Methods for Transcriptome Analysis with Tiling Arrays and mRNA-Seq

Gunnar Rätsch

Friedrich Miescher Laboratory
Max Planck Society
Tübingen, Germany

Talk at the University of Toronto
July 17, 2008



Friedrich Miescher Laboratory
of the Max Planck Society



Research Topics

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*
4. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*

Research Topics

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*
4. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*

Research Topics

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*
4. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*

Research Topics

1. Machine learning methods
   $\Rightarrow$ *Develop fast, accurate and interpretable learning methods*
2. Genome annotation
   $\Rightarrow$ *Predict features encoded on DNA*
3. Biological networks
   $\Rightarrow$ *Understand interactions between gene products*
4. Analysis of polymorphisms
   $\Rightarrow$ *Discover polymorphisms and associate them with phenotypes*

Machine Learning Methods

⇒ Develop fast, accurate and interpretable learning methods

1. **Large scale sequence classification**

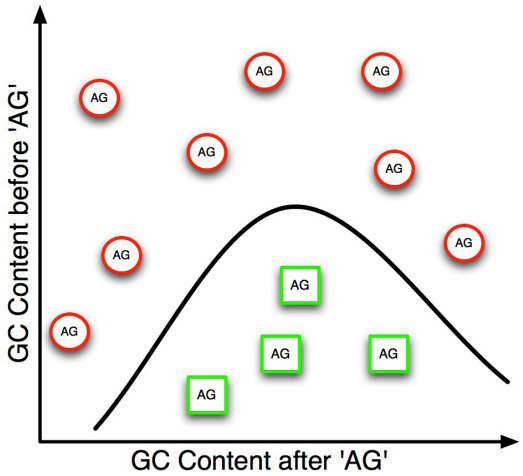
   with Sonnenburg (Fraunhofer, Berlin) & Schölkopf (MPI Biol. Cybernetics)

2. Analysis and explanation of learning result

   with Sonnenburg (Fraunhofer, Berlin)

3. Sequence segmentation

   with Altun (MPI Biol. Cybernetics)



[e.g. Sonnenburg et al., 2007, Rätsch et al., 2006, Rätsch and Sonnenburg, 2007]

Machine Learning Methods

⇒ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification

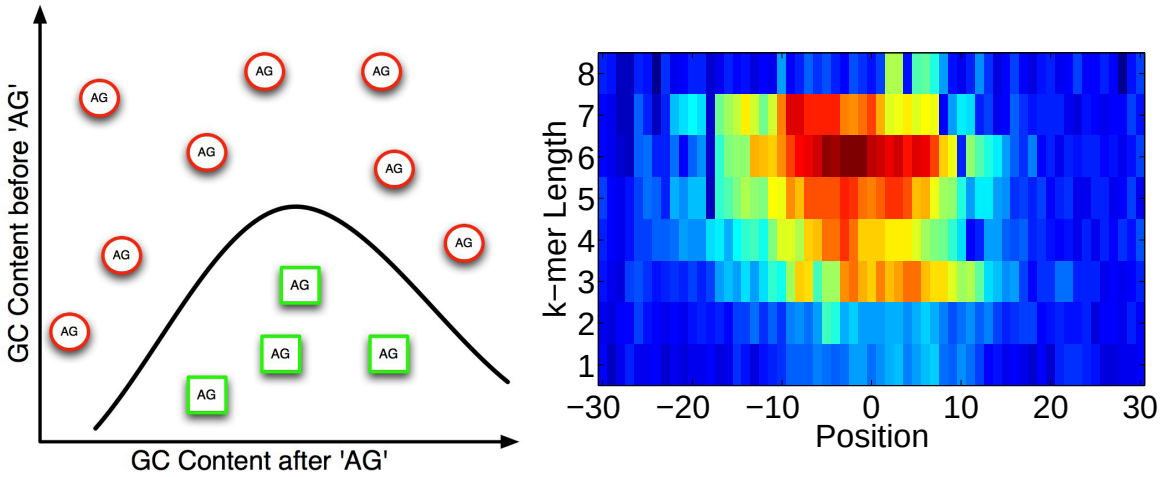
   with Sonnenburg (Fraunhofer, Berlin) & Schölkopf (MPI Biol. Cybernetics)

2. Analysis and explanation of learning result

   with Sonnenburg (Fraunhofer, Berlin)

3. Sequence segmentation

   with Altun (MPI Biol. Cybernetics)



[e.g. Sonnenburg et al., 2007, Rätsch et al., 2006, Rätsch and Sonnenburg, 2007]

Machine Learning Methods

⇒ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification

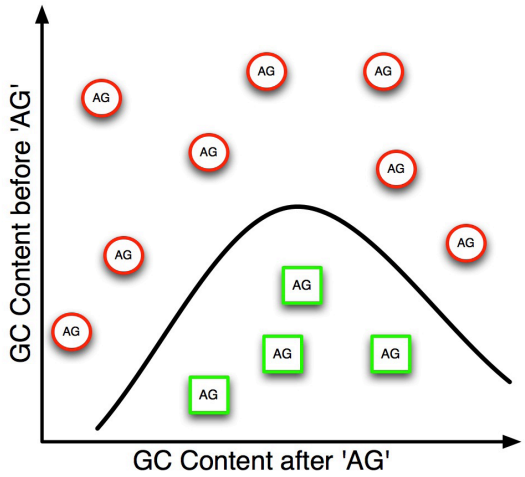
   with Sonnenburg (Fraunhofer, Berlin) & Schölkopf (MPI Biol. Cybernetics)

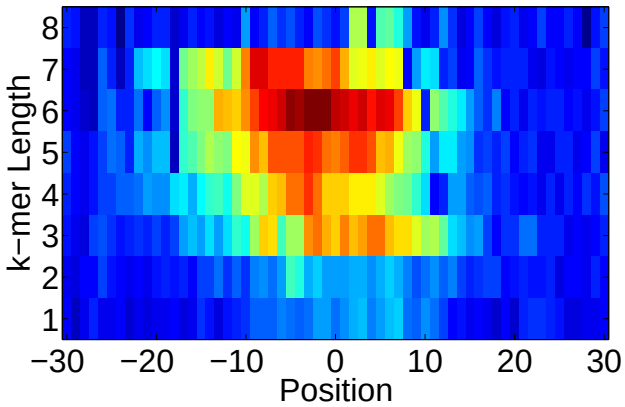
2. Analysis and explanation of learning result

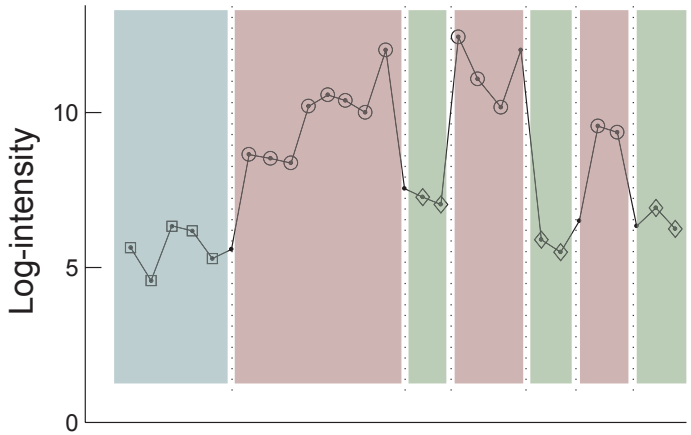
   with Sonnenburg (Fraunhofer, Berlin)

3. Sequence segmentation

   with Altun (MPI Biol. Cybernetics)







[e.g. Sonnenburg et al., 2007, Rätsch et al., 2006, Rätsch and Sonnenburg, 2007]

Genome annotation

⇒ Predict features encoded on DNA

1. *Ab initio* gene finding and prediction of alternative splicing
   1. *C. remanei/briggsae/japonica/brenneri* with Stein (CSHL)
   2. *P. pacificus* with Sommer (MPI Developmental Biology)
   3. Many fungal genomes with Güldener (MIPS)
   4. *V. carteri* with Hallmann (U. Bielefeld)
   5. Future: *A. lyrata*, *D. melanogaster*, *D. rerio*, human, ...
2. Transcriptome tiling arrays
   with Weigel (MPI Developmental Biology)
3. Alignment methods for short read sequencing
   with Weigel (MPI Developmental Biology)
4. Prediction of RNA subcellular localization and secondary structure

[e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]

Genome annotation

⇒ Predict features encoded on DNA



1. *Ab initio* gene finding and prediction of alternative splicing
   1. *C. remanei/briggsae/japonica/brenneri* with Stein (CSHL)
   2. *P. pacificus* with Sommer (MPI Developmental Biology)
   3. Many fungal genomes with Güldener (MIPS)
   4. *V. carteri* with Hallmann (U. Bielefeld)
   5. Future: *A. lyrata*, *D. melanogaster*, *D. rerio*, human, ...
2. Transcriptome tiling arrays

   with Weigel (MPI Developmental Biology)
3. Alignment methods for short read sequencing

   with Weigel (MPI Developmental Biology)
4. Prediction of RNA subcellular localization and secondary structure

[e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]

Genome annotation

⇒ Predict features encoded on DNA



1. *Ab initio* gene finding and prediction of alternative splicing
   1. *C. remanei/briggsae/japonica/brenneri* with Stein (CSHL)
   2. *P. pacificus* with Sommer (MPI Developmental Biology)
   3. Many fungal genomes with Güldener (MIPS)
   4. *V. carteri* with Hallmann (U. Bielefeld)
   5. Future: *A. lyrata*, *D. melanogaster*, *D. rerio*, human, ...
2. Transcriptome tiling arrays

   with Weigel (MPI Developmental Biology)
3. Alignment methods for short read sequencing

   with Weigel (MPI Developmental Biology)
4. Prediction of RNA subcellular localization and secondary structure

[e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]

Genome annotation

⇒ Predict features encoded on DNA



1. *Ab initio* gene finding and prediction of alternative splicing
   1. *C. remanei/briggsae/japonica/brenneri* with Stein (CSHL)
   2. *P. pacificus* with Sommer (MPI Developmental Biology)
   3. Many fungal genomes with Güldener (MIPS)
   4. *V. carteri* with Hallmann (U. Bielefeld)
   5. Future: *A. lyrata*, *D. melanogaster*, *D. rerio*, human, ...
2. Transcriptome tiling arrays

   with Weigel (MPI Developmental Biology)
3. Alignment methods for short read sequencing

   with Weigel (MPI Developmental Biology)
4. Prediction of RNA subcellular localization and secondary structure

[e.g. Rätsch et al., 2007, Zeller et al., 2008b, De Bona et al., 2008]

Biological networks

⇒ Understand interactions between gene products

1. Identification of Transcription factor targets

   with Lohmann (MPI Developmental Biology)

2. Network motif discovery

   with Tsuda (MPI Biol. Cybernetics) and Dittman (MIPS)

3. Future: Quantitative modeling of networks

[e.g. Georgii et al., 2008, Schultheiss et al., 2008]

Biological networks

⇒ Understand interactions between gene products

1. Identification of Transcription factor targets

   with Lohmann (MPI Developmental Biology)

2. Network motif discovery

   with Tsuda (MPI Biol. Cybernetics) and Dittman (MIPS)

3. Future: Quantitative modeling of networks

[e.g. Georgii et al., 2008, Schultheiss et al., 2008]

Biological networks

⇒ Understand interactions between gene products

1. Identification of Transcription factor targets

   with Lohmann (MPI Developmental Biology)

2. Network motif discovery

   with Tsuda (MPI Biol. Cybernetics) and Dittman (MIPS)

3. Future: Quantitative modeling of networks

[e.g. Georgii et al., 2008, Schultheiss et al., 2008]

Analysis of Polymorphisms

⇒ Predict polymorphisms and associate them with phenotypes



1. Array-based resequencing for polymorphism discovery
   1. *A. thaliana* with Weigel & Schölkopf (MPI Biol. Cybernetics)
   2. *O. sativa* with Rice consortium & Weigel (MPI Devel. Biology)
   3. *M. musculus* with Eskin (UCLA)
2. Future: Genome-wide association studies/environmental effects
   1. *A. thaliana* with Weigel (MPI Developmental Biology)
   2. Human diseases with Lawrence (U. Manchester) and Tsuda (MPI Biol. Cybernetics)

[e.g. Clark et al., 2007, Zeller et al., 2008a]

Analysis of Polymorphisms

⇒ Predict polymorphisms and associate them with phenotypes



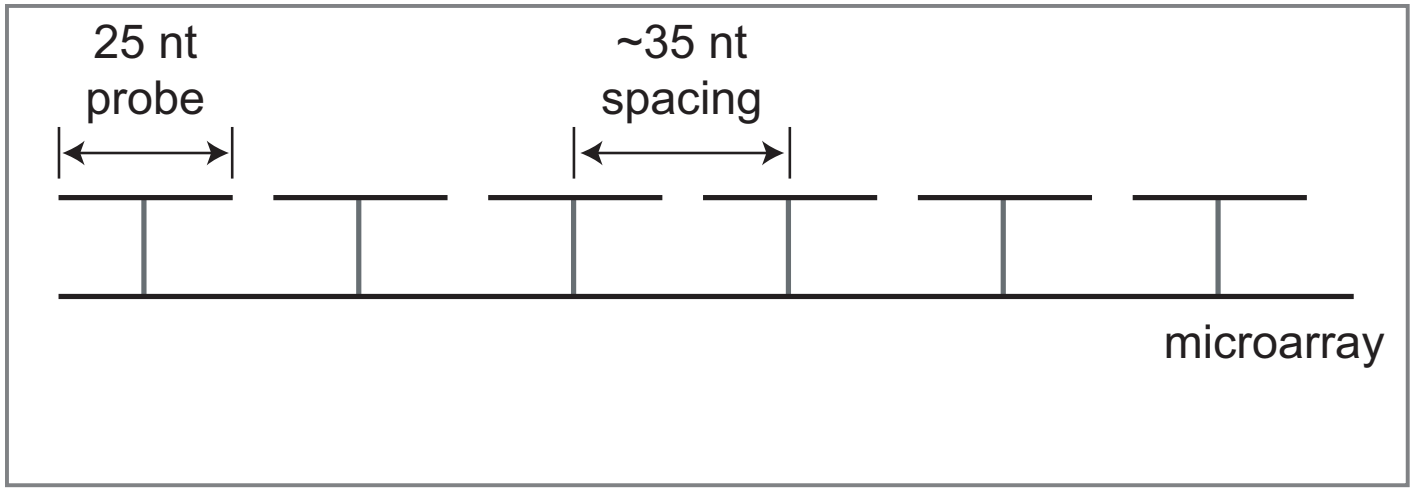
1. Array-based resequencing for polymorphism discovery
   1. *A. thaliana* with Weigel & Schölkopf (MPI Biol. Cybernetics)
   2. *O. sativa* with Rice consortium & Weigel (MPI Devel. Biology)
   3. *M. musculus* with Eskin (UCLA)
2. Future: Genome-wide association studies/environmental effects
   1. *A. thaliana* with Weigel (MPI Developmental Biology)
   2. Human diseases with Lawrence (U. Manchester) and Tsuda (MPI Biol. Cybernetics)

[e.g. Clark et al., 2007, Zeller et al., 2008a]

Talk Overview

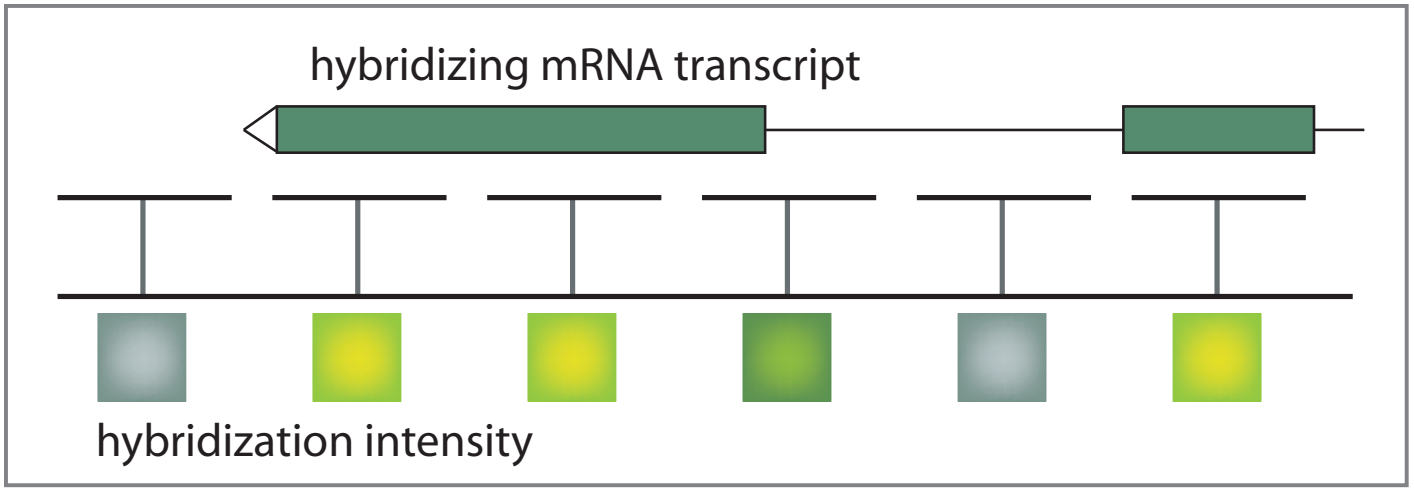
1. Transcriptome analysis with tiling arrays (50%)
   $\Rightarrow$ Identification of transcribed regions & alternative splicing
2. Spliced Alignments of Short Reads (40%)
   $\Rightarrow$ Accurate alignments using side information
3. Gene Finding with Tiling Arrays & mRNA-seq (10%)
   $\Rightarrow$ Transcriptome measurements improve gene predictions

Tiling Arrays for Transcriptome Analysis



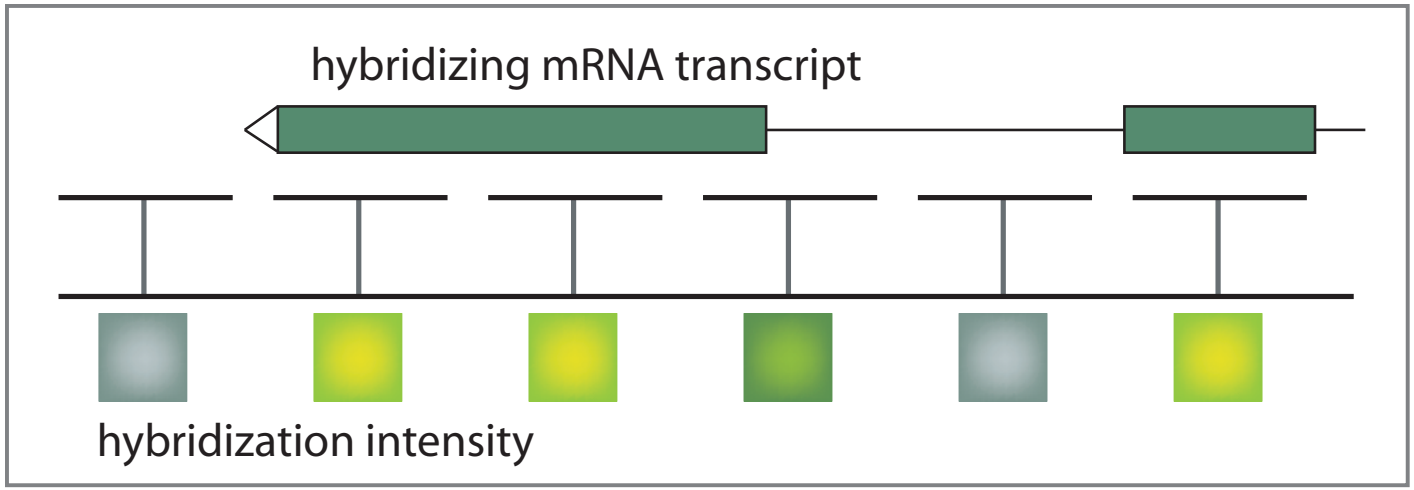
- Whole-genome quantitative measurements
- Cost-effective
  $\Rightarrow$ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  $\Rightarrow$ Do not rely on annotations or known cDNAs

Tiling Arrays for Transcriptome Analysis



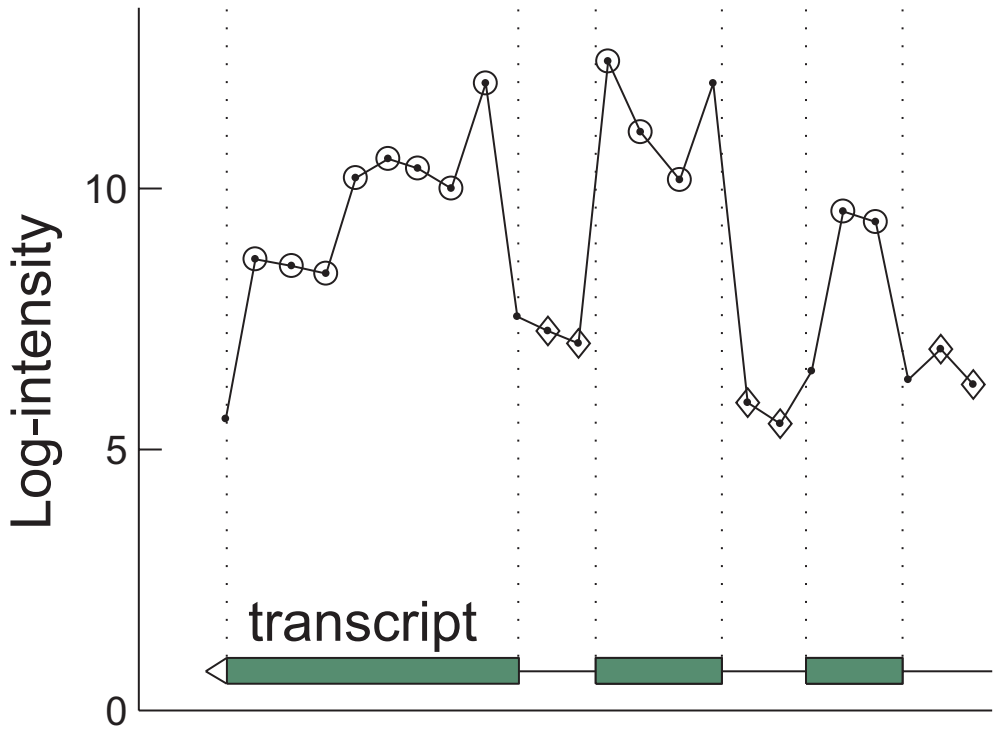
- Whole-genome quantitative measurements
- Cost-effective
  $\Rightarrow$ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  $\Rightarrow$ Do not rely on annotations or known cDNAs

Tiling Arrays for Transcriptome Analysis



- Whole-genome quantitative measurements
- Cost-effective
  $\Rightarrow$ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  $\Rightarrow$ Do not rely on annotations or known cDNAs

Intensities are Noisy Measurements

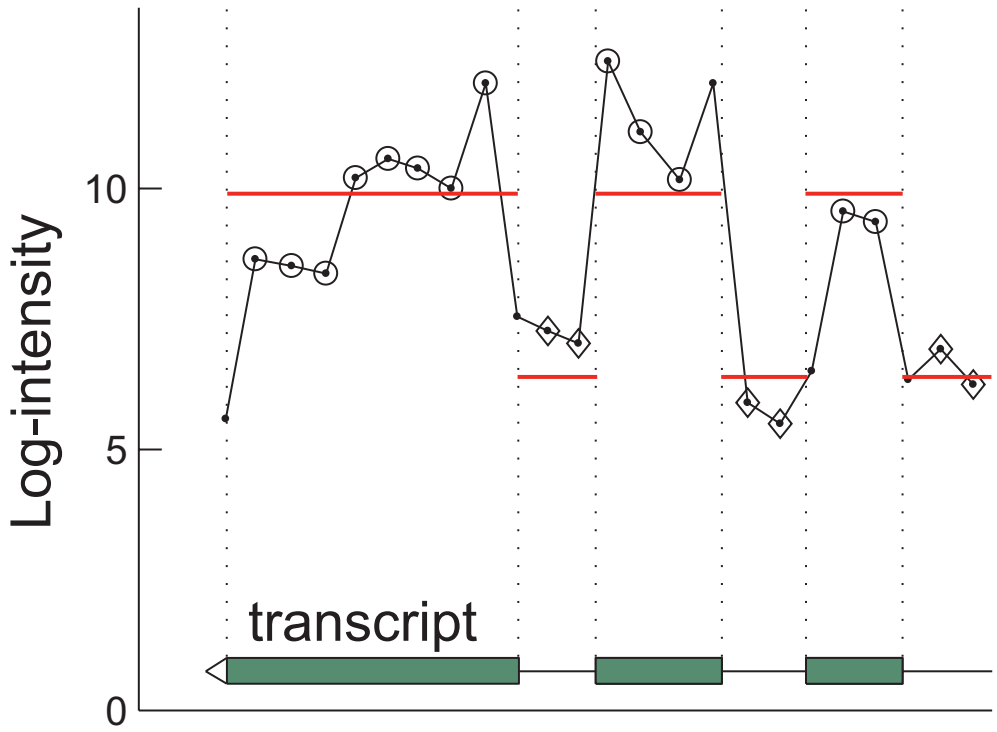


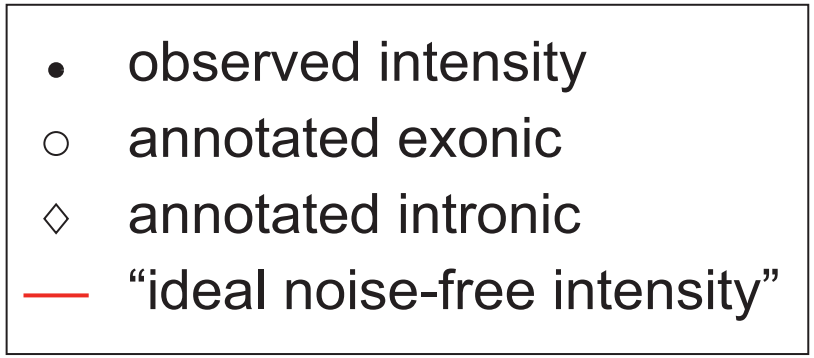
- observed intensity
- annotated exonic
- annotated intronic

Systematic bias induced by probe sequence effects

$\Rightarrow$ model effect for normalization

Intensities are Noisy Measurements

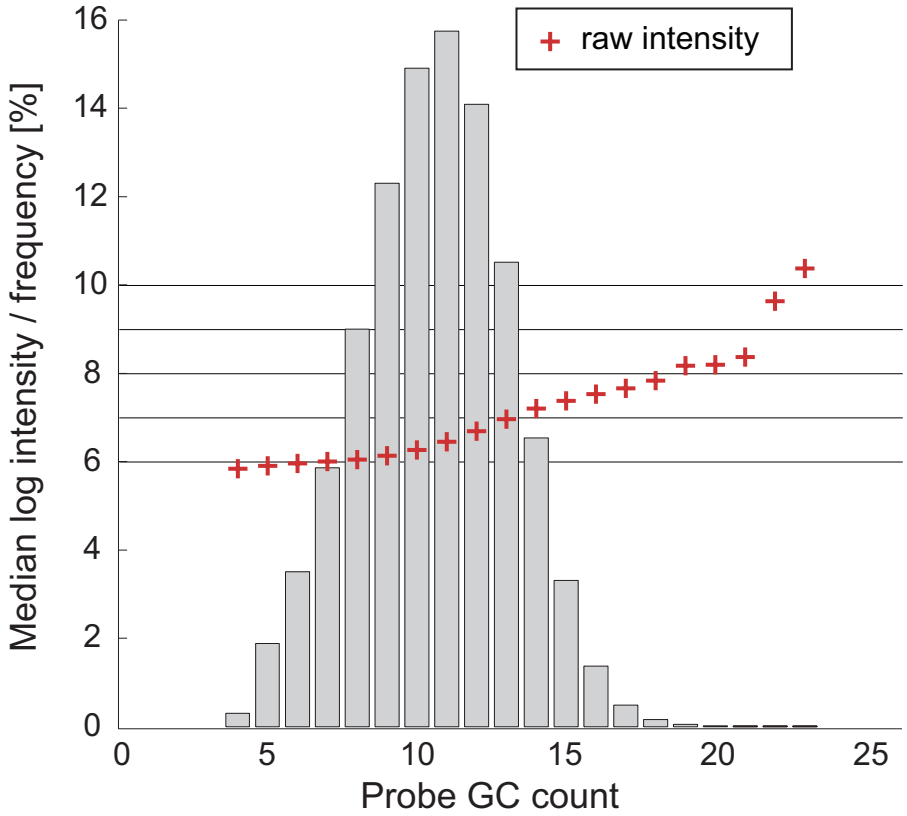




Systematic bias induced by probe sequence effects

$\Rightarrow$ model effect for normalization

Intensity Depends on Probe Sequence



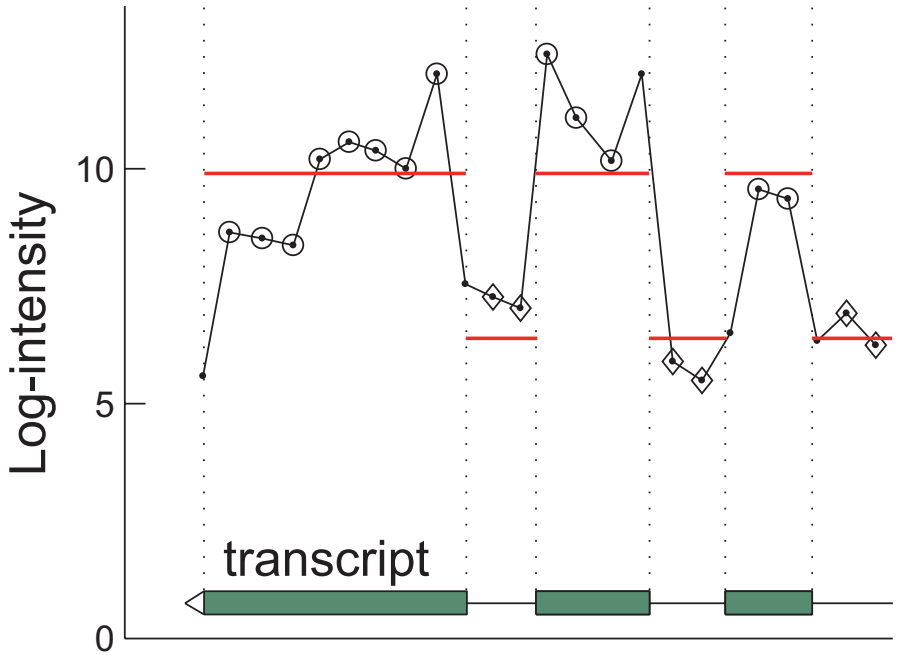
Results for the hybridization of polyadenylated RNA root tissue samples from *Arabidopsis thaliana*.

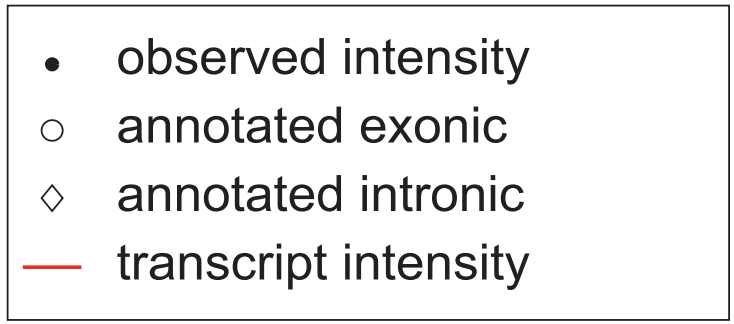
Previously proposed: Sequence Quantile Normalization (SQN)
[Royce et al., 2007]

[Zeller et al., 2008b]

Transcript Normalization

- Assume constant transcript intensities $\overline{y}_i$ (median estimate)
- Learn intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Model effect depending on probe sequence $\mathbf{x}_i$ and y_i:
  $f(\mathbf{x}_i, y_i) \approx \delta_i$ using quantilized linear regression

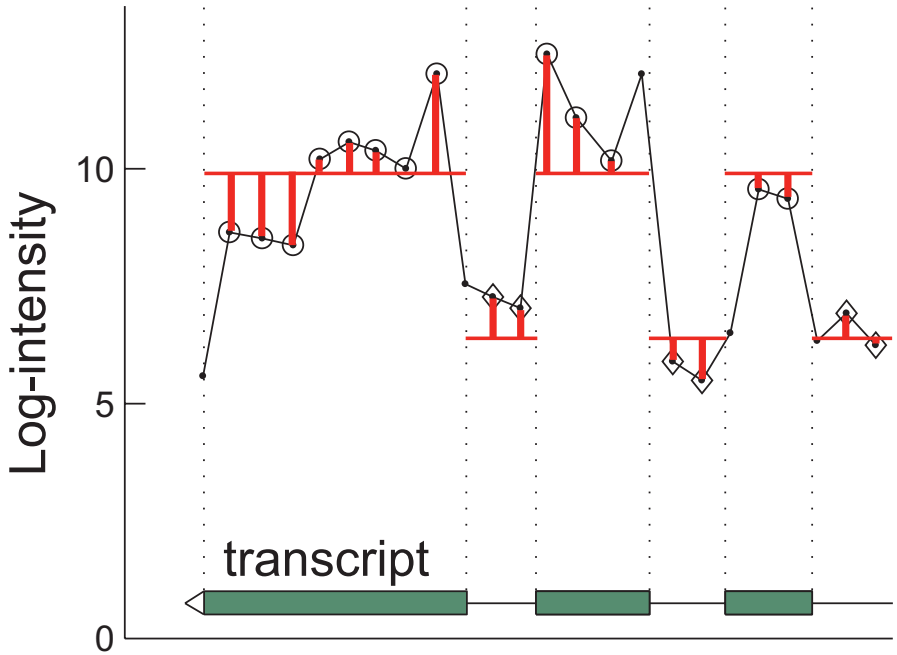




[Zeller et al., 2008b]

Transcript Normalization

- Assume constant transcript intensities $\overline{y}_i$ (median estimate)
- Learn intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Model effect depending on probe sequence $\mathbf{x}_i$ and y_i:
  $f(\mathbf{x}_i, y_i) \approx \delta_i$ using quantilized linear regression

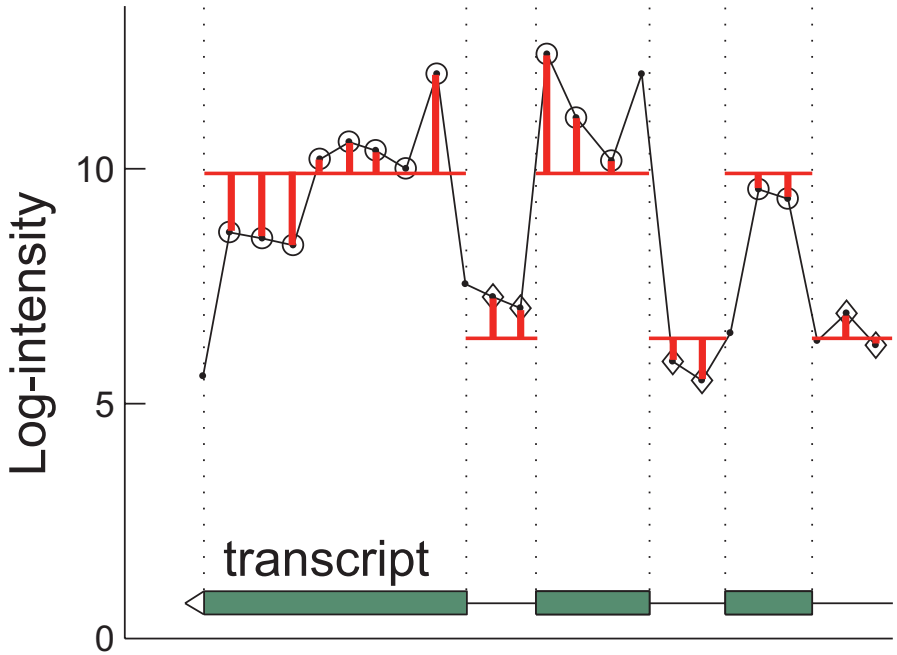


observed intensity
annotated exonic
annotated intronic
transcript intensity
fold difference δ between observed and transcript intensity

[Zeller et al., 2008b]

Transcript Normalization

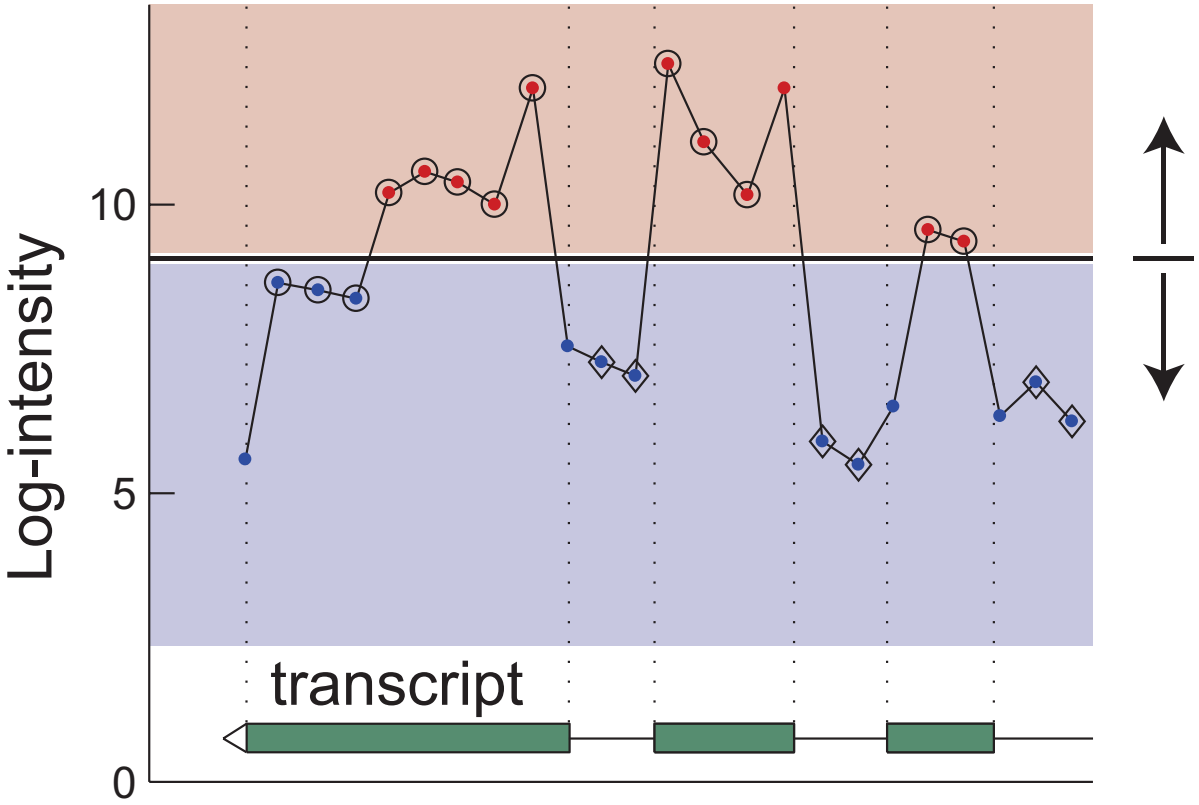
- Assume constant transcript intensities $\overline{y}_i$ (median estimate)
- Learn intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Model effect depending on probe sequence $\mathbf{x}_i$ and y_i:
  $f(\mathbf{x}_i, y_i) \approx \delta_i$ using quantilized linear regression



observed intensity
annotated exonic
annotated intronic
transcript intensity
fold difference δ between observed and transcript intensity

[Zeller et al., 2008b]

Exon/Background Probe Separation

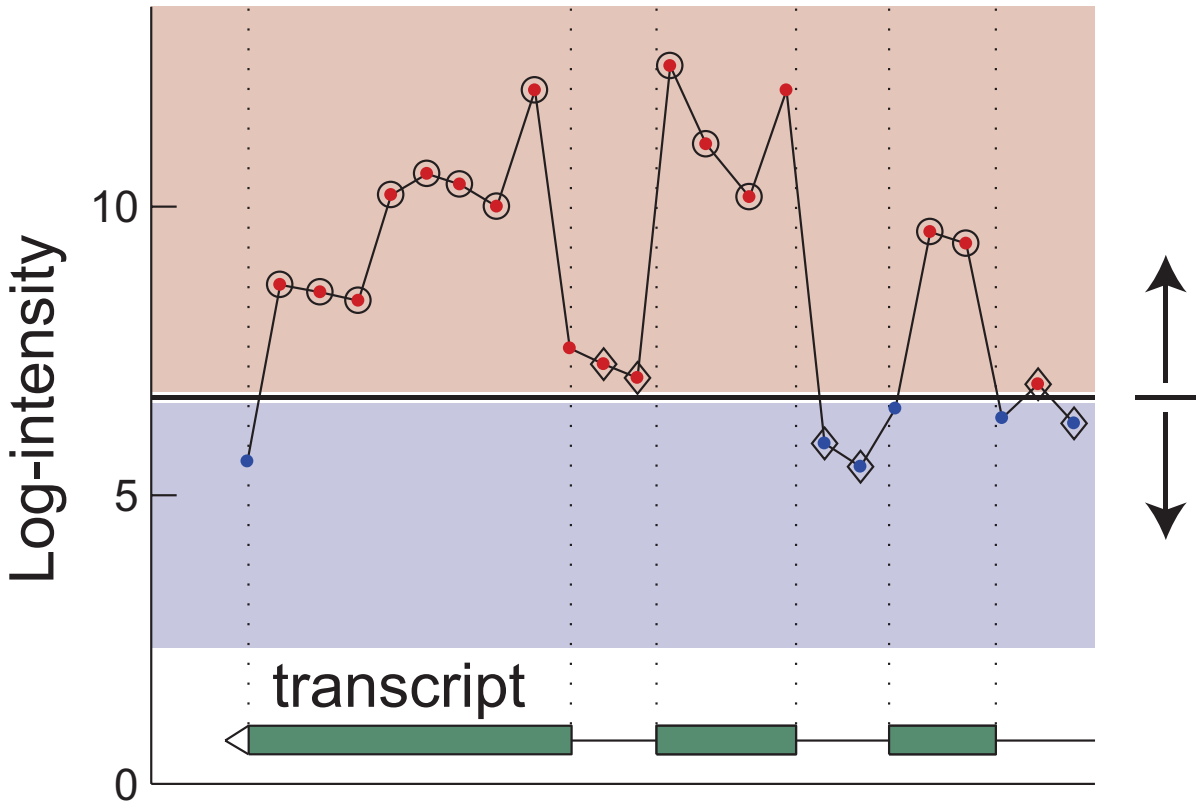


Global thresholding of probe intensities

$\Rightarrow$ bi-partition into exonic and intronic/intergenic probes

[Zeller et al., 2008b, Eichner et al., 2008]

Exon/Background Probe Separation



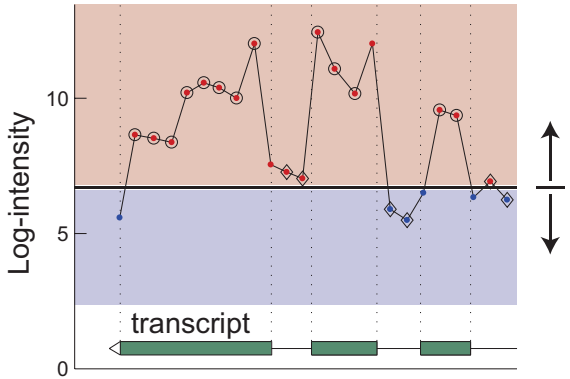
Global thresholding of probe intensities

$\Rightarrow$ bi-partition into exonic and intronic/intergenic probes

[Zeller et al., 2008b, Eichner et al., 2008]

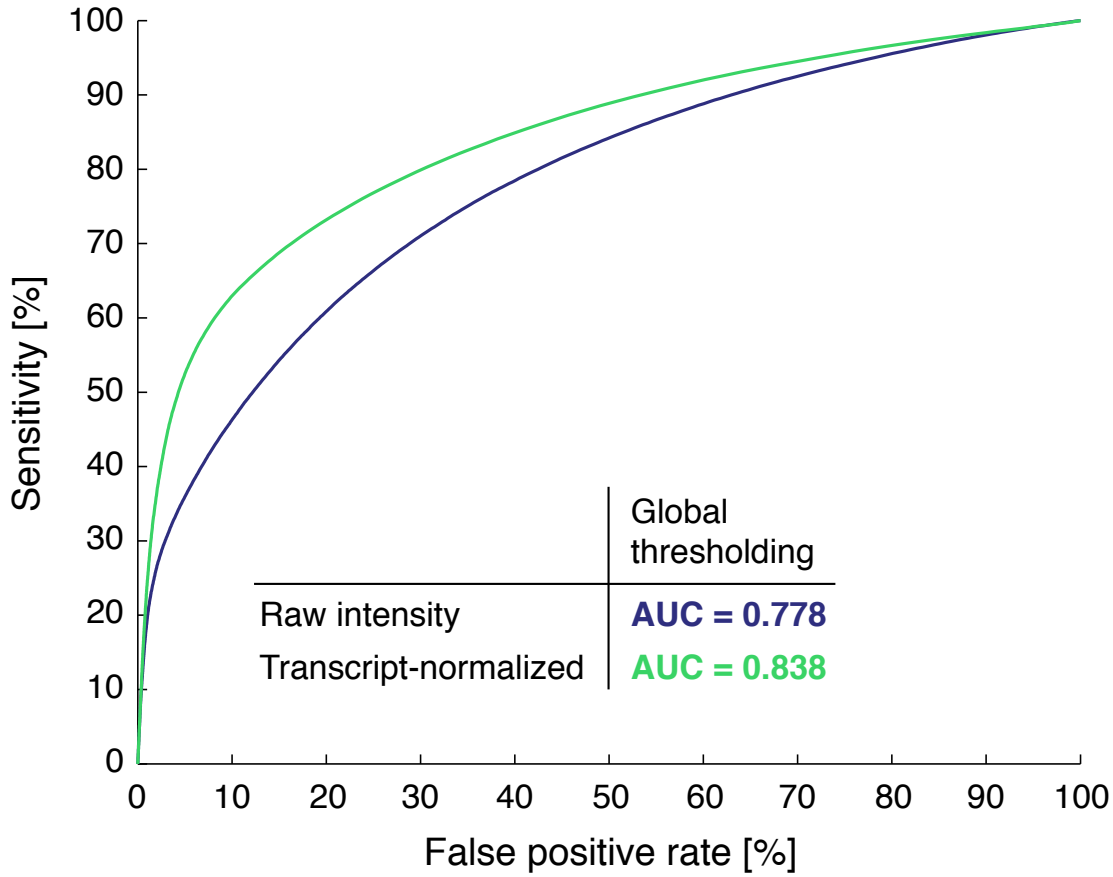
Exon/Background Probe Separation

- Global thresholding



- Support Vector Machines (SVMs) to discriminate exons from introns

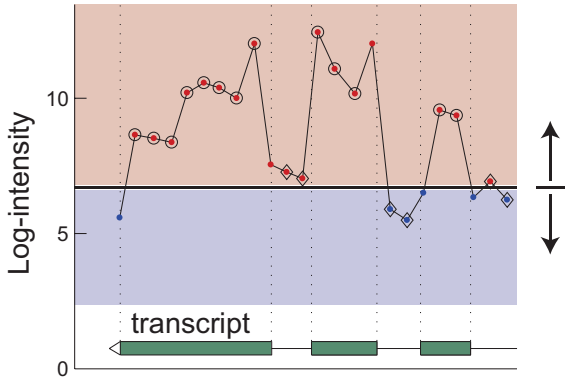
Segments typically consist of *several* probes
⇒ drastically improved separation



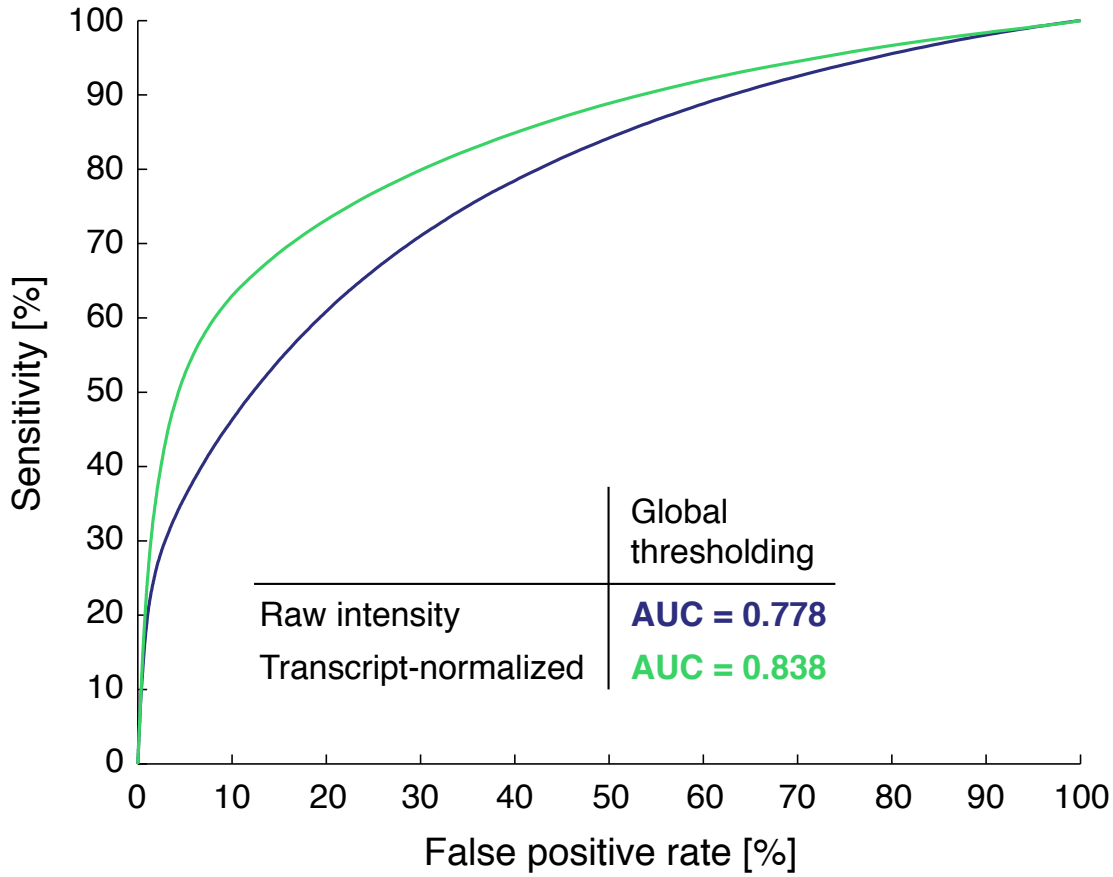
[Zeller et al., 2008b, Eichner et al., 2008]

Exon/Background Probe Separation

- Global thresholding



- Support Vector Machines (SVMs) to discriminate exons from introns



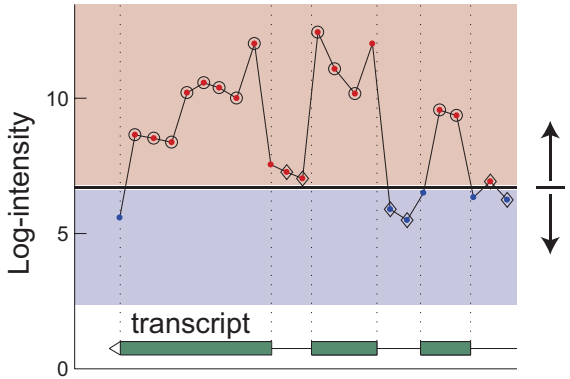
Segments typically consist of *several* probes

$\Rightarrow$ drastically improved separation

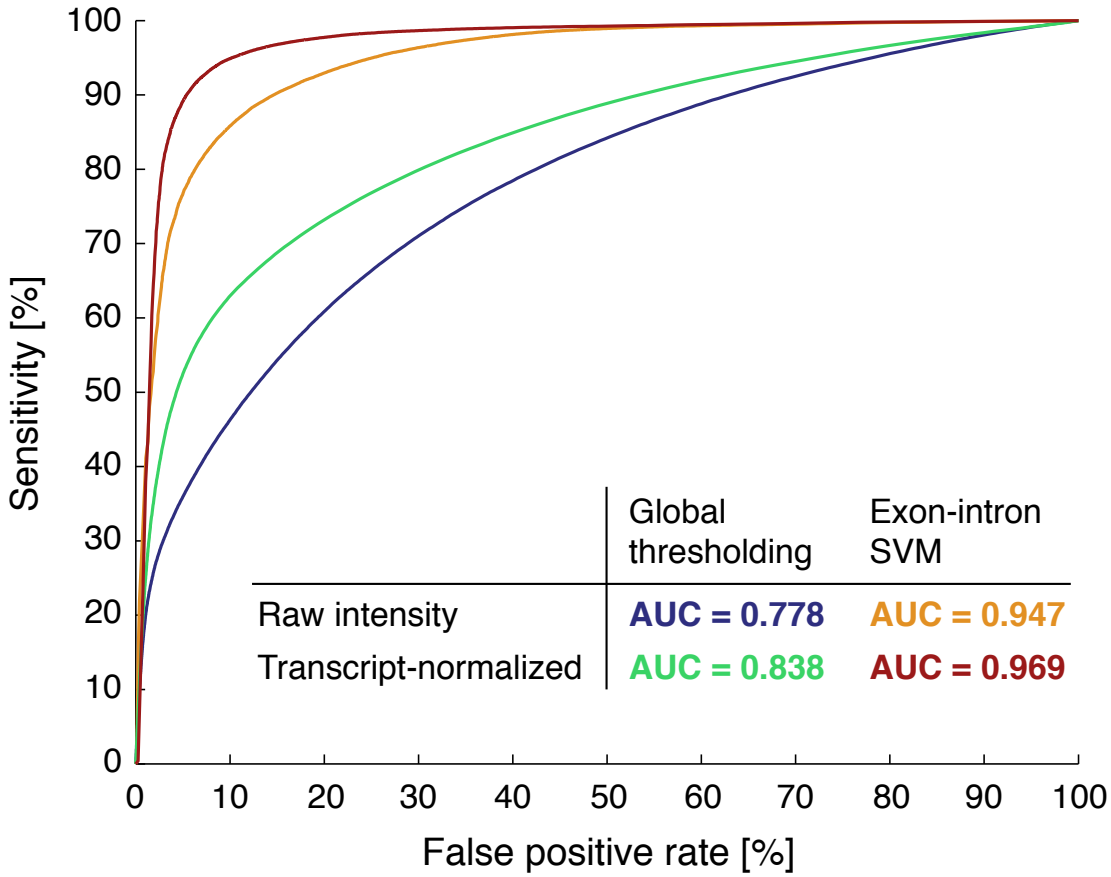
[Zeller et al., 2008b, Eichner et al., 2008]

Exon/Background Probe Separation

- Global thresholding



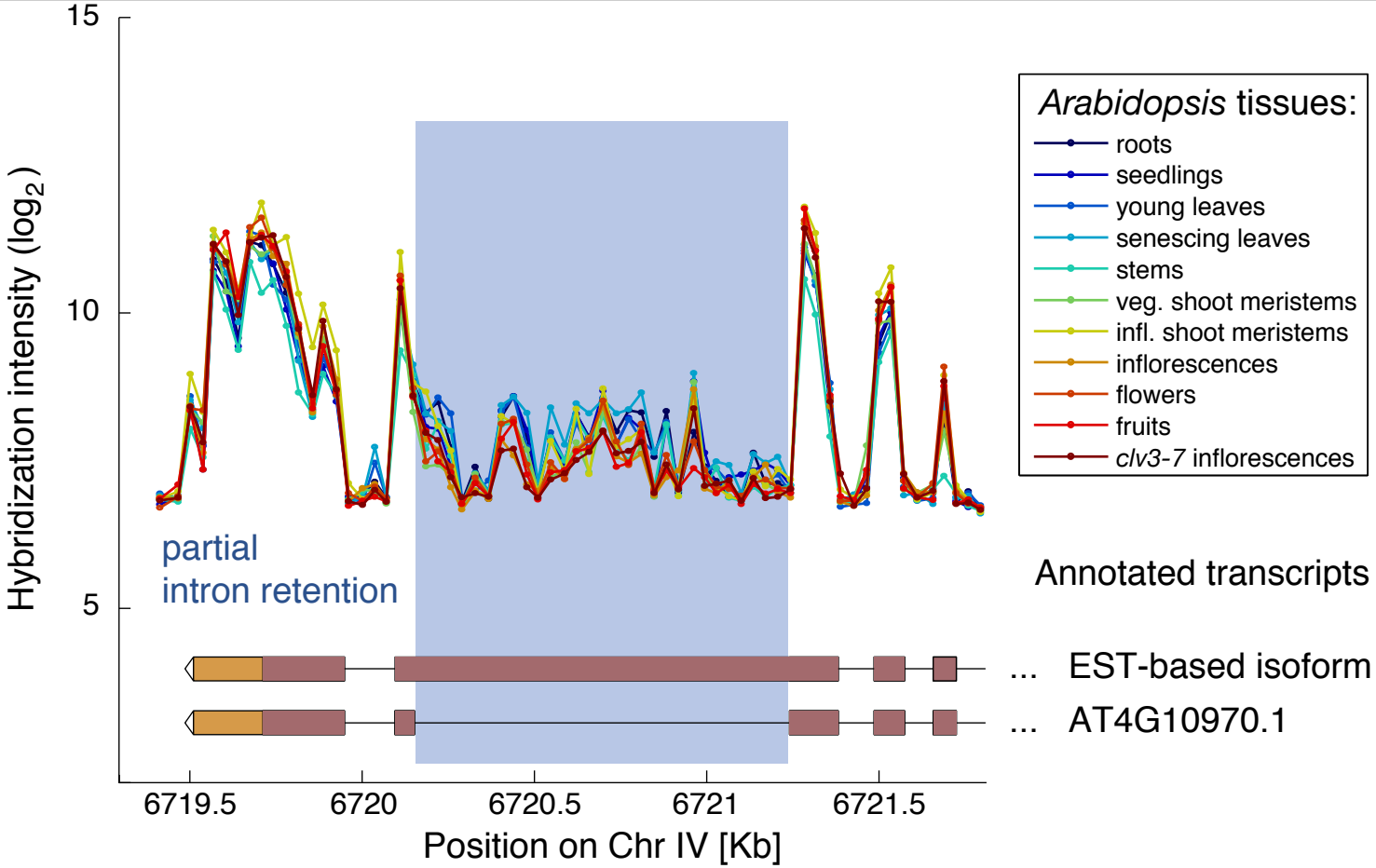
- Support Vector Machines (SVMs) to discriminate exons from introns



Segments typically consist of *several* probes
$\Rightarrow$ drastically improved separation

[Zeller et al., 2008b, Eichner et al., 2008]

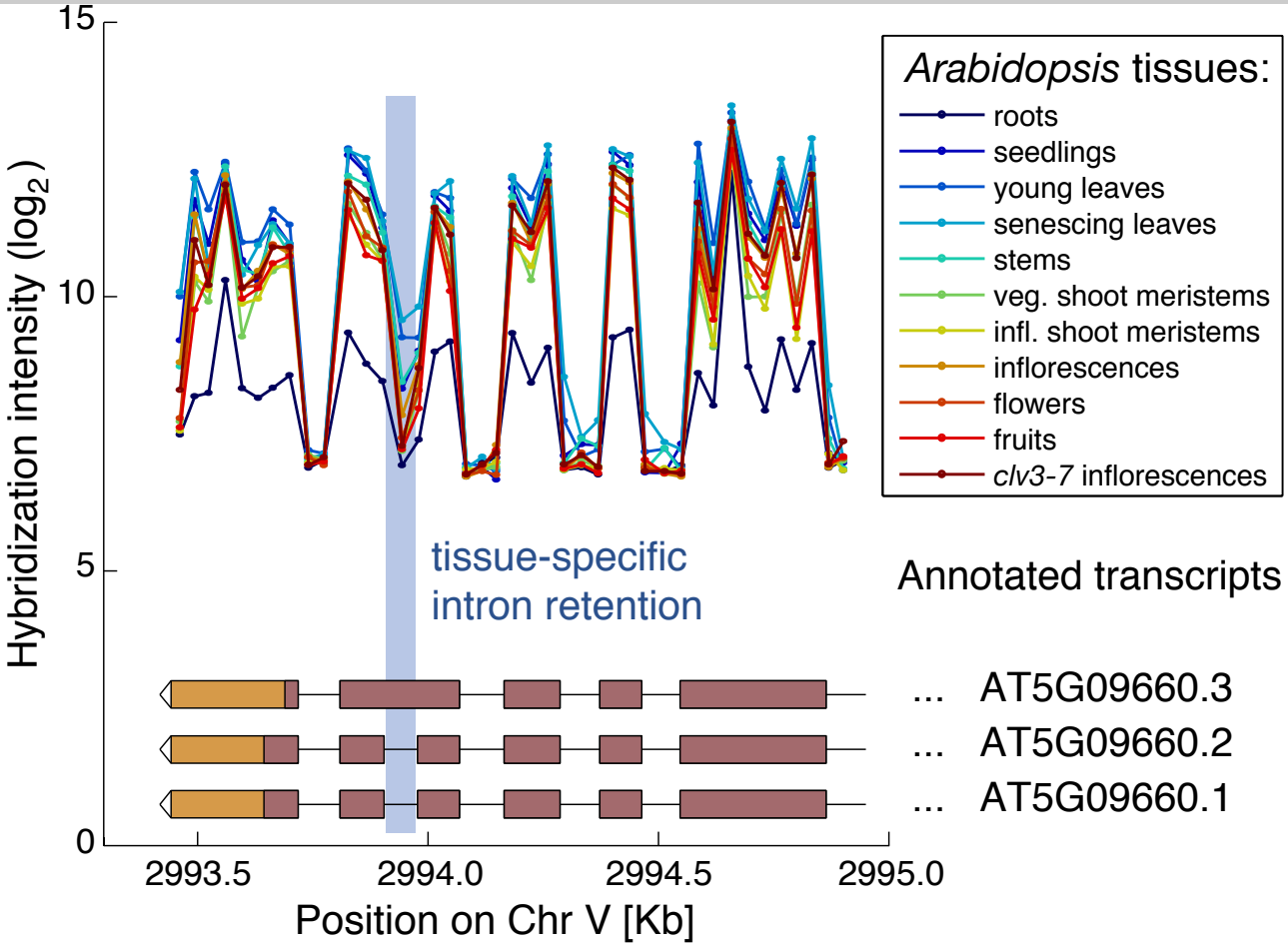
Alternatively Spliced Genes



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

[Eichner et al., 2008]

Differentially Spliced Genes

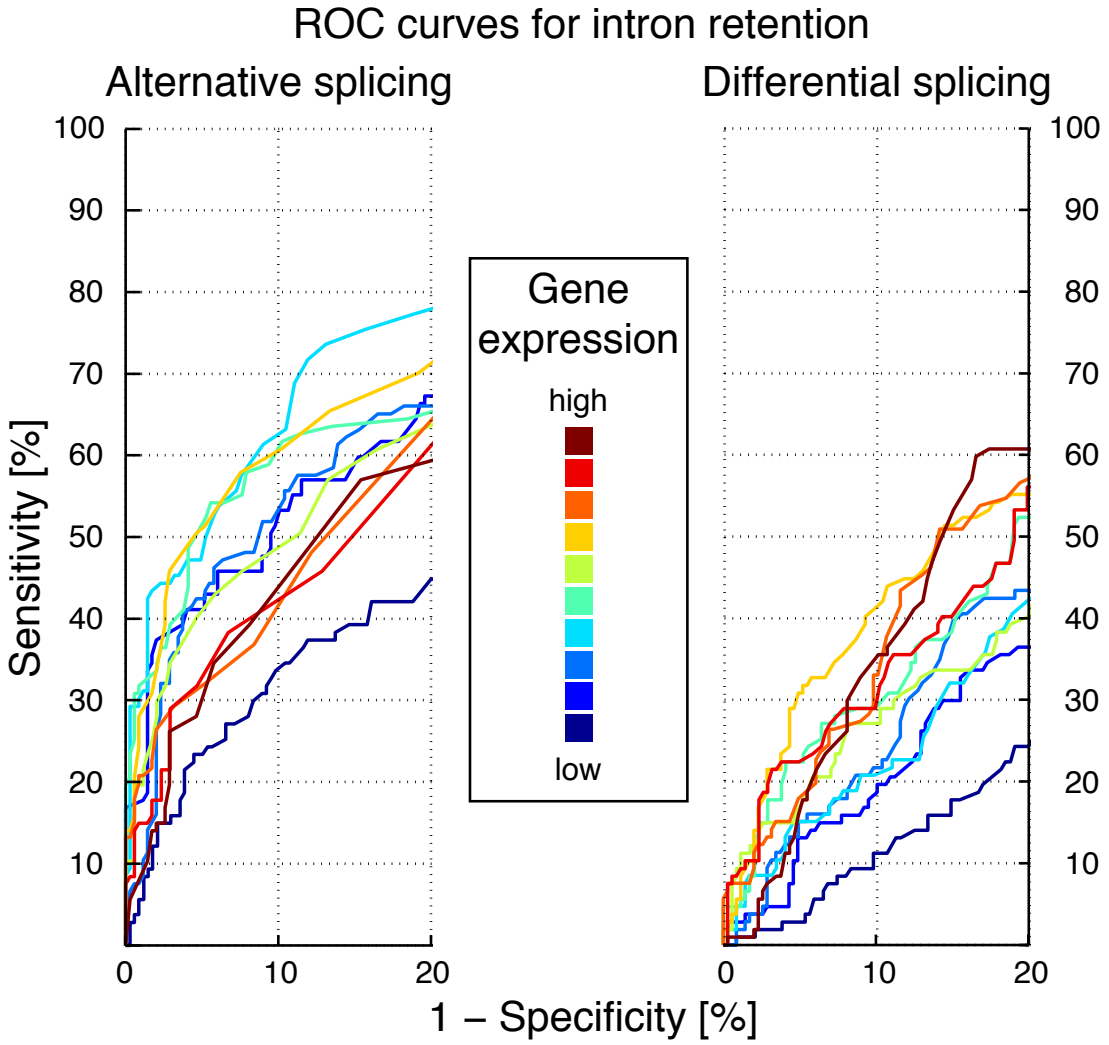


Goal: Identify exon/intron segments that are *differentially* spliced in the analyzed samples.

[Eichner et al., 2008]

Alternative vs. Differential Splicing

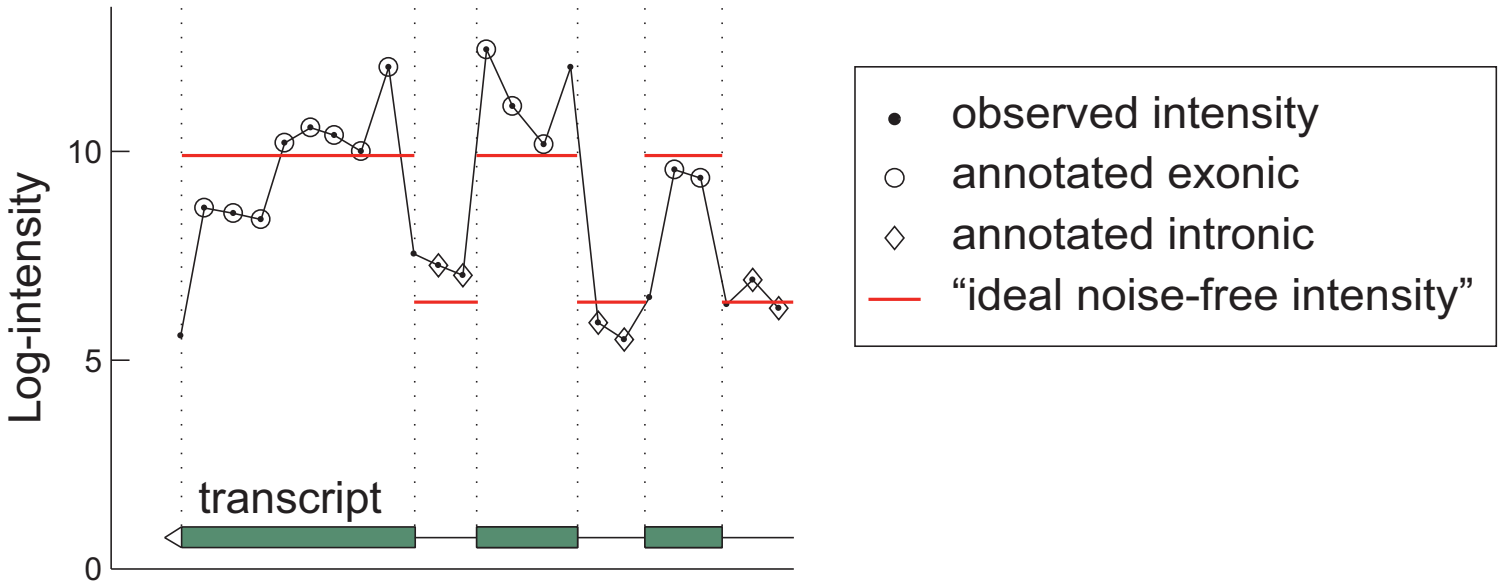
A Comparison with EST/cDNA-based Information



[Eichner et al., 2008]

Tiling Array Segmentation

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

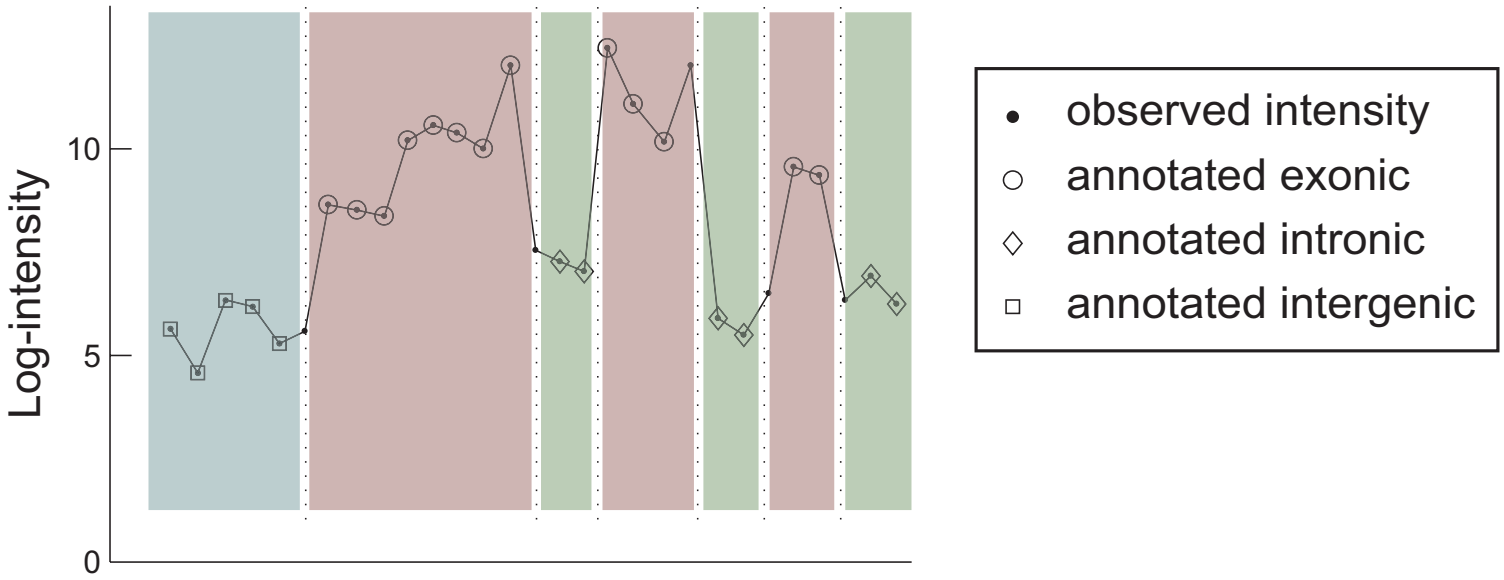


margin-based segmentation of tiling array data (**mSTAD**)
extends a segmentation method by Huber et al. [2006]

- very flexible noise model
- accounts for spliced transcripts
- parameters are learned on tiling array data from regions of known transcripts

Tiling Array Segmentation

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

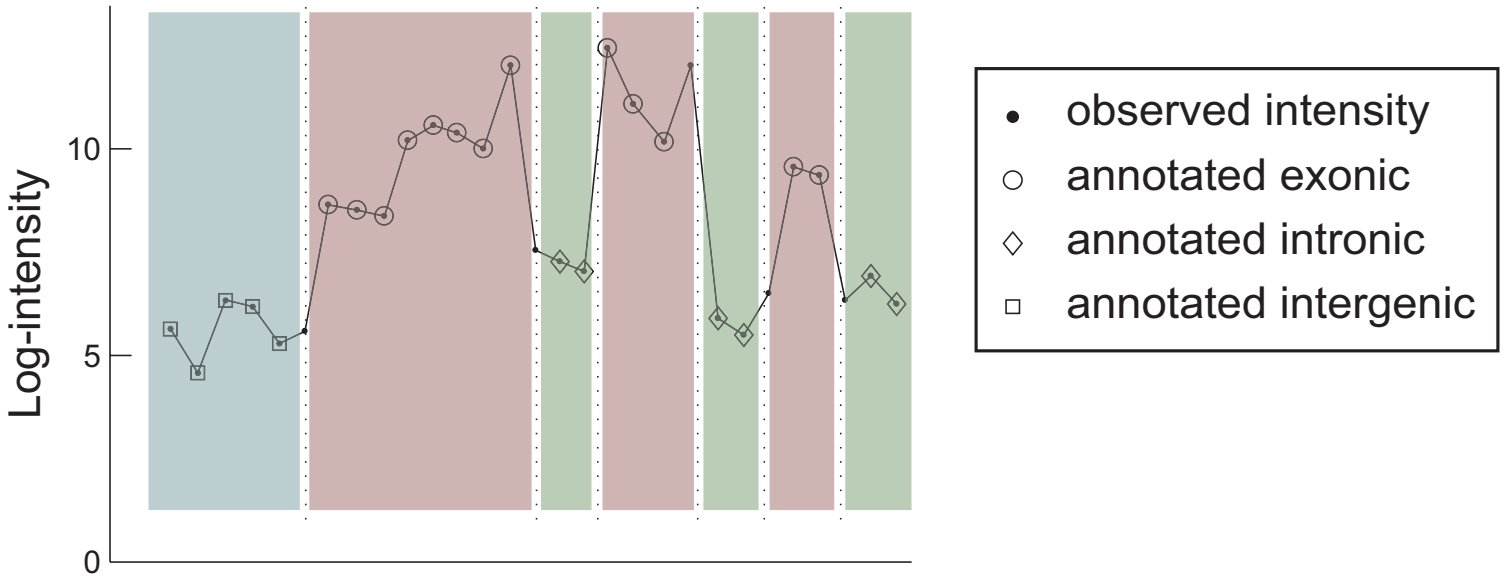


margin-based segmentation of tiling array data (**mSTAD**)
extends a segmentation method by Huber et al. [2006]

- very flexible noise model
- accounts for spliced transcripts
- parameters are learned on tiling array data from regions of known transcripts

Tiling Array Segmentation

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

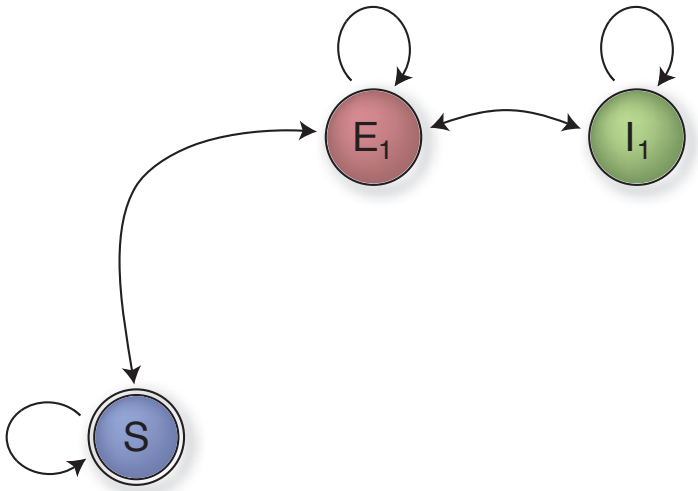


margin-based segmentation of tiling array data (**mSTAD**) extends a segmentation method by Huber et al. [2006]

- very flexible noise model
- accounts for spliced transcripts
- parameters are learned on tiling array data from regions of known transcripts

Tiling Array Segmentation

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



intergenic exonic intronic

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

$Q = 20$ discrete expression levels

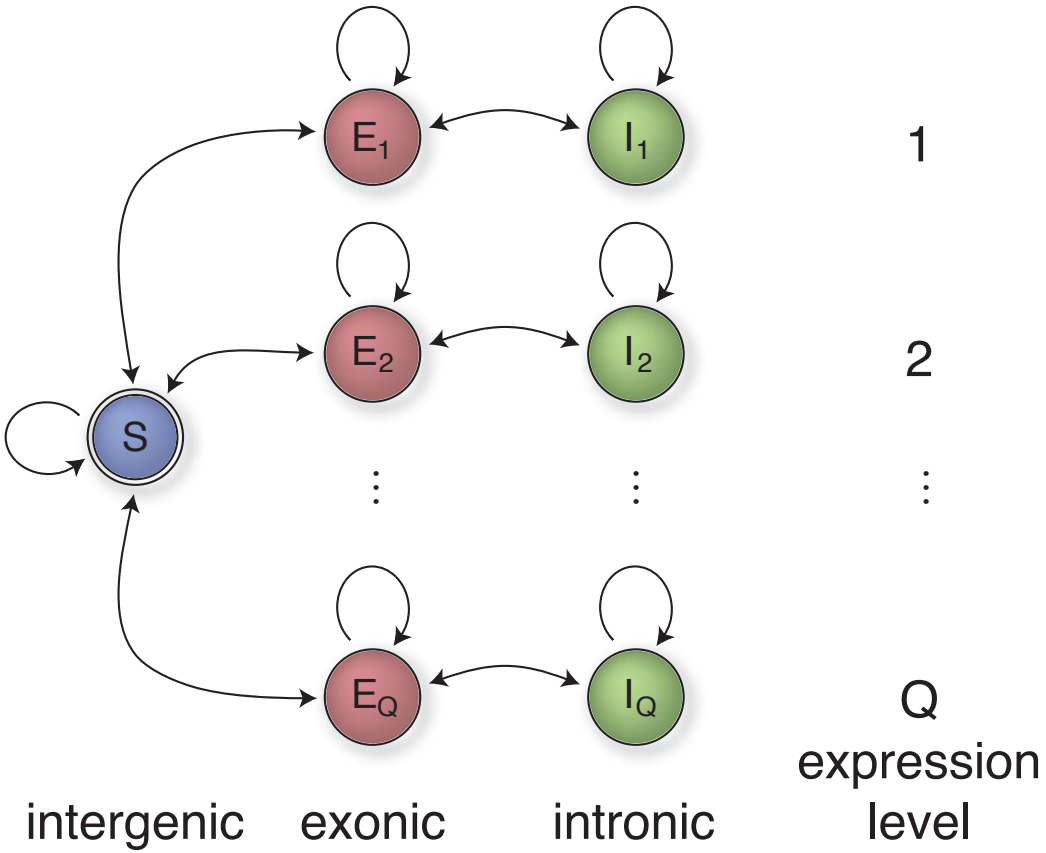
Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006]

[Zeller et al., 2008b]

Tiling Array Segmentation

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



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

$Q = 20$ discrete expression levels

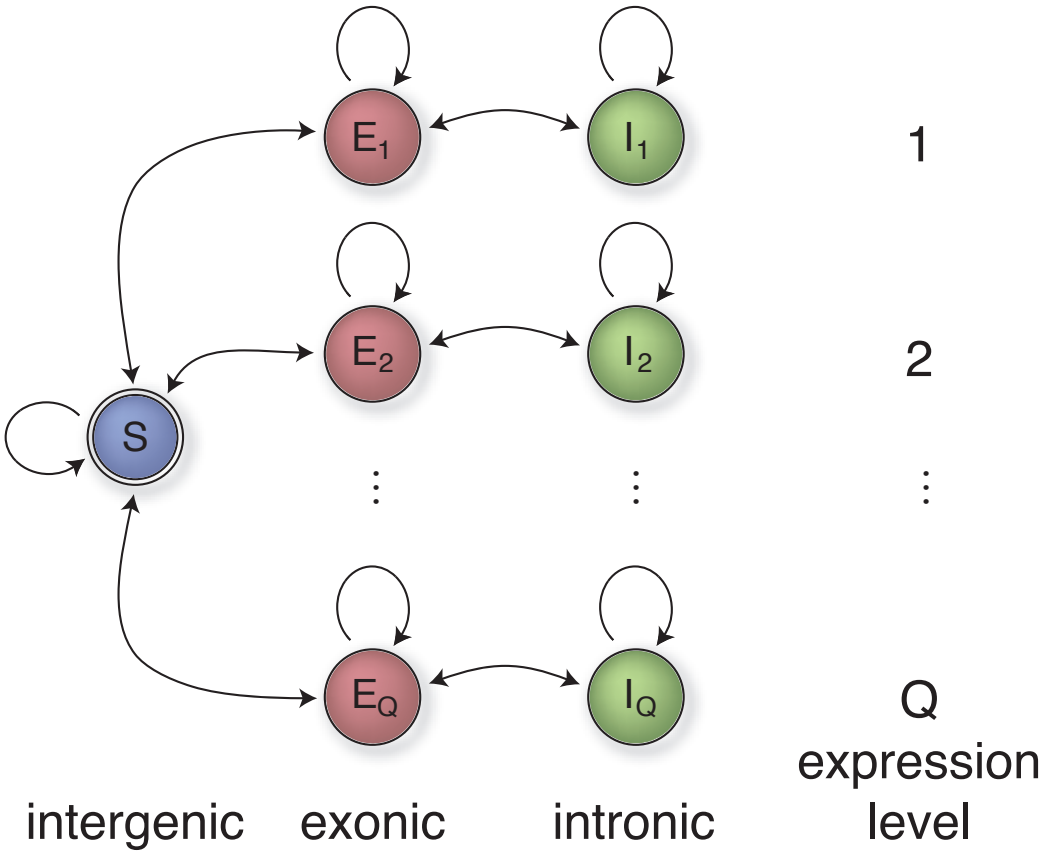
Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006]

[Zeller et al., 2008b]

Tiling Array Segmentation

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



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

$Q = 20$ discrete expression levels

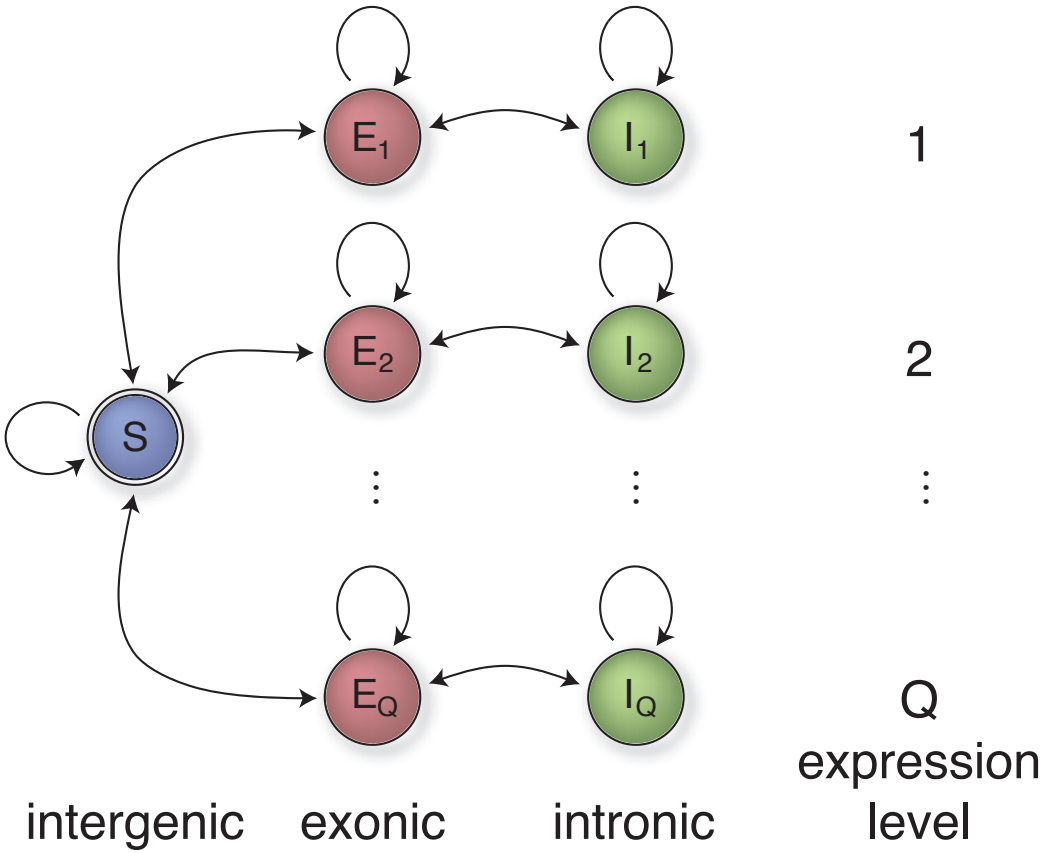
Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006]

[Zeller et al., 2008b]

Tiling Array Segmentation

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



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

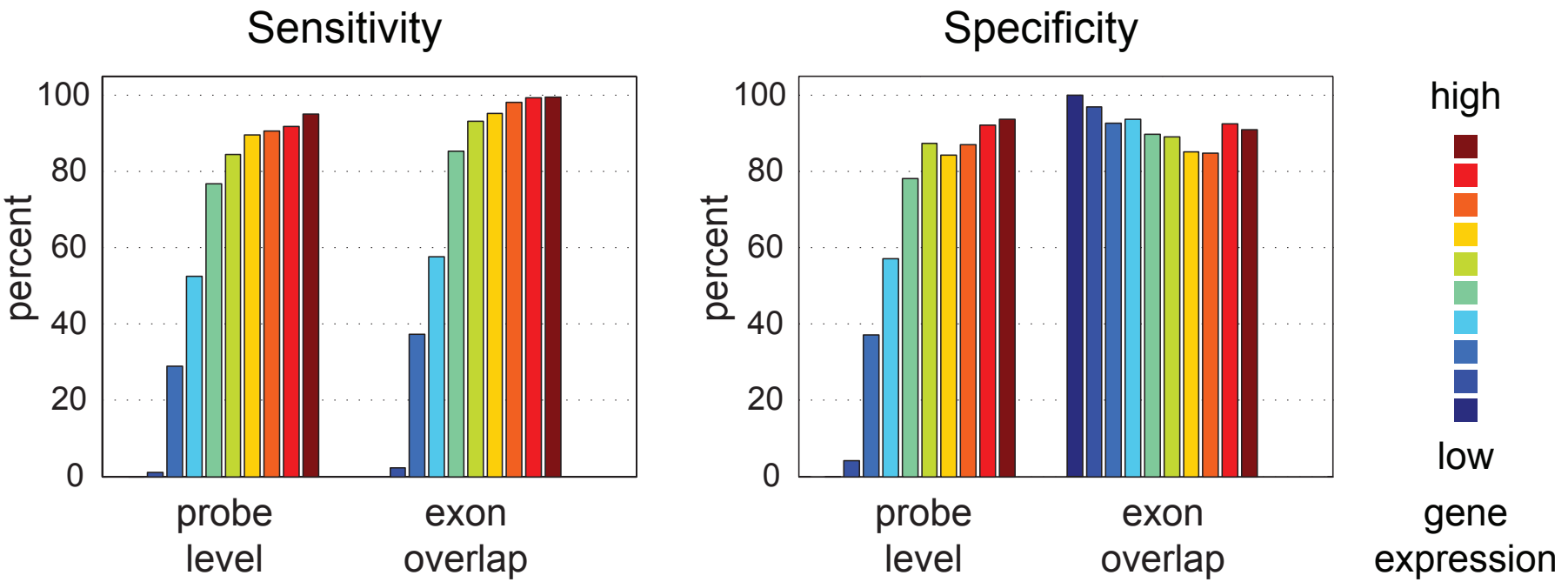
$Q = 20$ discrete expression levels

Use regions around annotated genes (TAIR7) for training.

Similar to GenRate model [Frey et al., 2006]

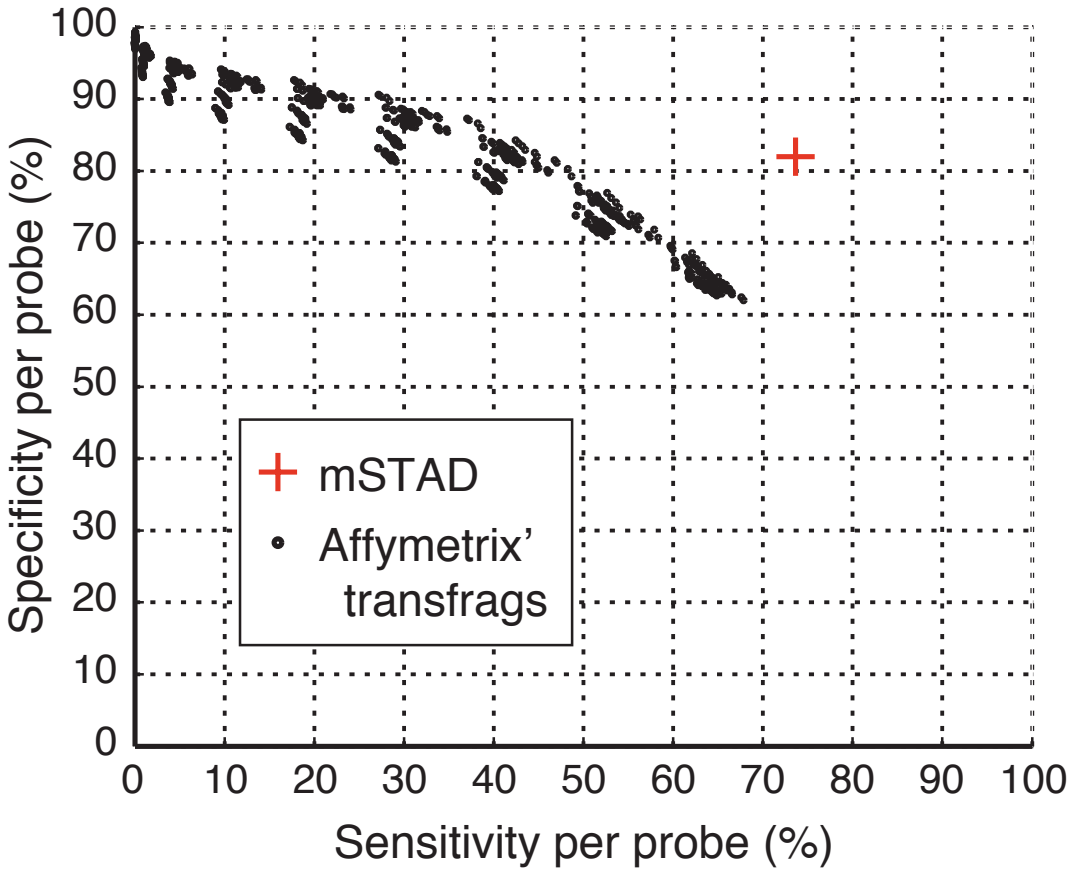
[Zeller et al., 2008b]

Segmentation Accuracy



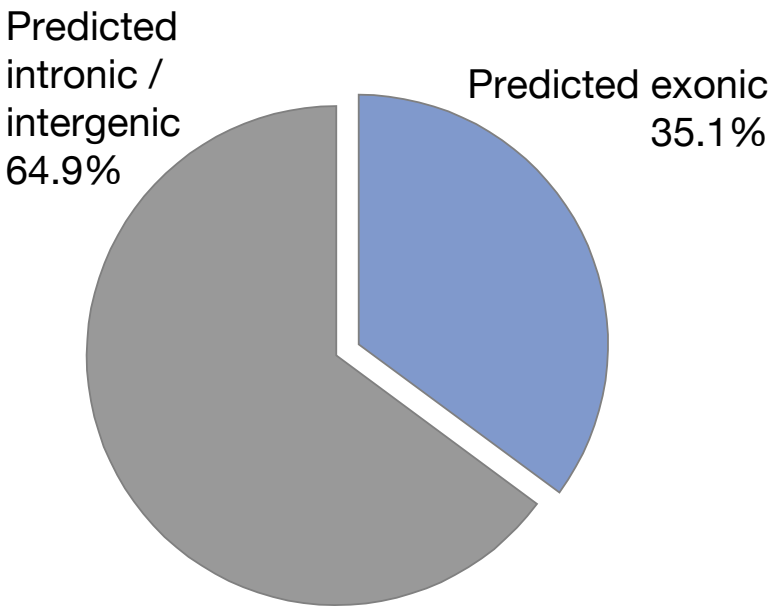
[Zeller et al., 2008b]

Comparison to Affymetrix's Transfrags



[Laubinger et al., 2008b]

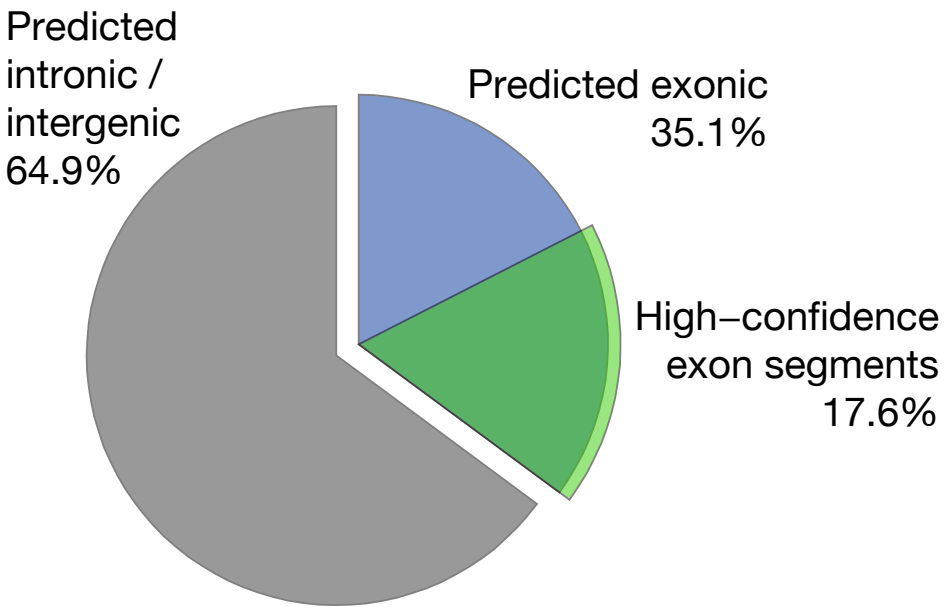
Discovering New Transcripts



- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]

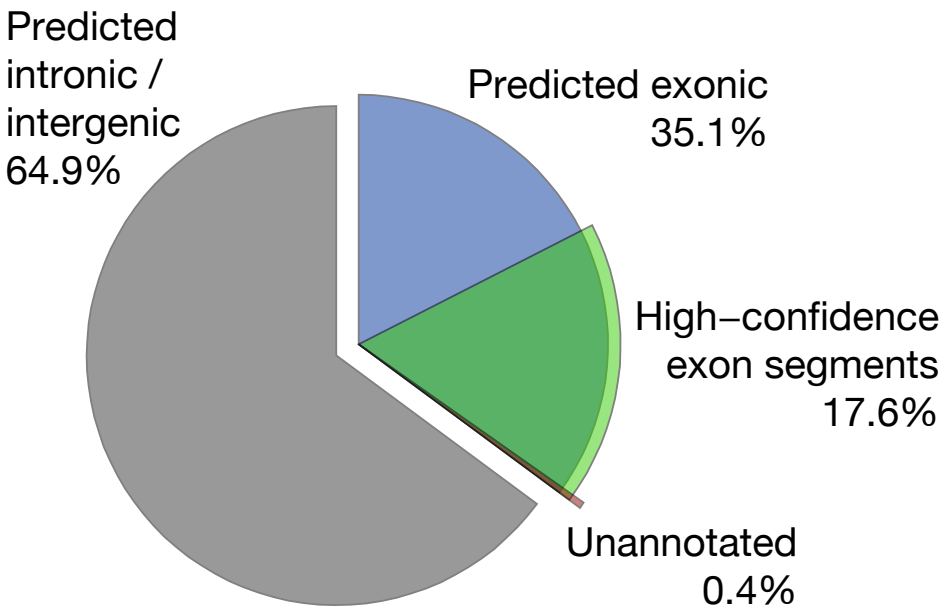
Discovering New Transcripts



- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]

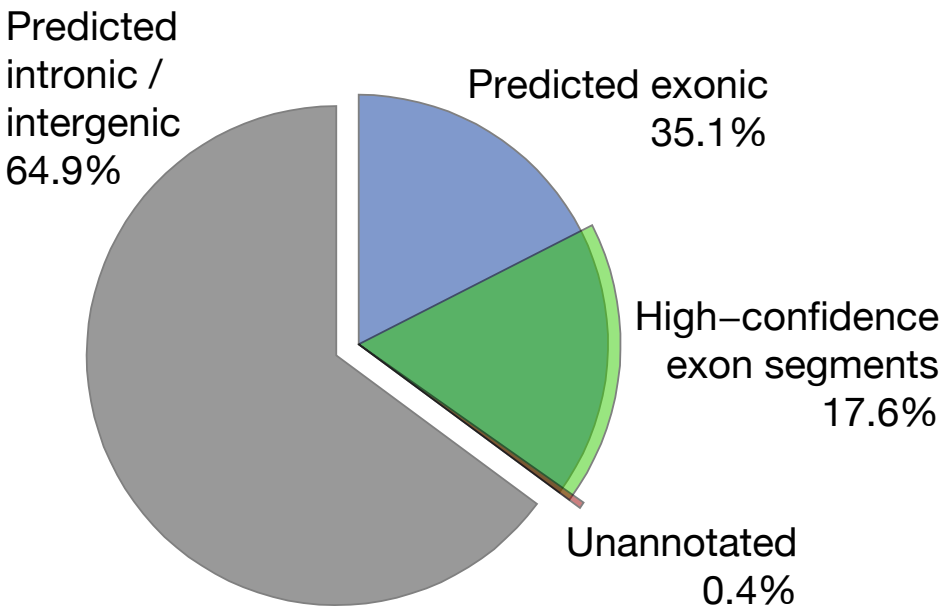
Discovering New Transcripts



- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]

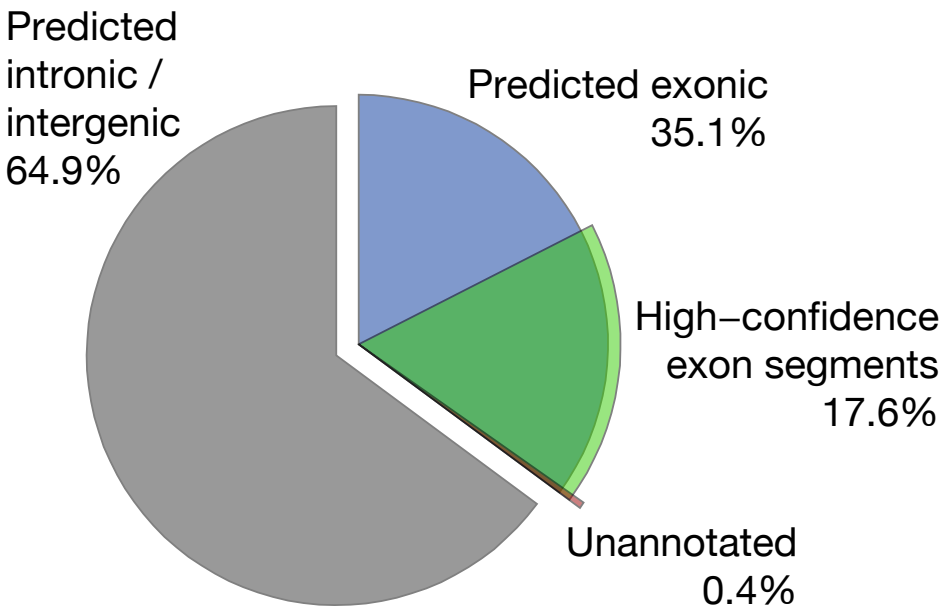
Discovering New Transcripts



- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]

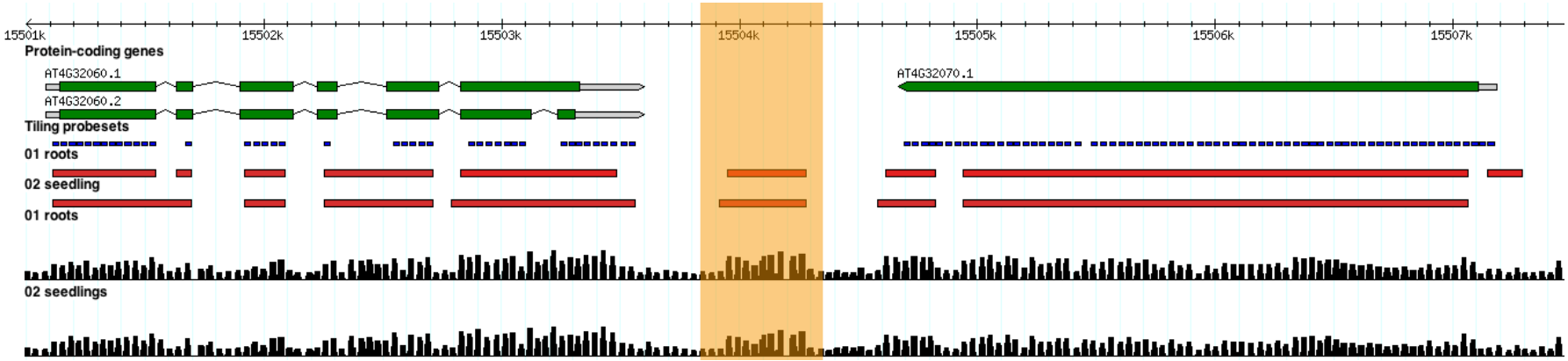
Discovering New Transcripts



- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]

Discovering New Transcripts

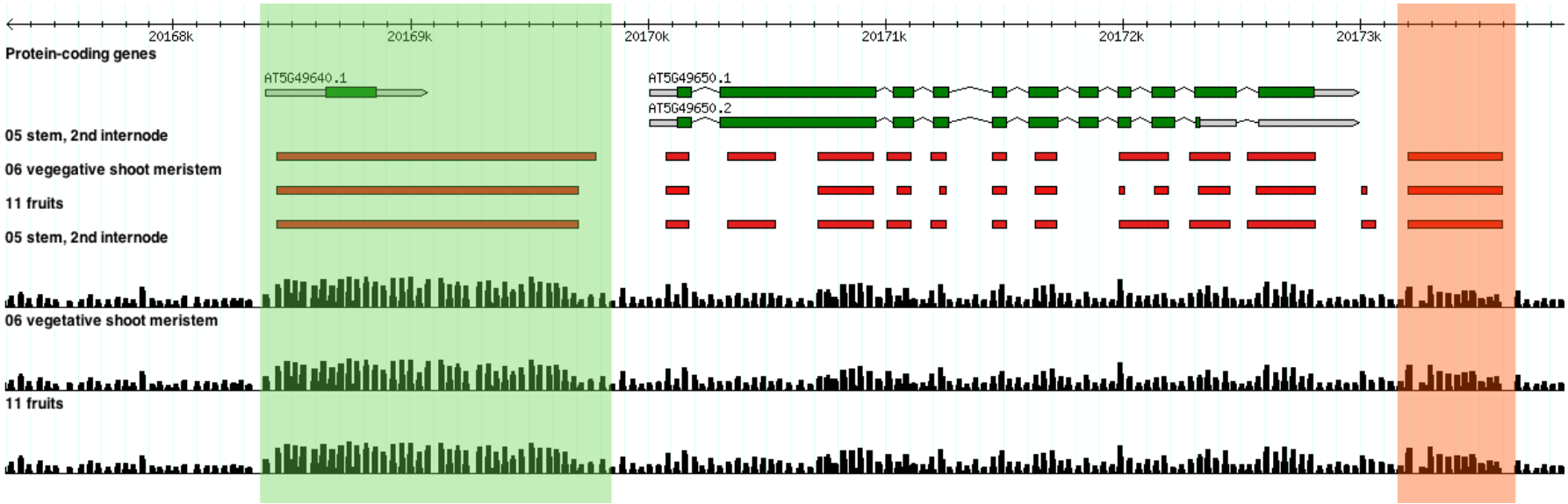


new transcript, not previously apparent from ESTs or cDNAs

- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]

Discovering New Transcripts



segment supported by ESTs; new exon segment (no ESTs)

- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

http://www.weigelworld.org/resources/microarray/at-tax [Laubinger et al., 2008b]

Outlook: Incorporate Sequence Information

- Incorporate sequence-based splice site predictions into **mSTAD**
  $\Rightarrow$ improved recognition of exon-intron boundaries
  $\Rightarrow$ no bias against non-coding transcripts
- Use tiling array data as feature for *ab initio* gene finder **mGene**
  $\Rightarrow$ highly accurate gene predictions for (protein-coding) genes with expression support.

Outlook: Incorporate Sequence Information

- Incorporate sequence-based splice site predictions into **mSTAD**
  $\Rightarrow$ improved recognition of exon-intron boundaries
  $\Rightarrow$ no bias against non-coding transcripts
- Use tiling array data as feature for *ab initio* gene finder **mGene**
  $\Rightarrow$ highly accurate gene predictions for (protein-coding) genes with expression support.

Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays
- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and *shorter* fragments
  - Much more *errors*
- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome sequencing
  - ...
- Transcriptome sequencing
  - Discovery of new genes
  - Identification of alternative splice forms
  - ...

Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays
- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and *shorter* fragments
  - Much more *errors*
- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome sequencing
  - ...
- Transcriptome sequencing
  - Discovery of new genes
  - Identification of alternative splice forms
  - ...

Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays
- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and *shorter* fragments
  - Much more *errors*
- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome sequencing
  - ...
- Transcriptome sequencing
  - Discovery of new genes
  - Identification of alternative splice forms
  - ...

Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays
- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and *shorter* fragments
  - Much more *errors*
- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome sequencing
  - ...
- Transcriptome sequencing
  - Discovery of new genes
  - Identification of alternative splice forms
  - ...

Next Generation Sequencing

- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays
- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and *shorter* fragments
  - Much more *errors*
- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome sequencing
  - ...
- Transcriptome sequencing
  - Discovery of new genes
  - Identification of alternative splice forms
  - ...

Spliced vs. Unspliced Alignments



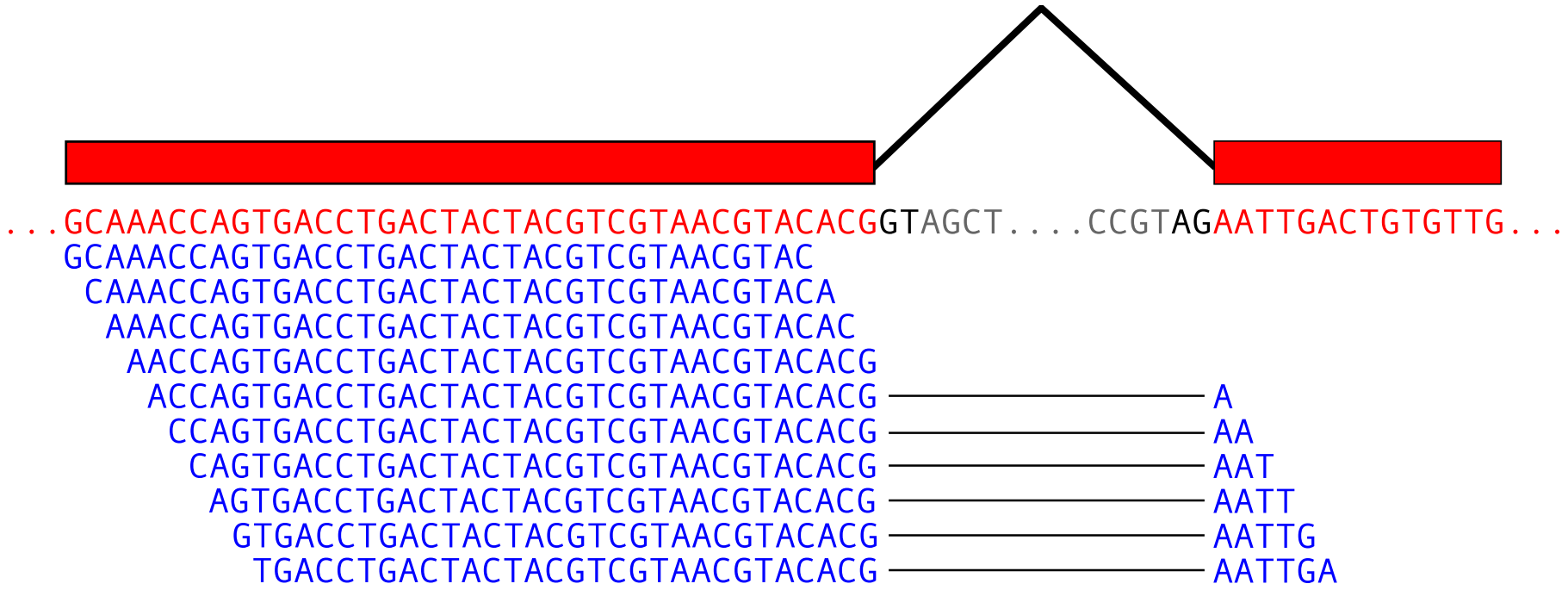
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

Spliced vs. Unspliced Alignments

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

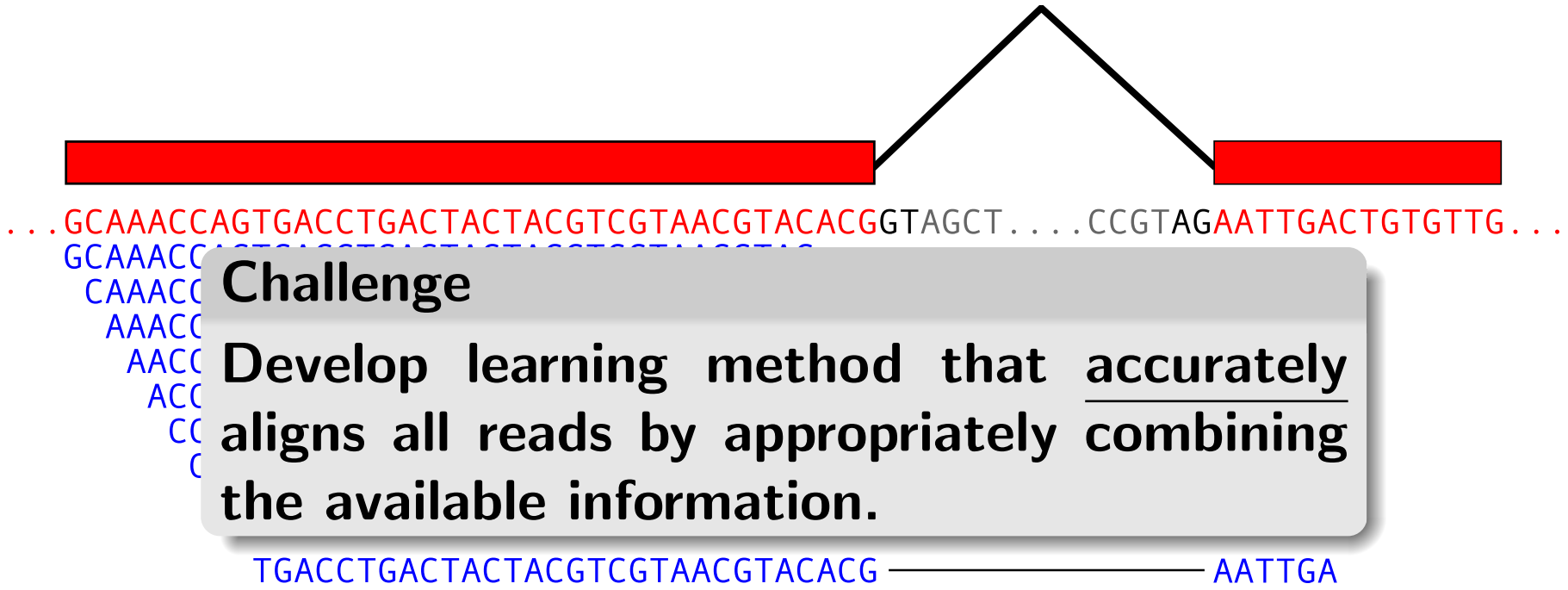
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

Spliced vs. Unspliced Alignments



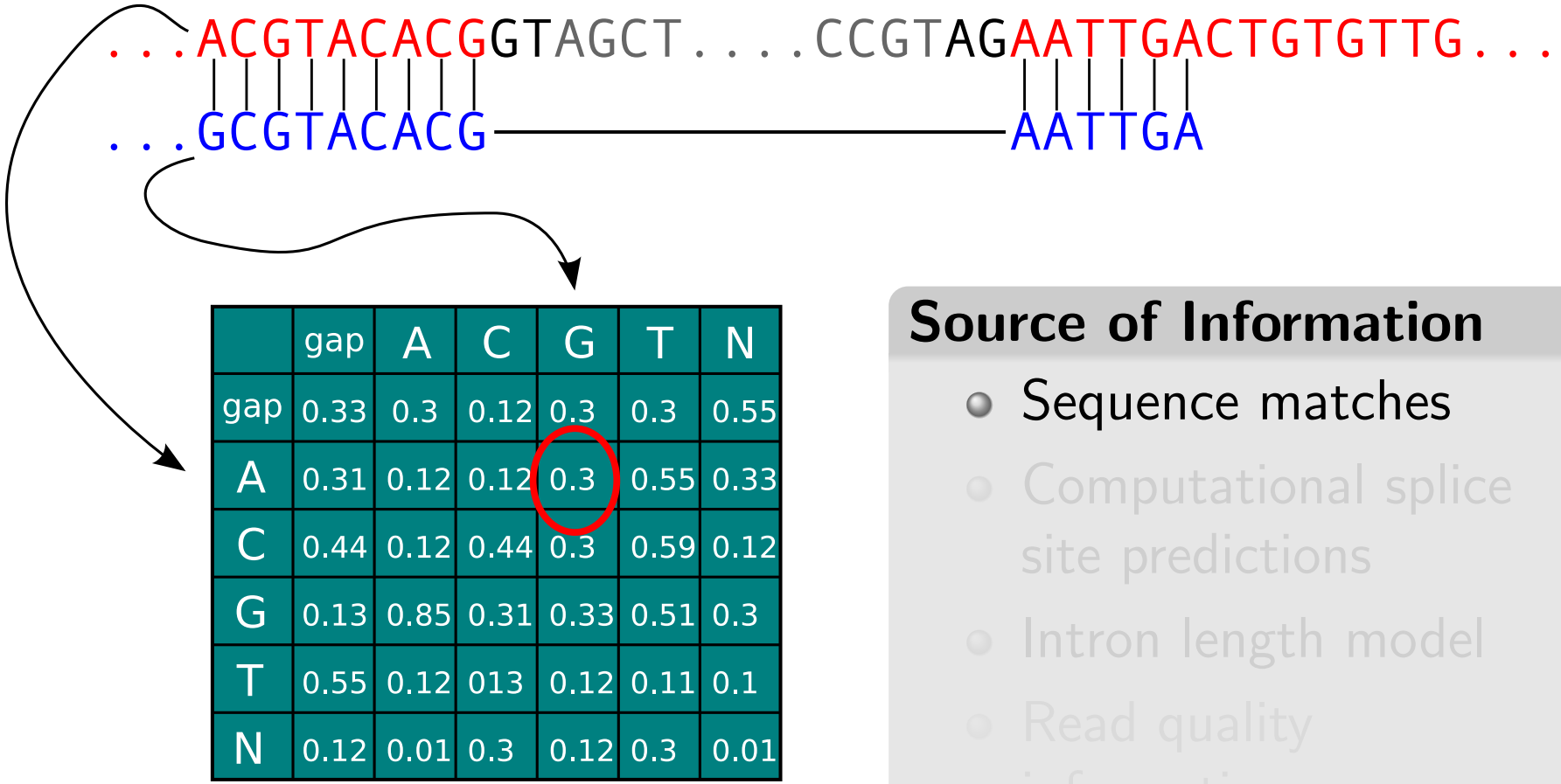
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

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

Alignment Scoring Function



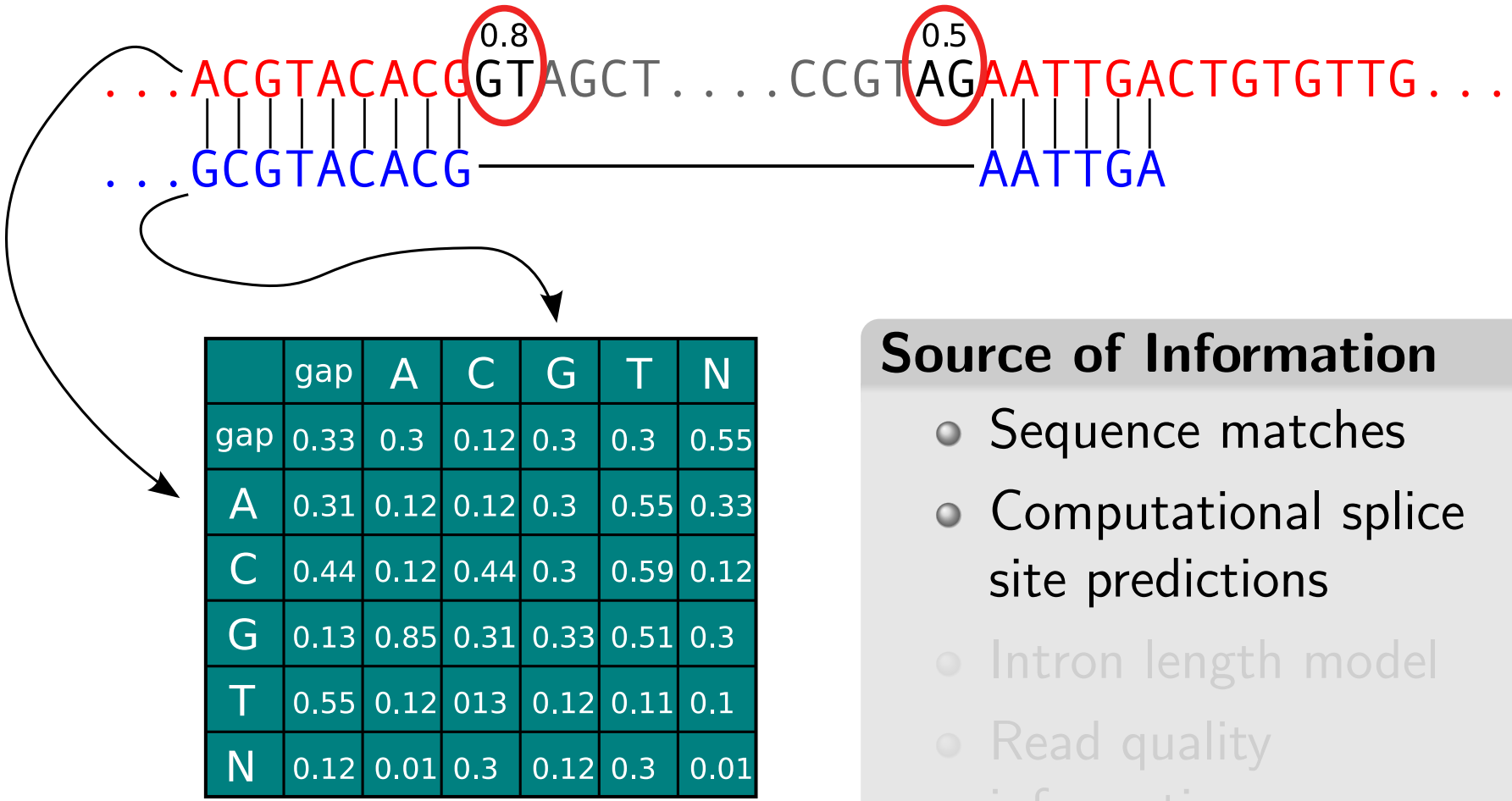
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Alignment Scoring Function



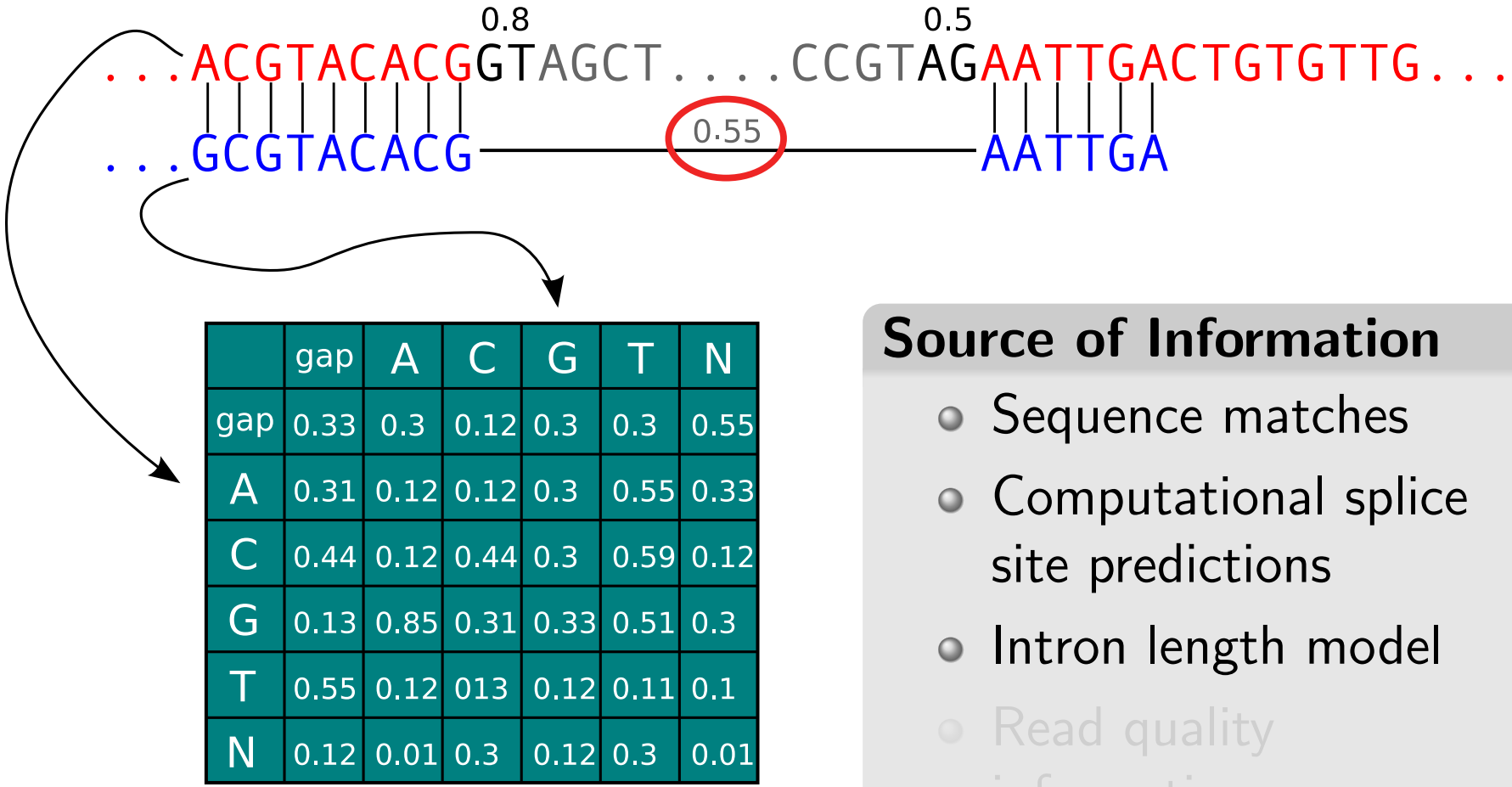
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Alignment Scoring Function



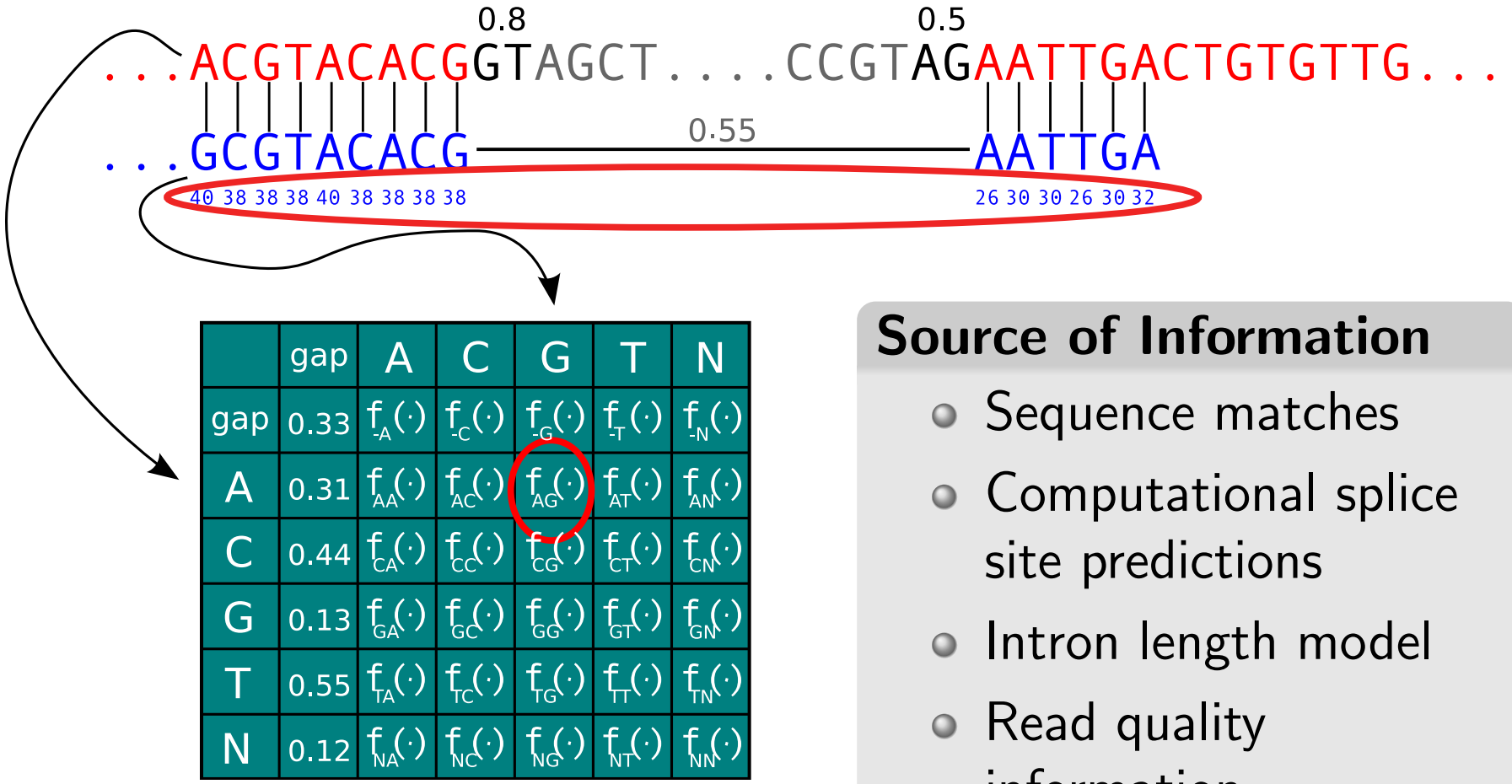
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Alignment Scoring Function

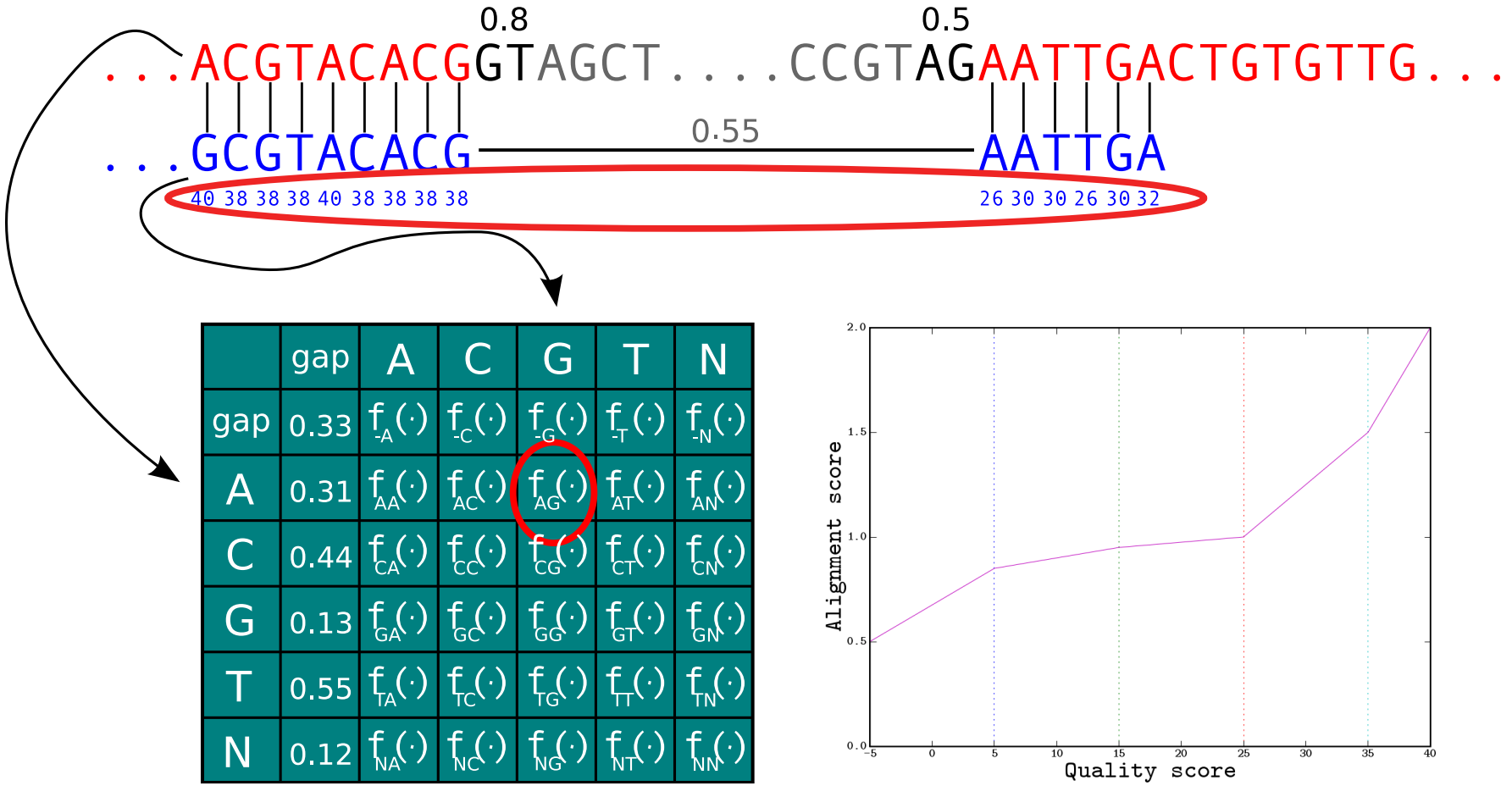


Source of Information

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

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

Alignment Scoring Function



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

Solving the Inverse Alignment Problem



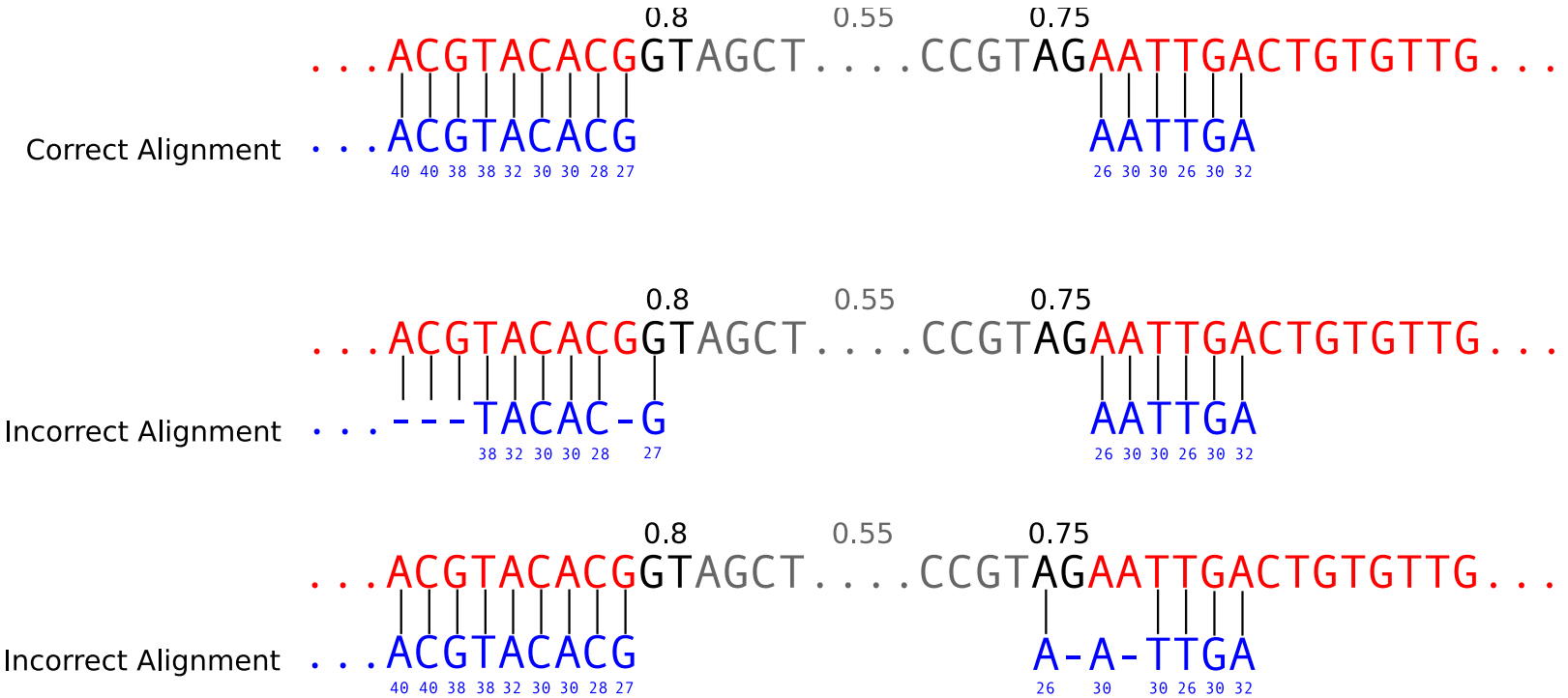
- How do we jointly optimize the 336 parameters?
- What are optimal parameters?

Example: three possible alignments

Solving the Inverse Alignment Problem

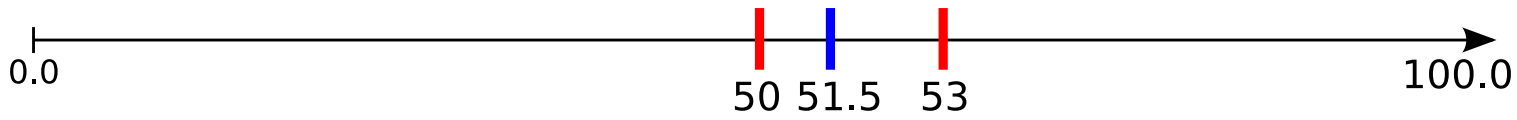
- How do we jointly optimize the 336 parameters?
- What are optimal parameters?

Example: three possible alignments



Cartoon

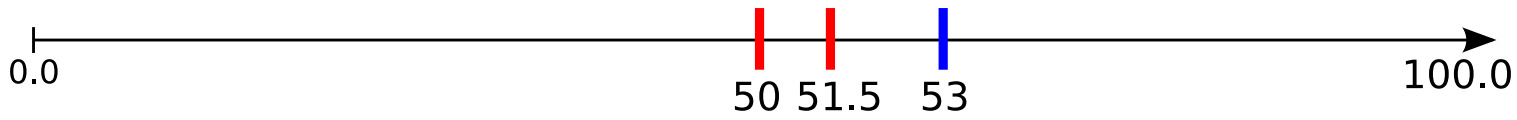
Scoring of the three alignments:



1. Correct alignment is **not** highest scoring one
2. Better parameters: now it is highest scoring. Can we do better?

Cartoon

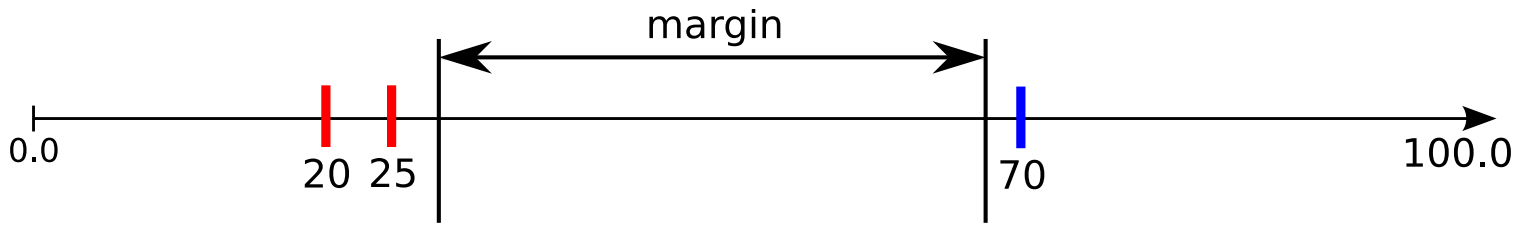
Scoring of the three alignments:



1. Correct alignment is **not** highest scoring one
2. Better parameters: now it is highest scoring. Can we do better?

Cartoon

Scoring of the three alignments:

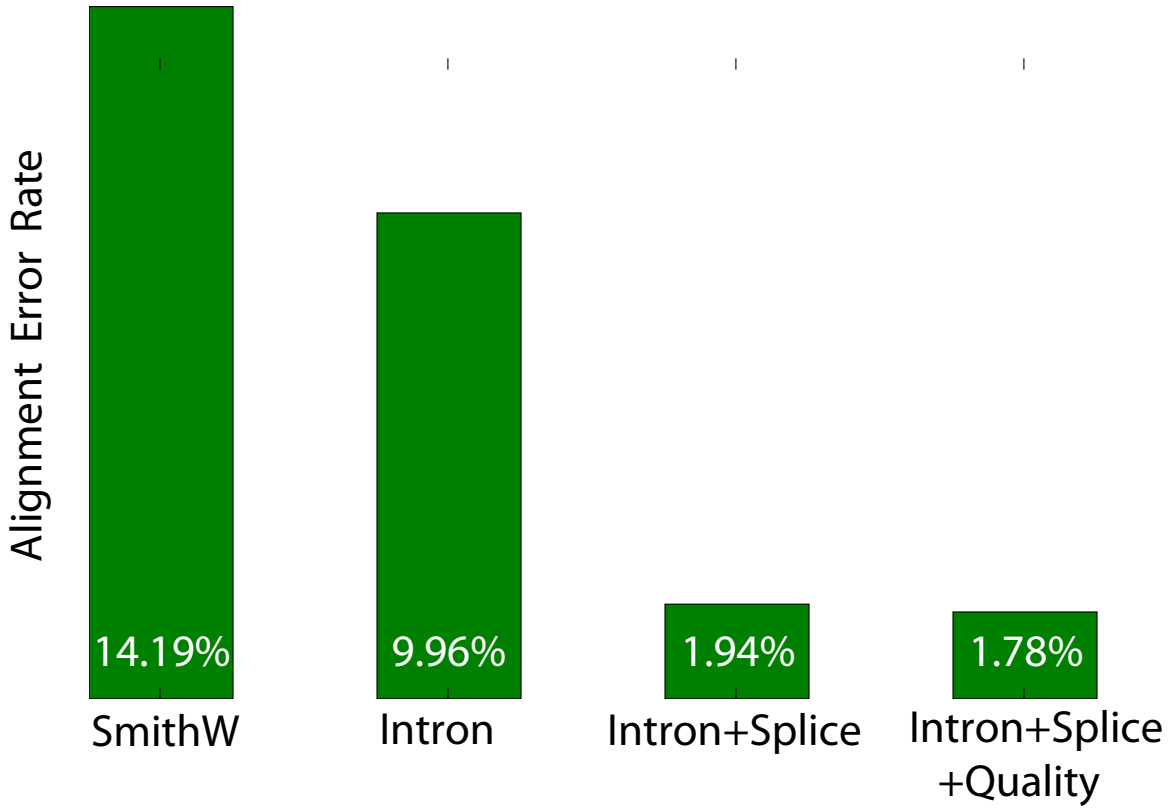


- Idea: Enforce a margin between correct and incorrect examples
- One has to solve a large quadratic optimization problem

First Experiment

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]

A Pipeline for Efficient Alignments

1. Run-time complexity of alignment $O(m \cdot n)$
2. Many reads will be fully contained in an exon

- Can we find smaller seed regions to align to?
- How do we discriminate between spliced/unspliced reads?

Pipeline Workflow (Example with $\approx$ 2.6 million reads)

1. Find seed regions
   ($\approx$ 4h for $2,586,170$ reads; 179 reads/second)
2. First run an approximation of the full model
   ($\approx$ 17min for $2,180,858$ reads; 417 reads/second)
3. Use the full model for the candidate spliced reads
   ($\approx$ 8h for $441,579$; 15 reads/second)

De Bona et al. [2008]

A Pipeline for Efficient Alignments

1. Run-time complexity of alignment $O(m \cdot n)$
2. Many reads will be fully contained in an exon

- Can we find smaller seed regions to align to?
- How do we discriminate between spliced/unspliced reads?

Pipeline Workflow (Example with $\approx$ 2.6 million reads)

1. Find seed regions
   ($\approx$ 4h for 2,586,170 reads; 179 reads/second)
2. First run an approximation of the full model
   ($\approx$ 17min for 2,180,858 reads; 417 reads/second)
3. Use the full model for the candidate spliced reads
   ($\approx$ 8h for 441,579; 15 reads/second)

De Bona et al. [2008]

A Pipeline for Efficient Alignments

1. Run-time complexity of alignment $O(m \cdot n)$
2. Many reads will be fully contained in an exon

- Can we find smaller seed regions to align to?
- How do we discriminate between spliced/unspliced reads?

Pipeline Workflow (Example with $\approx$ 2.6 million reads)

1. Find seed regions
   ($\approx$ 4h for $2,586,170$ reads; 179 reads/second)
2. First run an approximation of the full model
   ($\approx$ 17min for $2,180,858$ reads; 417 reads/second)
3. Use the full model for the candidate spliced reads
   ($\approx$ 8h for $441,579$; 15 reads/second)

De Bona et al. [2008]

Outlook: Methods for NG Sequencing

fml

So far:

- Adapted to Illumina 1G Genome Analyzer
  $\Rightarrow$ works similarly for other platforms
- Evaluation on artifically spliced reads
  $\Rightarrow$ how does it work in the real-world?

Working on:

- Getting it faster
- Include seed-finding in learning
- Methods for constructing splice graphs

Outlook: Methods for NG Sequencing

fml

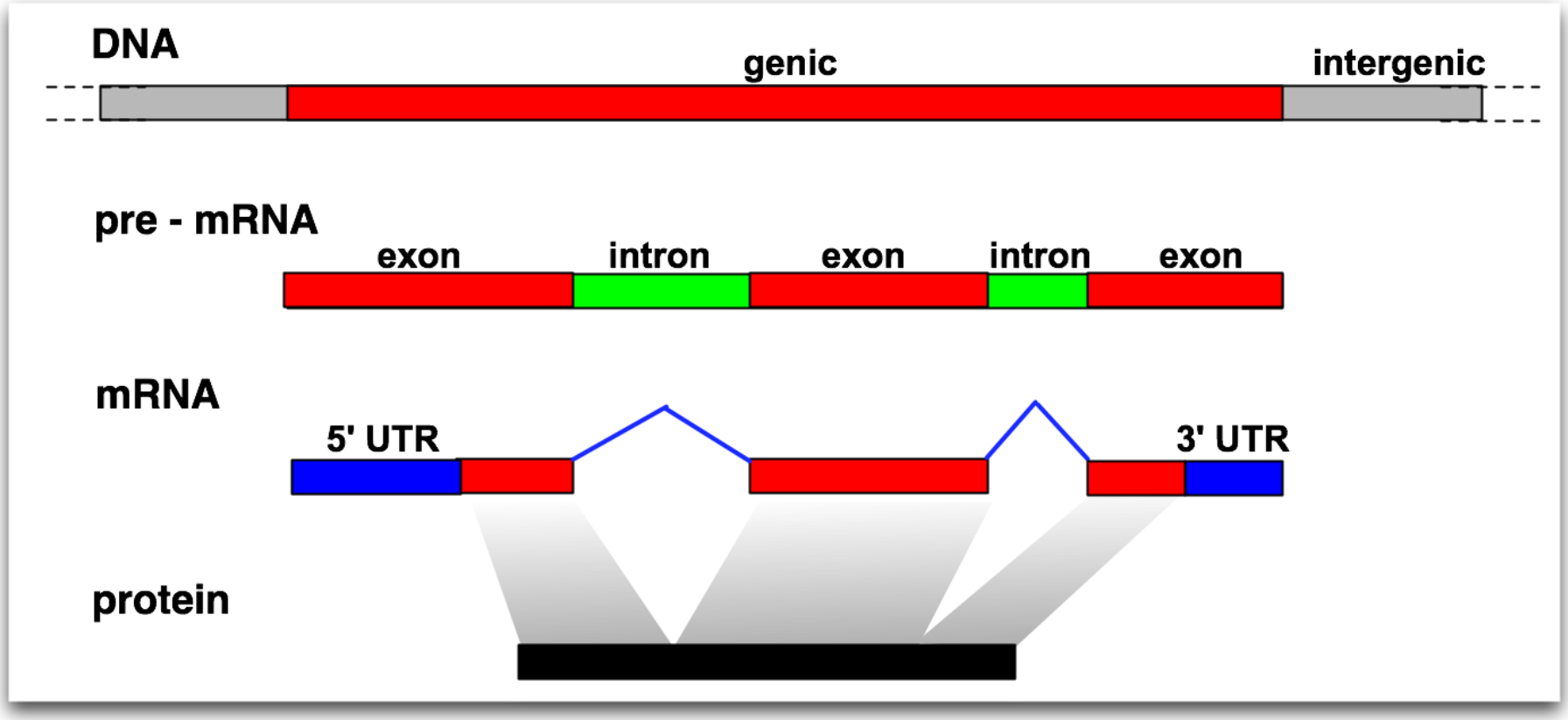
So far:

- Adapted to Illumina 1G Genome Analyzer
  $\Rightarrow$ works similarly for other platforms
- Evaluation on artifically spliced reads
  $\Rightarrow$ how does it work in the real-world?

Working on:

- Getting it faster
- Include seed-finding in learning
- Methods for constructing splice graphs

Gene Finding: Predict a Label Sequence

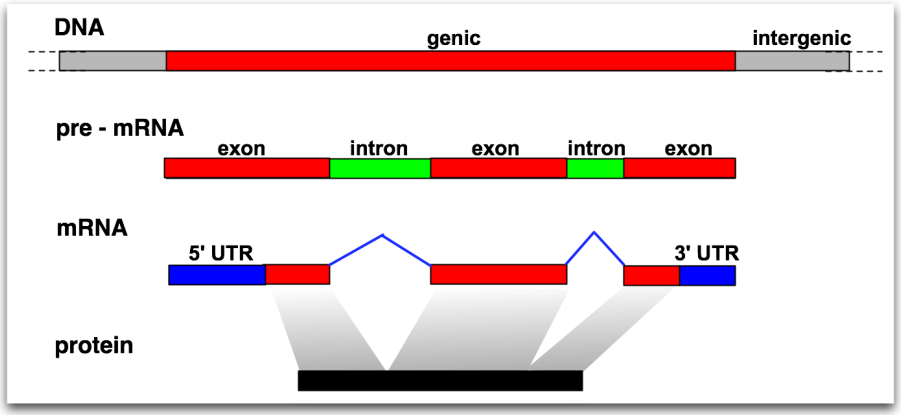


- Given a DNA sequence $\mathbf{x} \in \{\text{'A','C','G','T'}\}^L$
- Find the correct **label sequence y** $= y_1 y_2 \ldots y_L$
  ($y_i \in Y = \{$'intergenic', 'exon', 'intron', $\ldots\}$)

Standard Approach: HMMs

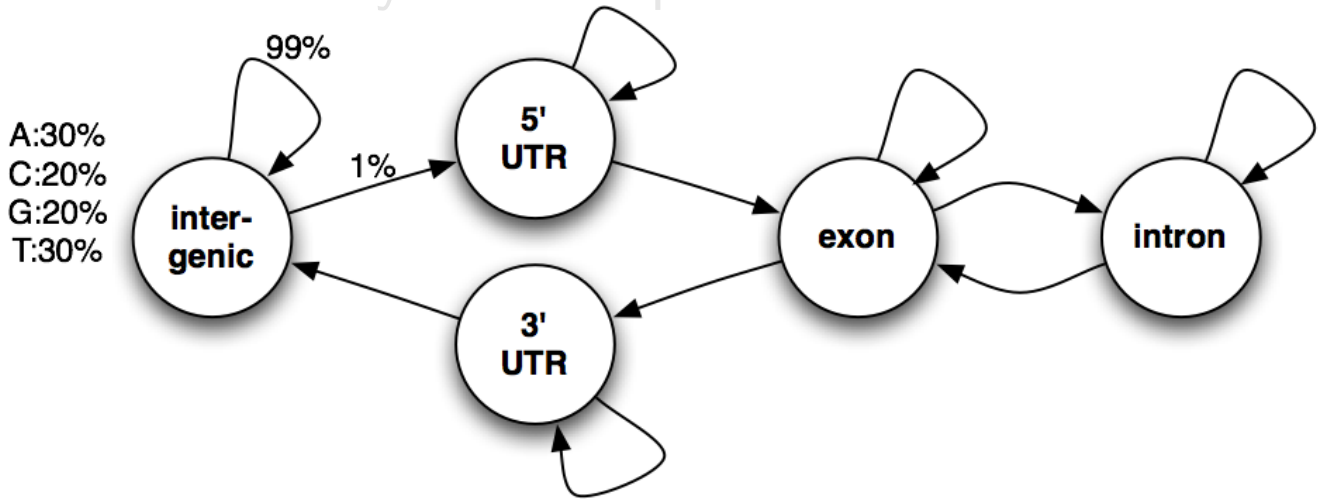
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

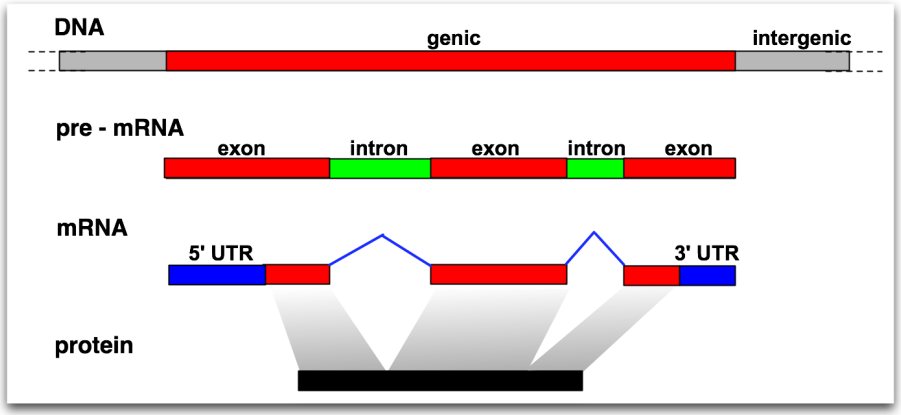
- Given DNA, find most likely state sequences



Standard Approach: HMMs

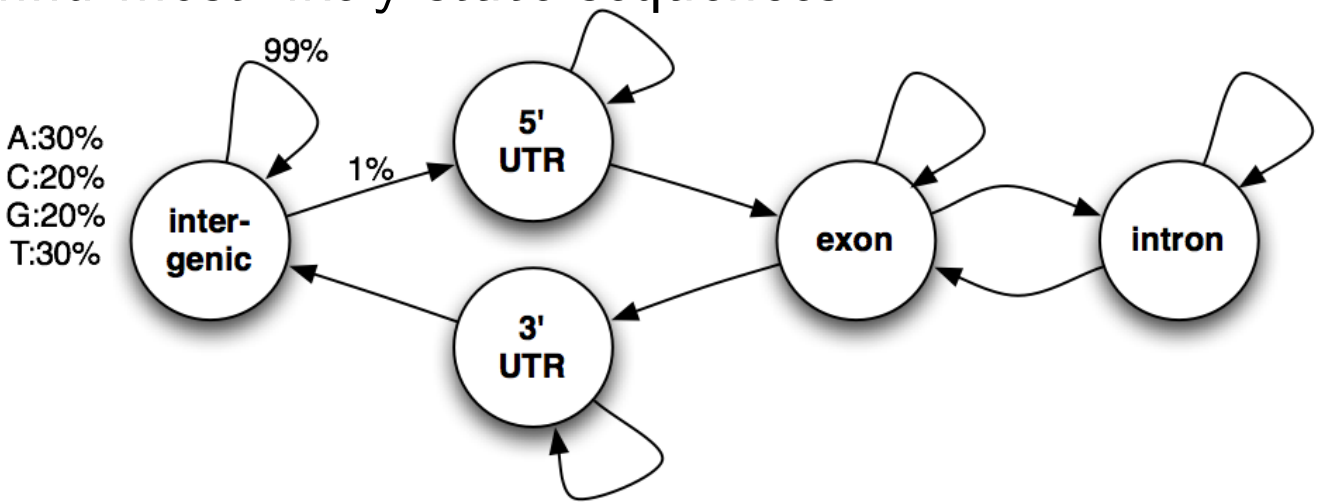
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



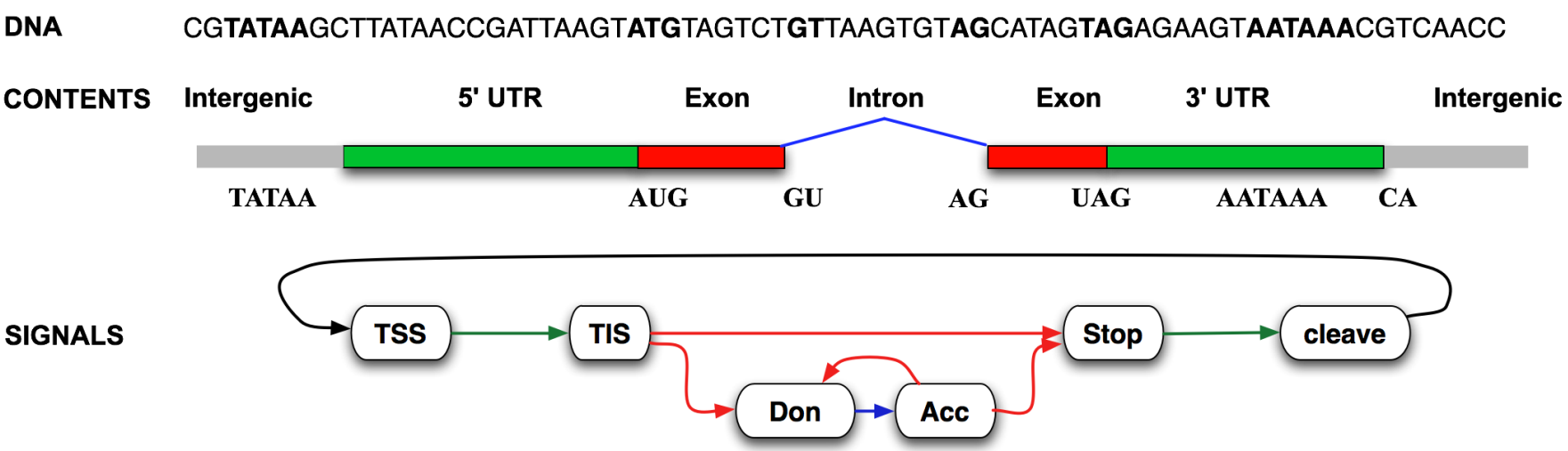
Prediction:

- Given DNA, find most likely state sequences



mGene: Modeling Signals & Content

- States correspond to sequence signals
  - Depends on recognition of signals on the DNA
- Transitions correspond to segments
  - Model length and content of segment



Recognition of Signals and Content

Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

- Exons
- Introns
- Intergenic

Distinguish one content type from all others

Typical approach: PSSMs or higher order Markov chains

We use Support Vector Machines

Recognition of Signals and Content



Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

- Exons
- Introns
- Intergenic

Distinguish one content type from all others

Typical approach: PSSMs or higher order Markov chains

We use Support Vector Machines

Recognition of Signals and Content



Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

- Exons
- Introns
- Intergenic

Distinguish one content type from all others

Typical approach: PSSMs or higher order Markov chains

We use Support Vector Machines

Recognition of Signals and Content



Sensors to recognize signals:

- Transcription start and cleavage site, polyA site
- Translation initiation site and stop codon
- Donor and acceptor splice sites

Discriminate true signal positions against all other positions

Sensors to recognize contents:

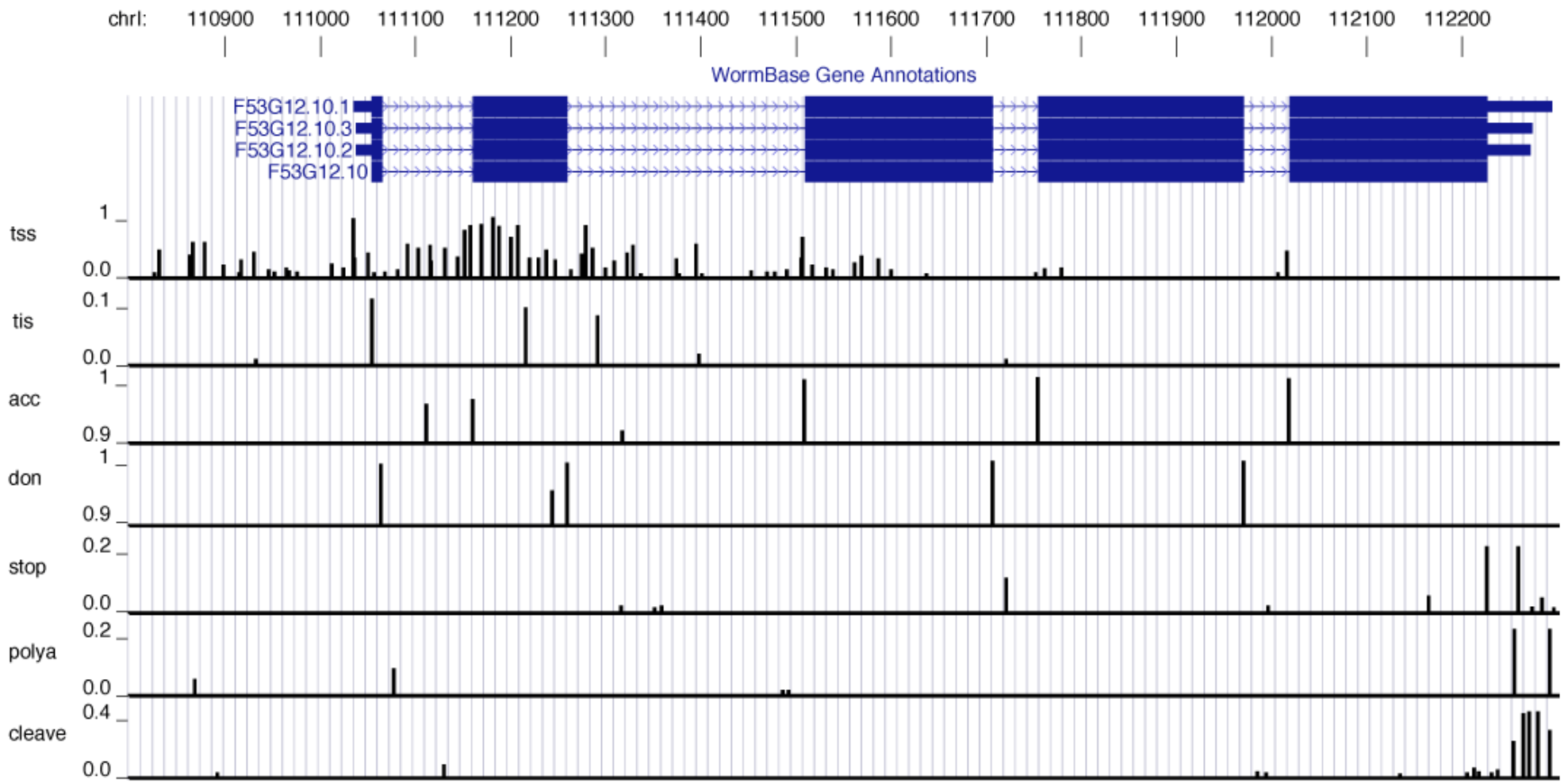
- Exons
- Introns
- Intergenic

Distinguish one content type from all others

Typical approach: PSSMs or higher order Markov chains

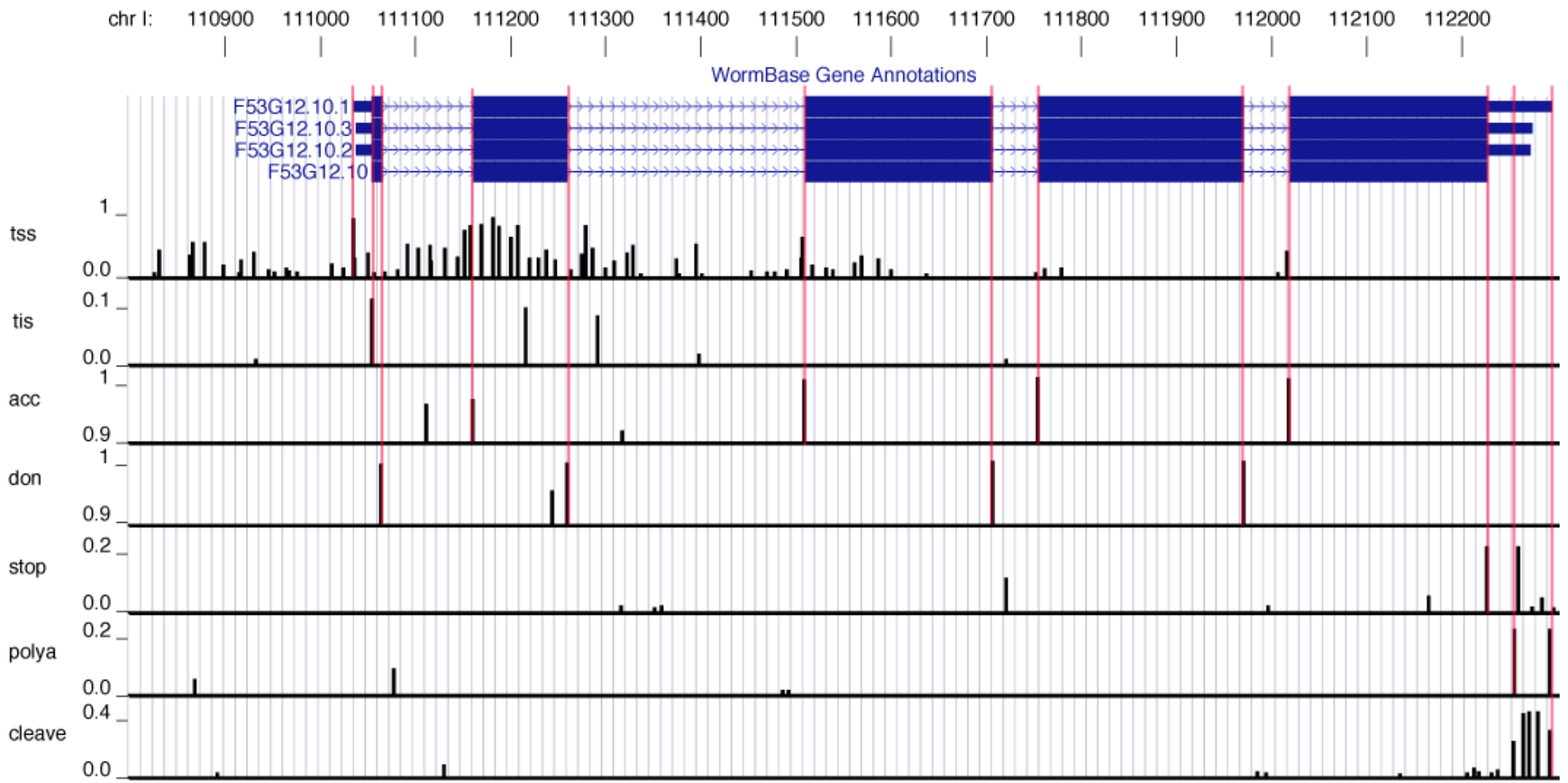
We use Support Vector Machines

Example: Predictions in UCSC Browser



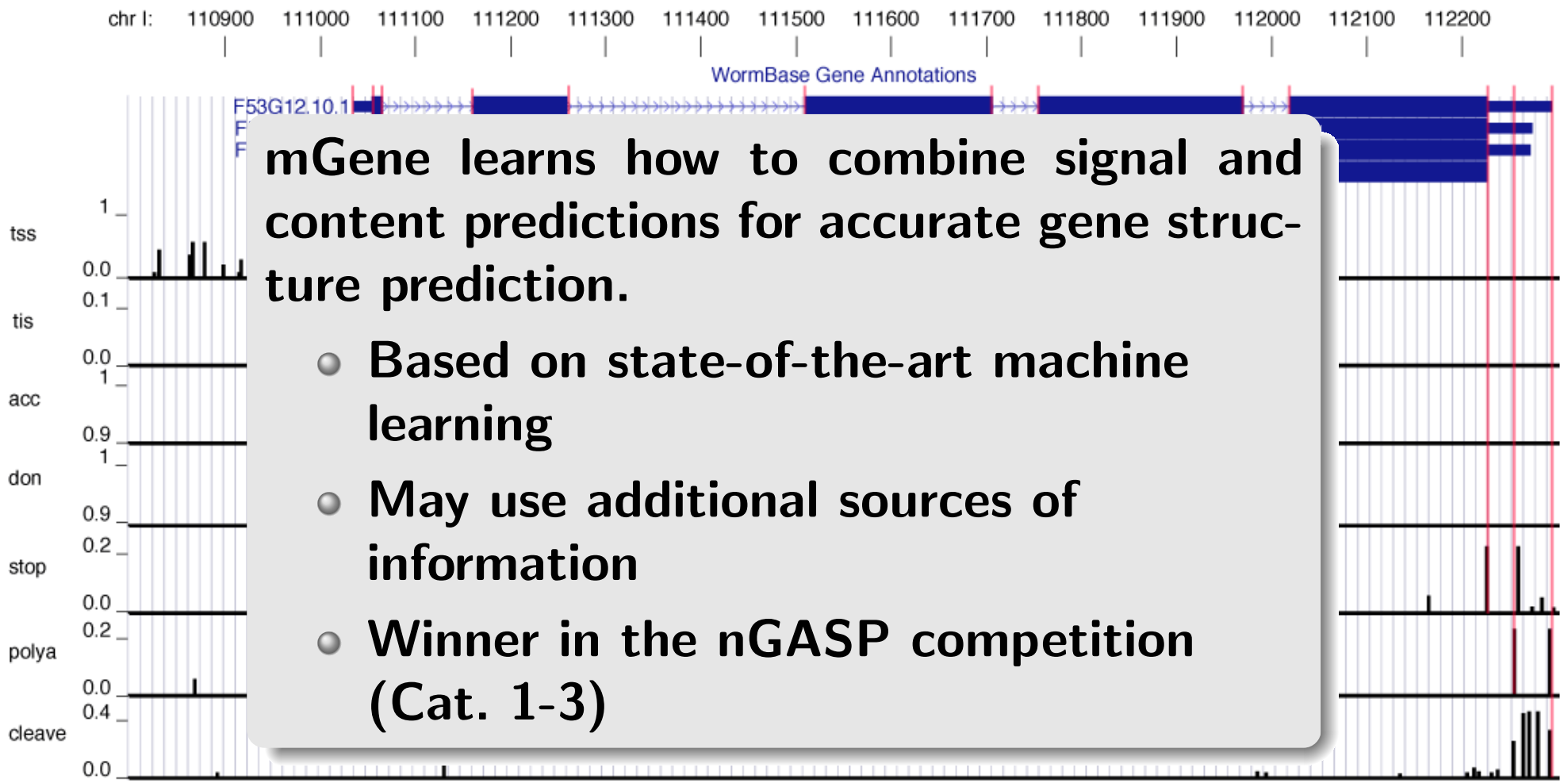
Schweikert et al. [2008]

Example: Predictions in UCSC Browser



Schweikert et al. [2008]

Example: Predictions in UCSC Browser

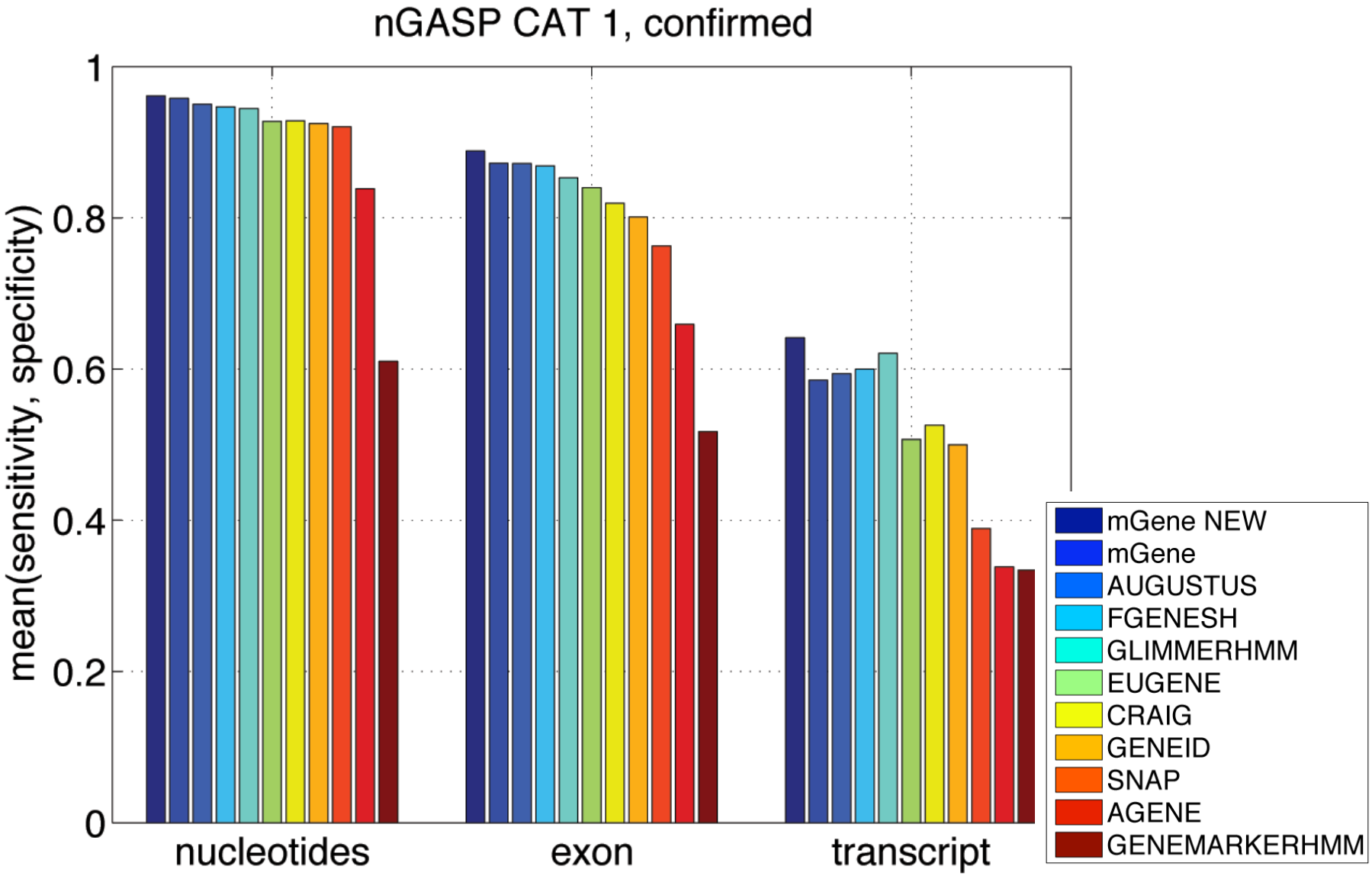


- **mGene learns how to combine signal and content predictions for accurate gene structure prediction.**
  - **Based on state-of-the-art machine learning**
  - **May use additional sources of information**
  - **Winner in the nGASP competition (Cat. 1-3)**

Schweikert et al. [2008]

Results of nGASP Competition (Cat. 1)

(Training and Testing on 10% of the C. elegans Genome)



Transcriptome Measurements for Improved Gene Finding



Ideas: Improve mGene by using

1. tiling array measurements as "content"-sensor track
2. base pair read coverage as "content"-sensor track
3. aligned spliced reads as high-confidence intron predictions

So far for A. thaliana: (preliminary)

1. *ab initio* mGene
   transcript level performance: **mean(SN, SP) = 74.3%**
2. Tiling array measurements in several tissues/conditions
   transcript level performance: **mean(SN, SP) = 78.3%**
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: **mean(SN, SP) = 76.5%**

Behr et al. [2008]

Transcriptome Measurements for Improved Gene Finding



Ideas: Improve mGene by using

1. tiling array measurements as "content"-sensor track
2. base pair read coverage as "content"-sensor track
3. aligned spliced reads as high-confidence intron predictions

So far for A. thaliana: (preliminary)

1. *ab initio* mGene
   transcript level performance: **mean(SN, SP) = 74.3%**
2. Tiling array measurements in several tissues/conditions
   transcript level performance: **mean(SN, SP) = 78.3%**
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: **mean(SN, SP) = 76.5%**

Behr et al. [2008]

Transcriptome Measurements for Improved Gene Finding

Ideas: Improve mGene by using

1. tiling array measurements as "content"-sensor track
2. base pair read coverage as "content"-sensor track
3. aligned spliced reads as high-confidence intron predictions

So far for A. thaliana: (preliminary)

1. *ab initio* mGene
   transcript level performance: **mean(SN, SP) = 74.3%**
2. Tiling array measurements in several tissues/conditions
   transcript level performance: **mean(SN, SP) = 78.3%**
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: **mean(SN, SP) = 76.5%**

Behr et al. [2008]

Transcriptome Measurements for Improved Gene Finding



Ideas: Improve mGene by using

1. tiling array measurements as "content"-sensor track
2. base pair read coverage as "content"-sensor track
3. aligned spliced reads as high-confidence intron predictions

So far for A. thaliana: (preliminary)

1. *ab initio* mGene
   transcript level performance: **mean(SN, SP) = 74.3%**
2. Tiling array measurements in several tissues/conditions
   transcript level performance: **mean(SN, SP) = 78.3%**
3. mRNA-seq (15x) for base pair read coverage
   transcript level performance: **mean(SN, SP) = 76.5%**

Behr et al. [2008]

Conclusions

- Analysis of Tiling Array Data
  - Proper normalization helps downstream analyses
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
- Short Read Alignments
  - Integrates splice site predictions & quality information
  - Novel technique to learn how to combine information
- Transcriptome Measurements
  - Lead to improved gene finding
  - Allow us to validate our assumptions in gene finding
  - Give rise to interesting computational challenges
  - Help to uncover the full complexity of transcriptomes

Conclusions

- Analysis of Tiling Array Data
  - Proper normalization helps downstream analyses
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
- Short Read Alignments
  - Integrates splice site predictions & quality information
  - Novel technique to learn how to combine information
- Transcriptome Measurements
  - Lead to improved gene finding
  - Allow us to validate our assumptions in gene finding
  - Give rise to interesting computational challenges
  - Help to uncover the full complexity of transcriptomes

Conclusions

- Analysis of Tiling Array Data
  - Proper normalization helps downstream analyses
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
- Short Read Alignments
  - Integrates splice site predictions & quality information
  - Novel technique to learn how to combine information
- Transcriptome Measurements
  - Lead to improved gene finding
  - Allow us to validate our assumptions in gene finding
  - Give rise to interesting computational challenges
  - Help to uncover the full complexity of transcriptomes

Acknowledgments

Tiling Arrays

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

Short Read Alignments

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

Gene Finding

- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Alex Zien (FML & FIRST)
- Georg Zeller (FML & MPI)

More Information

- `http://www.fml.mpg.de`
- Slides are available online

Thank you!

Acknowledgments

Tiling Arrays

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

Short Read Alignments

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

Gene Finding

- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Alex Zien (FML & FIRST)
- Georg Zeller (FML & MPI)

More Information

- `http://www.fml.mpg.de`
- Slides are available online

Thank you!

Acknowledgments

Tiling Arrays

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

Short Read Alignments

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

More Information

- `http://www.fml.mpg.de`
- Slides are available online

Gene Finding

- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Alex Zien (FML & FIRST)
- Georg Zeller (FML & MPI)

Thank you!

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. Submitted to the CSHL Genome Informatics Meeting, May 2008.

A. Ben-Hur, C.S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 2008. under revision.

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. Optimal spliced alignments of short sequence reads. In *Bioinformatics*. Oxford University Press, 2008. URL `http://www.fml.tuebingen.mpg.de/raetsch/projects/qpalma`. in press.

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, July 2008.

L. David et al. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA*, 103:5320–5325, 2006.

References II

B.J. Frey, Q.D. Morris, and T.R. Hughes. Genrate: A generative model that reveals novel transcripts in genome-tiling microarray data. *Journal of Computational Biology*, 13(2): 200–214, 2006.

E. Georgii, S. Dietmann, T. Uno, P. Pagel, and K. Tsuda. Enumeration of condition-dependent dense modules in protein interaction networks. submitted, May 2008.

W. Huber, J. Toedling, and L. M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(6):1963–1970, 2006.

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.

S. Laubinger, G. Zeller, SR Henz, T Sachsenberg, CK Widmer, N Naouar, M Vuylsteke, B Schölkopf, G Rätsch, and D. Weigel. At-tax: a whole genome tiling array resource for developmental expression analysis and transcript identification in arabidopsis thaliana. *Genome Biology*, 9(7):R112, July 2008b.

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.

References III

G Rätsch and S Sonnenburg. Large scale hidden semi-markov svms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 1161–1168, Cambridge, MA, 2007. MIT Press. URL `http://www.fml.tuebingen.mpg.de/raetsch/projects/HSMSVM`.

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

G. Rätsch, S. Sonnenburg, and C. Schäfer. Learning interpretable svms for biological sequence classification. *BMC Bioinformatics*, 7(Suppl 1):S9, February 2006.

G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.R. Müller, R. Sommer, and B. Schölkopf. Improving the caenorhabditis elegans genome annotation using machine learning. *PLoS Comput Biol*, 3(2):e20, 2007.

T.E. Royce, J.S. Rozowsky, and M.B. Gerstein. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, 23(8):988–997, 2007. doi: 10.1093/bioinformatics/btm052. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/8/988`.

S.J. Schultheiss, W. Busch, J.U. Lohmann, O. Kohlbacher, and G. Rätsch. Kirmes: Kernel-based identification of regulatory modules in euchromatic sequences. In *Proceedings of the German Conference on Bioinformatics*. GI, Springer Verlag, 2008. URL `http://www.fml.tuebingen.mpg.de/raetsch/projects/kirmes`. accepted.

References IV

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: A novel discriminative gene finding system. fortcoming, July 2008.

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.

S. Sonnenburg, G. Rätsch, and K. Rieck. *Large Scale Kernel Machines*, chapter Large Scale Learning with String Kernels. MIT Press, 2007.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–480, 2006.

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000. URL `http://dx.doi.org/10.1038/35048692`.

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.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, 2008b.

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.