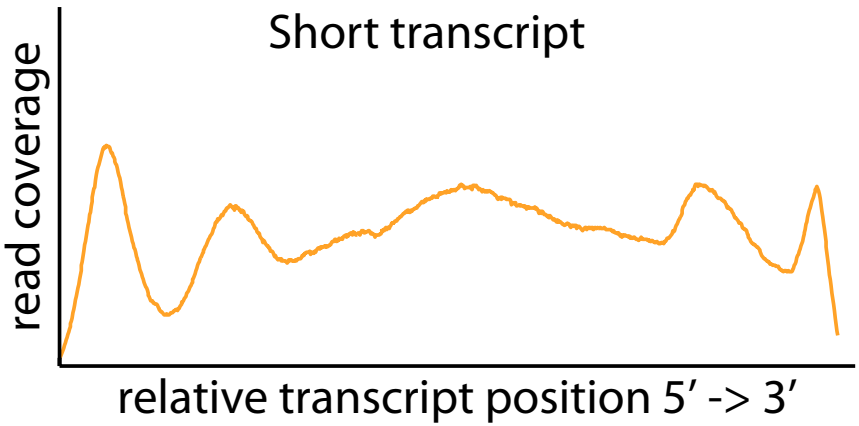



RNA-Seq data (*A. thaliana*)

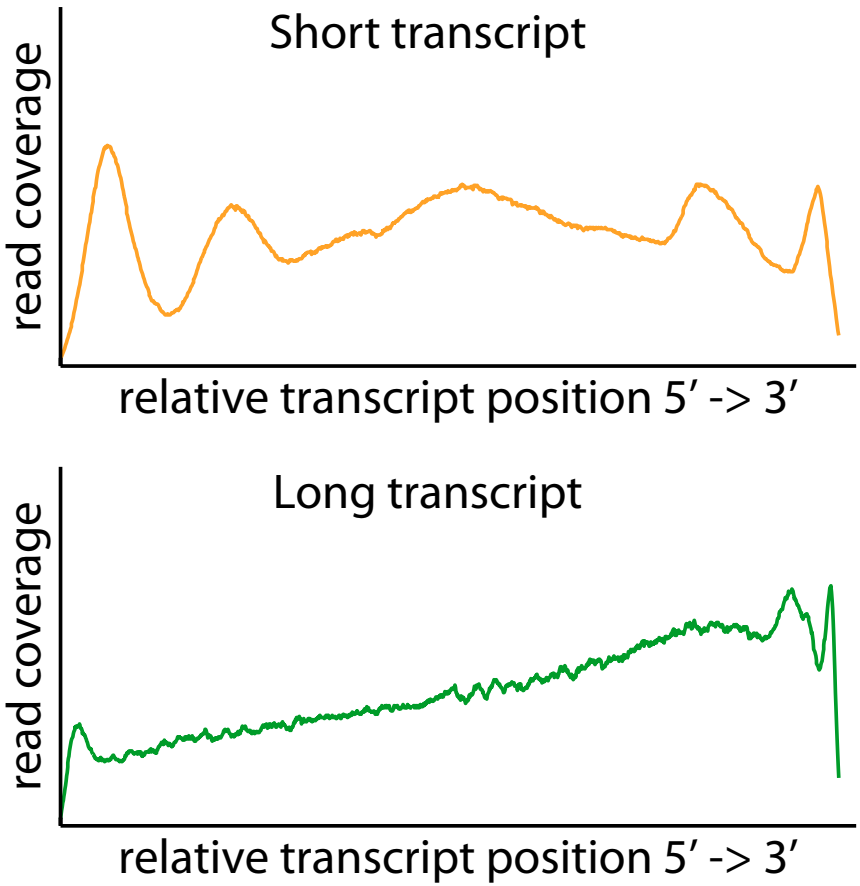
- Exon boundaries

# rQuant – Basic Idea

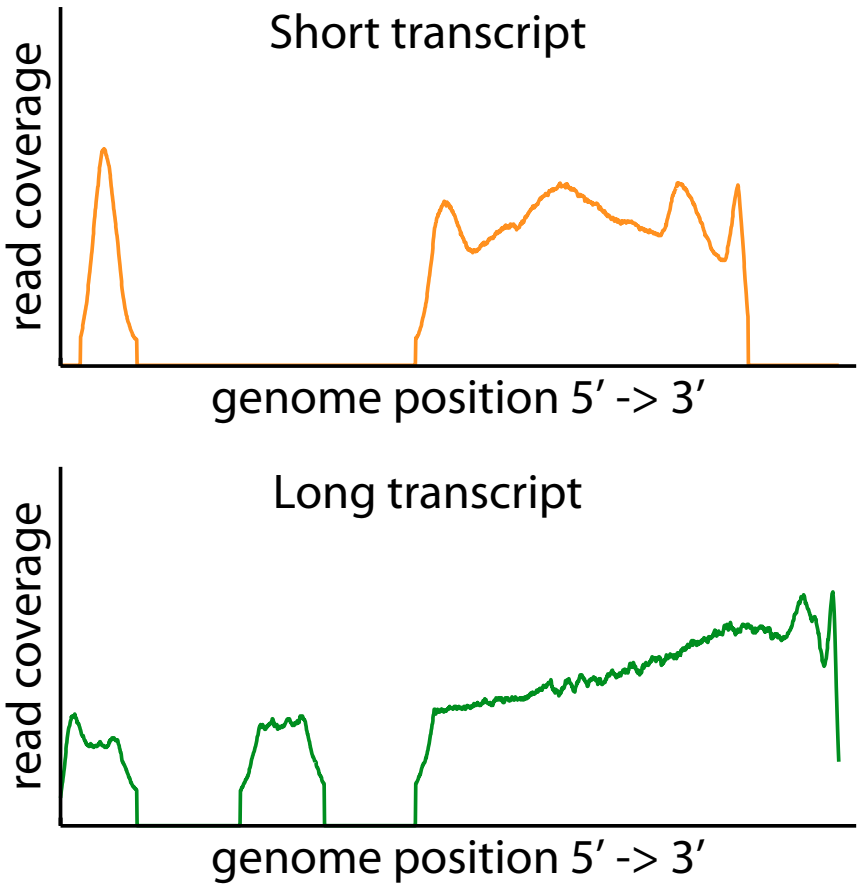




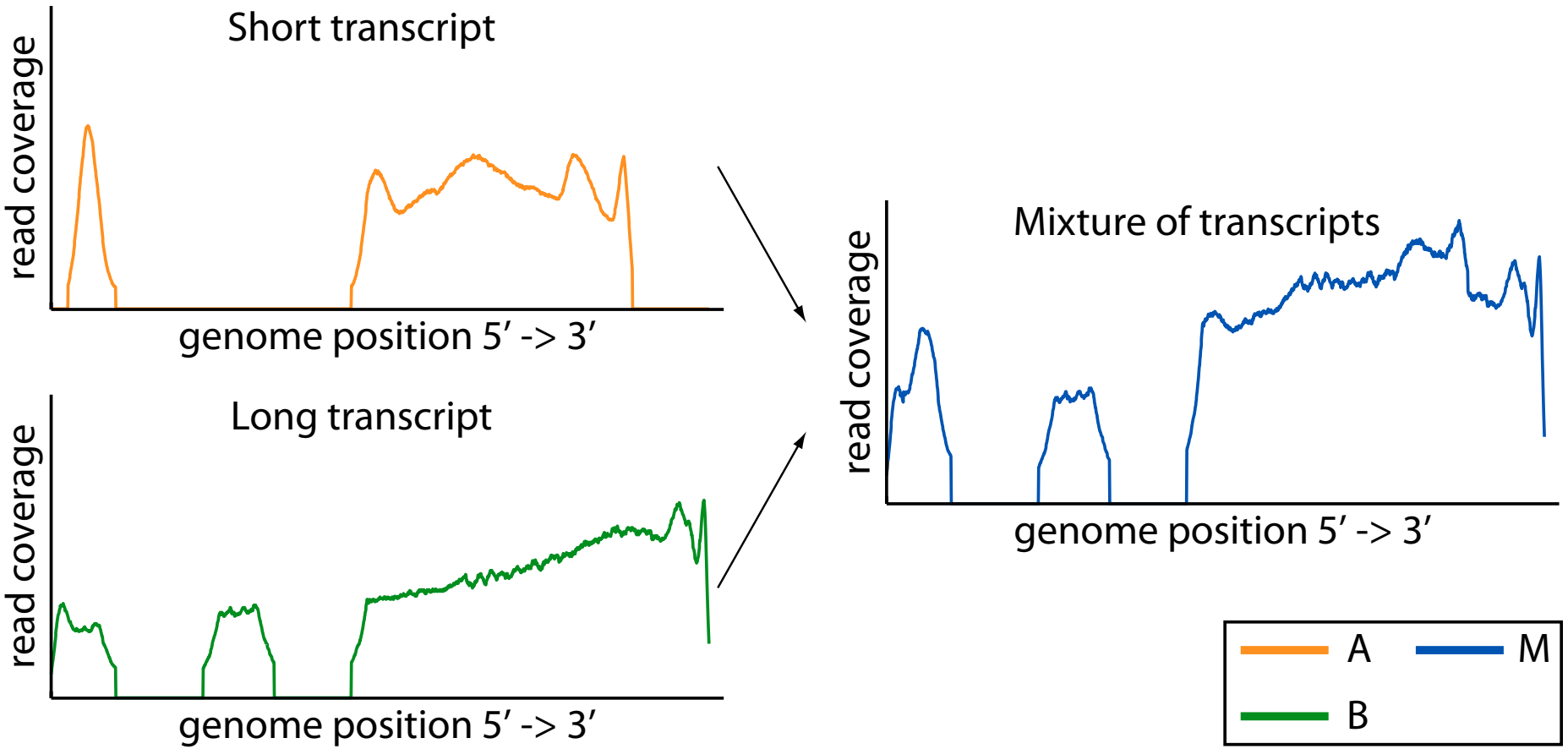

# rQuant – Basic Idea



# rQuant – Basic Idea

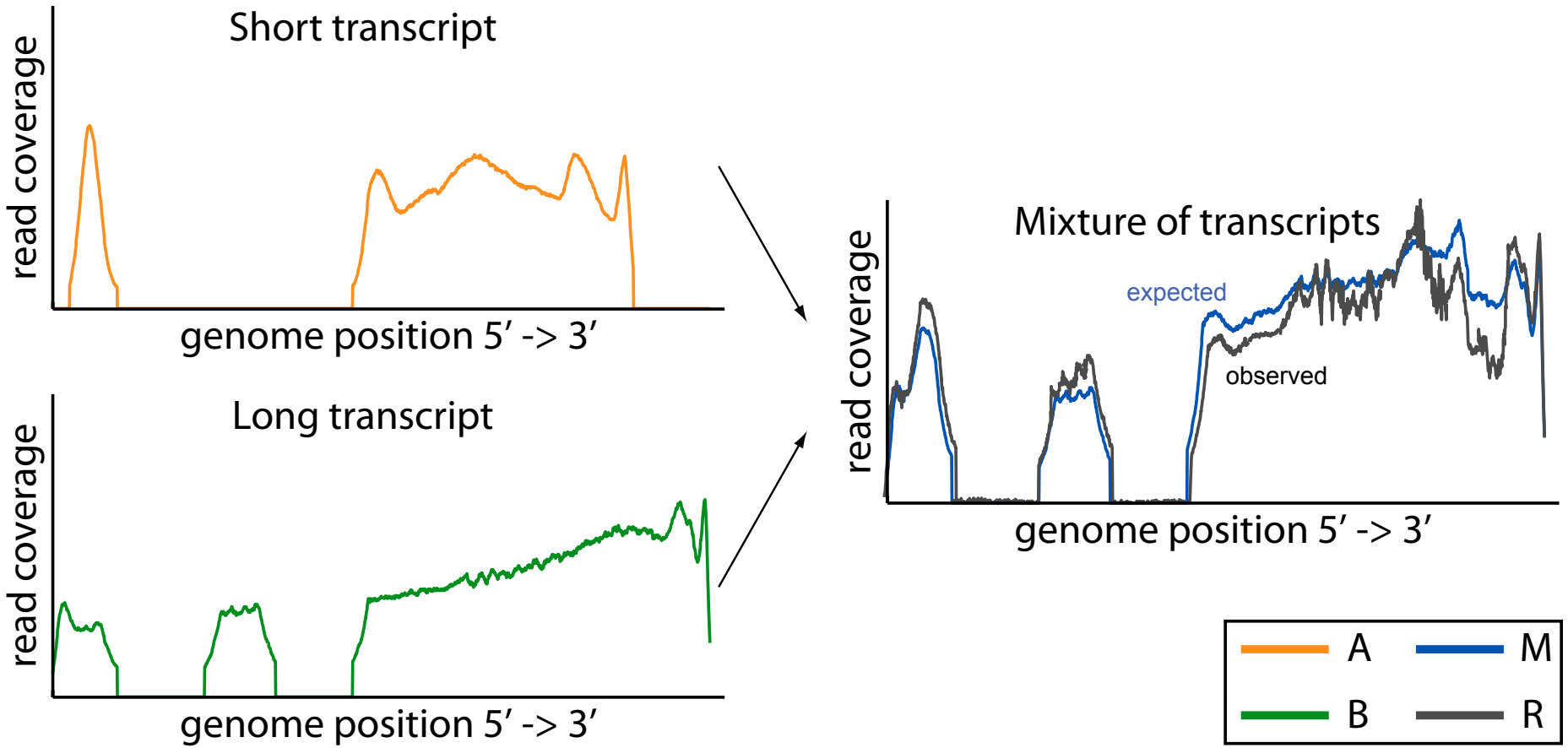


# rQuant – Basic Idea




$$M_i = w_A A_i + w_B B_i$$

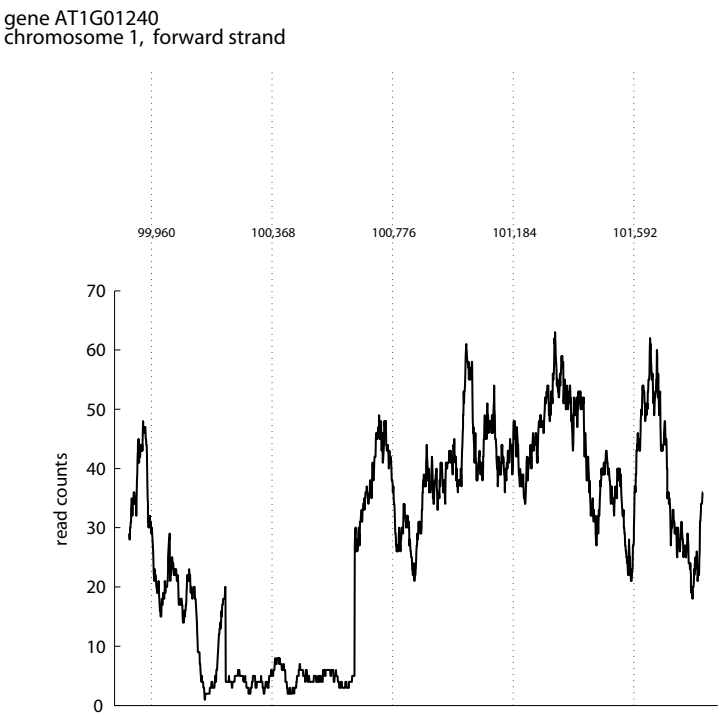
# rQuant – Basic Idea




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

# rQuant – Iterative Algorithm

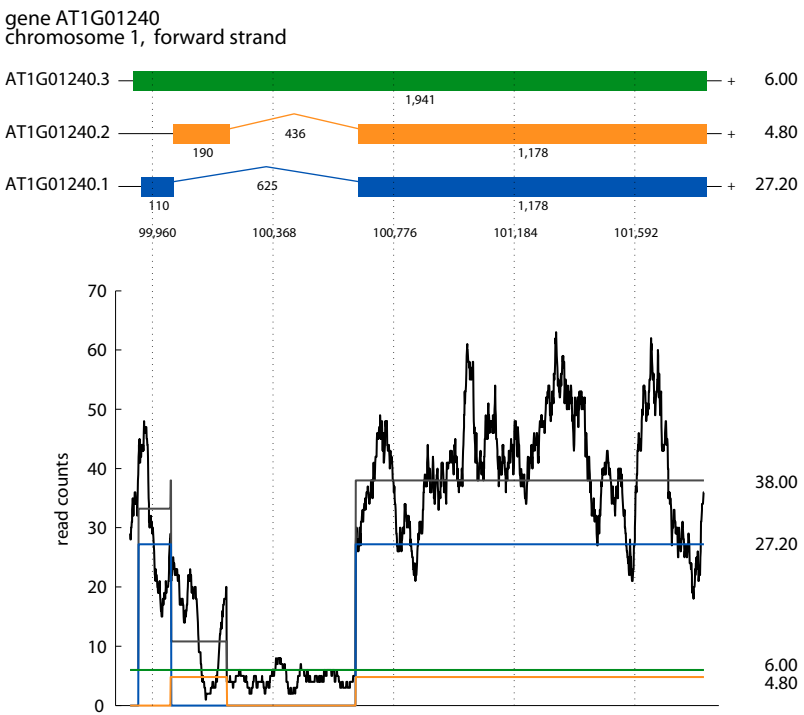
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$




[Bohnert et al., 2009]

# rQuant – Iterative Algorithm

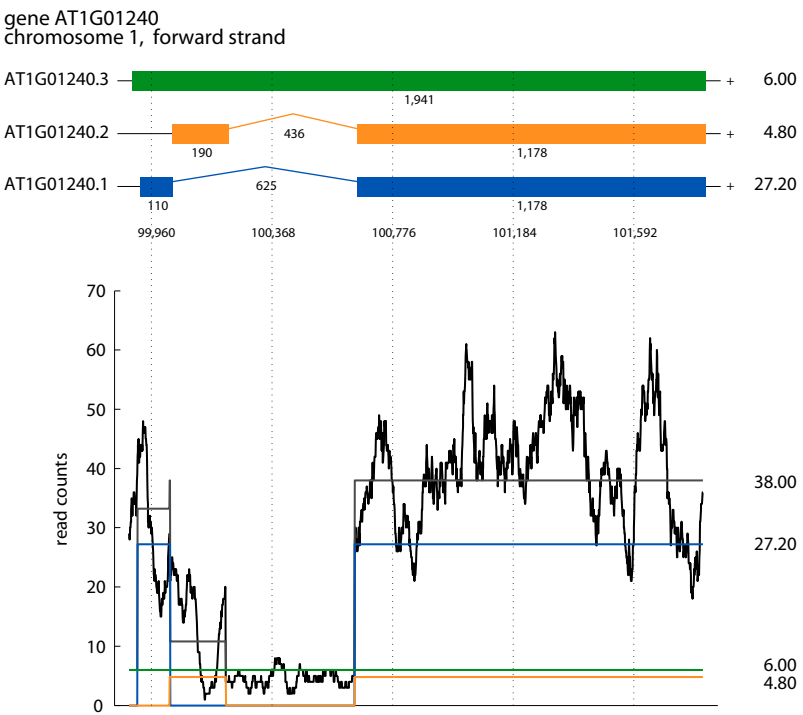
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$

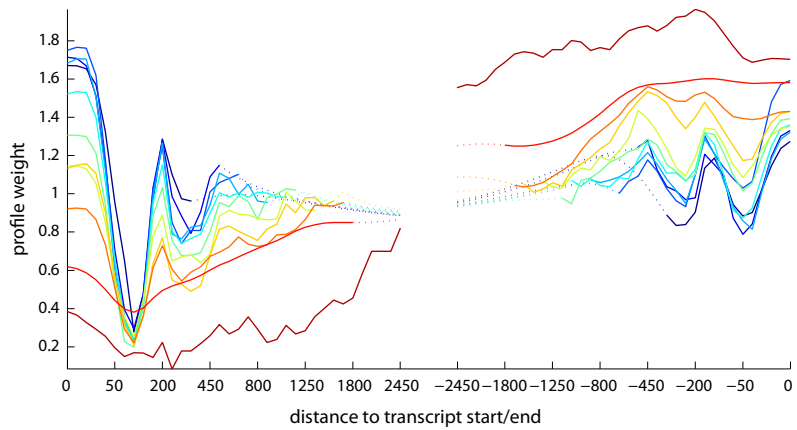



[Bohnert et al., 2009]

# 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: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$

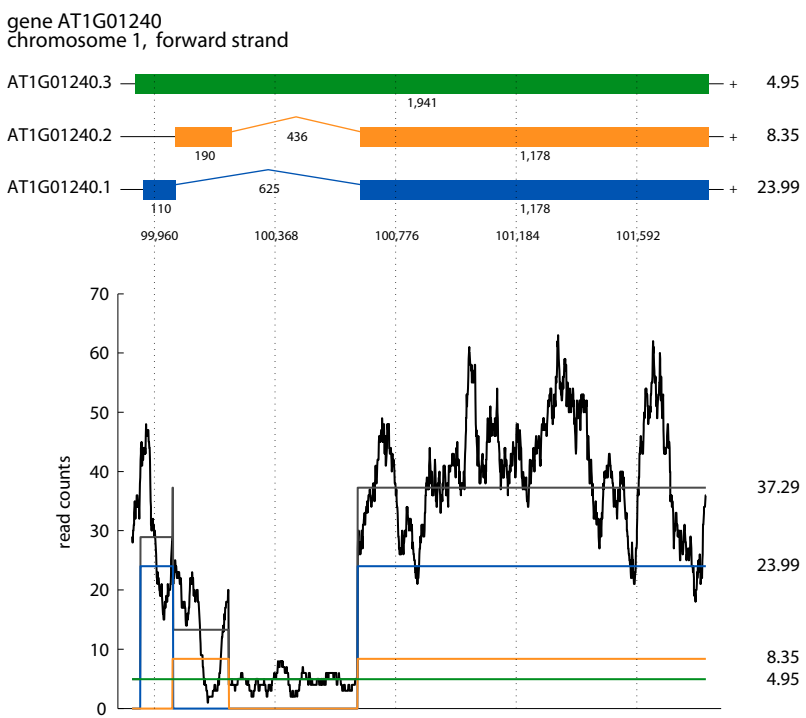


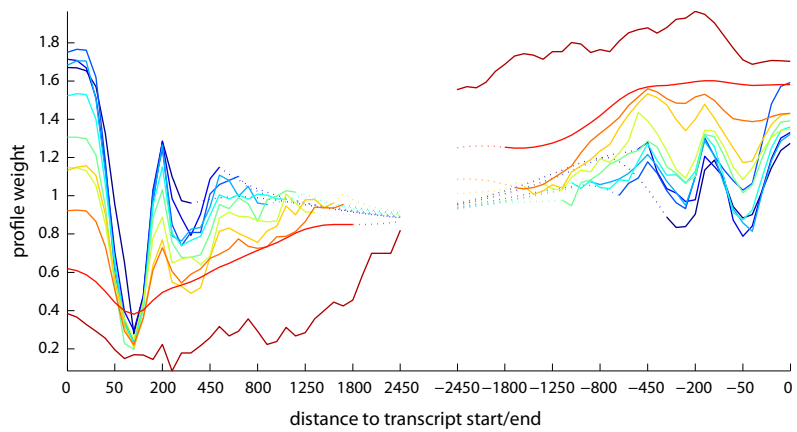




[Bohnert et al., 2009]

# 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.




[Bohnert et al., 2009]

# Evaluation I

Our method rQuant: Position-wise, with profiles

compared to

- Position-wise, without profiles
- Segment-wise, without profiles [Jiang and Wong, 2009]
- Segment-wise, with profiles [Sammeth, 2009a]

# Evaluation I

Our method rQuant: Position-wise, with profiles

compared to

- Position-wise, without profiles
- Segment-wise, without profiles [Jiang and Wong, 2009]
- Segment-wise, with profiles [Sammeth, 2009a]

Estimate transcript abundances

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

# Evaluation I

Our method rQuant: Position-wise, with profiles

compared to

- Position-wise, without profiles
- Segment-wise, without profiles [Jiang and Wong, 2009]
- Segment-wise, with profiles [Sammeth, 2009a]

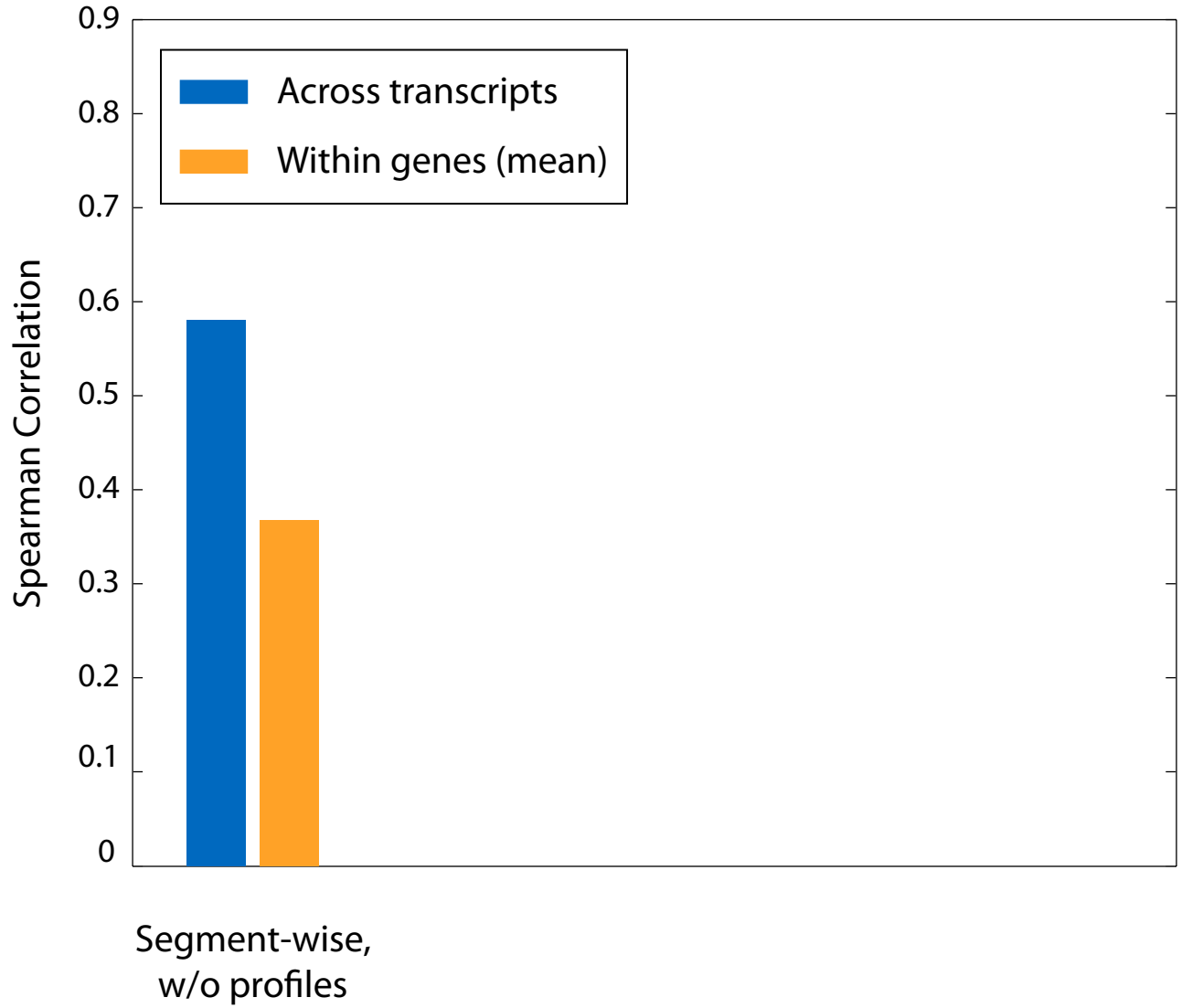
Estimate transcript abundances

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

Evaluation: Spearman correlation between

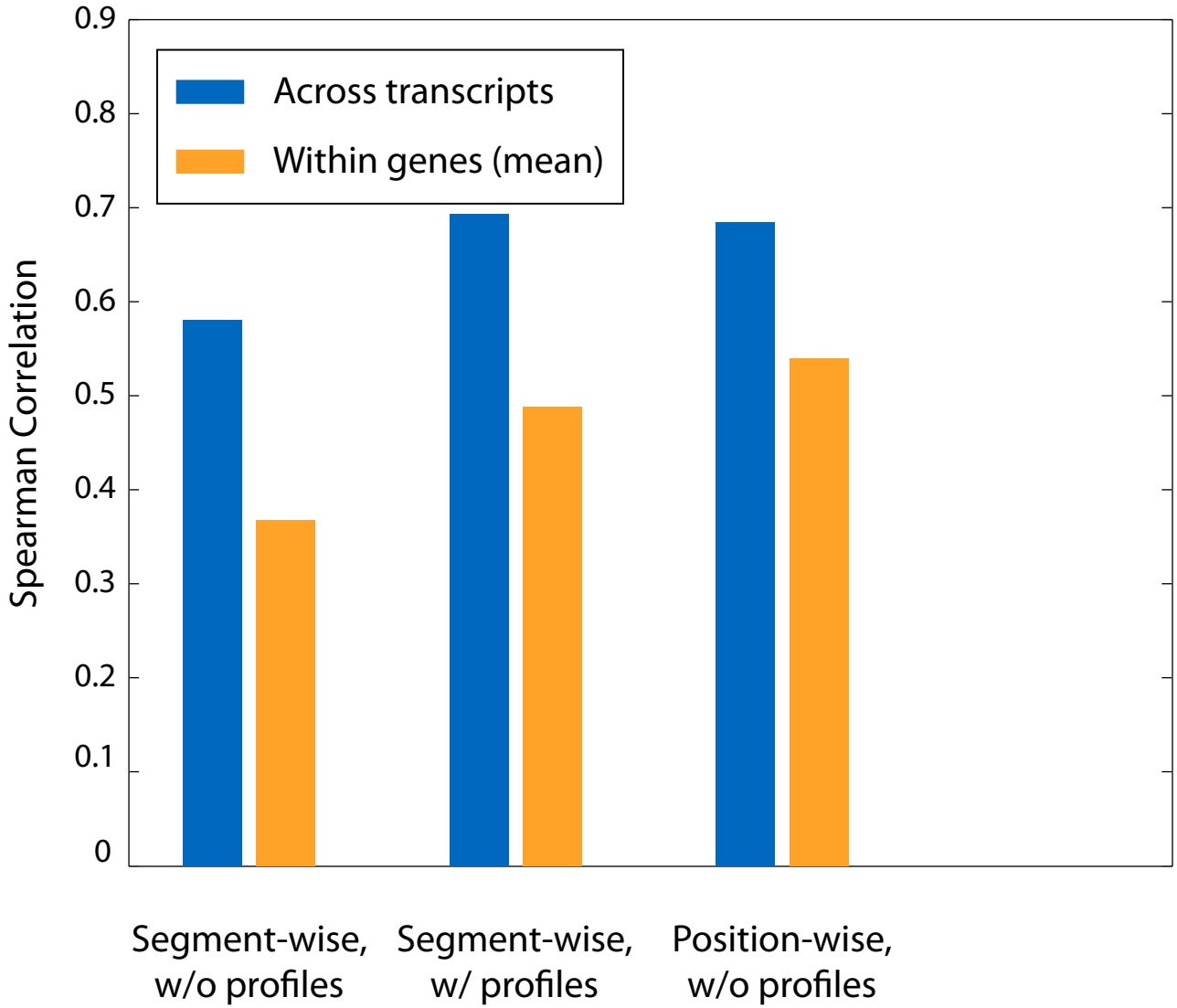
- Simulated RNA expression level and
- Predicted transcript weights

# Evaluation II



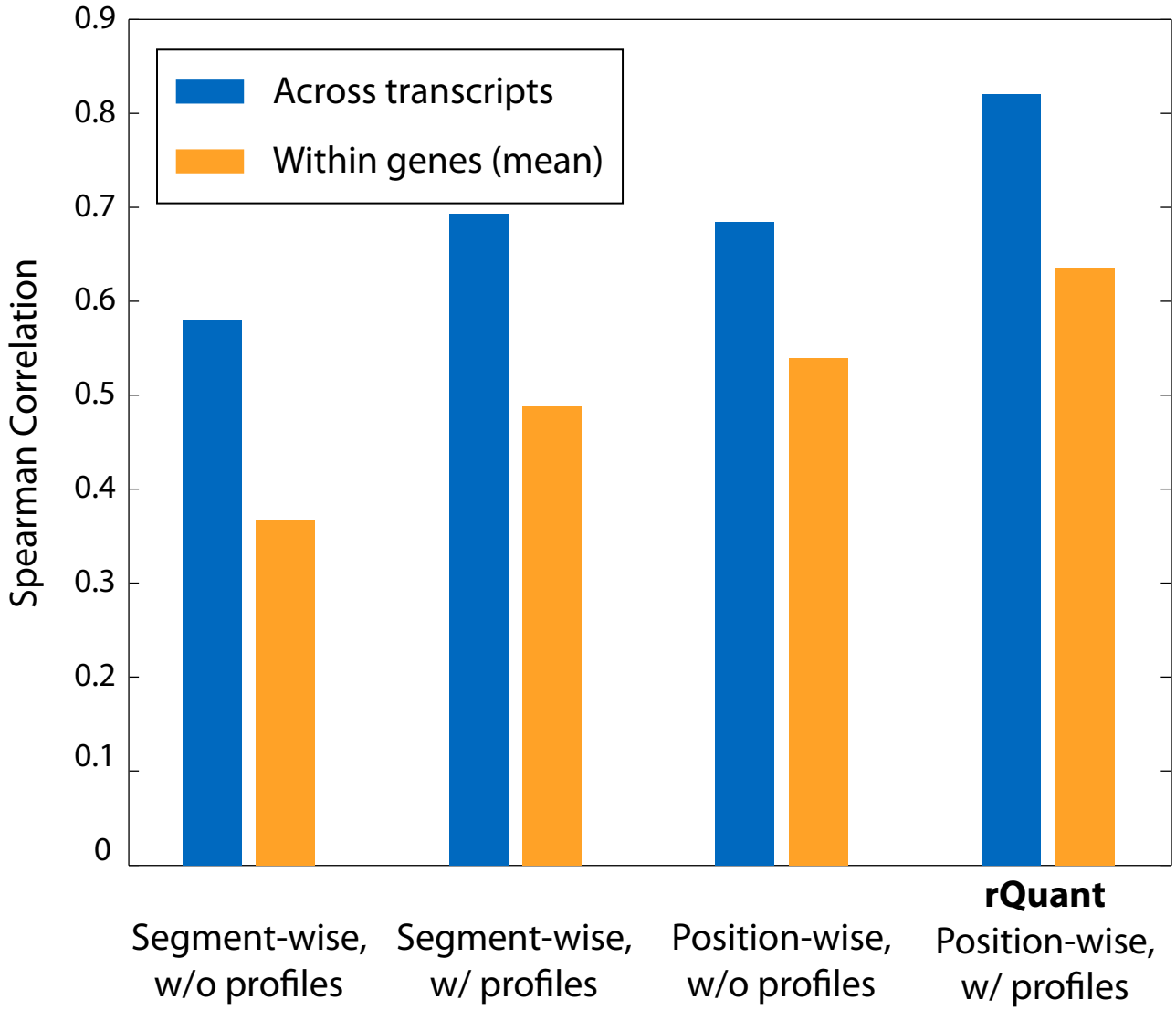

[Bohnert et al., 2009]

# Evaluation II




[Bohnert et al., 2009]

# Evaluation II




[Bohnert et al., 2009]

# Conclusions

- Biases in Tiling array and RNA-Seq data are severe

# Conclusions

- Biases in Tiling array and RNA-Seq data are severe
- Computational models help to deal with the biases and to increase the value of the data

# Conclusions

- Biases in Tiling array and RNA-Seq data are severe
- Computational models help to deal with the biases and to increase the value of the data
- Computational models of how the gene expression/processing works helps to more accurately deal with the measurements

# Conclusions

- Biases in Tiling array and RNA-Seq data are severe
- Computational models help to deal with the biases and to increase the value of the data
- Computational models of how the gene expression/processing works helps to more accurately deal with the measurements
- Proper learning-based techniques often perform considerably more accurate than *ad hoc* methods

# Conclusions

- Biases in Tiling array and RNA-Seq data are severe
- Computational models help to deal with the biases and to increase the value of the data
- Computational models of how the gene expression/processing works helps to more accurately deal with the measurements
- Proper learning-based techniques often perform considerably more accurate than *ad hoc* methods
- More information on `www.fml.mpg.de/raetsch/research/transcriptomics`
- The slides can be found on `www.fml.mpg.de/raetsch/lectures`

# Conclusions




- Biases in Tiling array and RNA-Seq data are severe
- Computational models help to deal with the biases and to increase the value of the data
- Computational models of how the gene expression/processing works helps to more accurately deal with the measurements
- Proper learning-based techniques often perform considerably more accurate than *ad hoc* methods
- More information on `www.fml.mpg.de/raetsch/research/transcriptomics`
- The slides can be found on `www.fml.mpg.de/raetsch/lectures`

Thank you for your attention!

# Acknowledgments


## Tiling Arrays

- **Georg Zeller** (FML & MPI)
- Johannes Eichner (FML)
- **Sascha Laubinger** (MPI)
- Stefan Henz (MPI)
- Detlef Weigel (MPI)

## Short Read Analysis

- **Fabio De Bona** (FML)
- **Regina Bohnert** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)
- Jun Cao (MPI)


## More Information

- `www.fml.mpg.de/raetsch/research/transcriptomics`
- Slides with references are available online:
  `www.fml.mpg.de/raetsch/lectures`

# 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. URL `http://www.fml.tuebingen.mpg.de/raetsch/lectures/RaetschGenomeInformatics08.pdf`. Oral presentation at the CSHL Genome Informatics Meeting, September 2008.

R. Bohnert, J. Behr, and G. Rätsch. Quantitative detection of alternative transcripts with rna-seq data. Oral presentation at the Short Read SIG at ISMB/ECCB 2009 in Stockholn, 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.

Lior David, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis, and Lars M Steinmetz. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA*, 103(14):5320–5, Apr 2006. doi: 10.1073/pnas.0601091103.

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

# References II


J. Eichner. Analysis of alternative transcripts in arabidopsis thaliana with whole genome arrays. Master's thesis, University of Tübingen, Sand 13, 72076 Tübingen, Germany, June 2008.

J. Eichner, G. Zeller, S. Laubinger, D. Weigel, and G. Rätsch. Analysis of alternative transcripts in arabidopsis thaliana with whole genome arrays. forthcoming, March 2009.

Yaniv Erlich, Partha P Mitra, Melissa delaBastide, W Richard McCombie, and Gregory J Hannon. Alta-cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods*, 5(8):679–682, Aug 2008. ISSN 1548-7105 (Electronic). doi: 10.1038/nmeth.1230.

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, February 2009. doi: 10.1093/bioinformatics/btp113.

S Laubinger, T Sachsenberg, G Zeller, W Busch, JU Lohmann, G Rätsch, and D Weigel. Dual roles of the nuclear cap-binding complex and serrate in pre-mrna splicing and microrna processing in arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 105(25):8795–8800, 2008a. ISSN 1091-6490 (Electronic). doi: 10.1073/pnas.0802493105.

Sascha Laubinger, Georg Zeller, Stefan R Henz, Timo Sachsenberg, Christian K Widmer, Naïra Naouar, Marnik Vuylsteke, Bernhard Schölkopf, Gunnar Rätsch, and Detlef Weigel. At-tax: a whole genome tiling array resource for developmental expression analysis and transcript identification in arabidopsis thaliana. *Genome Biol*, 9(7):R112, 2008b. doi: 10.1186/gb-2008-9-7-r112.

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*, Jun 2009. doi: 10.1093/bioinformatics/btp336.

# References III


Todd C Mockler, Simon Chan, Ambika Sundaresan, Huaming Chen, Steven E Jacobsen, and Joseph R Ecker. Applications of dna tiling arrays for whole-genome analysis. *Genomics*, 85 (1):1–15, Jan 2005. doi: 10.1016/j.ygeno.2004.10.005.

Ossowski. Next generation sequencing. Oral presentation at the PhD Symposium in Tübingen, Germany, November 2007.

S. Ossowski, K. Schneeberger, R. Clark, C. Lanz, N. Warthmann, and D. Weigel. Sequencing of natural strains of arabidopsis thaliana with short reads. *Genome Research*, 18(2024–2033), 2008.

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.

Thomas E Royce, Joel S Rozowsky, and Mark B Gerstein. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, 23(8): 988–97, Apr 2007. doi: 10.1093/bioinformatics/btm052. URL `http://bioinformatics.oxfordjournals.org/cgi/content/full/23/8/988`.

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.

# References IV


Manoj Pratim Samanta, Waraporn Tongprasit, Himanshu Sethi, Chen-Shan Chin, and Viktor Stolc. Global identification of noncoding rnas in saccharomyces cerevisiae by modulating an essential rna processing pathway. *Proc Natl Acad Sci USA*, 103(11):4192–7, Mar 2006. doi: 10.1073/pnas.0507669103.

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.

G. Schweikert, G. Zeller, A. Zien, J. Behr, C.S. Ong, P. Philips, A. Bohlen, R. Bohnert, F. De Bona, S. Sonnenburg, and G. Rätsch. mGene: Accurate computational gene finding with application to nematode genomes. under revision for Genome Research, March 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.

I. Sutskever. Arachne: A whole genome shotgun assembler. oral presentation, 2008.

Junshi Yazaki, Brian D Gregory, and Joseph R Ecker. Mapping the genome landscape using tiling array technology. *Current Opinion in Plant Biology*, 10(5):534–42, Oct 2007. doi: 10.1016/j.pbi.2007.07.006. URL `http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VS4-4PG2S31-1&_user=29041&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000003178&_version=1&_urlVersion=0&_userid=29041&md5=80ec1d3e091fd96a9f662289f6584c05`.

# References V


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, 2008a. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

G Zeller, S Henz, C Widmer, T Sachsenberg, G Rätsch, D Weigel, and S Laubinger. Stress-induced changes in the arabidopsis thaliana transcriptome analyzed using whole genome tiling arrays. *Plant J*, Feb 2009. doi: 10.1111/j.1365-313X.2009.03835.x.

Georg Zeller, Stefan R Henz, Sascha Laubinger, Detlef Weigel, and Gunnar Rätsch. Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 527–38, Jan 2008b.

D.R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18:828–829, 2008.

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.