Re-Annotating The Arabidopsis Transcriptome With Tiling Arrays

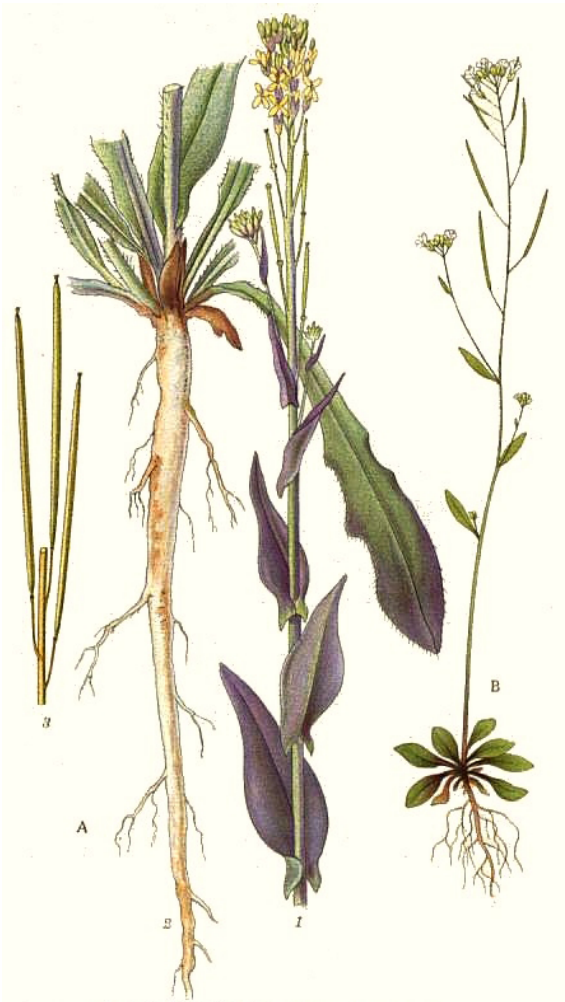
Georg Zeller, Johannes Eichner, Sascha Laubinger, Timo Sachsenberg, Detlef Weigel & Gunnar Rätsch

Friedrich Miescher Laboratory of the Max Planck Society
and
Max Planck Institute for Developmental Biology,
Dept. for Molecular Biology
Tübingen, Germany

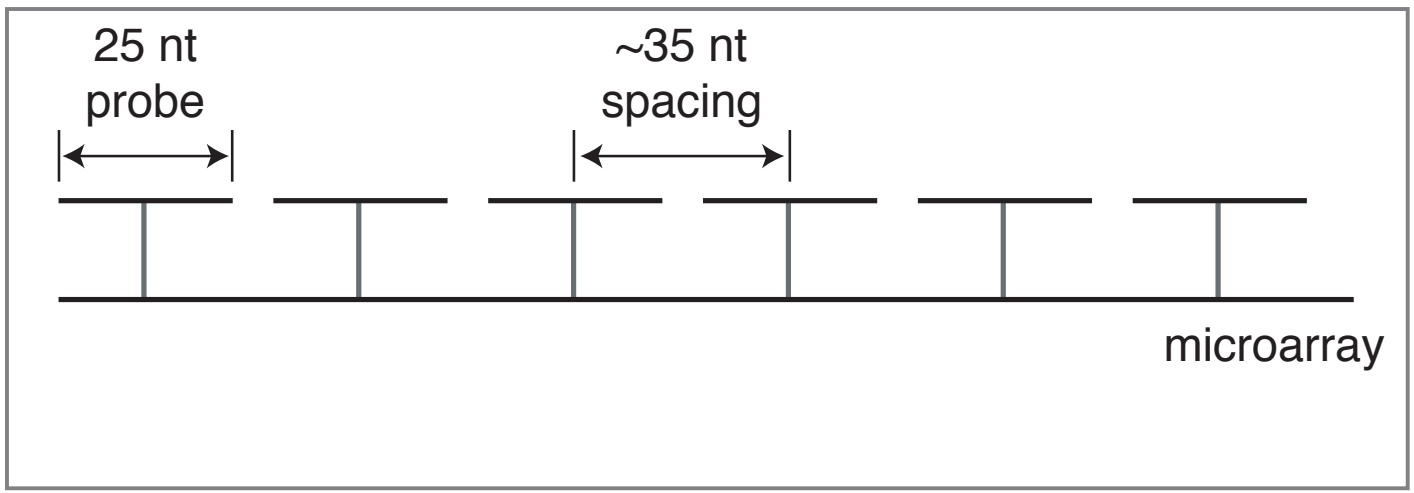
SCS4 in Toronto, July, 2008

Whole-Transcriptome Analyses

- Detect alternative splicing:
  skipped exons and retained introns of known genes
- Identify new transcripts
  absent from annotation and EST-data
- Tiling array data for 11 *Arabidopsis* tissues
  additionally stress treatments, pathogen infections, mutants, ...

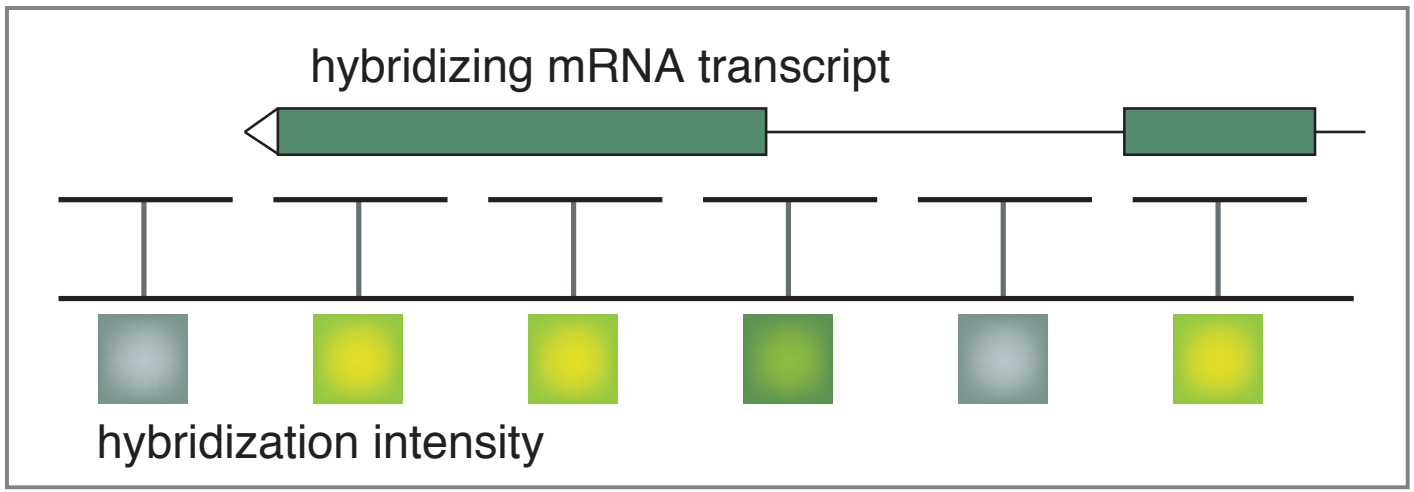


Tiling Arrays for Transcriptome Analysis



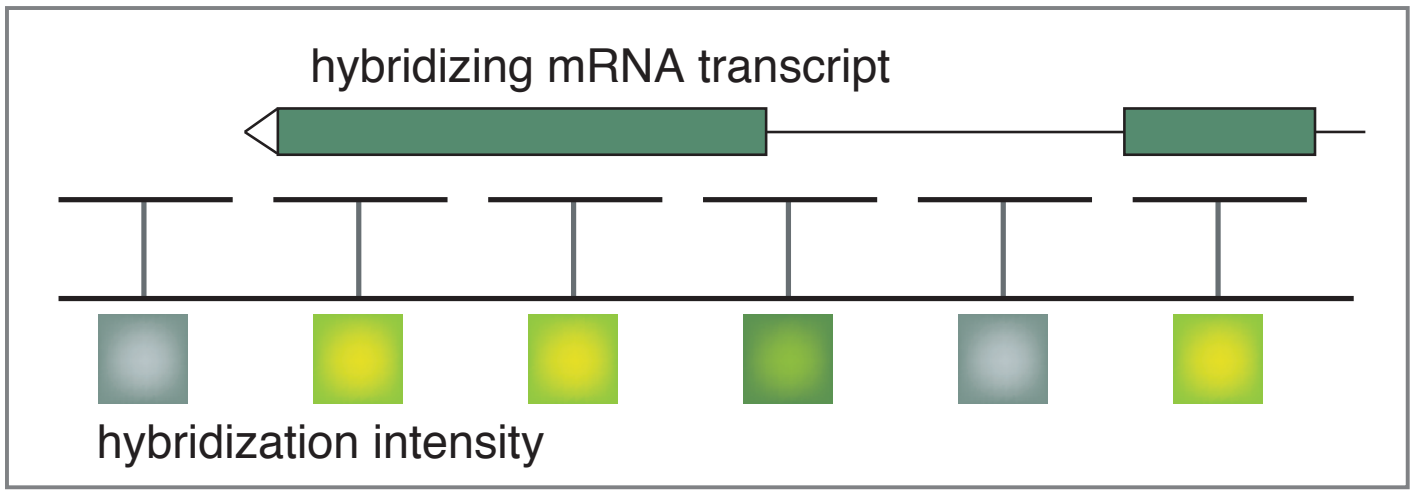
- Whole-genome quantitative measurements ($>$ 3 million features)
- Cost-effective
  $\Rightarrow$ replicates affordable, many tissues / mutants / conditions
- Unbiased
  $\Rightarrow$ does not rely on annotations or known cDNAs

Tiling Arrays for Transcriptome Analysis



- Whole-genome quantitative measurements ($>$ 3 million features)
- Cost-effective
  $\Rightarrow$ replicates affordable, many tissues / mutants / conditions
- Unbiased
  $\Rightarrow$ does not rely on annotations or known cDNAs

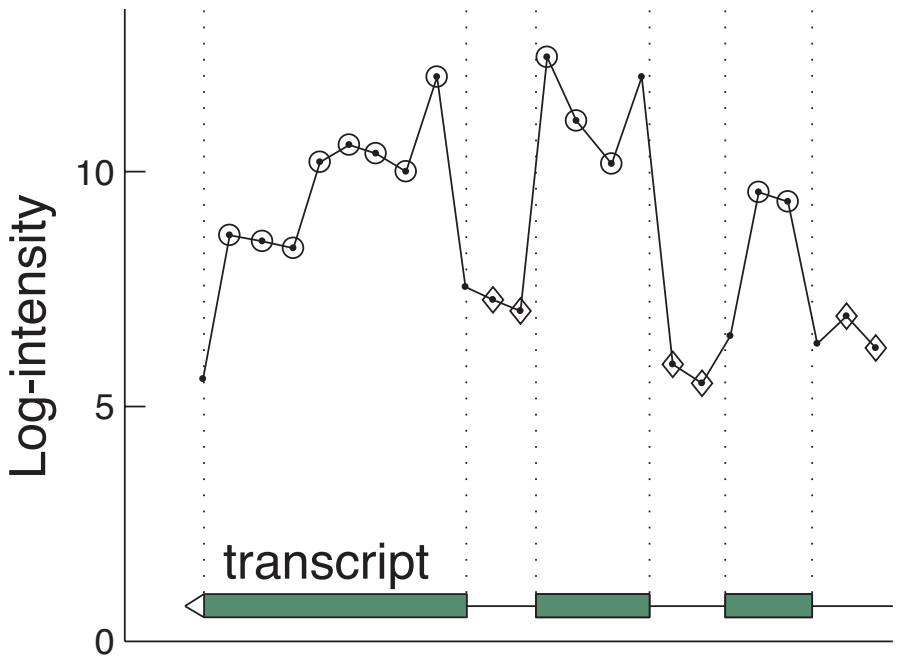
Tiling Arrays for Transcriptome Analysis



- Whole-genome quantitative measurements ($>$ 3 million features)
- Cost-effective
  $\Rightarrow$ replicates affordable, many tissues / mutants / conditions
- Unbiased
  $\Rightarrow$ does not rely on annotations or known cDNAs

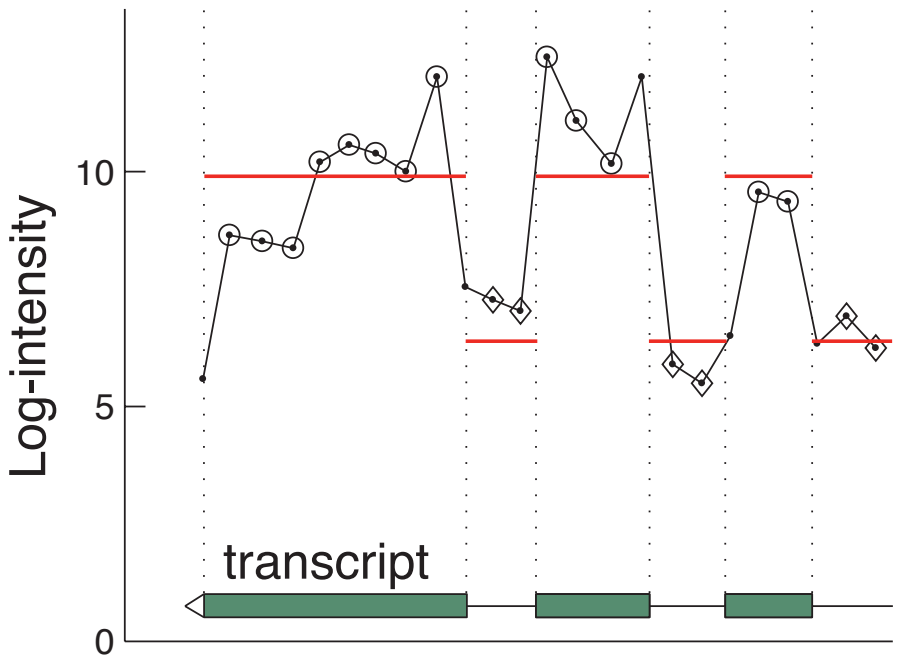
Transcript Normalization

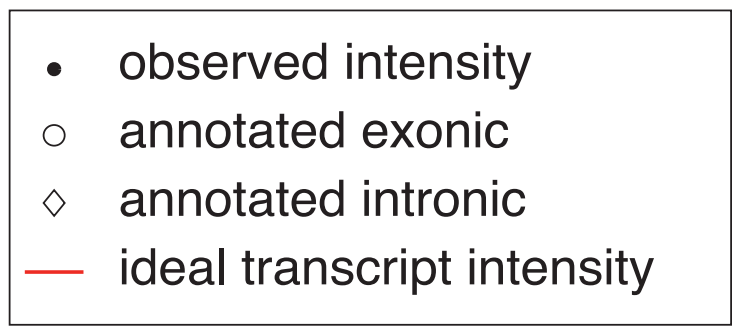
- Signal intensity are noisy measurements, mainly due to systematic probe sequence effects
- Model sequence effects for normalization using quantilized linear regression



Transcript Normalization

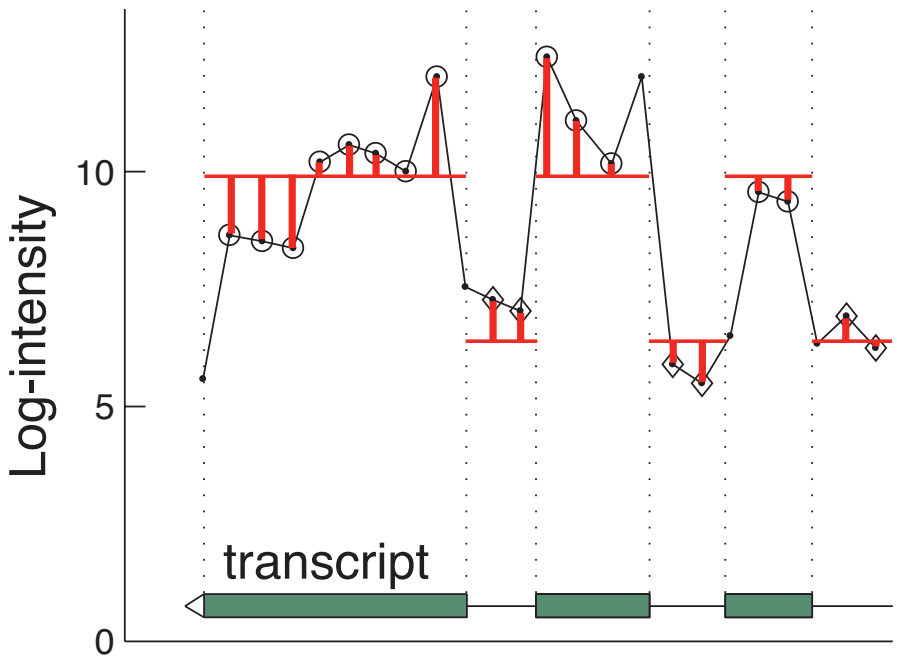
- Signal intensity are noisy measurements, mainly due to systematic probe sequence effects
- Model sequence effects for normalization using quantilized linear regression





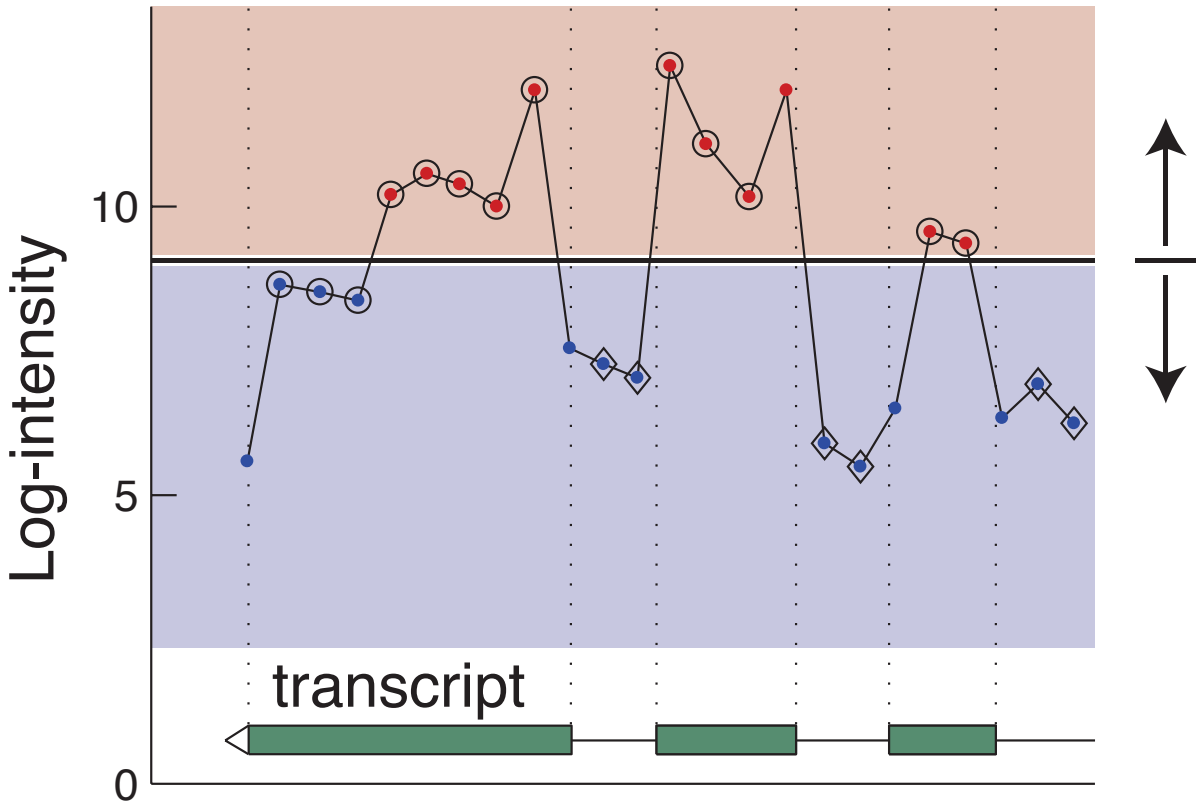
Transcript Normalization

- Signal intensity are noisy measurements,
  mainly due to systematic probe sequence effects
- Model sequence effects for normalization
  using quantilized linear regression



- observed intensity
- annotated exonic
- annotated intronic
- ideal transcript intensity
- fold difference between observed and transcript intensity

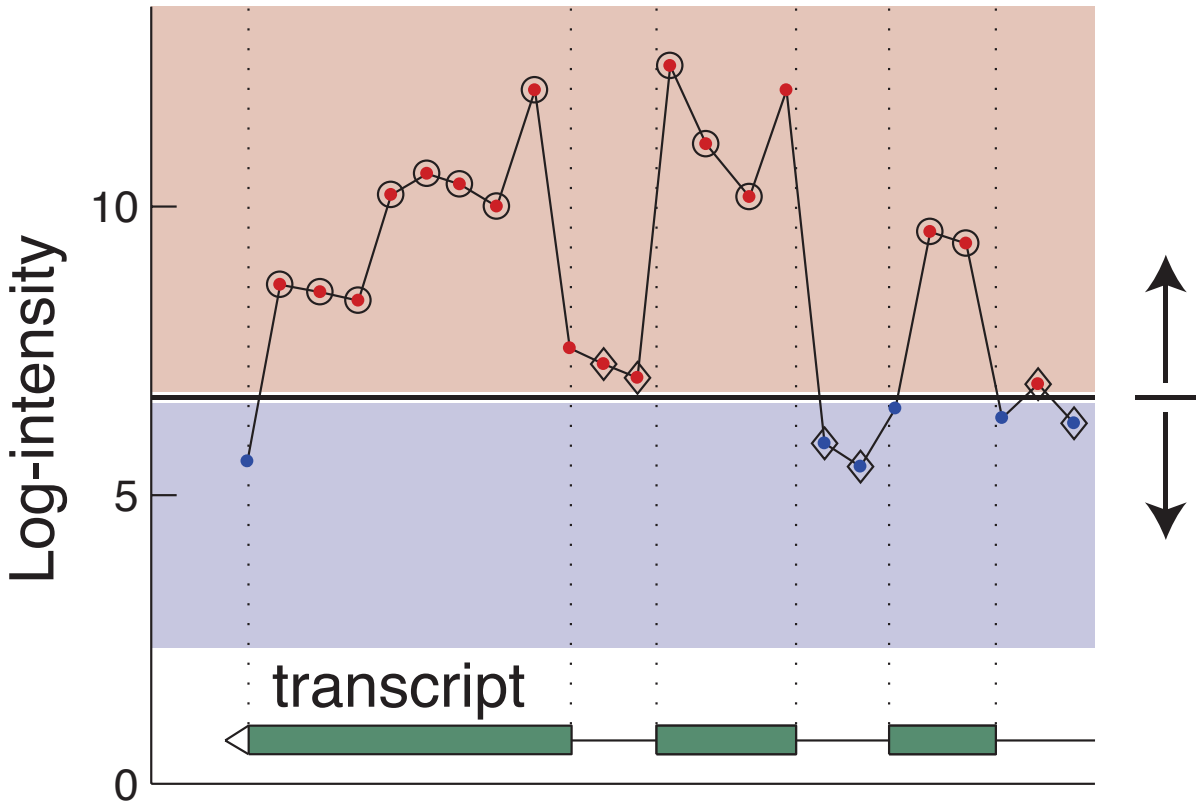
Exon/Background Probe Separation



Global thresholding of probe intensities

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

Exon/Background Probe Separation

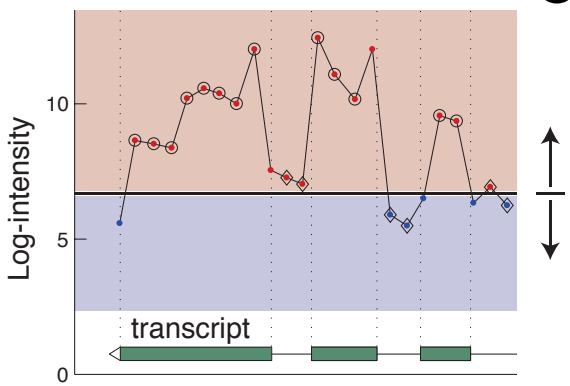


Global thresholding of probe intensities

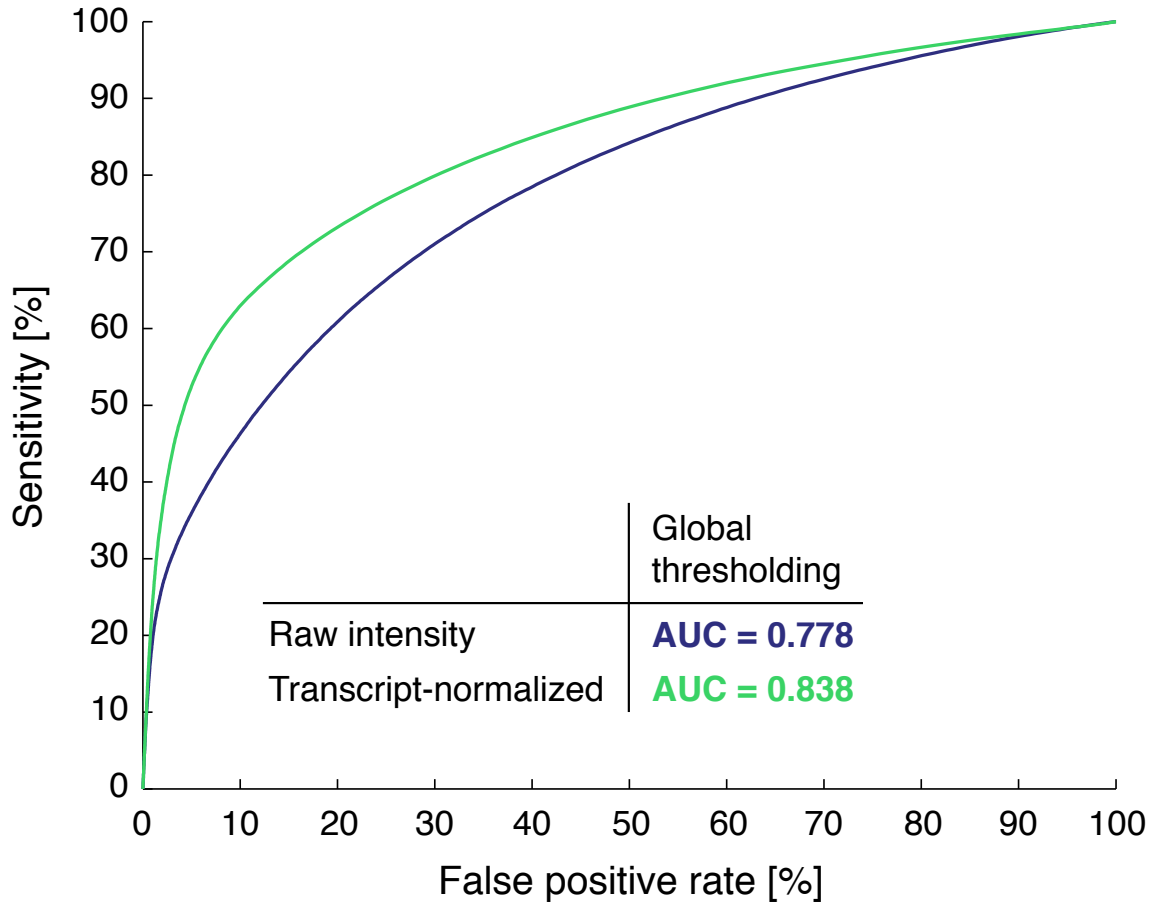
$\Rightarrow$ bi-partition into exonic and intronic/intergenic probes

Exon/Background Probe Separation

- Global thresholding



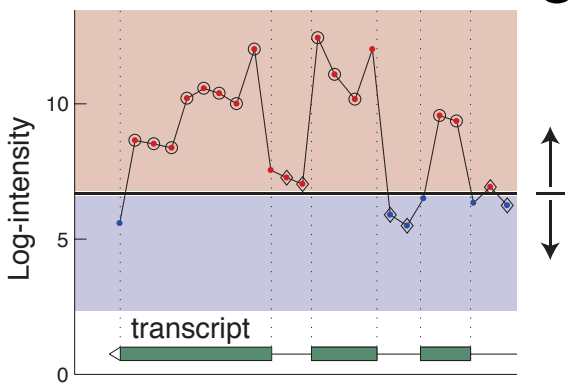
- Support Vector Machines (SVMs) to discriminate exons from introns



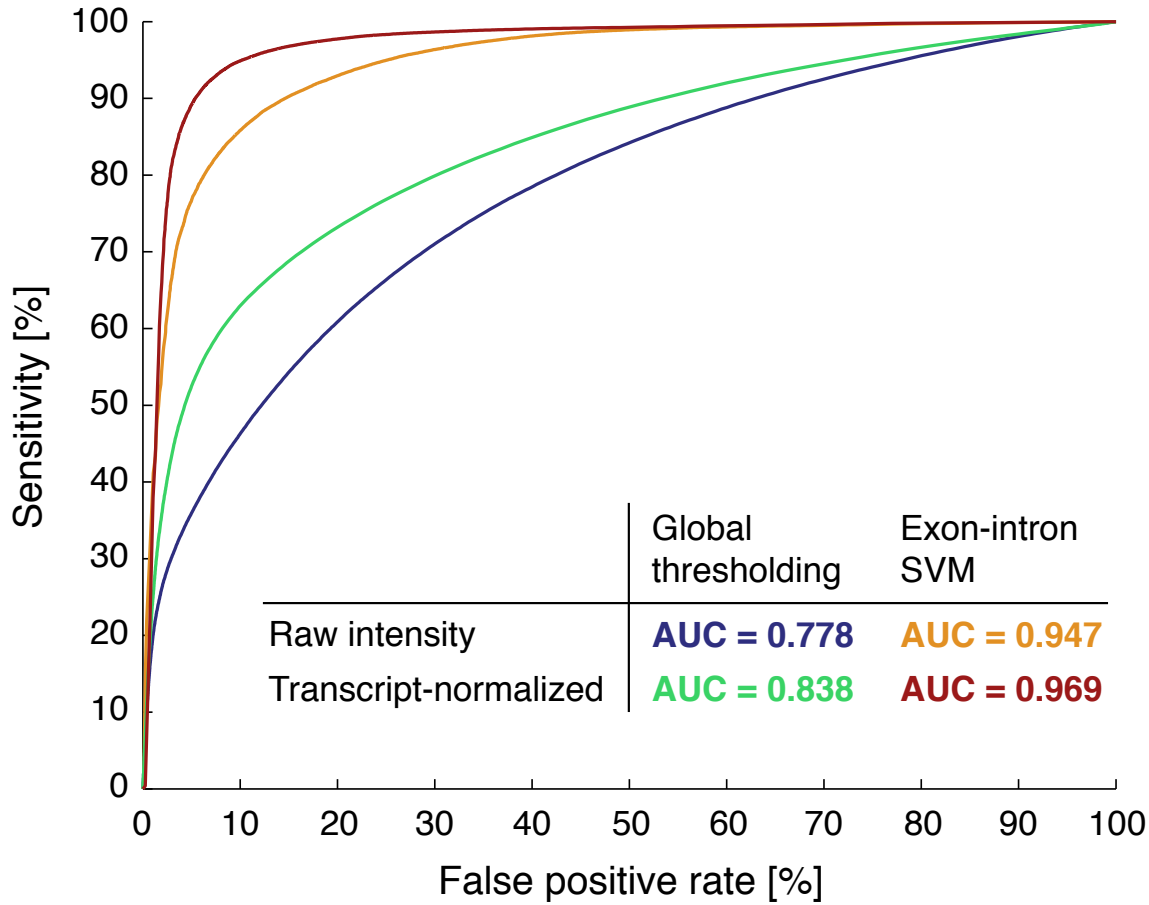
Known exon/intron segments typically consist of several probes
$\Rightarrow$ drastically improved separation

Exon/Background Probe Separation

- Global thresholding

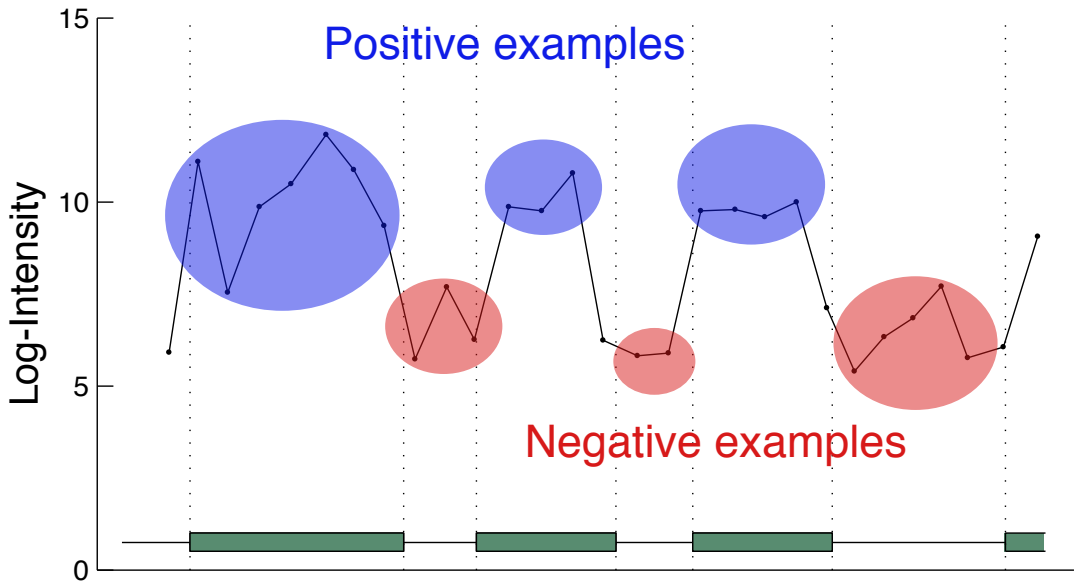


- Support Vector Machines (SVMs) to discriminate exons from introns



Known exon/intron segments typically consist of several probes
$\Rightarrow$ drastically improved separation

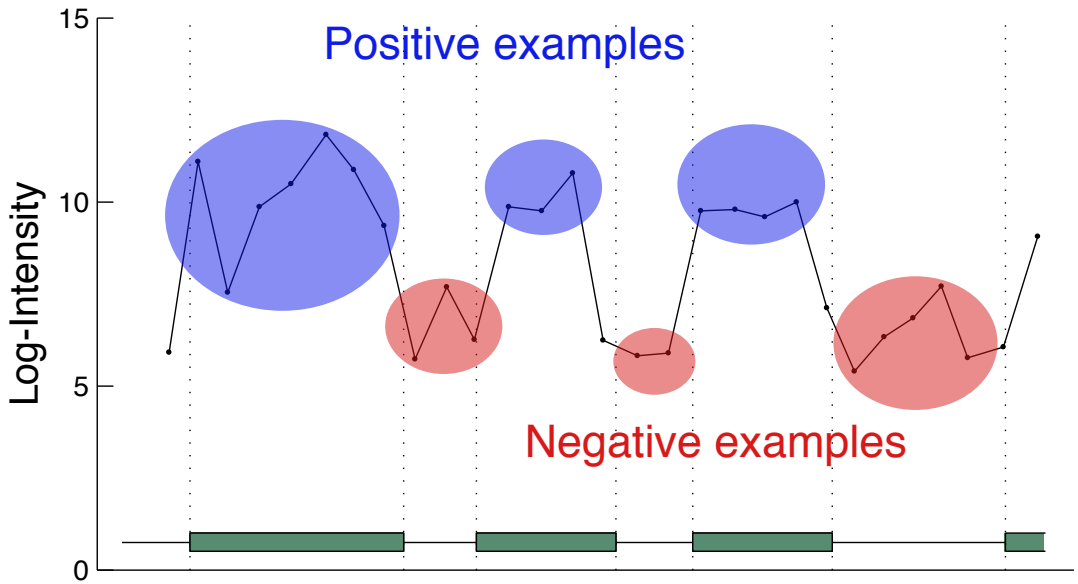
Exon-Intron Classification
using linear Support Vector Machines (SVMs)



- Input: Features derived from signal intensities for a single exon (intron) segment from a single tissue.
- Output: Real-valued score which is transformed in to an exon (inclusion) probability.

How to combine these predictions across several tissues?

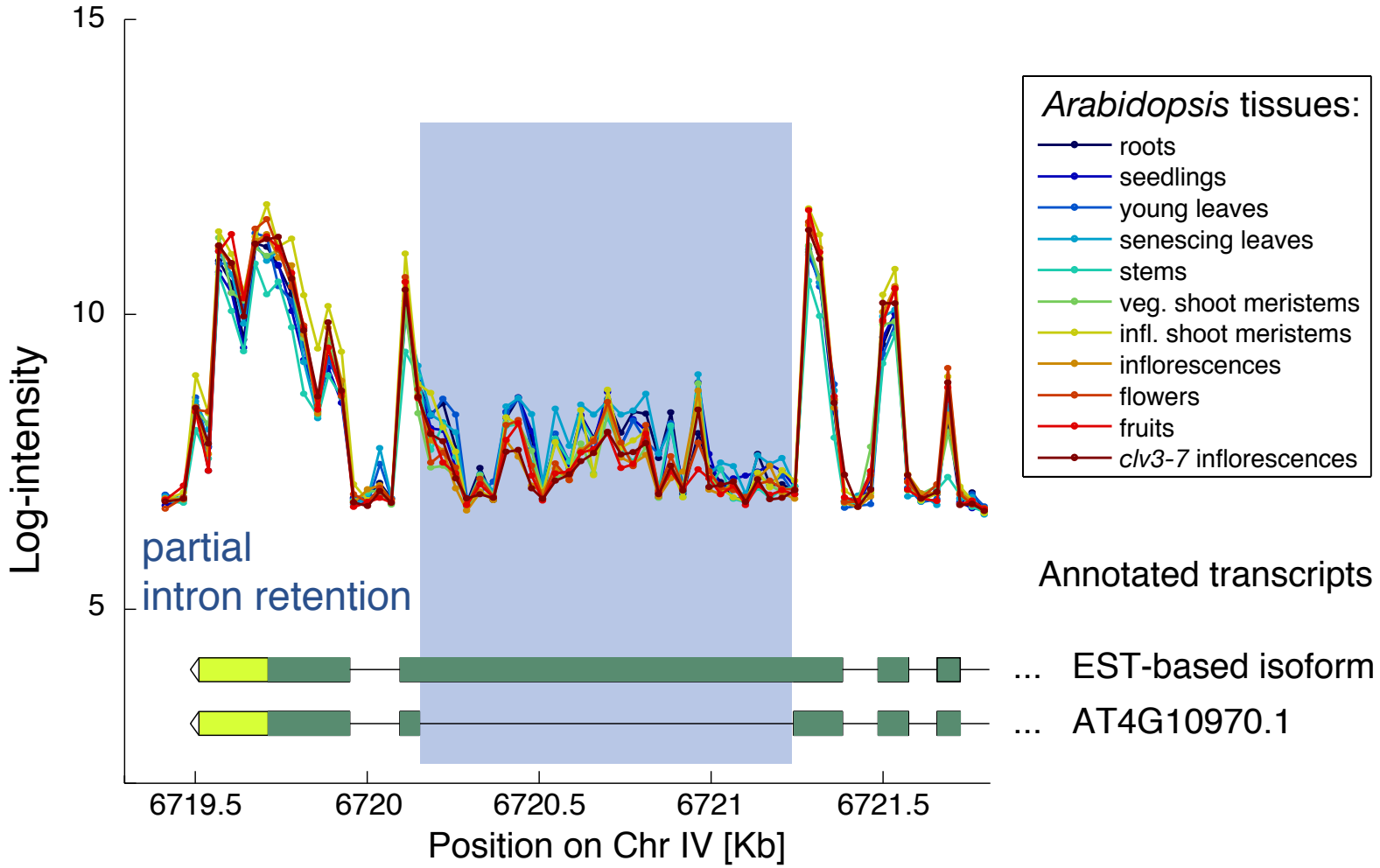
Exon-Intron Classification
using linear Support Vector Machines (SVMs)



- Input: Features derived from signal intensities for a single exon (intron) segment from a single tissue.
- Output: Real-valued score which is transformed in to an exon (inclusion) probability.

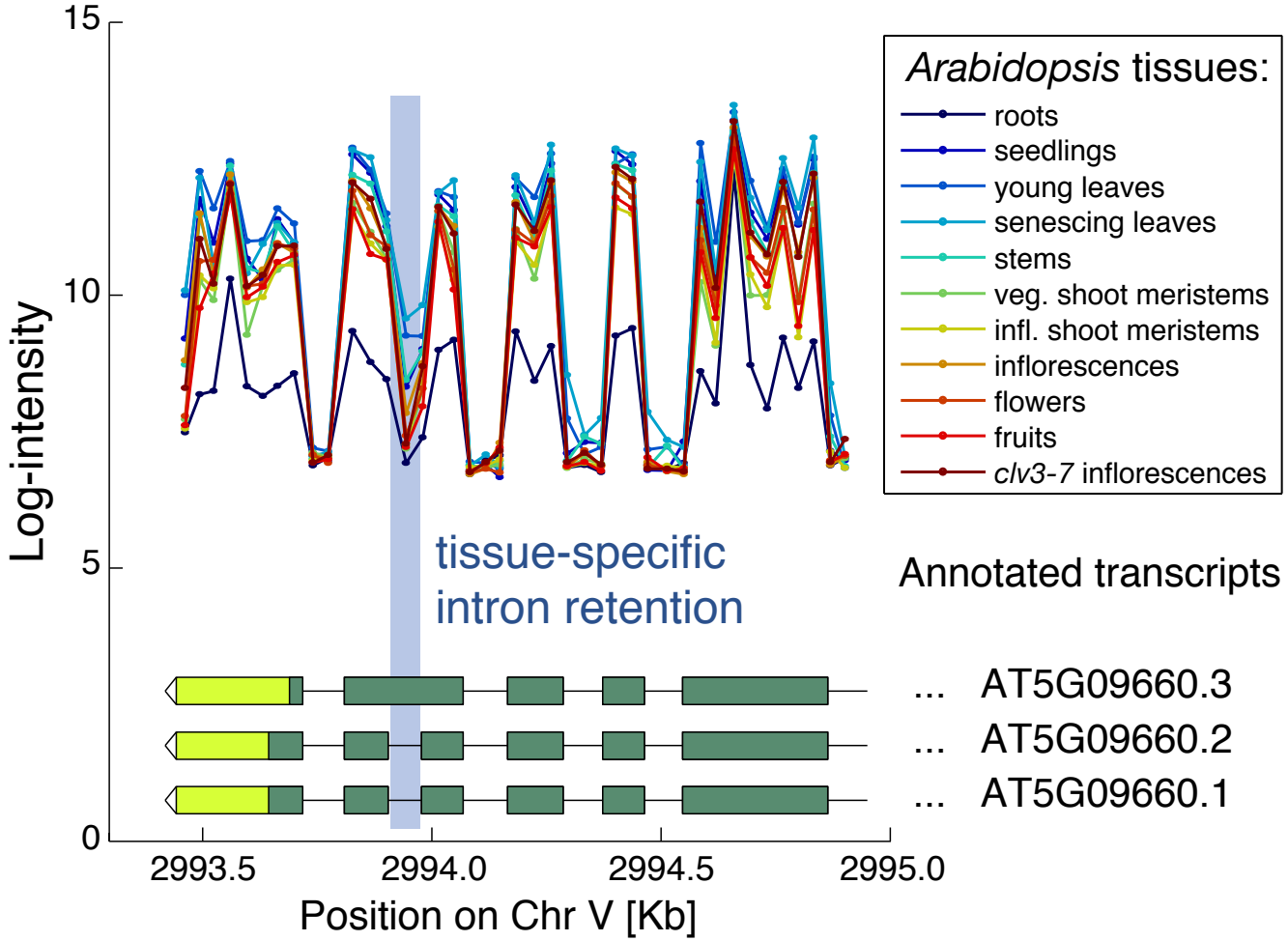
How to combine these predictions across several tissues?

Alternatively Spliced Genes



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

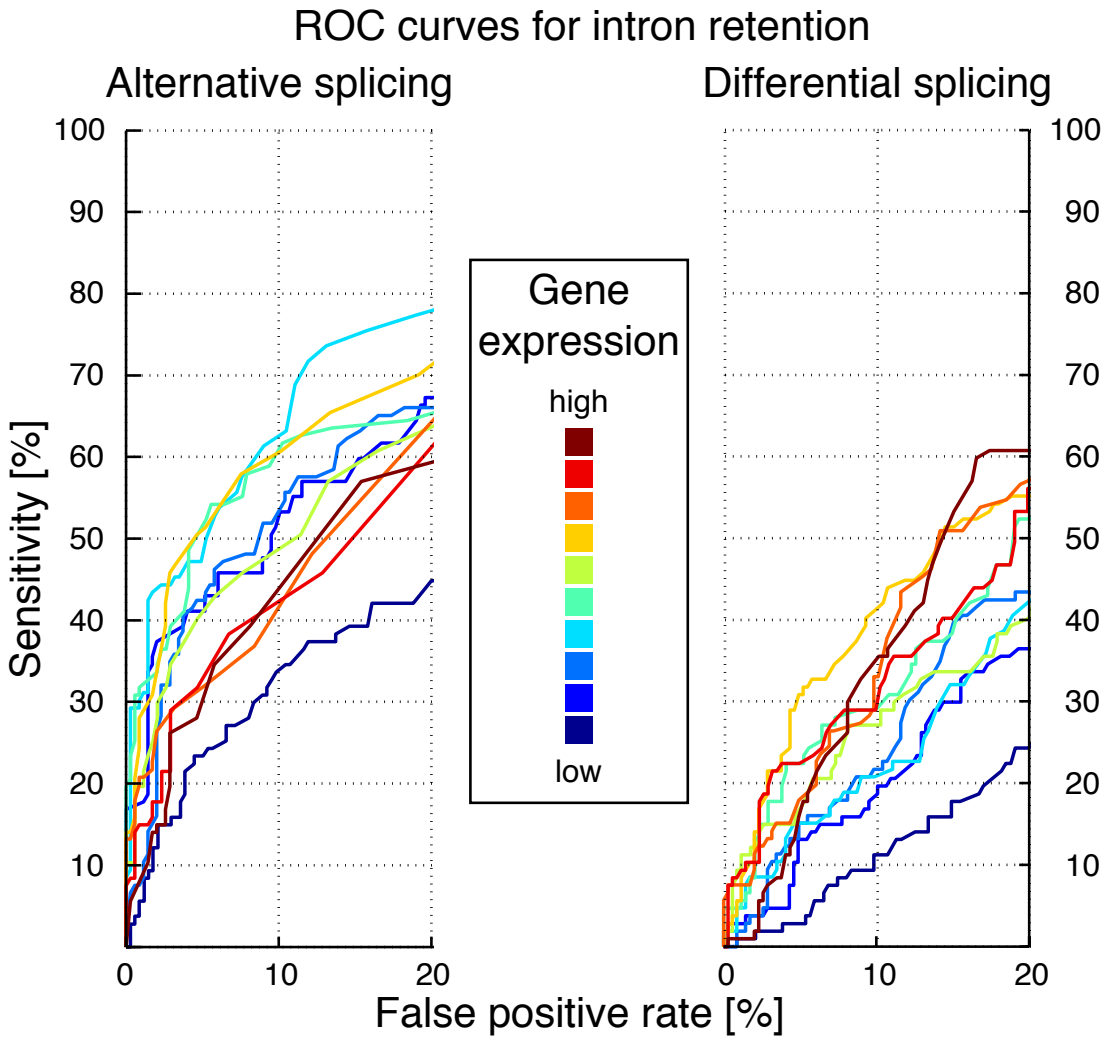
Differentially Spliced Genes



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

Alternative vs. Differential Splicing

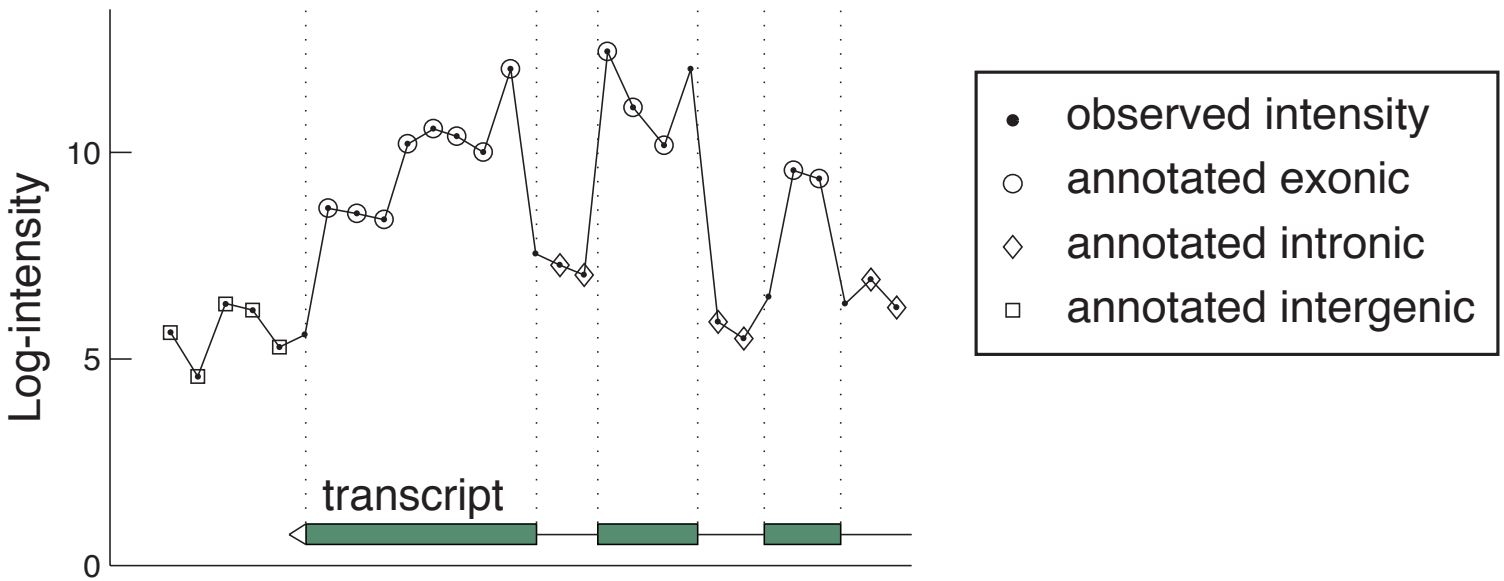
A Comparison with EST/cDNA-based Information



Evaluating tiling array data from 11 *Arabidopsis* tissues.

Tiling Array Segmentation

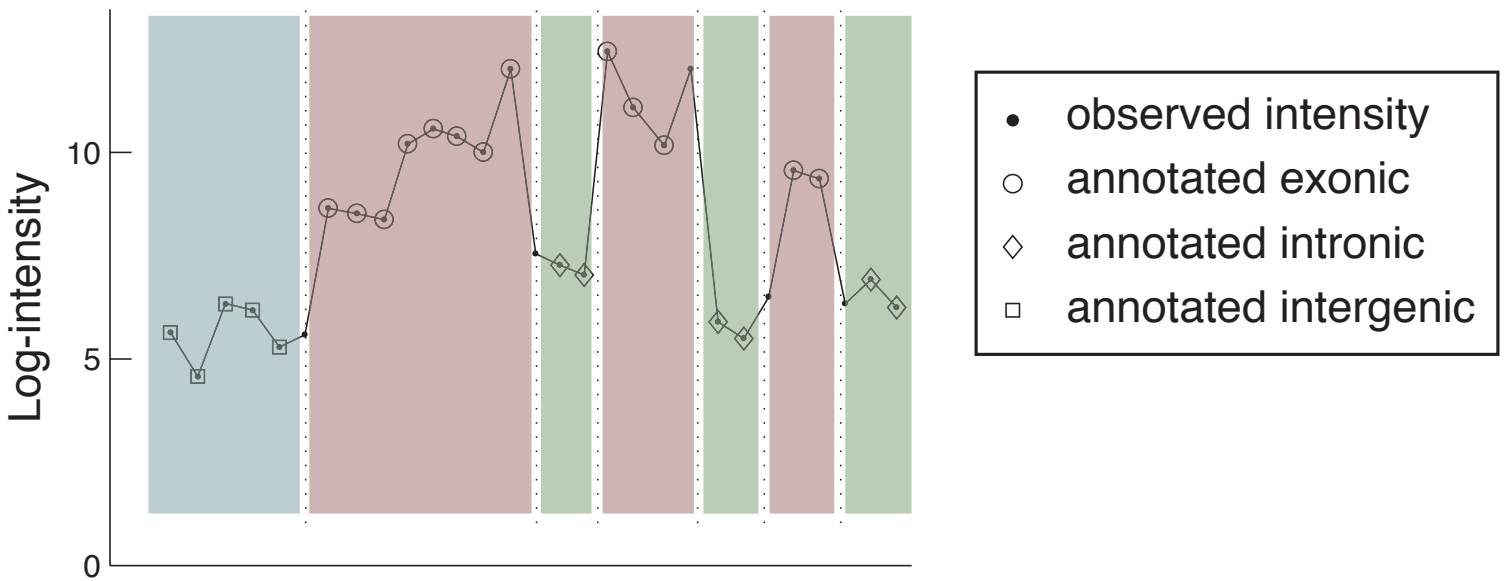
Goal: Characterize each probe as either intergenic, exonic or intronic



margin-based segmentation of tiling array data (**mSTAD**)

Tiling Array Segmentation

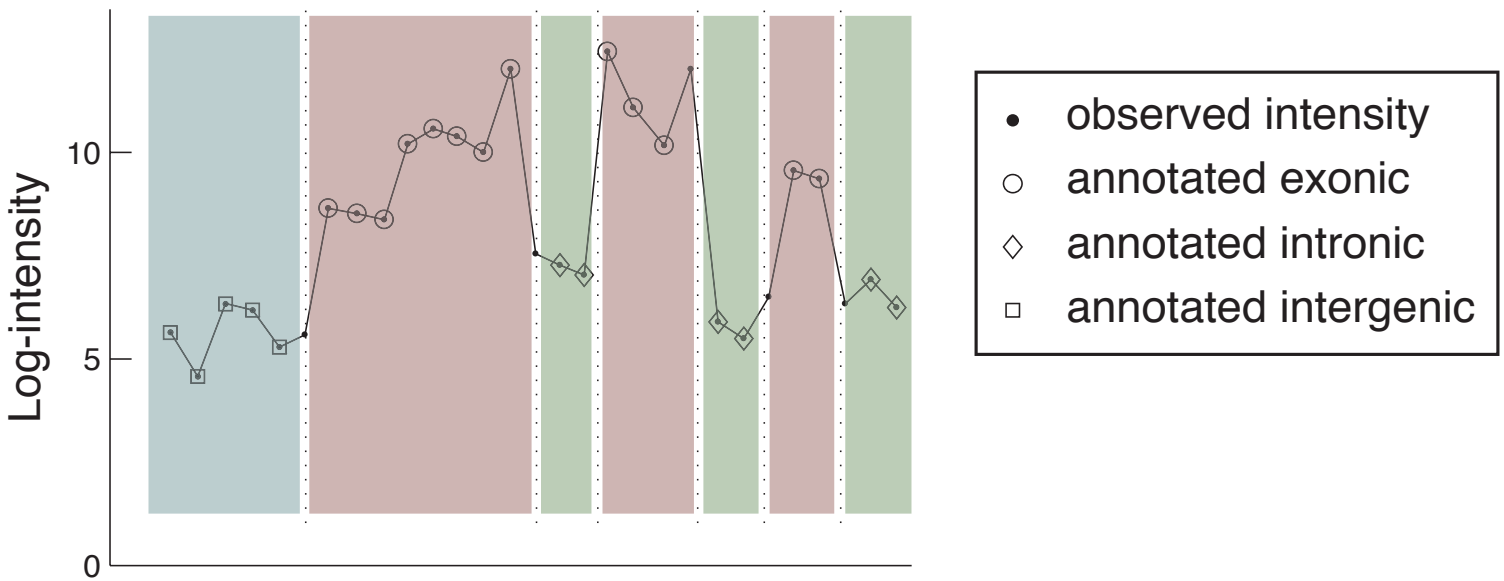
Goal: Characterize each probe as either intergenic, exonic or intronic



margin-based segmentation of tiling array data (**mSTAD**)

Tiling Array Segmentation

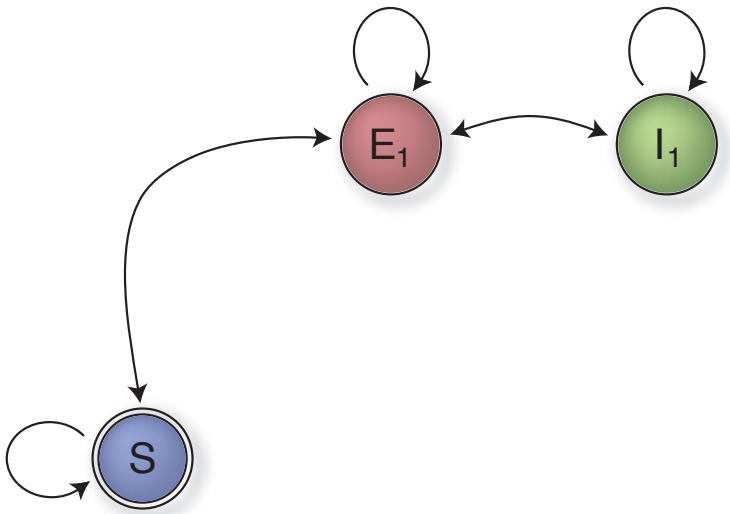
Goal: Characterize each probe as either intergenic, exonic or intronic



margin-based segmentation of tiling array data (**mSTAD**)

Tiling Array Segmentation

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

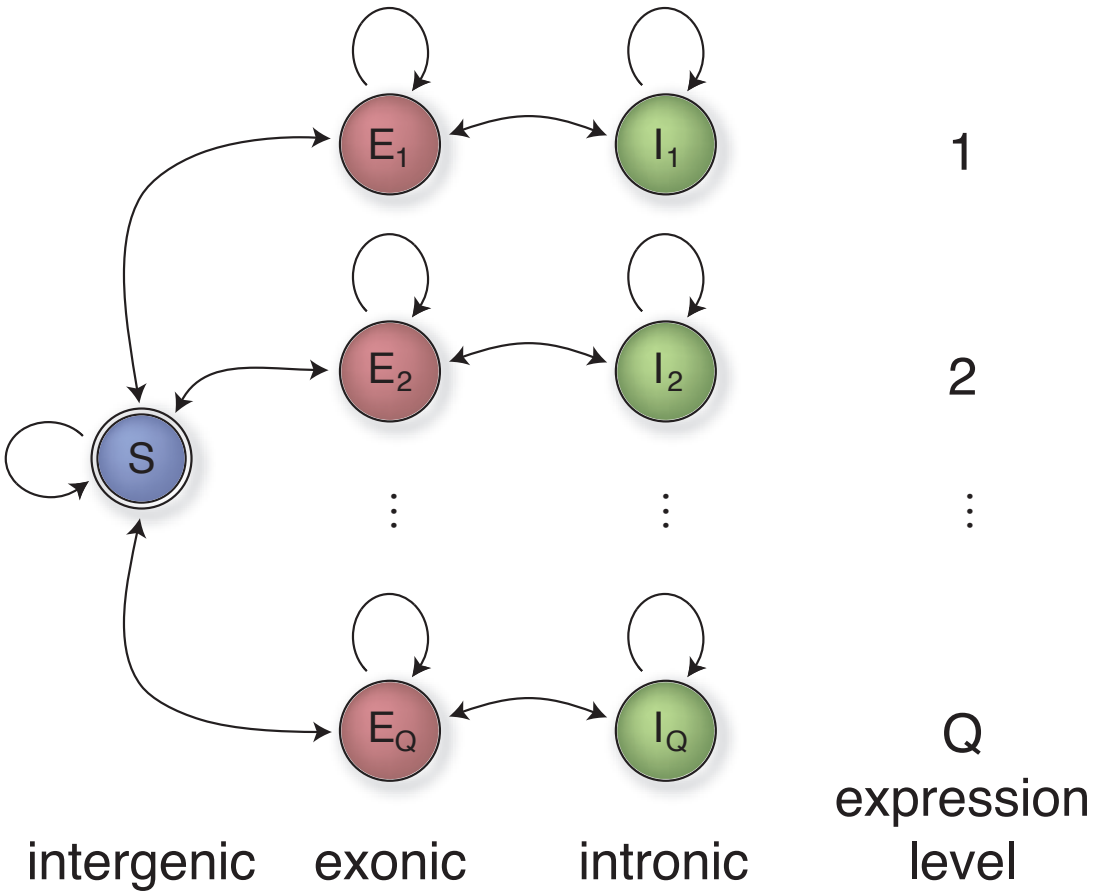


- Learn to associate a state with each probe given its hybridization signal and local context
- $Q = 10$ discrete expression levels

intergenic exonic intronic

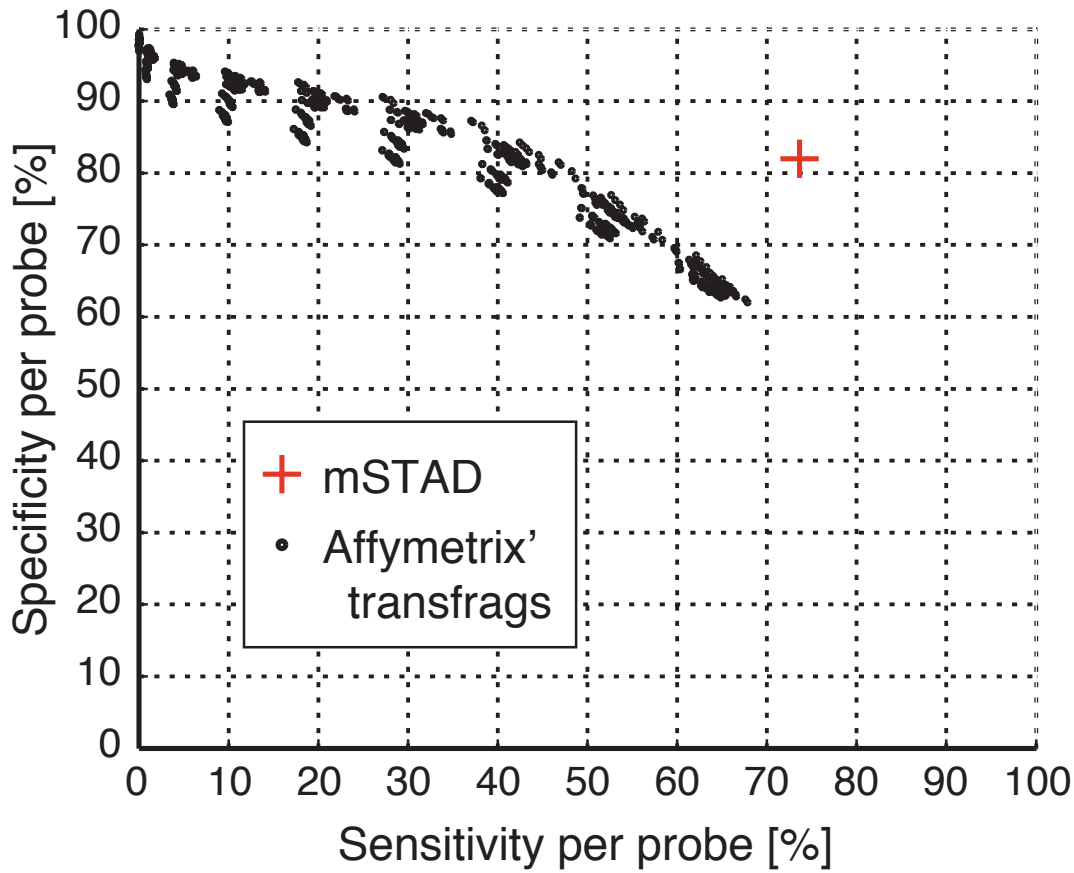
Tiling Array Segmentation

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

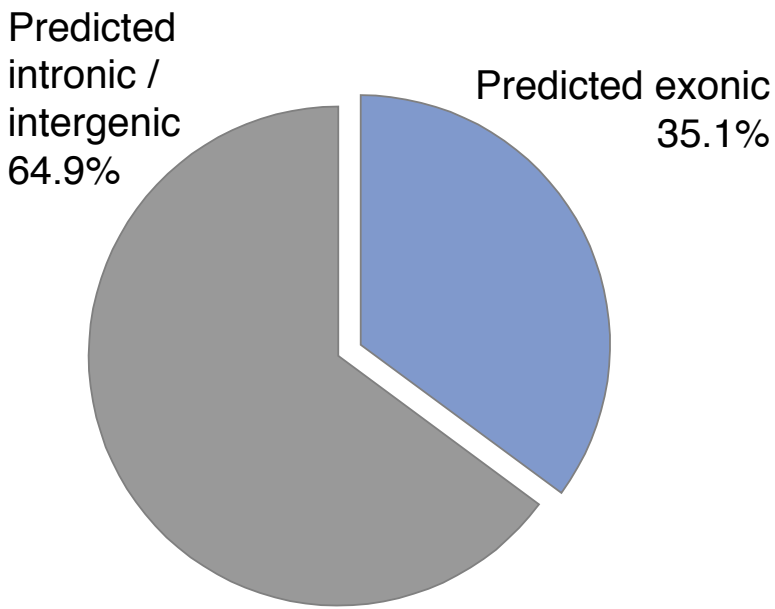


- Learn to associate a state with each probe given its hybridization signal and local context
- $Q = 10$ discrete expression levels

Comparison to Affymetrix's Transfrags

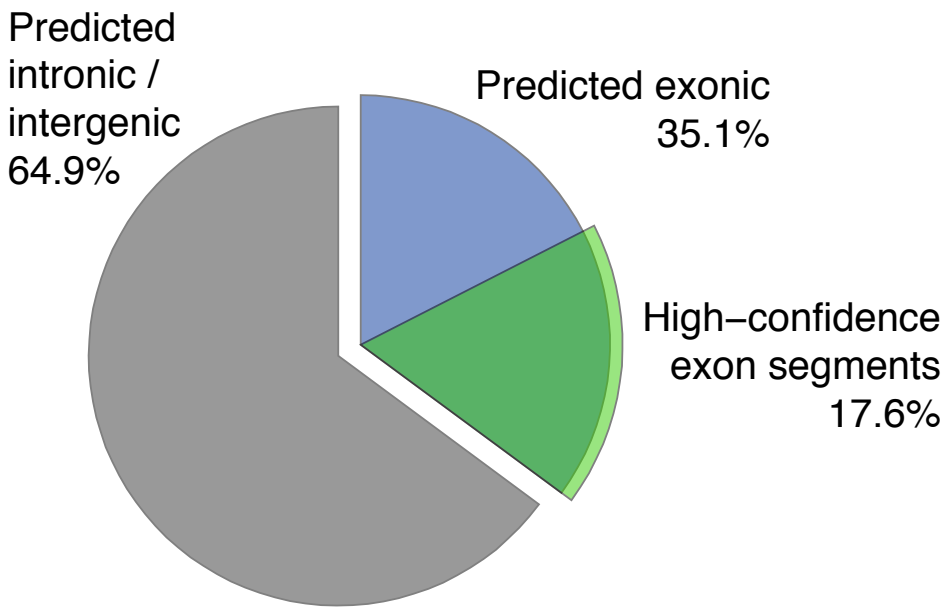


Novel Transcripts



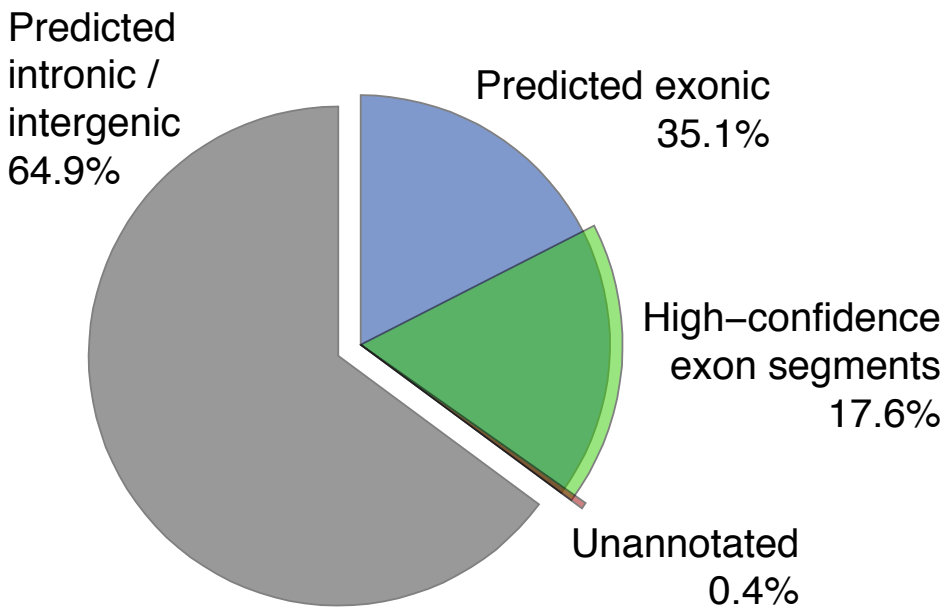
- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

Novel Transcripts



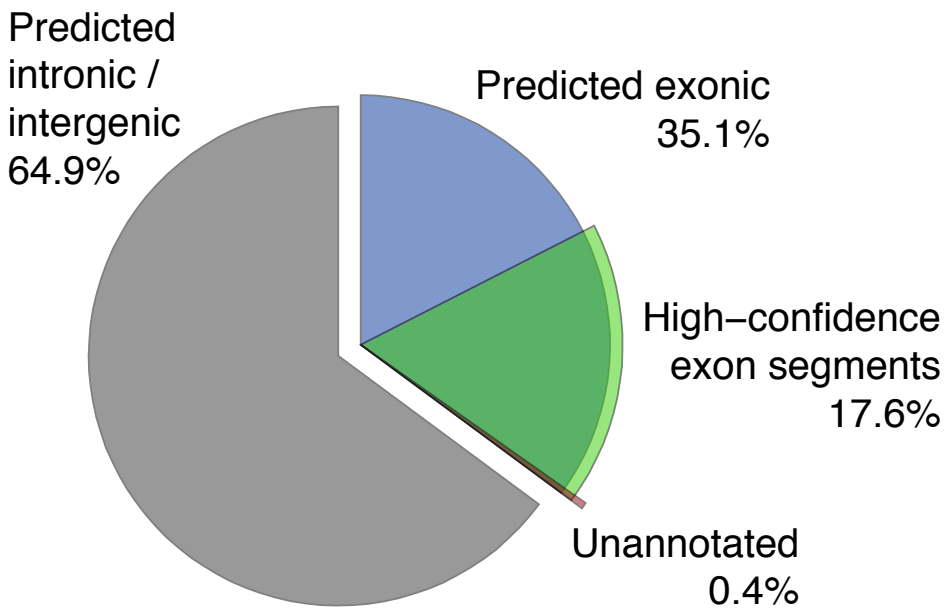
- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

Novel Transcripts



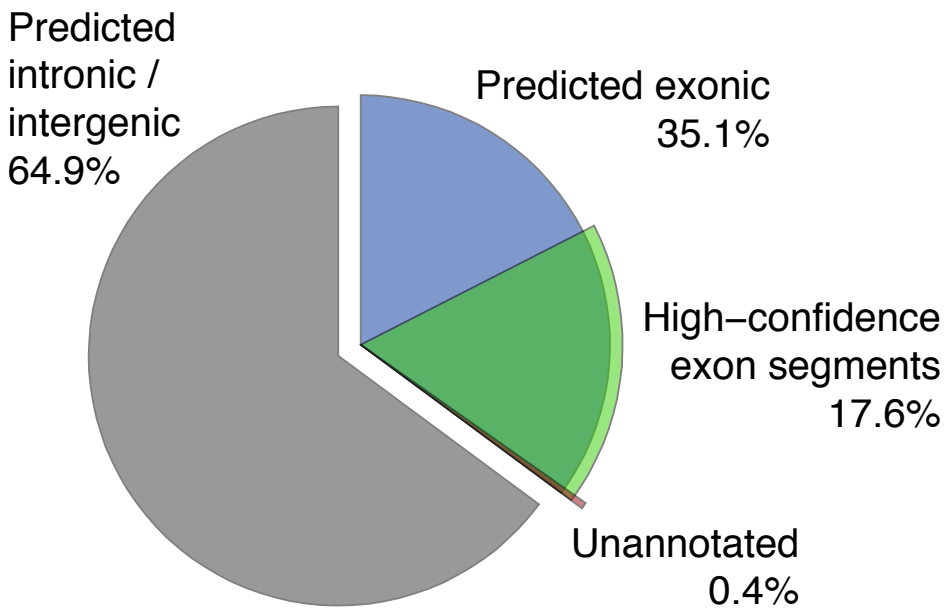
- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

Novel Transcripts



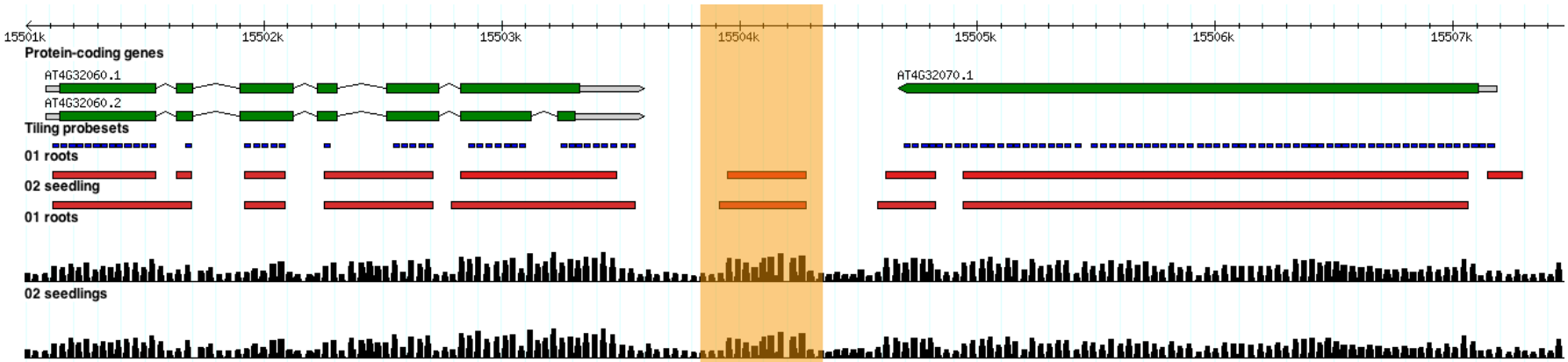
- Between 1,107 and 1,947 predicted high-confidence exons per sample (total length 242 to 406 kb) are absent from annotation and not covered by ESTs/cDNAs.
- 37 of 47 (>75%) RT-PCR validations successful.

Novel 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.

Novel 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.

Conclusions



- Identification of alternative and differential splicing
  - Good agreement with known ESTs and cDNAs
  - Adds tissue/condition information to splicing events
- Segmentation of tiling array data
  - Precisely identifies exons of annotated transcripts
  - Detects $>$ thousand new transcripts/exons of known genes
  - Small-scale RT-PCR validation suggests low false discovery rate
- Details, visualizations, data & code available from
  `www.fml.mpg.de/raetsch/projects/PSBtiling`
  `www.weigelworld.org/resources/microarray/at-tax`

Thank you

and hope to see you at the poster (also ISMB poster A17)!

Conclusions



- Identification of alternative and differential splicing
  - Good agreement with known ESTs and cDNAs
  - Adds tissue/condition information to splicing events
- Segmentation of tiling array data
  - Precisely identifies exons of annotated transcripts
  - Detects $>$ thousand new transcripts/exons of known genes
  - Small-scale RT-PCR validation suggests low false discovery rate
- Details, visualizations, data & code available from
  `www.fml.mpg.de/raetsch/projects/PSBtiling`
  `www.weigelworld.org/resources/microarray/at-tax`

Thank you

and hope to see you at the poster (also ISMB poster A17)!

Conclusions



- Identification of alternative and differential splicing
  - Good agreement with known ESTs and cDNAs
  - Adds tissue/condition information to splicing events
- Segmentation of tiling array data
  - Precisely identifies exons of annotated transcripts
  - Detects $>$ thousand new transcripts/exons of known genes
  - Small-scale RT-PCR validation suggests low false discovery rate
- Details, visualizations, data & code available from
  `www.fml.mpg.de/raetsch/projects/PSBtiling`
  `www.weigelworld.org/resources/microarray/at-tax`

Thank you

and hope to see you at the poster (also ISMB poster A17)!

Conclusions



- Identification of alternative and differential splicing
  - Good agreement with known ESTs and cDNAs
  - Adds tissue/condition information to splicing events
- Segmentation of tiling array data
  - Precisely identifies exons of annotated transcripts
  - Detects $>$ thousand new transcripts/exons of known genes
  - Small-scale RT-PCR validation suggests low false discovery rate
- Details, visualizations, data & code available from
  `www.fml.mpg.de/raetsch/projects/PSBtiling`
  `www.weigelworld.org/resources/microarray/at-tax`

Thank you

and hope to see you at the poster (also ISMB poster A17)!