Machine Learning Methods for RNA-seq-based Transcriptome Reconstruction

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society, Tübingen, Germany

NGS Bioinformatics Meeting, Paris (March 24, 2010)

Friedrich Miescher Laboratory
of the Max Planck Society

MAX-PLANCK-GESELLSCHAFT

Discovery of the Nuclein
(Friedrich Miescher, 1869)

Discovery of Nuclein:

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

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

Discovery of the Nuclein
(Friedrich Miescher, 1869)

Tübingen, around 1869

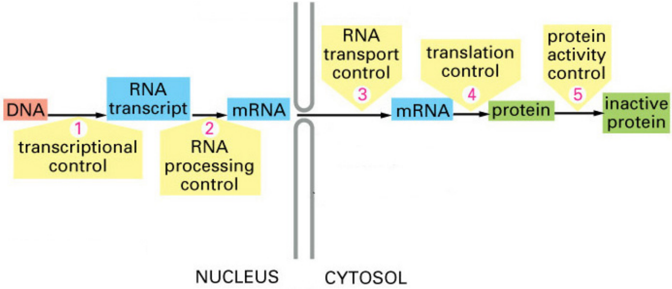
Discovery of Nuclein:

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

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

Learning about the Transcriptome

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



Computational Point of View

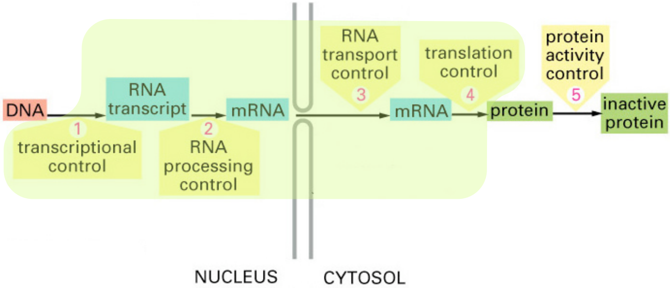
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

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 Point of View

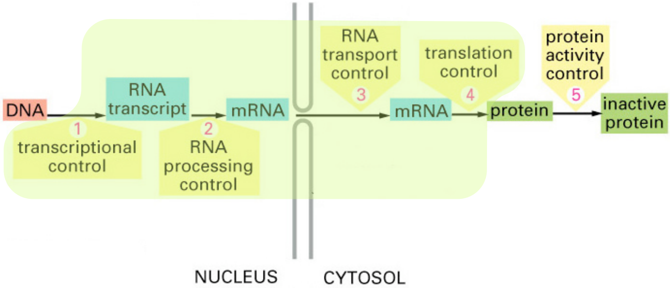
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

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 Point of View

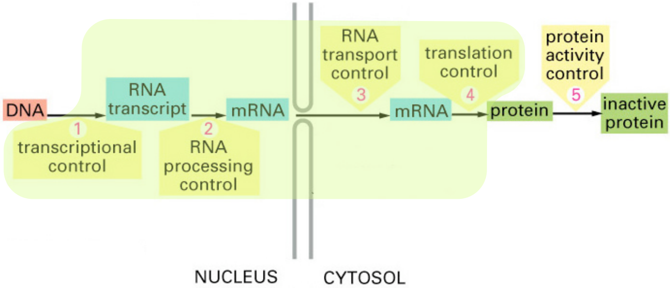
- How to learn to predict what these processes accomplish?
- How well can we predict it from the available information?

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 Point of 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
Learning from empirical observations

Given: Observations of some complex phenomenon

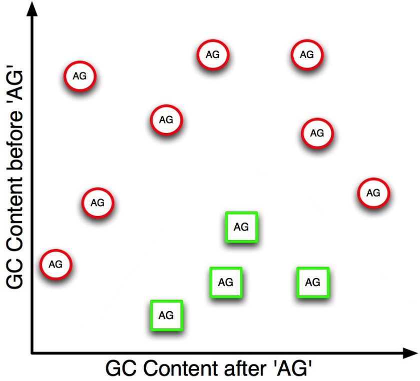
Goal: Learn from data & build predictive models

Machine Learning
Learning from empirical observations

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

Example:



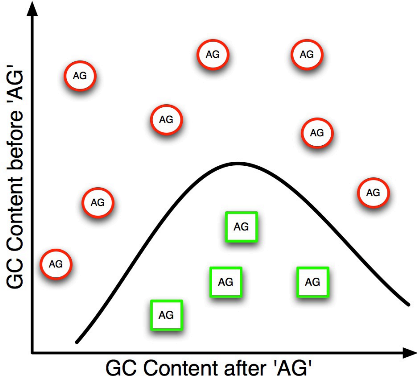
Two different classes of observations

Machine Learning
Learning from empirical observations

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

Example:



Inferred classification rule

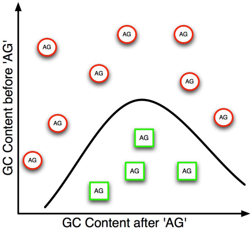
Machine Learning

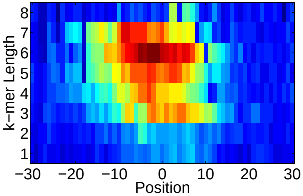
Learning from empirical observations

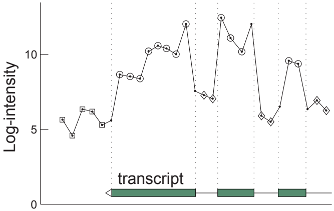
Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

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







Deep RNA Sequencing (RNA-Seq)

RNA-Seq allows . . .

- High-throughput transcriptome measurements
- Qualitative studies
  - New transcripts
  - Improved gene models
- Quantitative studies at high resolution
  - Differential expression in tissues, conditions, genotypes, etc.

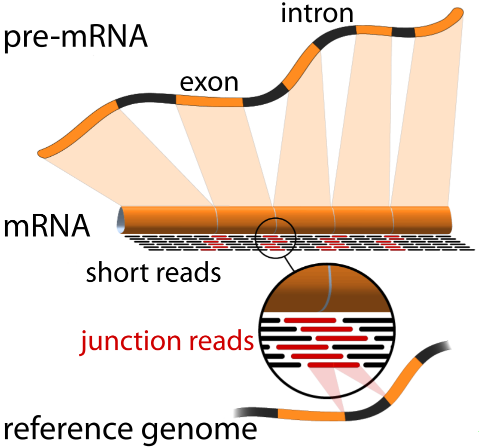


Figure adapted from Wikipedia

Goal: Obtain complete transcriptome for further analyses

Deep RNA Sequencing (RNA-Seq)

RNA-Seq allows . . .

- High-throughput transcriptome measurements
- Qualitative studies
  - New transcripts
  - Improved gene models
- Quantitative studies at high resolution
  - Differential expression in tissues, conditions, genotypes, etc.

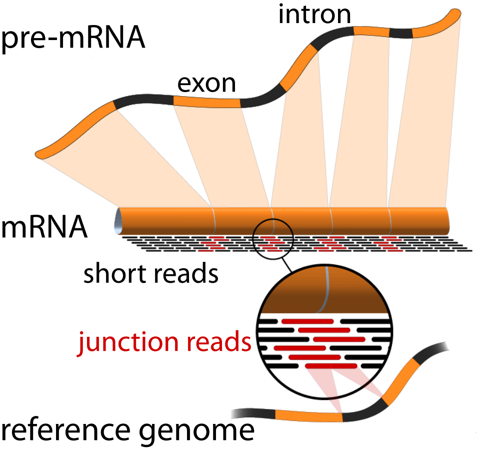


Figure adapted from Wikipedia

Goal: Obtain complete transcriptome for further analyses

Deep RNA Sequencing (RNA-Seq)

RNA-Seq allows . . .

- High-throughput transcriptome measurements
- Qualitative studies
  - New transcripts
  - Improved gene models
- Quantitative studies at high resolution
  - Differential expression in tissues, conditions, genotypes, etc.

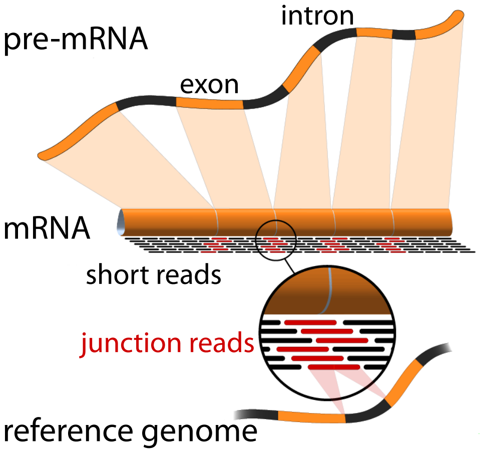


Figure adapted from Wikipedia

Goal: Obtain complete transcriptome for further analyses

Deep RNA Sequencing (RNA-Seq)

RNA-Seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
  - New transcripts
  - Improved gene models
- Quantitative studies at high resolution
  - Differential expression in tissues, conditions, genotypes, etc.

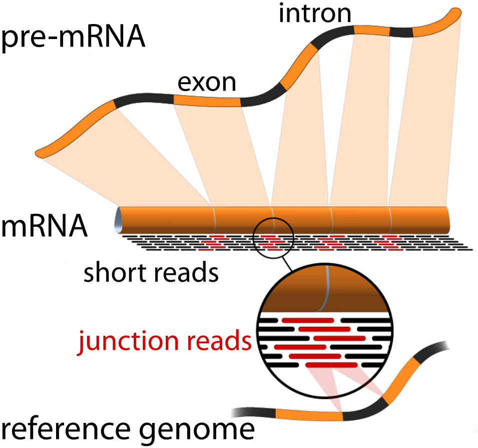
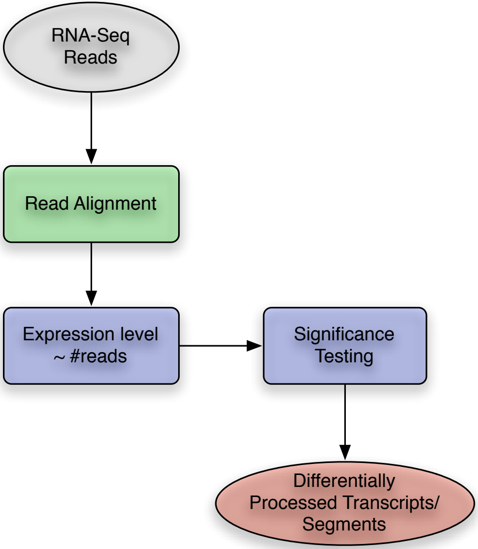


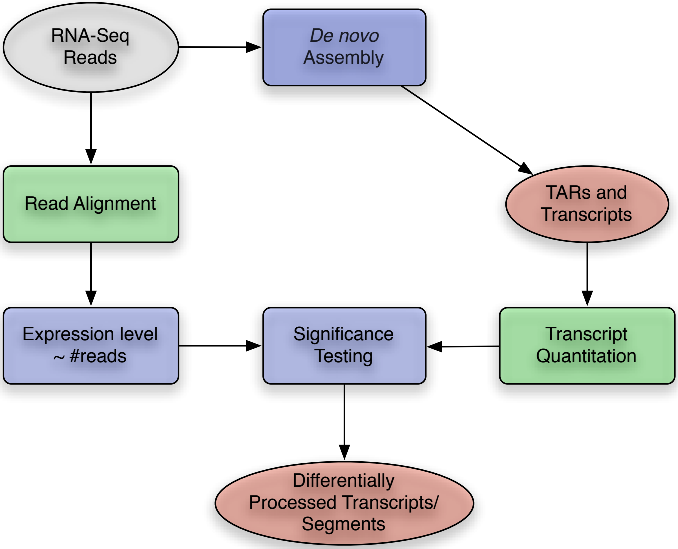
Figure adapted from Wikipedia

Goal: Obtain complete transcriptome for further analyses

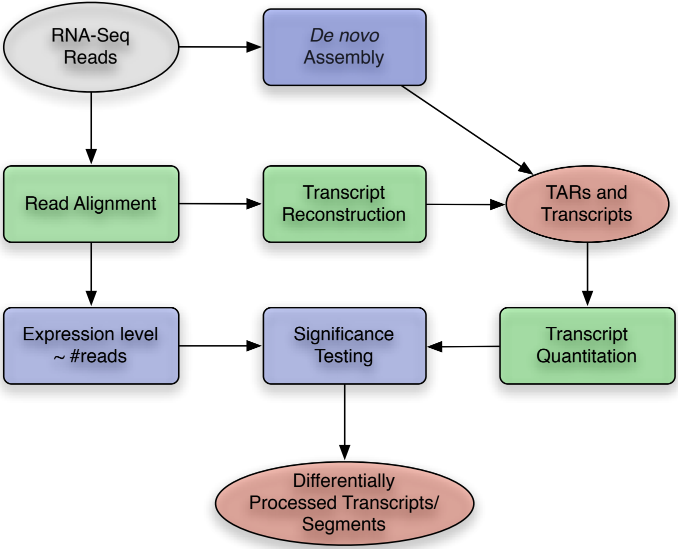
Common RNA-Seq Analysis Steps


RNA-Seq Reads
Read Alignment
Expression level ~ #reads
Significance Testing
Differentially Processed Transcripts/ Segments

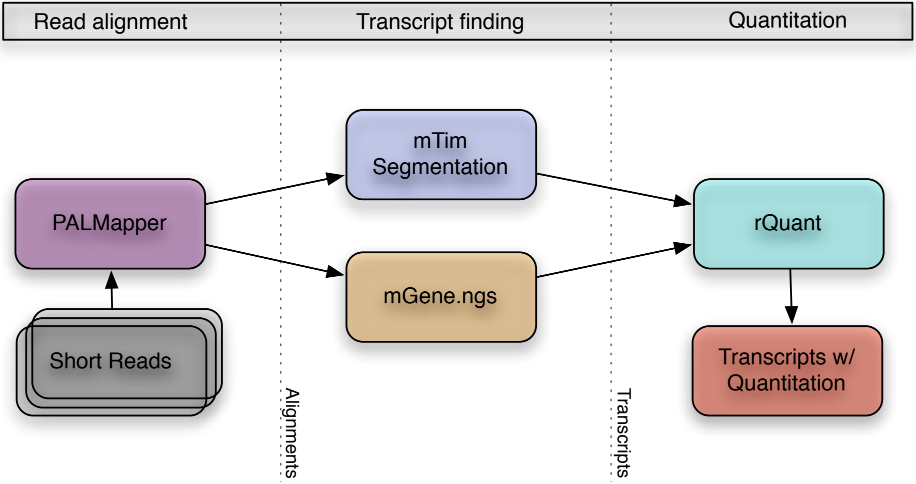
Common RNA-Seq Analysis Steps



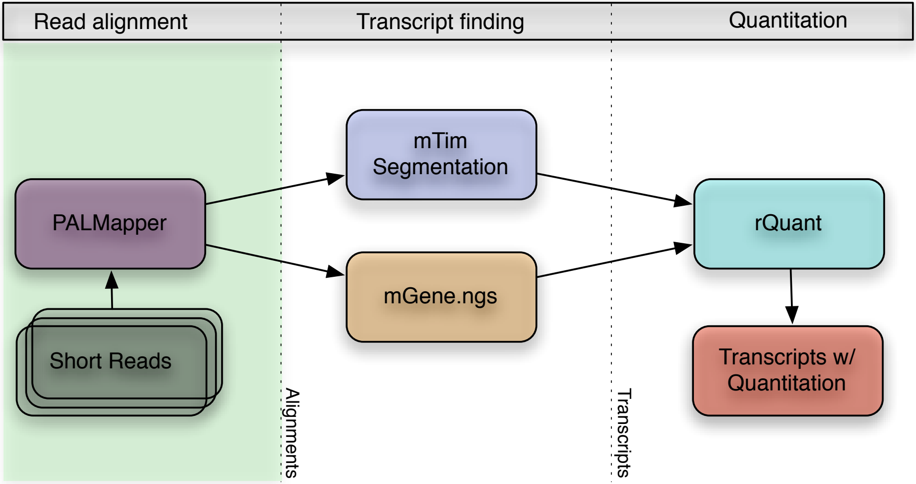
Common RNA-Seq Analysis Steps



RNA-Seq Pipeline Overview



RNA-Seq Pipeline Overview



Step 1: PALMapper Read Alignment

(PALMapper = QPALMA + GenomeMapper)

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes

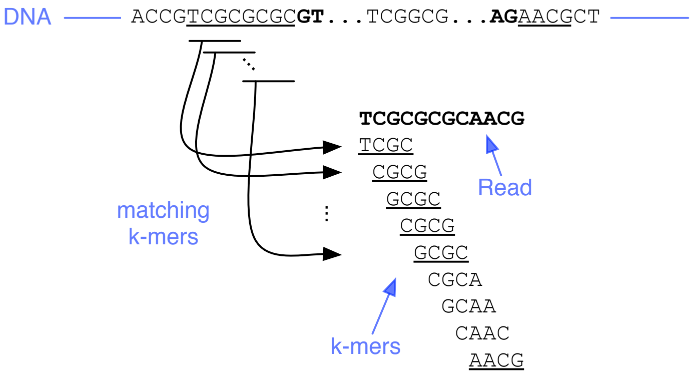
More info: `http://fml.mpg.de/raetsch/suppl/palmapper`

Step 1: PALMapper Read Alignment

(PALMapper = QPALMA + GenomeMapper)

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes

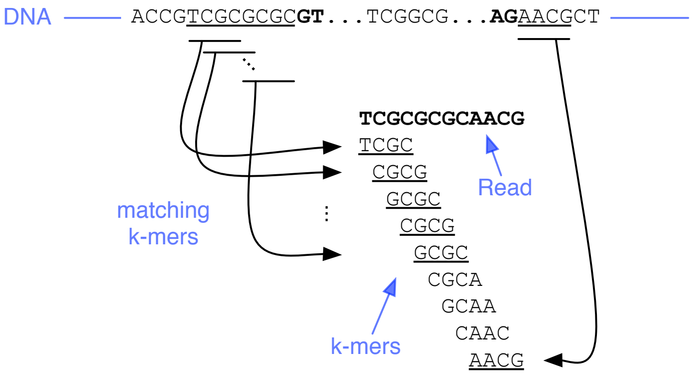


Step 1: PALMapper Read Alignment

(PALMapper = QPALMA + GenomeMapper)

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes



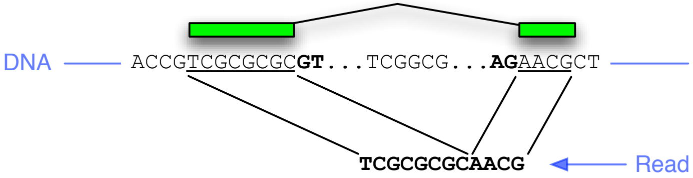
Step 1: PALMapper Read Alignment
(PALMapper = QPALMA + GenomeMapper)

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:

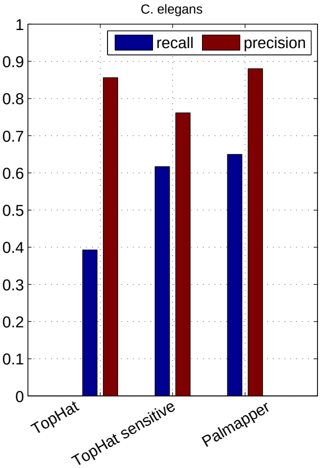
- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA [De Bona et al., 2008]



More info: `http://fml.mpg.de/raetsch/suppl/palmapper`

PALMapper Accuracy Evaluation

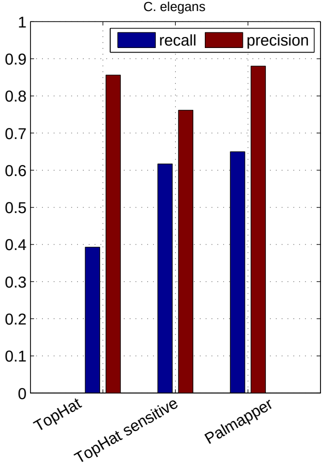
How accurately can PALMapper identify introns?

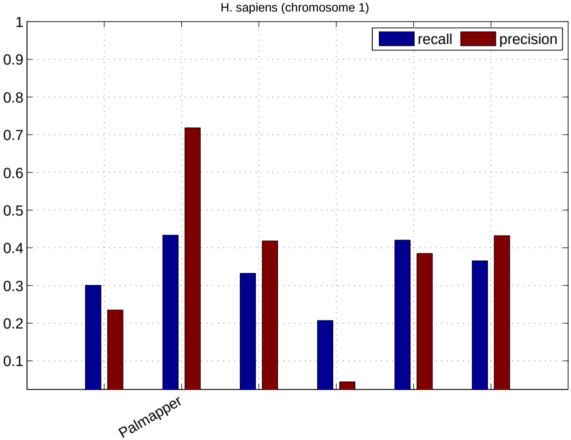


PALMapper (3.5h) and TopHat (3.5h/10h) aligning 24M reads

PALMapper Accuracy Evaluation

How accurately can PALMapper identify introns?

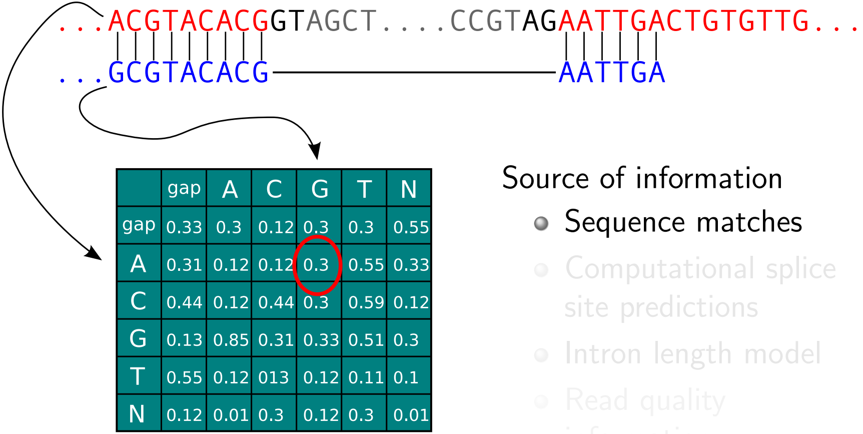




PALMapper (3.5h) and TopHat (3.5h/10h) aligning 24M reads

Comparison of PALMapper with other alignment programs within the RGASP project (preliminary)

QPALMA: Extended Smith-Waterman Scoring



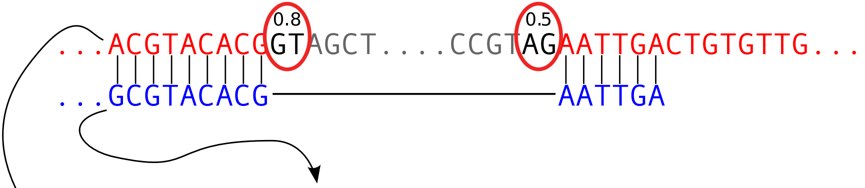
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

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

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

QPALMA: Extended Smith-Waterman Scoring



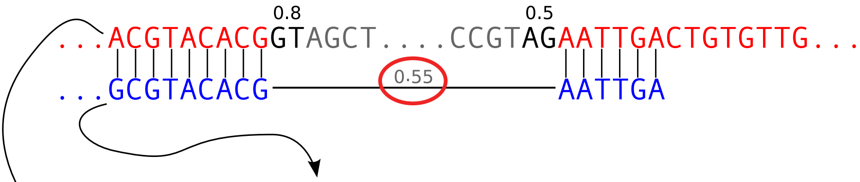
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

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

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

QPALMA: Extended Smith-Waterman Scoring



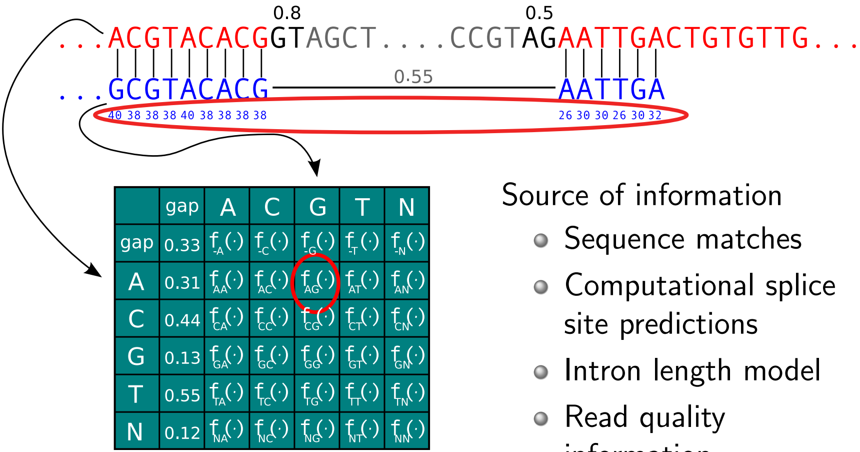
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

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

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

QPALMA: Extended Smith-Waterman Scoring



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

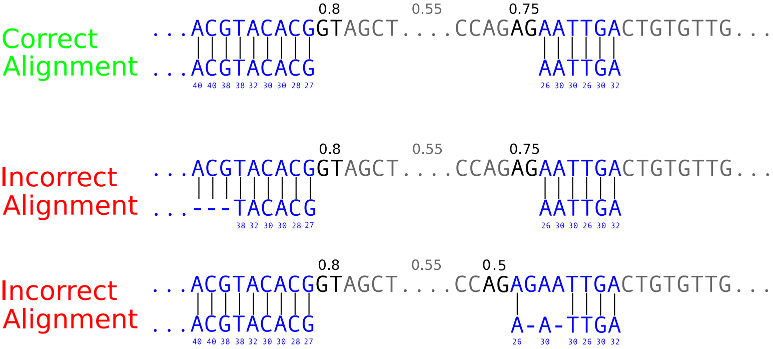
[De Bona et al., 2008]

Scoring Parameter Inference

- What are optimal parameters?
- How do we jointly optimize the 336 parameters?

Scoring Parameter Inference

- What are optimal parameters?
- How do we jointly optimize the 336 parameters?



Correct Alignment

```
                0.8       0.55     0.75
...ACGTACACGGTAGCT....CCAGAGAATTGACTGTGTTG...
   |||||||||                  ||||||
...ACGTACACG                  AATTGA
   40 40 38 38 32 30 30 28 27   26 30 30 26 30 32
```

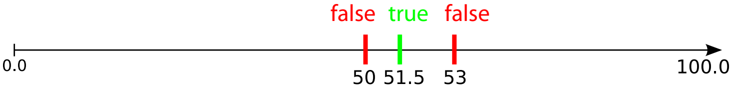
Incorrect Alignment

```
                0.8       0.55     0.75
...ACGTACACGGTAGCT....CCAGAGAATTGACTGTGTTG...
   |||||||||                  ||||||
...---TACACG                  AATTGA
   38 32 30 30 28 27           26 30 30 26 30 32
```

Incorrect Alignment

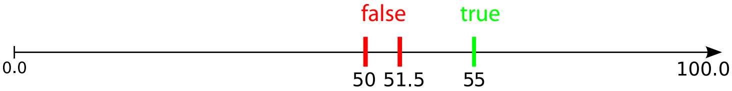
```
                0.8       0.55   0.5
...ACGTACACGGTAGCT....CCAGAGAATTGACTGTGTTG...
   |||||||||                |   ||||
...ACGTACACG                A-A-TTGA
   40 40 38 38 32 30 30 28 27   26 30 30 26 30 32
```

Cartoon: Maximize the Margin



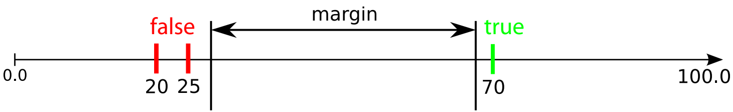
- Correct alignment is **not** highest scoring one
- Correct alignment is highest scoring one
- Can we do better?

Cartoon: Maximize the Margin



- Correct alignment is **not** highest scoring one
- Correct alignment is highest scoring one
- Can we do better?

Cartoon: Maximize the Margin



- Technique motivated by SVMs ("large-margin")
- Enforce a margin between correct and incorrect examples
- One has to solve a big quadratic problem

How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads with known origin

Strategy:

1. Estimate relationship between quality score and error probability from given reads
2. Use annotation of a few genes to simulate spliced reads
3. Introduce errors according to error model using quality strings from given read set
4. Train QPalma on generated read set with known alignments

How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads with known origin

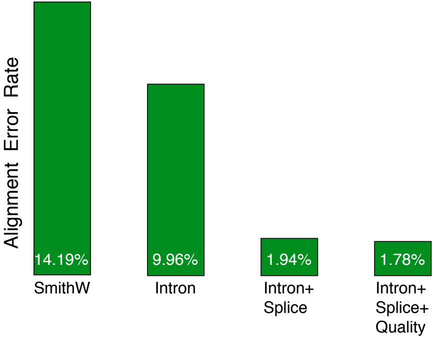
Strategy:

1. Estimate relationship between quality score and error probability from given reads
2. Use annotation of a few genes to simulate spliced reads
3. Introduce errors according to error model using quality strings from given read set
4. Train QPalma on generated read set with known alignments

QPALMA RNA-Seq Read Alignment

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

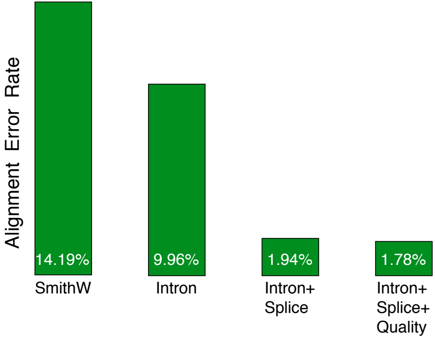


[De Bona et al., 2008]

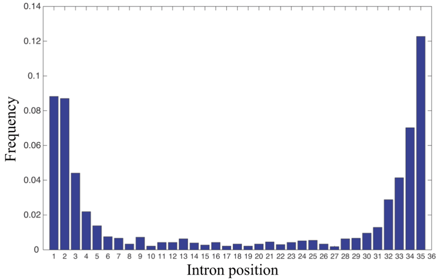
QPALMA RNA-Seq Read Alignment

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



Error vs. intron position

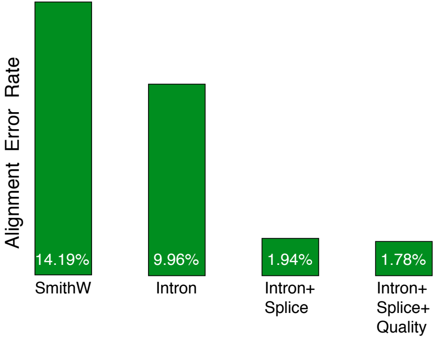


[De Bona et al., 2008]

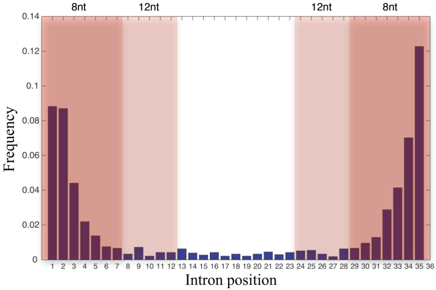
QPALMA RNA-Seq Read Alignment

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

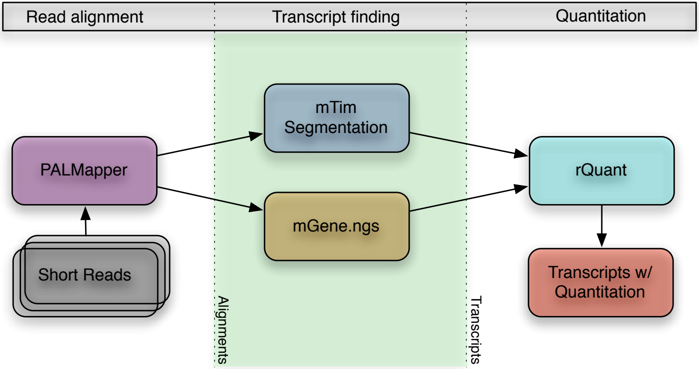


Error vs. intron position



[De Bona et al., 2008]

Step 2: Transcript Prediction

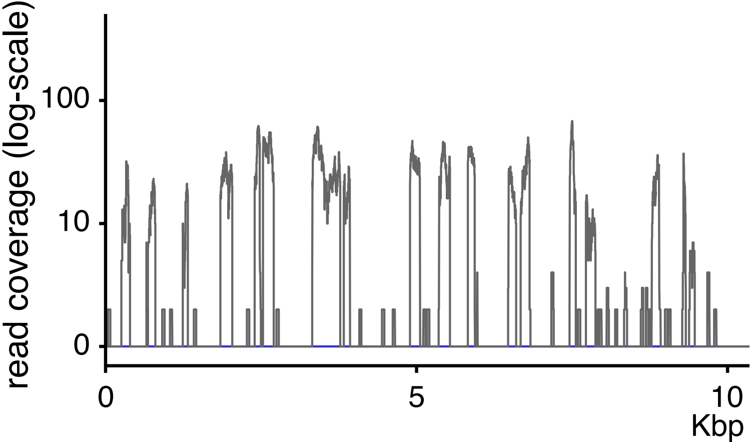


A. Coverage segmentation algorithm **mTiM** for general transcripts (no coding bias/assumption)

B. Extension of the mGene gene finding system to use NGS data for protein coding transcript prediction (**mGene.ngs**)

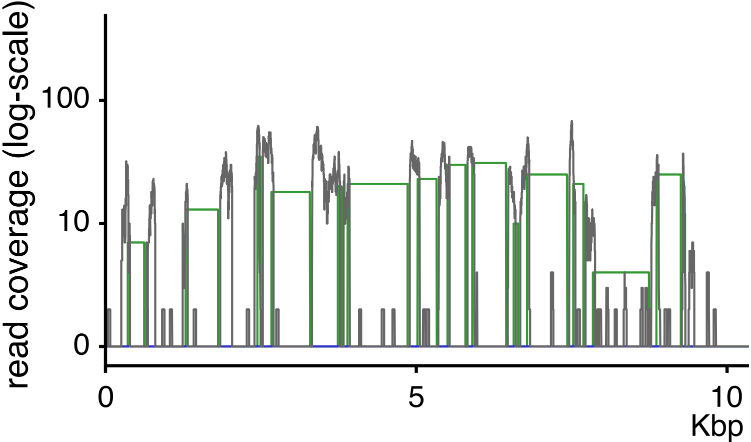
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



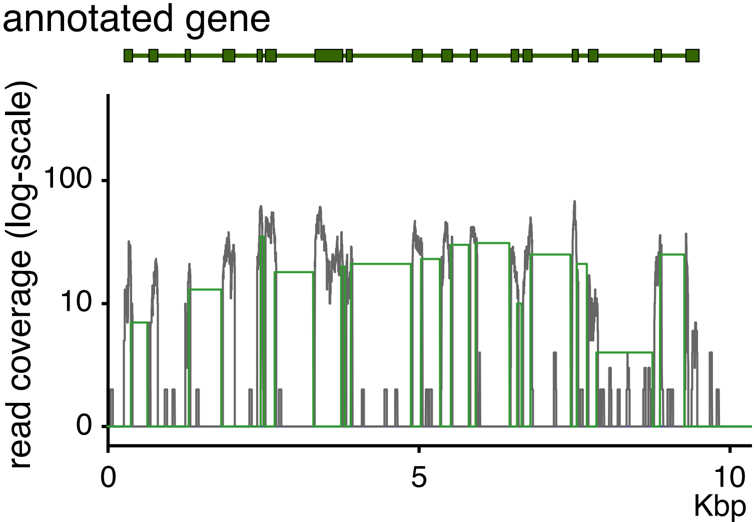
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*

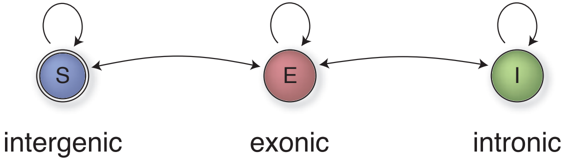


mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



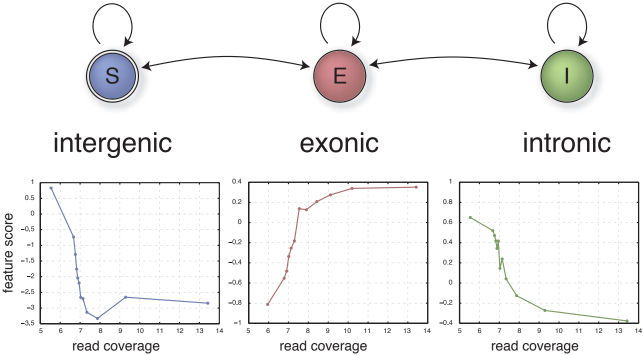
The mTiM Segmentation Approach



- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal → score
- Extension: Score spliced reads and splice sites

(G. Zeller et al., 2008; G. Zeller et al., in prep., 2009)

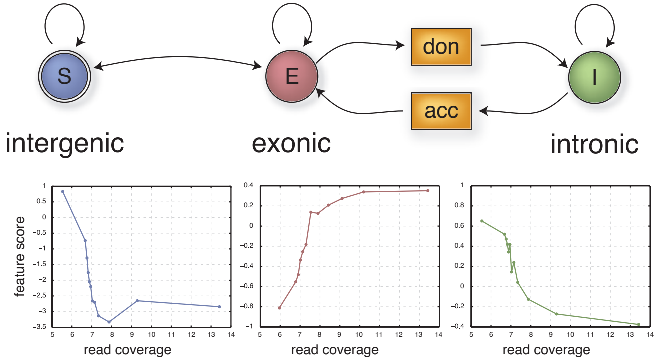
The mTiM Segmentation Approach



- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal → score
- Extension: Score spliced reads and splice sites

(G. Zeller et al., 2008; G. Zeller et al., in prep., 2009)

The mTiM Segmentation Approach

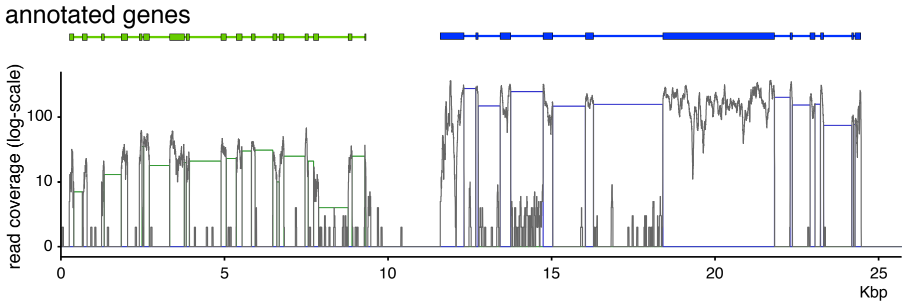


- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal → score
- Extension: Score spliced reads and splice sites

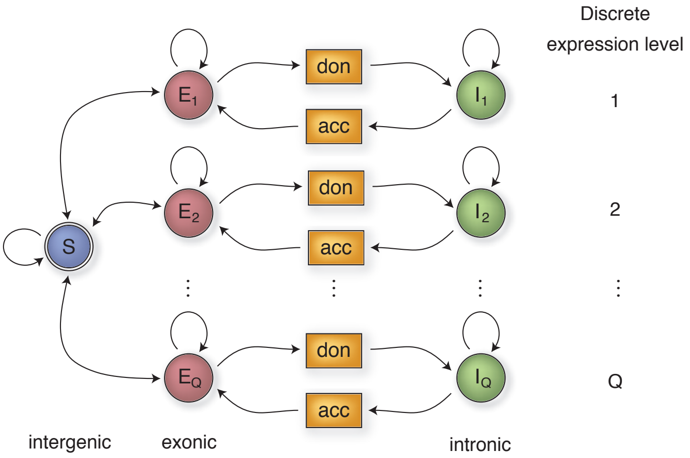
(G. Zeller et al., 2008; G. Zeller et al., in prep., 2009)

The mTiM Segmentation Approach

Idea: Assume uniform read coverage within exons of same transcript



The mTiM Segmentation Approach



Carry “expression level” information between exons of same transcript

(G. Zeller et al., 2008; G. Zeller et al., in prep., 2010)

Discriminative training of HM-SVMs

$f : \mathbb{R}^\star \rightarrow \Sigma^\star$

given a sequence of hybridization measurements $\boldsymbol{\chi} \in \mathbb{R}^\star$
predicts a state sequence (path) $\boldsymbol{\sigma} \in \Sigma^\star$

Discriminant function $F_{\boldsymbol{\theta}} : \mathbb{R}^\star \times \Sigma^\star \rightarrow \mathbb{R}$ such that for decoding: $f(\chi) = \underset{\boldsymbol{\sigma} \in \mathcal{S}^\star}{\operatorname{argmax}} F_{\boldsymbol{\theta}}(\chi, \boldsymbol{\sigma})$.

Training:

For each training example $(\chi^{(i)}, \boldsymbol{\sigma}^{(i)})$, enforce a large margin of separation

$$F_\theta(\chi^{(i)}, \boldsymbol{\sigma}^{(i)}) - F_\theta(\chi^{(i)}, \overline{\boldsymbol{\sigma}}) \geq \rho$$

between the correct path $\boldsymbol{\sigma}^{(i)}$ and *any* other wrong path $\overline{\boldsymbol{\sigma}} \neq \boldsymbol{\sigma}^{(i)}$.

A quadratic programming problem (QP) is solved to optimize $\boldsymbol{\theta}$.

[Altun et al., 2003, Rätsch et al., 2007, Zeller et al., 2008b]

Discriminative training of HM-SVMs

$f : \mathbb{R}^\star \rightarrow \Sigma^\star$

given a sequence of hybridization measurements $\boldsymbol{\chi} \in \mathbb{R}^\star$ predicts a state sequence (path) $\boldsymbol{\sigma} \in \Sigma^\star$

Discriminant function $F_{\boldsymbol{\theta}} : \mathbb{R}^\star \times \Sigma^\star \rightarrow \mathbb{R}$ such that for decoding: $f(\boldsymbol{\chi}) = \underset{\boldsymbol{\sigma} \in \mathcal{S}^\star}{\operatorname{argmax}} F_{\boldsymbol{\theta}}(\boldsymbol{\chi}, \boldsymbol{\sigma})$.

Training:

For each training example $(\chi^{(i)}, \boldsymbol{\sigma}^{(i)})$, enforce a large margin of separation

$$F_\theta(\chi^{(i)}, \boldsymbol{\sigma}^{(i)}) - F_\theta(\chi^{(i)}, \overline{\boldsymbol{\sigma}}) \geq \rho$$

between the correct path $\boldsymbol{\sigma}^{(i)}$ and *any* other wrong path $\overline{\boldsymbol{\sigma}} \neq \boldsymbol{\sigma}^{(i)}$.

A quadratic programming problem (QP) is solved to optimize $\boldsymbol{\theta}$.

[Altun et al., 2003, Rätsch et al., 2007, Zeller et al., 2008b]

Discriminative training of HM-SVMs

$f : \mathbb{R}^\star \rightarrow \Sigma^\star$

given a sequence of hybridization measurements $\boldsymbol{\chi} \in \mathbb{R}^\star$ predicts a state sequence (path) $\boldsymbol{\sigma} \in \Sigma^\star$

Discriminant function $F_{\boldsymbol{\theta}} : \mathbb{R}^\star \times \Sigma^\star \rightarrow \mathbb{R}$ such that for decoding: $f(\boldsymbol{\chi}) = \underset{\boldsymbol{\sigma} \in \mathcal{S}^\star}{\operatorname{argmax}} F_{\boldsymbol{\theta}}(\boldsymbol{\chi}, \boldsymbol{\sigma})$.

Training:

For each training example $(\boldsymbol{\chi}^{(i)}, \boldsymbol{\sigma}^{(i)})$, enforce a large margin of separation

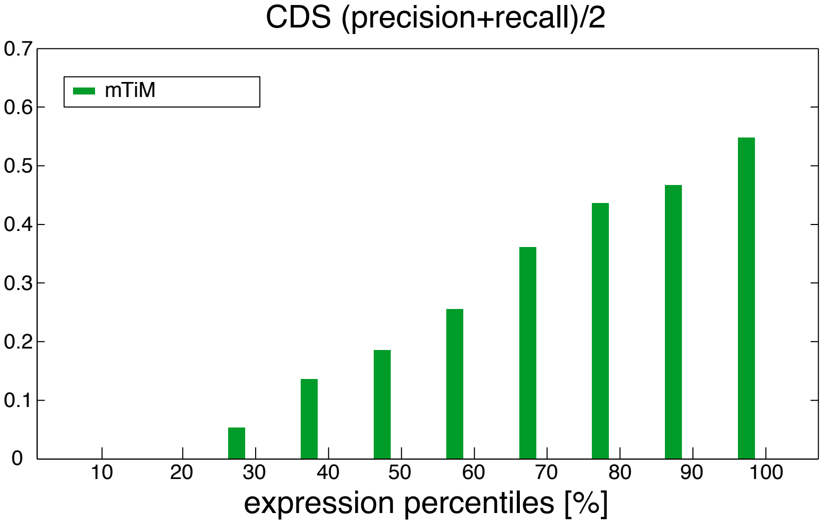
$$F_\theta(\boldsymbol{\chi}^{(i)}, \boldsymbol{\sigma}^{(i)}) - F_\theta(\boldsymbol{\chi}^{(i)}, \overline{\boldsymbol{\sigma}}) \geq \rho$$

between the correct path $\boldsymbol{\sigma}^{(i)}$ and *any* other wrong path $\overline{\boldsymbol{\sigma}} \neq \boldsymbol{\sigma}^{(i)}$.

A quadratic programming problem (QP) is solved to optimize $\boldsymbol{\theta}$.

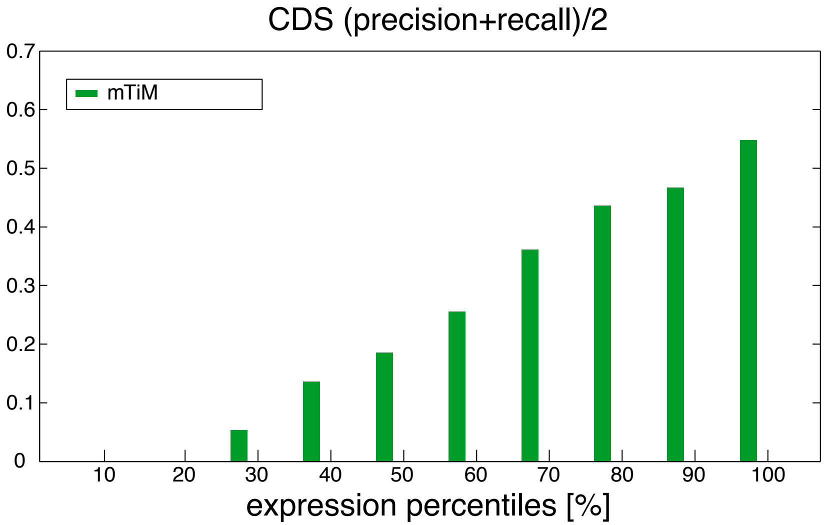
[Altun et al., 2003, Rätsch et al., 2007, Zeller et al., 2008b]

Preliminary Evaluation (*C. elegans*)



Sensitivity heavily depends on read density

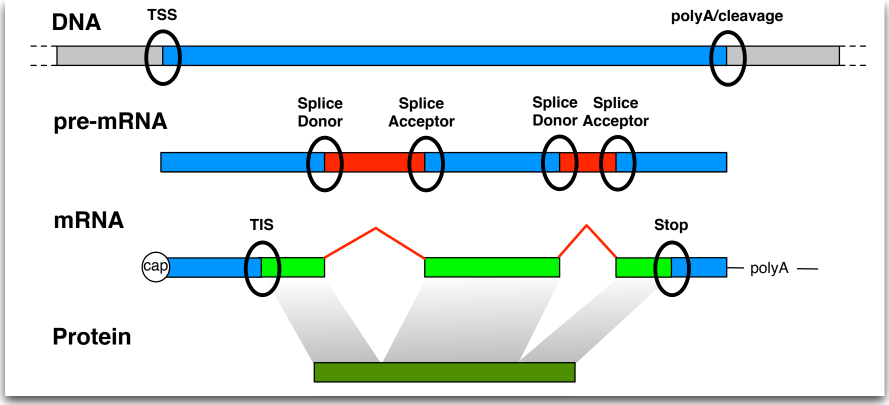
Preliminary Evaluation (*C. elegans*)



Sensitivity heavily depends on read density

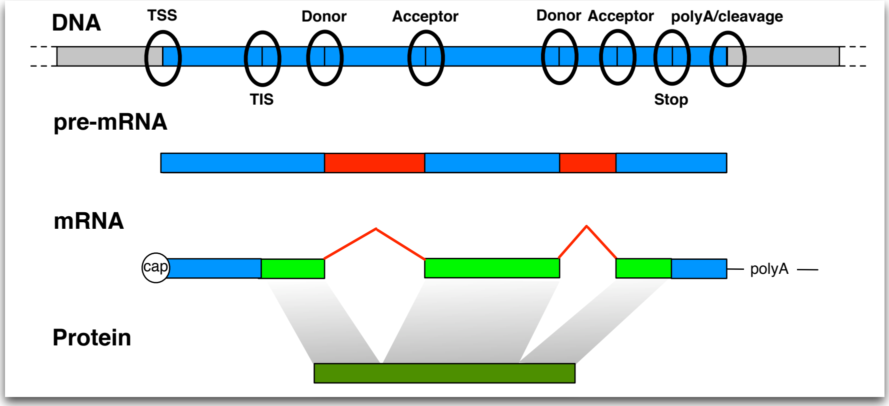
Computational Gene Finding

⇝ Labeling the Genome



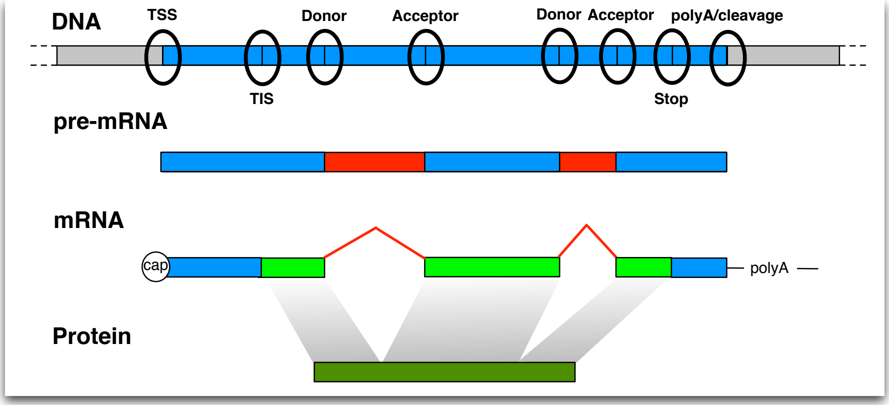
Computational Gene Finding

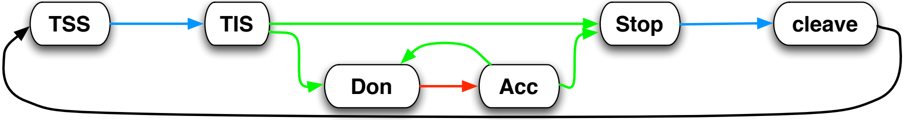
⇝ Labeling the Genome



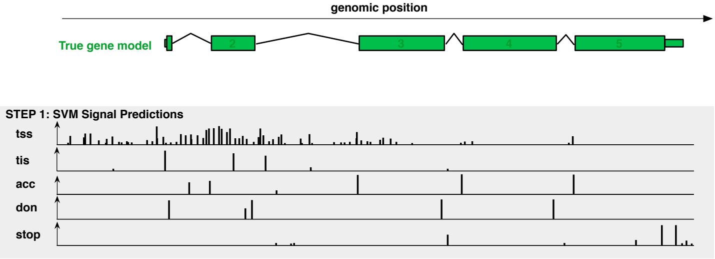
Computational Gene Finding

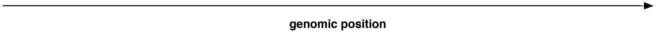
⇝ Labeling the Genome



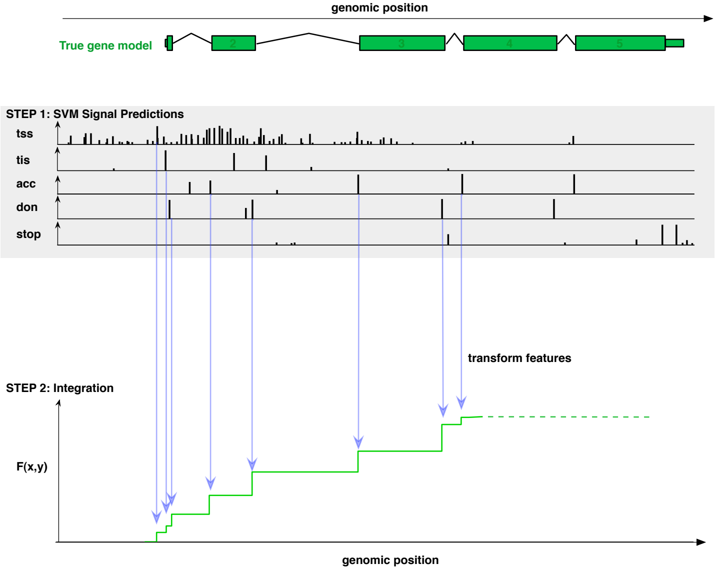


mGene-based Transcript Prediction

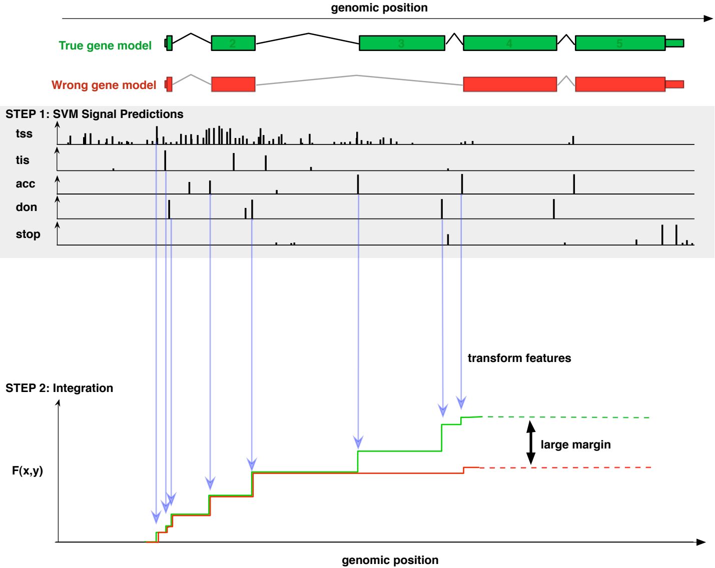




mGene-based Transcript Prediction



mGene-based Transcript Prediction



Learning to use Expression Measurements

Two approaches:

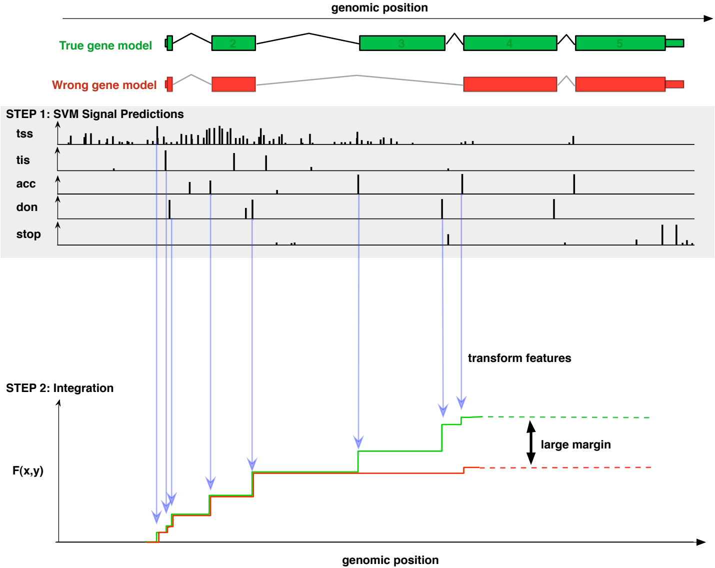
- Heuristic to incorporate ESTs/reads/tiling array measurements to *refine predictions*
- Directly *use evidence during learning* to learn to appropriately weight its importance

	Exon Level			Transcript Level		
	SN	**SP**	**F**	**SN**	**SP**	**F**
ab initio	82.3	82.6	82.5	43.1	49.5	46.1
ESTs heuristic	85.3	84.7	85.0	49.5	56.4	52.7
ESTs trained	84.8	85.8	**85.3**	50.5	57.8	**53.9**

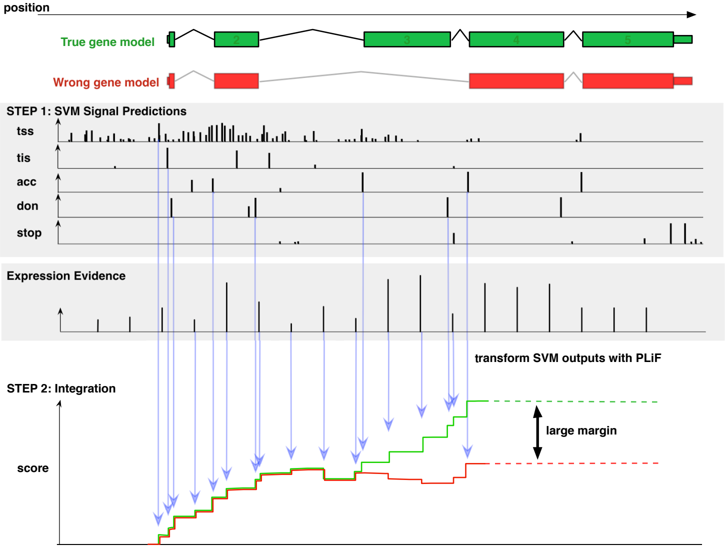
Gene prediction in *C. elegans* (CDS evaluation)

Behr et al., in pre., 2010

mGene-based Transcript Prediction



mGene-based Transcript Prediction



Learning to use Expression Measurements

Two approaches:

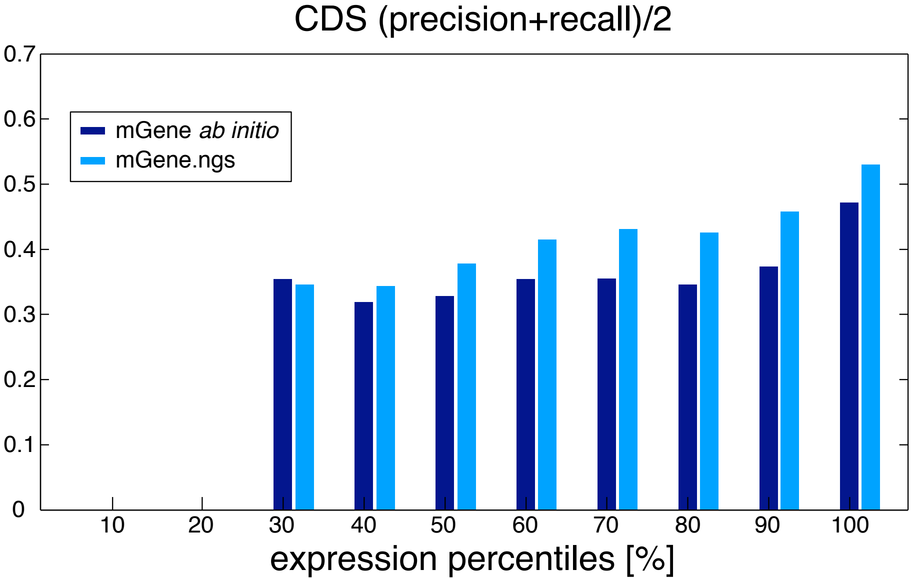
- Heuristic to incorporate ESTs/reads/tiling array measurements to *refine predictions*
- Directly *use evidence during learning* to learn to appropriately weight its importance

	Exon Level			Transcript Level		
	SN	**SP**	**F**	**SN**	**SP**	**F**
ab initio	82.3	82.6	82.5	43.1	49.5	46.1
ESTs heuristic	85.3	84.7	85.0	49.5	56.4	52.7
ESTs trained	84.8	85.8	85.3	50.5	57.8	53.9
RNA-Seq trained	84.6	84.9	84.8	49.1	55.2	52.0
RNA-Seq/ESTs trained	84.7	**86.9**	85.8	50.3	60.5	**54.9**

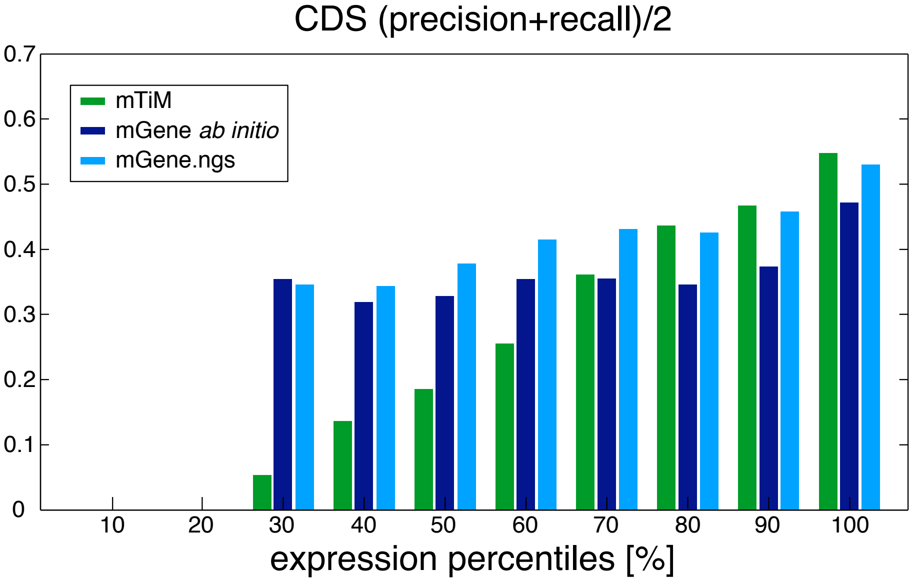
Gene prediction in *C. elegans* (CDS evaluation)

Behr et al., in prep., 2010

Preliminary Evaluation (*C. elegans*)



Preliminary Evaluation (*C. elegans*)



Digestion

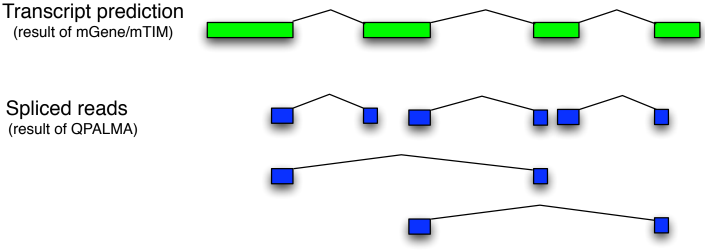
- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

Digestion

- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits “uniformity” of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

Digestion

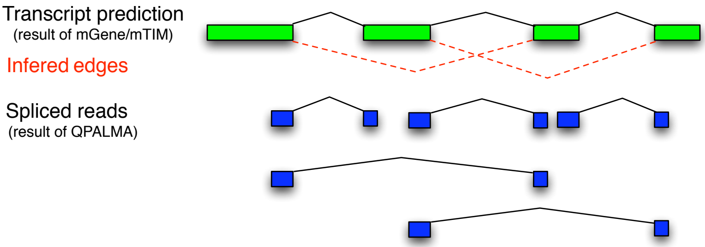
- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



- Paths through splicing graph define *alternative transcripts*

Digestion

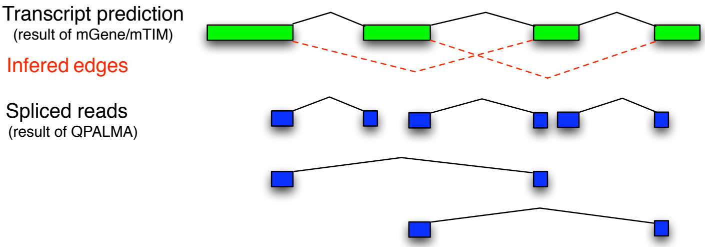
- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits “uniformity” of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alt. Transcripts:** Spliced reads for splicing graph completion:



- Paths through splicing graph define *alternative transcripts*

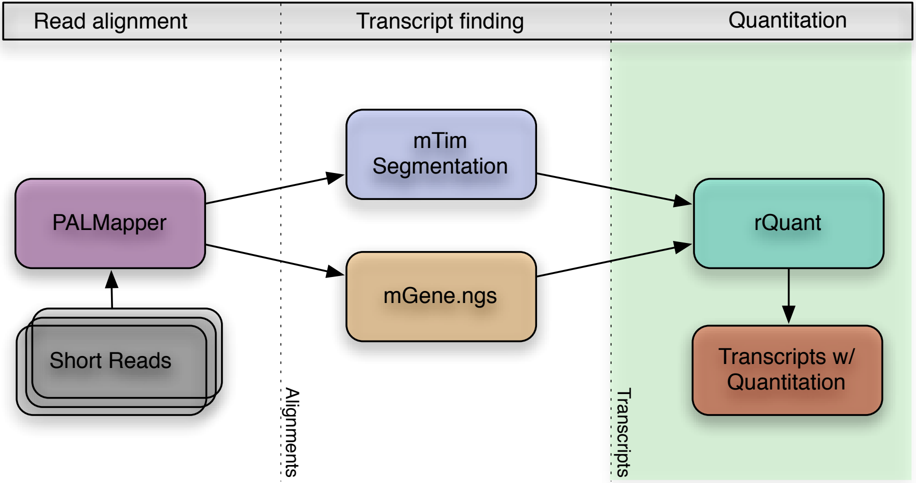
Digestion

- **mTiM** and **mGene.ngs** predict single transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **mGene.ngs** uses more assumptions on structure of transcripts
- **Alt. Transcripts:** Spliced reads for splicing graph completion:

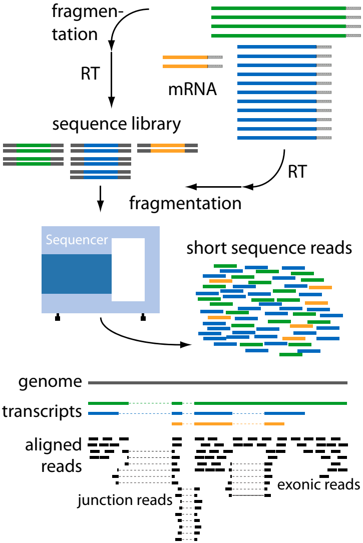


- Paths through splicing graph define *alternative transcripts*

RNA-Seq Pipeline Overview

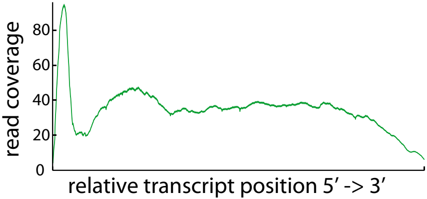


RNA-Seq Biases and Quantitation



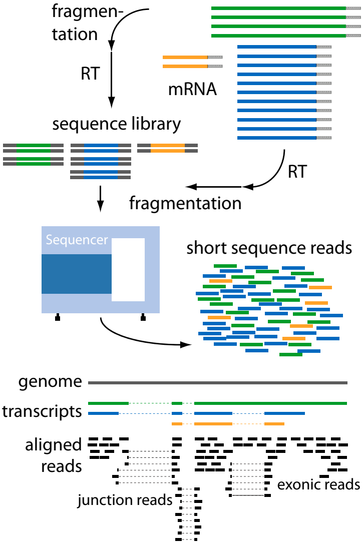
Biases due to . . .

- cDNA library construction
- Sequencing
- Read mapping



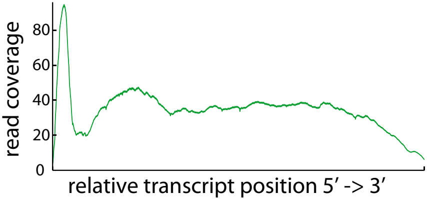
(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

RNA-Seq Biases and Quantitation



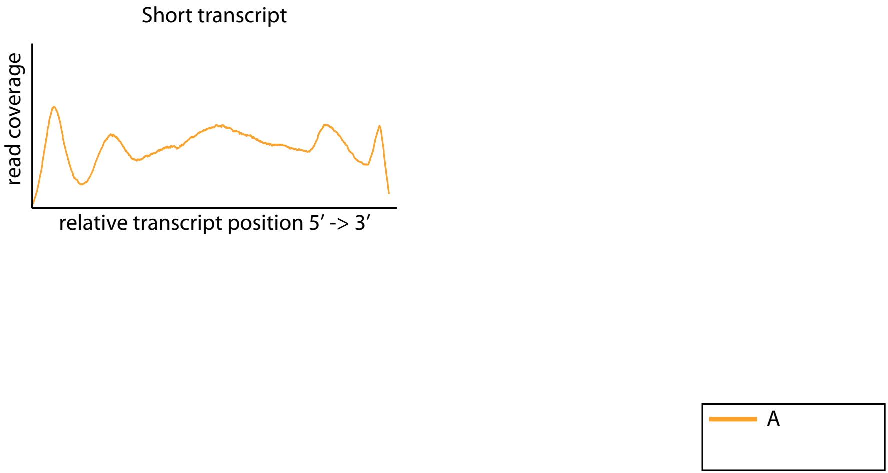
Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping



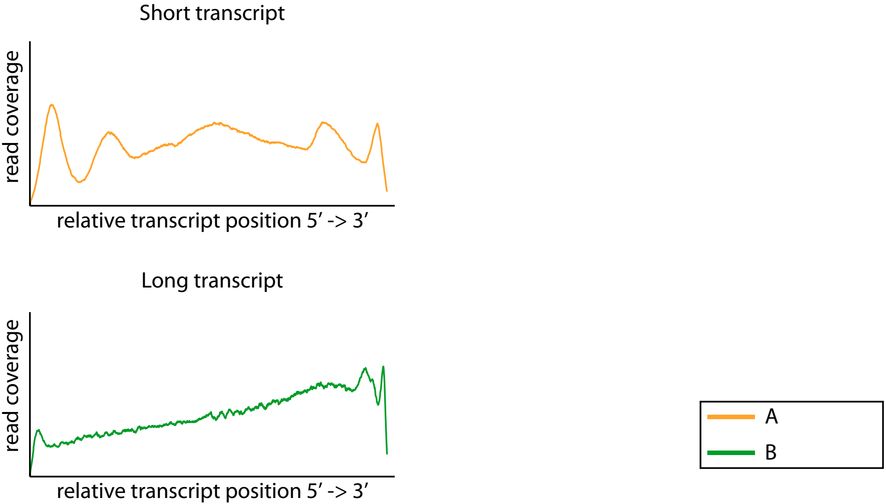
(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

rQuant – Basic Idea



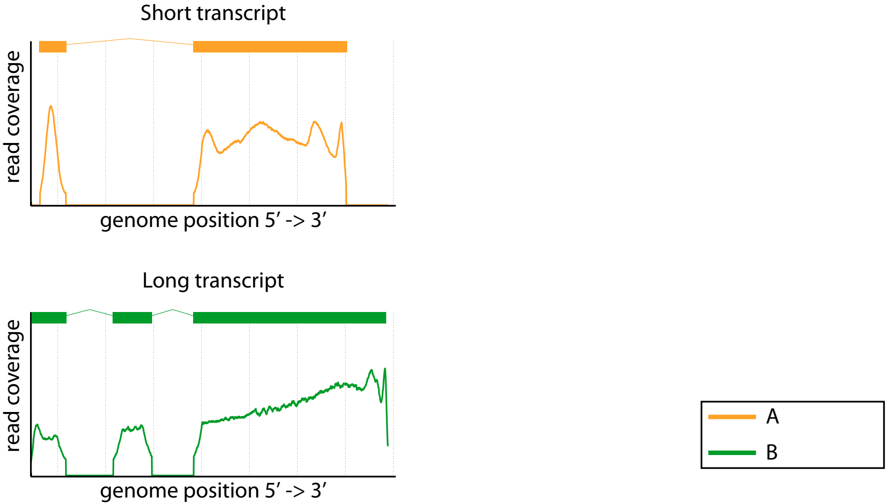
$$M_i = w_A A_i + w_B B_i \qquad \Rightarrow \qquad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea



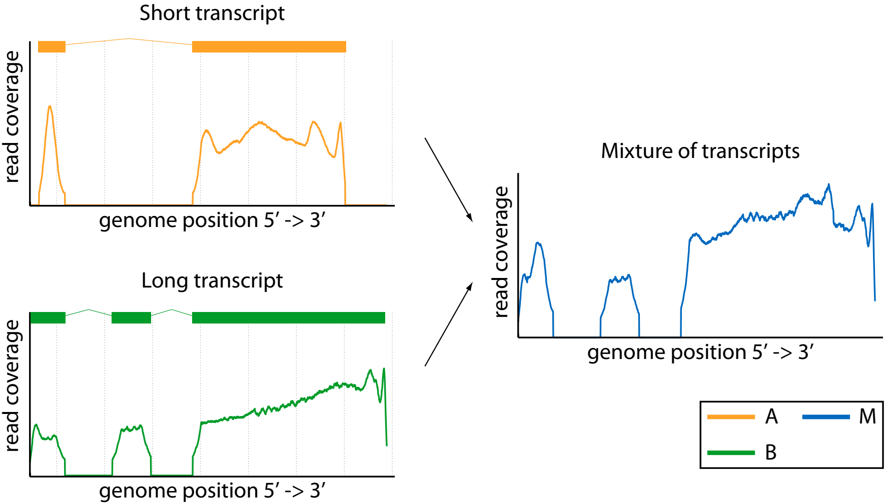
$$M_i = w_A A_i + w_B B_i \qquad \Rightarrow \qquad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea



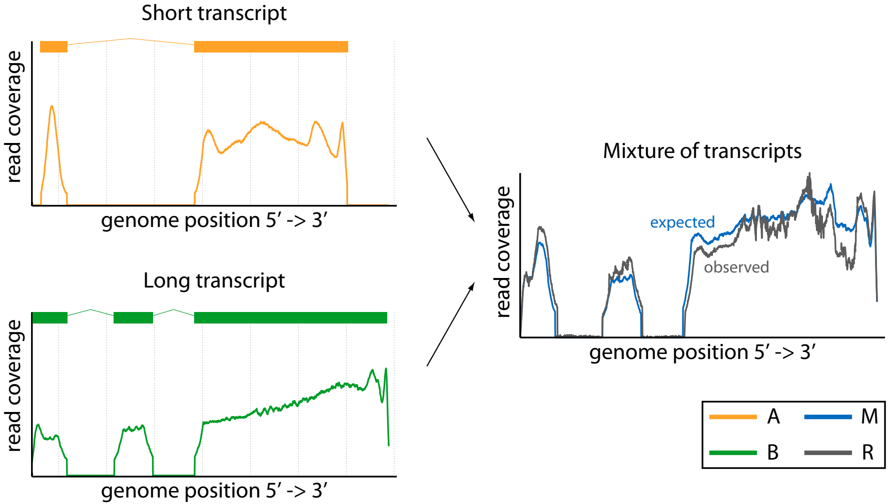
$$M_i = w_A A_i + w_B B_i \qquad \Rightarrow \qquad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

rQuant – Basic Idea



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

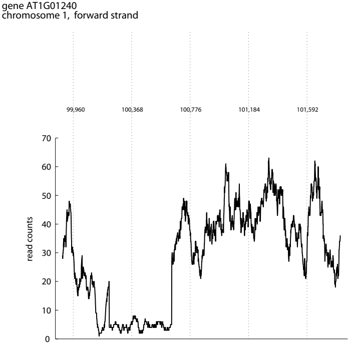
rQuant – Basic Idea



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

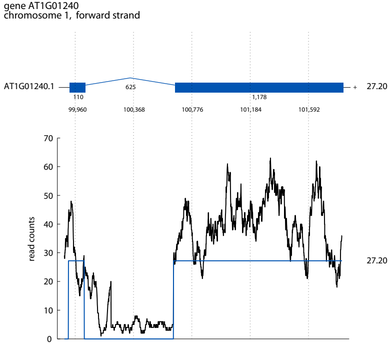
rQuant – Iterative Algorithm

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.



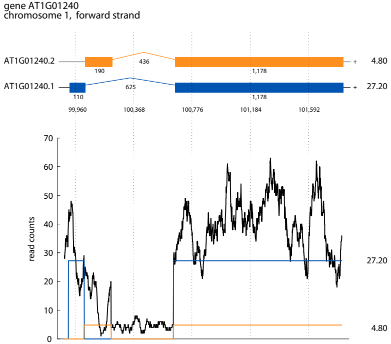
rQuant – Iterative Algorithm

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.



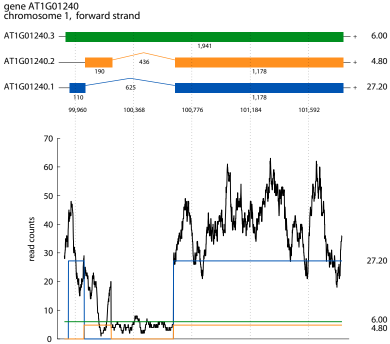
rQuant – Iterative Algorithm

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.



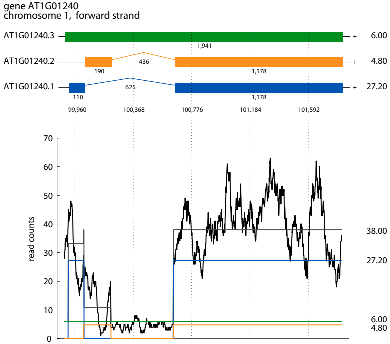
rQuant – Iterative Algorithm

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.



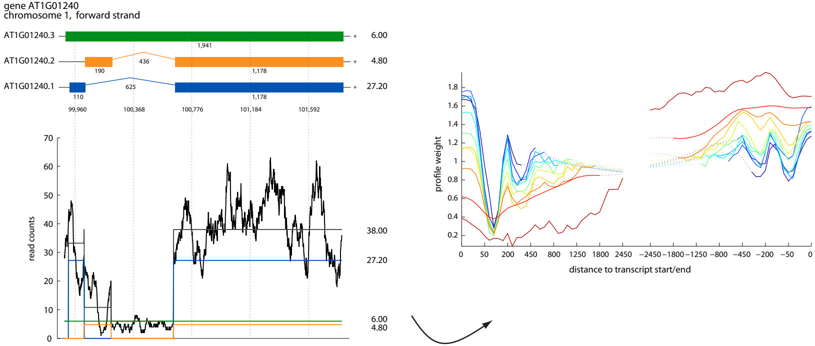
rQuant – Iterative Algorithm

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.



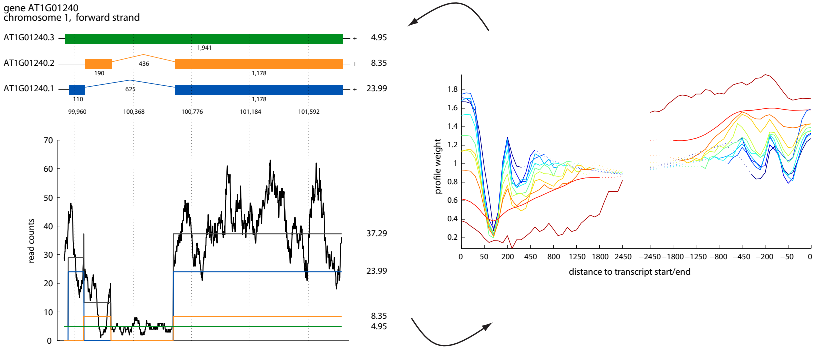
rQuant – Iterative Algorithm

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.



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.



rQuant Evaluation I

rQuant: Position-wise with profiles (estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g. Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

rQuant: Position-wise with profiles (estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g. Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

rQuant: Position-wise with profiles (estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g. Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

rQuant: Position-wise with profiles (estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g. Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

rQuant: Position-wise with profiles (estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g. Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation I

rQuant: Position-wise with profiles (estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009])
- Segment-wise, with profiles (e.g. Flux Capacitor [Sammeth, 2009a])

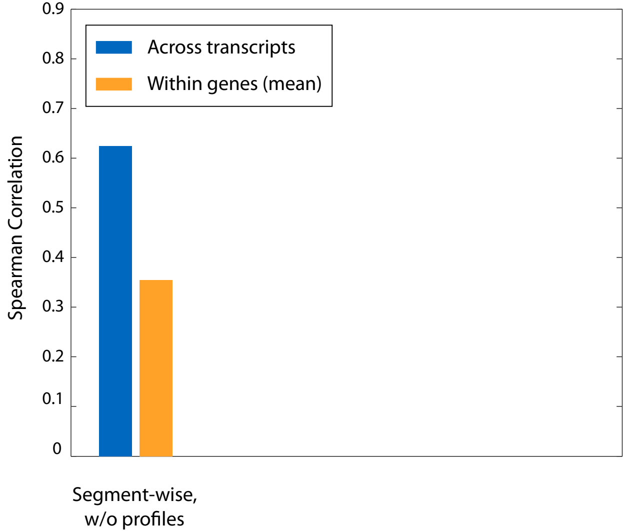
Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

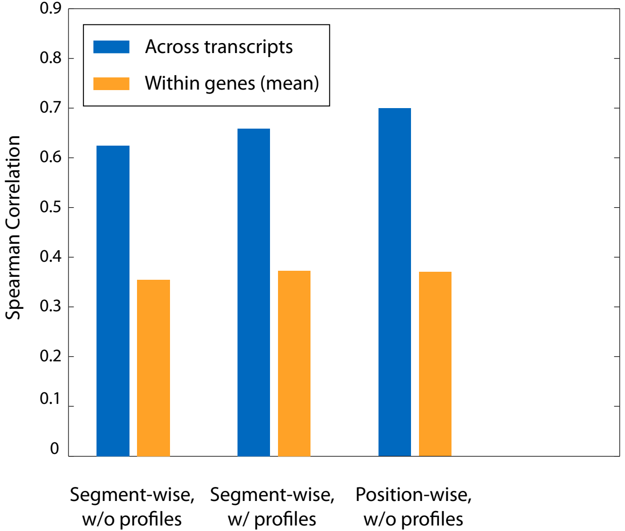
- Simulated RNA expression level and
- Predicted transcript weights

rQuant Evaluation II



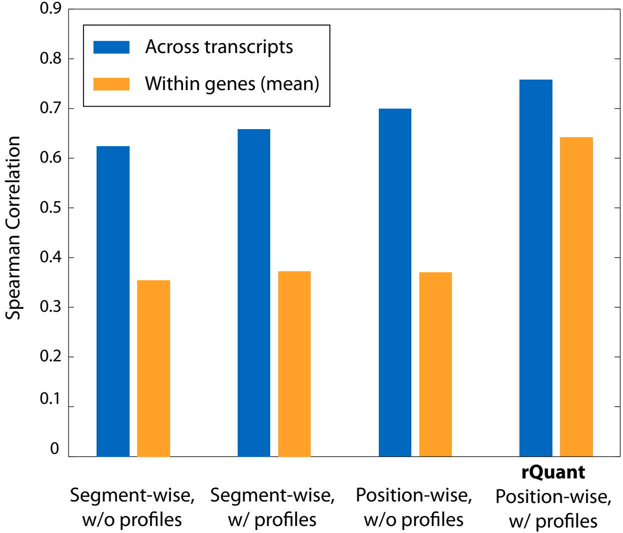
(Bohnert et al., submitted, 2010)

rQuant Evaluation II



(Bohnert et al., submitted, 2010)

rQuant Evaluation II

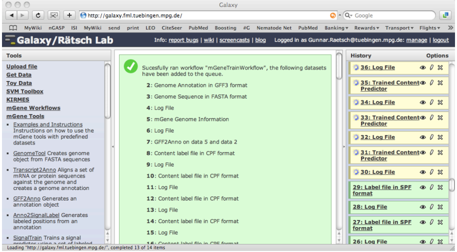


(Bohnert et al., submitted, 2010)

Galaxy-based Web Services for NGS Analyses

Galaxy-based web service http://galaxy.fml.mpg.de

- PALMapper http://fml.mpg.de/raetsch/suppl/palmapper
- mGene http://mgene.org/web
- mTIM http://fml.mpg.de/raetsch/suppl/mtim (in prep.)
- rQuant http://fml.mpg.de/raetsch/suppl/rquant/web


Galaxy
http://galaxy.fml.tuebingen.mpg.de/
Google
MyWiki nGASP ISI MyWiki send print LEO CiteSeer PubMed Boosting #G Nematode Net PubMed Banking Rewards Transport Flights
Galaxy/Rätsch Lab
Info: report bugs | wiki | screencasts | blog
Logged in as Gunnar.Raetsch@tuebingen.mpg.de: manage | logout
Tools
History
Options

Upload file
Get Data
Toy Data
SVM Toolbox
KIRMES
mGene Workflows
mGene Tools

- Examples and Instructions Instructions on how to use the mGene tools with predefined datasets
- GenomeTool Creates genome object from FASTA sequences
- Transcript2Anno Aligns a set of mRNA or protein sequences against the genome and creates a genome annotation
- GFF2Anno Generates an annotation object
- Anno2SignalLabel Generates labeled positions from an annotation
- SignalTrain Trains a signal predictor using a set of labeled

✓ Sucesfully ran workflow "mGeneTrainWorkflow", the following datasets have been added to the queue.

2: Genome Annotation in GFF3 format
3: Genome Sequence in FASTA format
4: Log File
5: mGene Genome Information
6: Log File
7: GFF2Anno on data 5 and data 2
8: Content label file in CPF format
9: Log File
10: Content label file in CPF format
11: Log File
12: Content label file in CPF format
13: Log File
14: Content label file in CPF format
15: Log File
16: Content label file in CPF format

36: Log File
35: Trained Content Predictor
34: Log File
33: Trained Content Predictor
32: Log File
31: Trained Content Predictor
30: Log File
29: Label file in SPF format
28: Log File
27: Label file in SPF format
26: Log File

Loading "http://galaxy.fml.tuebingen.mpg.de/", completed 13 of 14 items

(Rätsch et al., in preparation, 2010)

Summary

- PALMapper
  - Splice site predictions improve alignment performance
  - Outperforms many other read mappers in intron accuracy
- mTiM
  - High specificity, sensitivity depends on read coverage
  - Better for identifying transcripts specific to experimental data
- mGene
  - High sensitivity (also for lowly expressed genes)
  - Identifies also non-expressed genes ⇒ good for annotation
- rQuant
  - Models library prep., sequencing, alignment biases
  - Accurately quantifies transcripts
- Galaxy instance
  - Easy use of these tools

Summary

- PALMapper
  - Splice site predictions improve alignment performance
  - Outperforms many other read mappers in intron accuracy
- mTiM
  - High specificity, sensitivity depends on read coverage
  - Better for identifying transcripts specific to experimental data
- mGene
  - High sensitivity (also for lowly expressed genes)
  - Identifies also non-expressed genes ⇒ good for annotation
- rQuant
  - Models library prep., sequencing, alignment biases
  - Accurately quantifies transcripts
- Galaxy instance
  - Easy use of these tools

Summary

- PALMapper
  - Splice site predictions improve alignment performance
  - Outperforms many other read mappers in intron accuracy
- mTiM
  - High specificity, sensitivity depends on read coverage
  - Better for identifying transcripts specific to experimental data
- mGene
  - High sensitivity (also for lowly expressed genes)
  - Identifies also non-expressed genes ⇒ good for annotation
- rQuant
  - Models library prep., sequencing, alignment biases
  - Accurately quantifies transcripts
- Galaxy instance
  - Easy use of these tools

Summary

- PALMapper
  - Splice site predictions improve alignment performance
  - Outperforms many other read mappers in intron accuracy
- mTiM
  - High specificity, sensitivity depends on read coverage
  - Better for identifying transcripts specific to experimental data
- mGene
  - High sensitivity (also for lowly expressed genes)
  - Identifies also non-expressed genes ⇒ good for annotation
- rQuant
  - Models library prep., sequencing, alignment biases
  - Accurately quantifies transcripts
- Galaxy instance
  - Easy use of these tools

Summary

- PALMapper
  - Splice site predictions improve alignment performance
  - Outperforms many other read mappers in intron accuracy
- mTiM
  - High specificity, sensitivity depends on read coverage
  - Better for identifying transcripts specific to experimental data
- mGene
  - High sensitivity (also for lowly expressed genes)
  - Identifies also non-expressed genes ⇒ good for annotation
- rQuant
  - Models library prep., sequencing, alignment biases
  - Accurately quantifies transcripts
- Galaxy instance
  - Easy use of these tools

Acknowledgements

Fabio De Bona

Alignments

Jonas Behr

Gene finding

Georg Zeller

Segmentation

Regina Bohnert

Quantitation

RGASP Team

- Jonas Behr (FML)
- Georg Zeller (FML & MPI)
- Regina Bohnert (FML)

Funding by DFG & Max Planck Society.

Thank you for your attention!

Acknowledgements

Fabio De Bona

Alignments

Jonas Behr

Gene finding

Georg Zeller

Segmentation

Regina Bohnert

Quantitation

RGASP Team

- Jonas Behr (FML)
- Georg Zeller (FML & MPI)
- Regina Bohnert (FML)

Funding by DFG &
Max Planck Society.

Thank you for your attention!

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

RM Clark, G Schweikert, C Toomajian, S Ossowski, G Zeller, P Shinn, N Warthmann, TT Hu, G Fu, DA Hinds, H Chen, KA Frazer, DH Huson, B Schölkopf, M Nordborg, G Rätsch, JR Ecker, and D Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–342, 2007. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1138632.

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

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, April 2009.

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.

References II

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. 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`.

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.

Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biol*, 10(9):R98, Jan 2009. doi: 10.1186/gb-2009-10-9-r98. URL `http://genomebiology.com/2009/10/9/R98`.

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*, 2009. 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.

References III

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. Henz, S. Laubinger, D. Weigel, and G. Rätsch. Transcript normalization and segmentation of tiling array data. In *Proc. PSB 2008*. World Scientific, 2008b.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *BioInformatics*, 16(9): 799–807, September 2000.