Easier Workflows & Tool comparison with oqtans+

Vipin T. Sreedharan <vipin@cbio.mskcc.org>
Rätsch Lab cBio, Memorial Sloan-Kettering Cancer Center, New York, USA.
MLB AG-Rätsch, FML of the Max Planck Society, Tübingen, Germany.
Center for Bioinformatics, University of Tübingen, Germany.

Galaxy Community Conference, 26 July 2012, Chicago
Deep sequencing data

Specificities:

- Rapidity of production
- Low cost
- Small fragments (reads)
Deep sequencing data

Applied to many scientific contexts:
RNA-Seq Analysis

Common analysis tasks

- compare two samples (wild type, mutant)
- identify new transcripts

Tang et al. 2011 *Nature Methods*
Grabherr et al. 2011 *Nature Biotech*
Li et al. 2011 *Science*
Gerstein et al. 2010 *Science*

Lee et al. 2011 *Nucleic Acids Res*
Yamashita et al. 2011 *Genome Res*
Daines et al. 2011 *Genome Res*
Ramani et al. 2010 *Genome Res*
RNA-Seq Analysis

- Sequencing
- Read Mapping
- Transcript Prediction
- Quantification

Common issues
- Reproducibility
- Tool availability
- Scalability
Schultheiss et al. 2011 BMC Bioinformatics

*developed by Rätsch Lab, at cBio MSKCC
Transcriptome analysis toolsuite

- **Sequencing** → **PALMapper** → **Transcript Prediction** → **Quantification**

- **PALMapper**: highly accurate short-read mapper using base quality and splice site predictions

  G. Jean et al. 2010 *Curr Protoc Bioinformatics*
mTIM: reconstructs exon-intron structure from alignments and splice site predictions

oqtans+
Transcriptome analysis toolsuite

- **rQuant**: estimates biases in library prep, sequencing, and read mapping; accurately determines the abundances of transcripts

R. Bohnert & G. Rätsch 2010 *Nucleic Acids Res*
rDiff/DESeq: determines significant differences in transcript/gene expression between experiments using statistical tests

O. Stegle et al. 2010 Nature Preceedings
S. Anders and W. Huber 2010 Genome Biology
Accuracy of read alignments

- C. elegans
- 75 nt RNA-seq reads (24 million)
Intron and transcript accuracy evaluation

- **Sequencing**
  - PALMapper
  - TopHat
  - mTIM
  - Cufflinks

- **C. elegans**
- 75 nt RNA-seq reads (24 million)

**F-Score**
- Introns: 0.58
- Transcripts: 0.37
oqtans+ Workflow

- Multiple reference genomes and transcriptomes for *A. thaliana*

- Illumina, 78 nt RNA-seq reads
  - Columbia accession (Col-0) (1.2 million)
  - Canary Island accession (Can-0) (4.9 million)

Gan et al. 2011 Nature
oqtans+ on AWS cloud

- Compute resources: 2 xxlarge
- Time:
  - Alignments: 20 minutes
  - Quantitative analysis: 10 minutes
- Cost on Amazon EC2: approx. $2.82
# oqtans+: Package contents

<table>
<thead>
<tr>
<th>Read Mapping</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>PALMapper</td>
<td>0.4</td>
</tr>
<tr>
<td>BWA</td>
<td>0.5.7</td>
</tr>
<tr>
<td>TopHat</td>
<td>1.5.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transcript Prediction</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>mTIM</td>
<td>0.2</td>
</tr>
<tr>
<td>Cufflinks</td>
<td>1.3.0</td>
</tr>
<tr>
<td>Trinity</td>
<td>r2012-06-08</td>
</tr>
<tr>
<td>Scripture</td>
<td>Beta-2</td>
</tr>
<tr>
<td>ASP</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Read Alignment Filtering</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAFT</td>
<td>0.2</td>
</tr>
<tr>
<td>Multi-Mapper Resolution</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantification</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>rQuant</td>
<td>2.2</td>
</tr>
<tr>
<td>rDiff</td>
<td>0.2</td>
</tr>
<tr>
<td>Cufflinks/Cuffdiff</td>
<td>1.3.0</td>
</tr>
<tr>
<td>DESeq</td>
<td>1.6.1</td>
</tr>
<tr>
<td>topGO</td>
<td>0.1</td>
</tr>
<tr>
<td>Genesetter</td>
<td>0.1</td>
</tr>
</tbody>
</table>
How to resolve the requirements of OS specific packages?
oqtans+ Availability: Our Server

- Public compute cluster
- 12 nodes, 112 CPUs
- our Galaxy test instance
- All tools described here, and more!

http://bioweb.me/mlb-galaxy
oqtans+ Availability: Source Code

- Free, open-source packages of our own tools
  - Including Galaxy Tool Wrappers
  - [http://oqtans.org](http://oqtans.org)
- Fabric scripts to install on any Galaxy instance
- Community Tool Shed
  - [http://toolshed.g2.bx.psu.edu](http://toolshed.g2.bx.psu.edu)
oqtans+ Availability: Cloud Computing

- Demo cloud instance with all oqtans+ tools
  - [http://cloud.oqtans.org](http://cloud.oqtans.org)
- AMI at Amazon Web Services for EC2
- Cloudman to launch any number of instances as a compute cluster
http://oqtans.org

Jonas Behr, Regina Bohnert, Philipp Drewe, Nico Görnitz, Géraldine Jean

André Kahles, Pramod Mudra, Sebastian Schultheiss, Georg Zeller, Gunnar Rätsch