Novel Inference Methods for Computational Transcriptome Analysis

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society

Tübingen, Germany

Pattern Recognition in Bioinformatics

September 7, 2009



Friedrich Miescher Laboratory
of the Max Planck Society



Discovery of the Nuclein

(Friedrich Miescher, 1869)









Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein" (Miescher, 1874) [Dahm, 2008]

Discovery of the Nuclein

(Friedrich Miescher, 1869)









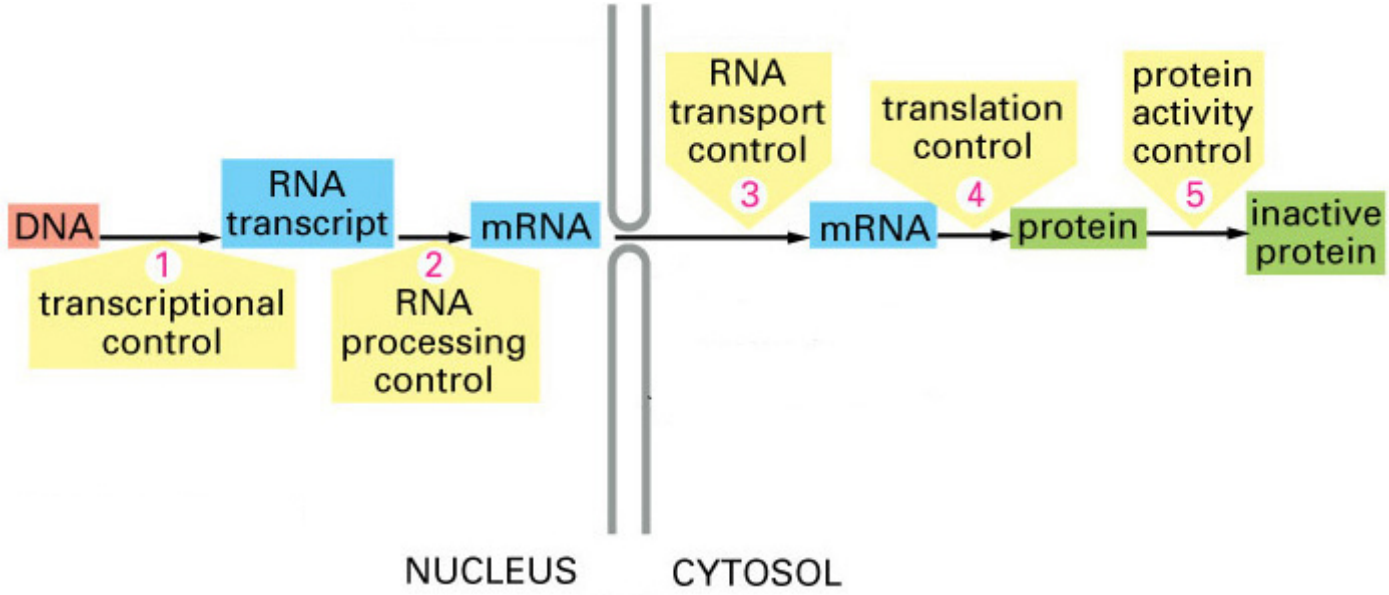
Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein" (Miescher, 1874) [Dahm, 2008]

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



Machine Learning View

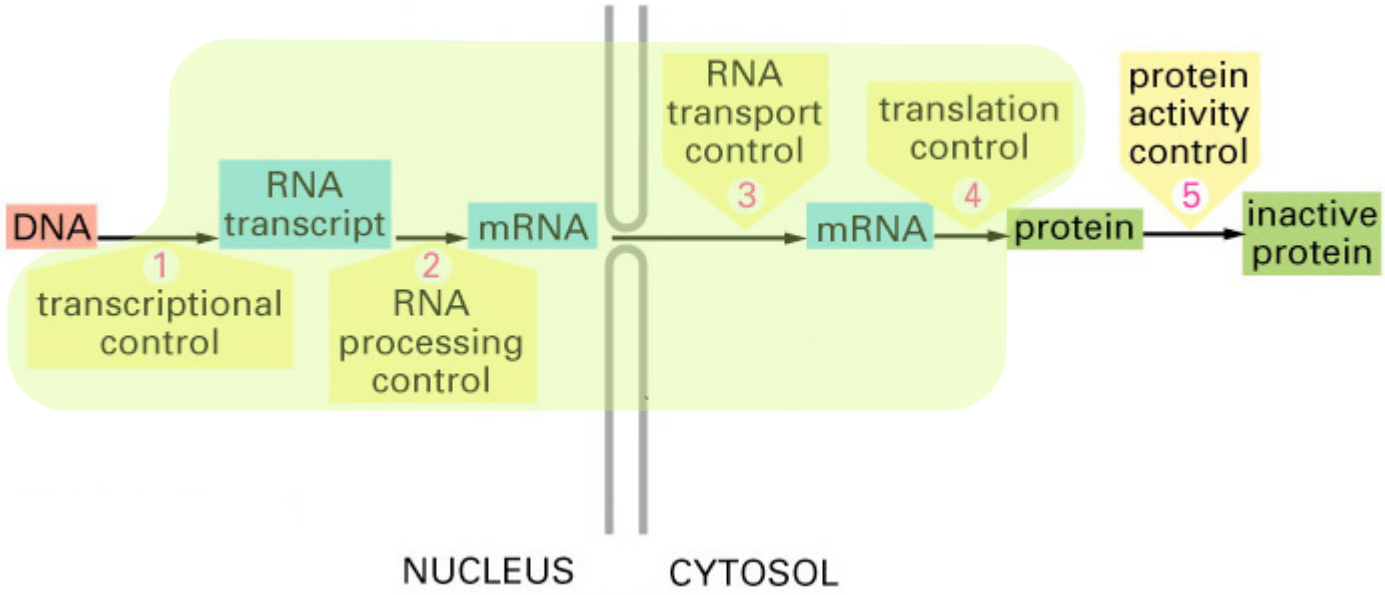
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

1. What can we not predict yet? What is missing?
2. Can we derive a deeper understanding of these processes?

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



Machine Learning View

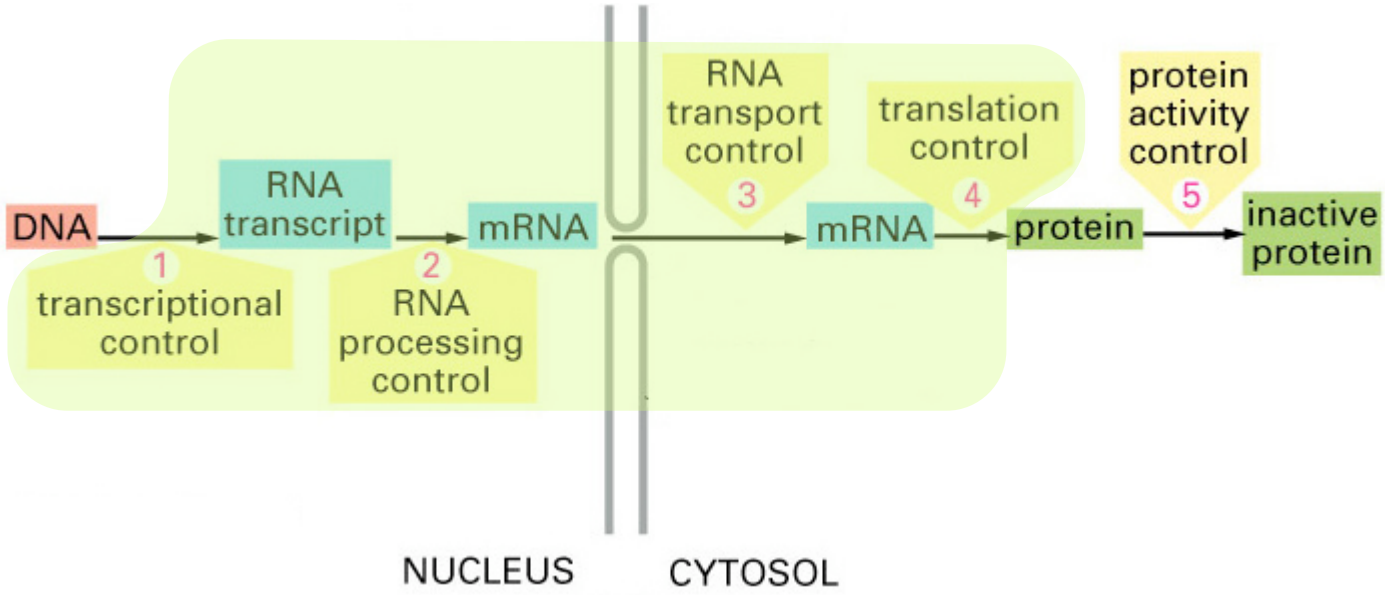
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

1. What can we not predict yet? What is missing?
2. Can we derive a deeper understanding of these processes?

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



Machine Learning View

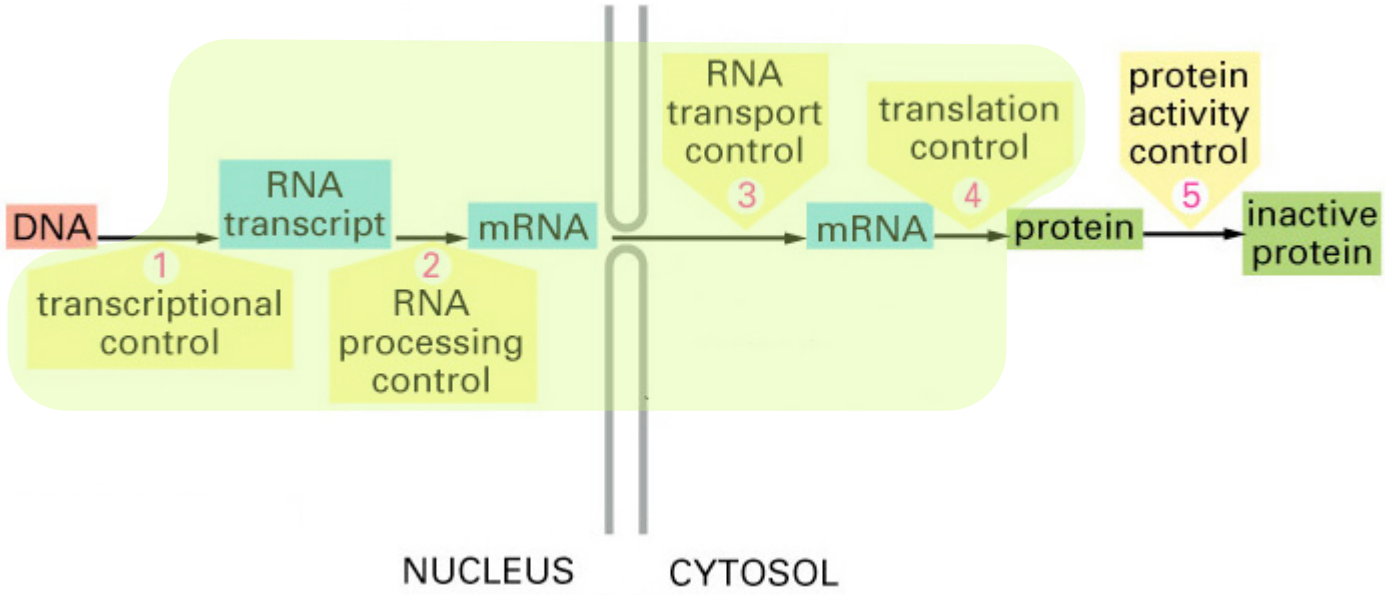
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

1. What can we not predict yet? What is missing?
2. Can we derive a deeper understanding of these processes?

Learning about the Transcriptome

⇝ What is encoded on the genome and how is it processed?



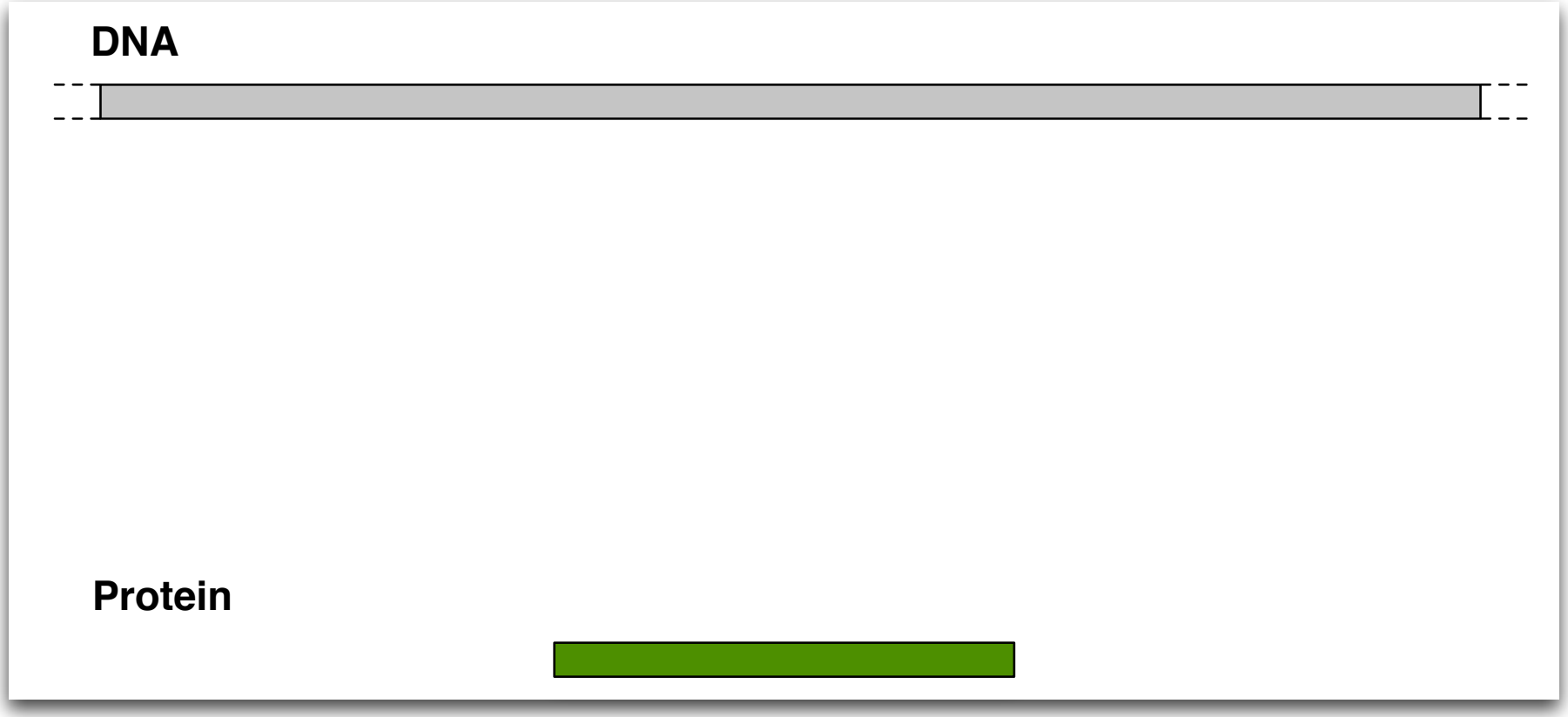
Machine Learning View

- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

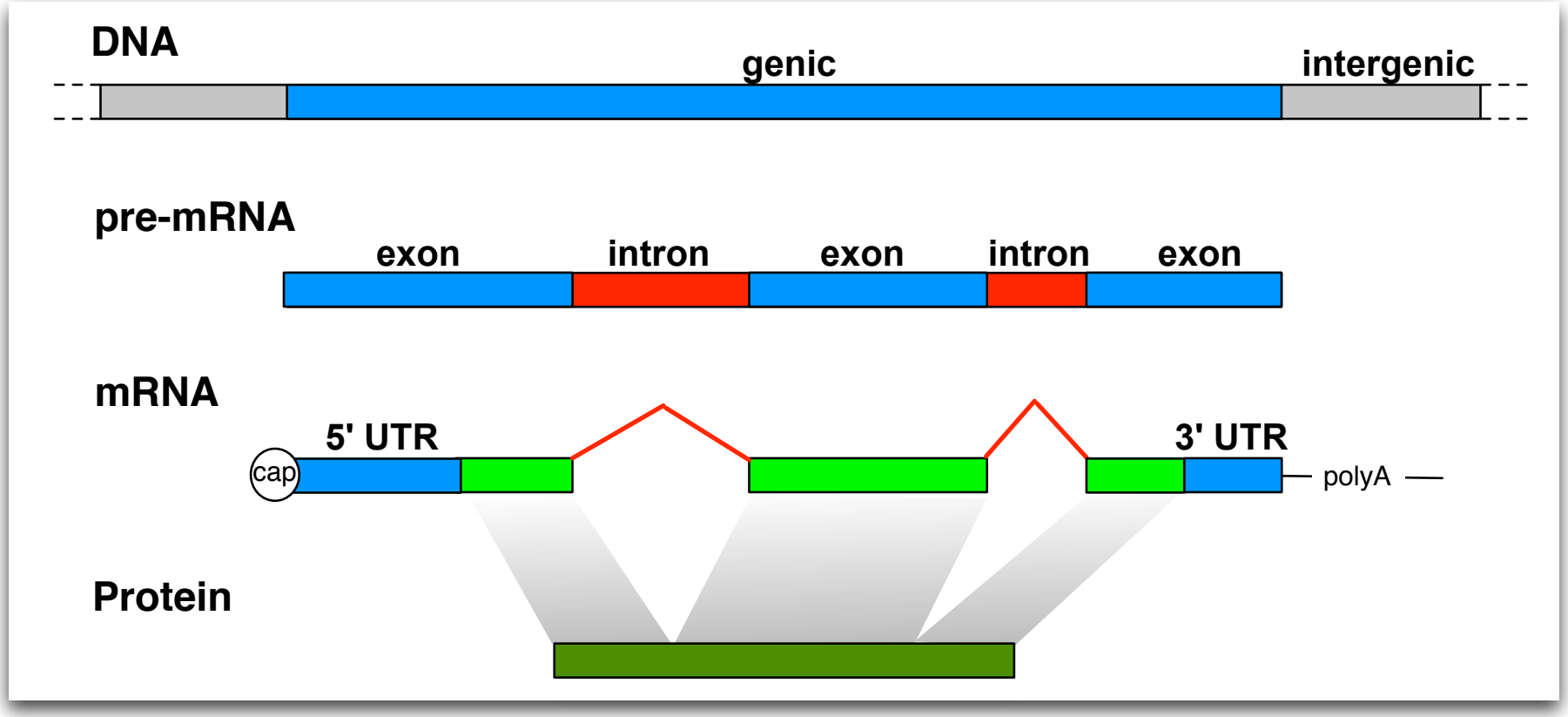
1. What can we not predict yet? What is missing?
2. Can we derive a deeper understanding of these processes?

1st Step: Computational Gene Finding



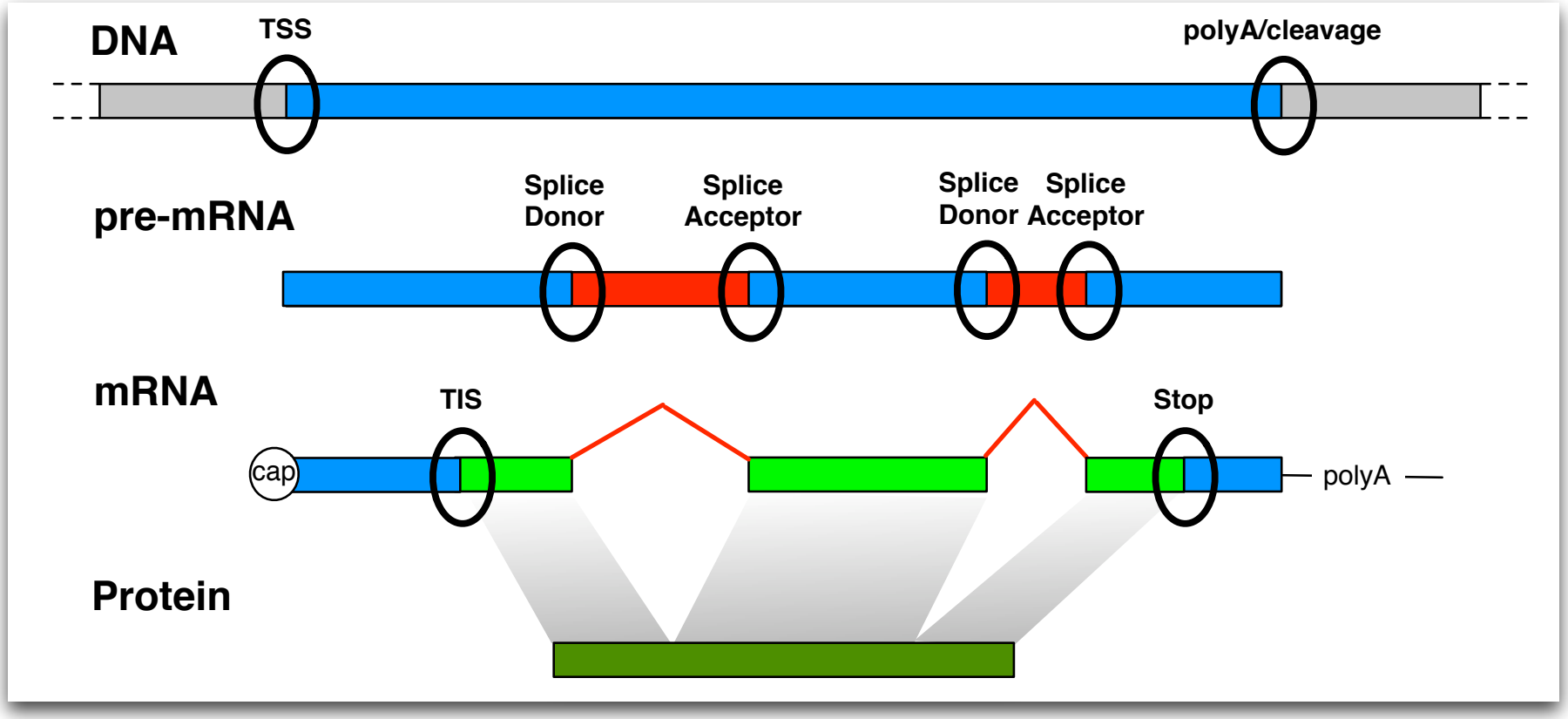
- Given a piece of DNA sequence
- Predict gene products

1st Step: Computational Gene Finding



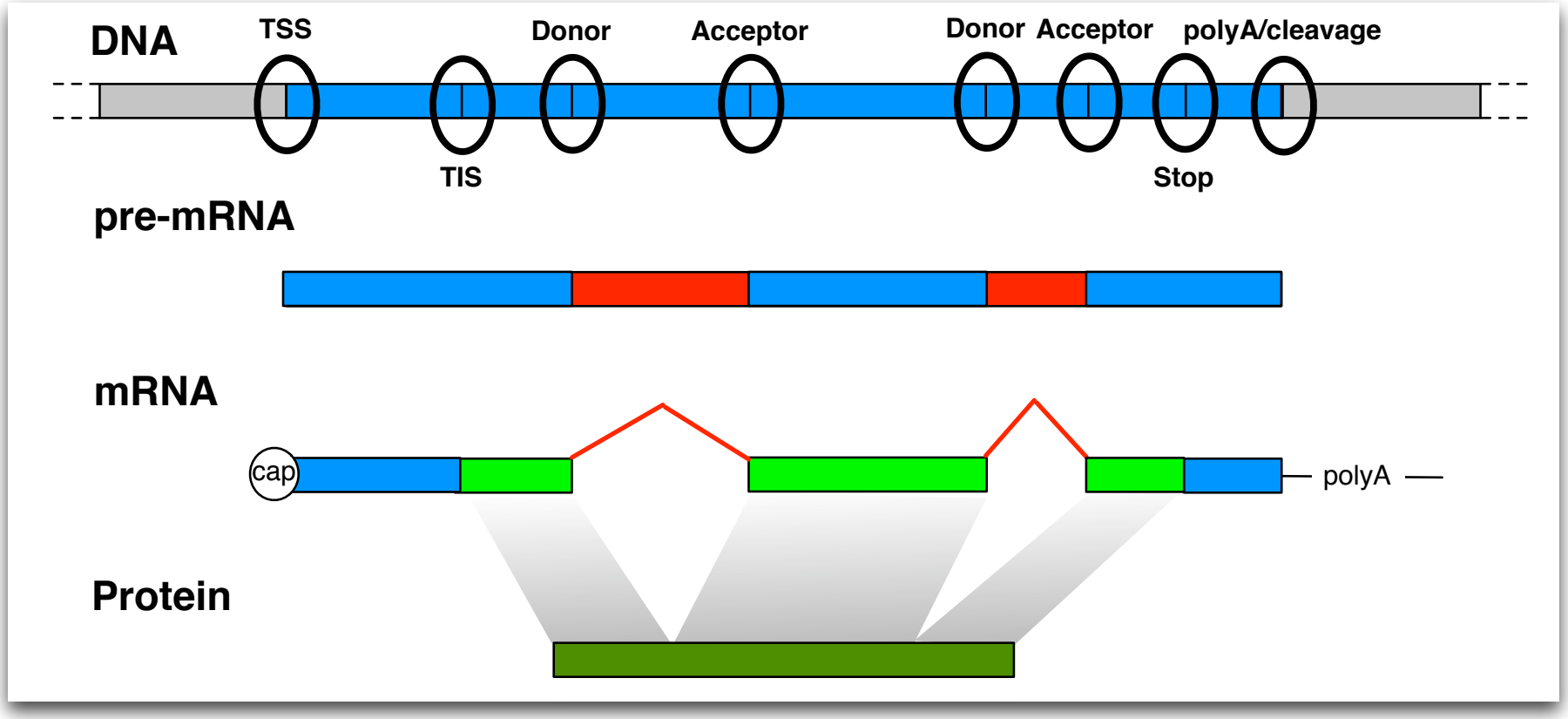
- Given a piece of DNA sequence
- Predict gene products inluding the intermediate processing steps

1st Step: Computational Gene Finding



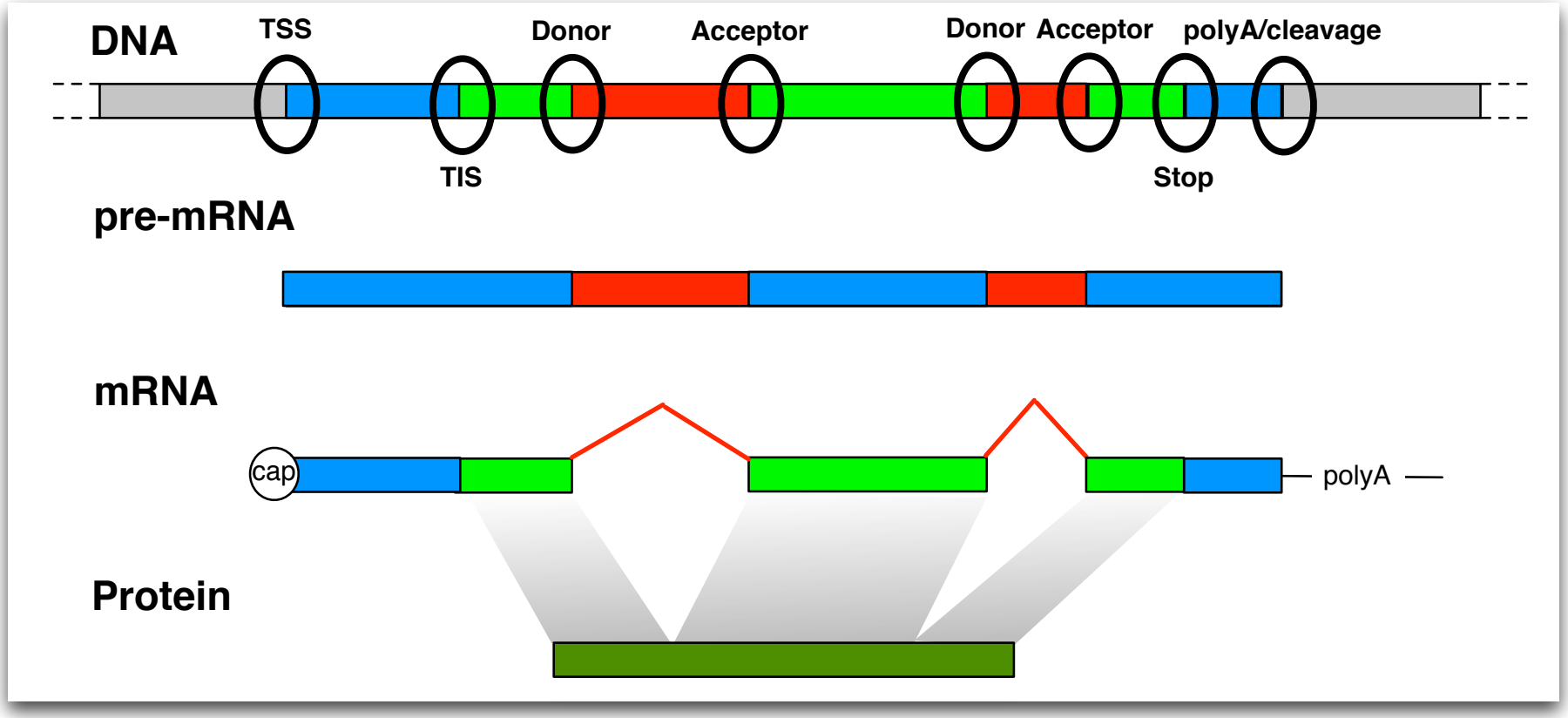
- Predict signals used during processing

1st Step: Computational Gene Finding



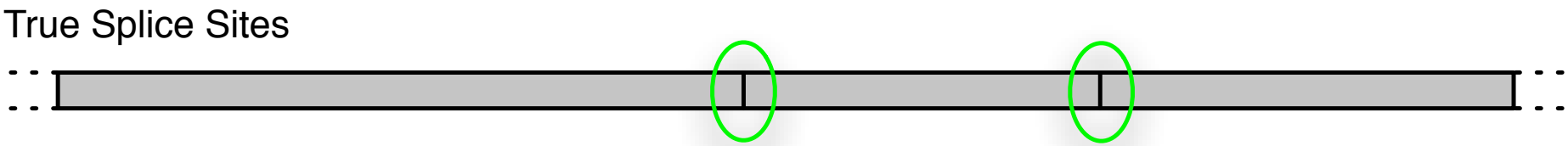
- Predict signals used during processing

1st Step: Computational Gene Finding

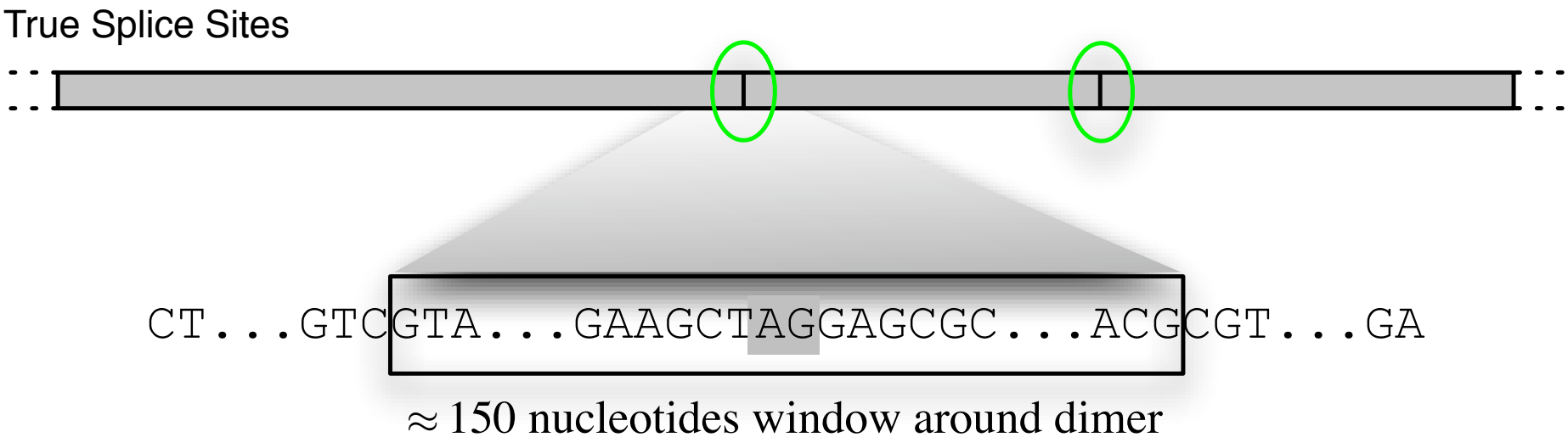


- Predict signals used during processing
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

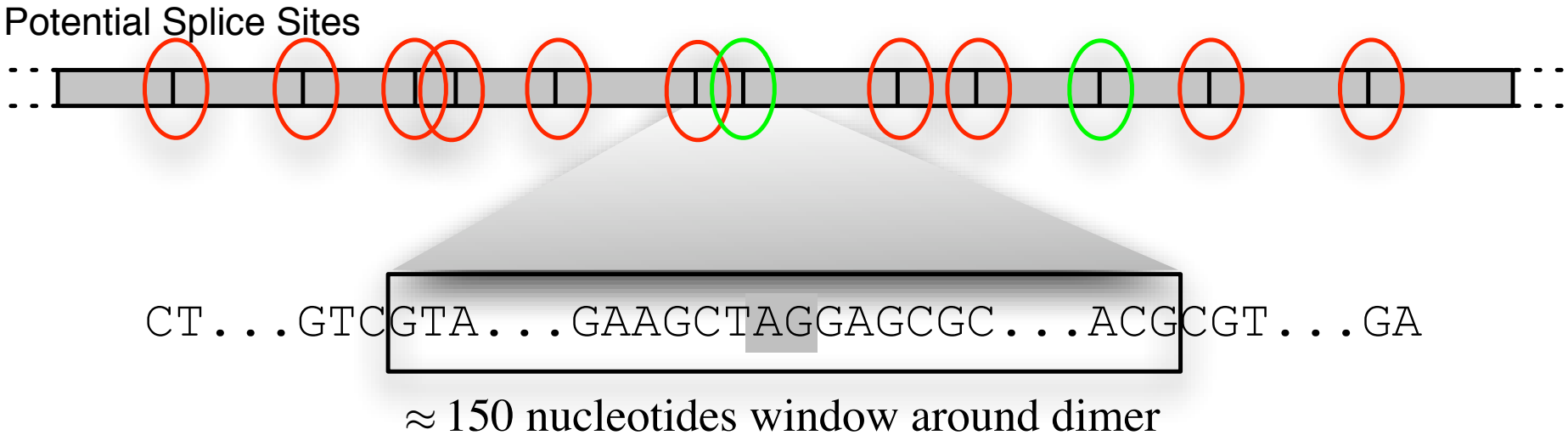
Example: Splice Site Recognition



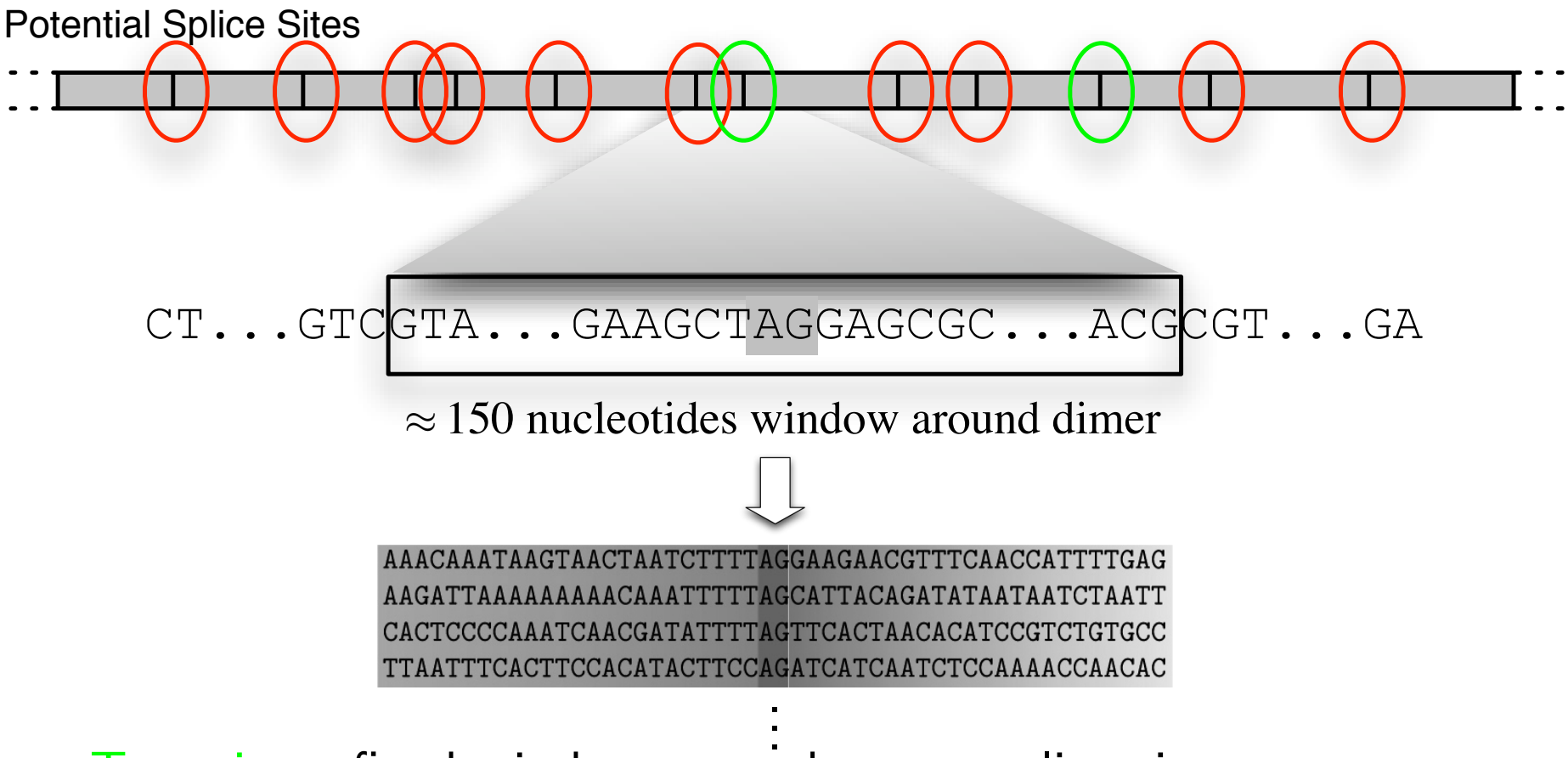
Example: Splice Site Recognition



Example: Splice Site Recognition



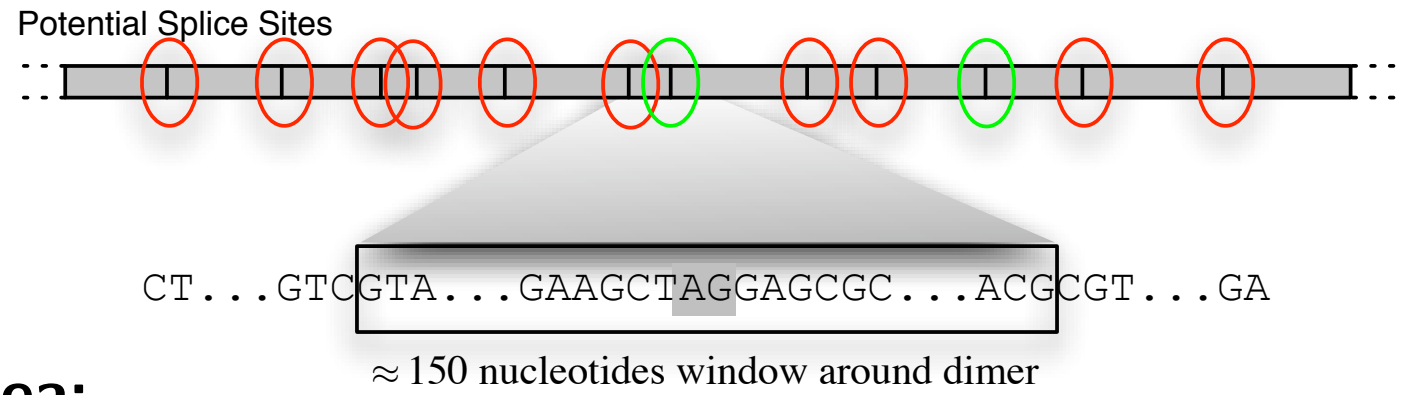
Example: Splice Site Recognition



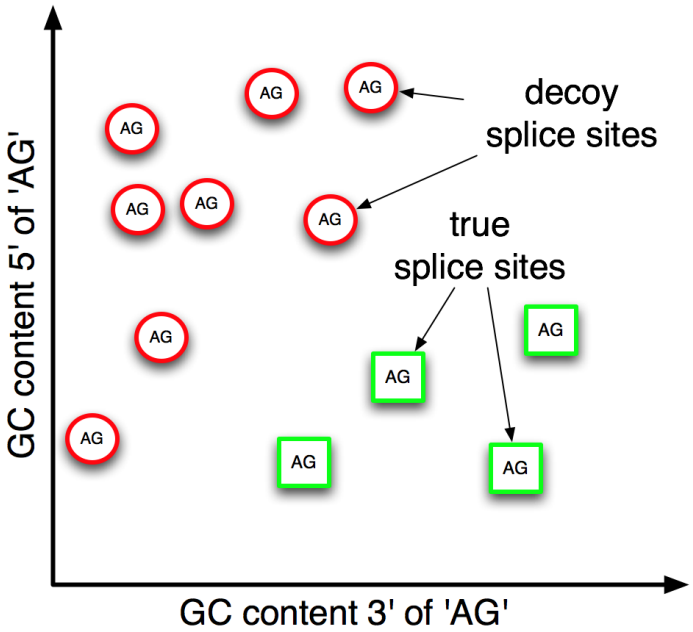
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

$\Rightarrow$ **Millions of examples from EST databases**

Example: Splice Site Recognition



Basic idea:



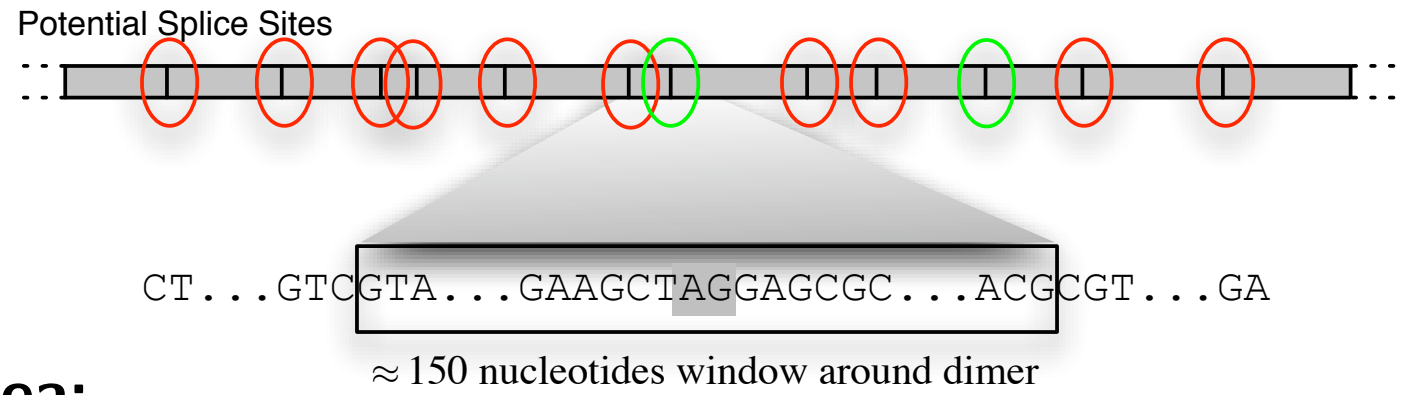
For instance, exploit that exons have higher GC content

or

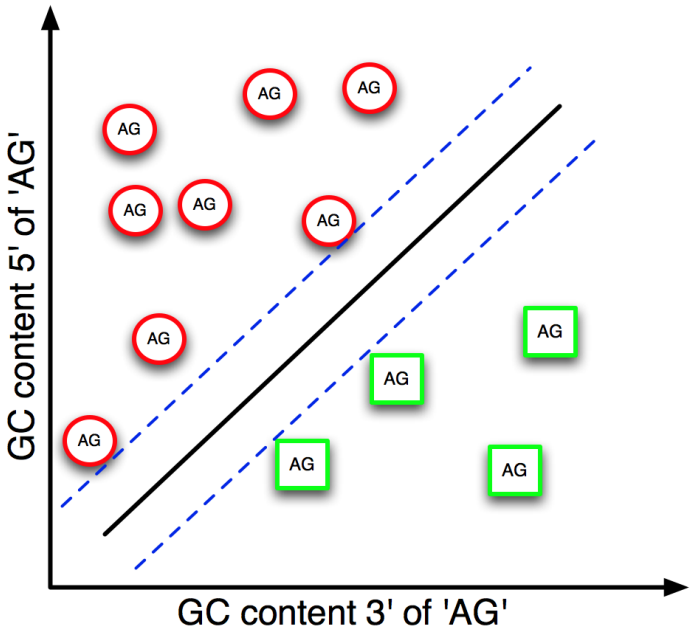
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007b]

Example: Splice Site Recognition



Basic idea:

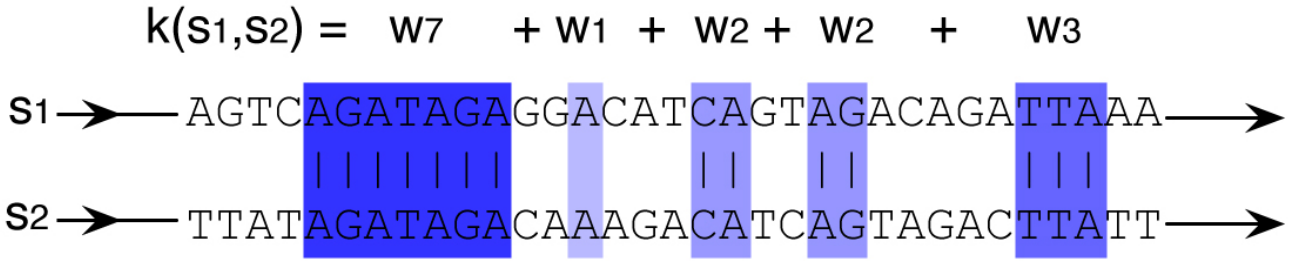


In practice: Use one feature per possible substring (e.g. $\leq$20) at all positions

$150 \cdot (4^1 + \ldots + 4^{20}) \approx 2 \cdot 10^{14}$ features

Kernels and String Data Structures

Good news: Use **WD kernel** to implicitly consider features:



[Rätsch et al., 2007, Rätsch and Sonnenburg, 2004]

Drawback: Training too expensive for millions of examples

Needs more sophisticated algorithms:

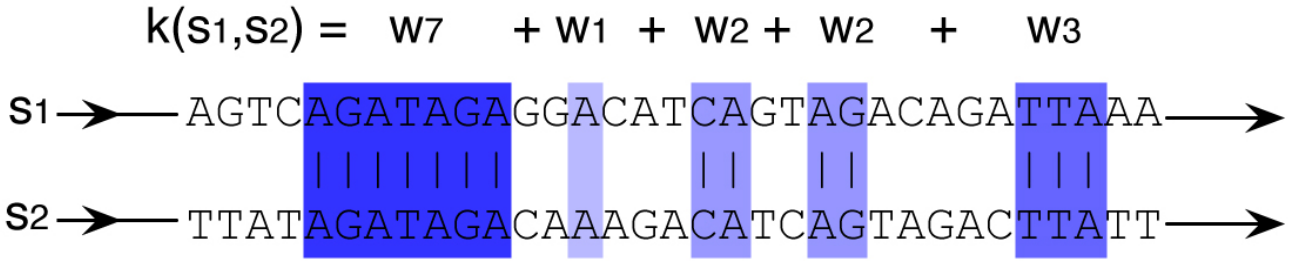
- Exploit that features are sparse and can be explicitly represented
  - String indexing data structures [Sonnenburg et al., 2007a]
- Use novel optimization techniques
  - Use subgradient-based optimization [Franc and Sonnenburg, 2009]

Implemented in a freely available software package

www.shogun-toolbox.org **[Sonnenburg et al., 2007a]**

Kernels and String Data Structures

Good news: Use **WD kernel** to implicitly consider features:



[Rätsch et al., 2007, Rätsch and Sonnenburg, 2004]

Drawback: Training is too expensive for millions of examples

Needs more sophisticated algorithms:

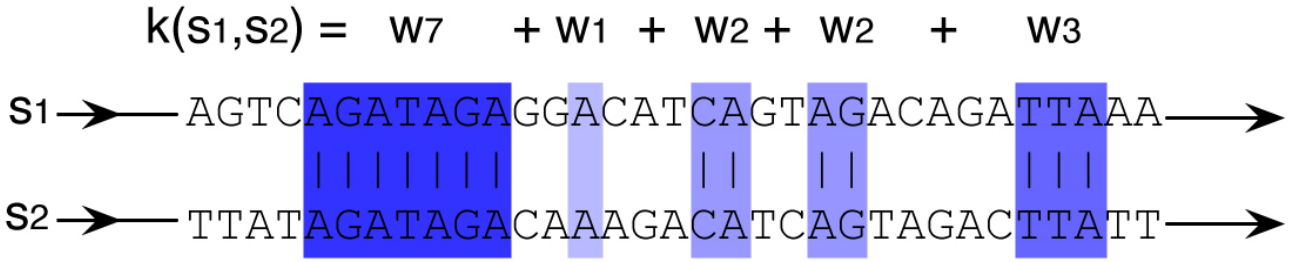
- Exploit that features are sparse and can be explicitly represented
  - String indexing data structures [Sonnenburg et al., 2007a]
- Use novel optimization techniques
  - Use subgradient-based optimization [Franc and Sonnenburg, 2009]

Implemented in a freely available software package

www.shogun-toolbox.org [Sonnenburg et al., 2007a]

Kernels and String Data Structures

Good news: Use **WD kernel** to implicitly consider features:



[Rätsch et al., 2007, Rätsch and Sonnenburg, 2004]

Drawback: Training was too expensive for millions of examples

Needs more sophisticated algorithms:

- Exploit that features are sparse and can be explicitly represented
  - String indexing data structures [Sonnenburg et al., 2007a]
- Use novel optimization techniques
  - Use subgradient-based optimization [Franc and Sonnenburg, 2009]

Implemented in a freely available software package

www.shogun-toolbox.org **[Sonnenburg et al., 2007a]**

Machine Learning Open Source Software



mlss.org

Contains more than 200 open software packages implementing machine learning algorithms.

jmlr.csail.mit.edu/mloss

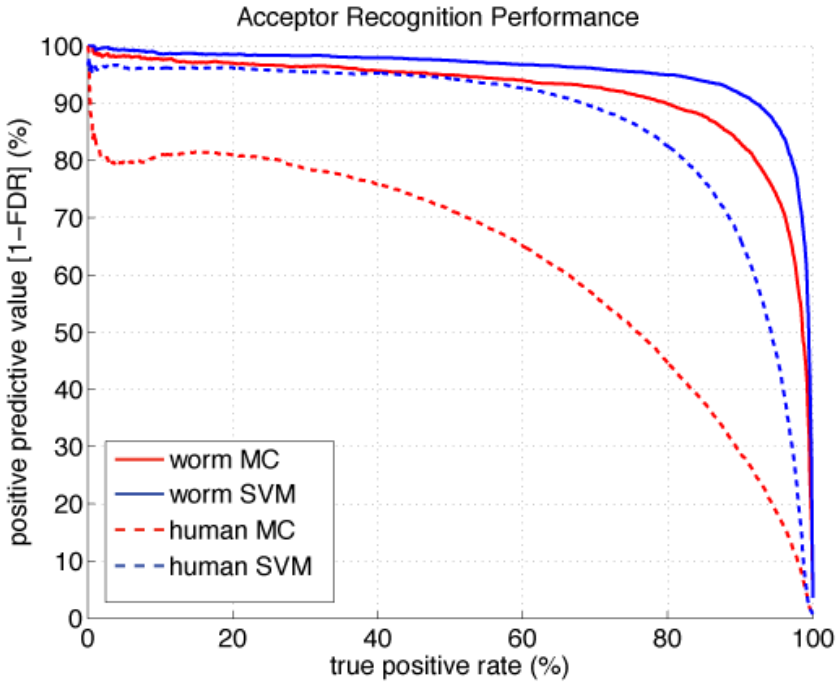
JMLR supports track on open source software, e.g.

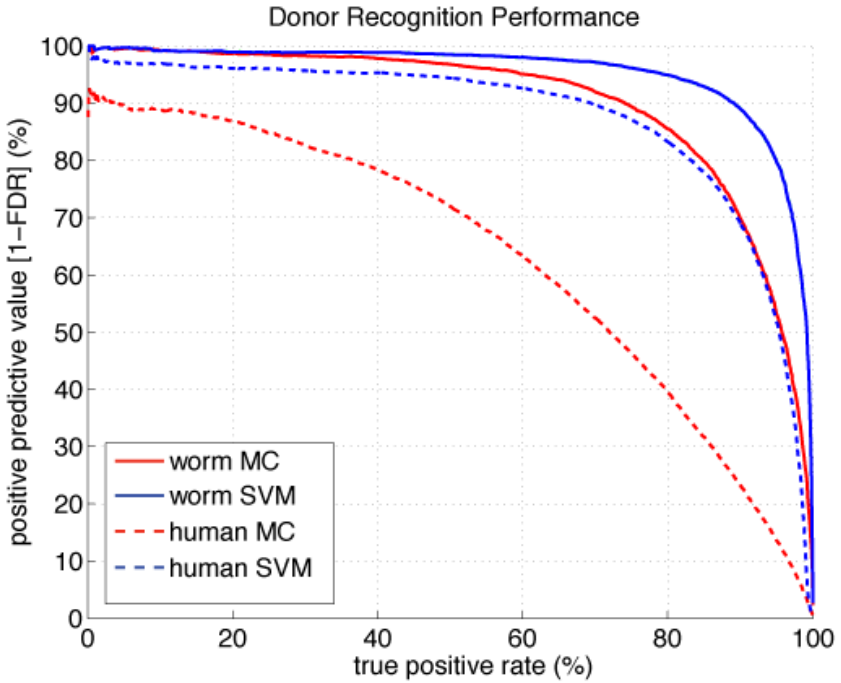
- implementations of machine learning algorithms
- toolboxes
- languages for scientific computing

Check out the JMLR paper "The Need for Open Source Software in Machine Learning" by Sonnenburg et al. [Sonnenburg et al., 2007c]

Results on Splice Site Recognition

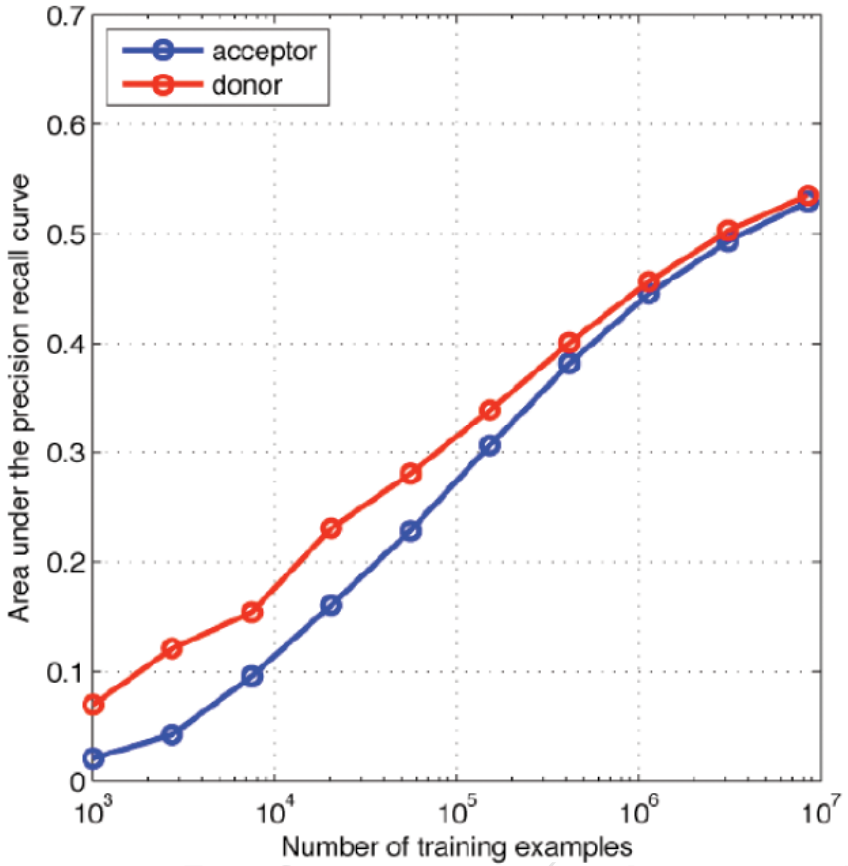
	Worm		Fly		Cress		Fish		Human	
	Acc	Don	Acc	Don	Acc	Don	Acc	Don	Acc	Don
Markov Chain auPRC(%)	92.1	90.0	80.3	78.5	87.4	88.2	63.6	62.9	16.2	26.0
SVM auPRC(%)	95.9	95.3	86.7	87.5	92.2	92.9	86.6	86.9	54.4	56.9

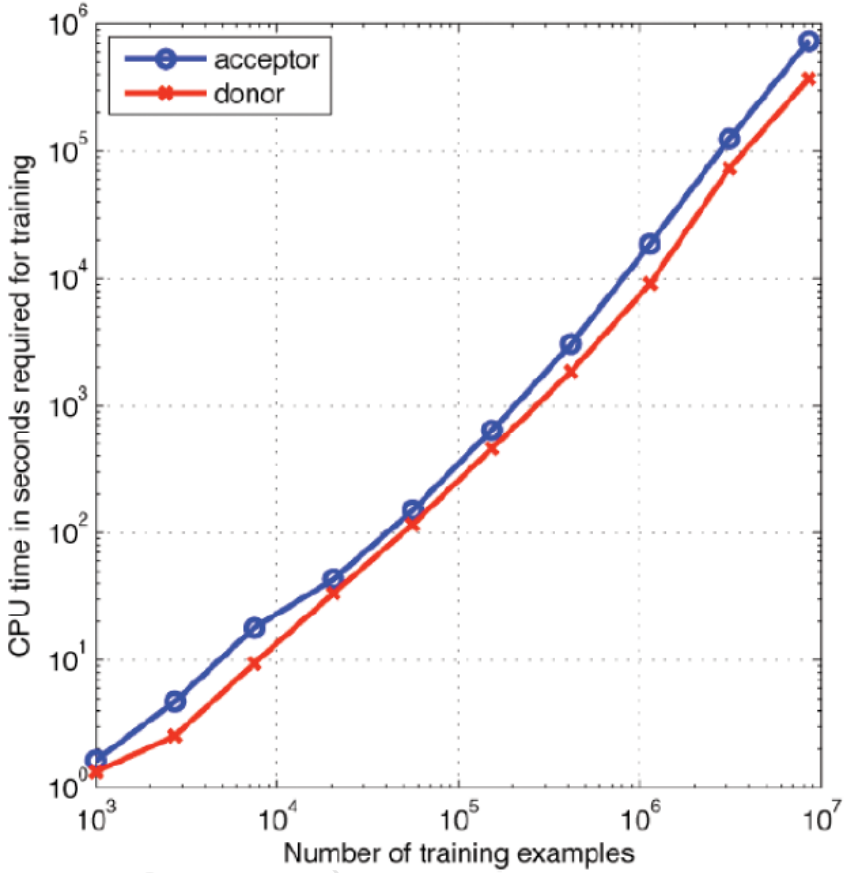




[Sonnenburg, Schweikert, Philips, Behr, Rätsch, 2007]

Human Splice Sites with WD Kernel



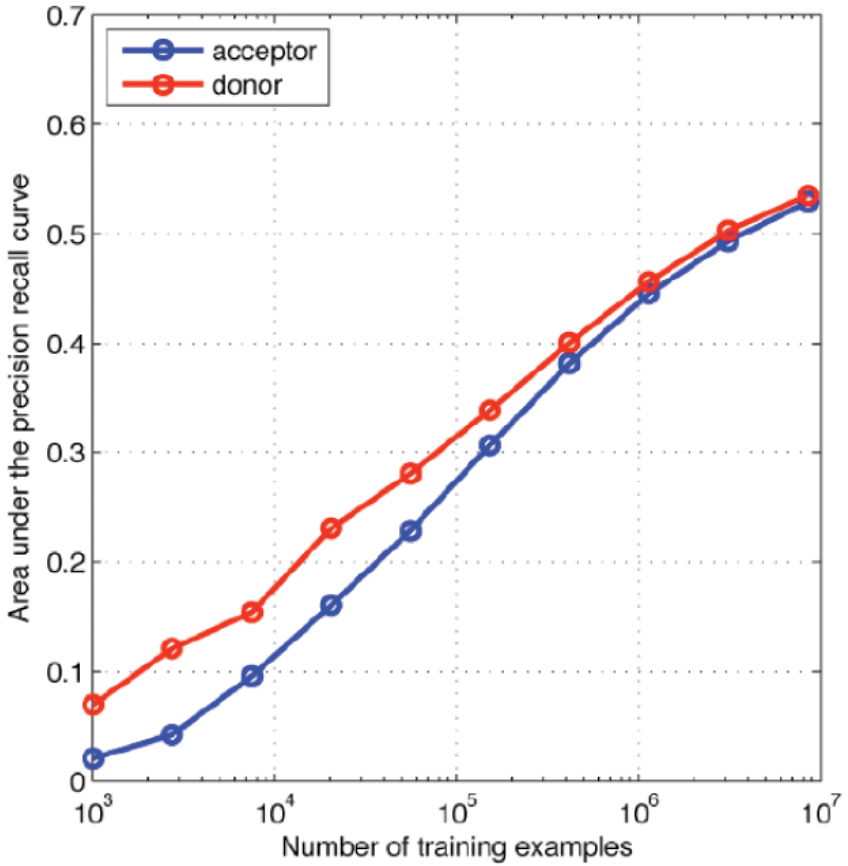


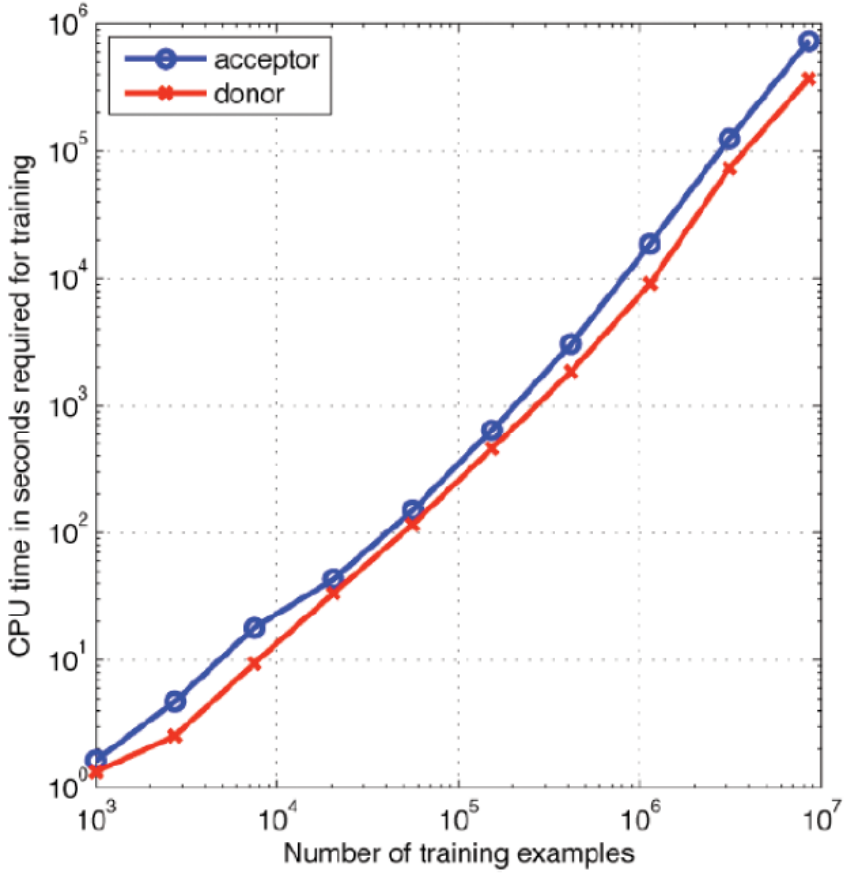
Performance (=Area under PRC curve) keeps increasing.

⇒ **Pays off to use all available data!**

[Sonnenburg et al., 2007b]

Human Splice Sites with WD Kernel



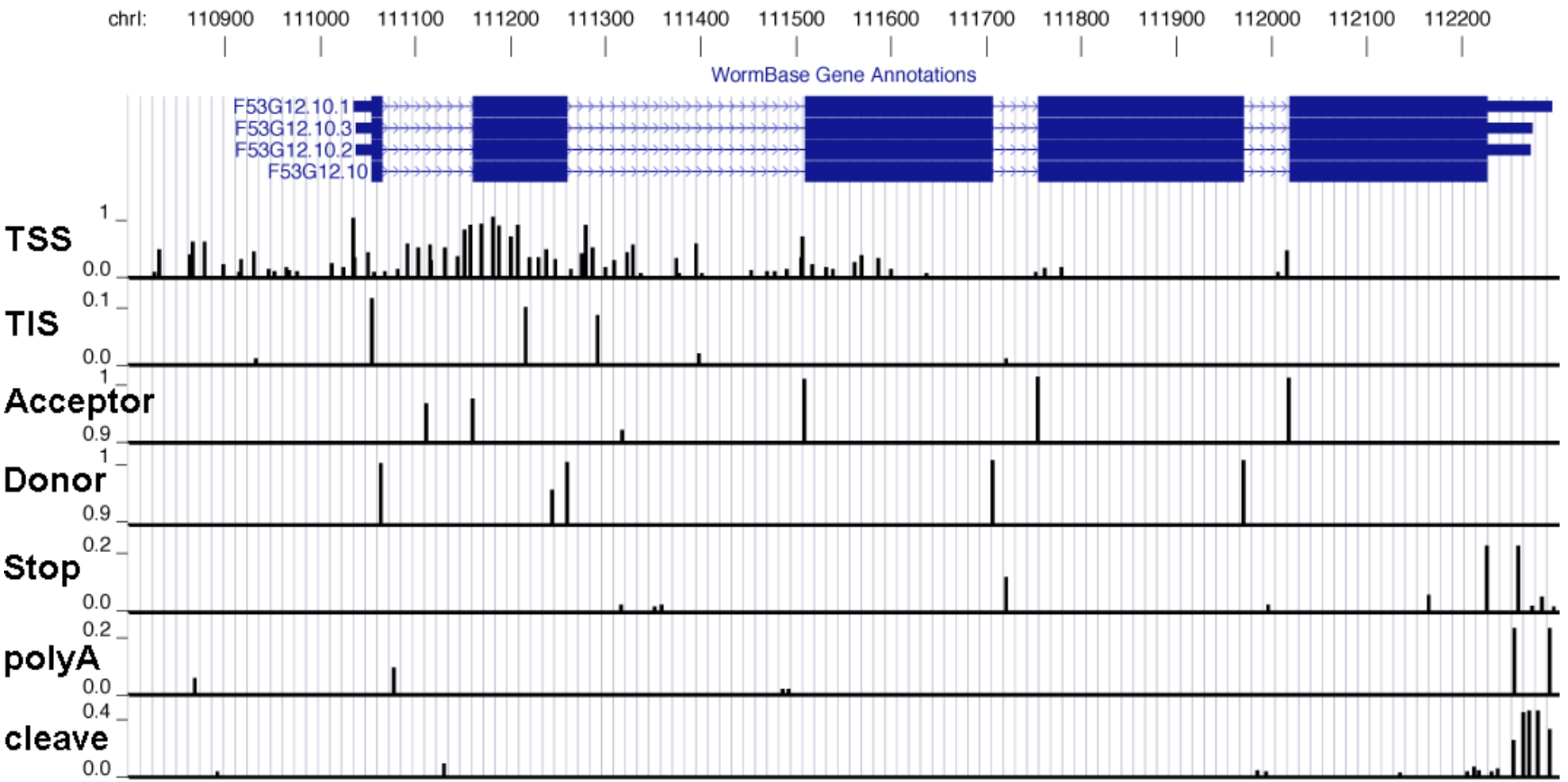


Performance (=Area under PRC curve) keeps increasing.

$\Rightarrow$ **Pays off to use all available data!**

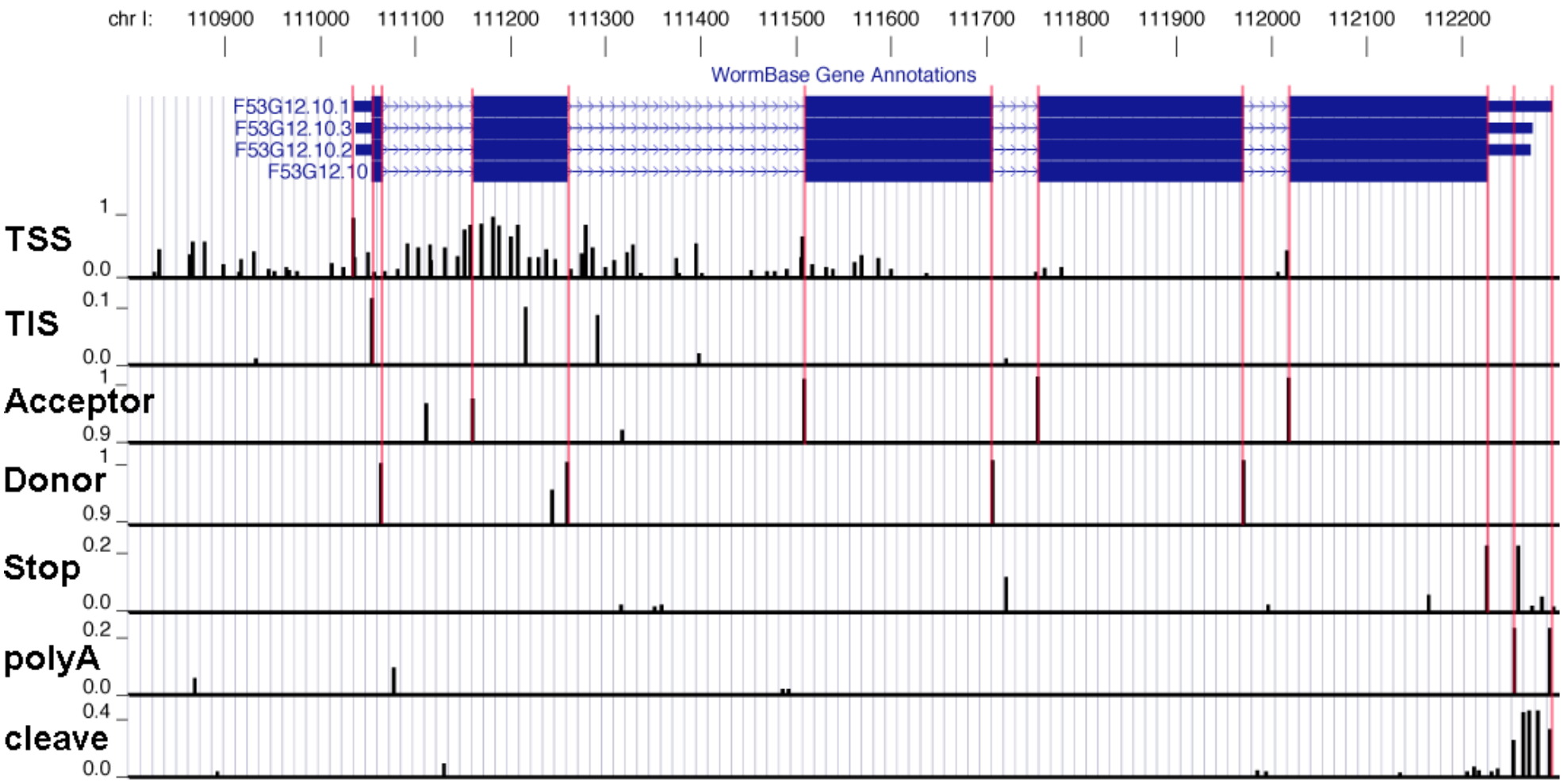
[Sonnenburg et al., 2007b]

Example: Predictions in UCSC Browser



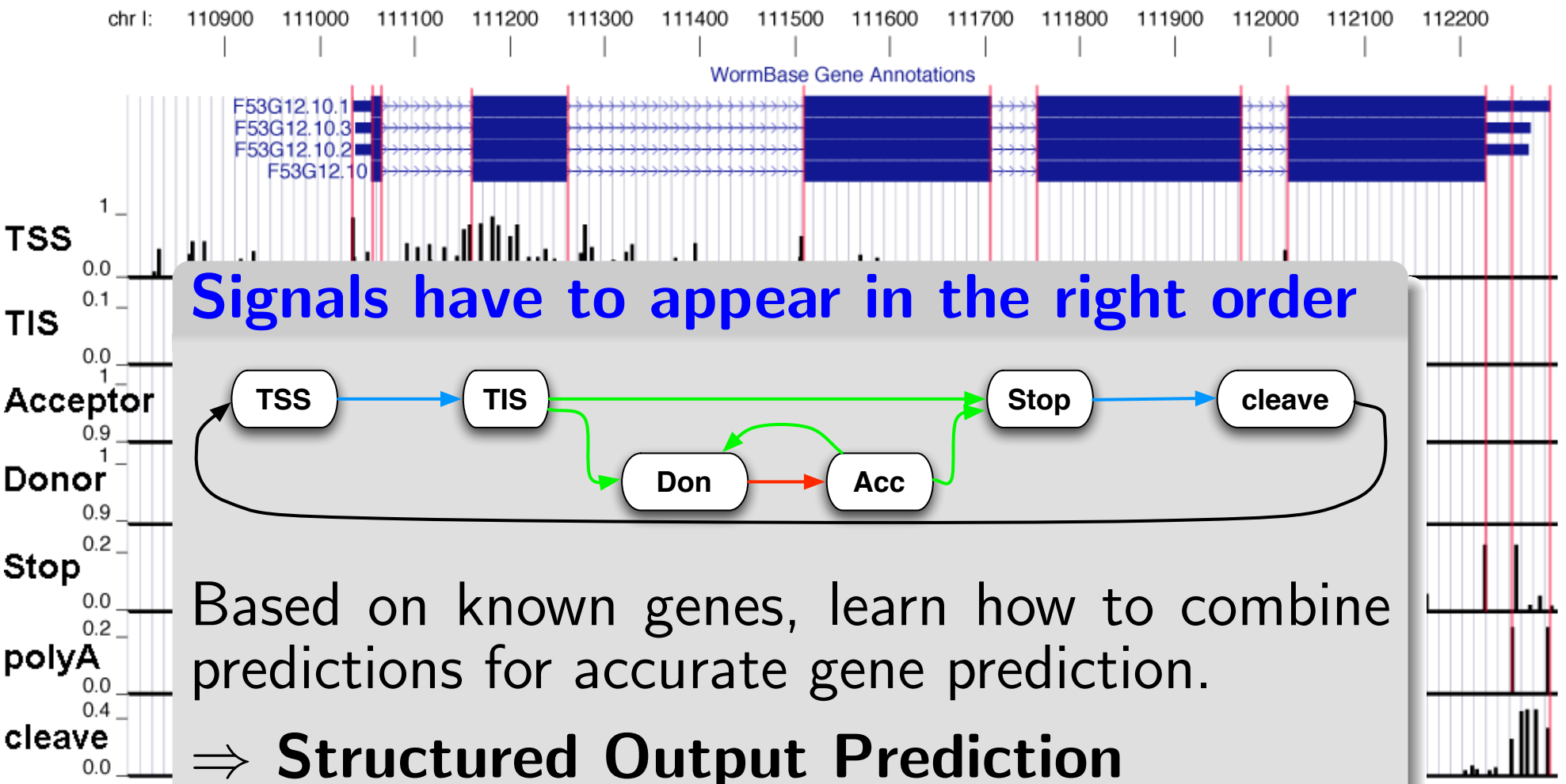
[Rätsch et al., 2007, Schweikert et al., 2009b]

Example: Predictions in UCSC Browser



[Rätsch et al., 2007, Schweikert et al., 2009b]

Example: Predictions in UCSC Browser



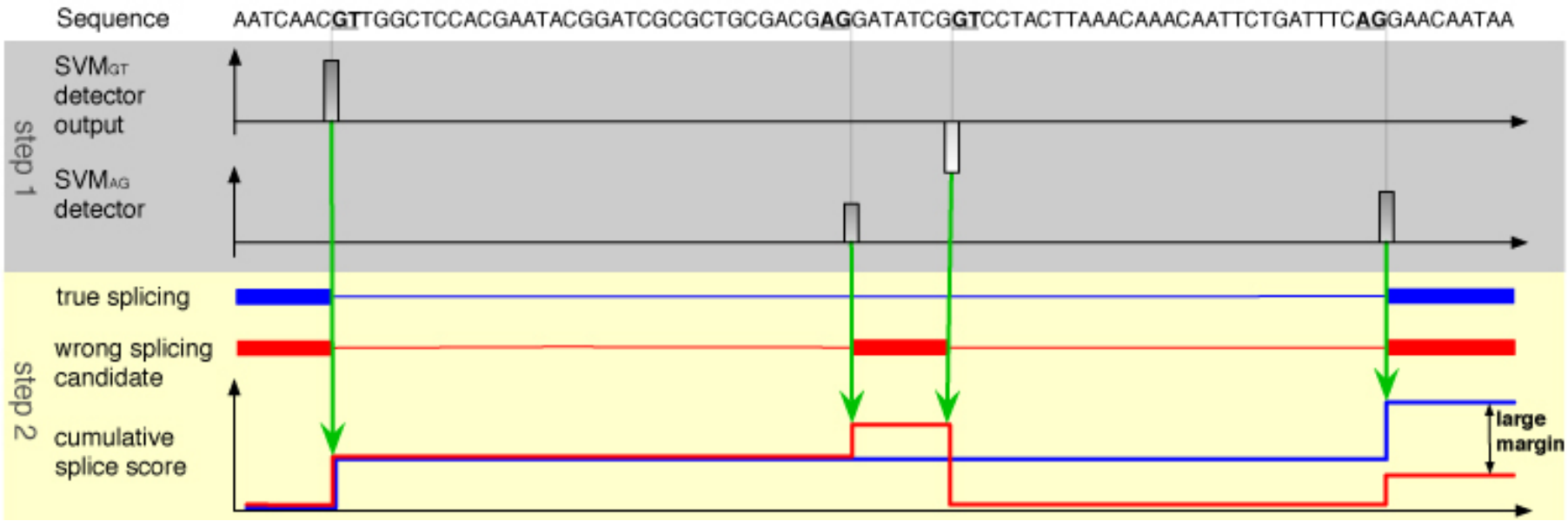
Signals have to appear in the right order

Based on known genes, learn how to combine predictions for accurate gene prediction.

⇒ **Structured Output Prediction**

[Rätsch et al., 2007, Schweikert et al., 2009b]

Discriminative Gene Prediction (simplified)



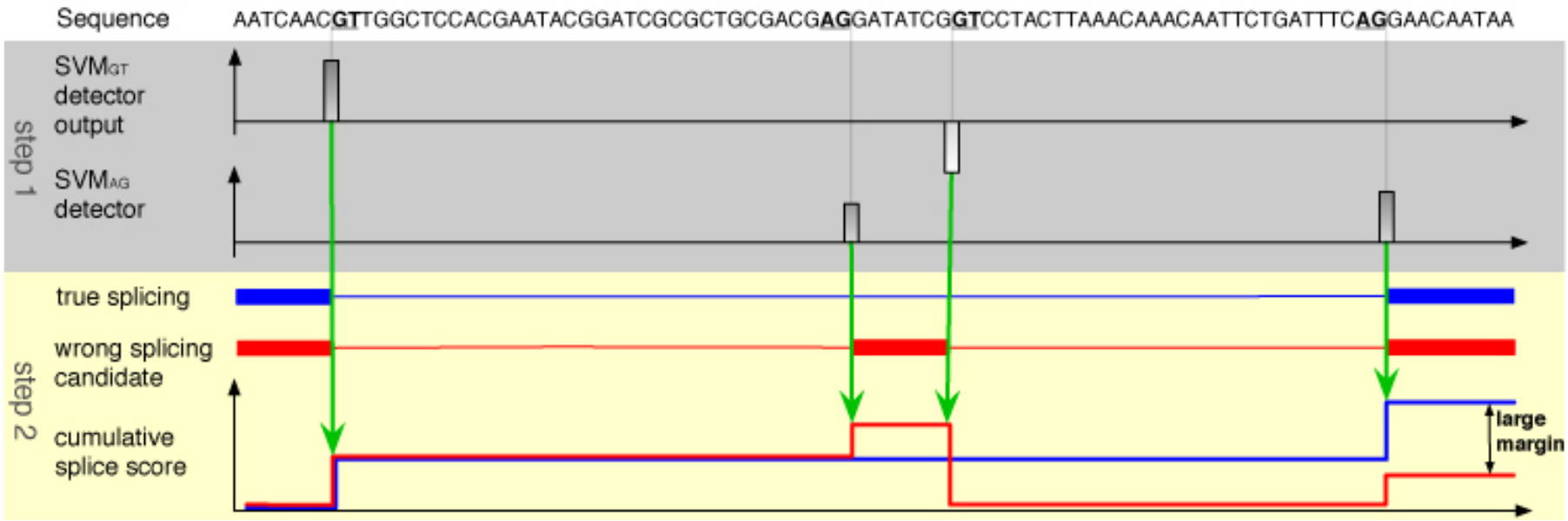
[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

$$F(\mathbf{y}) := \underbrace{\sum_{j=1}^{J-1} S_{GT}(f_j^{GT}) + \sum_{j=2}^{J} S_{AG}(f_j^{AG})}_{\text{Splice signals}} + \underbrace{\sum_{j=1}^{J-1} S_{L_I}(p_{j+1} - q_j) + \sum_{j=1}^{J} S_{L_E}(q_j - p_j)}_{\text{Segment lengths}}$$

Tune free parameters by solving **quadratic optimization problem**.

Discriminative Gene Prediction (simplified)



[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

$$F(\mathbf{y}) := \underbrace{\sum_{j=1}^{J-1} S_{GT}(f_j^{GT}) + \sum_{j=2}^{J} S_{AG}(f_j^{AG})}_{\text{Splice signals}} + \underbrace{\sum_{j=1}^{J-1} S_{L_I}(p_{j+1} - q_j) + \sum_{j=1}^{J} S_{L_E}(q_j - p_j)}_{\text{Segment lengths}}$$

Tune free parameters by solving **quadratic optimization problem**.

Discriminative Gene Prediction (simplified)



Schölkopf, 2007]

Simplified

$F(\mathbf{y}) := \sum_{j=1}^{J-1}$ … $p_j)$

Tune free … oblem.

Take-Home Message

- **Avoid solving a huge optimization problem by splitting it into two parts**
- **Exploit state-of-the-art signal predictions**
- **Predict more accurately than generative models**

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

	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., 2009b]

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 *ab initio* gene predictions [Schweikert et al., 2009a]

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

	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., 2009b]

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 *ab initio* gene predictions [Schweikert et al., 2009a]

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

	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., 2009b]

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 *ab initio* gene predictions [Schweikert et al., 2009a]

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

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

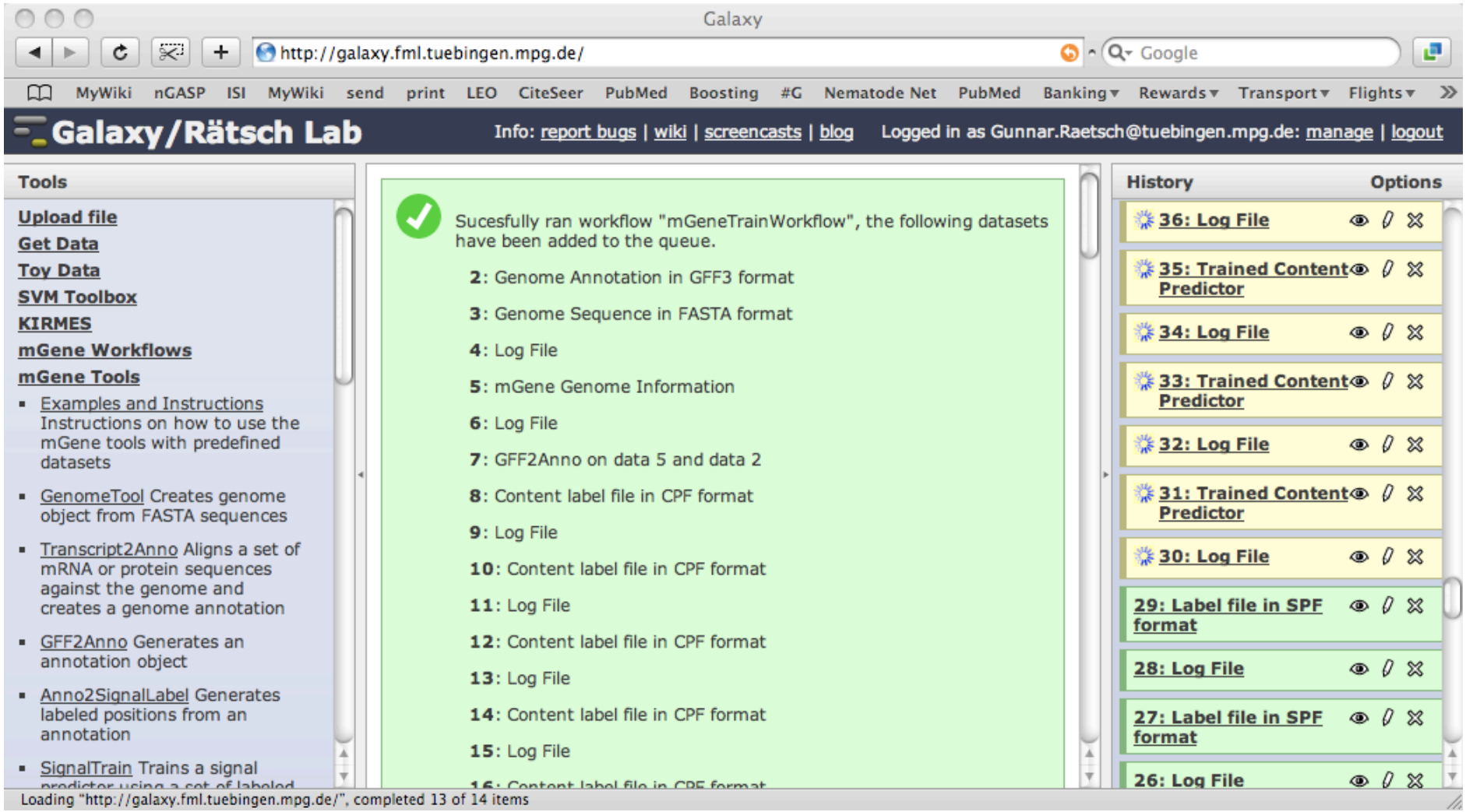
- Annotation of other nematode genomes: [Schweikert et al., 2009b]

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 *ab initio* gene predictions [Schweikert et al., 2009a]

mGene.web: Gene Finding for Everybody ;-)

(Schweikert et al., NAR Webserver Issue, 2009)



mgene.org/web

Limitations/Extensions

- Gene finding (by far) still not perfect
  - New *ab initio* techniques more accurate
  - Many genes mis-predicted (*C. elegans*: $\approx 50\%$; human: $\approx 80\%$)
  - What is missing?
    - Not enough examples for training?
    - Model complexity?
    - Other information?
- Transcriptome is the result of many complex processes
  - *Ab initio* methods cannot predict alternative transcripts
  - Current methods also ignore other important information:
    - Chromatin structure and methylation patterns
    - RNA structure and processing regulation ...

Limitations/Extensions

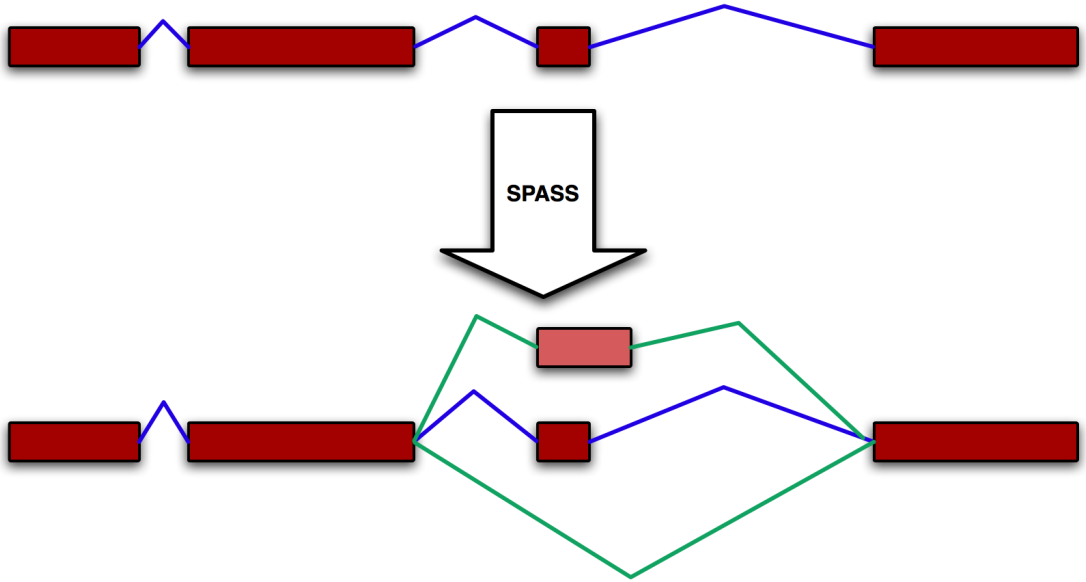
- Gene finding (by far) still not perfect
  - New *ab initio* techniques more accurate
  - Many genes mis-predicted (*C. elegans*: $\approx 50\%$; human: $\approx 80\%$)
  - What is missing?
    - Not enough examples for training?
    - Model complexity?
    - Other information?
- Transcriptome is the result of many complex processes
  - *Ab initio* methods cannot predict alternative transcripts
  - Current methods also ignore other important information:
    - Chromatin structure and methylation patterns
    - RNA structure and processing regulation ...

Limitations/Extensions

- Gene finding (by far) still not perfect
  - New *ab initio* techniques more accurate
  - Many genes mis-predicted (*C. elegans*: $\approx 50\%$; human: $\approx 80\%$)
  - What is missing?
    - Not enough examples for training?
    - Model complexity?
    - Other information?
- Transcriptome is the result of many complex processes
  - *Ab initio* methods cannot predict alternative transcripts
  - Current methods also ignore other important information:
    - Chromatin structure and methylation patterns
    - RNA structure and processing regulation ...

Alternative Transcripts: First Steps

Predictions of alternative splicing



- Predict novel alternative splicing as independent events
- Only use sequence information [Rätsch et al., 2005]
- Requires known gene structures: Use *ab initio* predictions

Alternative Transcripts: More Steps



- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:

    Input: DNA sequence

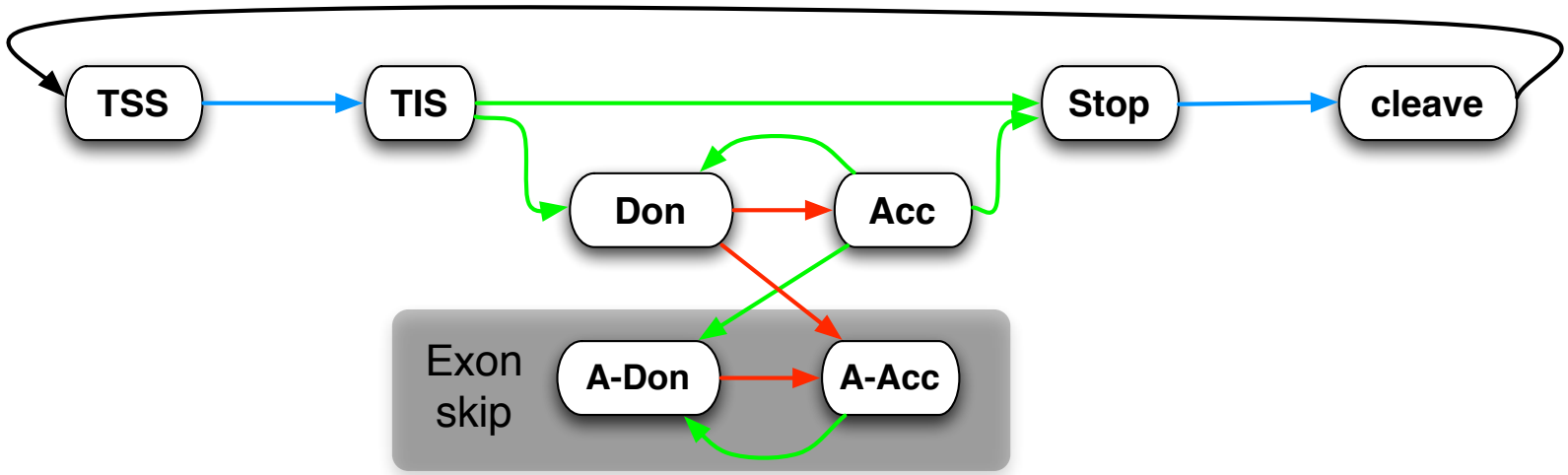
    Output: Splicing graph

- mGene approach can be extended to
  - Predict **"simple splicing graphs"**

- But, without **clean data** we cannot train the model ...

Alternative Transcripts: More Steps

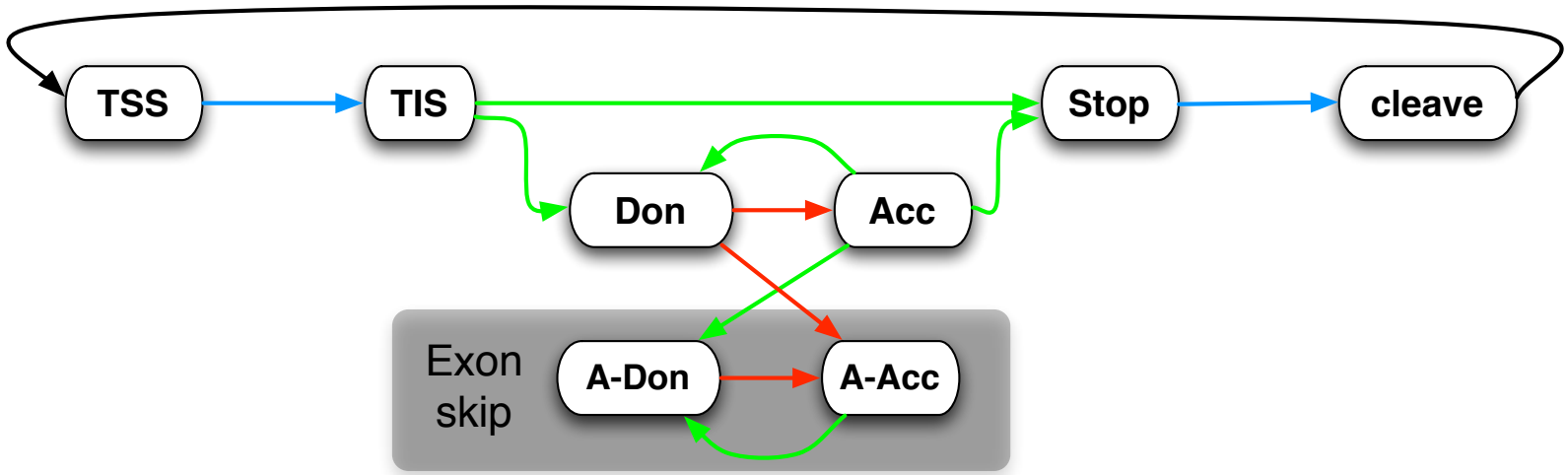
- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:
    Input: DNA sequence
    Output: Splicing graph
- mGene approach can be extended to
  - Predict **"simple splicing graphs"**



- But, without **clean data** we cannot train the model ...

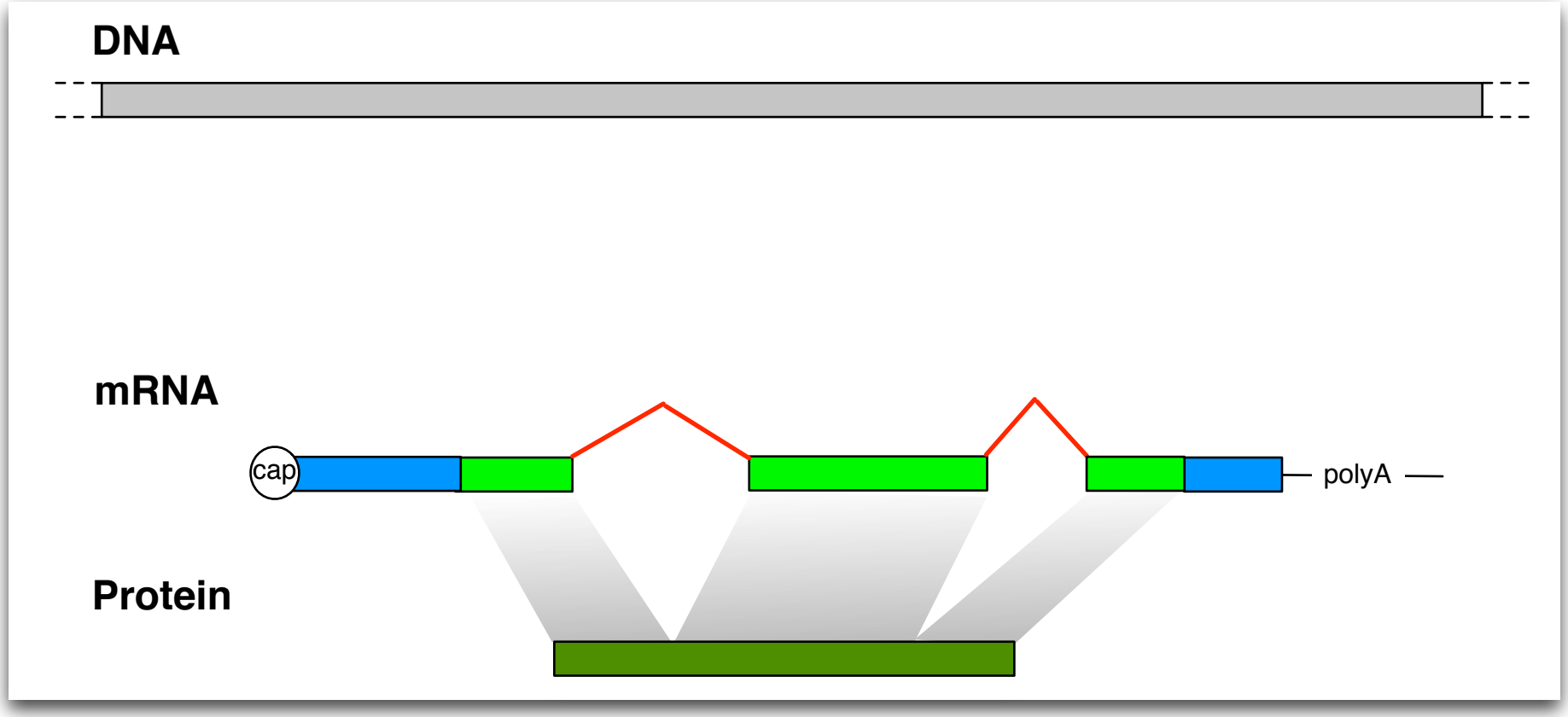
Alternative Transcripts: More Steps

- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:
    Input: DNA sequence
    Output: Splicing graph
- mGene approach can be extended to
  - Predict **"simple splicing graphs"**



- But, without **clean data** we cannot train the model ...

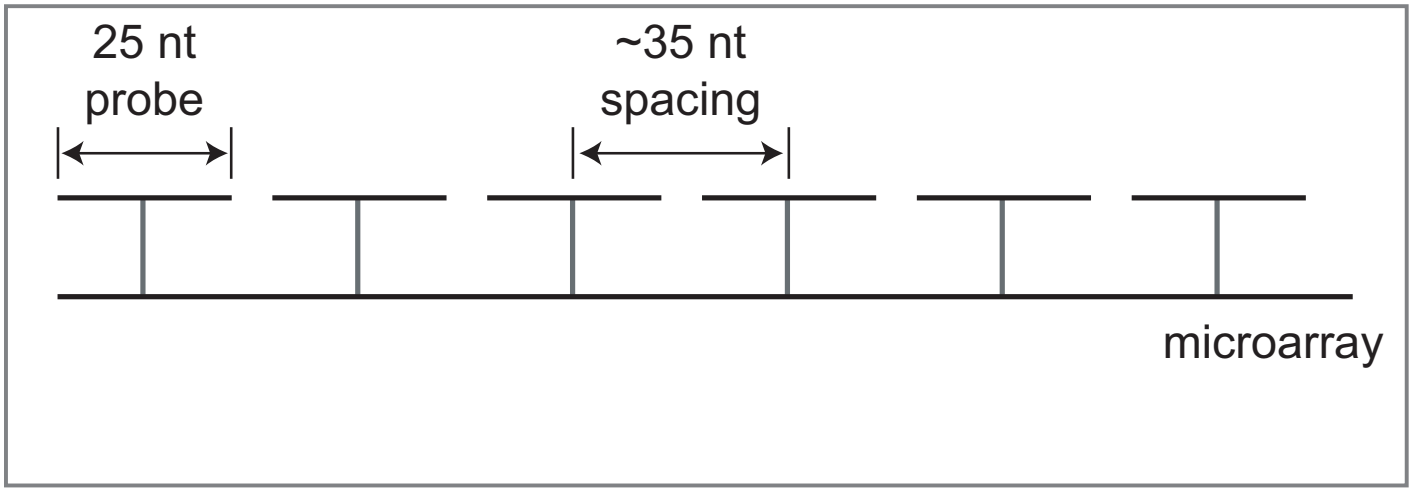
From Transcriptome to Products



Directly measure the transcriptome

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

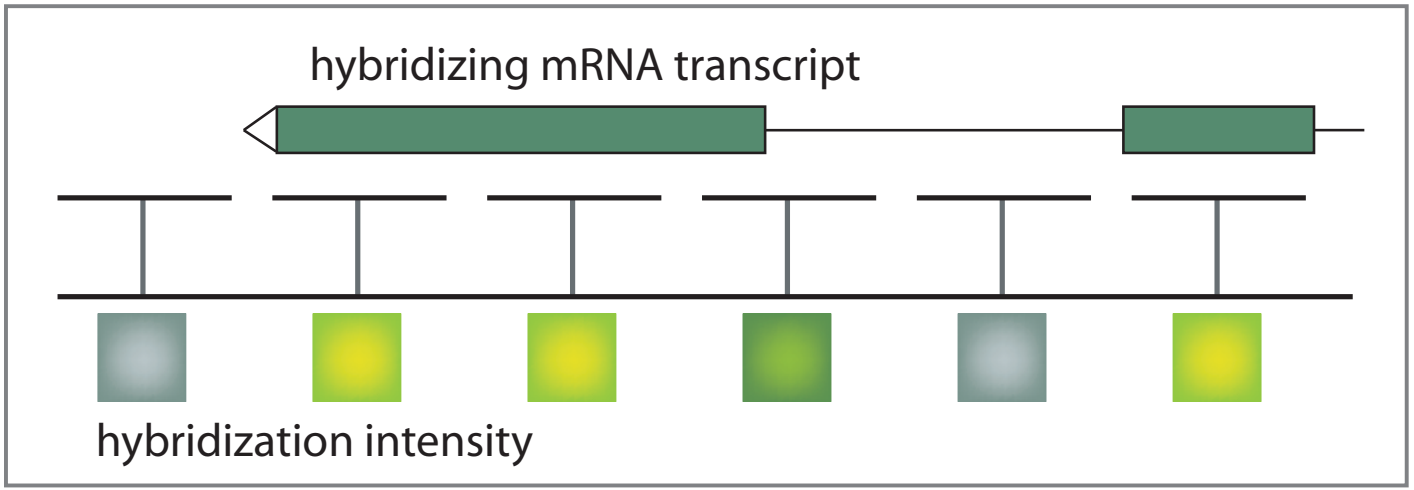
Tiling Arrays for Transcriptome Analysis



- 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)

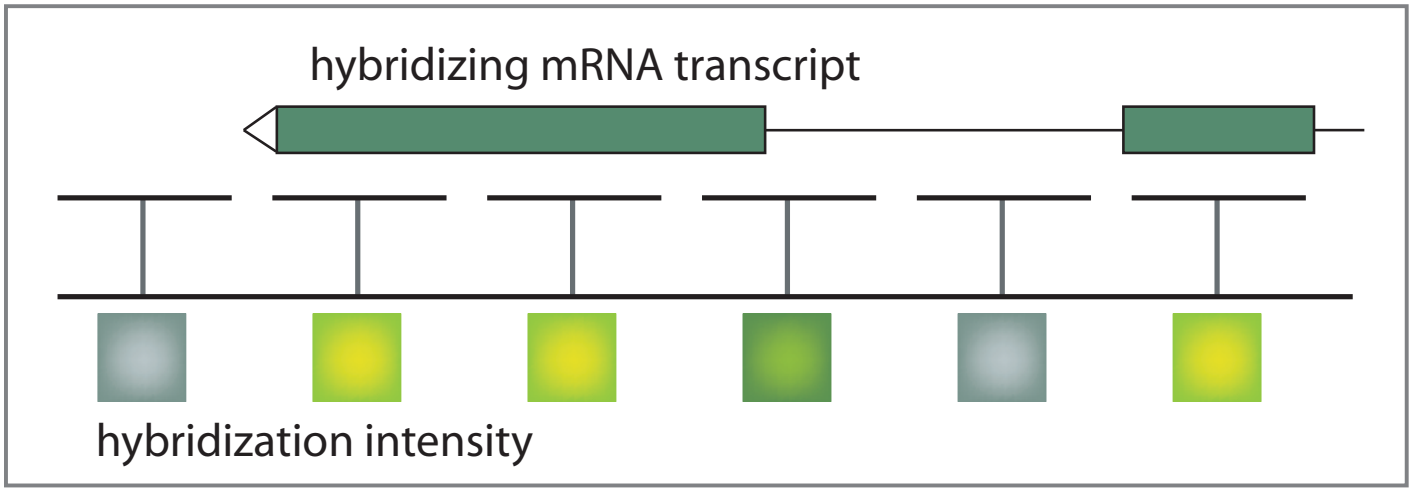
Tiling Arrays for Transcriptome Analysis



- 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 traditional microarrays

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

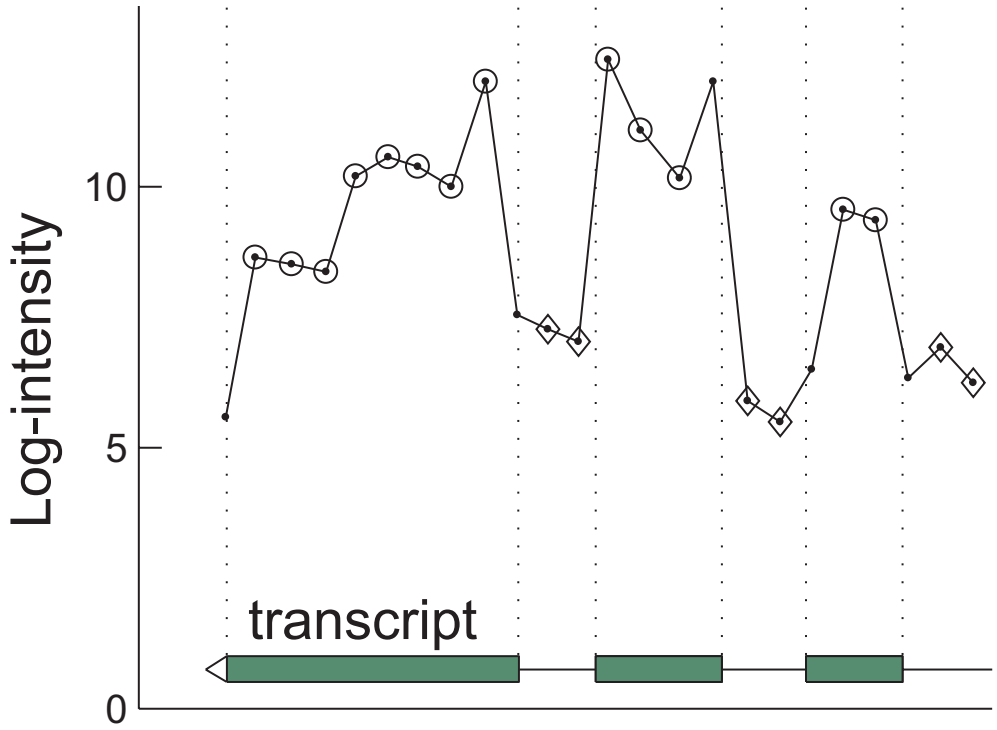
Tiling Arrays for Transcriptome Analysis



- 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]

Intensities are Noisy Measurements

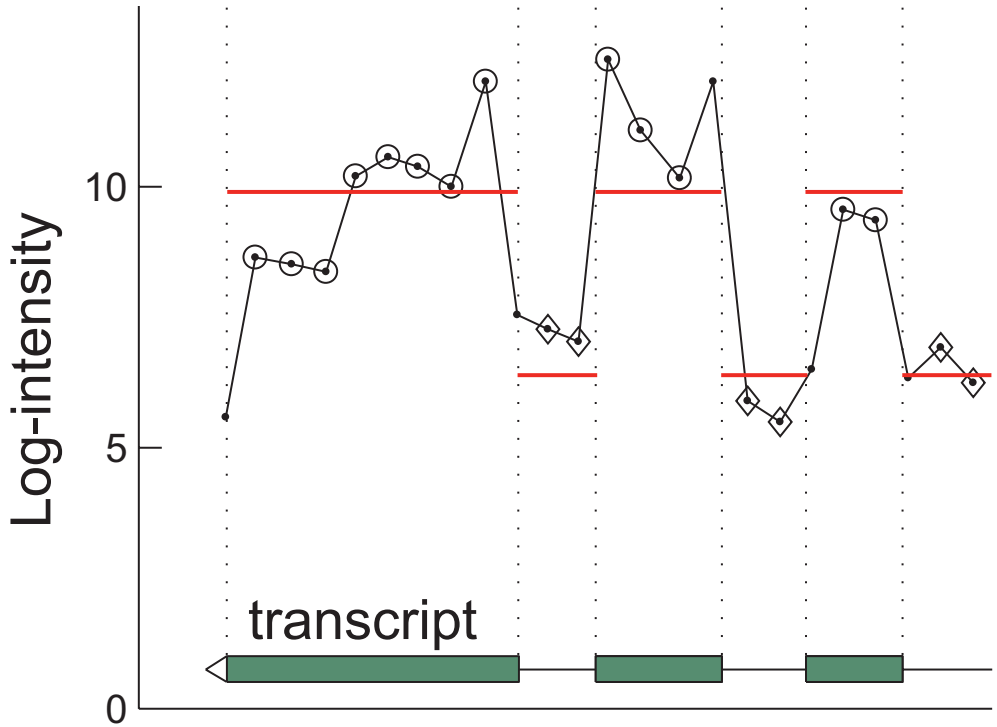


- observed intensity
- annotated exonic
- annotated intronic

Systematic bias induced by probe sequence effects

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

Intensities are Noisy Measurements

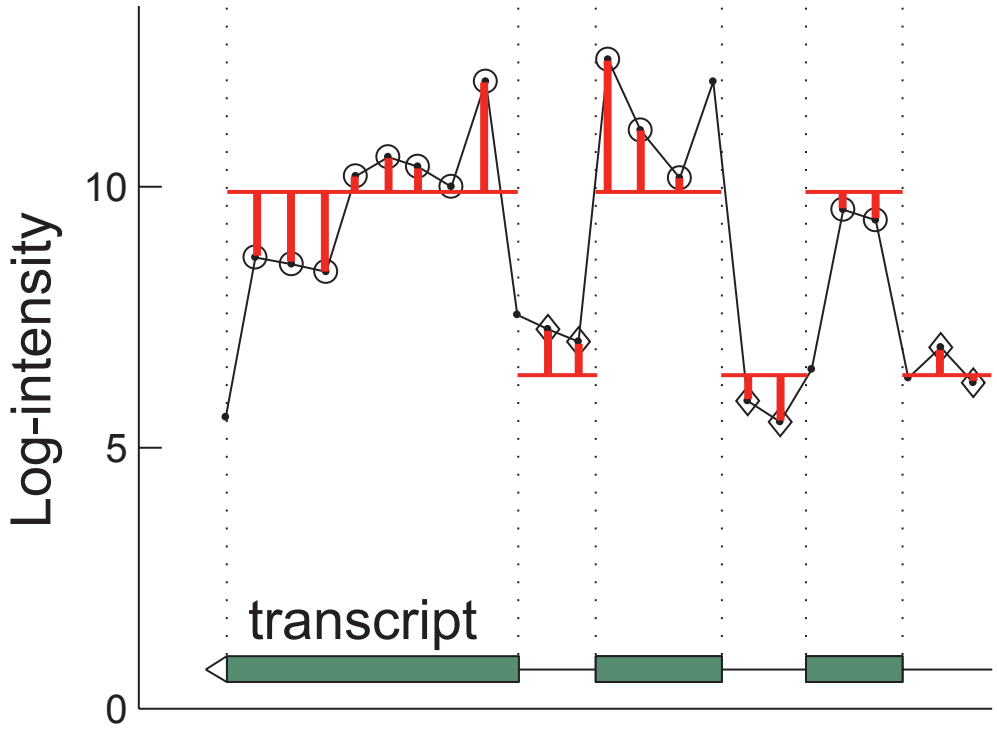


- observed intensity
- annotated exonic
- annotated intronic
- "ideal noise-free intensity"

Systematic bias induced by probe sequence effects

⇒ Model effect for normalization [Zeller et al., 2008b]

Intensities are Noisy Measurements



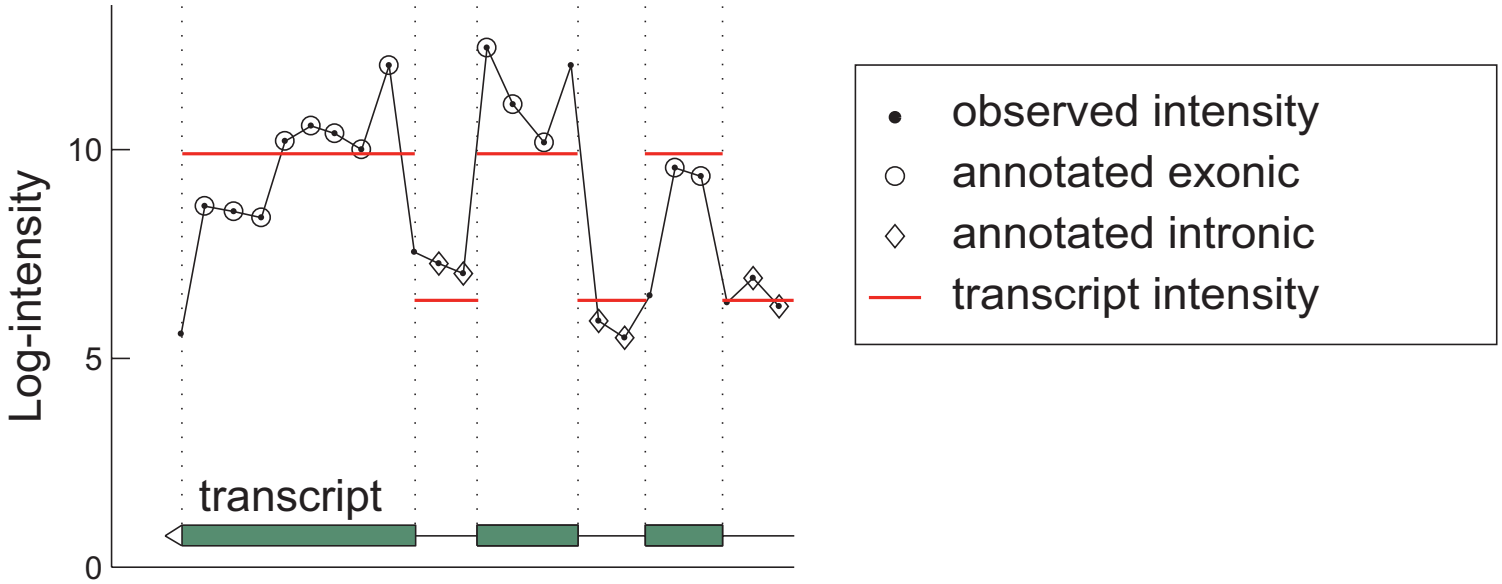
- observed intensity
- annotated exonic
- annotated intronic
- "ideal noise-free intensity"
- fold difference δ between observed and transcript intensity

Systematic bias induced by probe sequence effects

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

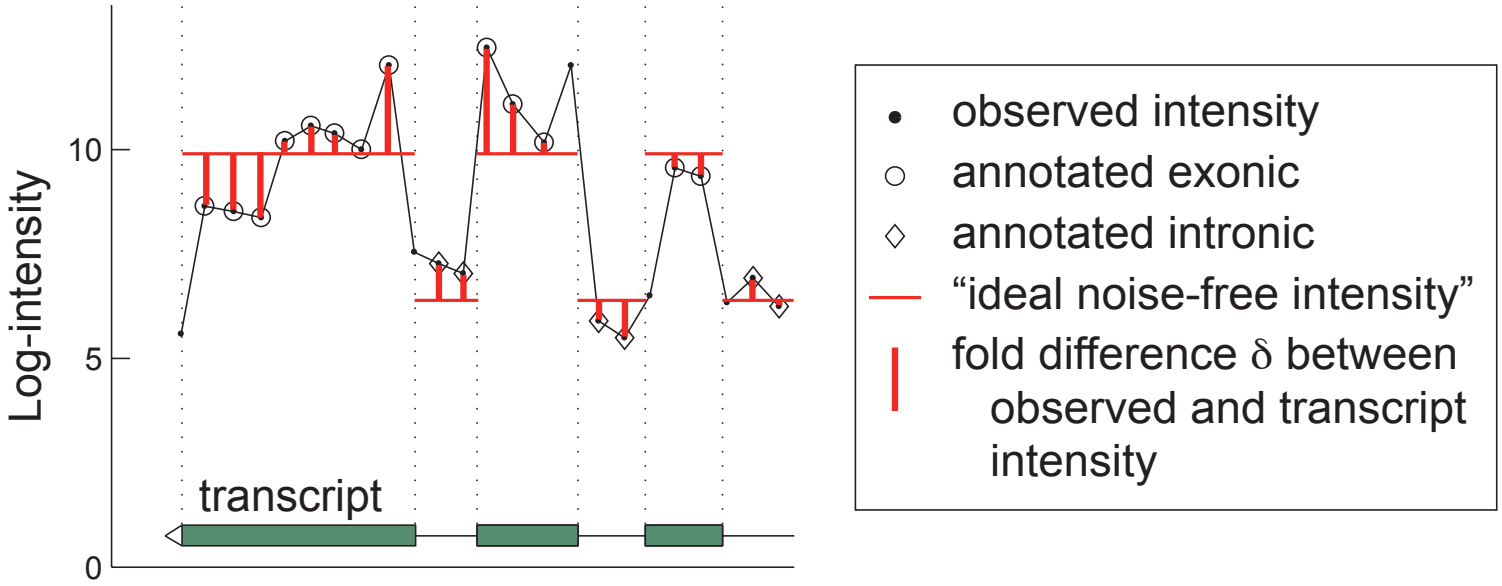
Tiling Array Normalization

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



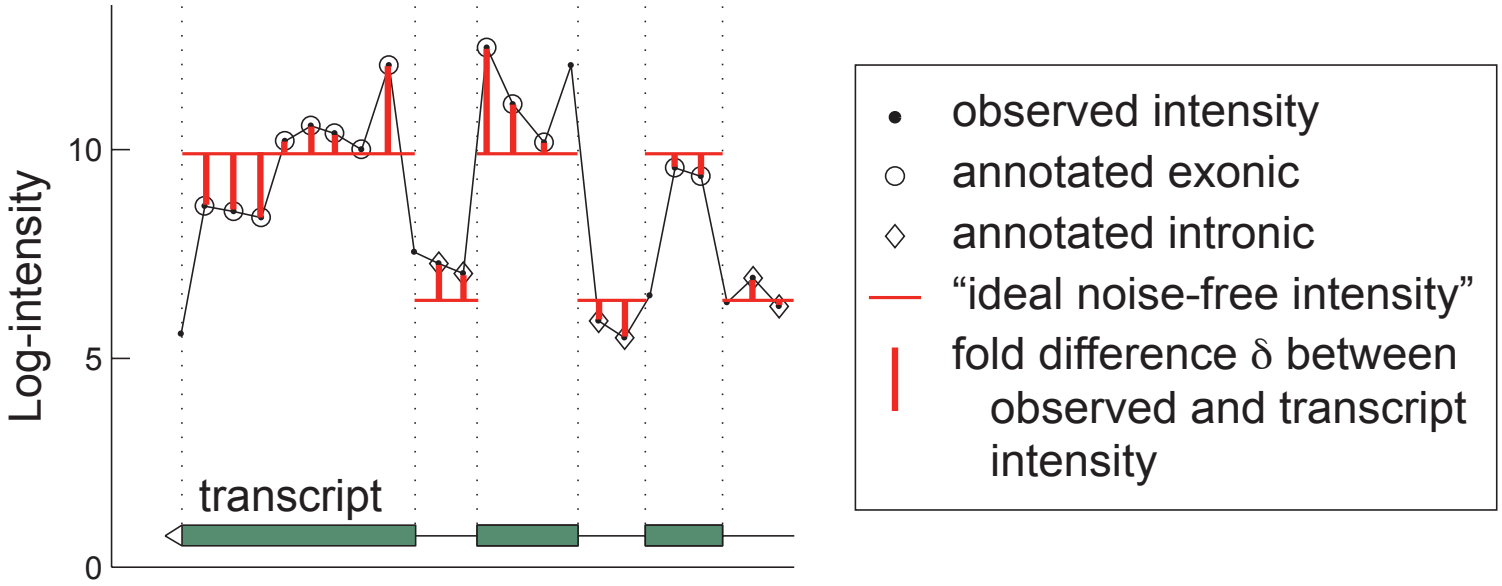
Tiling Array Normalization

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



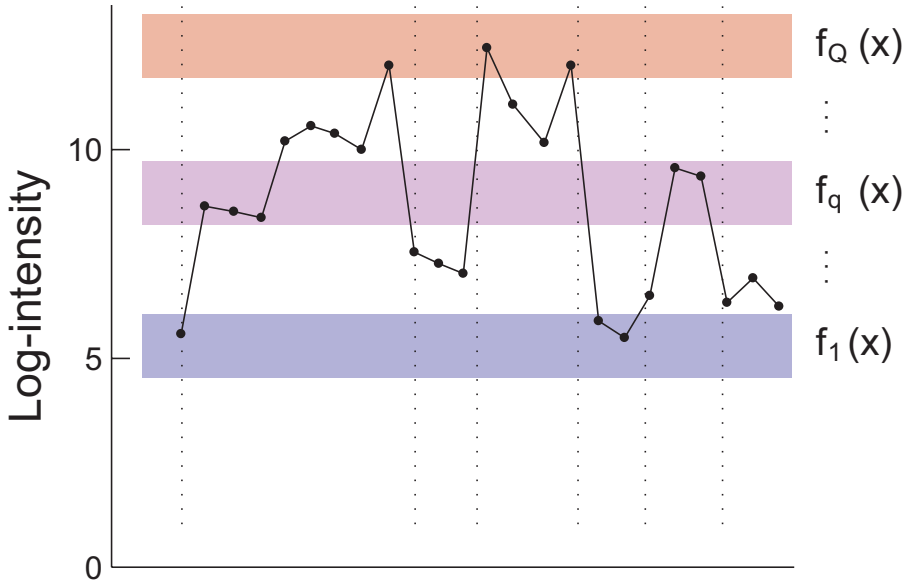
Tiling Array Normalization

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



Tiling Array Normalization

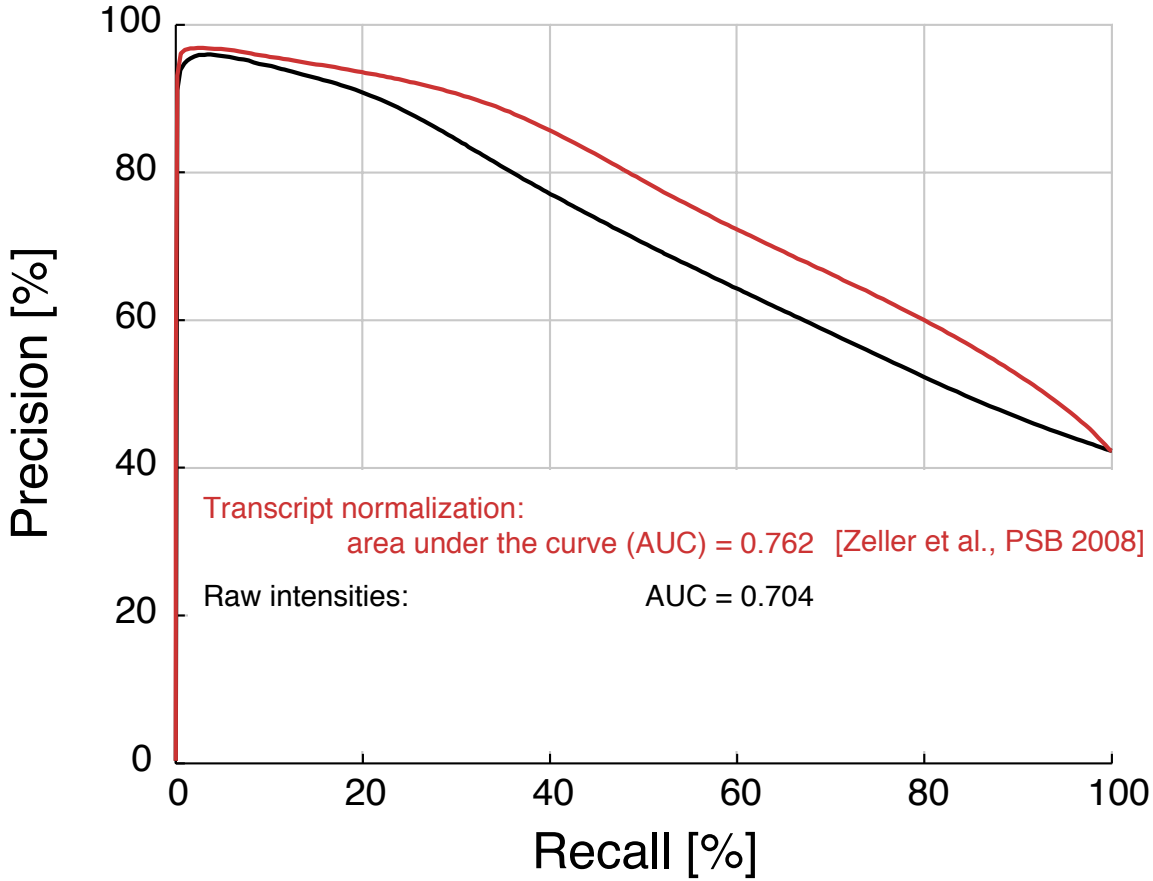
- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008b]
- Learns intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Takes probe sequence $\mathbf{x}_i$ (positional information on mono-, di- and tri-mer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



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

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

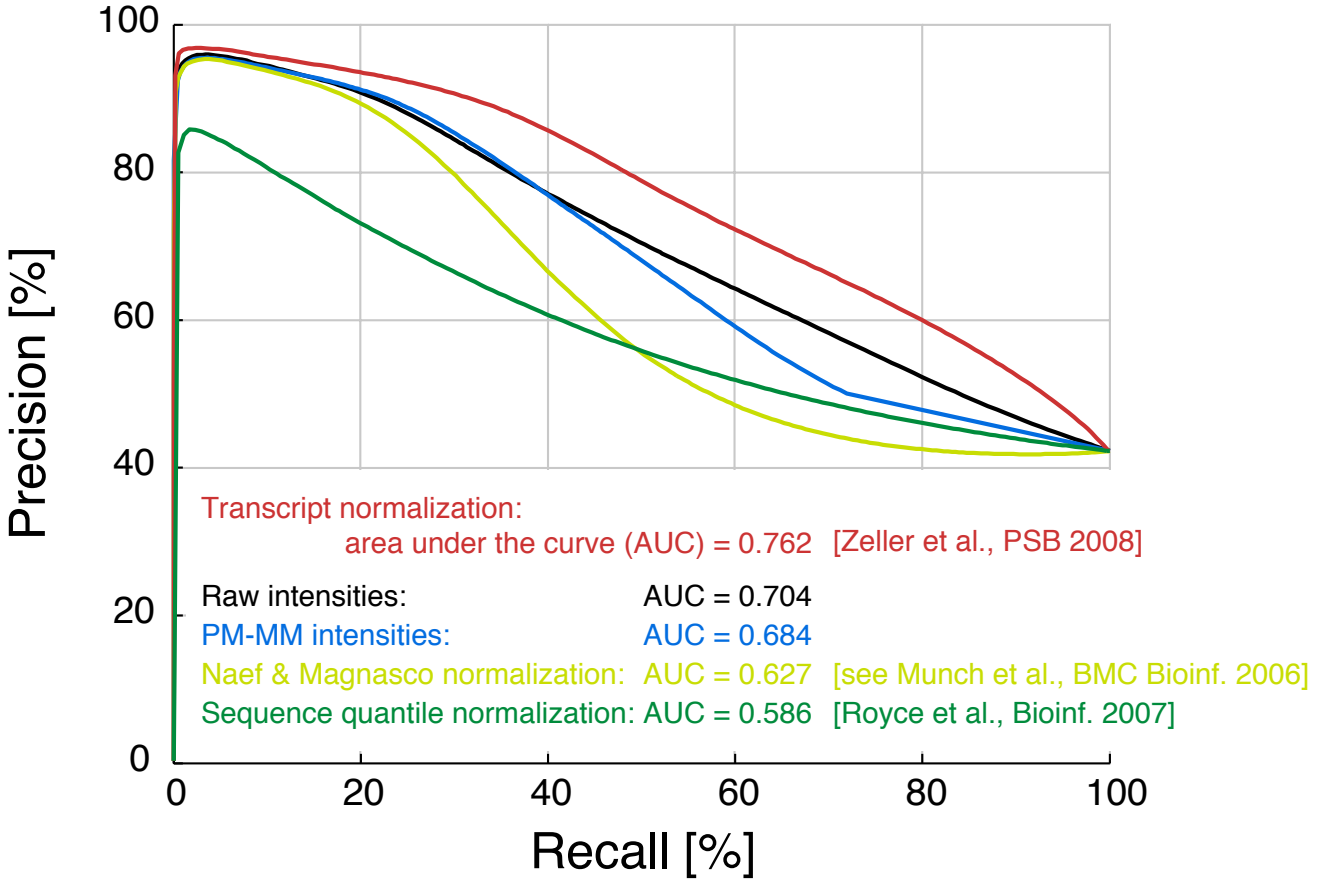
Transcript Normalization Results



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

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

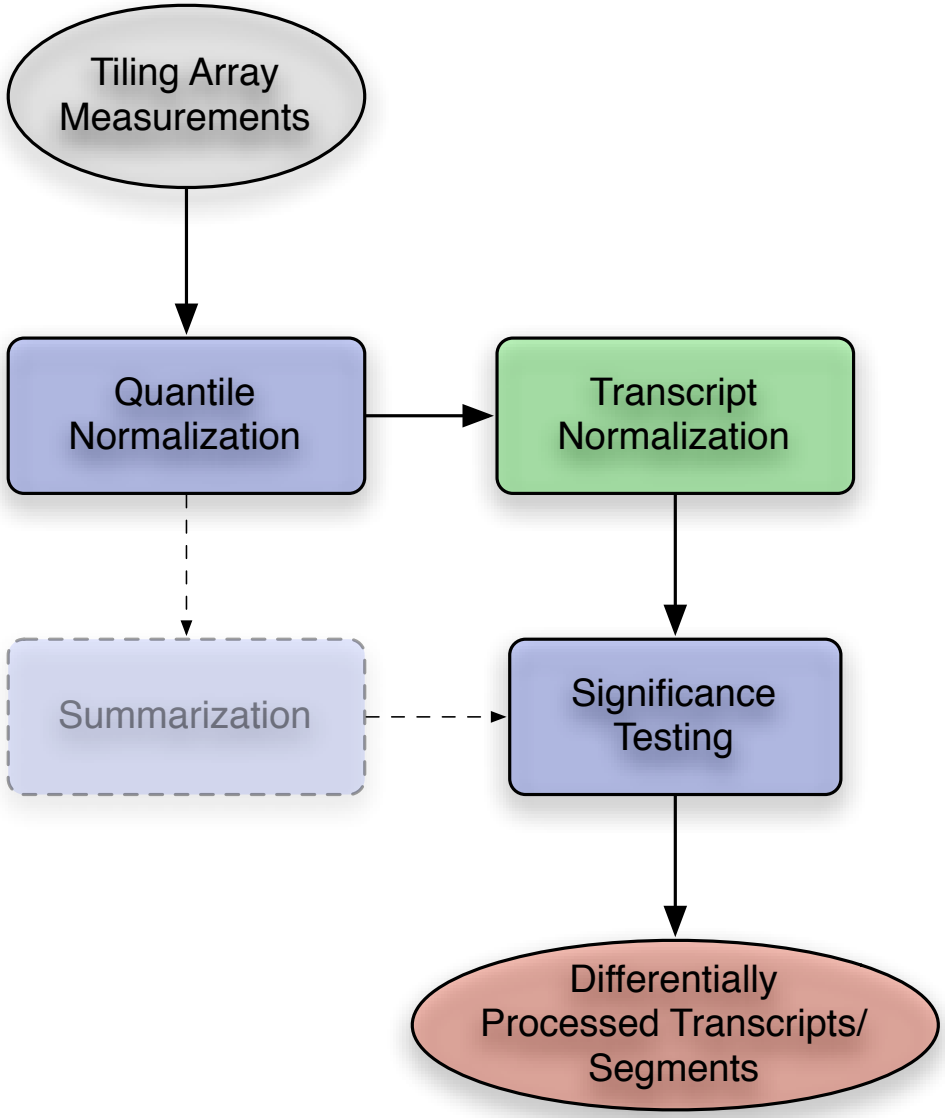
Transcript Normalization Results



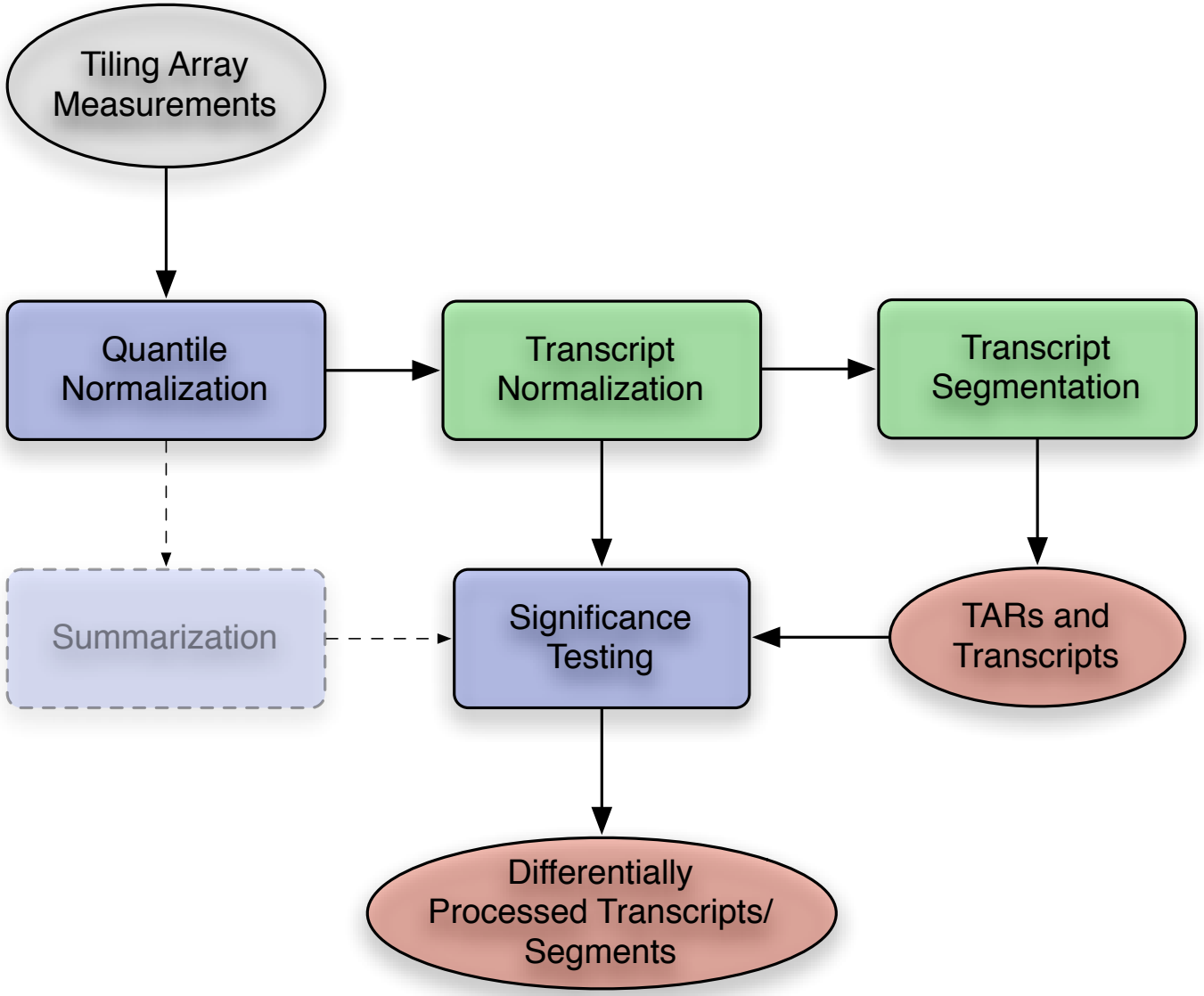
Transcript normalization (TN) results in better separation between background and exon signals $\longrightarrow$ improves transcript identification

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

Analysis Pipelines

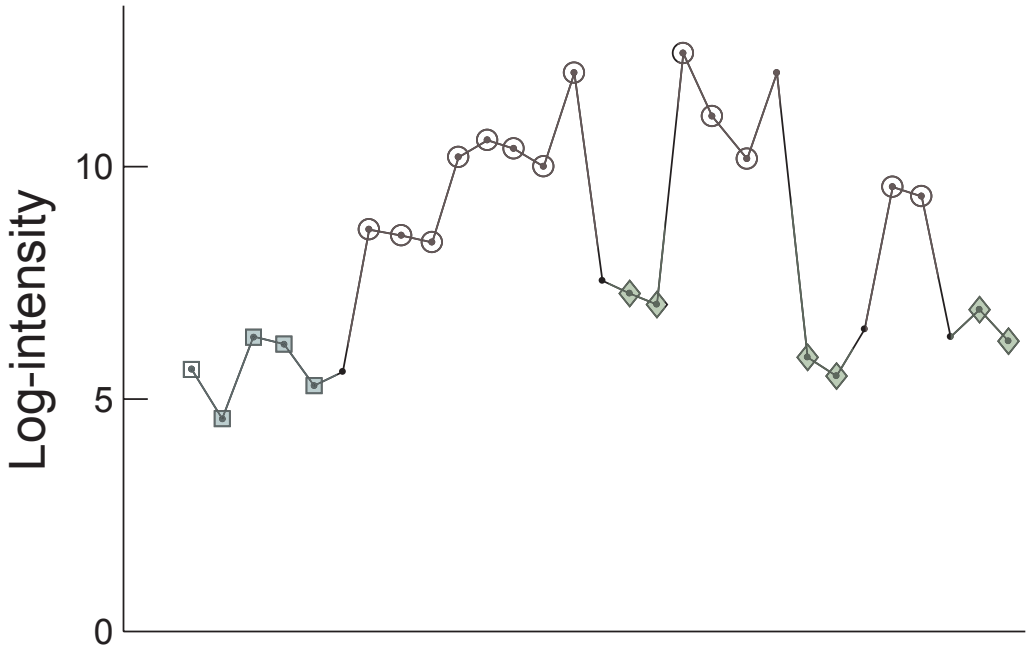


Analysis Pipelines



Tiling Array Segmentation

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



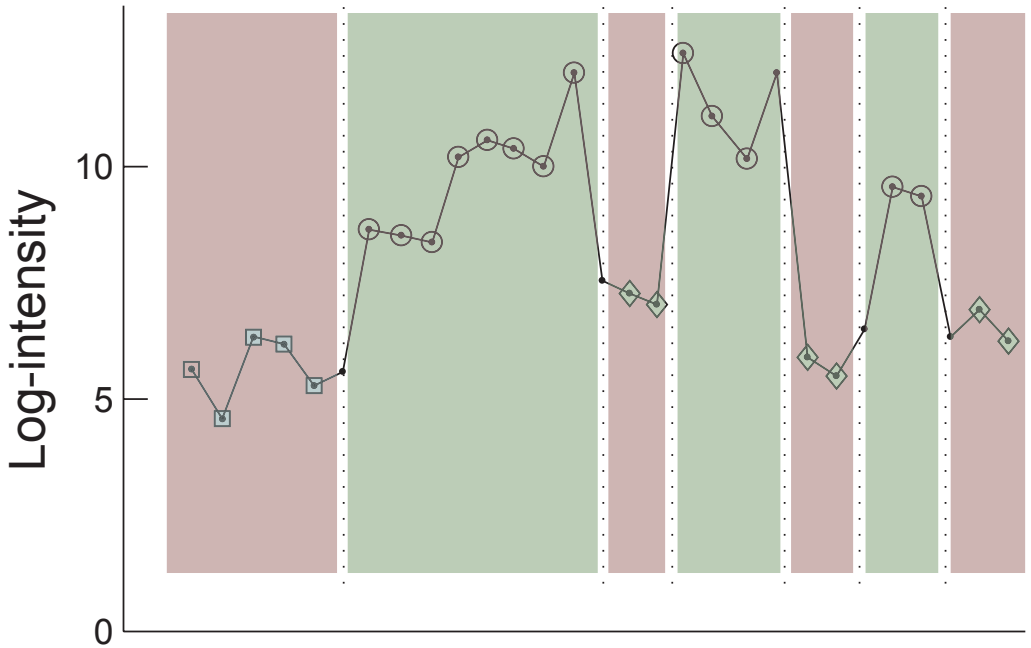
Novel segmentation method ("mSTAD")

- 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

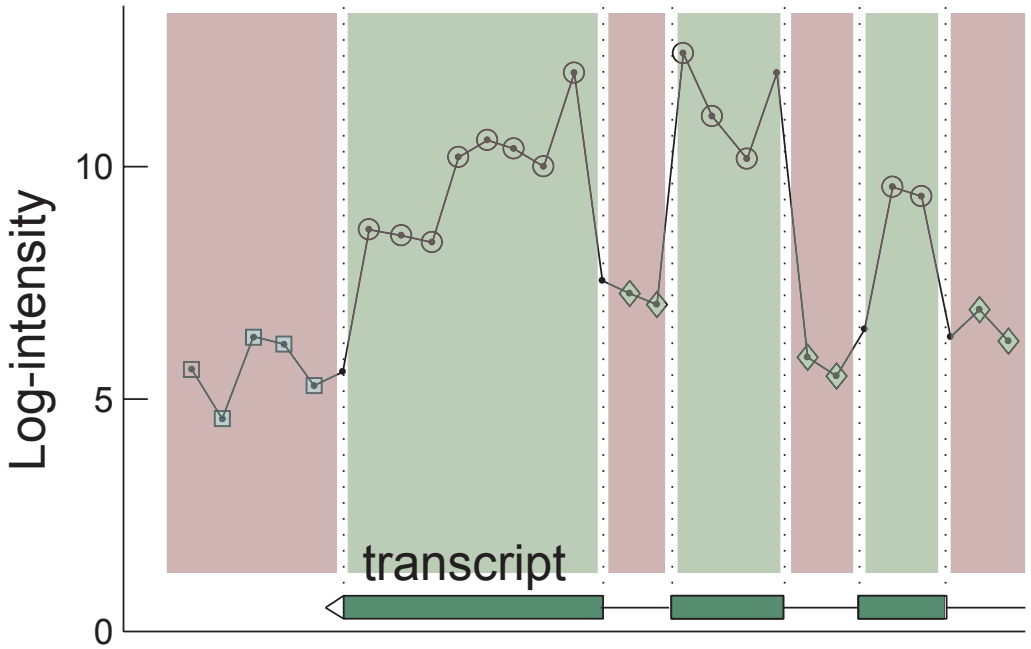
Novel segmentation method ("mSTAD")

- 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

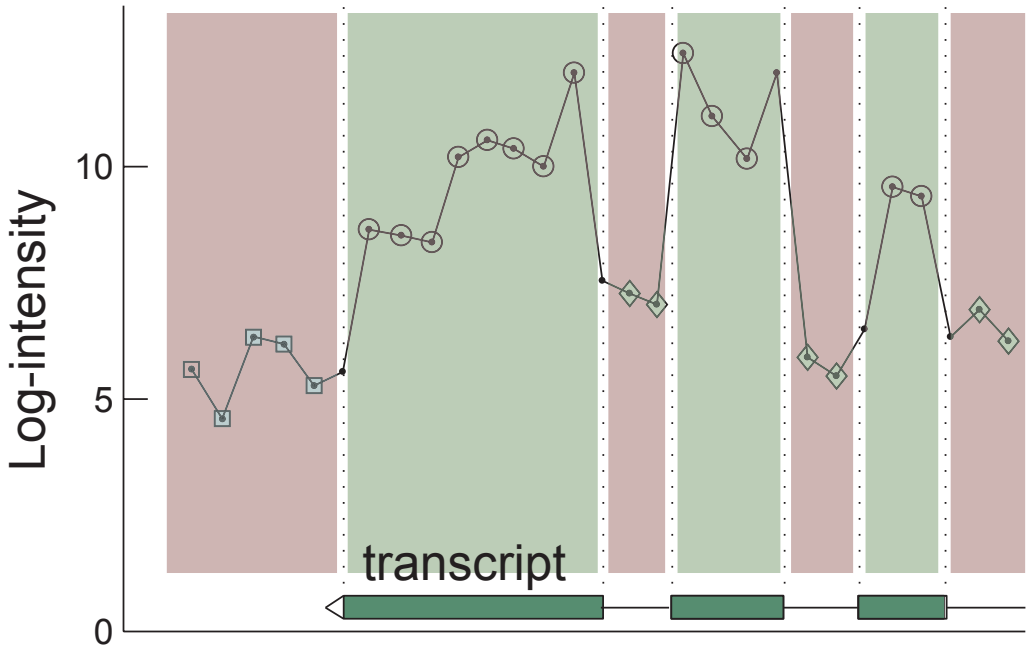
Novel segmentation method ("mSTAD")

- 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



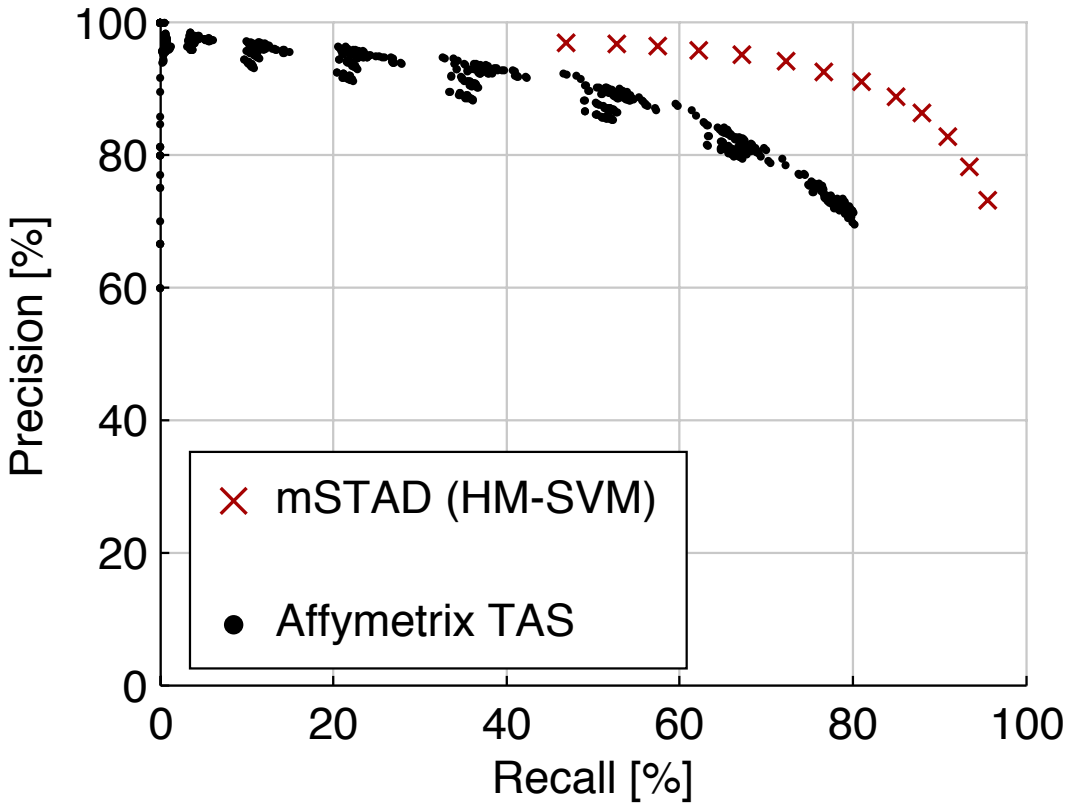
- observed intensity
- annotated exonic
- annotated intronic
- annotated intergenic

Novel segmentation method ("mSTAD")

- accounts for spliced transcripts
- provides very accurate predictions

[Zeller et al., 2008b]

TAR Identification: Method Comparison
(Combining Normalization and Segmentation Method)



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

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

Genome-wide Analyses of Alternative Transcripts

- *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., 2009, in prep.]
- Profiling cap-binding/miRNA processing mutants (CBP80/CBP20)

⇒ Influence splicing efficiency

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

Other analyses: under stresses & in small RNA biogenesis mutants [Zeller et al., 2009, Laubinger et al., 2009, submitted]

Genome-wide Analyses of Alternative Transcripts



- *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., 2009, in prep.]

- Profiling cap-binding/miRNA processing mutants (CBP80/CBP20)

⇒ Influence splicing efficiency

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

Other analyses: under stresses & in small RNA biogenesis mutants

[Zeller et al., 2009, Laubinger et al., 2009, submitted]

Genome-wide Analyses of Alternative Transcripts

- *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., 2009, in prep.]

- Profiling cap-binding/miRNA processing mutants (CBP80/CBP20)

$\Rightarrow$ Influence splicing efficiency

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

Other analyses: under stresses & in small RNA biogenesis mutants

[Zeller et al., 2009, Laubinger et al., 2009, submitted]

Genome-wide Analyses of Alternative Transcripts



- *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., 2009, in prep.]

- Profiling cap-binding/miRNA processing mutants (CBP80/CBP20)

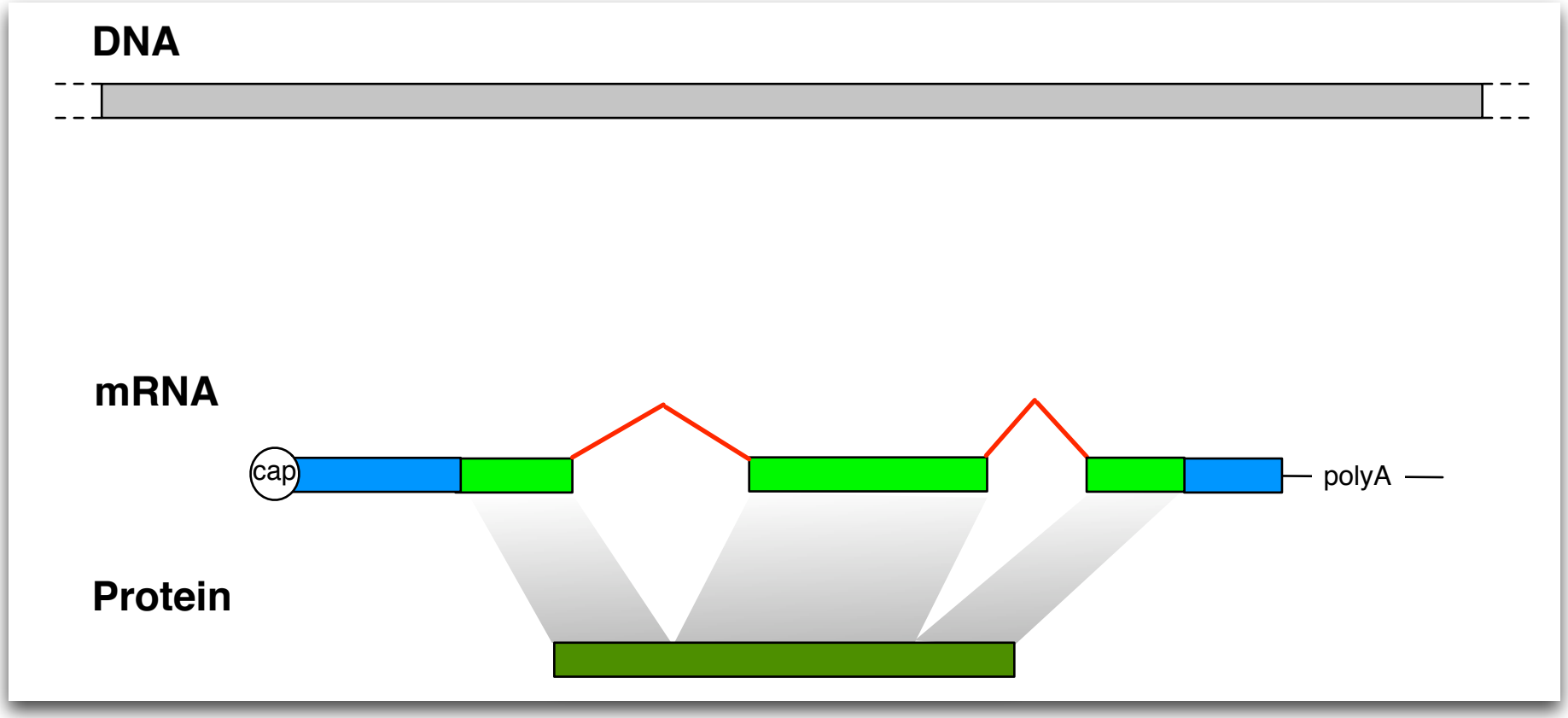
$\Rightarrow$ Influence splicing efficiency

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

Other analyses: under stresses & in small RNA biogenesis mutants

[Zeller et al., 2009, Laubinger et al., 2009, submitted]

From Transcriptome to Products



Directly measure the transcriptome

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

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



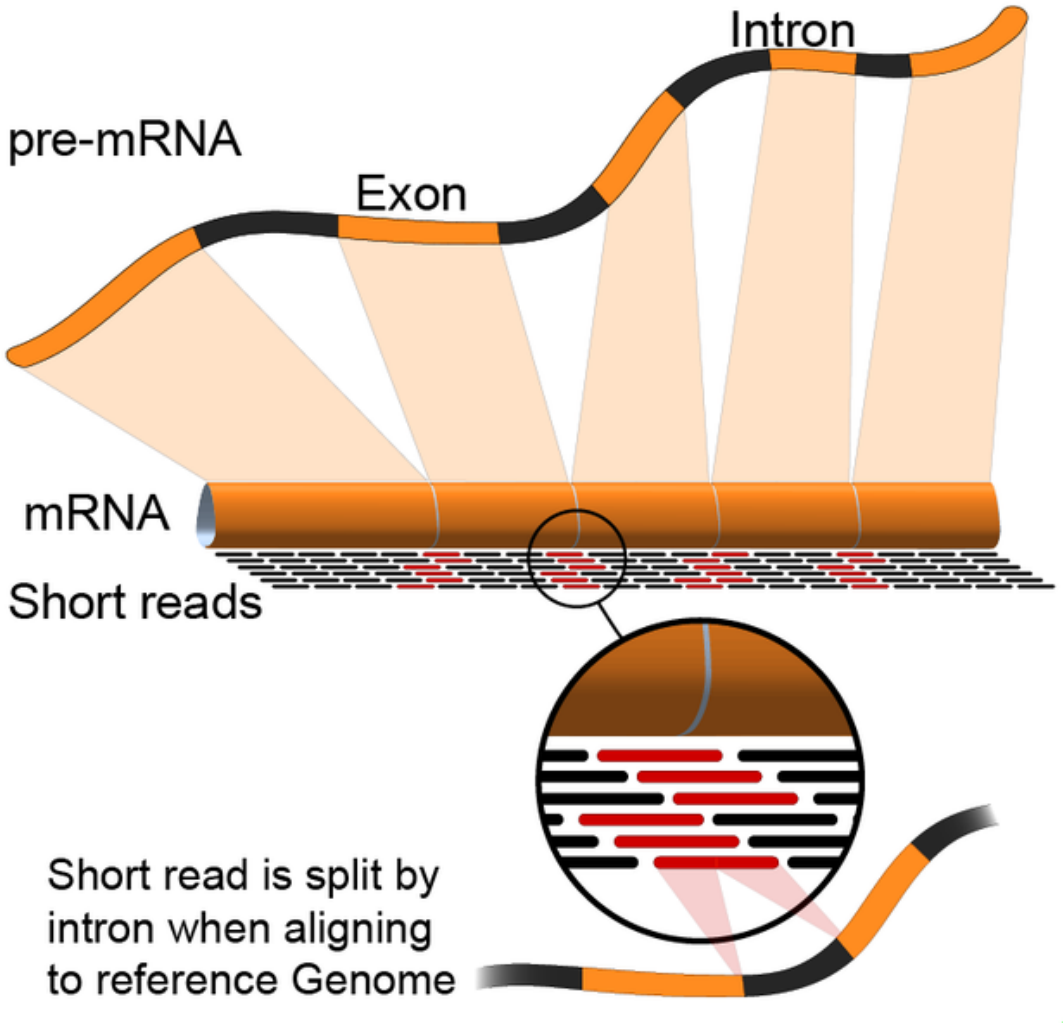
- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays
- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and *shorter* fragments
  - Much more *errors*
- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome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

Next Generation Sequencing



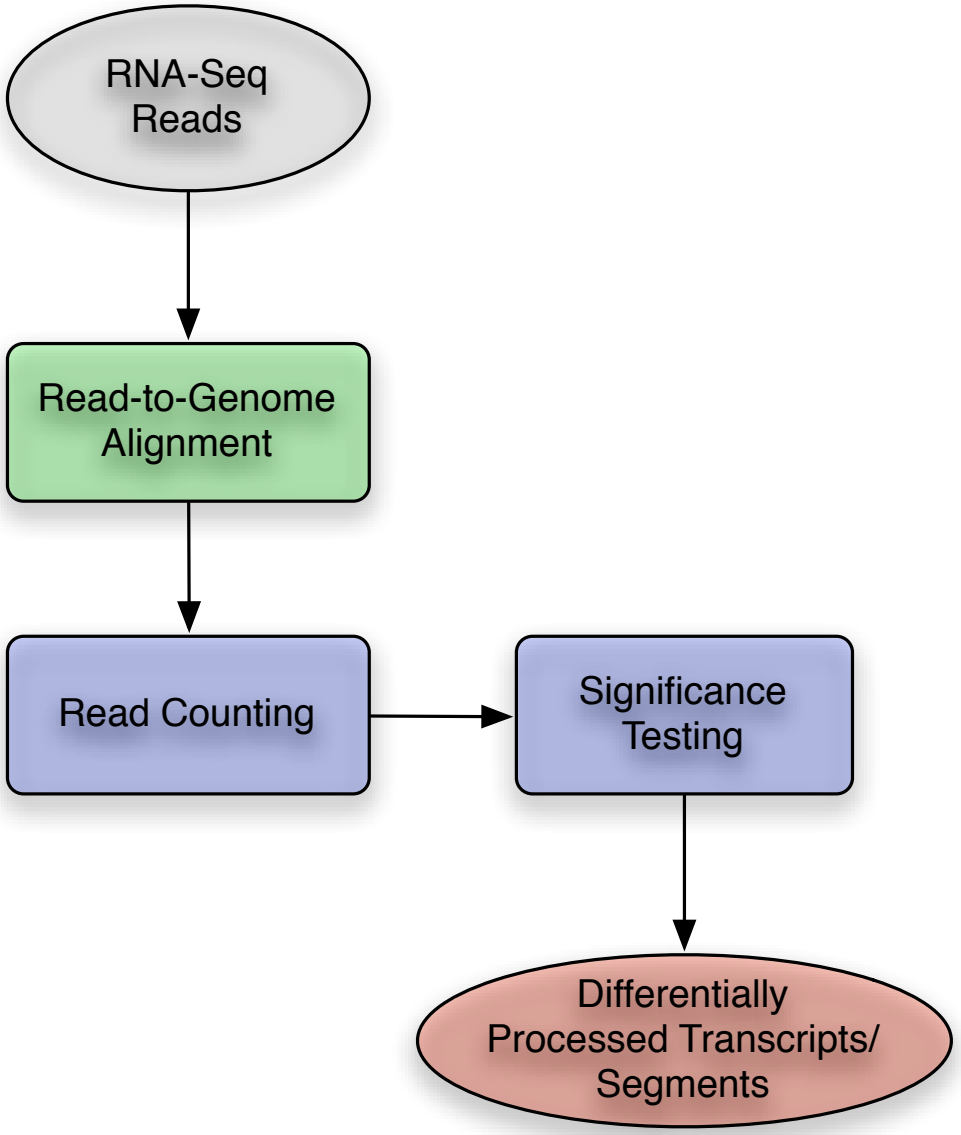
- Produces huge amounts of data
- Competes with Sanger sequencing and tiling arrays
- Differences to Sanger sequencing:
  - Much faster and cost effective per base
  - Much more and *shorter* fragments
  - Much more *errors*
- Genome (re-)sequencing
  - Identification of polymorphisms
  - *De novo* genome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

mRNA Deep Sequencing

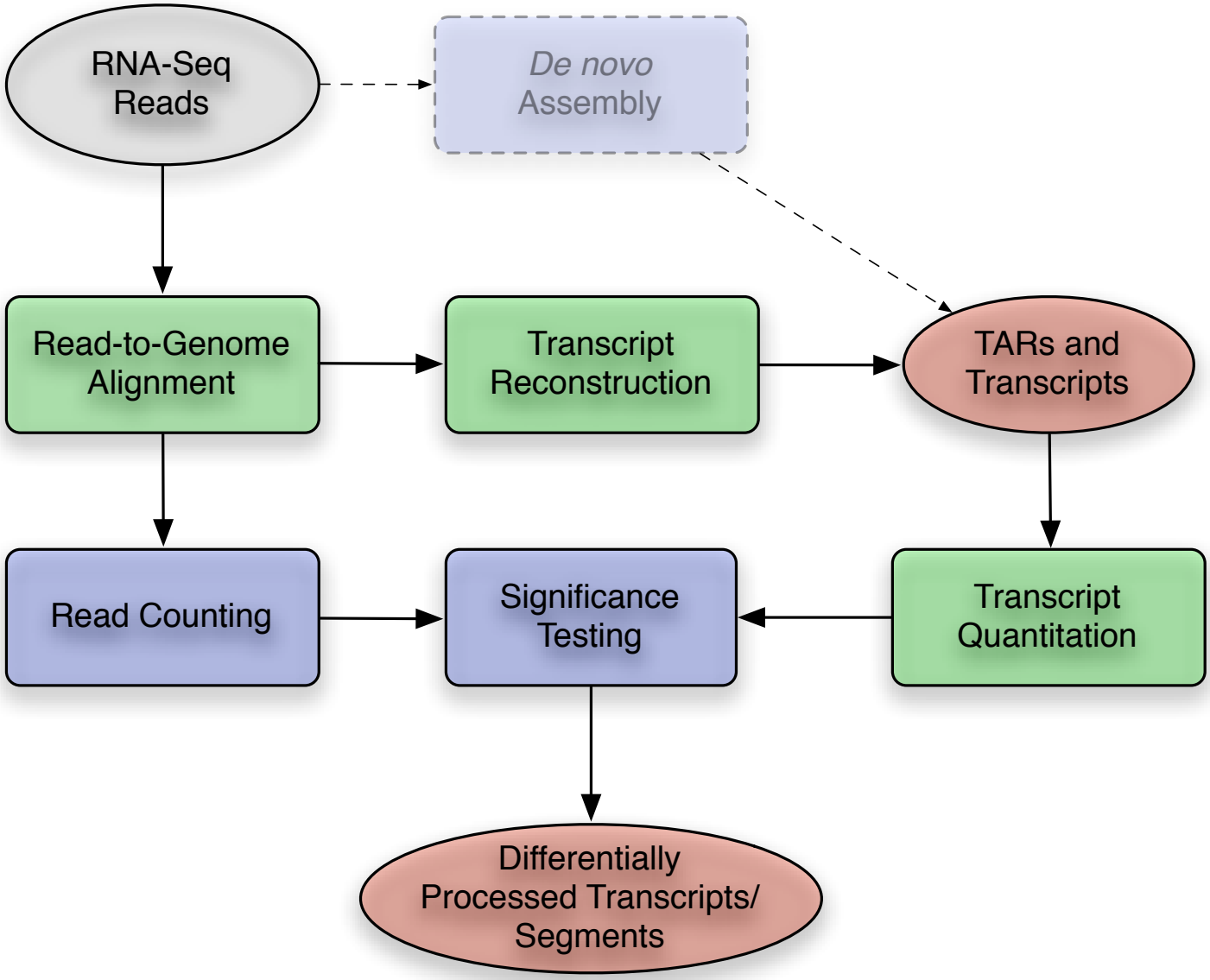


[Wikipedia]

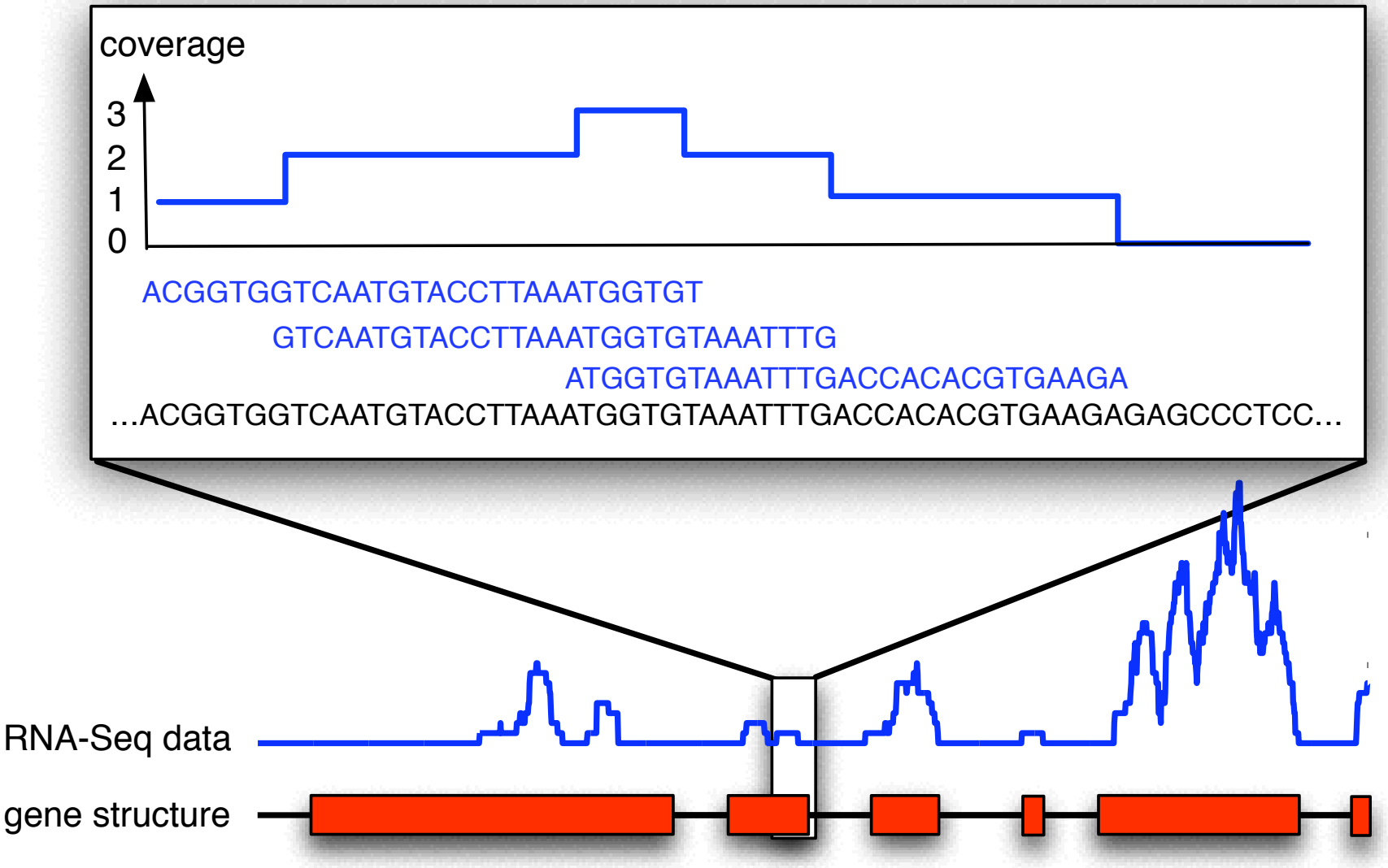
RNA-Seq Analysis Steps



RNA-Seq Analysis Steps

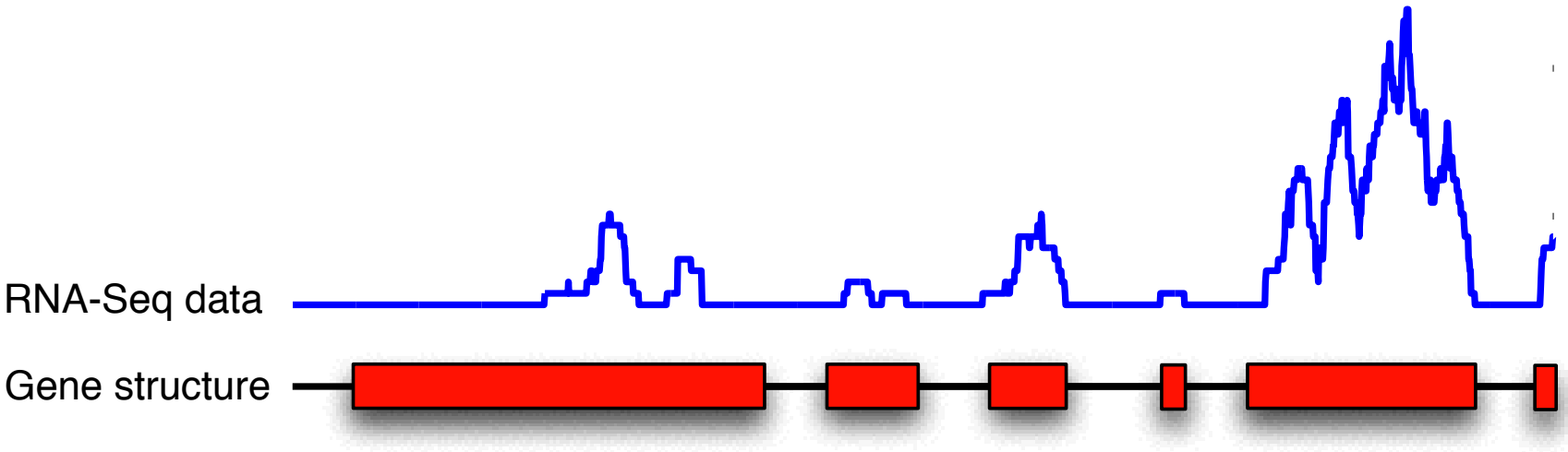


RNA-Seq Read Alignment



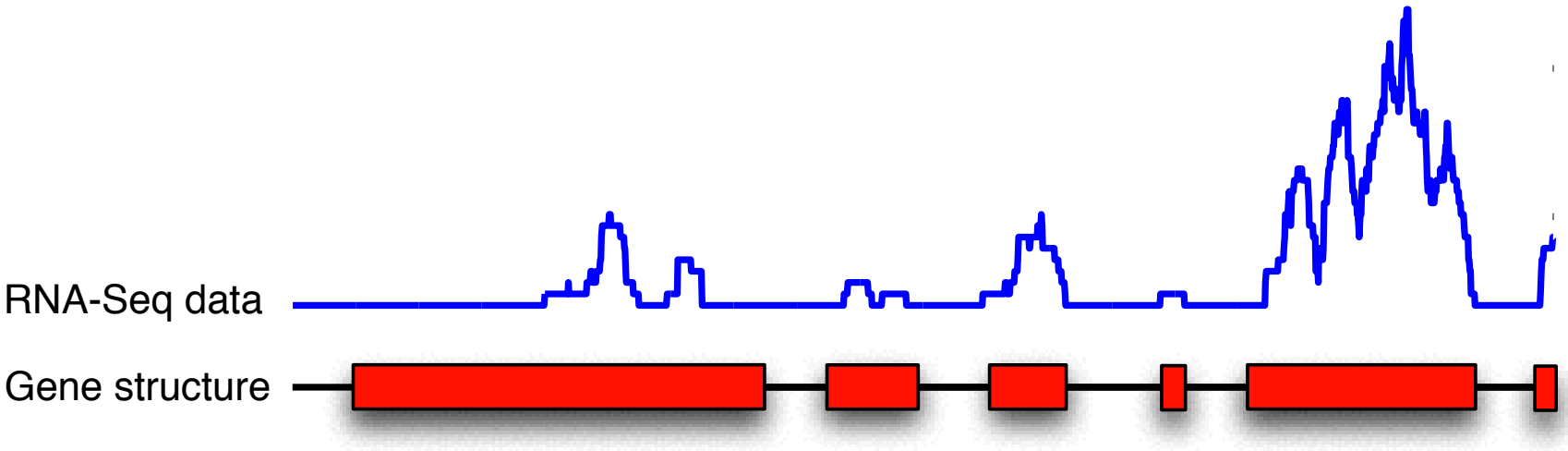
RNA-Seq Read Alignment

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

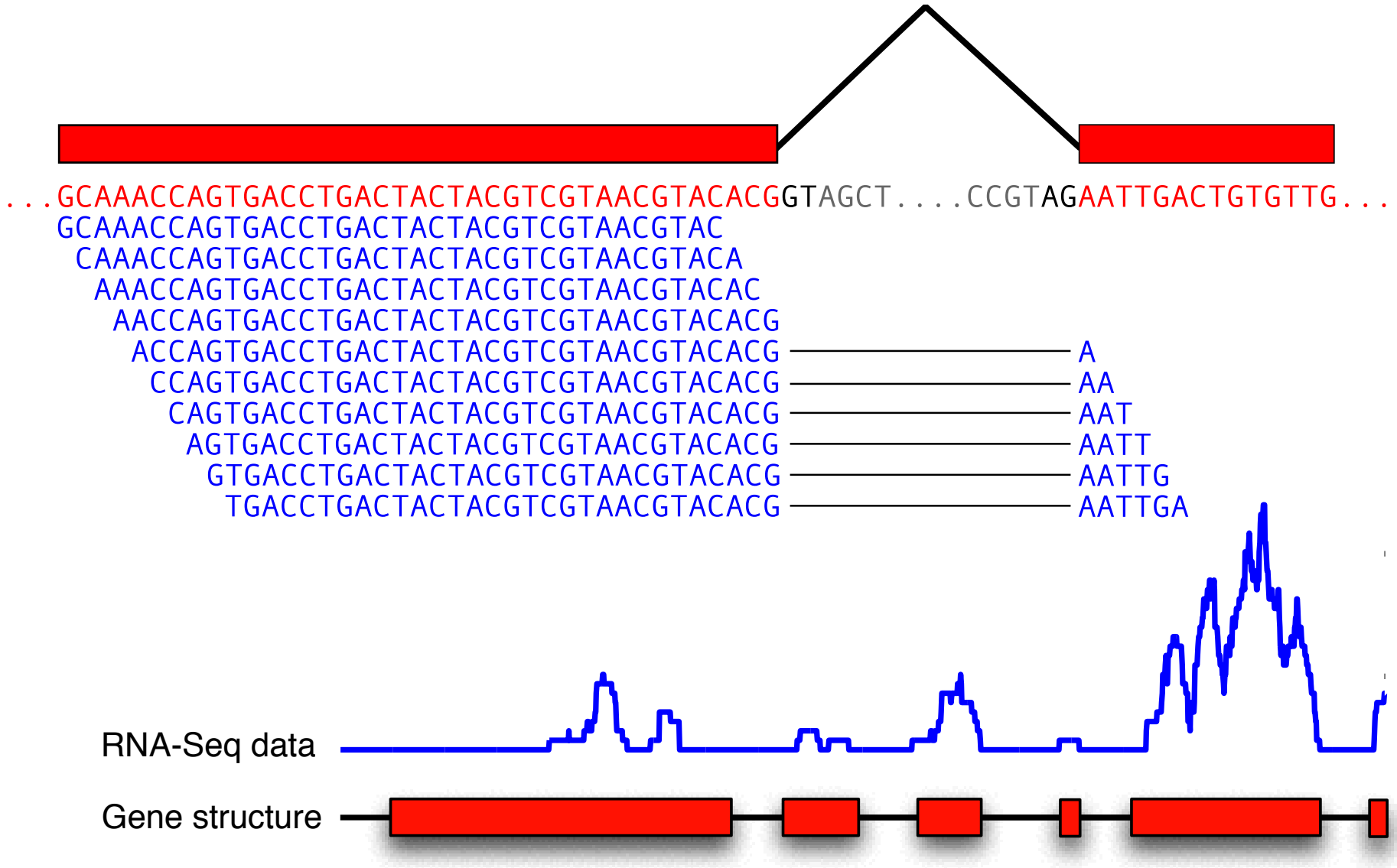


RNA-Seq Read Alignment

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



RNA-Seq Read Alignment



Extended Smith-Waterman Algorithm

⇒ Combine heterogeneous evidence for accurate alignment



	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

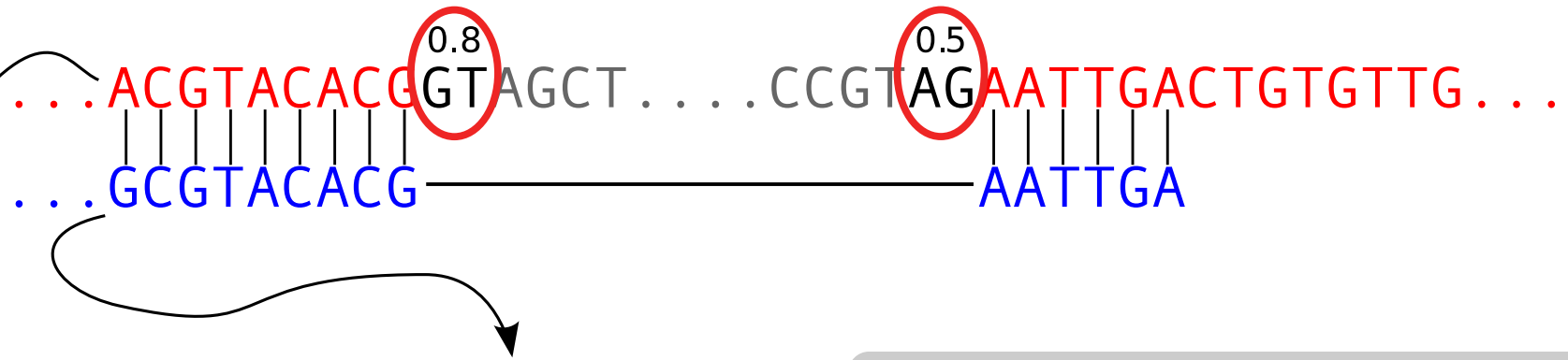
Source of Information

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

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

Extended Smith-Waterman Algorithm

⇒ Combine heterogeneous evidence for accurate alignment



	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

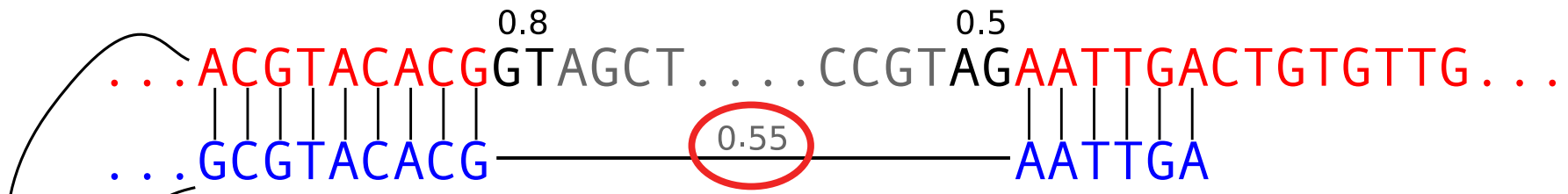
Source of Information

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

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

Extended Smith-Waterman Algorithm

⇒ Combine heterogeneous evidence for accurate alignment



	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

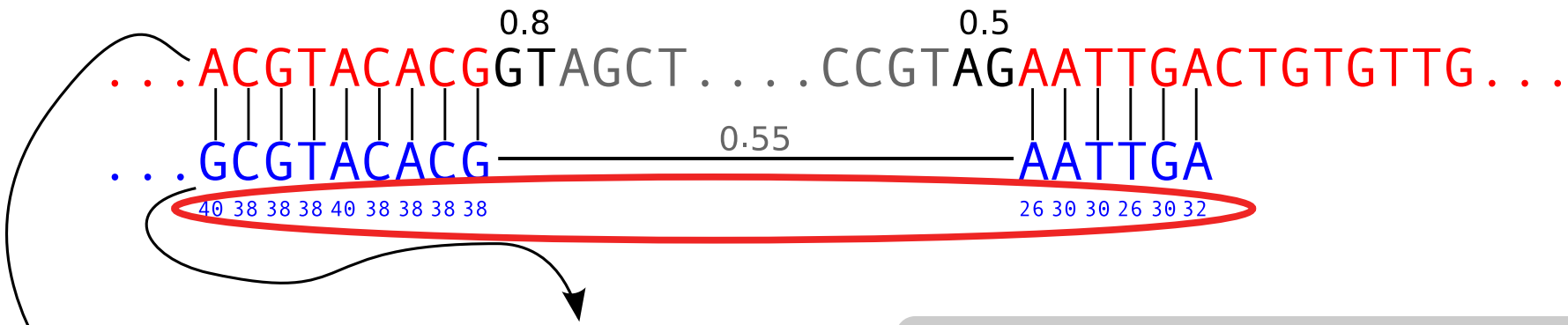
Source of Information

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

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

Extended Smith-Waterman Algorithm

⇒ Combine heterogeneous evidence for accurate alignment



	gap	A	C	G	T	N
gap	0.33	$f_{-A}(\cdot)$	$f_{-C}(\cdot)$	$f_{-G}(\cdot)$	$f_{-T}(\cdot)$	$f_{-N}(\cdot)$
A	0.31	$f_{AA}(\cdot)$	$f_{AC}(\cdot)$	$f_{AG}(\cdot)$	$f_{AT}(\cdot)$	$f_{AN}(\cdot)$
C	0.44	$f_{CA}(\cdot)$	$f_{CC}(\cdot)$	$f_{CG}(\cdot)$	$f_{CT}(\cdot)$	$f_{CN}(\cdot)$
G	0.13	$f_{GA}(\cdot)$	$f_{GC}(\cdot)$	$f_{GG}(\cdot)$	$f_{GT}(\cdot)$	$f_{GN}(\cdot)$
T	0.55	$f_{TA}(\cdot)$	$f_{TC}(\cdot)$	$f_{TG}(\cdot)$	$f_{TT}(\cdot)$	$f_{TN}(\cdot)$
N	0.12	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$

Source of Information

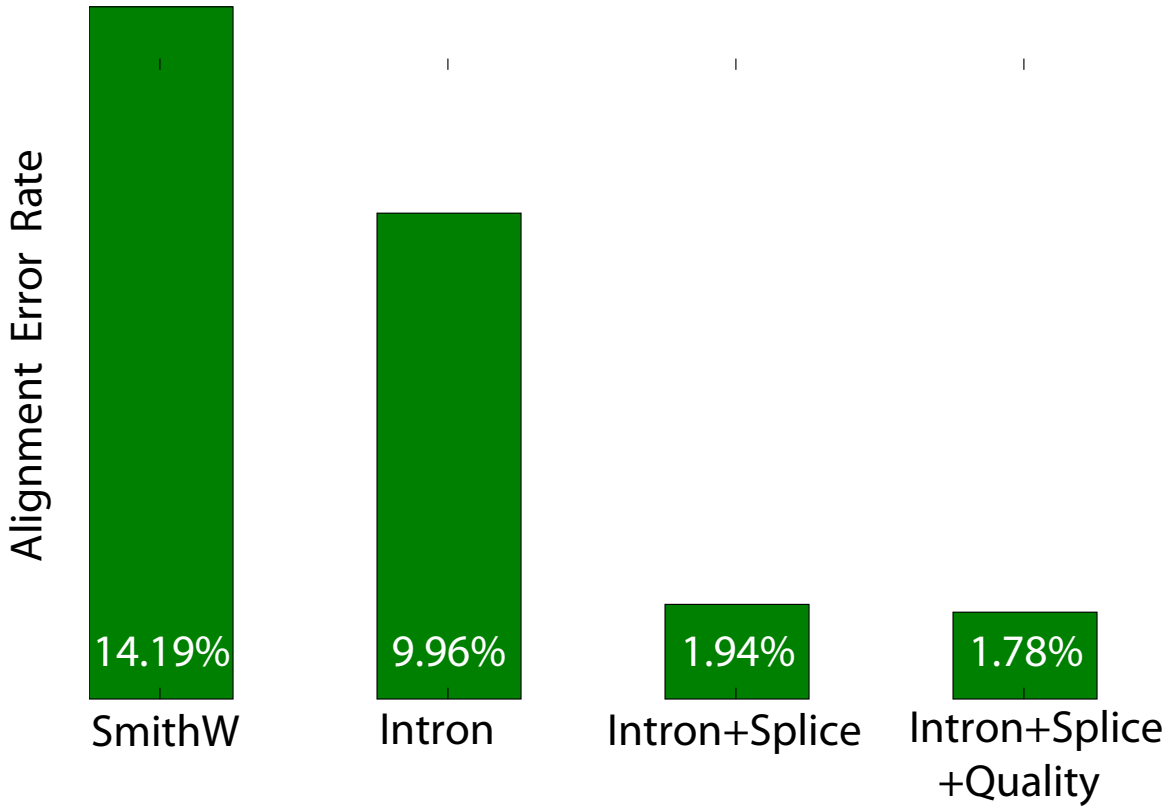
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

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

QPalma's Accurate Alignments

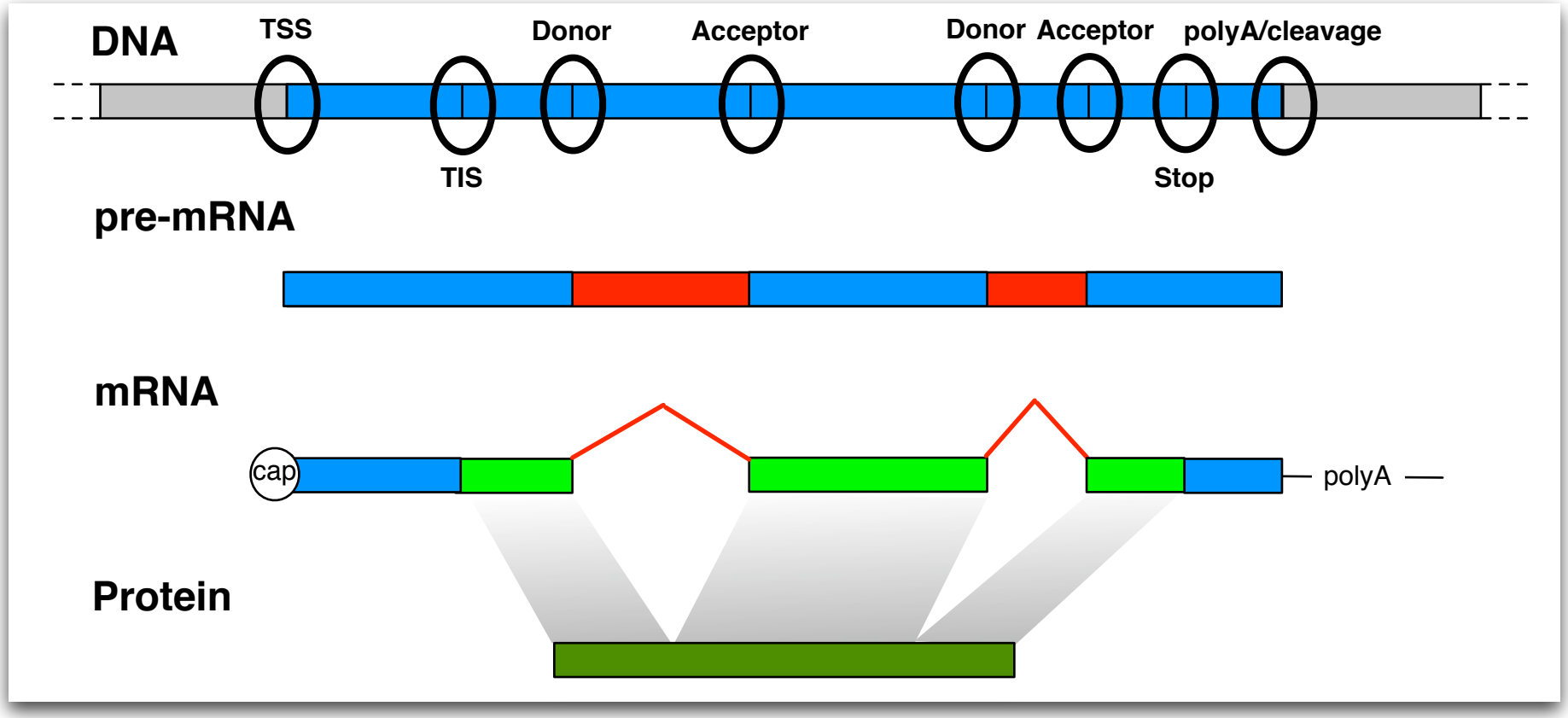
Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing



[De Bona et al., 2008]

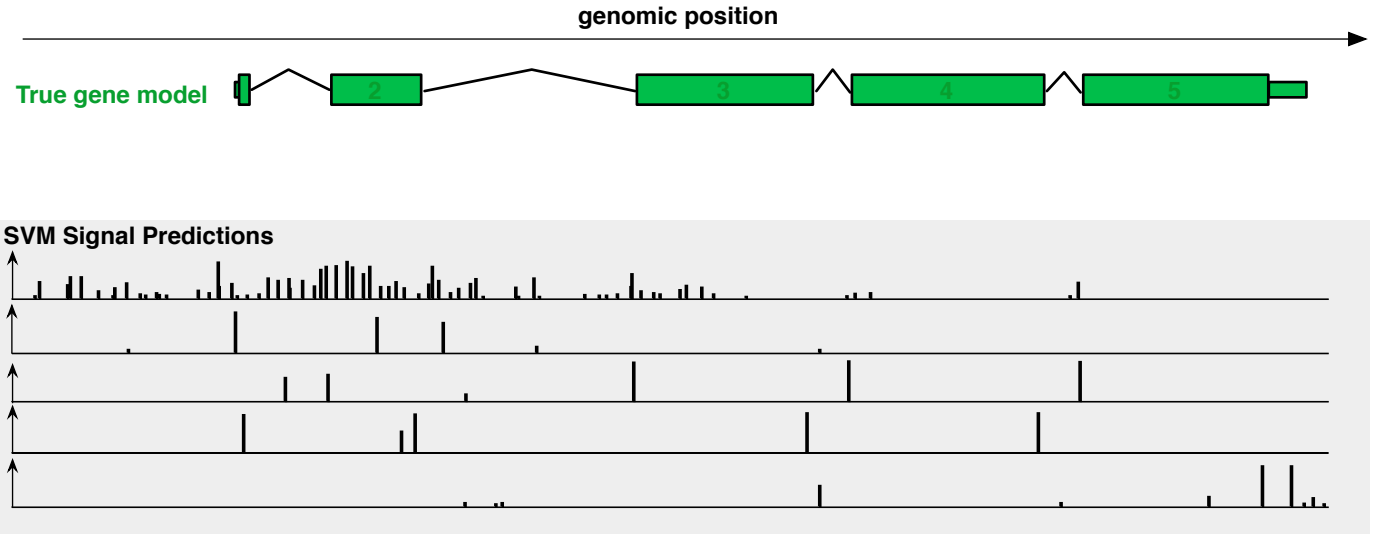
Gene Finding + Transcriptome Measurements



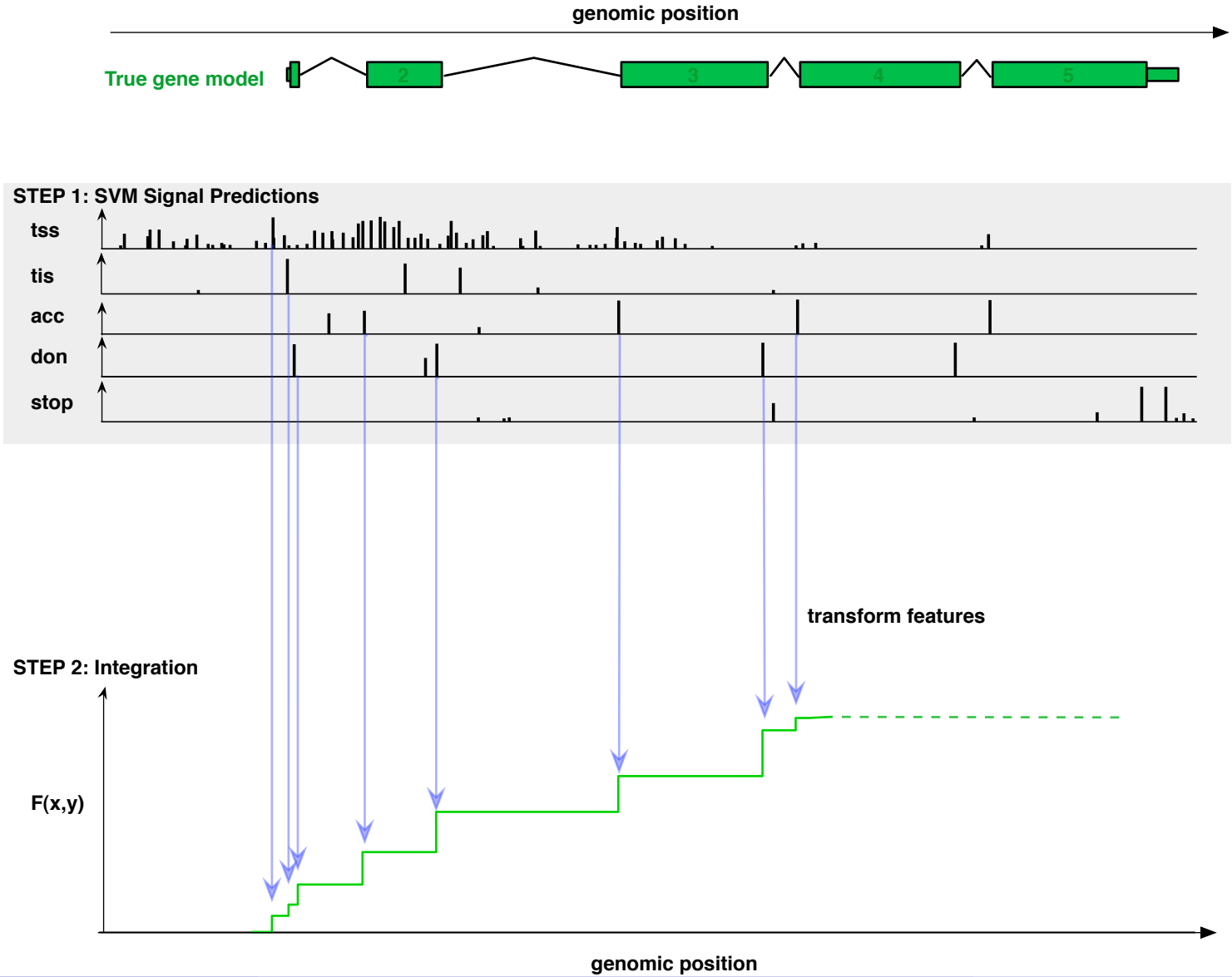
Combination

- Higher accuracy
- Condition/tissue specific predictions

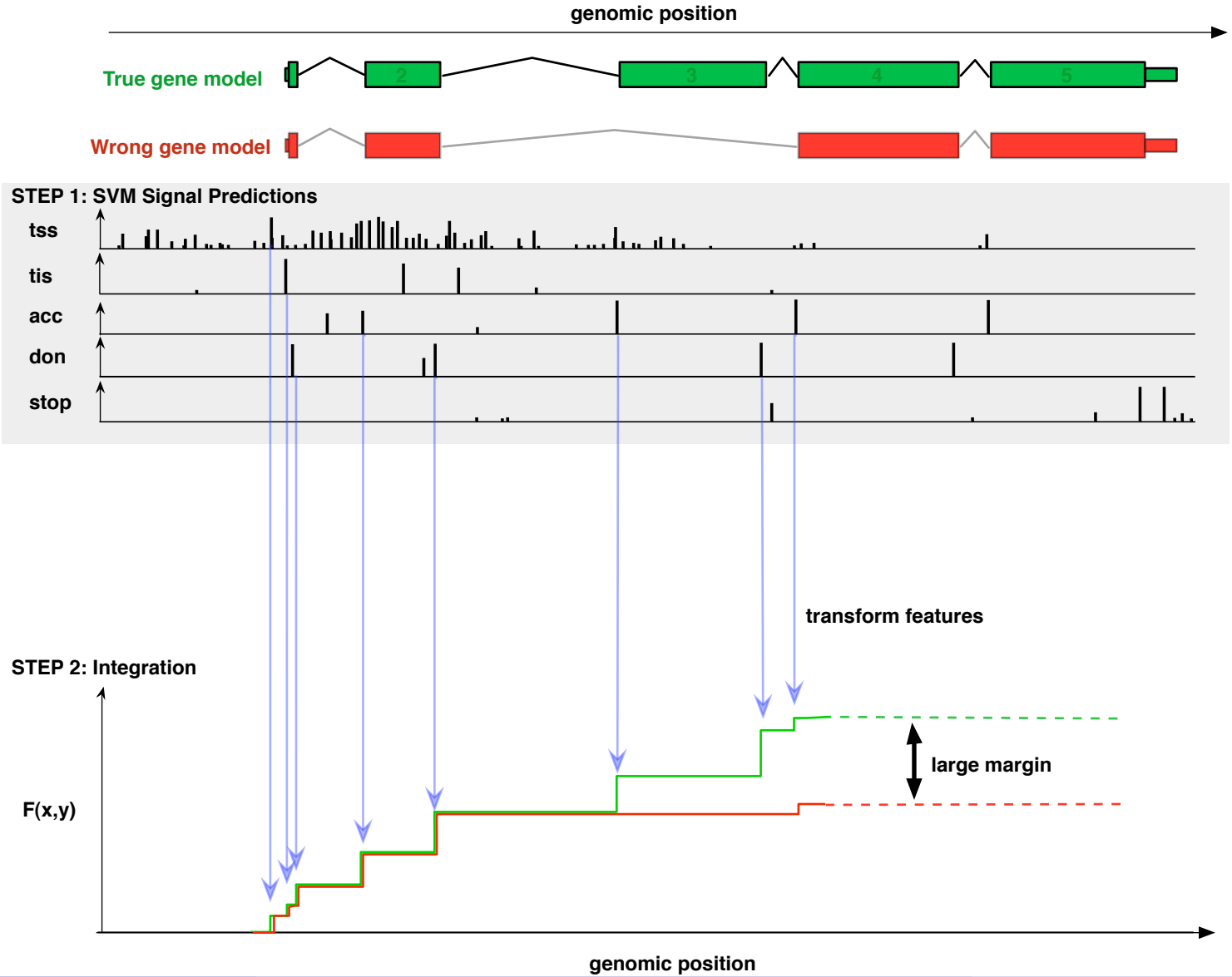
Learning to Integrate Data Sources



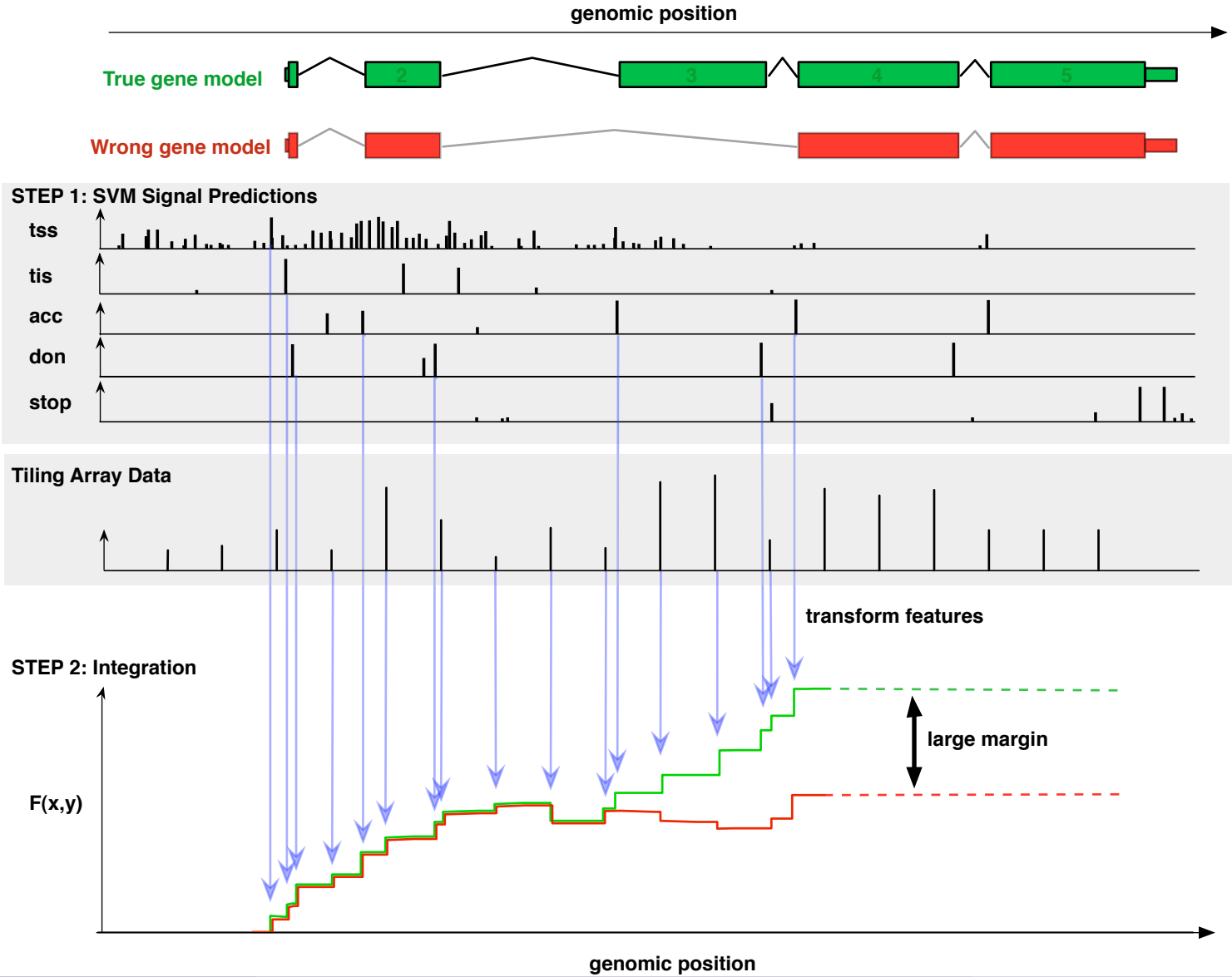
Learning to Integrate Data Sources



Learning to Integrate Data Sources



Learning to Integrate Data Sources



Gene Finding + Transcriptome Measurements

[Behr et al., 2009, in prep.]



Idea: Use transcriptome measurements to enhance recognition of exonic regions.

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

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

Gene Finding + Transcriptome Measurements

[Behr et al., 2009, in prep.]



Idea: Use transcriptome measurements to enhance recognition of exonic regions.

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

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

Gene Finding + Transcriptome Measurements

[Behr et al., 2009, in prep.]



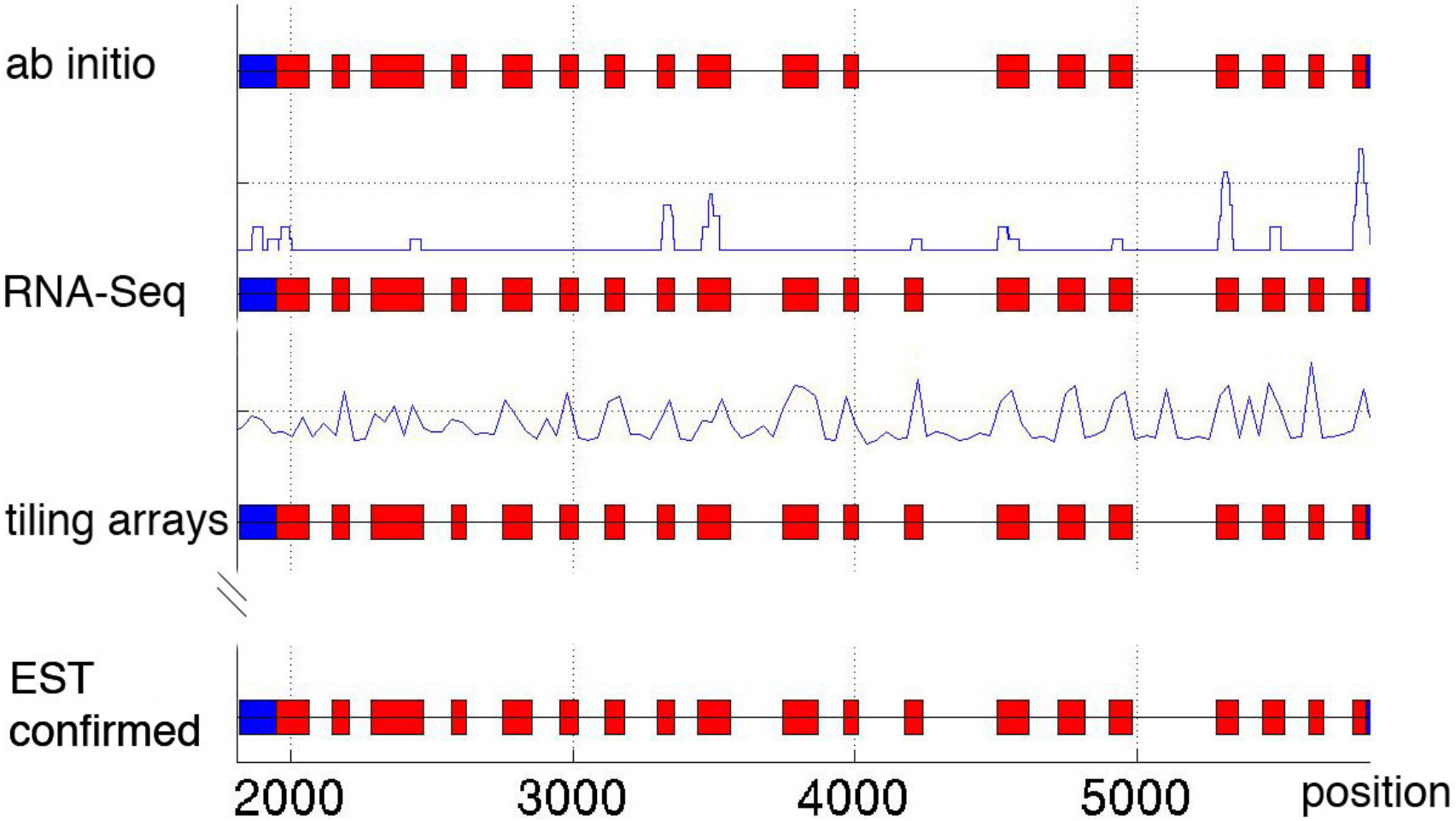
Idea: Use transcriptome measurements to enhance recognition of exonic regions.

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

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

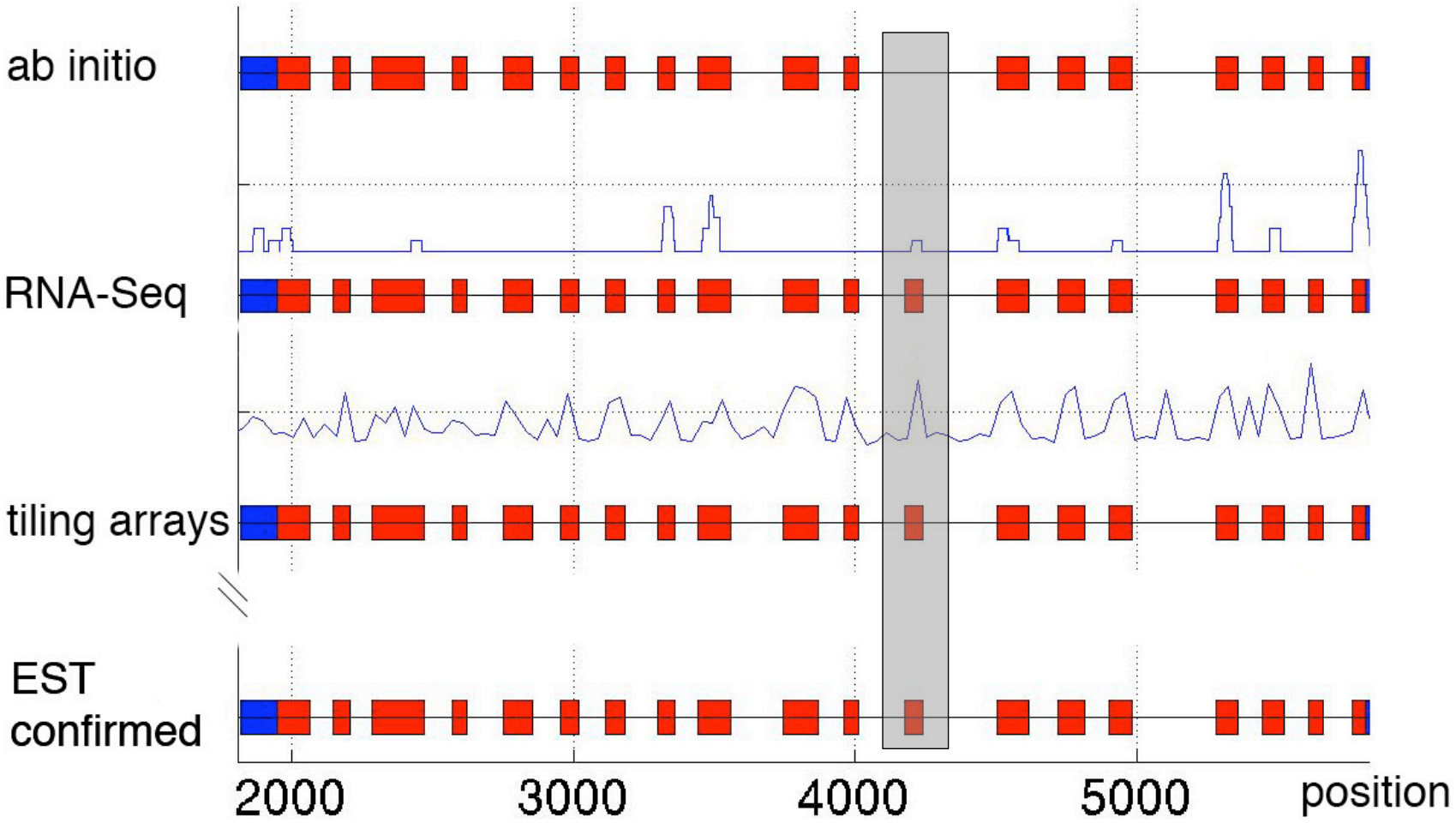
Results: Combining helps!

[Behr et al., 2009, in prep.]



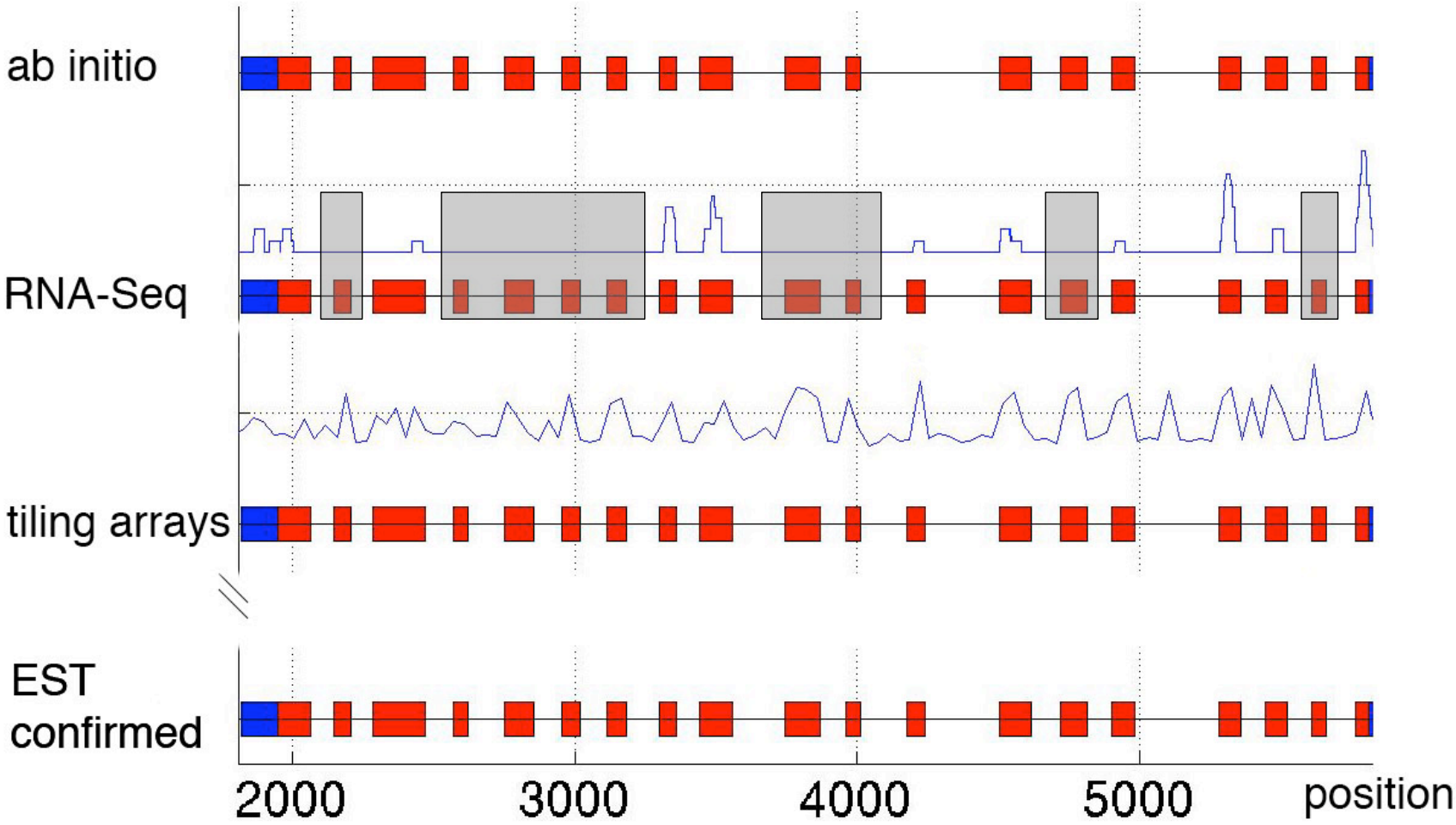
Results: Combining helps!

[Behr et al., 2009, in prep.]



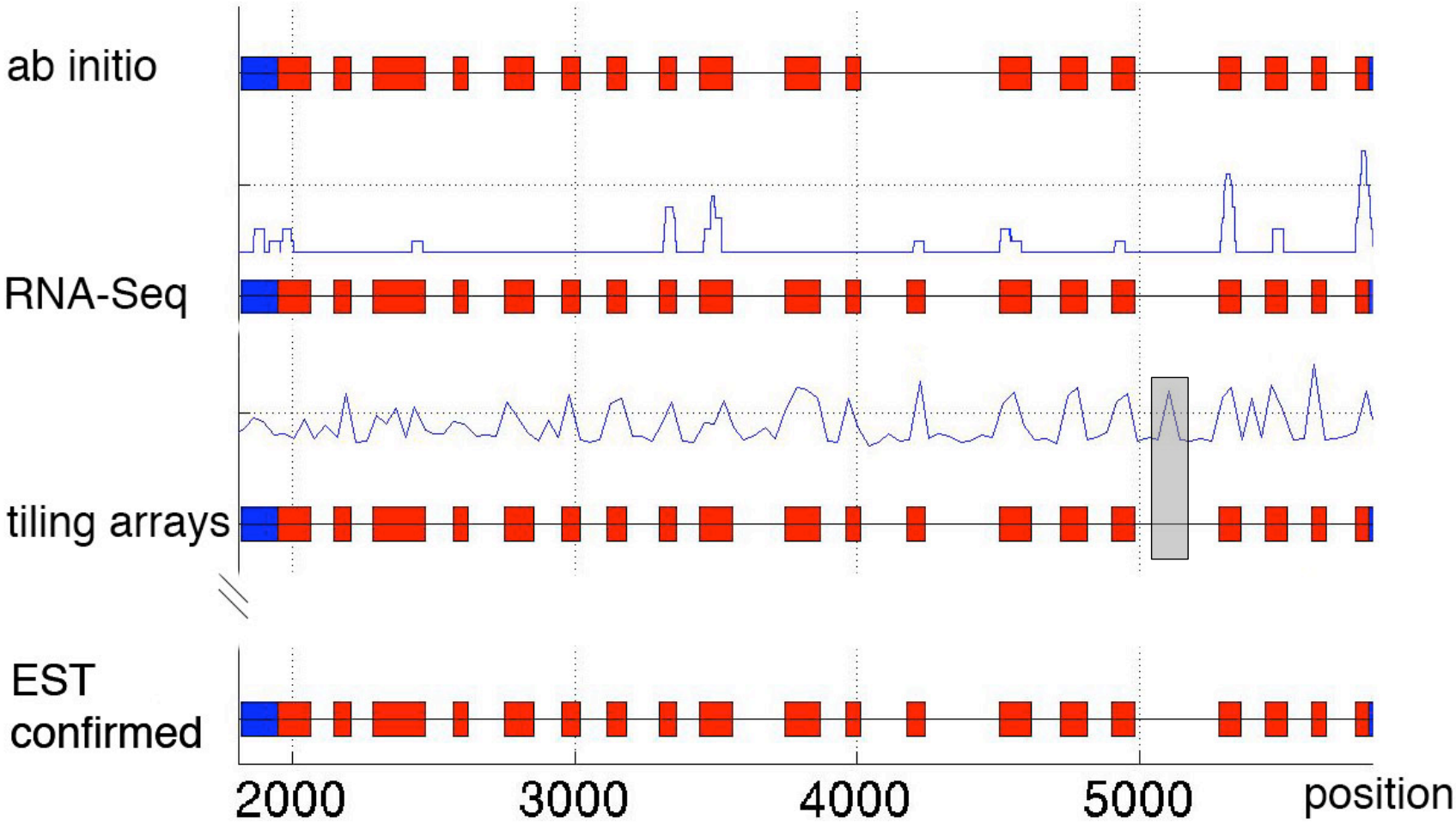
Results: Combining helps!

[Behr et al., 2009, in prep.]



Results: Combining helps!

[Behr et al., 2009, in prep.]



Gene Finding + Auxiliary Measurements

[Behr et al., 2009, in prep.]



So far: Transcriptome measurements

Is there other information that can improve prediction results?

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ...with tiling arrays (11 tissues)	**82.1%**
3. ...with mRNA-seq (1 tissue)	**81.1%**
4. ...with DNA methylation (1 tissue)	**76.1%**
5. ...with Nucleosome position predictions	**78.0%**
6. ...with RNA secondary structure predictions	**76.7%**
7. ...with all information combined	**84.3%**

Gene Finding + Auxiliary Measurements

[Behr et al., 2009, in prep.]



So far: Transcriptome measurements

Is there other information that can improve prediction results?

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ...with tiling arrays (11 tissues)	**82.1%**
3. ...with mRNA-seq (1 tissue)	**81.1%**
4. ...with DNA methylation (1 tissue)	**76.1%**
5. ...with Nucleosome position predictions	**78.0%**
6. ...with RNA secondary structure predictions	**76.7%**
7. ...with all information combined	**84.3%**

Gene Finding + Auxiliary Measurements

[Behr et al., 2009, in prep.]



So far: Transcriptome measurements

Is there other information that can improve prediction results?

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ...with tiling arrays (11 tissues)	**82.1%**
3. ...with mRNA-seq (1 tissue)	**81.1%**
4. ...with DNA methylation (1 tissue)	**76.1%**
5. ...with Nucleosome position predictions	**78.0%**
6. ...with RNA secondary structure predictions	**76.7%**
7. ...with all information combined	**84.3%**

Gene Finding + Auxiliary Measurements

[Behr et al., 2009, in prep.]



So far: Transcriptome measurements

Is there other information that can improve prediction results?

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ... with tiling arrays (11 tissues)	**82.1%**
3. ... with mRNA-seq (1 tissue)	**81.1%**
4. ... with DNA methylation (1 tissue)	**76.1%**
5. ... with Nucleosome position predictions	**78.0%**
6. ... with RNA secondary structure predictions	**76.7%**
7. ... with all information combined	**84.3%**

Gene Finding + Auxiliary Measurements

[Behr et al., 2009, in prep.]



So far: Transcriptome measurements

Is there other information that can improve prediction results?

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ... with tiling arrays (11 tissues)	**82.1%**
3. ... with mRNA-seq (1 tissue)	**81.1%**
4. ... with DNA methylation (1 tissue)	**76.1%**
5. ... with Nucleosome position predictions	**78.0%**
6. ... with RNA secondary structure predictions	**76.7%**
7. ... with all information combined	**84.3%**

Conclusions

- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Improved accuracy due to novel inference methods
  - Limitations: no alternative transcripts or expression information
- Analysis of transcriptome measurements
  - Identification of alternative and differential transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - 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 due to novel inference methods
  - Limitations: no alternative transcripts or expression information
- Analysis of transcriptome measurements
  - Identification of alternative and differential transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - 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 due to novel inference methods
  - Limitations: no alternative transcripts or expression information
- Analysis of transcriptome measurements
  - Identification of alternative and differential transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - 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 due to novel inference methods
  - Limitations: no alternative transcripts or expression information
- Analysis of transcriptome measurements
  - Identification of alternative and differential transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

Acknowledgments

Gene Finding

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

Short Read Analysis

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

Tiling Arrays

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

Machine Learning

- Sören Sonnenburg (FhG Berlin)
- Bernhard Schölkopf (MPI)

Nematode Material/Methods

- Adrian Streit (MPI)

Acknowledgments

Gene Finding

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

Tiling Arrays

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

Short Read Analysis

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

Machine Learning

- Sören Sonnenburg (FhG Berlin)
- Bernhard Schölkopf (MPI)

Nematode Material/Methods

- Adrian Streit (MPI)

Acknowledgments

Gene Finding

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

Short Read Analysis

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

Tiling Arrays

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

Machine Learning

- Sören Sonnenburg (FhG Berlin)
- Bernhard Schölkopf (MPI)

Nematode Material/Methods

- Adrian Streit (MPI)

Acknowledgments



Gene Finding

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

Short Read Analysis

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

Tiling Arrays

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

Machine Learning

- Sören Sonnenburg (FhG Berlin)
- Bernhard Schölkopf (MPI)

Nematode Material/Methods

- Adrian Streit (MPI)

Acknowledgments



Gene Finding

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

Tiling Arrays

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

Short Read Analysis

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

Machine Learning

- Sören Sonnenburg (FhG Berlin)
- Bernhard Schölkopf (MPI)

Nematode Material/Methods

- Adrian Streit (MPI)

Slides are available online: www.fml.mpg.de/raetsch/talks

Acknowledgments



Gene Finding

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

Short Read Analysis

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

Tiling Arrays

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

Machine Learning

- Sören Sonnenburg (FhG Berlin)
- Bernhard Schölkopf (MPI)

Nematode Material/Methods

- Adrian Streit (MPI)

Positions available: www.fml.mpg.de/raetsch/offers

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. Presented at the CSHL Genome Informatics Meeting, May 2008.

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. Presented at the CSHL Genome Informatics Meeting, July 2009, in prep.

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.

Ralf Dahm. Discovering dna: Friedrich miescher and the early years of nucleic acid research. *Hum Genet*, 122(6):565–581, Jan 2008. ISSN 1432-1203 (Electronic). doi: 10.1007/s00439-007-0433-0.

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.

References II

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

V. Franc and S. Sonnenburg. Optimized cutting plane algorithm for large-scale risk minimization. *Journal of Machine Learning Research*, 2009. in press.

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.

S. Laubinger, G. Zeller, S. Henz, G. Rätsch, and D. Weigel. Transcriptome analysis of small rna biogenesis mutants. submitted, June 2009, submitted.

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.

References III

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.

Gabriele Schweikert, Jonas Behr, Alexander Zien, Georg Zeller, Cheng Soon Ong, Sören Sonnenburg, and Gunnar Rätsch. mgene.web: a web service for accurate computational gene finding. *Nucleic Acids Research*, Web Server Issue, 2009a. URL `http://mgene.org/web`. Advance Access published on June 3, 2009.

Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krüger, Sören Sonnenburg, and Gunnar Rätsch. mgene: Accurate svm-based gene finding with an application to nematode genomes. *Genome Research*, 2009b. URL `http://genome.cshlp.org/content/early/2009/06/29/gr.090597.108.full.pdf+html`. Advance access June 29, 2009.

S. Sonnenburg, G. Rätsch, A. Jagota, and K.-R. Müller. New methods for splice-site recognition. In *Proc. International Conference on Artificial Neural Networks*, 2002.

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

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, 2007b. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-8-S10-S7.

References IV

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

Sören Sonnenburg, Mikio L Braun, Cheng Soon Ong, Samy Bengio, Leon Bottou, Geoffrey Holmes, Yann LeCun, Klaus-Robert Müller, Fernando Pereira, Carl Edward Rasmussen, Gunnar Rätsch, Bernhard Schölkopf, Alexander Smola, Pascal Vincent, Jason Weston, and Robert C Williamson. The need for open source software in machine learning. *Journal of Machine Learning Research*, 8:2443–2466, 2007c. URL `http://jmlr.csail.mit.edu/papers/v8/sonnenburg07a.html`.

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.

Georg Zeller, Stefan R Henz, Christian K Widmer, Timo Sachsenberg, Gunnar Rätsch, Detlef Weigel, and Sascha Laubinger. Stress-induced changes in the arabidopsis thaliana transcriptome analyzed using whole genome tiling arrays. *Plant Journal*, 2009. URL `http://www3.interscience.wiley.com/resolve/doi/abs?DOI=10.1111/j.1365-313X.2009.03835.x`.

References V

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.