High level distributed processing pipelines with Galaxy

Brad Chapman, Bioinformatics Core at Harvard School of Public Health

Galaxy Community Conference, 27 July 2012
Complex pipelines

Goal: simple interface to complex analyses

- Steps, lots of them
- Intermediate files
- Branching logic: if/else
- Idempotent
- Transactional
- Parallel: by record, by region
- Associated resources
- Experimental metadata

http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3
(At least) two types of users
Levels of abstraction

- Code
- Libraries
- Command Line Tools
- Web Interface wrapper
- Workflow
- Complex Custom Pipeline

Most control and most complexity  
Purpose built
Approaches

- Custom Galaxy interfaces
  - Variant calling pipeline
  - Galaxy views and controllers
- Web tools
  - BioCloudCentral
  - CloudMan
- External data upload
  - ISA-Tab experiments
  - Data Libraries and API
- Interoperable tools
  - Variant analysis platform
  - GenomeSpace
Variant pipeline interface

https://bitbucket.org/hbc/galaxy-central-hbc
Implementation approach

- Integrated into Galaxy fork
- Custom view (http://your.galaxy.org/pipeline)
- Custom controller
- Calls external server for processing
- Results upload to Galaxy Data Libraries

Demo with movies: http://j.mp/uNXZY6
Backend variant processing
### BioCloudCentral

Easily launch [CloudMan](http://biocloudcentral.org), [CloudBioLinux](http://biocloudcentral.org) and [Galaxy](http://biocloudcentral.org) platforms on Cloud Computing resources (including [Amazon Web Services](http://biocloudcentral.org)).

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster name</td>
<td>Name of your cluster used for identification. This can be any name you choose.</td>
</tr>
<tr>
<td>Password</td>
<td>Your choice of password, for the CloudMan web interface and accessing the instance via ssh or FreeNX.</td>
</tr>
<tr>
<td>Cloud</td>
<td>Choose from the available clouds. The credentials you provide below must match (ie, exist on) the chosen cloud.</td>
</tr>
<tr>
<td>Access key</td>
<td>Your Access Key ID. For the Amazon cloud, available from the security credentials page.</td>
</tr>
<tr>
<td>Secret key</td>
<td>Your Secret Access Key. For the Amazon cloud, also available from the security credentials page.</td>
</tr>
<tr>
<td>Instance type</td>
<td>Type (ie, virtual hardware configuration) of the instance to start.</td>
</tr>
</tbody>
</table>

[Show advanced startup options](http://biocloudcentral.org)

[Start an instance](http://biocloudcentral.org)
CloudMan: automate setup

http://usecloudman.org/
## Run analysis

<table>
<thead>
<tr>
<th>Analysis type</th>
<th>Pipeline</th>
<th>Variant calling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data</td>
<td>test.fastq</td>
</tr>
<tr>
<td></td>
<td>Multiplex</td>
<td>Illumina</td>
</tr>
<tr>
<td></td>
<td>Coverage Depth</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>Genome regions targeted</td>
<td>exome</td>
</tr>
<tr>
<td></td>
<td>Organism</td>
<td>hg19</td>
</tr>
<tr>
<td></td>
<td>Bait file</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Target file</td>
<td></td>
</tr>
</tbody>
</table>

Results will be available as a folder in the Shared Data Libraries, chapmanb@50mail.com/120722_WHVGGL.
Experimental metadata: ISA-Tab

http://isatab.sourceforge.net/
Data Library integration

Stem Cell Discovery Engine:
http://discovery.hsci.harvard.edu/

https://github.com/hbc/projects/blob/master/scde_deploy/scripts/bii_datasets_to_galaxy.py
Interoperable tools

- https://github.com/chapmanb/bcbio.variation
- http://validationprotocol.org
- http://genomespace.org
Answers and Questions

- Build off core Galaxy features
  - Data Libraries
  - API
  - CloudMan
- Maximize interoperability
- Automate aggressively

Would love to hear your experiences.