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)
Discovery of the Nuclein
(Friedrich Miescher, 1869)

Tübingen, around 1869

“...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 Nuclein:
- from lymphocyte & salmon
- “multi-basic acid” ($\geq 4$)
Motivation

Discovery of the Nuclein
(Friedrich Miescher, 1869)

Tübingen, around 1869

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

Example:

Two different classes of observations
**Motivation**

**Machine Learning**
Learning from empirical observations

**Given:** Observations of some complex phenomenon

**Goal:** Learn from data & build predictive models

**Example:**

Inferred classification rule
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
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.

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.

Goal: Obtain complete transcriptome for further analyses

Figure adapted from Wikipedia
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.

Goal: Obtain complete transcriptome for further analyses

Figure adapted from Wikipedia
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.

Goal: Obtain complete transcriptome for further analyses

Figure adapted from Wikipedia
Common RNA-Seq Analysis Steps

1. RNA-Seq Reads
2. Read Alignment
3. Expression level $\sim$ #reads
4. Significance Testing
5. Differentially Processed Transcripts/Segments
Common RNA-Seq Analysis Steps

1. RNA-Seq Reads
2. Read Alignment
3. Expression level (~ #reads)
4. Significance Testing
5. Transcript Quantitation
6. De novo Assembly
7. TARs and Transcripts
8. Differentially Processed Transcripts/Segments
RNA-Seq Pipeline

Common RNA-Seq Analysis Steps

RNA-Seq Reads → De novo Assembly

RNA-Seq Reads → Read Alignment

Read Alignment → Transcript Quantitation

Read Alignment → Expression level ~ #reads

Expression level ~ #reads → Significance Testing

Significance Testing → Transcript Reconstruction

Transcript Reconstruction → TARs and Transcripts

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Differentially Processed Transcripts/Segments

De novo Assembly → TARs and Transcripts

TARs and Transcripts → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruction

Transcript Reconstruction → Transcript Reconstruc
RNA-Seq Pipeline Overview

<table>
<thead>
<tr>
<th>Read alignment</th>
<th>Transcript finding</th>
<th>Quantitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Reads</td>
<td>PALMapper</td>
<td>mTim Segmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mGene.ngs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rQuant</td>
</tr>
<tr>
<td></td>
<td>Alignments</td>
<td>Transcripts w/ Quantitation</td>
</tr>
</tbody>
</table>
## RNA-Seq Pipeline Overview

<table>
<thead>
<tr>
<th>Read alignment</th>
<th>Transcript finding</th>
<th>Quantitation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short Reads</strong></td>
<td><strong>mGene.ngs</strong></td>
<td><strong>rQuant</strong></td>
</tr>
<tr>
<td><strong>PALMapper</strong></td>
<td><strong>mTim Segmentation</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Transcripts w/ Quantitation</strong></td>
</tr>
</tbody>
</table>

**PALMapper**

**mGene.ngs**

**mTim Segmentation**

**rQuant**

**Transcripts w/ Quantitation**
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
GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project
- $k$-mer based index, well suited for smaller genomes

```
DNA       ACCGTCGCGCGCCT...TCGGCG...AGAACGCT

matching k-mers
```

```
TCGCGCGCAACG
TCGC
CGCG
GCGC
CGCG
GCGC
CGCA
GCAA
CAAC
AACG
```

Matching k-mers

```
DNA       [ACCGTCGCGCGCCT...TCGGCG...AGAACGCT]

Read
TCGCGCGCAACG
TCGC
CGCG
GCGC
CGCG
GCGC
CGCA
GCAA
CAAC
AACG
```

[Schneeberger et al., 2009]
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
- *k*-mer based index, well suited for smaller genomes

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

© Gunnar Rätsch (FML, Tübingen)
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
- \( k \)-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:
- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA

More info: [http://fml.mpg.de/raetsch/suppl/palmapper](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?

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

PALMapper (3.5h) and TopHat (3.5h/10h) aligning 24M reads
QPALMA: Extended Smith-Waterman Scoring

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

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

<table>
<thead>
<tr>
<th></th>
<th>gap</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>gap</td>
<td>0.33</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.3</td>
<td>0.55</td>
</tr>
<tr>
<td>A</td>
<td>0.31</td>
<td>0.12</td>
<td>0.12</td>
<td>0.3</td>
<td>0.55</td>
<td>0.33</td>
</tr>
<tr>
<td>C</td>
<td>0.44</td>
<td>0.12</td>
<td>0.44</td>
<td>0.3</td>
<td>0.59</td>
<td>0.12</td>
</tr>
<tr>
<td>G</td>
<td>0.13</td>
<td>0.85</td>
<td>0.31</td>
<td>0.33</td>
<td>0.51</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>0.55</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
<td>0.11</td>
<td>0.1</td>
</tr>
<tr>
<td>N</td>
<td>0.12</td>
<td>0.01</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.01</td>
</tr>
</tbody>
</table>

...ACGTACACGGTAGCT...CCGTAGAATTGACTGTGTGG...

...GCCTACACCG________________________AATTGGA
QPALMA: Extended Smith-Waterman Scoring

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

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

<table>
<thead>
<tr>
<th></th>
<th>gap</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>gap</td>
<td>0.33</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.3</td>
<td>0.55</td>
</tr>
<tr>
<td>A</td>
<td>0.31</td>
<td>0.12</td>
<td>0.12</td>
<td>0.3</td>
<td>0.55</td>
<td>0.33</td>
</tr>
<tr>
<td>C</td>
<td>0.44</td>
<td>0.12</td>
<td>0.44</td>
<td>0.3</td>
<td>0.59</td>
<td>0.12</td>
</tr>
<tr>
<td>G</td>
<td>0.13</td>
<td>0.85</td>
<td>0.31</td>
<td>0.33</td>
<td>0.51</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>0.55</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
<td>0.11</td>
<td>0.1</td>
</tr>
<tr>
<td>N</td>
<td>0.12</td>
<td>0.01</td>
<td>0.3</td>
<td>0.12</td>
<td>0.3</td>
<td>0.01</td>
</tr>
</tbody>
</table>
QPALMA: Extended Smith-Waterman Scoring

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

Source of information
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information
QPALMA: Extended Smith-Waterman Scoring

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

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

[De Bona et al., 2008]
What are optimal parameters?
How do we jointly optimize the 336 parameters?
What are optimal parameters?

How do we jointly optimize the 336 parameters?
Correct alignment is **not** highest scoring one

Correct alignment is highest scoring one

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

Correct alignment is highest scoring one

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

Alignment Error Rate
- SmithW: 14.19%
- Intron: 9.96%
- Intron+: 1.94%
- Intron+ Splice: 1.78%

[De Bona et al., 2008]
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

Alignment Error Rate

<table>
<thead>
<tr>
<th>8nt</th>
<th>12nt</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmithW</td>
<td>14.19%</td>
</tr>
<tr>
<td>Intron</td>
<td>9.96%</td>
</tr>
<tr>
<td>Intron+Splice</td>
<td>1.94%</td>
</tr>
<tr>
<td>Intron+Splice+Quality</td>
<td>1.78%</td>
</tr>
</tbody>
</table>

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**)
Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*
**Goal:** Characterize each base as *intergenic*, *exonic*, or *intrinsic*
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic* annotated gene
Learn to associate a state with each position given its read coverage and local context

- HM-SVM training: Optimize transformations: signal $\rightarrow$ score
- Extension: Score spliced reads and splice sites

(G. Zeller et al., 2008; G. Zeller et al., in prep., 2009)
Learn to associate a state with each position given its read coverage and local context

HM-SVM training: Optimize transformations: signal $\rightarrow$ score

Extension: Score spliced reads and splice sites

(G. Zeller et al., 2008; G. Zeller et al., in prep., 2009)
Learn to associate a state with each position given its read coverage and local context

- HM-SVM training: Optimize transformations: signal $\rightarrow$ score
- Extension: Score spliced reads and splice sites

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

\( f : \mathbb{R}^* \rightarrow \Sigma^* \)

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

Discriminant function \( F_\theta : \mathbb{R}^* \times \Sigma^* \rightarrow \mathbb{R} \) such that for

decoding: \( f(\chi) = \arg \max_{\sigma \in \Sigma^*} F_\theta(\chi, \sigma) \).

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

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

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

A quadratic programming problem (QP) is solved to optimize \( \theta \).

[Altun et al., 2003, Rätsch et al., 2007, Zeller et al., 2008b]
Discriminative training of HM-SVMs

\( f : \mathbb{R}^* \rightarrow \Sigma^* \)

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

Discriminant function \( F_\theta : \mathbb{R}^* \times \Sigma^* \rightarrow \mathbb{R} \) such that for decoding:

\( f(\chi) = \arg\max_{\sigma \in S^*} F_\theta(\chi, \sigma). \)

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

\[
F_\theta(\chi^{(i)}, \sigma^{(i)}) - F_\theta(\chi^{(i)}, \sigma) \geq \rho
\]

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

A quadratic programming problem (QP) is solved to optimize \( \theta \).

[Altun et al., 2003, Rätsch et al., 2007, Zeller et al., 2008b]
Discriminative training of HM-SVMs

\[ f : \mathbb{R}^* \rightarrow \Sigma^* \]

given a sequence of hybridization measurements \( \chi \in \mathbb{R}^* \)

predicts a state sequence (path) \( \sigma \in \Sigma^* \)

Discriminant function \( F_\theta : \mathbb{R}^* \times \Sigma^* \rightarrow \mathbb{R} \) such that for

decoding: \( f(\chi) = \arg \max_{\sigma \in \Sigma^*} F_\theta(\chi, \sigma) \).

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

\[ F_\theta(\chi^{(i)}, \sigma^{(i)}) - F_\theta(\chi^{(i)}, \bar{\sigma}) \geq \rho \]

between the correct path \( \sigma^{(i)} \) and any other wrong path \( \bar{\sigma} \neq \sigma^{(i)} \).

A quadratic programming problem (QP) is solved to optimize \( \theta \).

[Altun et al., 2003, Rätsch et al., 2007, Zeller et al., 2008b]
Sensitivity heavily depends on read density.
Preliminary Evaluation \((C. \text{ elegans})\)

Sensitivity heavily depends on read density

CDS (precision+recall)/2

Sensitivity heavily depends on read density
Next Generation Gene Finding

Computational Gene Finding
⇝ Labeling the Genome

DNA
- TSS
- polyA/cleavage

Pre-mRNA
- Splice Donor
- Splice Acceptor

mRNA
- TIS
- Stop
- cap
- polyA

Protein
Computational Gene Finding

Labeling the Genome

DNA

TSS

Donor

Acceptor

TIS

Stop

polyA/cleavage

pre-mRNA

mRNA

cap

polyA

Protein

© Gunnar Rätsch (FML, Tübingen)
Computational Gene Finding

Labeling the Genome

DNA

TSS

Donor

Acceptor

Donor

Acceptor

polyA/cleavage

pre-mRNA

TIS

mRNA

cap

polyA

Protein

TSS

TIS

Stop

cleave

Don

Acc

Methods for Transcriptome Analysis

NGS Bioinformatics, Paris
mGene-based Transcript Prediction

STEP 1: SVM Signal Predictions

tss
tis
acc
don
stop
mGene-based Transcript Prediction

STEP 1: SVM Signal Predictions
- tss
- tis
- acc
- don
- stop

STEP 2: Integration
- F(x,y)

transform features
mGene-based Transcript Prediction

STEP 1: SVM Signal Predictions
- tss
- tis
- acc
- don
- stop

STEP 2: Integration
- transform features
- large margin

True gene model
Wrong gene model

genomic position

True gene model
Wrong gene model

large margin
F(x,y)
transform features

© Gunnar Rätsch (FML, Tübingen)
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

<table>
<thead>
<tr>
<th></th>
<th>Exon Level</th>
<th></th>
<th>Transcript Level</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN</td>
<td>SP</td>
<td>F</td>
<td>SN</td>
</tr>
<tr>
<td><strong>ab initio</strong></td>
<td>82.3</td>
<td>82.6</td>
<td>82.5</td>
<td>43.1</td>
</tr>
<tr>
<td>ESTs heuristic</td>
<td>85.3</td>
<td>84.7</td>
<td>85.0</td>
<td>49.5</td>
</tr>
<tr>
<td>ESTs trained</td>
<td>84.8</td>
<td>85.8</td>
<td>85.3</td>
<td>50.5</td>
</tr>
</tbody>
</table>

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

Behr et al., in pre., 2010
mGene-based Transcript Prediction

1. SVM Signal Predictions
   - tss
   - tis
   - acc
   - don
   - stop

2. Integration
   - transform features
   - large margin

Genomic position
mGene-based Transcript Prediction

STEP 1: SVM Signal Predictions
- tss
- tis
- acc
- don
- stop

STEP 2: Integration

Expression Evidence

transform SVM outputs with PLiF

large margin
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

<table>
<thead>
<tr>
<th></th>
<th>Exon Level</th>
<th></th>
<th>Transcript Level</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN</td>
<td>SP</td>
<td>F</td>
<td>SN</td>
</tr>
<tr>
<td>ab initio</td>
<td>82.3</td>
<td>82.6</td>
<td>82.5</td>
<td>43.1</td>
</tr>
<tr>
<td>ESTs heuristic</td>
<td>85.3</td>
<td>84.7</td>
<td>85.0</td>
<td>49.5</td>
</tr>
<tr>
<td>ESTs trained</td>
<td>84.8</td>
<td>85.8</td>
<td>85.3</td>
<td>50.5</td>
</tr>
<tr>
<td>RNA-Seq trained</td>
<td>84.6</td>
<td>84.9</td>
<td>84.8</td>
<td>49.1</td>
</tr>
<tr>
<td>RNA-Seq/ESTs trained</td>
<td>84.7</td>
<td><strong>86.9</strong></td>
<td>85.8</td>
<td>50.3</td>
</tr>
</tbody>
</table>

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

Behr et al., in prep., 2010
### Preliminary Evaluation (C. elegans)

Bar chart showing CDS (precision+recall)/2 expression percentiles for:
- mGene \textit{ab initio}
- mGene.ngs

Expression percentiles range from 10 to 100%.
Preliminary Evaluation (*C. elegans*)

CDS (precision+recall)/2

- mTiM
- mGene *ab initio*
- mGene.ngs

expression percentiles [%]

© Gunnar Rätsch (FML, Tübingen)
• **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*
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
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
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

<table>
<thead>
<tr>
<th>Read alignment</th>
<th>Transcript finding</th>
<th>Quantitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PALMapper</td>
<td>mTim Segmentation</td>
<td>rQuant</td>
</tr>
<tr>
<td>Short Reads</td>
<td>mGene.ngs</td>
<td>Transcripts w/ Quantitation</td>
</tr>
</tbody>
</table>

Alignments

© Gunnar Rätsch (FML, Tübingen) Methods for Transcriptome Analysis NGS Bioinformatics, Paris
RNA-Seq Biases and Quantitation

Biases due to . . .
- cDNA library construction
- Sequencing
- Read mapping

Biases in read coverage:
- Relative transcript position 5' -> 3'
- 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 $\approx 1$kb for the *C. elegans* SRX001872 dataset)
rQuant – Basic Idea

Short transcript

\[ M_i = w_A A_i + w_B B_i \Rightarrow \min_{w_A, w_B} \sum_i \ell (M_i, R_i) \]
**rQuant – Basic Idea**

### Short transcript

<table>
<thead>
<tr>
<th>Relative transcript position</th>
<th>Read coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' -&gt; 3'</td>
<td>🌋</td>
</tr>
</tbody>
</table>

### Long transcript

<table>
<thead>
<tr>
<th>Relative transcript position</th>
<th>Read coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' -&gt; 3'</td>
<td>🌋</td>
</tr>
</tbody>
</table>

\[
M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell (M_i, R_i)
\]
rQuant – Basic Idea

\[ M_i = w_A A_i + w_B B_i \Rightarrow \min_{w_A, w_B} \sum_i \ell (M_i, R_i) \]
**rQuant – Basic Idea**

- **Short transcript**
- **Long transcript**

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

© Gunnar Rätsch (FML, Tübingen)
rQuant – Basic Idea

Short transcript

Long transcript

Mixture of transcripts

\[ M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell (M_i, R_i) \]
rQuant – Iterative Algorithm

1. Optimise transcript weights: \( \min_w \sum_i \ell \left( \sum_t w(t)p^{(t)}_i, R_i \right) \)

2. Optimise profile weights: \( \min_p \sum_i \ell \left( \sum_t w(t)p^{(t)}_i, R_i \right) \)

3. Repeat 1. and 2. until convergence.

gene AT1G01240
chromosome 1, forward strand

<table>
<thead>
<tr>
<th>Read Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>99,960</td>
</tr>
</tbody>
</table>

© Gunnar Rätsch (FML, Tübingen)  NGS Bioinformatics, Paris
rQuant – Iterative Algorithm

1. Optimise transcript weights: \( \min_w \sum_i \ell \left( \sum_t w^{(t)} p_i^{(t)}, R_i \right) \)

2. Optimise profile weights: \( \min_p \sum_i \ell \left( \sum_t w^{(t)} p_i^{(t)}, R_i \right) \)

3. Repeat 1. and 2. until convergence.
rQuant – Iterative Algorithm

1. Optimise transcript weights: \( \min_w \sum_i \ell \left( \sum_t w(t) p_i(t), R_i \right) \)

2. Optimise profile weights: \( \min_p \sum_i \ell \left( \sum_t w(t) p_i(t), R_i \right) \)

3. Repeat 1. and 2. until convergence.
rQuant – Iterative Algorithm

1. Optimise transcript weights: \( \min_w \sum_i \ell \left( \sum_t w(t)p_i^{(t)}, R_i \right) \)

2. Optimise profile weights: \( \min_p \sum_i \ell \left( \sum_t w(t)p_i^{(t)}, R_i \right) \)

3. Repeat 1. and 2. until convergence.
rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_w \sum_i \ell \left( \sum_t w^{(t)} p_i^{(t)}, R_i \right)$

2. Optimise profile weights: $\min_p \sum_i \ell \left( \sum_t w^{(t)} p_i^{(t)}, R_i \right)$

3. Repeat 1. and 2. until convergence.
rQuant – Iterative Algorithm

1. Optimise transcript weights: \( \min_w \sum_i \ell \left( \sum_t w(t) p_i(t), R_i \right) \)

2. Optimise profile weights: \( \min_p \sum_i \ell \left( \sum_t w(t) p_i(t), R_i \right) \)

3. Repeat 1. and 2. until convergence.
rQuant – Iterative Algorithm

1. Optimise transcript weights: \( \min_w \sum_i \ell \left( \sum_t w^{(t)} p_i^{(t)}, R_i \right) \)

2. Optimise profile weights: \( \min_p \sum_i \ell \left( \sum_t w^{(t)} p_i^{(t)}, R_i \right) \)

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: 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)

- 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 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

Spearman Correlation
Across transcripts
Within genes (mean)

Segment-wise, w/o profiles

(Bohnert et al., submitted, 2010)
rQuant Evaluation II

Spearman Correlation

- Across transcripts
- Within genes (mean)

Segment-wise, w/o profiles
Segment-wise, w/ profiles
Position-wise, w/o profiles

(Bohnert et al., submitted, 2010)
rQuant Evaluation II

![Graph showing Spearman Correlation](graph.png)

- Across transcripts
- Within genes (mean)

Segment-wise, w/o profiles
Segment-wise, w/ profiles
Position-wise, w/o profiles
Position-wise, w/ profiles

(Bohnert et al., submitted, 2010)
Galaxy-based web service: http://galaxy.fml.mpg.de

- PALMMapper: 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

(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
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 $\Rightarrow$ 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 $\Rightarrow$ 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!


