Efficient tool deployment to the Galaxy Cloud: An RNA-seq workflow case study

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

- 927 web services (NAR Web Server Issues)
- Collection of 45 data on every service
- Survey among authors
- Problems:
  - URL change
  - missing example data
  - program on server not functional
  - undocumented changes from updates

Schultheiss et al. 2011, PLoS ONE i.r.
Web Service Availability

- Collection of 45 data on every service
- Survey among authors
- Problems:
  - URL change
  - Missing example data
  - Program on server not functional
  - Undocumented changes from updates

Caution: Published results may not be reproducible
Web Services Availability

- Redirected from URL stated in abstract: 13%
- Unreachable, only via search engine: 7%
- Unreachable and not found: 9%
- Correct URL in abstract: 72%

Evaluations:
- June 2009
- August 2010
- October 2010

Schultheiss et al. 2011, PLoS ONE i.r.
Galaxy Approach

- Persistent, reproducible approach to bioinformatics research
- Integration of tools made simple
- Source code release, VM/AMI, cloud instances, Galaxy pages

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
Giardine et al. 2005
Our Galaxy Tools:

- Machine Learning-powered for quantitative analyses of RNA-seq experiments
- Workflow can be adjusted to your needs
Common experimental setups:

- Identification of new transcripts
- Comparison of samples
- **PALMapper**: highly accurate short-read mapper using base quality and splice site predictions

G. Jean et al. 2010
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
Performance at or above state-of-the-art

Tophat, DESeq, Cufflinks, CuffDiff for comparison

A. Roberts et al. 2011
S. Anders and W. Huber 2010
C. Trapnell et al. 2010
MLB group tools into any Galaxy installation

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

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

Available from Galaxy Tool Shed

Adjusted to Ubuntu on Galaxy Cloud Image
Tools with Fabric Scripts

Joint work: Enis Afgan & Galaxy Team
James Taylor, Anton Nekrutenko
AG Rätsch

- Machine Learning Toolbox *Shogun*
- EasySVM tools for easy-to-use SVM classifications
- All oqtans tools
- GFF Toolkit for pre-processing annotations
- KIRMES regulatory modules identification
- WebLogo interface
External cluster (21 nodes, 168 CPUs) handles requests to our Galaxy instance

Fabrics to install tools from the shed

AMI instance created

‘Instant on’

Cloudman to launch as many as you need
oqtans Availability

- MLB Group Galaxy Instance
  http://galaxy.fml.mpg.de

- EC2 Cloud Instance & AMI
  ami-228a7e4b “Oqtans-Galaxy”

- Fabric scripts install Oqtans on your Galaxy
  http://community.g2.bx.psu.edu/

- Source code releases of all tools, email info
  http://fml.mpg.de/oqtans