Online Quantitative Transcriptome Analysis

Gunnar Rätsch

Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

July 9, 2010

Hitseq 2010, Boston

http://galaxy.fml.mpg.de
Online Quantitative Transcriptome Analysis (Oqtans)

Gunnar Rätsch

Friedrich Miescher Laboratory of the Max Planck Society

Tübingen, Germany

July 9, 2010

Hitseq 2010, Boston

http://galaxy.fml.mpg.de
Online Quantitative Transcriptome Analysis (Oqtans)

Gunnar Rätsch

Friedrich Miescher Laboratory of the Max Planck Society
Tübingen, Germany

July 9, 2010
Hitseq 2010, Boston

http://galaxy.fml.mpg.de
# RNA-seq Analysis Pipeline(s)

<table>
<thead>
<tr>
<th>Read alignment</th>
<th>Transcript definition</th>
<th>Quantitative analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignments</td>
<td>Transcripts</td>
<td></td>
</tr>
</tbody>
</table>
RNA-seq Analysis Pipeline(s)

Read alignment | Transcript definition | Quantitative analysis

- Short reads
- Read Mapping (PALMapper)
- Alignments
- Transcripts
- Quantitative analysis

ShortSIG 2008: Palmapper (≈ QPalma) (De Bona et al., 2008)
RNA-seq Analysis Pipeline(s)

Read alignment  |  Transcript definition  |  Quantitative analysis

- Short reads
- Read Mapping (PALMapper)
- Annotation
- Transcript quantitation (rQuant)
- Quantified Transcriptome

ShortSIG 2008: Palmapper (∼ QPalma)  
ShortSIG 2009: rQuant  

(De Bona et al., 2008)  
(Bohnert and Rätsch, 2010)
RNA-seq Analysis Pipeline(s)

ShortSIG 2008: Palmapper (~ QPalma) (De Bona et al., 2008)
ShortSIG 2009: rQuant (Bohnert and Rätsch, 2010)
Hitseq 2010: rDiff (Stegle et al., 2010)

Gunnar Rätsch  (FML, Tübingen)
Introduction

RNA-seq Analysis Pipeline(s)

<table>
<thead>
<tr>
<th>Read alignment</th>
<th>Transcript definition</th>
<th>Quantitative analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short reads</td>
<td>Read Mapping (PALMapper)</td>
<td>Annotation</td>
</tr>
<tr>
<td>Read Mapping (PALMapper)</td>
<td>Gene Finding (mGene.ngs)</td>
<td>Differential express. testing (rDiff)</td>
</tr>
<tr>
<td>Short SIG 2008: Palmapper (≈ QPalma)</td>
<td>Short SIG 2009: rQuant</td>
<td>Transcript quantitation (rQuant)</td>
</tr>
<tr>
<td>Short SIG 2009: rQuant</td>
<td>Hitseq 2010: rDiff</td>
<td>Quantified Transcriptome</td>
</tr>
</tbody>
</table>

(De Bona et al., 2008)

(Bohnert and Rätsch, 2010)

(Stegle et al., 2010)
RNA-seq Analysis Pipeline(s)

**Introduction**

**Read alignment**
- Short reads
  - Read Mapping (PALMapper)

**Transcript definition**
- Annotation
  - Transcript prediction (mTim)
- Gene Finding (mGene.ngs)

**Quantitative analysis**
- Differential express. testing (rDiff)
- Transcript quantitation (rQuant)
- Transcripts
- Quantified Transcriptome

**Annotation**

**Software and Tools**

- ShortSIG 2008: Palmapper (∼ QPalma) (De Bona et al., 2008)
- ShortSIG 2009: rQuant (Bohnert and Rätsch, 2010)
- Hitseq 2010: rDiff (Stegle et al., 2010)
Galaxy is great!

Galaxy/Rätsch Lab

This is a customized version of the Galaxy framework, extended with machine learning based tools for sequence and tiling array data analysis. It provides tools developed by members of the Machine Learning in Biology (MLiB) Group at the Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany. For problems with any of the non-standard tools, please contact the MLiB Galaxy Support Team.

Currently we provide:

- mGene for computational gene finding,
- PALMapper for accurate spliced alignment of RNA-seq reads,
- rQuant for transcriptome quantitation from RNA-seq experiments,
- KIRMES for promoter analysis, and
- SVM Toolbox for classification of sequences and vectorial data.

Learn more about our Galaxy and the lab

- Galaxy: Galaxy is a beautiful framework developed by the Galaxy team at Penn State University. It is easy to use and convenient for tool developers as well.
- Shogun: Most of our tools are based on the machine learning toolbox Shogun for solving large-scale classification problems.
- Projects: Find a list of other projects we are working on here.

News

- March 22, 2010: Mosek Agreement We settled an agreement with MOSEK that allows us to offer all our MOSEK-based tools (mGene, PALMapper, rQuant, etc) also to non-academic users via Galaxy.
- February 21, 2010: Maintenance This server will be down for maintenance on Monday, February 22, 2010 between 9-12am CEST.
- February 19, 2010: rQuant We added a new version of the rQuant tools for transcriptome quantitation from short read data (see NGS: Quantitation Tools).

This work is funded by...

(Taylor et al., 2007; Blankenberg et al., 2010a)
Galaxy is great!

**Tools**

Galaxy/Rätsch Lab

This is a customized version of the Galaxy framework, extended with machine learning based tools for sequence and tiling array data analysis. It provides tools developed by members of the Machine Learning in Biology (MLB) Group at the Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany. For problems with any of the non-standard tools, please contact the MLB Galaxy Support Team.

Currently we provide:

- mGene for computational gene finding,
- PALMapper for accurate spliced alignment of RNA-seq reads,
- rQuant for transcriptome quantification from RNA-seq experiments,
- KIRMES for promoter analysis, and
- SVM Toolbox for classification of sequences and vectorial data.

Learn more about our Galaxy and the Lab:

- **Galaxy**: Galaxy is a beautiful framework developed by the Galaxy team at Penn State University. It is easy to use and convenient for tool developers as well.
- **Shogun**: Most of our tools are based on the machine learning toolbox Shogun for solving large-scale classification problems.
- **Projects**: Find a list of other projects we are working on [here](#).

**News**

- **March 22, 2010**: Moskex Agreement We settled an agreement with MOSEK that allows us to offer all our MOSEK-based tools (mGene, PALMapper, rQuant, etc.) also to non-academic users via Galaxy.
- **February 21, 2010**: Maintenance This server will be down for maintenance on Monday, February 22, 2010 between 9-12am CEST.
- **February 19, 2010**: rQuant We added a new version of the rQuant tools for transcriptome quantification from short read data (see NGS: Quantitation Tools).

This work is funded by

[Taylor et al., 2007; Blankenberg et al., 2010a](#)
Introduction

Galaxy is great!

(Taylor et al., 2007; Blankenberg et al., 2010a)
Galaxy is great!

Working area

(Taylor et al., 2007; Blankenberg et al., 2010a)
Example 1: rQuant Transcript Quantitation

Inputs:
- Genome sequence (fasta)
- Annotation (gff3)
- Alignments (sam)
Example 1: rQuant Transcript Quantitation

**Inputs:**
- Genome sequence (fasta)
- Annotation (gff3)
- Alignments (sam)
Example 1: rQuant Transcript Quantitation

Inputs:
- Genome sequence (fasta)
- Annotation (gff3)
- Alignments (sam)

Output: Transcripts with RPKM values
Example 2: rDiff Differential Transcript Expression

Inputs:
- Annotation (gff3)
- Alignments Sample 1 (sam)
- Alignments Sample 2 (sam)

Output: List of transcripts with p-values
Example 3: *De Novo* Genome Annotation

**Inputs:**
- Genome sequence (contigs)
- RNA-seq sequences

**Output:**
- Annotated genome

**NB:** No previous annotation needed

(Schweikert et al., 2009d; Behr et al., 2010)
Transcript Prediction vs. Read Coverage

Annotated genes sorted by expression level
Transcript Prediction vs. Read Coverage

Annotated genes sorted by expression level
Transcript Prediction vs. Read Coverage

Annotated genes sorted by expression level
Label generation from RNA-seq data

Input:
- RNA-seq read alignments

Tools:
- RNA-seq label generation

Output:
- gene structures (high expressed genes)
Label generation from RNA-seq data

Input
- RNA-seq read alignments
- Genomic DNA sequence

Tools
- RNA-seq label generation
- MGene.ngs training

Output
- Gene structures (high expressed genes)
- Trained gene predictor
Label generation from RNA-seq data

Input:
- RNA-seq read alignments
- Genomic DNA sequence

Tools:
- RNA-seq label generation
- mGene.ngs training
- mGene.ngs prediction

Output:
- Gene structures (high expressed genes)
- Trained gene predictor
- Gene structures
Transcript Prediction

Two approaches:

1. Coverage segmentation algorithm $mTIM$ for general transcripts (no coding bias/assumption)

2. Extension of $mGene$ gene finding system to use RNA-seq data for protein coding transcript prediction
Transcript Prediction

Two approaches:

1. Coverage segmentation algorithm \textit{mTIM} for general transcripts (no coding bias/assumption)

2. Extension of \textit{mGene} gene finding system to use RNA-seq data for protein coding transcript prediction

Inputs:

- Genome sequence
- Alignments in BAM format
- Preliminary annotation (training only)
Strategy of mGene.ngs

STEP 1: SVM Signal Predictions

True gene model

Genome Annotation
Strategy of mGene.ngs

STEP 1: SVM Signal Predictions

tss

tis

acc
don

stop

True gene model

Score

genomic position
Strategy of mGene.ngs

STEP 1: SVM Signal Predictions

tss

tis

acc
don

stop

RNA-seq coverage

Score

genomic position
Strategy of mGene.ngs

STEP 1: SVM Signal Predictions

tss

tis

acc
don

stop

RNA-seq coverage

intron support from spliced reads

Score
**De Novo Annotation Workflow**

- **Genome Sequence**
  - Input dataset
  - output

- **RNA-seq FASTQ**
  - Input dataset
  - output

- **SAM-to-BAM**
  - Convert SAM file
  - Using reference file
  - output1 (bam)

- **mGeneTrain**
  - Genomic sequence
  - Set of known genes for training in GFF3 format
  - Trained_mGene_Predictor (tmgp)
  - Log_file (txt)

- **mGenePredict**
  - Genomic sequence
  - gff file (gff3)
  - Log_file (txt)

- **Map with TopHat**
  - Series 1 > source file
  - series
  - junctions (bed)
  - coverage (wig)
  - expr_file (tabular)
  - accepted_hits (sam)
  - log_report (txt)
  - run_log (txt)

- **BAM2Anno**
  - Genome information file
  - BAM alignment file
  - Annotation (gann)
  - Log_file (txt)

- **Anno2GFF**
  - Genome Annotation in AGS format
  - gff_output (gff3)
  - Log_file (txt)

Next step: visualize using Galaxy's Trackster or a genome browser.
De Novo Annotation Workflow

Next step: visualize using Galaxy’s Trackster or a genome browser
Results for *C. elegans*

![Graph showing expression percentile vs. F-score for mGene.ngs - only sequence](graph.png)

Gunnar Rätsch (FML, Tübingen)  
Quantitative Transcriptome Analysis

*(Behr et al., 2010)*
Results for *C. elegans*

![Graph](image)

- **mGene.ngs – only sequence**
- **mGene.ngs – no subsampling**

(Behr et al., 2010)
Results for *C. elegans*

![Graph showing F-score expression percentile for different conditions: mGene.ngs only sequence, mGene.ngs no subsampling, mGene.ngs subsampling.](image)

*Behr et al., 2010*
Results for *C. elegans*

![Graph showing F-score results for different conditions.](image)

- mGene.ngs – only sequence
- mGene.ngs – no subsampling
- mGene.ngs – subsampling
- mGene.ngs – using annotation

*(Behr et al., 2010)*
Results for *C. elegans*

![Graph showing expression percentile vs. F-score for different methods: mGene.ngs – only sequence, mGene.ngs – no subsampling, mGene.ngs – subsampling, mGene.ngs – using annotation, cufflinks – Trapnell et al. 2010.](image-url)
Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, ...)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)
Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, ...)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)

- Genome annotation with RNA-seq works well also for newly sequenced genomes
Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, …)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)

- Genome annotation with RNA-seq works well also for newly sequenced genomes

- Limitations for some tools w.r.t. genome size (RAM)

- Works well for few users in parallel, but has limits (CPU, disks)
Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, ...)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)

- Genome annotation with RNA-seq works well also for newly sequenced genomes

- Limitations for some tools w.r.t. genome size (RAM)

- Works well for few users in parallel, but has limits (CPU, disks)

- Deployment on Amazon-Cloud using Galaxy’s new cloud features
Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, ...)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)

- Genome annotation with RNA-seq works well also for newly sequenced genomes

- Limitations for some tools w.r.t. genome size (RAM)

- Works well for few users in parallel, but has limits (CPU, disks)

- Deployment on Amazon-Cloud using Galaxy’s new cloud features

- Software packages and Galaxy bindings will be made available
Acknowledgements

Fabio De Bona
Alignments

Jonas Behr
Gene finding

Regina Bohnert
Quantitation

Philipp Drewe
Diff. Testing

Andre Kahles
Testing

Vipin Sreedharan
Bug fixes

Gunnar Rätsch (FML, Tübingen)
Quantitative Transcriptome Analysis

The End
Acknowledgements

Fabio De Bona
- Alignments

Jonas Behr
- Gene finding

Regina Bohnert
- Quantitation

Philipp Drewe
- Diff. Testing

Andre Kahles
- Testing

Vipin Sreedharan
- Bug fixes

Funding by DFG & Max Planck Society.

Thank you for your attention.

Poster J50, Monday