Oqtans: Online Quantitative Transcriptome Analysis

Regina Bohnert[1] Jonas Behr[1] Géraldine Jean[1] André Kahles[1] Georg Zeller[1,2] Gunnar Rätsch[1]

[1] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany

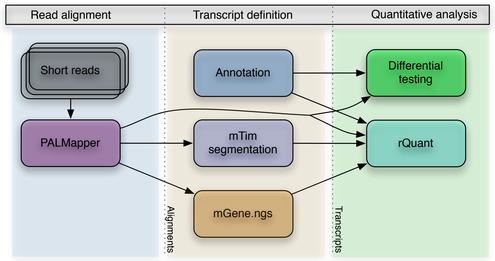
[2] Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany

Introduction & Motivation

Aspects of the Transcriptome Studied

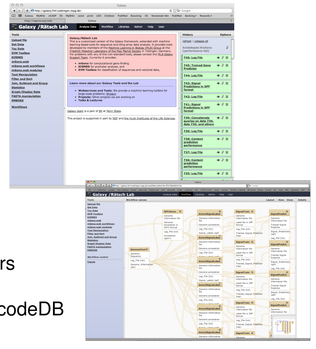
- Identification and quantification of alternative transcripts
- Discovery of new genes and transcripts
- Improve the accuracy of existing automatic annotation (methods)
- Web service available at:

`http://galaxy.fml.mpg.de/`



The Galaxy Framework

- **Galaxy: The framework for compute services [9]**
  Easy integration of command line tools
- **Exchange between users**
  Workflows can be exchanged among users and still can be modified and improved
- **Workflow editor**
  Graphical User Interface for combining tools to complex pipelines
- **Large number of bioinformatics tool**
  Including: EMBOSS, short reads tools, statistical tools, ...
- **Data import**
  Data can be uploaded by users or can be imported directly from UCSC, BioMart, and EncodeDB
- **NGS-Tools**
  Tools for manipulation and statistical examination of next generation sequencing data
- **Genome Size Data**
  Handles large data sets and distributes computations on computing cluster
- **Additional Packages from Tübingen:**
  - **KIRMES**
    Promoter analysis from ChIP-chip or ChIP-Seq data
  - **SVM Toolbox**
    Generic interface for classification of sequences and vectorial data with SVMs



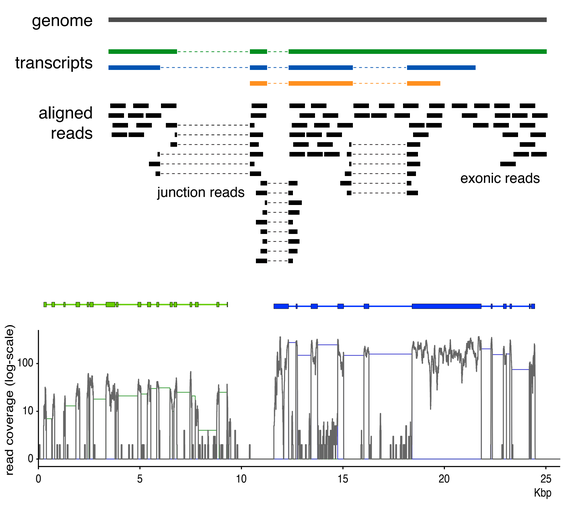
More information at

`http://www.fml.mpg.de/raetsch/suppl/oqtans`

Experimental Data

RNA Sequencing (RNA-Seq)

- Profiles transcripts in a **digital** manner
- Generate **RNA-Seq reads** that need to be mapped to the genome
- Exhibits various **biases** leading to distortions of the underlying transcript abundances

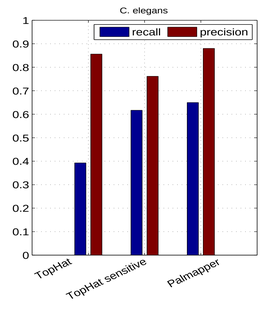


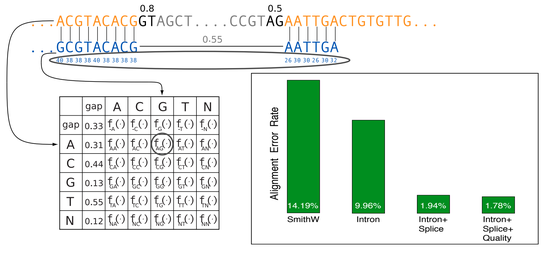
Mapping Short Reads with PALMapper

PALMapper is a combination of GenomeMapper [2] for fast read mapping and QPALMA [1] for accurate spliced alignment, incorporating

- read sequence and quality
- splice site information

during the alignment.





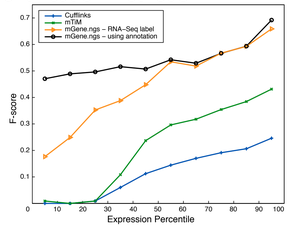
`http://fml.mpg.de/raetsch/suppl/palmapper`

Transcript Identification

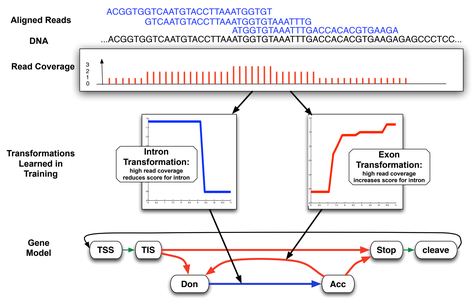
- mGene.NG: Gene finding system with RNA-Seq features:
  (+) rich set of sequence features (+) low expressed coding genes
- mTim: Segmentation of RNA-Seq coverage including splice sites:
  (+) less assumptions (+) noncoding transcripts
  (+) very accurate for sufficiently expressed transcripts

Comparison of transcript identification methods

- *C. elegans* RNA-Seq data from Waterston Lab, UW



De novo Gene Prediction (mGene)



Results

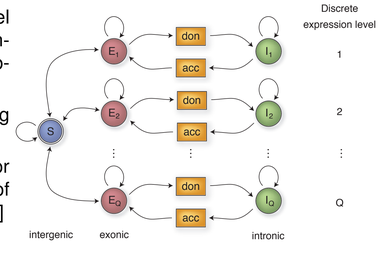
- Highly accurate *ab initio* predictions
- Impressive improvements with transcriptome measurements

A. thaliana	gene level		
	SN	SP	F
ab initio	71.7	74.8	73.3
RNA-Seq	80.6	82.2	81.4

`http://www.mgene.org/`

Segmentation of RNA-Seq Data (mTiM)

- Segmentation of read coverage data into exons, introns and intergenic regions
- State model with expression-dependent sub-models
- HM-SVM training algorithm
- Based on ideas for segmentation of tiling array data [5]



States (squares) model acceptor (acc) and donor (don) splice sites

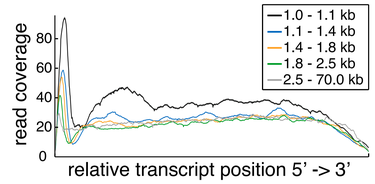
`http://www.fml.mpg.de/raetsch/suppl/mtim`

Transcript Quantification

RNA-Seq and Biases

The outcome of RNA-Seq depends on the experimental settings:

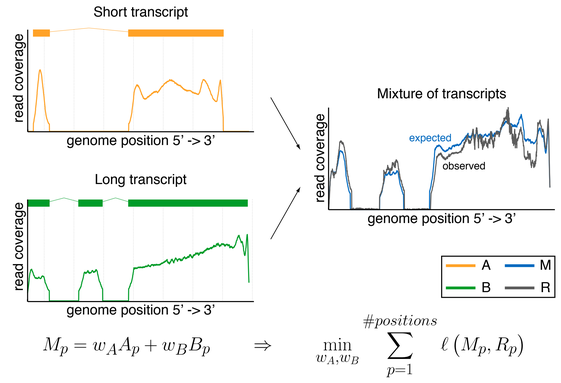
- cDNA library construction [7]
- Sequencing
- Read mapping



C. elegans SRX001872, R. Waterston Lab, UW

Transcript Quantification Problem (rQuant)

How can we infer transcript abundances from the observed read coverage? [6]



$$M_p = w_A A_p + w_B B_p \quad \Rightarrow \quad \min_{w_A, w_B} \sum_{p=1}^{\#positions} \ell\left(M_p, R_p\right)$$

`http://www.fml.mpg.de/raetsch/suppl/rquant`

References

[1] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch: Optimal Spliced Alignments of Short Sequence Reads. *Bioinformatics* 24(16):174–80 (2008).

[2] K. Schneeberger, J. Hagmann, S. Ossowski, N. Warthmann, S. Gesing, O. Kohlbacher, and D. Weigel: Simultaneous alignment of short reads against multiple genomes. *Genome Biology* 10(9):R98 (2009).

[3] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, N. Krüger, S. Sonnenburg, and G. Rätsch: mGene: Accurate SVM-based gene finding with an application to nematode genomes. *Genome Research* 19:2133-2143 (2009).

[4] G. Schweikert, J. Behr, A. Zien, G. Zeller, C. S. Ong, S. Sonnenburg, and G. Rätsch: mGene.web: a web service for accurate computational gene finding. *Nucleic Acids Research* 37:W312–W316 (2009).

[5] G. Zeller, S.R. Henz, S. Laubinger, D. Weigel, and G. Rätsch: Transcript normalization and segmentation of tiling array data. *Proceedings Pacific Symposium on Biocomputing* 13:527-538 (2008).

[6] R. Bohnert, J. Behr, and G. Rätsch: Transcript quantification with RNA-Seq data. *BMC Bioinformatics* 10(S13):P5 (2009).

[7] S. E. V. Linsen, E. de Wit, G. Janssens, S. Heater, L. Chapman, R. K. Parkin, B. Fritz, S. K. Wyman, E. de Bruijn, E. E. Voest, S. Kuersten, M. Tewari, and E. Cuppen: Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods* 6(7):474–476 (2009).

[8] The RNAseq Genome Annotation Assessment Project. `http://www.sanger.ac.uk/PostGenomics/encode/RGASP.html` (2009)

[9] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W.J. Kent, and A. Nekrutenko Galaxy: a platform for interactive large-scale genome analysis *Genome Research*, 15(10):1451-1455, 2005.