RNA-seq and Tiling Arrays for Improved Gene Finding

Gunnar Rätsch



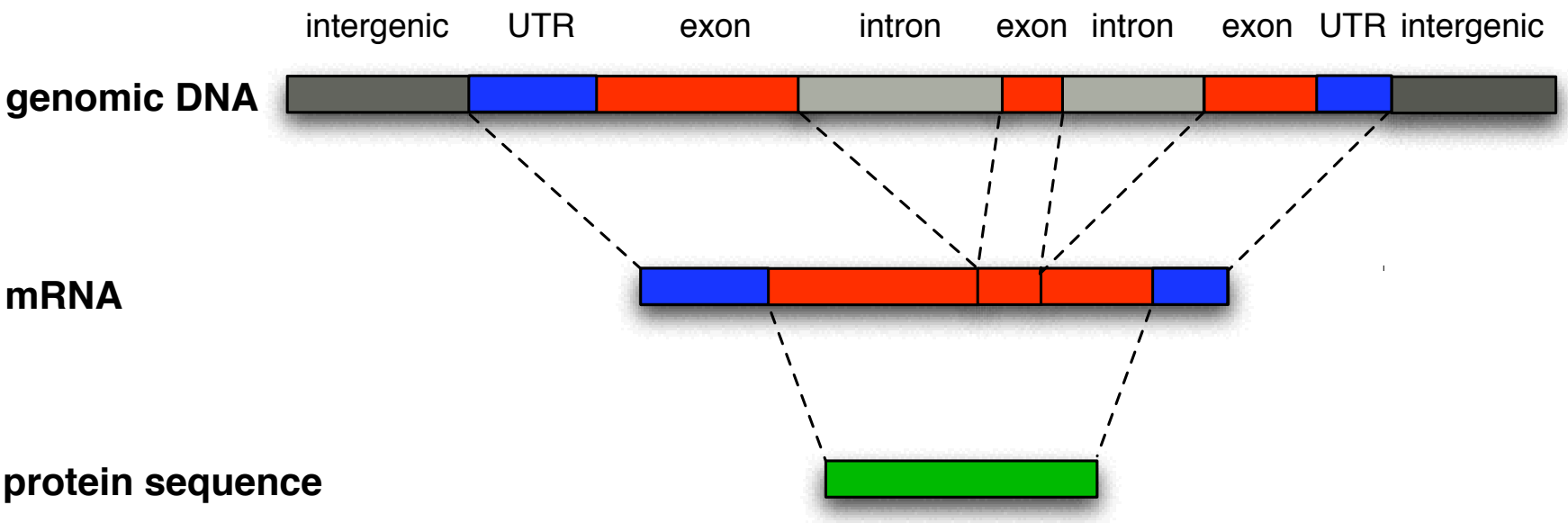
Friedrich Miescher Laboratory
of the Max Planck Society

Genome Informatics Meeting
September 14, 2008

Gene Finding

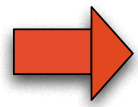
Given the genomic DNA sequence:

- Localize genes within the sequence
- Predict (alternative) transcript(s)

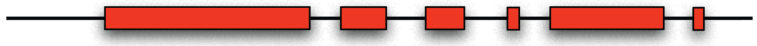


Motivation

- Genomic sequence
- Conservation
- cDNAs/ESTs/Proteins
- Tiling array data
- RNA-seq data

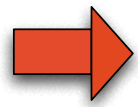


Accurate Gene Predictions

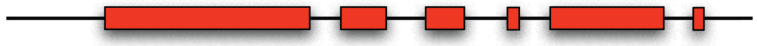


Motivation

- Genomic sequence
- Conservation
- cDNAs/ESTs/Proteins
- Tiling array data
- RNA-seq data

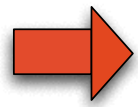


Accurate Gene Predictions

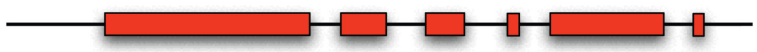


Motivation

- Genomic sequence
- Conservation
- cDNAs/ESTs/Proteins
- Tiling array data
- RNA-seq data



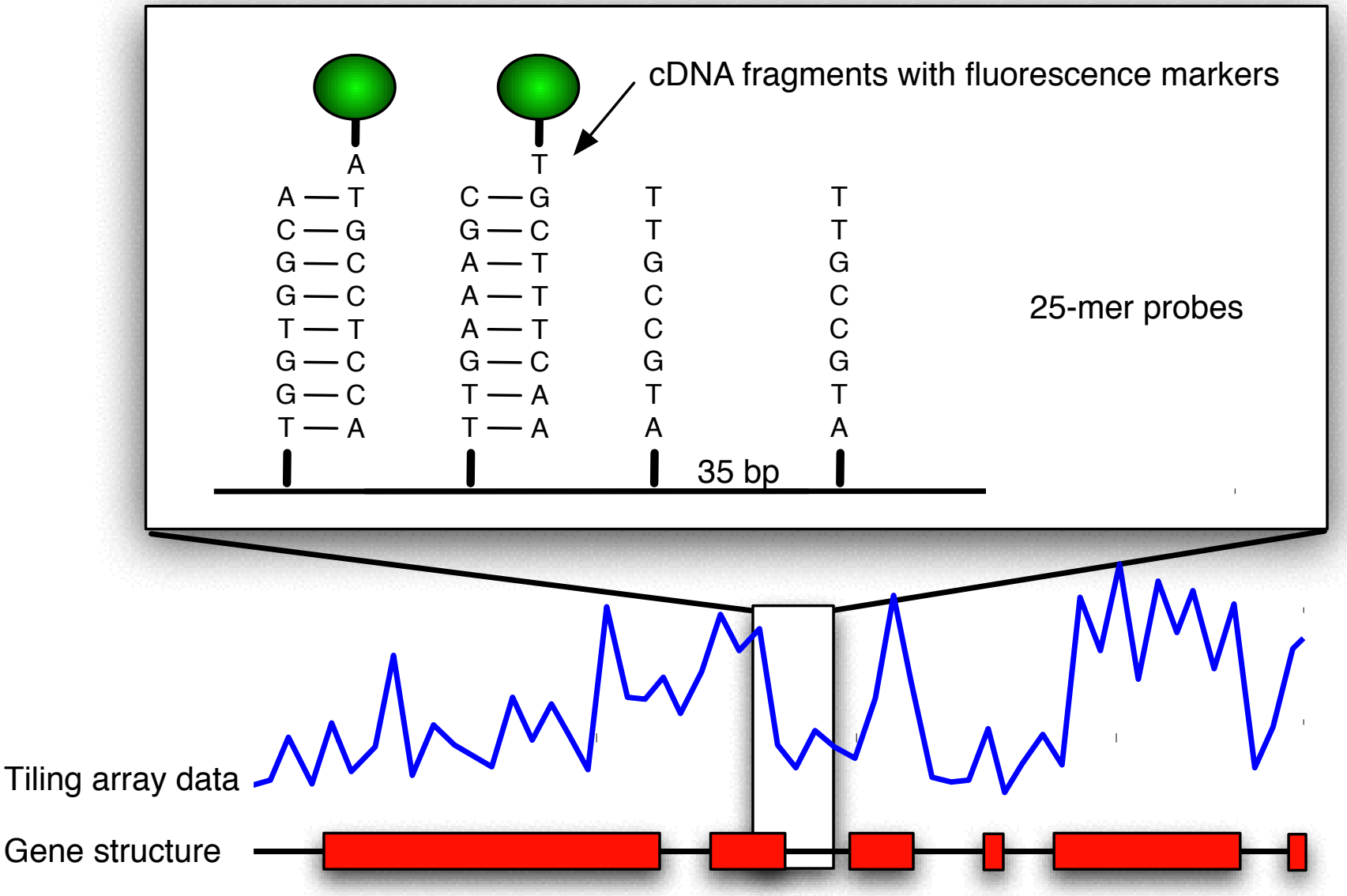
Accurate Gene Predictions



Roadmap

- Data
- Method
- Results
- Alternative transcripts

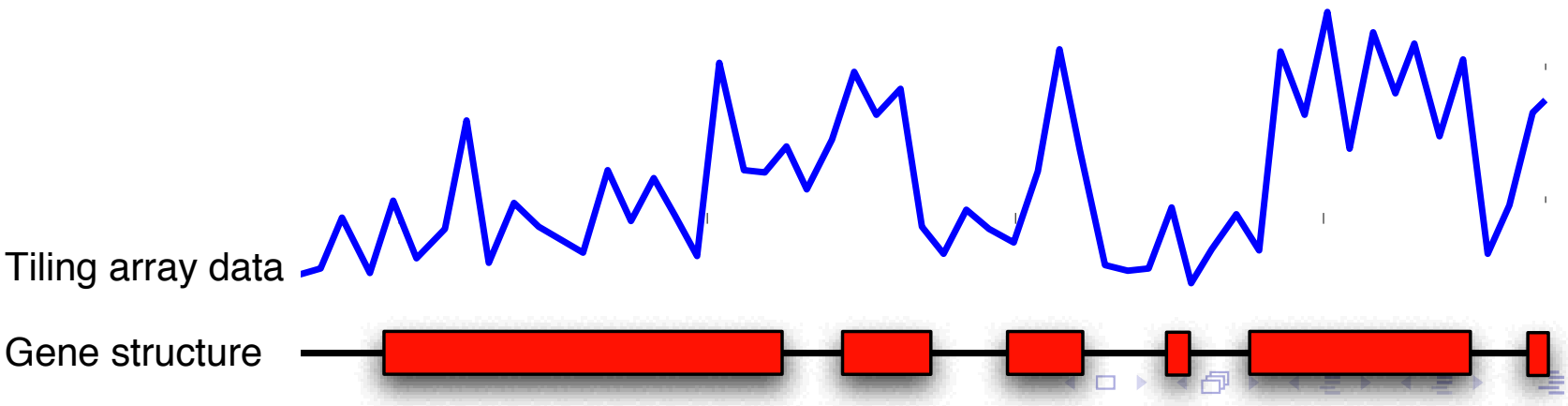
Tiling Array Data



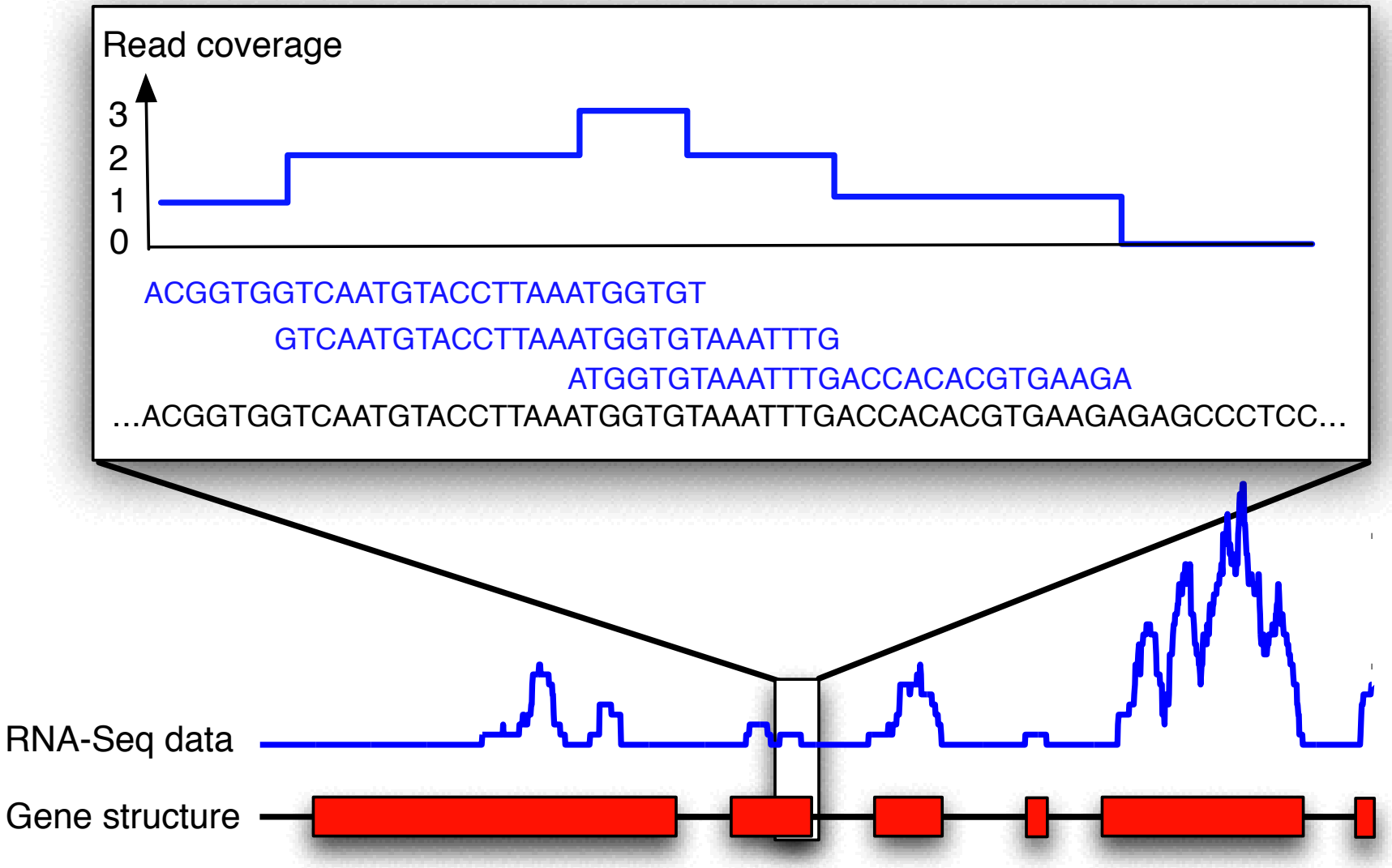
Tiling Array Data

Tiling arrays for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

- 25nt probes, 35nt spacing, 3 replicates
- Strand unspecific, polyA enriched
- 12 different tissues (young leaves, root, etc.)
- 12 conditions/mutants (e.g. abiotic stresses)
- Quantile and sequence dependent normalization (Zeller et al., 2008)



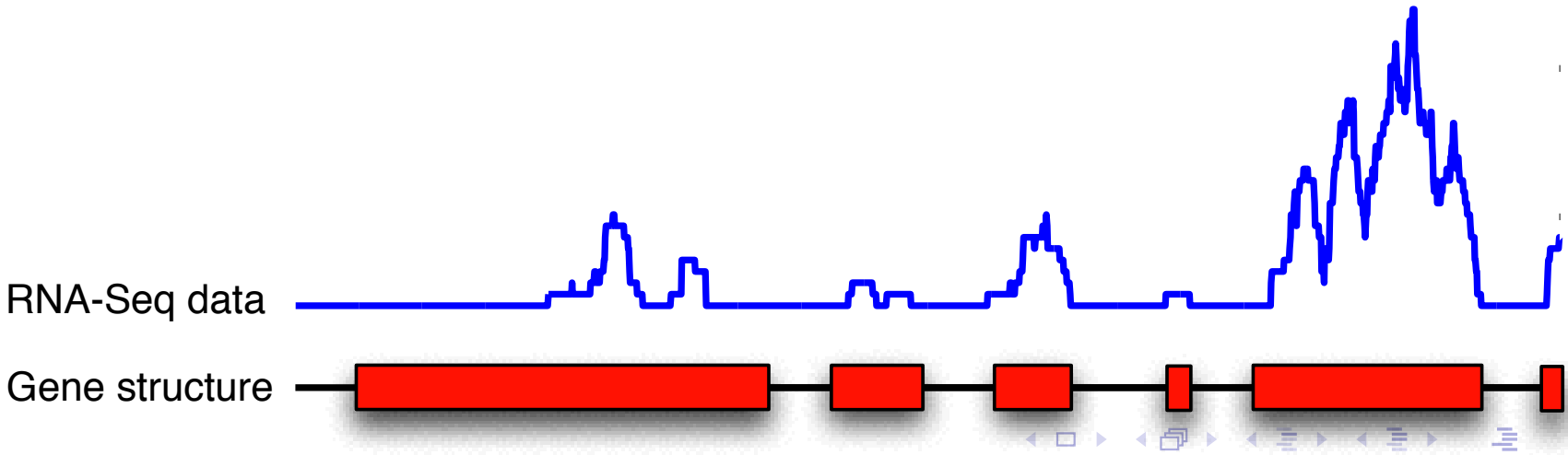
RNA-seq Data


Read coverage
3
2
1
0
ACGGTGGTCAATGTACCTTAAATGGTGT
GTCAATGTACCTTAAATGGTGTAAATTTG
ATGGTGTAAATTTGACCACACGTGAAGA
...ACGGTGGTCAATGTACCTTAAATGGTGTAAATTTGACCACACGTGAAGAGAGCCCTCC...
RNA-Seq data
Gene structure

RNA-seq Data

RNA-seq reads for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

- 4 lanes from Illumina Genome Analyzer
- 38nt reads, polyA enriched
- Strand unspecific, young leaves
- Read mapping using **ShoRe** (Ossowski et al., 2008)
- Spliced alignments using **QPalma** (De Bona et al., Bioinformatics, 2008)
- ≈30 million unspliced and ≈1 million spliced reads (≈50x coverage)



mGene: SVM-based Gene Finding

Two-stage modeling process:

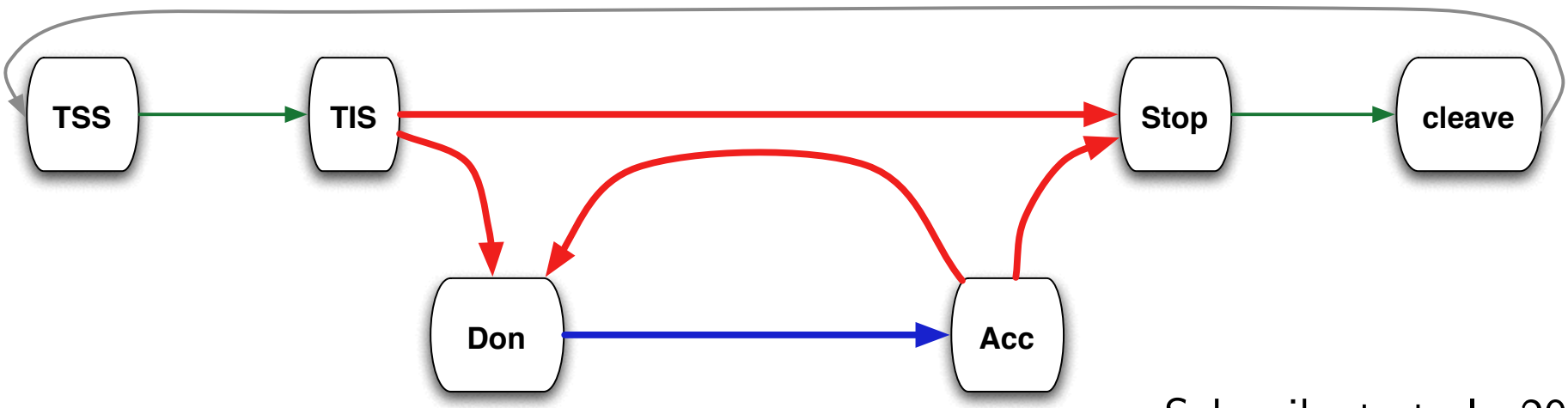
- Train detectors for functional elements
  - Splice sites, translation start/stop, transcription start/stop
- Train to optimally combine signals for prediction of gene structures additionally using
  - Content sensors (coding preferences, etc.)
  - Transcriptome measurements

Schweikert et al., 2008

mGene: SVM-based Gene Finding

Two-stage modeling process:

- Train detectors for functional elements
  - Splice sites, translation start/stop, transcription start/stop
- Train to optimally combine signals for prediction of gene structures additionally using
  - Content sensors (coding preferences, etc.)
  - **Transcriptome measurements**

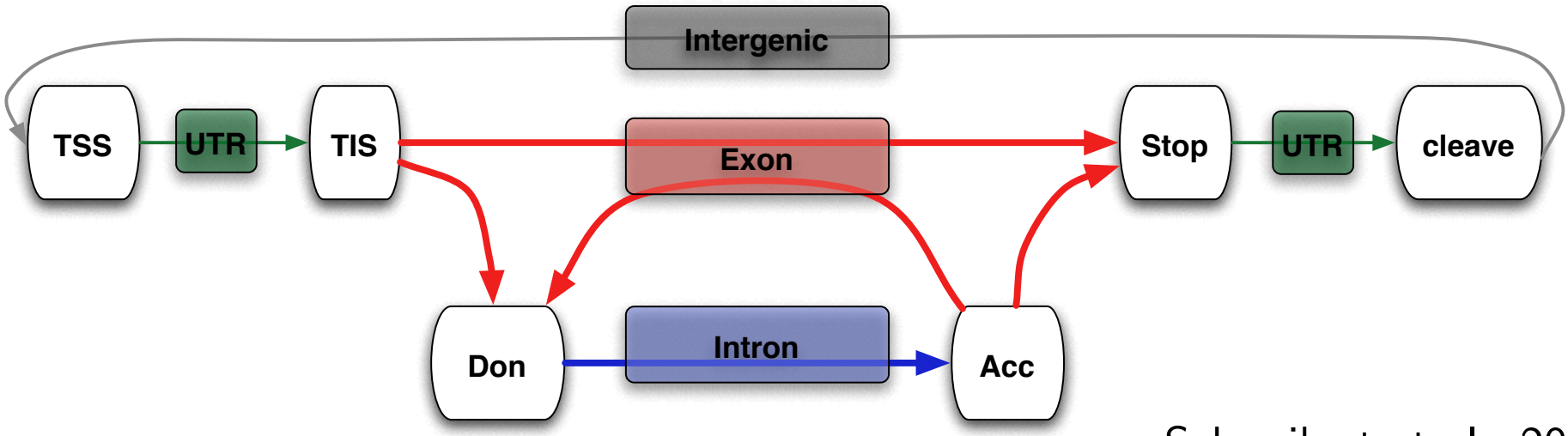


Schweikert et al., 2008

mGene: SVM-based Gene Finding

Two-stage modeling process:

- Train detectors for functional elements
  - Splice sites, translation start/stop, transcription start/stop
- Train to optimally combine signals for prediction of gene structures additionally using
  - Content sensors (coding preferences, etc.)
  - Transcriptome measurements

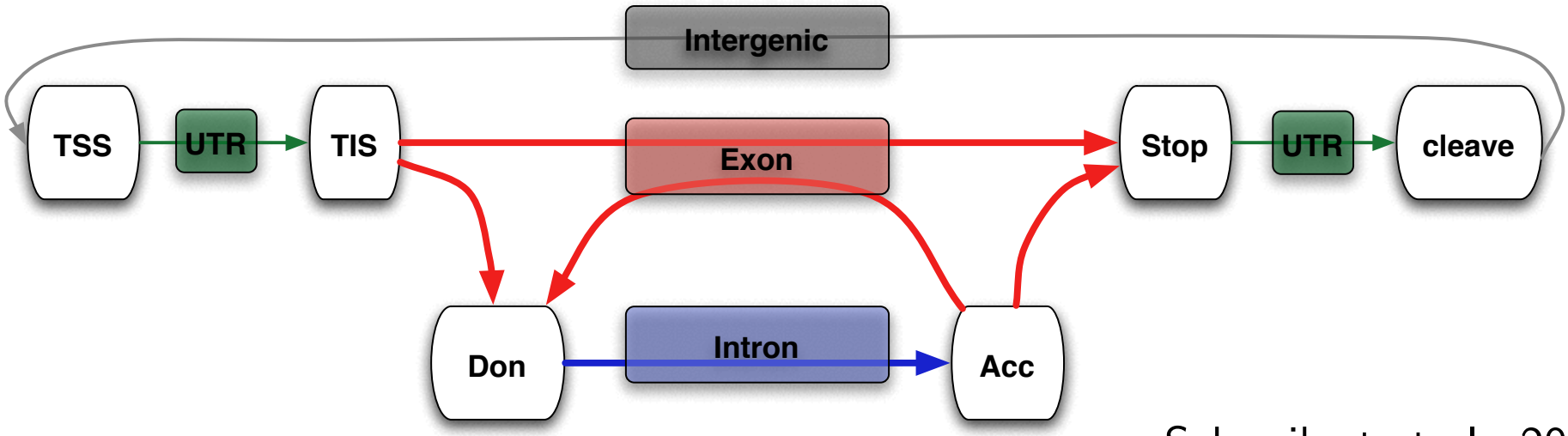


Schweikert et al., 2008

mGene: SVM-based Gene Finding

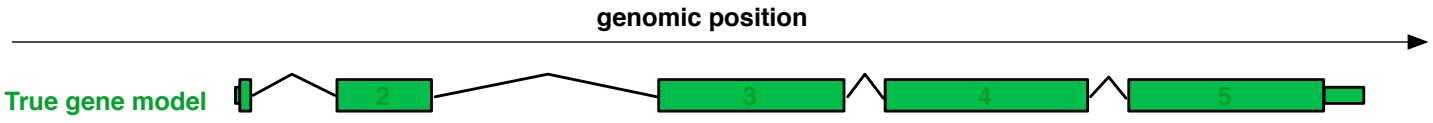
Two-stage modeling process:

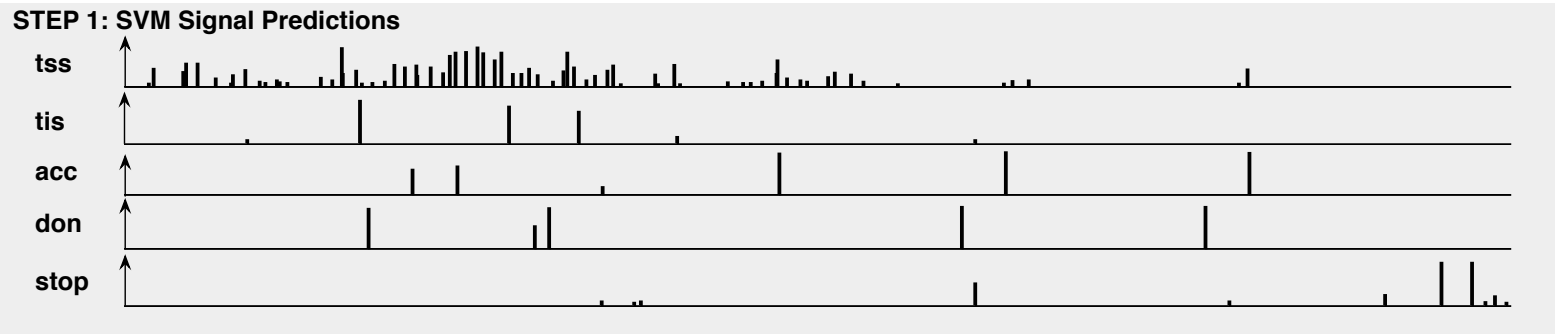
- Train detectors for functional elements
  - Splice sites, translation start/stop, transcription start/stop
- Train to optimally combine signals for prediction of gene structures additionally using
  - Content sensors (coding preferences, etc.)
  - **Transcriptome measurements**



Schweikert et al., 2008

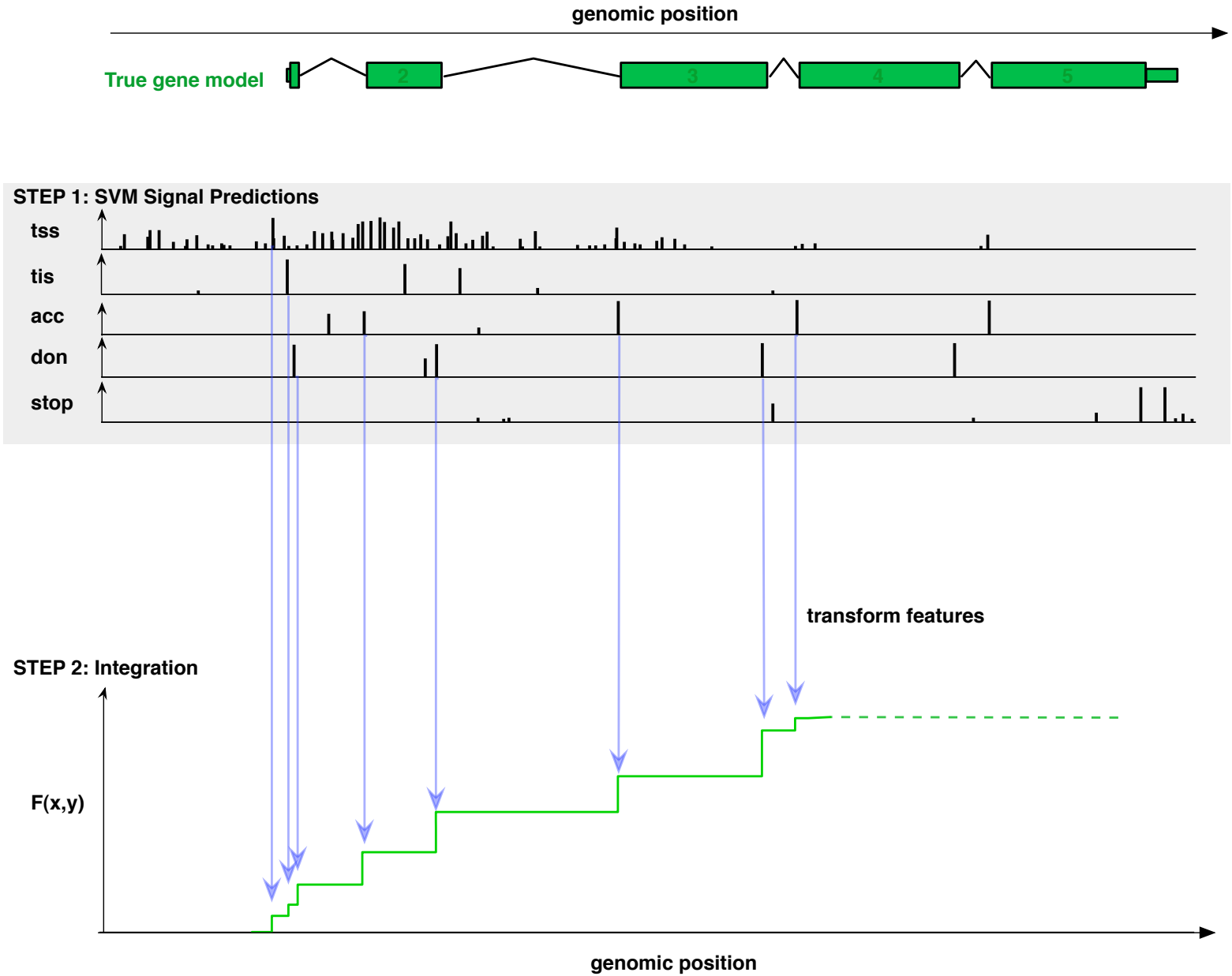
Learning to Integrate Data Sources



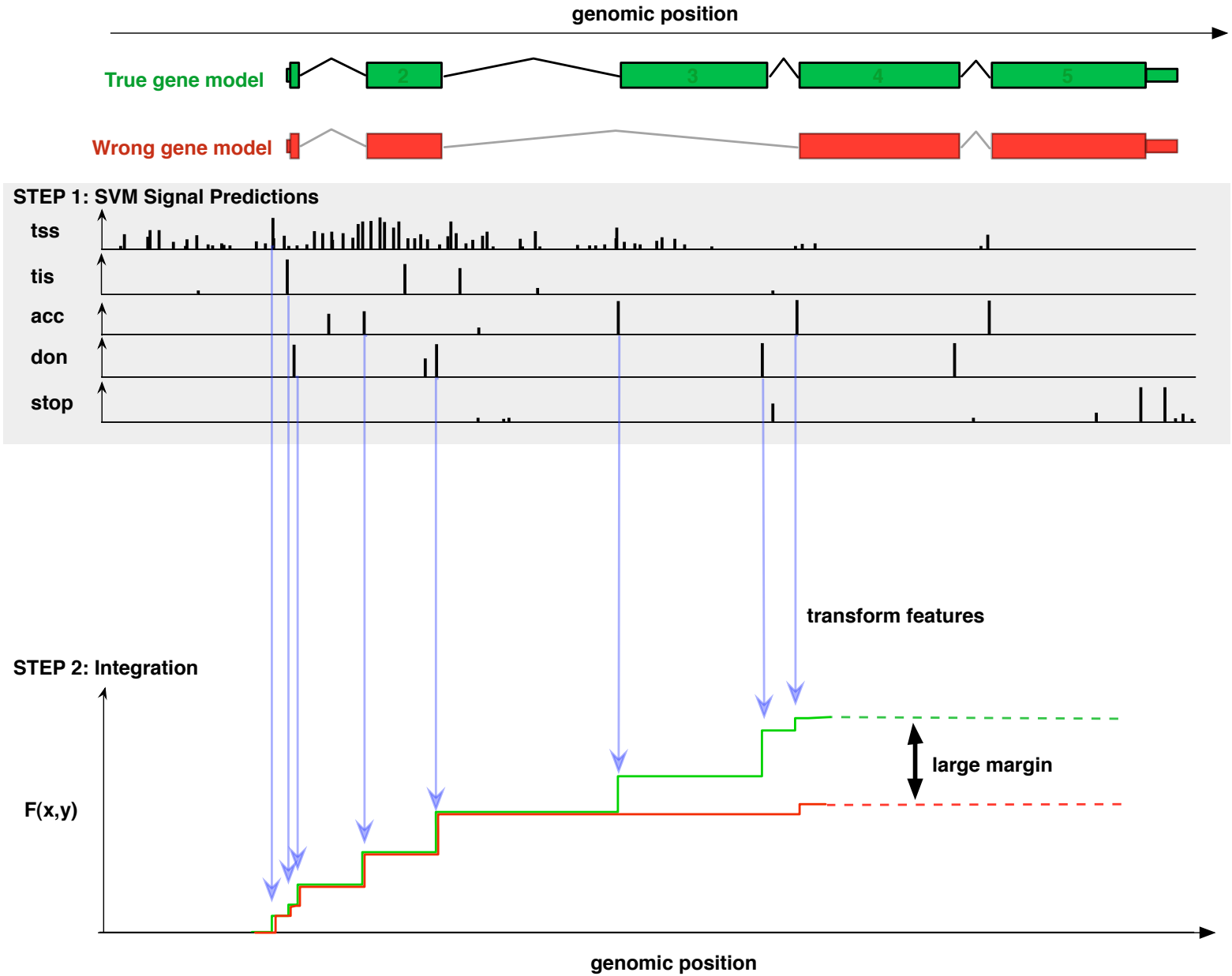




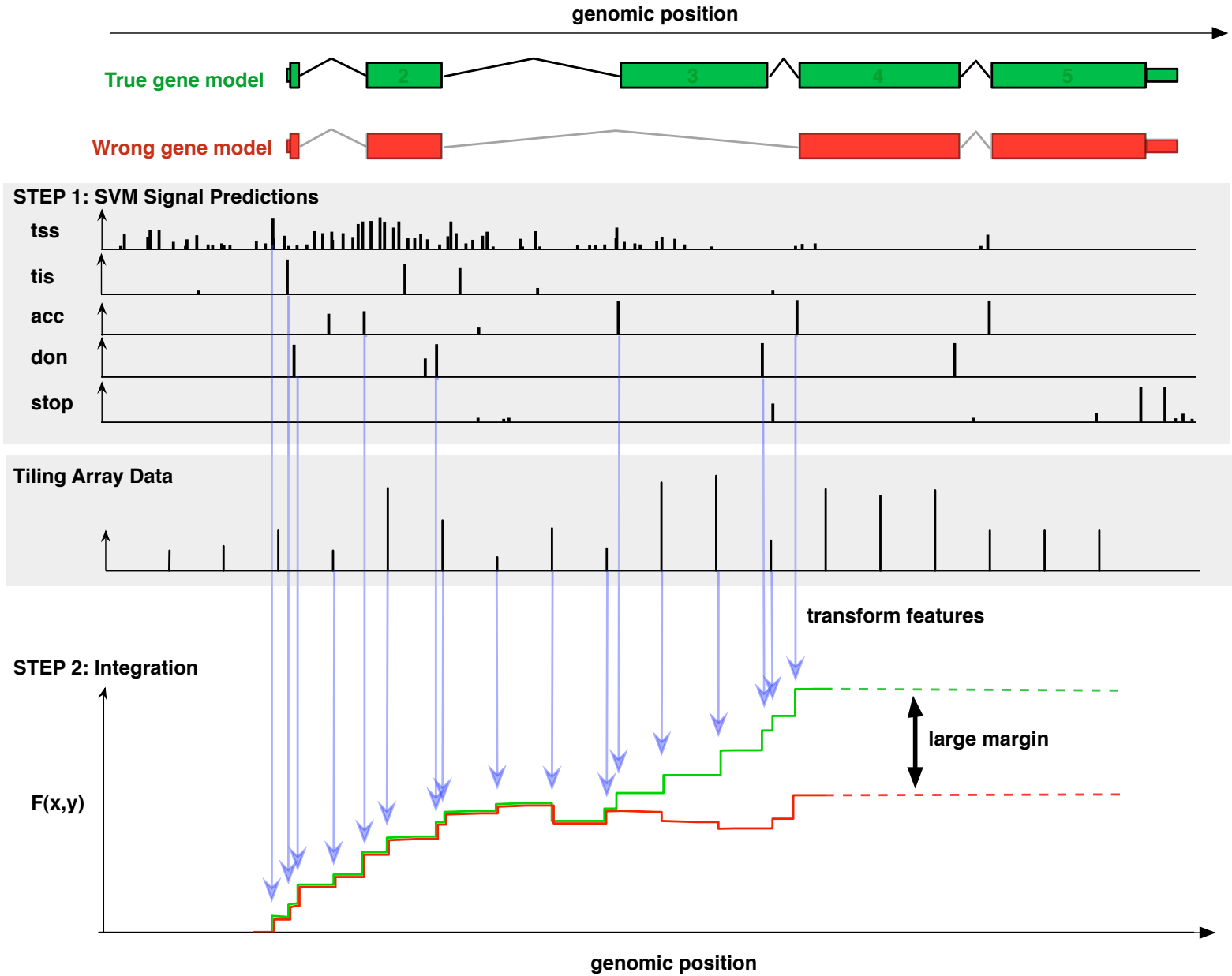
Learning to Integrate Data Sources



Learning to Integrate Data Sources



Learning to Integrate Data Sources



Results

Experimental setup (*Arabidopsis thaliana*):

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

Results

Experimental setup (*Arabidopsis thaliana*):

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

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

Results

Experimental setup (*Arabidopsis thaliana*):

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

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

Results

Experimental setup (*Arabidopsis thaliana*):

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

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

Results

Experimental setup (*Arabidopsis thaliana*):

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

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

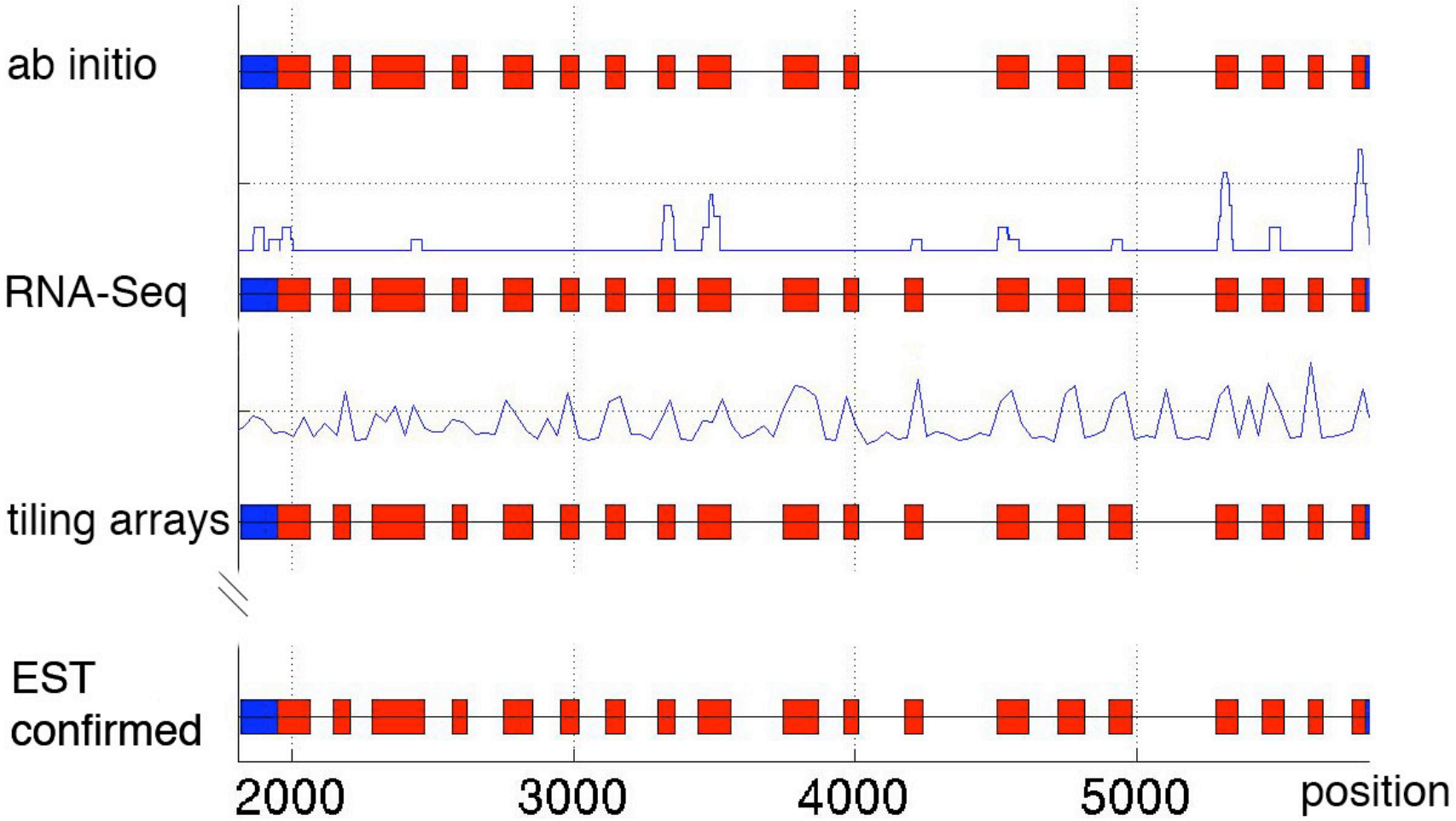
Results

Experimental setup (*Arabidopsis thaliana*):

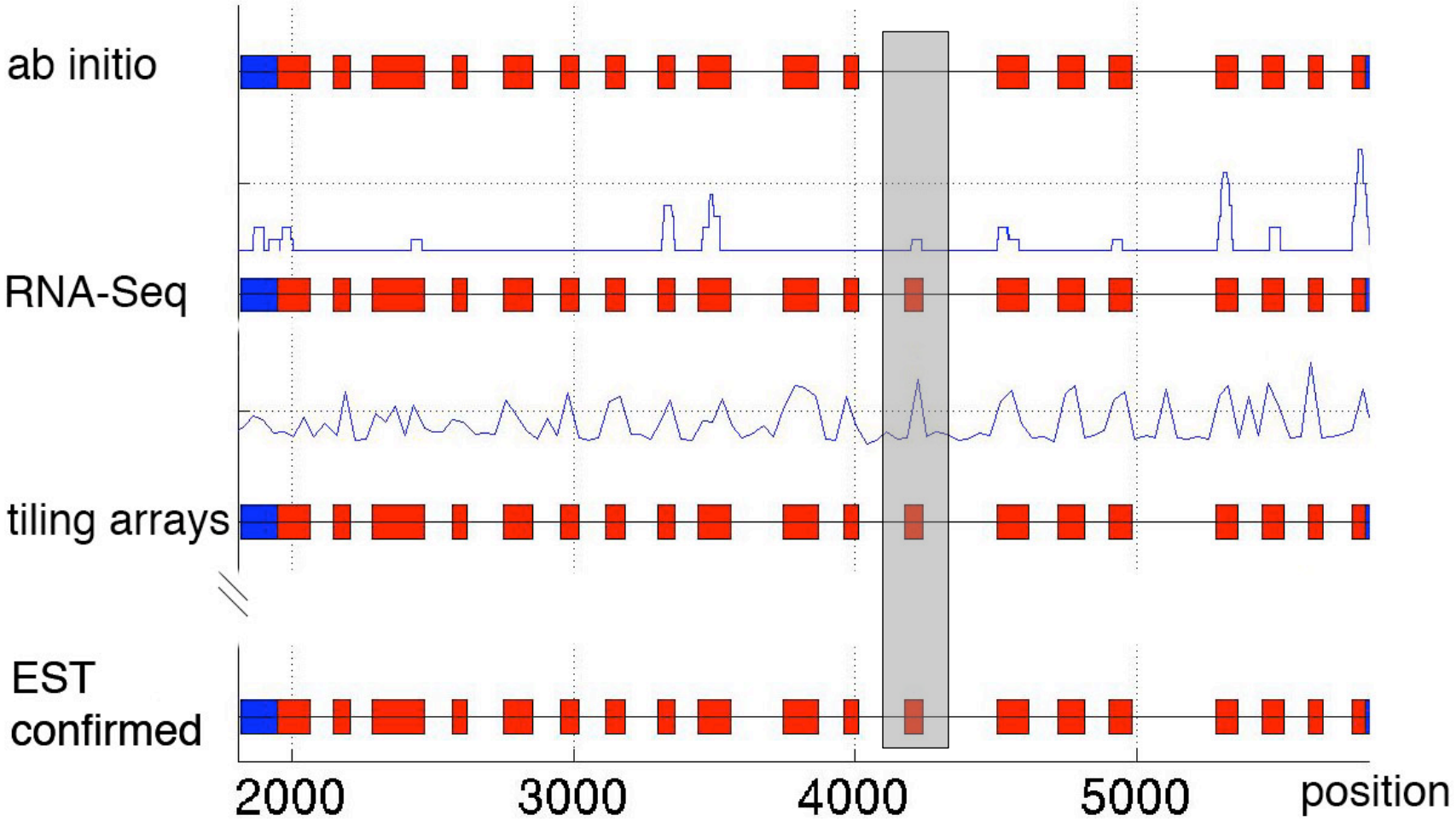
- 60% of known genes for training signals in step 1
- 400 genes for training of step 2
- 300 regions around cDNA confirmed genes for evaluation

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

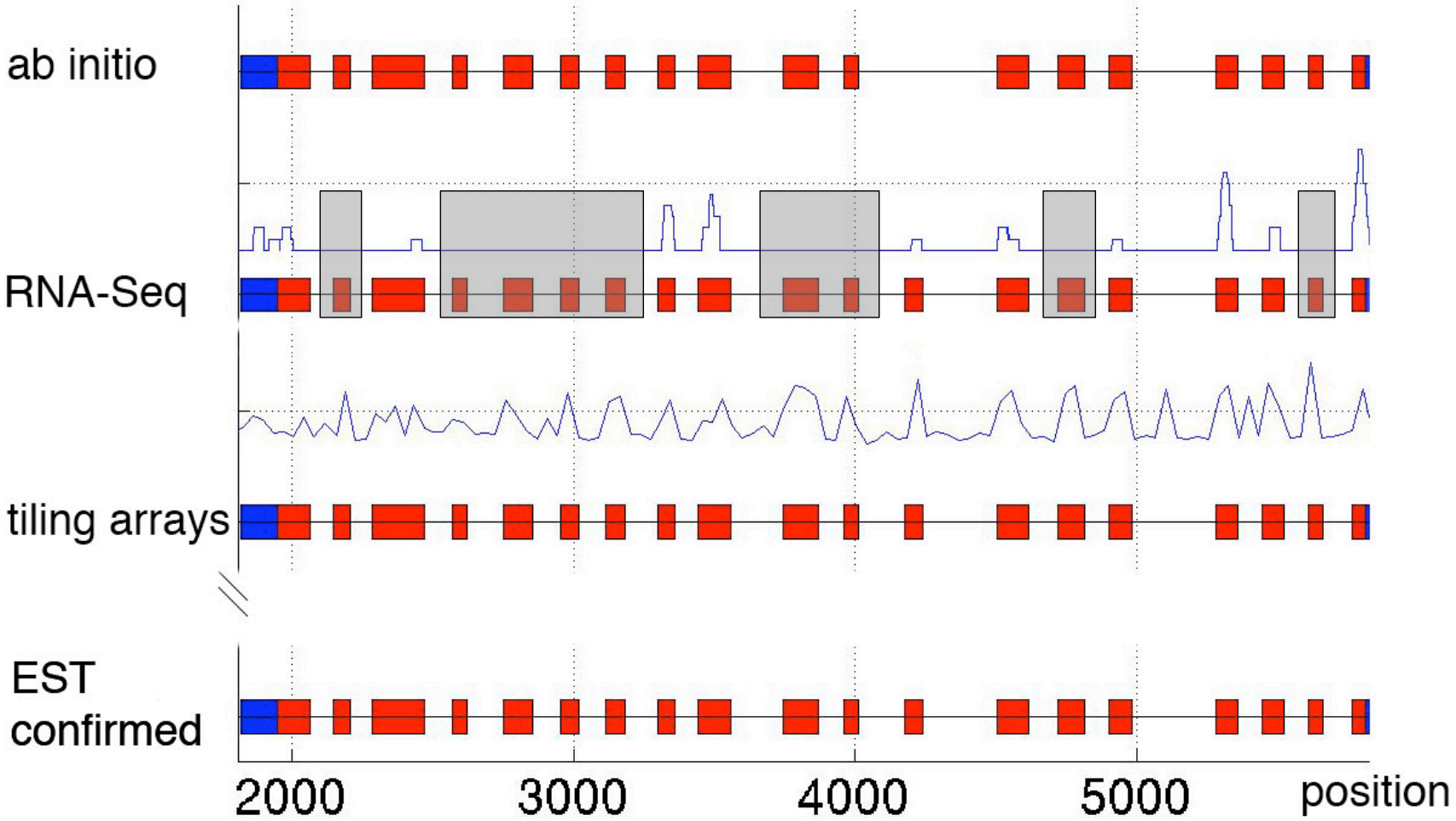
Results


ab initio
RNA-Seq
tiling arrays
EST
confirmed
2000
3000
4000
5000
position

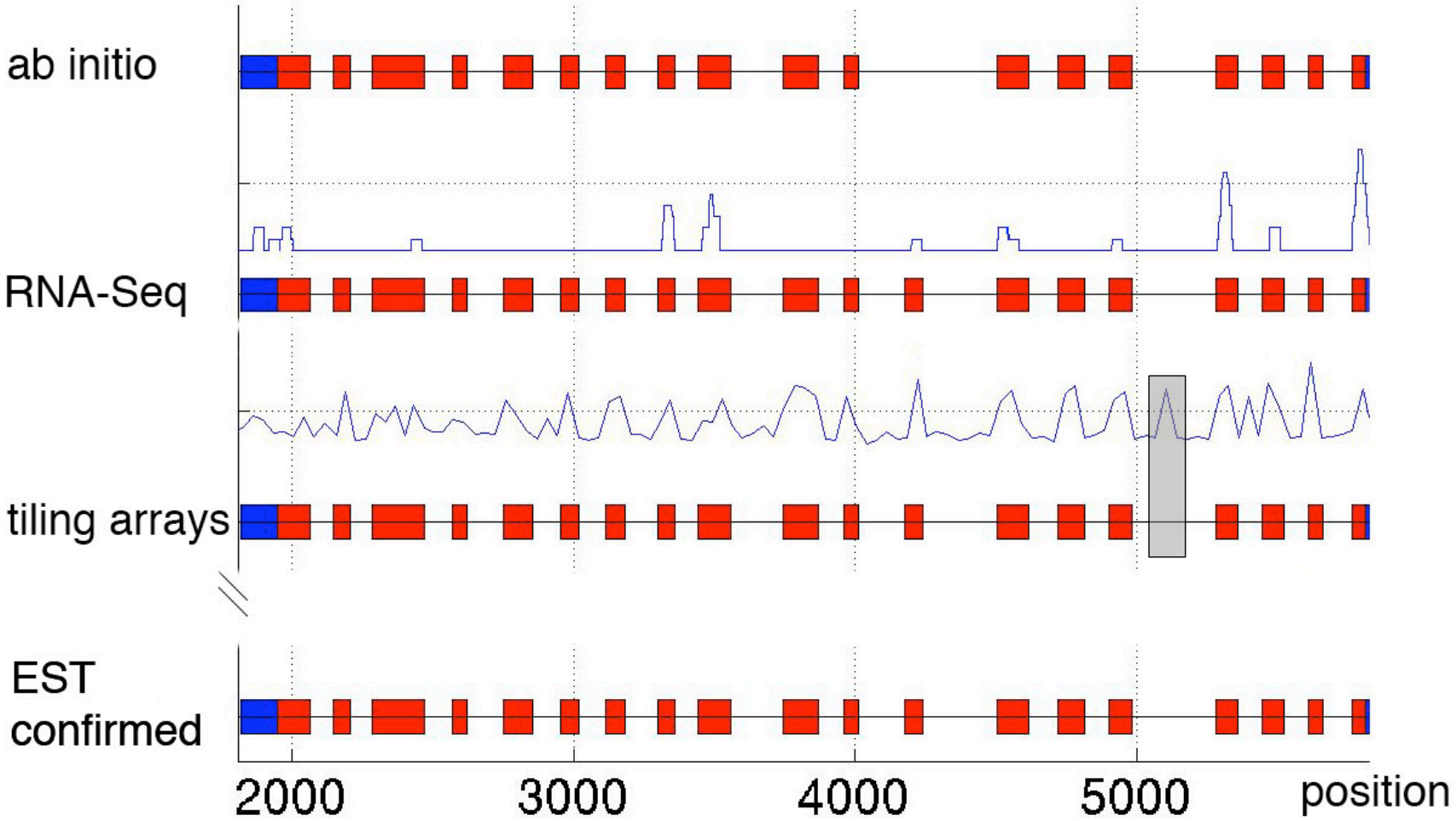
Results


ab initio
RNA-Seq
tiling arrays
EST
confirmed
2000
3000
4000
5000
position

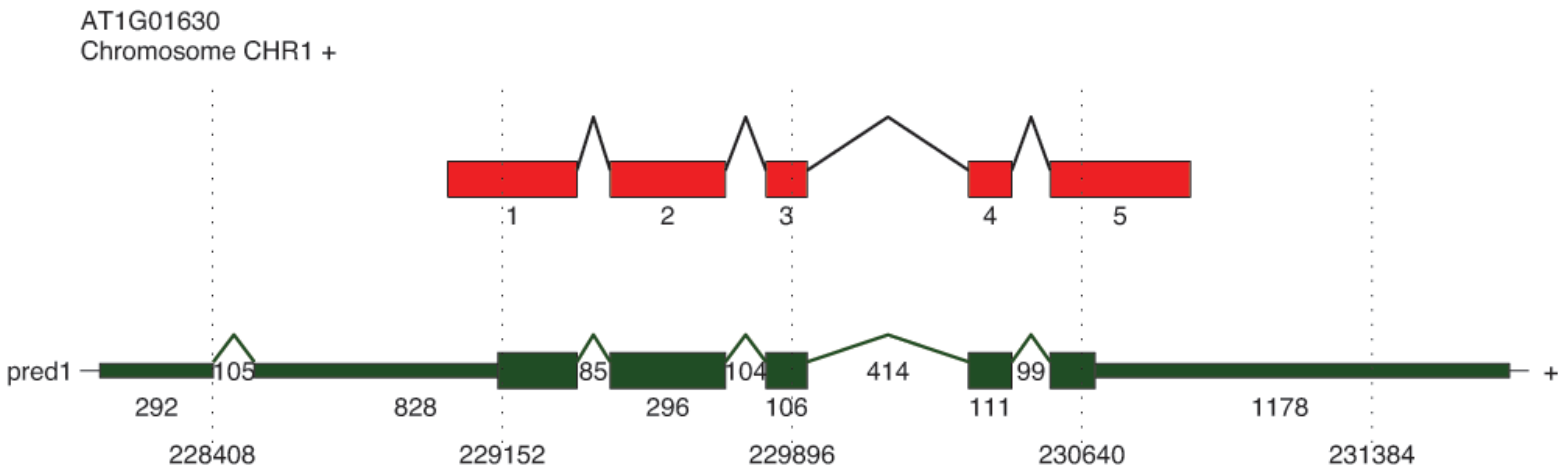
Results


ab initio
RNA-Seq
tiling arrays
EST
confirmed
2000
3000
4000
5000
position

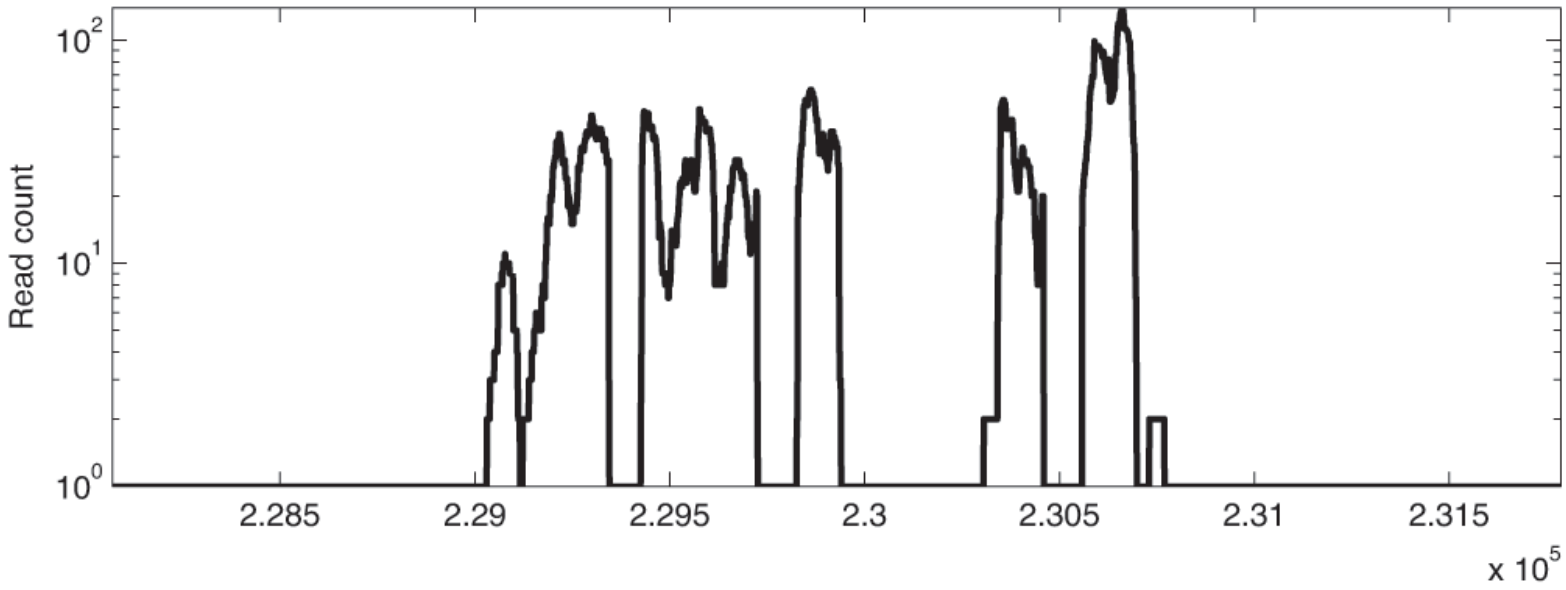
Results


ab initio
RNA-Seq
tiling arrays
EST
confirmed
2000
3000
4000
5000
position

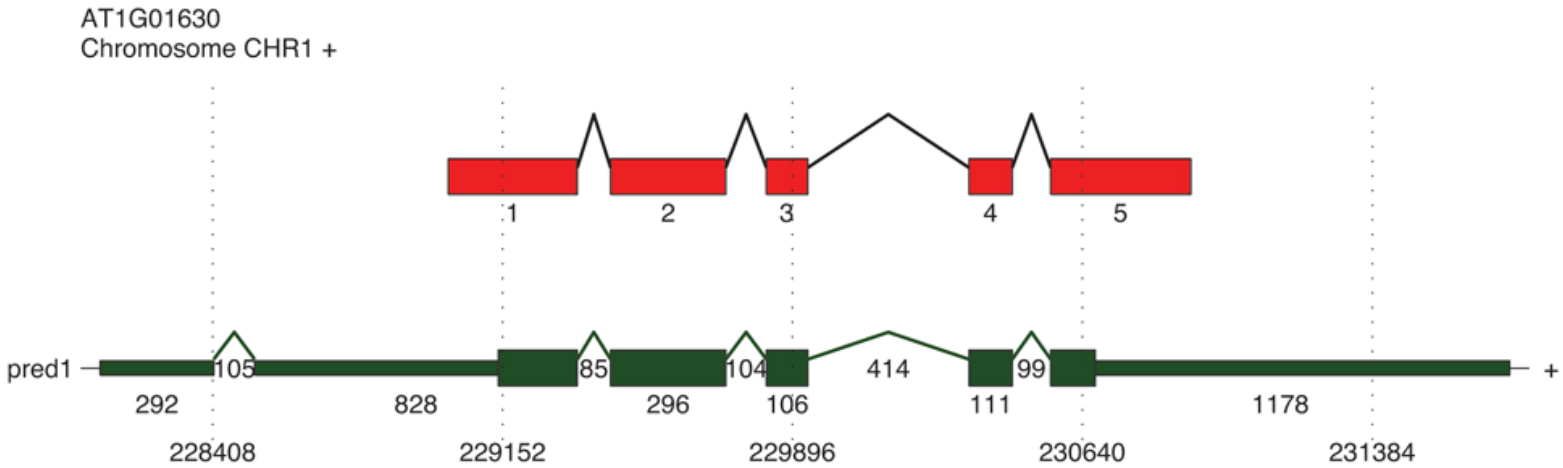
Extension to Alternative Transcripts



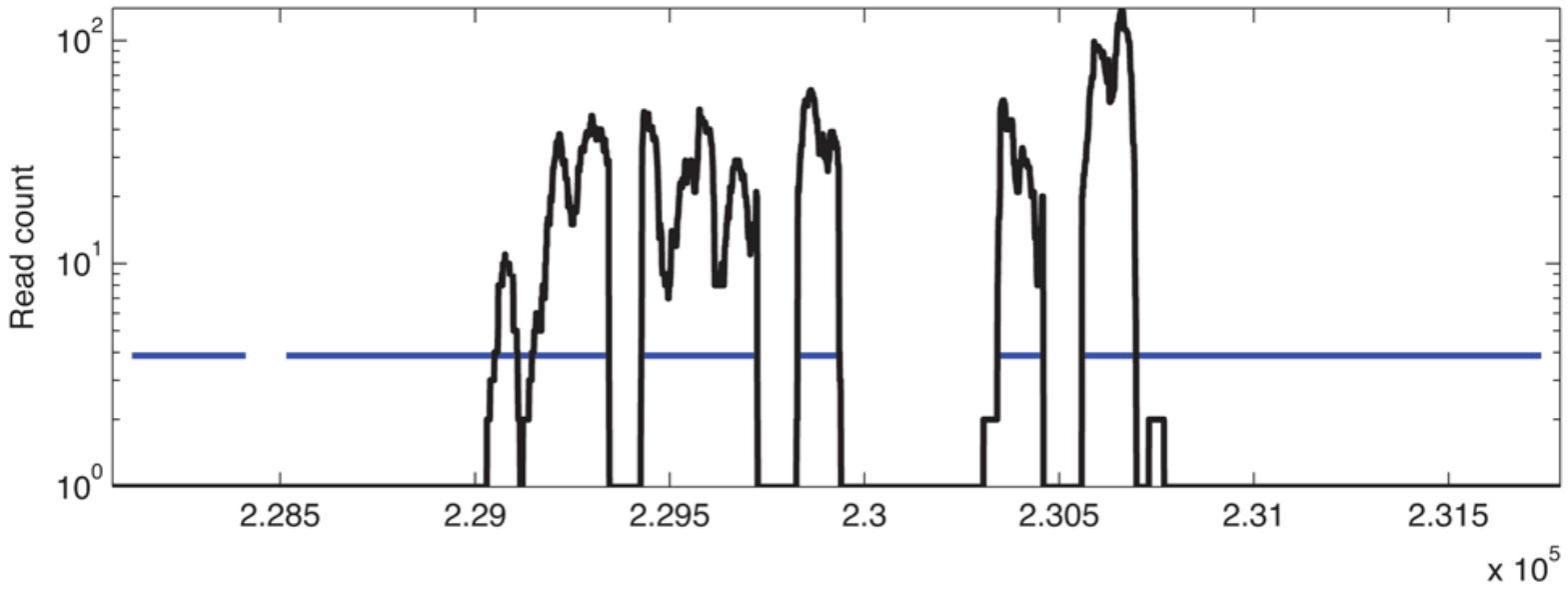
Prediction does not fit the data very well.



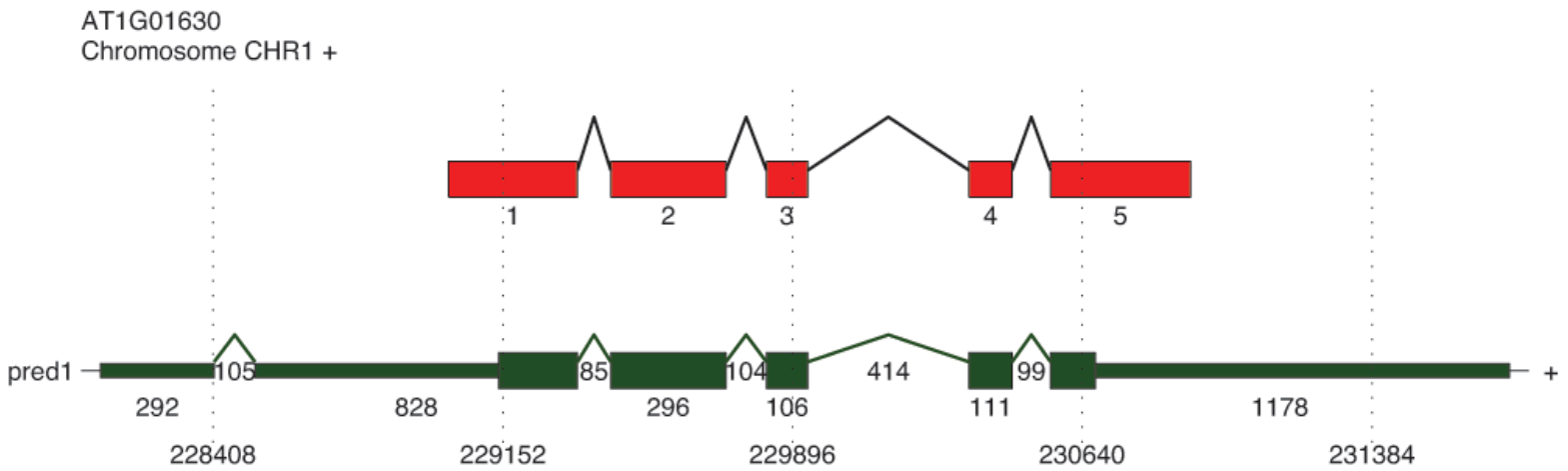
Extension to Alternative Transcripts



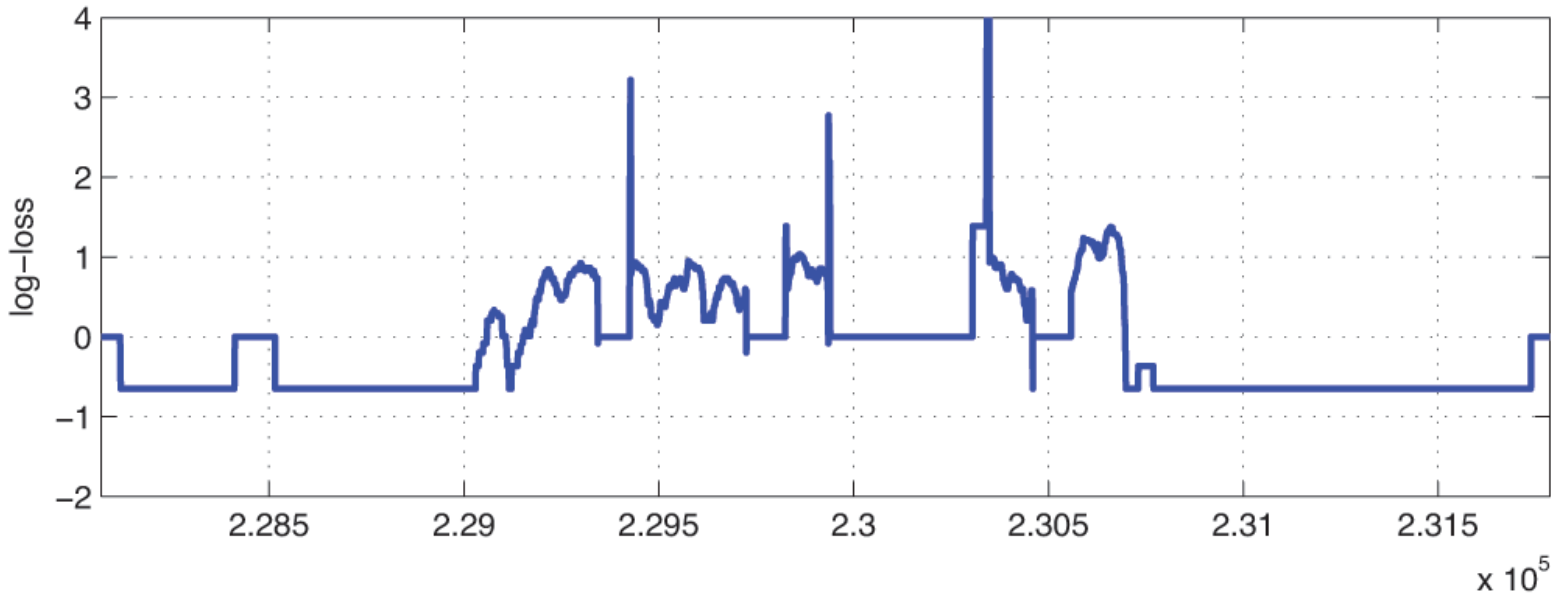
Determine best fitting transcript abundance.



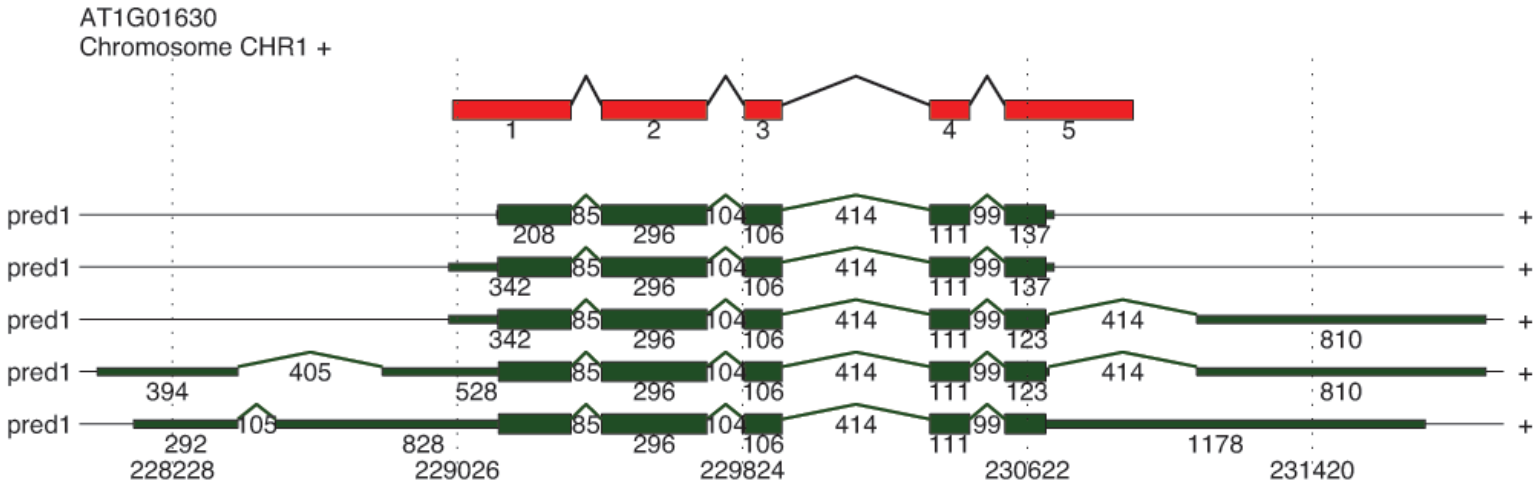
Extension to Alternative Transcripts



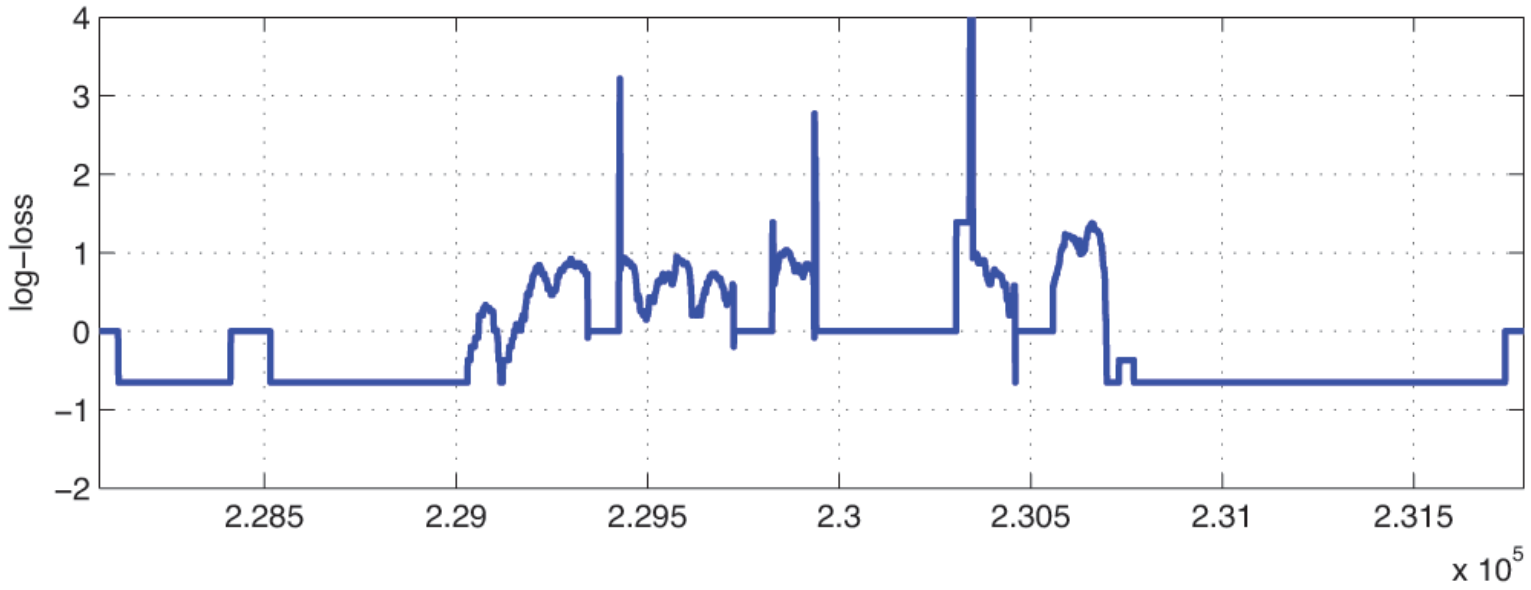
Compute residual log-**loss per position.**



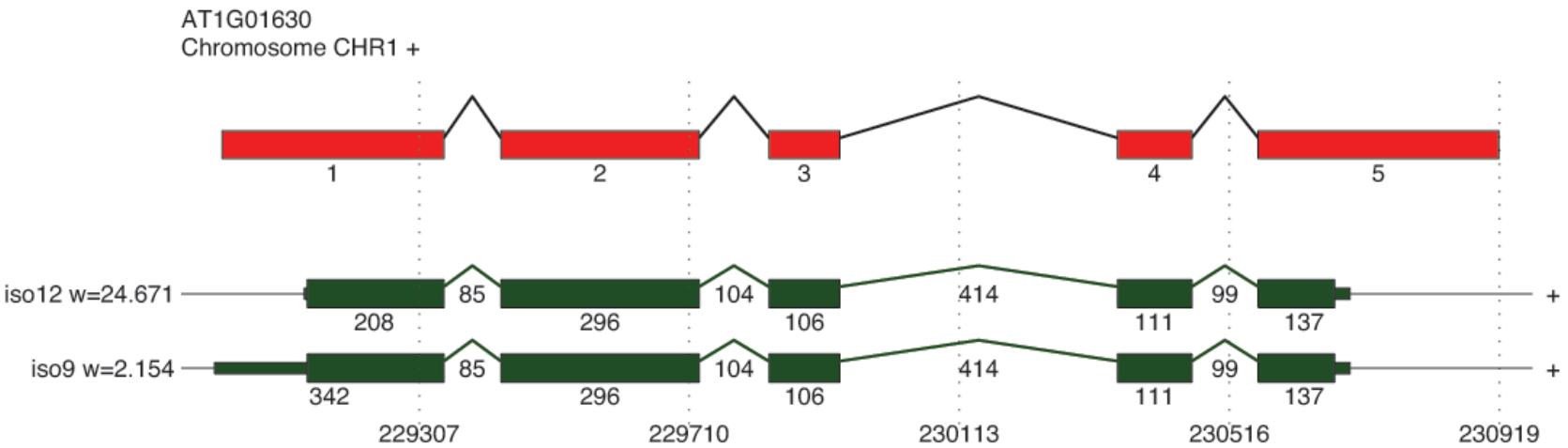
Extension to Alternative Transcripts



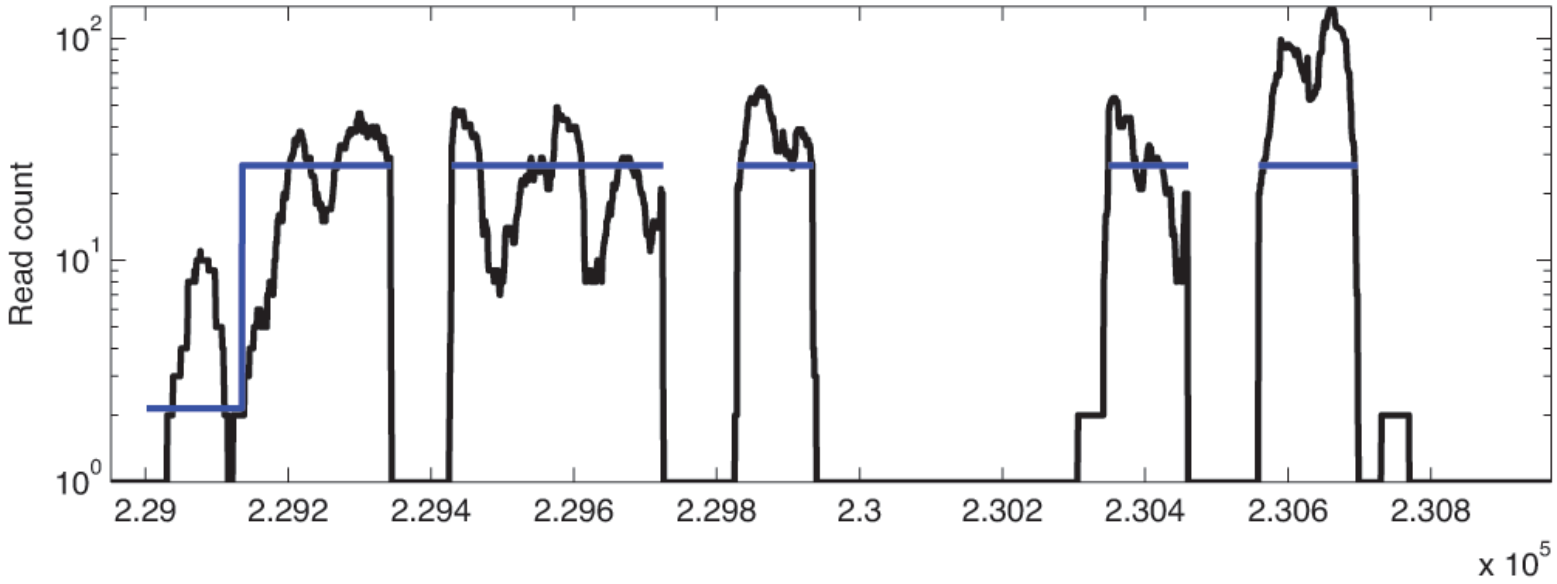
Modified dynamic program to find gene model fitting residuals.



Extension to Alternative Transcripts



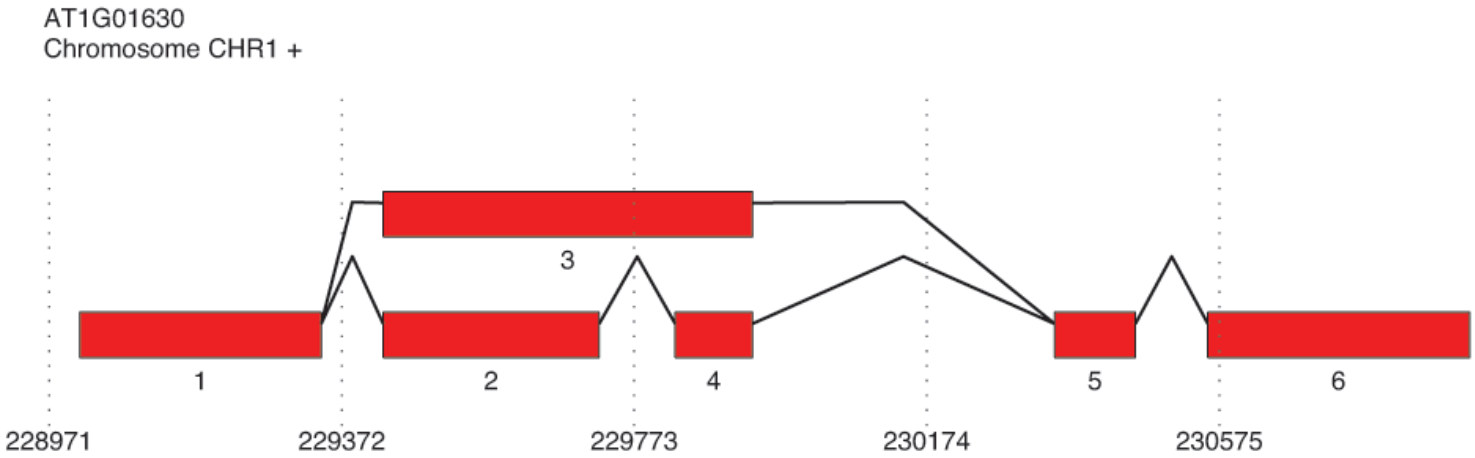
Compute transcript abundances better fitting observations.



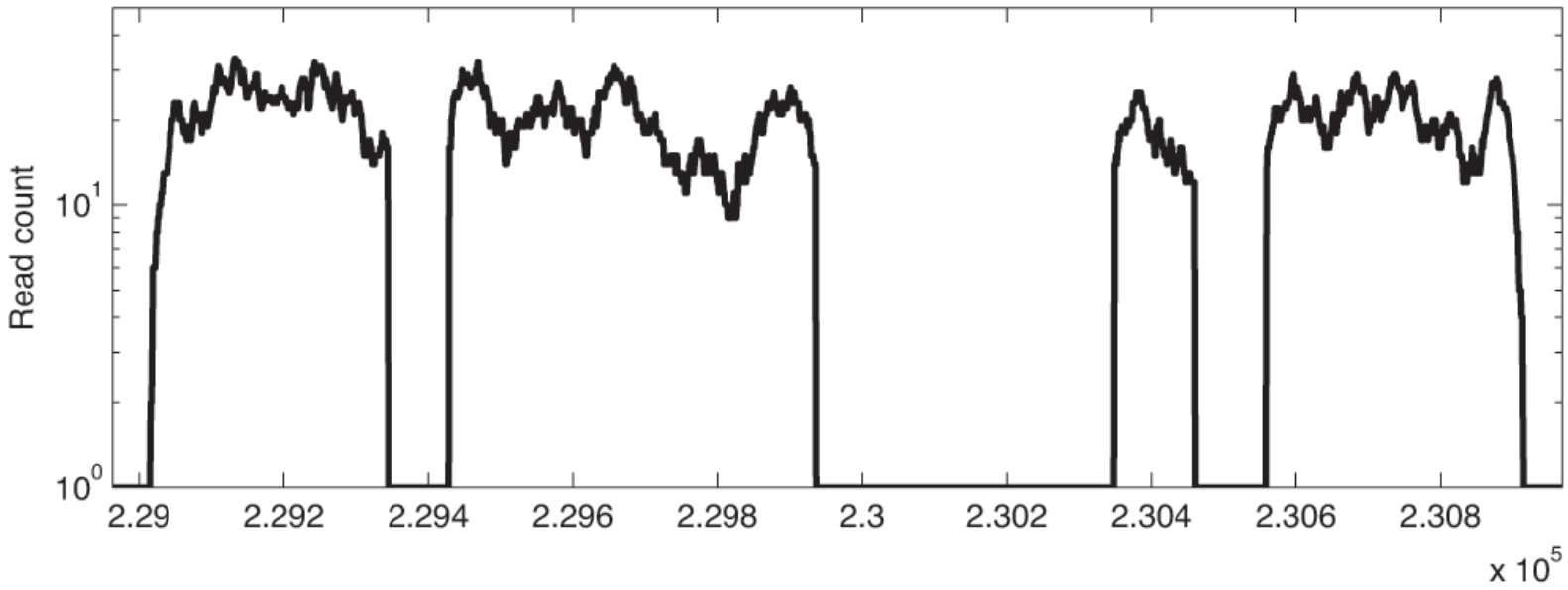
Extension to Alternative Transcripts

Does it work for the identification and quantification of alternatively spliced transcripts?

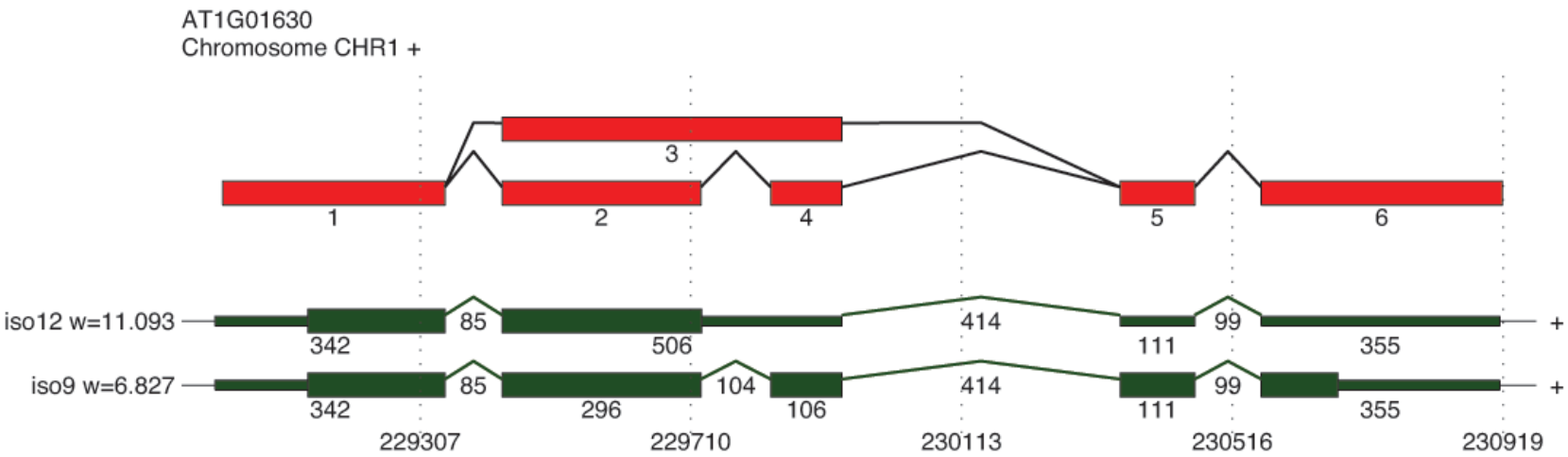
Extension to Alternative Transcripts



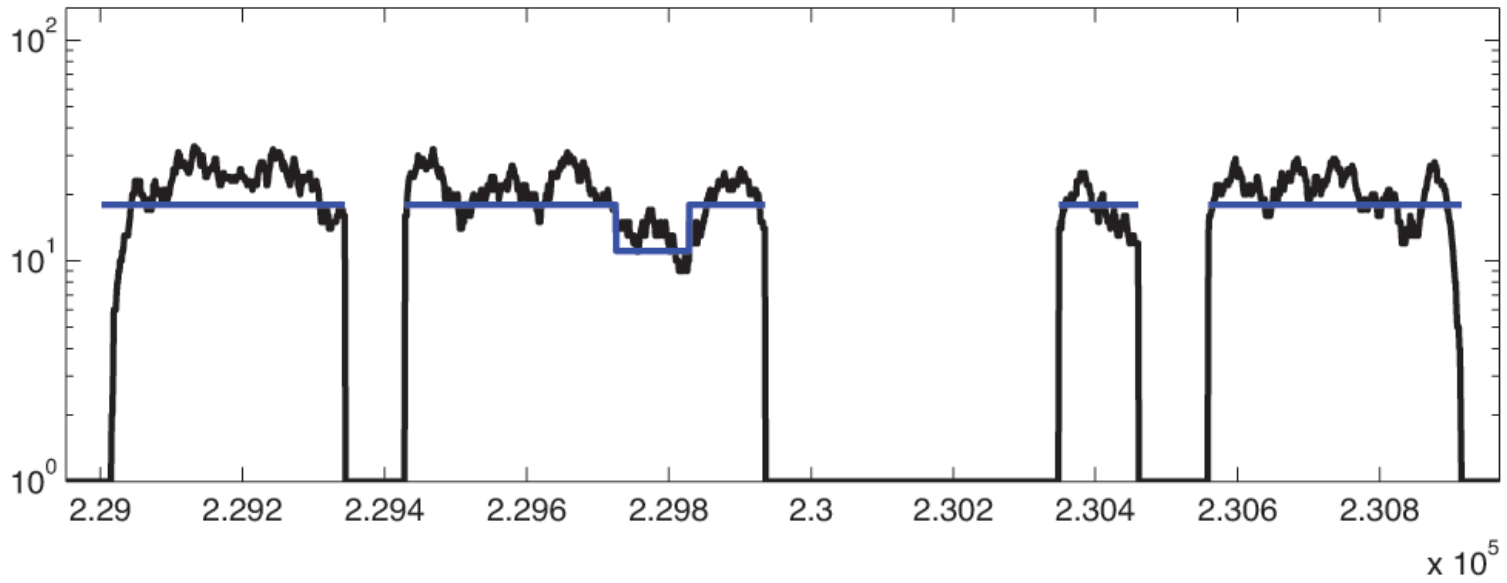
Simulated reads from two transcripts with equal abundance.



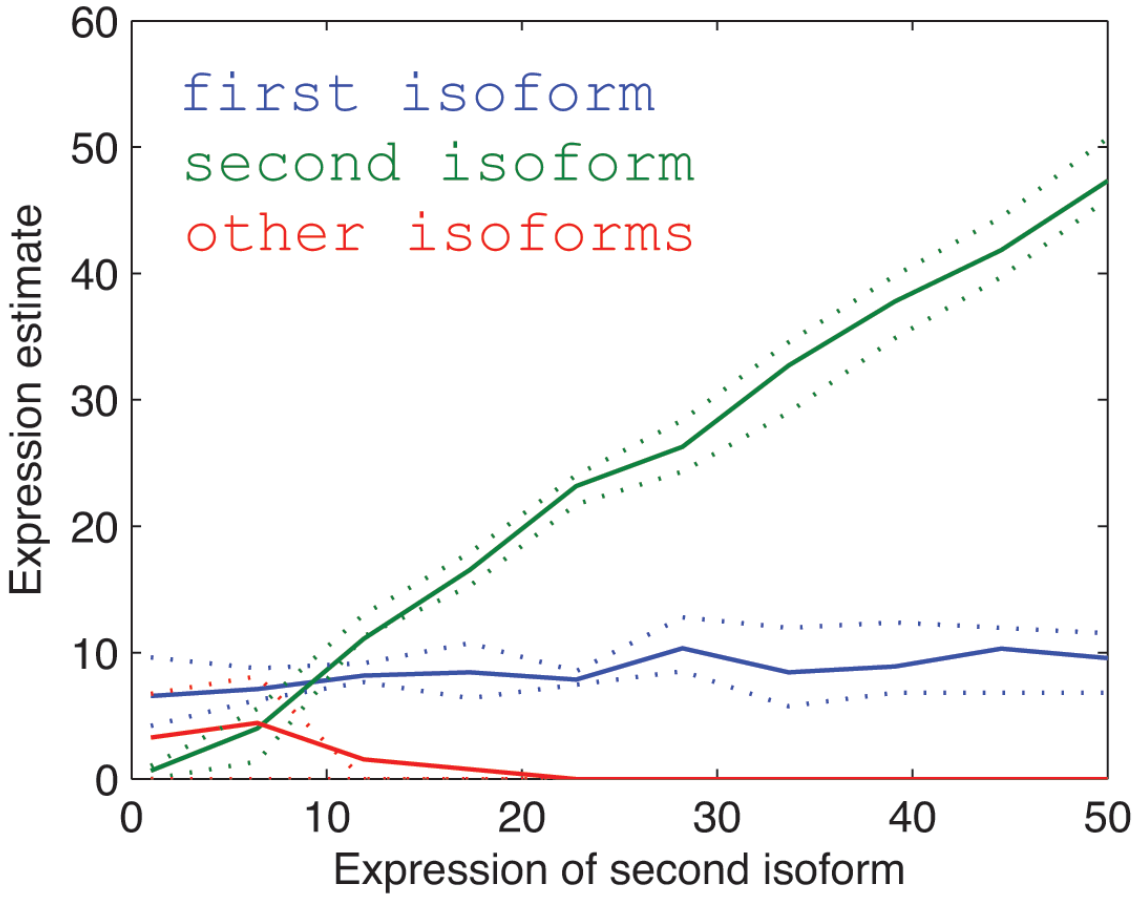
Extension to Alternative Transcripts



Correctly determined isoforms and good abundance estimate.



Extension to Alternative Transcripts



Abundance of 1^{st} transcript is 10, of second is varied. Shown are mean, 25 and 75 percentile of 10 runs.

Summary

- Trained **mGene** for accurate gene prediction in *A. thaliana*
- Extended system to integrate *tiling array* and *RNA-seq* data
- Both data types help to considerably *improve transcript level accuracy*
- Even tiny amounts of RNA-seq read help to improve accuracy
- Can use it to make *tissue specific predictions*
- Iterative residual fitting gets predictions closer to observations
- Prediction (of a few) alternative isoforms including abundances seems feasible

Summary

- Trained **mGene** for accurate gene prediction in *A. thaliana*
- Extended system to integrate *tiling array* and *RNA-seq* data
- Both data types help to considerably *improve transcript level accuracy*
- Even tiny amounts of RNA-seq read help to improve accuracy
- Can use it to make *tissue specific predictions*
- Iterative residual fitting gets predictions closer to observations
- Prediction (of a few) alternative isoforms including abundances seems feasible

Summary

- Trained **mGene** for accurate gene prediction in *A. thaliana*
- Extended system to integrate *tiling array* and *RNA-seq* data
- Both data types help to considerably *improve transcript level accuracy*
- Even tiny amounts of RNA-seq read help to improve accuracy
- Can use it to make *tissue specific predictions*
- Iterative residual fitting gets predictions closer to observations
- Prediction (of a few) alternative isoforms including abundances seems feasible

Summary

- Trained **mGene** for accurate gene prediction in *A. thaliana*
- Extended system to integrate *tiling array* and *RNA-seq* data
- Both data types help to considerably *improve transcript level accuracy*
- Even tiny amounts of RNA-seq read help to improve accuracy
- Can use it to make *tissue specific predictions*
- Iterative residual fitting gets predictions closer to observations
- Prediction (of a few) alternative isoforms including abundances seems feasible

Thank You for Your Attention!

Acknowledgments & links:

- mGene: *Jonas Behr*, *Gabriele Schweikert*, Georg Zeller, Alexander Zien, Sören Sonnenburg
- Tiling arrays and normalization: *Georg Zeller*, Sascha Laubinger, Detlef Weigel
- RNA-seq and mapping: *Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger*, Jun Cao, Christa Lanz, Detlef Weigel
  ShoRe: `http://1001genomes.org`
  QPalma: `http://www.fml.mpg.de/raetsch/projects/qpalma`
- A large scale open source Machine Learning toolbox:
  `http://www.shogun-toolbox.org`
- We are hiring ...
  `http://www.fml.mpg.de/raetsch/offers`