oqtans

A Galaxy-Integrated Workflow for
Quantitative Transcriptome Analysis
from NGS Data

Sebastian J. Schultheiss <sebi@tue.mpg.de>
Machine Learning in Biology, Rätsch Lab, FML of the Max Planck Society
Tübingen, Germany

Presented at the 7th ISCB Student Council Symposium 2011, Vienna, Austria
Web Services Availability

- 927 web services (NAR Web Server Issues)
- Check availability of every service, 4 times
- Survey among authors (274 respondents)
- Problems the authors perceive:
  - web address change (28%)
  - missing example data (30%)
  - program on server not functional (>17%)
  - undocumented changes (>1%)
Web Service Availability

- 927 web services (NAR Web Server Issues)
- Check availability of every service, 4 times
- Survey among authors (274 respondents)
  - Problems the authors perceive:
    - web address change (28%)
    - missing example data (30%)
    - program on server not functional (>17%)
    - undocumented changes (>1%)

Caution: Published results may not be reproducible
Galaxy Approach

- Persistent, transparent, reproducible approach to bioinformatics research
- Integration made simple with xml wrappers of command line tools
- Accessible: web service, download, cloud

J. Goecks et al. 2010
D. Blankenberg et al. 2010
E. Afgan et al. 2010
S. Koskovsky Pond et al. 2009

W. Miller et al. 2007
J. Taylor et al. 2007
D. Blankenberg et al. 2007
M. Giardine et al. 2005
Common analysis tasks

- compare two samples (wildtype, mutant)
- identify new transcripts
Galaxy Tools

- Sequencing
- Read Mapping
  - PALMapper*
  - BWA
  - TopHat
- Transcript Prediction
  - mTIM*
  - SplAdder*
  - ASP*
  - mGene.web*
  - Cufflinks
  - GFF Toolikt*
- Quantification
  - rQuant*
  - rDiff*
  - Cufflinks/Cuffdiff
  - DESeq
  - GOrilla
  - SAFT*

* developed by the MLB Rätsch Research Group, FML, Tübingen
PALMapper: highly accurate short-read mapper using base quality and splice site predictions

G. Jean et al. 2010
oqtans

Machine Learning Tools

- **mTIM**: reconstructs exon-intron structure from alignments and splice site predictions
- **SplAdder**: adds isoforms to known annotation based on splice graph

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

R. Bohnert et al. 2009
rDiff: determines significant differences in transcript expression between experiments using statistical tests

O. Stegle et al. 2010
oqtans Workflow in Galaxy

Map with PALMapper
- FASTQ file
- Acceptor splice site predictions in SPF format
- Donor splice site predictions in SPF format
  - unspliced_output (bed)
  - spliced_output (bed)
  - sam_output (sam)
  - bam_output (bam)
  - logfile (txt)
  - coverage_output (wig)
  - junctions_output (gff3)

Select first
- from
  - out_file1

Select last
- from
  - out_file1

rDiff
- Genome annotation
- Sample 1 BAM alignment file
- Sample 2 BAM alignment file
  - rdiff_out (txt)
  - Log_File (txt)

Predict using mTIM
- Genome sequence in FASTA format
- RNA-seq read alignments in BAM format
- Acceptor splice site predictions in SPF format
- Donor splice site predictions in SPF format
  - mtim_gff3_out (gff3)
  - mtim_gann_out (gann)
  - Log_File (txt)

rQuant
- Genome annotation in GFF3 format
- BAM alignment file
  - rquant_out (gff3)
  - profiles_fn_out (txt)
  - intron_dists_fn_out (txt)
  - Log_File (txt)
oqtans Availability: Our Server

- External compute cluster
- 21 nodes, 168 CPUs
- our Galaxy instance
- All tools described here, and more!

http://galaxy.fml.mpg.de
oqtans Availability: Source Code

- Packaged releases of our own tools
  - Including Galaxy Tool Wrappers
  - http://oqtans.org
- Fabric scripts to install on any Galaxy instance
- Community Tool Shed
  http://community.g2.bx.psu.edu/
MLB group tools into any Galaxy installation

Python Fabric scripts: used to manage automation of a remote server

```python
install_cmd("wget %s" % self.tool_env['url'])
install_cmd('chown -R %s %s' % (env.user, install_dir_root))
```

Adjusted to Ubuntu on Galaxy Cloud Image
oqtans Availability: Cloud Computing

- Demo cloud instance with all oqtans tools
- AMI at Amazon Web Services for EC2
  - Cloudman to launch any number of instances as a compute cluster
- AMI in your own virtualizer (e.g. Virtual Box)
  - search for “oqtans”
oqtans Availability: Cloud Computing

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove 'worker' nodes for running jobs.

- Terminate cluster
- Add nodes
- Remove nodes
- Access Galaxy

Status

- Cluster name: OqtansFMLCluster
- Disk status: 2.7G / 20G (14%)
- Worker status: Idle: 0 Available: 0 Requested: 0
- Service status: Applications · Data ·

Autoscaling is off. Turn on?