Novel Methods for Computational Transcriptome Analysis

Gunnar Rätsch

Friedrich Miescher Laboratory
Max Planck Society
Tübingen, Germany

NBIC Conference 2009
March 17, 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)









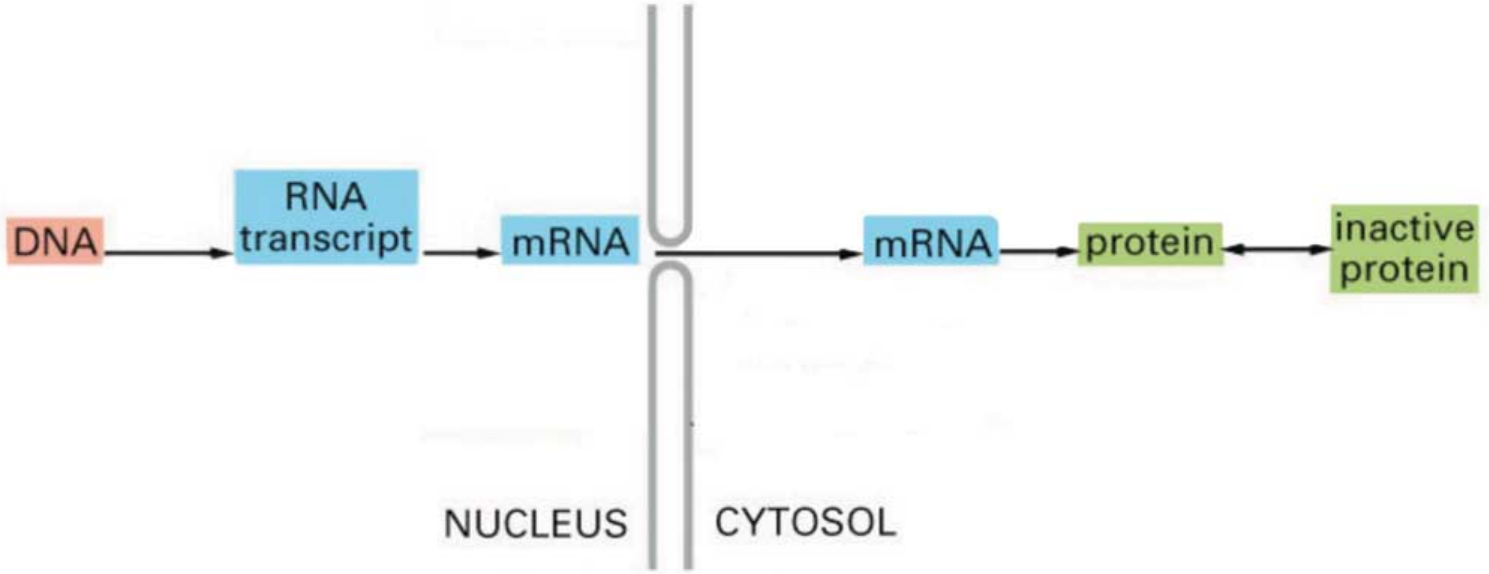
Discovery of Nuclein:

- 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]

Transcriptome Analysis

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



Then we can (try to) understand:

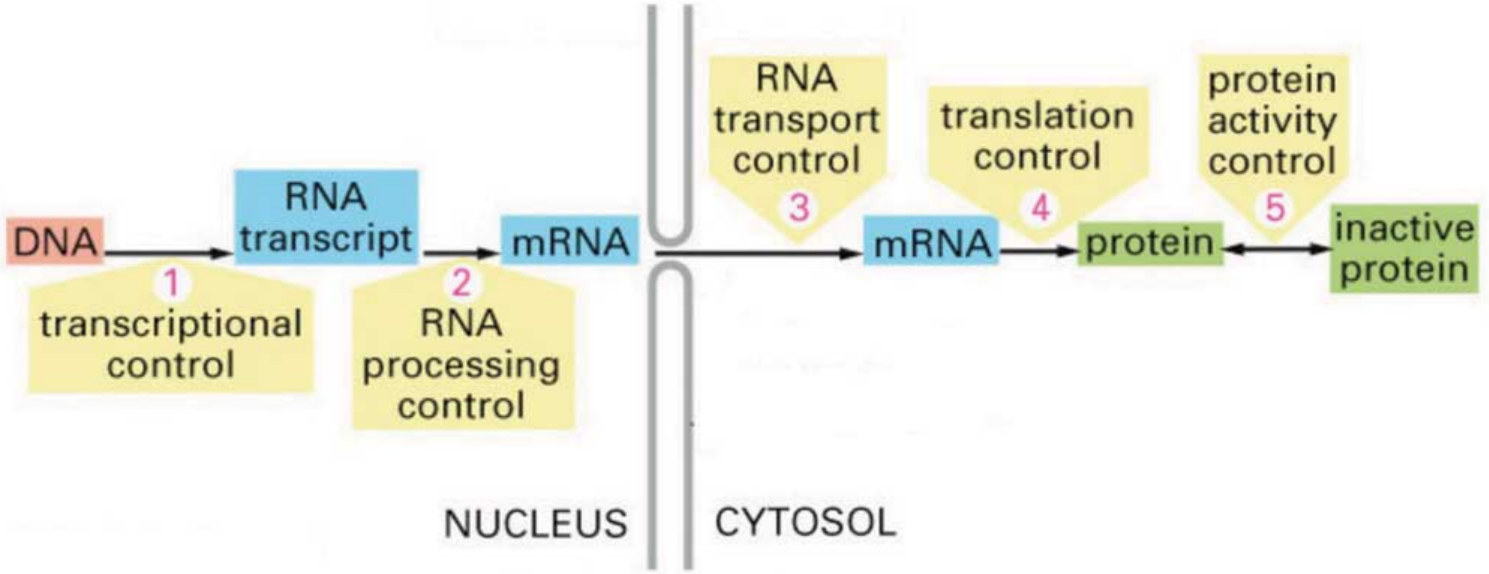
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

- Infer transcriptome from genome DNA
- Measure properties of transcriptome
- Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

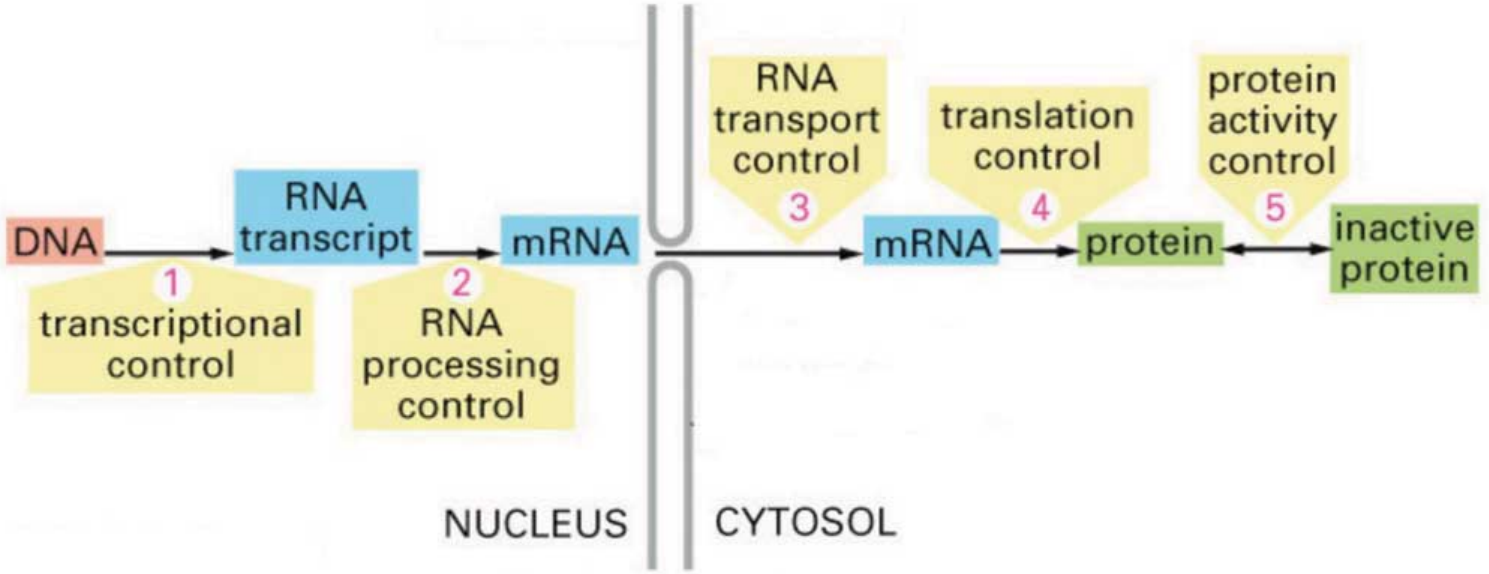
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

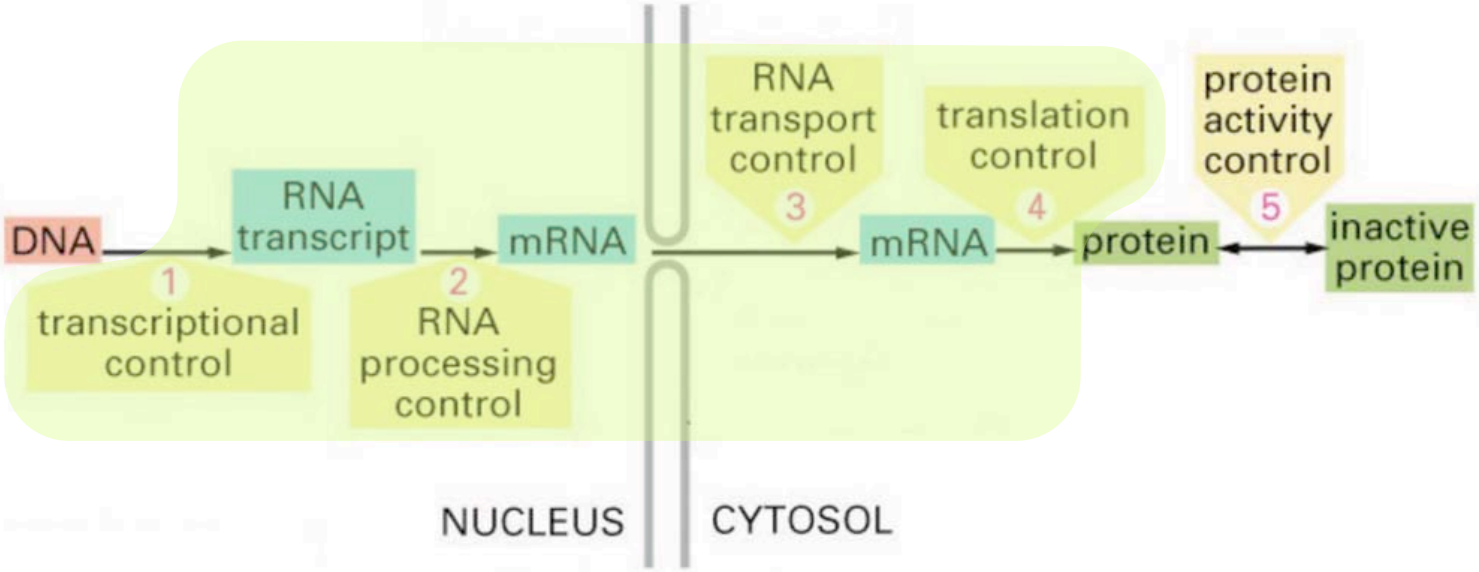
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

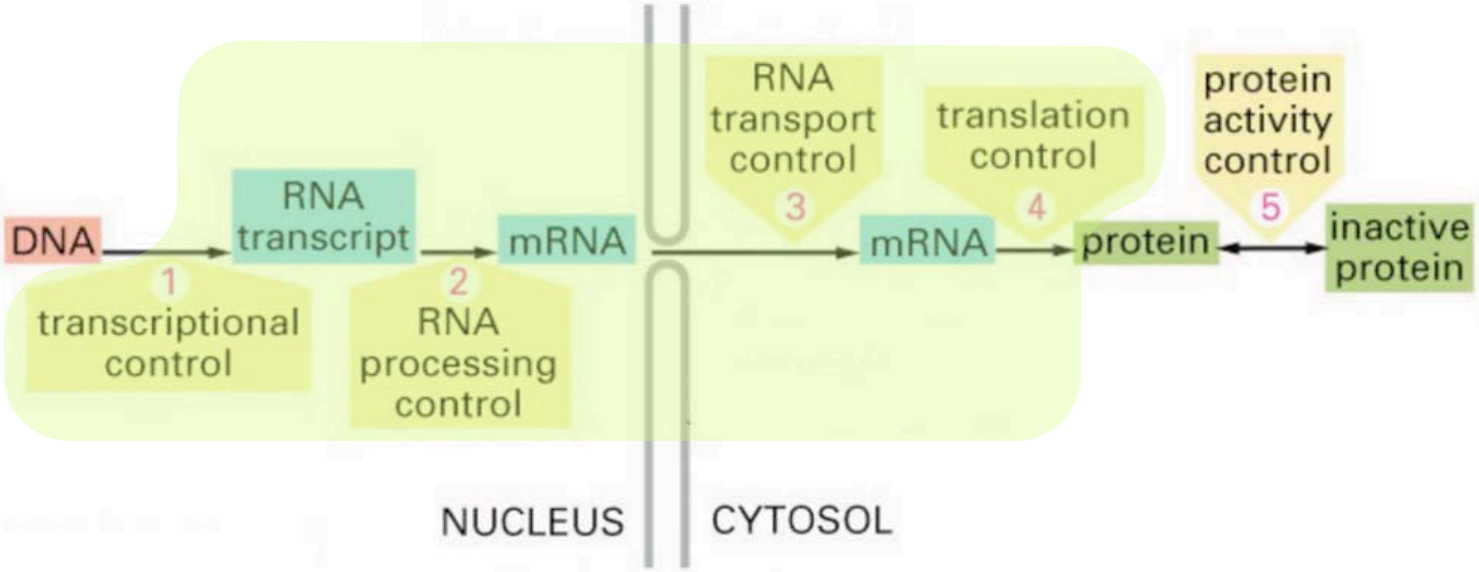
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

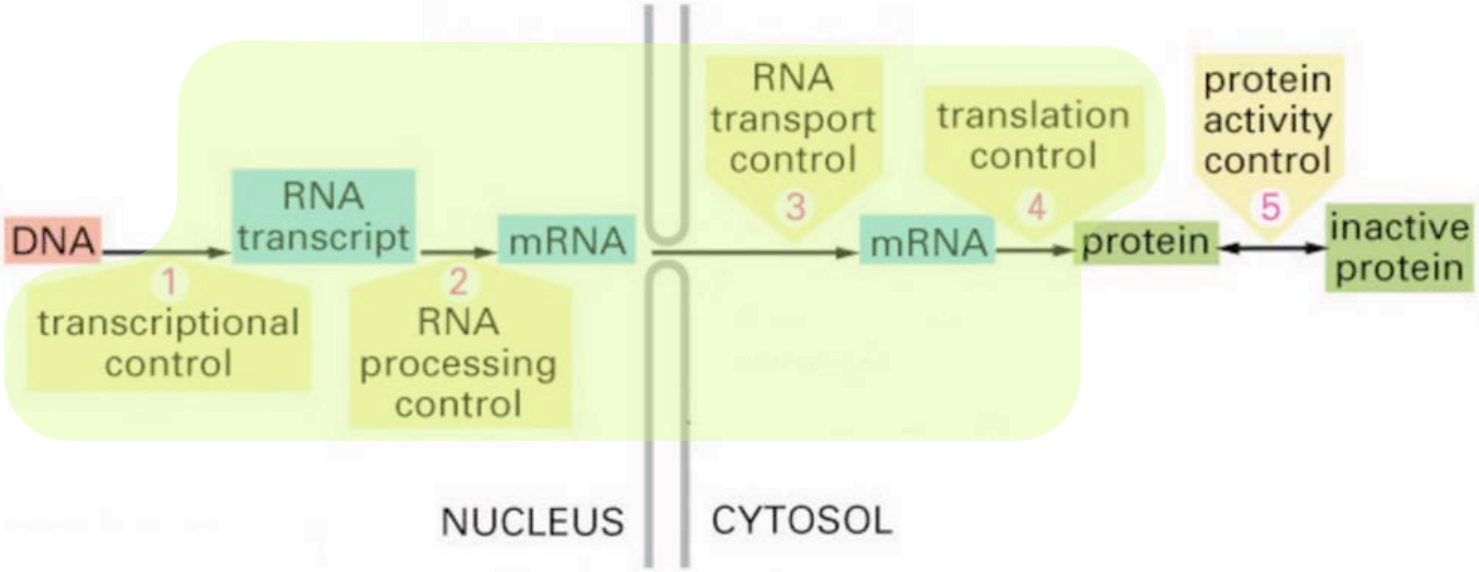
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

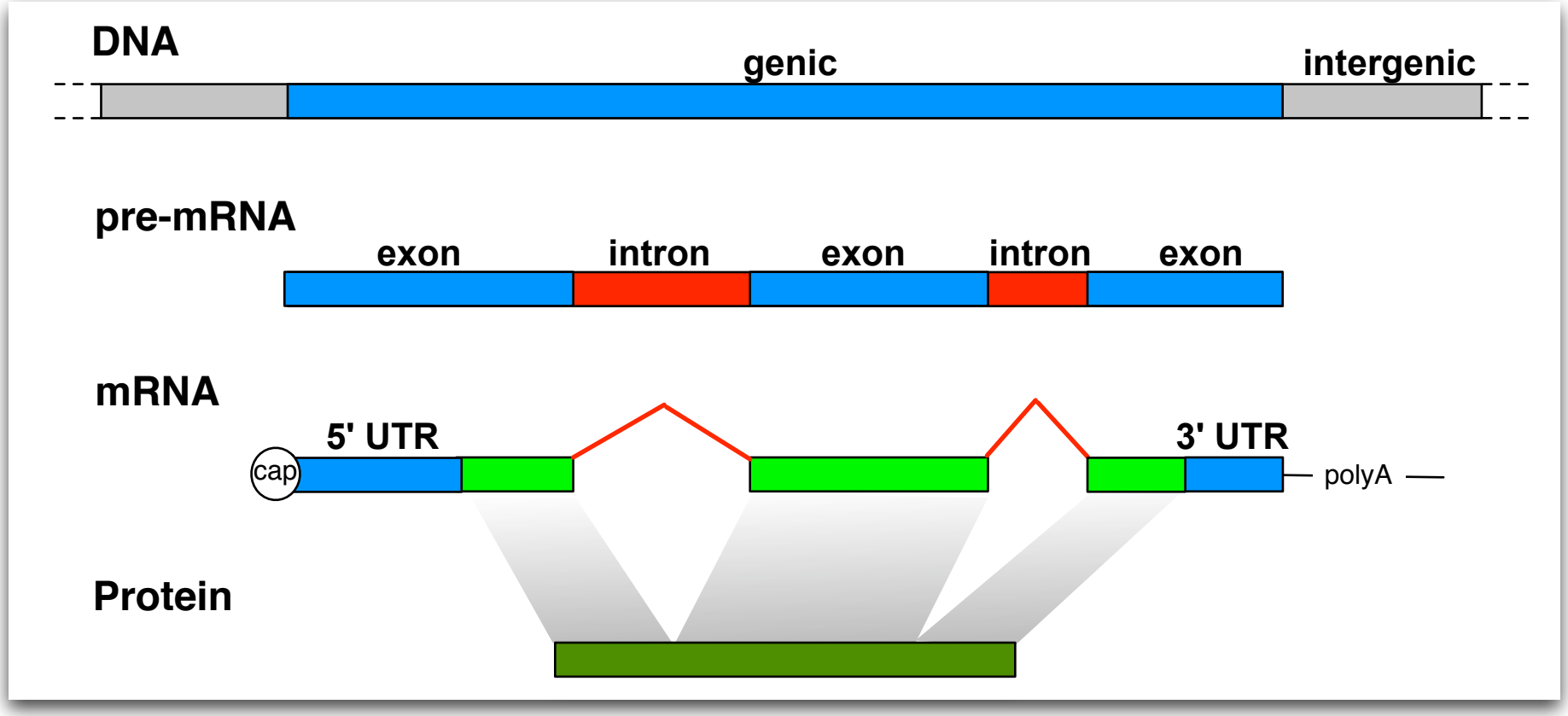
Computational Gene Finding



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

Computational Gene Finding

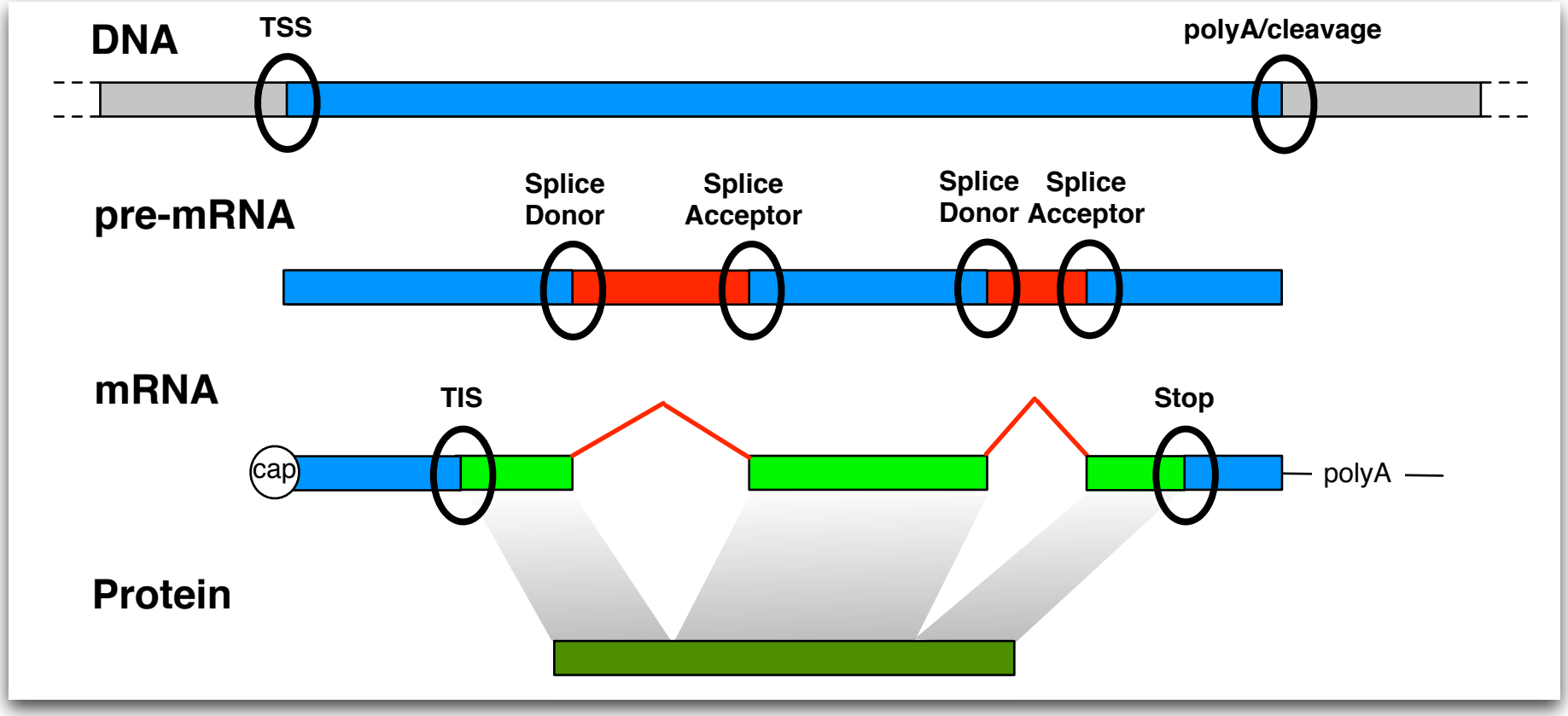
⇝ Labeling the Genome



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

Computational Gene Finding

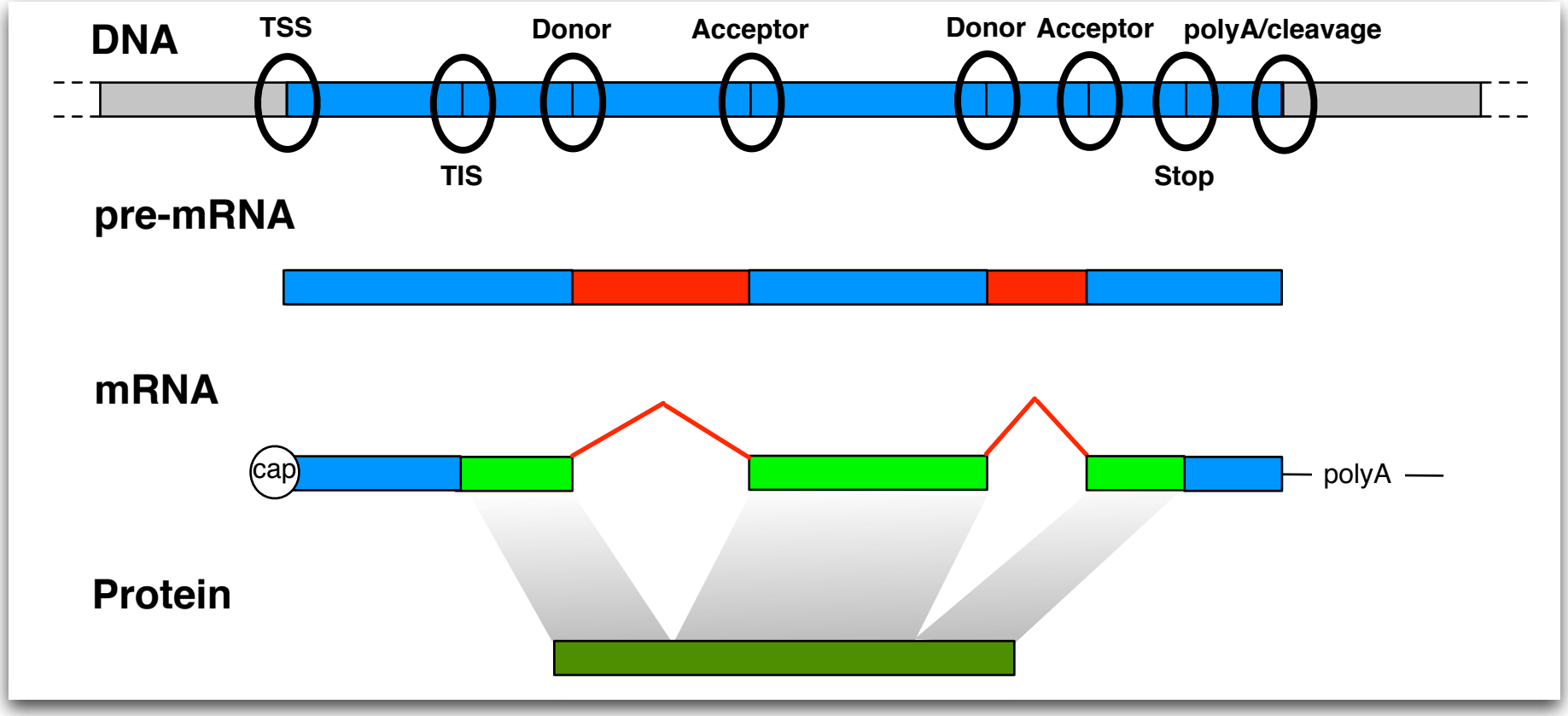
⇝ Labeling the Genome



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

Computational Gene Finding

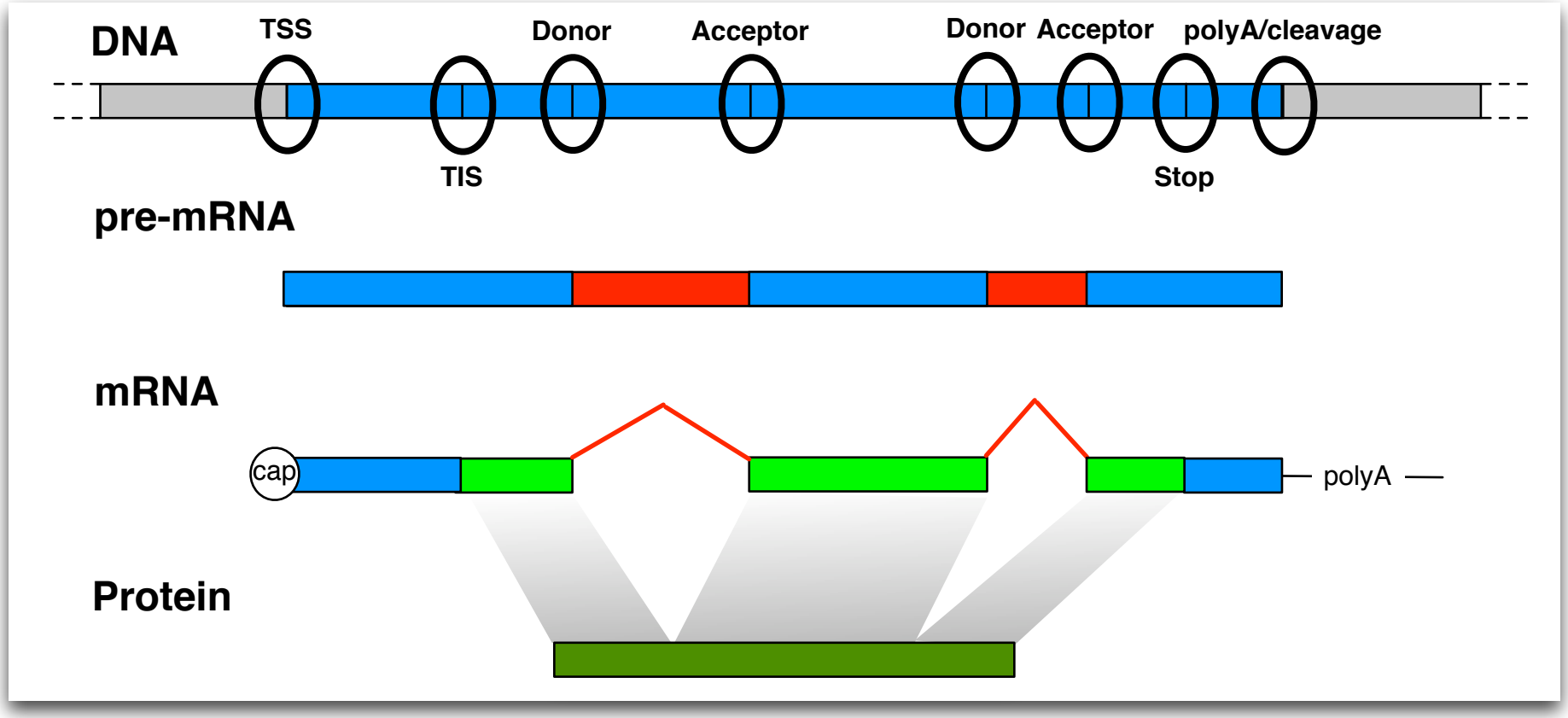
⇝ Labeling the Genome

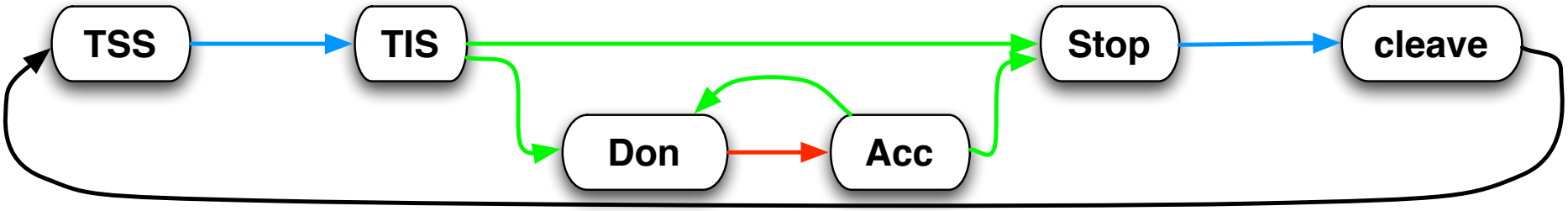


- Given a piece of DNA sequence
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

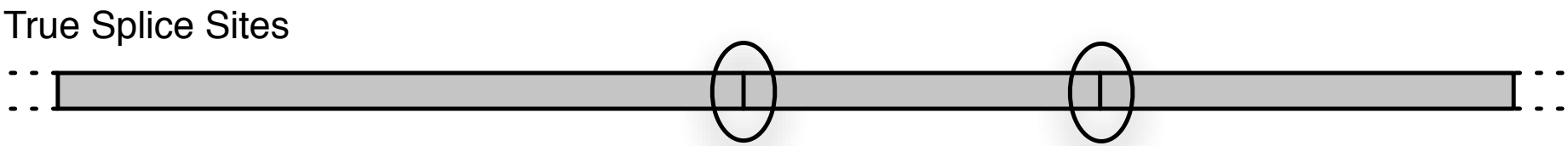
Computational Gene Finding

⇝ Labeling the Genome





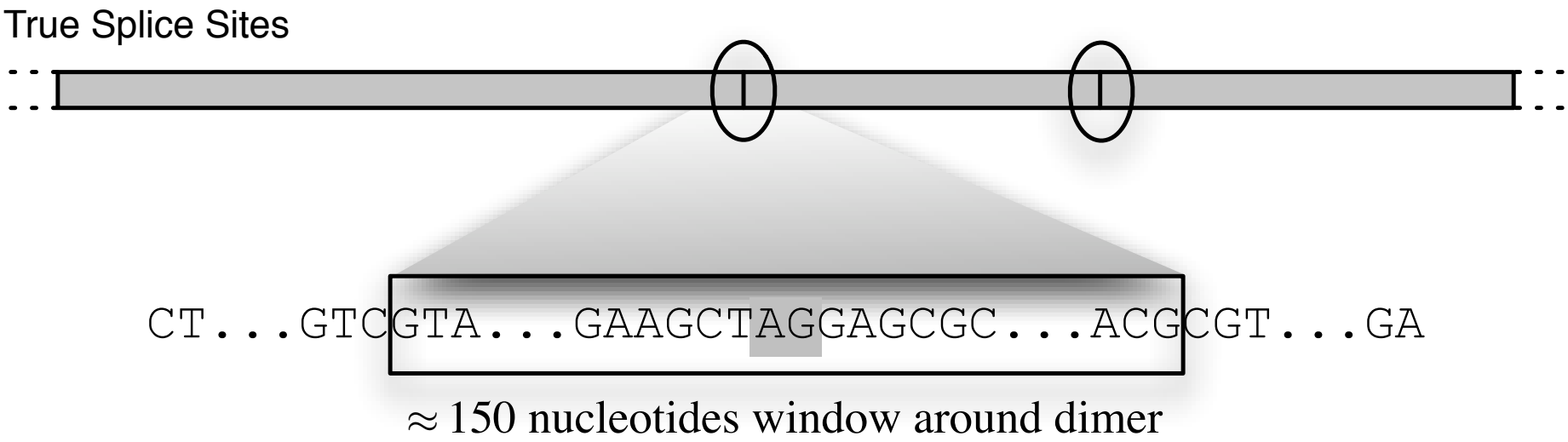
Example: Splice Site Recognition



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

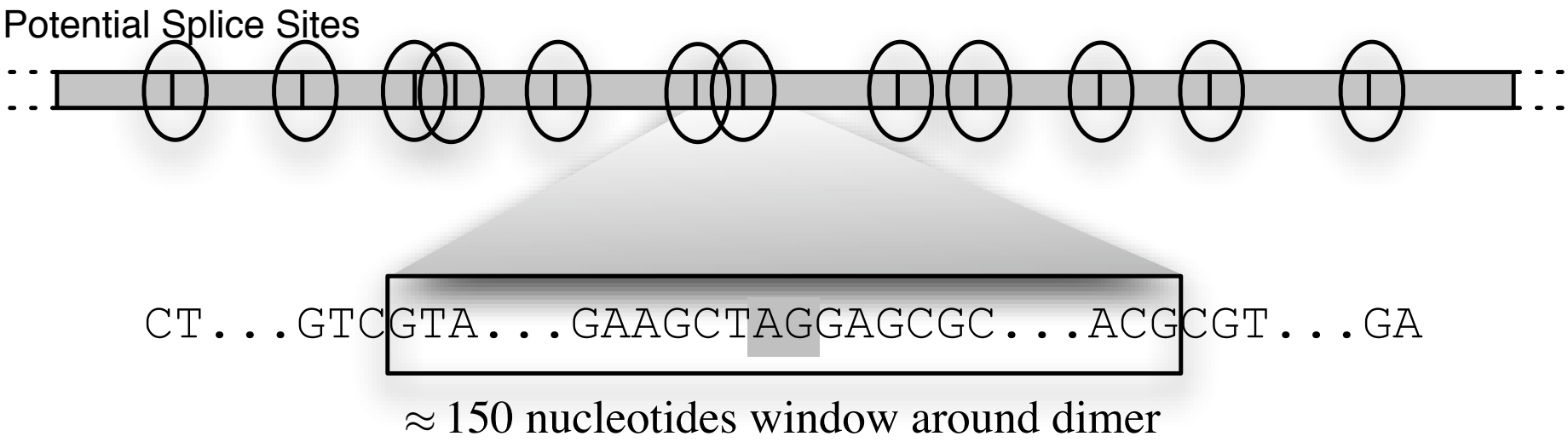
$\Rightarrow$ **Millions of labeled instances from EST databases**

Example: Splice Site Recognition



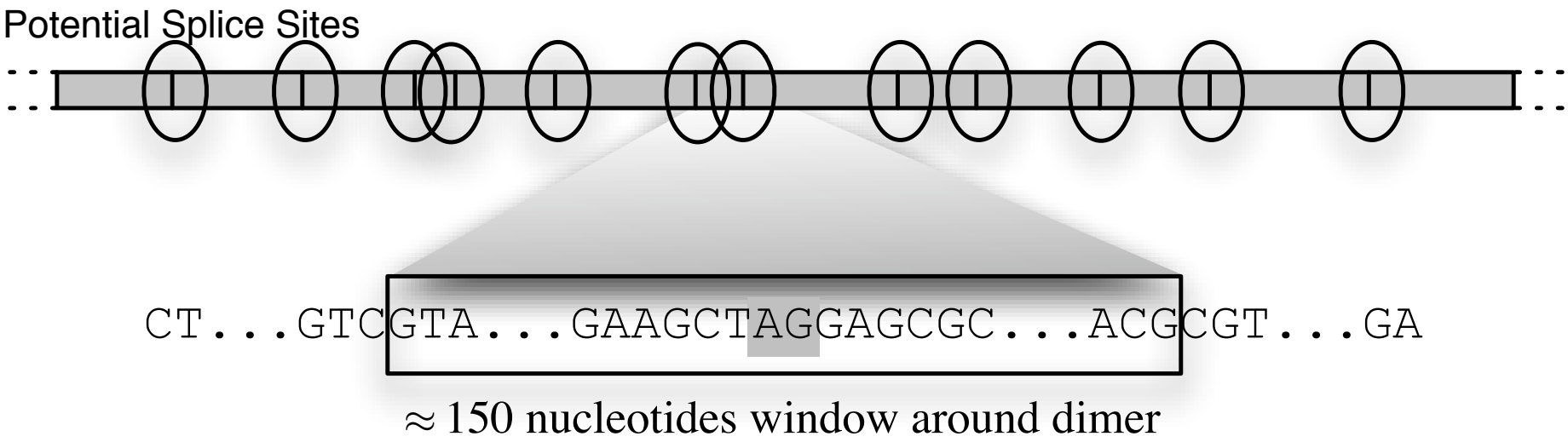
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

Example: Splice Site Recognition



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

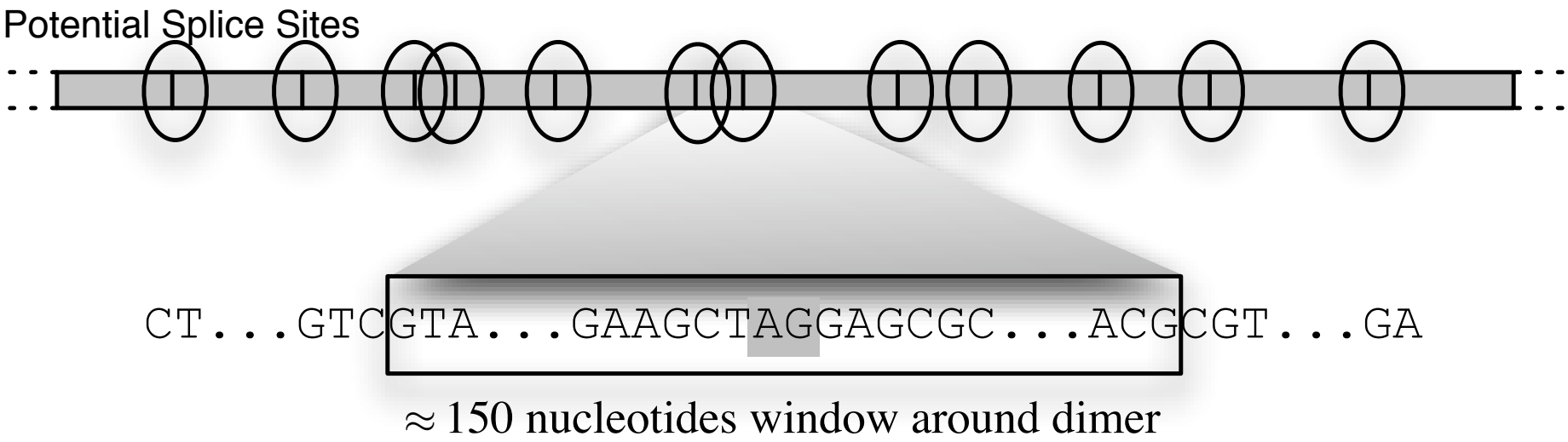
Example: Splice Site Recognition



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

$\Rightarrow$ Millions of labeled instances from EST databases

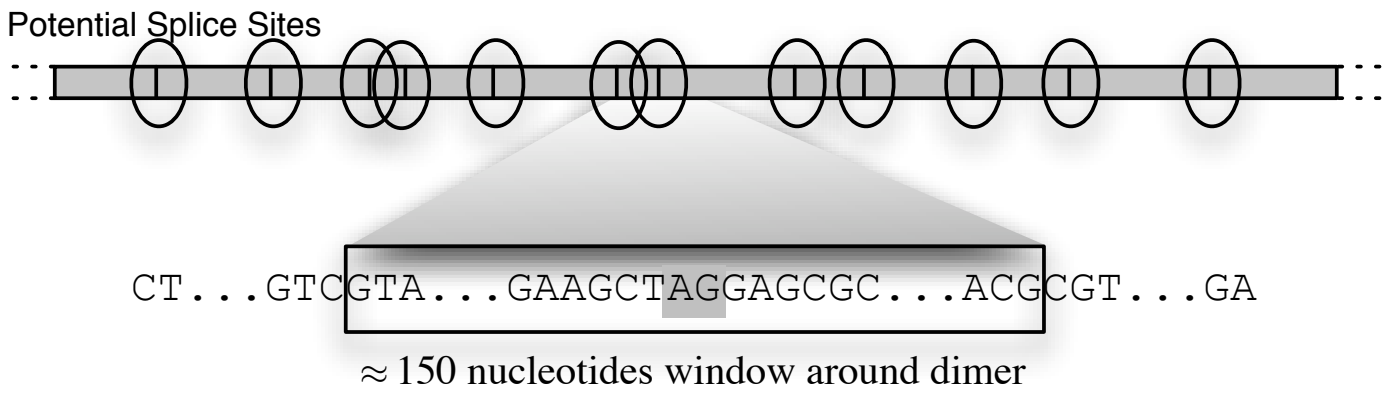
Example: Splice Site Recognition



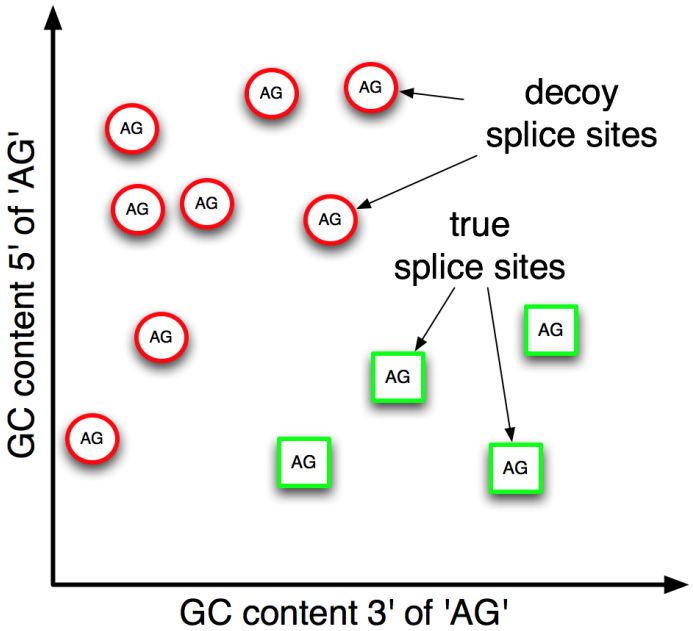
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

$\Rightarrow$ **Millions of labeled instances from EST databases**

Example: Splice Site Recognition



Basic idea:



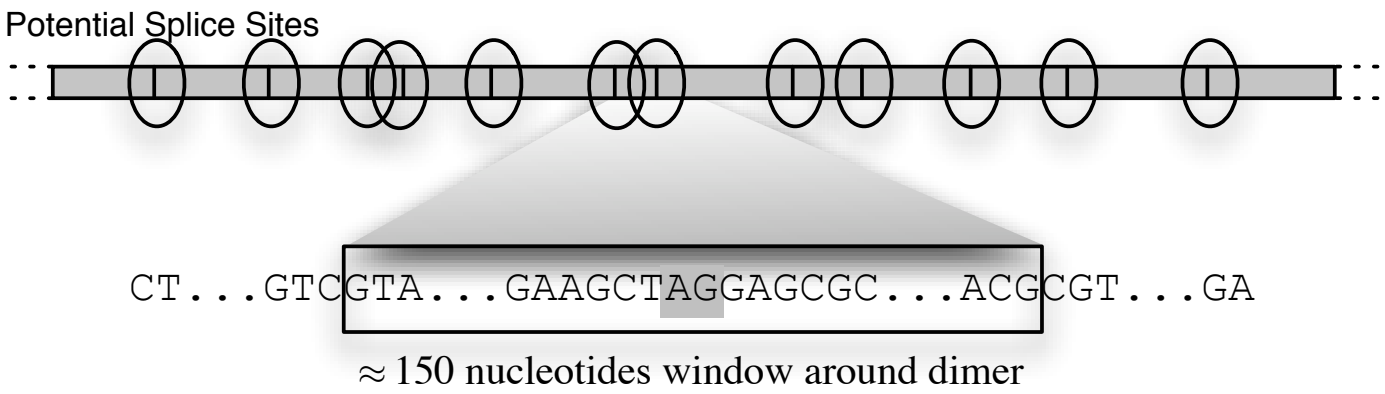
For instance, exploit that exons have higher GC content

or

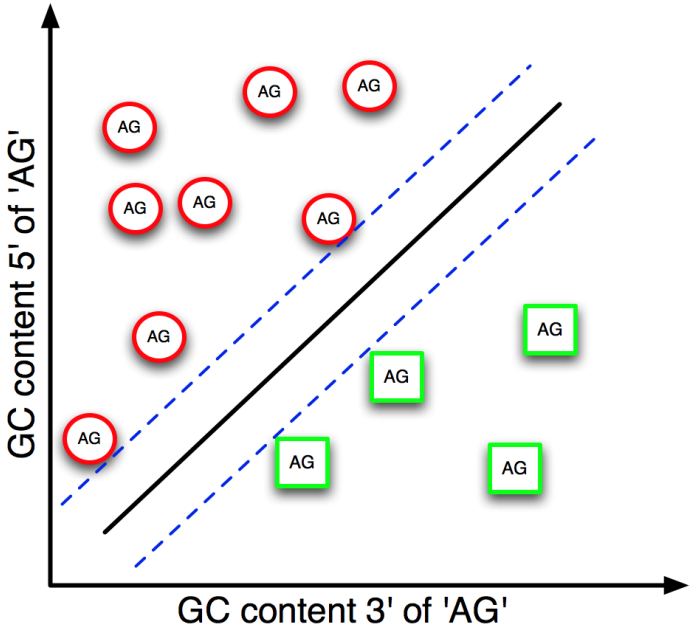
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

Example: Splice Site Recognition



Basic idea:



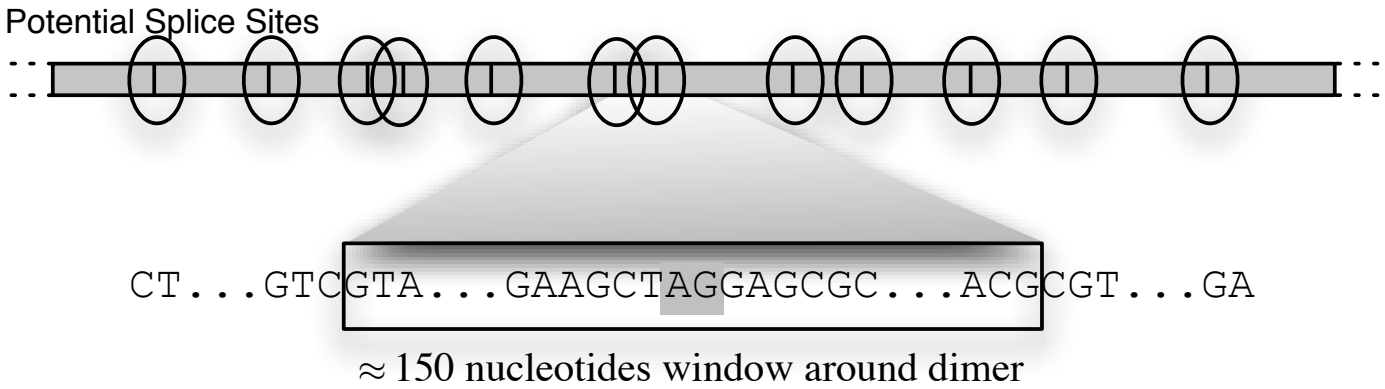
For instance, exploit that exons have higher GC content

or

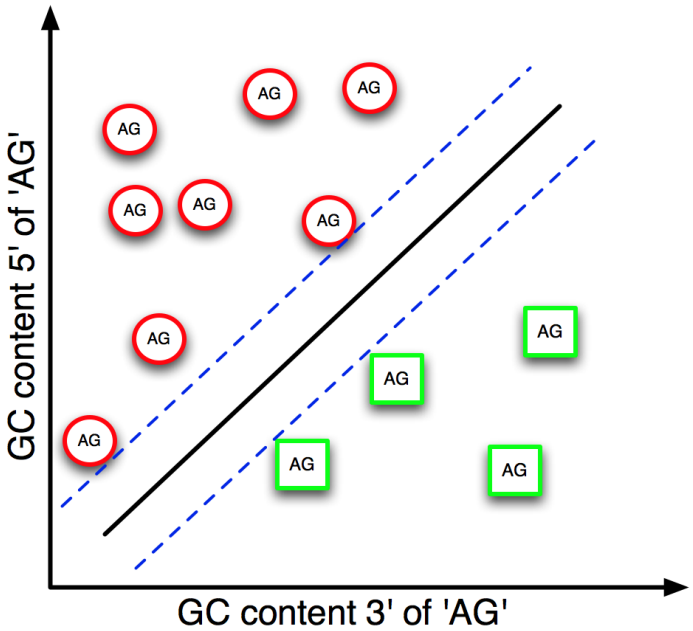
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

Example: Splice Site Recognition



Basic idea:

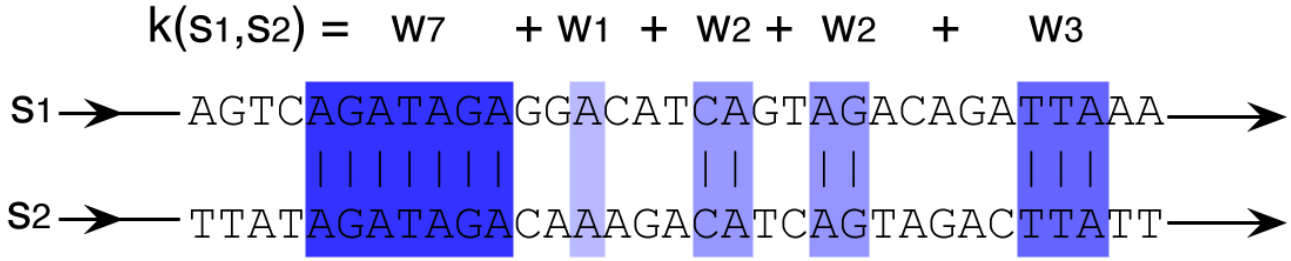


In practice: Use one feature per possible substring (e.g. $\leq$20) at all positions

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

Kernels and String Data Structures

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



[Rätsch & Sonnenburg, 2004; Rätsch et al., 2007]

Bad news: Need to compute kernel between all pairs of sequences

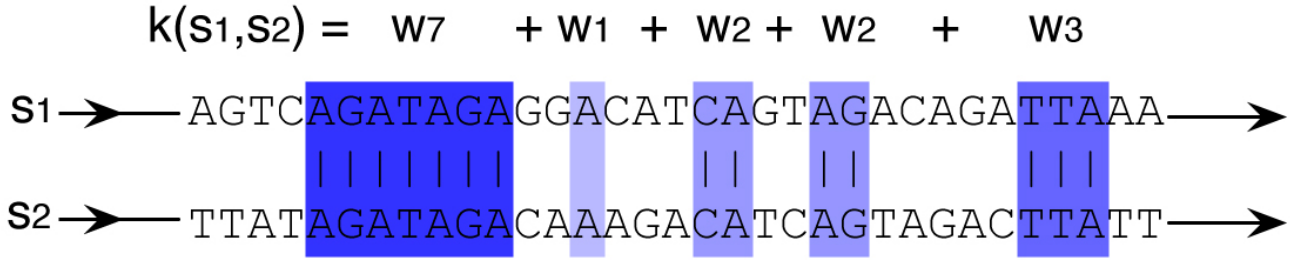
- $1,000,000^2$ examples $\rightsquigarrow$ 10^{12} **kernel elements :-(**
- Needs sophisticated algorithms
  - Iteratively consider smaller sub-problems [Schölkopf & Smola, 2006]
    - SMO type optimization
  - Exploit that features are sparse
    - String indexing data structures

Implemented in a freely available software package

`http://www.shogun-toolbox.org` [Sonnenburg, Rieck, Rätsch, 2007]

Kernels and String Data Structures

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



[Rätsch & Sonnenburg, 2004; Rätsch et al., 2007]

Bad news: Need to compute kernel between all pairs of sequences

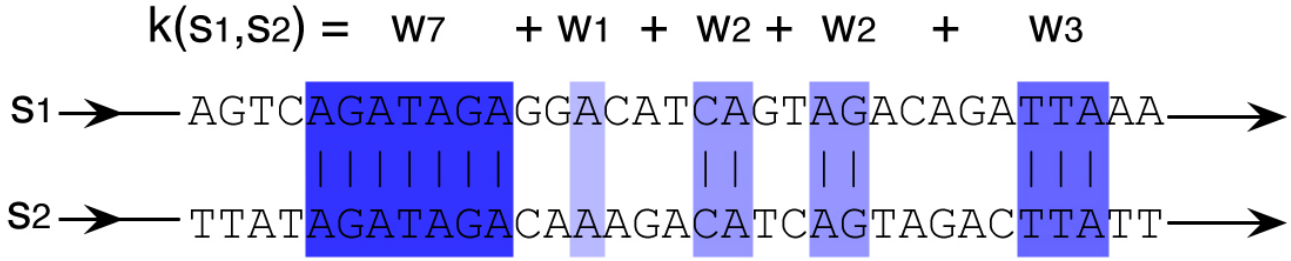
- $1,000,000^2$ examples $\rightsquigarrow$ 10^{12} **kernel elements :-(**
- Needs sophisticated algorithms
  - Iteratively consider smaller sub-problems [Schölkopf & Smola, 2006]
    - SMO type optimization
  - Exploit that features are sparse
    - String indexing data structures

Implemented in a freely available software package

`http://www.shogun-toolbox.org` [Sonnenburg, Rieck, Rätsch, 2007]

Kernels and String Data Structures

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



[Rätsch & Sonnenburg, 2004; Rätsch et al., 2007]

Bad news: Need to compute kernel between all pairs of sequences

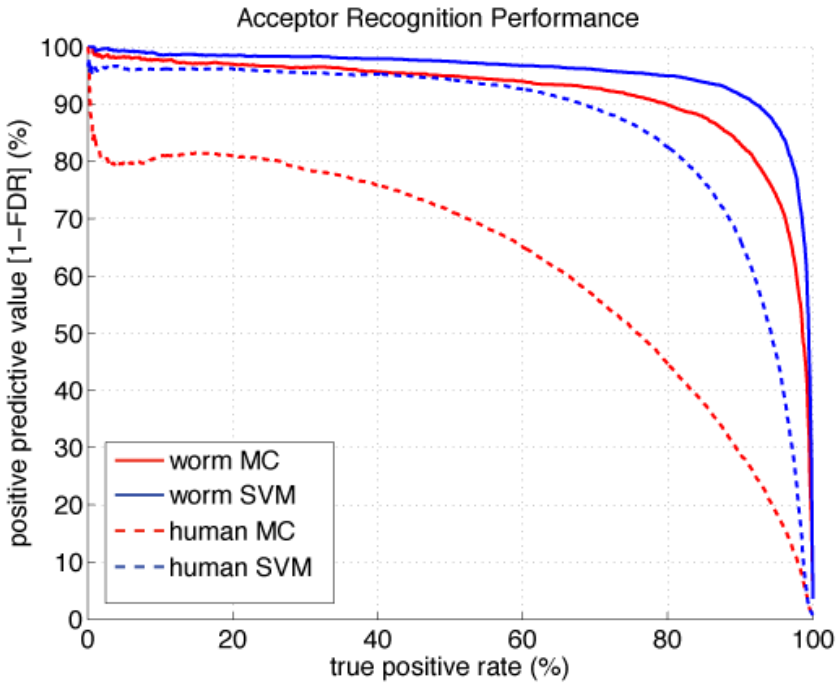
- $1,000,000^2$ examples $\rightsquigarrow$ 10^{12} **kernel elements :-(**
- Needs sophisticated algorithms
  - Iteratively consider smaller sub-problems [Schölkopf & Smola, 2006]
    - SMO type optimization
  - Exploit that features are sparse
    - String indexing data structures

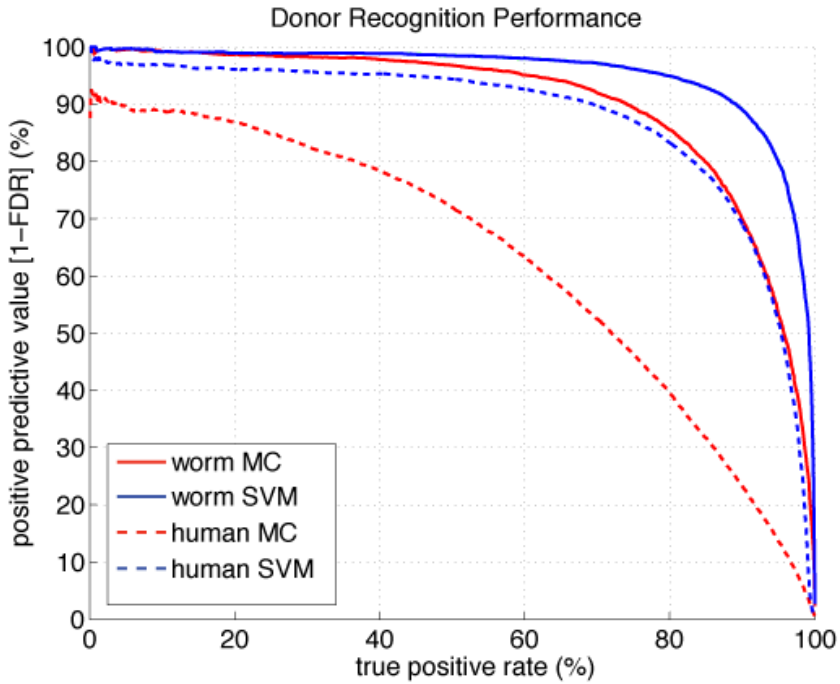
Implemented in a freely available software package

`http://www.shogun-toolbox.org` [Sonnenburg, Rieck, Rätsch, 2007]

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

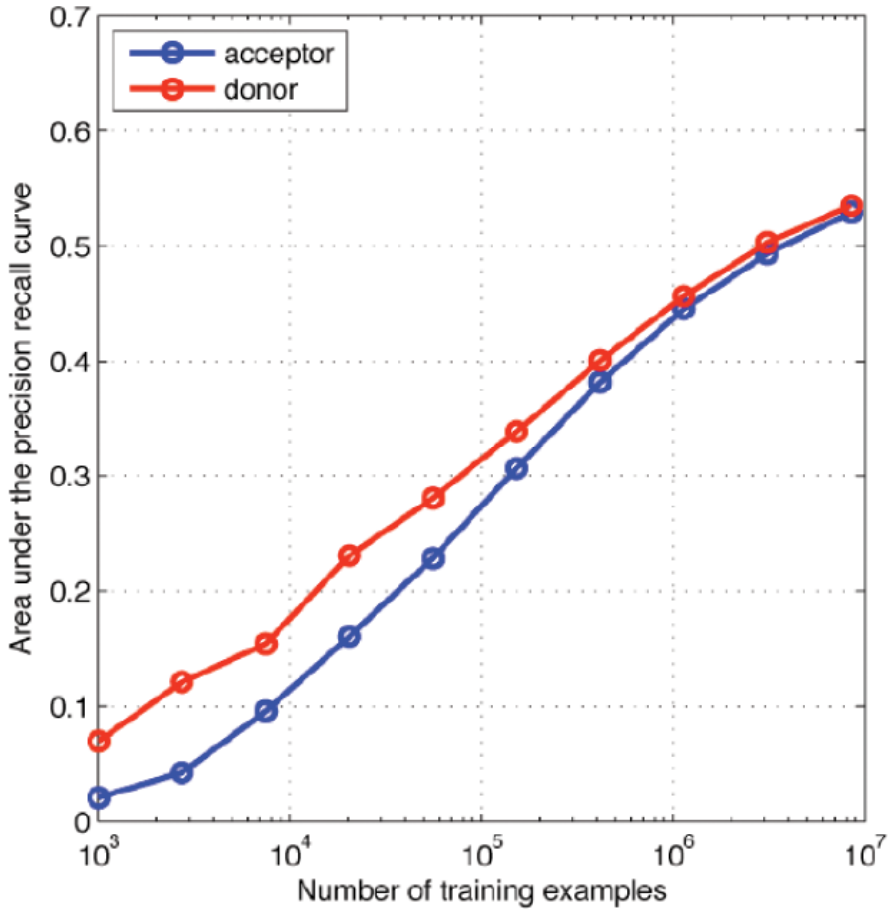


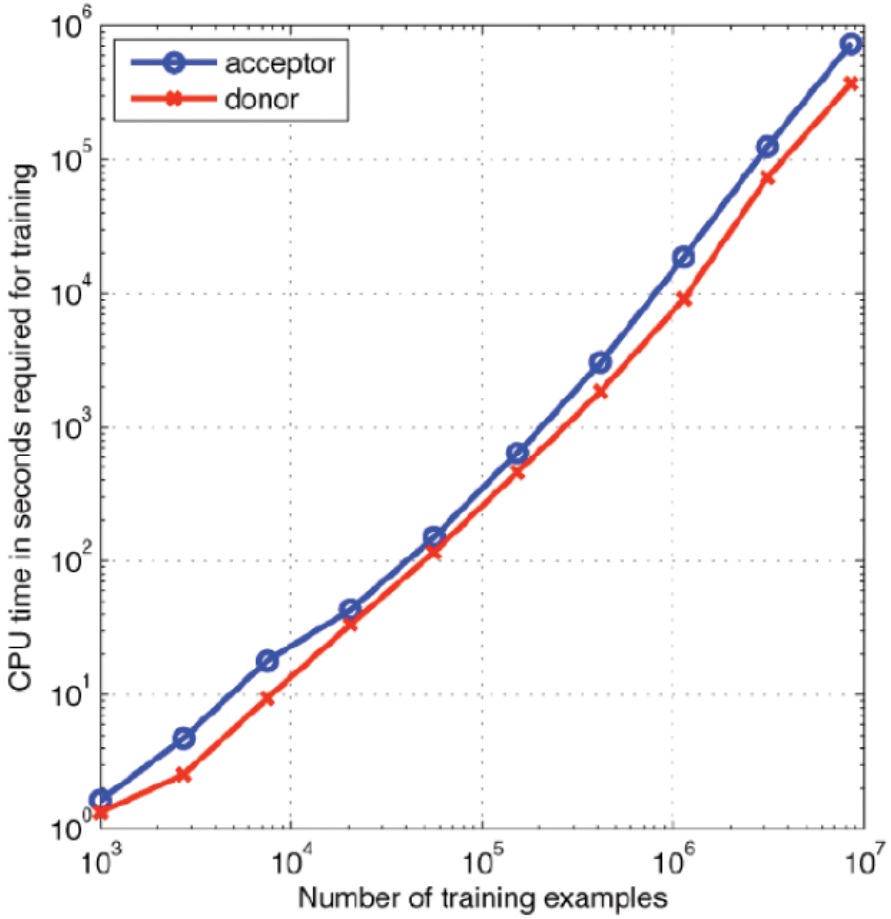


[Sonnenburg, Schweikert, Philips, Behr, Rätsch, 2007]

Human Splice Sites with WD Kernel

[Sonnenburg et al., 2007]



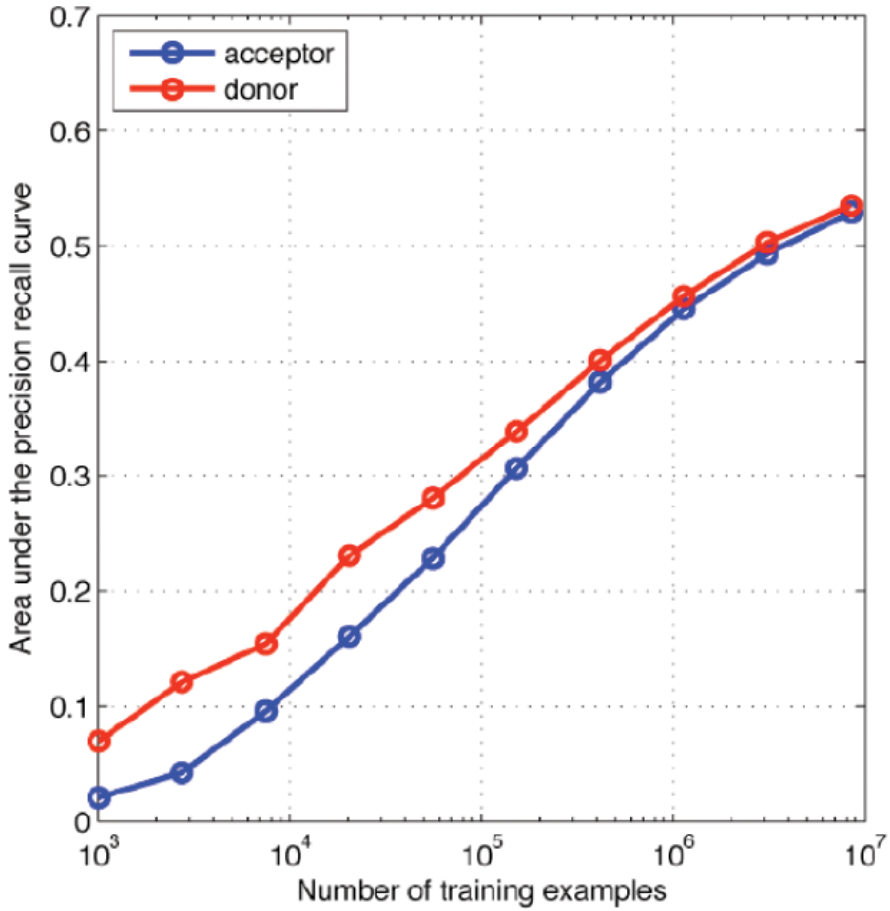


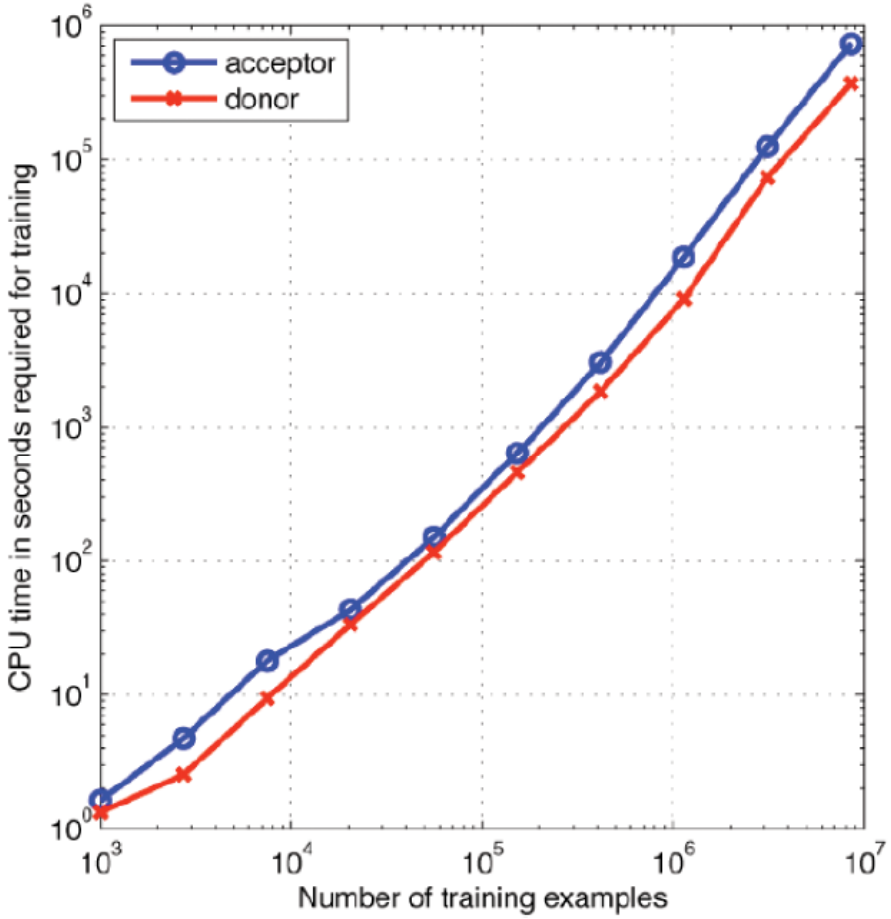
Performance (=Area under PRC curve) keeps increasing.

$\Rightarrow$ **Use all available data!**

Human Splice Sites with WD Kernel

[Sonnenburg et al., 2007]

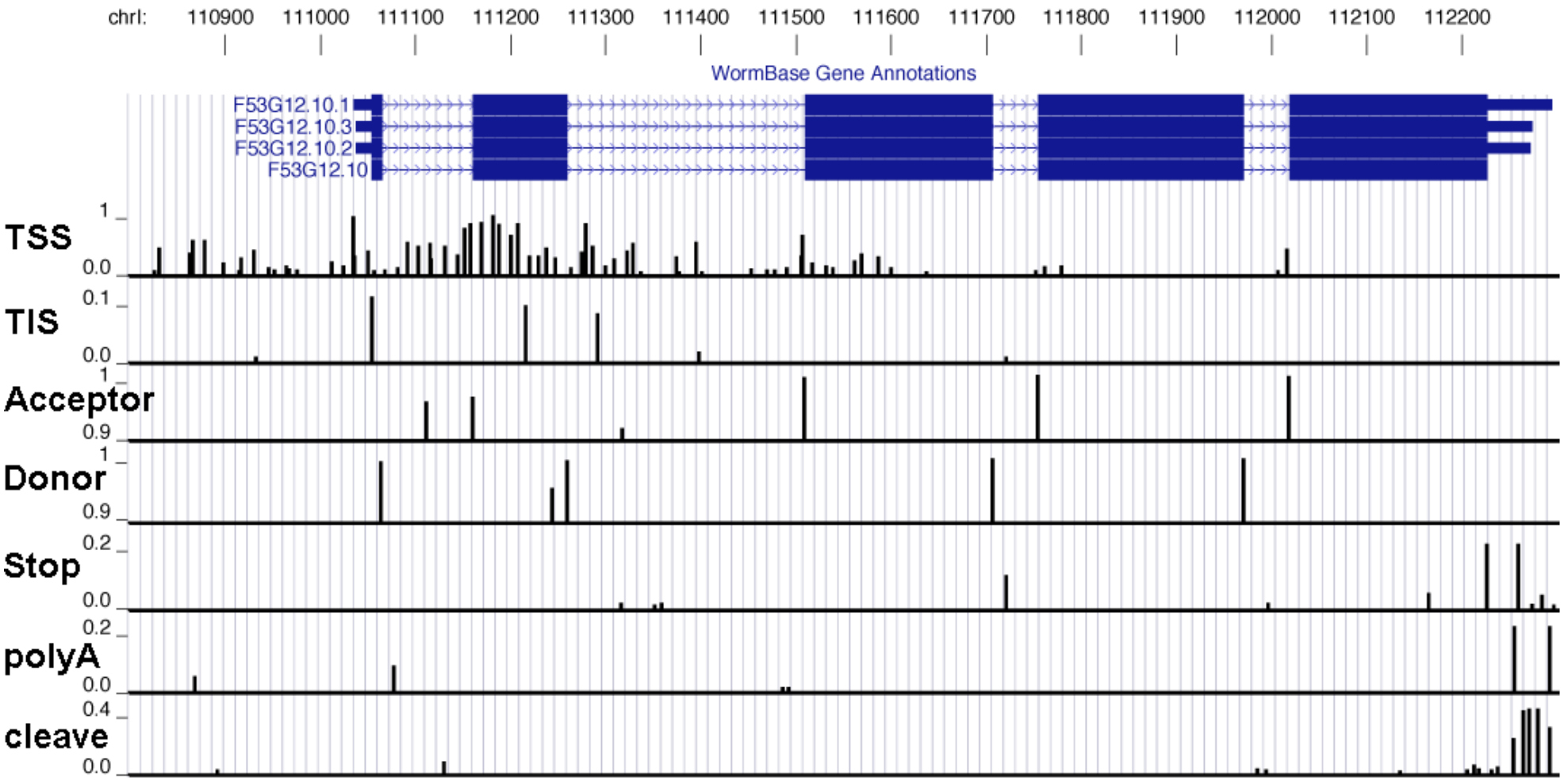




Performance (=Area under PRC curve) keeps increasing.

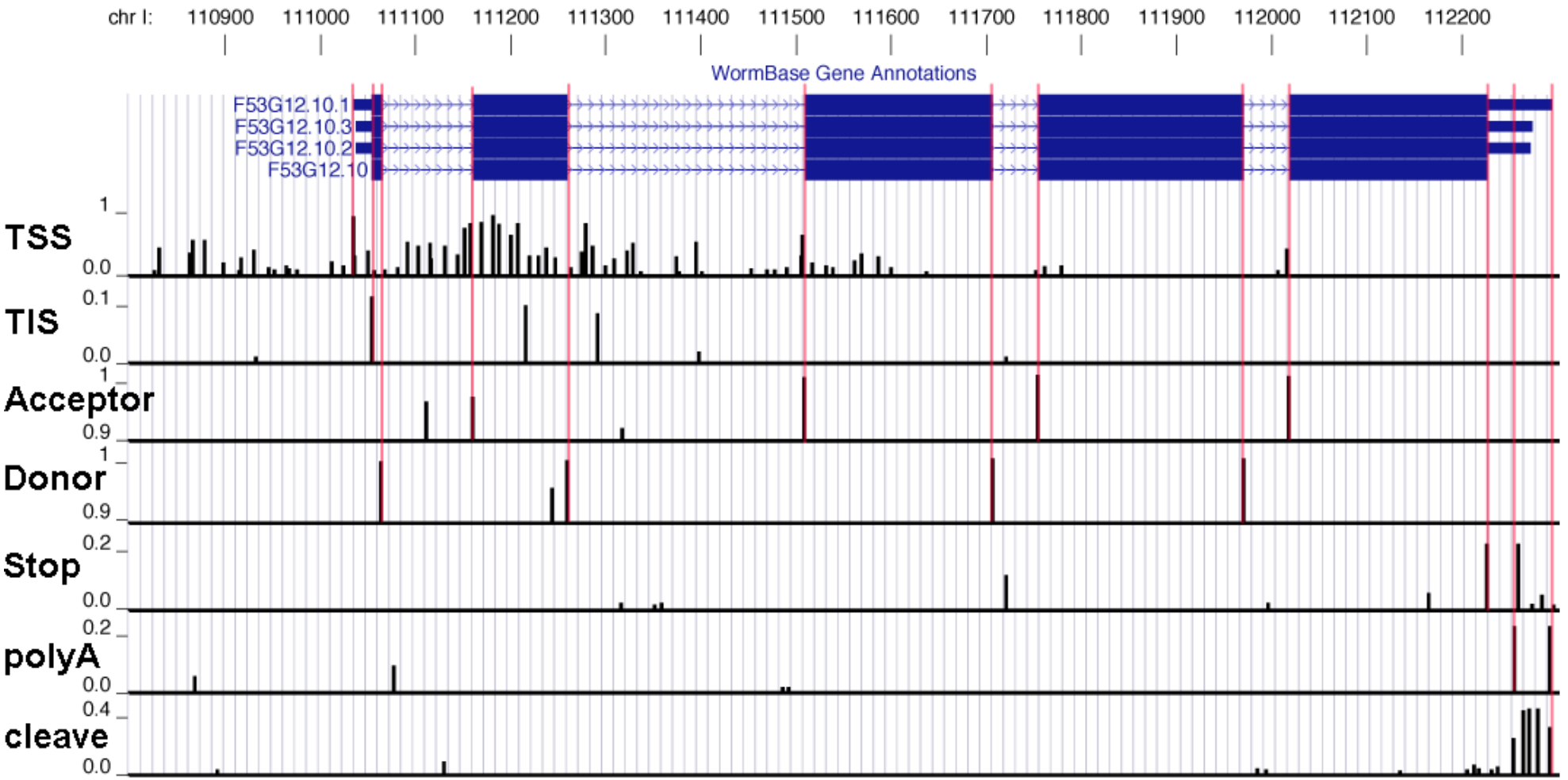
$\Rightarrow$ **Use all available data!**

Example: Predictions in UCSC Browser



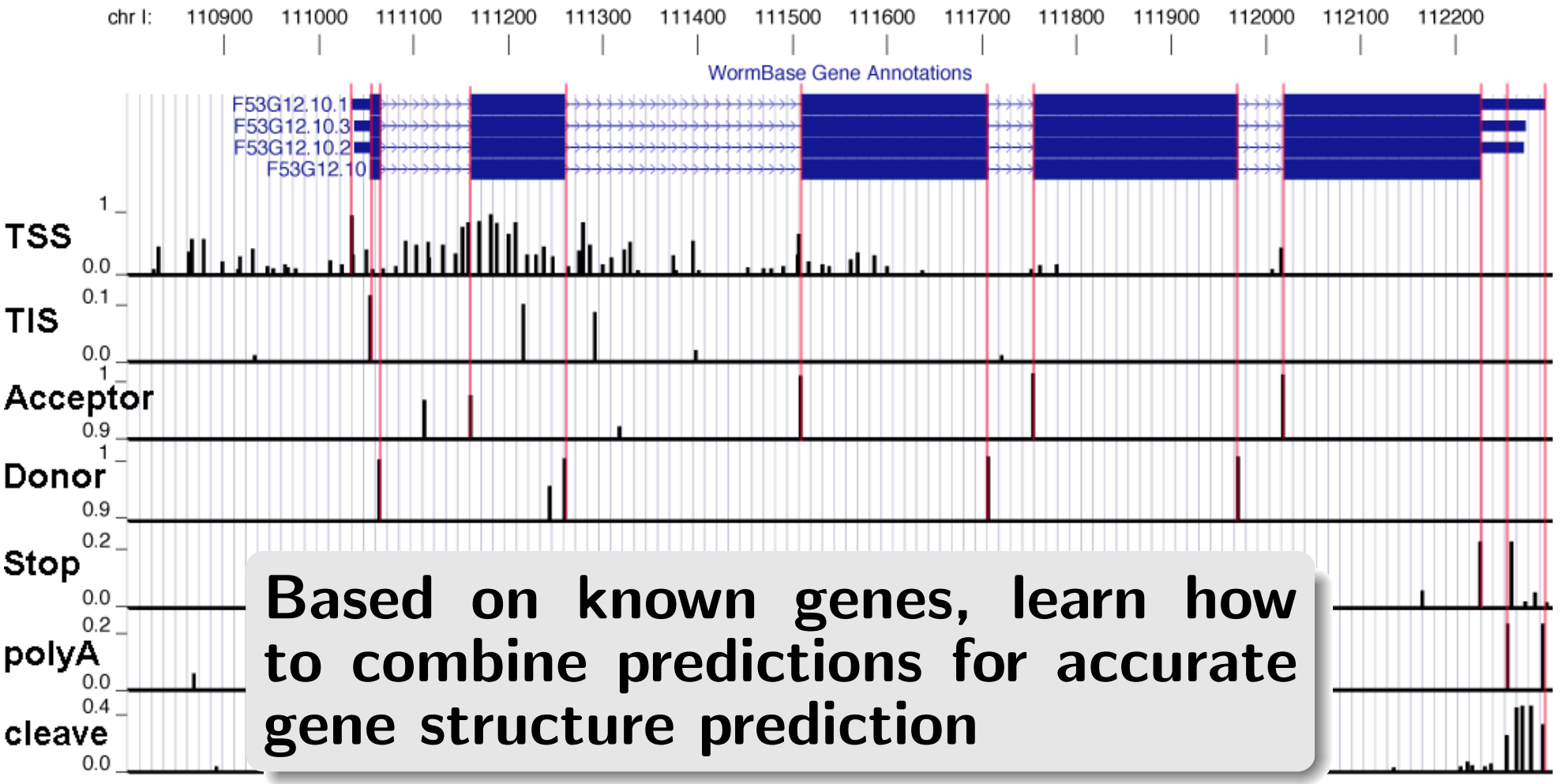
[Schweikert et al., 2009]

Example: Predictions in UCSC Browser



[Schweikert et al., 2009]

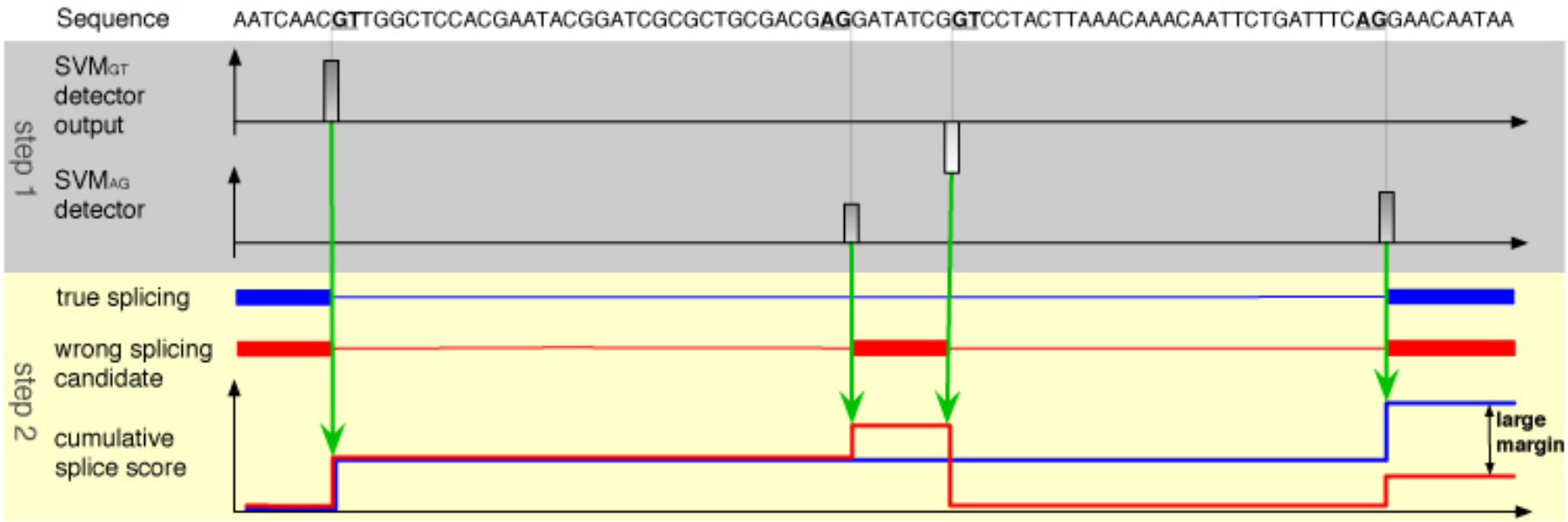
Example: Predictions in UCSC Browser



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

[Schweikert et al., 2009]

Discriminative Gene Prediction (simplified)



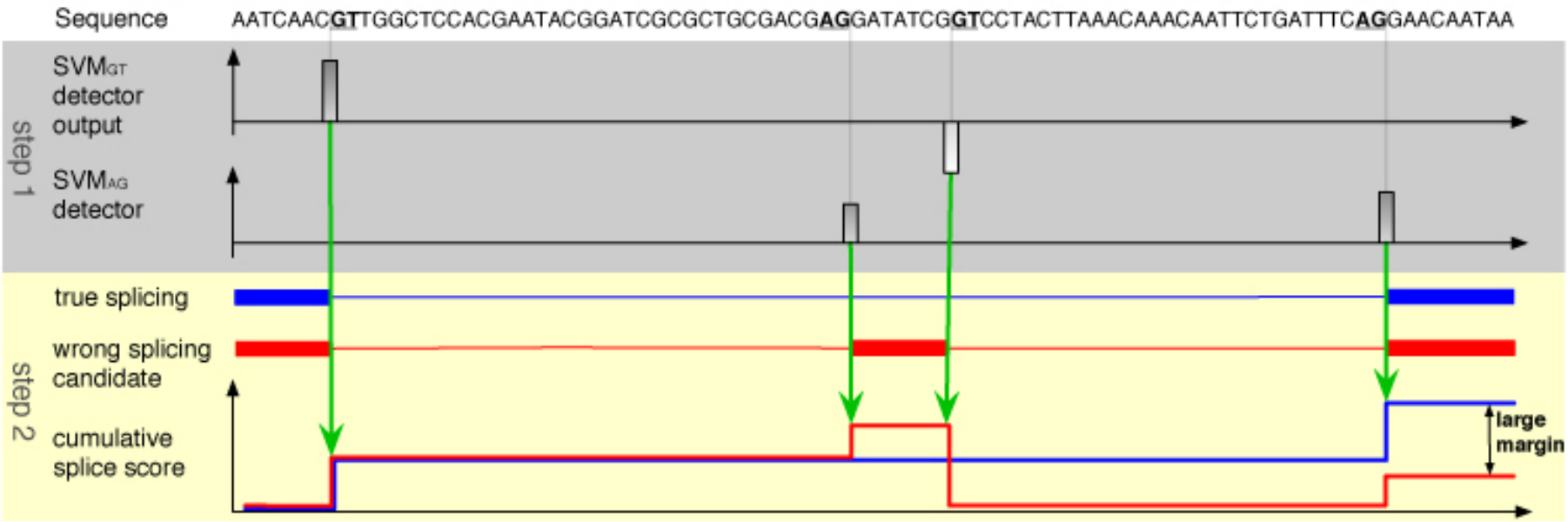
[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

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 (in functions $S_{GT}, S_{AG}, S_{L_E}, S_{L_I}$) by solving **linear program** using training set with known splice forms

Discriminative Gene Prediction (simplified)



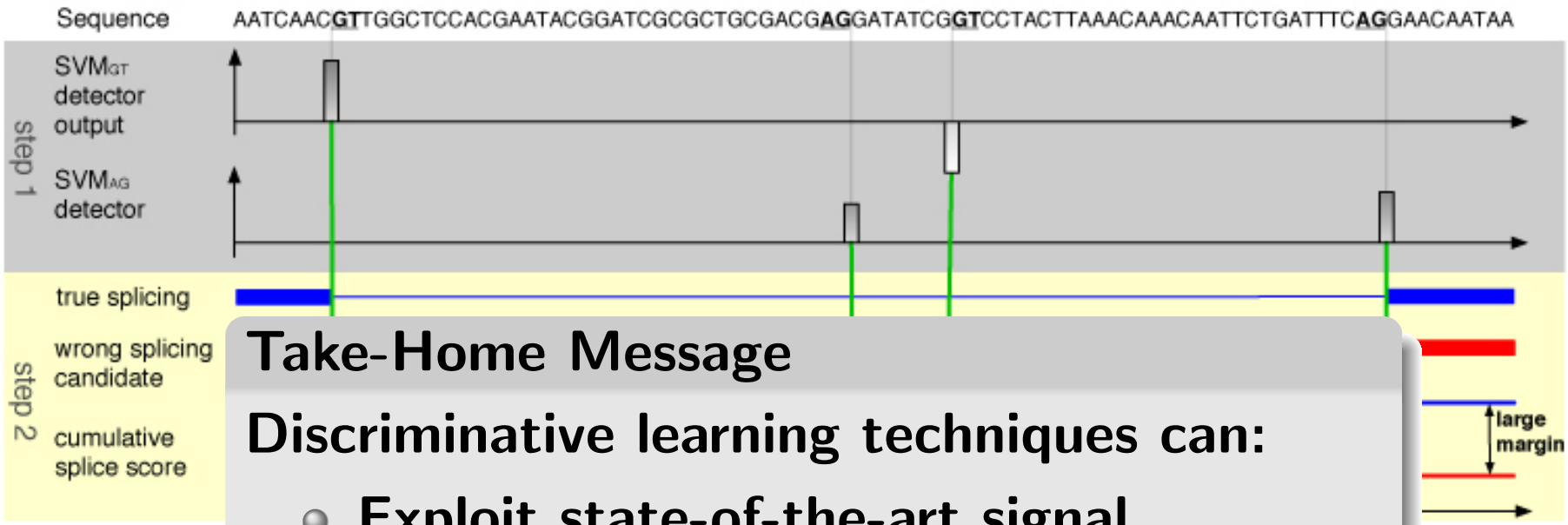
[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

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 (in functions $S_{GT}, S_{AG}, S_{L_E}, S_{L_I}$) by solving **linear program** using training set with known splice forms

Discriminative Gene Prediction (simplified)



Schölkopf, 2007]

Take-Home Message

Discriminative learning techniques can:

- **Exploit state-of-the-art signal predictions**
- **Benefit from auxiliary data**
- **Predict accurately**

Simplified

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

Splice signals | Segment lengths

Tune free parameters (in functions $S_{GT}, S_{AG}, S_{L_E}, S_{L_I}$) by solving **linear program** using training set with known splice forms

Results using mGene

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

	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., 2009]

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%

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge (*C. elegans*) [Coghlan et al., 2008]

- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009]

	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., 2009]

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%

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge (*C. elegans*) [Coghlan et al., 2008]

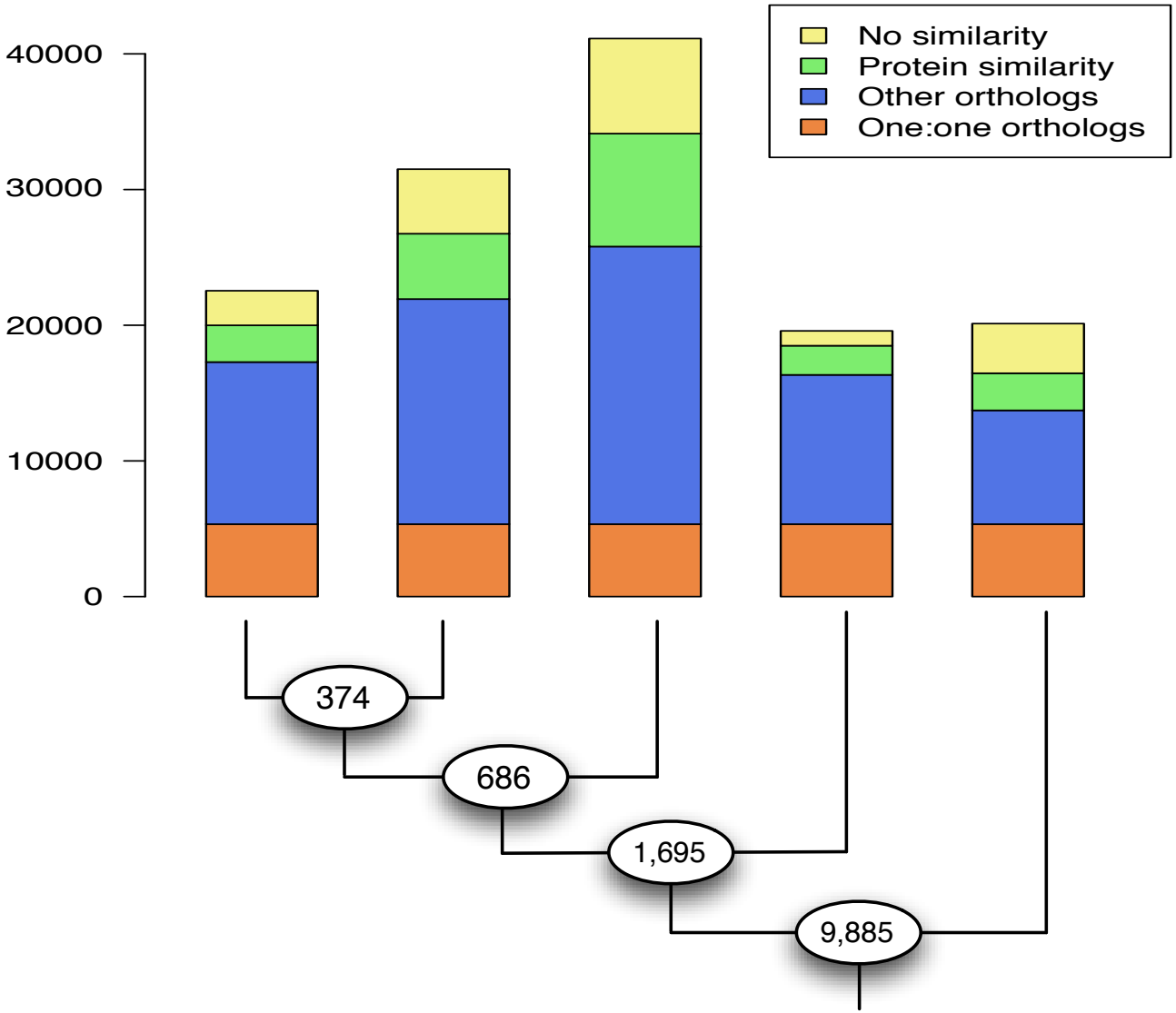
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009]

	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., 2009]

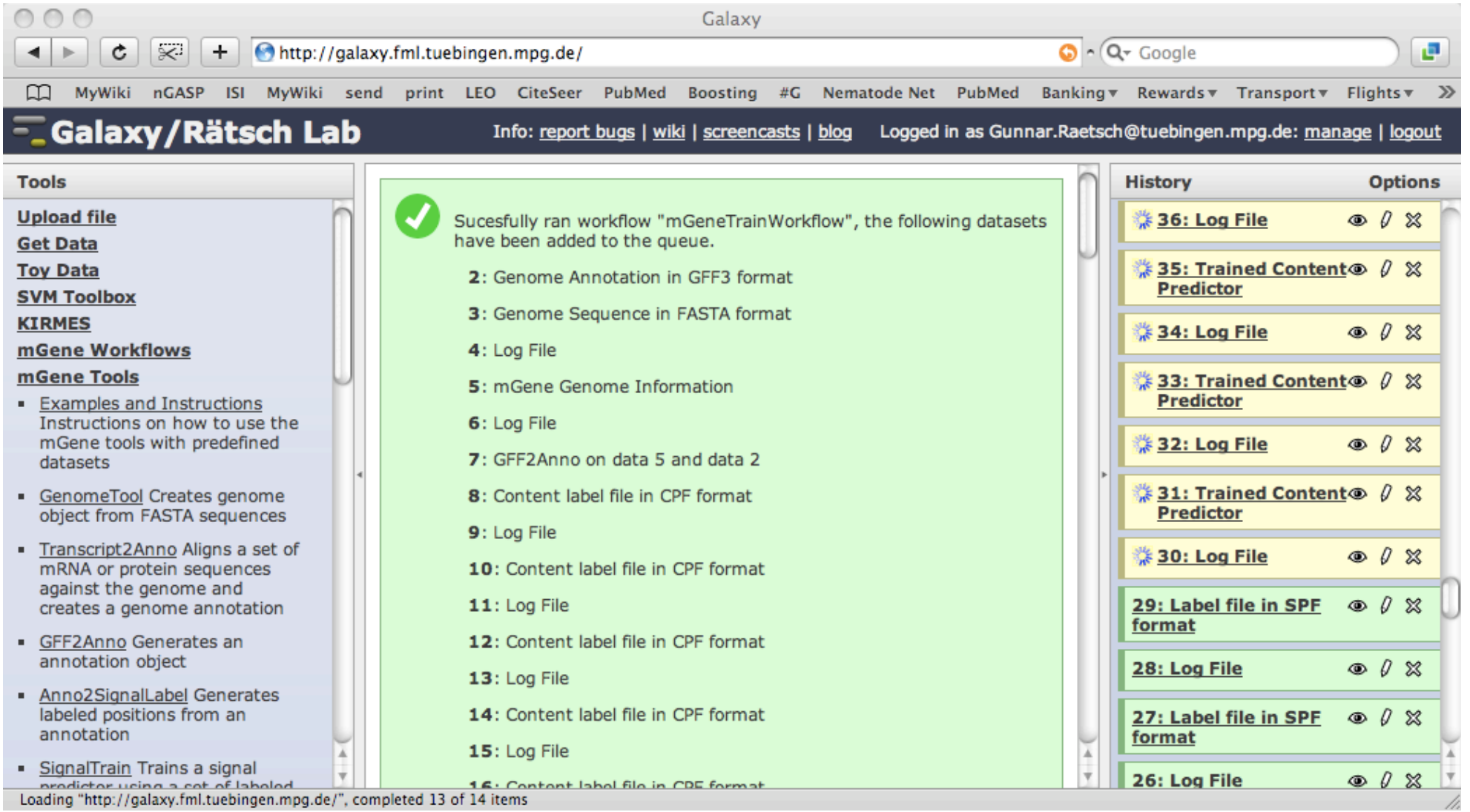
Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

Transcriptome/Proteome Comparison
(Schweikert et al., 2009)



mGene.web: Gene Finding for Everybody ;-)

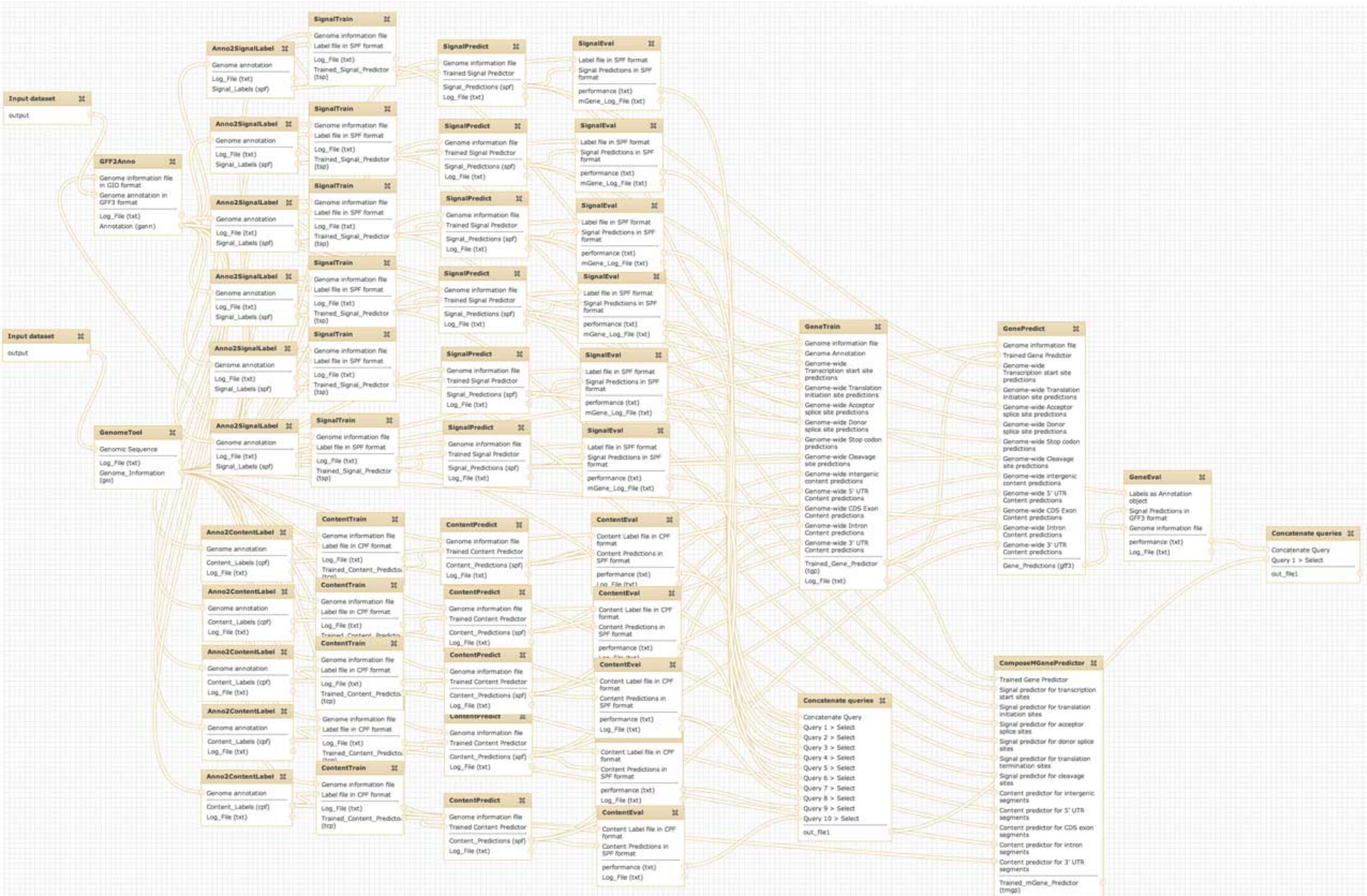
(Schweikert et al., 2009, submitted)



http://mgene.org/webservice

mGene.web: Gene Finding for Everybody ;-)

(Schweikert et al., 2009, submitted)



Limitations/Extensions

- Gene finding accuracy still far from perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when transcripts are expressed/modified/degraded...
- Accurate enough to accurately predict the effects of SNPs?
  - Annotate all the newly sequenced variations of genomes
  - 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., 2008]

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

Limitations/Extensions

- Gene finding accuracy still far from perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when transcripts are expressed/modified/degraded...
- Accurate enough to accurately predict the effects of SNPs?
  - Annotate all the newly sequenced variations of genomes
  - 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., 2008]

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

Limitations/Extensions

- Gene finding accuracy still far from perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when transcripts are expressed/modified/degraded...
- Accurate enough to accurately predict the effects of SNPs?
  - Annotate all the newly sequenced variations of genomes
  - 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., 2008]

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

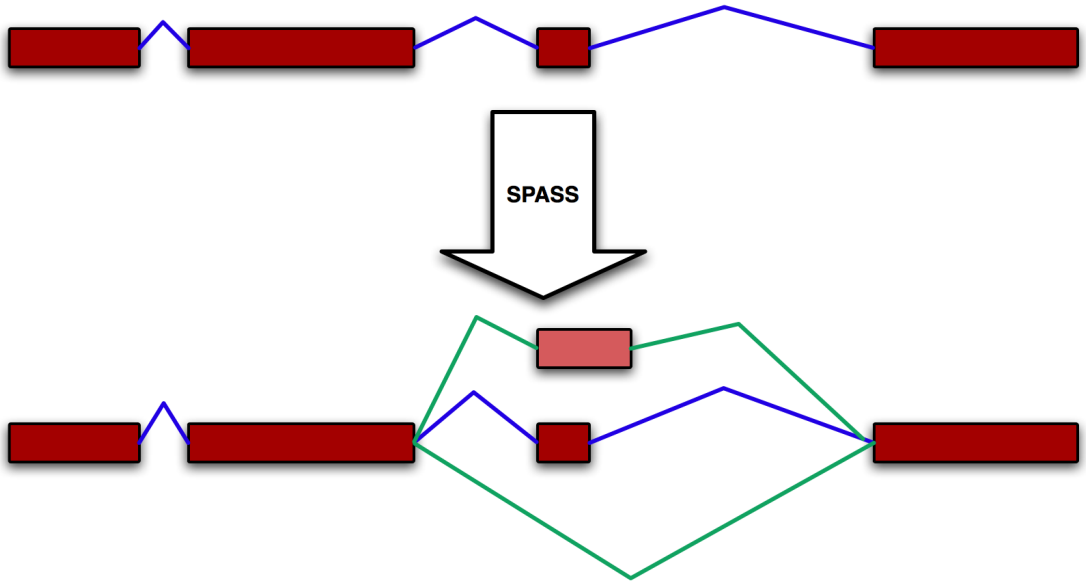
Limitations/Extensions

- Gene finding accuracy still far from perfect
  - Misses genes, predicts incorrect gene models
  - Does not (yet) predict alternative transcripts
  - Cannot predict when transcripts are expressed/modified/degraded...
- Accurate enough to accurately predict the effects of SNPs?
  - Annotate all the newly sequenced variations of genomes
  - 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., 2008]

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

Alternative Transcripts: First Steps

Predictions of alternative splicing



- Predict novel alternative splicing as independent events
- Only use sequence information (e.g. Rätsch et. al, ISMB'05)
- Requires known gene structures: Use *ab initio* predictions

Alternative Transcripts: More Steps

- Combine gene finding with prediction of single alternative splicing events
  - Predict the **splice graph** of a gene
  - Machine learning challenge:
    - Input: DNA sequence
    - Output: Splice graph
- mGene approach can be extended to
  - Include predictions of alternative splicing events
  - Predict **"simple splice graphs"**
    $\Rightarrow$ We tried ... It is difficult!
- Predicting arbitrary graphs is considerably harder
  - With more data this may become feasible

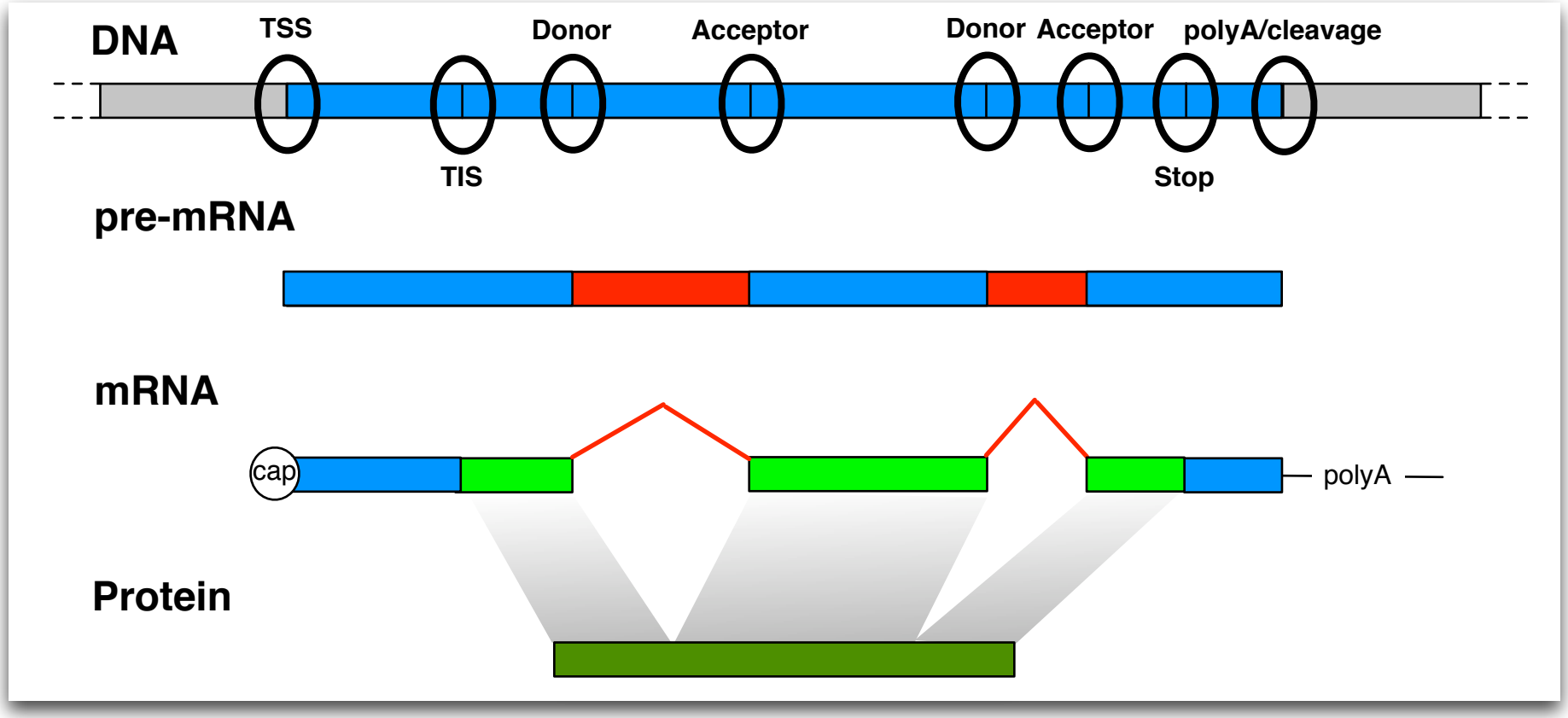
Alternative Transcripts: More Steps

- Combine gene finding with prediction of single alternative splicing events
  - Predict the **splice graph** of a gene
  - Machine learning challenge:
    - Input: DNA sequence
    - Output: Splice graph
- mGene approach can be extended to
  - Include predictions of alternative splicing events
  - Predict **"simple splice graphs"**
    $\Rightarrow$ We tried ... It is difficult!
- Predicting arbitrary graphs is considerably harder
  - With more data this may become feasible

Alternative Transcripts: More Steps

- Combine gene finding with prediction of single alternative splicing events
  - Predict the **splice graph** of a gene
  - Machine learning challenge:
    - Input: DNA sequence
    - Output: Splice graph
- mGene approach can be extended to
  - Include predictions of alternative splicing events
  - Predict **"simple splice graphs"**
    $\Rightarrow$ We tried ... It is difficult!
- Predicting arbitrary graphs is considerably harder
  - With more data this may become feasible

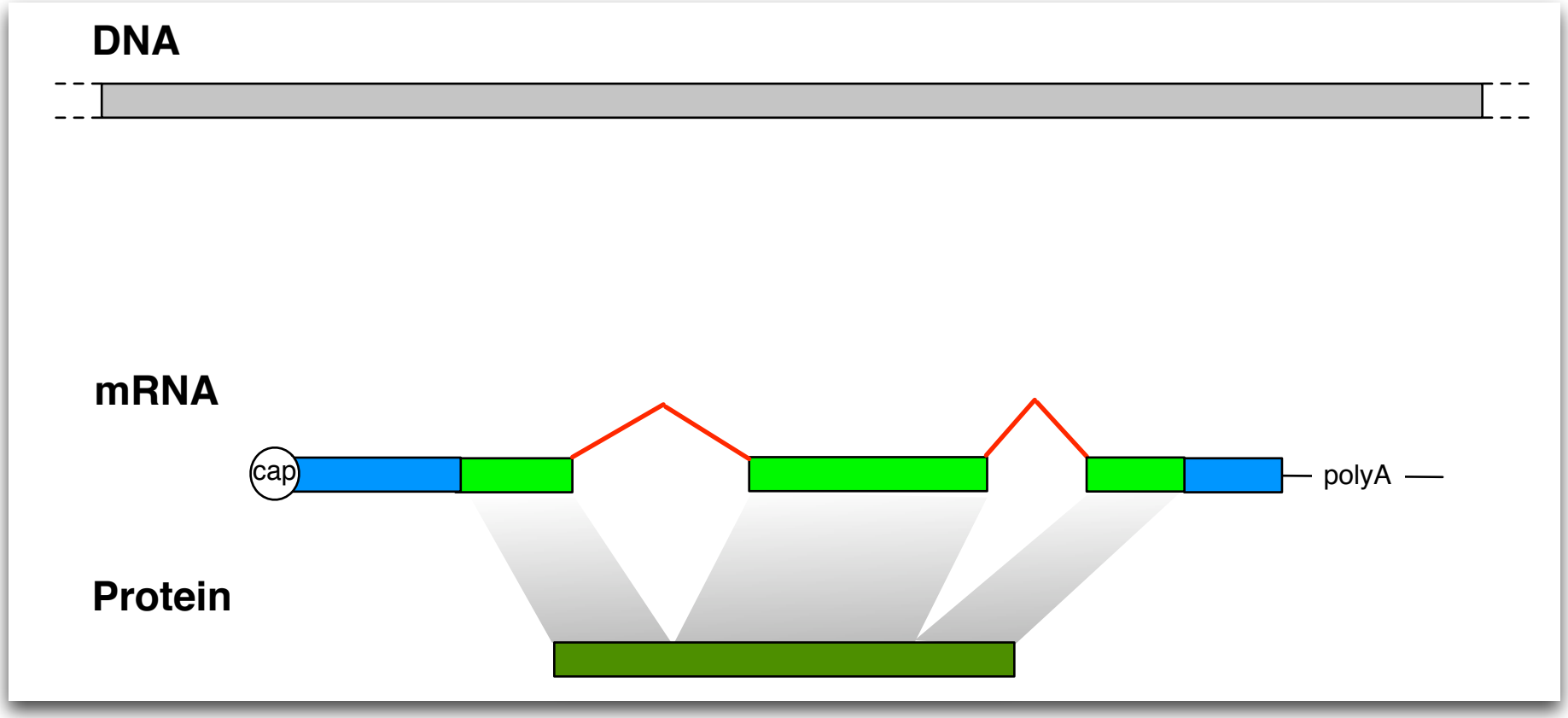
From Genome to Proteins



Directly measure the transcriptome

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

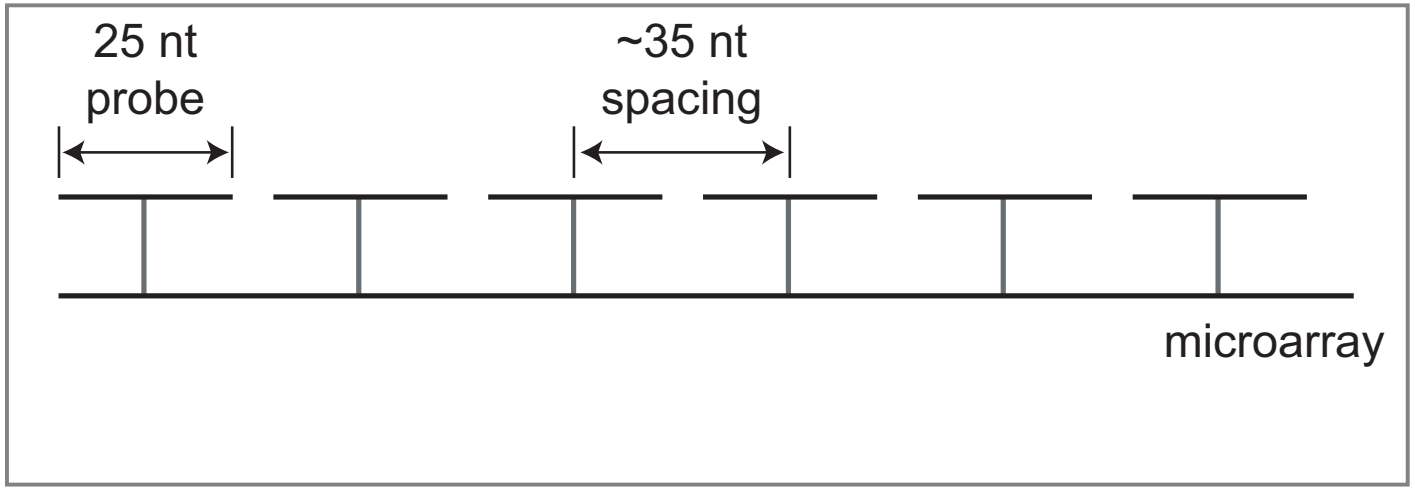
From Transcriptome to Proteins



Directly measure the transcriptome

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

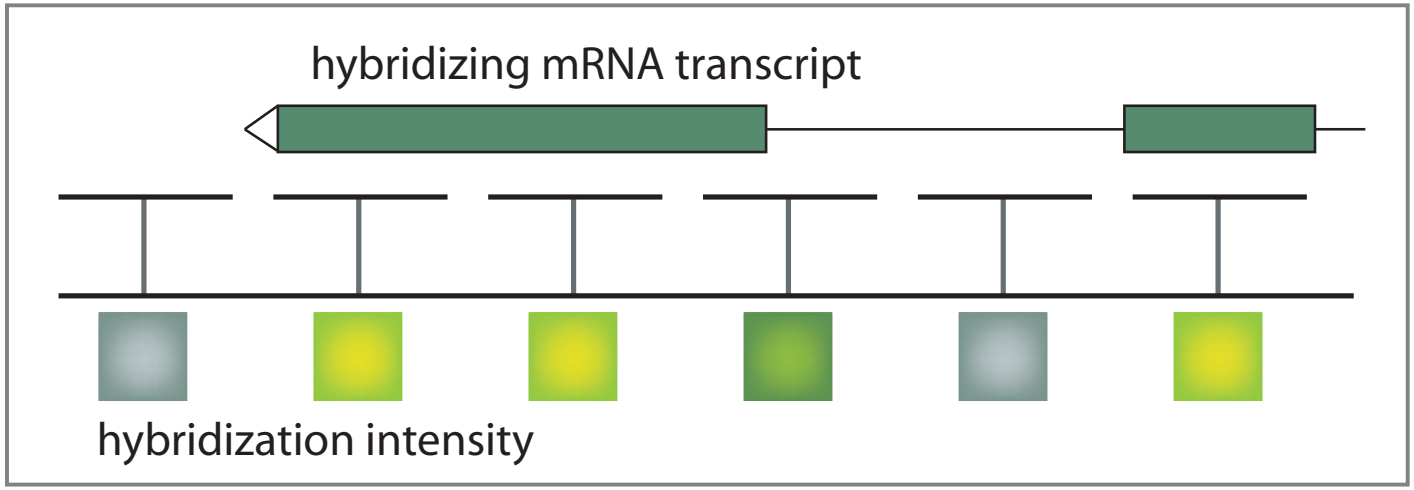
Tiling Arrays for Transcriptome Analysis



- 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)

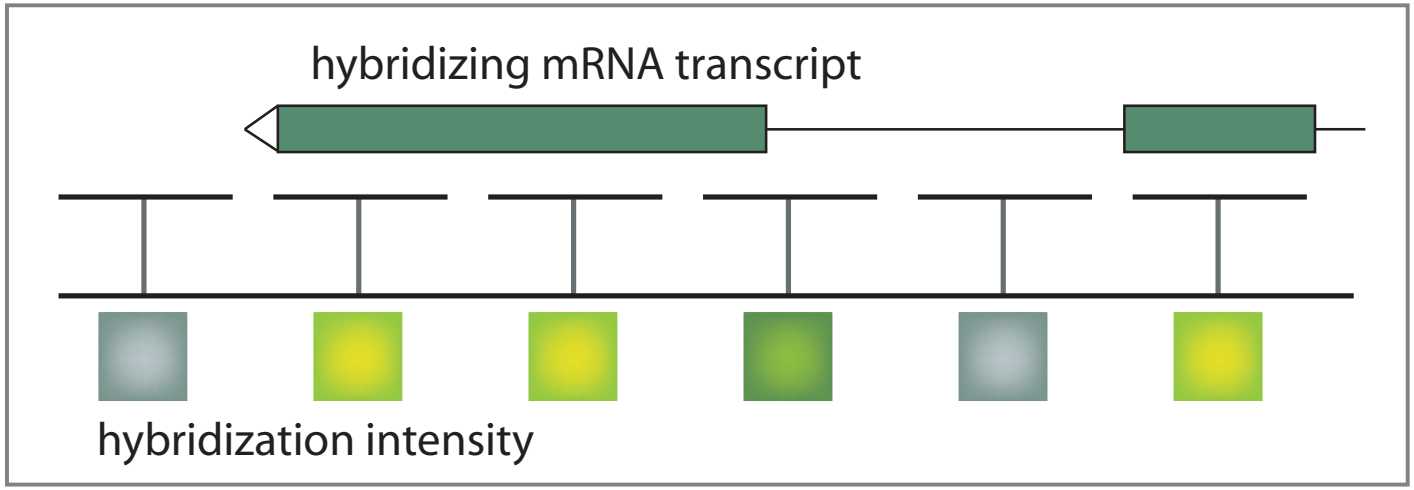
Tiling Arrays for Transcriptome Analysis



- 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 classical microarrays

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

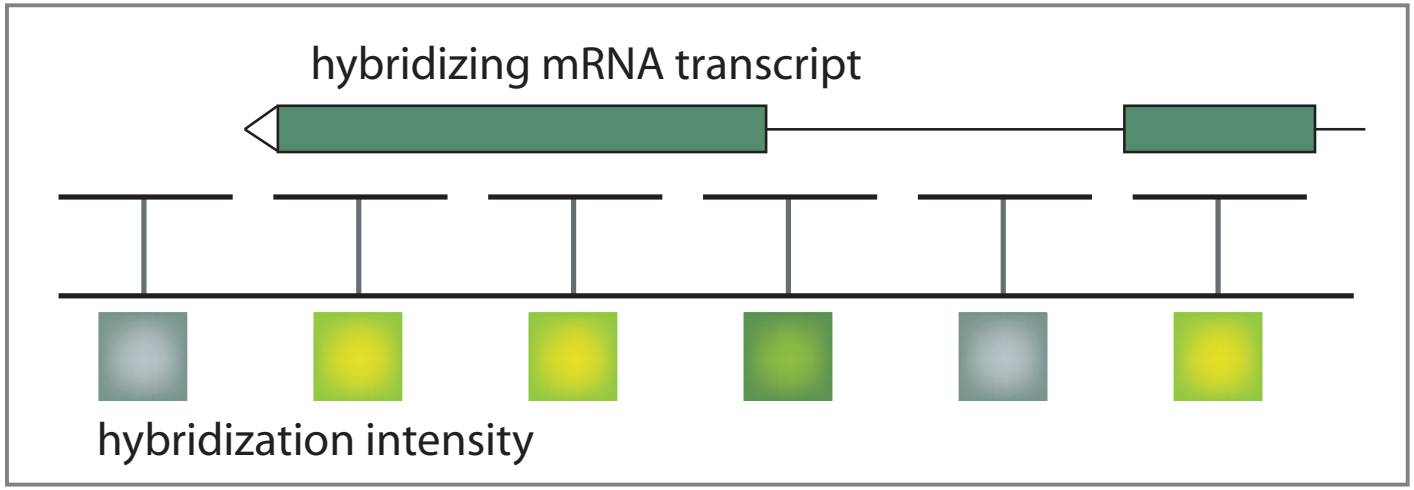
Tiling Arrays for Transcriptome Analysis



- 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]

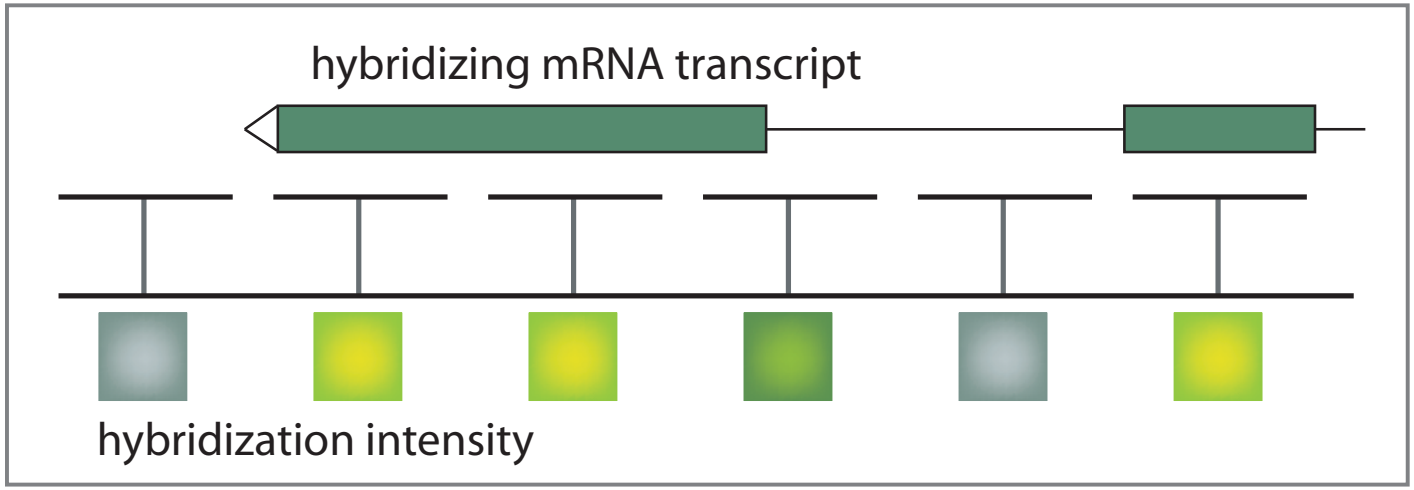
Tiling Arrays for Transcriptome Analysis



- 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]

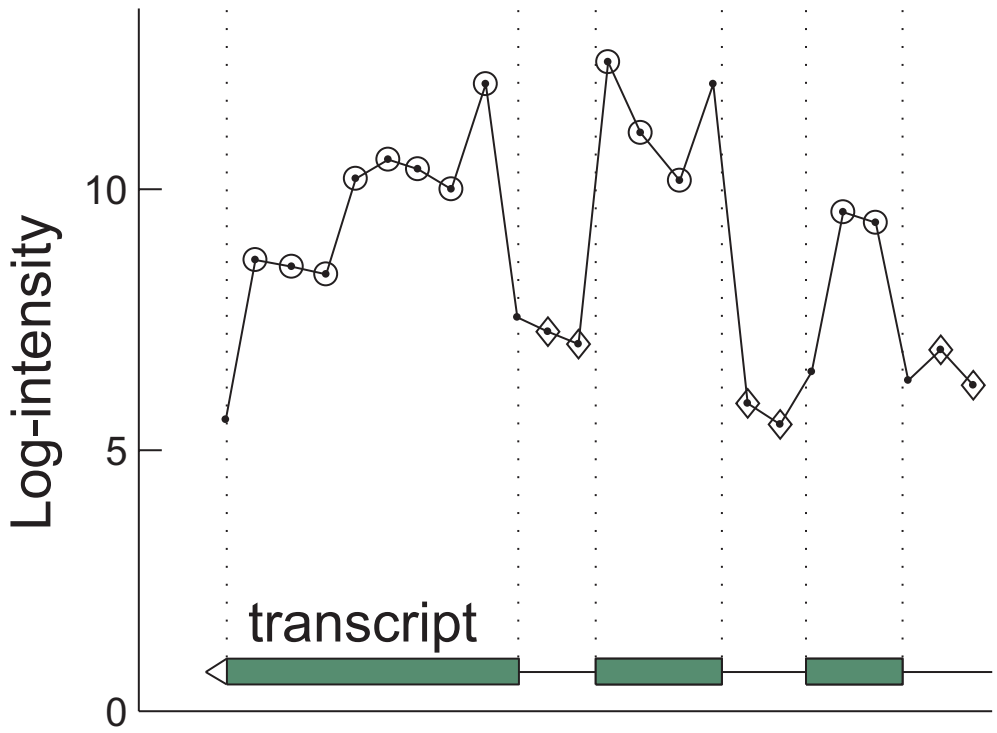
Tiling Arrays for Transcriptome Analysis



- 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



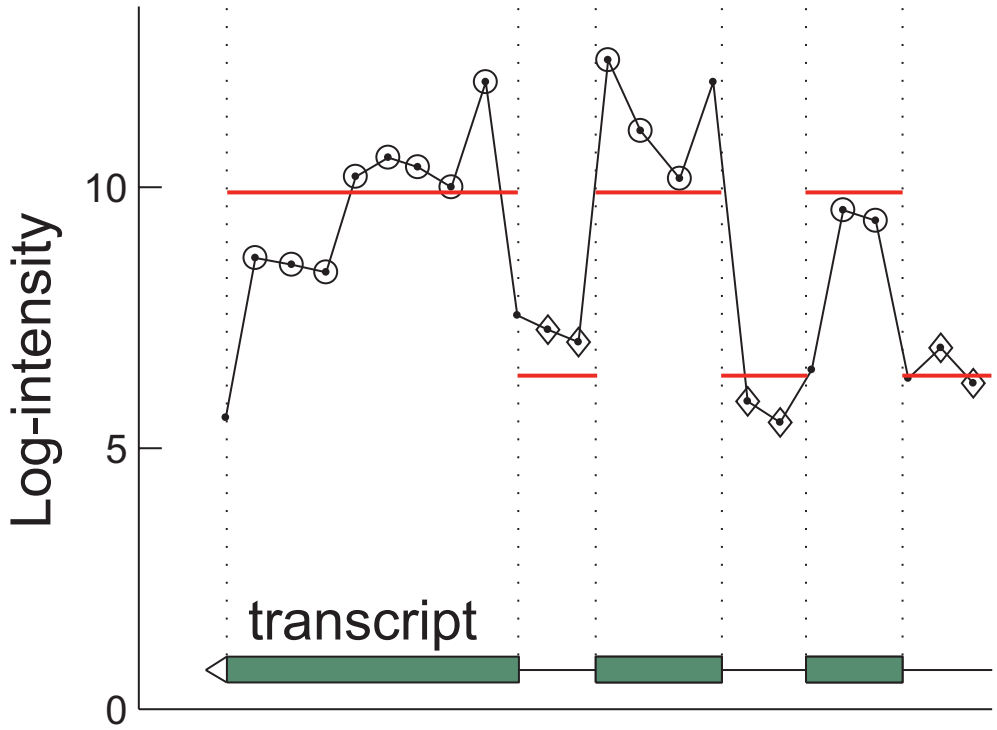
- observed intensity
- annotated exonic
- annotated intronic

Systematic bias induced by probe sequence effects

$\Rightarrow$ Model effect for normalization

[Zeller et al., 2008b]

Intensities are Noisy Measurements

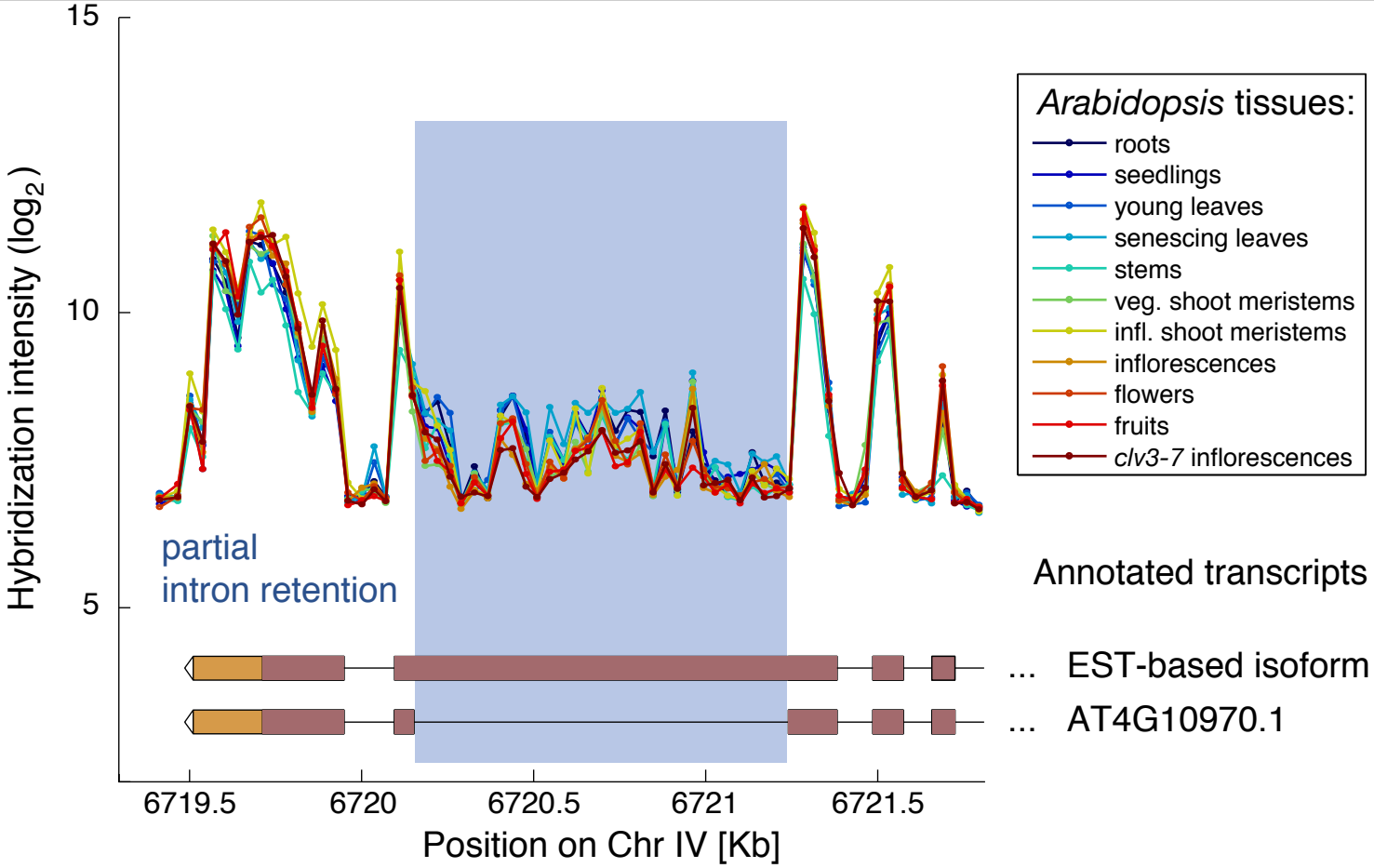


- observed intensity
- annotated exonic
- annotated intronic
- "ideal noise-free intensity"

Systematic bias induced by probe sequence effects

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

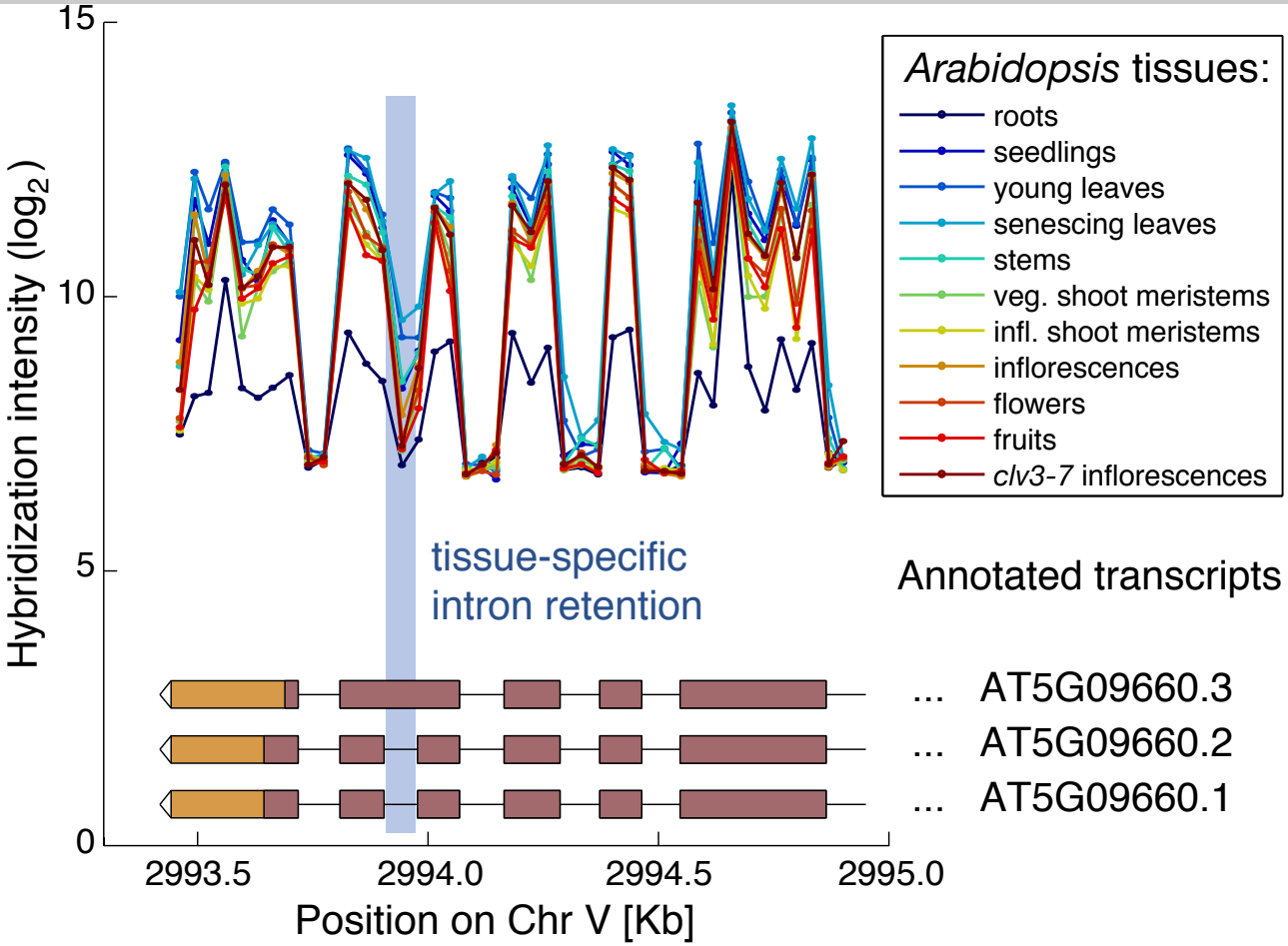
Alternatively Spliced Transcripts



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

[Eichner et al., 2009]

Differentially Spliced Transcripts



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

[Eichner et al., 2009]

Genome-wide Analyses of Alternative Splicing

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

- Identified few thousand intron retentions (est. FDR=14%)
- Many of them show differential splicing [Eichner et al., 2009]

- 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)
- First exon/intron-resolution analysis in plants [Laubinger et al., 2008b]

- Identified few thousand intron retentions (est. FDR=14%)
- Many of them show differential splicing [Eichner et al., 2009]

- 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)
- First exon/intron-resolution analysis in plants [Laubinger et al., 2008b]

- Identified few thousand intron retentions (est. FDR=14%)
- Many of them show differential splicing [Eichner et al., 2009]

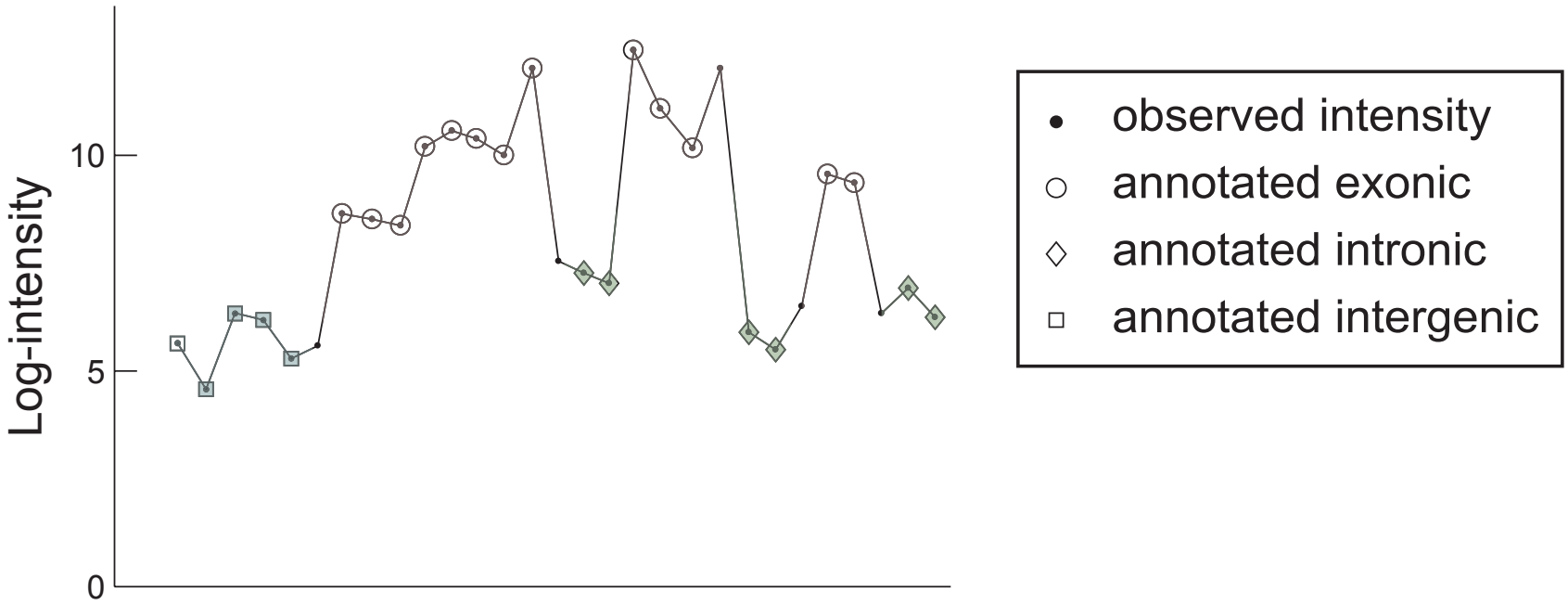
- Profiling cap-binding/miRNA processing mutants

$\Rightarrow$ 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



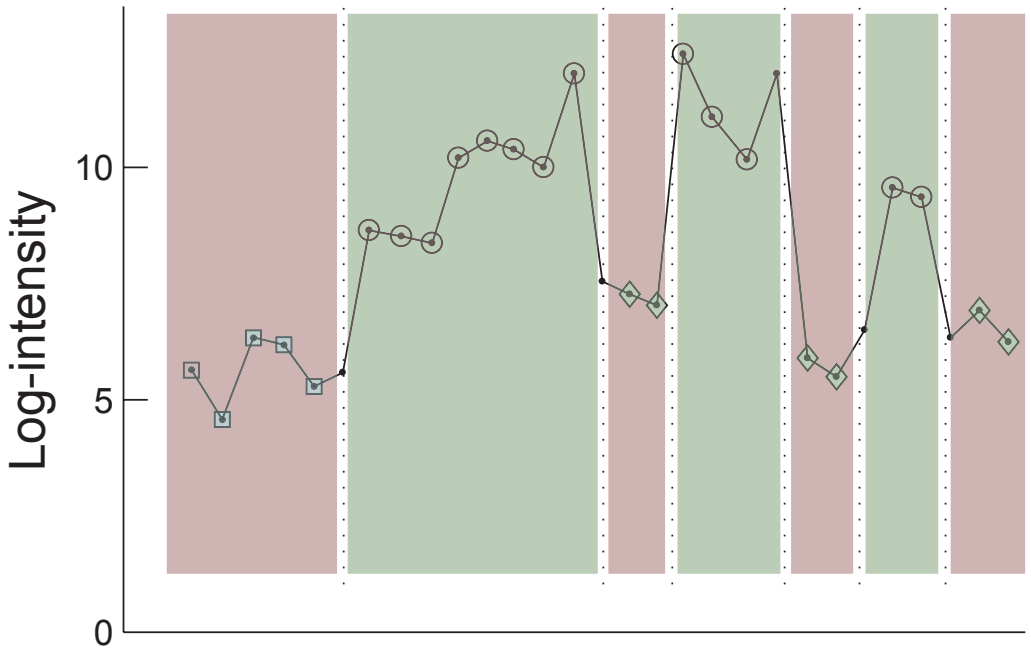
Novel segmentation method ("mSTAD")

- 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

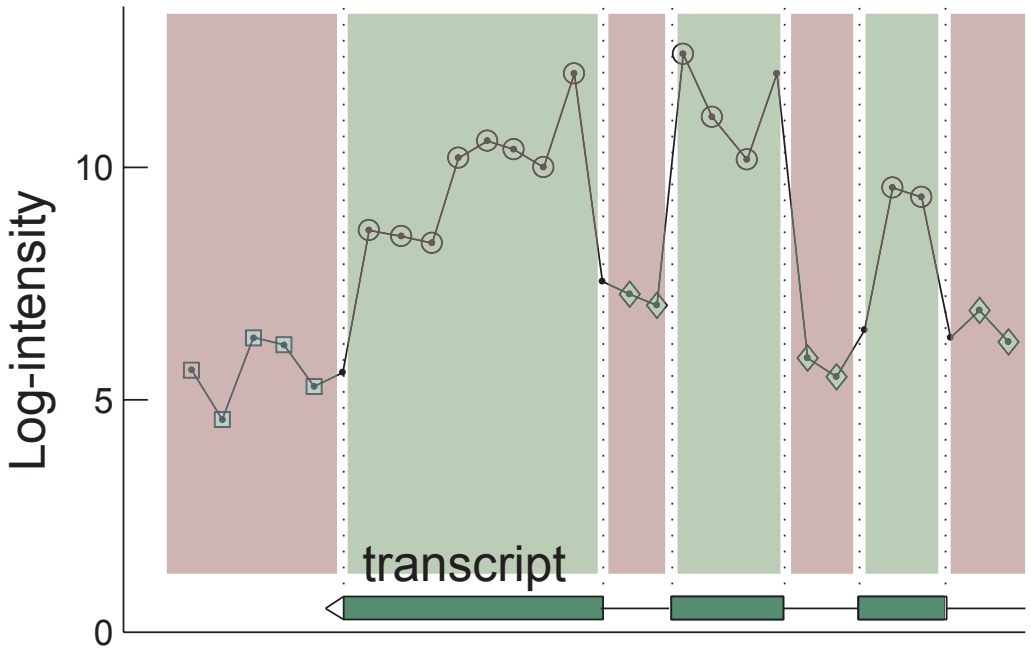
Novel segmentation method ("mSTAD")

- 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

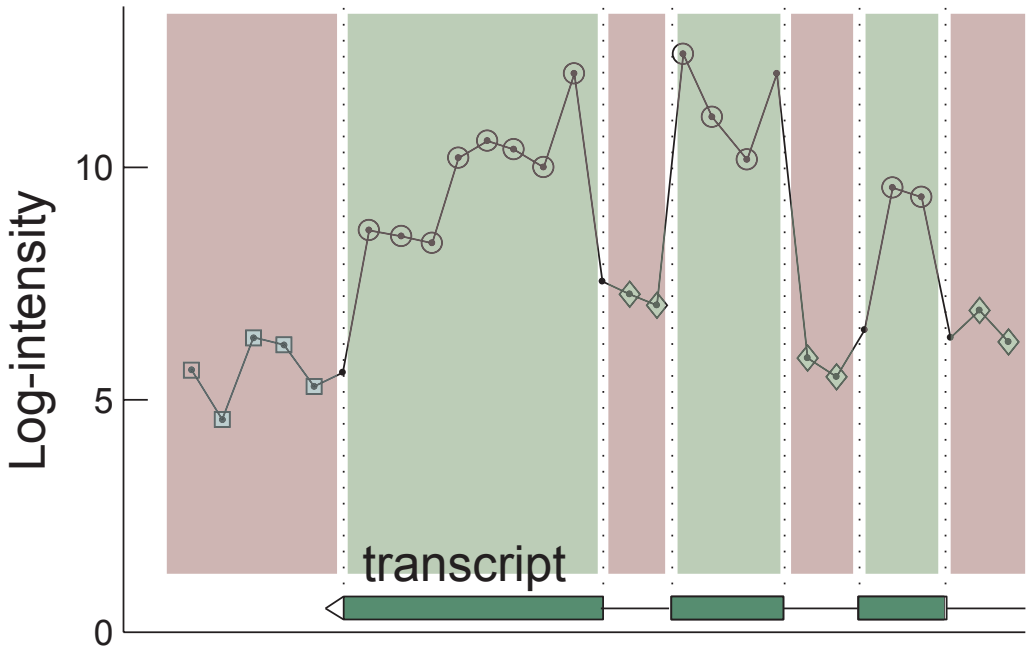
Novel segmentation method ("mSTAD")

- 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



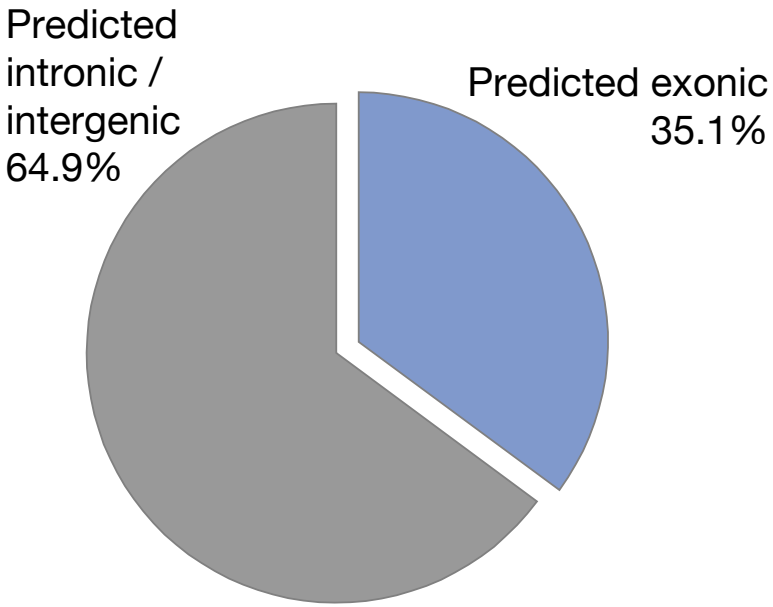
- observed intensity
- annotated exonic
- annotated intronic
- annotated intergenic

Novel segmentation method ("mSTAD")

- accounts for spliced transcripts
- provides very accurate predictions

[Zeller et al., 2008b]

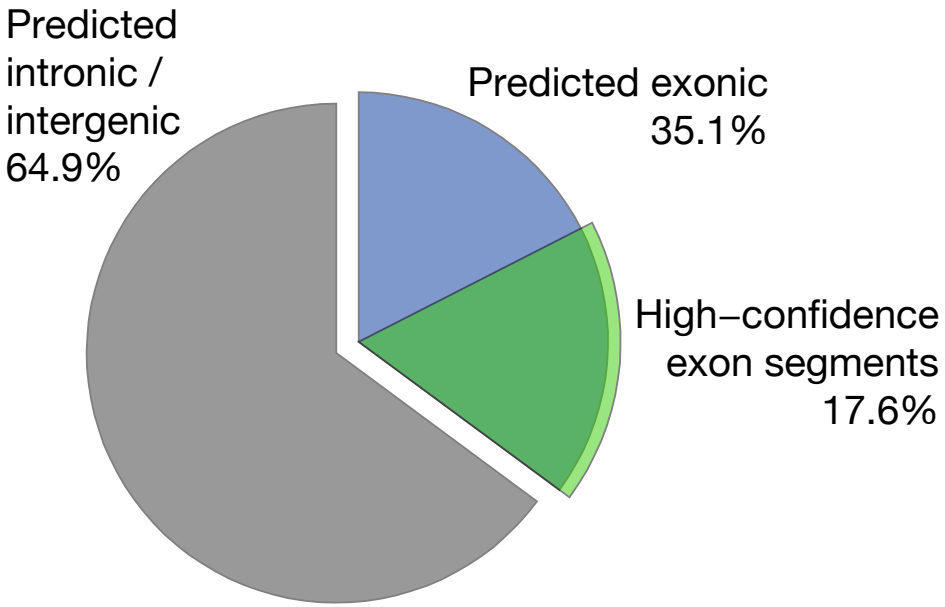
Discovery of New Transcripts



- 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]

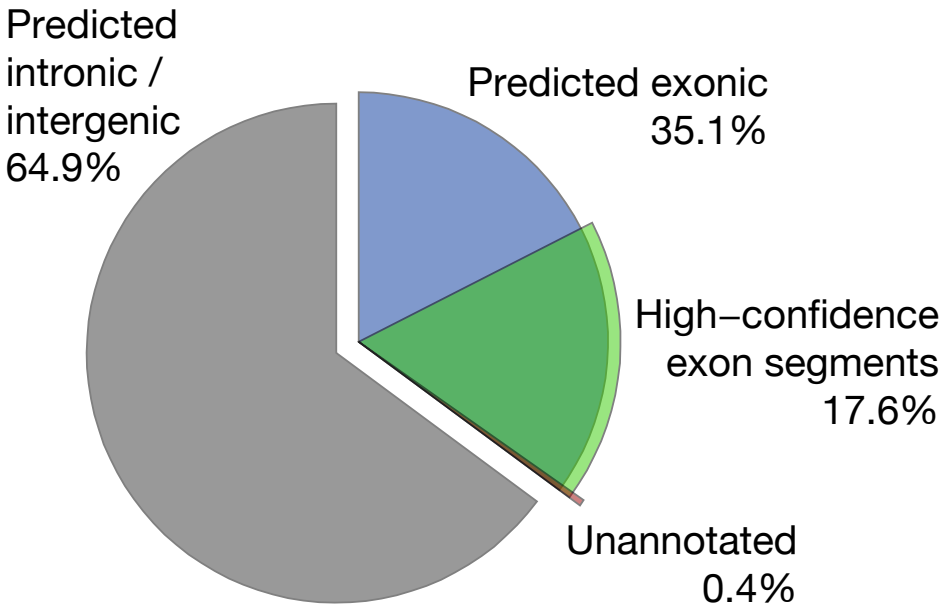
Discovery of New Transcripts



- 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]

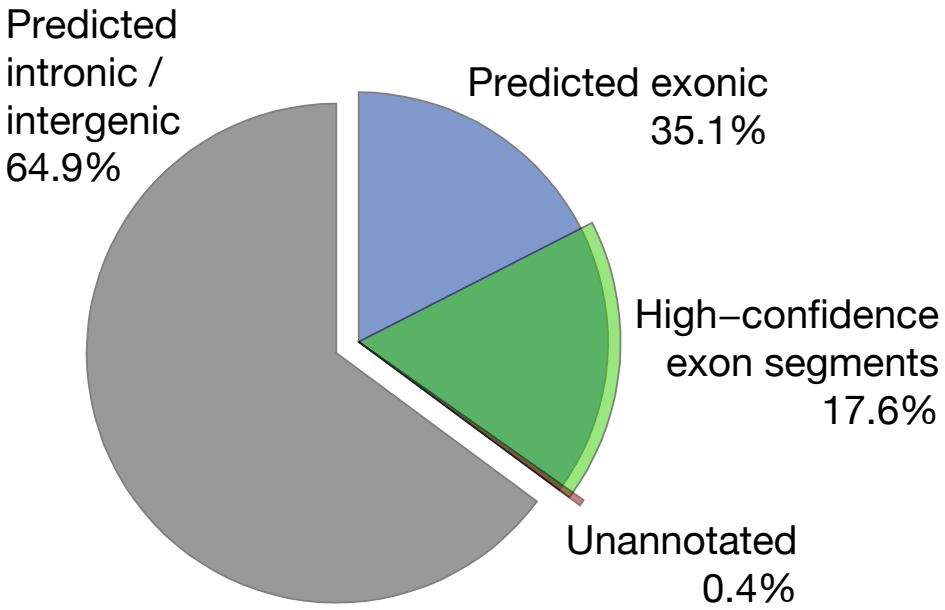
Discovery of New Transcripts



- 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]

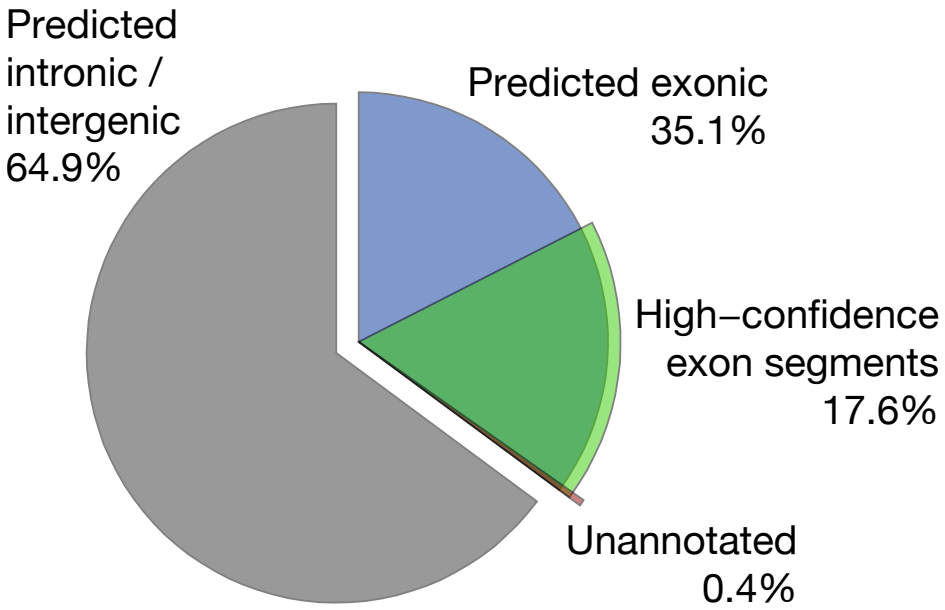
Discovery of New Transcripts



- 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]

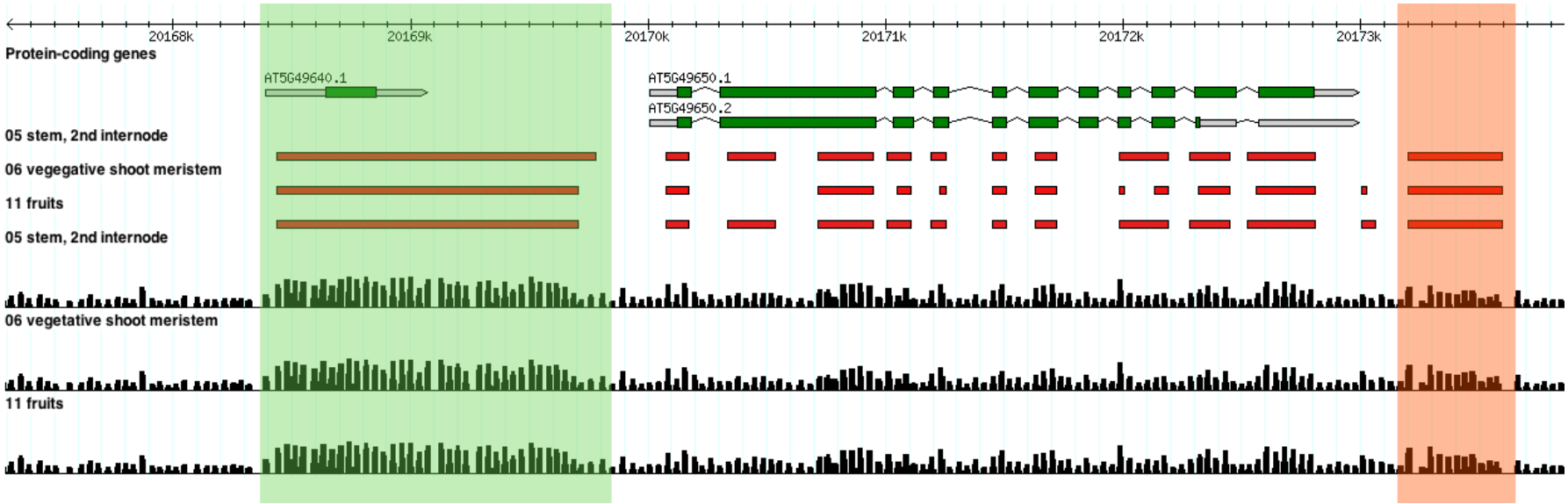
Discovery of New Transcripts



- 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]

Next Generation Sequencing

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

Next Generation Sequencing

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

Next Generation Sequencing

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

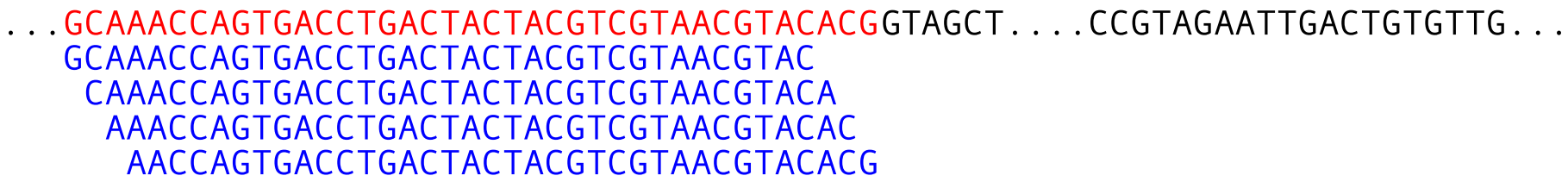
Spliced vs. Unspliced Alignments




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

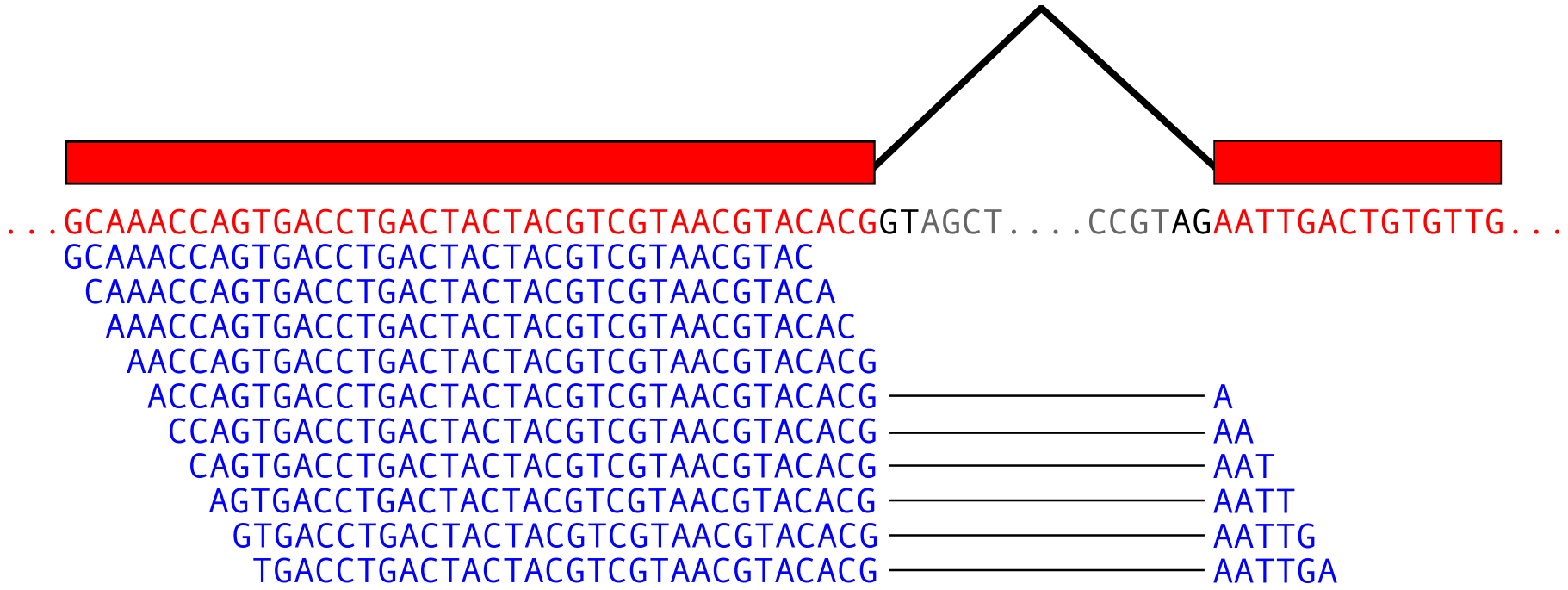
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

Spliced vs. Unspliced Alignments



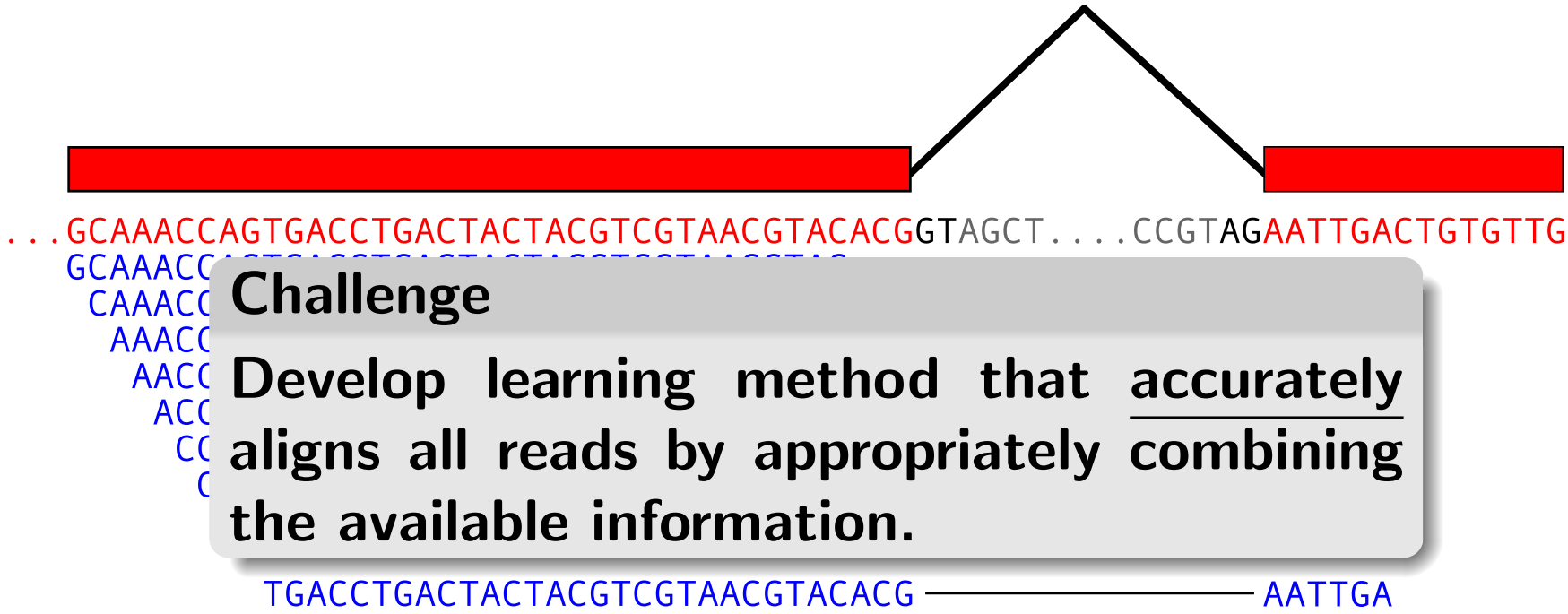
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

Spliced vs. Unspliced Alignments



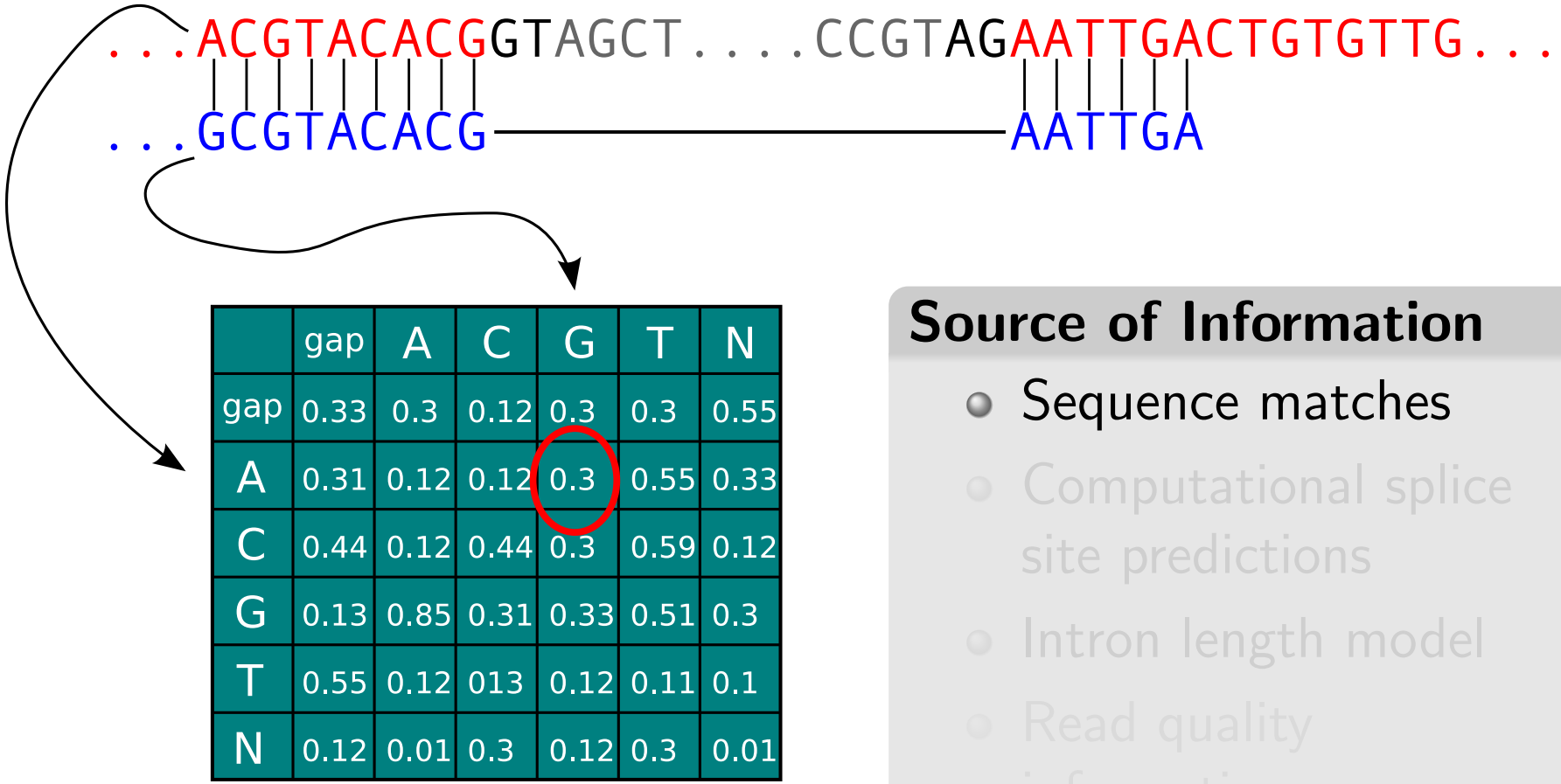
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

Spliced vs. Unspliced Alignments



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

Alignment Scoring Function



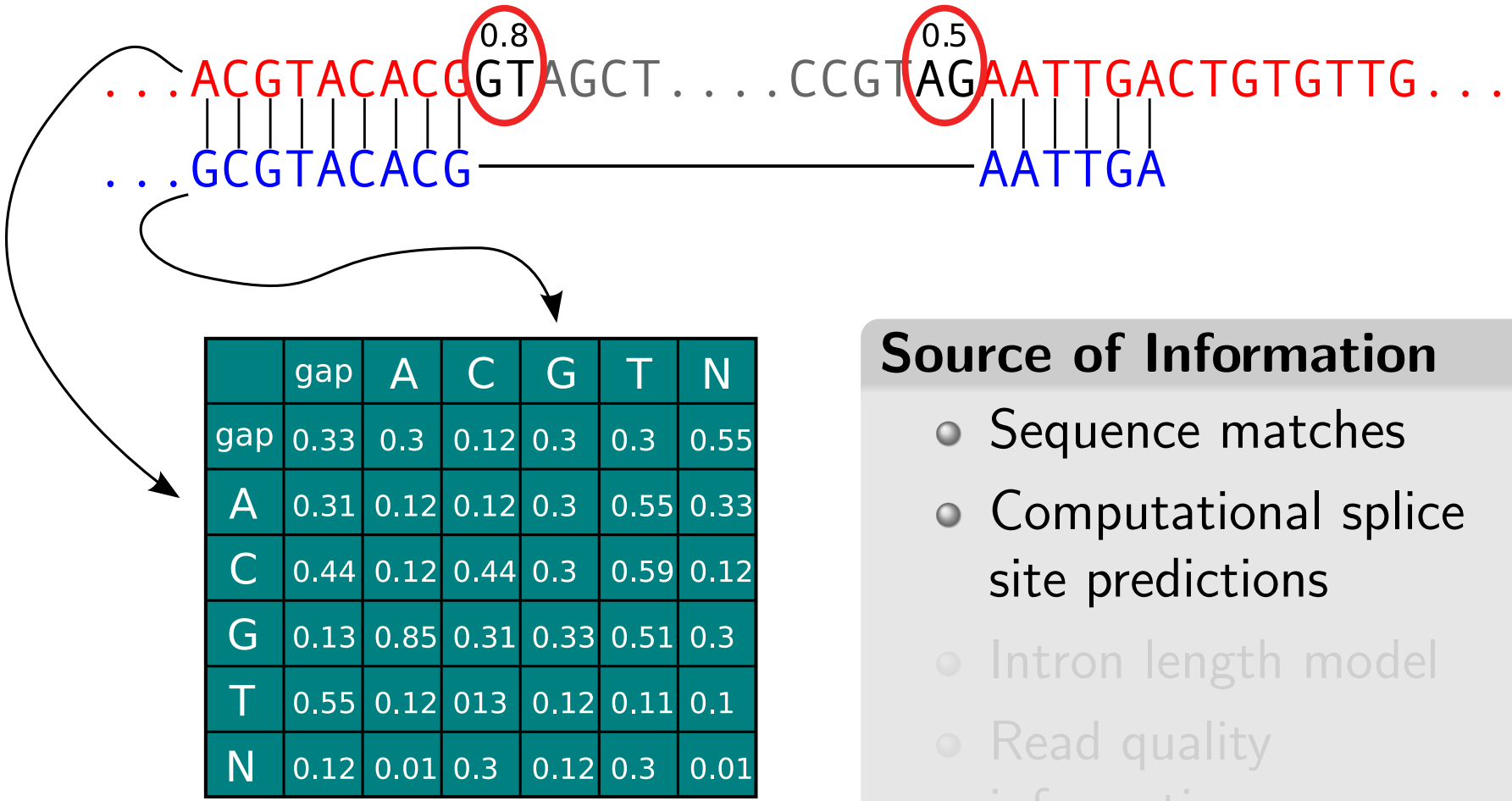
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Alignment Scoring Function



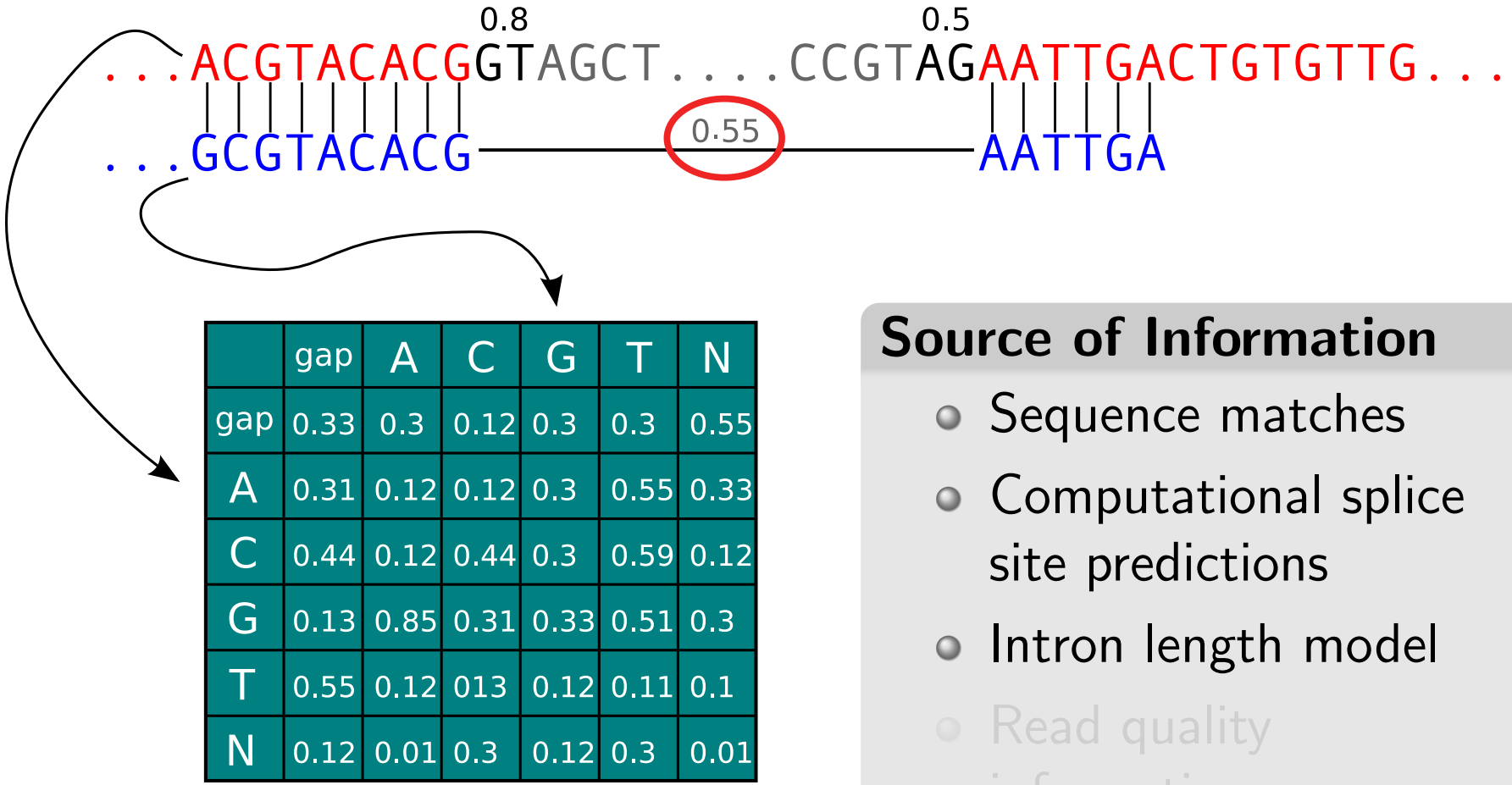
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Alignment Scoring Function



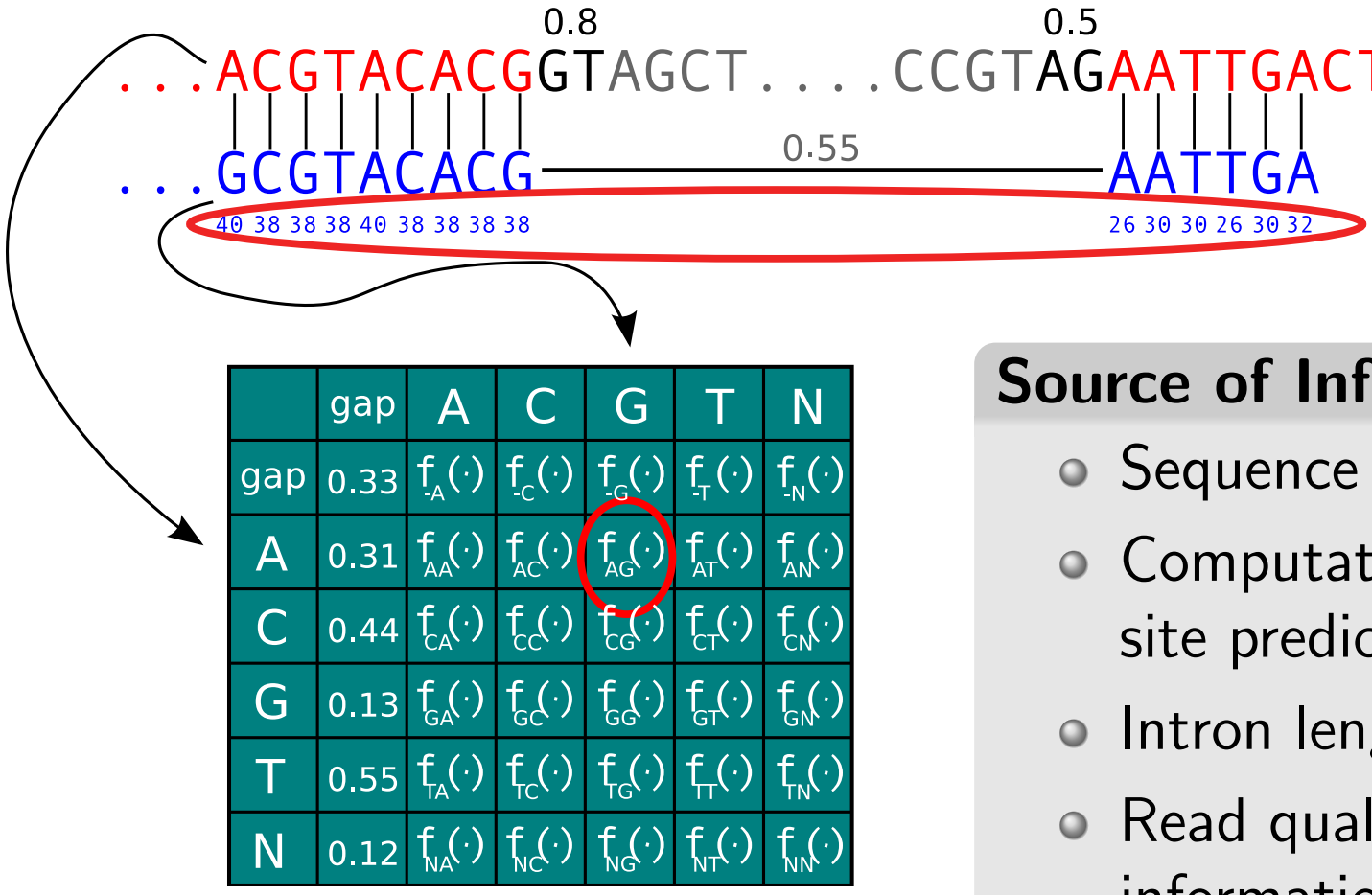
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Alignment Scoring Function



Source of Information

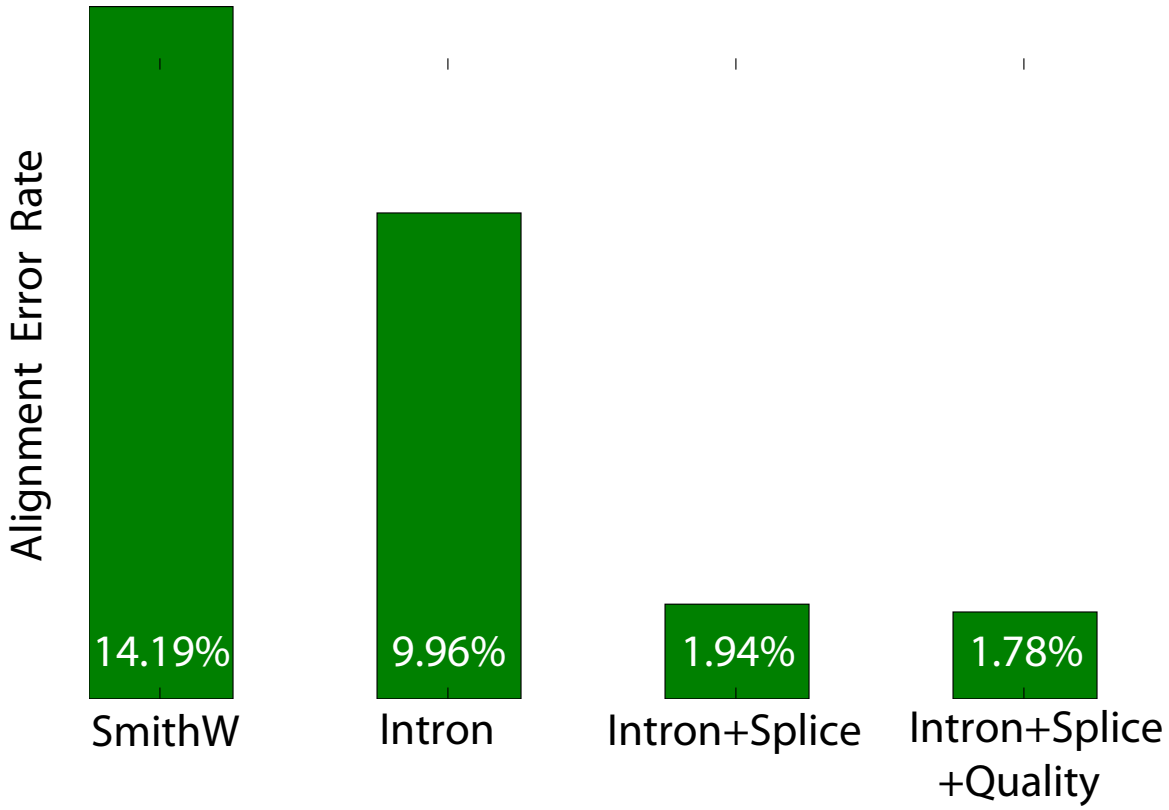
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

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

Proof of Concept

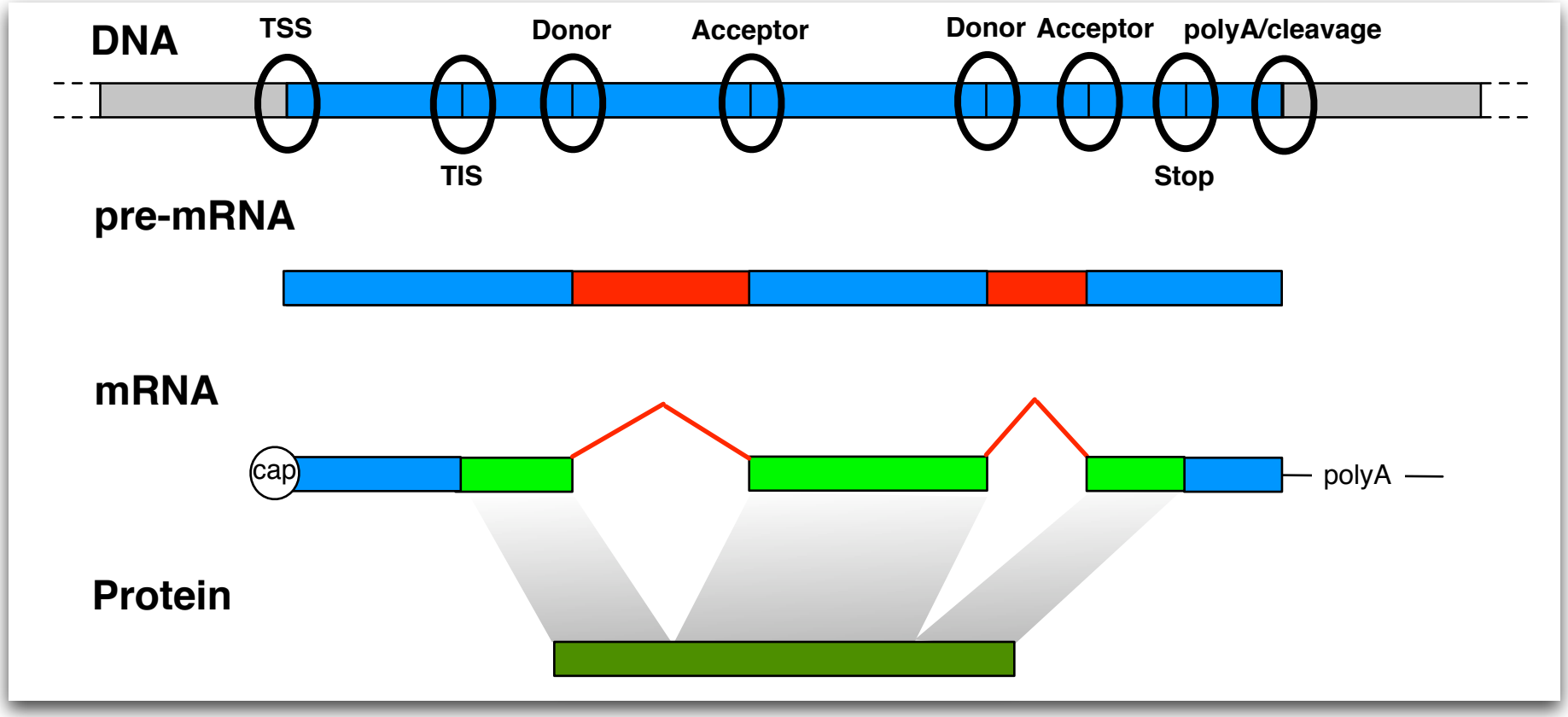
Generate set of artificially spliced reads

- 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]

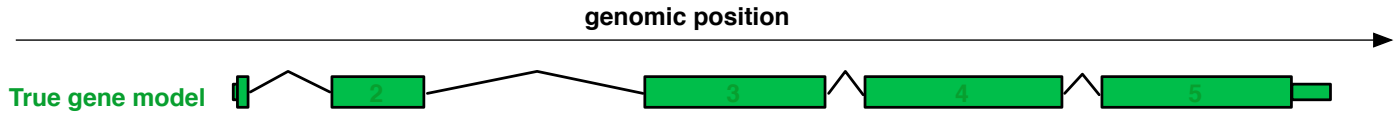
From Genome & Transcriptome to Proteins

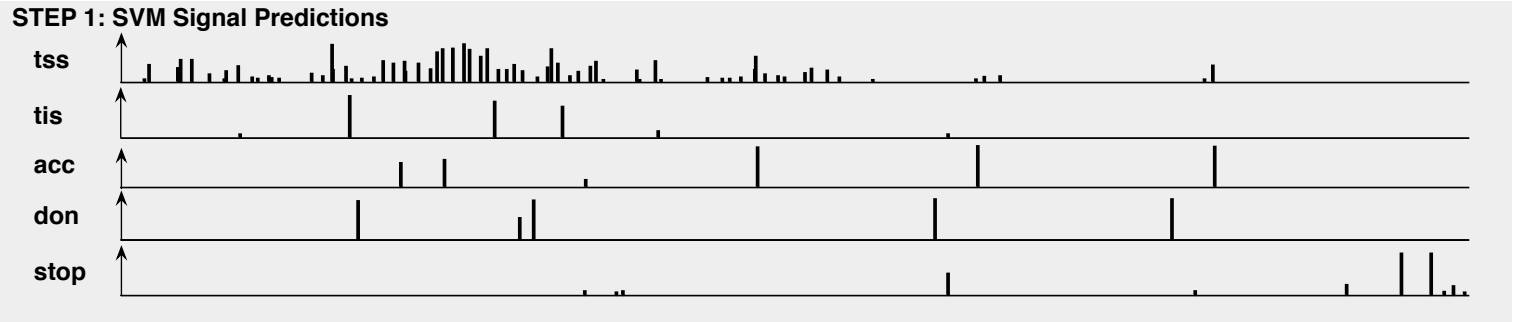


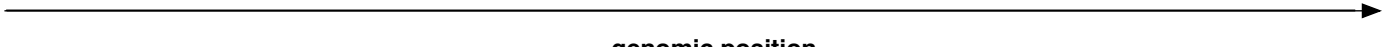
Combine *ab initio* predictions with transcriptome measurements

- Higher accuracy
- Condition/tissue specific predictions

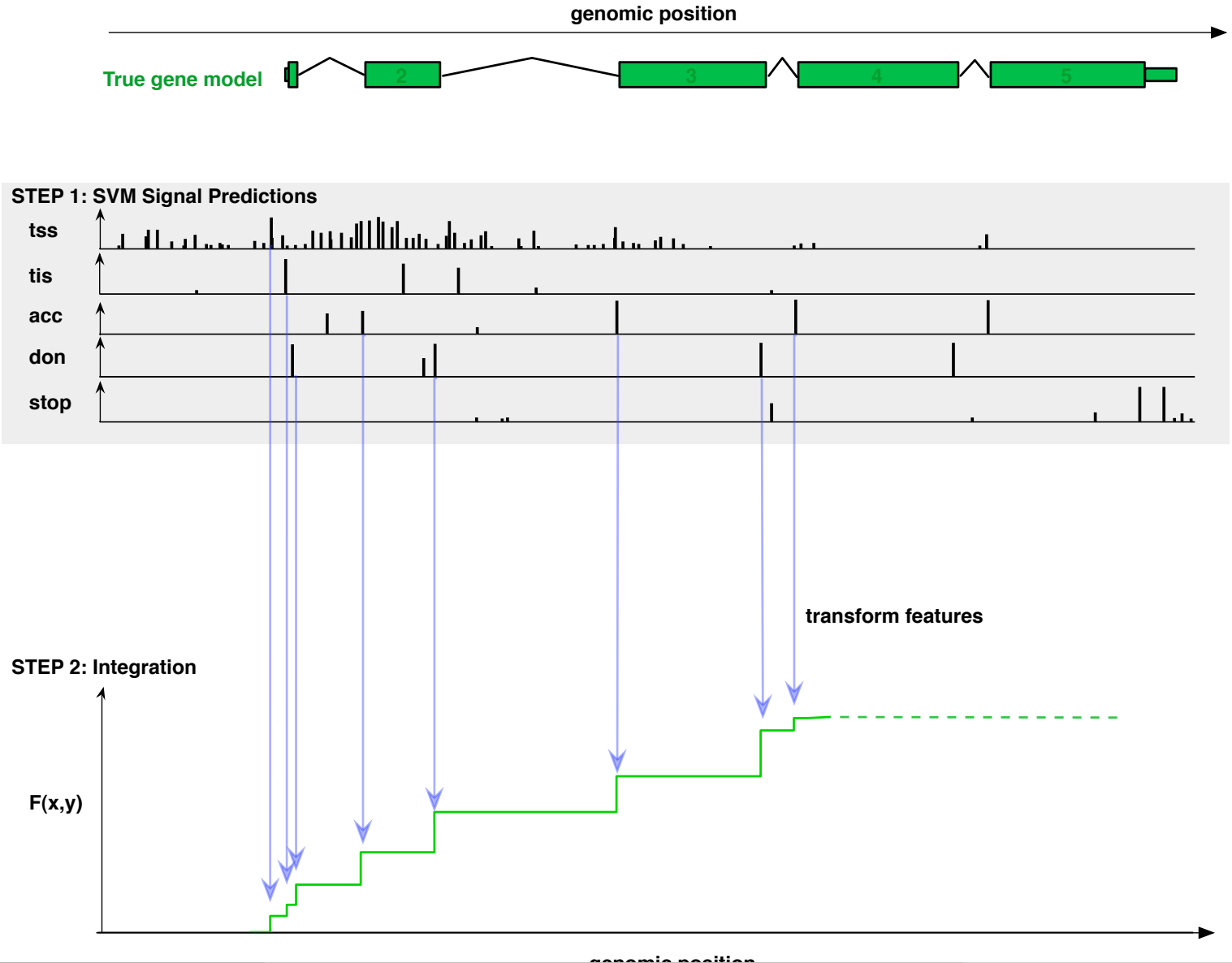
Learning to Integrate Data Sources
[Behr et al., 2008]



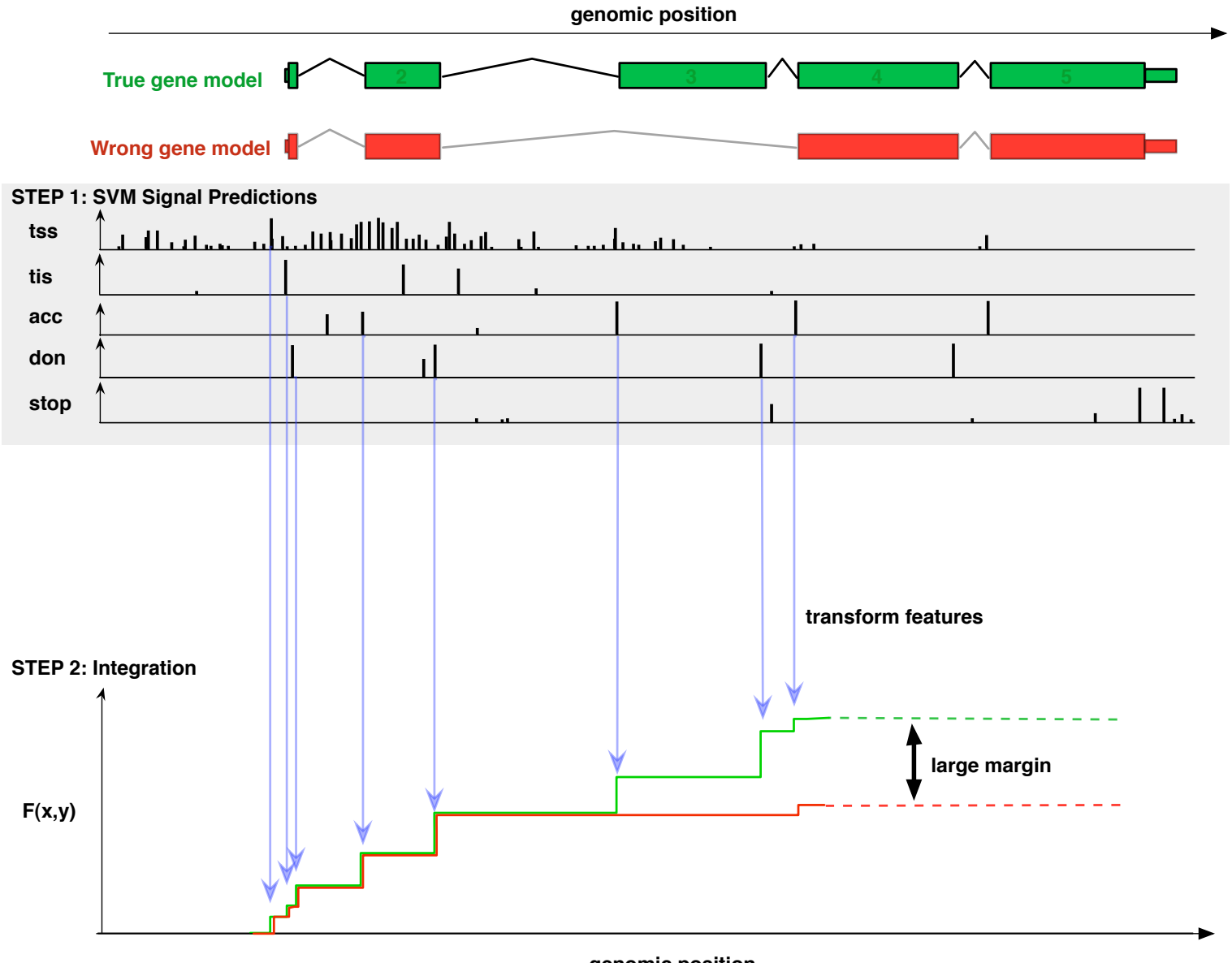




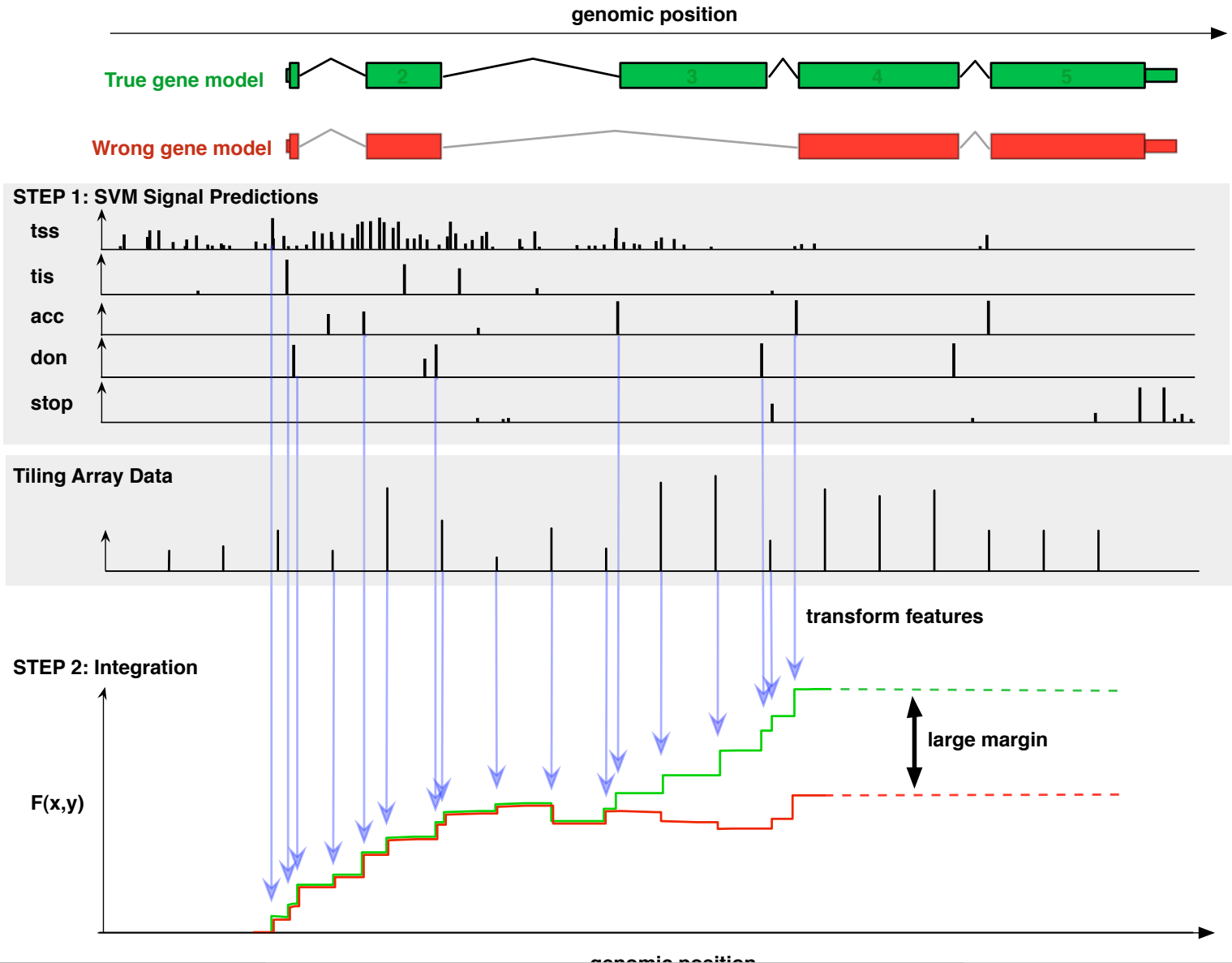
Learning to Integrate Data Sources
[Behr et al., 2008]



Learning to Integrate Data Sources
[Behr et al., 2008]



Learning to Integrate Data Sources
[Behr et al., 2008]



Results: How much does the data help?
[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of step 2
- 300 regions around cDNA confirmed genes for evaluation

Results: How much does the data help?
[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of step 2
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
Tiling arrays (young leaves)	78.0	82.9	**80.4**
Tiling arrays (inflorescence)	77.1	81.6	**79.3**
Tiling arrays (root)	76.2	81.5	**78.9**
Tiling arrays (combined)	77.4	81.4	**79.4**
RNA-seq (w/o spliced reads)	76.8	80.8	**78.7**
RNA-seq (with spliced reads)	79.6	82.1	**80.8**

Results: How much does the data help?
[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of step 2
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
Tiling arrays (young leaves)	78.0	82.9	**80.4**
Tiling arrays (inflorescence)	77.1	81.6	**79.3**
Tiling arrays (root)	76.2	81.5	**78.9**
Tiling arrays (combined)	77.4	81.4	**79.4**
RNA-seq (w/o spliced reads)	76.8	80.8	**78.7**
RNA-seq (with spliced reads)	79.6	82.1	**80.8**

Results: How much does the data help?
[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of step 2
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
Tiling arrays (young leaves)	78.0	82.9	**80.4**
Tiling arrays (inflorescence)	77.1	81.6	**79.3**
Tiling arrays (root)	76.2	81.5	**78.9**
Tiling arrays (combined)	77.4	81.4	**79.4**
RNA-seq (w/o spliced reads)	76.8	80.8	**78.7**
RNA-seq (with spliced reads)	79.6	82.1	**80.8**

Results: How much data is needed?
[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of step 2
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
RNA-seq 1/128	75.5	78.7	**77.1**
RNA-seq 1/64	76.8	79.5	**78.1**
RNA-seq 1/32	76.2	79.6	**77.9**
RNA-seq 1/16	77.1	79.1	**78.1**
RNA-seq 1/8	78.6	80.1	**79.4**
RNA-seq 1/4	77.4	79.6	**78.5**
RNA-seq 1/2	79.3	82.1	**80.6**
RNA-seq	79.6	82.1	**80.8**

Results: How much data is needed?
[Behr et al., 2008]

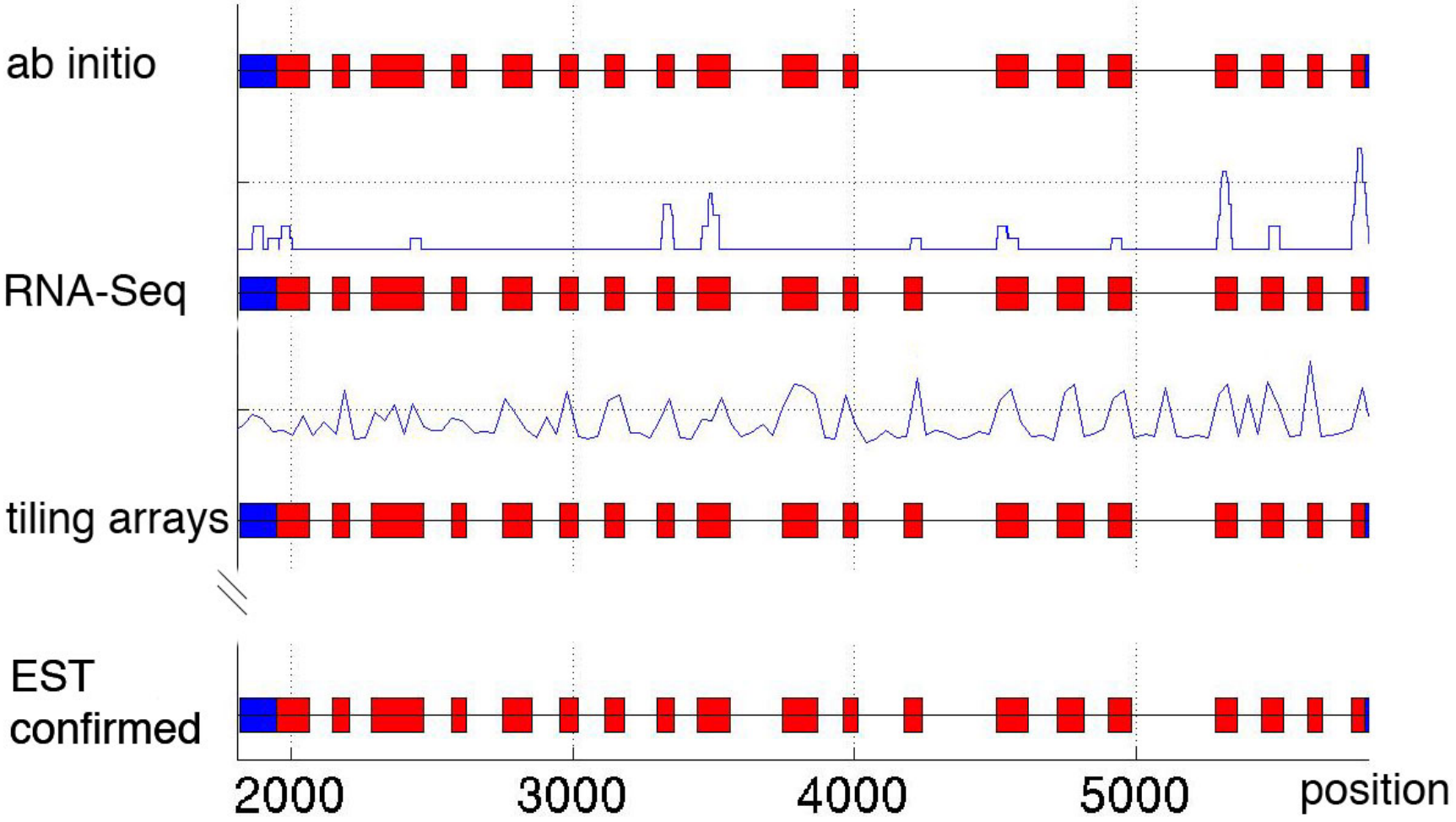
Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of step 2
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
RNA-seq 1/128	75.5	78.7	**77.1**
RNA-seq 1/64	76.8	79.5	**78.1**
RNA-seq 1/32	76.2	79.6	**77.9**
RNA-seq 1/16	77.1	79.1	**78.1**
RNA-seq 1/8	78.6	80.1	**79.4**
RNA-seq 1/4	77.4	79.6	**78.5**
RNA-seq 1/2	79.3	82.1	**80.6**
RNA-seq	79.6	82.1	**80.8**

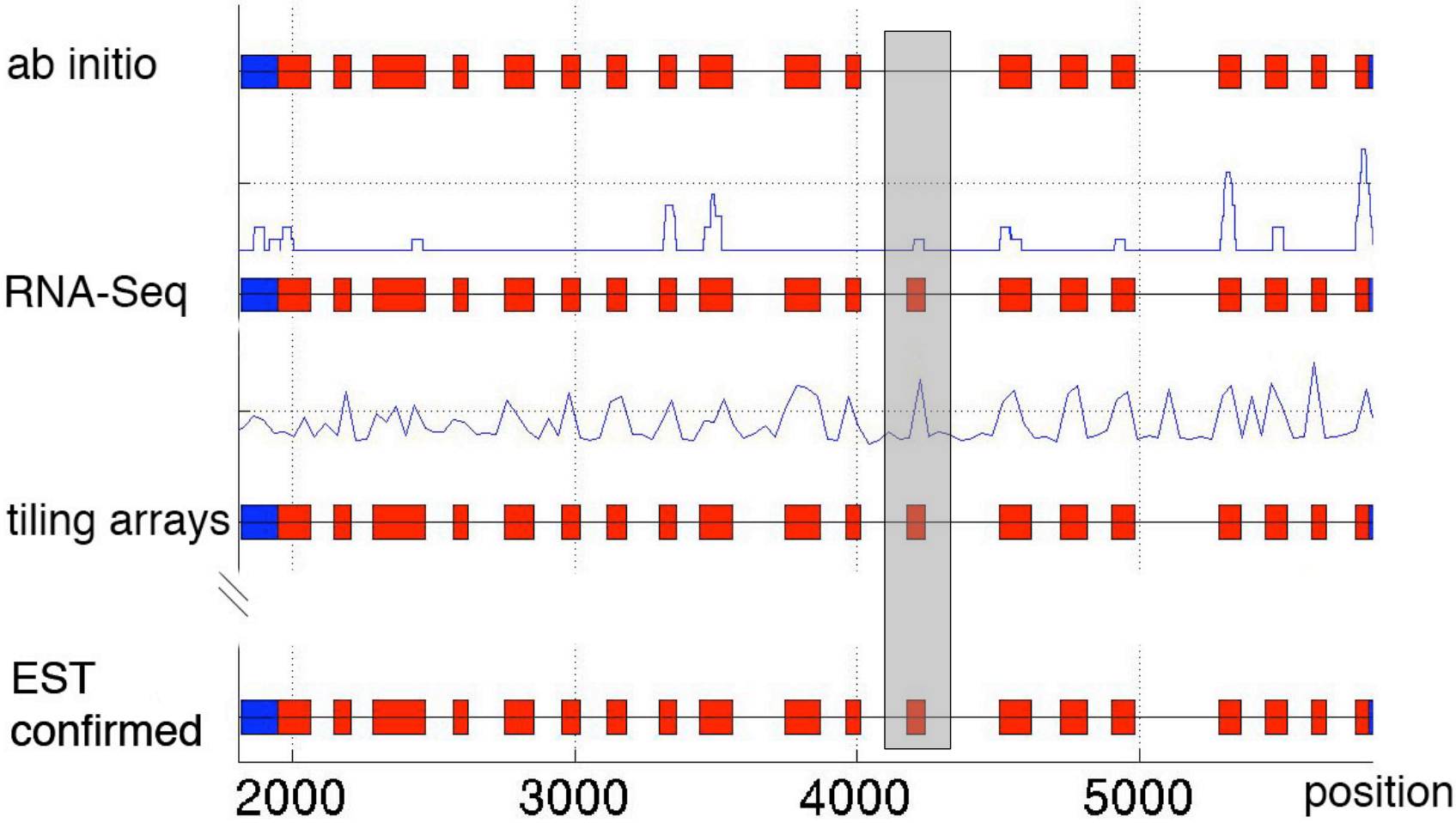
Results: Combining helps!

[Behr et al., 2008]



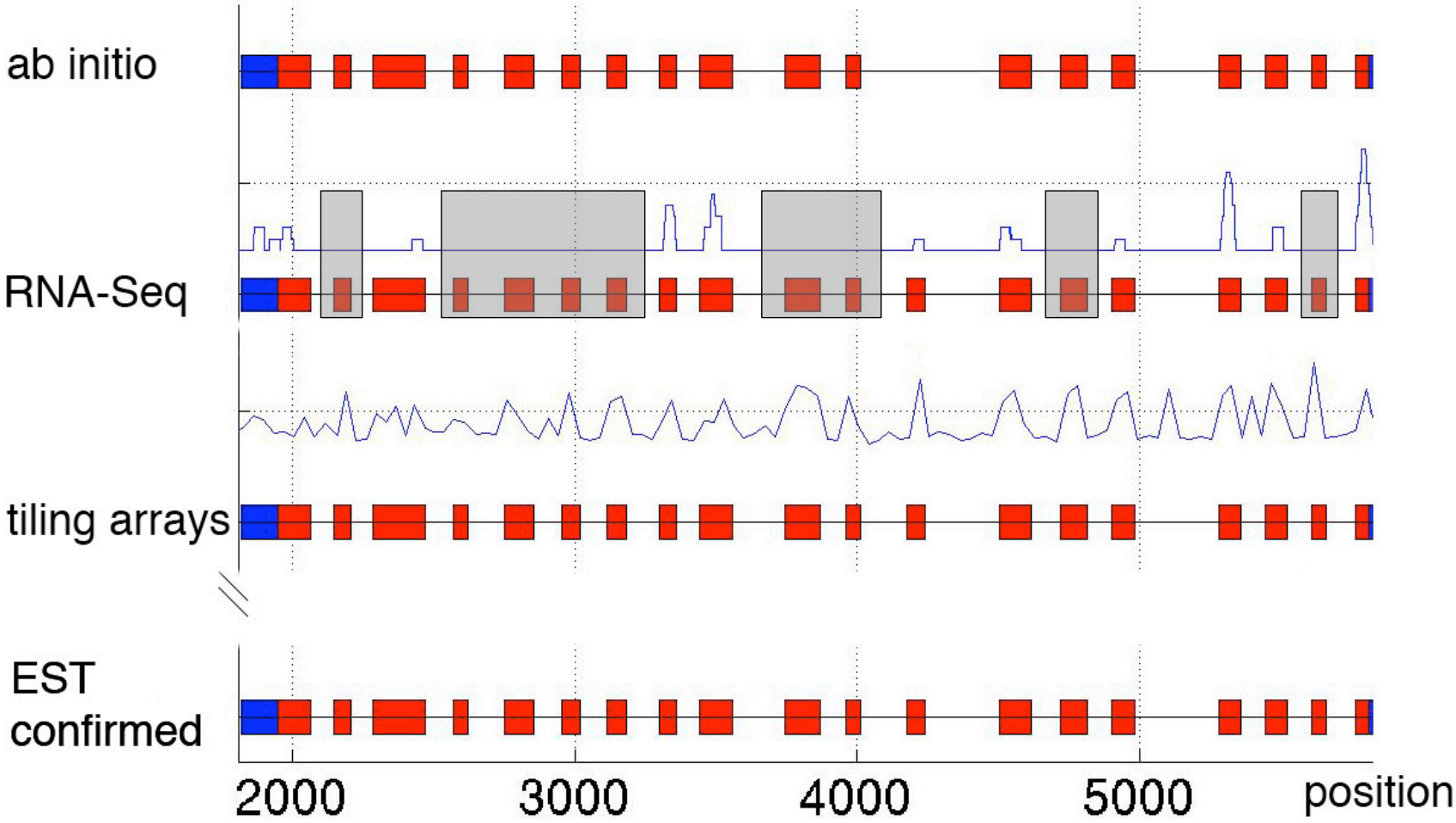
Results: Combining helps!

[Behr et al., 2008]



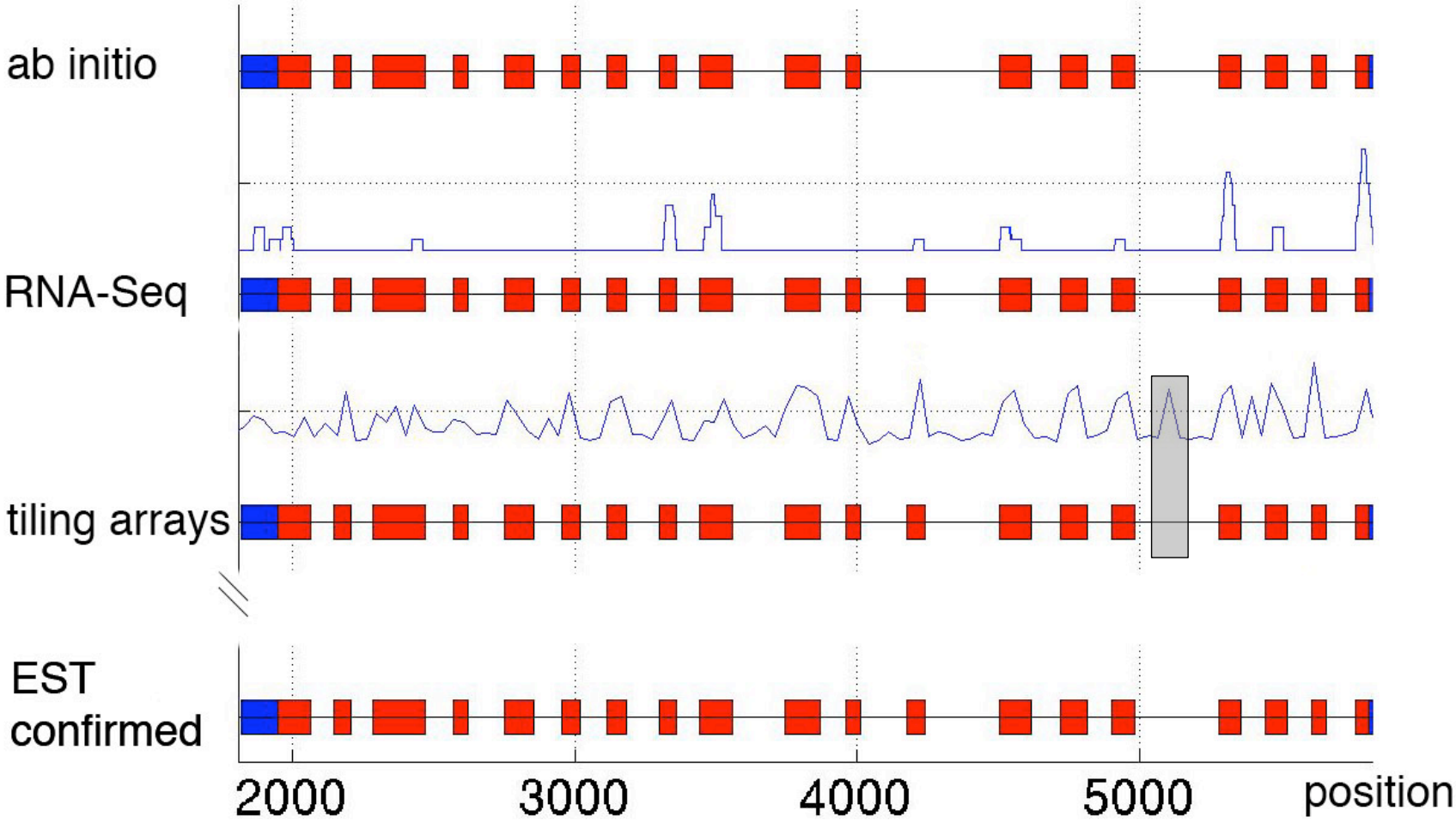
Results: Combining helps!

[Behr et al., 2008]

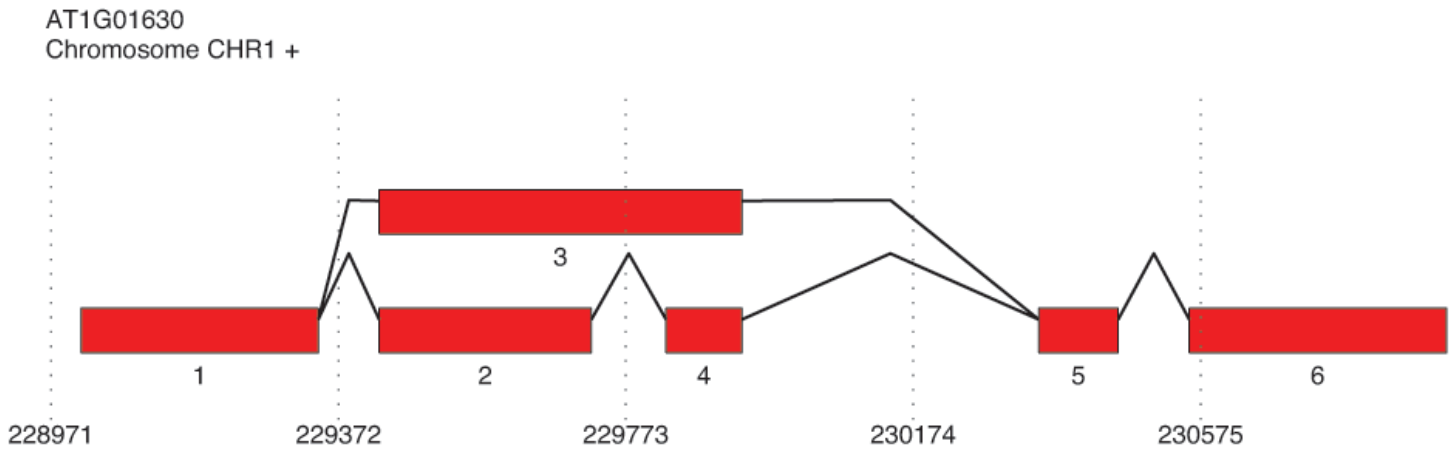


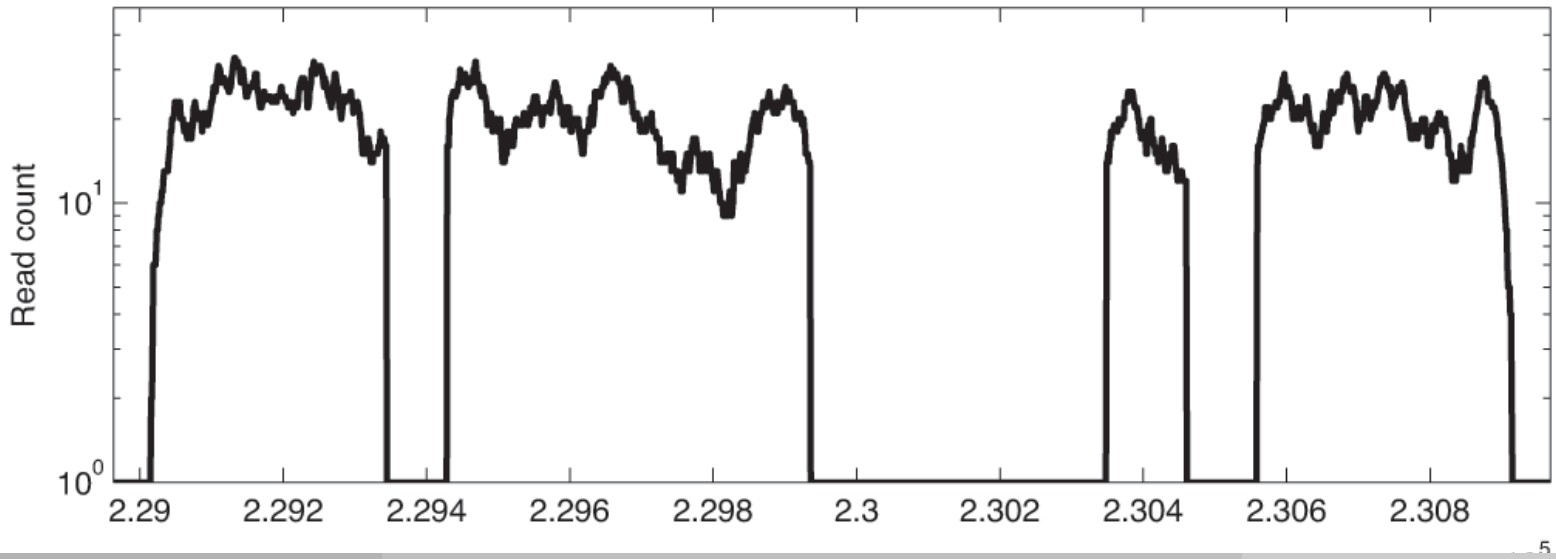
Results: Combining helps!

[Behr et al., 2008]

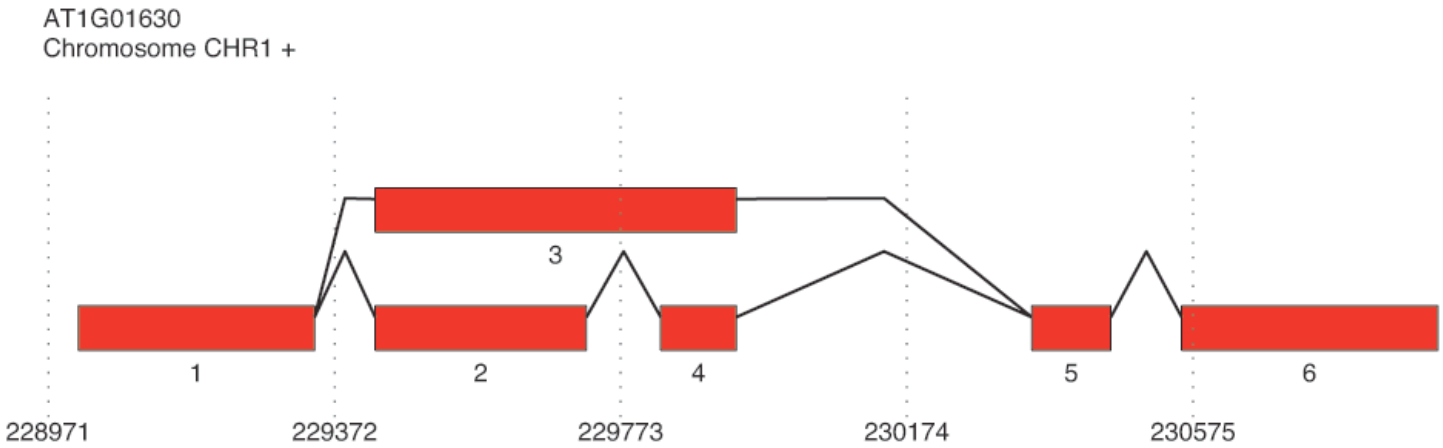


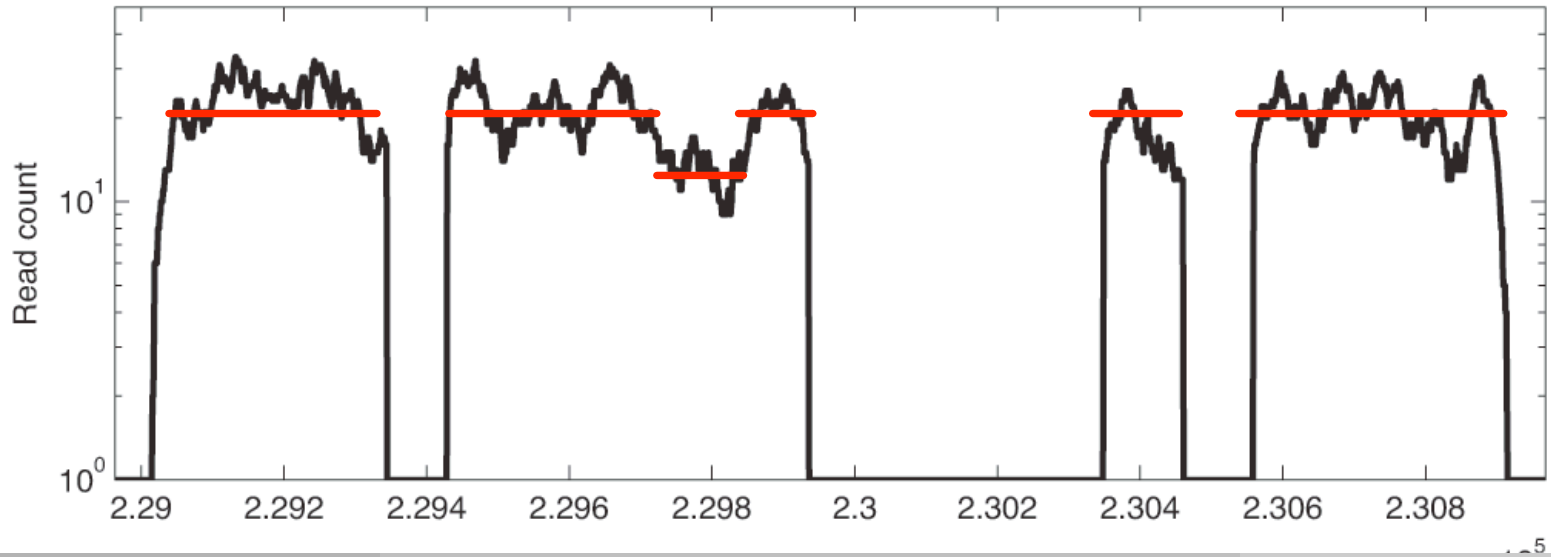
Extension to Alternative Transcripts?
(Preliminary)



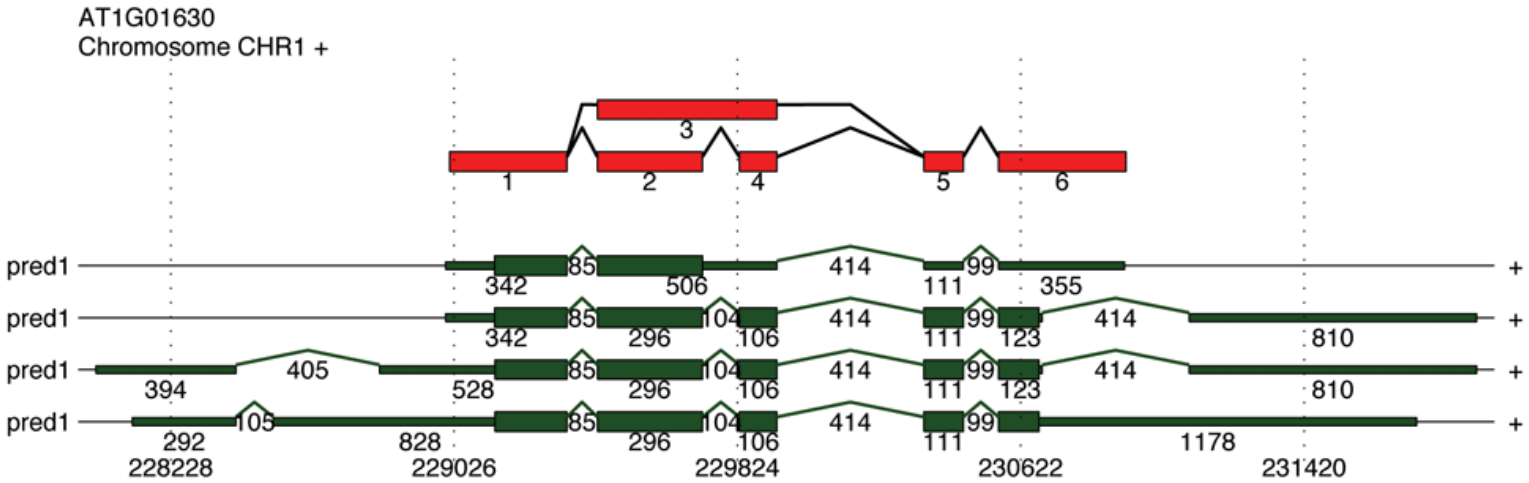


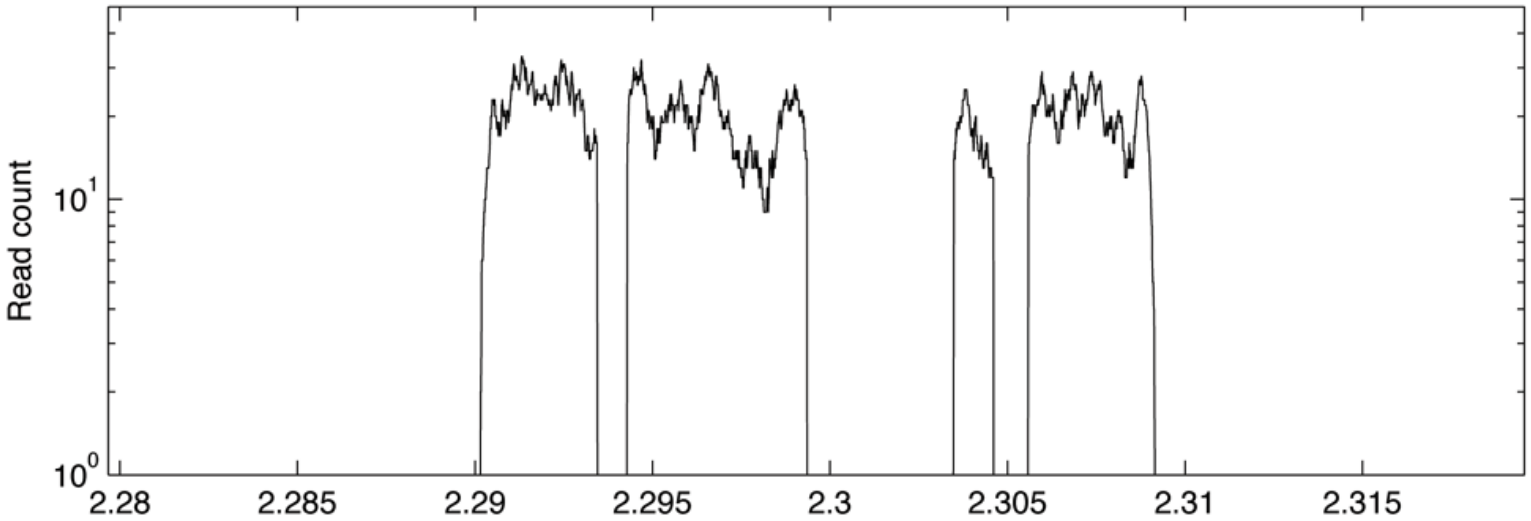
Extension to Alternative Transcripts?
(Preliminary)



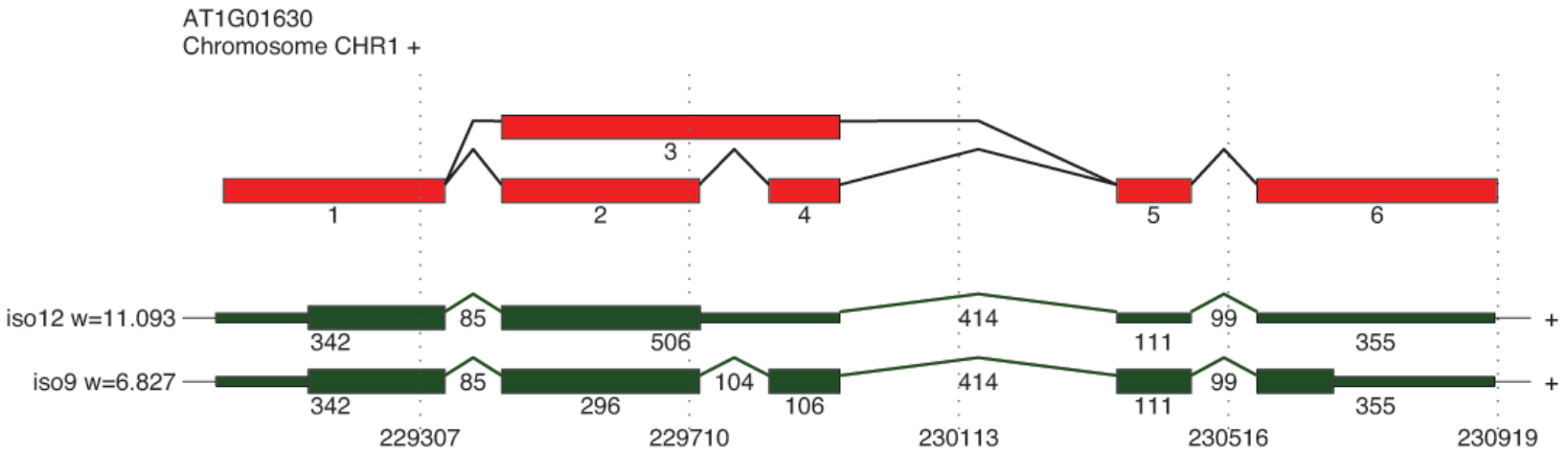


Extension to Alternative Transcripts?
(Preliminary)

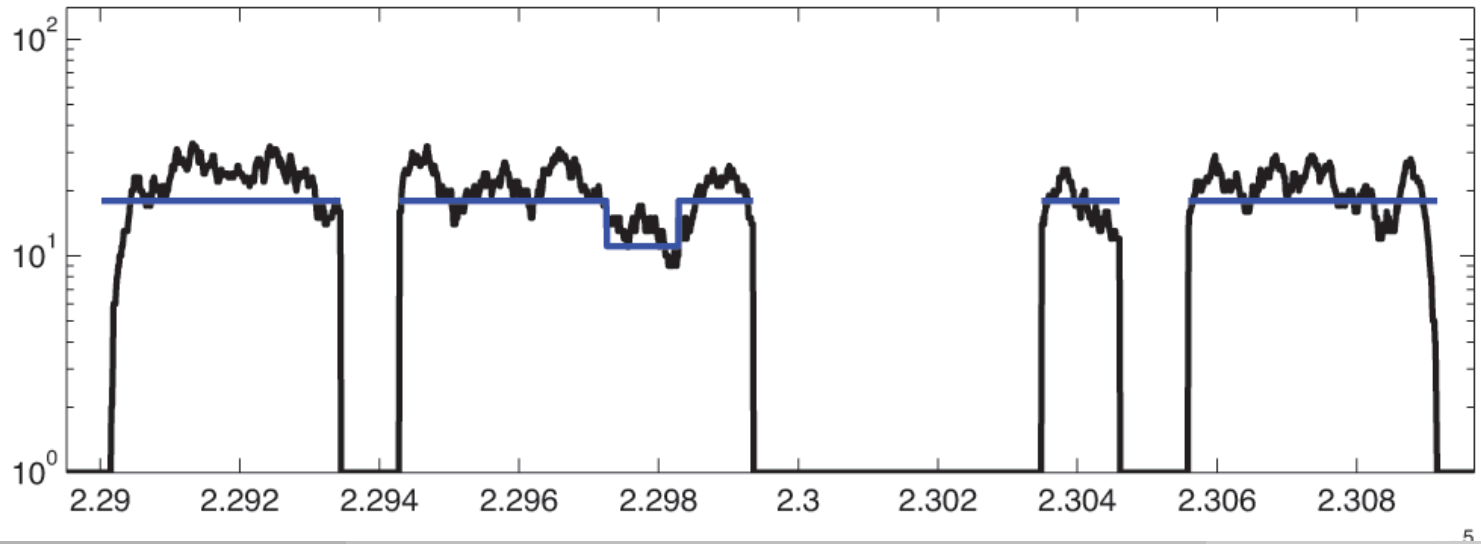




Extension to Alternative Transcripts?
(Preliminary)



Determine isoforms and abundance estimate simultaneously.



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 tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- 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 due to novel inference methods
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- 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 due to novel inference methods
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- 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 due to novel inference methods
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- 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/MPI)
Jonas Behr (FML)
Sören Sonnenburg (FML/FIRST)
Alex Zien (FML & FIRST)
Georg Zeller (FML/MPI)

Short Read Analysis

Fabio De Bona (FML)
Stephan Ossowski (MPI)
Korbinian Schneeberger (MPI)
Jun Cao (MPI)
Detlef Weigel (MPI)

Tiling Arrays

Georg Zeller (FML/MPI)
Johannes Eichner (FML)
Stefan Henz (MPI)
Sascha Laubinger (MPI)
Detlef Weigel (MPI)

More Information

- http://www.fml.mpg.de/raetsch
- http://www.shogun-toolbox.org
- http://www.mgene.org/webservice
- Slides with references are available online

Thank you!

Acknowledgments

Gene Finding

Gabi Schweikert (FML/MPI)
Jonas Behr (FML)
Sören Sonnenburg (FML/FIRST)
Alex Zien (FML & FIRST)
Georg Zeller (FML/MPI)

Short Read Analysis

Fabio De Bona (FML)
Stephan Ossowski (MPI)
Korbinian Schneeberger (MPI)
Jun Cao (MPI)
Detlef Weigel (MPI)

Tiling Arrays

Georg Zeller (FML/MPI)
Johannes Eichner (FML)
Stefan Henz (MPI)
Sascha Laubinger (MPI)
Detlef Weigel (MPI)

More Information

- `http://www.fml.mpg.de/raetsch`
- `http://www.shogun-toolbox.org`
- `http://www.mgene.org/webservice`
- Slides with references are available online

Thank you!

Acknowledgments

Gene Finding

Gabi Schweikert (FML/MPI)
Jonas Behr (FML)
Sören Sonnenburg (FML/FIRST)
Alex Zien (FML & FIRST)
Georg Zeller (FML/MPI)

Short Read Analysis

Fabio De Bona (FML)
Stephan Ossowski (MPI)
Korbinian Schneeberger (MPI)
Jun Cao (MPI)
Detlef Weigel (MPI)

Tiling Arrays

Georg Zeller (FML/MPI)
Johannes Eichner (FML)
Stefan Henz (MPI)
Sascha Laubinger (MPI)
Detlef Weigel (MPI)

More Information

- `http://www.fml.mpg.de/raetsch`
- `http://www.shogun-toolbox.org`
- `http://www.mgene.org/webservice`
- 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. Presented at 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.

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.

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.

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, March 2009.

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

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

G. Rätsch and S. Sonnenburg. Accurate splice site detection for *Caenorhabditis elegans*. In K. Tsuda B. Schoelkopf and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

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

G. Rätsch, S. Sonnenburg, 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.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

References III

G. Schweikert, G. Zeller, A. Zien, J. Behr, C.S. Ong, P. Philips, A. Bohlen, R. Bohnert, F. De Bona, S. Sonnenburg, and G. Rätsch. mGene: Accurate computational gene finding with application to nematode genomes. under revision, March 2009.

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, 2008. 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.

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/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.