Methods for Quantitative RNA-seq-based Transcriptome Analyses

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society, Tübingen

Sloan-Kettering Institute
New York City (December 6, 2010)



Friedrich Miescher Laboratory
of the Max Planck Society



Discovery of the Nuclein
(Friedrich Miescher, 1869)





[Dahm, 2005]





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]

Discovery of the Nuclein
(Friedrich Miescher, 1869)





[Dahm, 2005]





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]

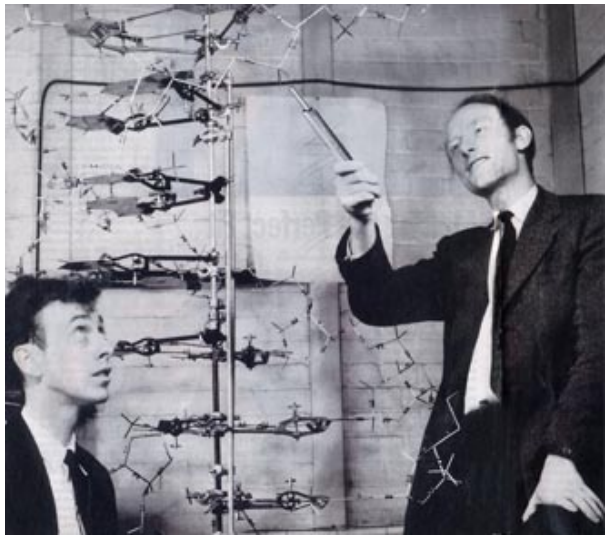
Discovery of the Nuclein
(Friedrich Miescher, 1869)





[Dahm, 2005]





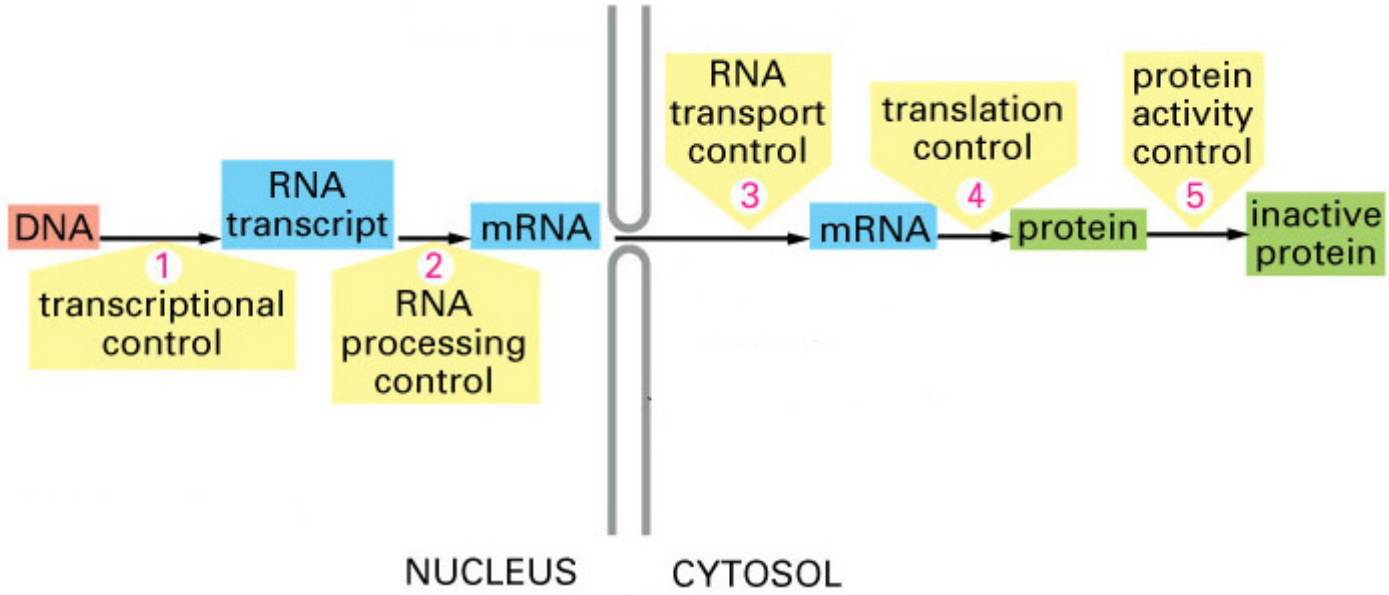
[Google images "watson crick"]

$\approx$ 80 years later ...

[Watson and Crick, 1953]

Learning about the Transcriptome

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



Computational View

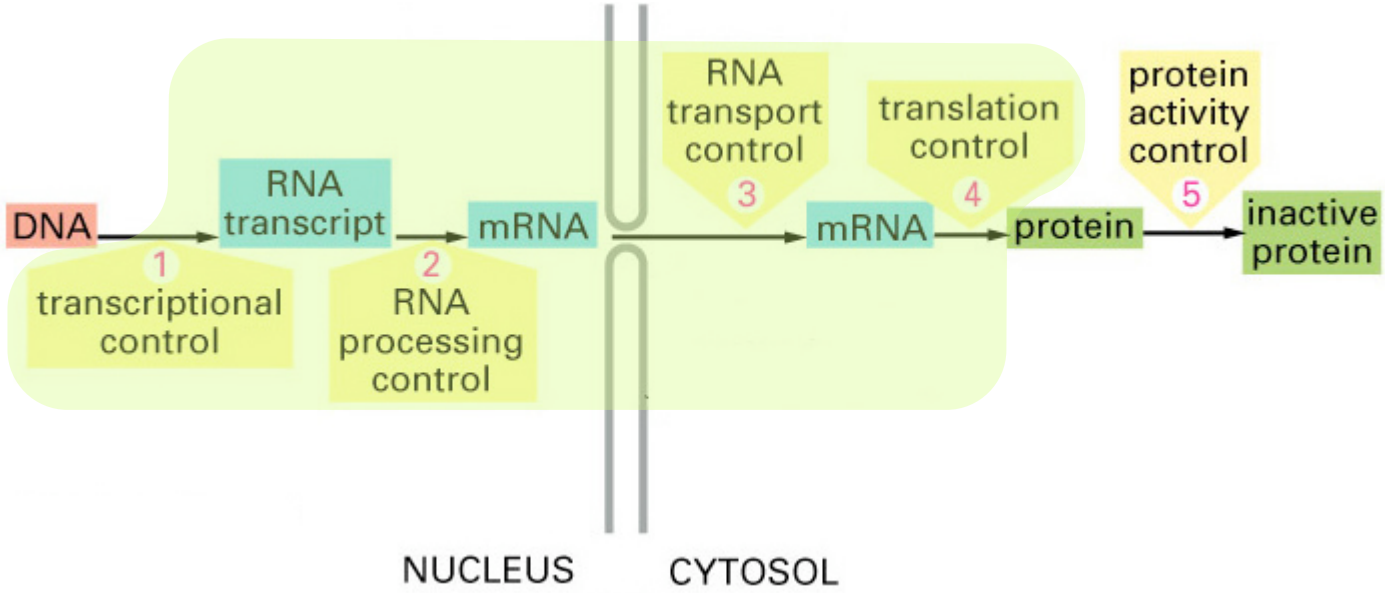
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

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

Learning about the Transcriptome

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



Computational View

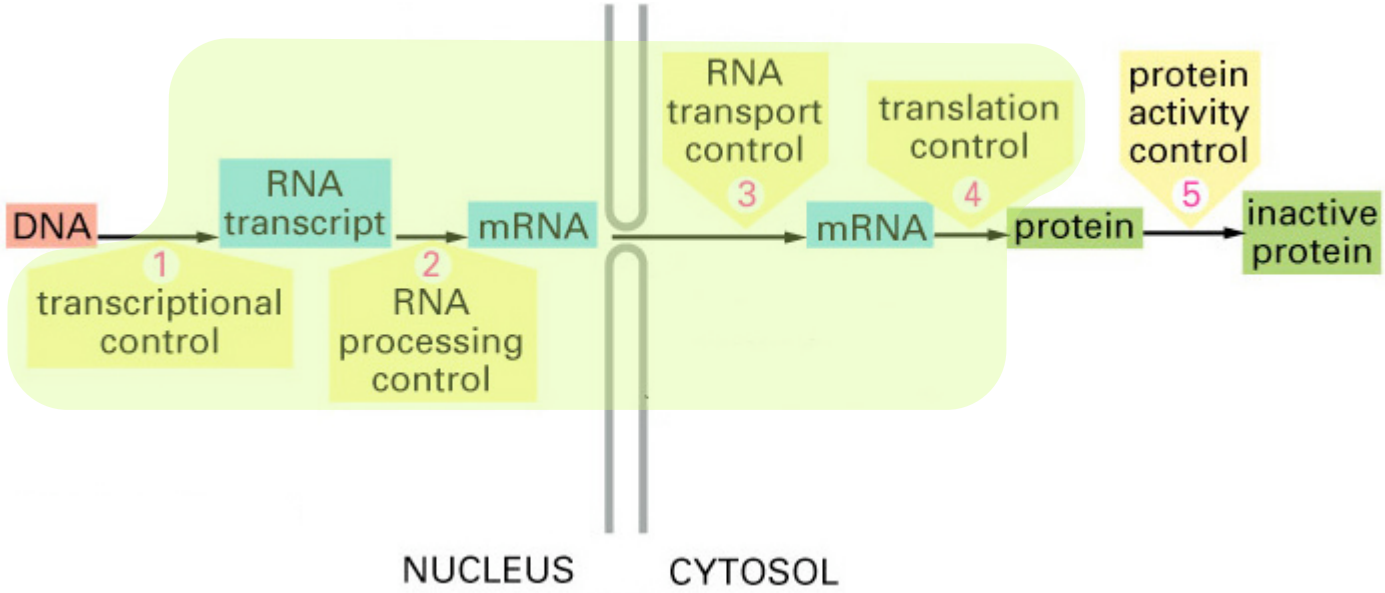
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

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

Learning about the Transcriptome

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



Computational View

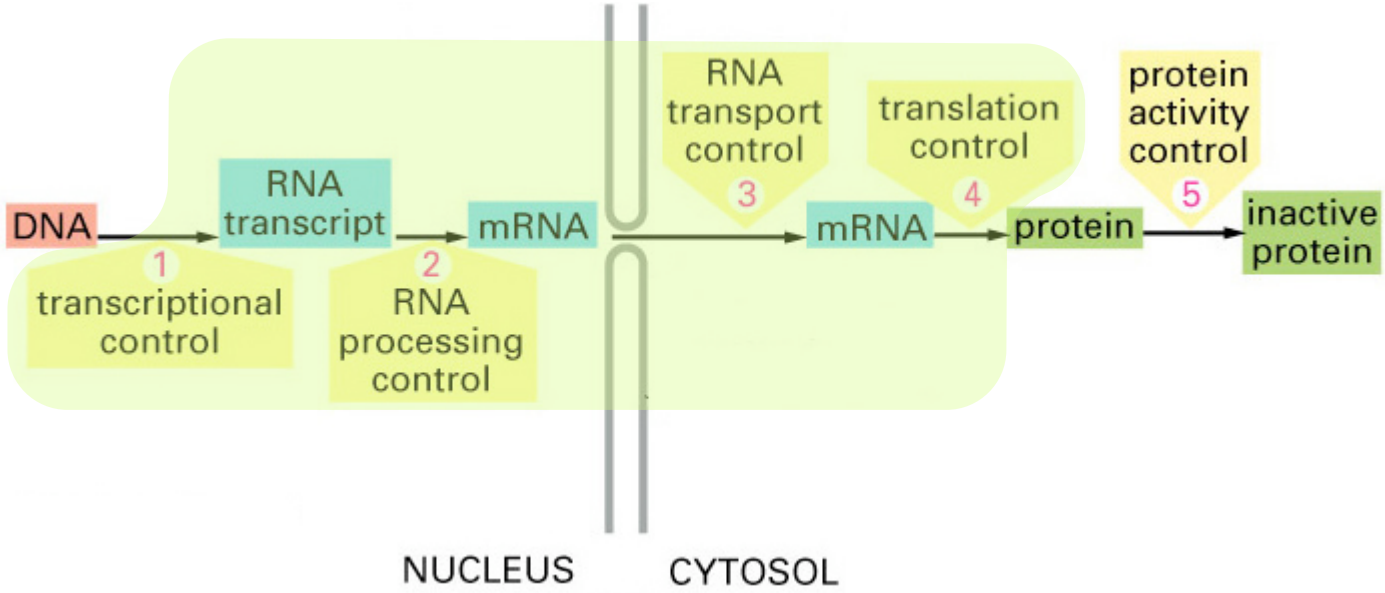
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

Biological View

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

Learning about the Transcriptome

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



Computational View

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

Biological View

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

Machine Learning in Biology
Research Topics



Machine Learning

1. Large scale sequence analysis
   ⇒ *Develop fast and accurate classification methods*
2. Predicting for structured outputs
   ⇒ *Accurate segmentation and structure prediction*

Molecular Biology

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

Machine Learning Methods

⇝ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification

   [Rätsch et al., 2006a, Sonnenburg et al., 2007a, Toussaint et al., 2010, Sonnenburg et al., 2010]

2. Analysis and explanation of learning results

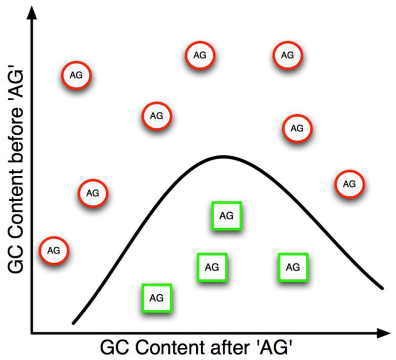
   [Rätsch et al., 2006b, Sonnenburg et al., 2008, Zien et al., 2009]

3. Sequence segmentation & structure prediction

   [Zeller et al., 2008, Rätsch et al., 2007, Schweikert et al., 2009b]

4. Transfer learning

   [Schweikert et al., 2008, Widmer et al., 2010b,a]



Algorithms implemented in Shogun toolbox (GPL, $\geq$ 1000 users)

Machine Learning Methods

⇝ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification

   [Rätsch et al., 2006a, Sonnenburg et al., 2007a, Toussaint et al., 2010, Sonnenburg et al., 2010]

2. Analysis and explanation of learning results

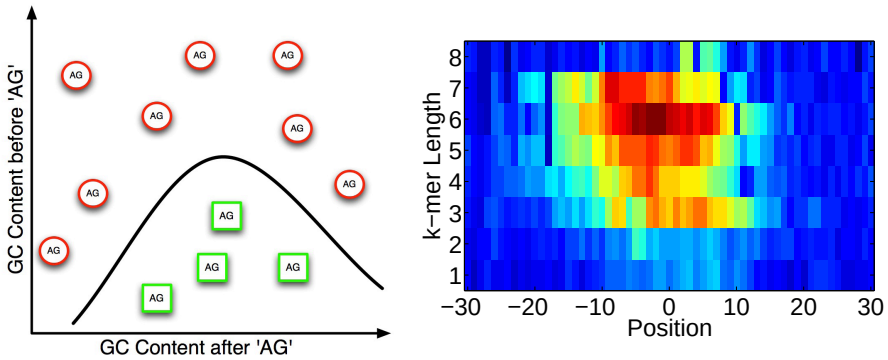
   [Rätsch et al., 2006b, Sonnenburg et al., 2008, Zien et al., 2009]

3. Sequence segmentation & structure prediction

   [Zeller et al., 2008, Rätsch et al., 2007, Schweikert et al., 2009b]

4. Transfer learning

   [Schweikert et al., 2008, Widmer et al., 2010b,a]



Algorithms implemented in Shogun toolbox (GPL, ≥ 1000 users)

Machine Learning Methods

⇝ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification

   [Rätsch et al., 2006a, Sonnenburg et al., 2007a, Toussaint et al., 2010, Sonnenburg et al., 2010]

2. Analysis and explanation of learning results

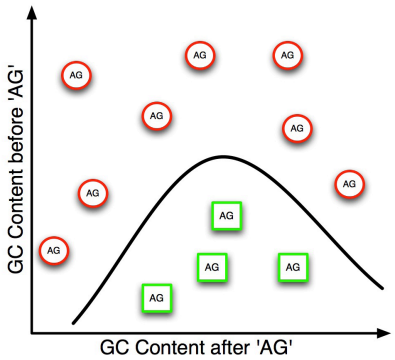
   [Rätsch et al., 2006b, Sonnenburg et al., 2008, Zien et al., 2009]

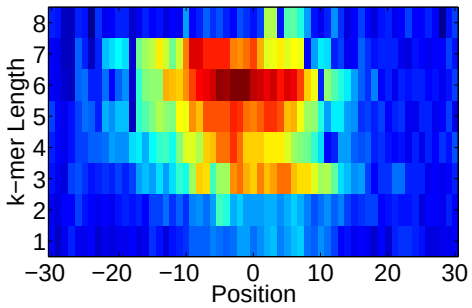
3. Sequence segmentation & structure prediction

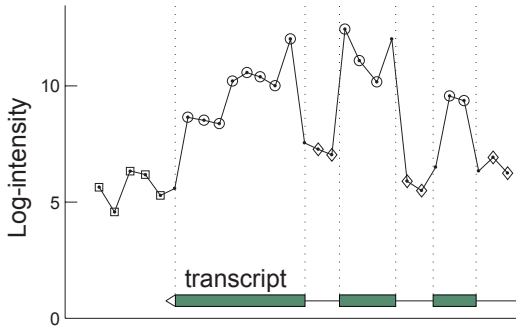
   [Zeller et al., 2008, Rätsch et al., 2007, Schweikert et al., 2009b]

4. Transfer learning

   [Schweikert et al., 2008, Widmer et al., 2010b,a]







Algorithms implemented in Shogun toolbox (GPL, ≥ 1000 users)

Machine Learning Methods

⇝ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification

   [Rätsch et al., 2006a, Sonnenburg et al., 2007a, Toussaint et al., 2010, Sonnenburg et al., 2010]

2. Analysis and explanation of learning results

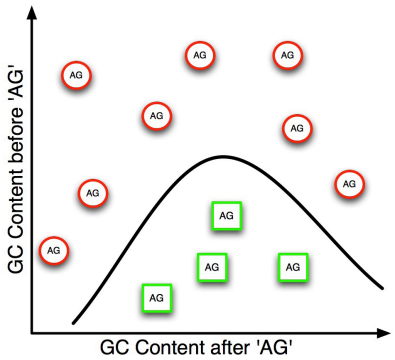
   [Rätsch et al., 2006b, Sonnenburg et al., 2008, Zien et al., 2009]

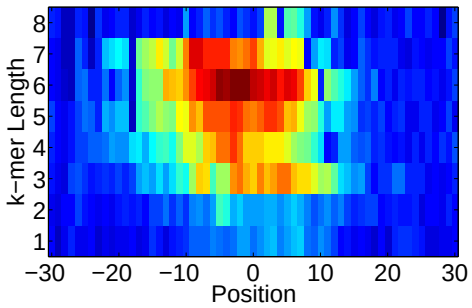
3. Sequence segmentation & structure prediction

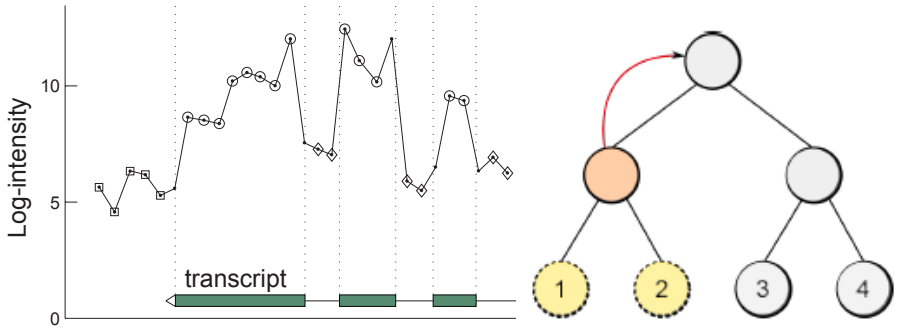
   [Zeller et al., 2008, Rätsch et al., 2007, Schweikert et al., 2009b]

4. Transfer learning

   [Schweikert et al., 2008, Widmer et al., 2010b,a]







Algorithms implemented in Shogun toolbox (GPL, ≥ 1000 users)

Machine Learning Methods

⇝ Develop fast, accurate and interpretable learning methods

1. Large scale sequence classification

   [Rätsch et al., 2006a, Sonnenburg et al., 2007a, Toussaint et al., 2010, Sonnenburg et al., 2010]

2. Analysis and explanation of learning results

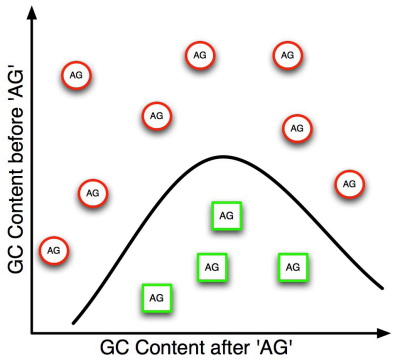
   [Rätsch et al., 2006b, Sonnenburg et al., 2008, Zien et al., 2009]

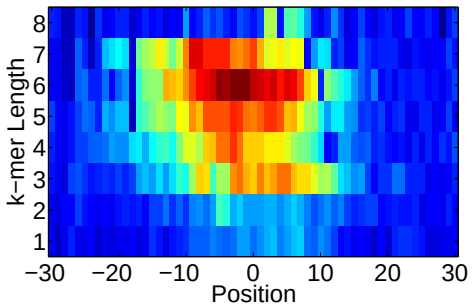
3. Sequence segmentation & structure prediction

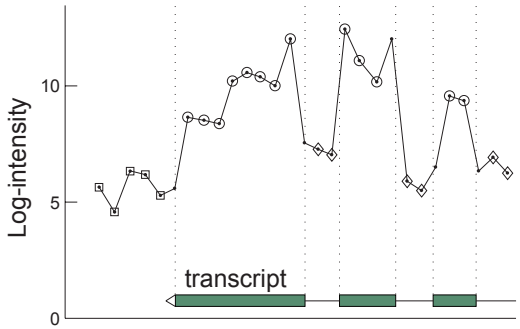
   [Zeller et al., 2008, Rätsch et al., 2007, Schweikert et al., 2009b]

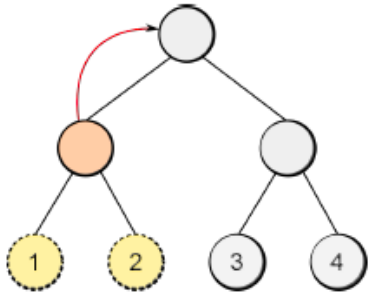
4. Transfer learning

   [Schweikert et al., 2008, Widmer et al., 2010b,a]



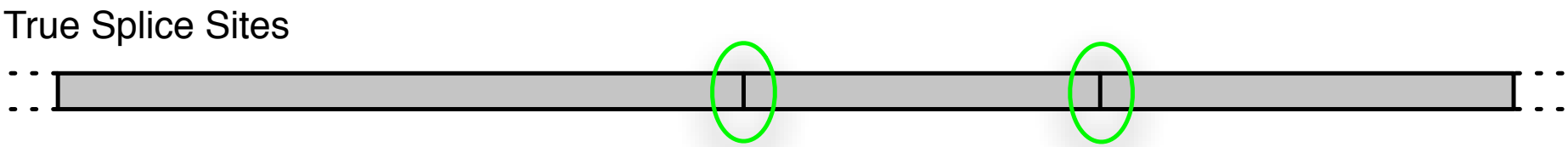




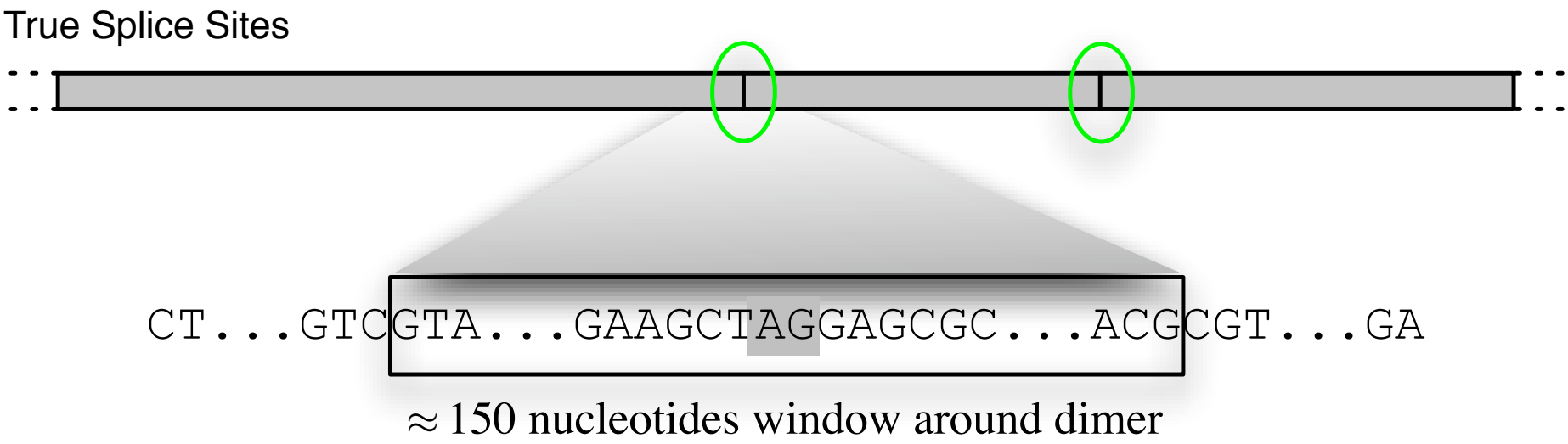


Algorithms implemented in Shogun toolbox (GPL, $\geq$ 1000 users)

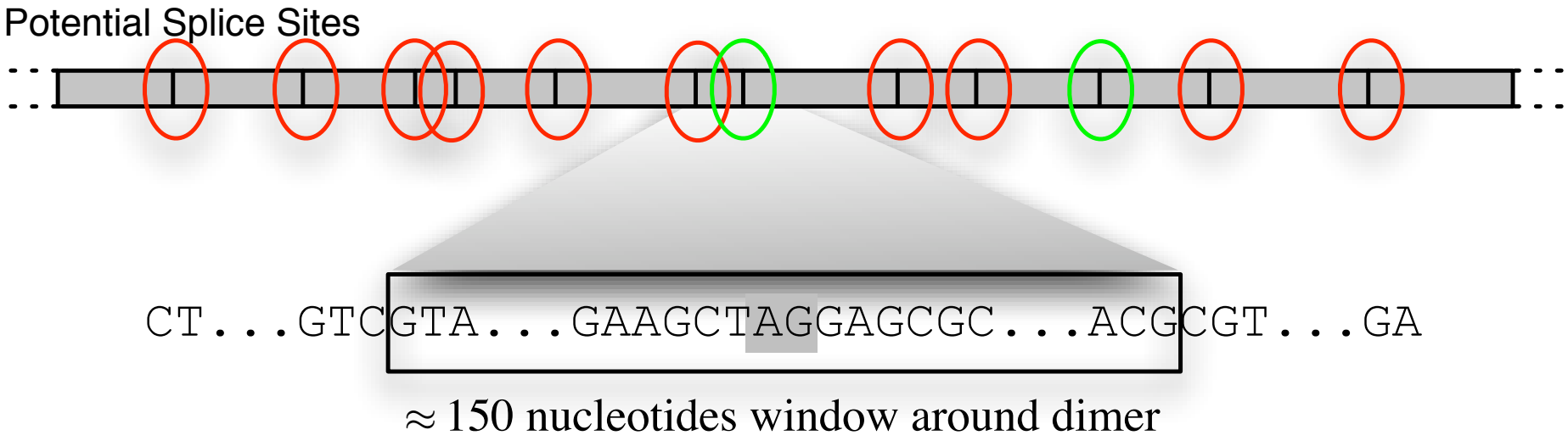
Example: Splice Site Recognition



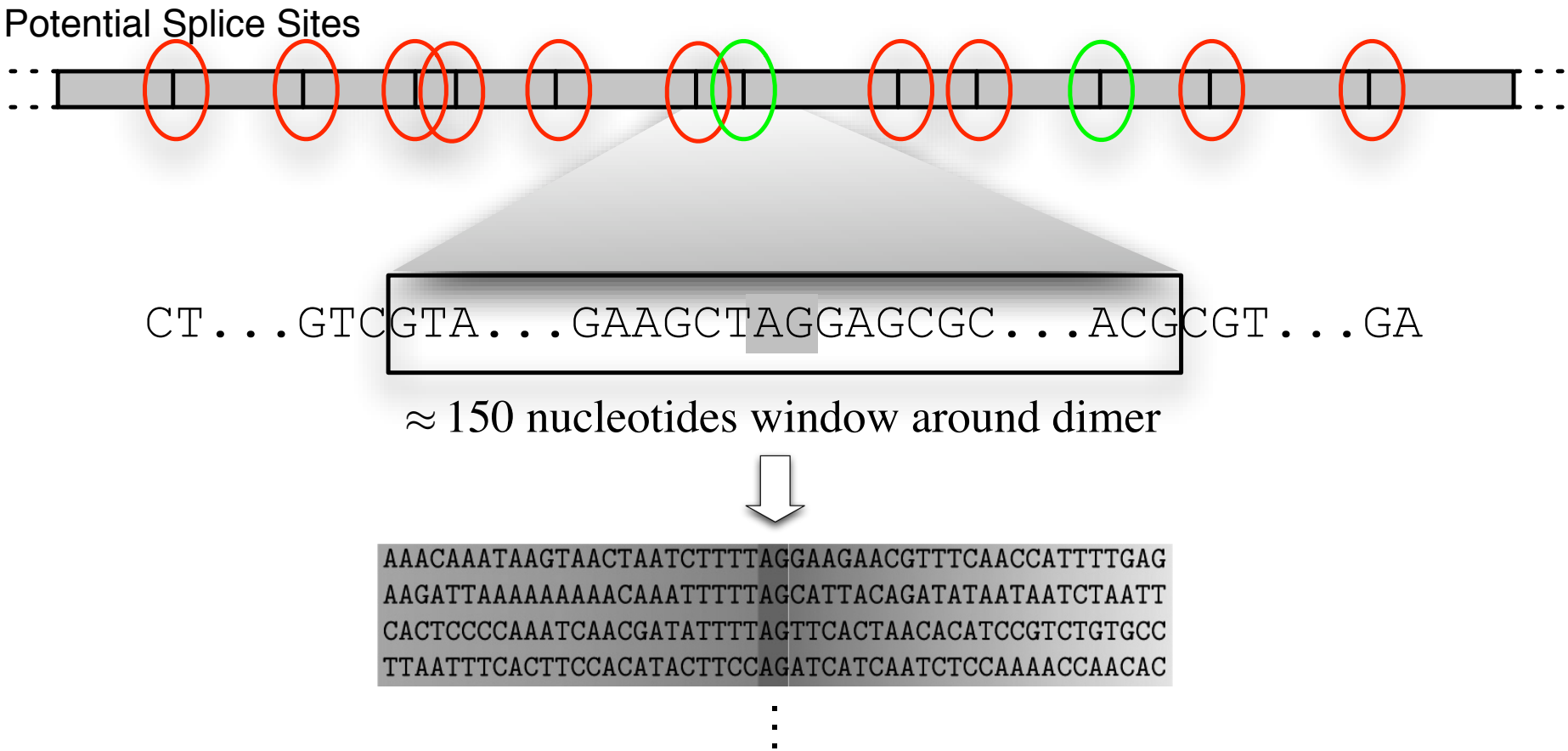
Example: Splice Site Recognition



Example: Splice Site Recognition



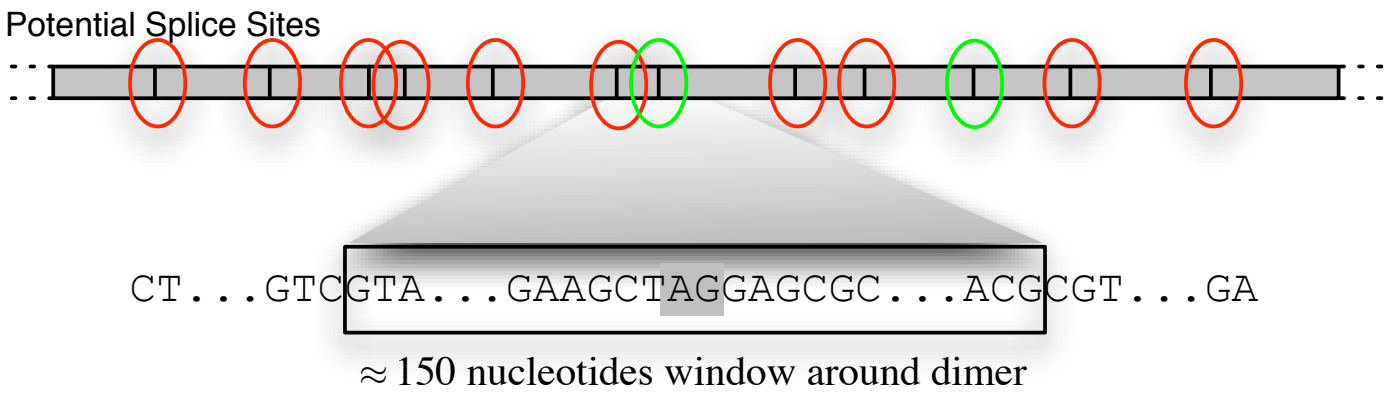
Example: Splice Site Recognition



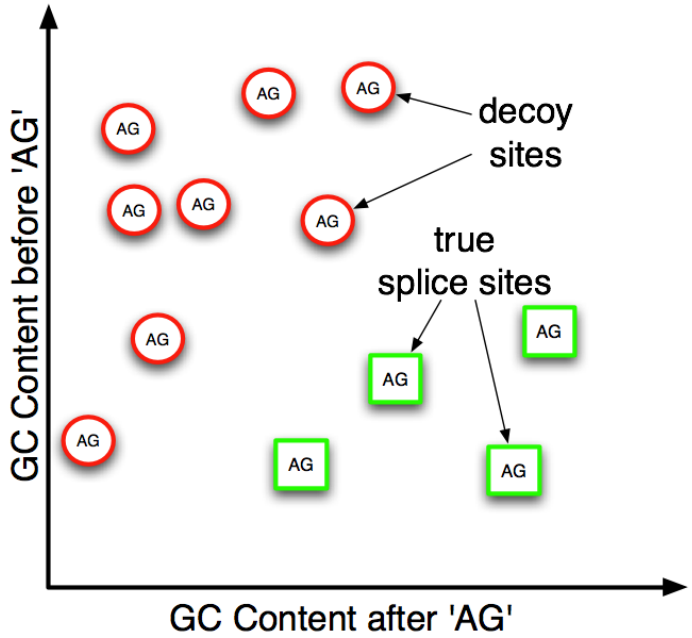
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

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

Example: Splice Site Recognition



Basic idea:



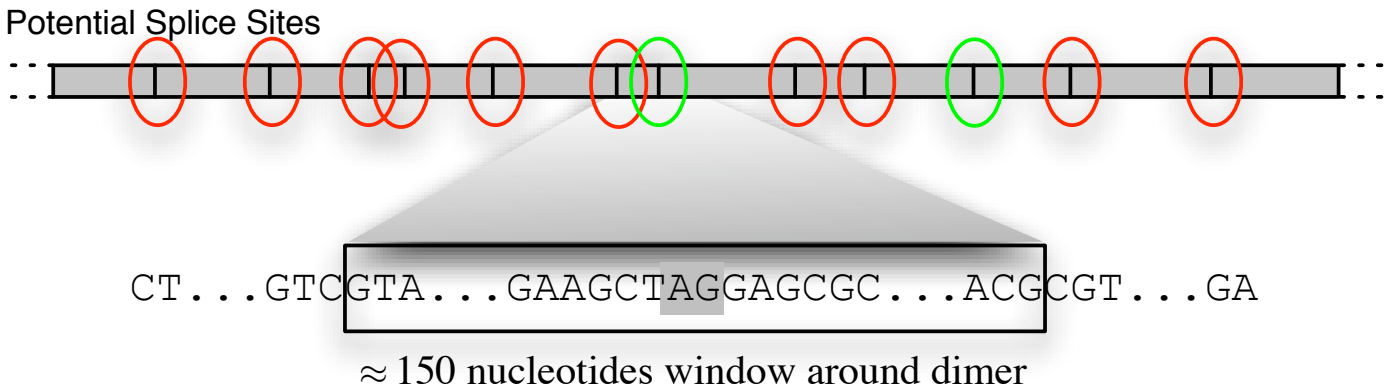
For instance, exploit that exons have higher GC content

or

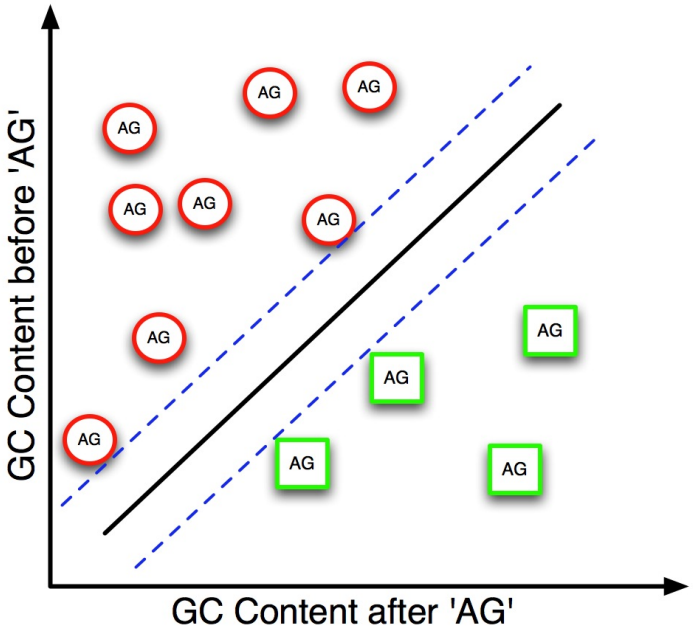
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007b]

Example: Splice Site Recognition



Basic idea:

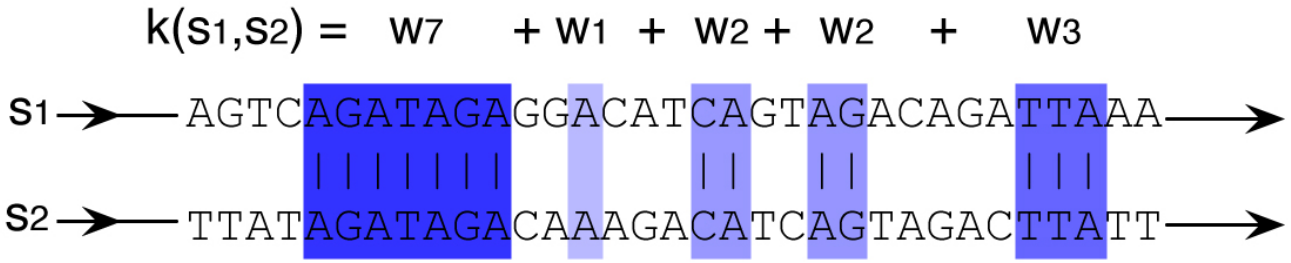


In practice: Use one feature per possible substring (e.g. $\leq$20) at all positions

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

Kernels and String Data Structures

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



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

Drawback: Training is too expensive for millions of examples

Needs more sophisticated algorithms:

- Exploit that features are sparse and can be explicitly represented
  - String indexing data structures

    [Sonnenburg et al., 2007a, Sonnenburg and Franc, 2010]
- Use novel optimization techniques
  - Subgradient-based optimization

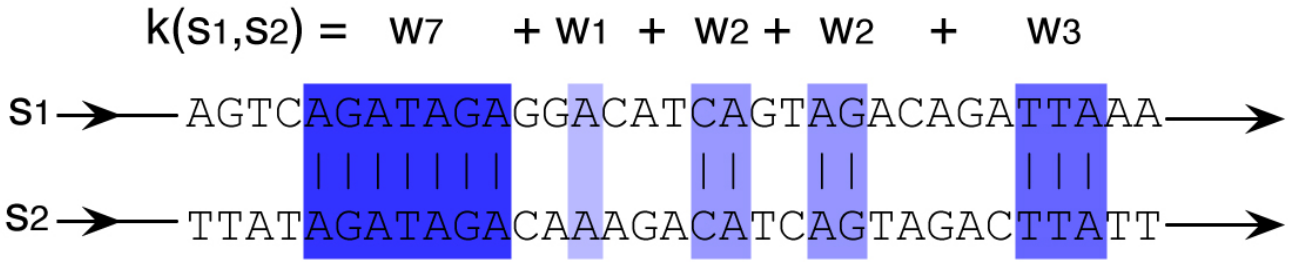
    [Franc and Sonnenburg, 2009]

Implemented in Shogun toolbox

[Sonnenburg et al., 2010]

Kernels and String Data Structures

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



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

Drawback: Training is too expensive for millions of examples

Needs more sophisticated algorithms:

- Exploit that features are sparse and can be explicitly represented
  - String indexing data structures

    [Sonnenburg et al., 2007a, Sonnenburg and Franc, 2010]
- Use novel optimization techniques
  - Subgradient-based optimization

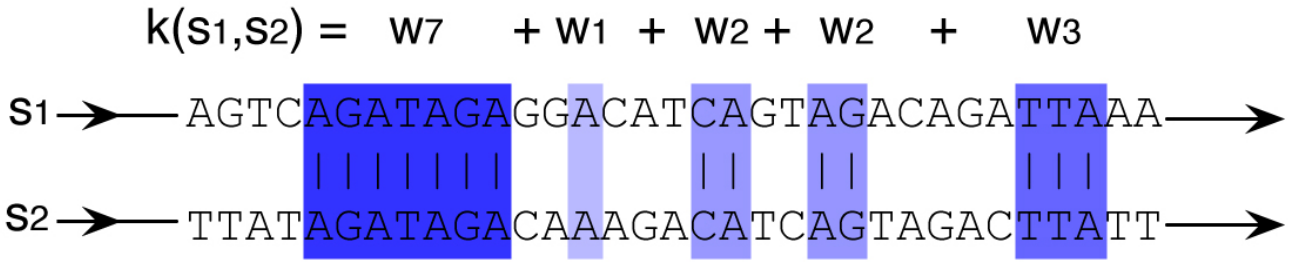
    [Franc and Sonnenburg, 2009]

Implemented in Shogun toolbox

[Sonnenburg et al., 2010]

Kernels and String Data Structures

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



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

Drawback: Training was too expensive for millions of examples

Needs more sophisticated algorithms:

- Exploit that features are sparse and can be explicitly represented
  - String indexing data structures

    [Sonnenburg et al., 2007a, Sonnenburg and Franc, 2010]
- Use novel optimization techniques
  - Subgradient-based optimization [Franc and Sonnenburg, 2009]

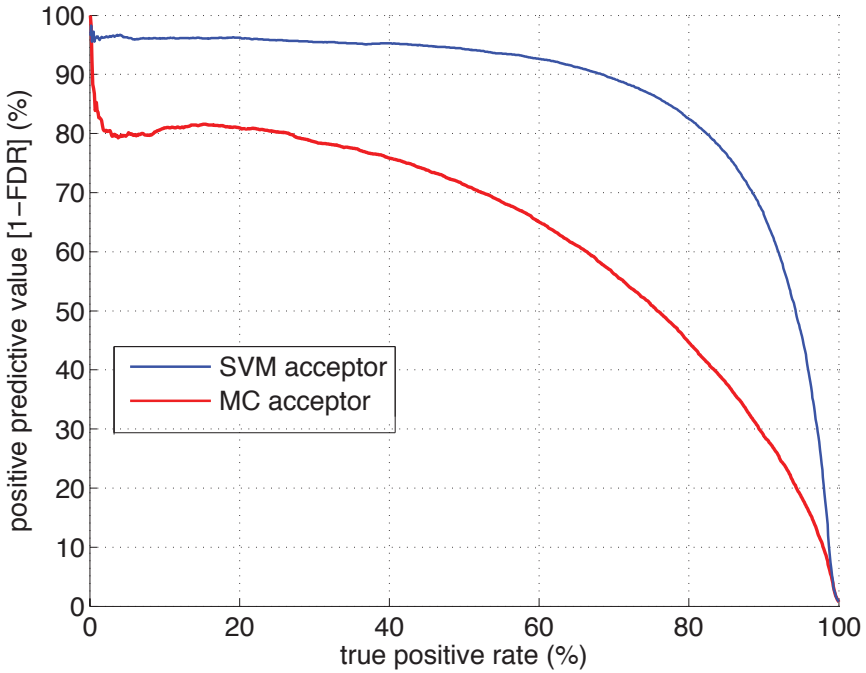
Implemented in Shogun toolbox [Sonnenburg et al., 2010]

Results on Splice Site Recognition

[Sonnenburg et al., 2007b]

	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

Acceptor Splice Site Recognition:

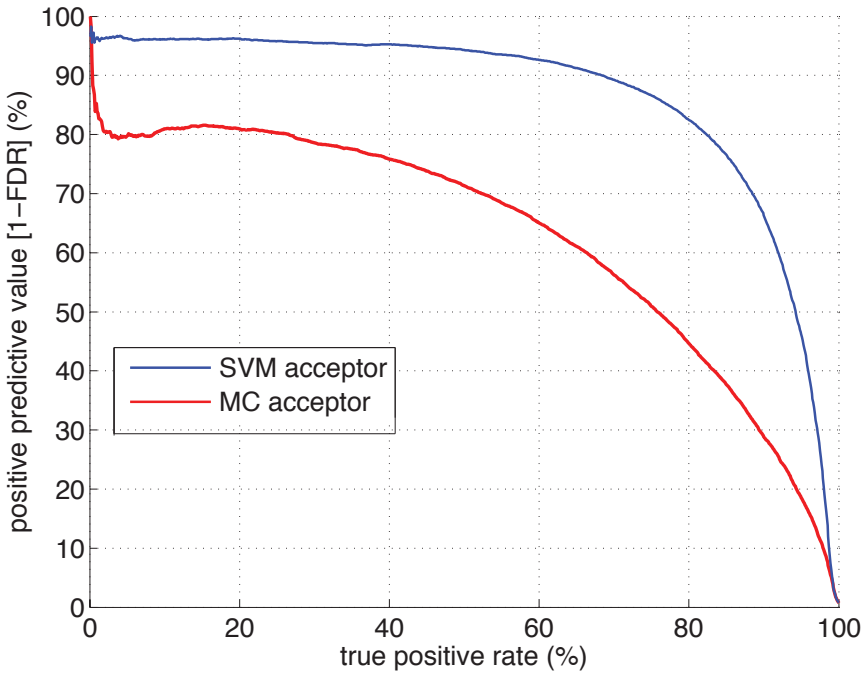


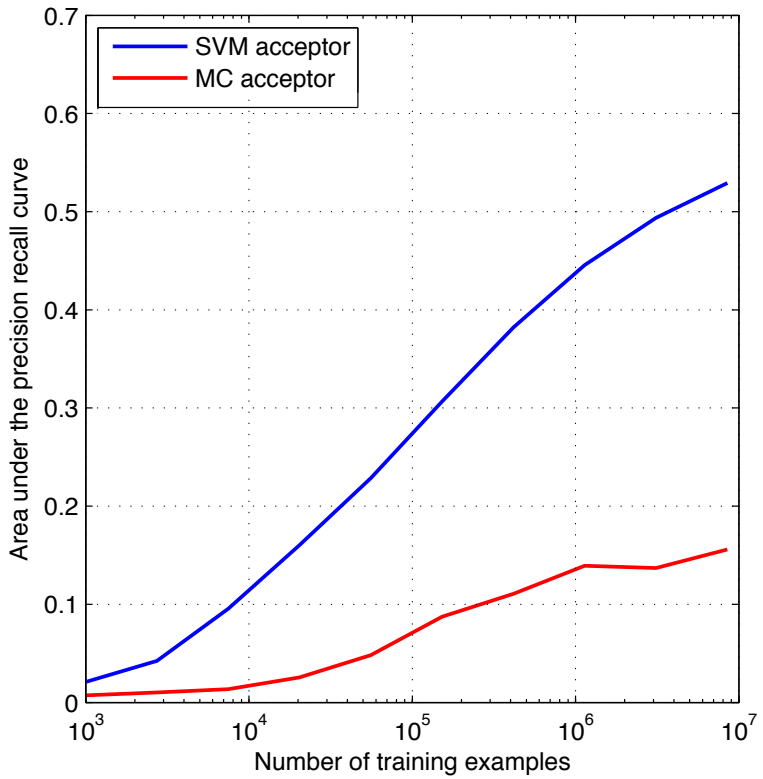
Results on Splice Site Recognition

[Sonnenburg et al., 2007b]

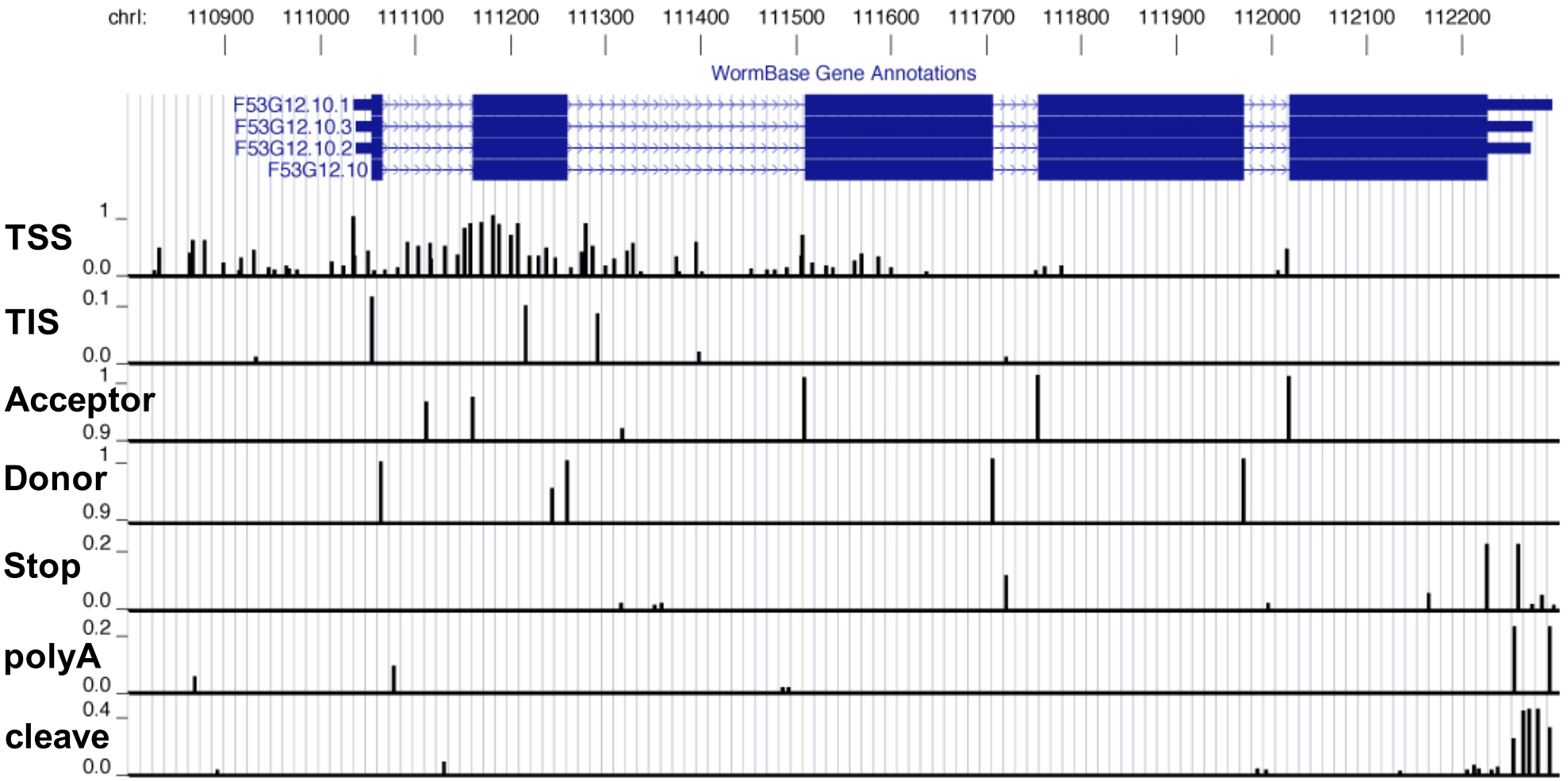
	Worm		Fly		Cress		Fish		Human	
	Acc	Don	Acc	Don	Acc	Don	Acc	Don	Acc	Don
Markov Chain auPRC(%)	92.1	90.0	80.3	78.5	87.4	88.2	63.6	62.9	16.2	26.0
SVM auPRC(%)	95.9	95.3	86.7	87.5	92.2	92.9	86.6	86.9	54.4	56.9

Acceptor Splice Site Recognition:



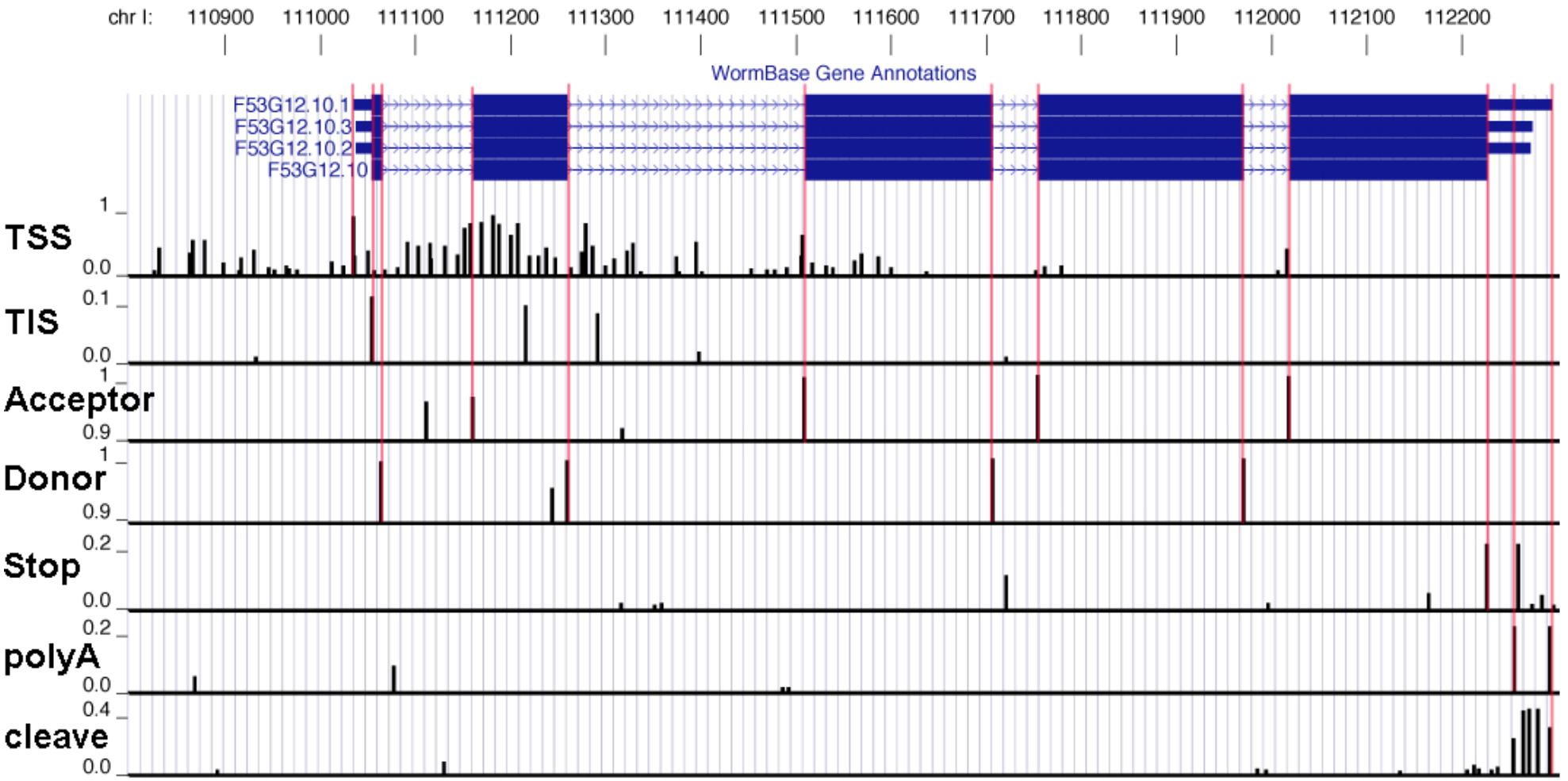


Genome-wide Signal Predictions



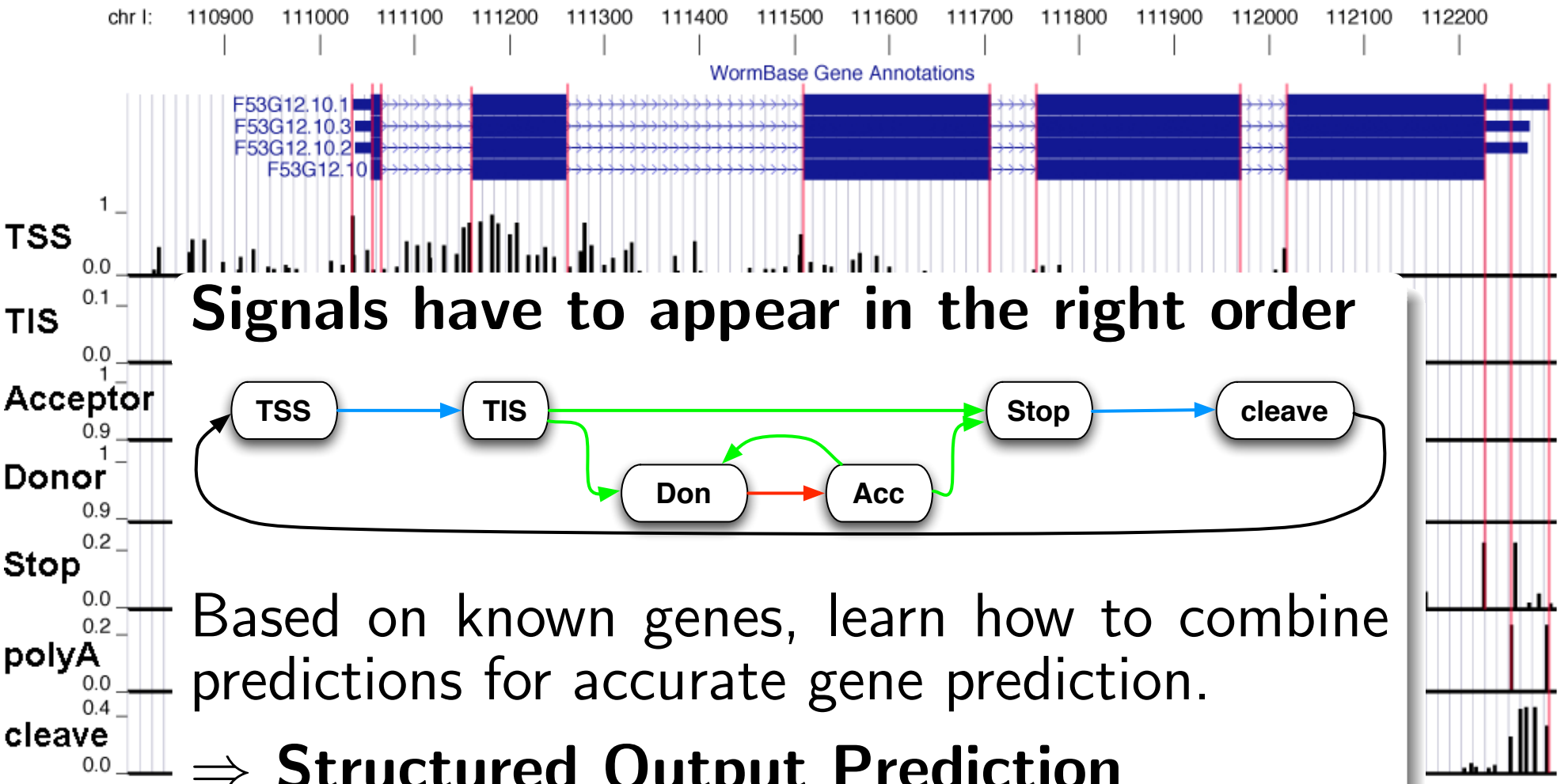
[Rätsch et al., 2007, Schweikert et al., 2009b]

Genome-wide Signal Predictions



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

Genome-wide Signal Predictions



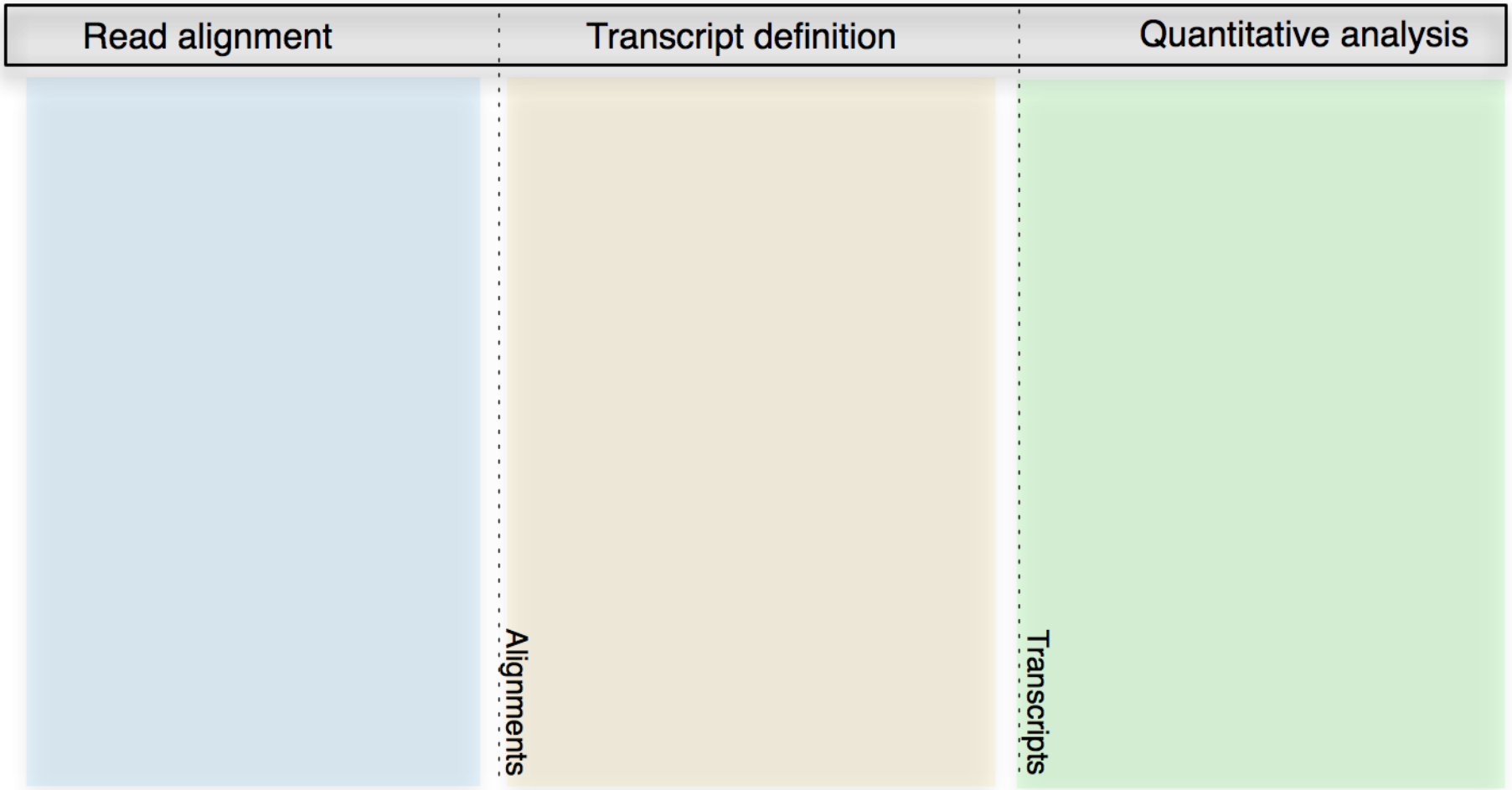
Signals have to appear in the right order

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

$\Rightarrow$ **Structured Output Prediction**

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

RNA-seq-based Transcriptome Analyses

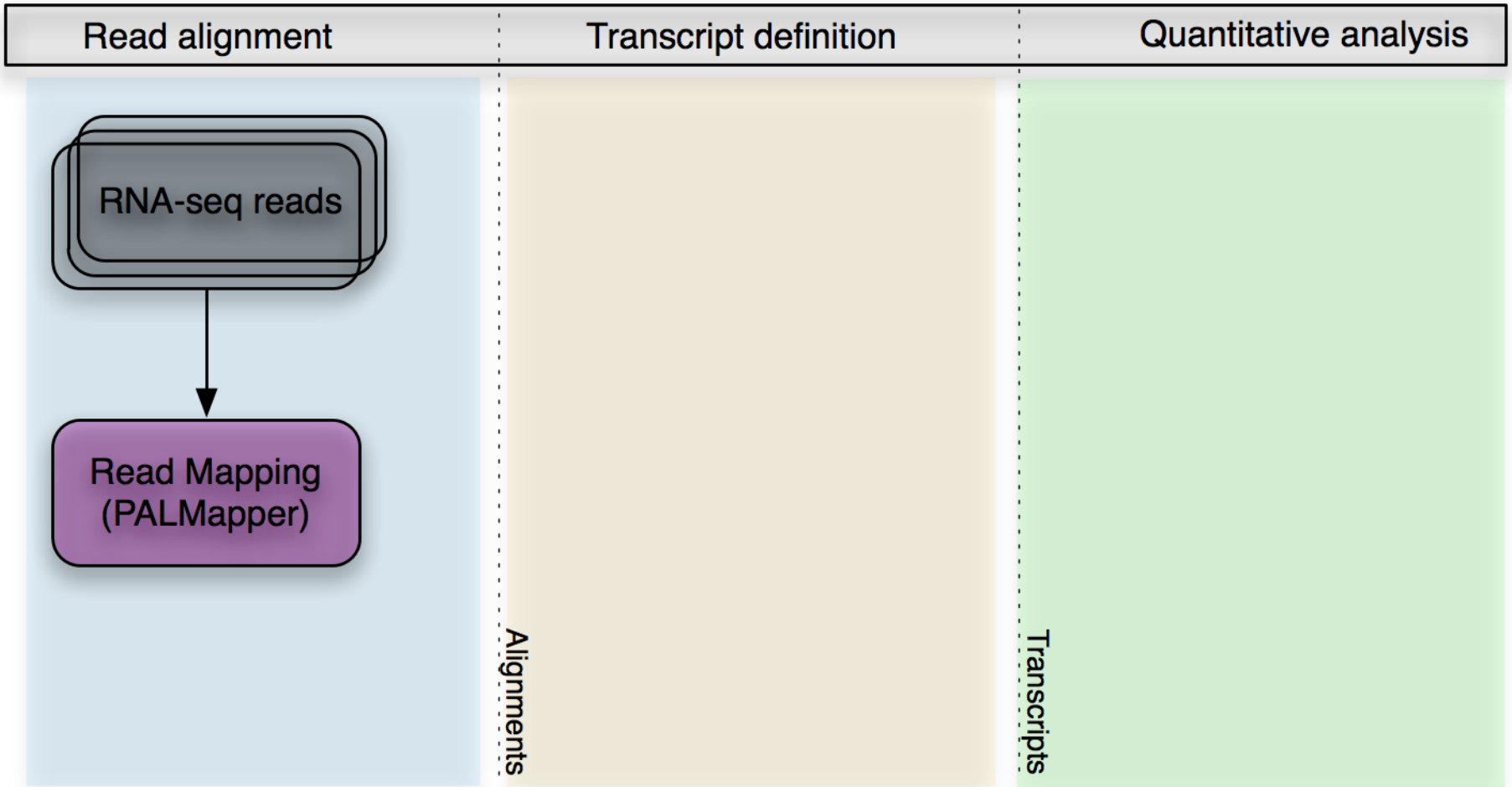


ShortSIG 2008: QPalma ($\sim$ Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff [Stegle et al., 2010a]

RNA-seq-based Transcriptome Analyses

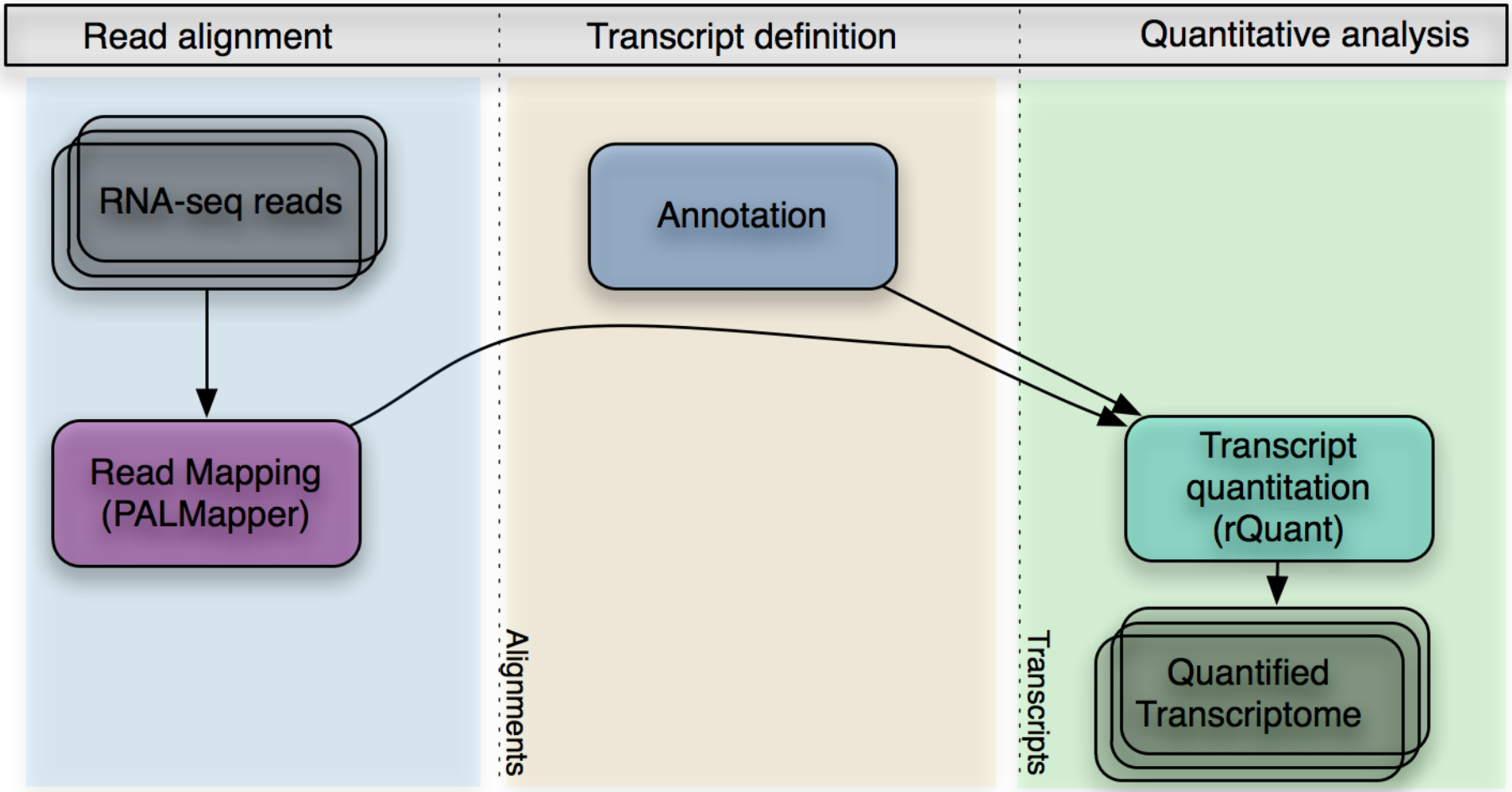


ShortSIG 2008: QPalma ($\sim$ Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff [Stegle et al., 2010a]

RNA-seq-based Transcriptome Analyses

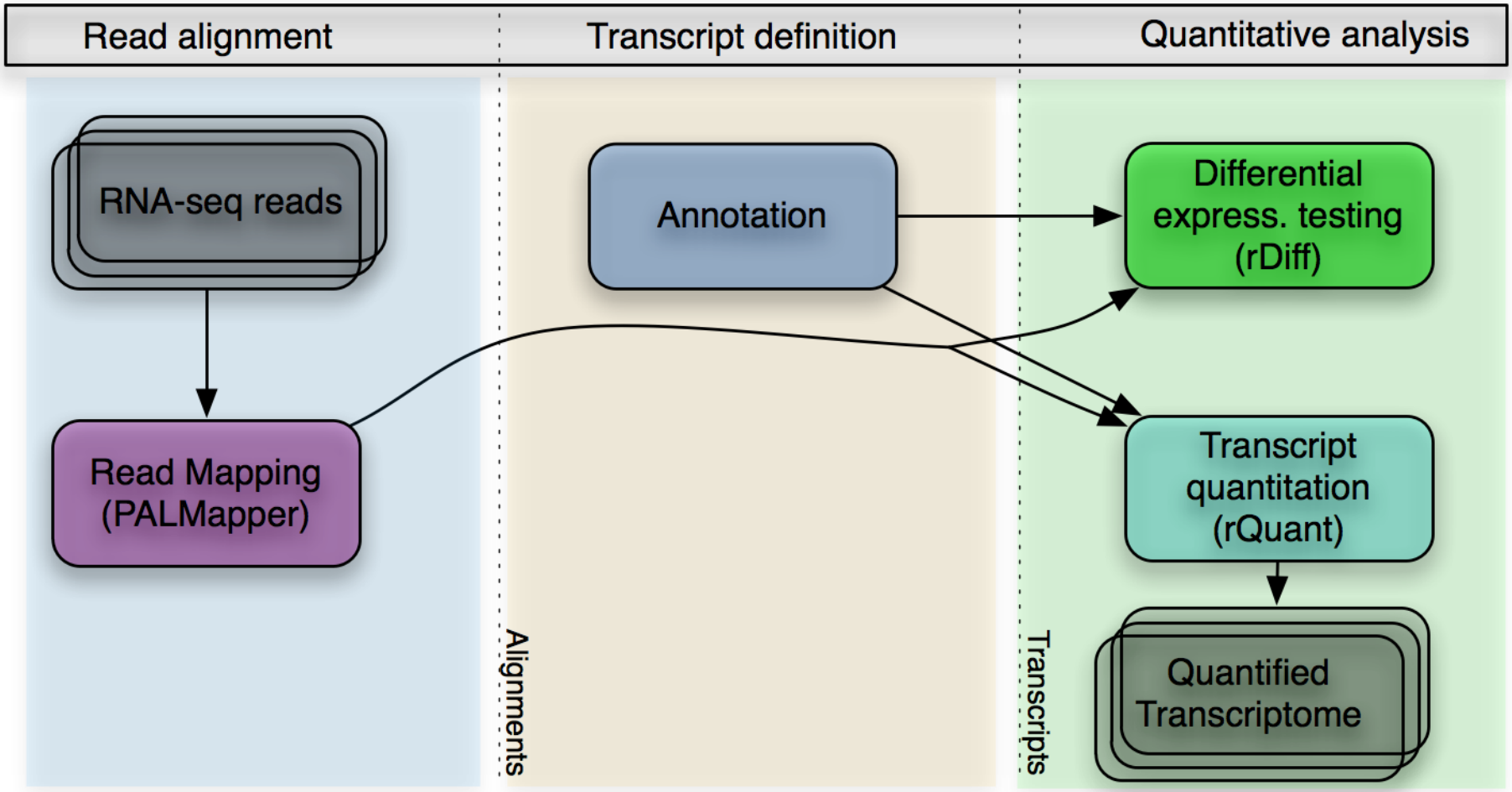


ShortSIG 2008: QPalma ($\sim$ Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff [Stegle et al., 2010a]

RNA-seq-based Transcriptome Analyses

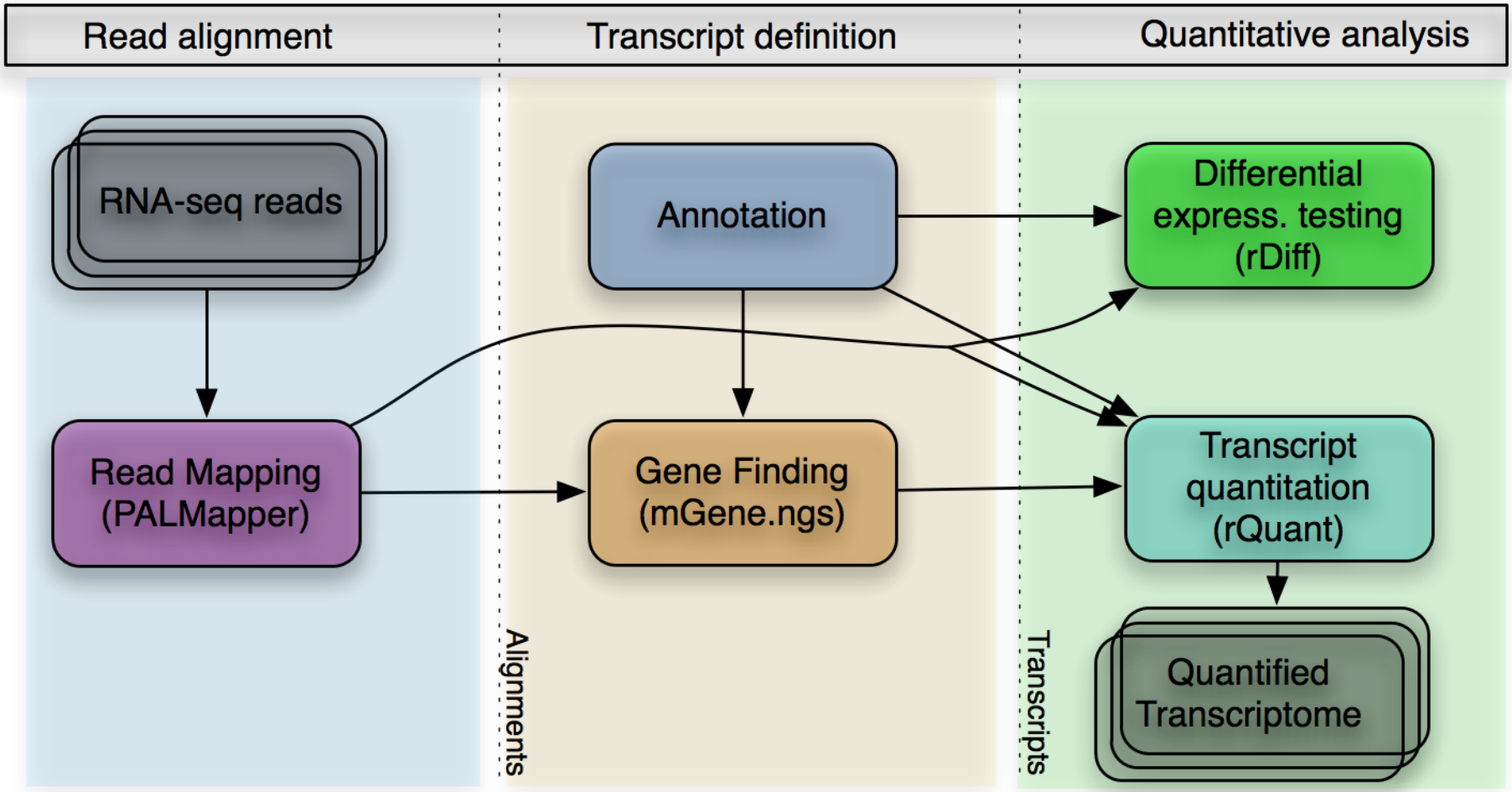


ShortSIG 2008: QPalma ($\sim$ Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff [Stegle et al., 2010a]

RNA-seq-based Transcriptome Analyses

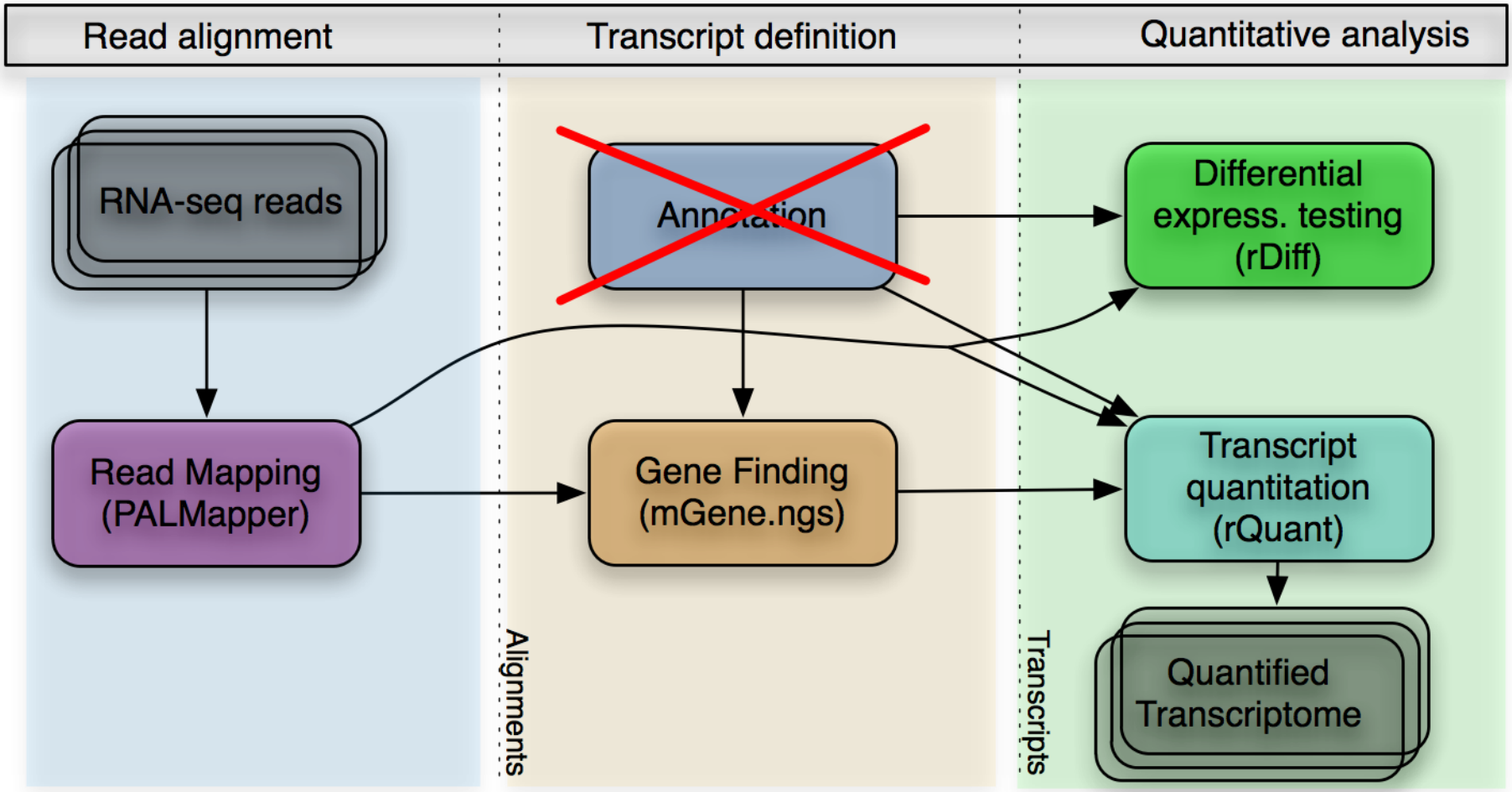


ShortSIG 2008: QPalma ($\sim$ Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff [Stegle et al., 2010a]

RNA-seq-based Transcriptome Analyses



ShortSIG 2008: QPalma ($\sim$ Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff [Stegle et al., 2010a]

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{"large margin"})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{"large margin"})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{"large margin"})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{“large margin”})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

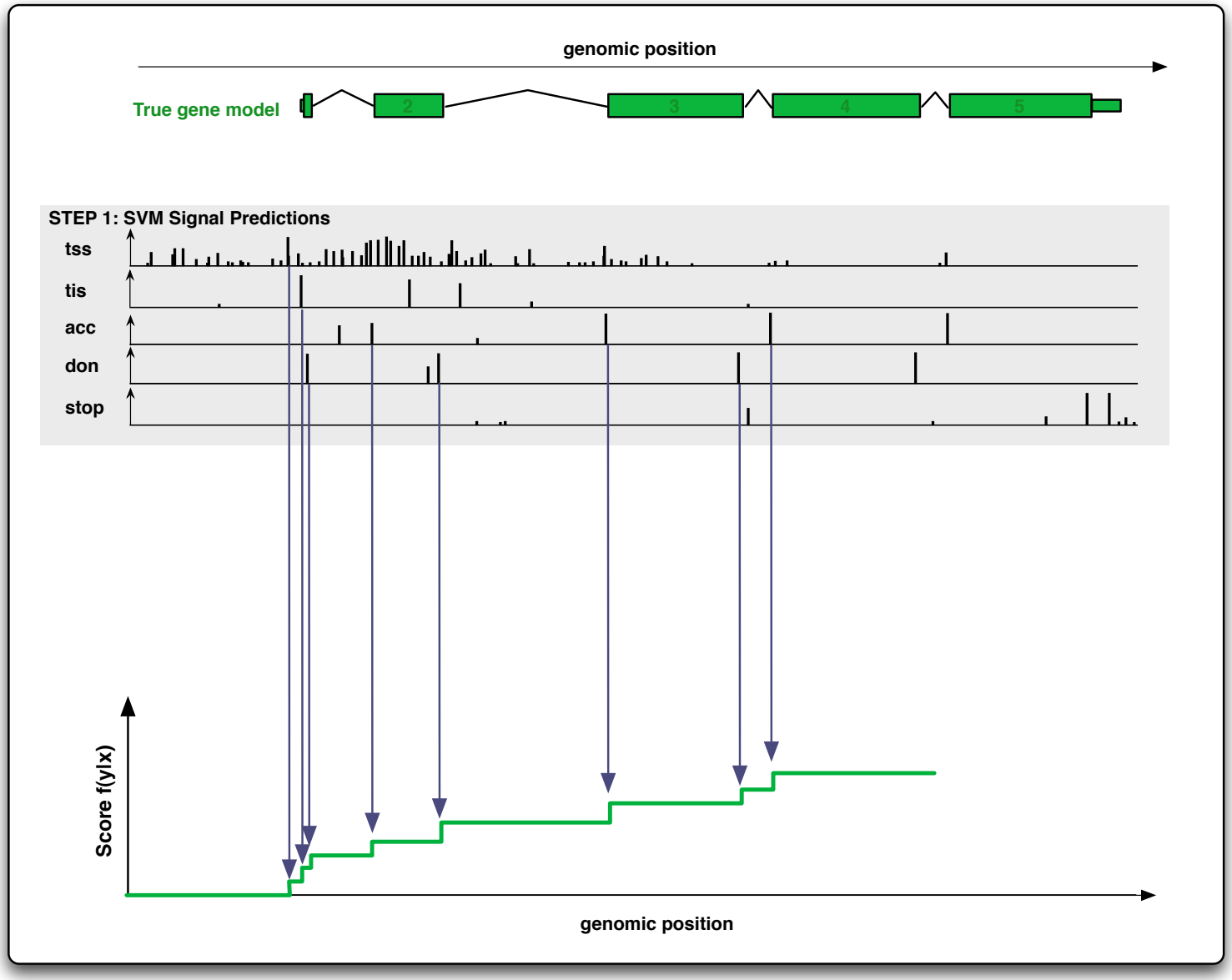
[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

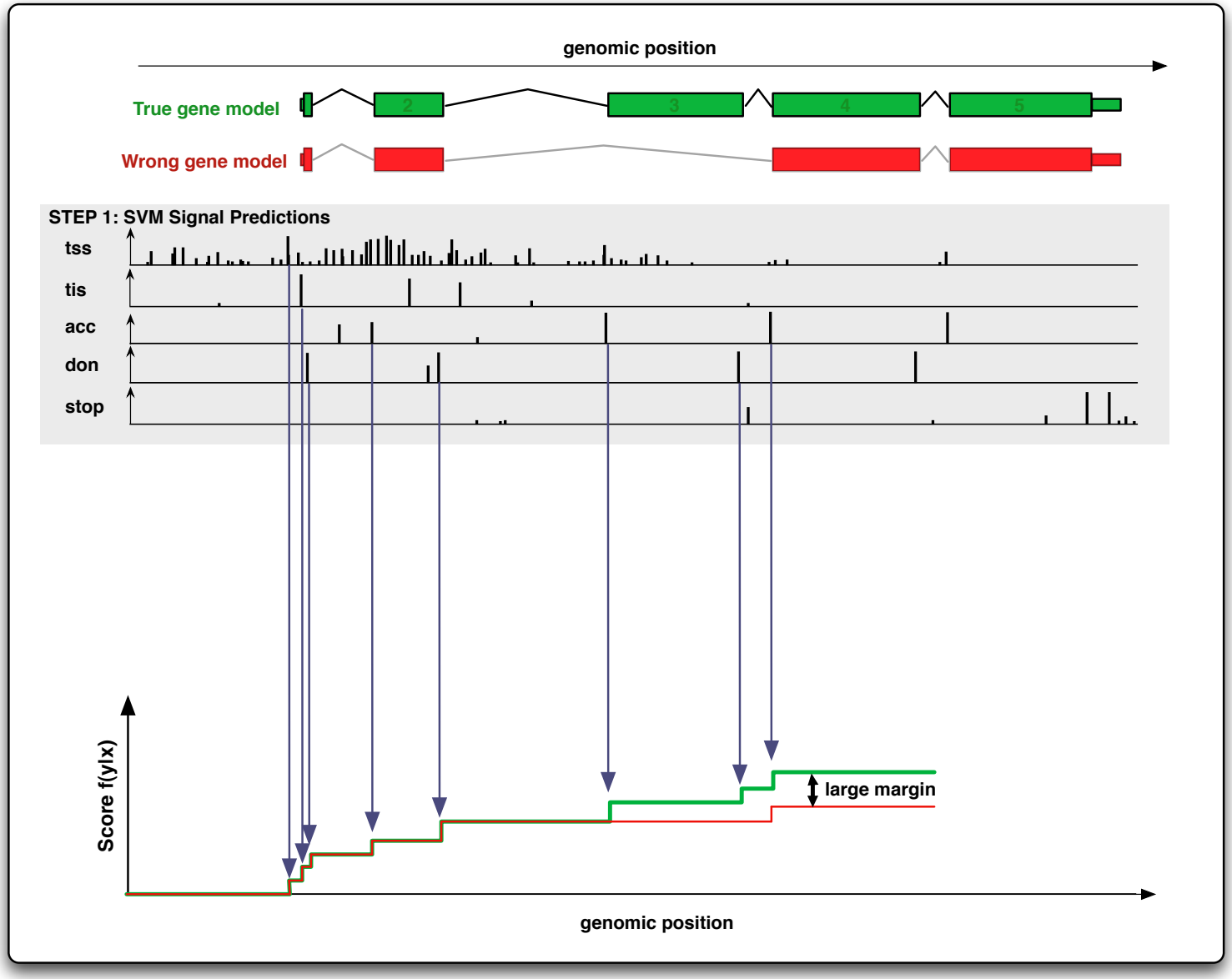
mGene.ngs Parameter Estimation



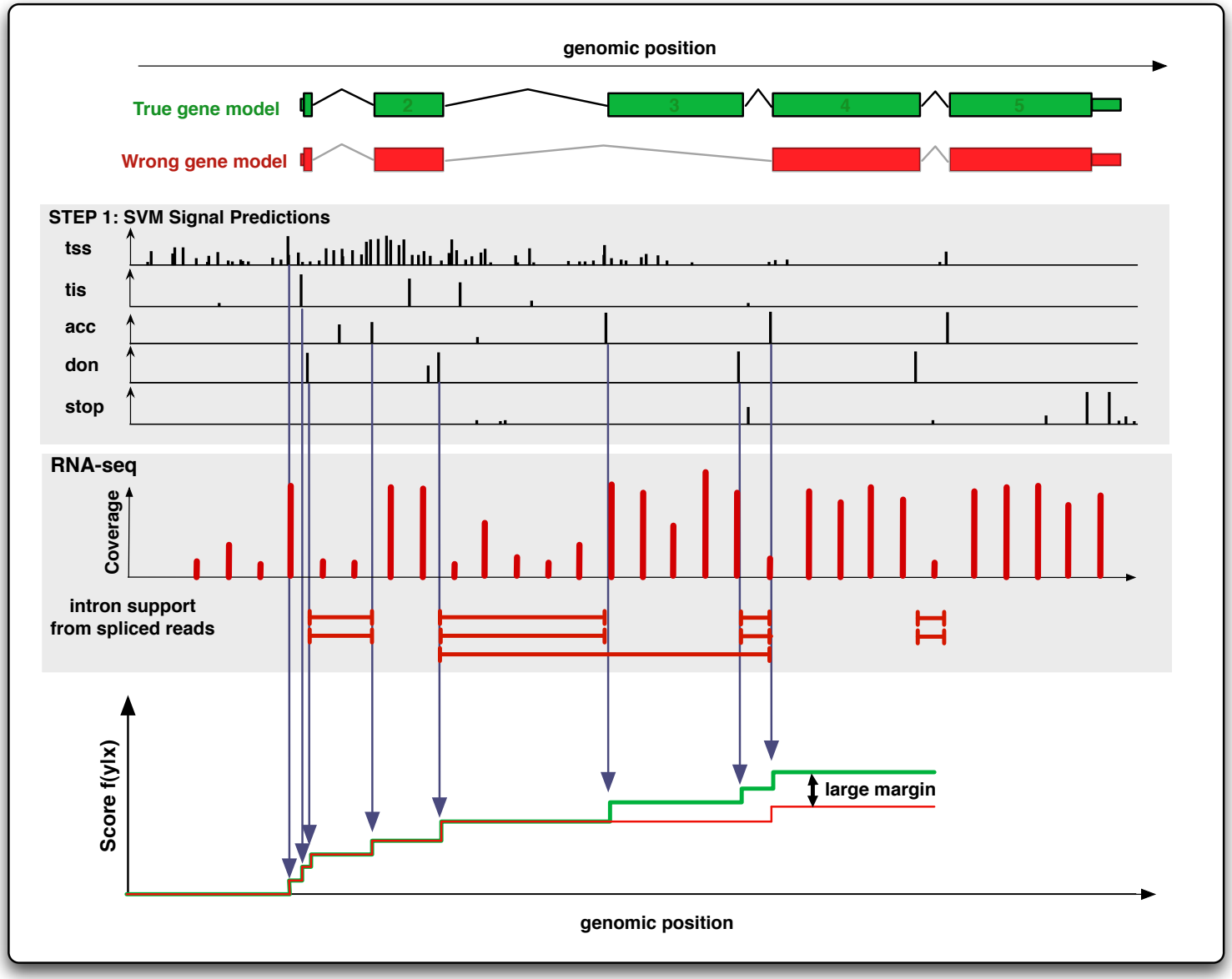
mGene.ngs Parameter Estimation



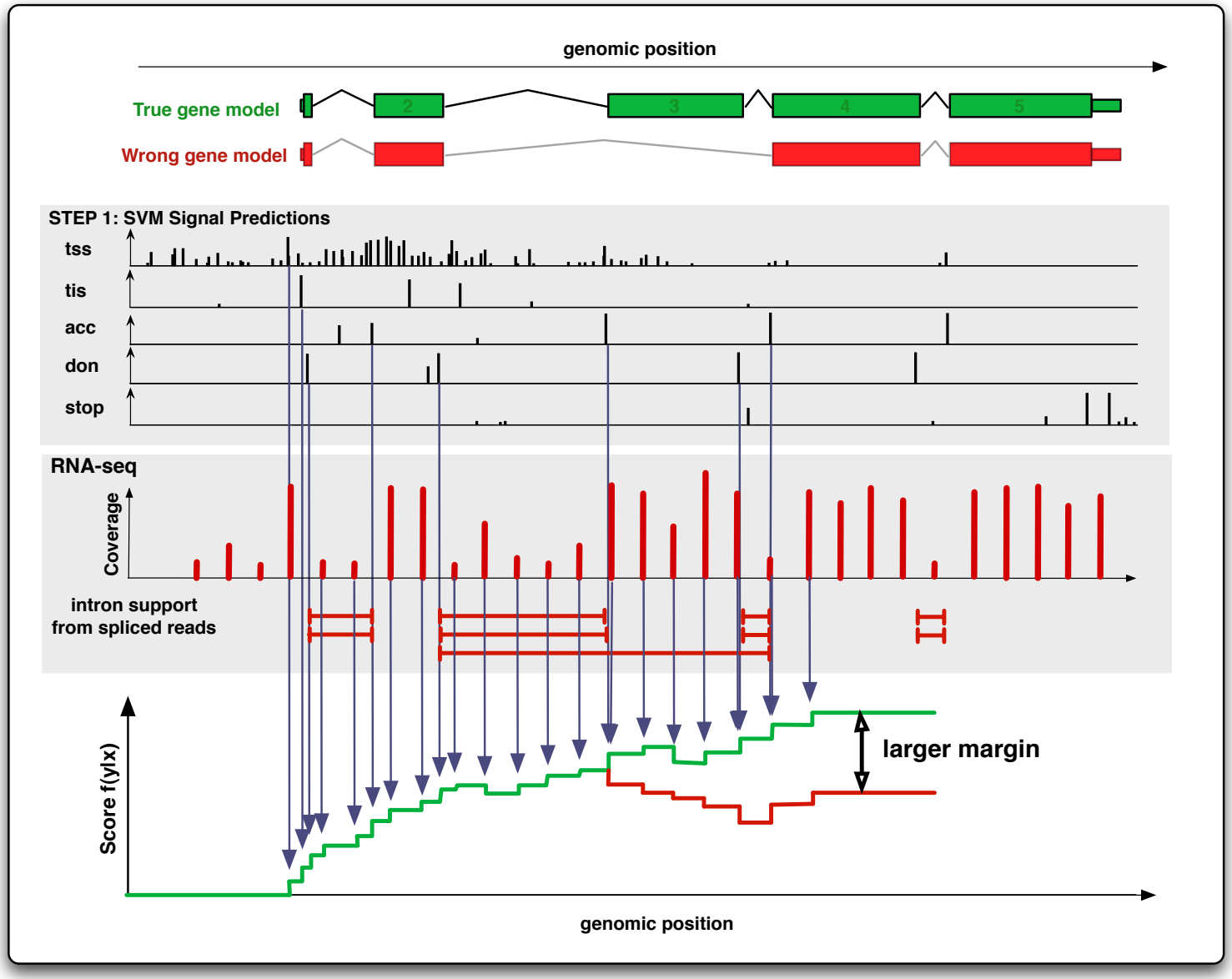
mGene.ngs Parameter Estimation



mGene.ngs Parameter Estimation



mGene.ngs Parameter Estimation



Results for C. elegans

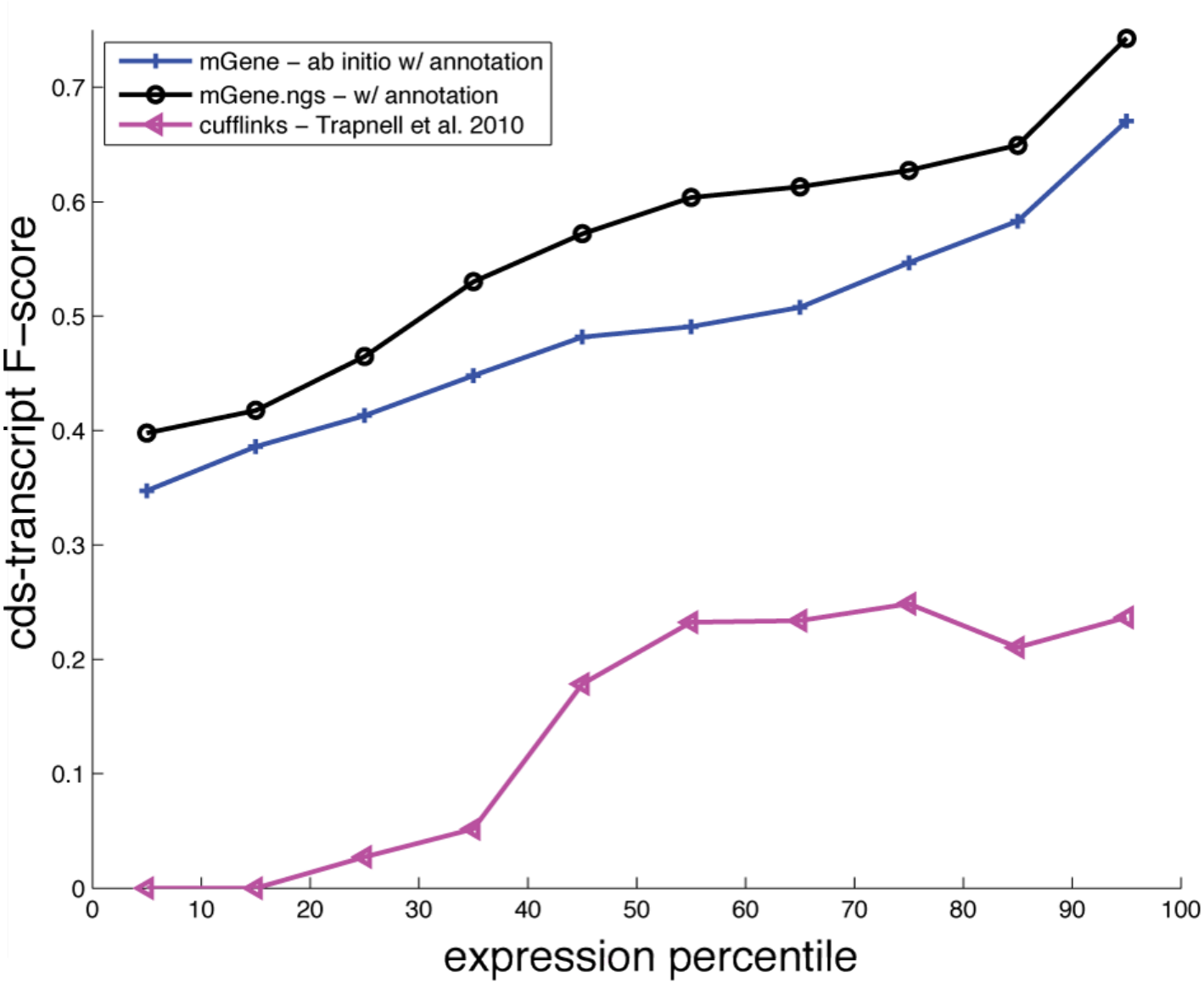
RNA-seq:

- paired-end, strand-specific RNA ligation based protocol
- 76bp reads, 50 million reads
- Alignment with Palmapper

Evaluation:

- Transcript-level F-score of coding transcripts
  ... for different expression levels
- Compare mGene (*ab initio*), mGene.ngs, cufflinks

Results for C. elegans



Digestion

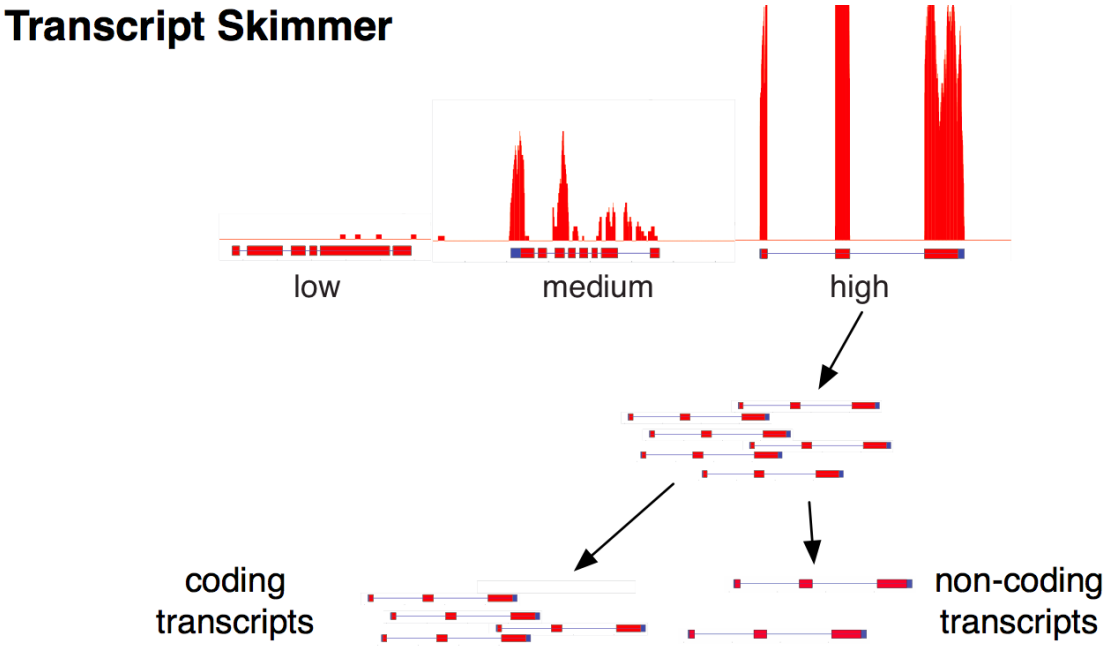
Observations:

- RNA-seq helps to improve performance
- Genomic signals help much (see cufflinks)

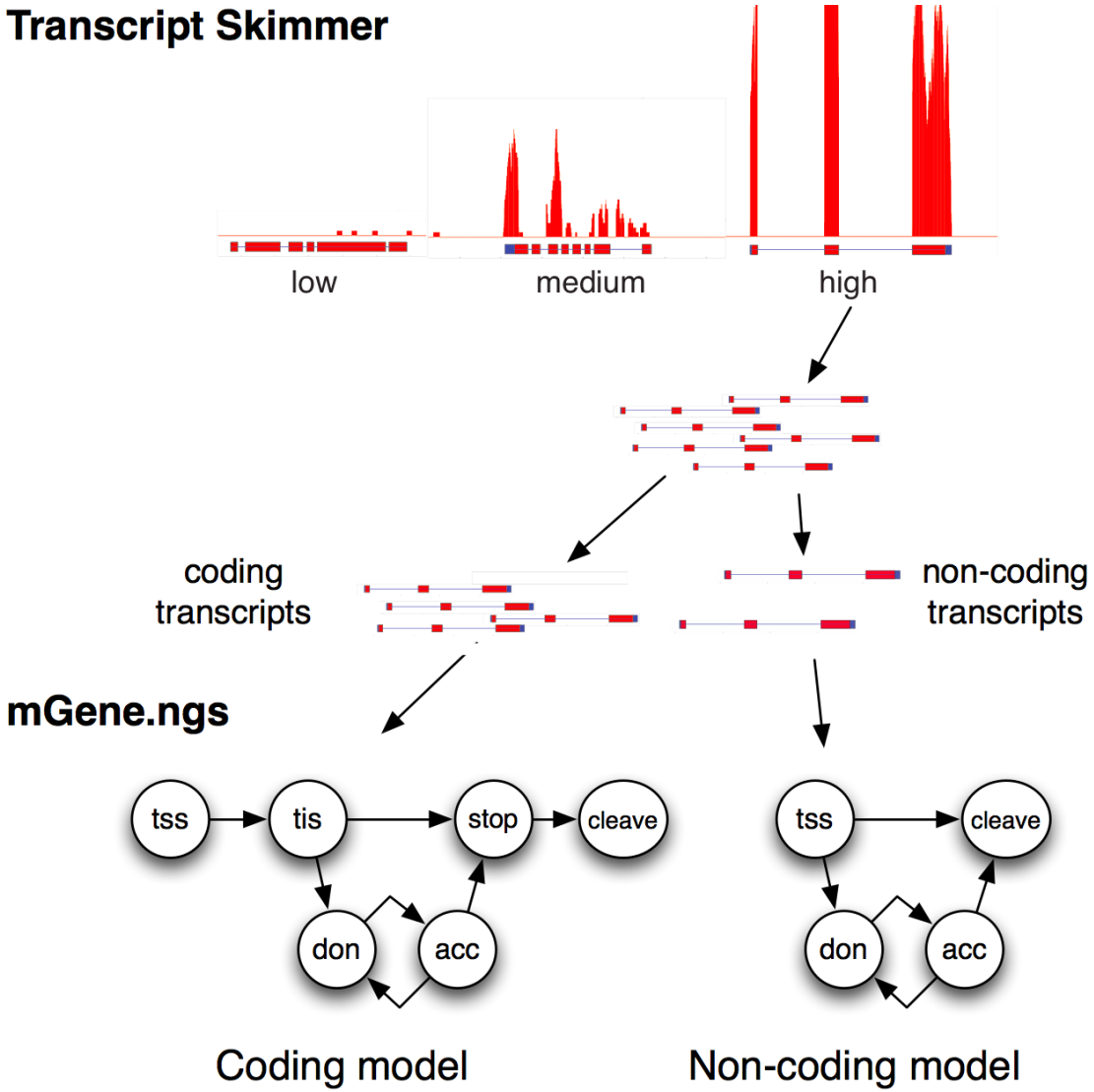
Problems:

- Need existing annotation for training
- Cannot predict non-coding transcripts

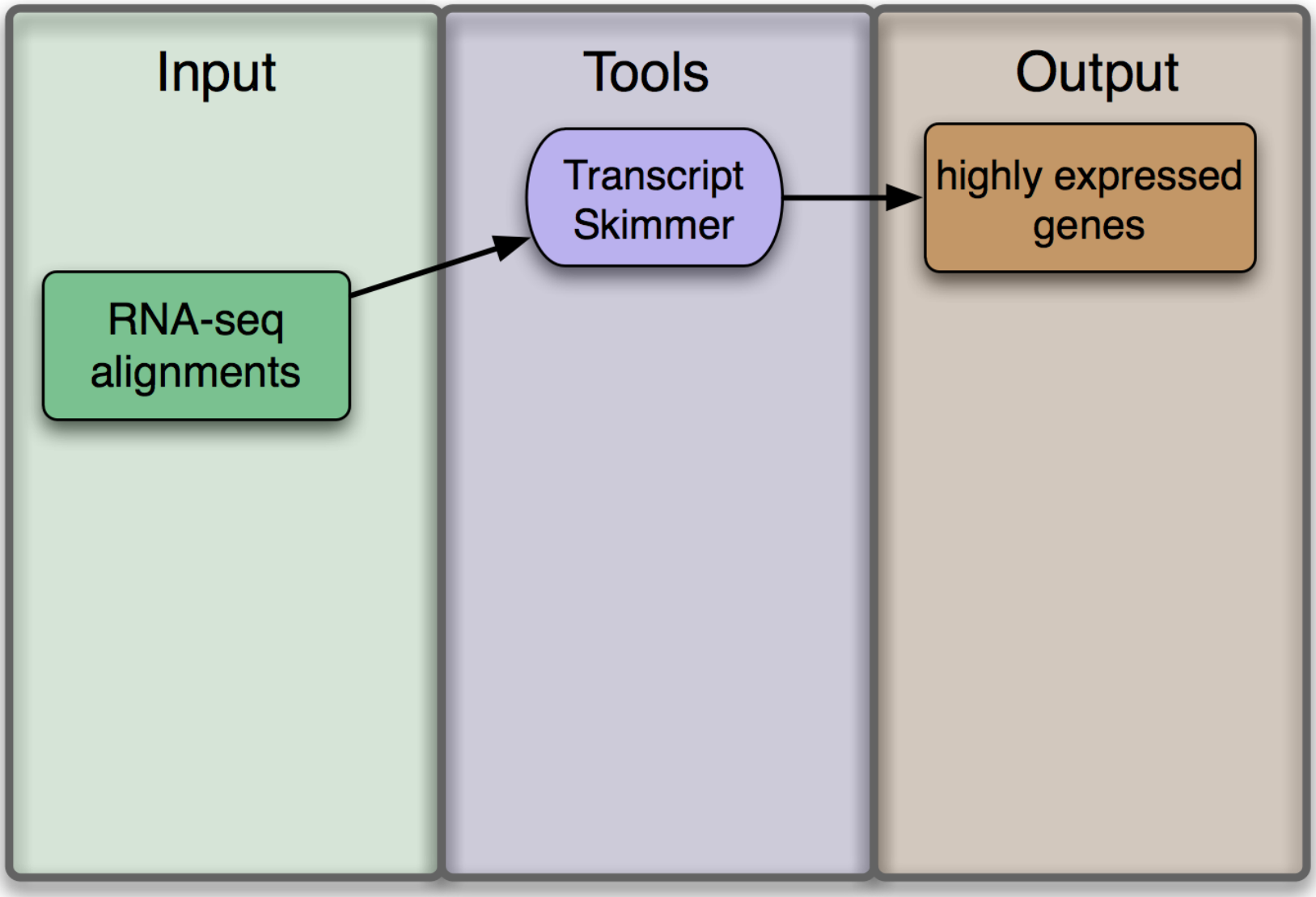
Skimming and Non-coding Transcripts



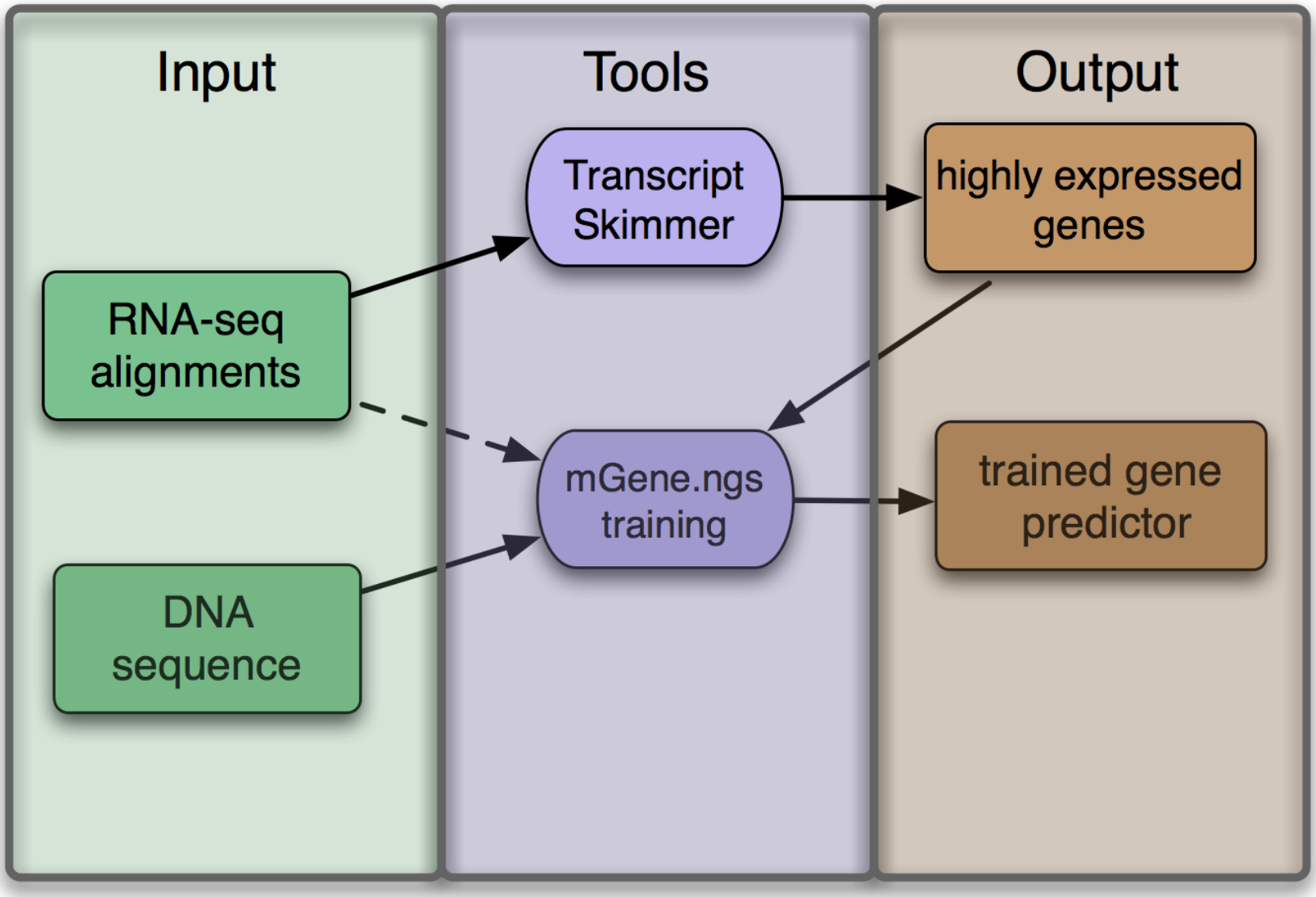
Skimming and Non-coding Transcripts



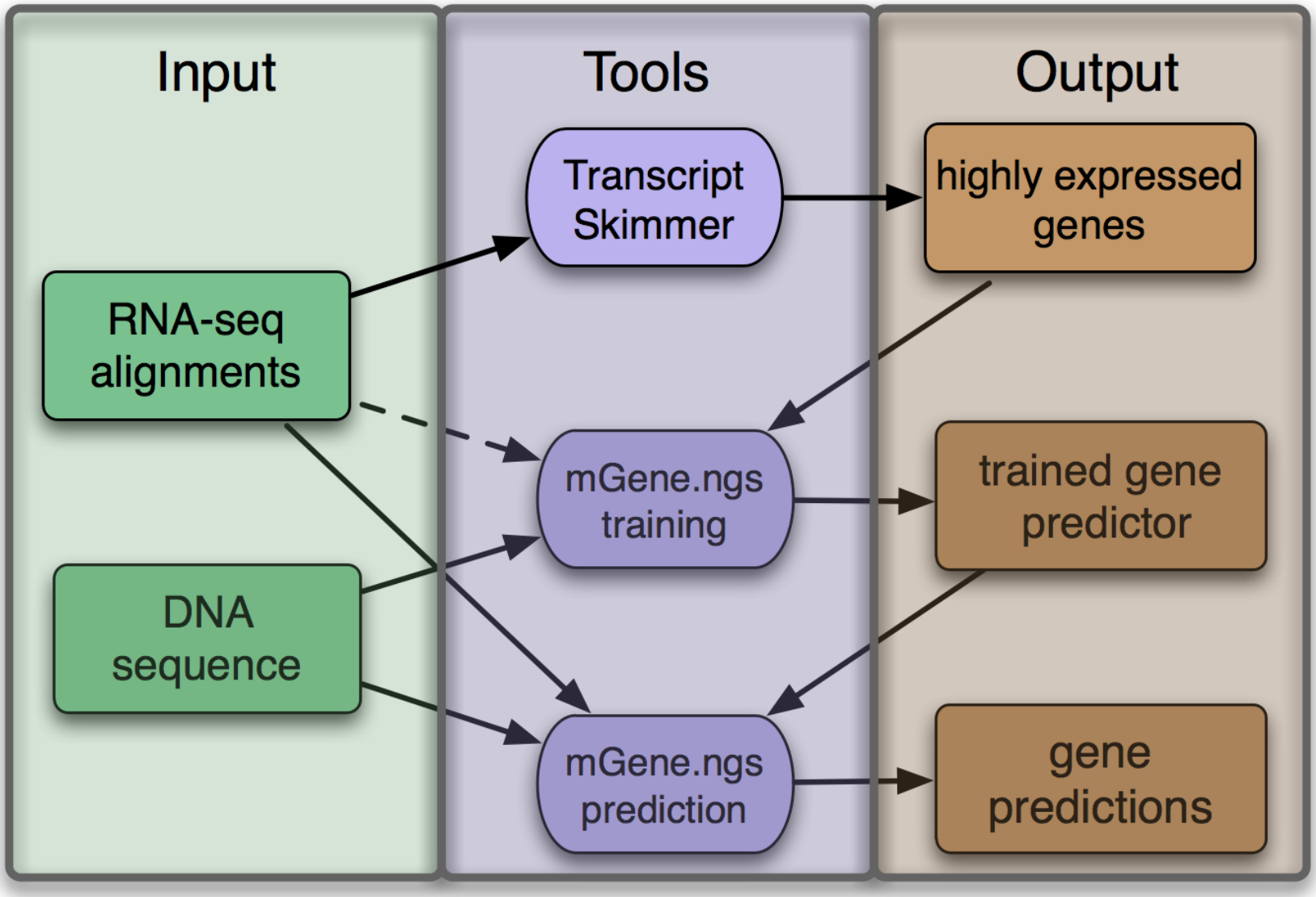
Learning Strategy



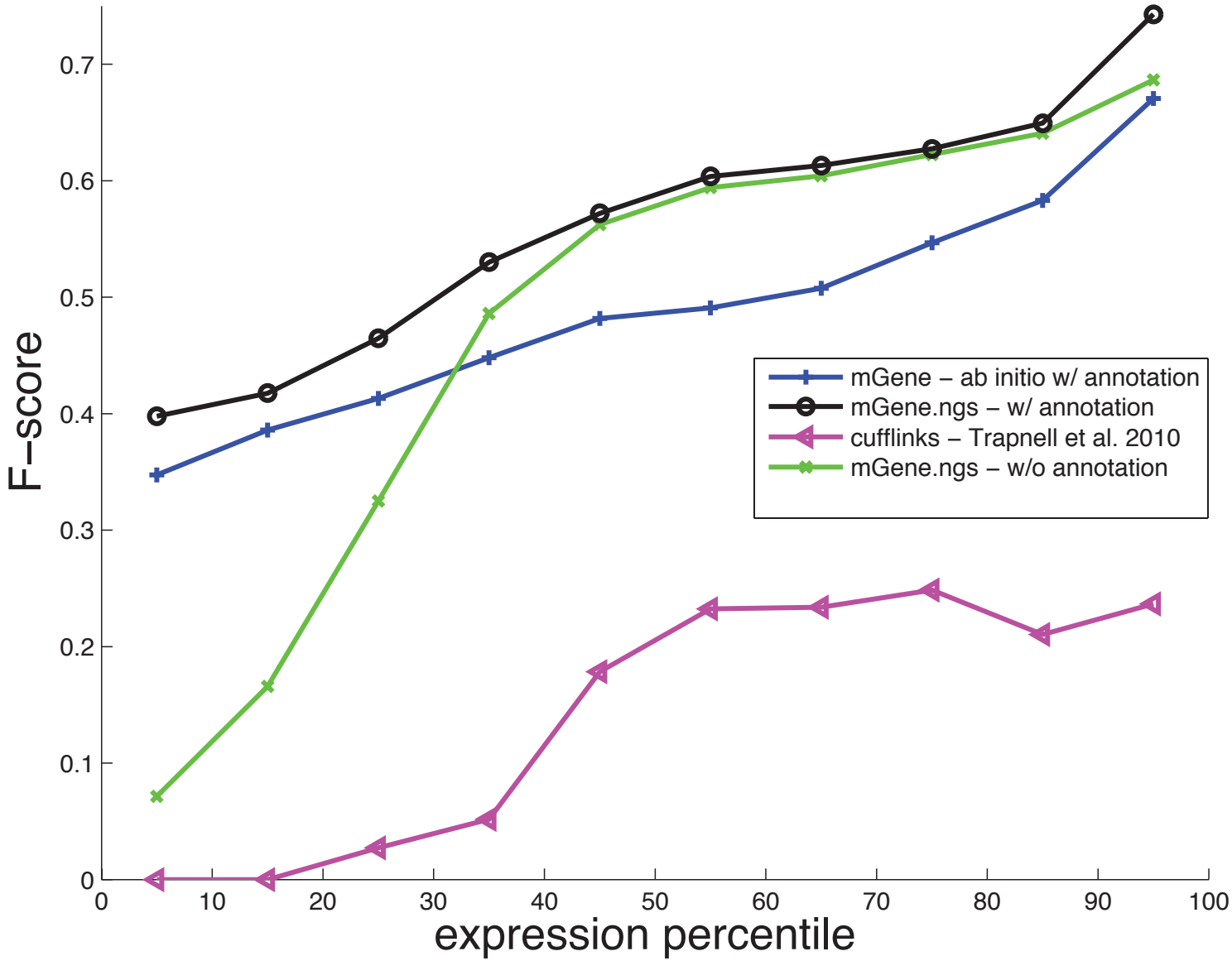
Learning Strategy



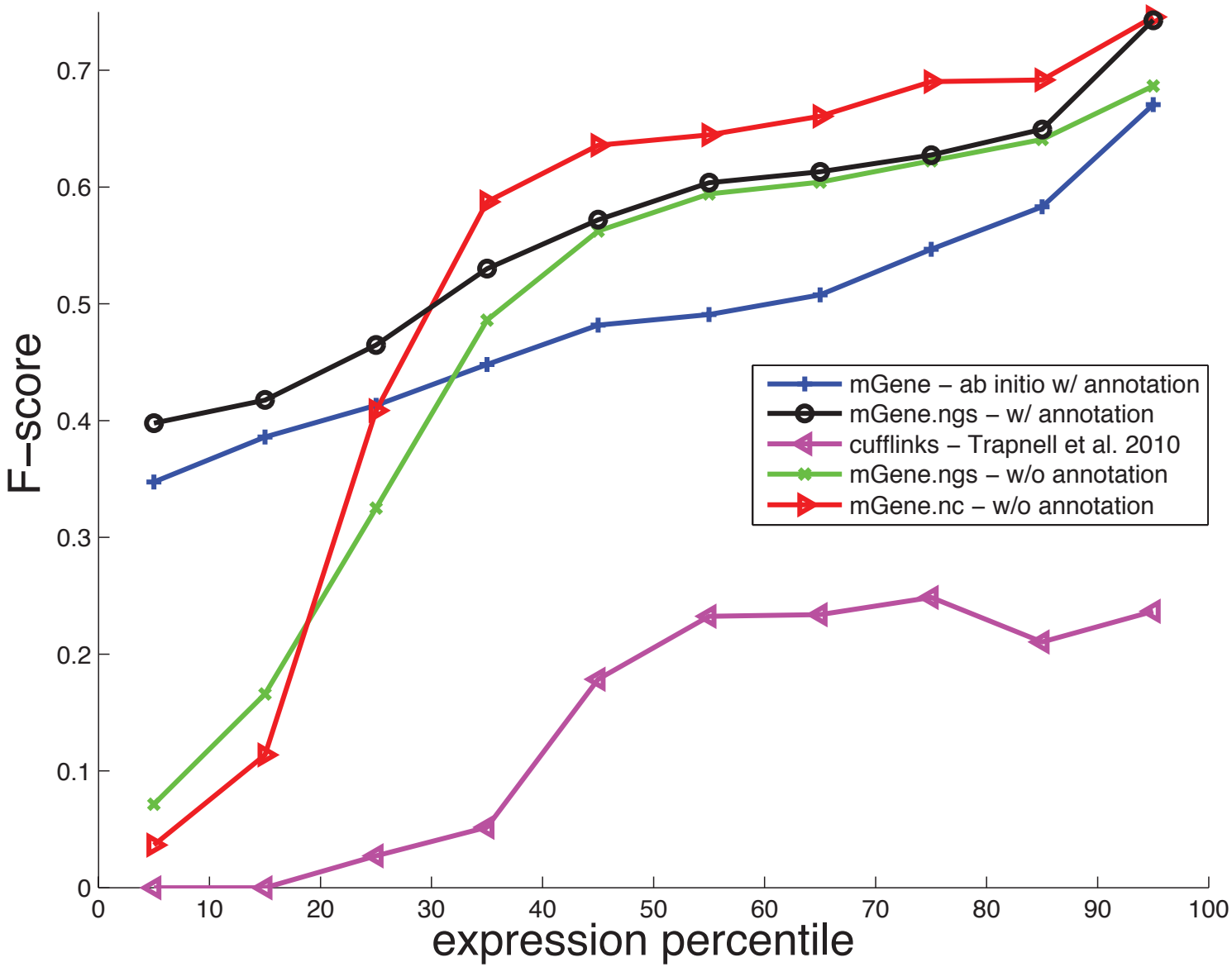
Learning Strategy



Results for C. elegans



Results for C. elegans



Discussion

Observations:

- 60-70% of expressed genes correct, RNA-seq improves by 15%
- Lowly expressed genes cannot be predicted (many GPCRs)
- Disagreements often in genes with alternative isoforms

Why just 60% correct (20% for human)?
What is missing?

- Current methods ignore other important information:
  - Chromatin structure and methylation patterns,
  - RNA structure and processing regulation . . .
- *Ab initio* methods cannot predict alternative transcripts
- How much of the annotation is actually exactly correct?

Discussion

Observations:

- 60-70% of expressed genes correct, RNA-seq improves by 15%
- Lowly expressed genes cannot be predicted (many GPCRs)
- Disagreements often in genes with alternative isoforms

Why just 60% correct (20% for human)?
What is missing?

- Current methods ignore other important information:
  - Chromatin structure and methylation patterns,
  - RNA structure and processing regulation . . .
- *Ab initio* methods cannot predict alternative transcripts
- How much of the annotation is actually exactly correct?

Discussion

Observations:

- 60-70% of expressed genes correct, RNA-seq improves by 15%
- Lowly expressed genes cannot be predicted (many GPCRs)
- Disagreements often in genes with alternative isoforms

Why just 60% correct (20% for human)?
What is missing?

- Current methods ignore other important information:
  - Chromatin structure and methylation patterns,
  - RNA structure and processing regulation ...
- *Ab initio* methods cannot predict alternative transcripts
- How much of the annotation is actually exactly correct?

Gene Finding + Auxiliary Measurements

[Behr et al., 2008]

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%**
5. . . . with tiling arrays (11 tissues) **82.1%**
6. . . . with all information combined **84.3%**



[Wikipedia]

Measuring vs. modeling transcriptomes

Gene Finding + Auxiliary Measurements

[Behr et al., 2008]

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%**
5. . . . with tiling arrays (11 tissues) **82.1%**
6. . . . with all information combined **84.3%**



[Wikipedia]

Measuring vs. modeling transcriptomes

Gene Finding + Auxiliary Measurements

[Behr et al., 2008]

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%**
5. . . . with tiling arrays (11 tissues) **82.1%**
6. . . . with all information combined **84.3%**



[Wikipedia]

Measuring vs. modeling transcriptomes

Gene Finding + Auxiliary Measurements

[Behr et al., 2008]

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%**
5. . . . with tiling arrays (11 tissues) **82.1%**
6. . . . with all information combined **84.3%**



[Wikipedia]

Measuring vs. modeling transcriptomes

Gene Finding + Auxiliary Measurements

[Behr et al., 2008]

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%**
5. . . . with tiling arrays (11 tissues) **82.1%**
6. . . . with all information combined **84.3%**

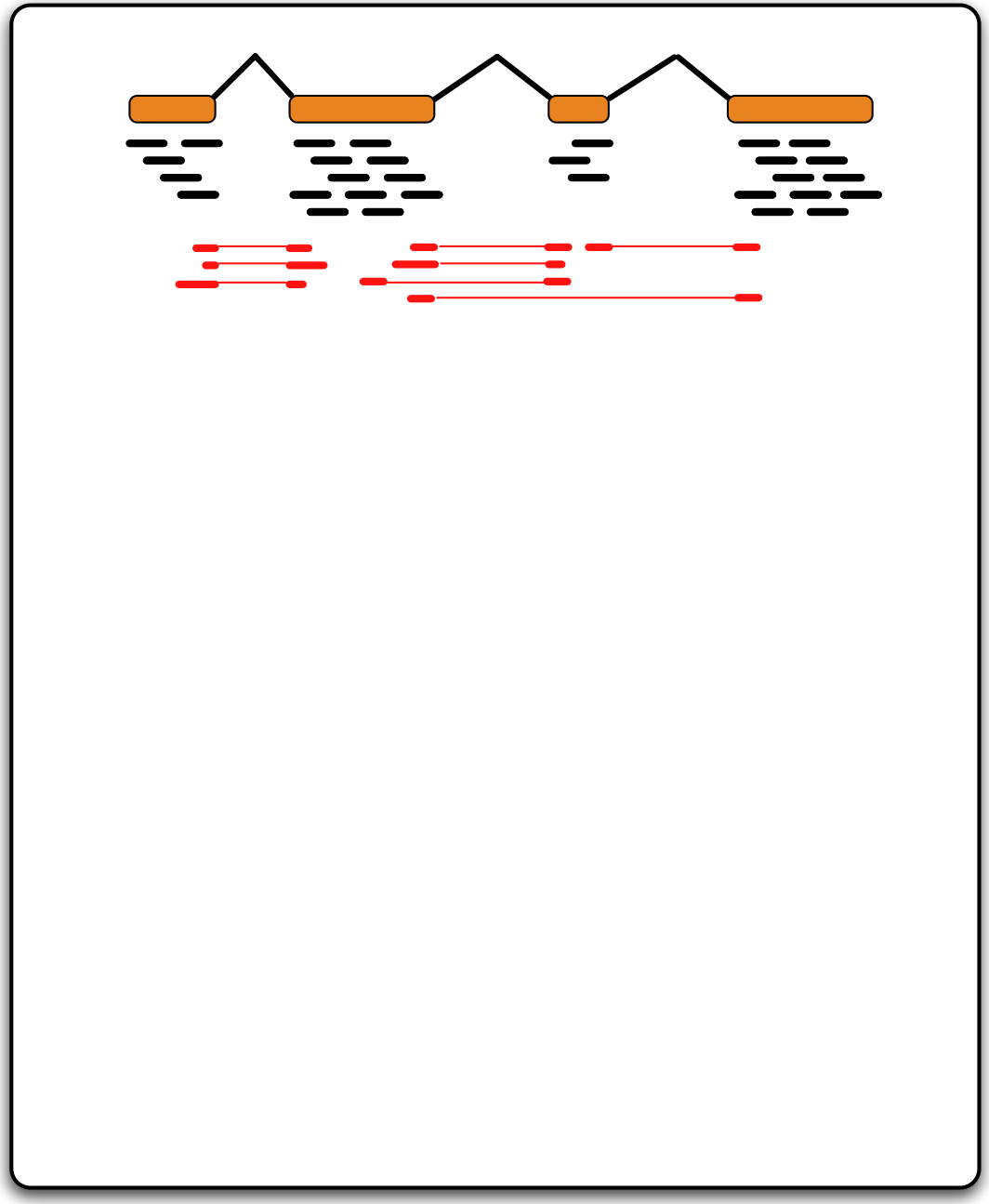


[Wikipedia]

Measuring vs. modeling transcriptomes

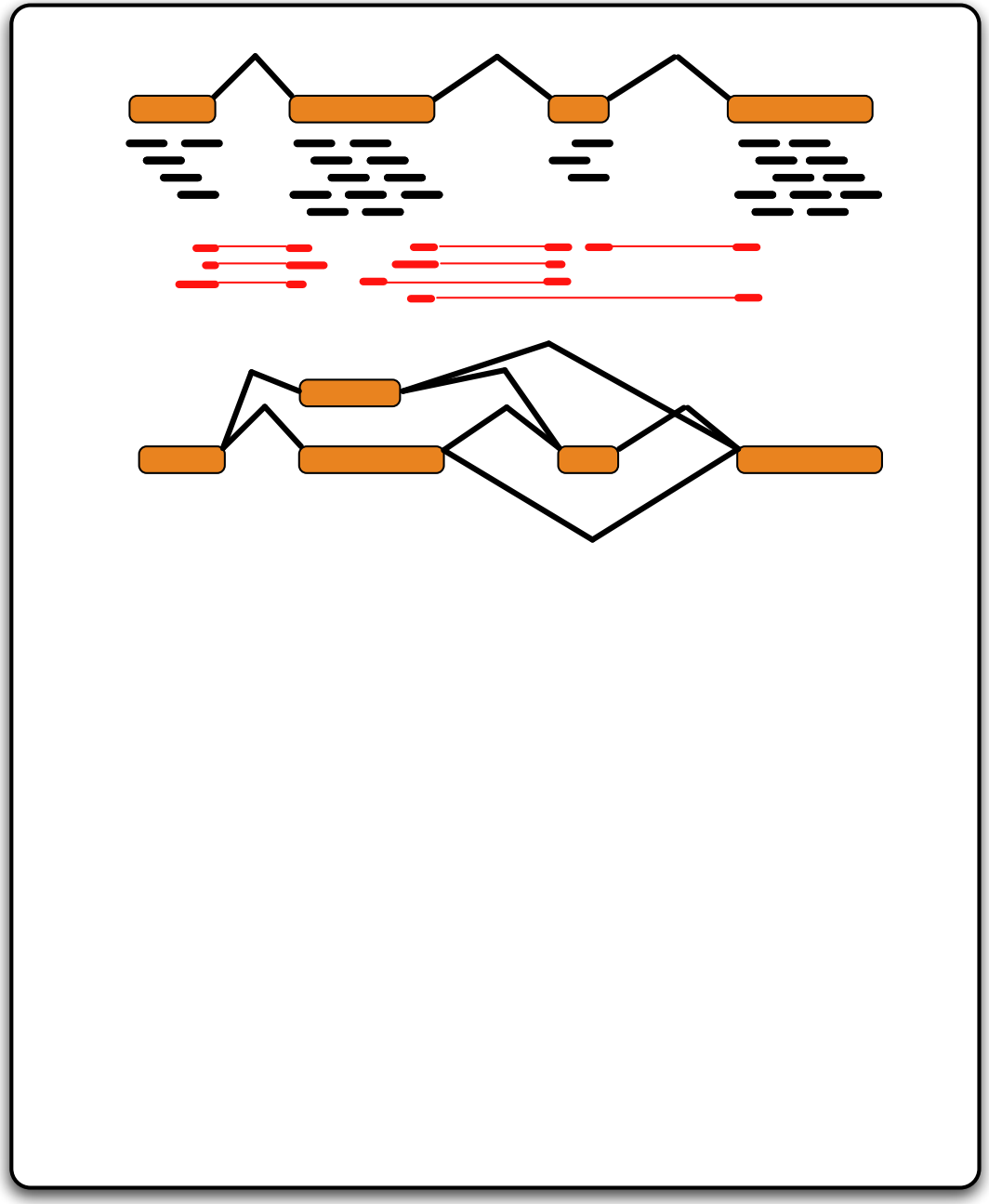
Infer alternative isoforms

- mGene.ngs prediction
- Build splicing graph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



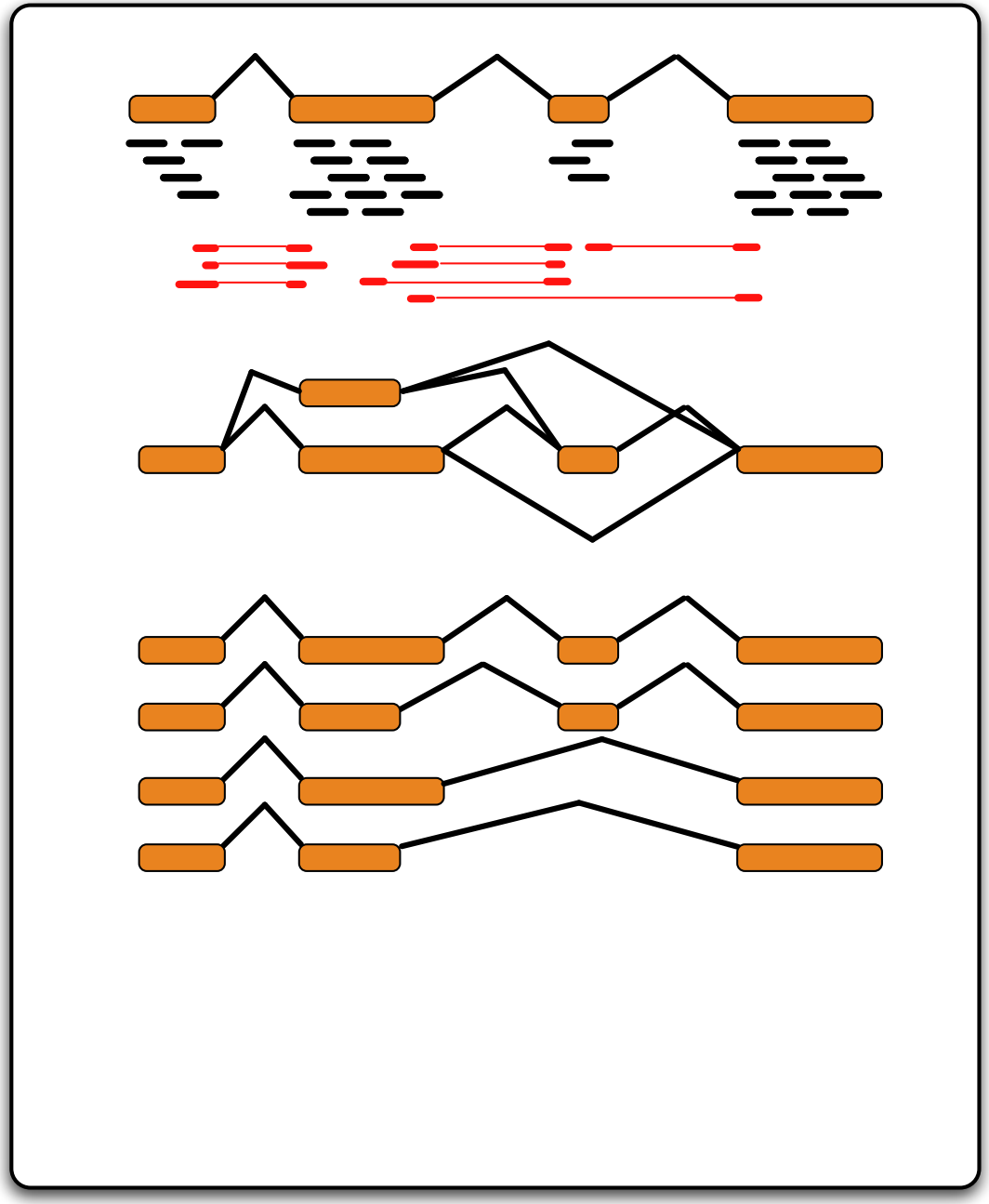
Infer alternative isoforms

- mGene.ngs prediction
- Build splicing graph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



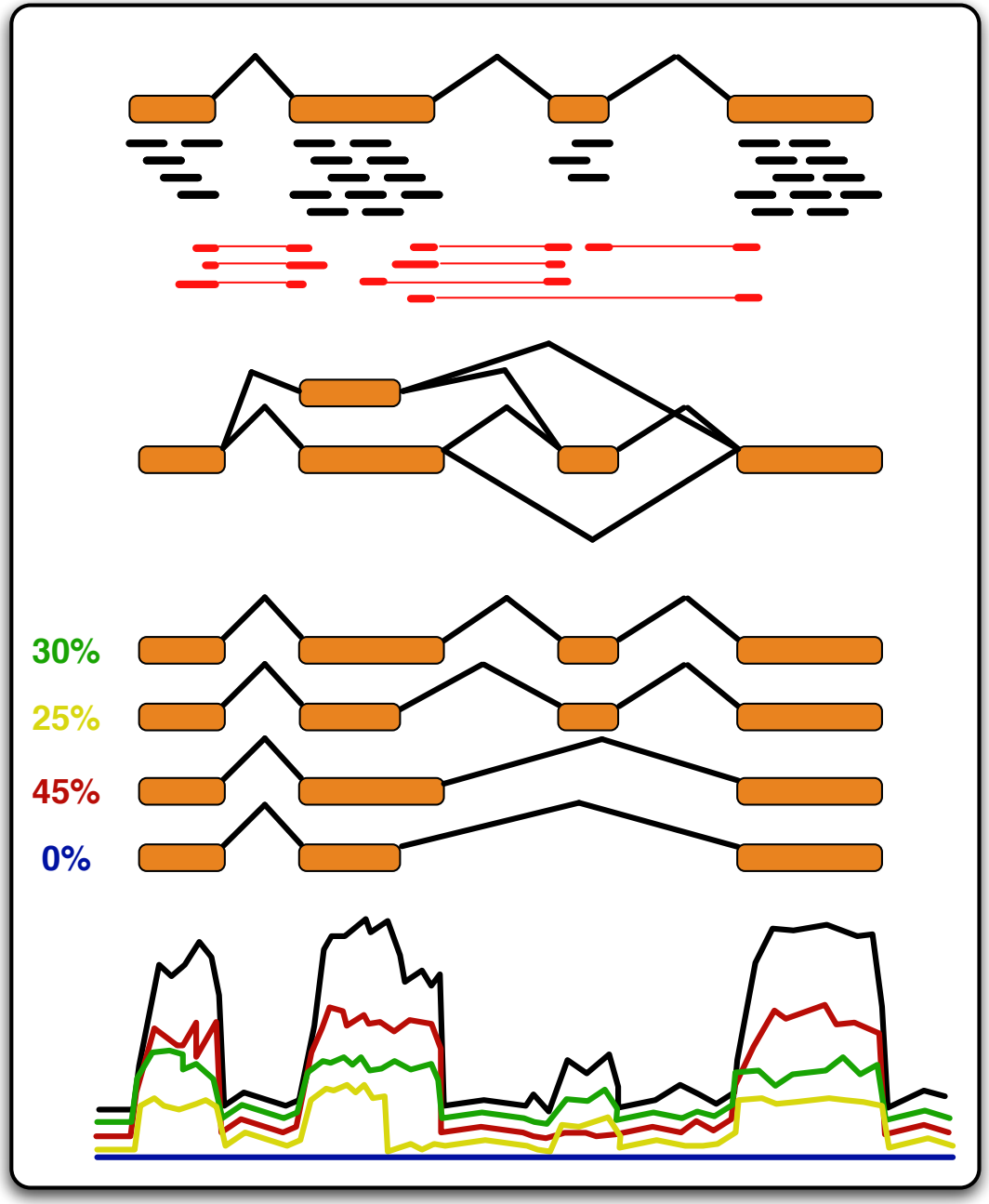
Infer alternative isoforms

- mGene.ngs prediction
- Build splicing graph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



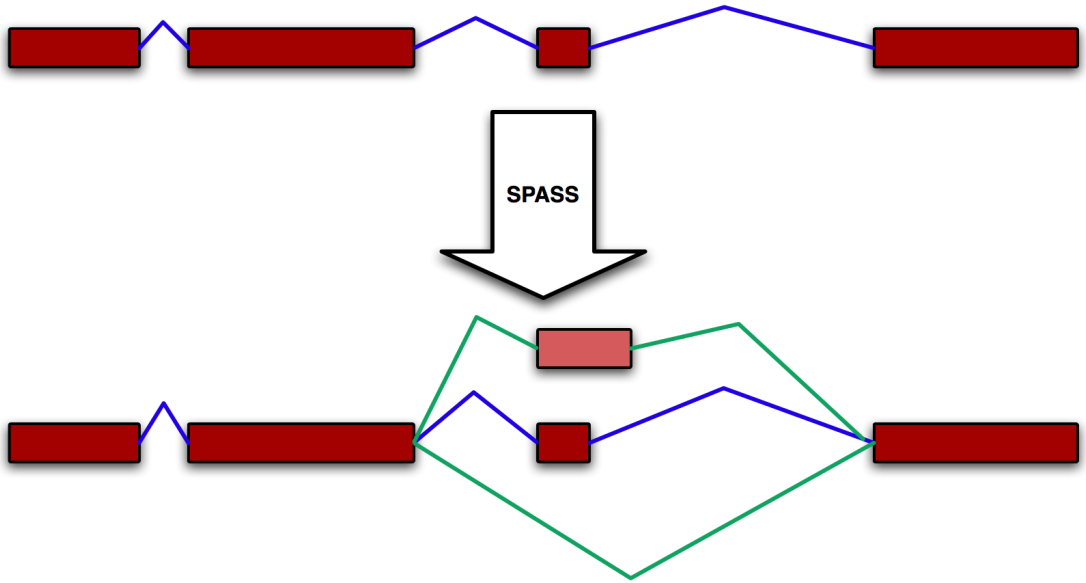
Infer alternative isoforms

- mGene.ngs prediction
- Build splicing graph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



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

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"**

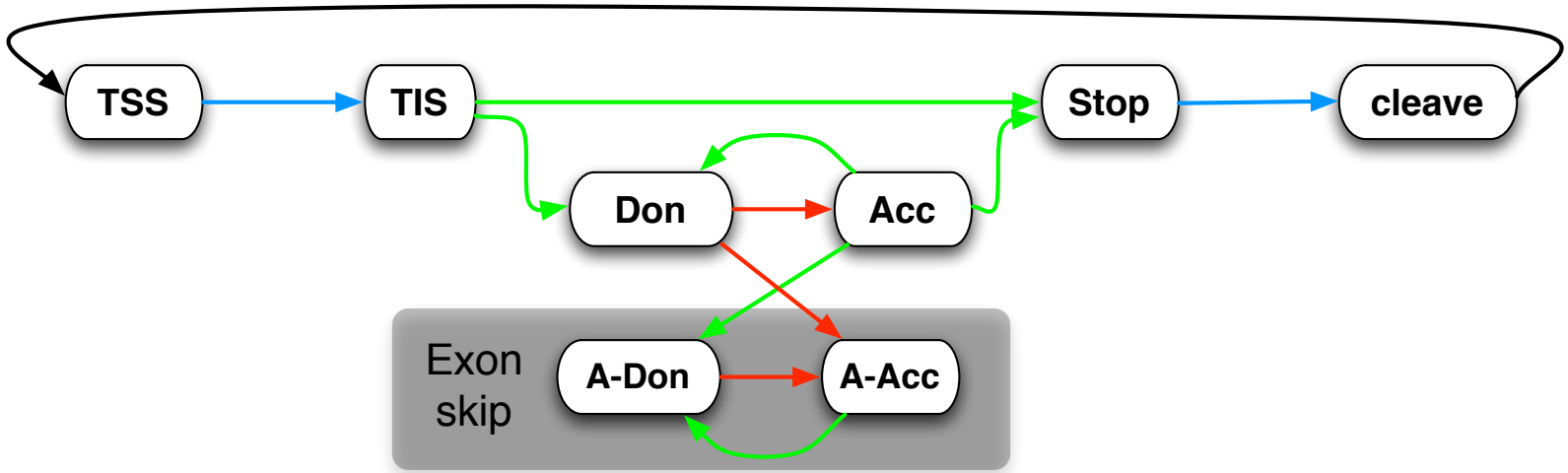
- Now we can use RNA-seq data to train such models ...

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"**



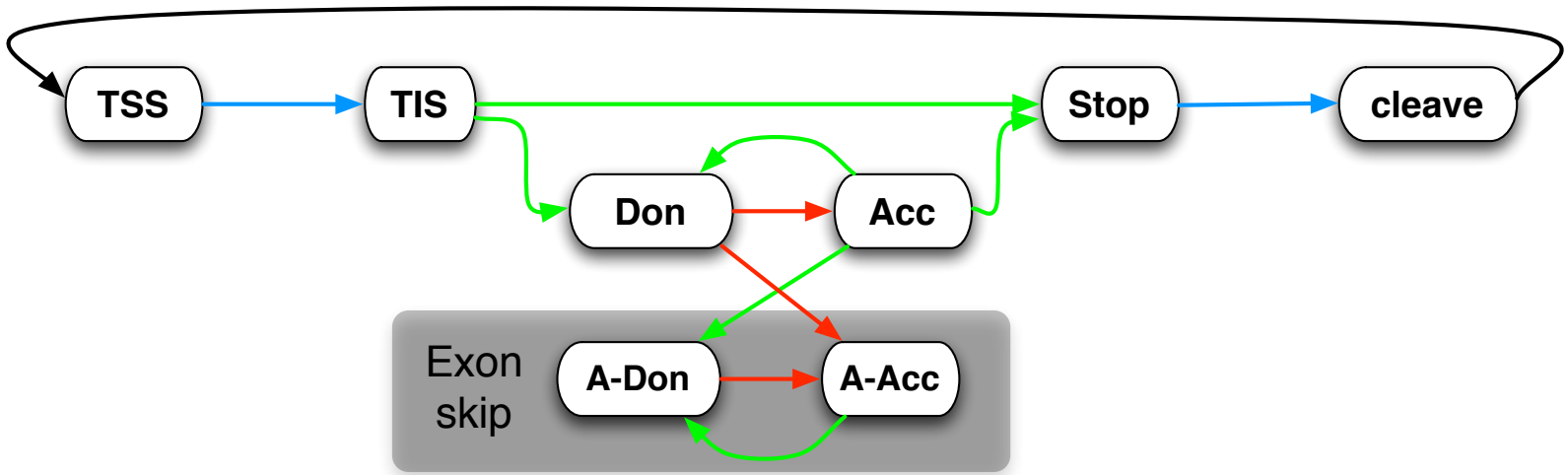
- Now we can use RNA-seq data to train such models ...

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"**



- Now we can use RNA-seq data to train such models ...

Application: Annotate 19 Arabidopsis Strains

Collaborative project:

- Richard Mott (U Cambridge), Richard Clark (U Utah), ...
- Data
  - DNA Sequencing and reference-guided assembly of 18 *Arabidopsis thaliana* genomes + reference ("Col_0")
    - founder strains for large collection of multi-parent (MAGIC) lines
  - RNA sequencing of young seedling RNA of
    - 19 founder strains (2 biological replicates)
    - 22 MAGIC lines (1 measurement)
- Questions
  - Extent of polymorphism, variant numbers, genomic patterns
  - Understand functional consequences of polymorphisms
    - Large- and small-effect changes
    - Loss-of-function or alternative annotation?
  - Understanding differential gene expression
    - Patterns, effect of polymorphisms, eQTLs, heritability

Application: Annotate 19 Arabidopsis Strains

Collaborative project:

- Richard Mott (U Cambridge), Richard Clark (U Utah), ...
- Data
  - DNA Sequencing and reference-guided assembly of 18 *Arabidopsis thaliana* genomes + reference ("Col_0")
    - founder strains for large collection of multi-parent (MAGIC) lines
  - RNA sequencing of young seedling RNA of
    - 19 founder strains (2 biological replicates)
    - 22 MAGIC lines (1 measurement)
- Questions
  - Extent of polymorphism, variant numbers, genomic patterns
  - Understand functional consequences of polymorphisms
    - Large- and small-effect changes
    - Loss-of-function or alternative annotation?
  - Understanding differential gene expression
    - Patterns, effect of polymorphisms, eQTLs, heritability

Application: Annotate 19 Arabidopsis Strains

Collaborative project:

- Richard Mott (U Cambridge), Richard Clark (U Utah), ...
- Data
  - DNA Sequencing and reference-guided assembly of 18 *Arabidopsis thaliana* genomes + reference ("Col_0")
    - founder strains for large collection of multi-parent (MAGIC) lines
  - RNA sequencing of young seedling RNA of
    - 19 founder strains (2 biological replicates)
    - 22 MAGIC lines (1 measurement)
- Questions
  - Extent of polymorphism, variant numbers, genomic patterns
  - Understand functional consequences of polymorphisms
    - Large- and small-effect changes
    - Loss-of-function or alternative annotation?
  - Understanding differential gene expression
    - Patterns, effect of polymorphisms, eQTLs, heritability

Large-effects and Loss-of-function



Large effects of polymorphisms:

- Signal site consensus disruption
- Frame-shifts or premature stop codons

To which extend do "large effects" lead to disrupted gene products?

- Are there alternative, plausible gene models?

Strategy:

- Find gene with large effect change
- Predict gene model for locus using mGene.ngs
- Compare gene product in reference genome and alternative gene product on strain's genome (amino acid similarity $\geq 90\%$)

Large-effects and Loss-of-function



Large effects of polymorphisms:

- Signal site consensus disruption
- Frame-shifts or premature stop codons

To which extend do "large effects" lead to disrupted gene products?

- Are there alternative, plausible gene models?

Strategy:

- Find gene with large effect change
- Predict gene model for locus using mGene.ngs
- Compare gene product in reference genome and alternative gene product on strain's genome (amino acid similarity $\geq 90\%$)

Large-effects and Loss-of-function



Large effects of polymorphisms:

- Signal site consensus disruption
- Frame-shifts or premature stop codons

To which extend do "large effects" lead to disrupted gene products?

- Are there alternative, plausible gene models?

Strategy:

- Find gene with large effect change
- Predict gene model for locus using mGene.ngs
- Compare gene product in reference genome and alternative gene product on strain's genome (amino acid similarity $\geq 90\%$)

Summary

- Accurate classification of millions/billions of sequences is possible
- Analysis tools for RNA-seq data: from alignments to significant results
- *De novo* annotation of newly sequenced genomes by combining accurate transcriptome models and RNA-seq
- Auxiliary measurements improve transcriptome models
- Alternative gene models for large effects polymorphisms often exist
- *cis*-effects. Next: *trans*-effects and quantitative modeling

Summary

- Accurate classification of millions/billions of sequences is possible
- Analysis tools for RNA-seq data: from alignments to significant results
- *De novo* annotation of newly sequenced genomes by combining accurate transcriptome models and RNA-seq
- Auxiliary measurements improve transcriptome models
- Alternative gene models for large effects polymorphisms often exist
- *cis*-effects. Next: *trans*-effects and quantitative modeling

Summary

- Accurate classification of millions/billions of sequences is possible
- Analysis tools for RNA-seq data: from alignments to significant results
- *De novo* annotation of newly sequenced genomes by combining accurate transcriptome models and RNA-seq
- Auxiliary measurements improve transcriptome models
- Alternative gene models for large effects polymorphisms often exist
- *cis*-effects. Next: *trans*-effects and quantitative modeling

Summary

- Accurate classification of millions/billions of sequences is possible
- Analysis tools for RNA-seq data: from alignments to significant results
- *De novo* annotation of newly sequenced genomes by combining accurate transcriptome models and RNA-seq
- Auxiliary measurements improve transcriptome models
- Alternative gene models for large effects polymorphisms often exist
- *cis*-effects. Next: *trans*-effects and quantitative modeling

Summary

- Accurate classification of millions/billions of sequences is possible
- Analysis tools for RNA-seq data: from alignments to significant results
- *De novo* annotation of newly sequenced genomes by combining accurate transcriptome models and RNA-seq
- Auxiliary measurements improve transcriptome models
- Alternative gene models for large effects polymorphisms often exist
- *cis*-effects. Next: *trans*-effects and quantitative modeling

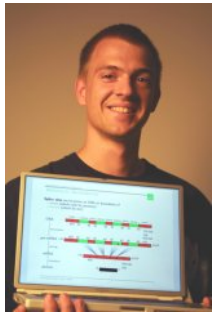
Summary

- Accurate classification of millions/billions of sequences is possible
- Analysis tools for RNA-seq data: from alignments to significant results
- *De novo* annotation of newly sequenced genomes by combining accurate transcriptome models and RNA-seq
- Auxiliary measurements improve transcriptome models
- Alternative gene models for large effects polymorphisms often exist
- *cis*-effects. Next: *trans*-effects and quantitative modeling

Further information

Slides available at:

fml.mpg.de/raetsch/lectures



Sören Sonnenburg



Jonas Behr



Gabriele Schweikert

Acknowledgments

Large scale learning:	**Sören Sonnenburg (TU Berlin)**
Gene finding:	**Jonas Behr, Gabriele Schweikert, Georg Zeller (EMBL)**
Quantification:	**Regina Bohnert**
Library preparation:	**Lisa Hartmann, Lisa M. Smith (MPI Dev. Biology)**
QTLs:	**Oliver Stegle (MPI Dev. Biology)**

Funding: Max Planck Society and German Research Foundation

References I

Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov Support Vector Machines. In *Proc. 20th Int. Conf. Mach. Learn.*, pages 3–10, 2003.

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. Oral presentation at the CSHL Genome Informatics Meeting, September 2008. URL `http://www.fml.tuebingen.mpg.de/raetsch/lectures/RaetschGenomeInformatics08.pdf`.

R. Bohnert and G. Rätsch. rQuant.web: a tool for RNA-seq-based transcript quantitation. *NAR Webserver Issue*, 38(Suppl):W348–51, 2010.

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.

R Dahm. Friedrich miescher and the discovery of dna. *Developmental Biology*, 278(2):274–288, 2005.

References II

F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Qpalma: Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24:i174–i180, 2008.

V. Franc and S. Sonnenburg. Optimized cutting plane algorithm for large-scale risk minimization. *Journal of Machine Learning Research*, 10:2157–2192, 2009.

G Jean, A Kahles, VT Sreedharan, F De Bona, and G Rätsch. Rna-seq read alignments with palmapper. *Current Protocols in Bioinformatics*, 2010.

F. Miescher. Die spermatozoen einiger wirbelthiere. ein beitrag zur histochemie. *Verh. Nat.forsch. Ges. Basel*, 1874.

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 and S Sonnenburg. Large scale hidden semi-markov svms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NIPS'06)*, volume 19, pages 1161–1168, Cambridge, MA, 2007. MIT Press. URL `http://www.fml.tuebingen.mpg.de/raetsch/projects/HSMSVM`.

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, and C. Schäfer. Learning interpretable svms for biological sequence classification. *BMC Bioinformatics*, 7(Suppl 1):S9, February 2006a.

References III

G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. Müller, R. Sommer, and B. Schl̈kopf. Improving the c. elegans genome annotation using machine learning. *PLoS Computational Biology*, 3(2):e20, 2007. URL `http://dx.doi.org/10.1371/journal.pcbi.0030020.eor`.

Gunnar Rätsch, Sören Sonnenburg, and Christin Schäfer. Learning interpretable svms for biological sequence classification. *BMC Bioinformatics*, 7 Suppl 1:S9, 2006b. doi: 10.1186/1471-2105-7-S1-S9.

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.

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

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

References IV

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

S. Sonnenburg, G. Rätsch, and K. Rieck. Large scale learning with string kernels. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*. MIT Press, 2007a.

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

S. Sonnenburg, G. Rtsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. De Bona, A. Binder, C. Gehl, and V. Franc. The shogun machine learning toolbox. *Journal of Machine Learning Research*, 11:1799–1802, 2010.

Sören Sonnenburg and Vojtech Franc. COFFIN: a computational framework for linear SVMs. In *Proceedings of the 27nd International Machine Learning Conference*, 2010.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):e472–480, 2006.

Sören Sonnenburg, Alexander Zien, Petra Philips, and Gunnar Rätsch. Poims: positional oligomer importance matrices–understanding support vector machine-based signal detectors. *Bioinformatics*, 24(13):i6–14, Jul 2008. doi: 10.1093/bioinformatics/btn170.

References V

O. Stegle, P. Drewe, R. Bohnert, K. Borgwardt, and G. Rätsch. Statistical tests for detecting differential rna-transcript expression from read counts. Technical report, Nature Preceedings, 2010a.

O. Stegle, L. Parts, R. Durbin, and J.M. Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Computational Biology*, 2010b.

NC Toussaint, C Widmer, O Kohlbacher, and G Rätsch. Exploiting physico-chemical properties in string kernels. *BMC Bioinformatics*, 11(Suppl. 8):S7, 2010. URL `http://www.biomedcentral.com/1471-2105/11/S8/S7`.

JD Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 4356:737, 1953.

C Widmer, NC Toussaint, Y Altun, and Gunnar Rätsch. Inferring latent task structure for multi-task learning by multiple kernel learning. *BMC Bioinformatics*, 11(Suppl. 8):S5, 2010a. URL `http://www.biomedcentral.com/1471-2105/11/S8/S5`.

Christian Widmer, Jose Leiva, Yasemin Altun, and Gunnar Rätsch. Leveraging sequence classification by taxonomy-based multitask learning. In Bonnie Berger, editor, *RECOMB*, volume 6044 of *Lecture Notes in Computer Science*, pages 522–534. Springer, 2010b. ISBN 978-3-642-12682-6.

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, 2008. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

References VI

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.

A Zien, N Krämer, S Sonnenburg, and G Rätsch. The feature importance ranking measure. In W. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, editors, *Proc. ECML PKDD*, volume 5782/2009 of *Lecture Notes in Artificial Intelligence*, pages 694–709, Springer Berlin / Heidelberg, 2009. Springer.

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]