Towards the Understanding of Transcriptome Complexity and Plasticity

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society

Tübingen, Germany

Future of Computational Biology

September 22, 2009



Friedrich Miescher Laboratory
of the Max Planck Society



Machine Learning in Biology

Research Topics

Machine Learning

1. Inference methods for structured data
   ⇒ *Develop fast and accurate learning methods*

Molecular Biology

2. Genome annotation
   ⇒ *Predict features encoded on DNA*
3. Biological networks
   ⇒ *Understand interactions between gene products*
4. Analysis of polymorphisms
   ⇒ *Discover polymorphisms and associate with phenotypes*

Machine Learning

$\rightsquigarrow$ Develop fast, accurate and interpretable learning methods



Given: Observations of some complex phenomenon
Goal: Learn from data & build predictive models

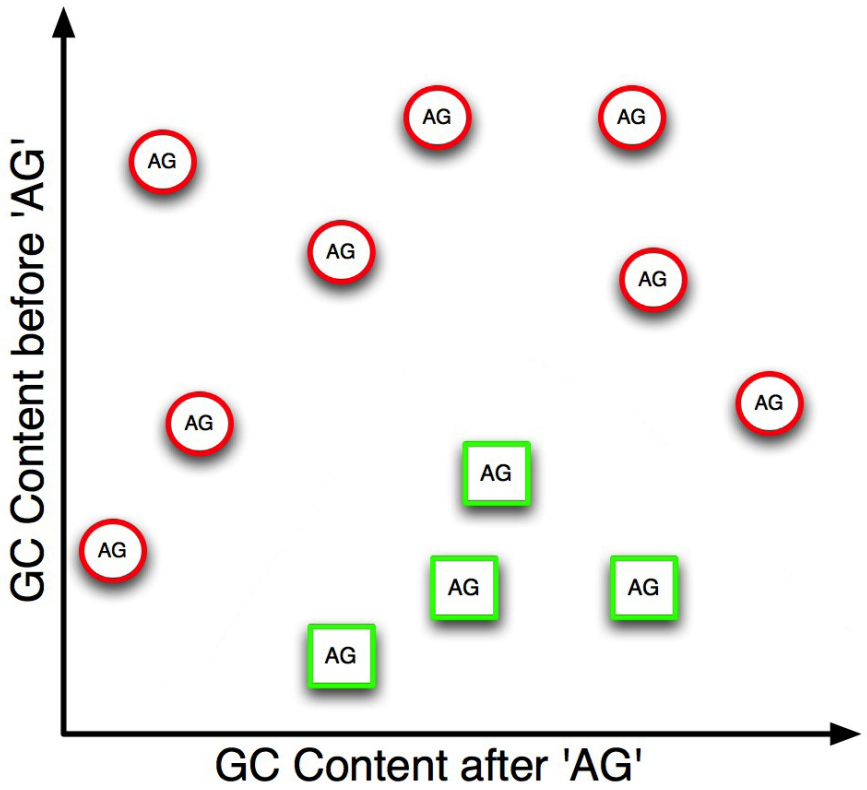
Machine Learning

⇝ Develop fast, accurate and interpretable learning methods

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

Example:



Two different classes of observations

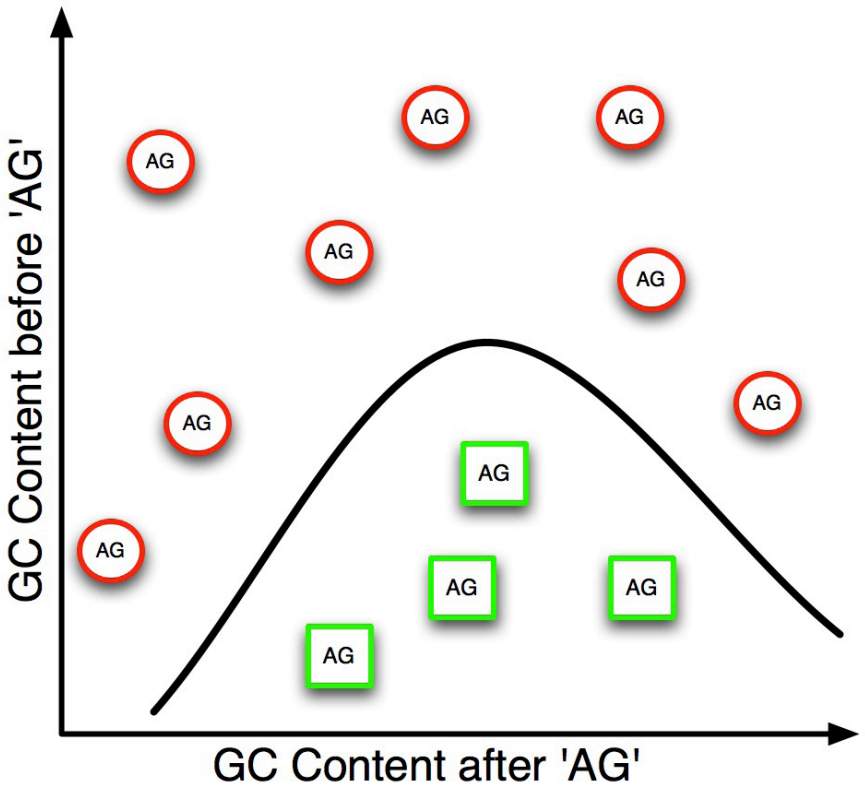
Machine Learning

⇝ Develop fast, accurate and interpretable learning methods

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

Example:



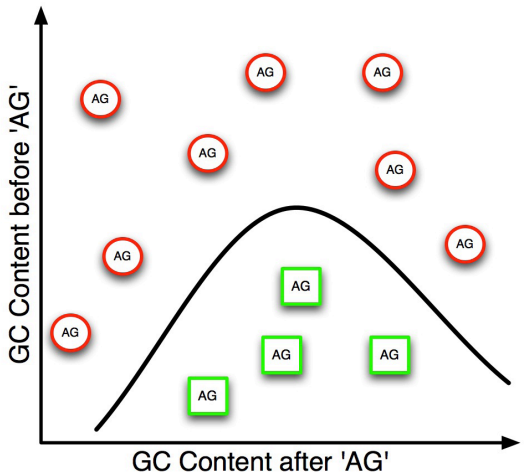
Inferred classification rule

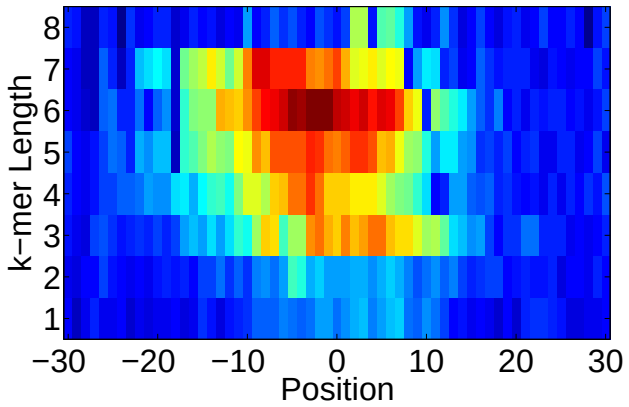
Machine Learning

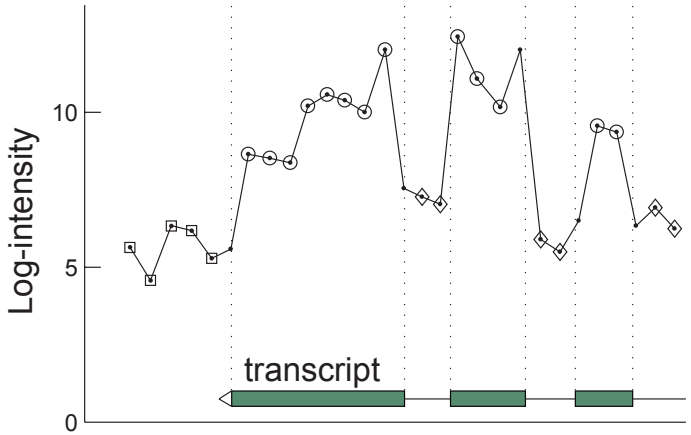
⇝ Develop fast, accurate and interpretable learning methods

Given: Observations of some complex phenomenon
Goal: Learn from data & build predictive models

1. Large scale sequence classification
2. Analysis and explanation of learning results
3. Sequence segmentation & structure prediction

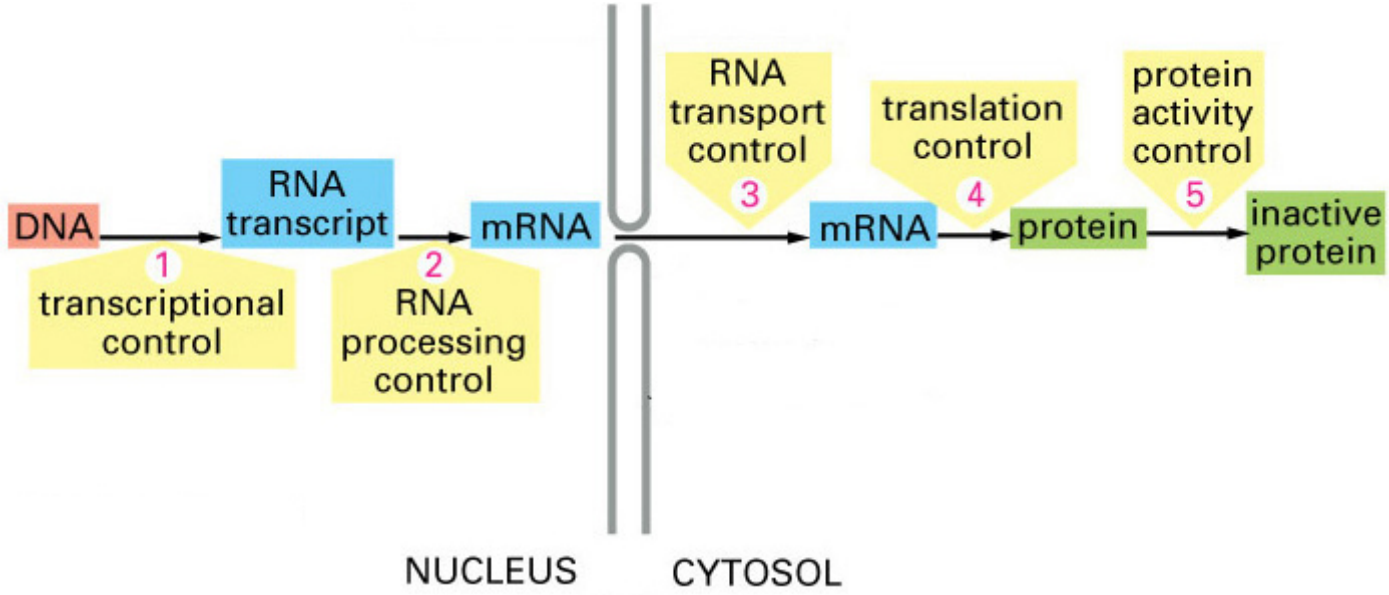






Learning about the Transcriptome

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



Machine Learning View

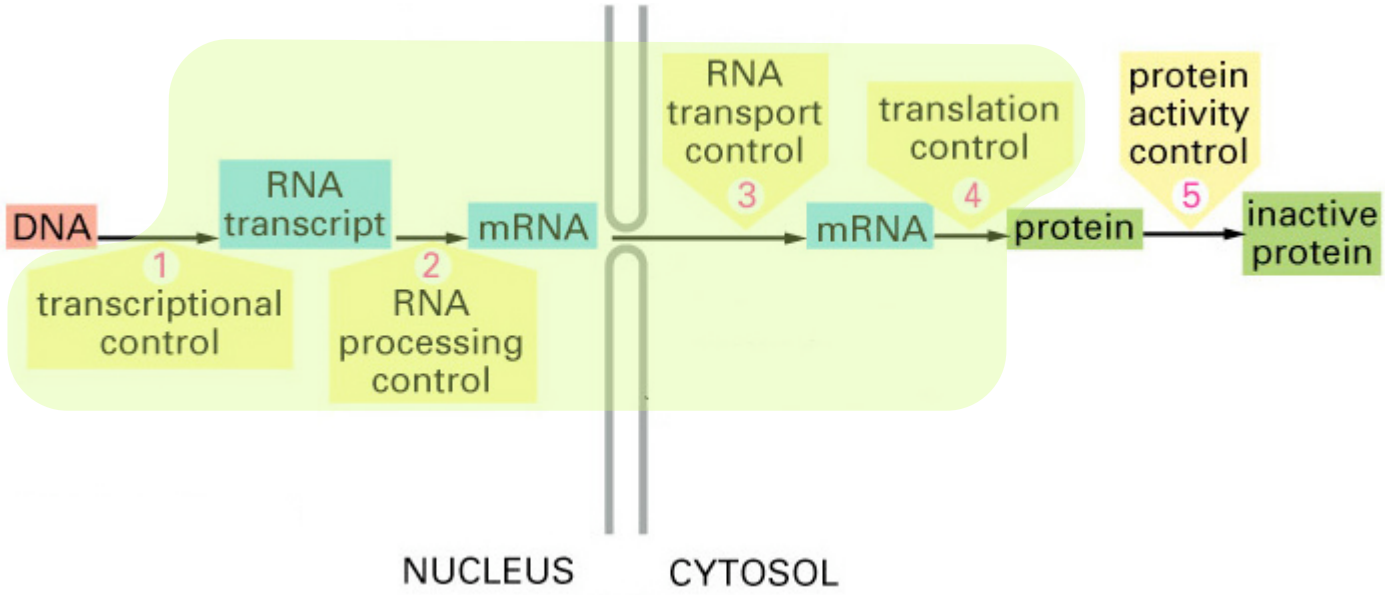
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

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

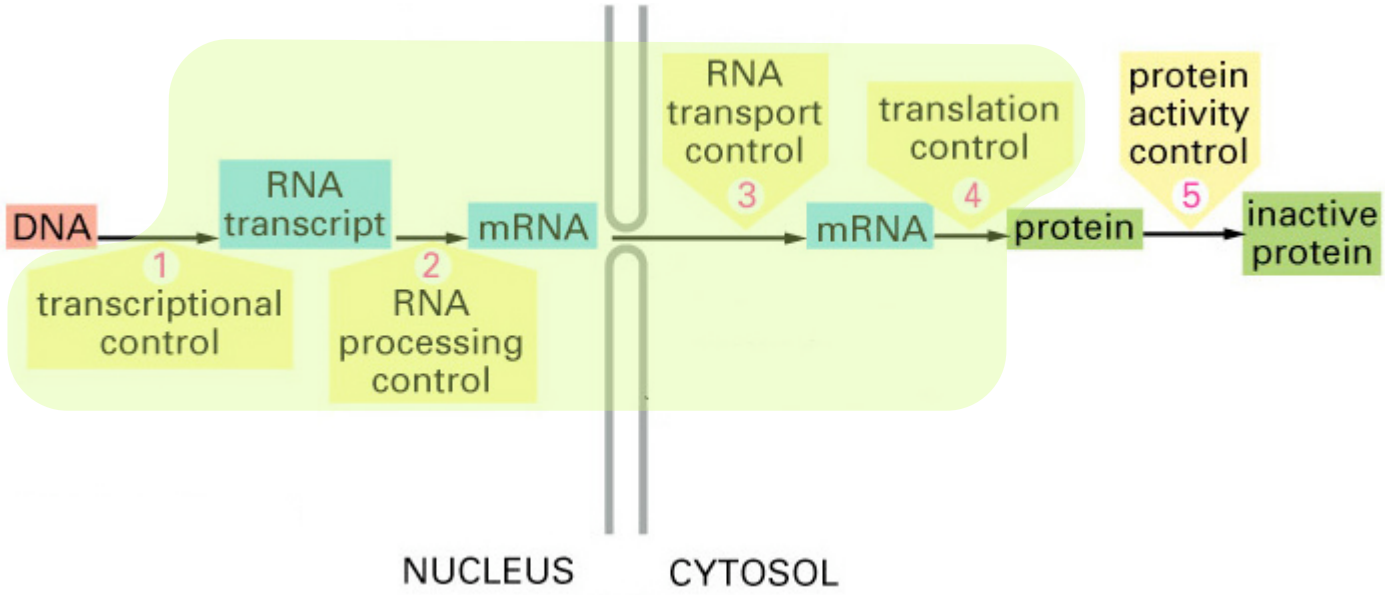
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

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

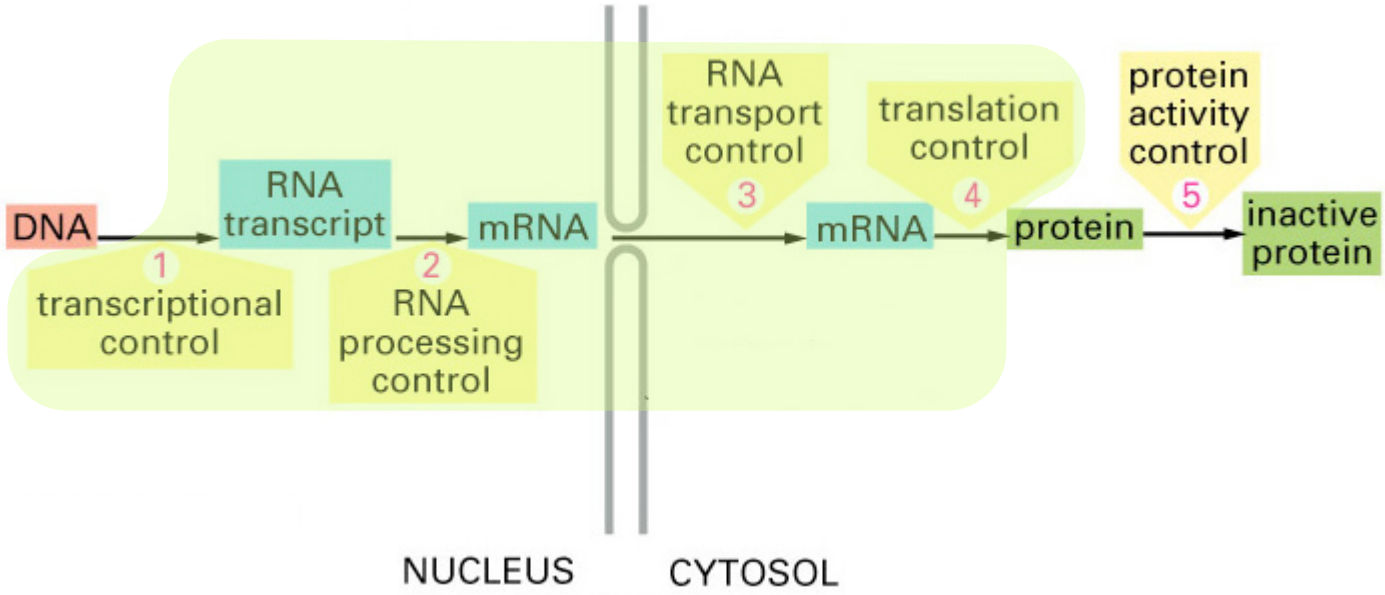
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

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?



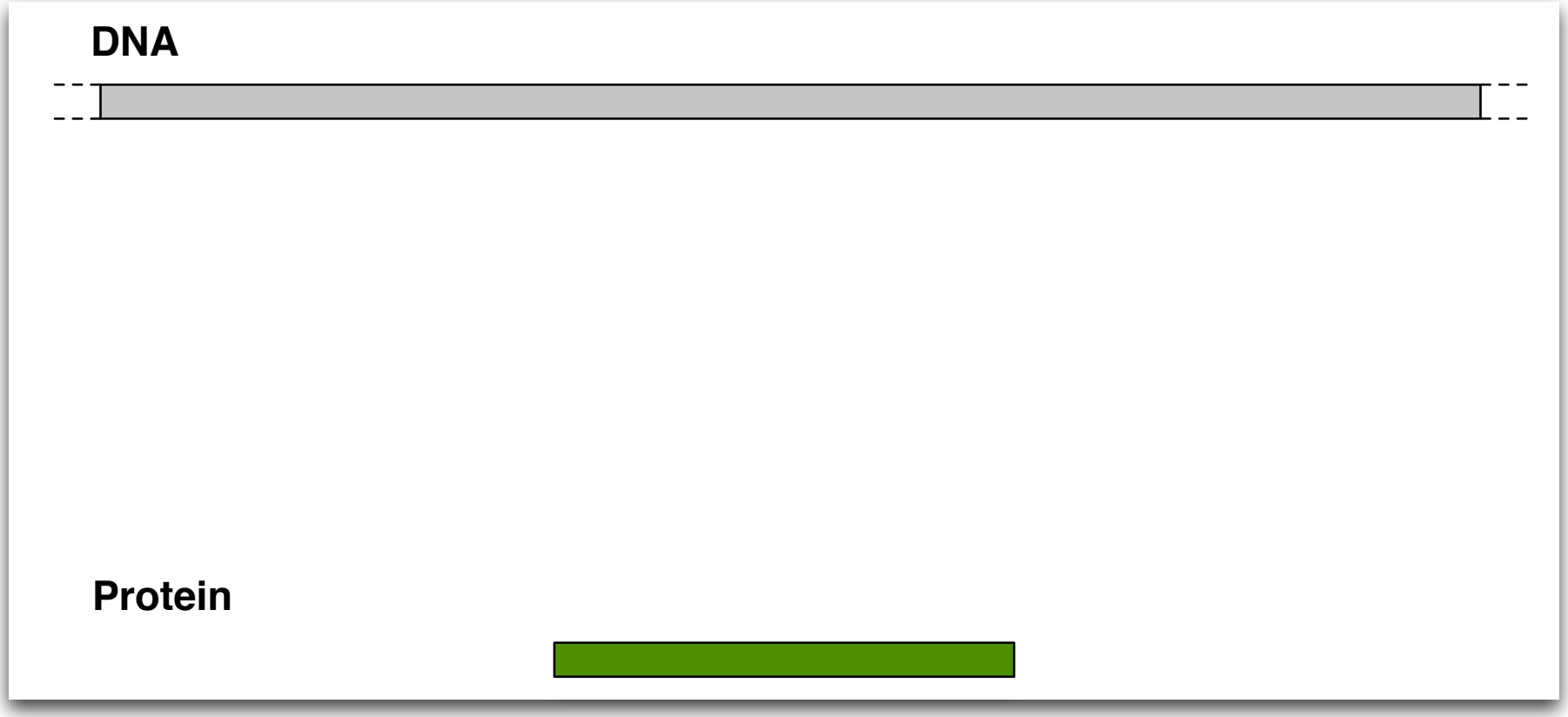
Machine Learning View

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

Biological View

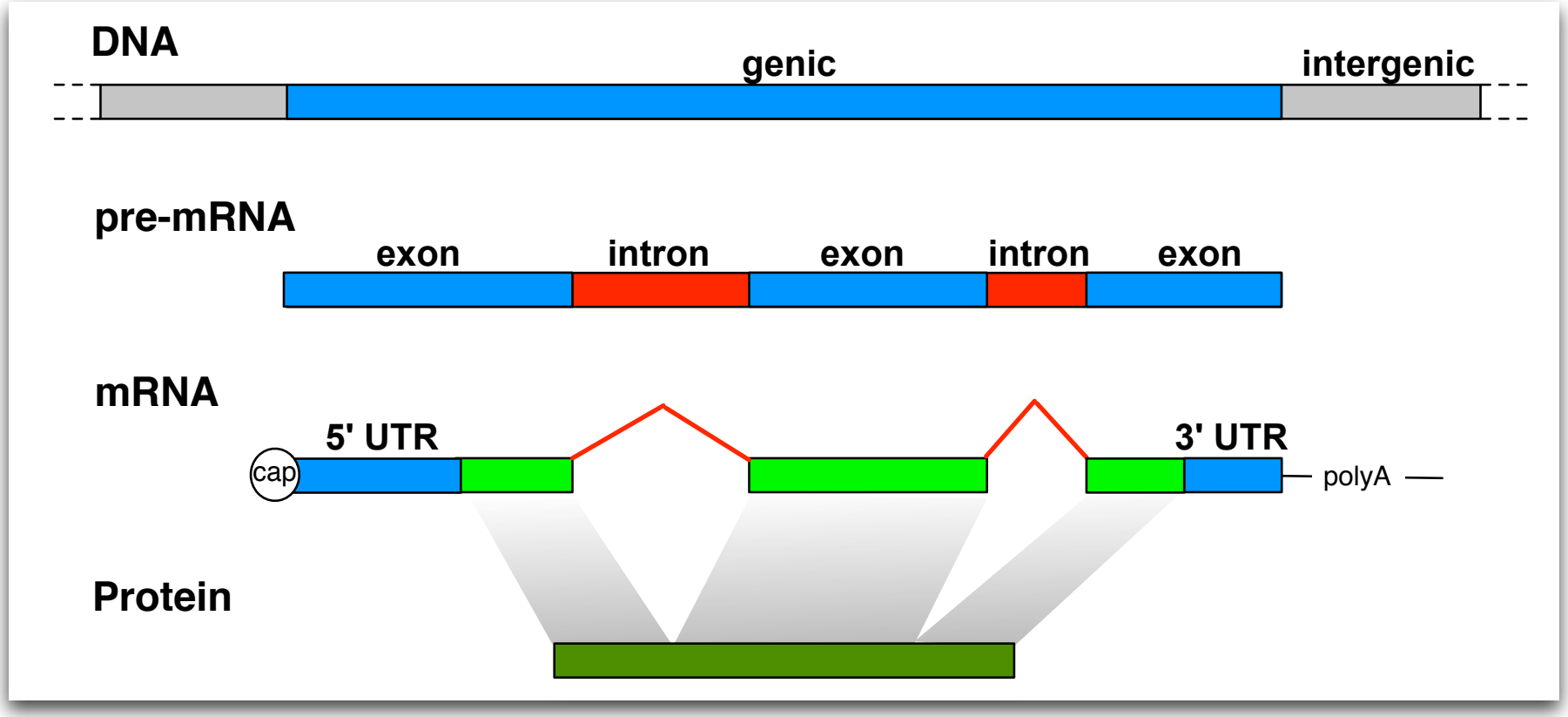
1. What can we not predict yet? What is missing?
2. Can we derive a deeper understanding of these processes?

1st Step: Computational Gene Finding



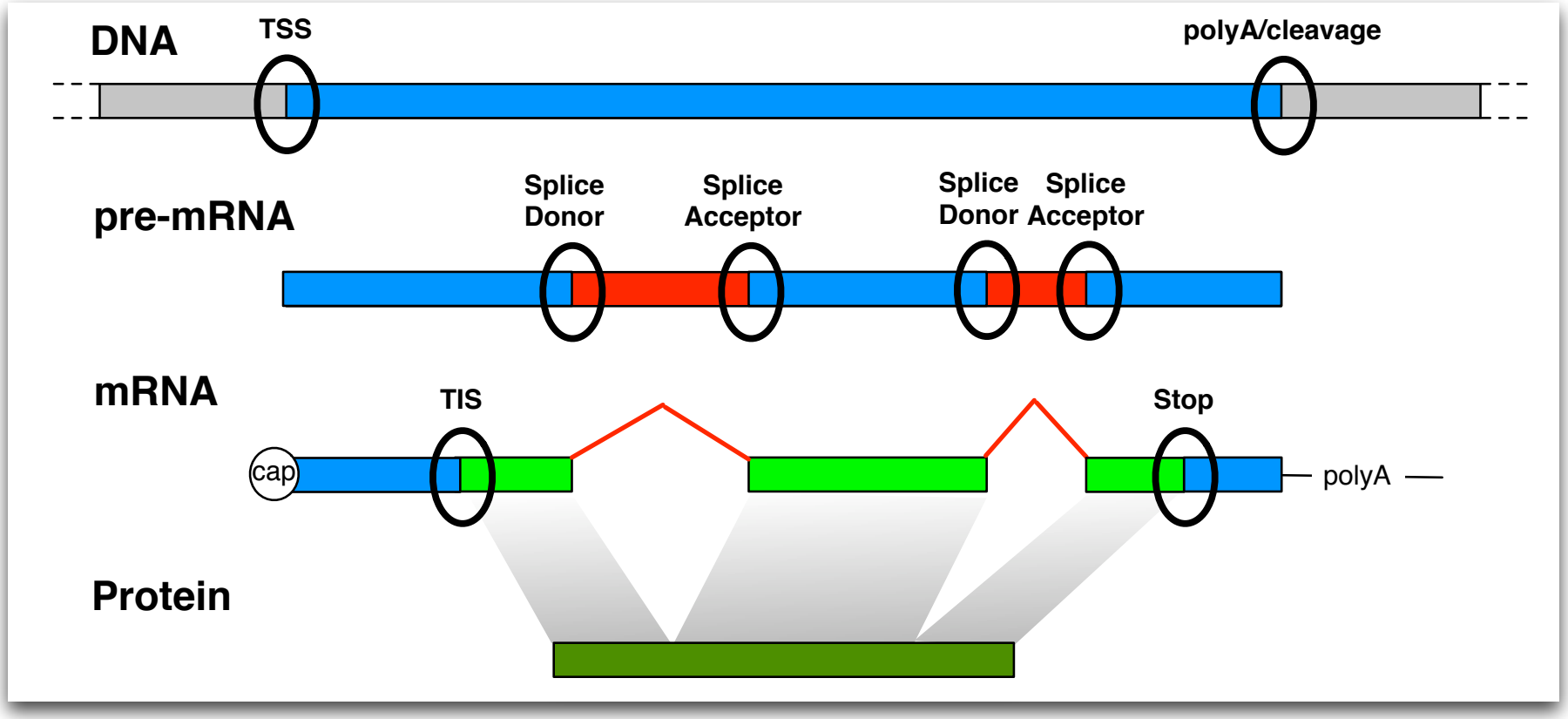
- Given a piece of DNA sequence
- Predict gene products

1st Step: Computational Gene Finding



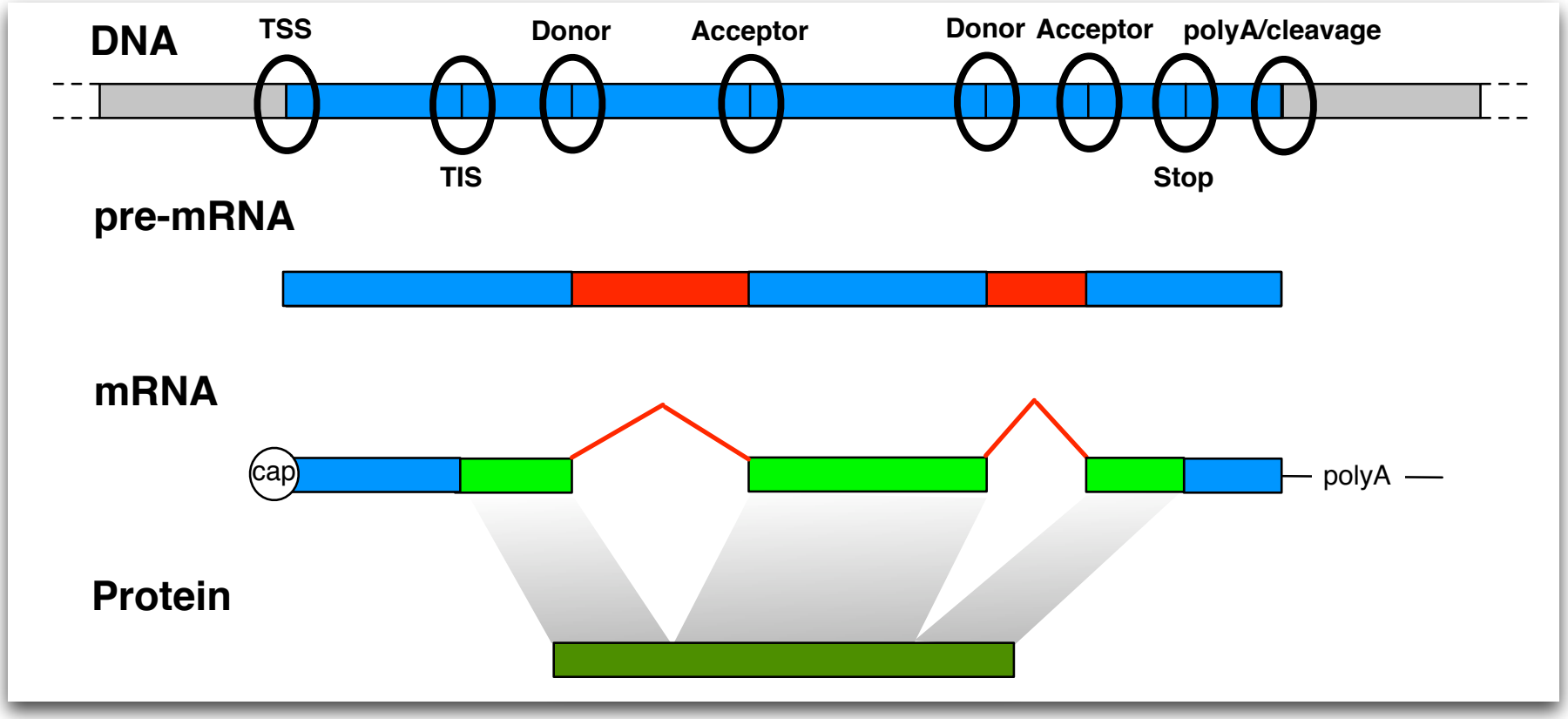
- Given a piece of DNA sequence
- Predict gene products inluding the intermediate processing steps

1st Step: Computational Gene Finding



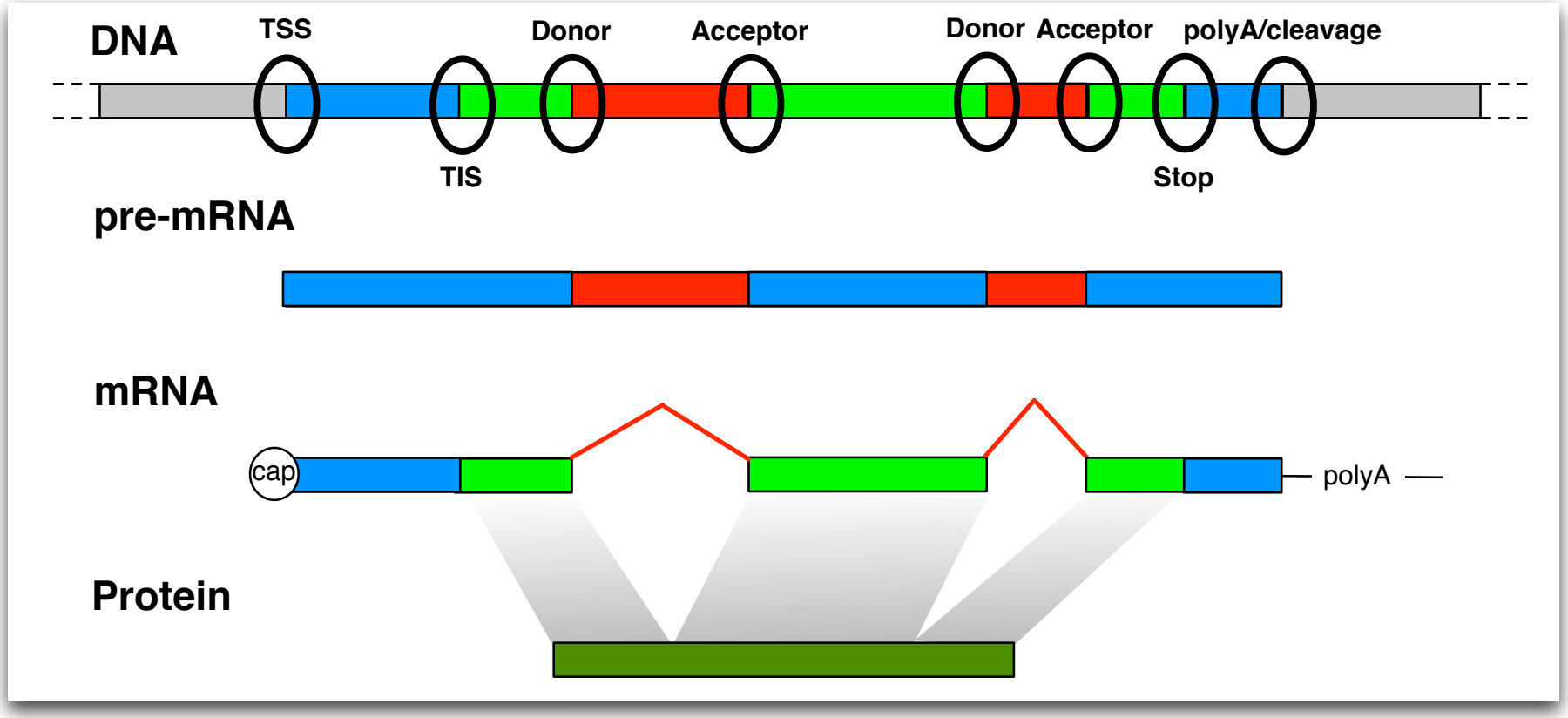
- Predict signals used during processing

1st Step: Computational Gene Finding



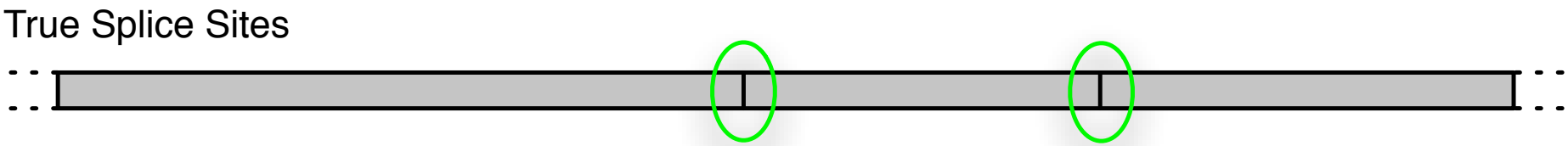
- Predict signals used during processing

1st Step: Computational Gene Finding

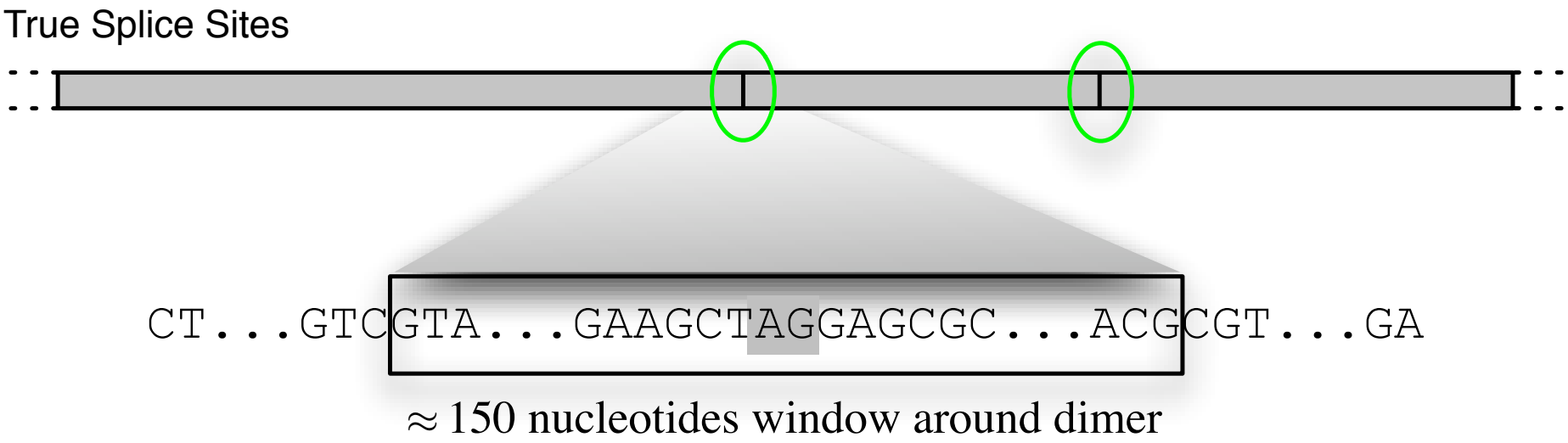


- Predict signals used during processing
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

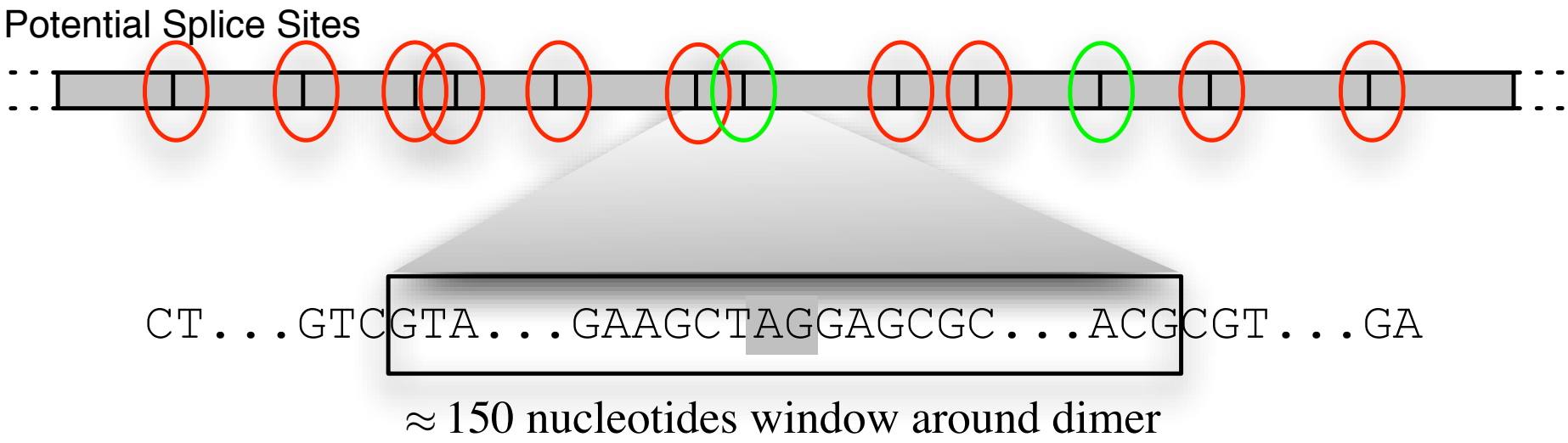
Example: Splice Site Recognition



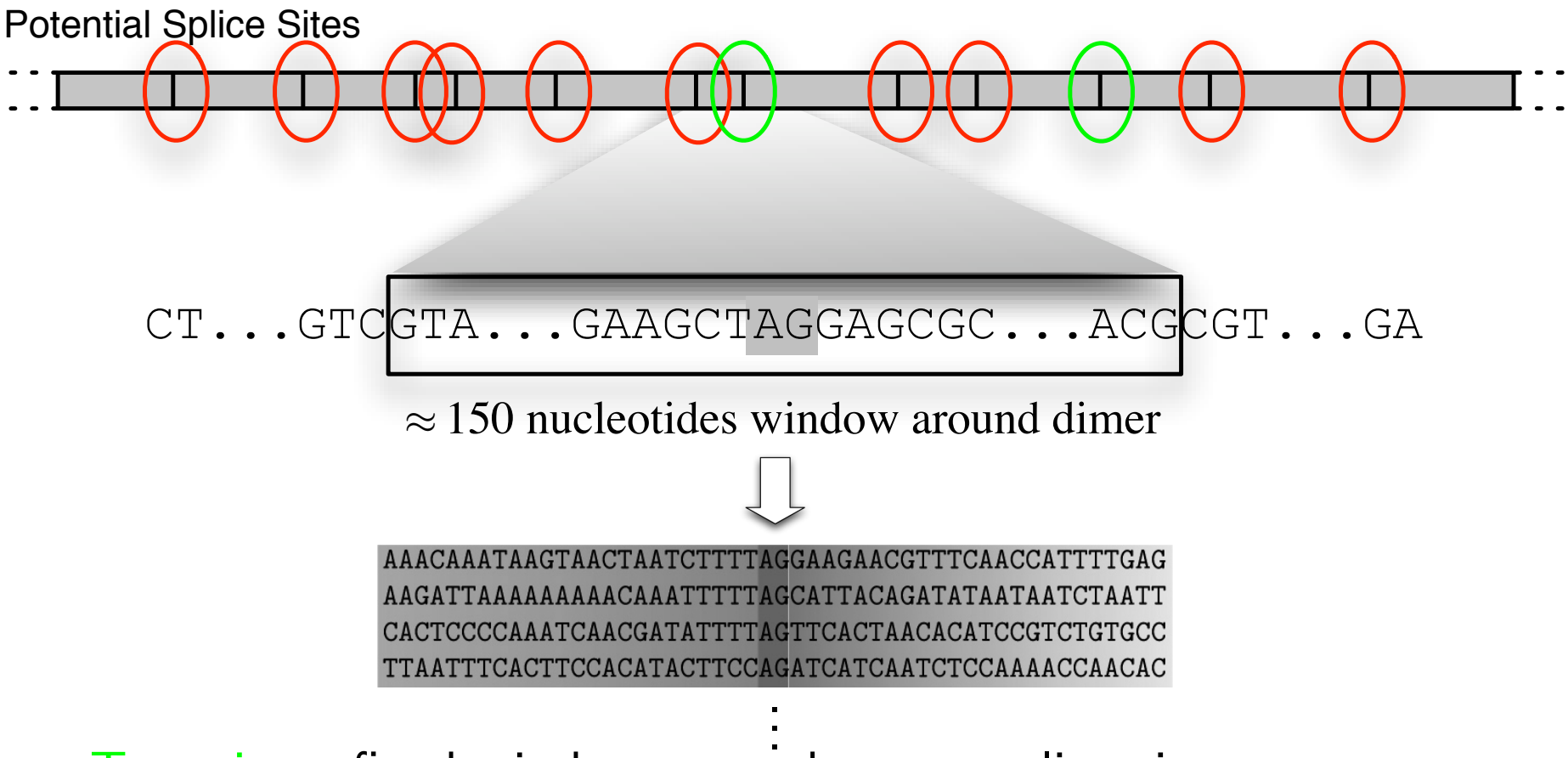
Example: Splice Site Recognition



Example: Splice Site Recognition



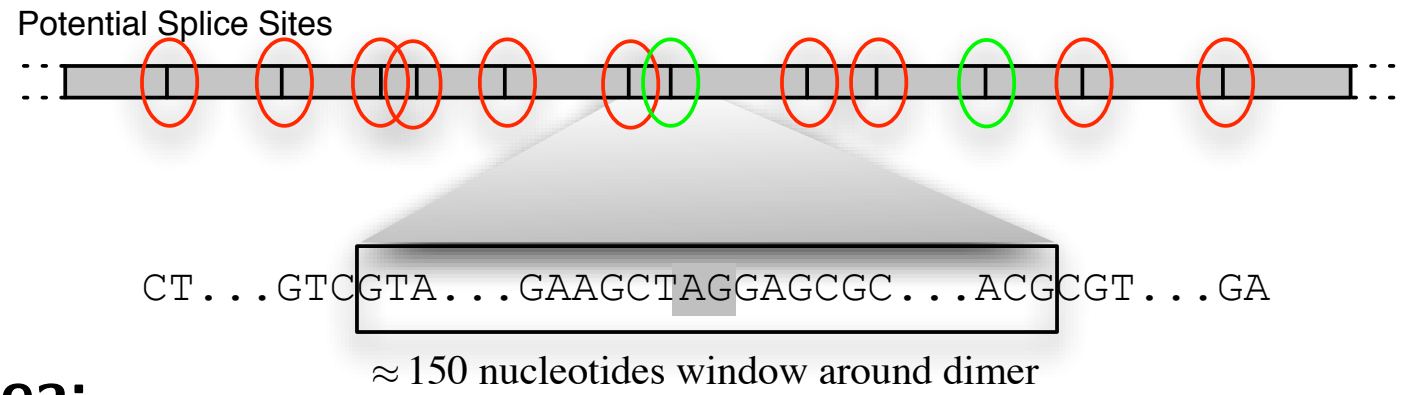
Example: Splice Site Recognition



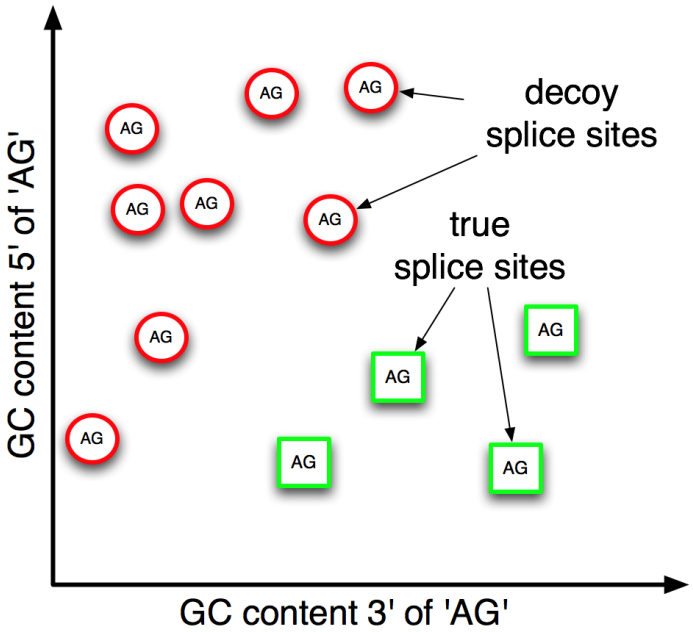
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

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

Example: Splice Site Recognition



Basic idea:



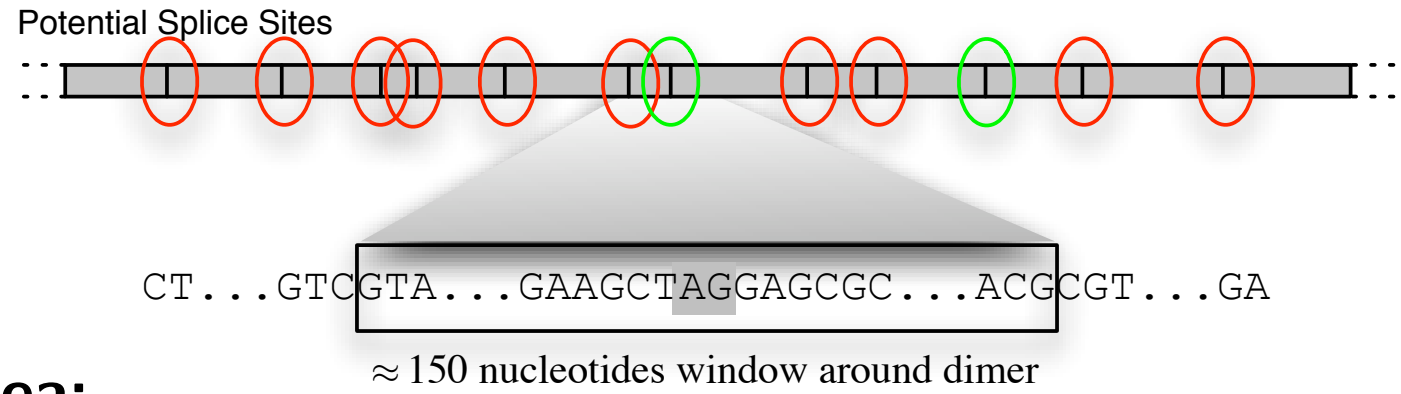
For instance, exploit that exons have higher GC content

or

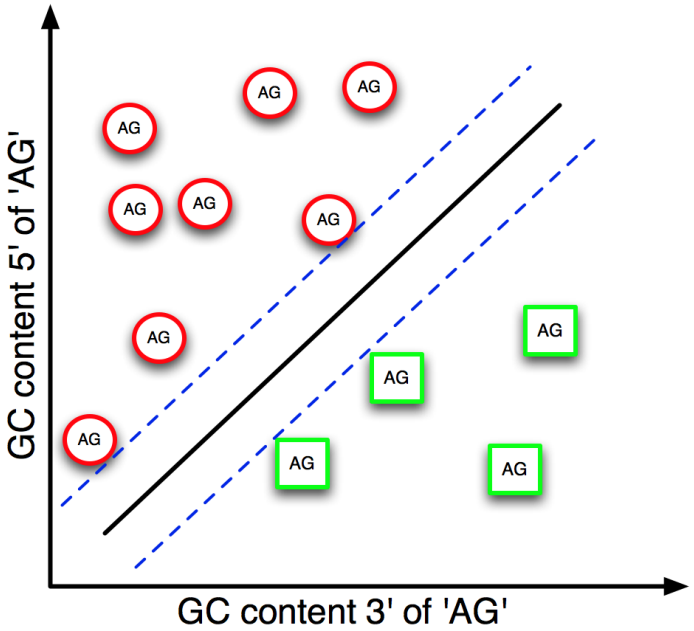
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

Example: Splice Site Recognition



Basic idea:

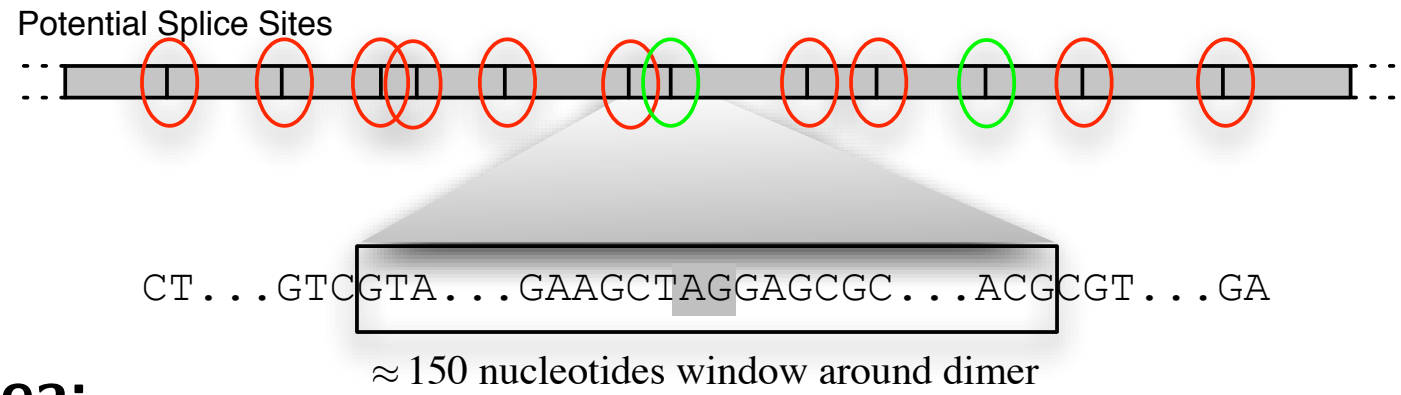


In practice: Use one feature per possible substring (e.g. ≤ 20) at all positions

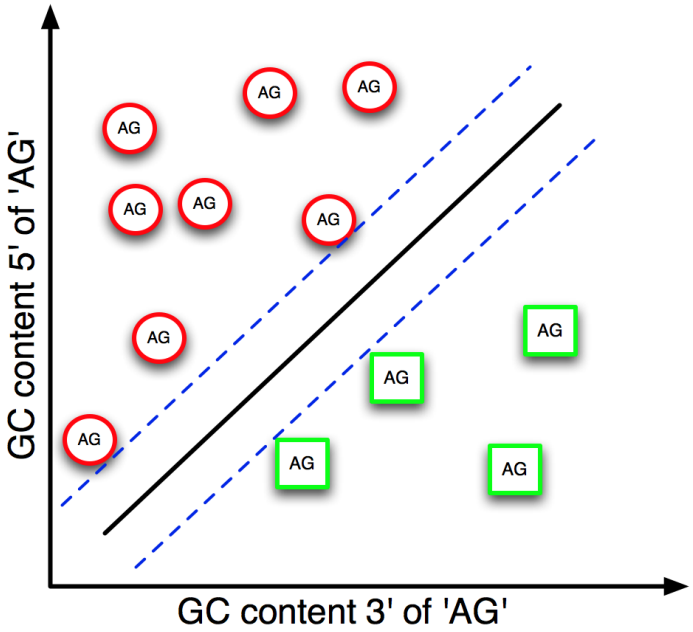
$150 \cdot (4^1 + \ldots + 4^{20}) \approx 2 \cdot 10^{14}$ features

- Needs efficient algorithms
- Leads to most accurate predictors that currently exist

Example: Splice Site Recognition



Basic idea:

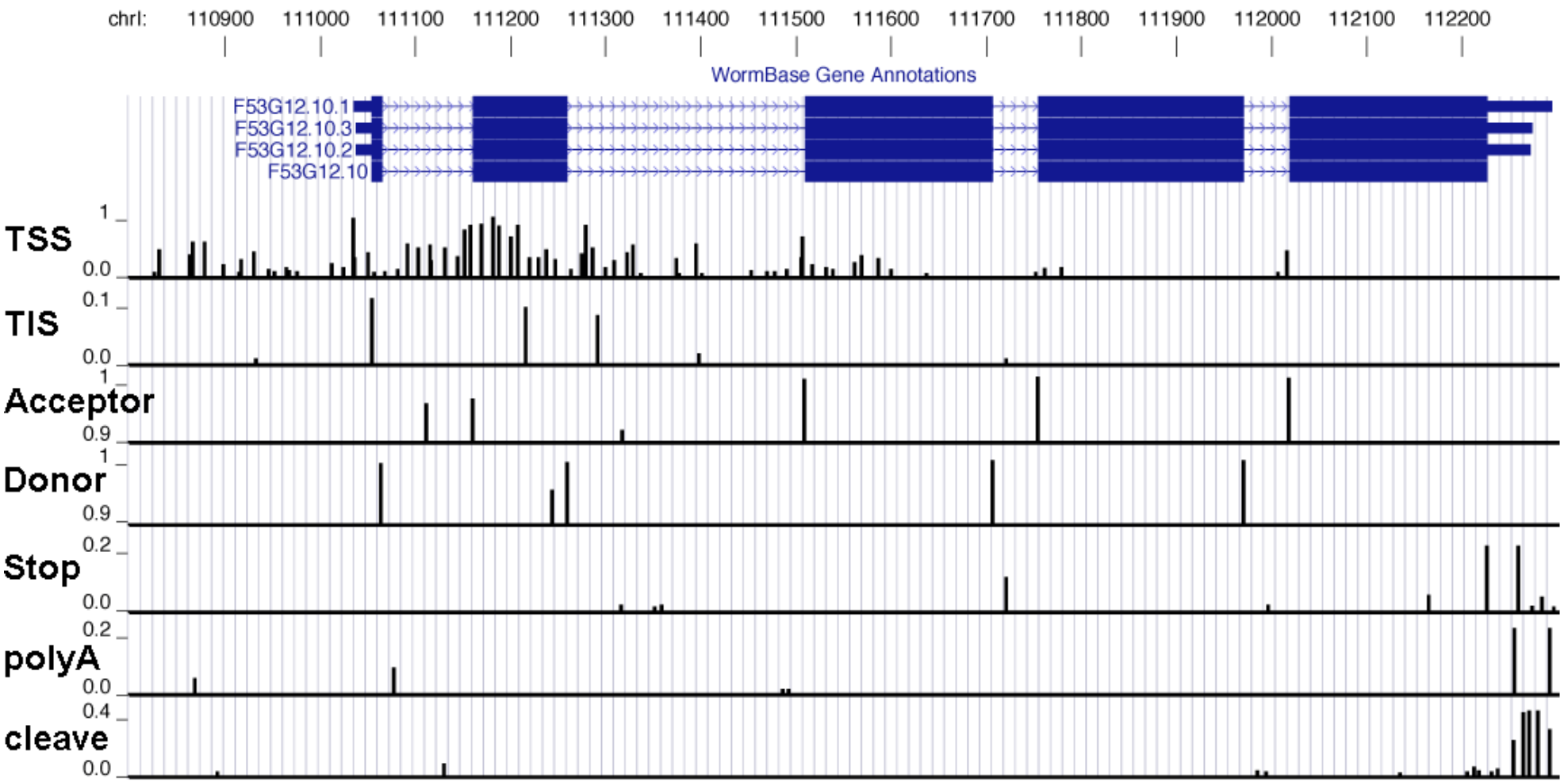


In practice: Use one feature per possible substring (e.g. ≤ 20) at all positions

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

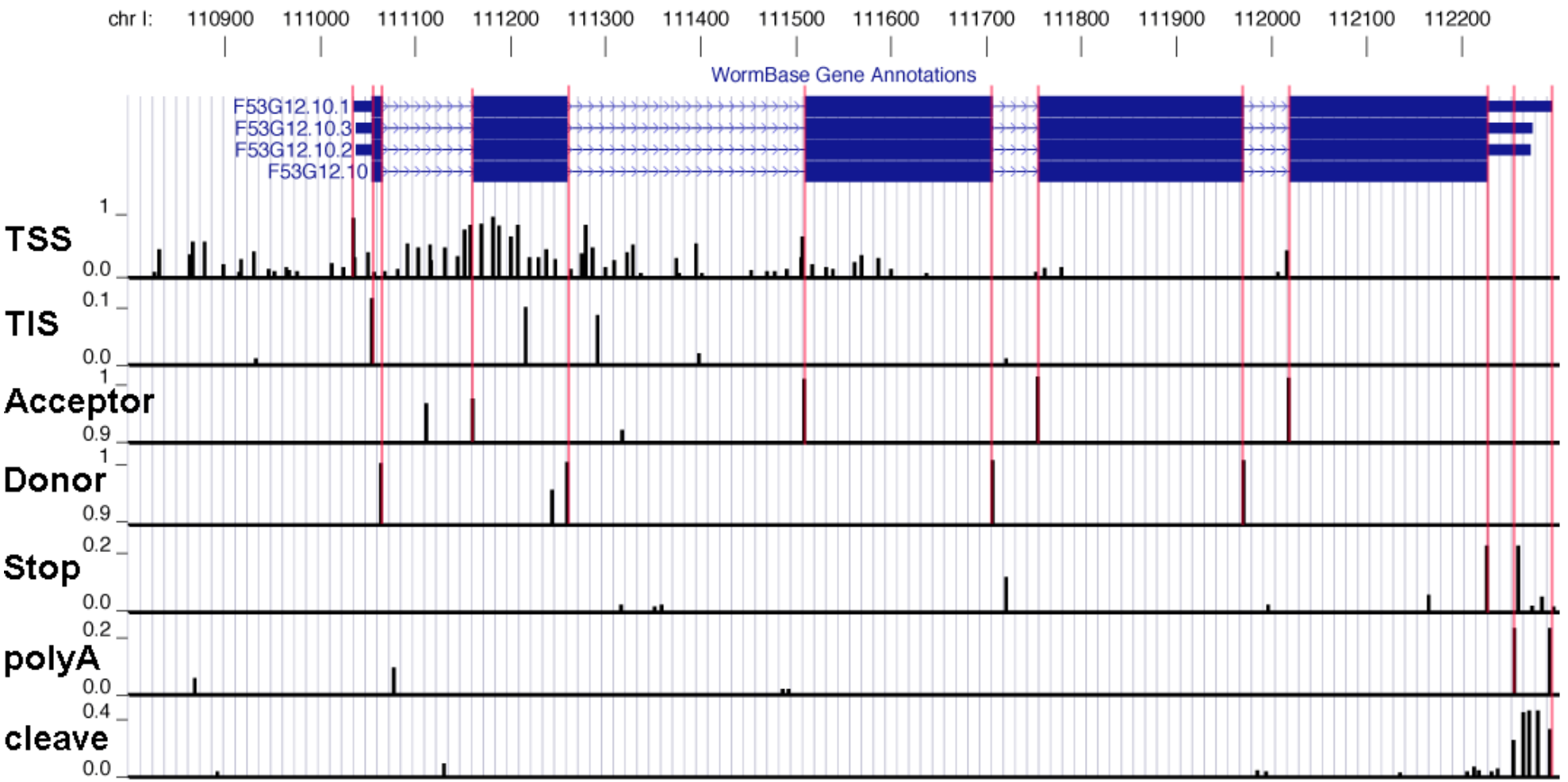
- Needs efficient algorithms
- Leads to most accurate predictors that currently exist

Example: Predictions in UCSC Browser



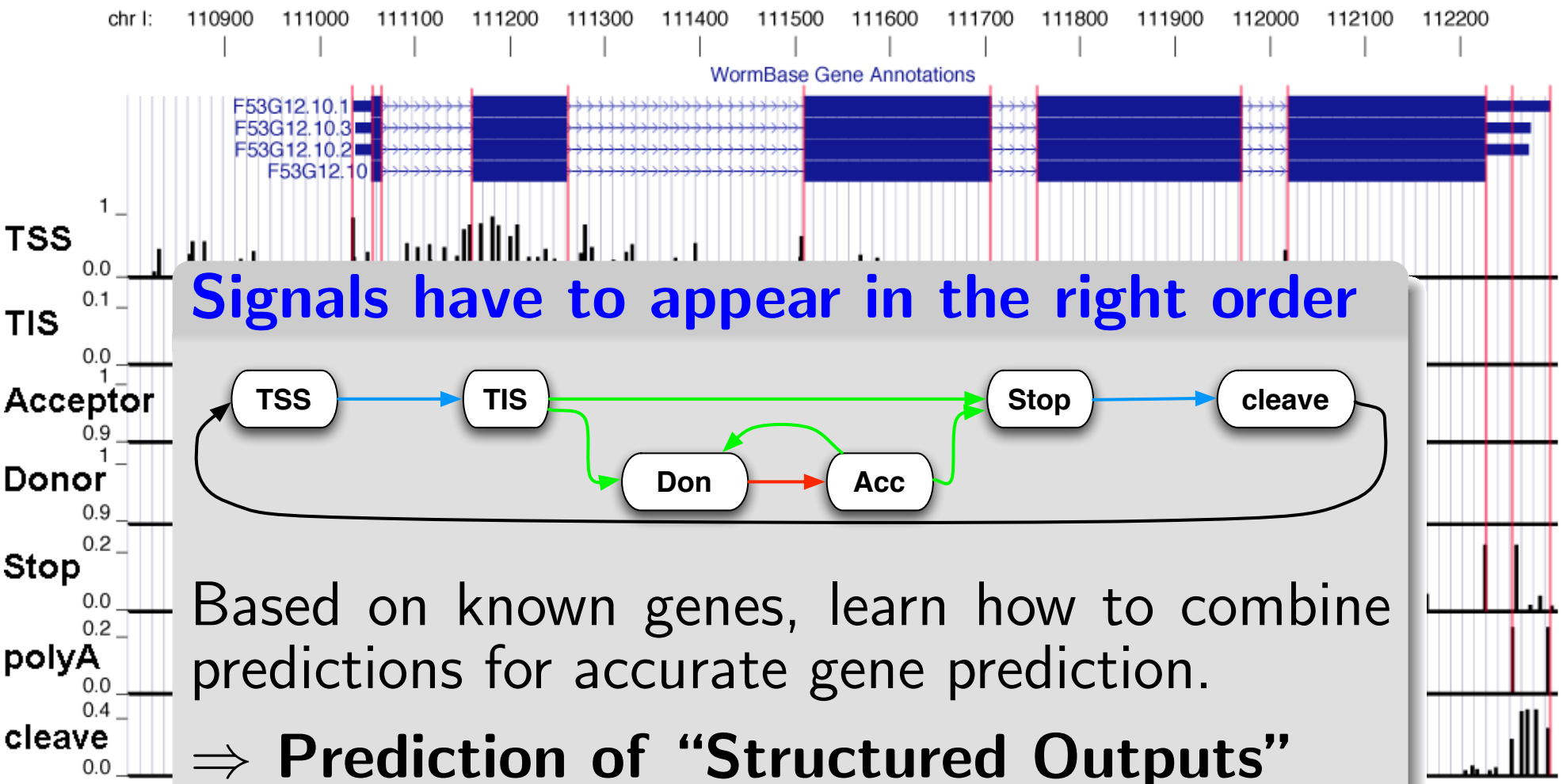
[Rätsch et al., 2007, Schweikert et al., 2009b]

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.

⇒ **Prediction of "Structured Outputs"**

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

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]

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
  - Current methods also ignore other important information:
    - Chromatin structure and methylation patterns
    - RNA structure and processing regulation ...
  - *Ab initio* methods cannot predict alternative transcripts

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
  - Current methods also ignore other important information:
    - Chromatin structure and methylation patterns
    - RNA structure and processing regulation ...
  - *Ab initio* methods cannot predict alternative transcripts

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
  - Current methods also ignore other important information:
    - Chromatin structure and methylation patterns
    - RNA structure and processing regulation ...
  - *Ab initio* methods cannot predict alternative transcripts

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
  - Current methods also ignore other important information:
    - Chromatin structure and methylation patterns
    - RNA structure and processing regulation ...
  - *Ab initio* methods cannot predict alternative transcripts

Gene Finding + Auxiliary Measurements

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



Is there other information that may be used by cellular processes that can improve prediction results?

Preliminary study: (*A. thaliana*) transcript level $(SN + SP)/2$

1. mGene (*ab initio*) . . . **73.3%**
2. . . . with DNA methylation (1 tissue) **76.1%**
3. . . . with Nucleosome position predictions **78.0%**
4. . . . with RNA secondary structure predictions **76.7%**

In progress:

- Study effect of other information sources for human gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Measurements

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



Is there other information that may be used by cellular processes that can improve prediction results?

Preliminary study: (*A. thaliana*) transcript level $(SN + SP)/2$

1. mGene (*ab initio*) . . . **73.3%**
2. . . . with DNA methylation (1 tissue) **76.1%**
3. . . . with Nucleosome position predictions **78.0%**
4. . . . with RNA secondary structure predictions **76.7%**

In progress:

- Study effect of other information sources for human gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Measurements

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



Is there other information that may be used by cellular processes that can improve prediction results?

Preliminary study: (*A. thaliana*) transcript level $(SN + SP)/2$

1. mGene (*ab initio*) . . . **73.3%**
2. . . . with DNA methylation (1 tissue) **76.1%**
3. . . . with Nucleosome position predictions **78.0%**
4. . . . with RNA secondary structure predictions **76.7%**

In progress:

- Study effect of other information sources for human gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Measurements

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

Is there other information that may be used by cellular processes that can improve prediction results?

Preliminary study: (*A. thaliana*) transcript level $(SN + SP)/2$

1. mGene (*ab initio*) . . . **73.3%**
2. . . . with DNA methylation (1 tissue) **76.1%**
3. . . . with Nucleosome position predictions **78.0%**
4. . . . with RNA secondary structure predictions **76.7%**

In progress:

- Study effect of other information sources for human gene prediction
- Ideally, consider measurements from the same sample

Gene Finding + Auxiliary Measurements

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



Is there other information that may be used by cellular processes that can improve prediction results?

Preliminary study: (*A. thaliana*) transcript level $(SN + SP)/2$

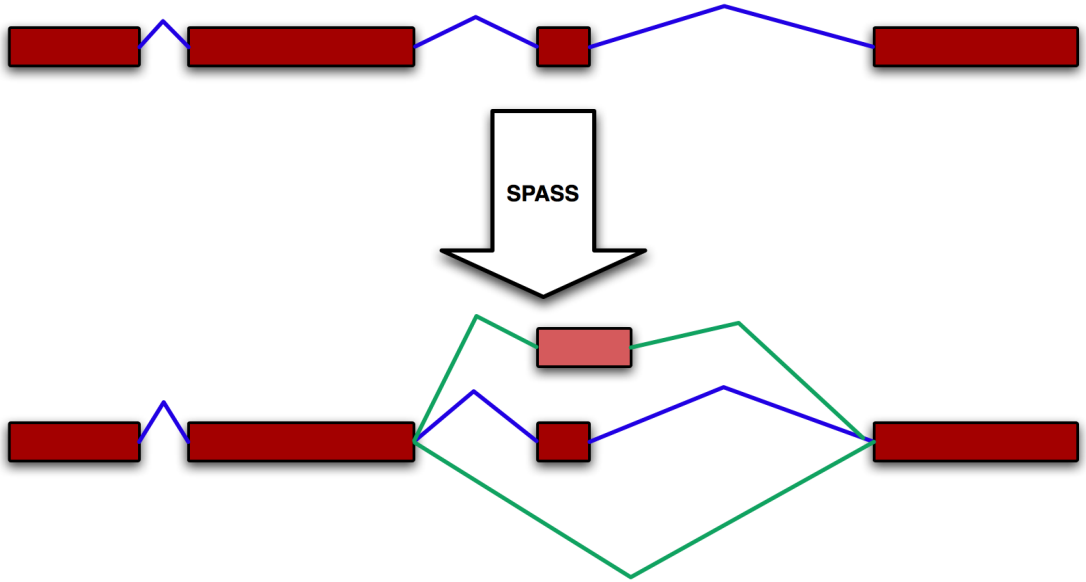
1. mGene (*ab initio*) . . . **73.3%**
2. . . . with DNA methylation (1 tissue) **76.1%**
3. . . . with Nucleosome position predictions **78.0%**
4. . . . with RNA secondary structure predictions **76.7%**

In progress:

- Study effect of other information sources for human gene prediction
- Ideally, consider measurements from the same sample

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

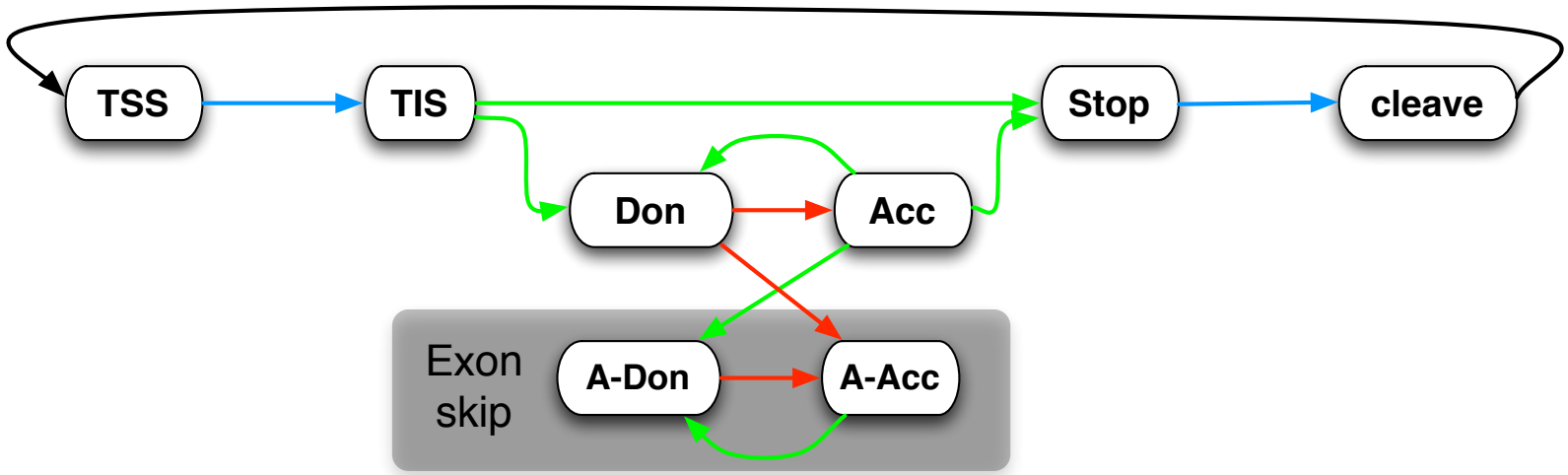
    Output: Splicing graph

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

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

Alternative Transcripts: More Steps

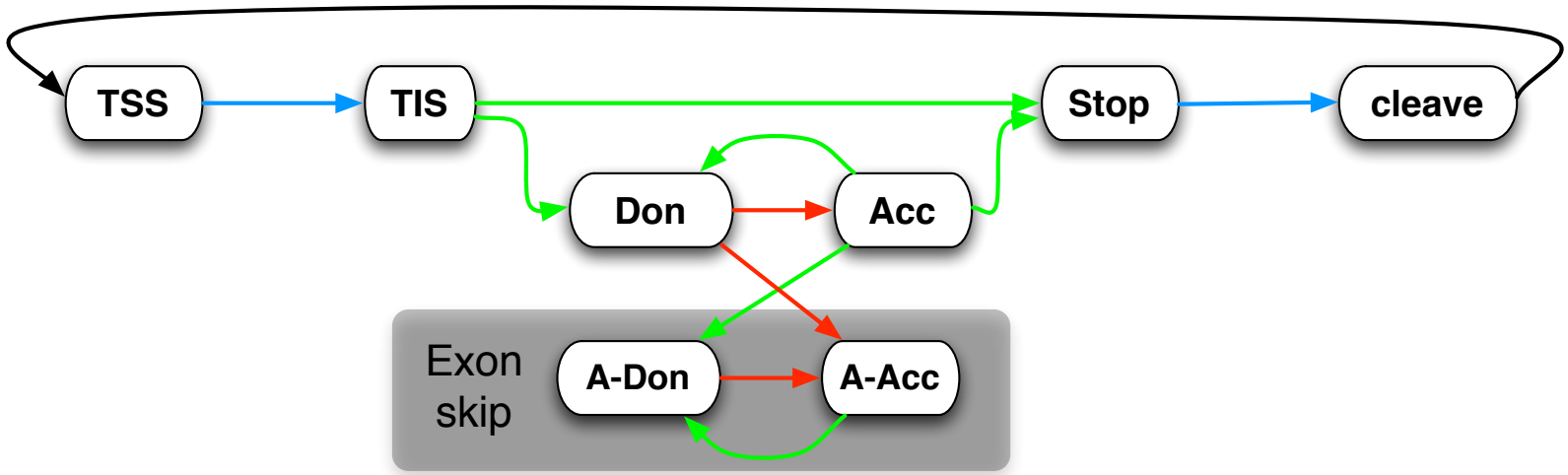
- Combine gene finding with prediction of alternative splicing
  - Machine learning challenge:
    Input: DNA sequence
    Output: Splicing graph
- mGene approach can be extended to
  - Predict **"simple splicing graphs"**



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

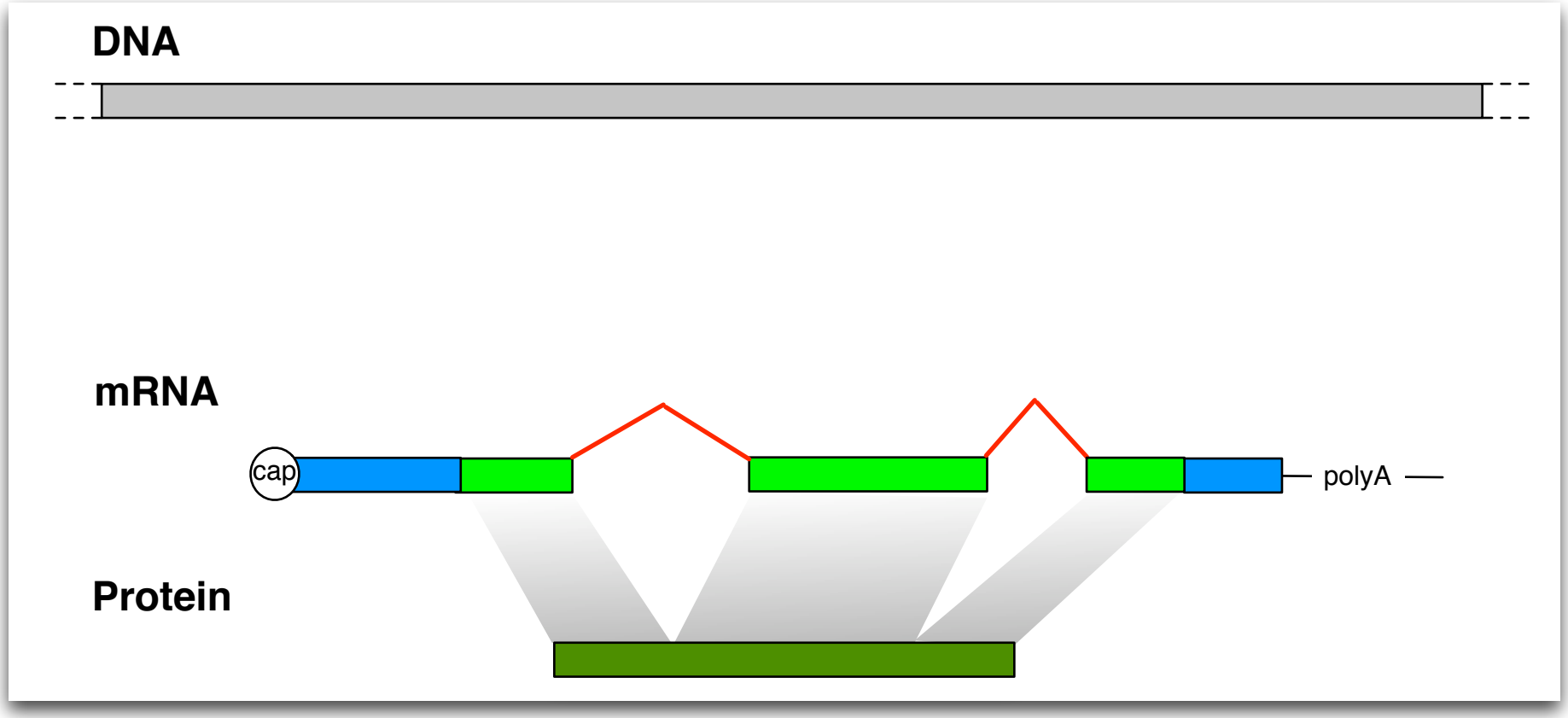
Alternative Transcripts: More Steps

- 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 ...

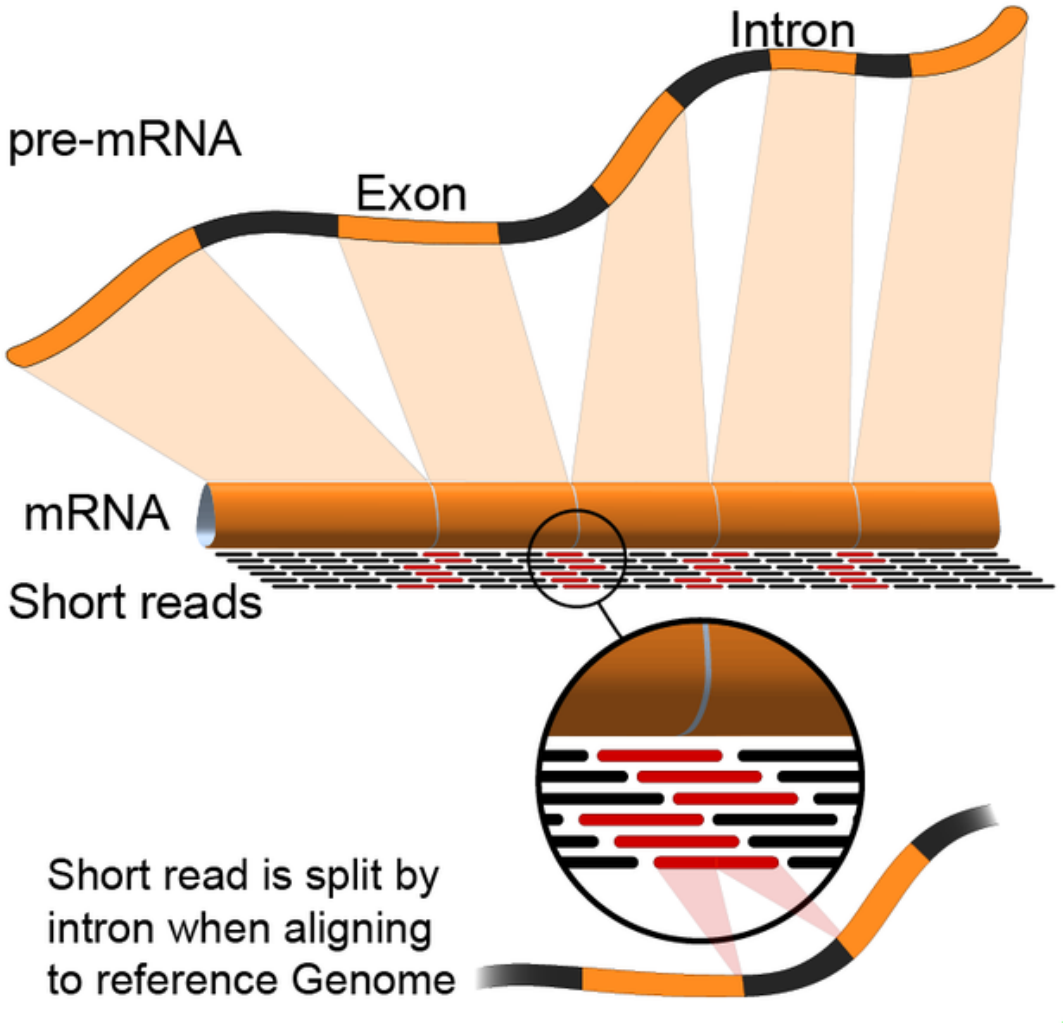
Reconstructing the Transcriptome



Directly measure the transcriptome

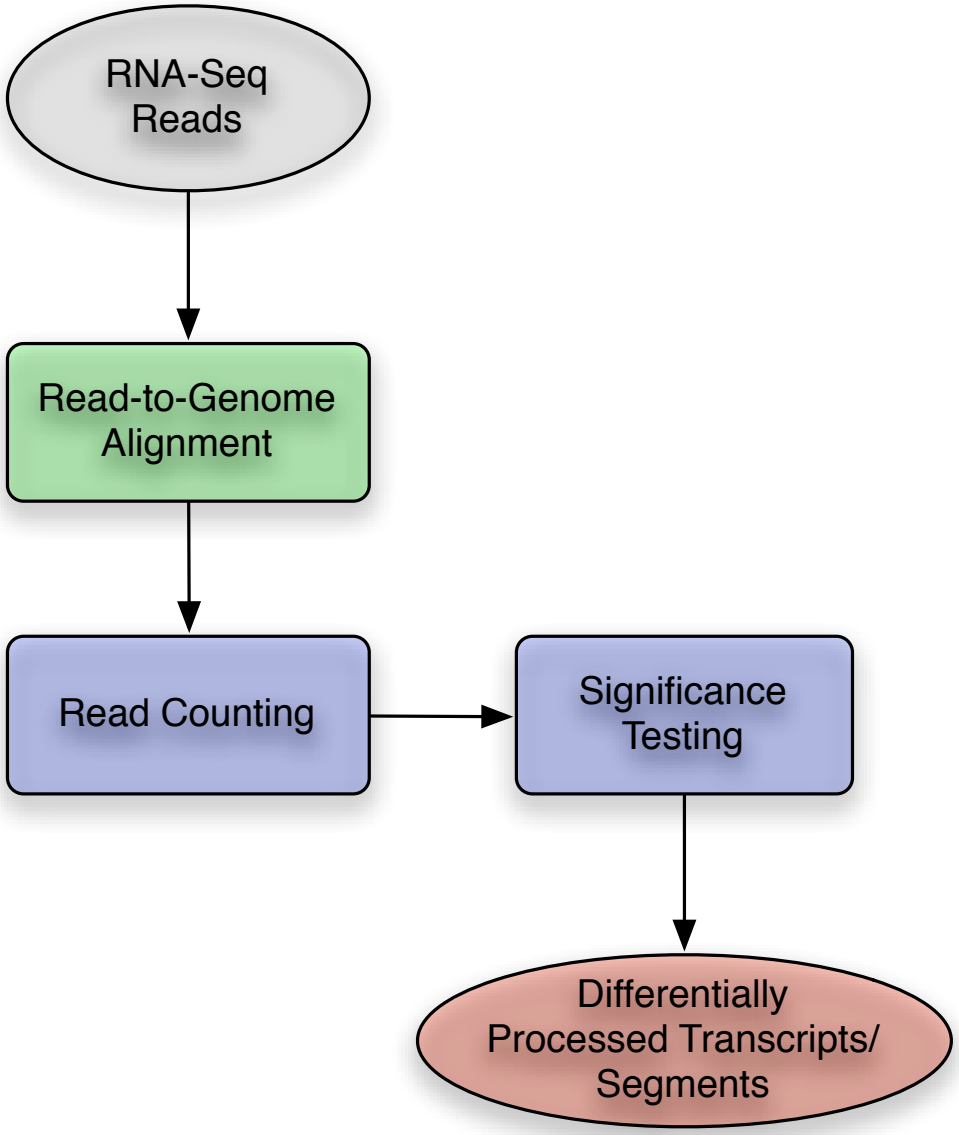
- Whole genome tiling arrays
- **Transcriptome sequencing** (Sanger or Next Generation Sequencing)

mRNA Deep Sequencing

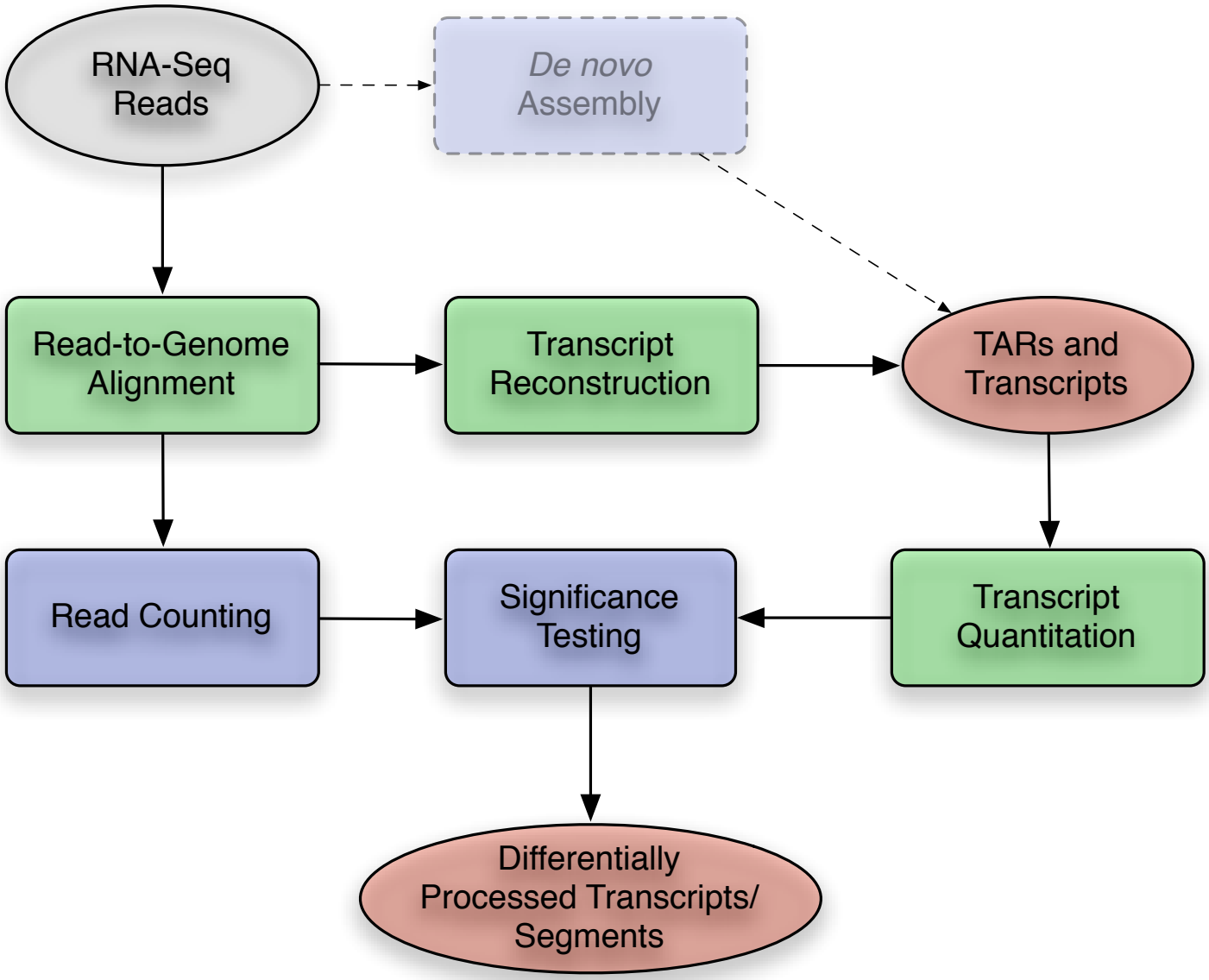


[Wikipedia]

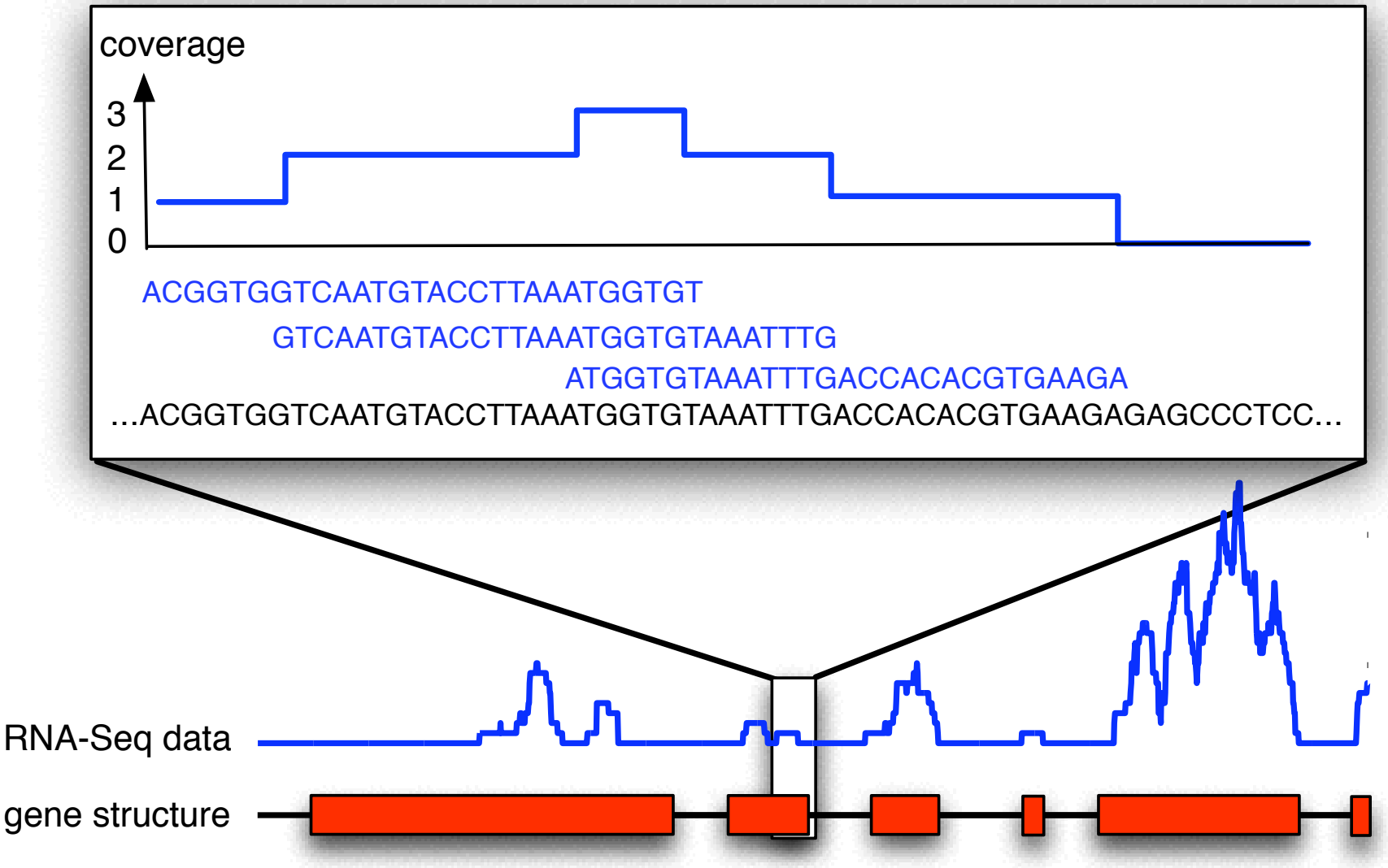
RNA-Seq Analysis Steps



RNA-Seq Analysis Steps

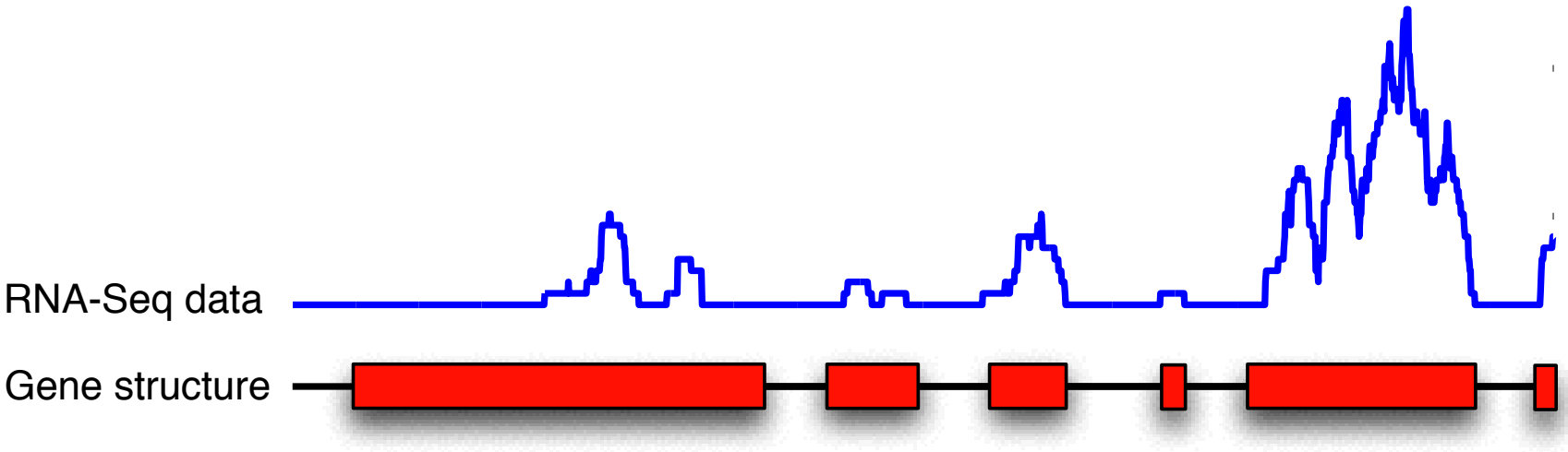


RNA-Seq Read Alignment



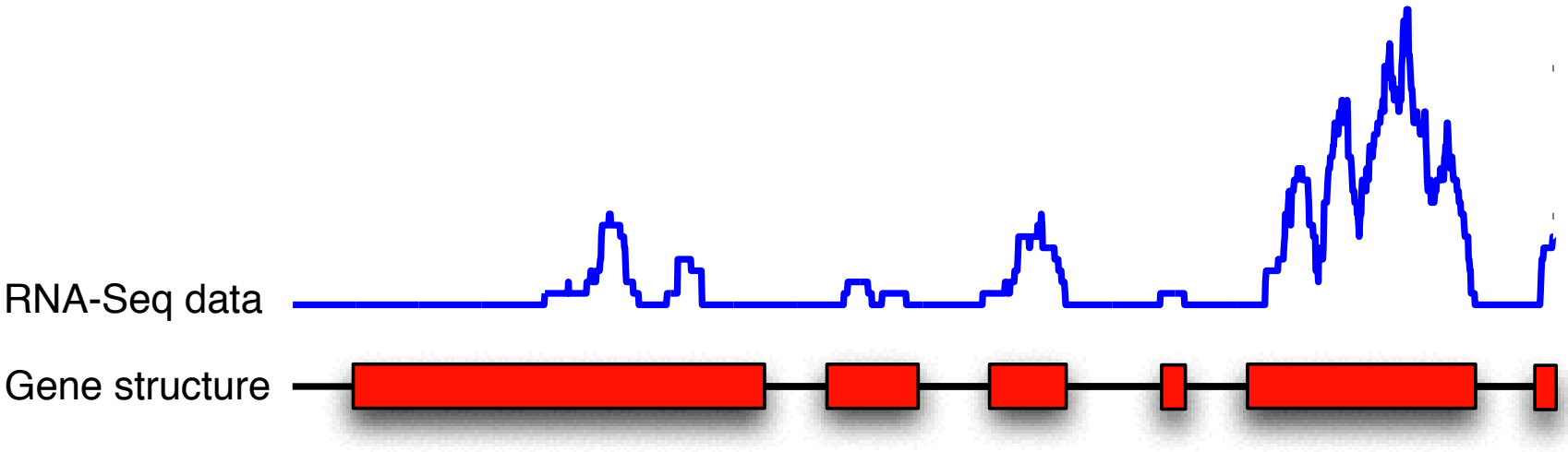
RNA-Seq Read Alignment

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

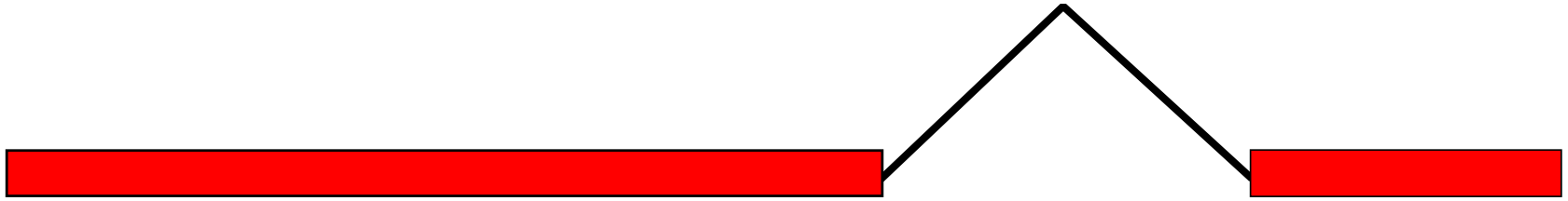


RNA-Seq Read Alignment

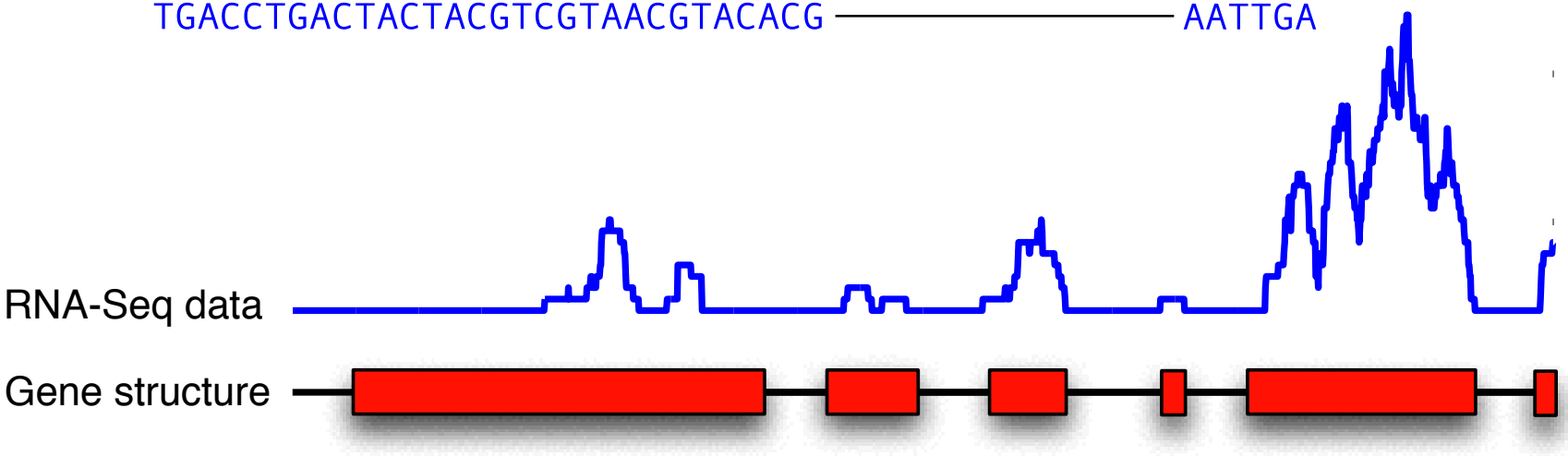
```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
```



RNA-Seq Read Alignment



...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
AACCAGTGACCTGACTACTACGTCGTAACGTACACG
ACCAGTGACCTGACTACTACGTCGTAACGTACACG A
CCAGTGACCTGACTACTACGTCGTAACGTACACG AA
CAGTGACCTGACTACTACGTCGTAACGTACACG AAT
AGTGACCTGACTACTACGTCGTAACGTACACG AATT
GTGACCTGACTACTACGTCGTAACGTACACG AATTG
TGACCTGACTACTACGTCGTAACGTACACG AATTGA



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

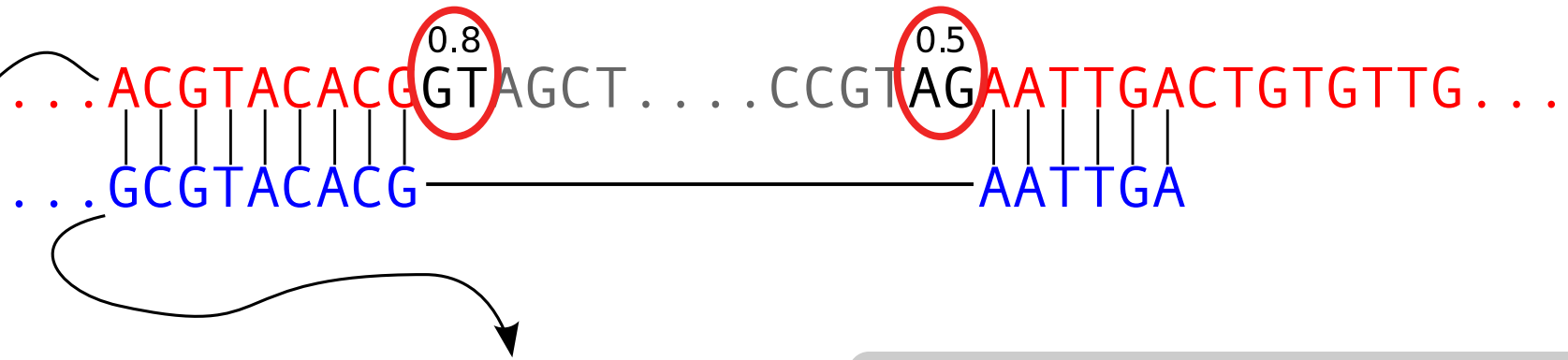
Source of Information

- 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

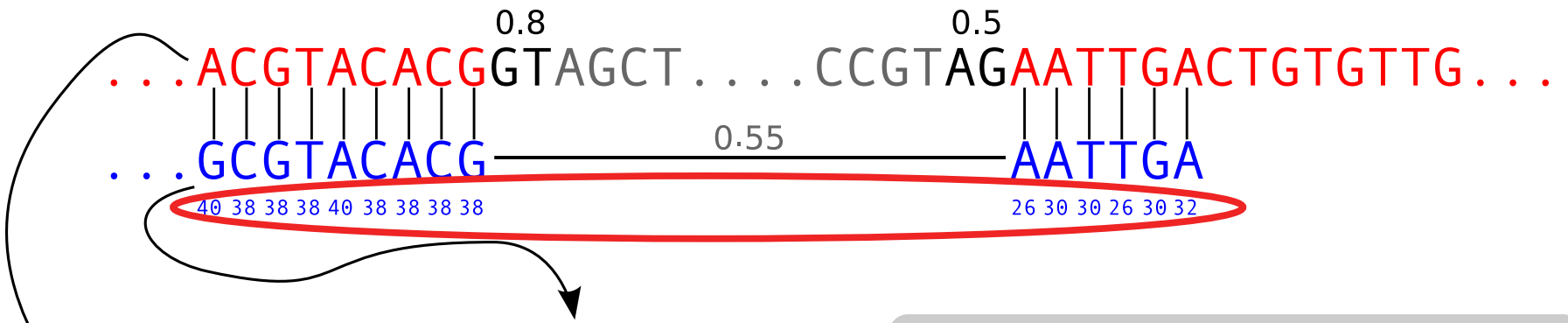
Source of Information

- 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

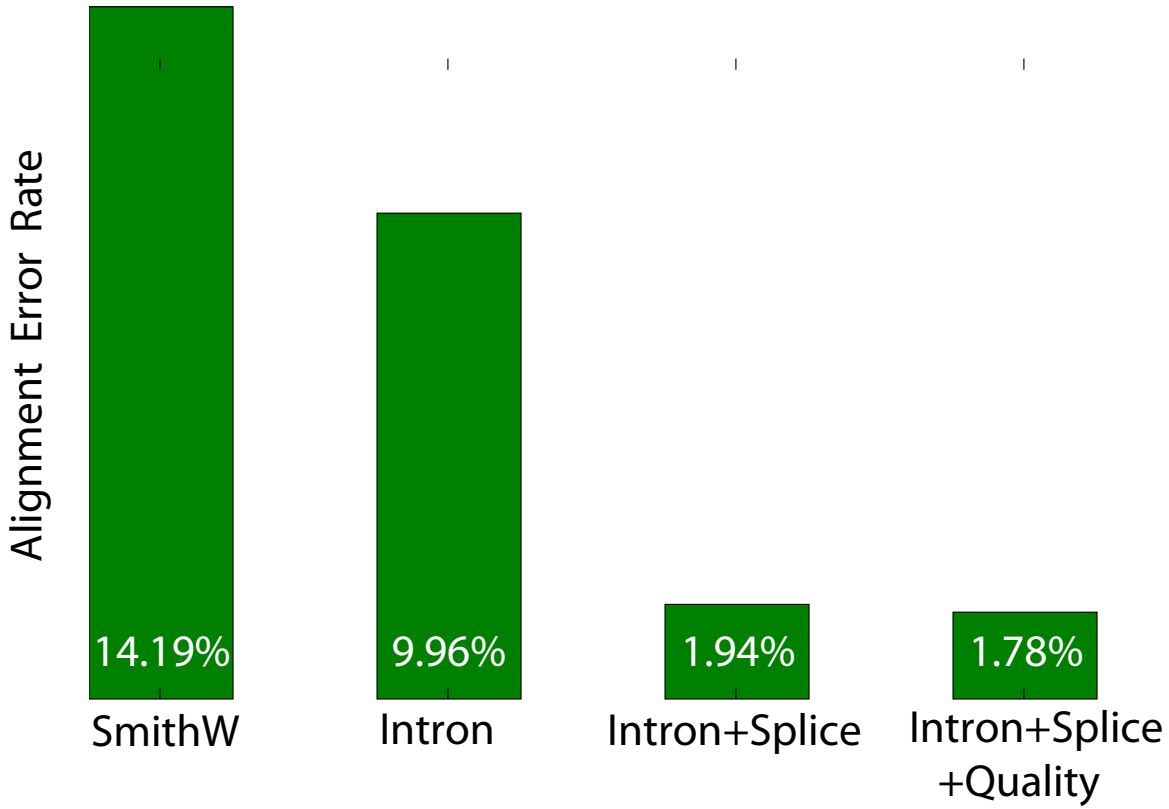
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

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

QPalma's Accurate Alignments

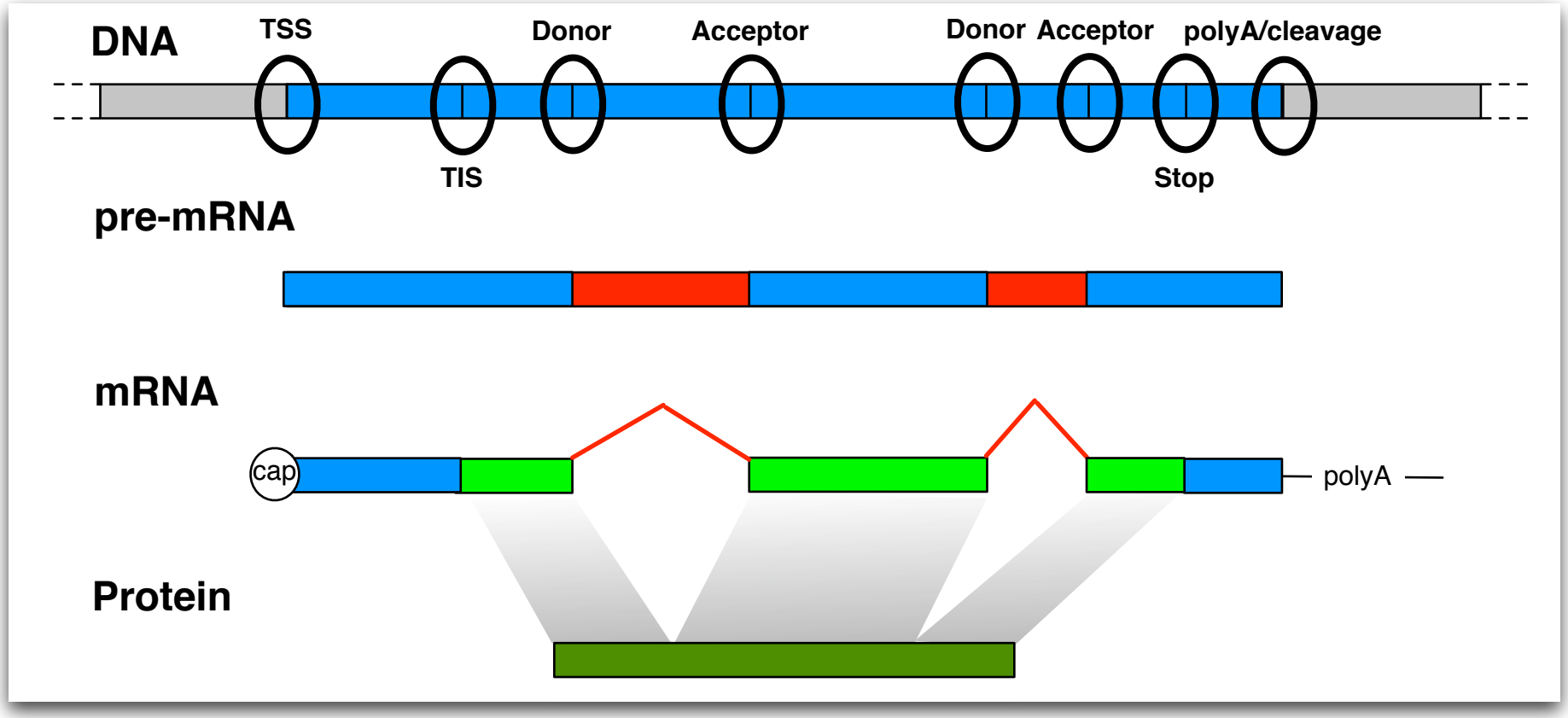
Generate set of artificially spliced reads

- 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

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%**

Transcriptome measurements certainly help, but predictions are still quite far from perfect.

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%**

Transcriptome measurements certainly help, but predictions are still quite far from perfect.

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%**

Transcriptome measurements certainly help, but predictions are still quite far from perfect.

Gene Finding + Transcriptome Measurements

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



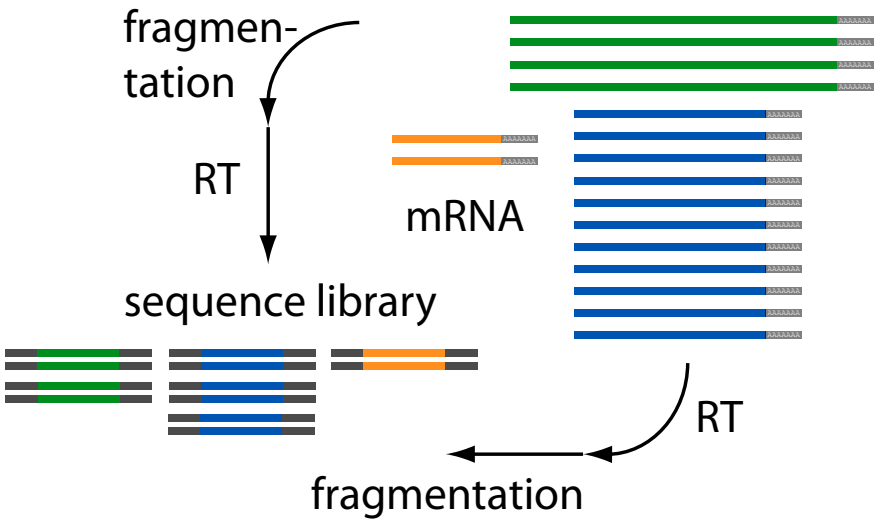
Idea: Use transcriptome measurements to enhance recognition of exonic regions.

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%**

Transcriptome measurements certainly help, but predictions are still quite far from perfect.

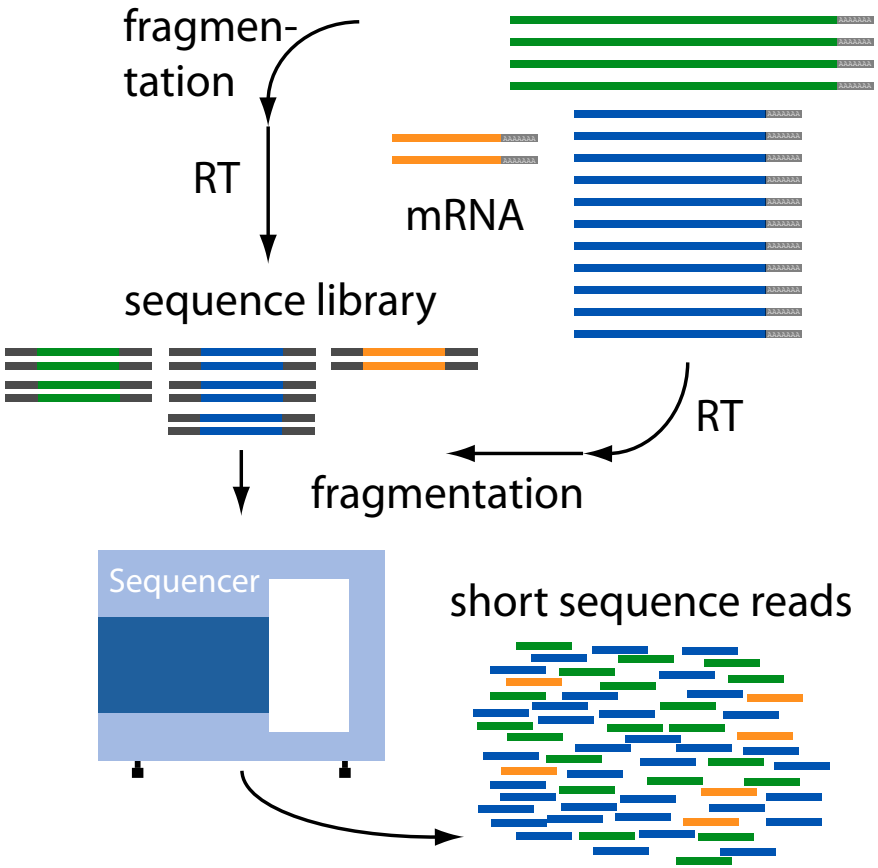
RNA Quantification and Biases



Biases due to . . .

- cDNA library construction
- Sequencing
- Read mapping

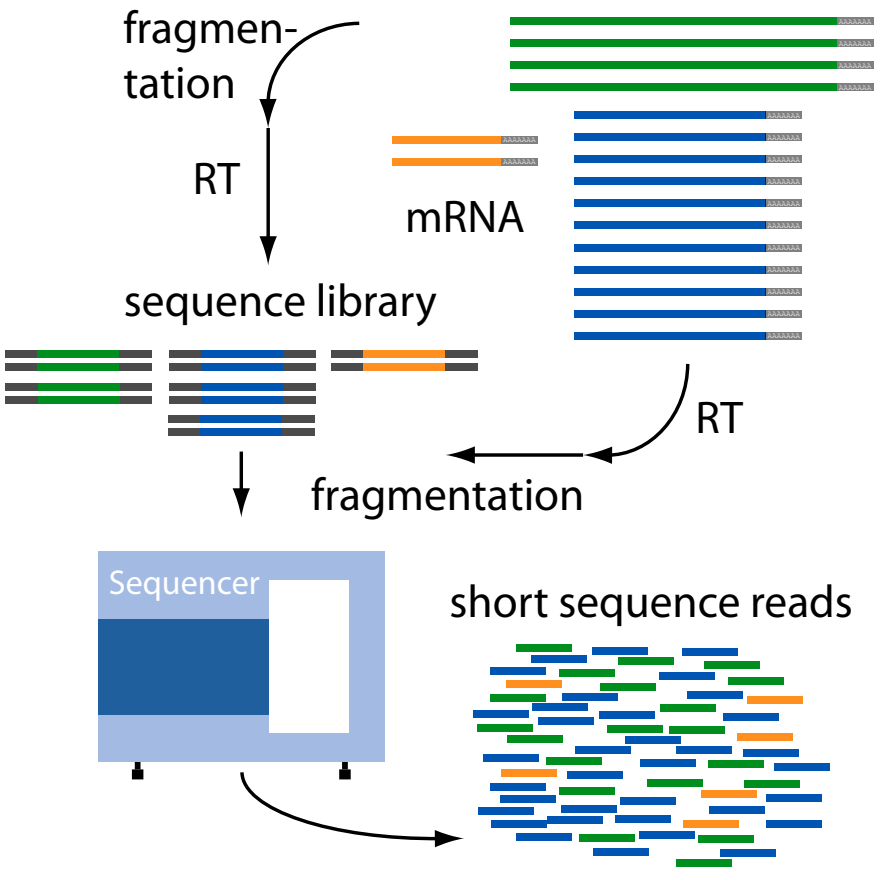
RNA Quantification and Biases



Biases due to . . .

- cDNA library construction
- Sequencing
- Read mapping

RNA Quantification and Biases



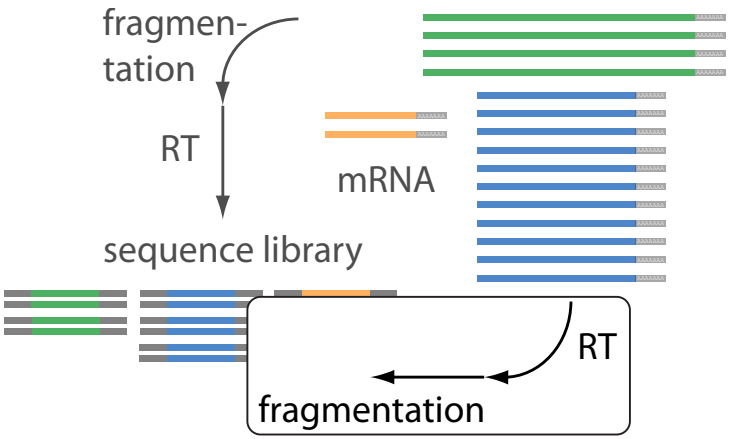
Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping

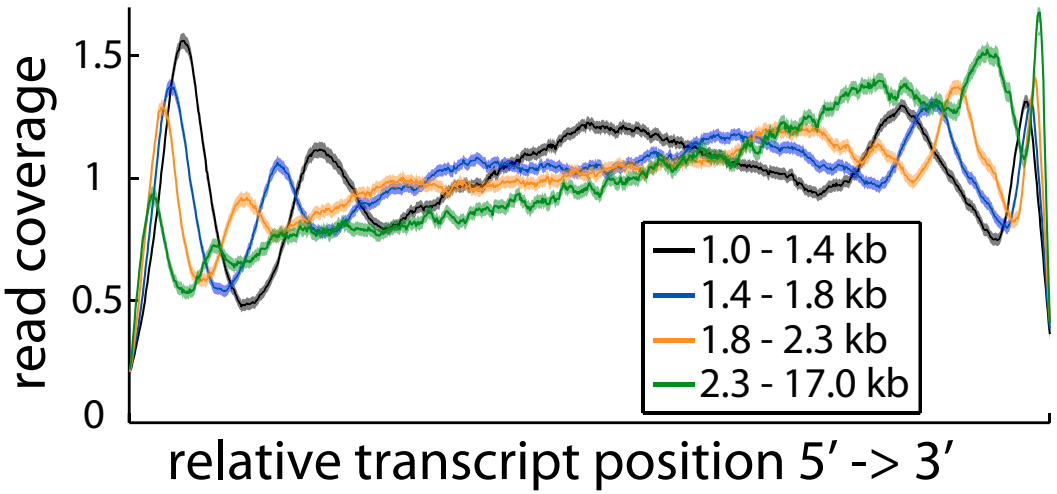


Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript



Transcript profiles for different lengths

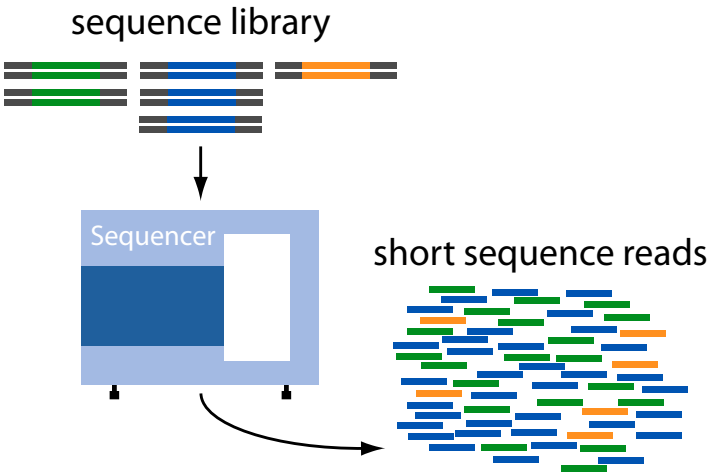


Data generated for *A. thaliana* with Flux Simulator v1.0 [Sammeth, 2009b]

- Oligo-dT priming
- Physical cDNA fragmentation

Sequencing Bias

PCR Bias



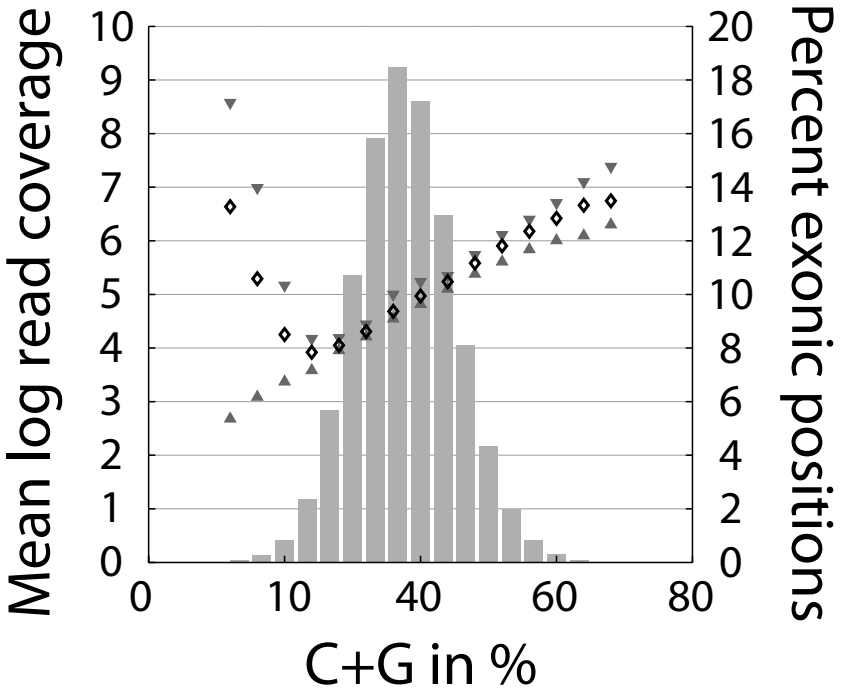


Figure provided by Georg Zeller

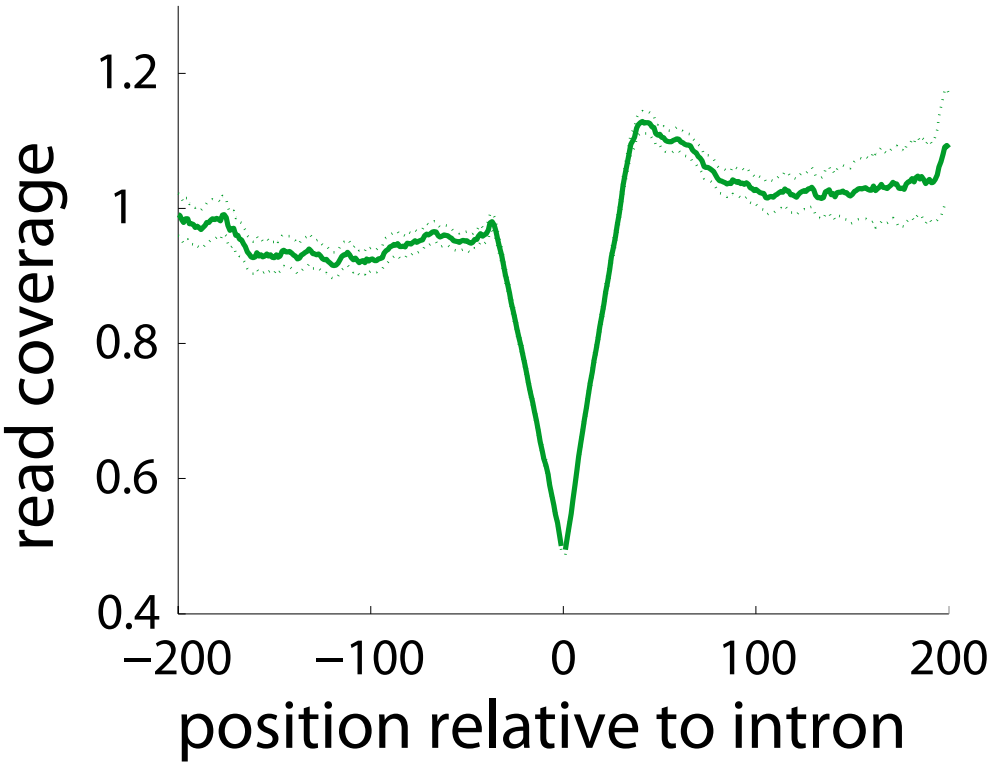
RNA-Seq data (*A. thaliana*)

- Exonic GC content

Mapping Bias

Read Mapping Bias

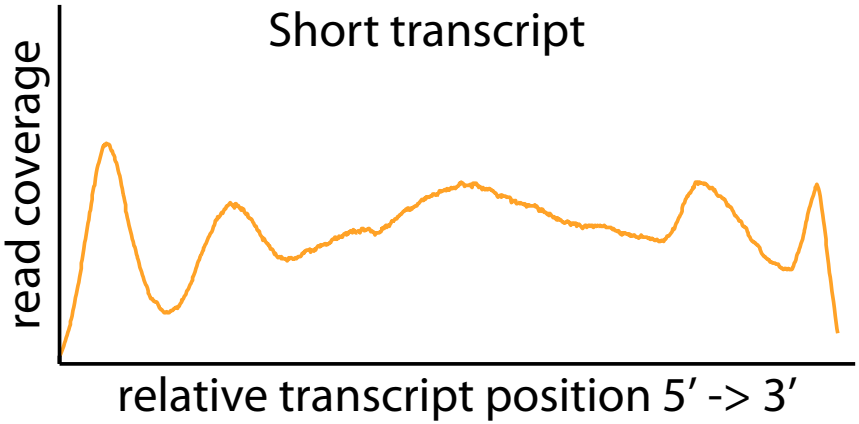




RNA-Seq data (*A. thaliana*)

- Exon boundaries

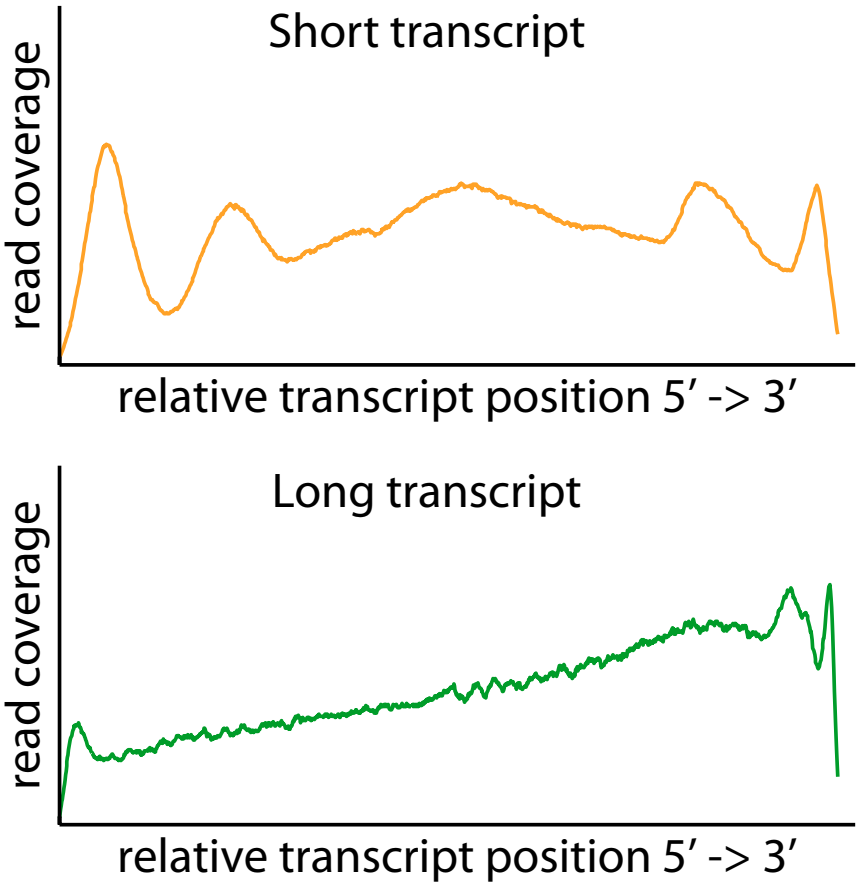
rQuant – Basic Idea





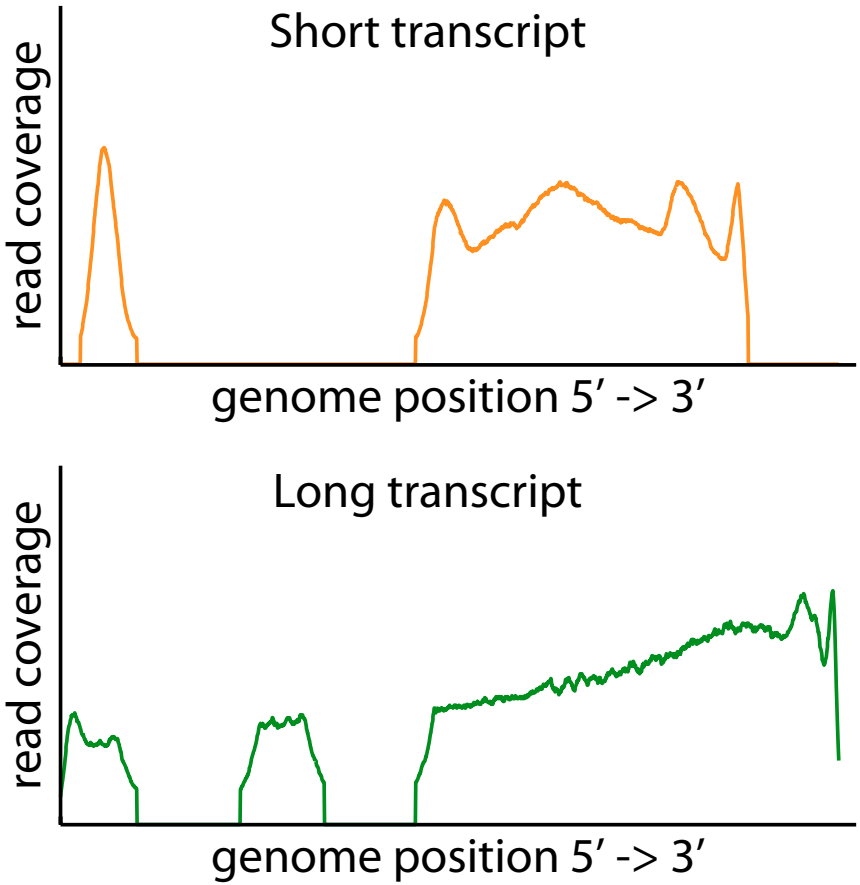
$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$

rQuant – Basic Idea



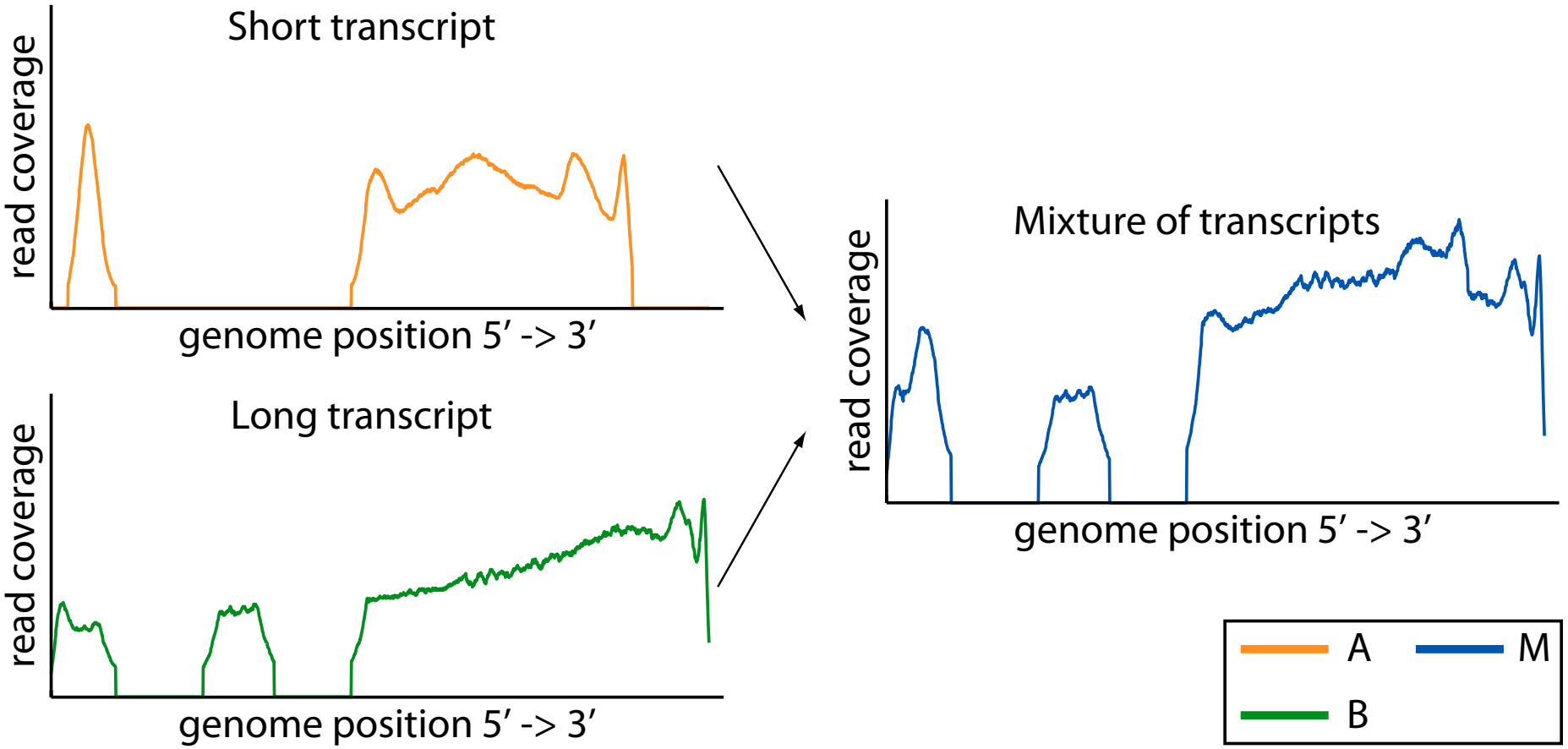
$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea



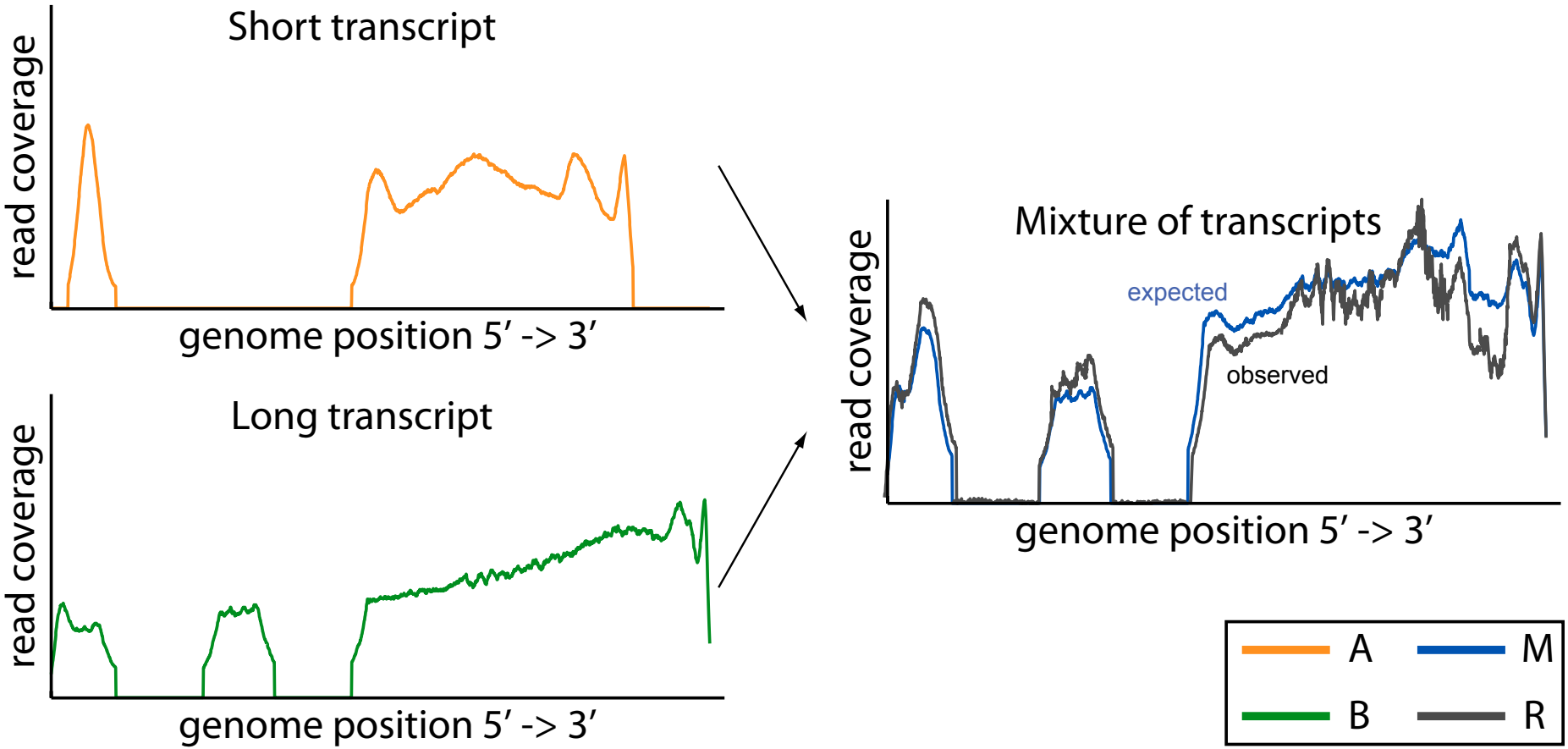
$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea



$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea



$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

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 transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - Allows us to train more sophisticated models predicting alternative transcripts
  - 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 transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - Allows us to train more sophisticated models predicting alternative transcripts
  - 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 transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - Allows us to train more sophisticated models predicting alternative transcripts
  - 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 transcripts
  - Quantification of transcript abundances
- Combination of predictions and measurements
  - Measurements help to predict more accurately
  - Allows us to train more sophisticated models predicting alternative transcripts
  - 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)

Machine Learning

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

Short Read Analysis

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

Machine Learning

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

Short Read Analysis

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

Machine Learning

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

Short Read Analysis

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

Machine Learning

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

Short Read Analysis

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

Machine Learning

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

Short Read Analysis

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

Machine Learning

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

Short Read Analysis

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

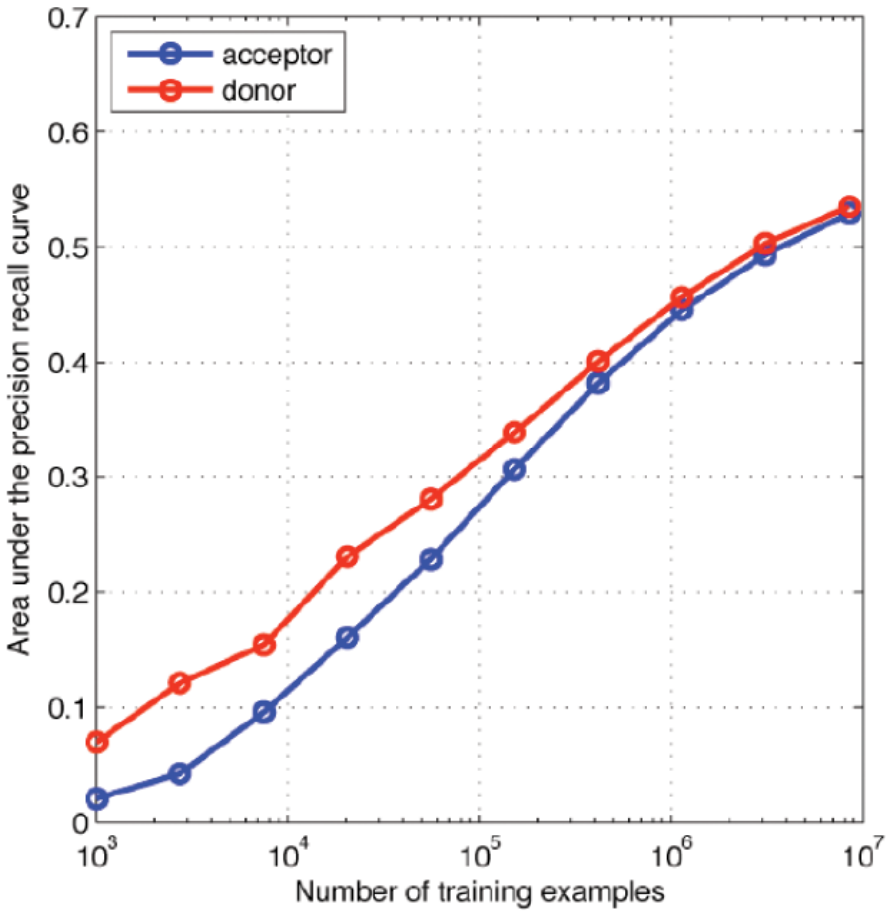
Nematode Material/Methods

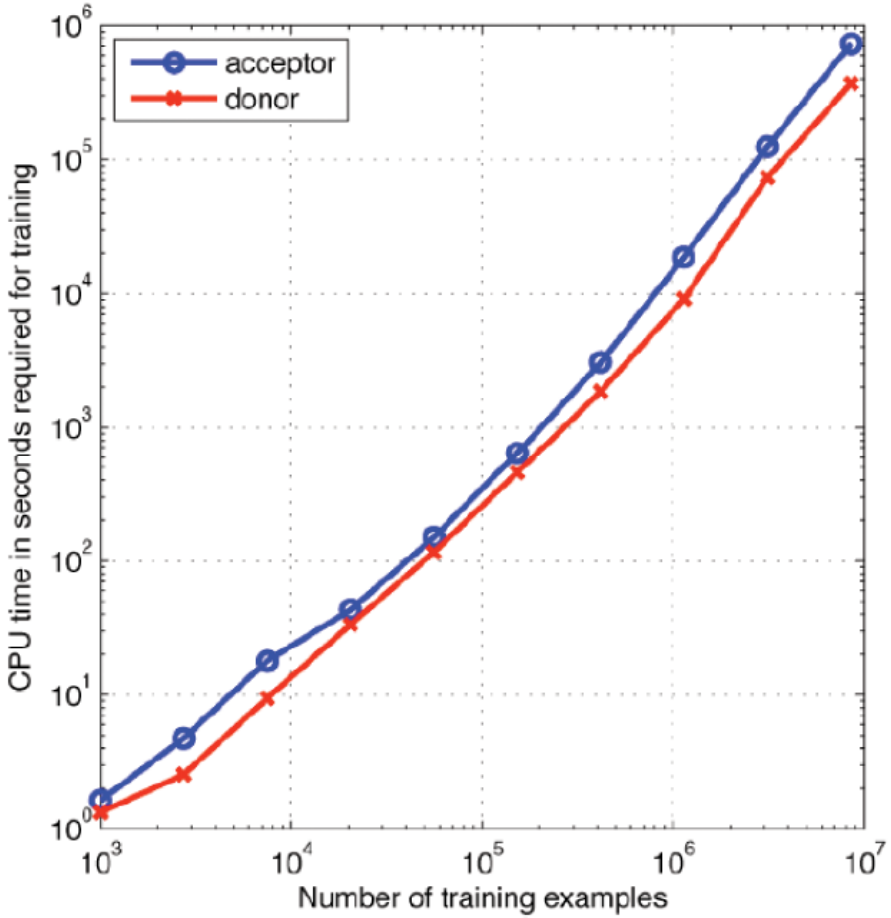
- Adrian Streit (MPI)

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

Thank you for your attention!

Human Splice Sites with WD Kernel
[Sonnenburg et al., 2007]

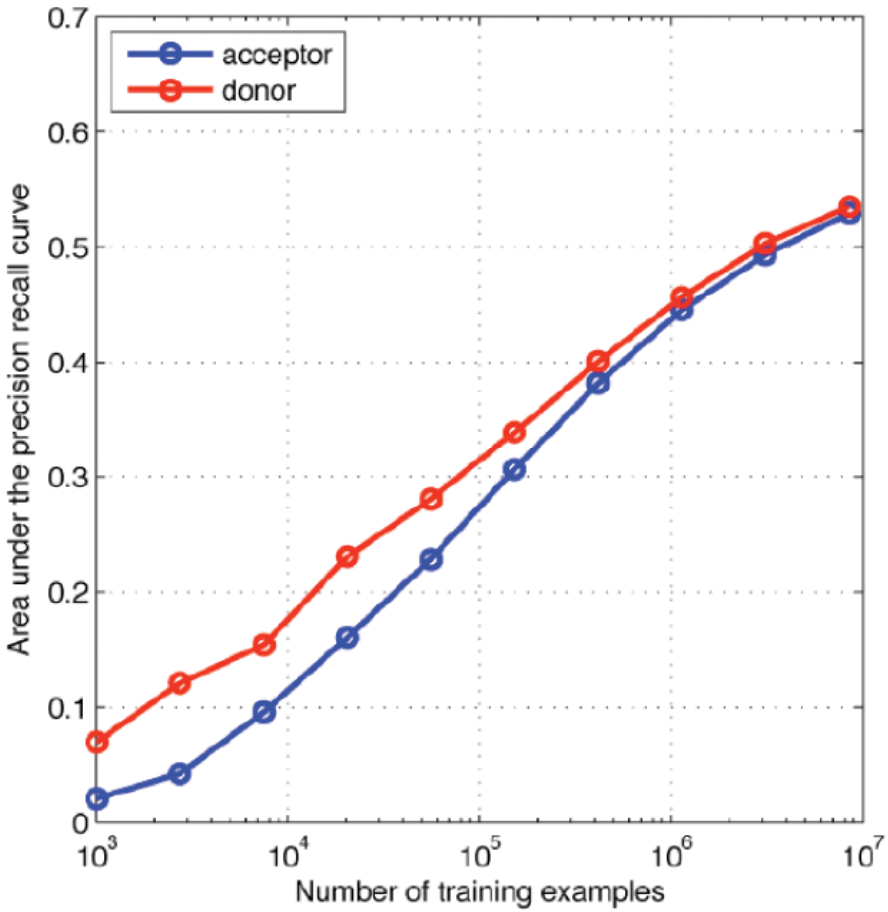


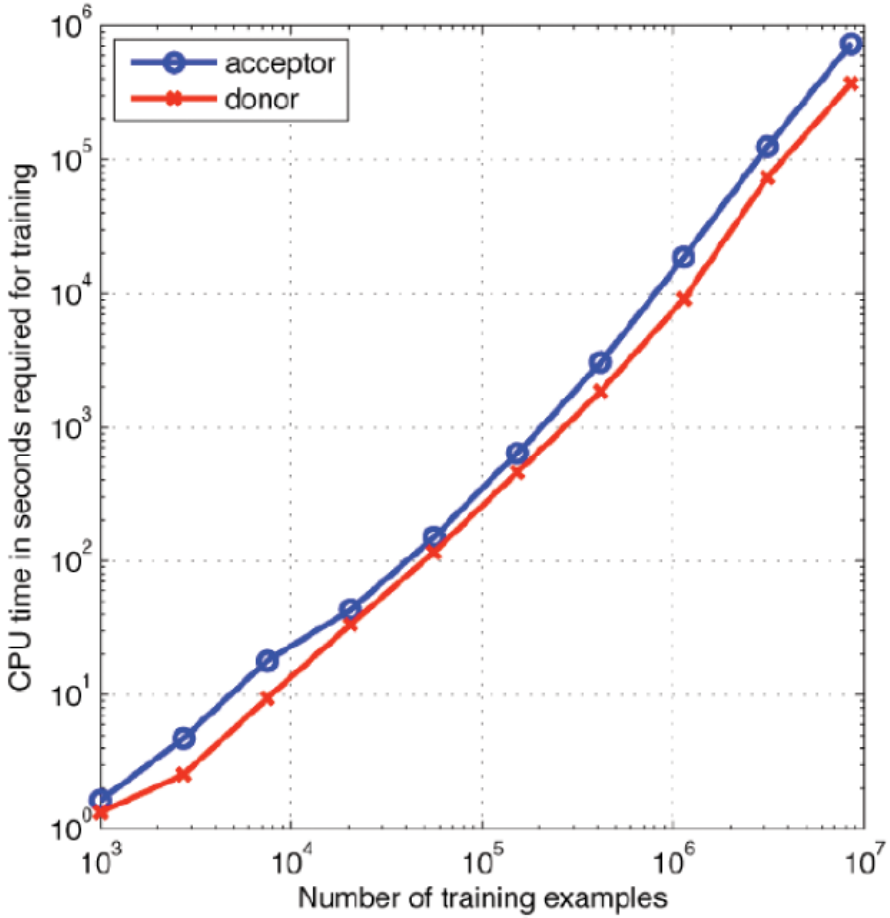


Performance (=Area under PRC curve) keeps increasing.

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

Human Splice Sites with WD Kernel
[Sonnenburg et al., 2007]



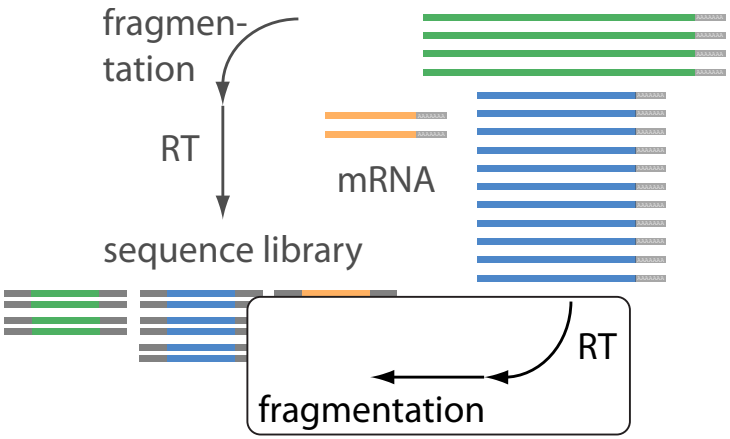


Performance (=Area under PRC curve) keeps increasing.

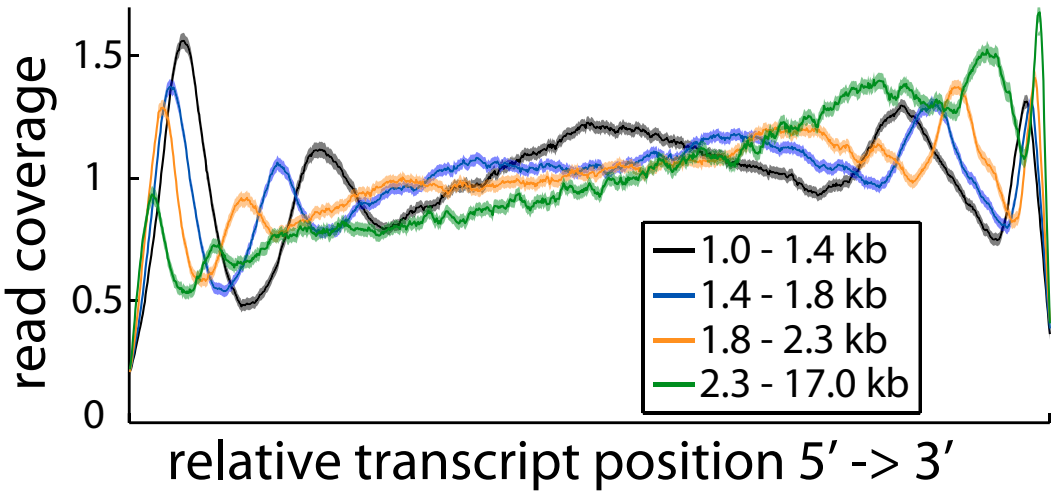
$\Rightarrow$ **Use all available data!**

Priming and Fragmentation Biases

Profile: normalised positional read coverage along the transcript



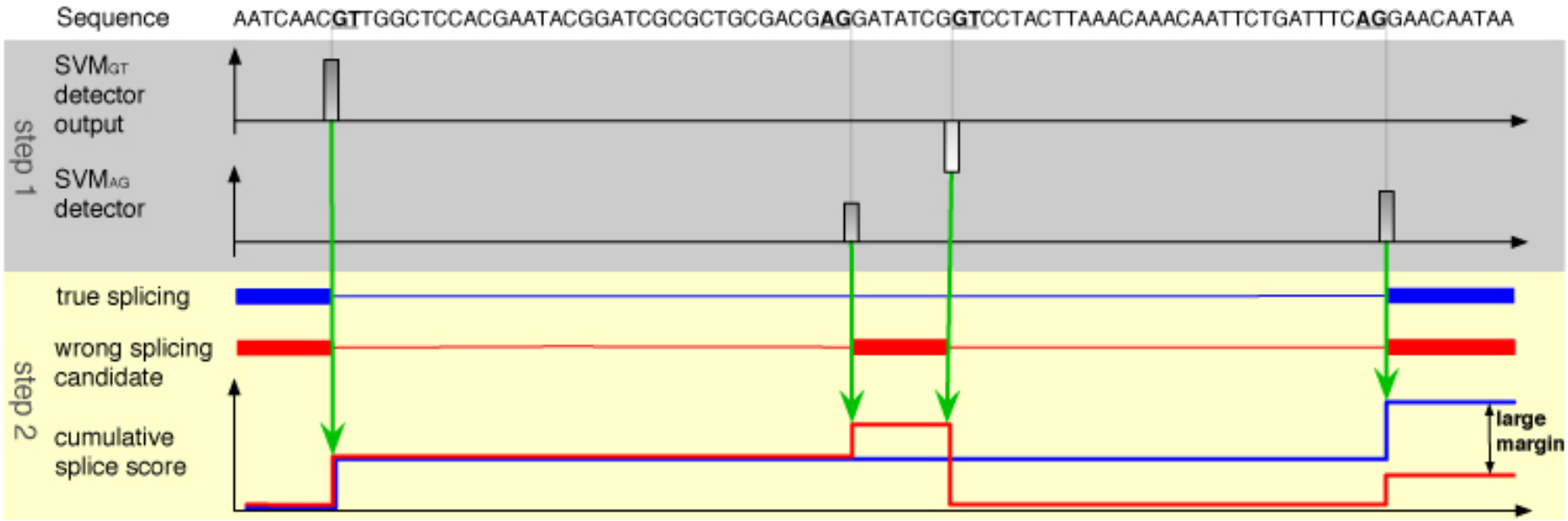
Transcript profiles for different lengths



Data generated for *A. thaliana* with Flux Simulator v1.0 [Sammeth, 2009b]

- Oligo-dT priming
- Physical cDNA fragmentation

Discriminative Gene Prediction (simplified)



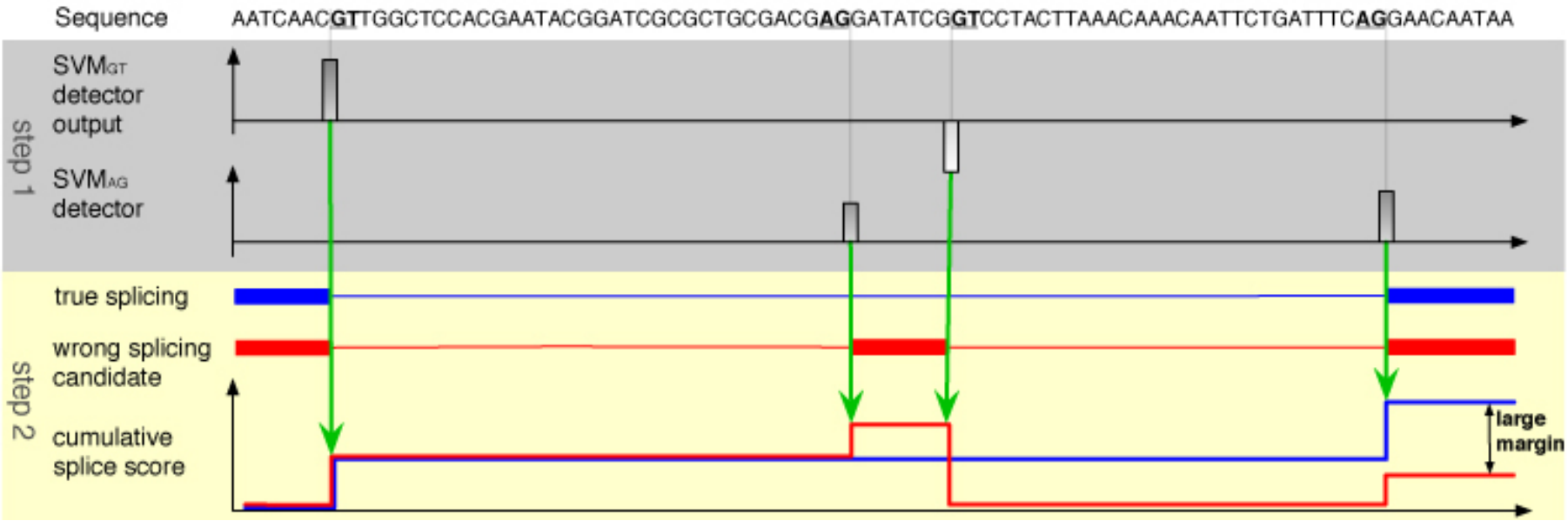
[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

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]

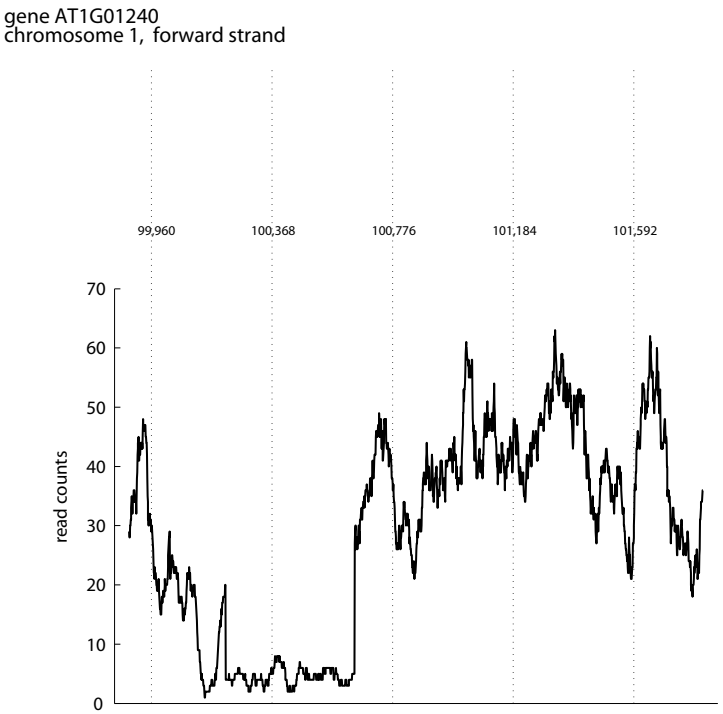
Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

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

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

rQuant – Iterative Algorithm

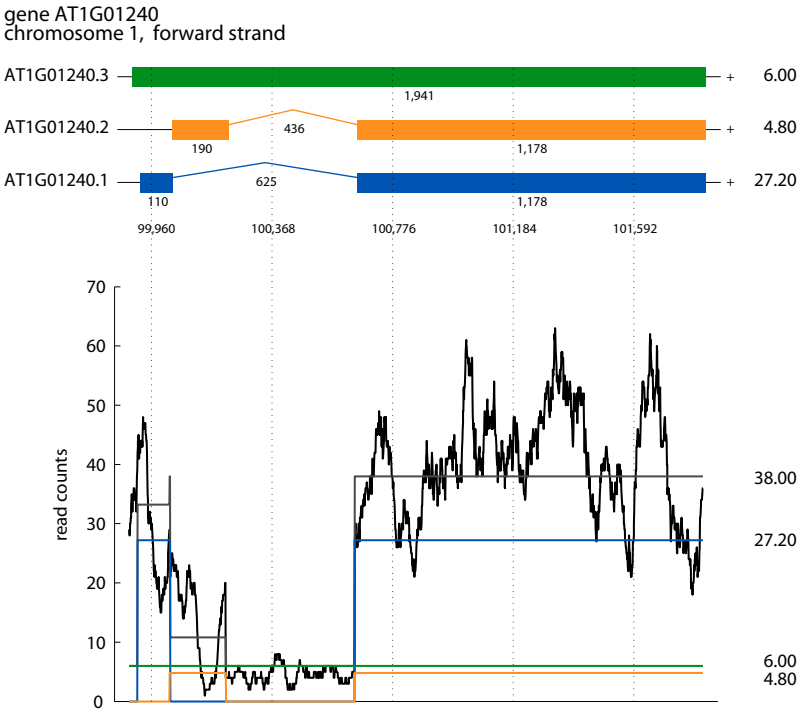
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



[Bohnert et al., 2009]

rQuant – Iterative Algorithm

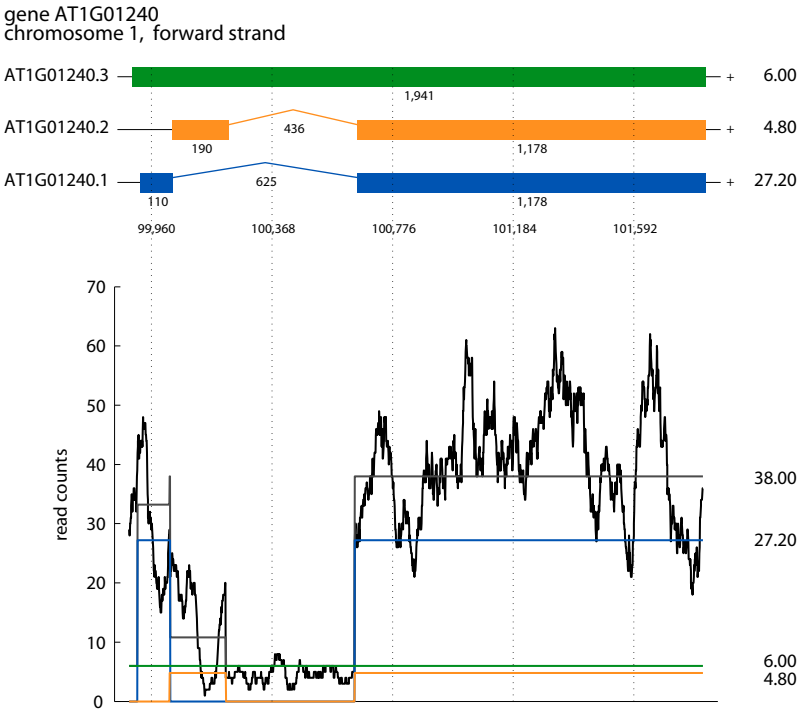
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.

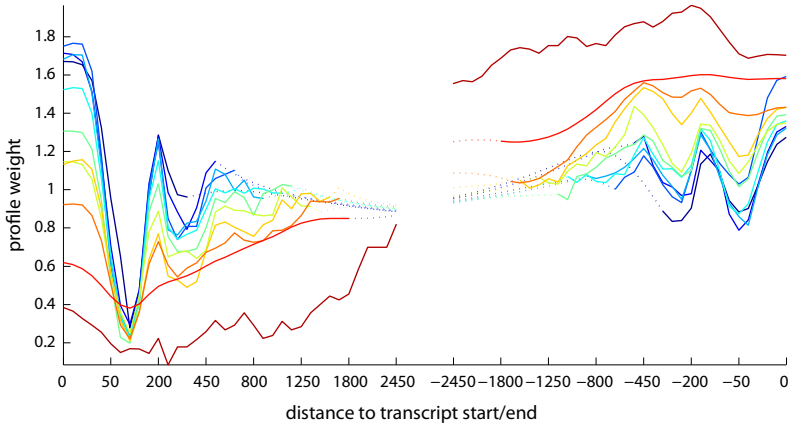


[Bohnert et al., 2009]

rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.

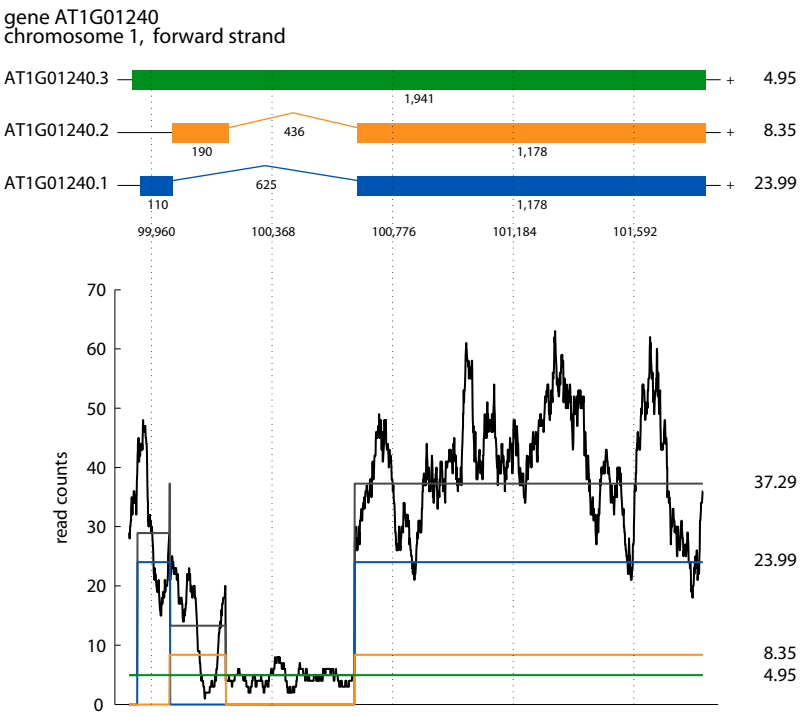


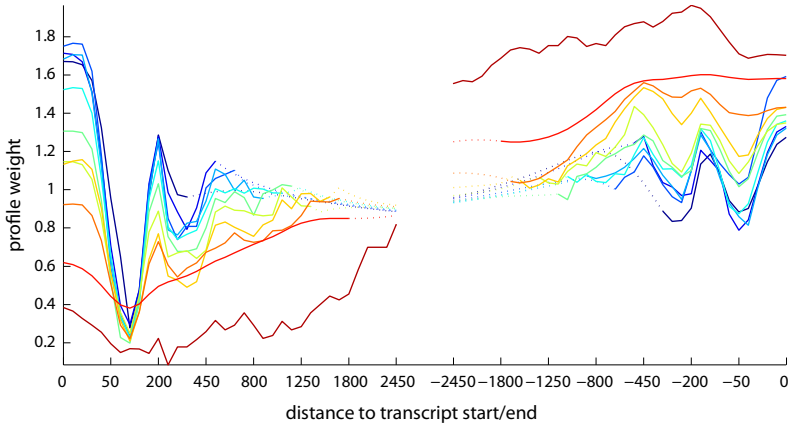


[Bohnert et al., 2009]

rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.





[Bohnert et al., 2009]

Evaluation I

Our method rQuant: Position-wise, with profiles

compared to

- Position-wise, without profiles
- Segment-wise, without profiles [Jiang and Wong, 2009]
- Segment-wise, with profiles [Sammeth, 2009a]

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator)
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

Evaluation I

Our method rQuant: Position-wise, with profiles

compared to

- Position-wise, without profiles
- Segment-wise, without profiles [Jiang and Wong, 2009]
- Segment-wise, with profiles [Sammeth, 2009a]

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator)
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

Evaluation I

Our method rQuant: Position-wise, with profiles

compared to

- Position-wise, without profiles
- Segment-wise, without profiles [Jiang and Wong, 2009]
- Segment-wise, with profiles [Sammeth, 2009a]

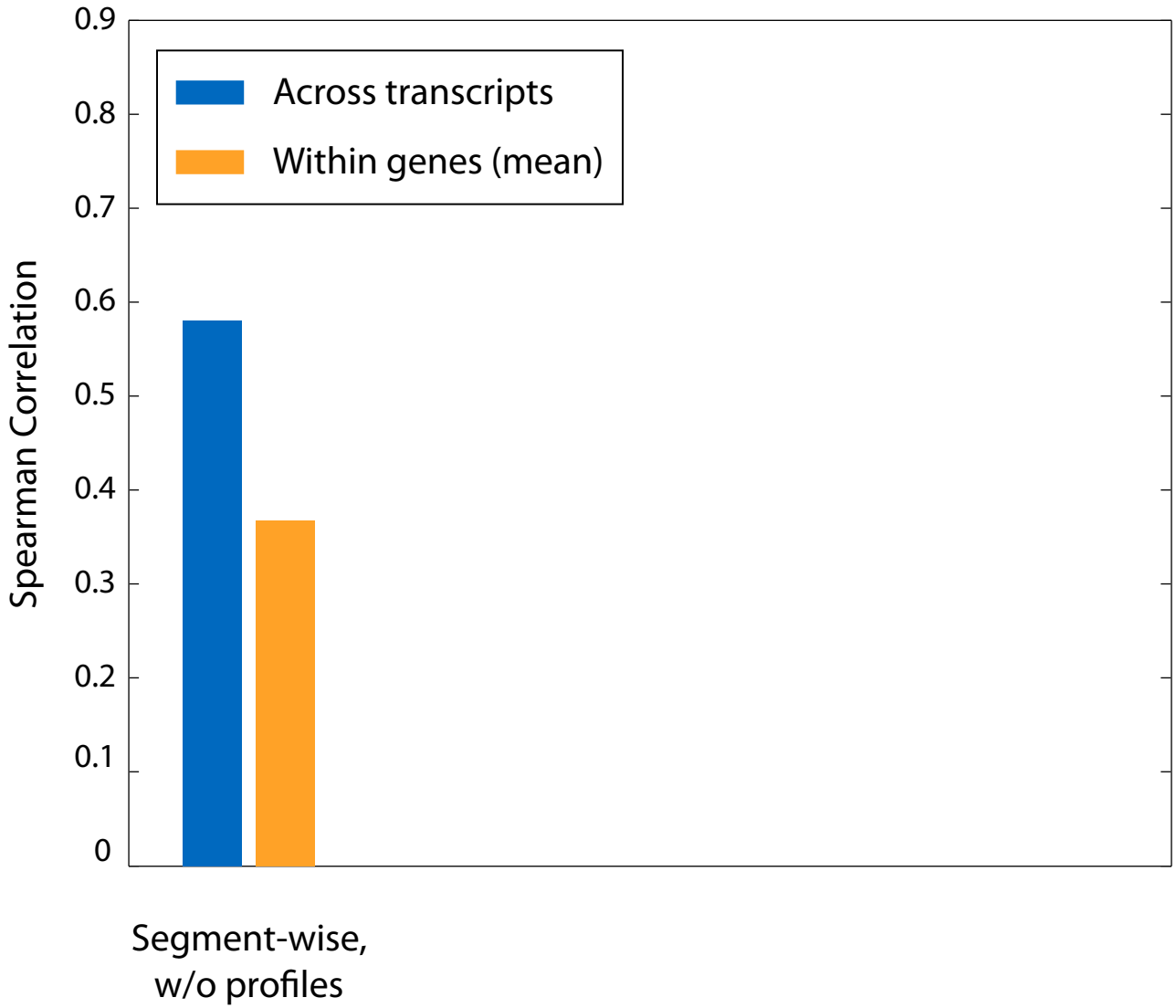
Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator)
- Subset of alternatively spliced genes

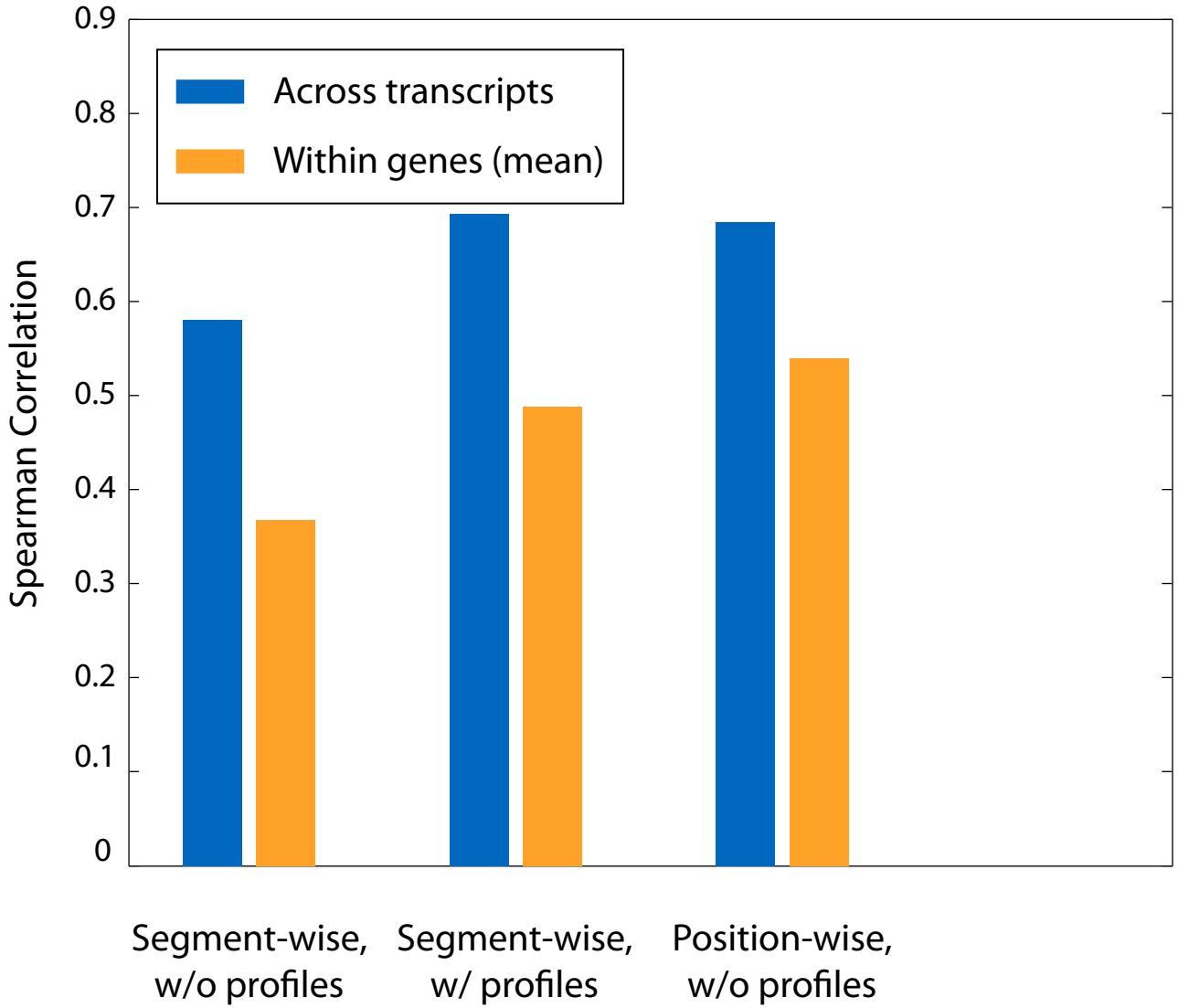
Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

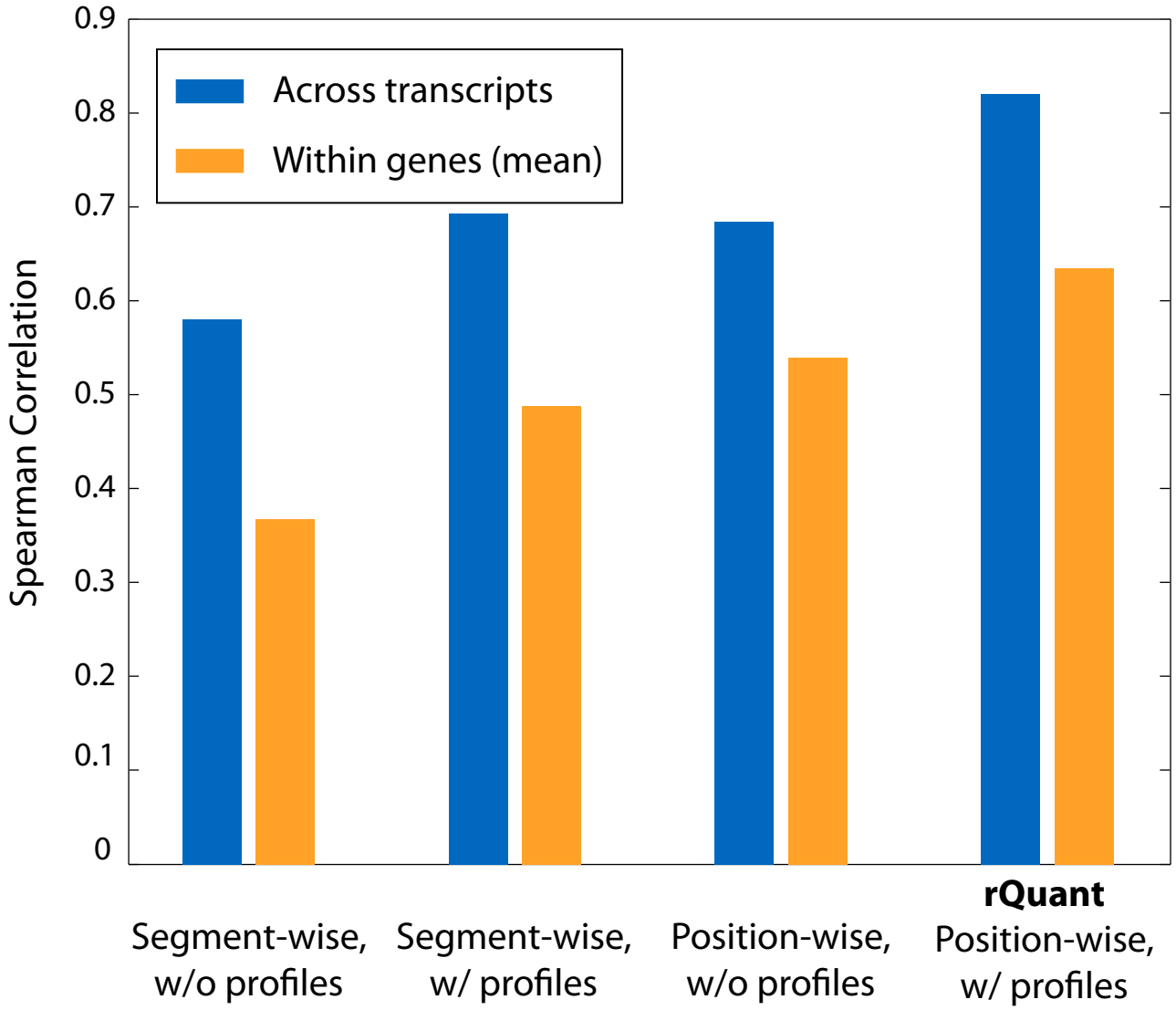
Evaluation II



Evaluation II



Evaluation II



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.

R. Bohnert, J. Behr, and G. Rätsch. Quantitative detection of alternative transcripts with rna-seq data. Oral presentation at the Short Read SIG at ISMB/ECCB 2009 in Stockholn, July 2009.

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.

References II

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.

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, February 2009. doi: 10.1093/bioinformatics/btp113.

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.

M. Sammeth. The Flux Capacitor. *Website*, 2009a. http://flux.sammeth.net/capacitor.html.

M. Sammeth. The flux simulator. *Website*, 2009b. http://flux.sammeth.net/simulator.html.

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.

References III

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 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.

References IV

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.