Novel Methods for Computational Transcriptome Analysis

Gunnar Rätsch

Friedrich Miescher Laboratory
Max Planck Society
Tübingen, Germany

1^{st} Symposium Lifesciences in Tübingen
November 06, 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*

Transcriptome Analysis

Ideally, we would like to know the concentration of all RNAs and proteins in the cell.

Then we can (try to) understand:

- What are differences of active components between organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
   - Computational gene finding
2. Measure properties of transcriptome
   - Tiling arrays: alternative splicing and new genes
3. Combining predictions with measurements

   $\Rightarrow$ Accurate tissue specific transcriptome predictions

Transcriptome Analysis

Ideally, we would like to know the concentration of all RNAs and proteins in the cell.

Then we can (try to) understand:

- What are differences of active components between organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
   - Computational gene finding
2. Measure properties of transcriptome
   - Tiling arrays: alternative splicing and new genes
3. Combining predictions with measurements

   $\Rightarrow$ Accurate tissue specific transcriptome predictions

Transcriptome Analysis

Ideally, we would like to know the concentration of all RNAs and proteins in the cell.

Then we can (try to) understand:

- What are differences of active components between organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
   - Computational gene finding
2. Measure properties of transcriptome
   - Tiling arrays: alternative splicing and new genes
3. Combining predictions with measurements

   $\Rightarrow$ Accurate tissue specific transcriptome predictions

Transcriptome Analysis

Ideally, we would like to know the concentration of all RNAs and proteins in the cell.

Then we can (try to) understand:

- What are differences of active components between organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
   - Computational gene finding
2. Measure properties of transcriptome
   - Tiling arrays: alternative splicing and new genes
3. Combining predictions with measurements

   $\Rightarrow$ Accurate tissue specific transcriptome predictions

Transcriptome Analysis

Ideally, we would like to know the concentration of all RNAs and proteins in the cell.

Then we can (try to) understand:

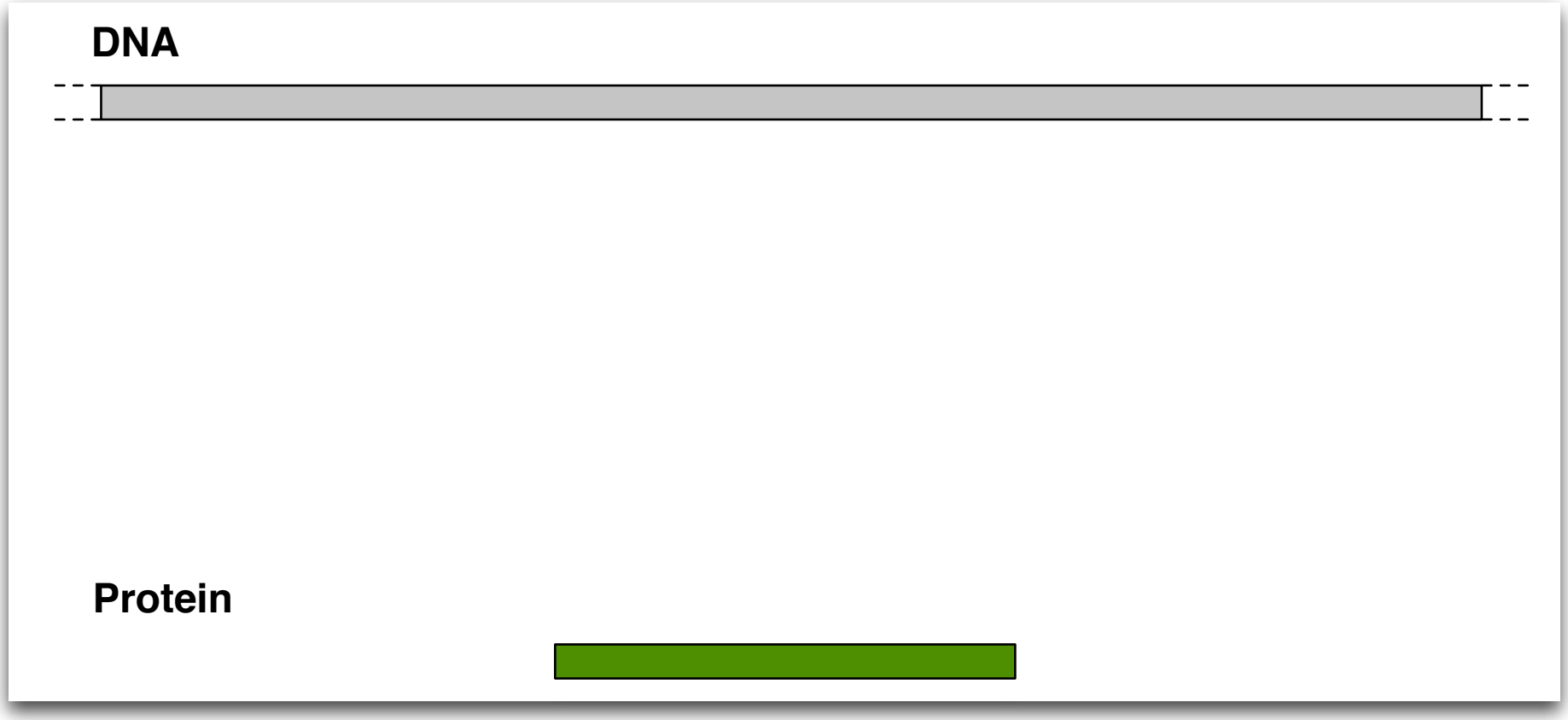
- What are differences of active components between organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
   - Computational gene finding
2. Measure properties of transcriptome
   - Tiling arrays: alternative splicing and new genes
3. Combining predictions with measurements

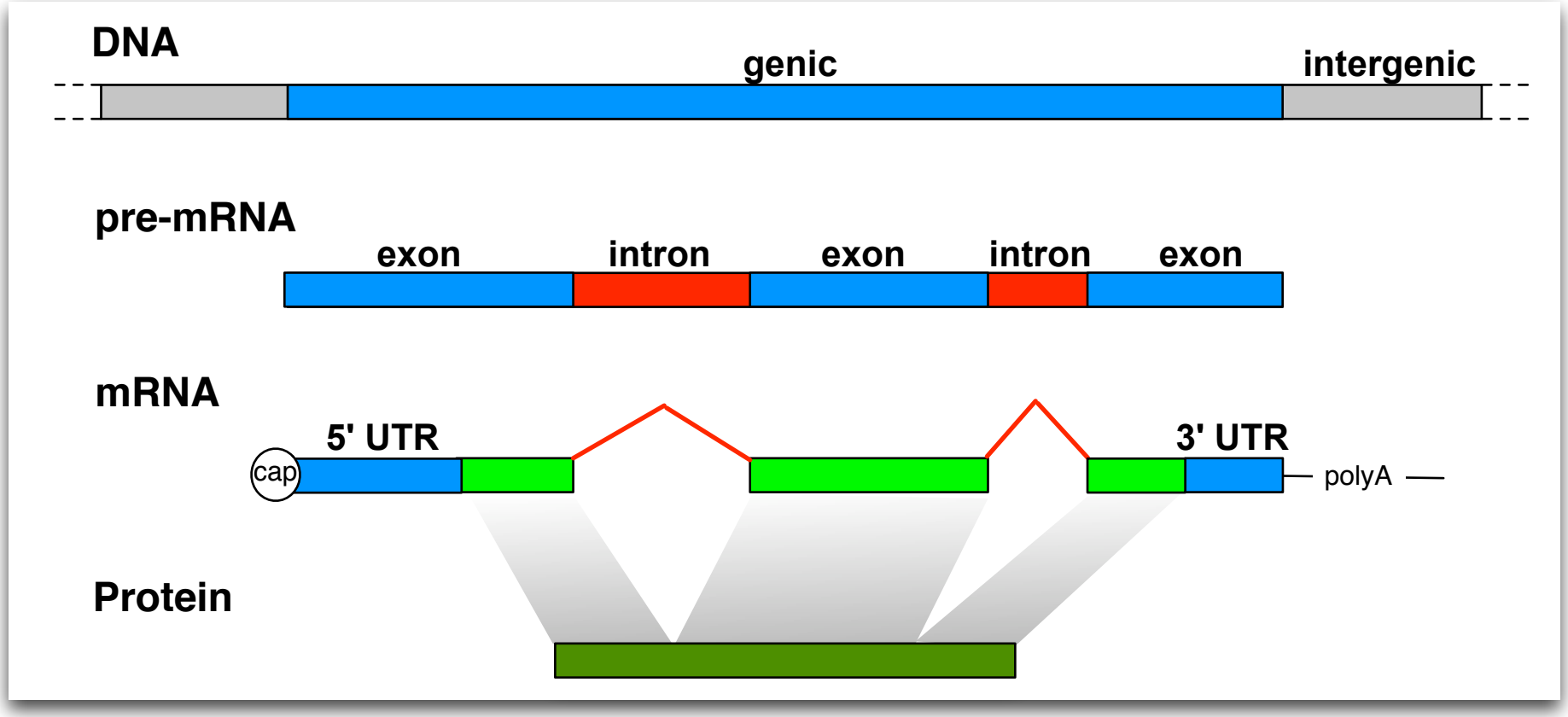
   $\Rightarrow$ Accurate tissue specific transcriptome predictions

Gene Finding



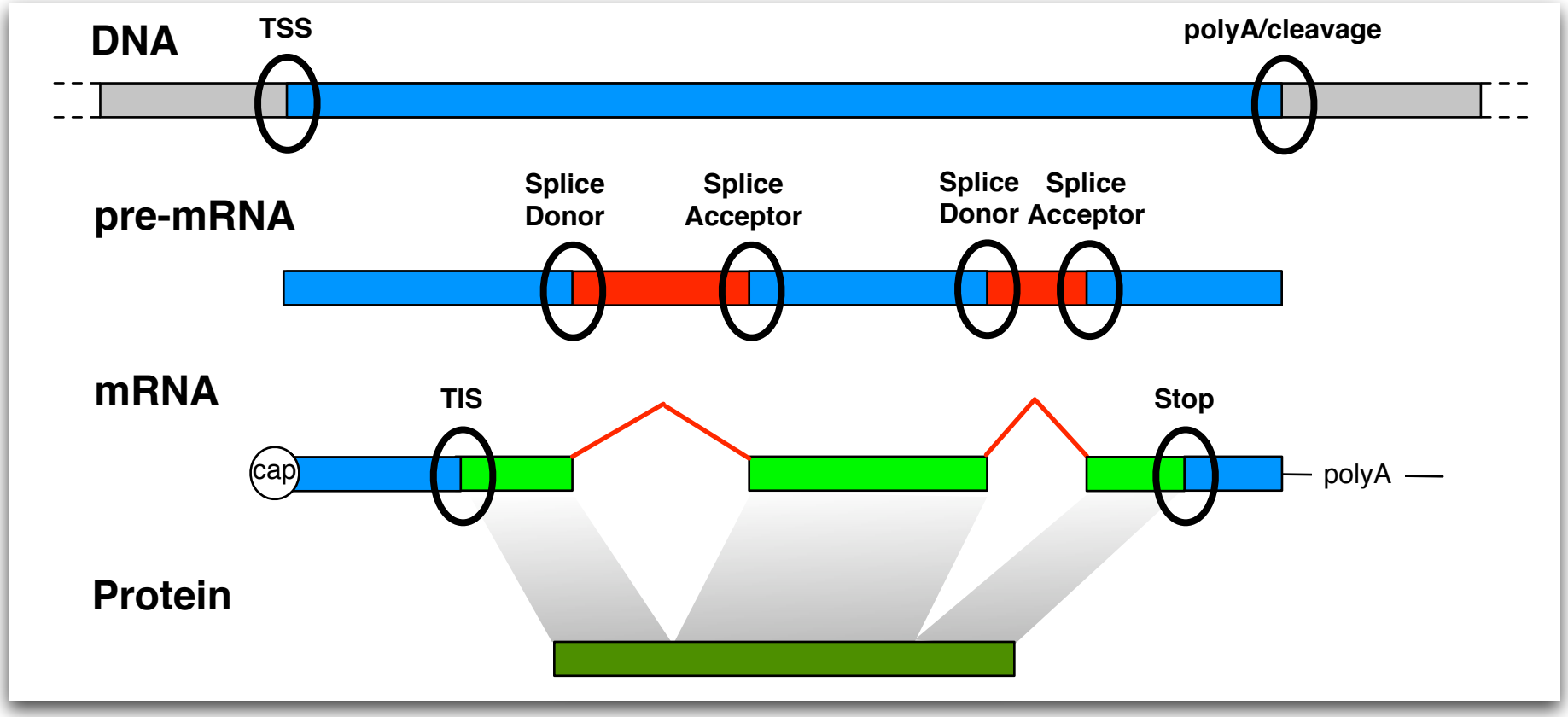
- Given a piece of DNA sequence of length L
- Predict proteins (or non-coding RNAs)

Gene Finding: Labeling the Genome



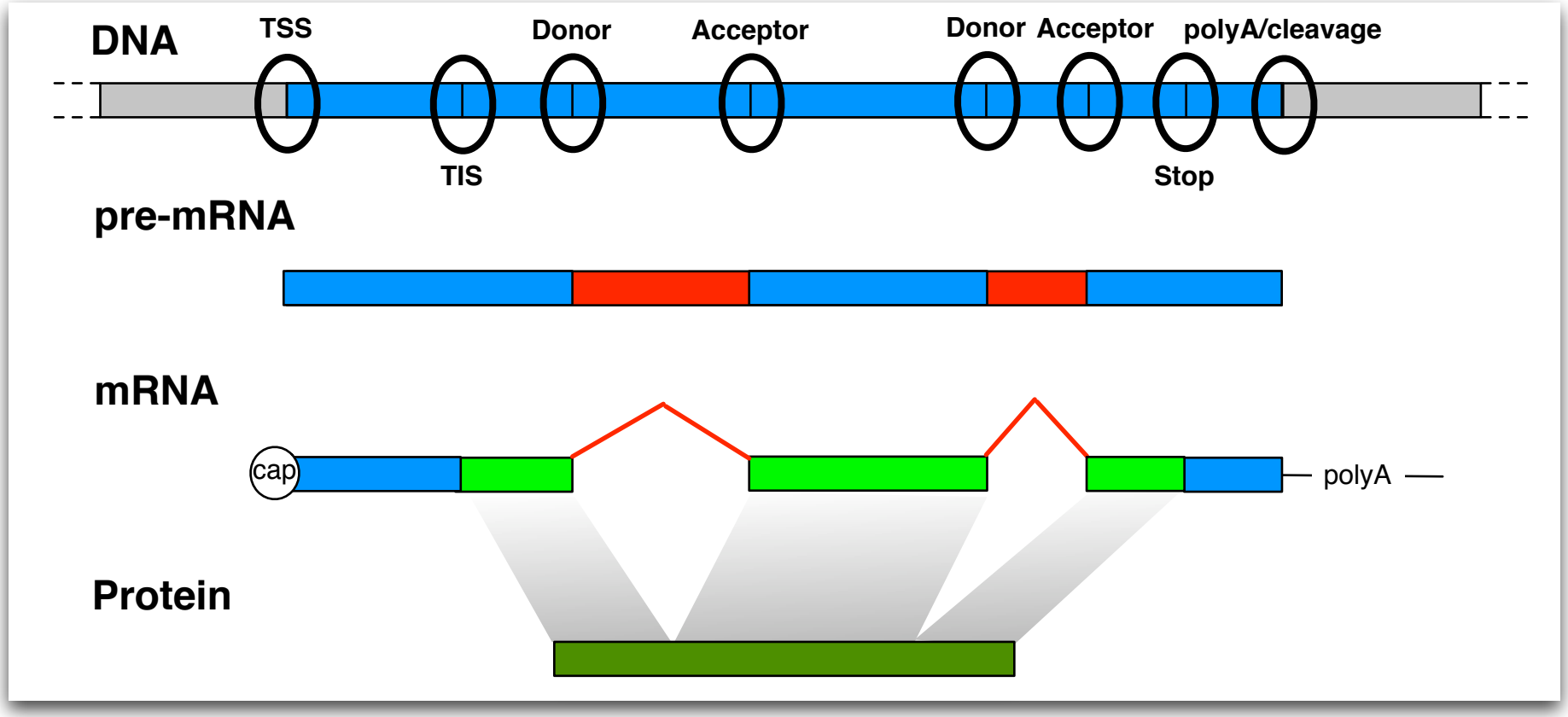
- Given a piece of DNA sequence of length L
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

Gene Finding: Labeling the Genome



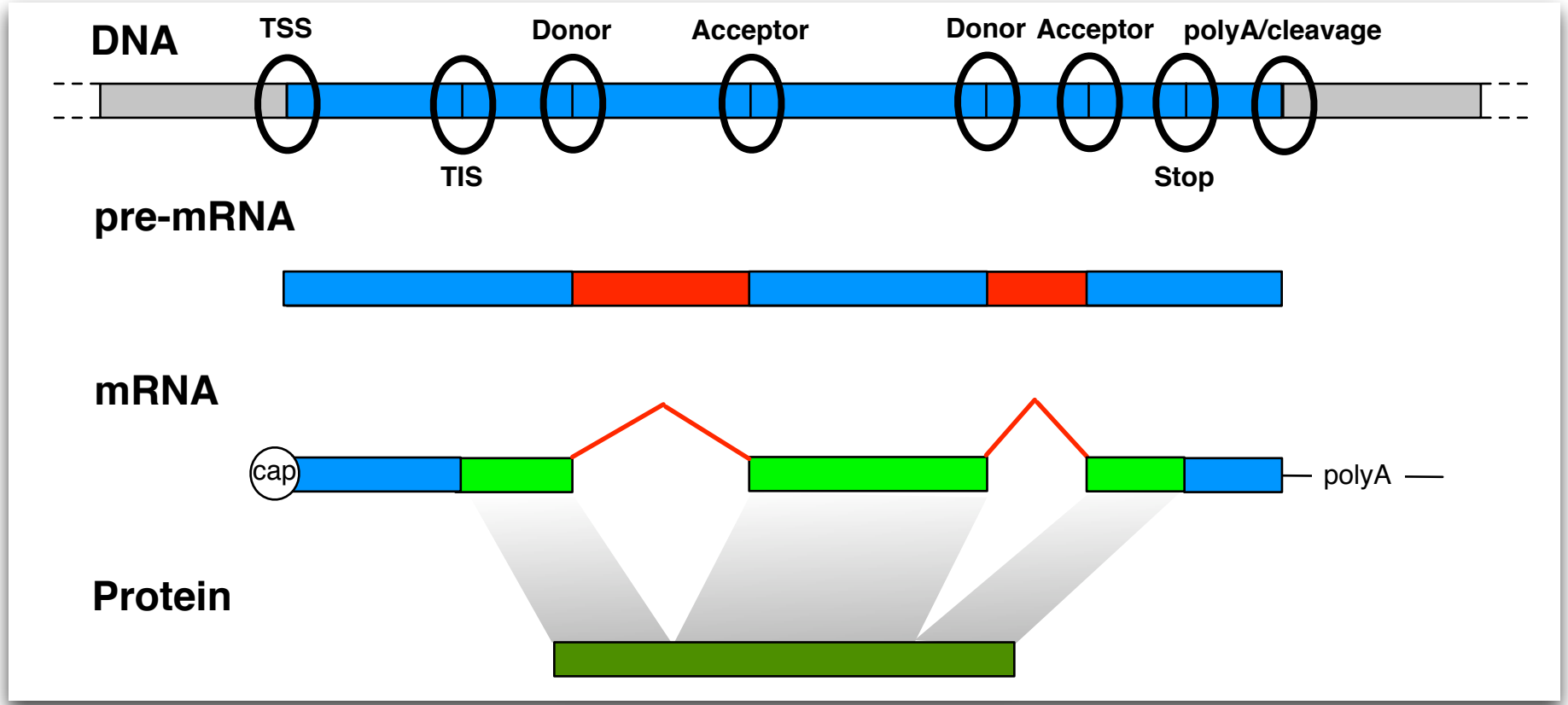
- Given a piece of DNA sequence of length L
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

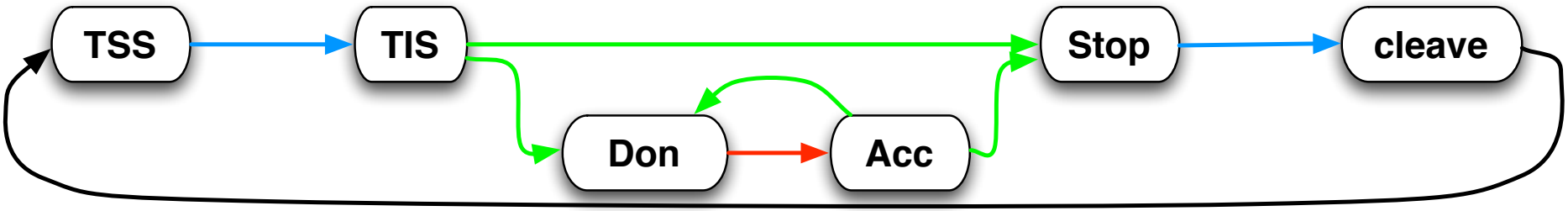
Gene Finding: Labeling the Genome



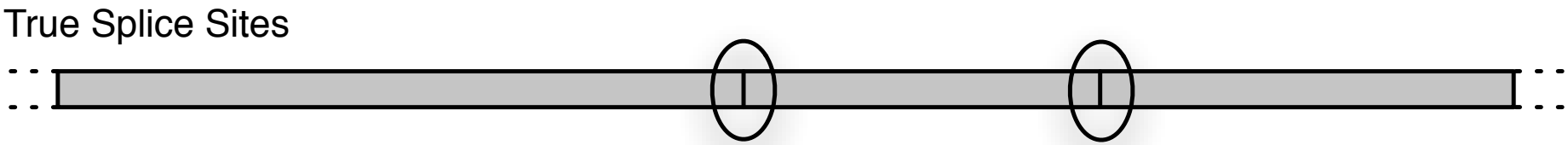
- Given a piece of DNA sequence of length L
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

Gene Finding: Labeling the Genome

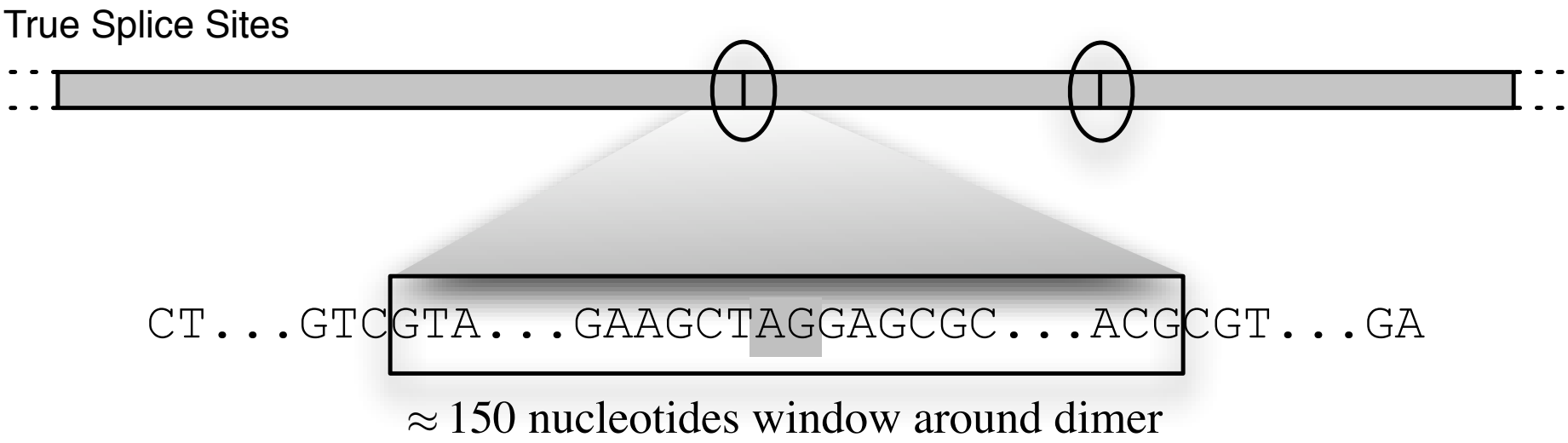




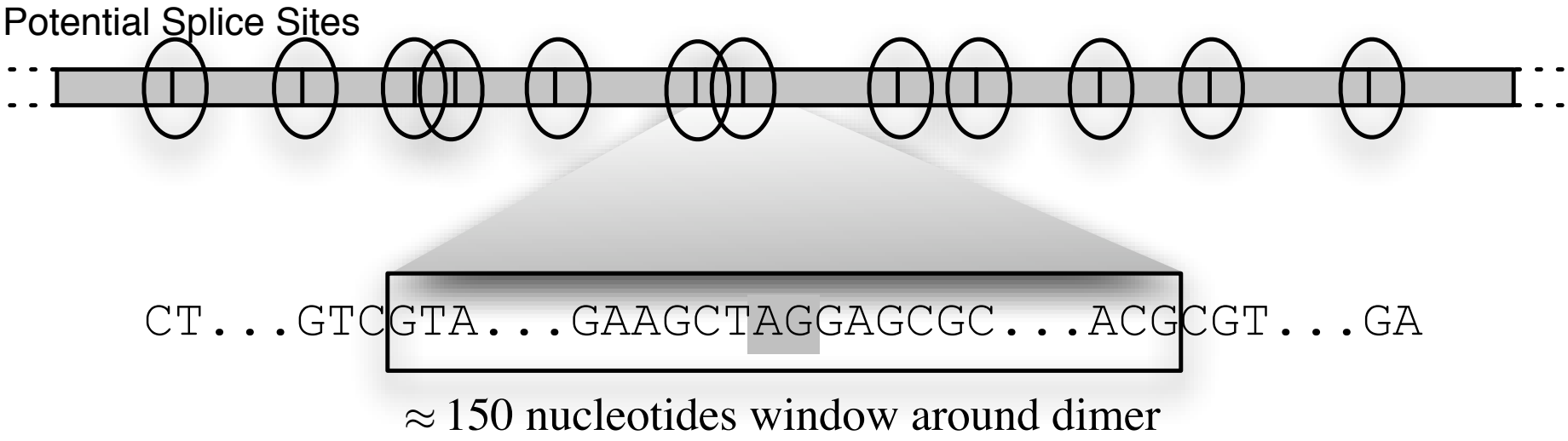
Example: Splice Site Recognition



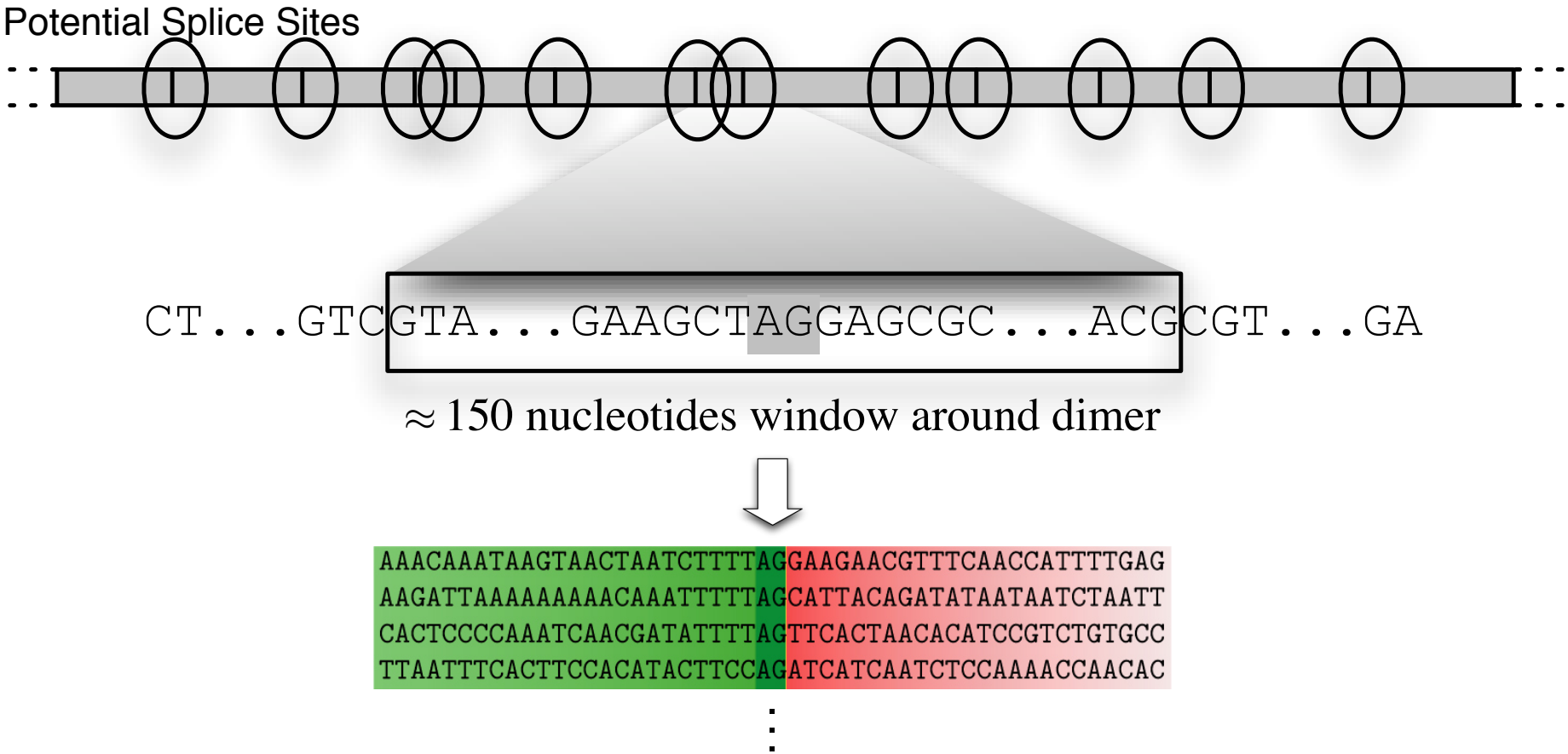
Example: Splice Site Recognition



Example: Splice Site Recognition

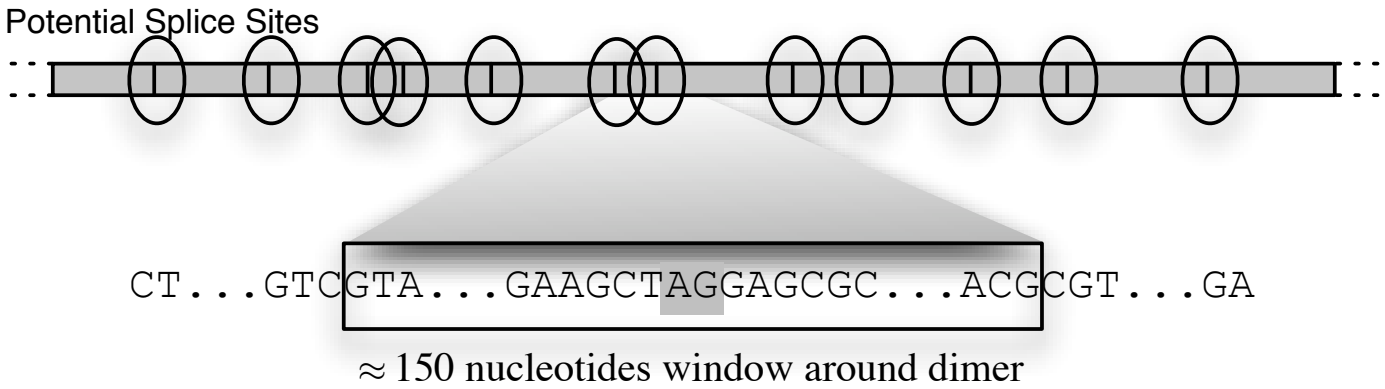


Example: Splice Site Recognition

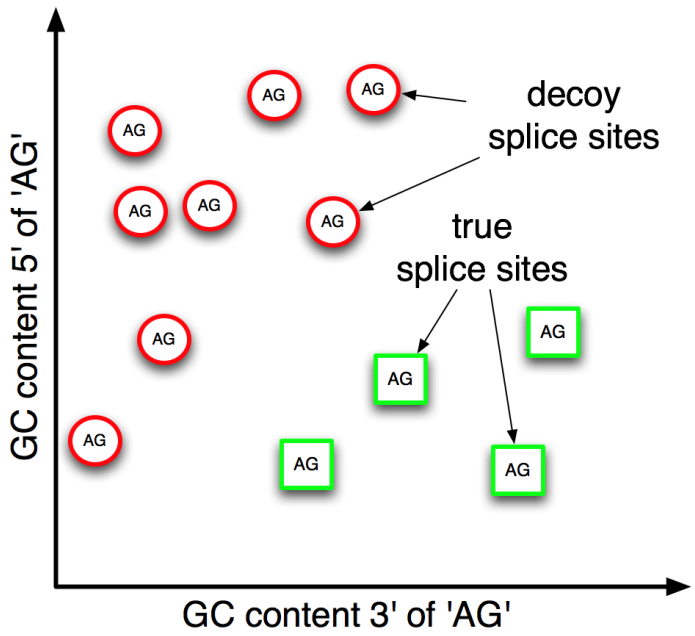


- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

Example: Splice Site Recognition



Basic idea:



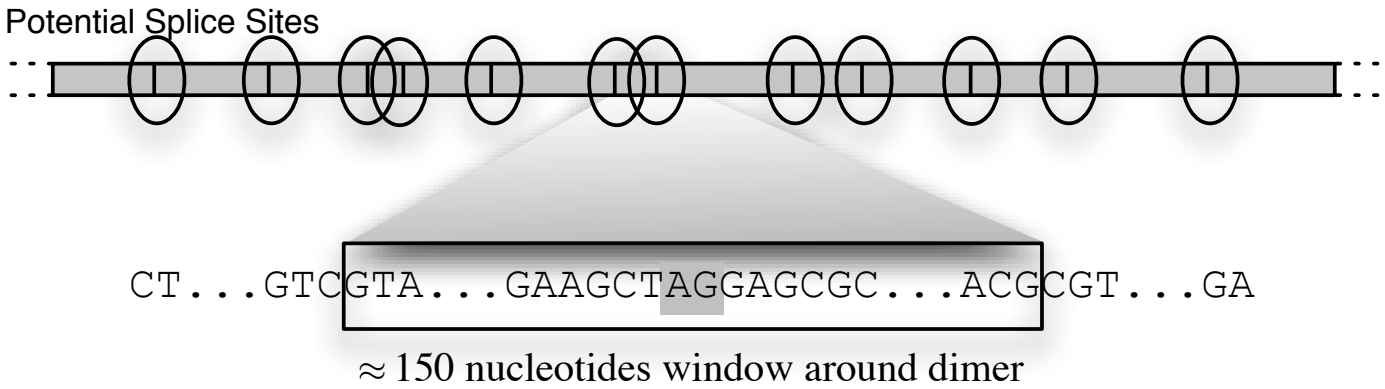
For instance, exploit that exons have higher GC content

or

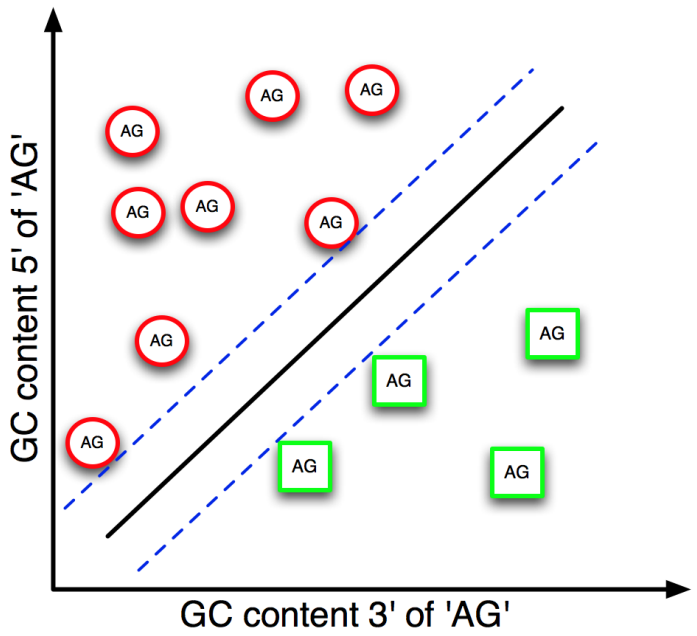
that certain motifs appear near splice sites.

[Sonnenburg et al., 2007]

Example: Splice Site Recognition



Basic idea:



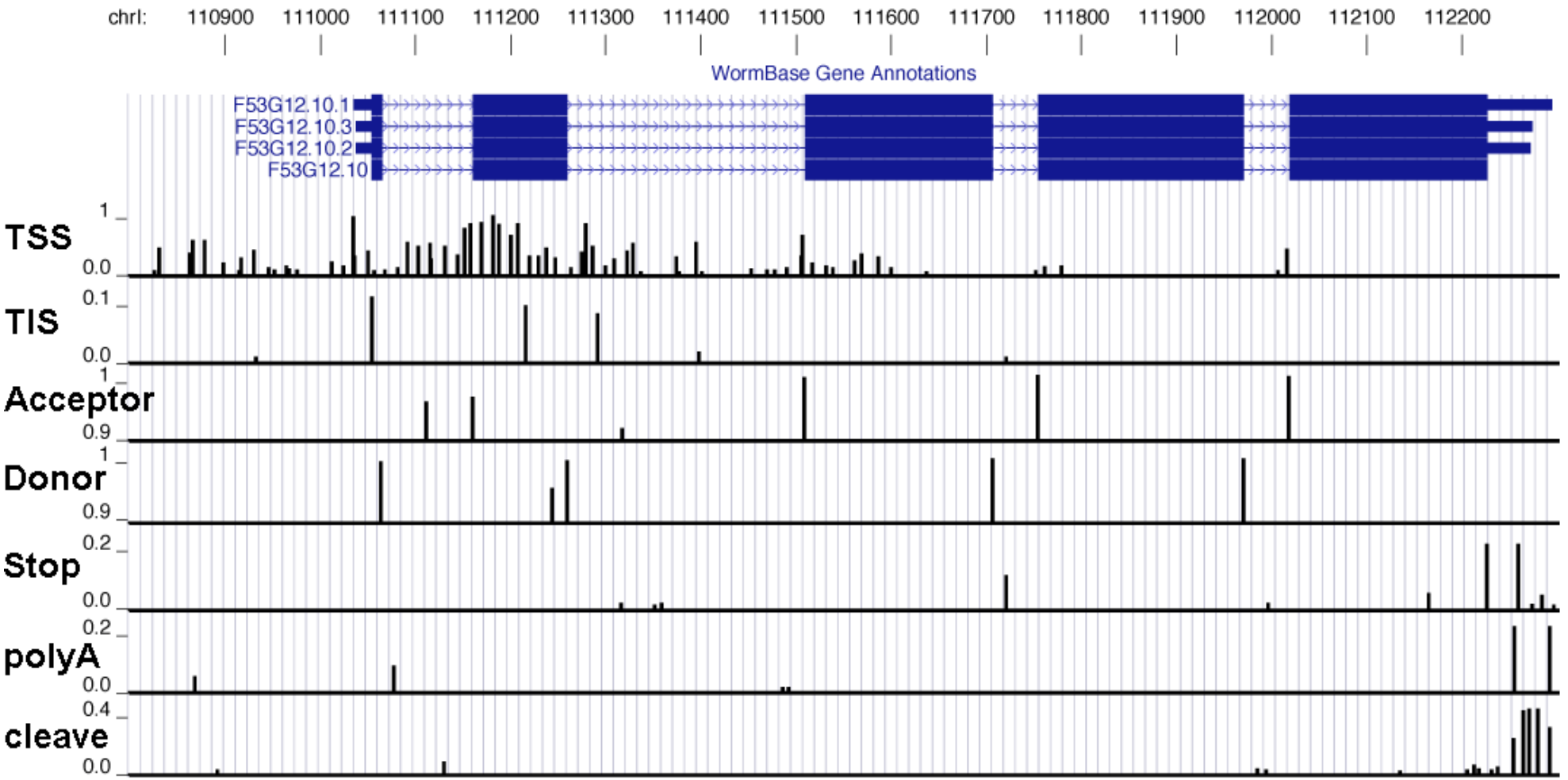
For instance, exploit that exons have higher GC content

or

that certain motifs appear near splice sites.

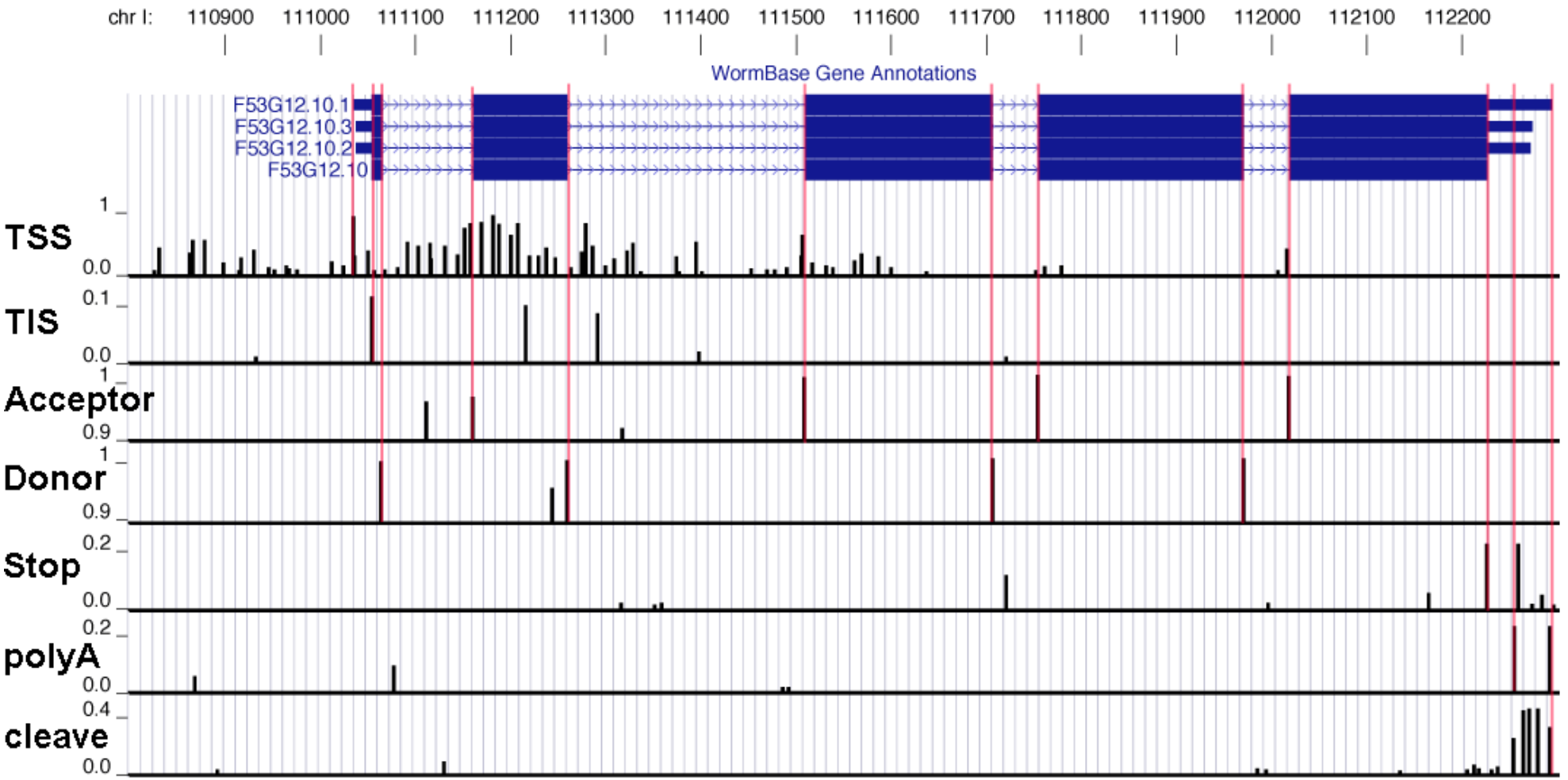
[Sonnenburg et al., 2007]

Example: Predictions in UCSC Browser



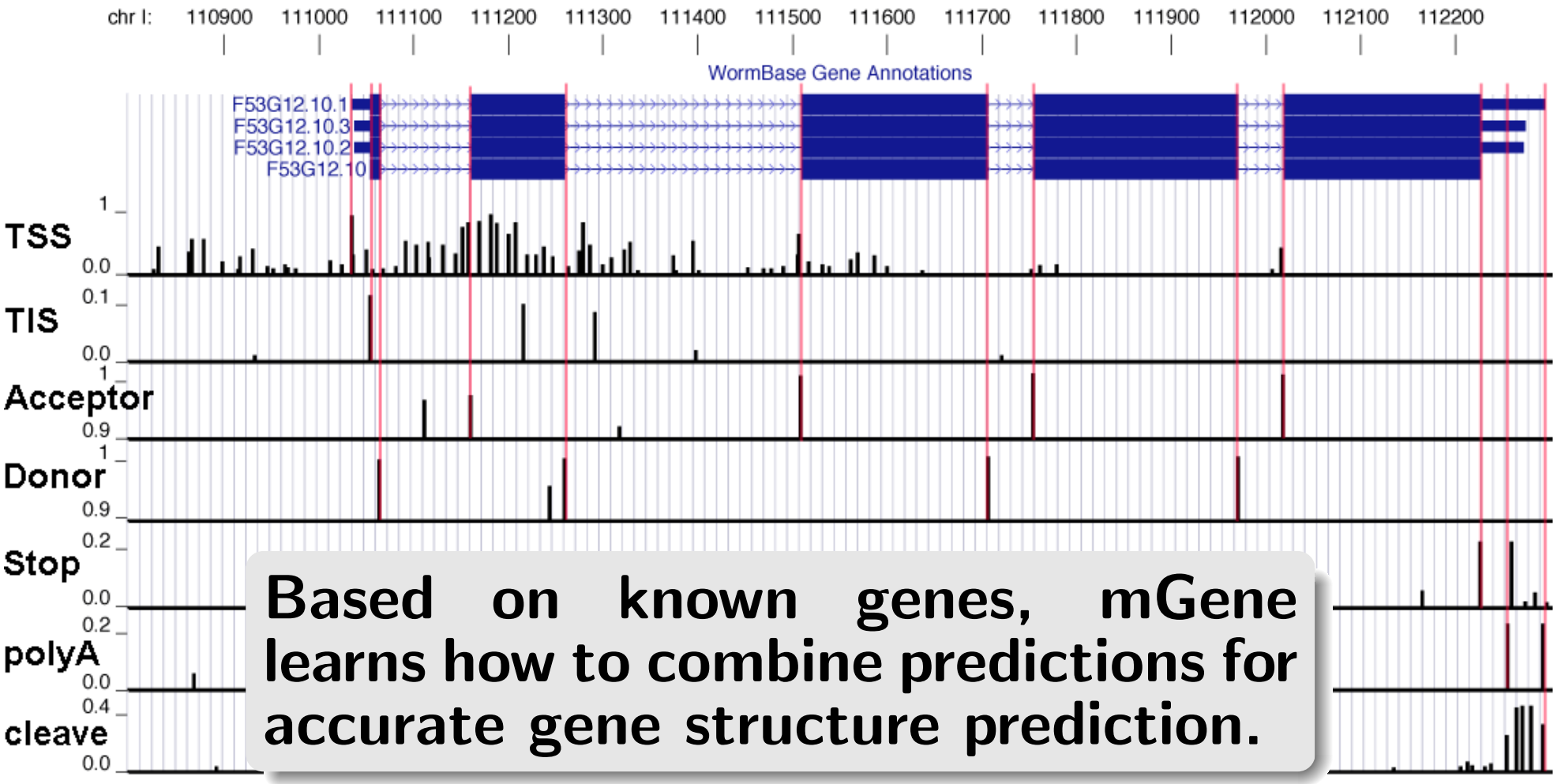
[Schweikert et al., 2008a]

Example: Predictions in UCSC Browser



[Schweikert et al., 2008a]

Example: Predictions in UCSC Browser



Based on known genes, mGene learns how to combine predictions for accurate gene structure prediction.

[Schweikert et al., 2008a]

Results using mGene

- Most accurate method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2008a]

	No. of genes	No. of genes analyzed	Frac. of genes w/ expression
New genes	2,197	57	$\approx 42\%$
Missing unconf. genes	205	24	$\approx 8\%$

- Annotation of other nematode genomes: [Schweikert et al., 2008a]

Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

- Web service for gene predictions of small/medium sized genomes

Results using mGene

- Most accurate method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2008a]

	No. of genes	No. of genes analyzed	Frac. of genes w/ expression
New genes	2,197	57	$\approx 42\%$
Missing unconf. genes	205	24	$\approx 8\%$

- Annotation of other nematode genomes: [Schweikert et al., 2008a]

Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

- Web service for gene predictions of small/medium sized genomes

Results using mGene

- Most accurate method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2008a]

	No. of genes	No. of genes analyzed	Frac. of genes w/ expression
New genes	2,197	57	$\approx 42\%$
Missing unconf. genes	205	24	$\approx 8\%$

- Annotation of other nematode genomes: [Schweikert et al., 2008a]

Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

- Web service for gene predictions of small/medium sized genomes

Results using mGene

- Most accurate method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2008a]

	No. of genes	No. of genes analyzed	Frac. of genes w/ expression
New genes	2,197	57	$\approx 42\%$
Missing unconf. genes	205	24	$\approx 8\%$

- Annotation of other nematode genomes: [Schweikert et al., 2008a]

Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

- Web service for gene predictions of small/medium sized genomes

Limitations/Extensions

- Gene finding still not perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when genes are expressed
- Accurate enough to accurately predict the effects of SNPs?
  - Consensus site changes just the first step [Clark et al., 2007]
- Needs to be adapted to new genomes
  - Requires sufficient number of known gene models for "training"
  - Develop methods that exploit *evolutionary information* and gene models from other genomes [Schweikert et al., 2008b]

Limitations/Extensions

- Gene finding still not perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when genes are expressed
- Accurate enough to accurately predict the effects of SNPs?
  - Consensus site changes just the first step [Clark et al., 2007]
- Needs to be adapted to new genomes
  - Requires sufficient number of known gene models for "training"
  - Develop methods that exploit *evolutionary information* and gene models from other genomes [Schweikert et al., 2008b]

  $\Rightarrow$ Model and understand the differences in transcription, RNA processing & translation.

Limitations/Extensions

- Gene finding still not perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when genes are expressed
- Accurate enough to accurately predict the effects of SNPs?
  - Consensus site changes just the first step [Clark et al., 2007]
- Needs to be adapted to new genomes
  - Requires sufficient number of known gene models for "training"
  - Develop methods that exploit *evolutionary information* and gene models from other genomes [Schweikert et al., 2008b]

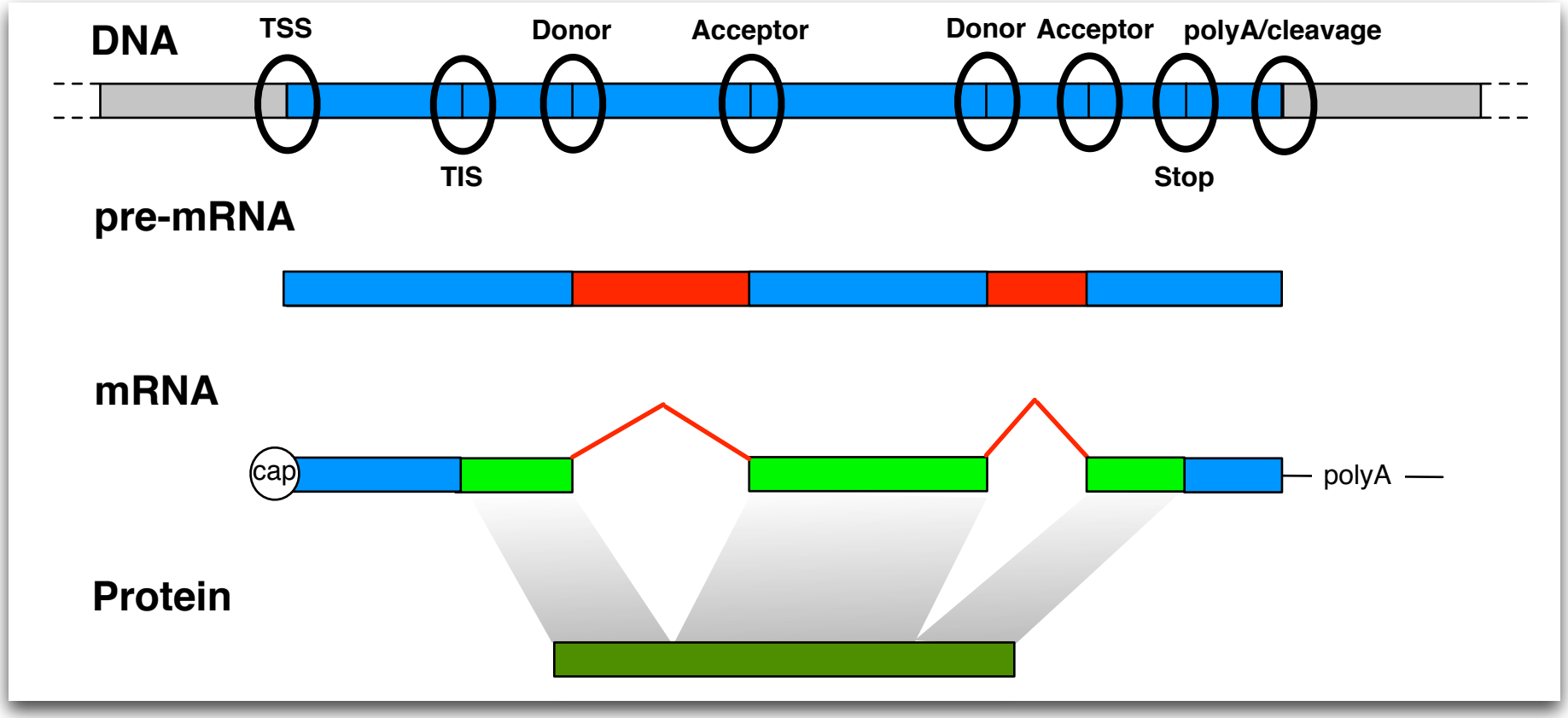
$\Rightarrow$ Model and understand the differences in transcription, RNA processing & translation.

Limitations/Extensions

- Gene finding still not perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when genes are expressed
- Accurate enough to accurately predict the effects of SNPs?
  - Consensus site changes just the first step [Clark et al., 2007]
- Needs to be adapted to new genomes
  - Requires sufficient number of known gene models for "training"
  - Develop methods that exploit *evolutionary information* and gene models from other genomes [Schweikert et al., 2008b]

  $\Rightarrow$ Model and understand the differences in transcription, RNA processing & translation.

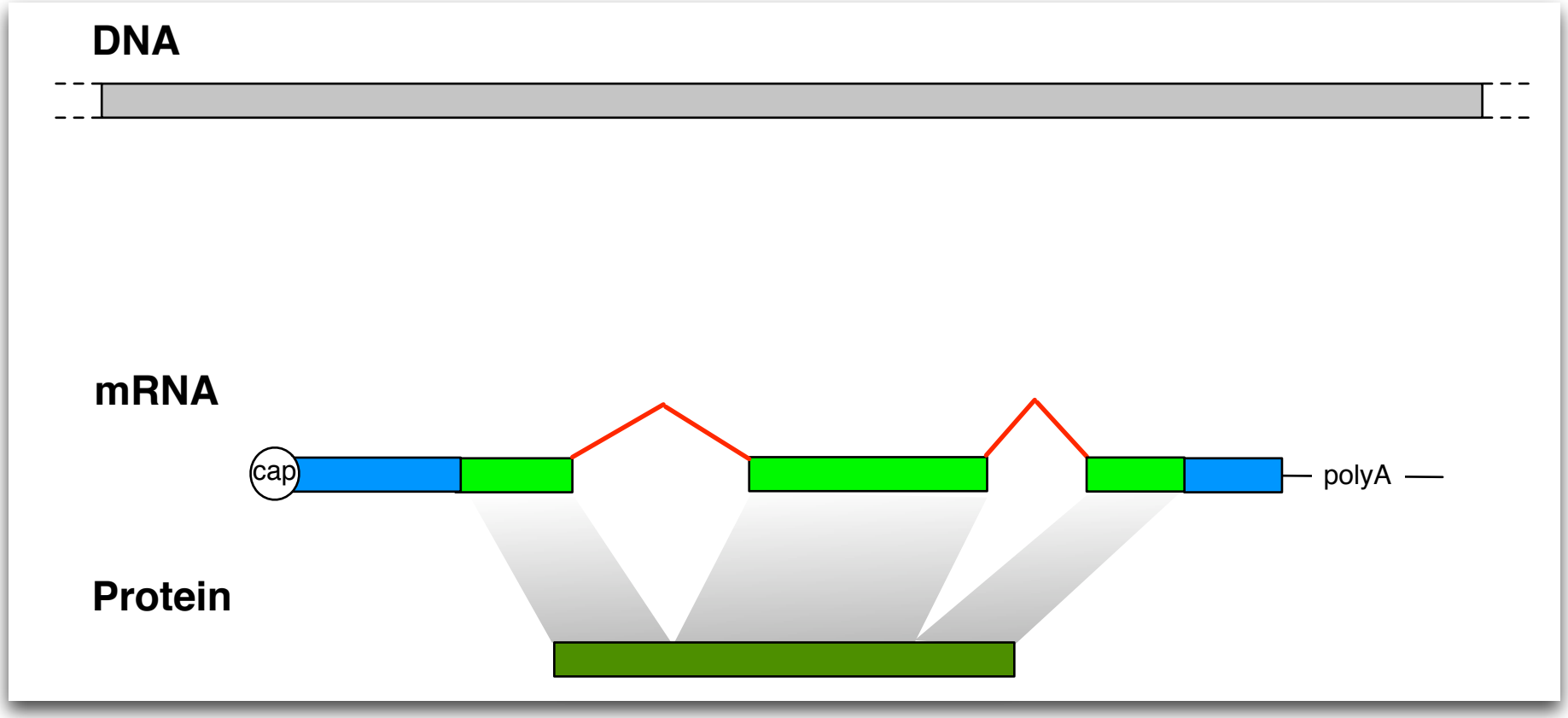
From Genome to Proteins



Directly measure the transcriptome

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

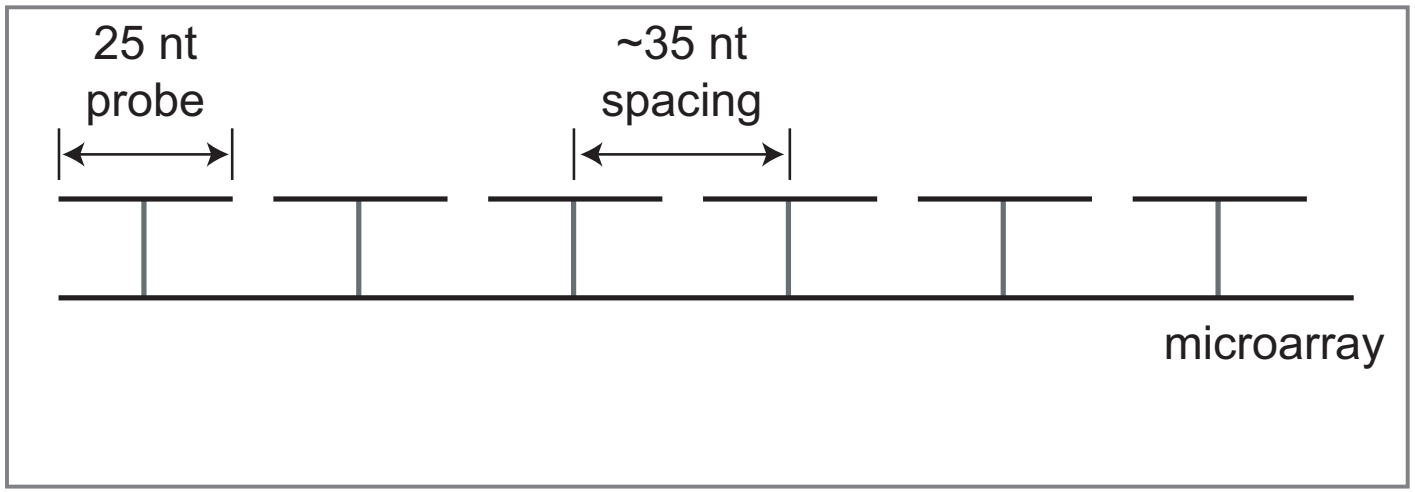
From Transcriptome to Proteins



Directly measure the transcriptome

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

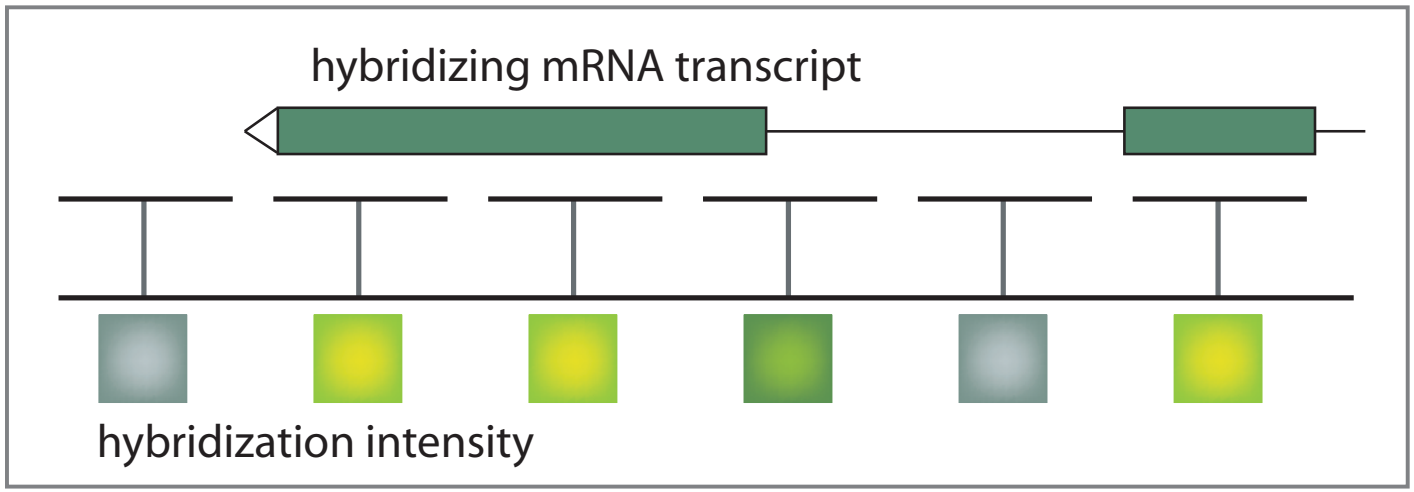
Tiling Arrays for Transcriptome Analysis



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

(Work in collaboration with Detlef Weigel's lab, Max Planck Institute for Developmental Biology)

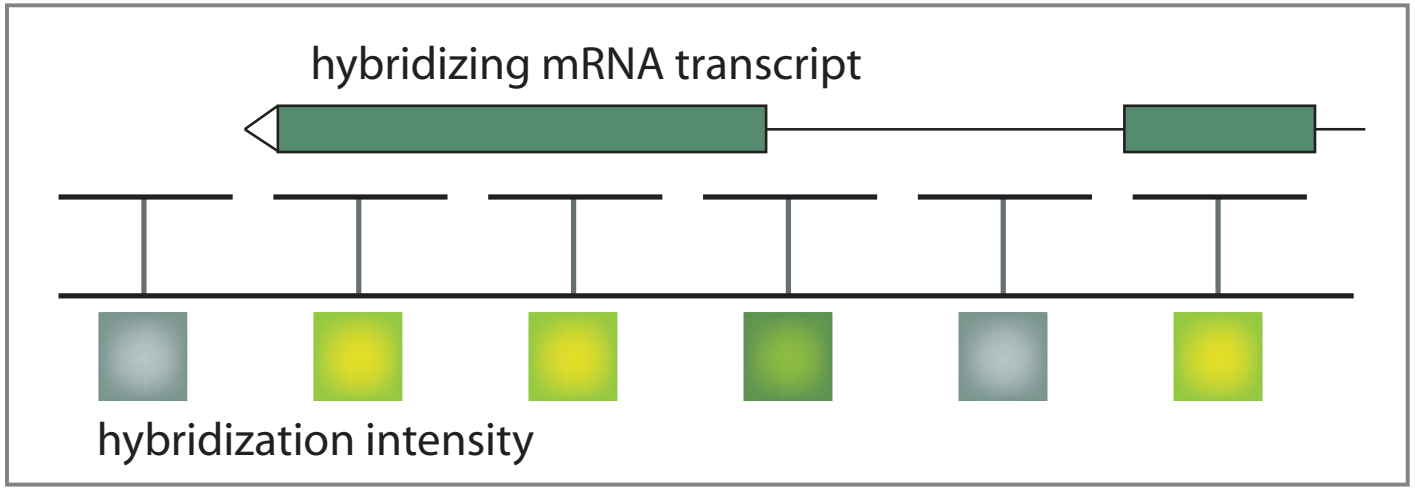
Tiling Arrays for Transcriptome Analysis



- Whole-genome quantitative measurements
- Cost-effective
  - ⇒ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  - ⇒ Do not rely on annotations or known cDNAs
  - ⇒ For known genes gives similar results as dedicated microarrays

(Work in collaboration with Detlef Weigel's lab, Max Planck Institute for Developmental Biology)

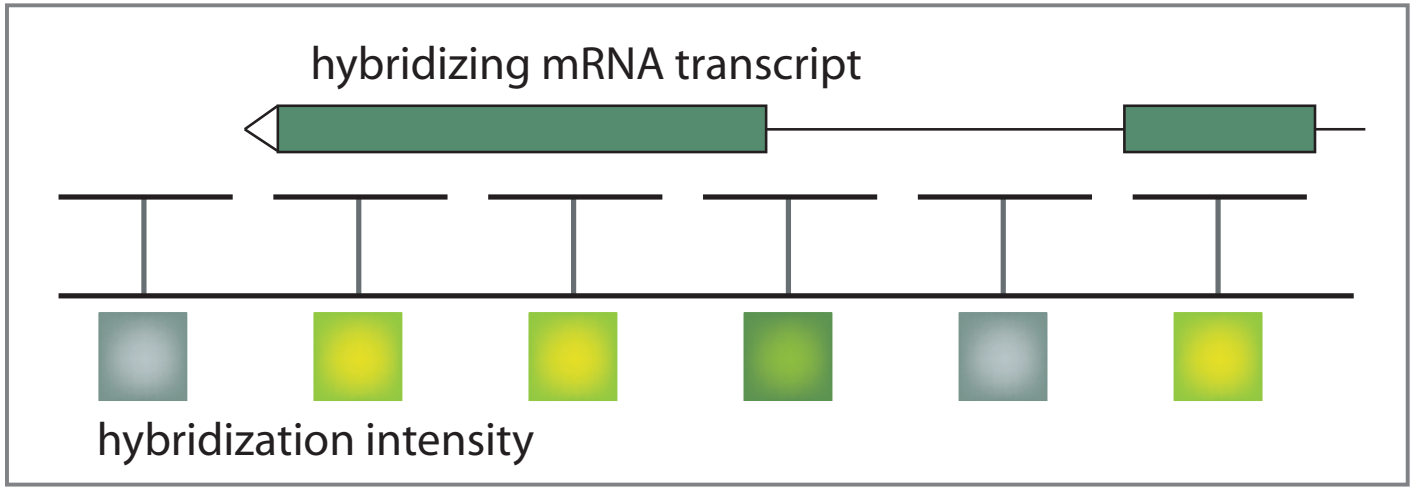
Tiling Arrays for Transcriptome Analysis



- Whole-genome quantitative measurements
- Cost-effective
  - ⇒ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  - ⇒ Do not rely on annotations or known cDNAs
  - ⇒ For known genes, gives similar results as traditional microarrays

[Laubinger et al., 2008b]

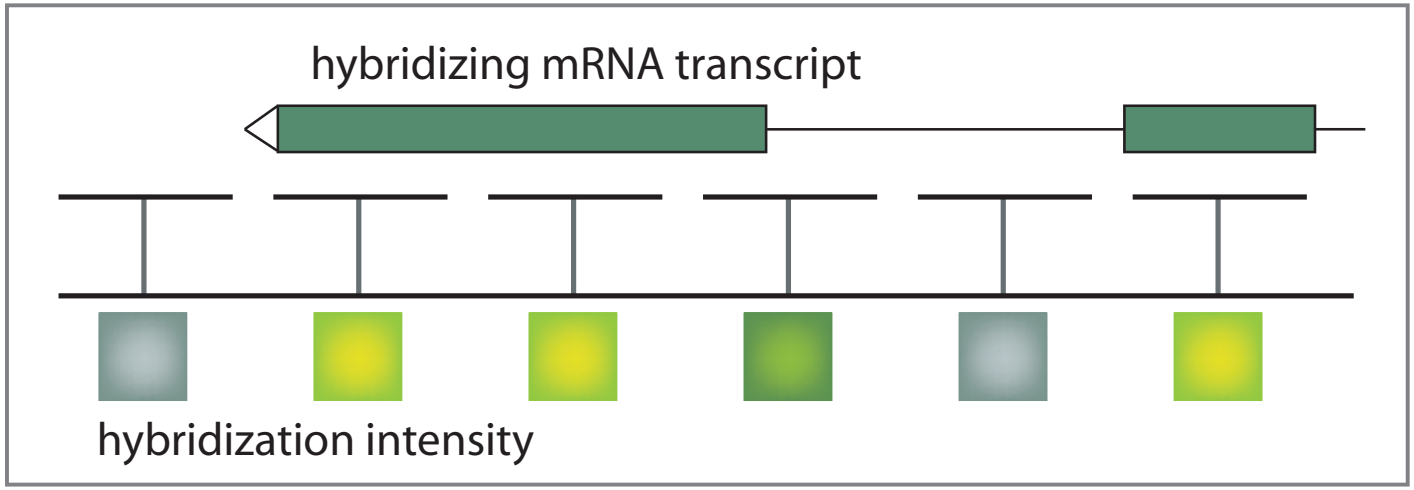
Tiling Arrays for Transcriptome Analysis



- Whole-genome quantitative measurements
- Cost-effective
  - ⇒ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  - ⇒ Do not rely on annotations or known cDNAs
  - ⇒ For known genes, gives similar results as traditional microarrays

[Laubinger et al., 2008b]

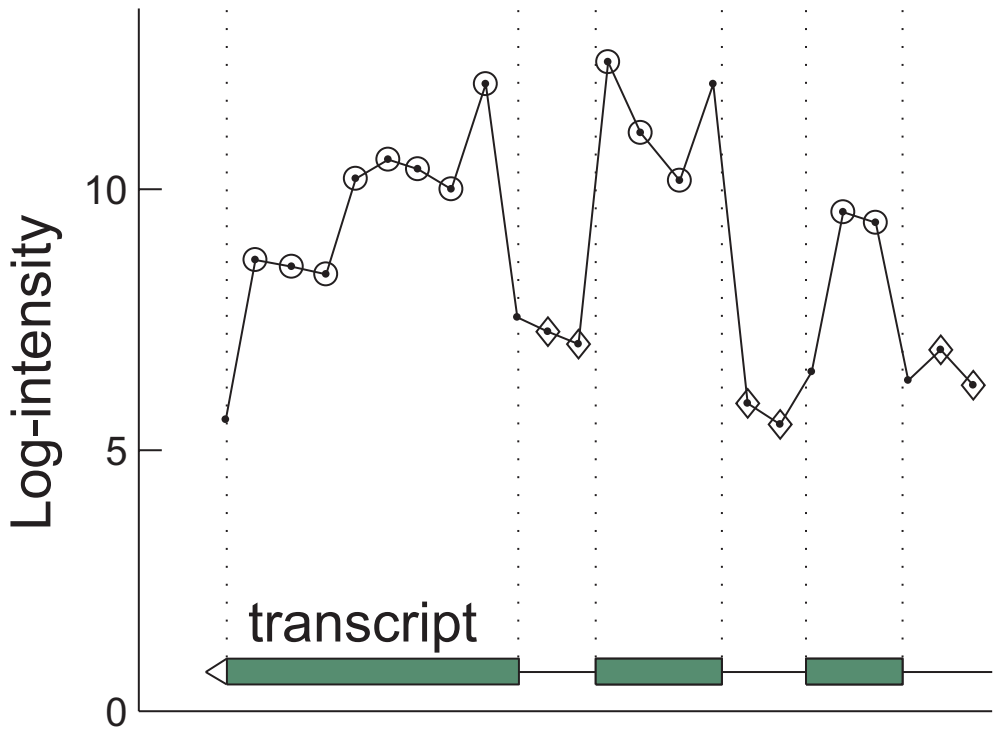
Tiling Arrays for Transcriptome Analysis



- Whole-genome quantitative measurements
- Cost-effective
  - ⇒ Replicates affordable, many tissues / mutants / conditions
- Unbiased
  - ⇒ Do not rely on annotations or known cDNAs
  - ⇒ For known genes, gives similar results as traditional microarrays

[Laubinger et al., 2008b]

Intensities are Noisy Measurements



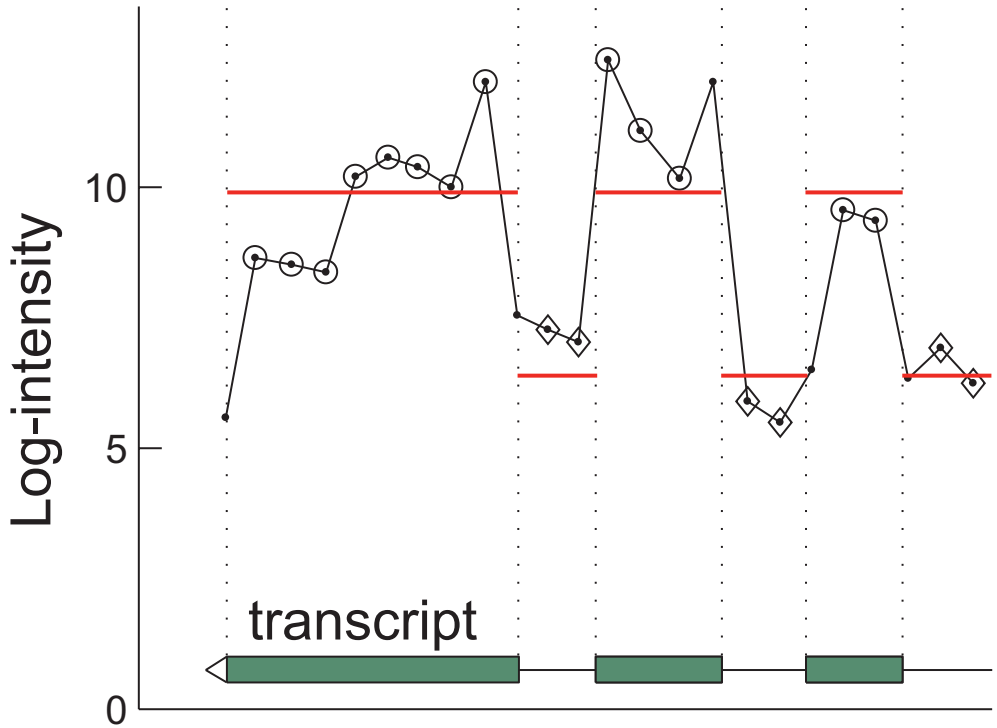
- observed intensity
- annotated exonic
- annotated intronic

Systematic bias induced by probe sequence effects

$\Rightarrow$ Model effect for normalization

[Zeller et al., 2008b]

Intensities are Noisy Measurements

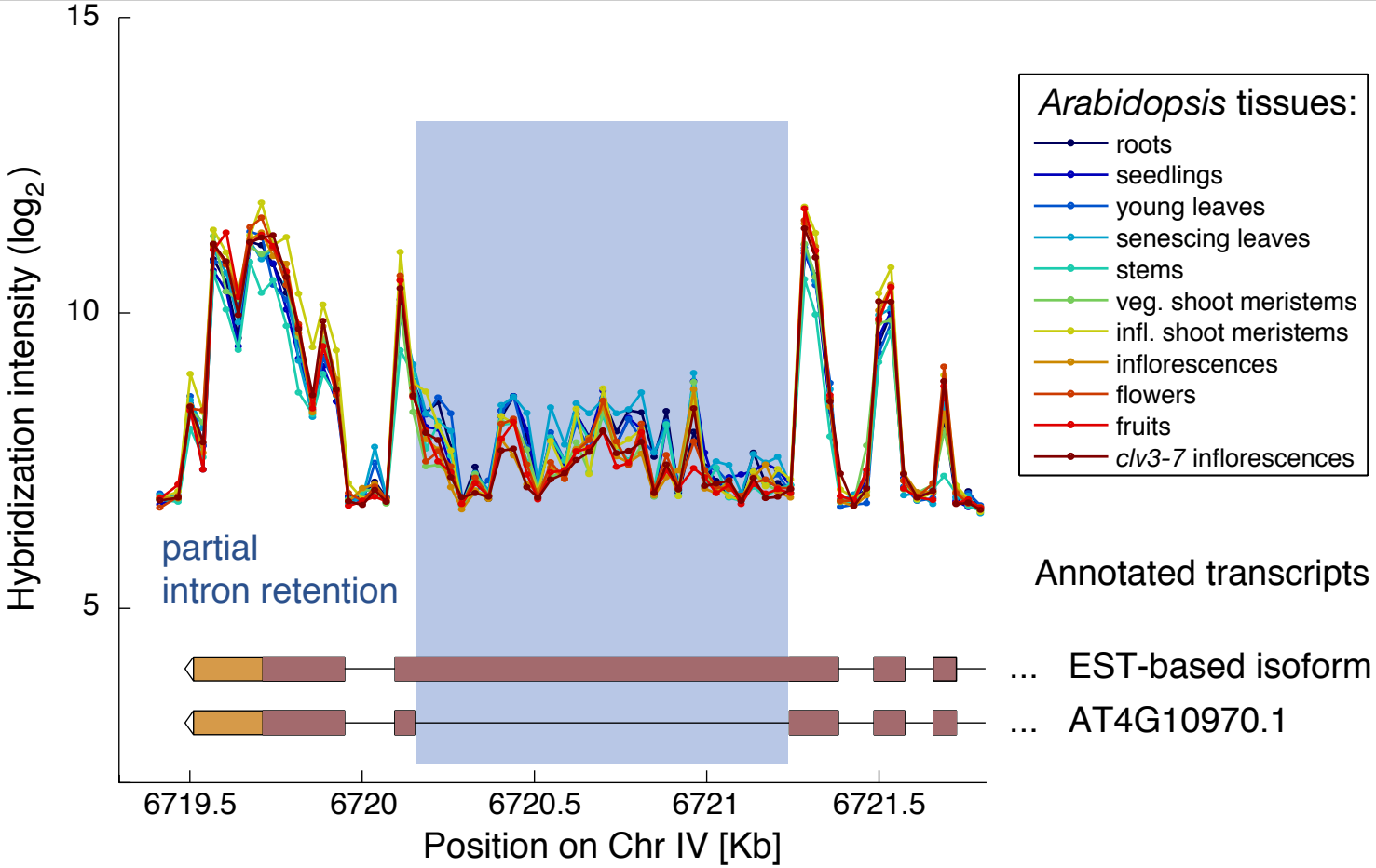


- observed intensity
- annotated exonic
- annotated intronic
- "ideal noise-free intensity"

Systematic bias induced by probe sequence effects

$\Rightarrow$ Model effect for normalization [Zeller et al., 2008b]

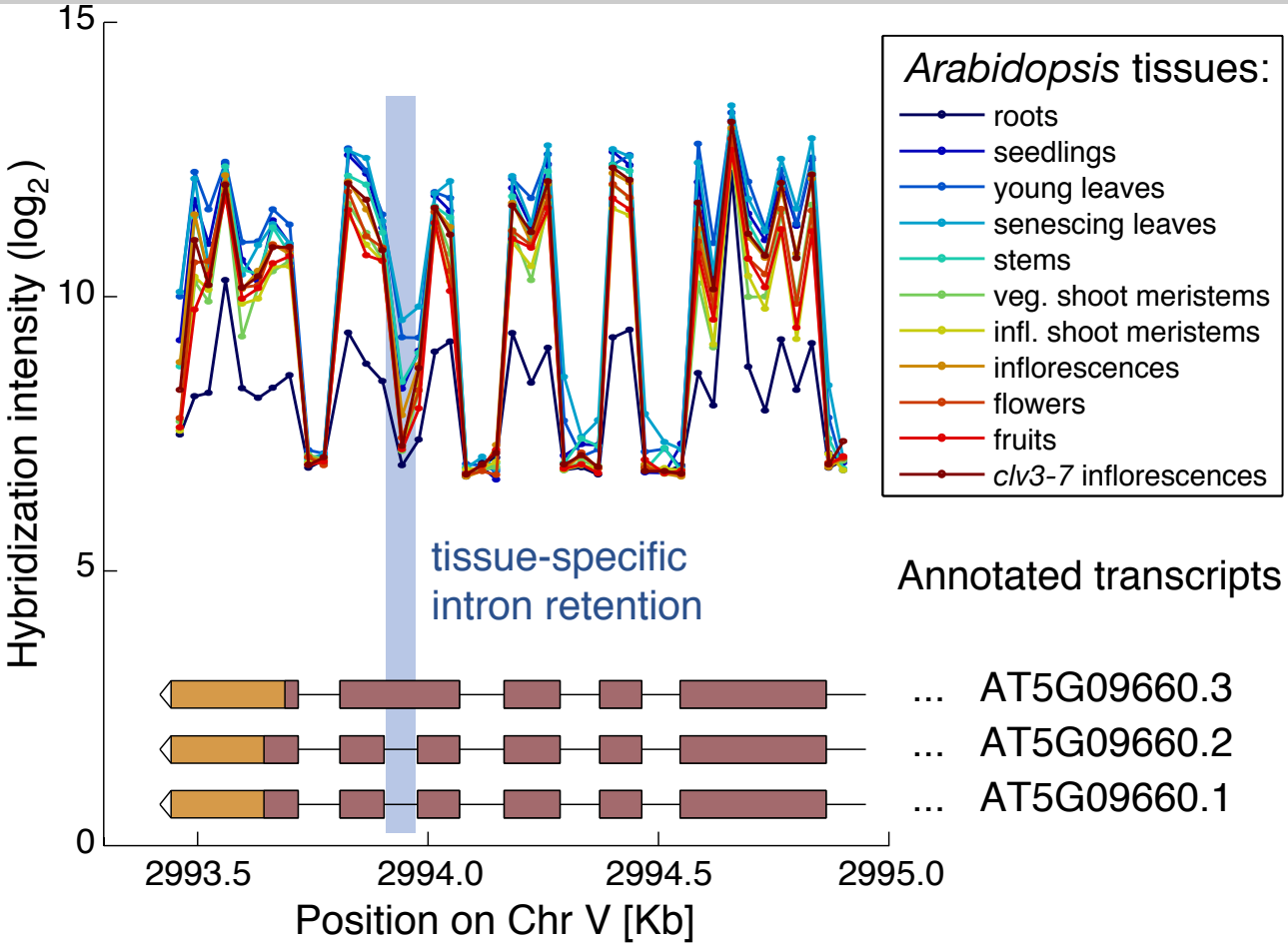
Alternatively Spliced Genes



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]

Genome-wide Analyses of Alternative Splicing

- *A. thaliana* transcriptomes in 11 different tissues, 5 different abiotic stresses, RNA processing mutants (3 replicates)
- 1^{st} exon/intron-resolution analysis in plants [Laubinger et al., 2008b]

- Identified ≈7,000 cases of intron retentions (est. FDR=14%)
- Many of them show differential splicing [Eichner et al., 2008]

- Profiling cap-binding/miRNA processing mutants

$\Rightarrow$ Influence splicing efficiency

$\Rightarrow$ 1^{st} intron often retained [Laubinger et al., 2008a]

Genome-wide Analyses of Alternative Splicing



- *A. thaliana* transcriptomes in 11 different tissues, 5 different abiotic stresses, RNA processing mutants (3 replicates)
- 1^{st} exon/intron-resolution analysis in plants [Laubinger et al., 2008b]

- Identified ≈7,000 cases of intron retentions (est. FDR=14%)
- Many of them show differential splicing [Eichner et al., 2008]

- Profiling cap-binding/miRNA processing mutants

$\Rightarrow$ Influence splicing efficiency

$\Rightarrow$ 1^{st} intron often retained [Laubinger et al., 2008a]

Genome-wide Analyses of Alternative Splicing



- *A. thaliana* transcriptomes in 11 different tissues, 5 different abiotic stresses, RNA processing mutants (3 replicates)
- 1^{st} exon/intron-resolution analysis in plants [Laubinger et al., 2008b]

- Identified ≈7,000 cases of intron retentions (est. FDR=14%)
- Many of them show differential splicing [Eichner et al., 2008]

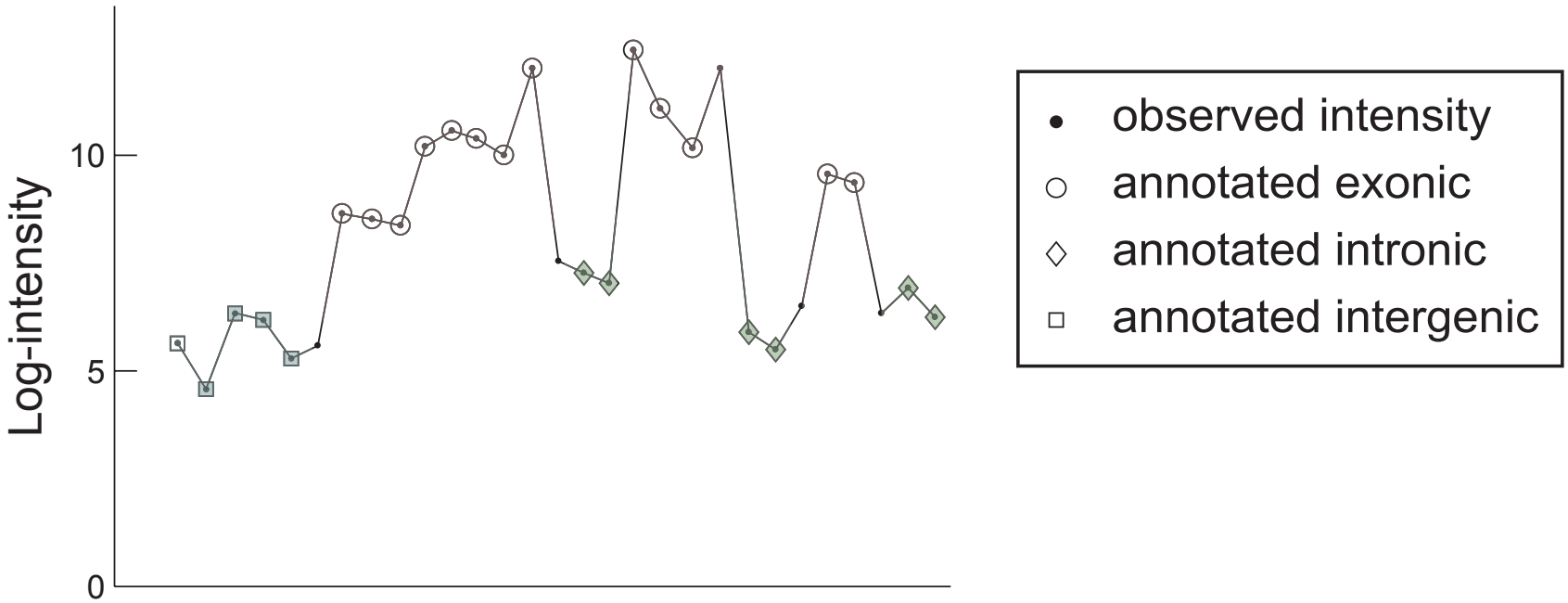
- Profiling cap-binding/miRNA processing mutants

$\Rightarrow$ Influence splicing efficiency

$\Rightarrow$ 1^{st} intron often retained [Laubinger et al., 2008a]

Tiling Array Segmentation

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



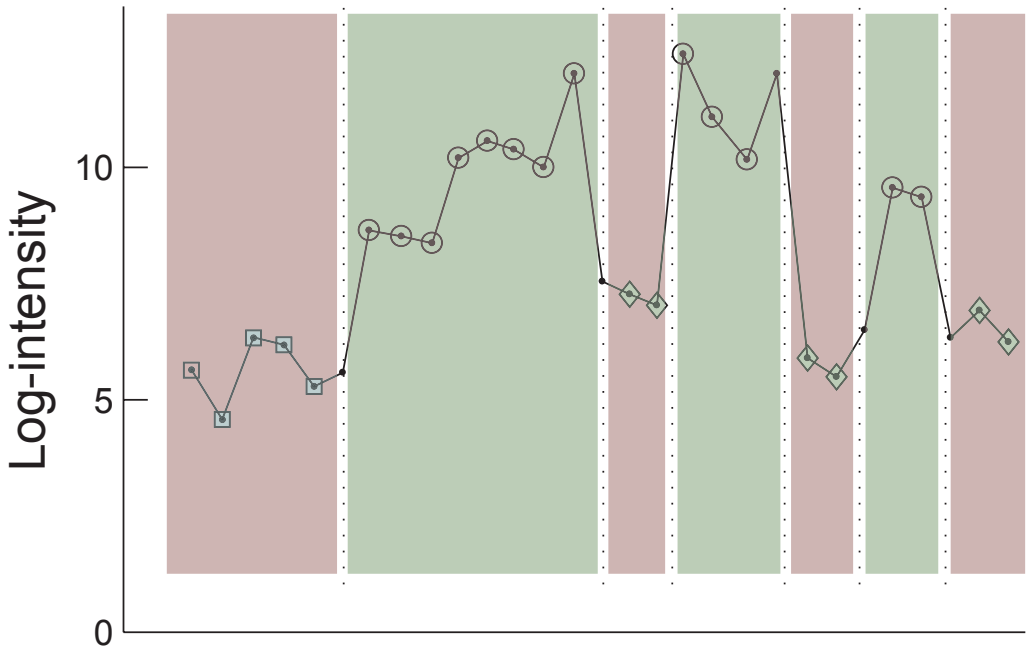
Novel segmentation method ("mSTAD")

- accounts for spliced transcripts
- provides very accurate predictions

[Zeller et al., 2008b]

Tiling Array Segmentation

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



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

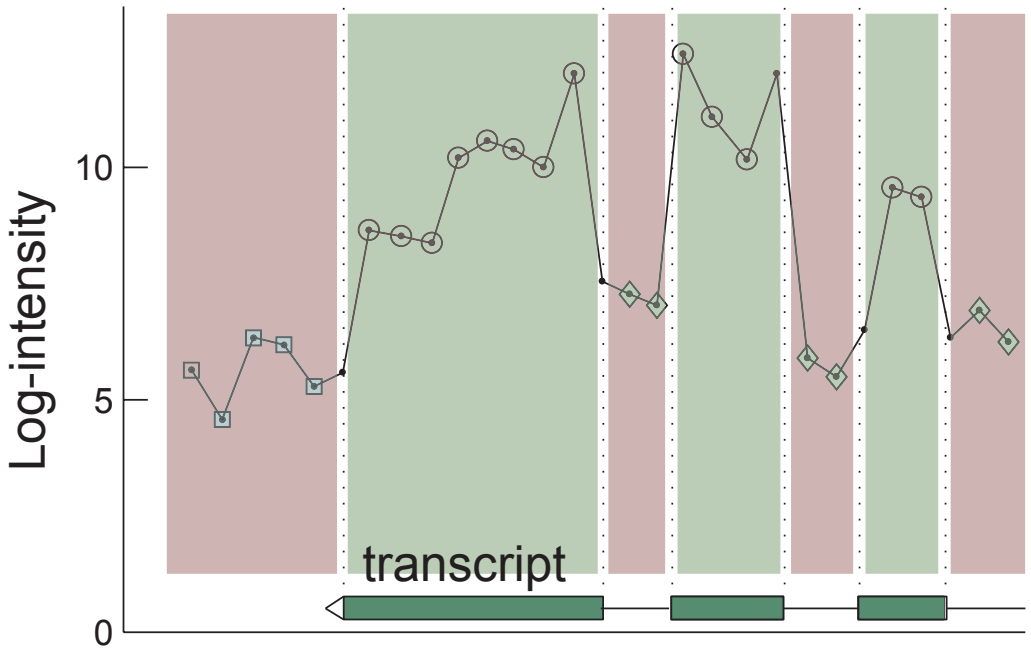
Novel segmentation method ("mSTAD")

- accounts for spliced transcripts
- provides very accurate predictions

[Zeller et al., 2008b]

Tiling Array Segmentation

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



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

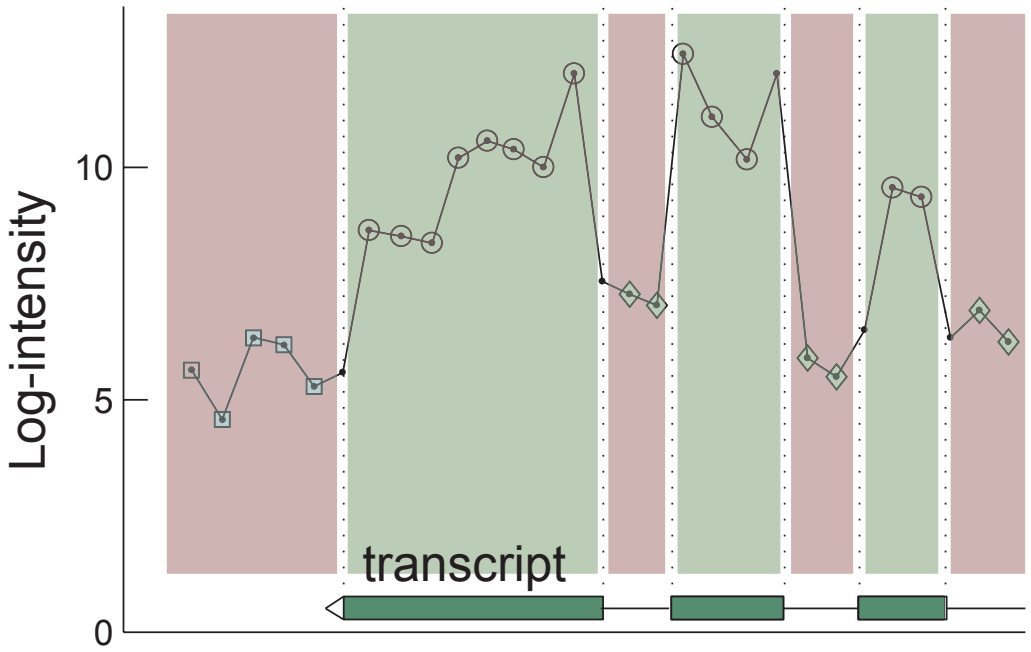
Novel segmentation method ("mSTAD")

- accounts for spliced transcripts
- provides very accurate predictions

[Zeller et al., 2008b]

Tiling Array Segmentation

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



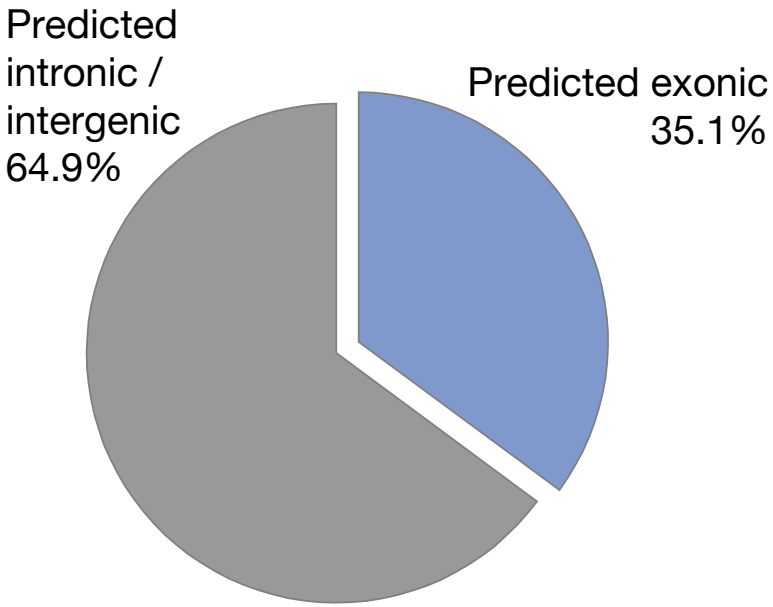
- observed intensity
- annotated exonic
- annotated intronic
- annotated intergenic

Novel segmentation method ("mSTAD")

- accounts for spliced transcripts
- provides very accurate predictions

[Zeller et al., 2008b]

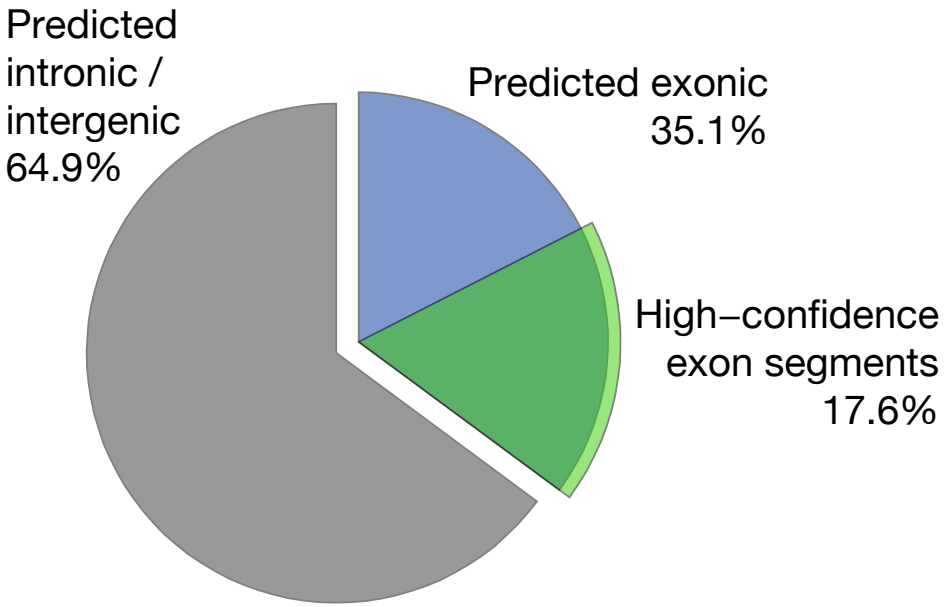
Discovery of New Transcripts



- 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]

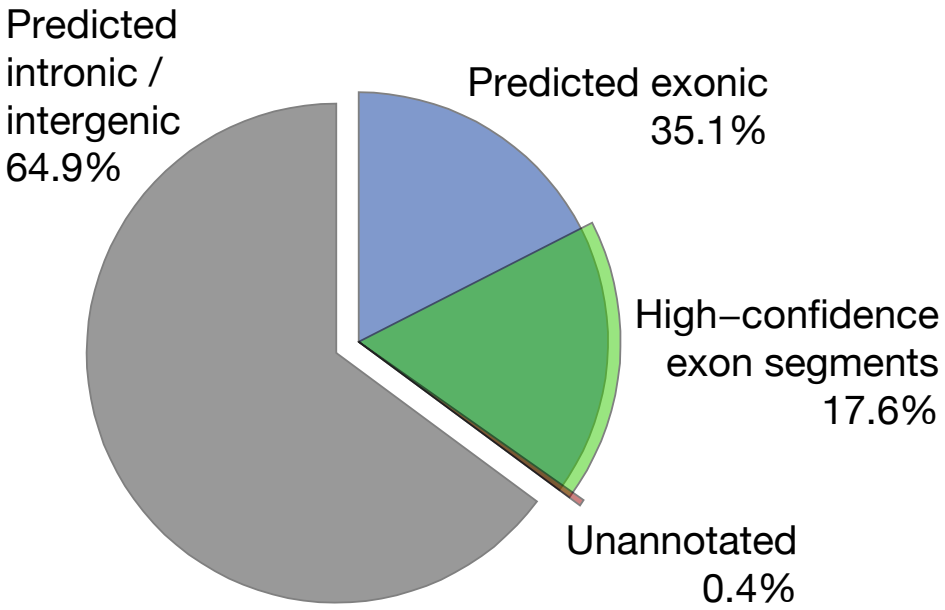
Discovery of New Transcripts



- 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]

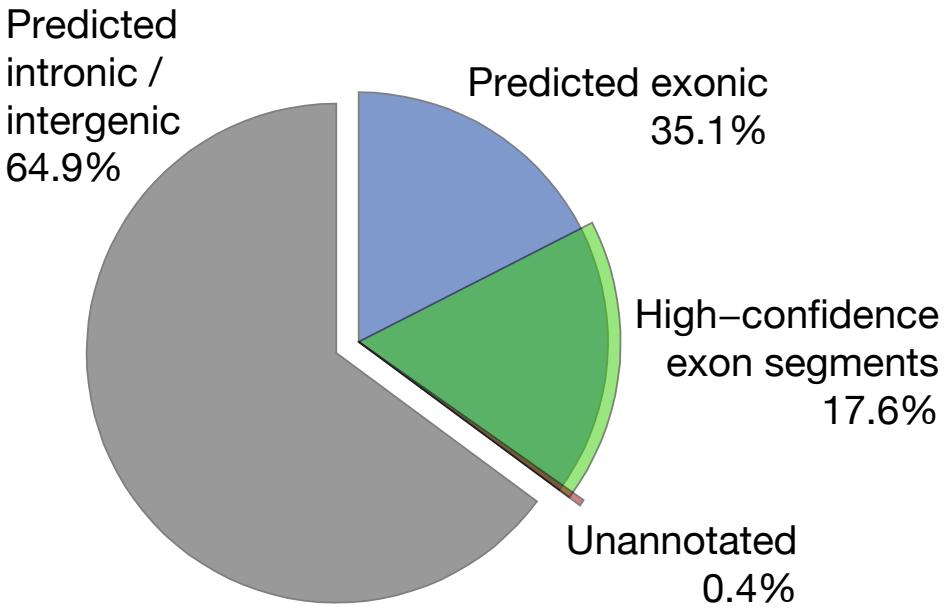
Discovery of New Transcripts



- 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]

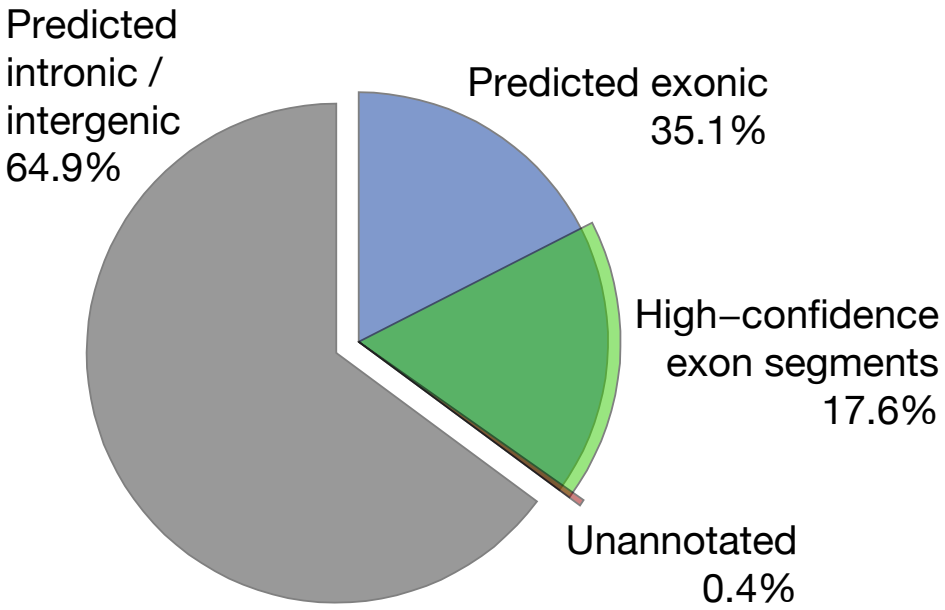
Discovery of New Transcripts



- 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]

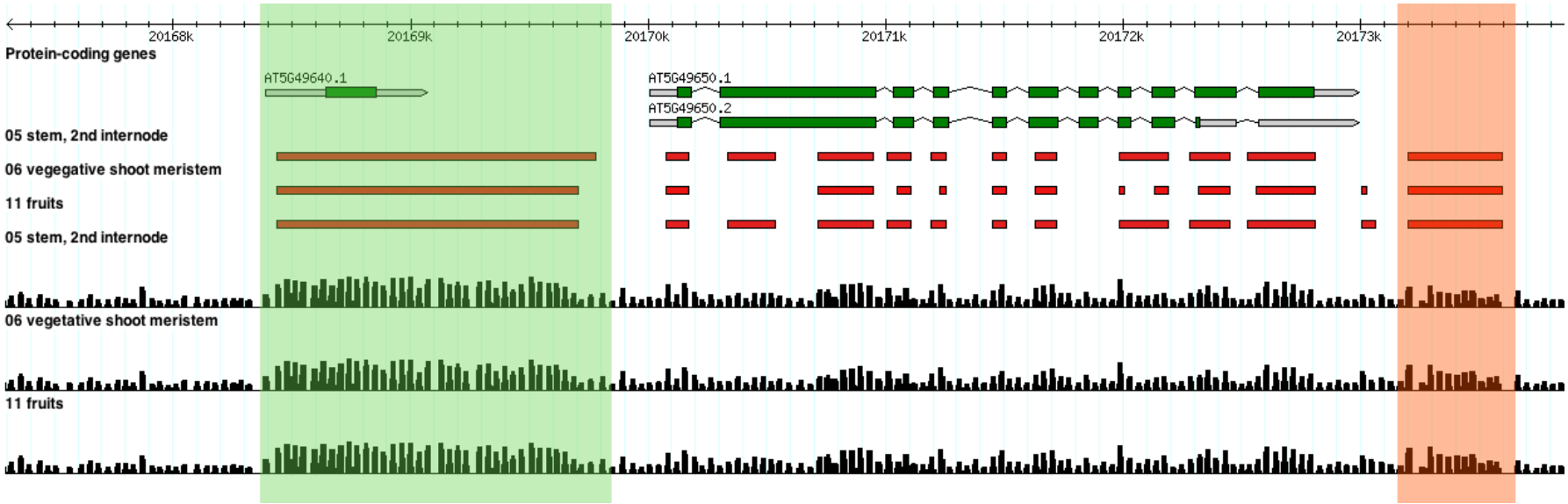
Discovery of New Transcripts



- 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]

Discovery of New Transcripts

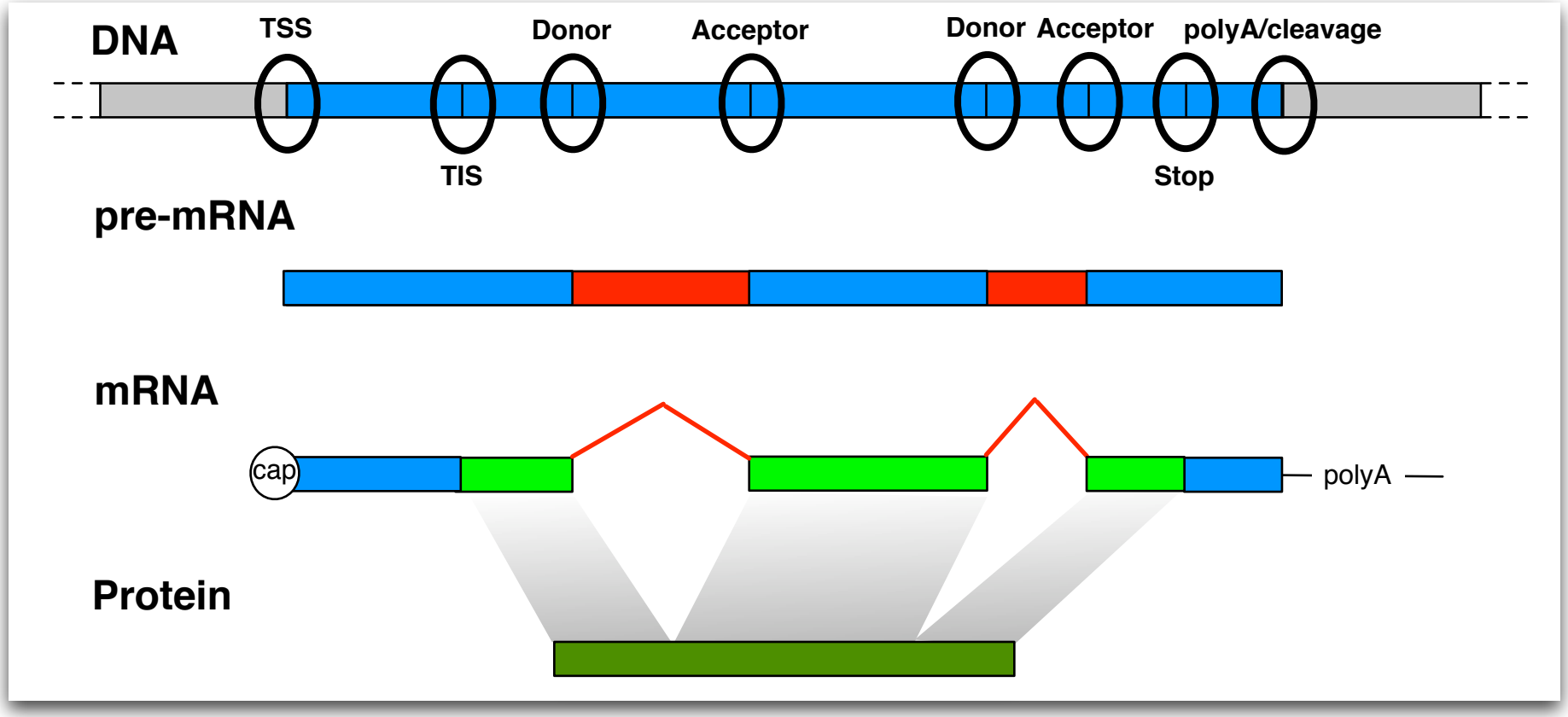


Segment supported by ESTs; new exon segment (no ESTs)

- 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]

From Genome & Transcriptome to Proteins



Combine *ab initio* predictions with transcriptome measurements

- Higher accuracy
- Condition/tissue specific predictions

Transcriptome Measurements for Improved Gene Finding



Idea: Improve mGene by using tiling array measurements to enhance recognition of exonic regions.

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

1. mGene (*ab initio*)
   transcript level performance: **mean(SN, SP) = 76.0%**
2. with tiling arrays *(young leaves)*
   transcript level performance: **mean(SN, SP) = 80.4%**
3. with tiling arrays *(inflorescence)*
   transcript level performance: **mean(SN, SP) = 77.1%**
4. with tiling arrays *(root)*
   transcript level performance: **mean(SN, SP) = 76.2%**

[Behr et al., 2008]

Transcriptome Measurements for Improved Gene Finding



Idea: Improve mGene by using tiling array measurements to enhance recognition of exonic regions.

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

1. mGene (*ab initio*)
   transcript level performance: **mean(SN, SP) = 76.0%**
2. with tiling arrays *(young leaves)*
   transcript level performance: **mean(SN, SP) = 80.4%**
3. with tiling arrays *(inflorescence)*
   transcript level performance: **mean(SN, SP) = 77.1%**
4. with tiling arrays *(root)*
   transcript level performance: **mean(SN, SP) = 76.2%**

[Behr et al., 2008]

Transcriptome Measurements for Improved Gene Finding



Idea: Improve mGene by using tiling array measurements to enhance recognition of exonic regions.

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

1. mGene (*ab initio*)
   transcript level performance: **mean(SN, SP) = 76.0%**
2. with tiling arrays *(young leaves)*
   transcript level performance: **mean(SN, SP) = 80.4%**
3. with tiling arrays *(inflorescence)*
   transcript level performance: **mean(SN, SP) = 77.1%**
4. with tiling arrays *(root)*
   transcript level performance: **mean(SN, SP) = 76.2%**

[Behr et al., 2008]

Conclusions

- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Improved accuracy
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
- Combination of predictions and transcriptome measurements
  - Lead to improved gene finding
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

Conclusions

- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Improved accuracy
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
- Combination of predictions and transcriptome measurements
  - Lead to improved gene finding
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

Conclusions

- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Improved accuracy
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
- Combination of predictions and transcriptome measurements
  - Lead to improved gene finding
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

Acknowledgments

Gene Finding

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

Tiling Arrays

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

More Information

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

Thank you!

Acknowledgments

Gene Finding

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

Tiling Arrays

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

More Information

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

Thank you!

Acknowledgments

Gene Finding

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

Tiling Arrays

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

More Information

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

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.

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.

A. Coghlan, T.J. Fiedler, S.J. McKay, P. Flicek, T.W. Harris, D. Blasiar, The nGASP Consortium, and L.D. Stein. ngasp: the nematode genome annotation assessment project. *BMC Bioinformatics*, 2008. submitted.

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.

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.

References II

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.

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. 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 2008a.

Gabriele Schweikert, Christian Widmer, Bernhard Schölkopf, and Gunnar Rätsch. An empirical analysis of domain adaptation algorithms. In *Proc. NIPS 2008*, Advances in Neural Information Processing Systems, 2008b. accepted.

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 Schweikert, P Philips, J Behr, and G Rätsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S7, 2007. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-8-S10-S7.

References III

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

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: $10.1101/\mathrm{gr}.070169.107$.

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.